The Use of Phenome-Wide Association Studies (PheWAS) for Exploration of Novel Genotype-Phenotype Relationships and Pleiotropy Discovery

S.A. Pendergrass,[1] K. Brown-Gentry,[1] S.M. Dudek,[1] E.S. Torstenson,[1] J.L. Ambite,[2] C.L. Avery,[3] S. Buyske,[4,5] C. Cai,[2] M.D. Fesinmeyer,[6] C. Haiman,[7] G. Heiss,[3] L.A. Hindorff,[8] C.-N. Hsu,[2] R.D. Jackson,[9] C. Kooperberg,[6] L. Le Marchand,[10] Y. Lin,[6] T.C. Matise,[5] L. Moreland,[11] K. Monroe,[7] A.P. Reiner,[6,12] R. Wallace,[13] L.R. Wilkens,[10] D.C. Crawford,[1,14] and M.D. Ritchie[1,14]*

[1]*Center for Human Genetics Research, Vanderbilt University, Nashville, Tennessee*
[2]*Information Sciences Institute, University of Southern California, Marina del Rey, California*
[3]*Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina*
[4]*Department of Statistics, Rutgers University, Piscataway, New Jersey*
[5]*Department of Genetics, Rutgers University, Piscataway, New Jersey*
[6]*Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington*
[7]*Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California*
[8]*National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland*
[9]*Ohio State University, Columbus, Ohio*
[10]*Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii*
[11]*University of Pittsburgh, Pittsburgh, Pennsylvania*
[12]*Department of Epidemiology, University of Washington, Seattle, Washington*
[13]*Departments of Epidemiology and Internal Medicine, University of Iowa, Iowa City, Iowa*
[14]*Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, Tennessee*

The field of phenomics has been investigating network structure among large arrays of phenotypes, and genome-wide association studies (GWAS) have been used to investigate the relationship between genetic variation and single diseases/outcomes. A novel approach has emerged combining both the exploration of phenotypic structure and genotypic variation, known as the phenome-wide association study (PheWAS). The Population Architecture using Genomics and Epidemiology (PAGE) network is a National Human Genome Research Institute (NHGRI)-supported collaboration of four groups accessing eight extensively characterized epidemiologic studies. The primary focus of PAGE is deep characterization of well-replicated GWAS variants and their relationships to various phenotypes and traits in diverse epidemiologic studies that include European Americans, African Americans, Mexican Americans/Hispanics, Asians/Pacific Islanders, and Native Americans. The rich phenotypic resources of PAGE studies provide a unique opportunity for PheWAS as each genotyped variant can be tested for an association with the wide array of phenotypic measurements available within the studies of PAGE, including prevalent and incident status for multiple common clinical conditions and risk factors, as well as clinical parameters and intermediate biomarkers. The results of PheWAS can be used to discover novel relationships between SNPs, phenotypes, and networks of interrelated phenotypes; identify pleiotropy; provide novel mechanistic insights; and foster hypothesis generation. The PAGE network has developed infrastructure to support and perform PheWAS in a high-throughput manner. As implementing the PheWAS approach has presented several challenges, the infrastructure and methodology, as well as insights gained in this project, are presented herein to benefit the larger scientific community. *Genet. Epidemiol*. 35:410–422, 2011. © 2011 Wiley-Liss, Inc.

Key words: genetic epidemiology; high throughput; phenomics; genetics; PheWAS

Contract grant sponsor: National Human Genome Research Institute (NHGRI); Contract grant numbers: U01HG004803; U01HG004798; U01HG004802; U01HG004790; U01HG004801; Contract grant sponsor: NHGRI PAGE; Contract grant numbers: U01HG004798-01; U01HG004802; U01HG004790; U01HG004803; Contract grant sponsors: Centers for Disease Control and Prevention; Vanderbilt University Center for Human Genetics Research; Computational Genomics Core; NIH; National Institute of Neurological Disorders and Stroke; National Institutes of Mental Health; National Center for Research Resources; Contract grant number: M01-RR00425; Contract grant sponsor: National Institute of Diabetes and Digestive and Kidney Diseases; Contract grant number: DK063491; Contract grant sponsor: NHLBI; Contract grant numbers: N01-HC65233; N01-HC65234; N01-HC65235; N01-HC65236; N01-HC65237; U01 HL65520; U01 HL41642; U01 HL41652; U01 HL41654; U01 HL65521; Contract grant sponsor: National Heart, Lung, and Blood Institute; Contract grant numbers: N01-HC-55015; N01-HC-55016; N01-HC-55018; N01-HC-55019; N01-HC-55020; N01-HC-55021; N01-HC-55022; N01-HC-95095; N01-HC-48047; N01-HC-48048; N01-HC-48049; N01-HC-48050; N01-HC-45134; N01-HC-05187; N01-HC-45205; N01-HC-85079; N01-HC-85086; N01-HC-35129; N01-HC-15103; N01 HC-55222; N01-HC-75150; N01-HC-45133; U01HL080295; R01 HL087652; Contract grant sponsor: U.S. Department of Health and Human Services; Contract grant numbers: N01WH22110; 24152; 32100-2; 32105-6; 32108-9;

32111-13; 32115; 32118-32119; 32122; 42107-26; 42129-32; 44221; Contract grant sponsor: National Cancer Institute; Contract grant numbers: R37CA54281; R01 CA63; P01CA33619; U01CA136792; U01CA98758.

*Correspondence to: M.D. Ritchie, Vanderbilt University, Center for Human Genetics Research, 2215 Garland Avenue, 519 Light Hall, Nashville, TN 37232-0700. E-mail: marylyn.ritchie@vanderbilt.edu

Received 11 February 2011; Revised 1 April 2011; Accepted 3 April 2011
Published online 18 May 2011 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.20589

INTRODUCTION

Much of the current focus on investigating the relationship between genetic variation and disease is through the use of genome-wide association studies (GWAS). This approach has been an important workhorse in genetic epidemiology for the past 5 years, as hundreds of SNPs have been associated with the risk of complex diseases, such as type 2 diabetes and osteoporosis, as well as with health-related intermediate traits such as elevated circulating low density lipoprotein (LDL) cholesterol levels [Hindorff et al., 2009]. While these associations have explained a small proportion of the heritability of these traits, they have provided new leads toward a better understanding of the etiology and underlying biology of disease [Lou et al., 2009; Moffatt et al., 2007; Musunuru et al., 2010].

One could argue that the current paradigm of GWAS is limited and suffers from important shortcomings that inhibit the quest to further the understanding of the genetic contribution to disease and traits. Despite the decrease in cost over time, GWAS are generally limited in scope due to the cost of genotyping hundreds of thousands millions of SNPs for each individual within a study. GWAS usually focus on a specific phenotypic outcome or series of measurements. The focus on a limited phenotypic domain, such as the specific presence or absence of a single disease in a GWAS, neglects the potential power gained through the use of intermediate phenotypes, sub-phenotypes, biomarkers, and endophenotypes that may more closely reflect a gene's mechanism, as well as the relationship between genetic variation and multiple diseases and phenotypes (pleiotropy). Finally, most published GWAS have been performed in populations of European-decent, and there is still little characterization of the relationship between risk variants found in GWAS and disease and/or phenotypes in other racial/ethnic groups.

A complementary approach to GWAS is the "phenome-wide association study" (PheWAS). In PheWAS, the association between a number of common genetic variations and a wide variety and large number of phenotypes are systematically characterized. Phenotypic/genotypic resources amenable to the PheWAS approach are large-scale epidemiologic studies with comprehensive collections of well-characterized phenotypic measurements and environmental exposures recorded prospectively or retrospectively for thousands of participants, such as the studies of the Population Architecture using Genomics and Epidemiology (PAGE) network. Electronic medical record (EMR) resources coupled to genotypic data can also be used for PheWAS, as they also contain rich resources of phenotypic and genotypic information.

The EMR PheWAS approach was recently employed successfully by Denny et al. in a proof-of-concept investigation using BioVU [Ritchie et al., 2010; Roden et al., 2008], Vanderbilt University's biobank. In this PheWAS, five SNPs selected from the candidate gene and GWAS literature were tested for associations with ICD-9 codes in several thousand patients. The investigators were able to detect both the original associations (for atrial fibrillation, Crohn's disease, carotid artery stenosis, coronary artery disease, multiple sclerosis, systemic lupus erythematosus, and rheumatoid arthritis), as well as potentially novel associations with other clinical outcomes/conditions in the EMR [Denny et al., 2010].

While ground-breaking, the sole PheWAS published in the literature has limitations that can be addressed through the use of data from the studies of the (PAGE) network. Like initial GWAS studies in the literature, the initial PheWAS focused primarily on binary disease traits in European Americans from a clinic setting. The studies of PAGE that include quantitative measurements, along with detailed disease status (incident or prevalent depending on the study), and in most cases longitudinal follow-up information, have the potential for increased power to detect genotypic/phenotypic relationships, as well as characterize those relationships across race/ethnicity in a population-based manner. In addition, the phenotypic measurements/outcomes of PAGE are measured using standardized protocols, thus reducing measurement error and facilitating phenotype harmonization between studies.

As stated above, the PheWAS analysis represents tests of association between a large number of SNPs and phenotypes and traits available in PAGE, and is meant to be high-throughput. As such, results of these first-pass analyses are considered hypothesis-generating and require additional scrutiny before the findings are further considered for follow-up. This hypothesis-generating exercise is unlike the directed, a priori hypothesis-testing within PAGE whereby specific SNPs hypothesized to be associated with specific phenotypes are tested only for those phenotypes. These a priori analyses include careful phenotype harmonization for traits and outcomes that overlap across two or more PAGE studies, as well as considerable investigation of the possible effect of covariates such as age, gender, and environmental exposure(s) on the association between genetic variation and phenotypic outcome. Unlike PheWAS, the advantage of these more carefully directed analyses is the potential for identification and characterization of genetic modifiers and accurate effect estimates. Also, because much effort has been expended to harmonize phenotypes across studies at the initiation of the study, less effort is required to scrutinize the results from these analyses compared with PheWAS. Despite the advantages of the more a priori driven approach, the hypothesis generating PheWAS promises the opportunity to identify unsuspected genotypic and phenotypic relationships for further investigation that can include these forms of more thorough model characterization.

We describe herein the conceptual framework and design of the high-throughput, first-pass analysis of the relationship between all risk variants genotyped thus far within PAGE and the comprehensive phenotypic resources of the PAGE network, using this PheWAS approach. The results of this scan across SNPs and phenotypes can be used to discover novel relationships between SNPs, phenotypes, and networks of phenotypes to foster hypothesis generation. This manuscript presents the infrastructure and methodology, as well as insights gained in this PAGE-directed project, to benefit the larger scientific community.

PAGE

The PAGE network, funded by the National Human Genome Research Institute (NHGRI) in 2008, is a collaboration of four study sites, representing eight population-based studies, and a coordinating center [Matise et al., 2010; in press]. All PAGE network studies are diverse and include multiple racial/ethnic populations such as European Americans, African Americans, Hispanic Americans, Asian Americans, and Native Americans. The PAGE network consists of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study accessing the National Health and Nutritional Examination Surveys (NHANES) [CDC, 2008; McQuillan et al., 2003], the Multiethnic Cohort Study (MEC) [Kolonel et al., 2000], the Women's Health Initiative (WHI) [1998], and the Causal Variants Across the Life Course (CALiCo) that is a consortium of five cohort studies: Atherosclerosis Risk in Communities [1989], Coronary Artery Disease Risk Development in Young Adults (CARDIA) [Hughes et al., 1987], Cardiovascular Health Study (CHS) [Fried et al., 1991], the Strong Heart Cohort Study [Lee et al., 1990], the Strong Heart Family Study, and the Hispanic Community Health Study/Study of Latinos.

The primary focus of the PAGE network is the deep characterization of well-replicated genetic risk variants identified in GWAS (~400 to date) and their relationships to specific phenotypes (e.g., lipids, diabetes, heart disease, cancers). Included in the characterization process is (1) replication of the original association in a population of similar genetic ancestry as the discovery population (which to date has been largely European-descent), (2) generalization of the association to diverse populations such as African Americans, Hispanics/Mexican Americans, American Indians, and other groups, (3) identification of modifiers involved in gene-gene and gene-environment interactions, and (4) identification of pleiotropy.

Table I shows several examples of phenotypes available in the studies of the PAGE network, demonstrating the breadth of phenotypes across PAGE. Table II contains more information about the following specific studies in this PheWAS from PAGE: WHI, EAGLE, MEC, CALiCo-ARIC, CALiCo-CHS, and CALiCo-CARDIA. Additional details about the various PAGE network studies are described in Matise et al. [2010; in press].

METHODS FOR IMPLEMENTING PheWAS IN PAGE

The intent of PheWAS is to keep data processing high-throughput and relatively simple while also annotating results with respect to quality control (general workflow in Fig. 1). Study samples are genotyped and quality control procedures are followed at each site independently. In addition all tests of association are calculated independently for each study. All analyses are stratified by race/ethnicity, assuming an additive genetic model where the homozygous genotype for the coded allele (reference allele chosen to be standard across race/ethnicity) is coded as "2," the heterozygous genotype coded as "1," and the homozygous genotype for the non-coded allele is coded as "0." All data analyses in PheWAS are standard logistic or linear regression analysis, dependent on design or whether the phenotypes are continuous or binary. For variables with multiple categories, binning is used to create new variables of the form "A vs. not A." An example would be a measurement of menarche, where study participants indicated where they fell within a series of age ranges (9–10, 11–12, 13–14, etc.). In PheWAS, these results would be broken into the age range of 9–10 vs. all other respondents, then a separate binary variable of the age range of 11–12 vs. all other respondents, etc. In addition all continuous phenotypes are natural log transformed after a constant of one is added to each quantitative trait measurement (as some traits have values of "0"). After this transformation, the associations between all SNPs and the transformed phenotypes are calculated. Data transformation will not be necessary for some phenotypes; however, to maintain a high-throughput workflow, we perform the same analyses on all continuous variables. Depending on the phenotype, results from linear regressions on untransformed or

TABLE I. Example survey categories, measurements, and phenotypes within PAGE-PheWAS

Example survey categories	Biomarker, cancer, renal function, neurology, nutrition, infectious disease, inflammation
Example measurements	Respiratory, skin, bone, muscle, joint, speech, hearing, biochemistry tests, cardiovascular, anthropometric, ocular
Example phenotypes	Systolic and diastolic blood pressure, body mass index, gingival bleeding, number tooth carries, apolipoprotein B level, wheal length alternaria allergy scratch test, plasma glucose level, fibrinogen level, monocyte number, QRS interval, QT interval, platelet count, forced vital capacity (FVC), self-reported presence/absence diabetes, self-reported presence/absence osteoporosis, self-reported presence/absence cardiovascular disease, follicle stimulating hormone, tympanometry measurement, balance measurement, bone mineral density, keratometry, self-reported presence/absence eczema, glycohemoglobin measurement, self-reported presence/absence asthma, cotinine levels, c-reactive protein level, HDL levels, LDL Levels, white blood cell count, hemoglobin, hematocrit, urine creatinine, serum folate, serum globulin, luteinizing hormone
Results reported	SNP ID, beta (linear regression), *P*-value, odds ratios (logistic regression), allele frequencies, associated phenotype and previously associated phenotype, phenotype summary information specific to the regression

TABLE II. Background for the six studies represented in the PAGE-PheWAS

Study	Cohort description	Study description	Age range	Sex	References
WHI	Study of post-menopausal women's health consisting of four clinical trials and an observational study	Prospective and longitudinal collection of biospecimens, intermediate outcomes, and phenotypic characteristics, such as bone mineral density, hormone concentrations, breast density, and inflammation.	50–79 at baseline	Women	[WHI, 1998]
CALiCo—ARIC	Ongoing population-based cohort of Caucasian and African-American males and females, selected using probability sampling from four United States communities	Participants recruited in 1987–1989 to examine cardiovascular and pulmonary disease, patterns of medical care, and disease variation over time. Standardized physical examinations and interviewer-administered questionnaires conducted at baseline, and at three triennial follow-up examinations. Participant follow-up through annual telephone interviews, review of hospitalization records and vital status is ongoing	45–64	Men and women	[ARIC, 1989]
CALiCo—*CHS*	CHS is a prospective, longitudinal cohort study of risk factors for cardiovascular disease in the elderly in Caucasian and African-American adults	Starting in 1989 and continuing through 1999, adults were sampled from four United States communities. Participants underwent extensive yearly annual clinical examinations measuring traditional risk factors such as blood pressure and lipids as well as measures of subclinical disease	65–100 at baseline	Men and women	[Fried et al., 1991]
CALiCo—CARDIA	The Coronary Artery Risk Development in Young Adults (CARDIA) is an ongoing longitudinal cohort study initiated in 1984 to investigate factors that influence coronary heart disease during young adulthood in Caucasians and African Americans	Participants were recruited from four urban areas: Birmingham, Alabama; Chicago, Illinois; Minneapolis, Minnesota, and Oakland, California. Examinations included standardized measurements of major risk cardiovascular disease risk factors as well as assessments of psychosocial, dietary, and exercise-related characteristics	18–30 years at baseline	Men and women	[Hughes et al., 1987]
MEC	Population-based cohort study of men and women in Hawaii and California with a biorepository of blood and urine. Five US racial/ethnic groups: African Americans, European Americans, Japanese Americans, Latinos and Native Hawaiians	Participants included in PAGE are those from nested case-control studies of various cancers and of type 2 diabetes. Prospectively collected risk factor (e.g., diet, smoking, physical activity), biomarker and clinical data	45–75	Men and women	[Kolonel et al., 2000]
EAGLE	EAGLE accesses National Health and Nutrition Examination Survey (NHANES). NHANES is a cross-sectional survey of Americans ascertained regardless of health status that over-samples minorities (non-Hispanic blacks and Mexican Americans), the young, and the elderly	NHANES contains detailed demographic, health, lifestyle, laboratory, clinical, and physical examination data. Genetic NHANES included in this PheWAS are DNA samples collected during NHANES III (phase 2; 1991–1994), NHANES 1999–2000, and NHANES 2001–2002	12–95	Men and women	[CDC 2008; McQuillan et al., 2003]

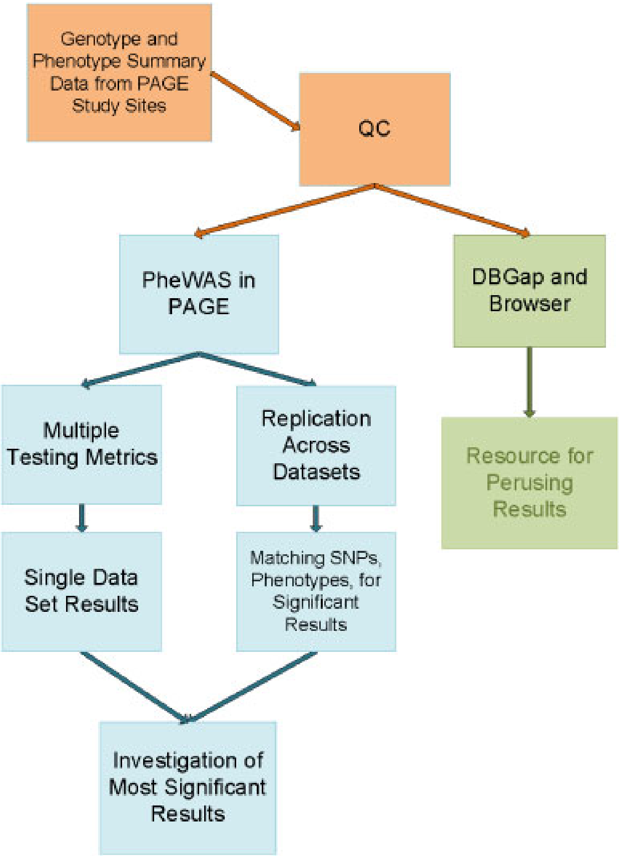
Genotype and Phenotype Summary Data from PAGE Study Sites
QC
PheWAS in PAGE
DBGap and Browser
Multiple Testing Metrics
Replication Across Datasets
Resource for Perusing Results
Single Data Set Results
Matching SNPs, Phenotypes, for Significant Results
Investigation of Most Significant Results

Fig. 1. Workflow for PheWAS. Study samples are genotyped and quality control procedures are followed at each site independently. In addition all tests of association are calculated independently for each study. The PheWAS quality control step flags results as described in the text. As a resource for PAGE investigators, the most significant associations from PheWAS can be explored in the PAGE browser across race/ethnicity (green). For investigators outside of PAGE, aggregate data from PheWAS will be available via dbGAP. For the PAGE PheWAS analysis, results for six studies will be used and where SNPs and phenotypes exist across data sets, replication will be sought (blue). Significant results can then be selectively investigated and characterized further.

transformed data can be later selected for interpretation and further downstream analyses. Because each PAGE study site performs their own tests of association, a variety of statistical packages are used, such as SAS version 9.2 and R version 2.11.0 (The R Foundation for Statistical Computing).

These analyses are not computationally intensive. The bigger challenge has been the surrounding infrastructure necessary to run and process this many tests of associations in a high-throughput fashion, which has necessitated the writing of software scripts/programs for automation purposes. It became clear early in the PheWAS analyses that if all results were to be stored, combined, and shared across sites, the use of a standardized template for results would be critical. Thus, the PAGE Coordinating Center developed a standard template for SNP specific information, phenotype-specific information, and SNP-phenotype association results. While using standardized templates has required each site to adapt their workflow to format results into the template, this process has enabled the group to efficiently collect, integrate, and analyze results across studies using a database.

Some basic quality control and result flagging procedures are also implemented by the PAGE Coordinating Center when the individual results from different studies are submitted. After the minimal curation steps described below, the results of associations can then be examined and explored. For continuous variables, skewness of the data is a potential issue; thus, degree of skewness is calculated for all continuous phenotypes. Skewness is also calculated for transformed phenotypes. For most continuous traits, the outliers should not impact the skewness of the data given the relatively large sample sizes for this study. Results are ''flagged'' where the phenotype for a given set of linear regression results have an absolute skewness value greater than 2 ($|\text{skewness}|>2$) for the transformed and/or untransformed phenotypes. Figure 2 shows the results of skewness before and after natural log transformation for 351 continuous phenotypes within WHI, stratified by race/ethnicity [African Americans (AA), European Americans (EA), Hispanics/Latinos (H)]. A horizontal red line is at skewness of $+/-2$. Extreme skew values are visible in the untransformed data. After data transformation, there is a shift of skewness away from extreme values.

For logistic regression and proportional hazards regression, odds ratios that are greater than 3 are flagged. While these results may be simply be very significant results, more likely, the phenotype may need more inspection or the result may reflect poor estimation because of small cell sizes. Results with an absolute value of beta divided by the standard deviation of the phenotype used in the regression greater than 2.5 ($|\text{beta}|/\text{SD}>2.5$) are also flagged. Much like the OR flag, these may be very significant results or perhaps some aspect of the association that should be investigated further. Finally, regression results where the Hardy-Weinberg Equilibrium (HWE) *P*-value of the SNP used in the regression is $P<0.00001$ are flagged. Because PAGE is genotyping and characterizing SNPs originally described in studies of European Americans, HWE may differ considerably if the population is admixed, which is expected in many of the racial/ethnic groups studied in PAGE. Both genetic ancestry (including admixture) and poor quality genotyping data are viable explanations for deviations from HWE, both of which will require further inspection of the data. The process of flagging results that do not pass certain criteria provides information to investigators so that characteristics of specific phenotype, SNP genotype, and regression results can be easily identified and more carefully considered before further data interpretation.

PheWAS EXPLORATION

The generation of PheWAS results has the potential to provide unique research opportunities, and one of the greatest challenges for PheWAS is not the implementation, but rather the exploration and sharing of the results. Two avenues for disseminating and exploring the PAGE PheWAS results are being used: dbGaP [Mailman et al., 2007] and the PAGE Browser. For all analyses, aggregate data will be deposited into dbGAP for secondary access by

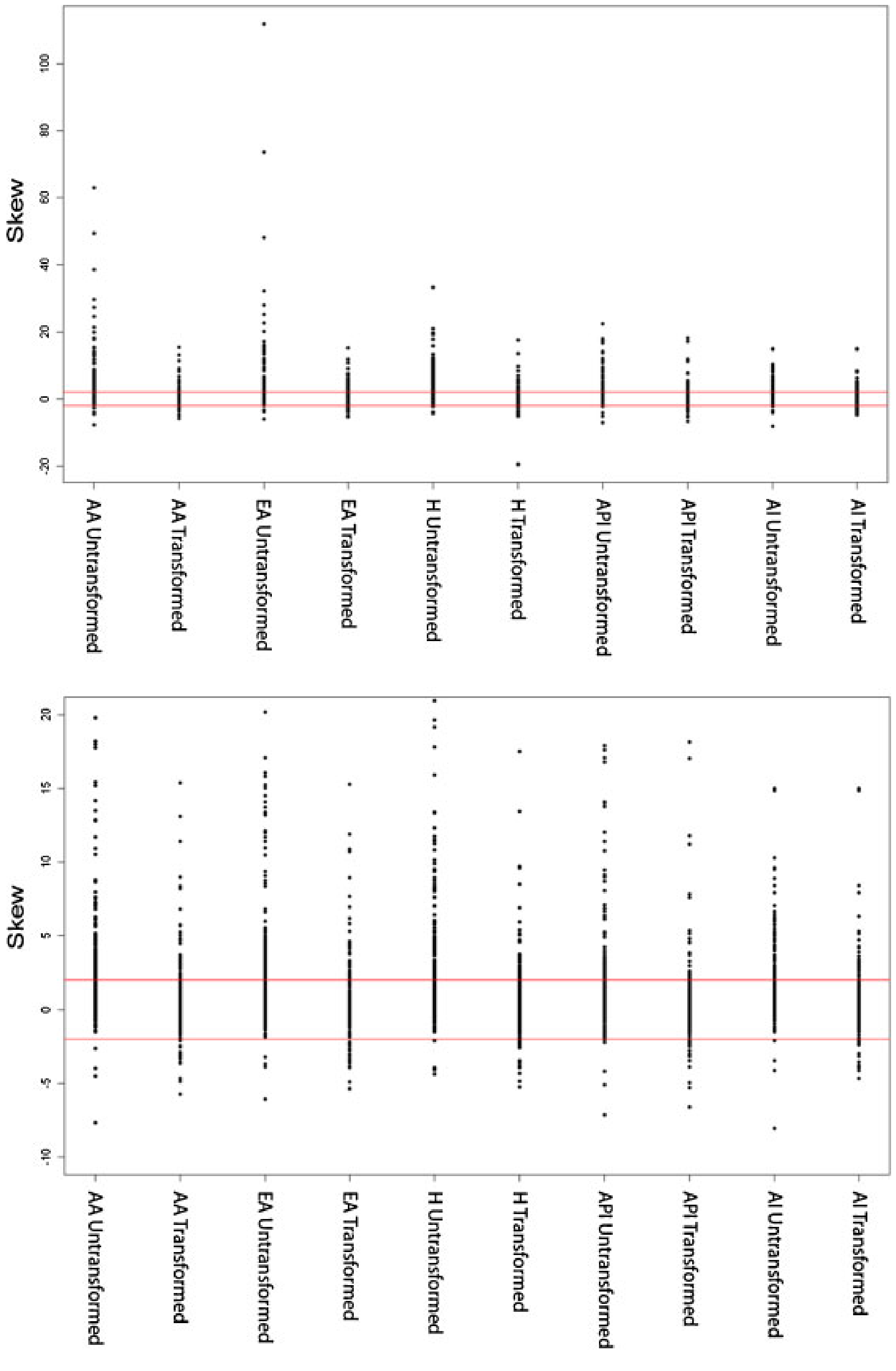

Fig. 2. Skewness and transformation. Skewness measured for 351 phenotypes, stratified by race/ethnicity, before transformation and after adding 1 then natural log transforming each continuous variable. Results are presented in the upper and lower panel for: African American (AA) untransformed, AA transformed, European American (EA) untransformed, EA transformed, Hispanic (H) untransformed, H transformed, Asian Pacific Islander (API) transformed, API untransformed, American Indians (AI) untransformed, AI transformed. Horizontal red lines are at positive two and negative two. The lower panel shows a "zoomed in" version of the upper panel, showing the compression of points after the transformation of variables.

approved investigators. Furthermore, results will be available and searchable through the PAGE browser. The PAGE browser is being developed by the PAGE Coordinating Center as a tool for searching, filtering, and visualizing the results from PheWAS, as well as other PAGE analyses. The PAGE browser includes information on SNPs, phenotypes, study, ancestry, and allows for PheWAS result filtering and display options depending on the approach of interest, whether by study, race, ethnicity, or phenotype. In addition to tabular data, the browser provides data visualization through the use of box plots and bar plots of summary statistics for quantitative and categorical variables, forest plots that display confidence intervals and effect sizes for individual SNP-phenotype results, and heat map plots indicating significance levels for large numbers of SNP-phenotype results. The results are currently provided to PAGE investigators and PAGE collaborators as a resource for investigating results on an SNP basis or phenotype basis using the interactive browser tool developed by the PAGE Coordinating Center. In the future, it is anticipated that outside investigators will be able to request access to the resource as a way to generate hypotheses and insights that could direct further research into genetic variation and phenotypes (Fig. 3).

ACCOUNTING FOR MULTIPLE TESTING IN PheWAS–USING REPLICATION IN PAGE STUDIES

A major challenge of PheWAS is adjusting for testing multiple hypotheses to protect from type I errors. Much has been written concerning the multiple hypothesis testing problem in GWAS and possible solutions [Rice et al., 2008]. However, unlike GWAS, PheWAS has an added dimension to this problem: hundreds to thousands of phenotypes in addition to potentially hundreds to thousands of SNPs. Depending on the total number of phenotypic and genotypic variables and the number of racial/ethnic groups, hundreds of thousands to millions of tests of association can be performed. As an example, in the WHI analysis there were 95 SNPs, 1,514 phenotypes, and 5 racial/ethnic groups, resulting in ~200,000 tests of association. Calculation of the degree of false-positive results due to multiple hypothesis testing is not straightforward as there are potential correlations between phenotypes and between SNPs. Also, a proportion of the genotype-phenotype correlations in the data set represent already-known associations, and, as such, are positive controls rather than experimental tests of association.

Despite the best efforts to determine the number of statistical tests performed, simple adjustments based on these counts (such as a Bonferroni correction) may not be appropriate. It is conceivable that alternative approaches such as false discovery rate, using principle component analysis to determine the effective number of independent tests, or use of permutation testing maybe utilized. However, within PAGE, only summary data are shared between studies, resulting in significant challenges for implementing these forms multiple testing correction. Thus, for the initial PAGE-PheWAS, in terms of prioritization of the first set of results, we will be reporting the most statistically significant results that are observed in 2 or more PAGE studies for the same SNP, phenotype and/or phenotype category, and race/ethnicity. This requirement for replication between these independent PAGE data sets lessens the chance of a result found by chance alone, particularly for results found to replicate in more than two studies. Recall that PheWAS is a hypothesis generating exercise; therefore, we are less concerned with the effect estimates and precision of the level of significance. Our goal is to identify novel SNP-phenotype associations that have evidence of replication in multiple PAGE studies, to consider in further in future studies. In addition, because multiple race/ethnicities are represented within these data, we intend to describe the presence/absence of generalization of PheWAS results across race/ethnicity. We define generalization here as a significant association in more than one population, in addition to a similar direction of effect.

INTERPRETATION OF RESULTS

The ultimate challenge of PheWAS data is interpretation. Multiple approaches for prioritizing and interpreting results are available for the exploration of results for PheWAS (Fig. 4). Such approaches, as well as related questions when generating hypotheses from these results, include the following:

SORTING RESULTS BY *P*-VALUE

As mentioned before, the most significant *P*-value results can be chosen for further inspection and analysis. However, the most significant results may not be the results of greatest interest depending on the phenotype and/or SNP.

IS THE RESULT EXPECTED OR UNEXPECTED?

One way of considering a result "expected," is through observing replication of a previously identified/reported genotype-phenotype association. Observing expected results provides positive controls for this high-throughput experiment. Another example of results possibly considered more "expected" are those for an SNP with a previously identified genotype-phenotype association also associated with closely related and/or correlated phenotypes. Although discovery of new gene-phenotype associations in PheWAS is both likely and important, many of the findings from PheWAS may reflect the inter-relationships existing among the phenotypes measured in epidemiological studies. In fact, some results may be more about the relationship between particular phenotypes and disease outcomes than about the impact of genetic variation on an independent phenotype. For example, a known type II diabetes-associated SNP may be associated with type II diabetes in PheWAS, as well as with related type II diabetes phenotypes such as fasting glucose levels, hyperinsulinemia, obesity, and other diabetes-related phenotypes. Another example of a more "expected result" would be for SNPs in high linkage disequilibrium (LD) with SNPs that have shown previously identified genotype-phenotype associations with a phenotype used in the PheWAS analysis.

More "novel" results may occur in multiple ways and exemplify pleitropy. As an example, an SNP that has been significantly associated with asthma in the literature is found in PheWAS to have an association with various blood cytokine levels, allergy skin tests results, and fungal infections. While these are related phenotypes, this novel insight may provide further information about the relationship between genetic variation and allergic/inflammatory responses. If two phenotypes are not known to be related, such as might be identified by observing significant associations between a single SNP and two apparently independent diseases, this may indicate the impact of genetic variation on multiple unrelated

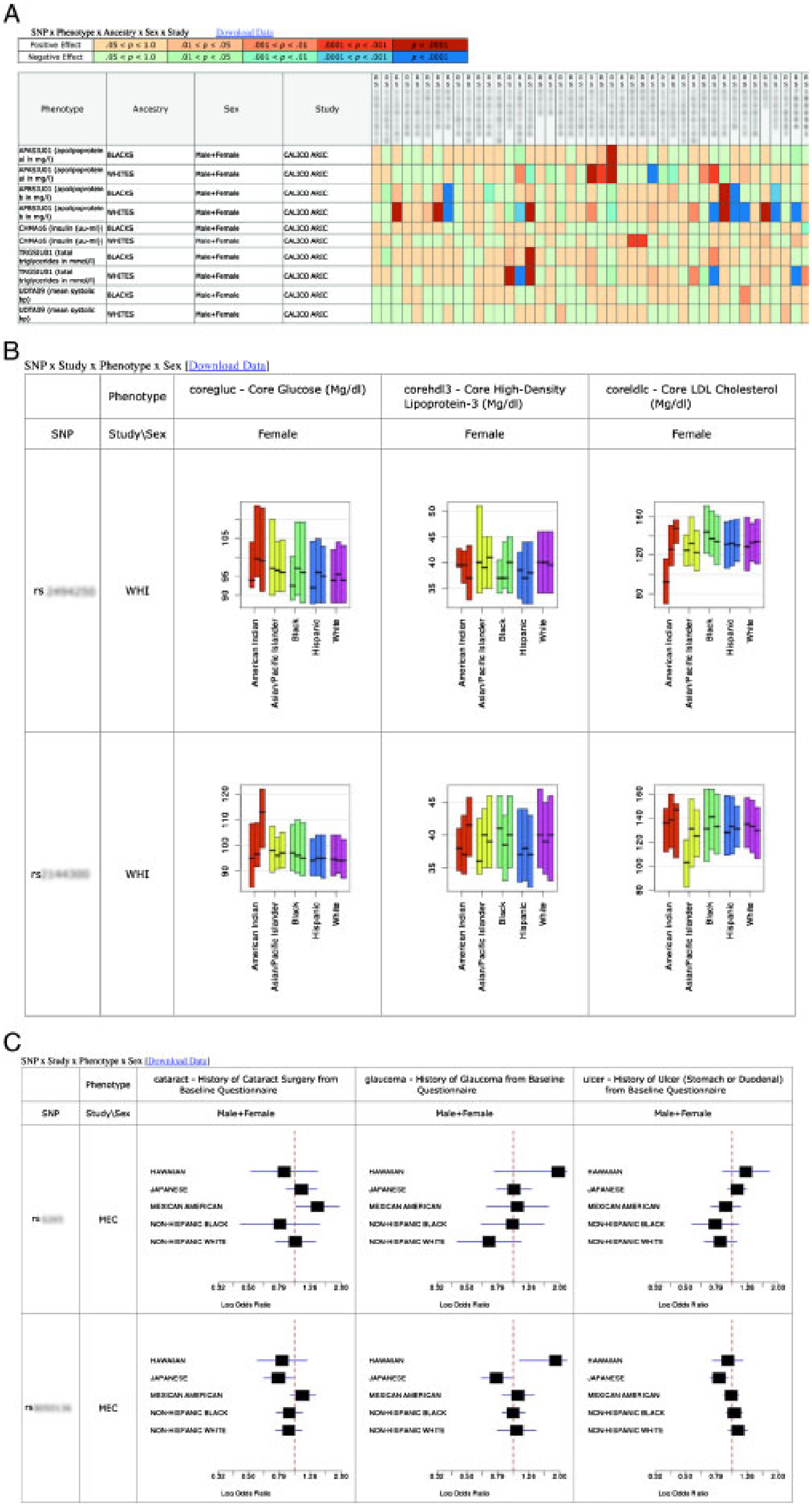
A
SNP x Phenotype x Ancestry x Sex x Study
Download Data
Positive Effect
.05 < p < 1.0
.01 < p < .05
.001 < p < .01
.0001 < p < .001
p < .0001
Negative Effect
.05 < p < 1.0
.01 < p < .05
.001 < p < .01
.0001 < p < .001
p < .0001
Phenotype
Ancestry
Sex
Study
APASSU01 (apolipoprotein al in mg/l)
BLACKS
WHITES
Male+Female
CALICO ARIC
APBSSU01 (apolipoprotein b in mg/l)
CHMA06 (insulin (au-ml))
TRGS01U01 (total triglycerides in mmol/l)
UDTA09 (mean systolic bp)
B
SNP x Study x Phenotype x Sex [Download Data]
Phenotype
coregluc - Core Glucose (Mg/dl)
corehdl3 - Core High-Density Lipoprotein-3 (Mg/dl)
coreldlc - Core LDL Cholesterol (Mg/dl)
SNP
Study\Sex
Female
rs [redacted]
WHI
American Indian
Asian/Pacific Islander
Black
Hispanic
White
C
SNP x Study x Phenotype x Sex [Download Data]
Phenotype
cataract - History of Cataract Surgery from Baseline Questionnaire
glaucoma - History of Glaucoma from Baseline Questionnaire
ulcer - History of Ulcer (Stomach or Duodenal) from Baseline Questionnaire
SNP
Study\Sex
Male+Female
rs [redacted]
MEC
HAWAIIAN
JAPANESE
MEXICAN AMERICAN
NON-HISPANIC BLACK
NON-HISPANIC WHITE
0.32
0.50
0.79
1.26
2.00
Log Odds Ratio

Fig. 3. PAGE browser. Different views of the PAGE browser provide different possibilities for investigators examining the results of the analyses. (A) The heat map view, where association results are plotted in a "heat map" format. SNPs are along *x*-axis, phenotypes are along the *y*-axis. The color of each cell corresponding to the individual SNP-by-Phenotype results is either a shade of yellow or red if the direction of effect is positive, or a shade of green or blue if the direction of the effect is negative. Increasing significance is indicated by shade respective to the direction of effect. (B) Phenotypic summary information across race/ethnicity. The box-plots show the 1st, 2nd, and 3rd quantiles by genotype for quantitative phenotypes. (C) For associations, forest plots are available, stratified across race/ethnicity.

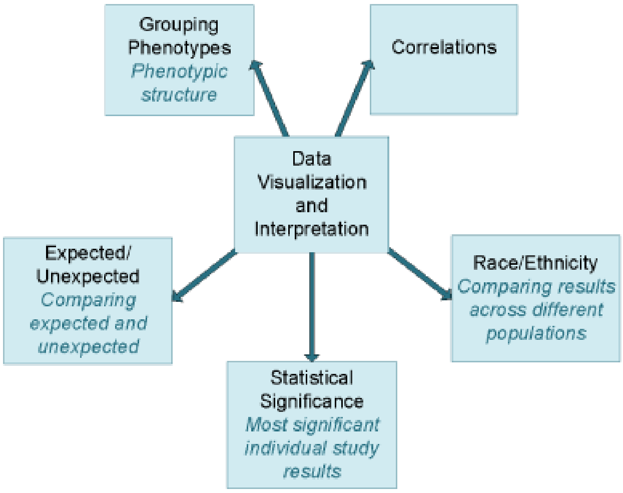
Grouping Phenotypes
Phenotypic structure
Correlations
Data Visualization and Interpretation
Expected/ Unexpected
Comparing expected and unexpected
Race/Ethnicity
Comparing results across different populations
Statistical Significance
Most significant individual study results

Fig. 4. Approaches for visualizing and interpreting results of PheWAS. There are multiple ways the results of PheWAS can be explored. Correlations between quantitative phenotypes can be calculated to investigate some of the relationships between phenotypes. Results across race/ethnicity can be compared. Results can be ranked by significance and then explored one at a time for further replication and characterization with further phenotype harmonization. The expected/unexpected nature of results can also be investigated, as some genotype-phenotype associations are previously known. Grouping of phenotypes is possible, as well as further exploration of phenotypic structure.

phenotypes (pleiotropy) or may expose a previously unknown relationship between two phenotypic outcomes. An example of this would be that of a known diabetes-associated SNP significantly associated with arthritis. This could potentially reflect the relationship between autoimmunity, diabetes, and arthritis, and/or the contribution of genetic variation of that gene to both phenotypes.

CORRELATION BETWEEN PHENOTYPES

In an extension to some of the discussions of related phenotypes, one possible approach to gain more information about and potentially uncover novel phenotypic relationships in PheWAS is the calculation of correlation matrices for all continuous phenotypes within each PAGE study. The results can then be explored in a heat-map, with higher correlations (positive and negative) having a lighter color (yellow) and lower correlations in a darker color (blue) (Fig. 5). Analogous to LD between SNPs, correlations between phenotypes can be used to interpret results from tests of association as both significant association results and strongly correlated phenotypes can be detected visually. In the example of Figure 5, the PheWAS results and correlation matrix are from EAGLE/NHANES data and the plot shows three areas of high correlation. For the allergy skin tests, measurements of the allergy response were recorded for each participant tested. Among these measurements, there is a high correlation between response to *Alternaria alternata* (a major plant pathogen) and Bermuda grass. In a similar vein, the bone mineral content and bone mineral density measurements in participants are highly correlated. Viewing these correlations has the potential for providing information for interpretation of some of the association results. Relationships between various categorical variables and phenotypes can be investigated on a case by case basis. For example, when comparing a specific continuous phenotype to a binary trait, a simple *t*-test can be used to determine if there are different distributions of the continuous variable between cases and controls. Likewise, certain binary traits can be investigated to see if there is a common overlap between cases/controls for one binary variable vs. another. This is a potential option to explore, not a necessary component of the PheWAS workflow.

RACE/ETHNICITY

Most GWAS and candidate gene studies to date have been performed using populations of European-descent. All PheWAS analyses here are being performed in multiple race/ethnicities; thus, results can be characterized across race/ethnicities to better understand how specific associations may or may not generalize across populations of differing genetic ancestry.

After initial exploration of PheWAS results, further work will be necessary to fully characterize phenotype-genotype associations of interest, and PAGE has the resources and expertise for follow through analyses for specific PheWAS results. In the case of results where the SNP and phenotype exist across two or more studies, initial follow-up work includes: determining the consistency of direction-of-effect, examining the *P*-values across the studies, examining how the results vary across race/ethnicity, and determining the sample size and resultant power across the studies. Further work includes directed, traditional hypothesis-testing, as performed already within PAGE for previously known genotype-phenotype associations, such as the work of the PAGE Lipids Project Group [Dumitrescu et al., 2010; submitted]. For phenotype-genotype associations that only exist in a single PAGE study because either the SNP or phenotype was not evaluated in other PAGE studies, replication of the result in a data set outside the PAGE network would be useful. An example would be replicating PheWAS cardiac trait results using data from the Framingham Heart Study GWAS [Cupples et al., 2007]. In addition, as PAGE moves forward, an SNP found to be of interest may become a target SNP for future a priori driven work in PAGE, and may be proposed for study-wide genotyping, increasing the sample sizes available for PAGE analysis.

FUTURE DIRECTIONS FOR PheWAS

Using the information gained from the broad interrogation of interrelated phenotypes and the relationship to genetic variation provides an opportunity to consider and explore the networks of phenotypes that comprise disease, as well as identify pleiotropy. The emerging field of phenomics has been championing "deep-phenotyping," the broad collection of diverse and detailed phenotypic information and investigating network structure among those arrays of phenotypes [Bilder et al., 2009; Ghebranious et al., 2007; Rzhetsky et al., 2007]. There have been discussions of approaches for relating these phenotypic data back to genotypic variation [Jones et al., 2005], as well as visualization and interpretation of phenotypic networks [Lanktree et al., 2010; Rzhetsky et al., 2007]. PheWAS is novel in the promise of harnessing and interrogating both phenotypic structure and large numbers of genotypes

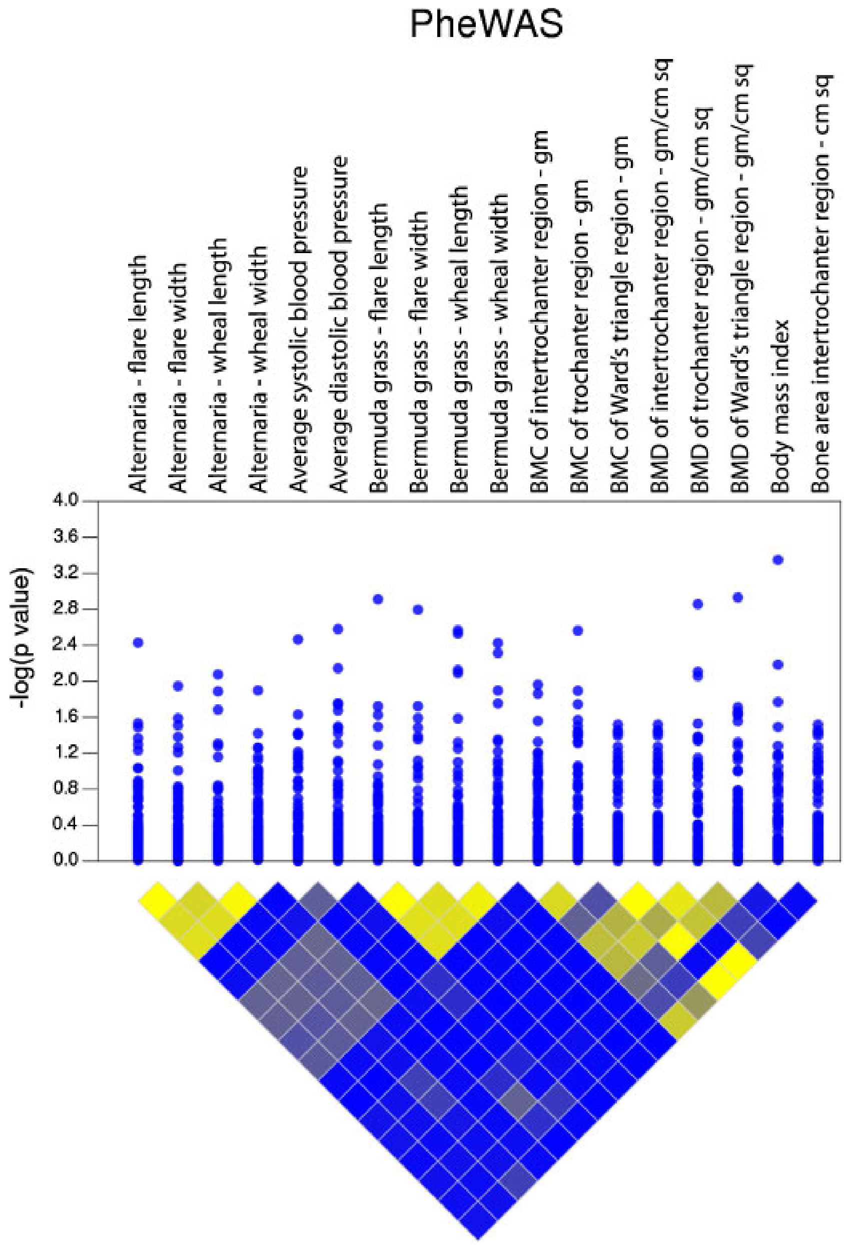

Fig. 5. Correlation Heat Map from EAGLE phenotypic and PheWAS data. Phenotypes are along the *x*-axis. The middle track shows the −log10(*P*-value) results for each genotype-phenotype test of association. Below the *P*-value significance track, a heat map of correlation results is plotted. Pearson correlations between quantitative phenotypes are calculated and the absolute value of each correlation is plotted. Relative significance is displayed by the shade of the color (that is, the more significant the correlation, the brighter yellow the cell). In this figure, bone mineral content and bone mineral density measurements are correlated, along with body mass index.

together in a high-throughput fashion, truly allowing for the investigation of the larger structure of and interrelationships between phenotypic and genotypic diversity.

Throughout the initial development of PheWAS in PAGE, approaches are being developed for both grouping phenotypes, as well as characterizing phenotypic structure. Thus, there is current work within PAGE to use computational algorithms to bring together related phenotypes across studies of PAGE. As mentioned previously, several known phenotype relationships exist between the phenotypes, intermediate phenotypes, sub-phenotypes, biomarkers, and endophenotypes that exist within the studies of PAGE. One approach for characterizing some of the relationships between phenotypes within these data is the possibility of visualizing correlations between continuous phenotypes. Another approach is to group related phenotypes into categories. This structure can then be related back to significant genotype-phenotype associations, potentially providing more information about the larger relationship between genetic variation and phenotypic variability. An additional step is to link the phenotypes of PAGE to specific phenotypic domains and data resources that have been developed that characterize phenotype relationships, as well as resources that summarize and monitor published GWAS results [Hindorff et al., 2009]. Through linking to these resources that include automated text mining and data summation of what has been published to date, further phenotypic structure and phenotype-genotype relationships can provide additional information when interpreting and delving into PheWAS results. Some approaches for mapping these phenotype networks are starting to be discussed and developed in the field of phenomics [Lanktree et al., 2010; Rzhetsky et al., 2007].

STRENGTHS AND LIMITATIONS OF THE PheWAS APPROACH

The main strength of this approach is the potential to make novel discoveries and to identify new relationships between phenotypes and genetic variation, different from the more common approach of examining the relationship between one or a handful of disease outcomes/phenotypes and genetic variation. In addition, PheWAS analyses include known genotype-phenotype associations in addition to novel associations, and these known associations serve as "positive controls" for this high-throughput analysis. Another strength of this PAGE-PheWAS is the diversity of the PAGE studies, as both novel and known genotype-phenotype associations can be examined across multiple racial/ethnic groups for generalization. As this work is being performed in parallel across multiple groups, as long as both the genotype and phenotype exist in more than one study, replication of associations across PAGE studies is possible. Heterogeneity in how some phenotypes across PAGE studies are measured is a potential limitation for the PAGE-PheWAS. However, PheWAS is intended to be exploratory and significant results of interest can be investigated further in future work with careful phenotype harmonization.

As previously mentioned, a major limitation of PheWAS is the potential for identifying false-positive associations, as thousands of genotype-phenotype relationships are being investigated. Sample sizes do differ between the various studies of PAGE, which can impact power and the detection of replication across studies. Another limitation of the PAGE PheWAS analysis is related to the SNP selection process. While the number of SNPs genotyped in Year 1 of PAGE is relatively large (currently ~400, and numbers will increase over time as more SNPs are genotyped), the genetic variation studied is much more limited than that of GWAS. The SNPs used in PAGE PheWAS are well-characterized SNPs, but are selected because of their well-replicated previous association with specific phenotypes (for example, LDL cholesterol-associated SNPs). One way to increase the SNP coverage is to perform a GWAS with an extremely rich phenotype data set (PheWAS-GWAS). While this would add additional complexity to the PheWAS analysis, this would truly be the optimal way to investigate both a comprehensive set of SNPs and phenotypes.

CONCLUSIONS

The steps and workflow described here have been developed and defined from the initial work on the PAGE PheWAS project. PheWAS is a novel resource for mining phenotypically rich data sets to provide insight on the interrelationships between genetic variation and networks of phenotypes. The nature of investigating a large number of phenotypes in PheWAS, ranging from outcome phenotypes, such as disease status to intermediate phenotypes, biomarkers, quantitative traits, and risk factors, has the potential for defining a more complete picture of genetic variation and disease, generating hypotheses, and identifying pleiotropy.

ACKNOWLEDGMENTS

The Population Architecture Using Genomics and Epidemiology (PAGE) program is funded by the National Human Genome Research Institute (NHGRI), supported by U01HG004803 (CALiCo), U01HG004798 (EAGLE), U01HG004802 (MEC), U01HG004790 (WHI), and U01HG004801 (Coordinating Center). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The complete list of PAGE members can be found at http://www.pagestudy.org.

The data and materials included in this report result from a collaboration between the following studies:

The "Epidemiologic Architecture for Genes Linked to Environment (EAGLE)" is funded through the NHGRI PAGE program (U01HG004798-01). The study participants derive from the National Health and Nutrition Examination Surveys (NHANES), and these studies are supported by the Centers for Disease Control and Prevention. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work. The EAGLE/NHANES DNA samples are stored and plated by the Vanderbilt DNA Resources Core. Genotyping was performed by Ping Mayo, Melissa Allen, and Dr. Nathalie Schnetz-Boutaud in the laboratory of Dr. Jonathan Haines and Hailing Jin and Nila Gillani under the direction of Dr. Holli Dilks in the Vanderbilt DNA Resources Core.

The Multiethnic Cohort study (MEC) characterization of epidemiological architecture is funded through the NHGRI PAGE program (U01HG004802). The MEC study is funded through the National Cancer Institute (R37CA54281, R01 CA63, P01CA33619, U01CA136792, and U01CA98758).

Funding support for the "Epidemiology of putative genetic variants: The Women's Health Initiative" study is provided through the NHGRI PAGE program (U01HG004790). The WHI program is funded by the National Heart, Lung, and Blood Institute; NIH; and U.S. Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-32119, 32122, 42107-26, 42129-32, and 44221. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whiscience.org/publications/WHI_investigators_shortlist.pdf.

Funding support for the Genetic Epidemiology of Causal Variants Across the Life Course (CALiCo) program was provided through the NHGRI PAGE program (U01HG004803). The following studies contributed to this manuscript and are funded by the following agencies: The Atherosclerosis Risk in Communities (ARIC) Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, N01-HC-55022. The Coronary Artery Risk Development in Young Adults (CARDIA) study is supported by the following National Institutes of Health, National Heart, Lung and Blood Institute contracts: N01-HC-95095; N01-HC-48047; N01-HC-48048; N01-HC-48049; N01-HC-48050; N01-HC-45134; N01-HC-05187; and N01-HC-45205. The Cardiovascular Health Study (**CHS**) is supported by contracts N01-HC-85079 through N01-HC-85086, N01-HC-35129, N01-HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133, grants U01HL080295 and R01 HL087652 from the National Heart, Lung, and Blood Institute, with additional contribution from the National Institute of Neurological Disorders and Stroke. CHS GWAS DNA handling and genotyping was supported in part by National Center for Research Resources grant M01-RR00425 to the Cedars-Sinai General Clinical Research Center Genotyping core and National Institute of Diabetes and Digestive and Kidney Diseases grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) is supported by NHLBI contracts N01-HC65233, N01-HC65234, N01-HC65235, N01-HC65236, and N01-HC65237. The following Institutes/Centers/Offices contribute to the HCHS/SOL through a transfer of funds to the NHLBI: National Center on Minority Health and Health Disparities, the National Institute of Deafness and Other Communications Disorders, the National Institute of Dental and Craniofacial Research, the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Neurological Disorders and Stroke, and the Office of Dietary Supplements. The Strong Heart Study (SHS) is supported by NHLBI grants U01 HL65520, U01 HL41642, U01 HL41652, U01 HL41654, and U01 HL65521. The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Indian Health Service.

Assistance with data management, integration and dissemination, and general study coordination was provided by the PAGE Coordinating Center (U01HG004801-01). The National Institutes of Mental Health also contributes to the support for the Coordinating Center.

REFERENCES

Bilder RM, Sabb FW, Cannon TD, London ED, Jentsch JD, Parker DS, Poldrack RA, Evans C, Freimer NB. 2009. Phenomics: the systematic study of phenotypes on a genome-wide scale. Neuroscience 164:30–42.

CDC. 2008. National Health and Nutrition Examination Survey (NHANES) Stored Biologic Specimens: Guidelines for Proposals to Use Samples and Proposed Cost Schedule. Fed Regist 73:51487–51489.

Cupples LA, Arruda HT, Benjamin EJ, D'Agostino Sr RB, Demissie S, DeStefano AL, Dupuis J, Falls KM, Fox CS, Gottlieb DJ, Govindaraju DR, Guo CY, Heard-Costa NL, Hwang SJ, Kathiresan S, kiel DP, Laramie JM, Larson MG, Levy D, Liu CY, Lunetta KL, Mailman MD, Manning AK, Meigs JB, Murabito JM, Newton-Cheh C, O'Connor GT, O'Donnell CJ, Pandy M, Seshadri S, Vasan RS, Wang ZY, WIlk JB, Wolf PA, Yang Q, Atwood LD. 2007. The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. BMC Med Genet 8:S1.

Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26:1205–1210.

Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, O'Leary D, Psaty B, Rautaharju P, Tracy R, Weller P. 1991. The Cardiovascular Health Study: design and rationale. Ann Epidemiol 1:263–276.

Ghebranious N, McCarty CA, Wilke RA. 2007. Clinical phenome scanning. Personalized Med 4:175–182.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367.

Hughes GH, Cutter G, Donahue R, Friedman GD, Hulley S, Hunkeler E, Jacobs Jr DR, Liu K, Orden S, Pirie P, Tucker B, Wagenknecht L. 1987. Recruitment in the coronary artery disease risk development in young adults (cardia) study. Control Clin Trials 8:68S–73S.

Jones R, Pembrey M, Golding J, Herrick D. 2005. The search for genenotype/phenotype associations and the phenome scan. Paediatr Perinat Epidemiol 19:264–275.

Kolonel LN, Henderson BE, Hankin JH, Nomura AM, Wilkens LR, Pike MC, Stram DO, Monroe KR, Earle ME, Nagamine FS. 2000. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. Am J Epidemiol 151:346–357.

Lanktree MB, Hassell RG, Lahiry P, Hegele RA. 2010. Phenomics: expanding the role of clinical evaluation in genomic studies. J Investig Med 58:700–706.

Lee ET, Welty TK, Fabsitz R, Cowan LD, Le NA, Oopik AJ, Cucchiara AJ, Savage PJ, Howard BV. 1990. The Strong Heart Study. A study of cardiovascular disease in American Indians: design and methods. Am J Epidemiol 132:1141–1155.

Lou H, Yeager M, Li H, Bosquet JG, Hayes RB, Orr N, Yu K, Hutchinson A, Jacobs KB, Kraft P, et al. 2009. Fine mapping and functional analysis of a common variant in MSMB on chromosome 10q11.2 associated with prostate cancer susceptibility. Proc Natl Acad Sci USA 106:7933–7938.

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, et al. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39: 1181–1186.

Matise T, Ambite JL, Buyske S, Carlson C, Cole S, Crawford D, Haiman C, Heiss G, Kooperberg C, Le Marchand L, Manolio T, North K, Pters U, Ritchie M, Hindorff L, Haines J. 2010. The next PAGE in understanding complex traits: study design for analysis of population architecture using genetics and epidemiology. American Journal of Epidemiology, in press.

McQuillan GM, Porter KS, Agelli M, Kington R. 2003. Consent for genetic research in a general population: the NHANES experience. Genet Med 5:35–42.

Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, et al. 2007. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448:470–473.

Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM, et al. 2010.

From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466:714–719.

Rice TK, Schork NJ, Rao DC. 2008. Methods for handling multiple testing. Adv Genet 60:293–308.

Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, Basford MA, Brown-Gentry K, Balser JR, Masys DR, et al. 2010. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 86:560–572.

Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. 2008. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 84:362–369.

Rzhetsky A, Wajngurt D, Park N, Zheng T. 2007. Probing genetic overlap among complex human phenotypes. Proc Natl Acad Sci USA 104:11694–11699.

The ARIC investigators. 1989. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. Am J Epidemiol 129: 687–702.

The Women's Health Initiative Study Group. 1998. Design of the Women's Health Initiative clinical trial and observational study. Control Clin Trials 19:61–109.