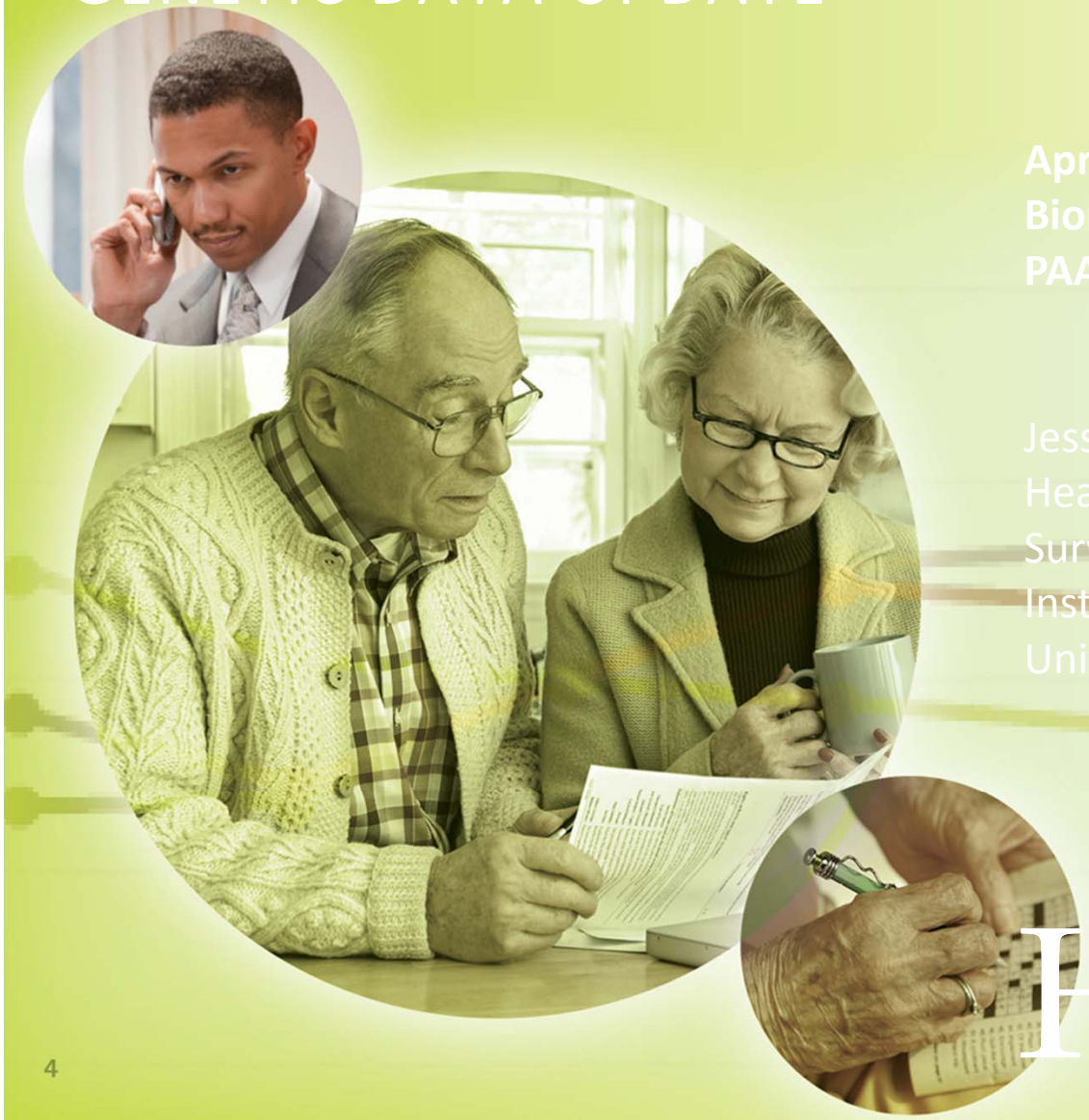
THE HEALTH AND RETIREMENT STUDY: GENETIC DATA UPDATE



April 30, 2014
Biomarker Network Meeting
PAA

Jessica Faul, Ph.D., M.P.H.
Health and Retirement Study
Survey Research Center
Institute for Social Research
University of Michigan

HRS | HEALTH AND RETIREMENT STUDY

HRS – EXPANSION INTO GENETICS

- HRS began collecting DNA in 2006
- Continued to ask for DNA from new respondents and prior refusals
- HRS awarded ARRA grants (RC2 AG036495; RC4 AG039029) to **genotype** our samples collected to date
- Genotyping performed by CIDR
- Illumina HumanOmni2.5 BeadChip v1 – 2.45 million single nucleotide polymorphisms (SNP)
    - SNP and copy number variation (CNV) analysis
    - Sex chromosomes
    - Mitochondrial DNA (< 100)
- Common and rare variants targeting down to 2.5% MAF selected from the 1000 Genomes Project

HRS – EXPANSION INTO GENETICS

- Data deposited into NIH database of Genotypes and Phenotypes (dbGaP)
- Data are free; application required
- **Version 1** - 12,500+ posted to date (2006 + 2008)
  - 1136 Hispanic (9%)
  - 1665 African-American (13%)
  - Posting includes measured SNPs and imputations using 1000 Genomes reference panel (22 million SNPs)
- **Version 2** – 15,600+ samples (2006-2010)
  - Expansion of minority sample
  - 1000 Genomes imputation (22 million SNPs)
  - KING-robust Relationship Matrix

KING-ROBUST ALGORITHM

- Accurate specification of familial relationships is crucial for population-based GWAS with unknown family structure.
- Previous algorithms for relationship inference have a strong assumption of homogeneous population structure which can lead to biased results - systematically inflating the degree of relatedness among individuals of the same racial group.
- KING-robust method performs properly even under extreme population stratification.
- The estimated pedigree information provided by KING (such as kinship coefficients) can be used to verify relationships, reconstruct pedigrees and conduct genetic association tests without relying on self-reported pedigree information.

EXOME DATA

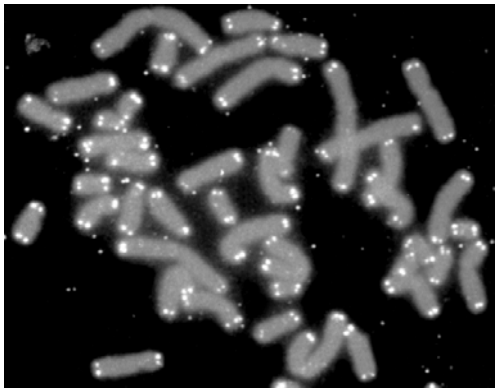
- Illumina HumanExome Beadchip v1.1
- **Exome** - measures genetic variants within the protein-coding region of the genome
- Focused coverage of exonic regions, but does not include coverage outside of coding regions
- The exonic content consists of > 240,000 markers including rare amino acid substitutions, nonsense mutations, splice site variations, and insertion/deletions as well as common variants
- The mutations for this chip selected from over 12,000 individual exome and while genome sequences.
- Representing diverse populations—including European, African, Chinese, and Hispanic individuals—and a range of common conditions (2 diabetes, cancer, metabolic, and psychiatric disorders).
- Can be used for new studies focused on identifying functionally relevant associations.

2012 SAMPLES

- An additional 3,500 samples currently being genotyped by CIDR using a new Illumina platform
  - Illumina HumanOmni2.5 Exome-8
  - Combination of 2.5M beadchip and Exome v1.2
  - Data should be released by the end of 2014

TELOMERE DATA

- **Telomere length -** (data released Dec 2013)
  - marker of cellular aging
  - telomere ends of chromosome become shorter with each cell division
  - 2008 samples only (n=5,808)
  - HRS data product
  - “Sensitive Health” data product; not restricted



CANDIDATE GENE / SNP FILES

- Not all users need/want full genotype data
- Creation of candidate gene / SNP "sets" for ease of use
- HRS restricted data products
- Initially, 2 broad phenotypic-related sets will be available:

  1) **Cognition**, and 2) **Longevity**

** These do not include a complete list of genes or SNPs potentially associated with these phenotype categories; they is merely a selection of the most biologically promising candidate genes. **

CANDIDATE GENE / SNP FILES

Within each phenotype category, there are multiple groups of data:

"Greatest Hits" SNPs: This file includes specific SNPs that have been identified from the literature as being associated with phenotype. The file contains SNPs from multiple genes.

Full Genes: These files include one data file for each gene. All SNPs within the gene itself and within 5,000 base pairs (5kb) on either side of the gene are included in the data file. All genes that had a SNP included in the "Greatest Hits" file also have a full gene file. The data files are referred to by their gene name.

Phenotype-specific files

Alzheimer SNPs: This file includes the SNPs that were identified as being associated with late-onset Alzheimer's disease in a recent genome-wide association study and meta-analysis conducted of 74,046 individuals (Lambert, et al. 2013. Nature Genetics 45: 1452-1458). The file contains SNPs from multiple genes.

APOE variants: This file includes information on the two SNPs that comprise the ε2, ε3, and ε4 isoforms of the ApoE protein, as well as the best guess genotype for the isoform itself.

CANDIDATE GENE / SNP FILES

Within each data group, there are three file types:

Documentation

Data Files

The value provided for each SNP is its dosage. The dosage for a person is the number of coded alleles that the person has (ranging from 0 to 2). The coded allele receives a value of "1" and the non-coded allele receives a value of "0".

Annotation Files

Column Name	Description	Type of Field
SNP	rs number of the SNP	Text
chr	chromosome that the SNP is on	Numeric
position	base-pair location of the SNP on the chromosome	Numeric
coded_allele	allele that is coded as a "1" in the dosage files	A, C, T, or G
non_coded_allele	allele that is coded as a "0" in the dosage files	A, C, T, or G
exp_freq_coded_allele	frequency of the coded allele in the full HRS sample	Percentage (ranging from 0 to 1)
info	measure of the observed statistical information associated with the allele frequency estimate (a measure of SNP imputation quality)	Numeric (ranging from 0 to 1)
certainty	average certainty (posterior probability) of best-guess genotypes (a measure of SNP imputation quality)	Percentage (ranging from 0 to 1)

CANDIDATE GENE / SNP FILES

Recommendations for the next phenotype for a gene / SNP set?

- Stress?
- Cardiovascular diseases?

THANK YOU!

- Visit our website: hrsonline.umich.edu

HRS

ADDITIONAL INFO – GENOTYPING QC AND IMPUTATION

- SHAPEIT2 was used to pre-phase haplotypes, and IMPUTE2 was used for imputation to the 1000 Genomes Project Version 1 integrated variant set.
- The following quality control filters were applied to the genotype data prior to imputation: SNP Hardy-Weinberg Equilibrium (HWE) p-value: $p<10^{-4}$, missing call rate >= 2%, minor allele frequency (MAF)<0.01.
- No filters were applied to the data following imputation.
- Position information for genes was obtained using the ENCODE/GENCODE Complete Version 17 genome build from the UCSC Genome Browser (genome.UCSC.edu).

HRS