A Multi-Ancestry Study of Gene-Lifestyle Interactions for Cardiovascular Traits in 610,475 Individuals from 124 Cohorts: Design and Rationale

D.C. Rao, Ph.D.[1], Yun Ju Sung, Ph.D.[1], Thomas W. Winkler, Ph.D.[2], Karen Schwander, M.S.[1], Ingrid Borecki, Ph.D.[3], L. Adrienne Cupples, Ph.D.[4], W. James Gauderman, Ph.D.[5], Kenneth Rice, Ph.D.[6], Patricia B. Munroe, Ph.D.[7,8], and Bruce Psaty, M.D., Ph.D.[9] on behalf of the CHARGE Gene-Lifestyle Interactions Working Group

Running Title: **Multi-Ancestry Study of Gene-Lifestyle Interactions**

[1] Division of Biostatistics, Washington University in St. Louis, School of Medicine, St. Louis, MO

[2] Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany

[3] Division of Statistical Genomics in the Center for Genome Sciences of the Washington University, St. Louis, USA.

[4] Department of Biostatistics, Boston University School of Public Health, Boston, MA; NHLBI Framingham Heart Study, Framingham, MA

[5] Division of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, CA

[6] Department of Biostatistics, University of Washington, Seattle, WA

[7] Clinical Pharmacology, William Harvey Research Institute, Queen Mary University of London, London, London, UK;

[8] NIHR Barts Cardiovascular Biomedical Research Unit, Queen Mary University of London, London, UK

[9] Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA; Group Health Research Institute, Group Health Cooperative, Seattle, WA.

Address correspondence to:

Dr. D. C. Rao
Division of Biostatistics
Washington University in St. Louis
660 S. Euclid Avenue, Campus Box 8067
St. Louis, MO 63110, USA
e-Mail: rao@wustl.edu
Phone: 314-362-3608
Fax: 314-362-2693 (ATTN: DC Rao)

Abstract

Background -- Several consortia have pursued genome-wide association studies for identifying novel genetic loci associated with various diseases and disease related risk factors including blood pressure (BP), lipids, hypertension, type 2 diabetes. They demonstrated the power of collaborative research through meta-analysis of study-specific results.

Methods -- The Gene-Lifestyle Interactions Working Group was formed to facilitate and promote the first large, concerted, multi-ancestry study to systematically evaluate gene-lifestyle interactions. In Stage 1, genome-wide interaction analysis is carried out in (up to) 53 cohorts with a total of 149,684 individuals from multiple ancestries. In Stage 2 involving an additional (up to) 71 cohorts with 460,791 individuals from multiple ancestries, focused analysis is carried out for a subset of the most promising variants from Stage 1. In all, the study involves up to 124 cohorts with 610,475 individuals. Current focus is on cardiovascular traits including blood pressure and lipids, and lifestyle factors including smoking, alcohol, education (as a surrogate for socio-economic status), physical activity, psychosocial variables, and sleep. The total sample sizes vary among projects due to missing data. Large scale gene-lifestyle or more generally gene-environment interaction (GxE) meta-analysis studies can be cumbersome and challenging. This paper describes the design and some of the approaches pursued in the interaction projects led by the Working Group.

Conclusions – The Gene-Lifestyle Interactions Working Group provides an excellent framework for understanding the lifestyle context of genetic effects and to identify novel trait loci through analysis of interactions. An important and novel feature of our study is that the gene-lifestyle interaction (GxE) results may improve our knowledge about the underlying mechanisms for novel as well as already known trait loci.

Key Words: Gene-Lifestyle Interactions, GWAS, GxE, Meta-analysis

Introduction

Remarkable advances in genomics, including the Human Genome Project (HGP) and 1000 Genomes (1000G) Project, have revolutionized methods for genetic dissection of common complex diseases and disease traits. Using Genome-Wide Association Studies (GWAS), large consortia such as CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology[1], ICBP (International Consortium of Blood Pressure), AGEN Asian Genetic Epidemiology Network), GLGC (Global Lipids Genetics Consortium), and DIAGRAM (Diabetes Genetics Replication and Meta-Analysis) have identified hundreds of common genetic variants associated with many common complex disease traits (https://www.genome.gov/26525384/catalog-of-published-genomewide-association-studies/). However, most of the identified genetic variants explain small proportions of the trait heritability, mostly through small main effects of common variants. It has been recognized that this focus on main effects may have become a barrier to further progress[2,3].

Hypertension and dyslipidemia are common complex disorders that contribute to two of the leading causes of death (cardiovascular and cerebrovascular disease) and exhibit significant patterns of health disparity among racial/ancestral groups in the US[4,5]. While lifestyle factors have long been recognized as risk factors, modulation of the effects of genetic variants by lifestyle factors, and the underlying candidate pathobiological mechanisms have not received much attention. Understanding these genetic modifiers is important because it may provide valuable clues for lifestyle-based interventions which may result in a more successful management of these health conditions through personalized therapies, and may explain part of the "missing heritability" [2,6].

The Gene-Lifestyle Interactions Working Group (hereafter referred to as "this study") investigates gene-lifestyle interactions for uncovering more of the unexplained genetic variance in BP and lipids and for gaining insights into the biological mechanisms influencing these important morbid conditions. We will do this by leveraging the extensive resources of existing studies in multiple ancestries that have data on phenotypes, lifestyle factors, and dense genotype data from both common variants (GWAS) and rare variants (Exome chip). We will also use the organizational infrastructure of the CHARGE consortium.

Research involving gene-environment (GxE) interactions is now being reported[7,8]. Our own work[9] and other studies have demonstrated the promise of GxE interactions for identifying genetic variants with large effects[10-13]. For example, mean triglyceride levels are 23 mg/dL lower in

physically active versus sedentary individuals (88 vs 111 md/dL) who carry a C-allele at rs2070744 in *NOS*3, but there is little difference by physical activity status in TT homozygotes[11]. This shows the utility of GxE interactions for using genetic information to identify subpopulations in whom modifying the environmental factors is beneficial[14-16], and that the main effect (of the genetic variant) alone is inadequate to inform lifestyle interventions that need to be personalized based on genotype[17,18]. In addition, GxE interactions may provide additional insight into biological mechanisms and pathways.

This is the first large, concerted, multi-ancestry study to systematically evaluate gene-lifestyle interactions using data from 610,475 individuals. Large scale GxE meta-analysis studies can be cumbersome and challenging. This paper describes the design and some of the approaches pursued in our ongoing Gene-Lifestyle Interaction projects.

Study Design

The CHARGE Consortium: This study leverages the infrastructure created by the CHARGE consortium[1]. CHARGE has created many resources including multiple phenotype-specific Working Groups (WGs), an analysis committee, an internal wiki site, guidelines for collaboration and authorship, and periodic CHARGE meetings where WGs meet in person.

The Gene-Lifestyle Interactions WG: With support from CHARGE leadership, a new WG has been established for pursuing the major goals of this study. The WG includes investigators and analysts from the large group of studies participating in **Stage 1** (Genome-Wide Discovery) as discussed later. Another large group of studies participates in **Stage 2** (Focused Discovery/Replication). The WG is assisted by a Coordinating Center (CC) at Washington University in St. Louis.

This study operates through the Working Group (WG), which serves as a steering committee, an Analysis Committee, a Harmonization Committee, and multiple Project Teams. The WG meets twice a year as part of the CHARGE meetings and meets by conference call twice a month. Overall research direction and priorities are set by the WG. The analysis and harmonization committees meet together once a year and by conference calls twice a month. All harmonization and analytical issues are resolved by these committees. There are multiple Project Teams, each leading interaction analyses for a combination of the phenotypes (BP or Lipids) and lifestyle domains (smoking, alcohol, education, PA, Psychosocial, Sleep).

Mission and Aim: The overall mission of the WG is to promote and facilitate large collaborative analysis of gene-lifestyle interactions on disease traits across a large number of cohorts from multiple ancestries. Primarily, the WG aims to better understand the lifestyle context of genetic effects and to discover new trait loci through analysis of interactions thereby explaining part of the missing heritability[2] in the disease traits. An important and novel feature of our study is that the gene-lifestyle interaction (GxE) results may improve our knowledge about the underlying mechanisms for novel as well as already known trait loci.

Primary Hypothesis: We hypothesize that lifestyle (environment) variables modulate some of the genetic effects on cardiovascular traits (equivalently, that genetic variants modify effects of environmental variables) and that accounting for lifestyle factors and gene-lifestyle interactions in genome-wide scans will identify multiple novel genetic variants.

Phenotypes and Lifestyle Variables: The primary cardiovascular (CV) risk factors include BP and lipids. An analysis plan in the online supplement discusses data definitions and adjustments. Future initiatives may consider other cardio-metabolic traits, such as diabetes and its risk factors, in collaboration with other working groups.

The primary BP phenotypes are resting/sitting Systolic Blood Pressure (SBP) (mmHg) and Diastolic Blood Pressure (DBP) (mmHg). For individuals taking any anti-hypertensive (BP lowering) medications, their SBP and DBP values are first adjusted by adding 15 mmHg to SBP and adding 10 mmHg to DBP. Mean Arterial Pressure (MAP) and Pulse Pressure (PP) are also derived, using the adjusted SBP and DBP values:

a. MAP = DBP + (SBP – DBP)/3, and

b. PP = SBP – DBP

The primary lipids phenotypes are High-density lipoprotein cholesterol (HDL, mg/dL), Triglycerides (TG, mg/dL) and Low-density lipoprotein cholesterol (LDL, mg/dL), either directly assayed (LDL_{da}) or derived using the Friedewald equation (LDL_F). For individuals with TG > 400 mg/dL, only directly assayed LDL (LDL_{da}) is used. When using non-fasting samples or fasting < 8 hours, only LDL_{da} and HDL are used (not LDL_F or TG). Log transformations are used for HDL and TG, and LDL is adjusted for statin use (see the analysis plan in the supplementary materials).

The initial set of dichotomized lifestyle are: Smoking (current smoking and ever smoking), Alcohol Consumption (Current Drinking, Current Regular Drinking, and Quantity of Drinks (>7 drinks per week)), Education (as a measure of socioeconomic status, SES; Some College, and Graduated

College), Physical Activity (Physically Inactive), Psychosocial Attributes (Depression, Trait Anxiety, and Social Support), and Sleep Duration (Short Sleep and Long Sleep). Future initiatives may consider other domains such as diet, and more detailed variables from the same lifestyle domains such as pack years, cigarettes per day, ounces of alcohol intake.

GWAS Data: Dosages derived from 1000 Genomes (1000G) imputation are the primary resource for GWAS analysis. 1000G imputations are based on the ALL ancestry panel from 1000G Phase I Integrated Release Version 3 Haplotypes (2010-11 data freeze, 2012-03-14 haplotypes) that contains haplotypes of 1,092 individuals of all ancestral backgrounds. Dosages based on HapMap Phase II / III reference panel is used if 1000G imputations are not available for a specific study. In general, rare variants (mean allele frequency (MAF) < 1%) and poorly imputed variants (Rsq <0.1) are excluded. Variants mapping to sex chromosomes or mitochondria have also been excluded. Although we refer to SNP (Single Nucleotide Polymorphism) variants, the imputed data also include indels (insertions and deletions).

Participating Studies and Ancestry Groups: Five ancestry groups are represented: European (EA), African (AA), Hispanic (HA), Asian (AS), and Brazilian admixed (BR). Men and women between the ages of 18-80 are included in the analyses. Although the participating studies are based on different study designs and populations, most of them have data on BP and lipid traits, a range of lifestyle variables, and genotypes across the genome. In total, this study comprises up to 610,475 individuals.

Stage 1 (Genome-Wide Discovery): A total of 32 studies with data on 53 cohorts (see **Table 1**) participate in the discovery phase (Stage 1), which involves genome-wide interaction analyses. In total, this stage includes up to 95,911 EA, 27,116 AA, 8,805 HA, 13,438 AS, and 4,414 BR individuals, to an overall total of 149,684 individuals in Stage 1.

Stage 2 (Focused Discovery/Replication): A total of 46 studies with data on 71 cohorts (see **Table 2**) participate in Stage 2, which involves analyses of small sets of variants that were identified in Stage 1 as either genome-wide *significant* (with $p < 10^{-8}$) or *suggestive* (with $p < 10^{-6}$). In total, this stage includes up to 290,552 EA, 7,785 AA, 13,522 HA, and 148,932 AS individuals, to a total of 460,791 individuals in Stage 2. There are no BR cohorts in Stage 2.

Analysis Models

The participating studies have considerable prior experience contributing to GWAS-based consortia studying the main effects of common variants, i.e. effects of genetic variants without regard to lifestyle exposures or interactions. For GxE work, existing analysis pipelines had to be modified. Based on extensive discussions with the Analysis Committee and the Working Group, an Analysis Plan was developed, addressing critical issues including: data preparation, analysis models, analysis methods, software packages, and procedures for uploading all results centrally onto a central server made available by the CHARGE Consortium at the University of Washington, Seattle. Individual project teams made appropriate modifications to the analysis plan as needed. The most critical elements are summarized below. An example of a full analysis plan (Education-Lipids) is provided in the **online supplement**.

We consider three different analysis models, each with slightly different purposes.

Joint model (Model 1): This is our primary model which features joint analysis of the *effects of the SNP, lifestyle, and their interaction*. For each combination of phenotype (Y) and lifestyle exposure variable (E), each study fits the following linear model, separately by ancestry:

Y ~ E + SNP + E * SNP + C, or more formally,

$$E(Y) = \beta_0 + \beta_E E + \beta_G SNP + \beta_{GE} E * SNP + \beta_C C$$

where SNP is the dosage of the genetic variant and C is the set of covariates (age, sex, principal components for controlling stratification effects, and other study-specific covariates, and therefore β_C is a vector; body mass index (BMI) was specifically excluded as a covariate so that lifestyle interactions with related pathway genes (such as inflammation genes) can be identified). Participating studies provide estimates of β_G and β_{GE} along with their covariance matrix. If E is dichotomous (E= 0 or 1), the SNP effect (β_G) represents the SNP effect in those who are unexposed (environmental variable E=0), and thus needs to be interpreted with caution. If E is continuous, it is often desirable to center it on its sample mean, so that β_G approximates the overall effect of the SNP on Y (as is estimated by Model 2). In either case, the SNP effect is context-dependent and therefore should not be interpreted as the ”main effect”.

Model 1 was used by all studies in both stages. In addition to model 1, each study in stage 1 (only) uses at least one of two additional models presented below, depending on the specific needs of the respective project.

Main effects model (Model 2): Analysis of the *main effect only*: For each Phenotype (Y), each study fits the following linear model, separately by ancestry:

Y ~ SNP + C, or more formally,

$$E(Y) = \lambda_0 + \lambda_G SNP + \lambda_C C$$

Model 2 is used as a benchmark to identify which of our discoveries from the joint model would be found using analysis of main effects alone. Some projects also fit this model separately in the exposed and unexposed groups (i.e. they performed stratified analysis) and provide a 1 degree of freedom (df) test of the interaction term as well as a 2 df joint test of the SNP and interaction effects[19,20]. For each analysis, participating studies provide estimates of β_G and and its standard error. Stratified analysis and the joint analysis using model 1 in stage 1 cohorts have been shown to yield largely similar results[21]. Stratified analysis can help reduce inflation of type I error rates by fitting separate covariate effects and error variances by strata[22-24].

Refined main effects model (Model 3): Analysis of the *SNP and lifestyle effects*, without interaction. For each Phenotype (Y) and lifestyle exposure variable (E), each study fits the following linear model, separately by ancestry:

Y ~ E + SNP + C, or more formally,

$$E(Y) = \gamma_0 + \gamma_E E + \gamma_G SNP + \gamma_C C$$

Model 3 is used to identify which of our discoveries from the joint model would be missed when the interaction term is not used. For each analysis, participating studies provide estimates of β_G and and its standard error.

Analysis Methods

Analysis Methods for Low Frequency and Common Variants: Through the use of efficient methods with large sample sizes, we believe that our study is poised to identify multiple novel associations, some of which may have large effect sizes, depending on lifestyle factors. We identify novel loci through SNP effects or SNP*E interaction effects, or both. For continuous traits, the joint test of the SNP and SNP*E interaction effects is known to be powerful for this aim[20, 25,26]. Since our interaction projects involve many studies, we rely on existing methods and software,

such as ProbABEL, Sandwich, and MMAP (see the analysis plan in the online supplement), or those that are straightforward to implement using these tools.

Testing the significance of the SNP and the SNP*E interaction effects: In Model 1, the focus is on the test of the interaction effect and the joint effects of the SNP and the interaction. The interaction effect (β_{GE}) is evaluated using a 1 degree of freedom (df) Wald test. The effects of both SNP (β_G) and interaction (β_{GE}) are tested jointly, using a 2 df Wald test[25]. In model 2, which does not include E or SNP*E terms, β_G is the familiar main effect of the SNP which is tested using a 1-df Wald test. A 1 df Wald test is also used in model 3 for evaluating the SNP effect (β_G) in the presence of E, which may be referred to as the refined SNP effect or E-adjusted SNP effect or context-dependent SNP effect. In all cases, we will use the "robust" Wald tests by using robust estimates of the standard errors (SEs) and covariances to protect against misspecification of the mean model[27,28]. When the SNP effect is weak and the SNP*E interaction effect is moderate, the joint 2 df test has been shown to be more powerful than either the 1 df test of the SNP effect or the 1 df test of the interaction effect alone[25]. The increase in power for the 2 df over either 1 df test can be particularly dramatic, especially when the type I error rate is controlled at very low levels (e.g., $5x10^{-8}$) as in this project[29].

Analyses needed from each cohort: Each cohort carries out a genome-wide analysis of the SNP and SNP*E interaction effects and provides estimates of betas, robust estimates of the corresponding standard errors (SEs) and covariance, and p-values from the joint 2 df test separately for each ancestry group. Because the model is based on a standard regression framework, software to compute the relevant statistics is widely available. For studies of unrelated individuals, standard commands and the R sandwich package[30] implement bivariate robust covariance estimates for SNP-specific analyses. To implement the analyses for all SNPs, the R interface in PLINK[31] may be used; ProbABEL[32] also provides appropriate utilities. For family studies in which relatedness must be taken into account, programs such as GenABEL/MixABEL[33] and MMAP (O'Connell, unpublished; personal communication) implement mixed models that allow for relatedness. All cohorts analyze their data using these methods/software following a pre-specified Analysis Plan, that spells out all analysis steps in detail. They then upload results to a secure server.

Meta-analysis for combining results across studies: To combine estimates of the betas and their corresponding 2x2 covariance matrix provided by each cohort, we use the joint meta-analysis method developed by Manning et al[26] who modified METAL[34] to handle this joint 2 df meta-analysis. The joint meta-analysis provides inference on the SNP and SNP*E interaction effect pooled across all cohorts. Manning[7] used this approach and demonstrated power enhancement for detecting

interactions. We use the modified METAL for the joint meta-analysis and use METAL for carrying out meta-analysis of the 1 df analyses (interaction effect in model 1, main effect in model 2, and refined SNP effect in model 3). We use a genome-wide significance threshold of $5x10^{-8}$ for identifying significant results and use 10^{-6} for identifying suggestive results.

Quality Control: Quality assurance in imputations is emphasized by preparing very detailed analysis plans with step by step instructions for preparing and analyzing data, and formatting results for uploading (see the Education-Lipids analysis plan included in the supplementary materials for details). Extensive QC measures are used for processing all study-specific results centrally by each project team, at two levels. "Study-level" QC involved reviewing and harmonization of each individual result file separately. "Meta-level" QC involved reviewing and harmonizing results files across all available discovery cohorts for a single analysis (e.g., comparing summary statistics across all SBP-Current Smoking-Model1 discovery cohorts). Some of the specifics are discussed as part of the supplementary materials. QC was performed using customized EasyQC scripts that provide a wide variety of QC checks for GWAS results[35].

Analysis of interactions involving rare variants: While the joint test of SNP and SNP*E is applicable to analysis of rare variants, its power for testing individual rare variants is limited primarily due to their low frequency. Burden tests[36-39] collapse all rare variants in a genomic region (typically a gene) into a single burden variable (essentially a "mega variant", giving each subjects' total dosage across a gene) and regress the phenotype on the burden variable to test for the combined effects of all rare variants in the region/gene. We use the burden test that collapses rare variants with MAF < 0.01 in the genomic region (gene) into a single burden variable (i.e., T1). We apply the 2 df test directly to each T1 burden variable. Since MAF varies across cohorts, the pooled MAF is computed by the CC as a weighted average of MAFs from all cohorts. Each cohort creates the T1 burden variables by collapsing variants within the genomic regions using the pooled MAF < 0.01, instead of the cohort-specific MAF. Analysis uses the 2 df joint test. We then perform meta-analysis of these T1-based results, similar to the meta-analysis of results from common variants described earlier but now with as many T1 burden variables as the number of genomic regions. To assess the significance for the analysis of rare variants, we will use a Bonferroni-corrected significance threshold ($\alpha = 0.05/N_b$ where N_b= number of burden variables). The CHARGE consortium has provided detailed analysis guidelines for exome chip data and the CC has used some of these rare variant methods[40-43].

Analysis of Stage 1 and Stage 2 Results: Primary publications resulting from the various analyses in stages 1 and 2 are pursuing two approaches shown in **Figure 1:** combined analysis of Stages 1 and 2 and traditional discovery/replication.

Combined analysis of stages 1 and 2: This approach can be more powerful than other approaches[44]. For a given combination of phenotype and lifestyle, all significant and suggestive results (with $\alpha = 10^{-6}$) from stage 1 cohorts and the corresponding results from stage 2 cohorts are pooled through meta-analysis (first within each stage and then meta-analyzing the two stage-specific meta-analyses) separately by ancestry. A significance threshold of $\alpha = 5\times10^{-8}$ is used to identify significant results from the combined stage 1 and 2 meta-analysis results. Finally, all ancestry-specific meta-analyses are meta-analyzed as an approximate trans-ancestry analysis for identifying additional associations (if any) that are missed by ancestry-specific analyses.

Traditional Discovery/ Replication Analysis: In this approach, all genome-wide significant results are identified from stage 1 results only, separately by ancestry, using a significance threshold of $\alpha = 5\times10^{-8}$. Stage 2 results are then used to formally replicate the stage 1 findings, using appropriate Bonferroni correction such as 0.05 divided by the number of independent novel loci discovered in stage 1. Variants that are suggestive but not significant in Stage 1 are only considered in the combined analysis approach.

The combined approach is more powerful than the traditional approach. However, the traditional approach can identify additional novel validated loci missed by the combined approach (as shown most recently using a slight variation of this approach[45]). This justifies using both approaches. If only one approach were to be used, the combined one is the method of choice.

Statistical power for detecting associations: With the overall sample size used, this study is well powered for identifying novel discoveries even with moderately small effect sizes. To demonstrate this, we illustrate the sample sizes required to achieve at least 80% power to identify the genetic (G) effect and the GxE interaction effect using the 2 df joint test for a range of model parameters. We used QUANTO[46], which computes power and sample size for both disease and quantitative trait studies of genes (G), environment factors (E), and GxE interactions. For our study of quantitative traits, the required sample sizes depend on the proportions of variance explained by the G (R2G), the lifestyle factor (R2E) and their interaction effect (R2GE). A wide range of R2E values yielded similar results, and therefore we fixed R2E = 0.1% and examined the effect of varying the other 2 parameters. Although low frequency variants explain large proportions of

variance in some cases[47], we limited this investigation to lower R2G values of 0.01%, 0.02%, 0.05%, and 0.1% because most variants identified through GWAS have much smaller effect sizes. **Figure 2** shows the sample sizes required for a range of R2GE values corresponding to each of the four values for R2G using a significance threshold of $5x10^{-8}$. These values are smaller than what we found in our preliminary studies (not reported), suggesting that our power estimates may be conservative.

The sample sizes should be more than adequate for 80% power in EA and AA using Stage 1 samples alone, so long as the SNP effect is not very small (e.g., R2G > 0.05%). In fact, for R2G = 0.05%, significance level of $5x10^{-8}$, and the stage 1 sample sizes shown in Table 1, the minimum detectable R2GE at 80% power are < 0.01%, 0.11%, 0.44%, and 0.27% for EA, AA, HA, and AS, respectively. When Stages 1 and 2 are combined, even smaller effect sizes are detectable (although the exact calculations are complex since Stage 2 studies did not carry out genome-wide interaction analyses). In any case, the combined sample size of stages 1 and 2 appears well poised for powerful discoveries even with smaller effect sizes than assumed in these estimates.

Discussion

Current Status and Anticipated Benefits: Our study has made considerable progress to date. Four projects have completed all analyses in stages 1 and 2 and are processing the final results for publications (Smoking-BP, Smoking-Lipids, Alcohol-BP, and Alcohol-Lipids). In addition, three other projects (Education-BP, Education-Lipids, and PA-Lipids) have completed stage 1 analyses and are in advanced stages of preparing results for stage 2 analyses, and two projects (Psychosocial-BP and Psychosocial-Lipids) have completed stage 1 analyses and are undergoing extensive QC. Still more projects are getting underway. We believe that these projects will make major contributions to the genetic dissection of cardiovascular traits and that the GxE analysis can help improve understanding of the mechanisms underlying the novel as well as known loci which have been identified previously through main effects.

What are the unique benefits of our approach? How critical is the consideration of lifestyle and interactions (models 1 and 3)? Emerging results indicate that a large proportion of novel findings originate from models 1 and 3, i.e. results that would be missed by limiting analyses to main effects (model 2). This suggests that inclusion of the lifestyle context and/or gene-lifestyle interaction is important for identifying novel signals.

Collaboration levels are unprecedented: In an area where direct competition among study groups was the norm until about a decade ago, collaborative GWAS-based consortia such as CHARGE represent an innovative model for research. Through working together, the contributing studies have achieved much more than they could have working alone. The Gene-Lifestyle Interactions Working Group takes this model further, assembling 610,475 subjects in 124 cohorts. While the collaborative nature of the work requires some compromises (e.g. using standard software, and meta-analysis of relatively simple analyses) the results should substantially deepen what has already been learned from GWAS.

Acknowledgments and Sources of Funding

This multi-ancestry study of gene-lifestyle interactions is sponsored by R01HL118305 from the National Heart, Lung, and Blood Institute (NHLBI), national Institutes of Health (NIH). The CHARGE infrastructure on which this study is based is also sponsored by another NHLBI grant HL105756. The authors wish to thank several investigators in the Gene-Lifestyle Interactions Working Group (WG) for their important contributions to the WG and the various projects, notably Hugues Aschard, Sharon Kardia, Ruth Loos, Alisa Manning, Jeff O'Connell, Michael Province, Patricia Peyser, Jerome Rotter, Xiaofeng Zhu, among others. The full list of the WG members can be found at: http://depts.washington.edu/chargeco/wiki/Gene-Lifestyle_Interactions. The authors also wish to thank Matthew Brown for his critical help as part of the Data Coordinating Center in preparing some of the materials for this publication. Study descriptions and study-specific acknowledgments are included in the Supplemental Materials along with an example analysis plan. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Disclosures

Dr. Psaty serves on the DSMB of a clinical trial funded by Zoll LifeCor and on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson.

Other authors: No conflicts

References

1. Psaty BM, O'Donnell CJ, Gudnason V, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet. 2009, Feb;2(1):73-80. doi: 10.1161/CIRCGENETICS.108.829747. PMID: 20031568.

2. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature, 2009 Oct 8;461(7265):747-53. doi: 10.1038/nature08494. PMID: 19812666.

3. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet 2012; 90(1):7-24. PMID: 22243964

4. Murphy SL, Xu J, Kochanek KD. *Deaths: Preliminary Data for 2010*, in *National Vital Statistics Reports*2012, National Center for Health Statistics: Hyattsville, MD.

5. Roger VL, Go AS, Lloyd-Jones DM, et al. Heart disease and stroke statistics--2012 update: a report from the American Heart Association. *Circulation.* 2012; 125: e2-e220.

6. Zheng JS, Arnett DK, Lee YC, et al. Genome-wide contribution of genotype by environment interaction to variation of diabetes-related traits. PloS One 2013; 8(10):e77442. PMID: 24204828

7. Manning AK, Hivert MF, Scott RA, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet.* 2012; 44: 659-69

8. Hutter CM, Mechanic LE, Chatterjee N, et al. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. Genet Epidemiol. 2013. 37(7):643-57. PMID: 24123198

9. Sung YJ, de las Fuentes L, Schwander KL, et al. Gene-smoking interactions identify several novel blood pressure loci in the Framingham Heart Study. *Am J Hypertens.* 2015, 28(3):343-354. PMID: 25189868.

10. Montasser ME, Shimmin LC, Hanis CL, et al. Gene by smoking interaction in hypertension: identification of a major quantitative trait locus on chromosome 15q for systolic blood pressure in Mexican-Americans. *J Hypertens.* 2009; 27: 491-501.

11. Higashibata T, Hamajima N, Naito M, et al. eNOS genotype modifies the effect of leisure-time physical activity on serum triglyceride levels in a Japanese population. *Lipids Health Dis.* 2012; 11: 150.

12. Grarup N, Andreasen CH, Andersen MK, et al. The -250G>A promoter variant in hepatic lipase associates with elevated fasting serum high-density lipoprotein cholesterol modulated by interaction with physical activity in a study of 16,156 Danish subjects. *J Clin Endocrinol Metab.* 2008; 93: 2294-9.

13. Parnell LD, Blokker BA, Dashti HS, et al. CardioGxE, a catalog of gene-environment interactions for cardiometabolic traits. BioData Min. 2014 Oct 26;7:21. doi: 10.1186/1756-0381-7-21. eCollection 2014. PMID: 25368670

14. Hunter DJ. Gene-environment interactions in human diseases. *Nat Rev Genet.* 2005; 6: 287-98.

15. Murcray CE, Lewinger JP, Gauderman WJ. Gene-Environment Interaction in Genome-Wide Association Studies*. American Journal of Epidemiology.* 2009; 169: 219-226.

16. Thomas D. Gene-environment-wide association studies: emerging approaches*. Nature Reviews Genetics.* 2010; 11: 259-272.

17. Taylor JY, Maddox R, Wu CY. Genetic and Environmental Risks for High Blood Pressure Among African American Mothers and Daughters*. Biological Research for Nursing.* 2009; 11: 53-65

18. Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011; 470: 204-13.

19. Randall JC, Winkler TW, Kutalik Z, et al. PLoS Genet. 2013 Jun;9(6):e1003500. doi: 10.1371/journal.pgen.1003500. Epub 2013 Jun 6. PMID: 23754948

20. Aschard H, Hancock DB, London SJ, Kraft P. Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. Hum Hered. 2010;70(4):292-300. PMID: 21293137

21. Sung YJ, Winkler TW, Manning AK, et al. An Empirical Comparison of Joint and Stratified Frameworks for Studying G × E Interactions: Systolic Blood Pressure and Smoking in the CHARGE Gene-Lifestyle Interactions Working Group. *Genet Epidemiol*. 2016; Jul;40(5):404-15. doi: 10.1002/gepi.21978. Epub 2016 May 27.

22. VanderWeele TJ, Yi-An Ko, and Bhramar Mukherjee. Environmental Confounding in Gene-Environment Interaction Studies. *Am J Epidemiol*. 2013;178(1):144–152.

23. Dudbridge F and Fletcher O. Gene-Environment Dependence Creates Spurious Gene-Environment Interaction. *Amer J Hum Genet*, 95, 2014.

24. Keller MC. Gene x Environment Interaction Studies Have Not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution. *Biol Psychiatry* 2014; 75: 18-24.

25. Kraft P, Yen YC, Stram DO, et al. Exploiting gene-environment interaction to detect genetic associations. Hum Hered. 2007;63(2):111-9. Epub 2007 Feb 2. PMID: 17283440

26. Manning AK, LaValley M, Liu CT, et al. Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP x environment regression coefficients*. Genet Epidemiol.* 2011; 35: 11-8. PMID: 21181894

27. Tchetgen EJ, Kraft P. On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified*. Epidemiology.* 2011; 22: 257-61.

28. Voorman A, Lumley T, McKnight B, Rice K. Behavior of QQ-plots and genomic control in studies of gene-environment interaction*. PLoS One.* 2011; 6: e19416.
29. Morris N, Elston R. A Note on Comparing the Power of Test Statistics at Low Significance Levels*. American Statistician.* 2011; 65: 164-166.

30. Zeileis A. Object-oriented computation of sandwich estimators. Journal of Statistical Software. 2006; 16.

31. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses*. Am J Hum Genet.* 2007; 81: 559-75.

32. Aulchenko YS, Struchalin MV, van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics.* 2010; 11: 134.

33. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics.* 2007; 23: 1294-6.

34. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics.* 2010; 26: 2190-1.

35. Winkler TW, Day FR, Croteau-Chonka DC, et al. Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* 2014; 9(5):1192-212.

36. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res.* 2007; 615: 28-56.

37. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008; 83: 311-21.

38. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5: e1000384.

39. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol.* 2010; 34: 188-93.

40. Sung YJ, Rice TK, Rao DC. Application of collapsing methods for continuous traits to the Genetic Analysis Workshop 17 exome sequence data. *BMC Proc.* 2011; 5 Suppl 9: S121.

41. Sun YV, Sung YJ, Tintle N, Ziegler A. Identification of genetic association of multiple rare variants using collapsing methods. *Genet Epidemiol.* 2011; 35 Suppl 1: S101-6.

42. Malllaney C, Sung YJ. Rare variant analysis of blood pressure phenotypes in the Genetic Analysis Workshop 18 whole genome sequencing data using SKAT. *BMC Proc.* 2013.

43. Sung YJ, Basson J, Rao DC. Whole genome sequence analysis of the simulated SBP in Genetic Analysis Workshop 18 family data: Long term average and collapsing methods. *BMC Proc.* 2013.

44. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet.* 2006; Feb;38(2):209-13. PMID: 16415888.

45. Surendran P, Drenos F, Young R, et al. Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. Nat Genet. 2016; Oct;48(10):1151-61. doi: 10.1038/ng.3654. Epub 2016 Sep 12. PMID: 27618447

46. Gauderman W, Morrison J. *QUANTO 1:1: A computer program for power and sample size calculations for genetic-epidemiology studies,* 2006.

47. Bowden DW, An SS, Palmer ND, et al. Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS Family Study. *Hum Mol Genet.* 2010; 19: 4112-20.

Table 1. Studies and ancestry groups participating in Stage 1 (Genome-wide discovery)

No	Study/ Cohort	Type of Study	European Ancestry	African Ancestry	Hispanic Ancestry	Asians	Brazilian Admixed
1	AGES	Population study of GxE in elderly	2,410	-	-	-	-
2	ARIC	Population-based study of Atherosclerosis	9,465	2,862	-	-	-
3	Baependi	Family-based study of CVD traits	-	-	-	-	873
4	CARDIA	Population-based study of CVD traits	1,649	945	-	-	-
5	CHS	Population-based study of CVD traits	2,975	734	-	-	-
6	CROATIA	Population-based study of Croatians: Vis	483	-	-	-	-
		Population-based study of Croatians: Korcula	456	-	-	-	-
7	Fam HS	Family study of CVD related traits	3,683	617	-	-	-
8	FHS	Longitudinal family study of CVD traits	8,195	-	-	-	-
9	GENOA	Sibling study of Atherosclerosis and HT	1,064	941	-	-	-
10	GenSalt	Family study of salt sensitivity	-	-	-	1,835	-
11	GENSCOT	Population-based study in Scotland	6,439	-	-	-	-
12	GOLDN	Family-based study of HT & CVD traits	820	-	-	-	-
13	HANDLS	Diversity study of aging and CVD traits	-	903	-	-	-
14	Health ABC	Study of health, aging and body comp	1,663	1,136	-	-	-
15	HERITAGE	Fam study of responses to exercise	499	-	-	-	-
16	HUFS	Family study of hypertension in AA	-	1,686	-	-	-
17	HyperGEN	Family-based study of HT & CVD traits	1,251	1,240	-	-	-
18	JHS	Population-based study of CVD traits	-	2,134	-	-	-
19	Maywood-L	Population study of CVD traits in AA	-	75	-	-	-
20	Maywood-N	Study of CVD traits in Nigerians	-	1,229	-	-	-
21	MESA	Family-based study of Atherosclerosis	2,591	1,594	1,455	748	-
22	Mt. Sinai IPM	Hospital-based / Biobank patients	1,480	3,101	3,973	-	-
23	NEO	Population-based study of obesity related traits	5,735	-	-	-	-
24	Pelotas	Population-based birth cohort in Brazil	-	-	-	-	3,541
25	RS	Rotterdam study of CVD traits: RS1	4,990	-	-	-	-
		Rotterdam study of CVD traits: RS2	1,998	-	-	-	-
		Rotterdam study of CVD traits: RS3	2,966	-	-	-	-
		Rotterdam family study of CVD traits: RS-ERF	2,491	-	-	-	-
26	SCES	Singapore Chinese eye study	-	-	-	1,848	-
27	SCHS	Singapore Chinese Health Study: Cases	-	-	-	674	-
		Singapore Chinese Health Study: Controls	-	-	-	1,218	-
28	SiMES	Singapore Malay eye study	-	-	-	2,531	-
29	SINDI	Singapore Indian eye study	-	-	-	2,491	-
30	SP2	Singapore 2: 1M	-	-	-	949	-
		Singapore 2: 610	-	-	-	1,144	-
31	WGHS	Popn-based; genomics; women's health	22,983	-	-	-	-
32	WHI	Popn-based study of women's health	-	7,919	3,377	-	-
		Popn-based study of women's health: GARNET	4,423	-	-	-	-
		Popn-based study of women's health: WHIMS	5,202	-	-	-	-
TOTALS			**95,911**	**27,116**	**8,805**	**13,438**	**4,414**

Note: Sample sizes may vary across phenotype-exposure combinations due to missing data.

Table 2. Studies and ancestry groups participating in Stage 2 (Focused Discovery/Replication)

No	Study/ Cohort	Type of Study	European Ancestry	African Ancestry	Hispanic Ancestry	Asians	Brazilian Admixed
1	AADHS	Case-Control study of diabetes in AAs	-	584	-	-	-
2	ASCOT-SC	Population-based study of cardiac outcomes	2,389	-	-	-	-
3	BBJ	Population-based biobank in Japan	-	-	-	126,413	-
4	BES	Population-based study of eye disease:610	-	-	-	601	-
		Popn-based study of eye disease:OmniExpress	-	-	-	545	-
5	BRIGHT	Population-based study of hypertension	1,823	-	-	-	-
6	CAGE	Popn-based study of CVD traits: Amagasaki	-	-	-	952	-
7	CARL	Family-based study of auditory traits in Italy	462	-	-	-	-
8	CFS	Family-based study of sleep apnea in AA	-	561	-	-	-
9	DESIR1	Epidemiological study on insulin resistance	697	-	-	-	-
10	DFTJ	Popn-based study of health and retirement	-	-	-	1,406	-
11	DHS	Family-based study of diabetes	1,173	-	-	-	-
12	DR's EXTRA	Unrelated study of exercise training	1,230	-	-	-	-
13	EGCUT	Popn-based biobank in Estonia:OmniExpress	5,937	-	-	-	-
		Popn-based biobank in Estonia:CoreExome	4,911	-	-	-	-
		Popn-based biobank in Estonia:Human370CNV	1,870	-	-	-	-
14	EPIC	Popn-based study of cancer/nutrition in Europe	20,458	-	-	-	-
15	Fenland	Popn-based study of metabolic traits: GWAS	1,345	-	-	-	-
		Popn-based study of metabolic traits: OMICS	8,471	-	-	-	-
16	FUSION	Case-Control Study of NIDDM:CASES	674	-	-	-	-
		Case-Control Study of NIDDM:CONTROLS	277	-	-	-	-
17	FVG	Family-based study of auditory traits in Italy	951	-	-	-	-
18	GeneSTAR	Family study of atherosclerosis risk	1,699	1,107	-	-	-
19	GLACIER	Population-based study of lobular cardinoma	5,909	-	-	-	-
20	GRAPHIC	Population-based study of arterial pressure	1,010	-	-	-	-
21	HRS	Population-based study of health & retirement	8,367	1,993	-	-	-
22	HyperGEN	Family-based study of HT & CVD traits:AXIOM	-	418	-	-	-
23	InterAct	Case-contrl study of T2DM:CoreExome:CASES	3,996	-	-	-	-
		CC study of T2DM:CoreExome:SUBCOHORT	6,405	-	-	-	-
		Case-control study of T2DM:GWAS:CASES	2,793	-	-	-	-
		CC study of T2DM:GWAS:SUBCOHORT	3,188	-	-	-	-
24	IRAS	Popn-based study of atherosclerosis:IRASC	-	-	185	-	-
		Family-based study of atherosclerosis:IRASFS	-	-	957	-	-
25	JUPITER	Population-based study of lipids and statin use	8,400	1,606	-	-	-
26	KORA	Population-based German research cohort:S3	3,095	-	-	-	-
		Population-based German research cohort:S4	3,770	-	-	-	-
27	LBC	Lothian Birth Cohort study:1921	511	-	-	-	-
		Lothian Birth Cohort study:1936	996	-	-	-	-
28	Lifelines	Biobank cohort in the Netherlands	12,323	-	-	-	-
29	LLFS	Family-based study on aging	3,133	-	-	-	-
30	LOLIPOP	London Population study of CVD traits: EW610	927	-	-	-	-
		London Population study of CVD traits: EWA	582	-	-	-	-
		London Population study of CVD traits: EWP	644	-	-	-	-
		London Population study of CVD traits: IA317	-	-	-	2,059	-
		London CC study of CVD traits: IA610-case	-	-	-	2,791	-
		London CC study of CVD traits: IA610-ctrl	-	-	-	3,757	-

		London Population study of CVD traits: IAP	-	-	-	501	-
		London Popn study of CVD traits: OmniEE	-	-	-	899	-
31	LOYOLA	Population-based Jamaican cohort of BP:GXE	-	612	-	-	-
		Population-based Jamaican health cohort:SPT	-	904	-	-	-
32	METSIM	Men-only unrelated study; metabolic syndrome	8,353	-	-	-	-
33	OBA	Unrelated French obese cases	669	-	-	-	-
34	PROCARDIS	Case-control study of CAD:Cases	5,651	-	-	-	-
		Case-control study of CAD:Controls	1,668	-	-	-	-
35	RHS	Popn-based cohort of metabolic syndrome	-	-	-	2,468	-
36	SHEEP	Case-control study of CVD traits:Cases	1,165	-	-	-	-
		Case-control study of CVD traits:Controls	1,528	-	-	-	-
37	SHIP	Population-based health study:0 Cohort	4,046	-	-	-	-
		Population-based health study:Trend Cohort	982	-	-	-	-
38	SMWHS	Population-based men/women health study	-	-	-	3,862	-
39	SOL	Hispanic community health study	-	-	12,380	-	-
40	TAICHI	Popn-based study of atherosclerosis:Zhonghua	-	-	-	1,505	-
41	THRV	Population-based Taiwan study of hypertension	-	-	-	287	-
42	TRAILS	Population-based study of adolescents	1,266	-	-	-	-
43	TUDR	Population-based study of diabetes	-	-	-	886	-
44	TWINGENE	Family-based study of twins in Sweden	5,358	-	-	-	-
45	UK Biobank	Population-based Biobank in the UK	137,426	-	-	-	-
46	YFS	Population-based CV study of young adults	2,024	-	-	-	-
TOTALS			**290,552**	**7,785**	**13,522**	**148,932**	**0**

Note: Sample sizes may vary across phenotype-exposure combinations due to missing data.

Figure legends

Figure 1. Overall flow of analyses. Combined analysis leverages the full power of Stages 1 and 2. The traditional discovery and replication approach identifies additional loci missed by the combined approach. Both approaches can be used for maximizing discovery.

Figure 2. Sample sizes needed for 80% power using the 2 df joint test. Sample size (Y-axis) is plotted as a function of the percent variance explained by the interaction (R2GE; X-axis), for each of 4 different values of the percent variance explained by the genetic effect (R2G); that due to the lifestyle factor (R2E) is fixed at 0.1% (see the text).

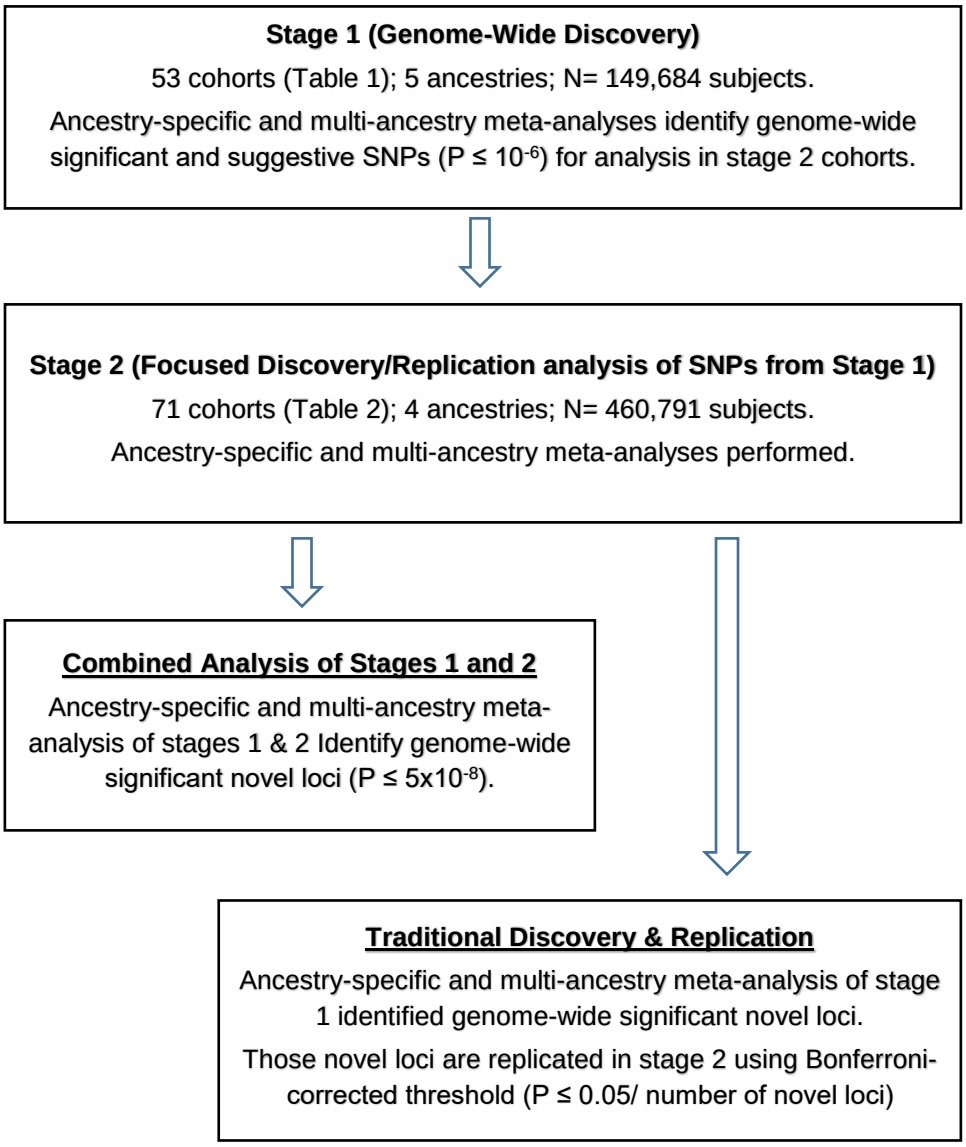

Figure 1. Overall flow of analyses. Combined analysis leverages the full power of Stages 1 and 2. The traditional discovery and replication approach identifies additional loci missed by the combined approach. Both approaches can be used for maximizing discovery.

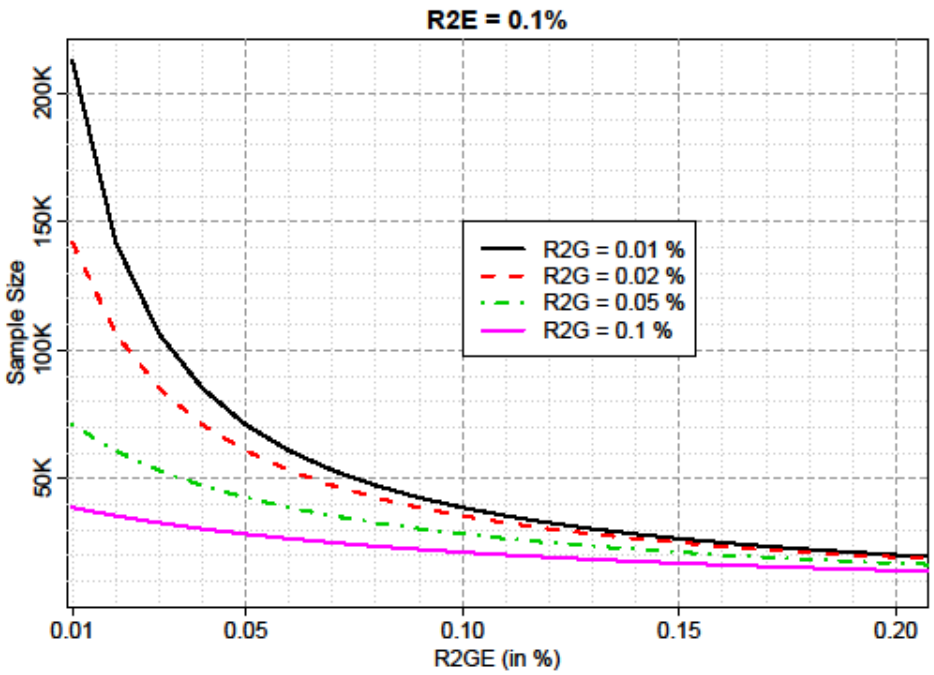

Figure 2. Sample sizes needed for 80% power using the 2 df joint test. Sample size (Y-axis) is plotted as a function of the percent variance explained by the interaction (R2GE; X-axis), for each of 4 different values of the percent variance explained by the genetic effect (R2G); that due to the lifestyle factor (R2E) is fixed at 0.1% (see the text).