|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT310 | |
| **Submission Date** | 10/30/2018 | |
| **Project Title** | Evaluate in silico prediction of pathogenicity of missense variants using sequence data and EHR phenotypes of the eMERGE III sequencing cohort | |
| **Tentative Lead Investigator** *(first author)* | Haicang Zhang and Xiao Fan | |
| **Tentative Senior Author**  *(last author)* | Wendy Chung and Yufeng Shen | |
| **All Other Authors** | Katherine Crew, Chunhua Weng, George Hripcsak, Ning Shang, Emily Groopman, Krzysztof Kiryluk, Lynn Petukhova | |
| **Sites Participating** | Any other interested eMERGE sites | |
| **Background / Significance** | Missense variants are the most common type of the protein coding variants. The ACMG guidelines use an ad hoc rule to determine pathogenicity or likely pathogenicity of rare missense variants based on concurrence of multiple prediction methods. The guideline is optimized for positive predictive value (PPV), while leaving many variants classified as variants of uncertain significance (VUS). Recent massively parallel cellular functional readout data suggest that more than 1/3 of the reported VUS in *BRCA1* in ClinVar could be pathogenic. We have preliminary results showing that new *in silico* prediction methods, developed by our team (*MVP*) and other groups, can improve prediction sensitivity while maintaining PPV, and have the potential to resolve VUS in a substantial fraction of patients. In this study, we propose to use all eMERGEseq data and phenotypes from the entire eMERGE III network to compare the performance of new methods in resolving VUS and assessing the clinical validity of new classification procedures based on these methods. | |
| **Outline of Project** | 1. Annotation of rare variants from the entire eMERGEseq data set using MVP and other new methods. We will pay particular attention to breast cancer related genes such as BRCA1/2, PTEN, TP53, PALB2, CHEK2, ATM. We will also assess variants in MC4R that are associated with obesity, HNF1B mutations associated with diabetes, and BMPR2 mutations associated with pulmonary hypertension. 2. EHR Phenotypes of the eMERGEseq participants. We will focus on breast cancer phenotypes, defined using ICD9 codes, and body mass index. We will first try to retrieve all the phenotype data in eRecordCounter and then contact individual sites for additional data if they are not in eRecordCounter. 3. Genetic analysis.    1. We will compare the performance of different methods based on the ability to distinguish patients and non-patients of the phenotype/diseases known to be associated with the genes.    2. We will predict likely pathogenic missense variants using a new procedure based on optimal threshold of MVP or combination of methods. and assess the clinical validity as a genetic screen in general or specific populations. For example, we will predict likely pathogenic variants that are otherwise classified as VUS by current practice in BRCA1/2, and then using ICD9 based on breast cancer/ovarian cancer status to calculate the relative risk in women in different age groups. | |
| **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)*  **hemoglobin A1C** | |
| **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 breast cancer, cancer  No | |
| **Planned Statistical Analyses** | 1. Data quality control 2. Calculating relative risk 3. Population ancestry inference from SNP array data 4. Machine learning methods to optimize prediction | |
| **Ethical Considerations** | Genomics data and phenotypic data will be de-identified to protect confidentiality. | |
| **Target Journal** | Nature Genetics, AJHG, Genetics in Medicine, Genome Medicine. | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | 1. Complete QC and annotation of eMERGEseq data by Dec 2018 2. Obtain and curate phenotype data of eMERGEseq subjects from the CC by Feb 2019 3. Statistical analysis by April 2019 4. Manuscript by June 2020 | |

**\*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