**eMERGE Network Proposal for Analysis**

Project/Manuscript Concept Sheet

|  |  |
| --- | --- |
| **Reference Number** | NT278 |
| **Submission Date** | 3/30/2018 |
| **Project Title** | **Genetic resistance to common life-threatening infections and its evolutionary effects on susceptibility to complex immune traits.** |
| **Tentative Lead Investigator (first author)** | Atlas Khan, Ning Shang |
| **Tentative Senior Author (last author)** | Krzysztof Kiryluk |
| **All other authors** | Mitchell Elkind, Chunhua Weng, George Hripscak, Ali G. Gharavi  ***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** | Bacterial, viral, and parasitic infections have provided strong selective pressures throughout the history of human evolution. Although parasitic infections are nowadays rare in developed countries, serious bacterial and viral infections are common and continue to contribute to significant morbidity and mortality. With few exceptions, genetic factors that convey pathogen susceptibility or resistance have not yet been comprehensively studied across the full spectrum of life-threatening infections.  This study will explore the hypothesis that common life-threatening infections provide a positive selective pressure on genomic loci regulating immune responses to specific pathogens. We also hypothesize that some infection-associated loci are subject to balancing selection – while certain common alleles convey resistance to specific infectious pathogens, they may also increase the risk of complex immune-mediated disease of autoimmune and inflammatory spectrum.  In this study, we will consider only the most common community-acquired infection types that can be considered life-threatening or would have been life-threatening in the pre-antibiotic era. We will design a single OMOP-based multi-infection phenotype that will consider infection types (e.g. respiratory, urinary, skin), along with identification of a specific causal organism. The following will be considered:   1. Bacterial urinary tract infections 2. Bacterial respiratory infections 3. Bacterial blood infections, endocarditis, and sepsis 4. Bacterial meningitis 5. Influenza infection   Our multi-infection phenotype will be constructed fully in OMOP to allow for efficient data extraction and implementation across the entire eMERGE network. The phenotype will be based on a combination of diagnostic codes and microorganism culture and serology data that defines causal organisms.  First, we will plan to map individual genetic loci underlying resistance to specific infections using GWAS approach. Because we expect to see multiple associations within the HLA region, we will also perform detailed analysis of the HLA region, including tests of association with classical HLA alleles imputed data across the eMERGE-3 dataset.  Next, using PheWAS approach, we will test the hypothesis if the newly discovered “infection resistance” have an opposed effect on the risk of other human diseases. We will also perform global (genome-wide) genetic correlation analyses between infection, autoimmune, and inflammatory traits with the hypothesis of a negative correlation between specific infection types and autoimmune/inflammatory diseases. The Columbia autoimmunity phenotype, which is currently being implemented network-wide, will be used in combination with the multi-infection phenotype to accomplish this task.  For all of the newly discovered “infection resistance” loci, we will test for signatures of positive selection and balancing selection. We will also study the distribution of resistance alleles across worldwide populations (HGDP, PAGE, HAPMAP) and test for correlations with local pathogen exposures and infection prevalence. |
| **Outline of Project** | The project will be conducted in several stages:   1. Construction and validation of the Columbia multi-infection phenotype (written fully in OMOP to facilitate network-wide implementation) 2. Implementation and quality control of the phenotype algorithm across all individuals with available GWAS datasets network-wide. 3. Genome-wide association analyses for major infection types and their causal organisms. 4. Classical HLA allele association analyses, including conditional analyses to define the most likely causal HLA alleles (including 4-digit classical alleles and AA positions). 5. PheWAS to discover pleiotropic associations for significant loci 6. Global genetic correlation analyses between specific infectious and autoimmunity phenotypes (testing for excess negative correlation) 7. Testing for signatures of natural selection at major infection resistance loci, geospatial correlation analyses of worldwide distribution of resistance alleles vs. pathogen exposures. 8. Manuscript preparation and submission |
| **Desired**  **Variables (essential for analysis**  **indicated by \*)** | Implementation of standardized Columbia EMERGE phenotypes related to infection across all sites with available GWAS datasets:   * Columbia multi-infection phenotype * Columbia autoimmunity phenotype * Age, sex, race/ethnicity\* |
| **Desired data** | * Columbia infection and autoimmunity phenotypes * ICD9/ICD10 codes for PheWAS of significant loci * Imputed EMERGE-III GWAS data * Imputed classical HLA genotype data, including AA positions |
| **Planned Statistical Analyses** | Primary analysis will involve standard case-control GWAS, where cases are defined by the infection type and causal organism (e.g. Gram-negative urinary tract infection, Streptococcal pharyngitis, Staphylococcal cellulitis, etc.) and controls are defined as infection-free. All analyses will be performed using dosage method on imputed eMERGE-III dataset after stratification into 3 major continental ancestral groups (European, Asian, African). The analysis of each ancestral group will be adjusted for significant PCs of ancestry within each stratum, then summary statistics from all strata will be combined using fixed effects meta-analyses.  Secondary analyses will include conditional analyses to refine new GWAS loci and classical HLA association analyses to refine associations across the HLA region. Targeted PheWAS analysis will be performed for all genome-wide significant loci to assess for their pleiotropic effects. Genetic correlations will be studied using the stratified LD-score regression methods. Genomic signatures of natural selection will be examined using a variety of existing methods, including extended haplotype methods. Global frequencies of risk alleles will be studied in existing cohorts, such as HGDP, PAGE, HapMap, and 30 other international control cohorts available to the investigators. Geospatial correlations will be tested between allelic frequency gradients and global pathogen prevalence data collected by the GIDEON (Global Infectious Diseases & Epidemiology Online Network). |
| **Ethical considerations** | There are no additional risks involved. The EMR and genomic data will be stored at a secure location in the Columbia data storage system. No data will 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** | TBA, depending on the results |
| **Milestones\*\*** | Total Duration of the study: 2 years  Completion of study design/approvals: April 2018  Implementation of phenotyping algorithms: January 2019  Implementation of GWAS analyses: July 2019  Draft of manuscript to authors: September 2019  First submission: January 2020 |

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