Introduction to Add Health GWAS Data
Part I

Christy Avery
Department of Epidemiology
University of North Carolina at Chapel Hill
Outline

• Introduction to genome-wide association studies (GWAS)
• Research enabled by GWAS
• Obtaining Add Health data
• Further considerations
Genetics: Difficult to Escape
Genetics: Difficult to Escape

**Genetic Research Definition**: Research into the cause, transmission, amelioration, elimination, or enhancement of inherited disorders and traits.
"Inherited Disorders" Encompasses a Broad Spectrum of Diseases and Traits

IMMEDIATE COMMUNICATION

Genome-wide association study of alcohol dependence: significant findings in African- and European-Americans including novel risk loci

LETTER

Genome-wide association study identifies 74 loci associated with educational attainment

Defining the role of common variation in the genomic and biological architecture of adult human height

Using genome-wide data from 289,285 individuals, we identified 687 variants at genome-wide significance that together explained one-sixth of the heritability for adult height. By testing different numbers of variants in independent studies, we show that the most strongly associated -2,000, -5,700 and -9,500 variants explained 20%, 24% and 29% of phenotypic variance. Furthermore, all common variants together captured 60% of heritability. The AFRI variants clustered in 425 loci were enriched for genes, pathways and processes known to be involved in growth and development, particularly genes and pathways not highlighted in earlier efforts, such as skeletal growth factors, WNT/β-catenin and chondrocyte-related genes. We identified several genes and pathways not previously connected with human skeletal growth, including WNT/beta-catenin and binding motifs, axolotl, and the insulin-like growth factor family. Our results indicate the genomic architecture for human height is also mediated by a large number (thousands) of causal variants.

Genome-Wide Association Study of Proneness to Anger
Definition: Genome-Wide Association Study (GWAS)

- **One** of many contemporary tools to evaluate the genetic basis of disease/phenotypes
- Study that surveys *most* of the genome for genetic causal variants.
- Capitalizes on the strengths of association studies without having to guess the identity of candidate genes.
- Enables testing of multiple, genome-wide (~40 million) variants without any prior hypothesis (other than the trait is heritable)
- GWAS genetic metric: the SNP

Single Nucleotide Polymorphisms (SNPs)

- Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered.
- Millions of SNPs in the genome!
Genome-wide association study identifies 74 loci associated with educational attainment

Our meta-analysis identified 74 genome-wide significant loci for educational attainment (EDT) in the genome-wide association study (GWAS) for educational attainment (EDT) in a recent study (1). These loci were identified using a large-scale meta-analysis of educational attainment across 20 studies. Each locus was associated with educational attainment with a nominal p-value of less than 5 x 10^-8. The loci were located on chromosomes 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, and 20. The loci were found to be associated with educational attainment after adjustment for age, sex, and family history. The loci were also found to be associated with other educational outcomes such as IQ and academic performance. The results of our meta-analysis are consistent with previous findings and support the genetic basis of educational attainment.

Figure 1: Manhattan plot for educational attainment (EDT) association study. The x-axis represents the genomic position of each locus, and the y-axis represents the negative log10 of the p-value. Each point represents a single SNP. The loci are colored by the significance level, with red indicating genome-wide significance (p < 5 x 10^-8). The red arrow indicates the 74 genome-wide significant loci identified in our meta-analysis.
**Figure 1 | Manhattan plot for EduYears associations ($n = 293,723$).**

The $x$ axis is chromosomal position, and the $y$ axis is the significance on a $-\log_{10}$ scale (two-tailed test). The black dashed line shows the genome-wide significance level ($5 \times 10^{-8}$). The red crosses are the 74 approximately independent genome-wide significant associations (lead SNPs). The black dots labelled with rs numbers are the three SNPs identified in ref. 1.
Published GWAS through 01/2016
SNP Data Enable a Wide Range of Investigations in Addition to Genome-Wide Scans
E.g, Limit Inference to Specific Genes
E.g., Polygenic Scores (PGS)

• One-variable summary score constructed from SNPs previously associated with phenotype/disease of interest (i.e. via large GWAS)

• AKA genetic risk score
  • Estimation strategy:
    • Locate dataset with large-scale genotyping and measure of phenotype of interest (e.g. blood pressure in Add Health)
    • Identify published GWAS, including associated lead SNPs and effect estimates
    • Predict participant-specific phenotype (e.g. blood pressure) using participant-specific genotypes and published lead SNP effect estimates
Estimation of a Cardiovascular PGS

Equation 1

Cardiovascular genetic risk score = \sum_{i=1}^{n} \frac{1}{\text{odds Ratio}_{SNP_i}} (SNP_i \text{ dosage})

where \( i \) is the index of SNPs included in Appendix A, Table 2.

Appendix A, Table 2: Lead SNPs and GIs for CHD used to calculate the cardiovascular genetic risk score (eGRS)\(^{16}\)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Lead SNP</th>
<th>Odds Ratio for coronary heart disease</th>
<th>Risk allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>1p31.3 (SORT1)</td>
<td>rs464776</td>
<td>1.19</td>
<td>T</td>
</tr>
<tr>
<td>1p32.3 (PPAP2B)</td>
<td>rs17114036</td>
<td>1.17</td>
<td>A</td>
</tr>
<tr>
<td>1q23.3 (PCOS9)</td>
<td>rs11206510</td>
<td>1.15</td>
<td>T</td>
</tr>
<tr>
<td>1q42 (MIA3)</td>
<td>rs17465637</td>
<td>1.14</td>
<td>C</td>
</tr>
<tr>
<td>2q35.1 (VDR)</td>
<td>rs8725887</td>
<td>1.17</td>
<td>C</td>
</tr>
<tr>
<td>6p21.31 (ANK5I)</td>
<td>rs17605940</td>
<td>1.07</td>
<td>G</td>
</tr>
<tr>
<td>6p21.31 (KCAM4)</td>
<td>rs0049879</td>
<td>1.12</td>
<td>G</td>
</tr>
<tr>
<td>6q23.2 (TCF2)</td>
<td>rs12190287</td>
<td>1.06</td>
<td>C</td>
</tr>
<tr>
<td>8q25.3 (F4A)</td>
<td>rs3798220</td>
<td>1.47</td>
<td>C</td>
</tr>
<tr>
<td>8q25.3 (F4A)</td>
<td>rs10455872</td>
<td>1.70</td>
<td>G</td>
</tr>
<tr>
<td>7q32.3 (C3H2C)</td>
<td>rs1556924</td>
<td>1.09</td>
<td>C</td>
</tr>
<tr>
<td>9q34.2 (CDK5R2)</td>
<td>rs4977574</td>
<td>1.19</td>
<td>G</td>
</tr>
<tr>
<td>9q34.2 (ABG)</td>
<td>rs4111489</td>
<td>1.10</td>
<td>T</td>
</tr>
<tr>
<td>10p11.21 (CCL2)</td>
<td>rs2746048</td>
<td>1.17</td>
<td>C</td>
</tr>
<tr>
<td>10q24.22 (CYP7A1)</td>
<td>rs12413409</td>
<td>1.12</td>
<td>G</td>
</tr>
<tr>
<td>11q23.3 (APOE)</td>
<td>rs964184</td>
<td>1.15</td>
<td>G</td>
</tr>
<tr>
<td>12q2.4 (HMN1A)</td>
<td>rs2259516</td>
<td>1.08</td>
<td>T</td>
</tr>
<tr>
<td>12q24.11 (SNX8)</td>
<td>rs8184504</td>
<td>1.18</td>
<td>T</td>
</tr>
<tr>
<td>12q5.4 (CDH8)</td>
<td>rs4772144</td>
<td>1.07</td>
<td>G</td>
</tr>
<tr>
<td>14q32.2 (HMN1L)</td>
<td>rs2999611</td>
<td>1.07</td>
<td>C</td>
</tr>
<tr>
<td>15q25.1 (ADAMTS7)</td>
<td>rs12836070</td>
<td>1.06</td>
<td>T</td>
</tr>
<tr>
<td>17q11.2 (R45D1)</td>
<td>rs12096587</td>
<td>1.07</td>
<td>G</td>
</tr>
<tr>
<td>17q11.3 (SMAD)</td>
<td>rs215172</td>
<td>1.07</td>
<td>C</td>
</tr>
<tr>
<td>17q21.31 (PDE2C)</td>
<td>rs465222</td>
<td>1.06</td>
<td>T</td>
</tr>
<tr>
<td>19q13.1 (LOXL)</td>
<td>rs1212608</td>
<td>1.15</td>
<td>G</td>
</tr>
<tr>
<td>21q22.11 (KCNV2)</td>
<td>rs98982601</td>
<td>1.10</td>
<td>T</td>
</tr>
</tbody>
</table>
PGS: Application to Kidney Disease

- One-variable summary score constructed from SNPs associated with phenotype/disease of interest
Polygenic risk for coronary artery disease is associated with cognitive ability in older adults

Saskia P. Hagenaaars,1,2,3 Sarah E. Harris,1,4 Toni-Kim Clarke,3 Lynsey Hall,3 Michelle Luciano,1,2 Ana Maria Fernandez-Pujals,3 Gail Davies,1,2 Caroline Hayward,4 Generation Scotland,4 John M. Starr,1,5 David J. Porteous,1,4 Andrew M. McIntosh1,3 and Ian J. Deary1,2,6

1Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK
2Department of Psychology, University of Edinburgh, Edinburgh, UK
3Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK
4Institute for Genetics and Molecular Medicine, Western General Hospital, University of Edinburgh, Edinburgh, UK and 5Geriatric Medicine Unit, University of Edinburgh, Royal Infirmary of Edinburgh, Edinburgh, UK

*Corresponding author. Centre for Cognitive Ageing and Cognitive Epidemiology, Department of Psychology, University of
Edinburgh, 7 George Square, Edinburgh EH8 9JU, UK. Email: i.deary@ed.ac.uk

Abstract

Background: Coronary artery disease (CAD) is associated with cognitive decrements and risk of later dementia, but it is not known if shared genetic factors underlie this association. We tested whether polygenic risk for CAD was associated with cognitive ability in community-dwelling cohorts of middle-aged and older adults.

Methods: Individuals from Generation Scotland: Scottish Family Health Study (GS:SFHS, N = 9866) and from the Lothian Birth Cohorts of 1921 (LBC1921, N = 617) and 1936 (LBC1936, N = 7055) provided cognitive data and genome-wide genotype data. Polygenic risk profile scores for CAD were calculated for all of the cohorts using the largest available genome-wide association studies (GWAS) data set, the CARDIoGRAM consortium (22,233 cases and 84,762 controls). Polygenic risk profile scores for CAD were then tested for their association with cognitive abilities in the presence and absence of manifest cardiovascular disease.

Results: A meta-analysis of all three cohorts showed a negative association between CAD polygenic risk and fluid cognitive ability ($\beta = 0.022$, $P = 0.016$), verbal intelligence ($\beta = 0.024$, $P = 0.011$) and memory ($\beta = 0.021$, $P = 0.028$).

Conclusions: Increased polygenic risk for CAD is associated with lower cognitive ability in older adults. Common genetic variants may underlie some of the association between age-related cognitive decrements and the risk for CAD.
E.g., Gene-Environment Interaction

Genetic variants in ABCB1 and CYP2C19 and cardiovascular outcomes after treatment with clopidogrel and prasugrel in the TRITON-TIMI 38 trial: a pharmacogenetic analysis

Jessica L. Muga*, Sandra L. Chua*, Stephanie D. Wicks, Lei Shen, Joseph W. Williams, Deborah Simon, Elliott M. Antman, Eugene Braunwald, Marc S. Schoenfeld

Summary
Background: Clopidogrel and prasugrel are subject to efflux via P-glycoprotein encoded by ABCB1, also known as MDR1. ABCB1 polymorphisms, particularly 3435C→T, may affect drug transport and efficacy. We aimed to assess the effect of this polymorphism by itself and in combination with CYP2C19 polymorphisms on cardiovascular outcomes in patients treated with clopidogrel or prasugrel in TRITON-TIMI 38. We also assessed the effect of genotype on the pharmacodynamic and pharmacokinetic properties of these drugs in healthy individuals.

Methods: We genotyped ABCB1 and CYP2C19 in 4982 patients with acute coronary syndromes undergoing percutaneous intervention who were treated with clopidogrel (n=1473) or prasugrel (n=1469) in the TRITON-TIMI 38 trial. We evaluated the association between ABCB1 3435C→T and rates of the primary efficacy endpoint (cardiovascular death, myocardial infarction, or stroke) until 15 months. We then assessed the combined effect of ABCB1 3435C→T genotype and reduced-function alleles of CYP2C19. 321 healthy individuals were also genotyped, and we tested the association of genetic variants with reduction in maximum platelet aggregation and plasma concentrations of active drug metabolites.

Findings: In patients treated with clopidogrel, ABCB1 3435C→T genotype was significantly associated with the risk of cardiovascular death, myocardial infarction, or stroke (p=0.004). TT homozygotes had a 72% increased risk of the primary endpoint compared with C/TCC individuals (Kaplan-Meier event rates 12.9% [95% CI 12.4-13.4] vs 7.8% [90 of 1057 participants]; HR 1.72, 95% CI 1.22-2.44, p=0.002). ABCB1 3435C→T and CYP2C19 genotypes were significant, independent predictors of the primary endpoint, and 581 (47%) of the 1454 genotyped patients taking clopidogrel who were either CYP2C19 reduced-function allele carriers, ABCB1 3435T homozygotes, or both were at increased risk of the primary endpoint (HR 1.97, 95% CI 1.48-2.62, p<0.002). In healthy participants, 3435T homozygotes had an absolute reduction in maximum platelet aggregation with clopidogrel that was 7.3 percentage points less than for C/TCC individuals (p=0.017). ABCB1 genotypes were not significantly associated with clinical or pharmacological outcomes in patients with an acute coronary syndrome or healthy individuals treated with prasugrel, respectively.

Interpretation: Individuals with the ABCB1 3435T TT genotype have reduced platelet inhibition and are at increased risk of recurrent ischemic events during clopidogrel treatment. In patients with acute coronary syndromes who have undergone percutaneous intervention, where both ABCB1 and CYP2C19 are taken into account, nearly half of the population carries a genotype associated with increased risk of major adverse cardiovascular events while on standard doses of clopidogrel.

Funding: Daiichi Sankyo Company Ltd and Eli Lilly and Company.
E.g., Gene-Environment Interaction

**Figure 1:** ABCB1 3435C→T and cardiovascular outcomes in patients treated with clopidogrel
Cumulative risk of cardiovascular death, myocardial infarction (MI), or stroke for each genotype, with a p value across genotype.
How Can You Obtain Add Health GWAS Data?
dbGaP: Add Health Genotype Warehouse

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.
Genome-wide Association Study of Adiposity in Samoans

dbGaP Study Accession: phs000914.v1.p1

The research goal of this study is to identify genetic variation that increases susceptibility to obesity and cardiometabolic phenotypes among adult Samoans using genome-wide association (GWAS) methods. DNA from peripheral blood and phenotypic information were collected from 3,119 adult Samoans, 23 to 70 years of age. The participants reside throughout the independent nation of Samoa, which is experiencing economic development and the nutrition transition. Genotyping was performed with the Affymetrix Genome-Wide Human SNP 6.0 Array using a panel of approximately 900,000 SNPs. Anthropometric, fasting blood biomarkers, and detailed dietary, physical activity, health and socio-demographic variables were collected. We are replicating the GWAS findings in an independent sample of 2,500 Samoans from earlier studies. After replication of genomic regions and informative SNPs in these regions, we will determine sequences of the important genes, and determine the specific genetic variants in the sequenced genes that are associated with adiposity and related cardiometabolic conditions. We will also identify gene by environment interactions, focusing on dietary intake patterns and nutrients.

Study Types: Cross-Sectional, Population

Number of study subjects that have individual level data available through Authorized Access: 3501
Subjects with Phenotypes: 3501  Subjects with Genotypes: 3189

Important Links and Information

- Request access via Authorized Access
  - Instructions for requestors
  - Data Use Certification (DUC) Agreement
  - Talking Glossary of Genetic Terms
Genome-wide Association Study of Adiposity in Samoans

dbGaP Study Accession: phs000914.v1.p1

Study Description

The research goal of this study is to identify genetic variation that increases susceptibility to obesity and cardiometabolic phenotypes among adult Samoans using genome-wide association (GWAS) methods. DNA from peripheral blood and phenotypic information were collected from 3,119 adult Samoans, 23 to 70 years of age. The participants reside throughout the independent nation of Samoa, which is experiencing economic development and the nutrition transition. Genotyping was performed with the Affymetrix Genome-Wide Human SNP 6.0 Array using a panel of approximately 900,000 SNPs. Anthropometric, fasting blood biomarkers and detailed dietary, physical activity, health and socio-demographic variables were collected. We are replicating the GWAS findings in an independent sample of 2,500 Samoans from earlier studies. After replication of genomic regions and informative SNPs in these regions, we will determine sequences of the important genes, and determine the specific genetic variants in the sequenced genes that are associated with adiposity and related cardiometabolic conditions. We will also identify gene by environment interactions, focusing on dietary intake patterns and nutrients.

- Study Types: Cross-Sectional, Population
- Number of study subjects that have individual level data available through Authorized Access: 3501
  
  Subjects with Phenotypes: 3501
  Subjects with Genotypes: 3189

Important Links and Information

- Request access via Authorized Access
  - Instructions for requestors
  - Data Use Certification (DUC) Agreement
  - Talking Glossary of Genetic Terms
Add Health Phenotype Data Are Available Through the Add Health Study (UNC)
Analysis Best Practices/Hints

- Do not discount the large number of existing resources!
- Team science/consortia
- Statistical power
- Race/ethnicity heterogeneity and admixture
- Intergenic regions
- Family structure/clustering
- Analytic pipeline
Summary Results from Many Large Consortia Are Available Online

Data available for Download

We are releasing the summary data from our genome-wide meta analyses of glycaemic traits, in order to enable other researchers to examine particular variants or loci of their interest for association with these traits. The files include p-values and direction of effect at over 2 million directly genotyped or imputed single nucleotide polymorphisms (SNPs), as well as frequency information from the HapMap project (release 27).

Acknowledging the data

When using data from the downloadable meta-analyses results please acknowledge the source of the data as follows: "Data on glycaemic traits have been contributed by MAGIC investigators and have been downloaded from www.magicinvestigators.org".

In addition to the above acknowledgement, please cite the relevant paper.

Downloading the data

The data can be downloaded from the magic directory on the Sanger FTP site:
Summary Results from Many Large Consortia Are Available Online
Summary Results from Many Large Consortia Are Available Online

LD Hub is a centralised database of summary-level GWAS results and a web interface for LD score regression. Currently v1.0.1

- 1.4 Billion SNP-Phenotype associations
- 1.5 million Number of individuals
- 36 GWAS consortia
- 177 GWAS studies
Summary Results from Many Large Consortia Are Available Online

For large-scale or genome-wide genetic studies, we feel that widespread availability of the complete set of results is highly desirable. Authors of manuscripts describing new genome-wide association or similar data must indicate in their cover letter whether at least a minimal set of summary results (p value and direction of effect) will be made freely available for all variants, either as supplementary material, by being publicly posted, or by being deposited in a database that is accessible to researchers with minimal restrictions on access. Making these results available will not necessarily be required in all cases for acceptance of a manuscript for publication, but the availability of results after publication will be considered in decisions regarding publication. Accordingly, authors are strongly encouraged to make available complete lists of summary statistics for large scale genetic studies.
Team Science/Consortia

• Joining a consortium is often the first step in **GWAS**
Team Science/Consortia

• Joining a consortium is often the first step in GWAS

But, wouldn’t it be simpler, or at least faster, to just go it alone?
Statistical Power: Gene-Environment GWAS

N=80,000
Statistical Power: Gene-Environment GWAS

N=80,000
Statistical Power: Gene-Environment GWAS

N=80,000

![Graph showing statistical power for Gene-Environment GWAS with N=80,000]
Statistical Power: Gene-Environment GWAS

Why is statistical power so challenging?

N=80,000
Statistical Power: Gene-Environment GWAS

We need to correct for 1,000,000 statistical tests when interrogating genome!

\[ \alpha = \frac{0.05}{1M} \text{ or } 5 \times 10^{-8} \]

Correction for only 1M tests given correlation in human genome
Race/Ethnicity Heterogeneity

Genomics for the world

Medical genomics has focused almost entirely on those of European descent. Other ethnic groups must be studied to ensure that more people benefit, say Carlos D. Bustamante, Esteban González Burchard and Francisco M. De La Vega.

In the past decade, researchers have dramatically improved our understanding of the genetic basis of complex chronic diseases, such as Alzheimer's disease and type 2 diabetes, through more than 12,000 genome-wide association studies (GWAS). These scan the genomes of thousands of people for known genetic variants, to find out which are associated with a particular condition.

Yet the findings from such studies are likely to have less relevance than was previously thought for the world's population as a whole. Ninety-six per cent of subjects included in the GWAS conducted so far are people of European descent (see "Sampling bias"). And a recent Nature survey suggests that this bias is likely to persist in the upcoming efforts to sequence people's entire genomes.

Geneticists worldwide must investigate a much broader ensemble of populations, including racial and ethnic minorities. If we do not, an biased picture will emerge of which variants are important, and genomics medicine will largely benefit a privileged few.
Race/Ethnicity Heterogeneity

Why are genomic studies in non-European populations necessary?

Medical genomics has focused almost entirely on those of European descent. Other ethnic groups must be studied to ensure that more people benefit, say Carlos D. Bustamante, Esteban González Burchard and Francisco M. De La Vega.

In the past decade, researchers have dramatically improved our understanding of the genetic basis of complex chronic diseases, such as Alzheimer's disease and type 2 diabetes, through more than 1000 genome-wide association studies (GWAS). These scan the genomes of thousands of people for known genetic variants, to find out which are associated with a particular condition.

Yet the findings from such studies are likely to have less relevance than we previously thought for the world’s population as a whole. Ninety-six percent of subjects included in the GWAS conducted so far are people of European descent (see “Sampling bias”). And a recent Nature survey suggests that this bias is likely to persist in the upcoming efforts to sequence people’s entire genomes.

Geneticists worldwide must investigate a much broader ensemble of populations, including racial and ethnic minorities. If we do not, the biased picture will emerge of which variants are important, and genomic medicine will largely benefit a privileged few.
Limited Studies Suggest that Genes Generalize Across Global Populations...

A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry

Genome-wide association studies (GWAS) have identified 36 loci associated with body mass index (BMI), predominantly in populations of European ancestry. We conducted a meta-analysis to examine the association of >3.2 million SNPs with BMI in 39,144 men and women of African ancestry and followed up the most significant associations in an additional 32,268 individuals of African ancestry. We identified one new locus at 5q33 (GALNT10, rs7708584, \( P = 3.4 \times 10^{-11} \)) and another at 7p15 when we included data from the GIANT consortium (MIR146A-NFE2L3, rs10261878, \( P = 1.2 \times 10^{-10} \)). We also found suggestive evidence of an association at a third locus at 6q16 in the African-ancestry sample (KLHL32, rs974417, \( P = 6.9 \times 10^{-6} \)). Thirty-two of the 36 previously established BMI variants showed directionally consistent effect estimates in our GWAS (binomial \( P = 9.7 \times 10^{-2} \)), five of which reached genome-wide significance. These findings provide strong support for shared BMI loci across populations, as well as for the utility of studying ancestrally diverse populations. Of GALNT10, \( P = 8.02 \times 10^{-8} \)), has not been previously associated with BMI at genome-wide significant levels in any population.

We subsequently selected the 1,500 most significantly associated SNPs from stage 1 (\( P < 1.19 \times 10^{-3} \)) and examined associations with BMI in an independent sample of 6,817 men and women of African ancestry from seven additional studies (stage 2) (Online Methods, Supplementary Tables 1–3 and Supplementary Note). Of these 1,500 SNPs, 179 replicated at nominal significance (\( P < 0.05 \)) and had effects that were directionally consistent with those in stage 1 (Supplementary Table 4). A meta-analysis of stages 1 and 2 revealed a second new locus, 6q16 (rs974417, located in an intronic region of KLHL32; stage 2 \( P = 3.5 \times 10^{-3} \), combined stages 1 and 2 \( P = 2.2 \times 10^{-8} \)), and confirmed our finding at rs7708584 on 5q33 near GALNT10 (stage 2 \( P = 9.4 \times 10^{-3} \), combined stages 1 and 2 \( P = 2.2 \times 10^{-10} \)). We further examined the associations of these two variants in a third stage composed of 25,451 individuals of African ancestry from an additional 12 studies. We found support for an association with both variants, although the strength of the association was
Limited Studies Suggest that Genes Generalize Across Global Populations…

Generalization of Associations of Kidney-Related Genetic Loci to American Indians


Summary

Background and objectives CKD disproportionately affects American Indians, who similar to other populations, show genetic susceptibility to kidney outcomes. Recent studies have identified several loci associated with kidney traits, but their relevance in American Indians is unknown.

Design, setting, participants, & measurements This study used data from a large, family-based genetic study of American Indians (the Strong Heart Family Study), which includes 94 multigenerational families enrolled from communities located in Oklahoma, the Dakotas, and Arizona. Individuals were recruited from the Strong Heart Study, a population-based study of cardiovascular disease in American Indians. This study selected 25 single nucleotide polymorphisms in 23 loci identified from recently published kidney-related genome-wide association studies in individuals of European ancestry to evaluate their associations with kidney function (estimated GFR; individuals 18 years or older, up to 322 individuals) and albuminuria (urinary albumin to creatinine ratio; n = 3552) in the Strong Heart Family Study. This study also examined the association of single nucleotide polymorphisms in the APOL1 region with estimated GFR in 1121 Strong Heart Family Study participants. GFR was estimated using the abbreviated Modification of Diet in Renal Disease Equation. Additive genetic models adjusted for age and sex were used.

Results This study identified significant associations of single nucleotide polymorphisms with estimated GFR in or near PRKAG2, SLCA13, UBEC2G, PIP5K1B, and WDR72 (P < 2.1 × 10^-8 to account for multiple testing). Single nucleotide polymorphisms in these loci explained 2.2% of the estimated GFR total variance and 2.9% of its heritability. An intronic variant of BCAS3 was significantly associated with urinary albumin to creatinine ratio. APOL1 single nucleotide polymorphisms were not associated with estimated GFR in a single variant test or haplotype analyses, and the at-risk variants identified in individuals with African ancestry were not detected in DNA sequencing of American Indians.

Conclusion This study extends the genetic associations of loci affecting kidney function to American Indians, a population at high risk of kidney disease, and provides additional support for a potential biologic relevance of these loci across ancestries.
Limited Studies Suggest that Genes Generalize Across Global Populations…

A meta-analysis of 36 loci associated with body mass index in populations of European ancestry failed to show genetic susceptibility to kidney traits, but their relevance in American Indians is unknown.

Design, setting, participants, & measurements This study used data from American Indians (the Strong Heart Family Study), which includes participants from communities located in Oklahoma, the Dakotas, and Arizona. The Strong Heart Study, a population-based study of cardiovascular disease, identified 25 single nucleotide polymorphisms in 23 loci identified from previous association studies in populations of European ancestry estimated GxE interactions of individuals 18 years or older, up to 95% of the estimated (47 of 186 in the Strong Heart Family Study. This study included 25 single nucleotide polymorphisms in the APOE region with estimated G x E interactions. G x E was estimated using the abbreviated models as adjusted for age and sex were used.

Results This study identified significant associations of single nucleotide polymorphisms in PRKAG2, SLC6A1, CP2Q2, PIP5K2B, and WDR7 (P < 0.05). Single nucleotide polymorphisms in these loci explained 2.3% of the heritability. An intronic variant of BCA3 was significantly associated with APOE single nucleotide polymorphisms that were not associated with the age and sex risk variants identified in individual GWASs, whereas the age and sex variants identified in individual GWASs.

Conclusion This study extends the genetic associations of loci affecting kidney disease, and provides additional tests for testing these loci across ancestries.

Abstract For the past five years, genome-wide association studies (GWASs) have identified hundreds of common variants associated with human diseases and traits, including high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C), and triglyceride (TG) levels. Approximately 95 loci associated with these traits have been identified primarily among populations of European ancestry. The Population Architecture using Genomics and Epidemiology (PAGE) study was established in 2008 to characterize GWAS-identified variants in diverse population-based studies. We genotyped 49 GWAS-identified SNPs associated with one or more of these traits in at least two PAGE studies and across six racial/ethnic groups. We performed an age-and sex-stratified analysis testing for SNP associations with fasting HDL-C, LDL-C, and TG levels in self-identified European American, African American (<0.0001), African American (<0.0001), African American (<0.0001), and younger adults, regardless of age or other demographic factors. We found no association with either TG or LDL-C levels in self-identified European American adults, and no association with either TG or LDL-C levels in self-identified European American adults.
Limited Studies Suggest that Genes Generalize Across Global Populations…

Generalization of Associations of Genetic Loci to American Indians

Nora Franceschini, Karin Haack, Laura Aimay, Sandra Laston, Elena Leventis, Jean W. McCue, Barbara V. Howard, Jason G. Uman, **, and **

Summary

Background and objectives CKD disproportionately affects American Indians, who show genetic susceptibility to kidney outcomes. Recent studies have identified loci associated with kidney traits, but their relevance in American Indians is unknown.

Design, participants, & measurements This study used data from 36 loci associated with body mass index in the NHGWAS consortium. In a genome-wide association study of American Indians (the Strong Heart Family Study), which included 32,268 individuals of African American ancestry and 39,144 individuals of full-blood ancestry in 15 populations, we identified 11 loci associated with kidney traits, but their relevance in American Indians is unknown.

Introduction

A meta-analysis of genome-wide association studies (GWAS) for body mass index in populations of European ancestry identified 25 loci associated with kidney traits, but their relevance in American Indians is unknown.

Results

This study used data from the NHGWAS consortium and established BMI-related genetic models for the utility of studying ancestry.

Conclusion

This study extends the genetic associations of loci affecting kidney disease in American Indians and provides additional loci across ancestries.
GENETIC ANALYSIS OF AFRICAN POPULATIONS: HUMAN EVOLUTION AND COMPLEX DISEASE

Sarah A. Tishkoff* and Scott M. Williams†$†$

Africa is one of the most ethnically and genetically diverse regions of the world. It is thought to be the ancestral homeland of all modern humans, and is the homeland of millions of people of the recent African diaspora. Because of the central role of African populations in human history, characterizing their patterns of genetic diversity and linkage disequilibrium is crucial for reconstructing human evolution and for understanding the genetic basis of complex diseases.
The Human Heredity and Health in Africa (H3Africa) Initiative aims to facilitate a contemporary research approach to the study of genomics and environmental determinants of common diseases with the goal of improving the health of African populations. To accomplish this, the H3Africa Initiative aims to contribute to the development of the necessary expertise among African scientists, and to establish networks of African investigators.
Genetic Determinants of Lipid Traits in Diverse Populations from the Population Architecture using Genomics and Epidemiology (PAGE) Study

Logan Dumitrescu,1, Cara L. Carty,1, Kira Taylor,2, Fredrick R. Schumacher,3, Lucia A. Hindorff,4, José L. Ambite5, Garnet Anderson,2, Lyle G. Best,2, Kristin Brown-Gentry,1, Petra Bužkova,6, Christopher S. Carlson7, Barbara Cochran,8, Shelley A. Cole,9, Richard B. Devereux,1, Dave Duggan1,10, Charles B. Eaton6,11, Myriam Fornage,12,13, Nora Franceschini,12, Jeff Haessler,12, Barbara V. Howard12, Karen C. Johnson12, Sandra Laston,12 Laurence N. Kolonel12, Elisa T. Lee12, Jean W. MacCusker12, Teri A. Manolio8, Sarah A. Pendergrass9, Miguel Quiñera12, Ralph V. Sholette12, Lynne R. Willkens12, Christopher A. Haiman1,10, Loïc Le Marchand12,10, Steven Buyske2,12, Charles Kooperberg12, Karl E. North12,13, Dana C. Crawford12,14

1Center for Human Genetics Research, Vanderbilt University, Nashville, Tennessee, United States of America. 2Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America. 3Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina, United States of America. 4Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America. 5Office of Population Genomics, National Human Genome Research Institute, Bethesda, Maryland, United States of America. 6Division of Biostatistics, University of Washington, Seattle, Washington, United States of America. 7Department of Epidemiology, University of Southern California, Los Angeles, California, United States of America. 8Missouri Breaks Industries Research, Timber Lake, South Dakota, United States of America. 9Department of Preventive Medicine, University of Southern California, Los Angeles, California, United States of America. 10Department of Community Health Sciences, University of Houston, Houston, Texas, United States of America. 11Division of Biostatistics, Vanderbilt University, Nashville, Tennessee, United States of America. 12Division of Biostatistics, Vanderbilt University, Nashville, Tennessee, United States of America. 13Division of Biostatistics, Vanderbilt University, Nashville, Tennessee, United States of America. 14Department of Preventive Medicine, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America. 15Division of Biostatistics, University of Washington, Seattle, Washington, United States of America. 16Division of Biostatistics, University of Mississippi Medical Center, Jackson, Mississippi, United States of America. 17Division of Biostatistics, University of Hawaii, Honolulu, Hawaii, United States of America. 18Division of Biostatistics, University of Kentucky, Lexington, Kentucky, United States of America. 19University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, United States of America. 20School of Public Health, University of North Carolina, Chapel Hill, North Carolina, United States of America. 21Center for Cardiovascular Research, Department of Medicine, John A. Burns School of Medicine, University of Hawaii, Honolulu, Hawaii, United States of America. 22Department of Statistics and Biostatistics, Rutgers University, Piscataway, New Jersey, United States of America. 23Department of Cellular and Molecular Medicine, University of California, San Diego, California, United States of America. 24Department of Molecular Pharmacology and Biophysics, Vanderbilt University, Nashville, Tennessee, United States of America.

Abstract
For the past five years, genome-wide association studies (GWAS) have identified hundreds of common variants associated with human diseases and traits, including high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), and triglycerides (TG) levels. Approximately 95 loci associated with lipid levels have been identified primarily among populations of European ancestry. The Population Architecture using Genomics and Epidemiology (PAGE) study was established in 2008 to characterize GWAS-identified variants in diverse population-based studies. We genotyped 40 GWAS-identified SNPs associated with one or more lipid traits in at least two PAGE studies and across six racial/ethnic groups. We performed a meta-analysis testing for SNP levels in self-identified European American (~20,000), African American (~9,000), American Indian (~6,000), Mexican American (2,500), Japanese/East Asian (~500), and Pacific Islander/Native Hawaiian (~150) adults, regardless of lipid-lowering medication use. We replicated 56 of 66 (85%) SNP associations tested in European Americans at p<0.05. Despite sufficient power, we were unable to replicate ABCA1 rs1419268 and rs1883052, CETP rs1884163, and TCDDrs157354 previously associated with HDL-C and MAF rs1005695 previously associated with LDL-C, based on significance (p<0.05) and consistent direction of effect, a majority of replicated genotype-phenotype associations for HDL-C, LDL-C, and TG in European Americans generalized to African Americans, Japanese, Chinese, Mexican Americans, and Native Hawaiians, although significant genotypic effects varied by 50%–80%.

Overall, 16 associations generalized across all three populations. For the associations that did not generalize, differences in effect sizes, allele frequencies, and linkage disequilibrium offer clues to the next generation of association studies for these traits.
Figure 1: *ABCB1* 3435C→T and cardiovascular outcomes in patients treated with clopidogrel
Cumulative risk of cardiovascular death, myocardial infarction (MI), or stroke for each genotype, with a p value across genotype.
The rs1045642 A Allele: Substantial Variation Across Global Populations

Approximately **33%** (i.e. \(0.57^2\)) of the SAS population is homozygous for the causal allele compared to **2.3%** (i.e. \(0.15^2\)) of the AFR population.
Linkage Disequilibrium (LD)

- Non-random assortment of alleles at 2+ SNPs
- Population-specific!
- The closer the SNPs, the stronger the LD since recombination will have occurred at a lower rate
- Two markers are in LD if knowing the allele at one marker allows you to predict the allele at the other marker
  - E.g. in a population where there are AB, Ab, and aB haplotypes at adjacent markers, but no ab haplotypes, if we know an individual has a b allele, we know that s/he also has at least one A allele.
Linkage Disequilibrium (LD): SNPs are Inherited in Blocks
Linkage Disequilibrium (LD)

• Non-random assortment of alleles at 2+ SNPs
• Population-specific!
• The closer the SNPs, the stronger the LD since recombination will have occurred at a lower rate
• Two markers are in LD if knowing the allele at one marker allows you to predict the allele at the other marker
  • E.g. in a population where there are AB, Ab, and aB haplotypes at adjacent markers, but no ab haplotypes, if we know an individual has a b allele, we know that s/he also has at least one A allele.
Linkage Disequilibrium (LD)

- Non-random assortment of alleles at 2+ SNPs
- Population-specific!
- The closer the SNPs, the stronger the LD since recombination will have occurred at a lower rate
- Two markers are in LD if knowing the allele at one marker allows you to predict the allele at the other marker
  - E.g. in a population where there are AB, Ab, and aB haplotypes at adjacent markers, but no ab haplotypes, if we know an individual has a b allele, we know that s/he also has at least one A allele.
Linkage Disequilibrium (LD)

- Non-random assortment of alleles at 2+ SNPs
- Population-specific!
- The closer the SNPs, the stronger the LD since recombination will have occurred at a lower rate
- Two markers are in LD if knowing the allele at one marker allows you to predict the allele at the other marker
  - E.g. in a population where there are AB, Ab, and aB haplotypes at adjacent markers, but no ab haplotypes, if we know an individual has a b allele, we know that s/he also has at least one A allele.
Linkage Disequilibrium (LD)

- Non-random assortment of alleles at 2+ SNPs
- Population-specific!
- The closer the SNPs, the stronger the LD since recombination will have occurred at a lower rate
- Two markers are in LD if knowing the allele at one marker allows you to predict the allele at the other marker
  - E.g. in a population where there are AB, Ab, and aB haplotypes at adjacent markers, but no ab haplotypes, if we know an individual has a b allele, we know that s/he also has at least one A allele.
Linkage Disequilibrium (LD)

• Non-random assortment of alleles at 2+ SNPs
• Population-specific!
• The closer the SNPs, the stronger the LD since recombination will have occurred at a lower rate
• Two markers are in LD if knowing the allele at one marker allows you to predict the allele at the other marker
  • E.g. in a population where there are AB, Ab, and aB haplotypes at adjacent markers, but no ab haplotypes, if we know an individual has a b allele, we know that s/he also has at least one A allele.
Linkage Disequilibrium (LD)

• Non-random assortment of alleles at 2+ SNPs
• Population-specific!
• The closer the SNPs, the stronger the LD since recombination will have occurred at a lower rate
• Two markers are in LD if knowing the allele at one marker allows you to predict the allele at the other marker
  • E.g. in a population where there are AB, Ab, and aB haplotypes at adjacent markers, but no ab haplotypes, if we know an individual has a b allele, we know that s/he also has at least one A allele.
Take-home messages:

1 – Genes generalize, but variation in SNPs exist.
2 – Studies in non-European populations are needed.
   A. Implications for gene-environment?
3 – Genetic analyses should be population-specific.
   A. Analyses also need to address within-population variation (e.g. with ancestral principal components.)
Genes mirror geography within Europe

John Novembre\(^1,2\), Toby Johnson\(^4,5,6\), Katarzyna Bryc\(^7\), Zoltán Kutalik\(^4,6\), Adam R. Boyko\(^7\), Adam Auton\(^7\), Amit Indap\(^7\), Karen S. King\(^8\), Sven Bergmann\(^4,6\), Matthew R. Nelson\(^8\), Matthew Stephens\(^2,3\) & Carlos D. Bustamante\(^7\)

Understanding the genetic structure of human populations is of fundamental interest to medical, forensic and anthropological sciences. Advances in high-throughput genotyping technology have markedly improved our understanding of global patterns of human genetic variation and suggest the potential to use large samples to uncover variation among closely spaced populations\(^1-5\). Here we characterize genetic variation in a sample of 3,000 European individuals genotyped at over half a million variable DNA sites in the human genome. Despite low average levels of genetic differentiation among Europeans, we find a close correspondence between genetic and geographic distances; indeed, a geographical map of Europe arises naturally as an efficient two-dimensional summary of genetic variation in Europeans. The results emphasize that when mapping the genetic basis of a disease phenotype, spurious associations can arise if genetic structure is not properly accounted for. In addition, the results are relevant to the prospects of genetic ancestry testing\(^7\); an individual's DNA can be used to infer their geographic origin with surprising accuracy—often to within a few hundred kilometres.

The resulting figure bears a notable resemblance to a geographic map of Europe (Fig. 1a). Individuals from the same geographic region cluster together and major populations are distinguishable. Geographically adjacent populations typically abut each other, and recognizable geographical features of Europe such as the Iberian peninsula, the Italian peninsula, southeastern Europe, Cyprus and Turkey are apparent. The data reveal structure even among French-, German- and Italian-speaking groups within Switzerland (Fig. 1b), and between Ireland and the United Kingdom (Fig. 1a, IE and GB). Within some countries individuals are strongly differentiated along the principal component (PC) axes, suggesting that in some cases the resolution of the genetic data may exceed that of the available geographic information.

When we quantitatively compare the geographic position of countries with their PC-based genetic positions, we observe few prominent differences between the two (Supplementary Fig. 1), and those that exist can be explained either by small sample sizes (for example, Slovakia (SK)) or by the coarseness of our geographic data (a problem for large countries, for example, Russia (RU)); see
Population Structure within Europe

PMID:18758442
Do Not Ignore Intergenic Regions

PMID:25337070
Coronary Heart Disease GWAS and 9p21

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

A Common Allele on Chromosome 9 Associated with Coronary Heart Disease


A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction

9p21.3: Replicated Locus with Zero Prior Biologic Plausibility

The risk interval narrowed to a block approximately 58 kb wide that did not contain any annotated genes.
9p21.3: Replicated Locus with Zero Prior Biologic Plausibility

**Coronary Heart Disease**

9p21.3 Coronary Artery Disease Risk Variants Disrupt TEAD Transcription Factor–Dependent Transforming Growth Factor β Regulation of p16 Expression in Human Aortic Smooth Muscle Cells

Nai A. M. Almendrashiri, PhD; Darielle Antoine, MSE; Xin Zhou, MSE; Ragnar O. Vilumkudser, MSE; Sean X. Zhang; Kennedy N. Hao; Hsiao-Hsiin Chen, PhD; Alexandre F. R. Stewart, PhD

**Background**—The mechanism whereby the 9p21.3 locus confers risk for coronary artery disease remains incompletely understood. Risk alleles are associated with reduced expression of the cell cycle suppressor genes CDKN2A (p16) and p14) and CDKN2B (p15) and increased vascular smooth muscle cell proliferation. We asked whether risk alleles disrupt transcription factor binding to account for this effect.

**Methods and Results**—A quantitative screen was used to predict which of 59 single nucleotide polymorphisms at the 9p21.3 locus disrupt transcription factor binding sites. Electrophoretic mobility shift and luciferase reporter assays examined the binding and functionality of the predicted regulatory sequences. Primary human aortic smooth muscle cells (HaSMCs) were genotyped for 9p21.3 and HaSMCs homozygous for the risk allele showed reduced p15 and p16 levels and increased proliferation. rs1081666 and rs4977257 disrupted functional TEAD1 (TEAD) domain (TEAD) transcription factor binding sites. TEAD3 and TEAD4 overexpression induced p16 in HaSMCs homozygous for the normal allele, but not for the risk allele. Transforming growth factor β, known to activate p16 and also to interact with TEAD factors, failed to induce p16 or to inhibit proliferation in HaSMCs homozygous for the risk allele. Knockdown of TEAD3 blocked transforming growth factor β–induced p16 mRNA and protein expression, and dual knockdown of TEAD3 and TEAD4 further reduced p16 expression in HaSMCs homozygous for the 9p21.3 risk allele.

**Conclusions**—Here, we identify a novel mechanism whereby sequence changes at the 9p21.3 risk locus disrupt TEAD factor binding and TEAD–dependent transforming growth factor β–induction of p16 in HaSMCs. This mechanism accounts, in part, for the 9p21.3 coronary artery disease risk.

Key Words: atherosclerosis; coronary disease; genes; molecular biology; smooth muscle cells
Family Structure/Clustering

• Add Health GWAS data has a non-negligible number of related participants
  • Failure to address lack of independence between family members leads to anti-conservative \( P \)-values
  • Most “canned” software (e.g. PLINK, ProbAbel) does not address relatedness

• Option 1 (easiest): exclude all but one member of each first-degree relative set (kinship matrix provided on dbGap) and proceed as unrelated.

• Option 2 (more work, more power): model the family structure

• School clustering also requires extension of models to include additional variance components
Analytic Pipeline: Addresses Add Health Data Challenges

• GWAS tools have been published that can accommodate Add Health analysis challenges

• Implementation may be challenging if modest Unix/R/python expertise

• Scalability remains a challenge in GWAS setting.
  • Linear mixed models run locally can be used when examining a limited number of SNPs
Analytic Pipeline: Addresses Add Health Data Challenges

GWAS tools have been published that can accommodate Add Health analysis challenges. Implementation may be challenging if modest Unix/R/python expertise is required. Scalability remains a challenge in GWAS setting.

Linear mixed models run locally can be used when examining a limited number of SNPs.
Analytic Pipeline: Addresses Add Health Data Challenges

- GWAS tools have been published that can accommodate Add Health analysis challenges.
- Implementation may be challenging if modest Unix/R/python expertise.
- Scalability remains a challenge in GWAS setting.
- Linear mixed models run locally can be used when examining a limited number of SNPs.

GWASTools

Tools for Genome Wide Association Studies

Bioconductor version: Release (3.3)

Classes for storing very large GWAS data sets and annotation, and functions for GWAS data cleaning and analysis.

Author: Stephanie M. Gogarten, Cathy Laurie, Tushar Bhangele, Matthew P. Conomos, Cecelia Laurie, Cellin McHugh, Ian Painter, Xiwen Zheng, Jess Shen, Rohit Swarnkar, Adrienne Stilo, Sarah Nelson

Maintainer: Stephanie M. Gogarten <sdmorris at u.washington.edu>, Adrienne Stilo <astilp at u.washington.edu>

Citation (from within R, enter citation("GWASTools")):


Conclusions

• Add Health GWAS data offer a wealth of opportunities for advancing the understanding of human phenotypes and traits
  • Very unique resource: few studies of nationally representative populations beginning in adolescence are available

• Genomics data are challenging at first to use, but numerous resources exist
  • Consider establishing relationships with existing consortia/engaging a genetic epidemiologist etc.

• Genetics of “social science” traits, gene-environment interactions etc. remain largely unexplored