|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT301 | |
| **Submission Date** | July 23, 2018 | |
| **Project Title** | eMERGE Harvard Site Top 6 Genes | |
| **Tentative Lead Investigator** *(first author)* | S. Asgari | |
| **Tentative Senior Author**  *(last author)* | S. Raychaudhuri, S. Weiss, J. Smoller | |
| **All Other Authors** | E. Karlson, S. Chu. J. Su, P. Natarajan, CY Chen, A Zheutlin | |
| **Sites Participating** | We invite each site to join | |
| **Background / Significance** | The discovery and clinical use of genetic variants associated with both rare Mendelian and more common complex diseases promises to dramatically change the practice of medicine. We have developed highly specific machine learning algorithms to identify a) cardiovascular disease (CVD) (coronary heart disease, congestive heart failure, hypercholesterolemia), b) neuropsychiatric diseases (bipolar disease, depression), and c) immune-mediated diseases (rheumatoid arthritis (RA), inflammatory bowel disease (IBD), and multiple sclerosis (MS)) and are developing new algorithms for COPD, asthma/COPD overlap (ACO), and schizophrenia**.**  As a part of aim 2 in our eMERGE 3 proposal (see below) we proposed to include 6 genes for sequencing in the eMERGE-Seq dataset, 2 genes from each of three disease areas describe above a) *ANGPTL3, CORIN for* association with hypercholesterolemia, coronary heart disease, and stroke, b) *CACNA1C, TCF4* for association with ADHD, Bipolar disease, depression, and schizophrenia; and c) *TYK2, VDR* for association with immune phenotypes RA, IBD, and MS, and asthma, and ACO. We plan to study the burden, penetrance and pleiotropic effects of rare variants in these genes. We will use gene burden analysis and PheWAS in the eMERGE-Seq dataset (N=30,000) to assess the burden of rare variants in these genes in patients compare to unaffected individuals, to detect other EMR phenotypes associated with rare variants in these genes and to confirm penetrance by comparing to the chart of patients enrolled in the Partners HealthCare Biobank from the Harvard site (N=2500).  **Aim 2 (Penetrance and Pleiotropy): *Hypothesis:* Sequencing a set of established genes or loci will allow us to discover additional variation, and define penetrance and pleiotropy using EMR phenotypes**. | |
| **Outline of Project** | **A. PheWAS**.As part of Aim 2, we will conduct a hypothesis driven PheWAS using sequencing data from genes known to be associated with 13 phenotypes defined by eMERGE algorithms to study rare variant associations within each group of conditions, CVD, neuropsychiatric NPD, and immune diseases, as well as for individual phenotypes. We will use the methodologies described below including CMC, RC-SKAT, and case control haplotype testing1 with focused testing for RCH-specific effects. We will perform a combined, gene-based analysis of rare and common variants using the Combined Multivariate and Collapsing (CMC) method for case-control data2. We will assess rare variants based on the SKAT framework that was recently implemented (RC-SKAT)3. We also have experience with other rare variant methods. Deleterious coding variants can interact with common regulatory variants, either via haplotype-specific effects or via non-additive epistatic effects. Thus, we will also assess the impact of potential Regulatory-Coding Haplotypes (RCHs) by using phased regulatory and coding haplotypes available from genome-wide genotyping. Association testing will be performed using case control haplotype testing1 with focused testing for RCH-specific effects (a hypothesis-focused approach, rather than testing all allelic combinations). In addition to stratifying by ethnicity, we have successfully utilized principal components analysis4 and genomic control methodologies5 with sequence and genotype data to effectively address population stratification and ethnic diversity. We have developed new methods to address this problem particularly for rare variants6,7. We will also conduct an unbiased PheWAS for association of sequencing variants with EMR phenotypes*.* Diagnoses will be defined using ICD9 codes manually grouped into clinically relevant diseases as PheWAS codes by a team of physicians (e.g. ICD9 codes 411and 414, representing different types of CHD are grouped into an “ischemic heart disease” PheWAS code) and utilized in PheWAS studies by eMERGE groups8-14. We will include PheWAS codes with a prevalence of ≥1% in N=30,000 subjects. We will fit logistic regression models using each variant and presence or absence of the EMR-algorithm phenotype or PheWAS code as the outcome adjusting for age, sex and race. We will set the p-value cut-off for a significant association using the Bonferroni correction. For PheWAS codes significantly associated with a variant, we will review a random set of 20 charts at Partners HealthCare to determine the accuracy (PPV) of the code. This will allow us to assess penetrance.  **B. Assessing penetrance of rare variants.** For newly discovered rare variants from 6 genes we will define penetrance as the proportion of subjects with the variant shown to be associated with a disease, for whom they have evidence in the EMR for that disease according to 1) EMR-algorithm phenotype or, 2) PheWAS code verified by chart review.  **C. Assessing pleiotropy.** Allelic pleiotropy, where one genetic variant influences several distinct phenotypes, is has been show in psychiatric and immune-mediated diseases15-26. For example, our group recently demonstrated how different amino acid resides in the same *HLA-DRB1* position (position 13) confers risk of different immune conditions27. We propose to comprehensively investigate pleiotropy using PheWAS analysis to define associations with multiple PheWAS groups for each variant. In addition, we will define pleiotropy as association with multiple phenotypes within each group of CVD, NPD, and immune disease. We and others have demonstrated the value of this approach in successfully replicating results from CVAS and assessing pleiotropic effects 10,11,13,14,28.  **C. Assessing deleterious or beneficial variants.** We will rank genes/variants based on effect size, predicted functional impact and population-attributable risk (PAR) which is a combination of effect size and allele frequency. Since there could be many rare LoF variants within a single gene we can create a gene based effect size and PAR (or benefit) estimate of impact by summing all of the rare variant effects within a gene. Alternatively, we can use a variant- based assessment. | |
| **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  a) hypercholesterolemia, coronary heart disease, stroke  b) ADHD, Bipolar disease, depression, and schizophrenia  c) RA, asthma, and ACO (asthma/COPD overlap  No | |
| **Planned Statistical Analyses** | Logistic regression  Generalized linear mixed models  PheWAS  CMC  SKAT  RCH | |
| **Ethical Considerations** | This study makes use of data that is already obtained from electronic health records. The chart reviews will be performed in the Partners HealthCare Biobank subjects, a process that is IRB approved at Partners HealthCare. | |
| **Target Journal** | TBD | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | Aug-Oct | Nov-Jan | Jan-Mar | | Apr-Jun | | | A. PheWAS and burden testing |  |  | |  | |  | | | B. Chart reviews |  |  | |  | |  | | | C. Penetrance |  |  | |  | |  | | | D. Pleitropy | |  | |  | |  | | | H. Manuscript writing |  |  | |  | |  | | | |

**\*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