Clarifying a few things from last week

Patrick Turley, 17 June 2018

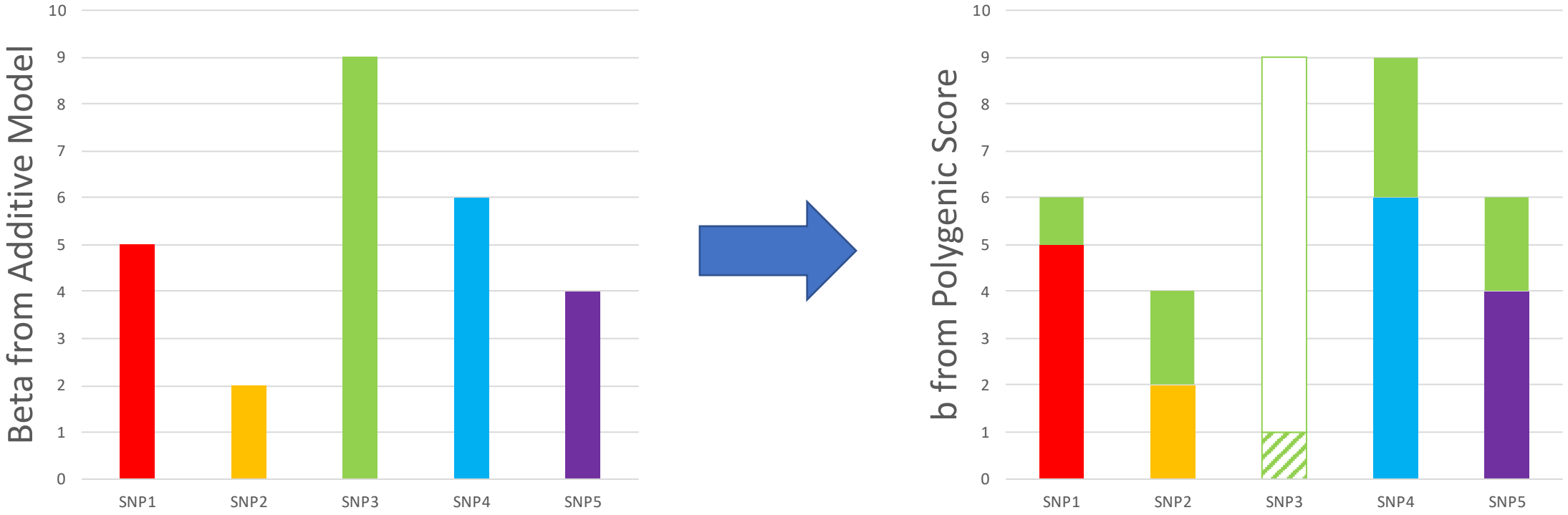
Several Small Things

- Slides for Polygenic Score lecture have been corrected
  - Should have been r_g^2 rather than r_g for expected R^2 when there is cross-cohort heterogeneity (or you are predicting across phenotypes)
- If SNP weights for a PGS, $\hat{b}_j$, are unbiased estimates of the optimal weights, $\hat{b}$, then the PGS, $\hat{A}_{SNP,i}$, will be an unbiased estimate of the true PGS (i.e., the additive genetic factor from SNPs), $A_{SNP,i}$.

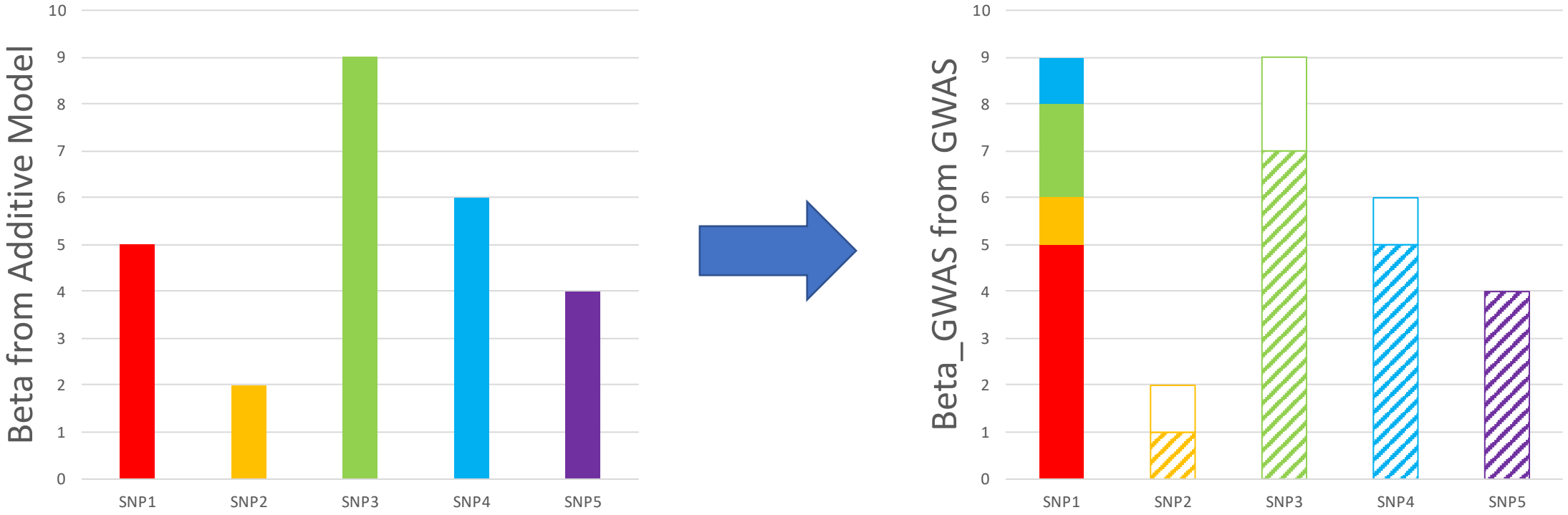
$$E\left(\hat{A}_{SNP,i}\middle|A_{SNP,i}\right) = A_{SNP,i}$$

- The PGS should probably **NOT** be considered a "genetic endowment"
  - Remember Jenck's critique

From Additive Genetic Factor to PGS



From Additive Genetic Factor to GWAS



$$\beta_{\mathrm{GWAS},j} = \sum \beta_k r_{jk}$$

Why is R^2 less than h^2_{SNP}?

- SNP-heritability is the explanatory power of the "true polygenic score"

$$A_{\mathrm{SNP},i} = \sum_j x_{ij} b_j$$

  - Since $A_{\mathrm{SNP},i}$ is the best linear predictor, the $A_{\mathrm{SNP},i}$ maximizes the R^2 over all other linear combinations of measured SNPs

- Because we only have imperfect estimates, $\hat{b}_j = b_j + e_j \neq b_j$, the predictive power of the (estimated) polygenic score must be less than the predictive power of $A_{\mathrm{SNP},i}$
  - We can calculate how much less using the Daetwyler formula

What happens to R^2 when you add more SNPs?

- Consider adding one addition SNP, x_{ij}, with effect size b_j
- Adds signal
  - Related to $\mathrm{Var}(x_{ij}b_j) = 2p_j(1-p_j)b_j^2$
  - Less than this if other SNPs already capture this signal
- Adds noise
  - Related to $\mathrm{Var}(x_{ij}) = 2p_j(1-p_j)$
- Holding p_j constant, SNPs with large b_j will tend to contribute more signal but will contribute a constant amount of noise
- For some traits, it may make sense to omit SNPs with large p-values/small b_j
  - In practice for behavioral traits, R^2 is maximized when you use all SNPs

Measurement Error Correction

- Two reasons why “units” of the regressions with a PGS are hard to interpret
  - Attenuation bias due to error in the PGS
  - Inflation of coefficients because PGS is usually standardized

$$Var\left(\hat{A}_{SNP,i}\right) = Var\left(A_{SNP,i}\right) + Var(U_i) > Var\left(A_{SNP,i}\right)$$

- Ideally, we’d want regressions that correspond to what we would have gotten if we had just used $A_{SNP,i}$ instead of $\hat{A}_{SNP,i}$
- Both of the obstacles to interpretation depend on $Var(U_i)$
  - Since we can estimate this value, we should be able to ”correct” our estimates

Two Sources of Stratification in PGSs

- Assume that the mean phenotype is higher in Population 1 then Population 2 for non-genetic reasons
- Prediction sample
  - Differences in allele frequencies across trait-associated SNPs may make one population have a different mean -> PGS captures non-genetic differences
- Discovery
  - SNPs that are more common in Population 1 will have effect estimates (and therefore PGS weights) that are biased upward
- Together
  - Effect estimates are biased upwards exactly where there are mean differences in allele frequencies
  - Compound such that mean differences in PGSs between groups imply a much larger difference in phenotype than actually is true

Bonferroni Correction vs M_e

- Genome-wide significance is 5×10^{-8}
  - Equivalent to if we performed 1M tests on uncorrelated markers
- M_e (which determines how much the predictive power of a PGS is attenuated) is 60,000
  - Equivalent to if we made a PGS with 60k uncorrelated genetic markers
- Why the difference?
  - Fuzzy explanation: Statistical testing is related to the total number of **SNPs** included in the analysis whereas prediction is more closely related to the number of **haplotypes** in the data.