**eMERGE Network Proposal for Analysis**

Project/Manuscript Concept Sheet

|  |  |
| --- | --- |
| **Reference Number** | NT261 |
| **Submission Date** | 10/30/2017 |
| **Project Title** | Evaluating global and local ancestry and admixture across PheWAS phenotypes in eMERGE-III |
| **Core Team**  (alphabetic order) | Samira Asgari, Digna R Velez Edwards, Jacob Keaton, Rachel Knevel,  Soumya Raychaudhuri |
| **All other authors** (alphabetic order) | David Crosslin, Lea Davis, Joshua Denny, Todd L Edwards, Jacklyn Hellwege, Elizabeth Karlson, Ian Stanaway, Dan Roden,  plus  ***any additional eMERGE authors interested in participating*** |
| **Sites Involved** | A network-wide study (all sites invited to participate). |
| **Background / Significance** | The demographic history of many modern-day populations is shaped by multiple migration and admixture events in the past. For example, after settlement by Native Americans, American populations experienced admixture with individuals of European, African and Asian ancestry. Such admixture is traditionally regarded as a confounding variable in genetic association studies as it incorporates cultural, geographical, socio-economical and biological dimensions.[1,2] Therefore, the problem of admixture is typically avoided by studying genetically homogeneous populations.  This practice of creating homogeneous populations in studies is unfortunate, as genetic population differences are highly relevant from the clinical perspective. Current approaches to precision medicine focus on a patient’s clinical history and often include known genetic risk factors. As multiple diseases have shown strong prevalence differences across racial groups, race has been incorporated into clinical practice as a factor to consider in personalizing a treatment plan. This practice views race as a genetic risk identifier and as a proxy for environmental factors that influence disease risk (access to healthcare, air pollution, diet, exercise etc.). To optimize treatment and interventions, disentangling genetic from environmental risk is needed. In addition, race is genetically more complex than the separation of people into groups. Given that there has been admixture between the populations, people can carry risk factors of several different races.  Admixture mapping, which tests the local ancestry-phenotype correlation, enables us to obtain new insights into this network of risk factors. As shown before [7], when ancestral populations differ both in allele frequencies and disease prevalence, studying admixed populations can help us understand the underlying genetic factors that contribute to disease risk. In addition, as admixture created the presence of population specific genotypes in a variety of social structures, the confounding of population genetics by environmental factors can be overcome.  Within this project we leverage the rich phenotypic information available from large electronic health record (EHR) biobanks and the large multi-ethnic, admixed population gathered through the eMERGE network to evaluate the relationship of disease risk with genetic ancestry, to identify genetic risk factors that are independent of environmental factor and to identify clinical patient characteristics that interact with genetic ancestry.  Additionally, we will use 1000 genomes to not only map disease risk to specific genetic ancestries but use this information to better understand geographic origins of disease. We will accomplish this through the following analyses:  **1. Define global ancestry proportions** Ancestry will be determined for all subjects (regardless of reported race) with existing genome-wide association study (GWAS) using ADMIXTURE software, and proportions of ancestry specific to each reference population from 1000 Genomes and Simons diversity project data[1] will be determined. *We hypothesize that the eMERGE patient population is highly admixed and that self-reported race only modestly correlates with genetic ancestry.*  **2. Identify diseases associated with differences in global and/or local ancestry proportions** Using a data from a large EHR biobank, we will conduct global ancestry analyses using diseases with existing PheWAS codes. These diseases will then be evaluated for association with proportions of genetic ancestry for each reference population. For those phenotypes with evidence of an ancestry related association, we then test local admixture mapping adjusted for available disease specific candidate covariates, including information on socio-economic status. *We hypothesize that racial disparities in disease prevalence are driven by differences in specific genetic ancestries.*  **3. Identify loci that show evidence of global and/or local ancestry associating with disease risk interacting with patient clinical characteristics using eMERGE.** We will conduct an ancestry by clinical characteristic (AncxCC) interaction analysis evaluating interactions with body mass index (BMI), sex, and smoking status (if available), and limiting to diseases that demonstrate ancestral differences in Aim 1. We will evaluate multivariable regression analyses testing for AncxCC interactions, and assess interaction effects with likelihood ratio tests. This will identify diseases where global ancestry interacts with clinical characteristics to modify risk for disease. For those models with evidence of an interaction we then run admixture mapping AncxCC analyses using subjects with existing GWAS data to identify genomic regions where local ancestry interacts with clinical characteristics, using RFMIX. Finally, those genomic regions with strong evidence of AncxCC will be further evaluated with focused single SNP association analyses using available GWAS data. *We hypothesize that clinical characteristics of patients interact with both global and local genetic ancestry to modify risk for disease.*  The significance of this study is three-fold. First, this study will facilitate the eMERGE Network in its whole: The estimated ancestry proportions from our study can be used by all groups to select specific populations with greater accuracy or to perform a statistical correction for population stratification in large multiethnic studies. Second, the identification of genetic risk factors for diseases that are currently viewed as life-style diseases, will facilitate disease prevention and treatment. Third, the identification of interactions between clinical characteristics and ancestry will optimize personalized medicine. |
| **Outline of Project** | ANALYSIS PART I   1. Global ancestry inference   The eMERGE 3 data will be used to select variants imputed to high quality across all patient cohorts. These variants will be used for further analyses. Using 1000 genome[8] and Simons[1] diversity project data as reference populations, the global ancestry proportions for each patient will be calculated.   1. Ancestry phenotype association   Test for global ancestry associations with PheWAS phenotypes.   1. Local ancestry inference   For each subject, the ancestry of local genetic regions will be determined  ANALYSES PART II   1. Socioeconomic status   Socioeconomic status will be determined by linking zipcodes to US Census data .[10]   1. Phenotype selection   Using only those diseases with evidence of global ancestry associating with disease risk and where environmental factors are thought to play an important role.   1. Admixture mapping   Using the selected phenotypes logistic regression with phenotype status as dependent variable and local ancestry as independent variable to identify the regions that contribute to phenotype risk. Appropriate covariates, e.g. SES, age, gender, etc. will be included.  ANALYSES PART III   1. Phenotype selection   Using only those diseases with evidence of global ancestry associating with disease risk, test for interactions between global ancestry and BMI, sex, and smoking status on risk for association with PheWAS phenotypes   1. Covariates-Ancestry interaction   Test for interactions between BMI, sex, and smoking status and local ancestry on risk for individual PheWAS phenotypes that show from significant associations from analysis #B above   1. Local association analyses   Conduct local association analyses of regions where local ancestry interacts with a given clinical characteristic  PART IV   1. Manuscript preparation and submission. |
| **Desired**  **Variables (essential for analysis**  **indicated by \*)** | * All eMERGE imputed GWAS data from across all racial/ancestral groups * BMI, age/ year of birth, height, sex, smoking status (if available),and race/ethnicity\* * PheWAS codes * ICDcodes * Zipcodes (or other socio-economic proxies, if available) |
| **Planned Statistical Analyses** | ADMIXTURE [1]  Generalized linear models[9]  logistic regression  AncxCC  RFMIX  DNAnexus platform will be used to conduct several of the proposed analyses |
| **Ethical considerations** | There are no additional risks involved. The data will be stored in Dr. Velez Edwards servers at Vanderbilt. No data will be shared. We will also abide by the EMERGE guidelines.  This study makes use of data that is already obtained from electronic health records. The geocoding will be used to identify people with similar socio-economic circumstances. To ensure that genetic data cannot be traced back to individuals, only the conclusion of socio-economic similarity will be combined with the genetic data. |
| **Target Journal** | TBD |
| **Milestones\*\*** | Total Duration of the study: 1 year |

**References**

1. Mallick, S., et al. (2016). "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations." Nature **538**(7624): 201-206
2. Homburger, J. R., et al. (2015). "Genomic Insights into the Ancestry and Demographic History of South America." PLoS Genet **11**(12): e1005602.
3. https://www.cdc.gov/nchs/data/factsheets/factsheet\_disparities.htm
4. Chan JC, Malik V, Jia W, et al. Diabetes in Asia: epidemiology, risk factors, and pathophysiology. JAMA. 2009;301(20):2129-2140.
5. Christopher A.et al “Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer.” N Engl J Med 2006; 354:333-342January 26, 2006
6. Ayanian, J. Z., et al. (2014). "Racial and ethnic disparities among enrollees in Medicare Advantage plans." N Engl J Med 371(24): 2288-2297.
7. Genovese G, et al (2010). “Association of Trypanolytic Apol1 Variants with Kidney Disease in African Americans." Science. 2010 Aug 13;329(5993):841-5.
8. Clarke, L., et al. (2017). "The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data." Nucleic Acids Res **45**(D1): D854-D859.
9. Shriner, D. (2013). "Overview of admixture mapping." Curr Protoc Hum Genet **Chapter 1**: Unit 1 23.
10. https://www.census.gov