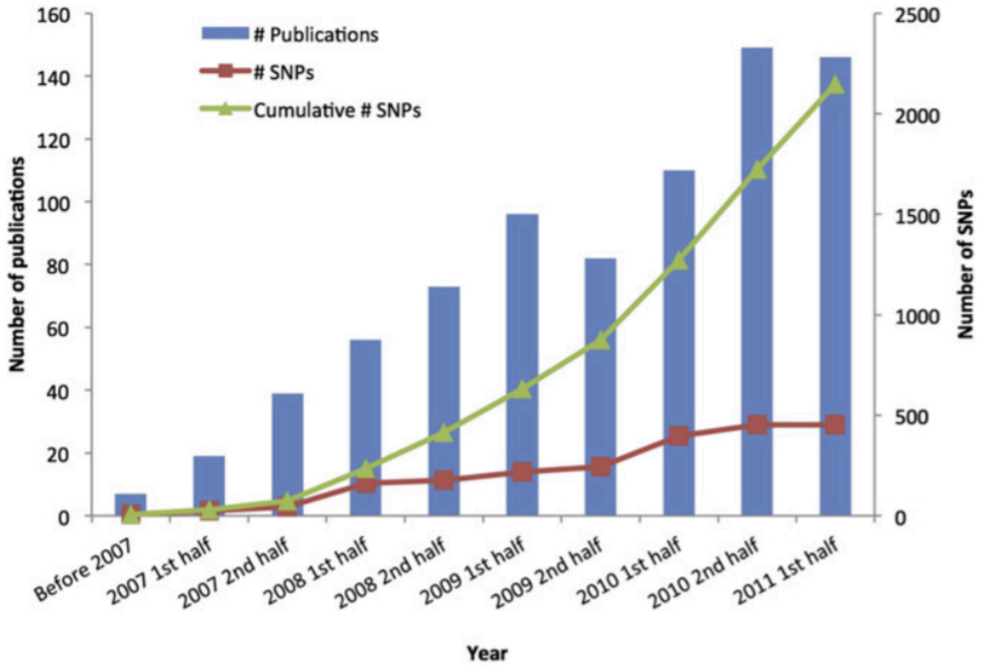
Machine Learning for large-scale genomics

Sriram Sankararaman

Computer Science, Human Genetics, Computational Medicine
UCLA

Genetic architecture of complex phenotypes

$$Phenotype = f(Genotype, Environment)$$



Visscher et al. AJHG 2012

Growth of Biobanks

Machine Learning for Biobank-scale data

How can we learn about genetic architecture of complex traits and diseases from datasets that contain millions of genomes and thousands of traits ?

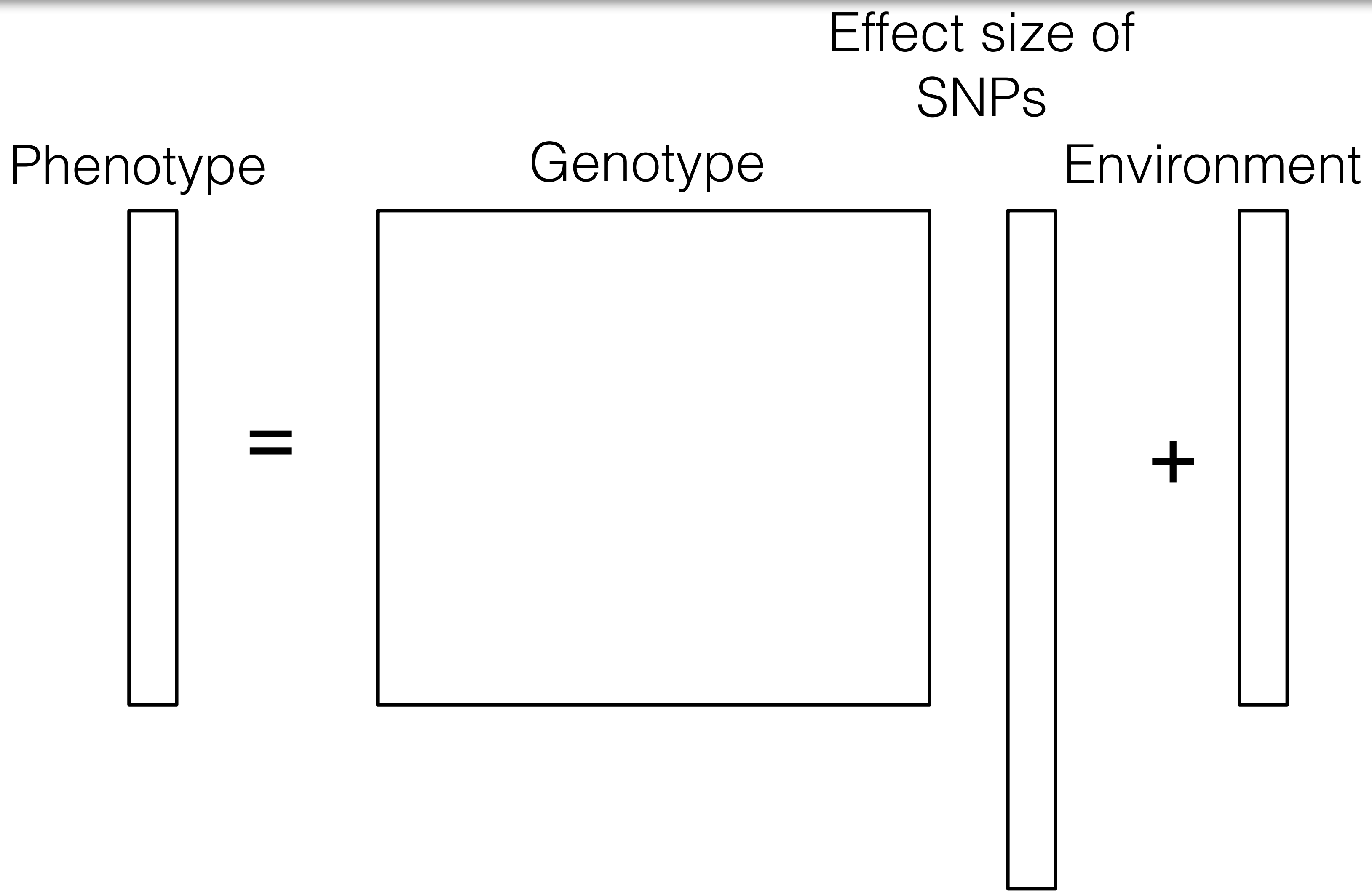
Statistical

Privacy

Computational

Interpretability

(Narrow-sense) Heritability



$$Y = X \beta + \epsilon$$

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

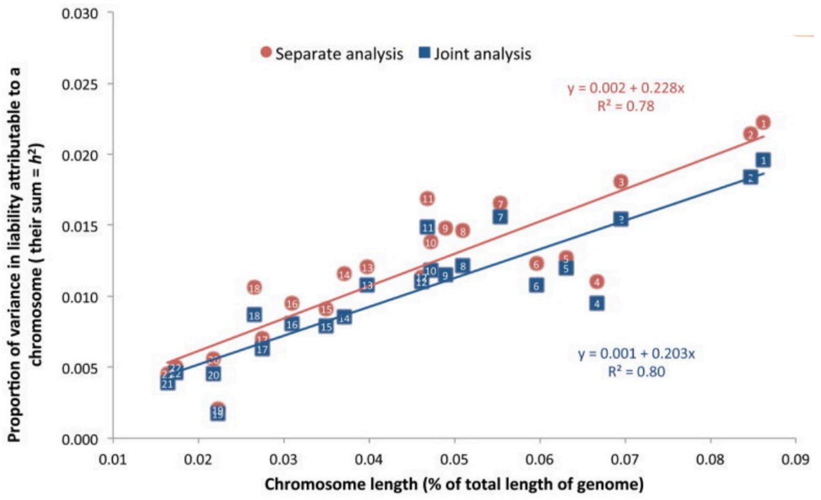
SNP heritability

$$\beta_m \sim \mathcal{N}(0, \frac{\sigma_g^2}{M}) \quad \epsilon_n \sim \mathcal{N}(0, \sigma_e^2)$$

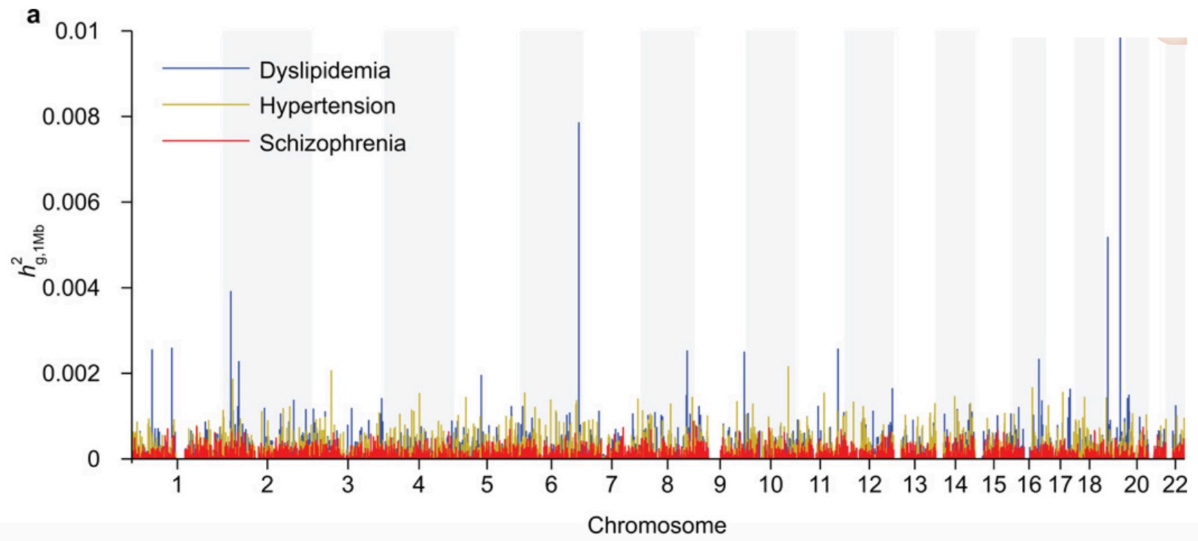
Environmental variance component

Genetic variance component

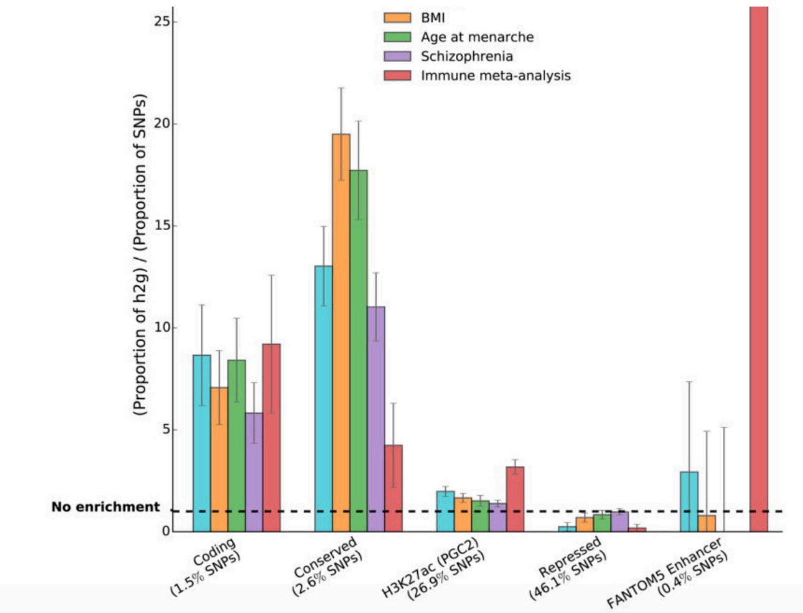
Beyond Heritability



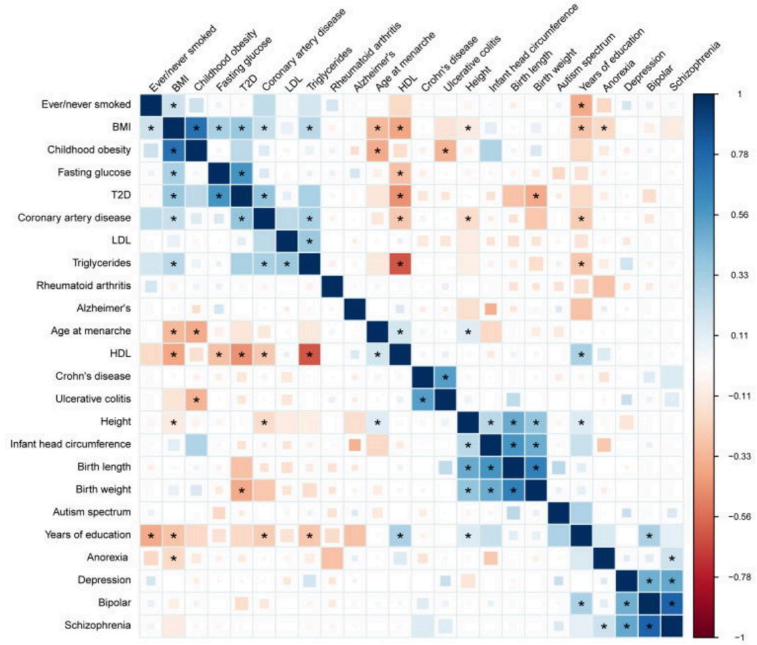
Lee et al. 2012



Loh et al. 2015



Finucane et al. 2015



Bulk-Sullivan et al. 2015

UCLA

Sriram Sankararaman
Machine Learning and Genomics Lab

Variance components model

$$\boldsymbol{y} = \sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon}$$

$$\boldsymbol{\beta}_k \sim \mathcal{N}(\mathbf{0}, \frac{\sigma_k^2}{M_k} \boldsymbol{I}_{M_k}), k \in \{1, \ldots, K\}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_N)$$

Goal

Estimate variance components $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2, \sigma_e^2)$

Estimating variance components

Maximum likelihood

$$\begin{aligned}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_K^2, \hat{\sigma}_e^2) &= argmax_{(\sigma_1^2,\ldots,\sigma_K^2,\sigma_e^2)} \mathcal{L}\mathcal{L}(\sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2) \\ &= argmax_{(\sigma_1^2,\ldots,\sigma_K^2,\sigma_e^2)} P(\boldsymbol{y}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K, \sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2)\end{aligned}$$

Computationally expensive

Scales as $\mathcal{O}(N^3)$

Challenging to apply to Biobank-scale data

Lippert et al. Nature Methods 2012
Zhou and Stephens, Nature Genetics 2012
Loh et al. Nature Genetics 2015

Randomized HE-regression (RHE-mc)

Combines randomization with a method-of-moments estimator

Work with a "sketch" of the genotype

Multiply the genotype matrix with B random vectors

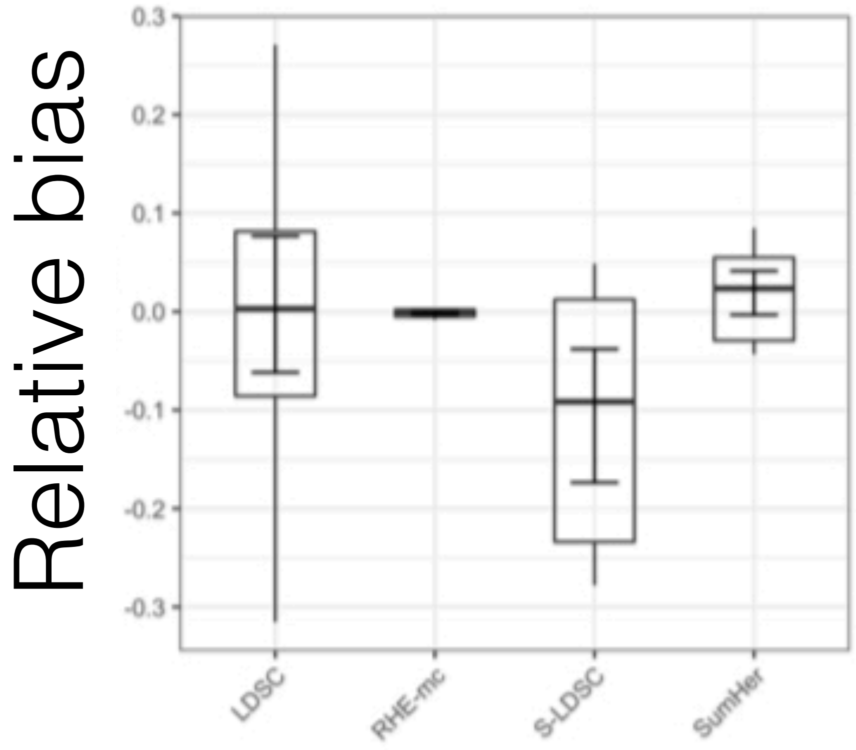
- Efficiency depends on B: $\mathcal{O}(\frac{MNB}{\log_3(\max(N,M))})$
- Accurate for B as small as 10

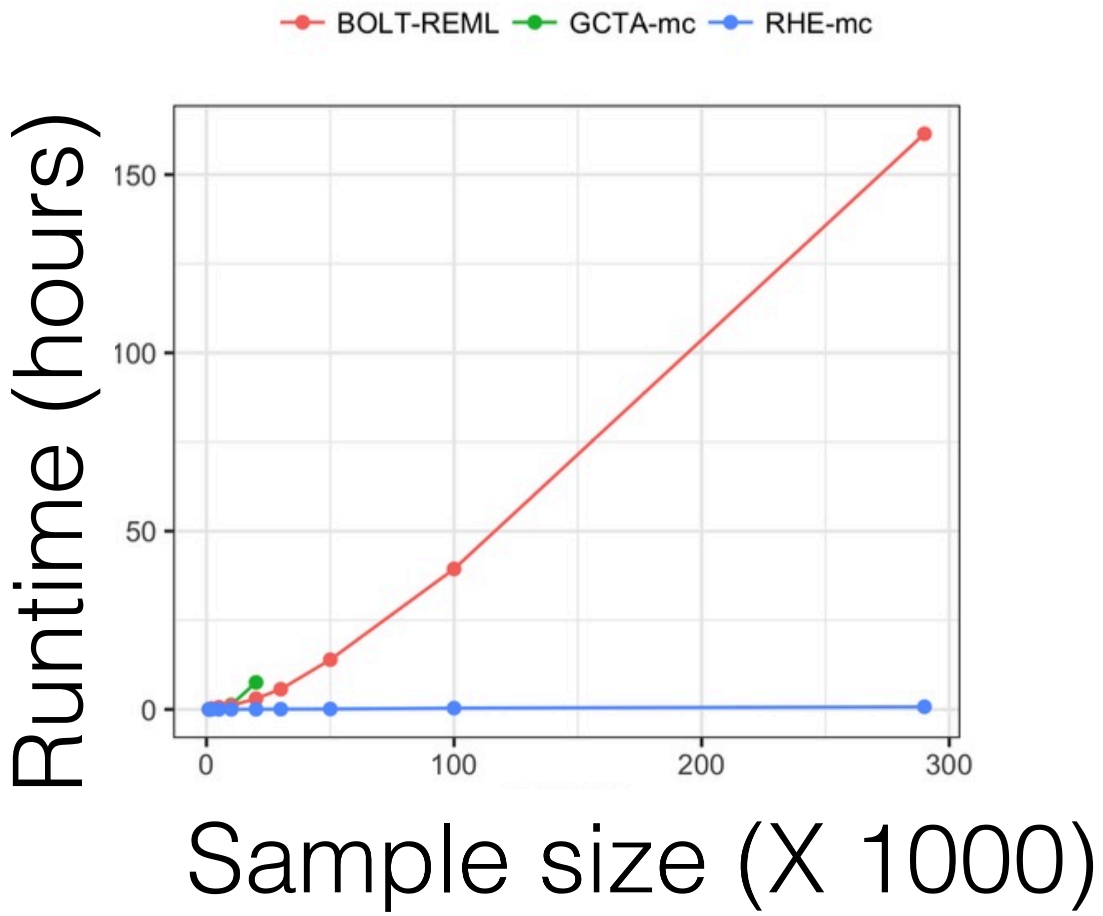
Hutchinson 1989

Wu et al. Bioinformatics 2018

Pazokitoroudi et al. RECOMB 2019, Nature Communication 2020

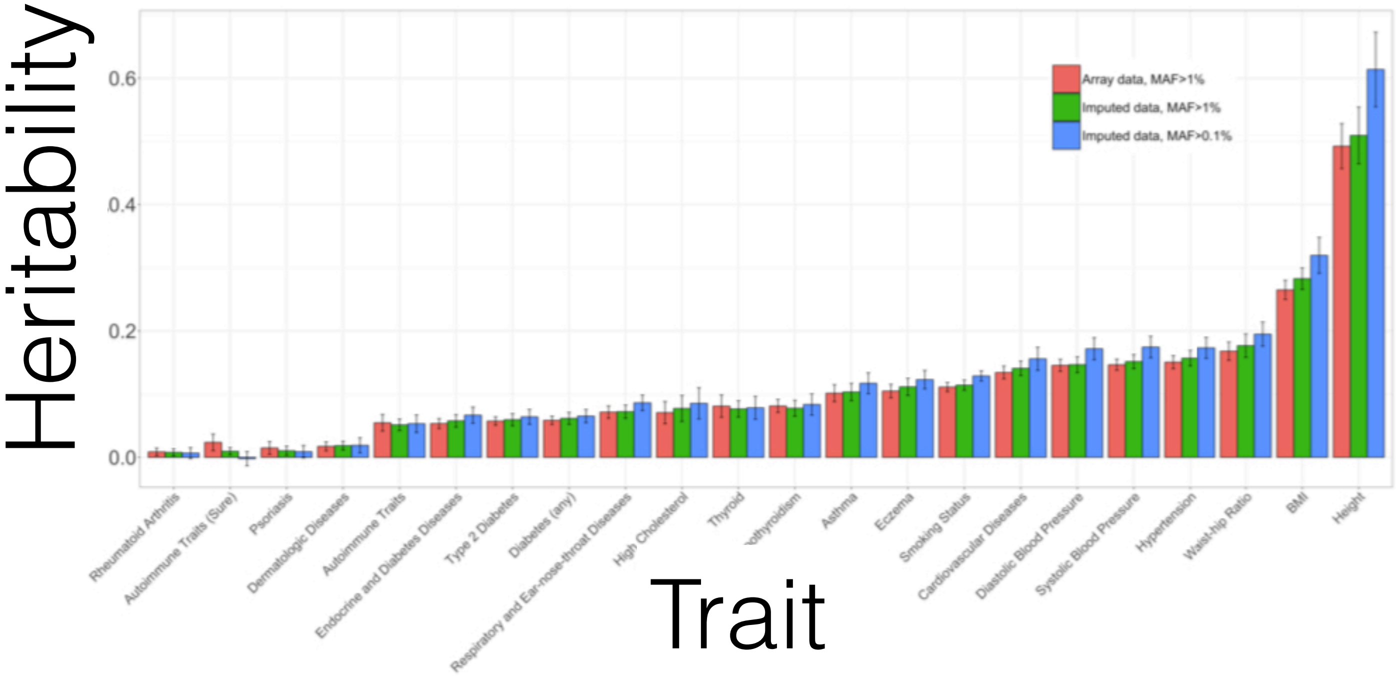
Comparisons of RHE-mc

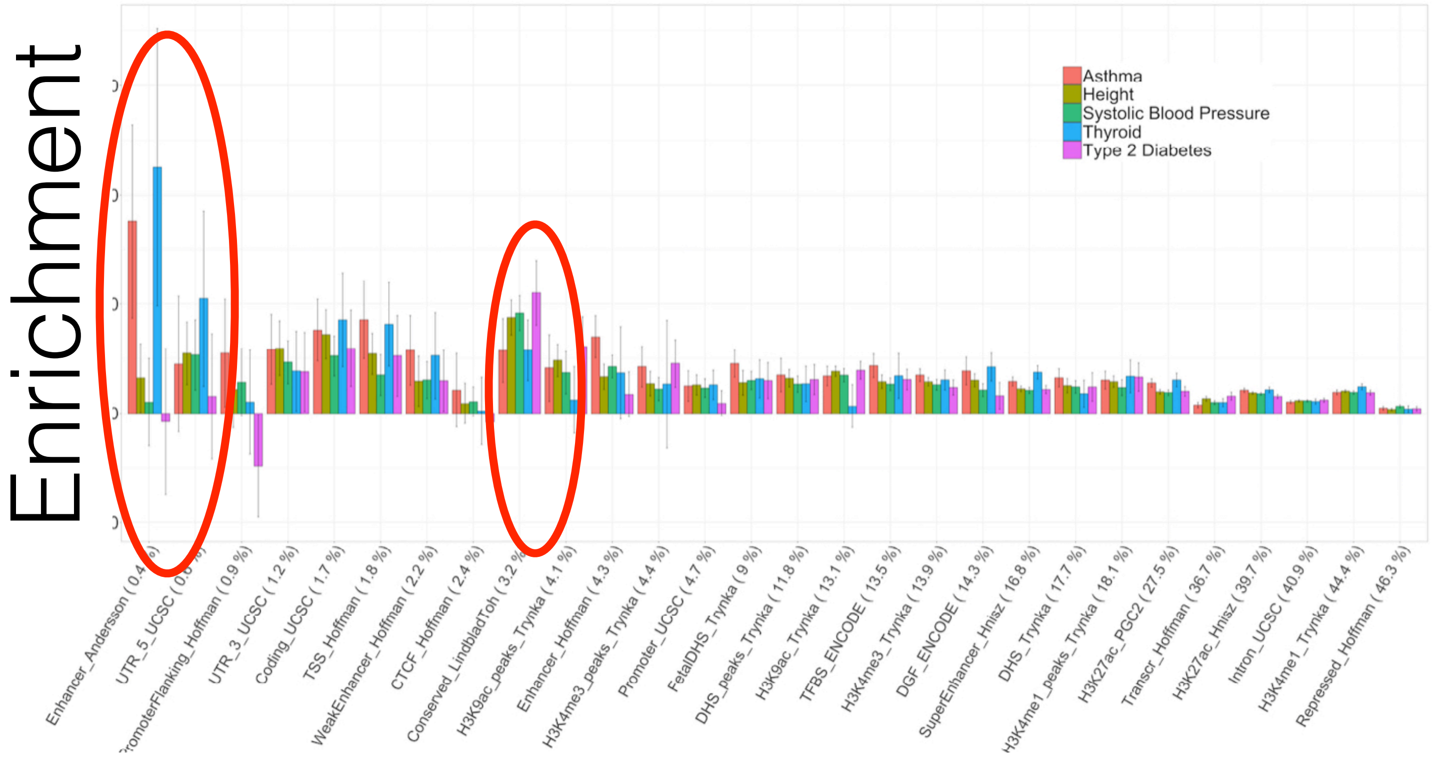


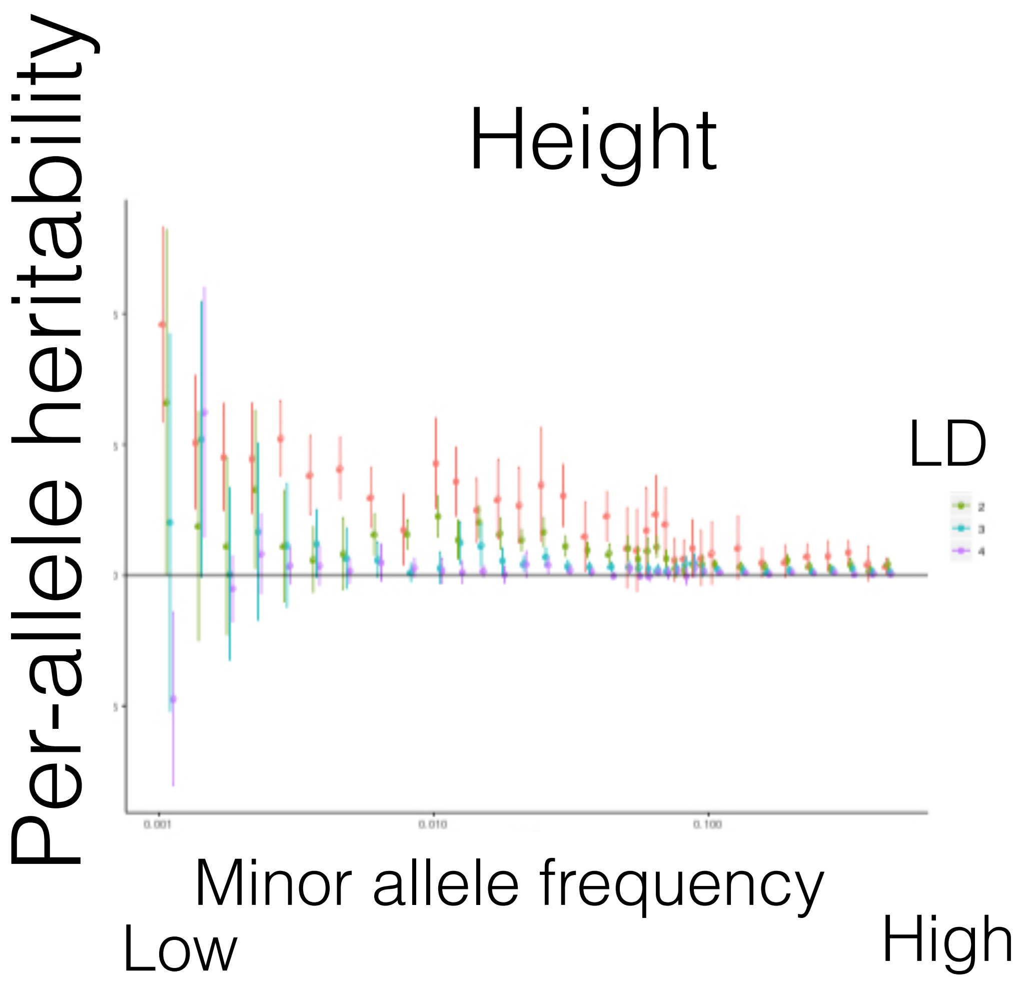


Pazokitoroudi et al. RECOMB 2019, Nature Communication 2020

Insights from Biobank-scale analysis







Pazokitoroudi et al. Nature Communication 2020

Beyond heritability

What is the contribution of non-linear effects ?

What is the contribution of environmental interactions ?

How are genetic effects shared across traits ?

Dominance deviation effects

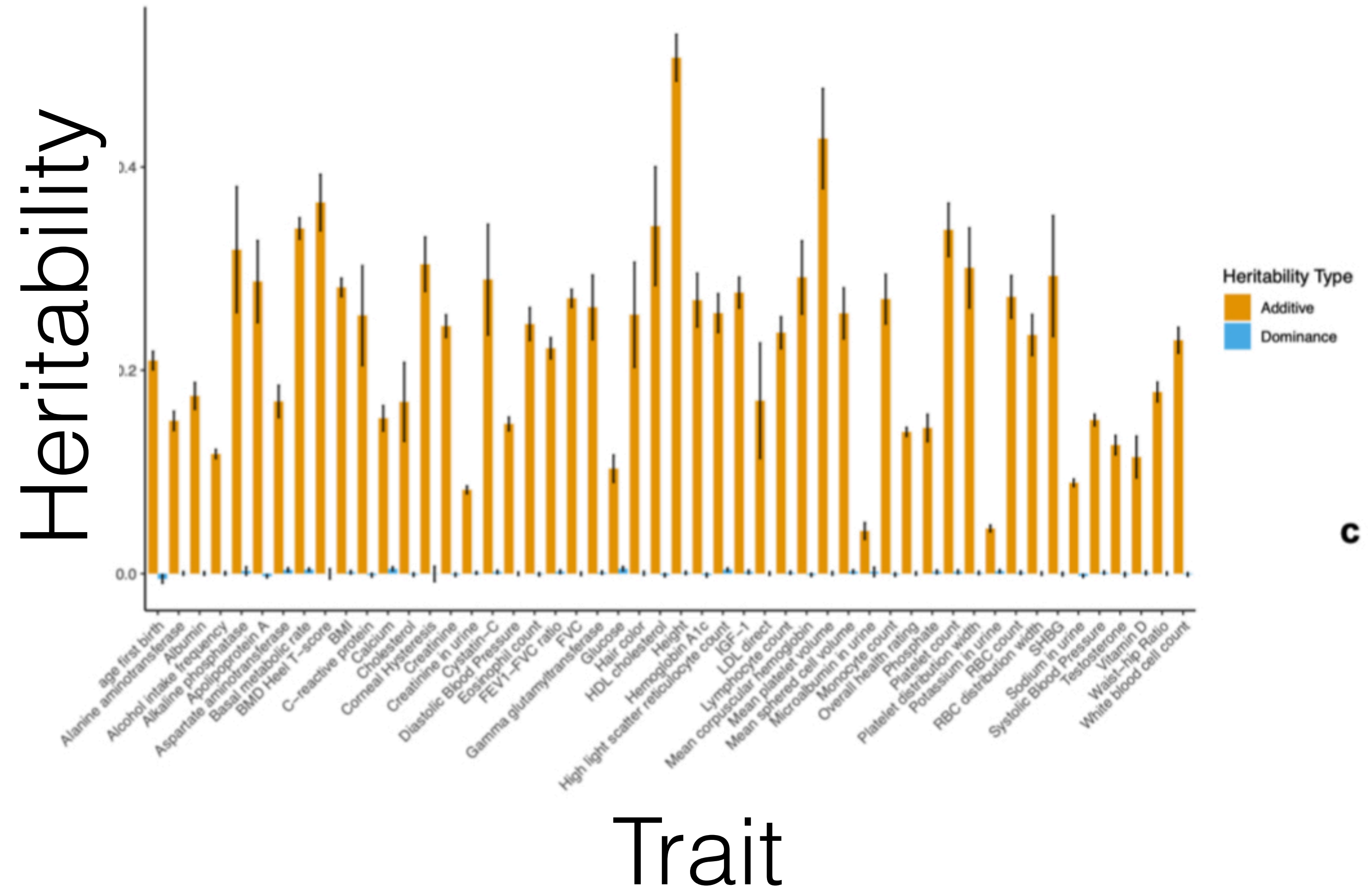
Additive variance component

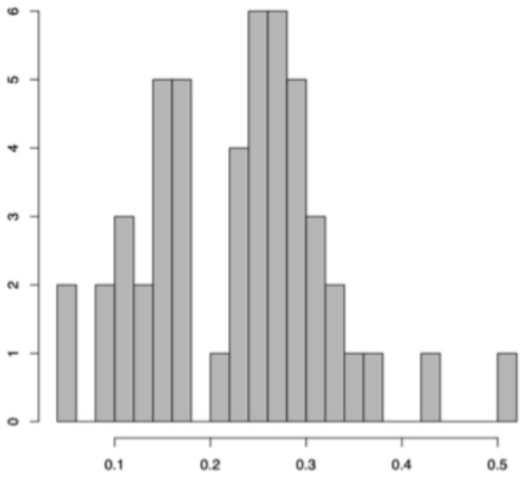
Dominance variance component

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

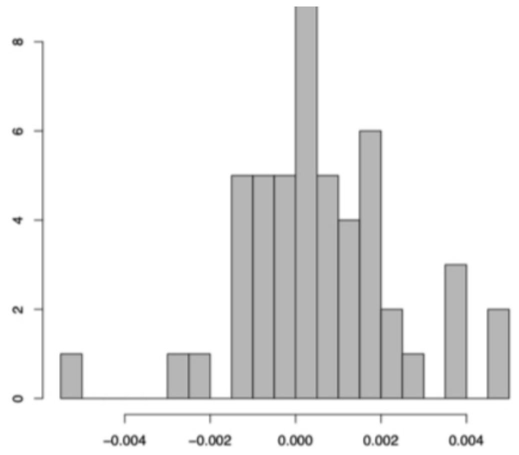
$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \frac{\sigma_a^2}{M}\boldsymbol{I}_M)$$

$$\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{0}, \frac{\sigma_d^2}{M}\boldsymbol{I}_M)$$





Additive heritability



Dominance heritability

Hivert et al. AJHG (to appear)

Pazokitoroudi et al. AJHG (to appear)

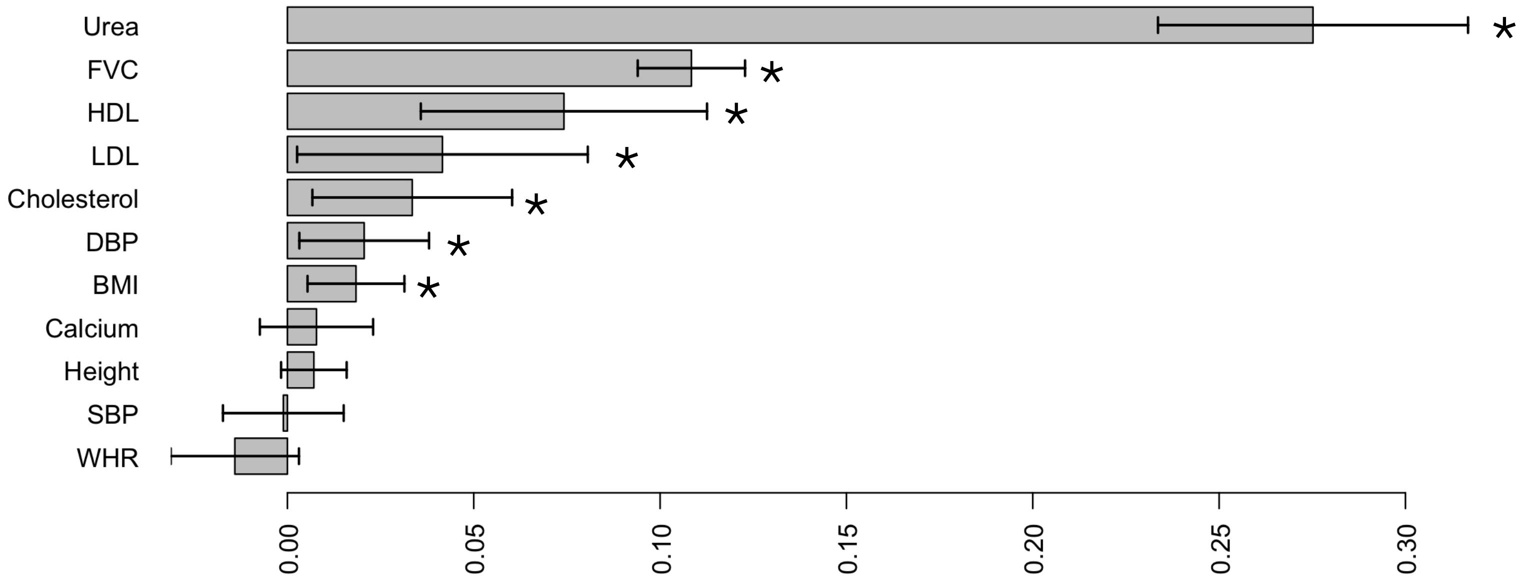
Gene-environment interactions (GxE)

$$y = X\beta + X \odot E\delta + \epsilon$$

$$\delta \sim \mathcal{N}(0, \frac{\sigma^2_{GE}}{ML} I)$$

GxE variance component

E = Smoking

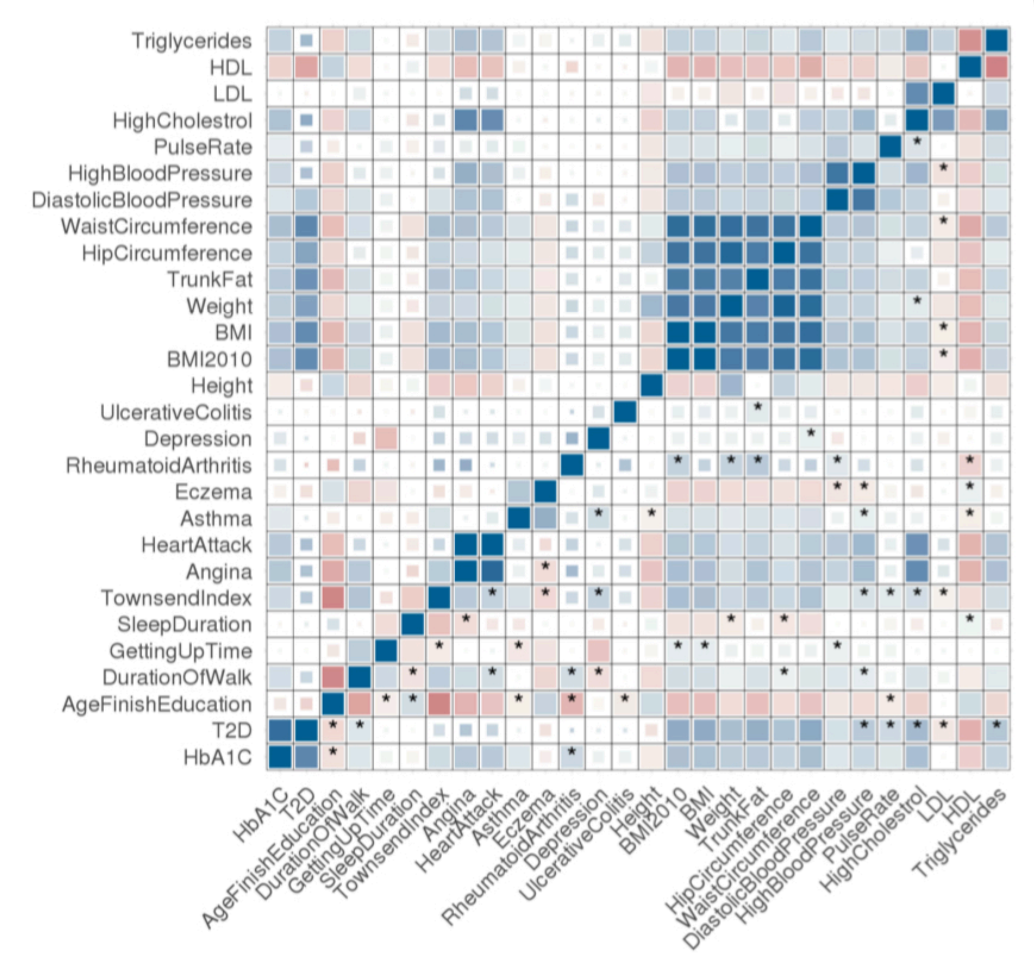


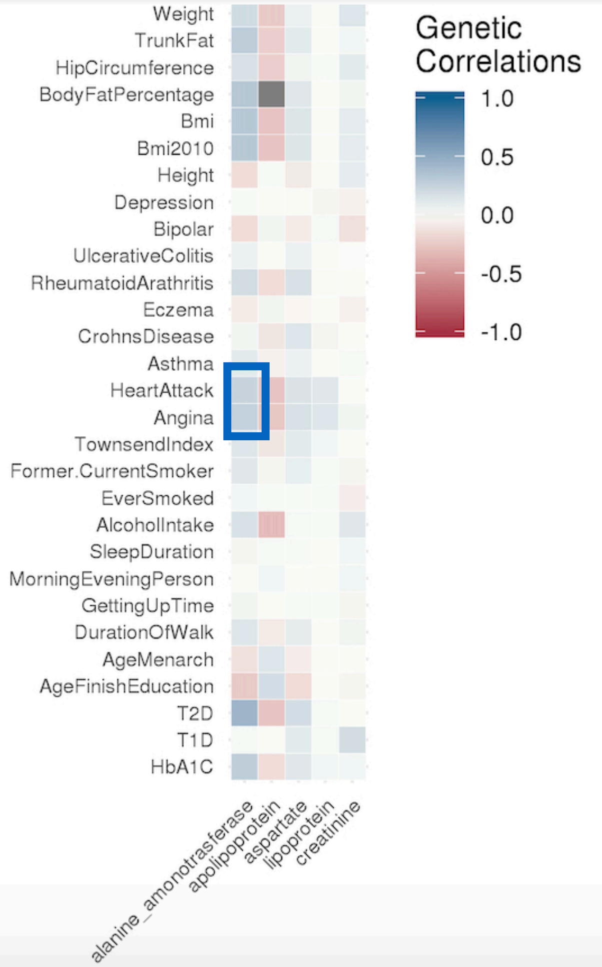
$$\frac{GxE \quad heritability}{Additive \quad heritability}$$

Kerrin and Marchini AJHG 2020

Pazokitoroudi et al. RECOMB 2021

Genetic correlation





Novel genetic correlation of coronary artery disease and serum liver enzyme

Wu et al. RECOMB 2019, BioRXiv 2020

Scaling Machine Learning to Biobanks

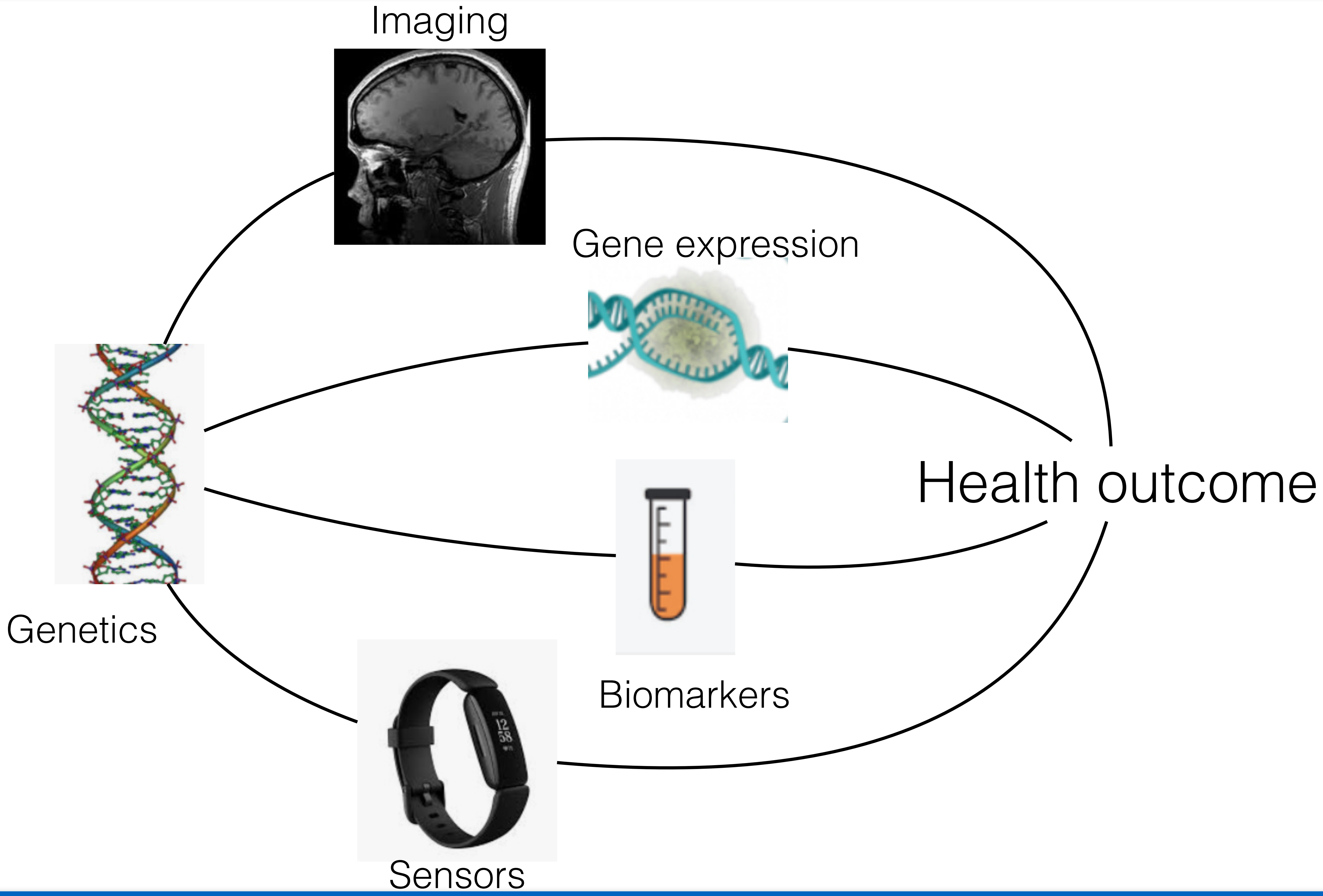
Effectiveness of randomization

Approximate inference

Distributed inference

Promises and challenges

Multi-modal data



Imaging

Gene expression

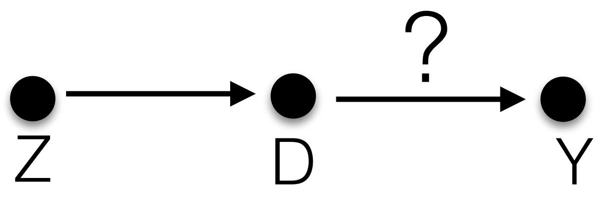
Genetics

Biomarkers

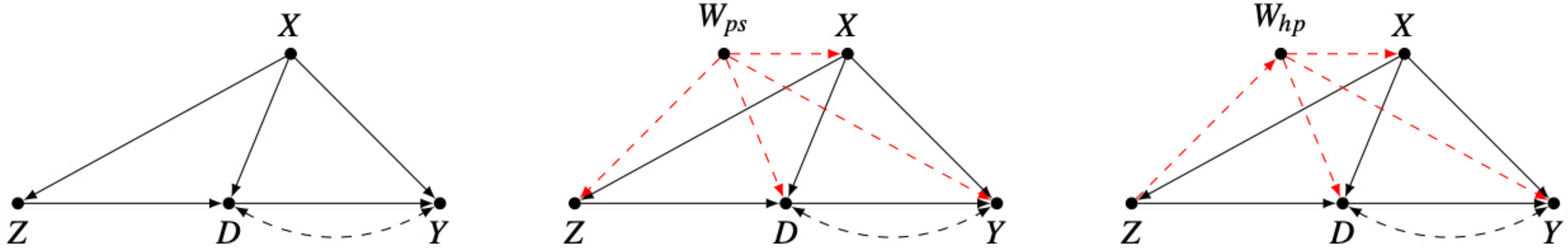
Health outcome

Sensors

Causal inference



Genetic variant Exposure Outcome

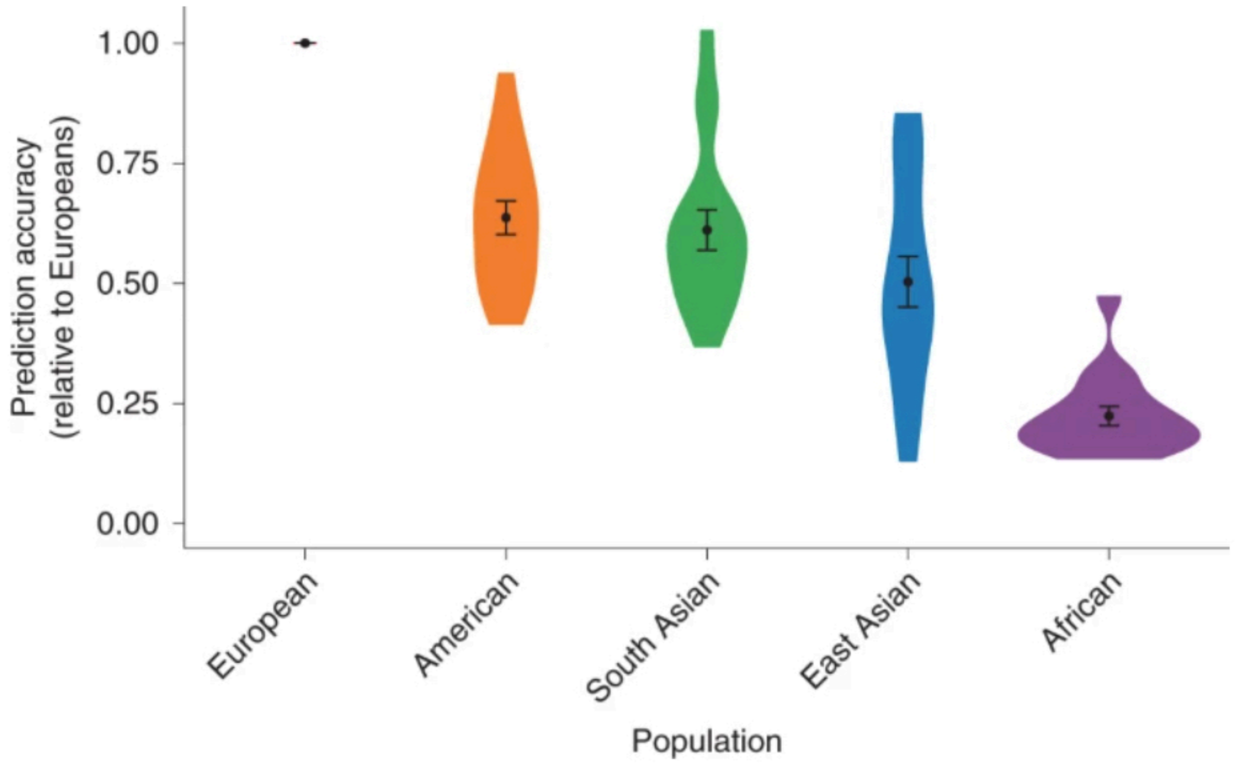


Katan et al. Lancet 1986

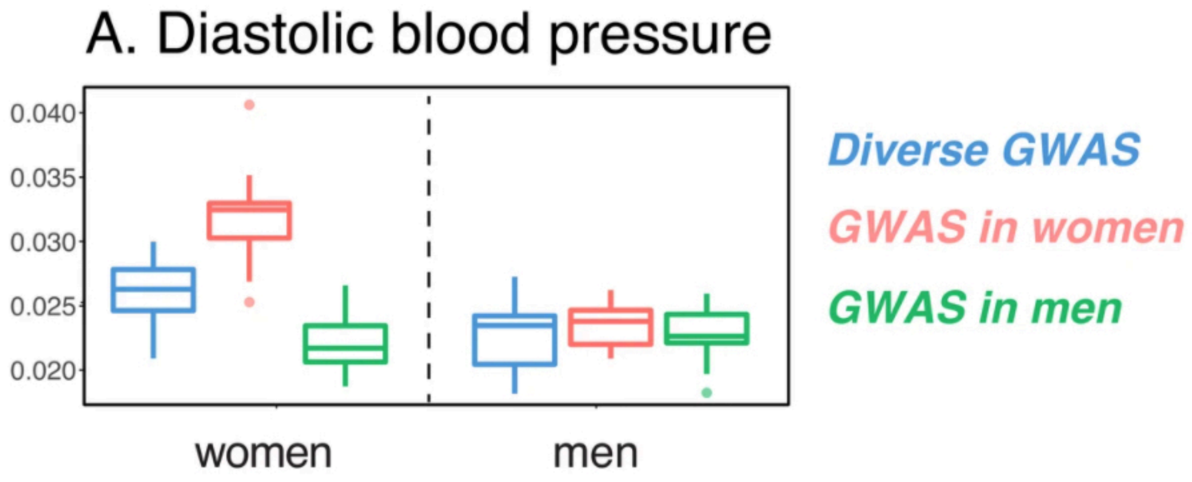
Brown et al. BioRxiv 2020

Cinelli et al. BioRxiv 2020

Generalizability

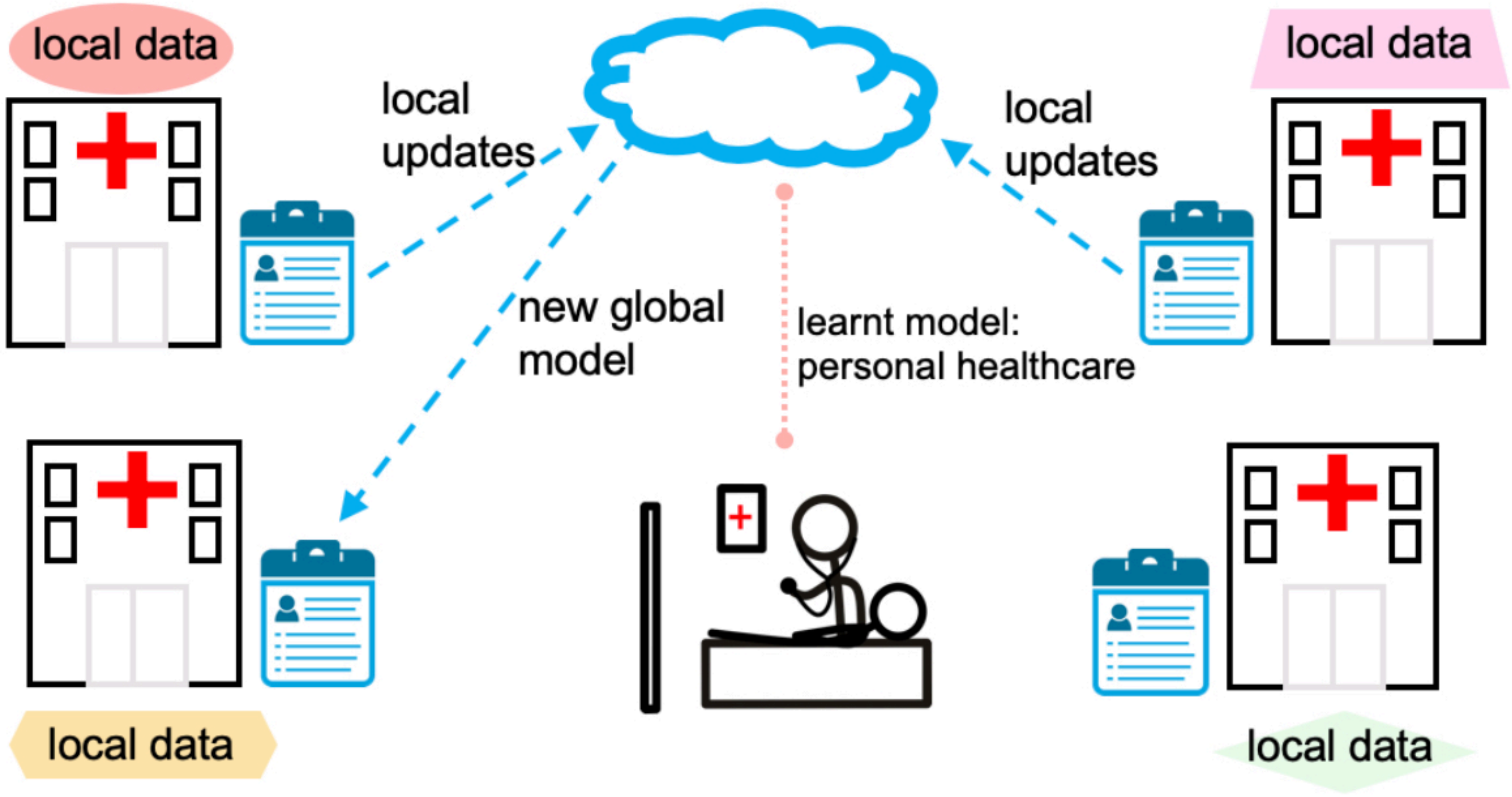


Martin et al. Nature Genetics 2019



Mostafavi et al. eLife 2020

Distributed and federated ML



https://blog.ml.cmu.edu/2019/11/12/federated-learning-challenges-methods-and-future-directions/

Acknowledgments

Funding

Microsoft Research

Alfred P. Sloan Foundation

Hellman Fellows Fund

NSF

NIH National Institute of General Medical Sciences

The Okawa Foundation for Information and Telecommunications