Rare Variant Burden Tests

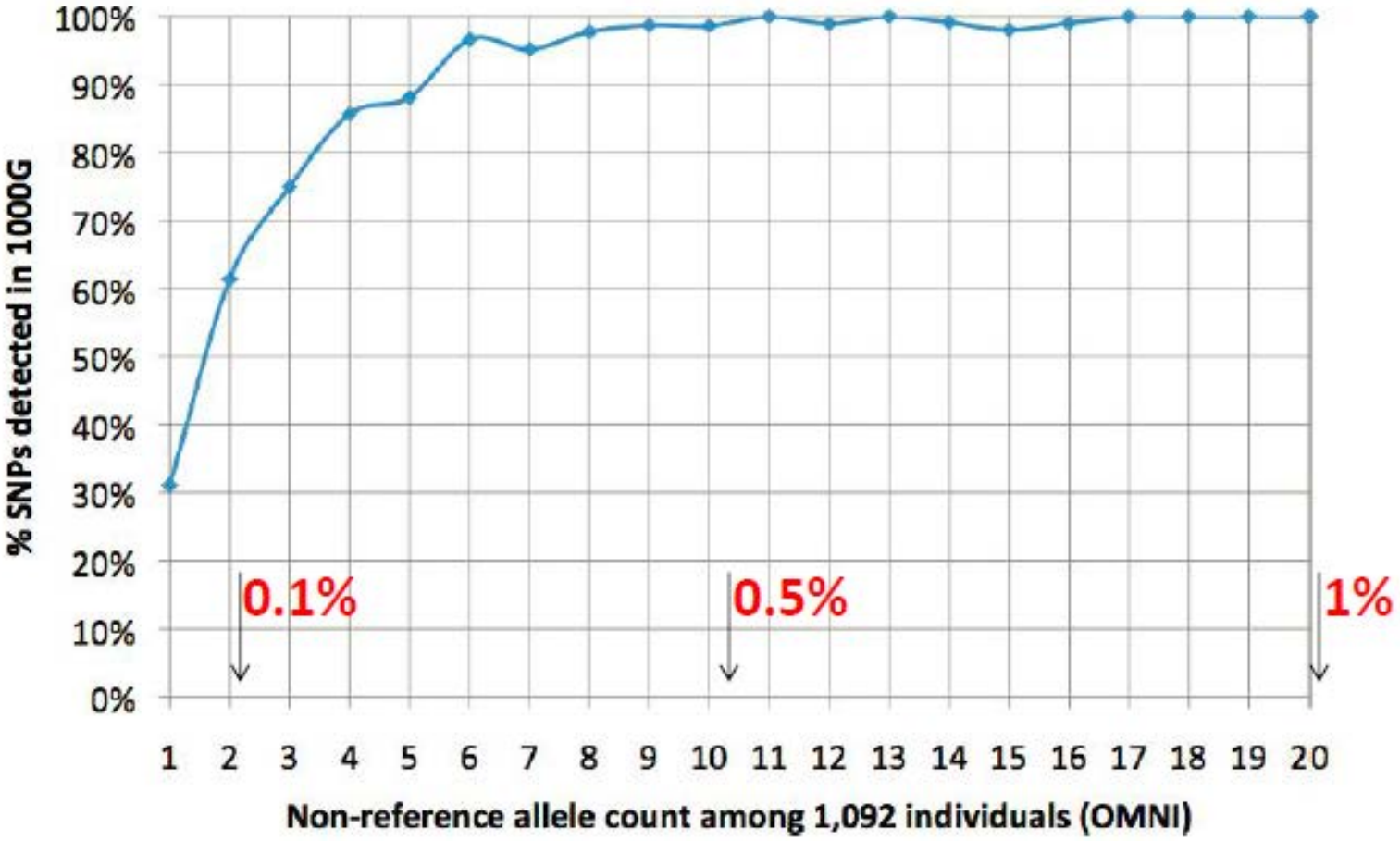
Biostatistics 666

Last Lecture

- Analysis of Short Read Sequence Data
- Low pass sequencing approaches
  - Modeling haplotype sharing between individuals allows accurate variant calls for shared variants
- Assembly Based Analyses
  - Conveniently allow many different types of variation to be analyzed in the same framework

Variants Discovered in Low Pass Analysis

As Function of Allele Frequency



In 1000 Genomes Project Phase I (1094 samples @ 4x), Hyun Min Kang

Today

- Exome Sequencing
- Association Analysis Of Rare Coding Variants
  - Single Variant Analysis
  - Burden Tests
  - Weighted Burden Tests
  - Allowing for Direction of Effect
- Example of an exome sequencing study

Why Study Rare Variants?

COMPLETE GENETIC ARCHITECTURE OF EACH TRAIT

- **Are there additional susceptibility loci to be found?**
- **What is the contribution of each identified locus to a trait?**
  - Sequencing, imputation and new arrays describe variation more fully
  - Rare variants are plentiful and should identify new susceptibility loci

UNDERSTAND FUNCTION LINKING EACH LOCUS TO A TRAIT

- **Do we have new targets for therapy?**
  What happens in gene knockouts?
  - Use sequencing to find rare human “knockout” alleles
  - Good: Results may be more clear than for animal studies
  - Bad: Naturally occurring knockout alleles are extremely rare

Why Study Rare Variants?

COMPLETE GENETIC ARCHITECTURE OF EACH TRAIT

- Are there additional susceptibility loci to be found?
- W

Coding Variants Especially Useful!

lly
ci

UNDERSTAND FUNCTION LINKING EACH LOCUS TO A TRAIT

- **Do we have new targets for therapy?**
  What happens in gene knockouts?
  - Use sequencing to find rare human “knockout” alleles
  - Good: Results may be more clear than for animal studies
  - Bad: Naturally occurring knockout alleles are extremely rare

The Scale of Rare Variation

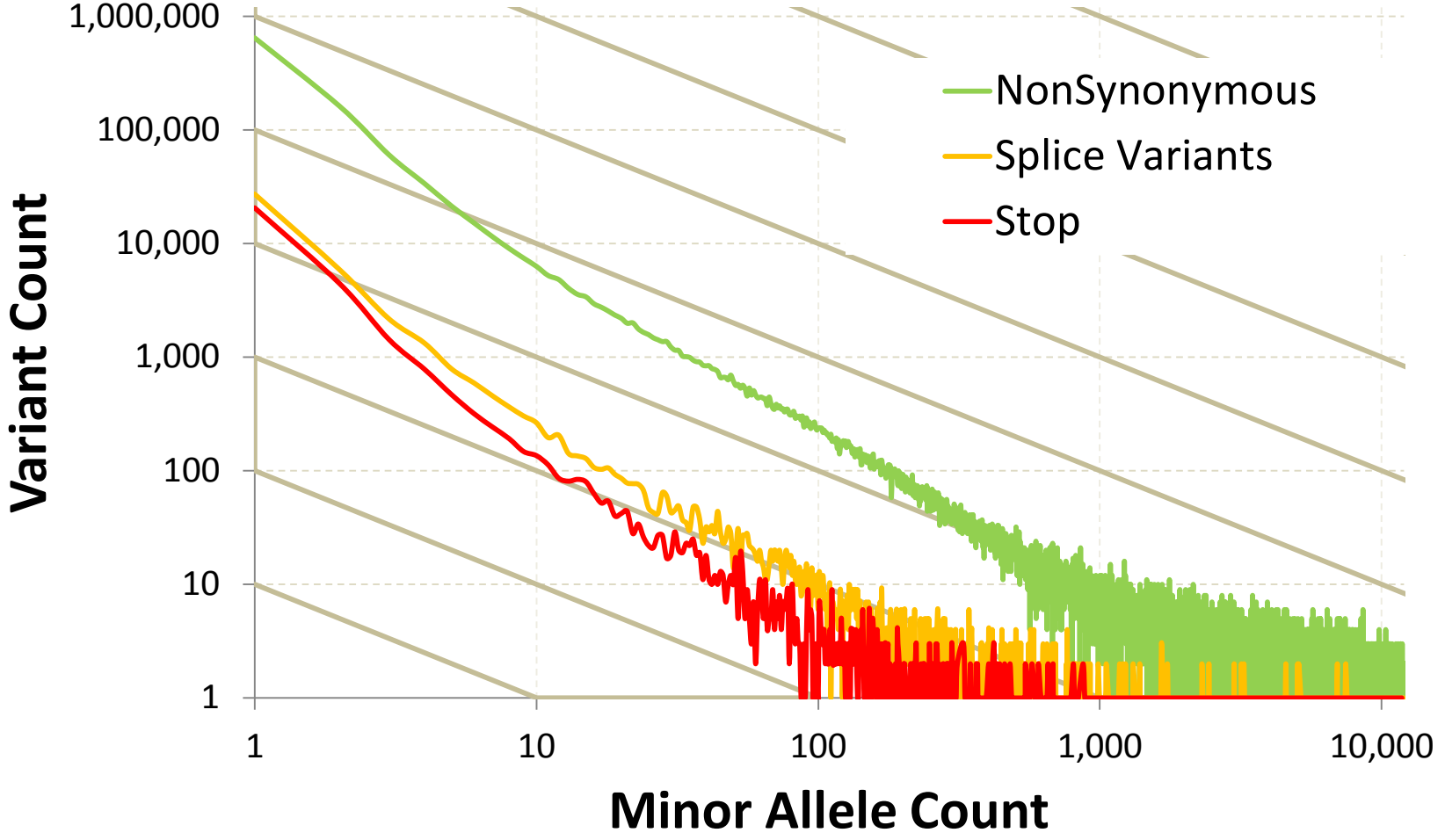
Lots of Rare Functional Variants to Discover

SET	# SNPs	Singletons	Doubletons	Tripletons	>3 Occurrences
Synonymous	270,263	128,319 (47%)	29,340 (11%)	13,129 (5%)	99,475 (37%)
Nonsynonymous	410,956	234,633 (57%)	46,740 (11%)	19,274 (5%)	110,309 (27%)
Nonsense	8,913	6,196 (70%)	926 (10%)	326 (4%)	1,465 (16%)
Non-Syn / Syn Ratio		1.8 to 1	1.6 to 1	1.4 to 1	1.1 to 1

There is a very large reservoir of extremely rare, likely functional, coding variants.

NHLBI Exome Sequencing Project

Allele Frequency Spectrum (After Sequencing 12,000+ Individuals)



http://genome.sph.umich.edu/wiki/Exome_Chip_Design

How Much Variation Might Rare Variants Explain?

- All variation neutral, population size constant
  - MAF<0.1% variants explain 0.2% of heritability
  - MAF<1.0% variants explain 2.0% of heritability
  - MAF<5.0% variants explain 10% of heritability

- Nonsynonymous frequency spectrum from 12,000 exomes
  - MAF<0.1% variants explain 3.6% of heritabilty
  - MAF<1.0% variants explain 10.6% of heritability
  - MAF<5.0% variants explain 22.7% of heritability

- Assuming rare variants effect sizes are ~2x larger on average
  - Above estimates increase to about 8.6, 25.4 and 54.0%

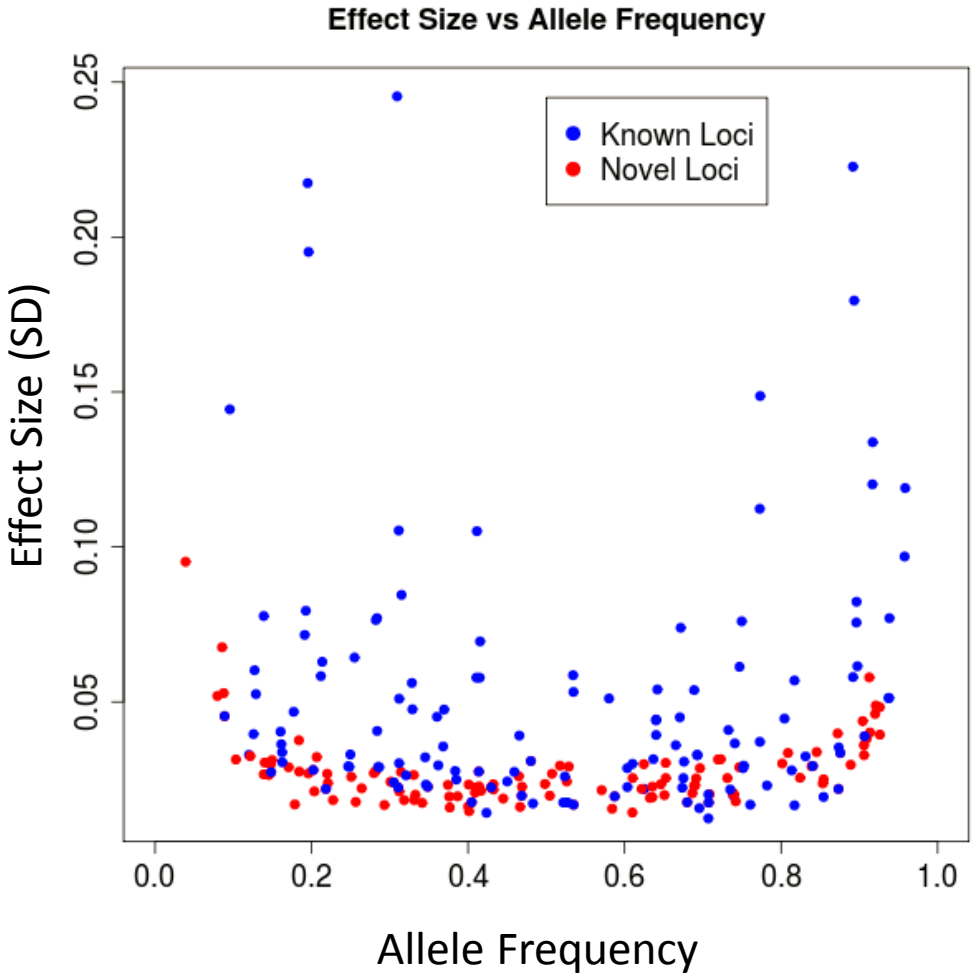
- Assuming rare variants effect sizes are ~3x larger on average
  - Above estimates increase to about 11.6, 34.1 and 72.6%

Do Rare Variants Have Large Effects?

- The main driver is natural selection
- Most variants that impact function are expected to be deleterious
  - Natural selection will prevent them from becoming common
- Good evidence that non-synonymous variants are depleted among common variant lists

Rare Variants Have Large Effects More Often

Lipid Associated Variants in 200,000 individuals



Results from analysis of >190,000 individuals

Sengupta et al (unpublished)

Genome Scale Approaches To Study Rare Variation

- **Deep whole genome sequencing**
  - Can only be applied to limited numbers of samples
  - Most complete ascertainment of variation

- **Exome capture and targeted sequencing**
  - Can be applied to moderate numbers of samples
  - SNPs and indels in the most interesting 1% of the genome

- **Low coverage whole genome sequencing**
  - Can be applied to moderate numbers of samples
  - Very complete ascertainment of shared variation

- **New Genotyping Arrays and/or Genotype Imputation**
  - Examine low frequency coding variants in 100,000s of samples
  - Current catalogs include 97-98% of sites detectable by sequencing an individual

Genome Scale Approaches To Study Rare Variation

- **Deep whole genome sequencing**
  - Can only be applied to limited numbers of samples
  - Most complete ascertainment of variation

- **Exome capture and targeted sequencing**
  - Can be applied to moderate numbers of samples
  - SNPs and indels in the most interesting 1% of the genome

- **Low coverage wh ome sequencing**
  - Can
  - Ver

- **New G**
  - Examine low frequency coding variants in 100,000s of samples
  - Current catalogs include 97-98% of sites detectable by sequencing an individual

Our Focus For Today

SNPs Per Individual

Primarily European Ancestry

European Ancestry	# SNP	# HET	# ALT	# Singletons	Ts/Tv
SILENT	10127	6174	3953	38.2	5.10
MISSENSE	8541	5184	3357	72.2	2.16
NONSENSE	86	57	29	2.1	1.70

Primarily African Ancestry

African Ancestry	# SNP	# HET	# ALT	# Singletons	Ts/Tv
SILENT	12028	8038	3990	53.2	5.19
MISSENSE	9870	6502	3367	94.2	2.16
NONSENSE	92	57	35	2.4	1.57

Rare Variant Association Testing

- Consider variant with frequency of ~0.001
- Significance level of $5x10^{-6}$
  - Corresponds to ~100,000 independent tests
- Disease prevalence of ~10%
- Detecting a two-fold increase in risk, requires ~33,000 cases and ~33,000 controls!
- Detecting a three-fold increase in risk requires ~11,000 cases and ~11,000 controls!

Rare Variant Association Testing

- Consider variant with frequency of ~0.001
- Significance level of 5×10^{-6}
  - Corresponds to ~100,000 independent tests
- Disease prevalence of ~10%
- Detecting a two-fold increase in risk, requires ~33,000 cases and ~33,000 controls!
- Detecting a three-fold increase in risk requires ~11,000 cases and ~11,000 controls!

Power Depends Both On:

Frequency

Effect Size

Even with large effects, rare variants can only be detected in large samples

Alternatives to Single Variant Tests

Collapsing Rare Variants

- Instead of testing rare variants individually, group variants likely to have similar function
- Score presence or absence of rare variants per individual
  - Use rare variant score to predict trait values
- If all variants are causal, leads to large increase in power
- In practice, success depends on:
  - Number of associated variants,
  - Number of neutral variants diluting signals
  - Whether direction of effect is consistent within gene

Li and Leal (2008) *Am J Hum Genet* **83:**311-321

Burden vs. Single Variant Tests

	Single Variant Test	Combined Test
10 variants / all have risk 2 / All have frequency .005	.05	.86
10 variants / all have risk 2 / Unequal Frequencies	.20	.85
10 variants / average risk is 2, but varies / frequency .005	.11	.97

- Power tabulated in collections of simulated data, for 250 cases and 250 controls
- Combining variants can greatly increase power
- Currently, appropriately combining variants is expected to be key feature of rare variant studies.

Li and Leal (2008) *Am J Hum Genet* **83:**311-321

Impact of Null Alleles

	Single Variant Test	Combined Test
10 disease associated variants	.05	.86
10 disease associated variants + 5 null variants	.04	.70
10 disease associated variants + 10 null variants	.03	.55
10 disease associated variants + 20 null variants	.03	.33

- Power tabulated in collections of simulated data
- Including non-disease variants reduces power
- Power loss is manageable, combined test remains preferable to single marker tests

Li and Leal (2008) *Am J Hum Genet* **83:**311-321

Impact of Missing Disease Alleles

	Single Variant Test	Combined Test
10 disease associated variants	.05	.86
10 disease associated variants, 2 missed	.05	.72
10 disease associated variants , 4 missed	.05	.52
10 disease associated variants , 6 missed	.04	.28
10 disease associated variants, 8 missed	.03	.08

- Power tabulated in collections of simulated data
- Missing disease associated variants loses power

Li and Leal (2008) *Am J Hum Genet* **83:**311-321

Refining Rare Variant Tests

- The original Li and Leal (2008) test simply “collapses” rare variants into one allele

- Multiple refinements have been proposed since…
  - Counting the number of rare variants per individual
  - Weighting rare variants according to frequency
  - Weighting rare variants according to function
  - Including imputed variants in the analysis

- Each of these methods may improve power, but few practical examples provide guidance

CMAT: Combined Minor Allele Test

Consider gene with k variants in sample of N cases and N controls.

For polymorphism i define:

- w_i, a weight based on functional annotation, minor allele frequency, imputation accuracy
- g_{ij}, the expected posterior minor allele count in individual j.
- Set $m_A = \sum_{i=1}^{k} w_i \sum_{j=case} g_{ij}$ $M_A = \sum_{i=1}^{k} w_i \sum_{j=case} (2 - g_{ij})$

The test statistic is then $\sum_{CMAT} = \frac{m_A M_U - m_U M_A}{N(m_A + m_U)(M_A + M_U)}$

Significance of the test statistic evaluated by permutation of affection status.

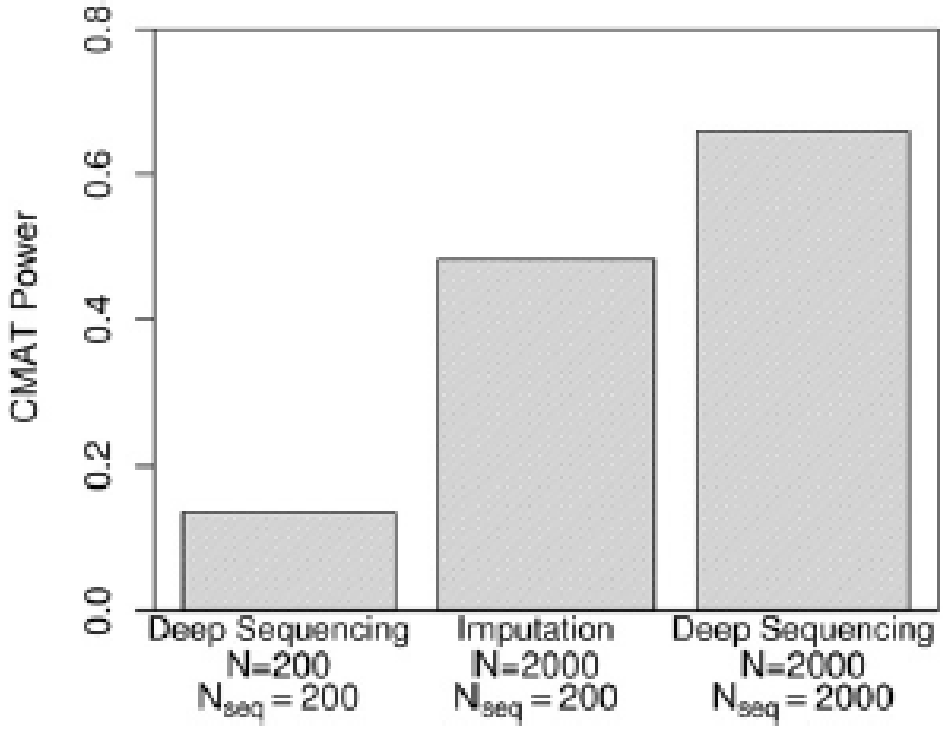
Zawistowski et al (2010)

Weights

- Use computational algorithms to prioritize functional variants
  - Based on conservation
  - Based on biochemical properties

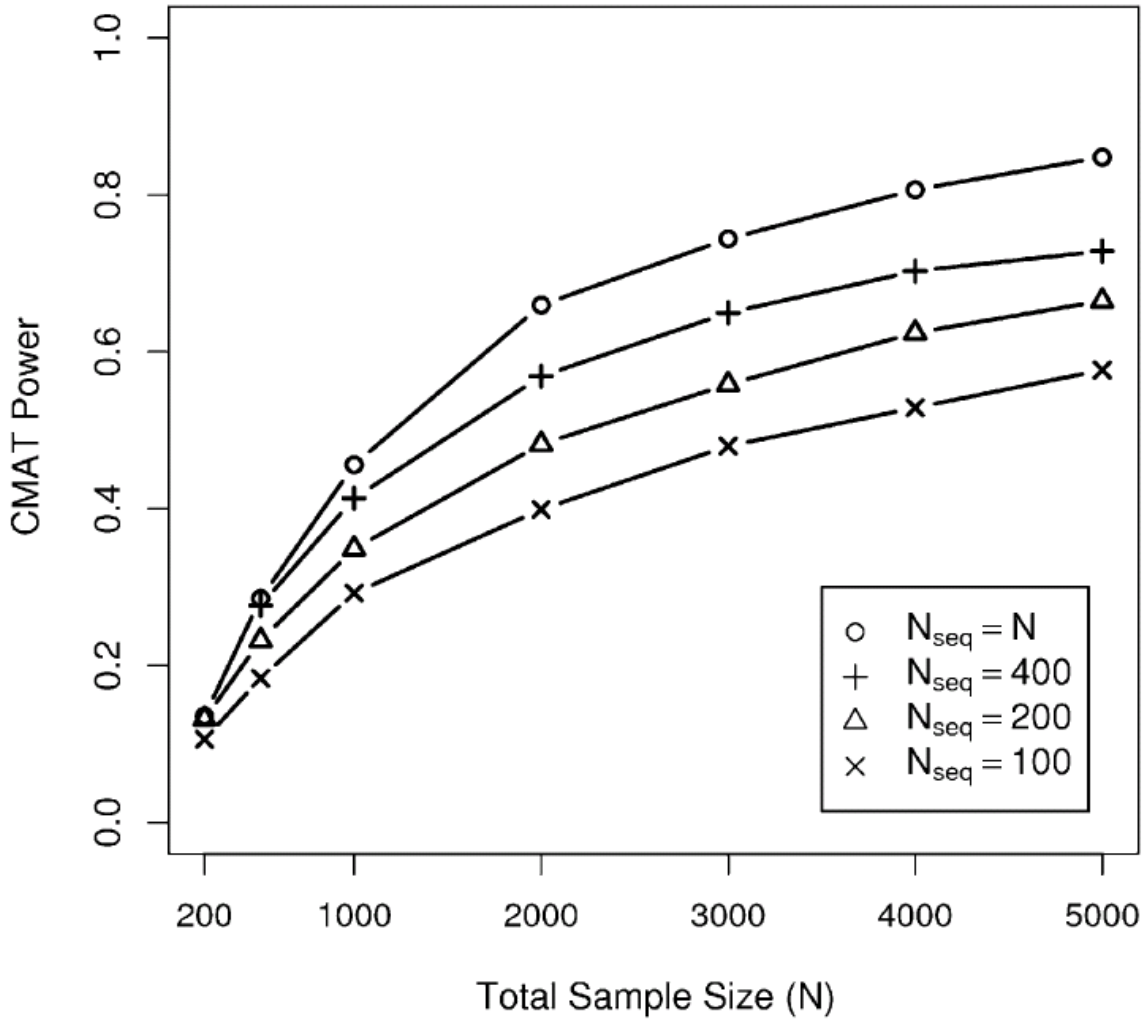
- Frequency is an independent predictor of functional consequence.

Imputation in Rare Variant Burden Tests



Zawistowski et al (2010)

Power as a Function of No. of Sequenced and Genotyped Samples



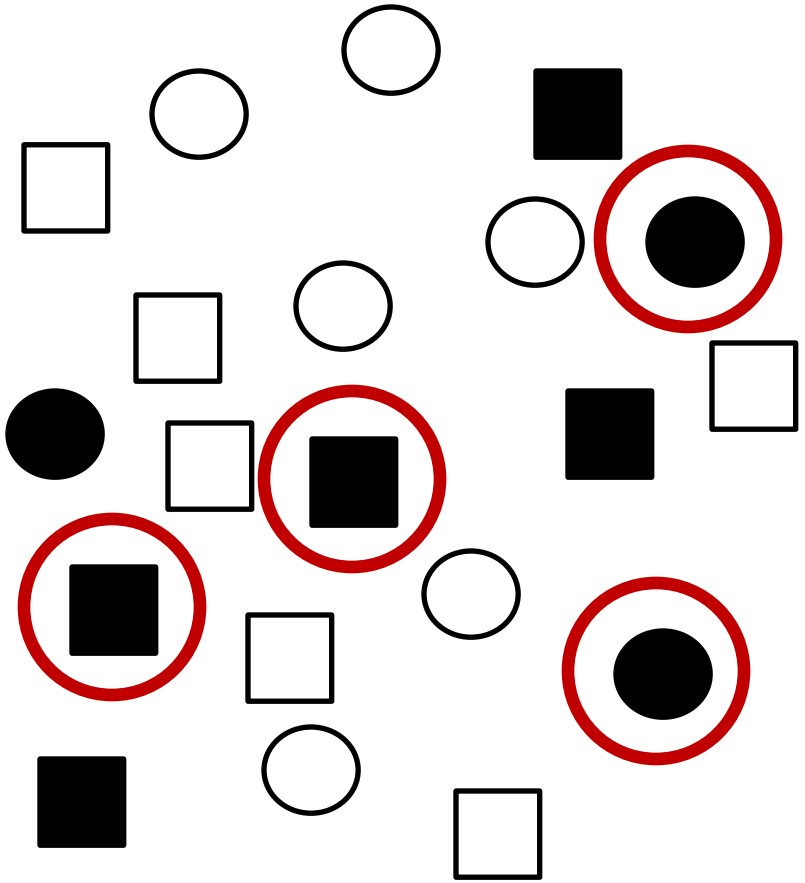
Zawistowski et al (2010)

Maximizing the Power

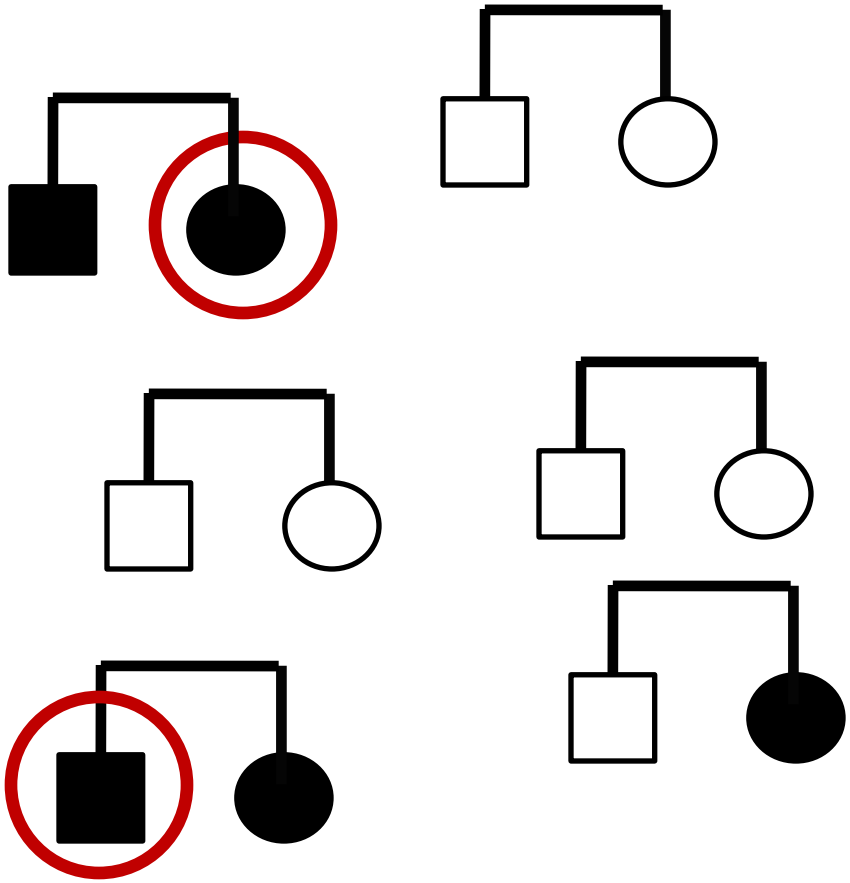
- Power depends on summed frequency – choose threshold for defining rare carefully.
- Enriching functional variants in cases increases power – perhaps by focusing on loss of function variants only.
- For quantitative traits, focus on individuals with extreme trait values.
- For discrete traits, focus on individuals with family history of disease.

Enriching based on familial risk

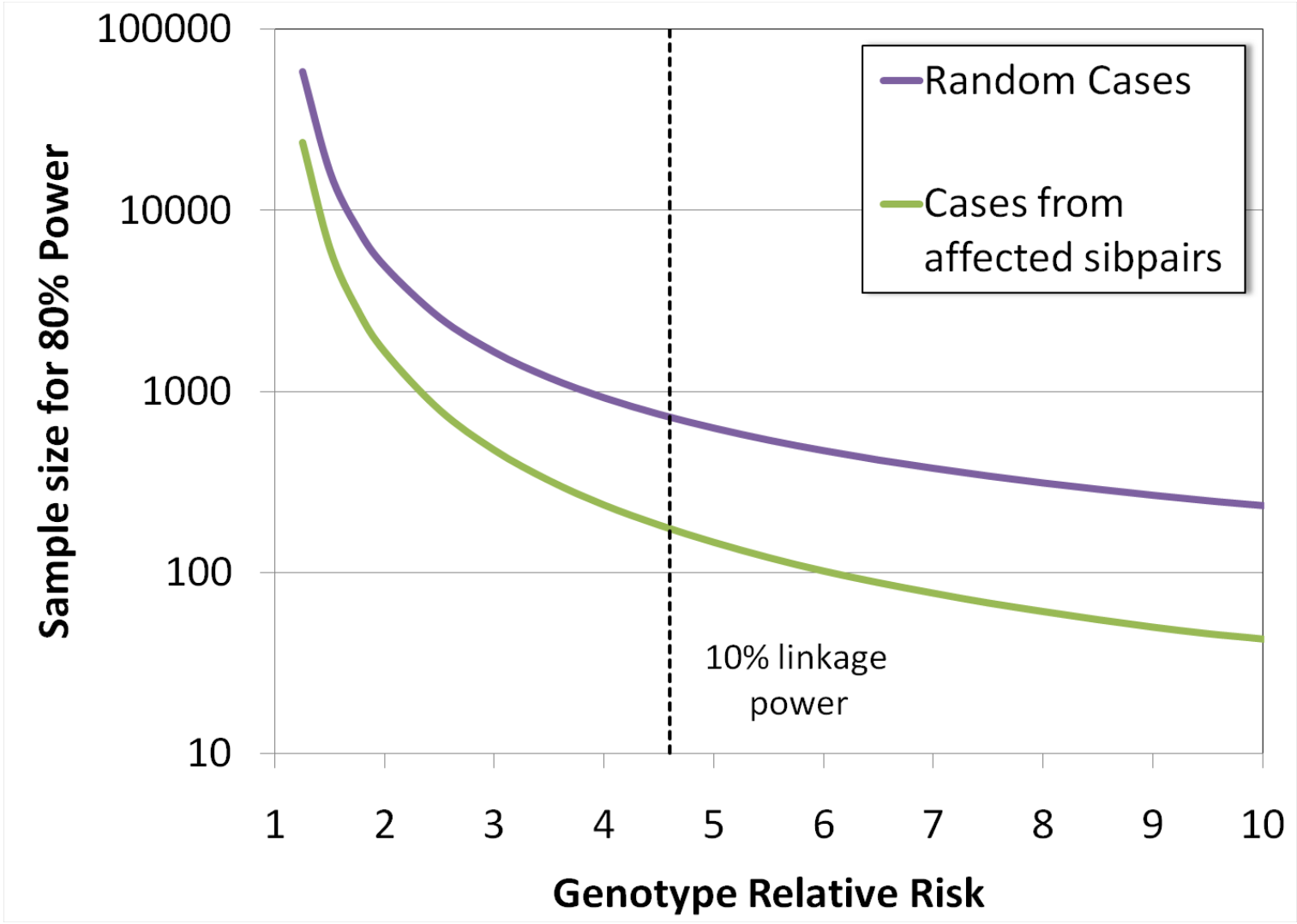
Classic Case-Control



Familial enrichment



Benefits of Favoring Family History of Disease



Practical Example: Exome Sequencing and Burden Tests

NHLBI Exome Sequencing Project
University of Washington and Broad Institute

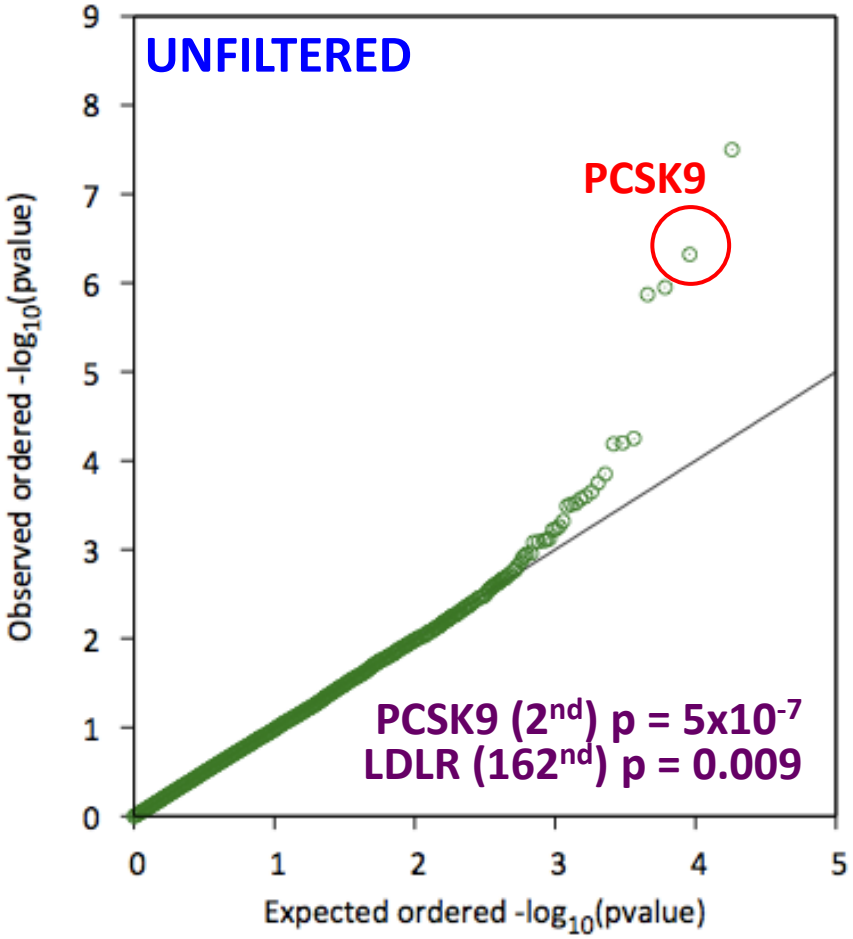
Cristen Willer and Leslie Lange

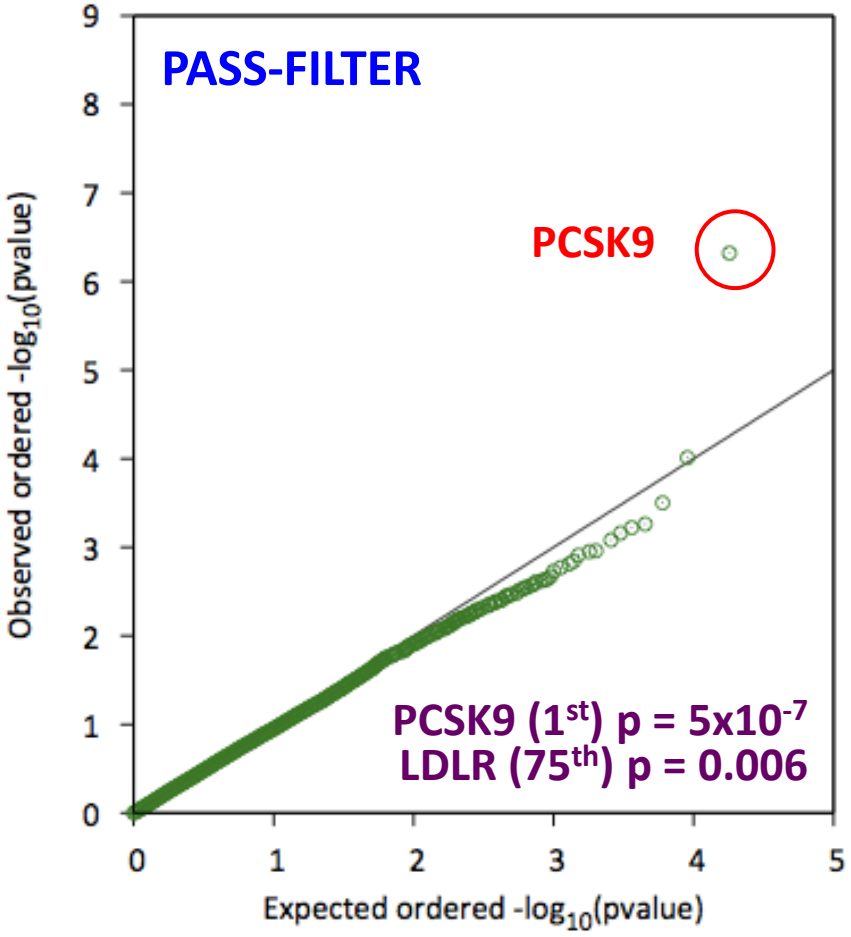
Exome Sequencing Project

- The NHLBI Exome Sequencing Project is studying heart, lung and blood related traits
- One of the traits of interest is LDL, a major risk factor for cardiovascular disease
- Let's review their preliminary findings, in analysis of …
  - 400 selected from top and bottom 2% of population
  - 1,600 individuals selected without consideration of LDL

LDL Results – Burden Test, MAF < 5%

(logistic regression adjusted by PC1, PC2, age, gender, center)

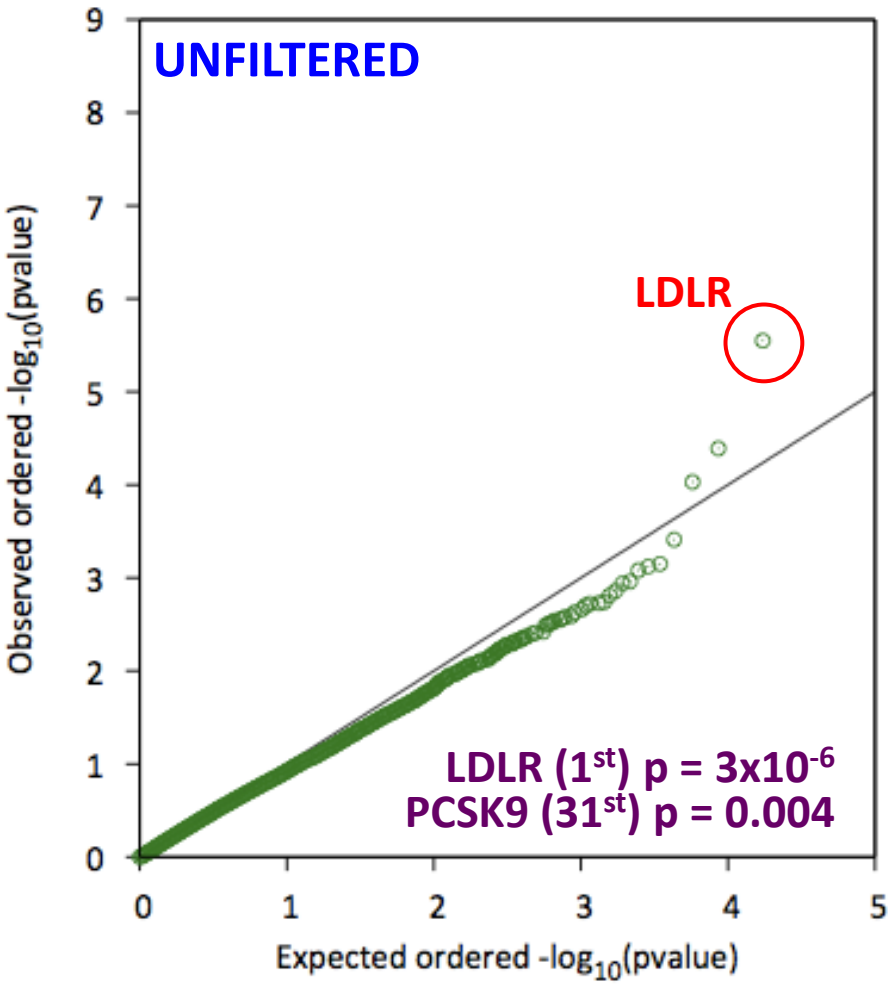


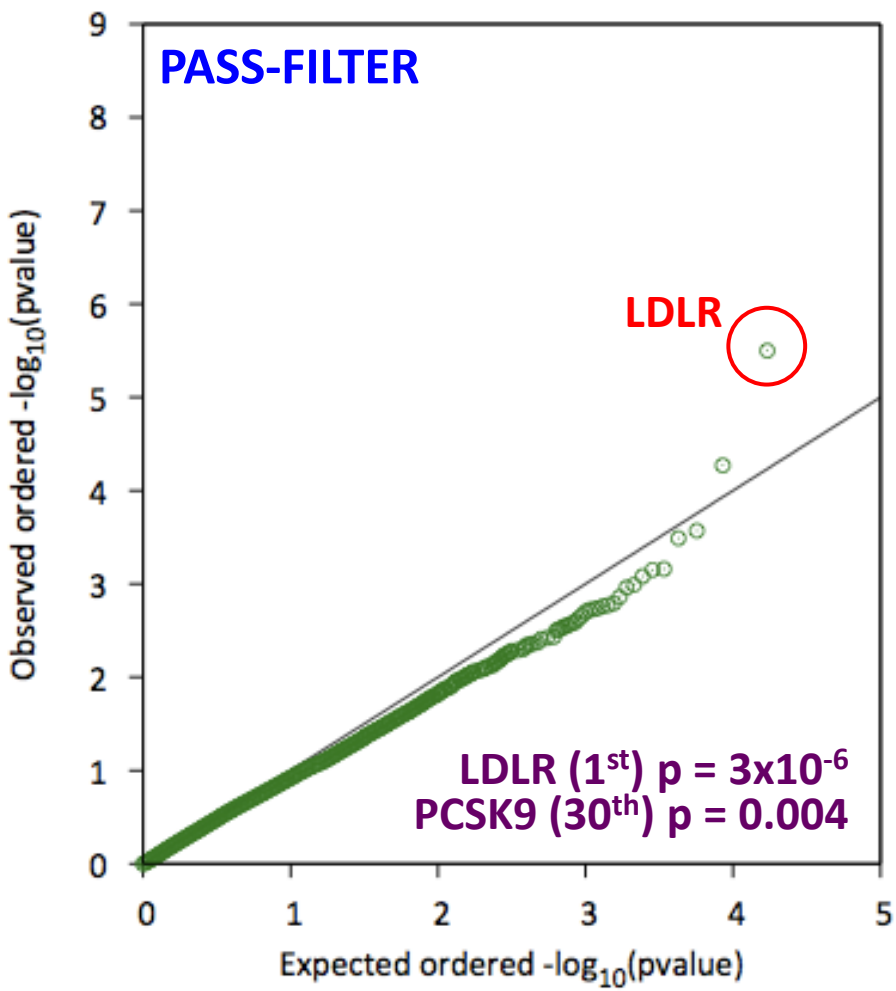


Cristen Willer and Leslie Lange, NHLBI Exome Sequencing Project

LDL Results – Burden Test, MAF < 0.1%

(logistic regression adjusted by PC1, PC2, age, gender, center)

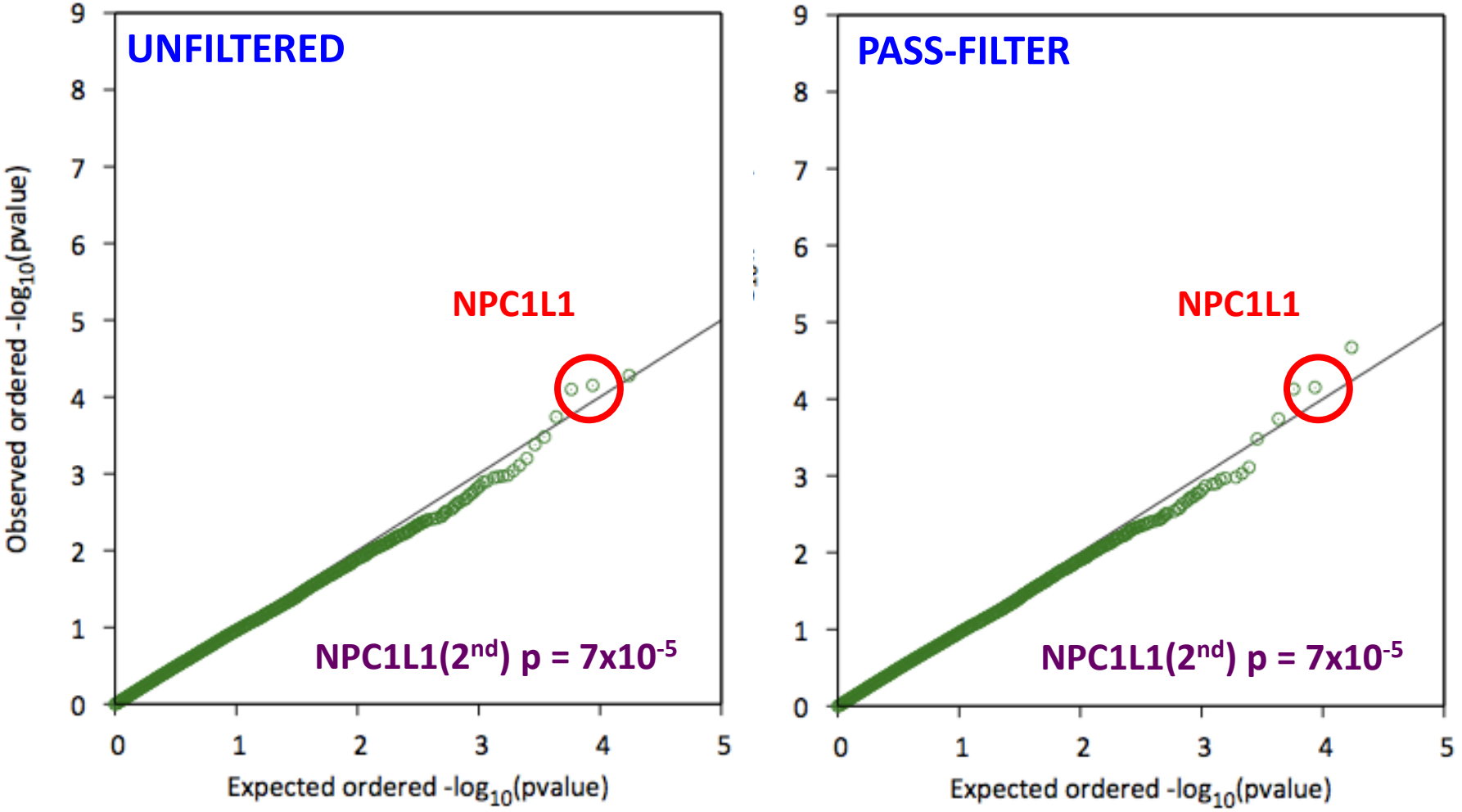




Cristen Willer and Leslie Lange, NHLBI Exome Sequencing Project

LDL Results – Burden Test, MAF < 0.5%

(logistic regression adjusted by PC1, PC2, age, gender, center)



Cristen Willer and Leslie Lange, NHLBI Exome Sequencing Project

Variable Threshold Tests

- Different definitions of “rare” lead to different signals
- Conducting multiple analyses quickly becomes hard to manage
- What to do?
- Variable threshold tests consider all possible thresholds for each gene and search for maximum test statistic
  - Evaluate significance by permutation

Price et al (2010) AJHG **86**:832-838

Variable Threshold Tests

- Price et al (2010) originally suggested using permutations for evaluating significance of variable threshold association tests
- Lin and Tang (2011) showed that statistics using different thresholds could be described using a multivariate normal distribution…
- … allowing for p-value calculation without permutations.

Lin and Tang (2011) AJHG **89**:354-367

Additional Complications!

- What to do if a gene includes some rare alleles that increase risk, others that decrease it?
- What sort of signal do you expect?
- What sort of strategies might identify these signals?

ARTICLE

Extending Rare-Variant Testing Strategies: Analysis of Noncoding Sequence and Imputed Genotypes

Matthew Za
and Sebasti

ARTICLE

Pooled Association Tests for Rare Variants in Exon-Resequencing Studies

Alkes L. Price,[1,2,3,6] Gregory V. Kryukov,[3,4,6] Paul I.W. de Bakker,[3,4] Shaun M. Purcell,[3,5] Jeff Staples,[3,4] Lee-Jen Wei,[2] and Shamil R. Sunyaev[3,4,*]

PLoS GENETICS

A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic

Bo Eskerod Mac

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

A Covering Method for Detecting Genetic Associations between Rare Variants and Common Phenotypes

Gaurav Bhatia[1,2*], Vikas B
Vineet Bafna[1,5]

OPEN ACCESS Freely available online

PLoS GENETICS

A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions

Dajiang J. Liu[1,2], Suzanne M. Leal[1,2*]

Analysing biological pathways in genome-wide association studies

nature

REVIEWS

Kai Wang[*1], *Mingyao Li*[§] *and Hakon Hakonarson*[*||]

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ran
Augustine Kong[11], Leonid Krug
Alice S. Whittemore[16], Michael
Trudy F. C. Mackay[22], Steven A

Hum Genet (2010) 128:627–633
DOI 10.1007/s00439-010-0889-1

ORIGINAL INVESTIGATION

Rare variation at the *TNFAIP3* locus and susceptibility to rheumatoid arthritis

John Bowes ·
Gisela Oroze
UKRAG · W

Annals of human genetics

doi: 10.1111/j.1469-1809.2010.00566.x

Common Susceptibility Variants Examined for Association with Dilated Cardiomyopathy

Evadnie Rampersaud[1*], Daniel D. Kinnamon[1*], Kara Hamilton[1], Sawsan Khuri[2], Ray E. Hershberger[3] and Eden R. Martin[1]

Summary

- Analysis of individual rare variants requires very large samples.
- Power may be increased substantially by combining information across variants.
  - Strategy for combining information across variants allows for many tweaks.
- This is an extremely active research area.

Recommended Reading

- Li and Leal (2008) *Am J Hum Genet* **83:**311-321

- Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S (2010) *Am J Hum Genet* **87**:604-617