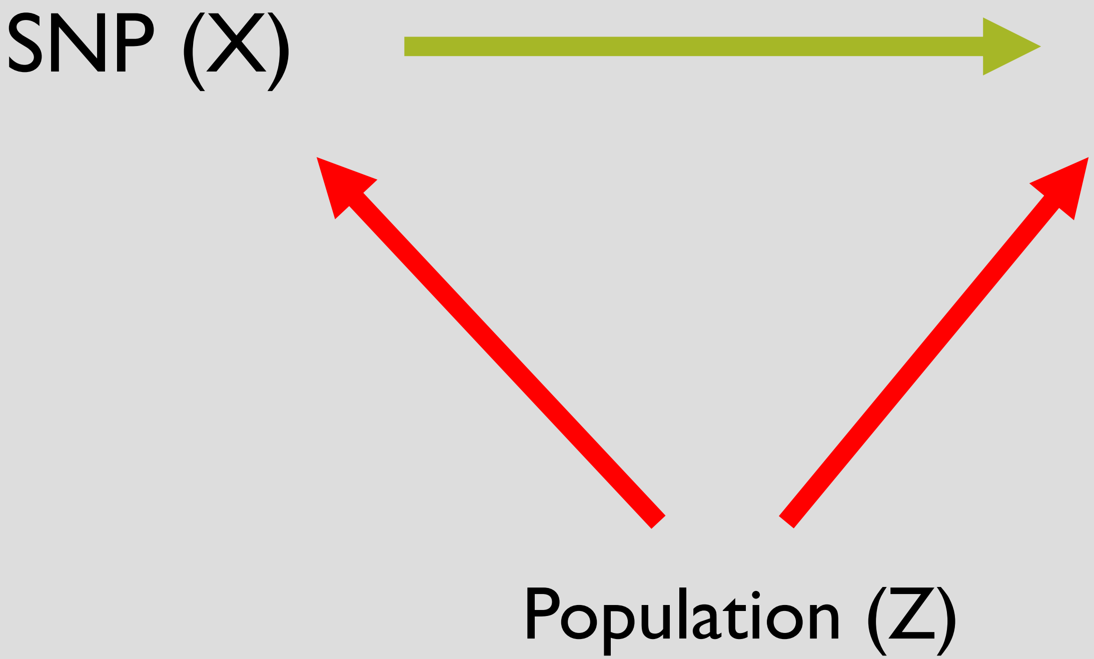
GWAS 6

Matti Pirinen
University of Helsinki

11.11.2020

CONFOUNDER

We want to study X-Y relationship…



SNP (X)

Disease (Y)

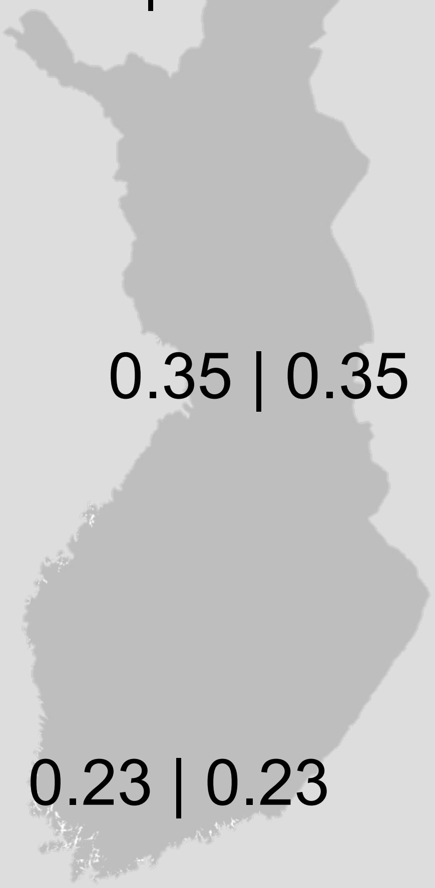
Population (Z)

… 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 ?



Frequencies

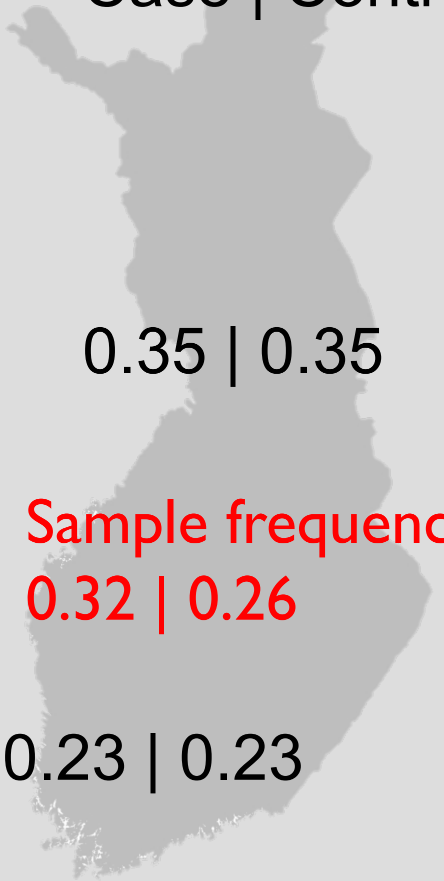
Case | Control

0.35 | 0.35

0.23 | 0.23

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



Frequencies

Case | Control

0.35 | 0.35

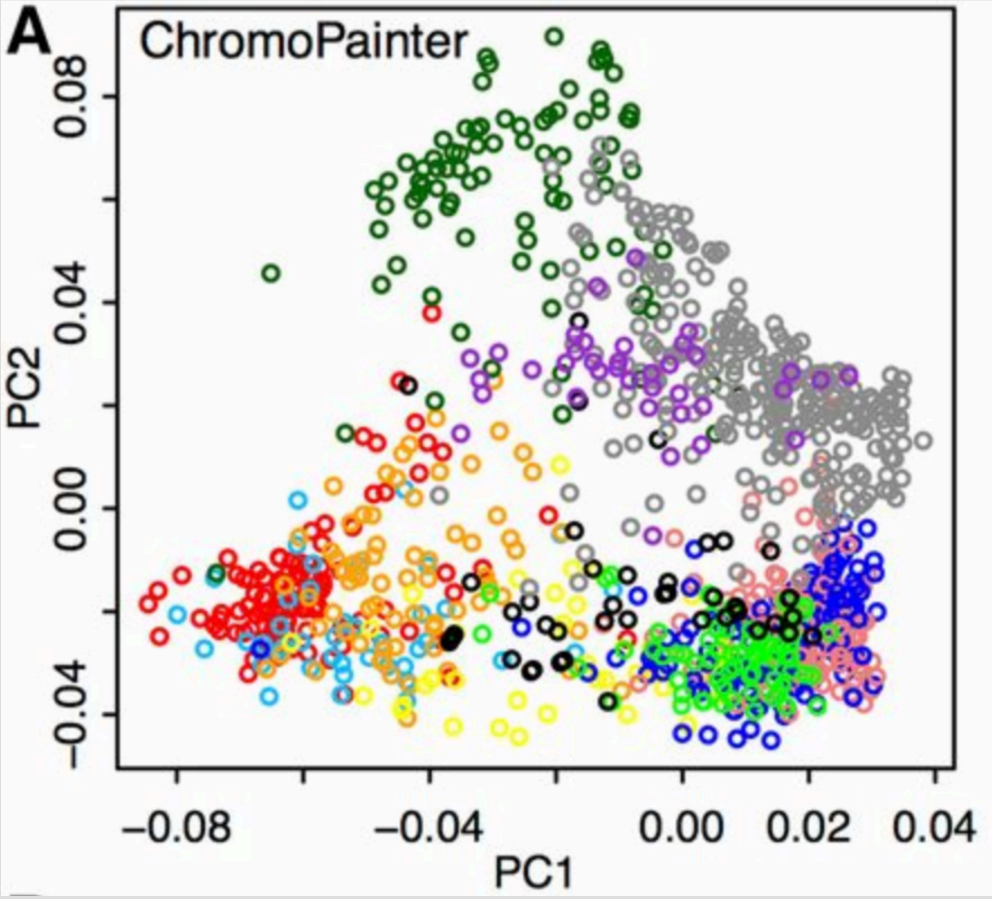
Sample frequencies:

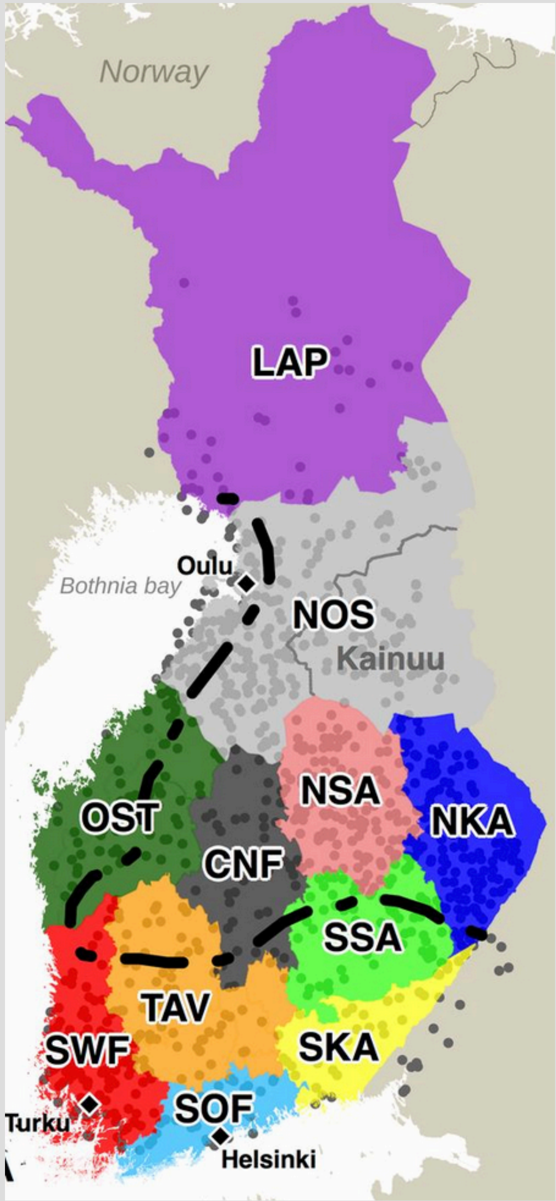
0.32 | 0.26

0.23 | 0.23

USING LEADING PCS TO MATCH CASES & CONTROLS

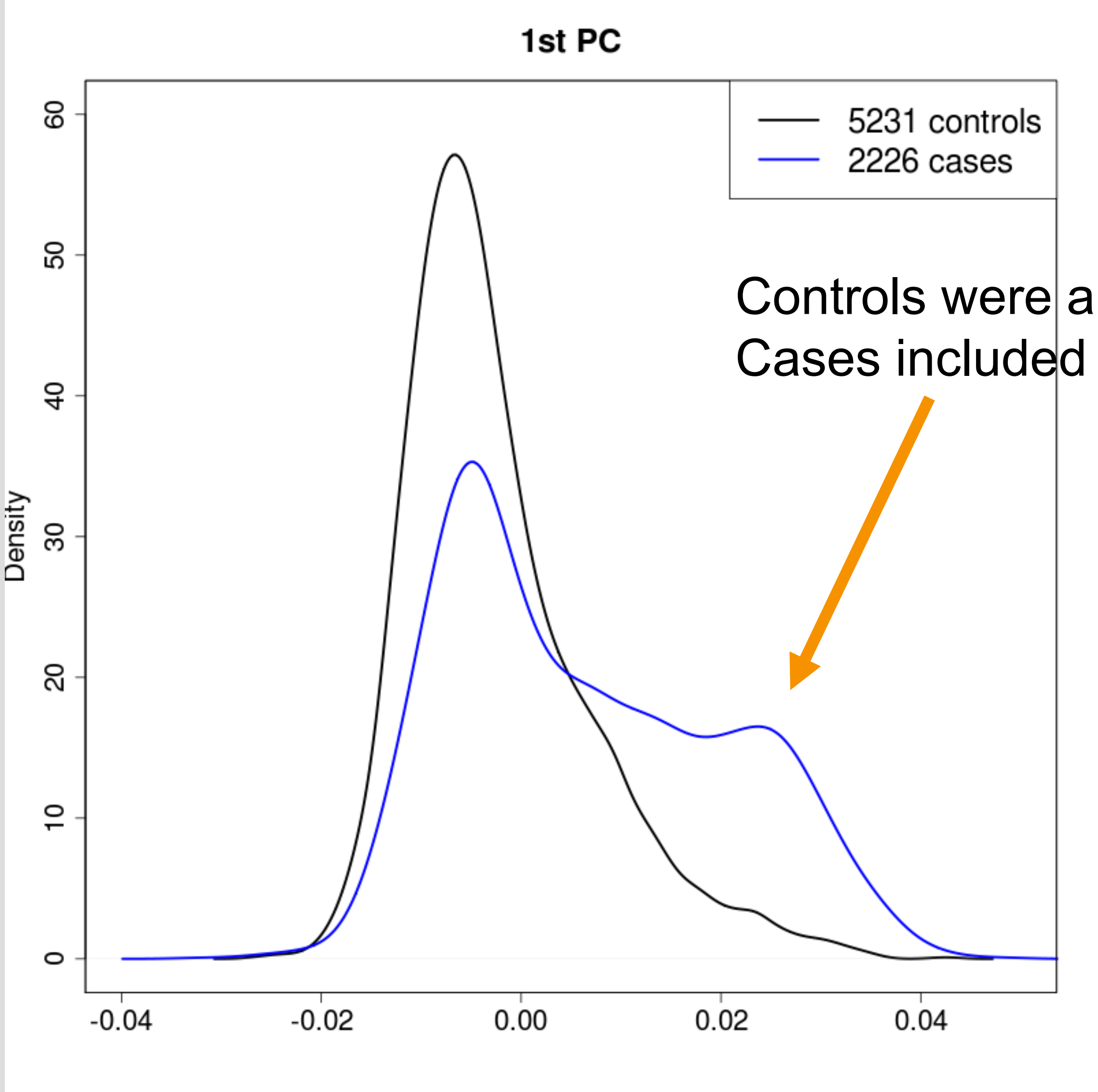
- Often we do not know regional origins of samples or they may not be informative of genetic background
- But we can infer genetic similarity and adjust the analyses for that by taking leading PCs of the genetic correlation matrix and use them as covariates (= additional predictors) in the regression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





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

EXAMPLE FROM A PSORIASIS STUDY IN UK

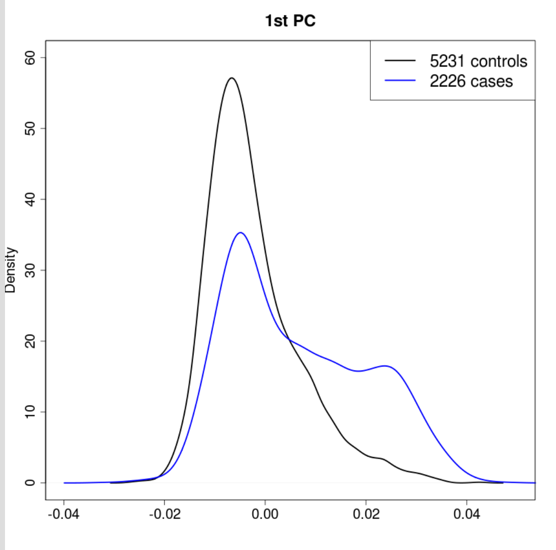


Controls were all from the UK.
Cases included 500 Irish samples.

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

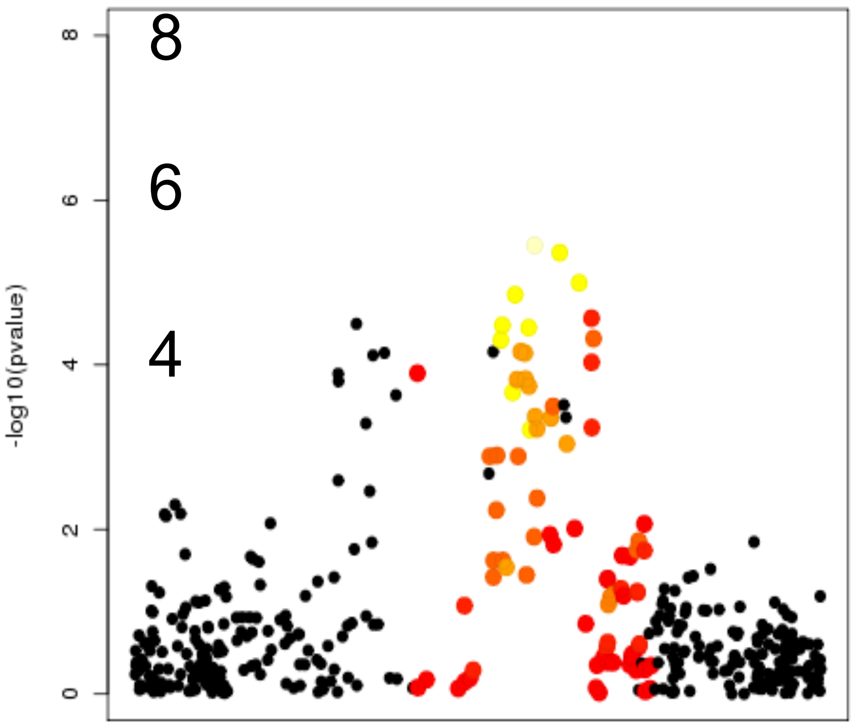
Strange et al. 2011 Nature Genetics

EXAMPLE FROM A PSORIASIS STUDY IN UK



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

We have P-values 1-e6 in this region.



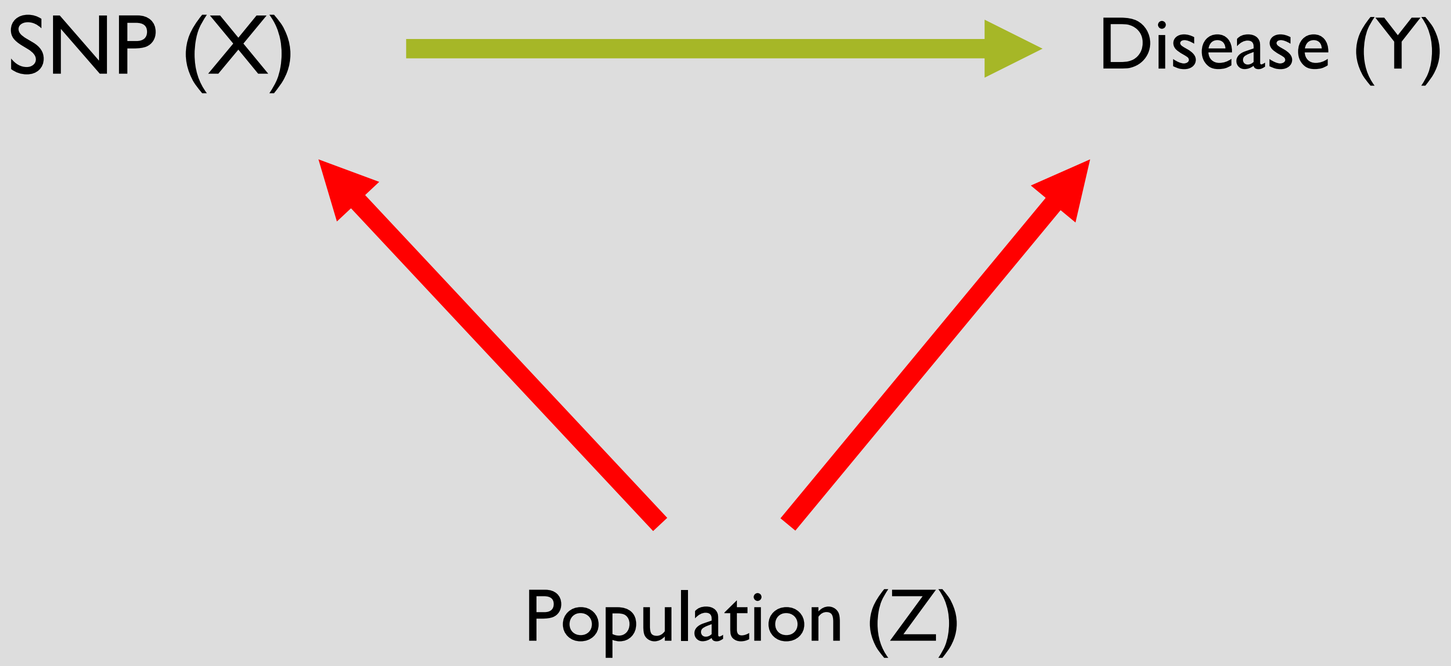
Region around lactase gene

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



SNP (X)

Disease (Y)

Population (Z)

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.

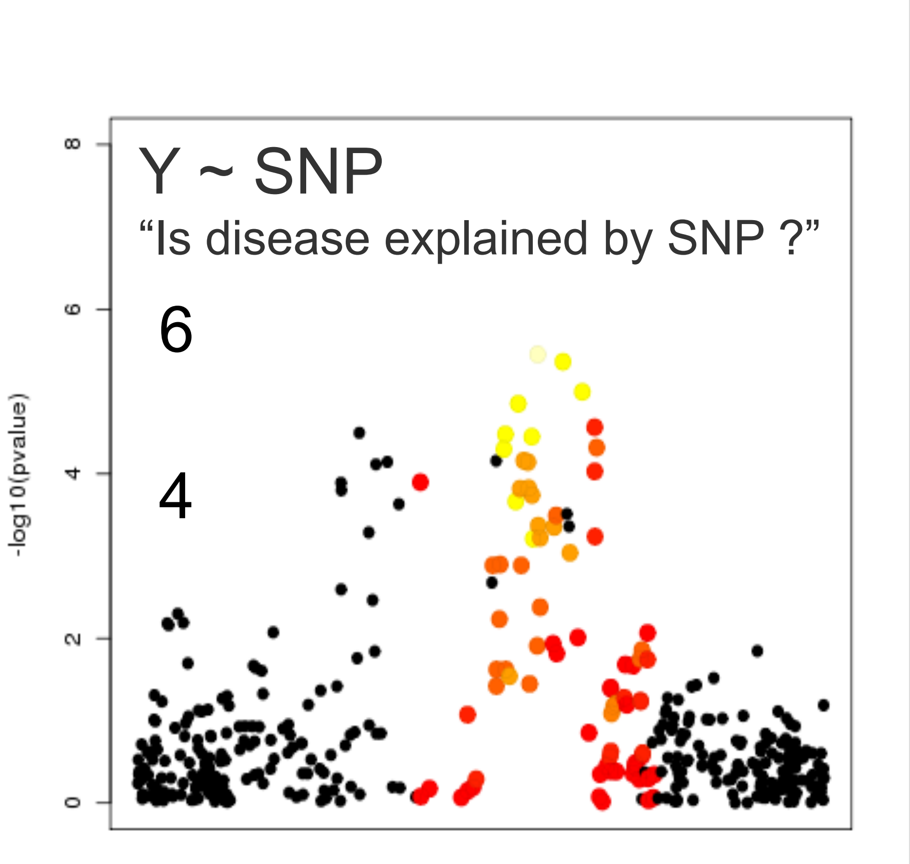
b = effect of SNP for a fixed value of Z

b* = effect of SNP without accounting for Z

EXAMPLE FROM A PSORIASIS STUDY IN UK

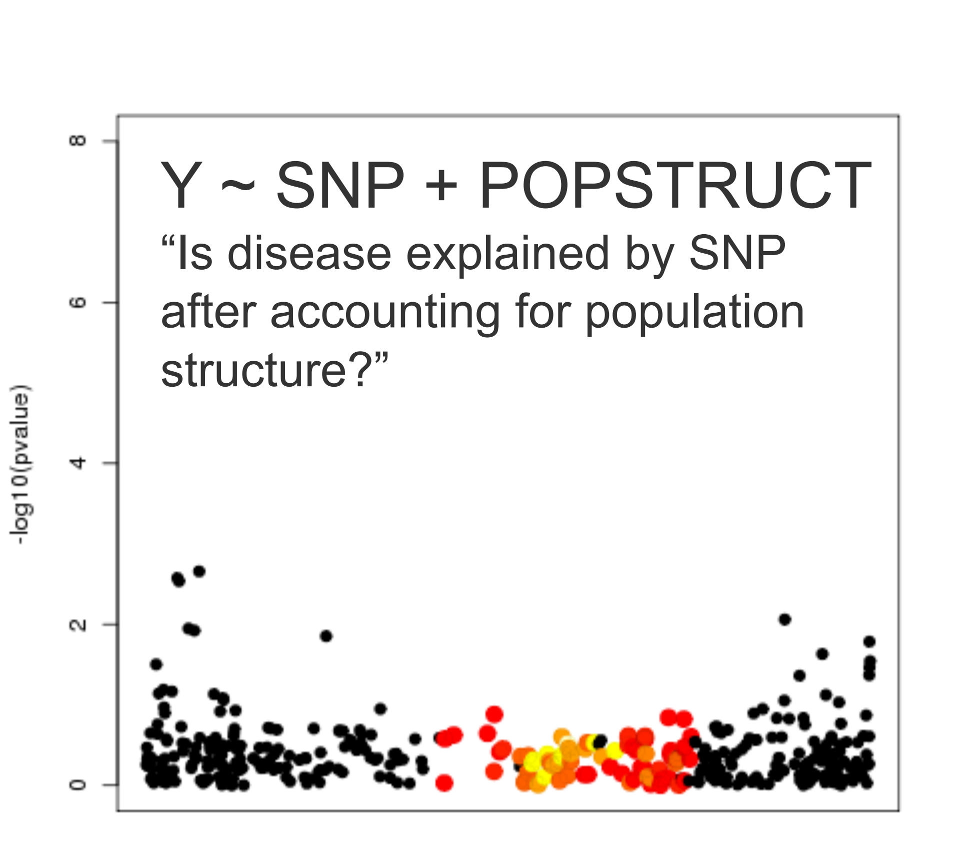
Does lactase gene really affect psoriasis susceptibility?

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



Y ~ SNP

“Is disease explained by SNP ?”



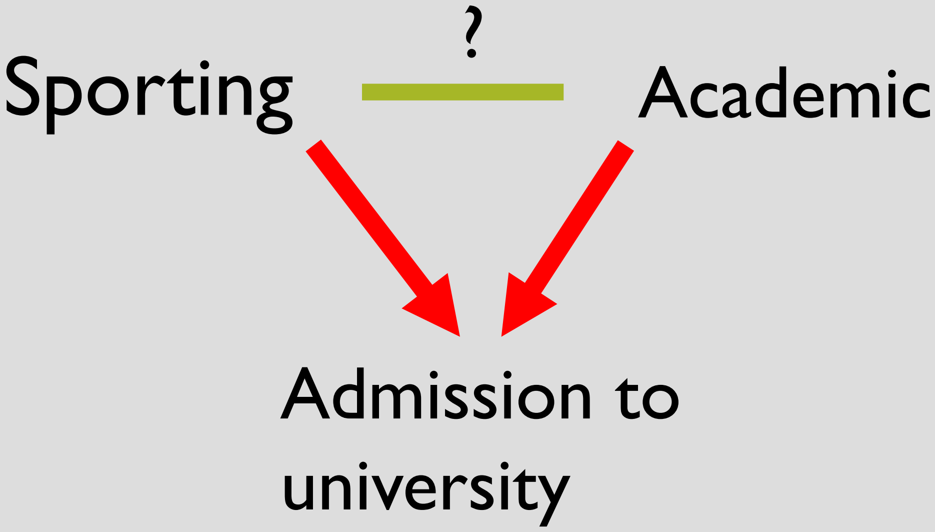
Y ~ SNP + POPSTRUCT

“Is disease explained by SNP after accounting for population structure?”

Strange et al. 2011
Nature Genetics

EXAMPLE OF COLLIDER BIAS

Nat Communications 11: 5479 (2020)

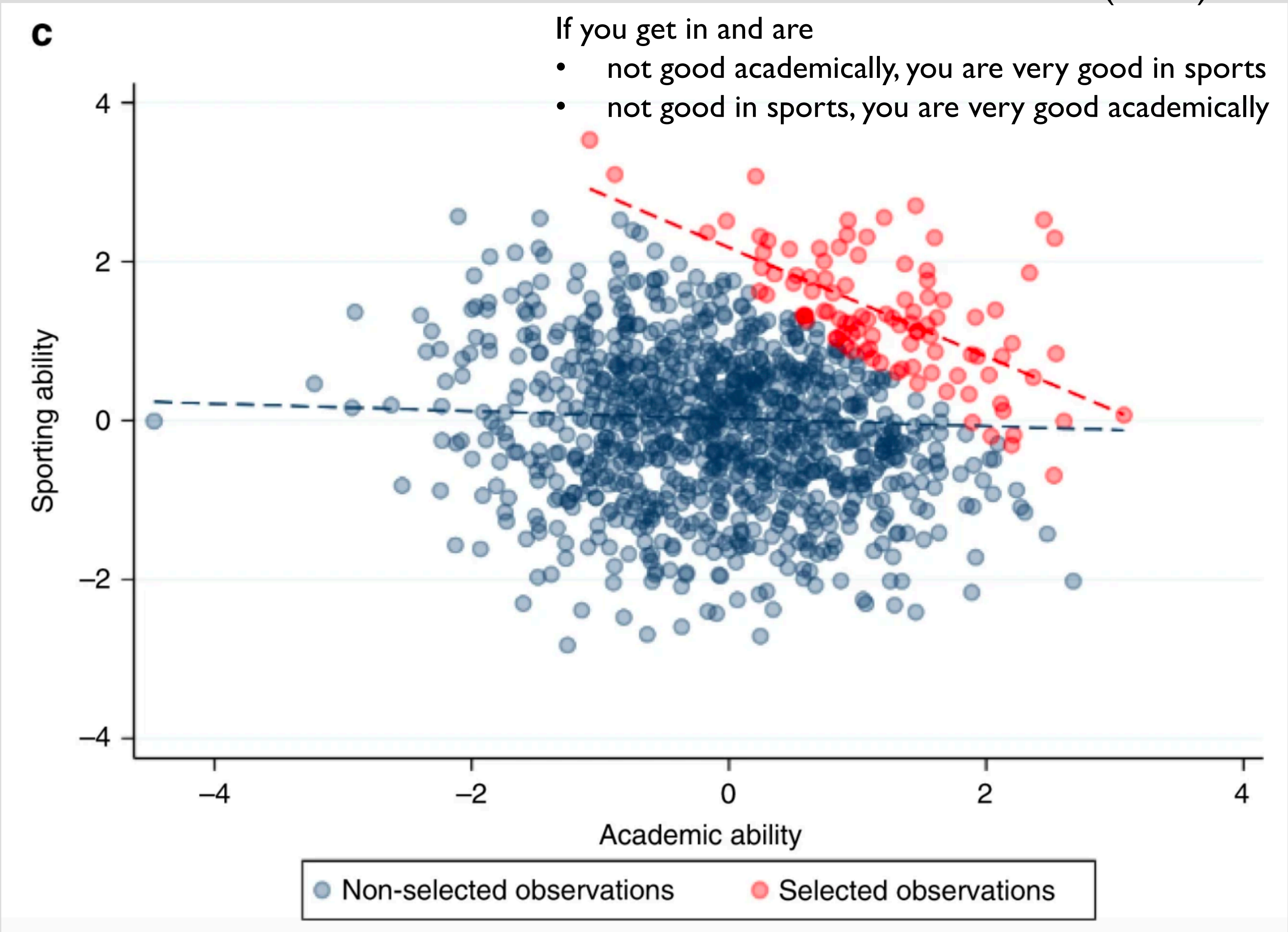


Sporting

?

Academic

Admission to university



If you get in and are

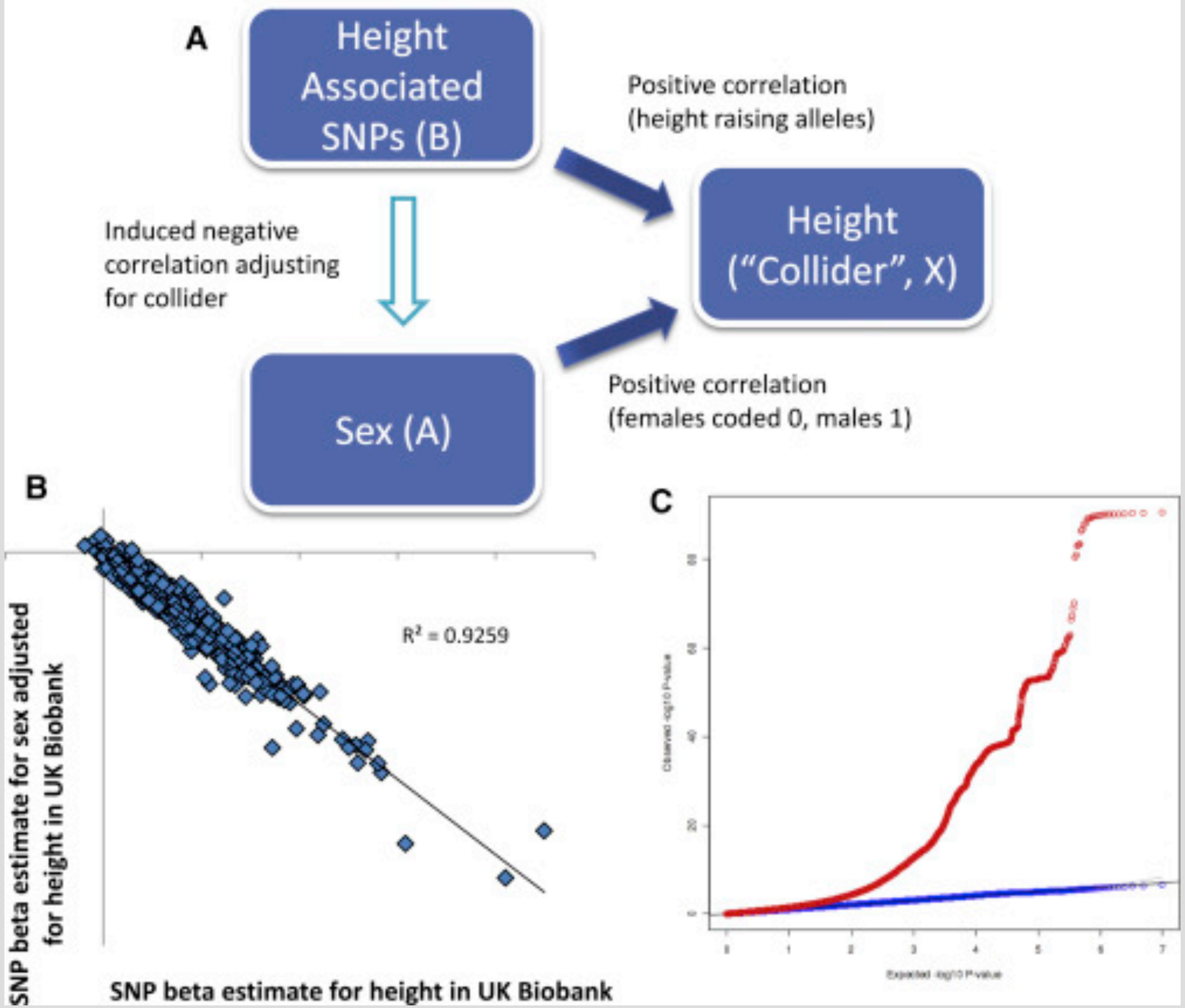
- not good academically, you are very good in sports
- not good in sports, you are very good academically

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

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

COLLIDER BIAS USING UK BIOBANK

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



- 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 height adjusted analysis, blue from the analysis without height as covariate
- 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 INDEPENDENT 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 ($\beta = 0$), X and W are independent and the population follows the model:

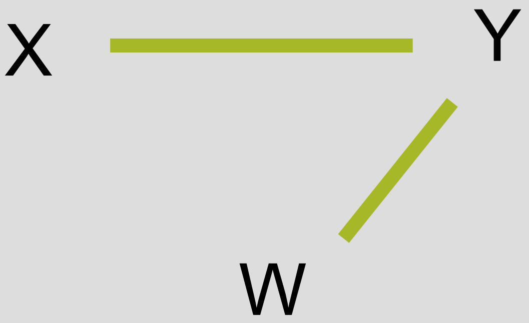
  $Y \sim \mu + X \cdot 0 + W\gamma$

  then also according to the regression model

  $Y \sim \mu^* + X\beta^*$,

  $\beta^* = 0$
- 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, so which model should we use?

Non-confounding covariate

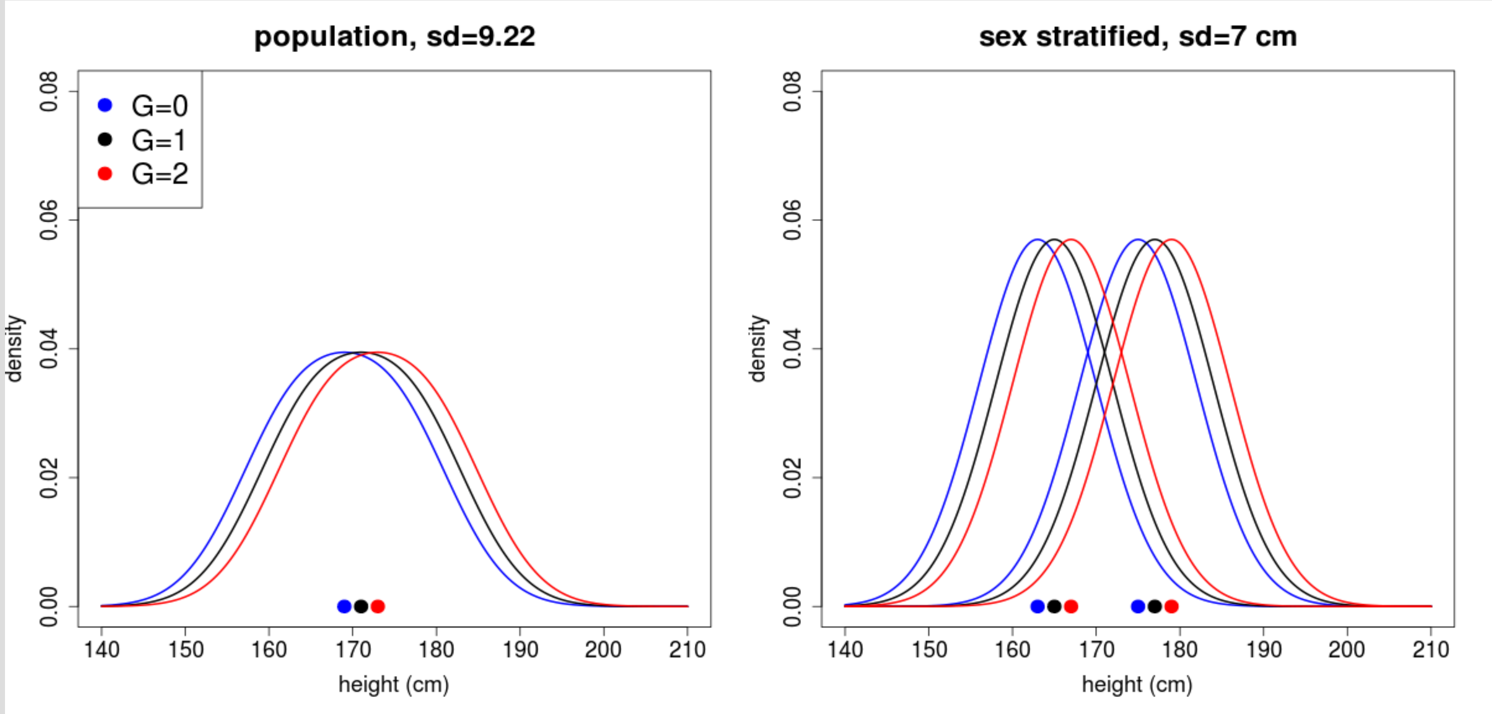


INDEPENDENT COVARIATES

- We consider two models when X and W are independent
  - Model M: $Y \sim \mu + X\beta + W\gamma$
  - Model M*: $Y \sim \mu^* + X\beta^*$
- 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

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

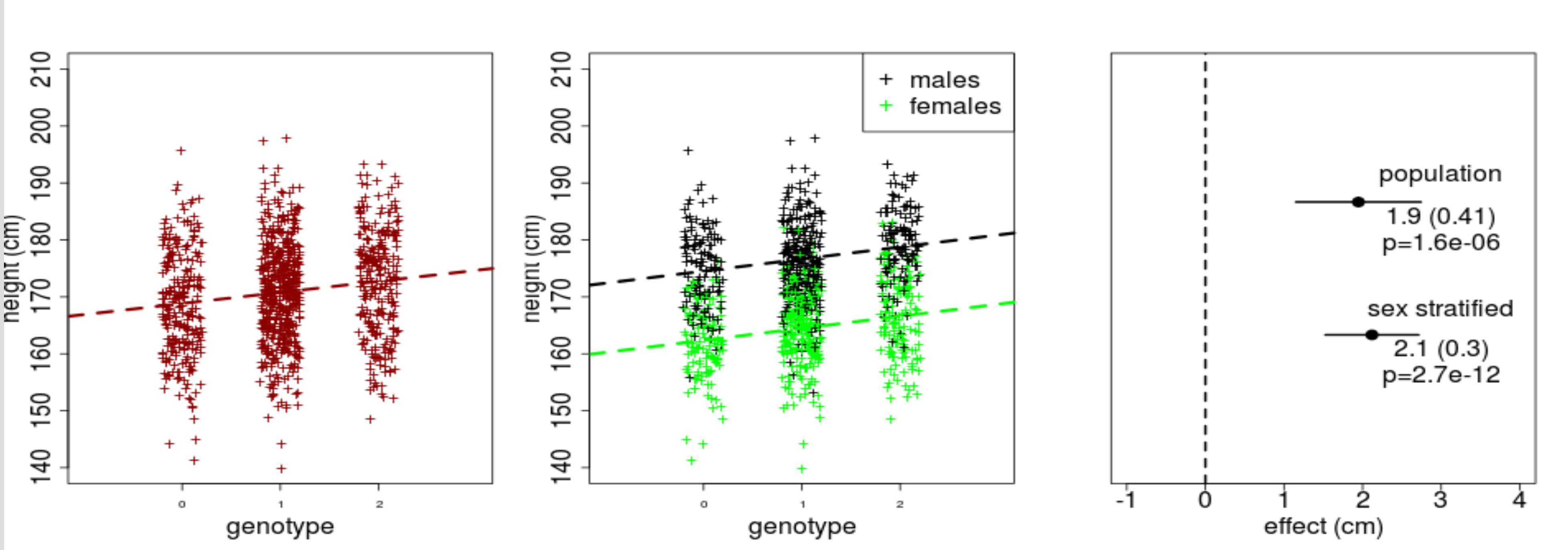


	X=0	X=1	X=2
Female	163	165	167
Male	175	177	179
Population	169	171	173

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 covariate has smaller SE and hence higher power to detect non-zero β than the model without 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.

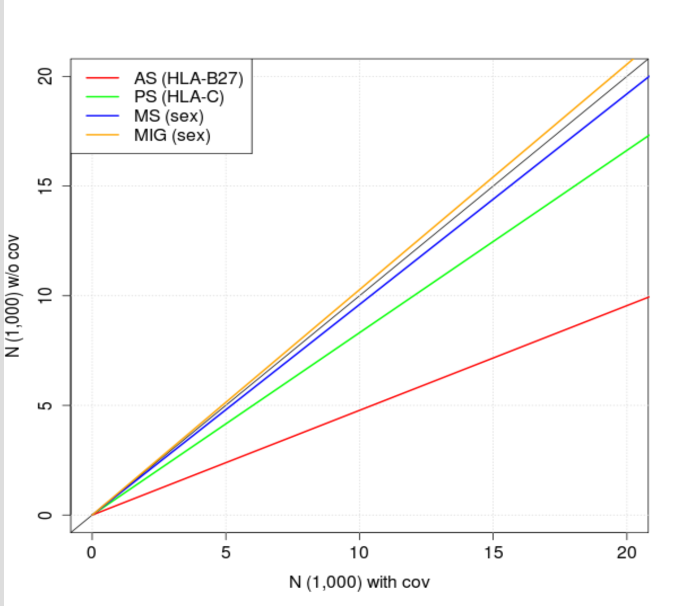
NON-CONFOUNDING COVARIATE & LOGISTIC REGRESSION OF POPULATION DATA

- This setting is relevant for population biobanks but is **not a typical GWAS setting** which would be the case-control ascertainment (next slide)
- It can be shown, that when applying logistic regression in a population sample where X and W are independent, then
  - $|\beta^*| \leq |\beta|$
  - SE of β^* $\leq$ SE of β
  - Power to detect that β is non-zero is larger than that β^* is non-zero
- If we are interested in effect size, then the model should be chosen based on whether the covariate-adjusted effect size is more relevant than the effect size without covariate adjustement
- In GWAS setting, where 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 low-prevalence diseases

NON-CONFOUNDING COVARIATES AND POWER IN CASE-CONTROL STUDIES

	Prevalence	Risk factor	Odds-ratio	Frequency	SSM
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

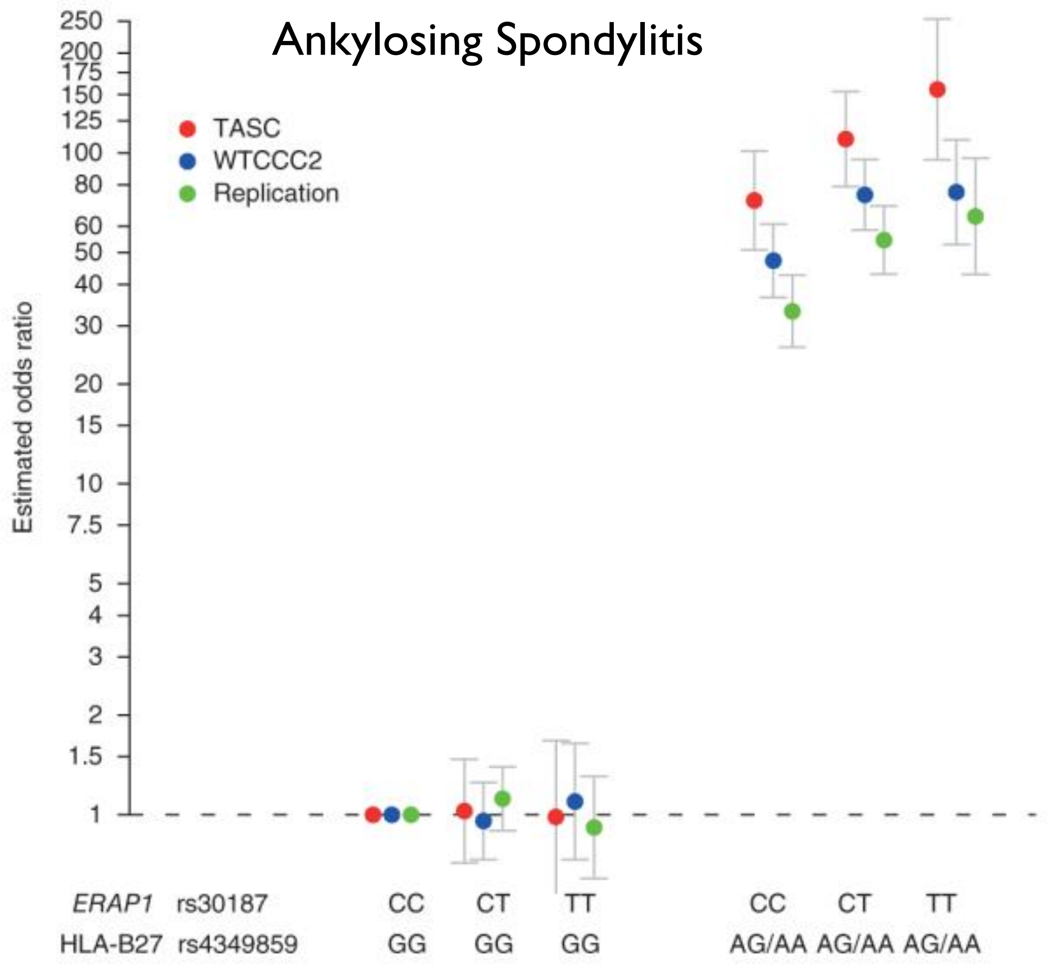
Which model to prefer depends on the disease prevalence in population!



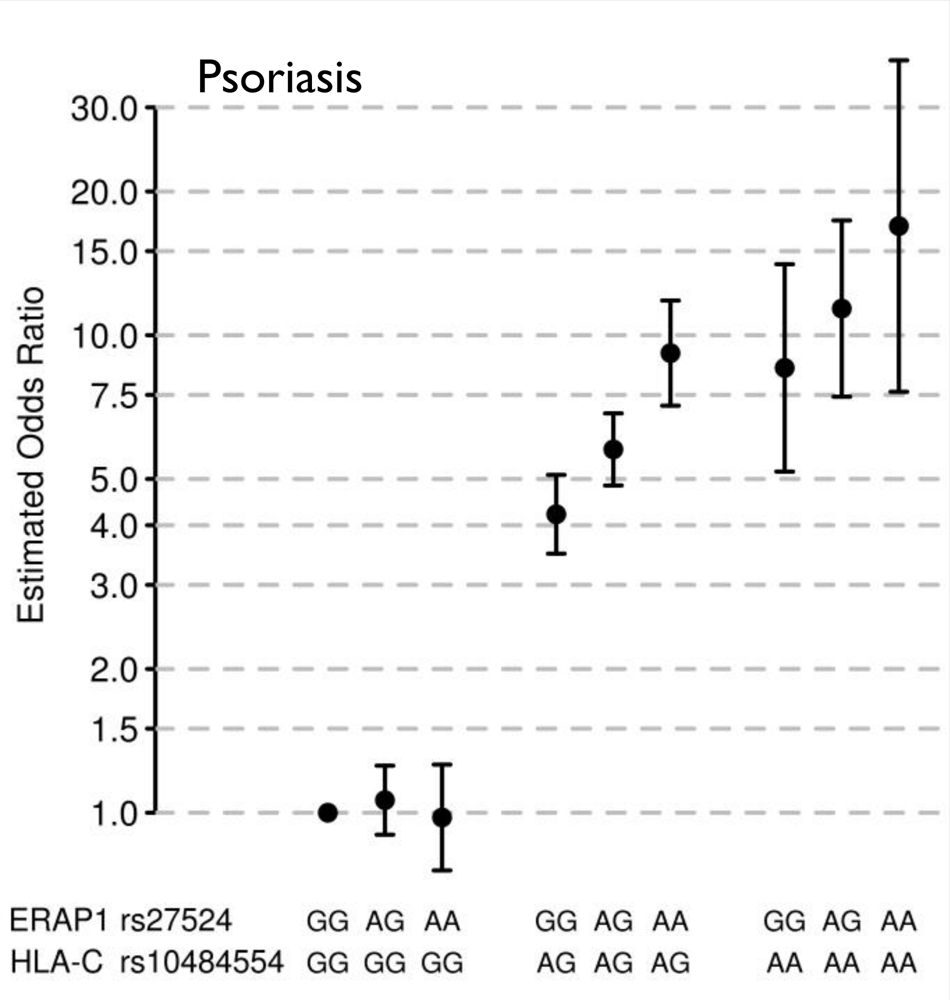
Sample size multiplier (SSM) is the ratio of sample sizes for the same power without and with the covariate. (Same as ratio of NCPs.)

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

INTERACTION



Evans et al. Nat Gen 2011



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