Session 10:
Rare variant association studies

Side note…

R packages and other notes

Some notes

- There are MANY, MANY R packages to conduct genetic analyses. There is very likely a package that does what you want to do.
- Input data is key – make sure the format is correct!
- Spend time reading manuals, google examples, google errors.
- My most important advice is to take programming classes that focuses on programming only (doesn't matter which language). Programming + content learning (e.g., genetics) might distract each other.
- There is no shortcut to learn programming
  - Even though you get sample code for one specific analysis/problem, chances are you won't be able to use the same code next time.

General packages

- HardyWeinberg
  - Contains tools for exploring Hardy-Weinberg equilibrium for bi and multi-allelic genetic marker data.
  - https://cran.r-project.org/web/packages/HardyWeinberg/index.html
- SNPlocs.Hsapiens.dbSNP144.GRCh37
  - SNP locations and alleles for Homo sapiens extracted from NCBI dbSNP
  - http://bioconductor.org/packages/release/data/annotation/html/SNPlocs.Hsapiens.dbSNP144.GRCh37.html
  - https://bioconductor.org/packages/release/data/annotation/html/SNPlocs.Hsapiens.dbSNP151.GRCh38.html

GWAS

- GWAStools
  - Classes for storing very large GWAS data sets and annotation, and functions for GWAS data cleaning and analysis.
  - https://www.bioconductor.org/packages/release/bioc/html/GWASTools.html
- GENESIS
  - Methodology for estimating, inferring, and accounting for population and pedigree structure in genetic analyses. Performs a Principal Components Analysis on genome-wide SNP data for the detection of population structure in a sample that may contain known or cryptic relatedness. Functions are provided to perform mixed model association testing for both quantitative and binary phenotypes.
  - https://bioconductor.org/packages/release/bioc/html/GENESIS.html

Gene-Environment Interactions

- GxEScanR
  - Genome-wide association study (GWAS) and genome-wide by environmental interaction study (GWEIS) scans using imputed genotypes stored in the BinaryDosage format. The phenotype to be analyzed can either be a continuous or binary trait. The GWEIS scan performs multiple tests that can be used in two-step methods.
  - https://github.com/USCbiostats/GxEScanR

Rare variant analyses

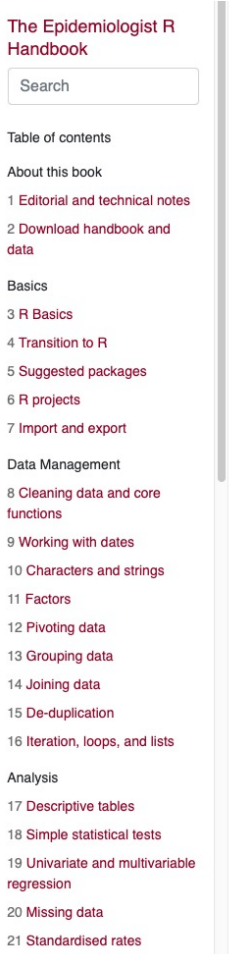
- Rvtests
  - Rare variant test software for next generation sequencing data
  - http://zhanxw.github.io/rvtests/
- SKAT
  - https://cran.r-project.org/web/packages/SKAT/index.html

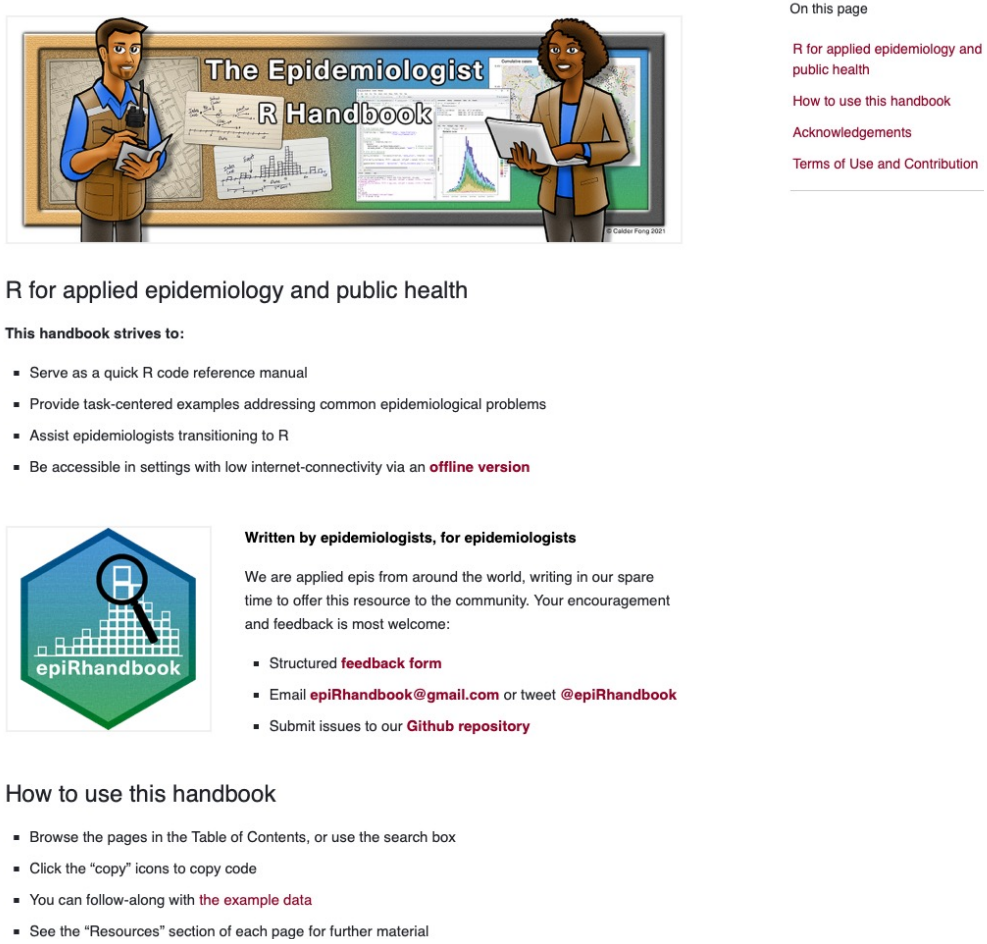
Mendelian Randomization

- Encodes several methods for performing Mendelian randomization analyses with summarized data. Summarized data on genetic associations with the exposure and with the outcome can be obtained from large consortia. These data can be used for obtaining causal estimates using instrumental variable methods.
- https://cran.r-project.org/web/packages/MendelianRandomization/index.html

The Epidemiologist R Handbook

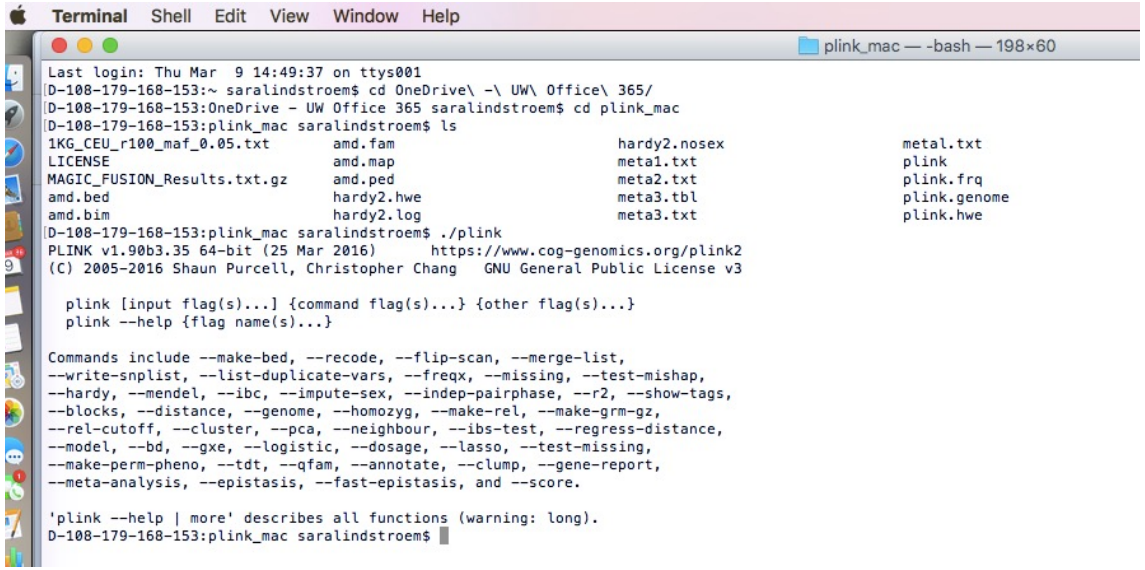
- https://epirhandbook.com/index.html
- Serve as a quick R code reference manual
- Provide task-centered examples addressing common epidemiological problems
- Assist epidemiologists transitioning to R
- Be accessible in settings with low internet-connectivity via an **offline version**
- Basics, Data Management, Analysis, Data Visualization, Reports and dashboards, Miscellaneous: writing functions, directory interactions, version control and collaboration with Git and Github, common errors, getting help, R on network drives, data table





Some additional software beyond R: PLINK

- What is PLINK?
  - Statistical software for analyzing phenotype/genotype data
  - Purcell S, et al. **PLINK**: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007.
  - Chang CC, et al. Second-generation **PLINK**: rising to the challenge of larger and richer datasets. Gigascience. 2015.
- Why PLINK?
  - It is arguably the most commonly used software for large-scale genetic association studies
  - It is fast
  - It is designed to conduct data quality control steps as well as generate descriptive statistics and run association analysis (just to mention a few things)
  - Many GWAS datasets are created in PLINK files (.bed, .bim, .fam)



Risk Prediction

- LDPred
  - LDpred is a Python based software package that adjusts GWAS summary statistics for the effects of linkage disequilibrium (LD).
  - https://github.com/bvilhjal/ldpred
- PRSics-2
  - PRSice (pronounced 'precise') is a Polygenic Risk Score software for calculating, applying, evaluating and plotting the results of polygenic risk scores (PRS) analyses.
  - http://www.prsice.info

Some additional software: Large scale data

- GATK
  - Variant Discovery in High-Throughput Sequencing Data. Includes multiple tools with a primary focus on variant discovery and genotyping.
  - https://gatk.broadinstitute.org/hc/en-us
- Hail
  - Python library that simplifies genomic data analysis. It provides powerful, easy-to-use data science tools that can be used to interrogate even biobank-scale genomic data (e.g., UK Biobank, **gnomAD**, TopMed, FinnGen, and Biobank Japan).
  - https://hail.is
- BOLT-LMM
  - The BOLT-LMM software package currently consists of two main algorithms, the BOLT-LMM algorithm for mixed model association testing, and the BOLT-REML algorithm for variance components analysis (i.e., partitioning of SNP-heritability and estimation of genetic correlations).
  - https://alkesgroup.broadinstitute.org/BOLT-LMM/BOLT-LMM_manual.html

One note about big data analyses

- Many research groups do their analysis on unix-based clusters
  - Firewalls, computational capacity, data storage
- Often packages come with great tutorials for analyses. The challenge will be for you to figure out how to run it on your cluster environment
- These are often specific things you will learn when you join a particular research group

Programming courses

- SISG 2021
  - https://si.biostat.washington.edu/suminst/sisg2021/modules
  - **Module 3: Introduction to R**
  - **Module 13: Association Mapping: GWAS and Sequencing Data**
  - **Module 16: Computational Pipeline for WGS Data**
- edX
  - https://www.edx.org/learn/computer-programming

R exercises

- http://faculty.washington.edu/tathornt/SISG2020.html

Time	Topic	Lecture	Exercises/Discussion
8:00am-9:20am	1. Introduction, Case Control Association Testing	Slides: (Intro) [.pdf], (Lecture) [.pdf], video	Exercises [.pdf], video , R Script:[.R] , Key: [.html], [.Rmd]
9:40am-11:00am	2. Association Testing with Quantitative Traits	Slides: [.pdf], video	Exercises: [.pdf], video R Script:[.R], Key: [.html], [.Rmd]
11:30am-12:50pm	3. Introduction to the PLINK Software for GWAS	Slides [.pdf], video	Exercises [.pdf], video, Plink Script: [.txt], R Script:[.R], Key (Rscript) : [.R]
1:10am-2:30pm	4. Gene and Pathway Level Analysis of Genetic Association Studies.	Slides [.pdf], video	Exercises [.pdf], video, Plink and R Script: [.txt]
Tuesday, July 28th			
Time	**Topic**	**Lecture**	**Exercises/Discussion**
8:00am-9:20am	5. Population Structure Inference	Slides [.pdf], video	Exercises [.pdf], video ,R Script: [.R], Key: [.html], [.Rmd]
9:40am-11:00am	6. GWAS in Samples with Structure	Slides [.pdf], video	Exercises [.pdf],video ,R Script:[.R]
11:30am-12:50pm	7. Interaction Analysis	Slides [.pdf], video	Exercises [.pdf], video , R Script:[.txt]
1:10am-2:30pm	8. Introduction to Rare Variant Analysis and Collapsing Tests	Slides [.pdf], video	Exercises [.pdf], video , Key: R Script:[.txt]
Wednesday, July 29th			
Time	**Topic**	**Lecture**	**Exercises/Discussion**
8:00am-9:20am	9. Rare Variant Analysis: Kernel (Variance Component) Tests and Omnibus Tests	Slides [.pdf], video	Exercises [.pdf], video , R Script:[.txt]
9:40am-11:00am	10. Power and Sample Size, Design Considerations, and Emerging Issues	Slides [.pdf], video	Exercises [.pdf], video , R Script:[.txt]

Datasets

A zipped folder with the genetic relatdness matrix (GRM) and other files for exercise 6, where a linear mixed model analysis is performed, can be $

Note: New link to zipped file with LMM files on dropbox posted below. The previously posted LMM zipped file was corrupted.

LMM_FILES_NEW.zip

All individual data files below can be downloaded as a single zipped folder from dropbox. This file can be downloaded here:

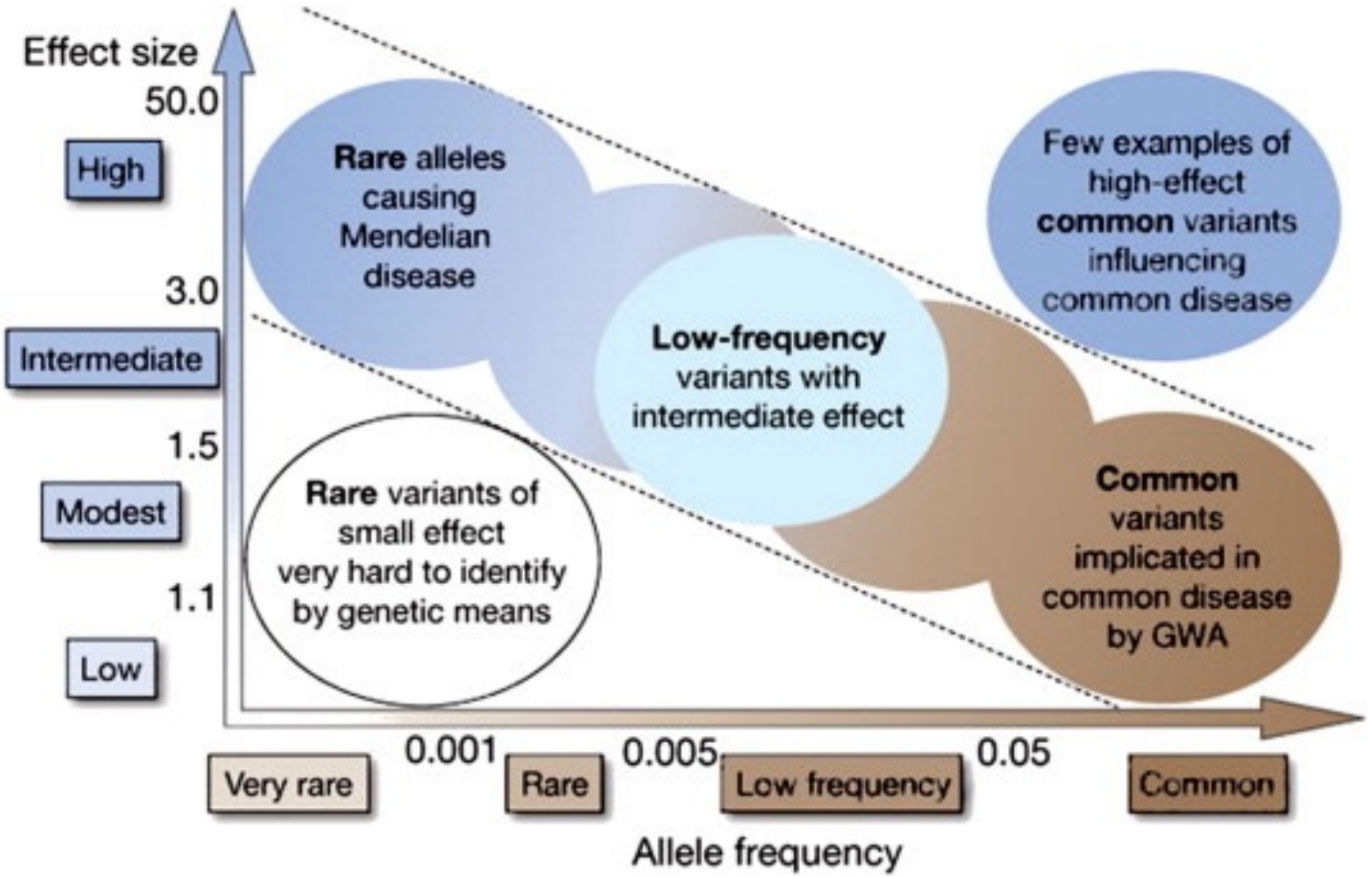
Note: New link to zipped file with the Data files on dropbox are posted below. The previously posted zipped file was corrupted.

SISG2020Data_NEW.zip

Alternatively, you can download each of the data files below. ***Before trying to read data into an R or PLINK session, we recommend looking at it first, in a text editor. Is the data comma- or tab-delimited? Does it have a 'header' row containing variable names?***

- **bpdata.csv**
- **Ht.pheno**
- **LHON.txt**
- **Population_Sample_Info.txt**
- **SNPlistHeight.txt**
- **SNPlistTransferrin.txt**
- **Tr.pheno**
- **YRI_CEU_ASW_MEX_NAM.bed**
- **YRI_CEU_ASW_MEX_NAM.bim**
- **YRI_CEU_ASW_MEX_NAM.fam**

Identifying genetic variation associated with disease



Manolio et al, Nature 2009

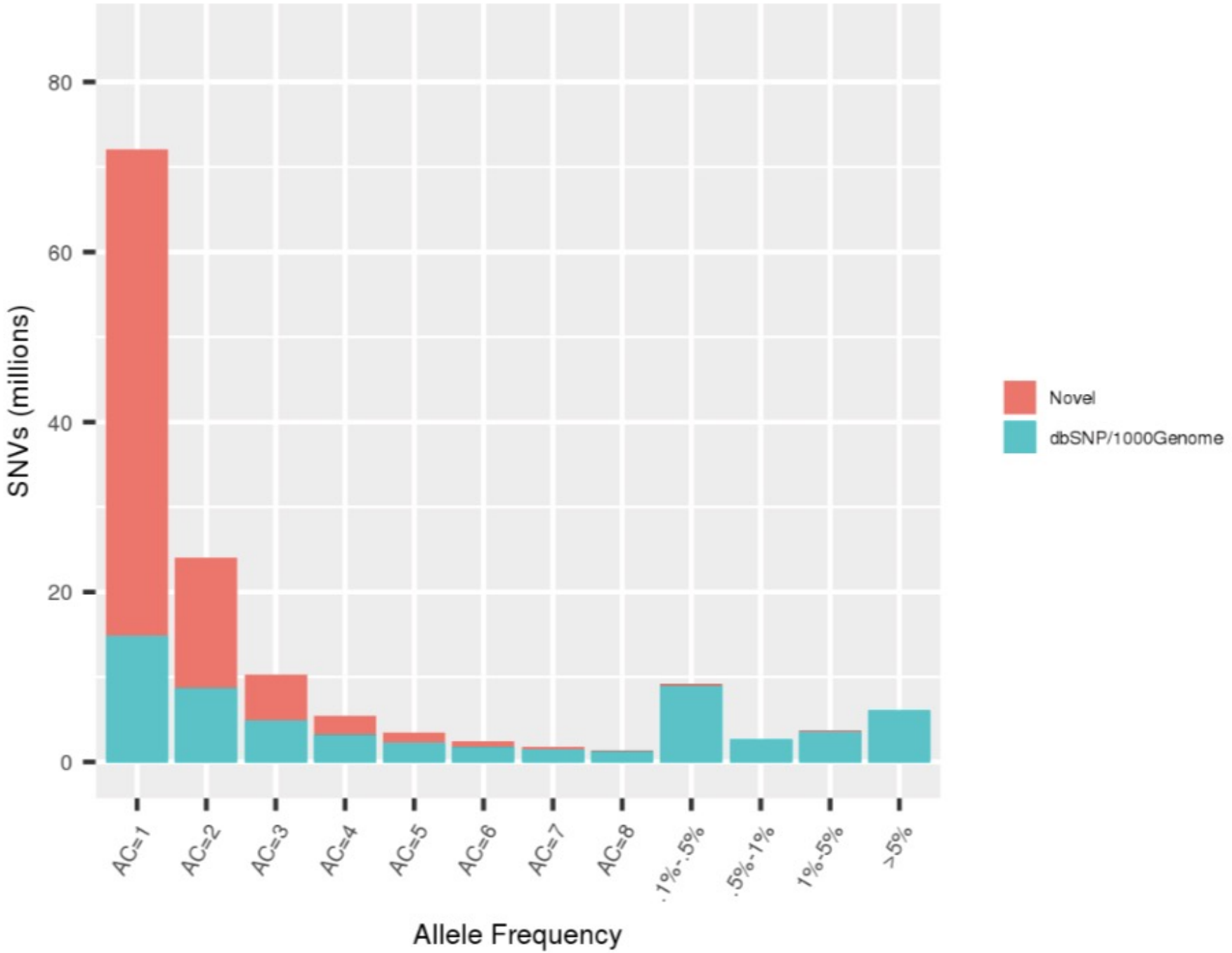
Introduction – Rare variants

- Usually less than 1% (depending on who you ask)

- Traditional single variant association analysis have low statistical power and/or are not valid
  - MAF=1% in 1,000 cases and 1,000 controls implies 40 minor alleles
  - Low cell counts lead to invalid statistical tests/low power

- Because the genome has many more rare variants than common variants, more stringent significance levels might be required, further reducing power

A recent study sequenced 10,545 human genomes and found more than 150 million variants



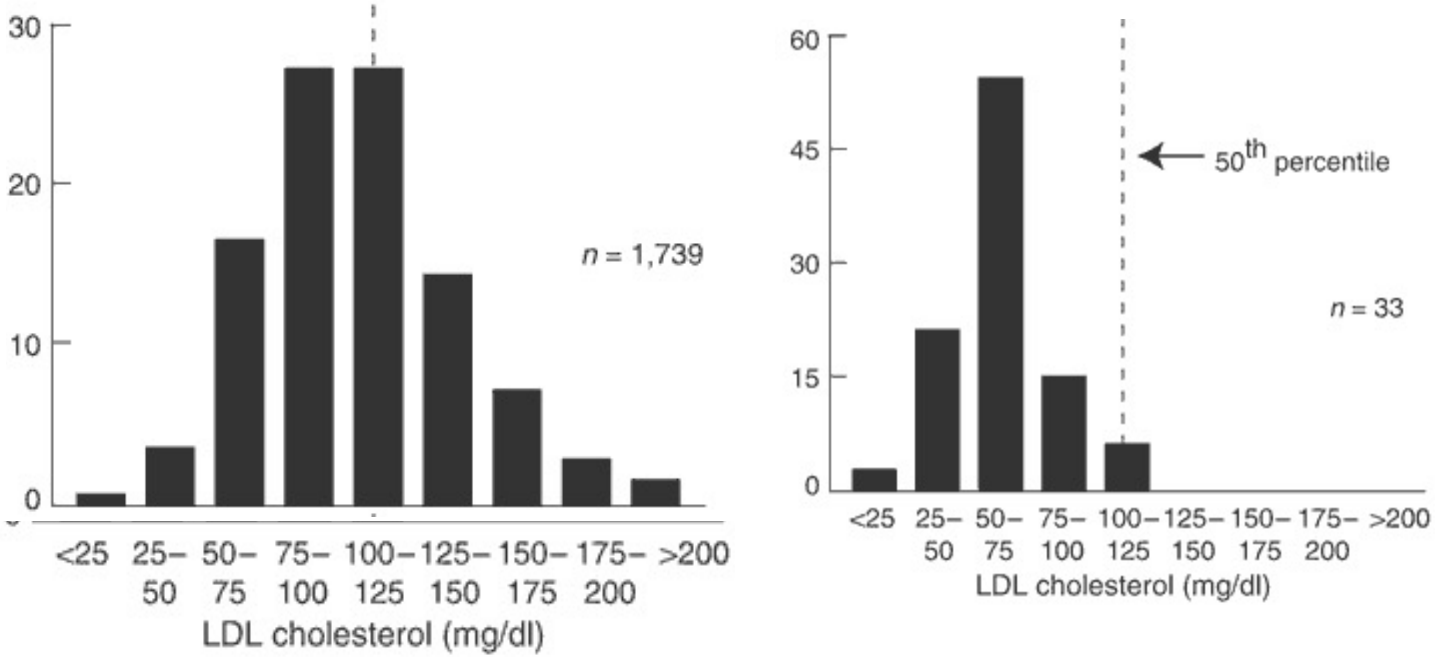
Telenti, PNAS 2016

Poll: Why study rare variants?

Why do we care about rare variants when they only affect a small proportion of the population?

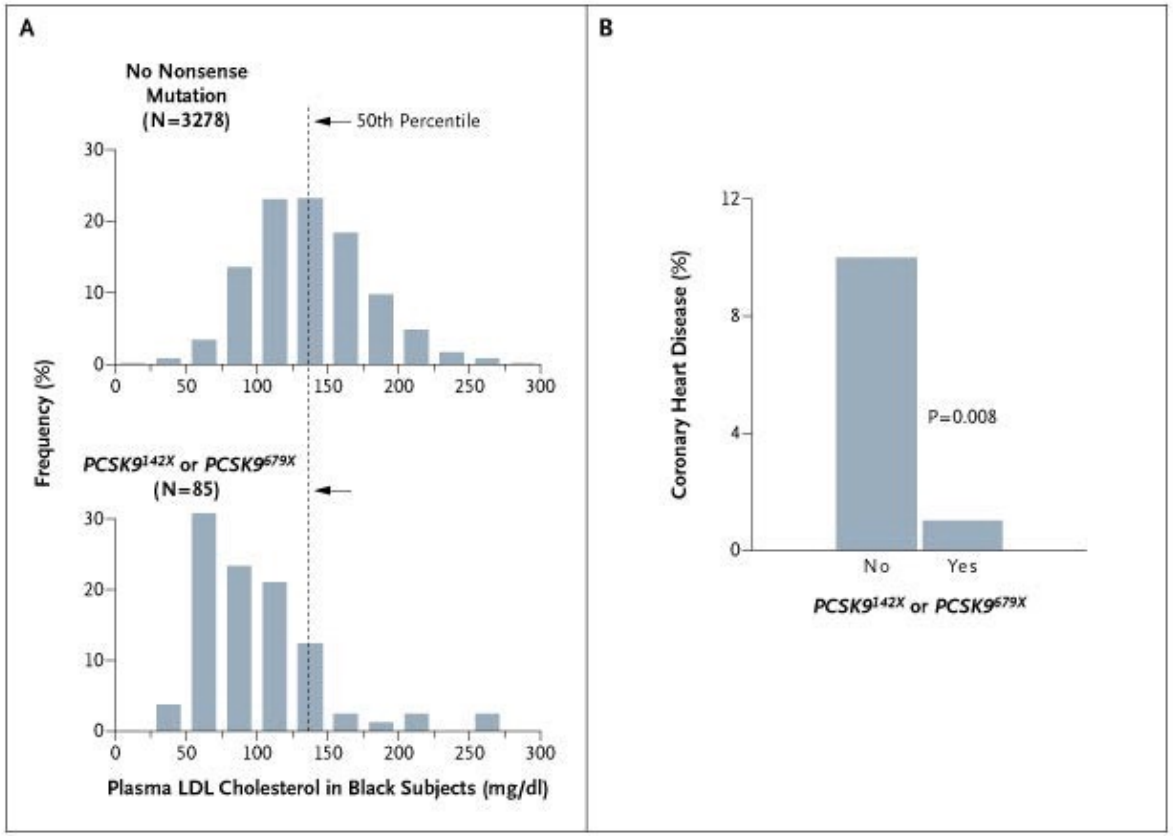
PCSK9 and LDL cholesterol

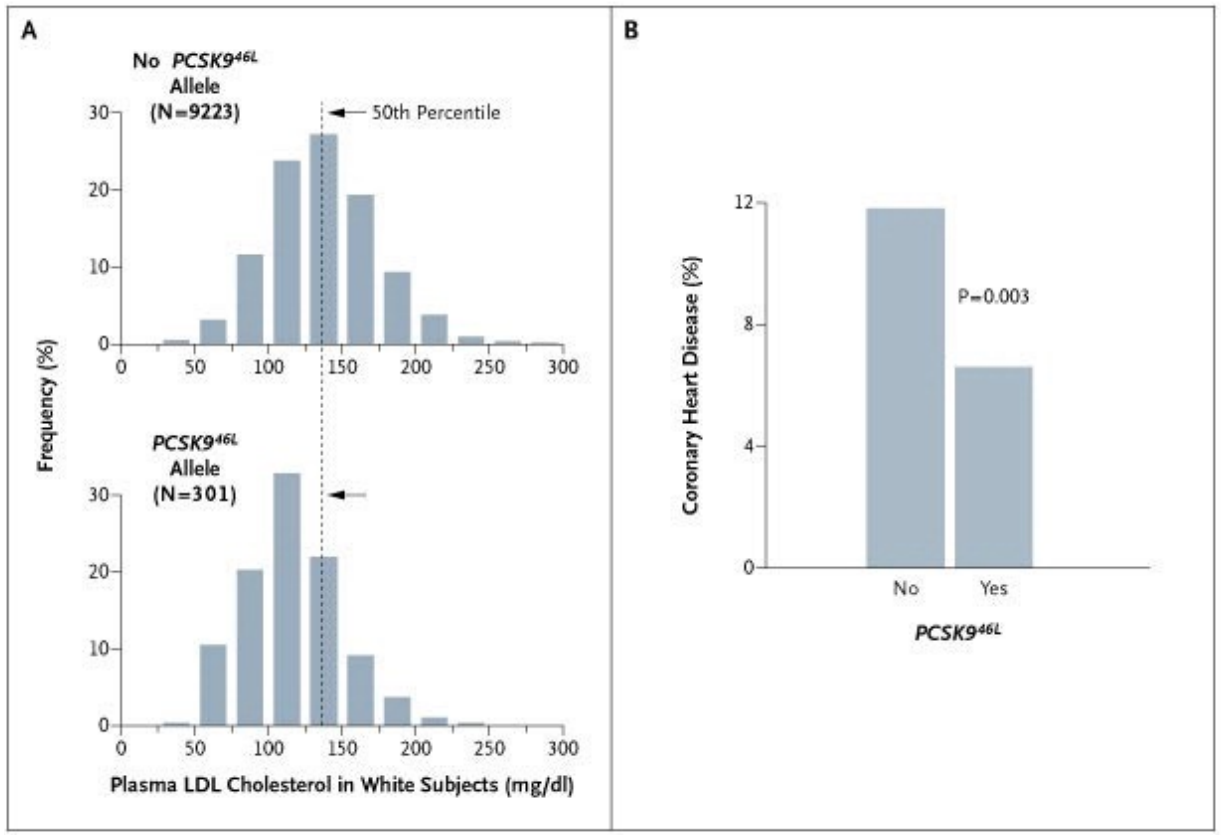
Plasma LDL-C levels in African American subjects without (left) and with (right) a nonsense mutation in *PCSK9*.



Cohen, Nat Genet 2005

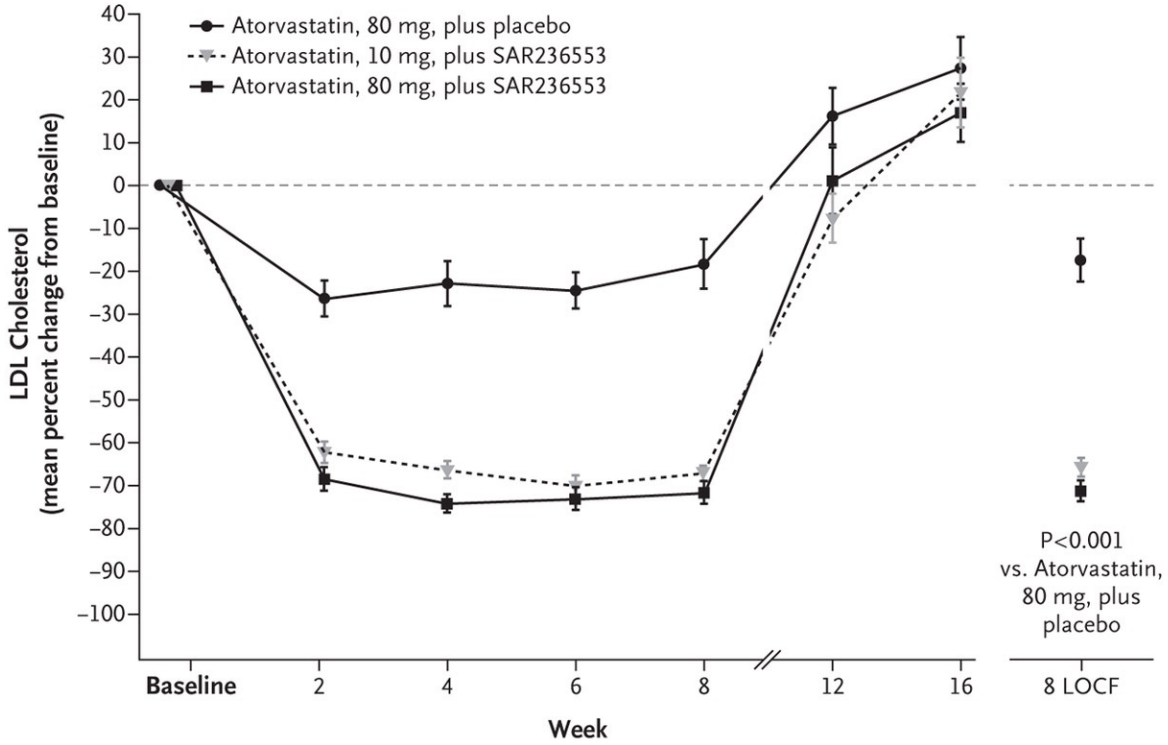
PCSK9 mutations and coronary heart disease





Cohen, NEJM 2005

A *PCSK9* antibody decreases LDL (8-week trial)



No. of Patients								
Atorvastatin, 80 mg, plus placebo	29	29	27	25	24	25	21	29
Atorvastatin, 10 mg, plus SAR236553	29	29	26	26	27	27	24	29
Atorvastatin, 80 mg, plus SAR236553	30	29	28	28	28	27	27	30

Roth, NEJM 2012

Study design for rare variant analysis

	Advantage	Disadvantage
High-depth WGS	can identify nearly all variants in the genome with high confidence	very expensive
Low-depth WGS	cost-effective and useful approach for association mapping	has limited accuracy for rare-variant identification and genotype calling; compared to deep sequencing, is subject to power loss if the same number of subjects is sequenced
Whole-exome sequencing	can identify all exonic variants; is less expensive than WGS	is limited to the exome
GWAS chip and imputation	inexpensive	has lower accuracy for imputed rare variants Will miss any variants unique to your sample
Exome chip (custom array)	much cheaper than exome sequencing	provides limited coverage for very rare variants and for non-Europeans is limited to target regions

Lee, AJHG 2014

Breakout room discussion

- You just got a large grant to identify rare variants associated with type 2 diabetes. You have colleagues around the world that can give you access to DNA from their case-control studies. If you were to design a study to identify rare (allele frequency <1%) variants associated with type 2 diabetes, what approach would you take and why?

  - High-depth whole genome sequencing
  - Low-depth whole genome sequencing
  - Whole exome sequencing
  - GWAS chip and imputation
  - Exome chip (custom array)

	Advantage	Disadvantage
High-depth WGS	can identify nearly all variants in the genome with high confidence	very expensive
Low-depth WGS	cost-effective and useful approach for association mapping	has limited accuracy for rare-variant identification and genotype calling; compared to deep sequencing, is subject to power loss if the same number of subjects is sequenced
Whole-exome sequencing	can identify all exonic variants; is less expensive than WGS	is limited to the exome
GWAS chip and imputation	inexpensive	has lower accuracy for imputed rare variants Will miss any variants unique to your sample
Exome chip (custom array)	much cheaper than exome sequencing	provides limited coverage for very rare variants and for non-Europeans is limited to target regions

Rare variant analysis - What to do?

- Many different rare variant tests are available.
  - Some are based on aggregating variants (“burden” tests)
    - CMC (Li and Leal, 2008)
    - WSS (Madsen and Browning, 2009)
    - Variable Threshold approach (Price, 2010)
  - Some are based on studying the distribution of variants
    - C-alpha (Neale, 2011)
    - SKAT (Wu, 2011)

Burden tests

- Collapse many variants into a single risk score
  - Combine minor allele counts into one variable

- Collapsing approach
  - Gene, pathways, functional annotations, etc
  - Much more straight-forward for coding regions

- Weighing
  - Variant type (predicted function)
  - Variant frequency

The Cohort Allelic Sums Test - CAST

<u>Main Idea:</u> Combine rare variants according to some (arbitrary) feature (gene, genetic region, functional category) and assess the new variable

Step 1: Create an indicator variable X for individual *j:*

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

Step 2: $\ln\left(\frac{p}{1-p}\right) = \alpha + \beta X$ (logistic regression)

Morgenthaler, Mutat Res 2007

Variant Collapsing – 2 approaches

i)

Subject	V1	V2	V3	V4	X
1	1	0	0	0	**1**
2	0	1	0	0	**1**
3	0	0	0	0	**0**
4	0	0	0	0	**0**
5	0	0	0	0	**0**
6	0	0	0	0	**0**
7	0	0	1	1	**1**
8	0	0	0	1	**1**

ii)

Subject	V1	V2	V3	V4	X
1	1	0	0	0	**1**
2	0	1	0	0	**1**
3	0	0	0	0	**0**
4	0	0	0	0	**0**
5	0	0	0	0	**0**
6	0	0	0	0	**0**
7	0	0	1	1	**2**
8	0	0	0	1	**1**

The Weighted Sum Statistic (WSS) – often called "Madsen-Browning"

- **Main idea:** Variants are grouped according to function (e.g., gene), and each individual is scored by a weighted sum of the variant counts.

- Use permutation to test for an excess of variants in affected individuals.

- Variants of all frequencies can be included, but variants are weighted according to their frequency in unaffected individuals.

$$\widehat{w}_i = 1/\sqrt{q_i(1-q_i)}$$

q_i is the estimated MAF in controls

Madsen and Browning, PLoS Genetics, 2009

Drawback with burden tests

- Assume all variants in a set are causal and associated with a trait in the same direction. The common assumption is often that the rare allele increases disease risk

- If this is not true, power is lost.

- Solution: Tests that look at the distribution of rare variants

The C-alpha test

- Main idea: Test whether observed variants either increase or decrease risk (or have no effect). Risk variants are expected to be more common in cases; protective variants more common in controls.

- If there is no association, variants are distributed randomly between cases and controls following a binomial (n,p) distribution.

- For example, if the case:control ratio is 1:1, a variant seen twice (doubleton) would be observed in cases y times where y is either 0, 1 and 2 with probability ¼, ½ and ¼, respectively.

Neale, PLoS Genetics 2011

***APOB* variant counts in individuals with high/low triglyceride levels.**

Position	Annotation	High Lipid Level	Low Lipid Level
21078358	Ala4481Thr	2	5
21078359	Ile4314Val	3	0
21078990	Arg4270Thr	6	3
21079417	Val4128Met	1	7
21083082	Thr3388Lys	2	1
21083637	Ser3203Tyr	6	0
21086035	Leu2404Ile	2	3
21086072	Glu2391Asp	2	2
21086127	Thr2373Asn	2	2
21086308	Val2313Ile	2	1
21087477	His1923Arg	6	12
21087504	Asn1914Ser	0	5
21087634	Asp1871Asn	2	0
21091828	Pro1143Ser	0	6
21091872	Arg1128His	0	3
21091918	Asp1113His	1	3
21106140	Thr498Asn	2	0
Singletons		6	4

Nonsynonymous variants discovered via targeted pooled sequencing in 192 individuals with extreme triglyceride levels. High counts represent the number of copies of the variant discovered in 96 individuals who have high triglycerides (defined as exceeding the 5% upper tail of the population distribution). Low counts represent the number of copies of the variant discovered in 96 individuals who have low triglycerides (lower 5% tail). The singletons are grouped together and listed as the penultimate row because its total count is second largest (10, versus 18 for the His1923Arg). For details about pooled sequencing, see Text S1.
doi:10.1371/journal.pgen.1001322.t001

SKAT: sequence kernel association test

- In contrast to the C-alpha test, SKAT is regression-based and thereby allows for adjustment of covariates.
- Uses a variance-component score test in a mixed-model framework to assess regression coefficients for rare variants.

$$logit\ P(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X}_i + \boldsymbol{\beta}'\boldsymbol{G_i}$$

$\boldsymbol{y_i}$: case-control status; α_0: intercept; $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_m]'$ is the vector of regression coefficients for the m covariates; $\boldsymbol{X_i}$: fixed effects of covariates; $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]'$ is the vector of regression coefficients for the p observed gene variants in the region; $\mathbf{G_i}$: $(G_{i1}, G_{i2}, \ldots, G_{ip})$ genotypes for the p variants within the region

$$H_0: \boldsymbol{\beta} = \mathbf{0} \text{ or } \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

Wu, AJHG 2011

Combined tests

- SKAT-O
  - Picks the best combination of SKAT and a burden test, and then corrects for the flexibility afforded by this choice. Specifically, if the SKAT statistic is Q1, and the squared score for a burden test is Q2, SKAT-O considers tests of the form (1-rho)*Q1 + rho*Q2, where rho is between 0 and 1.

Lee, AJHG 2012

Table 2. Summary of Statistical Methods for Rare-Variant Association Testing

	Description	Methods	Advantage	Disadvantage	Software Packages[a]
Burden tests	collapse rare variants into genetic scores	ARIEL test,[50] CAST,[51] CMC method,[52] MZ test,[53] WSS[54]	are powerful when a large proportion of variants are causal and effects are in the same direction	lose power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	EPACTS, GRANVIL, PLINK/SEQ, Rvtests, SCORE-Seq, SKAT, VAT
Adaptive burden tests	use data-adaptive weights or thresholds	aSum,[55] Step-up,[56] EREC test,[57] VT,[58] KBAC method,[59] RBT[60]	are more robust than burden tests using fixed weights or thresholds; some tests can improve result interpretation	are often computationally intensive; VT requires the same assumptions as burden tests	EPACTS, KBAC, PLINK/SEQ, Rvtests, SCORE-Seq, VAT
Variance-component tests	test variance of genetic effects	SKAT,[61] SSU test,[62] C-alpha test[63]	are powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	are less powerful than burden tests when most variants are causal and effects are in the same direction	EPACTS, PLINK/SEQ, SCORE-Seq, SKAT, VAT
Combined tests	combine burden and variance-component tests	SKAT-O,[64] Fisher method,[65] MiST[66]	are more robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants	can be slightly less powerful than burden or variance-component tests if their assumptions are largely held; some methods (e.g., the Fisher method) are computationally intensive	EPACTS, PLINK/SEQ, MiST, SKAT
EC test	exponentially combines score statistics	EC test[67]	is powerful when a very small proportion of variants are causal	is computationally intensive; is less powerful when a moderate or large proportion of variants are causal	no software is available yet

Lee, AJHG 2014

Issues in rare variant analysis (i)

- Which variants to include?
  - All variants
  - Only those we think are deleterious

- How to group variants?
  - For exome analysis, rare variants are often grouped by gene making variant grouping straight-forward.
  - For whole-genome analysis, alternative approaches such as sliding window or additional functional annotations (conserved regions, regulatory regions etc.) can be used

Issues in rare variant analysis (ii)

- Which association test to use?
  - If there are multiple variants with risk-increasing effects, burden tests are most powerful
  - If there is a mixture of risk increasing and risk decreasing variants and/or most variants do not have an effect, variance-component methods are most powerful
  - If no prior information is available, conduct both burden and variance component tests. Have to consider multiple testing.

- Population stratification
  - It is not clear how effective PCs are for dealing with population stratification

Issues in rare variant analysis (iii)

- In general, rare variants are more difficult to impute

- Replication is more complex for rare variants:
  - Since the variants are by definition rare, they might be unique to the discovery population

  - Replication of single variants is straightforward: genotype the variant in the replication population

  - For gene-based association tests: Sequencing the gene (or region) can identify additional variants

  - KEY STRATEGY: Maximize number of samples in your replication!