Simulation – software, applications, and approaches

Luke Evans

CU Boulder & IBG

March 8, 2019

LOCATION: Faculty folders/luke/2019/Friday_simulation_practical

Copy that into your directory, cd into it.

Postdoctoral & Graduate Student Positions Available at University of Colorado Institute for Behavioral Genetics

- Complex Trait Statistical Genetics
- Nicotine use, mental health
- Contact me or Matt Keller
  - Luke.m.evans@Colorado.edu
  - Matthew.c.keller@gmail.com

Why would you want to make up your own data?

What do you think?

1. Idea 1
2. Idea 2
3. …

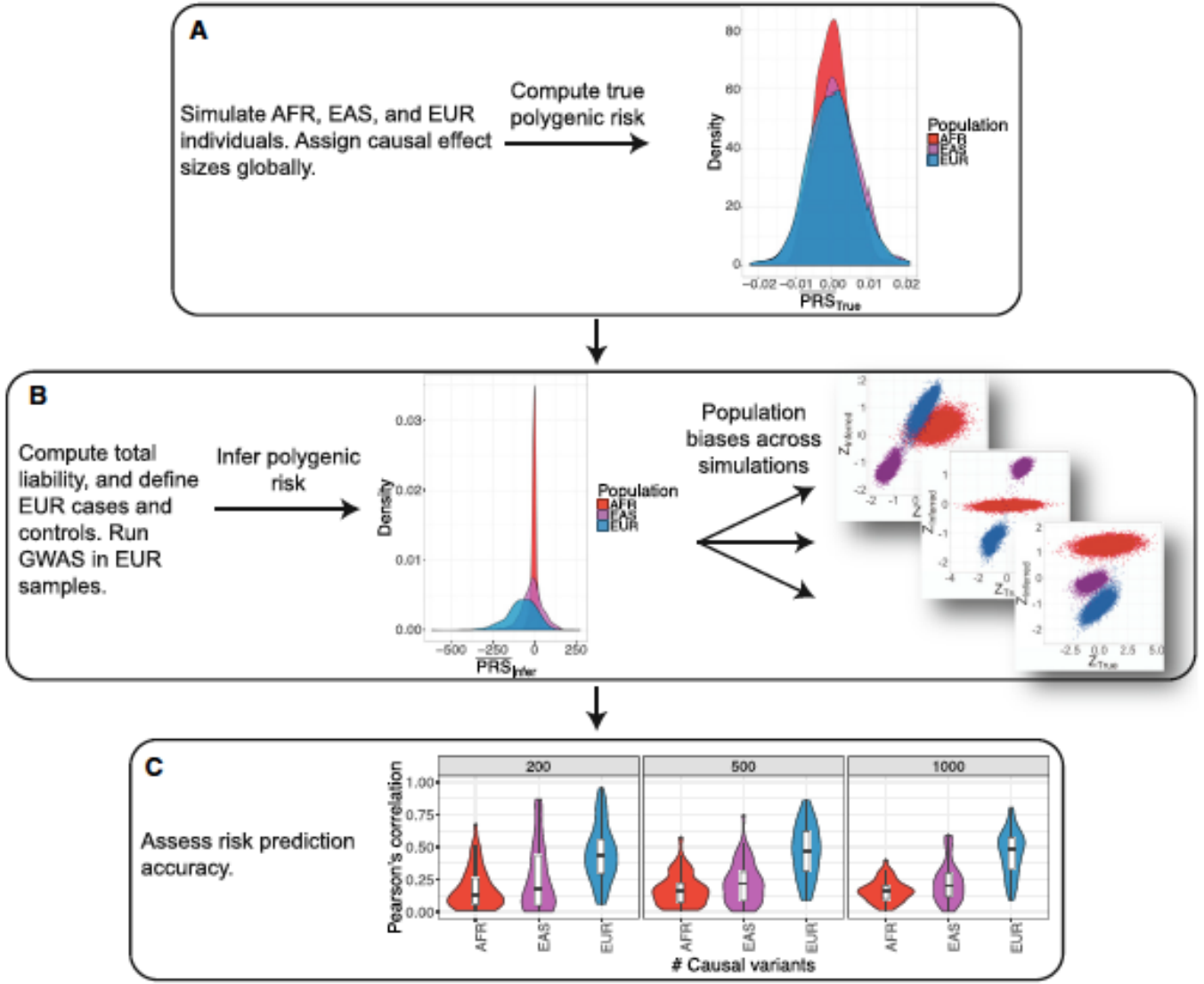
Why would you want to make up your own data?

Some reasons I've thought of:

1. Test new methods or compare across methods with the same data (equal footing to compare)
2. Identify the key assumptions for an analysis
3. Test how robust an analysis is to violations of those assumptions
4. Introduce new variables that you might be interested in
5. Get a sense of statistical power, or how variable something might be by chance alone.

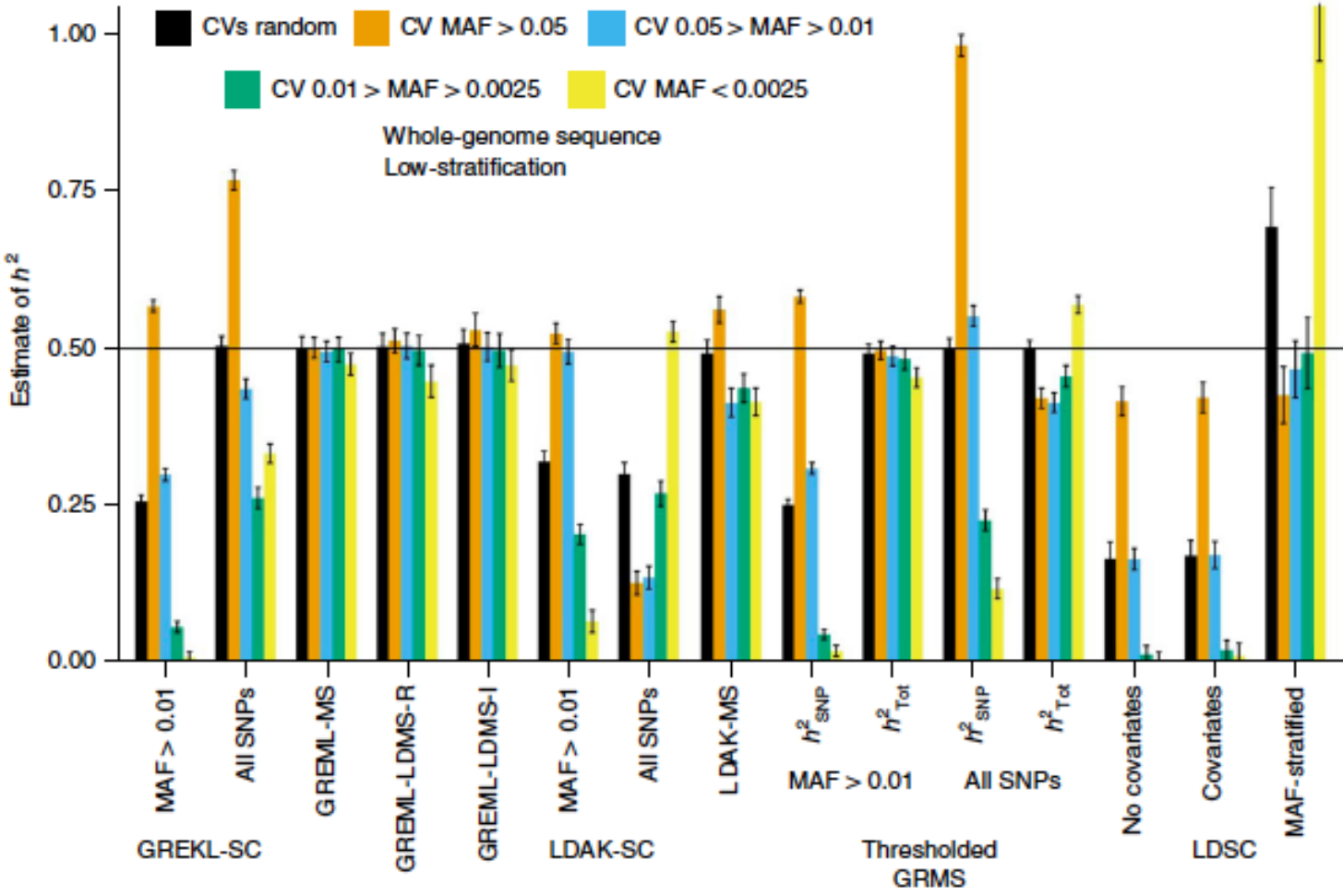
Example #1: How does human demographic history impact genetic risk prediction?

- Alicia Martin et al. 2017 *AJHG* 100:635-649
- Used coalescent simulations based on previously-estimated human demography to generate many replicates of human populations
- Created phenotypes, assessed associations in EUR population, tested PRS in a OTHER populations



Example #2: How do different h^2_{SNP} methods perform across varying complex trait genetic architectures?

- Evans et al. 2018 *Nat. Gen.* 50:737-745.
- Used whole genome sequence data to simulate phenotypes from different genetic architectures
- Methods to estimate h^2_{SNP} all have key assumptions, how robust are these approaches when those assumptions are not met?



How would you actually simulate data?

- Lots of different ways!
- R – rbinom function, for example, or more complex ways if you want linkage, relatives, demography, mutation, etc.
- Sequence simulators
  - Forward-time or coalescent
  - Hudson's ms, MSPRIME, GeneEvolve just a couple of examples (see papers in directory)
  - Scale to whole genomes & large sample sizes? Existing LD patterns or just simulate based on average recombination rates? Mutation rates? Demography? Long time frames or only several generations? Environmental influences? Selection?
  - Many things to consider!
- Just use real genotype data & assign effect sizes to particular variants
  - Maintain existing structure of the data – LD, polymorphism, etc.
  - What data are available? Samples & sample sizes? Biases, ascertainment, etc.?

1. Just make up some new genotypes & phenotypes in R

- Generate some genotypes & phenotypes in R with the rbinom function
- FILE: Genotype_simulation_1.R
  - Run through 1a, 1b, 1c (you'll have to do a few things yourself in the R code)
  - 1d is another example, simulating genotypes in different populations – optional if you want

1. Just make up some new genotypes & phenotypes in R

- Generate some genotypes & phenotypes in R with the rbinom function
- FILE: Genotype_simulation_1.R
  - Run through 1a, 1b, 1c (you'll have to do a few things yourself in the R code)
  - 1d is another example, simulating genotypes in different populations – optional if you want

- Simplest way to simulate genotypes!
- What sources of variation influence genotypes?
- What other things might you change or add that could influence your trait?

2. Real Data (or in this case, a simulated genotype dataset, but for our purposes, it's the same thing)

- Genotypes obtained from some real data set
  (here simulated from GeneEvolve).
- Phenotypes generated from some model (you decide! Your choice matters!), either in R yourself, or in PLINK, GCTA, etc.
- What kinds of data? What are the impacts of simulating from any of the following?
  - Sequence data: 1KGv3, TOPMed, HRC
  - Imputed data: UK Biobank, maybe your own?
  - Array data: Lots! Just look through dbGaP.

2. Real Data (or in this case, a simulated genotype dataset, but for our purposes, it's the same thing)

- Options to simulate phenotypes from existing genotypes
  - GCTA, Plink, others
- Can specify
  - Causal variants – how do you choose?
  - Trait heritability – what impact might this have?
  - Allelic effect sizes – what distribution might you draw effect sizes from?
  - Case control - specify prev. & ascertainment

2. Real Data (or in this case, a simulated genotype dataset, but for our purposes, it's the same thing)

Practical #2 – generate some phenotypes from randomly chosen causal variants of different MAF ranges

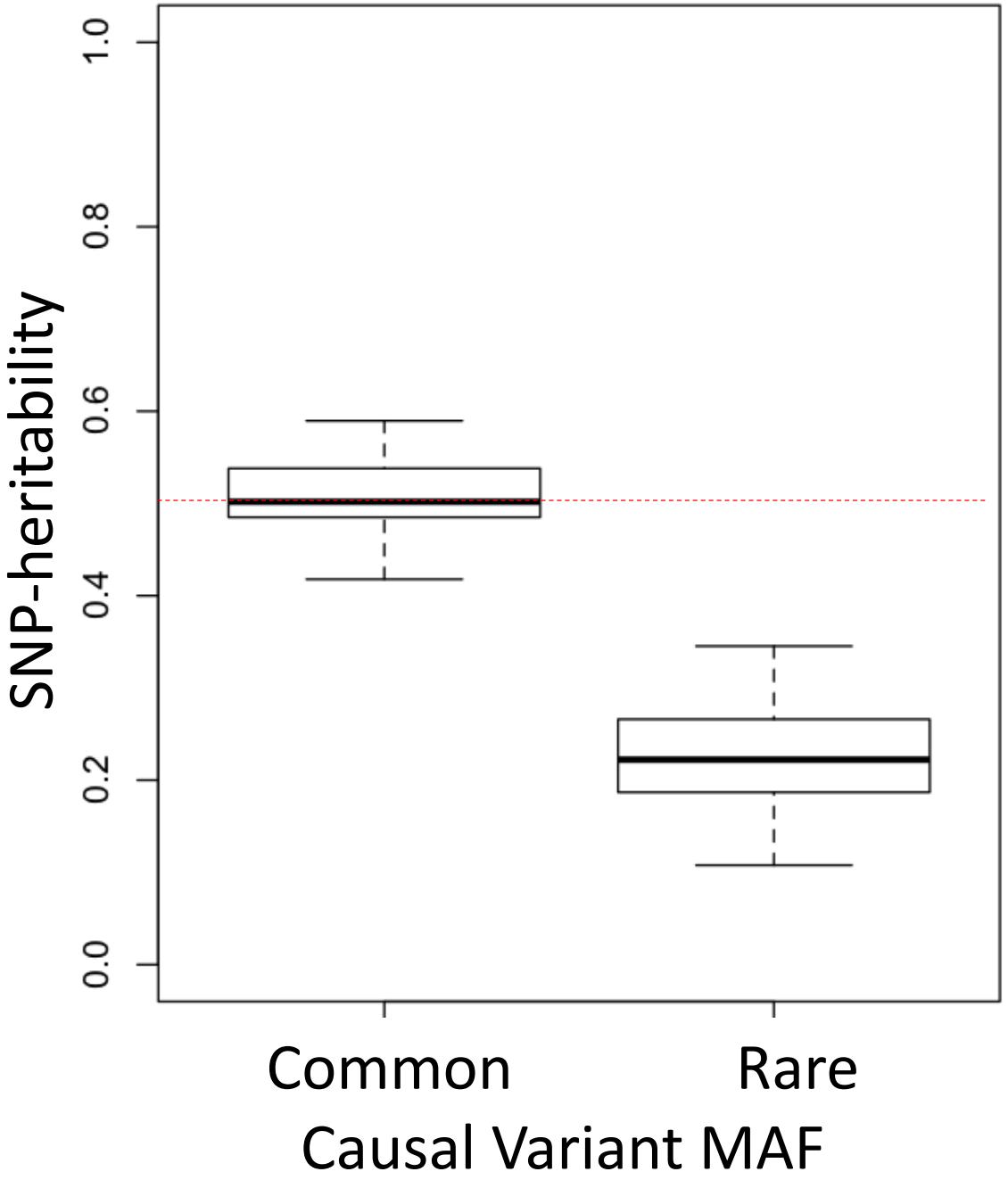
- Step 1: Randomly choose some CVs based on some criteria (here, MAF)
- Step 2: Use that list of CVs to create phenotypes with GCTA from a real plink binary file set.
- Step 3: Do some test on the simulated phenotypes, to see what the influence of your assumptions are.
  - Here, we're going to test the influence of CV MAF on your estimates of GCTA-based heritability – You've already done something like this during Loic's session, so it should be familiar to you.
  - The single GRM is from common variants, kind of the basic h2SNP model (previously computed, in the simulation_2/ directory)

- **FILE:** simulation_2.bash
- **NOTE:** In the terminal run command "bash" first, before you copy & paste any of the commands, so that you're running in bash rather than another shell

Simulation 2: Varying MAF of causal variants:

Why does it the analysis underestimate the true trait heritability when the causal variants are rare (or at least rarer than the markers)?

- See Loic's talk & practical



What are your assumptions? What are their impacts? Vary them in simulations & find out.

- What are you assuming in your analysis?

What are your assumptions? What are their impacts? Vary them in simulations & find out.

- What are you assuming in your analysis?
  - MAF-effect size relationship or none?
  - LD-effect size relationship or none?
  - Panmictic population or stratification effects?
  - Polygenicity?
  - Shared environments or independent environments for all individuals in the sample?
  - Combinations of these and others?
  - There are lots of assumptions for even a simple GWAS…

What are your assumptions? What are their impacts? Vary them in simulations & find out.

- What are you assuming in your analysis?
  - MAF-effect size relationship or none?
  - LD-effect size relationship or none?
  - Panmictic population or stratification effects?
  - Polygenicity?
  - Shared environments or independent environments for all individuals in the sample?
  - Combinations of these and others?
  - There are lots of assumptions for even a simple GWAS…
- Alter these relationships, then test what that will do to the phenotype, results, and your conclusions.

What are your assumptions? What are their impacts? Vary them in simulations & find out.

- What are you assuming in your analysis?
  - MAF-effect size relationship or none?
  - LD-effect size relationship or none?
  - Panmictic population or stratification effects?
  - Polygenicity?
  - Shared environments or independent environments for all individuals in the sample?
  - Combinations of these and others?
  - There are lots of assumptions for even a simple GWAS…
- Alter these relationships, then test what that will do to the phenotype, results, and your conclusions.
- What kinds of data? What are the impacts of simulating from any of the following?
  - Sequence data: Biases from sequencing itself?
  - Imputed data: What variants are imputed? Imputation quality?
  - Array data: Why are particular SNPs on genotyping chips?

3. Vary assumptions and test your models

- Practical #3 – generate some genotypes & phenotypes, varying the assumptions (the generating model), and see how it impacts your conclusions
- FILES:

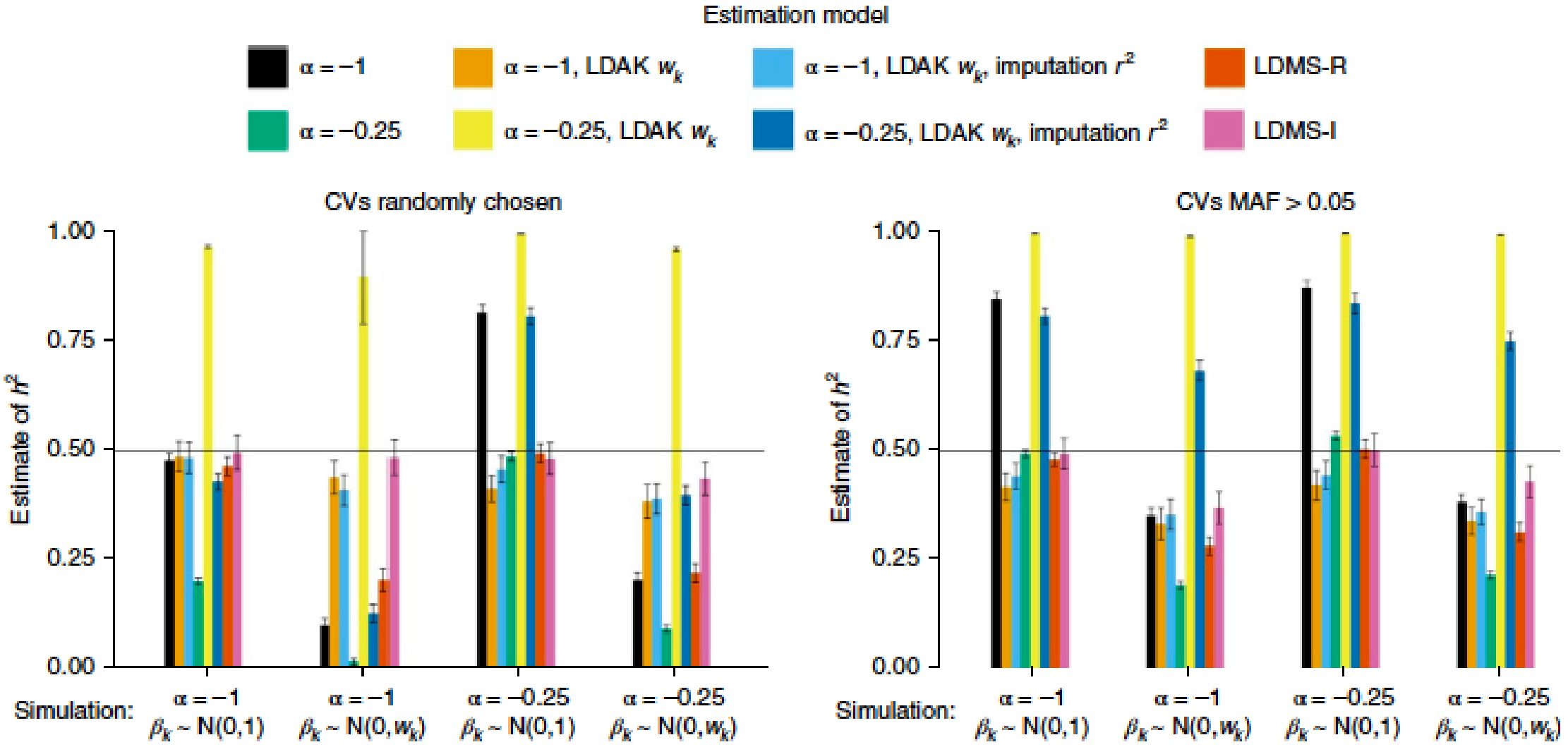
  Step 1 - FILE simulation_3.R – generate effect sizes with different assumptions, specifically making betas related to MAF.

  Step 2 - FILE simulation_3.bash - Bash script to run GCTA to first make the phenotypes with your MAF-scaled betas, then estimate heritability (think back to Loic's talk – what is the impact of MAF & LD on estimates?).

  Step 3 - FILE simulation_3.R – at the end of the script, visualize how all of these different simulations (all with the same TRUE h2SNP (=0.5) impact the ESTIMATED h2SNP

  Extra exercises & things to think about at the end of the R script.

Varying model assumptions and simulation parameters – test how robust your estimates may be.



Simulation 3: Varying MAF & beta distributions

Impacts of model assumptions?
Need more replicates?

Note: The “rare” category is still pretty common. Bigger impacts of the MAF-scaling happen when variants are really rare.