ARIC Manuscript Proposal # 1644B

PC Reviewed: 5/14/13 **Status: A** **Priority: 2**
SC Reviewed: _____ **Status: _____** **Priority: ____**

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: _M08_ Submission Date : _____ Approval Date: _____

Title of Proposed Ms.: Multi-ancestral generalization and Metabochip-wide association study of C-reactive protein in PAGE

I. INVESTIGATOR INFORMATION:

Name of Lead Author: Jonathan Kocarnik | **Junior Investigator? Y**
Email Address: jkocarni@fhcrc.org
Telephone Number: 206-667-5257

Name of Corresponding Author (**if different**): Ulrike Peters
Email Address: upeters@fhcrc.org
Telephone Number: 206-667-2450

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

N, N	Affiliation in PAGE	Email
Reiner, Alex	WHI	areiner@fhcrc.org
Carty, Cara	WHI	ccarty@whi.org
Haessler, Jeff	WHI	jhaessle@whi.org
Ambite, Jose Luis	CC	ambite@isi.edu
Crawford, Dana	EAGLE-BioVU	Crawford@chgr.mc.vanderbilt.edu
Le Marchand, Loic	MEC	loic@cc.hawaii.edu
Cheng, Iona	MEC	Iona.cheng@cpic.org
Hindorff, Lucia	NHGRI	hindorffl@mail.nih.gov
Chen, Dongquan	CARDIA	dongquan@uab.edu
Graff, Misa	ARIC	migraff@email.unc.edu
Avery, Christy	ARIC	christya@email.unc.edu
Pereira, Rocio	SOL	rocio.pereira@ucdenver.edu
North, Kari	SOL	Kari_north@unc.edu

Partner studies in PAGE not collaborating in this ms. proposal:

Study	Contacted? Y/N	Declined? / Other?

Names, affiliations, email address of non-PAGE investigators proposed as co-authors:

II. SCIENTIFIC RATIONALE (Please be specific and concise)

Inflammation is an important health outcome related to many common complex diseases, several of which have differential disease burden by ethnicity. C-reactive protein (CRP) is a circulating biomarker indicative of systemic inflammation. Genome-wide association studies (GWAS) have successfully identified susceptibility loci important to inflammation and inflammation-related diseases. However, most of these variants were discovered primarily in populations of European ancestry, and so may not represent the causal variants in other populations. To identify whether genetic variants associated with inflammation generalize in other populations, we will evaluate the association between serum CRP levels and the large number of variants on the Metabochip array in the various race/ethnicity groups available in the PAGE study. In regions demonstrating an association signal, we will perform fine-mapping to identify race/ethnicity-specific signals and conditional analyses to identify additional independent associations.

III. OBJECTIVES AND PLAN (Please be specific and concise)

a. **Study Questions/Hypotheses**.
We aim to investigate genetic associations with serum C-reactive protein (CRP) levels in the multiple race/ethnicity groups within the PAGE consortium.

b. **Study populations, study design for each**

All PAGE study populations with Metabochip data and measured C-reactive protein levels.
Studies included: ARIC, EAGLE-BioVU, CARDIA, CHS, MEC, WHI, HCHS/SOL, Tai Chi.
Race/ethnicities included: African American, Hispanic/Latino, Asian/Pacific Islander/Japanese/Hawaiian, American Indian.
Additional studies and race/ethnicity populations may be added as they become available.

c. **Variant/SNPs (Specify)**

All variants on the Metabochip that pass quality control filters (to be performed at PAGE CC). No locus was targeted specifically for CRP on the chip, however several regions targeted for other phenotypes contain SNPs previously associated with CRP which will be useful for fine-mapping.

d. **Phenotype(s) (Specify)**

Serum C-reactive protein (CRP). This should be reported in mg/L, and should be natural-log-transformed due to its skewed distribution.

e. **Covariates (Specify)**

Age, sex, and top principal components of genetic ancestry, as well as any needed study-specific variables (such as study center or region). Additional adjustment for local ancestry estimates will be performed if necessary and available.

f. **Main statistical analysis methods**

Association testing: The association between lnCRP and each variant on the Metabochip will be examined in each study using linear regression stratified by race/ethnicity. Each variant will be coded assuming an additive model (with 0, 1, or 2 copies of the coded allele). The coded allele will be defined by the primary analyst and sent to all analysts to prevent strand flipping issues for analyses conducted by studies where centralized data-sharing cannot occur (i.e. TAI CHI). Models will be minimally adjusted for age, sex, top 3 principal components of genetic ancestry, and any necessary study-specific variables (such as study center or region). Analyses in SOL will account for sampling weights and participant relatedness (see Appendix 1).

Meta-analysis: The overall association with CRP for each variant will be obtained by combining the beta estimates and standard errors from the study-specific analyses using inverse-variance weighted fixed-effect meta-analyses. Cochran's Q statistic and I^2 will be calculated to measure between-study heterogeneity. The statistical significance of an association will be determined using a Bonferroni-corrected p-value of 2.5e-7 (0.05 / 200,000). Initial meta-analyses will be specific to each race/ethnicity group (i.e., meta-analyzing the results from each study for the association of a given variant with CRP in African Americans). Depending on the heterogeneity observed, meta-analyses may also be conducted to examine the overall across-race/ethnicity association with CRP for a given variant. Any cross-race/ethnicity analyses may utilize additional methodologies currently under development, such as MANTRA.

Fine-mapping: Variants that demonstrate a statistically significant association with CRP and are located within one of the targeted fine-mapping regions of the metabochip will be utilized as an index variant for fine-mapping. It is expected that variants associated with CRP in other race/ethnicity 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 population information from the 1000 Genomes Project. Results for these regions will be graphically displayed using LocusZoom. To evaluate multiple signals, we will perform conditional analyses which include the most significant variant and a variant of interest (i.e. correlated SNPs in the fine-mapping region or the initial GWAS variant) 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 ($r^2 < 0.02$). 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.

Functional annotation: We will utilize bioinformatic tools to evaluate the potential function of any variants highlighted by our metabochip-wide or fine-mapping analyses.

g. **Ancestry information used? No __ Yes_X__ How is it used in the analyses?**
We will adjust for principle components of genetic ancestry in our analyses, and perform association analyses stratified by race/ethnicity.

h. **Anticipated date of draft manuscript to P&P: _September 2013_____**

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 __ 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:**

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

[X] EAGLE; **[X]** CALiCO; **[X]** MEC; **[X]** WHI; **[ ]** CC; **[X]** Other:_Tai Chi___
If CALiCo, specify **[X]** ARIC; **[X]** CARDIA; **[ ]** CHS; **[ ]** SHS-Fam; **[ ]** SHS-Cohort; **[X]** SOL

I, _(initials)_, affirm that this proposal has been reviewed and approved by all listed investigators.

V. REFERENCES

1. …….

Appendix 1 – Analysis of HCHS/SOL data

The Hispanic population of the United States is bi- and tri-admixed with Spanish, Native American, and African ancestry. Genetic mixing began with first contact in the early 1500's. As shown previously, there exists considerable heterogeneity of genetic structure among the major Hispanic populations (Hanis et al., 1991). This, of necessity, requires specialized analytic methods beyond the more general structure analysis. For this reason, we will implement a two stage strategy encompassing overall admixture structure using a principal components approach as implemented in programs such as Eigenstrat and localized structure surrounding all markers to be analyzed. Such analyses not only appropriately incorporate ancestral history, but they actually enhance the ability to detect genetic effects by enabling pooling of those regions of similar ancestral structure and exploiting the admixture disequilibrium that contributes to associations (Zhu et al., 2005; Basu et al., 2008). We note that these methods are appropriate for candidate gene studies and GWAS, and scale to next generation sequence data. We also note that there are additional opportunities afforded by the household based sampling employed in the SOL where a combination of identity by descent and admixture disequilibrium mapping will be implemented.

Estimation of Relatedness
The SOL study sampling design allows for related individuals. Preliminary SOL study data show that the range of cohort members drawn from a household is 1-14, with a median of 1, a mean of 1.7, and an upper quartile value of 2 per household. Twins older than 45 years would have been enrolled; those younger would have been subject to age-specific subsampling. In addition, we expect substantial relatedness of individuals across households in a Hispanic community. Therefore, we will estimate an empirical kinship matrix based on allele sharing at a set of independent autosomal SNPs, to be used to account for relatedness among individuals in subsequent association analyses.

We will estimate the average identity-by-descent (IBD) coefficients, Z0, Z1, and Z2, for the probability of sharing 0, 1 and 2 alleles, respectively (Hartl and Clark, 1997) for all the participants in the HCHS/SOL dataset. The investigators in this proposal have experience using genotyped SNP data to account for relatedness among individuals without specifying exact pedigree relationships and there are now multiple statistical programs available to accomplish this task. However, depending on the observed relatedness in these data, it is possible that we may not be able to exactly reconstruct all the pedigrees, particularly with respect to more distantly related extended family members. Whereas the identification of parent – offspring pairs and full sibling pairs is fairly straightforward, second-degree relationships usually require the incorporation of additional data, such as participant age to discern relationships. Third-degree relationships are even more difficult to specify and may not be discernible. Yet if unresolvable relationships are identified, there are multiple options for moving the analysis forward while still accounting for the correlations among study participants. The most computationally efficient approach would be to specify the strongest correlations in these data and to not consider the distant unresolvable relationships. Then statistical methods for analyzing a large number of relatively small families can be applied. On the other hand, we could directly incorporate the empirical kinship matrix into our analysis pipeline. Such an approach would necessitate analyzing each ethnic group as, effectively, a single pedigree with their relationships defined by the empirical kinship matrix. This would necessitate a variance-component (VC) approach as the analyses would require likelihood maximization. The VC association models (Amos, 1994; Amos et al., 1996; Almasy and Blangero, 1998; Abecasis et al., 2000; Diao and Lin, 2005) are widely used in the genetic analysis of quantitative traits in family studies. This approach has been extended to binary, ordinal and age-of-onset traits (Burton et al., 1999; Diao and Lin, 2006; Diao and Lin, 2010). An alternative approach is Generalized Estimating Equations (GEE) (Liang and Zeger, 1986), which is applicable to binary, ordinal and quantitative traits and has been extended to age-of-onset traits (Lee, Wei and Amato, 1992). The VC approach can directly incorporate the (empirical) kinship coefficients and tends to be more powerful than the GEE approach. However, the GEE approach is more robust (i.e., less model-dependent) and computationally simpler.

The genotype for each SNP will be coded as the number of copies of the minor allele that each individual carries (additive genetic model). This variable will be entered into a VC or GEE model and a p-value representing the significance of the association between the genotype variable and the trait will be obtained. The existing VC and GEE methods assume that the families are sampled randomly from the underlying population.

Accounting for sampling schema in HCHS/SOL
The HCHS/SOL cohort was selected through a stratified two-stage area probability sampling design (LaVange et al., 2010). Thus, the subjects were not selected with equal probabilities. The use of sampling weights that reflect unequal probabilities of selection must be incorporated into all analyses to obtain appropriate estimates of population characteristics and the corresponding standard errors. We will modify the VC and GEE methods by multiplying each subject's contribution

to the estimating function by his/her sampling weight and estimating the standard error by the sandwich variance estimator (Liang and Zeger, 1986; Binder, 1993; Lin, 2000). Failure to properly account for the sampling weights will lead to both biased hypothesis testing and parameter estimation.

Template Tables

Table 1 – Distribution of study characteristics, overall and by study.

Characteristic		OVERALL			HCHS/SOL			WHI *(etc.)*		
		Male	Female	Overall	Male	Female	Overall	Male	Female	Overall
Number of participants										
Age	Mean(SD)									
	Range(min-max)									
CRP	Mean(SD)									
	Median(IQR)									
	Range(min-max)									
BMI	Mean(SD)									
	Range(min-max)									
Smoking	% Current									
	% Former									

Table 2 – Distribution of allele frequency of SNPs in the fine-mapping regions. *Note: table should be repeated for each fine-mapping region identified by the Metabochip-wide CRP association analyses.*

(*list index SNP*)	n	%
Allele frequency[a]		
0.14% - 1%		
>1% - 5%		
>5% - 10%		
>10% - 25%		
>25%		
Total		

Exonic SNPs[b]
Total SNPs in exons
Synonymous SNPs
Missense SNPs

[a] Based on allele frequency for all AA participants
[b] Based on data from the Exome Variation Server (http://snp.gs.washington.edu/EVS)

Table 3 – Association between *(insert fine-mapping region here)* and C-reactive protein for all studies combined, stratified by race/ethnicity. *Note: this table should be repeated for each fine-mapping region identified in the Metabochip-wide CRP association analyses.*

(List fine-mapping region) – (list index SNP) – (list race/ethnicity group)									
		Alleles			% change in CRP per coded allele				
rs#	Position[a]	Coded allele	Non-coded allele	CAF[b]	Beta	SE	Nominal P-value	Adjusted P-value	P-hetero-geneity
Most significant variants in fine-mapping region									
List results									
Variants highlighted in previous studies									
List results									

[a] SNPposition based on build 37

[b] CAF = coded allele frequency (risk estimates provide the log additive effect per copy of the coding allele)

Supplemental Table 1 – Association between *(insert fine-mapping region here)* and C-reactive protein, stratified by study and race/ethnicity. *Note: this table should be repeated for each fine-mapping region identified in the Metabochip-wide CRP association analyses.*

(List fine-mapping region) – (list index SNP) – (list race/ethnicity group)									
(index SNP)		Alleles			% change in CRP per coded allele				
rs#	Position[a]	Coded allele	Non-coded	CAF[b]	Beta	SE	Nominal P-value	Adjusted P-value	P-hetero-

allele	geneity
HCHS/SOL	
Most significant variants in fine-mapping region	
List results	
Variants highlighted in previous studies	
List results	
WHI *(etc.)*	
Most significant variants in fine-mapping region	
List results	
Variants highlighted in previous studies	
List results	

[a] SNPposition based on build 37

[b] CAF = coded allele frequency (risk estimates provide the log additive effect per copy of the coding allele)

Supplemental Table 2 – Risk estimates for SNPs correlated with the most significant SNP *(list index SNP)* and conditional analyses with *(index SNP)* across studies, stratified by race/ethnicity. *Note: table should be repeated for each fine-mapping region identified in the Metabochip-wide CRP analyses.*

(List fine-mapping region) – (list index SNP) – (list race/ethnicity group)				
SNP	Alleles	% change in CRP per coded allele[c]	Correlation with *(index SNP)*	Results adjusted for *(index SNP)*[d]

rs#	Position[a]	Coded allele	Non-coded allele	CAF[b]	Beta	SE	P-value	r^2 in EA	r^2 in *(race/ethnicity)*	P-value for listed SNP	P-value for *(index SNP)*
SNPs correlated at r2>0.2 with *(index SNP)*											
List results											
Variants highlighted in previous studies											
List results											

[a] SNPposition based on build 37

[b] CAF = coded allele frequency (risk estimates provide the log additive effect per copy of the coding allele)

[c] Association analysis that only includes the listed SNP

[d] Association analysis with *(index SNP)* and the listed SNP in the same model