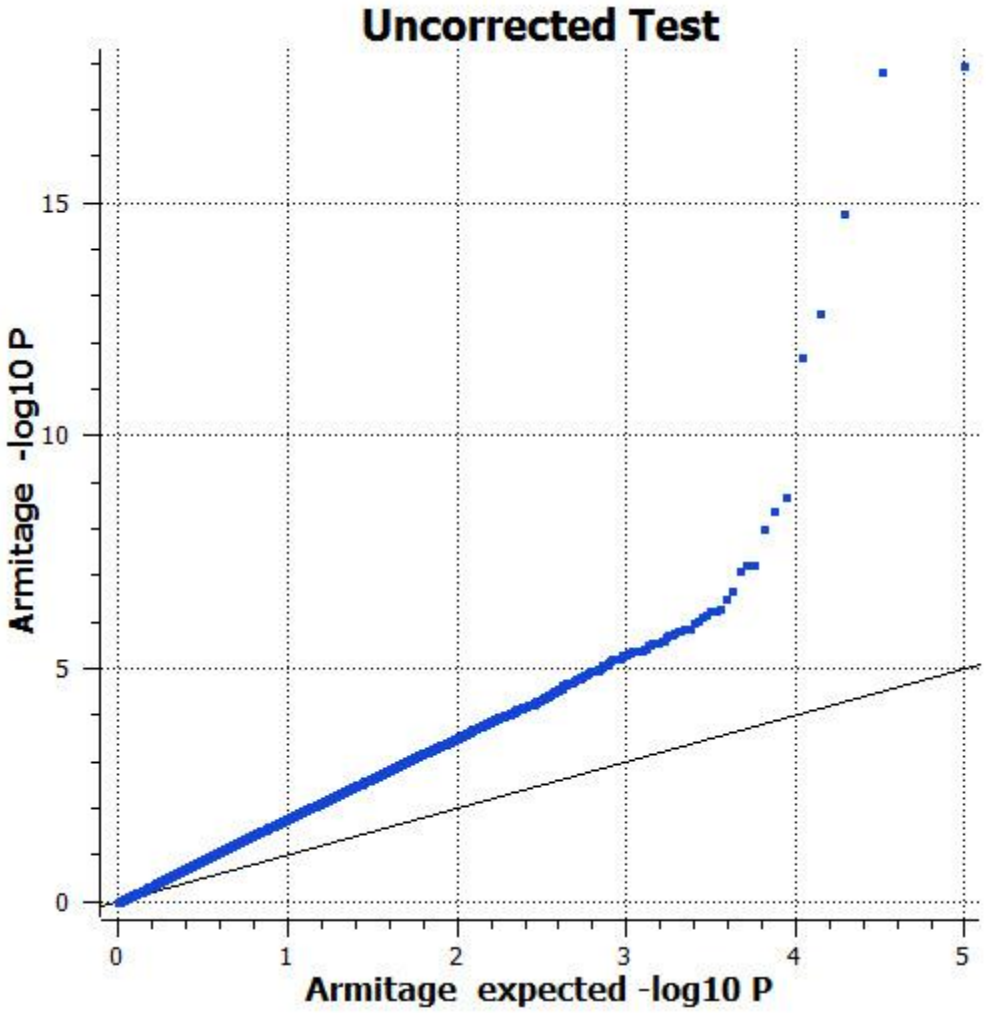
Population stratification

Why it is bad when we are doing association studies and what measures we can take to make sure it does not affect our results in ways we don't want it to affect them.

What is population stratification and why is it bad?

- Differences between cases and controls that happen not-quite-by chance but are picked up by association studies.
- Will affect overall distribution of test statistics.
- Association models often assume HWE or something similar:
  - Constant allele frequencies
  - Random mating
  - Infinite population size
  - No mutations
  - No migration/flow
- Of these, random mating is most often violated and introduces population structure/stratification.



https://blog.goldenhelix.com/population-structure-genetic-background-environment-mixed-model/mlm3/

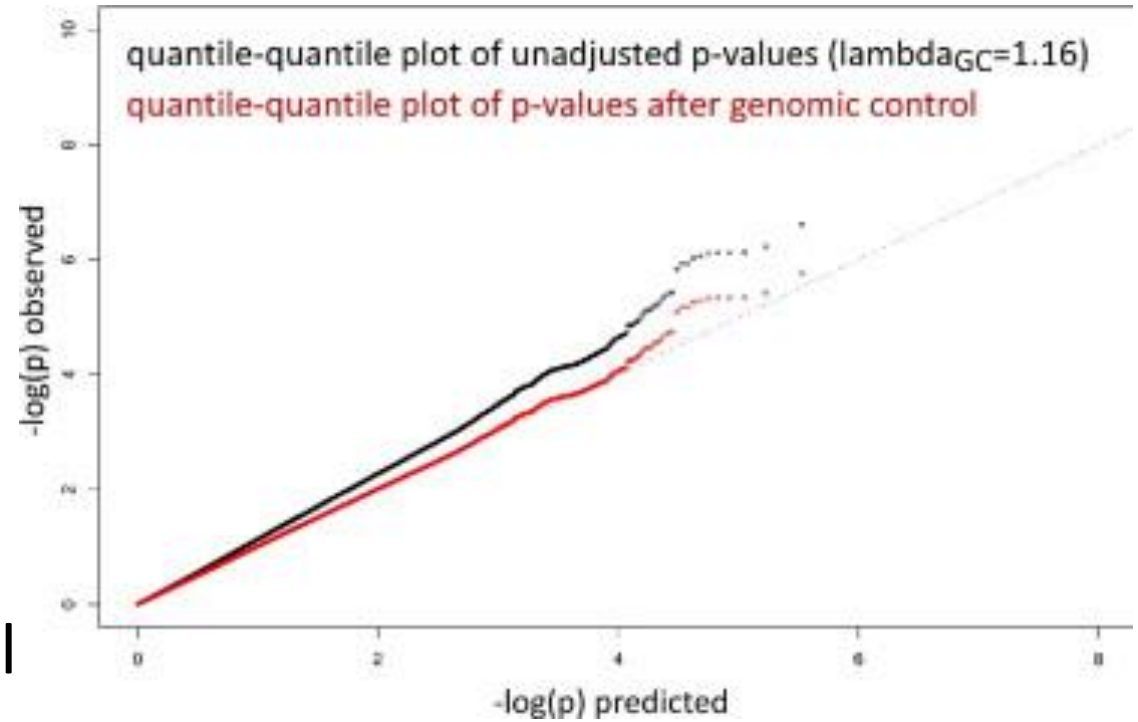
Genomic control

- The classical approach (Devlin and Roeder 1999)
- Define a **genomic inflation factor**

  λ = median(χ^2)/0.456

  and adjust the test statistics:

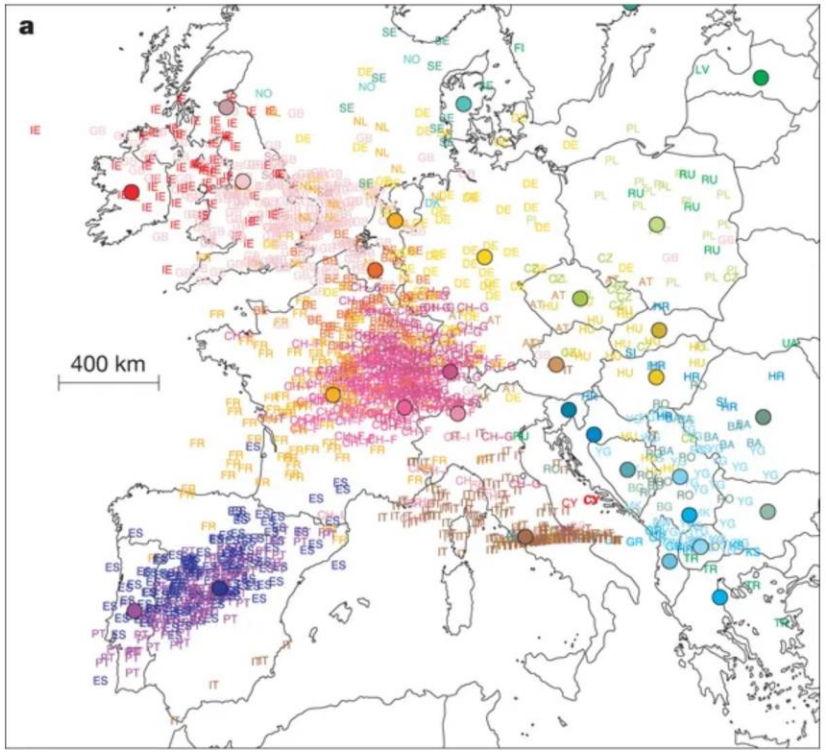
  $\chi^2_{adj} = \chi^2/\lambda$
- Has some assumptions that may fail
  - "Few" SNPs are truly causal
  - All SNPs have the same inflation factor

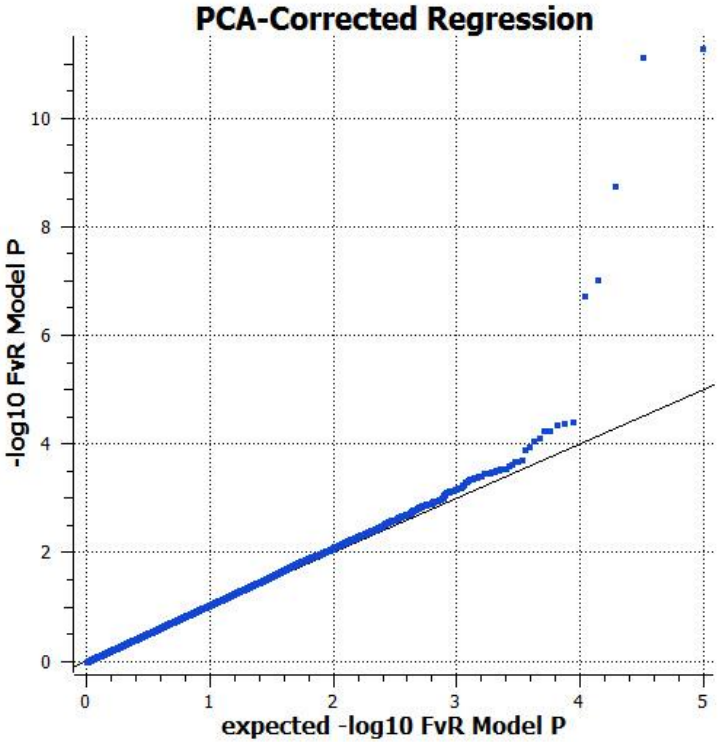


Park et al. (2013) Asian Australas. J. Anim. Sci. 26:1523-1528

Principal components

- Principal components from the population are added as covariates to the model.
- Principal components usually capture large parts of population stratification and cryptic relatedness.



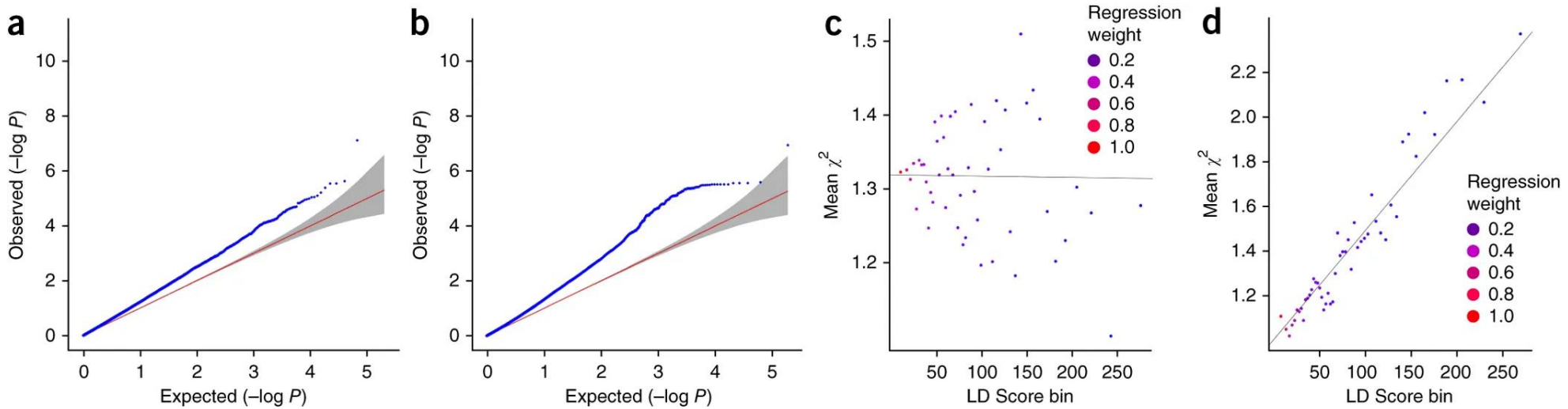


https://blog.goldenhelix.com/population-structure-genetic-background-environment-mixed-model/mlm4/

Novembre, J., Johnson, T., Bryc, K. *et al.* Genes mirror geography within Europe. *Nature* 456, 98–101 (2008) doi:10.1038/nature07331

LDscore regression

- In the polygenic model, $E[\chi^2|s_j] = Nh^2s_j/M + Na + 1$, where
  - N is the number of individuals;
  - M is the number of SNPs, such that h^2/M is the average heritability per SNP;
  - a is the contribution of confounding biases;
  - $s_j = \Sigma_k r^2_{jk}$ is the LD-score of variant j.
- Individual SNPs weighted by their LD-score
- Distinguishes between polygenicity and bias:



Bulik-Sullivan, B., Loh, P., Finucane, H. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47, 291–295 (2015) doi:10.1038/ng.3211

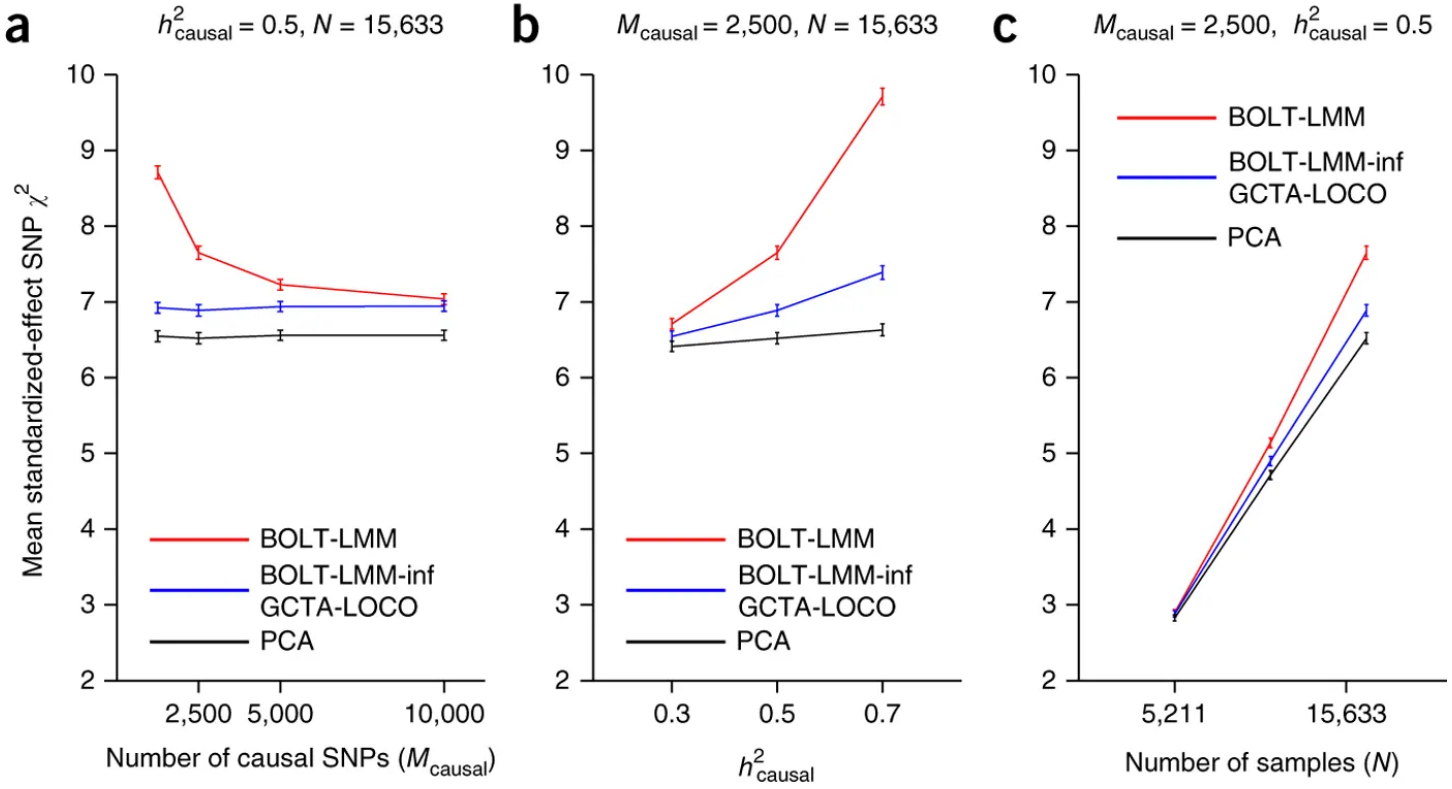
Mixed models

- Mixed models allow us to
  - model relatedness/population stratification,
  - jointly model all SNPs.
- We model phenotype as

  $\mathbf{y} = \mathbf{X\beta} + \mathbf{Wu} + \boldsymbol{\varepsilon}$,

  with

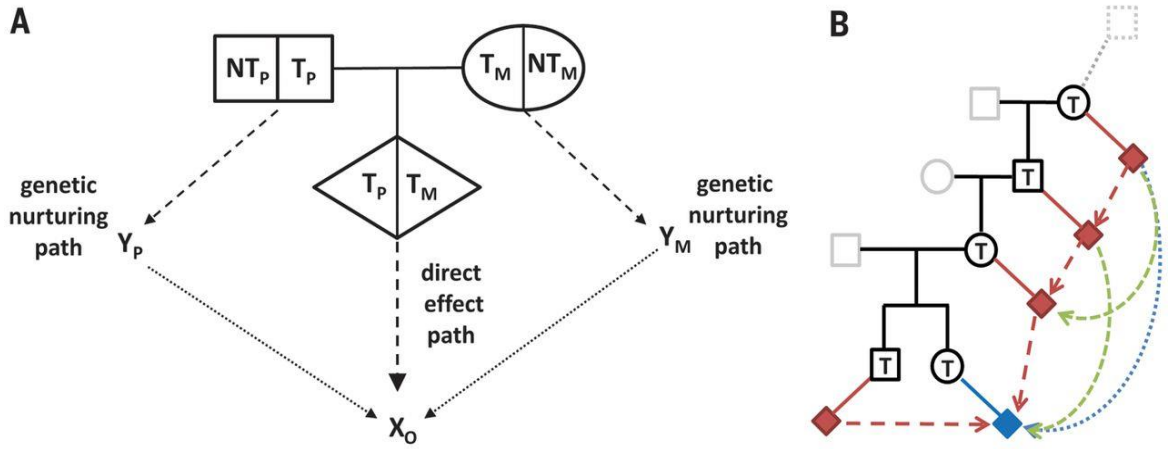
  $\mathrm{Var}(\mathbf{y}) = \mathbf{WW}^{T}\sigma^{2}_{u} + \mathbf{I}\sigma^{2}_{\varepsilon}$.
- BOLT-LMM works around the computational difficulties.
- Assuming a non-infinitesimal model greatly increases power.
- SNPs assumed to have mixed-normal prior effect sizes

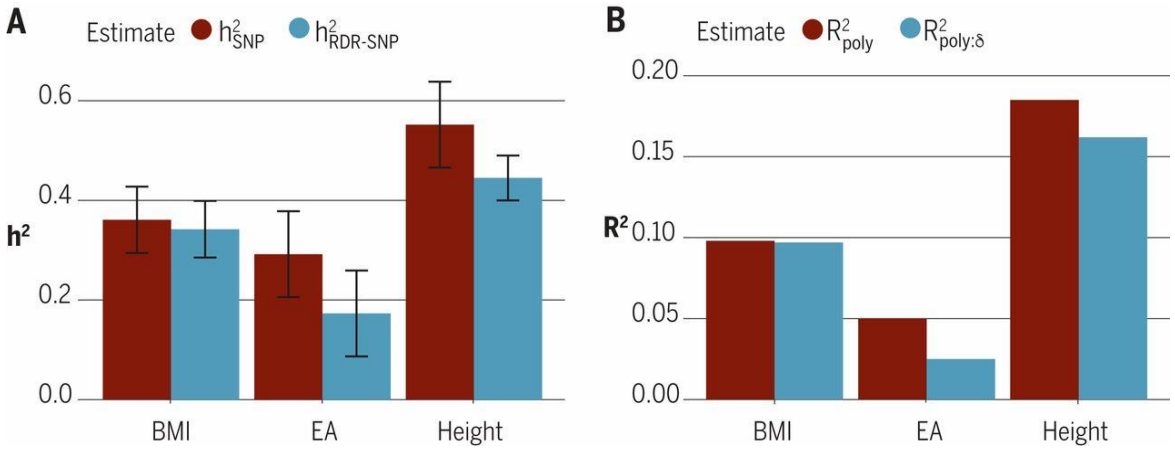


Loh, P., Tucker, G., Bulik-Sullivan, B. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 47, 284–290 (2015) doi:10.1038/ng.3190

But what are the biasesand limitations we're missing?

- We don't really know what all but the top few PCs are capturing
- Ascertainment bias in race/geography/socio-economic status
- Indirect genetic effects: We do not include information on families
- And probably like a million more





ALEXANDER I. YOUNG, STEFANIA BENONISDOTTIR, MOLLY PRZEWORSKI, AUGUSTINE KONG, *SCIENCE*, 2019