ARIC Manuscript Proposal #3803

PC Reviewed: 3/9/21	**Status: _____**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Proteomics of diabetes and glycemic biomarkers in the ARIC Study

b. Abbreviated Title (Length 26 characters): **Proteomics of diabetes**

2. Writing Group:

Writing group members: Mary Rooney, Jingsha Chen, Olive Tang, Christie Ballantyne, Eric Boerwinkle, Justin Echouffo Tcheugui, Chiadi Ndumele, Ryan Demmer, Jim Pankow, Pam Lutsey, Morgan Grams, Joe Coresh, Liz Selvin (others welcome; replication and specialized analyzes will lead to adding individuals who make substantive contributions)

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

First author: **Mary Rooney**
Address: 2024 E. Monument St. Suite 2-600
Baltimore, MD 21287
E-mail: mroone12@jhu.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: Elizabeth Selvin
Address: 2024 E. Monument St. Suite 2-600
Baltimore, MD 21287
E-mail: eselvin@jhu.edu

3. Timeline: We aim to complete the manuscript within 1 year from the time of approval and availability of SOMAScan data at visit 2. Meanwhile, we will conduct preliminary analysis using visit 3 as baseline for the primary/discovery analysis.

4. Rationale:

Diabetes is highly prevalent and is projected to increase in prevalence globally.[1] Type 2 diabetes accounts for the majority of diabetes diagnoses and is characterized by hyperglycemia, primarily due to insulin resistance.[2] However, the etiology of type 2 diabetes is not fully understood.[3] Furthermore, improved risk prediction for diabetes will be useful for improving the definition and utility of pre-diabetes and its treatment.

Improvements in proteomic measurements provide an opportunity to explore biologic pathways and improve risk assessment for diabetes onset. Previous serum (or plasma) proteomics of diabetes analyses have been conducted,[4-6] two of which included prospective analyses.[4,5] Among 1,367 Swedish older men (mean age 73 years),[4] seven proteins (cathepsin D, leptin, renin, interleukin-1 receptor antagonist, hepatocyte growth factor, fatty acid–binding protein 4, and tissue plasminogen activator) of 92 measurable proteins were statistically associated with greater insulin resistance (estimated using HOMA-IR). Two of these proteins—interleukin-1 receptor antagonist and tissue plasminogen activator—were positively associated with incident diabetes, but these proteins were no longer statistically significant after

adjusting for baseline fasting glucose. In the AGES-Reykjavik, 536 proteins of 4,137 proteins measured were associated with diabetes (437 associated with only prevalent diabetes, 16 associated with only incident diabetes, 83 associated with both prevalent and incident diabetes). A recent proteomics of diabetes paper (using data from Framingham and the Malmö Diet and Cancer Study) identified 146 plasma proteins associated with incident diabetes, replicating previously identified hits such as adiponectin and vitamin E binding glycoprotein afamin.[7] The few prospective studies that have examined proteomics of diabetes have generally been of smaller sample size (all less than 3,000 participants) or had fewer measurable proteins. None have examined the proteomics signatures associated with different levels of glycemic biomarkers.

Using SOMAScan data with ~5000 proteins measured in ARIC participants at multiple time points, we will examine the proteomic signatures associated with 1) incident diabetes, 2) prevalent diabetes, and 3) glycemic biomarkers. We will compare associations across glycemic markers (glucose, HbA1c, glycated albumin, fructosamine, and 1,5-anhydrogluticol) to identify proteins common across glycemic markers and those that are specific to different measures of glycemia. By examining associations across multiple markers, we hypothesize that we will be able to distinguish between trait-specific (i.e., diabetes) associations vs marker-specific associations, providing novel insights into the biological underpinnings of diabetes. We will also assess the improvement in risk compared to traditional risk factors.

We propose to use ARIC visit 2 for our discovery analysis in midlife. While SOMAScan is not yet available for visit 2 (where we have all diabetes biomarkers measured including HbA1c), we will use visit 3 as baseline for our preliminary analyses. Because risk factors and pathophysiology for diabetes may vary across the life course, we will also conduct a separate discovery analysis in late life (visit 5).

5. Main Hypothesis/Study Questions:

Aim 1: What are the proteomic signatures of incident diabetes?
Aim 2: What are the proteomic signatures of prevalent diabetes?
Aim 3: What are the proteomic signatures of glycemic biomarkers?

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).

Design: Prospective (Aim 1) and cross-sectional (Aims 2 and 3). Midlife (Visit 3 until visit 2 data are available) and late life (Visit 5) analyses will be conducted separately. The data will be divided into a 2/3 discovery and 1/3 validation sample.

Exclusions:

- Standard ARIC race-center exclusions
- Missing SOMAscan data at visit 2 or failed QC
- Missing covariate data
- For glycemic markers, exclude non-fasting blood draw and glucose-lowering medication use.
- For prospective analyses, exclude prevalent diabetes at baseline (visit 2)
- Participants with a known diagnosis of diabetes prior to age 30

Exposures: SOMAScan proteins (~5,000 proteins measured using SOMAScan's aptamer-based proteomics platform). We will be guided by the QC document following the current recommended data cleaning, focus on log protein levels, and use flag 2 to exclude unreliable data (flag2=0 for N = 4877 aptamers excluding flag 1 cat1-3: Fc mouse/Contaminants, or var<0.01 or soma qc cvba>50% at v3 or v5) – this is the suggested list of aptamers to analyze).

Covariates: age, sex, race-center, eGFR, family history of diabetes, body mass index, current smoking, total cholesterol, HDL-cholesterol, systolic blood pressure, hypertension medication use, physical activity (sport index)

Outcomes: For our first aim, the outcome will be incident diagnosed diabetes (diagnosis or glucose-lowering medication use) after midlife (visit 2 or visit 3; separate analysis for late-life, visit 5). The outcome for our second aim will be prevalent diagnosed diabetes, and the outcome for our third aim will be levels of glycemia biomarkers (available at visits 5 and 2: HbA1c, fasting glucose, glycated albumin, fructosamine, 1,5-anhydroglucitol), stratified by diagnosed diabetes.

Statistical Analysis

We will log-transform protein levels as needed and when comparisons across proteins is desirable, we will analyze protein levels on a standardized scale (mean=0, SD=1). When incident diabetes (Aim 1) is the outcome, we will use Cox regression. For prevalent diabetes as the outcome (Aim 2), we will use logistic regression. We will use diabetes-stratified linear regressions where glycemic biomarkers (Aim 3) – HbA1c, fasting glucose, glycated albumin, fructosamine, 1,5-anhydroglucitol—are the outcomes of interest. For all Aims, we will use a similar set of hierarchical models, shown below.

- M0: unadjusted
- M1: age, sex, race-center, eGFR
- M2: M1 + family history of diabetes, BMI, current smoking, total cholesterol, HDL-cholesterol, systolic blood pressure, hypertension medication use, physical activity
- For aims 1 & 2 we will further adjust for hyperglycemia to discover novel markers:
  M3: M2 + most recently available HbA1c and fasting glucose

To determine statistical significance threshold, we will apply the Bonferonni correction in the most conservative analysis. FDR ($q<0.05$) will be used to examine a broader list of proteins for pathway analyses.

Replication within ARIC: In the replication analysis, we will examine which Bonferonni significant "hits" in the development (2/3) sample were also significant in the (1/3) validation sample at $p<0.05/\#hits$ tested. This will be done separately for midlife and late-life since we hypothesize some risk factors may vary across the lifespan. We will also test which proteins from midlife validate as late life diabetes predictors. We will examine prevalent and incident diabetes separately. Similar analyses will be conducted cross-sectionally for glycemic biomarkers (HbA1c, fasting glucose, glycated albumin, fructosamine, 1,5-anhydroglucitol).

Mendelian Randomization: We will conduct bi-directional Mendelian randomization using SNPs identified from published GWAS to validate potential causal pathways and their direction (causes of diabetes should have protein SNPs also lead to DM risk vs. consequences of diabetes where diabetes GWAS SNPs will lead to alterations in the protein) as we have done for dementia and kidney disease.

Other interactions: We will explore the consistency of proteomic associations by **sex and follow-up time**. Men and women differ markedly in body composition, a key determinant of glucose metabolism, suggesting some pathways for diabetes risk may differ by sex. We hypothesize that most "hits" will be shared across men and women while some proteins will play a greater role in women vs. men. The more specific approach to this will be to test interaction of validated hits in their associations across sex. The more sensitive approach will be to conduct separate discovery in the subgroups and compare the results.

The latter approach has lower specificity and its findings will require validation. A similar comparison could be made across race but we hypothesize that racial differences will be a product of social and behavioral differences overlaid on top of a similar biology.

The long duration of follow-up in ARIC (~30 years from midlife) means that risk over the full follow-up includes both short and long term risk factors which may not be the same. We will test whether the strength of association varies by decade of follow-up. We hypothesize that most proteomic risk factors in the first decade will also be risk factors in subsequent decades but the magnitude of the association will often diminish. As a result, we will focus on 10-year incidence in predictive models.

Prediction: We will develop predictive models in the development sample and test their discrimination and calibration in the validation sample. We will examine 10-year risk of diabetes as the primary outcome. We will consider a comparison of M3 above (full risk factor model including measures of hyperglycemia) as the comparison model to be improved (testing change in C-statistic and categorical NRI). We will also examine discrimination and calibration among individuals currently defined as pre-diabetic based on fasting glucose of HbA1c. The analyses in midlife at visit 3 will be limited by not having HbA1c at that visit.

Limitations: The type of diabetes is not known; given the age-range of participants, it is likely most have type 2 rather than type 1. In efforts to address this, we can exclude the small number of participants with a known diagnosis of diabetes prior to age 30. The lack of a separate cohort for external validation is another important potential limitation – we will seek external replication in collaboration with Peter Ganz and others working with Soma Logic.

7.a. Will the data be used for non-CVD analysis in this manuscript? ____ Yes __x__ 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? __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)?

No prior proteomics of diabetes proposals.

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* _ AS2017.27___)

___ 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

1. International Diabetes Federation. *IDF Diabetes Atlas.* 2019.
2. American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2020. *Diabetes Care.* 2020;43(Suppl 1):S14-s31.
3. Ingelsson E, McCarthy MI. Human Genetics of Obesity and Type 2 Diabetes Mellitus: Past, Present, and Future. *Circ Genom Precis Med.* 2018;11(6):e002090.
4. Nowak C, Sundström J, Gustafsson S, et al. Protein Biomarkers for Insulin Resistance and Type 2 Diabetes Risk in Two Large Community Cohorts. *Diabetes.* 2016;65(1):276-284.
5. Gudmundsdottir V, Zaghlool SB, Emilsson V, et al. Circulating Protein Signatures and Causal Candidates for Type 2 Diabetes. *Diabetes.* 2020;69(8):1843-1853.
6. Beijer K, Nowak C, Sundström J, Ärnlöv J, Fall T, Lind L. In search of causal pathways in diabetes: a study using proteomics and genotyping data from a cross-sectional study. *Diabetologia.* 2019;62(11):1998-2006.
7. Ngo D, Benson MD, Long JZ, et al. Proteomic profiling reveals novel biomarkers and pathways in type 2 diabetes risk. *JCI Insight.* 2021.