Probabilistic models for large-scale human genomic data

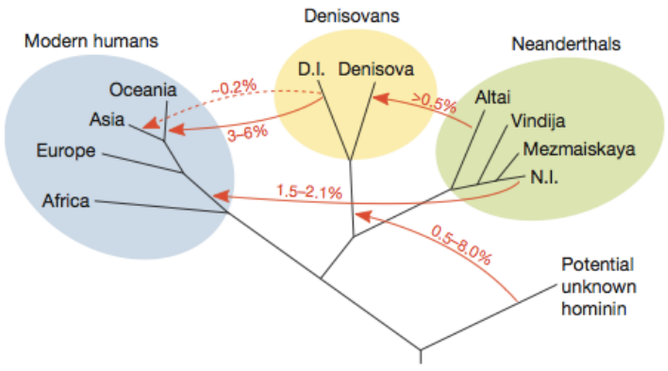
Sriram Sankararaman

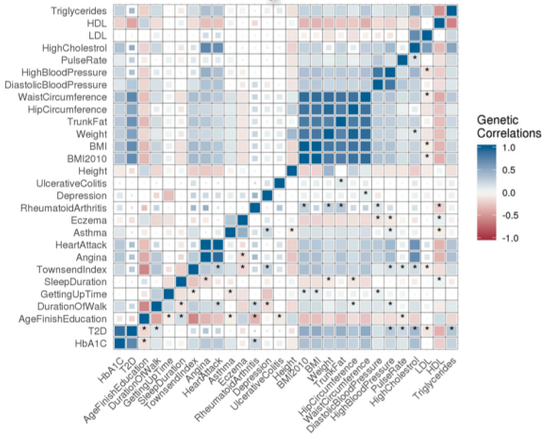
Computer Science, Human Genetics, Computational Medicine
UCLA

Machine learning for genomic data

Evolution

Clinical



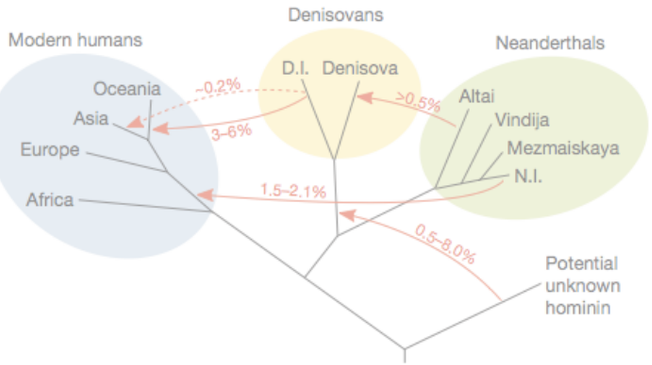


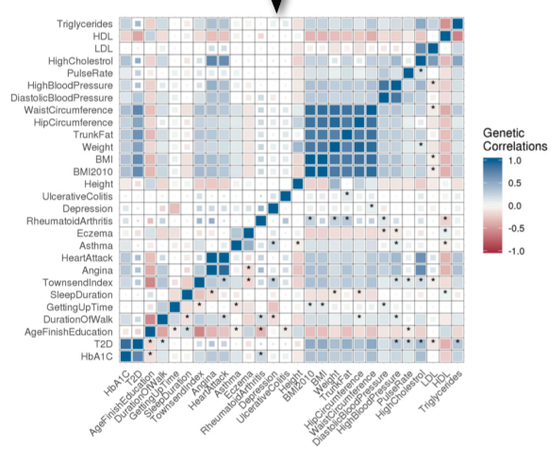
Complex traits

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

Machine learning for genomic data

Evolution





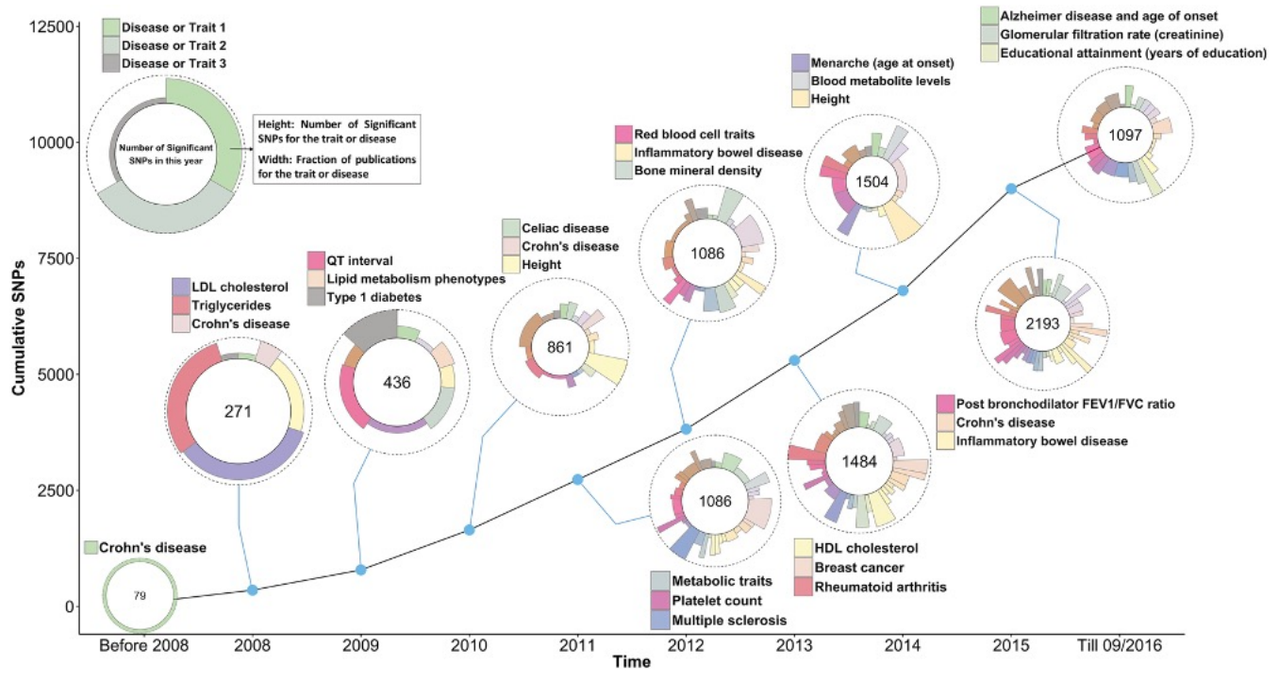
Clinical

Complex traits

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

Genetic architecture of complex phenotypes

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



Visscher et al. AJHG 2017

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

Growth of Biobanks

biobank uk

FINNGEN

MVP

BioMe BIOBANK

BioVU

BIOBANK JAPAN

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

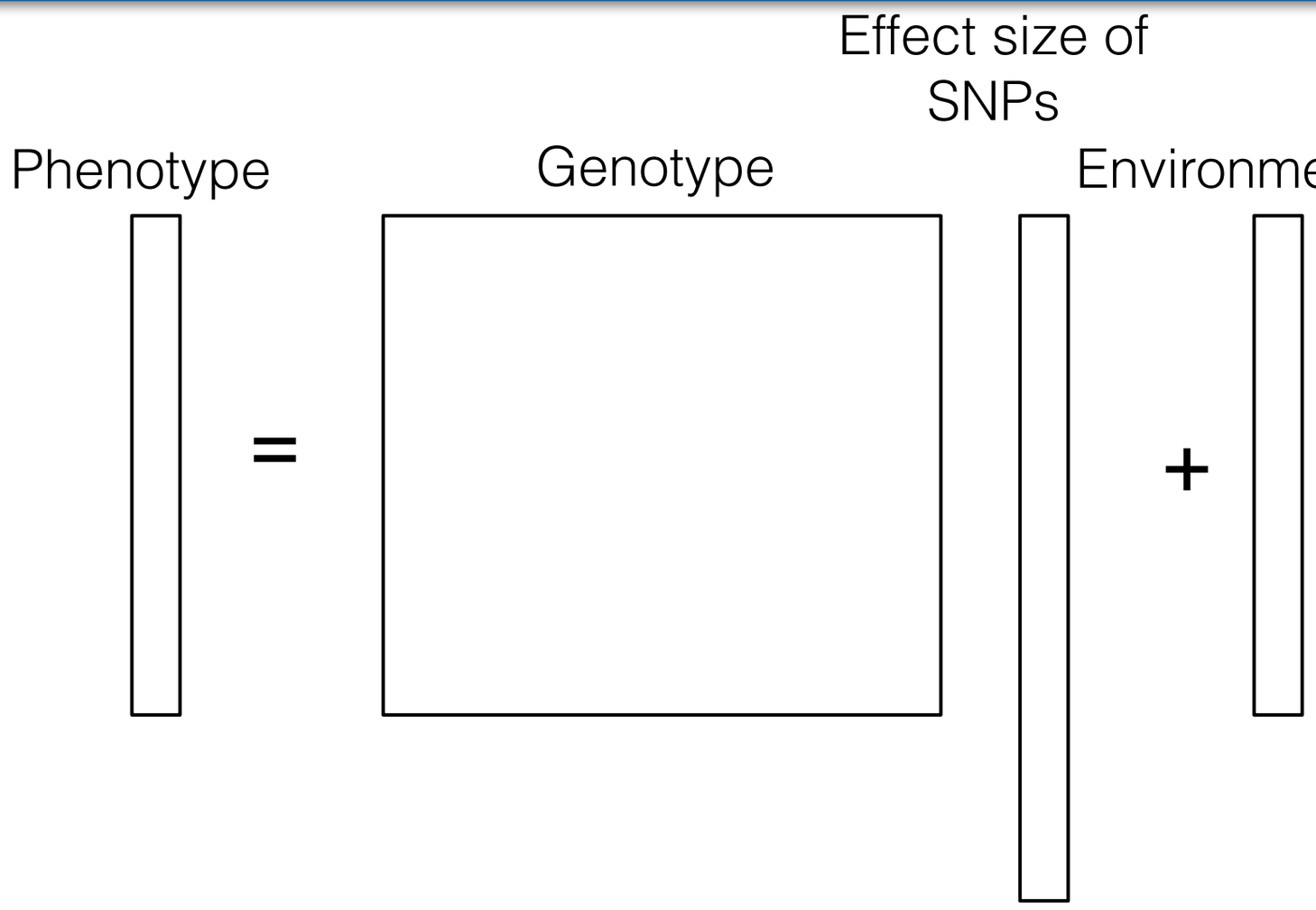
Privacy

Computational

Interpretability

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

(Narrow-sense) Heritability



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

SNP heritability

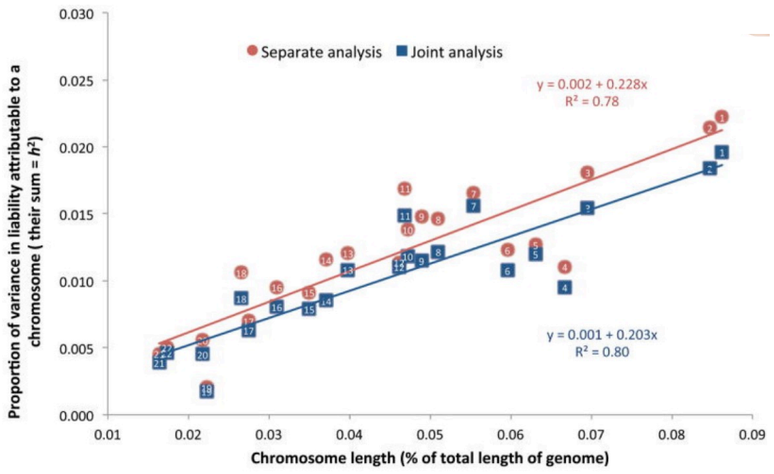
$$Y = X \beta + \epsilon$$

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

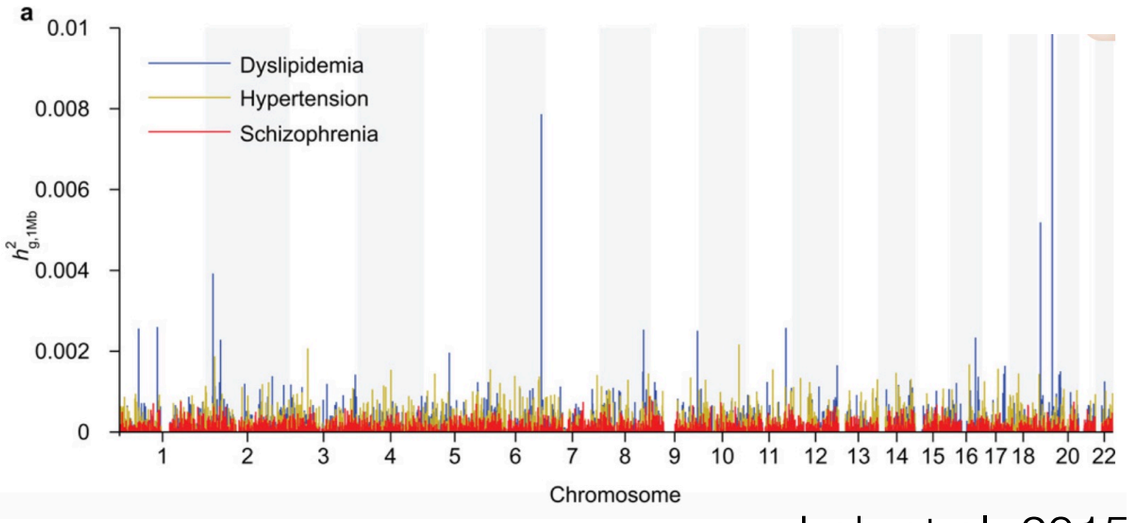
Genetic variance component

Environmental variance component

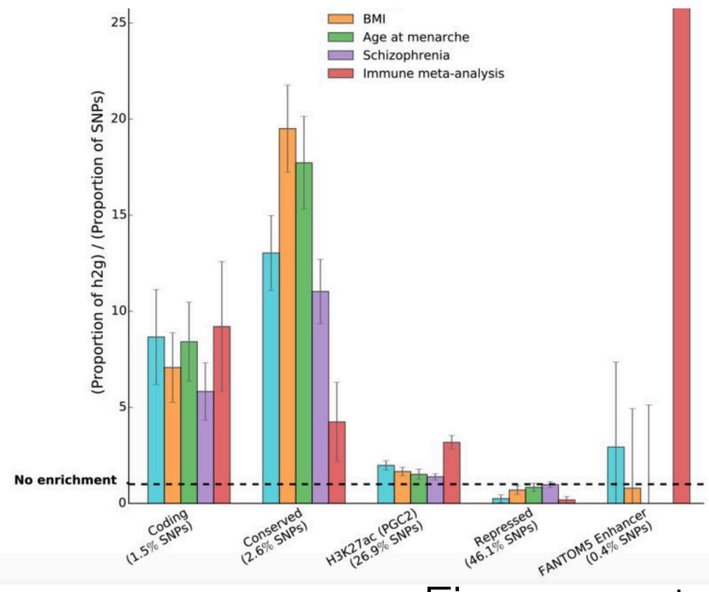
Beyond Heritability



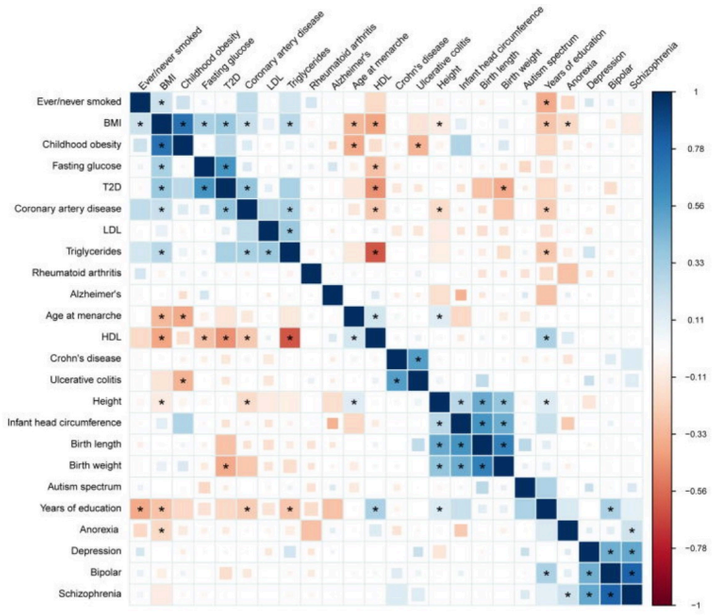
Lee et al. 2012



Loh et al. 2015



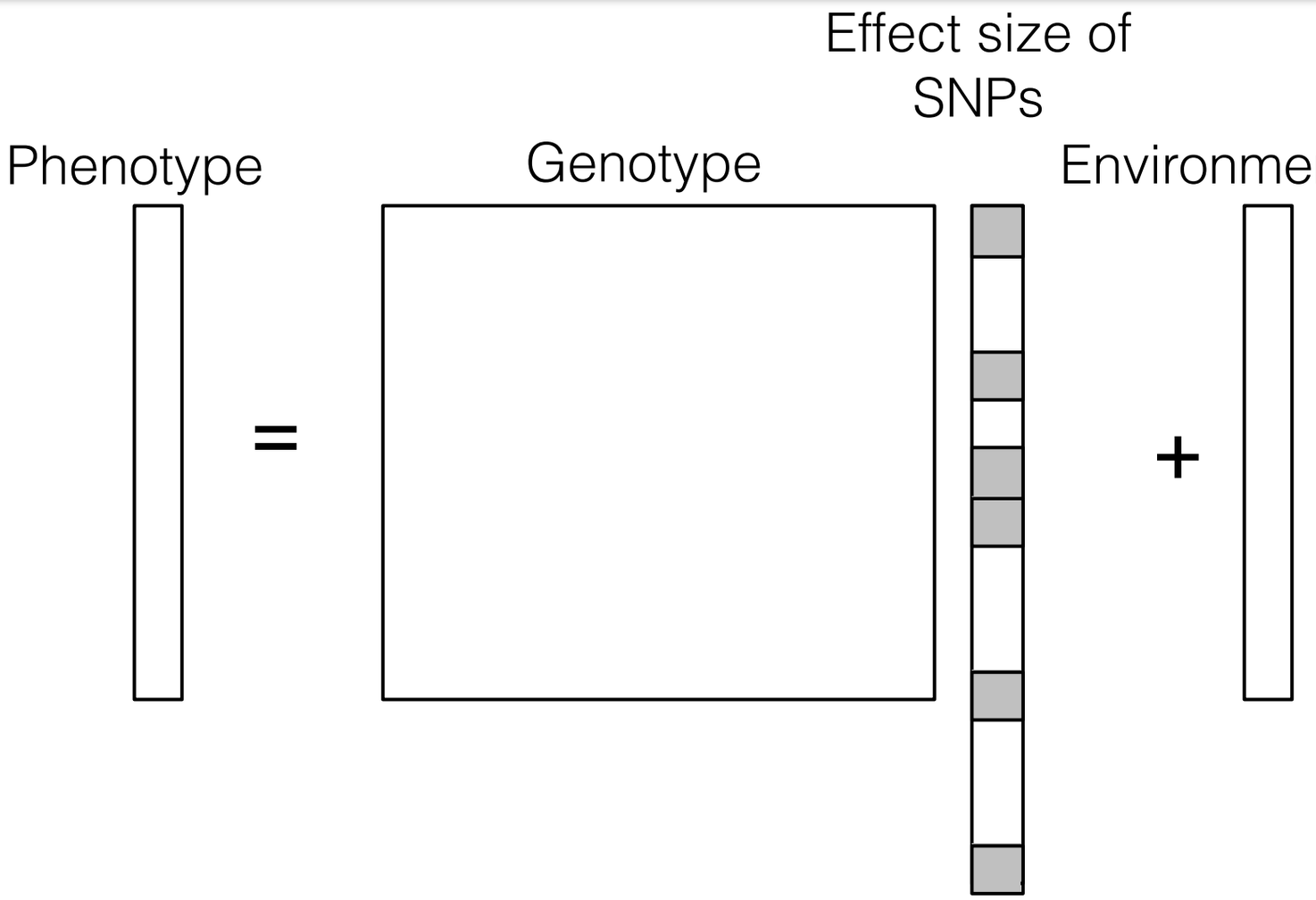
Finucane et al. 2015

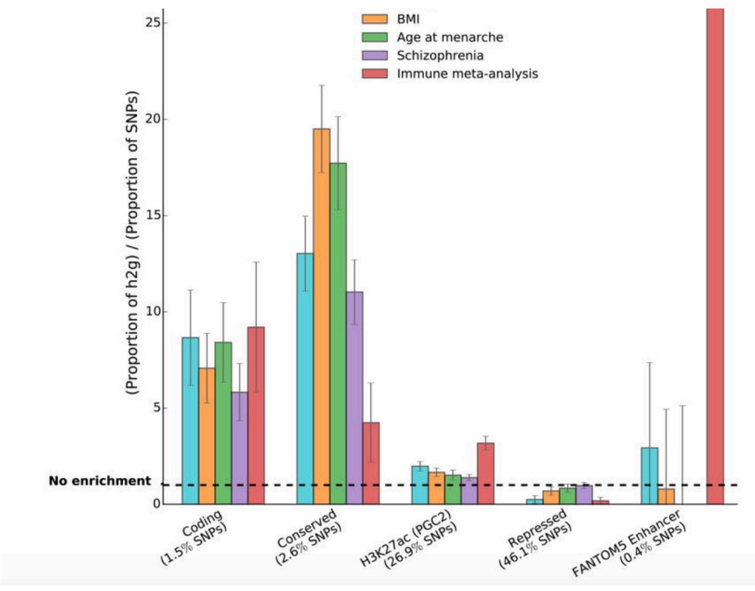


Bulk-Sullivan et al. 2015

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

Variance components model





Finucane et al. Nature Genetics 2015

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

$$\beta_m \sim \mathcal{N}(0, \sigma^2_{Gene}) \text{ if m is in Gene}$$

$$\beta_m \sim \mathcal{N}(0, \sigma^2_{NonGene}) \text{ otherwise}$$

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

$$(\hat{\sigma_1^2}, \hat{\sigma_2^2}, \ldots, \hat{\sigma_K^2}, \hat{\sigma_e^2}) = argmax_{(\sigma_1^2,\ldots,\sigma_K^2,\sigma_e^2)} \mathcal{LL}(\sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2)$$
$$= argmax_{(\sigma_1^2,\ldots,\sigma_K^2,\sigma_e^2)} P(\boldsymbol{y}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K, \sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2)$$

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

Alternate estimator

Method of Moments (HE-regression)

$$cov(\boldsymbol{y}) = \sum_k \sigma_k^2 \frac{1}{M_k} \boldsymbol{X}_k \boldsymbol{X}_k^T + \sigma_e^2 \boldsymbol{I}_N$$

$$= \sum_k \sigma_k^2 \boldsymbol{K}_k + \sigma_e^2 \boldsymbol{I}_N$$

$$\approx \boldsymbol{y}\boldsymbol{y}^T$$

Closed-form solution

Still needs computing GRMs $\boldsymbol{K}_k = \frac{1}{M_k} \boldsymbol{X}_k \boldsymbol{X}_k^T$

Scales as $\mathcal{O}(N^2 M)$

Haseman and Elston, AJHG 1972
Zhou Annals of Applied Statistics 2017

Randomized HE-regression (RHE-mc)

Avoid computing the GRM

$$tr(\boldsymbol{K}_l\boldsymbol{K}_k) \approx \frac{1}{B}\sum_b \boldsymbol{z}_b^T\boldsymbol{K}_l\boldsymbol{K}_k\boldsymbol{z}_b$$

$$= \frac{1}{M^2B}\sum_b \boldsymbol{z}_b^T\boldsymbol{X}_l\boldsymbol{X}_l^T\boldsymbol{X}_k\boldsymbol{X}_k^T\boldsymbol{z}_b$$

Multiply the genotype matrix with B random vectors

Hutchinson 1989

Wu et al. Bioinformatics 2018

Pazokitoroudi et al. Nature Communication 2020

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: $\mathcal{O}(\frac{MNB}{\log_3(\max(N,M))})$

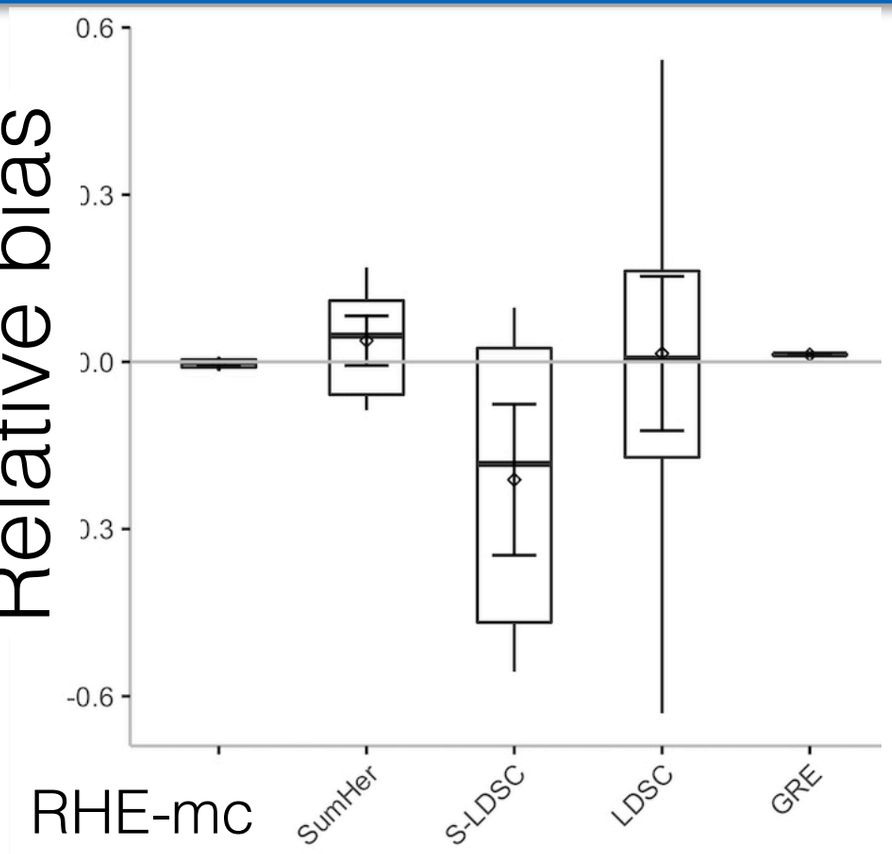
Accurate for B as small as 10

Hutchinson 1989

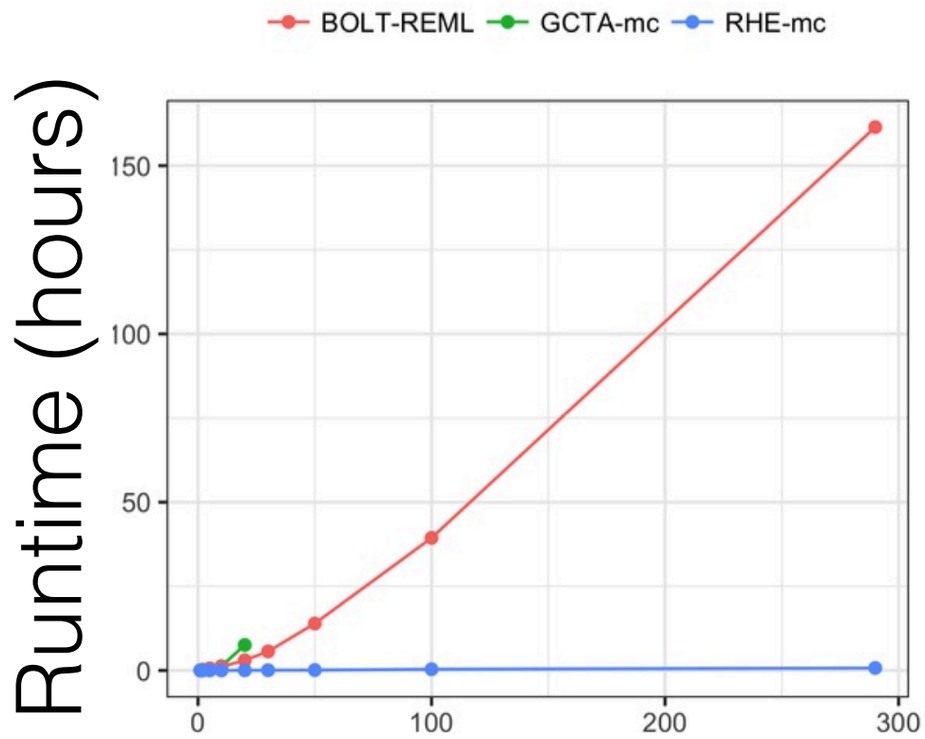
Wu et al. Bioinformatics 2018

Pazokitoroudi et al. RECOMB 2019, Nature Communication 2020

RHE-mc is accurate and scalable



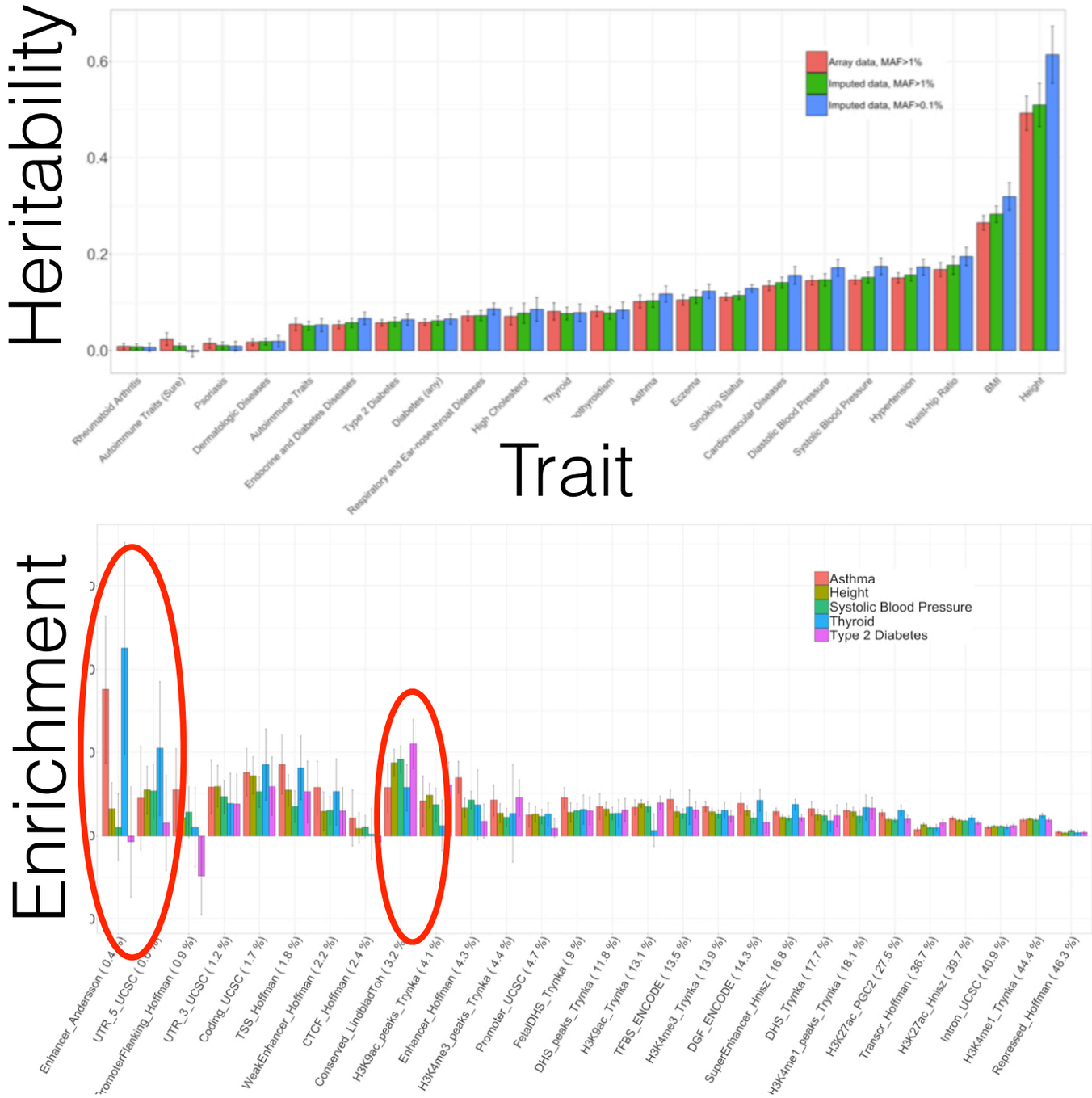
Summary-statistic methods

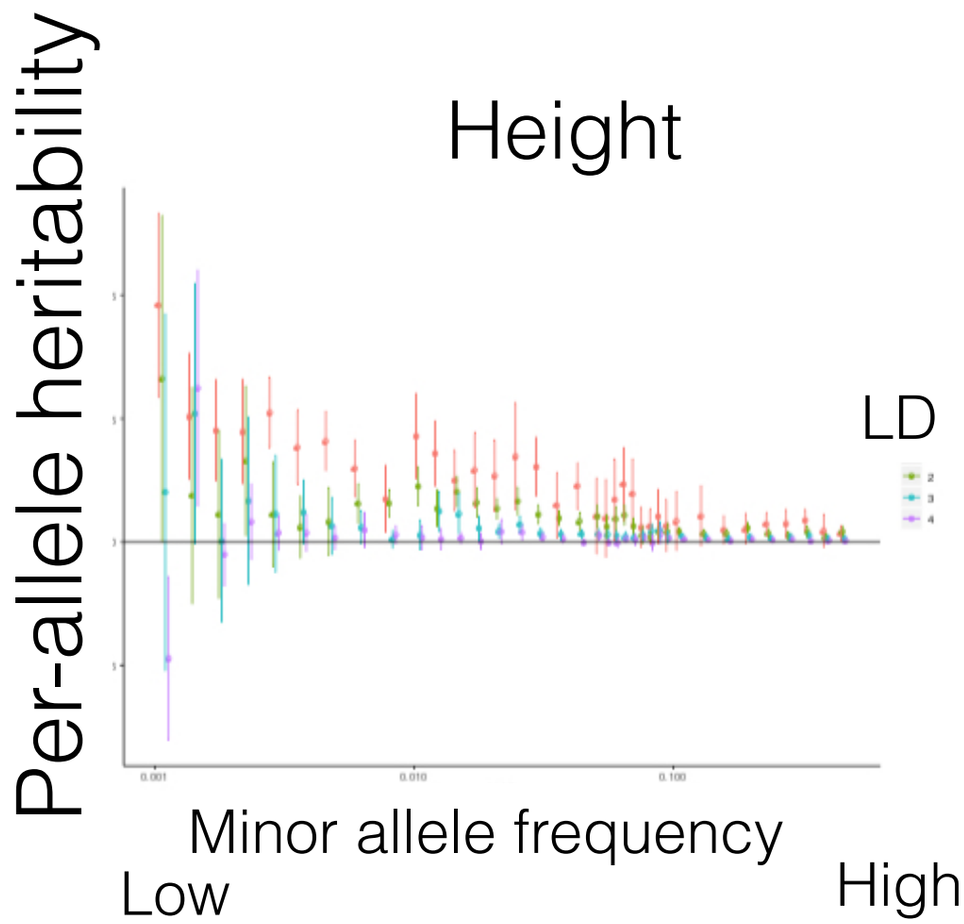


Parameters			Running time (hour)		
N	M	K	RHE-mc	GCTA-mc	BOLT-REML
10,000	459,792	22	< 1	1.3	1
100,000	459,792	22	< 1	-	40
291,273	459,792	22	< 1	-	162
291,273	459,792	300	< 1	-	-
291,273	4,824,392	8	3.2	-	-
1,000,000	1,000,000	8	3	-	-
1,000,000	1,000,000	100	12.4	-	-

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

Insights from Biobank-scale analysis

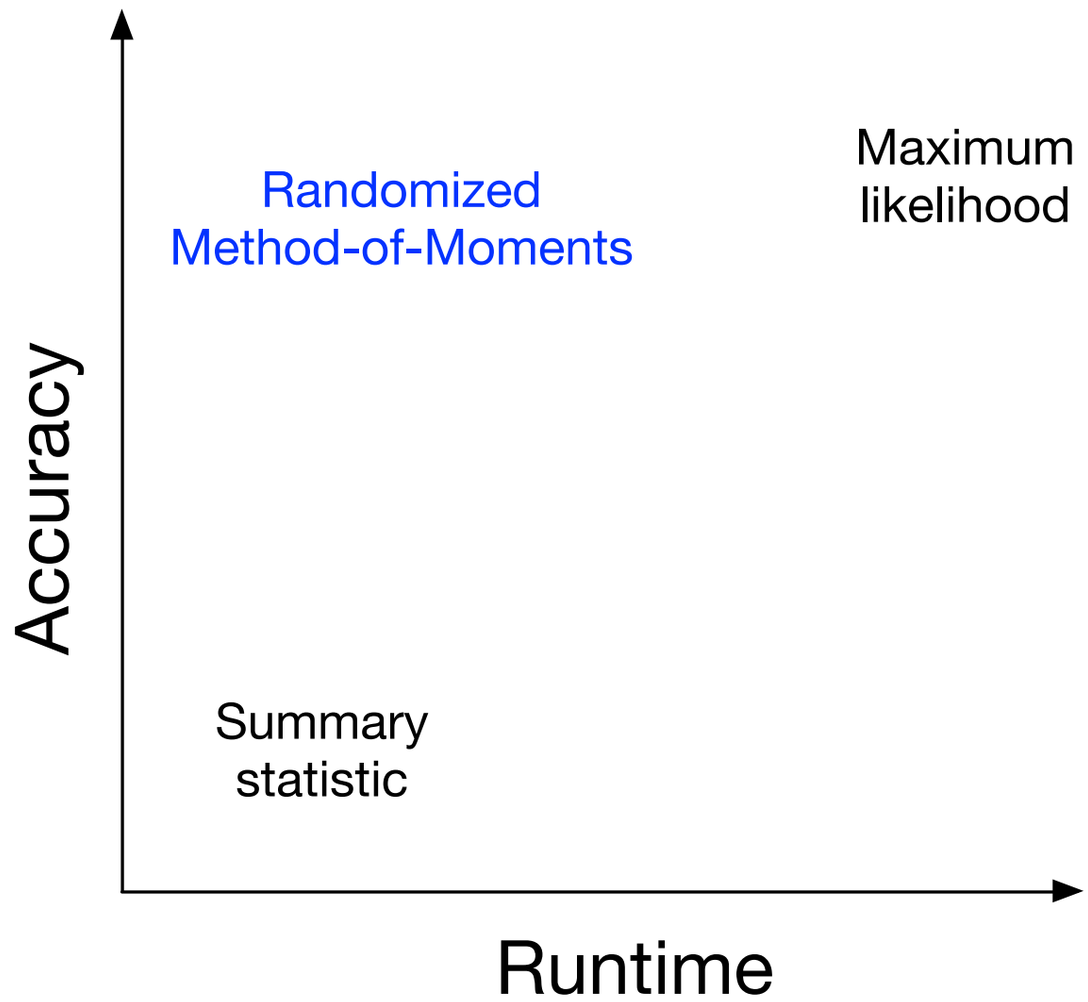




Pazokitoroudi et al. Nature Communication 2020

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

RHE-mc


Accuracy
Randomized
Method-of-Moments
Maximum
likelihood
Summary
statistic
Runtime

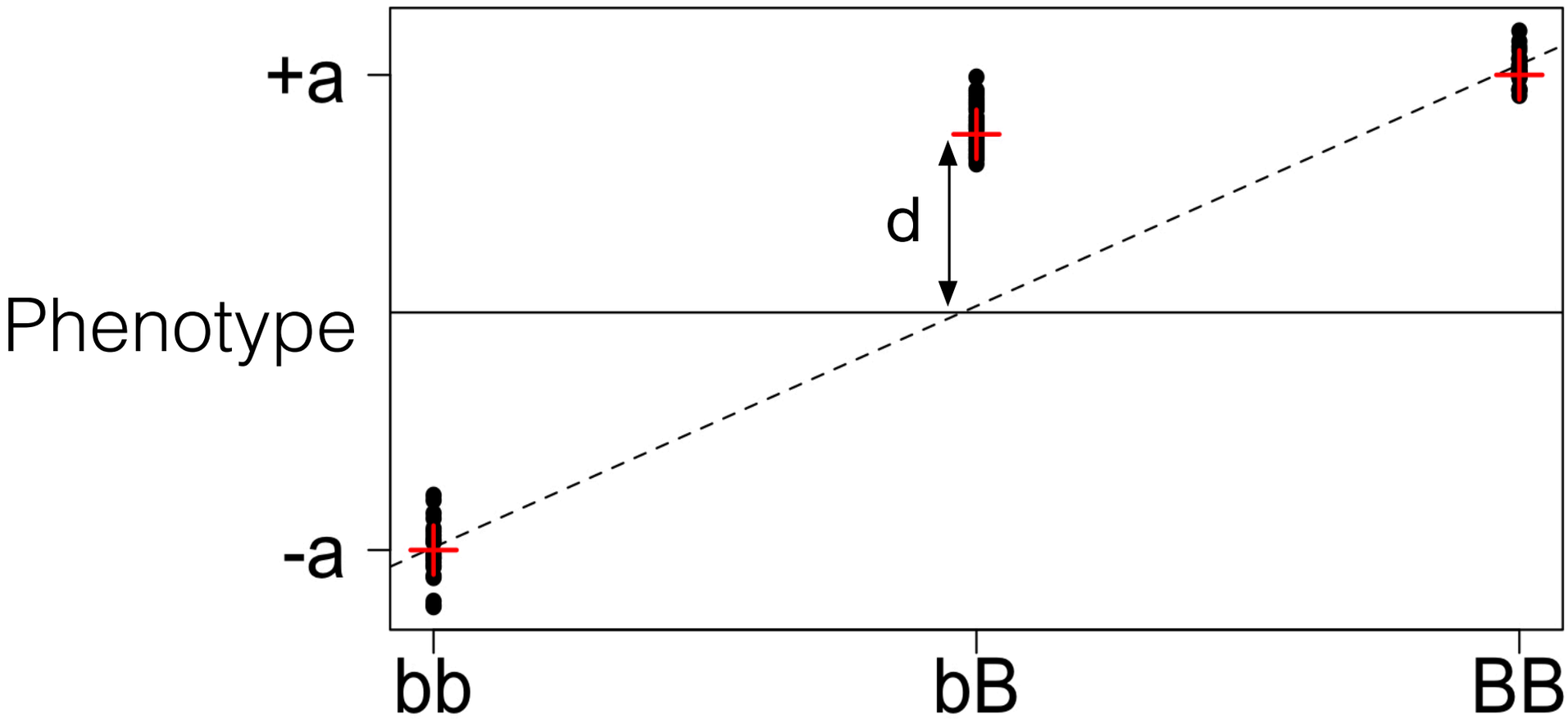
Beyond heritability

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



Dominance deviation effects

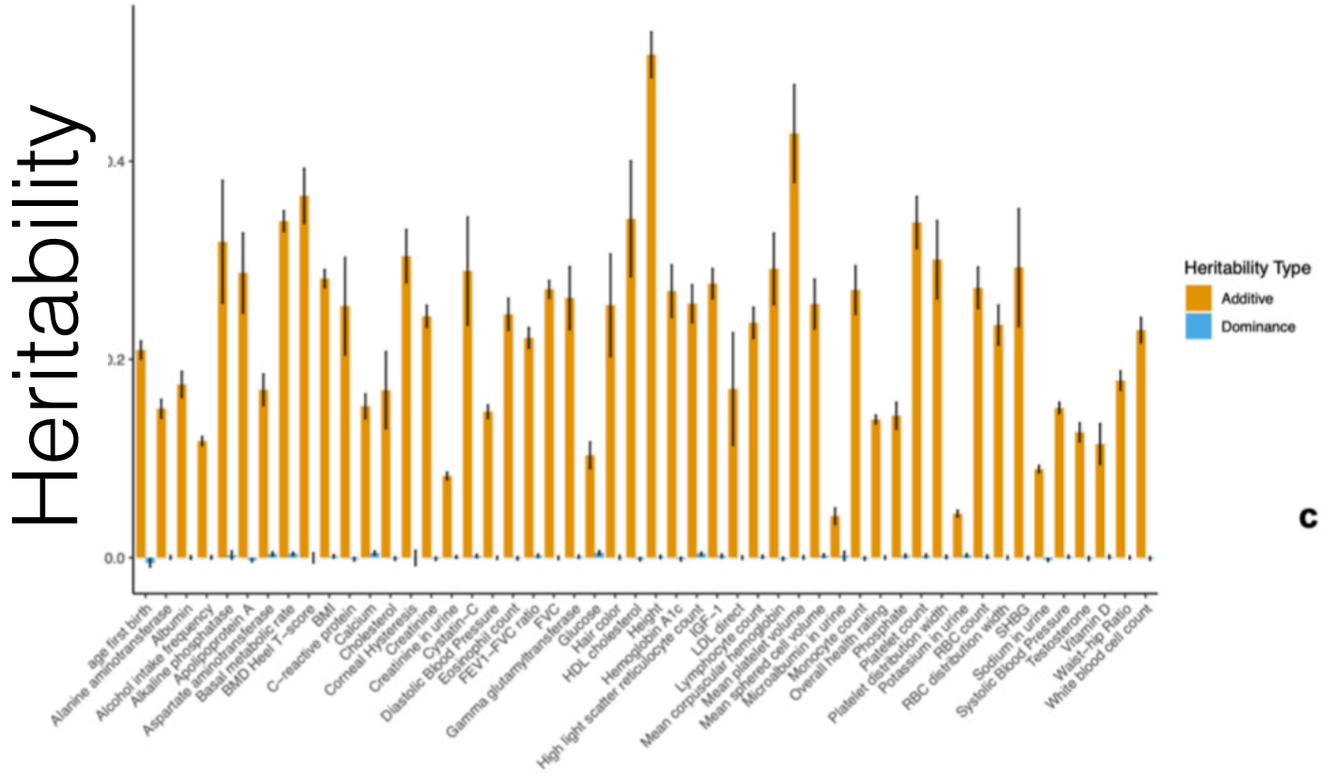
Additive variance component

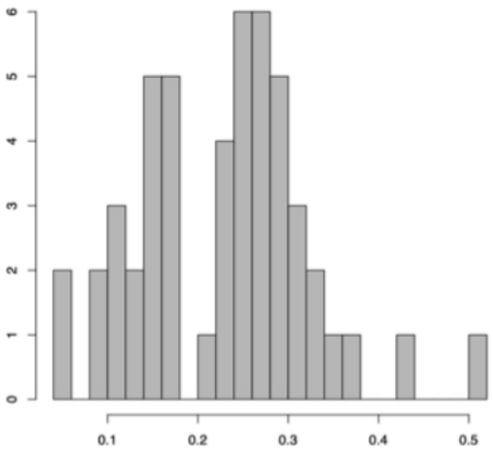
Dominance variance component

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

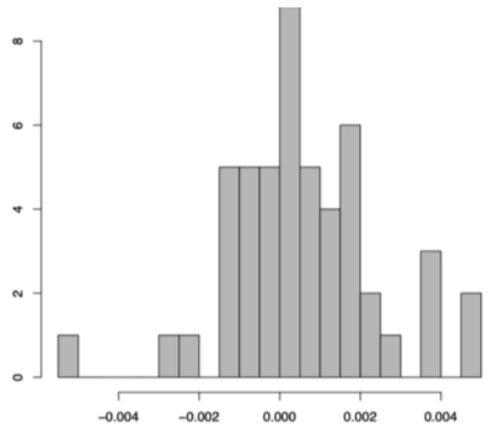
$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \frac{\sigma_a^2}{M}\boldsymbol{I}_M)$$

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





Additive heritability

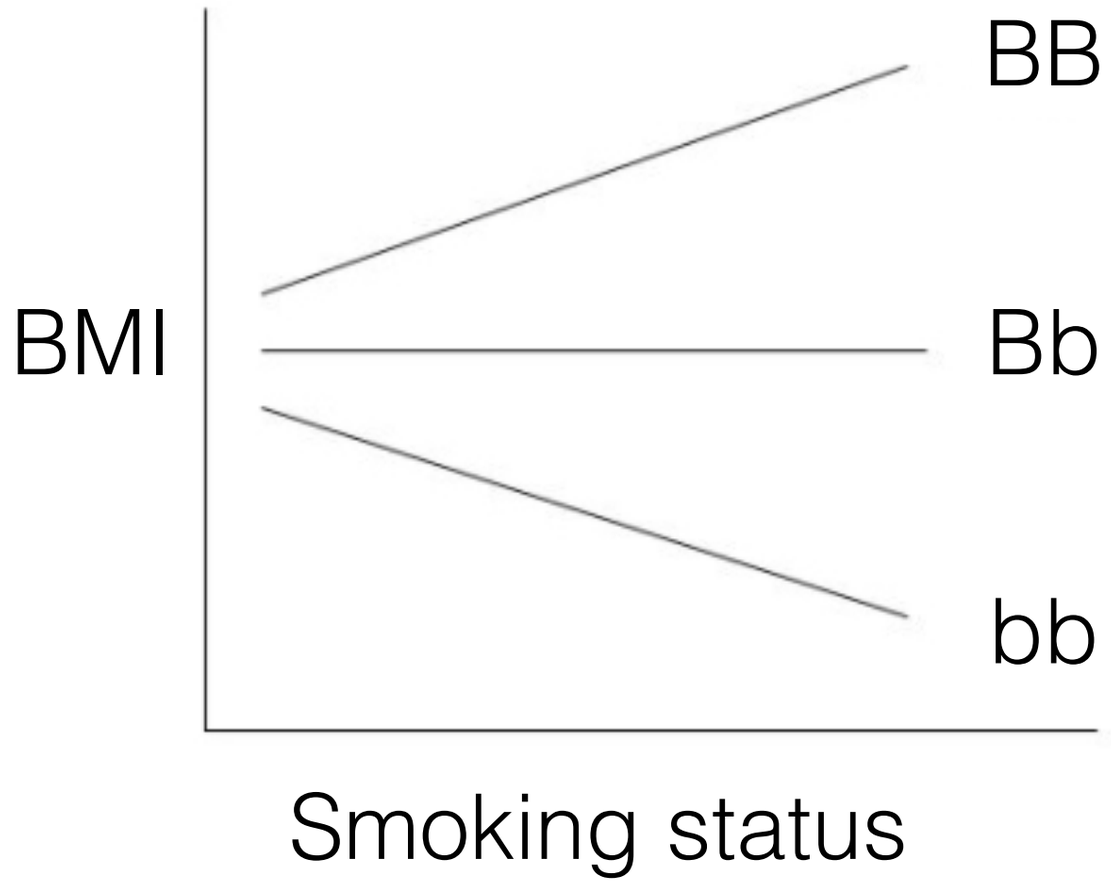


Dominance heritability

Hivert et al. AJHG 2021

Pazokitoroudi et al. AJHG 2021

Gene-environment interactions (GxE)



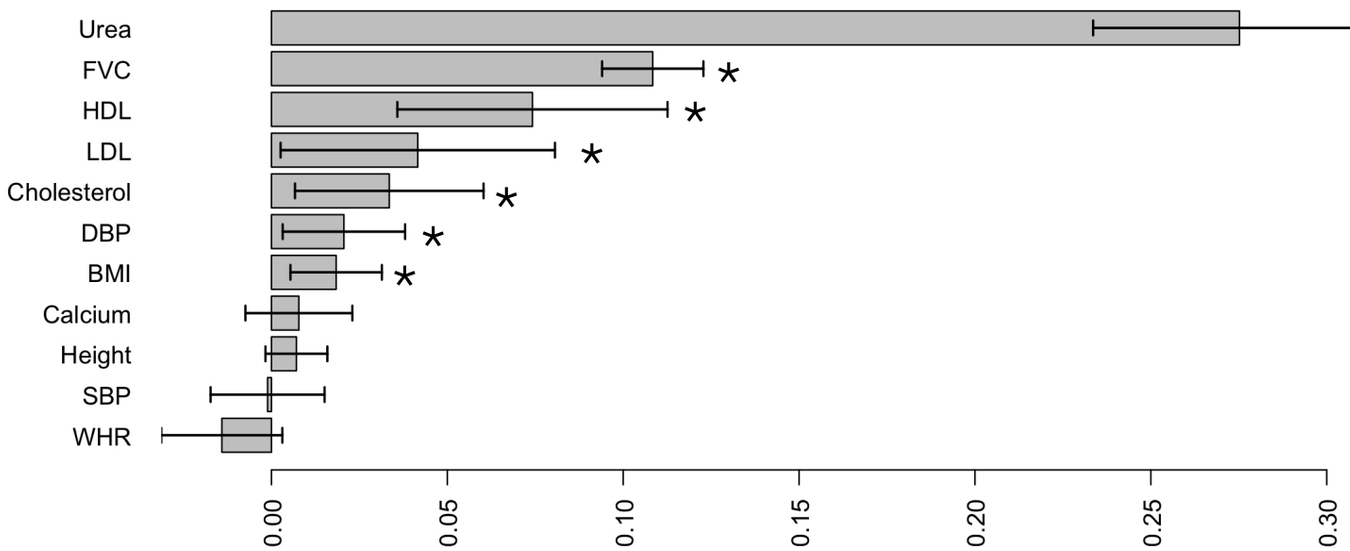
Gene-environment interactions (GxE)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X} \odot \boldsymbol{E}\boldsymbol{\delta} + \boldsymbol{\epsilon}$$

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

GxE variance component

E = Smoking



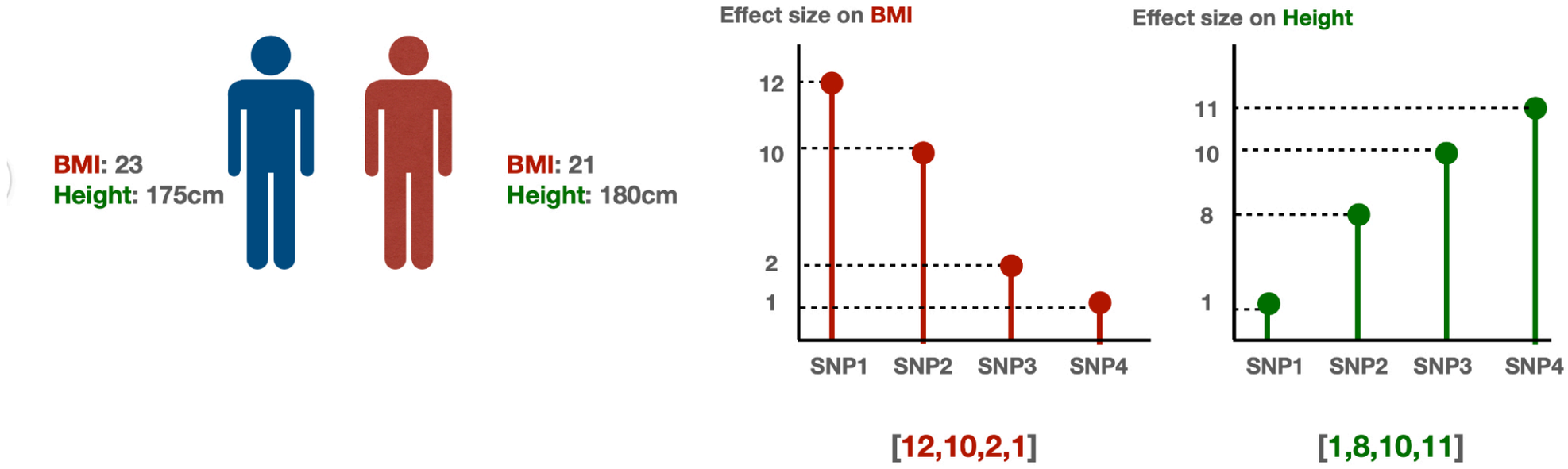
$$\frac{GxE \quad heritability}{Additive \quad heritability}$$

Kerrin and Marchini AJHG 2020

Pazokitoroudi et al. RECOMB 2021

Genetic effects shared across traits

Genetic correlation



Estimating genetic correlation

$$\boldsymbol{y_1} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{\epsilon_1}$$

$$\boldsymbol{y_2} = \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon_2}$$

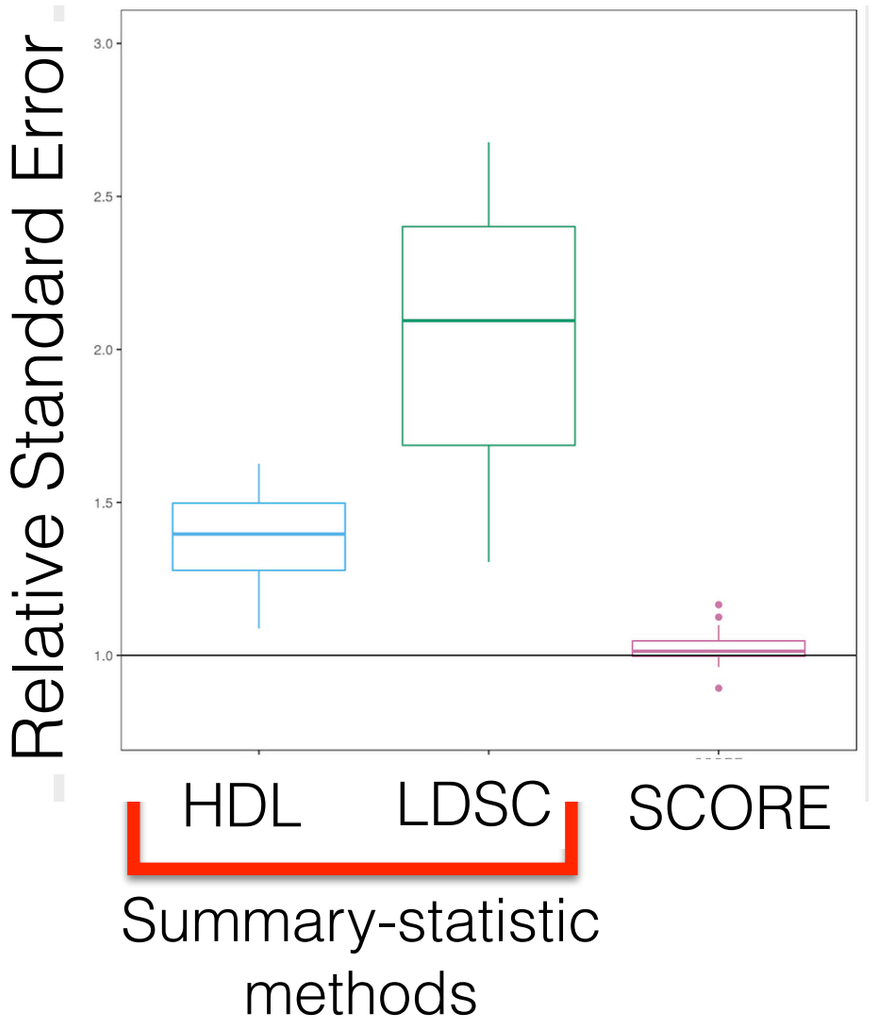
$$\epsilon_{1n}, \epsilon_{2n} \sim \mathcal{N}(\mathbf{0}, \begin{bmatrix} \frac{\sigma_{e1}^2}{N} & \frac{\gamma_e}{N} \\ \frac{\gamma_e}{N} & \frac{\sigma_{e2}^2}{N} \end{bmatrix})$$

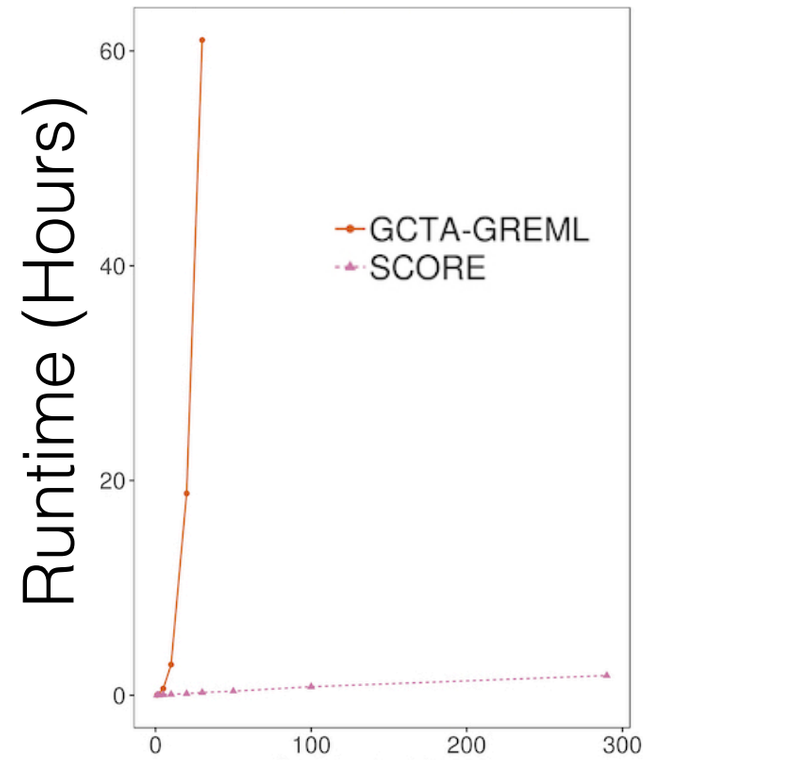
$$\beta_{1m}, \beta_{2m} \sim \mathcal{N}(\mathbf{0}, \begin{bmatrix} \frac{\sigma_{g1}^2}{M} & \frac{\gamma_g}{M} \\ \frac{\gamma_g}{M} & \frac{\sigma_{g2}^2}{M} \end{bmatrix})$$

Coupling of genetic effects

Wu et al. RECOMB 2019, BioRXiv 2020

Scalable genetic correlation estimator (SCORE)

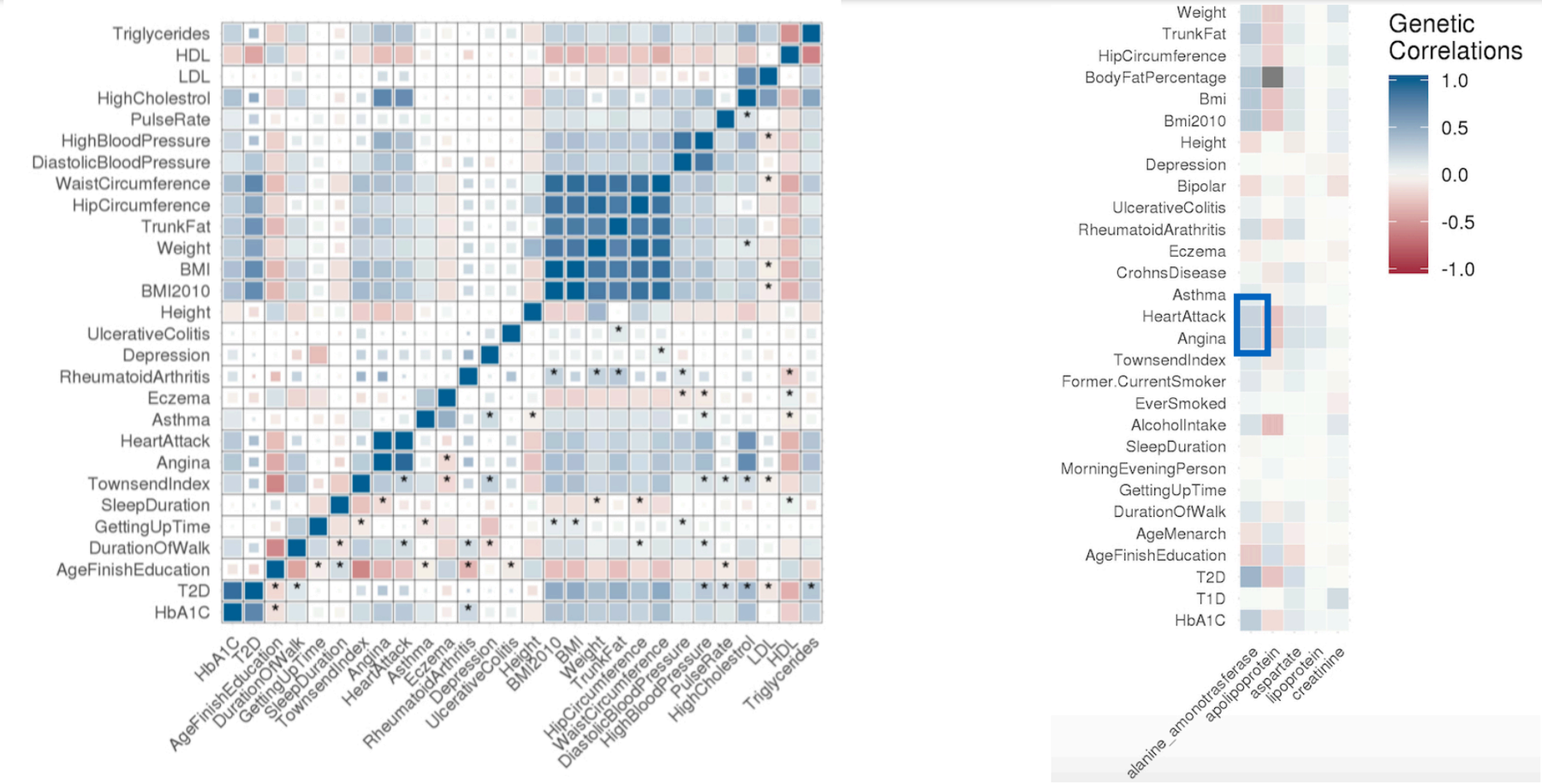




Wu et al. RECOMB 2019, BioRXiv 2020

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

Genetic correlation



Wu et al. RECOMB 2019, BioRXiv 2020

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

Promises and challenges of Biobank-scale datasets

Challenging problems of statistical inference

Variance components analysis

- Heritability estimation and partitioning
- Non-linear contributions: Dominance and GxE
- Genetic correlation

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

Promises and challenges of Biobank-scale datasets

Challenging problems of statistical inference

Variance components analysis

Action of negative selection on complex traits

Substantial additive heritability, negligible dominance heritability with substantial GxE for specific gene-environment combinations

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

Promises and challenges of Biobank-scale datasets

Challenging problems of statistical inference

Variance components analysis

Ongoing work

Biobank-scale association testing and prediction

Multi-ethnic analyses

Combining data modalities

UCLA
Sriram Sankararaman
Machine Learning and Genomics Lab

Acknowledgments

UCLA
Bogdan Pasaniuc
Kathryn Burch
Paivi Pajukanta
Noah Zaitlen

University of Chicago
Andy Dahl

FIMM
Andrea Ganna

Tel Aviv University
Saharon Rosset

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