ARIC Manuscript Proposal Form

ARIC Publication Admin Use Only

Publication Committee Review Date: [_**01/09/24**_]
ARIC Manuscript Proposal Number: **#4390**

1.a. Full Title [Validation of the SomaScan 11k platform and comparisons to SomaScan 5k, Olink, and targeted immunoassays]

b. Abbreviated Title (Length 26 characters): **Soma 11k validation**]

2. Writing Group [please provide a middle initial if available; EX: Adam L Williams]:

Writing group members: [Mary R. Rooney, Jingsha Chen, Christie M. Ballantyne, Ron C. Hoogeveen, Eric Boerwinkle, Bing Yu, Keenan Walker, Morgan E. Grams, Pascal Schlosser, Elizabeth Selvin, Nilanjan Chatterjee, David Couper, Josef Coresh. Others welcome. Authors will be added if additional platforms are added (e.g. Olink 3k; NULISA neuro panel)]

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

First author [please provide a middle initial; EX: Adam L Williams]:

[**Mary**] [**R**] [**Rooney**]
First name Middle initial Last name

Address: [2024 E. Monument St., Suite #2-600; Baltimore, MD 21287]

Phone: 630-995-5772]
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 (The ARIC author should be involved enough in ARIC to be able to point the lead author to appropriate ancillary study PIs and to be able to search ARIC manuscript proposals if the lead author doesn't have the access needed to do such a search).

Name: Josef Coresh]
Address: [TRB 227 E. 30th Street 7th, 707; New York, NY 10003]

Phone: 646-501-2503]
E-mail: josef.coresh@nyulangone.org]

3. Timeline: We anticipate submitting a manuscript for review by ARIC (and SomaLogic) within 3 months of receipt of all data. Soma 11k (mednorm) data is available and have been cleaned. We anticipate that Soma11k (ANML) and Olink HT 5k data will be available by early 2024.

4. Rationale:

The capture of proteomic technology is expanding rapidly resulting in measurements of an unprecedented number of human proteins. In fall 2023, SomaLogic released the SomaScan 11k platform, which enables measurement of an unprecedented >10,000 plasma proteins using aptamer-based technology. (For context, in humans, there are ~18,000 known protein-encoding genes.) As the number of measurable proteins expands, there is continued need for validation studies of proteomic biomarkers, both in terms of comparisons to earlier versions of the same platform and across different technologies.

Multiple studies have reported a wide range of correlations for plasma proteins measured using earlier versions of the SomaScan platform (1-5k proteins measured) vs Olink immunoassay technology,[1-5] including findings from ARIC. We have published ARIC findings on the cross-platform comparisons of the SomaScan 5k (v4, 5284 proteins measured) with 5 panels of the Olink 96 (460 proteins measured on CVD II, CVD III, inflammation, organ damage, and cardio-metabolic panels) among ~400 ARIC participants at visit 5. We also compared proteins measured on the SomaScan 5k with targeted immunoassays in a subset of up to 116 participants ("ARIC calibration panel"). Across the platforms, overall, we found that approximately half of the overlapping proteins tend to correlate moderately or highly ($r>0.5$) and the other half having poor correlations ($r<0.5$).

In 2023, we received SomaScan 11k data among the same ~116 ARIC participants at visits 2 and 5. Olink HT 5k data are planned (expected early 2024). Using data from up to 116 ARIC participants (part of the ARIC calibration panel), we will provide estimates of validity and reliability for the SomaLogic 11k protein measurements. Specifically, we will provide (a) estimates of precision (CVs) based on duplicate measurements of samples; (b) comparisons of SomaScan 11k to targeted immunoassays (clinical or research use); (c) comparisons of proteins to the earlier SomaScan 5k platform from visits 2 and 5; and (d) cross-platform agreement of SomaScan 11k vs Olink HT 5k assay.

5. Main Hypothesis/Study Aims:

Our aims are:

1) To characterize the validity and reliability of plasma proteins quantified on the SomaScan 11k within the ARIC Study at visits 2 and 5.
2) To assess the agreement of proteins measured on the SomaScan 11k platform vs. targeted immunoassays within the ARIC Study at visits 2 and 5.
3) To compare the relative abundance of the 5284 overlapping plasma proteins measured on the latest SomaScan 11k platform vs the SomaScan 5k platform at visits 2 and 5.
4) To assess the comparability of proteins measured using aptamer-based SomaScan 11k assay versus the immunoassay-based Olink HT 5k (measurements to be started soon and

data expected early 2024) at ARIC visit 5. We will additionally explore correlations with Olink 3072 (as well as the NULISA neurology panel) and may include these results in the paper if helpful rather than cluttering the paper.

6. Design and analysis - please address the following aspects:

a) inclusion/exclusion
b) study design
c) outcome and other variables of interest with specific reference to the time of their collection
d) summary of data analysis
e) Any anticipated methodologic limitations or challenges if present

Study Design. This study will include up to 116 ARIC participants as described in our prior publication reporting comparisons of the SomaScan v4 (~5k proteins measured) and targeted immunoassay[2] (hereafter called the "ARIC calibration panel"). Participants with prevalent coronary heart disease (CHD), stroke, or heart failure at visit 5 were not included in this calibration panel. Participants with incident CHD, stroke, or heart failure after visit 5 could be cases and controls included those without incident CHD, stroke, or heart failure, and did not die within 5 years of visit 5. The ARIC calibration panel included 58 cases, 58 controls, and 26 Baylor QC pools. Cases were balanced by age (< or ≥ median age of 73), sex (M/F), race (B/W) and eGFR (≥60 or <60). Controls were frequency matched to the age (±10 years), sex, race and eGFR groupings of cases.

We recently obtained data quantifying >10,000 plasma proteins (SomaLogic; SomaScan 11k) on 116 individuals from visit 5 (sent in duplicate) as well as the same individuals at visit 2 (sent in duplicate, 12 were missing plasma samples at visit 2) plus 26 Baylor Calibration panel samples. All samples from visit 2 (n=104) and visit 5 (n=116; ~6 samples did not meet QC on the 5k assay and may be excluded from this study) will be analyzed using the new SomaScan 11k platform. Plasma samples were sent in duplicate to allow estimation of assay precision based on both ARIC visits 2 and 5.

Table 1. Summary of proteomic data available in ARIC at visits 2 and 5

	Visit 2 (n=14348)	**Visit 5 (n=6538)**
SomaScan 5k	X	X
SomaScan 11k (new)	X (n=104)*	X (n=116)*
Olink 96		X (n=500)**
Olink 3k (new)		X (n=116)*
Olink 5k (planned)		X (n=116)*
NULISA neurology panel (planned)		X (n=116)*

*ARIC Calibration panel. **~30 of the participants with Olink 96 data overlap with the ARIC calibration panel

Statistical Analysis

Aim 1: We will use adaptive normalization maximum likelihood (ANML) transformed data from SomaScan for most analyses. Separately for visits 2 and 5, we will flag aptamers measured on

the SomaScan 11k with a CV_{BA} >50%, a variance <0.01 on the log scale, or binding to mouse Fc-fusion, contaminants, or nonproteins (flag2=1). We will exclude protein outliers >5 SD from the mean. Using blind duplicates of the specimens, we will derive coefficients of variation using a method by Bland and Altman (CV_{BA}= e√ mean(variance(ln (X))) − 1).[6] We will calculate the reliability coefficient (intraclass correlation coefficient) based on the blind duplicates at each visit. We will summarize CV_{BA}'s and ICC's in histograms and using means/percentiles. In supplemental analyses, we will summarize (means, medians, p5, p95) the correlations of the 11k SomaScan proteins using the mednorm data vs proteins that have undergone ANML.

Aim 2: We will include scatterplots for the relative abundance of proteins measured on the SomaScan 11k platform vs absolute measurement of proteins quantified using targeted immunoassays (log2 scale similar to SomaScan proteins). We will generate Spearman's correlations of proteins measured on targeted immunoassays vs SomaScan 11k (ANML) as listed in the table below (additional relevant targeted immunoassays at visits 2 and 5 can be included for comparison where helpful to other ARIC investigators). Additionally, we will compare the correlations of proteins measured on targeted immunoassays vs SomaScan 11k (mednorm and unnormalized at visits where these versions of the SomaScan data are available).

Table 2. List of immunoassays for targeted comparison against SomaScan 11k

	Targeted Immunoassays (clinical or research use)
Visit 2	-Albumin -B2M -Cystatin-C -CRP -Troponin T -NTproBNP
Visit 5	-Albumin -B2M -Cystatin-C -CRP -Troponin T -NTproBNP -GDF15 (Roche) -Multiplex analytes (research use): GDF-15, ST2, Osteopontin, IL-6, MMP-1, TIMP-1, MMP-3, MMP-7, MCP-1, IL-10, VCAM-1, ICAM-1, IL-18 TNF-α

Aim 3: We will conduct analyses separately at visits 2 and 5 for the comparison of aptamers measured on the SomaScan 11k vs 5k (5284 overlapping aptamers). SomaScan data will be ANML transformed. We will generate scatterplots for each SOMAmer ID that overlaps on the 11k and 5k platforms, obtain Spearman's correlation and p-value for each aptamer pair and report the intercept, slope from linear regression. We will include correlations and mean RFU for aptamers as measured on the SomaScan 11k vs 5k on the scatterplots.

Aim 4: We anticipate >4000 overlapping proteins on the SomaScan 11k and Olink 5k (>2500 proteins overlap with the Olink 3k). After excluding outliers (>3 standardized residual from linear regression), we will obtain Spearman's correlation and p-value for each protein pair and report the intercept, slope from linear regression. When there are multiple measurements of the same protein within a platform (e.g. >1 SOMAmer ID), we will calculate and summarize the pairwise correlations for all protein comparisons. We will summarize the correlations using

histograms and using summary statistics (mean, median, percentiles). We will generate scatterplots for each protein pair (Olink [y-axis] vs Soma [x-axis]) with linear regression and loess regression overlaid (likely random selection of 5-10 protein comparisons per 0.25 increment in Olink vs SomaScan correlation). We will provide summary statistics for each protein (as quantified on the SomaScan and on the Olink platforms) using mean, SD, and SD/mean. We will also graph a scatter plot of the correlation coefficients for each protein pair with vs without excluding outliers. In sensitivity analysis, we will also calculate and summarize the correlations for each unique overlapping protein (after taking the mean of the pairwise correlations for each protein).

We will conduct analyses to try to characterize which assay is more biologically relevant by plotting correlations of the overlapping proteins on respective platforms with kidney function (eGFR is related to approximately one third of plasma proteins). Separate ARIC papers will more deeply explore the biologic relevance of proteins.

Anticipated Limitations

We recognize that the sample size is relatively small. However, this subset of up to 116 individuals has extensive clinical and omic phenotyping and, for this analysis, we are interested in characterizing agreement of protein measurements across platforms rather than discovering risk associations. Our results for SomaScan 11k vs Olink will provide important information on cross-platform correlations for ARIC and non-ARIC investigators (eg UK Biobank has Olink data but not SomaScan).

Additionally, SomaScan provides relative (not absolute concentrations) abundance of the proteins and proteins may not be on the same scale (for 5k vs 11k or for 11k vs Olink).

f) Will the author need Limited data to complete the proposed manuscript? ☒ Yes, Limited data is needed (Provide a brief (2-3 sentences) justification for requesting PHI data) This proposal requires the use of proteomics/SomaLogic data, which falls under the limited data category. Other limited data (including dates, CMS, genomic, geocoded) are not required. **☐ No, De-identified data will be sufficient.**

*Please note, Limited dataset access is strict and rarely provided. Limited data includes identifiable information such as dates (birthdays, visit dates, etc.). CMS, Genomic, Geocoded, *Proteomics/Somalogic*, and other -omic data all fall under the limited data category. De-identified data does not include dates. All dates are date adjusted to "Days since Visit 1".

7.a. Will the data be used for non-ARIC analysis or by a for-profit organization in this manuscript? (Non-ARIC analysis means that the authors are not regarded as ARIC investigators and the "ARIC author" is essentially just a facilitator rather than an integral part of the writing group.) ☐ Yes ☒ No

b. If Yes, is the author aware that the current derived consent file ICTDER05 must be used to exclude persons with a value RES_OTH and/or RES_DNA = "ARIC only" and/or "Not for Profit" ? ☐ Yes ☐ No

(The file ICTDER is distributed to ARIC PIs annually, 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 current derived consent file ICTDER05 must be used to exclude those with value RES_DNA = "No use/storage DNA"? ☐ **Yes** ☐ **No**

9. The lead author or the "sponsoring" ARIC 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 website at: https://aric.cscc.unc.edu/aric9/proposalsearch [ARIC Website Publications Proposal Search]

☒ 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)?
Rooney MS #3835 Validation of proteomic measurements across platforms

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or does it use current [or ongoing] ancillary study data (this includes ACHIEVE)? ☒ **Yes** ☐ **No → Skip to question 12**

11.b. If yes to 11.a., is the proposal

☒ **A. primarily the result of an ancillary study**

☐ **B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables)**

11.c. If yes to 11.a., list number* 2017.27

*ancillary studies are listed by number https://aric.cscc.unc.edu/aric9/researchers/ancillary_studies/approved_ancillary_studies [ARIC Website Ancillary Studies Approved Ancillary Studies]

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 https://aric.cscc.unc.edu/aric9/publications/policies_forms_and_guidelines [ARIC Website Publications Publication Policies, Forms, and Guidelines]. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.

References:

To view publications materials, click "Log in" at the top right of the ARIC website. Click "Forgot Password" if you are experiencing issues with logging in.

1. Raffield LM, Dang H, Pratte KA, et al. Comparison of Proteomic Assessment Methods in Multiple Cohort Studies. *Proteomics.* 2020;20(12):e1900278.
2. Rooney MR, Chen J, Ballantyne CM, et al. Comparison of Proteomic Measurements Across Platforms in the Atherosclerosis Risk in Communities (ARIC) Study. *Clinical chemistry.* 2023;69(1):68-79.
3. Pietzner M, Wheeler E, Carrasco-Zanini J, et al. Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat Commun.* 2021;12(1):6822.
4. Haslam DE, Li J, Dillon ST, et al. Stability and reproducibility of proteomic profiles in epidemiological studies: comparing the Olink and SOMAscan platforms. *Proteomics.* 2022;22(13-14):e2100170.
5. Eldjarn GH, Ferkingstad E, Lund SH, et al. Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature.* 2023;622(7982):348-358.
6. Bland JM, Altman DG. Measurement error proportional to the mean. *BMJ (Clinical research ed).* 1996;313(7049):106.