Published in final edited form as:
Genet Epidemiol. 2011 April ; 35(3): 159–173. doi:10.1002/gepi.20564.

Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience

Siiri N. Bennett[1,*], **Neil Caporaso**[2], **Annette L. Fitzpatrick**[1,3], **Arpana Agrawal**[4], **Kathleen Barnes**[5], **Heather A. Boyd**[6], **Marilyn C. Cornelis**[7], **Nadia N. Hansel**[5], **Gerardo Heiss**[8], **John A. Heit**[9], **Jae Hee Kang**[10], **Steven J. Kittner**[11], **Peter Kraft**[12], **William Lowe**[13], **Mary L. Marazita**[14], **Kristine R. Monroe**[15], **Louis R. Pasquale**[10], **Erin M. Ramos**[16], **Rob M. van Dam**[17], **Jenna Udren**[1], and **Kayleen Williams**[1] **for the GENEVA Consortium**

[1]Collaborative Health Studies Coordinating Center, Department of Biostatistics, University of Washington, Seattle, WA

[2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD

[3]Department of Epidemiology, University of Washington, Seattle, WA

[4]Department of Psychiatry, Washington University School of Medicine, Saint Louis, MO

[5]Johns Hopkins University School of Medicine, Baltimore, MD

[6]Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark

[7]Harvard School of Public Health, Boston, MA

[8]Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, NC

[9]Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN

[10]Harvard Medical School, Boston, MA

[11]Department of Neurology, University of Maryland School of Medicine and Baltimore Veterans Affairs Medical Center, Baltimore, MD

[12]Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, MA

[13]Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL

[14]Center for Craniofacial and Dental Genetics, Department of Oral Biology, School of Dental Medicine, University of Pittsburgh; Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh; Clinical and Translational Science Institute and Department of Psychiatry, School of Medicine, University of Pittsburgh, Pittsburgh, PA

[15]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA

[16]Office of Population Genomics, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD

[17]Department of Epidemiology and Public Health and Medicine, Faculty of Medicine, National University of Singapore, Singapore; Department of Nutrition, Harvard School of Public Health, Boston, MA

[*]Correspondence to: Siiri N. Bennett, MD, Collaborative Health Studies Coordinating Center, Building 29, Suite 310, 6200 NE 74th Street, Seattle, WA 98115. siirib@u.washington.edu.

Abstract

Genome-wide association study (GWAS) consortia and collaborations formed to detect genetic loci for common phenotypes or investigate gene-environment (G*E) interactions are increasingly common. While these consortia effectively increase sample size, phenotype heterogeneity across studies represents a major obstacle that limits successful identification of these associations. Investigators are faced with the challenge of how to harmonize previously collected phenotype data obtained using different data collection instruments which cover topics in varying degrees of detail and over diverse time frames. This process has not been described in detail. We describe here some of the strategies and pitfalls associated with combining phenotype data from varying studies. Using the Gene Environment Association Studies (GENEVA) multi-site GWAS consortium as an example, this paper provides an illustration to guide GWAS consortia through the process of phenotype harmonization and describes key issues that arise when sharing data across disparate studies. GENEVA is unusual in the diversity of disease endpoints and so the issues it faces as its participating studies share data will be informative for many collaborations. Phenotype harmonization requires identifying common phenotypes, determining the feasibility of cross-study analysis for each, preparing common definitions, and applying appropriate algorithms. Other issues to be considered include genotyping timeframes, coordination of parallel efforts by other collaborative groups, analytic approaches, and imputation of genotype data. GENEVA's harmonization efforts and policy of promoting data sharing and collaboration, not only within GENEVA but also with outside collaborations, can provide important guidance to ongoing and new consortia.

Keywords

phenotype; harmonization; genome-wide association studies; GENEVA; consortia

Introduction and overview

The vast majority of genome-wide association studies (GWAS) have focused on the main effects of gene variants at specific loci on disease outcomes or traits, although most associations identified so far account for only a small portion of the phenotype variations seen [McCarthy et al., 2008; Hindorff et al., 2009]. To extend these findings GWAS consortia and collaborations have formed in which investigators share data to achieve adequate power to identify genetic loci associated with secondary phenotypes and increase the power to study less common outcomes [Psaty et al., 2009; Manolio et al., 2007]. There have been several single center studies of gene-environment (G*E) interactions in complex disease [Kang et al., 2010; Cornelis et al., 2009] but the development of consortia also offers the opportunity to enhance power to discover and verify G*E interactions. Investigators planning new studies and looking to establish consortia or collaborations can share information on how phenotypes are being measured, what questions are being asked, and how responses will be coded so that similar data can be combined with relative ease. Tools such as the PhenX Toolkit of standardized, high priority measures (https://www.phenxtoolkit.org) are increasingly available to investigators planning new studies, though most current collaborations involve existing studies whose phenotypes and data collection instruments are already defined.

While considerable effort has gone into reducing genotype measurement error and ensuring genotype accuracy and consistency of results [de Bakker et al., 2008], phenotype heterogeneity, in both outcomes and covariates across studies, represents a major challenge to successful GWAS analysis of common traits [Zeggini and Ioannidis, 2009; Seminara et al., 2007] or genome-wide assessment of G*E interactions in complex diseases. The recent discovery of markers that are specifically associated with estrogen-receptor negative breast cancer highlights the potential importance of specific, harmonized phenotypes: earlier GWAS of a general breast cancer phenotype did not identify these markers, due to lack of power [Kraft and Haiman, 2010]. In contrast, cross-study analysis groups, such as those established by the Psychiatric Genetics Consortium [Psychiatric GWAS Consortium Coordinating Committee et al., 2009], have begun to analyze GWAS results related to differing psychiatric diagnoses that are known to have shared genetic underpinnings (e.g. Schizophrenia and Bipolar Disorder, Bipolar and Major Depressive Disorder) [International Schizophrenia Consortium et al., 2009; Liu et al., 2011]. The existing literature suggests that phenotype harmonization may reveal novel loci for disease subtypes as well as shared variants among broad categories of disease. These examples illustrate how phenotype harmonization is a crucial first step and a key component of successful multi-study collaboration, where investigators are faced with the challenge of harmonizing previously collected phenotype data derived from diverse data collection instruments over different timeframes and in varying degrees of detail.

Phenotype harmonization in GWAS consortia and collaborations involves a multi-stage process of 1) identifying commonalities and differences in phenotype data to assess the feasibility and potential benefit of combining data; 2) developing common data definitions for phenotypes of interest; and 3) creating and applying study-specific algorithms to convert data into a common format. The goal is to maximize the comparability and compatibility of data across datasets and to minimize inconsistencies and misclassification [Seminara et al., 2007]. The process must balance the need to augment the sample size which increases power for gene discovery with the likelihood of increased data heterogeneity which will decrease power to detect real effects.

GENEVA (Gene Environment Association Studies) is a multi-site collaborative program initiated in 2006 as part of the National Institutes of Health (NIH) Genes, Environment and Health Initiative (GEI) that aims to accelerate the understanding of genetic and environmental contributions to health and disease [Cornelis et al., 2010]. The GENEVA consortium, led by the National Human Genome Research Institute (NHGRI) and working closely with representatives from several institutes at the NIH, includes a Coordinating Center (CC), two genotyping centers, and fourteen independently-designed GWAS whose primary outcomes of interest include addiction, blood pressure, cardiovascular disease, chronic obstructive pulmonary disease, dental caries, type 2 diabetes, lung cancer, maternal metabolism and birth weight interactions, oral clefts, premature birth, primary open-angle glaucoma, prostate cancer, stroke, and venous thrombosis. Two additional GWAS examining coagulation and melanoma joined the consortium in 2009. Each study has collected relevant environmental exposure data. The participating studies vary widely in design: some of the longitudinal studies have been ongoing for years or even decades, others have international data collection sites, and some are family-based studies with phenotype and genotype data available on multiple family members. Using GENEVA as a practical example, this paper provides an illustration to guide GWAS consortia through the process of phenotype harmonization and describes key phenotype harmonization issues that arise when sharing data across disparate studies. The principles outlined here can also be applied to other data sharing contexts.

Roles and responsibilities

Phenotype harmonization must be recognized early on as a key requirement for cross-study collaboration. Key personnel should be identified and, depending on the complexity, size, number, and primary endpoints of the participating studies, a group or committee should be established or identified that takes responsibility for directing and overseeing phenotype harmonization activities. In GENEVA, three groups are involved in phenotype harmonization: a Phenotype Harmonization Subcommittee (PHS); the phenotype-specific Working Groups (WGs); and the Coordinating Center (CC) (Figure 1).

Given the diverse nature of studies supported by GENEVA and their lack of familiarity with each others' phenotypes, the GENEVA Steering Committing agreed that phenotype harmonization was a core task warranting investigator participation and leadership, so a PHS was created. The PHS includes individual study, NIH and CC representatives who are interested in harmonization of specific phenotypes and exposures or in increasing sample size for improving power of analyses in their areas of investigation. The Subcommittee provides a forum for discussion of common problems, sets policies, identifies phenotypes common across studies, encourages data sharing, establishes and oversees phenotype-specific WGs, and provides advice, direction and feedback to the CC regarding phenotype harmonization and related issues.

Phenotype WGs consist of representatives from each study contributing data to that phenotype's cross-study analyses. These representatives generally have expertise in the subject area and are aware of the intricacies involved in categorizing or characterizing the phenotype. The WGs also have representatives from the NIH and the CC. Each WG is led by an investigator with a specific interest in the phenotype who manages and coordinates the group's activities. The WGs evaluate the feasibility and logistics for cross-study analyses related to specific phenotypes of interest. They identify and define variables to be shared, identify covariates, recommend the most appropriate analytic methods, draft analysis plans using a template developed by the PHS, and identify the investigators performing the analyses, usually members of the WG.

The CC is responsible for phenotype data organization and coordinating activities related to data collection and management and facilitating cross-study analyses. The CC assists the PHS with identifying potential areas of common interest, harmonizes covariates, and establishes and manages a centralized relational database which serves as a common phenotype/genotype repository providing working groups with cross-study data upon request. The CC assists working groups with phenotype harmonization and provides them with statistical summaries of phenotype data and a central file-sharing site. The CC may also assist working group investigators with data analysis, if requested.

Process of Phenotype Harmonization

Phenotype harmonization involves a number of integrated activities (Table 1). Note that the elements below may be addressed concurrently and not necessarily in the order given.

Identifying common phenotypes

Phenotype harmonization starts with an inventory of phenotypes that investigators are interested in pursuing during cross-study analyses. For consortia involving only a few studies, or for consortia in which all studies have a common interest (e.g. stroke), this may only require a review of data collected by different studies and discussion among investigators. For collaborations such as GENEVA that incorporate studies of various designs and diverse outcomes of interest, a more extensive (formal survey) process is required.

In GENEVA, the CC reviewed the initial phenotype submission plans and data collection forms for each study. The CC identified overlapping phenotype areas for which data had been collected, and created a web-based survey (see Supplementary Material, Appendix A) in which study investigators were asked to indicate, for 13 broad phenotype categories, if (a) their study had collected specific data, and if so, (b) the level of sharing, i.e. would they share it solely with other GENEVA investigators or would they share their data with authorized researchers through the controlled-access Database of Genotypes and Phenotypes (dbGaP) [Mailman et al., 2007]. Investigators were invited to list other phenotypes for which data were available and which could be shared in cross-study analyses. The CC tabulated responses and provided the information back to the PHS for review and discussion. If there was interest in a phenotype across three or more studies, the Subcommittee solicited a volunteer to lead a WG for that phenotype and solicited nominations for WG members from those studies interested in participating in cross-study analyses of those phenotypes. If only two studies were able to share data on a specific phenotype, they could still collaborate and perform cross-study analyses, but a formal WG might not be required.

Determining feasibility of cross-study analyses

The most important task of the phenotype-specific WGs is to determine the feasibility and logistics for cross-study analyses related to a given phenotype. The WGs' reviews need to take into account each study's data scope, consent limitations, and study design, including the actual questions asked, data collection protocols, phenotype definitions, possible values or responses, estimated number of individuals for whom there would be phenotype data, and any other factors that might influence analyses. The primary considerations fall into seven main categories:

1. Identifying participating studies. An individual study's participation may depend on how many subjects can participate in the analysis and whether these subjects are representative of the cohort. This is a complex judgment dependent upon study size, design, depth and quality of data supporting the phenotype and whether the primary study outcome is related to the phenotype of interest.

2. Identifying numbers of subjects. It is important to determine whether the number of subjects resulting from combining studies provides sufficient statistical power to detect associations between genetic variants and the phenotype of interest. The estimate of the number of subjects that each study can contribute will help determine whether the increase in power will be sufficient to justify efforts associated with harmonization. The combined total number of subjects to be included in the cross-study analysis also needs to be reassessed and reevaluated after the following considerations are reviewed.

3. Reviewing consent status. Studies' consent forms define and limit how an individual's data can be used. The consent limitations for each study contributing data need to be reviewed by the WG to ensure that the planned analyses are compatible with the studies' consent limitations. In all instances, the interpretation of the consent and decision as to which analyses the data can contribute should be determined by each study's primary investigator in conjunction with the appropriate Institutional Review Board (IRB), if relevant.

4. Reviewing each study's inclusion/exclusion criteria. Each study's inclusion and exclusion criteria need to be considered and their impact on any cross-study analysis evaluated. In some cases appropriate analytic strategies such as sampling fractions or adjustments may allow participation.

5. Reviewing data definitions. Phenotype and covariate data from different studies may differ in units of measurement (e.g. kilograms versus pounds), manner of collection, mode of administration, how questions are worded, and how quantities or qualities are measured. Each WG needs to decide if the different data sources and values are compatible and comparable. If they are compatible and comparable,

or can be converted to a common definition (e.g. number of alcoholic drinks per day to ethanol intake in grams per week), then the WG defines the phenotype variable for analysis and the possible values.

- **a.** *Measured versus self-reported*. Data on measured phenotypes (e.g. weight, blood pressure, serum glucose) will reflect each study's measurement protocols, laboratory standards, specimen processing, and population exposures. For example, differences in instrument calibration can introduce individual-level and study-specific variance. Self-reported measures such as tobacco use or dietary intake may be more prone to error, misclassification or misspecification in the questions asked and possible responses, as well as being more prone to recall bias and varying degrees of accuracy that reflect subjects' motivation and education, and questionnaire design, quality and completeness. Such measures may derive the most benefit from standardization, though hard measures such as height or glucose levels may also benefit. However, validation of some self-reported phenotypes such as weight may exist that allow these phenotypes to be harmonized with directly measured quantitative traits such as body mass index (BMI) [Rimm et al., 1990].

- **b.** *Variations in data definitions*. Minor variations in the actual questions asked by different studies can have significant impact. For instance, in GENEVA, while some studies asked respondents to report the number of alcoholic beverages they consumed in a single typical week, others asked how many drinks they drank in a day, but not how many days they drank each week.

- **c.** *Differences in source populations*. The nature of some cohorts can influence responses such that participants may respond in a manner not representative of other populations. For example, alcohol consumption among drug users may be sufficiently different from that of the general population that the data cannot be compared.

6. Comparing data distributions. Preliminary statistical summaries of the data are useful for identifying ranges of responses to determine if different studies' data are comparable and for identifying possible outliers or extreme values. In some instances, investigators may decide that the data distribution across studies differs to such an extent that any predictors of the distribution would be unlikely to be the same and cross-study analyses would be inappropriate. These distributions can also be used to help investigators define cases and controls for cross-study analyses.

7. Determining if data are being used in similar analyses by other consortia. Large, well-known longitudinal studies often contribute genotype and phenotype data to and participate in multiple collaborations and consortia, and this needs to be considered when the WG plans its analyses. While analyses done within different collaborations and consortia may share certain characteristics, they may differ in when data were collected, number of participants and outcomes, technology used to assess genetic data, number and quality of endpoints, covariate mix, and so on. Thus, anticipated analyses in one consortium will likely differ from those of another, and explicit duplication of phenotype and genotype data in analyses should not be an issue. However, where the planned investigations of the WG and those of the other consortium overlap, a decision needs to be made regarding the appropriateness of using the same data in separate but similar analyses. In GENEVA, one large longitudinal study is part of both GENEVA and another consortium, and approximately two-thirds of the individuals in GENEVA are included in the other consortium's analyses. In this case, the WG and study representatives decided to exclude from cross-study analyses individuals included in the other consortium's analyses so that duplication of populations would be avoided in the analyses performed by the different consortia.

Preparing common definitions

Once the review of the data definitions and values described above indicates that data from different studies are comparable, common definitions and values need to be agreed upon. WGs may combine categories into larger units (e.g. drinks per day combined with drinking days per week to get drinks per week), stipulate inclusion and exclusion criteria (e.g. excluding those who have never had a drink), create a dichotomous variable (e.g. ever or never drinker, or use longitudinal and cross-sectional data to define whether the respondent is a current smoker or has quit smoking), or assign a standard measure to be used (e.g. BMI calculated from a variety of self-reported or laboratory assessments of height and weight). In general, the more tightly defined the phenotype, the greater the likelihood that one or more studies may be unable to contribute to the analyses; the looser the definition, the greater the likelihood that more subjects and more studies can be included. An example of how data from various studies might be combined is shown in Tables 2a and 2b.

For continuously distributed outcomes, GENEVA WGs, like other large-scale collaborative meta-analyses [Lindgren et al., 2009; Thorgeirsson et al., 2010], have applied the same transformation to all datasets for a single outcome. These transformations are discussed in a phenotype-specific manner for each meta-analysis and vary across phenotypes. Logical and extreme outliers are deleted or recoded before transformations are applied to avoid non-normal error distributions. The removal or recoding of outliers, application of a common transformation and a strategy for covariate selection (see below) ensures comparability and interpretation of the estimates from the contributing studies.

Analytic methods

Phenotype WGs determine phenotype definitions, identify covariates, and set inclusion/exclusion criteria. But they must also consider the advantages and disadvantages of different analytic approaches, i.e. whether to perform meta-analyses of summary data provided by each study, analyses of pooled individual-level data, or both. If investigators choose to perform meta-analyses of summary data, the WG will need to agree on the subgroup analyses, data format, and analysis method used by each study to produce its summary statistics. Investigators need to address the following points:

1. Analytic approaches. In consortia such as GENEVA where participating studies have different designs and participant populations as well as different ethical and administrative constraints, many investigators prefer to conduct their own analyses of their study data, generate summary data, and then perform meta-analyses of the summary statistics. Many study investigators feel that they are best positioned to control for or otherwise anticipate irregularities in their data that would potentially be missed in a pooled analysis on one large, combined dataset of individual-level data conducted centrally. In addition, meta-analyses of summary statistics are much easier to do than creating and applying study-specific algorithms to convert data from each study into a harmonized dataset that includes data from all participating studies. Logistic requirements and ethical concerns (such as sharing individual-level data) are fewer, and guidelines for meta-analysis of GWAS are available [de Bakker et al., 2008; Zeggini and Ioannidis, 2009]. However, statistical methods for G*E interactions are still being developed and non-standard methods requiring specific statistical programs may be more difficult to conduct across studies using meta-analysis if expertise is not available for all individual studies. Also, use of

summary statistics reduces the opportunities for stratification and in-depth analyses that might be possible by pooling individual level data.

2. Coding disparate variables. If meta-analyses are planned, then each contributing study's investigators must convert their raw data to conform to the agreed upon definitions and values before calculating summary statistics. If, instead, an analysis of pooled, individual-level data is planned, the application of algorithms to convert individual study data to the common definitions and values can be done by individual study investigators, and the converted data pooled at a central location, such as a coordinating center, or the coordinating center may apply the appropriate algorithm to each study's data.

3. Controlling for admixture. Individual studies should be encouraged to consult with their analysts for control of admixture during cross-study analyses. In GENEVA it is now standard practice during the genotype cleaning process to identify genetic outliers and exclude them from analysis, and include an analysis of population effects of principal components to further control for admixture.

Selecting and harmonizing covariates

The issues surrounding phenotype harmonization also apply to the selection and harmonization of covariates. For GENEVA, WGs have identified a standard set of covariates (e.g. sex and age) applicable to all studies. Other covariates relevant to the phenotype and that can be easily harmonized across contributing studies (e.g. smoking as a covariate for caffeine consumption) are also included. Where covariates are relevant to the characteristics of one study but not to others (e.g. menopause in a sample of older women) adjustment may be inappropriate and stratified analysis or exclusion may be a more appropriate strategy. Studies that have divergent assessment protocols for cases and control subjects have elected, in some instances, to analyze cases and controls separately. This strategy of uniform covariate selection is similar to that used by several other consortia in their large-scale meta-analyses [Lindgren et al., 2009; Thorgeirsson et al., 2010] and has led to comparability of the resulting parameter estimates from individual analyses. However, it can also result in some study-specific confounds. In GENEVA, it is the responsibility of the individual study investigators to assess the feasibility and acceptability of covariate adjustments.

Creating a centralized relational database

Creation of a centralized database and repository for all phenotype and genotype data provides investigators with the opportunity to easily access the data as they identify new areas of interest for cross-study analyses. In GENEVA, once each study's genotype and phenotype data have gone through a standardized quality control cleaning process, the CC adds each study's phenotype data, data dictionary, individual-level consent and public use or GENEVA-only use status to a centralized relational database. As the WGs define common phenotypes and covariates for cross-study analyses, the CC applies study-specific algorithms to each study's phenotype data to create a dataset and data dictionary for the harmonized variables and covariates, and these are added to the centralized relational database.

Drafting an analysis plan and performing cross-study analyses

While investigators work through the process of phenotype harmonization, they should simultaneously develop a preliminary analysis plan as issues will likely arise that affect harmonization decisions. In GENEVA, each WG drafts a plan for analysis using guidelines created by the PHS (Table 1), defining the variables or outcomes of interest and their type (e.g. whether they are discrete, ordinal or continuous variables), identifying covariates needed for the analyses, stating inclusion and exclusion criteria (e.g. race or ethnicity, value limits), describing the planned subgroup analyses and proposed analytic approach, and specifying individuals' roles in analysis. These plans are refined as preliminary analyses of the selected phenotypes are conducted by study investigators, and the current plans are posted on the consortium's website.

For WGs that have decided to conduct meta-analyses of summary data provided by each study, phenotypes are defined and variables recoded (if necessary) by all studies so that genome-wide analyses can be done on comparable measures. Analyses of individual study data are performed by each study's investigators as soon as that study's phenotype and genotype data have completed the cleaning process. The summary statistics are then provided to the WG member leading the meta-analysis. Where the WG has decided on an analysis of pooled individual-level data, the CC applies the appropriate algorithms to each dataset to harmonize each study's data according to the WG's definition, pools the data into one combined dataset, and provides this to the investigators doing the analysis.

In general, a majority of meta-analyses of gene effects conducted by GENEVA groups have adopted a fixed-effects model, with selective testing for random-effects in subsequent follow-up analyses. Kraft, Zeggini & Ioannidis [Kraft et al., 2009] provide evidence for reduced power and deflation of the heterogeneity parameter in standard meta-analyses – hence, heterogeneity testing is restricted to the most promising signals.

Issues related to phenotype harmonization

Genotyping timeframes

Genotyping of the magnitude required for a consortium such as GENEVA, which includes data on over 80,000 individuals from 16 disparate studies, does not take place on all studies simultaneously. One study's genotyping results arrive for cleaning and quality control checks while other studies' samples are being prepared for or are undergoing genotyping. Therefore genotyping timelines become a critical consideration when planning cross-study analyses. WGs must consider that GWAS results are emerging in the context of sharp competition among various research groups for publication, and there should be an ongoing process to decide when sufficient numbers are available to conduct a study that represents a meaningful advance. WGs may decide to proceed with data analyses when only a few studies are complete and if statistical thresholds are achieved because discoveries relating to the phenotype of interest are emerging so rapidly. The data from studies whose genotype results will be completed later can be used for replication analyses. Alternate strategies might be to wait until more studies' genotype results are available, refine the phenotype definitions so the group can conduct more specific, stratified analyses of key phenotypes, or collaborate with outside consortia doing similar investigations.

Working with external collaborative groups

There are a large number of consortia undertaking GWAS analyses (Table 3) and the phenotypes under investigation by various consortia often overlap. Negotiations with other consortia WGs can lead to collaborations where GWAS investigators can contribute to data analyses being led by other consortia. This benefits both groups—investigators are able to contribute to primary analyses of an important phenotype and the collaboration's primary analyses now include data on additional individuals. Studies that are still being genotyped can contribute to later analyses. Alternatively, investigators may decide to serve as a replication study or look at other related or intermediate phenotypes or associations and G*E analyses beyond the initial ones being conducted. There are, of course, implications of overlapping data across independent meta-analyses and in some instances, such overlap is unavoidable. In those instances, appropriate corrections, such as those described by Lin & Sullivan [Lin DY and Sullivan PF, 2009], might be considered.

The sample size-phenotype heterogeneity paradox

Inevitably an increase in sample size achieved by pooling samples across studies introduces phenotype heterogeneity that may alter the power to discover genes or G*E interactions for a complex trait of interest. During gene discovery, the balance between power achieved by larger sample size versus loss of power produced by introduced phenotype heterogeneity during phenotype harmonization may favor the former. For example, little is known about the genetic architecture of primary open-angle glaucoma (POAG) although the sib relative risk is 10 [Wolfs et al., 1998]. Icelandic investigators reported two SNPs between CAV1 and CAV2 associated with POAG with an OR of ~1.3 and with a p-value of **5E-10** [Thorleifsson et al., 2010]. Several replication datasets were provided and phenotype definitions varied widely. Not all replication sets achieved statistical significance and several had wide confidence intervals. When the replication sets were combined they did achieve significance with an effect size and confidence comparable to that reported in the discovery set. This suggests that the association between CAV1 and CAV2 gene variants and POAG represents a true positive and that pooling samples during the phenotype harmonization process can provide power even in the face of phenotype heterogeneity during gene discovery, although this might not be generalizeable to all complex traits.

On the other hand, when trying to identify G*E interactions for a phenotype with known genetic architecture, the balance between power gained by adding more samples versus power degradation produced by phenotype heterogeneity may favor the latter. The tradeoff between sample size and phenotypic heterogeneity of exposure data is modeled in Figure 2 [Lindstrom et al., 2009]. This hypothetical example considers a rare disease (prevalence 1 in 1,000), no main effect for the binary genetic factor (with 20% prevalence), an odds ratio of 1.5 for the exposure, an interaction odds ratio of 1.35, and a Type I error rate of 5E-8. This illustrates the power of a case-control study to detect a G*E interaction (departure from a multiplicative odds model) when the binary exposure is measured perfectly or via a good proxy with 77% specificity and 99% sensitivity (roughly analogous to self-reported versus measured overweight status). Figure 2 illustrates that even modest misclassification can greatly decrease the power of tests for G*E interaction (and the relative decrease is greater for rare exposures). On the other hand Figure 2 also illustrates that a large study using the proxy can have greater power than a smaller study using the perfect measure, although the power gain is modest at best. Nevertheless, the modest power gain may be important when the perfect measure is prohibitively expensive or only available on a small fraction of samples, while the good measure is relatively inexpensive or already available on many samples.

Genotype imputation

In many consortia or collaborations, different studies' samples may be genotyped on different platforms or may be genotyped using different versions of the same technology. This is particularly true when large studies are conducted over time because the technology is continuously evolving. Imputation tries to address these differences [Li et al., 2009], but this, too, can create more uncertainties in data comparability when investigators use different software packages to impute data. In GENEVA, genotyping is being performed on both Affymetrix and Illumina platforms with varying degrees of single nucleotide polymorphism (SNP) coverage. In addition, the two platforms detect different sets of SNPs. Plans to conduct imputation vary widely across studies, and there has been considerable discussion regarding the comparability of imputation results performed at different sites using different HapMap reference panels [http://hapmap.ncbi.nlm.nih.gov] and software packages (MACH [http://www.sph.umich.edu/csg/abecasis/MACH/index.html; Li et al., 2009; Li et al., 2006], Impute [https://mathgen.stats.ox.ac.uk/impute/impute.html; Marchini et al., 2007; Howie et al., 2009] and BEAGLE [https:faculty.washington.edu/browning/beagle/beagle.html; Browning and Browning, 2009] being the most common). Most WGs are utilizing individual investigators' imputation results and reviewing preliminary analyses of individual study data to ensure that analyses yield similar findings. Another solution for consortia, and one adopted by GENEVA, is to have all genotype data imputed centrally using a standard methodology.

Converging on an analytic framework

As previously mentioned, pivotal issues WGs face when combining or harmonizing data from different studies are determining the analytic plan and the precise modeling strategy, adjusting for confounders, and performing stratified analysis. WGs continually consider, for instance: (a) whether waves of longitudinal data should be combined for comparability to cross-sectional studies; (b) whether analyses should be adjusted for covariates, such as ethnicity, or conducted separately in each group; (c) whether analyses should remove, truncate, or adjust for outliers (e.g. by normalizing distributions) and how these outliers are defined; (d) which statistical models accommodate the nuances of each dataset; and (e) whether studies should focus on main effects of individual SNPs, candidate genes, on gene systems or on gene-by-gene (G*G) and G*E analyses. Overarching these phenotype issues are comparable concerns regarding uniform quality control metrics for genotype data and comparable imputation statistics which add inherent complexity to the process. A pre-determined analytic strategy for the main gene discovery stage as well as approaches to G*E and G*G associations requires special planning as there is no standard approach though all agree that very large numbers are required to support the enhanced power requirements of G*E investigations [Garcia-Closas and Lubin, 1999; Kraft and Hunter, 2005].

Phenotype harmonization when using controls for cases

Many investigators are now looking at using controls from one study as cases in another study as a way of expanding the size of an investigation without the cost of genotyping additional individuals. However, it is only possible when the subjects come from studies whose primary interests are not associated with each other. This becomes complex because it requires well-characterized phenotypes for selection of cases and controls, yet many of the phenotypes are secondary measures collected by the contributing studies and thus may not be well-characterized. Such phenotypes also may represent only a snapshot of the individual's phenotype at one point in time and may not be correlated with the development of future disease or conditions. GENEVA currently has a WG looking specifically at this issue, and its investigation is still in progress.

Progress

The infrastructure and processes described here have enabled GENEVA investigators to address common problems, facilitated the generation of new ideas for research and cross-study analyses, and provided GENEVA investigators with a means of turning these questions into active areas of study. GENEVA initially convened seven phenotype WGs. These included anthropometry, alcohol use, smoking, caffeine use, female reproductive history, psychiatric history, and oral health. As groups met and discussed the details of how and what data were collected, a few collaborations were determined to be not feasible (e.g. psychiatric history was not adequately addressed across most of the studies) and one group found that the distribution of data between two studies differed to such an extent that cross-study analyses were inappropriate. However, as investigators have become familiar with the various studies and with each other, and as new studies have joined the consortium, several new areas of mutual interest (e.g. sleep, protective effects, and physical activity) were identified and prompted the formation of new WGs. Each WG has identified one or more key phenotypes for which three or more studies are contributing data (Table 4). Thus far, eleven studies have genotype and phenotype data that have undergone consortium-level quality control and assurance. Six studies have so far contributed data to cross-study analyses, and an additional six have plans to contribute data to cross-study analyses.

Summary

Large consortia have emerged to correlate phenotypes with specific variants at certain genetic loci but the strategies and pitfalls associated with combining phenotype data from varying studies have not previously been described in detail. A key recent critique of GWAS reveals a principal concern with effect sizes [Goldstein, 2009] suggesting that meta-analyses of complex traits may be the avenue to successful gene discovery. Consortia-based studies that incorporate the approaches described here will be essential to achieving the power to support investigations of effect sizes demonstrated to be typical [Garcia-Closas and Lubin, 1999].

GENEVA is unusual in the diversity of its studies, yet the issues it has had to face in sharing data are common to all collaborations. GENEVA's harmonization efforts have important applications to current consortia that struggle with study heterogeneity but that seek to maximize the use and value of their data or extend their current data to explore the genetic architecture of novel but relevant traits.

Furthermore, GENEVA's policy to promote data sharing and collaboration, not only within GENEVA but also with other consortia and collaborations outside of GENEVA, has presented unique challenges in harmonization. Our experience demonstrates that systematic planning by a knowledgeable Coordinating Center, a team of collaborative investigators, and engaged program staff from the funding agency, are critical for maximizing the scientific return from large-scale GWAS. Collaborations such as GENEVA can stimulate development of novel methodologies for phenotype harmonization which, hopefully, will actively translate into steps towards identification of genetic loci important for traits related to health and disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding support for the GENEVA genome-wide association studies was provided through the NIH Genes, Environment and Health Initiative (GEI). Some studies also received support from individual NIH Institutes. Barnes (U01HG004738); Beaty (NIDCR: U01DE018993, NIH contract: HHSN268200782096C); Bierut (U01HG004422, NIAAA: U10AA008401, NIDA: P01CA089392, R01DA013423, NIH contract: HHSN268200782096C); Boerwinkle (U01HG004402, NHLBI: N01HC55015, N01HC55016, N01HC55018, N01HC55019, N01HC-55020, N01HC-55021, N01HC55022, R01HL087641, R01HL59367, R01HL086694, NIH contract: HHSN268200625226C, NIH Roadmap for Medical Research: UL1RR025005); Caporaso (Z01CP010200); Fornage (U01HG004729); Hu (U01HG004399); Haiman (U01HG004726, NCI: CA63464, CA54281, CA136792); Heit (U01HG004735); Lowe (U01HG004415); Marazita (NIDCR:U01DE018903, NIH contract: HHSN268200782096C); Mitchell (U01HG004436, NINDS and Office for Research in Women's Health: R01NS45012); Murray (U01HG004423); Pasquale [U01HG004728, NEI: R01EY015473, NEI: R01EY015872 (J.L.Wiggs)]. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Genotyping was performed at the Broad Institute of MIT and Harvard, with funding support from the NIH GEI (U01HG04424), and Johns Hopkins University Center for Inherited Disease Research, with support from the NIH GEI (U01HG004438) and the NIH contract “High throughput genotyping for studying the genetic contributions to human disease”

(HHSN268200782096C). Funding support was also provided, in part, by the Intramural Research Program of the NIH, National Library of Medicine. M.C. Cornelis is a recipient of a Canadian Institutes of Health Research Fellowship. A. Agrawal is supported by grants from the NIDA. K. C. Barnes is supported in part by the Mary Beryl Patch Turnbull Scholar Program. L. R. Pasquale is supported by a Research to Prevent Blindness Physician Scientist Award.

The GENEVA consortium thanks the staff and participants of all GENEVA studies for their important contributions.

References

Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 2009;84:210–223. [PubMed: 19200528]

Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, Beaty TH, Bennett SN, Bierut LJ, Boerwinkle E, Doheny KF, Feenstra B, Feingold E, Fornage M, Haiman CA, Harris EL, Hayes MG, Heit JA, Hu FB, Kang JH, Laurie CC, Ling H, Manolio TA, Marazita ML, Mathias RA, Mirel DB, Paschall J, Pasquale LR, Pugh EW, Rice JP, Udren J, van Dam RM, Wang X, Wiggs JL, Williams K, Yu K, for the GENEVA Consortium. The Gene, Environment Association Studies Consortium (GENEVA): Maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genet Epidemiol 2010;34:364–372. [PubMed: 20091798]

Cornelis MC, Qi L, Kraft P, Hu FB. TCF7L2, dietary carbohydrate, and risk of type 2 diabetes in US women. Am J Clin Nutr 2009;89:1256–1262. [PubMed: 19211816]

de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 2008;17(R2):R122–128. [PubMed: 18852200]

Garcia-Closas M, Lubin JH. Power and sample size calculations in case-control studies of gene-environment interactions: Comments on different approaches. Am J Epidemiol 1999;149:689–692. [PubMed: 10206617]

Goldstein DB. Common genetic variation and human traits. N Engl J Med 2009;360(17):1696–1698. [PubMed: 19369660]

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. PNAS 2009;106(23):9362–9367. [PubMed: 19474294]

Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 2009;5(6):e1000529. [PubMed: 19543373]

International Schizophrenia Consortium. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009;460:748–752. [PubMed: 19571811]

Kang JH, Wiggs JL, Rosner BA, Hankinson SE, Abdrabou W, Fan BJ, Haines J, Pasquale LR. Endothelial nitric oxide synthase gene variants and primary open-angle glaucoma: Interactions with sex and postmenopausal hormone use. Invest Ophthalmol Vis Sci 2010;51(2):971–979. [PubMed: 19815736]

Kraft P, Haiman CA. GWAS identifies a common breast cancer risk allele among BRCA1 carriers. Nat Genet 2010;42(10):819–820. [PubMed: 20877320]

Kraft P, Hunter D. Integrating epidemiology and genetic association: The challenge of gene-environment interaction. Phil Trans R Soc B 2005;360:1609–1616. [PubMed: 16096111]

Kraft P, Zeggini E, Ioannidis JPA. Replication in genome-wide association studies. Stat Sci 2009;24(4):561–573. [PubMed: 20454541]

Li Y, Abecasis GR. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet 2006:S79–2290.

Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet 2009;10:387–406. [PubMed: 19715440]

Lin DY, Sullivan PF. Meta-analysis of genome-wide association studies with overlapping subjects. Am J Hum Genet 2009;85:862–872. [PubMed: 20004761]

Lindgren CM, Heid IM, Randall JC, Lamina C, Steinthorsdottir V, Qi L, Speliotes EK, Thorleifsson G, Willer CJ, Herrera BM, Jackson AU, Lim N, Scheet P, Soranzo N, Amin N, Aulchenko YS, Chambers JC, Drong A, Luan J, Lyon HN, Rivadeneira F, Sanna S, Timpson NJ, Zillikens MC, Zhao JH, Almgren P, Bandinelli S, Bennett AJ, Bergman RN, Bonnycastle LL, Bumpstead SJ, Chanock SJ, Cherkas L, Chines P, Coin L, Cooper C, Crawford G, Doering A, Dominiczak A, Doney ASF, Ebrahim S, Elliott P, Erdos MR, Estrada K, Ferrucci L, Fischer G, Forouhi NG, Gieger C, Grallert H, Groves CJ, Grundy S, Guiducci C, Hadley D, Hamsten A, Havulinna AS, Hofman A, Holle R, Holloway JW, Illig T, Isomaa B, Jacobs LC, Jameson K, Jousilahti P, Karpe F, Kuusisto J, Laitinen J, Lathrop GM, Lawlor DA, Mangino M, McArdle WL, Meitinger T, Morken MA, Morris AP, Munroe P, Narisu N, Nordström A, Nordström P, Oostra BA, Palmer CNA, Payne F, Peden JF, Prokopenko I, Renström F, Ruokonen A, Salomaa V, Sandhu MS, Scott LJ, Scuteri A, Silander K, Song K, Yuan X, Stringham HM, Swift AJ, Tuomi T, Uda M, Vollenweider P, Waeber G, Wallace C, Walters GB, Weedon MN, The Wellcome Trust Case Control Consortium; Witteman JCM, Zhang C, Zhang W, Caulfield MJ, Collins FS, Smith GD, Day INM, Franks PW, Hattersley AT, Hu FB, Jarvelin MR, Kong A, Kooner JS, Laakso M, Lakatta E, Mooser V, Morris AD, Peltonen L, Samani NJ, Spector TD, Strachan DP, Tanaka T, Tuomilehto J, Uitterlinden AG, van Duijn CM, Wareham NJ, Watkins H, for the PROCARDIS Consortia; Waterworth DM, Boehnke M, Deloukas P, Groop L, Hunter DJ, Thorsteinsdottir U, Schlessinger D, Wichmann HE, Frayling TM, Abecasis GR, Hirschhorn JN, Loos RJF, Stefansson K, Mohlke KL, Barroso I, McCarthy MI, for the GIANT consortium. Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. PLoS Genet 2009;5(6):e1000508. [PubMed: 19557161]

Lindström S, Yen YC, Spiegelman D, Kraft P. The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. Hum Hered 2009;68:171–181. [PubMed: 19521099]

Liu Y, Blackwood DH, Caesar S, de Geus EJC, Farmer A, Ferreira MAR, Ferrier IN, Fraser C, Gordon-Smith K, Green EK, Grozeva D, Gurling HM, Hamshere ML, Heutink P, Holmans PA, Hoogendijk WJ, Hottenga JJ, Jones L, Jones IR, Kirov G, Lin D, McGuffin P, Moskvina V, Nolen WA, Perlis RH, Posthuma D, Scolnick EM, Smit AB, Smit JH, Smoller JW, St Clair D, van Dyck R, Verhage M, Willemsen G, Young AH, Zandbelt T, Boomsma DI, Craddock N, O'Donovan MC, Owen MJ, Penninx BWJH, Purcell S, Sklar P, Sullivan PF, Wellcome Trust Case-Control Consortium. Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. Mol Psychiatry 2011 Mar 30;16:2–4. [PubMed: 20351715]

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007;39(10):1181–1186. [PubMed: 17898773]

Manolio TA, Rodriguez LL, Brooks L, Abecasis G, the Collaborative Association Study of Psoriasis; Ballinger D, Daly M, Donnelly P, Faraone SV, the International Multi-Center ADHD Genetics Project; Frazer K, Gabriel S, Gejman P, the Molecular Genetics of Schizophrenia Collaboration; Guttmacher A, Harris EL, Insel T, Kelsoe JR, the Bipolar Genome Study; Lander E, McCowin N, Mailman MD, Nabel E, Ostell J, Pugh E, Sherry S, Sullivan PF, the Major Depression Stage 1 Genomewide Association in Population-Based Samples Study; Thompson JF, Warram J, the Genetics of Kidneys in Diabetes (GoKinD) Study. Wholley D, Milos PM, Collins FS. New models of collaboration in genome-wide association studies: The Genetic Association Information Network. Nat Genet 2007;39(9):1045–1051. [PubMed: 17728769]

Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39(7):906–913. [PubMed: 17572673]

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nature 2008;9:356–369.

Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JI, Uitterlinden AG, Harris TB, Witteman JCM, Boerwinkle E, on behalf of the CHARGE Consortium. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2009;2:73–80. [PubMed: 20031568]

Psychiatric GWAS Consortium Coordinating Committee. Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, Lehner T, Levinson DF, Moran A, Sklar P, Sullivan PF. Genomewide association studies: History, rationale, and prospects for psychiatric disorders. Am J Psychiatry 2009;166(5):540–556. [PubMed: 19339359]

Rimm EB, Stampfer MJ, Colditz GA, Chute CG, Litin LB, Willett WC. Validity of self-reported waist and hip circumferences in men and women. Epidemiology 1990;1(6):466–473. [PubMed: 2090285]

Seminara D, Khoury MJ, O'Brien TR, Manolio T, Gwinn ML, Little J, Higgins JPT, Bernstein JL, Boffetta P, Bondy M, Bray MS, Brenchley PE, Buffler PA, Casas JP, Chokkalingam AP, Danesh J, Smith GD, Dolan S, Duncan R, Gruis NA, Hashibe M, Hunter D, Jarvelin MR, Malmer B, Maraganore DM, Newton-Bishop JA, Riboli E, Salanti G, Taioli E, Timpson N, Uitterlinden AG, Vineis P, Wareham N, Winn DM, Zimmern R, Ioannidis JPA, for the Human Genome Epidemiology Network and the Network of Investigator Networks. The emergence of networks in human genome epidemiology: Challenges and opportunities. Epidemiology 2007;18(1):1–8. [PubMed: 17179752]

Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, Sulem P, Rafnar T, Esko T, Walter S, Gieger C, Rawal R, Mangino M, Prokopenko I, Mägi R, Keskitalo K, Gudjonsdottir IH, Gretarsdottir S, Stefansson H, Thompson JR, Aulchenko YS, Nelis M, Aben KK, den Heijer M, Dirksen A, Ashraf H, Soranzo N, Valdes AM, Steves C, Uitterlinden AG, Hofman A, Tönjes A, Kovacs P, Hottenga JJ, Willemsen G, Vogelzangs N, Döring A, Dahmen N, Nitz B, Pergadia ML, Saez B, De Diego V, Lezcano V, Garcia-Prats MD, Ripatti S, Perola M, Kettunen J, Hartikainen AL, Pouta A, Laitinen J, Isohanni M, Huei-Yi S, Allen M, Krestyaninova M, Hall AS, Jones GT, van Rij AM, Mueller T, Dieplinger B, Haltmayer M, Jonsson S, Matthiasson SE, Oskarsson H, Tyrfingsson T, Kiemeney LA, Mayordomo JI, Lindholt JS, Pedersen JH, Franklin WA, Wolf H, Montgomery GW, Heath AC, Martin NG, Madden PAF, Giegling I, Rujescu D, Järvelin MR, Salomaa V, Stumvoll M, Spector TD, Wichmann HE, Metspalu A, Samani NJ, Penninx BW, Oostra BA, Boomsma DI, Tiemeier H, van Duijn CM, Kaprio J, Gulcher JR, The ENGAGE Consortium. McCarthy MI, Peltonen L, Thorsteinsdottir U, Stefansson K. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat Genet 2010;42(5):448–453. [PubMed: 20418888]

Thorleifsson G, Walters GB, Hewitt AW, Masson G, Helgason A, DeWan A, Sigurdsson A, Jonasdottir A, Gudjonsson SA, Magnusson KP, Stefansson H, Lam DSC, Tam POS, Gudmundsdottir GJ, Southgate L, Burdon KP, Gottfredsdottir MS, Aldred MA, Mitchell P, St Clair D, Collier DA, Tang N, Sveinsson O, Macgregor S, Martin NG, Cree AJ, Gibson J, MacLeod A, Jacob A, Ennis S, Young TL, Chan JCN, Karwatowski WSS, Hammond CJ, Thordarson K, Zhang M, Wadelius C, Lotery AJ, Trembath RC, Pang CP, Hoh J, Craig JE, Kong A, Mackey DA, Jonasson F, Thorsteinsdottir U, Stefansson K. Common variants near CAV1 and CAV2 are associated with primary open-angle glaucoma. Nat Genet 2010;42(10):906–909. [PubMed: 20835238]

Wolfs RCW, Klaver CCW, Ramrattan RS, van Duijn CM, Hofman A, de Jong PTVM. Genetic risk of primary open-angle glaucoma. Arch Ophthalmol 1998;116:1640–1645. [PubMed: 9869795]

Zeggini E, Ioannidis JPA. Meta-analysis in genome-wide association studies. Pharmacogenomics 2009;10(2):191–201. [PubMed: 19207020]

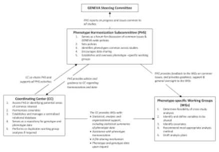
GENEVA Steering Committee
PHS reports on progress and issues common to all studies
Phenotype Harmonization Subcommittee (PHS)
1. Serves as a forum for discussion of common issues & GENEVA-wide policies
2. Sets policies
3. Identifies phenotypes common across studies
4. Encourages data sharing
5. Establishes and oversees phenotype –specific working groups
CC co-chairs PHS and supports all PHS activities
PHS provides advice and guidance to CC regarding harmonization and data
PHS provides feedback to the WGs on common issues, and provides guidance, support & general oversight to the WGs
Coordinating Center (CC)
1. Assists PHS in identifying potential areas of common interest
2. Harmonizes covariates
3. Establishes and manages a centralized relational database
4. Serves as a repository for genotype and phenotype data
5. Performs or duplicates working group analyses if required
The CC provides WGs with:
• Statistical, analytic and organizational support, including statistical summaries of phenotype data
• Assistance with phenotype harmonization
• A file sharing mechanism
• Phenotype and genotype data upon request
Phenotype-specific Working Groups (WGs)
1. Determine feasibility of cross-study analyses
2. Identify and define variables to be shared
3. Identify covariates
4. Recommend most appropriate analysis method
5. Draft analysis plans

Figure 1. Phenotype harmonization roles and responsibilities in GENEVA

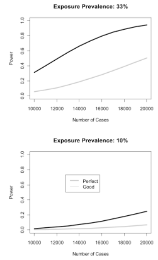

Figure 2.

Power of a case-control study to detect a gene-environment interaction (departure from a multiplicative odds model) when the binary exposure is measured perfectly or via a good proxy with 77% specificity and 99% sensitivity (roughly analogous to self-reported versus measured overweight status)

This figure illustrates several points: a) large samples sizes are needed to detect gene-environment interactions; b) even modest misclassification can greatly decrease the power of tests for gene-environment interaction (and the relative decrease is greater for rare exposures); yet c) a large study using the proxy can have greater power than a smaller study using the perfect measure. This last point is important when the perfect measure is prohibitively expensive or only available on a small fraction of samples, while the good measure is relatively inexpensive or already available on many samples. Power calculations were performed using the methods described in Lindstrom et al. (2009), assuming a rare disease (prevalence 1 in 1,000), no main effect for the binary genetic factor (with 20% prevalence), an odds ratio of 1.5 for the exposure, an interaction odds ratio of 1.35, and a Type I error rate of 5×10^{-8}.

Table 1
Phenotype harmonization and analysis plan check list

Items may be addressed concurrently and not necessarily in this order.

Establish a group and group leader

- Identify a leader interested and willing to drive the process forward
- Include representatives from all studies and groups contributing data and any other groups (such as a coordinating center) contributing to the work

Identify common phenotypes of interest (including covariates)

- Review available data—what was collected and what can be shared
  - Review data collection forms and questionnaires
  - Create a spreadsheet of phenotypes showing which studies collect which variables
  - Create a web-based or other survey, if appropriate

Determine feasibility of cross-study analyses

- Identify participating studies
  - Are the subjects representative of the cohort
- Identify approximate numbers of subjects to be included in cross-study analyses
  - Is the combined total number of subjects sufficient to detect a significant association
- Review consent status of each contributing study
  - Do planned analyses fall within the scope of the data use statements of the consent forms used by each study
- Review each study's inclusion and exclusion criteria
  - Are there studies for which the inclusion/exclusion criteria make their inclusion in the cross-study analysis not useful
- Review data definitions from each contributing study
  - What exactly was asked (or measured)—compare the wording of the actual questions
  - Are response options comparable
- Compare data distributions
  - If data distributions are not similar, i.e. if there is little overlap, the data may not be comparable
- Confirm that the data are not being used for similar analyses by another collaboration or consortium

Prepare common definitions

- Prepare phenotype definition
  - Define the outcome of interest and values
  - Define the outcome type—whether discrete, ordinal, continuous

Code variables

- Create algorithms for converting each study's raw data to conform to the agreed upon definitions

Draft an analysis plan

- Describe the phenotype, covariates, and inclusion/exclusion criteria
- List participating studies and anticipated number of subjects contributing to the analyses
- Describe the planned subgroup analyses
- Describe the analysis approach
  - How will outliers be handled
    - ♦ Will outliers be Windsorized (i.e. equated to the next highest/lowest value)

    - Will extreme values be truncated (e.g. to 4 standard deviations of sample mean)

- How will missing values be handled
- How will covariates be adjusted for
- How will data from longitudinal studies (studies with data from multiple time points) be combined with data from cross-sectional studies
    - Will a date range be chosen for which data values to include
- Determine which statistical model fits best; often several models are needed
- Identify type of analysis, i.e. meta-analyses of summary statistics or analyses of pooled individual data
- Describe the statistical support required and who will provide it
- Describe the imputation plan
    - Who will be responsible for imputation, i.e. will this be done by individual studies or will it be done centrally
- Identify any candidate genes that can be examined
- Identify key personnel and their roles
    - Who at each site and/or the CC will be doing analyses
    - Who will be the primary author
    - Will any assistance be required from the CC
    - What will be the role of the CC
- Develop a timeline—genotype and phenotype data for different studies may not be ready at the same time
    - When can analyses start and what studies will be included
    - How will studies whose data will be ready later be incorporated into the analysis
- Relationship to work being done in other consortia/collaborations
    - How do these analyses fit in with what other consortia are doing
    - Can collaborations be established with another consortium to combine data for a primary analysis
    - Can the other consortium's data serve as a replication study
    - What specific associations and G*E analyses can be performed that have not been already examined

Table 2

Table 2a. Examples of possible smoking-related questions

Study (N)	Smoking-related questions	Possible responses
Study 1 (2,500)	1. Do you currently smoke cigarettes? 2. If yes, how many cigarettes per day?	Y/N ###
Study 2 (1,200)	1. Have you smoked more than 100 cigarettes in your lifetime? 2. If yes, do you currently smoke? 3. If yes, how many packs per day do you smoke?	Y/N Y/N ##
Study 3 (8,500)	1. Have you ever smoked?	Y/N
Study 4 (1,250)	1. Do you currently smoke?	Y/N
Study 5 (4,200)	1. Do you smoke? 2. When did you first start smoking regularly?	Y/N Past year; 1-5 years ago; >5 years ago
Study 6 (6,600)	1. Have you smoked tobacco in the past month?	Y/N
Study 7 (800)	1. Have you ever smoked regularly? 2. If yes, do you still smoke? 3. If yes, how much do you smoke a day?	Y/N Y/N 1-10 cigarettes, 11-20 cigarettes, 21-30 cigarettes, >30 cigarettes

Table 2b. Examples of possible new variables for cross-study analyses

New variable	Studies that could contribute data	Total N	Comment
Cigarettes per day	Study 1	2,500	Data from Study 7 might also be included if specific values were assigned to each response category, e.g. 5 for category ‘1-10 cigarettes’, 15 for category ‘11-20 cigarettes’, and so on
Packs per day	Study 1 (if convert cigarettes/day to packs/day) Study 2 Study 7 (if convert categories to packs/day)	4,500	
Former smoker	Study 2 Study 7	2,000	
Ever smoker	Study 2 Study 3 Study 7	10,500	Requires ability to determine if subjects are former smokers
Current smoker	Study 1 Study 2 Study 4	16,550	

Table 2b. Examples of possible new variables for cross-study analyses

New variable	Studies that could contribute data	Total N	Comment
	Study 5 Study 6 (if current smoker is defined as having smoked in the past month) Study 7		

Table 3

Selected Consortia and Collaborations

Study name and website	Brief description
1. **BPC3** (Breast and Prostate Cancer and Hormone-Related Gene Variant Study)	This study pools data and biospecimens from 10 large prospective cohorts to conduct research on gene-environment interactions in cancer etiology
2. **CADISP** (Cervical Artery Dissections and Ischemic Stroke Patients) http://www.chazard.org/cadisp	A European Consortium performing research on ischemic stroke in the young and in particular on cervical artery dissection (the most common cause of ischemic stroke in young people)
3. **CALiCo Consortium** (Genetic Epidemiology of Causal Variants Across the Life Course) https://www.pagestudy.org/index.php/studies/58-calico	A consortium of well characterized population based studies and a central genotyping and resequencing core laboratory, to accelerate the understanding of the role and population impact of putative causal genetic variants related to complex diseases
4. **CARe** (Candidate-gene Association REsource) http://www.broad.mit.edu/gen_analysis/care/index.php/Main_Page	An NHLBI collaboration of up to 50,000 participants from nine cohort studies whose outputs will be analytic results from statistical and computational methods used to perform large-scale candidate gene association studies of phenotypes across multiple cohorts, whole-genome association studies, and tests for gene-gene and gene-environment interactions
5. **CGASP** (Consortium of Genetic Association of Smoking Related Phenotypes)	A consortium sponsored by NIDA (National Institute on Drug Abuse) to conduct a meta-analysis as well as integrate data on the genetics of nicotine addiction, lung cancer, and COPD
6. **CHARGE** (Cohorts for Heart and Aging Research in Genomic Epidemiology) http://web.chargeconsortium.com	Formed to facilitate genome-wide association study meta-analyses and replication opportunities among multiple large and well-phenotyped longitudinal cohort studies
7. **CKDGen Consortium**	A team of researchers from the United States and Europe investigating the role of genes in the etiology of common forms of kidney disease
8. **COGENT** (COlorectal cancer GENeTics)	An international consortium to study the role of polymorphic variation on the risk of colorectal cancer
9. **DentalSCORE** (Dental Strategies Concentrating on Risk Evaluation)	A collaborative study adding an oral and dental exam component to HeartSCORE, a community-based longitudinal study focusing on racial and socioeconomic disparities in cardiovascular risk
10. **DGI** (Diabetes Genetics Initiative) http://www.broad.mit.edu/diabetes	A collaboration of the Broad Institute of MIT and Harvard, Lund University, and Novartis Institutes for BioMedical Research to identify the genetic determinants of type 2 diabetes. This collaboration aims to collect and analyze samples from type 2 diabetic patients from nations across the globe, performing whole genome scans to provide a comprehensive view of the DNA sequence variants associated with the disease
11. **DIAGRAM** (Diabetes Genetics Replication And Meta-analysis Consortium) http://www.well.ox.ac.uk/DIAGRAM/index.html	A genome-wide association study of type 2 diabetes (T2D) from the FUSION, DGI, and WTCCC/UKT2D groups
12. **eMERGE** (Electronic Medical Records & Genomics) https://www.mc.vanderbilt.edu/victr/dcc/projects/acc	A national consortium formed to develop, disseminate, and apply approaches to research that combine DNA biorepositories with electronic medical record (EMR) systems for large-scale, high-throughput genetic research
13. **ENGAGE** (European Network of Genomic and Genetic Epidemiology) http://www.euengage.org	A 5-year research project started in January 2008 and funded with 12 million euros by the European Commission under the 7th Framework Programme-Health Theme. The ENGAGE Consortium includes 23 research organizations and two biotechnology and pharmaceutical companies across Europe and in Canada and Australia. Its goal is to identify large numbers of novel susceptibility genes that influence metabolic, behavioral and cardiovascular traits, and to study the interactions between genes and life style factors. It is led by Leena Pettonen at the University of Helsinki
14. **EUROCRAN** (European Collaboration on Craniofacial Anomalies) http://www.eurocran.org/default.asp	Funded by the European Commission (EC) under the Quality of Life theme of Framework Programme V, EUROCRAN brings together researchers from a range of clinical / scientific disciplines from 19 European centers with the shared aim of improving the management and understanding of craniofacial anomalies (CFA)

Study name and website	Brief description
15. **GAPPS** (Global Alliance to Prevent Prematurity and Stillbirth) http://gappsseattle.org	GAPPS collaborates to help catalyze research to understand causes and find solutions to preterm delivery. GAPPS is developing a repository of data and specimens from diverse women to provide a resource for researchers around the world
16. **GARNET** (Genomics and Randomized Trials Network, or Genome-wide Association Research Network into Effects of Treatment) http://www.garnetstudy.org/Default.aspx	A series of genome-wide association studies of treatment response in randomized clinical trials that looks to identify genetic variants associated with response to treatments for conditions of clinical or public health significance
17. **GEFOS** (Genetic Factors of Osteoporosis Consortium) http://www.gefos.org	A large international collaboration that proposes to capitalize on the success of GENOMOS by using Genome Wide Association (GWA) analysis with high density SNP arrays to identify common risk gene variants for osteoporosis
18. **GENEVA** (GENe EnVironment Association studies) http://www.genevastudy.org	The genetics component of an NIH-wide initiative that aims to accelerate understanding of genetic and environmental contributions to health and disease, its aims are to identify genetic variants related to common, complex diseases, identify variations in gene-trait associations related to environmental exposures, and address potential pathways to disparities in health outcome
19. **GIANT** (Genome-wide Investigation of ANThropometric measures) http://www.helmholtz-muenchen.de/epi/beitraege-zu-netzwerken/giant-genomewide-investigation-of-anthropometric-measures/index.html	An international assembly of study investigators and genetic epidemiologists which was established to pool genome-wide association (GWA) results on anthropometric parameters such as body-mass-index and height. The GIANT consortium has the aim to determine all relevant genes that are involved in modulating weight, height and more complex measures of obesity
20. **Global BPGen Consortium**	A multinational collaboration with investigators from across the US and Europe that seeks to identify genetic variants in known and novel genes that influence blood pressure in the general population
21. **Global Lipid Genetics Consortium**	The aim of this consortium is to study the genetic determinants of blood low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol and triglycerides
22. **ILCCO** (International Lung Cancer Consortium) http://ilcco.iarc.fr	An international group of lung cancer researchers established in 2004 with the aim of sharing comparable data from ongoing lung cancer case-control and cohort studies from different geographical areas and ethnicities
23. **International Type 2 Diabetes Consortium** http://csg.sph.umich.edu/consortium	A consortium of groups mapping genes for NIDDM in diverse populations that has come together to carry out a joint analysis of their linkage data
24. **ISGC** (International Stroke Genetics Consortium) http://www.strokegenetics.org	A consortium formed in early 2007 for the purpose of facilitating collaborative efforts to perform large scale, multi-center genetic studies in stroke. The aims include: 1) assembling a large well-characterized sample of stroke subjects with DNA, all of whom have been carefully phenotyped; and 2) harmonizing phenotype information to allow easy and reliable primary and secondary analyses of the combined cohort
25. **MAGIC** (The Meta-Analyses of Glucose and Insulin-related traits Consortium)	A collaborative effort to combine data from multiple GWAS to identify additional loci that affect glycemic and metabolic traits. The genetic studies of fasting glucose levels were originally organized as four distinct consortia: (i) European Network for Genetic and Genomic Epidemiology (ENGAGE), combining data from deCODE, Northern Finland Birth Cohort 1966 (NFBC1966), Netherlands Twins Register/Netherlands Study of Depression and Anxiety (NTR/NESDA) and the Rotterdam Study; (ii) Genetics of Energy Metabolism (GEM), a meta-analysis of the Lausanne (CoLaus) and TwinsUK scans; (iii) DFS, involving the Diabetes Genetics Initiative (DGI), Finland-United States Investigation of NIDDM Genetics (FUSION) and SardiNIA scans; and (iv) the Framingham Heart Study (FHS)
26. **NEIGHBOR** (National Eye Institute Glaucoma Human Genetics CollaBORation)	A consortium aiming to identify genetic determinants for primary open angle glaucoma (POAG)
27. **NGFN** (German National Genome Research Network)	An endeavor to research diseases that have a high incidence in Germany or that are particularly important for health policy due to the prolonged suffering and premature death of the people affected. These diseases include cancer, cardiovascular diseases, diseases of the nervous system, infections and inflammation as well as diseases linked to environmental factors
28. **P3G Consortium** (Public Population Project in Genomics) http://www.p3g.org	A non-profit international consortium founded in 2003 to respond to the growing needs and demands of the population genomics community, its goal is to foster harmonization of research tools developed by its members

Study name and website	Brief description
29. **PAGE** (Population Architecture using Genomics and Epidemiology) https://www.pagestudy.org	The purpose of this program is to provide support for the investigation, in well-characterized population studies, of genetic variants identified as potentially causally associated with complex diseases in genome-wide association (GWA) and other genetic studies, with the aim of widespread sharing of the resulting population-based descriptive and association data to accelerate the understanding of genes related to complex diseases
30. **PREGENIA** (Preterm Birth and Genetics International Alliances)	A consortium established to conduct research to understand genetic predisposition in preterm birth, a leading cause of neonatal morbidity and mortality. The aim is to promote research to better understand the genetic basis of preterm birth, and to develop genetic tools to predict women who are at risk for having preterm labor
31. **SHARe** (SNP Health Association Research) http://public.nhlbi.nih.gov/GeneticsGenomics/home/share.aspx	The NHLBI SHARe Project will conduct genome wide association studies and analyses in several large NHLBI Cohort studies to identify genes underlying cardiovascular and lung disease and other disorders like osteoporosis and diabetes
32. **SpiroMeta Consortium**	A consortium to facilitate large-scale meta-analysis of GWAS of lung function.
33. **SUNLIGHT Consortium** (Study of Underlying Genetic Determinants of Vitamin D and Highly Related Traits)	A consortium of 15 epidemiologic studies from the U.S., U.K., Canada, Netherlands, Sweden and Finland investigating the genetic contribution to vitamin D status in almost 32,000 white individuals of European descent
34. **TAG** (The Tobacco, Alcohol and Genetics Consortium)	The goal of the TAG Consortium is to conduct a genome-wide association study (GWAS) meta-analysis for smoking-related phenotypes using genotype and smoking phenotype data from existing GWAS of other traits e.g., heart disease, diabetes, etc. The sample size of the TAG Consortium may approach 100,000 subjects and is a multi-national collaboration which has the potential to identify novel genetic loci for smoking
35. **WTCCC** (Wellcome Trust Case-Control Consortium) http://www.wtccc.org.uk	A group of 50 research groups across the UK established in 2005, the WTCCC aims are to exploit progress in understanding of patterns of human genome sequence variation along with advances in high-throughput genotyping technologies, and to explore the utility, design and analyses of genome-wide association (GWA) studies

Table 4

GENEVA Phenotype Working Groups

Working Group	**Key variable(s) defined**	**No. of contributing studies so far**	**Approx. N**	**Comment**
Alcohol use	Drinks per week, reported maximum, as a continuous measure	7	18,000	First studies used for replication; Meta-analysis to be considered when data on later studies are available
Anthropometric measures	Height (m), continuous	7	42,000	
	Weight (kg) at study entry	6	36,000	
	Body Mass Index (BMI)	3	36,000	
	Weight (kg) at late adolescence or early adulthood	3	25,000	To be analyzed as part of a broader consortium with other, non-GENEVA studies
	Adult weight gain	3	25,000	To be analyzed as part of a broader consortium with other, non-GENEVA studies
	Waist circumference (cm) at study entry (also waist adjusted for BMI and height)	3	24,000	
	Hip circumference (cm) at study entry	2	23,000	
Caffeine use	Caffeine intake per day (mg/day)	3	25,000	Meta-analysis, including findings from subjects not part of GENEVA
Diabetes	*On hold*			
Hypertension	*Only two studies have actual blood pressure measures; discussing collaboration with another consortium*			
Oral health	Various dental caries and periodontal disease variables (under development)	3	TBD	
Physical activity and exercise	Metabolic equivalent moderate/vigorous leisure activity min/day (continuous); Engagement in moderate/vigorous leisure activity (dichotomized)	6	TBD	Planned meta-analysis, In-silico/de novo replication
Protective effects	Under development	10	TBD	
Psychiatric history	*Too much variance in the instruments to develop a meaningful harmonized variable across studies*			
Reproductive history	Age at menarche	8	18,000	Four studies collaborated with another consortium for analysis
	Age at menopause		8,500	
	Spontaneous loss	TBD	TBD	
Sleep	*New working group, under development*			
Tobacco use	G*E interactions	4	15,000	Under development

TBD = To be determined