|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT362 | |
| **Submission Date** | 9/19/2019 | |
| **Project Title** | Exome-by-phenome-wide rare variant gene burden association with electronic health record phenotypes | |
| **Tentative Lead Investigator** *(first author)* | Joseph Park | |
| **Tentative Senior Author**  *(last author)* | Marylyn D Ritchie/Daniel J Rader | |
| **All Other Authors** | Any interested authors from eMERGE sites, and the Penn Medicine team (Nathan Katz, Xinyuan Zhang, Anastasia M Lucas, Yuki Bradford, Anurag Verma, Renae L Judy, Rachel L Kember, Scott M Damrauer, Jinbo Chen) | |
| **Sites Participating** | We are interested in including all eMERGE sites in the eMERGE I-III datasets. | |
| **Background / Significance** | By coupling large-scale exome sequencing with electronic health records (EHR), “genome-first” approaches can enhance our understanding of the contribution of rare genetic variants to disease. We’ve previously shown that aggregating rare loss-of-function variants in a candidate gene into a “gene burden” for association with EHR phenotypes is valuable for identifying both known and novel clinical implications for the gene in human disease. However, this methodology has not yet been applied on an exome-wide scale, and the clinical ontologies of rare loss-of-function variants in many genes have yet to be described. | |
| **Outline of Project** | We leveraged exome sequencing data in the Penn Medicine Biobank (PMBB; N=11,451) to address on an exome-wide scale the association of each gene’s rare loss-of-function variants with diverse EHR phenotypes using a phenome-wide association study (PheWAS) approach. We collapsed rare (minor allele frequency (MAF) ≤ 0.1%) predicted loss-of-function (pLOF) variants (*i.e.* frameshift insertions/deletions, gain/loss of stop codon, or splice site disruption) per gene to perform gene burden PheWAS. We’d like to be able to replicate significant gene-disease relationships identified in PMBB by testing low-frequency to common pLOFs and predicted deleterious missense variants on a univariate basis in the eMERGE dataset. | |
| **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   No | |
| **Planned Statistical Analyses** | We will conduct hypothesis-driven, targeted studies of association between low-frequency to common (MAF > 0.1%) pLOF variants and missense variants with REVEL≥ 0.5 in our significant genes with their corresponding associated Phecodes identified during our discovery phase in PMBB. We will use a logistic regression model adjusted for age (year of birth), sex, and principal components of ancestry. | |
| **Ethical Considerations** | There are no ethical concerns as we will be using de-identified datasets to interrogate gene-phenotype relationships. | |
| **Target Journal** | *Nature*, *Nature Genetics* | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | As soon as this application is approved, we plan to conduct the eMERGE replication studies within the following 1-2 weeks and submit our manuscript a couple weeks following that. | |

**\*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