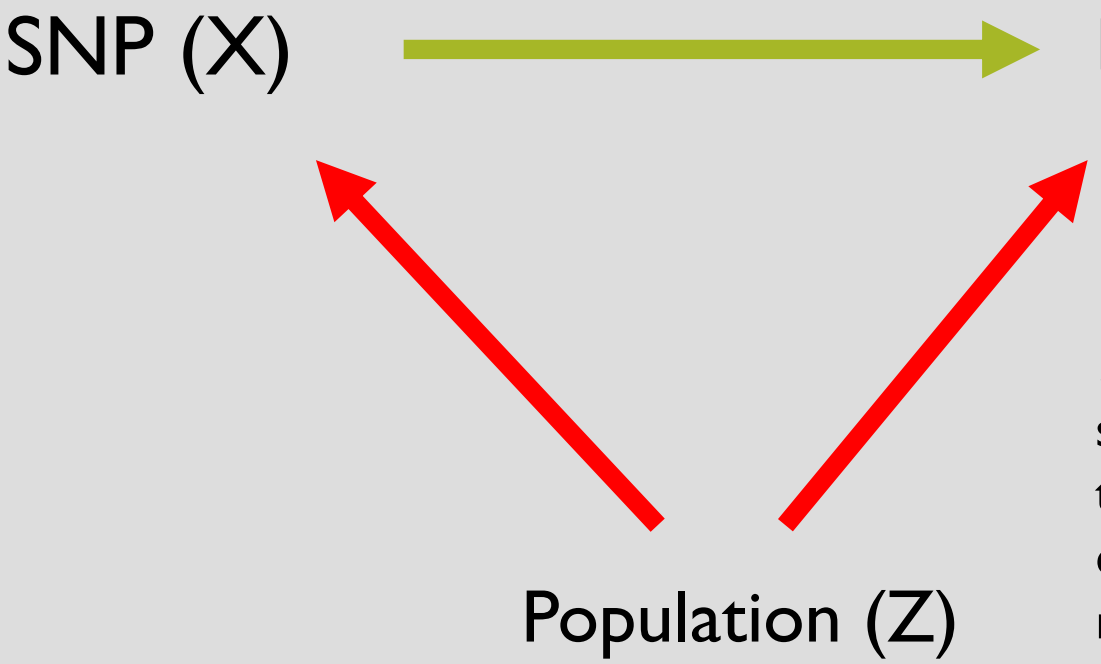
GWAS 6

Matti Pirinen
University of Helsinki
Week 3-4

CONFOUNDER

We want to study X-Y relationship…



… but if there are associations between some 3rd variable Z and both X and Y, then Z may cause an observable X-Y association even if there is no **direct/causal** relationship between X and Y

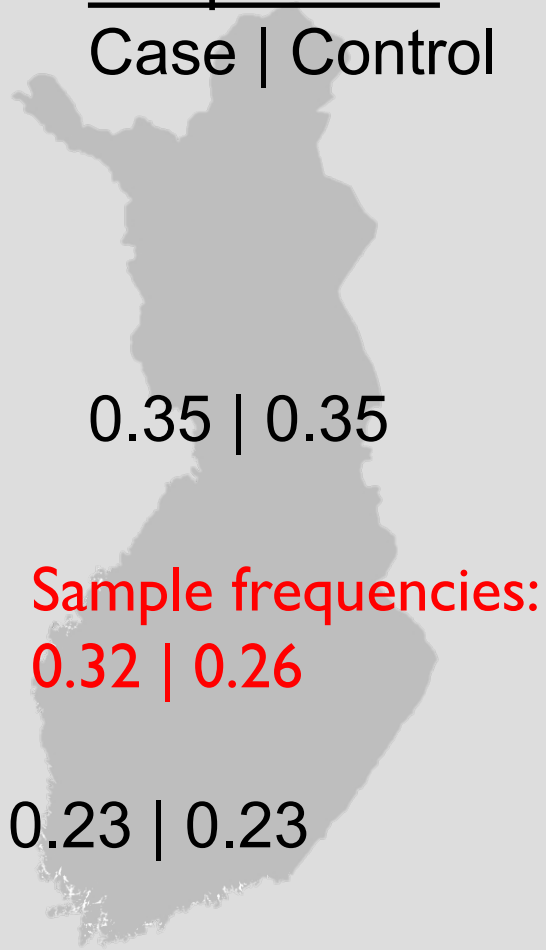
Z is **confounder** of X-Y association

CONFOUNDING BY ANCESTRY

- Consider a genetic variant that has no effect on heart disease but has different regional frequencies
  - Variant "A" frequency 0.23 in Helsinki region
  - Variant "A" frequency 0.35 in Oulu region
- Does not show association with disease in Helsinki or in Oulu (because there is none)
- What happens if we do not match well regions of origins for cases and controls ?

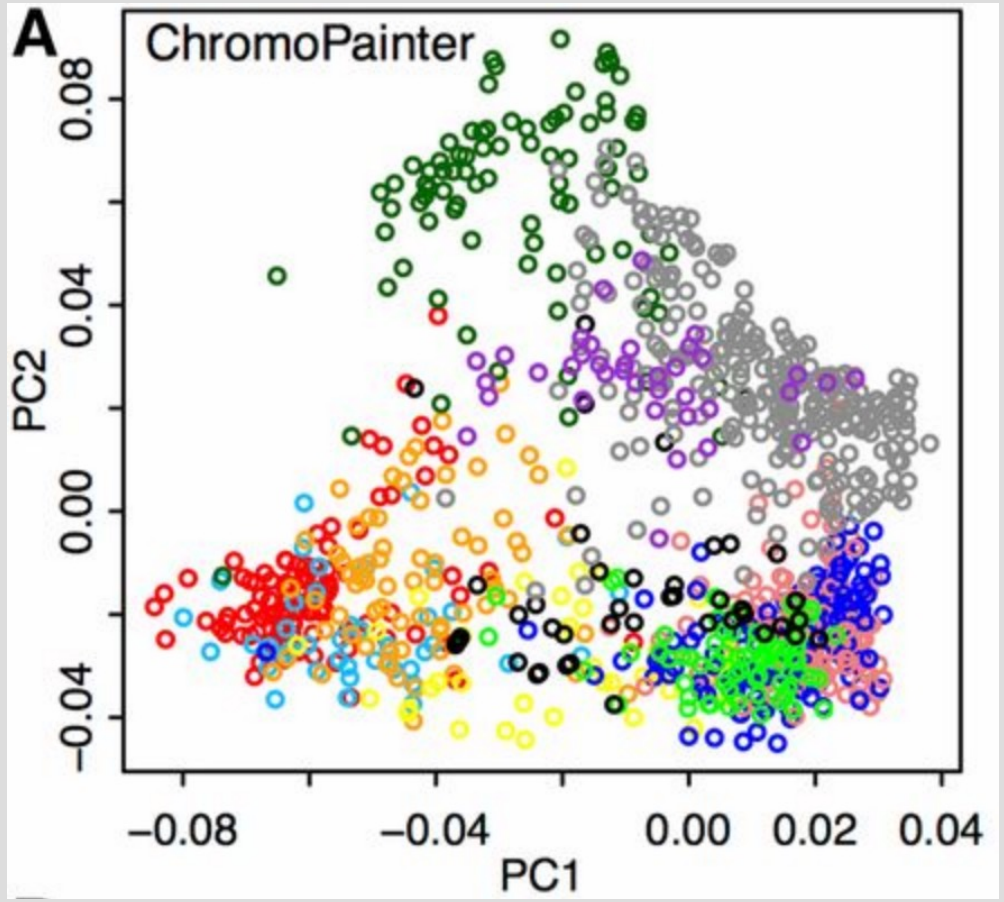
CONFOUNDING BY ANCESTRY

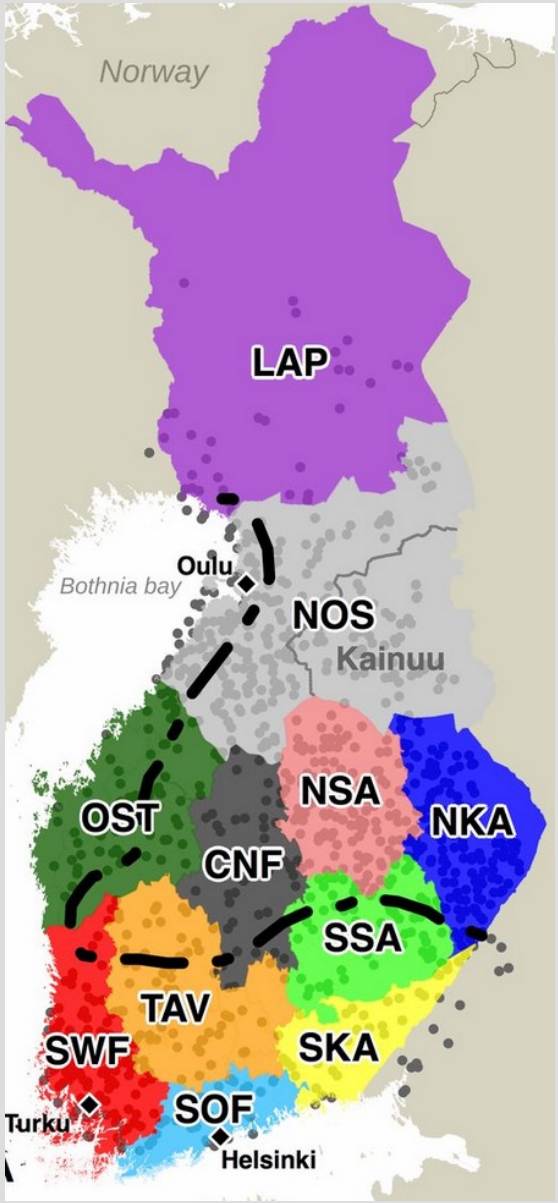
- SNP that has no effect on heart disease but has different regional frequencies
  - Variant ”A” frequency 0.23 in Helsinki region
  - Variant ”A” frequency 0.35 in Oulu region
- Consider sampling
  - 2000 cases (500 from H and 1500 from O).
    - ”A” frequency in cases is 0.32
  - 2000 controls (1500 from H and 500 from O).
    - ”A” frequency in cases is 0.26
- False association that variant ”A” increases risk for heart disease !
- Different ancestries confound the analysis



USING LEADING PCS TO MATCH CASES & CONTROLS

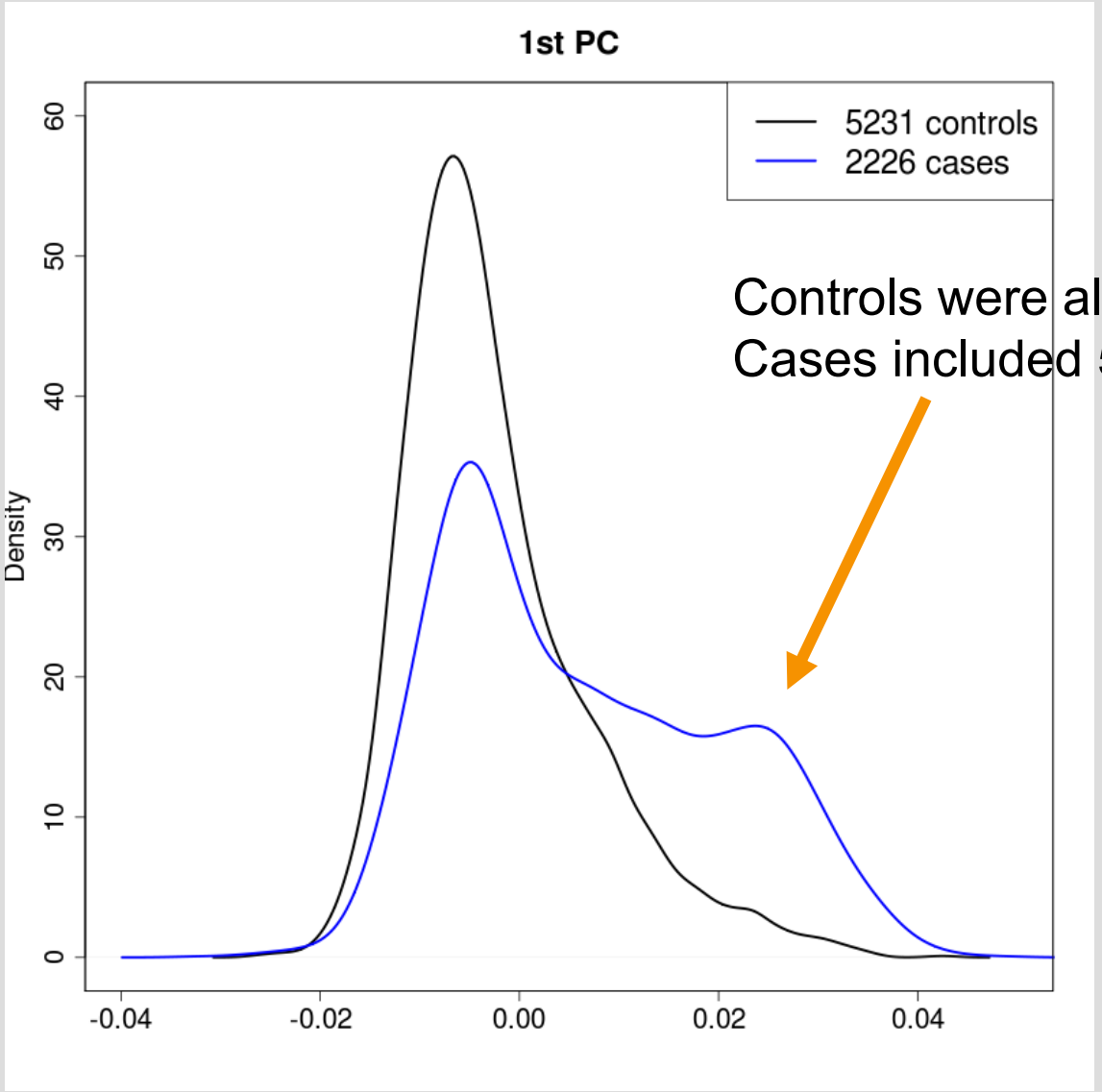
- Often we do not know regional origins of samples or they may not be informative of genetic background
- But we can infer genetic similarity and adjust the analyses for that by taking leading PCs of the genetic correlation matrix and use them as covariates (= additional predictors) in the regression model to remove confounding





Kerminen et al. 2017 G3: http://www.g3journal.org/content/7/10/3459

EXAMPLE FROM A PSORIASIS STUDY IN UK

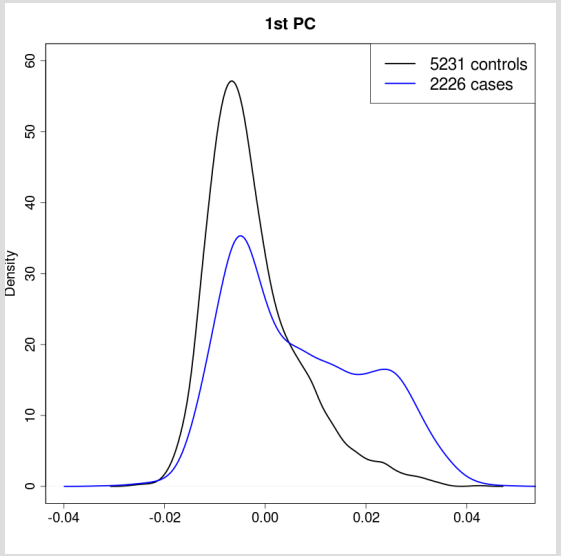


Controls were all from the UK.
Cases included 500 Irish samples.

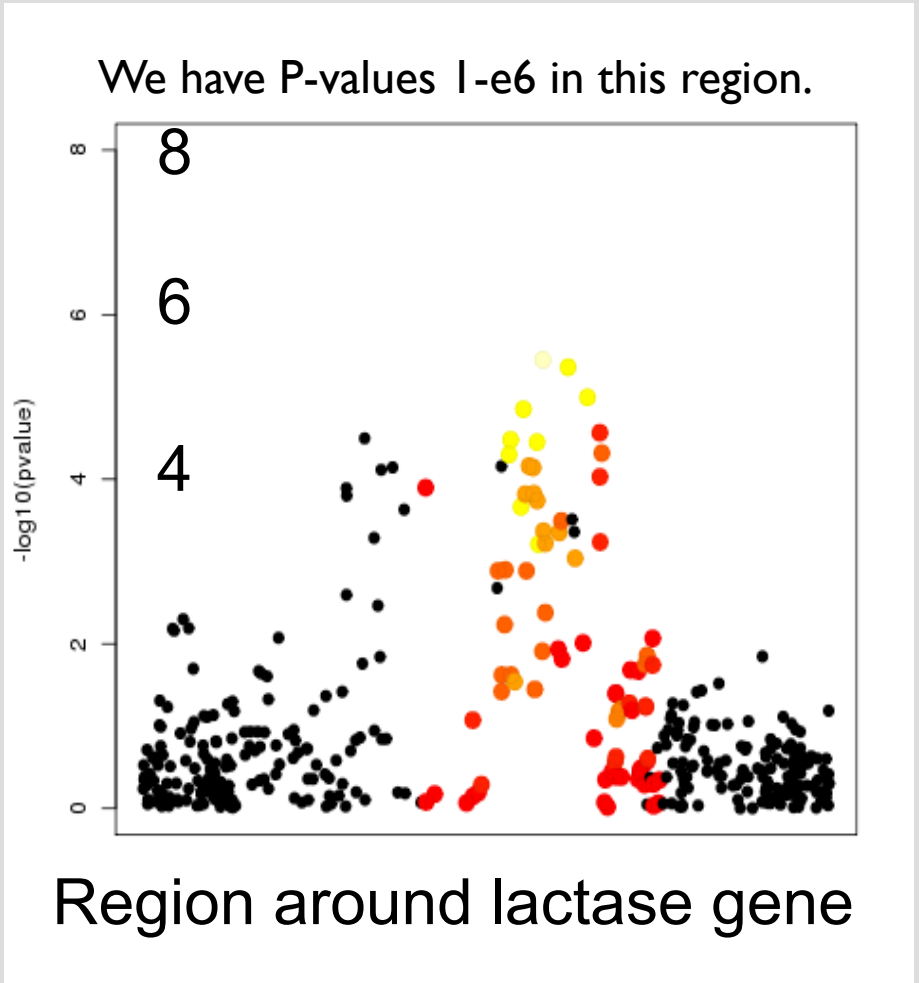
- Clear mismatch in ancestry profiles btw cases / controls!
- If we just analyze these data for association between genetic variants and psoriasis what comes up?

Strange et al. 2011 Nature Genetics

EXAMPLE FROM A PSORIASIS STUDY IN UK



Controls were all from the UK.
Cases included 500 Irish samples.



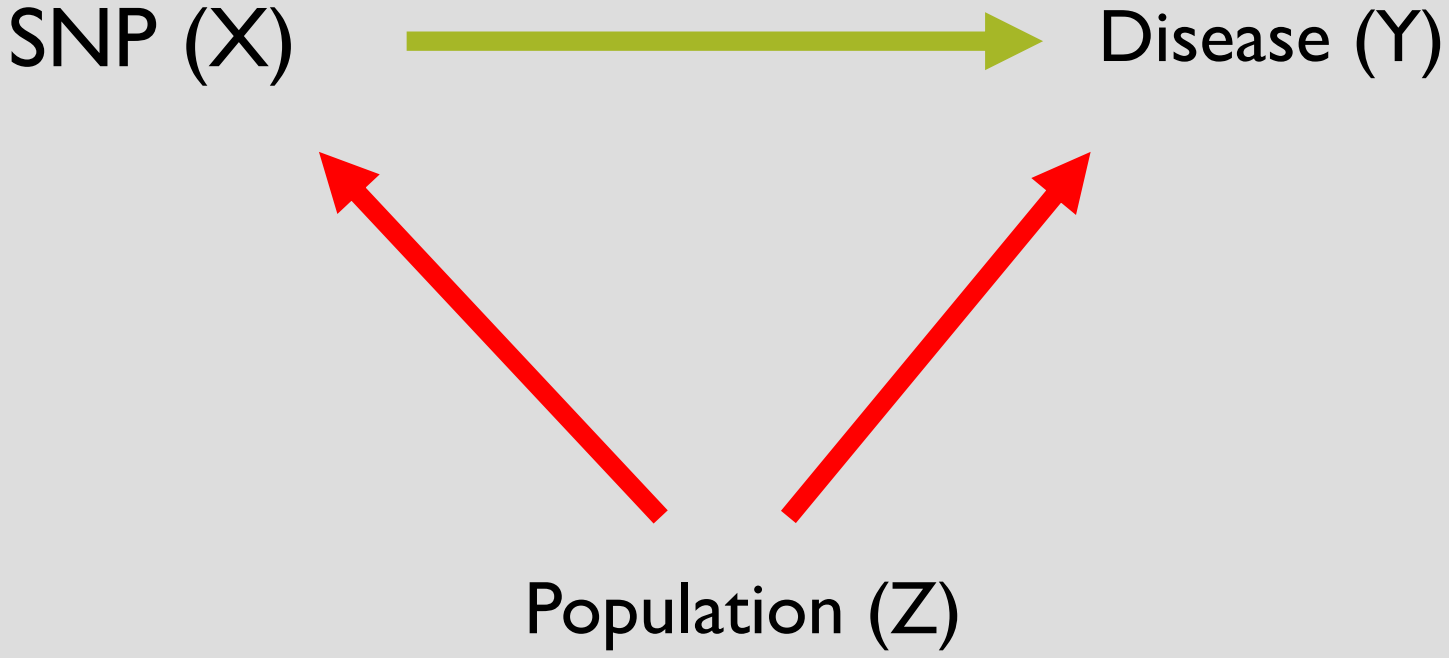
Region around lactase gene

Does lactase persistance variant really affect psoriasis susceptibility ?

(Or is it just in different frequencies in the UK and Ireland, and we are seeing a spurious association with psoriasis in this unmatched sample?)

Strange et al. 2011 Nature Genetics

CONFOUNDER AND REGRESSION



Logistic regresison model M:
$Y \sim a + X\,b + Z\,c$

Logistic regression model M*:
$Y \sim a^* + X\,b^*$

M adjusts for confounder Z whereas M* does not.

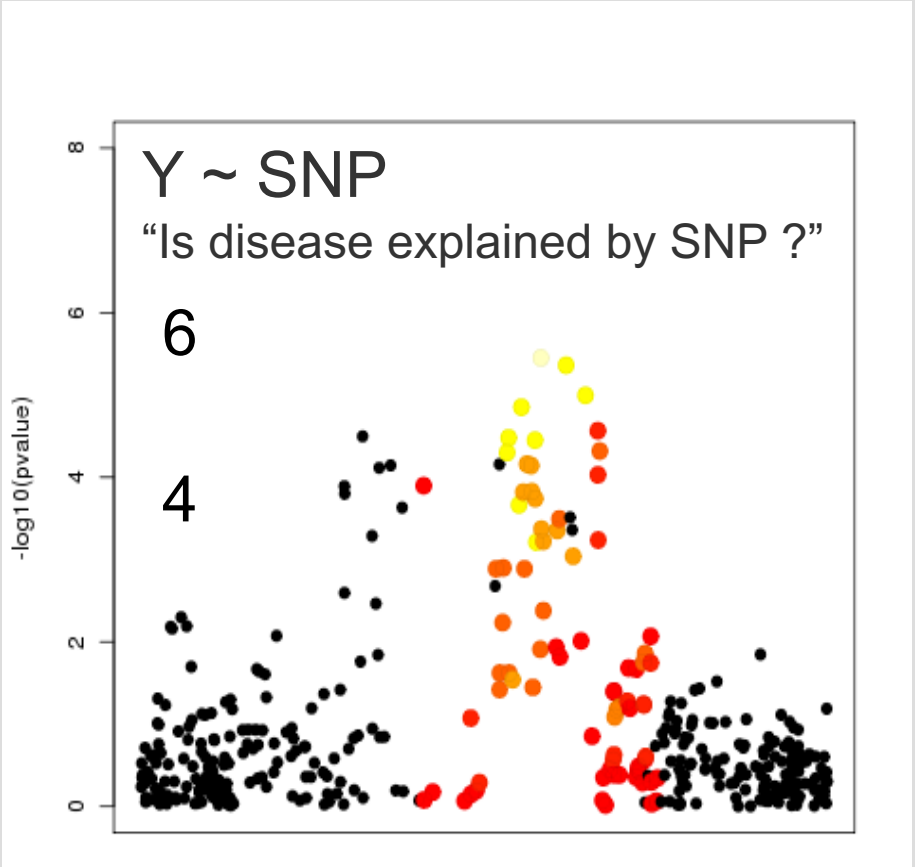
b = effect of SNP for a fixed value of Z

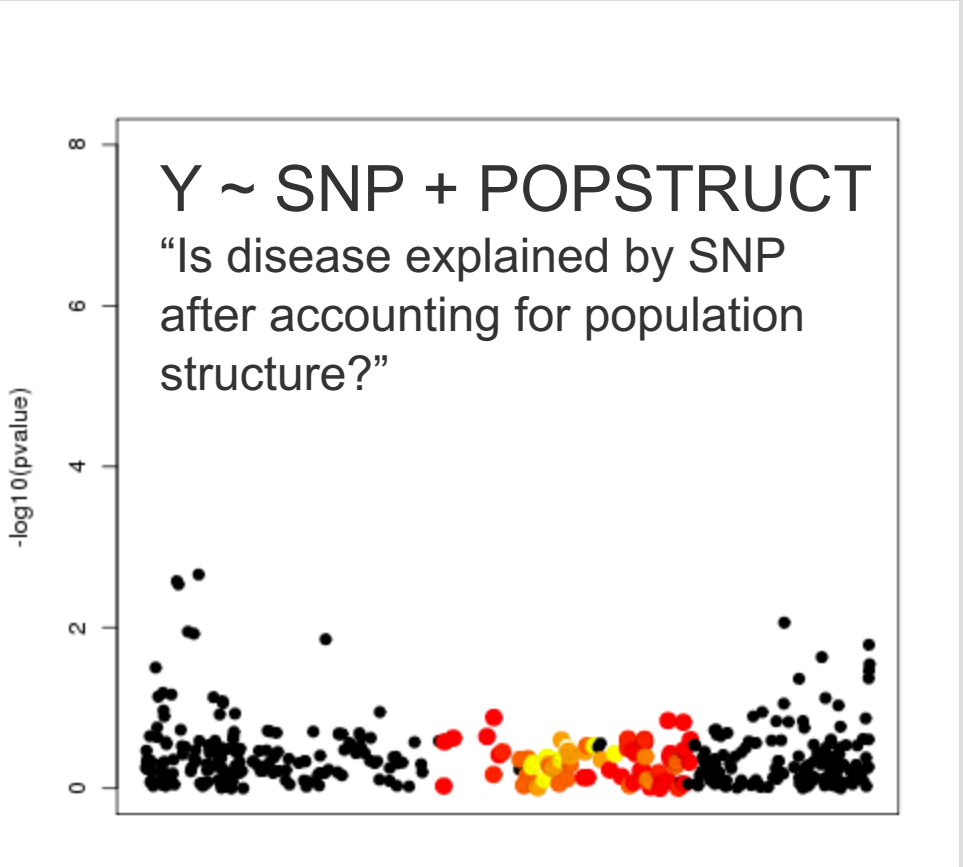
b* = effect of SNP without accounting for Z

EXAMPLE FROM A PSORIASIS STUDY IN UK

Does lactase gene really affect psoriasis susceptibility?

Probably not, since the signal can be completely explained by ancestry (1st PC) and goes away when PC1 is included in the logistic regression model

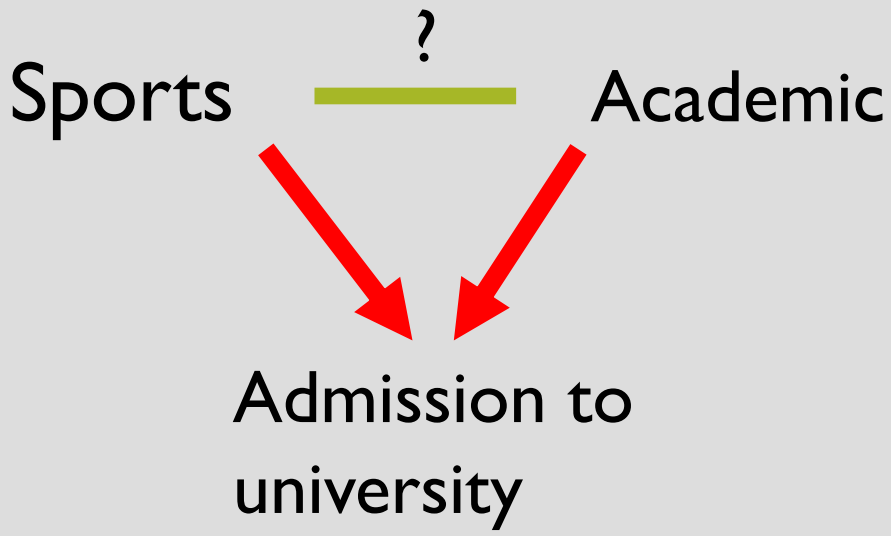




Strange et al. 2011 Nature Genetics

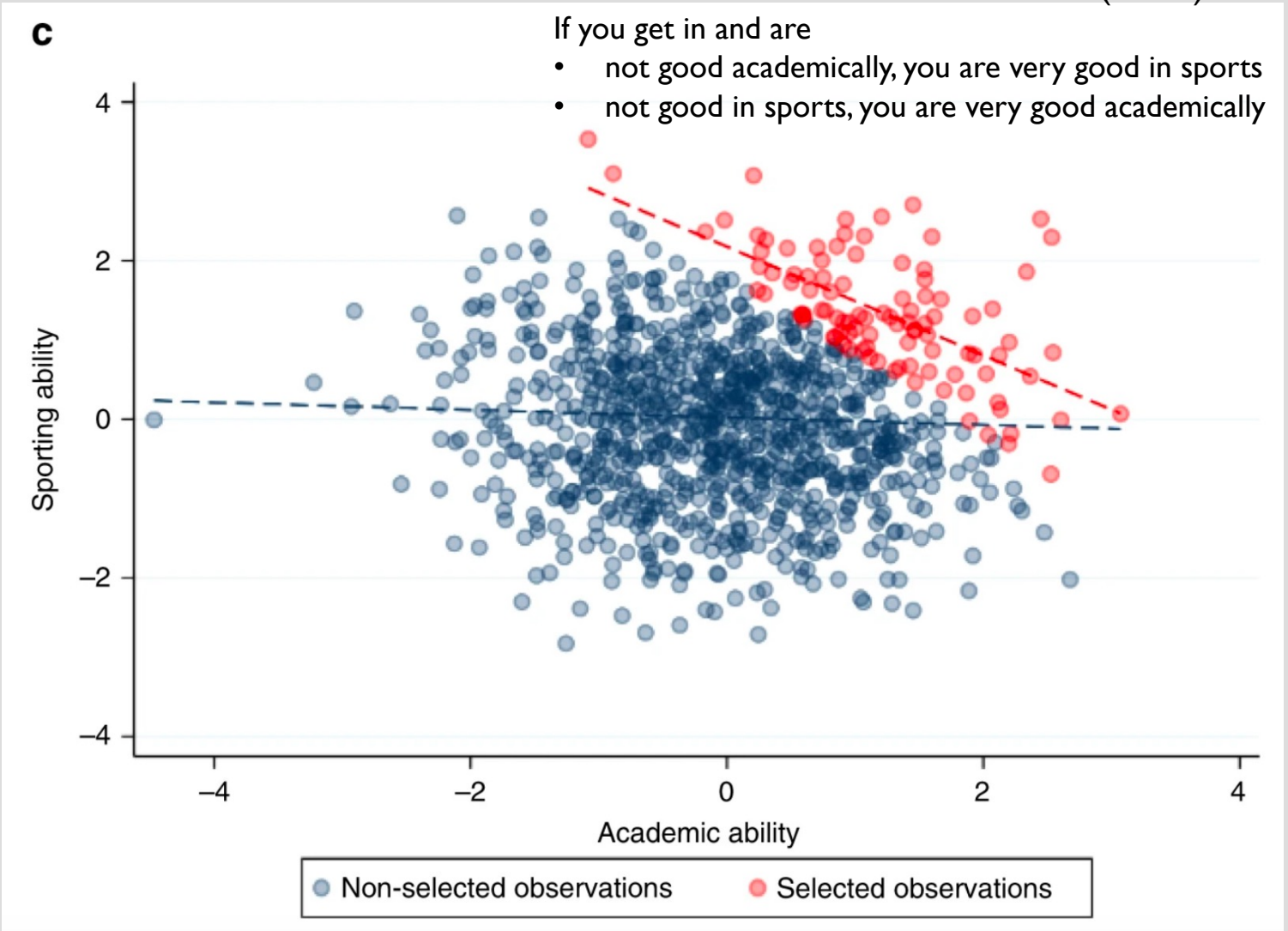
EXAMPLE OF COLLIDER BIAS

Nat Communications 11: 5479 (2020)

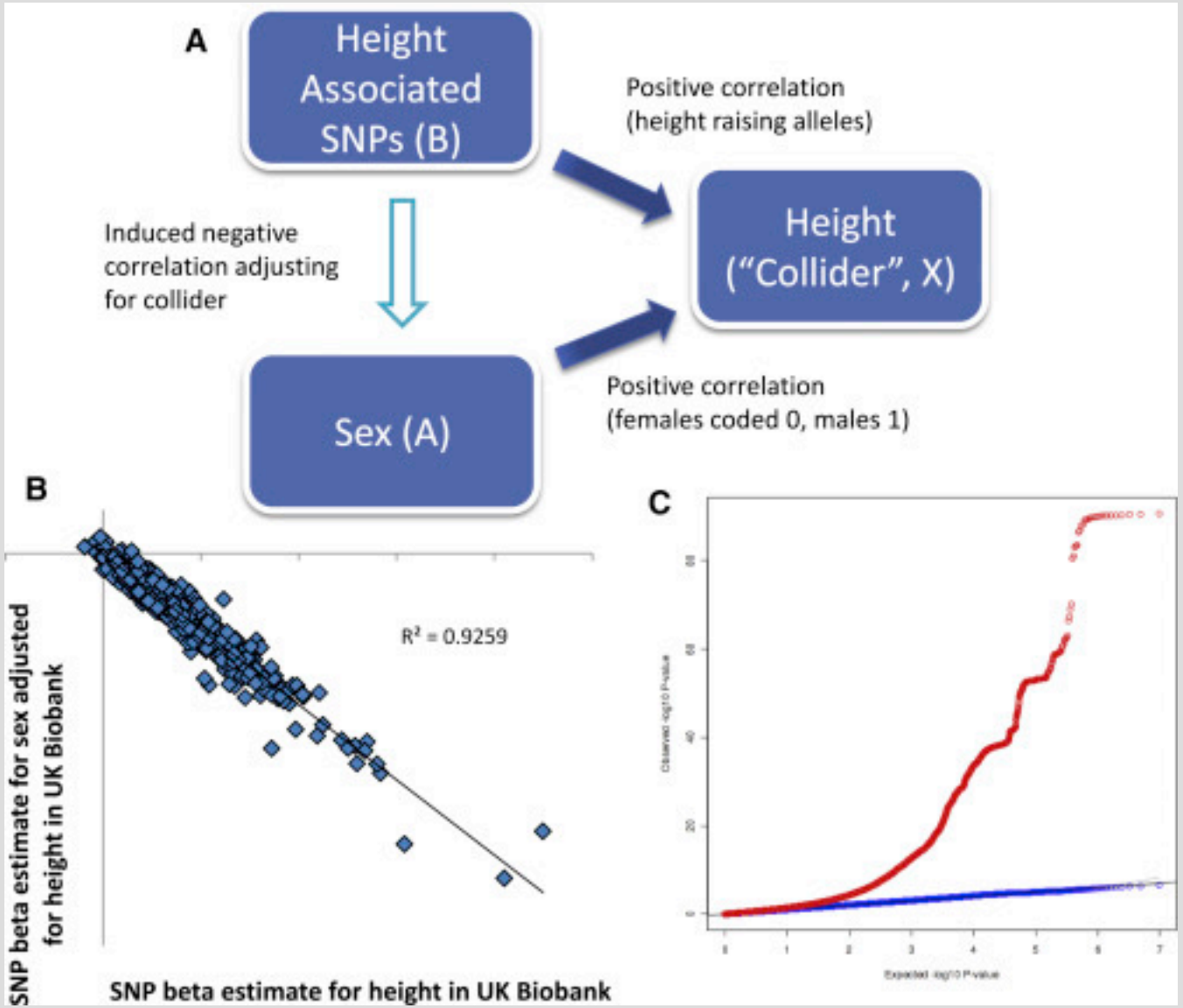


Two ways to get in to a university:
Excel in sports or excel academically.

If relationship between athletic and academic abilities is studied within the admitted students then a negative correlation is seen. This is due to collider bias as in the general population no such relationship exists.



Day et al. AJHG (2016) 98(2): 392-393



COLLIDER BIAS USING UK BIOBANK

- N=142,000 (~50% males)
- In GWAS of sex, no SNP reached GWS
  - All known ~700 height-SNPs followed null
- When adjusting GWAS of sex on height, 222 of the known height SNPs had P<5e-8
  - Red QQ-plot is from the height adjusted analysis, blue from the analysis ignoring height
- Each height-increasing allele showed "lower probability of being male" as expected under collider bias
- Outside of known height-SNPs no other SNPs gave GWS results
  - Collider bias affects only variants associated with the collider

INDEPENDENT COVARIATES

- Genotype X and covariate W are <u>INDEPENDENT</u> in the population
  - X = autosomal variant and W = sex
  - X = SNP in chr 17 and W = SNP in HLA region on chr 6
- If genotype has no effect (β = 0), X and W are independent and the population follows the model:

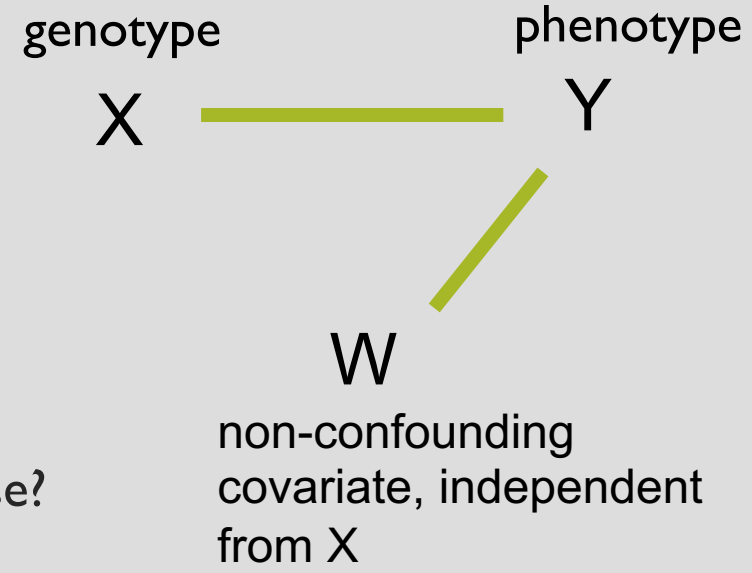
$$Y \sim \mu + X \cdot 0 + W \cdot \gamma$$

then also according to the regression model

$$Y \sim \mu' + X \cdot \beta',$$

$$\beta' = 0.$$

- When X and W are independent, we do not create a spurious X-Y association by leaving W out from the model
- We have a choice between the models, so which model should we use?

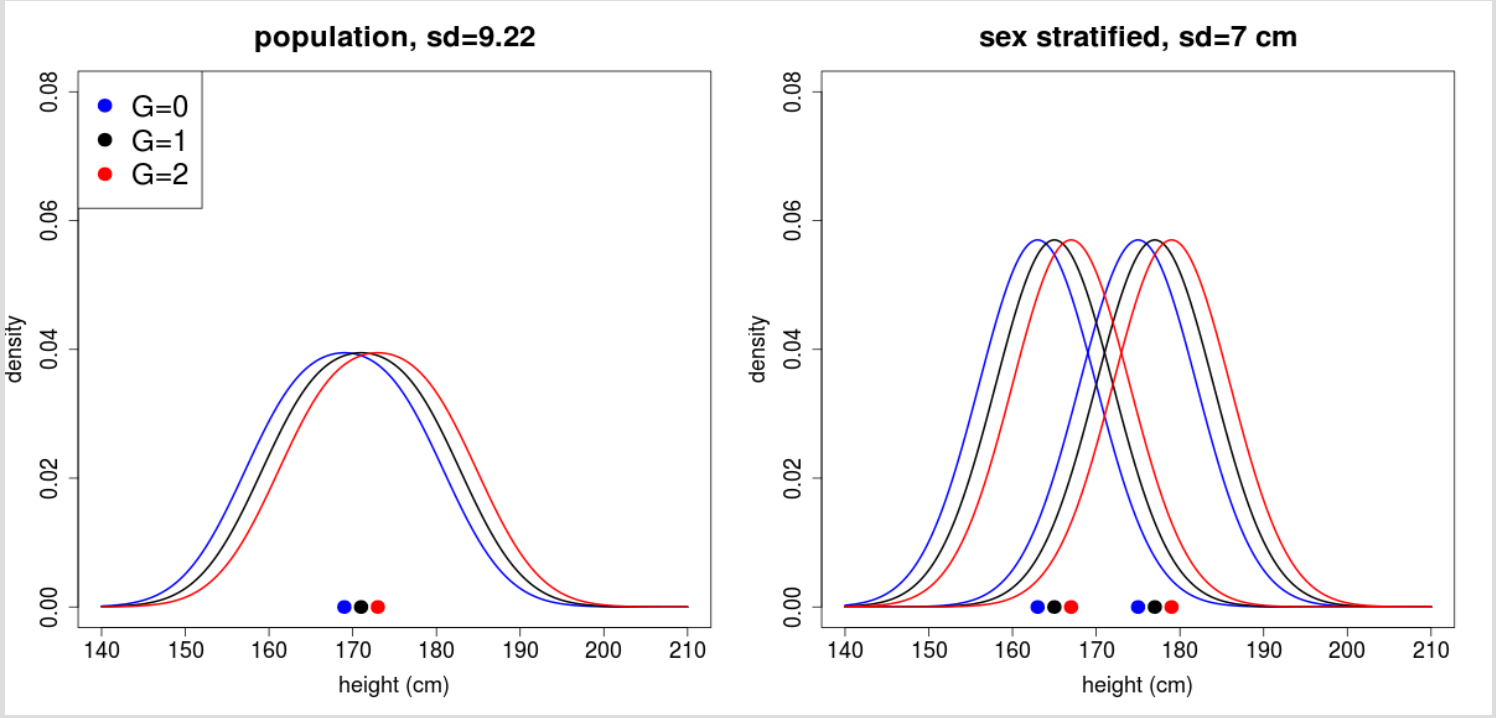


INDEPENDENT COVARIATES

- We consider two models when X and W are independent
  - Model M: $Y \sim \mu + X \cdot \beta + W \cdot \gamma$
  - Model M': $Y \sim \mu' + X \cdot \beta'$
- What is the relationship between β and β' ?
- What is the uncertainty in the estimates of β and β' ?
- What is the information in model M and model M' about hypothesis that X does not have an effect ($\beta = \beta' = 0$) ?
  - What is the statistical power for detecting a non-zero genetic effect by using these two models?
- Answers depend on whether we conside linear or logistic model and whether we consider population or case-control analysis

LINEAR MODEL & INDEPENDENT COVARIATE

- $Y \sim \mu + X \cdot \beta + W \cdot \gamma + \varepsilon$
- *Y* height, *W* sex (0 = female, 1 = male), *X* hypothetical SNP
- $\mu = 163, \gamma = 12, \beta = 2, \varepsilon \sim N(0, 7^2)$

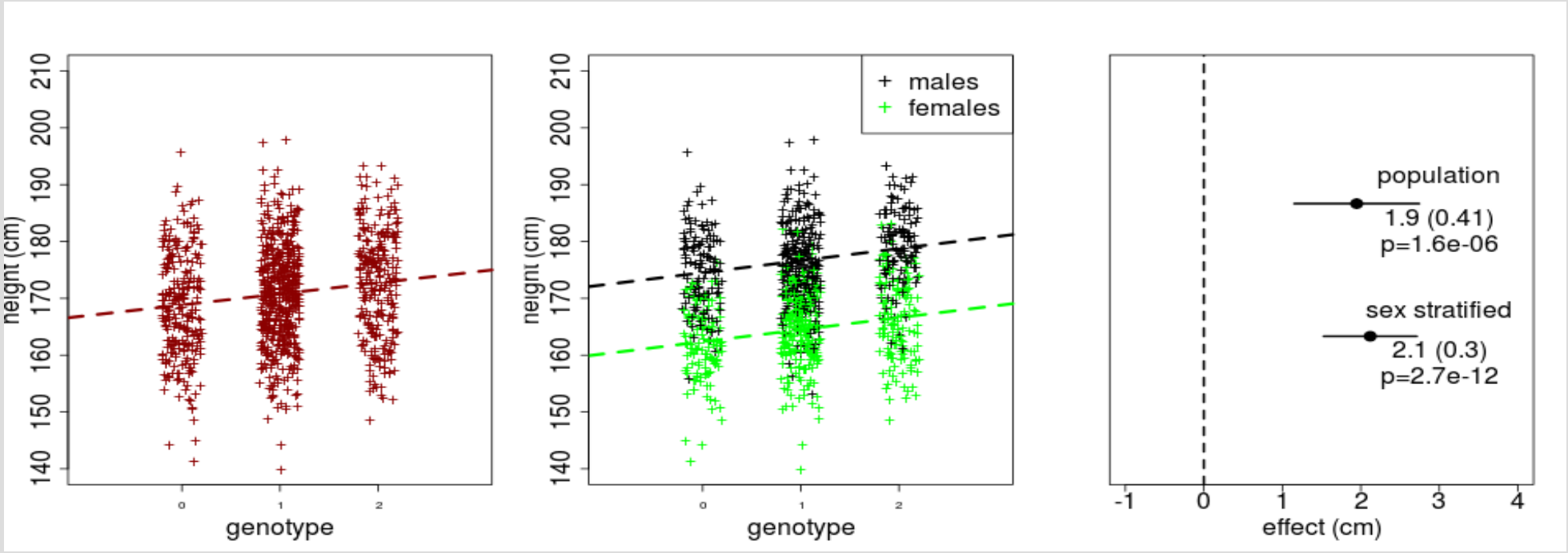


	X=0	X=1	X=2
Female	163	165	167
Male	175	177	179
Population	169	171	173

The effect of SNP is the same in all three samples of individuals.

COVARIATE REDUCES NOISE

- Both models estimate the same effect (β = β')
- SE is $\frac{\sigma_\varepsilon}{\sqrt{2nf(1-f)}}$ and decreases when covariate explains away some noise in σ_ε
- The model with the covariate has a smaller SE and hence a higher power to detect a non-zero β than the model without the covariate



Using sex as covariate is like analysing males and females separately and then combining the estimates.

This results in smaller SE than fitting regression to all of data at once since variation due to sex is just noise when estimating genotype's effect.

NON-CONFOUNDING COVARIATE & LOGISTIC REGRESSION OF POPULATION DATA

- X and W are independent
  - M: $Y \sim \mu + X \cdot \beta + W \cdot \gamma$
  - M': $Y \sim \mu' + X \cdot \beta'$

- For logistic regression in a population sample where *X* and *W* are independent,
  - $|\beta'| \leq |\beta|$
  - $\mathrm{SE}(\widehat{\beta'}) \leq \mathrm{SE}(\hat{\beta})$
  - Power to detect that β is non-zero is larger than that β' is non-zero
- If we are interested in the effect size, then the model should be chosen based on whether the covariate-adjusted effect size is more relevant than the effect size without the covariate adjustement
- In the GWAS setting, where the power to detect non-zero effects is the primary goal, the covariate adjusted model should be used as it is the more powerful model
  - Difference between models become tiny for diseases with low prevalence
- This setting is relevant for population biobanks but is **not a typical GWAS setting** which would be the case-control ascertainment (next slide)

NON-CONFOUNDING COVARIATES AND POWER IN CASE-CONTROL STUDIES

	Prevalence	Risk factor	Odds-ratio	Frequency	NCP/
Ankylosing Spondylitis	0.25%	HLA-B27	49	0.08	0.48
Psoriasis	1%	HLA-C	6.4	0.24	0.83
Multiple sclerosis	0.1%	Female	2.3	0.5	0.96
Migraine	20%	Female	4	0.5	1.04

NCP/ is the ratio of NCPs between models M (adjusted) and M' (unadjusted).

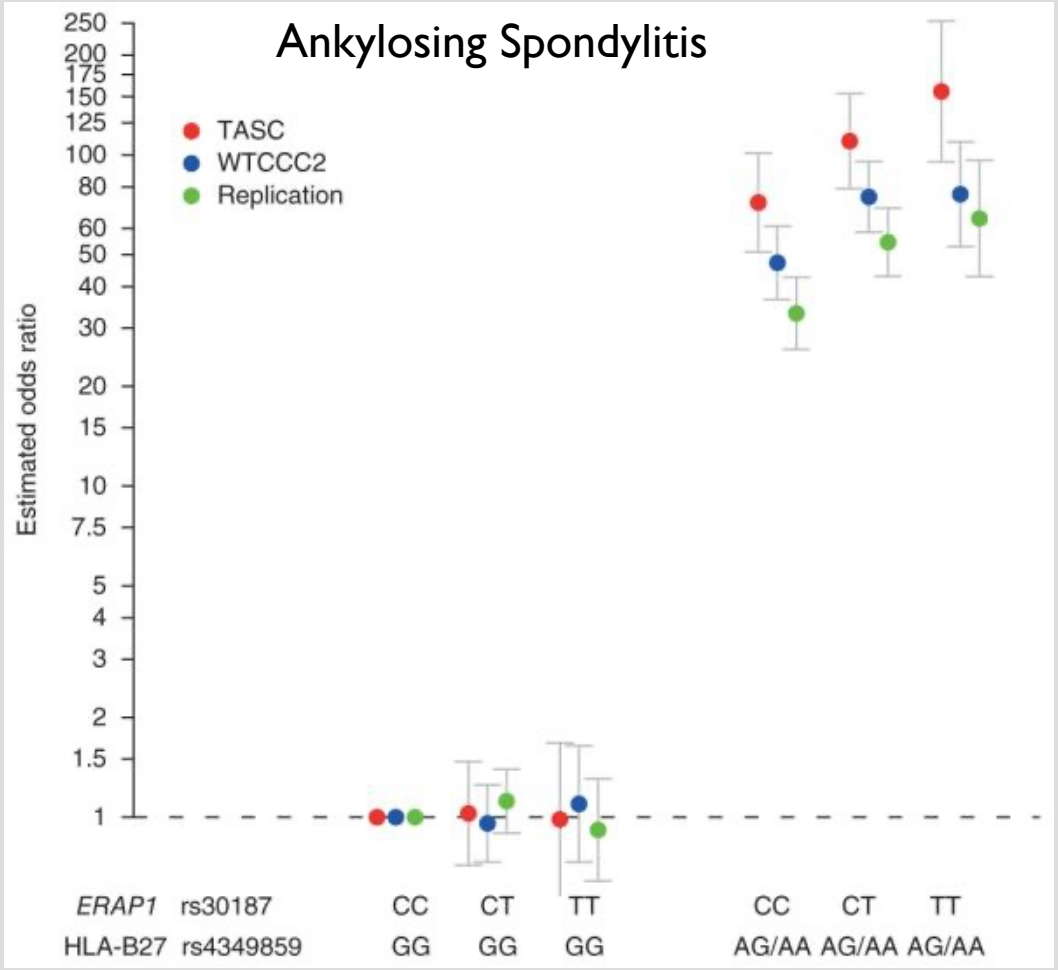
Which model to prefer depends on the disease prevalence in the population!

For a low prevalence, prefer the unadjusted model.

For a high prevalence, prefer the adjusted model.

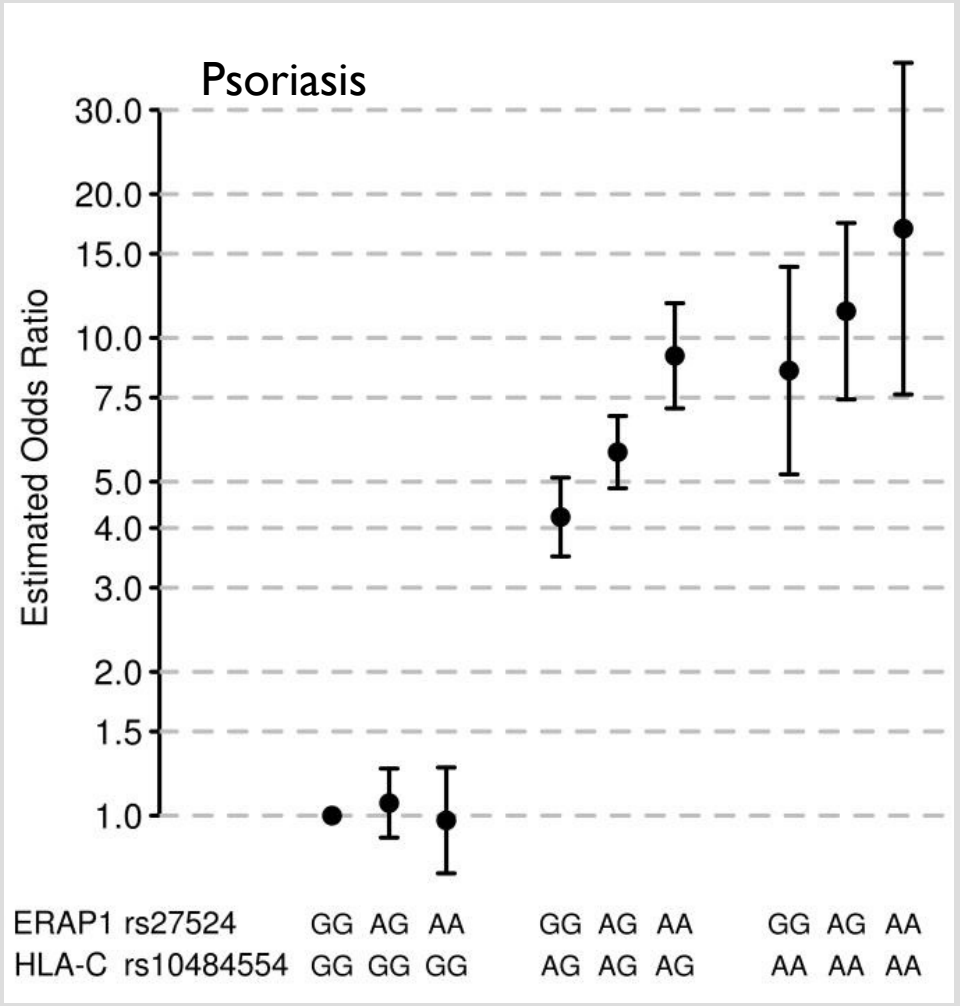
For details see:
Pirinen et al. 2012
Including known covariates can reduce power to detect genetic effects in case-control studies. ***Nat Genet*** 44:848-851.

INTERACTION



Evans et al. Nat Gen 2011

If HLA-B27 genotype is GG, then *ERAP1* SNP has no effect. But when individual carries at least one A, then *ERAP1* SNP has an effect.



Strange et al. Nat Gen 2010

Interaction or *modifier* effect:

Genetic effect varies with the value of the covariate

Here covariate is HLA-allele tagged by a SNP