|  |  |  |
| --- | --- | --- |
| **eMERGE Network: External Collaborator Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT374 | |
| **Submission Date** | 01/23/2020 | |
| **Project Title** | Two-stage join analysis of top association for CAD in the Million Veteran Program | |
| **Tentative Lead Investigator** *(first author)* | Catherine Tcheandjieu | |
| **Tentative Senior Author**  *(last author)* | Themistocles (Tim) L. Assimes | |
| **eMERGE Site Sponsor & Contact** | Iftikhar Kullo | |
| **All Other Authors** | Partial list of lead junior and senior authors from MVP (there will be many others):  Junior: Austin Hilliard, Xiang Zhu, Shoa Clarke, Shining Ma, Valerio Napolioni  Senior: KM Chang, Kelly Cho, Peter Wilson, Phil Tsao, Chris O’Donnell  eMERGE authors: Ozan Dikilitas, Daniel Schaid, Iftikhar Kullo | |
| **Sites Participating** | All | |
| **Background / Significance** | Coronary artery disease (CAD) is a heritable condition with ~160 loci having been identified to date through GWAS involving predominantly white/European, South Asian, and East Asians populations. No locus has been identified to date among African Americans (AA) and Hispanics (HISP). | |
| **Outline of Project** | Investigators have conducted a large scale multi-ethnic GWAS of CAD in the Million Veteran Program involving up to 118,731 cases and 281,064 controls and have identified >70 novel loci through meta-analysis with existing datasets (e.g. Cardiogram+C4D and UK Biobank). However, a large majority of loci have reached genome wide significance only in whites or in our trans-ethnic meta-analysis. MVP investigators are seeking ‘replication’ of their most promising findings in their GWAS of African Americans (17,202 cases, 59,507 controls) and Hispanics (6,378 cases, 24,270 controls) in external datasets. | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  CPT codes  Phecodes  BMI | Common Variable Labs  Common Variable Meds  Other: Case/Control status on Phase I and Phase II phenotypes |
| **Other Desired Data *(Available from participating sites)*** | *Please specifically list out any data elements that participating sites would collect or extract from clinical or other sources for this project (i.e. not common variables above)* | |
| **Desired Genetic Data** | eMERGE I-III Merged set (HRC imputed, GWAS)  eMERGE PGx/PGRNseq data set  eMERGEseq data set (Phase III)  eMERGE Whole Genome sequencing data set  eMERGE Exome chip data set  eMERGE Whole Exome sequencing data set  Other (not listed above): | |
| **Does project pertain to an existing eMERGE Phenotype?** | Yes, if so please list    Coronary heart disease as defined by the presence of either   1. a revascularization or 2. a myocardial infarction (MI) 3. +/- softer CAD outcome as per eMERGE if available (see MVP definition below)   In MVP, an individual was classified as a case if he or she had ≥1 admission to a VA hospital with primary or secondary discharge diagnosis of acute myocardial infarction (AMI ICD 9 410, ICD-10: I21, I22)) OR ≥1 procedure code for revascularization of the coronary arteries OR ≥ 2 ICD outpatient codes for CAD on ≥ 2 dates (see list below). Individuals with only 1 ICD code for CAD on 1 outpatient encounter and no discharge diagnoses for AMI or revascularization procedures were excluded from the analyses. The remaining subjects were classified as controls.  ICD9 ICD10  410 I21  411.0 I22  411.1 I23  411.81 I24  411.89 I25.1  412 I25.2  414.00 I25.5  414.01 I25.6  414.02 I25.70  414.03 I25.71  414.04 I25.72  414.05 I25.73  414.2 I25.79  414.3 I25.810  414.4 I25.82  414.8 I25.83  414.9 I25.84  V45.81 I25.89  V45.82 I25.9  Z95.1  Z98.61    No | |
| **Planned Statistical Analyses** | MVP investigators identified SNPs for replication in external cohorts like eMERGE through the application of 2 analytic methods to their genetic data in AA and Hispanics:  1) classical GWAS analysis where all independent lead SNPs with p<1x 10-5 are carried forward along with ‘candidate’ causal SNPs that are in moderate-high LD (r2 >0.6) with lead SNPs even if their p value of some of these candidate SNPs in LD fall below <1x10-5  2) using XBEP. XPEB takes as input P-value summary statistics from two GWAS, a target-GWAS typically a smaller minority population of primary interest and a base-GWAS typically a much larger GWAS of Europeans and reprioritizes variants in the target population to compute local false discovery rates (for details of methods see Coram MA et al.  Am J Hum Genet. 2015).  The number of SNPs identified for replication in African Americans is ~430 SNPs while for Hispanics it is ~2750.  As a starting point, we will leverage existing harmonized phenotyping for CHD involving subjects with MI and revascularization procedures in eMERGE for the multi-ethinc polygenic risk score analyses to generate genetic association results for these SNPs of the top SNPs from Million Veteran Program GWAS of African Americans and Hispanics, respectively. Both prevalent and incident cases will be included in the analyses. We will then proceed with a second analyses that will include additional “soft” cases of CAD with definitions similar to those used in MVP. Logistic regression analyses will be conducted for each SNP on the CHD outcome adjusting for sex and the first 10 PCs for the respective race/ethnic group. Age adjustment is optional and if included in analyses should represent the age at the first ever ICD diagnostic or procedure code for CAD in the EHR for cases and age at the last EHR clinical encounter for controls. The minimum summary statistics for each SNP that will be provided to MVP investigators will include (minimum):  ID  CHROM (build 37)  POS (build 37)  MARKER\_NAME (rsID)  REF (reference allele)  ALT (alternate allele)  ALT\_AF (alternate allele freq)  ALT\_AC (alternate allele count)  N\_Cases (number of cases)  N\_controls (number of controls)  BETA  SE  PVALUE  Info (imputation quality)  MVP investigators will then perform a 2-stage joint analysis of these SNPs with the data fro MVP and other replication cohorts. A locus will be considered genome wide significant if the joint p value < 5 x 10-8. | |
| **Ethical Considerations** | None | |
| **Available Funding or Resources** |  | |
| **Target Journal** | Nature Medicine | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Analyses/look ups in January or early February 2020  Submit results to MVP in February. | |

**\*Common Variables available across all datasets:**

* Demographics: sex, year of birth, decade of birth, race, ethnicity
* Codes: (repeated values & age at event): ICD, CPT, Phecodes
* BMI: (repeated value & age at event) height, weight, BMI
* Labs: (lab name, repeated lab value & age at event) Serum total cholesterol, LDL, HDL, Triglycerides, Glucose fasting/non-fasting/unknown, & White Blood Cell count
* Medications: (medication name, repeated, & age at event) Cerivastatin sodium, Rosuvastatin, Simvastatin, Fluvastatin, Pravastatin, Lovastatin, Atorvastatin, & Pitavastatin
* Other: Case/Control status on Phase I and Phase II phenotype: only on GWAS dataset participants