TESTING POPULATION-SPECIFIC QUANTITATIVE TRAIT ASSOCIATIONS FOR CLINICAL OUTCOME RELEVANCE IN A BIOREPOSITORY LINKED TO ELECTRONIC HEALTH RECORDS: *LPA* AND MYOCARDIAL INFARCTION IN AFRICAN AMERICANS

LOGAN DUMITRESCU

Center for Human Genetics Research, Vanderbilt University, 519 Light Hall, 2215 Garland Avenue, Nashville, TN 37232, USA
Email: logandumitrescu@gmail.com

KIRSTEN E. DIGGINS

Cancer Biology, Vanderbilt University, 742 Preston Research Building, 2220 Pierce Avenue, Nashville, TN 37232, USA
Email: kirsten.e.diggins@vanderbilt.edu

ROBERT GOODLOE

Center for Human Genetics Research, Vanderbilt University, 519 Light Hall, 2215 Garland Avenue, Nashville, TN 37232, USA
Email: robert.goodloe@gmail.com

DANA C. CRAWFORD

Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University, Wolstein Research Building, 2103 Cornell Road, Suite 2527, Cleveland, OH 44106, USA
Email: dana.crawford@case.edu

Previous candidate gene and genome-wide association studies have identified common genetic variants in *LPA* associated with the quantitative trait Lp(a), an emerging risk factor for cardiovascular disease. These associations are population-specific and many have not yet been tested for association with the clinical outcome of interest. To fill this gap in knowledge, we accessed the epidemiologic Third National Health and Nutrition Examination Surveys (NHANES III) and BioVU, the Vanderbilt University Medical Center biorepository linked to de-identified electronic health records (EHRs), including billing codes (ICD-9-CM) and clinical notes, to test population-specific Lp(a)-associated variants for an association with myocardial infarction (MI) among African Americans. We performed electronic phenotyping among African Americans in BioVU ≥40 years of age using billing codes. At total of 93 cases and 522 controls were identified in NHANES III and 265 cases and 363 controls were identified in BioVU. We tested five known Lp(a)-associated genetic variants (rs1367211, rs41271028, rs6907156, rs10945682, and rs1652507) in both NHANES III and BioVU for association with myocardial infarction. We also tested *LPA* rs3798220 (I4399M), previously associated with increased levels of Lp(a), MI, and coronary artery disease in European Americans, in BioVU. After meta-analysis, tests of association using logistic regression assuming an additive genetic model revealed no significant associations ($p<0.05$) for any of the five *LPA* variants previously associated with Lp(a) levels in African Americans. Also, I4399M rs3798220 was not associated with MI in African Americans (odds ratio

= 0.51; 95% confidence interval: 0.16 – 1.65; p=0.26) despite strong, replicated associations with MI and coronary artery disease in European American genome-wide association studies. These data highlight the challenges in translating quantitative trait associations to clinical outcomes in diverse populations using large epidemiologic and clinic-based collections as envisioned for the Precision Medicine Initiative.

1. Introduction

Labs ordered in a clinical setting provide valuable diagnostic and prognostic data at the individual patient level. In a research setting, labs can be studied to better understand the biological basis of clinical outcomes. As an example, lipid labs such as low-density lipoprotein cholesterol (LDL-C) are frequently ordered in a clinical setting to monitor the cardiovascular disease risk in patients. In turn, these labs or quantitative traits have been extensively studied in genomic research settings to identify genetic variants predictive of extreme LDL-C levels and cardiovascular disease risk [1].

A major advantage of quantitative trait genetic studies compared with case-control outcome studies is sample size resulting in statistical power [2]. As a result, there are more or larger genome-wide association studies (GWAS) and significant findings for lipid traits compared with cardiovascular disease outcomes [1], particularly for diverse populations. The emergence of electronic health records (EHRs) linked to biorepositories, however, provides contemporary opportunities to apply quantitative trait genetic variants to assess clinical relevance with an eye towards precision medicine.

We describe here the application of *LPA* genetic variants, previously associated with Lp(a) levels [3], to assess myocardial infarction associations in both an epidemiologic and clinical African American population. Lipoprotein (a) [Lp(a)] is considered an emerging biomarker or risk factor for cardiovascular disease [4-6] whose relationship with cardiovascular disease varies across races/ethnicities. Elevated plasma Lp(a) levels have been reported to be associated with cardiovascular disease in European Americans but have not been clearly documented in African Americans [7]. Paradoxically, among participants with no previous history of cardiovascular disease, the mean Lp(a) level is two- to three-fold higher in African Americans compared with European Americans [8,9]. The underlying cause(s) for this difference has not yet been determined.

Recent studies have identified common SNPs in *LPA* as strongly associated with Lp(a) levels, explaining up to 36% of the trait variance in populations of European-descent [10,11]. In a recent epidemiologic study conducted in the Third National Health and Nutrition Examination Survey (NHANES III), we demonstrated that *LPA* common genetic variants were associated with Lp(a) levels in a population-specific manner [3]. *LPA* SNP rs3798220 (I4399M) has also been associated with cardiovascular disease [11-14] and severe cardiovascular disease [15] in several European-descent populations. Thus, common genetic variants in *LPA* are strong predictors of both Lp(a) levels and cardiovascular disease risk in at least one population. We test here whether

LPA variants associated with Lp(a) levels in African Americans are associated with myocardial infarction in African Americans ascertained from epidemiologic and clinical settings.

2. Methods

2.1. *Study population*

The study populations presented here include the epidemiologic Third National Health and Nutrition Examination Survey (NHANES III) and the clinical BioVU, Vanderbilt University Medical Center's biorepository linked to de-identified electronic health records. NHANES III is a cross-sectional survey conducted between 1988 and 1994 by the National Center for Health Statistics at the Centers for Disease Control and Prevention. NHANES ascertained non-institutionalized Americans regardless of health status. Demographic, health, and lifestyle data were collected on NHANES participants through surveys, labs, and physical exams in the Mobile Examination Center (MEC). DNA is available on consenting phase 2 participants (ascertained between 1991 and 1994). The present study was approved by the CDC Ethics Review Board. Because the study investigators did not have access to personal identifiers, this study was considered non-human subjects research by the Vanderbilt University Internal Review Board.

BioVU operations [16] and ethical oversight [17] have been previously described. Briefly, DNA is extracted from discarded blood drawn for routine clinical care at Vanderbilt outpatient clinics in Nashville, Tennessee and surrounding areas. The DNA samples are linked to a de-identified version of the patient's EHR. The de-identified version of the EHR is referred to as the "Synthetic Derivative." The data in this study were de-identified in accordance with provisions of Title 45, Code of Federal Regulations, part 46 (45 CFR 46); therefore, this study was considered non-human subjects research by the Vanderbilt University Internal Review Board.

2.2. *Phenotyping*

Race/ethnicity in NHANES III is self-identified, which is concordant with global genetic ancestry for non-Hispanic whites and non-Hispanic blacks [18]. Myocardial infarction (MI) case status in NHANES III was based on data collected from a physical examination, administered by a physician, in the MEC. A continuous cardiac infarction/injury score (CIIS) was calculated based on 12 lead electrocardiogram (ECG) multiplied by 10. Those participants with a CIIS ≥ 20 were considered to have probable infarction/injury and those with a CIIS < 20 but ≥ 15 were considered to have possible infarction/injury. These thresholds correspond to an estimated specificity level of 98% and 95% [19], respectively. Our NHANES III MI case definition included participants classified as having possible or probable infarction/injury.

Race/ethnicity in BioVU is administratively assigned, which is highly concordant with genetic ancestry for European Americans and African Americans [20,21]. The de-identified EHR in BioVU contains both structured (International Classification of Diseases, Ninth Revision, Clinical

Modification billing codes [ICD-9-CM]; current procedural terminology codes; problems lists; labs) and unstructured (clinical free text) data that are accessible for electronic phenotyping. We explored five different electronic phenotyping strategies to identify cases of MI using mentions of ICD-9-CM codes (Table 1) among African American adults ≥ 40 years of age. MI case review was performed in 2013 using the browser search function in the Synthetic Derivative user interface to find the following keywords in the patient's clinical notes: myocardial infarction, MI, infar, STEMI, and NSTEMI. If none of the keywords were found in the record, the case reviewer searched for ICD-9-CM code 410 in the record and extracted the clinic visit date associated with the ICD-9-CM code. The case reviewer then searched the remainder of the patient's records on that clinic visit date for evidence of an MI. The ECG records of all possible cases were also accessed for review. Patients were considered unconfirmed for MI if EHR review failed to identify evidence of MI in the patient's medical history. Unconfirmed cases of MI were excluded from genotyping as cases. Positive predictive values (PPVs) were calculated as the total number of confirmed cases divided by the total number of potential cases. A total of 311 MI cases were identified for genotyping in BioVU.

Controls in BioVU were defined as African American adults ≥ 40 years of age with no mention of ICD-9-CM codes of MI (410) or any other codes relating to ischemic heart disease (ICD-9-CM 411-414). A total of 5,883 potential controls were identified in BioVU. Controls were frequency matched to cases by age and sex prior to selection for genotyping.

2.3. *Genotyping*

Genotyping in NHANES III was performed using the Illumina GoldenGate assay (as part of a custom 384 OPA) by the Center for Inherited Disease Research (CIDR) through the National Heart Lung and Blood Institute's Resequencing and Genotyping Service, as previously described [3]. Vanderbilt Technologies for Advanced Genomics (VANTAGE) genotyped BioVU samples for six *LPA* SNPs (rs3798220, rs41271028, rs6907156, rs10945682, rs1652507, and rs1367211) using Sequenom. Genotyping quality control for NHANES III was performed using SAS version 9.2 and for BioVU using PLINK [22].

2.4. *Statistical methods*

Tests of association in NHANES III were performed using SAS version 9.2 (SAS Institute, Cary, NC). Each *LPA* variant associated with Lp(a) levels in non-Hispanic blacks [3] was tested for association with MI status (dependent variable) using logistic regression assuming an additive genetic model adjusting for 1) age and sex and 2) age, sex, and ln(Lp[a]+1). Data was accessed remotely from the CDC's Research Data Center (RDC) in Hyattsville, Maryland using Analytic Data Research by Email (ANDRE) [23]. In BioVU, tests of association between MI (the dependent variable) and *LPA* SNPs were performed with PLINK using logistic regression assuming an additive genetic model and adjusting for age and sex (Lp[a] levels are not available in

BioVU). Meta-analyses were performed with METAL using a fixed-effects inverse-variance weighted approach [24], and results were visualized using Synthesis-View [25,26].

3. Results

As noted in the Methods section, cases of MI in NHANES III were identified using a continuous cardiac infarction/injury score applied to ECGs, an exam administered by public health professionals as part of the survey.

Table 1. Phenotyping criteria and case review results for five definitions of myocardial infarction based on mentions of billing codes. Overall, a total of 311 individual cases of confirmed MI were identified and 297 had sufficient DNA for genotyping. Abbreviations: International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM); positive predictive value (PPV).

Case definition	Phenotyping criteria	Potential cases	Confirmed cases	PPV
1	ICD-9-CM code 410.* on 3 consecutive days	108	107	99.1%
2	ICD-9-CM code 410.* on 2 consecutive days	159	158	99.4%
3	≥ 3 ICD-9-CM codes 410.* ever	159	158	99.4%
4	≥ 2 ICD-9-CM codes 410.* ever	209	205	98.1%
5	≥ 1 ICD-9-CM codes 410.* ever	355	311	87.6%

In contrast, we used electronic phenotyping approaches to extract cases of MI from EHRs of African American patients. We used ICD-9-CM billing codes in various combinations in an attempt to achieve the largest samples size possible with acceptable PPV. As might be expected, the most stringent case definitions (Table 1; definitions 1 and 2) where codes for MI were required on three and two consecutive days identified the fewest number of cases at PPVs >99% after manual review. These cases of MI likely represent incident inpatient cases of MI in BioVU at the time of data extraction. Equally high in PPV but low in case count was case definition 3 where three or more ICD-9-CM codes were required. The least stringent case definitions 4 and 5 yielded the most confirmed cases (205 and 311, respectively) at acceptably high PPVs (Table 1). The high PPVs observed here are consistent with other studies examining the accuracy of using ICD-9-CM codes to identify cases of acute MI [27-30]. Of the 311 total cases identified in the Synthetic

Derivative, only 265 passed quality control after genotyping (12 had insufficient DNA for genotyping; two were compromised samples; 14 failed genotyping).

Table 2. Association between myocardial infarction and Lp(a)-associated SNPs in non-Hispanic blacks from NHANES III. A total of 19 SNPs were tested for an association with Lp(a) levels [3] and MI in non-Hispanic blacks from NHANES III. *LPA* SNPs associated with MI at $p < 0.05$ are shown here. MI case status was defined as participants with possible or probable cardiac infarction/injury (CIIS score ≥ 15). Associations with MI and transformed Lp(a) levels were performed unweighted using logistic and linear regression, respectively. [1]Adjusted for age and sex. [2]Adjusted for age, sex, and Lp(a) levels. [3]Associations between Lp(a) and *LPA* in non-Hispanic blacks in NHANES III as reported in Dumitrescu et al 2011 [3]. Abbreviations: confidence interval (CI); odds ratio (OR).

SNP	**Lp(a) levels[1, 3]** n = 1,711		**MI[1]** n_{cases} = 93 $n_{controls}$ = 522		**MI[2]** n_{cases} = 91 $n_{controls}$ = 498	
	β (95% CI)	**p-value**	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**
rs41271028	-0.06 (-0.18, 0.07)	0.3608	2.12 (1.01, 4.46)	***0.0470***	2.12 (1.01, 4.45)	***0.0476***
rs6907156	0.15 (0.05, 0.25)	0.0031	0.53 (0.30, 0.92)	***0.0231***	0.53 (0.30, 0.92)	***0.0241***
rs10945682	-0.14 (-0.21, -0.06)	0.0003	1.41 (1.00, 1.98)	***0.0481***	1.37 (0.97, 1.93)	0.0725
rs1652507	-0.45 (-0.59, -0.32)	1.06 x10^{-10}	1.88 (1.06, 3.34)	***0.0308***	1.79 (1.00, 3.2)	0.0510
rs1367211	-0.27 (-0.34, -0.20)	3.67 x10^{-14}	1.46 (1.04, 2.07)	***0.0306***	1.42 (1.00, 2.01)	0.0518

We previously tested 19 *LPA* SNPs for an association with Lp(a) levels in non-Hispanic blacks in NHANES III, 12 of which were associated at $p < 0.0001$ [3]. We tested here the same 19 *LPA* SNPs for an association with MI in non-Hispanic blacks from NHANES III. Despite the limitation of small sample size (n_{cases} = 93), five *LPA* variants (rs1367211, rs1652507, rs6907156, rs10945682, and rs41271028) were associated with risk of MI in non-Hispanic blacks at p < 0.05 (Table 2; Figure 1). Interestingly, four of the five alleles associated with *increased* risk of MI were also associated with *decreased* Lp(a) levels at p < 0.003.

To determine if the association of *LPA* variants and MI in non-Hispanic blacks from NHANES III was due to the putative causal role of increased Lp(a) levels in coronary artery disease, we adjusted for Lp(a) level (Table 2). This adjustment attenuated the associations for three of the five single-SNP associations with MI (p = 0.05, 0.05, and 0.07 for rs1367211, rs1652507, and rs10945682, respectively). In contrast, two *LPA* SNPs (rs6907156 and rs41271028) remained associated with MI at p < 0.05 after adjustment for Lp(a) levels. The first SNP, rs6907156, was originally associated with decreased Lp(a) levels (p = 0.0031) while the second, rs41271028, was not (p = 0.36).

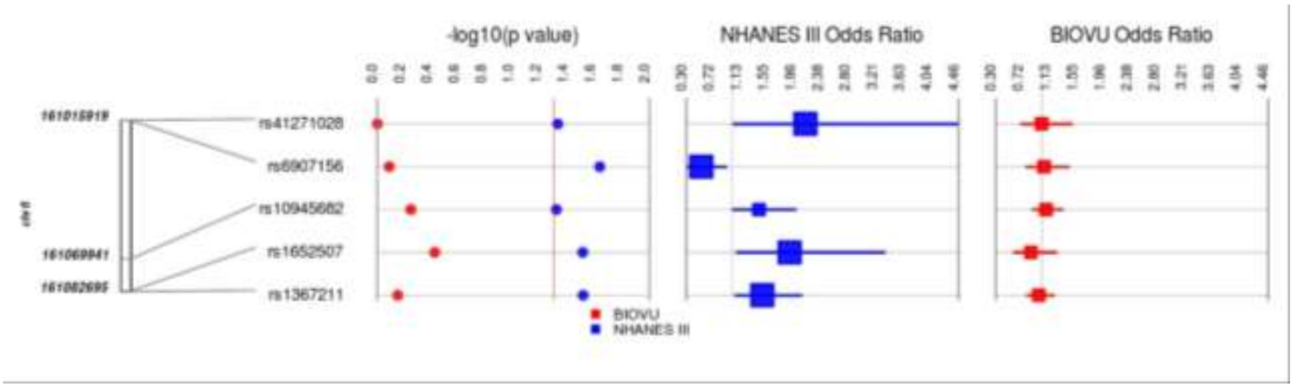

Figure 1. Synthesis-View plot of associations between Lp(a)-associated variants and myocardial infarction in non-Hispanic blacks from NHANES III and African Americans from BioVU. Tests of association were performed in non-Hispanic blacks NHANES III (n = 93 cases; n = 522 controls) and African Americans from BioVU (n = 265 cases; n = 343 controls) for myocardial infarction and each of five *LPA* SNPs previously associated with Lp(a) levels in NHANES III non-Hispanic blacks. Analyses were performed using logistic regression assuming an additive genetic model adjusted for age and sex. Odds ratios, 95% confidence intervals, and $-\log_{10}$(p-values) are plotted by study (NHANES III in blue and BioVU in red) in a forest plot generated by Synthesis-View. The red line denotes a level of significance at p = 0.05. Significant odds ratios are denoted by the larger squares.

To replicate the NHANES III findings, we genotyped all five *LPA* SNPs from Table 2 in BioVU African American cases (n = 265) of MI and controls (n = 343). None of the five tests of association were significant at p < 0.05 (Figure 1). We meta-analyzed NHANES III and BioVU tests of association and found two (rs10945682 and rs41271028) of the five associations had consistent directions of effect. However, none of the five *LPA* SNPs were associated with MI in the meta-analysis at p < 0.05.

Previous studies have suggested that rs3798220 (I4399M) is associated with higher Lp(a) levels and coronary artery disease risk [31,32] in European-descent populations. In the present study, we genotyped rs3798220 in BioVU African American MI cases and controls to determine if the association generalized to African-descent populations. In unadjusted tests of association, rs3798220 was not associated with MI in BioVU African Americans (odds ratio = 0.51; 95% CI: 0.16 – 1.65; p-value = 0.26).

4. Discussion

Genome-wide association studies have identified thousands of common variants significantly associated with quantitative traits, and a fraction of these, in turn, are associated with risk for specific clinical outcomes [33]. The emergence of EHRs linked to biorepositories is enabling larger clinical outcome association studies, an important step in translating quantitative trait associations into precision medicine efforts. In the present study, we tested genetic variants associated with Lp(a) levels, an emerging risk factor for cardiovascular disease, for associations with MI in African Americans ascertained from epidemiologic and clinic-based collections. Overall, Lp(a)-associated genetic variants were not associated with MI in this small sample of African Americans, highlighting the challenges of translating strong genetic associations identified for quantitative traits to clinically relevant outcomes such as cardiovascular disease.

The lack of statistical associations may be due to a combination of imprecise phenotyping and small sample size. Indeed, precise phenotyping and phenotype harmonization across studies is a major challenge for genetic association studies. We used both an epidemiologic cross-sectional survey and an EHR-based biorepository to identify cases and controls of MI. Prevalent NHANES III cases were based on ECG scores (as opposed to self-report), and BioVU cases were likely a mixture of prevalent and incident cases identified using primarily billing codes. As detailed above, we deliberately applied the least stringent MI case definition to identify the largest number of cases for manual review, a strategy that identifies "silver standard" cases that can then be combined with gold standard cases to potentially boost statistical power [34]. This silver standard strategy is likely to play a major role in the Precision Medicine Initiative as it is anticipated that the one million participants will be a combination of prevalent and incident cases of various common disease drawn from epidemiologic and clinic-based collections [35].

In the meta-analysis, we assumed that the case and control definitions were roughly equivalent. While there were a few tests of heterogeneity with p-values < 0.05 (such as *LPA* rs1652507 at p = 0.03), none were statistically significant when accounting for multiple tests. Despite the lack of evidence for gross differences in case-control definitions between the two collections, it is likely that subtle differences and possible case misclassification impacted statistical power. The NHANES III case-control definition has been shown to have high specificity, most likely resulting in good control classification (i.e., ruling out MI). Conversely, the BioVU case-control definition has high PPV or precision. Neither case-control definition addresses the underlying genetic and environmental heterogeneity typical of complex, common human diseases that, like misclassification of case-control status, decreases statistical power [36].

The present study had a small sample size, which in combination with imprecise phenotyping, led to low statistical power. The small sample size available for MI cases and controls in both the epidemiologic and EHR-based collection was disappointing given that both were drawn from relatively large collections of DNA linked to demographic and health information. Although there were 2,108 African American participants with biospecimens, NHANES III had relatively few cases of MI. At the time this study was conducted, BioVU contained DNA samples linked to EHRs for ~12,000 African Americans. However, from among these patients, only 311 cases of MI were identified through billing codes and of these, 265 were available for genotyping. It is possible that additional cases could have been identified using the clinical notes and more sophisticated natural language processing techniques, but it is doubtful that a sufficient number of additional cases would have been identified for a substantial increase in power. It is widely noted that genome-wide studies in general are not conducted in diverse populations [37], and those that are available generally have smaller sample sizes compared with their European counterparts [38]. This trend is unlikely to change with the rise in EHRs linked to biorepositories, and even concerted efforts such as the proposed Precision Medicine Initiative with one million participants [35] will be hard-pressed to muster sufficient sample sizes for clinically relevant outcomes in diverse populations.

In addition to imprecise phenotyping and low case-control counts available for study, low allele frequencies also contributed to low statistical power for specific tests of association. Overall, *LPA* rs3798220 is not common in African Americans, with a minor allele frequency of 2% in the present study. This minor allele frequency is similar to African Americans in third phase of the International HapMap Project as well as to European-descent cases and controls of coronary artery disease meta-analyzed by the CARDIoGRAM consortium [31]. The CARDIoGRAM consortium reported an odds ratio of 1.51, and assuming the same coded allele (C) and the same minor allele frequency (2%), we had ~17% power to detect an association with rs3798220 at $p < 0.05$.

Despite the numerous limitations, this study had several strengths. This study accessed both epidemiologic and clinic-based collections to identify cases and controls for MI among African Americans. Continued case-control collection for this and other clinically relevant outcomes is sorely needed to better translate genetic associations identified using quantitative traits to prevention, diagnosis, and treatment options for MI and other forms of cardiovascular disease at the bedside.

5. Acknowledgments

Genotyping services were provided by the Johns Hopkins University under federal contract number (N01-HV-48195) from NHLBI. We would like to thank Dr. Geraldine McQuillan and Jody McClean for their help in accessing the Genetic NHANES III data. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. The dataset(s) used for the analyses described were also obtained from Vanderbilt University Medical Center’s BioVU, supported by institutional funding and by the Vanderbilt CTSA grant funded by the NCRR Grant UL1 RR024975-01 (now the NCATS, Grant 2 UL1 TR000445-06).

References

1. Global Lipids Genetics Consortium: **Discovery and refinement of loci associated with lipid levels.** *Nat Genet* 2013, **45:** 1274-1283.
2. Turner SD, Crawford DC, Ritchie MD: **Methods for optimizing statistical analyses in pharmacogenomics research.** *Expert Rev Clin Pharmacol* 2009, **2:** 559-570.
3. Dumitrescu L, Glenn K, Brown-Gentry K, Shephard C, Wong M, Rieder MJ *et al.*: **Variation in LPA Is Associated with Lp(a) Levels in Three Populations from the Third National Health and Nutrition Examination Survey.** *PLoS ONE* 2011, **6:** e16604.
4. Bennet A, Di Angelantonio E, Erqou S, Eiriksdottir G, Sigurdsson G, Woodward M *et al.*: **Lipoprotein(a) Levels and Risk of Future Coronary Heart Disease: Large-Scale Prospective Data.** *Arch Intern Med* 2008, **168:** 598-608.
5. Berglund L, Ramakrishnan R: **Lipoprotein(a): An Elusive Cardiovascular Risk Factor.** *Arterioscler Thromb Vasc Biol* 2004, **24:** 2219-2226.
6. Danesh J, Collins R, Peto R: **Lipoprotein(a) and coronary heart disease. Meta-analysis of prospective studies.** *Circulation* 2000, **102:** 1082-1085.
7. Heiss G, Schonfeld G, Johnson JL, Heyden S, Hames CG, Tyroler HA: **Black-white differences in plasma levels of apolipoproteins: the Evans County Heart Study.** *Am Heart J* 1984, **108:** 807-814.
8. The Emerging Risk Factors Collaboration: **Lipoprotein(a) Concentration and the Risk of Coronary Heart Disease, Stroke, and Nonvascular Mortality.** *JAMA: The Journal of the American Medical Association* 2009, **302:** 412-423.
9. Guyton JR, Dahlen GH, Patsch W, Kautz JA, Gotto AM, Jr.: **Relationship of plasma lipoprotein Lp(a) levels to race and to apolipoprotein B.** *Arterioscler Thromb Vasc Biol* 1985, **5:** 265-272.
10. Ober C, Nord AS, Thompson EE, Pan L, Tan Z, Cusanovich D *et al.*: **Genome-wide association study of plasma lipoprotein(a) levels identifies multiple genes on chromosome 6q.** *J Lipid Res* 2009, **50:** 798-806.
11. Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, Heath SC *et al.*: **Genetic variants associated with Lp(a) lipoprotein level and coronary disease.** *N Engl J Med* 2009, **361:** 2518-2528.
12. Luke MM, Kane JP, Liu DM, Rowland CM, Shiffman D, Cassano J *et al.*: **A polymorphism in the protease-like domain of apolipoprotein(a) is associated with severe coronary artery disease.** *Arterioscler Thromb Vasc Biol* 2007, **27:** 2030-2036.
13. Shiffman D, O'Meara ES, Bare LA, Rowland CM, Louie JZ, Arellano AR *et al.*: **Association of gene variants with incident myocardial infarction in the Cardiovascular Health Study.** *Arterioscler Thromb Vasc Biol* 2008, **28:** 173-179.
14. Shiffman D, Kane JP, Louie JZ, Arellano AR, Ross DA, Catanese JJ *et al.*: **Analysis of 17,576 Potentially Functional SNPs in Three Case-Control Studies of Myocardial Infarction.** *PLoS ONE* 2008, **3:** e2895.

15. Luke MM, Kane JP, Liu DM, Rowland CM, Shiffman D, Cassano J *et al.*: **A Polymorphism in the Protease-Like Domain of Apolipoprotein(a) Is Associated With Severe Coronary Artery Disease.** *Arterioscler Thromb Vasc Biol* 2007, ATVBAHA.
16. 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:** 362-369.
17. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR: **Principles of Human Subjects Protections Applied in an Opt-Out, De-identified Biobank.** *Clinical and Translational Science* 2010, **3:** 42-48.
18. Oetjens M, Brown-Gentry K, Goodloe R, Dilks HH, Crawford DC. **Population stratification in the context of diverse epidemiologic surveys** (in preparation).
19. Rautaharju PM, Warren JW, Jain U, Wolf HK, Nielsen CL: **Cardiac infarction injury score: an electrocardiographic coding scheme for ischemic heart disease.** *Circulation* 1981, **64:** 249-256.
20. 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:** 648-650.
21. 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:** e99161.
22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.*: **PLINK: a tool set for whole-genome association and population-based linkage analysis.** *Am J Hum Genet* 2007, **81:** 559-575.
23. Bush WS, Boston J, Pendergrass SA, Dumitrescu L, Goodloe R, Brown-Gentry K *et al.*: **Enabling high-throughput genotype-phenotype associations in the Epidemiology Architecture for Genes Linked to Environment (EAGLE) project as part of the Population Architecture using Genomics and Epidemiology (PAGE) study.** *Pac Symp Biocomput* 2013, **18:** 373-384.
24. Willer CJ, Li Y, Abecasis GR: **METAL: fast and efficient meta-analysis of genomewide association scans.** *Bioinformatics* 2010, **26:** 2190-2191.
25. Pendergrass S, Dudek SM, Roden DM, Crawford DC, Ritchie MD: **Visual integration of results from a large DNA biobank (BioVU) using synthesis-view.** *Pac Symp Biocomput* 2011, 265-275.
26. Pendergrass S, Dudek S, Crawford D, Ritchie M: **Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis.** *BioData Mining* 2010, **3:** 10.
27. Varas-Lorenzo C, Castellsague J, Stang MR, Tomas L, Aguado J, Perez-Gutthann S: **Positive predictive value of ICD-9 codes 410 and 411 in the identification of cases of acute coronary syndromes in the Saskatchewan Hospital automated database.** *Pharmacoepidem Drug Safe* 2008, **17:** 842-852.
28. Pladevall M, Goff DC, Nichaman MZ, Chan F, Famsey D, Ortiz C *et al.*: **An Assessment of the Validity of ICD Code 410 to Identify Hospital Admissions for Myocardial infarction:**

The Corpus Christi Heart Project. *International Journal of Epidemiology* 1996, **25:** 948-952.

29. Petersen LA, Wright S, Normand SL, Daley J: **Positive Predictive Value of the Diagnosis of Acute Myocardial Infarction in an Administrative Database.** *J Gen Intern Med* 1999, **14:** 555-558.
30. Thygesen S, Christiansen C, Christensen S, Lash T, Sorensen H: **The predictive value of ICD-10 diagnostic coding used to assess Charlson comorbidity index conditions in the population-based Danish National Registry of Patients.** *BMC Medical Research Methodology* 2011, **11:** 83.
31. Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, Holm H *et al*.: **Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease.** *Nat Genet* 2011, **43:** 333-338.
32. Tregouet DA, Konig IR, Erdmann J, Munteanu A, Braund PS, Hall AS *et al*.: **Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease.** *Nat Genet* 2009, **41:** 283-285.
33. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H *et al*.: **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.** *Nucleic Acids Research* 2014, **42:** D1001-D1006.
34. McDavid A, Crane PK, Newton KM, Crosslin DR, McCormick W, Weston N *et al*.: **Enhancing the Power of Genetic Association Studies through the Use of Silver Standard Cases Derived from Electronic Medical Records.** *PLoS ONE* 2013, **8:** e63481.
35. Collins FS, Varmus H: **A New Initiative on Precision Medicine.** *N Engl J Med* 2015, **372:** 793-795.
36. Manchia M, Cullis J, Turecki G, Rouleau GA, Uher R, Alda M: **The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases.** *PLoS ONE* 2013, **8:** e76295.
37. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M: **Genome-wide association studies in diverse populations.** *Nat Rev Genet* 2010, **11:** 356-366.
38. Kaufman JS, Dolman L, Rushani D, Cooper RS: **The Contribution of Genomic Research to Explaining Racial Disparities in Cardiovascular Disease: A Systematic Review.** *American Journal of Epidemiology* 2015, **181:** 464-472.