Analytic Issues Associated with Rare Variant Testing

Benjamin Neale

March 10th 2011

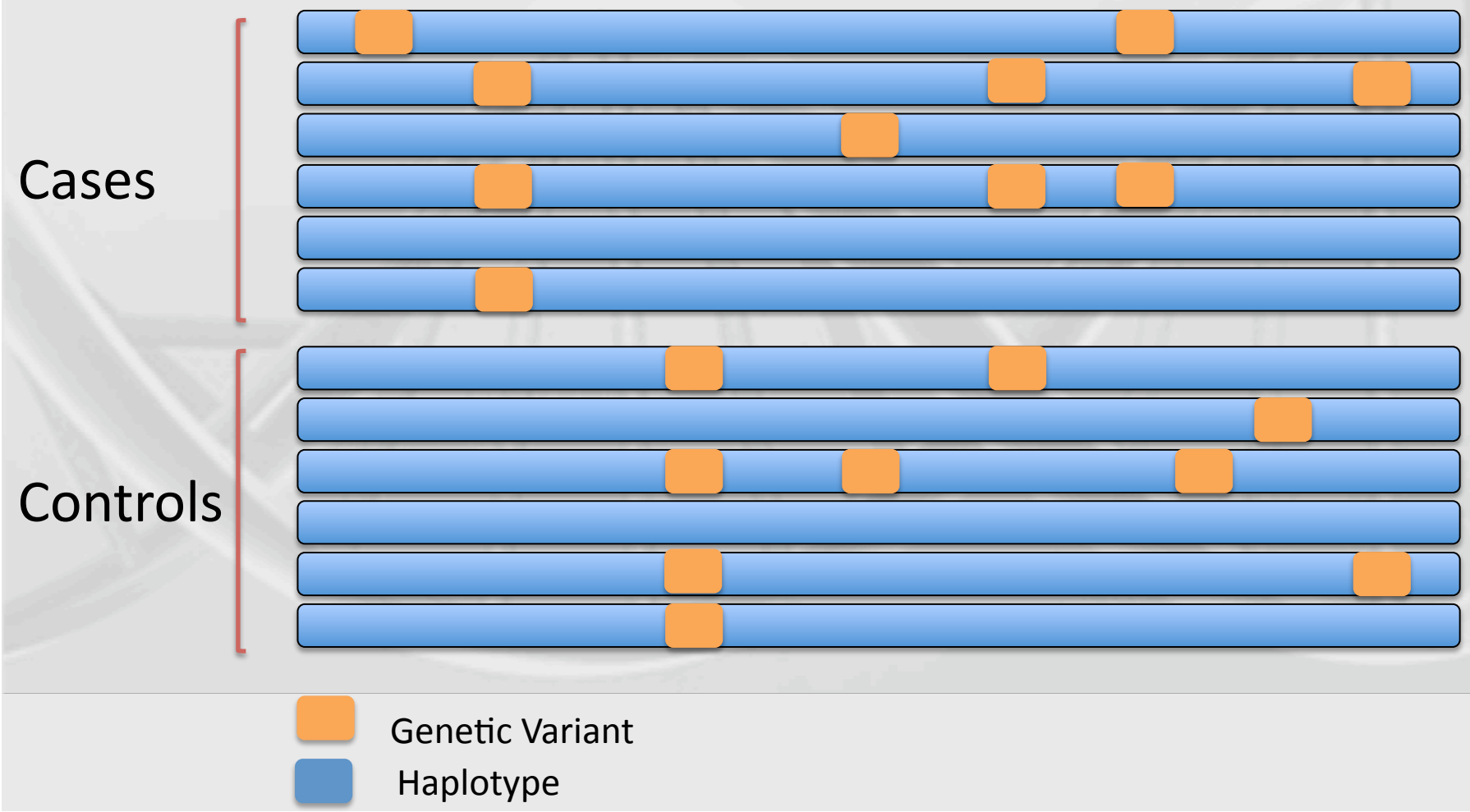
Boulder Workshop

Massachusetts General Hospital
Harvard Medical School
BROAD INSTITUTE

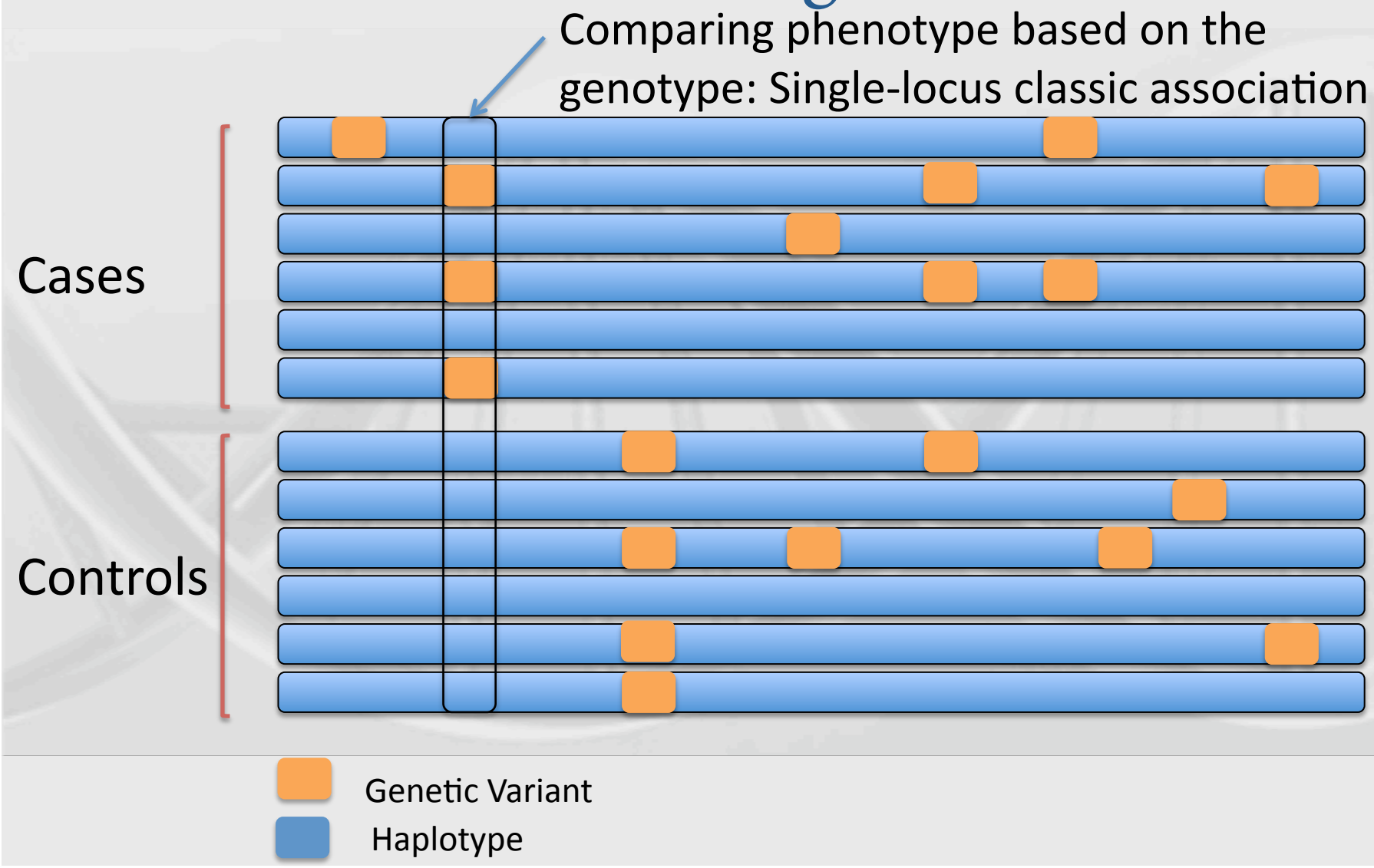
Slide of Contents

- Single Locus Testing
  - Difficulties/asymptotics
- Field-wide development of analytic approaches for sequence data
  - Cohort Allelic Sum Test (CAST; Hobbs, Cohen and others)
  - Li and Leal (AJHG)
  - Madsen and Browning (PLoS Genetics)
- C-alpha and sequence analysis
- Power comparisons
- Extensions

Visual representation of sequence data testing



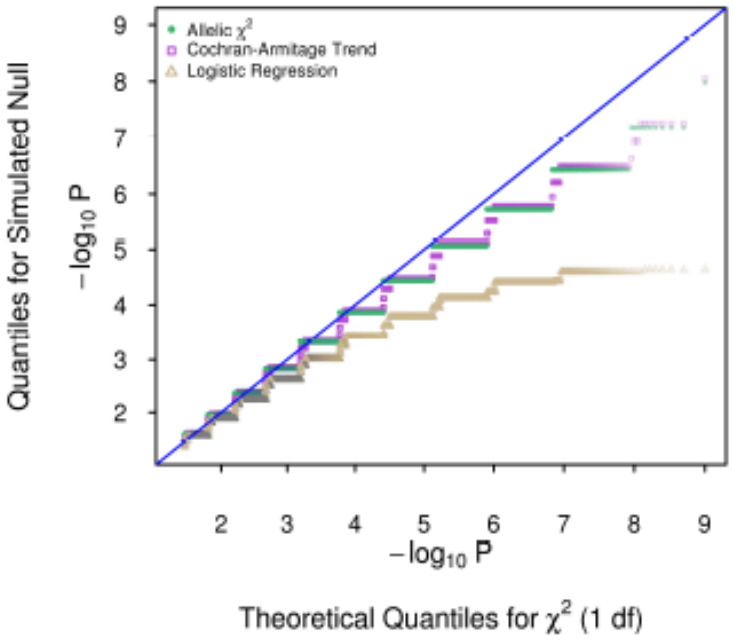
Visual representation of sequence data testing



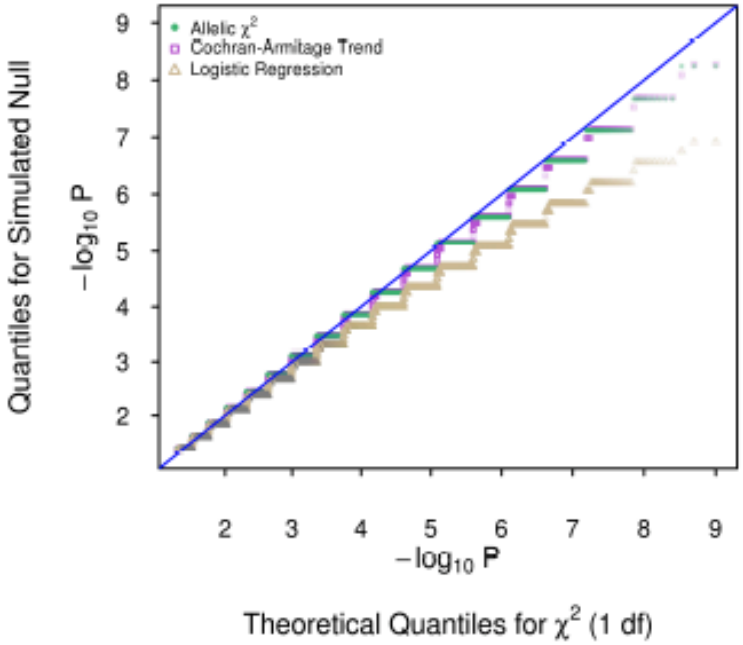
Single Locus Association

- As we've seen, many different tests can be used for association
- With few observations, the maximum significance achievable can be quite large
- If I have a doubleton, even with 10,000 cases and 10,000 controls my best two-sided P-value is 0.5
- We can explore this a bit further

Difficulties of common tests



(a) 40 minor alleles in $N = 2,000$

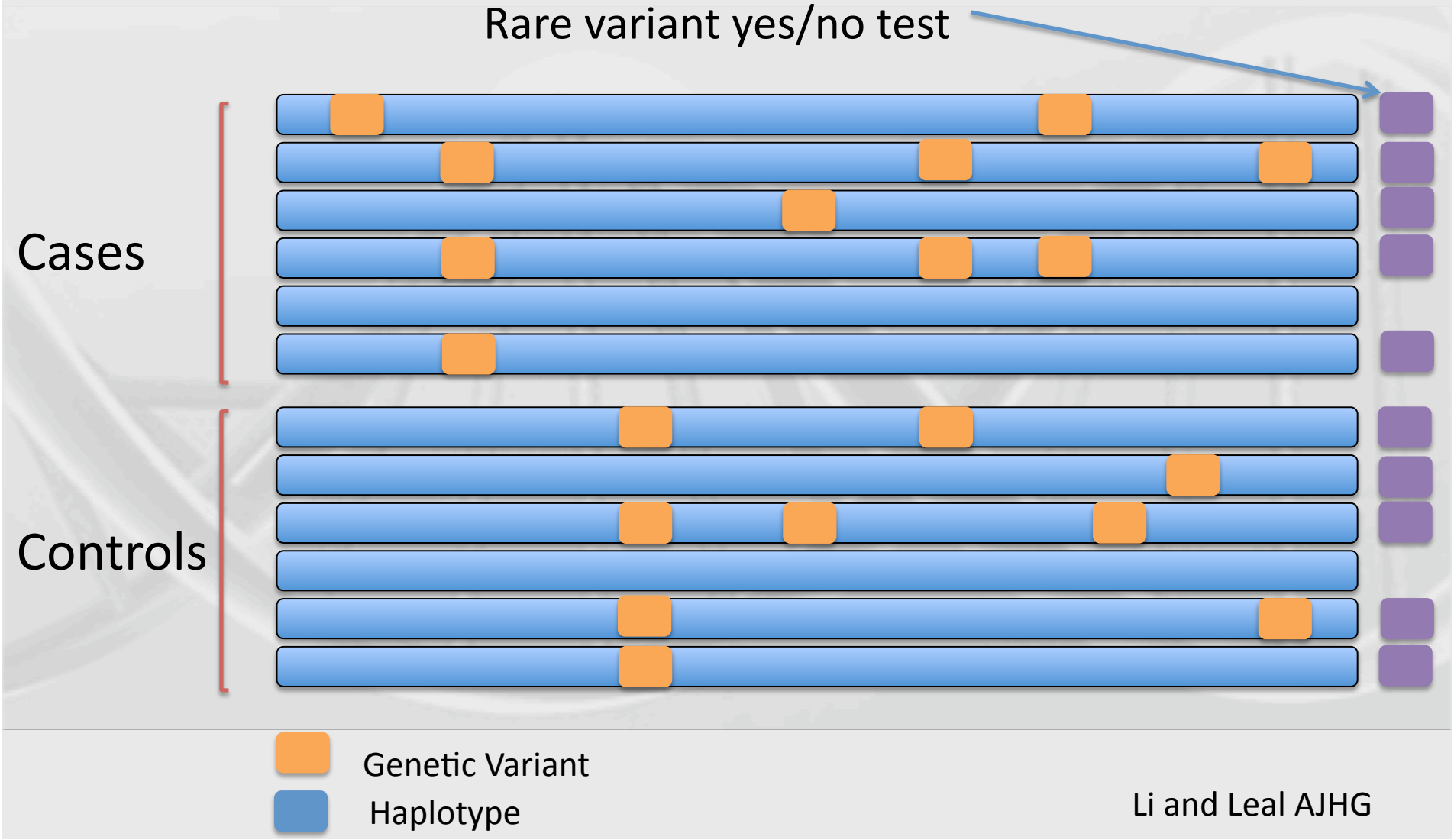


(b) 80 minor alleles in $N = 10,000$

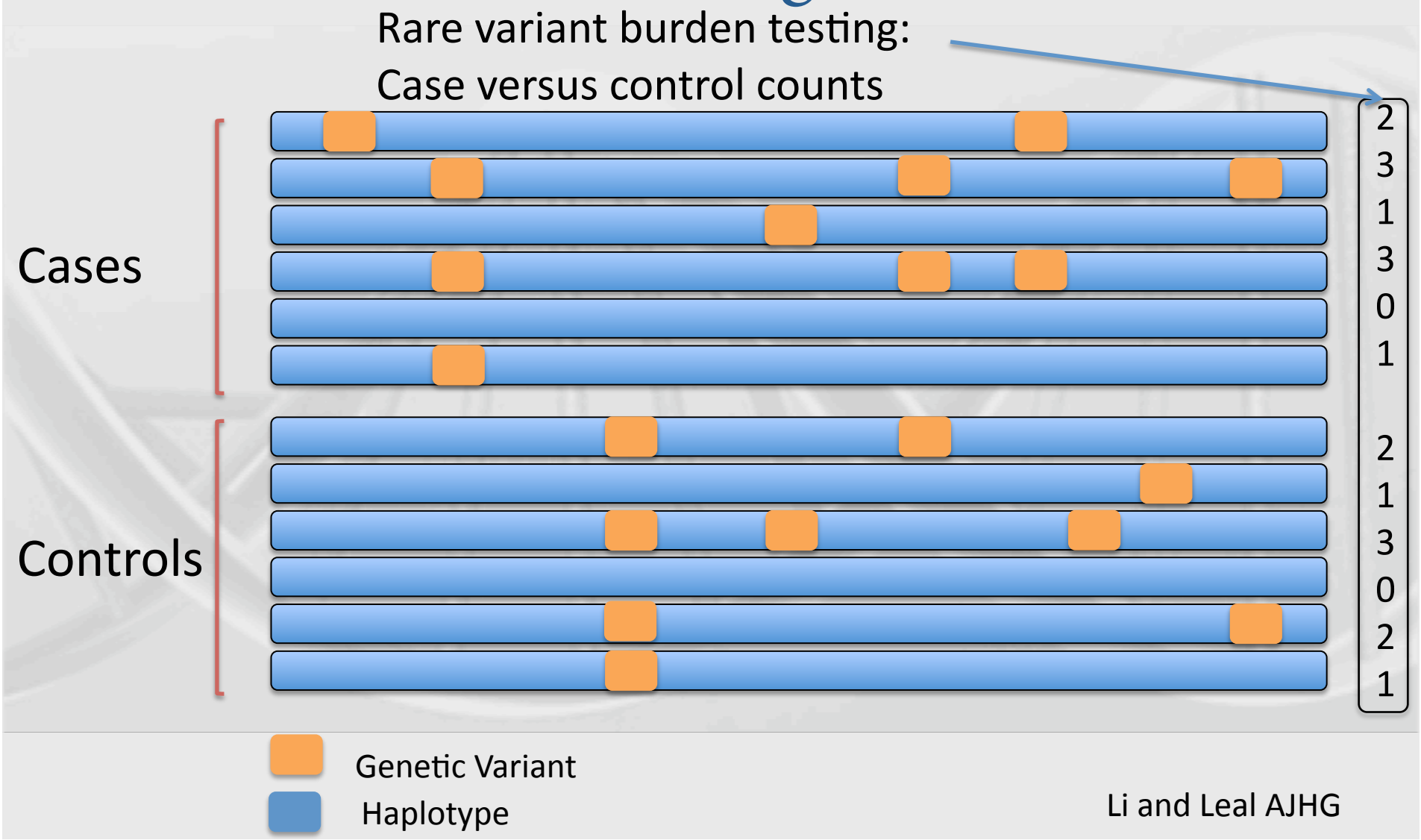
Problem:
We do not observe enough copies of rare variants to use traditional association testing approaches.

Solution:
We group variants together to increase the amount of testable variation and to improve power to detect association

Visual representation of sequence data testing



Visual representation of sequence data testing



Cohort Allelic Sum Test

- Conduct a burden test
  - # case alleles vs. # control alleles
- For a gene stratify the variation
  - 'Functional' vs. not [e.g. missense/nonsense vs. synonymous]
- Test for difference in functional vs. not.

Combined Multivariate and Collapsing

- Define an allele frequency threshold
  - Variation above this gets modeled in regression
  - Variation below gets collapsed into a single predictor [burden test]
- Improves power over collapsing everything or testing all variation separately
- Test the variants you can and collapse the ones you can't

Observation:
Neutral variation reduces power to detect, so restricting to functional variation or weighting variation may yield improved power

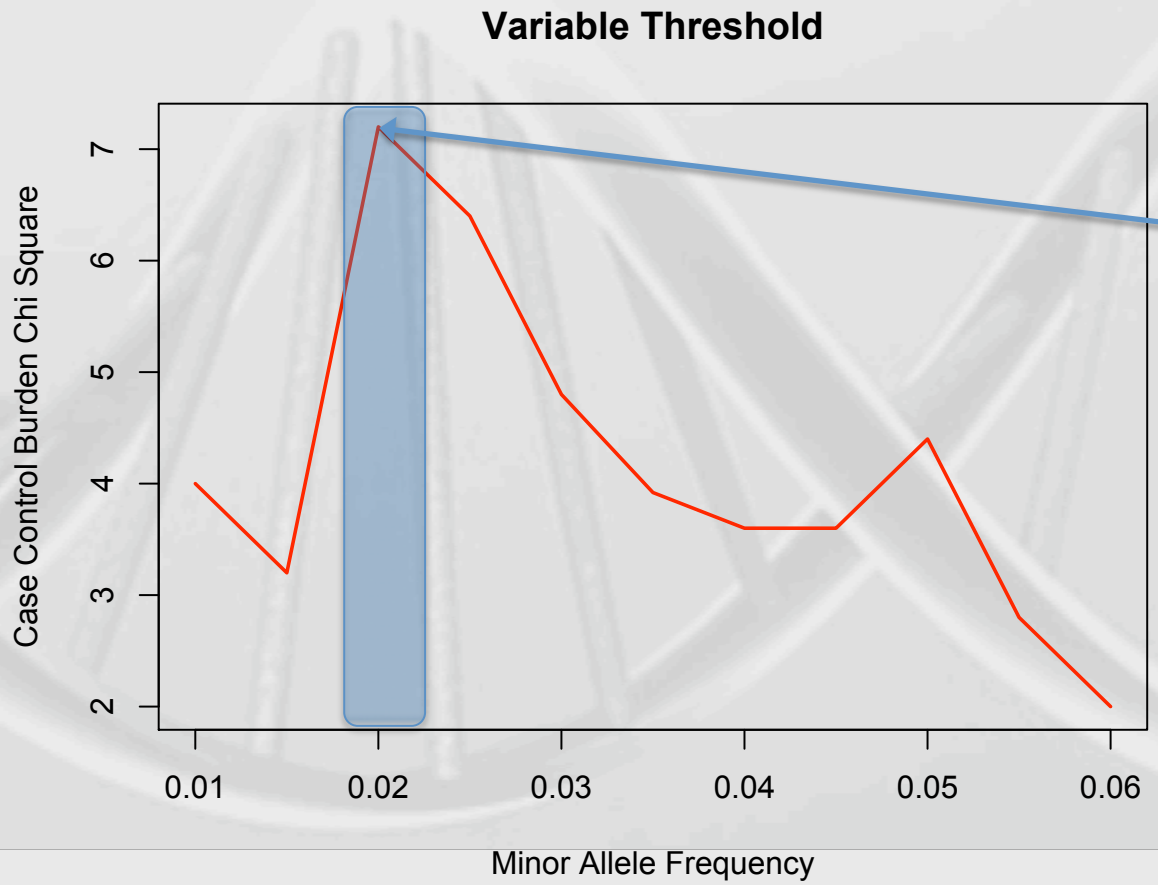
Additional methods

- Madsen and Browning (PLoS Genetics)
  - Sum of allele count, weighted by inverse of the binomial variance (i.e. rarer alleles carry more weight)
  - Functional weighting (e.g. POLYPHEN) may improve power
- Variable Threshold (Alkes Price, Shamil Sunyaev)
  - Adaptation of burden tests to optimally select allele frequency cut-off

Madsen & Browning Test

- In a region sum rare variation in a weighted fashion
- Derive these weights from the inverse of the control allele frequency
  - In practice this is v. similar to pooled allele frequency
- Conduct a rank sum test
  - Similar to the Wilcoxon test

Variable Threshold

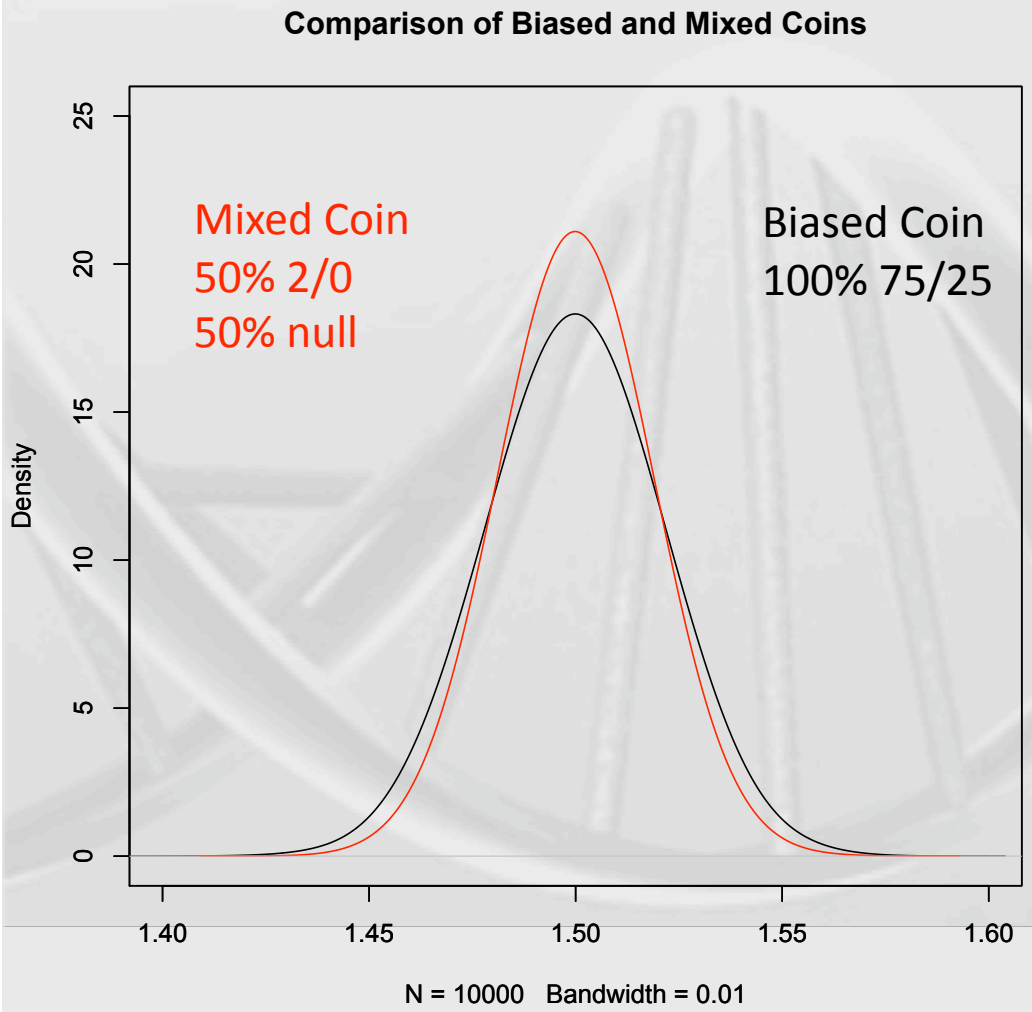


Identify threshold that gives the largest test statistic

Use permutation to assess significance

Observation:
Variation in a region that influences phenotype is potentially a mixture of neutral, risk, and protective variation

It's a mixture we're looking for



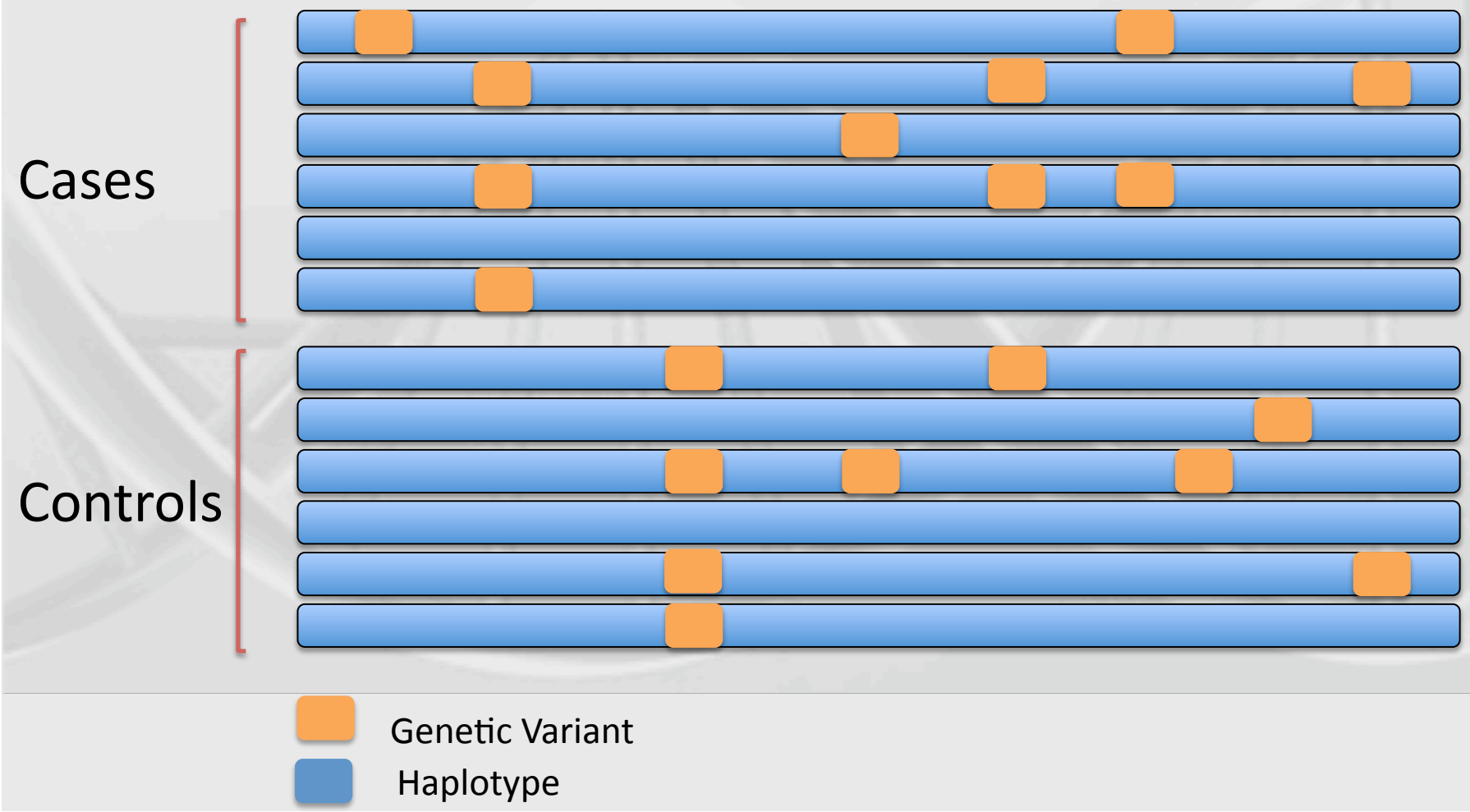
Coin	2/0	1/1	0/2
Mixed Coin	0.625	0.25	0.125
Biased Coin	0.5625	0.375	0.0625

Problem:
What is the most powerful approach for testing whether a mixture of neutral, risk, and protective variants are present?

Solution: C-alpha

- Neyman and Scott. On the use of c(α) optimal tests of composite hypotheses. 1966
  - Developed a set of functions defined for testing for mixtures
- Zelterman and Chen. Homogeneity tests against central-mixture alternatives. 1988
  - Applied c(α) approach for binomial mixtures in a score test fashion

Visual representation of sequence data testing

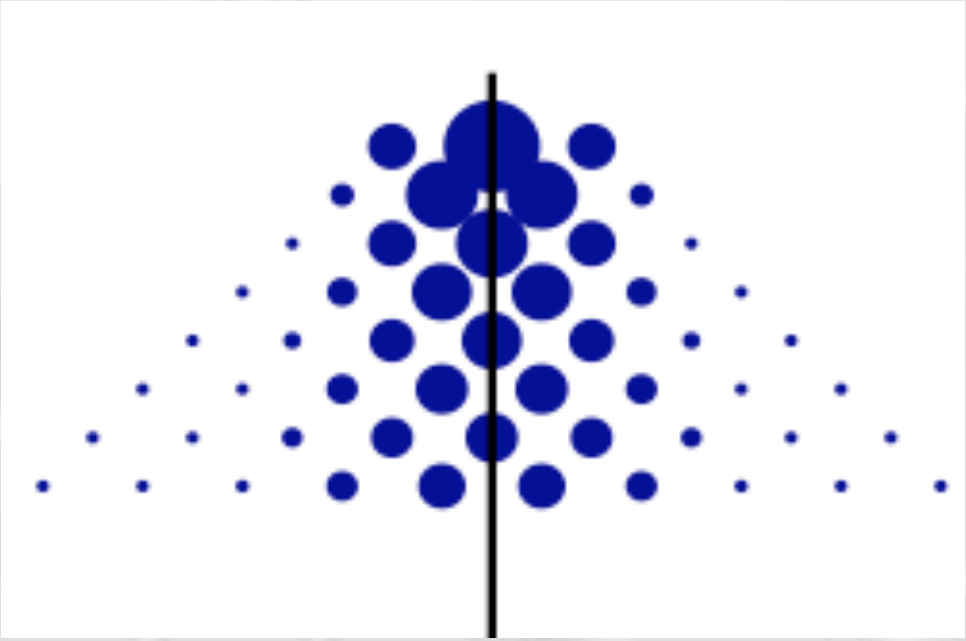


C-alpha: variant distribution in cases versus controls

- From our example the data can be reduced down to:

	Case Count / Control Count
Singles :	1/0 0/1 0/1
Doubles:	1/1 2/0 1/1
Triples:	3/0 2/1
Quadruples:	0/4
Quintuples:	-
….	

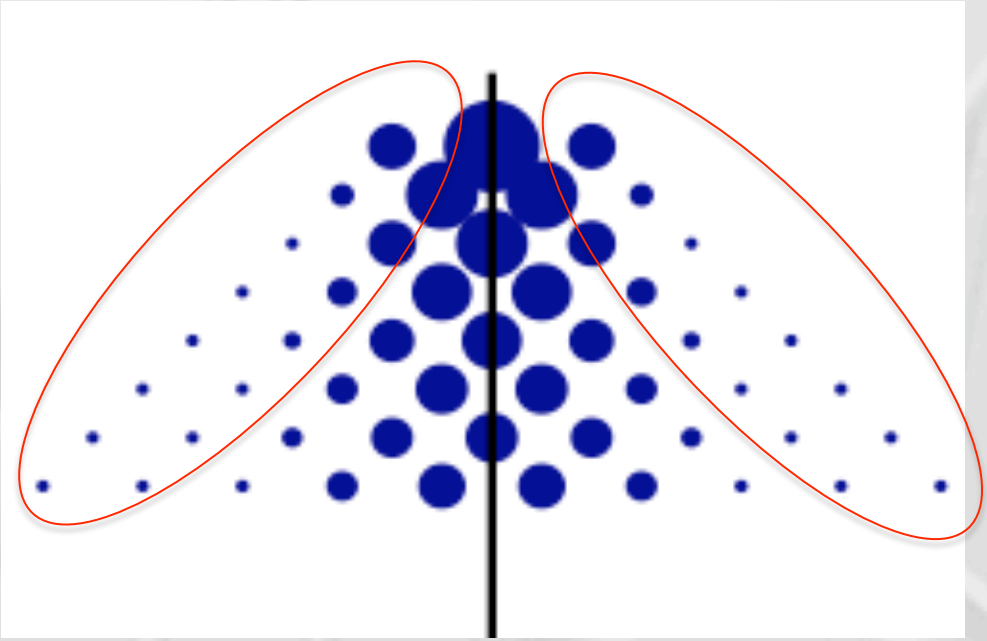
Binomial Expectation



Each row shows a different number of copies of the rare variant [2-9]

We align the variation by counting # in cases vs. # controls

What does signal look like?



We are trying to find overdispersion, or increase in variance.
So we will observe more 2:0 and 0:2 than 1:1.

Expectation vs. Deviation

	1
1 copy	1 1
2 copies	1 2 1
3 copies	1 3 3 1
4 copies	1 4 6 4 1
	.....

Expectation vs. Deviation

	1	1
1 copy	1 1	1 1
2 copies	1 2 1	1 2 1
3 copies	1 3 3 1	1 3 3 1
4 copies	1 4 6 4 1	1 4 6 4 1
	.....	.....

Red cells are more likely, blue less

Expectation vs. Deviation

	1	1
1 copy	1 1	1 1
2 copies	1 2 1	1 2 1
3 copies	1 3 3 1	1 3 3 1
4 copies	1 4 6 4 1	1 4 6 4 1
	.....	.....

This pattern of observations is characterized by overdispersion of the distribution. Specifically, any mixture of binomials will generate an inflated variance

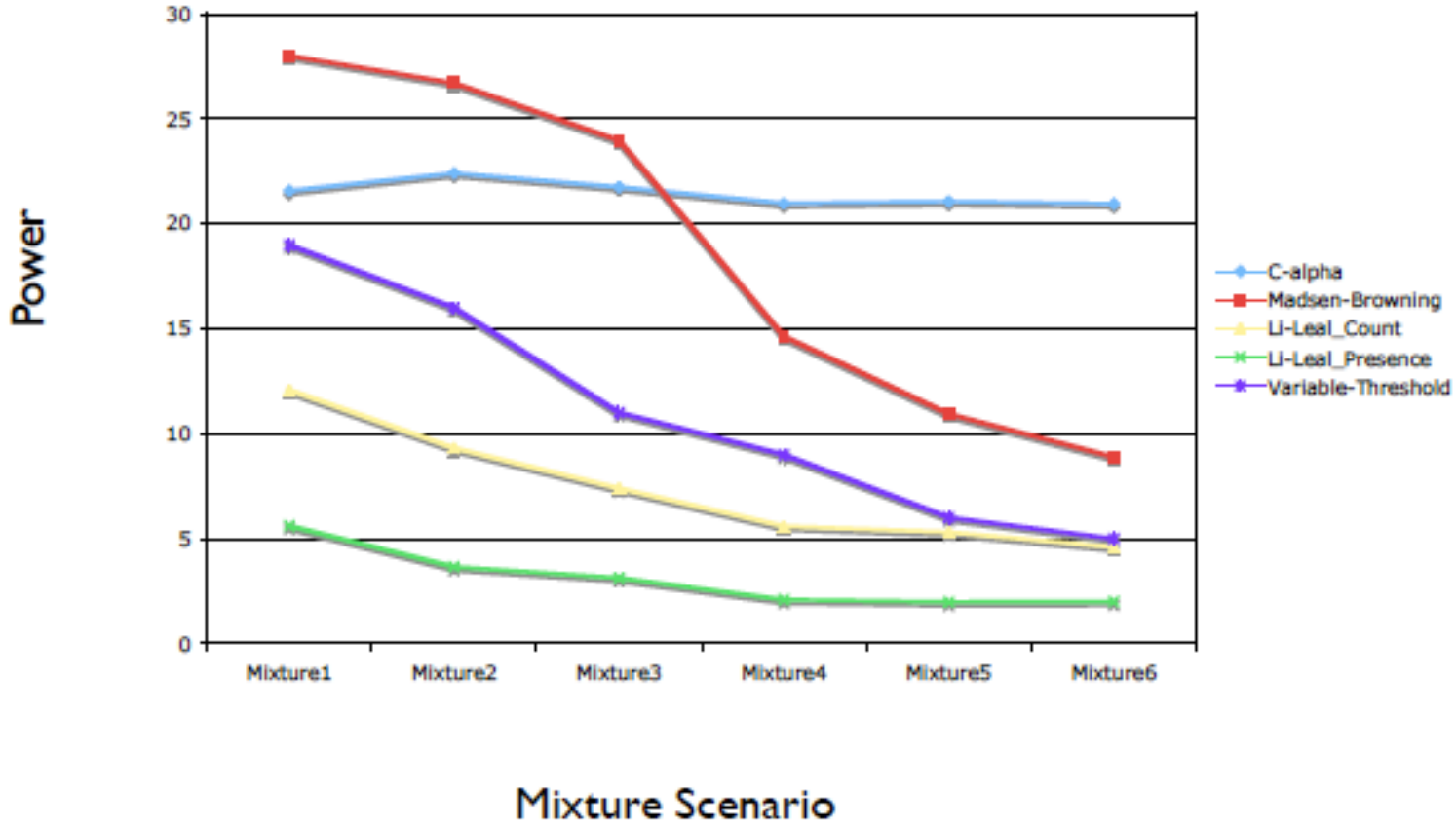
Power Comparisons - Simulations

- Population Genetics Model (Kryukov et al. 2007)
- Empirical allele frequency distribution
- Liability threshold model
- Each functional variant is under selection pressure, implying functional variants tend to be rarer
- Unselected controls

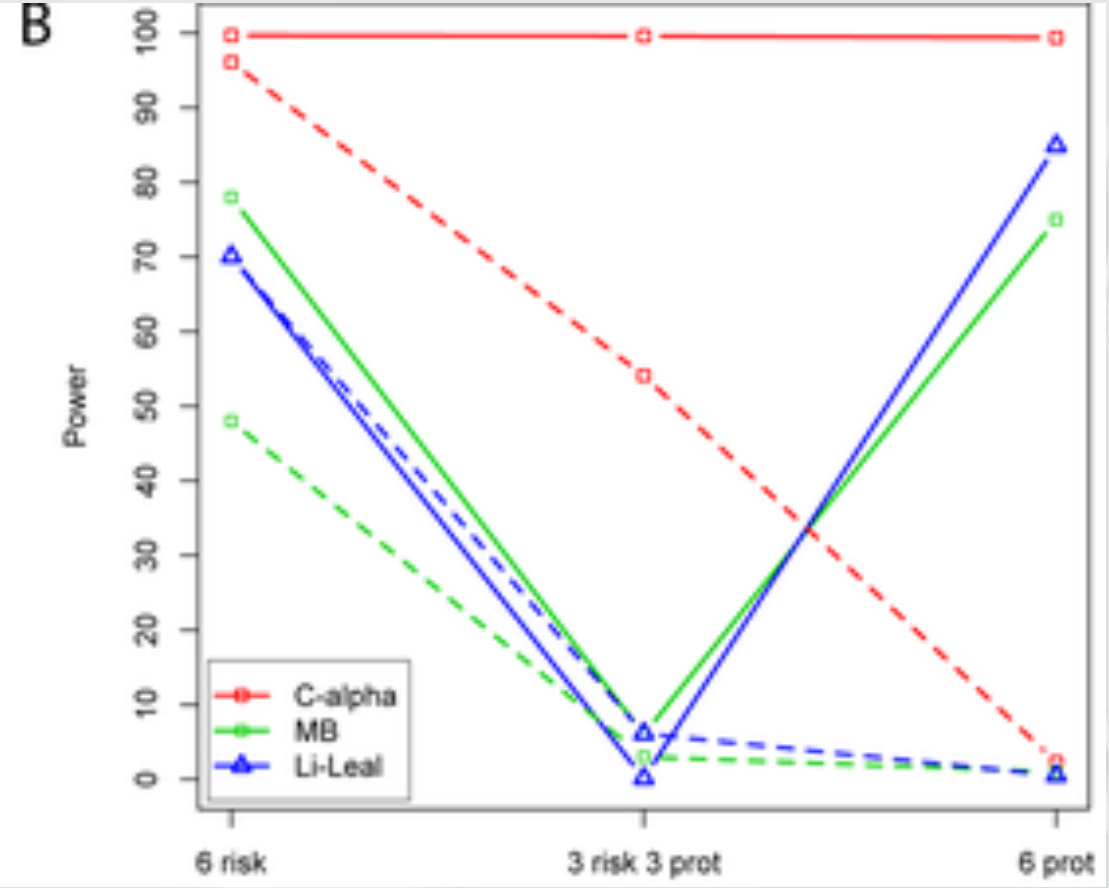
Power Comparisons - Simulations

- Equal variance explained by each functional variant
- Empirical allele frequency distribution
- Converts variance explained to relative risk (RR) via Fisher's biometrical model and liability threshold
- Uses Bayes Theorem to generate P(G|D), given RR
- Effect size increases as minor allele frequency decreases
- Selected and Unselected Controls

Power Comparisons I



Power Comparison II



Dashed refers to unselected controls
Solid refers to selected controls

Real Data – Crohn's Disease

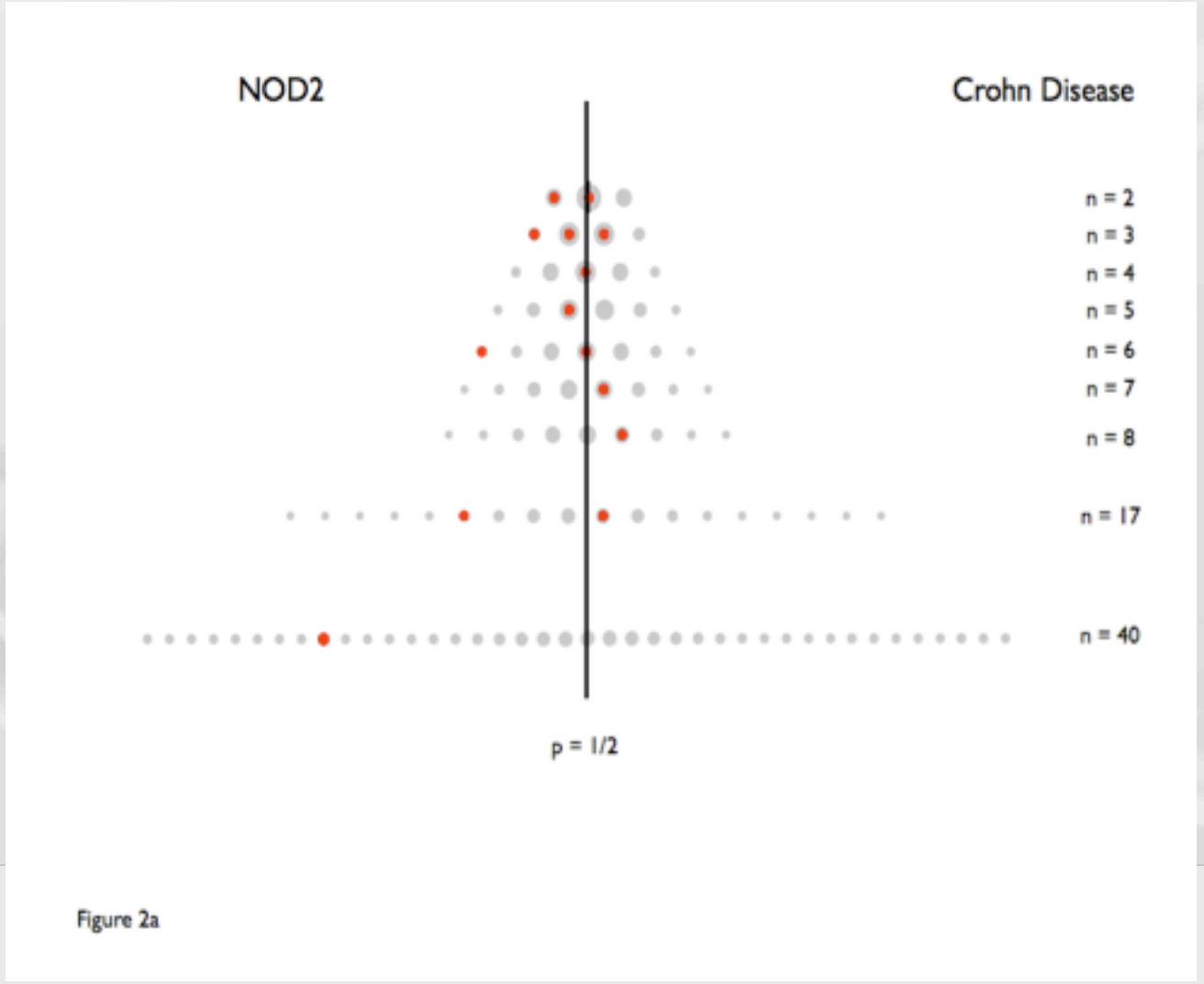


Figure 2a

Real Data – LDL Cholesterol

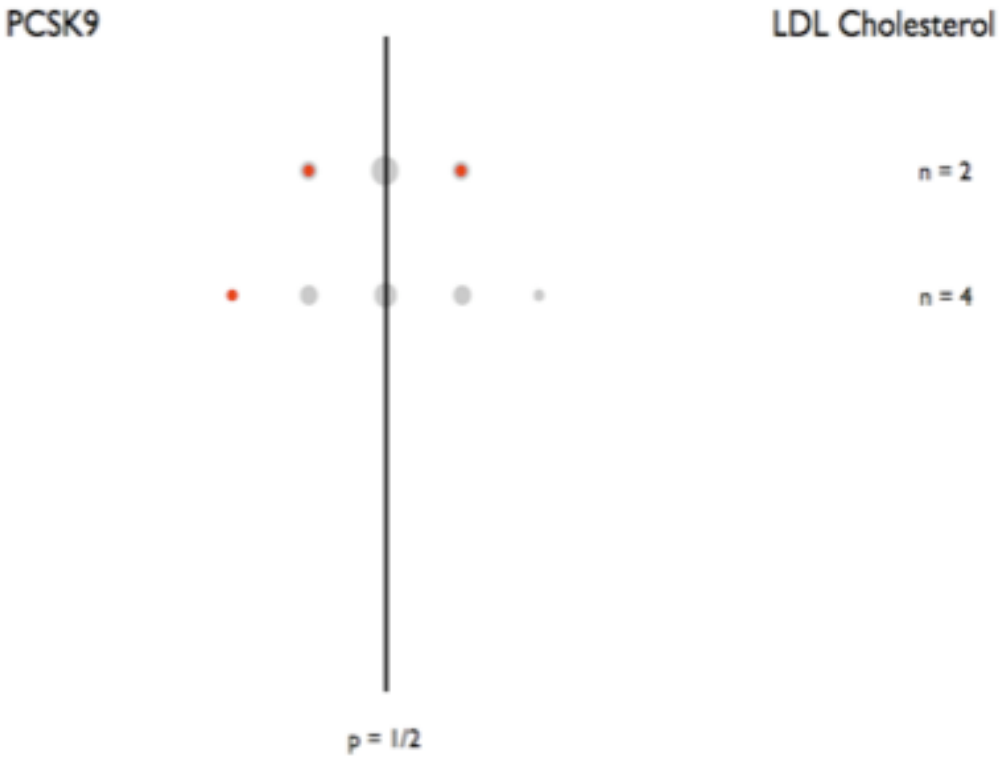


Figure 2b

APOB – Triglycerides

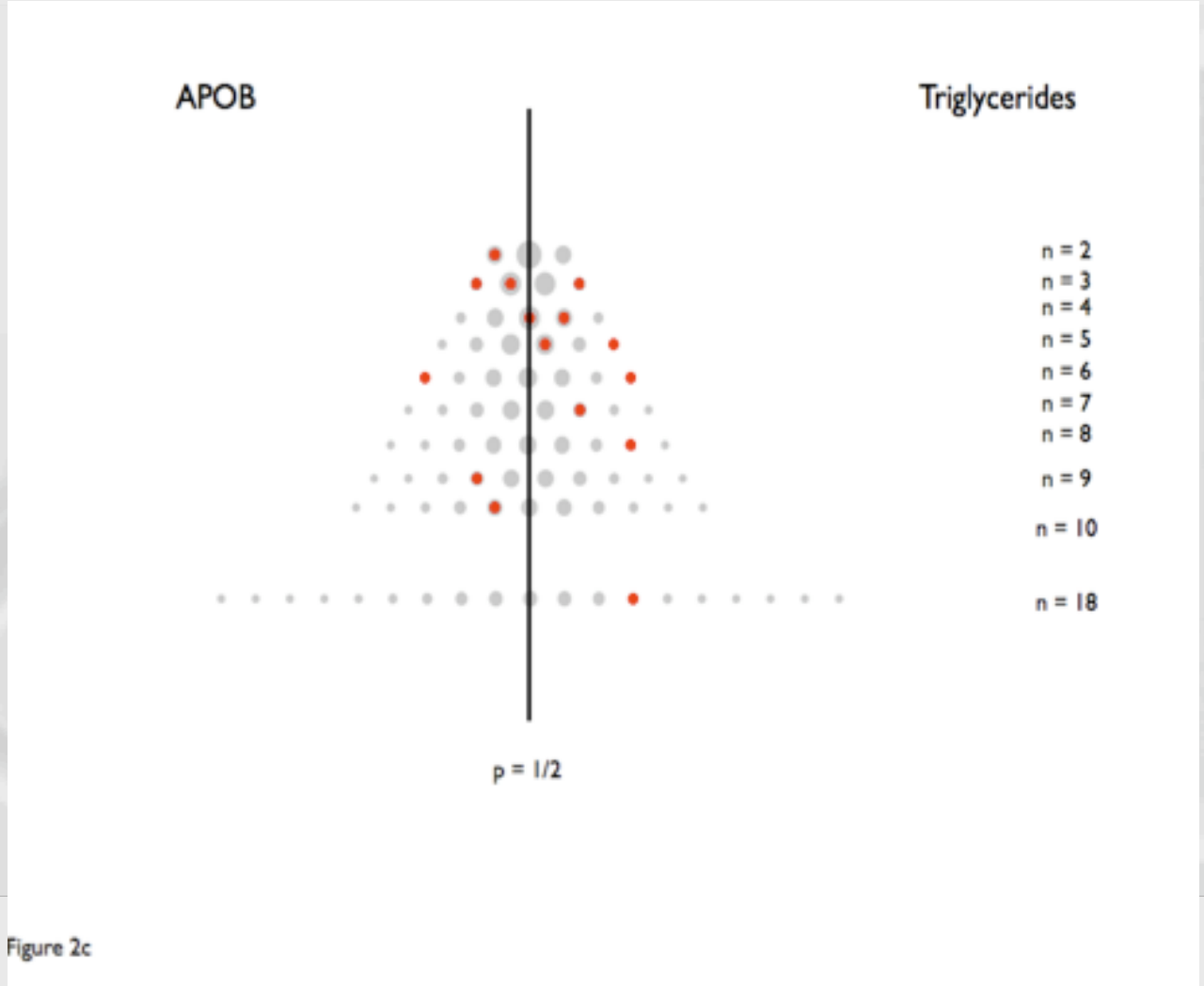


Figure 2c

APOB – Triglycerides

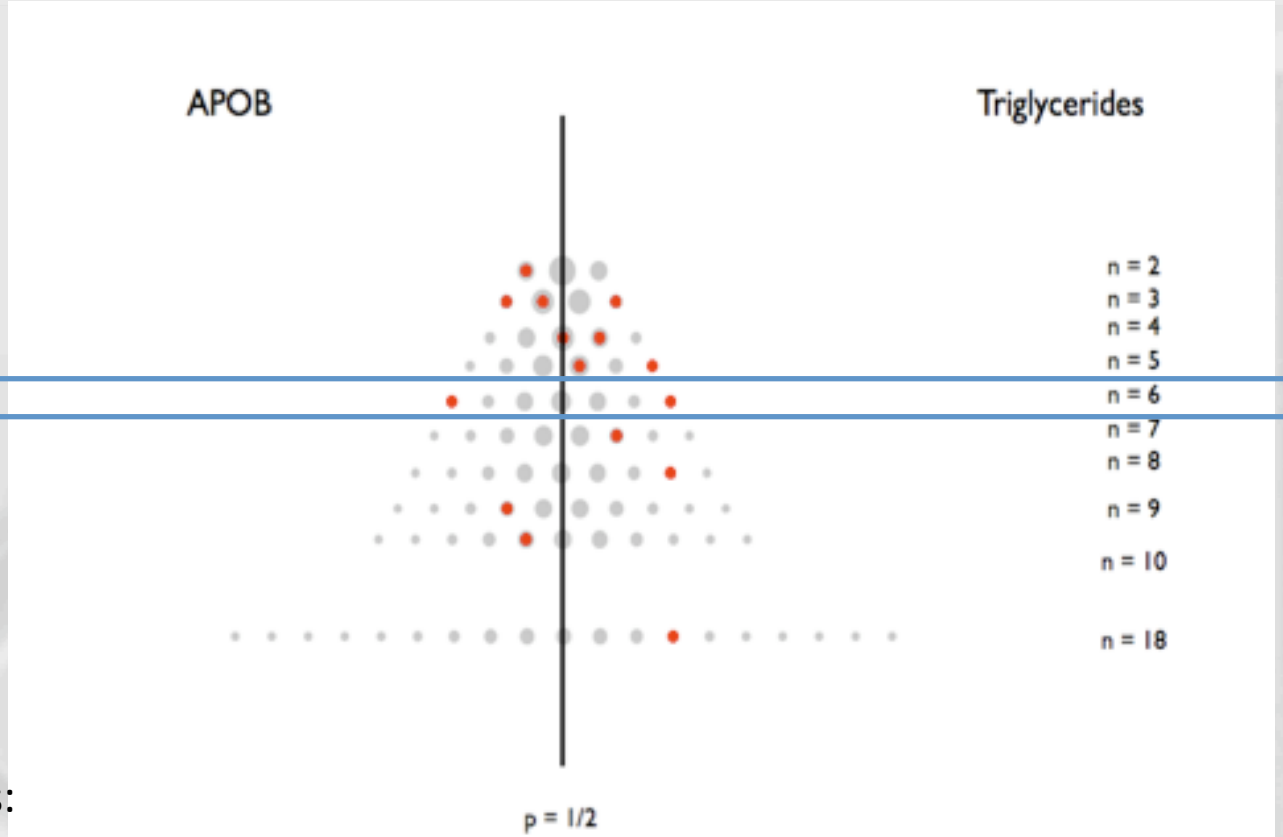


Figure 2c

Two Variants:

6 copies in cases and none in controls

&

6 copies in controls and none in cases

What about singletons?

Singletons

- Much of the variation in exomes and genomes will be singletons [i.e. observed in one person]
- C-alpha either excludes singletons or combines them into a single case/control measure [30 singletons -> 30-ton variant]
- Madsen and Browning are most heavily defined by singletons

Extensions to C-alpha

- Weights
  - Easily incorporated
  - Essential to be divorced from data
    - Madsen-Browning derives from sample -> bias
- Estimation of mixtures
  - Conditional on 'signal' can estimate risk, protective effects
  - Generates posterior probability of each variant being risk, neutral, or protective
- Pathways
  - Variants for analysis can be selected based on any grouping criteria (pathway, functionality, etc.)

Additional advantages

- C-alpha works on counts
  - Facilitates meta-analysis
  - Can be applied to pooled data
- Well-distributed statistic
  - Fast to calculate [useful for whole exome/genome]
- Sensitive to stratification
  - Use synonymous variation to calibrate
  - Permutation will not solve this problem

Other methods

OPEN ACCESS Freely available online

PLoS GENETICS

A New Testing Strategy to Identify Rare Variants with Either Risk or Protective Effect on Disease

Iuliana Ionita-Laza[1]*, Joseph D. Buxbaum[2], Nan M. Laird[3☯], Christoph Lange[3,4,5☯]

1 Department of Biostati
New York, United State
Mathematics, University

OPEN ACCESS Freely available online

PLoS one

Comprehensive Approach to Analyzing Rare Genetic Variants

Thomas J. Hoffmann[1], Nicholas J. Marini[2], John S. Witte[1]*

1 Department of Epidemiology and Biostatistics and Institute of Human Genetics, University of California San Francisco, San Francisco, California, United States of America
2 Departr

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]*

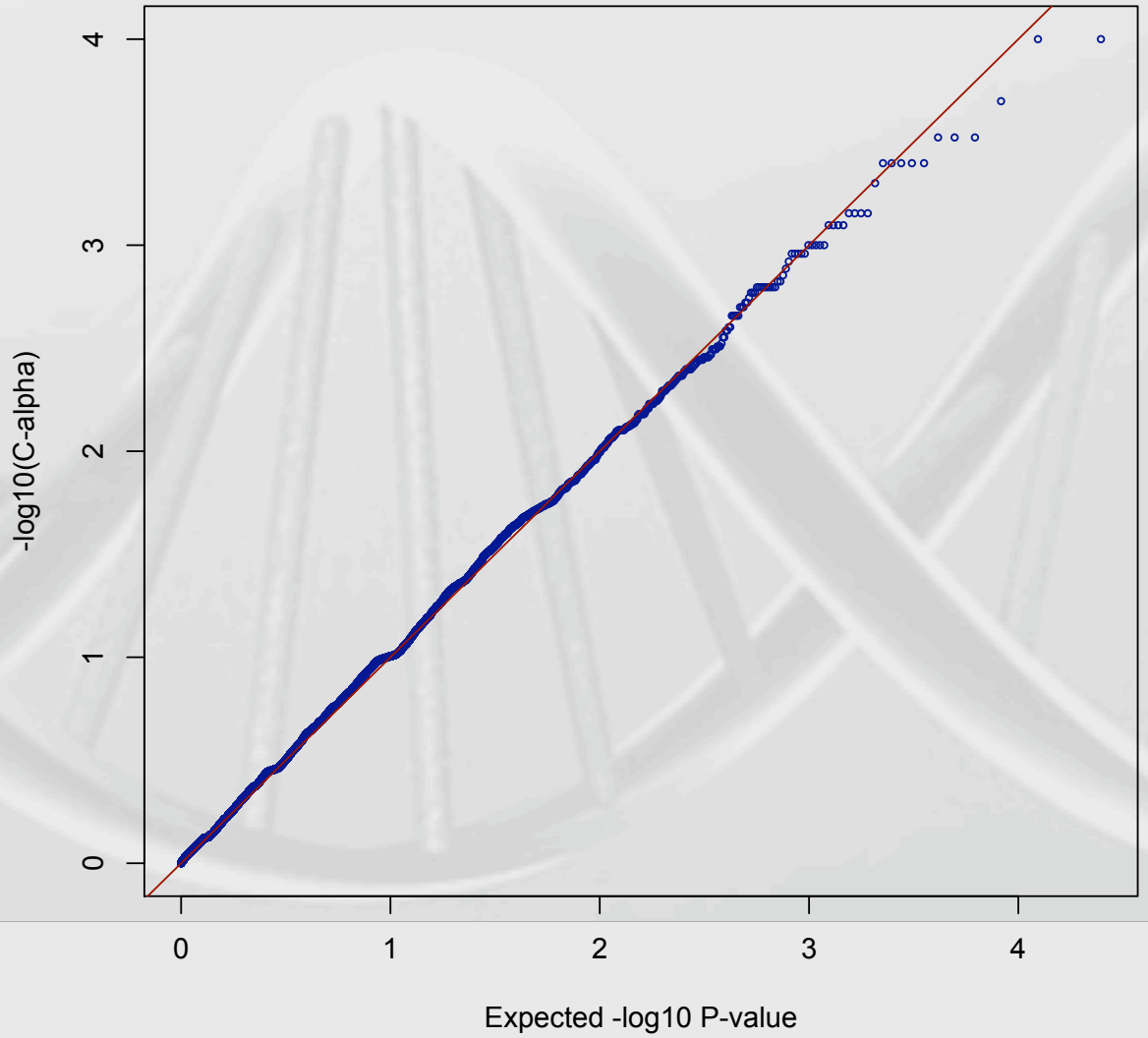
1 Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, United States of America, 2 Department of Statistics, Rice University, Houston, Texas, United States of America

Dimensionality of the exome or why I want more variants

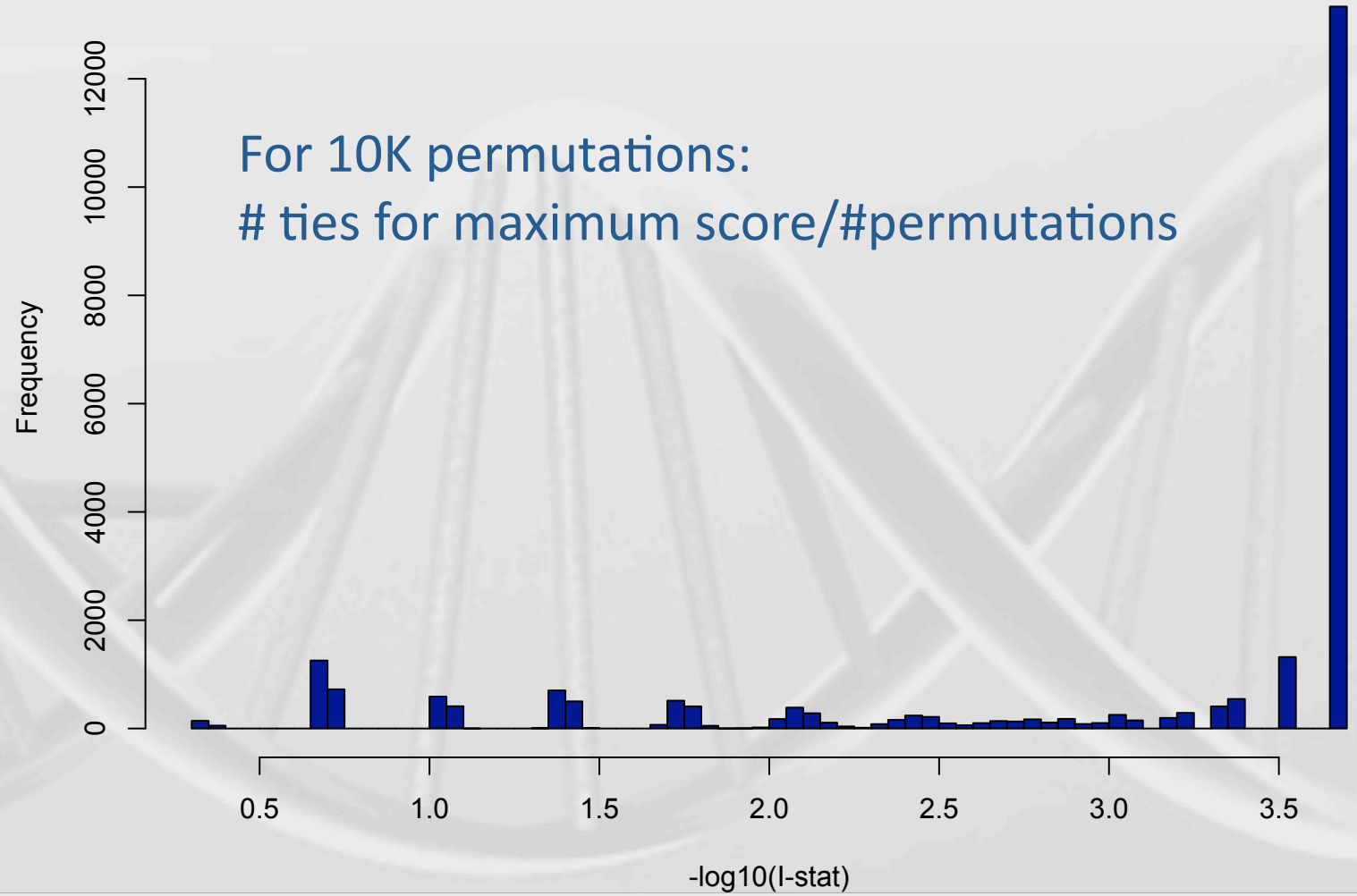
C-alpha approach

- Aims to identify overdispersion of distribution of allele gene-by-gene
- Condition minor allele count between 2 and 40
- Presented analysis on 701 passing sample
- From real autism analysis data

QQ Plot of Gene Tests

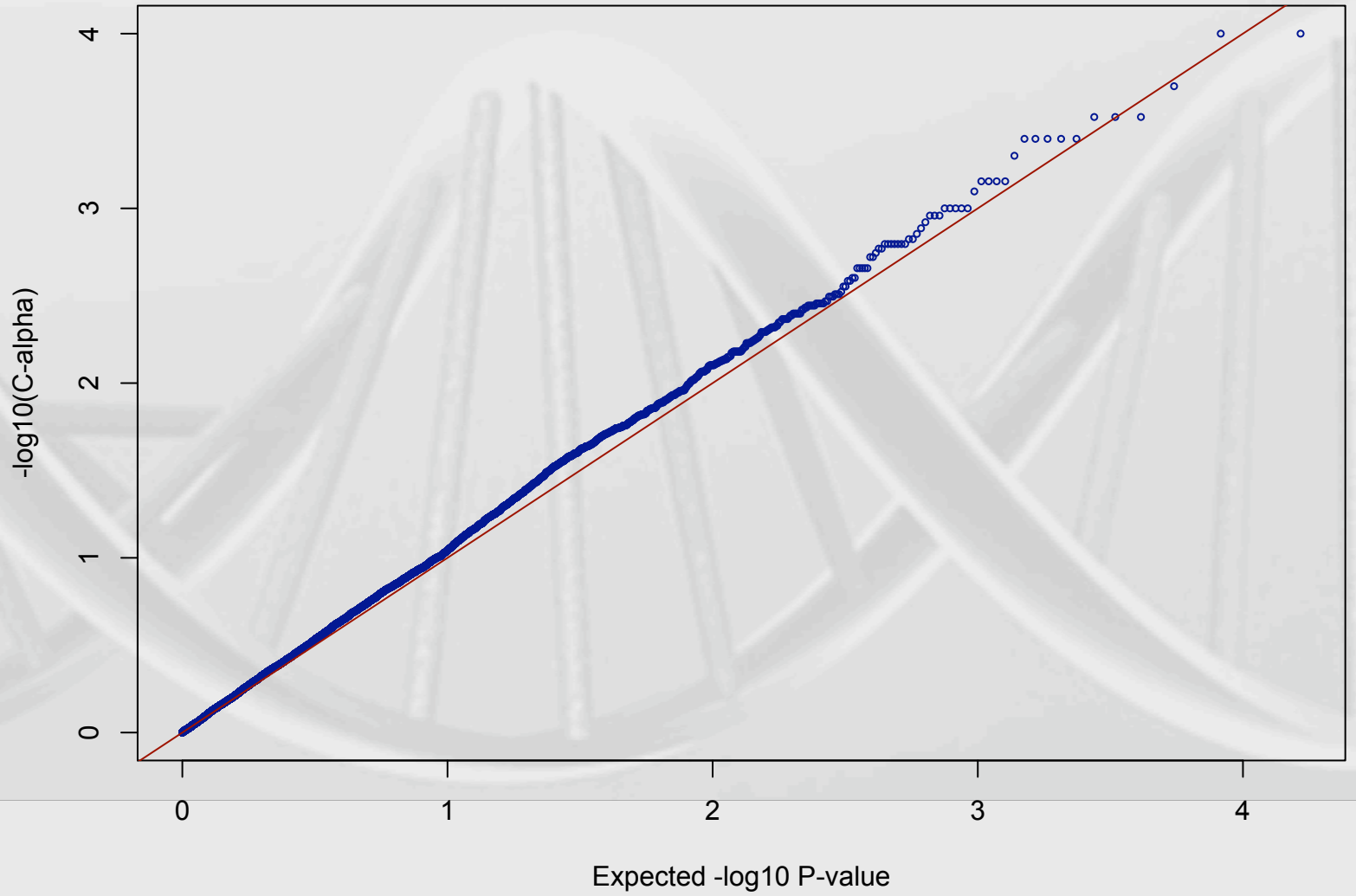


Maximum Permutation P-values



~10% P > 1e-1 ~22% P > 1e-2 ~34% P > 1e-3

QQ Plot I-stat reduced [1e-3]



Conclusions

- Much work remains for defining the optimal approach
- Different models perform well under different scenarios
- Plenty of room for extending and developing the work
  - E.g. families, pathways, etc.

References

- C-alpha http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1001322
- Variable Threshold http://www.cell.com/AJHG/abstract/S0002-9297(10)00207-7
- Madsen Browning http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000384
- Li and Leal http://www.cell.com/AJHG/abstract/S0002-9297(08)00408-4