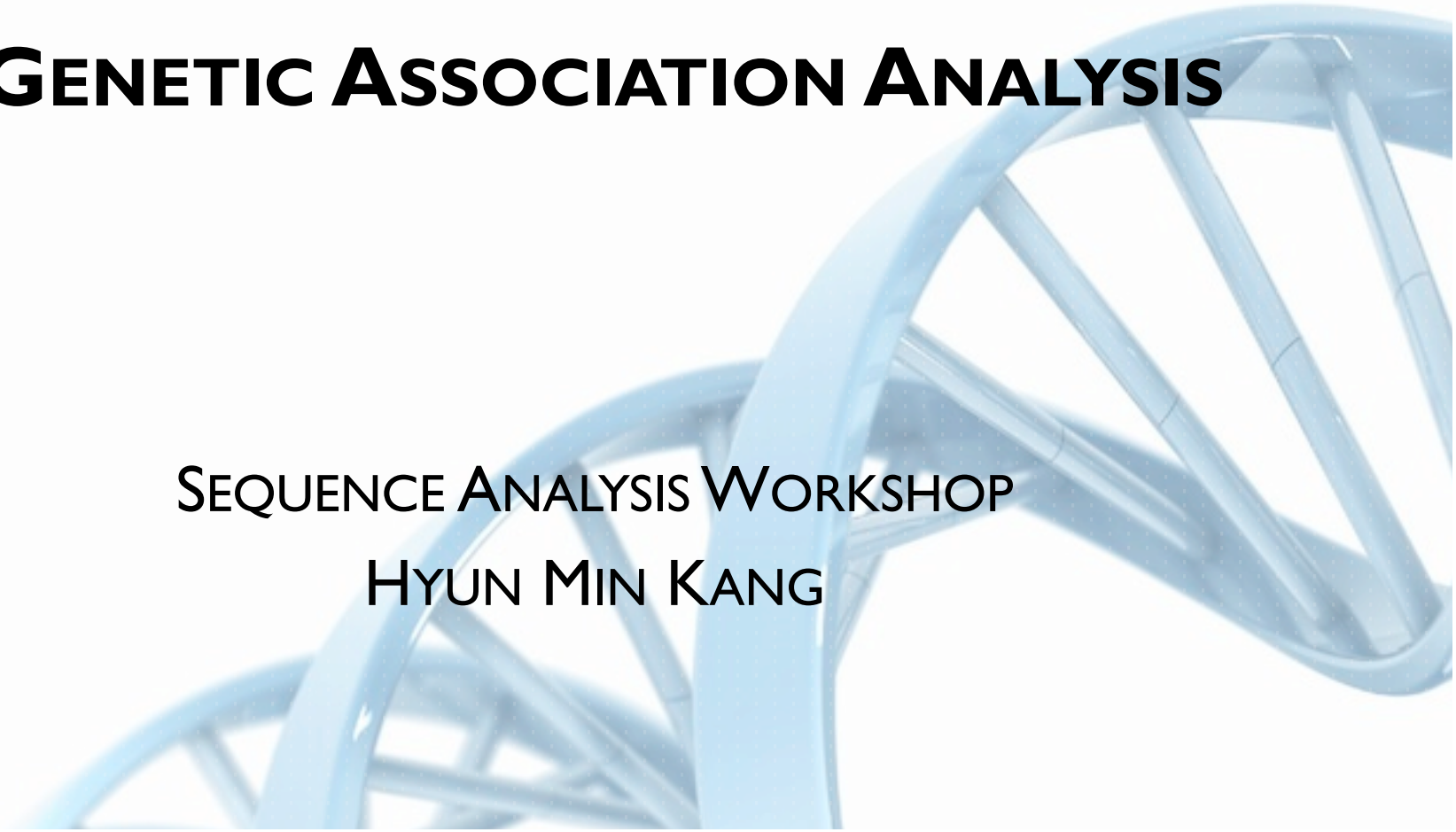
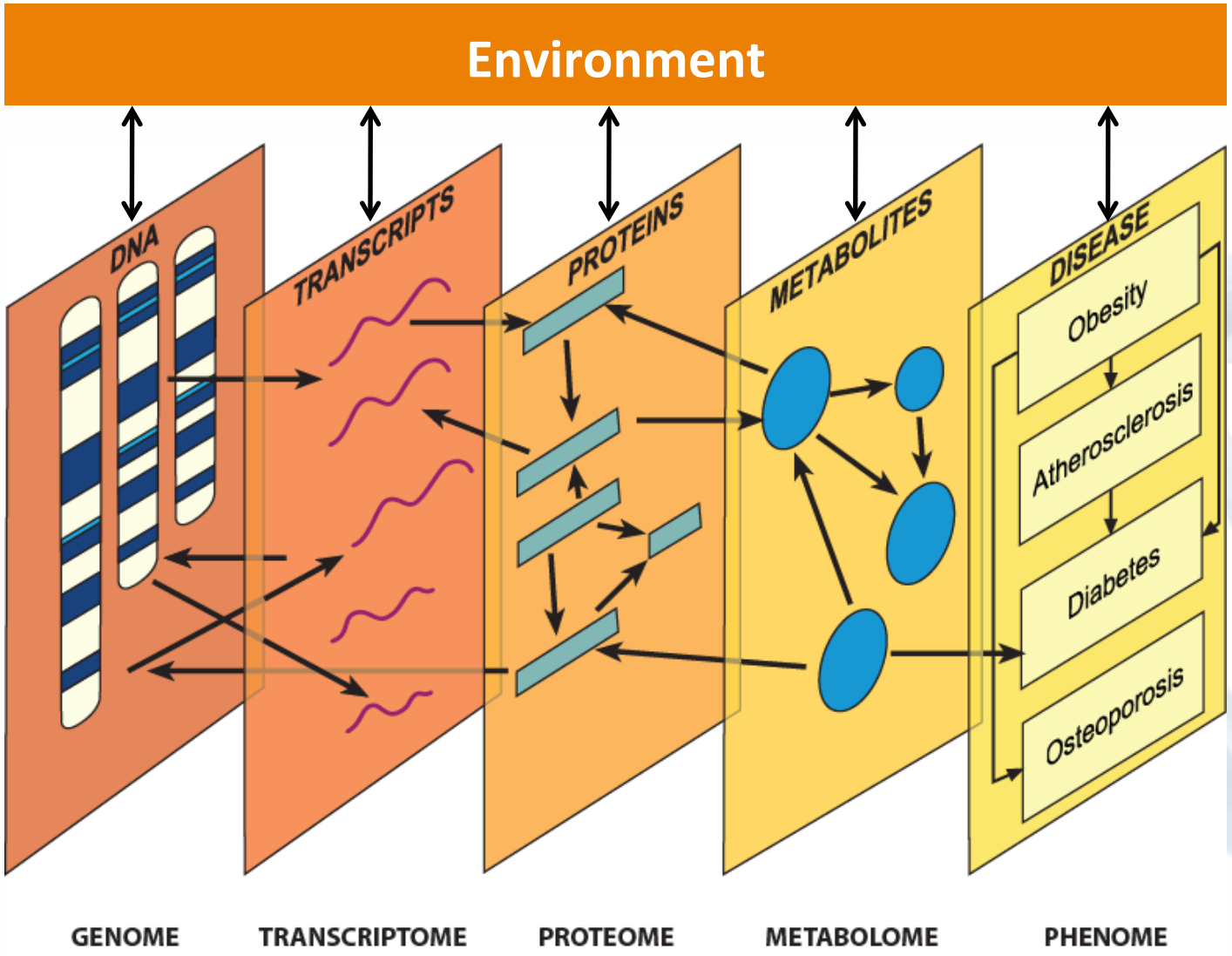
GENETIC ASSOCIATION ANALYSIS



SEQUENCE ANALYSIS WORKSHOP

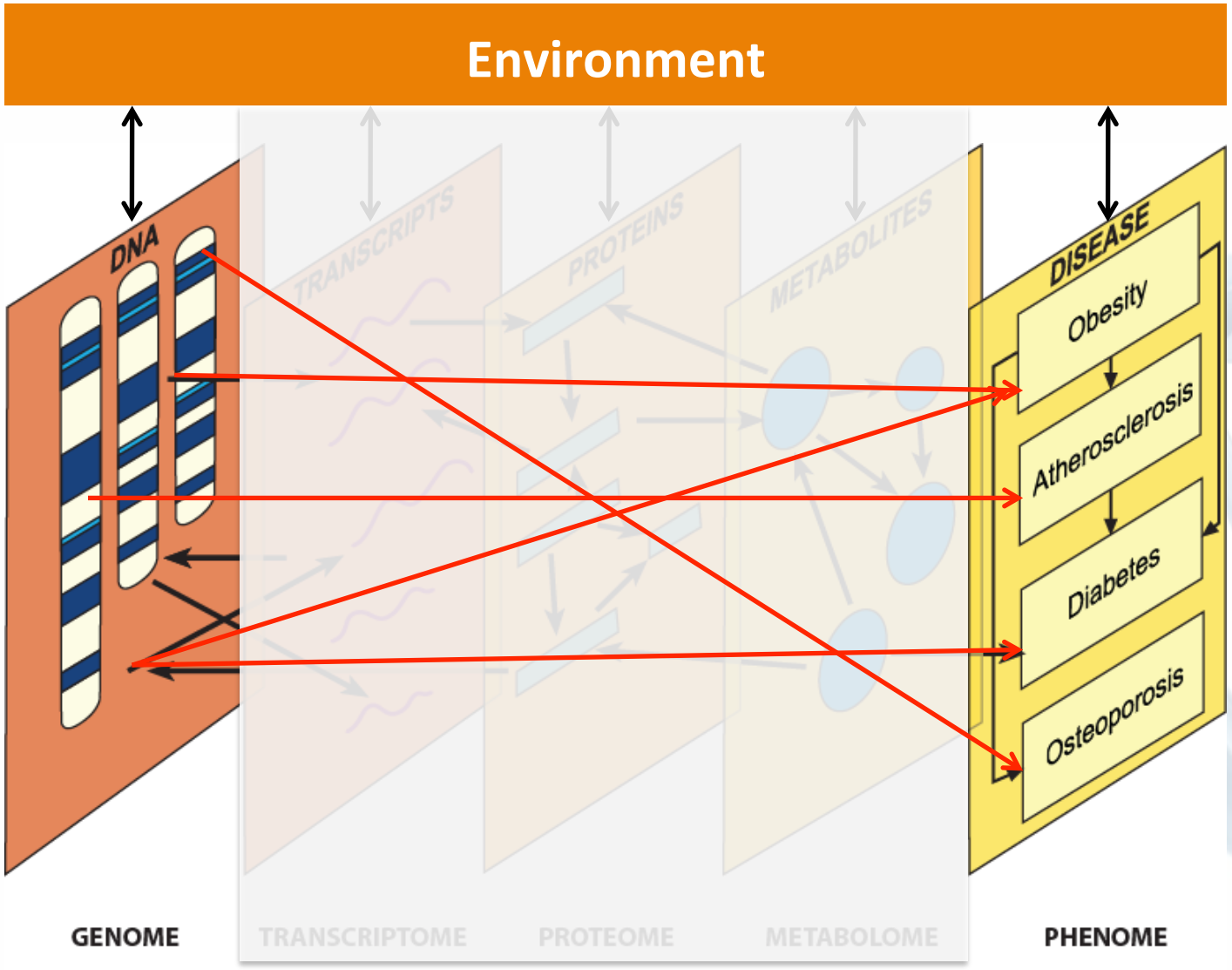
HYUN MIN KANG

GENETIC ARCHITECTURE OF COMPLEX TRAITS



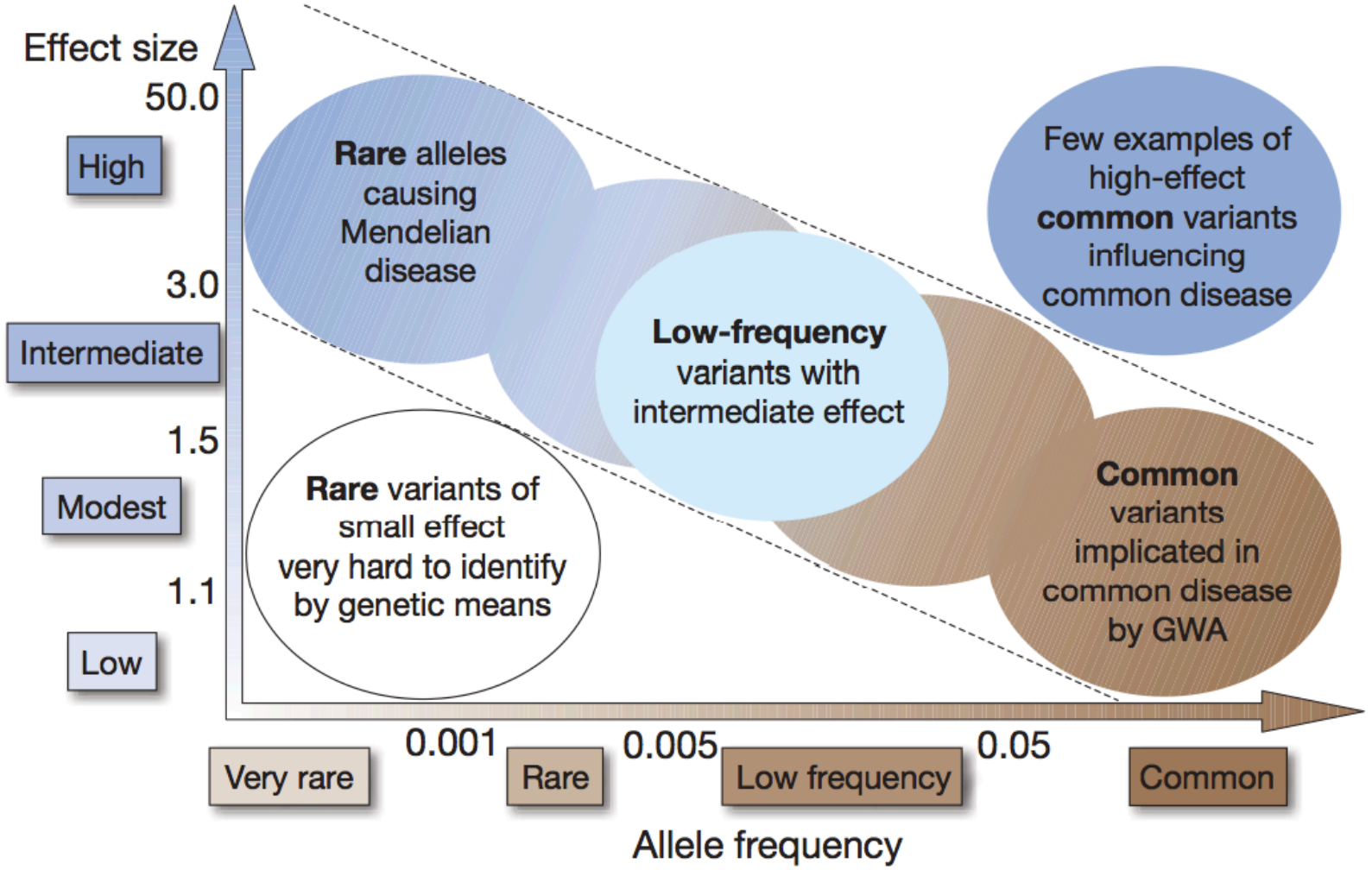
Slide partly due to Aldon Jake Lusis at UCLA

TODAY : GENETIC ASSOCIATION STUDIES

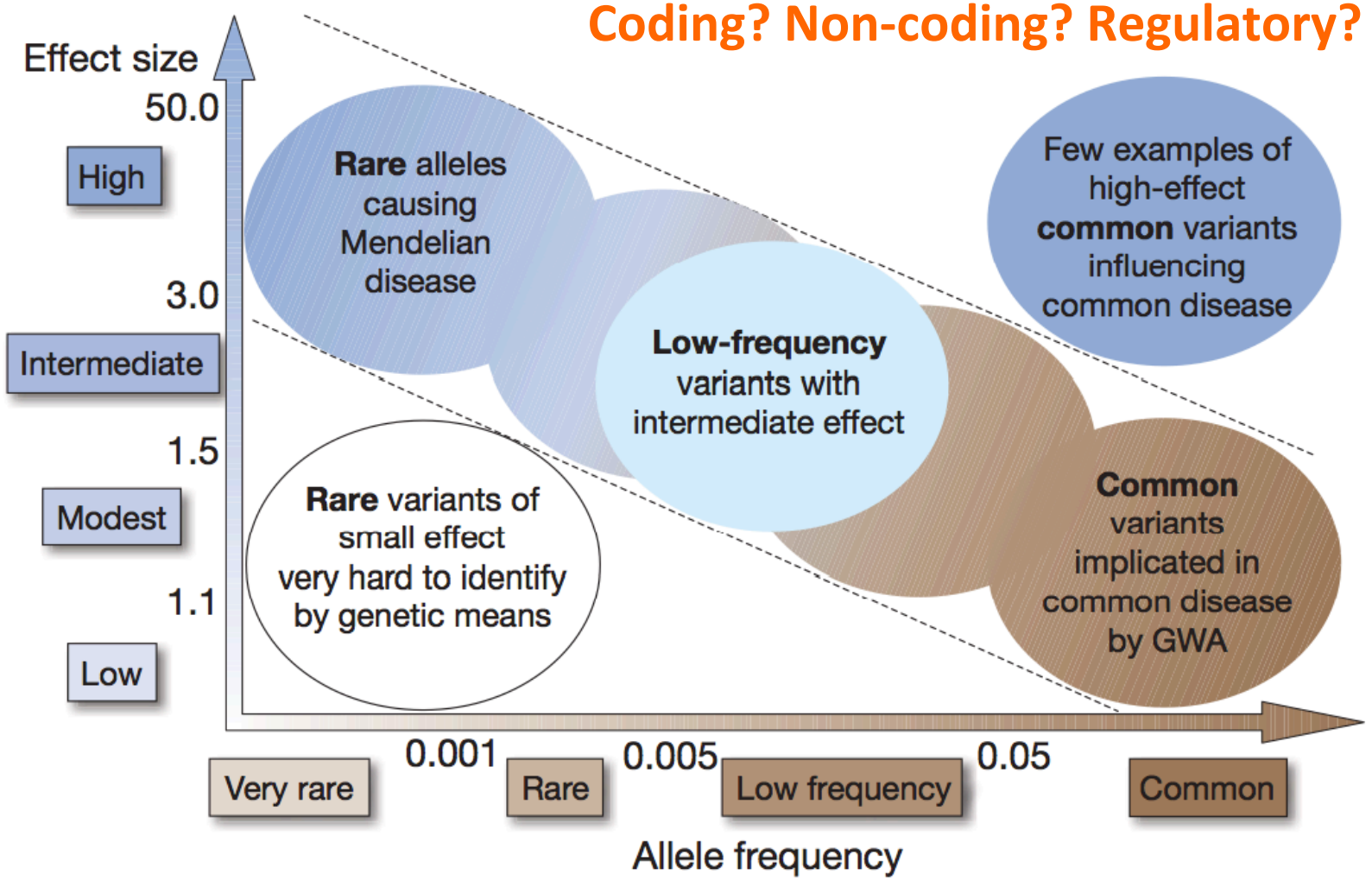


Slide partly due to Aldon Jake Lusis at UCLA

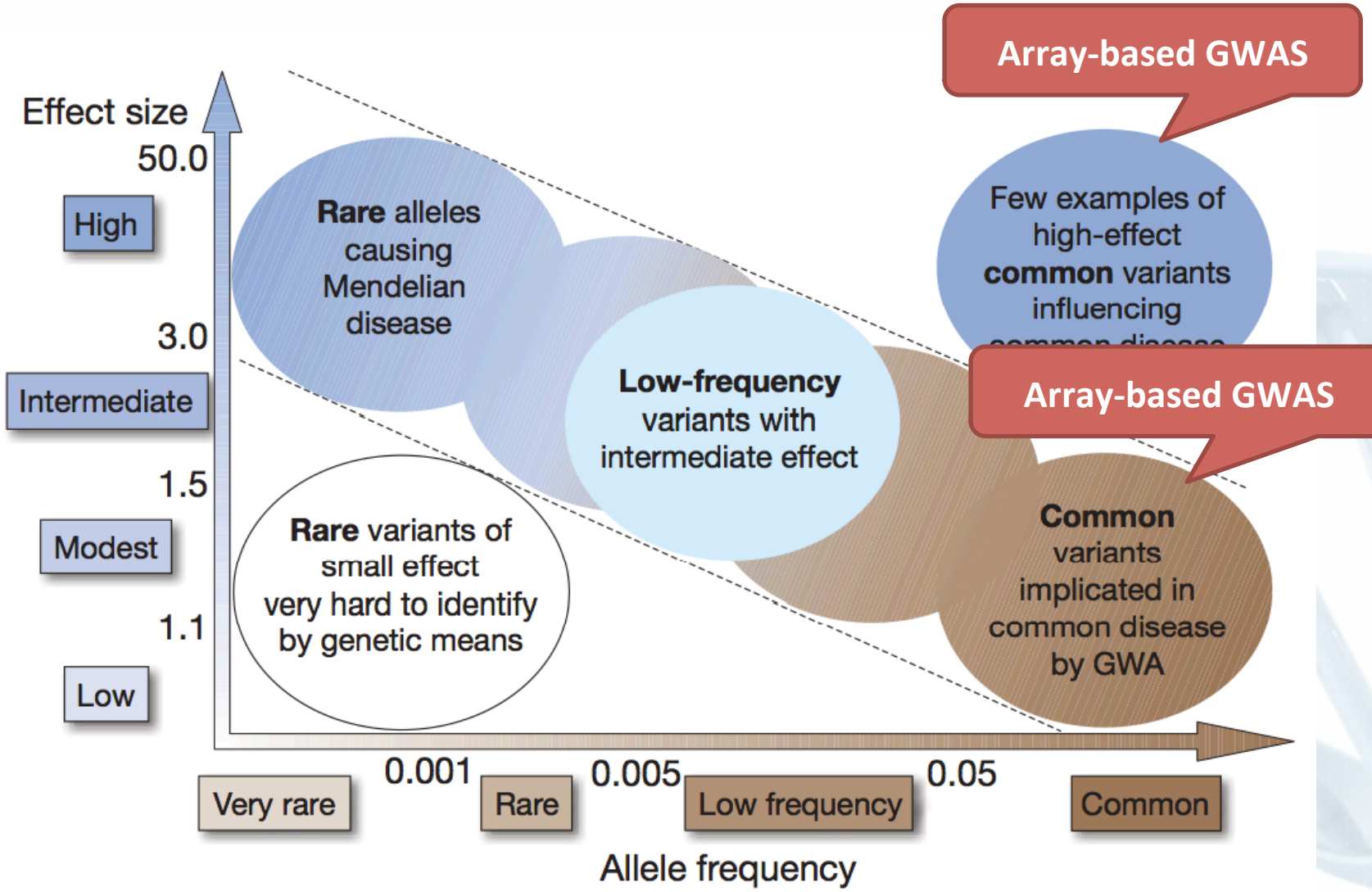
GENETIC ARCHITECTURE OF COMPLEX TRAITS



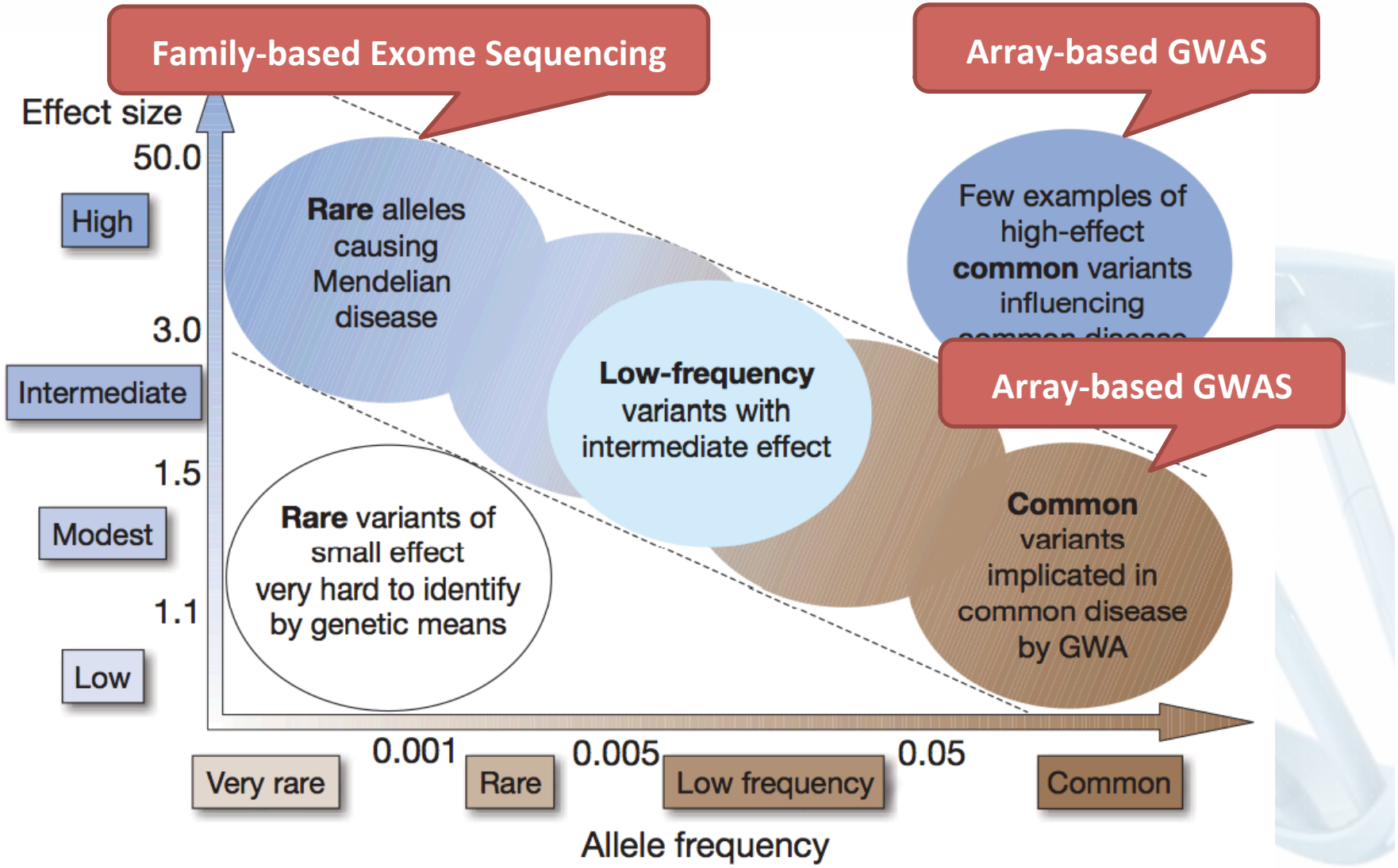
GENETIC ARCHITECTURE OF COMPLEX TRAITS



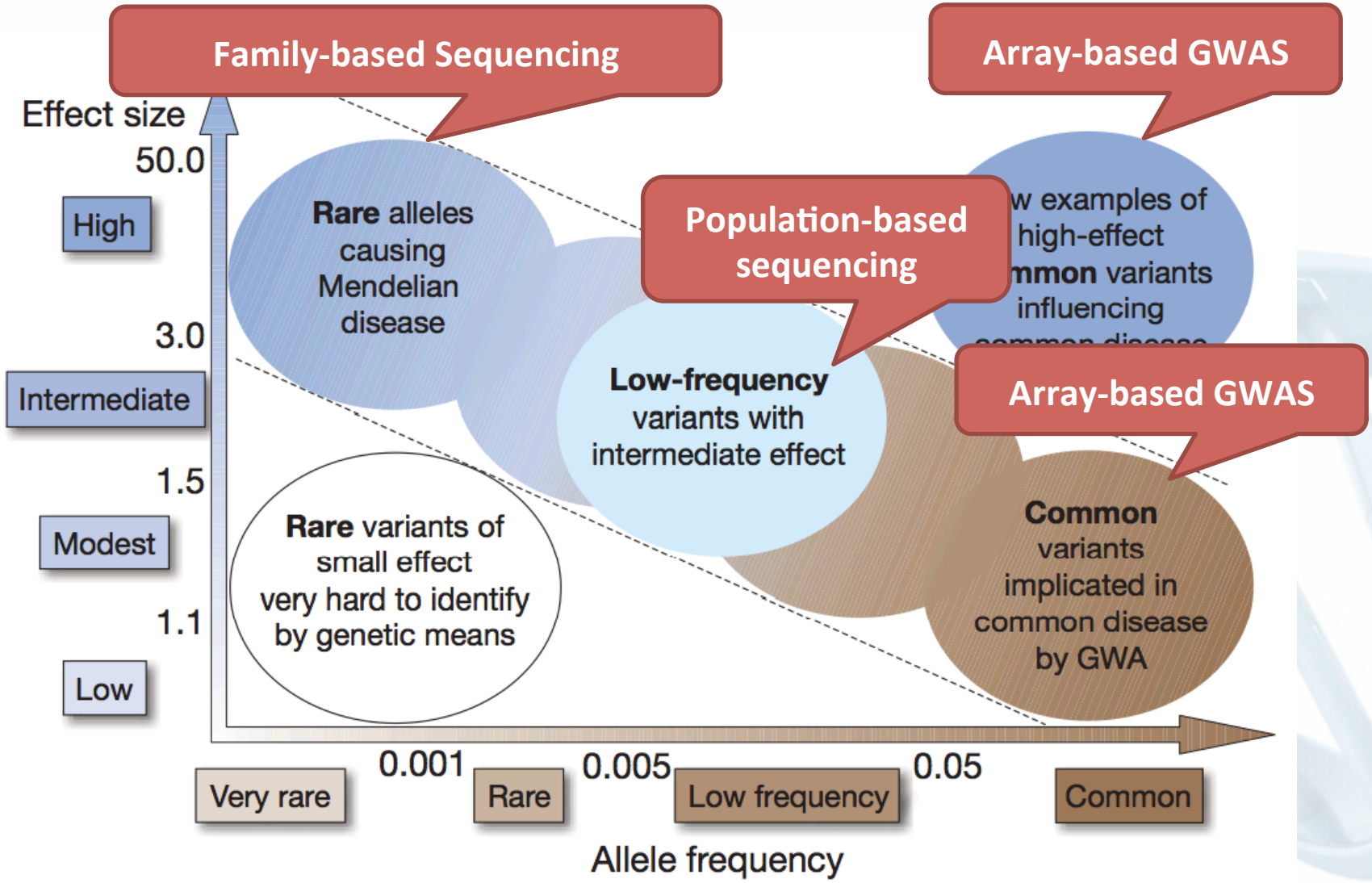
GENETIC ARCHITECTURE OF COMPLEX TRAITS



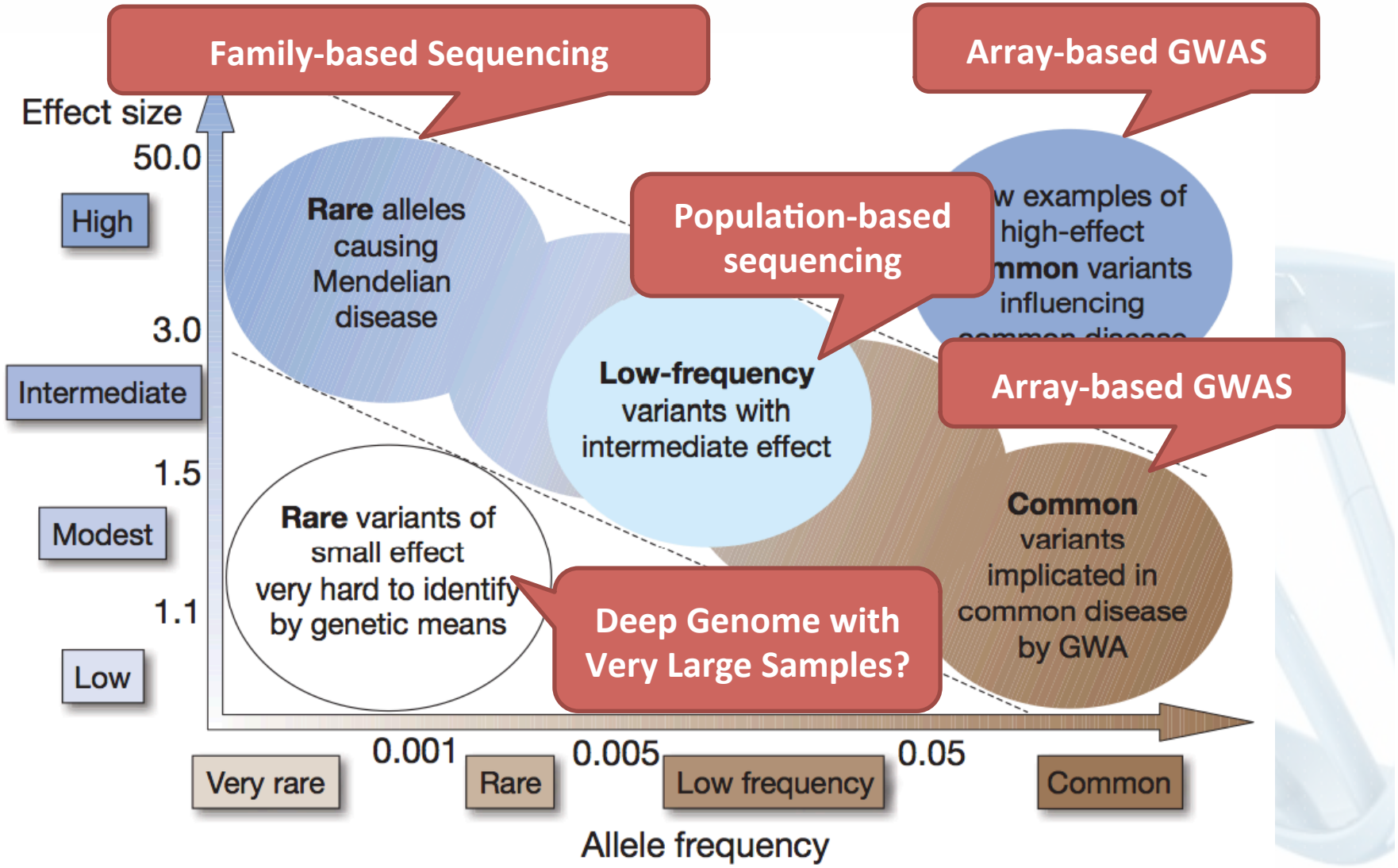
Genetic Architecture of Complex Traits



Genetic Architecture of Complex Traits



Genetic Architecture of Complex Traits

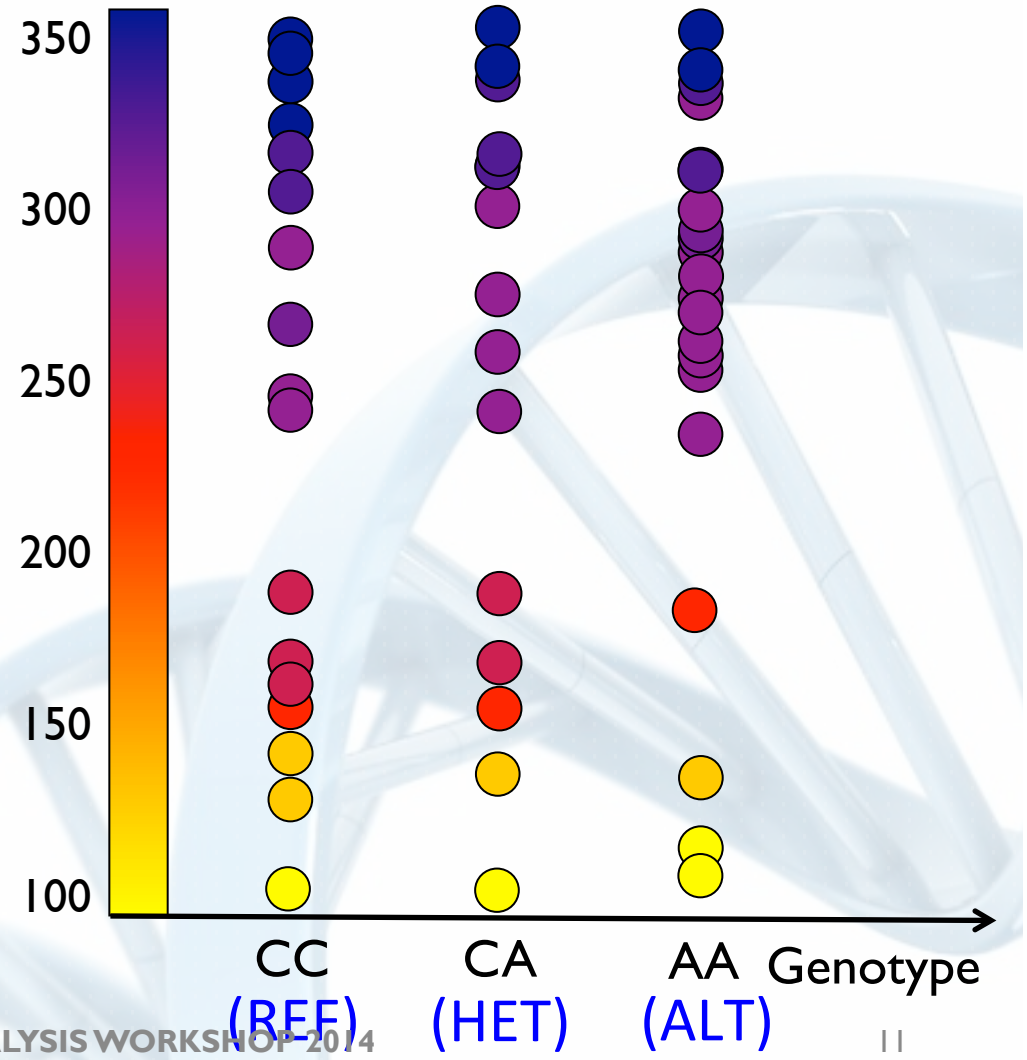


A Single Marker Association Test

- Simplest strategy to detect genetic association
- Compare frequencies of particular alleles, or genotypes, in set of cases and controls
- Typically, use contingency table tests…
  - Chi-squared Goodness-of-Fit Test
  - Cochran-Armitage Trend Test
  - Likelihood Ratio Test
  - Fisher's Exact Test
- … or regression based tests.
  - More flexible modeling of covariates

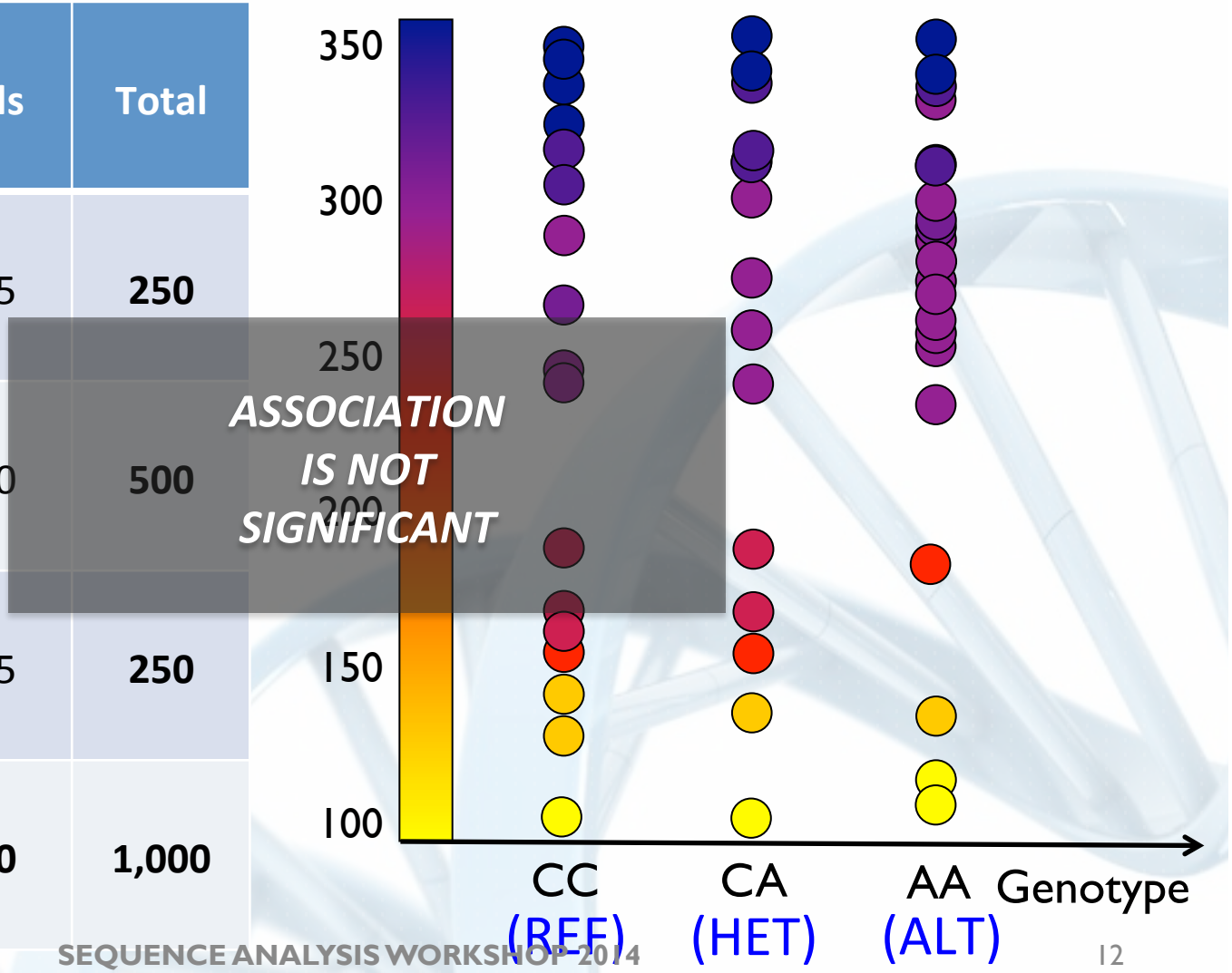
MAPPING GENOTYPE-PHENOTYPE ASSOCIATIONS

Geno-type	Cases	Ctrls	Total
CC (REF)	125	125	**250**
CA (HET)	250	250	**500**
AA (ALT)	125	125	**250**
Total	**500**	**500**	**1,000**



MAPPING GENOTYPE-PHENOTYPE ASSOCIATIONS

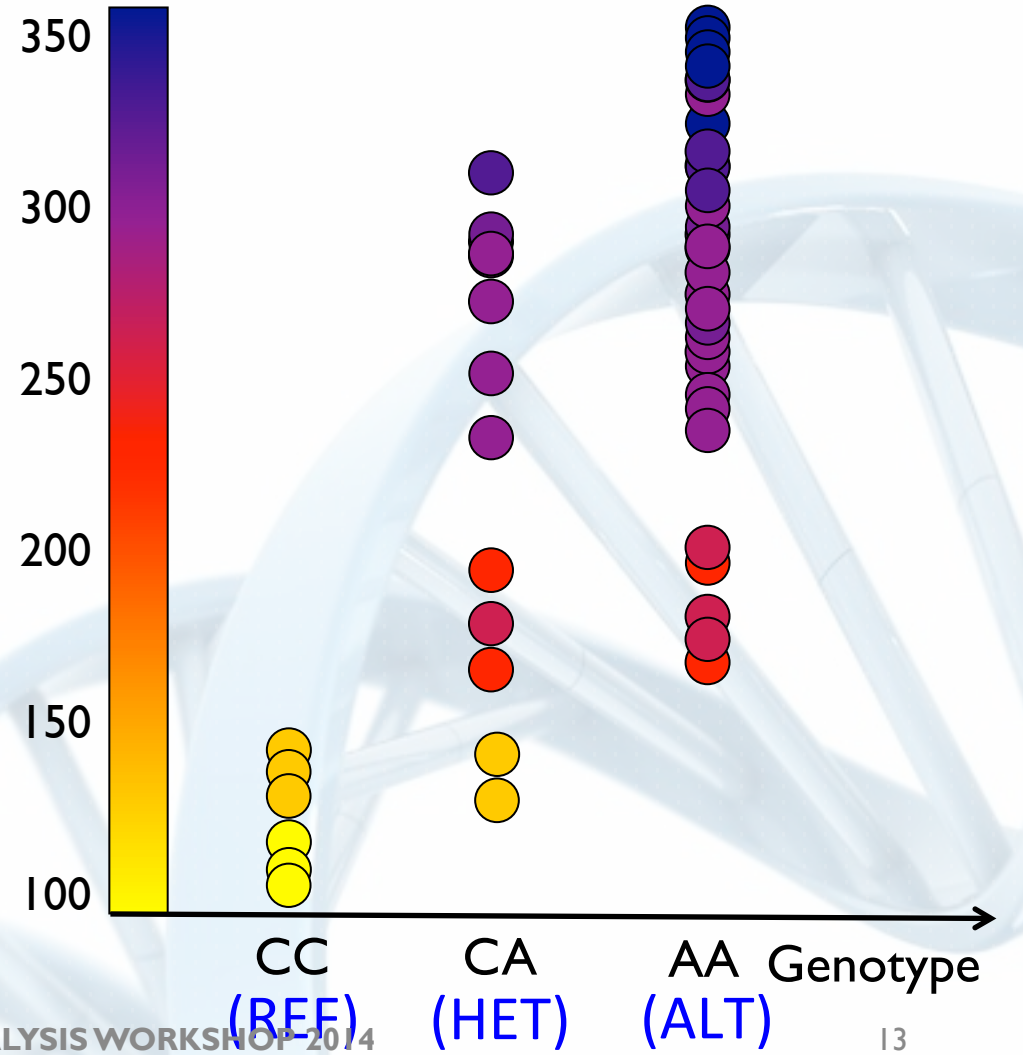
Geno-type	Cases	Ctrls	Total
CC (REF)	125	125	**250**
CA (HET)	250	250	**500**
AA (ALT)	125	125	**250**
Total	**500**	**500**	**1,000**



ASSOCIATION IS NOT SIGNIFICANT

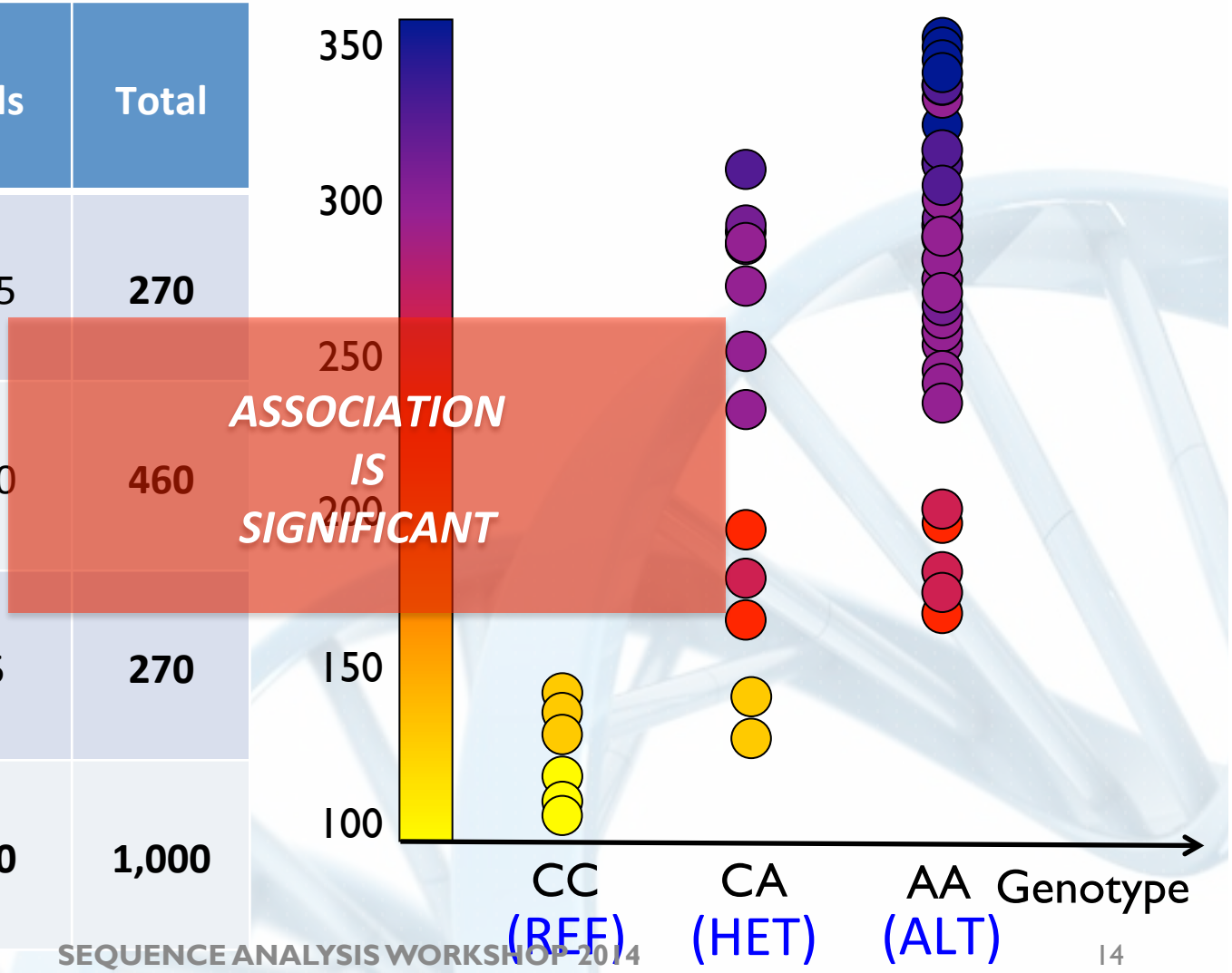
MAPPING GENOTYPE-PHENOTYPE ASSOCIATIONS

Geno-type	Cases	Ctrls	Total
CC (REF)	65	205	**270**
CA (HET)	230	230	**460**
AA (ALT)	205	65	**270**
Total	**500**	**500**	**1,000**

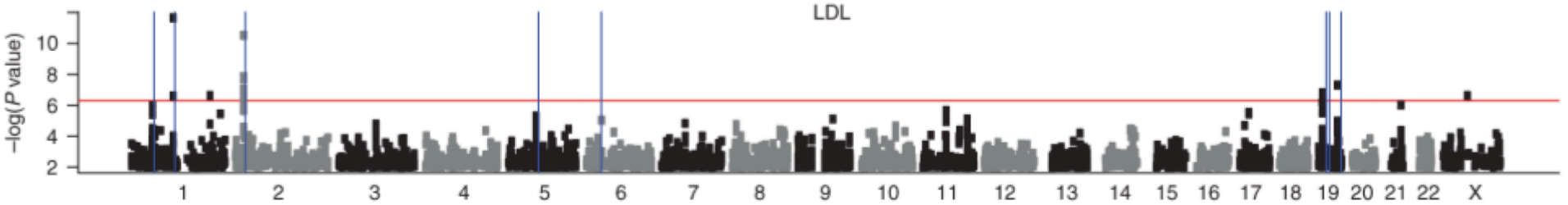


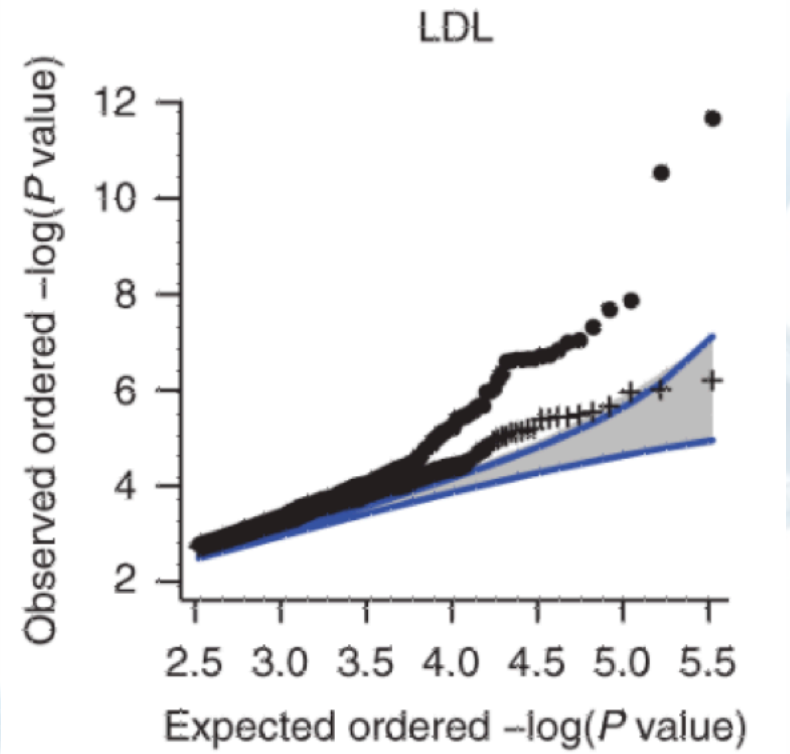
MAPPING GENOTYPE-PHENOTYPE ASSOCIATIONS

Geno-type	Cases	Ctrls	Total
CC (REF)	65	205	**270**
CA (HET)	230	230	**460**
AA (ALT)	205	65	**270**
Total	**500**	**500**	**1,000**



GENOME-WIDE ASSOCIATION STUDIES





Sabatti et. al (2009)

TEST STATISTIC (FOR BALANCED CASE-CONTROL STUDIES)

$$z = \frac{\hat{p}_+ - \hat{p}_-}{\sqrt{[\hat{p}_+(1-\hat{p}_+) + \hat{p}_-(1-\hat{p}_-)]/(2N)}}$$

- $\hat{p}_+$ is observed case allele frequency
- $\hat{p}_-$ is observed control allele frequency
- N is the number of cases and controls

Distribution Under the Null

- Under the null hypothesis $p_+ = p_-$
- Z is distributed as Normal(0, 1) under the null
- Using Inverse Normal Cumulative Distribution Function
- Derive P-value thresholds for target significance level α
  - α = 0.05 leads to cutoff = $-\Phi^{-1}(0.05/2) = 1.96$
  - α = 5×10^{-8} leads to cutoff = $-\Phi^{-1}(5 \times 10^{-8}/2) = 5.45$

Distribution Under the Alternative

- For a specific set of expected case and control allele frequencies..

- We can calculate the expected value of test statistic

$$\mu = \frac{p_+ - p_-}{\sqrt{[p_+(1-p_+) + p_-(1-p_-)]/(2N)}}$$

- Under the alternative, statistic is Normal(μ,1)

POWER

- To calculate power, we first calculate:
  - Significance threshold ***C***
  - Expected test statistic **μ**

- Use normal cumulative distribution function **Φ**

- $$\Pr(|Z| > C) = \Pr(Z > C) + \Pr(Z < -C)$$
$$= 1 - \Phi(C - \mu) + \Phi(-C - \mu)$$

Power calculation is important for designing association studies

Sources of Association

- Causal association
  - Genetic marker alleles influence susceptibility

best

- Linkage disequilibrium
  - Genetic marker alleles associated with other nearby alleles that influence susceptibility

useful

- Population stratification
  - Genetic marker is unrelated to disease alleles

misleading

Example of Spurious Association due to Population Stratification

	Population 1	Population 2	Combined
Allele Frequencies			
p_1	0.20	0.80	0.50
p_2	0.80	0.20	0.50
Genotype Frequencies			
p_{11}	0.04	0.64	0.34 (0.25 Expected)
p_{12}	0.32	0.32	0.32 (0.50 Expected)
p_{22}	0.64	0.04	0.40 (0.25 Expected)

Example of Spurious Association due to Population Stratification

Population 1

	Allele 1	Allele 2
Affected	50	200
Unaffected	25	100

$\chi^2 = 0.00$ p-value = 1.0

Population 2

	Allele 1	Allele 2
Affected	100	25
Unaffected	200	50

$\chi^2 = 0.00$ p-value = 1.0

Combined

	Allele 1	Allele 2
Affected	150	225
Unaffected	225	150

$\chi^2 = 29.2$ p-value = 6.5×10^{-8}

THE STRATIFICATION PROBLEM HAPPENS..

- If..
  - Phenotypes differ between populations
  - and allele frequencies have drifted apart
- Then..
  - Unlinked markers exhibit association
  - Not very useful for gene mapping!
- For example, Glaucoma has prevalence of ~2% in elderly Caucasians, but ~8% in African-Americans

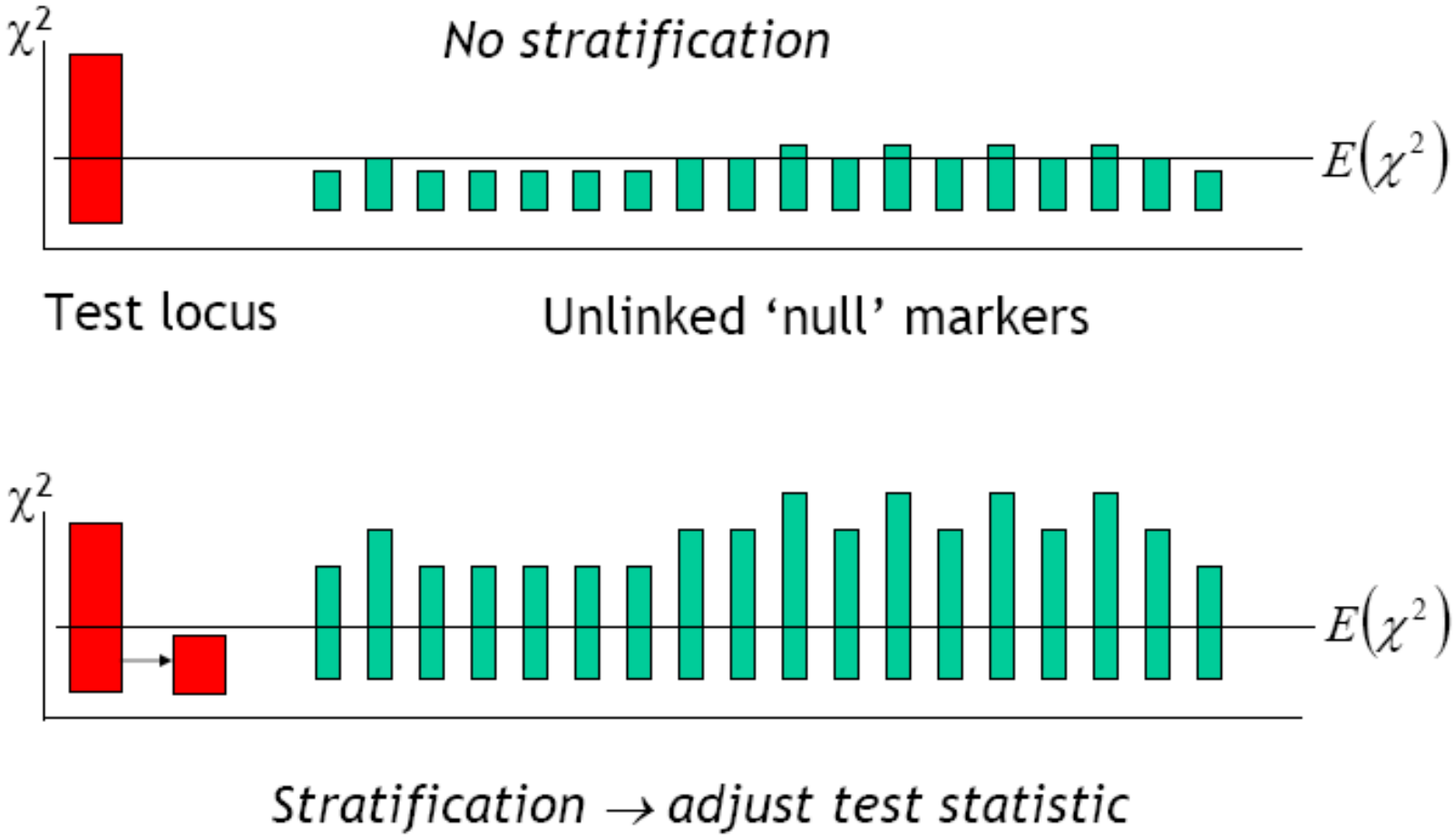
Possible Solutions For Population stratification

- Avoid stratification by design
  - Collect a better matched sample by ancestry
  - Use family-based controls
    - and apply Transmission Disequilibrium Test (TDT)

- Analyze association by population groups
  - Using self reported ethnicity or genetic markers
  - Carry out association analysis within each group

- Account for inflated false-positive rate
  - Many different ways exist

GENOMIC CONTROL



(Figure courtesy Shaun Purcell, Harvard, and Pak Sham, HKU)

Define Inflation Factor

- Compute chi-squared for each marker
- Inflation factor λ
  - Average observed chi-squared
  - Median observed chi-squared / 0.456
  - Should be >= 1
- Adjust statistic at candidate markers
  - Replace χ^2_{biased} with $\chi^2_{fair} = \chi^2_{biased}/\lambda$

QUESTIONS

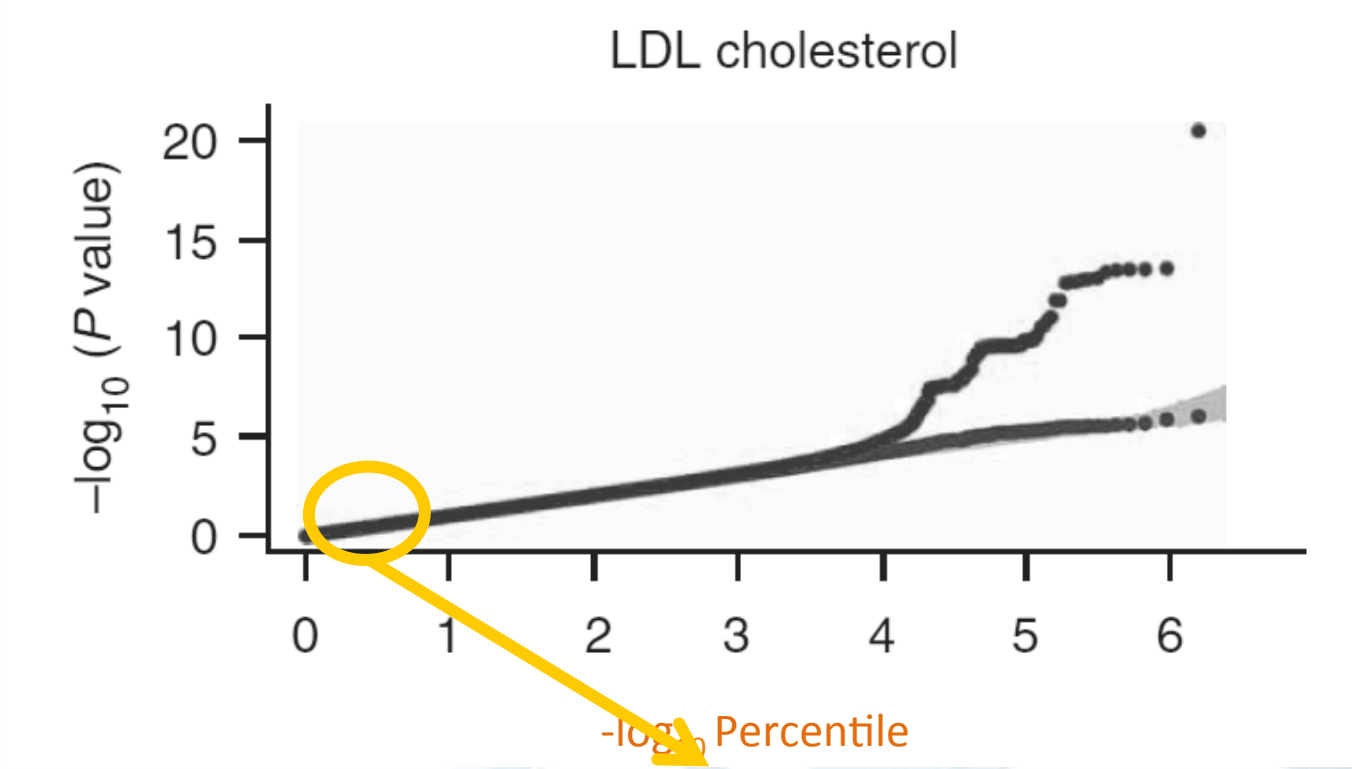
- When defining the inflation factor λ …
- Why do we use a lower bound of 1?
- What might be the advantages of using the median rather than the mean?

Applying Genomic Control

- Simple and convenient approach…
  - Easily adapted to other test statistics, such as those for quantitative trait and haplotype tests

- Under the null, stratification always inflates evidence for association…
  - Is this also true under the alternative?
  - What might be the consequences?

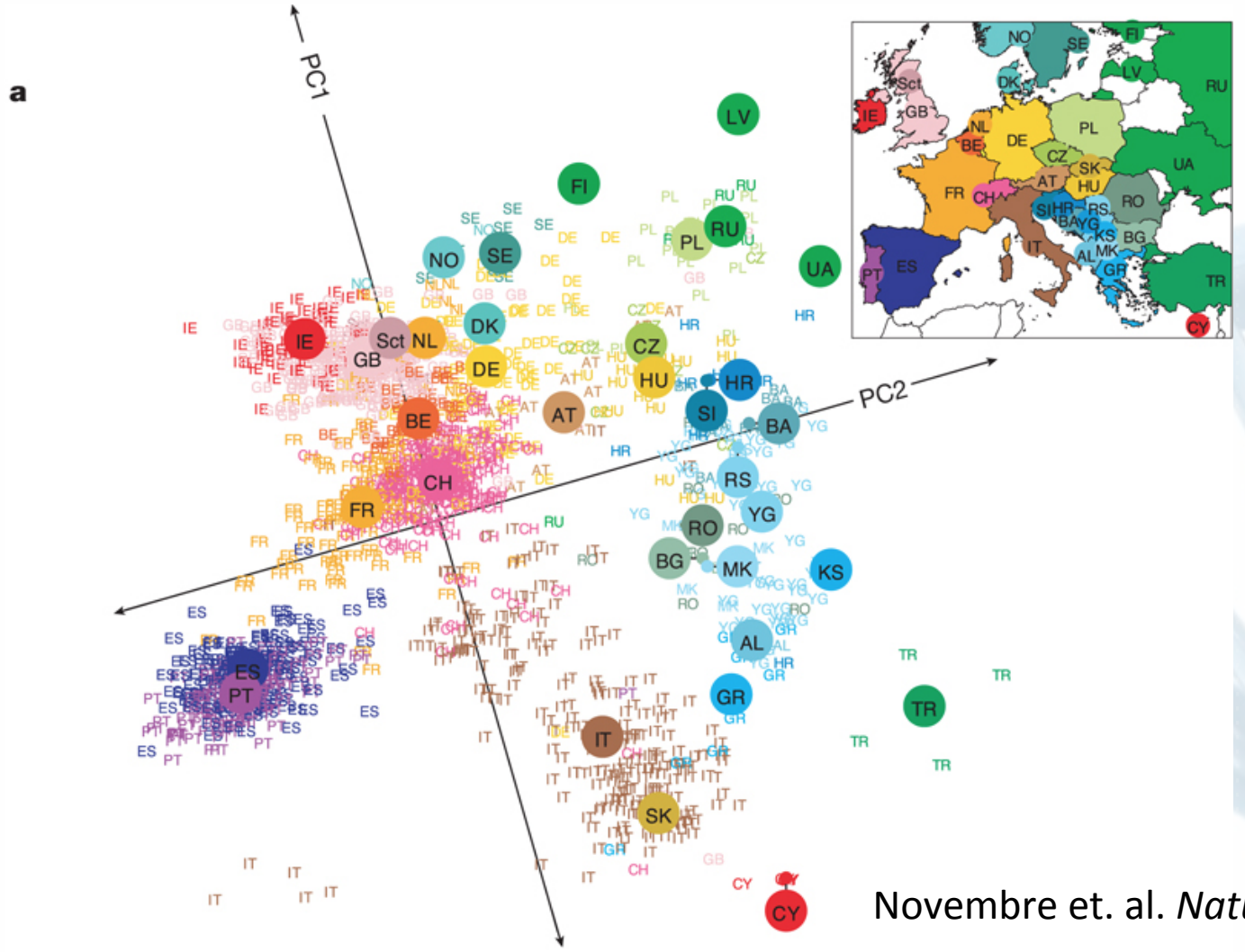
Q-Q Plots: A Useful Diagnostic



The genomic control value examines markers with little evidence for association. If these large p-values were to deviate from expected, there is a problem! In this case, λ=1.02.

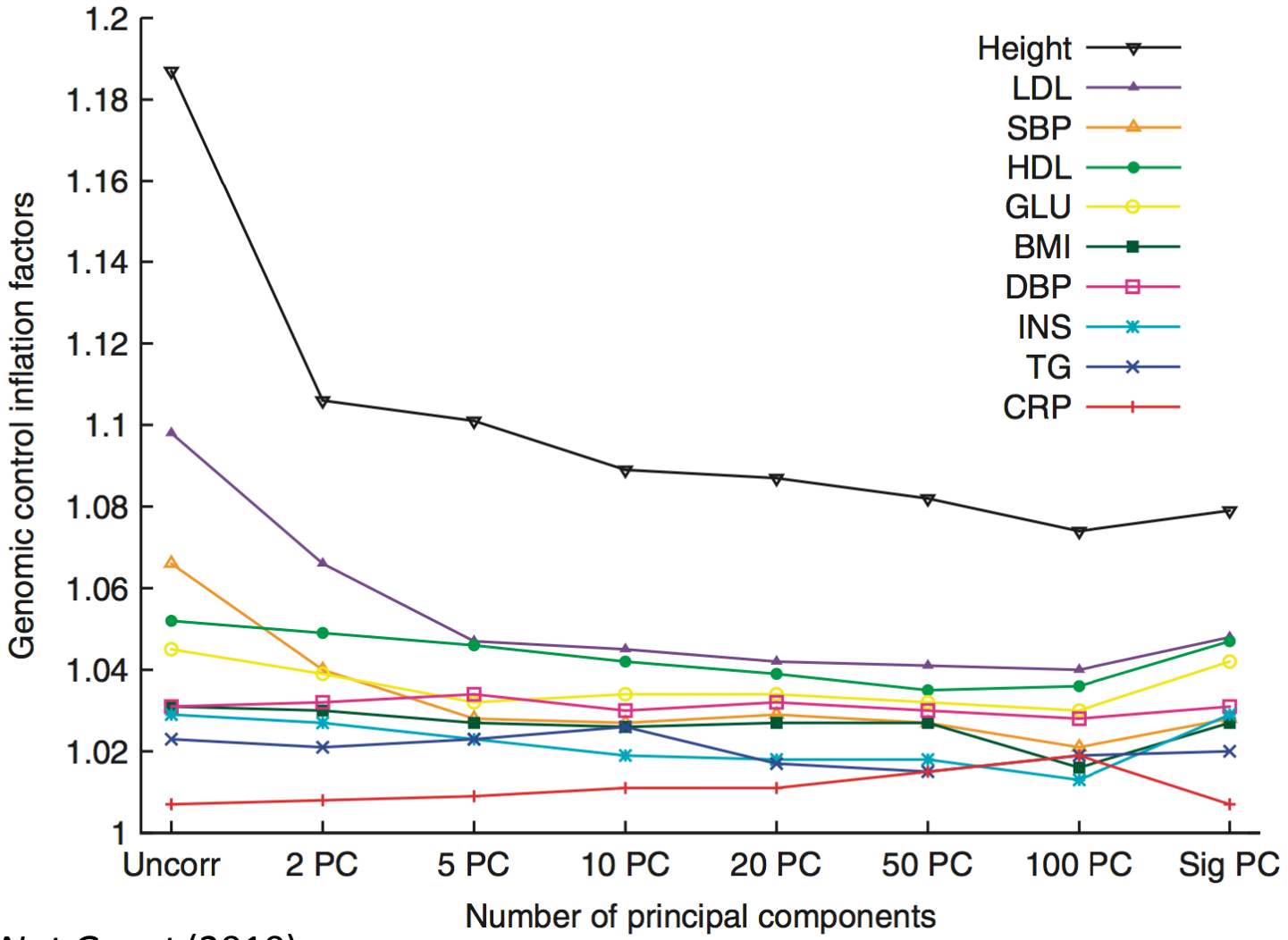
Willer et al, *Nature Genetics*, 2008

Principal Components Mirror European Geography



Novembre et. al. *Nature* 2008

Correcting for Population Structure using Principal Components



Kang et. al. *Nat Genet* (2010)

VARIANCE COMPONENT MODEL FOR FAMILY-BASED ASSOCIATION TEST

- Population-based analysis assumes uncorrelated phenotypes between individuals under the null

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I)$$

VARIANCE COMPONENT MODEL FOR FAMILY-BASED ASSOCIATION TEST

- Population-based analysis assumes uncorrelated phenotypes between individuals under the null

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I)$$

- Family-based analysis assumes phenotypes are correlated with relatives' phenotypes

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma_g^2 K + \sigma_e^2 I)$$ K_{ij} : kinship coefficient

VARIANCE COMPONENT MODEL FOR FAMILY-BASED ASSOCIATION TEST

- Population-based analysis assumes uncorrelated phenotypes between individuals under the null

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I)$$

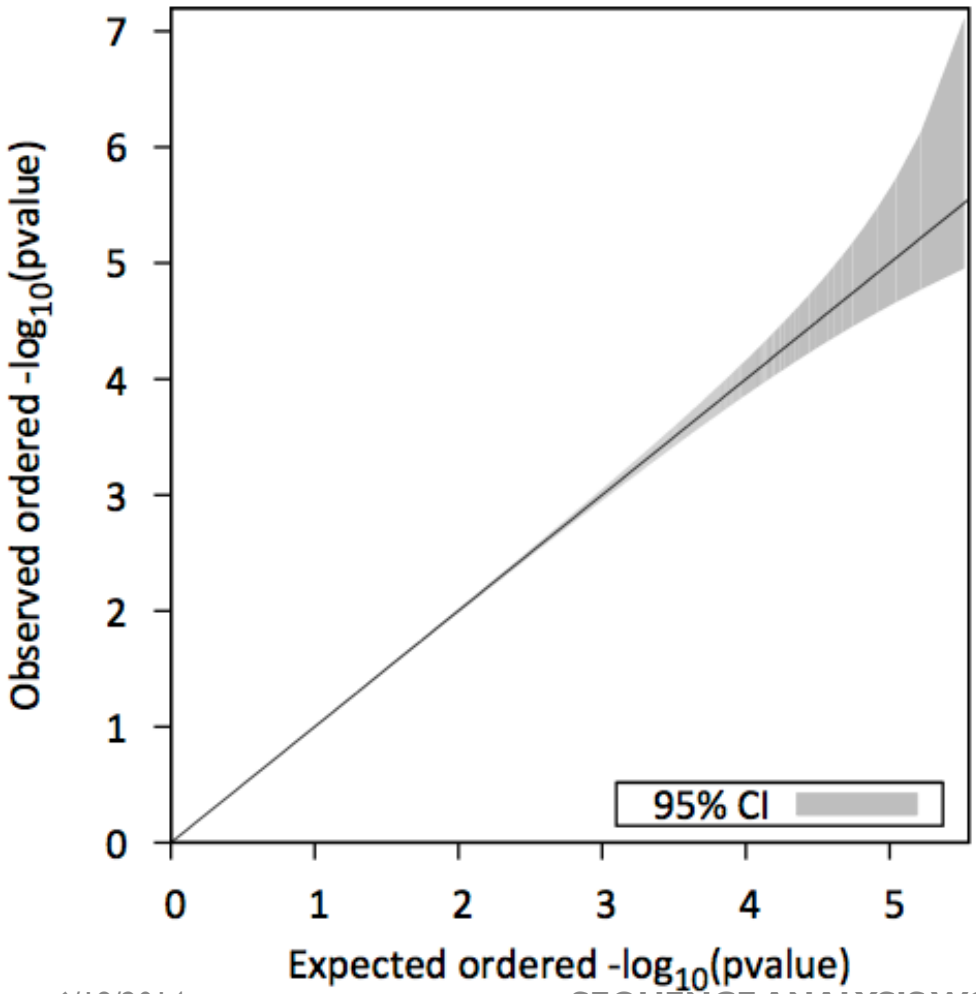
- Family-based analysis assumes phenotypes are correlated with relatives' phenotypes

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma_g^2 K + \sigma_e^2 I)$$ K_{ij} : kinship coefficient

- Similar model for population-based analysis to account for distant relationship inferred from dense SNP arrays

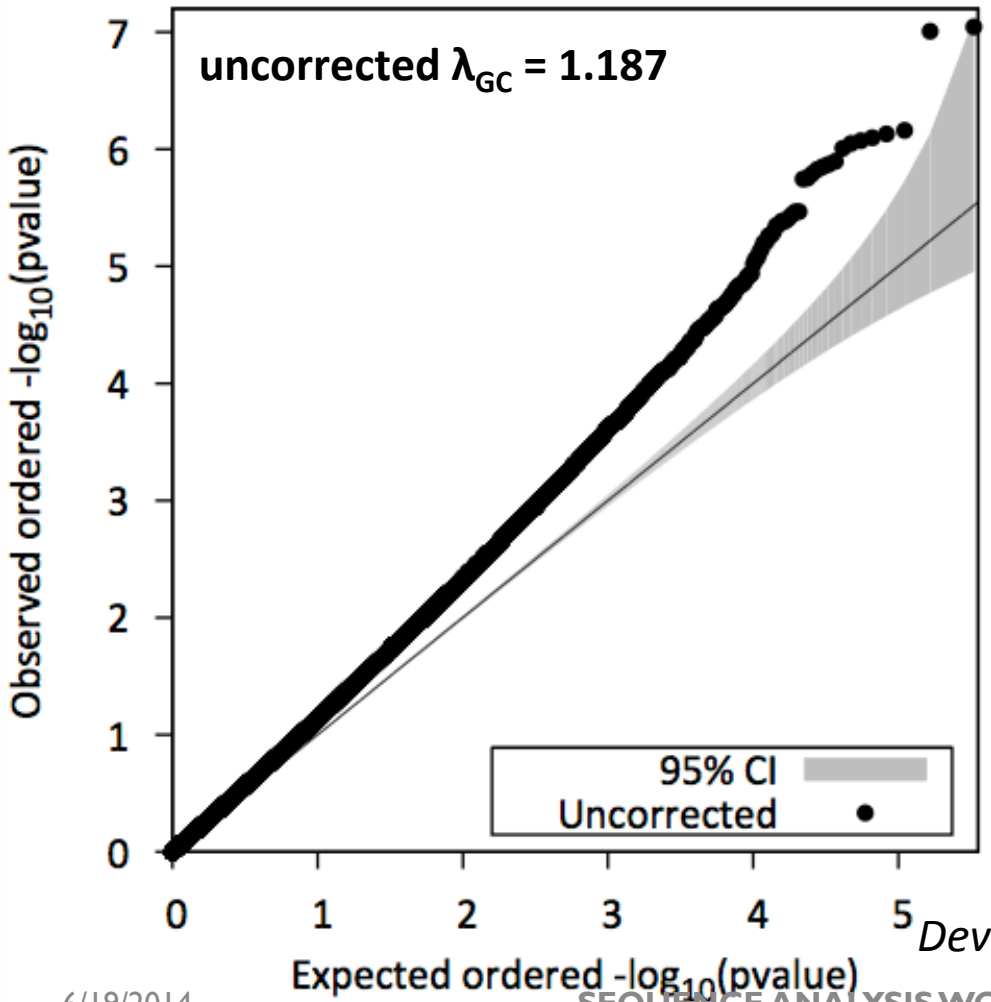
$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma_g^2 \hat{K} + \sigma_e^2 I)$$ $\hat{K}_{ij}$: marker-based kinship coefficient

GENOME-WIDE ASSOCIATION OF HUMAN HEIGHT



- NFBC 1966 birth cohort
  - *Sabatti et al, Nat Genet (2008) 41:35-46*
- Illumina 370,000 SNPs
- 5,326 unrelated individuals

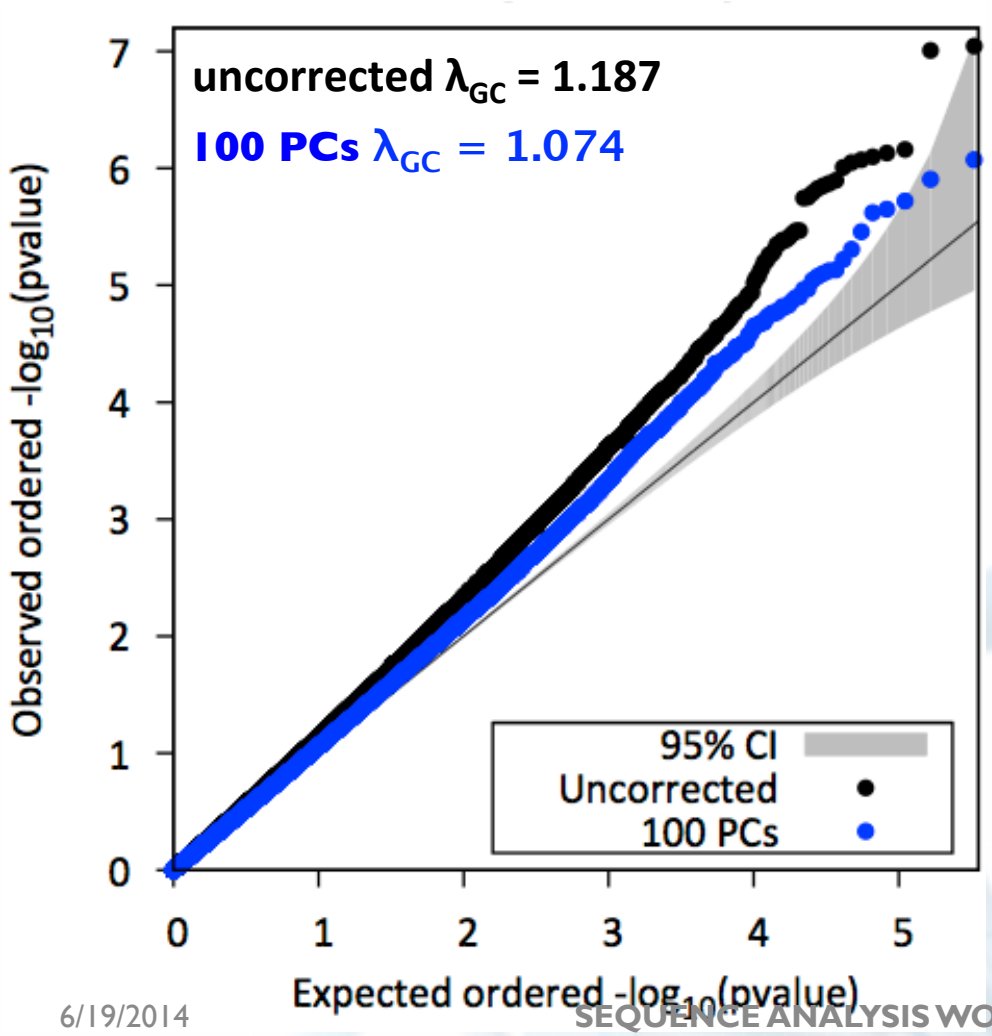
UNCORRECTED ANALYSIS
- OVERDISPERSION OF TEST STATISTICS -



$$\lambda_{GC} = \frac{\text{median}\{T_1, T_2, \cdots, T_n\}}{\mathbf{E}[\text{median}\{T\}]}$$

Devlin & Roeder Biometrics (1999) 55:997-1004

CONDITIONING ON PRINCIPAL COMPONENTS
- OVERDISPERSION STILL EXISTS -

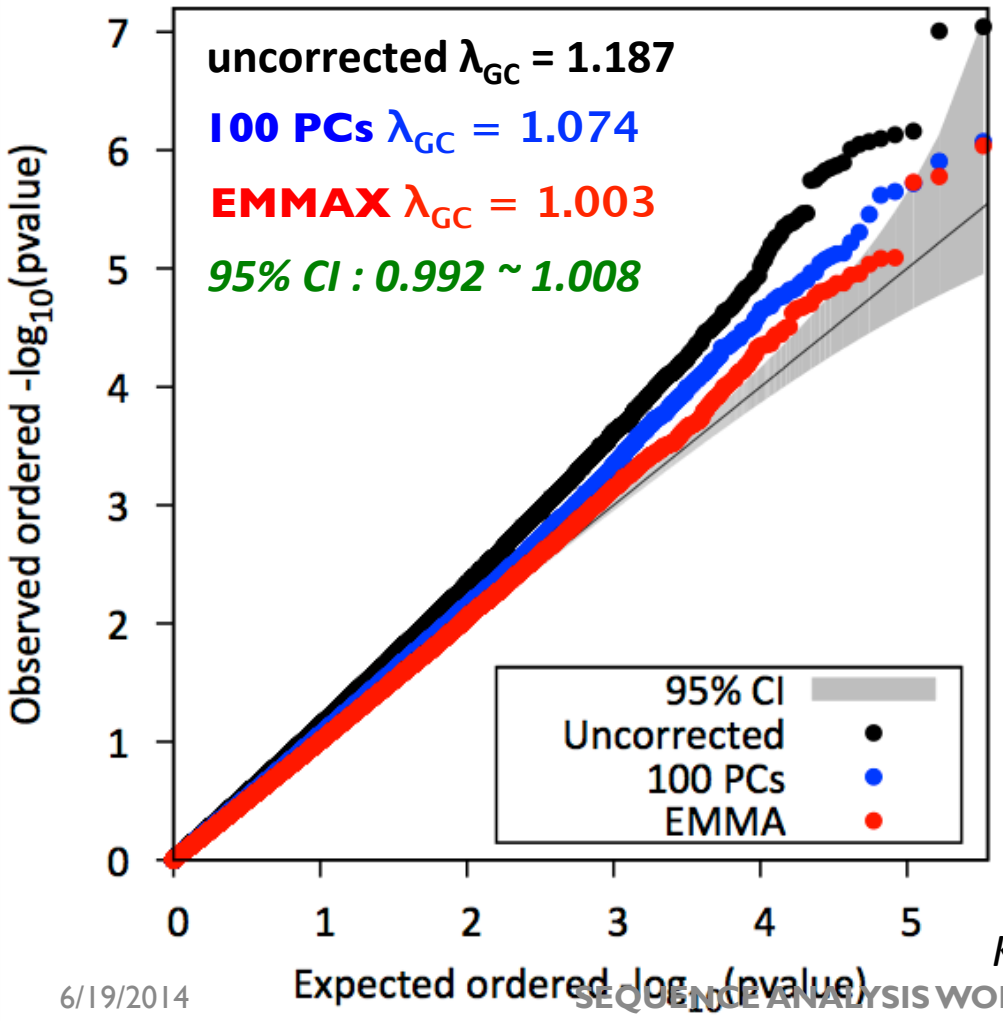


$$\mathbf{y} = \mu + \mathbf{x}\beta + G\gamma + \mathbf{e}$$

- G is top k(=100) eigenvectors of kinship matrix K
- λ_{GC} from 1.187 to 1.074
- λ_{GC} is still substantially higher than expected
- Corrects for population structure, but not hidden relatedness

Price AL et al, Nat Genet (2006) 38:904-909

VARIANCE COMPONENT MODEL
- OVERDISPERSION RESOLVED -



$$\begin{aligned} \mathbf{y} &= \mu + \mathbf{x}\beta + \mathbf{u} + \mathbf{e} \\ \mathrm{Var}(\mathbf{u}) &= \sigma_g^2 K \\ \mathrm{Var}(\mathbf{e}) &= \sigma_e^2 I \end{aligned}$$

- Using EMMAX reduced λ_{GC} from 1.187 to 1.003
- λ_{GC} falls into 95% confidence intervals

Kang HM et al, Nat Genet (2010) 42:348-54

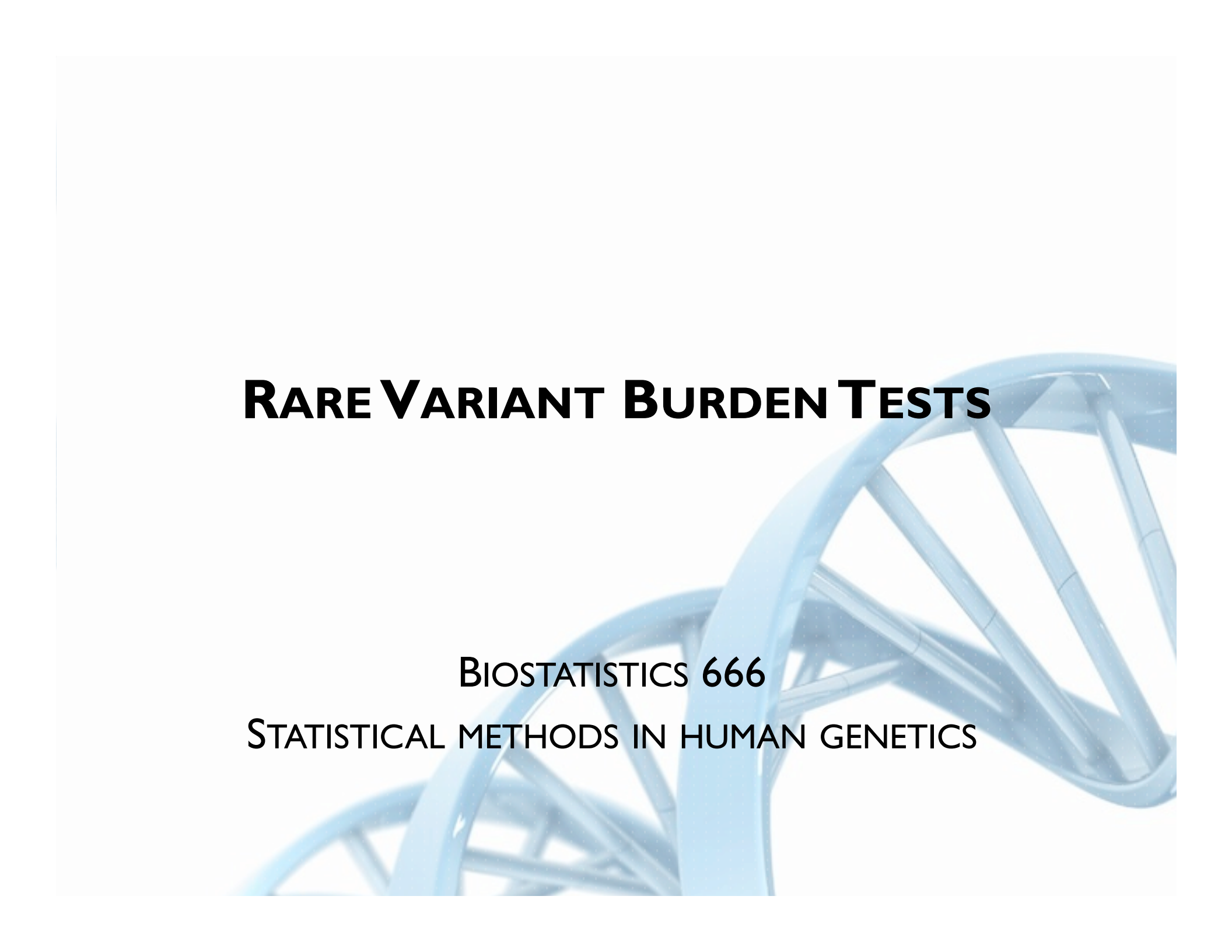
SUMMARY

- Genome-wide single variant test can identify regions of genome associated with disease traits
- Understanding power of your study design based on the genetic architecture of traits are important.
- Accounting for population structure and cryptic relatedness is important to avoid misleading results

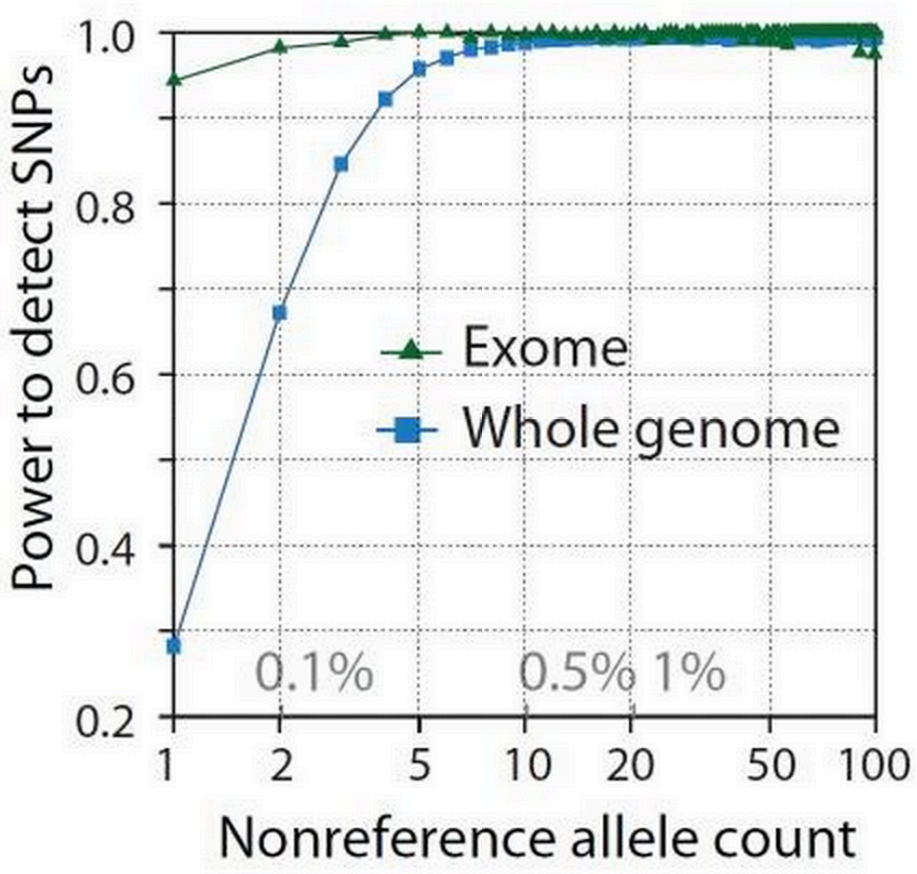
RARE VARIANT BURDEN TESTS

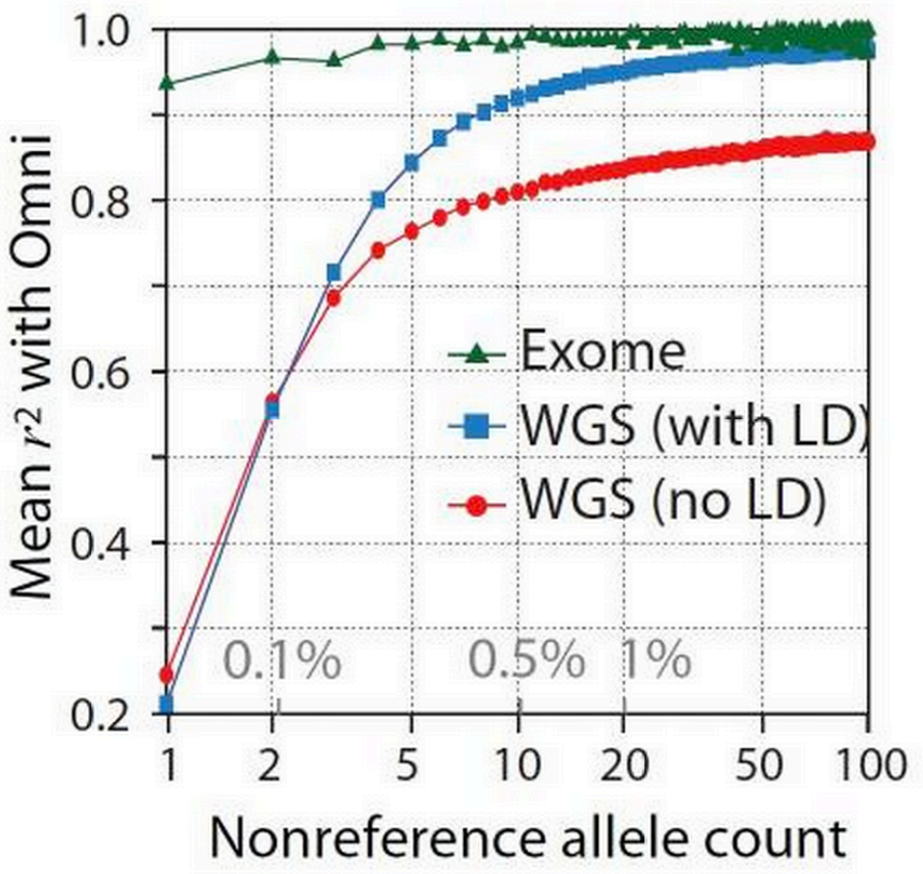
BIOSTATISTICS 666

STATISTICAL METHODS IN HUMAN GENETICS



POWER TO DETECT VARIANTS FROM SEQUENCE DATA





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
  - Rare varia… …entiful and should identify new susceptibility loci

Coding Variants Especially Useful!

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

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

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 whole genome sequencing
  - Ca
  - Ver
- 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

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

ASSOCIATION TEST OF SINGLE RARE VARIANT

- 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^{-7}
  - 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 two-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

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

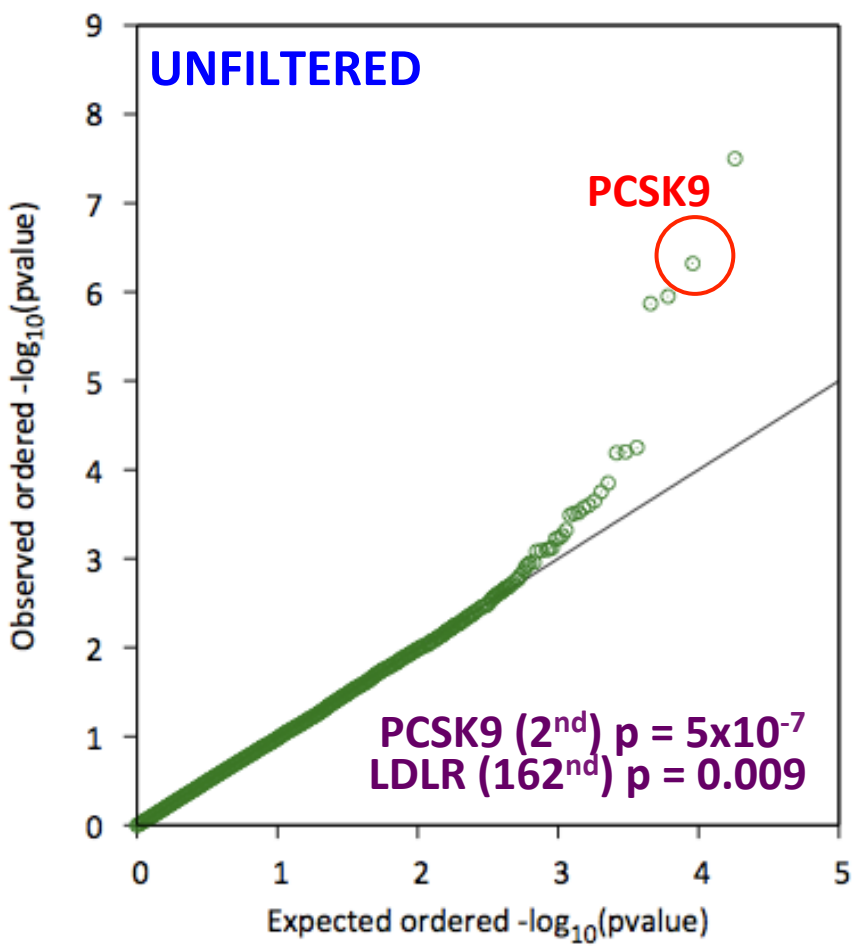
Li and Leal (2008) *Am J Hum Genet* **83:**311-321

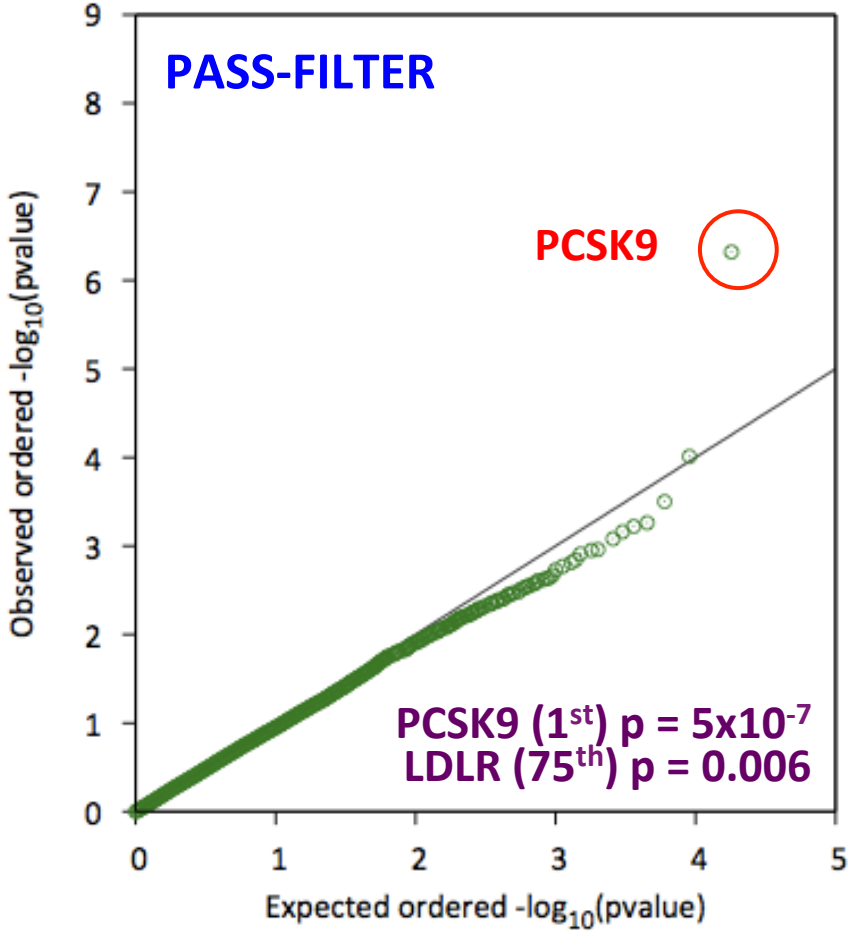
EXOME SEQUENCING PROJECT

- 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%

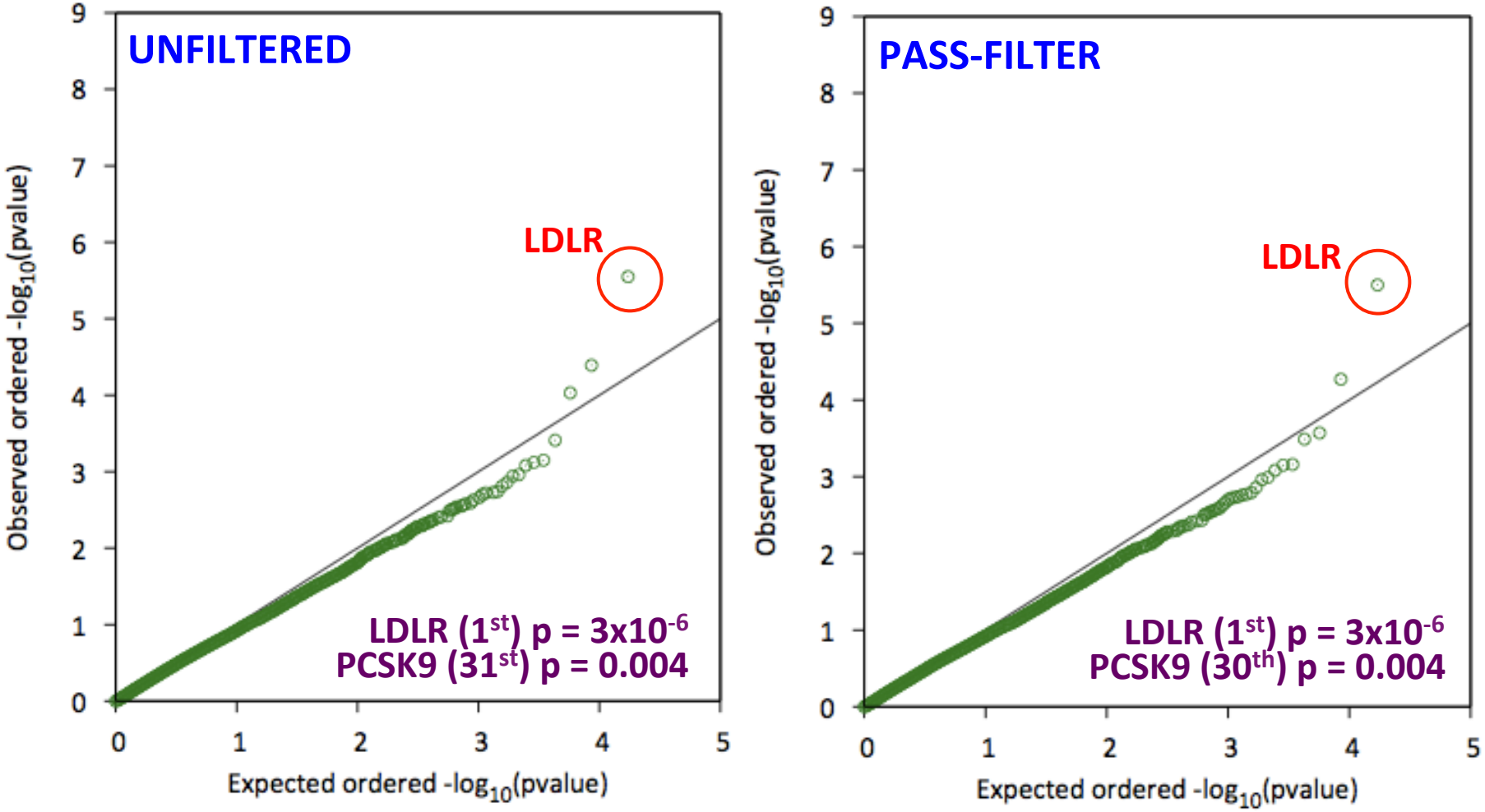
(logistic regression adjusted by PC1, PC2, age, gender, center)





LDL Results – Burden Test, MAF < 0.1%

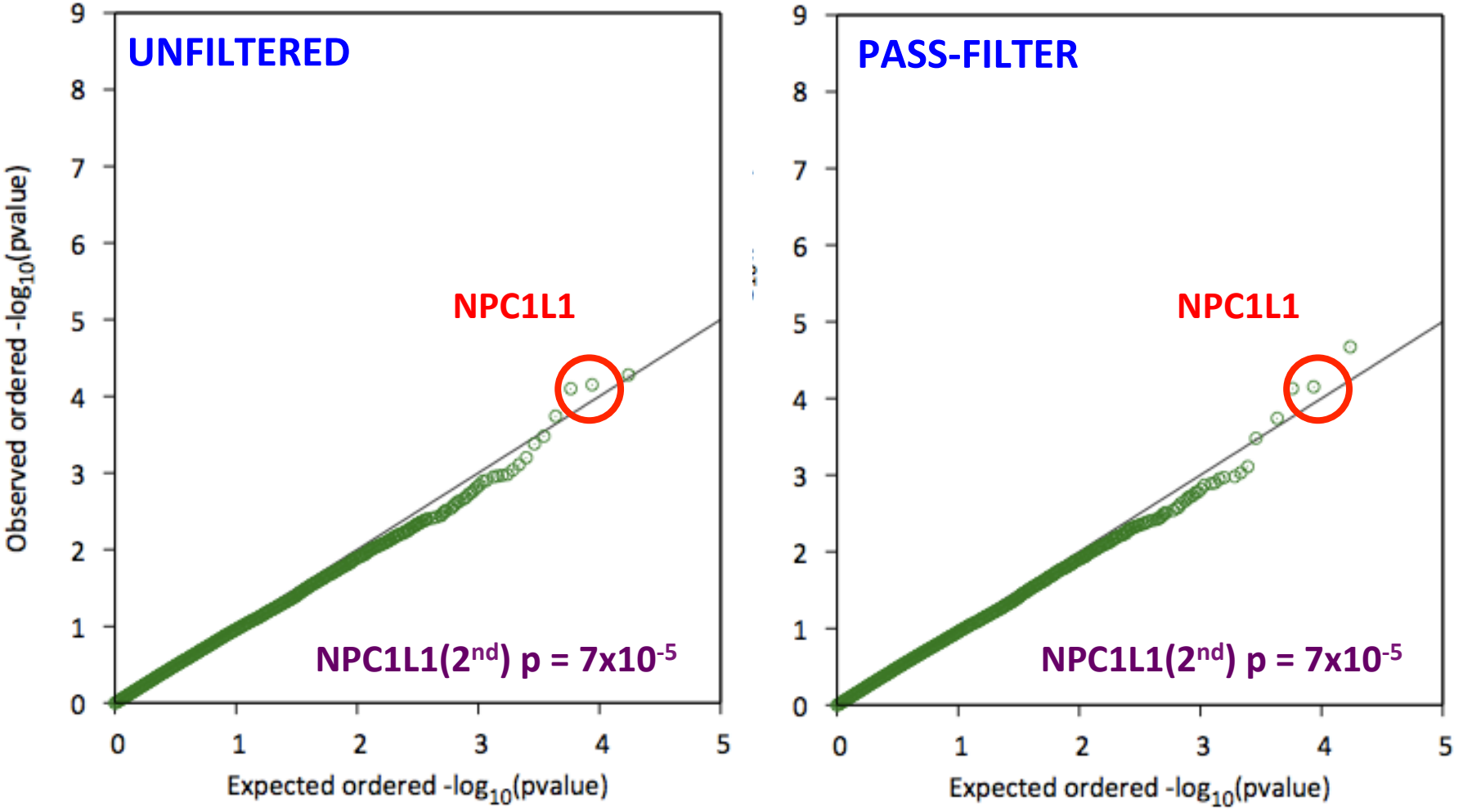
(logistic regression adjusted by PC1, PC2, age, gender, center)



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 Sebastia

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

Kai Wang[*‡], Mingyao Li[§] and Hakon Hakonarson*[*‖]*

nature REVIEWS

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 Oroz
UKRAG · W

Annals of human genetics

doi: 10.1111/j.1469-1809.2010.00566.x

Common Susceptibility Variants Examined for Association with Dilated Cardiomyopathy

Vadnie 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.