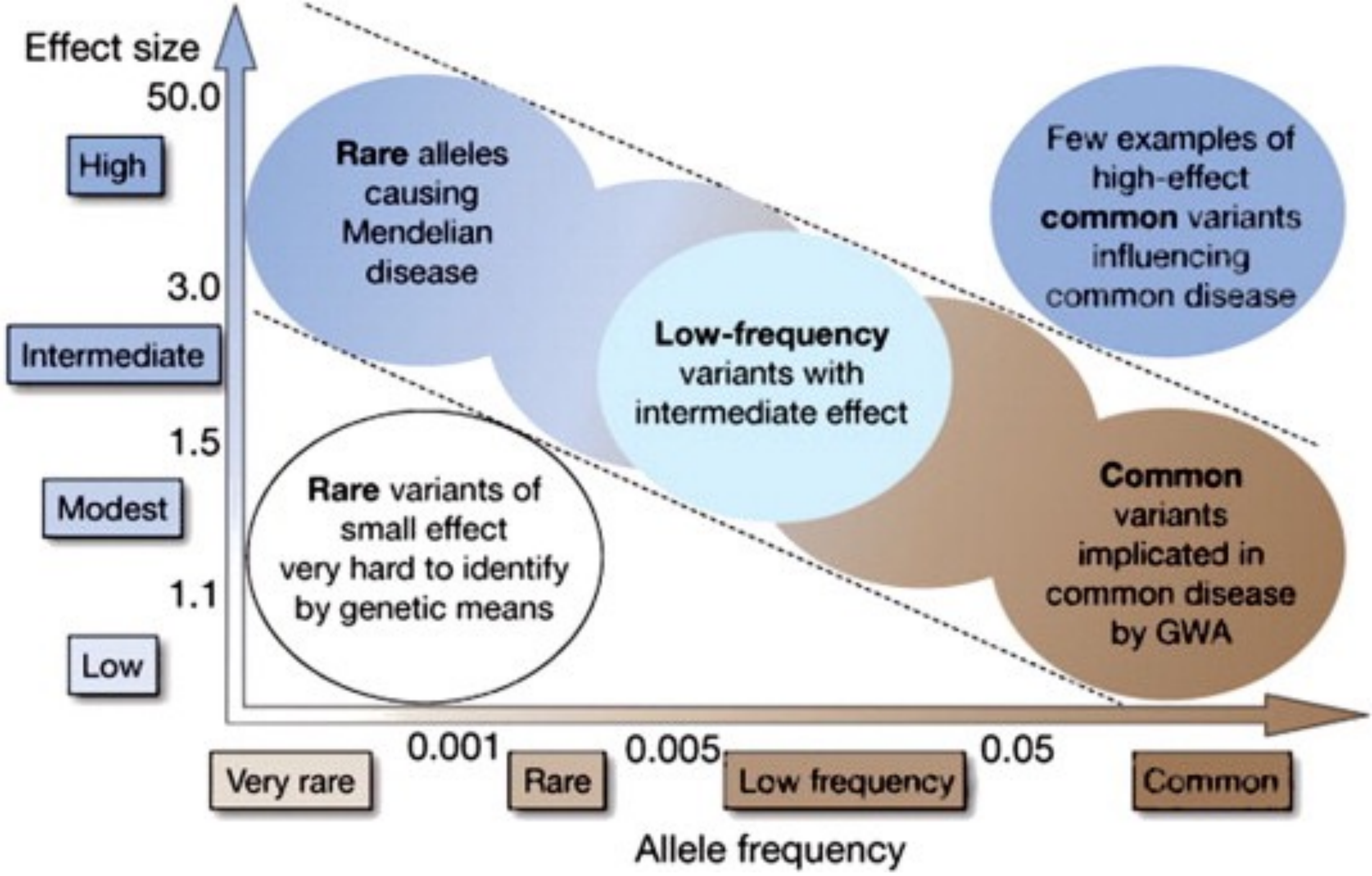
SCHOOL OF PUBLIC HEALTH
EPIDEMIOLOGY
UNIVERSITY *of* WASHINGTON

Session 8:
Rare variant association studies

Identifying genetic variation associated with disease


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Manolio et al, Nature 2009

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Introduction – Rare variants

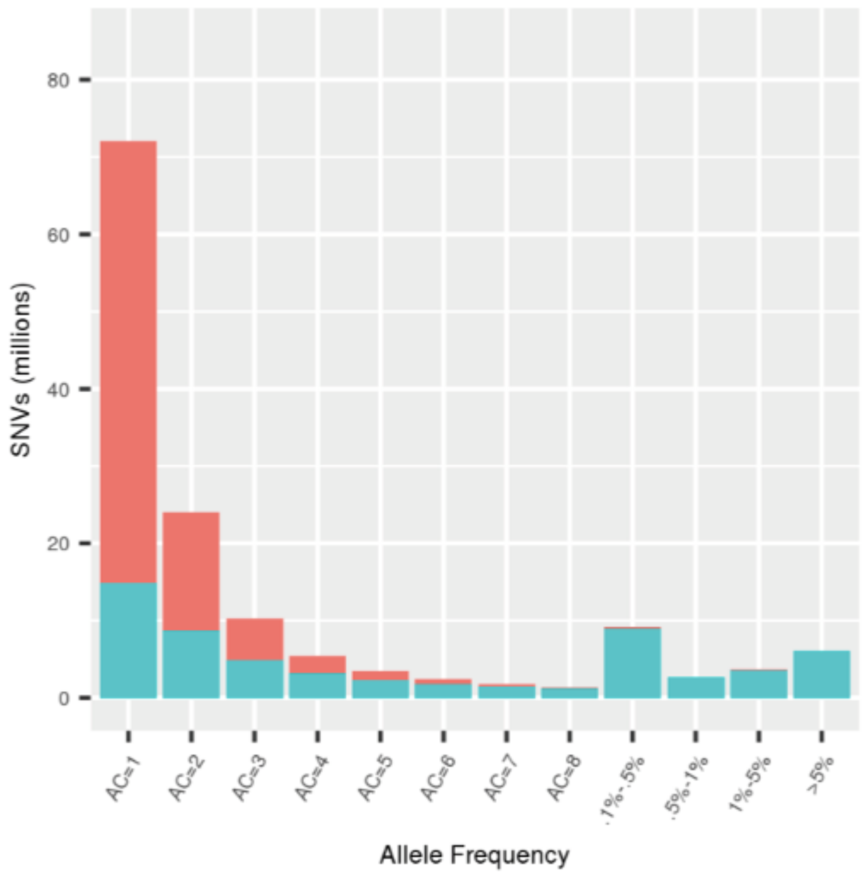
- Rare variants: Genetic variants with a minor allele frequency (MAF) less than 1% (sometimes < 0.5% depending on who you ask)

- Traditional single variant association analyses have low statistical power and/or are not valid
  - MAF = 1% in 1,000 individuals translates to a total of 20 minor alleles
  - Low cell counts lead to invalid statistical tests/low power

- As the total number of rare variants is far greater than the number of common variants, more stringent significance levels may be required, further reducing power to detect associations

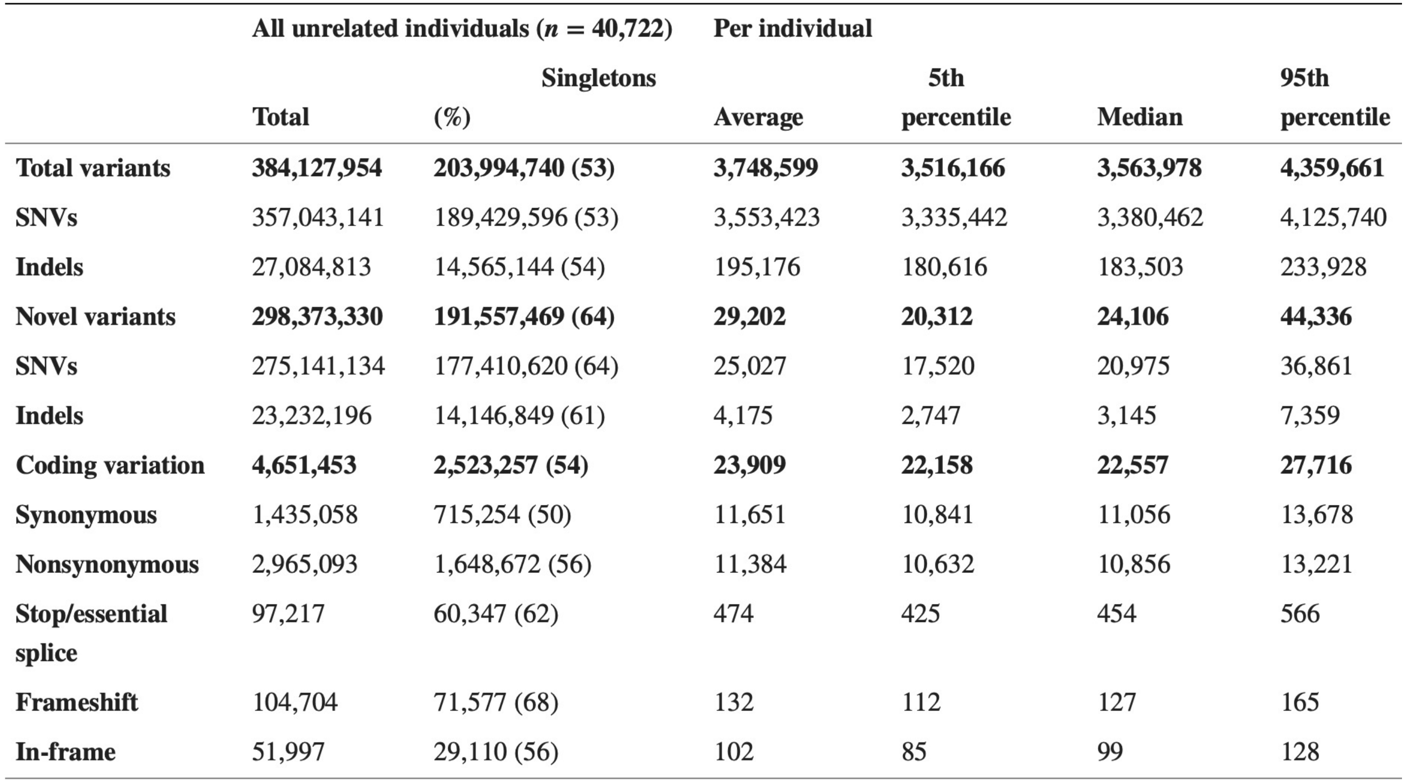
Most of the human genetic variation is rare

N=10,545 genomes, 150 million variants



Telenti, PNAS 2016

N=40,722 genomes, 384 million variants

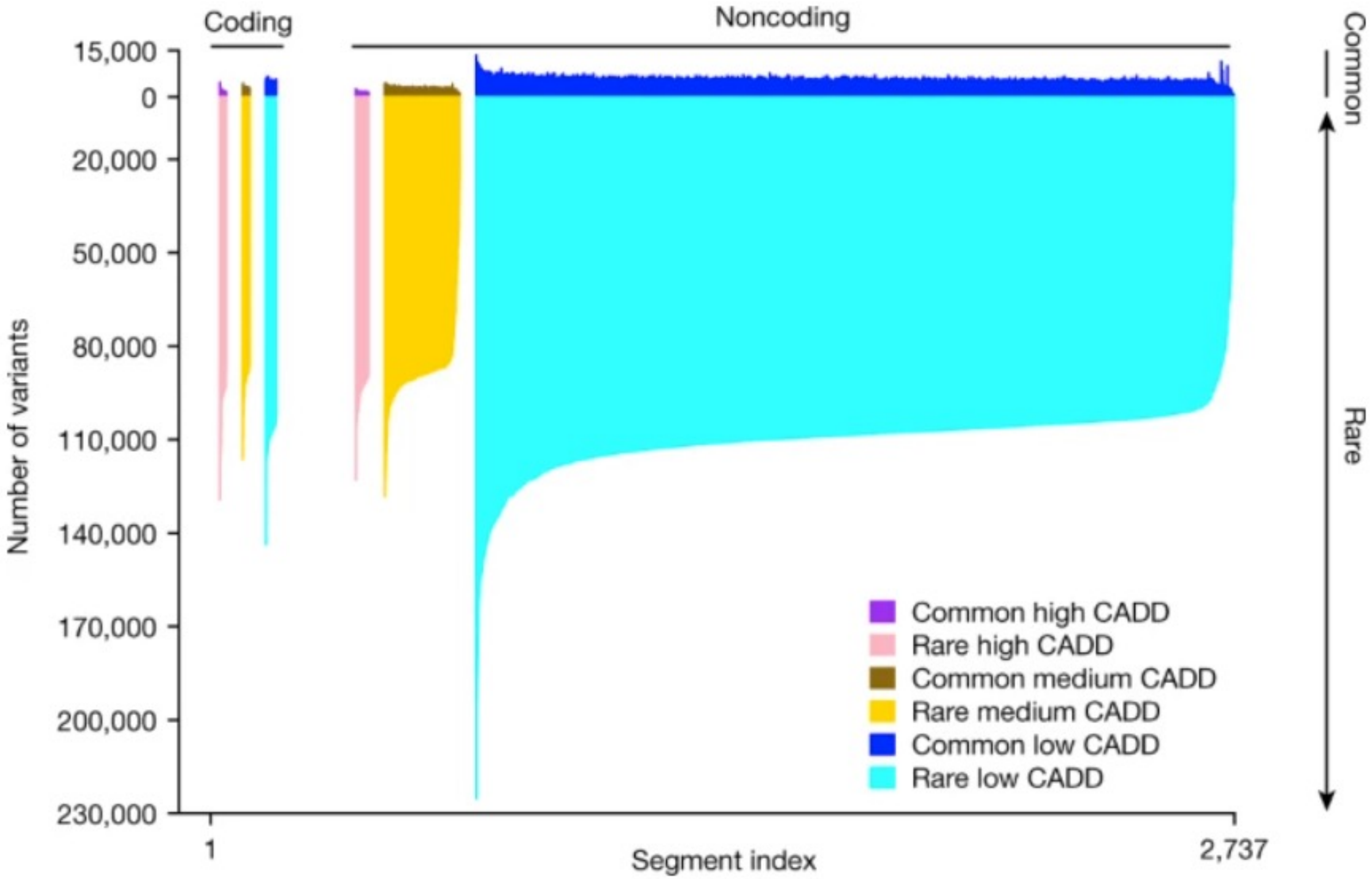


	All unrelated individuals (*n* = 40,722)		Per individual			
	Total	Singletons (%)	Average	5th percentile	Median	95th percentile
Total variants	**384,127,954**	**203,994,740 (53)**	**3,748,599**	**3,516,166**	**3,563,978**	**4,359,661**
SNVs	357,043,141	189,429,596 (53)	3,553,423	3,335,442	3,380,462	4,125,740
Indels	27,084,813	14,565,144 (54)	195,176	180,616	183,503	233,928
Novel variants	**298,373,330**	**191,557,469 (64)**	**29,202**	**20,312**	**24,106**	**44,336**
SNVs	275,141,134	177,410,620 (64)	25,027	17,520	20,975	36,861
Indels	23,232,196	14,146,849 (61)	4,175	2,747	3,145	7,359
Coding variation	**4,651,453**	**2,523,257 (54)**	**23,909**	**22,158**	**22,557**	**27,716**
Synonymous	1,435,058	715,254 (50)	11,651	10,841	11,056	13,678
Nonsynonymous	2,965,093	1,648,672 (56)	11,384	10,632	10,856	13,221
Stop/essential splice	97,217	60,347 (62)	474	425	454	566
Frameshift	104,704	71,577 (68)	132	112	127	165
In-frame	51,997	29,110 (56)	102	85	99	128

Taliun, Nature 2021

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Distribution of genetic variants across TOPMed genomes



What are some things that we notice about the distribution variants?

CADD is a score for the predicted effect of a variant (high CADD = predicted deleterious)

Taliun et al., Nature 2021

Common (allele frequency ≥ 0.5%) and rare (allele frequency < 0.5%)

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

SCHOOL OF PUBLIC HEALTH
EPIDEMIOLOGY
UNIVERSITY *of* WASHINGTON

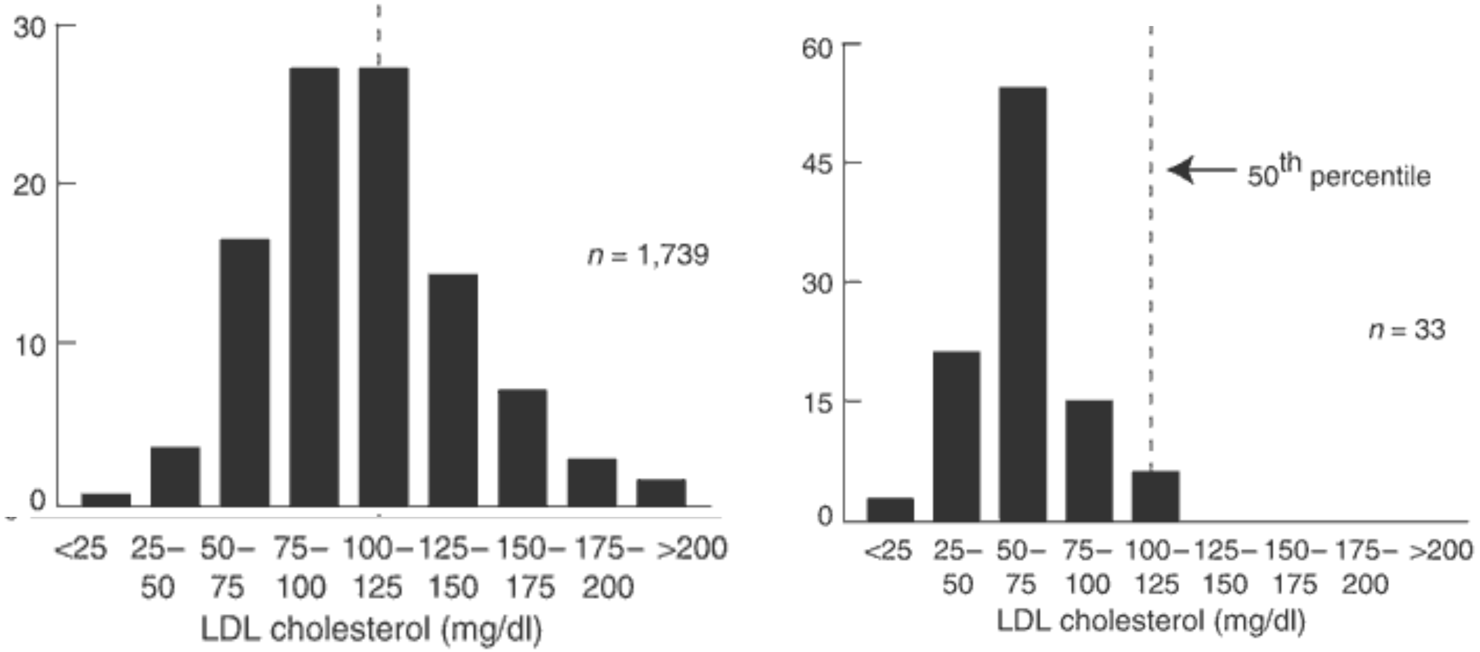
Rare variants are individually rare but collectively common and make up the vast majority of total human genetic variation

Why study rare variants?

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

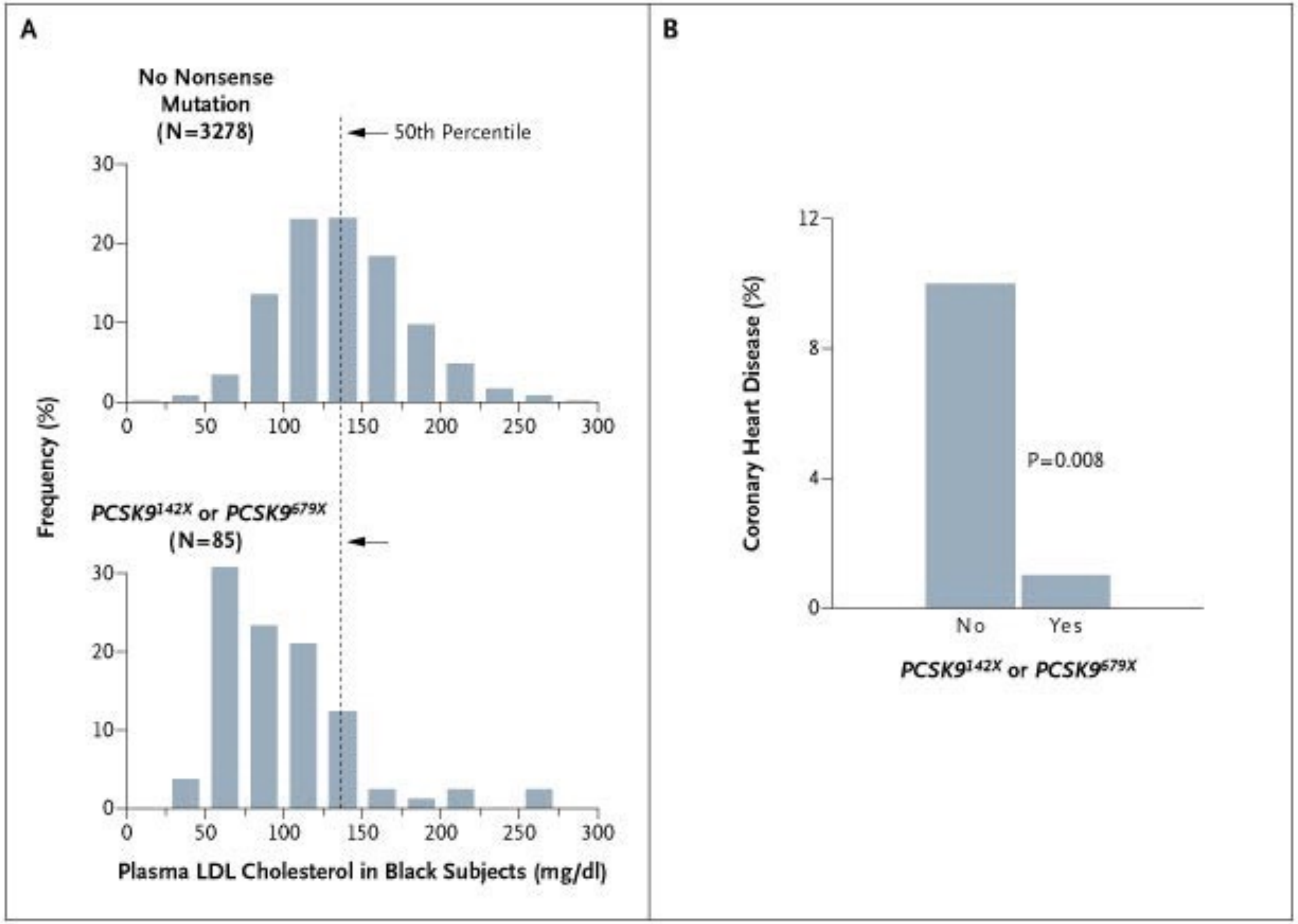
PCSK9 and LDL cholesterol

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



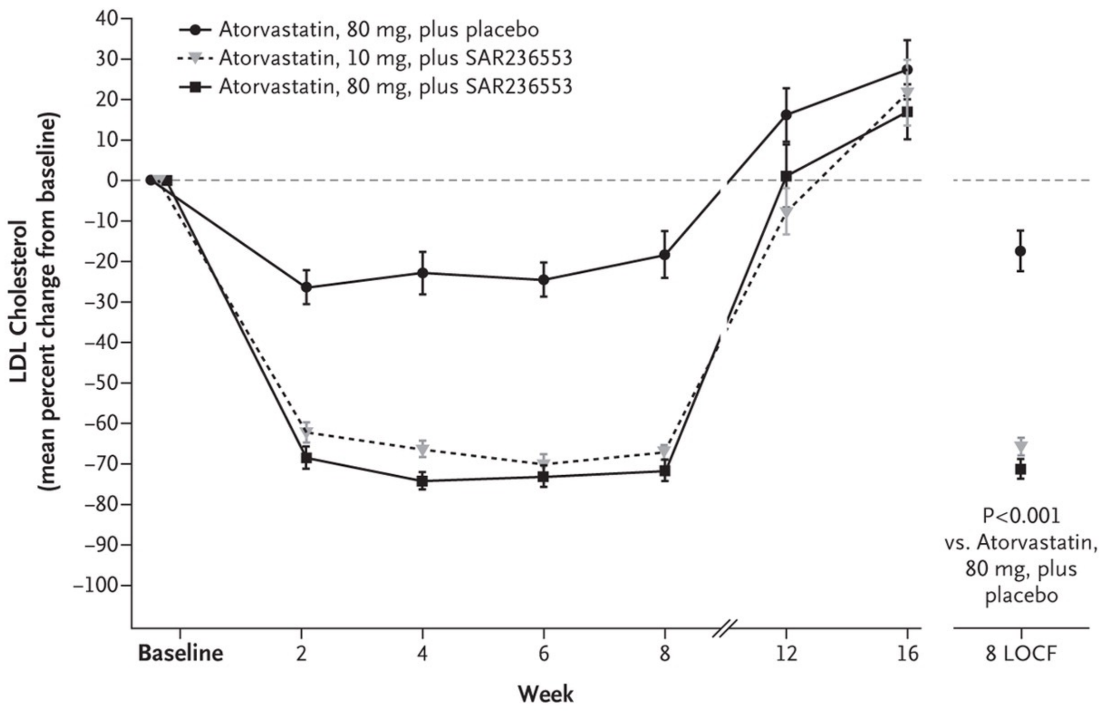
Cohen, Nat Genet 2005

PCSK9 mutations and coronary heart disease



Cohen, NEJM 2005

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



P<0.001 vs. Atorvastatin, 80 mg, plus placebo

No. of Patients

	Baseline	2	4	6	8	12	16	8 LOCF
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

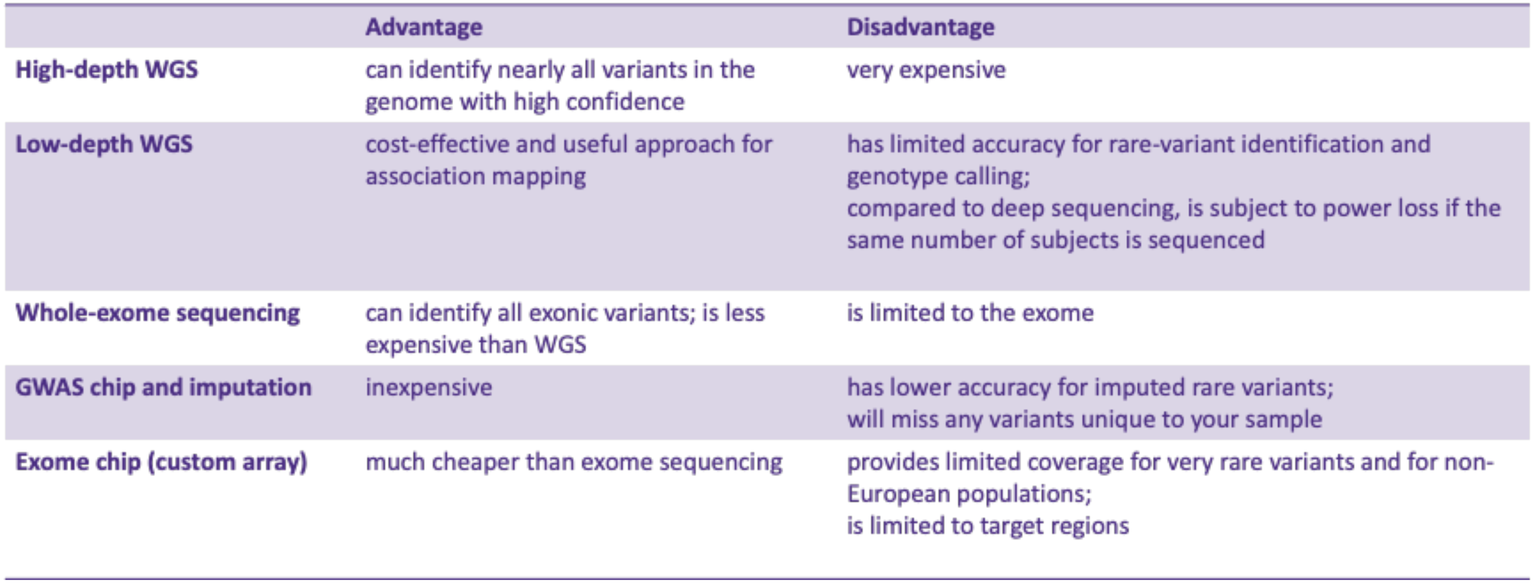
Study design for rare variant analysis

	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-European populations; is limited to target regions

Lee, AJHG 2014

Breakout Discussion

- If you were to design a study to identify rare (allele frequency <1%) variants associated with ovarian cancer, 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-European populations; is limited to target regions

Analyses of rare variants

- Many different rare variant tests are available, but most fall into one of two major categories
  - 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

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)

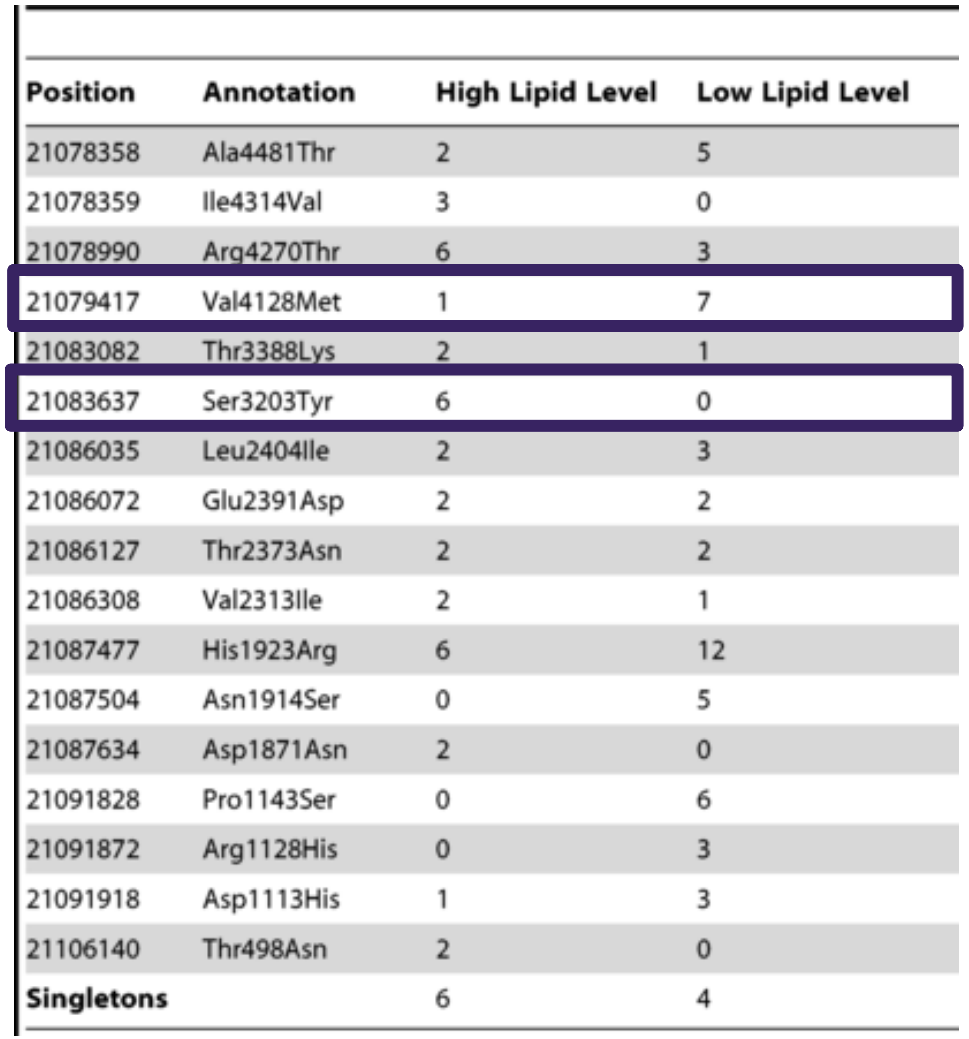
Subject	V1	V2	V3	V4	X
1	1	0	0	0	**1**
2	0	1	0	0	**1**
3	0	0	0	0	**0**
4	0	0	0	0	**0**
5	0	0	0	0	**0**
6	0	0	0	0	**0**
7	0	0	1	1	**2**
8	0	0	0	1	**1**

Disadvantage of burden tests

> Burden tests assume that all variants in a set are causal and associated with the outcome in the same direction. If this is not true, power is lost.

> Potential solution: Use a test that assesses the distribution of rare variants in a set.

***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

Neale, PLoS Genetics 2011

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

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.

$$logit\, P(y_i = 1) = \alpha_0 + \alpha' X_i + \beta' G_i$$

***y**_i_*: case-control status; α_0: intercept; **α** = $[\alpha_1, \ldots, \alpha_m]'$ is the vector of regression coefficients for the *m* covariates; ***X**_i_*: fixed effects of covariates; **β** = $[\beta_1, \ldots, \beta_p]'$ is the vector of regression coefficients for the *p* observed gene variants in the region; **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

Tradeoffs between SKAT and burden tests

> Burden tests tend to have higher power when a larger proportion of variants in a set have an effect on the outcome AND most variants have consistent direction of association.

> SKAT tends to have higher power when a smaller proportion of variants in a set have an effect on the trait OR the directions of associations are inconsistent

> Both scenarios are biologically plausible for a given set of variants. We typically do not know *a priori* if a burden or SKAT test will be more powerful.

Combined test: SKAT-O

- Picks the best combination of SKAT and a burden test, and then corrects for the flexibility afforded by this choice.
  - If the SKAT statistic is Q_1, and the squared score for a burden test is Q_2, SKAT-O considers tests of the form

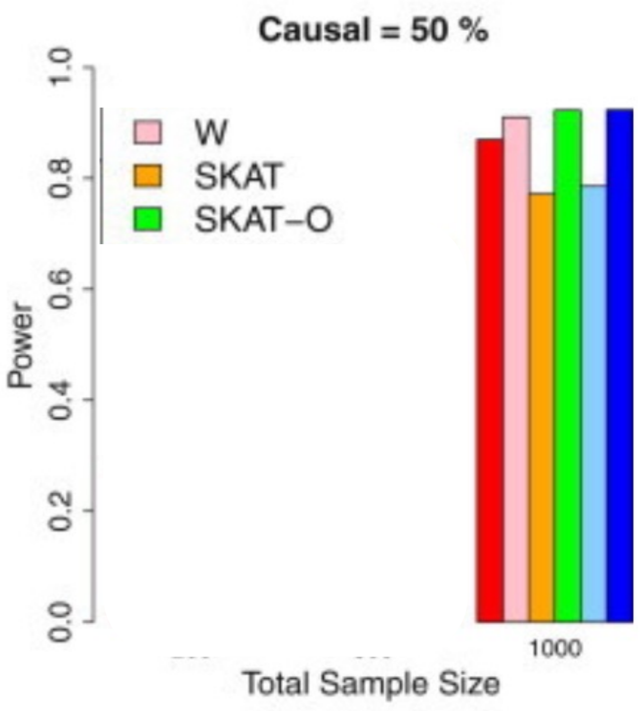
$$(1-\rho) \times Q_1 + \rho \times Q_2, \text{ where } \rho \text{ is between 0 and 1}$$

- ρ is selected to maximize the power of the test for each variant set
- When $\rho = 1$, SKAT-O is a burden test
- When $\rho = 0$, SKAT-O is a SKAT test
- When $0 < \rho < 1$, SKAT-O is a linear combination of a burden and SKAT test

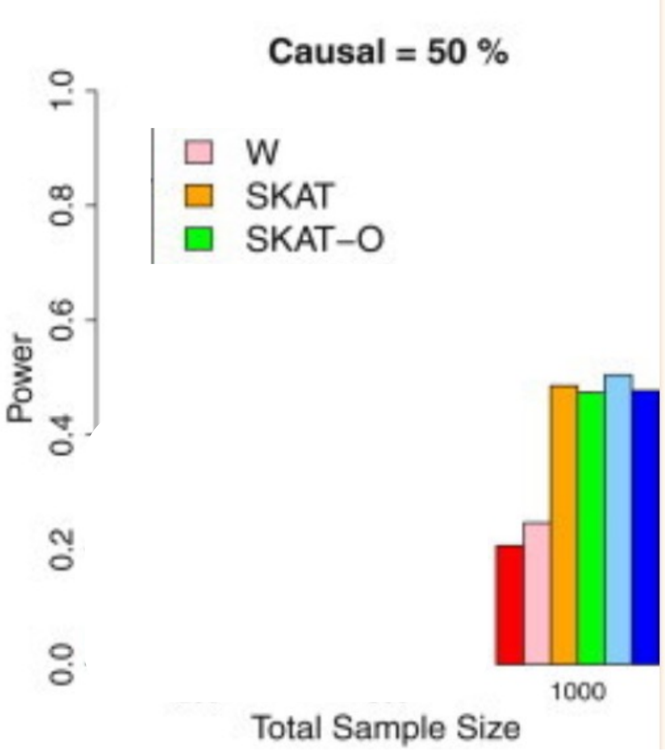
Lee, AJHG 2012

Statistical Power

100% of causal variants are deleterious (0% protective)



80% of causal variants are deleterious (20% protective)



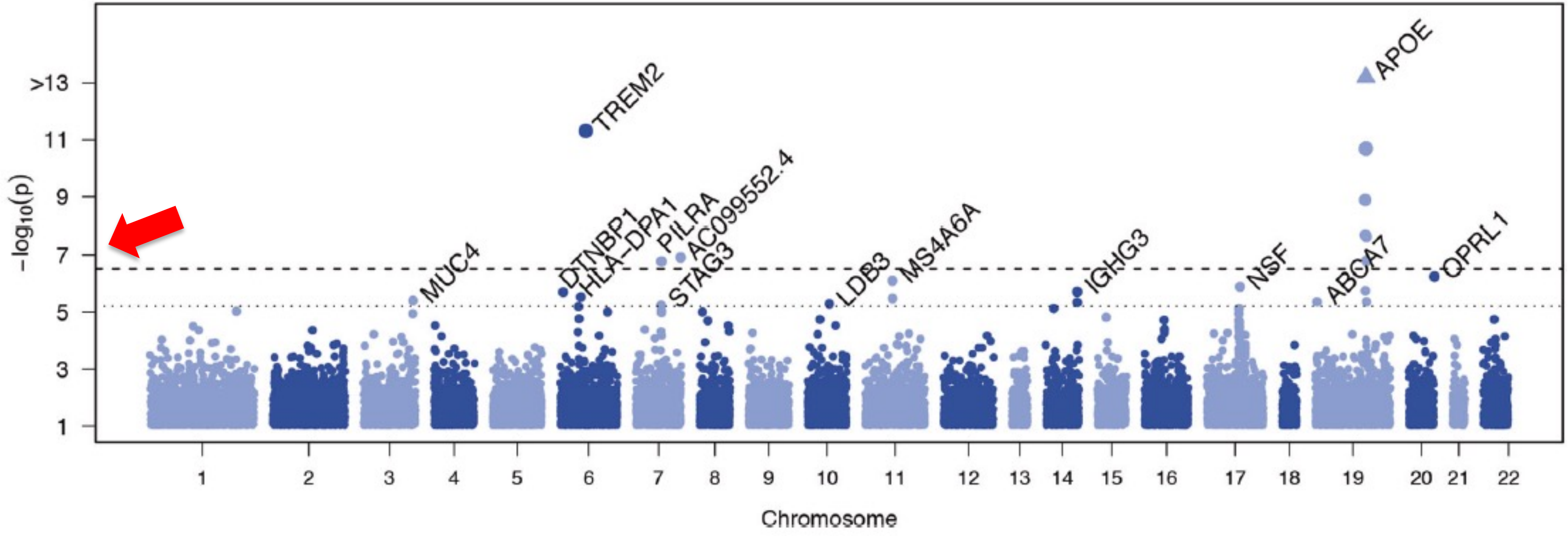
Key points:

- The power of burden and SKAT tests depend on the features of the variant set being tested
- In theory, the power of a SKAT-O test will be similar to the power of the best individual test in each scenario

Lee, AJHG 2012

Example SKAT-O analysis (Alzheimer's disease diagnosis)

Whole exome sequencing data from 5,740 AD cases and 5,096 controls



The significance threshold is lower than typical genome-wide significance for a GWAS. There are also fewer points that we usually see on Manhattan plots. Any thoughts on why this might be?

Bis et al., Mol. Psychiatry 2017

W EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

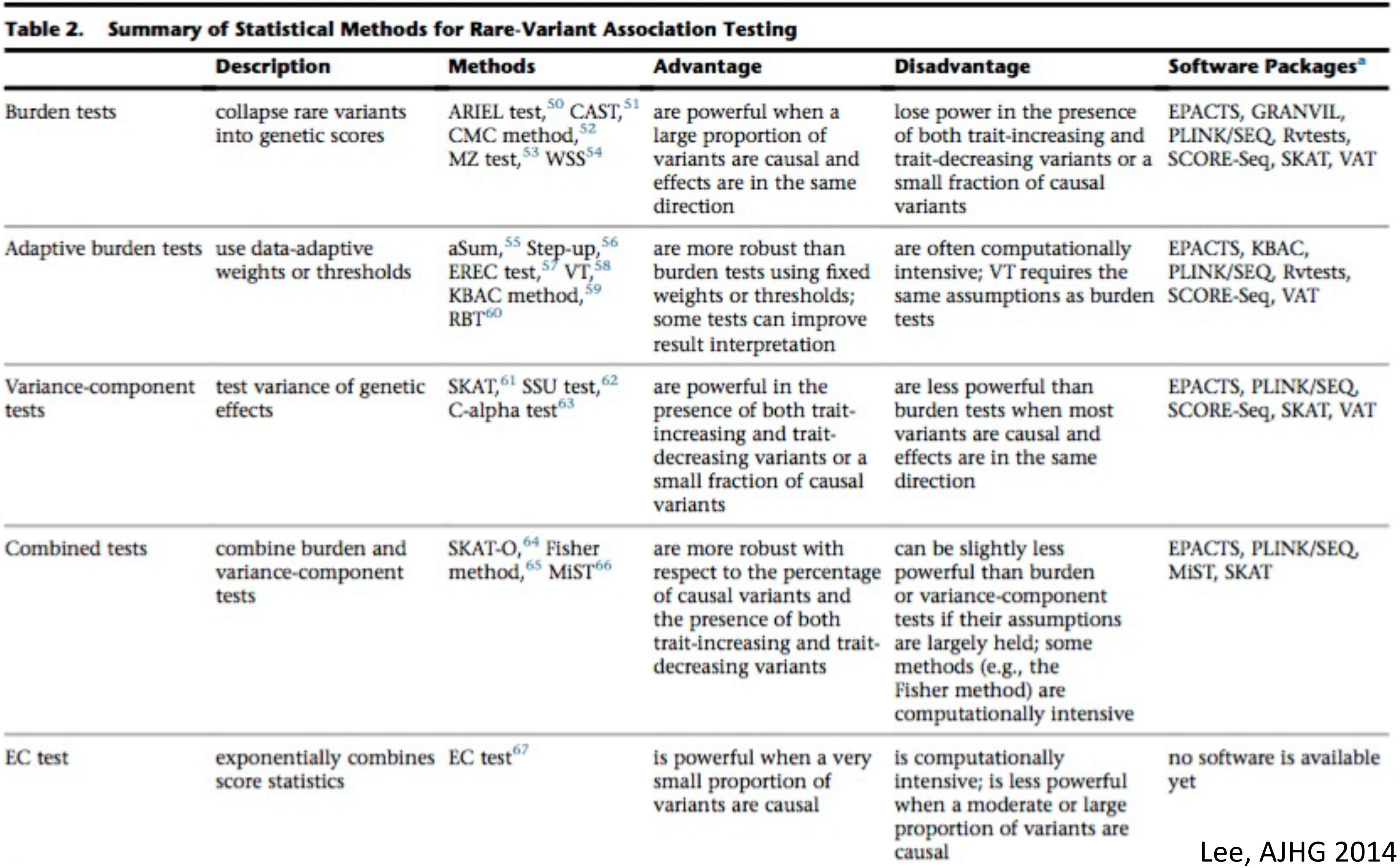


Table 2. Summary of Statistical Methods for Rare-Variant Association Testing

	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

Rare variant analyses software

- Rvtests (http://zhanxw.github.io/rvtests/)
- SKAT (https://cran.r-project.org/web/packages/SKAT/index.html)
- GENESIS (https://bioconductor.org/packages/devel/bioc/vignettes/GENESIS/inst/doc/assoc_test_seq.html)
- SAIGE-GENE+ (https://github.com/saigegit/SAIGE)

Issues in rare variant analysis (i)

- Which variants do we include?
  1. All variants
     - Most variants likely have no effect on our outcome
  2. Only those we think are deleterious
     - How do we determine/predict deleteriousness?
     - What if we get rid of some variants that have effects on our outcome?

- How should we group variants?
  - Rare variants are often grouped by their functional unit such as by gene. This makes variant grouping straight-forward in exome studies
  - 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)

- In general, rare variants are more difficult to impute compared to rare variants
- Replication is more complex for rare variants since the variants of interest might not be shared across datasets
- Adjusting for population stratification and cryptic relatedness may be more critical and more complicated for rare variant analyses (GRM is often recommended)
- Rare variants tend to be more recent mutational events and tend to be more geographically localized than common variants