Population Architecture using Genomics and Epidemiology (PAGE)  
Ver. 06/14/10

PAGE Manuscript Proposal Template
Submit proposals by email to the PAGE Coordinating Center at Rwilliams@biology.rutgers.edu

All sections must be completed; incomplete applications will be returned. 
Do not exceed 3 pages in length (not including references).

PAGE Ms. Number: 102 Submission Date: 8/05/2016 [Approval Date: ______]

Title of Proposed MS: Analysis of genetic variants associated with pancreatitis in US minorities

Abbreviated Title of Proposed MS: MEGA analysis of pancreatitis

I. INVESTIGATOR INFORMATION:

Name of Lead Author: V. Wendy Setiawan
Junior Investigator? N
Email Address: vsetiawa@usc.edu
Telephone Number: 323-442-7806

Name of Corresponding Author (if different): NA
Email Address:
Telephone Number:

Names, affiliations and email address of PAGE Investigators proposed as co-authors:

<table>
<thead>
<tr>
<th>N, N</th>
<th>Affiliation in PAGE</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dan Stram</td>
<td>MEC</td>
<td><a href="mailto:stram@usc.edu">stram@usc.edu</a></td>
</tr>
<tr>
<td>Charles Kooperberg</td>
<td>WHI</td>
<td><a href="mailto:clk@fredhutch.org">clk@fredhutch.org</a></td>
</tr>
<tr>
<td>Kari North</td>
<td>ARIC</td>
<td><a href="mailto:Kari_north@unc.edu">Kari_north@unc.edu</a></td>
</tr>
<tr>
<td>Ruth Loos</td>
<td>BioMe</td>
<td><a href="mailto:ruth.loos@mssm.edu">ruth.loos@mssm.edu</a></td>
</tr>
<tr>
<td>Chris Haiman</td>
<td>MEC</td>
<td><a href="mailto:haiman@usc.edu">haiman@usc.edu</a></td>
</tr>
</tbody>
</table>
### II. SCIENTIFIC RATIONALE

Pancreatitis is a significant medical and financial problem in the US. Over the last two decades, hospitalizations for pancreatitis have increased by 100%\(^1\). In 2012, acute pancreatitis (AP) was the third most common cause of gastrointestinal-related hospitalization in the US with 275,000 hospital admissions each year incurring $2.6 billion in hospitalization costs\(^2\). Recurrent AP can lead to chronic pancreatitis, a serious condition which severely impacts quality of life\(^3\) and can lead to serious complications including diabetes and pancreatic cancer\(^4\). Currently there is no available treatment for pancreatitis, and thus understanding the etiology of pancreatitis is paramount to prevent this disease. Racial disparity in pancreatitis incidence is striking. The risk of pancreatitis is 2- to 3-fold higher among African Americans than Whites\(^4\). Among people aged 35-65, African Americans are nearly 10 times more likely to develop pancreatitis compared to other racial/ethnic populations in the same age group. Despite the striking racial disparity, epidemiologic studies for pancreatitis in African Americans or in other US minorities are scarce. Whether genetic variants are responsible for the observed differences in pancreatitis susceptibility is unknown\(^5\). The only GWAS for pancreatitis conducted to date was performed in individuals of European ancestry and reported two loci associated with pancreatitis: *CLDN2-MORC4* and *PRSS1-PRSS2*\(^6\). The SNPs in these loci have only been replicated in Japanese\(^7\) and Indian\(^8\) populations. An investigation of genetic variation at known susceptibility regions and genome-wide in non-European ancestry populations is likely to reveal novel susceptibility loci for pancreatitis that may contribute to ethnic differences in risk and which may not be identified in populations of European ancestry.

In this project, we aim to perform a discovery and fine fine-mapping studies on pancreatitis among US minorities using the Multi-Ethnic Genotyping Array (MEGA). This array has improved variant coverage across multiple ethnicities, assay ing approximately 1.7 million SNP markers. MEGA was designed to facilitate fine-mapping and functional discovery analyses at known risk loci for metabolic, cardiovascular, renal, inflammatory, anthropometric and a variety of lifestyle traits. We will attempt to identify novel pancreatitis loci and conduct targeted fine-mapping of known pancreatitis loci in African Americans, Native Americans, Asian/Pacific Islanders and Hispanics from the PAGE II study. Given the very limited data available on pancreatitis genetic susceptibility in non-whites, the results from this study will be a significant contribution to the literature in regards to characterizing and describing genetic risk in minority populations.

### III. OBJECTIVES AND PLAN

a. **Study Questions/Hypotheses.**

1. To discover novel common and rare genetic variants influencing pancreatitis risk by testing all variants on MEGA as well as imputed variants.
2. To test and refine known pancreatitis loci previously identified in European Americans in non-European populations.

b. Study populations, study design for each

This analysis will include approximately 50,000 African Americans, Native Americans, Asian/Pacific Islanders and Hispanics from PAGE studies that have been genotyped on the MEGA array or are planning to impute variants from GWAS data. The anticipated studies include: ARIC, BioMe, MEC, and WHI. We will include individuals with genotyping, covariate and pancreatitis phenotype information. We will also include all available participants of European descent from the same studies as confirmation/replication for the newly identified findings.

c. Variant/SNPs (Specify)

All MEGA variants that pass quality control filters (to be performed at PAGE CC/UW) will potentially be included. Genotyping on the Metabochip, Exomechip, and various GWAS arrays are available for many studies participating in PAGE and will be included in this analysis as appropriate (e.g., replication, meta-analysis). For samples with GWAS data but no MEGA data, MEGA variants subset from 1000 Genomes Project-based imputed data will be used in studies that were not directly genotyped with MEGA. In addition, we will carry out analyses using all successfully \(r^2>0.3\) imputed variants after the imputation of the MEGA array data is complete.

d. Phenotype(s) (Specify)

We will perform case-control analyses, where pancreatitis cases define the case group and individuals without pancreatitis diagnosis define the control group. Both prevalent or incident cases will be included in the case group. We will perform stratified analyses by pancreatitis type (acute, recurrent, chronic) if numbers permit. Counts and method of case identification by participating studies are provided in table below.

<table>
<thead>
<tr>
<th>Pancreatitis</th>
<th>Total pancreatitis cases</th>
<th>Case Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute pancreatitis</td>
<td>1794</td>
<td>ICD-9 from Medicare claims</td>
</tr>
<tr>
<td>Recurrent acute pancreatitis</td>
<td>424</td>
<td>ICD-9</td>
</tr>
<tr>
<td>Chronic pancreatitis</td>
<td>213</td>
<td>Self report</td>
</tr>
<tr>
<td></td>
<td>571</td>
<td>ICD-9</td>
</tr>
</tbody>
</table>

e. Covariates (Specify)

1) Age (continuous)
2) Sex (binary)
3) BMI (continuous)
4) Ethanol (g/day)
5) Smoking status (never, past, current)
6) Diabetes status (yes, no)
7) Global ancestry (estimated via principal components)
8) Study center (indicators/dummy variables)

Global ancestry: Principal components analysis will be conducted centrally for all studies using unrelated individuals. The top 8 ancestry informative principal components (PCs) will be used in regression models to control for population substructure. We will include HapMap, 1000 Genomes, and other groups to improve interpretation.

For rare variants (MAF <1%), there is concern that methods for controlling for population stratification using common variants is not appropriate. It is uncertain how applicable this concern is to study of complex traits in modern-day outbred populations as opposed to composites of highly-stratified endogamous groups localized geographically. Keeping this concern in mind, we will keep up with this literature and be careful of over-interpreting the results for very rare variants. We may also consider modifying our use of PCs, (perhaps increasing the number of PCs used in rare variant versus common variant analyses) or modify linear mixed models (LMMs), or generalized estimating equations (GEEs), possibly by including rarer SNPs in the calculation of the genetic relatedness matrix.

f. Main statistical analysis methods

Aim 1 analysis: Discovery of novel common and rare genetic variants influencing pancreatitis risk

Common variants: For autosomal variants with MAF ≥1% coded for additive effects, we will conduct race/ethnicity stratified logistic regression analysis. Our logistic regression model will be: pancreatitis status ~ Genotype + Covariates + Global PCs. Our baseline model will adjust for age, sex, and PC 1-8. Additional covariates (e.g. study site, smoking status, BMI, alcohol intake, diabetes) will be explored. Novel loci are defined as those that reach genome-wide significance and remain significant after adjusting for known loci up or downstream of the potential novel loci. Significance will be defined as P < 0.05/N (the number of variants being tested). To test for multiple independent signals within a single locus, we will perform a series of sequential conditional analyses at loci that showed evidence of association at P < 0.05/N.

Rare variants: Rare variants will be defined within each racial/ethnic population, where a variant is considered rare if the MAF is reported as less than 1% in the corresponding PAGE racial/ethnic group. Single variant tests based on score statistics will be conducted for rare variants where the A2 allele is observed in a sufficient number (n>100) of participants (score statistic tests can be better powered than the Wald test).

We will also aggregate rare autosomal variants by gene or loci to conduct burden and/or non-burden association tests. Rare variants will be mapped to genes or loci based on the appropriate build using annotations from the UCSC Table Browser. As tissue specific enhancer definitions (and gene-enhancer links using gene expression) will become available, we will also expand aggregated analyses to those in relevant tissues. We will use the Sequence Kernel Association Test (SKAT), the optimal SKAT test (SKAT-O) or alternative methods to perform this rare-variant association testing on unrelated individuals. We will run the models separately for each race/ethnic group. Equal weights will be applied to all variants or variants will be weighted by MAF or functional scores (such as CADD). A Bonferroni correction for the number of genes tested will be applied.

Aim 2 analysis: Refine known pancreatitis loci previously identified in European Americans.

To refine previously identified signals in European Americans, we will use the African American, Hispanic, Native American and Asian populations from PAGE and leverage the fine-mapped pancreatitis loci on MEGA.
**Fine mapping**: Variants that demonstrate a (genome-wide) statistically significant association with pancreatitis will be utilized as an index variant for fine-mapping. It is expected that variants associated with pancreatitis in other racial/ethnic groups will be correlated with the index variant found in populations of European ancestry. Therefore, for each of the index variants identified above, we will identify all SNPs that are correlated ($r^2 > 0.2$) with the original index variant in that region, using the CEU (or whatever population is relevant for the original GWAS identifying the locus) population information from the 1000 Genome Project. Results for these regions will be graphically displayed using LocusZoom. To test for multiple independent signals within a single locus, we will perform a series of sequential conditional analyses at loci that showed evidence of association at $p < 0.05/\#$ of variants at the locus), where the most significant variant and a correlated variant of interest will be included in the same model. P-values and changes in effect estimates will be evaluated to investigate which variant shows a stronger signal and to comment on the more likely functional variant.

**Second signals**: We will search for independent second signals among variants that are in the same fine-mapping region but are not correlated with the index variant ($r^2 < 0.2$). Within each region, the statistical significance of potential second signals will be determined by a region-specific Bonferroni-correction ($0.05 / \#$ correlated variants in that region). Conditional analyses will again be performed to identify whether the top signals in the region appear to be independent. Conditional analyses will be performed adjusting for successive variants until no variants with $p$-values lower than the Bonferroni-corrected threshold remain.

g. Ancestry information used? No ___ Yes X ___ How is it used in the analyses? We will adjust for population substructure with global ancestry estimates (i.e. PCs).

h. Anticipated date of draft manuscript to P&P: 2017

i. What manuscript proposals listed on www.pagestudy.org/index.php/manuscripts/ are most related to the work proposed here? Approved PAGE ms. numbers:

   – If any: Have the lead authors of these proposals been contacted for comments and/or collaboration? Yes X No ___

**IV. SOURCE OF DATA TO BE USED** (Provide rationale for any data whose relevance to this manuscript is not obvious): Check all that apply:

Genotype data
We will use ~1.7 million variants on the MEGA array. Genotyping on the MetaboChip, ExomeChip, and various GWAS arrays are available for many studies participating in PAGE and will be included in this analysis as appropriate (e.g., replication, meta-analysis).

Aggregate/summary data to be generated by investigators of the study(ies) mentioned:

[ ] ISMMS; [ ] CALICO; [ ] MEC; [ X ] WHI; [ ] CC; [ ] Other: ________________

If CALiCo, specify [ X] ARIC; [ ] CARDIA; [ ] SHS-Fam; [ ] SHS-Cohort; [ ] SOL

I, (VWS), affirm that this proposal has been reviewed and approved by all listed investigators.
V. REFERENCES

VI. IF USING SOL DATA (Please provide the information below)
   a. Keywords:
   b. Using biomarker data? Yes __ No __
   c. Where will the SOL data analyses be performed?