Genetic and environmental associations with disease risk and drug response in Alaska Native people, and a responsive justice approach to reconciling statistical and ethical research demands.

Alison E. Fohner

A dissertation submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Kenneth Thummel, Chair

Timothy Thornton

Kelly Edwards

Program Authorized to Offer Degree:

Public Health Genetics

University of Washington

Abstract

Genetic and environmental associations with disease risk and drug response in Alaska Native people, and a responsive justice approach to reconciling statistical and ethical research demands.

Alison E. Fohner

Chair of the Supervisory Committee:
Kenneth E. Thummel, Professor and Chair
Department of Pharmaceutics

Genetic research with diverse and underserved communities is important for expanding the benefits of genetic discoveries and their application to personalized medicine. In partnership with Alaska Native communities, we identified and characterized novel and known variation in *CYP2C9, VKORC1, CYP4F2, CYP4F11,* and *GGCX,* 5 genes known to affect warfarin disposition and response, in 350 Yup'ik people and 365 customer-owners of Southcentral Foundation. Through resequencing and targeted genotyping, we identified two common novel variants *M1L* and *N218I* in *CYP2C9* and high frequencies of the *VKORC1* haplotype (-1639G>A and 1173C>T) that are expected to result in increased warfarin

sensitivity. We also observed high frequencies of *CYP4F2*3*, which may increase vitamin K conservation and necessitate a higher warfarin dose to achieve the desired anticoagulation effect. Individual patient needs will depend on a complex mixture of genetic traits.

Because of seasonal variation in sunlight exposure, vitamin D deficiency is a concern for people living in circumpolar regions. We characterized 25(OH)D_3 concentrations in 743 Yup'ik people living in the Yukon Kuskokwim delta of southwestern Alaska and identified sources of inter-individual variability, including season of sample collection, genetic variation in *CYP2R1* and *DHCR7*, age, gender, body mass index, the degree of consumption of the traditional diet, and inland or coastal geography of the community. Yup'ik participants on average had adequate concentrations of 25(OH)D_3 (31.1 +/- 0.9 ng/mL), but a younger age (< 33 years) was significantly associated with lower concentration of 25(OH)D_3 (24.5 ng/mL compared to 37.5 ng/mL for ages $\geq$ 33 years), lower levels of a biomarker of traditional food consumption, and greater fluctuation with changes in sunlight exposure. Younger Yup'ik participants may be at increased risk for adverse outcomes associated with vitamin D deficiency, especially during seasons of low sunlight exposure.

Finally, in genetic epidemiology research, genome-wide markers or pedigree information is used to adjust for population substructure and prevent confounding of results. Population substructure and the concerns of communities with respect to the methods used to adjust for it are described. A responsive justice framework is suggested as a tool to approach these conflicting demands of research, presenting as an example the stratification of participants by self-identified ancestral language group to approximate the statistical adjustments needed for population stratification.

Acknowledgments:

I thank the customer-owners at Southcentral Foundation and the Yup'ik people of the Yukon Kuskokwim Delta for their partnership and participation in these research studies. I also thank collaborating scientists at University of Washington (Kenneth Thummel, Timothy Thornton, Kelly Edwards, Barbara Kavanaugh, Jesse Tsai, Patricia Janssen, Zhican Wang), University of Alaska Fairbanks and the Center for Alaska Native Health Research (Bert Boyer, Scarlett Hopkins, Jynene Black, Diane O'Brien), Southcentral Foundation (Denise Dillard, Renee Robinson, Burhan Khan), and the Yukon Kuskokwim Health Corporation (Joseph Klejka), who all contributed greatly to this research. I acknowledge financial support from an Achievement Rewards for College Scientist fellowship and the Northwest – Alaska Pharmacogenomics Research Network grant (National Institute of General Medical Sciences U01 GM092676-05).

Table of Contents

CHAPTER 1: INTRODUCTION

1.1 Introduction

Genetic technology can increase understanding of variability in medical outcomes and improve care for populations and individuals. For all people to benefit from medical advances in personalized medicine resulting from genetic research, however, more people with diverse ancestries must be included in discovery research and application development. Historically, European populations have been oversampled in genetic research. This selection bias means that genetic variation that has been found in European populations is the focus of further research and that any resulting medical advances are thus preferentially beneficial for people of European ancestry. Because patterns of genetic variation differ with ancestry, people of non-European descent may not benefit in the same way from this research. In fact, many common complex traits and diseases are associated with genetic variants that occur at different frequencies in different populations, meaning that biased sampling leads to skewed research and unequal health benefits for people of non-European ancestries [1].

Small identifiable groups, such as indigenous communities, may be hesitant to participate in the genetic research that would inform these medical advances because of distrust resulting from a history of abuses at the hands of the government and researchers [2] and because of the potential for harms resulting from research. By not participating in research, community members are protected from potential harms, but also are prevented from obtaining equitable benefit from medical genetics, thereby widening a gap of health disparities. Until 2008, only 12 studies in the previous 30 years had included indigenous people of the Americas in pharmacogenetic research, and none of these populations were in the United States [1, 3]. As a consequence, working with these communities to expand genetic research is important for reducing health disparities.

The Northwest-Alaska Pharmacogenomics Research Network (NWA-PGRN) is expanding genetic research to include historically marginalized populations through Community-based Participatory Research (CBPR) [2, 4] in partnership with Southcentral Foundation (SCF) in Anchorage, AK, the Center for Alaska Native Health Research (CANHR) in Fairbanks, AK, the Yukon Kuskokwim Health Corporation (YKHC) in Bethel, AK, the American Indian and Alaska Native (AI/AN) people of southcentral Alaska, the Yup'ik people of the Yukon-Kuskokwim Delta (Y-K Delta), the Confederated Salish and Kootenai Tribes of western Montana, and the University of Montana. Alaska Native communities historically have been excluded from medical research, and geographic isolation has led to substantial portions of AI/AN people in Alaska being medically under-served and having considerable health disparities compared to other populations [5].

SCF is a tribally owned and operated regional health corporation, providing pre-paid healthcare services to 58,000 AI/AN patients, who are considered "customer-owners" of SCF. The Anchorage Service Unit served by SCF is comprised of both urban and rural areas, including Anchorage, the Matanuska-Susitna Borough, and 60 outlying communities (most with fewer than 500 residents). It provides primary care services to ~ 55% of the total AN population at 6 SCF primary care clinics on the Alaska Native Medical Center (ANMC) campus. Tertiary care is provided at the 150-bed ANMC hospital, which is co-owned and co-managed by SCF and Alaska Native Tribal Health Consortium (ANTHC). CANHR is based at the University of Alaska Fairbanks and has ongoing genetic research partnerships with 11 of the 58 rural communities in the Yukon-Kuskokwim River Delta (Y-

K Delta) that are served by the YKHC. The YKHC provides healthcare to about 23,000 Yup'ik people. To date, there has been no research conducted with the indigenous populations of Alaska to characterize variation in genes associated with interindividual differences in drug response – so called "pharmacogenes".

Effects of genetic variation in pharmacogenes on drug disposition and response have been identified and are being applied to clinical care. The most notable drugs are those with a small therapeutic window between efficacy and toxicity, and whose metabolism are dominated by enzymes with variable catalytic efficiency resulting from genetic variation. The enzymes of the Cytochrome P450 superfamily, which dominate the clearance of many drugs, are highly polymorphic and patterns of variation differ between global populations [3, 6]. Genetic polymorphisms that alter enzyme function lead to "super-metabolizers", "intermediate metabolizers", and "slow metabolizers" based on the catalytic efficiency of an enzyme, which can lead to differences in the efficacy and toxicity of the same drug dose between individuals.

Examples of genetic variation in pharmacogenes already being implemented in standard medical care include the effects of genetic polymorphisms in *thiopurine S-methyltransferase* (*TPMT*) and *cytochrome P450 2C19* (*CYP2C19*) on the metabolism and dosing of 6-mercaptopurine (6MP) and clopidogrel, respectively. 6MP is an immunosuppressive drug used to treat acute lymphocytic leukemia and inflammatory bowel disease and acts by interfering with DNA replication in cell division. The TPMT protein methylates 6MP, inactivating the drug, and, as a result, genetic variation in *TPMT* can lead to potentially life-threatening toxicities as the drug is cleared more slowly and circulating concentrations increase [7]. For developing the toxicity of leucopenia with standard 6MP treatment, heterozygosity for a reduced-function variant is associated with an odds ratio of 4.29 and homozygosity for a reduced-function variant is associated with an odds ratio of 20.84 compared to homozygous wildtype [7]. As a result, since 2004 the FDA has recommended genetic testing prior to initiation of therapy, and it has become routine medical practice to do so in Europe [8] and at many US cancer treatment centers [9].

Clopidogrel is another example of a drug affected by genetic variation. It is used during coronary surgeries and to treat acute coronary syndrome [10]. While 6MP is inactivated by TPMT, clopidogrel is a prodrug, activated by CYP2C19. Variation in *CYP2C19* has been associated with altered drug response, leading to the 2010 FDA "Black Box" warning; individuals with reduced enzyme function at increased risk for adverse cardiovascular events because their concentration of active drug are below therapeutic concentrations [11]. The most common reduced-function variant, *CYP2C19*2*, is found at high, but variable frequencies across populations, ranging from 15% in Caucasians to 35% in Asians, and results in a hazard ratio of 1.55 for heterozygotes and 1.76 for homozygotes for failed treatment and thrombotic events [12]. The *CYP2C19* gene has over 25 variants, including some leading to increased enzymatic activity (**17*). Alternative treatment can be given to patients with reduced CYP2C19 activity, making genetic testing clinically useful prior to initiating therapy.

Based on the importance of genetic variation in altering drug response, NWA-PGRN, SCF, CANHR, and their partnering communities developed pharmacogenetic research questions to address community health concerns as defined in part by the community members themselves. In CBPR, communities are involved in all steps of the research process, as a transformed practice of respect for persons and of developing long term research partnerships between communities and researchers [2]. At SCF, the questions developed with the community include genetic variation affecting appropriate dose for warfarin treatment for prevention of thromboembolic events and for tamoxifen treatment for breast cancer. At CANHR, these questions include variation in genes affecting cancer risk and treatment, genetic variation affecting warfarin dose and response, and factors affecting vitamin D concentrations in the communities.

In Chapter 2, I present a study of genetic variation affecting warfarin disposition and response in Alaska Native people. Anecdotal evidence suggests that the average warfarin dose needed for Alaska Native people to achieve a stable, target anticoagulation effect (eg, INR = 2-3) is lower than what is needed on average in other populations [13]. Warfarin is prescribed commonly to prevent thromboembolic events, but too high of a dose can lead to dangerous bleeds [14-17]. The 5 genes studied were *CYP2C9, VKORC1, CYP4F2, CYP4F11,* and *GGCX,* which have been associated with warfarin activity and the vitamin K–dependent clotting cascade [18]. Through resequencing, this study identified common and rare candidate gene variation in Alaska Native populations and then determined frequencies of the most prevalent variants thought to affect associated enzyme activities and warfarin anticoagulation response. Better understanding of genetic variation in the Alaska Native population may inform more appropriate warfarin dosing and reduce the risk of adverse bleeding and thrombotic events.

In Chapter 3, I present a study characterizing the associations of genetics, diet, demographics, and seasonal sunlight variation with serum 25-hydroxy vitamin D [25(OH)D] concentrations in an opportunistic sample of Yup'ik people living in the Y-K Delta. Vitamin D deficiency is a concern in Alaska because of the seasonal changes in sunlight exposure [19]. Vitamin D deficiency has been associated with increased risk for many diseases, including rickets and colon cancer [20-24]. Additionally, vitamin D levels affect expression of intestinal cytochrome P450 3A4 (CYP3A4) [25, 26], an enzyme involved in the first-pass metabolism of some drugs, including tacrolimus [27], tamoxifen [28], midazolam [25] and certain statins [29-33]. In many populations, over 90% of vitamin D that the body receives is synthesized in the skin from UVB radiation, with smaller amounts obtained through the diet [34], but it is thought that Yup'ik people historically maintained sufficient concentrations of $25(OH)D_3$ through a diet high in vitamin D_3 [35]. Dietary patterns are changing, however, and may expose the population to greater risk of vitamin D deficiency and related diseases [19, 36], as well as variability in drug disposition and response. Characterizing 25(OH)D concentrations and the factors affecting them in the Yup'ik population (and other AI/AN populations in Alaska) is important for understanding disease risk associated with vitamin D deficiency and variability in drug response.

In chapter 4, I discuss a challenge to conducting genetic research with relatively small, historically underserved populations and analyze a possible approach to addressing these concerns using a responsive justice framework. Population substructure of study participants can confound research results if genetic and phenotypic patterns differ between cases and controls [37, 38]. I present the scientific importance of adjusting for population substructure in statistical analyses and the ethical challenges of doing so when working with historically underserved and marginalized populations. I then analyze the use of ancestral linguistic relationships of subpopulations as a possible tool for reconciling these conflicting scientific and ethical demands of population genetic research. This type of approach is an example of how responsive justice can be used when working with communities in genetic research to maximize scientific validity of results without compromising the ethical responsibility of researchers. By including communities in research conversations and decisions up front, more of them may engage in research and enable the expansion of benefits resulting from genetic research.

In summary, I address genetic research with historically underserved communities in 3 ways: 1) by presenting an example of how genetic variation could affect warfarin anticoagulation response in Alaska Native people; 2) by presenting genetic and environmental associations with $25(OH)D_3$ concentrations in the Yup'ik people and how changing dietary patterns could increase the risk of vitamin D deficiency and instability in this population; and 3) by analyzing ancestral linguistic relationships of subpopulations as an example in applying responsive justice to expand genetic research to include more underserved populations.

1.2 References

1. Goering, S., S. Holland, and K. Fryer-Edwards, *Transforming Genetic Research Practices with Marginalized Communities: A case for representative justice.* Hastings Center Report, 2008. **38**(2): p. 43-58.
2. Boyer, B., et al., *Ethical issues in developing pharmacogenetic research partnerships with American Indigenous communities.* Clinical pharmacology and therapeutics, 2011. **89**(3): p. 343-5.
3. Jaja, C., et al., *Cytochrome p450 enzyme polymorphism frequency in indigenous and native american populations: a systematic review.* Community genetics, 2008. **11**(3): p. 141-9.
4. Boyer, B.B., et al., *Building a community-based participatory research center to investigate obesity and diabetes in Alaska Natives.* International Journal of Circumpolar Health International Journal of Circumpolar Health, 2005. **64**(3).
5. Alaska Department of Labor and Workforce Development, R.a.A.S., *Alaska Population Overivew: 2012 Estimates*, 2013. p. 128.
6. Polimanti, R., et al., *Human genetic variation of CYP450 superfamily: analysis of functional diversity in worldwide populations.* Pharmacogenomics, 2012. **13**(16): p. 1951-60.
7. Lennard, L., *Implementation of TPMT testing.* Br J Clin Pharmacol, 2014. **77**(4): p. 704-14.
8. Paugh, S.W., et al., *Cancer pharmacogenomics.* Clin Pharmacol Ther, 2011. **90**(3): p. 461-6.
9. Hoffman, J.M., et al., *PG4KDS: a model for the clinical implementation of pre-emptive pharmacogenetics.* Am J Med Genet C Semin Med Genet, 2014. **166C**(1): p. 45-55.
10. Levitt, M.R., J.W. Osbun, and L.J. Kim, *The Parmacogenomics of Clopidogrel.* World Neurosurgery News, 2012. **77**(3/4): p. 402-407.
11. Ford, N.F. and D. Taubert, *Clopidogrel, CYP2C19, and a Black Box.* J Clin Pharmacol, 2013. **53**(3): p. 241-8.
12. Scott, S.A., et al., *Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update.* Clin Pharmacol Ther, 2013. **94**(3): p. 317-23.
13. Schilling, B. *Anticoagulation Care for Alaska Native Customer-Owners within the Nuka Model of Care.* in *7th Dawn AC Anticoagulation Management Software North American User Group Meeting*. 2013. La Jolla, CA.
14. Stafford, R.S. and D.E. Singer, *Recent national patterns of warfarin use in atrial fibrillation.* Circulation, 1998. **97**(13): p. 1231-1233.
15. Birman-Deych, E., et al., *Use and effectiveness of warfarin in Medicare beneficiaries with atrial fibrillation.* Stroke, 2006. **37**(4): p. 1070-1074.
16. Shapiro, S.S., *Treating thrombosis in the 21st century.* N Engl J Med, 2003. **349**(18): p. 1762-1764.
17. Hart, R.G., L.A. Pearce, and M.I. Aguilar, *Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation.* Ann Intern Med, 2007. **146**(12): p. 857-867.

18. McDonagh, E., et al., *From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource.* Biomarkers in medicine, 2011. **5**(6): p. 795-806.

19. Gessner, B.D., J. Plotnik, and P.T. Muth, *25-Hydroxyvitamin D levels among healthy children in Alaska.* The Journal of Pediatrics, 2003. **143**(4): p. 434-437.

20. Ahn, J., et al., *Genome-wide association study of circulating vitamin D levels.* Human Molecular Genetics, 2010. **19**(13): p. 2739-2745.

21. Ramos-Lopez, E., et al., *CYP2R1 (vitamin D 25-hydroxylase) gene is associated with susceptibility to type 1 diabetes and vitamin D levels in Germans.* Diabetes/Metabolism Research and Reviews, 2007. **23**: p. 631-636.

22. Levin, G.P., et al., *Genetic variants and associations of 25-hyroxyvitamin D concentrations with major clinical outcomes.* JAMA, 2012. **308**(18): p. 1898-1905.

23. Zhang, Z., et al., *An analysis of the association between the vitamin D pathway and serum 25-hydroxyvitamin D levels in a healthy Chinese population.* J Bone Miner Res, 2013. **28**(8): p. 1784-92.

24. Sharma, S., et al., *Vitamin D deficiency and disease risk among aboriginal Arctic populations.* Nutr Rev, 2011. **69**(8): p. 468-78.

25. Thirumaran, R.K., et al., *Intestinal CYP3A4 and midazolam disposition in vivo associate with VDR polymorphisms and show seasonal variation.* Biochem Pharmacol, 2012. **84**(1): p. 104-12.

26. Thummel, K.E., et al., *Transcriptional control of intestinal cytochrome P-4503A by 1alpha,25-dihydroxy vitamin D3.* Molecular pharmacology, 2001. **60**(6): p. 1399-406.

27. Lindh, J.D., et al., *Seasonal variation in blood drug concentrations and a potential relationship to vitamin D.* Drug Metab Dispos, 2011. **39**(5): p. 933-7.

28. Teft, W.A., et al., *CYP3A4 and seasonal variation in vitamin D status in addition to CYP2D6 contribute to therapeutic endoxifen level during tamoxifen therapy.* Breast Cancer Res Treat, 2013. **139**(1): p. 95-105.

29. Robien, K., et al., *Drug-vitamin D interactions: a systematic review of the literature.* Nutr Clin Pract, 2013. **28**(2): p. 194-208.

30. Schwartz, J.B., *Effects of vitamin D supplementation in atorvastatin-treated patients: a new drug interaction with an unexpected consequence.* Clin Pharmacol Ther, 2009. **85**(2): p. 198-203.

31. Bhattacharyya, S., K. Bhattacharyya, and A. Maitra, *Possible mechanisms of interaction between statins and vitamin D.* QJM, 2012. **105**(5): p. 487-91.

32. Elens, L., et al., *Novel CYP3A4 intron 6 single nucleotide polymorphism is associated with simvastatin-mediated cholesterol reduction in the Rotterdam Study.* Pharmacogenet Genomics, 2011. **21**(12): p. 861-6.

33. Gao, Y., L.R. Zhang, and Q. Fu, *CYP3A4*1G polymorphism is associated with lipid-lowering efficacy of atorvastatin but not of simvastatin.* Eur J Clin Pharmacol, 2008. **64**(9): p. 877-82.

34. Luick, B., A. Bersamin, and J.S. Stern, *Locally harvested foods support serum 25-hydroxyvitamin D sufficiency in an indigenous population of Western Alaska.* Int J Circumpolar Health, 2014. **73**.

35. Lehmann, B. and M. Meurer, *Vitamin D metabolism.* Dermatologic Therapy, 2010. **23**: p. 2-12.

36. Singleton, R., et al., *Rickets and vitamin D deficiency in Alaska native children.* Journal of Pediatric Endocrinology and Metabolism, 2015. **0**(0).

37. Hoyos-Giraldo, L.S., et al., *The effect of genetic admixture in an association study: genetic polymorphisms and chromosome aberrations in a Colombian population exposed to organic solvents.* Ann Hum Genet, 2013. **77**(4): p. 308-20.

38. Wacholder, S., N. Rothman, and N. Caporaso, *Counterpoint: Bias from Population Stratification Is Not a Major Threat to the Validity of Conclusions from Epidemiological Studies of Common Polymorphisms and Cancer.* Cancer Epidemiology, Biomarkers, & Prevention, 2002. **11**(513-520).

CHAPTER 2: VARIATION IN GENES CONTROLLING WARFARIN DISPOSITION AND RESPONSE IN AMERICAN INDIAN AND ALASKA NATIVE PEOPLE: *CYP2C9, VKORC1, CYP4F2, CYP4F11, GGCX*

This work was published previously as Fohner, et al. *Pharmacogenetics and Genomics.* 2015.

2.1 Introduction

Genetic variation affects the pharmacokinetics and pharmacodynamics of medical drugs by altering enzyme and transporter expression and function, resulting in differences in drug efficacy and safety between individuals [1-5]. The prevalence and frequency of genetic variation affecting pharmacologic response is diverse across racial and ethnic populations [6-8]. As a result, populations that are rarely included in medical research, such as American Indian and Alaska Native (AI/AN) people, are less likely to benefit from genome-based, personalized, drug therapy [6-8]. Thus, pharmacogenetic research in these understudied populations is needed to improve both the selection of drugs and their dosages, and to reduce the number of adverse effects [6-8].

The incidence of stroke is disproportionately high in AN communities, with AN people more likely to be affected by stroke at an earlier age than other populations [9-11] and death rates due to stroke being approximately 20% greater than for individuals in other racial and ethnic groups in the United States [9]. An oral vitamin K antagonist, warfarin, is used to prevent thromboembolic events, but it has a narrow therapeutic index and is affected by wide inter-individual and inter-ethnic dose variability (up to 20-fold), requiring intensive dose management to prevent adverse bleeding events, which can be difficult when patients live in remote communities [12-15].

Over the last 7 years, the average stable anticoagulation dose used in long-term care (> 22 weeks of treatment) for the AI/AN people receiving drug therapy at Southcentral Foundation (SCF) in Anchorage, AK was found to be lower than the average for other sites in North America and Europe, using the same DAWN AC Anticoagulation Management Software (4S Information Systems Ltd., Cumbria, England). For a target INR of 2.5, with a mean INR of 2.0 to 3.0, the average dose at SCF was 4.5 mg/day, with a percent time within INR of 69.7%. For all other locations, the average dose was 4.9 mg/day, with a percent time within INR of 73.0% [16]. These differences in dose may be attributed to inter-individual and inter-ethnic differences, including differences in dietary vitamin K consumption, drug-drug interactions, age, body surface area, gender, concurrent health conditions (e.g., diabetes), and genetic polymorphisms [17-20].

To explore the sources of this reported lower dose among AI/AN people, we characterized the variation in 5 genes that have been associated with altered warfarin response. These include the genes that encode the following enzymes: 1) the pharmacological target of warfarin (*VKORC1),* which reduces vitamin K-epoxide to its active form; 2) the major cytochrome P450 enzyme that clears (S)-warfarin (*CYP2C9*); 3) two enzymes that catabolize vitamin K (*CYP4F2* and *CYP4F11*); and 4) the g-carboxylase that activates vitamin K-dependent clotting factors (e.g., Factors II, VII IX and X) (*GGCX*). It is estimated that by combining knowledge of *VKORC1, CYP2C9*, and *CYP4F2* genotypes with readily accessible clinical factors, including age, gender, and body mass index (BMI), more than 60% of the variance in warfarin dosage can be explained in European-American populations [21].

To assess novel variation in *CYP2C9, VKORC1, CYP4F2*, *CYP4F11*, and *GGCX*, each gene was resequenced in a sample of AI/AN study participants. Population frequencies were determined for alleles that had been previously associated with the dose of warfarin in other populations and for novel non-synonymous alleles that were discovered during resequencing. We hypothesized that there could be novel, function-disrupting variation or higher frequencies of known gene variants in this population that could reduce warfarin dose requirement and impact bleeding/thrombotic risk.

2.2 Methods

Setting. As of 2012, 106,260 AI/AN people live in Alaska, with approximately 1/3 living in more densely populated areas such as Anchorage, Fairbanks, and Juneau, and 2/3 living primarily in rural communities with populations of 50 to 1000 people, with many of the communities accessible only by air or water [22]. Geographic isolation leads to substantial portions of AI/AN people in Alaska being medically under-served and having considerable health disparities compared to other populations [22].

Southcentral Foundation (SCF), a tribally owned and operated regional health corporation, provides pre-paid healthcare services to 58,000 AI/AN patients, who are considered “customer-owners” of SCF. The Anchorage Service Unit (ASU) served by SCF is comprised of both urban and rural areas, including Anchorage, the Matanuska-Susitna Borough, and 60 outlying communities (most with fewer than 500 residents). It provides primary care services to ~ 55% of the total AN population at 6 SCF primary care clinics on the Alaska Native Medical Center (ANMC) campus. Tertiary care is provided at the 150-bed ANMC hospital, which is co-owned and co-managed by SCF and Alaska Native Tribal Health Consortium (ANTHC).

The Center for Alaska Native Health Research (CANHR) is based at the University of Alaska Fairbanks. CANHR has an ongoing genetic research partnership with 11 of the 58 rural communities in the Yukon-Kuskokwim River Delta (Y-K Delta) that are served by the Yukon-Kuskokwim Health Corporation (YKHC). The YKHC is based in Bethel and provides healthcare to about 23,000 Yup'ik people. CANHR collaborates with Yup'ik communities in community-based participatory research focused on understanding, preventing, and reducing health disparities.

Institutional Review Board (IRB) Approval. The Alaska Area Institutional Review Board (IRB), and the SCF and ANTHC tribal review boards approved work conducted at SCF on the ANMC campus. The YKHC Executive Board of Directors and the University of Alaska Fairbanks IRB approved the work conducted in the Y-K Delta by CANHR. The University of Washington (UW) IRB approved the overall research project, as UW is the academic home of the grant funding this research (Pharmacogenetics in Rural and Underserved Populations) and its principal investigators. The National Institute of General Medical Sciences and the Indian Health Service granted a Certificate of Confidentiality for protection of participant information, and the respective Alaska IRBs approved forms for written consent prior to initiating research. Research questions were developed through community-based participatory research at SCF and CANHR.

Study Participants. A convenience sample of study participants (n = 380) was obtained through recruitment by research staff members at SCF's primary care clinics. Any AI/AN person ≥18 years of age and receiving care at SCF was eligible to participate in the study. Surveys collected self-reported gender, date of birth, and tribal affiliation. A representative subset (n=188) was used for resequencing of *CYP2C9*, *VKORC1*, *CYP4F2*, *CYP4F11*, and *GGCX* genes.

A convenience sample of 350 residents of the Y-K Delta, ≥18 years of age, was recruited using written and oral advertisement during research-focused community visits by the CANHR research personnel. All CANHR participants self-identified as Yup'ik. A subset of 94 individuals was chosen for targeted resequencing of *CYP2C9, VKORC1, CYP4F2, CYP4F11*, and *GGCX*; because all participants were recruited from the same communities, these 94 individuals were selected from the set of 350 individuals on the basis of being unrelated according to either pedigree-based kinship coefficients obtained from available genealogical information [23] as well as empirical kinship coefficients calculated using the KING (kinship-based inference for GWAS)-robust method [24] for sample individuals with genome-wide SNP genotyping data available.

Specimen Processing and Storage. Buffy coats were extracted from blood that was collected into EDTA-coated tubes (BD Vacutainer® CPT™), centrifuged (900 x *g*, 15 min) at room temperature, incubated with Puregene RBC Lysis Solution for 10 minutes, and centrifuged again (1800 x *g*, 10 min) at room temperature. White blood cells from CANHR samples were then re-suspended in 10 mL Puregene Cell Lysis Solution until DNA purification. At CANHR, genomic DNA was isolated using the Gentra Puregene kit (Qiagen, Valencia, California, USA) prior to shipment to UW investigators. White blood cells from SCF samples were washed in phosphate-buffer saline (1X PBS), centrifuged again (800 x *g*, 6 min) at room temperature, then re-suspended in PBS and frozen (-80°C) until shipped to UW investigators for DNA isolation. Genomic DNA from the samples of SCF participants was isolated using a QIAamp DNA Blood Midi/Maxi kit (Qiagen, Valencia, California, USA). Quality and concentration of DNA were determined using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Wilmington, Delaware, USA).

Gene Resequencing Methods. Exons, 50-100 base pairs into each adjacent intronic region, 1000-4000 base pairs in the 5' flanking region, and 150-300 base pairs into 3' flanking regions were resequenced using PCR amplicons ~500–600 bp in size, with amplicons containing overlapping segments of ~150 bp to validate primer binding sites and to prevent allele-specific amplification [25]. The PCR primers were standardized with a universal M13-tailed PCR sequence, and used in conventional Sanger sequencing reactions using BigDye chemistry under standard conditions and separated on an ABI 3730 DNA analyzer (Life Technologies, Grand Island, NY, USA). Chromatograms were analyzed using Phrap software (UW, Seattle, Washington, USA) for base calling and quality assignment, and Consed software (UW) was used for assembly and editing [26]. Single nucleotide polymorphisms (SNPs) and small insertions/deletions were identified through pairwise comparison of chromatogram peak heights/intensities using the PolyPhred program (version 5.0; UW) [27, 28], which produces chromatograms averaging greater than 500 bp and averaging a Phred quality greater than 40 (corresponding to a 1/10,000 probability of incorrect base assignment) [29]. Data were verified using second-strand confirmation.

Automated scripts were used to map variants onto the intron–exon gene structure. For the *CYP2C9* and *VKORC1* haplotype analysis, sites were based on human reference sequence *AL359672*.

Coding variants *CYP2C9 M1L*, *N218I*, and *P279T* (*CYP2C9*29*) were analyzed for PolyPhen2 and Grantham scores to predict the phenotypic effect of the amino acid change on enzyme function [30, 31].

Genotyping Methods. We genotyped DNA samples from all study participants for novel coding variants identified through resequencing and for those variants, both intronic and coding, that have published phenotypes. This included 9 SNPs in *CYP2C9*, 3 SNPs in *VKORC1*, 6 SNPs in *CYP4F2,* 2 SNPs in *CYP4F11*, and 3 SNPs in *GGCX*. DNA samples were pre-amplified following Fluidigm's (South San Francisco, CA) Specific Target Amplification protocol to increase available template DNA. TaqMan SNP Genotyping Assays (Applied Biosystems, Inc.) were run on 96.96 Dynamic Genotyping Arrays (Fluidigm) according to the BioMark™ 96.96 Genotyping protocol. Dynamic Arrays were primed and loaded on the Fluidigm HX and thermal cycled on the Fluidigm FC1 controller following pre-set programs in the instruments. End-point fluorescence was read on a BioMark™ Real-Time PCR System (Fluidigm) and analyzed using SNP Genotyping Analysis software (Fluidigm). Samples with overall call rates below 95% were removed from further analysis. Of DNA samples selected for genotyping, 22 from the SCF cohort were excluded due to call rates below 95%. For samples from participants included in both sequencing and genotyping, concordance between calls for the 2 methods was over 99.5%.

Population Substructure Analysis. Genealogical information for the participants from the CANHR sample set (Yup'ik residents of the Y-K Delta in southwestern Alaska) were used to calculate pairwise kinship coefficients between each participant [23]. Using these pedigree relationships, allele frequencies and confidence intervals in the CANHR dataset were calculated according to the best linear unbiased estimator (BLUE) of allele frequency [32] to account for the non-independence of these samples resulting from family structure. This adjustment appropriately weights correlated genotypes based on kinship coefficients.

For participants from SCF, neither pedigree nor genome-wide marker information was collected, so this kinship adjustment could not be calculated. To account for population substructure within the SCF cohort, participants were asked for self-reported tribal affiliation. Participants were grouped based on geographic and language similarities of these affiliations, clustered based on linguistic studies by Krauss [33, 34]. These regions are Northern (Inupiaq), Interior (Athabascan subgroups), Southeastern (Tlingit, Tsimshian, Haida, Eyak), Southwestern (Aleut/Unangan), and Western (Central Yup'ik, Cup'ik, Sugpiaq/Alutiiq). Participants also were given the option of choosing affiliation with multiple groups and affiliation with tribes in the lower 48 states of the US, which resulted in 7 total subgroups of participants. All subgroups are represented in the SCF cohort, with each individual subgroup comprising no more than 17% of the total. Between subgroups of study participants at SCF, Analysis of Variance (ANOVA) was performed at SNPs with known phenotypic effects, with a significance threshold of 0.05.

Comparisons with Other Populations. Population frequencies were compared to the 1000 Genomes Database, which documents the distribution of common genetic variants in geographically and historically diverse populations [35]. The populations used for comparison were Admixed American (AMR), African (AFR), Asian (ASN), and European (EUR). Confidence intervals for allele frequencies were calculated based on the number of individuals included in the 1000 Genomes Database for each SNP, as accessed on September 17, 2014.

Statistical Analysis. Allele frequencies were compared using RStudio version 0.97.551 (RStudio, Inc., Boston, MA) and Haploview 4.2 software [36]. All SNPs identified were tested for deviations from Hardy–Weinberg equilibrium using a χ^2-test. Pairwise linkage disequilibrium (LD) was calculated using Haploview 4.2 software [36]. The r^2 values were used to determine the LD between all non-monomorphic SNPs. The LD display was generated using Haploview 4.2 software [36].

2.3 Results

Resequencing for SNP Identification. The exons and bordering intronic regions of the 5 genes *CYP2C9, VKORC1, CYP4F2, CYP4F11,* and *GGCX* were resequenced in 94 CANHR participants and 188 SCF participants to identify any novel population-specific variation. All SNPs identified in the SCF and CANHR samples are listed in Table 2-1. Novel SNPs not found in the 1000 genomes database as of November 5, 2014 are labeled rsNA, as they do not have rs numbers.

For *CYP2C9,* 33 SNPs were identified in the samples from SCF participants, including 2 novel SNPs. In the samples from CANHR participants, 25 SNPs were identified, including the same 2 novel SNPs. One novel SNP predicted a non-synonymous change from asparagine to isoleucine at amino acid 218 (*N218I* allele). The other was discovered in the first codon, resulting in a change from methionine to leucine (*M1L* allele). The sequencing chromatograms identifying *N218I* are found in Figure 2-1 and those for *M1L* are found in Figure 2-2. This *M1L* SNP was found at frequency of 9.7% (+/- 4.3%) of chromosomes in the 94 CANHR samples subjected to resequencing. *M1L* was also identified in the samples from SCF participants, though at a lower frequency of 1.0% (+/- 0.7%). A known SNP, rs182132442, resulting in a proline to threonine substitution at amino acid 279 (*CYP2C9*29*) is not well characterized and was found in the CANHR cohort only [37]. PolyPhen and Grantham scores predicted a deleterious effect on protein function for all 3 variants [30, 31, 38]. The *M1L* variant had a PolyPhen score of 0.904, predicting a severe effect on protein function based on likely truncation. The *N218I* variant had a Grantham score of 149 and the CYP2C9*29 variant had a Grantham score of 38, predicting severe effects due to chemical dissimilarities of the affected amino acids.

For *VKORC1*, 10 SNPs were identified in resequencing the samples from SCF participants. Of these, 3 were novel synonymous changes. Only 2 SNPs were identified in the CANHR participants, neither of them novel. While the -1639 SNP differentiating the major *VKORC1* haplotypes was assessed, the 1173 base was outside of the sequencing range, though both sites were assessed in subsequent genotyping.

For *CYP4F2*, 34 SNPs were identified in the samples from SCF participants, with 4 of those being novel. One of these novel SNPs changed the splice site of exon 1 (*exon+splice* allele). Within the CANHR participants, 22 SNPs were identified, with the only novel SNP being the *exon+splice*.

For *CYP4F11*, 28 SNPs were identified in the samples from SCF participants, including 5 novel SNPs. One SNP that was novel at the time of sequencing, but has since been named rs199657164, predicted a glycine to arginine change at amino acid 12 (*G12R* allele). One of these five novel SNPs found in the samples from SCF participants predicted a coding change from asparagine to aspartic acid at amino acid 285. In the CANHR participants, 25 SNPs were identified, including 4 novel SNPs, 3 of which were also in with the samples from SCF participants.

Resequencing of *GGCX* identified 21 SNPs in the samples from SCF participants. These SNPs included 3 novel SNPs, including a predicted alanine to glycine change at amino acid 421 (*A421G* allele). Of the SNPs identified in the samples from SCF participants, 11 of those were identified in the samples from CANHR participants, including 1 of the novel SNPs. No unique SNPs were identified in the CANHR cohort that were not found in the SCF cohort.

Genotyping for Population Frequencies. A summary of the characteristics of study participants for whom we recovered DNA producing ≥ 95% genotyping call rates is presented in Table 2-2. Genotyping at specific SNPs was performed to verify the findings from resequencing and to establish better estimates of population frequencies (Table 2-3). The SNPs chosen for genotyping either are SNPs that have a published phenotype, or are non-synonymous SNPs that were discovered during resequencing. Allele frequencies of the samples from the CANHR cohort were adjusted for the kinship between study participants using BLUE [32]. All SNPs were in Hardy-Weinberg equilibrium.

Of the 9 SNPs genotyped in *CYP2C9*, 6 were previously known alleles (**2, *3, *8, *11, *13, *14,* and **29*) and 2 were the *M1L* and *N218I* novel non-synonymous SNPs identified in resequencing. The frequencies of the **29* allele and both novel variants *M1L* and *N218I* were significantly higher (p < 0.05) in the CANHR cohort, and the frequency of *CYP2C9*2* was higher in samples from SCF participants. All other SNPs, with the exception of the *CYP2C9*3* allele, were found at frequencies below 1% of alleles in both cohorts.

Of the 3 SNPs related to *VKORC1* activity, 2 are the SNPs that differentiate the major haplotype groups and typically are seen in complete linkage disequilibrium [39, 40]. The other (rs28940302) predicts a substitution of leucine for valine at amino acid 29 (*V29L*) and has been associated with warfarin resistance [41]. The *V29L* variant was found at less than 1% in both cohorts. Both SNPs designating the major *VKORC1* haplotype associated with lower warfarin dose were found at significantly higher frequencies in the CANHR cohort than in the SCF cohort (p < 0.05). The *VKORC1* and *CYP2C9* diplotypes of the CANHR and SCF cohorts are reported in Table 2-4 and predict phenotypes for warfarin metabolism.

At the *CYP4F* locus, 6 previously identified alleles were assessed, as well as 2 novel non-synonymous changes. The *CYP4F2*3* (*V433M* rs2108622) allele was found at significantly lower frequency in the SCF cohort than in the CANHR cohort. However, the frequency of this variant was high in both populations and the frequency in the CANHR cohort was one of the highest reported for a population to date – similar to the 53.2% reported in the Saudi Arabian tribal subgroup of the Kuwaiti population [42]. Samples from SCF participants had a significantly higher frequency of the *CYP4F2*2* allele (*W12G* rs3093105) than the samples from CANHR participants. The same was true of the *CYP4F2*

M519L (rs3093200) substitution, the *CYP4F2* amino acid 185 change from glycine to valine (*G185V* rs3093153), and the *CYP4F11* amino acid 276 change from arginine to cysteine (*R276C* rs8104361).

The 2 *GGCX* SNPs described previously [43-45] were both found at significantly different frequencies in the SCF and CANHR cohorts ($p < 0.05$). The SNP rs11676382 was less common in the samples from CANHR participants, and the missense SNP (rs699664) was less common in the samples from SCF participants.

Linkage Disequilibrium. Linkage disequilibrium (LD) was calculated between every non-monomorphic SNP in each gene. For *CYP2C9*, LD was low between all SNPs, in both the SCF and CANHR cohorts. In the *VKORC1* gene, the 1173C>T (rs9934438) and -1639G>A (rs9923231) SNPs, which differentiate the 2 major haplotypes of *VKORC1*, were in tight LD in both the samples from SCF participants ($r^2 = 0.96$) and the samples from CANHR participants ($r^2 = 0.97$).

At the *CYP4F* locus (Figure 2-3), the samples from both SCF and CANHR participants show moderate levels of LD between the rs2108622 and rs2189784 SNPs. The samples from SCF participants show moderate LD between the rs2108622 (*3 allele) and the rs3093105 (*2 allele). LD was low between all other SNPs.

The SNPs in *GGCX* were not in LD.

Allele Frequency Comparison Between AN Regional Subgroups. While the samples in the CANHR cohort all were collected from participants living in the Y-K Delta and self-reporting Yup'ik ancestry, many of the participants in the SCF cohort self-identify with tribes historically distributed throughout Alaska but now live in Anchorage or make visits there for healthcare. The allele frequencies of the genotyped SNPs were determined for each of the self-identified regional subgroups of Alaska (Table 2-3). Because subgroup frequencies could not be adjusted for population substructure, the frequency variance is expected to be slightly larger than presented here, but the frequency estimate should not be greatly affected. Based on ANOVA results, SCF subgroups had significantly different allele frequencies at the *VKORC1 -1639, VKORC1 1173,* and *CYP4F2*3* loci.

Allele Frequency Comparison to Other Populations. Comparisons to the 4 main continental groups of the 1000 Genomes Database are presented in Table 2-5 [35]. Generally, the frequencies of variant alleles *CYP2C9*2* and *CYP2C9*3* were low in the samples from SCF participants by global comparison and even lower in the samples from CANHR participants. For *VKORC1*, the frequency of the lower warfarin dose associated AT haplotype (A at rs9934438 and T at rs9923231) was high in the SCF cohort (~60%) and higher in the CANHR cohort (~80%), though both frequencies were higher than in the EUR and AMR and lower than in the ASN populations. At the *CYP4F2* locus, the frequency of the *CYP4F2*3* SNP was higher in both the SCF and CANHR cohorts than in any of the 1000 Genomes Database populations. The allele frequency of *CYP4F2*2* in the CANHR cohort was low compared to the 1000 Genomes Database populations, whereas that of the SCF cohort was similar to the EUR, AMR, and ASN populations. For *GGCX*, the variant allele frequencies in the SCF and CANHR cohorts were similar to the 1000 Genomes Database populations.

2.4 Discussion

The most significant new findings in the AN subpopulations studied were (1) the presence of two, previously unreported, relatively high frequency coding variants in the *CYP2C9* gene (*M1L* and *N218I*) in most, but not all, regional subgroups; (2) a high frequency of the low warfarin dose associated non-coding variants in the *VKORC1* gene, especially in the Yup'ik population living in the Y-K Delta; and 3) a relatively high frequency of the higher warfarin dose associated *CYP4F2*3* variant in some, but not all, regional subgroups. These results are generally consistent with an observed requirement of lower warfarin doses to achieve target INR values in an AI/AN population receiving healthcare in Anchorage, in comparison to that for the non-indigenous population of the US [16].

The identification of relatively common, novel, potentially function-disrupting variants in *CYP2C9* illustrates how pharmacogenetic discoveries made from studies of "representative" world populations do not always capture variation that could be important for other, historically geographically isolated populations, such as the AN people. Population-specific pharmacogenetic studies are necessary to guide anticoagulation therapy in the AI/AN community if clinical testing becomes standard of care and is to be implemented effectively. A prime example is the ATG to TTG change in the first codon of the *CYP2C9* gene (resulting in a predicted methionine to leucine substitution at amino acid position 1) that would be expected to adversely affect mRNA translation and protein synthesis. Although we have not yet confirmed the phenotype of this variant, *M1L* is predicted to disrupt mRNA translation, by alteration of the first codon. Similar disruption of codon 1 (*M1V* or *M1L*) in other genes is associated with a highly penetrant, loss of function phenotype [46-49]. Indeed, the extremely rare *M1V* variant of CYP2C9

(*CYP2C9*36*) was recently described in a Chinese population, and its recombinant expression was found to result in low accumulation of the variant enzyme relative to wild-type in COS cells [50].

The novel asparagine to isoleucine substitution at amino acid 218 was also predicted to have a deleterious effect on CYP2C9 enzyme function. *N218I* is located between helices F and G in an area of the enzyme known to be important for catalytic activity [51]. To our knowledge, no protein variant at this position has been prepared and tested for function, but the neighboring, inter-helical, *Q214L* variant is *CYP2C9*28*, which expressed well in COS-7 cells, but exhibited no detectable S-warfarin 7-hydroxylation activity [52].

Another known, but little studied variant, a proline to threonine change at position 279 is predicted to reduce the catalytic function of CYP2C9. The *CYP2C9 P279T* variant (*CYP2C9*29*) resides between the H and I helices of *CYP2C9* [51], and has been found at low frequencies in Japanese and Chinese populations [37, 50]. The recombinant enzyme expressed well in mammalian cells and exhibited a ~30% reduction in tolbutamide hydroxylation activity relative to wild-type [50]. A more detailed kinetic analysis of S-warfarin 7-hydroxylation found no difference in K_m compared to wild-type CYP2C9 and a ~50% decrease in V_{max}, similar to that found for *CYP2C9*2* [52], which is an established risk factor for warfarin sensitivity.

Combined with the known, functionally important *CYP2C9* loss of function coding variants (**2* and **3* alleles), these novel potentially loss of function variants were found at a frequency that ranged from 9.5% in the Northern to 14.8% in the Interior AN subgroups. Importantly, individuals carrying *CYP2C9*2* and *3 alleles are at increased risk of major bleeding events following the initiation for warfarin therapy [53]. If relying on variants identified in European populations, the much lower frequencies of *2 and *3 in AN populations, especially in the Yup'ik population, compared to European populations would underestimate overall *CYP2C9* coding variation. Novel *CYP2C9* variants (*M1L* and *N218I*) first identified in this paper, and a relatively new variant (*P279T*), may confer a similar risk in AN subgroups. Thus, while the exact composition of *CYP2C9* coding variation differs between the AN regional subgroups and other populations globally, the overall impact of these changes on warfarin dose requirement and bleeding risk could be the same.

Variation seen in the *VKORC1* gene is also likely to have a significant effect on warfarin dose requirement in regional AN subgroups. This is especially true for the Yup'ik people of the Y-K Delta, for whom we observed the highest frequency of the low dose haplotype. Although the specific mechanism is unknown, these non-coding SNPs are consistently associated with reduced gene transcription and protein expression, and with lower warfarin dose requirements in other populations [54, 55].

Although *CYP4F2* variation is thought to contribute less to inter-individual differences in warfarin dose requirement than do variation in *CYP2C9* and *VKORC1*, it may have a greater generalized impact on warfarin dose requirements in several AN regional subpopulations because of the observed relatively high frequency of the **3* allele. The *CYP4F2*3* variant is associated with reduced metabolic clearance of vitamin K, increased vitamin K levels, and increased warfarin dose [56, 57]. It would be expected to counteract to some degree the effect of the reduced function *CYP2C9* and *VKORC1* variants on average dose requirement, although a given individual could fall anywhere along a wide continuum of warfarin dose sensitivity based on overall genetic constitution. Selective pressure may have acted on the *CYP4F2* gene to conserve vitamin K as a result of inconsistent access to greens (e.g., traditional tundra greens, beach greens) throughout the year. Alternatively, it could be the result of a founder effect, with ancestral blocks of DNA preserved over time. Of particular relevance is the recent report of reduced bleeding risk following variation in vitamin K consumption in individuals carrying the *CYP4F2*3* allele who receive long-term warfarin therapy [53]. If changes in vitamin K consumption and accumulation in the body affect the risk of a major bleeding event in individuals receiving warfarin, the relatively high frequency of the *CYP4F2*3* allele may provide some resiliency against that adverse event.

Variation in *GGCX* and *CYP4F11* is not expected to have implications for warfarin therapy in the AN subpopulations. While genes were sequenced to look for novel SNPs, none were discovered at a frequency expected to affect enzyme function at a population level. Known variation in these genes has not been associated with warfarin dose or response.

We sought to identify and characterize variation in 5 genes associated with warfarin dose requirement and drug response in two cohorts of AN people. This study is the first to partner with a diverse population to systematically sequence these genes to identify population-specific variation. Furthermore, by genotyping these and other SNPs relevant for warfarin response in cohorts of AN subgroups, we have established robust population frequencies that can be used to guide patient care. Although this study does not have information on drug dose or bleeding events for participants or functional studies on novel variation, based on known genotype-phenotype relationships, our phenotypic predictions of these genetic findings are well supported.

The presence of novel *CYP2C9* gene variants and relatively high frequencies of variant alleles in the *VKORC1* and *CYP4F2* genes support our hypothesis that pharmacogenetic research in understudied populations is needed and suggests the possibility of significant associations with warfarin dose requirement and bleeding risk in the AN subpopulations. Warfarin is still the drug of choice for chronic therapy in the prevention of thromboembolic events for at-risk individuals receiving health care at SCF and YKHC. Although treatment is challenging for all of the reasons that apply to other populations [12-15], major bleeding risk associated with alternative direct thrombin inhibitors and the absence of an antidote for those drugs is a deterrent to changing treatment in patients who often have restricted access to emergency care because of geographic isolation. Thus, the findings we report may advance the development of genetic tests for optimizing the initiation of warfarin therapy and for identifying individuals at increased risk of major bleeding events during chronic therapy.

2.5 Figures

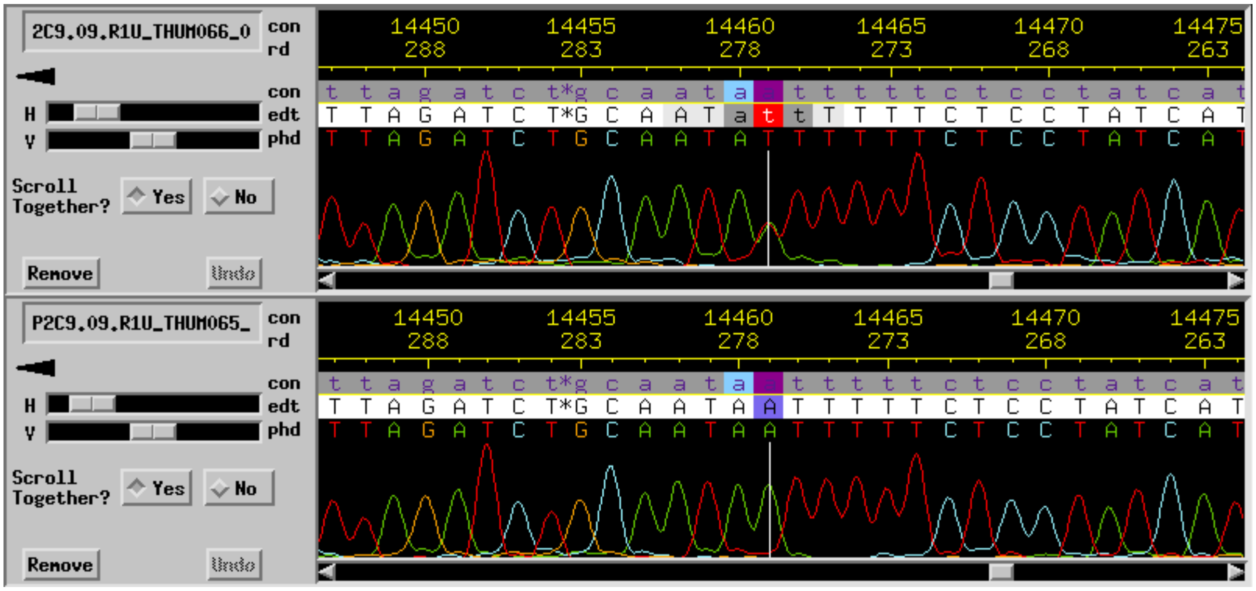
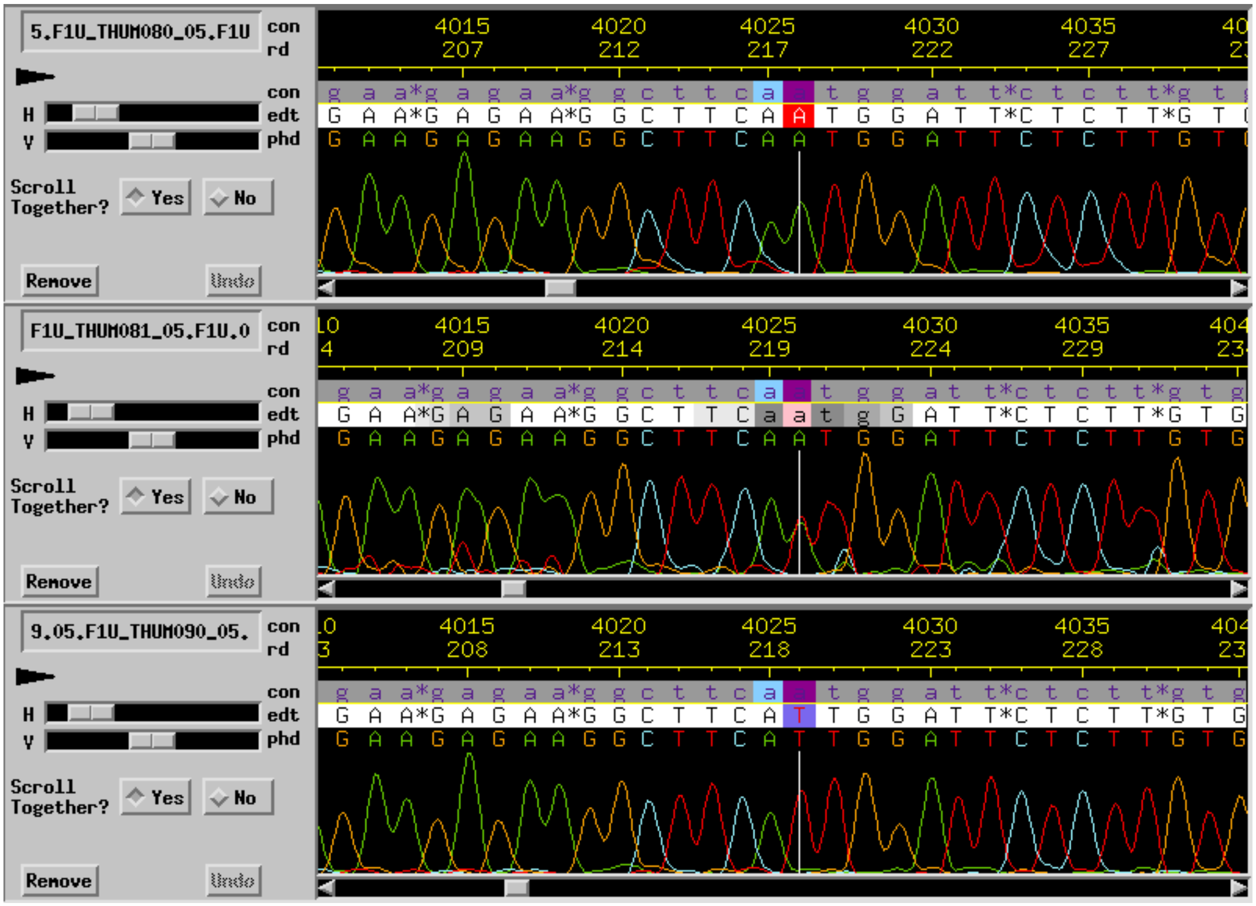

2C9.09.R1U_THUM066_0
con
rd
con
edt
phd
H
V
Scroll Together?
Yes
No
Remove
Undo
14450 288
14455 283
14460 278
14465 273
14470 268
14475 263
t t a g a t c t*g c a a t a t t t t t c t c c t a t c a t
T T A G A T C T*G C A A T a t t T T T T C T C C T A T C A T
T T A G A T C T G C A A T A T T T T T C T C C T A T C A T
P2C9.09.R1U_THUM065_
con
rd
con
edt
phd
H
V
Scroll Together?
Yes
No
Remove
Undo
14450 288
14455 283
14460 278
14465 273
14470 268
14475 263
t t a g a t c t*g c a a t a t t t t t c t c c t a t c a t
T T A G A T C T*G C A A T A A T T T T T C T C C T A T C A T
T T A G A T C T G C A A T A A T T T T T C T C C T A T C A T

Figure 2-1: Electrochromatogram of novel SNP *N218I*, showing homozygous and heterozygous samples at that position. The color of florescence at each base indicates which base is at a sequence location, with the height of a peak indicating intensity of florescence. Overlapping peaks represent heterozygosity and high peaks of 1 color represent homozygosity of the corresponding base. Electrochromatograms produced using Phred/Phrap/Polyphred/Consed system from Sanger sequencing[26]. The top electrochromatogram shows heterozygosity at the N218I position. The bottom chromatogram shows homozygosity of the reference base at that position.



```
    4015      4020      4025      4030      4035      40
    207       212       217       222       227       2
con g a a*g a g a a*g g c t t c a a t g g a t t*c t c t t*g t
edt G A A*G A G A A*G G C T T C A A T G G A T T*C T C T T*G T
phd G A A G A G A A G G C T T C A A T G G A T T C T C T T G T
```

```
10  4015      4020      4025      4030      4035      404
4   209       214       219       224       229       23
con g a a*g a g a a*g g c t t c a a t g g a t t*c t c t t*g t g
edt G A A*G A G A A*G G C T T C a a t g G A T T*C T C T T*G T G
phd G A A G A G A A G G C T T C A A T G G A T T C T C T T G T G
```

```
10  4015      4020      4025      4030      4035      404
3   208       213       218       223       228       23
con g a a*g a g a a*g g c t t c a a t g g a t t*c t c t t*g t g
edt G A A*G A G A A*G G C T T C A T T G G A T T*C T C T T*G T G
phd G A A G A G A A G G C T T C A T T G G A T T C T C T T G T G
```

Figure 2-2: Electrochromatogram of Sanger sequencing output at the novel SNP *M1L*, showing homozygous and heterozygous samples at that position. Based on relative intensity of florescence tagged bases, the top chromatogram shows homozygosity for the allele, the middle chromatogram shows heterozygosity, and the bottom chromatogram shows homozygosity for the allele at the M1L position. Chromatograms produced with the Phred/Phrap/ Polyphred/Consed system.

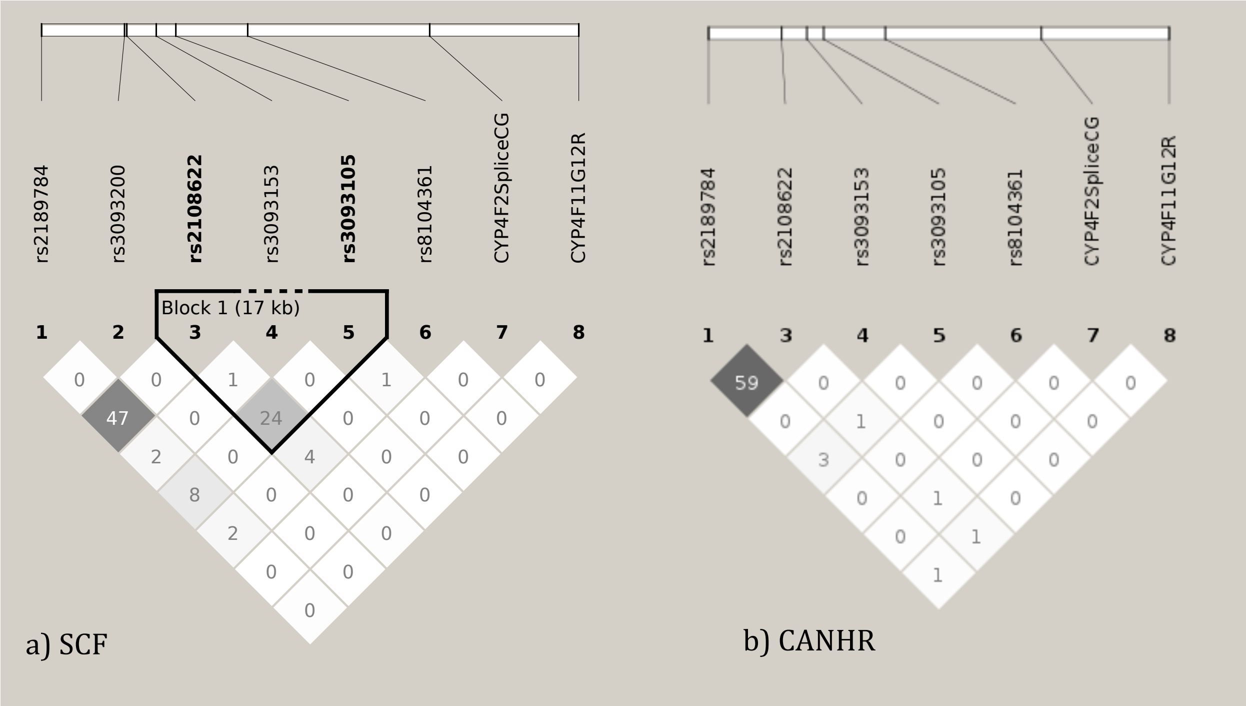


Figure 2-3: Linkage Disequilibrium (LD) in the *CYP4F* locus for all non-monomorphic SNPs in the participants from a) SCF and b) CANHR sample sets, by r^2 measure. Pairwise comparisons illustrate low LD between most SNPs. The rs2189784 and rs2108622 SNPs are found in tighter LD in the samples from CANHR participants than in the samples from SCF participants. These patterns illustrate the effects of potential founder effects and population isolation, resulting in differences in genetic patterns found in regional subgroups of Alaska. Variant pairs with LD scores closer to 100 were more often inherited together than not. Scores were calculated with Haploview version 4.2. Haplotype blocks determined by confidence intervals.

2.6 Tables

Table 2-1. Sequence variation in *CYP2C9, CYP4F2, VKORC1*, and *GGCX* in Yup'ik people living in the Y-K Delta. SNPs identified through resequencing of warfarin genes in SCF (n=188 individuals) and CANHR (n=94) cohorts. "X" indicates that the SNP was found in that population. A "1" indicates that the SNP was found on only one allele total, in the population marked.

CYP2C9

Allele	rs Number	Chromosome 10 position	SCF	CANHR
	9332092	96696529	X	X
	9332093	96696555	X	X
	61604699	96696903	X	X
	4918758	96697252	X	X
	4917636	96697344	X	X
	9332098	96697459	X	X
	9332100	96697820	X	X
	9332101	96697955	X	X
	9332102	96697956	X	X
M1L	NA	96698440	X	X
	9332104	96698690	X	X
	114071557	96701593	X	X
	9332119	96701601	X	X
	17847036	96701674	X	X
	9332120	96701850	X	X
*2 R144C	1799853	96702047	X	X
	114071557	96707696	1	
N218I	NA	96708875	X	X
	9332172	96731788	X	X
	9332197	96740908	X	
*3 I359L	1057910	96741053	X	X
	57811561	96745742	1	
	9332230	96745984	X	X
	28371688	96748492	X	X
	28371689	96748495	X	X
	1057911	96748737	X	X
	376114904	96748752	1	
*12 P489S	9332239	96748777	1	
	9332242	96748893	X	
	9332245	96749181	X	
	148135940	96749312	X	
	182792423	96749327	1	
*29 P279T	182132442	96731876		X

VKORC1

Allele	rs Number	Chromosome 16 position	SCF	CANHR
	7294	31102321	X	X
	61742233	31105922	1	
	55894764	31106015	1	
	NA	31106257	1	
	201733800	31106927	1	
	149536015	31106991	1	
	148249176	31107232	X	
	NA	31107503	X	
	9923231	31107689	X	X
	NA	31108008	1	

CYP4F2

Allele	rs Number	Chromosome 19 position	SCF	CANHR
	1126433	15989405	X	X
	3952538	15989523	X	
	NA	15989564	X	
M519L	3093200	15989589	X	
A483G	3952537	15989696	X	
G464V	NA	15990162	X	
M433V	2108622	15990431	X	X
Exon+1 splice	NA	15990573	X	X
P409S	200492423	15990598	X	
	2074900	15996820	X	X
	3093160	15996907	X	X
	3093158	16000166	X	X
G185V	3093153	16001215	X	
	3093114	16006413	X	X
	3093106	16008257	X	X
**3 G12W*	3093105	16008388	X	X
**2*	3093103	16008434	X	X
	201641652	16008435	1	
	3093100	16008469	X	X
	3093098	16008512	X	X
	4646498	16008643	X	X
	3093097	16008691	X	X
	NA	16008722	1	
	3093092	16009127	X	X
	3093091	16009133	X	X
	3093090	16009134	X	X
	147387603	16009153	X	X
	3093089	16009292	X	X
	146748375	16009294	X	X
	3093088	16009388	X	X
	3093086	16009607	X	X
	117961148	16009905	1	
	182792423	16009941	1	
	187396291	16010095	X	

CYP4F11

Allele	rs Number	Chromosome 19 position	SCF	CANHR
	1060467	16024538	X	X
	1064796	16024662	X	X
	2072269	16024739	X	X
N446D	1060463	16025176	X	X
	12985248	16025312	X	X
	NA	16025541	X	X
N285D	NA	16034687	1	
R276C	8104361	16034714	X	X
	3746153	16035474	X	X
	3746154	16035494	X	X
	3746156	16035517	X	X
	2219358	16038334	X	X
	2305804	16038365	X	X
	2305803	16038390	X	X
	NA	16040230	1	
	3765070	16040292	X	X
	NA	16040440		1
	4808414	16040473	X	X
	2305801	16045141	X	X
G12R	199657164	16045185	X	X
	148250072	16045186	X	
	2305800	16045294	X	X
	11879253	16045491	X	X
	12985091	16045749	X	X
	3826950	16045891	X	X
	3810428	16046478	X	X
	3810427	16046650	X	X
	NA	16047103	1	
	NA	16047237	X	X

GGCX

Allele	rs number	Chromosome 2 position	SCF	CANHR
	11676382	85777633	X	X
R498R	41290033	85779050	X	
A421G	NA	85780087	X	
T414T	10179904	85780107	X	
R406R	2592551	85780131	X	X
P337L	NA	85780500	1	
Q325R	699664	85780536	X	X
G279G	1254896	85781318	X	
	78185751	85782587	X	X
D113D	6751560	85786074	X	X
	1254898	85788140	X	X
	7568458	85788175	X	X
	NA	85788420	X	X
	78809601	85788476	1	
	115669754	85788738	X	
	148690568	85789101	X	X
	75830997	85789135	X	
	142935757	85789199	X	
	6707308	85789906	X	
	145257780	85789963	X	X
	11890182	85790165	X	X

Table 2-2. Demographic characteristics of genotyped study cohorts. SCF participants were classified by self-reported tribal affiliation, clustered by geographic region and linguistic similarities. Only participants for whom genotyping reached ≥ 95% call rate for all alleles tested were included.

Population	**Subpopulation**	**Males**	**Females**	**Total**
SCF		125	234	359
	Interior	17	20	37
	Northern	30	55	85
	Southeastern	16	26	42
	Southwestern	23	37	60
	Western	8	41	49
	Multiple	21	35	56
	Lower 48	10	20	30
CANHR		165	185	350

Table 2-3. Prevalence of *CYP2C9*, *CYP4F2*, *VKORC1*, and *GGCX* variant alleles in the SCF and CANHR AI/AN cohorts of Alaska, as determined using the Fluidigm genotyping platform. The SCF sample participants are presented in total (column 6) and divided into regional subgroups (columns 7-13). Reference allele (Ref) obtained from dbSNP [58]. Reported frequency is of the variant allele (Var) listed. Frequencies are reported in percentages, with 95% confidence intervals for the true population allele frequency in parentheses.

CYP2C9

Allele	rs number	Ref	Var	CANHR (n=350)	SCF (n=359)	Interior (n=37)	Northern (n=85)	South eastern (n=42)	South western (n=60)	Western (n=49)	Multiple (n=56)	Lower 48 (n=30)
**13 L90P*	72558187	T	C	0.0	0.0	0.0	0.0	1.2 (0 – 3.5)	0.0	0.0	0.0	0.0
**14 R125H*	72558189	G	A	0.0	0.0	0.0	0.0	1.2 (0 – 3.5)	0.0	0.0	0.0	0.0
**2 R144C*	1799853	C	T	0.3 (0 – 0.7)	5.2 (3.6 – 6.8)	5.4 (0.3 – 10.5)	4.1 (1.1 – 7.1)	7.1 (1.6 – 12.6)	5.8 (1.6 – 10.0)	6.1 (1.4 – 10.8)	1.8 (0 – 4.3)	8.3 (1.3 – 15.3)
**8 R150L*	7900194	G	A	0.0	0.1 (0 – 0.3)	0.0	0.0	0.0	0.0	0.0	0.0	1.7 (0 – 5.0)
**11 R355W*	28371685	C	T	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
**3 I359L*	1057910	A	C	2.1 (1.0 – 3.2)	3.4 (2.1 – 4.7)	4.1 (0 – 8.6)	1.2 (0 – 2.8)	3.6 (0 – 7.6)	5.9 (1.7 – 10.1)	0.0	3.6 (0.1 – 7.1)	8.3 (1.3 – 15.3)
M1L	NA	T	A	6.3 (4.5 – 8.1)	1.0 (0.3 – 1.7)	1.4 (0 – 4.1)	1.8 (0 – 3.8)	0.0	0.8 (0 – 2.4)	1.0 (0 – 3.0)	0.9 (0 – 2.6)	0.0
N218I	NA	A	T	3.8 (2.4 – 5.3)	1.4 (0.5 – 2.3)	5.4 (0.3 – 10.5)	2.4 (0.1 – 4.7)	0.0	0.8 (0 – 2.4)	1.0 (0 – 3.0)	0.0	0.0
**29 P279T*	182132442	C	A	2.1 (1.0 – 3.2)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

	VKORC1											
Allele	rs number	Ref	Var	CANHR (n=350)	SCF (n=359)	Interior (n=37)	Northern (n=85)	South eastern (n=42)	South western (n=60)	Western (n=49)	Multiple (n=56)	Lower 48 (n=30)
V29L	28940302	G	T	0.0	0.1 (0 – 0.3)	0.0	0.0	1.2 (0.0 – 3.6)	0.0	0.0	0.0	0.0
1173	9934438	G	A	77.5 (74.4 – 80.7)	59.7 (56.1 – 63.3)	52.7 (41.3 – 64.1)	64.7 (57.5 – 71.9)	53.6 (42.9 – 64.3)	55.0 (46.1 – 63.9)	68.4 (59.2 – 77.6)	64.3 (55.4 – 73.2)	50.0 (37.3 – 62.7)
-1639	9923231	C	T	77.8 (74.6 – 80.9)	59.7 (56.1 – 63.3)	54.1 (42.7 – 65.5)	64.1 (56.9 – 71.3)	53.6 (42.9 – 64.3)	54.2 (45.3 – 63.1)	68.4 (59.2 – 77.6)	64.3 (55.4 – 73.2)	50.0 (37.3 – 62.7)

CYP4F

Allele	rs number	Ref	Var	CANHR (n=350)	SCF (n=359)	Interior (n=37)	Northern (n=85)	South eastern (n=42)	South western (n=60)	Western (n=49)	Multiple (n=56)	Lower 48 (n=30)
4F2	2189784	G	A	39.6 (35.9 – 43.2)	31.0 (27.6 – 34.4)	29.7 (19.3 – 40.1)	34.1 (27.0 – 41.2)	32.1 (22.1 – 42.1)	28.0 (20.0 – 36.0)	34.7 (25.3 – 44.1)	27.7 (19.4 – 36.0)	30.0 (18.4 – 41.6]
4F2 *M519L*	3093200	C	A	0.0	2.7 (1.5 – 3.9)	2.7 (0 – 6.4)	1.8 (0 – 3.8)	2.4 (0 – 5.7)	0.8 (0 – 2.4)	1.1 (0 – 3.2)	5.4 (1.2 – 9.6)	6.7 (0.4 – 13.0)
*4F2*3* *V433M*	2108622	C	T	50.9 (47.2 – 54.7)	31.5 (28.1 – 34.9)	31.1 (20.6 – 41.6)	36.5 (29.3 – 43.7)	39.3 (28.9 – 49.7)	22.9 (15.4 – 30.4)	35.7 (26.2 – 45.2)	32.1 (23.5 – 40.7)	16.7 (7.3 – 26.1)
4F2 *G185V*	3093153	G	T	0.3 (0 – 0.6)	2.2 (1.1 – 3.3)	4.1 (0 – 8.6)	1.8 (0 – 3.8)	0.0	4.2 (0.6 – 7.8)	1.0 (0 – 3.0)	1.8 (0 – 4.3)	3.3 (0 – 7.8)
*4F2*2* *W12G*	3093105	T	G	3.7 (2.3 – 5.1)	11.0 (8.7 – 13.3)	16.2 (7.8 – 24.6)	10.6 (6.0 – 15.2)	22.6 (13.7 – 31.5)	7.5 (2.8 – 12.2)	6.1 (1.4 – 10.8)	8.0 (3.0 – 13.0)	10.0 (2.4 – 17.6)
4F2 *spliceCG*	NA	C	G	0.7 (0.1 – 1.4)	1.4 (0.5 – 2.3)	9.7 (3.0 – 16.4)	1.2 (0 – 2.8)	1.2 (0 – 3.5)	0.0	1.0 (0 – 3.0)	0.9 (0 – 2.6)	0.0
4F11 *R276C*	8104361	G	A	0.3 (0 – 0.7)	9.1 (7.0 – 11.2)	6.8 (1.1 – 12.5)	7.6 (3.6 – 11.6)	8.3 (2.4 – 14.2)	10.8 (5.2 – 16.4)	7.1 (2.0 – 12.2)	3.6 (0.1 – 7.1)	23.3 (12.6 – 34.0]
4F11 *G12R*	NA	C	G	1.7 (0.7 – 2.6)	0.8 (0.2 – 1.5)	1.4 (0 – 4.1)	0.0	2.4 (0 – 5.7)	0.8 (0 – 2.4)	1.0 (0 – 3.0)	0.9 (0 – 2.6)	0.0

GGCX

Allele	rs number	Ref	Var	CANHR (n=350)	SCF (n=359)	Interior (n=37)	Northern (n=85)	South eastern (n=42)	South western (n=60)	Western (n=49)	Multiple (n=56)	Lower 48 (n=30)
	11676382	C	G	0.3 (0 – 0.7)	3.8 (2.4 – 5.2)	5.4 (0.3 – 10.5)	2.4 (0.1 – 4.7)	2.4 (0 – 5.7)	4.2 (0.6 – 7.8)	4.1 (1.7 – 8.0)	2.7 (0 – 5.7)	8.3 (1.3 – 15.3)
G421A	NA	G	C	0.0	0.6 (0 – 1.2)	0.0	0.0	2.4 (0 – 5.7)	0.0	0.0	1.8 (0 – 4.3)	0.0
R325Q	699664	G	A	49.1 (45.3 – 52.9)	35.9 (32.4 – 39.4)	31.1 (20.6 – 41.6)	38.2 (30.9 – 45.5)	23.8 (14.7 – 32.9)	29.2 (21.1 – 37.3)	49.0 (39.1 – 58.9)	43.8 (34.6 – 53.0)	30.0 (18.4 – 41.6)

Table 2-4: Frequencies of predicted diplotypes known to affect warfarin dose in the CANHR and SCF datasets. Diplotypes of *VKORC1* and *CYP2C9* were calculated from observed 1173 and -1639-containing high and low dose *VKORC1* haplotypes, and *CYP2C9 *1, *2, *3* and the *M1L, N218I*, and *P279T* genotypes, assuming no LD between *CYP2C9* variants.

CANHR Diplotype Frequencies

VKORC1 diplotype	*CYP2C9* diplotype									
	**1/*1*	**1/*2*	**1/*3*	**2/*2*	**1/M1L*	**1/N218I*	**1/*29*	*M1L/M1L*	**3/M1L*	Total
Low dose homozygous (AT/AT)	153	1	6	0	25	20	12	2	3	222
Low dose heterozygous (AT/GC)	77	0	3	1	13	6	1	1	0	102
High dose homozygous (GC/GC)	17	1	2	0	4	1	0	1	0	26
Total	247	2	11	1	42	27	13	4	3	350

SCF Diplotype Frequencies

VKORC1 diplotype	*CYP2C9* diplotype									
	**1/*1*	**1/*2*	**1/*3*	**2/*2*	**1/M1L*	**1/N218I*	**2/N218I*	**2/*3*	**3/*3*	Total
Low dose homozygous (AT/AT)	101	10	5	0	4	3	0	1	0	124
Low dose heterozygous (AT/GC)	145	15	10	1	1	5	1	0	1	179
High dose homozygous (GC/GC)	41	7	5	0	1	1	0	1	0	56
Total	287	32	20	1	6	9	1	2	1	359

Table 2-5: Comparison of allele frequencies of SCF and CANHR cohorts to global populations from the 1000 genomes database[35]. Frequencies are reported as percentage of the variant allele, including 95% confidence interval. The populations used for comparison were Admixed American (AMR), African (AFR), Asian (ASN), and European (EUR) from the 1000 genomes database.

CYP2C9

Allele	rs number	Reference allele	Variant allele	SCF	CANHR	AMR	AFR	ASN	EUR
Number of alleles				718	700	362	492	572	758
**2 R144C*	1799853	C	T	5.2 (3.6 – 6.8)	0.3 (0 – 0.7)	12.4 (9.0 – 15.8)	1.8 (0.6 – 3.0)	0.3 (0 – 0.7)	12.3 (10.0 – 14.6)
**3 I359L*	1057910	A	C	3.4 (2.1 – 4.7)	2.1 (1.0 – 3.2)	5.8 (3.4 – 8.2)	0.6 (0 – 1.3)	4.0 (2.4 – 5.6)	6.1 (4.4 – 7.8)

VKORC1

Allele	rs number	Reference Allele	Variant allele	SCF	CANHR	AMR	AFR	ASN	EUR
Number of alleles				718	700	362	492	572	758
1173	9934438	G	A	59.7 (56.1 – 63.3)	77.5 (74.4 – 80.7)	43.9 (38.8 – 49.0)	6.5 (4.3 – 8.7)	91.8 (89.6 – 94.0)	40.1 (36.6 – 43.6)
-1639	9923231	C	T	59.7 (56.1 – 63.3)	77.8 (74.6– 80.9)	43.9 (38.8 – 49.0)	6.5 (4.3 – 8.7)	91.8 (89.6 – 94.0)	40.1 (36.6 – 43.6)

CYP4F

Allele	rs number	Reference Allele	Variant Allele	SCF	CANHR	AMR	AFR	ASN	EUR
Number of alleles				718	700	362	492	572	758
4F2	2189784	G	A	31.0 (27.6 – 34.4)	39.6 (35.9 – 43.2)	41.7 (36.6 – 46.8)	28.7 (24.7 – 32.7)	26.6 (23.0 – 30.2)	43.0 (39.5 – 46.5)
*4F2*3 V433M*	2108622	C	T	31.5 (28.1 – 34.9)	50.9 (47.2 – 54.7)	28.5 (23.8 – 33.2)	8.5 (6.0 – 11.0)	20.6 (17.3 – 23.9)	27.3 (24.1 – 30.5)
4F2 G185V	3093153	C	A	2.2 (1.1 – 3.3)	0.3 (0 – 0.6)	4.4 (2.3 – 6.5)	2.6 (1.2 – 4.0)	0	7.5 (5.6 – 9.4)
*4F2*2 W12G*	3093105	A	C	11.0 (8.7 – 13.3)	3.7 (2.3 – 5.1)	17.7 (13.8 – 21.6)	24.0 (20.2 – 27.8)	7.2 (5.1 – 9.3)	15.8 (13.2 – 18.4)
4F11 R276C	8104361	G	A	9.1 (7.0 – 11.2)	0.3 (0 – 0.7)	22.4 (18.1 – 26.7)	29.3 (25.3 – 33.3)	7.2 (5.1 – 9.3)	26.5 (23.4 – 30.0)

GGCX

Allele	rs number	Reference Allele	Variant Allele	SCF	CANHR	AMR	AFR	ASN	EUR
Number of alleles				718	700	362	492	572	758
	11676382	C	G	3.8 (2.4 – 5.2)	0.3 (0 – 0.7)	3.3 (1.5 – 5.1)	0.6 (0 – 1.3)	0	8.7 (6.7 – 10.7)
R325Q	699664	C	T	35.9 (32.4 – 39.4)	49.1 (45.3 – 52.9)	29.3 (24.6 – 34.0)	69.1 (65.0 – 73.2)	32.9 (29.0 – 36.8)	35.9 (32.5 – 39.3)

2.7 References

1. Shin, J., *Clinical pharmacogenomics of warfarin and clopidogrel.* J Pharm Pract, 2012. **25**(4): p. 428-438.
2. Bhathena, A. and B.B. Spear, *Pharmacogenetics: improving drug and dose selection.* Curr Opin Pharmacol, 2008. **8**(5): p. 639-646.
3. Thummel, K.E. and Y.S. Lin, *Sources of interindividual variability.* Methods Mol Biol, 2014. **1113**: p. 363-415.
4. Eichelbaum, M., M. Ingelman-Sundberg, and W.E. Evans, *Pharmacogenomics and Individualized Drug Therapy* Annual Review of Medicine, 2006. **57**(1).
5. Smart, A. and P. Martin, *The promise of pharmacogenetics: assessing the prospects for disease and patient stratification.* Stud Hist Philos Biol Biomed Sci, 2006. **37**(3): p. 583-601.
6. Boyer, B., et al., *Ethical issues in developing pharmacogenetic research partnerships with American Indigenous communities.* Clinical pharmacology and therapeutics, 2011. **89**(3): p. 343-5.
7. Fohner, A., et al., *Pharmacogenetics in American Indian populations: analysis of CYP2D6, CYP3A4, CYP3A5, and CYP2C9 in the Confederated Salish and Kootenai Tribes.* Pharmacogenetics and genomics, 2013. **23**(8): p. 403-14.
8. Shaw, J.L., et al., *Risk, reward, and the double-edged sword: perspectives on pharmacogenetic research and clinical testing among Alaska Native people.* Am J Public Health, 2013. **103**(12): p. 2220-2225.
9. Horner, R.D., et al., *Stroke mortality among Alaska Native people.* Am J Public Health, 2009. **99**(11): p. 1996-2000.
10. Go, A.S., et al., *Heart disease and stroke statistics--2014 update: a report from the American Heart Association.* Circulation, 2014. **129**(3): p. e28-e292.
11. Howard, B.V., et al., *All-cause, cardiovascular, and cancer mortality in western Alaska Native people: Westen Alaska Tribal Collaborative for Health (WATCH).* Am J Public Health, 2014. **104**(7): p. 1334:40.
12. Stafford, R.S. and D.E. Singer, *Recent national patterns of warfarin use in atrial fibrillation.* Circulation, 1998. **97**(13): p. 1231-1233.
13. Birman-Deych, E., et al., *Use and effectiveness of warfarin in Medicare beneficiaries with atrial fibrillation.* Stroke, 2006. **37**(4): p. 1070-1074.
14. Shapiro, S.S., *Treating thrombosis in the 21st century.* N Engl J Med, 2003. **349**(18): p. 1762-1764.
15. Hart, R.G., L.A. Pearce, and M.I. Aguilar, *Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation.* Ann Intern Med, 2007. **146**(12): p. 857-867.
16. Schilling, B. *Anticoagulation Care for Alaska Native Customer-Owners within the Nuka Model of Care.* in *7th Dawn AC Anticoagulation Management Software North American User Group Meeting*. 2013. La Jolla, CA.
17. Daly, A.K., *Pharmacogenetics of the major polymorphic metabolizing enzymes.* Fundam Clin Pharmacol, 2003. **17**(1): p. 27-41.
18. Daly, A.K., *Pharmacogenetics of the cytochromes P450.* Curr Top Med Chem, 2004. **4**(16): p. 1733-1744.
19. Ingelman-Sundberg, M., et al., *Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects.* Pharmacol Ther, 2007. **116**(3): p. 496-526.
20. Zanger, U.M., et al., *Functional pharmacogenetics/genomics of human cytochromes P450 involved in drug biotransformation.* Anal Bioanal Chem, 2008. **392**(6): p. 1093-108.

21. McDonagh, E., et al., *From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource.* Biomarkers in medicine, 2011. **5**(6): p. 795-806.
22. Alaska Department of Labor and Workforce Development, R.a.A.S., *Alaska Population Overivew: 2012 Estimates*, 2013. p. 128.
23. Bourgain, C. and Q. Zhang, *Kinship and Inbreeding coefficients computation in general pedigrees*, 2009, Free Software Foundation, Inc.: Boston, MA.
24. Manachaikul, A., et al., *Robust relationship inference in genome-wise associaiton studies.* Bioinformatics, 2010. **26**(22): p. 2867-2873.
25. Rieder M.J, et al., *Sequence variation in the human angiotensin converting enzyme.* Nature genetics, 1999. **22**(1): p. 59-62.
26. Gordon, D., C. Abajian, and P. Green, *Consed: A Graphical Tool for Sequence Finishing.* Genome Research, 1998. **8**(3): p. 195-202.
27. Nickerson, D.A., V.O. Tobe, and S.L. Taylor, *PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing.* Nucleic acids research., 1997. **25**(14): p. 2745.
28. Stephens, M., et al., *Automating sequence-based detection and genotyping of SNPs from diploid samples.* Nature genetics, 2006. **38**(3): p. 375-81.
29. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome research, 1998. **8**(3): p. 186-94.
30. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.
31. Grantham, R., *Amino Acid Difference Formula to Help Explain Protein Evolution.* Science, 1974. **185**(4154): p. 862-864.
32. McPeek, M.S., X. Wu, and C. Ober, *Best Linear Unbiased Allele-Frequency Estimation in Complex Pedigrees.* Biometrics, 2004. **60**: p. 359-367.
33. Krauss, M.E., *Alaska Native Languages: Past, Present, and Future*. Alaska Native Language Center Research Papers. Vol. 4. 1980, Fairbanks, AK: Alaska Native Language Center. 110.
34. Krauss, M.E., A.N.L.C.U.o. Alaska, and A.I.o.S.E. Research, *Indigenous peoples and languages of Alaska*, 2011, Alaska Native Language Center, University of Alaska Fairbanks: [Fairbanks].
35. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
36. Barrett, J., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics (Oxford, England), 2005. **21**(2): p. 263-5.
37. Maekawa, K., et al., *Four novel defective alleles and comprehensive haplotype analysis of CYP2C9 in Japanese.* Pharmacogenetics and Genomics, 2006. **16**: p. 497-514.
38. Shu, Y., et al., *Evolutionary conservation predicts function of variants of the human organic cation transporter, OCT1.* Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5902-7.
39. Limdi, N.A., et al., *Influence of CYP2C9 and VKORC1 on warfarin response during initiation of therapy.* Blood Cells Mol Dis, 2009. **43**(1): p. 119-28.
40. Limdi, N.A., et al., *Warfarin pharmacogenetics: a single VKORC1 polymorphism is predictive of dose across 3 racial groups.* Blood, 2010. **115**(18): p. 3827-34.
41. Scott, S.A., et al., *Warfarin pharmacogenetics: CYP2C9 and VKORC1 genotypes predict different sensitivity and resistance frequencies in the Ashkenazi and Sephardi Jewish populations.* Am J Hum Genet, 2008. **82**(2): p. 495-500.
42. Alsmadi, O., et al., *Genome at juncture of early human migration: a systematic analysis of two whole genomes and thirteen exomes from Kwaiti Population subgroup of inferred Saudi Arabian tribe ancestry.* PLOS ONE, 2014. **9**(6): p. e99069.

43. Wadelius, M., et al., *Common VKORC1 and GGCX polymorphisms associated with warfarin dose.* Pharmacogenomics J, 2005. **5**(4): p. 262-70.
44. King, C.R., et al., *Gamma-glutamyl carboxylase and its influence on warfarin dose.* Thromb Haemost, 2010. **104**(4): p. 750-4.
45. Kimura, R., et al., *Genotypes of vitamin K epoxide reductase, gamma-glutamyl carboxylase, and cytochrome P450 2C9 as determinants of daily warfarin dose in Japanese patients.* Thromb Res, 2007. **120**(2): p. 181-6.
46. Vijayasarathy, C., et al., *Molecular mechanisms leading to null-protein product from retinoschisin (RS1) signal-sequence mutants in X-linked retinoschisis (XLRS) disease.* Hum Mutat, 2010. **31**(11): p. 1251-60.
47. Gannage-Yared, M.H., et al., *Exome sequencing reveals a mutation in DMP1 in a family with familial sclerosing bone dysplasia.* Bone, 2014. **68**: p. 142-5.
48. Farrow, E.G., et al., *Molecular analysis of DMP1 mutants causing autosomal recessive hypophosphatemic rickets.* Bone, 2009. **44**(2): p. 287-94.
49. Caridi, G., et al., *A novel mutation in the albumin gene (c.1A>C) resulting in analbuminemia.* Eur J Clin Invest, 2013. **43**(1): p. 72-8.
50. Dai, D.P., et al., *CYP2C9 polymorphism analysis in Han Chinese populations: building the largest allele frequency database.* Pharmacogenomics J, 2014. **14**(1): p. 85-92.
51. Wester, M.R., et al., *The structure of human cytochrome P450 2C9 complexed with flurbiprofen at 2.0-A resolution.* J Biol Chem, 2004. **279**(34): p. 35630-7.
52. Niinuma, Y., et al., *Functional characterization of 32 CYP2C9 allelic variants.* Pharmacogenomics J, 2014. **14**(2): p. 107-14.
53. Roth, J.A., et al., *Genetic risk factors for major bleeding in patients treated with warfarin in a community setting.* Clin Pharmacol Ther, 2014. **95**(6): p. 636-43.
54. Consortium, I.W.P., *Estimation of the warfarin dose with clinical and pharmacogenetic data.* N Engl J Med, 2009. **360**: p. 753-764.
55. Rieder, M.J., et al., *Effect of VKORC1 Haplotypes on Transcriptional Regulation and Warfarin Dose.* New England Journal of Medicine, 2005. **352**(22).
56. Nahar, R., et al., *CYP2C9, VKORC1, CYP4F2, ABCB1 and F5 variants: influence on quality of long-term anticoagulation.* Pharmacol Rep, 2014. **66**(2): p. 243-9.
57. Edson, K.Z., et al., *Cytochrome P450-dependent catabolism of vitamin K: omega-hydroxylation catalyzed by human CYP4F2 and CYP4F11.* Biochemistry, 2013. **52**(46): p. 8276-85.
58. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.

CHAPTER 3: ASSOCIATIONS BETWEEN GENETICS, DIET, AGE, SEASON AND SERUM 25-HYDROXYLATED VITAMIN D_3 CONCENTRATION IN A YUP'IK STUDY POPULATION FROM SOUTHWESTERN ALASKA.

This work was previously submitted for publication at the *Journal of Nutrition:*

Fohner, et al.

3.1 Introduction

Vitamin D deficiency is linked to increased risk for multiple illnesses, including bone demineralization, rickets, multiple sclerosis, cardiovascular disease, colon cancer, some types of breast and prostate cancers, type 1 and type 2 diabetes, respiratory infections, influenza, active tuberculosis and depression [1-4]. The concentration of 25-hydroxy vitamin D [25(OH)D] in serum or plasma is used as the primary indicator of vitamin D sufficiency [5, 6]. Although there is some controversy [7], the Institute of Medicine considers a serum concentration of <12 ng/mL (30 nM) to be deficient, 12-20 ng/mL to be insufficient, and >20 ng/mL (50 nM) to be sufficient [8]. Sun exposure, diet, age, gender, body mass index (BMI), disease status, and use of some drugs have been associated with serum/plasma 25(OH)D concentration [5, 9]. Heritability of vitamin D concentrations has been estimated at 29-80%, with known genetic variants explaining 1 – 4% of that variation [10-14]. Specifically, common noncoding variants in *CYP2R1*, which encodes the enzyme that hydroxylates vitamin D into 25(OH)D, and in *DHCR7*, which encodes the enzyme that modulates the amount of vitamin D precursor in the skin for synthesis with sunlight, have been associated with serum 25(OH)D_3 concentration [1, 15, 16]. Indeed, a haplotype of

DHCR7 that is thought to reduce enzyme function has been found more commonly in northern latitudes and is suggested to confer an evolutionary advantage [16].

Vitamin D deficiency is increasingly prevalent among Alaska Native infants and children, with hospitalization due to rickets occurring more frequently than in the general US and at a rate of 2.23/100,000 children/year [17, 18]. Cancer is the leading cause of death among AN people [4, 19]. Colon cancer is of particular concern to the Yup'ik people in the Yukon-Kuskokwim Delta (Y-K Delta), and had an overall incidence in AN people of 102.6 per 100,000 for all years combined between 1999 and 2004 [20], with a relative rate of incidence of 2.03, compared to an overall incidence of 50.6 per 100,000 for non-hispanic white people living in the same regions [20]. Thus, vitamin D deficiency is a public health concern in Alaska.

For the Yup'ik people who live in the Y-K Delta of rural southwestern Alaska, adequate vitamin D may be obtained from the traditional diet, including fish, marine mammals, liver, and other organ meats [21]. In a study of Yup'ik communities in the Y-K Delta conducted between 2003 and 2005, average 25(OH)D_3 concentration was 2x the threshold for sufficiency and it was estimated that 90% of vitamin D came from traditional food sources [21]. However, reduced consumption of locally harvested foods may be leading to increased rates of 25(OH)D deficiency in these communities [21-23]. To better understand genetic, environmental, and demographic factors affecting circulating 25(OH)D_3 concentrations, we conducted a cross-sectional study of the population that included analysis of variants in *CYP2R1* and *DHCR7*, serum 25(OH)D_3 concentrations, and nitrogen isotope ratio in red blood cells (RBCs), a validated biomarker of the marine-based diet in the population [24, 25].

3.2 Methods

Research approval. This study emerged from a partnership between the University of Washington (UW), Center for Alaska Native Health Research (CANHR), Yukon Kuskokwim Health Corporation (YKHC), and Yup'ik communities in the Y-K Delta. The research questions were developed with Y-K communities in partnership with CANHR under Community-based Participatory Research (CBPR) guidelines [26]. Approval for research was received from the YKHC Executive Board of Directors and the University of Alaska Fairbanks IRB, in addition to UW IRB. This study was granted a Certificate of Confidentiality by the National Institute of General Medical Sciences to protect participant information.

Study population. The Y-K Delta in southwestern Alaska is home to approximately 23,000 people, 85% of whom are self-identified Alaska Native (AN), who live predominantly in remote communities and obtain healthcare through the YKHC [27]. A total of 743 male and female research participants, ≥ 14 years of age, were recruited for the study between September 2009 and December 2013 through written and oral advertisement. They represent convenience sampling from 10 communities in the Y-K Delta.

Sample collection and processing. Fasting venous blood samples were collected from each participant for isolation of red blood cells (RBCs), plasma, serum, and DNA. Blood was collected into silica-coated coated K2 EDTA tubes (BD Vacutainer®) and centrifuged (900 × *g*, 15 min) at room temperature. Buffy coats were incubated with Puregene RBC Lysis

Solution for 10 minutes, and centrifuged again (1800 × *g*, 10 min) at room temperature. White blood cells were then re-suspended in 10 mL Puregene Cell Lysis Solution until DNA purification using the Gentra Puregene kit (Qiagen, Valencia, California, USA). Serum was isolated from blood collected in a BD red-top Vacutainer and transferred to amber tubes for 25(OH)D_3 analysis. All samples collected in the field were stored in aliquots at -15^{o}C in a portable freezer while on site and then shipped to University of Alaska Fairbanks within 7 days and stored at -80^{o}C. Isolated DNA and serum were subsequently sent to UW for genetic and vitamin D analysis.

DNA isolation and genetic analysis. The *CYP2R1* SNPs genotyped were rs2060793, rs10741657, rs1993116, and rs11023374. The 4 SNPs informing *DHCR7* haplotypes were rs12785878, rs3794060, rs12800438, and rs4944957. These SNPs were chosen based on previous association with 25(OH)D levels [11-14, 16].

For *CYP2R1* genotyping, DNA samples were pre-amplified according to Fluidigm's (South San Francisco, CA) Specific Target Amplification protocol to increase template DNA for genotyping. TaqMan SNP Genotyping Assays (Applied Biosystems, Inc.) were run on 96.96 Dynamic Genotyping Arrays (Fluidigm) according to the manufacturer's established protocol for BioMark™ 96.96 Genotyping. 96.96 Dynamic Arrays were primed and loaded on the Fluidigm HX and thermal cycled on the Fluidigm FC1 controller according to the manufacturer's pre-loaded profiles. End-point fluorescence was read on a BioMark™ Real-Time PCR System (Fluidigm) and analyzed using SNP Genotyping Analysis software (Fluidigm).

For determining *DHCR7* haplotypes, DNA samples were genotyped using pre-designed 5'-nuclease SNP Genotyping Assays (Applied Biosystems /Life Technologies, Foster City, CA), which employ specific fluorogenic probes. The fluorescent 5'-nuclease assays were performed and analyzed on an ABI 7900HT Fast Real-Time PCR System (Applied Biosystems). The specific PCR reaction conditions were based on the general guidelines provided by the manufacturer and incorporated 10-25ng of genomic DNA template. Thermocycling parameters consisted of an initial incubation at 95°C for 10 minutes, followed by 40 cycles of 92°C for 15 seconds and 60°C for 1 minute.

Pairwise r^2 linkage disequilibrium (LD) patterns between the *CYP2R1* SNPs and the *DHCR7* SNPs were calculated using Haploview 4.2 software [28]. All SNPs were tested for deviations from Hardy–Weinberg equilibrium (HWE) using a χ^2-test, with a significance level of 0.05. To calculate population frequencies of the SNPs in *CYP2R1* and *DHCR7,* genealogic information contained in pedigrees were used to create a matrix of pairwise kinship coefficients between each participant, as described in Bourgain [29]. Estimates of population frequency and variance with adjustment for relatedness among sample individuals were obtained using the best linear unbiased estimator (BLUE) for allele frequency [30]. The statistical analysis for allele frequency estimation was performed using R statistical computing language [31].

Red blood cell nitrogen isotope ratio analysis. RBC nitrogen isotope ratios were prepared as described in O'Brien [24] and analyzed at the Alaska Stable Isotope Facility at UAF by continuous flow isotope ratio mass spectrometry, using a Costech ECS4010 Elemental Analyzer (Costech Scientific Inc., Valencia, CA) interfaced with a Delta V Plus isotope ratio mass spectrometer (Thermo Fisher Scientific, Inc., Waltham, MA, USA) via the Conflo IV Interface (Thermo Fisher Scientific, Inc., Waltham, MA, USA). Nitrogen isotope ratios are represented in delta notation as 'permil' abundance of heavy isotope relative to an international standard: [δ^{15}N value = (R_{sample} - $R_{standard}$)/$R_{standard}$] × 1000‰, where R is the ratio of heavy to light isotope, and the standard is atmospheric nitrogen. The RBC δ^{15}N value is a validated biomarker of RBC w-3 polyunsaturated fatty acid (PUFA) intake, because both δ^{15}N values and PUFAs are elevated in the fish and marine mammals that are a staple of the traditional Yup'ik diet [24, 25].

Serum 25(OH)D$_3$ measurement. The total (unbound and bound) concentrations of 25(OH)D_3 in serum samples were determined following a validated Liquid-Chromatography, Tandem Mass Spectrometry (LC-MS/MS) assay, as described in Wang [32]. Analytical standards were compared to reference 25(OH)D_3 National Institute of Standards and Technology standards [33] and found to be within 15% of the reference concentration across the standard curve range.

Preliminary analysis and assessment of non-normality of samples. The distributions of both 25(OH)D_3 and δ^{15}N values were assessed for normality. Log transformation of δ^{15}N values but not 25(OH)D_3 improved the normality of the distribution. Accordingly, 25(OH)D_3 concentrations were not transformed and δ^{15}N values were log transformed for further analysis (**Supplemental Figure 1**).

Regression of serum 25(OH)D_3 concentration against each genotype at each SNP was performed to determine which, if any, *CYP2R1* or *DHCR7* SNPs should be included in association analysis as an additive or recessive variant model. Each SNP was evaluated separately in a linear regression model and by an ANOVA test with 25(OH)D_3 concentration. Genotypes were also tested in each season to assess any seasonal variation in genetic contribution. Only the genotype at rs11023374 in *CYP2R1* was included for analysis, following a recessive inheritance model of the variant allele. None of the other SNPs were found to be significantly associated with 25(OH)D_3 concentration overall or by season.

Kinship adjustment and summary statistics. A kinship correlation matrix based on pairwise familial relationships was used to adjust for correlated 25(OH)D_3 concentration values among sample individuals in an association analysis with a linear mixed effects (LME) model. The pedigree information on the Yup'ik participants was used to create a kinship correlation matrix with the coxme R package [34], and the kinship matrix was incorporated in an LME regression analysis that was performed using the lmekin function from the same package.

Correlations with serum 25(OH)D_3 concentration. To determine associations with serum 25(OH)D_3 concentration, a maximum likelihood mixed effect multiple linear regression analysis was performed. To avoid spurious associations, the kinship correlation matrix was included to account for the random effects of the non-independence of samples resulting from any familial relationships. The fixed effects of the model included cofactors found to be significantly associated with 25(OH)D_3 concentration in previous studies [15, 21, 35]. These included age (continuous), gender (binary), BMI (continuous), yearly quarter of sample collection (factored), recessive *CYP2R1* rs11023374 genotype (binary), and $\log_{10}(\delta^{15}N$ value) (continuous). Inland or coastal geography of each community was also included as a binary variable. The previously mentioned lmekin function in the coxme R package was used to evaluate the mixed effects model using maximum likelihood analysis [34]. For this exploratory analysis, covariates were determined as predictors of 25(OH)D_3 concentration if the significance of its coefficient surpassed $p < 0.05$. Samples missing data were excluded.

Sinusoidal model analysis. A sinusoidal model previously developed to fit the seasonal pattern of 25(OH)D_3 concentration [36, 37] was used to model 25(OH)D_3 concentration over the course of the year. The model was developed in a population of people ages 45-84 years living in 6 communities across the lower 48 states of the USA [36]. The model was fit to the overall Yup'ik data and adjusted for age. It was also fit to subsets including 1) younger and older participants split at the median (33 years), 2) male and female, 3) $\log_{10}(\delta^{15}N$ value) split at the median (0.927), 4) genotype at rs11023374, and 5) coastal or inland location of community. Sinusoidal model analysis was performed with the cosinor.lm function of the cosinor R package [38].

Linear regression with subset of unrelated participants. A subset of 526 unrelated participants was selected by removing individuals from the kinship matrix who were related to others by the 3rd degree or closer. Simple linear regression with only these participants was performed using R [31] to determine variability in 25(OH)D_3 concentrations attributable to covariates. Regression was also performed on subsets of participants, split at the median age. T-tests were used to compare the demographics of the unrelated subset to the full data set.

3.3 Results

Population demographics and summary statistics. The complete dataset included 743 individuals, whose demographics and summary statistics are described in **Table 3-1**. The distributions of 25(OH)D_3 concentrations and $\delta^{15}N$ values, both untransformed and transformed, are shown in **Figure 3-1**. Serum 25(OH)D_3 concentrations ranged from 6.0 ng/mL to 68.7 ng/mL.

Adjusted statistics and analyses. Summary statistics of the sample, adjusted for kinship coefficients between participants, are presented in **Table 3-2**. Overall, participants younger than 33 years old had significantly lower BMI, $\log_{10}$($\delta^{15}N$ value), and 25(OH)D_3 concentrations than participants age 33 years and older ($p < 0.05$). The same was true of males compared to females. Concentration of 25(OH)D_3 and $\log_{10}$($\delta^{15}N$ value) by decade of age are presented in **Figure 3-2**. Mean concentration of 25(OH)D_3 and $\log_{10}$($\delta^{15}N$ value) by season are presented in **Figure 3-3**, showing significant differences in 25(OH)D_3 concentration but no variation in $\log_{10}$($\delta^{15}N$ value) by season. The correlation between $\log_{10}$($\delta^{15}N$ value) and 25(OH)D_3 concentration is presented in **Figure 3-4**, illustrating a correlation of 0.24 between the 2 metrics.

Overall, 22.9% of participants had 25(OH)D_3 concentrations below the 20 ng/mL threshold for insufficiency, as defined by the Institute of Medicine. Among the younger half of participants, 38.1% were insufficient, whereas among the older half of participants, 8.0% were insufficient.

CYP2R1 and DHCR7 population genotyping. The frequencies of minor alleles at each SNP in this Yup'ik study population are listed in **Table 3-3**, adjusted for kinship between study participants. Association of each SNP with 25(OH)D_3 concentration is also shown. Only the SNP rs11023374 in *CYP2R1* was significantly associated with 25(OH)D_3 concentrations (p=0.009). All SNPs were in Hardy-Weinberg equilibrium.

Three SNPs in *CYP2R1*, rs10741657, rs2060793, and rs1993116 were in high LD in this Yup'ik study population (r^2 = 0.97). The same 3 SNPs were in moderate LD with the fourth SNP, rs11023374 (r^2 = 0.35) (**Figure 3-5a**). A LD block reported for other populations [16] was also seen with the 4 SNPs informing the *DHCR7* haplotypes; r^2 between 0.94 and 0.97 for all pairwise comparisons of all 4 SNPs (**Figure 3-5b**). The haplotype associated with lower DHCR7 activity was identified at 48.3% in the sample (**Table 3-4**).

Modeled annual fluctuation in 25(OH)D_3 concentration. Based on a sinusoidal model describing the seasonal fluctuations in 25(OH)D_3 concentration [36], the average Yup'ik participant was predicted to have a mean 25(OH)D_3 concentration of 30.3 ng/mL (95% CI: 29.5 – 31.2) over the course of the year, with an amplitude of 7.97 ng/mL (95% CI: 6.27 – 9.66). This model predicted an average minimum concentration of 22.4 ng/mL and an average maximum concentration of 38.3 ng/mL throughout the year. The R^2 of the fit was 0.11, indicating other significant sources of variation, but consistent with the fit found in populations used in development and validation of the model [37]. Including age as a covariate in the model improved the fit with $R^2 = 0.45$. Twenty-four samples were excluded from the model due to missing 25(OH)D_3 data.

Older age, as evaluated by splitting the dataset at the median (33 years) (**Figure 3-6**), and higher $\log_{10}(\delta^{15}N$ value), as evaluated by splitting the dataset at the median (0.927), were both significantly associated with higher 25(OH)D_3 concentration over the course of the year ($p < 0.001$ for age split; $p < 0.001$ for $\log_{10}(\delta^{15}N$ value) split) in a model not adjusting for the other covariates. The predicted annual mean 25(OH)D_3 concentration for a younger member of the Yup'ik population was 24.0 ng/mL (95% CI: 23.0 – 25.0), with an amplitude of 8.19 ng/mL (95% CI: 6.09 – 10.3), whereas the predicted annual mean 25(OH)D_3 concentration for an older member of the population was 36.8 ng/mL (95% CI: 34.3 –39.3), with an amplitude of 5.74 ng/mL (95% CI: 3.87 – 7.61).

Age and $\log_{10}(\delta^{15}N$ value) were independently associated with the fit of the sinusoidal model, as both covariates were significant in the regression model that included both variables ($p < 0.001$) as predictors. Gender ($p = 0.10$) was not significantly associated with 25(OH)D_3 concentration, but recessive genotype at rs11027334 ($p = 0.03$) and coastal versus inland geographic location of the community ($p = 0.009$) were significantly associated with lower yearly average 25(OH)D_3 concentrations in the sinusoidal model.

Seasonal analysis. Participants were divided into seasonal quartiles, based on the distribution of 25(OH)D_3 concentration over the course of a year. Peak 25(OH)D_3 concentration occurred in September and the trough was in March, as determined from a fit of the sinusoidal model. This result yielded a seasonal breakdown of low 25(OH)D_3

concentration in February, March, and April; increasing concentration in May, June, and July; high concentration in August, September, and October; and decreasing concentration in November, December, and January. These divisions reflect the 20 day half-life of 25(OH)D, following seasonal patterns of sunlight exposure. On average, 25(OH)D_3 concentrations in older participants would not be expected to drop to insufficiency in the February-April trough, whereas 25(OH)D_3 concentrations in the average younger participant would be expected to do so following the nadir in sunlight exposure.

Associations of covariates with serum 25(OH)D_3 concentration. Based on the significance threshold of $p < 0.05$ in a linear mixed model regression, $\log_{10}(\delta^{15}N$ value), age, gender, BMI, community location, homozygosity of the variant allele at *CYP2R1* rs11023374, and season of blood draw were all significantly associated with 25(OH)D_3 concentration after accounting for the kinship coefficients between participants (**Table 3-5**). Based on proportional variability explained by the kinship matrix, heritability of 25(OH)D_3 concentration in these study participants was estimated to be 0.46 after adjusting for age, diet, BMI, season, and community effects.

Associations with 25(OH)D_3 concentration based on maximum unrelated subjects. In multiple linear regression including only the 526 unrelated participants, age, $\log_{10}(\delta^{15}N$ value), season, gender, BMI, community location, and genotype at *CYP2R1* rs11023374 were found to be significantly associated with 25(OH)D_3 concentration. These unrelated participants are representative of the entire dataset (**Table 3-6**). Overall, 45 samples were excluded due to missing data on 25(OH)D_3 concentration (24 samples) and/or dietary marker (42 samples) and/or genetic data (22 samples). All together, the variables included in the regression explained 52.8% of the variability in 25(OH)D_3 concentration of this sample set, with season, age, and $\log_{10}(\delta^{15}N$ value) explaining 45.2% of variability. Considered individually, season of blood draw accounted for 9.1%, age accounted for 36.5%, and $\log_{10}(\delta^{15}N$ value) accounted for 20.5% of variability (**Table 3-7**). In a combined analysis, age and $\log_{10}(\delta^{15}N$ value) accounted for 38.6% of variability in 25(OH)D_3.

In the younger half of participants, differences in the dietary marker accounted for 1.4% of variability in 25(OH)D_3 concentrations and differences in season of sample collection accounted for 13.0% (Table 3-7). In the older half of participants, differences in dietary marker accounted for 14.0% of variability in 25(OH)D_3 concentrations and differences in season of sample collection accounted for 6.4%.

3.4 Discussion

We characterized the serum 25(OH)D_3 concentration in a cross-section of Yup'ik people living in the Y-K Delta to better understand the role of genetics, diet and sun exposure in determining Vitamin D levels in that population. The average 25(OH)D_3 concentration among the Yup'ik study participants (31.0 +/- 0.1 ng/mL) was higher than the average reported for many populations [39], including healthy individuals living at lower latitudes (25.1 +/- 0.4 ng/mL) [37], who would arguably have greater opportunity to synthesize vitamin D from sunlight throughout the year. Greater intake of traditional foods rich in fats from marine mammals and fish, as measured by $\log_{10}(\delta^{15}N$ value), was associated with higher 25(OH)D_3 concentration in the Yup'ik study population. Not surprisingly, diet appears to be a more important source of vitamin D than sunlight in the

Yup'ik population, with older participants consuming more of the traditional foods rich in vitamin D [40]. Although the direction of the association varies by region, serum 25(OH)D_3 concentration has been associated with age in other populations and has been linked to age-related lifestyle differences that affect the amount of time outside and exposed to sunlight [39]. Because the association with age is found in the Yup'ik study participants even after adjusting for diet, there are likely age-related differences in lifestyle and variability in sources of vitamin D that are not being captured in this study.

While SNPs in *CYP2R1* have been associated with 25(OH)D_3 concentrations in other populations, only rs11023374 was associated with 25(OH)D_3 concentrations in the Yup'ik study participants. The SNP rs11023374 was in weaker LD compared to the other SNPs, suggesting that it may be more closely linked with the causal variant. The strong LD patterns between the other SNPs are similar to what has been found in other populations [1, 14]. Even so, genotype at rs11023374 accounted for only 1% of variability in 25(OH)D_3 concentrations. Furthermore, the *DHCR7* haplotype associated with increased production of vitamin D from sunlight was found at a lower frequency in the Yup'ik participants than in other populations living at northern latitudes [16], and none of the SNPs informing *DHCR7* haplotypes was associated with 25(OH)D_3 concentration. These results suggest that the haplotype does not confer a particular advantage for maintaining vitamin D concentrations in this population and that dietary sources were sufficient historically. Although the contribution of these 8 genetic variants was found to be minor, the heritability estimate of 0.46 for 25(OH)D_3 concentration suggests that uncharacterized genetic variation should still be considered for associations with disease risk and vitamin D status. While the estimate must be interpreted with caution, as it does not account for the shared environment of participants, it suggests that *CYP2R1, DHCR7*, and other candidate genes in the vitamin D pathway, such as for the vitamin D binding protein [11], should be sequenced to identify any novel variation or LD patterns in the Yup'ik population.

Consistent with the results of the regression analysis, modeling results indicated that greater consumption of the traditional diet provides sufficient vitamin D concentrations over the entire year. While the specific source may vary by season, traditional marine foods are available year round, either fresh, frozen or dried, as indicated by the consistency of the dietary marker throughout the year. However, the traditional diet of fish, marine mammals, birds, land mammals, and berries is being replaced by nutrient-poor imported foods in many indigenous arctic communities, especially among youth and young adults, and this transition has been linked with lower vitamin D concentrations [4, 22, 25, 40]. Because of lower consumption of traditional Yup'ik subsistence foods, younger people are at greater risk of being vitamin D deficient during the winter months (February – April) and of having more variable $25(OH)D_3$ concentration with changes in season. This deviation from the traditional diet among the younger demographic is prompting public health concern, as it increases the risk for illnesses associated with vitamin D deficiency. Future studies are needed to understand the health impacts of these dietary patterns and low vitamin D concentrations in Yup'ik communities.

Vitamin D concentrations differed significantly by gender and community location, suggesting demographic differences in vitamin D sources. In contrast to findings from other populations [41], female Yup'ik participants had higher $25(OH)D_3$ concentrations compared to males, likely reflecting differences in traditional food intake estimated from $\delta^{15}N$ values [25]. After adjusting for diet, however, men had higher concentrations of

$25(OH)D_3$, likely because of spending more time outside. Whereas $\delta^{15}N$ values were higher in coastal communities, suggesting greater intake of traditional marine foods, participants from inland communities had higher concentrations of $25(OH)D_3$. One possible explanation is that certain foods have higher vitamin D content in relation to $\delta^{15}N$ value and are eaten more frequently in inland communities.

Strengths of this study are that it combined demographic, genetic, seasonal, and dietary measures to more completely characterize sources of variability in $25(OH)D_3$ concentrations and risk of vitamin D deficiency in a population with highly variable serum $25(OH)D_3$ concentrations and experiencing a transition in dietary patterns. Moreover, it used an objective measurement of dietary intake of vitamin D and sampled concentrations throughout the year in a region experiencing drastic changes in sunlight availability. A limitation of this study is that it did not evaluate sources of vitamin D_2, such as supplementation in market foods, which would contribute to overall vitamin D status and to the risk of deficiency and insufficiency. Another important limitation of these data is that they are cross-sectional. A longitudinal study is needed to confirm the seasonal variation in $25(OH)D_3$ concentration within individuals, although this variation would be expected based on studies in other populations [36, 37]. Furthermore, due to convenience sampling, the dataset contains an abundance of samples collected in transitional seasons between maximum and minimum annual vitamin D concentrations (May – June and November – January), and has under-sampled the low season (February – April), especially. As a result, the average $25(OH)D_3$ concentrations may be skewed high.

In summary, average serum $25(OH)D_3$ concentrations were high relative to other populations and were highly correlated with dietary sources of vitamin D; however, they were also highly variable, with a larger portion of younger people insufficient and deficient for vitamin D concentrations compared to older participants. A public health campaign to raise awareness about potential health benefits of a traditional diet could possibly raise average serum 25(OH)D_3 concentrations and maintain stability during periods of low sunlight exposure. Efforts to increase serum 25(OH)D_3 concentrations could also help reverse recent increases in the incidence of vitamin D deficiency, such as among young pregnant women in AN communities, which has been associated with an increase in illnesses associated with vitamin D deficiency [17]. While supplementation and fortification of foods with vitamin D may reduce the risk for vitamin D deficiency in the Yup'ik population [4], promotion of a traditional diet could have positive health impacts that go beyond raising vitamin D concentrations in these communities.

3.5 Figures

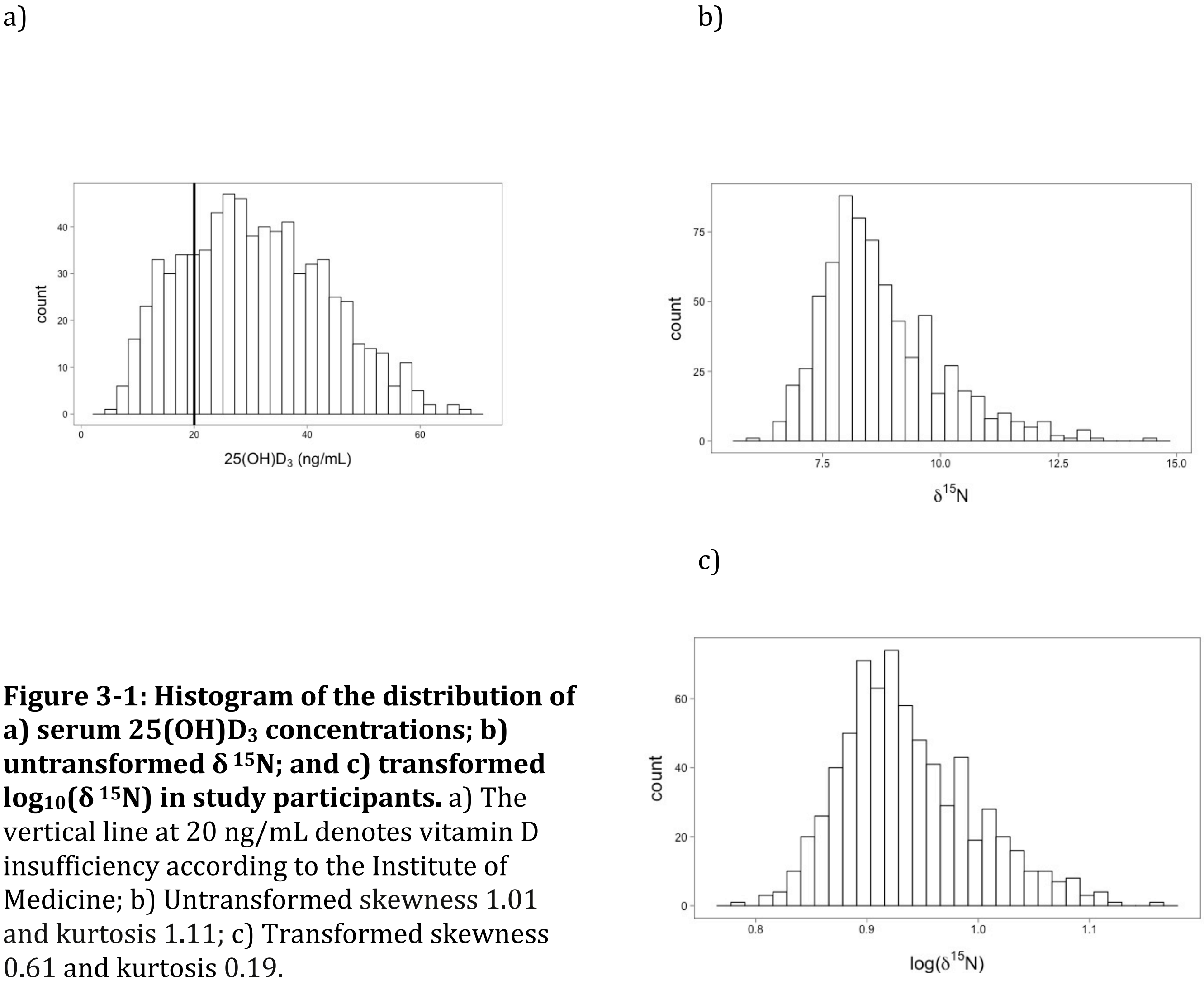


Figure 3-1: Histogram of the distribution of a) serum 25(OH)D_3 concentrations; b) untransformed δ ^{15}N; and c) transformed $\log_{10}$(δ ^{15}N) in study participants. a) The vertical line at 20 ng/mL denotes vitamin D insufficiency according to the Institute of Medicine; b) Untransformed skewness 1.01 and kurtosis 1.11; c) Transformed skewness 0.61 and kurtosis 0.19.

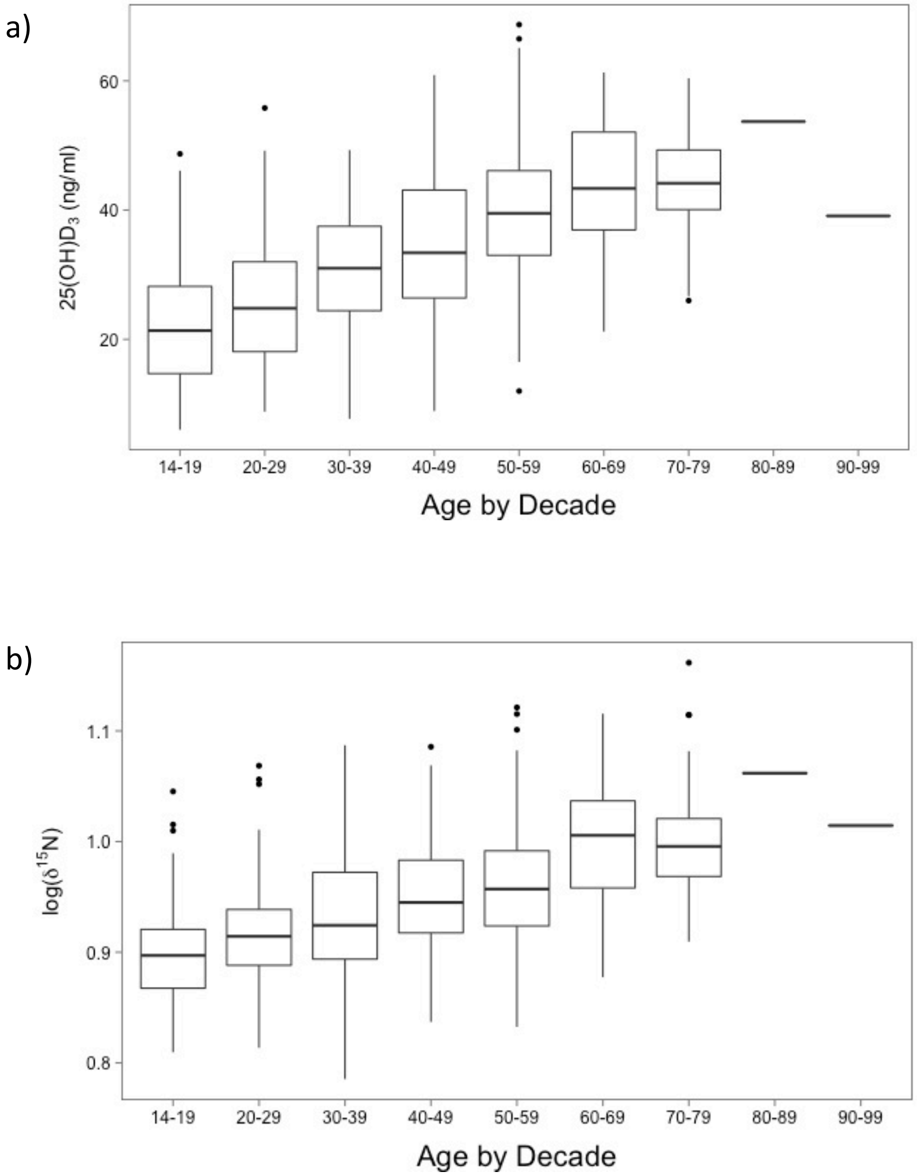
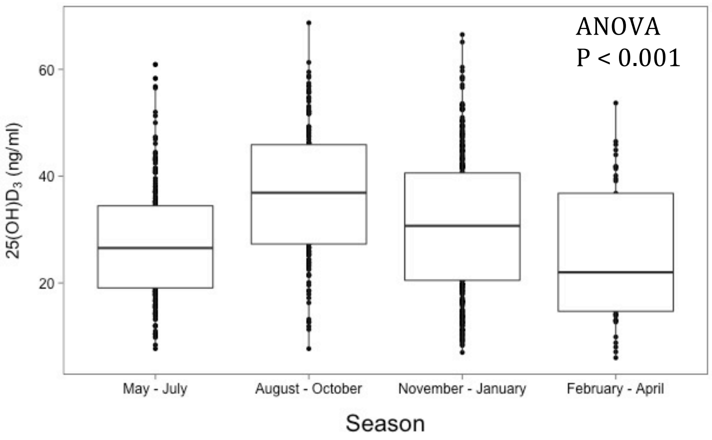


Figure 3-2: Concentration of a) $25(OH)D_3$ and b) $\log_{10}(\delta^{15}N$ value) by age of study participants, stratified by decade. As age increases, so does mean $25(OH)D_3$ concentration and mean $\log_{10}(\delta^{15}N$ value).

a)



b)

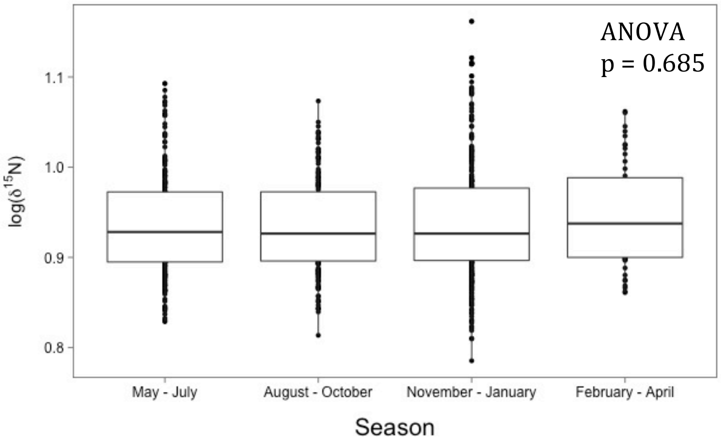


Figure 3-3: Seasonal comparisons of mean a) 25(OH)D_3 concentrations; and b) log_{10}(δ ^{15}N).

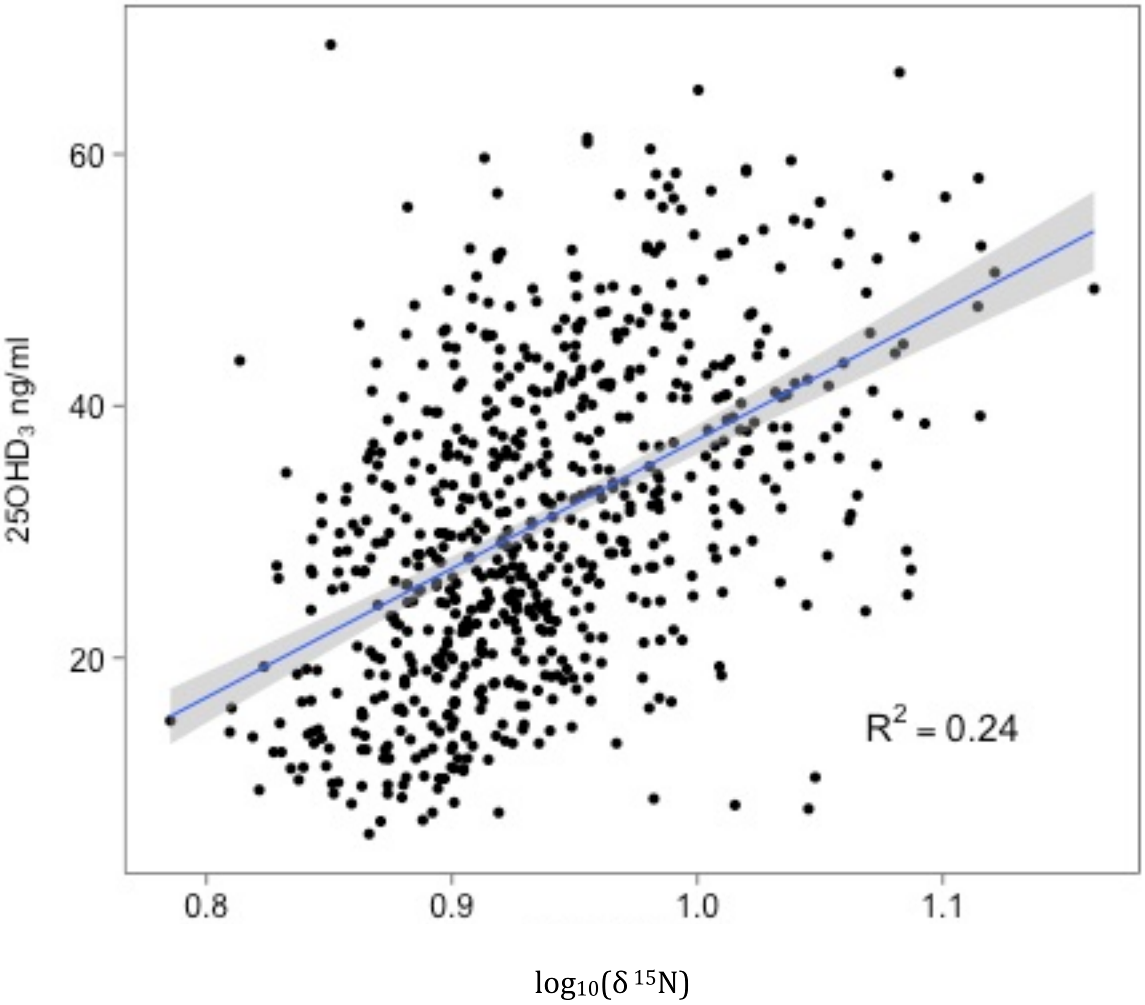
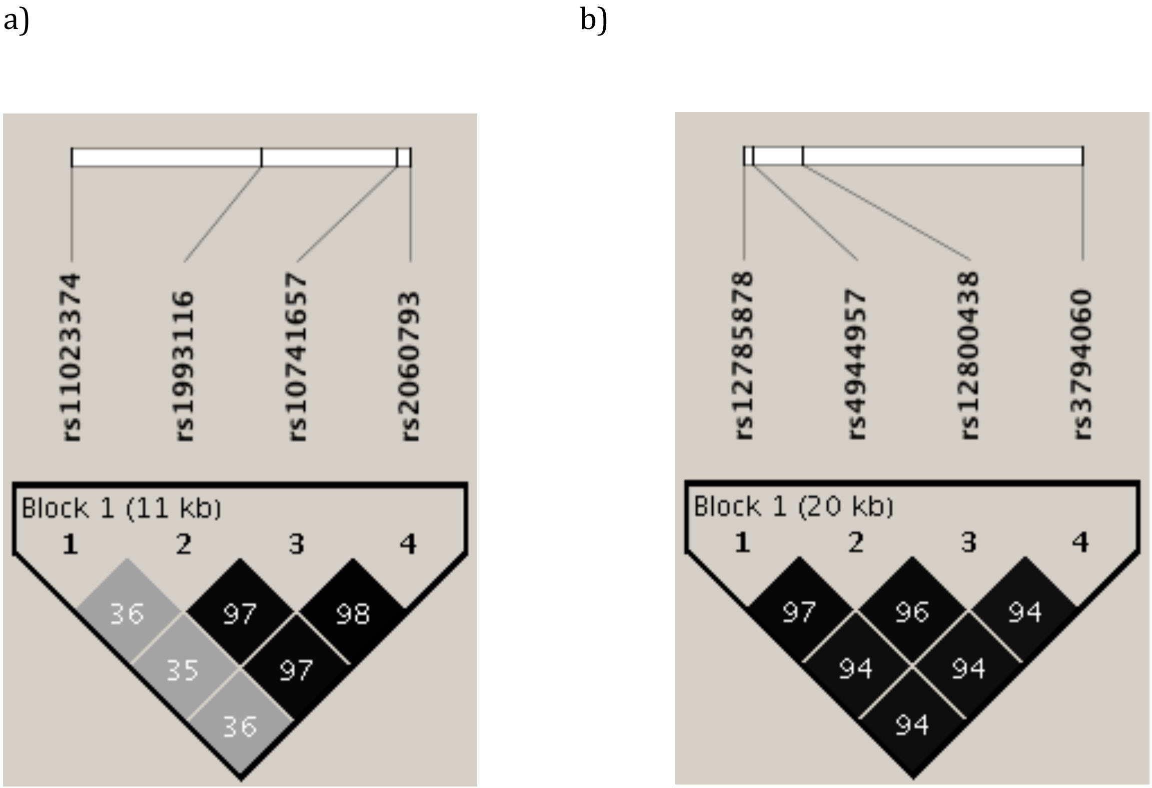


Figure 3-4. Correlation of the concentration of 25(OH)D_3 with the log of the Nitrogen Isotope Ratio. As the Nitrogen Isotope Ratio increases, so does 25(OH)D_3 concentration.

Figure 3-5: Linkage disequilibrium (LD) pattern of the SNPs genotyped in a) *CYP2R1*; and b) *DHCR7*, by r^2 measure. a) The 3 SNPs in *CYP2R1,* rs10741657, rs2060793, and rs1993116 are in nearly complete LD by both measures. All 3 are in moderate LD with the 4th SNP, rs11023374; b) The 4 SNPs in *DHCR7* are in tight LD, preserving the haplotype structure seen in other populations.



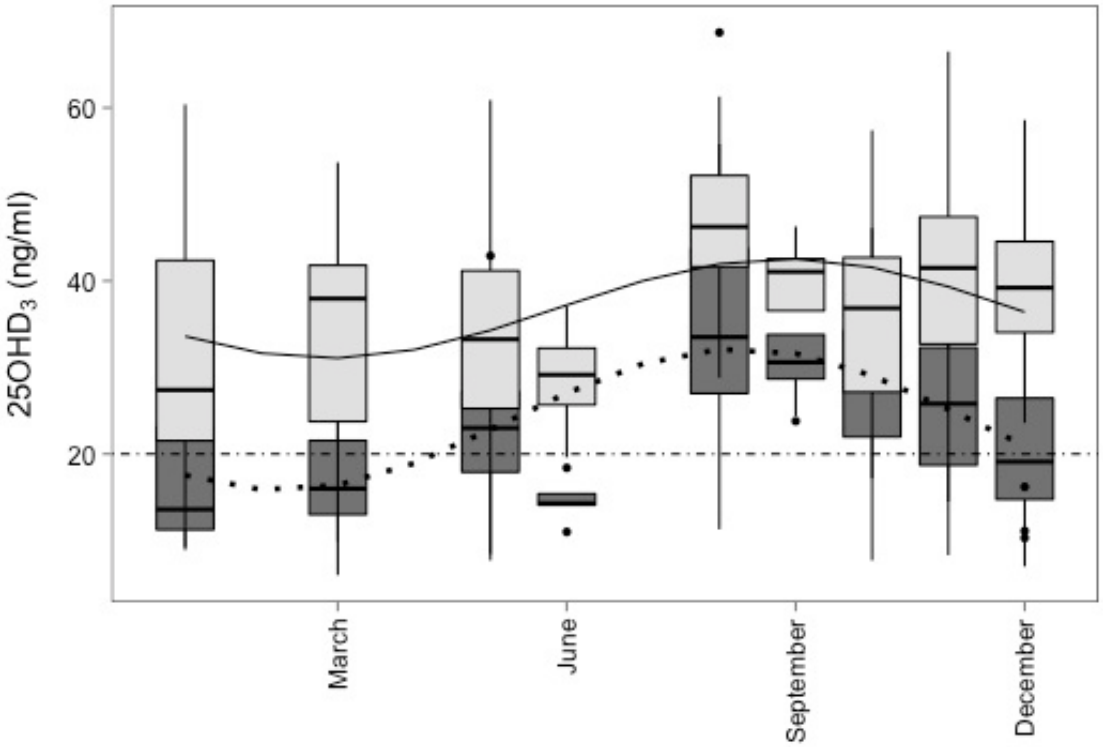


Figure 3-6: Sinusoidal model of variation in 25(OH)D$_3$ concentration in study participants over the course of the year. Participants are stratified by age into 2 groups divided at the median age to illustrate higher average 25(OH)D$_3$ concentration among older participants. The boxes represent the distribution of 25(OH)D$_3$ concentration during each month of sample collection, with younger participants represented by dark grey and older participants represented by light grey boxes. Sinusoidal lines represent the best fit of the model to the respective data sets, with the younger half of participants depicted by a dotted line and older participants by a solid line. The dashed horizontal line at 20 ng/mL indicates the Institute of Medicine threshold for vitamin D insufficiency.

3.6 Tables

Table 3-1: Participant demographics stratified by quarter of data collection. Study participants all self-identify as Yup'ik. Standard deviation (sd) is indicated in parentheses.

Time of Collection	Number	Male	Female	Mean Age (years) (sd)	Mean BMI (sd)	Mean $\log_{10}(\delta\ ^{15}N$ value) (sd)	Mean 25(OH)D_3 (ng/mL) (sd)	Number from Inland Communities	Number from Coastal Communities
Aug-Oct	167	79	88	40.9 (20.2)	25.8 (5.3)	0.934 (0.053)	36.7 (12.1)	117	50
Nov-Jan	310	171	139	35.5 (17.1)	26.2 (5.8)	0.939 (0.066)	30.9 (13.0)	161	149
Feb-April	48	19	29	35.1 (20.4)	26.3 (6.0)	0.945 (0.058)	25.6 (12.9)	0	48
May-July	218	121	97	34.7 (16.5)	26.6 (6.2)	0.935 (0.057)	27.4 (10.6)	63	155
Total	743	390	353	36.4 (18.0)	26.2 (5.8)	0.937 (0.060)	30.8 (12.6)	341	402

Table 3-2: Summary of demographics adjusted for correlation between participants using mixed effect model linear regression. Z-scores and p-values show significance of differences by each measure, stratified by age groups or gender.

	Number	**Mean Age (years)**	**Age z-score and p-value**	**Mean BMI**	**BMI z-score and p-value**	**Mean $\log_{10}(\delta\ ^{15}N)$**	**$\log_{10}(\delta\ ^{15}N$ value) z-score and p-value**	**Mean $25(OH)D_3$**	**$25(OH)D_3$ z-score and p-value**
<33	369	20.9	z = 45.17	24.3	z = 9.25	0.907	z = 16.49	24.5	z = 18.16
≥33	374	51.8	p < 0.001	28.0	p < 0.001	0.966	p < 0.001	37.7	p < 0.001
Male	390	34.4	z = -3.18	24.9	z = -6.79	0.919	z = -8.88	30.1	z = -2.41
Female	353	38.6	p = 0.002	27.7	p <0.001	0.957	p < 0.001	32.3	p = 0.016
Overall (standard error)	743	36.4 (0.7)		26.2 (0.2)		0.937 (0.002)		31.1 (0.5)	

Table 3-3: Minor allele frequencies at each SNP in *CYP2R1* and *DHCR7*, and significance of association with 25(OH)D_3 concentration.

SNP	Major/Minor Allele	Minor Freq (95% CI)	ANOVA with 25(OH)D_3
CYP2R1			
rs10741657	A/G	39.10% (36.54 – 41.66%)	0.166
r2060793	A/G	39.02% (36.45 – 41.59%)	0.269
rs11023374	T/C	38.68% (36.12 – 41.23%)	0.009**
rs1993116	A/G	19.25% (17.18 – 21.32%)	0.216
DHCR7			
rs12785878	G/T	49.04% (47.13 – 50.95%)	0.311
rs3794060	C/T	48.53% (46.62 – 50.44%)	0.402
rs12800438	G/A	48.81% (46.89 – 50.72%)	0.257
rs4944957	A/G	49.27% (47.34 – 51.20%)	0.333

**: indicates statistical significance, based on a $p < 0.05$ threshold.

Table 3-4: The frequencies of haplotypes of *DHCR7* identified in Yup'ik participants. The order of SNPs is rs12785878, rs4944957, rs12800438, rs3794060.

Haplotype	**Frequency**
TGAT*	0.483
GAGC	0.494
GGAC	0.005
GAAC	0.005
GAGT	0.005
TGGC	0.004
TAGC	0.002
TAGT	0.002

*Associated with greater vitamin D production

(Kuan V, Martineau AR, Griffiths CJ, Hypponen E, Walton R. DHCR7 mutation linked to higher vitamin D status allowed early human migration to Northern latitudes. Evolutionary Biology 2013;13).

Table 3-5: Linear mixed effects model regression coefficients and significance of the contribution of each variable to serum 25(OH)D_3 concentration. The model included all of the variables in the table. Significance was set as $p < 0.05$. Standard error (se) is indicated in parentheses. The reference category (ref) is indicated for all discrete variables.

Characteristic	**N**	**B coefficient from full model (*se*)**	**P value from full model**
Fully adjusted model	**743**		
Age		0.268 (0.023)	<0.001
Younger than 33	369		
Older than 33	374		
Season			
May-July	305	-3.906 (0.994)	<0.001
Aug-Oct	167	ref	
Nov – Jan	218	-2.751 (0.886)	0.002
Feb-April	49	-5.219 (1.621)	0.001
$\text{Log}_{10}(\delta\ ^{15}\text{N value})$		85.070 (7.423)	<0.001
Gender			
Male	390	1.905 (0.687)	0.006
Female	353	ref	
Community location			
Coastal	402	ref	
Inland	341	7.654 (0.794)	<0.001
BMI		-0.209 (0.0589)	<0.001
CYP2R1 rs11023374			
Homozygous variant	32	-3.900 (1.623)	0.016
At least one reference allele	207	ref	

Table 3-6: Demographics of unrelated subset of study participants compared to full dataset.

								Number in Each Season (% total)			
	Total	Male (% total)	Female (% total)	25(OH)D	Age	BMI	Log$_{10}$(δ^{15}N)	Aug-Oct	Feb-April	Nov-Jan	May-July
Unrelated subset	526	282 (53.6%)	244 (46.4%)	31.0	36.1	26.1	0.936	136 (25.9%)	33 (6.3%)	207 (39.3%)	150 (28.5%)
All	743	390 (52.5%)	353 (47.5%)	30.8	36.4	26.2	0.937	167 (22.5%)	48 (6.5%)	310 (41.7%)	218 (29.3%)
t.test of comparison unrelated subset to all participants				p = 0.78	p = 0.77	p = 0.76	p = 0.77				

Table 3-7: Variability explained in multiple linear regression with the subset of unrelated participants (n = 526), considered together and stratified by older and younger participants. Regressions included the full model of all variables, each variable alone, and age and $\log_{10}(\delta^{15}N)$ together. The significance of the association of each variable with $25(OH)D_3$ concentration is indicated by the p-value. The goodness of fit is reflected in the R^2 of the regression, on a scale of 0 to 1, and indicated in parentheses. The left side of the table shows regression with all participants combined. The right side of the table shows regression in the younger half of participants and older half of participants, split at 33 years.

Characteristic	**N**	**Significance in full model**	**Variability explained (R^2)**	**Younger Subset: significance in full model and (R^2)**	**Older subset: significance in full model and (R^2)**
Fully adjusted model	**526**	**$p < 0.001$**	**(0.528)**	**(0.287)**	**(0.417)**
Age		**$p < 0.001$**	**(0.365)**	**$p < 0.001$ (0.010)**	**$p < 0.001$ (0.161)**
Younger than 33	272				
Older than 33	254				
Season			**(0.091)**	**(0.130)**	**(0.064)**
May-July	150	$p < 0.001$		$p < 0.001$	$p = 0.089$
Aug-Oct	136	reference			
Nov – Jan	207	$p < 0.001$		$p < 0.001$	$p = 0.293$
Feb-April	33	$p < 0.001$		$p < 0.001$	$p = 0.948$
$\text{Log}_{10}(\delta^{15}N$ value)		**$p < 0.001$**	**(0.205)**	**$p < 0.001$ (0.014)**	**$p < 0.001$ (0.141)**
Gender		**$p = 0.007$**	**(0.00)**	**$p = 0.005$**	**$p = 0.251$**
Male	282				
Female	244				
Community location		**$p < 0.001$**	**(0.063)**	**$p < 0.001$ (0.072)**	**$p < 0.001$ (0.070)**
Coastal	270				
Inland	256				
BMI		**$p = 0.041$**	**(0.006)**	**$p = 0.014$**	**$p = 0.161$**
CYP2R1 rs11023374		**$p = 0.016$**	**(0.011)**	**$p = 0.362$**	**$p = 0.033$**
Homozygous variant	23				
At least one reference allele	503				
Age and $\log_{10}(\delta^{15}N$ value)			**(0.386)**	**(0.038)**	**(0.212)**

3.7 References

1. Ramos-Lopez, E., et al., *CYP2R1 (vitamin D 25-hydroxylase) gene is associated with susceptibility to type 1 diabetes and vitamin D levels in Germans.* Diabetes/Metabolism Research and Reviews, 2007. **23**: p. 631-636.
2. Levin, G.P., et al., *Genetic variants and associations of 25-hyroxyvitamin D concentrations with major clinical outcomes.* JAMA, 2012. **308**(18): p. 1898-1905.
3. Zhang, Z., et al., *An analysis of the association between the vitamin D pathway and serum 25-hydroxyvitamin D levels in a healthy Chinese population.* J Bone Miner Res, 2013. **28**(8): p. 1784-92.
4. Sharma, S., et al., *Vitamin D deficiency and disease risk among aboriginal Arctic populations.* Nutr Rev, 2011. **69**(8): p. 468-78.
5. Berry, D. and E. Hypponen, *Determinants of vitamin D status: focus on genetic variations.* Current Opinion in Nephrology and Hypertension, 2011. **20**: p. 331-336.
6. Lehmann, B. and M. Meurer, *Vitamin D metabolism.* Dermatologic Therapy, 2010. **23**: p. 2-12.
7. Holick, M.F., et al., *Evaluation, treatment, and prevention of vitamin D deficiency: an Endocrine Society clinical practice guideline.* J Clin Endocrinol Metab, 2011. **96**(7): p. 1911-30.
8. *Dietary Reference Intakes for Calcium and Vitamin D*, ed. A.C. Ross, et al. 2011: The National Academies Press.
9. Bolland, M.J., et al., *The effects of seasonal variation of 25-hydroxvitamin D and fat mass on a diagnosis of vitamin D sufficiency.* Am J Clin Nutr, 2007. **86**: p. 959-64.
10. Berry, D. and E. Hypponen, *Determinants of vitamin D status: focus on genetic variations.* Curr Opin Nephrol Hypertens, 2011. **20**(4): p. 331-6.
11. Ahn, J., et al., *Genome-wide association study of circulating vitamin D levels.* Hum Mol Genet, 2010. **19**(13): p. 2739-45.
12. Dastani, Z., R. Li, and B. Richards, *Genetic regulation of vitamin D levels.* Calcif Tissue Int, 2013. **92**(2): p. 106-17.
13. Hiraki, L.T., et al., *Exploring the genetic architecture of circulating 25-hydroxyvitamin D.* Genet Epidemiol, 2013. **37**(1): p. 92-8.
14. Wang, T.J., F. Zhang, and J.B. Richards, *Common genetic determinants of vitamin D insufficiency: a genome-wide association study.* Lancet, 2010.
15. Engelman, C.D., et al., *Vitamin D intake and season modify the effects of the GC and CYP2R1 genes on 25-hydroxyvitamin D concentrations.* J Nutr, 2013. **143**(1): p. 17-26.
16. Kuan, V., et al., *DHCR7 mutation linked to higher vitamin D status allowed early human migration to Northern latitudes.* Evolutionary Biology, 2013. **13**(144).
17. Singleton, R., et al., *Rickets and vitamin D deficiency in Alaska native children.* Journal of Pediatric Endocrinology and Metabolism, 2015. **0**(0).
18. Gessner, B.D., J. Plotnik, and P.T. Muth, *25-Hydroxyvitamin D levels among healthy children in Alaska.* The Journal of Pediatrics, 2003. **143**(4): p. 434-437.
19. Yin, L., et al., *Meta-analysis: longitudinal studies of serum vitamin D and colorectal cancer risk.* Aliment Pharmacol Ther, 2009. **30**(2): p. 113-25.
20. Perdue, D.G., et al., *Regional differences in colorectal cancer incidence, stage, and subsite among American Indians and Alaska Natives, 1999-2004.* Cancer, 2008. **113**(5 Suppl): p. 1179-90.

21. Luick, B., A. Bersamin, and J.S. Stern, *Locally harvested foods support serum 25-hydroxyvitamin D sufficiency in an indigenous population of Western Alaska.* Int J Circumpolar Health, 2014. **73**.
22. Frost, J.T. and L. Hill, *Vitamin D deficiency in a nonrandom sample of southeast Alaska Natives.* J Am Diet Assoc, 2008. **108**(9): p. 1508-11.
23. Kenny, D.E., et al., *Vitamin D content in Alaskan Arctic zooplankton, fishes, and marine mammals.* Zoo Biology, 2004. **23**(1): p. 33-43.
24. O'Brien, D.M., et al., *Red blood cell delta15N: a novel biomarker of dietary eicosapentaenoic acid and docosahexaenoic acid intake.* Am J Clin Nutr, 2009. **89**(3): p. 913-9.
25. Nash, S.H., et al., *Stable nitrogen and carbon isotope ratios indicate traditional and market food intake in an indigenous circumpolar population.* J Nutr, 2012. **142**(1): p. 84-90.
26. Boyer, B.B., et al., *Building a community-based participatory research center to investigate obesity and diabetes in Alaska Natives.* International Journal of Circumpolar Health International Journal of Circumpolar Health, 2005. **64**(3).
27. Boedeker, B. and S. Foster, *2010 Census Counts: American Indians/Alaska Natives Alone or in Combination with One or More other Races, Alaska*, I.H. Service, Editor 2011, Alaska Area Native Health Service: Anchorage, AK.
28. Barrett, J., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics (Oxford, England), 2005. **21**(2): p. 263-5.
29. Bourgain, C. and Q. Zhang, *Kinship and Inbreeding coefficients computation in general pedigrees*, 2009, Free Software Foundation, Inc.: Boston, MA.
30. McPeek, M.S., X. Wu, and C. Ober, *Best Linear Unbiased Allele-Frequency Estimation in Complex Pedigrees.* Biometrics, 2004. **60**: p. 359-367.
31. Team, R.C., *R: A language and environment for statistical computing.*, 2014, R Foundation for Statistical Computing: Vienna, Austria.
32. Wang, Z., et al., *Simultaneous measurement of plasma vitamin D(3) metabolites, including 4beta,25-dihydroxyvitamin D(3), using liquid chromatography-tandem mass spectrometry.* Anal Biochem, 2011. **418**(1): p. 126-33.
33. Phinney, K.W., et al., *Development and certification of a standard reference material for vitamin D metabolites in human serum.* Anal Chem, 2012. **84**(2): p. 956-62.
34. Therneau, T., *coxme: Mixed Effects Cox Models*, 2012: http://CRAN.R-project.org/package=coxme.
35. Chen, T.C., et al., *Factors that influence the cutaneous synthesis and dietary sources of vitamin D.* Arch Biochem Biophys, 2007. **460**(2): p. 213-7.
36. Sachs, M.C., et al., *Estimating mean annual 25-hydroxyvitamin D concentrations from single measurements: the Multi-Ethnic Study of Atherosclerosis.* Am J Clin Nutr, 2013. **97**(6): p. 1243-51.
37. Shoben, A.B., et al., *Seasonal variation in 25-hydroxyvitamin D concentrations in the cardiovascular health study.* Am J Epidemiol, 2011. **174**(12): p. 1363-72.
38. Sachs, M., *cosinor: Tools for estmating and predicting the cosinor model*, in http://cran.r-project.org/web/packages/cosinor/index.html R.p.v. 1.1, Editor.
39. Hilger, J., et al., *A systematic review of vitamin D status in populations worldwide.* Br J Nutr, 2014. **111**(1): p. 23-45.

40. Bersamin, A., et al., *Diet quality among Yup'ik Eskimos living in rural communities is low: the Center for Alaska Native Health Research Pilot Study.* J Am Diet Assoc, 2006. **106**(7): p. 1055-63.

41. Holick, M.F., *High prevalence of vitamin D inadequacy and implications for health.* Mayo Clin Proc, 2006. **81**(3): p. 353-73.

CHAPTER 4: A RESPONSIVE JUSTICE APPROACH TO RECONCILING THE STATISTICAL DEMANDS OF POPULATION STRATIFICATION AND THE ETHICAL DEMANDS IN GENETIC RESEARCH WITH UNDERSERVED COMMUNITIES

4.1 Introduction

When conducting genetic epidemiological research, associations are made between genetic markers and medical outcome. When assessed correctly, these associations can inform medical decisions regarding disease risk and medical treatment. By understanding the effects of these genetic variants, medical providers can adjust their care of individuals who have the variants, thereby providing better, more personalized treatment.

A genetic variant that is associated with a given medical response or outcome often is discovered through large population-wide analyses and case-control studies. When the "risk markers" are discovered within a study population, they inherently include information from the genetic structure and environment of that population. What this means is that associations are most valid in the population in which they are developed. As a result, associations that are valid in some groups of people are not valid in others. In trying to extend the benefits of genetic testing to all people and to improve health justice, all populations must be included in research. However, some groups may be hesitant to participate in the very research that could improve their care, as genetic research may put them and their community at risk.

Furthermore, the underlying genetic structure of human populations can complicate the validity of these medical genetic associations in different groups of people, even in the groups participating in the research. Statistical adjustments can be made to account for this complication, but the measures needed to make these adjustments often expose participating groups to greater risk of the harms associated with genetic research. Here, I explore solutions to maximizing scientific validity of genetic research and improving healthcare for underserved groups, while also minimizing risk to these same people. As part of this discourse, I present the need for adjusting for population substructure in genetic studies, the ethical barriers to this scientific standard, and possible solutions to finding a maximally desirable balance between the two through the lens of responsive justice.

4.2 Population Substructure

Before we begin, let us establish what we mean by "population." In defining populations globally, groups of people may be categorized by any arbitrary measure. One of these categories is race. While race itself is not the topic of this discussion, it is a concept often tied up in discussions of population genetics and can lead to misunderstanding. As Luigi Luca Cavalli-Sforza said of the classifications of race in his work *History and Geography of Human Genes*, "the level at which we stop our classification is completely arbitrary. Explanations are statistical, geographic, and historical. Statistically, genetic variation within clusters is large compared with that between clusters" [1]. Genetic boundaries between groups are weak, and there is a long-standing consensus among population geneticists that there is no biological basis for race [2, 3]. As such, the populations I discuss are divisions of people on a scale that could be useful in informing medical care. These populations can be defined geographically, or by medical care sought, or among nearly every line of division imaginable. In my case, I will discuss populations of historically marginalized and underserved communities. These are groups of people who share a similar history, often geographically isolated from other populations and, as a result, share genetic patterns that differ from the populations traditionally included in medical research.

Population structure is the pattern of genetic differences within a given group of people that reflects family relationships. The question is not whether or not there is population structure in historically geographically isolated communities, because there most certainly is. Population substructure is a concern in any sample that is not absolutely random and in which not every individual has the same relationship to each other and to the individuals in the population that the sample is representing. Between and within populations globally, small but statistically significant differences in genetic variation have been found [4]. Even within apparently homogenous populations, population substructure exists [5-7]. As Cardon, *et al.* argue, the challenge is not in showing that population substructure exists, but in accounting for it [6].

Population stratification arises from human natural history. Different genetic patterns in human populations across the world exist simply because of differences in migration patterns, mating practices, changes in population size affecting genetic drift, cultural practices, and new mutations, in addition to selection, random mutation, and chance [6]. Large-scale population substructure can be detected as a deviation from Hardy Weinberg Equilibrium (HWE) at genetic loci, but smaller scale structure may not affect HWE. As groups separated and ceased interbreeding freely, different genetic patterns developed differently across the globe, often driven by randomness, revealing a gradient of allele frequencies and variation. The frequency of genetic variation at a single spot in the genome can change unpredictably from one population to another, regardless of how population boundaries are defined, and as such, there is great genetic variation between all populations and people.

4.3 The Effect of Population Structure on Genetic Studies

Population substructure can lead to both Type 1 (false positive) and Type 2 (false negative) errors [8] depending on the selection of participants, and can affect the precision and accuracy of allele frequency and association studies. Especially in large genome wide association studies, the worry is that population substructure will lead to type 1 errors and identify associations that are not causative for the trait of interest, but will then be used to incorrectly inform medical intervention and risk assessment [9].

Just as failing to account for non-genetic factors can bias results of association studies, the genetic structure among a sample of research participants can lead to spurious associations and confounding of results. This confounding from population substructure happens when individuals in a group have phenotypes and genotypes that are correlated with each other more than would be expected by chance. When phenotypic differences vary between subgroups that also have different allele frequencies, an association between the two is found irrespective of causality. To confound results, 3 conditions must be met: 1) the frequency of the trait of interest must be different between subpopulations after adjusting for any cofactors, 2) the frequency of a given genetic marker must be different between subpopulations, and 3) the allele frequencies and the trait frequencies must follow the same directionality for reasons other than causality [9].

The most severe effect of population structure on study results is false association of genetic markers and outcome. For example, in a study of diabetes in the Pima Indian population, an HLA variant was associated with increased risk of diabetes, but later found to be an artifact of differences between the cases and controls in the degree of European ancestry resulting from population admixture [6].

Besides the confounding of associations themselves, the power in association studies can be reduced by population structure, with weaker associations more affected, even in the cases of true association [10] because population substructure is essentially a case of selection bias [11, 12]. A study of lung function in Mexico found that positive associations would be masked in failing to account for population substructure, an example of Type 2 error [13].

In the case of population allele frequency studies, population substructure causes over-dispersion. In over-dispersion, the predicted variance from the statistical model is smaller than the actual variance of the population [11].

The impact of population substructure on associations may be greater in groups with sparse or inaccurate understanding of ancestry or in groups with recent admixture, and in studies involving alleles or phenotypes that vary greatly between populations [6]. As a result, understanding and accounting for the population substructure of a study population improves the validity of conclusions from medical research, even when conducting studies within a population defined narrowly.

It seems that the greatest risk for confounding of results occurs in genome wide association studies, or when otherwise trying to identify new genetic targets. In genome wide studies, the large number of markers assayed increases the probability that a given marker will track with the trait of interest by chance and lead to a false positive result. However, in replicating associations found in other populations or in interrogating a few biologically plausible candidate genes, as of 2009 it was unknown whether ancestry affected the association of a given causal genotype that had been found in another population [14]. Because genome wide association studies look for an enrichment of a genetic marker in cases of an illness, for example, they may not detect a causative marker directly and instead might rely on linkage disequilibrium patterns, which can vary significantly from one population to another [15]. When studying causal genetic markers, however, differences in linkage disequilibrium patterns should not affect results. Therefore, while population substructure exists in all study populations and has the potential to confound results, its effect on results and conclusions likely depends on study design.

4.4 How Genetic Analyses Are Best Adjusted for Population Substructure

Adjusting statistical models can minimize this confounding from population substructure. The standard way to account for population substructure is to use genome-wide ancestry informative markers (AIMs) to cluster participants by genetic ancestry and then to adjust for differences using genomic control or Bayesian clustering with likelihood statistics [10]. Subgroups are clustered into more genetically homogenous clusters using principal component analysis (PCA) or multidimensional scaling (MDS) [16]. Both PCA and MDS separate genetic components to maximize the variance between samples. The clusters created through PCA and MDS can be used as covariates in association studies to reduce confounding by population substructure.

As few at 30 single nucleotide polymorphisms (SNPs) may be needed for the independent genome-wide informative markers used for genomic control when accounting for population substructure [11]. As would be expected, populations with less differentiation between them need data from more SNPs to uncover genetic substructure [16]. To distinguish populations on a global scale, just 1000 SNPs are needed, but to differentiate populations on a finer scale within Europe, 10,000 SNPs are needed [16]. Even within Switzerland, PCA could distinguish genetic differences between French, German, and Italian speakers [17] and population structure also could be detected in Sweden [18].

Without genome wide markers, pedigrees showing the kinship relationships between participants can be used to adjust both allele frequency estimates and association studies. The Best Linear Unbiased Estimator (BLUE) adjusts allele frequency estimates by weighting genotypes by kinship coefficients to account for the non-independence of samples [19]. Including a kinship coefficient matrix in mixed effects regression models of association studies also can account for this population structure by weighting individuals [20]. Without either AIMs or pedigree information, however, these analyses cannot be adjusted for population substructure, leaving results at risk of confounding and over-dispersion.

4.5 Hesitation of Underserved Populations to Participate in Genetic Research and to Collect Genome-Wide Markers of its Members

While becoming more interested and involved in genetic research, some historically marginalized and identifiable communities are hesitant to participate, especially to the extent of allowing collection of genome-wide markers. Before discussing alternatives to collecting either genome-wide markers or pedigrees to account for population substructure, I will present some concerns of communities with respect to genetic research in general, and how collection of genome-wide markers increases these risks.

When the Human Genome Diversity Project (HGDP) began to collect DNA from indigenous populations worldwide, indigenous leaders called it the "vampire project" [21]. A history of genetic research contributing to racial stereotyping, such as publication of a "risk-taking", or "warrior", gene that is associated with increased alcohol and tobacco use among Maori people, has formed a basis for hesitation to release genetic information [21]. After a long history of oppression and discrimination, communities are understandably hesitant to participate in research that could be misused to support stereotypes [22, 23].

While indigenous groups across the globe may recognize the benefits of genetic research, they may feel that the risks are too severe to participate [21, 24]. Even given the best intentions of researchers, research relationships can be delicate with identifiable communities, and can have negative consequences for the culture, community, and individuals of the people involved. This is especially true for populations vulnerable to exploitation, such as those with a history of marginalization or of exploitation by the government. For example, the Tuskegee syphilis experiment, in which the U.S. government prevented African American men from receiving normal care for syphilis so that the natural course of the disease could be studied, contributed to the establishment of ethical review boards and protocols, but has had a lasting negative effect on the willingness of the African American community to participate in medical research and to receive preventative health care [25]. In another prominent example, DNA samples donated by members of the Havasupai tribe in a population study of diabetes were also used to study historical migration patterns, inbreeding, and schizophrenia without tribal approval or knowledge, and led to a lawsuit against Arizona State University [22].

While all research has potential to harm participants, genetic research can be especially sensitive to research harms, as outcomes inherently apply to an entire community. These outcomes include those intended from the study as well as any unintended consequences and harms. Many of these harms can be avoided in research with less identifiable populations through privacy and confidentiality protections for individual research participants. However, because genetics is inherently unique and non-anonymous, privacy is not always possible at a group level when conducting genetic research [26]. Even if an individual's identity is protected, the community may still be identifiable [27].

Harms can be divided into tangible harms, those that have a measurable or physical outcome, and dignitary harms, those that affect emotion or self-perception [28]. Tangible harms in genetic research include discrimination, stigmatization, and loss of social opportunities and standing.

Discrimination and stigmatization can occur if genetic variation associated with increased occurrence of a disease is more prevalent in a community [29]. For example, the finding that sickle cell trait was more common among African Americans led to their discrimination in the 1970's [30]. In a study in Barrow, Alaska, increased prevalence of alcoholism was tied to a community of Alaska Native people. While the researchers had designed the study with the intention of helping the community to address alcoholism and related issues, presentation of the results became stigmatizing and affected the town's bond rating [27, 31]. While not exclusively so, stigmatization is a greater risk when the community is small and the trait studied is rare and severe, as smaller communities are easier to identify and to collectively stereotype [29, 32].

Stigmatization also can take the form of revealing socially disrespected classifications. For example, genetic research may reveal instances of consanguinity or inbreeding, which are generally disrespected, and so can have a negative impact on the status of the individuals studied [29]. The community with which these individuals identify can also be affected if consanguinity is seen as ubiquitous in the community.

With regard to loss of opportunities, studying genetic markers may show that some individuals in a tribal community are more “non-Native” than “Native”, and as a result, tribal standing may be challenged [28, 33, 34]. Tribal standing can be important in qualifying for resources of the tribe, such as healthcare or stake in tribal-owned operations, and in running for political office. Changing the tribal standing of someone as a result of genetic quotas can affect the social opportunities and status of an individual within the community.

Dignitary harms are those that violate a collective right or show disrespect, causing feelings of shame or humiliation. For example, challenges to understanding of place, spirituality, values, history, and autonomy may harm community or individual identity.

For example, in the Human Genome Diversity Project (HGDP) aiming to characterize human genetic diversity across the globe, indigenous groups “opposed HGDP because they questioned the goals of the project, e.g., mapping human migration, and worried about the possible outcomes of the project, such as discrimination against their group, stigmatization, effects on sovereignty, internal harms to the tribe, and the possibility that researchers or others would profit from the tribes’ biological materials” [24].

Genetic studies could challenge a community's claims to a land. Studies of population history and migration through genetics may challenge territorial integrity, affecting such things as community claims for repatriation of remains and artifacts found in regions of historical tribal lands [28, 29, 35]. This became relevant with Kennewick man, whom researchers argued was not an ancestor of the local tribes who requested repatriation and who identified religious significance of the remains [36]. Genetic information could have been used as evidence against (though also for) this tribal understanding of ancestral place.

Additionally, genetic research can harm community spirituality and identity as a result of cultural and epistemological differences between researchers and communities. For example, through the study of historical migration patterns, or through such things as characterizing spiritual leaders as schizophrenic, such as what happened in research with the Havasupai Tribe, this lack of recognition can violate a group's rights to its own culture [24, 27, 32]. As another example, the western researchers of the HGDP aimed to use genetic patterns to develop tools for archeology, cultural anthropology, historical linguistics, and "the evolution of our species" [23]. However, indigenous epistemologies differ from the western normative, as "communities know where they came from, who they are, and what their relations to the land are" [23, 24]. These harms of group identity can have an interesting impact on a community because they don't harm a particular person directly, but can diminish the outsider views of the group and the views of members of the group toward themselves.

These harms to identity and spiritual understanding also can arise in the form of publication or secondary use of specimens after the intended research is completed, even if the original study was executed respectfully [24]. Individuals and communities may lose control of the samples, any genetic data produced, or oversight over making sure data are used for acceptable purposes [24, 36]. The US Department of Health and Human Services Office of Human Research Protections considers de-identified secondary uses of data to be exempt from human subjects oversight, and secondary research with HGDP samples indeed has investigated human migration history and effects of human evolution [22]. This loss of control can be a violation of sovereignty and informed consent [29]. Additionally, donation of genetic samples raises concerns of data ownership, who controls any research, and who benefits from the community's "resources" [24].

As a summary of harms from genetic research, communities may be concerned about discrimination, stigmatization, loss of opportunities, and challenges to community identity, which are embodied in attacks on historic territory, spirituality, and epistemology, and in secondary use of specimens.

With the collection of more genetic markers, such as would be needed for calculating and adjusting for principal components, the risk for harm is greater. The more information collected across the genome, especially targeting regions that would be used to establish ancestry, the greater the ability to conduct the science that would lead to these perceived risks. Provided these data do not exist, control of them cannot be lost and they cannot be used beyond their original intention. Once these data are collected, loss of control opens the opportunity for these harms.

4.6 An Approach to Dealing with these Issues: Responsive Justice

Given that participation in genetic research is important for expanding benefits to include all people [37], but that some communities are hesitant to participate in genetic research, a balance is needed between pursuing the advancement of knowledge and protecting these groups from harm. In considering core ethical concepts, values and priorities are analyzed to weigh autonomy, beneficence, and justice in a way that evaluates whether the risks of research are worth the benefits, both at the level of the individual and the community. One way to approach this balance is through transformed practice of responsive justice, which is based on a collaboration of researchers and underserved communities to develop mutually beneficial research goals and plans [27].

Responsive justice seeks to reduce power disparity between researchers, who traditionally hold more power, and participating communities, who traditionally hold less power. This power redistribution transforms communities into partners in research, as opposed to subjects of research, and includes them consistently and wholly in research decisions, including what research will be done and how it will be conducted. This process relies on community based participatory research (CBPR) [38]. The three elements of responsive justice are recognition, responsibility, and redistribution [27]. By including communities in research conversations, researchers can understand the views and values of collaborating communities, thereby establishing a “fundamental awareness of and respect for the person or communities with whom research is being conducted, and to whom the clinical benefits of the research will be returned” [27].

The first element of responsive justice is recognition, which reflects the understanding by researchers of the needs and values of the participating community, as defined by the community's own members [27]. The opposite of recognition, misrecognition, is inherent in traditional structures of research, including priorities and research methods, and is not simply a lack of attention or care by researchers. A purposeful approach by researchers is important to have careful and conscientious discussions with a community about fears and hesitations surrounding participation in research, and how best to address these concerns to achieve mutual goals [21]. Recognition requires consistent and repeated conversations with each community throughout a research partnership to understand concerns specific to that community, at that time and place. By fully knowing *why* a specific community feels the way it does, researchers can respond with the most appropriate approaches and responses, and can adjust them as the relationship with the community matures and concerns change.

The second element of responsive justice is responsibility, which reflects a moral obligation of the researchers to their collaborating participants to act conscientiously in response to the elements of recognition. As the holders of greater power, the researchers are responsible for driving the conversation, ensuring that their engagement extends beyond the boundaries of their personal research goals to address further interests of the community. Researchers must examine and modify their behavior in response to conversations with communities, always acting toward justice [27].

The third element is redistribution, which reflects the fair distribution of benefits and burdens resulting from the research. Fair distribution is different from equal distribution, as it focuses on distribution and redistribution according to need and appropriateness, aiming to increase justice overall. The responsibility of researchers is to provide equal opportunities for access to benefits. Whether or not an individual or community chooses to accept those benefits is outside the scope of the researchers' obligation.

By applying responsive justice, researchers can work with communities to develop research methods and questions that are mutually beneficial, but which prioritize the needs and views of the community. Responsive justice requires researchers to acknowledge and respond to specific concerns of traditionally underserved communities. Communities may be more willing to participate in genetic and medical research if their fears and needs are addressed directly and they have ownership over the research process.

In the case of genetic research and collection of genome-wide markers, responsive justice can be applied in working with communities. Based on conversations of recognition with the community, research parameters can be established around how genetic samples are collected, who participates in a study, what genetic data are obtained, how data are housed and shared, who controls the data, how results will be shared with the community, and how the community will benefit from any resulting advances. These conversations can also include specific concerns of the community with respect to each of these topics, and how researchers will address each in developing a collaborative research plan. One specific conversation related to genetic research is whether or not to collect data on genome-wide markers of participants, which allows for adjustment by principal components in genetic association studies. Through the framework of responsive justice, researchers can recognize the concerns of the communities, and through responsible practice, engage communities in exploring the risks and benefits of using complete genetic adjustment, compared to possible alternatives.

One of these alternatives to collecting data on genome-wide markers is to cluster subsets of the community, such as individual tribes, by similarities of ancestral language. By assigning individuals to one of these clusters based on self-identified affiliation, membership in a given cluster may be able to be used as a covariate in association analyses, thereby approximating the first principal component that would emerge from clustering by genome-wide markers. In this way, researchers may be able to account for population substructure and improve the accuracy of results. Using these language relationships may provide an alternative to collecting genome-wide markers, but its usefulness depends on both the scientific validity and ethical utility of applying these clusters in genetic studies. A discussion of the scientific validity and ethical utility of stratifying research participants by linguistic clustering follows.

4.7 Scientific Validity of Using Language Relationships to Account for Population Substructure

It has been suggested that in the working guidelines of the Human Genome Diversity Project, researchers were encouraged to seek out groups with unique languages, as "communities with distinct languages are presumed to have the greatest potential for also bearing distinctive genetic material" [23]. Languages have been used to group people in anthropological studies since before the development of genetic technology, and it was thought that language relationships may approximate genetic relationships [39]. This connection of language and genetic history is logical, as language relationships say much about social and cultural connections and about the identity of a person [33]. As such, categorization of people by the relationship between languages of their ancestral groups may serve as an accurate surrogate for principal component analysis in cases of genetic studies in which colleting AIMs is not preferable.

The correlation between human language and genetic patterns is not new. In *On the Origin of Species,* Darwin proposed that the classification patterns would be the same between the two, essentially creating the same genealogy [1, 40]. At first glance, this concept seems to be upheld. Studies globally have found positive correlation between geography, genetics, and linguistics, even at a microgenetic level [1]. The complication is that the effects of geographic, cultural, and linguistic history must be separated from each other, and the relationship of each with genetic history is complicated. Genetic and linguistic patterns can be difficult to decipher, as groups mix and evolve, and transmission patterns of the two are not the same. Language learned travels horizontally between people through cultural transmission, while genetics adheres to vertical transmission between people through biological ancestry. To assess the use of language relationships in genetic studies, the scientific validity behind the theory that linguistic clustering can be used to approximate genetic clustering to account for population structure is explored here.

The strengths of this theory include the parallel evolutionary processes of language and genetics, and the mutual reinforcement of barriers. The similarities in the evolutionary process of genes and language are inherent in a population splitting and differentiating, leading to increased divergence of both language and genetics between groups over time. Thus, it seems reasonable that languages would diverge in accordance with increasing genetic divergence. Furthermore, groups that are nearby on the genetic tree often speak related languages [1].

In addition to following similar evolutionary paths, genetic and linguistic development may reinforce barriers between groups, continuously strengthening the parallels between the two patterns. The same geographic and ecological barriers and distance that limit genetic pools also limit cultural interactions, including the exchange of language. Once the cultural barriers develop to a degree that prevents intermixing, such as through inability to communicate, genetic intermixing also typically declines. The language an individual speaks is determined by culture, which often is rooted in a family environment defined by genetic relationships. Similarly, language creates barriers that can preserve distinct genetic features. As such, it is reasonable to assume that shared social identity may reflect a shared genetic history [41]. According to some views, linguistic and genetic relationships seem to reflect the same patterns at national and supranational language scales, but to be less reliable at smaller scales, where dialects and more subtle differences in language are found [1]. Other studies suggest the opposite, that genetic correlations only are upheld at a micro level [42]. This issue of scale is important for determining the usefulness of language patterns in approximating genetic patterns for use in medical and genetic research.

While language and genetic relationships broadly follow the same patterns, language groupings cannot be assumed to reflect genetic groupings on the scale needed for clinical research. Languages evolve very quickly compared to genetics, which can make connections between languages difficult to reconstruct accurately. The rate at which genes and languages diverge is not consistent, and differences between the two can be intensified by replacement, in the case of language, and substitutions, in the case of genetics [1].

Replacement and substitutions are products of admixture, in which groups mix after being separated for a long time.

Admixture weakens the similarities between linguistic and genetic patterns. The genetic effects of admixture can be assumed to be proportional to the relative contribution of ancestral groups [1]. However, language behaves as a more complete unit and often becomes more strongly singular over time. After an admixture event, however, DNA blocks from both ancestral groups remain across the genome. During mixing of one group into another or migration of individuals into a neighboring group, a complete linguistic or cultural transition can happen in less than 3 generations, whereas genetic signatures remain indefinitely in proportion to the degree of admixture [1, 43].

Also not supporting the comparability of genetic and linguistic genealogy, unlike genes, languages are not always transferred vertically in a family tree, but can transfer between unrelated individuals. Due to proximity of groups that may not share language, it is likely that at least a few mates will be exchanged over time, which will bring the groups closer genetically, while not greatly affecting language differences [44].

In conducting genetic analysis at specific loci in the genome, it is important to consider that this proportional genetic distance assumes average over the entire genome, whereas individual genes may show erratic and inconsistent patterns with respect to ancestral groups. Therefore, small portions of the genome, including mitochondrial DNA or Y-chromosomes, cannot be assumed to reveal the same population history as would be expected in the genome overall in the same population. Thus, there can be major deviations in the correlation of language and genes, both on the scale of an individual gene and the whole genome.

Furthermore, dialects merge into each other, creating a gradient over space that can correlate with genetic distance, but may make the division of groups by language arbitrary. For example, there is a gradient of dialects among indigenous languages in the Americas, spanning from northern Alaska to Greenland, including Alaska Inuit, Canadian Inuit, and Greenland Inuit [1]. Additionally, the Athabascan language family is comprised of 30 sublanguages stretching from eastern Alaska to California and Arizona, and in the rest of the Americas the Amerind language family includes almost 600 sublanguages [1].

Looking at studies in populations around the world that specifically address the question, some linguistic and genetic correlations between groups are found. Studies in Ethiopia and Peru found linguistic clustering to be correlated with genetic clustering [42, 45]. Genetic diversity in Ethiopia is large, about half of all genetic diversity in Africa, and the languages corresponded with cultural barriers sharing the same geographic area. The linguistic distances in Peru were correlated with how far away each population lived geographically from the others, and the authors caution that determining accurate language relationships was important for the validity of the genetic association, but also were difficult to access accurately.

Given the caveats of these studies, it is not surprising that the correlation of language and genetics is not consistent across populations. A study of linguistic, geographic, and genetic distance in South Asia, including 46 diverse tribal populations, found low correlations between geographical and genetic distance between tribal groups, but found correlations of language and genetics only at the scale of language family [43]. Similarly, studies of European language and genetic distances between groups found no correlation, but did identify some genetic deviations along cultural and ethnic barriers and by geographic distance [46, 47]. As such, cultural barriers may define gene pools more so than geographic proximity.

Looking specifically at the Americas, even Cavalli-Sforza, a champion for the correlation of linguistic and genetic relationships between groups, found that Amerindians seem to be quite genetically variable, with linguistic patterns not correlating well with genetic relationship patterns [48]. Admixture may be partially responsible for this deviation. A study by Chakraborty found no relationship between genetic and linguistic distances in a study of indigenous groups in the Andean highlands of Chile, which were clustered based on linguistic studies by Joseph Greenberg [44]. This study found a significant correlation (0.716) between genetic distance and geographic distance in South America, with geographic differences explaining 50% of variability in genetic differences [44]. While the correlation between geographic and genetic distance was significantly positive, the correlations between linguistic distance and genetic distance, and between linguistic distance and geographic distance were not significant [1, 44]. Other studies, including one in Central and South America studying 1381 individuals from 17 populations looking at genetic correlations with 8 different language classification patterns, and another looking at 3 American Indian tribes in the Pacific Northwest where linguistic diversity is especially high, found no correlation of genetic patterns with linguistic clusters [40, 49]. Yet another study of linguistic and genetic correlations among indigenous groups, this time in Mexico, found no correlation between the two [50]. However, the genetic signatures were consistent with historical population expansion and diverse genetic drift, including community isolation and founder effects. European contact in the Americas may have affected linguistic and genetic patterns due to admixture combined with the mass eradication of people, which resulted in severe reduction of the gene pool [1, 40].

Interestingly, many of these studies found no correlation between linguistic and genetic distance, but did find correlation between geographic and genetic distance. Any associations that were found between genetic and cultural differences, including language, may result from confounding by the geographic distance that accompanies cultural differences [40]. Broken into ranges of 200 miles, genetic distance increases with geographic distance between each group across the world [1]. On a global scale, reconstruction of population divergence based on genetic distance reflects the same relationship structure as reconstruction based on geographic distance [16]. On a smaller scale within Europe, individuals coded by country of origin were clustered by genetic distance, which created a 2-dimensional map of the populations that essentially depicted the geographical layout of the same populations over a map of Europe [17]. Of note, only individuals with all four grandparents from the same country were included in the study, thereby reducing the complications of admixture.

Similarly, a study in Mexico found a strong correlation between fine scale geographic distances and genetic distances among 20 indigenous and 11 mestizo communities [13]. The degree of variation found in this study was striking, showing some genetic divergence between these communities on the same scale as divergence between European and East Asian populations. Importantly, the population structure created from these geographic clusters was strong enough to affect the results of association studies. This study found significant association between ancestry and risk for lung disease, and supports the need for adjusting for genetic relationships in medical studies, even among communities that are geographically close together and that may share history [13]. Interestingly, another study in Mexico did not find correlation between geography and genetics in groups defined linguistically, a finding that the authors thought likely resulted from the constant migration of ancestral groups [51].

Correlations of linguistic and genetic distances are imperfect, and often attributable to geographic distribution of language and culture. While clustering by linguistic patterns can approximate genetic structure to some degree, available evidence suggests that these patterns are likely not reliable for clustering groups in lieu of genome-wide markers. While the relationship may be valid in some cases, full genetic analysis would be needed to confirm the accuracy of each case, which would require the exact data collection that this approach is trying to avoid. Perhaps a more appropriate approach would be to use geographic clusters, defined along the lines of dominant cultural separations. The utility of this tool must be considered for each case individually. Because kinship, social structure, and culture affect geographic expansion and boundaries just as much as geography affects the development of culture and the exchange of genetic material, a combination of geographic and cultural measurements may maximize scientific validity in approximating genetic relationships [52].

However, while geographic clustering, especially defined by cultural divisions, may be more appropriate than linguistic clustering alone, it still does not allow for the nuanced adjustment capable with genome-wide markers, especially in cases of admixture. The geography of a population is still based on a single location, connected to a relatively recent point in time and accounting for a single, often self-described ancestry of a person. As such, it will not match the finer chromosomal adjustment possible through AIMs, but using geographic clustering may allow for adjustment by a dominant principal component in association studies, and in that way be preferable to no adjustment at all.

4.8 Ethical Utility of Using Population Geographic and Linguistic Relationships to Account for Population Substructure

In this role as an approximation of genetic clustering, geographic and linguistic clustering of self-identified tribal affiliation can be a useful tool to remedy some, but not all, of the potential harms presented by genetic research. In fact, geographic clustering is preferable to linguistic clustering with respect to ethical challenges, as it does not classify people by measures that are not already obvious. Language development and evolution can be analyzed using similar theories as those applied to genetics, with older languages showing richer diversity. Language relationship patterns can reveal, and indeed have been used to reveal, migration patterns and historical relationships between groups. Geography, especially considered in present tense, does not reveal these patterns or relationships with respect to collective community identity of history, time, and place, or understandings of origins, and so may be preferable to linguistic classifications.

In considering the usefulness of linguistic and geographic classification to approximate principal components in genetic analysis, the scope here is limited to small, underserved communities in which genetic research for medical purposes could be valuable, but for which the genome-wide genetic information used to adjust for population substructure is unavailable. First, I will present harms that are avoidable through using geographic and linguistic groupings instead of genetic markers. Then I will discuss harms that are not avoidable by using geographic and linguistic groupings. Avoidable harms include any revelations of consanguinity, use of genetic quotas to affect tribal status of individuals, some challenges to community identity and history, and some inappropriate secondary uses of data.

First of all, some data that could lead to stigmatization will not be collected, such as data revealing inbreeding or consanguinity. The geographic groupings will not reveal family structure, but only self-identified membership within a larger community. No other sensitive or surprising relationships between individuals would be uncovered.

Additionally, because geography of a self-identified tribal affiliation is tied to a single social identity, it does not segregate or dilute with each generation like genetic signatures do, and it cannot be co-opted to reveal quantity of native genetic ancestry. Individuals may be incorporated into a community regardless of genetic ancestry, so using language and geography of self-identified tribal affiliation protects against political and social disenfranchisement, such as access to tribal benefits or ability to run for tribal office. Language and geography of self-identified tribal affiliation is known and obvious, and classification using this measure will not reveal any new identity or understanding of ancestral history.

Finally, with respect to secondary uses of data for unapproved and sometimes stigmatizing research, using language and geographic region to group individuals can remove some risk but not all risk. For example, any secondary research using genetic markers that would have been used to account for population substructure is now impossible, as data at those markers have not been collected. This means that secondary associations with disease or claims about individual ancestry cannot be made. As an example, if only limited genetic markers were collected in the Havasupai people, the markers then used to challenge community history and spirituality may not have been available.

Although using language and geography of self-identified tribal affiliation to group individuals offers important benefits, some of the potential harms from collecting genetic information to adjust for population substructure will not be avoided. These include some risk of discrimination from research results, some challenges to community understandings of history and identity, and some secondary uses of data.

Risk of discrimination resulting from research results tied to a particular community will not be completely avoided. Specific communities or groups of communities can still be identified through their language and location, so their privacy will not be preserved. This means that any trait or disease that could have led to discrimination if tied to a particular community through genetics can still be tied to that community through language and geographic classification and still can lead to discrimination. This identification also leaves communities at risk for harmful secondary uses of data and presentation of data. For example, because of pleiotropy, genetic data can be linked with traits or outcomes other than the ones specifically intended to be studied. As a result, some stigmatizing or discriminatory secondary analysis and conclusions may still be possible and able to be tied to communities. These secondary uses can also be used to cause dignitary harm if the results cause community members or outsiders to see the community differently.

The key concept in analyzing the impact on harms associated with genetic research from grouping individuals by linguistic and geographic clusters instead of genetic markers is that communities can still be identified, even as it masks relationships at the individual level [26]. As a result, any research findings can be attributed to the community, which is a necessary point of genetic health research. However, any other results can also be attributed to the community, which can have a negative impact on the identity of a community, both from the viewpoint of its own members and from those on the outside.

Although it is not always scientifically valid, in some instances, linguistic and geographic relationships may be a preferable tool to using genetic markers to cluster research participants. As with many ethical issues, this is not the one right and perfect answer for genetic research with communities hesitant to participate. The best way to decide whether this classification system could be a preferable tool for a community is to include the community in the decision-making process through responsive justice [24, 27, 30]. Community harms and priorities can be difficult to anticipate. Just as research agendas should be developed in collaboration with communities to foster community buy-in, benefit, and recognition of potential harms, conversation with each community can be used to tailor classification systems that align with the interests of the community. Trust is a major motivator against participation in genetic research, and trust can be built through engaging a community and through addressing the ethical concerns of its members [30, 36].

While not a perfect tool, using linguistic and geographic relationships to account for population substructure in genetic analyses can reduce the risks for some of the potential harms in genetic research with identifiable communities, especially with low levels of admixture. However, as admixture increases, leading to more complicated population structure and genomic patterns, the utility of stratifying participants by self-identified geographic and linguistic affiliation decreases.

4.9 Conclusion

Expanding the benefits of genetic research to address the health concerns of underserved populations requires new approaches to engaging communities. In balancing the impact of population stratification on the validity and utility of results with respect for the ethical concerns of communities, stratifying participants by cultural and geographic relationships of self-identified ancestry may be appropriate for populations with low rates of admixture. However, as admixture increases, the utility of this stratification will decrease. While studies can be adjusted for AIMs of participants, a responsive justice approach requires an exploration of alternative approaches to addressing community concerns. Communities may be hesitant to participate in genetic research, especially the collection of genome-wide markers, for a variety of reasons. The obligation lies with the researchers to recognize and address the specific concerns of each group and to include the group in decisions throughout the entire research process.

One possible concern of a group when participating in genetic research is the collection of genome-wide markers that are used to adjust for population substructure in order to reduce confounding in association studies. Collecting these data could leave the communities at greater risk for harms inherent in genetic analysis. Here, we have explored using language relationships of ancestral groups to cluster people in order to approximate population substructure and to avoid collecting genome-wide information, while still preserving the validity of research results. Although language clustering may be appropriate in some situations, it is generally not a reliable tool for approximating genetic relationships.

For some types of genetic research, however, adjusting for population stratification may not significantly impact results or change their utility for medical decisions. While not the most scientifically rigorous approach to population genetic research, not adjusting for AIMs may be the most ethically responsible choice and may still produce meaningful results. Most documented cases of confounding due to population stratification are in association studies that are looking for new SNP targets [5, 53]. For studies looking to replicate associations of genetic markers that have been found in other populations or that investigate candidate genes based on biological plausibility, the risk of false positive results may be minor. This is because confounding requires both the trait of interest and a genetic marker to occur at different frequencies in the case versus the control groups, and also requires both the trait and genetic marker to have the same directionality. The probability of confounding is smaller in candidate and replication studies than in genome wide studies due to the smaller number of markers assayed and the deliberate choice of gene targets. Future studies that specifically investigate the impact of adjusting and not adjusting for population stratification in these subsets of study design could inform the importance of adjusting for population substructure when working with underserved communities.

Here I have explored a possible approach to addressing a specific concern of some communities when participating in genetic research. However, the ultimate need is for researchers to employ responsive justice, involving the community in the collaborative development of a research proposal and throughout the research process. By including, acknowledging, and responding to the concerns and unique history of each population in this way, medical advances, including those resulting from genetic studies, can be accessible to more people.

4.10 References

1. Cavalli-Sforza, L.L., *History and Geography of Human Genes*.
2. Gould, S.J., *The mismeasure of man*. 1996, New York: Norton.
3. Graves, J.L., *The race myth : why we pretend race exists in America*. 2004, New York: Dutton.
4. Weiss, K.M. and S.M. Fullerton, *Racing around, getting nowhere.* Evolutionary Anthropology: Issues, News, and Reviews, 2005. **14**(5): p. 165-169.
5. Berger, M., et al., *Hidden population substructures in an apparently homogeneous population bias association studies.* Eur J Hum Genet, 2006. **14**(2): p. 236-44.
6. Cardon, L.R. and L.J. Palmer, *Population stratification and spurious allelic association.* The Lancet, 2003. **361**(9357): p. 598-604.
7. Redden, D.T. and D.B. Allison, *The effect of assortative mating upon genetic association studies: spurious associations and population substructure in the absence of admixture.* Behav Genet, 2006. **36**(5): p. 678-86.
8. Hoyos-Giraldo, L.S., et al., *The effect of genetic admixture in an association study: genetic polymorphisms and chromosome aberrations in a Colombian population exposed to organic solvents.* Ann Hum Genet, 2013. **77**(4): p. 308-20.
9. Wacholder, S., N. Rothman, and N. Caporaso, *Counterpoint: Bias from Population Stratification Is Not a Major Threat to the Validity of Conclusions from Epidemiological Studies of Common Polymorphisms and Cancer.* Cancer Epidemiology, Biomarkers, & Prevention, 2002. **11**(513-520).
10. He, Y., et al., *Correlation of population parameters leading to power differences in association studies with population stratification.* Ann Hum Genet, 2008. **72**(Pt 6): p. 801-11.
11. Bacanu, S.-A., B. Devlin, and K. Roeder, *Association Studies for Quantitative Traits in Structured Populations.* Genetic Epidemiology, 2002. **22**(1): p. 78-93.
12. Tian, C., P.K. Gregersen, and M.F. Seldin, *Accounting for ancestry: population substructure and genome-wide association studies.* Hum Mol Genet, 2008. **17**(R2): p. R143-50.
13. Moreno-Estrada, A., et al., *Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits.* Science, 2014. **344**(6189): p. 1280-5.
14. Via, M., E. Ziv, and E.G. Burchard, *Recent advances of genetic ancestry testing in biomedical research and direct to consumer testing.* Clin Genet, 2009. **76**(3): p. 225-35.
15. Ziv, E. and E.G. Burchard, *Human population structure and genetic association studies.* Pharmacogenomics, 2003. **4**(4): p. 431.
16. Wang, C., S. Zollner, and N.A. Rosenberg, *A quantitative comparison of the similarity between genes and geography in worldwide human populations.* PLoS Genet, 2012. **8**(8): p. e1002886.
17. Novembre, J., et al., *Genes mirror geography within Europe.* Nature, 2008. **456**(7218): p. 98-101.
18. Salmela, E., et al., *Swedish population substructure revealed by genome-wide single nucleotide polymorphism data.* PloS One, 2011. **6**(2).

19. McPeek, M.S., X. Wu, and C. Ober, *Best Linear Unbiased Allele-Frequency Estimation in Complex Pedigrees.* Biometrics, 2004. **60**: p. 359-367.
20. Pan, Z. and D.Y. Lin, *Goodness-of-fit methods for generalized linear mixed models.* Biometrics, 2005. **61**(4): p. 1000-9.
21. Kowal, E.E., *Genetic research in Indigenous health: significant progress, substantial challenges.* Med J Aust, 2012. **197**(1): p. 19-20.
22. Fullerton, S.M. and S.S. Lee, *Secondary uses and the governance of de-identified data: lessons from the human genome diversity panel.* BMC Med Ethics, 2011. **12**: p. 16.
23. Grounds, R.A., *The Yuchi Community and the Human Genome Diversity Project: Historic and Contemporary Ironies.* Cultural Survival Quarterly, 1996. **20**(2).
24. McGregor, J.L., *Population Genomics and Research Ethics with Socially Identifiable Groups.* Jounral of Law, Medicine, and Ethics, 2007: p. 356-370.
25. Corbie-Smith, G., *The Continuing Legacy of the Tuskegee Syphilis Study: Consideratios for Clinical Investigation.* The American Journal of Medical Sciences, 1999. **317**(1): p. 5-8.
26. Foster, M.W., et al., *The role of community review in evaluating the risks of human genetic variation research.* Am J Hum Genet, 1999. **64**(6): p. 1719-27.
27. Goering, S., S. Holland, and K. Fryer-Edwards, *Transforming Genetic Research Practices with Marginalized Communities: A case for representative justice.* Hastings Center Report, 2008. **38**(2): p. 43-58.
28. Sharp, R.R. and M.W. Foster, *Community Involvement in the Ethical Review of Genetic Research: Lessons from American Indian and Alaska Native Populations.* Environmental Health Perspectives, 2002. **110**: p. 145-148.
29. Underkuffler, L.S., *Human Genetics Studies: The Case for Group Rights.* Journal Of Law, Medicine, and Ethics, 2007: p. 383-395.
30. Foster, M.W., D. Bernsten, and T.H. Carter, *A model agreement for genetic research in socially identifiable populations.* Am J Hum Genet, 1998. **63**(3): p. 696-702.
31. Foulks, E.F., *Misalliances in the Barrow Alcohol Study.* American Indian and Alaska native mental health research : journal of the National Center, 1989. **2**(3): p. 7-17.
32. Tsosie, R., *Cultural Challenges to Biotechnology: Native American Genetic Resources and the Concept of Cultural Harm.* Jounral of Law, Medicine, and Ethics, 2007. **35**(3): p. 396-411.
33. Tallbear, K., *DNA, Blood, and Racializing the Tribe.* Wicaso Sa Review, 2003. **18**(1): p. 81-107.
34. TallBear, K. *Native American DNA : tribal belonging and the false promise of genetic science*. 2013; Available from: http://public.eblib.com/choice/publicfullrecord.aspx?p=1362022.
35. Tsosie, R., *Indigenous People and Epistemic Injustice: Science, Ethics, and Human Rights.* Washington Law Review, 2012. **87**: p. 1133-.
36. Schroeder, K.B., R.S. Malhi, and D.G. Smith, *Opinion: Demystifying Native American genetic opposition to research.* Evolutionary Anthropology: Issues, News, and Reviews, 2006. **15**(3): p. 88-92.
37. Bustamante, C.D., E.G. Burchard, and F.M. De la Vega, *Genomics for the world.* Nature, 2011. **475**(7355): p. 163-5.

38. Boyer, B.B., et al., *Building a community-based participatory research center to investigate obesity and diabetes in Alaska Natives.* International Journal of Circumpolar Health International Journal of Circumpolar Health, 2005. **64**(3).
39. Cavalli-Sforza, L.L., *Genes, Peoples, and Languages*. 2000, New York, NY: North Point Press. 228.
40. Hunley, K.L., et al., *A formal test of linguistic and genetic coevolution in native Central and South America.* Am J Phys Anthropol, 2007. **132**(4): p. 622-31.
41. Foster, M.W. and R.R. Sharp, *Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity.* Genome Res, 2002. **12**(6): p. 844-50.
42. Lewis, C.M., Jr., et al., *Land, language, and loci: mtDNA in Native Americans and the genetic history of Peru.* Am J Phys Anthropol, 2005. **127**(3): p. 351-60.
43. Krithika, S., S. Maji, and T.S. Vasulu, *A microsatellite study to disentangle the ambiguity of linguistic, geographic, ethnic and genetic influences on tribes of India to get a better clarity of the antiquity and peopling of South Asia.* Am J Phys Anthropol, 2009. **139**(4): p. 533-46.
44. Chakraborty, R., *Cultural, language, and geographical correlates of genetic variability in Andean highland Indians.* Nature, 1976. **264**: p. 350-352.
45. Pagani, L., et al., *Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool.* Am J Hum Genet, 2012. **91**(1): p. 83-96.
46. Harding, R.M. and R.R. Sokal, *Classification of the European language families by genetic distance.* Proc Natl Acad Sci U S A, 1988. **85**: p. 9370-9372.
47. Sims-Williams, P., *Genetics, linguistics, and prehistory: thinking big and thinking straight.* Antiquity, 1998. **72**(277): p. 505-527.
48. Cavalli-Sforza, L.L., *"Genes, Peoples, and Languages".*
49. Ward, R.H., et al., *Genetic and Linguistic Differentiation in the Americas.* Proceedings of the National Academny of Sciences of the United States of America, 1993. **90**(22): p. 10663-10667.
50. Sandoval, K., et al., *Linguistic and maternal genetic diversity are not correlated in Native Mexicans.* Hum Genet, 2009. **126**(4): p. 521-31.
51. Quinto-Cortes, C.D., et al., *Genetic characterization of indigenous peoples from Oaxaca, Mexico, and its relation to linguistic and geographic isolation.* Hum Biol, 2010. **82**(4): p. 409-32.
52. Jones, D., *Kinship and Deep History: Exploring Connections between Culture Areas, Genes, and Languages.* American Anthropologist, 2003. **105**(3): p. 501-514.
53. Knowler, W.C., *Diabetes melitus in the pima indians: Incidence, risk factors, and pathogenesis.* 1990.

CHAPTER 5: CONCLUSION

5.1 Conclusion

Unique genetic patterns and environmental exposures in Alaska Native communities affect average warfarin dose requirements and risk of vitamin D deficiency, illustrating the importance of expanding biomedical research to include more diverse populations. If historically underserved and marginalized communities are not included in research, results derived from research with other populations are used to inform medical decisions for all people, which can lead to sub-optimal care for overlooked groups if the patterns of genetic variation and environmental exposures are not the same across populations. This type of generalization can perpetuate health disparities. To include more communities in research and reverse this trend, however, ethical concerns of communities must be addressed. Responsive justice may be an appropriate framework to use in approaching community partnerships and re-evaluating research procedures.

Genetic variation affecting the warfarin dose-response in Alaska Native people is an example of how including historically underserved populations in research can reveal clinically important information. Genetic variation can alter critical enzyme functions and affect the warfarin dose-response relationship. As described in Chapter 2, we resequenced genes that affect the metabolism of and response to warfarin, and found two significant novel and common coding variants, *M1L* and *N218I,* in *CYP2C9*, both of which likely reduce or eliminate enzyme function. This type of variation would reduce warfarin clearance and necessitate a lower dose for the people carrying either variant, which would affect the treatment of approximately 10% of Yup'ik people and 3% of SCF customer-owners*.* However, known variants in *CYP2C9* that reduce enzyme function were found at lower frequencies in AI/AN populations compared to some other world populations, such that the overall fraction of all populations carrying loss of function *CYP2C9* alleles is about the same. Applying sequencing technology to identify the *CYP2C9 M1L* and *N218I* variants illustrates the need for including more populations in genetic research and for looking for variation beyond what has been identified and analyzed in other populations. If only variation that has been identified in the populations traditionally included in genetic research is used to inform warfarin dose requirement decisions, a significant fraction of the population carrying the potentially function-disrupting *M1L* and *N218I* variants would be miss-classified, and Alaska Native people with those variants would be dosed inappropriately.

Population frequencies of variants provide a useful starting point for estimating average warfarin dosing requirements, but subgroups and individuals within the population may respond quite differently from each other. For example, the reduced function haplotype in *VKORC1* was found at high frequency in Alaska Native communities, especially in Yup'ik communities, compared to other populations and would predict a lower warfarin dose requirement for carriers of that haplotype. In contrast, the vitamin K conserving variant *CYP4F2*3* was identified at one of the highest frequencies in any population studied to date and would be predicted to increase the required warfarin dose for individuals carrying the variant. In addition, any one of the reduced activity *CYP2C9* alleles seen in the population would predict a lower warfarin dose to achieve the target INR response. Personalizing care for an individual would require individual genotyping information to determine the composition of variation in *CYP2C9, VKORC1*, and *CYP4F2,* which would vary for each person and lead to different requirements for optimal warfarin therapy.

While this study identified novel variation with potentially important effects on enzyme function and drug response, no phenotypes were evaluated in our investigation. An ongoing study being conducted at the Southcentral Foundation should provide this critical translational genotype-phenotype information. However, besides genetic variation, warfarin therapy is affected by diet, especially vitamin K intake, as well as age, BMI, and gender, among other factors. An algorithm developed by the International Warfarin Pharmacogenetics Consortium includes race [1], a variable that attempts to capture uncharacterized population-specific variability that modifies the associations of genetic and demographic variation with dose. These crude racial categories are "White", "Black", "Asian," and "Mixed or missing," and it is unknown how the associations between genetics, demographics, diet and dose in Alaska Native populations might differ. Future studies that connect genotype, demographic and environmental factors and clinical dosing data in these populations are needed to confirm the phenotypic effects of the novel and known variants on warfarin dosing and bleed risk in Alaska Native people. Finally, in addition to evaluating the associations of known variation with warfarin dose, the effects of the two novel variants we discovered must be studied. These associations are important to establish to improve warfarin dose management in Alaska Native populations. Both in vitro (cell-based) testing and clinical testing studies in healthy volunteers by the investigative team are planned.

Although genetic analysis in two cohorts of Alaska Native people revealed new patterns of variation that have important implications for warfarin dosing and outcomes, the study of associations with serum 25(OH)D_3 concentrations in the Yup'ik people, presented in Chapter 3, show the importance of unique environmental exposures on phenotype and that genotype-phenotype associations found in one population are not necessarily as significant in other populations. While $25(OH)D_3$ concentrations in many populations vary most significantly with changes in sunlight exposure that affect vitamin D_3 synthesis in the skin, the serum concentrations of $25(OH)D_3$ in the Yup'ik study participants were found to vary most significantly with variation in the consumption of the traditional Yup'ik diet, which is rich in vitamin D_3. Additionally, although all 8 SNPs interrogated in *CYP2R1* and *DHCR7* have been associated with $25(OH)D_3$ concentrations in other populations [2-4], only one of them was found to be associated with $25(OH)D_3$ concentrations in the Yup'ik participants. Even then, that one SNP accounted for only 1% of variation in serum $25(OH)D_3$ concentration. Unlike the variation in sunlight exposure and genetic predictors that may be used to assess risk for vitamin D deficiency in other populations, the risk factors of concern for Yup'ik people are dietary, linked with deviation from a traditional diet. As a result, risk for deficiency is a greater concern for children and young adults because they consume more market-based foods associated with a western lifestyle, which are poor sources of vitamin D.

Understanding vitamin D concentrations and sources of variation are important for assessing disease risk and also for improving drug dosing in Yup'ik populations. Vitamin D deficiency is associated with increased risk for illnesses that are being seen in increasing prevalence among Alaska Native and Canadian First Nations populations, including rickets and colon cancer [5-9]. Both of these diseases are also associated with dietary patterns, including increased consumption of foods high in animal fat and poor in nutrients like vitamin D [10]. In addition to its role in disease, Vitamin D modulates the expression of CYP3A4 in the intestine, which affects the first pass metabolism and oral bioavailability of some common drugs, including statins used to treat high cholesterol and tamoxifen used to treat breast cancer [11, 12]. With vitamin D concentrations that fluctuate, the bioavailability of a drug can vary with changes in CYP3A4 expression, which can change efficacy and toxicity for a given dose [12]. Close to the arctic, variation in sunlight exposure affecting vitamin D synthesis can lead to fluctuations in CYP3A4 expression, but consistent dietary intake of vitamin D, such as in the traditional Yup'ik diet, can stabilize these concentrations. Indeed, the older Yup'ik participants who consume greater amounts of the traditional diet show less variability in 25(OH)D_3 concentrations by season. By recognizing the importance of diet in maintaining sufficient 25(OH)D_3 concentrations for people in the Y-K Delta, interventions highlighting the value of the traditional diet can prevent disease and stabilize drug response in these communities. Additional studies are needed to confirm the association of variable 25(OH)D_3 concentrations on drug metabolism and disease risk in this population.

While the association of genotype with 25(OH)D_3 concentrations was small in this study, the results of the warfarin study and the heritability measure of 0.46 for 25(OH)D_3 concentration suggest that novel genetic variation or unique linkage disequilibrium patterns in these two genes, or in others in the vitamin D pathway, may still affect serum 25(OH)D_3 concentrations in this population. Additionally, the study was not powered to detect gene-environment interactions, but these effects could be important determinants of 25(OH)D_3 concentrations and also could be specific to this population; for example, interactions between sunlight variation, diet, and genetic variation in relevant enzymes. This study highlights the need for biomedical and genetic research with individual communities, as the underlying causes of and associations with vitamin D deficiency are different in the Yup'ik population compared to other populations [13].

The studies presented in Chapters 2 and 3 on genetics related to warfarin metabolism in the Alaska Native population and the variables affecting vitamin D status in the Yup'ik people of the Y-K Delta illustrate how population-specific genetic variation and environmental exposures can affect medical care, and show the importance of including diverse populations in biomedical and genetic research. In order to conduct this type of population-specific research, however, researchers must use transformed research practices, which place greater value on the interests and concerns of the communities by making them partners in research. As an example of applying responsive justice to expand biomedical research to more communities, in Chapter 4, I explored the problem of adjusting for population substructure in genetic epidemiologic research with identifiable communities. While the statistical standard is to account for population substructure to avoid possible confounding of results, the preferred method increases risks for harms to the community and may make a community choose not to participate in the research. By applying responsive justice to acknowledge and address community interests and concerns, researchers can conduct meaningful research that also aligns with community values and priorities. While researchers may be able to approximate population stratification in a way that respects community needs, stratifying by ancestral language does not seem to be a reliable method. However, future studies on the impact of population stratification on research conclusions may clarify whether adjustment is needed at all for meaningful and actionable conclusions. Some types of study design, such as candidate gene analysis, may not be as susceptible to confounding due to population stratification.

Historically underserved and identifiable communities have valid reasons to be hesitant to participate in medical and, especially, genetic research. However, because of population specific genetic variation and environmental exposure patterns affecting drug response and disease risk, this research is important for reducing health disparities. Based on the identification of novel variation and unique genetic patterns, the average Alaska Native patient may need a lower warfarin dose than what is needed on average in other populations, and starting Alaska Native patients at a lower dose or genotyping individuals prior to initiating therapy could reduce the frequency of adverse bleeding events from warfarin overdose. Additionally, understanding the importance of traditional Yup'ik foods in maintaining adequate vitamin D levels illuminates an approach to helping members of the community improve their vitamin D status to prevent the illnesses associated with vitamin D deficiency and to manage unstable drug dosing from altered CYP3A4 expression in the intestine. If research partnerships are approached effectively, such as through a framework of responsive justice, historically underserved and marginalized communities can benefit from and be actively engaged in population-specific medical research.

5.2 References

1. Consortium, I.W.P., *Estimation of the warfarin dose with clinical and pharmacogenetic data.* N Engl J Med, 2009. **360**: p. 753-764.
2. Ramos-Lopez, E., et al., *CYP2R1 (vitamin D 25-hydroxylase) gene is associated with susceptibility to type 1 diabetes and vitamin D levels in Germans.* Diabetes/Metabolism Research and Reviews, 2007. **23**: p. 631-636.
3. Engelman, C.D., et al., *Vitamin D intake and season modify the effects of the GC and CYP2R1 genes on 25-hydroxyvitamin D concentrations.* J Nutr, 2013. **143**(1): p. 17-26.
4. Kuan, V., et al., *DHCR7 mutation linked to higher vitamin D status allowed early human migration to Northern latitudes.* Evolutionary Biology, 2013. **13**(144).
5. Singleton, R., et al., *Rickets and vitamin D deficiency in Alaska native children.* Journal of Pediatric Endocrinology and Metabolism, 2015. **0**(0).
6. Sharma, S., et al., *Vitamin D deficiency and disease risk among aboriginal Arctic populations.* Nutr Rev, 2011. **69**(8): p. 468-78.
7. Perdue, D.G., et al., *Regional differences in colorectal cancer incidence, stage, and subsite among American Indians and Alaska Natives, 1999-2004.* Cancer, 2008. **113**(5 Suppl): p. 1179-90.
8. Howard, B.V., et al., *All-cause, cardiovascular, and cancer mortality in western Alaska Native people: Westen Alaska Tribal Collaborative for Health (WATCH).* Am J Public Health, 2014. **104**(7): p. 1334:40.
9. Friborg, J.T. and M. Melbye, *Cancer patterns in Inuit populations.* The Lancet Oncology, 2008. **9**(9): p. 892-900.
10. Byers, T., *Nutrition and Cancer among American Indians and Alaska Natives.* American Cancer Society, 1996. **78**(7).
11. Goh, X.W., C.H. How, and S. Tavintharan, *Cytochrome P450 drug interactions with statin therapy.* Singapore Medical Journal, 2013. **54**(3): p. 131-135.
12. Teft, W.A., et al., *CYP3A4 and seasonal variation in vitamin D status in addition to CYP2D6 contribute to therapeutic endoxifen level during tamoxifen therapy.* Breast Cancer Res Treat, 2013. **139**(1): p. 95-105.
13. Hilger, J., et al., *A systematic review of vitamin D status in populations worldwide.* Br J Nutr, 2014. **111**(1): p. 23-45.