|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT465 | |
| **Submission Date** | August 29, 2017 | |
| **Project Title** | Ancestry and admixture in the eMERGE-III cohort of 83,717 individuals | |
| **Tentative Lead Investigator** *(first author)* | Keyue Ding, Xiao Fan | |
| **Tentative Lead Investigator Email Address** | Ding.Keyue@mayo.edu | |
| **Tentative Senior Author**  *(last author)* | Mariza de Andrade, Iftikhar Kullo (Mayo Clinic) | |
| **All Other Authors** | Daniel Schaid, Gabriel Shaibi, Elena DeFillipis, Stephen Thibodeau, and other investigators from eMERGE sites. | |
| **Sites Participating** | All eMERGE sites | |
| **Background / Significance** | Genetically defined ancestry (e.g., European, African, and Asian) is an important covariate for conducting genome-wide association studies. Inference of admixed populations (e.g., Hispanics or Latinos) has important implications for localizing disease genes in populations of recently mixed ancestry by admixture mapping. Hispanics or Latinos have an admixture of European, African, and Native American ancestries, i.e., a three-way admixture. However, it is challenging to identify admixed Hispanics or Latinos as self-reported Hispanic, or Latino ethnicity may not conform to genetically defined Hispanic or Latino ethnicity. | |
| **Outline of Project** | First, we will perform population structure analysis for the eMERGE cohort by combining sequencing data from Africans (*n*=661), Europeans (*n*= 503), East Asian (*n*=504) in the 1000 Genome Project and genotype data from a sample of Native Americans (*n*=401) that we have obtained. We will estimate the proportion of ancestry for each individual in eMERGE-III cohort. Second, we will infer genetically defined Latinos based on the proportions of ancestry from African, European and Native American genomes, and in the setting of two-way admixture (European and Native American). We will assess the correlation of self-reported and genetically defined Hispanic or Latinos ethnicity. Finally, local ancestry will be inferred using a discriminative modeling approach. We will use local-ancestry inference for admixture mapping for lipid traits and diabetes in Latinos. | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  CPT codes  Phecodes  BMI | Common Variable Labs  Common Variable Meds  Geocoding 2015 ACS variables  Other: Case/Control status |
| **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)*  - Pre-imputed phased genotypes, and imputed merged GWAS genotype data from eMERGE 3 and desired additional variables.  - Sequencing data from the 1000 Genome Project (downloaded)  - Genotype data from a sample of Native Americans (requested and available) | |
| **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  No | |
| **Planned Statistical Analyses** | Principal component analysis, model-based population structure analysis, local-ancestry inference using a discriminative modeling approach, and local ancestry mapping | |
| **Ethical Considerations** | None noted | |
| **Target Journal** | PLoS Genetics | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | August 2017: Proposal submission  September 2022-January 2023: Statistical analysis  March 2023: First manuscript draft  June 2023: Manuscript submission | |

**\*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