GWAS 6

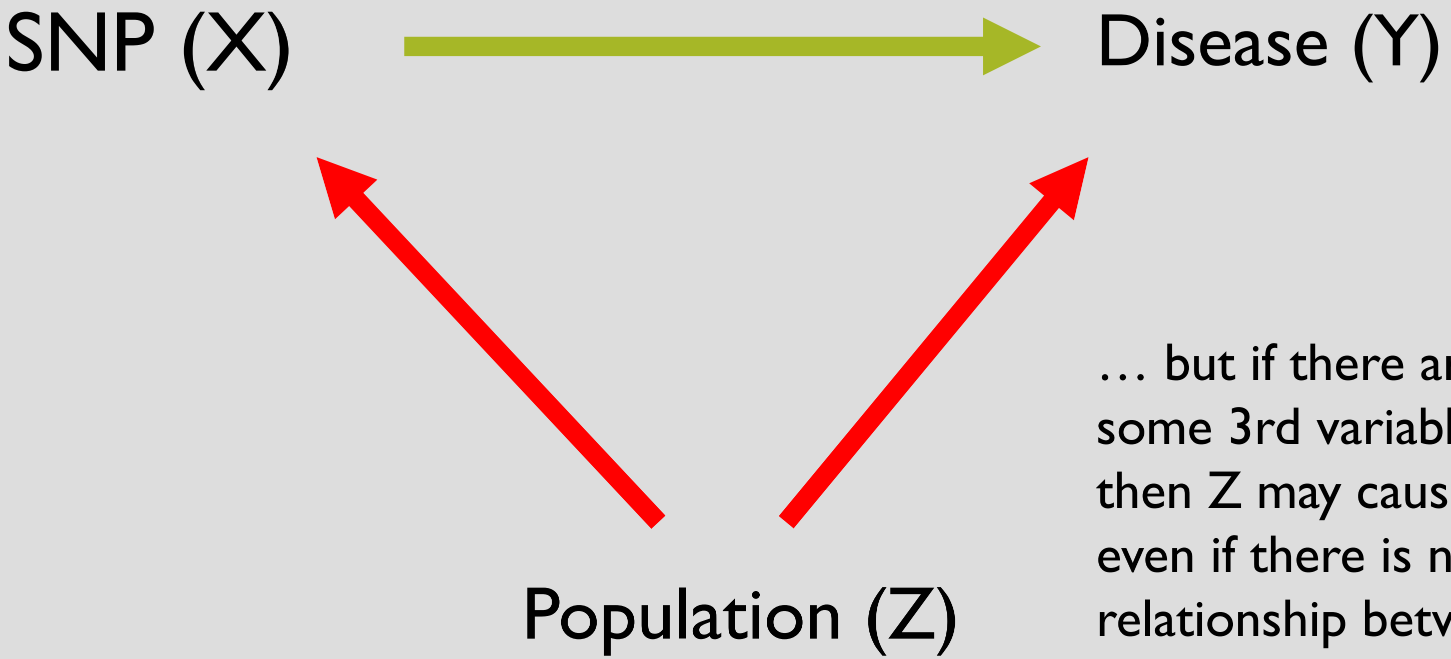
Matti Pirinen

University of Helsinki

30.1.2019

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

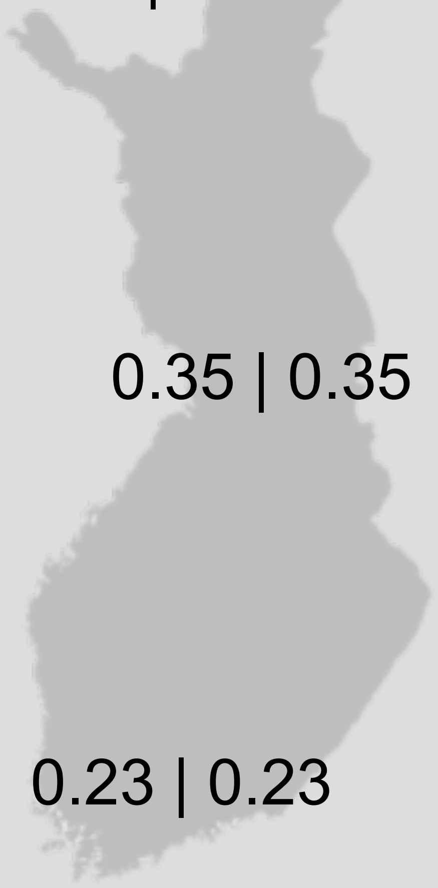
Z is **confounder** of X-Y association

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 ?

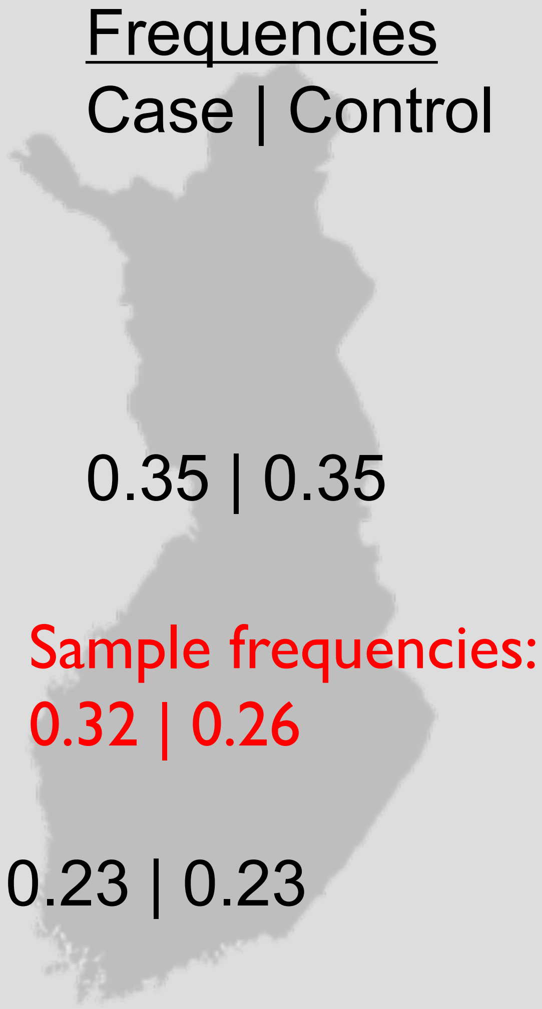
Frequencies

Case | Control



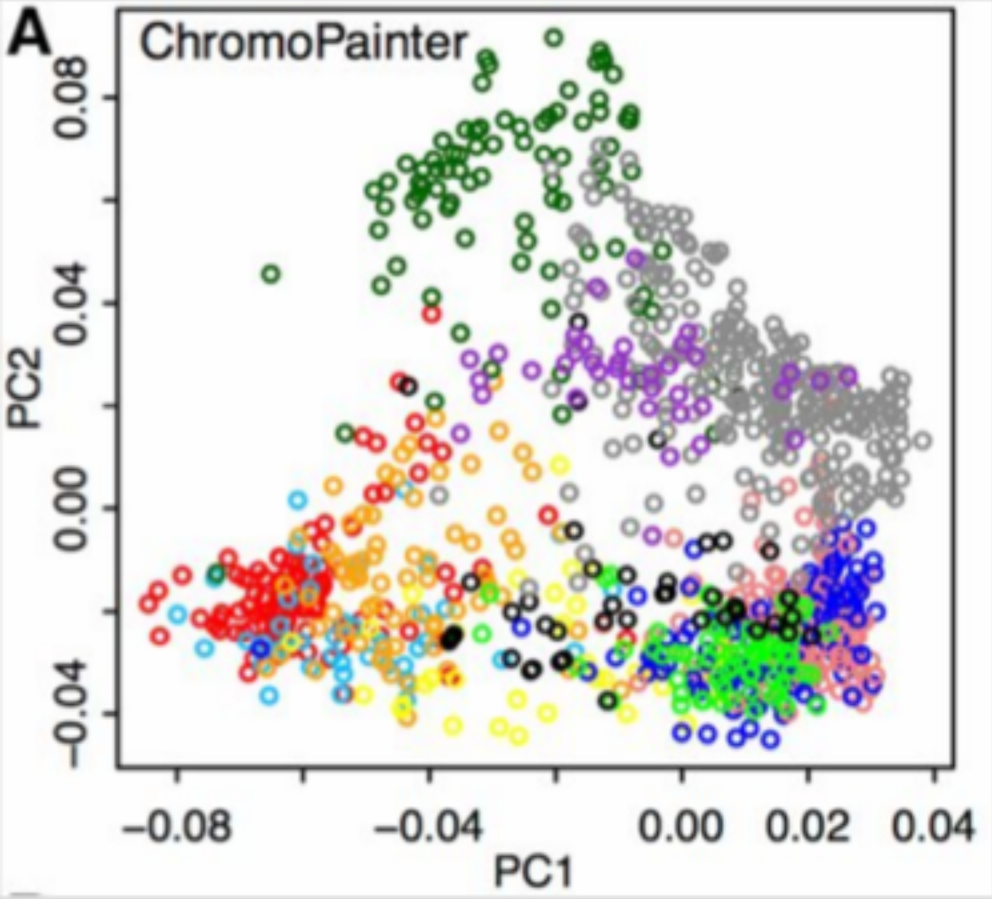
CONFOUNDING BY ANCESTRY

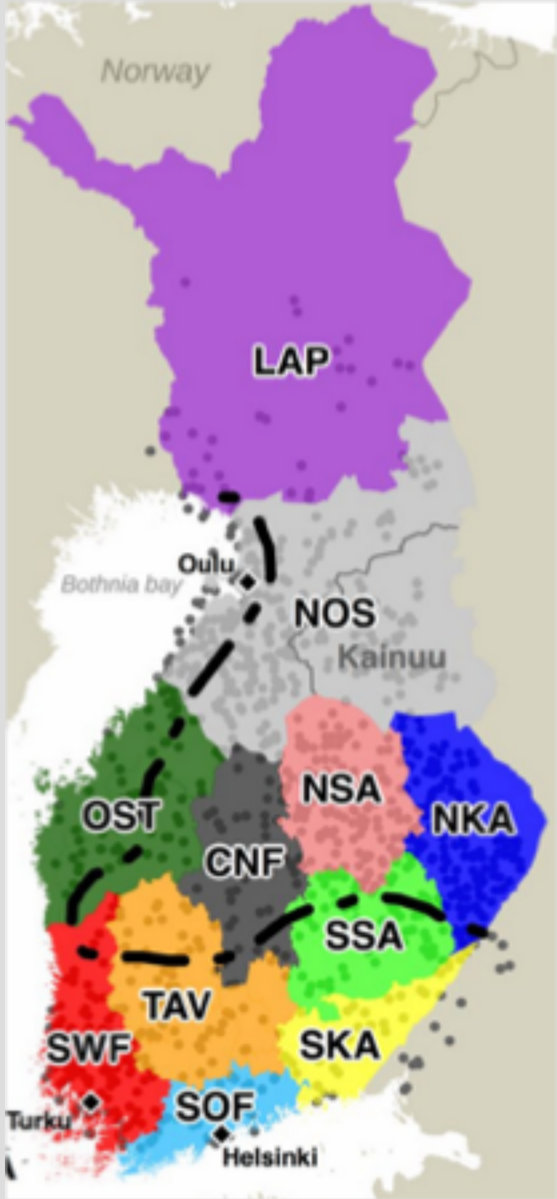
- SNP that has no effect on heart disease but has different regional frequencies
  - Variant ”A” frequency 0.23 in Helsinki region
  - Variant ”A” frequency 0.35 in Oulu region
- Consider sampling
  - 2000 cases (500 from H and 1500 from O).
    - ”A” frequency in cases is 0.32
  - 2000 controls (1500 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 confounds the analysis



USING LEADING PCS TO MATCH CASES & CONTROLS

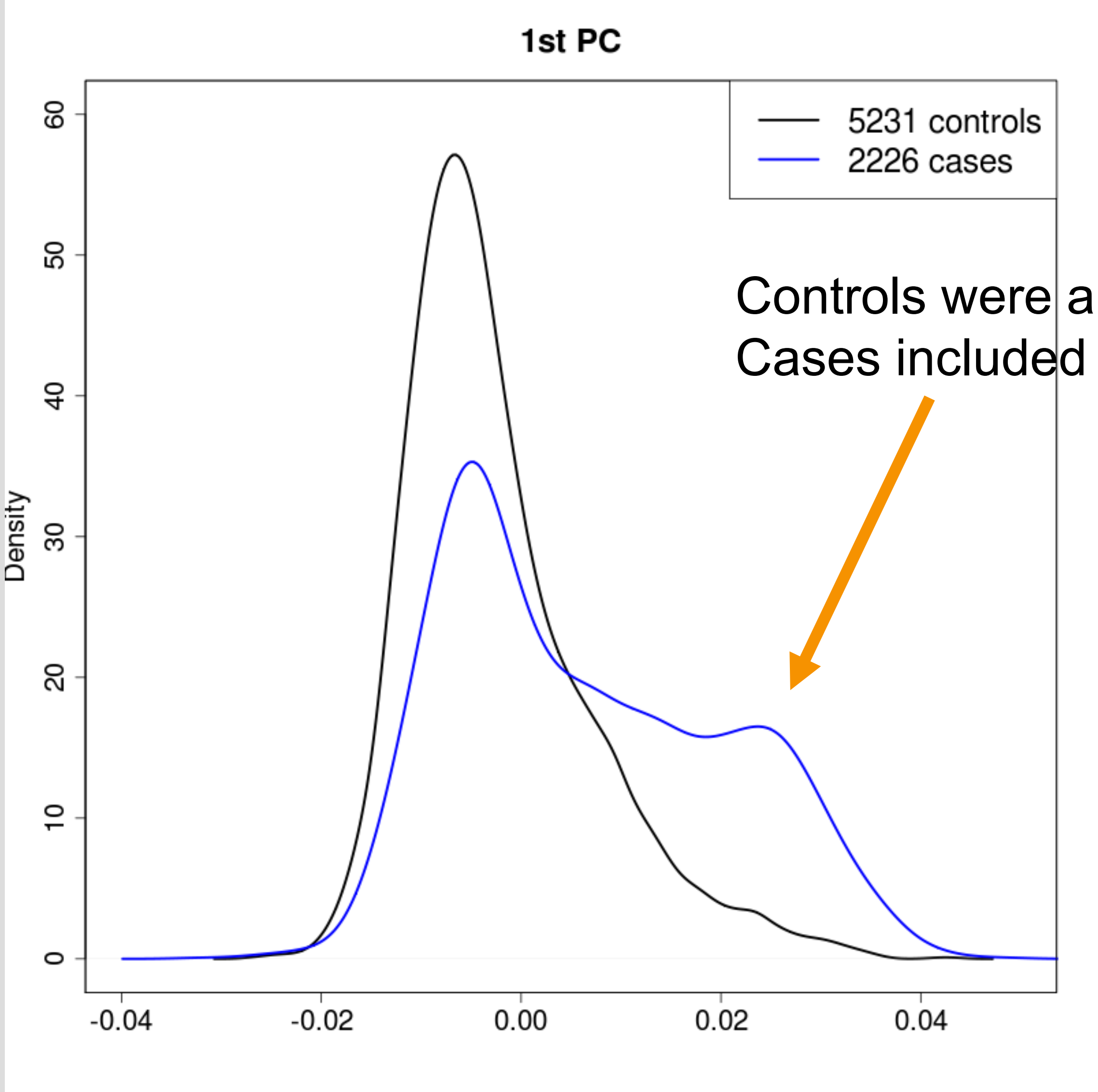
- Often we do not know regional origins of samples or they may not be informative of genetic background
- But we can infer genetic similarity and adjust the analyses for that by taking leading PCs of the genetic correlation matrix and use them as covariates (= additional predictors) in the regression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





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

EXAMPLE FROM A PSORIASIS STUDY IN UK

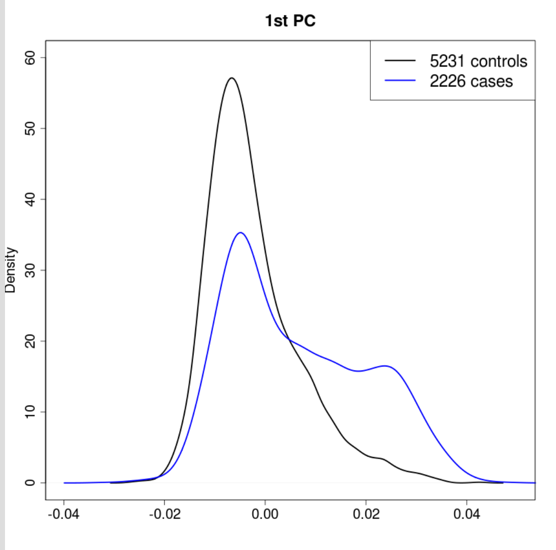


Controls were all from the UK.
Cases included 500 Irish samples.

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

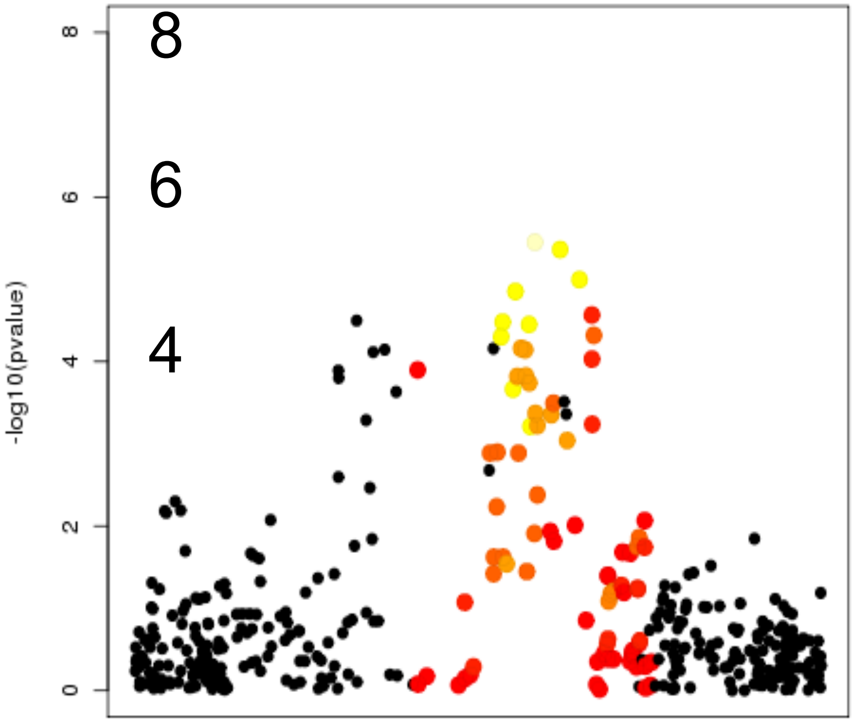
Strange et al. 2011 Nature Genetics

EXAMPLE FROM A PSORIASIS STUDY IN UK



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

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



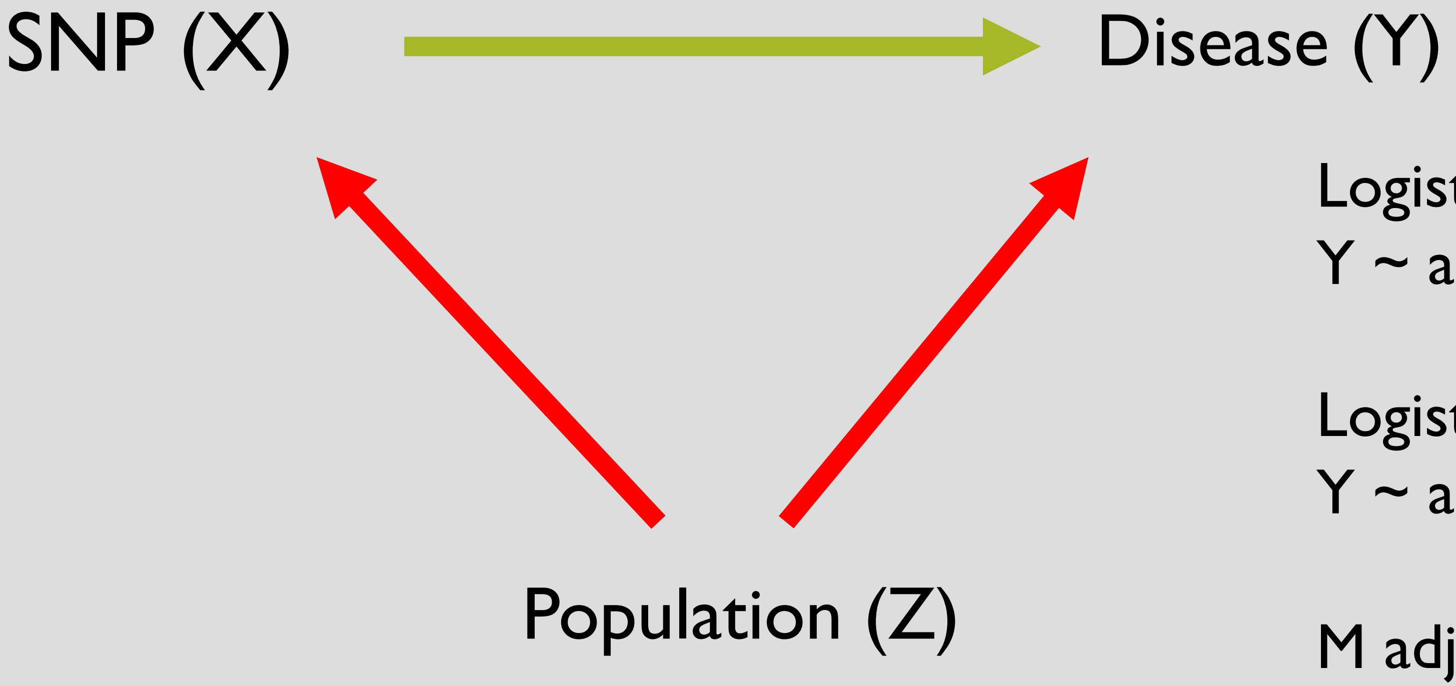
Region around lactase gene

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:
$Y \sim a + Xb + Zc$

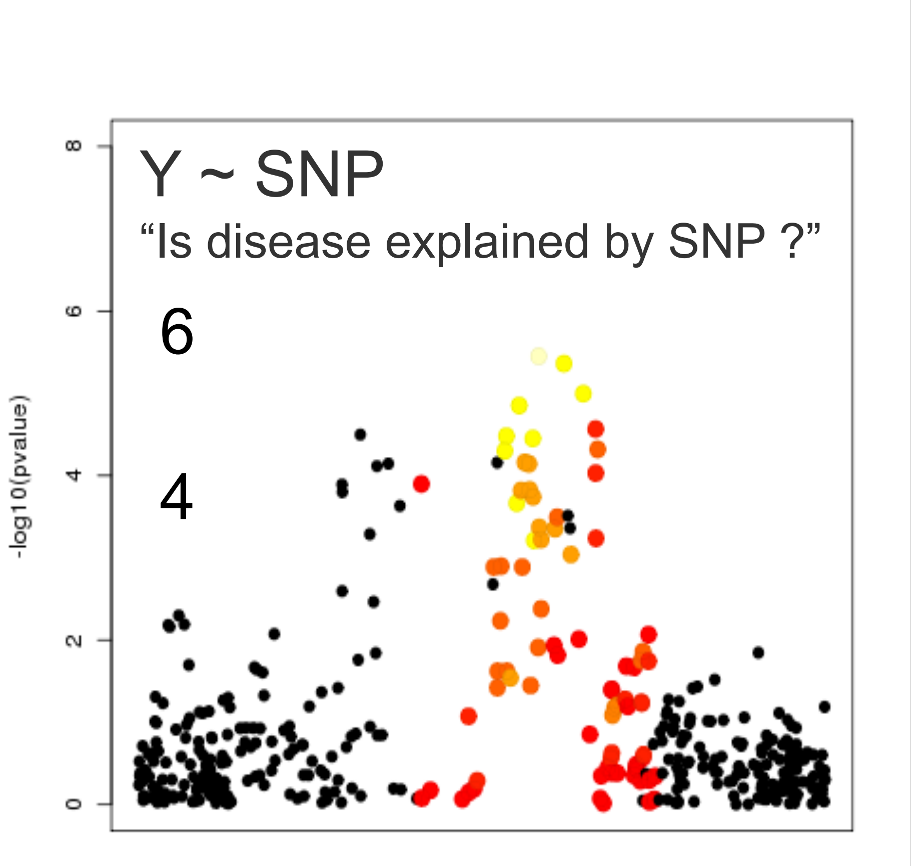
Logistic regression model M*:
$Y \sim a^* + Xb^*$

M adjusts for confounder Z whereas M* does not. Therefore a significant b* from M* may be caused by confounding

EXAMPLE FROM A PSORIASIS STUDY IN UK

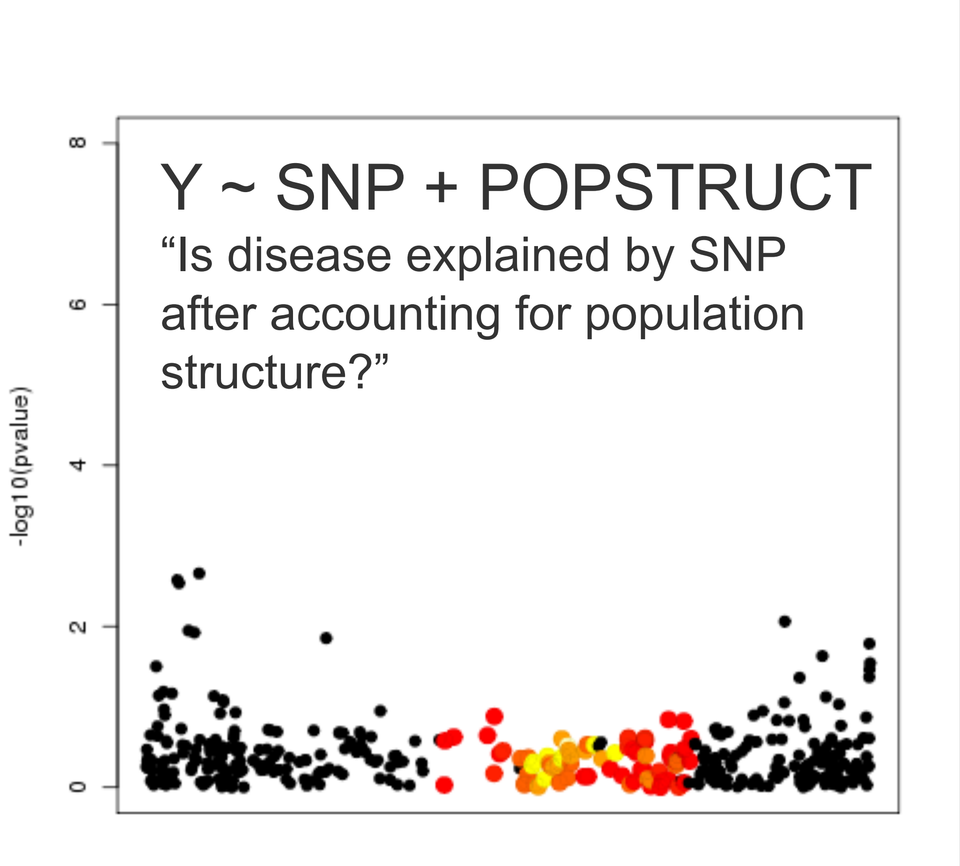
Does lactase gene really affect psoriasis susceptibility?

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