Title of Proposed Ms.:  *Pleiotropic Effects of Cancer Risk Variants on Non-Hodgkin Lymphoma Risk*

I. INVESTIGATOR INFORMATION:

Name of Lead Author:  Unhee Lim  
Junior Investigator?  Y
Email Address:  ULim@crch.hawaii.edu
Telephone Number:

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

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation in PAGE</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christian Caberto</td>
<td>Univ of Hawaii</td>
<td><a href="mailto:ccaberto@crch.hawaii.edu">ccaberto@crch.hawaii.edu</a></td>
</tr>
<tr>
<td>Riki Peters</td>
<td>WHI, FHCRC</td>
<td><a href="mailto:upeters@fhcrc.org">upeters@fhcrc.org</a></td>
</tr>
<tr>
<td>Kylee Spencer</td>
<td>Vanderbilt</td>
<td><a href="mailto:kylee.spencer@Vanderbilt.Edu">kylee.spencer@Vanderbilt.Edu</a></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Study</th>
<th>Contacted? Y/N</th>
<th>Declined? / Other?</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIC</td>
<td>Y</td>
<td>Number of cases is small – Will participate if desired</td>
</tr>
<tr>
<td>CHS</td>
<td>Y</td>
<td>Number of cases unknown – Will participate if desired</td>
</tr>
</tbody>
</table>

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

SCIENTIFIC RATIONALE
Cancer is the second leading cause of mortality in this country and, unlike for heart disease, little progress has been made to decrease the cancer burden of the US population. In recent years, genome-wide association studies (GWAS) have identified common risk variants for cancer. These studies have been cancer site-specific and almost exclusively conducted in populations of European descent. Among the many risk loci identified, several have been associated with multiple cancer sites. For example, the 8q24 region contains several loci associated with prostate, colon, breast, bladder and/or ovarian cancers. This region, which was
not previously suspected to play a role in cancer, is now thought to be an important regulatory region for the oncogene c-MYC and possibly for other genes. Similarly a locus on Chromosome 5 (5p15.33) that includes the telomerase gene (TERT) has been associated with lung adenocarcinoma, bladder and pancreatic cancers, as well as some hematological malignancies. Despite these striking examples, pleiotropic effects have not been systematically explored in cancer. We propose to do so in the cohorts assembled in the Population Architecture Using Genomics and Epidemiology (PAGE) study. GWAS hits reported for any cancer site (as of April 2010) have been genotyped and will be examined for association with each cancer site included in PAGE. These sites were selected based on reaching 1,000 incident cases across PAGE studies.

OBJECTIVES AND PLAN

In general, all of the papers in this proposal will explore the following general research questions

Study Questions/Hypotheses

i. Do the known Cancer GWAS risk variants identified in European-descent populations replicate in PAGE?

ii. Do the known GWAS risk variants identified separately for each cancer generalize to other cancer sites?

iii. Do the associations identified differ across ethnic/racial groups?

iv. Do these associations differ by host and disease characteristics?

v. Are these associations modified by known risk factors or other risk variants?

Study populations, study design

These analyses will be done in case-controls studies nested in each PAGE study with information on pathologically confirmed or registry-based cancer incidence data (MEC, WHI, ARIC, CHS) and combined in a meta-analysis.

Table 1: Number of Cancer Cases Genotyped in Wave 2 by Study

<table>
<thead>
<tr>
<th></th>
<th>WHI</th>
<th>MEC</th>
<th>ARIC</th>
<th>CHS</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>1,961*</td>
<td>3,231</td>
<td>384</td>
<td>182</td>
<td>5,768</td>
</tr>
<tr>
<td>Prostate</td>
<td>-</td>
<td>4,576</td>
<td>400</td>
<td>234</td>
<td>5,210</td>
</tr>
<tr>
<td>CRC</td>
<td>1,436</td>
<td>2,297</td>
<td>200</td>
<td>177</td>
<td>4,110</td>
</tr>
<tr>
<td>Lung</td>
<td>1,751</td>
<td>689</td>
<td>282</td>
<td>205</td>
<td>2,927</td>
</tr>
<tr>
<td>Endometrial</td>
<td>1,103</td>
<td>419</td>
<td>60</td>
<td>24</td>
<td>1,606</td>
</tr>
<tr>
<td>Ovarian</td>
<td>843</td>
<td>148</td>
<td>28</td>
<td>27</td>
<td>1,046</td>
</tr>
<tr>
<td>Melanoma</td>
<td>1,102</td>
<td>294</td>
<td>55</td>
<td>45</td>
<td>1,496</td>
</tr>
</tbody>
</table>
* Plus ~2,500 with GWAS hits for breast cancer

**Variant/SNPs:** See table in Appendix

**Phenotypes:** Outcomes: Incident invasive cancer diagnoses confirmed pathologically or through linkage to a cancer registry. Risk Factors (considered as potential confounders or modifiers): self-reported on baseline questionnaire.

**Covariates:** age, race/ethnicity, sex, education, weight, height, BMI, menopausal status, physical activity, cancer family history, screening history, smoking (smoking status, pack-years), dietary factors (e.g., calories from fat, red meat, alcohol, energy-adjusted nutrients from diet and diet+supplements), HRT use, reproductive history (age at menarche, age at menopause, parity, age at first birth, oral contraceptive use), medication (aspirin and other NSAIDs), sun exposure, eye/hair color, skin freckling, mole count, skin reaction to sunlight and ability to tan.

For cases: histology, anatomical subtype, age at diagnosis, stage at diagnosis, molecular subtype (i.e., hormone receptor status)

**Main statistical analysis methods**
Tests of main effects of each SNP within each ethnic population within each study will be carried out within a logistic regression model, controlling for covariates, such as age and ancestry (discussed below). For each variant, we will examine a wide range of genetic models (dominant, co-dominant, recessive). Meta-analyses will be done using random effects models and individual study summary estimates. Summary estimates and 95% confidence intervals will be obtained using the method described by DerSimonian and Laird. Fixed effect models will also be run in order to test heterogeneity across studies within sex-ethnic group. These tests, along with Forest plots, will be considered in understanding the level of heterogeneity. In the meta-analysis, we will test the heterogeneity of each SNP across ethnic groups by testing cross-product terms between genotype and ethnicity. While for true causal variants we would expect the strength of the associations observed to be relatively consistent across ethnic group, differences in LD (if it is not the causal variant) and modifying genetic or non-genetic risk factors may produce effect heterogeneity. A logistic regression approach will also be used to test for heterogeneity of effect across sub-groups defined by potentially important modifying factors, such as the main risk factors for the cancer site studied. A case-only approach will be used to improve power if the assumptions of independence of the SNP and exposure are met. Because this assumption may not always be met, we will explore with the Statistical Analysis Committee the use of newly described methods, such as the Empirical Bayesian proposed by Mukerjee and Chatterjee. Differences by disease characteristics (stage, histology, anatomic subsites) will be tested in a case-only analysis or polytomous logistic model. Type I errors will be controlled through adjustment of the critical alpha-level as well as through permutation testing. Analyses using a summary risk score (i.e., counting the number of risk alleles) will also be conducted. The predictive ability of a risk score based on the known risk alleles for a given cancer site will be compared to that of a risk score based on all markers described for all cancer sites.

We will work with the Statistical Analysis Committee to determine the best method for analysis of these data.

**Ancestry and how it is used in the analyses.** We will adjust for self-reported ethnicity
including parental ethnicity in this analysis. To account for residual confounding by race, we will also further adjust for main continental origin using principal components variables based on the genotyped AIMs.

Implementation of the Cancer Analyses in PAGE
Each PAGE site contributing to the Cancer analyses in PAGE will send genotyping and covariate data, using a common format, to the Coordinating Center which will assemble and curate a single analysis data file. Each paper leader will download the relevant data and conduct the analysis at his/her site. The CC will keep a record of which investigators received data and what data were received.

Following approval of the proposal by the PAGE P&P committee, a protocol for implementation of the analyses by each analysis group will be reviewed by all collaborators (in MEC, WHI, ARIC, CHS, EAGLE). This analysis plan will identify the following steps:

Centralized analyses
- Inclusion criteria in the creation of the data sets;
- Common and study-specific (if any) exclusions from the analyses;
- Distribution of a shell table for each analysis to document inclusions, exclusions and final number of observations for each study in the analysis dataset for this manuscript;
- A breakdown of the component criteria elements included in the definition of incident cancer, by age, sex, and race/ethnicity;
- A shell table for each analysis to list, for each study, the study population by demographic characteristics and the profile of study covariates at baseline by gender, race/ethnicity and incident cancer status, and the allele frequencies of the variants considered in these analyses;
- Establish a common approach for the estimation of population stratification and the inclusion of such estimates as covariates in the analyses;
- Specify the logistic regression model parameters to be run for each study in each stratum of race/ethnicity; with the progression in the inclusion of covariates and the criteria by which covariates are retained in a model;
- Specify the criteria to define effect modification and the a priori threshold of statistical significance for interaction.
- Indicate the preferred analytic approach to arrive at summary estimates and 95% confidence intervals (e.g., DerSimonian and Laird)

SOURCE OF DATA TO BE USED (Provide rationale for any data whose relevance to this manuscript is not obvious):

Check all that apply:

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

[X] EAGLE; [X] CALiCO; [X] MEC; [X] WHI; [X] CC; [ ] Other:

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