National Heart Lung and Blood Institute Exome Sequencing Project
(NHLBI-ESP)

NHLBI-ESP Study Design

Samples Selected from Large Studies

HeartGO	LungGO	WHISP
N>60,000	N>19,000	N>142,000
e.g. Jackson Heart Study, Framingham	e.g. Lung Health Study, Acute Lung Injury	i.e. Women's Health Initiative

Sequencing Performed at the Broad & University of Washington

Read Mapping and Alignment -BWA

Multi-Sample Variant Calling (glfMultiples –UMAKE)

European Americans (EA) N=4420

~6700 Exomes

African Americans (AA) N=2312

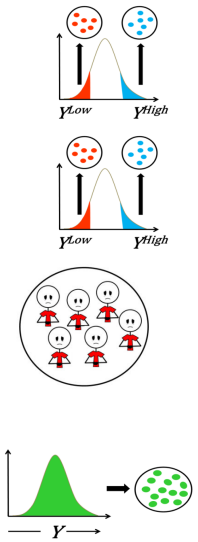
Selection for the 12 Primary Traits

Extreme quantitative trait values
- Low-density lipoprotein (N=657)
- Blood pressure (N=812)

Disease severity
- Asthma (N=190)
- Chronic obstructive pulmonary disease (N=623)

Disease endpoints
- Stroke (N=551)
- Early onset myocardial infarction (N=1007)

Deeply phenotyped individuals
- Randomly selected to be used as controls (N=964)



Extensive Secondary Phenotypic Data

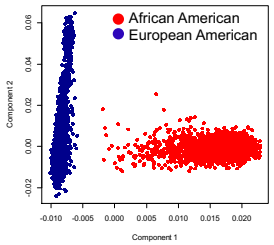
- C-reactive protein (N=3379)
- EKG measurements(N_{EKG-QT}= 3442)
- Fasting blood glucose (N=2470)
- Fibrinogen (N=2915)
- High-density lipoprotein (N=3770)
- Intima-media thickness (N=2079)
- Low-density lipoprotein (N=2685)
- Red blood cell count (N=1103)
- Systolic blood pressure (N=4423)
- Triglycerides (N=3728)
- Uric acid (N=2169)
- Waist-to-hip ratio (N=3853)
- White blood cell count (N=3792)
- von Willebrand factor (N=1587)

- 59 Secondary phenotypes*
  - 48 quantitative traits
  - 11 qualitative traits
- *Some traits are both primary and secondary
  - i.e. asthma, blood pressure, BMI, COPD, LDL, T2D

Data Quality Control & Association Analysis

Very different for rare variant sequence data than for common variants obtained from genotyping arrays

Analysis performed using Variant Association Tools
http://varianttools.sourceforge.net/VAT



Exome Data Quality Control

Variant Site Removal
Support Vector Machine

Variant Call Removal
Read Depth <10X

Variant Site Removal
>500X mean depth, Missing >10% genotypes

Sex Check

Duplicates & Related Sample Check

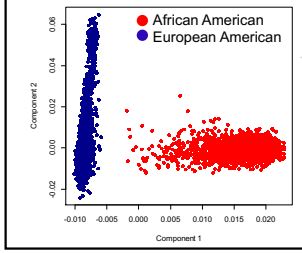
Designate EAs & AAs
Multidimensionality scaling (MDS)

Variant Site Removal
Deviation from HWE

Data Quality Control & Association Analysis

Very different for rare variant sequence data than for common variants obtained from genotyping arrays

Analysis performed using Variant Association Tools
http://varianttools.sourceforge.net/VAT



Exome Data Quality Control

Variant Site Removal
Support Vector Machine

Variant Call Removal
Read Depth <10X

Variant Site Removal
>500X mean depth, Missing >10% genotypes

Sex Check

Duplicates & Related Sample Check

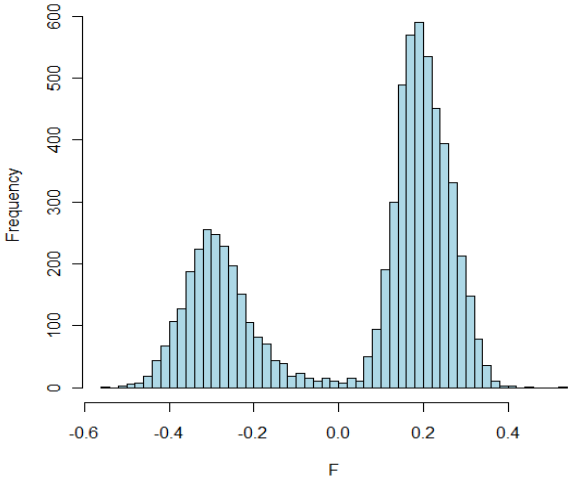
Designate EAs & AAs
Multidimensionality scaling (MDS)

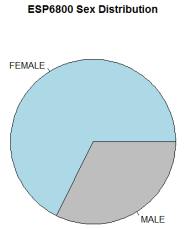
Variant Site Removal
Deviation from HWE

Support Vector Machine (SVM)

- A machine-learning algorithm, to separate likely true positive and false-positive variant sites.
- Uses VCF annotation related to quality of each SNV, including
  - Overall depth
  - Fraction of samples with coverage
  - Fraction of reference bases in heterozygous individuals (allele balance)
  - Inbreeding coefficient
  - In all 16 parameters were used
- Training set
  - False positives
    - SNVs that deviated significantly from expected values in three or more annotation categories
  - True positives
  - SNVs at HapMap polymorphic sites and Omni 2.5 array polymorphic sites in the 1000 Genomes project data
- The SVM classifier was used to identify all likely false positive sites
- Those variant sites which fail the support vector machine (SVM) (Likely false positive variant sites)
  - Are flagged and removed from further analysis

ESP6800 SEX CHECK





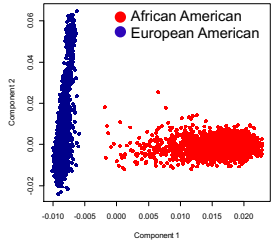
Sex	Report	Exome
Males	2608	2612
Females	4191	4211

*14 individuals inconsistent sex information

Data Quality Control & Association Analysis

Very different for rare variant sequence data than for common variants obtained from genotyping arrays

Analysis performed using Variant Association Tools
http://varianttools.sourceforge.net/VAT



Exome Data Quality Control

Variant Site Removal
Support Vector Machine

Variant Call Removal
Read Depth <10X

Variant Site Removal
>500X mean depth, Missing >10% genotypes

Sex Check

Duplicates & Related Sample Check

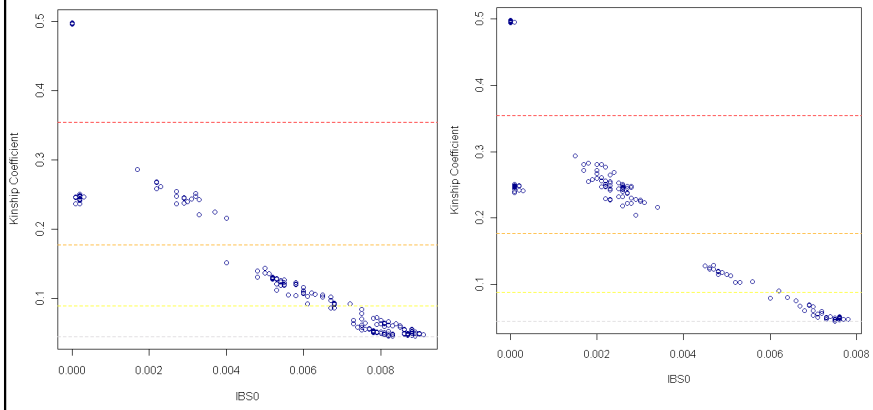
Designate EAs & AAs
Multidimensionality scaling (MDS)

Variant Site Removal
Deviation from HWE

KING Relatedness

African Americans

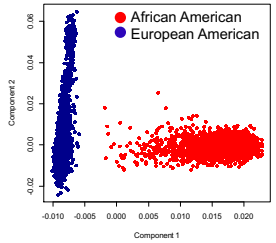
European Americans



Data Quality Control & Association Analysis

Very different for rare variant sequence data than for common variants obtained from genotyping arrays

Analysis performed using Variant Association Tools
http://varianttools.sourceforge.net/VAT



Exome Data Quality Control

Variant Site Removal
Support Vector Machine

Variant Call Removal
Read Depth <10X

Variant Site Removal
>500X mean depth, Missing >10% genotypes

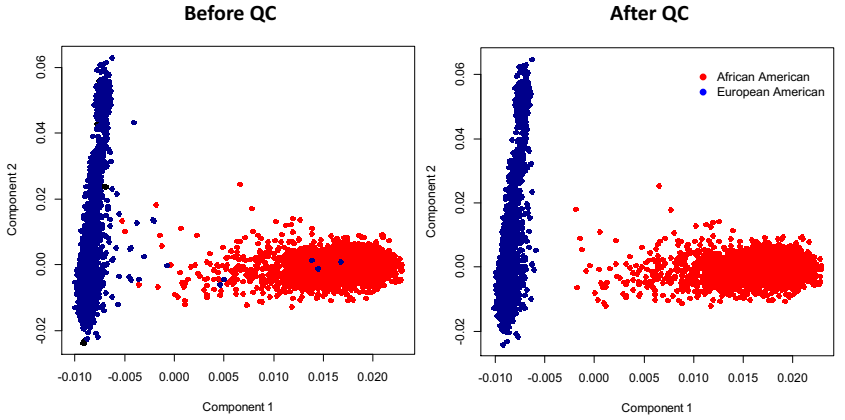
Sex Check

Duplicates & Related Sample Check

Designate EAs & AAs
Multidimensionality scaling (MDS)

Variant Site Removal
Deviation from HWE

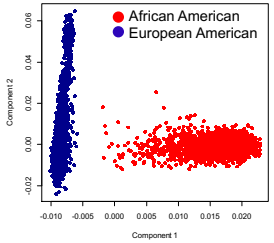
MDS Before and After Removal of Outliers



Data Quality Control & Association Analysis

Very different for rare variant sequence data than for common variants obtained from genotyping arrays

Analysis performed using Variant Association Tools
http://varianttools.sourceforge.net/VAT



Exome Data Quality Control

Variant Site Removal
Support Vector Machine

Variant Call Removal
Read Depth <10X

Variant Site Removal
>500X mean depth, Missing >10% genotypes

Sex Check

Duplicates & Related Sample Check

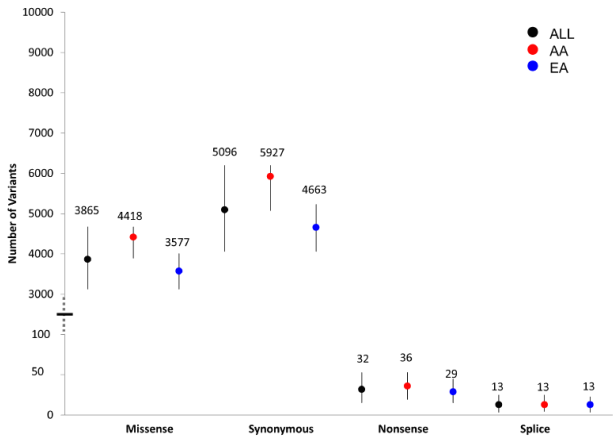
Designate EAs & AAs
Multidimensionality scaling (MDS)

Variant Site Removal
Deviation from HWE

Removal of Additional Variant Sites

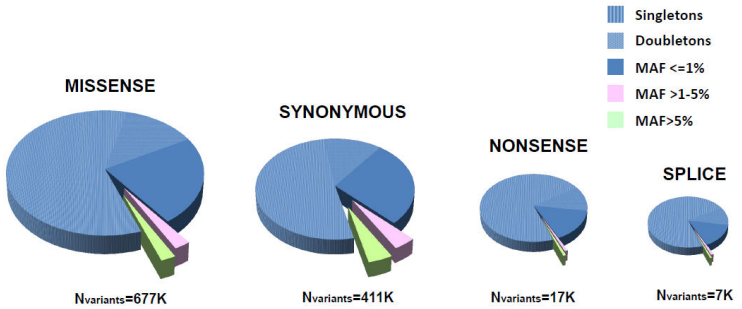
- Variant sites which deviate from HWE
  - Using a p-value $<1x10^{-7}$ criterion
    - Number of variant sites which deviate from HWE expectations:
      - EA: 2332
      - AA: 2663

Average Number of Variant Site Per Individual

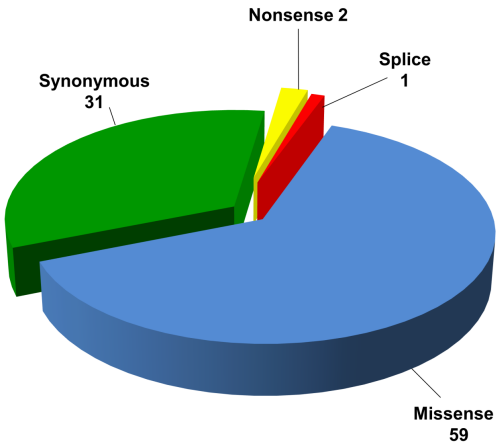


Intersect of 4 capture array targets - 16,206 genes

Most Variants are Rare



Average Number of Unique Variants per Individual



Analysis of Phenotypes and Exome Data

Extreme QTs
Dichotomized

Disease Traits
Case-control

QTs
Analyze QT values

Control for Population Substructure
Included population-specific C1 & C2

Selection of Covariates
Phenotype-specific model selection

Association Analysis

Single Variant Association Analysis
- All variant types included
- e.g. synonymous, missense, etc.

Rare Variant Aggregate Association Analysis
- CMC, SKAT (MAF ≤1%) and VT (MAF<5%)
- Variant types restricted within gene region
- i.e. missense, nonsense, spice site

Analysis of Phenotypes and Exome Data

Extreme QTs
Dichotomized

Disease Traits
Case-control

QTs
Analyze QT values

Control for Population Substructure
Included population-specific C1 & C2

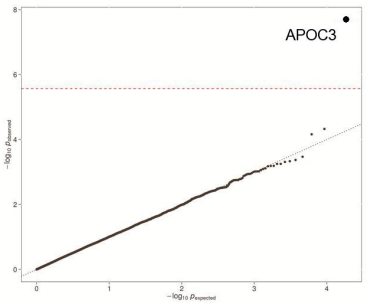
Selection of Covariates
Phenotype-specific model selection

Association Analysis

Single Variant Association Analysis
- All variant types included
- e.g. synonymous, missense, etc.

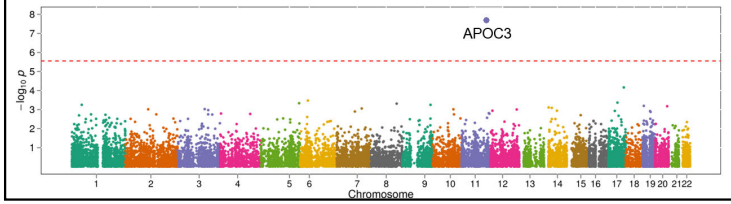
Rare Variant Aggregate Association Analysis
- CMC, SKAT (MAF ≤1%) and VT (MAF<5%)
- Variant types restricted within gene region
- i.e. missense, nonsense, spice site

Burden of Rare Variants *APOC3* Associated with Triglycerides Levels



APOC3 *	Beta	# Variant sites	P-value (CMC)
Meta Analysis			**3.0×10^{-8}**
EA (N=2074)	-0.598	7	3.9×10^{-6}
AA (N=1654)	-0.357	4	1.6×10^{-3}

*Rare variants reduce triglyceride levels
Triglyceride levels natural log transformed
Covariates: age, age^2, BMI & current smoker
3 variant sites overlap between EA & AA



Auer et al.
AJHG 2016

The Exome Chip

NHLBI-ESP the largest contributor of sequence data for the development of the exome chip

- ~240,000 missense, nonsense and splice site variants
- NHLBI-ESP findings are being followed up using the exome chip
- Novel findings are also being pursed
- More than 100,000 exome chips being genotyped and analyzed using samples from the ESP cohorts

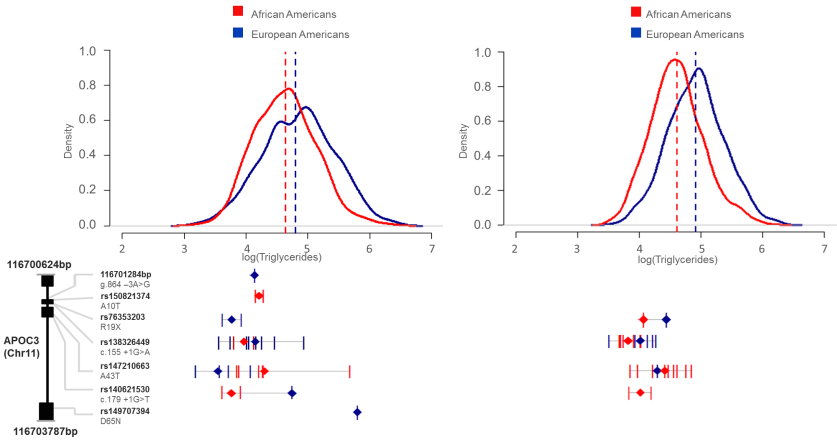
Replication with the Exome Chip
APOC3 Associated with Triglycerides Levels

APOC3 *	Sample Size	# Variant Sites	P-value
Meta Analysis	**8,069**		**1.7×10^{-18}**
Women's Health Initiative (WHI) Exome Chip			
Meta Analysis	4,341		9.4×10^{-12}
European Americans	2,301	3	1.3×10^{-6}
African Americans	2,041	4	1.6×10^{-6}
Exome Sequencing Project			
Meta Analysis	3,728		3.0×10^{-8}
European Americans	2,074	7	3.9×10^{-6}
African Americans	1,654	4	1.6×10^{-3}

*Reduces triglyceride levels
Triglyceride levels natural log transformed
Covariates: age, age^2, sex, BMI & current smoker

Triglyceride Levels for Carriers of APOC3 Rare Variants

Exome Sequencing Project



Women's Health Initiative

For additional information see Auer et al. 2016