**eMERGE Network Proposal for Analysis**

Project/Manuscript Concept Sheet

This is project is part of the Phenotyping Workgroup’s list of Network phenotypes and the algorithm(s) is expected to be implementation-ready by 06/2016.

|  |  |
| --- | --- |
| **Submission Date** | 1/21/2016 |
| **Project Title** | **Comprehensive genetic association study of kidney traits across the EMERGE network** |
| **Tentative Lead Investigator (first author)** | Nikol Mladkova |
| **Tentative Senior Author (last author)** | Krzysztof Kiryluk, Ali Gharavi |
| **All other authors** | George Hripcsak, Chunhua Weng, Ning (Sunny) Shang, Jun Zhang, Maddalena Marasa, others and “The eMERGE Network” plus ***any additional eMERGE authors interested in participating*** |
| **Sites Involved** | We propose a network-wide study (all sites invited to participate). The analyses will be led by Columbia University. |
| **Background / Significance** | We are studying the molecular genetics of chronic kidney disease (CKD) and its complications. The goal of this study is to identify new loci responsible for the phenotype of CKD and its various quantitative sub-phenotypes using EHR and GWAS data available through the EMERGE network.  Most patients with CKD eventually develop end-stage renal disease (ESRD). Currently, ESRD affects 1:700 Americans and is associated with a 20% yearly mortality. ESRD patients depend on renal replacement therapy (dialysis or transplantation) for survival. Among adults, the major causes of CKD remain diabetes mellitus, hypertension, and glomerulonephritis (GN). Among children, the major causes of CKD are congenital renal and urologic malformations (such as renal agenesis) and GN. Transplant recipients constitute a special group of CKD patients because they have a primary underlying cause of kidney disease but are also at risk for recurrent CKD due to allograft loss. In renal transplant patients, the most common cause of graft loss is allograft rejection, followed by recurrence of primary disease. There are currently no efficacious cures for most conditions leading to CKD.  Depending on the ethnic group surveyed, 10-30% of those with CKD have an affected first or second degree relative, suggesting strong hereditary contribution to disease. This familial aggregation of renal failure is frequently observed for complex traits that are not normally considered to be hereditary, such diabetic, hypertensive, or IgA nephropathy. Using some of the methodologies delineated in this protocol, we have previously identified multiple susceptibility loci for IgA nephropathy, the most common form of glomerulonephritis and a major cause of CKD worldwide (Kiryluk et al. *Nat Gen* 2014, Kiryluk et al*. PLoS Gen* 2012, Gharavi at al. *Nat Gen* 2011).  Here, we propose to perform a comprehensive genetic study of a number of CKD-related quantitative endophenotypes derived from EHR across the entire EMERGE network. We primarily concentrate on lab-value derived phenotypes (eGFR, proteinuria, electrolytes), because:   1. Kidney-related lab values (basic chemistries) are obtained as part of routine medical care and are available for a great majority of EMERGE subjects enabling very large sample size across the entire network. 2. Lab values are typically stored as structured data in EHR, thus lab-based EHR phenotypes are simple and accurate. 3. Longitudinal lab value measurements contained in the EHR will allow for testing of time trends in renal function, as well as derivation of more specific time-averaged estimates; our proposed analysis of such traits using mixed models will provide optimal power of the GWAS approach; this approach is novel and has not been previously implemented for kidney-related phenotypes genome-wide.   In addition to lab-derived phenotypes described above, we will design simple NLP-based phenotypes to determine kidney size (through NLP of kidney ultrasound reports) and glomerular disease subtype (through NLP of kidney biopsy reports). Kidney size is correlated with total nephron number, which represents a major risk factor for subsequent CKD and hypertension. NLP of kidney biopsy reports will allow for subset analysis of very specific CKD subtypes. Although the number of patients with available kidney biopsy reports for this analysis will be much smaller, the strength of this approach lies in defining a specific glomerular diagnosis. Notably, common variants with large effects have previously been described for several GN subtypes, and many have not been studied by GWAS previously.  In summary, we feel that the EMERGE network is ideally suited to comprehensively address the complex genetics of CKD and its related sub-phenotypes using GWAS approach. CKD is also the top priority phenotype for the Columbia site EMERGE-III sequencing effort. The phenotype algorithms will be generated for this purpose regardless of the network-wide GWAS proposed here. |
| **Outline of Project** | The project will be conducted in several stages:   1. Building kidney phenotypes and sub-phenotypes (also included as top priority Columbia EMERGE III phenotypes) 2. Implementation of phenotype algorithms for individuals with available GWAS datasets network-wide 3. Phenotype quality control analyses 4. Genome-wide association analyses 5. Secondary analyses 6. Manuscript preparation and submission |
| **Desired**  **Variables (essential for analysis**  **indicated by \*)** | Implementation of standardized Columbia EMERGE phenotypes related to chronic kidney disease (CKD) across all sites with available GWAS datasets:   * CKD phenotype\* * eGFR sub-phenotype\* (quantitative, longitudinal) * Proteinuria sub-phenotype\* (quantitative, longitudinal) * Electrolyte (Na/K/Cl/HC03/BUN/Mg/Ca/Phos) sub-phenotypes\* * Kidney size sub-phenotype\* (quantitative, based on renal US) * Glomerular disease sub-types\* (based on kidney biopsy reports) * Age, sex, race/ethnicity, albumin, diabetes phenotype, type of ascertainment (if disease-driven)\* |
| **Desired data** | * Standardized CKD-related phenotypes as above (appropriate standardized algorithms will be constructed to extract relevant phenotype data) * All genotype data from EMERGE sites * Imputed genotype data for all sites. |
| **Planned Statistical Analyses** | Standard case-control GWAS for binary traits such as CKD and glomerular subtypes, quantitative GWAS for kidney size, and mixed models GWAS for longitudinal eGFR, proteinuria, and electrolyte sub-phenotypes. Secondary analyses will include conditional analyses, haplotype analyses and rare-variant association scans. Depending on the nature of the newly discovered associations, some genetic loci may be followed by targeted functional studies in model organisms. |
| **Ethical considerations** | There are no additional risks involved. The EMR and genomic data will be stored at a secured location in the data storage system of Kiryluk and Gharavi labs. No data will be be shared with unauthorized third parties. Patient identity will not be compromised by the proposed analysis. We will also abide by the EMERGE guidelines in this regard. |
| **Target Journal** | TBD, depending on the findings |
| **Milestones\*\*** | Total Duration of the study: 3 years  Implementation of phenotyping algorithms: 2016  Implementation of GWAS analyses: 2017  Draft of manuscript to authors: 2018  First submission: 2018 |

**\*\*** This section should include: Timeline for completion of project, including approval, project duration, first and second draft of the paper and submission.