**eMERGE Network Proposal for Analysis**

Project/Manuscript Concept Sheet

|  |  |
| --- | --- |
| **Reference Number** | NT252 |
| **Submission Date** | August 29, 2017 |
| **Project Title** | Ancestry and admixture in the eMERGE-III cohort of 83,717 individuals |
| **Tentative Lead Investigator (first author)** | Xiao Fan |
| **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 Involved** | 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 in Latinos. |
| **Desired**  **Variables (essential for analysis**  **indicated by \*)** | Additional Variables:  Age  Gender  Self-reported race and Ethnicity  Height and Weight  LDL-C levels  Lipid-lowering medications |
| **Desired data** | - 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) |
| **Planned Statistical Analyses** | Principal component analysis, model-based population structure analysis, local-ancestry inference using a discriminative modeling approach, and admixture mapping |
| **Ethical considerations** | None noted |
| **Target Journal** | PLoS Genetics |
| **Milestones\*\*** | August 2017: Proposal submission  September-October 2017: Statistical analysis  October 2017: First manuscript draft  Dec 2017: Manuscript submission |

**\*\*** This section should include: Timeline for completion of project, including approval, project duration, first and second draft of the paper and submission.