Rare variant association studies

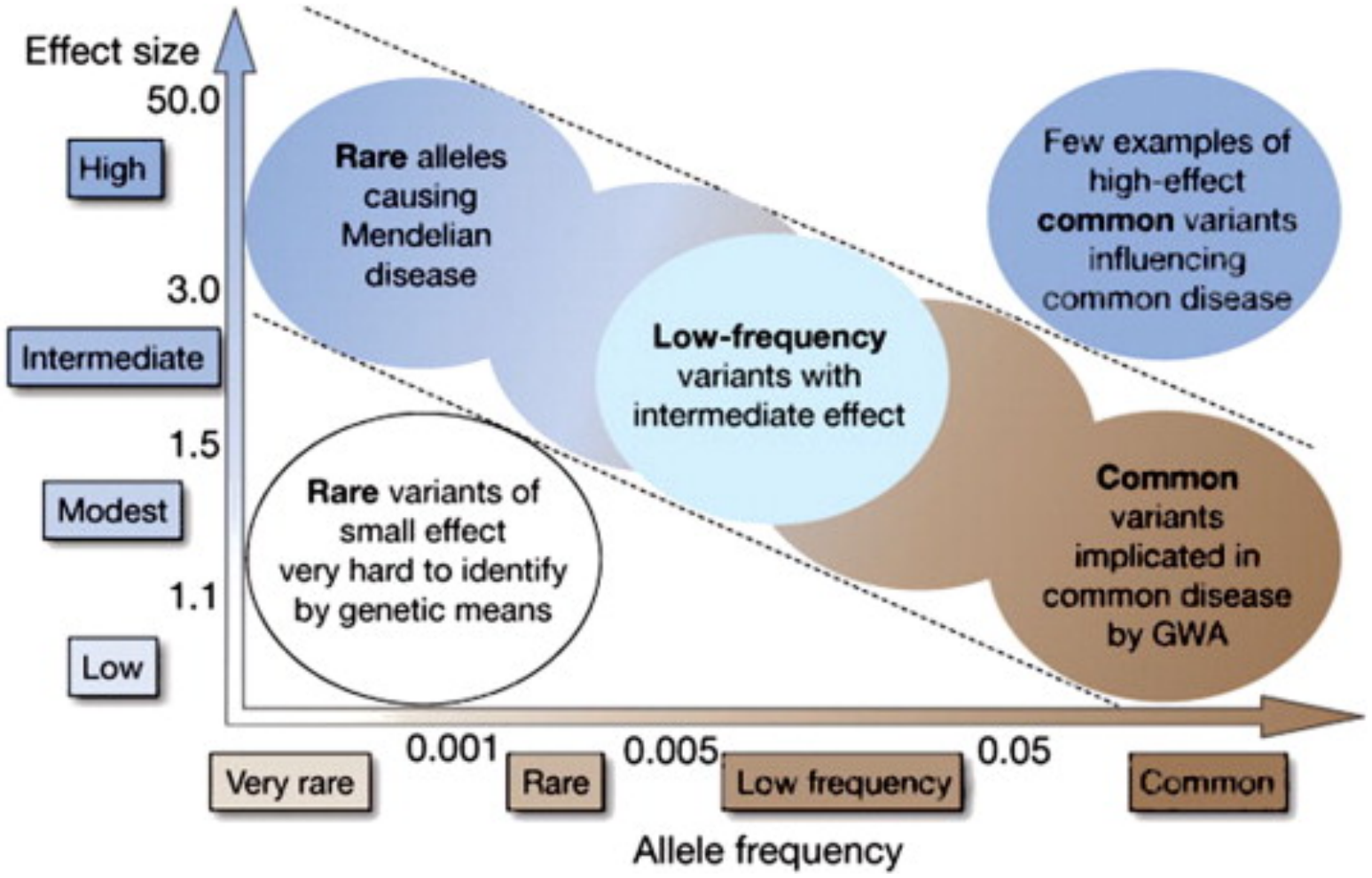
Module evaluation

- Please complete the online evaluation available by logging into your SISG account.

- After you complete the evaluation, you will be able to download your Certificate of Completion through the account.

- http://uwsurvey.qualtrics.com/jfe/form/SV_6A3uZcCSyPxMwfz

Identifying genetic variation associated with disease



Manolio et al, Nature 2009

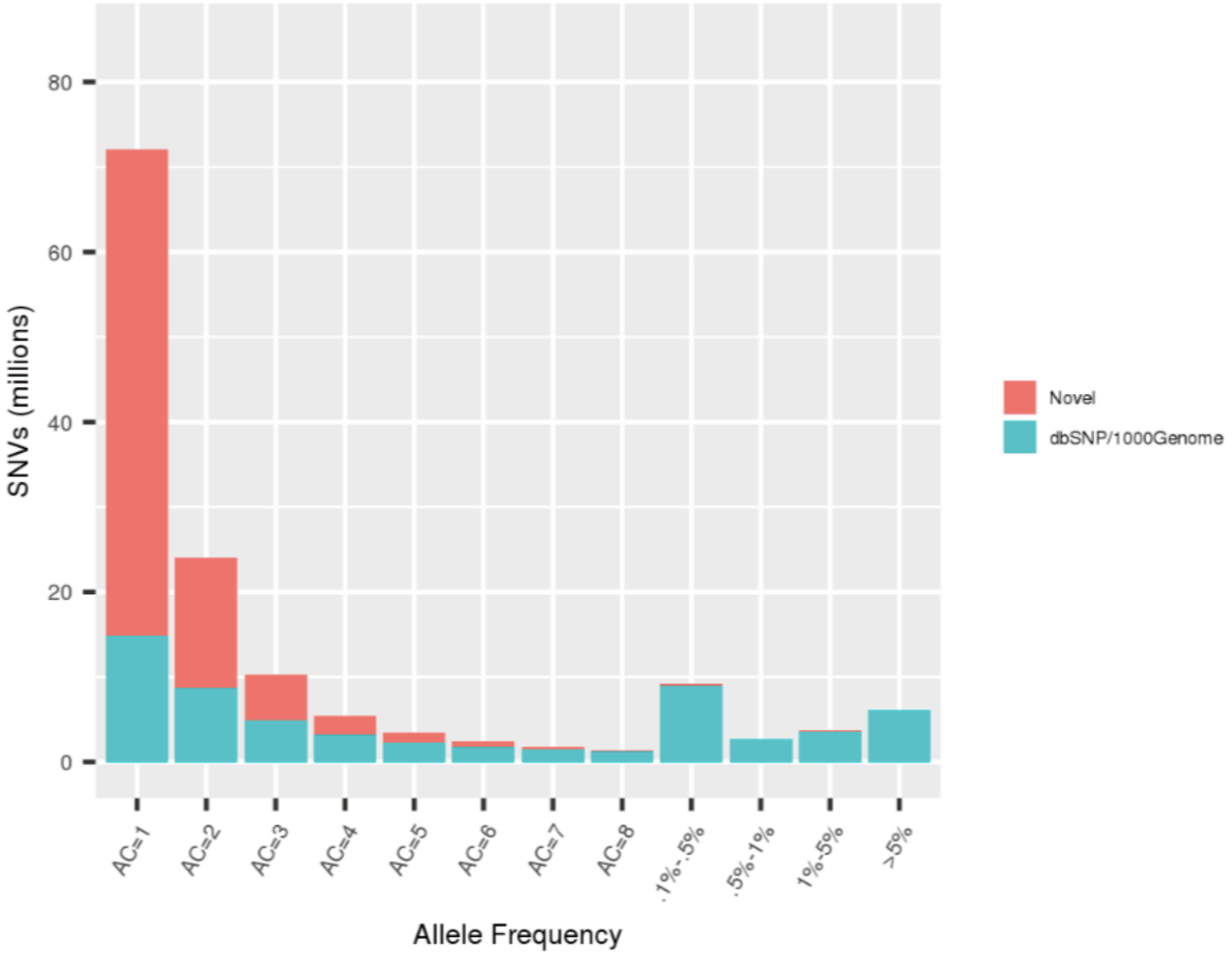
Introduction – Rare variants

- 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 number of rare variants is much larger than the number of 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



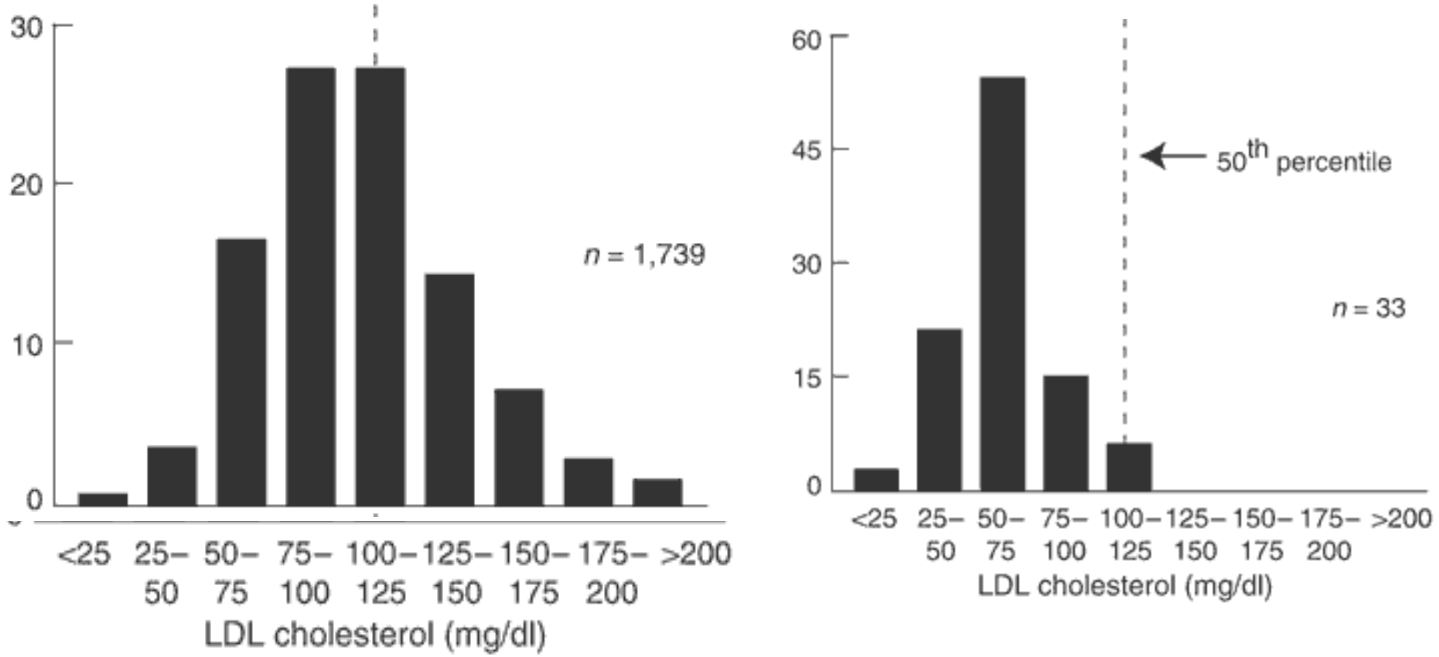
Telenti, PNAS 2016

Poll: Why do we care about rare variants?

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

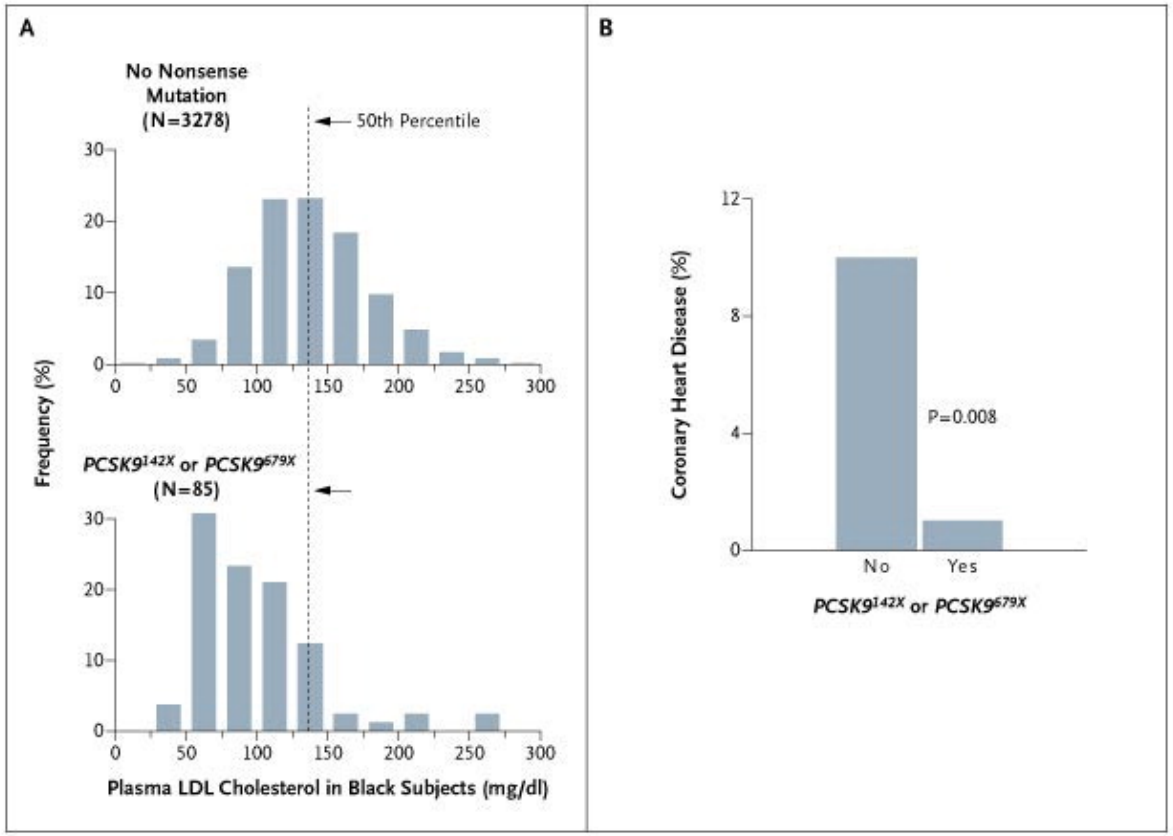
PCSK9 and LDL cholesterol

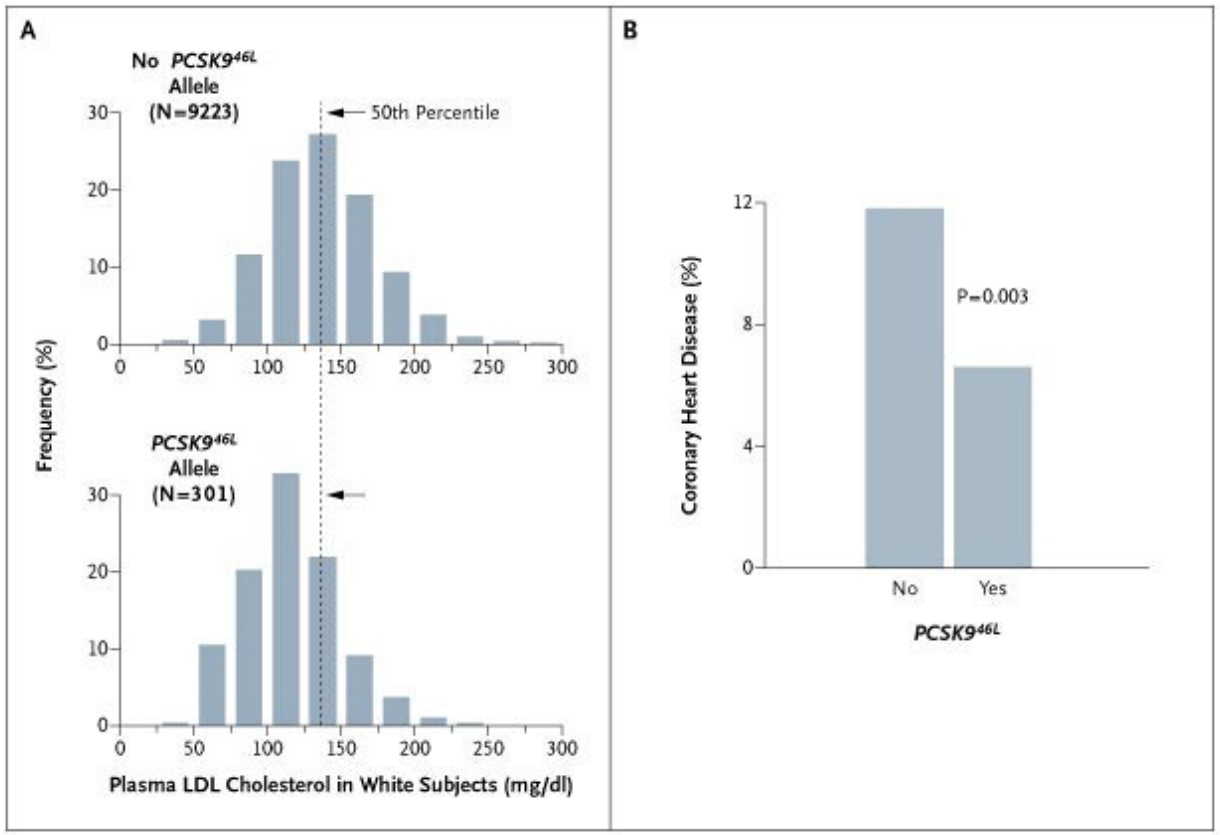
Plasma LDL-C levels in African American subjects without (left) and with (right) a nonsense mutation in *PCSK9*.



Cohen, Nat Genet 2005

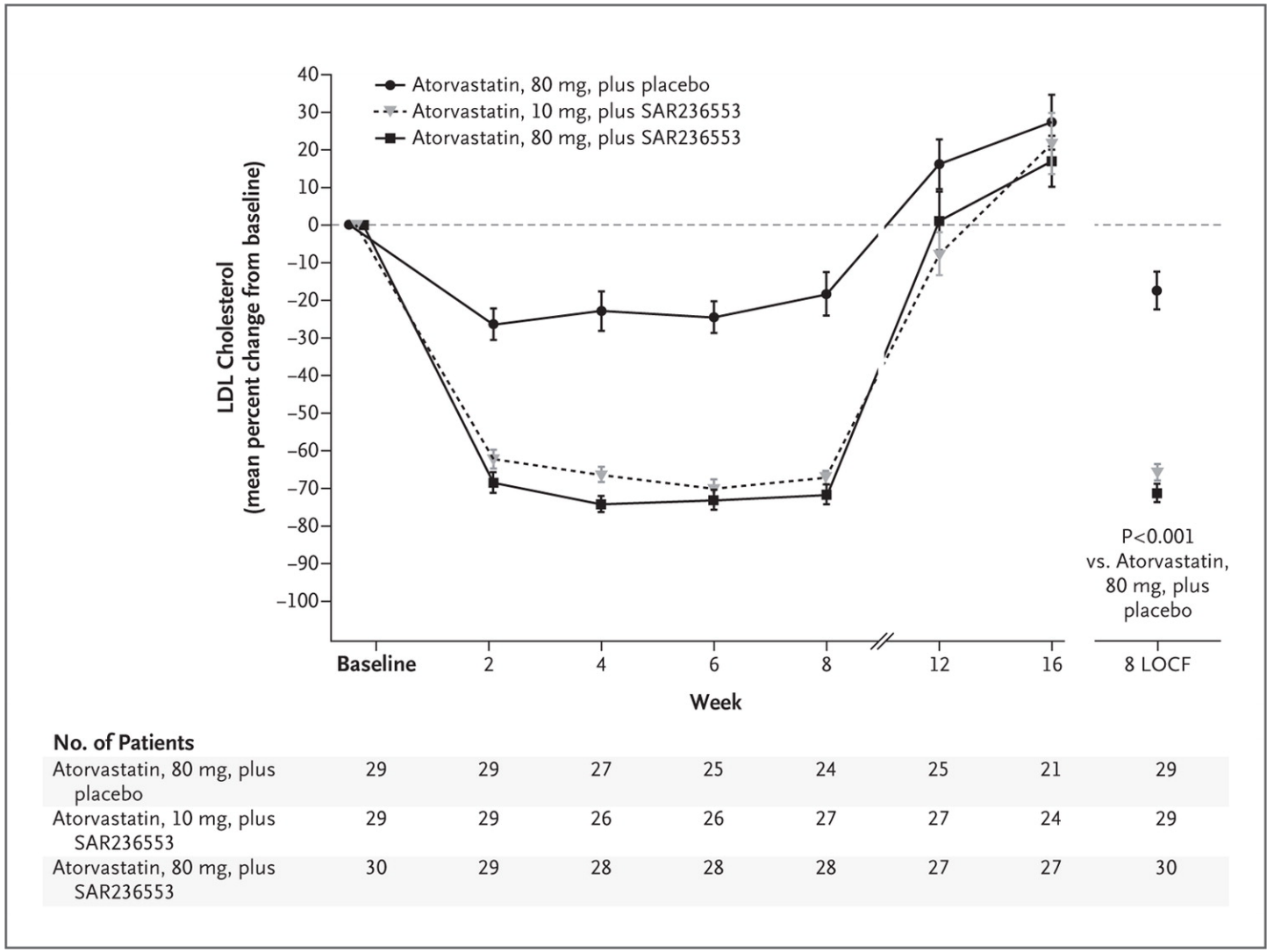
PCSK9 mutations and coronary heart disease





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 have a large, but not unlimited budget. You have colleagues around the world that can give you access to DNA from their breast cancer case/control studies. If you were to design a study to identify rare (allele frequency <1%) variants associated with breast cancer, what are the advantages and disadvantages of each approach? 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

- Limited power/invalid statistical tests due to low cell counts (only a few individuals in your population will carry the minor allele)

- The vast majority of variants in the genome are rare -> increased statistical burden

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

Main Idea: 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**

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.

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

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

Neale, PLoS Genetics 2011

C-alpha test

- 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.

- If there is an association, we typically will observe a higher proportion of doubletons with $y = 2$ and/or $y = 0$ than expected.

- C-alpha can be used to detect a pattern across the full set of rare variants. Under the null hypothesis, $p_i = p_0$. The alternative hypothesis is that p_i follows a mixture distribution across all *variants*, with some variants being detrimental ($p_i > p_0$), some neutral, and some protective ($p_i < p_0$).

Neale, PLoS Genetics 2011

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; $\boldsymbol{\alpha}$ = [α_1,…, α_m]' is the vector of regression coefficients for the *m* covariates; $\boldsymbol{X_i}$: fixed effects of covariates; $\boldsymbol{\beta}$ = [β_1,…,β_p]' is the vector of regression coefficients for the *p* observed gene variants in the region; $\mathbf{G_i}$: (G_{i1}, G_{i2}, …, G_{ip}) genotypes for the *p* variants within the region

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

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?
  - Rare variants are often grouped by gene making 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)

- 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 PCA or linear mixed models 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!