Extracting Country-of-Origin from Electronic Health Records for Gene-Environment Studies as Part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Study

Eric Farber-Eger, BS[1], Robert Goodloe, MS[1], Jonathan Boston, BS[1], William S. Bush, PhD. MS[2], Dana C. Crawford, PhD[2]

[1]Center for Human Genetics Research, Vanderbilt University, Nashville, USA;
[2]Department of Epidemiology and Biostatistics, Institute for Computational Biology, Case Western Reserve University, Cleveland, OH, USA.

Abstract

We describe here the extraction of country-of-origin, an acculturation variable relevant for gene-environment studies, in a biorepository linked to de-identified electronic health records (EHRs) assessed by the Epidemiologic Architecture for Genes Linked to Environment (EAGLE), a study site of the Population Architecture using Genomics and Epidemiology (PAGE) I study. We extracted country-of-origin from the unstructured clinical free text using regular expressions within the MySQL relational database system in a cohort of 15,863 subjects of mostly non-European descent (including 11,519 African Americans, 1,702 Hispanics, and 1,118 Asians). We performed searches for 231 world countries (including independent sovereign states, dependent areas, and disputed territories) and common misspellings in >14 gigabytes of data including >13 billion characters of clinical text. Manual review of a fraction of the initial country-of-origin assignments established rules for data cleaning and quality control to achieve final country-of-origin status for each subject. After data cleaning, a total of 1,911/15,893 (12.02%) subjects were assigned to a country-of-origin outside of the United States. Mexico was the most commonly assigned country outside of the United States (264 subjects; 13.8% of subjects with a foreign country-of-origin assignment). The distribution of the countries assigned followed expectations based on known migration patterns to the United States with an emphasis on the southeastern region. These data suggest country-of-origin can be successfully extracted from unstructured clinical text for downstream genetic association studies.

Introduction

Human complex disease risk and trait distributions such as body mass index are impacted both by genetics and the environment. Decades of genetic linkage studies followed by more current genetic association studies have been successful in identifying thousands of genetic variants contributing to outcome risk or trait distributions[1, 2]. And, epidemiologic studies have been invaluable in identifying lifestyle exposures that impact overall human health. Yet despite these successes, the marriage of genetics and epidemiology has been less successful in identifying gene-environment interactions impacting risk of disease or trait distributions.

One of the difficulties in conducting gene-environment studies is the lack of harmonized phenotypic and environmental data available across large datasets. Indeed, most recent large gene-environment studies have examined either commonly collected “environmental” exposures such as sex[3, 4] or age[5, 6] or have used laboratory measures and biomarkers[7] that are easily standardized across labs and studies. Recently, extensive efforts have been made to harmonize other variables such as acculturation, smoking status, and physical activity available within consortiums[8], and tools have been developed to collect new or map existing variables[9] in a standardized manner for eventual pooled or meta-analysis studies[10].

To contribute to the harmonization and use of environmental data for gene-environment studies, we as the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) group as part of the larger the Population Architecture using Genomics and Epidemiology (PAGE) I study[11] are accessing large-scale epidemiologic and clinical collections for gene-environment studies. Unlike epidemiologic collections, environmental exposure data available in clinical collections such as biorepositories linked to de-identified electronic health records (EHRs) are difficult to extract given most of these data reside in unstructured fields of the EHR and are collected in highly variable manners. In this report, we describe the successful extraction of an acculturation variable, country-of-origin, from a biorepository linked to de-identified EHRs, a necessary initial step in developing needed resources and datasets for powerful next-generation genetic association studies for complex human diseases.

Methods

Study Population

The study population consists of 15,863 mostly non-European American subjects from BioVU, the Vanderbilt University Medical Center (VUMC) biorepository linked to de-identified EHRs. A description of the operations and oversight of BioVU has been previously published[12, 13]. In brief, BioVU followed an opt-out model where DNA was extracted from discarded blood samples drawn for routine clinical care in the outpatient setting. The DNA sample is linked to a de-identified version of the patient's EHR, collectively known as the Synthetic Derivative.

All non-European American samples in BioVU as of 2011 were genotyped on the Illumina Metabochip[14, 15] by the Vanderbilt University DNA Resources Core for the EAGLE group, a study site of the PAGE I study[11]. A major goal of the PAGE I study was to characterize genome-wide association study (GWAS)-identified index variants and genomic regions for various complex diseases and traits in populations of non-European descent. To contribute towards this goal, EAGLE genotyped 11,519 African Americans, 1,702 Hispanics, and 1,118 Asians among other mostly non-European samples identified in BioVU. This subset of BioVU genotyped on Metabochip[14], a custom array of 200,000 variants chosen from GWAS-index associations and fine-mapped regions based on the first iteration of the 1000 Genomes Project[16], is hereto referred as "EAGLE BioVU."[17] Population stratification was evaluated using EIGENSOFT[18] as detailed in Buyske et al[15]. Race/ethnicity is administratively assigned in BioVU[19, 20] and is available as a structured field. In contrast, country-of-origin data broadly defined as an individual's place of birth or nationality, when available, is not available as a structured field in BioVU.

Extraction approach

We searched all free text in the EHR representing >14 gigabytes of data including >13 billion characters of clinical text for a total of 231 world countries, including independent sovereign states, dependent areas, and disputed territories using an on-line list of countries accessed in 2013[21]. A list of common country misspellings was generated for each of the 231 countries[22]. The putative country name and misspellings were outputted along with 30 characters on either side of the spelling. The search was performed using regular expressions within the MySQL relational database system. The query operations required approximately four weeks to complete, with the bulk of this time due to the combinatorial misspellings included in the search.

Table 1. Key words used to filter output to assign probable country-of-origin

Born	Grow	Orig	Up
From	Home	Raised	
Grew	Live	Resided	

Manual review of the resulting output was performed for select countries with the most EHR mentions (such as India and Italy) to determine the rules for the filtering process. Based on the manual review, we automated the filtering of the output using the words in Table 1. Using SAS v9.3, for each word in Table 1 we required a blank space before and/or after the word to qualify as a phrase containing country of origin data. After the initial filtering step, we then developed a list of keywords associated with false country-of-origin for further filtering (Table 2). A fraction (n=177) of EAGLE BioVU subjects were assigned more than one country after these two filtering steps. For subjects with only two country-of-origin entries in the EHR, the first entry was retained. For subjects with more than two country-of-origin entries in the EHR, the most frequently mentioned country was retained. EAGLE BioVU subjects not assigned a country-of-origin were assumed to be from the United States by default. Choropleth maps were generated in SASv9.3 using the PROC GMAP procedure.

Results

We searched for mention of country-of-origin status using a list of 231 countries and misspellings among 15,863 mostly non-European descent subjects with DNA samples linked to de-identified EHRs genotyped on the Illumina Metabochip (EAGLE BioVU). The average number of spellings per country was 9.7, ranging from 1 spelling to 305 spellings (Central African Republic). On the first pass extraction, 179/231 (77.5%) countries included in the search were found with at least one mention in EAGLE BioVU. Of the 52 countries without data, the majority were island nations (59.6%) followed by ten African nations (19.2%), six European nations (11.5%), and five Asian nations (9.6%). As expected, countries not found among EAGLE BioVU tended to be small with respect to population including three island nations with no permanent inhabitants (Bouvet Island, South Georgia and South Sandwich Islands, and Svalbard and Jan Mayen Islands). Interestingly, six countries not found in EAGLE BioVU are large with respect to population (>5 million inhabitants), including North Korea, Burkina Faso, United

Table 2. Key words used to filter output associated with false country-of-origin assignment

Words	Context
" medicine"	Medication from foreign countries
" army" " deploy" " disabl" " hospital" " military" " service" " vet" " veteran" " war" " wound"	Military service-related
" Chad" " Kenya" " Wanda" (common misspelling for "Rwanda)	Person's name
" travel" " cruise" " vacation" " back" " going to" " return" " trip"	Travel-related
" sandwich" " hunting"	Turkey (the edible bird)

Arab Emirates, Tajikistan, Kyrgyzstan, and Turkmenistan. The lack of mention of these countries in EAGLE BioVU may be representative of the migration patterns to the United States, particularly the southeastern region. The lack of mention could also be associated with the extraction process. For example, we searched for United Arab Emeritus and misspellings, but we did not search for UAE, a common abbreviation for this country.

Before quality control and further filtering, almost all EAGLE BioVU subjects were assigned at least one country-of-origin (15,775/15,863; 99.4%). On average, subjects were assigned 11.3 countries (3.95 standard deviation), with a range of 1 country to 28 countries. The most frequently mentioned countries were located on the continents of Asia (99.65%), Europe (99.63%), South America (97.95%), and African (92.37%). The most frequently assigned countries-of-origin included India, Italy, Chile, Greece, and China, each with >93% of subjects being assigned to one or more of these countries.

Given that EAGLE BioVU is majority African American most likely from the southeastern region of the United States, the pre-quality control results likely do not accurately reflect the actual country-of-origin of subjects in EAGLE BioVU. Therefore, to more accurately estimate the country-of-origin of subjects, we first inspected the country-of-origin output individual-level data to better understand why most subjects were assigned to one of five countries. For India, we observed 914,796 mentions of "India" or one of its 15 misspellings. Of the 16 possible spellings for "India", "indi" returned the most country-of-origin individual-level data (741,931 mentions or 81.1%). Because the string "indi" is found in words commonly used in clinical notes (*indi*cated, contra*indi*cated, f*indi*ngs, *indi*vidual, etc.), the overwhelming majority of output returned for "indi" represented false positive country assignments. Similar observations were made for Italy ("ital": hosp*ital*, dig*ital*ly, v*ital*, mar*ital*, gen*ital*, etc.), Chile ("chil": *chil*dren, *chil*ls, A*chil*les, etc.), Greece ("gree": Wal*gree*ns, a*gree*ment, de*gree*s, etc.), and China ("hia": hydrochlorot*hia*zide, psyc*hia*tric, brac*hia*l, etc.).

In addition to returning correctly spelled words commonly used in the clinical notes, we noted that the search for countries and misspellings returned commonly misspelled words in the clinical notes. For example, in the search for Chile, "chil" returned "chile" (a misspelling for "child") and "achiles" (a common misspelling for "Achilles"). Misspellings of "benign" (such as "benin", "bening", and "beningn") returned false positives for the African country of Benin. Other false country assignments were made when states or cities in the United States mimicked countries specified in the search (India for "Indiana"; Lebanon for "Lebanon, TN"; Colombia for "Columbia, TN" or "Columbia University") or when commonly used words or phrases mimicked countries (Hungary for "is hungry"; Turkey for "turkey sandwich"; Togo or Togolese Republic for "to go"; Reunion for "family reunion"; Greece for "cooking grease"; Guinea for "guinea pig").

We then filtered the country-of-origin output to remove highly likely false positive assignments using key words to extract qualifying phrases for further review (see Methods). After filtering, a total of 1,735 out of 15,863 (10.9%) EAGLE BioVU subjects were assigned a single country-of-origin. The remainder of subjects was assigned multiple countries (177) or no country (13,951). Subjects without a country-of-origin assignment were assumed to have originated from the United States (13,951/15,863; 87.9% of EAGLE BioVU).

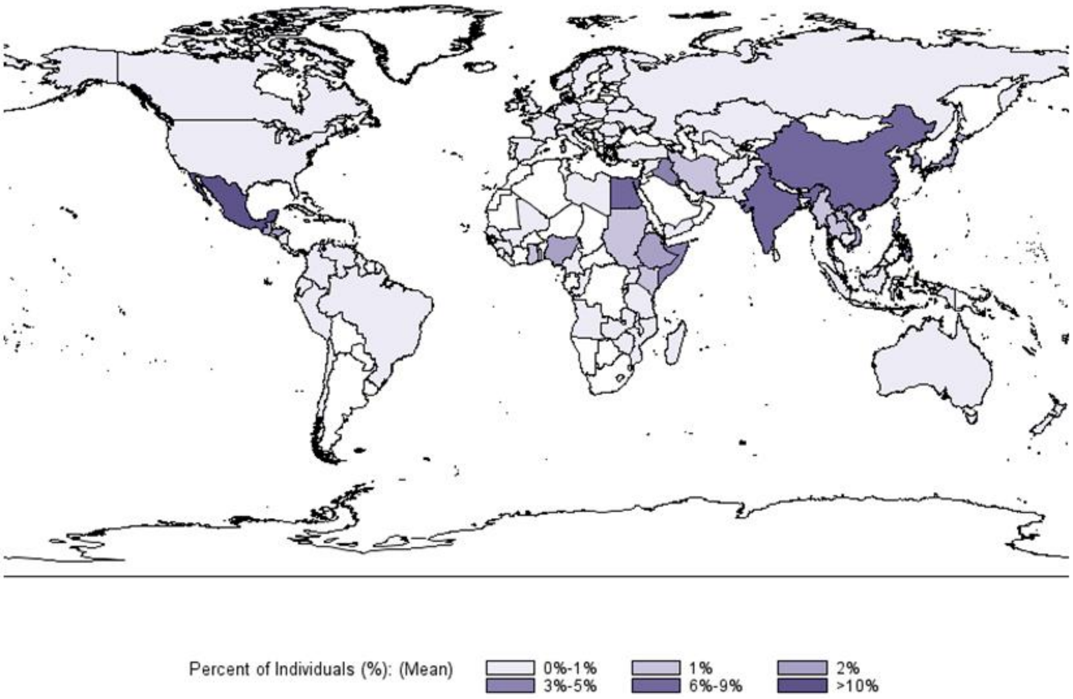

Figure 1. Percentage of EAGLE BioVU subjects assigned a foreign country-of-origin, by country.

Table 3. The most commonly assigned countries of origin in EAGLE BioVU outside the United States, by administratively assigned race/ethnicity

African Americans (n=353)	**Hispanics (n=438)**	**Asians (n=522)**	**Indians (n=60)**	**Others (n=539)**
Nigeria (13.0%)	Mexico (55.9%)	China (28.9%)	India (46.7%)	Egypt (25.0%)
Somalia (10.5%)	Honduras (10.3%)	South Korea (11.7%)	Egypt (8.3%)	India (9.9%)
Ethiopia (7.0%)	Guatemala (9.6%)	Vietnam (8.0%)	Bangladesh (6.7%)	Iraq (9.7%)
Ghana (5.7%)	Cuba (3.7%)	India (6.5%)	Mexico (5.0%)	Somalia (8.2%)
Haiti (4.8%)	Peru (2.7%)	Japan (5.7%)	Afghanistan (3.3%)	Iran (4.0%)
Jamaica (4.8%)	Ecuador (2.5%)	Myanmar (5.6%)	Ethiopia (3.3%)	Ethiopia (4.0%)
Kenya (4.5%)	Nicaragua (2.5%)	Laos (5.4%)	Iraq (3.3%)	Sudan (3.0%)

After assignment of a single country-of-origin to all possible EAGLE BioVU subjects, a total of 1,911 subjects were assigned to a country-of-origin outside of the United States (Figure 1). Perhaps not surprisingly given the region of the US, Mexico was the most commonly assigned country outside of the United States (264 subjects; 13.8% of EAGLE BioVU with a foreign country-of-origin assignment). The next most frequent assignment was made for China (8.8%), Egypt (7.9%), India (6.4%), and Somalia (4.2%). When stratified by administratively assigned race/ethnicity, the distribution of countries-of-origin differed by group (Table 2). For example, among African Americans, the most frequently assigned country other than the default assignment of "United States" was Nigeria (45) followed by Somalia (37), Ethiopia (24), Ghana (20), Haiti (17), Jamaica (17), and Kenya (16). For Hispanics, Asians, and Indians, 55.9%, 28.9%, and 46.6% of the EAGLE BioVU subjects with assigned country-of-origin outside the United States were assigned to Mexico, China, and India, respectively (Table 3).

Preliminary Evaluation

We reviewed 16 clinical records representing individuals assigned a country-of-origin outside the United States using the approach described above. Of these 16 records, three were not definite matches with the assigned country-of-origin. One may be a false positive related to travel ("Just returned from Australia and New Zealand"). Another may be a false positive related to language spoken ("Foreign language ALBANIAN"), but this could serve as an acceptable proxy for country-of-origin (that is, the patient speaks Albanian because he or she is from Albania). The possible mismatch is likely related to family history ("Mom is from Angola, speaks portugese" and "Maternal race: Angola"), although it is unclear from the clinical notes if the patient was born in the same country as the mother or born in the United States. The remaining 13 records were clearly born outside the United States (i.e., "originally from the Bahamas," "He immigrated from Mexico"). The positive predictive value of the approach outlined here ranges from 81.25 – 93.75% depending on how the three possible mismatches are classified.

Comparison to Genetic Ancestry

Most race/ethnicity in EAGLE/BioVU is administratively assigned, not self-identified[19, 20]. As evidenced by Table 3, many subjects labeled as "African American" are actually African. In addition to providing associated environmental data, acculturation variables such as country-of-origin can be useful in the quality control and evaluation of population stratification for downstream genetic association studies. As shown in in Figure 2, most subjects administratively assigned African American in EAGLE BioVU cluster together based on genetic ancestry, as expected given the predominant western African ancestry of the subjects. Nevertheless, country-of-origin was able to explain at least one of the outliers: an eastern African subject mislabeled as African American (Figure 2).

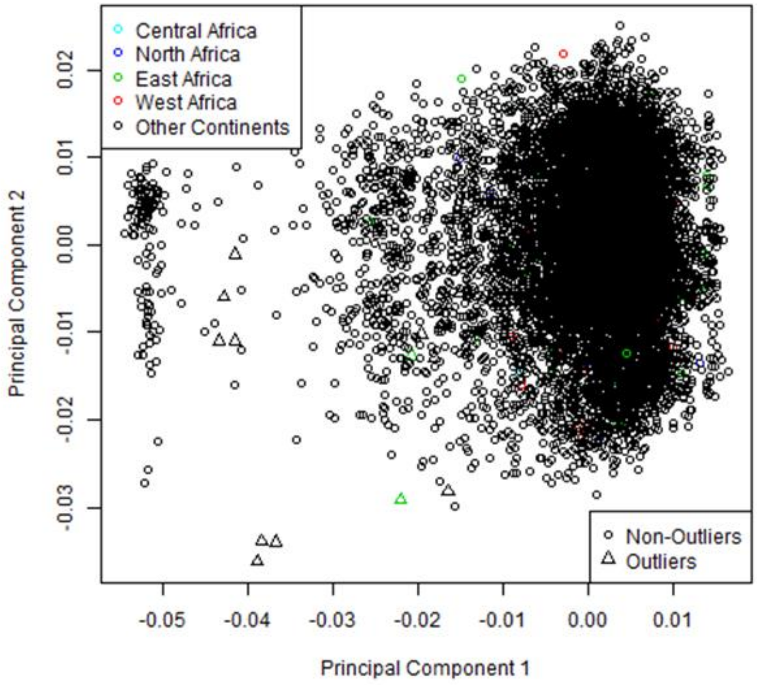

Figure 2. Evaluation of population stratification among African Americans in EAGLE BioVU. Principal components (PC) 1 (x-axis) and 2 (y-axis) from EIGENSOFT are plotted for each African American using HapMap Yoruba (representing west Africa) as the anchor. Country-of-origin is color coded, and PC outliers are marked with a triangle.

Conclusion

We have successfully developed bioinformatic approaches to assign country-of-origin to >15,000 subjects from a biorepository linked to de-identified EHRs genotyped on Metabochip for gene-environment studies in EAGLE. Overall, 1,911 of the 15,863 subjects in EAGLE BioVU were assigned a country-of-origin outside of the United States. The distribution of countries represented in race/ethnicity strata followed expectations based on current migration patterns to the southeastern United States. For example, more than half of the Hispanics in EAGLE BioVU were identified as originating from Mexico, slightly less than the 61% reported in the 2010 US Census for Davidson county Tennessee[23]. On a similar note, China and India were countries most frequently extracted for Asian and Indian EAGLE BioVU subjects assigned a country-of-origin outside the United States.

These country-of-origin data lay the foundation for gene-environment studies involving acculturation variables in EAGLE BioVU. Acculturation, defined as the modification of beliefs and/or behaviors by an individual or group when introduced to another group or environment, is an important exposure variable with a long history in epidemiologic studies. For example, acculturation studies of Latinos representing 20 Spanish-speaking countries have suggested that lower acculturation or adoption of American behaviors was associated with healthier diets and lower body-weight outcomes[24]. Greater acculturation or adoption of American behaviors has also been associated with increased risk for type 2 diabetes[24, 25], cardiovascular disease[26], and breast cancer[27] among other outcomes. In general, acculturation can add valuable data associated with socioeconomic status[28], diet, physical activity, and smoking and drinking habits for gene-environment studies.

Country-of-origin also provides valuable data towards the evaluation of population stratification and use of race/ethnicity in genetic association studies. Race/ethnicity in BioVU is administratively assigned and is not self-identified, the latter of which is a common data collection approach taken by epidemiologic studies. Previous work in BioVU demonstrated that the concordance for subjects assigned "white" (European American) or "black" (African American) compared with genetic ancestry assignments made using ancestry informative genetic markers (AIMs) was high[19, 20]. In the absence of routinely collected data for nationality, EAGLE BioVU subjects administratively assigned African American or Hispanic may actually be Ethiopian or Cuban, respectively, for example. These same DNA samples may appear as outliers in principal component analysis given that we often compare African Americans to HapMap Yoruba from west Africa and Hispanics to either HapMap Mexican Americans from Los Angeles or HapMap Han Chinese and Japanese. The added country-of-origin data assist in confirming outliers and help guide their reassignment so that all the data can be properly analyzed, particularly for trans-ethnic genetic association studies now emerging in the literature.

Despite success in extracting country-of-origin from this biorepository linked to de-identified EHRs, there are several limitations worth noting. First, the while a preliminary evaluation suggested the algorithm was adequate in correctly identifying foreign-born patients when mentioned in the clinical record, it is likely that other foreign-born subjects exist in this biorepository but were missed because the clinic does not routinely collect these data in a standardized manner. The PhenX Toolkit[10] recommends asking study participants "where were you born?". For those participants residing in the United States, the PhenX Toolkit recommends recording the state; for all others, the Toolkit recommends recording the U.S. territory or name of the foreign country. It is possible that the intake staff asked the same question as recommended by the PhenX Toolkit. However, given the question itself is not documented, it is unclear that all VUMC patients were asked the same question or asked a question about country-of-origin at all. In fact, it is probable that only a fraction of patients is asked a question about nationality or country of birth possibly based on a clinician's observation such as the patient's accent or the need for an interpreter. It is clear, however, that intake staff do not distinguish between American Indian (Native American) and Indian from the subcontinent of India highlighting the limitations of racial/ethnic labels in the clinic. It is also clear from the free text search that the recorded patient answers are highly variable and may depend on the patient's or intake staff's knowledge of geography or race/ethnicity. For example, multiple subjects were associated with more than one country-of-origin, and these countries bordered one another (Colombia/Venezuela; Guatemala/Mexico; Honduras/Mexico; Iran/Iraq; Kenya/Somalia; Somalia/Ethiopia; Sudan/Egypt). It may be that the patient mentioned his/her true country-of-origin but changed it to suite the knowledge of the intake staff. It may also be that these patients intentionally obfuscated his/her country of origin for some unknown reason. As noted in Methods, for these cases we used either the first mention of country of origin or the most frequently mentioned country of origin, neither of which can be confirmed even with manual review.

Another limitation of the search implemented in the present study is that it was limited to search terms using the English language. Not all subjects in EAGLE BioVU speak English as their primary language. In the event that non-English speaking patients describe his/her country-of-origin using a foreign language, our approach would not have identified them. An expanded list of country names in foreign languages could be used to search for these subjects, but most likely at the expense of computational time. At four weeks' running time, the list of countries limited to English and common misspellings was already lengthy and time consuming. Therefore, expansion of search terms would require additional computational resources and parallelization. Alternatively, we could apply more sophisticated name recognition entity approaches that have already been developed for multiple languages (such as those available at http://stanfordnlp.github.io/CoreNLP/).

Despite the numerous limitations, the strengths of this approach include the ease of its implementation and its relative accuracy of identifying countries-of-origin after minimal manual review. While these data were limited to EAGLE BioVU non-European descent samples, we envision expanding this effort to other EHRs as a resource for gene-environment studies.

Acknowledgements

This work was supported in part by NIH grant 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 funded by the National Center for Research Resources, Grant UL1 RR024975-01, which is now at the National Center for Advancing Translational Sciences, Grant 2 UL1 TR000445-06.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 2009;106(23):9362-7.
2. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Research. 2009;37(suppl 1):D793-D6.
3. Ober C, Loisel DA, Gilad Y. Sex-specific genetic architecture of human disease. Nat Rev Genet. 2008;9(12):911-22.
4. Weiss LA, Pan L, Abney M, Ober C. The sex-specific genetic architecture of quantitative traits in humans. Nat Genet. 2006;38(2):218-22.
5. Dumitrescu L, Brown-Gentry K, Goodloe R, Glenn K, Yang W, Kornegay N, et al. Evidence for Age As a Modifier of Genetic Associations for Lipid Levels. Annals of Human Genetics. 2011;75(5):589-97.
6. Graff M, Gordon-Larsen P, Lim U, Fowke JH, Love S, Fesinmeyer M, et al. The influence of obesity related SNPs on BMI across the life course: the PAGE Study. Diabetes. 2013;62(5):1763-7.
7. Dumitrescu L, Goodloe R, Brown-Gentry K, Mayo P, Allen M, Jin H, et al. Serum vitamins A and E as modifiers of lipid trait genetics in the National Health and Nutrition Examination Surveys as part of the Population Architecture using Genomics and Epidemiology (PAGE) study. Hum Genet. 2012;131(11):1699-708.
8. Bennett SN, Caporaso N, Fitzpatrick AL, Agrawal A, Barnes K, Boyd HA, et al. Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. Genetic Epidemiology. 2011;35(3):159-73.
9. Pendergrass SA, Verma SS, Holzinger ER, Moore CB, Wallace J, Dudek SM, et al. Next-generation analysis of cataracts: determining knowledge drive gene-gene interactions using Biofilter, and gene-environment interactions using the PhenX Toolkit. Pac Symp Biocomput. 2013:147-58.
10. Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, et al. The PhenX Toolki: get the most from your measures. Am J Epidemiol. 2011;174(3):253-60.
11. Matise TC, Ambite JL, Buyske S, Carlson CS, Cole SA, Crawford DC, et al. The Next PAGE in Understanding Complex Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. American Journal of Epidemiology. 2011;174(7):849-59.
12. Pulley JM, Denny JC, Peterson JF, Bernard GR, Vnencak-Jones CL, Ramirez AH, et al. Operational Implementation of Prospective Genotyping for Personalized Medicine: The Design of the Vanderbilt PREDICT Project. Clin Pharmacol Ther. 2012;92(1):87-95.
13. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. Clin Pharmacol Ther. 2008;84(3):362-9.

14. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. PLoS Genet. 2012;8(8):e1002793.

15. Buyske S, Wu Y, Carty CL, Cheng I, Assimes TL, Dumitrescu L, et al. Evaluation of the Metabochip Genotyping Array in African Americans and Implications for Fine Mapping of GWAS-Identified Loci: The PAGE Study. PLoS ONE. 2012;7(4):e35651.

16. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.

17. Crawford DC, Goodloe R, Farber-Eger E, Boston J, Pendergrass SA, Haines JL, et al. Leveraging epidemiologic and clinical collections for genomic studies of complex traits. Human Heredity. 2015;79(3-4):137-46.

18. Price AL. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38.

19. Dumitrescu L, Ritchie MD, Brown-Gentry K, Pulley JM, Basford M, Denny JC, et al. Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. Genet Med. 2010;12(10):648-50.

20. Hall JB, Dumitrescu L, Dilks HH, Crawford DC, Bush WS. Accuracy of Administratively-Assigned Ancestry for Diverse Populations in an Electronic Medical Record-Linked Biobank. PLoS ONE. 2014;9(6):e99161.

21. Countries and Regions of the World from A to Z. Available from: www.nationsonline.org/oneworld/countries_of_the_world.htm.

22. How Do You Spell. Available from: www.spellweb.com.

23. 2010 US Census.

24. Perez-Escamilla R. Acculturation, nutrition, and health disparities in Latinos. The American Journal of Clinical Nutrition. 2011;93(5):1163S-7S.

25. Perez-Escamilla R, Putnik P. The Role of Acculturation in Nutrition, Lifestyle, and Incidence of Type 2 Diabetes among Latinos. The Journal of Nutrition. 2007;137(4):860-70.

26. Daviglus ML, Talavera GA, Avil+¬s-Santa M. Prevalence of major cardiovascular risk factors and cardiovascular diseases among hispanic/latino individuals of diverse backgrounds in the united states. JAMA. 2012;308(17):1775-84.

27. John EM, Phipps AI, Davis A, Koo J. Migration History, Acculturation, and Breast Cancer Risk in Hispanic Women. Cancer Epidemiology Biomarkers & Prevention. 2005;14(12):2905-13.

28. Hollister BM, Restrepo NA, Farber-Eger E, Crawford DC, Aldrich MC, Non A. Development and performance of text-mining algorithms to extract socioeconomic status from de-identified electronic health records. Pacific Symposium on Biocomputing. 2016;22:230-241.