Exome Chip variants associated with Galectin-3 levels and clinical cardiac disease

b. Abbreviated Title (Length 26 characters): exome chip associations with galectin-3

2. Writing Group:
Writing group members:
Linda Polfus, Carlos Iribarren, David Aguilar, Vijay Nambi, Amil Shah, [OTHERS WELCOME], Ron Hoogeveen, John W. McEvoy, Elizabeth Selvin, Kunihiro Matsushita, Wensheng Sun, Eric Boerwinkle, Santhi Ganesh, Scott D. Solomon, Myriam Fornage, Christie M. Ballantyne

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

First author: Linda Polfus, PhD
Address: University of Texas
1200 Pressler Street, Room E411
Houston, TX 77030
Phone: (713)500-9012 Fax: 713-500-0900
Email: Linda.M.Whitaker@uth.tmc.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:  Eric Boerwinkle, PhD
Address: University of Texas
1200 Pressler Street, Room E411
Houston, TX 77030
Phone: 713-500-9803 Fax: 713-500-0900
E-mail: Eric.Boerwinkle@uth.tmc.edu

3. Timeline: Analysis to begin immediately. Manuscript to be written and sent for publication within one year of approval.
4. Rationale:

Galectin-3 is an established independent marker of heart failure as well as all cause mortality and is particularly useful in early heart failure detection to predict prognosis.\textsuperscript{1-3} As a member of the galectin family of carbohydrate binding proteins, this lectin has affinity for beta-galactoside binding via a carbohydrate recognition domain (CRD). In particular, galectin-3 contributes to heart failure via different physiological processes including fibrosis, inflammation, and proliferation by mediating cell-cell and cell-matrix interactions of the CRD to lactosamine-containing cell surface glyconjugates.\textsuperscript{4} Inhibition of galectin-3 by N-acetyllactosamine in rats and mice was found to attenuate cardiac fibrosis and left ventricular dysfunction with the potential to reverse heart failure with extensive fibrosis.\textsuperscript{4} In humans with heart failure, the prognostic value of galectin-3 appears to be most pronounced in HF patients with preserved left ventricular ejection fraction (LVEF) compared to reduced LVEF.\textsuperscript{5} In a preliminary ARIC study analyzing racial differences with follow up for 6.8 years for galectin-3 prediction of heart failure (N=1375 European Americans (EA) and N=434 African Americans (AA)), an association was established for EA’s but not AAs.\textsuperscript{6} From existing epidemiologic and biologic knowledge, we aim to garner the prognostic significance of galectin-3 to inform precision medicine.

Another mechanism of heart failure includes the cytokine hypothesis with proinflammatory cytokines (tumor necrosis factor α, interleukin-1, interleukin-6, and interleukin-18) produced by damaged myocardium, which activate monocytes to produce more cytokines, which in turn continue to impair myocardium function.\textsuperscript{7} Galectin-3 linkage to monocytes is exhibited by profound galectin-3 expression with activated macrophages differentiated from human monocytes.\textsuperscript{8} One proposed mechanism in which galectin-3 may modulate immune response is by inducing T cell death by binding to complement T cell surface glycoprotein receptors, CD45 and CD71.\textsuperscript{9} There is also evidence that galectin-3 may regulate leukocyte migration by inducing exposure of the common death signal phosphatidylserine (PS) independently of apoptosis.\textsuperscript{10} With systematic evaluation of variants, both common and rare, we will uncover genes which play an important role in galectin-3 activity and inhibition as part of the development of heart failure, especially by the regulation of T cell death and immune response.

Genome-wide association studies (GWAS) among participants of European descent from the PREVEND cohort (N=3,776 discovery and N=3,516 replication) of over 2.2 million common SNPs imputed to HapMap 2.0 found significant loci within the galectin-3 encoding, lectin, galactoside binding soluble 3 (\textit{LGALS3}) and \textit{ABO} genes.\textsuperscript{11} Together, the \textit{LGALS3} and \textit{ABO} loci (rs2274273, P=2.34x10^{-188}; rs6444234, P=4.52x10^{-18}, respectively) that were reported, explained 29.2\% of the phenotypic variance. Some confounding is likely present for the highly significant \textit{LGALS3} result, due to high linkage disequilibrium structure in the \textit{LGALS3} region for which the galectin-3 assay binds to tracer antibody. Interestingly, galectin-3 binding differences in the ABO epitopes
have been observed with binding to polysaccharides occurring more strongly to the A- or B- histoblood group epitopes compared to O group.\textsuperscript{12}

Thus, by identifying genetic associations with galectin-3 and potential inhibitors from the ARIC cohort and corroborated by the CARDIA cohort from exome chip followed by fine mapping with ARIC genetic sequencing data, we aim to identify novel genes influencing an intermediate biomarker of heart failure, galectin-3, as well as investigating whether naturally occurring genetic galectin-3 inhibitors (which would replicate the actions of a specific drug target) are associated with diminished disease.

5. Main Hypothesis/Study Questions:

The overarching goal of this proposed study is to determine and investigate novel, genetic associations with galectin-3 levels in European Americans and African Americans.

Associations between novel loci and surrogate (e.g., echocardiography traits with HFPEF and HFREF) clinical cardiovascular outcomes may provide further mechanistic evidence of a gene’s impact on disease progression. Pleiotropy among related immunologic indicators such as neutrophil, monocyte, and leukocyte counts may reinforce mechanistic postulations.

Our primary hypothesis is that systematic evaluation of variants, both common and rare, will uncover genes which play an important role in galectin-3 levels and inhibition as part of the development of heart failure.

A. Rare and common variants from exome chip will be discovered and corroborated from ARIC and CARDIA cohorts.
B. Analyses will be performed using both single variant and gene-based approaches for sequencing data surrounding suggestive exome chip results will be conducted in ARIC for exome sequence and whole genome sequence.
C. For Whole Genome Sequencing Data: Overlapping sliding 4 kb windows, with the skip length of 2 kb

**Exposure:** natural log transformation of Galectin-3 Activity from ARIC visit 4 measured by ELISA (BG Medicine, Inc.)

**Lookups from suggestive Galectin-3 replicated results (P<10\textsuperscript{-5}):**

- Echocardiographic variables (3D) from ARIC visit 5: Left ventricular mass g, Left ventricular wall thickness cm, ejection Fraction <0.50\%\textsuperscript{13}
- Calculate HF with preserved ejection fraction and HF with reduced ejection fraction
- ARIC Visit 5: Pulse wave velocity
- ARIC Visit 4: NT-proBNP
- ARIC: Neutrophil count, basophil count, monocyte count, and lymphocyte count of 10\textsuperscript{9} cells/L\textsuperscript{14-16}
- ARIC Carotid MRI Flow Cytometry measured for the CD45 cell surface markers of lymphocytes and monocytes
- Subtyping individual ABO blood group by SNP to infer differences in galectin-3 binding via strength of association\textsuperscript{15,17}

**Clinical covariates (Visit 4-5):** Self-reported ethnicity, sex, age at measurement, renal disease by eGFR, pulse pressure, smoking status, height, BMI and lipids (total cholesterol or lipid profile), height, weight, blood pressure, heart rate, history of hypertension, estimated glomerular filtration rate, diabetes, dyslipidemia, coronary artery disease, prior MI or revascularization procedure, heart failure and prior hospitalization for heart failure, prior stroke or TIA.

**Incident Event data (Outcomes):** Incident hospitalized heart failure, CHD and stroke (adjudicated incident event data through December 2013) will be evaluated. In addition, to evaluate a gene/SNP’s impact on incident heart failure, CHD and stroke, survival analysis in carriers compared to non-carriers will be conducted.

**Analysis Model:**
ARIC Exome Chip data will be analyzed Study led by Linda Polfus with collaboration with CARDIA Exome Chip to confirm suggestive findings. Fine mapping from ARIC exome sequence and whole genome sequence data will be conducted. Lookups of top galectin-3 level replicated variants will be conducted to assess use as an instrumental variable for Mendelian randomization of outcomes. A manuscript is estimated to be submitted in mid 2017.

**Table 1. ARIC visit 4 participants with Galectin-3 measured**

<table>
<thead>
<tr>
<th>Platform</th>
<th>EA</th>
<th>AA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exome sequence frz5</td>
<td>5,917</td>
<td>1,997</td>
<td>7,914</td>
</tr>
<tr>
<td>Exome Chip Array</td>
<td>8,104</td>
<td>2,243</td>
<td>10,347</td>
</tr>
</tbody>
</table>

**Analysis Plan:** We will perform a number of preliminary QC steps, including plotting the distribution of betas and standard errors by study and plotting the allele frequencies of variants compared to the weighted average of the allele frequencies for the ARIC sample. QQ-plots will be examined at different points in the analysis in order to identify potential issues that would require additional refining of sample and variant filters.

Rare variant analyses will be carried out using SeqMeta. For galectin-3 association, we will begin with using a single SNP analysis with a multiple linear regression approach. In addition to analyzing genetic variants individually, we will perform analyses collapsing rare variants over genes\textsuperscript{18}. Rare (<1% allele frequency) and uncommon (1-5% allele frequency) variants will be collapsed gene-by-gene using a burden test. In its simplest form, the CMC (Combined Multivariate and Collapsing) method (Li et al. 2008), sums all
the minor alleles in a specified window that are below a pre-specified threshold (typically <1% or <5%) for each individual. SNP-set (Sequence) Kernel Association Test (SKAT) is a variance components approach that takes into account heterogeneity in direction of effect for different variants in the same gene.\textsuperscript{19,20}

**Methods:** Gene-based SKAT and T5 burden tests (N=18,891 genes) will be conducted filtering for functional variants with a minor allele frequency (MAF) less than 5% and included splicing, stop-gain, stop-loss, nonsynonymous variants, or indels.

For Whole Genome sequencing data, all variants with MAF<0.05 for a 4 kb sliding window will also be collapsed for burden and SKAT tests (Bonferroni corrected threshold 7.47x10^{-8}).

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).

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

(This file ICTDER 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?  ____ 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”?  ____ 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.cscu.unc.edu/ARIC/search.php](http://www.cscu.unc.edu/ARIC/search.php)

 ____ 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)? None

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

11.b. If yes, is the proposal

_____ A. primarily the result of an ancillary study (list number* _________)

_____ 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.

Agree.
References


