ARIC Manuscript Proposal #2984

PC Reviewed: 5/9/17	**Status: _____**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: 1000 Genome Imputed GWAS of Anthropometric Traits: Giant Consortium

b. Abbreviated Title (Length 26 characters): **1000 Genome GWAS of Obesity Traits**

2. Writing Group: Kristin Young, Misa Graff, Anne Justice, Heather Highland, Kari North, Eric Boerwinkle, and Ellen Demerath, and other interested ARIC Co-Authors.

Giant Writing group members: Joel Hirschhorn, Elizabeth Speliotes, Ruth Loos, Cecelia Lindgren, Tim Frayling, Mark McCarthy, and other members of the GIANT Consortium

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

First author: Kristin Young
Assistant Professor
Department of Epidemiology
University of North Carolina
Chapel Hill, NC
919-966-8491; Fax 919-966-9800
kristin.young@email.unc.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

Kari E. North
Professor
Department of Epidemiology
University of North Carolina
Chapel Hill, NC
919-966-2148; Fax 919-966-9800
kari_north@unc.edu

3. Timeline:

Statistical analyses: July 2017
Manuscript preparation: July 2018
Manuscript revision: January 2019
Manuscript submission: July 2019

4. Rationale:

***Obesity is an enormous public health problem*[1] with suboptimal therapies and preventive measures.** Over 30% of adults in the United States are currently obese[2-4], with huge implications for future obesity-related health costs, morbidity and mortality[2,5-7]. The accompanying increase in diabetes is especially alarming; like obesity itself, the impact is greater for underrepresented minorities[8-10]. Diet and lifestyle changes have only modest long-term efficacy. Pharmacologic therapies can enhance lifestyle modifications, but have side effects and only moderate efficacy[11-14]. Bariatric surgery, while effective, carries risk and is reserved for the most severe cases[12,15,16]. Thus, in parallel with continued public health efforts, there is a clear unmet need for new, effective therapies and interventions for the vast majority of the obese and overweight population[17].

Despite initial successes, there remains an urgent need for additional genetic studies of obesity to identify relevant biology. GWAS have been performed for different measures of obesity, including BMI, WHR, and fat mass[18-38]; many of these were carried out by the GIANT consortium and with leadership by our team[18-23,33-38]. We have identified >200 loci that have begun to point to relevant biology. Our initial studies showed that many BMI loci contain genes expressed in the brain; more recently, we used DEPICT to highlight gene sets related to synaptic plasticity and neurotransmitter release[22,33]. Studies of WHR point to genes expressed in fat cells and to insulin action[34]. However, at many loci there is no connection to relevant gene sets, nor is the delineation of enriched biological pathways yet sharp enough to identify candidate genes with certainty. There are clearly enough biological connections across loci to know we are headed in the right direction. But, to generate relevant therapeutic hypotheses for obesity, we need both larger studies to expand on initial successes, and complementary approaches to attain additional biological insights.

Height is a classical model polygenic trait that has enabled the development, testing, and refinement of genetic methods. We study adult height to gain insights into normal and abnormal skeletal growth, and also to better understand how to decipher the genetic architecture of polygenic traits. Height has been a model trait since the dawn of genetics: it was used by Galton to coin the term "regression" and by Fisher to describe the concept of a polygenic trait[39,40]. More recently, genetic results for height (including our data) have been used to develop many methods, including better estimators of heritability[41-45], and to show that confounding by ancestry is an important consideration[45,46]. Because height is highly heritable and easily measured, GWAS of height have found more associated loci than studies of any other polygenic trait. Furthermore, much of the relevant biology is defined by >200 monogenic disorders of abnormal skeletal growth, and aligns well with results of GWAS[47,48]. This known biology provides an invaluable positive control for methods designed to move from genetic association to biological insight. Thus, genetic data for height remain highly valued by leaders in development of statistical genetic and computational methods.

Samples for GWAS. In total, we have already generated or received association results from >636,000 samples from 198 cohorts, imputed to Phase 1 of 1000 Genomes. We have agreements from many additional collaborators (N~800,000). We will incorporate data from the UK Biobank (N~500,000), for which we have access and approval to perform anthropometric analyses. We have a long track record of attracting new collaborators, so we expect the sample size to continue to grow beyond the currently committed ~1.6 million samples from five major ethnic/ancestry

groups (European ancestry, largely African ancestry, East Asian ancestry, South Asian ancestry, and Hispanic ethnicity). For example, the Million Veterans Program (N ~400,000) is committed to working with us during the timeframe of this proposal. Thus, we are confident that the meta-analysis of deeply imputed genotypes will encompass at least 2 million samples from multiple ancestry groups. This sample size, which is several times larger than any completed GWAS, will dramatically improve power for genetic discovery.

Reference panels for imputation. Imputation panels have been consistently improving along several axes: sample size (allowing imputation to lower allele frequencies), representation of multiple ancestries (allowing accurate imputation in more diverse samples) and inclusion of more types of variation (such as indels). Because of the constant improvement, there is a temptation to wait for the next and better panel; indeed, the GIANT consortium have decided not to publish a meta-analysis of currently available data imputed to Phase 1 of the 1000 Genomes[49], given the current availability of the HRC panel (30 times as large[50]) and Phase 3 of the 1000 Genomes (more diverse, and containing indels). In theory, we could continue to delay and wait for newer panels (such as those that will emerge from studies such as TOPMed). After discussion, we decided to proceed with the plan described in this proposal: impute all samples to Phase 3 of 1000 Genomes[51] (to provide a uniform framework with indels and allow inclusion of non-European ancestry samples) and impute European-ancestry samples to HRC (to access variants with lower allele frequencies). We will likely reimpute to larger, more diverse and more comprehensive panels at a later date. The existence of imputation servers, such as the one at the University of Michigan, should enable most groups to reimpute. Where necessary, we also can use methods[52-54] (some of which we helped develop) that estimate the results of association studies from more deeply imputed genotypes by combining LD relationships with the summary statistics from results imputed to a prior reference panel.

Infrastructure for imputation. An imputation server (https://imputationserver.sph.umich.edu) has been developed by the Michigan team that enables imputation from genotypes (typically VCF files) to several reference panels. For this proposal, the relevant panels are HRC version r1.1 and 1000 Genomes Phase 3, but the server will continue to be updated to allow future reimputation. In addition to providing free (to the user) imputation with standardized versions of the reference panels, the server automatically performs basic checks (SNP names and alleles, minor allele frequencies). If needed, the server also will perform prephasing using Eagle[55] (which uses reference panel haplotypes to improve phasing).

5. Main Hypothesis/Study Questions:

To conduct genome-wide association analyses of adiposity traits (BMI, WHR, and height) using the ARIC data to participate in the GIANT Consortium.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

Anthropometric phenotypes. In our analysis plans, we ask for analyses of height, body mass index (BMI, weight/height^2), and WHR (the ratio of waist to hip circumference). WHR is analyzed both adjusted for BMI and unadjusted to test for collider bias[56]. Individuals <18 yrs (not fully grown) and pregnant women are excluded. After correction for covariates (sex, age,

age[2], and study specific covariates such as recruitment center), phenotypes are inverse normal transformed to eliminate artifacts from outliers, especially key for rare variants.

Association analysis and meta-analysis for individual variants. There are a large number of association methods for studying individual variants. Because the deeper imputation panels will enable us to assay variants with a lower range of allele frequencies, we will ask cohorts to use a method that will also enable us to study combinations of variants (for example, gene-based tests of coding variants; see Aims 2 and 3 for more details). We currently ask analysts to use rvtests[57] (version 20151007), which allows us to perform meta-analysis of individual and aggregated sets of variants using RAREMETAL[58]. Analysts will provide both summary association statistics for each variant and also a large covariance matrix that captures study-specific LD relationships within windows spanning 500kb on either side of each variant.

Approach to accounting for ancestry within and across studies. Our previous work largely used principal components (PCs) of ancestry as covariates to account for population substructure (stratification[46,59]). This worked fairly well, resulting in only a small amount of inflation attributable to stratification for height[45,47,60,61]. Mixed model approaches (such as implemented in EMMAX[61] or BOLT-LMM[62]) are slightly more effective than PCs, especially in the setting of rare variation. Moving forward, we will request that analysts use a mixed model approach (implemented in rvtests through creation of a genetic relatedness matrix). We continue to monitor new developments in meta-analysis, including approaches that group studies by ancestry during analysis and facilitate identification and interpretation of signals that vary in intensity across ancestries (either because of allele frequency differences or because of other sources of heterogeneity), such as MANTRA[63].

Fine mapping signals of association from analyses of individual variants. Another advantage of requesting covariance matrices is that we will be able to use these in combination with summary association statistics to perform conditional analysis and identify which variant(s) might be driving association signals in any given locus. We will also use fine-mapping methods to generate credible sets of variants (one method that performs well in certain scenarios is CAVIAR-BF[64]; we will keep abreast of this rapidly moving literature).

Aggregate variants within genes and noncoding regions. We will perform the aggregate tests of rare and low frequency (RLF) variants that will be enabled by having requested variant covariance matrices, for RLF nonsynonymous and noncoding variants.

7.a. Will the data be used for non-CVD analysis in this manuscript? ____ Yes __X__ No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = "CVD Research" for non-DNA analysis, and for DNA analysis RES_DNA = "CVD Research" would be used? ____ Yes ____ No
(This file ICTDER has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript? __X__ Yes ____ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = "No use/storage DNA"? ___X_ Yes ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/ARIC/search.php

__X____ Yes _____ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

This is an extension of **MS1500** that has been used to publish several GIANT GWAS studies.

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? ____ Yes X____ No

11.b. If yes, is the proposal

___ A. primarily the result of an ancillary study (list number* _____)
___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* _____ _____ _____)

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PubMed Central** whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.

13. Per Data Use Agreement Addendum, approved manuscripts using CMS data shall be submitted by the Coordinating Center to CMS for informational purposes prior to publication. Approved manuscripts should be sent to Pingping Wu at CC, at pingping_wu@unc.edu. I will be using CMS data in my manuscript ____ Yes __X__ No.

1. Swinburn, B.A. *et al.* The global obesity pandemic: shaped by global drivers and local environments. *Lancet* **378**, 804-14 (2011).
2. Ogden, C.L., Carroll, M.D., Kit, B.K. & Flegal, K.M. Prevalence of childhood and adult obesity in the United States, 2011-2012. *JAMA* **311**, 806-14 (2014).
3. Flegal, K.M., Kruszon-Moran, D., Carroll, M.D., Fryar, C.D. & Ogden, C.L. Trends in Obesity Among Adults in the United States, 2005 to 2014. *JAMA* **315**, 2284-91 (2016).
4. Ogden, C.L., Carroll, M.D. & Flegal, K.M. Prevalence of obesity in the United States. *JAMA* **312**, 189-90 (2014).
5. Flegal, K.M., Graubard, B.I., Williamson, D.F. & Gail, M.H. Cause-specific excess deaths associated with underweight, overweight, and obesity. *JAMA* **298**, 2028-37 (2007).
6. Kim, D.D. & Basu, A. Estimating the Medical Care Costs of Obesity in the United States: Systematic Review, Meta-Analysis, and Empirical Analysis. *Value Health* **19**, 602-13 (2016).
7. Tsai, A.G., Williamson, D.F. & Glick, H.A. Direct medical cost of overweight and obesity in the USA: a quantitative systematic review. *Obes Rev* **12**, 50-61 (2011).
8. Boggs, D.A. *et al.* General and abdominal obesity and risk of death among black women. *N Engl J Med* **365**, 901-8 (2011).
9. Menke, A., Casagrande, S., Geiss, L. & Cowie, C.C. Prevalence of and Trends in Diabetes Among Adults in the United States, 1988-2012. *JAMA* **314**, 1021-9 (2015).
10. Ludwig, J. *et al.* Neighborhoods, obesity, and diabetes--a randomized social experiment. *N Engl J Med* **365**, 1509-19 (2011).
11. Barja-Fernandez, S., Leis, R., Casanueva, F.F. & Seoane, L.M. Drug development strategies for the treatment of obesity: how to ensure efficacy, safety, and sustainable weight loss. *Drug Des Devel Ther* **8**, 2391-400 (2014).
12. Kushner, R.F. Weight loss strategies for treatment of obesity. *Prog Cardiovasc Dis* **56**, 465-72 (2014).
13. Yanovski, S.Z. & Yanovski, J.A. Long-term drug treatment for obesity: a systematic and clinical review. *JAMA* **311**, 74-86 (2014).
14. Khera, R. *et al.* Association of Pharmacological Treatments for Obesity With Weight Loss and Adverse Events: A Systematic Review and Meta-analysis. *JAMA* **315**, 2424-34 (2016).
15. Marcotte, E. & Chand, B. Management and Prevention of Surgical and Nutritional Complications After Bariatric Surgery. *Surg Clin North Am* **96**, 843-56 (2016).
16. Arterburn, D.E. & Courcoulas, A.P. Bariatric surgery for obesity and metabolic conditions in adults. *BMJ* **349**, g3961 (2014).
17. Ahima, R.S. Digging deeper into obesity. *J Clin Invest* **121**, 2076-9 (2011).
18. Heid, I.M. *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet* **42**, 949-60 (2010).
19. Kilpelainen, T.O. *et al.* Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. *Nat Genet* **43**, 753-60 (2011).
20. Lindgren, C.M. *et al.* Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS Genet* **5**, e1000508 (2009).
21. Loos, R.J. *et al.* Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* **40**, 768-75 (2008).
22. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42**, 937-48 (2010).
23. Willer, C.J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* **41**, 25-34 (2009).
24. Wen, W. *et al.* Genome-wide association studies in East Asians identify new loci for waist-hip ratio and waist circumference. *Sci Rep* **6**, 17958 (2016).
25. Hagg, S. *et al.* Gene-based meta-analysis of genome-wide association studies implicates new loci involved in obesity. *Hum Mol Genet* **24**, 6849-60 (2015).
26. Horikoshi, M. *et al.* Discovery and Fine-Mapping of Glycaemic and Obesity-Related Trait Loci Using High-Density Imputation. *PLoS Genet* **11**, e1005230 (2015).
27. Usher, C.L. *et al.* Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nat Genet* **47**, 921-5 (2015).

28. Nettleton, J.A. *et al.* Gene x dietary pattern interactions in obesity: analysis of up to 68 317 adults of European ancestry. *Hum Mol Genet* **24**, 4728-38 (2015).
29. Liu, C.T. *et al.* Genome-wide association of body fat distribution in African ancestry populations suggests new loci. *PLoS Genet* **9**, e1003681 (2013).
30. Ahmad, S. *et al.* Gene x physical activity interactions in obesity: combined analysis of 111,421 individuals of European ancestry. *PLoS Genet* **9**, e1003607 (2013).
31. Graff, M. *et al.* Genome-wide analysis of BMI in adolescents and young adults reveals additional insight into the effects of genetic loci over the life course. *Hum Mol Genet* **22**, 3597-607 (2013).
32. Bradfield, J.P. *et al.* A genome-wide association meta-analysis identifies new childhood obesity loci. *Nat Genet* **44**, 526-31 (2012).
33. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).
34. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187-96 (2015).
35. Lu, Y. *et al.* New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nat Commun* **7**, 10495 (2016).
36. Berndt, S.I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet* **45**, 501-12 (2013).
37. Randall, J.C. *et al.* Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet* **9**, e1003500 (2013).
38. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192-212 (2014).
39. Fisher, R.A. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399-433 (1918).
40. Galton, f. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute* **15**, 246-263 (1886).
41. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47**, 1114-20 (2015).
42. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* **43**, 519-25 (2011).
43. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
44. Visscher, P.M. *et al.* Genome partitioning of genetic variation for height from 11,214 sibling pairs. *Am J Hum Genet* **81**, 1104-10 (2007).
45. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
46. Campbell, C.D. *et al.* Demonstrating stratification in a European American population. *Nat Genet* **37**, 868-72 (2005).
47. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-86 (2014).
48. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-8 (2010).
49. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
50. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).
51. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
52. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906-14 (2014).
53. Park, D.S. *et al.* Adapt-Mix: learning local genetic correlation structure improves summary statistics-based analyses. *Bioinformatics* **31**, i181-9 (2015).
54. Kwan, J.S., Li, M.X., Deng, J.E. & Sham, P.C. FAPI: Fast and accurate P-value Imputation for genome-wide association study. *Eur J Hum Genet* **24**, 761-6 (2016).
55. Loh, P.R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443-1448 (2016).
56. Aschard, H., Vilhjalmsson, B.J., Joshi, A.D., Price, A.L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am J Hum Genet* **96**, 329-39 (2015).

57. Zhan, X., Hu, Y., Li, B., Abecasis, G.R. & Liu, D.J. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423-6 (2016).
58. Feng, S., Liu, D., Zhan, X., Wing, M.K. & Abecasis, G.R. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* **30**, 2828-9 (2014).
59. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
60. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* **19**, 807-12 (2011).
61. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-54 (2010).
62. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-90 (2015).
63. Morris, A.P. Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* **35**, 809-22 (2011).
64. Chen, W. *et al.* Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**, 719-36 (2015).