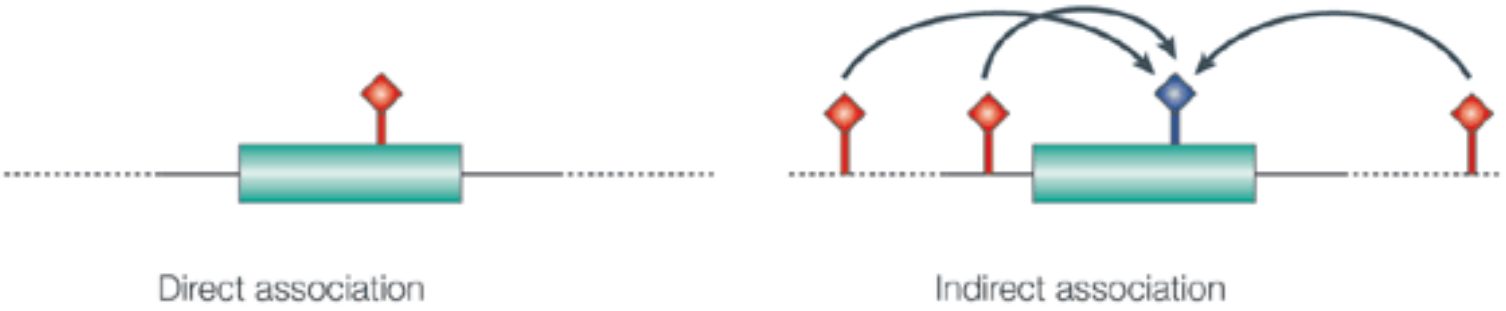
Introduction to common variation, quality control, GWAS, and PLINK (Part II)

Katrina Grasby and Lucia Colodro Conde

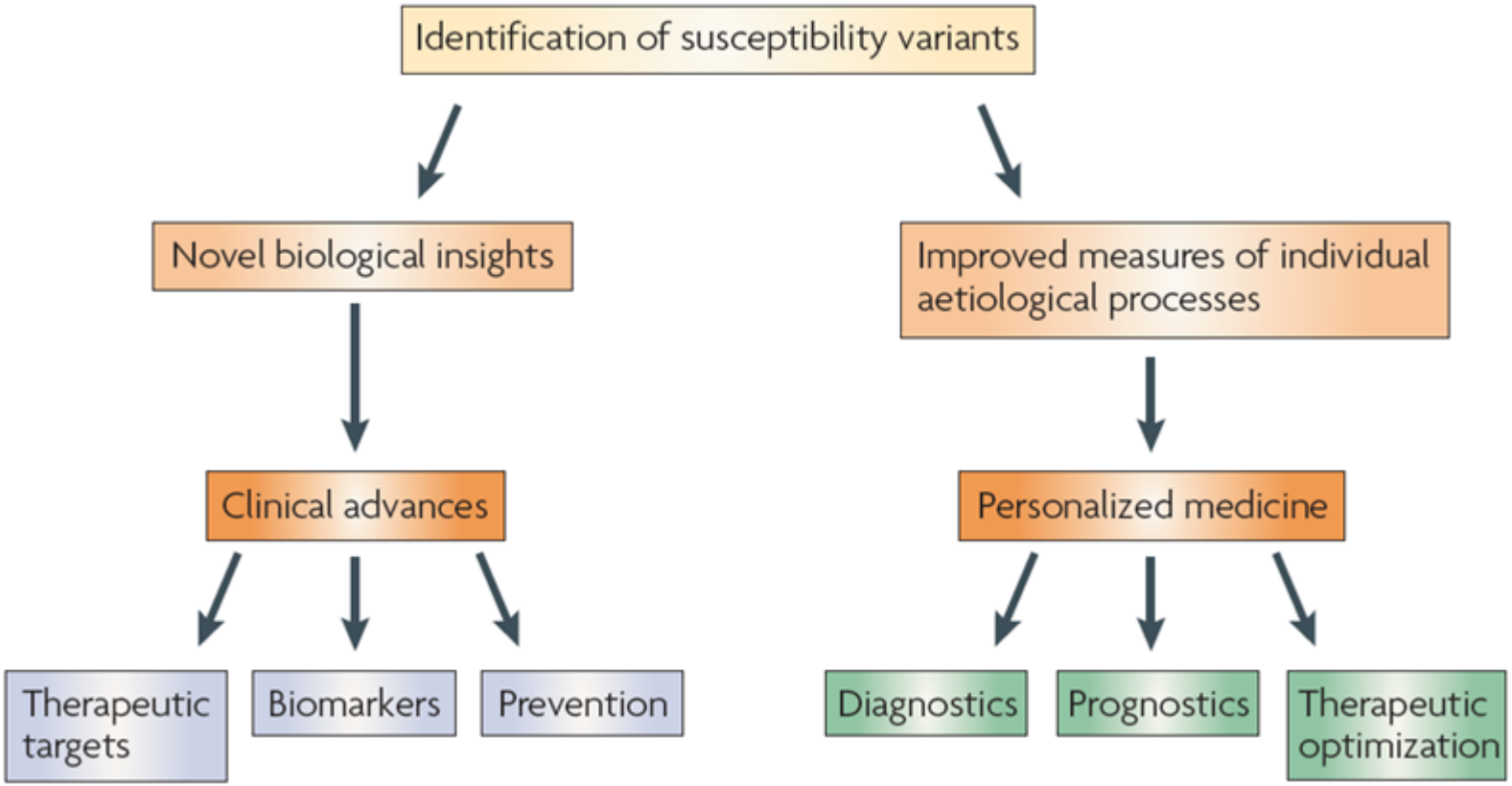
What is it?

- A hypothesis free study of genetic variation across the entire human genome
- Tests for genetic associations with continuous traits (e.g. height) or with the presence / absence of disease (e.g. cancer)
- With a focus on low penetrance & high frequency loci
- Tests indirect association



Hirschhorn & Daly. *Nat Rev Genet* (2014)

Why do it?



McCarthy et al. *Nat Rev Genet* (2008)

Quantitative Trait

Linear Regression

$$\hat{Y} = \alpha + \beta X$$

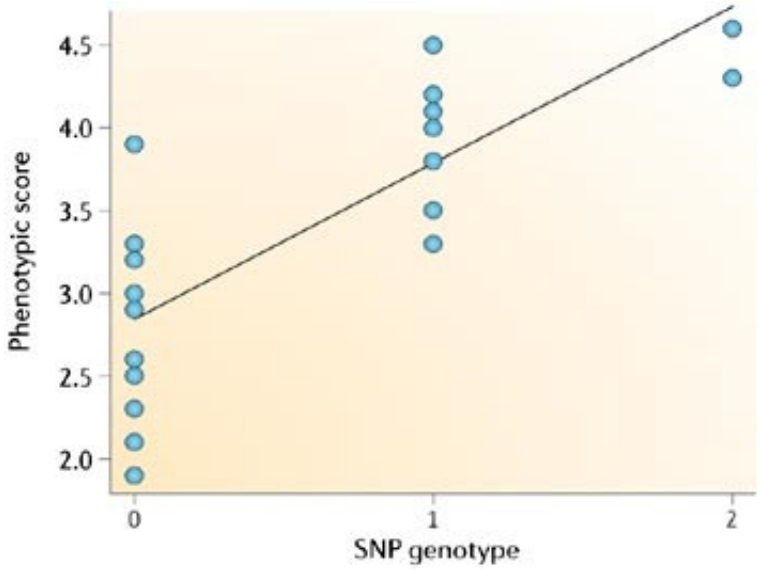
$\hat{Y}$ = score on phenotype

X = 0, 1 or 2 copies of allele (“G”)

$\beta = 0$ no association

$\beta > 0$ G allele increases trait

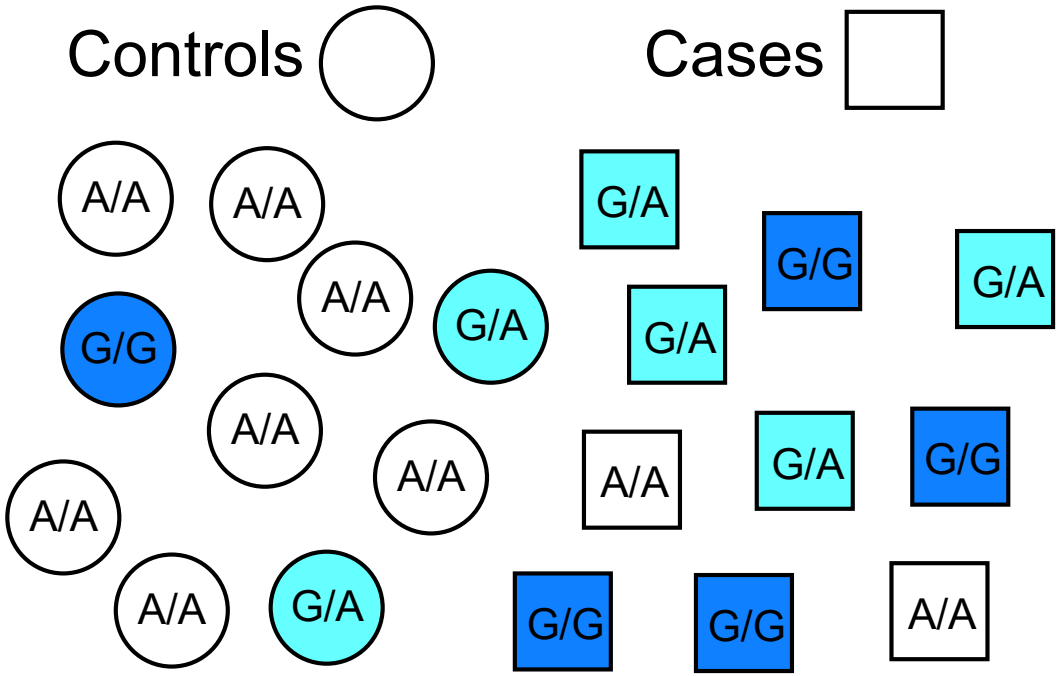
$\beta < 0$ G allele decreases trait

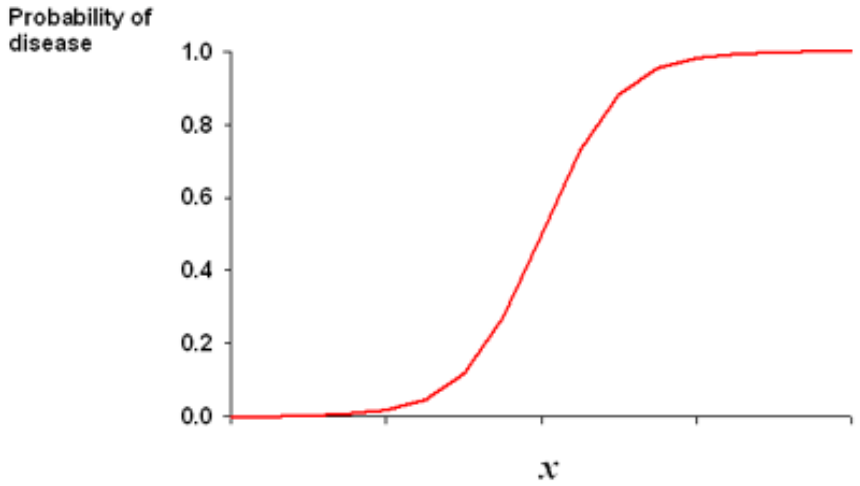


Balding. *Nat Rev Genet* (2006)

Case-Control

Logistic Regression





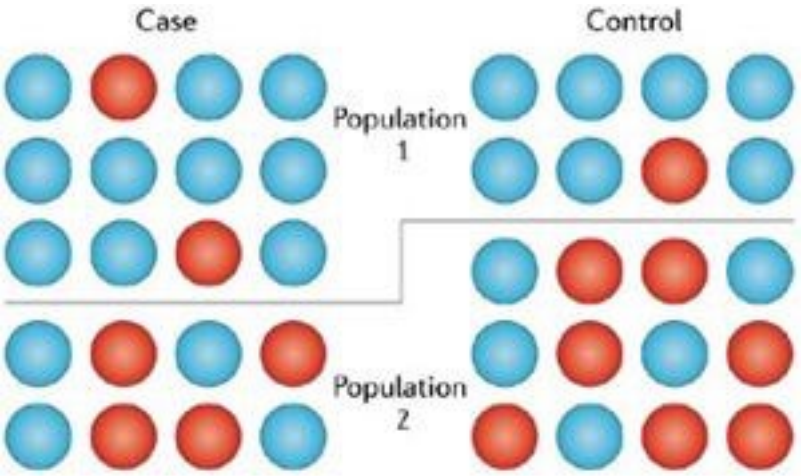
Allelic effect is an Odds Ratio (OR)

OR > 1 increases risk
OR < 1 decreases risk

The G allele is associated with disease

Confounders

- Population Stratification



Mean trait or case frequency differences between populations

Alleles with frequency differences between populations

False positive / negative associations

Balding. *Nat Rev Genet* (2006)

Multiple Testing Burden

$p < 5 \times 10^{-8}$

- Consider ancestry

~ 1 million independent tests in Caucasians (CEU)

~ 2 million in African (YRI)

Genetic Epidemiology 32: 227–234 (2008)

Estimation of Significance Thresholds for Genomewide Association Scans

Frank Dudbridge* and Arief Gusnanto

MRC Biostatistics Unit, Institute for Public Health, Cambridge, United Kingdom

Genetic Epidemiology 32: 381–385 (2008)

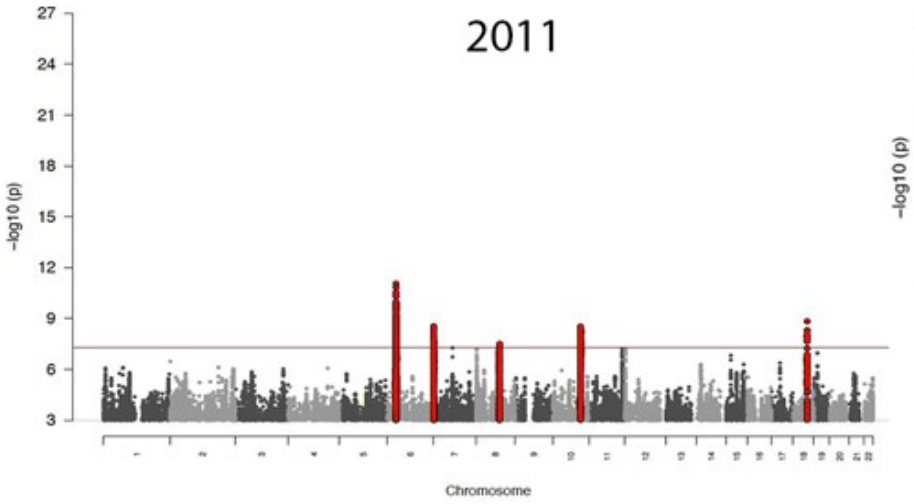
Brief Report

Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants

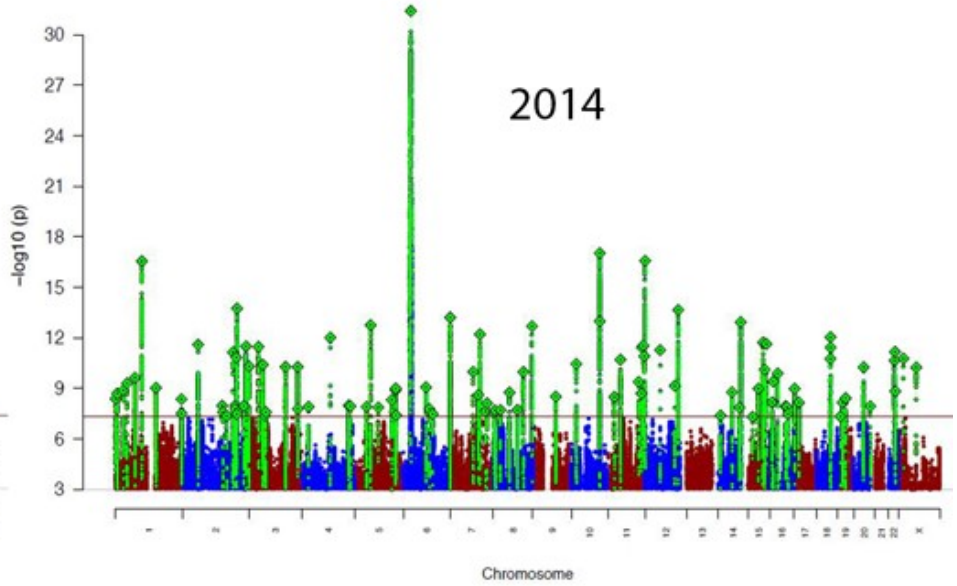
Itsik Pe'er,[1] Roman Yelensky,[2–4] David Altshuler,[2,3,5–7] and Mark J. Daly[2,5,8]*

Sample Size & Power

Schizophrenia Working Group of the Psychiatric Genomics Consortium.



9,394 cases
12,462 controls



36,989 cases
113,075 controls

Power Calculation Tools

Consider: Effect size, Sample size, Prevalence, MAF

(more on Power later in the week)

Purcell, Cherny, & Sham. *Bioinformatics,* 2003

http://zzz.bwh.harvard.edu/gpc/

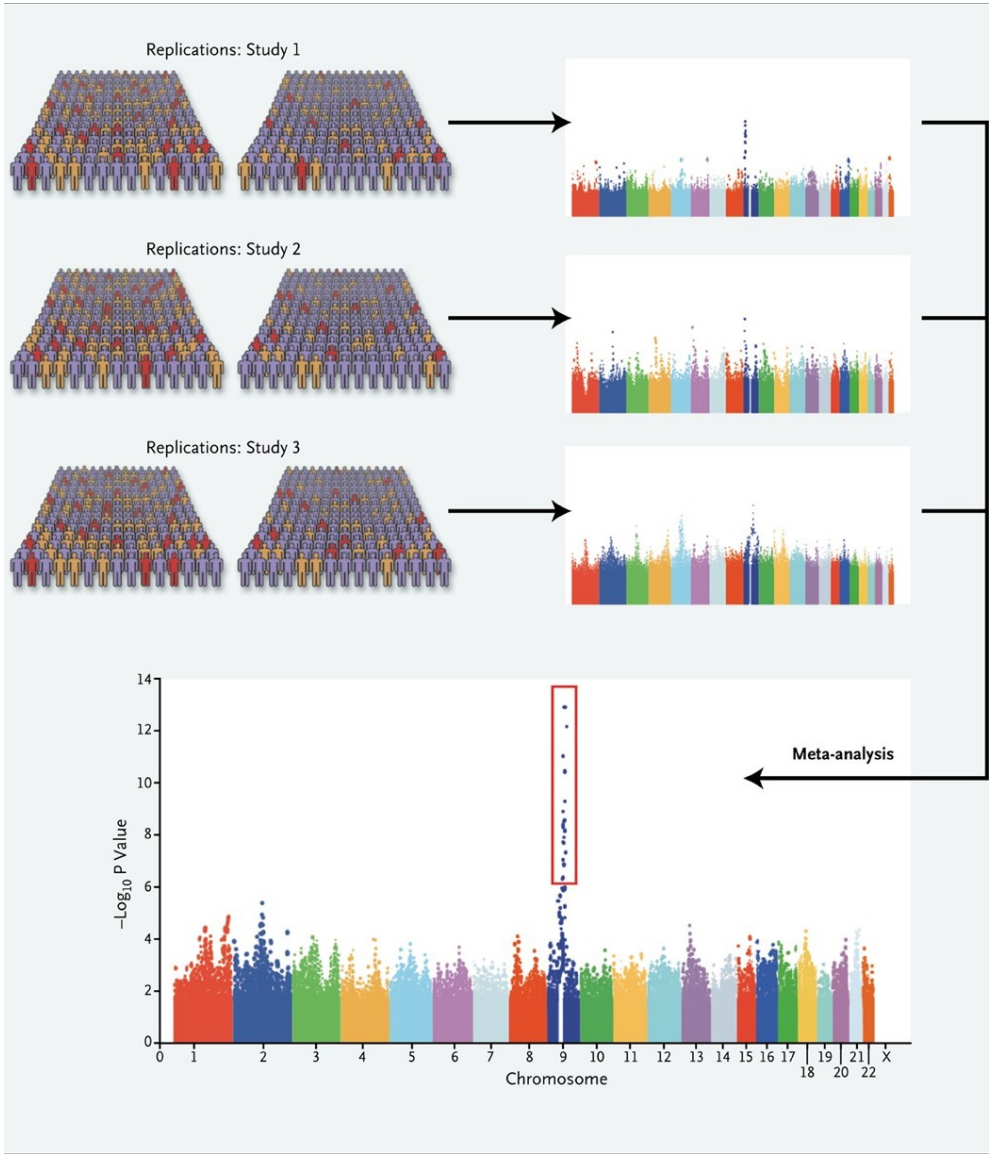
Johnson & Abecasis. *bioRxiv*, 2017

https://csg.sph.umich.edu/abecasis/gas_power_calculator/index.html

Replication

- Run GWAS in multiple samples & meta-analyze

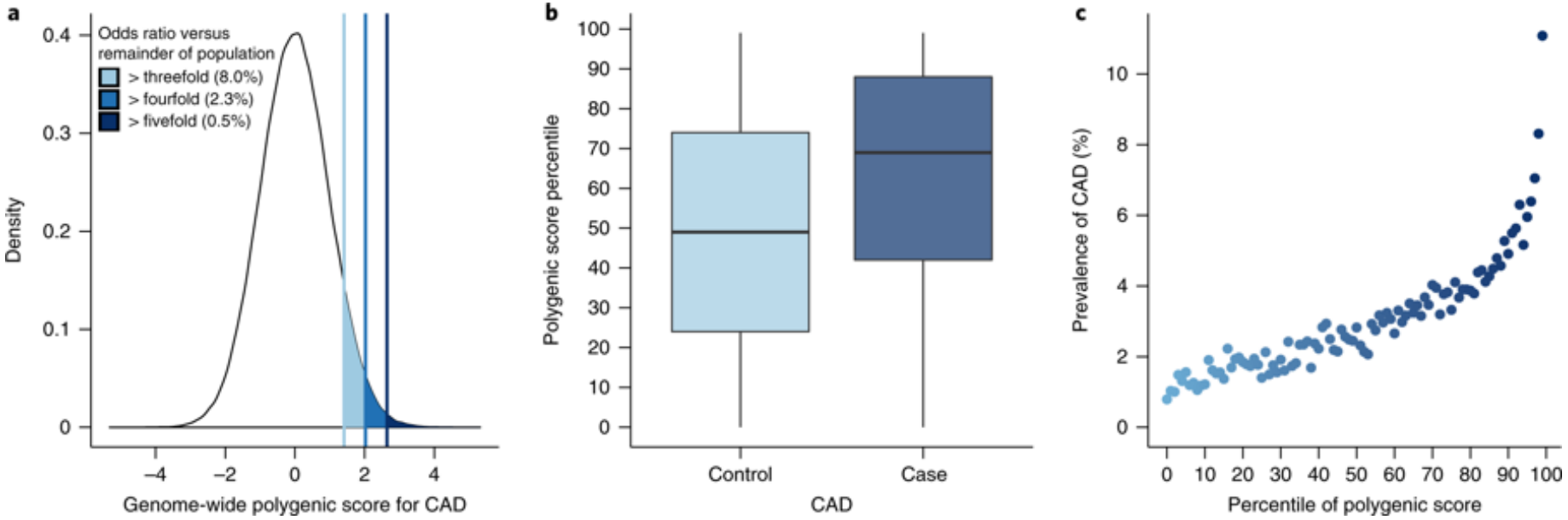
- Replicate the just the “top hits” (i.e. $p < 1e\text{-}5$)



Manolio. *N Engl J Med*, 2010

Key GWAS Findings (so far)

- Thousands of genetic variants
- Each has a very small effect
- Large samples required
- Can look at the cumulative effect…



Khera et al. *Nat Gen*, 2018

GWAS check list

1. Quality Control
   - Genotyping Call Rate, HWE, MAF, Sample Call Rate
2. Confounders
   - Population stratification, any systematic difference between cases & controls
3. Appropriate methods for individuals are related
   - mixed models (e.g. SAIGE later in the week)
4. Sample size large
5. Replication
6. Indirect association
   - be wary of over-interpreting biology, follow-up work is essential!