|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT381 | |
| **Submission Date** | 03/11/2020 | |
| **Project Title** | Supplemental analysis to ‘Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk’ | |
| **Tentative Lead Investigator** *(first author)* | Minta Thomas | |
| **Tentative Senior Author**  *(last author)* | Li Hsu | |
| **All Other Authors** | Elisabeth A. Rosenthal, David R. Crosslin, Ian Stanaway, Gail P. Jarvik, Lori C. Sakoda, Michael Hoffmeister, Jeffrey K. Lee, Franzel JB van Duijnhoven, Elizabeth A. Platz, Anna H. Wu, Christopher H. Dampier, Albert de la Chapelle, Alicja Wolk, Amit D. Joshi, Andrea Burnett-Hartman, Andrea Gsur, Annika Lindblom, Antoni Castells, Aung Ko Win, Bethany Van Guelpen, Catherine M. Tangen, Qianchuan He, Christopher I. Li, Clemens Schafmayer, Corinne E. Joshu, Cornelia M. Ulrich, D Timothy Bishop, Daniel D. Buchanan, David A. Drew, David C. Muller, David Duggan, Demetrius Albanes, Edward L. Giovannucci, Flora Qu, Graham G. Giles, Heather Hampel, Jane C. Figueiredo, Jeroen R. Huyghe, Jessica Minnier, Jenny Chang-Claude, Jochen Hampe, Kala Visvanathan, Keith R. Curtis, Kenneth Offit, Li Li, Loic LeMarchand, Ludmila Vodickova, Marc J. Gunter, Mark A. Jenkins, Martha L. Slattery,  Mathieu Lemire, Michael O. Woods, Mingyang Song, Neil Murphy, Noralane M.  Lindor, Paul D P Pharoah, Peter T. Campbell, Polly A. Newcomb, Roger L. Milne, Robert  J. MacInnis, Sergi Castellví-Bel, Shuji Ogino, Sonja I. Berndt, Stéphane Bézieau,  Stephen N. Thibodeau, Steven J. Gallinger, Syed H. Zaidi, Tabitha A. Harrison, Temitope O.  Keku, Thomas J. Hudson, Veronika Vymetalkova, Victor Moreno, Vicente Martín,  Volker Arndt, Yu-Ru Su, Richard B. Hayes, Emily White, Pavel Vodicka, Graham  Casey, Stephen B. Gruber, Robert E. Schoen, Andrew T. Chan, John D. Potter,  Hermann Brenner, Douglas A. Corley, Ulrike Peters | |
| **Sites Participating** | All eMERGE sites | |
| **Background / Significance** | **This work is done as a rapid supplemental analysis requested by a reviewer for a favorably reviewed methods paper by our collaborators Peters and Hsu. The reviewer’s requested additional data to support the authors’ claims regarding the PRSs.**  Three sets of polygenic risk scores (PRS) were developed for colorectal cancer (CRC), using three different methods, in subjects of European ancestry in non-eMERGE participants. Thse PRSs were compared for their discriminatory accuracy, using the area under the receiver operator curve (AUC). The best performing PRS utilized 1.2M SNPs and had a AUC=0.65, after adjusting for age and sex, which was significantly more accurate compared with all other PRSs. As multiple PRSs have been compared in non-eMERGE data, it is of great interest to validate that this PRS performs the best in an independent cohort. | |
| **Outline of Project** | As a supplemental analysis, we will calculate the AUC for the discriminatory accuracy of the PRS for CRC, adjusted for age, sex, the first ten principal components and site, using the European ancestry cohort from eMERGE. This supplement is for a favorably reviewed paper at American Journal of Human Genetics. We have a separate MCS for PRS development in the multiethnic emerge 4 cohort. Those separate analyses are underway.  The methods for imputation of genotype data and principal components analysis can be found in PMID: 30298529 (Stanaway et. al., The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. Genetic Epidemiology. 43(1):63-81).  The case control status for CRC was defined by an algorithm based on ICD9 codes 153 -153.9, 154 - 154.2,154.8 and ICD10 codes C18-C18.9, C19, C20, C21-C21.2, C21.8. An individual is considered a case if their medical record has any of these CRC codes appearing at least twice. That means an individual can have two different codes, each occurring one time, and that is sufficient. An individual is considered a control if they have none of these codes in their medical record. Only adults over age 18 were considered. | |
| **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 CRC  No | |
| **Planned Statistical Analyses** | Calculate the area under the curve (AUC) for each PRS prediction on CRC status, adjusting for censored age, sex, the first ten principal components of ancestry and site. | |
| **Ethical Considerations** | None | |
| **Target Journal** | American Journal of Human Genetics | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | February 2020 – completion of statistical analysis | |

**\*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