

**Notes on the National Institute of Aging (NIA) Population Genetics Conference Call for the Health and Retirement Study (HRS): September 19, 2011**  
**Jennifer Bragg-Gresham and Eimear Kenny**

**On the Call:**

Erica Spotts (*NIA Staff – call organizer*)  
Goncalo Abecasis (*Principal Investigator – University of Michigan*)  
Carlos Bustamante (*Principal Investigator – Stanford University*)  
John Novembre (*Principal Investigator – University of California, Los Angeles*)  
John Ioannidis (*Principal Investigator – Stanford University*)  
Eimear Kenny (*Post-doc, Bustamante Lab – Stanford University*)  
John Haaga (*NIA Staff*)  
Jennifer Harris (*NIA Staff*)  
Jonathan King (*NIA Staff*)  
Georgeanne Patmios (*NIA Staff*)  
John Phillips (*NIA Staff*)

**Purpose of this call:** On September 23-24<sup>th</sup> of 2010 an expert meeting was held to discuss using GWAS to explore fundamental questions about aging in the Health and Retirement Study (HRS). There were no experts in population genetics able to attend, so this call was convened to explore the potential work that can be done in this area. A full summary of the 2010 meeting was circulated to all participants on today's call for review.

**Major Points of Discussion:** (details below)

1. What data is available (collected and coming) and what is the data access procedure
2. Need for a position paper to highlight the NIA-HRS study for the research community
3. Features and opportunities of the study
4. Future possible directions

**Immediate goals:**

1. QC data and compare to other studies
2. Preliminary GWAS for selected traits.
3. Develop ideas for position paper/user's guide

**Actions:**

1. A position paper (intended venue Nature, Science or Nature Genetics) that will:
  - a. Outline the benefits of the HRS, but also include cautions concerning the data use
  - b. Include a few examples of results using the genetic data – possibly replicate known total cholesterol or body mass index (BMI) variants.
2. Dr. Spotts will electronically introduce the group to David Weir, the PI of the HRS.
3. Next call will be scheduled in approximately 6 weeks
4. Dr. Spotts will send email to the group when dbGaP data is available

## **Discussion:**

### **Opening:**

Dr. Spotts began the call by giving a brief introduction to the HRS study: The HRS is funded by the NIA and ongoing for two decades (since 1992). The study cohort consists of ~22,000 individuals over the age of 50, with longitudinal measures via survey of socio-economic (e.g., income, health care expenditure etc.) and health factors (e.g., physical health, cognitive function). A 2005 renewal proposed the collection of DNA from saliva samples and biomarkers from blood serum from respondents (e.g., total cholesterol levels, HbA1c), which was completed for the whole cohort in 2008. The cohort is a population-based sample ascertained from across the United States, with oversampling for African-Americans and Hispanics. In order to maximize the genetic component of this study for researchers, ~13K of the 22K individuals were genotyped on the Illumina 2.5M Omni chip, with further plans to genotype an additional 7K individuals collected in 2010 and 2012, including a large new oversampling of minorities. The goal of the HRS is to make this data available for analysis by the research community and the purpose of this conference call is to discuss the challenges and opportunities for achieving this goal.

### **1. What is the current status of the data?**

The genotypes were cleaned and curated by Dr. Sharon Kardia at the University of Michigan (UM). Dr. Bustamante led a discussion about the need for quality assurance analyses across the current dataset and data harmonization. In particular, to examine the genetic data for batch effects and other quality control issues. For example, it was pointed out that samples from the 2006 batch were collected from buccal swabs and that the quality of these samples should be especially examined since all other samples were collected from saliva.

### **2. At what stage will the data be ready for release?**

The data is currently 3-4 weeks away from being deposited to dbGaP. There are 13,055 samples being prepared which have been genotyped on the Illumina 2.5M Omni chip. These samples came primarily from respondents aged 55-75 years, with 78%, 13% and 9% of respondents reporting European-American, African-American and Hispanic ancestry, respectively. A small subset of the collected phenotypes will be submitted to dbGaP, and these will include a top-coded measure of episodic memory at one time point, age, and gender. Other non-sensitive HRS phenotype data are available over the web (<http://hrsonline.isr.umich.edu/>); linkages between HRS phenotypes and the genotype data deposited at dbGaP will be made available via a licensing arrangement with UM. Some highly restricted variables such as earnings histories and medical claims require further safeguards.

### **3. Procedure for obtaining the data?**

To obtain the data, a request first needs to be made to dbGaP. That request will then be forwarded to UM to obtain the remainder of the phenotypes. Normal IRB approval and review by a data use committee will be required before data is released. Attempts are being made to streamline the process of requesting data.

There is concern about the sensitivity of the phenotype data available. Particularly because there are Medicare claims data that include geographic information and zip codes, earnings histories from the SSA, location of employment, and timing and location of each claim.

#### **4. Need for a position paper to highlight the NIA-HRS study**

The group discussed the need to publish a position paper to describe the features of the NIA-HRS study as a resource for the research community. The paper should also note the limitations of the current dataset; in particular to guide the non-genetic research communities in its use. The group thought that the position paper should include genome-wide association study (GWAS) results for a subset of traits and suggested preliminary analysis should be performed on selected phenotypes.

#### **5. Features and opportunities of the study:**

The first sample of 13K individuals contains an oversampling of minorities (~3K), but over time a smaller number of minorities were maintained due to loss of follow-up. Therefore, minorities have again been oversampled (~3K individuals) in the newer data collected of 7K more individuals. This data is expected to be available in 2013. Ethnicity was self-identified and it was felt it would be very interesting to compare to information concerning genetic origins from the genetic data and to examine the admixture of the US sample. In particular, at the resolution of 2.5M SNP's, there is a chance that not only continental, but also, sub-continental sub-structure might be identified.

A feature of recruitment for the study was that both respondents and their spouses were interviewed and, therefore, a substantial fraction of the cohort are spousal pairs. The both Dr. Bustamante and Novembre emphasized the opportunity to examine whether there is correlation in ancestry among spousal pairs and address questions regarding mating, such as the degree of assortative pairing.

Dr. Abecasis suggested that GWAS be performed for known traits, for which there already exist large-scale meta-analysis, such as total cholesterol levels and HbA1c, to examine whether known hit regions can be replicated in the HRS dataset. He also suggested that the HRS become involved in the larger GWAS collaboration efforts that are ongoing with many of the traits available in the HRS sample.

The dataset also contains socio-economic variables that will be important to account for in future analyses. For a full description and list of variables available, the HRS website (<http://hrsonline.isr.umich.edu/>) was stated as the best resource. In particular, the study is enriched for some psycho-socio measures, such as ordinal personality measures, studies of which are currently under-represented in the genetic literature. However, concern was raised over the possible meta-analyzing studies that did not use the same instruments to gather personality phenotypes.

Finally, the population-based ascertainment of 20K individuals from the US population is a unique aspect of this study. Since the likelihood that many of the phenotypes of interest will vary across ancestries due to non-genetic factors and it will make attention to population stratification a high priority in any GWAS carried out with the sample.

#### **6. Future possibilities:**

There are limitations to what can be discovered using solely the Illumina 2.5M Omni chip, as it focuses mainly on common variants. Four possible directions were suggested for the future, as follows:

- A. Genotyping individuals on new arrays being designed with rare, but known functional variants. There have been approximately 250K functional rare variants (identified from sequencing efforts on 12K individuals of European and African-American ancestry) included on the chip. The chips cost approximately \$40 each and processing would range from \$10-\$20 per person. Only protein coding regions are covered.
- B. Sequencing the whole genome at low coverage (~6-8x) would yield approximately 30-40 million variants. This option is more expensive, approximately \$400 per individual, but would allow for the assessment of structural variation. It was suggested that a subsample of individuals could be sequenced, but Dr. Abecasis warned against selecting the sample based on one or a few traits, since it will hurt the chances of finding associations for other traits. The group was very interested in this, although low pass sequencing would not allow for analysis of single individuals mutations, rather it would target variants present in at least 4-5 people in the sequenced cohort.
- C. The third option discussed was whole exome sequencing, which currently costs \$500-\$1,000 per individual.
- D. High coverage whole genome sequencing was also discussed, but is currently cost prohibitive at \$2,000-\$2,500 per individuals. It was suggested that the group consider this option in 18 months to 2 years as prices should be lower. The advantages of this data would be to examine copy number variation (CNV) and structure in detail. There are currently arrays designed to assess CNV's, but they can only find large variants. CNVs occur in only 5% of the genome and they reside in areas that are hard to read and are usually outside of the arrays. Dr. Abecasis stated that the arrays miss the 20% of the genome that contain > 50% of the CNVs.

In general, the group was most enthusiastic about the future potential for high coverage whole genome sequencing (option D above). They also felt that increasing the number of genotyped individuals beyond the current and planned GWAS cohort (~13K and ~7K participants, respectively) would not be advantageous, but did suggest the strategy of genotyping a subset of the cohort using emerging array technologies that target rare and/or functional variation (option A above). Ultimately the group would like to have data available to examine methylation changes with aging. Blood samples would be needed to make this possible. The group would like to start getting things in place now to make this possible in the future.

In addition, the group asked if new variables could be added to data collection at future follow-up collections. In particular, the group felt it would be informative to know where the grandparents of the participants were born. It was suggested that possibly a pilot question could be added in an experimental group.

## **7. Potential issues:**

A short discussion occurred near the end of the call concerning the consent agreements signed by the patients who gave genetic samples. Dr. Spotts stated that it was quite broad. The group voiced concerns over being sensitive to ethnic groups such as the Native Americans. It was felt it is very important to keep the minority ethnicities involved in the study.