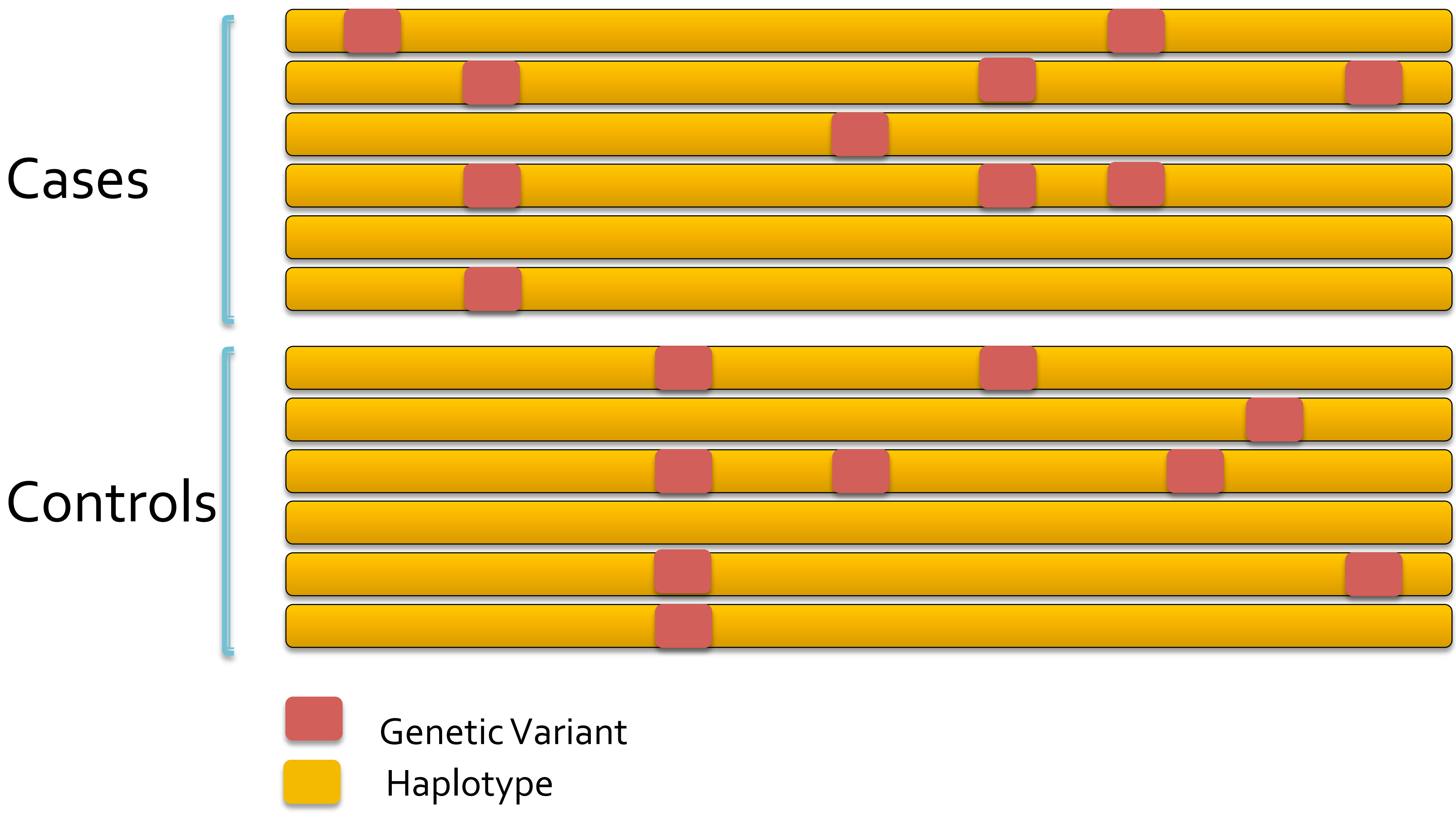
Benjamin Neale

Rare Variants I

Slide of Contents

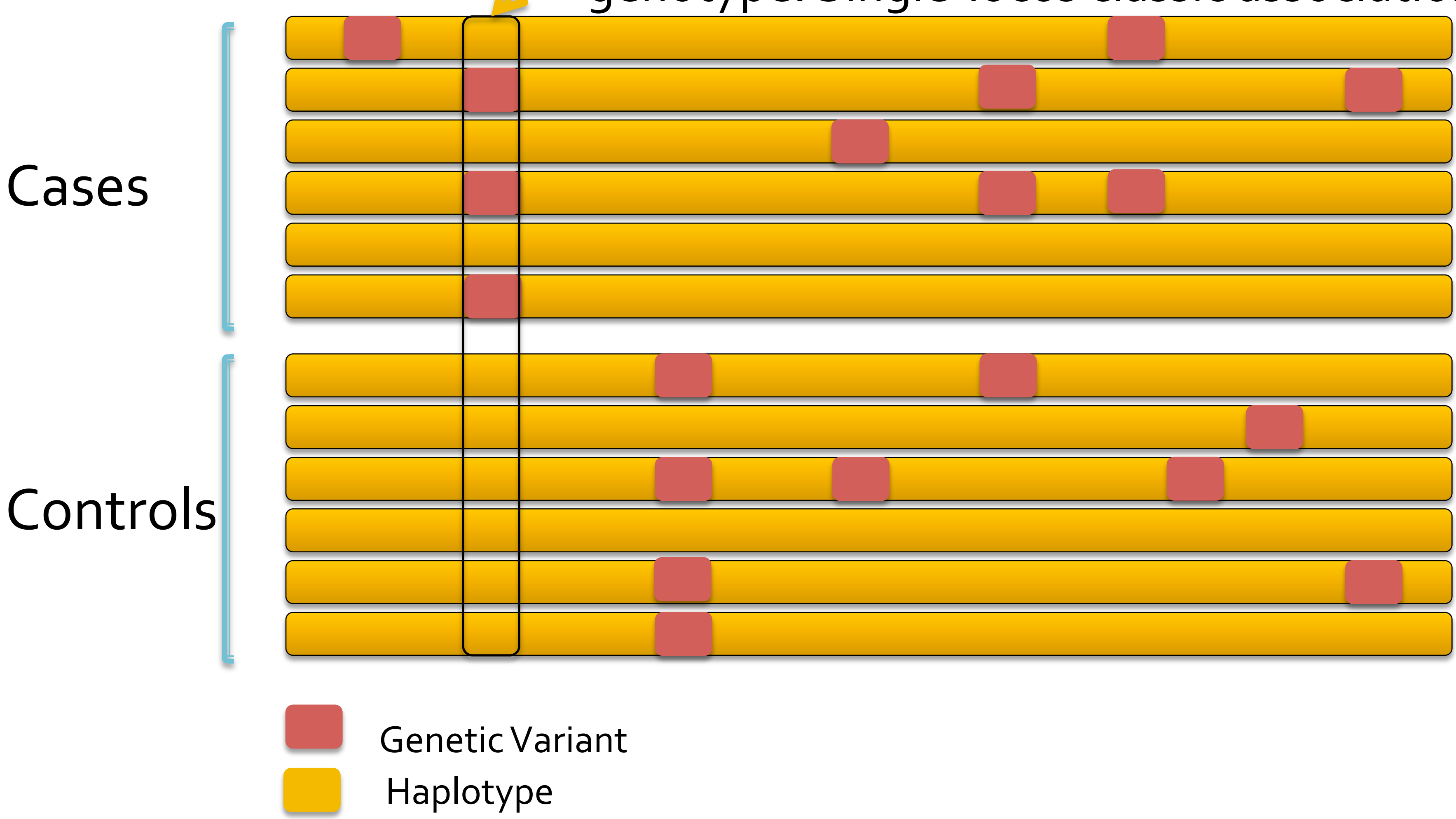
- Field-wide development of analytic approaches for sequence data
  - Cohort Allelic Sum Test (CAST; Hobbs, Cohen and others)
  - Li and Leal (AJHG)
  - Madsen and Browning (PLoS Genetics)
- C-alpha and sequence analysis
- Power comparisons
- Extensions

Visual representation of sequence data testing



Visual representation of sequence data testing

Comparing phenotype based on the genotype: Single-locus classic association



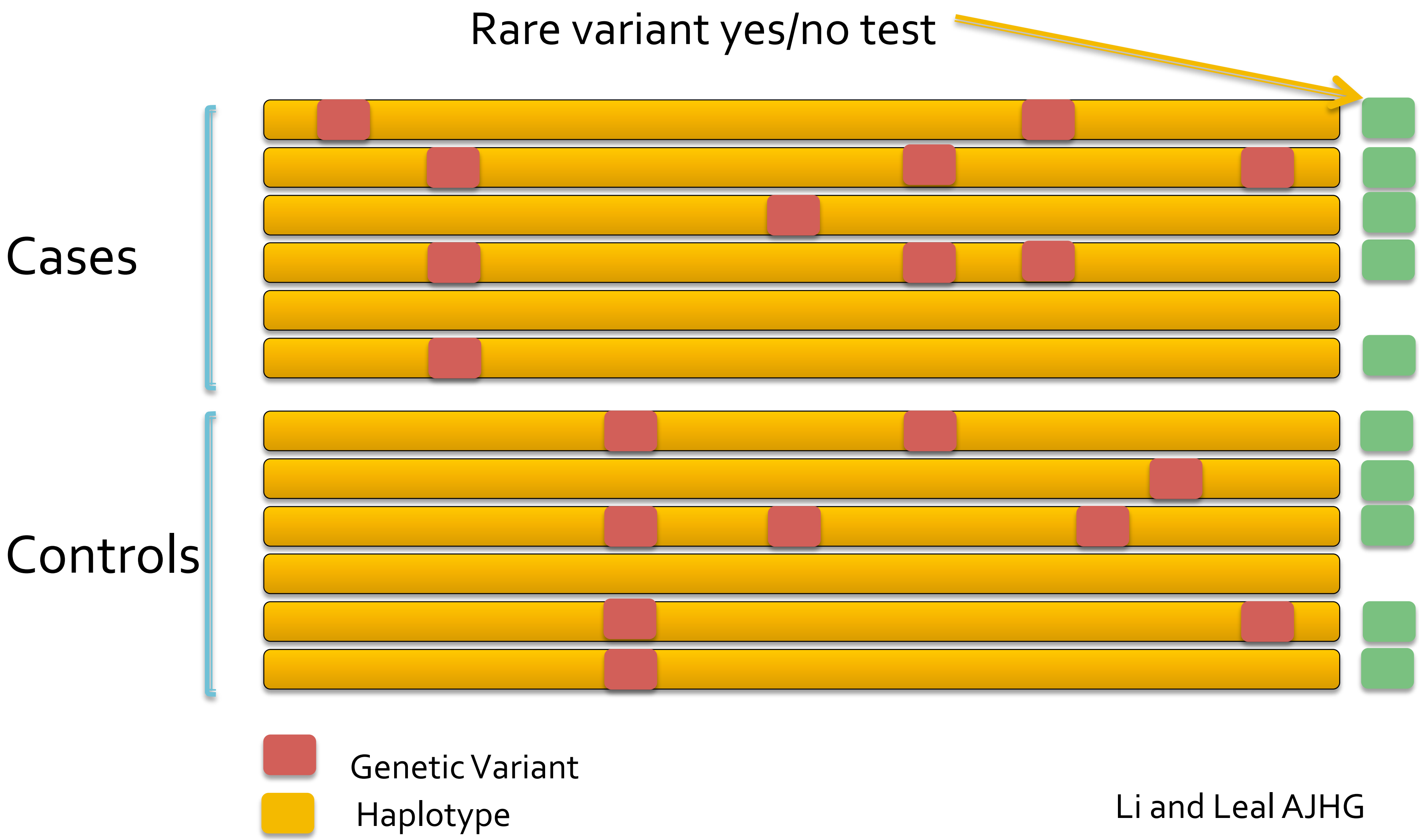
Problem:

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

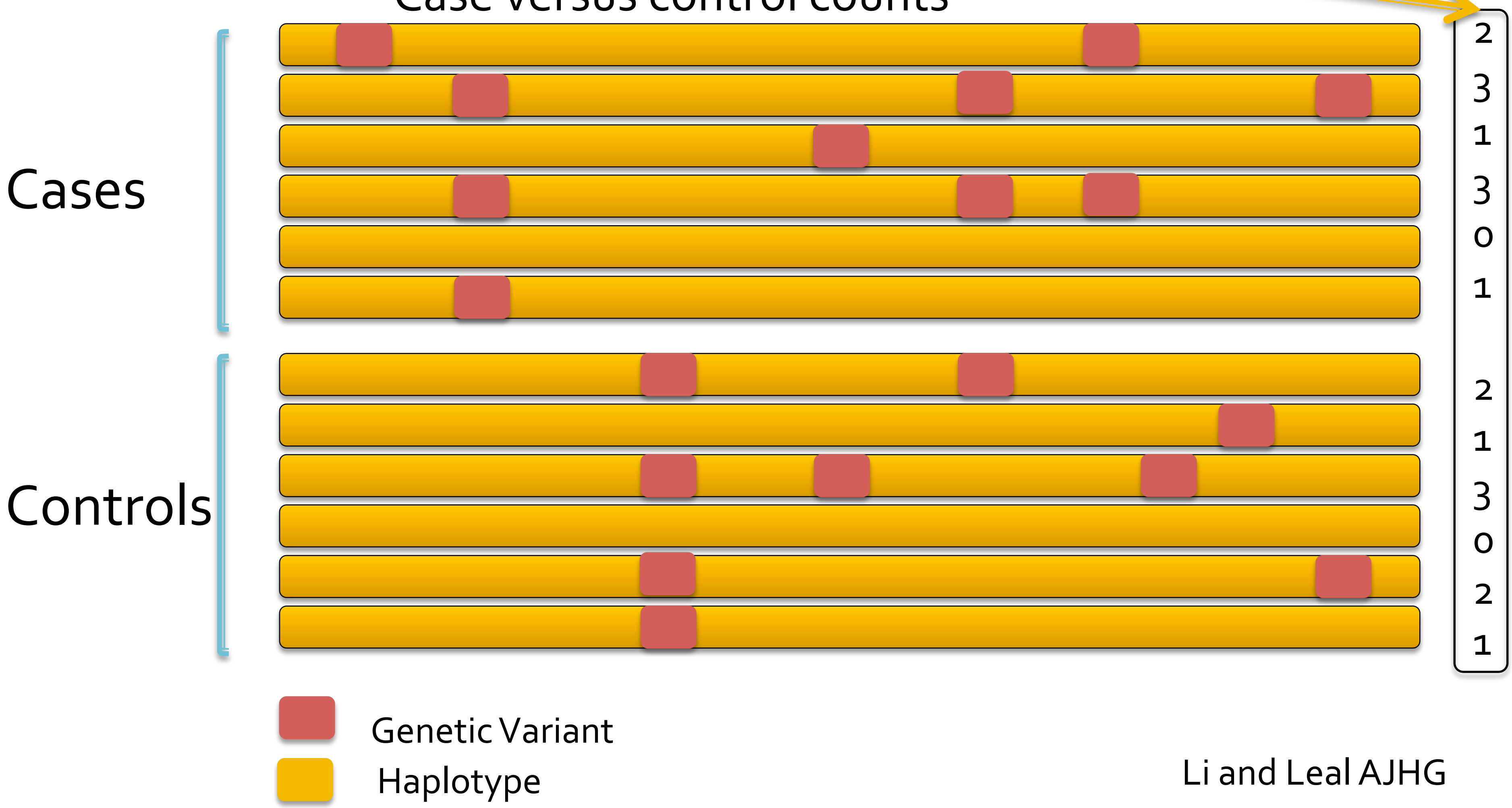
Visual representation of sequence data testing



Li and Leal AJHG

Visual representation of sequence data testing

Rare variant burden testing:
Case versus control counts



Li and Leal AJHG

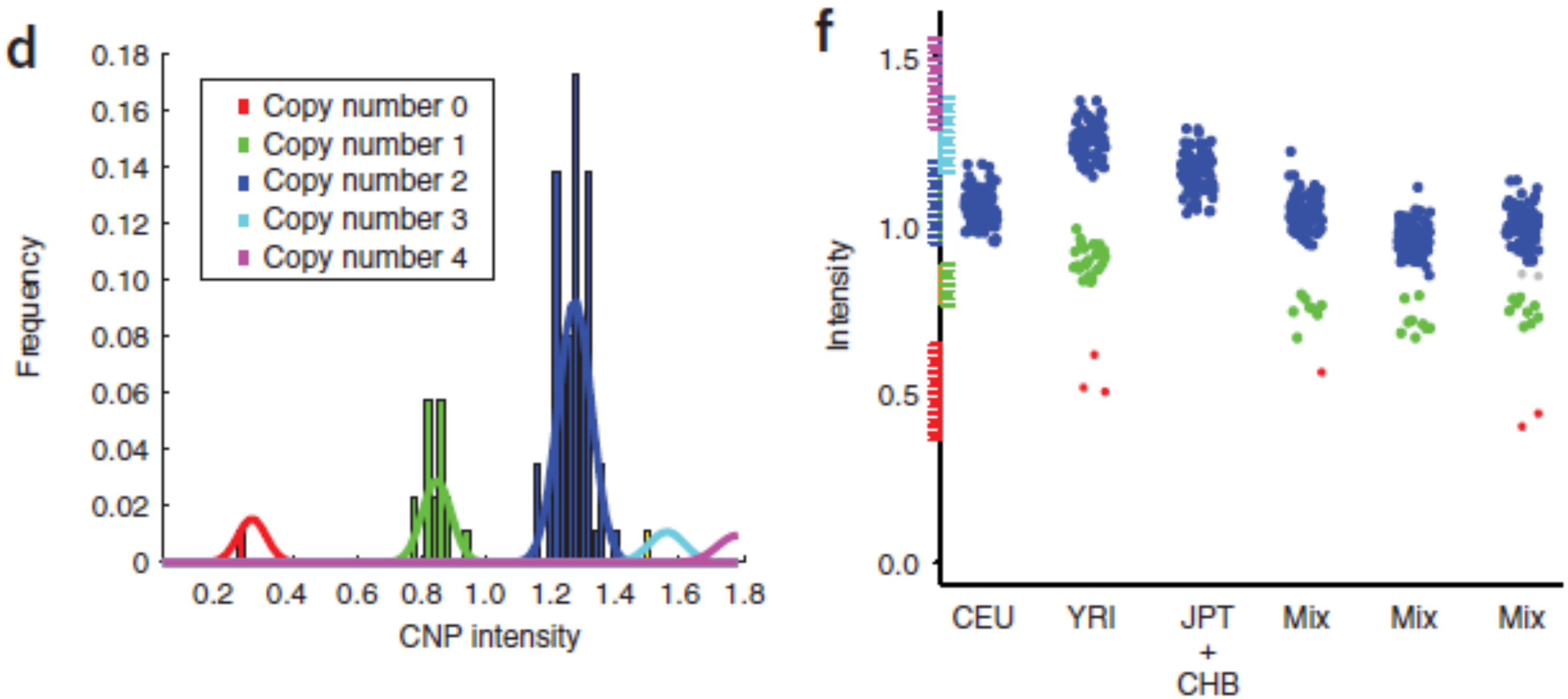
Observation:

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

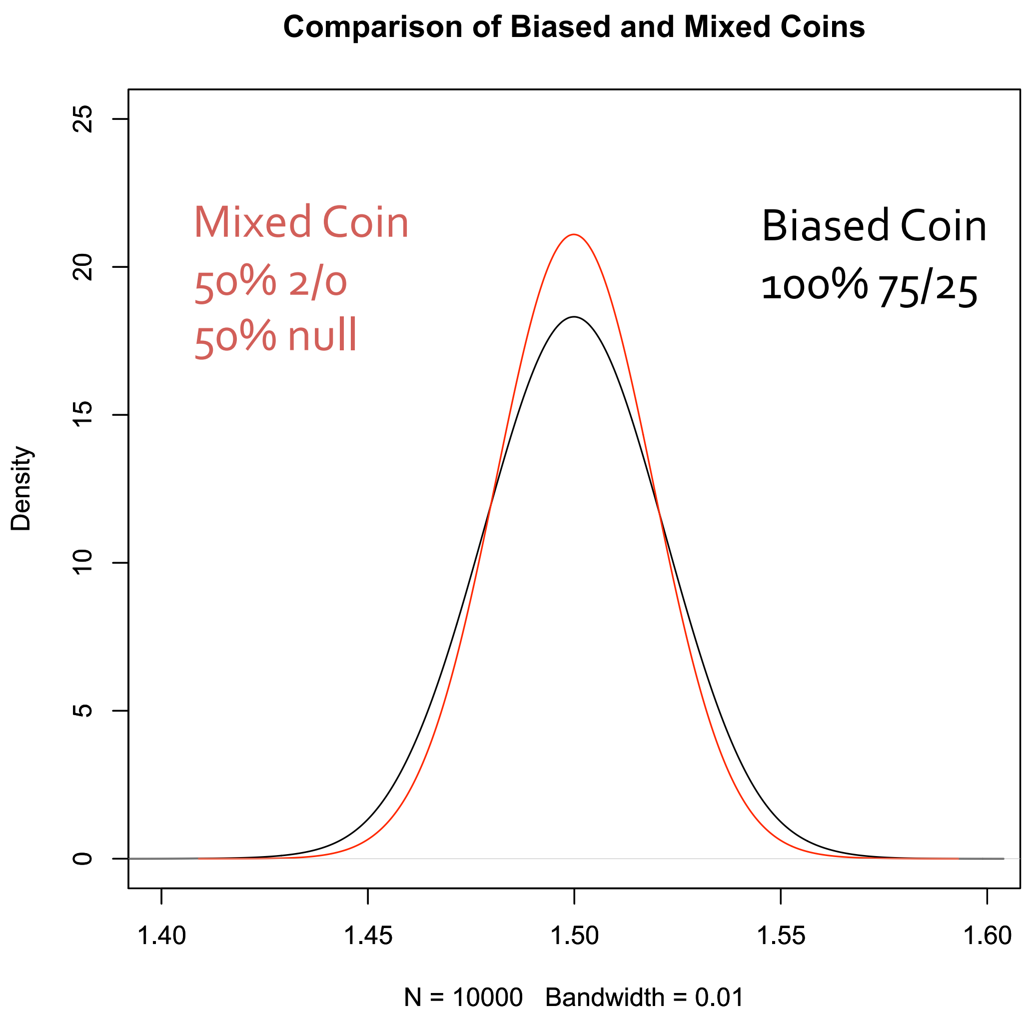
Mixtures in Genetics



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

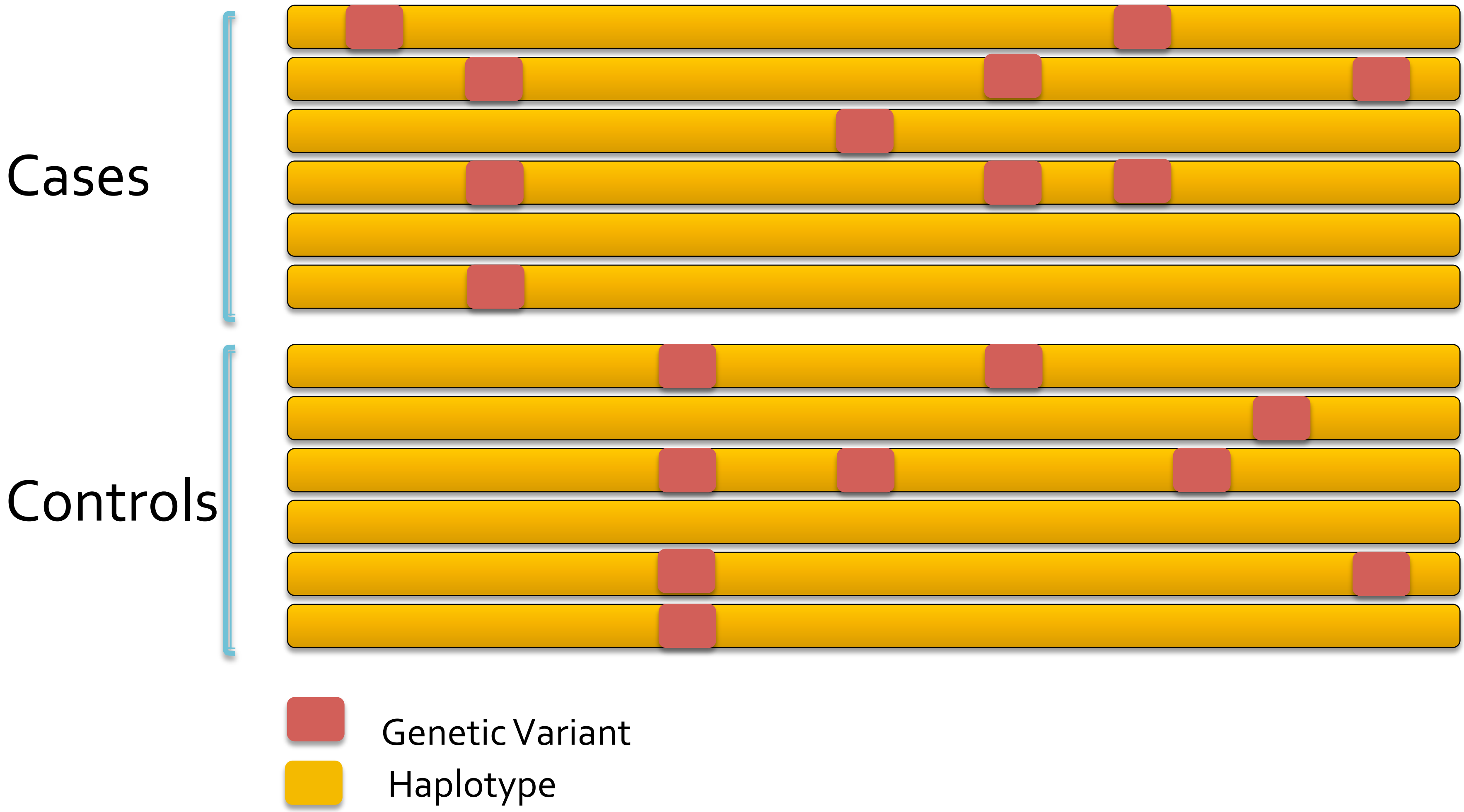
Problem:

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:	-
....	

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

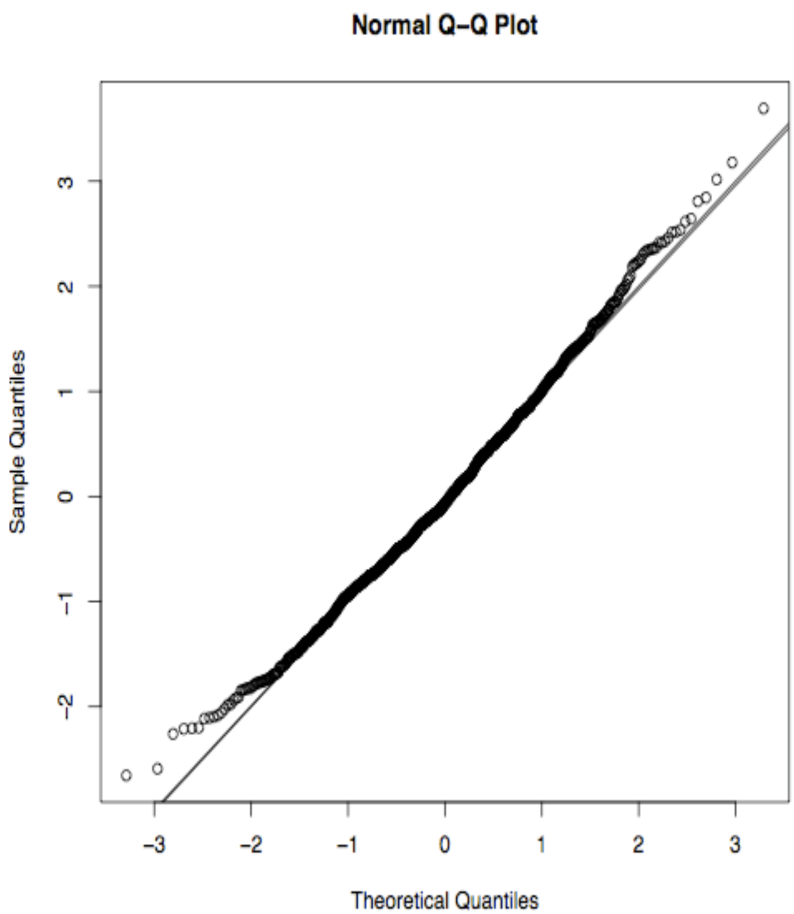
Power comparisons - methods

- Madsen and Browning (PLoS Genetics)
  - Sum of allele count, weighted by inverse of the binomial variance (i.e. rarer alleles carry more weight)
- Variable Threshold (Alkes Price, Shamil Sunyaev)
  - Adaptation of Madsen and Browning to select allele frequency cut-off
- Li and Leal (AJHG)
  - Presence: indicator variable for presence of rare variant as predictor in regression model
  - Burden: count of rare variants in the region as predictor in regression model

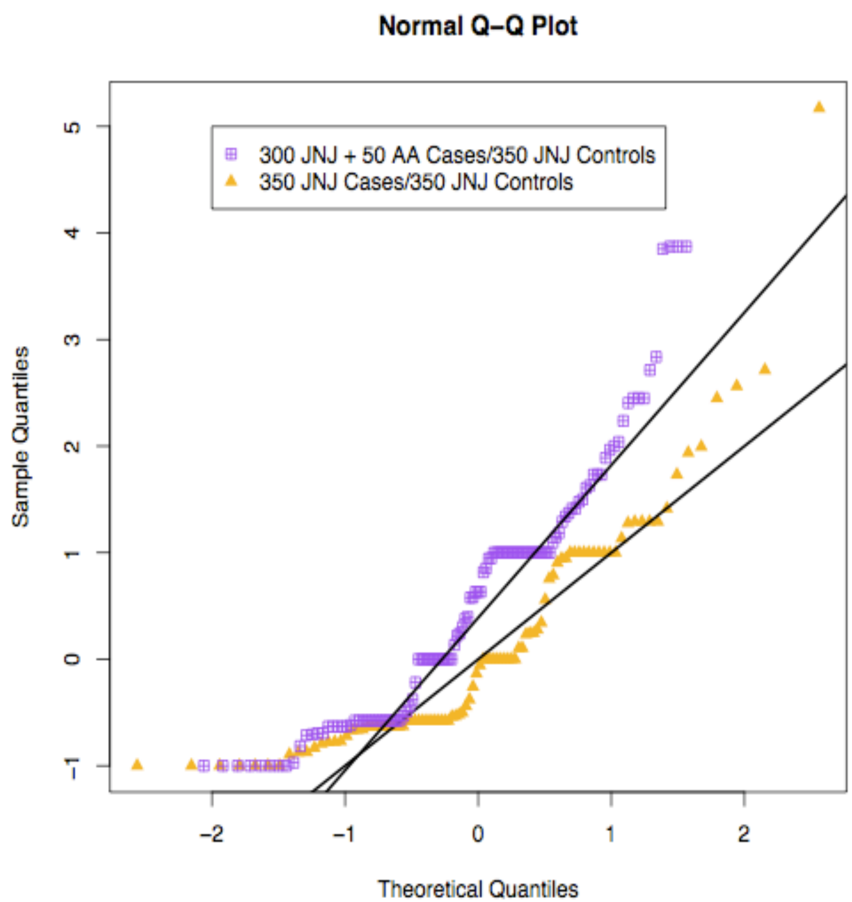
Extensions to C-alpha

- 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.)

Stratification?

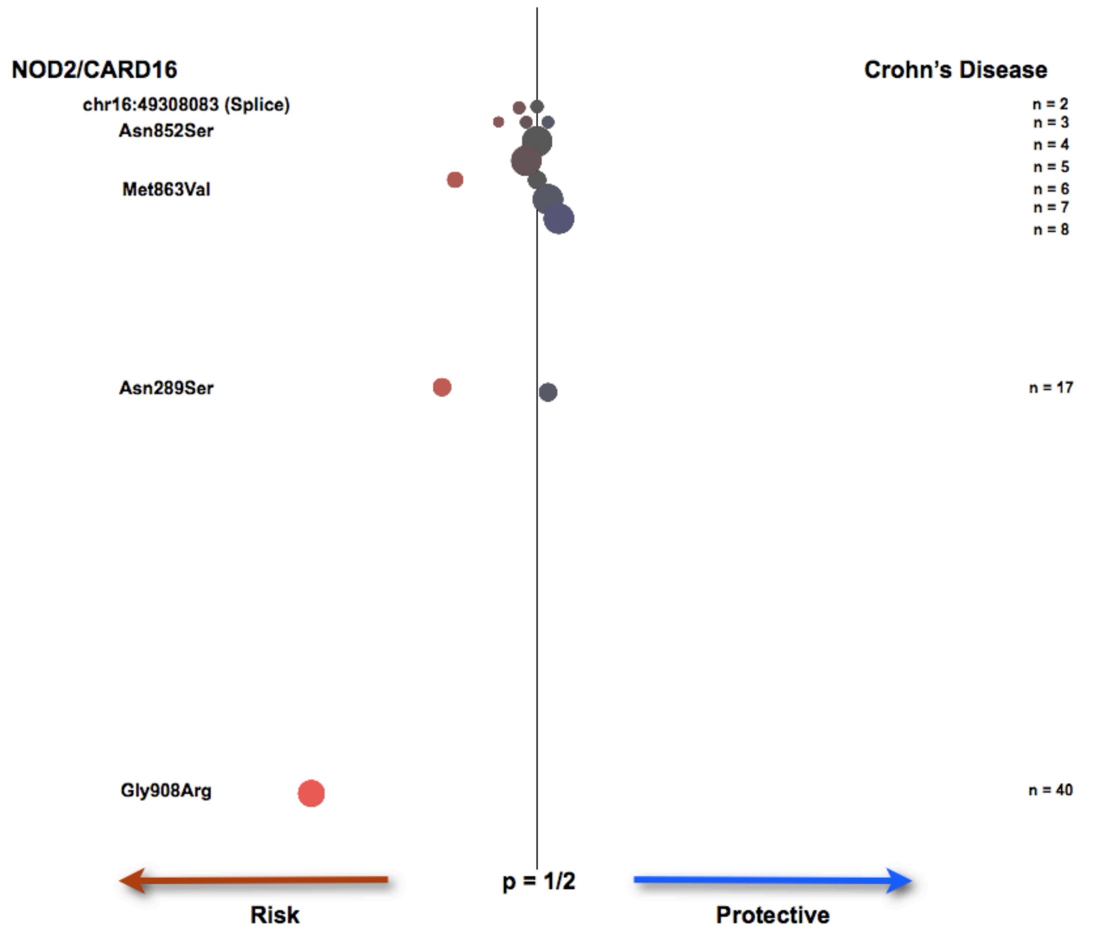


Null



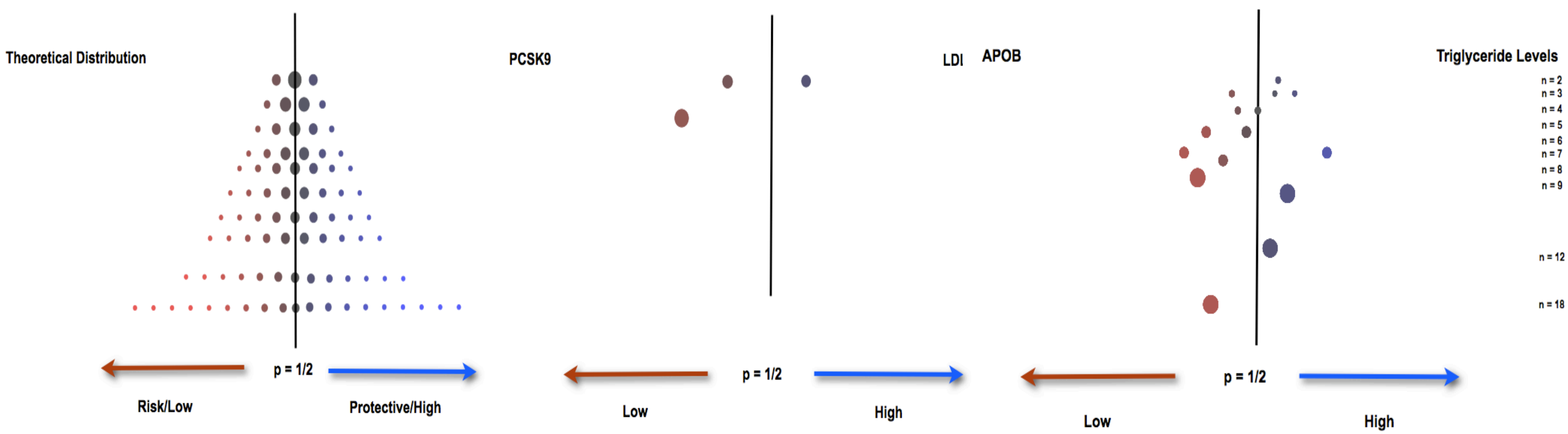
Stratified

What about real data?



C-alpha is robust to mixture of null variants, in the presence of a few large effects

More real data



Both PCSK9 and APOB show signal using C-alpha in the presence of clear variants which influence the phenotype in both directions

Additional advantages

- C-alpha works on counts
  - Facilitates meta-analysis
  - Can be applied to pooled data [as we've seen]
- 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

Acknowledgements

- Madsen and Browning: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2633048/
- Li and Leal: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2842185/?tool=pubmed
- Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR., Pooled Association Tests for Rare Variants in Exon-Resequencing Studies., Am J Hum Genet. 2010 May 12. [Epub ahead of print]
- Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 2006 Mar 23;354(12):1264-72.
- C-alpha
- Mark Daly
- Manny Rivas
- Kathryn Roeder
- Ben Voight
- Bernie Devlin
- Shaun Purcell
- David Altshuler
- Sekar Kathiresan
- Marju Orho-Melander
- For more information C-alpha contact me at bneale@broadinstitute.org