ARIC Manuscript Proposal Form

ARIC Publication Admin Use Only

Publication Committee Review Date: **08/15/24**
ARIC Manuscript Proposal Number: **#4519**

1.a. Full Title: Discovering plasma proteins associated with common solid cancer incidence in ARIC

b. Abbreviated Title (Length 26 characters): Proteins and solid cancers

2. Writing Group [please provide a middle initial if available; EX: Adam L Williams]:
Writing group members:
Ziqiao Wang (first)
Vernon A Burk
Nilanjan Chatterjee (co-senior)
Elizabeth A. Platz (co-senior)

We invite other ARIC Cancer and ARIC investigators:
Zhike (Coco) Lin
Corinne E. Joshu
Anna Prizment
Josef Coresh
Kenneth R. Butler
David J. Couper

We will invite other interested ARIC investigators.

We will also invite select members of the EPIC investigators for their contributions to confirmation of the top hits.

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

Elizabeth Platz prepared the manuscript proposal on behalf of the team. Ziqiao Wang will be first author of the manuscript, and Elizabeth and Nilanjan Chatterjee will be co-senior authors.

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

Ziqiao | | **Wang**

First name | Middle initial | Last name

Address: Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe St.
Baltimore, MD 21205

Phone:
E-mail: zwang389@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: Nilanjan Chatterjee
Address: Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe St.
Baltimore, MD 21205

Phone:
E-mail: nchatte2@jhu.edu

Name: Elizabeth A. Platz
Address: Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe St.
Baltimore, MD 21205

Phone: 443-850-4143
E-mail: eplatz1@jhu.edu

3. Timeline: 1 year – summer 2025

4. Rationale: One in two men and one in three women will develop cancer in the lifetimes, and cancer remains the second leading cause of cancer death in the US [1]. Public health approaches to reduce cancer burden in the at-risk population include preventing and intervening on cancer risk factors and early detection of cancer in asymptomatic individuals. Targeting these primary and secondary prevention interventions to individuals based on their predicted cancer risk is a strategy to increase benefit while reducing harms to individuals from the interventions as well as to use more judiciously limited healthcare resources. To implement such a strategy, tools are needed to predict cancer risk in the general population. We propose that plasma proteins,

reflecting both exogenous and endogenous factors, could serve as the basis for or components of tools for risk stratification to guide decision-making for primary and secondary prevention of cancer in adults. As a step toward developing such tools, in this study, we propose to discover plasma proteins that are associated with future risk of solid cancers. We will confirm the top hits in a case-cohort study (PI, Marc Gunter) nested in 4 of the countries in EPIC, a cohort of Europeans. EPIC used the 7K SomaScan®, which includes all of the proteins in the 5K SomaScan® used by ARIC. We will also perform a 'look up' for EPIC for their top hits. No participant data will be exchanged between ARIC and EPIC.

In ARIC (manuscript proposal #4144), we previously identified 212 proteins (of 4,712 plasma proteins or protein complexes [4,955 aptamers] that were associated with post-menopausal breast cancer risk at the conventional p-value of <0.05. We used Visit 2 proteins, and after a median follow-up of 23.3 years ascertained 340 incident breast cancer cases among 4,403 post-menopausal women. After correction for false discovery, two proteins were statistically significant (per doubling: LEG1 homolog [Q6P5S2] HR=1.45, 95% CI: 1.23-1.70, $p=7.47*10^{-6}$; ADP-dependent glucokinase [Q9BRR6] HR=2.52, 95% CI: 1.66-3.83, $1.50*10^{-5}$). The associations for both proteins were similar in Black and White participants, and were not notably changed after adjusting for breast cancer risk factors. These associations remained after excluding the first three years of follow-up, suggesting that these two proteins are not marking the presence of an undetected cancer. The two proteins were confirmed in the EPIC case-cohort study (LEG1 homolog, HR=1.24, 95% CI: 1.14-1.35, $p=9.79*10^{-7}$, ADP-dependent glucokinase, HR=1.13, 95% CI: 1.03-1.23, p=0.01). These preliminary data support the feasibility of the general approach that we will use for other common solid cancers.

We have previously investigated proteins and cancer in ARIC for confirmation of proteins associated with cancer risk, risk-factor-based discovery, and for the early detection of cancer. We describe here what we found and how we will compare and reconcile the findings from our proposed and prior work.

Confirmation: Using the SomaScan® data, we confirmed the top hits from published studies on genetically-predicted proteins and risk of pancreatic (confirmed 10 of 53; MS #3772; [2]) and prostate (confirmed 5 of 30; MS #3676; accepted in *The Prostate*) cancers. All of the confirmed plasma protein associations remained after excluding the first several years of follow up. In the proposed discovery studies, we will assess whether the identified proteins for pancreatic and prostate cancer are among those we confirmed and did not confirm in our past work. We will also assess whether the identified proteins are among the genetically-predicted proteins that were associated with these cancers, but were not statistically significant after correction for false discovery in prior publications ([3-5].

Risk factor-based discovery: We also developed a plasma protein score for detailed smoking history and evaluated the association of that score with smoking-associated cancers combined and lung cancer separately (ARIC manuscript proposal #3431). The approach we previously used differs conceptually from the proposed agnostic discovery of proteins associated with smoking-associated cancers. However, we expect overlap in the proteins when using the two approaches. In the discovery of the proteins, we will run models before and after adjustment for detailed smoking history. We will also compare the identified

proteins using our current and previous approaches to determine whether any proteins are indicative of smoking-associated cancers but are not marking detailed smoking history. Such proteins may mark other endogenous or exogenous factors that influence smoking-associated cancer risk.

Anna Prizment's team (includes Elizabeth Platz) developed a cancer-specific protein score for accelerated aging [6] and has investigated the score with cancer incidence (ancillary study 2021.06 "Proteomic aging clock and cancer: Atherosclerosis Risk in Community study", R21 funded). It is possible that the proteins we identify in this proposal will overlap with that score given the strong link between aging and cancer. After we complete the discovery, we will compare whether any of the proteins associated with cancer risk are included in the accelerating aging score.

Early detection: In this proposed study, we aim to identify plasma proteins that mark cancer risk (likelihood of developing a cancer in the future), not the likelihood that an undetected cancer is present (early detection). We have already conducted research on proteins that predict the current presence of a cancer under ARIC manuscript proposal #3909). The key analysis that differentiates between risk and early detection, is the lagged analysis: proteins that remain associated with risk after excluding the first several years of follow up are less likely to be produced by or in response to an undetected cancer. In the proposed study, for the top hits, in a sensitivity analysis, we will exclude the first 3 and 5 years of follow up.

To our knowledge, only one group has addressed this research question using a different large-scale proteomics approach, Olink. That study was conducted in the UK Biobank and evaluated 1,468 proteins in relation to multiple chronic disease/conditions linked with aging, including brain/CNS cancer, gynecological cancer, lung cancer, colorectal cancer, breast cancer, and prostate cancer, during 15 years of follow up and with correction for multiple testing (Bonferroni, $p<3.1*10^{-6}$ threshold) [7]. Very few proteins were associated (all were in the positive direction) with prostate (N=2, GP2, TSPAN1), colorectal (N=2, REG4, TFF3) and breast (N=3, AREG, STC2, CRLF1) cancer either before or after full adjustment for risk factors. In contrast, many proteins were associated with lung cancer (N=217) in minimally adjusted models, and several remained (N=33), 2 inversely (FASLG, TNR) and 31 positively (CEACAM5, CXCL17, GDF15, WFDC2, TNFRSF10B, PIGR, TNFSF13B, SPINT1, PLAUR, LAMP3, IL6, MMP12, IL6, IL6, SFTPA2, PRSS8, HGF, SFTPD, CDCP1, TREM2, ALPP, SFTPA1, ADM, ST6GAL1, CLEC4D, IL6, SERPINB8, HAVCR2, IL15, IGFBP4, EFNA1 – note that some proteins such as IL6 had multiple detection methods, so the total number of proteins is actually fewer), after adjusting for multiple risk factors (age, sex, BMI, alcohol consumption, social deprivation, educational attainment, smoking status and physical activity). They reported that batch, study center, and principal components were not associated with protein levels, and thus, did not adjust the analysis for these. For lung cancer, residual confounding could have remained from not taking into account detailed smoking history (which is available in ARIC). For breast cancer, all 3 of the top hits from the UK Biobank were in the SomaScan® in ARIC; only CRLF1 was associated with post-menopausal breast cancer (per doubling, HR=1.21, p=0.03) and the association was in the same direction. The top hits in ARIC, LEG1 homolog and ADP dependent glucokinase, were not measured in the Olink scan used in UK Biobank study. In addition to the use of different protein measurement platforms, differences between the two studies include a

longer follow-up time for breast cancer, restricting to post-menopausal breast cancer, and adjusting for reproductive risk factors in ARIC.

Thus, more work remains to be done to investigate plasma proteins as risk factors for common solid cancers. A strength of this proposal is that we have already developed the pipeline for this work and have made arrangements for a confirmation cohort. We anticipate 1 manuscript reporting on all common solid cancers other than breast.

5. Main Hypothesis/Study Aims:

Aim 1. To discover plasma proteins that are associated with risk of common solid cancers:

a. prostate
b. lung and bronchus
c. colorectal
d. pancreas
e. kidney
f. bladder

We hypothesize that each cancer site has a unique set of plasma proteins that mark risk and are not produced by or in response to an extant but undetected cancer.

Aim 2. To discover plasma proteins that are associated with risk of combinations of solid cancers:

a. all solid cancers
b. obesity-associated solid cancers
c. smoking-associated solid cancers
d. adenocarcinomas

Note: We will include breast cancer for all, obesity association, and adenocarcinomas.

We hypothesize that cancer sites with similar etiologies (exposures) or histologies will have sets of plasma proteins in common that mark endogenous and exogenous factors, and thus, risk.

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

a) inclusion/exclusion

Exclusions:

- Did not consent to studies on outcomes beyond CVD or to industry studies
- Do not have SomaScan data at Visit 2
- Had a history of cancer before Visit 2 (prevalent)
- Not Black or White

b) study design

- Prospective cohort design
- Follow up from Visit 2 through 12/31/2015

c) outcome and other variables of interest with specific reference to the time of their collection

Exposure: SomaScan 5K at Visit 2

Outcome: Incident cancer (first primary) after Visit 2 through 12/31/2015 (existing cancer case files)

- all solid cancers
- obesity-associated solid cancers (primary set with convincing evidence: breast [post-menopausal], colorectal, endometrial, esophageal [adenocarcinoma only], kidney, liver, pancreatic; secondary set with probable evidence: head and neck [mouth, pharynx, larynx], stomach [cardia], gallbladder, ovarian, prostate [advanced only] [8])
- smoking-associated solid cancers (primary set with sufficient evidence: bladder, cervical, colorectal, esophageal, kidney, laryngeal, liver, lung, oral cavity and pharyngeal, pancreatic, stomach [9]
- adenocarcinoma histology (most breast, most colorectal, most kidney, most pancreas, most prostate, most stomach, some lung, some esophagus)
- prostate
- lung and bronchus
- colorectal
- pancreas
- kidney
- bladder

Covariates at Visit 2 unless otherwise noted

- PEER factors
- age (Visit 1)
- sex (Visit 1)
- field center (Visit 1)
- race (Visit 1)
- height
- BMI
- physical activity (meeting guidelines)
- smoking status
- pack-years smoked
- smoking-related plasma protein score (developed by Meng Ru et al. under ARIC manuscript proposal #3431)
- alcohol drinking status
- ethanol intake
- use of post-menopausal hormones (females)
- diabetes status (diagnosed, undiagnosed, at risk, normoglycemia)
- eGFR
- clinical diagnosis of liver disease based on response to the question “Has a doctor ever told you that you have cirrhosis or another chronic liver disease” and from

routinely collected hospital discharge summaries for liver reasons (ICD-9: 571, ICD-10: K74)
- genetic instruments for protein levels
- consent variables

d) summary of data analysis

We will exclude participants without appropriate consent for this study on proteins and cancer. We will pre-process the proteomics data (flag 2) which are already $\log_2$ transformed, pre-process the cancer data and specific covariates related to each cancer site, and perform Cox proportional hazards regression for each of the 4,955 aptamers for each common solid cancer site, total cancer, obesity-associated cancers, and smoking-associated cancers. We will explore whether the associations differ by age (median among the cancer cases at diagnosis), sex (other than prostate), and race (Black, White), and if appropriate, test for interaction using the likelihood ratio test. For the top hits, we will visualize the shape of the association and test for linearity by including spline terms for the proteins in a model and compare the fit with the model for the protein entered as a continuous variable.

We will perform model diagnostics of the Cox proportional hazards assumption for the proteins by time, and if violation is identified, we will model a term for protein*time. We will include splines for age at Visit 2 in the Cox model to account for the violation of the proportional hazards assumption that we have already identified in the breast cancer analysis and in the developing the pipeline for this work.

All analyses will be performed with and without taking into account the PEER factors.

We will determine whether adjustment for eGFR or exclusion of participants with abnormal eGFR notably affects the findings. We noted that in Gadd et al. a very large number of proteins was associated with liver disease[7]. The liver is the source of 90% of the mass of plasma proteins. In a sensitivity analysis, we will exclude participants with liver disease or high liver fibrosis score [we previously calculated these for manuscript proposal #3178).

For lung cancer, we expect that many of the proteins that we identify will mark cigarette smoking, and as mentioned above, these proteins will likely overlap somewhat with the proteins that we previously identified (ARIC manuscript proposal #3431). We will use multiple adjustments for detailed smoking history, including adjusting for smoking status (current, former, never), pack-years smoked, and time since quitting, and also adjust. Any such proteins that do not overlap or remain after detailed smoking history adjustment may be capturing other endogenous or exogenous factors that influence lung cancer risk.

We will compare proteins we identify (before and after adjusting for multiple testing) with our previous findings for candidate proteins and prostate and pancreatic cancer risk (see Rationale).

Some cancers have several common histologies, such as lung cancer, and these histologies may have different risk factors (i.e., etiologic heterogeneity). We will perform the discovery work additionally separately by those histologies for a given cancer site (where possible based on sample size, e.g., lung cancer, non-small cell and small cell), and also for all adenocarcinomas combined (the most common one for breast, prostate, colon). Sample size will be smaller, so we will use the conventional p-value, confirm in EPIC, and be clear in the manuscript.

We seek to differentiate between plasma proteins as markers of risk versus early detection because this affects the nature of the intervention: preventing/intervening on exposures versus targeting for screening. The main strategy that we will use to identify proteins associated with risk (and not the current presence of a cancer) is to start follow-up 3 to 5 years after Visit 2 but still using Visit 2 proteins.

After the top hits are identified, we will share the protein names with the EPIC proteomics study PI, Marc Gunter for the 'look up' confirmation. As part of the collaboration, we will perform a look up for the proteins identified in EPIC. No data will be exchanged; only estimates.

As a second strategy for confirmation within ARIC, we will evaluate the association between genetic instruments (previously identified by Nilanjan Chatterjee's team [10]) for the top hit proteins (if exist) and cancer risk. In this analysis, we will additionally exclude participants RES_DNA = "No use/storage DNA".

e) Any anticipated methodologic limitations or challenges if present

We recognize that correction for multiple testing based on the total number of tests performed can be too conservative when high correlation exists among many of the proteins. Thus, we will consider a strategy that takes into account those high correlations when defining the threshold for correction for multiple testing. A less conservative strategy may be needed given the number of cancer cases in this single cohort study (as not to miss key proteins). We expect that this is reasonable, however, because we use an independent confirmation cohort.

Adjusting for the PEER factors can eliminate unmeasured confounding, for example, from assay factors by batch and the participant characteristics in a batch. However, PEER factor adjustment could also lead to overadjustment. We expect that it may be difficult Thus, we will present the findings without adjusting for the PEER factors and in a sensitivity analysis, after adjusting for the PEER factors.

Some of the plasma proteins that we identify may be marking exposures (e.g., smoking), including some that are known already. To be able to provide information that could lead to the discovery of new endogenous or exogenous factors that influence cancer risk, we will run models that are minimally adjusted for age, race*center, sex, and are adjusted for known risk factors (e.g., detailed smoking history).

Given that our research question about risk (not early detection), we will not be time updating the proteins in the analysis to minimize the likelihood that a nascent, but undetected cancer is influencing protein levels. However, we expect that a long time between protein measurement and cancer incidence may lead to an attenuation of associations, if plasma proteins are marking exposures that change over time (e.g., someone is thin at Visit 2 and then gains substantial weight subsequently). One strategy to minimize this problem is to model the mean of Visit 2 and Visit 3 protein levels. The time between these two visits is only 3 years, and for most cancers, we expect a longer time course for an influence on cancer risk with some exceptions (risk of lung cancer decreases with quitting albeit not back to never smoker status; risk of breast cancer drops precipitously after stopping post-menopausal hormone use likely because hormones are acting on early).

A contemporary issue is early onset cancer (<50 years old). ARIC has too few participants who are <50 years old at Visit 2 to perform this analysis.

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) . ☐ 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".

The SomaLogic data at Visit 2 is needed for this work given the goal of the analysis of discovering plasma proteins that are associated with common solid cancer incidence. Drs. Wang, Chatterjee and Platz, and Mr. Burk have already signed the SomaLogic form.

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)?

Many approved manuscript proposals indicate use of the SomaScan data or cancer data. Listed here are those that use both, and focus on cancer incidence or early detection:

#4144 Discovering proteins associated with malignant breast cancer incidence in post-menopausal women in the prospective ARIC study

#3676 Validation of candidate protein biomarkers identified in studies using genetic instruments for the directly measured levels in ARIC study with prostate cancer risk and prognosis

#3772 Validation of candidate protein biomarkers identified in studies using genetic instruments for the directly aptamer-measured levels in ARIC study with pancreatic cancer

#3909 Identification of plasma proteins for the early detection of cancer in ARIC

#3431 Proteomics, cigarette smoking, and mortality due to cigarette-smoking-related cancers

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 Proteomic longitudinal ARIC study: SOMAscan of multiple visits
2011.07 Enhancing ARIC Infrastructure to Yield a New Cancer Epidemiology Cohort
1995.04 Cancer Study

*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:

1. American Cancer Society. Cancer Facts & Figures 2024. In. Atlanta, GA: American Cancer Society; 2024.
2. Liu T, Joshu CE, Lu J, et al. External validation of genetically predicted protein biomarkers for pancreatic cancer risk using aptamer-based plasma levels: A prospective analysis in the Atherosclerosis Risk in Communities Study. *Int J Cancer.* 2023;153(6):1201-1216.
3. Zhu J, Shu X, Guo X, et al. Associations between Genetically Predicted Blood Protein Biomarkers and Pancreatic Cancer Risk. *Cancer Epidemiol Biomarkers Prev.* 2020;29(7):1501-1508.
4. Zhu J, Wu K, Liu S, et al. Proteome-wide association study and functional validation identify novel protein markers for pancreatic ductal adenocarcinoma. *Gigascience.* 2024;13.
5. Wu L, Shu X, Bao J, et al. Analysis of Over 140,000 European Descendants Identifies Genetically Predicted Blood Protein Biomarkers Associated with Prostate Cancer Risk. *Cancer Res.* 2019;79(18):4592-4598.
6. Wang S, Rao Z, Cao R, et al. Development and Characterization of Proteomic Aging Clocks in the Atherosclerosis Risk in Communities (ARIC) Study. *medRxiv.* 2023.
7. Gadd DA, Hillary RF, Kuncheva Z, et al. Blood protein assessment of leading incident diseases and mortality in the UK Biobank. *Nat Aging.* 2024;4(7):939-948.
8. World Cancer Research Fund / American Institute for Cancer Research. *Diet, Nutrition, Physical Activity, and Cancer: A Global Perspective.* Washington, DC2018.
9. US Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. *The Health Consequences of Smoking: 50 Years of Progress. A Report of the Surgeon General.* Atlanta, GA 2014.
10. Zhang J, Dutta D, Kottgen A, et al. Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nature genetics.* 2022;54(5):593-602.

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.