|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT350 | |
| **Submission Date** | 06/12/2019 (modified 6/27) | |
| **Project Title** | Phenotypic signatures of genome-wide polygenic risk scores for complex traits. | |
| **Tentative Lead Investigator** *(first author)* | Atlas Khan <ak4046@cumc.columbia.edu> | |
| **Tentative Senior Author**  *(last author)* | Krzysztof Kiryluk <kk473@columbia.edu> | |
| **All Other Authors** | Chunhua Weng, Nick Tatonetti, Ali Gharavi, George Hripcsak, Cong Liu, Phyllis Thangaraj, David Fasel. Any other interested eMERGE investigator | |
| **Sites Participating** | Open to all sites  Current participants: Columbia University | |
| **Background / Significance** | Genome-wide polygenic risk (GPS) scores have been demonstrated to perform well in predicting selected complex phenotypes with high SNP-based heritability, such as coronary artery disease (CAD), atrial fibrillation, type 2 diabetes, inflammatory bowel disease and breast cancer. However, the specificity of these scores, and their pleiotropic associations are poorly understood. In addition, we are interested in comprehensively testing interactions between GPSes in predicting specific disease risks. For example, is the effect of GPS for CAD modified by type 2 diabetes GPS, breast cancer GPS, or IBD GPS? Lastly, it is not known if there are phenotypic differences among the cases with high polygenic risk versus those with low polygenic risk. The detection of such differences could enable clinical sub-stratification of patients into distinct subgroups without the need for genotyping (e.g. we hypothesize younger age of disease onset, more severe disease, and more comorbid conditions in cases in the higher tail of the GPS compare to those in the lower tail). | |
| **Outline of Project** | Several GPSes have been recently constructed, optimized, and validated based on summary statistics from large scale GWAS studies with external validations in UKBB and other biobank datasets. For the purpose of this study, which will serve as a proof-of-concept, we will examine mainly the existing GPSes that have already been optimized and validated for Europeans. This includes GPSes for the following 8 traits:   1. Coronary Artery Disease (Khera et al. Nat Gen 2018) 2. Atrial Fibrillation (Khera et al. Nat Gen 2018) 3. Inflammatory Bowel Disease (Khera et al. Nat Gen 2018) 4. Type 2 Diabetes (Khera et al. Nat Gen 2018) 5. Breast Cancer (Khera et al. Nat Gen 2018) 6. BMI (Khera et al. Cell 2019) 7. Chronic Kidney Disease (Wuttke et al. Nat Gen 2019) 8. Hypertension (Evangelou et al. Nat Gen 2018)   We will implement these risk scores in the eMERGE dataset. First, we will verify the predictive performance of the scores separately in each major ancestral group. Second, we will perform phenome-wide association studies for each GPS stratified by ancestry, and we will examine shared phenotypic correlations between different GPSes. Third, we will test for pairwise interactions between different GPSes in predicting their respective phenotypes. Fourth, we will perform case-only analyses to define specific clinical and demographic features that correlate with each score (e.g. sex ratio, age of onset, comorbidity burden, disease features that are unique to the cases in the tails of the risk score distribution). For each phenotype, we will aim to construct a phenotypic signature that defines a high-GPS versus low-GPS case. These phenotypic signatures will be captured in GPS-phenotypic scores. We will validate our findings in the UKBB dataset (already set up in our lab), and other available biobank-based datasets, such as FinGEN, HUNT, MGI, Mt.Sinai. | |
| **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)*** | NA | |
| **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** | 1. Implementation of validated GPSes in the eMERGE I-III merged set. 2. Testing the predictive performance of the GPSes and their pairwise interactions. 3. Phenome-wide association studies for each GPS to define shared phenotypic correlations between different GPSes. 4. Case-only analyses to define specific phenotypic signatures of GPSes and capture them in the GPS-phenotypic scores for each complex trait. 5. Validations of phenotypic correlations, pairwise interactions, and GPS-specific phenotypic scores in different ancestral groups 6. Validations of phenotypic correlations, pairwise interactions, and GPS-specific phenotypic scores in external biobanks. | |
| **Ethical Considerations** | None | |
| **Target Journal** | Depending on the results Am J Hum Genetics, PLoS Genetics, Human Molecular Genetics, Nature Communications or similar. | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Overall timeline of 1 year includes the following milestones:  Gather data from coordinating center: 6-7/2019  Conduct statistical analyses: 7-12/2019  Validations: 1-5/2020  Write manuscript: 5-6/2020  Circulate and submit manuscript: 6/2020 | |