ARIC Manuscript Proposal # 3620

PC Reviewed: 5/12/20 **Status: _____** **Priority: 2**

SC Reviewed: _____ **Status: _____** **Priority: ____**

1.a. Full Title: Functional mutations in cystic fibrosis transmembrane conductance regulator *(CFTR)* gene and colorectal cancer

We have approval for ancillary study 2018.22. This is the first manuscript proposal from that ancillary study.

b. Abbreviated Title: *CFTR* gene and colorectal cancer

2. Writing Group:

ARIC co-authors: Anna Prizment, Nathan Pankratz, Patricia Scott, Weihong Tang, Guillaume Onyeaghala, Timothy Starr, David Couper, Corinne Joshu, Elizabeth Platz (other ARIC researchers are welcome to participate).

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

First author: Anna Prizment

Address:
Anna Prizment, PhD, MPH
Associate professor
work phone: 612-301-1860
e-mail:prizm001@umn.edu
University of Minnesota

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: Nathan Pankratz

Address:

Phone:

E-mail: pankr018@umn.edu

3. Timeline: About a year from approval date

4. Rationale:

Background

Despite advances in screening, colorectal cancer (CRC) remains the second leading cause of cancer-related death in the U.S. Establishing new genetic risk factors for CRC holds promise for screening and may serve as an important step towards reducing CRC burden. Recently, our collaborators at the U of MN and other researchers have identified that cystic fibrosis transmembrane conductance regulator (*CFTR*) gene, which is mutated in cystic fibrosis (CF), is a tumor suppressor in CRC.[1-2]

CFTR gene is an ion channel gene on chromosome 7q31.2, expressed mainly in the lung and gastrointestinal (GI) tract.[3-4] In the GI tract, loss of *CFTR* activity leads to loss of integrity of the epithelial layer, inflammation, and activation of Wnt/Beta-catenin signaling – the fundamental pathway in CRC development.[5-6] To date, more than 2,000 mutations have been identified in the *CFTR* gene. Among them, ~336 have been classified as causing CF (https://www.cftr2.org/). The most frequent CF-causing variant is F508del, which makes up 70% of CF-causing alleles in CF patients and is encountered at a frequency of 1.2%-1.4% in the general population. Other CF-causing mutations are much rarer (< 0.04% in the general population), but together with F508Del, all CF-causing mutations are encountered in 4% of individuals of European ancestry.

Mouse studies of our U of MN collaborators (Drs. Scott and Starr) and other animal studies [1, 7-8] linked *CFTR* mutations with CRC risk. In addition, a human study using MarketScan claims data from all over the U.S. found that increased CRC burden extends to CF carriers, i.e. individuals with one copy of mutant *CFTR* but without CF disease. [9] That MarketScan study included insurance claims from ~20,000 CF carriers and 99,010 controls. Compared to non-carriers, CF

carrier status was associated with increased odds ratio (OR, 95% CI) for GI cancers (including colorectal, stomach and cancers of GI organs combined; pancreatic cancer was excluded): OR=1.44 (1.01-1.02). The authors validated this association in a study of CF carriers -- mothers of CF children (before the child CF diagnosis was ascertained) and matched controls, and showed that being a carrier was associated with increased OR for GI cancer: 2.50 (0.94-6.66).[9] To our knowledge, that was the first large human study that examined the CF carrier status and GI cancer. However, the MarketScan study has several limitations: the study population was young (most of them below 50 years of age) and had a small number of GI cancer cases (n=40 among CF carriers and 178 GI cancers in the total study), which precluded researchers from specifically examining the CRC risk associated with CF carrier status. Also, the study could have been biased by including CF carriers into the non-carrier group due to the non-uniform screening for CF [9].

Elucidating the potential association between *CFTR* mutational status and CRC is important because, if confirmed, the burden to society is large, as there are more than 10 million CF carriers in the U.S, and CRC is the third most common cancer in the country. Now, there is a need in the study that will identify the heterozygous CF carriers using genetic data and examine its association with CRC risk using sound epidemiologic design. We propose to examine this association in two studies with existing genetic data – ARIC and UK Biobank. The strengths of these studies is that they are prospective, have well ascertained outcomes, and genome-wide association study (GWAS) and whole exome sequencing (WES) were done irrespective of indication as part of the cohort protocol. ARIC has also conducted high-quality whole genomic sequencing (WGS) in 12,400 participants that allows assessing rare CF-related variants. We will

use WGS to ascertain rare CF-causing mutations. Currently, the WGS data have been processed for 8400 participants only; thus, we will ascertain the most frequent Fdel508 using GWAS data imputed to TOPMed that has become recently available for all ARIC participants. We will also examine this association in the UK Biobank study of 3,680 CRC cases among 500,000 individuals. The bioinformatician on this study – Dr. Nathan Pankratz – assessed F508del to determine study feasibility – in the ARIC study using WGS data (the prevalence of F508del is 1.4% in European Americans and 0.2% in African Americans) and in the UK Biobank using existing GWAS data imputed to Haplotype Reference Consortium (the prevalence of mutation is ~1.5%).

5. Main Hypothesis/Study Questions:

The goal of this study is to determine the contribution of functional germline *CFTR* variants to the risk of CRC in the ARIC study.

Hypothesis: Functional germline mutations in the *CFTR* gene are associated with increased CRC risk**.**

Specific aim 1. Determine if most prevalent CF-causing mutation in the *CFTR* gene – F508del – is associated with CRC incidence.

Specific aim 2. Determine if a burden score of rare (minor allele frequency < 2%) CF-causing mutations in the *CFTR* gene is associated with CRC incidence.

Specific aim 3. Determine if a weighted polyvariant risk score composed of single nucleotide polymorphisms (SNPs) in the regulatory regions of the *CFTR* gene that affect *CFTR* expression levels are associated with CRC incidence.

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).

Inclusions: The analytic sample will include participants free of cancer at baseline, who gave consent to participate in non-CVD research and genetic studies and have genomic data.

Exposure:

Ascertainment of *CFTR* genetic variants. In the ARIC study, for the analyses of F508Del and polyvariant genetic score, we will use existing GWAS data, and for burden score, high-quality WGS data. GWAS in the ARIC study was conducted using the Affymetrix Genome-Wide Human SNP Array 6.0 and imputed to the 1000 Genomes Phase 3 reference panel. For the ascertainment of F508Del, we will use GWAS data recently imputed to another reference panel – TOPMed, since F508Del could not be detected using GWAS imputed to 1000 Genome or WES in the ARIC. Deep WGS was conducted at the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) for ~12,000 participants. About 8,400 samples were already processed by TOPMed (funded by NLBI), and about additional 4000 samples, sequenced using funding from NHGRI, should be ready for analysis by the end of 2020. Of all the samples, 100% passed sample quality control (QC).

Illumina X Ten technology, with PCR-free library construction (Illumina TruSeq), 30-40X coverage in 2x150 bp paired-end reads and including study trio and HapMap controls for QC. Data harmonization was conducted by TOPMed at the Informatics Resource Center (IRC) at the University of Michigan. The IRC made genotype calls for SNV and short indels for samples from all TOPMed Program studies to ensure consistency and rigor for all QC and quality assurance procedures. Dr. Pankratz obtained the TOPMed WGS data in VCF format from the

TOPMed Exchange Area on dbGaP. Population allele frequencies will be annotated using ANNOVAR and the effect of coding variants using the RefSeq and UCSC gene sets, which will delineate putative changes in the amino acid sequence, interference with a splice site junction, the creation or removal of a stop-codon, and structural interaction variants. Variants will be annotated based on their computationally predicted deleteriousness using information from dbNSFP. All variants will be annotated for their frequency and presence in multiple variant collections (e.g., dbSNP, 1000 Genomes, and Human Gene Mutation Database [HGMD]). Dr. Pankratz will conduct quality assurance before data analysis by considering the following criteria at the variant level: monomorphic sites, missing rate, mappability score, mean depth of coverage for all populations in the studies, allelic ratios, and genotype quality (GQ) scores. Sample-level QC metrics include missingness, possible contamination, outliers for mean depth, singleton count, heterozygote to homozygote ratio, and Ti/Tv ratio.

***CFTR* genetic variants.** Because all CF-causing variants, but F508del, are rare, we will only be able to study F508del individually. All other CF-causing mutations in the protein-coding region of the *CFTR* gene [i.e., non-synonymous polymorphisms with minor allele frequency <1% and loss of function mutations including frameshift, stop-gain, stop-loss, and splice variants; (https://www.cftr2.org/)] will be combined into a single variable – burden score to improve power and avoid multiple testing.

Burden scores will be created using group-based collapsing methods that have been developed and successfully implemented for the analysis of rare variants. Two burden scores will be constructed (**1**) All CF-causing mutations, i.e. the set of 336 CF-causing mutations in the *CFTR* gene (https://www.cftr2.org/) which includes a subset of 23 mutations used for screening

in the US. **(2)** Loss of function (LOF) mutations including frameshift, stop-gain, stop-loss, and splice variants as defined by TOPMed. The calculation of each burden score will be conducted by counting the number of minor alleles [i.e., CF causing alleles] for all polymorphic variants in the *CFTR* gene. In the absence of information on any meaningful/proven weights, we will assume that all rare missense variants have the same effect.

Polyvariant risk score will be generated by combining eight *CFTR* SNPs affecting expression as identified in Kerschner et al.,[10] which showed that regulatory *CFTR* variants are associated with *CFTR* expression.[10-11] SNPs will be presented as continuous variables and analyzed using an additive model. For each individual, this score will be computed by multiplying the minor allele count for a given variant by the effect estimate for that variant derived from Kerschner's study,[10] i.e., this score will be weighted and alleles with large effects will have a larger contribution to the final genetic score.

Table 2. Regulatory variants in CFTR that repress CFTR promoter and enhancer activity. MAF - minor allele frequency. Adapted from Kerschner, 2018

Regulatory region	Variant	dbSNP138	MAF* in 1000 Genomes	MAF* in CF patients	Promoter and enhancer activity (relative to intact element)
2kb promoter	intact 2kb promoter	----	----	----	1.00 (reference)
	c.-887C>T	rs34465975	0.164	1/160	0.77
	c.-812T>G	rs181008242	0.0012	3/160	0.48
	c.-410G>C	N/A	N/A	1/160	0.55
	c.-8G>C	rs1800501	0.0274	4/160	0.55
Intron 11 Intestinal enhancer	intact enhancer element	----	----	----	1.00 (reference)
	c.1679+566G>T	11 novel.1	N/A	2/160	0.50
	c.1679+1280G>A	rs213963	0.4265	95/160	0.63
	c.1679+1449A>G	rs13964	0.4263	95/160	0.37
	c.1679+1539T>C	11 novel.2	N/A	2/160	0.59

Outcomes: CRC incidence and mortality was ascertained from 1987 through 2015 using state cancer registries in Minnesota, North Carolina, Maryland, and Mississippi, and supplemented by abstraction of medical records and hospital discharge summaries.[12-13] A total of 435 incident CRC cases were ascertained over a maximum follow-up of 29.1 years.

Statistical analysis: Cox proportional hazards regression will be used to calculate hazard ratios and 95% confidence intervals to estimate associations of *CFTR* genetic variants (F508del and aggregate scores) with CRC risk. The proportional hazards assumption will be tested by including an interaction term between each measure and follow-up time in the Cox model. Person-years will be estimated from the baseline until the date of CRC diagnosis, death, loss to follow-up, or administrative censoring on December 31, 2015, whichever occurs first. The analyses will be stratified by age and sex although the power will be limited. We will also conduct sub-analyses restricted to Whites

Burden scores will be categorical variables and presented as (1) yes versus no CF-causing mutations and as (2) three categories: no mutation, 1 mutation and 2+ mutations versus no mutations (if there is a sufficient number of people with 2+ CF-causing mutations).

Polyvariant score will be weighted using beta-estimates (obtained from the associations between regulatory SNPs and *CFTR* expression[10] as described in the section "Exposure"); these beta estimates will serve as weights. In the Cox model, a linear relationship between polyvariant score and CRC risk will be tested using splines. If the linearity assumption is violated, additional analyses will be conducted by transforming the continuous variable or via categorizing polyvariant score into tertiles/quartiles.

All the analyses will be adjusted for age, sex, center, and PC ancestry (2 for individuals of European Ancestry and 4 for AAs). Because all the genetic variants are putatively functional and the power of our study is limited, we will combine the data from African Americans and Whites either using strata function in the proportional hazards regression, which includes individual hazard function for each race, or by conducting inverse variance weighted meta-analysis. Analyses will be conducted using SAS (version 9.4) or R (version 3.4.3).

Sensitivity analyses. We will conduct a sensitivity analysis to test for indication bias via two ways since CF careers could have other complications and could have used health provider more often. First, we will adjust for colonoscopy among ARIC participants who have this information. Second, we will test if the association changes after excluding earlier stage at diagnosis because indication bias would have affected early stage in the first turn. We will also test CRC mortality associated with CFTR mutations but the sample size will be even more limited.

Power calculation:

For F508del (frequency of minor allele =1.4%, prevalence of paricipants with this allele = 2.8%), assuming ~400 CRC cases among 14,000 individuals in the ARIC study, we will have 80% power (α=0.05, 2-sided) to detect an HR=2.3 (α=0.05, 2-sided).

For all CF-causing alleles combined (frequency of all minor allele = 4%, prevalence of people = 8%), assuming ~360 CRC cases among 12,400 individuals in the ARIC study, we will have 80% power to detect an HR=1.8 (α=0.05, 2-sided).

We will run parallel analyses in the ARIC and UK Biobank studies, evaluate consistency/heterogeneity between the two studies, and will meta-analyze the data if these is heterogeneity

Note: The proposal received funding from the Academic Health Center (UMN), as a faculty development grant.

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

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

(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

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

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 **Yes**

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)?

2766 Whole genome sequence analysis of heart, lung, blood, sleep and aging risk factors and disease endpoints. An omnibus manuscript proposal for CHARGE, CCDG and TOPMed

3103 Whole genome analysis of venous thromboembolism (VTE) (TOPMed)

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 ()

2018.22 Cystic fibrosis transmembrane conductance regulator (CFTR) gene and colorectal cancer.

2011.07 Enhancing ARIC Infrastructure to Yield a New Cancer Epidemiology Cohort

2012. 10 Whole-genome sequencing in ARIC

1995.04 Cancer Study

*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.

References

1. Than, B. L.; Linnekamp, J. F.; Starr, T. K.; Largaespada, D. A.; Rod, A.; Zhang, Y.; Bruner, V.; Abrahante, J.; Schumann, A.; Luczak, T.; Niemczyk, A.; O'Sullivan, M. G.; Medema, J. P.; Fijneman, R. J.; Meijer, G. A.; Van den Broek, E.; Hodges, C. A.; Scott, P. M.; Vermeulen, L.; Cormier, R. T., CFTR is a tumor suppressor gene in murine and human intestinal cancer. *Oncogene* **2016,** *35* (32), 4179-87.
2. Billings, J. L.; Dunitz, J. M.; McAllister, S.; Herzog, T.; Bobr, A.; Khoruts, A., Early colon screening of adult patients with cystic fibrosis reveals high incidence of adenomatous colon polyps. *Journal of clinical gastroenterology* **2014,** *48* (9), e85-e88.
3. Gelfond, D.; Borowitz, D., Gastrointestinal complications of cystic fibrosis. *Clin Gastroenterol Hepatol* **2013,** *11* (4), 333-42; quiz e30-1.
4. De Lisle, R. C.; Borowitz, D., The cystic fibrosis intestine. *Cold Spring Harb Perspect Med* **2013,** *3* (9), a009753.
5. Strubberg, A. M.; Liu, J.; Walker, N. M.; Stefanski, C. D.; MacLeod, R. J.; Magness, S. T.; Clarke, L. L., Cftr Modulates Wnt/beta-Catenin Signaling and Stem Cell Proliferation in Murine Intestine. *Cell Mol Gastroenterol Hepatol* **2018,** *5* (3), 253-271.
6. Zhang, J. T.; Wang, Y.; Chen, J. J.; Zhang, X. H.; Dong, J. D.; Tsang, L. L.; Huang, X. R.; Cai, Z.; Lan, H. Y.; Jiang, X. H.; Chan, H. C., Defective CFTR leads to aberrant beta-catenin activation and kidney fibrosis. *Sci Rep* **2017,** *7* (1), 5233.
7. Yamada, A.; Komaki, Y.; Komaki, F.; Micic, D.; Zullow, S.; Sakuraba, A., Risk of gastrointestinal cancers in patients with cystic fibrosis: a systematic review and meta-analysis. *The Lancet Oncology* **2018,** *19* (6), 758-767.
8. Starr, T. K.; Allaei, R.; Silverstein, K. A.; Staggs, R. A.; Sarver, A. L.; Bergemann, T. L.; Gupta, M.; O'Sullivan, M. G.; Matise, I.; Dupuy, A. J.; Collier, L. S.; Powers, S.; Oberg, A. L.; Asmann, Y. W.; Thibodeau, S. N.; Tessarollo, L.; Copeland, N. G.; Jenkins, N. A.; Cormier, R. T.; Largaespada, D. A., A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science (New York, N.Y.)* **2009,** *323* (5922), 1747-50.
9. Miller, A. C.; Comellas, A. P.; Hornick, D. B.; Stoltz, D. A.; Cavanaugh, J. E.; Gerke, A. K.; Welsh, M. J.; Zabner, J.; Polgreen, P. M., Cystic fibrosis carriers are at increased risk for a wide range of cystic fibrosis-related conditions. *Proceedings of the National Academy of Sciences* **2020,** *117* (3), 1621-1627.
10. Kerschner, J. L.; Ghosh, S.; Paranjapye, A.; Cosme, W. R.; Audrézet, M.-P.; Nakakuki, M.; Ishiguro, H.; Férec, C.; Rommens, J.; Harris, A., Screening for regulatory variants in 460 kb Encompassing the CFTR locus in cystic fibrosis patients. *The Journal of Molecular Diagnostics* **2019,** *21* (1), 70-80.
11. Giordano, S.; Amato, F.; Elce, A.; Monti, M.; Iannone, C.; Pucci, P.; Seia, M.; Angioni, A.; Zarrilli, F.; Castaldo, G., Molecular and functional analysis of the large 5′ promoter region of CFTR gene revealed pathogenic mutations in CF and CFTR-related disorders. *The Journal of Molecular Diagnostics* **2013,** *15* (3), 331-340.
12. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *American Journal of Epidemiology* **1989,** *129* (4), 687-702.
13. Joshu, C. E.; Barber, J. R.; Coresh, J.; Couper, D. J.; Mosley, T. H.; Vitolins, M. Z.; Butler, K. R.; Nelson, H. H.; Prizment, A. E.; Selvin, E., Enhancing the Infrastructure of the Atherosclerosis Risk in Communities (ARIC) Study for Cancer Epidemiology Research: ARIC Cancer. *Cancer Epidemiology and Prevention Biomarkers* **2017**, cebp. 0696.2017.