|  |
| --- |
| **eMERGE Network: Manuscript Concept Sheet** |
| **Reference Number** *(to be assigned by CC)* | NT335 |
| **Submission Date** | March 7, 2019 |
| **Project Title** | Evaluating the ‘Star Allele’ PGx nomenclature standard in the context of automated interpretation of panel, exome, and genome sequencing results |
| **Tentative Lead Investigator** *(first author)* | Adam Gordon |
| **Tentative Senior Author** *(last author)* | Bob Freimuth |
| **All Other Authors**  | Any interested eMERGE members |
| **Sites Participating** | All eMERGE sites |
| **Background / Significance** |  The process of translating an individual’s genotype data, NGS-derived or otherwise, into a specific combination of star alleles for interpretation relies heavily on a curated set of allele definition tables. Despite their widespread use, star alleles do not inherently convey how each named allele differs from the human reference: a ‘\*4’ allele could mean a single missense mutation, a haplotype of several variants, a whole gene duplication, or even no difference from the reference at all. This means that reporting a star allele may erroneously imply that genotype is known at sites that weren’t tested, particularly for star alleles defined by a single SNV.  Although careful curation and communication of allele definition tables could mitigate some of these issues, the influx of new variants and alleles requiring committee naming and curation will quickly become unmanageable as we continue to sequence thousands of individuals from diverse populations, many of which were not well represented in early research efforts from which these tables derive. Despite these inherent structural issues, the star allele system continues to pervade clinical pharmacogenetics, primarily via translation of individual genotypes in standardized genotyping or NGS reports into the historical star system. Though this translation step is becoming increasingly common, and increasingly automated, errors introduced by this process have never been quantified, described, and contrasted across sequencing platforms. Thus, understanding the sources of error and uncertainty introduced by attempts to harmonize contemporary genomic standards with the historical standard of the star system is critical as the field continues to move rapidly towards broad, algorithmic, CDS-linked implementation of pharmacogenetic results.  As a proof-of-concept, we previously developed the algorithm to assign star alleles from phased NGS data, and applied it to whole-exome sequencing data from NHLBI-ESP as a roadmap. These analyses uncovered a considerable amount of haplotype translation & match errors among 5 key pharmacogenes (*TPMT*, *CYP2C9*, *CYP2C19*, *CYP3A5*, and *SLCO1B1*), and, notably, that the distribution of these errors differs considerably across genes when stratifying by patient ancestry and type of error. From our preliminary experiences with this dataset, we believe eMERGE is particularly well-positioned to further explore these questions due to the wealth of sequencing data across multiple platforms (each with differing subsets of PGx loci) and the active efforts at multiple sites to return pharmacogenetic results to participants. |
| **Outline of Project** | 1. Obtain and encode allele definition tables (PharmVar hosted) for 5 critical PGx genes (*TPMT*, *CYP2C9*, *CYP2C19*, *CYP3A5*, and *SLCO1B1*)
2. Extract and phase genotypes for these 5 genes across 4 datasets representing common PGx use cases:
	1. panel sequencing, both full gene (PGRNseq, n=9010) and selected loci (eMERGEseq, n=25,000)
	2. exome sequencing (NHLBI-ESP, n=6503, already obtained from dbGaP)
	3. genome sequencing (eMERGE, n=900)
3. Algorithmically assign star allele diplotypes to each sample for all 5 genes by matching phased data with allele definitions
4. Analysis of assigned diplotypes: frequency and type of haplotypes, comparison across datasets, known vs unknown haplotypes, modelling phase errors
 |
| **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)*** | N/A |
| **Desired Genetic Data** | [ ] eMERGE I-III Merged set (HRC imputed, GWAS)[x] eMERGE PGx/PGRNseq data set [x] eMERGEseq data set (Phase III)[x] 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 [x] No |
| **Planned Statistical Analyses** | * Comparison of haplotype frequency spectra across genes & datasets
* Contrast measured diplotype frequencies with reported population frequencies
* Modeling phase errors (bootstrap phase of single/doubleton variants)
 |
| **Ethical Considerations** | n/a |
| **Target Journal** | Science Translational Med |
| **Milestones***(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Approval: March ‘19Data extraction and phasing: April ‘19Generate diplotypes: May ‘19Data analysis / manuscript draft: May/June ‘19Manuscript submission: July ‘19 |

**\*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