Biodemographic specifics of the effects of body-mass-index risk alleles identified in genome-wide association studies

Alexander M. Kulminski[1-2], Irina Culminskaya[1-2], Konstantin G. Arbeev[1-2], Liubov Arbeeva[1-2], Svetlana V. Ukraintseva[1-3], Eric Stallard[1-2], Deqing Wu[1-2], Kaare Christensen[4], and Anatoli I. Yashin[1-3]

[1] Center for Population Health and Aging, [2]Social Science Research Institute, [3]Duke Cancer Institute; Duke University, Durham, NC 27708-0408, USA.
[4] The Danish Aging Research Center, University of Southern Denmark, 5000 Odense C, Denmark

INTRODUCTION

Aging of populations in developed countries requires effective strategies to extend healthspan. A promising solution could be to yield insights into genetic predisposition to diseases, their precursors (called endophenotypes [EP]), and mortality. Genome-wide association studies (GWAS) have been thought as a major breakthrough in this endeavor. The optimism is, however, tempered because GWAS face difficulties reflecting important limitations in currently prevailing GWAS strategies [2, 3]. A fundamental source of difficulties in genetics of complex traits characteristic for modern societies is the lack of *direct* evolutionary selection against or in favor of such traits [4]. For example, chronic diseases in late (post reproductive) life cannot be major direct evolutionary force in principle because they do not affect reproductive fitness in the same individuals.

At a first glance, refocusing from genetics of traits in late life to EPs could be a promising solution. However, the same evolutionary constraints still hold because: (i) genes regulating EPs have not been selected against or in favor of pathological EPs causing diseases and (ii) genes regulating normal function of EPs were selected in principally different conditions than those in modern societies.

Evolutionary constraints imply that unconditional connections of genes with traits in late life are unlikely. Rather, genes should be linked to such traits in a complex fashion through different mechanisms specific for a given period of life. Accordingly, the linkage between genes and these traits should be strongly modulated by age-related processes in a changing environment, i.e., by the individuals' life course. Inherent sensitivity of genetic mechanisms of complex health traits to the life course is a key concern as long as genetic discoveries are aimed to improve human health.

Currently prevailing GWAS strategies heavily rely on studies gathering large samples. Basic hypothesis behind such strategy is that phenotypic variance can be explained by alleles with modest effects which gain statistical significance only in large samples. This logic implicitly assumes that genetic effects in different samples should be relatively homogeneous. Conversely, it is also argued that increasing the size of human disease cohorts merely increases the heterogeneity making it even harder to detect true risk alleles [5].

In this paper we re-analyze the associations of SNPs discovered as correlates of body mass index (BMI) in a large-scale meta-analysis [1]. The goal is to better understand advantages and disadvantages of standard GWAS strategy in studying complex traits using, as an example, three successive generations participating in the Framingham Heart Study (FHS), --a study which was a part of meta-analysis in [1]--, and two generations of the participants of the Long Life Family Study (LLFS), --the generations which match those in the FHS.

DATA AND METHODS

The FHS design has been previously described [6]. Briefly, the FHS original cohort was launched in 1948 (N=5,209; 28–62 years of age at baseline). This cohort has been biennially examined during 60 years. The FHS Offspring (FHSO) cohort was launched 22 years later and included respondents (N=5,124) aged 5–70 years at baseline who were mostly biological descendants and their spouses of the FHS participants. The FHSO respondents have been examined about every four years at eight

examinations. The 3rd Generation (3rd Gen) cohort was launched in 2001 and mostly included biological descendants (N=4,095) of the FHSO participants (one baseline examination is available via dbGaP). Measurements of weight and height are available at multiple examinations in FHS/FHSO and one measurement is available in 3rd Gen cohort. Biospecimens were mostly collected in the late 1980s and through 1990s from surviving participants, i.e., after some survival selection for participants of the FHS original and FHSO cohorts. Genotyping of 9,167 FHS participants available for this study was conducted using Affymetrix Gene-Chip Human Mapping 500K array.

The LLFS collected data in about equal proportions at four field centers in Boston, New York, Pittsburg, and Denmark on families showing exceptional longevity. The study eligibility criteria were described in [7]. Virtually all study participants were whites. Briefly, in the U.S., the families eligible for the LLFS must have two living siblings aged 80+ years, two living offspring of one or more of the siblings, and a living spouse of one of the offspring. In addition, the family must demonstrate exceptional longevity based on a Family Longevity Selection Score [8]. In Denmark, individuals who would be aged 90+ years during the study recruitment period were identified in the Danish National Register of Persons. They were contacted to assess the family's eligibility for participation in the LLFS using criteria parallel to that used in the U.S. Information on weight and height from the 4,954 LLFS participants was collected at baseline from 2006 to 2009. Biospecimens were collected at baseline. Genotyping of the LLFS participants was done using Human Omni 2.5 array. The data include information on long-living individuals (N=1,384, probands and siblings), their offspring (N=2,321), and 177 spouses of long-living individuals and 777 spouses of offspring. Due to small number of spouses of the long-living individuals, they were pooled together with spouses of offspring (N=954).

Selection of SNPs. We selected SNPs reported as correlates of BMI (kg/m^2) in [1]. This meta-analysis included FHS but not LLFS. We selected SNPs which were explicitly genotyped in FHS and LLFS. SNPs were matched between studies using proxy SNPs with strongest linkage disequilibrium ($r^2>0.9$) available from the 1000 Genomes project. SNPs have been selected after quality control (Hardy Weinberg Equilibrium $p>0.01$, Mendel's errors <2%, and call rate >90%).

Analysis. We used longitudinal measurements of BMI in the FHS assessed from: (i) 19 examinations in the FHS original cohort and (ii) 8 examinations in the FHSO. These examinations cover the entire range of follow up in these cohorts. Because the LLFS and the 3rd Gen cohorts have data available for examinations at baselines, only cross-sectional data were used. The associations were evaluated using a mixed effects regression model. We used a two-level model to account for potential within-family correlation using data from the LLFS and the 3rd Gen cohorts. For evaluation of cumulative effects a three-level mixed effects regression model was fitted to account for familial and repeated-measurements correlations. The effect size beta was evaluated using additive genetic model (following [1]) with minor allele as an effect allele for BMI. The models were adjusted: (i) for whether the DNA samples had been subject to whole-genome amplification (FHS) and (ii) for field centers (LLFS). All models were adjusted for cross-sectional age as well as for sex and cohort differences, when applicable. Meta-analysis was conducted using *plink* [9].

RESULTS

SNP-BMI associations in Nature meta-analysis and in FHS and LLFS

Table 1 shows the associations for each SNP with BMI reported in [1] as well those evaluated in the pooled sample of all cohorts from the FHS, in the pooled sample of all cohorts from the LLFS, and the associations from the meta-analysis of the results from these FHS and LLFS samples.

Table 1 shows that for three SNPs (rs2860323, rs17782313, and rs734597) the effect sizes look relatively homogeneous across FHS and LLFS, whereas for the other three SNPs (rs527248, rs8055543, and rs28670272) they are not. The effect size for rs527248 is larger in the FHS than in the LLFS whereas the effect sizes for rs8055543 and rs28670272 are larger in the LLFS than in the FHS. Given that the sample size of the FHS is about two-fold larger than that of the LLFS, these

observations show that sample size does not play a pivotal role in the associations of SNPs with such complex phenotype as BMI. Rather, understanding specifics of the studied populations may be more essential. The latter is particularly critical because without understanding such specifics just non-informatively increasing the sample sizes following traditional GWAS may not be efficient.

Table 1. Associations of SNPs with BMI from [1] and those evaluated in the FHS and LLFS

SNP	Nature meta-analysis			FHS, N=8,624			LLFS, N=4,445			FHS/LLFS meta	
	Beta	P	$N_{p=.05}$	Beta	SE	P	Beta	SE	P	Beta	P
rs2860323	-0.31	2.8E-49	13,325	-0.32	0.10	1.5E-03	-0.21	0.13	1.0E-01	-0.28	4.5E-04
rs17782313	0.23	6.4E-42	18,730	0.31	0.09	3.6E-04	0.18	0.12	1.4E-01	0.27	1.6E-04
rs527248	0.22	3.6E-23	24,266	0.19	0.09	4.6E-02	-0.02	0.14	8.7E-01	0.12	1.2E-01
rs8055543	-0.17	2.9E-21	55,314	-0.10	0.10	3.2E-01	-0.34	0.15	2.8E-02	-0.17	4.0E-02
rs734597	0.13	2.9E-20	72,485	0.14	0.10	1.3E-01	0.11	0.13	4.0E-01	0.13	8.7E-02
rs28670272	-0.13	1.2E-18	62,345	0.02	0.09	8.3E-01	-0.14	0.12	2.3E-01	-0.04	5.8E-01

$N_{p=.05}$ denotes sample size which is required to achieve nominal (p=0.05) significance for related individuals given sample sizes reported in [1] given population mean for BMI of 25.8 kg/m^2 (standard deviation is 4.8) observed in the FHS.

Indeed, meta-analysis (Table 1) of the FHS and LLFS results shows that increasing the sample size by pooling the results from these studies improves significance for half SNPs with relatively homogeneous effects, i.e., rs2860323, rs17782313, and rs734597. For the other half (rs527248, rs8055543, and rs28670272), increasing the sample size in non-informative way makes the estimates of p-values worse even in the most favorable case when heterogeneity between studies is disregarded (if not disregarded, the estimates are even worse). This meta-analysis implies that traditional GWAS strategy is at most only 50% effective in these populations.

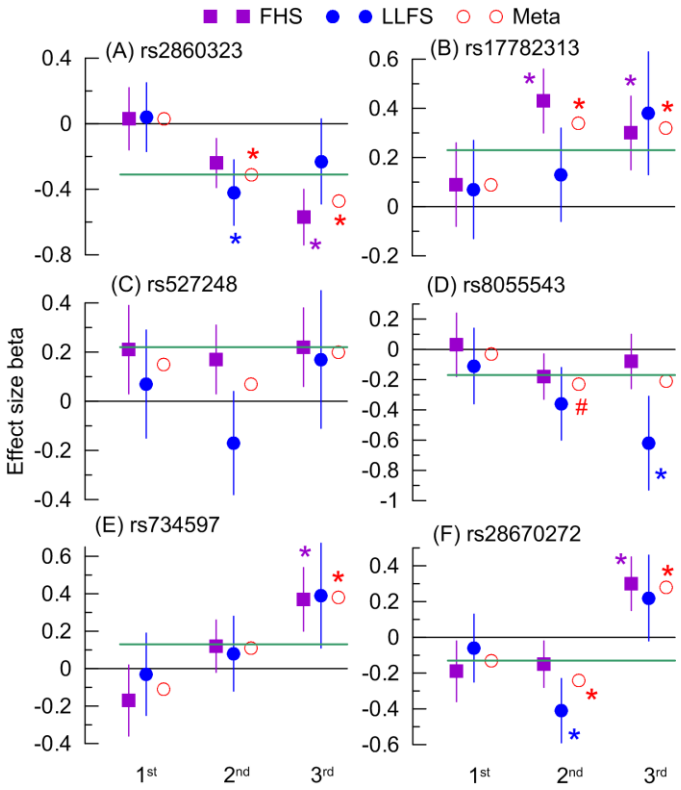

Fig. 1. **Associations of SNPs with BMI across FHS and LLFS cohorts**. Dots show the effect size beta for the minor allele. Solid horizontal green line depicts the effect size beta in [1]. Bars show standard errors. Asterisk and number symbol show significant (p≤0.05) and suggestive-effect (0.05<p≤0.1) associations, respectively.

Cohort-specific associations in the FHS and LLFS

Figure 1 shows the associations of the selected SNPs with BMI in the parental (1st), offspring (2nd), and 3rd Gen (3rd) cohorts of the FHS and the LLFS cohorts of the long-living individuals (1st), their offspring (2nd), and spouses of the long-living individuals and offspring (3rd). The choice of the LLFS cohorts reflects specifics of the LLFS population selected according to chances to live long lives (see Methods). Figure 1 shows dissimilar effects for all SNPs (but rs527248 in the FHS) across the studied cohorts. Notably, larger effect sizes tend to cluster in the 2nd and 3rd cohorts of each study.

Because dissimilarity in the effect sizes implies differences in sample sizes which are required to gain statistical significance, it is desirable to better understand the nature of the observed heterogeneity to improve efficiency of the analyses. One origin of dissimilarity is that it may be the result of stochastic noise. This origin is pursued in most GWAS with arguments that large samples are needed to get robust estimates. The results in the above section suggest, however, that non-informative increase of the

sample size succeeds for half SNPs whereas it does not succeed for the other half. Further, the estimates of the required samples to gain even nominal (p=0.05) significance given effect sizes reported in [1] are far from those in the FHS and LLFS for which significant associations at that level are observed. These observations imply that stochasticity may not be plausible explanation of heterogeneous origin of such a complex trait as BMI.

Another explanation is that the observed dissimilarities in genetic effects are the result of real processes of bio-demographic origin which include demographic processes and biological changes with age. These processes are embedded in specifics of FHS and LLFS cohorts. Cohort 1 in both studies includes largely overlapping birth cohorts. These birth cohorts, however, were subject to different selection processes (see Methods). The FHS cohorts 2 (FHSO) and 3 (3rd Gen) represent two subsequent generations of the parental cohort with the 3rd Gen being the youngest. The LLFS 2nd and 3rd cohorts represent offspring of long living parents and spouses, respectively. They are from about the same birth cohorts which overlap with the FHSO and 3rd Gen cohorts. Analysis of these specifics requires further efforts which are currently underway.

CONCLUSIONS

GWAS often claim the benefits of the large sample sizes achievable through collaboration for detecting risk alleles of complex traits. Implicitly, such strategy relies on existence of unconditional genetic risks that is, generally, questionable. As a consequence, this strategy ignores possible complexity of genetic effects that results in non-informativeness in increasing samples [5]. To better understand pros and cons of traditional GWAS strategy in the analysis of complex traits, we re-examined the associations of SNPs which were identified as correlates of BMI in a recent Nature meta-analysis [1]. Our results suggest that simplistic strategies on increasing sample sizes in large-scale GWAS are at least not efficient. They suggest that gaining insights into bio-demographic specifics of the studied populations may be crucial.

Ethic statement. This study uses de-identified data from the FHS and LLFS, which are available through dbGaP http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v22.p8, and http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000397.v1.p1). No new data were collected in this work.

References

1. Speliotes, E.K., et al., *Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index.* Nat Genet, 2010. **42**(11): p. 937-48.
2. Gibson, G., *Rare and common variants: twenty arguments.* Nat Rev Genet, 2011. **13**(2): p. 135-45.
3. Eichler, E.E., et al., *Missing heritability and strategies for finding the underlying causes of complex disease.* Nat Rev Genet, 2010. **11**(6): p. 446-50.
4. Kulminski, A.M., *Unraveling genetic origin of aging-related traits: evolving concepts.* Rejuvenation Res, 2013. **16**(4): p. 304-12.
5. MacRae, C.A. and R.S. Vasan, *Next-generation genome-wide association studies: time to focus on phenotype?* Circ Cardiovasc Genet, 2011. **4**(4): p. 334-6.
6. Cupples, L.A., et al., *Genetics Analysis Workshop 16 Problem 2: the Framingham Heart Study data.* BMC Proc, 2009. **3 Suppl 7**: p. S3.
7. Yashin, A.I., et al., *"Predicting" parental longevity from offspring endophenotypes: data from the Long Life Family Study (LLFS).* Mech Ageing Dev, 2010. **131**(3): p. 215-22.
8. Sebastiani, P., et al., *A family longevity selection score: ranking sibships by their longevity, size, and availability for study.* Am J Epidemiol, 2009. **170**(12): p. 1555-62.
9. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.