THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Understanding the genetic and environmental variations in human complex traits

Huanwei Wang

Bachelor of Science, Master of Science

0000-0002-6137-3391

A thesis submitted for the degree of Doctor of Philosophy at The University of Queensland in 2021

Institute for Molecular Bioscience

Abstract

Most human traits are complex as they are affected by many genetic and environmental factors as well as the potential interactions between them. In the field of quantitative genetics, it has been formulated that the phenotypic value of a complex trait (P) can be partitioned into the genetic component (G), the environmental component (E), and the genotype-by-environment interaction component (I_{GE}). Furthermore, the genetic variance component can be partitioned into the additive genetic component, the dominance genetic component, and the epistatic genetic component. Prior work in quantitative genetics has provided us powerful tools to understand human complex traits from more than a century ago (Chapter 1).

The genome-wide association study (GWAS), an experimental design to associate a trait of interest with genetic variants across the genome, has developed rapidly during the last decade, due to the advancement of genotyping technologies and the large samples accumulated through biobanks and research consortia. Apart from identifying trait-associated genetic variants, GWASs provide tremendous resources to answer many old but important questions in quantitative genetics, including estimating the proportion of phenotypic variance attributable to the genetic component (i.e., heritability estimation), estimating the genetic correlation between two traits, inferring causal relationship between exposure and outcome traits, predicting the trait based on genetic information, and so on (Chapter 1).

In this thesis, the original research results of three projects have been included. Firstly, I performed a genome-wide variance quantitative trait locus (vQTL) analysis to associate the genetic variants with the variance of a phenotype and demonstrated the identification of vQTLs can be used to infer genotype-by-environmental interaction (GEI) without environmental data (Chapter 2). Secondly, I quantified the inflation in the test-statistics for two interaction effects, i.e., GEI and genotype-by-genotype interaction, by theoretical derivation and simulation study, respectively (Chapter 3). Thirdly, I explored methods to integrate both the genetic and environmental information to improve the accuracy of phenotype prediction (Chapter 4).

The final chapter included the summary of the findings and a discussion of related future directions (Chapter 5).

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications included in this thesis

1. **Wang, H.**, Zhang, F., Zeng, J., Wu, Y., Kemper, K.E., Xue, A., Zhang, M., Powell, J.E., Goddard, M.E., Wray, N.R., Visscher, P.M., McRae, A.F., Yang, J.. Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. Science advances, 2019, 5(8), p.eaaw3538[1].

Submitted manuscripts included in this thesis

1. Hemani, G., Powell, J.E., **Wang, H.**, Shakhbazov, K., Westra, H., Esko, T., Henders, A.K., McRae, A.F., Martin, N.G., Metspalu, A., Franke, L., Montgomery, G.W., Goddard, M.E., Gibson, G., Yang, J., Visscher, P.M.. Testing for genetic interactions with imperfect information about additive causal effects. Nature (under review).

Other publications during candidature

Peer-reviewed papers

1. **Wang, H.**, Zhang, F., Zeng, J., Wu, Y., Kemper, K.E., Xue, A., Zhang, M., Powell, J.E., Goddard, M.E., Wray, N.R., Visscher, P.M., McRae, A.F., Yang, J. (2019) Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. Science Advances, 5(8), p.eaaw3538.[1]
2. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., **Wang, H.**, Zheng, Z., Magi, R., Esko, T., Metspalu, A., Wray, N.R., Goddard, M.E., Yang, J., Visscher, P.M. (2019) Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nature Communications, 10(1), pp. 5086.[2]
3. Revez, J.A., Lin, T., Qiao, Z., Xue, A., Holtz, Y., Zhu, Z., Zeng, J., **Wang, H.**, Sidorenko, J.J., Kemper, K.E., Vinkhuyzen, A.A., Frater, J., Eyles, D., Burne, T.H., Mitchell, B., Martin, N.G., Zhu, G., Visscher, P.M., Yang, J., Wray, N.R., McGrath, J.J. (2020) Genome-wide association study identifies 143 loci associated with 25 hydroxyvitamin D concentration. Nature Communications, 11(1), p. 1647.[3]
4. Wu, Y., Qi, T., **Wang, H.**, Zhang, F., Zheng, Z., Phillips-Cremins, J. E., Deary, I. J., McRae, A. F., Wray, N. R., Zeng, J., Yang, J. (2020) Promoter-anchored chromatin interactions predicted from genetic analysis of epigenomic data. Nature Communications, 11(1): p. 2061.[4]
5. Zeng J., Xue A., Jiang L., Lloyd-Jones L. R., Wu Y., Wang H., Zheng Z., Yengo L.,

Kemper K. E., Goddard M. E., Wray N. R., Visscher P. M., Yang J. (2021) Widespread signatures of natural selection across human complex traits and functional genomic categories. Nature Communications, 12: p. 1164.[5]

Conference abstracts

1. Quantifying the inflation in test-statistics for epistasis due to imperfect tagging using whole-genome sequence data (lightning talk). GeneMappers Conference 2018. Queensland, Australia.
2. Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank (Poster presentation). Gorden Research Conference Quantitative Genetics and Genomics 2019. Tuscany, Italy.
3. Integrating the genetic and environmental information to improve the phenotype prediction for body mass index (Poster presentation). The 6th International Conference of Quantitative Genetics Virtual 2020.

Contributions by others to the thesis

Several other people have made significant contributions to this thesis, including my principal advisor Jian Yang and all the co-authors including Futao Zhang, Jian Zeng, Yang Wu, Kathryn E. Kemper, Angli Xue, Longda Jiang, Julia Sidorenko, Min Zhang, Joseph E. Powell, Michael E. Goddard, Naomi R. Wray, Peter M. Visscher, Allan F. McRae.

Statement of parts of the thesis submitted to qualify for the award of another degree

No works submitted towards another degree have been included in this thesis.

Research involving human or animal subjects

No animal or human subjects were involved in this research.

Acknowledgments

My PhD study cannot be finished without the help from many amazing people. Firstly, I would like to give my sincere gratitude to my principal advisor Jian Yang, who provided excellent guidance for all three research projects included in this thesis, practical feedback and suggestion on the research progress, and valuable encourage to overcome the difficulties in both my work and life.

I also would like to thank my associate advisor Allan McRae, who is always supportive and friendly. Thank Naomi Wray and Lachlan Coin for being the committee members for all my three milestones. Thank all co-authors for their help for my research projects, including Futao Zhang, Jian Zeng, Yang Wu, Kathryn E. Kemper, Angli Xue, Longda Jiang, Julia Sidorenko, Min Zhang, Joseph E. Powell, Michael E. Goddard, Naomi R. Wray, Peter M. Visscher, Allan F. McRae, and Jian Yang.

I feel really honored and grateful to be part of PCTG, such a wonderful team with so many talented and extraordinary colleagues, where we can communicate, work and learn freely with each other regardless of possible existing language and culture barriers.

Acknowledge also goes to many supports from IMB and UQ, including the HDR officer Amanda Carozzi, the IT team from IMB and RCC, and all kinds of training provided by UQ student service/library/graduate school.

Finally, I would like to thank my wife Ranran Zhang for her company from China to Australia and my parents for their lifelong support.

Financial support

This research was supported by University of Queensland Training Tuition Fee Offset, University of Queensland Research Training Stipend, Research Higher Degree Top Up Scholarships, and the Australian National Health and Medical Research Council (1113400).

Keywords

Quantitative genetics, genome-wide association study, genotype-by-environment interaction, variance quantitative trait locus, epistasis, genetic prediction, polygenetic risk score

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060412, Quantitative Genetics, 100%

Fields of Research (FoR) Classification

FoR code: 0604, Genetics, 100%

Table of Content

List of Figures

List of Tables

List of Abbreviations used in the thesis

Abbreviations	
1KG3	1000 genomes project phase 3
ACC/AHA	American College of Cardiology/American Heart Association
ANOVA	Analysis of variance
BFP	Body fat percentage
BMD	Heel bone mineral density T-score, automated
BMI	Body mass index
BMR	Basal metabolic rate
BW	Birth weight
DET	Dispersion effect test
DGLM	Double generalized linear model
ERS	Environmental risk score
FEV1	Forced expiratory volume in 1-second
FFR	FEV1 and FVC ratio
FK test	Fligner-Killen test
FPR	False positive rate
FVC	Forced vital capacity
G6PDD	Gluscose-6-phosphate dehydrogenase deficiency
GEI	Genotype-by-environment interaction
GERS	Genetic and environmental risk score
GIF	Genomic inflation factor
GREML	GRM restricted maximum likelihood
GREML-LDMS	MAF and LD stratified GREML
GREML-MC	GREML with multiple components
GREML-SC	GREML with single component
GRM	Genomic relatedness matrix
GRS	Genetic risk score
GWAS	Genome-wide association study
GWS	Genome-wide significant
HC	Hip circumference
HE regression	Haseman-Elston regression
HRS	Haplotype reference consortium

HT	Standing height
IPAQ	International physical activity questionnaire
LD	Linkage disequilibrium
LDSR	LD score regression
LMM	Linear mixed model
LRT	Likelihood ratio test
MAF	Minor allele frequency
MGRS	Multiple GRS
MLR	Multiple linear regression
MPS	Multiple polygenic risk score
MR	Mendelian Randomization
NCI	Net reclassification improvement
PA	Physical activity
PC	Principal component
PKU	Phenylketonuria
PRS	Polygenic risk score
QTL	Quantitative trait locus
RINT	Rank-based inverse-normal transformation
SB	Sedentary behavior
SNP	Single nucleotide polymorphism
UKB	UK Biobank
vQTL	Variance quantitative trait locus
WC	Waist circumference
WGS	Whole genome sequence
WHR	Waist to Hip Ratio
WHRadjBMI	WHR adjusted for BMI
WTCCC	Wellcome Trust Case Control Consortium
XP	Xeroderma pigmentosum

1

Chapter 1: Introduction

Most human traits are complex as they are affected by many genetic and environmental factors as well as the potential interactions between them[6,7]. Quantitative genetics, founded more than one century ago[8,9], provides us great tools to understand human complex traits. In this chapter, I will introduce the basic concepts of quantitative genetics and its application in the era of genome-wide association study (GWAS). I will also introduce the background of three specific topics involved in the research studies presented in chapters 2-4, which are genotype-by-environment interaction (GEI), epistasis, and genetic prediction.

1.1 Quantitative genetics

Quantitative genetics focuses on the genetic study of quantitative traits that are continuously distributed in the population (e.g. human height) in contrast to discrete or qualitative traits (e.g. pea color)[6,7]; furthermore, the theories and methods developed in quantitative genetics can also be extended for complex diseases[10], such as diabetes[11], cardiovascular disease[12], and psychiatric disease[13]. The heredity of quantitative traits usually involves many genetic variants (e.g. single nucleotide polymorphism or SNPs) with small effects in comparison with Mendelian traits[14] that are controlled by one or a few genetic variants with large effects.

The foundation of quantitative genetics can be traced back to more than one century ago[8,9]. With the rediscovery of Mendel's laws in the 1900s and also the work of Francis Galton in the 1880s[15], there were conflicting views between Mendelians supporting discontinuous evolution and biometricians supporting Darwinian evolution[16-18]. Fisher's 1918 paper[8] proposed the model of a very large number of Mendelian genes each with small effects (also called infinitesimal model later)[17] and finally reconciled the debate. The paper is seen as the founding paper of quantitative genetics and also introduces a series of statistical concepts (e.g. variance), which are still widely used in modern genetics and statistics.

In quantitative genetics, the phenotypic value or its variation is partitioned into different components which can be explained by observable data, such as genetic factors or environmental factors. The phenotypic value of a complex trait (P) is partitioned into genotype (G) and environment (E) components with the interaction between them (I_{GE}):

$$P = G + E + I_{GE}$$

Thus, the variance of a complex trait (V_P) is partitioned into components attributable to genotype (V_G), environment (V_E) and GEI effect (V_{GE}) with the covariance between genotype and environment (Cov_{GE}):

$$V_P = V_G + V_E + V_{GE} + 2Cov_{GE}$$

Here the genotype-environment correlation and GEI are different[19], although both terms are often ignored in genetic analysis of human complex traits. More discussion about GEI can be found in section 1.3 in Chapter 1, Chapter 2, and Chapter 3.

Furthermore, the genetic variance (V_G) is partitioned into the variance of additive (V_A), of dominance (interaction between two alleles at a single locus, V_D), and of epistatic (interaction between two or more loci, V_I) genetic effects. The dominance and epistatic genetic effects are also called non-additive genetic effects.

$$V_G = V_A + V_D + V_I$$

Then heritability[20] is defined as the proportion of the total phenotypic variance explained by genetic variance (broad-sense heritability, H^2) or additive genetic variance (narrow-sense heritability, h^2):

$$H^2 = \frac{V_G}{V_P}$$

$$h^2 = \frac{V_A}{V_P}$$

Heritability, usually referred to as narrow-sense heritability, is an important parameter in genetics as it defines the upper boundary of the accuracy of genetic prediction and is related to the response to selection. It can be estimated without any knowledge about the specific genes or genetic variants[20]. Traditionally, heritability is often estimated from the resemblances between relatives, such as regression between monozygotic and dizygotic twin pairs or parents and offspring pairs in simple and balanced designs, or a linear mixed model in more complicated and unbalanced designs[7,20]. For example, Polderman et al. meta-analyzed the estimated heritability from 2,748 papers, published from 1958 to 2012 including 14,558,903 monozygotic or dizygotic twin pairs for 17,804 traits, and found the mean heritability across all traits is 49%[21]. Another example is Lakhani et al. that estimated heritability using 56,396 twin pairs and 724,513 sibling pairs from health insurance data and found the mean estimated heritability is 31.1% across 560 disease-related phenotypes[22].

1.2 GWAS

Another equally (if not more) important goal for the genetic study of human complex traits is to identify the specific genetic variants influencing the traits. Early studies with either linkage analysis or candidate gene association analysis were largely unsuccessful[23]. Then people started to performed genome-wide association study (GWAS)[24,25] with genetic markers across the whole genome, facilitated by the efforts of consortiums like HapMap project[26] or 1000 Genomes Project[27] and technological innovation of genotyping arrays. GWAS can identify robust and replicable genotype-phenotype associations in a prior-free manner. While the first GWAS can be traced back to 2002[24], sometimes the Wellcome Trust Case Control Consortium (WTCCC) paper published in 2007[28] was taken as the start of GWAS[29,30] due to its good design and largest sample size at the time.

The basic statistical method of GWAS is a simple univariate linear or logistic regression model to associate genetic variants (usually SNPs) across the genome with a trait of interest. More sophisticated statistical methods based on linear mixed model (LMM) are used to control the confounders, such as the population stratification and relatedness [31-33], and handle large biobank-scale data[34-36]. To control the false positives, the data need to be quality controlled before the association test on individual level, genotype level, and phenotype level[37], and a genome-wide significant (GWS) threshold of p value $<5\times10^{-8}$ is usually used. To improve the power, the genetic variants not captured by genotyping arrays are usually imputed to a sequenced reference panel[38,39], and non-heritable covariates, such as age, sex, genotyping batch, need to be pre-fitted or fitted in the model[40]. Another way to increase the power of GWAS is to increase the sample size via establishing big consortia (such as Psychiatric GWAS Consortium (PGC)[41], GIANT consortium[42,43], SSGSC consortium[44]) to meta-analyse[45] data from multiple cohorts or single large biobank (such as the UK Biobank[46]). As of 13 August 2020, there are 196,813 variant-trait associations passing a GWS threshold[47] from 4,671 publications curated by the GWAS Catalog database[48].

After the GWS variants are initially identified, people find that the phenotypic variance explained by the GWS variants is much smaller than the heritability estimated from family or twin studies using traditional quantitative genetic methods, which is the so-called "missing heritability" problem[49,50]. Taking human height as an example[51], the variance explained by around 42 GWS SNPs identified with a total sample size of around 60,000 in 2008 was 5% in contrast to the estimate of heritability of about 80% estimated from family or twin studies. In 2010, Yang et al[52,53] proposed a method called genomic relatedness matrix (GRM) restricted maximum likelihood (GREML), which included all SNPs to fit a linear mixed model. They estimated that ~45% of variance in height could be explained by all common SNPs, suggesting that the heritability was not missing but hidden because of a large number of genetic variants with effect sizes too small to be detected in the previous GWAS[52]. The heritability estimated by all genetic variants across the genome is then called the SNP-based heritability[54].

The initial GREML method only included a single random component to model the effect sizes of genome-wide genetic variants (called GREML-SC). The method was subsequently extended to partition the heritability into contributions from different chromosomes or multiple sets of genetic variants stratified by functional annotations (called GREML-MC)[55], to reduce the estimation bias by stratifying genetic variants by MAF and LD (GREML-LDMS) [56], or to estimate the SNP-based heritability for disease traits[57,58]. A latest study incorporating both common and rare genetic variants measured by whole genome sequencing (WGS) using the GREML-LDMS method recovered most of the pedigree-based heritability estimation for human height[59].

There are other methods developed to estimate the SNP-based heritability. The GREML approach needs individual-level genotype and phenotype data, which are usually inaccessible due to the privacy or logistical issues[60]. In contrast, GWAS summary statistics data, which only contains estimated effect sizes, standard errors, p values, sample sizes, etc., are more convenient to share and access. Linkage disequilibrium score regression (LDSR) method[61,62], also based on the GREML model, only requires GWAS summary statistics data with LD information from a reference sample of the same ethnicity and can dramatically reduce the computation time. Apart from GREML model, there are other methods based on different models, including the LDAK model with a SNP-specific variance assumption and related summary-statistics version SumHer [63], mixture distribution model using Bayesian statistics[64,65], generalized random effect model with biobank-level data[66], and Haseman-Elston (HE) regression[67,68]. In addition to additive genetic variance, the SNP-based heritability method can also be extended to estimate dominance[69,70] and epistatic[71] genetic variances.

GWAS data can also be used to study the relationship between two different traits. Genetic correlation is a parameter to quantify the correlation of the genetic component between two traits[72]. Many SNP-based heritability estimation methods can be expanded for genetic correlation estimation, including the GREML approach[73] and LD score regression[74,75]. The genetic correlation can be caused by many sources, including causality (sometimes called vertical pleiotropy), pleiotropy (sometimes called horizontal pleiotropy), and others (e.g., LD-induced genetic correlation)[72]. Mendelian Randomization (MR) is a statistical method to draw causal inference between two traits (exposure and outcome), which takes genetic variant(s) as instrumental variable(s) under some strong assumptions[76,77]. The methodology of MR is still under active development, from one sample to two samples, from one genetic instrument to multiple genetic instruments, from individual-level data to GWAS summary data, from single exposure method for one exposure and one outcome to multivariate MR method including multiple exposures, and also to better handle the correlated and uncorrelated pleiotropy effects[78].

1.3 Genotype-by-environment interactions (GEI)

The GEI is an important component in the quantitative genetic model. It can be defined as that the effect of a genetic variant on a phenotype depends on environmental factor(s), or alternatively, the effect of an environmental factor on a phenotype depends on the genotype(s)[79]. Or statistically, GEI can be defined as the departure of the joint effect of genetic factor(s) and environmental factor(s) from the sum of their marginal effects[80]. GEI could be quantitative or qualitative[80]. For quantitative GEI (also called "non-crossover" interaction[81]), different environments alter the level, but not the direction of the genetic effect; whereas for qualitative GEI (also called "crossover" interaction), different environments alter the direction of the genetic effect. The quantitative GEI is tied to the scale of measurement, which means any quantitative GEI can be removed through a non-linear transformation[81].

The concept of GEI is thought[82,83] to be traced back to Archibald Garrod's paper in 1902[84] and J. B. S. Haldane's book in 1938[85]. The classic and well-characterized GEI examples are usually for Mendelian traits or diseases[83,86], including mutations in gene *PAH* interacted with phenylalanine intake in the diet on phenylketonuria (PKU), xeroderma pigmentosum (XP) interacted with exposure to sunlight on the risk of skin cancer, and so on. There are also examples for complex traits or diseases, including *NAT2* with smoking on bladder cancer[87], *FTO* with physical activity on body mass index[88], and *ALDH2* with alcohol on esophageal cancer[89].

Identifying GEI effects in humans is difficult[80,82,90,91], despite the limited number of successful examples mentioned above. One reason is environmental factors are challenging to assess[80,92]. Firstly, the broadest definition of environmental factors could include any factors apart from genetic variants, including exogenous factors (such as air pollution), lifestyle factors (such as diet, smoking, or physical activity), medicine-taking history, and so on. Secondly, environmental factors are usually multidimensional. For example, there are at least three dimensions to assess drinking (i.e., the frequency of drinking, the typical quantity of drinking, and binge drinking). Thirdly, many environmental factors change with time and are hard to record during a life-long course. Finally, there are usually measurement errors for environmental factors, especially using self-reported questionnaires.

Many statistical methods are developed to detect GEI effects and applicable depending on different assumptions and contexts[93]. One method is called vQTL (to be discussed in Chapter 2), which can be used to infer GEI, although there could be other sources contributing to phenotypic variability, including epistasis, environmental sensitivity, temporal fluctuation, and measurement errors[94].

There are also methods developed to estimate the overall contribution of GEI effects to phenotypic variance. Robinson et al. in 2017[95] used the method proposed by Yang et al.[53] (called GCI-GREML model), including a single genetic component and a GEI component with a discrete environmental factor or a continuous environmental factor stratified into discrete groups, and estimated that the genotype-age interaction contributes 8.1% of BMI variation and genotype-smoking interaction contributes 4.0% of BMI variation. Ni et al. in 2019 proposed a multivariate reaction norm model (MRNM)[96], which accounts for both GEI and genotype-environment correlation and allows a continuous environmental factor. The MRNM model was subsequently extended to be applicable for GWAS summary data (called GxEsum[97]). Dahl et al. in 2020 proposed a GxEMM model, which could accommodate arbitrary environmental factors and binary traits[98]. Kerin et al. in 2020[99] proposed a (Linear Environment Mixed Model Analysis) LEMMA model to allow multiple environmental factors using an environmental score (called ES) by a linear combination.

1.4 Epistasis

Epistasis, the genetic interaction effect between two or more loci, is another component in Fisher's partition of genetic value or variance (see section 1.1 above), although this term was first used by Bateson in 1909[100,101]. It can be additive-by-additive, additive-by-dominance, dominance-by-dominance, or high-order epistasis with more than two loci.

It is difficult to identify the genetic variants with epistatic effects[102,103]. Firstly, the number of statistical tests for epistasis is proportional to the square of the total number of genetic variants, so it will create a high burden for multiple test correction. Secondly, in comparison with the additive effect that a marker variant can explain a proportion of LD r^2 of the genetic variance, a marker variant can explain a proportion of r^4 of the genetic variance for epistatic effect[104]. Therefore, a much larger sample size is needed to have enough power to detect genetic variants with epistatic effects. Thirdly, the statistical test methods for epistasis could be biased by other factors, like linkage disequilibrium (see Chapter 3 for one example of the inflation of epistasis test).

The overall contribution to phenotypic variation by epistatic effects in humans is expected to be low theoretically[105]. Because the three components in Fisher's partition of the genetic variation are not independent. The dominance and epistatic effects in the level of gene action can be largely captured by the additive component in the level of variance in the population, even when the epistatic value in gene action is high[105]. One latest empirical analysis is the meta-analysis of 14,558,903 monozygotic (MZ) and dizygotic (DZ) twin pairs by Polderman et al[21]. They found the correlations of MZ twins for most traits were close to twice of the correlations of DZ twins, which was consistent with a model that the genetic variance was mainly due to the additive genetic effect. Another latest estimation using SNP genotypes in 254,679 unrelated individuals from UK Biobank did not find evidence for an epistatic variance for 70 traits, although the sampling variances were large[71].

While the evidence in the literature about epistasis in humans is weak, there are multiple lines of evidence for epistatic effects from more controllable model species[106,107]. For example, Domingo et al.[108] conducted an experiment covering 5,184 genotypes in 10 positions of tRNA gene using yeast and found widespread pairwise and high-order epistatic effects on fitness. And another intuitive reason for the existence of epistasis is the interconnected complicated biological system, including protein-protein interactions and functional pathways[106].

1.5 Prediction of complex traits

Apart from partitioning the phenotypic variance into genetic and environmental components and identifying the genetic variants associated with traits, another active research direction is to predict human complex traits or diseases, which is key to achieve the goal of personalized and precision medicine[109,110]. In the early years of the GWAS era, genetic predictors were built based on GWS loci and shown relatively low prediction accuracy, mainly because only a small proportion of phenotypic variance was explained by the GWS loci[111]. To overcome this limitation, genetic risk prediction based on a polygenic model was proposed[112] and used[113], which accumulated the genetic effects of many variants across the genome.

The simplest way to construct a genetic predictor (called polygenic risk score, PRS) across M independent genetic markers is the weighted sum of genotype values ($x_1 \dots x_M$):

$$\hat{y}_M = \sum_{i=1}^{M} b_i x_i$$

where b_i is the estimated genetic SNP effect by GWAS. Then the prediction accuracy evaluated by the proportion of variance explained by the genetic predictor (R^2) can be quantified as[114,115]:

$$R^2 = \frac{h_M^2}{1 + M/(Nh_M^2)} < h_M^2$$

where N is the discovery sample size and h_M^2 is the variance explained by the genetic markers included. So the accuracy of genetic prediction can be improved by increasing the sample size (*N*) with an upper boundary of SNP-heritability (h_M^2).

In addition, more sophisticated statistical methods and software are developed to improve the prediction accuracy. The simple method is called P+T method[113], which selects SNPs by LD pruning and p-value thresholding. A validation dataset is used to test a range of p-value thresholds and choose the p-value with the highest prediction accuracy. BLUP[116] (i.e. best linear unbiased predictor) is a method to estimate the parameters in the traditional linear mixed model and then extended to create a genetic predictor based on GWAS summary data (also called SBLUP[117]). LDpred[118] is a Bayesian method using a point-normal mixture prior to model the GWAS summary data and the extended LDpred-funct[119] claims to further improve the prediction accuracy by incorporating the functional annotations. BayesR[120] and the following summary-data version SBayesR[2] are also Bayesian methods based on a mixture distribution with multiple components (four components as default). Other genetic prediction methods include lassosum[121], PRS-CS[122], and NPS[123].

The improving prediction accuracy has made the practical utility of PRS possible[110,124-126]. For example, Khera et al.[124] constructed PRS for five common diseases and found much more individuals based on PRS in comparison with an equivalent risk based on relative monogenic mutations. Now more studies are trying to combine the PRS with established epidemiology risk factors/models (see Chapter 4 for more information about combining genetic and environmental predictors).

2

Chapter 2: Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank

This chapter has been published in *Science Advances* in 2019

The following publication has been incorporated as Chapter 2:

1. **Wang, H.**, Zhang, F., Zeng, J., Wu, Y., Kemper, K.E., Xue, A., Zhang, M., Powell, J.E., Goddard, M.E., Wray, N.R., Visscher, P.M., McRae, A.F., Yang, J.. Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. Science advances, 2019, 5(8), p.eaaw3538[1].

Contributor	Statement of contribution	%
Huanwei Wang (candidate)	Design the experiment Perform data analysis Write the manuscript	20% 100% 60%
Futao Zhang	Develop the software tool	100%
Jian Zeng	Assistance or guidance	10%
Yang Wu	Assistance or guidance	10%
Kathryn E. Kemper	Assistance or guidance	10%
Angli Xue	Assistance or guidance	10%
Min Zhang	Assistance or guidance	10%
Joseph E. Powell	Critical advice	25%
Michael E. Goddard	Critical advice	25%
Naomi R. Wray	Critical advice	25%
	Contribute resources and funding	20%
Peter M. Visscher	Critical advice	25%
	Contribute resources and funding	20%
Allan F. McRae	Conceive the study	20%
	Design the experiment	20%
	Assistance or guidance	20%
Jian Yang	Conceive the study	80%
	Design the experiment	60%
	Assistance or guidance	30%
	Contribute resources and funding	60%
	Write the manuscript	40%

Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank

2.1 Abstract

Genotype-by-environment interaction (GEI) is a fundamental component in understanding complex trait variation. However, it remains challenging to identify genetic variants with GEI effects in humans largely because of the small effect sizes and the difficulty of monitoring environmental fluctuations. Here, we demonstrate that GEI can be inferred from genetic variants associated with phenotypic variability in a large sample without the need of measuring environmental factors. We performed a genome-wide variance quantitative trait locus (vQTL) analysis of ~5.6 million variants on 348,501 unrelated individuals of European ancestry for 13 quantitative traits in the UK Biobank, and identified 75 vQTLs for 9 traits passing an experiment-wise significant threshold of $P<2.0\times10^{-9}$ (GWS threshold of 5×10^{-8} divided by the effective number of independent traits; see Methods), especially for those related to obesity. Direct GEI analysis with five environmental factors showed that the vQTLs were strongly enriched with GEI effects. Our results indicate pervasive GEI effects for obesity-related traits and demonstrate the detection of GEI without environmental data.

2.2 Introduction

Most human traits are complex because they are affected by many genetic and environmental factors as well as potential interactions between them[6,7]. Despite the long history of effort[82,84,85], there has been limited success in identifying genotype-by-environment interaction (GEI) effects in humans[80,82,90,91]. This is likely because many environmental exposures are unknown or difficult to record during the life course, and because the effect sizes of GEI are small given the polygenic nature of most human traits[13,127,128] so that the sample sizes of most previous studies are not large enough to detect the small GEI effects. For model complex traits such as body mass index (BMI), GEI analyses have been limited to GEI tests at known BMI loci[88,129,130] or estimation of GEI variance captured by all common SNPs[55,95].

GEI effect of a genetic variant on a quantitative trait could lead to differences in variance of the trait among groups of individuals with different variant genotypes (Figure 2-1a-b and Supplementary Note 2-1). GEI can therefore be inferred from a variance quantitative trait locus (vQTL) analysis[131], although there are other explanations for an observed vQTL such as direct effect on phenotypic dispersion (e.g., induced by selection[132]), epistasis[131], and phantom vQTL[133,134]. Unlike the classical quantitative trait locus (QTL) analysis that tests the allelic substitution effect of a variant on the mean of a phenotype (Figure 2-1c), vQTL analysis tests the allelic substitution effect on the trait variance (Figure 2-1b or d). In comparison to the analyses that perform direct GEI tests, vQTL analysis is more flexible because it does not require measures of environmental factors and thus can be performed in a very large sample where the environmental factors are unknown, unavailable or incomplete[135]. Of course, the vQTL test is less powerful than the direct GEI test if the corresponding environmental factor has indeed been measured on all the genotyped individuals in the sample[131]. Although there had been empirical evidence for the genetic control of phenotypic variance in livestock for decades[136,137], it was not until recent years that genome-wide vQTL analysis was applied in humans[131,138,139], and only a handful of vQTLs have been identified for a limited number of traits (e.g. the *FTO* locus for BMI[139]) owing to small effect sizes of vQTLs. The availability of data from large biobank-based genome-wide association studies (GWAS)[46,140] provide an opportunity to interrogate the genome for vQTLs for a range of phenotypes in cohorts with unprecedented sample size.

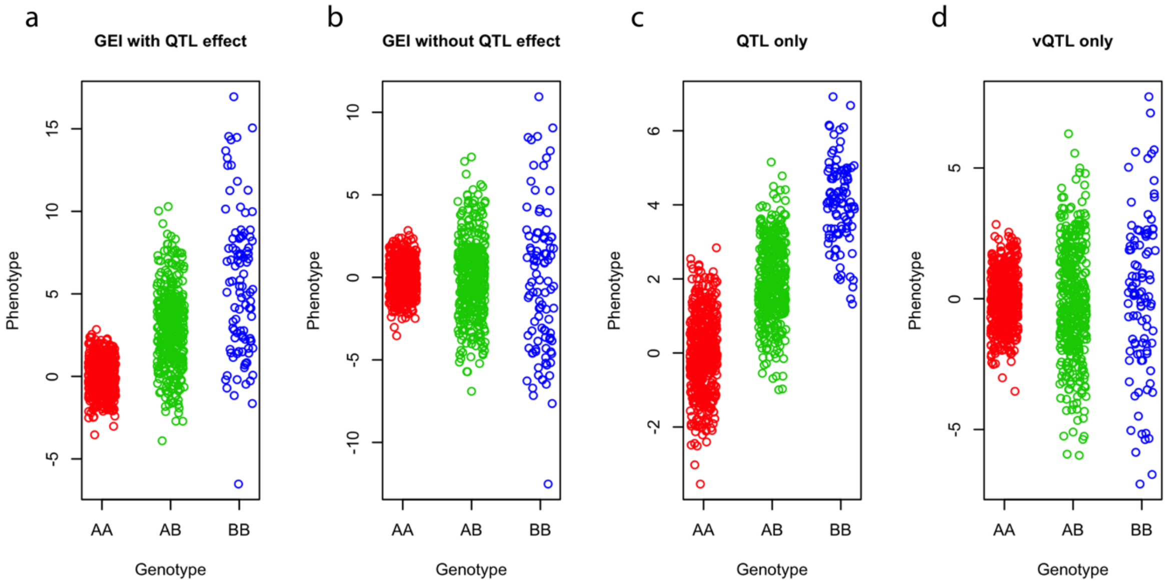


Figure 2-1 Schema of the differences in mean or variance among genotype groups in the presence of GEI, QTL and vQTL effects.

The phenotypes of 1,000 individuals were simulated based on a genetic variant (MAF = 0.3) with a) both QTL and GEI effects, (b) GEI effect only (no QTL effect), (c) QTL effect only (no GEI or vQTL effect), or (d) vQTL only (no QTL effect).

On the other hand, statistical methods for vQTL analysis are not entirely mature[135]. There have been a series of classical non-parametric methods[141], originally developed to detect violation of the homogeneous variance assumption in linear regression model, which can be used to detect vQTLs, including the Bartlett's test[142], the Levene's test[143,144] and the Fligner-Killen (FK) test[145]. Recently, more flexible parametric models have been proposed, including the double generalized linear model (DGLM)[94,146,147] and the likelihood ratio test for variance effect (LRT_V)[133]. In addition, it has been shown that transformation of phenotype that alters phenotype distribution also has an influence on the power and/or false positive rate (FPR) of a vQTL analysis[138,148].

In this study, we calibrated the most commonly used statistical methods for vQTL analysis by extensive simulations. We then used the best performing method to conduct a genome-wide vQTL analysis for 13 quantitative traits in 348,501 unrelated individuals using the UK Biobank (UKB) data[46]. We further investigated whether the detected vQTLs are enriched for GEI by conducting a direct GEI test for the vQTLs with five environmental factors (or covariates).

2.3 Results

Evaluation of the vQTL methods by simulation

We used simulations to quantify the FPR and power (i.e., true positive rate) for the vQTL methods and phenotype processing strategies (Methods). We first simulated a quantitative trait based on a simulated single nucleotide polymorphism (SNP), i.e., a single-SNP model, under a number of different scenarios, namely: 1) five different distributions for the random error term (i.e., individual-specific environment effect); 2) four different types of SNP with or without the effect on mean or variance (Methods). We used the simulated data to compare the four most widely used vQTL methods, namely Bartlett's test[142], Levene's test[143,144], the FK test[145] and the DGLM[94,146,147]. We observed no inflation in FPR for the Levene's test under the null (i.e., no vQTL effect) regardless of the skew or kurtosis of the phenotype distribution or the presence or absence of SNP effect on the mean (Figure 2-2a). These findings are in line with the results from previous studies[138,141,149] that the Levene's test is robust to the distribution of the phenotype. The FPR of the Bartlett's test or DGLM was inflated if the phenotype distribution was skewed or heavy-tailed (Figure 2-2a). The FK test seemed to be robust to kurtosis but vulnerable to skewness of the phenotype distribution (Figure 2-2a). Since the Levene's test performed the best in the simulations, for this test we investigated the impact of non-linear transformations of the phenotype by considering logarithm (log(y)), square (y^2), cube (y^3) and rank-based inverse-normal transformation (RINT) and found that these non-linear transformations could result in inflated FPR (Figure 2-2b). The non-linear transformation, including RINT, could create a departure from the pure additive genetic model and give rise to inflated false positive rate for the vQTL test.

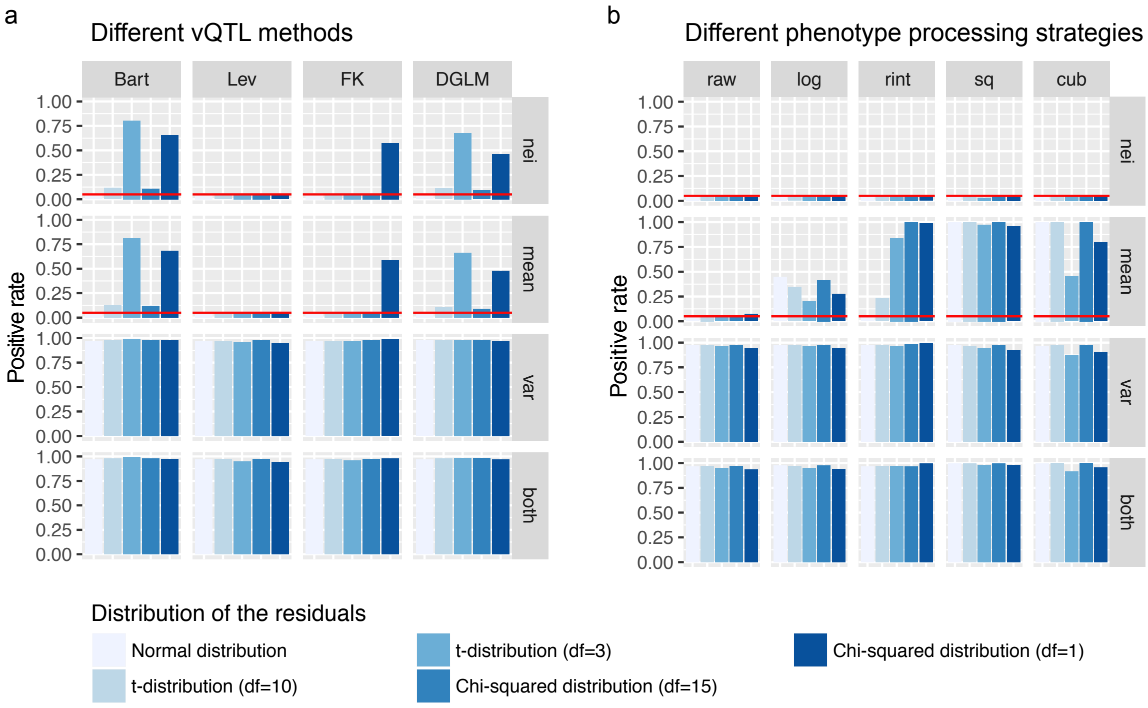


Figure 2-2 Evaluation of (a) statistical methods and (b) phenotype processing strategies for vQTL analysis by simulation based on a single-SNP model.

Phenotypes of 10,000 individuals were simulated based on one SNP and one error term in a single-SNP model (Methods). The SNPs effects were simulated under four scenarios: 1) effect on neither mean nor variance (nei), 2) effect on mean only (mean), 3) effect on variance only (var), or 4) effect on both mean and variance (both). The error term was generated from 5 different distributions: normal distribution, t-distribution with degree of freedom (df) = 10 or 3, or χ^2 distribution with df = 15 or 1. Four statistical test methods, i.e. the Bartlett's test (Bart), the Levene's test (Lev), the Fligner-Killen test (FK) and the DGLM, were used to detect vQTLs. In panel b, the Levene's test was used to analyse phenotypes processed using five strategies, i.e., raw phenotype (raw), raw phenotype adjusted for covariates (adj), rank-based inverse-normal transformation after adj (rint), logarithm transformation after adj (log), square transformation after adj (sq), and cube transformation after adj (cub). Positive rate is defined as the number of vQTLs with $p < 0.05$ divided by the total number of tests across 1,000 simulations, which is the FPR under the null ("nei" and "mean") and power under the alternative ("var" and "both"). The red horizontal line represents an FPR of 0.05.

To simulate more complex scenarios, we used a multiple-SNP model with two covariates (age and sex) with effects on both mean and variance (see Methods), and different numbers of independent SNPs (Figure 2-3). The results were similar to those described above, although the power of the Levene's test decreased with an increase of the number of causal SNPs (Figure 2-3a). Non-linear transformations led to an inflated FPR when the variance explained by a QTL effect (i.e., SNP effect on mean) was relatively large and a loss of power of vQTL detection when the per-QTL variance explained was relatively small although logarithm transformation did not seem affect power (Figure 2-3b). These results also suggested that pre-adjusting the phenotype by covariates slightly increased the power (Figure 2-3b). Based on the results of these simulations we used the Levene's test, a one-way analysis of variance (ANOVA) to test for absolute deviations from the medians (Methods), for real data analysis with the phenotypes pre-adjusted for covariates without any non-linear transformation.

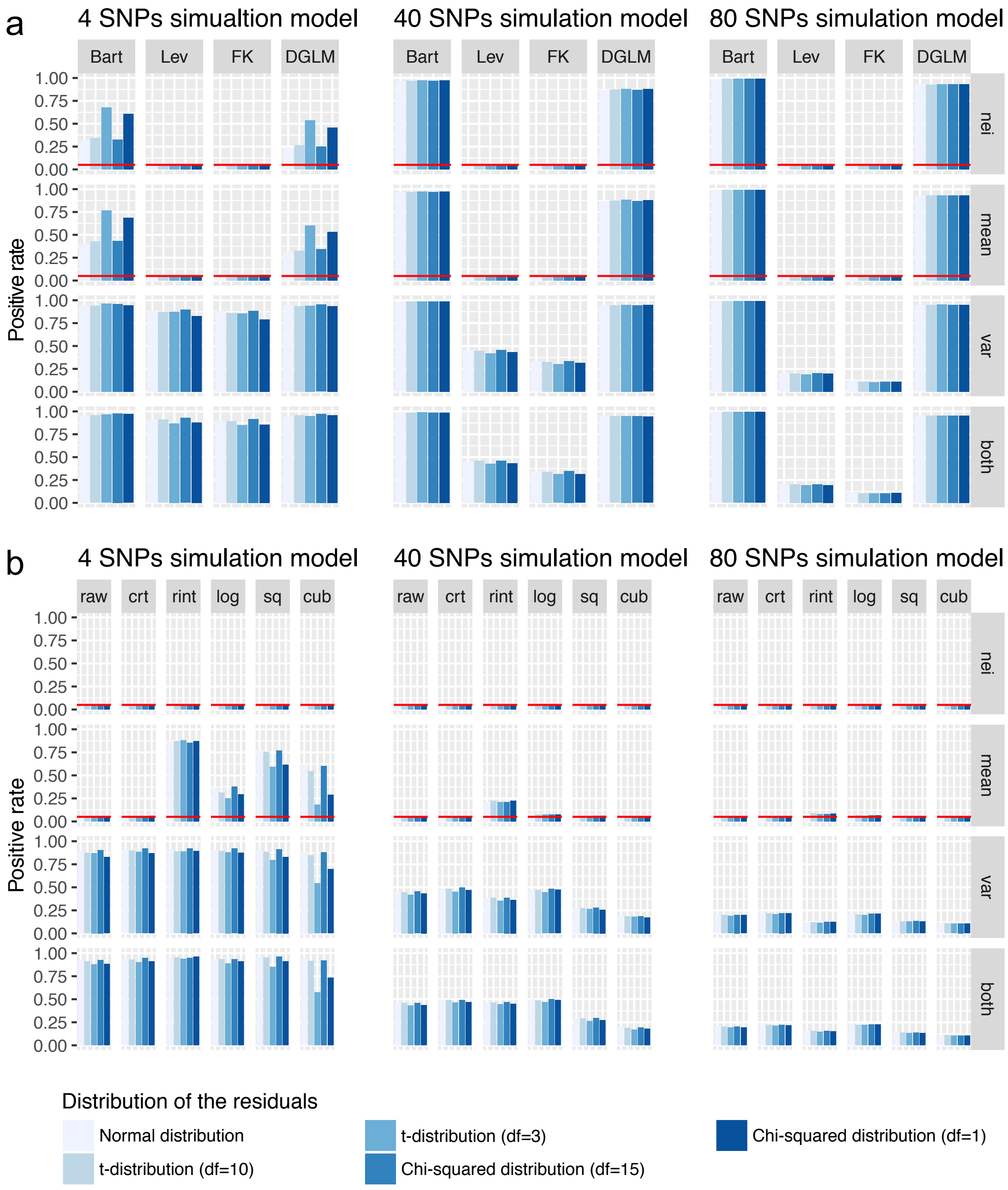


Figure 2-3 Evaluation of (a) statistical methods and (b) phenotype processing strategies for vQTL analysis by simulation based on a multiple-SNP model.

Phenotypes of 10,000 individuals were simulated based on different number of causal independent SNPs (i.e. 4, 40 or 80), two covariates (i.e. sex and age) and one error term in a multiple-SNP model (Methods). The SNP effects were simulated under four scenarios: 1) effect on neither mean nor variance (nei), 2) effect on mean only (mean), 3) effect on variance only (var), or 4) effect on both mean and variance (both). The error term was generated from five different distributions: normal distribution, *t*-distribution with df = 10 or 3, or χ^2 distribution with df = 15 or 1. In panel a, four statistical test methods, i.e., the Bartlett's test (Bart), the Levene's test (Lev), the Fligner-Killen test (FK) and the DGLM, were used to detect vQTLs. In panel b, the Levene's test was used to analyse phenotypes processed using six strategies, i.e., raw phenotype (raw), raw phenotype adjusted for covariates (adj), rank-based inverse-normal transformation after adj (rint), logarithm transformation after adj (log), square transformation after adj (sq), and cube transformation after adj (cub). Positive rate is defined as the number of vQTLs with $p < 0.05$ divided by the total number of tests across 1,000 simulations, which is the FPR under the null ("nei" and "mean") and power under the alternative ("var" and "both"). The red horizontal line represents an FPR of 0.05.

Genome-wide vQTL analysis for 13 UKB traits

We performed a genome-wide vQTL analysis using the Levene's test with 5,554,549 genotyped or imputed common variants on 348,501 unrelated individuals of European ancestry for 13 quantitative traits in the UKB[46] (Methods, Table 2-1 and Figure 2-4). For each trait, we pre-adjusted the phenotypic mean for age and the first 10 principal components (PCs, derived from SNP data) and standardised the residuals to z-scores (i.e., mean 0 and variance 1) in each gender group (Methods). This process removed not only the effects of age and the first 10 PCs on the phenotype but also the differences in mean and variance between the two genders. We excluded individuals with adjusted phenotypes more than 5 standard deviations (SD) from the mean and removed SNPs with minor allele frequency (MAF) smaller than 0.05 to avoid potential false positive associations due to the coincidence of a low-frequency variant with an outlier phenotype (see Figure 2-5 for an example). We acknowledge that this process could potentially result in a loss of power, but this can be compensated for by the use of a very large sample (*n* ~ 350,000).

Table 2-1 Descriptive summary of the quantitative traits and used in this study from the UKB.

Trait	Description	Sample size	UDI[a]
HT	Standing height	347,086	50-0.0

FVC	Forced vital capacity	317,222	3062-0.0
FEV1	Forced expiratory volume in 1-second	317,285	3063-0.0
FFR[b]	FEV1 and FVC ratio	316,614	NA
BMD	Heel bone mineral density T-score, automated	197,261	78-0.0
BW	Birth weight	197,758	20022-0.0
BMI	Body mass index (BMI)	346,393	21001-0.0
WC	Waist circumference	347,158	48-0.0
HC	Hip circumference	346,781	49-0.0
WHR[c]	Waist to Hip Ratio	347,134	NA
WHRadjBMI[d]	WHR adjusted for BMI	346,535	NA
BFP	Body fat percentage	341,632	23099-0.0
BMR	Basal metabolic rate	341,584	23105-0.0

Note: a) UDI, the Unique Data Identifier in the UKB dataset; b) FFR is the ratio of FEV1 to FVC; c) WHR is the ratio of waist circumference to hip circumference; d) WHRadjBMI is the residual after adjusting WHR for BMI.

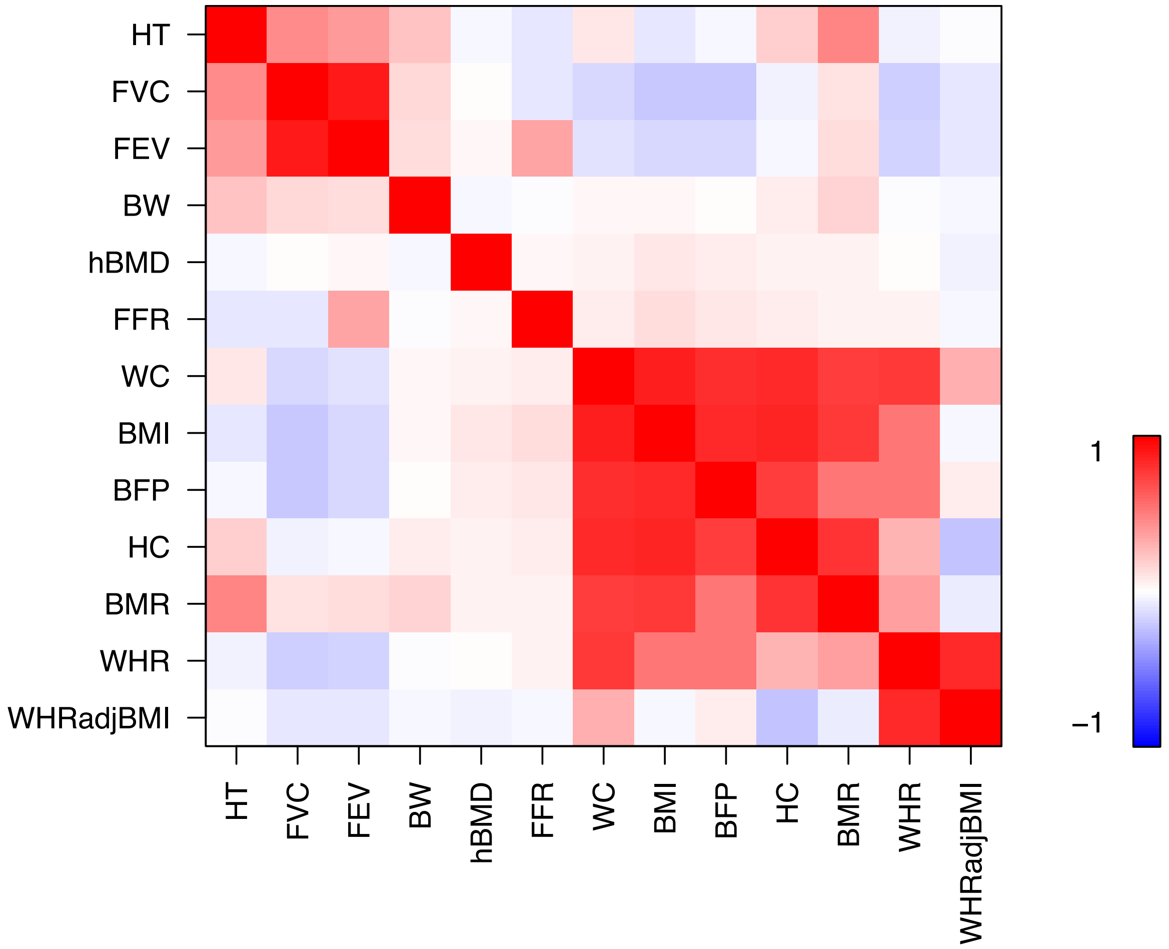


Figure 2-4 Phenotypic correlations among 13 quantitative traits in the UKB.

The Pearson's correlation coefficient was calculated between each pair of (a) the processed phenotypes. The order shown on the plot above was determined by hierarchical cluster analysis using the R function *hclust()*.

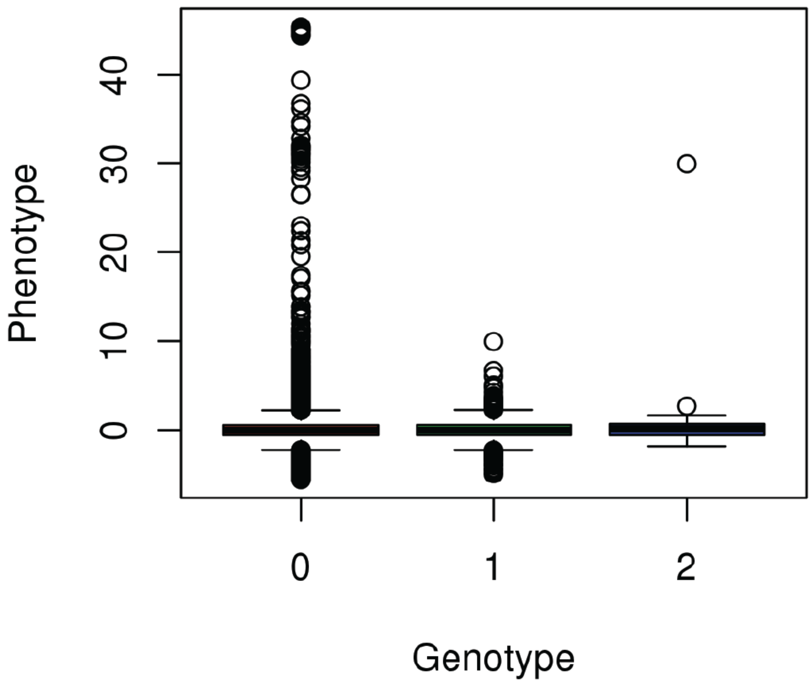


Figure 2-5 Spurious vQTL association due to the coincidence of a minor allele with a phenotypic outlier.

This is an example that a spurious vQTL signal ($P_{\text{vQTL}} = 4.48\times10^{-9}$) at a low-MAF variant (MAF = 0.012) is caused by the coincidence of a minor allele with a phenotypic outlier for FVC. The variance of the phenotype (after covariates adjustment and standardisation) are 1.00, 0.83 and 20.20 in the three genotype groups of rs11102024 respectively. Note that for all the other vQTL results presented in this paper are from analyses excluding individuals with adjusted phenotypes more than 5 SD from the mean and SNPs with MAF < 0.05.

With an experiment-wise significant threshold 2.0×10^{-9} (i.e., $1\times10^{-8}/5.0$ with 1×10^{-8} being a more stringent genome-wide significant threshold recommended by recent studies[150,151] and 5.0 being the effective number of independent traits (Supplementary Note 2-4)), we identified 75 vQTLs (independent to linkage disequilibrium (LD) $r^2 < 0.01$ within trait) across the 9 traits (Figure 2-6, Table 2-2, and Table 2-3). There was no vQTL for height, consistent with the observation in a previous study[139]. We identified more than 15 vQTLs for each of the three obesity-related traits, i.e., BMI, waist circumference (WC), and hip circumference (HC) (

Table 2-2). The 75 vQTLs were located at 41 near-independent loci after excluding one of each between-trait pair of top vQTL SNPs (i.e., the SNP with lowest vQTL p-value at each vQTL association peak) with LD $r^2 > 0.01$, suggesting that some of the loci were associated with the phenotypic variance of multiple traits. For example, the *FTO* locus was associated with the phenotypic variance of WC, HC, BMI, body fat percentage (BFP) and basal metabolic rate (BMR) (Figure 2-7) and the vQTL associations were likely to be driven by a shared causal variant having pleiotropic vQTL effects on multiple traits (Table 2-4). For the lung-function-related traits, there was no significant vQTL for forced expiratory volume in one second (FEV1) and forced vital capacity (FVC) but were 3 vQTLs for FEV1/FVC ratio (FFR). There was no evidence for an effect of MAF on vQTL test-statistic at the 41 independent loci (Figure 2-8), consistent with the observation in a previous study[139].

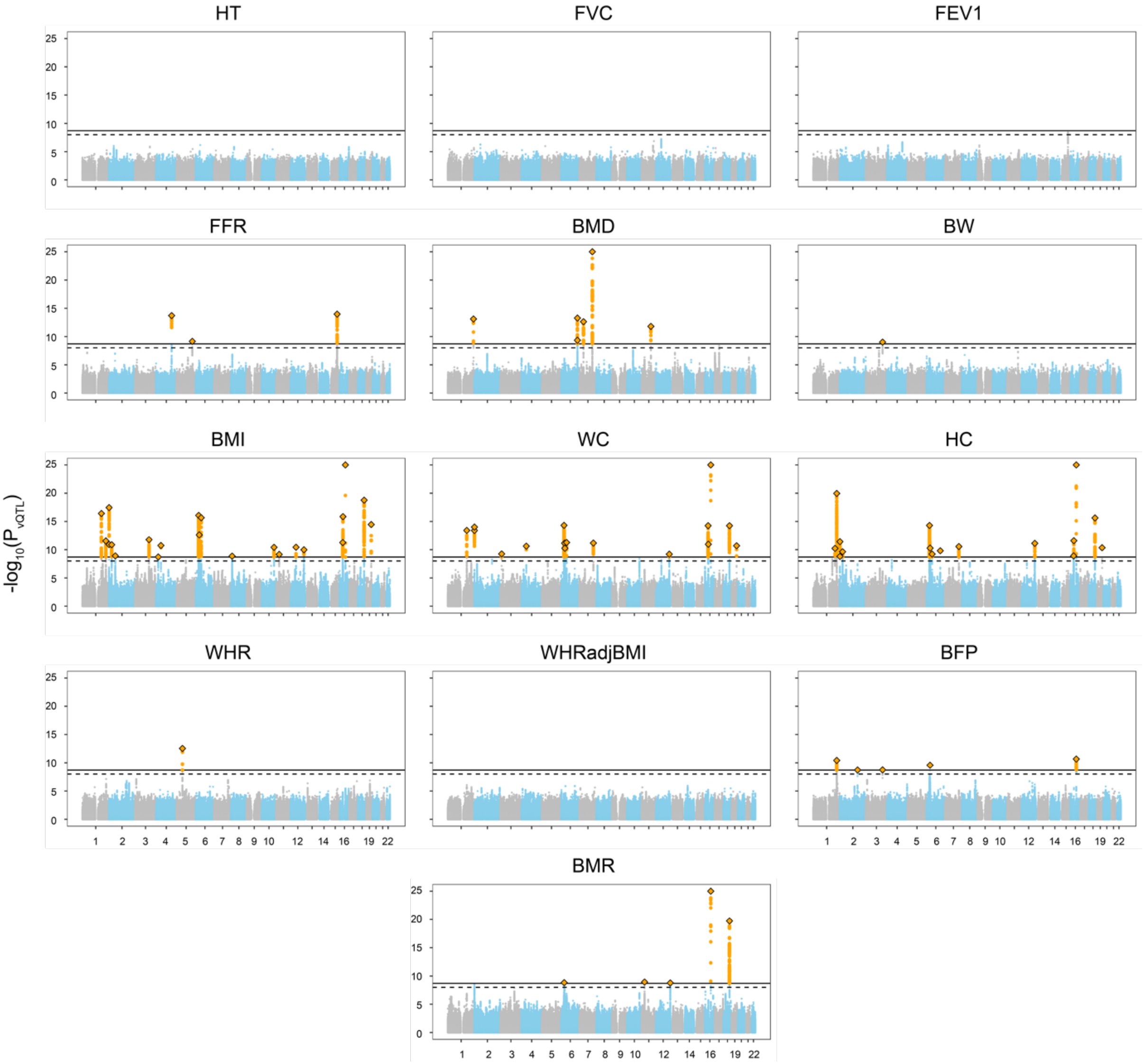


Figure 2-6 Manhattan plots of genome-wide vQTL analysis for 13 traits in the UKB.

For each of the 13 traits (see

Table 2-2 for full names of the traits), test statistics (-log$_{10}$(P_{vQTL})) of all common (MAF ≥ 0.05) SNPs from the vQTL analysis are plotted against their physical positions. The dash line represents the genome-wide significance level 1.0×10^{-8} and the solid line represents the experiment-wise significance level 2.0×10^{-9}. For graphical clarity, SNPs with $P_{\text{vQTL}} < 1\times10^{-25}$ are omitted, SNPs with $P_{\text{vQTL}} < 2.0\times10^{-9}$ are colour-coded in orange, the top vQTL SNP for each locus is represented by a diamond, and the remaining SNPs on odd and even chromosome are colour-coded in grey and blue, respectively.

Table 2-2 The number of experiment-wise significant vQTLs or QTLs for the 13 UKB traits.

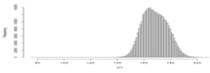
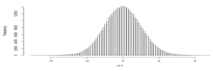
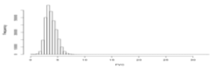
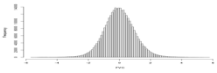
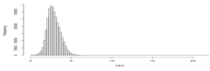
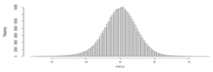
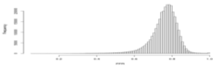
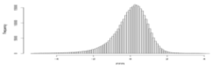
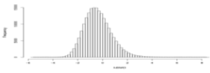
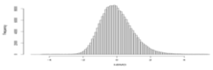
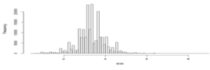
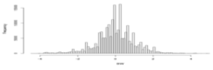
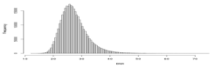
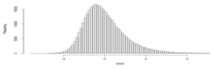
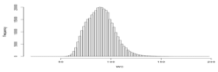
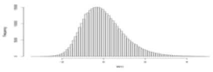
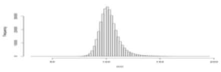
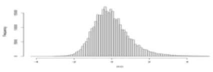
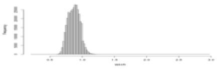
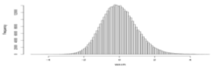
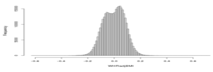
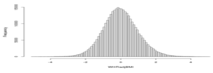
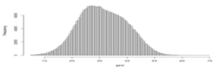
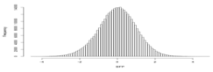
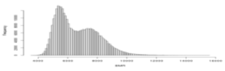
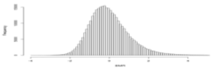
Trait	Description	Distribution of raw phenotype	Distribution of processed phenotype	Number of independent vQTLs for each trait	Number of independent QTLs for each trait
HT	Standing height			0	1063
FVC	Forced vital capacity			0	325
FEV1	Forced expiratory volume in 1-second			0	266
FFR	FEV1 and FVC ratio			3	221
BMD	Heel bone mineral density T-score, automated			6	267
BW	Birth weight			1	57
BMI	Body mass index			22	271
WC	Waist circumference			16	196
HC	Hip circumference			16	249
WHR	Waist to Hip Ratio			1	157
WHRadj BMI	WHR adjusted for BMI			0	187
BFP	Body fat percentage			5	249
BMR	Basal metabolic rate			5	465
Total				75	3,973

Table 2-3 Seventy-five experiment-wise significant vQTLs for 9 UKB traits.

Trait	CHR	SNP	bp	Nearest Gene	MAF	vQTL p-value	QTL p-value	Phenotypic variance in each genotype group	Phenotypic mean in each genotype group
FFR	4	rs6537292	145469968	HHIP	0.394	1.97E-14	3.58E-122[a]	1.0217,0.9936,0.9561	-0.0453,0.0091,0.0787
	5	rs12374521	147836880	FBXO38	0.456	7.10E-10	1.60E-58	1.0223,0.9978,0.97	-0.039,0.0055,0.0417
	15	rs56077333	78899003	CHRNA3	0.325	1.09E-14	2.11E-06	0.9757,1.0107,1.0588	0.0072,-0.0019,-0.0225
BMD	1	rs1414660	240586695	GREM2	0.192	7.83E-14	1.28E-94	0.977,1.0362,1.0452	-0.0322,0.0523,0.1304
	6	rs9371221	151885986	CCDC170	0.101	4.59E-10	1.30E-76	1.0097,0.9502,0.9408	0.02,-0.0817,-0.1479
	6	rs3020332	152008924	ESR1	0.45	5.42E-14	8.94E-130	0.966,0.997,1.0429	-0.074,0.0126,0.0795
	7	rs4576334	38153747	STARD3NL	0.196	2.36E-13	2.60E-86	0.9784,1.0308,1.0684	-0.0325,0.0511,0.1152
	7	rs10254825	120956440	WNT16	0.391	2.01E-45	0	0.9279,1.0107,1.057	-0.1455,0.05,0.1903
	11	rs603140	86884615	TMEM135	0.312	1.61E-12	4.48E-98	1.0149,0.9924,0.9333	0.0417,-0.0204,-0.1142
BW	3	rs13322435	156795468	CCNL1	0.402	9.71E-10	6.21E-48	1.0287,0.9847,0.9742	0.0376,-0.0072,-0.0585
BMI	1	rs545608	177899121	SEC16B	0.206	3.88E-17	1.97E-63	0.9801,1.0251,1.0835	-0.0202,0.0282,0.0847
	1	rs6689335	219628682	LYPLAL1	0.419	2.86E-12	4.73E-08	1.0249,0.9907,0.972	0.0106,-0.0013,-0.0167
	2	rs62104180	466003	FAM150B	0.05	1.22E-11	3.57E-51	1.0054,0.9461,0.8598	0.0083,-0.075,-0.1488
	2	rs6751993	635864	TMEM18	0.167	3.50E-18	3.31E-65	1.0155,0.9707,0.9188	0.0197,-0.0361,-0.0912
	2	rs10203386	25136866	ADCY3	0.452	1.33E-11	8.45E-43	0.9768,0.9994,1.0333	-0.0272,3e-04,0.0404
	2	rs1641155	58965211	FANCL	0.311	1.25E-09	4.42E-17	0.9872,1.0092,1.0266	-0.0141,0.0108,0.0266

	3	rs1225053	131642852	CPNE4	0.264	1.69E-12	2.85E-17	0.9863,1.009,1.0577	-0.0109,0.0073,0.0444
	4	rs10016841	20213781	SLIT2	0.133	1.95E-09	2.11E-13	0.9898,1.0272,1.0621	-0.007,0.0187,0.0468
	4	rs12507026	45181334	GNPDA2	0.434	1.84E-11	6.78E-41	0.9762,0.9998,1.0381	-0.0243,-6e-04,0.0435
	6	rs34817112	27176628	PRSS16	0.134	8.48E-17	3.51E-08	0.9857,1.0416,1.0635	-0.0054,0.0154,0.026
	6	rs3132947	32176782	GPSM3	0.218	2.36E-13	8.53E-15	0.9834,1.0214,1.0563	-0.0096,0.0119,0.038
	6	rs987237	50803050	TFAP2B	0.18	2.18E-16	7.51E-43	0.9842,1.0249,1.0845	-0.0148,0.0247,0.0833
	8	rs17150703	9745798	MSRA	0.104	1.38E-09	2.05E-11	0.9925,1.0243,1.1486	-0.0053,0.0196,0.0583
	10	rs4132670	114767771	TCF7L2	0.312	3.88E-11	2.75E-15	1.0205,0.986,0.9606	0.0133,-0.0084,-0.0263
	11	rs2049045	27694241	BDNF	0.187	6.91E-10	8.20E-42	1.0115,0.9794,0.9461	0.0162,-0.0288,-0.0563
	12	rs7132908	50263148	BCDIN3D	0.385	3.73E-11	3.94E-32	0.9791,1.0024,1.0429	-0.0211,0.0046,0.0392
	12	rs11057413	124489162	ZNF664-FAM101A	0.334	1.05E-10	6.30E-09	0.981,1.0071,1.0456	-0.0104,0.0064,0.0173
	16	rs4072402	28937259	RABEP2	0.337	5.55E-12	2.72E-28	0.9802,1.0076,1.0463	-0.0185,0.0081,0.0393
	16	rs12716979	31011821	STX1B	0.375	1.40E-16	7.30E-24	1.031,0.9897,0.9517	0.0186,-0.0053,-0.0328
	16	rs11642015	53802494	FTO	0.404	1.73E-73	7.43E-217	0.9398,1.0013,1.1095	-0.0555,0.005,0.1062
	18	rs10871777	57851763	MC4R	0.236	1.73E-19	3.01E-81	0.9767,1.0232,1.0751	-0.0248,0.0262,0.0897
	19	rs2238691	46179043	GIPR	0.194	3.46E-15	2.31E-32	1.0176,0.9706,0.9309	0.0142,-0.0231,-0.0537
WC	1	rs10913469	177913519	SEC16B	0.205	3.80E-14	4.50E-44	0.9848,1.0189,1.0695	-0.0166,0.0229,0.0724
	2	rs62104180	466003	FAM150B	0.05	3.93E-14	4.02E-44	1.0061,0.9417,0.8124	0.0077,-0.0689,-0.1472
	2	rs13412194	653245	TMEM18	0.172	9.76E-15	1.39E-55	1.0134,0.9726,0.9343	0.0176,-0.0341,-0.0761
	3	rs7649970	12392272	PPARG	0.121	5.60E-10	5.30E-10	0.9915,1.0245,1.0873	-0.0057,0.0186,0.0314

	4	rs12507026	45181334	GNPDA2	0.434	2.39E-11	9.40E-31	0.9757,1.0016,1.0355	-0.0207,-8e-04,0.0377
	6	rs13198716	26582035	ABT1	0.109	4.89E-15	0.0305	0.99,1.0363,1.0714	-0.002,0.0078,0.0046
	6	rs1062070	32148031	RNF5	0.199	7.20E-12	6.06E-10	0.9862,1.0221,1.043	-0.0079,0.0132,0.0217
	6	rs4472337	34769765	UHRF1BP1	0.155	5.60E-11	1.78E-23	0.9893,1.0237,1.0573	-0.0105,0.024,0.0493
	6	rs987237	50803050	TFAP2B	0.18	5.43E-12	1.09E-34	0.9874,1.0199,1.0669	-0.0137,0.0239,0.0663
	7	rs12667251	130449458	KLF14	0.436	6.82E-12	1.91E-05	1.025,0.9962,0.9637	0.0096,-0.0023,-0.0109
	12	rs7133378	124409502	CCDC92	0.318	6.25E-10	0.506	0.9845,1.0071,1.0384	0.001,1e-04,-0.0034
	16	rs8056890	28897452	ATP2A1	0.355	5.57E-15	8.85E-40	0.9759,1.009,1.0433	-0.0234,0.0094,0.0433
	16	rs34898535	31025641	STX1B	0.378	1.11E-11	6.24E-22	1.0246,0.991,0.9616	0.0173,-0.0047,-0.0316
	16	rs1421085	53800954	FTO	0.404	3.27E-52	3.21E-166	0.9501,1.0048,1.0807	-0.0481,0.0038,0.0936
	18	rs11152213	57852948	MC4R	0.236	5.62E-15	1.39E-70	0.9828,1.0153,1.0646	-0.0224,0.0224,0.0898
	19	rs1800437	46181392	GIPR	0.194	2.05E-11	1.19E-24	1.0137,0.9791,0.93	0.0124,-0.0203,-0.0445
HC	1	rs6685593	203516075	OPTC	0.495	5.99E-11	2.47E-12	0.9682,1.0032,1.0238	-0.016,-3e-04,0.0181
	1	rs2605098	219643649	LYPLAL1	0.338	1.11E-20	3.87E-38	0.9723,1.0085,1.0687	-0.0209,0.0081,0.0483
	2	rs62104180	466003	FAM150B	0.05	1.58E-09	5.56E-45	1.0054,0.9447,0.9202	0.0078,-0.07,-0.1422
	2	rs6751993	635864	TMEM18	0.167	3.84E-12	1.43E-58	1.0142,0.9714,0.932	0.0186,-0.0337,-0.0883
	2	rs10200566	25130462	ADCY3	0.451	2.32E-10	2.29E-18	0.9811,0.9976,1.0342	-0.0171,0,0.026
	6	rs34158769	26336572	BTN3A2	0.104	5.06E-15	4.20E-13	0.9881,1.0436,1.0925	-0.0061,0.0238,0.0417
	6	rs3132947	32176782	GPSM3	0.218	5.34E-11	2.52E-24	0.9851,1.019,1.049	-0.0132,0.0176,0.0429
	6	rs72891717	50858235	TFAP2B	0.169	6.03E-10	1.10E-36	0.9869,1.0226,1.0786	-0.0134,0.0247,0.0784
	6	rs141783576	127439897	RSPO3	0.067	1.57E-10	4.72E-32	1.0064,0.9482,0.994[b]	0.0079,-0.0512,-0.0833

	7	rs17789506	130445574	KLF14	0.493	2.80E-11	1.87E-18	0.9676,1.0001,1.0321	-0.0189,-0.0019,0.0234
	12	rs10846580	124415453	CCDC92	0.337	7.64E-12	9.27E-17	0.9793,1.0122,1.0302	-0.016,0.0106,0.0209
	16	rs8056890	28897452	ATP2A1	0.355	1.27E-09	1.71E-41	0.9762,1.0105,1.0362	-0.0239,0.0096,0.0444
	16	rs34898535	31025641	STX1B	0.378	2.57E-12	1.58E-27	1.0248,0.9945,0.9502	0.0198,-0.0054,-0.0352
	16	rs1421085	53800954	FTO	0.404	1.65E-48	2.05E-152	0.9486,1.0029,1.0909	-0.0462,0.0039,0.0893
	18	rs11152213	57852948	MC4R	0.236	2.39E-16	1.44E-72	0.98,1.0189,1.0704	-0.0237,0.0257,0.0817
	19	rs2238691	46179043	GIPR	0.194	4.52E-11	9.80E-20	1.0162,0.9727,0.9364	0.0105,-0.0164,-0.0467
WHR	5	rs459193	55806751	C5orf67	0.253	2.86E-13	1.75E-19	0.9859,1.0102,1.0584	-0.0128,0.0129,0.0354
BFP	1	rs2820468	219673705	LYPLAL1	0.345	3.76E-11	3.46E-21	0.9824,1.0063,1.0364	-0.0164,0.0066,0.0328
	2	rs1128249	165528624	GRB14	0.392	1.93E-09	2.32E-18	0.9819,1.0045,1.0279	-0.0165,0.0039,0.0275
	3	rs900399	156798732	CCNL1	0.397	1.82E-09	0.000121	1.0198,0.9932,0.9746	0.0065,-4e-04,-0.0138
	6	rs2523625	31315648	HLA-B	0.331	2.69E-10	0.0215	0.9852,1.0071,1.0331	-0.0049,0.0039,0.0041
	16	rs62033406	53824226	FTO	0.411	2.15E-11	1.44E-91	0.9806,0.9991,1.0359	-0.0367,0.0025,0.0677
BMR	6	rs10456362	28221816	ZKSCAN4	0.161	1.48E-09	1.63E-09	0.99,1.0252,1.0064	-0.0069,0.0163,0.0186
	11	rs80083564	27733143	BDNF	0.136	1.15E-09	2.32E-22	0.9895,1.0273,1.0718	-0.0092,0.0242,0.0657
	12	rs78719460	133395038	GOLGA3	0.31	1.66E-09	1.92E-12	0.984,1.0063,1.0467	-0.0108,0.0054,0.0289
	16	rs1421085	53800954	FTO	0.404	8.95E-47	9.23E-154	0.9561,1.0008,1.0805	-0.0469,0.0041,0.09
	18	rs476828	57852587	MC4R	0.237	1.87E-20	1.35E-148	0.9775,1.0195,1.0716	-0.0351,0.0388,0.1125

Note: a) p values smaller than 2.0×10^{-9} are highlighted in pink; b) vQTLs with non-additive genetic effect on variance are highlighted in yellow.

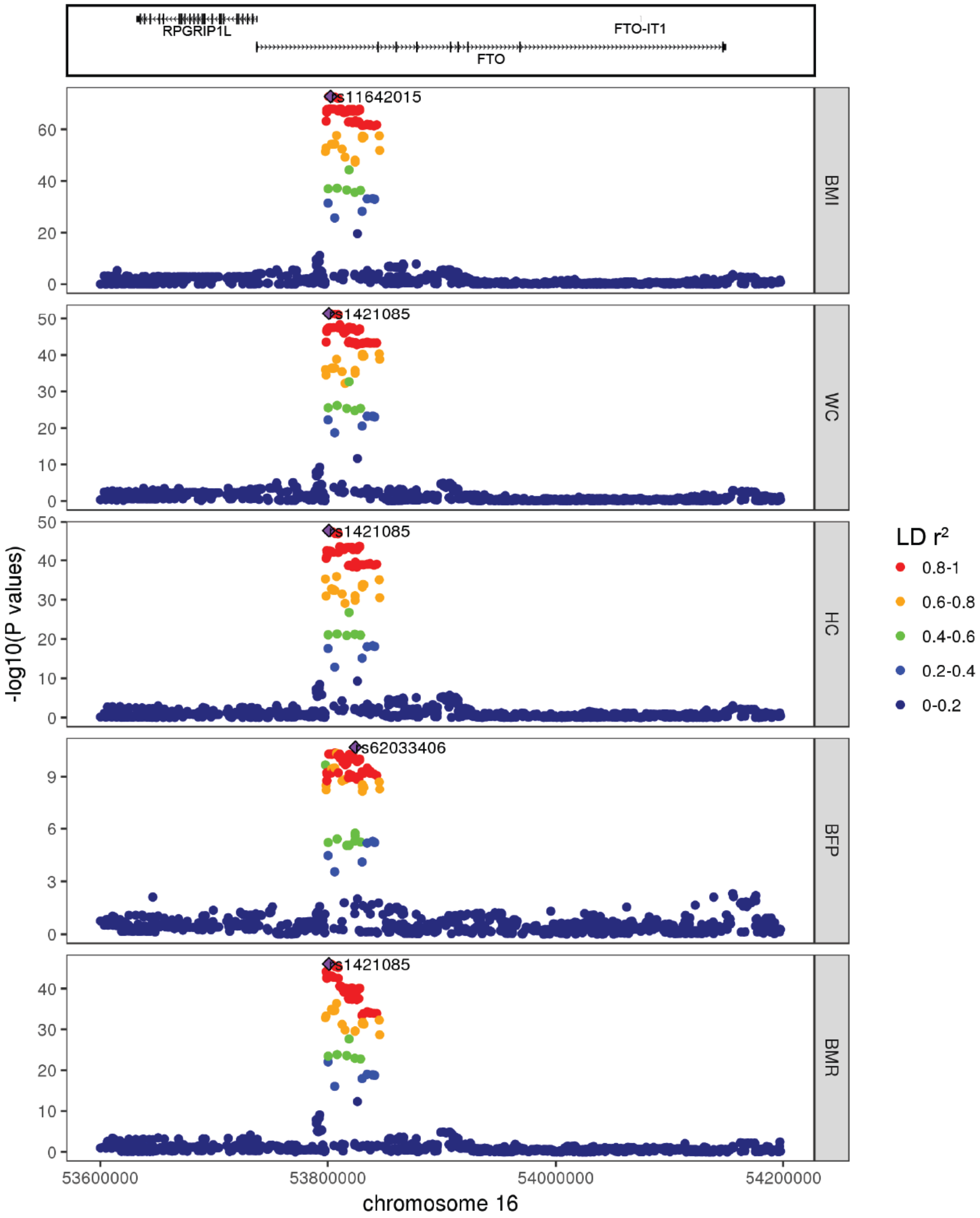


Figure 2-7 The vQTL regional plot at the *FTO* locus for 5 traits.

For each of the 5 traits for which the phenotypic variance is significantly associated with the *FTO* locus, vQTL test statistics (-log$_{10}$(P_{vQTL})) are plotted against SNP positions surrounding the top vQTL SNP (represented by a purple diamond) at the *FTO* locus. SNPs in different levels of LD with the top vQTL SNP are shown in different colours. The RefSeq genes in the top panel are extracted from the UCSC Genome Browser (URLs).

Table 2-4 Colocalization and HEIDI tests for the vQTL associations at the *FTO* locus for the 5 traits.

	BMI	WC	HC	BFP	BMR
BMI	-	99.50%	99.60%	96.30%	99.60%
WC	0.959	-	99.50%	96.30%	99.50%
HC	0.834	0.958	-	96.30%	99.60%
BFP	0.663	0.535	0.689	-	96.30%
BMR	0.793	0.887	0.867	0.463	-

We used the COLOC[152] method implemented in *R* and the HEIDI method[153] implemented in SMR (URLs) to test whether the vQTL associations at the *FTO* locus for the 5 traits as shown in panel (b) are due to the same underlying causal variant. The COLOC and HEIDI analyses were performed for each pair of traits. Note that we convert vQTL p-values to vQTL effect sizes and standard errors using the method described in Zhu et al. [153] (with the direction of each vQTL effect determined by comparing the phenotypic variance among the genotype classes of a SNP) for the HEID analysis. The COLOC PP4 values (up-right off-diagonal), the posterior possibility for hypothesis 4 (i.e., association signals at a locus for two traits are driven by a shared causal variant), were all greater than 80%, and the HEIDI p-values (down-left off-diagonal), testing against the null hypothesis that the association signals for two traits at a locus are driven by the same set of causal variants, were all larger than 0.05.

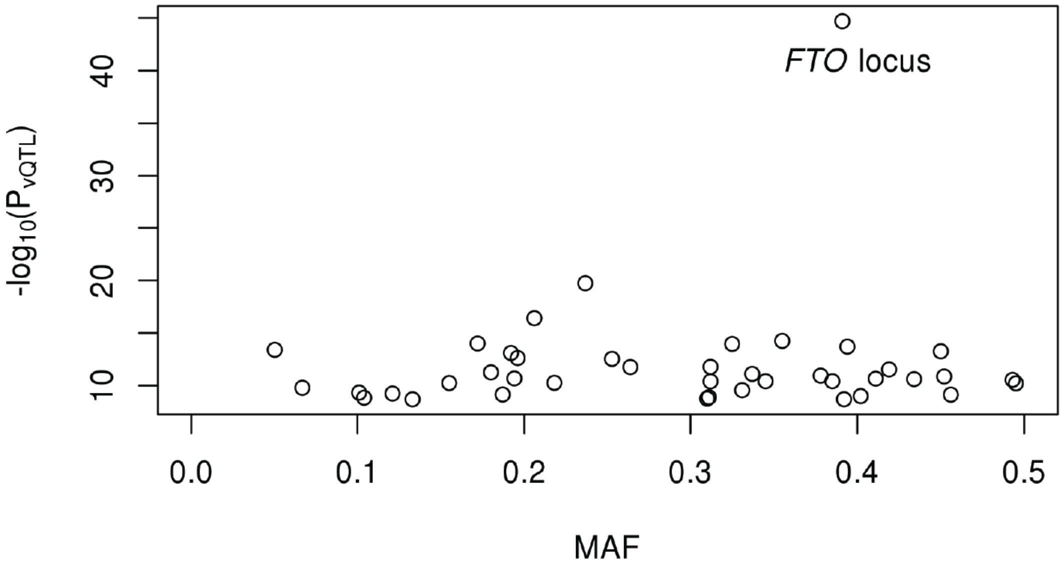


Figure 2-8 A plot of test statistic (-log$_{10}$(P_{vQTL})) against MAF for the 41 independent vQTLs across traits.

The Levene’s test assesses the difference in variance among three genotype groups free of the assumption about additivity (i.e., the vQTL effect of carrying two copies of the effect allele is not assumed to be twice that carrying one copy). We found two vQTLs (i.e., rs141783576 and rs10456362) potentially showing non-additive genetic effect on the variance of HC and BMR, respectively (Table 2-3).

To demonstrate the vulnerability of vQTL analysis to non-linear transformations in real data, we performed genome-wide vQTL analysis for height squared and cubed. There was no genome-wide significant vQTL for height squared but one genome-wide significant vQTL for height cubed, which was very likely to be driven by a strong QTL signal for height ($P_{QTL(Height)}$=4.35×10^{-150}) (Figure 2-9 and Figure 2-10), consistent with our simulation results that non-linear transformations could inflate the vQTL test-statistics in the presence of a strong QTL signal (Figure 2-2b and Figure 2-3b). Although we have not applied any non-linear transformation to the UKB traits, some of them are non-linear functions of other traits, i.e., BMI (= WT/HT^2), FFR (= FEV1/FVC) and WHR (= WC/HC). We therefore explored whether the BMI, FFR and WHR vQTLs were driven by the non-linear functions by testing the variance effects of the BMI, FFR and WHR vQTLs on 1/HT^2, 1/FVC and 1/HC, respectively. There were 26 tests in total, none of which reached the experiment-wise significance level (i.e., 2.0×10^{-9}) used to claim vQTLs in this study and 23 of which had a p-value larger than 0.05 (Table 2-5), suggesting that the BMI, FFR and WHR vQTLs were not driven by the non-linear functions. Although the variance effect of an FFR vQTL (rs56077333) on 1/FVC was significant after correcting for 26 tests ($p = 5.11\times10^{-6}$; Table 2-5), the effect of rs56077333 on the variance of 1/FVC was not large enough to drive the vQTL signal for FFR and rs56077333 has a known GEI effect on lung function (see below for more details).

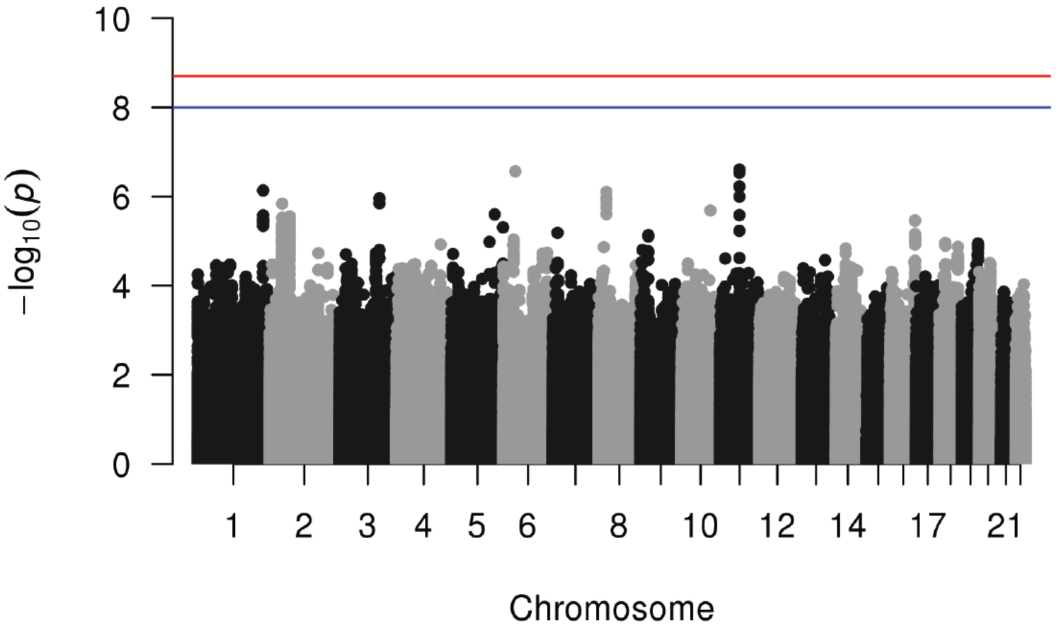


Figure 2-9 Manhattan plots of genome-wide vQTL analysis for height squared in the UKB.

Test statistics (-log$_{10}$(P_{vQTL})) of all common (MAF≥0.05) SNPs from the vQTL analysis are plotted against their physical positions. The blue horizontal line represents the genome-wide significance level 1.0×10^{-8} and the red horizontal line represents the experiment-wise significance level 2.0×10^{-9}.

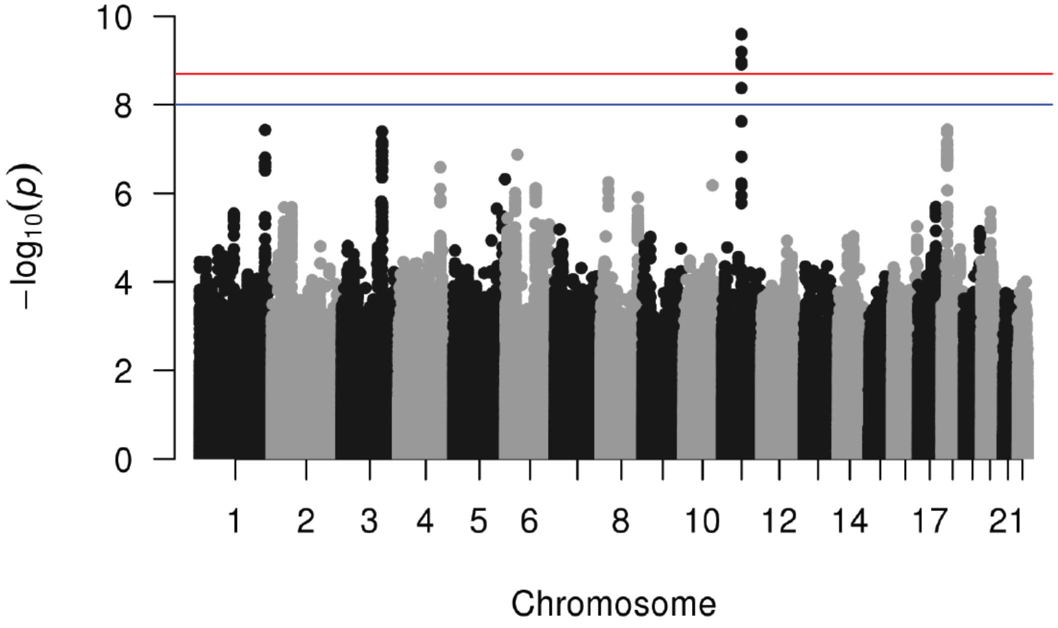


Figure 2-10 Manhattan plots of genome-wide vQTL analysis for height cubed in the UKB.

Test statistics ($-\log_{10}(P_{\mathrm{vQTL}})$) of all common (MAF≥0.05) SNPs from the vQTL analysis are plotted against their physical positions. The blue horizontal line represents the genome-wide significance level 1.0×10^{-8} and the red horizontal line represents the experiment-wise significance level 2.0×10^{-9}.

Table 2-5 Testing for the variance effects of the BMI, WHR and FFR vQTLs on $1/HT^2$, 1/HC and 1/FVC respectively.

vQTL test	SNP	vQTL p-value
BMI - $1/HT^2$	rs545608	4.42E-01
	rs6689335	2.00E-01
	rs62104180	4.31E-02
	rs6751993	2.97E-01
	rs10203386	6.71E-03
	rs1641155	6.74E-01
	rs1225053	1.90E-01

	rs10016841	3.98E-01
	rs12507026	9.08E-01
	rs34817112	2.08E-01
	rs3132947	3.90E-01
	rs987237	7.39E-01
	rs17150703	1.24E-01
	rs4132670	7.29E-01
	rs2049045	6.00E-01
	rs7132908	9.29E-01
	rs11057413	1.71E-01
	rs4072402	1.99E-01
	rs12716979	1.73E-01
	rs11642015	7.65E-01
	rs10871777	8.46E-01
	rs2238691	9.75E-01
WHR - 1/HC	rs459193	9.98E-01
FFR - 1/FVC	rs6537292	5.82E-01
	rs12374521	5.89E-01
	rs56077333	5.11E-06

GWAS analysis for the 13 UKB traits

To investigate whether the SNPs with effects on variance also have effects on mean, we performed GWAS (or genome-wide QTL) analyses for the 13 UKB traits described above. We identified 3,973 QTLs at an experiment-wise significance level (i.e., $P_{\mathrm{QTL}} < 2.0\times10^{-9}$) for the 13 traits in total, a much larger number than that of the vQTLs (

Table 2-2 and Figure 2-11). Among the 75 vQTLs, the top vQTL SNPs at 9 loci did not pass the experiment-wise significance level in the QTL analysis (Table 2-3). For example, the *CCDC92* locus showed a significant vQTL effect but no significant QTL effect on WC (Table 2-3 and Figure 2-12), whereas the *FTO* locus showed both significant QTL and vQTL effects on WC (Figure 2-12). For the 66 vQTLs with both QTL and vQTL effects, the vQTL effects were all in the same directions as the QTL effects, meaning that for any of these SNPs the genotype group with larger phenotypic mean also tends to have larger phenotypic variance than the other groups. For the 9 loci with vQTL effects only, it is equivalent to a scenario where a QTL has a GEI effect with no (or a substantially reduced) effect on average across different levels of an environmental factor (Figure 2-1b).

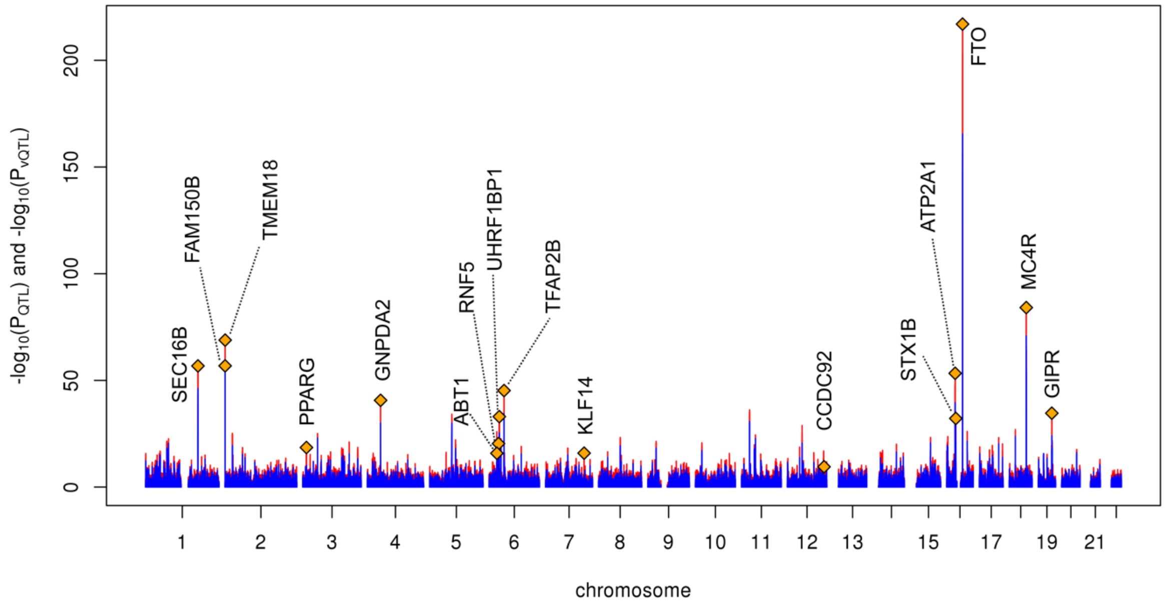


Figure 2-11 Manhattan Sunset plot of genome-wide vQTL and QTL analyses for waist circumference in the UKB.

Test statistics (-log$_{10}$(*P* values)) of all common SNPs from vQTL (red bars) and QTL (blue bars) analysis are plotted against their physical positions. The top vQTL SNP is represented by an orange diamond and the name of the nearest protein-coding gene is indicated for each significant vQTL locus ($P_{vQTL} < 2.0\times10^{-9}$).

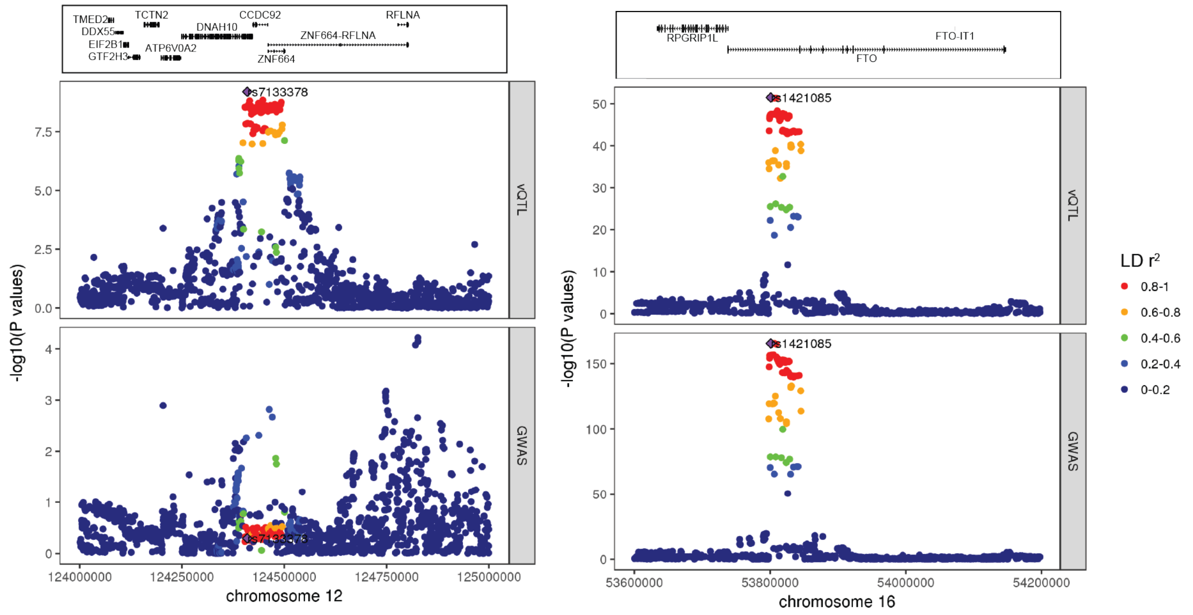


Figure 2-12 QTL and vQTL regional plots at the *CCDC92* or *FTO* locus for waist circumference.

The QTL and vQTL test statistics (i.e., -log$_{10}$(P values)) for waist circumference are plotted against SNP positions surrounding the top vQTL SNP at the *CCDC92* (panel a) or *FTO* locus (panel b). The top vQTL SNP is represented by a purple diamond. SNPs in different levels of LD with the top vQTL SNP are shown in different colours. The RefSeq genes in the top panel are extracted from the UCSC Genome Browser (URLs).

vQTL and GEI

To further investigate whether the associations between vQTLs and phenotypic variance can be explained by GEI, we performed a direct GEI test based on an additive genetic model with an interaction term between a top vQTL SNP and one of five environmental factors/covariates in the UKB data (Methods). The five environmental factors/covariates are sex, age, physical activity (PA), sedentary behaviour (SB), and ever smoking (Supplementary Note 2-5, Figure 2-13 and Table 2-6). We observed 16 vQTLs showing a significant GEI effect with at least one of five environmental factors after Bonferroni correction for multiple tests ($p < 1.33\times10^{-4} = 0.05/(75\times5)$; Figure 2-14a and Table 2-7).

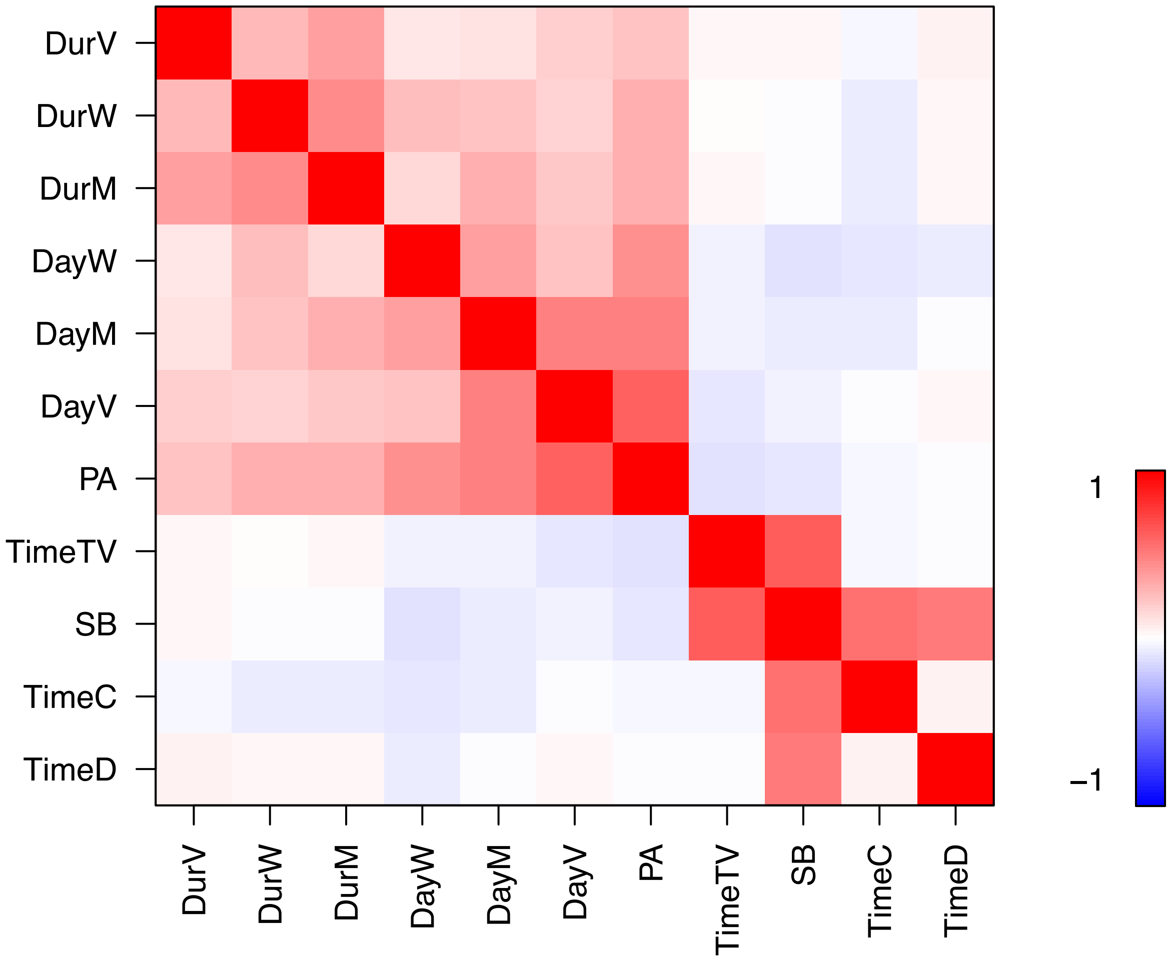


Figure 2-13 Phenotypic correlations among PA and SB measures in the UKB.

The Pearson's correlation coefficient was calculated between each pair of the PA and SB measures. The order shown on the plot above was determined by hierarchical cluster analysis using the R function *hclust()*.

Table 2-6 Descriptive summary of the environmental data used in this study from the UKB.

Item	Description	UDI
Sex	Sex	31-0.0
Age	Year of birth	34-0.0

DayW	Number of days/week walked 10+ minutes	864-0.0
DurW	Duration of walks	874-0.0
DayM	Number of days/week of moderate physical activity 10+ minutes	884-0.0
DurM	Duration of moderate activity	894-0.0
DayV	Number of days/week of vigorous physical activity 10+ minutes	904-0.0
DurV	Duration of vigorous activity	914-0.0
TimeD	Time spent driving	1090-0.0
TimeC	Time spent using computer	1080-0.0
TimeTV	Time spent watching television (TV)	1070-0.0
CurS	Current tobacco smoking	1239-0.0
PastS	Past tobacco smoking	1249-0.0

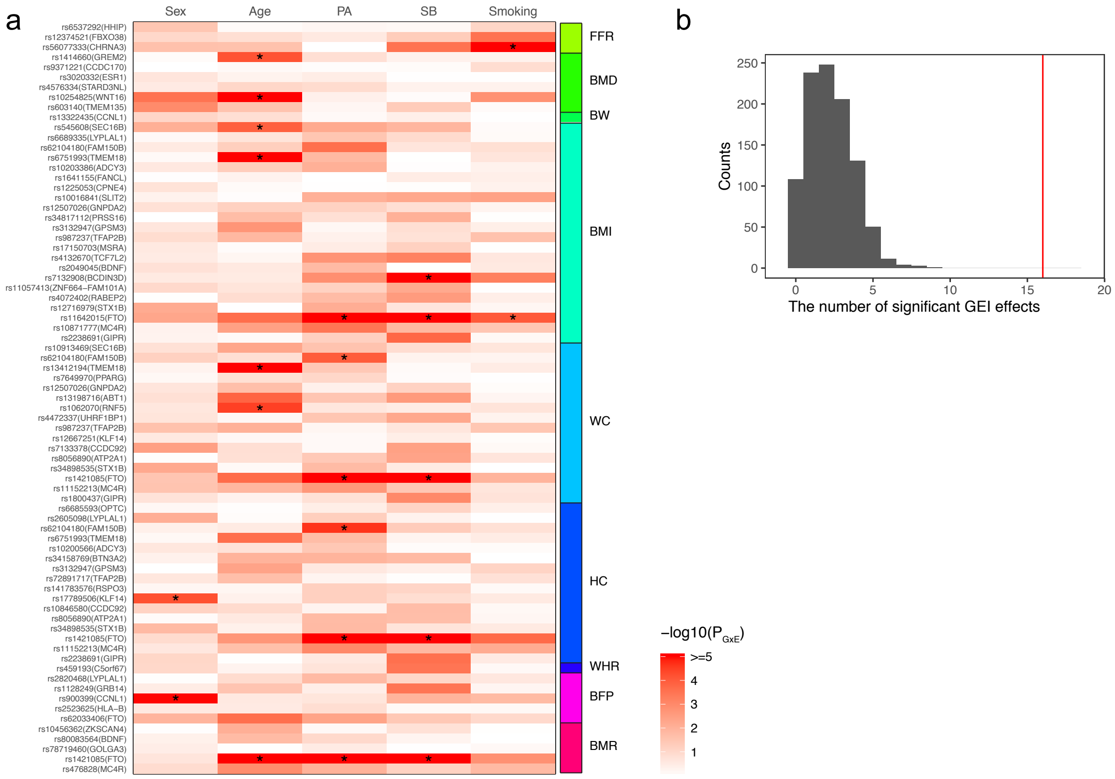


Figure 2-14 Enrichment of GEI effects among the 75 vQTLs in compared with a random set of QTLs.

Five environmental factors/covariates, i.e., sex, age, physical activity (PA), sedentary behaviour (SB), and smoking, were used in the GEI analysis. (a) The heatmap plot of GEI test statistics (-$\log_{10}(P_{\mathrm{GEI}})$) for the 75 top vQTL SNPs. “*” denotes significant GEI effects after Bonferroni correction ($P_{\mathrm{GEI}} < 1.33\times10^{-4} = 0.05/(75\times5)$). (b) The distribution of the number of significant GEI effects for 75 top QTL SNPs randomly selected from all the top QTL SNPs with 1000 repeats (mean 2.25 and SD 1.49). The red line represents the number of significant GEI effects for the 75 top vQTL SNPs (i.e., 16).

Table 2-7 GEI analyses with five environmental factors/covariates in the UKB.

Trait	CHR	SNP	BP	Nearest Gene	P values of GEI analyses with				
					Sex	Age	PA	SB	Smoking
FFR	4	rs6537292	145469968	HHIP	3.13E-02	7.20E-01	5.91E-01	6.39E-01	8.14E-02
	5	rs12374521	147836880	FBXO38	9.46E-02	1.43E-01	2.63E-01	4.79E-02	2.88E-04
	15	rs56077333	78899003	CHRNA3	2.36E-02	2.52E-02	9.88E-01	3.02E-04	4.55E-25
BMD	1	rs1414660	240586695	GREM2	7.60E-01	7.09E-05	1.45E-01	4.26E-01	4.55E-01
	6	rs9371221	151885986	CCDC170	9.81E-01	7.05E-01	8.33E-01	7.69E-01	1.28E-01
	6	rs3020332	152008924	ESR1	2.08E-01	4.76E-01	3.28E-01	9.46E-01	9.53E-01
	7	rs4576334	38153747	STARD3NL	2.85E-01	9.08E-02	1.17E-01	4.58E-01	4.45E-01
	7	rs10254825	120956440	WNT16	3.06E-04	1.16E-07	4.02E-01	7.83E-01	1.59E-03
	11	rs603140	86884615	TMEM135	1.07E-03	2.75E-02	5.74E-01	4.58E-02	7.51E-01
BW	3	rs13322435	156795468	CCNL1	8.46E-02	1.44E-01	6.87E-01	3.69E-01	9.36E-01
BMI	1	rs545608	177899121	SEC16B	8.59E-03	1.24E-04	6.11E-03	1.27E-02	7.51E-01
	1	rs6689335	219628682	LYPLAL1	6.65E-01	1.90E-01	3.15E-02	1.13E-01	7.38E-01
	2	rs62104180	466003	FAM150B	2.75E-01	5.36E-02	2.52E-04	2.07E-01	1.48E-01
	2	rs6751993	635864	TMEM18	7.08E-01	1.01E-07	1.49E-02	9.67E-01	2.06E-01
	2	rs10203386	25136866	ADCY3	2.52E-01	4.94E-02	8.81E-03	8.92E-01	4.51E-01
	2	rs1641155	58965211	FANCL	8.94E-01	4.64E-01	9.82E-01	3.09E-01	3.73E-01
	3	rs1225053	131642852	CPNE4	1.78E-01	7.44E-01	9.53E-01	8.74E-01	4.52E-01

	4	rs10016841	20213781	SLIT2	4.41E-01	8.52E-01	8.64E-03	5.42E-03	3.88E-03
	4	rs12507026	45181334	GNPDA2	1.59E-01	6.19E-02	6.40E-02	1.46E-01	5.21E-01
	6	rs34817112	27176628	PRSS16	9.93E-01	1.91E-02	1.52E-01	9.24E-03	9.21E-01
	6	rs3132947	32176782	GPSM3	2.18E-01	1.92E-03	6.16E-01	1.41E-01	3.66E-01
	6	rs987237	50803050	TFAP2B	1.24E-01	1.41E-02	4.31E-01	1.68E-01	2.49E-02
	8	rs17150703	9745798	MSRA	2.62E-01	7.88E-01	3.15E-01	6.25E-02	9.09E-01
	10	rs4132670	114767771	TCF7L2	3.03E-01	5.72E-01	1.73E-03	6.84E-04	2.36E-01
	11	rs2049045	27694241	BDNF	1.67E-01	2.66E-01	1.59E-02	9.22E-01	2.62E-01
	12	rs7132908	50263148	BCDIN3D	2.73E-01	2.94E-01	1.36E-03	2.15E-07	5.88E-04
	12	rs11057413	124489162	ZNF664-FAM101A	1.02E-01	2.03E-01	6.54E-02	5.81E-03	9.57E-01
	16	rs4072402	28937259	RABEP2	9.25E-01	1.30E-01	1.82E-02	2.72E-03	3.58E-01
	16	rs12716979	31011821	STX1B	6.74E-03	9.39E-01	7.48E-03	2.07E-01	5.89E-01
	16	rs11642015	53802494	FTO	5.01E-03	2.35E-04	1.28E-10	1.64E-09	9.24E-05
	18	rs10871777	57851763	MC4R	4.63E-01	3.72E-03	3.52E-04	1.41E-02	3.22E-02
	19	rs2238691	46179043	GIPR	4.83E-01	7.04E-01	5.95E-02	1.74E-04	6.53E-01
WC	1	rs10913469	177913519	SEC16B	1.63E-01	4.96E-03	2.74E-02	1.31E-01	2.15E-01
	2	rs62104180	466003	FAM150B	6.08E-02	1.46E-01	1.04E-04	4.77E-01	5.19E-01
	2	rs13412194	653245	TMEM18	4.87E-01	1.88E-07	3.70E-02	7.16E-01	3.15E-01
	3	rs7649970	12392272	PPARG	6.35E-01	1.49E-01	8.87E-02	7.91E-01	9.55E-01
	4	rs12507026	45181334	GNPDA2	2.12E-01	2.10E-02	4.08E-01	6.70E-02	8.13E-01
	6	rs13198716	26582035	ABT1	1.52E-01	1.50E-04	3.11E-02	2.45E-03	5.51E-01

	6	rs1062070	32148031	RNF5	2.30E-01	3.46E-05	2.32E-01	3.13E-01	1.96E-01
	6	rs4472337	34769765	UHRF1BP1	2.03E-01	6.15E-01	3.14E-02	5.56E-03	4.73E-01
	6	rs987237	50803050	TFAP2B	2.52E-02	9.32E-03	5.61E-01	2.19E-01	3.32E-02
	7	rs12667251	130449458	KLF14	2.61E-01	6.30E-01	6.99E-01	3.50E-01	8.05E-01
	12	rs7133378	124409502	CCDC92	3.59E-03	1.45E-01	7.05E-01	3.62E-03	6.79E-01
	16	rs8056890	28897452	ATP2A1	8.75E-01	1.52E-01	3.93E-02	4.24E-03	2.10E-01
	16	rs34898535	31025641	STX1B	5.92E-03	7.08E-01	1.69E-02	2.92E-01	9.87E-01
	16	rs1421085	53800954	FTO	3.04E-02	2.17E-04	1.44E-07	2.84E-08	1.10E-02
	18	rs11152213	57852948	MC4R	2.80E-02	1.36E-02	2.21E-03	3.39E-02	2.14E-01
	19	rs1800437	46181392	GIPR	1.84E-01	4.59E-01	1.62E-01	9.50E-04	1.81E-01
HC	1	rs6685593	203516075	OPTC	6.22E-01	8.84E-01	3.48E-01	7.42E-02	3.80E-01
	1	rs2605098	219643649	LYPLAL1	8.00E-03	8.14E-01	5.37E-02	4.52E-01	9.48E-01
	2	rs62104180	466003	FAM150B	3.95E-01	2.89E-01	2.27E-05	4.70E-02	1.62E-01
	2	rs6751993	635864	TMEM18	6.50E-01	2.35E-04	1.87E-02	5.09E-01	3.11E-01
	2	rs10200566	25130462	ADCY3	2.35E-01	1.73E-01	3.57E-02	7.90E-01	7.74E-01
	6	rs34158769	26336572	BTN3A2	4.31E-01	1.05E-02	1.14E-02	1.60E-02	9.93E-01
	6	rs3132947	32176782	GPSM3	9.79E-01	4.23E-03	2.53E-01	3.71E-01	7.46E-02
	6	rs72891717	50858235	TFAP2B	2.32E-01	2.19E-02	5.21E-01	8.67E-01	1.67E-01
	6	rs141783576	127439897	RSPO3	5.75E-01	5.59E-01	5.95E-02	8.08E-02	6.84E-01
	7	rs17789506	130445574	KLF14	6.09E-05	4.28E-01	5.89E-02	1.36E-01	2.73E-01
	12	rs10846580	124415453	CCDC92	6.97E-02	1.10E-01	4.89E-01	1.93E-02	6.48E-01

	16	rs8056890	28897452	ATP2A1	7.53E-01	2.66E-01	1.68E-02	2.05E-02	5.89E-01
	16	rs34898535	31025641	STX1B	1.60E-02	4.20E-01	1.62E-02	1.94E-01	2.77E-01
	16	rs1421085	53800954	FTO	1.18E-01	2.05E-03	5.32E-07	2.17E-06	1.94E-04
	18	rs11152213	57852948	MC4R	2.38E-01	2.19E-02	9.80E-04	1.41E-02	6.68E-03
	19	rs2238691	46179043	GIPR	8.71E-02	9.12E-01	2.66E-01	2.64E-04	2.85E-01
WHR	5	rs459193	55806751	C5orf67	9.25E-02	4.82E-01	1.92E-01	3.48E-04	6.67E-01
BFP	1	rs2820468	219673705	LYPLAL1	8.76E-01	1.19E-01	1.60E-02	8.66E-02	2.40E-01
	2	rs1128249	165528624	GRB14	3.18E-01	2.91E-02	3.55E-01	3.41E-04	5.58E-01
	3	rs900399	156798732	CCNL1	1.04E-05	2.63E-01	2.15E-01	1.13E-02	1.64E-02
	6	rs2523625	31315648	HLA-B	2.66E-01	2.87E-01	1.35E-01	6.76E-01	6.38E-01
	16	rs62033406	53824226	FTO	1.14E-02	2.47E-04	4.43E-03	3.52E-02	1.02E-01
BMR	6	rs10456362	28221816	ZKSCAN4	8.90E-01	8.38E-03	6.68E-01	1.60E-01	9.13E-01
	11	rs80083564	27733143	BDNF	4.20E-01	1.69E-02	1.00E-01	5.22E-01	9.89E-01
	12	rs78719460	133395038	GOLGA3	3.41E-01	8.69E-01	3.04E-01	9.40E-01	6.67E-01
	16	rs1421085	53800954	FTO	2.07E-01	2.60E-07	1.52E-06	1.45E-07	1.54E-03
	18	rs476828	57852587	MC4R	1.20E-01	1.22E-03	9.98E-03	6.47E-02	1.62E-02

Note: p values smaller than 1.33×10^{-4} are highlighted in pink.

To test whether the GEI effects are enriched among vQTLs in comparison with the same number of QTLs, we performed GEI test for 75 top GWAS SNPs randomly selected from all the QTLs and repeated the analysis 1000 times. Of the 75 top SNPs with QTL effects, the number of SNPs with significant GEI effects was 2.25 averaged from the 1000 repeated samplings with a SD of 1.49 (Figure 2-14b), significantly lower than the number (16) observed for the vQTLs (the difference is larger than 9 SDs, equivalent to p = 1.4×10^{-20}). This result shows that SNPs with vQTL effects are much more enriched with GEI effects compared to those with QTL effects. To exclude the possibility that the GEI signals were driven by phenotype processing (e.g., the adjustment of phenotype for sex and age), we repeated the GEI analyses using raw phenotype data without covariates adjustment; the results remain largely unchanged (Figure 2-15).

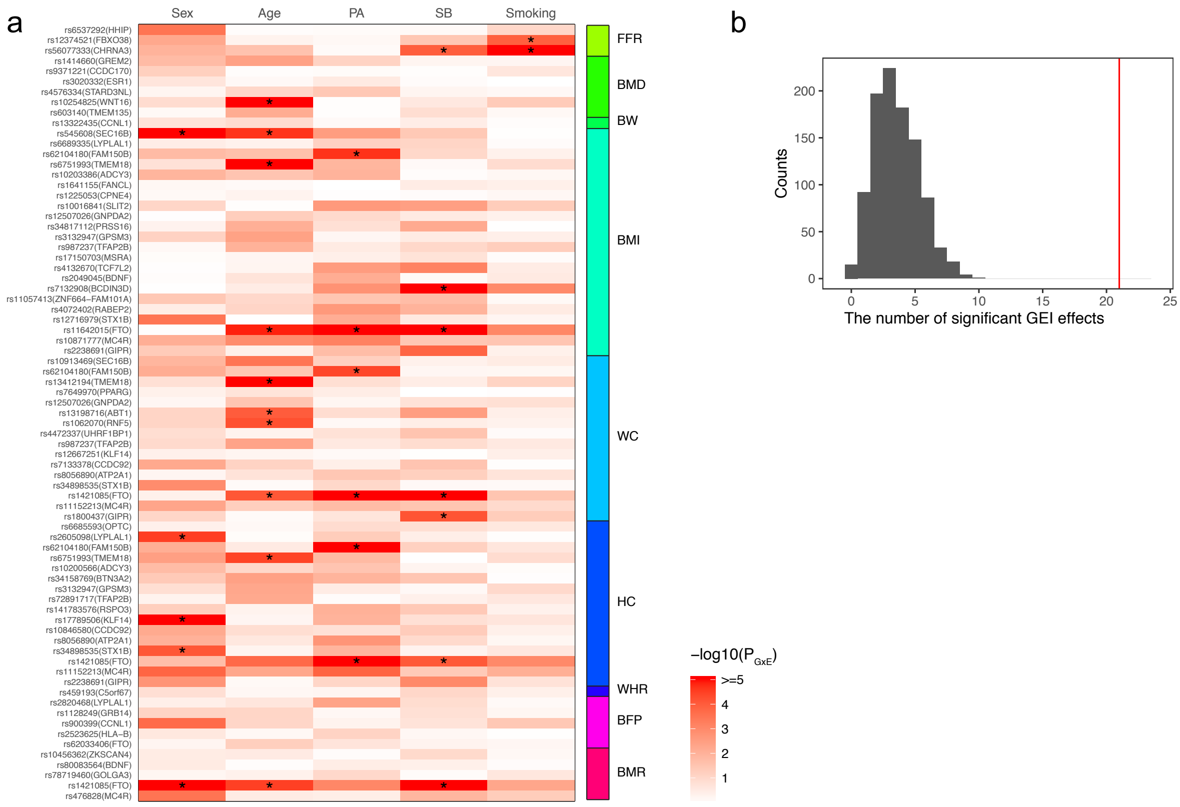


Figure 2-15 Enrichment of GEI effects among the 75 vQTLs in compared with a random set of QTLs using the raw phenotypic values.

Five environmental factors, i.e., sex, age, physical activity (PA), sedentary behaviour (SB), and smoking, were used in the GEI analysis. (a) The heatmap plot of GEI test statistics (-log$_{10}$(P_{GEI})) for the 75 top vQTL SNPs. “*” denotes significant GEI effects after Bonferroni correction ($P_{\mathrm{GEI}} < 1.33\times10^{-4} = 0.05/75/5$). (b) The distribution of the number of significant GEI effects for 75 top QTL SNPs randomly selected from all the top QTL SNPs with 1000 repeats (mean 3.56 and SD 1.76). The red line represents the number of significant GEI effects for the 75 top vQTL SNPs (i.e., 21).

2.4 Discussion

In this study, we leveraged the genetic effects associated with phenotypic variability to infer GEI. We calibrated the most commonly used vQTL methods by simulation. We found that the FPR of the Levene's test was well-calibrated across all simulation scenarios whereas the other methods showed an inflated FPR if the phenotype distribution was skewed or heavy-tailed under the null hypothesis (i.e., no vQTL effect), although the Levene's test appeared to be less powerful than the other methods in particular when the per-variant vQTL effect was small (Figure 2-2 and Figure 2-3). Parametric bootstrap or permutation procedures have been proposed to reduce the inflation in the test-statistics of DGLM and LRTv, both of which are expected to be more powerful than the Levene's test[133,149], but bootstrap and permutation are computationally inefficient and thus not practically applicable to biobank data such as the UKB. We observed inflated FPR for the Levene's test in the absence of vQTL effects but in the presence of QTL effects if the phenotype was non-linearly transformed (e.g., logarithm transformation or RINT). We therefore recommend the use of the Levene's test in practice without non-linear transformation of the phenotype. In addition, a very recent study by Young et al.[154] developed an efficient algorithm to perform a DGLM analysis and proposed a method (called dispersion effect test (DET)) to remove confounding in vQTL associations (identified by DGLM) due to QTL effects. We showed by simulation that when the number of simulated causal variants was relatively large (note that the DET test is not applicable to oligogenic traits), the Young et al. method (DGLM followed by DET) performed similarly as the Levene's test with differences depending on how the phenotype was processed (Figure 2-16).

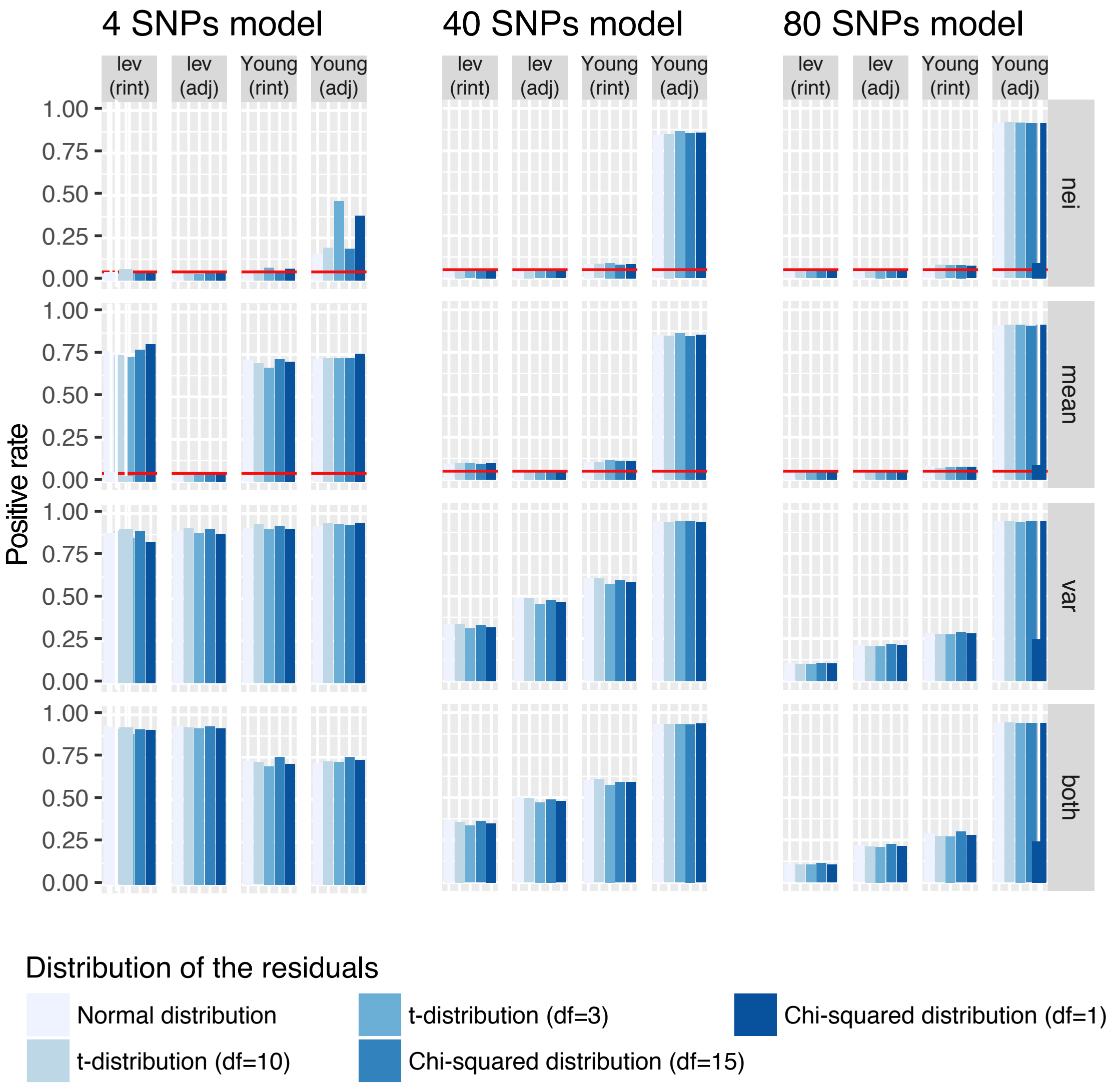
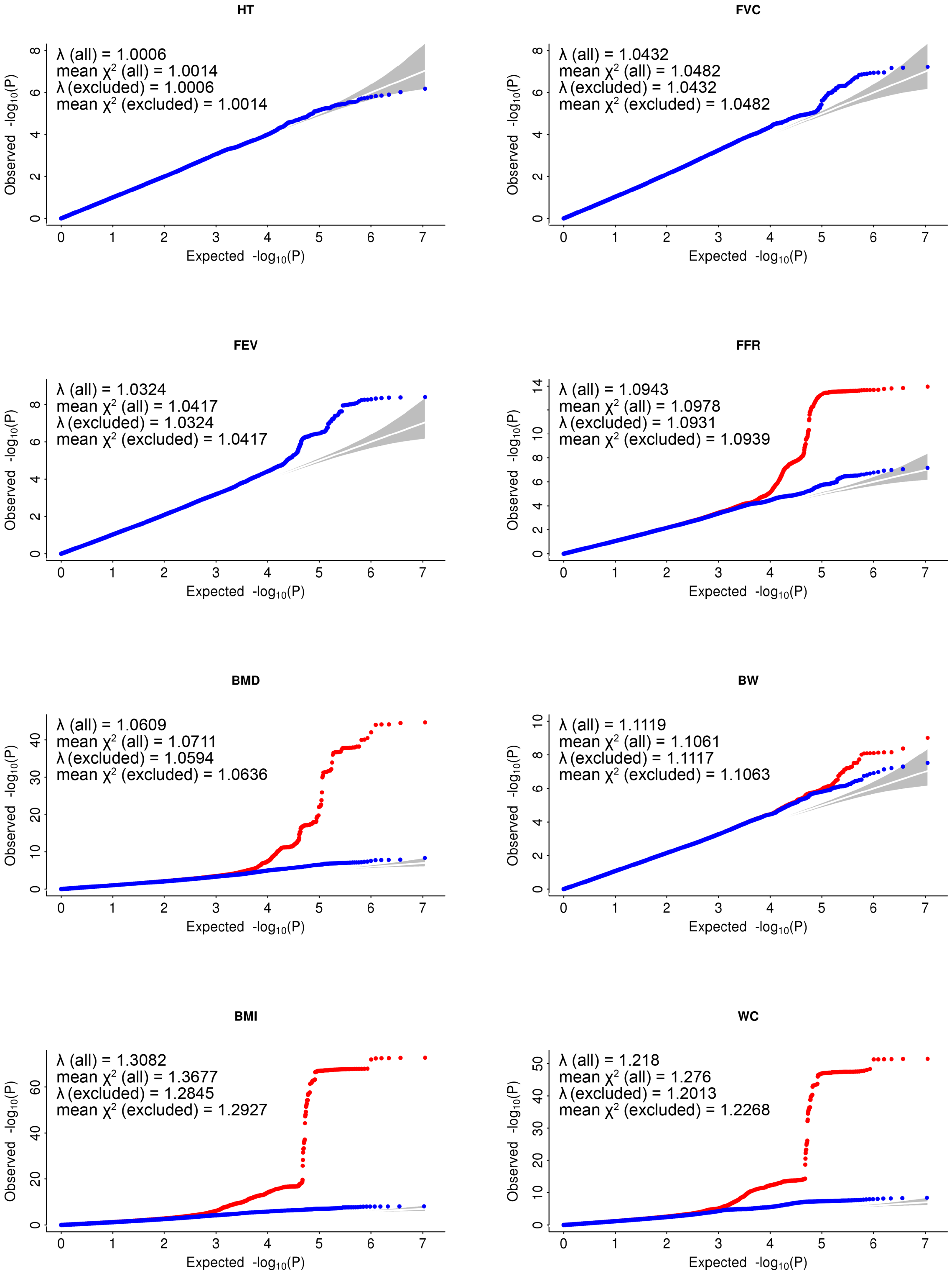


Figure 2-16 Comparison of the Young et al. method with the Levene's test by vQTL simulation.

While we were preparing the manuscript, a very recent study from Young et al.[40] developed an efficient algorithm for fitting DGLM (called heteroskedastic linear mixed model or HLMM) and proposed a dispersion effect test (DET) to remove the impact of the QTL effects on the vQTL signals. We used our multiple-SNP simulation setting (Figure 2-3 and Methods) to quantify the FPR and power of the Young et al. method (HLMM + DET) in comparison with the Levene's test based on the phenotype after 1) covariate adjustment ("adj") or 2) covariate adjustment followed by rank-based inverse-normal transformation ("rint"). For the Levene's test, the positive rate (FPR or power) was computed as the number of vQTLs with $p < 0.05$ divided by the total number of tests across 1,000 simulations. For the analysis with the Young et al. method, the positive rate (FPR or power) was computed as the number of vQTLs with DET $p < 0.05$ divided by the total number of tests across simulations.

We demonstrated in the analysis of the UKB data that a number of vQTLs (with enriched GEI effects) can be detected by an appropriate analytical strategy in a very large sample. Traits with a larger number of vQTLs detected at the experiment-wise significance level tended to have a higher genomic inflation factor (GIF, defined as the mean or median chi-squared statistic divided by its expected value) even after excluding the top vQTLs as well as SNPs in LD with them (Figure 2-17), consistent with a polygenic model of variance effect[61,155], suggesting a large number of vQTLs with small variance effects yet to be discovered in larger samples in the future.


HT
λ (all) = 1.0006
mean χ² (all) = 1.0014
λ (excluded) = 1.0006
mean χ² (excluded) = 1.0014
FVC
λ (all) = 1.0432
mean χ² (all) = 1.0482
λ (excluded) = 1.0432
mean χ² (excluded) = 1.0482
FEV
λ (all) = 1.0324
mean χ² (all) = 1.0417
λ (excluded) = 1.0324
mean χ² (excluded) = 1.0417
FFR
λ (all) = 1.0943
mean χ² (all) = 1.0978
λ (excluded) = 1.0931
mean χ² (excluded) = 1.0939
BMD
λ (all) = 1.0609
mean χ² (all) = 1.0711
λ (excluded) = 1.0594
mean χ² (excluded) = 1.0636
BW
λ (all) = 1.1119
mean χ² (all) = 1.1061
λ (excluded) = 1.1117
mean χ² (excluded) = 1.1063
BMI
λ (all) = 1.3082
mean χ² (all) = 1.3677
λ (excluded) = 1.2845
mean χ² (excluded) = 1.2927
WC
λ (all) = 1.218
mean χ² (all) = 1.276
λ (excluded) = 1.2013
mean χ² (excluded) = 1.2268
Observed -log10(P)
Expected -log10(P)
0 1 2 3 4 5 6 7
0 2 4 6 8
0 2 4 6 8 10 14
0 10 20 30 40
0 20 40 60
0 10 20 30 40 50

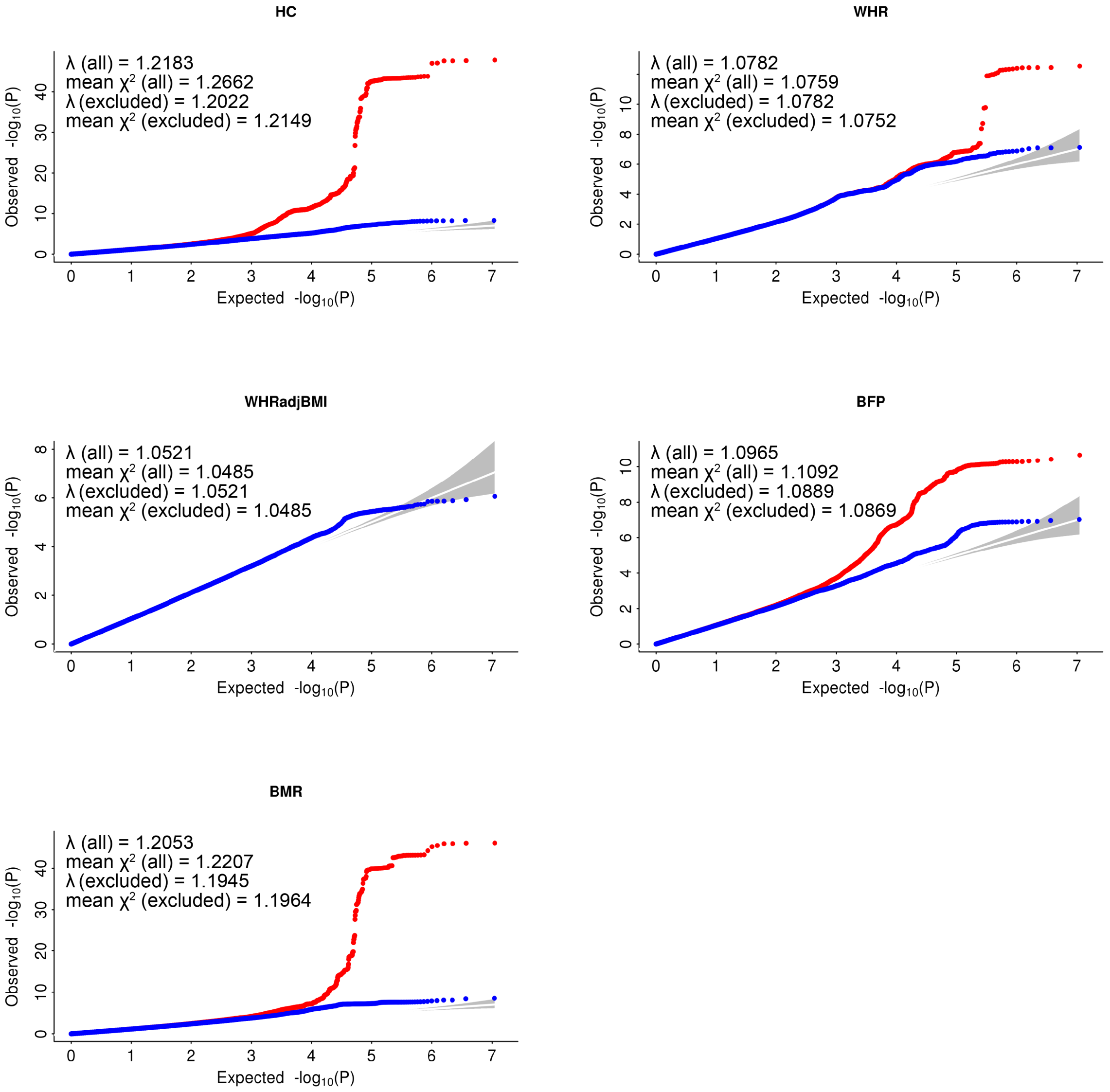


Figure 2-17 Quantile-Quantile plots of vQTL associations for the 13 UKB traits.

For each trait, we shown the QQ plots for all SNPs including (red) or excluding (blue) the top vQTLs and SNPs in LD with them (determined by GCTA-LDF[53]). The area highlighted in grey is the 95% confidence interval.

There are several vQTLs for which the GEI effect has been reported in previous studies. The first example is the interaction effect of the *CHRNA5-A3-B4* locus (rs56077333) with smoking for lung function (as measured by FFR ratio, i.e., FEV1/FVC), $P_{\text{vQTL}} = 1.1\times10^{-14}$ and $P_{\text{GEI(smoking)}} = 4.6\times10^{-25}$ (Table 2-8). The *CHRNA5-A3-B4* gene cluster is known to be associated with smoking and nicotine dependence[156-158]. However, results from recent GWAS

studies[159-161] do not support the association of this locus with lung function. We hypothesize that the effect of the *CHRNA5-A3-B4* locus on lung function depends on smoking[162] (Table 2-8). The vQTL signal at this locus remained ($P_{\text{vQTL}} = 5.2\times10^{-12}$) after adjusting the phenotype for array effect, which was reported to affect the QTL association signal at this locus[46]. The second example is the interaction of the *WNT16-CPED1* locus with age for BMD (rs10254825: $P_{\text{vQTL}} = 2.0\times10^{-45}$ and $P_{\text{GEI(age)}} = 1.2\times10^{-7}$). The *WNT16-CPED1* locus is one of the strongest BMD-associated loci identified from GWAS[163,164]. We observed a genotype-by-age interaction effect at this locus for BMD (
Table 2-9), in line with the results from previous studies that the effect of the top SNP at *WNT16-CPED1* on BMD in humans[165] and the knock-out effect of *Wnt16* on bone mass in mice[166] are age-dependent. The third example is the interaction of the *FTO* locus with physical activity and sedentary behaviour for obesity-related traits ($P_{\text{vQTL}} < 1\times10^{-10}$ for BMI, WC, HC, BFP and BMR; $P_{\text{GEI(PA)}} = 1.3\times10^{-10}$ for BMI, 1.4×10^{-7} for WC, 5.3×10^{-7} for HC and 2.6×10^{-7} for BMR). The *FTO* locus was one of the first loci identified by the GWAS of obesity-related traits[167] although subsequent studies[168,169] show that *IRX3* and *IRX5* (rather than *FTO*) are the functional genes responsible for the GWAS association. The top associated SNP at the *FTO* locus is not associated with physical activity but its effect on BMI decreases with the increase of physical activity level[88,170], consistent with the interaction effects of the *FTO* locus with physical activity or sedentary behaviour for obesity-related traits identified in this study (
Table 2-10 and
Table 2-11). In addition, 5 of the 22 BMI vQTLs were in LD ($r^2 > 0.5$) with the variants (identified by a recently developed multiple-environment GEI test) showing significant interaction effects at FDR < 5% (corresponding to $p < 1.16\times10^{-3}$) with at least one of 64 environmental factors for BMI in the UKB[171].

It should be noted that GEI is sufficient but not necessary to generate a vQTL. For the vQTLs that did not show a direct GEI effect in our GEI analysis, we cannot distinguish whether they are due to undetected GEI or direct effects on phenotypic dispersion although GEI is a more likely explanation because of the enrichment of GEI (Figure 2-14), hence these traits and loci are candidates for follow-up studies to identify putative environmental risk factors that may be amendable to lifestyle modification.

Table 2-8 GEI effect between the *CHRNA5-A3-B4* locus and smoking on FFR

Phenotype	Top vQTL SNP	Effect size (Standard error)		P values		
		never smokers (n= 188,860)	ever smokers (n= 160,488)	vQTL analysis	QTL analysis	GEI test
FFR	rs56077333	0.0105 (0.0035)	-0.0453 (0.0042)	1.09E-14	2.11E-06	4.55E-25

Table 2-9 GEI effect between the WNT16-CPED1 locus and age on BMD

Phenotype	Top vQTL SNP	Effect size (Standard error)				*P* values		
		Age group 1: 40-49 years (n = 59,734)	Age group 2: 50-59 years (n = 108,736)	Age group 3: 60-69 years (n = 156,173)	Age group 4: 70-74 years (n = 23,250)	vQTL analysis	QTL analysis	GEI test
BMD	rs10254825	0.1448 (0.0081)	0.1650 (0.0059)	0.1907 (0.0050)	0.1765 (0.0119)	2.01E-45	0	1.16E-07

Table 2-10 Associations of *FTO* locus with obesity-related traits stratified by physical activity (PA) levels

Phenotype	Top vQTL SNP	Effect size (Standard error)			*P* values		
		Low PA group (n = 103,374)	Intermediate PA group (n = 145,889)	High PA group (n = 97,506)	vQTL analysis	QTL analysis	GEI test
BMI	rs11642015	0.1018 (0.0049)	0.0715 (0.0037)	0.0609 (0.0041)	1.73E-73	7.43E-217	1.28E-10
WC	rs1421085	0.0858 (0.0048)	0.0652 (0.0037)	0.0524 (0.0042)	3.27E-52	3.21E-166	1.44E-07
HC	rs1421085	0.0825 (0.0049)	0.0623 (0.0037)	0.0505 (0.0042)	1.65E-48	2.05E-152	5.32E-07

BMR	rs1421085	0.0842 (0.0049)	0.0608 (0.0037)	0.0531 (0.0043)	8.95E-47	9.23E-154	1.52E-06

Table 2-11 Associations of *FTO* locus with obesity-related traits stratified by sedentary behaviour (SB) levels

		Effect size (Standard error)			*P* values		
Phenotype	Top vQTL SNP	SB group 1: 0-5 hours (*n* = 244,215)	SB group 2: 6-11 hours (*n* = 89,712)	SB group 3: 12-17 hours (*n* = 5,445)	vQTL analysis	QTL analysis	GEI test
BMI	rs11642015	0.0694 (0.0027)	0.1001 (0.0052)	0.1085 (0.0234)	1.73E-73	7.43E-217	1.64E-09
WC	rs1421085	0.0593 (0.0028)	0.0879 (0.0050)	0.1089 (0.0223)	3.27E-52	3.21E-166	2.84E-08
HC	rs1421085	0.0577 (0.0028)	0.0815 (0.0052)	0.1199 (0.0230)	1.65E-48	2.05E-152	2.17E-06
BMR	rs1421085	0.0576 (0.0028)	0.0849 (0.0052)	0.1024 (0.0233)	8.95E-47	9.23E-154	1.45E-07

In conclusion, we systematically quantified the FPR and power for four commonly used vQTL methods by extensive simulations and demonstrated the robustness of the Levene's test. We also showed that in the presence of QTL effects the Levene's test statistic could be inflated if the phenotype was non-linearly transformed. We implemented the Levene's test as part of the OSCA software package[172] (URLs) for efficient genome-wide vQTL analysis. We applied OSCA-vQTL to 13 quantitative traits in the UKB and identified 75 vQTL (at 41 near-independent loci) associated with 9 traits, 9 of which did not show a significant QTL effect. As a proof-of-principle, we performed GEI analyses in the UKB with 5 environmental factors/covariates and demonstrated the enrichment of GEI effects among the detected vQTLs. Hence, the vQTL trait-loci combinations we have identified, could be investigated for as-yet-undetermined but measurable environmental risk factors generating GEI, as these factors could be amenable to lifestyle change interventions. However, the conclusions from this study may be only applicable to quantitative traits of polygenic architecture. We caution vQTL analysis for binary or categorical traits, or molecular traits (e.g., gene expression or DNA methylation), for which the methods need further investigation.

2.5 Methods

Simulation study

We used a DGLM[94,146,147] to simulate the phenotype based on two models with simulated SNP data in a sample of 10,000 individuals, i.e., a single-SNP model and multiple-SNP model with two covariates (i.e. age and sex). The single-SNP model can be written as

$y = w\beta_g + e$ with $log(\sigma_e^2) = w\phi_g + log(\sigma^2)$

and the multiple-SNP model can be expressed as

$y = \sum_{j=1}^{l} c_j \, \beta_{c_j} + \sum_{k=1}^{m} w_k \, \beta_{g_k} + e$ with $log(\sigma_e^2) = \sum_{j=1}^{l} c_j \, \phi_{c_j} + \sum_{k=1}^{m} w_k \, \phi_{g_k} + log(\sigma^2)$,

where y is a simulated phenotype; w or w_k is a standardized SNP genotype, i.e., $w = (x - 2f)/\sqrt{2f(1-f)}$ with x being the genotype indicator variable coded as 0, 1 or 2, generated from binomial(2, f) and f being the MAF generated from uniform(0.01, 0.5); c_j is a standardized covariate with c_1 (sex) generated from binomial(1, 0.5) and c_2 (age) generated from uniform(20, 60); e is an error term with mean 0 and variance σ_e^2. To simulate the error term with different levels of skewness and kurtosis, we generated e from five different distributions, including normal distribution, t-distribution with degree of freedom (df) = 10 or 3 and χ^2 distribution with df = 15 or 1. β (ϕ) is the effect on mean (variance) generated from $N(0,1)$ if exists, 0 otherwise. $log(\sigma^2)$ is the intercept of the second linear model which was set to 0. We re-scaled the different components to control the variance explained, i.e., 0.1 and 0.9 for the genotype component and error term, respectively, in the single-SNP model, and 0.2, 0.4 and 0.4 for the covariate component, genotype component and error term, respectively, in the multiple-SNP model. We simulated the SNP effects in four different scenarios: 1) effect on neither mean nor variance (nei), 2) effect on mean only (mean), 3) effect on variance only (var), or 4) effect on both mean and variance (both). We simulated only one causal SNP in the single-SNP model and 4, 40 or 80 causal SNPs in the multiple-SNP model.

We performed vQTL analyses using the simulated phenotype and SNP data to compare four vQTL methods, including the Bartlett's test[142], the Levene's test[144], the Fligner-Killeen test[145] and the DGLM (Supplementary Note 2-2). We also performed the Levene's test with six phenotype process strategies, including raw phenotype (raw), raw phenotype adjusted for covariates (adj), RNIT after adj (rint) (Supplementary Note 2-3), logarithm transformation after adj (log), square transformation after adj (sq), and cube transformation after adj (cub). We repeated the simulation 1,000 times and calculated the FPR and power at $p < 0.05$ at a single SNP level.

The UK Biobank data

The full release of the UKB data comprised of genotype and phenotype data for ~500,000 participates across the UK[46]. The genotype data were cleaned and imputed to the Haplotype Reference Consortium (HRC)[39] and UK10K[173] reference panels by the UKB team. Genotype probabilities from imputation were converted to hard-call genotypes using PLINK2[174] (--hard-call 0.1). We excluded genetic variants with MAF < 0.05, Hardy-Weinberg equilibrium test p value < 1×10^{-5}, missing genotype rate > 0.05 or imputation INFO score < 0.3 and retained 5,554,549 variants for further analysis.

We identified a subset of individuals of European ancestry (n = 456,422) by projecting the UKB PCs onto those of 1000 Genome Project (1KGP)[175]. We then removed one of each pair of individuals with SNP-derived (based on HapMap 3 SNPs) genomic relatedness > 0.05 using GCTA-GRM[53] and retained 348,501 unrelated European individuals for further analysis.

We selected 13 quantitative traits for our analysis (Table 2-1 and Figure 2-4). We adjusted the raw phenotype values for age and the first 10 PCs, excluded from the analysis phenotype values that were more than 5 SD from the mean, and then standardized to z-scores with mean 0 and variance 1 in each gender group.

Genome-wide vQTL analysis

The genome-wide vQTL analysis was conducted using the Levene's test implemented in the software tool OSCA[172] (URLs). The Levene's test used in the study (also known as the median-based Levene's test or the Brown-Forsythe test[144]) is a modified version of the original Levene's test[143] developed in 1960 that is essentially an one-way ANOVA test of the variable $z_{ij} = |y_{ij} - \widetilde{y_i}|$, where y_{ij} is phenotype of the j-th individual in the i-th group and $\widetilde{y_i}$ is the median of the i-th group. The Levene's test statistic

$$\frac{(n-k)}{(k-1)} \frac{\sum_{i=1}^{k} n_i \, (z_{i.} - z_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\, z_{ij} - z_{i.})^2}$$

approximately follows a F distribution with $k-1$ and $n-k$ degrees of freedom under the null hypothesis, where n is the total sample size, k is the number of groups ($k = 3$ in vQTL analysis), n_i is the sample size of the i-th group, i.e. $n = \sum_{i=1}^{k} n_i$, $z_{ij} = |y_{ij} - \widetilde{y_i}|$, $z_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}$, and $z_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} z_{ij}$.

The experiment-wise significance level was set to 2.0×10^{-9}, which is the genome-wide significance level (i.e., 1×10^{-8})[150,151] divided by the effective number of independent traits (i.e. 5.00 for our 13 traits). The effective number of independent traits was estimated based on the phenotypic correlation matrix[176] (Supplementary Note 2-4). To determine the number of near-independent vQTLs, we performed an LD clumping analysis for each trait using PLINK2[174] (--clump option with parameters --clump-p1 2.0e-9 --clump-p2 2.0e-9 --clump-r2 0.01 and --clump-kb 5000). To visualize the results, we generated the Manhattan and regional association plots using the ggplot2 package in R.

GWAS analysis

The GWAS (or genome-wide QTL) analysis was conducted using PLINK2[174] (--assoc option) using the same data as used in the vQTL analysis (note that the phenotype had been pre-adjusted for covariates and PCs). The other analyses, including LD clumping, and visualization, were performed using the same pipelines as those for genome-wide vQTL analysis described above.

GEI analysis

Five environmental factors/covariates (i.e., sex, age, PA, SB and smoking) were used for the direct GEI tests. Sex was coded as 0 or 1 for female or male. Age was an integer number ranging from 40 to 74. PA was assessed by a three-level categorical score (i.e., low, intermediate and high) based on the short form of the International Physical Activity Questionnaire (IPAQ) guideline[177]. SB was an integer number defined as the combined time (hours) spent driving, non-work-related computer using or TV watching. The smoking factor "ever smoked" was coded as 0 or 1 for never or ever smoker. More details about the definition and derivation of environmental factor PA, SB and smoking can be found in the Supplementary Note 2-5, Figure 2-13 and Table 2-6.

We performed a GEI analysis to test the interaction effect between the top vQTL SNP and one of the five environmental factors based on the following model

$y = \mu + \beta_g x_g + \beta_E x_E + \beta_{gE} x_g x_E + e,$

where y is phenotype, μ is the mean term, x_g is mean-centred SNP genotype indicator, and x_E is mean-centred environmental factor. We used a standard ANOVA analysis to test for β_{gE} and applied a stringent Bonferroni-corrected threshold 1.33×10^{-4} (i.e., 0.05/(75×5)) to claim a significant GEI effect.

2.6 URLs

OSCA, http://cnsgenomics.com/software/osca

PLINK2, http://www.cog-genomics.org/plink2

GCTA, http://cnsgenomics.com/software/gcta

UCSC Genome Browser, https://genome.ucsc.edu/

SMR, http://cnsgenomics.com/software/smr

The UKB data, http://www.ukbiobank.ac.uk/

vQTL summary statistics for the 13 UKB traits, http://cnsgenomics.com/software/osca/#DataResource

2.7 Acknowledgements

This research was supported by the Australian Research Council (DP160101343 and

DP160101056, FT180100186), the Australian National Health and Medical Research Council (1078037, 1078901, 1113400, 1107258 and 1083656), and the Sylvia & Charles Viertel Charitable Foundation. This study makes use of data from the UK Biobank (project ID: 12514). A full list of acknowledgments of this data set can be found in Supplementary Note 2-6. Competing Interests: The authors declare that they have no competing interests. Author contributions: JY and AFM conceived the study. JY, HW and AFM designed the experiment. FZ developed the software tool. HW performed simulations and data analyses under the assistance or guidance from JY, JZ, YW, KK, AX and MZ. JEP, MEG, NRW and PMV provided critical advice that significantly improved the experimental design and/or interpretation of the results. PMV, NRW and JY contributed resources and funding. HW and JY wrote the manuscript with the participation of all authors. The vQTL summary statistics of all SNPs for the 13 UKB traits are available at the OSCA website (URLs). All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

2.8 Supplementary Notes

Supplementary Note 2-1 The theoretical derivation of vQTL as a consequence of GEI

It has been shown by Pare et al.[131] that the interaction of a genetic variant with a genetic or environmental factor for a trait (e.g., GEI) can lead to differences in variance of the trait across genotype classes of the variant. Take GEI as an example. Under a GEI model, a phenotype y is affected by a genetic variant x_g, an environmental factor x_E, and an interaction term $x_gx_{E,}$, i.e.,

$$y = \mu + \beta_g x_g + \beta_E x_E + \beta_{gE} x_g x_E + e$$

where μ is the intercept term, $\beta_g, \beta_E, \beta_{gE}$ are the effects of x_g, x_E and x_gx_E, respectively, and e is the residual. The phenotypic variance conditional on the genotype of the variant is

$$\begin{aligned} Var(y|x_g) &= Var(\mu + \beta_g x_g + \beta_E x_E + \beta_{gE} x_g x_E + e) \\ &= Var((\beta_E + \beta_{gE} x_g) x_E + \mu + \beta_g x_g + e) \\ &= (\beta_E + \beta_{gE} x_g)^2 Var(x_E) + Var(e) \end{aligned}$$

, assuming that x_g, x_E and e are independent of each other. This equation shows that the phenotypic variance given a genotype is dependent on the genotype in the presence of GEI (i.e., $\beta_{gE} \neq 0$).

Supplementary Note 2-2 The Bartlett's test, the Fligner-Killeen test, and the double generalized linear model (DGLM) test

We evaluated four variance quantitative trait locus (vQTL) methods by simulation. Details of the Levene's test have been described in the Methods section of the main text, and details of the other three methods are described below.

The Bartlett's test[142] is one of the earliest methods used to test the inequality of variance but known to be sensitive to the violation of normality assumption[141]. The Bartlett's test-statistic is

$$\frac{(n-k)ln(S_p^2) - \sum_{i=1}^{k}(n_i - 1)ln(S_i^2)}{1 + \frac{1}{3(k-1)}(\sum_{i=1}^{k}(\frac{1}{n_i - 1}) - \frac{1}{n-k})} \sim \chi_{k-1}^2$$

where *n* is the total sample size; *k* is the number of groups; n_i is the sample size of the *i*-th group, $n = \sum_{i=1}^{k} n_i$; S_i^2 is the sample variance in the *i*-th group; S_p^2 is the pooled estimate of the variance, $S_p^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)S_i^2$. We used the *bartlett.test()* function in R for data analysis.

The Fligner-Killeen (median) test[145] is a rank-based method with similar performance to the Levene's test. The Fligner-Killeen test-statistic is

$$\frac{\sum_{i=1}^{k} n_i\, (\overline{A}_i - \overline{a})^2}{V^2} \sim \chi_{k-1}^2$$

where *n* is the total sample size; *k* is the number of groups; n_i is the sample size of the *i*-th group, $n = \sum_{i=1}^{k} n_i$; *a* is the "rank score" assigned by $\Phi^{-1}(\frac{1+\frac{j}{n+1}}{2})$ with *j* being the rank of all observations based on $|y_{ij} - \widetilde{y}_i|$, $\widetilde{y}_i$ being the median of the *i*-th group and Φ^{-1} being the standard normal quantile function; $\overline{A}_i$ is the mean rank score of the *i*-th group; $\overline{a}$ is the mean rank score of all observations; V^2 is the sample variance of rank scores of all observations. We used the *fligner.test()* function in R for data analysis.

Ronnegard et al.[94,146] proposed a double generalized linear model (DGLM)[147] that contained two linear predictors, one for the effect on the trait mean and the other for the effect on the trait variance:

$$E(y|u, u_d) = \mu;\ \mu = Xb + Zu$$

$$var(y|u, u_d) = \phi;\ log(\phi) = X_d b_d + Z_d u_d$$

where *y* is the phenotype; *u* and u_d are the random effects on the mean and variance (dispersion), respectively; *b* and b_d are the fixed effects on the mean and variance (dispersion), respectively. We used “dglm” package in R for data analysis.

Supplementary Note 2-3 Rank-based inverse-normal transformation

We used the simulated data to compare several phenotype processing strategies. Rank-based inverse-normal transformation (RINT) was conducted based on the formula below[178,179]

$$y_i^t = \Phi^{-1}\left(\frac{r_i - c}{n - 2c + 1}\right)$$

where r_i is the ordinary rank of the *i*-th observation; *n* is the total number of observations; *c* is a constant value (set to 0.5 in this study); Φ^{-1} is the standard normal quantile function; y_i^t is the transformed value for the *i*-th observation. For RINT after covariate adjustment, we first adjusted the phenotypes for covariates and then transformed the residuals by RINT.

Supplementary Note 2-4 The effective number of independent traits

As some phenotypes were correlated with each other (Figure 2-4), we used an eigendecomposition analysis to estimate the effective number of independent traits[176]. Let y be a vector of *p* phenotypes and V be the variance-covariance matrix of vector y. The eigen decomposition of matrix V is

V = Q’ΛQ

where Q is the matrix of eigenvectors and Λ is the diagonal matrix comprised of the ordered eigenvalues $\lambda_1 \dots \lambda_p$. The effective number of *p* phenotypes can be estimated as[176]:

$$\frac{(\sum_{k=1}^{p} \lambda_k)^2}{\sum_{k=1}^{p} \lambda_k^2}$$

Supplementary Note 2-5 Definitions of the three environmental factors - PA, SB and smoking

Physical activity (PA) was assessed based on the questions from International Physical Activity Questionnaire (IPAQ)[180], including the number of days per week of walking (DayW), the number of days per week of moderate physical activity (DayW), the number of days per week of vigorous physical activity more than 10 minutes (DayV), the duration of walking (DurW), the duration of moderate physical activity (DurM), and the duration of vigorous physical activity (DurV) (Table 2-6). According to the IPAQ analysis guideline[177], the metabolic equivalents (MET) minutes for walking (METW), moderate physical activity

(METM), vigorous physical activity (METV), and the total MET (METT) minutes were calculated by

$\text{METW} = 3.3 \times \text{DayW} \times \text{DurW}$

$\text{METM} = 4.0 \times \text{DayM} \times \text{DurM}$

$\text{METV} = 8.0 \times \text{DayV} \times \text{DurV}$

$\text{METT} = \text{METW} + \text{METM} + \text{METV}$

The physical activity level was then labelled as 1) “high” (coded as 3) when “DayV≥3 and METT≥1500” or “DayW+DayM+DayV≥7 and METT≥3000”; 2) “moderate” (coded as 2) when “DayV≥3 and DurV≥20” or “DayM≥5 and DurM≥30” or “DayW≥5 and DurW≥30” or “DayW+DayM+DayV≥5 and METT≥600”; 3) “low” (coded as 1) when no activity or some activity was reported but not enough to meet the criteria above.

Sedentary behaviour (SB) was defined as the sum of the time spent driving (TimeD), non-work-related computer using (TimeC) or TV watching (TimeTV) (Table 2-6). We removed outliers 5 SD from the mean; the remaining data ranged from 0 to 17 hours.

Smoking was assessed based on the answers to two questions about current tobacco smoking (CurS) and past tobacco smoking (PastS) (Table 2-6). Individuals were classified as “never smoker” (coded as 0) if CurS = “no” and PastS = “tried once or twice” or “never”. Individuals were classified as “ever smoker” (coded as 1) if CurS = “most days” or “occasionally”, or PastS = “most days” or “occasionally”.

Supplementary Note 2-6 Acknowledgements

This study has been conducted using the UK Biobank resource under Application Number 12505. The UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government and the Northwest Regional Development Agency. It has also had funding from the Welsh Assembly Government, British Heart Foundation and Diabetes UK.

3

Chapter 3: Inflation in test-statistics for genotype-by-environment and genotype-by-genotype interactions due to linkage disequilibrium

Part of this chapter has been published in *Science Advances* in 2019 and part has been incorporated in a manuscript under review.

Inflation in test-statistics for genotype-by-environment and genotype-by-genotype interactions due to linkage disequilibrium

3.1 Abstract

Most genetic analyses of human complex traits mainly focus on additive genetic effects, partially because of the complexity of statistical modeling for non-additive effects. Linkage disequilibrium (LD) is known to be able to create phantom signals for genotype-by-environment interaction (GEI) or genotype-by-genotype interaction (i.e., epistasis) test. Here, we performed theoretical derivations to quantify inflation due to LD in test-statistics for vQTL analysis, which can be used to infer GEIs. We examined the vQTLs identified in UKB in Chapter 2 based on our derivation and found no evidence for them to be phantom vQTLs. For epistasis test, we used simulations based on whole genome sequence (WGS) data from the UK10K project to quantify the inflation for different genotyping strategies. We found the level of inflation was related to genotyping strategies and increased almost linearly with the increase of the variance explained by the additive causal variant. Our study quantifies the inflation due to LD in test-statistics for vQTL and epistasis test under a range of scenarios and provides an important caveat for the analysis of interaction effects using genotyped or imputed variants in traits for which there is a large additive genetic effect.

3.2 Introduction

Linkage disequilibrium (LD), the correlation between alleles at different loci, is a double-edged sword for the genetic study of human complex traits. On one side, LD enables the utility of observed genetic marker variants (usually SNPs) as proxies to tag unobserved causal variants in the association study for human complex traits[181]. On the other side, however, it is difficult to distinguish causal variants from highly correlated marker variants due to LD (the so-called fine-mapping problem[182]). In addition, LD can also be a source of phantom association for studies to detect interaction effects, such as GEI and epistasis.

The vQTL approach can be used to search genetic variants involving GEIs without measuring environmental factors, which has been demonstrated in Chapter 2. LD can produce phantom vQTL signals in presence of a causal variant with only an additive effect, which has been demonstrated by a simulation study and observed in real vQTL analysis on DNA methylation[133,134]. In other words, if the underlying causal QTL is not well imputed or not well tagged by a genotyped/imputed variant, the untagged variation at the causal QTL will inflate the vQTL test-statistic, potentially leading to a spurious vQTL association.

The phantom signal created by LD can also be seen for the epistasis test. In 2014, Hemani et al.[183] performed an exhaustive epistasis test on every pair of 528,509 genotyped SNPs for 7,339 gene expressions in a discovery sample of 864 individuals, identified and replicated a few SNP pairs with significant interaction effects. Following this publication, however, Wood et al.[184] found that the significance of interaction effects identified in Hemani et al. could be removed by including a third genetic variant detected using WGS dataset more strongly associated with the gene expression, which demonstrated that the epistasis test involving an imperfect tagged genetic variant can be inflated by a causal variant with an additive effect[185].

The explicit statistical mechanisms of phantom vQTL and phantom epistasis mentioned above are still elusive. For phantom vQTL, the quantitative relationship between LD and phantom vQTL signals and to what extent it will affect the vQTL analysis are largely unknown. For phantom epistasis, Wood et al.[184] explained that the inflation was raised by two genetic variants in a moderate level of linkage disequilibrium (LD), corresponding to the cis-cis interactions. Furthermore, de los Campos et al.[185] also claimed that the inflation for epistasis test required two genetic marker variants and the causal additive variant mutually in LD. However, neither Wood et al. nor de los Campos et al. could explain the cis-trans interactions, which accounted for the majority (462/501) of the SNP pairs with significant epistatic effect discovered in Hemani et al.[183].

Here we used theoretical derivation and simulation study to quantify the inflation level of phantom vQTL and phantom epistasis. This work partially explains why interaction effects are harder to study than additive genetic effects and raises caution when performing association scanning for interaction effects.

3.3 Results

Model overview

Let us consider two genetic variants: causal variant A and marker variant B. The LD between variant A and B can be measured by a few parameters, including D, D', and r^2, based on their allele frequency and haplotype frequency[186] (Supplementary Note 3-1). Suppose the causal variant A has an additive genetic effect (b_c) on phenotype *y*:

$$y = \mu + b_c x_a + e$$

, where μ is the intercept, x_a is the genotype value of variant A (coded as 0, 1, 2), and e is the residual term assumed with a mean 0 and variance σ^2. The causal variant A is usually not observed in practice, so the marker variant B can be used to tag the causal variant A. And the test-statistic for additive genetic association (or GWAS) at marker variant B is a function of sample size, the phenotypic variance explained by variant A, and the LD r^2 between variant A and B[187,188] (Supplementary Note 3-2, Supplementary Note 3-3, and Supplementary Note 3-4).

Phantom vQTL

It is less acknowledged that the marker variant B not only can tag the additive genetic effect of causal variant A, but also has a variance heterogeneity. We derived the expected genetic effect on phenotypic mean (b_m) and phenotypic variance (β_m) at marker variant B (Supplementary Note 3-3) based on genotype frequencies and conditional genotype frequencies of these two variants (Supplementary Note 3-2):

$$b_m = \frac{b_c(p_{AB} - p_A p_B)}{p_B(1 - P_B)}$$

$$\beta_m = \frac{b_c^2[(1 - 2p_B)p_{AB}^2 + (2p_A p_B + p_B - 1)p_B p_{AB} + (1 - p_A - p_B)p_A p_B^2]}{p_B^2(1 - p_B)^2}$$

. This variance heterogeneity could be detected by vQTL test. We derived that the Levene's test-statistic due to the phantom vQTL effect was a function of sample size *n*, variance explained by the causal variant q_c^2, allele frequency of the causal variant p_A, allele frequency of the marker variant p_B, and the haplotype frequency p_{AB} (Supplementary Note 3-5):

$$F_{\text{Levene}} = \frac{n}{\pi - 2}(1 - \frac{[\sqrt{{\sigma^2}_m} {p_B}^2 + \sqrt{{\sigma^2}_m + \beta_m} 2p_B(1 - p_B) + \sqrt{{\sigma^2}_m + 2\beta_m}(1 - p_B)^2]^2}{{\sigma^2}_m + 2(1 - p_B)\beta_m})$$

, where ${\sigma^2}_m = \sigma^2 + \frac{2{b_c}^2(p_B - p_{AB})p_{AB}}{p_B^2}$. This formula was confirmed by simulation shown in Figure 3-1. Our theory was consistent with the observation of pervasive phantom vQTLs for molecular traits with large-effect QTLs (e.g., DNA methylation[134]).

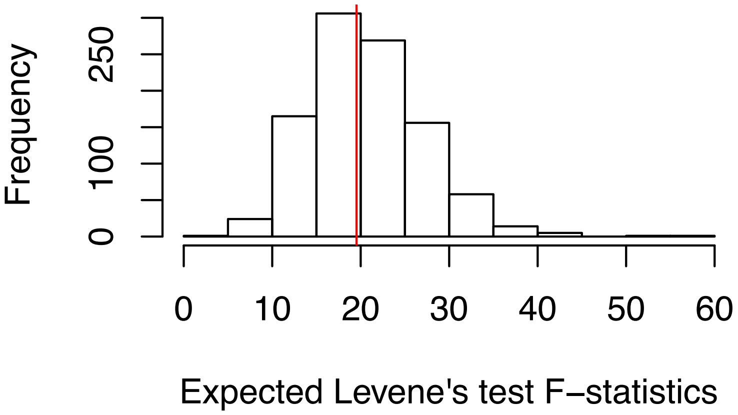


Figure 3-1 Verification of the expected Levene's test *F*-statistic due to phantom vQTL effect by simulation.

We simulated two variants A (p_A = 0.7) and B (P_B = 0.6) in LD (P_{AB} = 0.6, LD r^2 = 0.64, and LD D' = 1) from multinomial(2, (P_{AB}, P_{Ab}, PaB, Pab)) and a phenotype based on the causal variant A explaining 5% variance in 350,000 individuals. Shown is the distribution of *F*-statistics from the Levene's test using the simulated data with 1,000 replicates. The red line indicates the theoretical value.

To investigate whether there were phantom vQTLs for the vQTLs identified in Chapter 2 using UKB dataset, we then computed F_{Levene} given a number of parameters including p_{AB} (equivalent to D' ranging from -1 to 1), p_a (ranging from 0.001 to 0.5, equivalent to p_A from 0.999 to 0.5), p_b (ranging from 0.05 to 0.5, equivalent to p_B from 0.95 to 0.5), q_c^2 (= 0.005, 0.01 or 0.02) and n (= 350,000) (Figure 3-2). The result showed that for a causal QTL with q^2 < 0.005 and MAF > 0.05, the largest possible phantom vQTL *F*-statistic was smaller than 2.69 (corresponding to a p-value of 6.8×10^{-2}; Figure 3-2). This explains why there were thousands of genome-wide significant QTLs but no significant vQTL for height (

Table 2-2 and Figure 2-6). This result also suggests that the vQTLs detected in Chapter 2 are very unlikely to be phantom vQTLs because the estimated variance explained by their QTL effects were all smaller than 0.005 except for rs10254825 at the *WNT16* locus on BMD (q^2 = 0.014) (Figure 3-3). However, our numerical calculation also indicated that for a QTL with MAF > 0.3 and q^2 < 0.02, the largest possible phantom vQTL *F*-statistic was smaller than 5.64 (corresponding to a p-value of 3.6×10^{-3}), suggesting rs10254825 is also unlikely to be a phantom vQTL. Note that we used the variance explained estimated at the top GWAS SNP to approximate q^2 of the causal QTL so that q^2 was likely to be underestimated because of imperfect tagging. However, considering the extremely high imputation accuracy for common variants[39], the strong LD between the causal QTLs and the GWAS top SNPs observed in a previous simulation study based on whole-genome-sequence data[150], and the overestimation of variance explained by the GWAS top SNPs because of winner's curse, the underestimation in causal QTL q^2 is likely to be small. In addition, we re-ran the vQTL analysis with the phenotype adjusted for the top GWAS variants within 10Mb of the top vQTL SNP; the vQTL signals after this adjustment were highly concordant with those without adjustment (Figure 3-4).

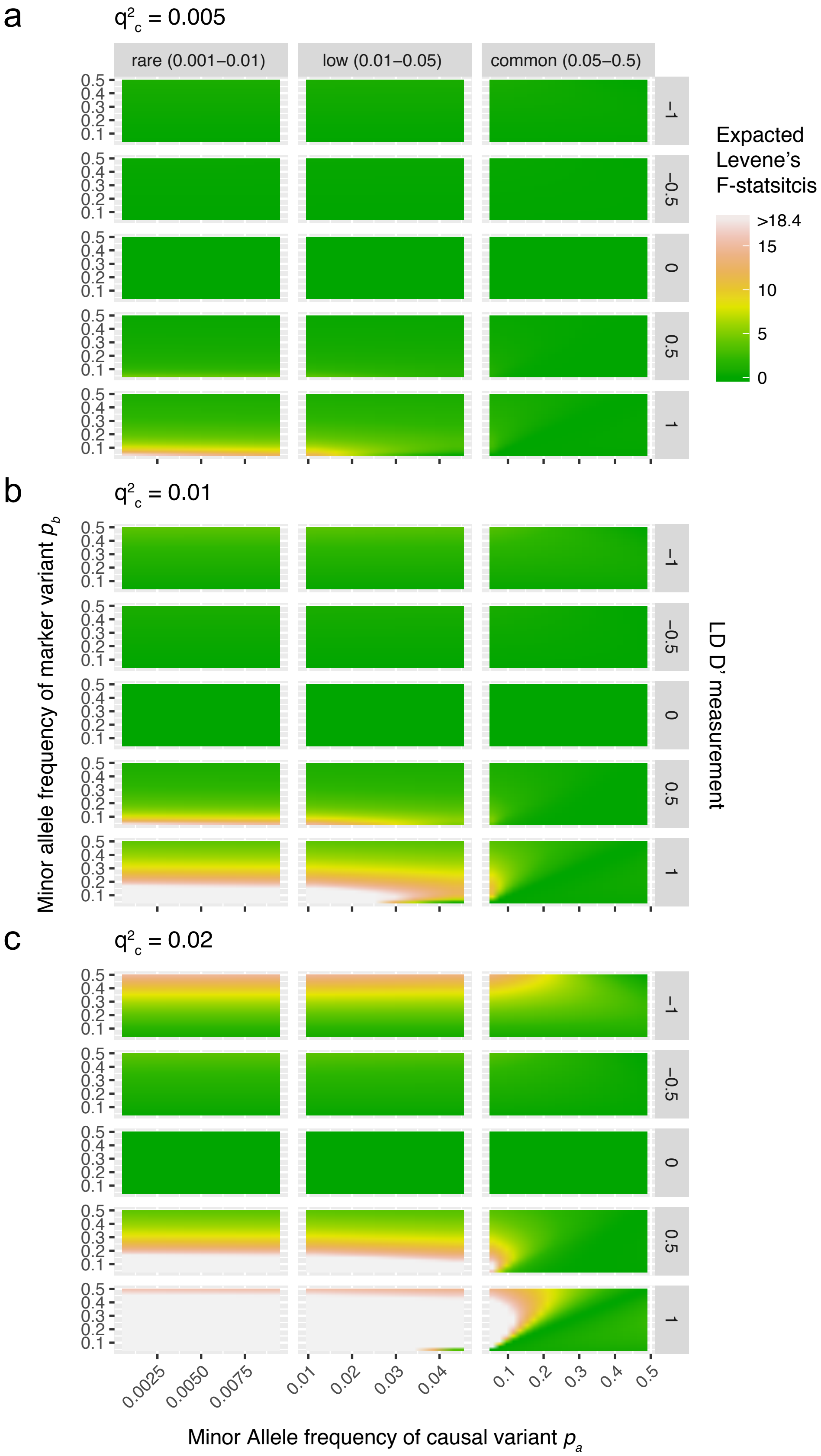


Figure 3-2 Expected phantom vQTL *F*-statistics from Levene's test.

We calculated the expected phantom vQTL F-statistics given a number of parameters including p_{AB} (equivalent to LD D' from -1 to 1), p_a (ranging from 0.001 to 0.5), p_b (ranging from 0.05 to 0.5), q_c^2 (= 0.005, 0.01 or 0.02) and n (= 350,000). An F value of 18.4 is equivalent to a genome-wide significant p-value of 1×10^{-8}.

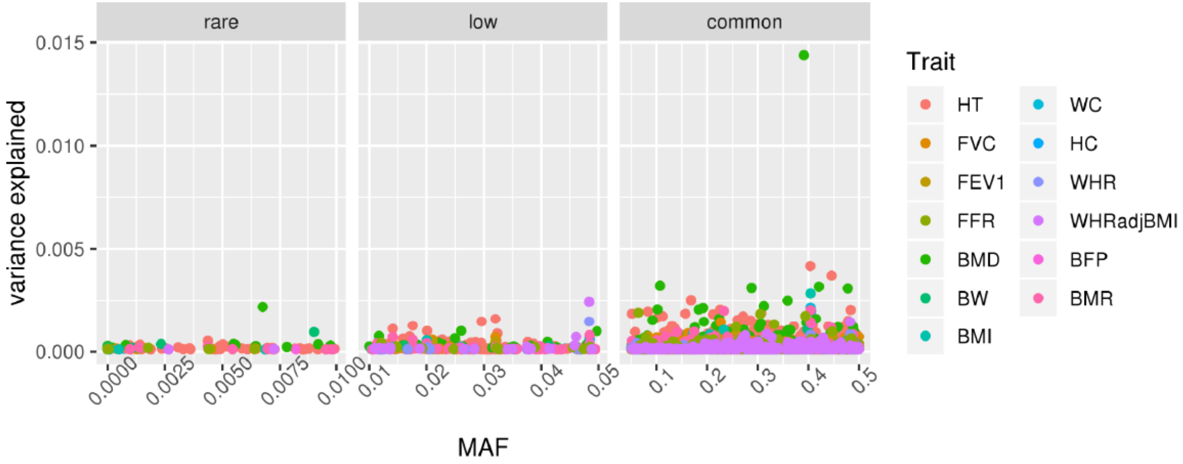


Figure 3-3 Estimated variance explained by top QTL SNPs for the 13 UKB traits.

Note that because the phantom vQTL signals at common SNPs can be induced by rare (MAF≤0.01) or low-frequency (0.01≤MAF<0.05) variants, we extended our GWAS analysis to all 44,741,800 imputed variants (MAF<0.05). The estimated variance explained by each GWAS top SNP is plotted against its MAF.

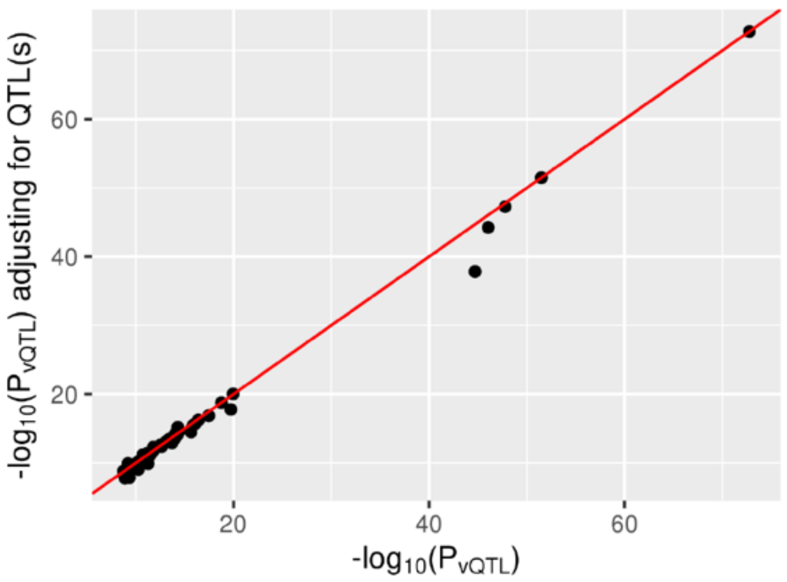


Figure 3-4 vQTL test statistics (-log$_{10}$(P$_{vQTL}$)) from analyses with and without adjusting the phenotype for the QTL effect(s) of the top GWAS SNP(s) within 10Mb of the top vQTL SNP.

The red line represents the line with slope 1 and intercept 0.

We further showed that there was no evidence for epistatic interactions, which could be another source for vQTL signals, between the top vQTL SNPs and any other SNP located more than 10 Mb away or on a different chromosome (Figure 3-5). Note that we did not perform epistatic test for SNP pairs within 10 Mb to avoid phantom epistatic signals by LD[184].

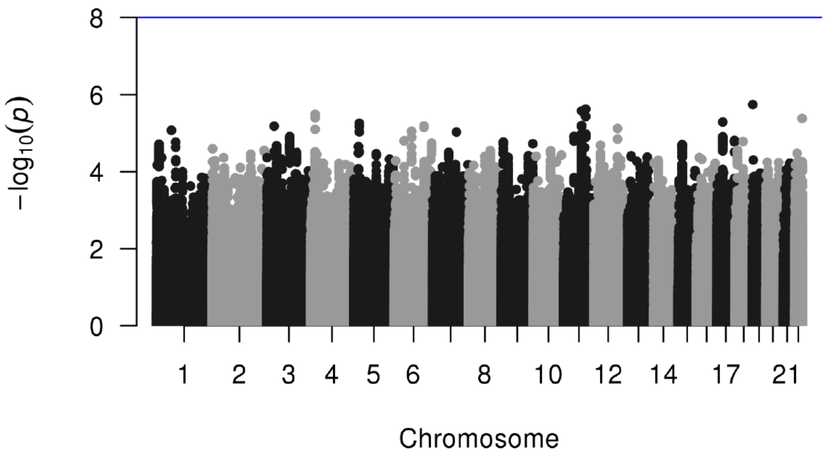


Figure 3-5 Manhattan plot of epistasis analysis for one of top vQTL SNPs.

We conducted epistasis analysis between each of 75 top vQTL SNPs and any other SNPs in more than 10 Mb distance or on a different chromosome for the relevant trait using PLINK2[174] (--epistasis option). The blue horizontal line represents the genome-wide significance level (i.e., p-value = 1×10^{-8}). Shown are the results from the epistasis analysis with the top vQTL SNP rs10913469 for waist circumference (WC).

Phantom epistasis

LD-induced phantom signals can also be seen for epistasis tests. Let us consider the marker variant B (x_b), which is in LD with causal variant A (x_a), is included in an epistasis test with variant C (x_c). One simple test for epistasis between variant B and C is to include a product term in a multiple linear regression model:

$$y = \mu + b_1 x_b + b_2 x_c + b_3 x_b x_c + e$$

, and then the test statistics for epistasis is the t-test or partial F-test for the term $x_b x_c$. We explored the situation where variant B and variant C were in no LD, which was not well studied previously[184,185].

We performed simulation study using WGS data from UK10K. Firstly a quantitative phenotype was simulated based on one causal variant on chromosome 21 with only the additive genetic effect (Methods section). Secondly, we conducted association study across genetic variants on chromosome 21 captured using four different genotyping strategies (i.e. WGS variants (WGS), array-based genotyped variants (array), and imputed variants based on reference panel HapMap project (Hapmap) or 1000 genome project (1KG3)) (Methods). Finally, we performed the epistasis test between the top variant on chromosome 21 (cis variant) and each of common (MAF≥0.01) array-based genetic variants on chromosome 22 (trans variants) and evaluated the inflation using genomic inflation factor. The means of genomic inflation factor for the interaction term across 540 replicates were calculated and shown in Figure 3-6.

We found that the test statistics were not inflated for WGS genotype data but inflated for array or imputed genotype data. The level of inflation for array-variants was higher than that for imputed-variants. This was consistent with the level of imperfect tagging qualified using LD r^2 between top variant and causal variant (Figure 3-7). And also, the level of inflation increased with the increase of variance explained by the additive effect of the causal variant (Figure 3-6). We explored four additional models (Methods) aiming to correct the epistasis test inflation. However, none of them can correct the inflation, and the model 3 even deflated the test-statistics (Figure 3-8), suggesting that the scale of x_b or x_c would not affect the test-statistic of the interaction term $x_b x_c$.

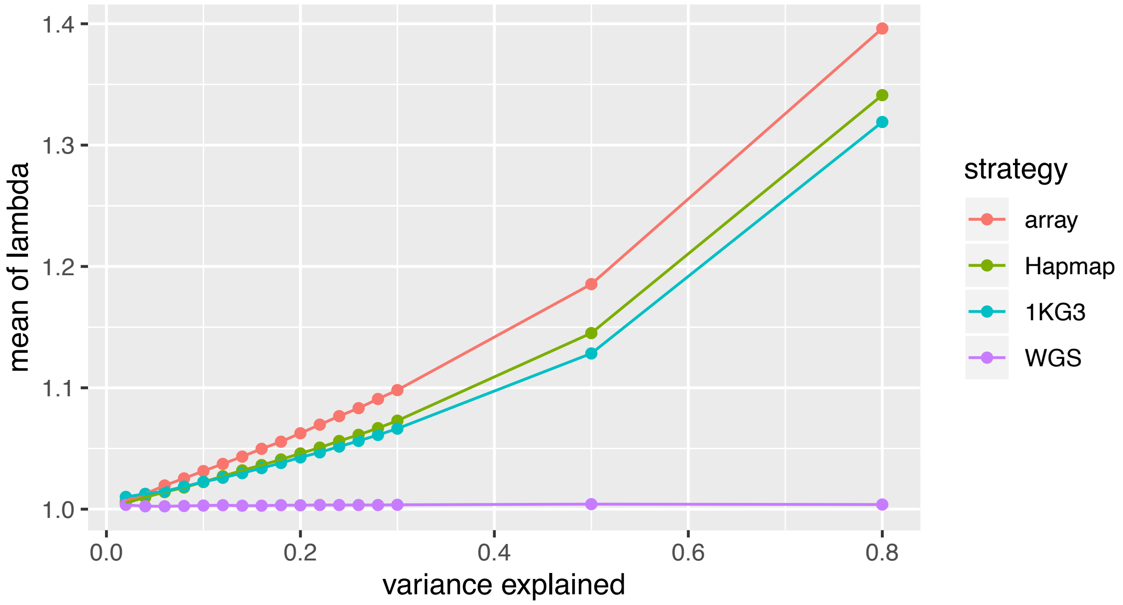


Figure 3-6 The inflation of test-statistics for epistasis test for simulated phenotype with different variance explained using four different genotyping strategies.

The phenotype was simulated based on one causal variant on chromosome 21 with variance explained ranging from 0.02 to 0.8. The epistasis test was conducted between the top variant on chromosome 21 identified using four different genotyping strategies and all common (MAF≥0.01) array-based genetic variants on chromosome 22. The means of genomic inflation factors of test statistics for interaction term across 540 replicates were reported on y-axis.

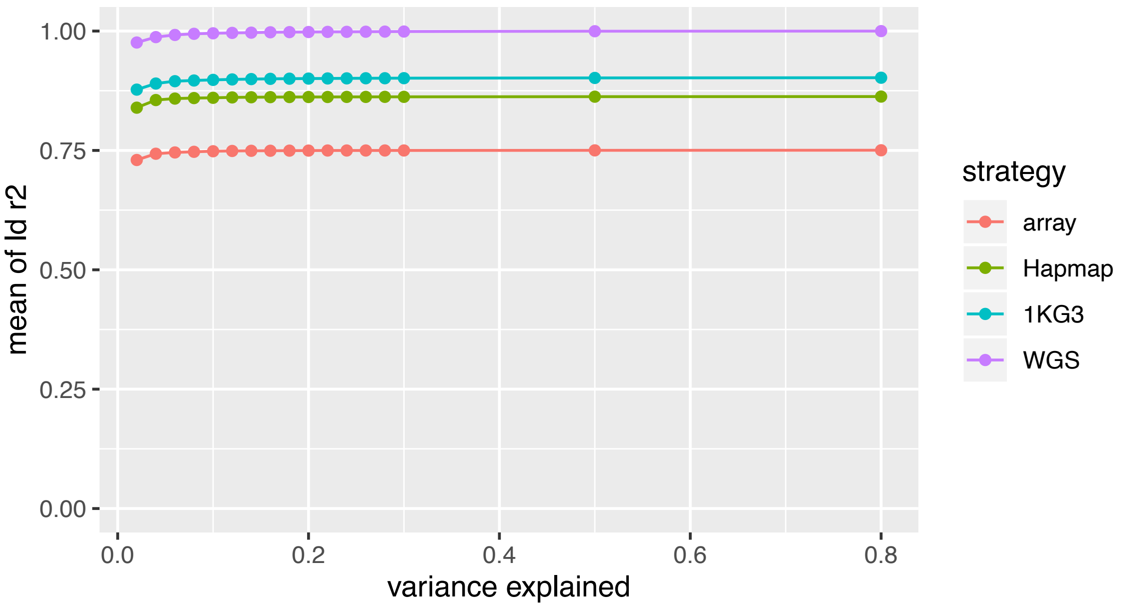


Figure 3-7 The LD r^2 between the causal variant and top variant identified using four different genotyping strategies with different variance explained.

The means of LD r^2 across 540 replicates were reported on y-axis.

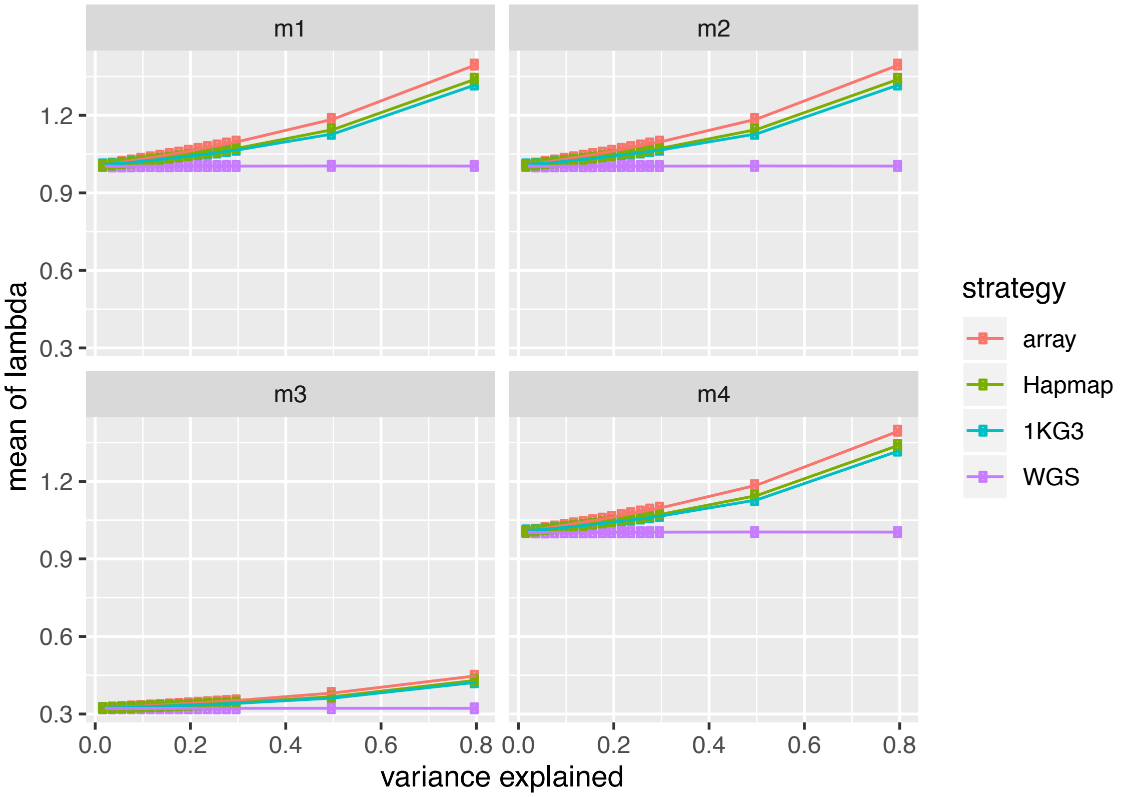


Figure 3-8 The inflation of test-statistics for epistasis test using four additional models.

The means of genomic inflation factor λ across 540 replicates were reported on y-axis. See Methods section for the details about the four additional models for epistasis tests.

3.4 Discussion

In this chapter, we quantified the inflation in test-statistics for phantom vQTL and phantom epistasis. We used theoretical derivation to calculate the expected value of test-statistics for median-based Levene's test, which was the method we chose to conduct vQTL analysis in Chapter 2. We found vQTLs we identified in chapter 2 in UKB cannot be explained by phantom vQTLs. We further pre-fitted the top QTL and found consistent test-statistics before and after fitting. We did not find any evidence for these vQTLs being explained by epistasis.

For phantom epistasis, we performed a simulation study to demonstrate the inflation of epistasis test between two genetic variants not in LD, which was corresponding to the cis-trans interactions observed in Hemani et al.[183] but not explained in either Wood et al.[184] or de Los Campos et al.[185]. We further quantified the level of epistasis inflation and found it was a function of variance explained by the causal variant and LD between the tagging variant and causal variant. We had not found any solutions for correcting this inflation except the causal variants were captured by the genotype data (like WGS data in our simulation). The reason for discrepancies between our study and de los Campos et al. paper[185] needs further investigation.

The phantom signals are caused by the imperfect tagging between the causal and marker variants. With more WGS data available, supposed to capture almost all genetic variants, the LD between causal and marker variants would diminish, which is likely to greatly solve the problem of phantom signals. Without WGS, it is a suboptimal way to detect the phantom signals by pre-fitting the largest QTL and looking at the test-statistics before and after fitting, although a consistency of test-statistics could not completely exclude the possibility of phantom signals, as the largest QTL observed is not necessarily the causal variant or the marker variant perfectly tagging the causal variant. Our analysis showed that the phantom signals for both vQTL and epistasis analysis are positively related to the variance explained by the additive QTL. So, this problem is more likely to occur for the traits with big effect variants, such as gene expression or DNA methylation. Therefore, we did not observe in our vQTL analysis for 13 traits in UKB with a sample size as large as ~350,000. The heritability analysis to estimate the overall variance explained by epistatic effect[71] for complex traits seems unlikely to be affected either. But with a larger sample, the phantom signal may become more concerning, as it also increases with the sample size.

3.5 Methods

Genotype in simulation study for phantom epistasis

The genotype data was generated based on WGS data from UK10K project[173] using four different strategies, which has been described previously[56,150]: 1) the WGS data from UK10K project containing about 17.6 million variants for 3,642 unrelated European individuals after quality control (WGS); 2) the subset of WGS variants captured by the array of Illumina CoreExome (array); 3) and 4) the imputed variants based on array-variants using the software IMPUTE2[189] with phase 2 of HapMap project[190] (Hapmap) or phase 3 of 1000 Genomes Project[191] (1KG3) as the reference panel. In this simulation study, we only used the genetic variants on chromosomes 21 and 22. The number of common (MAF≥0.01) variants and all variants for different genotypes can be found in Table 3-1.

Table 3-1 The number of variants on chromosomes 21 and 22 for genotype data generated using different strategies.

		Chromosome 21		Chromosome 22	
Strategy	Chromosome	Common variants	All variants	Common variants	All variants
Hapmap	21	31872	32942	30578	32399
Array	21	3813	4415	3886	4842
UK10K	21	116907	241712	114272	233568
1KG3	21	128546	273274	130050	289916

Simulated Phenotype

We used the method described before[150] to simulate the phenotype, which was based on one causal variant randomly chosen from WGS variants:

$$y = wu + e$$

, where w was $\frac{x-2f}{\sqrt{2f(1-f)}}$ with x being the genotype value (coded as 0, 1, 2) and f being the allele frequency, and e was the error term generated from a normal distribution $N(0, var(wu)\left(\frac{1}{q^2} - 1\right))$ with q^2 being the variance explained by the genetic value ranging from 0.02 to 0.8. The previous study[150] replicated the simulation with 50,000 common and 50,000 rare causal variants chosen across the whole genome. We only retained the replicates with 719 common (MAF≥0.01) variants on chromosome 21 in this study.

Association study

The association analysis was performed on all genetic variants on chromosome 21 using the software PLINK2[174] (option "--assoc"). We selected the one with the largest R^2 value as the top variant for different strategies, as there could be more than one variant with p-values equal to 0 when the simulated variance explained was large.

Epistasis test

The epistasis test was the multiple linear regression model with a cross-product/interaction term to model the additive-by-additive epistatic effect:

$$y = \mu + b_1 x_b + b_2 x_c + b_3 x_b x_c + e$$

, where x_{b} is genotype value (coded by 0, 1, 2) of the top variant in association analysis on chromosome 21, and x_c is genotype value for each of the 3,886 common array-based variants on chromosome 22. The test for the interaction term (*t*-test or partial *F*-test) was performed using function *lm()* in R language.

To correct the inflation, we investigated four additional models trying to remove the correlation between x_b or x_c with the interaction term $x_b x_c$ by centralizing x_b and/or x_c, or two-step fitting. For example, $Cov(x_b, x_b x_c) = Var(x_b)E(x_c)$, so $Cov(x_b, x_b x_c^c) = 0$.

More specifically, model 1 included one genotype value (x_c) centralized (x_c^c)

$$y \sim x_b + x_c^c + x_b x_c^c$$

, and model 2 included two genotype values (x_b and x_c) centralized (x_b^c and x_c^c)

$$y \sim x_b^c + x_c^c + x_b^c x_c^c$$

; in model 3, we fitted x_b and x_c first and then took the residual (y') to be regressed on the interaction term $x_b x_c$

$$y' \sim x_b x_c;\ y' = resid(y \sim x_b + x_c)$$

, and in model 4, we first regressed $x_b x_c$ on x_b and x_c and then took the residual (x_{bc}') to be fitted with x_b and x_c on y

$$y \sim x_b + x_c + x_{bc}{}'; \ x_{bc}{}' = resid(x_b x_c \sim x_b + x_c)$$

We quantified the inflation level using the genomic inflation factor $\lambda = median(\chi^2)/0.455$ for interaction term. In practice, we excluded the simulation replicates with no solution for epistasis test due to collinearity, in two situations where $x_b x_c$ were all 0s, or $x_b x_c$ were all the same as x_b or x_c, and eventually 540 replicates were analyzed.

3.6 Supplementary Notes

Supplementary Note 3-1 LD between two variants

Allele and haplotype frequencies

	Variant B		
Variant A	Major allele B	Minor allele b	Allele frequency
Major allele A	p_{AB}	$p_{Ab} = p_A - p_{AB}$	p_A
Minor allele a	$p_{aB} = p_B - p_{AB}$	$p_{ab} = 1 - p_A - p_B + p_{AB}$	$p_a = 1 - p_A$
Allele frequency	p_B	$p_b = 1 - p_B$	1

The haplotype frequency p_{AB} and LD between variant A and B as a function of p_A and p_B

Measures	Definition	Maximum value	Minimum value
p_{AB}	-	$min[p_A, p_B]$	$p_A + p_B - 1$
D	$D = p_{AB} - p_A \times p_B$	$min[p_A(1-p_B), p_B(1-p_A)]$	$-(1-p_A)(1-p_B)$
D'	$D' = \frac{D}{\min[p_A(1-p_B), p_B(1-p_A)]}$, if D > 0 $D' = \frac{D}{min[p_A p_B, (1-p_A)(1-p_B)]}$, if D < 0	1	-1
r^2	$r^2 = \frac{D^2}{p_A p_B (1-p_A)(1-p_B)}$	$min[\frac{p_A(1-p_B)}{(1-p_A)p_B}, \frac{(1-p_A)p_B}{p_A(1-p_B)}]$	0

Supplementary Note 3-2 Genotype frequencies of the two variants

Genotype frequencies of the two variants

	Genotype BB	Genotype Bb	Genotype bb	Genotype Frequency
Genotype AA	$p_{AABB} = p_{AB}^2$	$p_{AABb} = 2p_{AB}p_{Ab}$ $= 2p_{AB}(p_A - p_{AB})$	$p_{AAbb} = p_{Ab}^2 = (p_A - p_{AB})^2$	p_A^2
Genotype Aa	$p_{AaBB} = 2p_{AB}p_{aB}$ $= 2p_{AB}(p_B - p_{AB})$	$p_{AaBb} = 2(p_{AB}p_{ab} + p_{Ab}p_{aB})$ $= 2[p_{AB}(1 - p_A - p_B + p_{AB})$ $+(p_A - p_{AB})(p_B - p_{AB})]$	$p_{Aabb} = 2p_{Ab}p_{ab}$ $= 2(p_A - p_{AB})(1 - p_A - p_B$ $+ p_{AB})$	$2p_A(1-p_A)$
Genotype aa	$p_{aaBB} = p_{aB}^2$	$p_{aaBb} = 2p_{aB}p_{ab}$	$p_{aabb} = p_{ab}^2$	$(1-p_A)^2$

	$= (p_B - p_{AB})^2$	$= 2(p_B - p_{AB})(1 - p_A - p_B + p_{AB})$	$= (1 - p_A - p_B + p_{AB})^2$	
Genotype Frequency	p_B^2	$2p_B(1 - p_B)$	$(1 - p_B)^2$	1

Genotype frequency of variant A conditioning on variant B

Genotype	AA	Aa	aa
BB	$P(\, AA \mid BB\,) = \dfrac{p_{AABB}}{p_{BB}}$ $= \dfrac{p_{AB}^2}{p_B^2}$	$P(\, Aa \mid BB\,) = \dfrac{p_{AaBB}}{p_{BB}}$ $= \dfrac{2p_{AB}(p_B - p_{AB})}{p_B^2}$	$P(\, aa \mid BB\,) = \dfrac{p_{aaBB}}{p_{BB}}$ $= \dfrac{(p_B - p_{AB})^2}{p_B^2}$
Bb	$P(\, AA \mid Bb\,) = \dfrac{p_{AABb}}{p_{Bb}}$ $= \dfrac{p_{AB}(p_A - p_{AB})}{p_B(1 - p_B)}$	$P(\, Aa \mid Bb\,) = \dfrac{p_{AaBb}}{p_{Bb}}$ $= \dfrac{p_{AB}(1 - p_A - p_B + p_{AB}) + (p_A - p_{AB})(p_B - p_{AB})}{p_B(1 - p_B)}$	$P(\, aa \mid Bb\,) = \dfrac{p_{aaBb}}{p_{Bb}}$ $= \dfrac{(p_B - p_{AB})(1 - p_A - p_B + p_{AB})}{p_B(1 - p_B)}$
bb	$P(\, AA \mid bb\,) = \dfrac{p_{AAbb}}{p_{bb}}$ $= \dfrac{(p_A - p_{AB})^2}{(1 - p_B)^2}$	$P(\, Aa \mid bb\,) = \dfrac{p_{Aabb}}{p_{bb}}$ $= \dfrac{2(p_A - p_{AB})(1 - p_A - p_B + p_{AB})}{(1 - p_B)^2}$	$P(\, aa \mid bb\,) = \dfrac{p_{aabb}}{p_{bb}}$ $= \dfrac{(1 - p_A - p_B + p_{AB})^2}{(1 - p_B)^2}$

Supplementary Note 3-3 The expected phenotypic mean and variance for variant A and B

The expected phenotypic mean and variance for variant A (causal) can be found in the table below.

Genotype	Code (x_a)	$E(y\|x_a)$	$Var(y\|x_a)$	$E(y^2\|x_a)$

AA	0	μ	σ^2	$\sigma^2+\mu^2$
Aa	1	$\mu+b_c$	σ^2	$\sigma^2+(\mu+b_c)^2$
aa	2	$\mu+2b_c$	σ^2	$\sigma^2+(\mu+2b_c)^2$

The expected phenotypic mean and variance given a genotype of variant B (marker) can be found in the table below.

Genotype	Code (x_b)	$E(y\|x_b)$
BB	0	$\mu+\frac{2b_c(p_B-p_{AB})}{p_B}$
Bb	1	$\mu+\frac{b_c(2p_B-p_{AB}-2p_B^2+2p_Bp_{AB}-p_Ap_B)}{p_B(1-p_B)}$ $=\mu+\frac{b_c[(p_B-p_{AB})(1-p_B)+(1-p_A-p_B+p_{AB})p_B]}{p_B(1-p_B)}$
bb	2	$\mu+\frac{2b_c(1-p_A-p_B+p_{AB})}{1-p_B}$

$Var(y\|x_b)$
$\sigma^2+\frac{2{b_c}^2(p_B-p_{AB})p_{AB}}{p_B^2}$
$\sigma^2+\frac{{b_c}^2(p_Bp_{AB}-p_{AB}^2+2p_Bp_{AB}^2-3p_B^2p_{AB}+p_Ap_B^2-2p_B^2p_{AB}^2+2p_Ap_B^2p_{AB}+2p_B^3p_{AB}-p_Ap_B^3-p_A^2p_B^2)}{p_B^2(1-p_B)^2}$ $=\sigma^2+\frac{{b_c}^2[(p_B-p_{AB})p_{AB}(1-p_B)^2+(1-p_A-p_B+p_{AB})(p_A-p_{AB})p_B^2]}{p_B^2(1-p_B)^2}$

$$\sigma^2 + \frac{2{b_c}^2(1 - p_A - p_B + p_{AB})(p_A - p_{AB})}{(1 - p_B)^2}$$

We therefore can observe an additive effect on both mean (b_m) and variance (β_m) at the marker (variant B):

$$y \sim (\mu_m + b_m x_b, {\sigma^2}_m + \beta_m x_b)$$

where

$$\mu_m = \mu + \frac{2b_c(p_B - p_{AB})}{p_B}$$

$$b_m = \frac{b_c(p_{AB} - p_A p_B)}{p_B(1 - P_B)}$$

$${\sigma^2}_m = \sigma^2 + \frac{2{b_c}^2(p_B - p_{AB})p_{AB}}{p_B^2}$$

$$\beta_m = \frac{b_c^2[(1 - 2p_B)p_{AB}^2 + (2p_A p_B + p_B - 1)p_B p_{AB} + (1 - p_A - p_B)p_A p_B^2]}{p_B^2(1 - p_B)^2}$$

Supplementary Note 3-4 QTL test-statistics at the marker variant B

Assuming phenotypic variance of 1 (i.e., var(*y*) = 1), the variance explained by the marker variant (q_m^2) and the non-centrality parameter (NCP) of a chi-squared test for QTL effect at the marker can be written as

$$q_m^2 = 2p_B(1 - p_B)b_m^2 = 2p_B(1 - p_B)\frac{b_c^2(p_{AB} - p_A p_B)^2}{p_B^2(1 - p_B)^2} = 2p_A(1 - p_A)b_c^2\frac{(p_{AB} - p_A p_B)^2}{p_A(1 - p_A)p_B(1 - p_B)} = q_c^2 r^2$$

$$\text{NCP} = \frac{nq_m^2}{1 - q_m^2} = \frac{nq_c^2 r^2}{1 - q_c^2 r^2}$$

where n is the sample size, q_c^2 is the variance explained by the causal variant, and r^2 is the LD between the causal and the marker variants. This derivation is consistent with that in previous studies[187,188].

Supplementary Note 3-5 The vQTL test statistic at the marker variant B

Under normality assumption, the distribution of the phenotype with respect to the marker variant can be written as:

$$y \sim N(\mu_m + b_m x_b, {\sigma^2}_m + \beta_m x_b)$$

We then have

$$y - E(y|x_b) \sim N(0, {\sigma^2}_m + \beta_m x_b)$$

, and $z = |y - \widetilde{y}\,|$

$$z = |y - \widetilde{y}\,| = |y - E(y|x_b)| \sim \text{ Folded Normal Distribution}(\sqrt{\frac{2}{\pi}({\sigma^2}_m + \beta_m x_b)}, (1 - \frac{2}{\pi})({\sigma^2}_m + \beta_m x_b))$$

Genotype	Code (x_b)	E(z|x_b)	var(z|x_b)	$E(z^2\|x_b)$
BB	0	$\sqrt{\frac{2}{\pi}{\sigma^2}_m}$	$(1 - \frac{2}{\pi}){\sigma^2}_m$	${\sigma^2}_m$
Bb	1	$\sqrt{\frac{2}{\pi}({\sigma^2}_m + \beta_m)}$	$(1 - \frac{2}{\pi})({\sigma^2}_m + \beta_m)$	${\sigma^2}_m + \beta_m$

bb	2	$\sqrt{\frac{2}{\pi}(\sigma^2{}_m + 2\beta_m)}$	$(1 - \frac{2}{\pi})(\sigma^2{}_m + 2\beta_m)$	$\sigma^2{}_m + 2\beta_m$

$$E(z) = E(z|x_b = 0)P(x_b = 0) + E(z|x_b = 1)P(x_b = 1) + E(z|x_b = 2)P(x_b = 2)$$
$$= \sqrt{\frac{2}{\pi}\sigma^2{}_m}{p_B}^2 + \sqrt{\frac{2}{\pi}(\sigma^2{}_m + \beta_m)}2p_B(1 - p_B) + \sqrt{\frac{2}{\pi}(\sigma^2{}_m + 2\beta_m)}(1 - p_B)^2$$

$$E(z^2) = E(z^2|x_b = 0)P(x_b = 0) + E(z^2|x_b = 1)P(x_b = 1) + E(z^2|x_b = 2)P(x_b = 2)$$
$$= \sigma^2{}_m{p_B}^2 + (\sigma^2{}_m + \beta_m)2p_B(1 - p_B) + (\sigma^2{}_m + 2\beta_m)(1 - p_B)^2$$
$$= \sigma^2{}_m + 2(1 - p_B)\beta_m$$

$$var(z) = E(z^2) - [E(z)]^2 = \sigma^2{}_m + 2(1 - p_B)\beta_m - [E(z)]^2$$

The Levene's test is essentially one-way ANOVA test on the variable *z* (see the Methods section). We therefore have

$$E(SST) = E[\textstyle\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - z_{..})^2] = Var(z)n = (\sigma^2{}_m + 2(1 - p_B)\beta_m - [E(z)]^2)n;$$

$$E(SSE) = E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - z_{i.})^2\right]$$

$$= (1-\frac{2}{\pi})\sigma^2{}_m n p_B{}^2 + (1-\frac{2}{\pi})(\sigma^2{}_m + \beta_m) n 2 p_B(1-p_B) + (1-\frac{2}{\pi})(\sigma^2{}_m + 2\beta_m)(1-p_B)^2$$

$$= (1-\frac{2}{\pi})(\sigma^2{}_m + 2(1-p_B)\beta_m)n;$$

$$E(SSR) = E(SST - SSE) = [\frac{2}{\pi}(\sigma^2{}_m + 2(1-p_B)\beta_m) - [E(z)]^2]n;$$

$$F_{\text{Levene}} = \frac{(n-3)E(SSR)}{(3-1)E(SSE)} \approx \frac{n}{2}\frac{E(SSR)}{E(SSE)} = \frac{n}{2}\frac{\frac{2}{\pi}(\sigma^2{}_m + 2(1-p_B)\beta_m) - [E(z)]^2}{(1-\frac{2}{\pi})(\sigma^2{}_m + 2(1-p_B)\beta_m)}$$

$$= \frac{n}{\pi - 2}(1 - \frac{[\sqrt{\sigma^2{}_m} p_B{}^2 + \sqrt{\sigma^2{}_m + \beta_m} 2p_B(1-p_B) + \sqrt{\sigma^2{}_m + 2\beta_m}(1-p_B)^2]^2}{\sigma^2{}_m + 2(1-p_B)\beta_m})$$

where F_{Levene} is the Levene's *F*-statistic; *SST*, *SSR* and *SSR* are the total sum of squares, regression sum of squares and error sum of squares, respectively, as defined in an ANOVA analysis.

Given that var(y) = 1, we can replace b_c^2 with $\frac{q_c^2}{2p_A(1-p_A)}$, and σ^2 with $1 - q_c^2$:

$$\beta_m = \frac{q_c^2[(1-2p_B)p_{AB}^2 + (2p_A p_B + p_B - 1)p_B p_{AB} + (1 - p_A - p_B)p_A p_B^2]}{2p_A(1-p_A)p_B^2(1-p_B)^2}$$

$$\sigma^2{}_m = 1 - q_c^2 + \frac{(p_B - p_{AB})p_{AB}}{p_A(1-p_A)p_B^2} q_c^2$$

$$F_{\text{Levene}} = \frac{n}{\pi-2}(1 - \frac{[\sqrt{\sigma^2{}_m} p_B{}^2 + \sqrt{\sigma^2{}_m + \beta_m} 2p_B(1-p_B) + \sqrt{\sigma^2{}_m + 2\beta_m}(1-p_B)^2]^2}{\sigma^2{}_m + 2(1-p_B)\beta_m})$$

4

Chapter 4: Integrating genetic and environmental information to improve phenotype prediction for body mass index

This work is unpublished.

Integrating genetic and environmental information to improve phenotype prediction for body mass index

4.1 Abstract

Predicting human complex traits or diseases can be achieved by using genetic information (genetic risk score, GRS) or environmental information (environmental risk score, ERS). However, it has been less studied to integrate both genetic and environmental information (genetic and environmental risk score, GERS). Here we took body mass index (BMI) as a model trait, generated the GRS based on 1,317,930 HapMap3 SNPs, generated ERS based on eight environmental factors, and explored different methods to construct the GERS using 348,501 unrelated European individuals of UK Biobank (UKB). We found GERS could improve the prediction accuracy in comparison with GRS only, with GERS based on a multiple linear regression (MLR) performing best (i.e. R^2 increased from 13.1% for GRS to 18.1% for GERS_mlr). In addition, integrating GRS with the genetic components of environmental factors (multiple GRS, MGRS) could not improve the prediction accuracy if all GRSs were based on the same training dataset using both real data and simulation study. Our results indicate the value of integrating both genetic and environmental information for predicting BMI and other complex traits or diseases.

4.2 Introduction

The prediction of human complex traits or diseases is key for personalized prevention, intervention, and treatment[126,192-195]. One notable example is Framingham risk score/pooled cohort equation for cardiovascular disease (CVD)[196,197]. The latest American College of Cardiology/American Heart Association (ACC/AHA) recommends lipid-lowering treatment for individuals with risk > 7.5% in primary care[198]. Most risk scores/models are based on multiple risk factors/predictors, which could be demographic factors (e.g., age or sex), environmental factors (e.g. biomarkers or smoking), or genetic factors (e.g. family history or monogenic mutations). The current risk models, including CVD risk model, are usually imprecise estimates, so there are continuous efforts to search for new factors to be included and to improve the accuracy of risk prediction[199,200].

Recently, polygenic/genetic risk score (PRS/GRS) is promising, which accumulates the effects of many genetic variants across the whole genome. The prediction accuracy of GRS has achieved comparable prediction performance with other risk factors[112,115,124,125] and is approaching its theoretical maximum[114,201,202] due to the increasing sample size of GWAS study[43,46] and the development of more sophisticated statistical tools[2,113,118]. The genetic information, in contrast with environmental information, is determined at birth, so it is free of reverse causality and less vulnerable for confounding (except for confounders such as relatedness and population stratification[203] which can be well-controlled[31]). GRS is also cost-effective, because one genotype array of < 100 dollars can be used to predict many traits and diseases.

Early studies have tried to combine GRS with established risk models or risk factors[204] (e.g. CVD[205-208], breast cancer[205,209-212], and others[212]), while the added values of GRS are still under debate. For example, Mars et al.[205] and Mosley et al.[206] found non-significant improvement (assessed by net reclassification improvement (NCI)) of GRS adding to clinical risk scores using FINRISK, ARIC, and MESA cohorts to predict CVD, Elliott et al.[207] found significant but modest improvement using UKB data, and Riveros-Mckay et al.[208] found the significant and highest improvement using UKB data. In addition, more broad environmental factors (sometimes called exposomic factors)[213,214] are being examined for their potential to be included in the risk prediction model.

Here, we took BMI as a model trait and built GRS based on genome-wide genetic variants, ERS based on eight environmental factors, and GERS by combining GRS and ERS in different ways in the UKB. We evaluated whether and to what extent GERS could improve the prediction accuracy.

4.3 Results

Method overview

For an individual with m genetic variants ($G_1 \dots G_m$) and k environmental factors ($E_1 \dots E_k$), a simple form of GRS is the weighted sum of the genotype values:

$$GRS = w_1G_1 + w_2G_2 + \cdots + w_mG_m$$

, and then a GERS can be calculated by

$$GERS = u_0GRS + u_1E_1 + u_2E_2 + \cdots + u_kE_k$$

, where weights $w_1 \dots w_k$ are generated by SBayesR method[2] in this study, and weights $u_0 \dots u_k$ are generated by different methods under investigation: 1) $u_0 = 1$ and $u_1 \dots u_k$

using phenotypic correlation (r_p) (GERS_rp); 2) u_0 ... u_k estimated by a MLR (GERS_mlr); 3) $w_0 = 1$ and w_1 ... w_k using causal effect sizes of environmental factors on the phenotype inferred by a mendelian randomization analysis (GERS_mr) (Methods).

We used BMI as our target phenotype, 1,317,930 HapMap3 SNPs genotype data, and eight environmental factors (i.e. coffee intake (Coffee), educational attainment (EA), nap during day (Nap), salt added to food (Salt), SB, sleep duration (Sleep), smoking status (Smoking), and tea intake(Tea)) measured in 348,501 unrelated European individuals in UKB (Table 4-1 and Figure 4-1). We used SBayesR[2] method to construct the GRS and GSMR[215] method in the mendelian randomization analysis. We tried different strategies to split the UKB dataset into training, validation and testing datasets: 1) randomly (rrr); 2) young, old, and old individuals (yoo); 3) young, young, and old individuals (yyo); 4) first, first, and second BMI measurements accessible individuals (ffs).

Table 4-1 Phenotype and eight environmental factors in UKB

Name	Description	Sample size	Data field(s)	Unit or coding rule
BMI	Body mass index (BMI)	346,989	21001	kg/m^2
Coffee	Coffee intake	347,124	1498	cups/day
EA[a]	Education attainment	344,890	6138	EduYears
Nap	Nap during day	348,057	1190	never/rarely (1); sometimes (2); usually (3)
Salt	Salt added to food	348,199	1478	never/rarely (1); sometimes (2); usually (3); always (4)
SB[b]	Sedentary behaviour	339,330	1090, 1080, 1070	hours/day
Sleep	Sleep duration	346,161	1160	hours/day
Smoking[c]	Smoking status	346,407	1239, 1249	never (0); ever (1)
Tea	Tea intake	346,428	1488	cups/day

[a] based on the rules in Lee et al. 2018 paper[44] (Table S17).

[b] and [c] details of the definition can be found in Wang et al. 2019 paper[1] (Note S5). Briefly, SB is the total time for driving, computer using and TV watching,

and smoking is the status of smoking (never or ever).

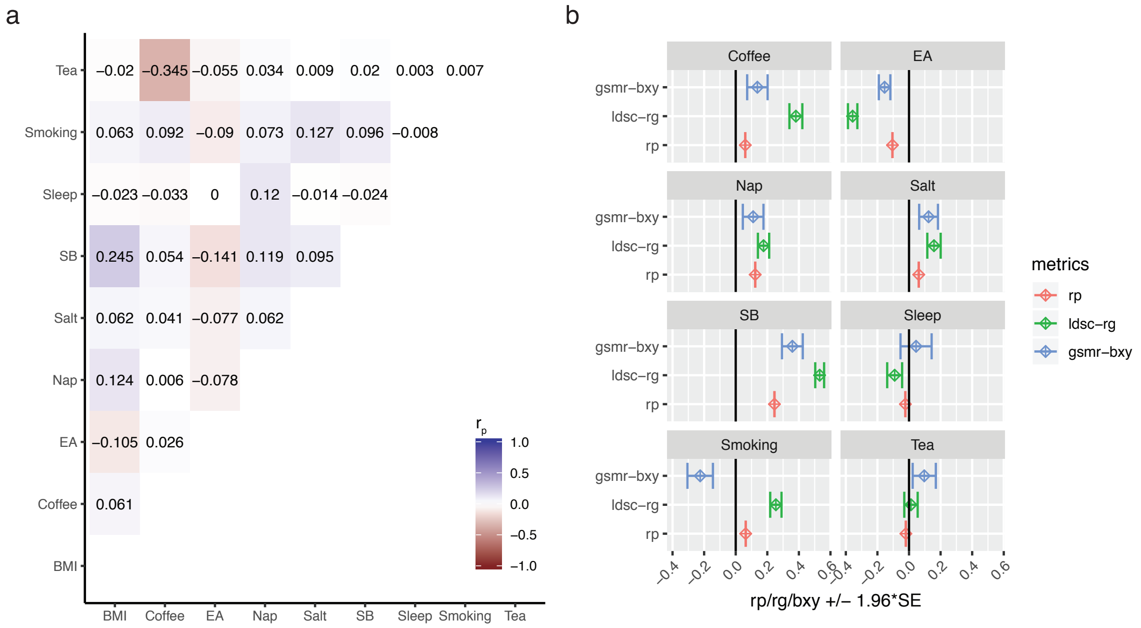


Figure 4-1 Phenotypic correlation (r_p), genetic correlation (r_g), and causal effect size (b_{xy}) for BMI and eight environmental factors.

a) The phenotypic correlation between each pair of BMI and eight environmental factors were calculated in the training dataset of "rrr" strategy. b) The phenotypic correlation with BMI, genetic correlation estimated by LDSR[61] method with BMI, and causal effect size inferred by GSMR[215] method on BMI for eight environmental factors in the training dataset of "rrr" strategy were plotted with point estimates as diamonds and standard error (SE) multiplied by 1.96 as lines (only r_g and b_{xy}).

GRS and GERSs

We found all GERSs built by three different methods performed better than the GRS across all four data splitting strategies (Figure 4-2). Among three GERSs, the GERS_mlr, which accounted both the correlation between GRS and environmental factors and the correlation within different environmental factors, performed best, and improved the prediction accuracy of GRS from 0.131 to 0.181 (38.2%), 0.117 to 0.169 (44.4%), 0.119 to 0.168 (41.2%), and 0.116 to 0.171 (47.4%), in "rrr", "yoo", "yyo", and "ffs" data splitting strategy, respectively. And the GERS_mr performed worst, even for the second BMI measurements ("ffs" strategy). For different data splitting strategies, the prediction accuracies of GRS and GERSs using the randomly splitting strategy ("rrr" strategy) were better than those using other strategies, which was consistent with previous study[216].

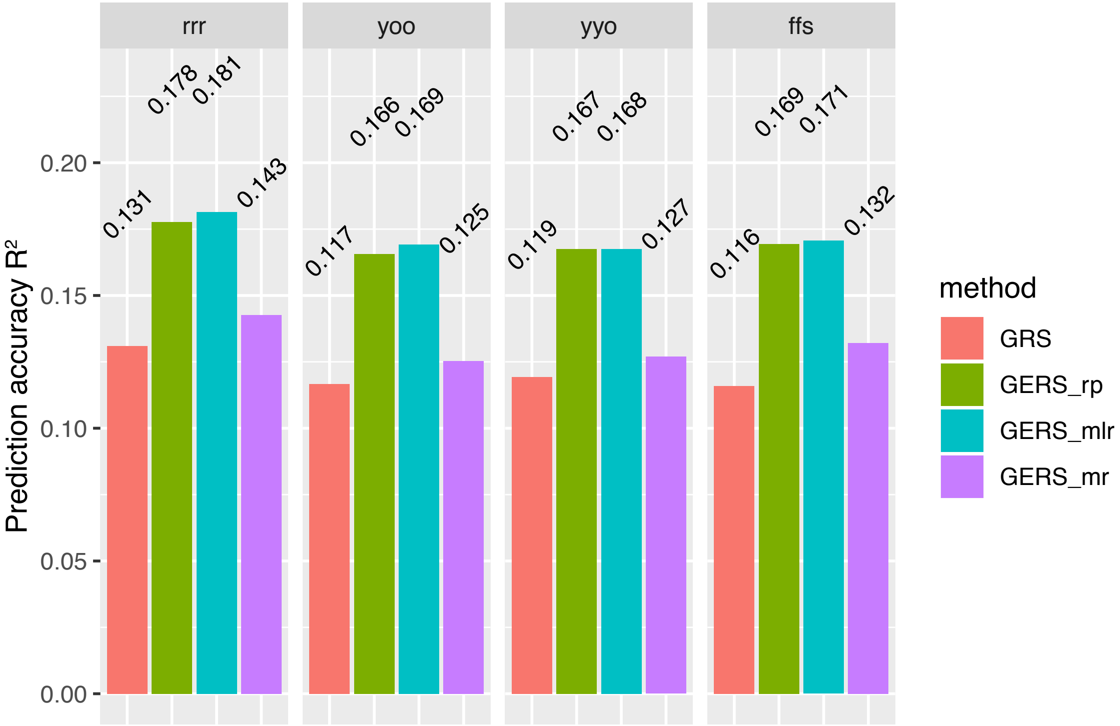


Figure 4-2 Prediction accuracy of GRS and GERSs with eight environmental factors in different data splitting strategies.

The prediction accuracy was evaluated using R^2 in the testing dataset for GRS and GERS_rp, GERS_mlr, and GERS_mr with eight environmental factors in different data splitting strategies (i.e., rrr, yoo, yyo, and ffs). See more details in Methods section.

We also assessed the performance of GERSs with each one of these eight environmental factors. We found that the prediction accuracy of GERS with SB was highest, followed by Nap and GERS with coffee, sleep and tea almost could not improve prediction accuracy (Figure 4-3).

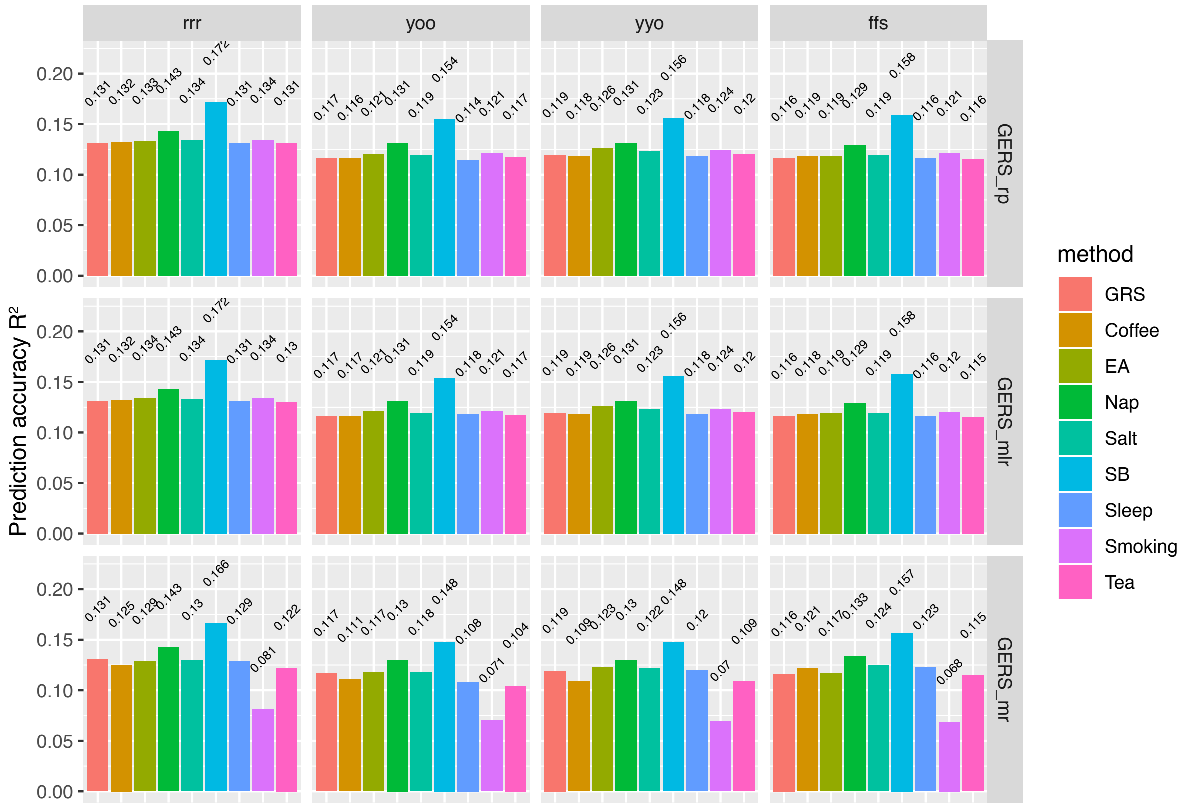


Figure 4-3 Prediction accuracy of GRS and GERSs with each one of eight environmental factors in different data splitting strategies.

The prediction accuracy was evaluated using R^2 in the testing dataset for GRS and GERS_rp, GERS_mlr, and GERS_mr with each of eight environmental factors in different data splitting strategies (i.e., rrr, yoo, yyo, and ffs). See more details in Methods section.

MGRS

Instead of requiring environmental information measured for GERS, there are other studies utilizing only the genetic component of environmental information, including multi-trait prediction[217-219], multiple polygenic risk scores (MPS)[220], and metaGRS[125], which we called MGRS here. We constructed a MGRS using a MLR (MGRS_mlr):

$$MGRS_{mlr} = v_0\text{GRS} + v_1 GRS_{E1} + \cdots + v_k GRS_{Ek}$$

. We found that the MGRS combining the GRSs of BMI and eight environmental factors cannot improve the prediction accuracy in comparison with BMI GRS only (Figure 4-4). We explained below by simulation study it was because all GRSs were based on the same training dataset.

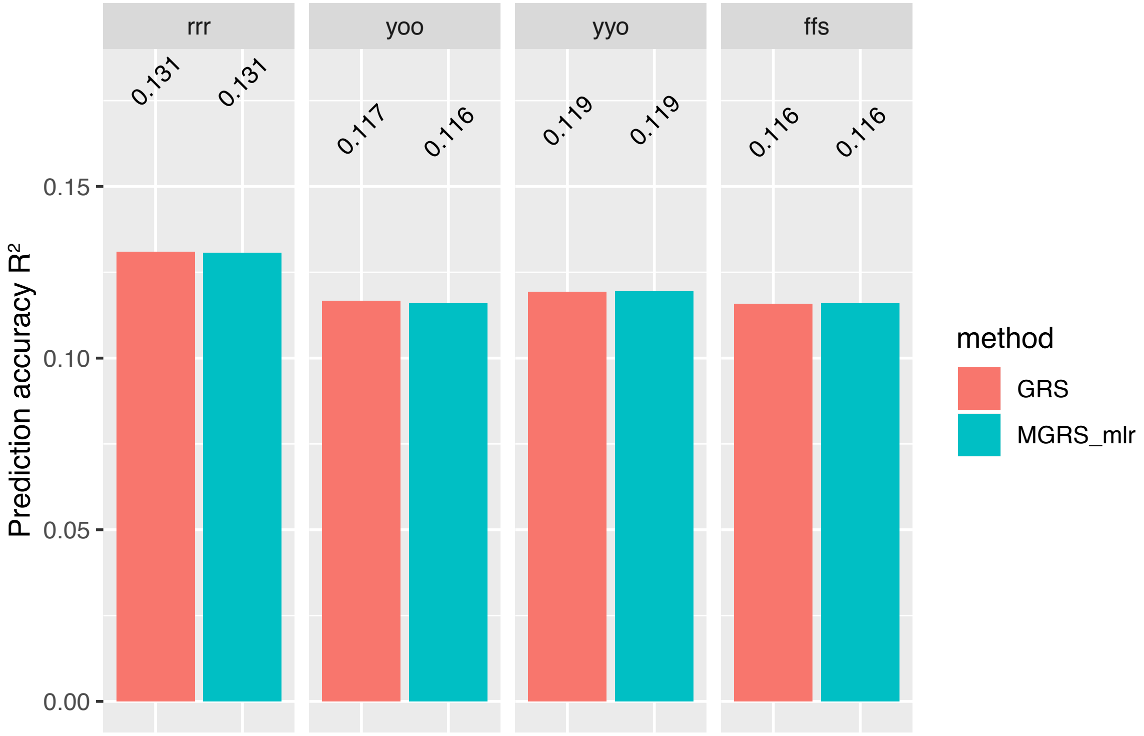


Figure 4-4 Prediction accuracy of GRS of BMI and MGRS of BMI and eight environmental factors in comparison with GRS in different data splitting strategies.

We simulated two quantitative traits (trait 1 and 2) with varying parameters: variance explained by the genetic component for trait 1 (h_1), variance explained by the genetic component for trait 2 (h_2), the training sample size for trait 1 (n_1), the training sample size for trait 2 (n_2), and the genetic correlation between trait 1 and trait 2 (r_g). We found that if trait 1 and trait 2 were based on one same dataset with sample size 500 (one500), 5,000 (one5k), and 50,000 (one50k), the prediction accuracy of MGRS_mlr was similar to that of GRS (the first 3 rows in Figure 4-5). However, if trait 1 and trait 2 were based on two different datasets with sample size 500 for trait 1 and sample size 500 (500+500), 5,000 (500+5k) and 50,000 (500+50k) for trait 2, the prediction accuracy of MGRS_mlr was improved in comparison to GRS and the level of improvement increased with the decrease of h_1 and n_1, and the increase of h_2, n_2 and r_g, which was consistent with the previous study[217].

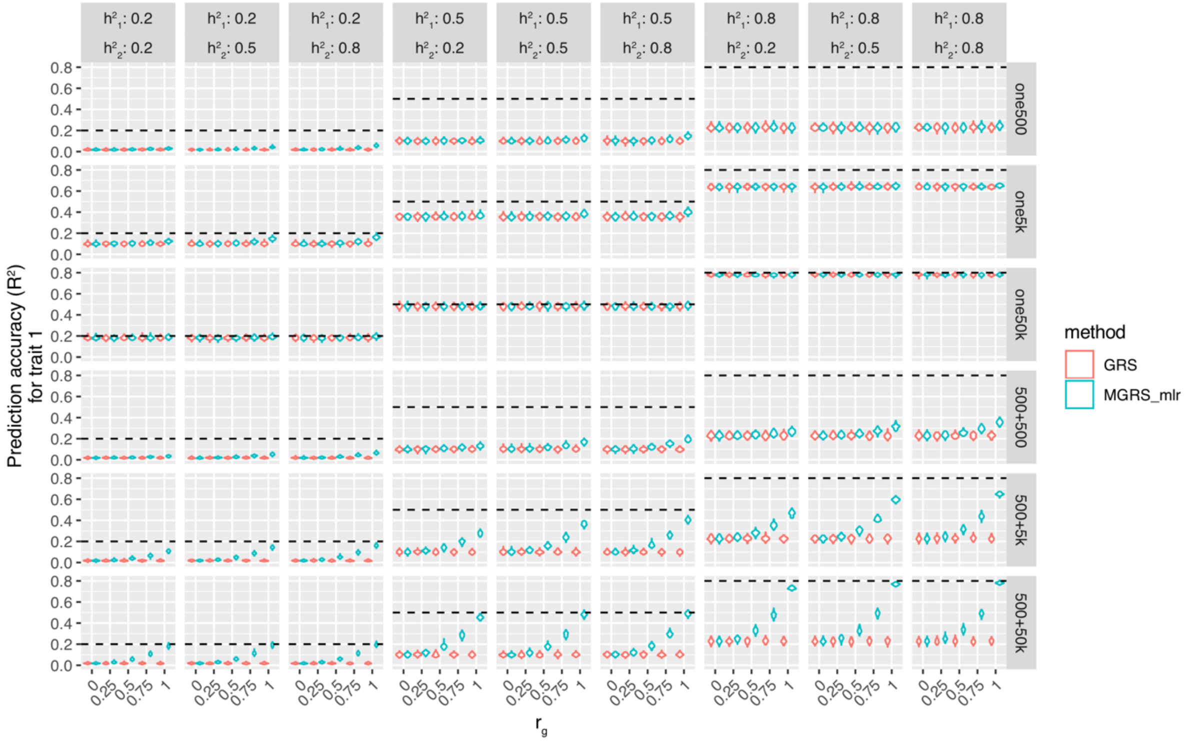


Figure 4-5 Simulation study evaluating the prediction accuracy for trait 1 of MGRS in comparison to GRS.

The following parameters were used with different values, including h_1^2 with 0.2, 0.5, 0.8, h_2^2 with 0.2, 0.5, 0.8, and r_g with 0, 0.25, 0.5, 0.75, 1, in one training dataset (i.e., one500, one5k, one50k) or two training datasets (i.e., 500+500, 500+5k, 500+50k) with different sample sizes. Each combination was replicated 100 times and prediction accuracies were presented in violin plots above. The black dash line represents the variance explained by the genetic component for trait 1 (h_1^2).

4.4 Discussion

We explored different methods to combine the genetic and environmental factors to improve the prediction accuracy for human complex traits. We took BMI as our model trait and built GRS, ERS, and GERS using different methods in UKB dataset. We found an improvement for GERS in comparison with GRS and the multiple linear regression method shown the highest improvement. We also performed real data and simulation study to demonstrate that an MGRS combined with different GRS based on the same dataset could not improve the prediction accuracy.

We tested three ways to construct the GERS. The GERS_rp and GERS_mlr were based on the simple and multiple linear regression, respectively, which were simple and classic textbook linear regression models. We also included a GERS_mr based on the MR method, as we hypothesized the causal effects estimated by MR are more robust to confounding and reverse causality effects than the regression coefficients estimated by linear regression models, so that might provide higher prediction accuracy. However, we observed a lower prediction accuracy for GERS based on MR than GERS based on linear regression models. It could be because the confounding and reverse causality effects also existed in the test dataset, leading to inflated prediction accuracy for MLR, although we tried different data splitting strategies. Another reason could be the immature methodology of MR to accurately estimate the causal effect size of environmental factors on the trait and the potential non-linearity of the causal effect.

There are a few limitations in this study. Firstly, we only included eight environmental factors to explore the potential of GERS, which were clearly not comprehensive, and more lifestyle factors (e.g. diet) could affect BMI. Secondly, we only considered additive effects and did not consider interaction effects. The genotype-by-genotype interaction was expected to be small, but the influence of environmental-by-environmental interaction and genotype-by-environmental interaction needs further study. Finally, more methods to build GERS (e.g. multivariate MR methods) and MGRS (e.g. MTAG[218] or SMTpred[217]) need to be tested.

4.5 Methods

UKB and data splitting

UKB is a large-scale data resource consisting of genotype and phenotype information for around 500,000 individuals aged 40-70 years old in United Kingdom[46]. We used the genotype of 1,317,930 HapMap 3 SNPs, BMI as our target phenotype, and eight environmental factors (Table 4-1) for 348,501 unrelated European individuals in UKB. We excluded the outlier values 5 standard deviation from the mean for BMI and each of eight environmental factors. More details about per-individual and per-SNP quality control can be found in Chapter 2.

The UKB dataset was then split into training, validation, and testing datasets based on different strategies (Table 4-2): 1) “rrr”: 328,501, 10,000 and 10,000 individuals were randomly sampled as training, testing and validation datasets, respectively. 2) "yoo": the youngest 327,893 individuals were selected as training dataset. For the remaining 20,000 individuals, we randomly sampled 10,000 individuals as validation dataset and 10,000 individuals as testing dataset. 3) "yyo": the youngest 337,893 individuals were randomly split to two datasets (327,893 individuals as training dataset and 10,000 individuals as testing dataset). The remaining 10,000 individuals were classified as testing dataset. 608 individuals without age information available were excluded in strategy "yoo" and "yyo". 4) "ffs": the 333,311 individuals with only first/initial BMI measurement available were randomly split to two datasets (323,311 individuals as training dataset and 10,000 individuals as validation dataset). The remaining 15190 individuals with second BMI measurement available were classified as testing dataset.

Table 4-2 The sample size of training, validation and testing dataset in different data splitting strategies.

Data splitting strategy	Training (*n*)	Validation (*n*)	Testing (*n*)
rrr	Random (328501)	Random (10000)	Random (10000)
yoo	Young (327893)	Old (10000)	Old (10000)
yyo	Young (327893)	Young (10000)	Old (10000)
ffs	First BMI (323311)	First BMI (10000)	Second BMI (15190)

GRS and GERSs

The GWAS analysis was conducted using PLINK2[174] (--assoc option) in the training dataset for BMI and eight environmental factors, whose values were pre-adjusted with age, sex, PC 1-10, inverse-normal transformed (BMI only), and standardized to *z* scores with mean zero and variance 1. The marginal effect sizes were re-estimated to joint effect sizes using a summary data-based Bayesian multiple regression method SBayesR[2] implemented in software GCTB (version 2), with parameters "--sbayes R --maf 0.01 --chain-length 21000 --burn-in 1000 --estimate-ps", and a sparse LD matrix reference (a subset of LD correlation values set to zero) generated using a randomly sampled 50,000 individuals in UKB. Then the GRS was generated using PLINK2 ("--score" option) in validation and testing dataset.

We explored different methods to combine GRS and eight environmental factors ($E_1 \dots E_8$) to construct the GERS. Firstly, we use the phenotypic correlations ($r_{p(1)} \dots r_{p(8)}$) between BMI and environmental factors estimated in the training dataset as weights (GERS_rp):

$$GERS_{rp} = GRS + \sum_{i=1}^{8} r_{p(i)}\, E_i$$

. Secondly, we use the weights ($u_0 \dots u_8$) estimated by MLR between GRS and environmental factors in the validation dataset (GERS_mlr):

$$GERS_{mlr} = u_0 GRS + \sum_{i=1}^{8} u_i\, E_i$$

. Thirdly, we use the causal effect sizes ($b_{xy(1)} \dots b_{xy(8)}$) of environmental factors on BMI inferred by a Mendelian randomization analysis (GSMR method) (GERS_mr):

$$GERS_{mr} = GRS + \sum_{i=1}^{8} b_{xy(i)}\, E_i$$

. We used the GSMR method implemented in GCTA[53] with GWAS summary data of environmental factors generated in the training dataset, GWAS summary data of BMI from the GIANT consortium[43], and LD reference based on randomly selected 10,000 individuals in the UKB (GSMR parameters: --gsmr2-beta --heidi-thresh 0.01 0.01 and --gsmr-snp-min 10). The values of BMI and eight environmental factors were also standardized into *z* score with mean 0 and variance 1 in the validation and testing dataset separately. The prediction accuracy was estimated using R^2 in the testing dataset.

MGRS methods

We constructed a MGRS to combine GRSs of BMI and eight environmental factors using the GWAS summary data calculated in the training dataset with weights estimated by a MLR in the validation dataset:

$$MGRS_{mlr} = v_0 \mathrm{GRS} + v_1 GRS_{E1} + \cdots + v_k GRS_{Ek}$$

. For the simulation study, we simulated two traits (trait 1 and trait 2),

$$y_1 = \sum_{i}^{m} \beta_{1(i)}\, w_{(i)} + e_1; y_2 = \sum_{i}^{m} \beta_{2(i)}\, w_{(i)} + e_2$$

, where y_1 and y_2 are trait 1 and trait 2; $w_{(i)}$ is the standardized genotype value for *i*-th out of *m* (=1000) SNPs, i.e., $w_{(i)} = (x_{(i)} - 2f_{(i)})/\sqrt{2f_{(i)}(1 - f_{(i)})}$, with $x_{(i)}$ being the genotype indicator variable coded as 0, 1, or 2 generated from binomial(2,$f_{(i)}$), and $f_{(i)}$ being the minor allele frequency (MAF) generated from uniform(0.1,0.5); β_1 and β_2 are effect sizes of i-th SNP on trait 1 and trait 2 generated from a binormal distribution $N(0, \begin{bmatrix} \frac{h_1^2}{m} & \frac{r_g\sqrt{h_1^2 h_2^2}}{m} \\ \frac{r_g\sqrt{h_1^2 h_2^2}}{m} & \frac{h_2^2}{m} \end{bmatrix})$, with h_1^2 being the variance explained for trait 1, h_2^2 being the variance explained for trait 2, and r_g being the genetic correlation between trait 1 and trait 2; e_1 and e_2 are the environmental terms for trait 1 and trait 2, generated by a binormal distribution $N(0, \begin{bmatrix} 1-h_1^2 & 0 \\ 0 & 1-h_2^2 \end{bmatrix})$. We varied the following parameters: h_1^2 = 0.2, 0.5, 0.8; h_2^2 = 0.2,0.5, 0.8; r_g = 0, 0.25, 0.5, 0.75, 1. We simulated one training dataset with sample size 500 (one500), 5,000 (one5k), 50,000 (one50k), or two independent training datasets with sample size 500 for trait 1 and sample size 500 (500+500), 5000 (500+5k), 50000 (500+50k) for trait 2, and 5000 individuals for validation dataset, and 5000 individuals for testing dataset. Each situation was replicated for 100 times.

4.6 URLs

PLINK2, http://www.cog-genomics.org/plink2

GCTB-SBayesR, https://cnsgenomics.com/software/gctb/#Overview

GCTA-GSMR, https://cnsgenomics.com/software/gcta/#GSMR

The UKB data, http://www.ukbiobank.ac.uk/

GIANT BMI summary data, https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

5

Chapter 5: Summary and discussion

The overall aim of this thesis is to better understand the genetic and environmental effects on human complex traits. Based on Fisher's quantitative genetics model more than 100 years ago, the phenotypic value or the phenotypic variance can be partitioned into genetic and environmental components, and also the potential GEI component, where the genetic component can be further partitioned into additive, dominance, and epistatic genetic effects. Three research projects have been conducted to study the GEI component (Chapter 2), inflation in test-statistics for GEI and epistasis (Chapter 3), and phenotype prediction using both genetic and environmental factors (Chapter 4). The main findings of these three research chapters will be summarized in this chapter and future directions of human complex traits study will be discussed.

5.1 GEI effects inferred from vQTL analysis

A range of methods can be used to associate genetic variants with phenotype variability. Firstly, we used simulations to evaluate the FPR and power of four methods, which were Barlett's test, Levene's test, FK test, and DGLM, and found that Levene's test had a good control of FPR and was robust to the distribution of the phenotype. In addition, we found all non-linear transformations (including RINT, logarithm, square, and cube transformations) could inflate Levene's test when there was a mean QTL effect. So Levene's test without transformation was chosen for real-data analysis.

We applied this genome-wide vQTL analysis to 13 quantitative traits in the UKB dataset and found 75 significant vQTLs for 9 traits, which included 64 with significant mean QTL effects. A further direct GEI analysis with covariates (i.e. age and sex) and environmental factors (i.e. PA, SB, and smoking status) demonstrated that GEI effects were enriched in vQTLs in comparison to randomly selected QTLs. The identified GEI effects included some examples consistent with the previous published results, including *CHRNA5-A3-B4* locus interacting with smoking status on FFR, *WNT16-CPED1* locus interacting with age on BMD, and *FTO* locus interacting with PA on obesity-related traits.

5.2 Inflation level in vQTL and epistasis test

One challenge of statistical tests for interaction effects is the inflated FPR caused by imperfect tagging of the causal genetic variant by the marker variants. We studied this phantom signal of statistical tests for two interaction effects, vQTL test and epistasis test. For phantom vQTL, we referred to a vQTL signal caused by a QTL effect via LD. We derived the expected phantom vQTL F-statistic and demonstrated the observed vQTL in Chapter 2 was unlikely to be phantom vQTL. In addition, we did not observe discordant vQTL test statistics before and after fitting a nearby QTL SNP. Overall, this study quantified the inflation level of phantom vQTL and found no evidence for the vQTL we detected in Chapter 2. We also performed a direct epistasis test between vQTL and genome-wide SNPs on the relevant trait and found no genome-wide significant epistatic interactions with the vQTLs.

Previous study[185] elaborated phantom epistasis on pairs of genetic variants with LD. We performed simulation studies based on whole genome sequencing data and demonstrated the existence of phantom epistasis on pairs of genetic variants without LD. Furthermore, we quantified the level of inflation was a function of variance explained by the additive causal genetic variant. And it was also related to genotyping strategies (array genotyping > array genotyping followed by imputation > WGS), which was consistent with the level of LD of imperfect tagging for different genotyping strategies. However, we explored four other models and none of them could fix the problem. The recognition and qualification of phantom epistasis on pairs of genetic variants without LD raise caution to interpret the epistasis and call for the genotyping strategy of whole genome sequencing, which can capture more (if not all) genetic variants.

5.3 Genetic and environmental risk score

Current established risk prediction models for complex traits or diseases rely on limited numbers of risk factors, usually environmental except family history or monogenetic variants for some cases. The recent development of polygenic/genetic risk score has brought its prediction accuracy comparable with established risk factors, so more studies are trying to assess the added value of genetic risk score combined with established risk prediction models. In addition, it is also worth exploring more broad environmental factors given more environmental factors have been collected based on questionnaires or electronic health records by the recent effort of large biobanks (e.g., UK Biobank).

We used BMI as a model trait, built genetic risk score (GRS) using state-of-the-art statistical method SBayesR, built environmental risk score (ERS) with eight environmental factors, and constructed genetic and environmental risk score (GERS) using UK Biobank data. We explored different ways to build the ERS and shown the improvement of prediction accuracy of GERS in comparison with GRS. The prediction accuracy of GERS based on the multiple linear regression was highest among different ways and reached R^2 of 18.1% tested in a randomly selected testing subset of UKB. And we also investigated the prediction of multiple GRS methods (MGRS), which combined the genetic components of environmental factors with GRS. We found no improvement of prediction accuracy over GRS, if phenotype and environmental factors were measured in the same training dataset. Our study demonstrated the value of combining genetic and environmental information for prediction and called for more environmental factors to be measured.

5.4 Future directions

To better understand the genetic and environmental influence on human complex traits, we need better datasets. The sample size is a key factor for many aspects of human complex traits analysis. A large sample size can increase the power for association mapping, including the traditional GWAS and also vQTL analysis described in Chapter 2. It can also increase the power for non-additive genetic association, while it is still elusive how large sample size is needed to identify robust, replicable, and biological meaningful signals for human complex traits, considering the statistical complexities mentioned in Chapter 3. In addition, the sample size is an important parameter to determine genetic prediction accuracy. The field has seen rapidly accumulating samples for the past decade with now reaching millions of individuals for some complex traits or diseases[44,221]. We can predict the sample size will continue to increase by the establishment of more biobanks (e.g. TOPMed[222], MVP[223], and All of Us[140]) and also by meta-analysis across cohorts in consortia (e.g. the PGC[41], GIANT[42,43], and SSGSC[44] consortia).

However, current GWAS datasets contain samples mainly collected from European ancestry and other ancestries, such as Asian, African, Latin are under-represented[224]. This problem of lack of diversity will create obstacles to transfer genetic analysis into practice across different ancestries. For example, it has been shown that the genetic prediction accuracy based on the training dataset of European individuals was far lower for non-European individuals than for European individuals[225]. So there is an increasing effort to collect more non-European samples, including GenomeAsia 100K project[226], Japan Biobank[227], African Genome Variation Project[228]. The diverse genetic datasets will also provide new opportunities for fine-mapping causal variants, GEI analysis, and other directions[229].

The project combining genetic and environmental information for predicting BMI in Chapter 4 shows the need to characterize human phenotypes/traits more broadly and comprehensively. Richer phenotypes are being collected via questionnaires, wearable devices[230], and electronic health records (EHRs)[231]. A multi-omics dataset can also be taken as phenotypes to provide another layer of information and study the genetic control of underlying molecular mechanisms, including transcriptomics[232], epigenomics[233], proteomics[234], microbiomes[235], and also these omics data in a single cell level. The rich phenotype information can be further extended to the temporal dimension. Longitudinal datasets[236] can help to distinguish the causal and reverse causal effects and exclude some confounding effects.

Another challenge we are facing is the imperfect tagging between marker variants and causal variants, which has been shown in Chapter 3. I hypothesize that sequencing data can help to solve the problem, as sequencing can capture more genetic variants with less error, especially for rare variants. The current genotyping strategy is mainly based on the SNP array followed by an imputation to a sequenced reference panel, because of its relatively low cost of < 100 dollars in comparison to WGS of around a thousand dollars per sample. However, with the anticipated further drop of sequencing price and the presence of alternative short-read sequencers (e.g. BGI sequencer[237]), we can foresee more and more individuals will be sequenced using WGS. And rare variants are expected to be included for more genetic analysis[59,238], despite many statistical analytical challenges ahead[239].

Overall, the understanding of human complex traits and diseases in the perspective of both genetic and environmental factors will be further improved with the accumulation of bigger and richer datasets and the development of novel statistical methods.

Bibliography

1. Wang H, Zhang F, Zeng J, et al. Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Sci Adv.* 2019;5(8):eaaw3538.
2. Lloyd-Jones LR, Zeng J, Sidorenko J, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat Commun.* 2019;10(1):5086.
3. Revez JA, Lin T, Qiao Z, et al. Genome-wide association study identifies 143 loci associated with 25 hydroxyvitamin D concentration. *Nat Commun.* 2020;11(1):1647.
4. Wu Y, Qi T, Wang H, et al. Promoter-anchored chromatin interactions predicted from genetic analysis of epigenomic data. *Nat Commun.* 2020;11(1):2061.
5. Zeng J, Xue A, Jiang L, et al. Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nat Commun.* 2021;12(1):1164.
6. Falconer DS, Mackay TFC. *Introduction to quantitative genetics.* 4th ed: Longman, Harlow; 1996.
7. Lynch M, Walsh B. *Genetics and analysis of quantitative traits.* Sinauer Associates, Sunderland, Ma; 1998.
8. Fisher RA. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh.* 1918;52(2):399-433.
9. Visscher PM, Bruce Walsh J. Commentary: Fisher 1918: the foundation of the genetics and analysis of complex traits. *Int J Epidemiol.* 2019;48(1):10-12.
10. Yang J, Wray NR, Visscher PM. Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genet Epidemiol.* 2010;34(3):254-257.
11. Flannick J, Florez JC. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet.* 2016;17(9):535-549.
12. Musunuru K, Kathiresan S. Genetics of Common, Complex Coronary Artery Disease. *Cell.* 2019;177(1):132-145.
13. Maier RM, Visscher PM, Robinson MR, Wray NR. Embracing polygenicity: a review of methods and tools for psychiatric genetics research. *Psychol Med.* 2017:1-19.
14. Hill WG. Understanding and using quantitative genetic variation. *Philos Trans R Soc Lond B Biol Sci.* 2010;365(1537):73-85.

15. Galton F. Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland.* 1886;15:246-263.

16. Gianola D, Rosa GJ. One hundred years of statistical developments in animal breeding. *Annu Rev Anim Biosci.* 2015;3:19-56.

17. Barton NH, Etheridge AM, Veber A. The infinitesimal model: Definition, derivation, and implications. *Theor Popul Biol.* 2017;118:50-73.

18. Provine WB. *The origins of theoretical population genetics.* 2nd ed., with a new afterword. ed. Chicago: University of Chicago Press; 2001.

19. Plomin R. Genotype-environment correlation in the era of DNA. *Behav Genet.* 2014;44(6):629-638.

20. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet.* 2008;9(4):255-266.

21. Polderman TJ, Benyamin B, de Leeuw CA, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet.* 2015;47(7):702-709.

22. Lakhani CM, Tierney BT, Manrai AK, Yang J, Visscher PM, Patel CJ. Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nat Genet.* 2019;51(2):327-334.

23. Claussnitzer M, Cho JH, Collins R, et al. A brief history of human disease genetics. *Nature.* 2020;577(7789):179-189.

24. Ozaki K, Ohnishi Y, Iida A, et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet.* 2002;32(4):650-654.

25. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science.* 2005;308(5720):385-389.

26. International HapMap Consortium. The International HapMap Project. *Nature.* 2003;426(6968):789-796.

27. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. *Nature.* 2010;467(7319):1061-1073.

28. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007;447(7145):661-678.

29. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet.* 2012;90(1):7-24.

30. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics.* 2017;101(1):5-22.

31. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet.* 2014;46(2):100-106.

32. Loh PR, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet.* 2015;47(3):284-290.

33. Yu J, Pressoir G, Briggs WH, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet.* 2006;38(2):203-208.

34. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. *Nat Genet.* 2018;50(7):906-908.

35. Jiang L, Zheng Z, Qi T, et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet.* 2019;51(12):1749-1755.

36. Mbatchou J, Barnard L, Backman J, et al. Computationally efficient whole genome regression for quantitative and binary traits. *bioRxiv.* 2020:2020.2006.2019.162354.

37. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc.* 2010;5(9):1564-1573.

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

39. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet.* 2016;48(10):1279-1283.

40. Aschard H, Vilhjalmsson BJ, Joshi AD, Price AL, Kraft P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am J Hum Genet.* 2015;96(2):329-339.

41. Sullivan PF. The psychiatric GWAS consortium: big science comes to psychiatry. *Neuron.* 2010;68(2):182-186.

42. Wood AR, Esko T, Yang J, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet.* 2014;46(11):1173-1186.

43. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature.* 2015;518(7538):197-206.

44. Lee JJ, Wedow R, Okbay A, et al. Gene discovery and polygenic prediction from a

genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet.* 2018;50(8):1112-1121.

45. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet.* 2013;14(6):379-389.

46. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562(7726):203-209.

47. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol.* 2008;32(4):381-385.

48. Buniello A, Macarthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research.* 2019;47(D1):D1005-D1012.

49. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009;461(7265):747-753.

50. Maher B. Personal genomes: The case of the missing heritability. *Nature.* 2008;456(7218):18-21.

51. Visscher PM. Sizing up human height variation. *Nat Genet.* 2008;40(5):489-490.

52. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* 2010;42(7):565-569.

53. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011;88(1):76-82.

54. Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM. Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet.* 2017;49(9):1304-1310.

55. Yang J, Manolio TA, Pasquale LR, et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet.* 2011;43(6):519-525.

56. Yang J, Bakshi A, Zhu Z, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet.* 2015;47(10):1114-1120.

57. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet.* 2011;88(3):294-305.

58. Lee SH, DeCandia TR, Ripke S, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet.* 2012;44(3):247-250.

59. Wainschtein P, Jain D, Zheng Z, et al. Recovery of trait heritability from whole genome sequence data. *bioRxiv.* 2021:588020.

60. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet.* 2017;18(2):117-127.

61. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 2015;47(3):291-295.

62. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet.* 2015;47(11):1228-1235.

63. Speed D, Holmes J, Balding DJ. Evaluating and improving heritability models using summary statistics. *Nat Genet.* 2020;52(4):458-462.

64. Zeng J, de Vlaming R, Wu Y, et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nat Genet.* 2018;50(5):746-753.

65. Zhang Y, Qi G, Park JH, Chatterjee N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat Genet.* 2018;50(9):1318-1326.

66. Hou K, Burch KS, Majumdar A, et al. Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat Genet.* 2019;51(8):1244-1251.

67. Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet.* 1972;2(1):3-19.

68. Wu Y, Sankararaman S. A scalable estimator of SNP heritability for biobank-scale data. *Bioinformatics.* 2018;34(13):i187-i194.

69. Zhu Z, Bakshi A, Vinkhuyzen AA, et al. Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet.* 2015;96(3):377-385.

70. Pazokitoroudi A, Chiu AM, Burch KS, Pasaniuc B, Sankararaman S. Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. *Am J Hum Genet.* 2021;108(5):799-808.

71. Hivert V, Sidorenko J, Rohart F, et al. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *Am J Hum Genet.* 2021;108(5):786-798.

72. van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. Genetic correlations of

polygenic disease traits: from theory to practice. *Nat Rev Genet.* 2019;20(10):567-581.

73. Cross-Disorder Group of the Psychiatric Genomics Consortium, Lee SH, Ripke S, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet.* 2013;45(9):984-994.

74. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet.* 2015;47(11):1236-1241.

75. Ni G, Moser G, Schizophrenia Working Group of the Psychiatric Genomics C, Wray NR, Lee SH. Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. *Am J Hum Genet.* 2018;102(6):1185-1194.

76. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol.* 2003;32(1):1-22.

77. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ.* 2018;362:k601.

78. Morrison J, Knoblauch N, Marcus JH, Stephens M, He X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat Genet.* 2020;52(7):740-747.

79. Ottman R. Gene–environment interaction: definitions and study design. *Preventive medicine.* 1996;25(6):764-770.

80. Thomas D. Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics.* 2010;11(4):259.

81. Wagenmakers EJ, Krypotos AM, Criss AH, Iverson G. On the interpretation of removable interactions: a survey of the field 33 years after Loftus. *Mem Cognit.* 2012;40(2):145-160.

82. Kraft P, Hunter D. Integrating epidemiology and genetic association: the challenge of gene–environment interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences.* 2005;360(1460):1609-1616.

83. Hunter DJ. Gene-environment interactions in human diseases. *Nat Rev Genet.* 2005;6(4):287-298.

84. Garrod AE. The incidence of alkaptonuria: a study in chemical individuality. *The Lancet.* 1902;160(4137):1616-1620.

85. Haldane J. Heredity and politics. In: WW Norton & Co., NY; 1938.

86. Ritz BR, Chatterjee N, Garcia-Closas M, et al. Lessons Learned From Past Gene-Environment Interaction Successes. *Am J Epidemiol.* 2017;186(7):778-786.

87. Figueroa JD, Han SS, Garcia-Closas M, et al. Genome-wide interaction study of smoking and bladder cancer risk. *Carcinogenesis.* 2014;35(8):1737-1744.

88. Kilpeläinen TO, Qi L, Brage S, et al. Physical activity attenuates the influence of FTO variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. *PLoS medicine.* 2011;8(11):e1001116.

89. Wu C, Kraft P, Zhai K, et al. Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. *Nat Genet.* 2012;44(10):1090-1097.

90. Aschard H, Lutz S, Maus B, et al. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. *Hum Genet.* 2012;131(10):1591-1613.

91. McAllister K, Mechanic LE, Amos C, et al. Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. *Am J Epidemiol.* 2017;186(7):753-761.

92. Patel CJ, Kerr J, Thomas DC, et al. Opportunities and Challenges for Environmental Exposure Assessment in Population-Based Studies. *Cancer Epidemiol Biomarkers Prev.* 2017;26(9):1370-1380.

93. Gauderman WJ, Mukherjee B, Aschard H, et al. Update on the State of the Science for Analytical Methods for Gene-Environment Interactions. *Am J Epidemiol.* 2017;186(7):762-770.

94. Ronnegard L, Valdar W. Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics.* 2011;188(2):435-447.

95. Robinson MR, English G, Moser G, et al. Genotype–covariate interaction effects and the heritability of adult body mass index. *Nature genetics.* 2017;49(8):1174.

96. Ni G, van der Werf J, Zhou X, Hypponen E, Wray NR, Lee SH. Genotype-covariate correlation and interaction disentangled by a whole-genome multivariate reaction norm model. *Nat Commun.* 2019;10(1):2239.

97. Shin J, Lee SH. GxEsum: genotype-by-environment interaction model based on summary statistics. *bioRxiv.* 2020:2020.2005.2031.122549.

98. Dahl A, Nguyen K, Cai N, Gandal MJ, Flint J, Zaitlen N. A Robust Method Uncovers Significant Context-Specific Heritability in Diverse Complex Traits. *Am J Hum Genet.* 2020;106(1):71-91.

99. Kerin M, Marchini J. Inferring Gene-by-Environment Interactions with a Bayesian

Whole-Genome Regression Model. *Am J Hum Genet.* 2020;107(4):698-713.

100. Bateson W. *Mendel's principles of heredity.* Cambridge: Cambridge University Press; 1909.

101. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet.* 2002;11(20):2463-2468.

102. Wei WH, Hemani G, Haley CS. Detecting epistasis in human complex traits. *Nat Rev Genet.* 2014;15(11):722-733.

103. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet.* 2009;10(6):392-404.

104. Weir BS. Linkage disequilibrium and association mapping. *Annu Rev Genomics Hum Genet.* 2008;9:129-142.

105. Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* 2008;4(2):e1000008.

106. Mackay TF. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet.* 2014;15(1):22-33.

107. Mackay TF, Moore JH. Why epistasis is important for tackling complex human disease genetics. *Genome Med.* 2014;6(6):124.

108. Domingo J, Diss G, Lehner B. Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature.* 2018;558(7708):117-121.

109. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics.* 2016;17(7):392.

110. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics.* 2018;19(9):581.

111. Kraft P, Hunter DJ. Genetic risk prediction—are we there yet? *New England Journal of Medicine.* 2009;360(17):1701-1703.

112. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome research.* 2007;17(10):1520-1528.

113. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature.* 2009;460(7256):748.

114. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One.* 2008;3(10):e3395.

115. Wray NR, Kemper KE, Hayes BJ, Goddard ME, Visscher PM. Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans:

Genomic Prediction. *Genetics.* 2019;211(4):1131-1141.

116. Robinson GK. That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science.* 1991;6(1):15-32.

117. Robinson MR, Kleinman A, Graff M, et al. Genetic evidence of assortative mating in humans. *Nature Human Behaviour.* 2017;1(1).

118. Vilhjálmsson BJ, Yang J, Finucane HK, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics.* 2015;97(4):576-592.

119. Marquez-Luna C, Gazal S, Kim S, Furlotte N, Auton A, Price A. Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *BioRxiv.* 2019.

120. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.* 2015;11(4):e1004969.

121. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol.* 2017;41(6):469-480.

122. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun.* 2019;10(1):1776.

123. Chun S, Imakaev M, Hui D, et al. Non-parametric Polygenic Risk Prediction via Partitioned GWAS Summary Statistics. *Am J Hum Genet.* 2020;107(1):46-59.

124. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics.* 2018;50(9):1219.

125. Inouye M, Abraham G, Nelson CP, et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *Journal of the American College of Cardiology.* 2018;72(16):1883-1893.

126. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet.* 2019;28(R2):R133-R142.

127. Yang J, Lee T, Kim J, et al. Ubiquitous polygenicity of human complex traits: genome-wide analysis of 49 traits in Koreans. *PLoS genetics.* 2013;9(3):e1003355.

128. Shi H, Kichaev G, Pasaniuc B. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics.* 2016;99(1):139-153.

129. Abadi A, Alyass A, du Pont SR, et al. Penetrance of polygenic obesity susceptibility

loci across the body mass index distribution. *The American Journal of Human Genetics.* 2017;101(6):925-938.

130. Nagpal S, Gibson G, Marigorta U. Pervasive modulation of obesity risk by the environment and genomic background. *Genes.* 2018;9(8):411.

131. Pare G, Cook NR, Ridker PM, Chasman DI. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genet.* 2010;6(6):e1000981.

132. Metzger BP, Yuan DC, Gruber JD, Duveau F, Wittkopp PJ. Selection on noise constrains variation in a eukaryotic promoter. *Nature.* 2015;521(7552):344.

133. Cao Y, Wei P, Bailey M, Kauwe JSK, Maxwell TJ. A versatile omnibus test for detecting mean and variance heterogeneity. *Genet Epidemiol.* 2014;38(1):51-59.

134. Ek WE, Rask-Andersen M, Karlsson T, Enroth S, Gyllensten U, Johansson A. Genetic variants influencing phenotypic variance heterogeneity. *Hum Mol Genet.* 2018;27(5):799-810.

135. Rönnegård L, Valdar W. Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC genetics.* 2012;13(1):63.

136. Van Vleck LD. Variation of milk records within paternal-sib groups. *Journal of Dairy Science.* 1968;51(9):1465-1470.

137. Hill WG, Mulder HA. Genetic analysis of environmental variation. *Genetics Research.* 2010;92(5-6):381-395.

138. Struchalin MV, Dehghan A, Witteman JC, van Duijn C, Aulchenko YS. Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. *BMC Genet.* 2010;11:92.

139. Yang J, Loos RJ, Powell JE, et al. FTO genotype is associated with phenotypic variability of body mass index. *Nature.* 2012;490(7419):267-272.

140. Collins FS, Varmus H. A new initiative on precision medicine. *New England Journal of Medicine.* 2015;372(9):793-795.

141. Conover WJ, Johnson ME, Johnson MM. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics.* 1981;23(4):351-361.

142. Bartlett MS. Properties of sufficiency and statistical tests. Paper presented at: Proc. R. Soc. Lond. A1937.

143. Levene H. Robust Tests for Equality of Variances. *In Ingram Olkin; Harold Hotelling; et al Contributions to Probability and Statistics: Essays in Honor of*

Harold Hotelling Stanford University Press, Stanford. 1960: 278–292.

144. Brown MB, Forsythe AB. Robust tests for the equality of variances. *Journal of the American Statistical Association.* 1974;69(346):364-367.

145. Fligner MA, Killeen TJ. Distribution-free two-sample tests for scale. *Journal of the American Statistical Association.* 1976;71(353):210-213.

146. Ronnegard L, Felleki M, Fikse F, Mulder HA, Strandberg E. Genetic heterogeneity of residual variance - estimation of variance components using double hierarchical generalized linear models. *Genet Sel Evol.* 2010;42:8.

147. Smyth GK. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society Series B (Methodological).* 1989:47-60.

148. Sun X, Elston R, Morris N, Zhu X. What is the significance of difference in phenotypic variability across SNP genotypes? *Am J Hum Genet.* 2013;93(2):390-397.

149. Corty RW, Valdar W. QTL Mapping on a Background of Variance Heterogeneity. *G3 (Bethesda).* 2018;8(12):3767-3782.

150. Wu Y, Zheng Z, Visscher PM, Yang J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol.* 2017;18(1):86.

151. Pulit SL, de With SA, de Bakker PI. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. *Genetic epidemiology.* 2017;41(2):145-151.

152. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 2014;10(5):e1004383.

153. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet.* 2016;48(5):481-487.

154. Young AI, Wauthier FL, Donnelly P. Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nature Genetics.* 2018;50(11):1608-1614.

155. Yang J, Weedon MN, Purcell S, et al. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics.* 2011;19(7):807.

156. Saccone SF, Hinrichs AL, Saccone NL, et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Human molecular genetics.* 2006;16(1):36-49.

157. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine

dependence, lung cancer and peripheral arterial disease. *Nature.* 2008;452(7187):638.

158. Fowler CD, Lu Q, Johnson PM, Marks MJ, Kenny PJ. Habenular α5 nicotinic receptor subunit signalling controls nicotine intake. *Nature.* 2011;471(7340):597.

159. Repapi E, Sayers I, Wain LV, et al. Genome-wide association study identifies five loci associated with lung function. *Nature genetics.* 2010;42(1):36.

160. Hancock DB, Eijgelsheim M, Wilk JB, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nature genetics.* 2010;42(1):45.

161. Wain LV, Shrine N, Artigas MS, et al. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nature genetics.* 2017;49(3):416.

162. Kaur-Knudsen D, Nordestgaard BG, Bojesen SE. CHRNA3 genotype, nicotine dependence, lung function and disease in the general population. *European Respiratory Journal.* 2012;40(6):1538-1544.

163. Estrada K, Styrkarsdottir U, Evangelou E, et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature genetics.* 2012;44(5):491.

164. Kemp JP, Morris JA, Medina-Gomez C, et al. Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nature genetics.* 2017;49(10):1468.

165. Medina-Gomez C, Kemp JP, Estrada K, et al. Meta-analysis of genome-wide scans for total body BMD in children and adults reveals allelic heterogeneity and age-specific effects at the WNT16 locus. *PLoS genetics.* 2012;8(7):e1002718.

166. Movérare-Skrtic S, Henning P, Liu X, et al. Osteoblast-derived WNT16 represses osteoclastogenesis and prevents cortical bone fragility fractures. *Nature medicine.* 2014;20(11):1279.

167. Frayling TM, Timpson NJ, Weedon MN, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science.* 2007;316(5826):889-894.

168. Smemo S, Tena JJ, Kim K-H, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature.* 2014;507(7492):371.

169. Claussnitzer M, Dankel SN, Kim K-H, et al. FTO obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine.* 2015;373(10):895-907.

170. Loos RJ, Yeo GS. The bigger picture of FTO—the first GWAS-identified obesity gene. *Nature Reviews Endocrinology.* 2014;10(1):51.

171. Moore R, Casale FP, Jan Bonder M, et al. A linear mixed-model approach to study multivariate gene–environment interactions. *Nature Genetics.* 2018.

172. Zhang F, Chen W, Zhu Z, et al. OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol.* 2019;20(1):107.

173. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature.* 2015;526(7571):82-90.

174. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.

175. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature.* 2010;467(7319):1061.

176. Bretherton CS, Widmann M, Dymnikov VP, Wallace JM, Bladé I. The effective number of spatial degrees of freedom of a time-varying field. *Journal of climate.* 1999;12(7):1990-2009.

177. IPAQ Research Committee. Guidelines for data processing and analysis of the International Physical Activity Questionnaire (IPAQ)-short and long forms. 2005.

178. Beasley TM, Erickson S, Allison DB. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav Genet.* 2009;39(5):580-595.

179. Peng B, Yu RK, Dehoff KL, Amos CI. Normalizing a large number of quantitative traits using empirical normal quantile transformation. *BMC Proc.* 2007;1 Suppl 1:S156.

180. Craig CL, Marshall AL, Sjostrom M, et al. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc.* 2003;35(8):1381-1395.

181. Collins FS, Guyer MS, Charkravarti A. Variations on a theme: cataloging human DNA sequence variation. *Science.* 1997;278(5343):1580-1581.

182. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet.* 2018;19(8):491-504.

183. Hemani G, Shakhbazov K, Westra HJ, et al. Detection and replication of epistasis influencing transcription in humans. *Nature.* 2014;508(7495):249-253.

184. Wood AR, Tuke MA, Nalls MA, et al. Another explanation for apparent epistasis. *Nature.* 2014;514(7520):E3-5.

185. de Los Campos G, Sorensen DA, Toro MA. Imperfect Linkage Disequilibrium

Generates Phantom Epistasis (& Perils of Big Data). *G3 (Bethesda).* 2019;9(5):1429-1436.

186. Wray NR. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Res Hum Genet.* 2005;8(2):87-94.

187. Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered.* 2003;56(1-3):18-31.

188. Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* 2009;5(5):e1000477.

189. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009;5(6):e1000529.

190. International HapMap Consortium, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 2007;449(7164):851-861.

191. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68-74.

192. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ.* 2009;338:b375.

193. Steyerberg EW. *Clinical prediction models.* Vol 381: Springer; 2009.

194. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-73.

195. Meisner A, Chatterjee N. Disease Risk Models. *Handbook of Statistical Genomics: Two Volume Set.* 2019:815-842.

196. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: A historical perspective. *The Lancet.* 2014;383(9921):999-1008.

197. Goff DC, Jr., Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation.* 2014;129(25 Suppl 2):S49-73.

198. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation.* 2019;140(11):e596-e646.

199. Force USPST, Curry SJ, Krist AH, et al. Risk Assessment for Cardiovascular Disease With Nontraditional Risk Factors: US Preventive Services Task Force Recommendation Statement. *JAMA.* 2018;320(3):272-280.

200. Lin JS, Evans CV, Johnson E, Redmond N, Coppola EL, Smith N. Nontraditional Risk Factors in Cardiovascular Disease Risk Assessment: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA.* 2018;320(3):281-297.

201. Visscher PM, Yang J, Goddard ME. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Res Hum Genet.* 2010;13(6):517-524.

202. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 2013;9(3):e1003348.

203. Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. *Statistical Science.* 2009;24(4):451-471.

204. Dudbridge F, Pashayan N, Yang J. Predictive accuracy of combined genetic and environmental risk scores. *Genet Epidemiol.* 2018;42(1):4-19.

205. Mars N, Koskela JT, Ripatti P, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med.* 2020;26(4):549-557.

206. Mosley JD, Gupta DK, Tan J, et al. Predictive Accuracy of a Polygenic Risk Score Compared With a Clinical Risk Score for Incident Coronary Heart Disease. *JAMA.* 2020;323(7):627-635.

207. Elliott J, Bodinier B, Bond TA, et al. Predictive accuracy of a polygenic risk score–enhanced prediction model vs a clinical risk score for coronary artery disease. *Jama.* 2020;323(7):636-645.

208. Riveros-Mckay Aguilera F, Weale ME, Moore R, et al. An integrated polygenic and clinical risk tool enhances coronary artery disease prediction. *medRxiv.* 2020:2020.2006.2001.20119297.

209. Wilcox AN, Choudhury PP, Gao C, et al. Prospective Evaluation of a Breast Cancer Risk Model Integrating Classical Risk Factors and Polygenic Risk in 15 Cohorts from

Six Countries. *medRxiv.* 2019:19011171.

210. Garcia-Closas M, Gunsoy NB, Chatterjee N. Combined associations of genetic and environmental risk factors: implications for prevention of breast cancer. *J Natl Cancer Inst.* 2014;106(11).

211. Lee A, Mavaddat N, Wilcox AN, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet Med.* 2019;21(8):1708-1718.

212. Kachuri L, Graff RE, Smith-Byrne K, et al. Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat Commun.* 2020;11(1):6084.

213. Zhou X, Lee SH. An integrative analysis of genomic and exposomic data for complex traits and phenotypic prediction. *bioRxiv.* 2020:2020.2011.2009.373704.

214. He Y, Lakhani CM, Manrai AK, Patel CJ. Poly-Exposure and Poly-Genomic Scores Implicate Prominent Roles of Non-Genetic and Demographic Factors in Four Common Diseases in the UK. *bioRxiv.* 2019:833632.

215. Zhu Z, Zheng Z, Zhang F, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat Commun.* 2018;9(1):224.

216. Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M. Variable prediction accuracy of polygenic scores within an ancestry group. *Elife.* 2020;9.

217. Maier RM, Zhu Z, Lee SH, et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat Commun.* 2018;9(1):989.

218. Turley P, Walters RK, Maghzian O, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet.* 2018;50(2):229-237.

219. Craig JE, Han X, Qassim A, et al. Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. *Nat Genet.* 2020;52(2):160-166.

220. Krapohl E, Patel H, Newhouse S, et al. Multi-polygenic score approach to trait prediction. *Mol Psychiatry.* 2018;23(5):1368-1374.

221. Surendran P, Feofanova EV, Lahrouchi N, et al. Discovery of rare variants associated with blood pressure regulation through meta-analysis of 1.3 million individuals. *Nat Genet.* 2020.

222. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv.* 2019:563866.

223. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: A mega-biobank

to study genetic influences on health and disease. *J Clin Epidemiol.* 2016;70:214-223.

224. Gurdasani D, Barroso I, Zeggini E, Sandhu MS. Genomics of disease risk in globally diverse populations. *Nat Rev Genet.* 2019;20(9):520-535.

225. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet.* 2019;51(4):584-591.

226. GenomeAsia 100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature.* 2019;576(7785):106-111.

227. Nagai A, Hirata M, Kamatani Y, et al. Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol.* 2017;27(3S):S2-S8.

228. Gurdasani D, Carstensen T, Tekola-Ayele F, et al. The African Genome Variation Project shapes medical genetics in Africa. *Nature.* 2015;517(7534):327-332.

229. Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature.* 2019;570(7762):514-518.

230. Price ND, Magis AT, Earls JC, et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat Biotechnol.* 2017;35(8):747-756.

231. Li R, Chen Y, Ritchie MD, Moore JH. Electronic health records and polygenic risk scores for predicting disease risk. *Nat Rev Genet.* 2020;21(8):493-502.

232. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science.* 2020;369(6509):1318-1330.

233. eGTEx Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat Genet.* 2017;49(12):1664-1670.

234. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. *Nature.* 2018;558(7708):73-79.

235. Rothschild D, Weissbrod O, Barkan E, et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature.* 2018;555(7695):210-215.

236. Caruana EJ, Roman M, Hernandez-Sanchez J, Solli P. Longitudinal studies. *J Thorac Dis.* 2015;7(11):E537-540.

237. Senabouth A, Andersen S, Shi Q, et al. Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing. *NAR Genomics and Bioinformatics.* 2020;2(2).

238. Fuchsberger C, Flannick J, Teslovich TM, et al. The genetic architecture of type 2 diabetes. *Nature.* 2016;536(7614):41-47.

239. Li X, Li Z, Zhou H, et al. Dynamic incorporation of multiple in silico functional

annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet.* 2020;52(9):969-983.