1.a. **Full Title**: Understanding the contribution of rare and common genetic variants on n-3 and n-6 fatty acid levels: A CHARGE Fatty Acid working group meta-analysis

b. **Abbreviated Title (Length 26 characters)**: PUFA and gene variants

2. **Writing Group**: numerous co-authors from contributing cohorts

   Writing group members: Nathan Tintle, Nathan Pankratz, Lyn M. Steffen, other contributing cohort members from CARDIA, CHS, FHS, MESA, METSIM, NHAPC, PIVUS, WGHS and WHI.

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

   **First author**: Nathan Tintle, PhD

   Address: Dordt College, Department of Mathematics, Computer Science and Statistics

   498 4th Ave NE, Sioux Center, IA 51250

   Phone: 712-722-6264       Fax:

   E-mail: nathan.tintle@dordt.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: Lyn Steffen

   Address: University of Minnesota School of Public Health

   Phone: 612-625-9307       Fax:

   E-mail: steffen@umn.edu

3. **Timeline**: 1 year

   August 2019 – July 2020

4. **Rationale**: Recent studies have identified numerous plausible causal loci associated with n-3 and n-6 fatty acid levels. For example, in a multi-ethnic meta-analysis, Hu et al. (2016) identified four novel loci (MYB, AGPAT4, DGAT2 and PPT2), while replicating other well-known loci (e.g., FADS1, ELOVL2, NTAN1, NRBF2 and GCKR) that have been previously identified. However, this and most previous studies (e.g., by Tintle et al. 2015, Kalsbeek et al. 2018, Lemaitre et al. 2011, among others) have focused primarily on common (MAF>5%) directly typed or imputed genetic variants. As is the case for many clinically relevant phenotypes, heritability estimates suggest a great deal of genetic variation in n-3 and n-6 fatty acid levels that is not fully explained by the loci identified to date (Harris et al. 2012). To potentially identify additional genetic loci associated
with n-3 and n-6 fatty acid levels we propose the following analytic plan to be coordinated via the CHARGE-fatty acids working group:

- Utilize cohorts for which Exome Chip data is available, along with plasma or RBC fatty acid measurements on n-3 and n-6 fatty acid levels.

- We will generally follow the approach taken in Peloso et al. (2014) and Lange et al. (2014) and use a combination of single-marker tests and gene-based tests on the approximately 250,000 measured markers from the ExomeChip in order to:

  a. Part 1. Conduct single marker analyses on the ~250,000 single markers (rare and common) available on the exome chip, adjusting for standard sets of covariates and with meta-analytic methods as conducted for prior meta-analysis CHARGE analyses with appropriate multiple testing adjustments

  b. Part 2. Conduct gene-based tests for all exonic regions. We will conduct both burden tests and covariate adjusted SKAT along with meta-analysis across cohorts using only variants with MAF<5%.

5. Main Hypothesis/Study Questions:

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).

Analytic details

Fatty acids We propose to investigate seven common, previously measured, n-3 and n-6 PUFA, with each evaluated as % total fatty acids of the following seven fatty acids: linoleic acid (18:2n6), arachidonic acid (20:4n6), adrenic acid (22:4n6), docosapentaenoic acid (22:5n3), alpha-linolenic acid (18:3n3), eicosapentaenoic acid (20:5n3; EPA), and docosahexaenoic acid (22:6n3; DHA) in plasma phospholipids and in erythrocyte membranes.

Covariates: Adjustment for potential confounders in model one will include age, sex, field center and other cohort-specific variables (e.g., family structure) as needed. Secondary models will include body mass index, lifestyle factors (e.g., smoking (current/former/never), physical activity (METS score or cohort specific variable—confirm with us first), and alcohol (current/former/never)), medications (e.g., lipid lowering drug, hypertensive drug, diabetes drug, and aspirin) and socioeconomic status (e.g., education—<HS, HS degree and College graduate). Analyses plans will be refined in conjunction with the CHARGE Inflammation Working Group, based on their expertise (Dehghan, 2009; Dehghan, 2011).

Population stratification: not applicable to ARIC, since only Caulcaisions had plasma fatty acids.

Part 1 specific analysis details

The R/seqMeta() package should be used for all analyses.

Genome-wide genotyping should be conducted across populations using standard calling algorithms and quality control (QC) checks. The following QC exclusions should be applied: subjects with call rates <95%, SNPs with call rates <95% or deviation from Hardy Weinberg expectations, excessive heterozygosity or high Mendelian error rates in family-based studies. We will only focus analyses on markers directly genotyped using the ExomeChip.
**Cohort-specific Analyses:** Prior to meta-analysis, analysts with population-specific expertise will implement jointly developed analysis as described in previous CHARGE projects. Linear regression models of the form: **Model 1:** \( y(fatty\ acid) = \alpha + \beta(SNP) + \beta(age) + \beta(sex) + \beta(field\ center) + \beta(cohort\ specifics) \) and **Model 2:** \( Model\ 1 + \beta(BMI) + \beta(lifestyle\ factors) + \beta(medications) + \beta(socioeconomics) \) will be used to assess direct SNP impact on fatty acid levels using an additive, allele dosage model (see covariate section for specifics). In these regressions, the SNP coefficient (SNP) represents the difference in fatty acid level per allele copy. From each analysis, we will record regression coefficients and standard errors for fatty acid terms (e.g., prepScores from seqMeta). This information is sufficient for basic and advanced meta-analysis models.

**Meta-analysis:** CHARGE was formed to facilitate meta-analysis and will provide the infrastructure for this project (Psaty, 2009). The \( \beta \)'s and SE (prepScores from seqMeta) recorded from each cohort are sufficient statistics for implementing meta-analyses. We will use R/seqMeta () to calculate summary effects using fixed effects meta-analysis, and combine regression coefficients using inverse variance weighting to produce a summary statistic. Heterogeneity will be assessed using the Cochran's Q-test. Trans-ethnic meta-analyses will be conducted with MANTRA software (Hong, 2016), as in our previous collaborative study (Hu, 2016).

**Multiple comparisons:** The large number of tests performed increases the probability of false positive results. To reduce type 1 error, we will conservatively set a threshold \( P \) value of 0.05/\( \approx 250000 \) (number of SNPs that vary) = 2x10\(^{-7} \) GWAS results for each of the 7 PUFA exposures, based on a Bonferroni correction for the ExomeChip.

**Part 2 specific analysis details**

**Genotyping of low frequency/rare protein coding variants:** Co-investigators for the current proposal have co-authored publications using rare variants genotyped in CHARGE cohorts (Peloso, 2014; Auer, 2015). Within CHARGE, ExomeChip genotyping was completed using the Illumina HumanExome BeadChip (ExomeChip). Genotypes were called using Gencall in Illumina’s Genome Studio software via the CHARGE Consortium joint calling effort and QC adhered to the CHARGE exome chip best practices (Grove, 2013). Based on the NHLBI Exome Sequencing Project design used for the ExomeChip, rare variants (MAF<1%) were captured (Wessel, 2015) but very rare variants excluded. However, the ExomeChip provides high-quality genotyping of the same variant set across studies, facilitating more consistent comparisons compared to whole genome sequencing.

**Annotation of variants for analyses:** Together with the CHARGE Consortium joint calling effort, CHARGE developed annotation of the ExomeChip variants including consortium-wide variant allele frequencies, race/ethnic-specific allele frequencies, and gene annotation for the ExomeChip variants. We will use the CHARGE annotation file (http://www.chargeconsortium.com/main/exomechip) to restrict analyses to rare/low-frequency variants with race/ethnic-specific MAF < 5% and cohort-specific MAF > 0. Variants will be grouped at the gene-level according to the CHARGE annotation file.

**Cohort-specific Analyses:** We will use the prepScores from the cohort specific analysis in part 1.

**Meta-analysis:** We will use R/seqMeta for rare variant meta-analysis. Primary analyses will be conducted using a rare variant-based test based on the SKAT framework (Wu, 2011). However, limitations of this approach include the lack of estimated effect sizes/directions, which are available in burden tests. We will address these issues in two ways. First, we will conduct a heterogeneity test for SKAT across studies using P-values. Second, in addition to SKAT, we will compute burden tests, providing additional information about the statistical significance.
observed in SKAT-based analyses, and facilitating an effect-based, heterogeneity analysis. Based on our previous experience using the R/seqMeta package, we are aware that computation can be extremely slow or infeasible for genes with very small cumulative minor allele count (cMAC). Therefore, we will filter on gene-level cMAC > 10.

**Multiple comparisons:** Statistical significance will be assessed using a Bonferroni correction for the number of gene-based tests (approximate significance threshold of \( P^* = 0.05 / 20,000 \) genes = 2.5 \( \times 10^{-6} \) for each of the 7 exposures (individual circulating PUFAs). For targeted analyses, the adjusted thresholds for Part 2 will be based on the number of gene-based tests.

**Data handling:**

Each cohort will conduct quality control on their specific cohort and use “recode_all.txt” from [http://www.chargeconsortium.com/main/exomechip](http://www.chargeconsortium.com/main/exomechip) to confirm all alleles are coded to the same strand.

Each group will run prepScores from the latest version of the seqMeta package and upload the output RData files. (Please include all data that passes QC including MAF > 0.05 and monomorphics since they will be used in single SNP analysis and for the MAF computations.)

Each group will upload ethnic specific prepscores (email us the proposed ethnic strata before beginning the analysis) for each fatty acid as a .Rdata file to CHARGE Fatty Acid working group google drive. These files should be named as follows Cohort_FattyAcid_date.Rdata ex (FHS_18_2N6_7.30.2018.Rdata)

Each cohort will also upload a readme file that includes fatty acid distribution information (mean, SD), any cohort specific relevant information, and contact information for lead analyst. These files should be named Cohort_readme.txt (ex. FHS_readme.txt)

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 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/mantrack/maintain/search/dtSearch.html](http://www.cscc.unc.edu/aric/mantrack/maintain/search/dtSearch.html)

_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)?

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* __Aaron Folcom fatty acids)

   ____  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 https://www2.cscc.unc.edu/aric/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 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.