|  |
| --- |
| **eMERGE Network: Manuscript Concept Sheet** |
| **Reference Number** *(to be assigned by CC)* | NT316 |
| **Submission Date** | 12/4/18 |
| **Project Title** | Novel encoding method EDGE offers enhanced ability to identify genetic interactions |
| **Tentative Lead Investigator** *(first author)* | Molly Hall |
| **Tentative Senior Author** *(last author)* | Marylyn Ritchie |
| **All Other Authors**  | John Wallace, Anastasia M. Lucas, Yuki Bradford, Shefali S. Verma, Bertram Mueller-Myhsok, Beibei Jiang, Kristel Van Steen, Murray Brilliant, Catherine McCarty, Jason H. Moore, other interested site investigators |
| **Sites Participating** | Geisinger, Grouphealth, Kaiser Permanente/University of Washington, Marshfield, Mayo, Mount Sinai, Northwestern, Vanderbilt |
| **Background / Significance** | Robust computational methods to explore genetic interactions are essential for elucidating the complex nature of common human phenotypes. Yet, certain assumptions are made about the biological action of a variant when choosing a traditional genetic encoding. For each encoding type, risk incurred by one copy of the alternate allele in relation to two copies varies: the heterozygous genotype is coded to bear 0%, 50%, and 100% the risk of homozygous alternate for recessive, additive, and dominant encodings, respectively. However, the heterozygous action for a given SNP may incur any portion of the risk of the homozygous alternate genotype. Further, SNPs across the genome are unlikely to demonstrate the same genetic action. Choosing just one encoding, therefore, is not flexible to the diversity of action in biology; yet running every encoding raises the multiple testing burden. We present a novel, robust alternative encoding for genetic interaction testing: the elastic data-driven genetic encoding (EDGE). Here, a heterozygous value is assigned, based on the genetic action each SNP demonstrates in a dataset, thereby providing an individualized encoding for every SNP. We compared power of this method to detect main effect and genetic interactions using a comprehensive combination of simulated genetic models and found the novel method to outperform the traditional methods for the highest number of underlying models across varying minor allele frequencies (MAFs). We further tested our method with null and main effect-only simulated datasets and found that our method maintained a low false positive rate for identifying an interaction when none existed, while the additive and dominant encodings demonstrated inflation. |
| **Outline of Project** | Application of EDGE GxG to eMERGE phenotypes: T2D, resistant hypertension, age-related catartact, and glaucoma* QC of phenotypes
* GWAS (with covariate adjustment: age, sex, BMI, site, genotype platform, and phenotype-appropriate number of principal components) using traditional encoding methods (additive, dominant, recessive, codominant) and comparison of results to EDGE-derived heterozygous value.
* Main effect filtered GxG using EDGE and comparing results to traditional encodings for all phenotypes with same covariate adjustment as for GWAS
 |
| **Desired Data - Common Variables\*** *(Available from the CC)* | [x] Demographics [ ] ICD9/10 codes[ ] CPT codes[ ] Phecodes[x] BMI  | [ ] Common Variable Labs[ ] Common Variable Meds[x] 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** | [x] 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?** | [x] Yes, if so please list [ ] NoT2D, resistant hypertension, AMD, cataract, glaucoma  |
| **Planned Statistical Analyses** | The mechanisms of the EDGE method are described in the following steps:1. Logistic or linear regression is run using a codominant (dummy) encoding (Equation 1A).
2. Using the betas from the heterozygous genotype (Het) and homozygous alternate genotype (HA) dummy encodings, a weighted value (α) for the heterozygous genotype is calculated, whereby the α corresponds to the relative risk the heterozygous genotype to homozygous alternate genotype when homozygous alternate risk is scaled to 1 (Equation 1B).
3. These EDGE encodings (homozygous referent = 0, heterozygous = α, homozygous alternate = 1) are used for SNP-SNP interaction analyses. A common approach for genetic interaction is performing a likelihood ration test (LRT) between a full and reduced model: Y = β0 + β1SNP1 + β2SNP2 + β3SNP1×SNP2 (full), Y = β0 + β1SNP1 + β2SNP2 (reduced). A significant LRT p-value indicates in this case that the beta for the interaction of the two SNPs (β3 of the full model) is significant above and beyond the added main effects of the two SNPs (β1 and β2). As such, any overfitting that could be at play by employing this methodology is adjusted for in the main effects of the models. The lack of inflation is validated in the conserved false positive rates demonstrated by EDGE in our rigorous simulation studies that will be described extensively in the results section of the manuscript.

**Equation 1.** **Equations used to assign a data-driven weighted value for the heterozygous genotype.** 1. Y ~ βHetSNPHet + βHASNPHA B. α = βHet / βHA

To determine the significance of a SNP-SNP interaction model above and beyond the main effects of both SNPs combined, we perform a likelihood ratio test (LRT) between the full (Y = β0 + β1SNP1 + β2SNP2 + β3SNP1×SNP2) and reduced (Y = β0 + β1SNP1 + β2SNP2) models and derived a LRT p-value. All analysis was performed using PLATO software (M.A. Hall *et al.*, 2017), which has employs EDGE, additive, dominant, recessive, and dominant encodings with user specification. |
| **Ethical Considerations** | None |
| **Target Journal** | *Nature Communications* |
| **Milestones***(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | The data is already available to our analytic team. Analysis has been streamlined as a previous version of the manuscript has been underway for several years (NT140) and the manuscript is already underway, so the timeline will be short.* Analysis completion: December 20, 2019
* Draft completion: January 2, 2019
* Submission for publication: January 15, 2019
 |

**\*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