|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT384 | |
| **Submission Date** | 3/17/2020 | |
| **Project Title** | A GWAS study for overall cancer susceptibility across eMERGE network | |
| **Tentative Lead Investigator** *(first author)* | Bahram Namjou, CCHMC-postdocs… | |
| **Tentative Senior Author**  *(last author)* | John B Harley, | |
| **All Other Authors** | eMERGE network | |
| **Sites Participating** | Open to all sites | |
| **Background / Significance** | Previous GWAS studies have identified SNPs near telomere-associated genes as markers of overall cancer risk, regardless of tissue. In fact, the critical role of telomeres and telomerase in carcinogenesis has led to the hypothesis that short telomere length (TL) is a risk factor for cancer, however other factors such as aging or chronic inflammatory conditions might influence TL and the TL-cancer associations were not consistent across studies. In eMERGE network while we do not have information on TL from cells, we noticed a high rate of cancer diagnosis. Indeed, we found that (1 out of 2) individuals had a history of tumor diagnosis (from benign lipoma to malignancy) and (1 out of 4) had a diagnosis of malignant tumor. Given this large sample size, we plan to apply a simple strategy as shown below to perform a series of GWAS analyses. | |
| **Outline of Project** | 1. A two-step GWAS analyses:   a) Using ICD9, ICD10 and Phecodes, we will identify all cases with malignancy regardless of tissue of origin as “case” and those with no history of cancer (benign or malignant) as “control” and adjust the GWAS results for age, PCs, sites and other relevant covariates. We will exclude benign tumors as well as those with only carcinoma in situ or suspected cases.  b) to perform individual GWAS for a few well-powered cancers (e.g., breast, lung, colon) in these cohorts and explore pairwise genetic correlation between cancers using LDSC approach. Ancestry specific evaluations also will be done.   1. Functional enrichment analyses of highly significant variants as well as PheWAS approach for analyses of pleiotropy. 2. We will primarily use previous publications and posted UK biobank results for confirmation of our findings. | |
| **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)*** |  | |
| **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** | GWAS-PheWAS | |
| **Ethical Considerations** | None | |
| **Target Journal** | TBD | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Duration: 6 months  Completion: 9/1/2020 | |

**\*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