ENABLING HIGH-THROUGHPUT GENOTYPE-PHENOTYPE ASSOCIATIONS IN THE EPIDEMIOLOGIC ARCHITECTURE FOR GENES LINKED TO ENVIRONMENT (EAGLE) PROJECT AS PART OF THE POPULATION ARCHITECTURE USING GENOMICS AND EPIDEMIOLOGY (PAGE) STUDY

WILLIAM S. BUSH[*]

Department of Biomedical Informatics, Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: william.s.bush@vanderbilt.edu

JONATHAN BOSTON*

Center for Human Genetics Research, Vanderbilt University, 1207 17th Avenue, Suite 300
Nashville, TN 37232, USA
Email: boston@chgr.mc.vanderbilt.edu

SARAH A. PENDERGRASS

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 503 Wartik Lab
University Park, PA 16802, USA
Email: sap29@psu.edu

LOGAN DUMITRESCU, ROBERT GOODLOE

Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: logan.dumitrescu@chgr.mc.vanderbilt.edu; robert.j.goodloe@vanderbilt.edu

KRISTIN BROWN-GENTRY, SARAH WILSON, BOB MCCLELLAN, JR

Center for Human Genetics Research, Vanderbilt University, 1207 17th Avenue, Suite 300
Nashville, TN 37232, USA
Email: kristin.brown@chgr.mc.vanderbilt.edu; sarah.wilson@chgr.mc.vanderbilt.edu; bob.mcclellan@chgr.mc.vanderbilt.edu

ERIC TORSTENSON

Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: eric.torstenson@chgr.mc.vanderbilt.edu

[*] Contributed equally to the work

MELISSA A. BASFORD

Office of Research, Office of Personalized Medicine, Vanderbilt University, 2525 West End Avenue
Nashville, TN 37203, USA
Email: melissa.basford@vanderbilt.edu

KYLEE L. SPENCER

Biology and Environmental Science, Heidelberg University, Bareis Hall 131, 310 East Market Street
Tiffin, OH 44883, USA
Email: kspencer@heidelberg.edu

MARYLYN D. RITCHIE

Center for System Genomics, Department of Biochemistry and Molecular Biology, , Pennsylvania State University,
512 Wartik Lab
University Park, PA 16802, USA
Email: marylyn.ritchie@psu.edu

DANA C. CRAWFORD

Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University,
2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: crawford@chgr.mc.vanderbilt.edu

Genetic association studies have rapidly become a major tool for identifying the genetic basis of common human diseases. The advent of cost-effective genotyping coupled with large collections of samples linked to clinical outcomes and quantitative traits now make it possible to systematically characterize genotype-phenotype relationships in diverse populations and extensive datasets. To capitalize on these advancements, the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) project, as part of the collaborative Population Architecture using Genomics and Epidemiology (PAGE) study, accesses two collections: the National Health and Nutrition Examination Surveys (NHANES) and BioVU, Vanderbilt University's biorepository linked to de-identified electronic medical records. We describe herein the workflows for accessing and using the epidemiologic (NHANES) and clinical (BioVU) collections, where each workflow has been customized to reflect the content and data access limitations of each respective source. We also describe the process by which these data are generated, standardized, and shared for meta-analysis among the PAGE study sites. As a specific example of the use of BioVU, we describe the data mining efforts to define cases and controls for genetic association studies of common cancers in PAGE. Collectively, the efforts described here are a generalized outline for many of the successful approaches that can be used in the era of high-throughput genotype-phenotype associations for moving biomedical discovery forward to new frontiers of data generation and analysis.

1. Introduction

In a typical genome-wide association study (GWAS), a single or limited number of traits or diseases are tested for association with common single nucleotide polymorphisms (SNPs) assayed regardless of presumed function across the human genome. Since 2005, GWAS has been successful in confirming already known and identifying novel genotype-phenotype associations relevant to the biomedical community. GWAS is now a mainstay discovery approach in human genetics.

With hundreds to thousands of genotype-phenotype associations now catalogued across the human genome(1,2), there is great interest in expanding the characterization of these associations beyond the initial population or phenotype studied. Indeed, the systematic characterization and fine-mapping of known GWAS-identified variants from European-descent populations has begun in earnest(3-10). In addition, large scale methods to identify pleiotropy, such as phenome-wide association studies (PheWAS) (11,12), are increasing in frequency. To propel research in these two avenues, the National Human Genome Research Institute founded the Population Architecture using Genomics and Epidemiology (PAGE) study in 2008. PAGE is a collection of large, diverse epidemiologic and clinical collections with DNA samples linked to hundreds of disease outcomes, quantitative traits, and exposures(13)(Figure 1). A major activity of the PAGE study is the systematic characterization of GWAS-identified genotype-phenotype relationships across populations and phenotypes. The Epidemiologic Architecture for Genes Linked to Environment (EAGLE) project, one of PAGE's four study sites, accesses the National Health and Nutrition Examination Surveys (NHANES) and the Vanderbilt University biorepository linked to de-identified electronic medical records (BioVU)(14) to pursue PAGE study goals.

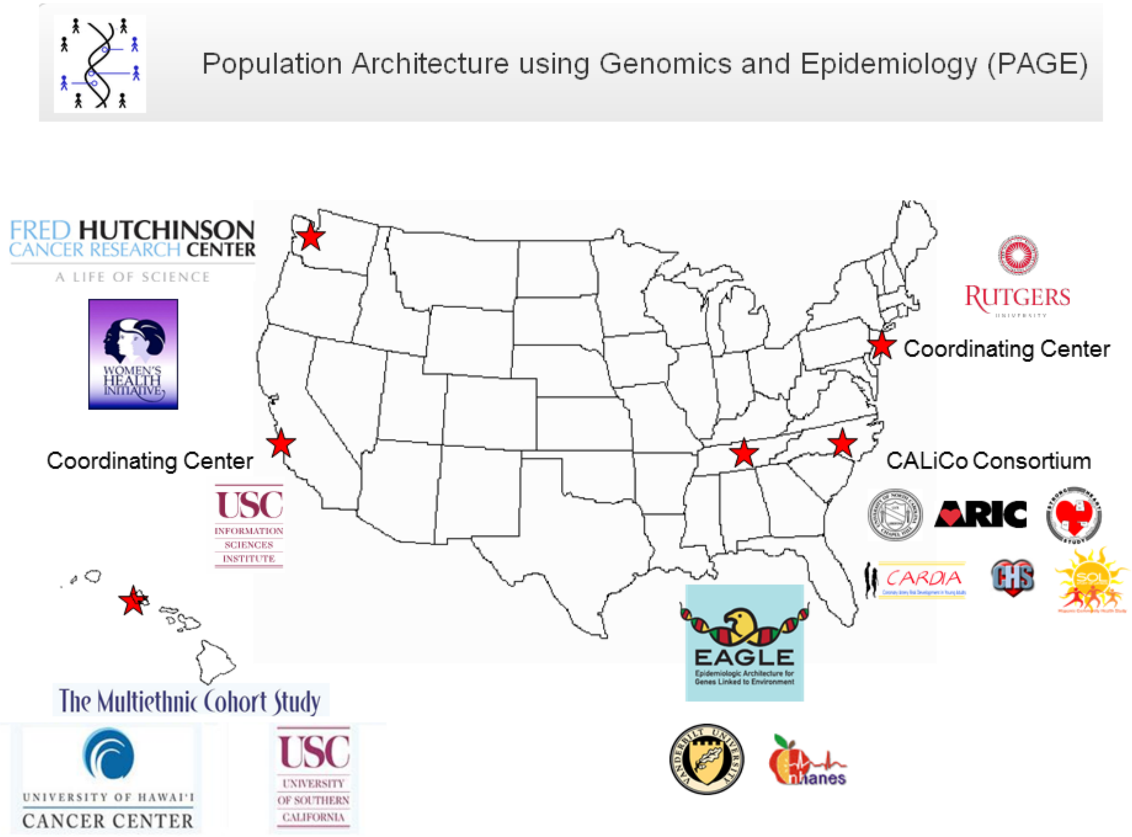
EAGLE participates in collaborative PAGE studies for disease and traits related to cardiovascular, metabolic, and cancer phenotypes among many others. To enable characterization of genotype-phenotype relationships in EAGLE and PAGE, EAGLE has developed high-throughput workflows customized to test GWAS-identified variants for all outcomes and traits in multiple populations available in both EAGLE collections. The development of a systematic workflow was and continues to be necessary to harmonize EAGLE analyses with analyses from other PAGE study sites and to facilitate meta-analysis across multiple studies. We describe herein each EAGLE collection, including characteristics of each data collection that impact both the workflow design for effective data analysis as well as data sharing, all crucial elements for collaborative high-throughput human genetic association studies for biomedical discovery.

2. Methods

1.1. *Study populations*

EAGLE currently accesses two diverse study populations as part of the PAGE study: the National Health and Nutrition Examination Surveys (NHANES) and BioVU, the Vanderbilt University biorepository linked to de-identified electronic medical records (EMRs). NHANES is a population-based survey conducted by the National Center for Health Statistics at the Centers for Disease Control and Prevention(15). NHANES ascertains Americans regardless of health status at the time of the survey. For each study participant, data on demographics, health, and lifestyle are collected. A physical exam is conducted by a CDC physician or health professional, and laboratory measures are assayed from blood and urine. DNA samples were collected on consenting participants for the Third NHANES (NHANES III) conducted between 1991 and 1994 (n=7,159), NHANES 1999-2000 (n=3,570), NHANES 2001-2002 (n=4,269), and NHANES 2007-2008 (n=4,615). A total of 19,613 DNA samples are available for research representing self-reported non-Hispanic whites (n=8,858), non-Hispanic blacks (n=4,325), Mexican Americans (n=4,768), and other race/ethnicities (n=1,662).

Figure 1. The Population Architecture using Genomics and Epidemiology (PAGE) study. The PAGE study, funded in 2008, consists of a coordinating center (Rutgers University and Information Sciences Institute at the University of Southern California) and four study sites: the Causal Variants Across the Life Course (CALiCo) consortium accessing the Atherosclerosis Risk in Communities (ARIC), Coronary Artery Risk in Young Adults (CARDIA), Cardiovascular Heart Study (CHS), Strong Heart Cohort and Family Studies (SHS/SHFS), and Study of Latinos (SOL); Epidemiologic Architecture for Genes Linked to Environment (EAGLE) accessing the National Health and Nutrition Examination Surveys (NHANES) and Vanderbilt University's biorepository linked to de-identified medical records (BioVU); the Multiethnic Cohort (MEC); and the Women's Health Initiative (WHI).


Population Architecture using Genomics and Epidemiology (PAGE)
FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE
WOMEN'S HEALTH INITIATIVE
Coordinating Center
USC INFORMATION SCIENCES INSTITUTE
RUTGERS UNIVERSITY
Coordinating Center
CALiCo Consortium
ARIC
CARDIA
CHS
SOL
EAGLE
Epidemiologic Architecture for Genes Linked to Environment
The Multiethnic Cohort Study
UNIVERSITY OF HAWAI'I CANCER CENTER
USC UNIVERSITY OF SOUTHERN CALIFORNIA
VANDERBILT UNIVERSITY
nhanes

In contrast to NHANES, BioVU is a clinic-based collection of patients visiting the outpatient clinics affiliated with Vanderbilt University in Nashville, Tennessee(14). DNA is extracted from discarded blood collected for routine outpatient clinic use and linked to a de-identified version of the electronic medical record known as the Synthetic Derivative (SD). The SD is updated routinely and contains outpatient as well as inpatient clinical structured and unstructured data including billing codes, procedure codes, labs, tumor registry entries, demographic data, vital signs, and text-based clinical notes. Because of extensive de-identification procedures, BioVU is considered non-Human Subjects research(16). As of June 2012, BioVU contained 143,993 DNA samples, 57% of which are from females and 10% from African Americans.

2.2. Genotyping

The majority of EAGLE's genotypic data are a result of *de novo* targeted genotyping. Briefly, SNPs were selected in 2008 to mid-2010 representing index genetic variants from GWAS of common diseases and traits such as HDL-C, LDL-C, triglycerides, total cholesterol, markers of inflammation, bone mineral density/osteoporosis, electrocardiographic traits, body mass index, complete blood count traits, type 2 diabetes and eight major cancers. SNPs were then genotyped using a variety of assays/platforms including TaqMan, TaqMan OpenArray, Illumina BeadXpress, and Sequenom. To date, EAGLE has submitted greater than 5.1 million genotypes to the CDC Genetic NHANES database, and these data are available for secondary analyses via NCHS/CDC.

2.3. Statistical analyses

In EAGLE (single site) and PAGE (multi-site) studies, genotype-phenotype association analyses are conducted as defined by the following "tiers"(13):

- Tier 1: High-throughput unadjusted linear or logistic regressions assuming an additive genetic model. For categorical phenotypes, binning was used to create new variables of the form "A versus not A" for each category, and logistic regression was used to model the new binary variable. All continuous phenotypes were natural log transformed, following a y to log (y+1) transformation of the response variable with +1 added to all continuous measurements before transformation to prevent variables recorded as zero from being omitted from analysis. All analyses are stratified by race/ethnicity. Statistical analyses are performed by each PAGE study site independently. The phenotypic and exposure variables are not harmonized across PAGE study sites.
- Tier 2: Low throughput unadjusted linear and/or logistic regressions performed for select genotypes and phenotypes of interest in a single PAGE study site. The genetic modeling and levels of stratification are dependent on a specific hypothesis or study question. The study subjects are carefully phenotyped and multiple covariates (also well-defined) are considered in the models.
- Tier 3: Low throughput unadjusted linear and/or logistic regressions performed for select genotypes and phenotypes of interest across PAGE study sites where the genetic modeling and levels of stratification are dependent on the hypothesis or study question. The study subjects are carefully phenotyped like Tier 2 analyses; however for Tier 3, phenotypes and exposures are harmonized across multiple PAGE study sites. Statistical analyses are performed by each PAGE study site independently, and aggregate results are shared across study sites for meta-analysis by the lead author(s).

All PAGE study results, regardless of Tier, must be available in aggregate form for the PAGE Coordinating Center browser(13) and possible dbGaP(17) deposition. To facilitate the uniform submission of PAGE study aggregate data by study site, the PAGE Coordinating Center created three "Results Template" files consisting of the phenotype file, the SNP file, and the Association file (version 8). The phenotype file currently consists of 32 column headers such as phenotype label, PAGE study site, phenotype units, information on transformation and analysis tier, type of variable (binary versus quantitative), types of covariates included in the models, race/ethnicity, gender, sample size, and descriptive statistics of the phenotype used in the analysis. The SNP files currently consists of 19 column headers such SNP ID (rs number), PAGE study site, race/ethnicity, gender, alleles and counts (including coded allele designation), genotypes and counts, Hardy Weinberg p-values, genotype call rates, and strand information. The Association file currently consists of 53 column headers such as SNP ID, phenotype, PAGE study site, race/ethnicity, gender, genetic effect size of association and standard errors and/or confidence intervals, modeling label (defined by lead of the analysis plan), p-values, sample sizes, alleles (included allele and frequency of coded allele), genotype counts by affection status, median values and quartiles of quantitative traits by genotype, and genetic model.

In EAGLE, all NHANES genotype-phenotype associations are performed using SAS v9.2 and SUDAAN v10.0(SAS Institute, Cary, NC) using the Analytic Data Research by Email (ANDRE) portal of the CDC Research Data Center (RDC) in Hyattsville, MD (further described below). EAGLE analyses accessing BioVU data are performed using a variety of software packages including PLINKv1.07(18), SASv9.3, and Rv2.14.1(19). The EAGLE workflows described here are supported by multiple scripts written in several computer languages such Ruby with Ruby on Rails framework and Javascript with Backbone framework.

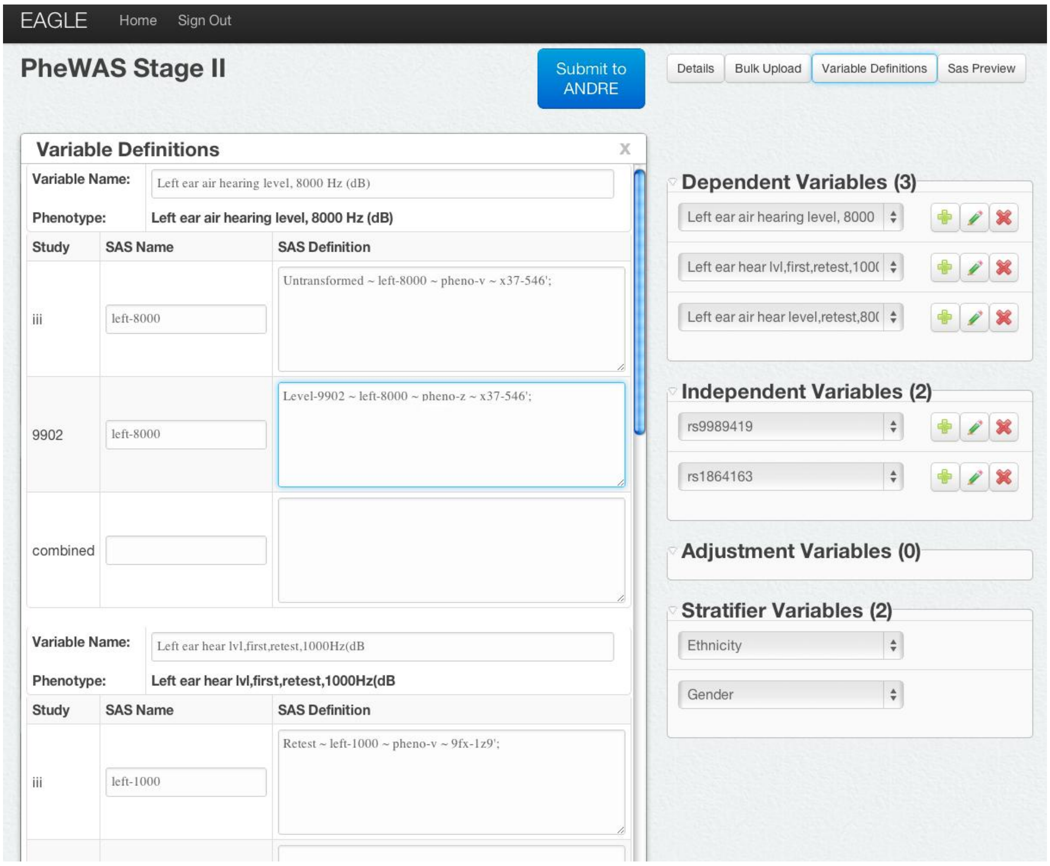
3. Workflow

3.1. The epidemiologic collection (NHANES)

Like many epidemiologic collections, NHANES consists of thousands of DNA samples linked to thousands of variables and, in the case of EAGLE, hundreds of genetic variants. To automate the high-throughput genotype-phenotype associations such as the PheWAS approach, the workflow for this and many epidemiologic collections must accommodate the fact that sample size, phenotypic/exposure variable list, and genetic variant content can vary substantially across the years of survey. Also, the workflow must acknowledge and work with various data access models that can range from open access to highly restricted access to individual level data within and across collaborating studies. Finally, the workflow must anticipate high volumes of structured data that will require accessible archival or storage for specialized searches.

Specifically for NHANES, EAGLE accesses up to 19,613 DNA samples that have anywhere from one to 1,100 genetic variants and approximately 3,500 phenotypic/demographic variables available for analysis. Due to concerns related to confidentiality even for aggregate data(20), genetic data are considered restricted variables by CDC and therefore cannot be linked to phenotypic variables and accessed outside of the CDC RDC firewall. To facilitate analyses such as genotype-phenotype association studies for research groups outside of CDC, the RDC created Analytic Data Research by Email (ANDRE). ANDRE is the remote server for CDC that accepts and runs analyses generated in Statistical Analysis System (SAS) or Survey Data Analysis (SUDAAN). ANDRE is an e-mail exchange that serves as an interface for processing code. Only analyses or SAS commands that result in aggregate data are allowed, and specific SAS commands and macros are explicitly forbidden. SAS output resulting from analyses sent to ANDRE by outside investigators are further inspected to ensure that counts fewer than five are redacted or suppressed from the output before the output is returned to outside investigators for consumption. And, ANDRE e-mail exchange is limited to outgoing files <20MB in size, which includes both the log and output files. The time elapsed between submitting code to ANDRE and receiving the output files from ANDRE via e-mail is typically less than 30 minutes, but this can range from two minutes to several hours.

Figure 2. EAGLE project web-based Experiment Designer. We developed a web-based Experiment Designer to assist EAGLE analysts in generating standard SAS code for high-throughput genotype-phenotype tests of association. The SAS designer allows each EAGLE analyst to create experiments by selecting pre-defined variables approved for study by CDC by NHANES dataset. EAGLE analysts can also specify dependent variables, independent variables, and stratification variables (gender and race/ethnicity) for linear or logistic regression modeling. The SAS Generator takes the experiment created with the Experiment Designer and generates the appropriate SAS code for submission to ANDRE.



PheWAS Stage II

Variable Definitions

Variable Name: Left ear air hearing level, 8000 Hz (dB)

Phenotype: Left ear air hearing level, 8000 Hz (dB)

Study	SAS Name	SAS Definition
iii	left-8000	Untransformed ~ left-8000 ~ pheno-v ~ x37-546';
9902	left-8000	Level-9902 ~ left-8000 ~ pheno-z ~ x37-546';
combined		

Variable Name: Left ear hear lvl,first,retest,1000Hz(dB

Phenotype: Left ear hear lvl,first,retest,1000Hz(dB

Study	SAS Name	SAS Definition
iii	left-1000	Retest ~ left-1000 ~ pheno-v ~ 9fx-1z9';

Dependent Variables (3)

- Left ear air hearing level, 8000
- Left ear hear lvl,first,retest,100(
- Left ear air hear level,retest,80(

Independent Variables (2)

- rs9989419
- rs1864163

Adjustment Variables (0)

Stratifier Variables (2)

- Ethnicity
- Gender

The restrictions posed by the RDC present several challenges for high-throughput genotype-phenotype associations in EAGLE and for data sharing with the PAGE study sites. To work within the restrictions and to minimize analyst workload, we created a web-based "Experiment Designer" and "SAS Generator". With the Experiment Designer (Figure 2), analysts create and edit the variables for an experiment that will be sent to ANDRE. Analysts can then select dependent and independent variables along with any adjustments and stratifications. The Experiment Designer allows analysts to focus on the data and desired results instead of the SAS code itself. The Experiment Designer also ensures uniform SAS coding of the genetic model (and coded allele), an important feature for large datasets accessed by three analysts at any one time. The SAS Generator then takes the experiment created with the Experiment Designer and generates the appropriate SAS code for submission to ANDRE. Each experiment can be queued and sent to ANDRE when output from the previous experiment is received by the analyst via e-mail. Thus, the SAS Generator ensures that there are no gaps between sending SAS code and receiving output from ANDRE. The SAS Generator ultimately saves the analyst time from constantly checking e-mail for receipt of ANDRE output. To date, EAGLE analyses for EAGLE and PAGE study analysis plans have generated >400 experiments resulting in >20,000,000 SAS output files each with approximately 50 lines of unstructured SAS data output.

Most tests of association performed in NHANES result in tens of thousands of SAS output files from ANDRE. With so many output files and lines of data per output file, a second major challenge is translating the output into a condensed, accessible, and readily available format. For each set of output we have developed the "Parser" software to do the following: 1) parse the file headers to classify the files (e.g. Linear Regression, SNP Frequency, etc), and 2) process the text of each SAS output file and extract the appropriate data values. The Parser can be utilized only when necessary, allowing EAGLE analysts to store the SAS output files and then process them in real-time, as needed. This also allows EAGLE analysts to view any single output file and also view the parsed results.

Once the SAS output file results are parsed, the data are compiled into the PAGE Coordinating Center Results Template file format. To automate this process, we created the "Template Generator" step. In this step, an experiment's SAS output files are parsed and combined into a template for submission to the PAGE Coordinating Center and to PAGE collaborators for meta-analysis or for visualization using Synthesis-View(21), PheWAS-View(22), or other software. Automation of this step results in analysis results required for meta-analysis or dbGaP submission.

The full epidemiologic workflow for EAGLE, from SAS code generation to Results Template file generation for data dissemination, is given in Figure 3. The code is open source and will be available on the EAGLE website (https://eagle.mc.vanderbilt.edu/).

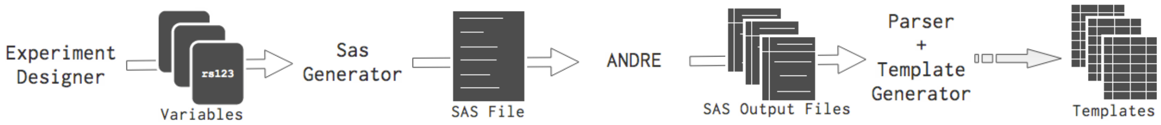


Figure 3. EAGLE project epidemiologic collection workflow. The epidemiologic collection workflow begins with the Experiment Designer, designed as a web-interface and accessed by EAGLE analysts. The analyst can easily use the Experiment Designer to create standardized SAS code based on parameters set by the analysts. The resultant ANDRE-friendly code is automatically generated. Once the code has been submitted, ANDRE will send censored output files back to the EAGLE analysts. These resultant files are first crudely parsed and stored in a database in preparation for "real-time" parsing by analysts. Finally, analysts use the "Template Generator" to create standard PAGE Results Template files for sharing data across PAGE study sites for meta-analysis.

3.2. The clinical collection (BioVU)

The epidemiologic collection of NHANES described above is an extensive and rich source of phenotypic and genotypic data for genetic association studies of quantitative traits; however, because of the wide age range and lack of health information for specific diseases, the collection is underpowered for many diseases, including common diseases such as cardiovascular disease, type 2 diabetes, and various cancers. To supplement EAGLE sample sizes for clinical outcomes in diverse populations, a clinical collection at Vanderbilt University known as BioVU was accessed.

Additional cancer cases and controls were first identified in BioVU using billing (ICD-9) codes. Specific cancers such as melanoma could be defined with high positive predictive values whereas others such as endometrial cancer could not. Therefore, to increase the positive predictive value of all EAGLE case/control definitions, data from the tumor registry were utilized. These data include primary site designations and histology information collected for clinical reporting purposes for the North America Association of Central Cancer Registries. A combination of the tumor registry data, along with ICD-9 billing codes, procedure codes, vital signs, and free text clinical notes, were used to identify cases for eight cancers among all patients aged 18 or greater in the SD with DNA samples using the following algorithms:

- Breast cancer: Three or more mentions of ICD-9 primary code 174 (malignant neoplasm of the female breast) and all sub-codes (denoted "*" here and throughout) on separate clinic visits OR a tumor registry entry for breast cancer AND female
- Colorectal cancer: Tumor registry entry for colorectal cancer.
- Endometrial cancer: Tumor registry entry for endometrial cancer with primary sites C540-C549, C559 AND histology not one of 9590-9989 AND female.

- Lung cancer: Tumor registry entry for lung cancer, any location and any type.
- Melanoma: Three or more mentions of ICD-9 codes 172.* (malignant melanoma of skin) OR tumor registry entry for melanoma.
- Non-Hodgkin's lymphoma: Tumor registry entry for non-Hodgkin's lymphoma with histology in ('9673', '9675', '9684', '9687', '9695', '9705', '9823', '9827'), OR (histology >= '9590' and histology <= '9596'), OR (histology >= '9670' and histology <= '9671'), OR (histology >= '9678' and histology <= '9680'), OR (histology >= '9689' and histology <= '9691'), OR (histology >= '9698' and histology <= '9702'), OR (histology >= '9708' and histology <= '9709'), OR (histology >= '9714' and histology <= '9719'), OR (histology >= '9727' and histology <= '9729').
- Ovarian cancer: Tumor registry entry for ovarian cancer AND female.
- Prostate cancer: Three or more mentions of ICD-9 codes 185.* (malignant neoplasm of prostate) OR tumor registry entry for prostate cancer.

Approximately two control samples were identified per case matched on sex, race/ethnicity, and age (within 5 years). Control samples were required to have at least two clinical narratives (clinical notes, discharge summaries, etc), with preference given to records with at least one fully documented history and physical. Records were excluded as controls if they had one or more codes for neoplasms, ICD-9 codes between 140.* and 239.*, had a tumor registry entry or had the one or more cancer related keywords in the problem list. For breast cancer, endometrial cancer, and ovarian cancer, male controls were also excluded, and for prostate cancer, female controls were excluded.

For specific cancers, controls with additional clinical data were desirable for anticipated analyses. For example, for breast cancer controls among women over 40 years of age, we required that records contain at least one mammography Bi-Rad score as 1 (negative) or 2 (benign). For colorectal cancer controls, we required for patients over 50 years of age the keyword "colonoscopy" in the problem list OR one of the following CPT codes: 45378 (colonoscopy, flexible, proximal to splenic flexure, diagnostic), 45379 (with removal), 45380 (with biopsy, single), 45381 (with directed), 45382 (with control), 45383 (with ablation of), 45384 (with removal of), 45385 (with removal of), 45386 (with dilation by), 45387 (with transendoscopic), 45391 (with endoscopic), and 45392 (with transendoscopic). Finally, for prostate cancer, we required male controls aged 40 years and greater to have at least one prostate specific antigen (PSA) level <4 and that the most recent PSA level is within the normal range.

With these algorithms implemented in the SD in late 2010/early 2011, we identified a total of 7,348 cancer cases for targeted genotyping. Race/ethnicity in the Vanderbilt University EMR and BioVU SD is administratively assigned, which we have shown is highly concordance with genetic ancestry determined by ancestry informative markers (AIMs)(23). As expected based on the overall demographics of BioVU, the majority of case samples were European American (87%). Approximately 4% of the samples were of unknown race/ethnicity and were assigned genetic ancestry via ancestry informative markers for downstream analyses (data not shown). For the first five cancers defined in the SD (breast, colorectal, melanoma, ovarian, and prostate cancers) we identified approximately two controls per case for genotyping as defined in the text above. A total of 8,996 controls were targeted for genotyping. Two controls per case of endometrial cancer, lung cancer, and non-Hodgkin's lymphoma were defined from among the genotyped control samples.

In addition to defining case and control status for genotyping, we have begun to define clinical covariates anticipated for analysis. As described above, screening data has been preferentially represented in controls for select cancers (breast, colorectal, and prostate) and is expected to be defined in cases. Environmental exposures are more difficult to define given that most of these data, if available, exist in the unstructured data (free text or clinical narrative) of the EMR. Work is on-going to define common exposures or other variables that reside in the clinical narrative such as alcohol use, physical activity, and family history using text mining and other approaches. For smoking status, we have applied an implementation of the CTAKES algorithm(24), and have also illustrated that ICD-based smoking definitions are highly specific for identifying smokers(25).

Unlike the epidemiologic collection (NHANES), the clinical collection (BioVU) is relatively free of data access restrictions. Therefore, the clinical collection workflow only utilizes the later stages of the workflow described in Figure 3. Output files from various statistical packages (such as PLINK) are parsed and Results Template files are generated for sharing among PAGE study sites and meta-analysis.

4. Conclusions

We describe here the epidemiologic (NHANES) and clinical (BioVU) collection workflows that enable high-throughput genotype-phenotype association studies and data sharing within EAGLE and the PAGE study. Both workflows were customized based on a variety of factors including data structure and data access. A major strength of this approach is that it provides the infrastructure to conduct systematic genetic analyses resulting in standardized files for data sharing and meta-analysis. A major weakness of this approach is that is requires substantial bioinformatics and computing resources and personnel to create, maintain, and implement the workflow. The preferential accessing of datasets with open access or fewer data use restrictions would assist in easing the effort required for the workflows. However, full access to local or collaborative datasets through dbGaP will still require substantial bioinformatics and computational support to fully mine the genotype-phenotype investments for high returns relevant to human disease and biology.

5. Acknowledgements

This work was supported by NIH U01 HG004798 and its ARRA supplements. The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center’s BioVU which is supported by institutional funding and by the Vanderbilt CTSA grant UL1 TR000445 from NCATS/NIH. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work. 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.

References

1. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S., Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, **106**, 9362-9367.

2. Lettre,G., Palmer,C.D., Young,T., Ejebe,K.G., Allayee,H., Benjamin,E.J., Bennett,F., Bowden,D.W., Chakravarti,A., Dreisbach,A., *et al.* (2011) Genome-Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project. *PLoS Genet*, **7**, e1001300.

3. Dumitrescu,L., Carty,C.L., Taylor,K., Schumacher,F.R., Hindorff,L.A., Ambite,J.-L., Anderson,G., Best,L.G., Brown-Gentry,K., Buzkova,P., *et al.* (2011) Genetic Determinants of Lipid Traits in Diverse Populations from the Population Architecture using Genomics and Epidemiology (PAGE) Study. *PLoS Genet*, **7**, e1002138.

4. Haiman,C.A., Fesinmeyer,M., Spencer,K.L., Buzkova,P., Voruganti,V.S., Wan,P., Haessler,J., Francheschini,N., Monroe,K., Howard,B.V., *et al.* (2012) Consistent direction of effect for established T2D risk variants across populations: The Population Architecture using Genomics and Epidemiology (PAGE) Consortium. *Diabetes*, **61**, 1642-1647.

5. Fesinmeyer,M.D., North,K.E., Ritchie,M.D., Lim,U., Franceschini,N., Wilkens,L.R., Gross,M.D., Buzkova,P., Glenn,K., Quibrera,M., *et al.* Genetic risk factors for body mass index and obesity in an ethnically diverse population: results from the Population Architecture using Genomics and Epidemiology (PAGE) Study *Obesity (Silver Spring)* (in press).

6. Zhang,L., Spencer,K.L., Voruganti,V.S., Jorgensen,N.W., Fornage,M., Best,L.G., Brown-Gentry,K.D., Cole,S.A., Crawford,D.C., Deelman,E., *et al.* Association of functional polymorphism rs2231142 (Q141K) in *ABCG2* gene with serum uric acid and gout in four US populations: the Population Architecture using Genomics and Epidemiology (PAGE) Study *Am J Epidemiol* (in press).

7. Carty,C.L., Buzkova,P., Fornage,M., Franceschini,N., Cole,S., Heiss,G., Hindorff,L.A., Howard,B.V., Mann,S., Martin,L.W., *et al.* (2012) Associations Between Incident Ischemic Stroke Events and Stroke and Cardiovascular Disease-Related Genome-Wide Association Studies Single Nucleotide Polymorphisms in the Population Architecture Using Genomics and Epidemiology Study. *Circulation: Cardiovascular Genetics*, **5**, 210-216.

8. N'Diaye,A., Chen,G.K., Palmer,C.D., Ge,B., Tayo,B., Mathias,R.A., Ding,J., Nalls,M.A., Adeyemo,A., Adoue,V.+., *et al.* (2011) Identification, Replication, and Fine-Mapping of Loci Associated with Adult Height in Individuals of African Ancestry. *PLoS Genet*, **7**, e1002298.

9. Chen,F., Chen,G.K., Millikan,R.C., John,E.M., Ambrosone,C.B., Bernstein,L., Zheng,W., Hu,J.J., Ziegler,R.G., Deming,S.L., *et al.* (2011) Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans. *Human Molecular Genetics*, **20**, 4491-4503.

10. Haiman,C.A., Chen,G.K., Blot,W.J., Strom,S.S., Berndt,S.I., Kittles,R.A., Rybicki,B.A., Isaacs,W.B., Ingles,S.A., Stanford,J.L., *et al.* (2011) Characterizing Genetic Risk at Known Prostate Cancer Susceptibility Loci in African Americans. *PLoS Genet*, **7**, e1001387.

11. Denny,J.C., Ritchie,M.D., Basford,M.A., Pulley,J.M., Bastarache,L., Brown-Gentry,K., Wang,D., Masys,D.R., Roden,D.M., Crawford,D.C. (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, **26**, 1205-1210.

12. Pendergrass,S.A., Brown-Gentry,K., Dudek,S.M., Torstenson,E.S., Ambite,J.L., Avery,C.L., Buyske,S., Cai,C., Fesinmeyer,M.D., Haiman,C., *et al.* (2011) The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. Epidemiol.*, **35**, 410-422.

13. Matise,T.C., Ambite,J.L., Buyske,S., Carlson,C.S., Cole,S.A., Crawford,D.C., Haiman,C.A., Heiss,G., Kooperberg,C., Marchand,L.L., *et al.* (2011) The Next PAGE in Understanding Complex

Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. *American Journal of Epidemiology*, **174**, 849-859.

14. Roden,D.M., Pulley,J.M., Basford,M.A., Bernard,G.R., Clayton,E.W., Balser,J.R., Masys,D.R. (2008) Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin Pharmacol Ther*, **84**, 362-369.

15. Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS) (2012).

16. Pulley,J., Clayton,E., Bernard,G.R., Roden,D.M., Masys,D.R. (2010) Principles of Human Subjects Protections Applied in an Opt-Out, De-identified Biobank. *Clinical and Translational Science*, **3**, 42-48.

17. Mailman,M.D., 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.

18. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J., Sham,P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analysis. *Am J Hum Genet*, **81**, 559-575.

19. R Development Core Team (2008) R Foundation for Statistical Computing, Vienna, Austria.

20. Homer,N., Szelinger,S., Redman,M., Duggan,D., Tembe,W., Muehling,J., Pearson,J.V., Stephan,D.A., Nelson,S.F., Craig,D.W. (2008) Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genet*, **4**, e1000167.

21. Pendergrass,S., Dudek,S., Crawford,D., Ritchie,M. (2010) Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis. *BioData Mining*, **3**, 10.

22. Pendergrass,S., Dudek,S., Crawford,D., Ritchie,M. (2012) Visually integrating and exploring high throughput Phenome-Wide Association Study (PheWAS) results using PheWAS-View. *BioData Mining*, **5**, 5.

23. Dumitrescu,L., Ritchie,M.D., Brown-Gentry,K., Pulley,J.M., Basford,M., Denny,J.C., Oksenberg,J.R., Roden,D.M., Haines,J.L., Crawford,D.C. (2010) Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genet Med*, **12**, 648-650.

24. Liu,M., Shah,A., Min,J., Peterson,N.B., Dia,Q., Aldrich,M.C., Chen,Q., Bowton,E.A., Liu,H., Denny,J.C., Xu,H. A study of transportability of an existing smoking status detection module across institutions *AMIA Annu Symp Proc* (in press).

25. Wiley,L.K., Shah,A., Xu,H., and Bush,W.S. (2012) ICD-9 tobacco use codes are effective identifiers of smoking status. Translational Bioinformatics Conference, October 14-17, Jeju Island, Korea.