ARIC Manuscript Proposal # 1964

PC Reviewed: 7/10/12	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Heterozygosity as a Predictor of Longevity

b. Abbreviated Title (Length 26 characters): Het and Longevity

2. Writing Group: ARIC Het Working Group

Writing group members: Nathan A. Bihlmeyer, Dan E. Arking, Joe Coresh, Jim Pankow, Nora Franceschini, Linda Kao

Other ARIC study members are welcome to participate.

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. NAB **[please confirm with your initials electronically or in writing]**

F**irst author**: **Nathan A. Bihlmeyer**
Address: 733 N Broadway, BRB 515
Baltimore, MD 21205

Phone: (443) 287-0251 Fax: (410) 614-8600
E-mail: nbihlme1@jhmi.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

Name: **Dan E. Arking**
Address: 733 N Broadway, BRB 447
Baltimore, MD 21205

Phone: (410) 502-4867 Fax: (410) 614-8600
E-mail: arking@jhmi.edu

3. Timeline: Analyses to be completed spring of 2013, and manuscript submitted by July 2013.

4. Rationale:

Positive Heterozygosity Fitness Correlations (HFCs) have been found in many organisms, including humans (Chapman et al. 2009). In particular, associations have been

found between heterozygosity at the Major Histocompatibility Complex (MHC) (a.k.a. Human Leukocyte Antigen, HLA) region and general health in humans (Lie et al. 2009). Several studies in humans have linked increased MHC heterozygosity in a male with facial and auxiliary odor preferences in females (Thornhill et al. 2003 and Roberts et al. 2005). It is thought that these influences on mate selection are adaptive (Piertney et al. 2006). In the case of the heterozygosity in the MHC region, the cause of a positive HFC being observed is believed to be the result of increased antibody diversity conveying robust pathogen resistance and therefore increased general health (Piertney et al. 2006).

However in the case of whole genome heterozygosity, the mechanism of action is less clear. The current thinking is there are two mechanisms that act at a genome level to influence fitness. One mechanism is to compensate for deleterious mutations in polyploidy organisms by being heterozygous at sites where there is one of these mutations (Charlesworth et al. 2009). The second mechanism that acts at a genome level to influence fitness is theorized to be overdominance/heterozygous advantage (Charlesworth et al. 2009). For this study, we wish to measure whole genome heterozygosity in Caucasians to determine if increased heterozygosity at a whole genome level conveys increased longevity.

If promising associations are found, then we will stratify by cause of death and perform the analysis again to see if one particular disease or other cause of death is particularly influenced by heterozygosity. This could lead to altered care for such individuals leading to longer life. Also, findings from this study will be replicated in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) cohort.

5. Main Hypothesis/Study Questions:
Here we ask if Genome Wide Heterozygosity (GWH) stratified by genomic location can be correlated with mean survival in humans.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).
Data from 8,509 Caucasians from the Atherosclerosis Risk in Communities (ARIC) cohort are requested for this study. Affymetrix SNP genotyping data from sites across the genome and basic phenotype data is needed for each person. Because of the ARIC cohort's design, the fitness parameter chosen for the HFC is time in years from study entry until death or date of last mortality survey (a.k.a. follow-up time). Additional, we will use time in years from birth until death or date of last mortality survey (a.k.a. age) as an alternate fitness parameter in order to parse apart potential interactions with age. In order to correlate this type of survival data to genome wide heterozygosity, a Cox proportional hazards (CoxPH) model will be used. Phenotype and technical covariates will be tested in the model to ensure we are accurately testing our hypothesis. Possible covariates include data collection site, age at time of entering study, gender, general education level, BMI, income, and other available socioeconomic parameters. Cross-

sectional analysis of age and heterozygosity against survival will also be done to ensure the assumptions inherent in the CoxPH are valid for this study.

We will use several methods to estimate genome wide heterozygosity. First will be to sum the number of heterozygous loci and divide by the total number of loci for which there is data in each individual. Additional methods involve the same calculation, the difference being which loci are included. We hypothesize that heterozygosity in different parts of the genome may have a different effect on survival. See the following table for a list of the inclusion criteria.

Table 1: Lists of nine sets of SNPs used to calculate genome wide heterozygosity

All SNPs	SNPs per Chromosome
SNPs Within an Exon	SNPs Within 20kb of an Exon
SNPs Not Within an Exon	SNPs Not Within 20kb of an Exon
SNPs Within a Gene	SNPs Within 20kb of a Gene
SNPs Not Within a Gene	SNPs Not Within 20kb of a Gene

In order to make a conservative list of these parts of the genome (where the genes/exons are), we will consider any location annotated by CCDS, Refseq, Ensembl, or UCSC as a genomic element to be a part of that genomic element.

We will use Plink (http://pngu.mgh.harvard.edu/~purcell/plink/) and R-Project (http://www.r-project.org/) to perform the following analyses (Purcell et al. 2007 and Therneau 2012). First we will check the quality of each sample's data. We will set the minimum call rate to be included in this study to 95%. Exclusion based on sample call rate is primarily to exclude defective SNP chips on the basis that chips that cannot make enough SNP calls cannot make accurate SNP calls either. Next we will check to see if a sample's genotypic sex matches its phenotypic sex. Excluding samples because of sex inconsistencies is done primarily because inconsistencies many indicate that the sample has been switched for another by mistake. If that were true, the sample's phenotype data would be effectively random compared to its genotype and therefore reduce any correlation. We will perform IBD analysis to find related samples. Related samples reduced correlations by adding structure to the data and lead to inflated test statistics. First degree relatives will be excluded from the study. Nearest Neighbor analysis will be conducted with Plink to find the five closest related samples to a particular sample. Then the number of times a sample showed up in the list is counted and plotted. Outliers are assumed to be samples contaminated with DNA from other samples. This kind of contamination would artificially result in more heterozygous calls and reduce IBD0 between samples. Since artificially increased heterozygosity would reduce a HFC, the strict cut off of will be used. Principal component analysis (PCA) will be performed to account for population substructure. These principal components will also be added to the final model as covariates.

Each SNP's integrity will analyzed as well. A stringent cut off will be used to ensure only the highest quality SNP calls will contribute to our heterozygosity metric. Also, only SNPs that can be uniquely mapped back to the genome will be used. This is because

SNPs that do not map uniquely could be called heterozygous when in fact there is a single nucleotide difference between the places it maps to. Only autosomal SNPs will be used in this study to reduce sex bias. Lastly, SNPs will be excluded for having a Hardy-Weinberg Equilibrium p-value of greater than 0.001. This is to exclude SNPs of great effect that are under high selection.

7.a. Will the data be used for non-CVD analysis in this manuscript? __X__ Yes ____ No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = "CVD Research" for non-DNA analysis, and for DNA analysis RES_DNA = "CVD Research" would be used? __X__ Yes ____ No
(This file ICTDER03 has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript? __X__ Yes ____ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = "No use/storage DNA"?
__X__ Yes ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/ARIC/search.php

__X____ Yes _____ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?
MP#1511 Genomewide association of event-free survival and other longevity related phenotypes: the CHARGE Consortium

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? __X_ Yes ____ No

11.b. If yes, is the proposal
_X__ A. primarily the result of an ancillary study (list number* 2006.03, 2007.02)
___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* _____ _____ _____)

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PUBMED Central** whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to Pubmed central.

References

Chapman, J.R., Nakagawa, S., Coltman, D.W., Slate, J., and Sheldon, B.C. (2009). A quantitative review of heterozygosity–fitness correlations in animal populations. Molecular Ecology *18*, 2746–2765.

Charlesworth, D., and Willis, J.H. (2009). The genetics of inbreeding depression. Nature Reviews Genetics *10*, 783–796.

Lie, H.C., Simmons, L.W., and Rhodes, G. (2009). Does Genetic Diversity Predict Health in Humans? PLoS ONE *4*, e6391.

Piertney, S.B., and Oliver, M.K. (2006). The evolutionary ecology of the major histocompatibility complex. Heredity (Edinb) *96*, 7–21.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., Bakker, P.I.W. de, Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics *81*, 559–575.

Roberts, S.C., Little, A.C., Gosling, L.M., Perrett, D.I., Carter, V., Jones, B.C., Penton-Voak, I., and Petrie, M. (2005). MHC-heterozygosity and human facial attractiveness. Evolution and Human Behavior *26*, 213–226.

Thornhill, R., Gangestad, S.W., Miller, R., Scheyd, G., McCollough, J.K., and Franklin, M. (2003). Major Histocompatibility Complex Genes, Symmetry, and Body Scent Attractiveness in Men and Women. Behavioral Ecology *14*, 668–678.