Title of proposed program: Planning a Synthetic Cohort for Longitudinal Analysis of Gene-Environment Interactions

Submitting Source: NIH

What is the major obstacle/challenge/opportunity that the Common Fund should address? There is a pressing need for a large-scale U.S. prospective cohort study of genes and the environment, with a minimal sample size estimated at 500,000 (Collins, 2004; Manolio et al., 2006). Even with a very large cohort, at least 7 to 15 years follow up is needed to accrue enough incident cases to adequately power studies of common diseases (UK Biobank calculations; Burton et al, 2009 IJE). The NIH supports a large number of high-quality cohort studies with rich longitudinal data on health, environmental, behavioral, and social phenotypes that, if strategically coordinated, could help elucidate how genes and environments interact to affect trajectories of health (Willett et al., 2007). NCI's Cohort Consortium demonstrates that very large scale coordination between cohorts is possible and scientifically productive beyond original intentions (e.g., de Gonzalez et al., 2010 *NEJM*). Discovery would be accelerated if the science behind the creation of even larger *synthetic cohorts* from existing NIH studies were supported in the Beyond the GWAS era to embrace novel approaches to discovery such as selection by genotype (aka PheWas or phenotype mining), the investigation of gene-gene interactions, and the systematic analysis of *change phenotypes* over the life course*.*

What would the goals of the program be? As Manolio et al. (2012, AJE) note, to reach their full potential, very large prospective and synthetic cohort studies require new approaches to the entire process of data collection, processing, and ascertainment. This initiative will evaluate design issues for a synthetic cohort study of genes, environment, health, and behavior based on existing NIH-supported cohort studies. These studies already provide rich longitudinal or early life data and are increasingly enhanced through the addition of genetic, biomarker, administrative linkage, and electronic medical records data. But they have not yet reached their potential for understanding trajectories of health and disease. To cite one important example, the p-value associating FTO with BMI and obesity is now infinitesimal, but we do not understand how, why, or when it has its influence on eating, how it operates in non-obesogenic environments, or what other variants may modulate its role. We should, however, be able to synthesize a large enough cohort to find out relatively quickly. (Note: FTO is now "FaT mass and Obesity associated protein." "Fused Toes" was the mouse phenotype caused by deletion of a nearby gene.)

Why is a trans-NIH strategy needed to achieve these goals? What initiatives might form the strategic plan for this topic? To date, attempts to harmonize or synthesize data collection efforts among studies have been modest, in part due to IC boundaries and resource constraints. Although implementing the synthetic cohort design would require modest costs for harmonization and data collection, there are existing platforms for the former (e.g., P3G, www.p3g.org) and a growing number of freely available, high quality measures of phenotypes and exposures developed by the NIH (e.g., instruments from PROMIS, PhenX, the NIH Toolbox and the Health and Retirement Study). Furthermore, HITECH and the ACA could lead to the widespread incorporation of these and other health risk measures in EHRs. We propose that the Common Fund seize this opportunity and invest in three crucial planning and support activities for a synthetic cohort project: (1) Form a harmonization plan for existing NIH cohort studies; (2) Develop the analytical and bioinformatics framework for such studies and support calibration studies to create crosswalks between measures not harmonized at the outset; and (3) Develop plans for data-sharing and consent policies and, crucially, detailed scenarios for the cost and longer term maintenance of synthetic cohorts of varying sizes and intensities. We estimate the cost of this initial developmental and feasibility initiative at $4.3 million per year for three years, with funds to be divided between RMS needs and supplements to cohort study investigators.

If a Common Fund program on this topic achieved its objectives, what would be the impact? The plan for a synthetic cohort for longitudinal analysis of gene-environment interactions would establish the feasibility and cost of an intended scalable synthetic national cohort of people for discovery research in health and disease. It would complement and leverage current and proposed sequencing projects by focusing on phenotype harmonization and large scale design issues to support future G2P and P2G projects as recommended at the 2011 NIH Innovation Brainstorm meeting. Because the founder cohorts are longitudinal, the synthetic cohort would provide rapid access to trajectory information as well as a richer characterization of factors influencing health. This planning activity could also inform the design of a *de novo* national cohort study, or determine if a synthetic cohort effort of five to seven years' duration could provide similar information while affording room for innovation within studies.