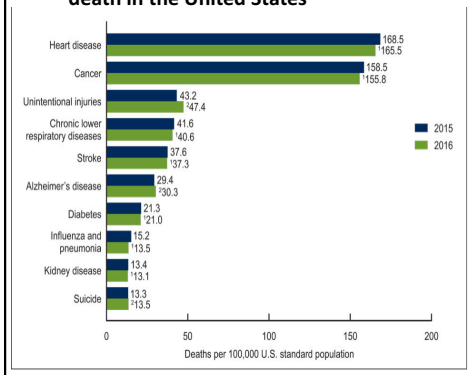
Complex Trait Association Analysis of Rare Variants Obtained from Sequence Data: Population-Based Data

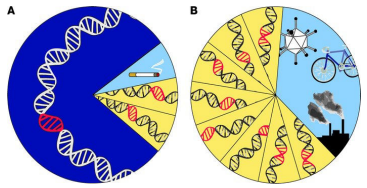
Complex Diseases (Traits)

Top 10 leading causes of death in the United States



D. Kenneth, et al. NCHS Date Brief No. 293, 2017

Genetic and environmental contribution to complex disorders

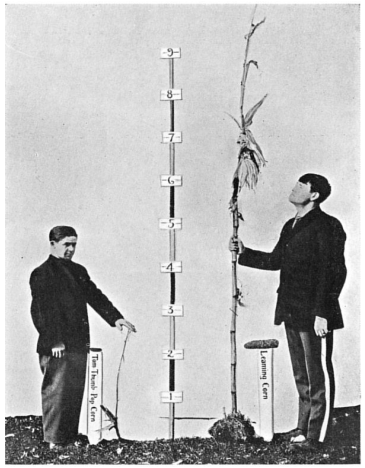


T.A. Manolio, et al. J clin Invest, 2001

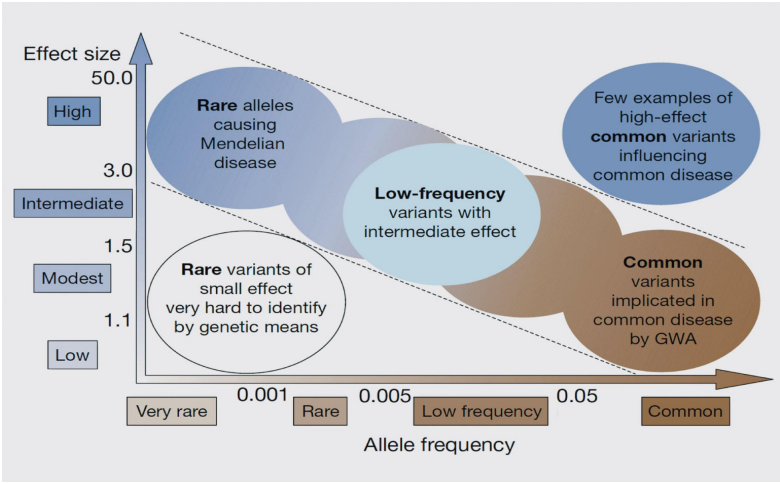
Heritability for Common Traits

Human height heritability is ~80%

- Strongly associated common variation explain 21—29%
- All common variation explains 60% of height heritability



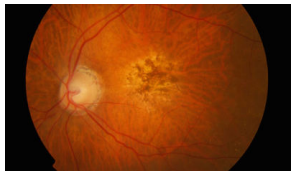
Allelic Architecture



T. A. Manolio et al. Nature, 2009

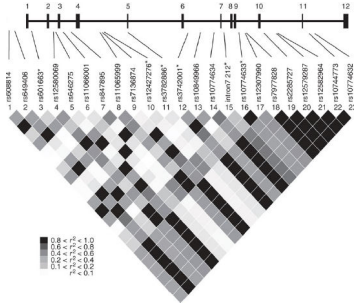
Complex Disease – Common Variant Associations

- Disease susceptibility is conferred by variants which are common within populations
  - Variants are old and widespread
- These variants have modest phenotypic effect
- This model is supported by a large number of replicated examples
  - Age Related Macular Degeneration (Klein et al. 2005)
    - Complement factor H (CFH) gene

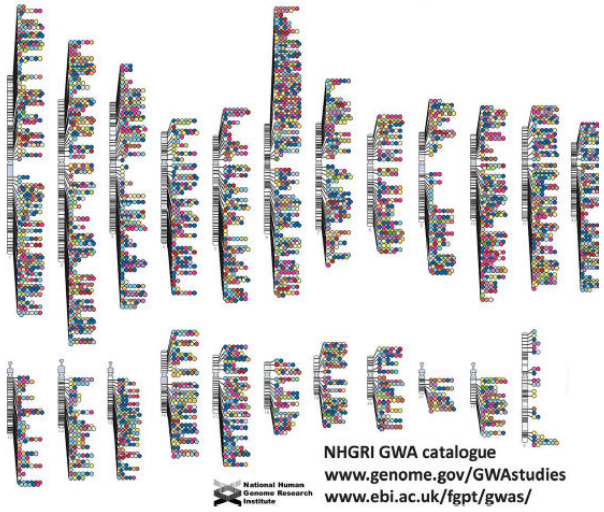


Studying Complex Traits – Common Variant Associations

- Hundreds of thousands of Single nucleotide polymorphism (SNPs) genotyped and analyzed
  - Indirect mapping
    - Markers usually had a minor allele frequency (MAF) > 0.05
    - Usually not pathogenic – tag SNPs
    - In linkage disequilibrium with disease susceptibility variant



Complex Trait – Common Variant Associations



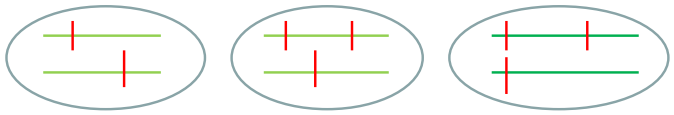
- Although highly successful in identifying thousands of complex trait loci
- Usually pathogenic susceptibility variant(s) not identified

Complex Disease – Rare Variant Associations

- Complex traits are the result of multiple rare variants
  - Although first thought to large effects, there effect sizes are usually small
- Although these variants are rare, e.g. MAF<0.005
  - Collectively they may be quite common
- Direct tests of this hypothesis where first reported >10 years ago
  - Dallas Heart Study
    - Small sample ~1,200 indiviuals
      - Multi-ethnic
      - Used “extreme” sampling
    - Plasma low density lipoprotein levels (Cohen et al. 2004)
      - NPC1L1

Rationale for Rare Variant Aggregate Association Tests

- Testing individual variants with low effect sizes and minor allele frequencies (MAFs)
  - Underpowered to detect associations
- Testing variants in aggregate increases MAFs
  - Improving the power to detect associations



Gene 1 Gene 2 Gene 3

Caveats - Aggregate Rare Variant Association Tests

- Misclassification of variants can reduce power
  - Inclusion of non-causal variants
  - Exclusion of causal variants
- Analysis is limited to
  - Genes
  - Genes within pathways
- Analysis outside of exonic regions is problematic
  - Unlikely a sliding window approach will work
    - Size of window unknown and will differ across the genome
  - A better understanding of functionality outside the coding regions is necessary
    - Predicted functional regions, enhancer regions, transcription factors, DNase I hypersensitivity sites, etc.

A Few Rare Variant Association Tests

- Combined Multivariate Collapsing (CMC)
  - Li and Leal AJHG 2008
- Burden of Rare Variants (BRV)
  - Auer, Wang, Leal Genet Epidemiol 2013
- Weighted Sum Statistic (WSS)
  - Madsen and Browning PloS Genet 2009
- Kernel based adaptive cluster (KBAC)
  - Liu and Leal PloS Genet 2010
- Variable Threshold (VT)
  - Price et al. AJHG 2010

Fixed Effect Tests

- Sequence Kernel Association Test (SKAT)
  - Wu et al. AJHG 2011

Random Effect Test

- SKAT-0
  - Lee et al. AJHG 2012

Optimal test

Types of Aggregate Analyses

- Frequency cut offs used to determine which variants to include in the analysis
  - Rare Variants (e.g. <1% frequency)
  - Rare and low (1-5%) frequency variants
- Maximization approaches
- Tests developed to detection associations when variants effects are bidirectional
  - e.g. protective and detrimental
- Incorporate weights based upon annotation
  - Frequency
    - e.g. gnomAD
  - Functionality
    - CADD c-scores

Methods to Detect Rare Variant Associations Using Variant Frequency Cut-offs

- Combined multivariate & collapsing (CMC)
  - Li & Leal, AJHG 2008
- Collapsing scheme which can be used in the regression framework
  - Can use various criteria to determine which variants to collapse into subgroups
    - Variant frequency
    - Predicted functionality

CMC

- Define covariate Xj for individual j as

$$X_j = \begin{cases} 1 & \text{if rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

- Compute Fisher exact test for 2x2 table

Number of cases for which one or more rare variants are observed e.g. nonsynonymous variants freq. ≤1%

Number of cases without a rare variants

	X=1	X=0
cases		
controls		

Number of controls for which one or more rare variants are observed

Number of controls without a rare variants

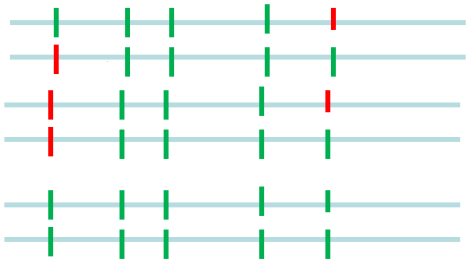
Can also use same coding in a regression framework

CMC

- Example of coding used in regression framework:
  - Binary coding

$$X_j = \begin{cases} 1 & \text{if rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

  - Gene region with 5 variant sites
  -



Individual	Coding
1	1
2	1
3	0

Rare Variant Sites

Green bars: Major allele is observed in the study subject

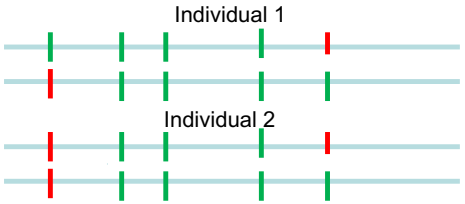
Red bars: Minor allele has been observed

Methods to Detect Rare Variant Associations Using Variant Frequency Cut-offs

- Gene-or Region-based Analysis of Variants of Intermediate and Low frequency (GRANVIL)
  - Aggregate number of rare variants used as regressors in a linear regression model
  - Can be extended to case-control studies
    - Morris & Zeggini 2010 Genet. Epidemiol
  - Test also referred to as MZ

GRANVIL

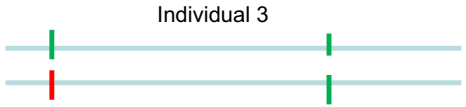
- Example of coding used in regression framework
  - Gene region with 5 variant sites – data available on all sites
  -



Coded 2/5 (0.4)

Coded 2/5 (0.4) Note same coding for heterozygous and homozygous genotypes

- Missing data for three of the five variant sites



Coded 1/2 (0.5)

Burden Rare Variant (BRV) extension (Auer et al. 2013 Genet Epidemiol)

- Individual 1: Coded 2
  Individual 2: Coded 3
  Individual 3: Coded 1

Methods to Detect Rare Variant Associations Weighted Approaches

- Group-wise association test for rare variants using the Weighted Sum Statistic (WSS)
  - Variants are weighted inversely by their frequency in controls (rare variants are up-weighted)
    - Madsen & Browning, PLoS Genet 2009
- Kernel based adaptive cluster (KBAC)
  - Adaptive weighting based on multilocus genotype
    - Liu & Leal, PLoS Genet 2010

Methods to Detect Rare Variant Associations Maximization Approaches

- Variable Threshold (VT) method
  - Uses variable allele frequency thresholds and maximizes the test statistic
  - Also can incorporate weighting based on functional information
    - Price et al. AJHG 2010
- RareCover
  - Maximizes the test statistic over all variants with a region using a greedy heuristic algorithm
    - Bhatia et al. 2010 PLoS Computational Biology

Methods to Detect Associations with Protective & Detrimental Variants within a Region

- C-alpha
  - Detects variants counts in cases and controls that deviate from the expected binomial distribution
    - For qualitative traits only
      - Neale et al. 2011 PLoS Genet
- Sequence Kernel Association Test (SKAT)
  - Variance components score test performed in a regression framework
    - Can also incorporate weighting
  - Wu et al. 2011 AJHG

Optimal Test

- SKAT-O
  - Maximizes power by adaptively using the data to combine an aggregate test and the sequence kernel association tests
    - Lee et al. 2012 AJHG

Significance Level for Rare Variant Association Tests

- For exome data where individual genes are analyzed usually a Bonferroni correction for the number of genes tested is used.
  - There is very little to no linkage disequilibrium between genes
- Often a Bonferroni correction for testing 20,000 genes is often used as the significance level cut-off
  - 2.5×10^{-6}

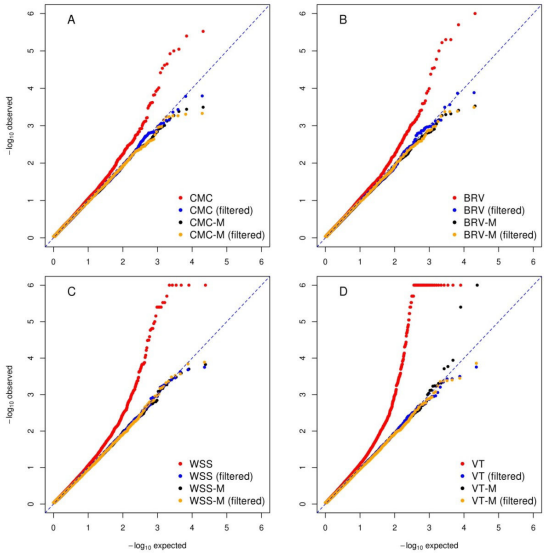
Determine MAF Cut-offs for Aggregate Rare Variant Association Tests

- MAF cut-offs are frequently used to determine which variants to analyze in aggregate rare variant association tests
- MAF from controls should not be used
  - Increases in type I error rates
- Determine variant frequency cut-offs from databases
  - ExAC
    - http://exac.broadinstitute.org/
  - gnomAD
    - http://gnomad.broadinstitute.org/

Problem of Missing Genotypes for Aggregate Rare Variant Association Tests

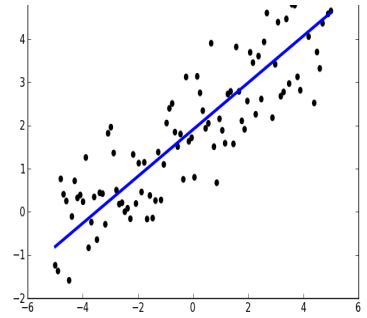
- Same frequency of missing variant calls in cases and controls
  - Decrease in power
- More variant calls missing for either cases or controls
  - Increase in Type I error
  - Decrease in power
- Remove variant sites which are missing genotypes, e.g. >10%
- Impute missing genotypes using observed allele frequencies
  - For the entire sample
    - Not based on case or control status
- Analyze imputed data using dosages

Results



Rare Variant Aggregate Methods

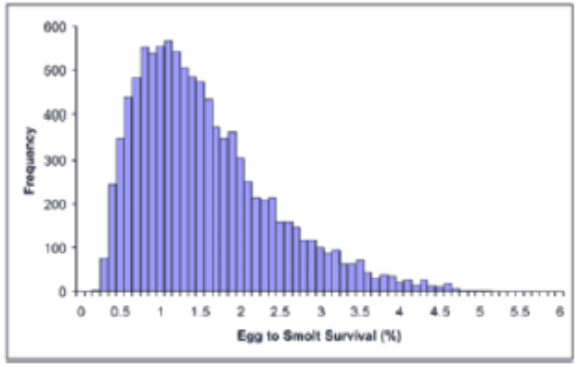
- Ideally should be performed in a regression framework
  - Logistic
  - Linear regression



- Almost all methods have been extended to be implemented within a regression framework

Analyzing Quantitative Variants

- Most rare variant aggregate analysis methods can be performed on quantitative traits
- If phenotype data includes outliers or deviates from normality
  - Can increase type I errors

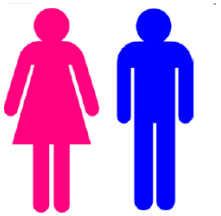


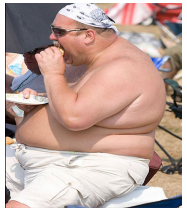
Analyzing Quantitative Variants

- For data that deviates from normality
  - Quantile-quantile normalization
- For data that includes outliers
  - Winsorize
- Don't winsorize and then normalize
- Instead of analyzing quantitative trait values
- Residual can be generated
  - If their our confounders which need to be controlled
    - Residuals are generated were confounders have been adjusted

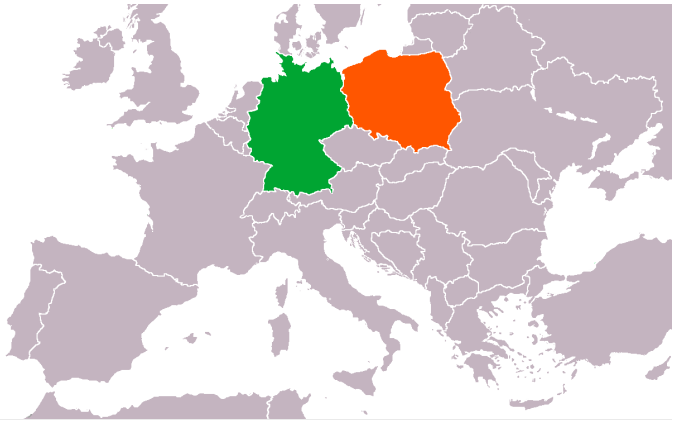
Rare Variant Aggregate Methods

- Can control for covariates in the analysis which are potential confounders
  - Age
  - Sex
  - Body Mass Index (BMI)
  - Smoking pack years





Confounder -Population Substructure and Admixture



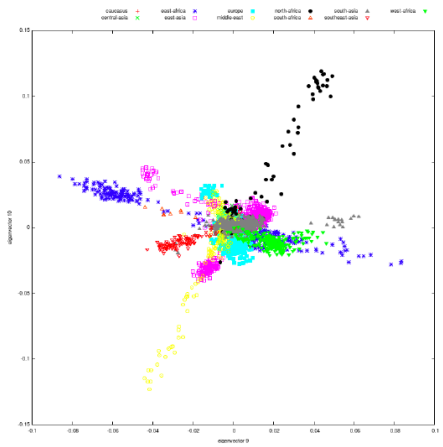
Rare Variant Aggregate Methods

- If proportion of cases and controls sampled from each populations is different
  - Can occur due to
    - Disease frequency is different between populations
    - Sloppy sampling
- Population substructure\admixture can cause detection of differences in variant frequencies within a gene which is due to sampling and not disease status
  - False positive findings can be increased

Example Pima Indians

Rare Variant Aggregate Methods

- Currently PCA or MDS are use to control for population substructure\admixture
  - Controls on the global level
  - May not be sufficient in particular for admixed populations



Rare Variant Aggregate Methods

- Best to obtain components to include in the regression model
  - using variants which are not in LD e.g. r^2<0.5 (prunned)
  - covering a wide range of the allelic frequency spectrum e.g. >0.1%
- Success of PCA\MDS in controlling for population substructure\admixture can be evaluated through lambda and examining Quantile-Quantile (QQ) plots

Linear Mixed Models and Generalized linear Mixed Models

- Linear mixed models and their extension Generalized linear mixed models for binary traits
  - Can offer better control of type I error than linear and logistic regression
    - When observations are not independent
      - Related or cryptically related individuals are included in the sample
      - Population structure
- Models both fixed and random effects