An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene-environment interactions.

Yun Ju Sung[1], Karen Schwander[1], Donna K. Arnett[2], Sharon L.R. Kardia[3], Tuomo Rankinen[4], Claude Bouchard[4], Eric Boerwinkle[5], Steven C. Hunt[6], DC Rao[1]

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

[2]Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL

[3]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI

[4]Human Genomics Laboratory, Pennington Biomedical Research Center, Baton Rouge, LA

[5]Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston TX

[6]Cardiovascular Genetics, University of Utah, Salt Lake City, UT

Corresponding author: Dr. Yun Ju Sung

Division of Biostatistics

Washington University School of Medicine

660 South Euclid Avenue, Campus Box 8067

St. Louis, MO 63110-1093

Email: yunju@wubios.wustl.edu

Phone: (314) 362-0053; Fax: (314) 362-2693

Abstract

Meta-analysis combining results from multiple studies is a standard practice in GWAS. For genetic main effects, meta-analysis has been shown to provide comparable results as mega-analysis that jointly analyzes the pooled data from the available studies. Gene-environment interaction (GEI) studies are an important component of genetic epidemiology research since they can explain a part of the missing heritability, elucidate the biological networks underlying disease risk, and identify individuals at high risk for disease. However, it is not known whether meta- and mega-analyses of interactions also yield comparable results. In this study, we investigate whether both approaches provide comparable results for identifying interaction effects using empirical data from 4 studies: the Framingham Heart Study, GENOA, HERITAGE and HyperGEN. We performed meta-analysis of cohort-specific results and mega-analysis by analyzing the pooled data from all 4 studies. We used the standard 1 degree of freedom (df) test of main effect only, the 1 df test of the interaction effect (in the presence of main effect), and the joint 2 df test of main and interaction effects. We found that the results from meta- and mega-analyses were highly consistent for all three tests. The correlation between -log (p) values from the two analyses was 0.89 for the 1 df main effect, 0.90 for the 1 df interaction test, and 0.91 for the joint 2 df test. Although mega-analysis provided slightly better results as expected, both yielded very similar results for the most promising SNPs. Moreover, mega-analysis is not always feasible especially in very large and diverse consortia since pooling of raw data may be limited by the terms of the informed consent. Our study illustrates that meta-analysis can be an effective approach also for identifying interactions in very large consortia without losing appreciable power.

Keywords: gene-environment interactions (GEI), meta-analysis, mega-analysis

Introduction

Genome wide association studies (GWAS) have identified hundreds of common genetic variants associated with many common complex disease traits (http://www.genome.gov). As of February 19, 2013, a total of 3,584 unique SNPs have been significantly associated with 500 traits through GWAS. The confidence and funding invested in GWAS approaches have led to unprecedented levels of collaboration, spawning a number of highly productive consortia such as the CHARGE (Cohorts for Heart and Aging Research in Genetic Epidemiology),[1] GIANT (Genetic Investigation of Anthropometric Traits),[2] and ICBP (International Consortium of Blood Pressure).[3] It is standard practice in these consortia to perform meta-analysis that combines the results from multiple studies for the same phenotype of interest.[4] Meta-analysis provides an efficient and practical strategy for overcoming the limitations of power that can compromise any individual study.

To increase sample sizes, one could obtain the individual level data from studies able to participate and jointly analyze the pooled individual-level data (such analysis is referred to here as mega-analysis). However, meta-analysis that combines summary results has several advantages as opposed to mega-analysis. First, obtaining individual leve data is challenging and limited by the terms of the informed consent within each study. Second, integration of very large genotype and phenotypic data sets from different studies is time-consuming and poses additional challenges with data management, storage, and harmonization issues. Third, meta-analysis allows for analyses of individual studies to account for local population substructure, relationships among subjects, study-specific covariates, and other ascertainment-related issues which may be optimally dealt with within each study.

Although GWAS have successfully identified hundreds of new susceptibility loci for complex traits, these identified loci have very subtle effects, thus explaining only a small fraction of the heritability of most complex traits.[5] It is increasingly recognized that the near-exclusive focus on main effects has become a barrier to the identification of additional genes underlying these disease traits. Increasingly greater emphasis is being placed in recent years on GxE interaction analyses, also relying on meta-analysis in large consortia. The identification of gene-environment (GxE) interactions (GEI) is important for many reasons.[6; 7] GxE interactions or more complex pathways involving multiple genes and environments can explain a fraction of missing heritability.[5; 8-10] They can further elucidate the

biological networks underlying complex disease risk and enable "profiling" of individuals at highest risk for disease. Although identifying GxE interactions is important, it requires larger sample sizes than those needed to identify genetic main effects alone.[11] Therefore, it is important to have very large sample sizes. Meta-analysis which has been useful for identifying genetic main effects will be even more useful for identifying GxE interactions.

Although mega-analysis is arguably more powerful than meta-analysis of summary results, in so far as analysis of main effects is concerned, it has been shown that both approaches yield highly comparable results.[12] However, whether meta- and mega-analyses yield comparable results also for GxE interaction effects is unknown. Therefore, we investigated this comparison using empirical data.

Methods

Studies

In this paper, we used four studies: the Framingham Heart Study (as obtained through the dbGaP), GENOA, HERITAGE and HyperGEN. All four studies obtained informed consent from participants and approval from the appropriate institutional review boards.

The Framingham Heart Study (FHS) is a longitudinal family-based study for identifying the factors that contribute to cardiovascular disease, sponsored by the National Heart, Lung, and Blood Institute (NHLBI). FHS began in 1948 with the recruitment of an original cohort of 5,209 men and women who were 28 to 62 years of age at entry.[13] Clinic examinations took place approximately every 2 years. In 1971, a second generation of study participants, 5,124 children and spouses of children of the original cohort, were enrolled.[14] Clinic examinations took place approximately every 4 years. Enrollment of the third generation cohort of 4,095 children of offspring cohort participants began in 2002.[15] In this analysis we used cross-sectional data on 6,686 subjects involving all 3 generations who have GWAS data.

The Genetic Epidemiology Network of Arteriopathy (GENOA) is a family-based study of hypertension and diabetes (e.g., see Daniels et al[16]). GENOA recruited sibships with at least two hypertensive siblings and any additional available siblings, regardless of hypertension status. In total, it has collected data on over 5,000 Caucasian-American, African-American subjects, and Hispanic-American subjects, from

Minnesota, Mississippi, and Texas, respectively. GENOA subjects were brought in for a complete physical exam and questionnaire. We used the 1,420 Caucasian subjects with GWAS data in this analysis.

The HERITAGE (Health, risk factors, exercise training and genetics) Family Study was designed to evaluate the role of genetic and non-genetic factors in cardiovascular, metabolic, and hormonal responses to aerobic exercise training.[17] Extensive data, including body composition, cardiovascular risk factors, and lifestyle habits, were gathered on approximately 800 subjects in over 100 families, both before and after 20-weeks of supervised training. We used 469 Caucasian subjects with GWAS data in this analysis.

The Hypertension Genetic Epidemiology Network (HyperGEN) is a multicenter family-based study for investigating the genetic causes of hypertension and related conditions.[18] HyperGEN recruited African American and Caucasian participants at five field centers. It recruited sibships with at least two hypertensive siblings, unmedicated offspring of these hypertensive siblings, and an age-matched random sample. Subjects were brought into the clinic for a one day exam, and data was collected from questionnaires, a physical exam, and blood and urine samples. In total, HyperGEN has collected data on 2,471 Caucasian-American subjects and 2,300 African-American subjects. We used 1,216 Caucasian subjects with GWAS data in this analysis.

Statistical Analyses

To compare meta-analysis and mega-analysis in the context of GxE interactions, we analyzed systolic blood pressure (SBP) and used pack-years of smoking as an interacting covariate. In addition, we included age, sex, body mass index (BMI) and use of anti-hypertensive medications as covariates. Using data from these four studies, we analyzed each study data separately and then meta-analyzed the results using METAL.[19] We also combined the raw phenotypic and GWAS data from all 4 studies and analyzed the pooled data (mega-analysis). For this pooled data set, we included study-specific intercepts. Our analysis was restricted to SNPs on chromosomes 16 and 18, for which the FHS data yielded significant results.

To compare meta-analysis and mega-analysis, we considered three analysis options. First, as most GWASs use a main-effect-only analysis, we also performed a main-effect-only analysis. The expected response trait (Y) has the regression model

$$E[Y] = \lambda_0 + \lambda_e E + \lambda_g G,$$

where λ_e is the environmental main effect and λ_g is the genetic main effect. We used the Wald test statistic that follows a chi-squared distribution with 1 degrees of freedom (df) under the hypothesis H_0: $\lambda_g=0$ (i.e., testing for the genetic main effect). Second, we followed the standard approach to identify GxE interactions. The regression model is generalized to

$$E[Y] = \alpha + \beta_e E + \beta_g G + \beta_{ge} GE,$$

where β_e and β_g respectively are the environmental and genetic main effects and β_{ge} is their multiplicative interaction effect. We used the Wald test statistic that also follows a chi-squared distribution with 1 df under the H_0: $\beta_{ge}=0$ (i.e., testing for the GxE interaction effect in the presence of the genetic main effect). Third, we performed the test proposed by Kraft et al[20] that jointly tests the genetic main and GxE interaction effects using the same regression model as in the second analysis. We used a Wald test statistic that follows a chi-squared distribution with 2 df under the H_0: $\beta_g=\beta_{ge}=0$. This Wald test statistic is based on estimates of β_g and β_{ge} and their corresponding 2x2 covariance matrix. Because all four studies are family studies in which relatedness must be taken into account, we used ProbABEL[21] and GenABEL/MixABEL.[22]

To combine the association evidence from the four studies, we performed meta-analysis. We used the inverse-variance weighting method in METAL[19] by computing inverse-variance-weighted coefficients. It is straightforward to combine the 1 df results of main-effect-only analysis and the 1 df results of interaction effect. To combine the joint 2 df results (estimates of β_g, β_{ge} and their corresponding 2x2 covariance matrix) from these 4 studies, we used the joint meta-analysis method developed by Manning et al[23] who modified METAL[19] to handle this joint 2df meta-analysis.

Results

Table 1 presents select characteristics of the four studies. Roughly half of the subjects are female within each study. The subjects in HERITAGE tended to be younger, relatively leaner and low pack years with no anti-hypertensive medications (since hypertension was an exclusion criterion). In contrast, the subjects in GENOA tended to be older and have high pack years and high SBP with over 55% on anti-

hypertensive medications. The HyperGEN study has characteristics similar to GENOA, although not as marked. The FHS has intermediate values for SBP, medication use and pack years.

Results were compared between meta-analysis and mega-analysis for all three analysis options: main-effect-only analysis, 1 df interaction effect in the presence of the main effect, and 2 df joint tests of main and interaction effects. As shown in Figure 1, the results [- log (p) values] between the meta- and mega-analyses correlated very well for all three analysis options. Surprisingly, the joint 2 df test provided the highest overall correlation (0.91). About 10% of the SNPs (shown in red) had heterogeneity p-values < 0.05; for these heterogeneous SNPs there was larger variation in the meta- and mega-analysis results resulting in a bit lower correlation between them (as expected). Similarly about 5% and 13% of SNPs had heterogeneity p-values from the 1 df main effect and 1 df interaction effect, respectively. Overall, although mega-analysis provided slightly better results as expected, both methods yielded very similar results for the most promising SNPs (with the smallest p values).

Discussion

Over the past several years, GWAS have identified hundreds of variants for complex traits, thereby accelerating the progress in the genetic dissection of complex human disease traits.[28] With the use of meta-analysis, the GIANT consortium identified multiple loci that modulate human body size and shape;[2] the CHARGE consortium identified multiple variants associated with several cardiovascular phenotypes.[1; 3] However, an important and sobering observation is that most of the identified (common) genetic variants have small main effects, explaining only small proportions of the trait variances. Increasingly greater emphasis is being placed in recent years on GxE interaction analyses, also relying on meta-analysis in large consortia.

While it is known that meta- and mega-analyses of main effects yield comparable results[12], the same is not known for GxE interactions, which forms the basis of our investigation here. In addition to the standard 1 degree of freedom test of interaction effect, we used the joint 2 df test of main and interaction effects by Kraft et al[20] and the joint meta-analysis by Manning et al.[23] When the genetic main effect is weak and GxE interaction effect is moderate, the joint 2 df test has been shown to be more powerful than either the 1 df test of the genetic main effect only or the 1 df test of the interaction effect alone.[20] The increase in power for the 2 df over either 1 df test can be dramatic when the type I error rate is

controlled at low levels as it is in genome-wide studies.[29] Because the joint 2 df test supplements standard marginal tests of genetic main effects with additional information from GxE interactions, the joint test can detect loci that are missed in marginal scans.[23; 30] Furthermore, Manning et al[23] developed the joint meta-analysis and modified METAL[19] to combine the joint 2 df tests across multiple studies. The joint meta-analysis provides inference on the genetic main effect and interaction effect pooled across all studies. Manning et al[30] used this approach and demonstrated power enhancement for detecting GxE interactions.

To combine the association evidence from the four studies, we used the inverse-variance weighting meta-analysis method in METAL[19] by computing inverse-variance-weighted coefficients. For the 1 df test of main effect and also for the 1 df test of interaction effect, we also performed another commonly-used approach that converts the direction of effect and p-value observed in each study into a signed Z-score; these Z-scores are then combined across studies in a weighted sum, with weights proportional to the square-root of the sample size for each study. The results between the inverse-variance weighting method and the sample size weighted signed Z score method were remarkably similar, providing a correlation of 0.98.

Meta-analysis of genetic main effects from multiple studies is known to yield results comparable to those from analysis of the combined raw data (mega-analysis). In most consortia, it is a standard practice to perform meta-analysis that combines results from multiple studies, since each study can analyze its own data optimally, by customizing data adjustments as needed. Moreover, even if the study investigators are able, mega-analysis is not readily feasible owing to several challenging issues including limitations of the informed consent for sharing raw data, data management and storage issues in dealing with huge data sets, and data harmonization may become a bottleneck. All these challenges render meta-analysis as the convenient and expedient method of choice. It is comforting to note that, as we illustrated using real data, results from meta- and mega-analyses appear to be highly consistent even for identifying GxE interaction effects. Therefore meta-analysis may continue to be leveraged for investigating interaction effects in very large consortia without losing much power.

Table 1. Summary statistics for the four studies

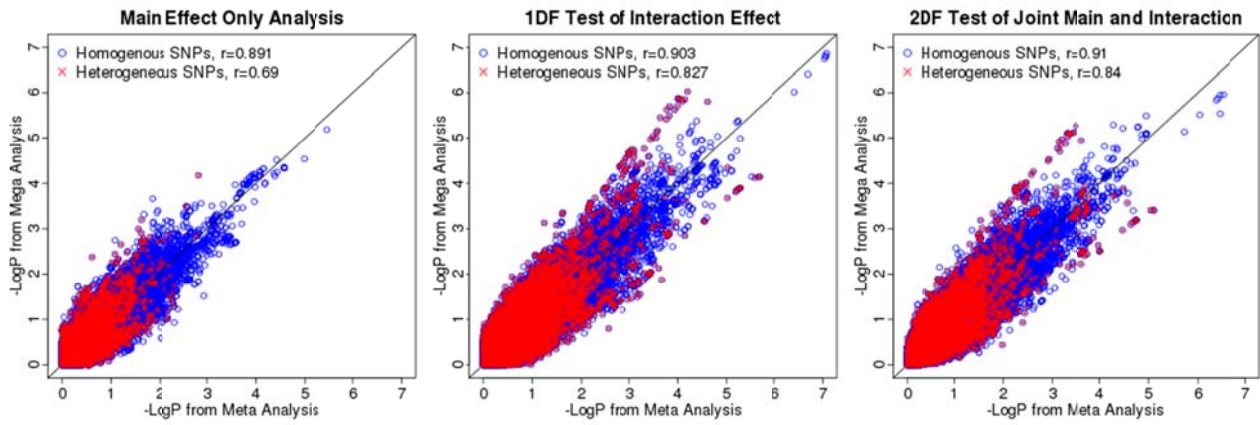|	FHS	GENOA	HERITAGE	HyperGEN
# Subjects	6,686	1,420	469	1,216
% Male	46.63	44.93	48.61	49.67
Age (average)	49.05 ± 13.57	54.67 ± 10.70	36.86 ± 14.24	49.49 ± 13.83
BMI (average)	27.44 ± 5.44	30.36 ± 6.24	26.04 ± 5.01	29.33 ± 5.94
% on Anti-HT Meds	19.04	55.70	0.00	45.97
SBP (average)	120.41 ± 16.40	133.53 ± 19.57	116.53 ± 11.15	123.19 ± 19.19
Pack Years (average)	9.87 ± 17.59	13.57 ± 21.57	4.99 ± 10.88	7.17 ± 16.26

Figure 1. Scatterplots showing mega-analysis versus meta-analysis for main-effect-only analysis, 1 df interaction effect in the presence of the main effect, and 2 df joint test of the main and interaction effects.

References

1. Psaty, B.M., O'Donnell, C.J., Gudnason, V., Lunetta, K.L., Folsom, A.R., Rotter, J.I., Uitterlinden, A.G., Harris, T.B., Witteman, J.C., Boerwinkle, E., et al. (2009). Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circulation Cardiovascular genetics 2, 73-80.
2. Willer, C.J., Speliotes, E.K., Loos, R.J., Li, S., Lindgren, C.M., Heid, I.M., Berndt, S.I., Elliott, A.L., Jackson, A.U., Lamina, C., et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nature genetics 41, 25-34.
3. Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C., Hwang, S.J., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature.
4. de Bakker, P.I., Ferreira, M.A., Jia, X., Neale, B.M., Raychaudhuri, S., and Voight, B.F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Human molecular genetics 17, R122-128.
5. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747-753.
6. Le Marchand, L., and Wilkens, L.R. (2008). Design considerations for genomic association studies: importance of gene-environment interactions. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 17, 263-267.
7. Thomas, D. (2010). Gene--environment-wide association studies: emerging approaches. Nature reviews Genetics 11, 259-272.
8. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am J Hum Genet 90, 7-24.
9. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews Genetics 11, 446-450.
10. McCarthy, M.I., and Hirschhorn, J.N. (2008). Genome-wide association studies: potential next steps on a genetic journey. Human molecular genetics 17, R156-165.
11. Thomas, D. (2010). Gene-environment-wide association studies: emerging approaches. Nature reviews Genetics 11, 259-272.
12. Lin, D.Y., and Zeng, D. (2010). Meta-Analysis of Genome-Wide Association Studies: No Efficiency Gain in Using Individual Participant Data. Genetic epidemiology 34, 60-66.
13. Dawber, T.R., Meadors, G.F., and Moore, F.E., Jr. (1951). Epidemiological approaches to heart disease: the Framingham Study. American journal of public health and the nation's health 41, 279-281.
14. Feinleib, M., Kannel, W.B., Garrison, R.J., McNamara, P.M., and Castelli, W.P. (1975). The Framingham Offspring Study. Design and preliminary data. Preventive medicine 4, 518-525.
15. Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B., Sr., Fox, C.S., Larson, M.G., Murabito, J.M., et al. (2007). The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. American journal of epidemiology 165, 1328-1335.

16. Daniels, P.R., Kardia, S.L., Hanis, C.L., Brown, C.A., Hutchinson, R., Boerwinkle, E., Turner, S.T., and Genetic Epidemiology Network of Arteriopathy, s. (2004). Familial aggregation of hypertension treatment and control in the Genetic Epidemiology Network of Arteriopathy (GENOA) study. The American journal of medicine 116, 676-681.
17. Bouchard, C., Leon, A.S., Rao, D.C., Skinner, J.S., Wilmore, J.H., and Gagnon, J. (1995). The HERITAGE family study. Aims, design, and measurement protocol. Medicine and science in sports and exercise 27, 721-729.
18. Williams, R.R., Rao, D.C., Ellison, R.C., Arnett, D.K., Heiss, G., Oberman, A., Eckfeldt, J.H., Leppert, M.F., Province, M.A., Mockrin, S.C., et al. (2000). NHLBI family blood pressure program: methodology and recruitment in the HyperGEN network. Hypertension genetic epidemiology network. Annals of epidemiology 10, 389-400.
19. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190-2191.
20. Kraft, P., Yen, Y.C., Stram, D.O., Morrison, J., and Gauderman, W.J. (2007). Exploiting gene-environment interaction to detect genetic associations. Human heredity 63, 111-119.
21. Aulchenko, Y.S., Struchalin, M.V., and van Duijn, C.M. (2010). ProbABEL package for genome-wide association analysis of imputed data. BMC bioinformatics 11, 134.
22. Aulchenko, Y.S., Ripke, S., Isaacs, A., and van Duijn, C.M. (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics 23, 1294-1296.
23. Manning, A.K., LaValley, M., Liu, C.T., Rice, K., An, P., Liu, Y., Miljkovic, I., Rasmussen-Torvik, L., Harris, T.B., Province, M.A., et al. (2011). Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP x environment regression coefficients. Genetic epidemiology 35, 11-18.
24. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science 291, 1304-1351.
25. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.
26. International HapMap, C. (2005). A haplotype map of the human genome. Nature 437, 1299-1320.
27. Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. Science 322, 881-888.
28. Hindorff, L.A., Junkins, H.A., Mehta, J.P., and Manolio, T.A. (2009). A Catalog of Published Genome-Wide Association Studies. In. (
29. Morris, N., and Elston, R. (2011). A Note on Comparing the Power of Test Statistics at Low Significance Levels. Am Stat 65, 164-166.
30. Manning, A.K., Hivert, M.F., Scott, R.A., Grimsby, J.L., Bouatia-Naji, N., Chen, H., Rybin, D., Liu, C.T., Bielak, L.F., Prokopenko, I., et al. (2012). A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet 44, 659-669.