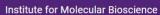
KITP Polygenic adaptation

# Evidence for selection using human GWAS data

Naomi R Wray

naomi.wray@uq.edu.au







Queensland Brain Institute





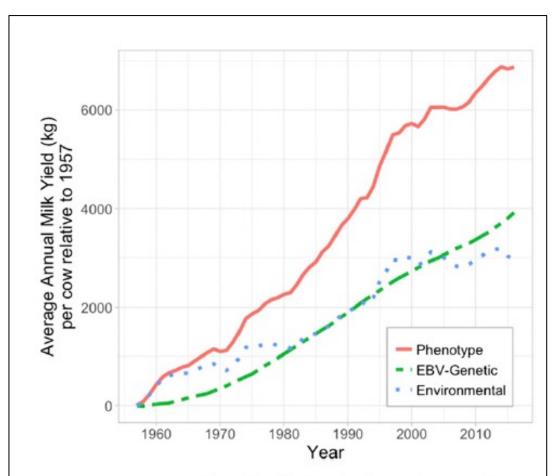


# PERSPECTIVES

# Polygenic adaptation: a unifying framework to understand positive selection

Neda Barghi, Joachim Hermisson, and Christian Schlötterer

# Genetic Improvement in Dairy Cattle



**Figure 2** Increase in milk yield in black and white Holstein cattle since 1957. The mean EBV has increased by 3916 or 66 kg per cow per year. The phenotypic and genetic SD of milk yield in 1957 were ~1200 and ~600 kg. Hence, the genetic contribution to milk yield has increased by ~6.5 genetic SD since 1957. Source: Council on Dairy Cattle Breeding (https://queries.uscdcb.com/eval/summary/trend.cfm)

Complex Trait Prediction from Genome Data:
Contrasting EBV in Livestock to PRS in Humans

Naomi R. Wray,\*\*\* Kathryn E. Kemper,\* Benjamin J. Hayes,\* Michael E. Goddard,\*\*\*
and Peter M. Visscher\*\*

The genetic contribution to milk yield has increased 6.5 genetic SD since 1957

70 years; generation interval ~ 5 years

25 million B&W cattle worldwide Effective population size: ~75







# Challenging selection theory with experiments



Quant Gen 101

$$A_{child} = \frac{1}{2}A_{dad} + \frac{1}{2}A_{mum} + A_{seg}$$

$$V(A_{seg}) = \frac{1}{2}V(A)$$





Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

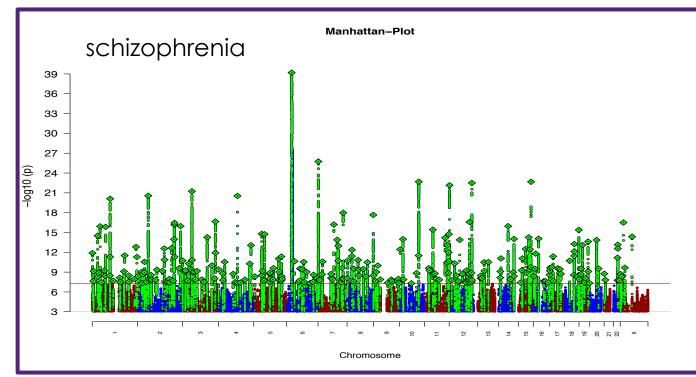
Can more be learned from selection experiments of value in animal breeding programmes? Or is it time for an obituary?

W.G. Hill



# Common diseases are polygenic





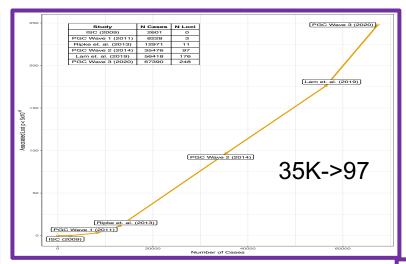
248 risk loci identified at genome-wide significance level

We predict thousands are associated









67K ->248



#### Article

Mapping genomic loci implicates genes and synaptic biology in schizophrenia

Nature 2022

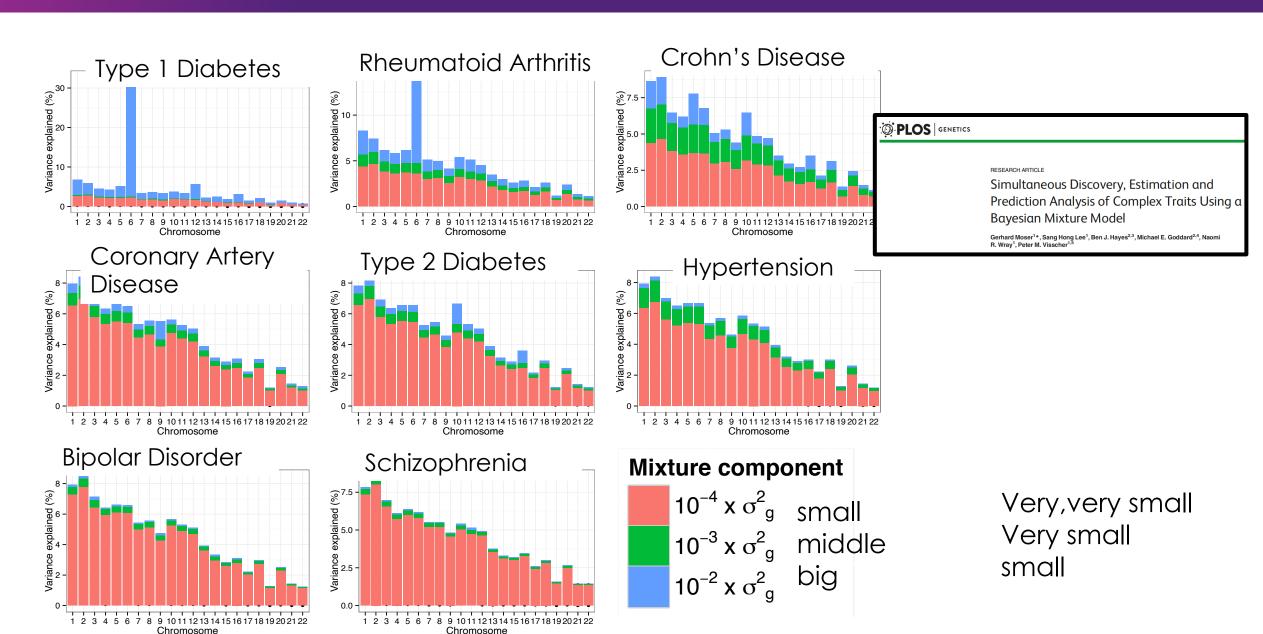






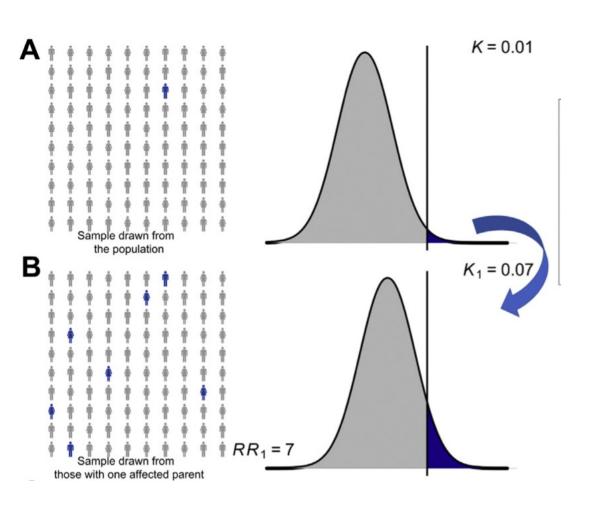
## Many polygenic genetic architectures

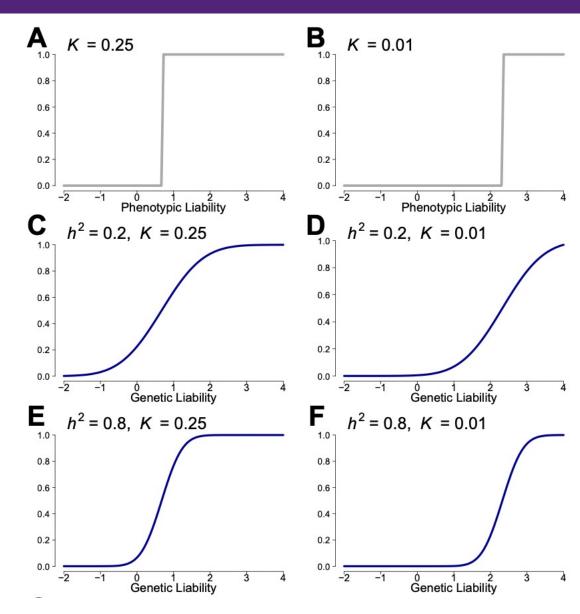




## Robustness model



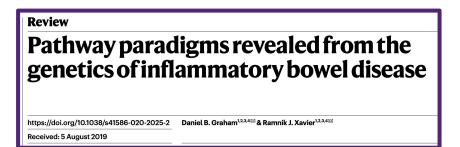


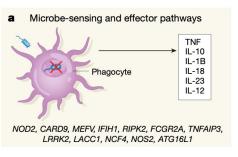


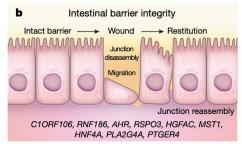
Risk in Relatives, Heritability, SNP-Based Heritability, and Genetic Correlations in Psychiatric Disorders: A Review

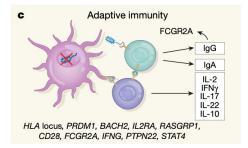
## Common diseases are complex

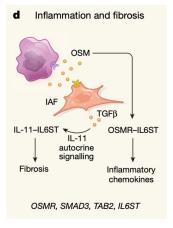


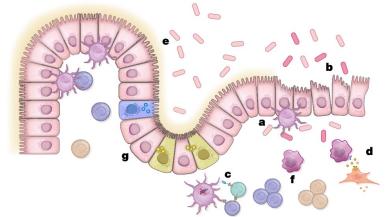


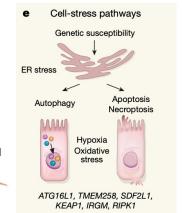


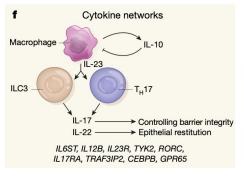


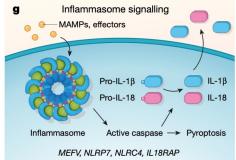










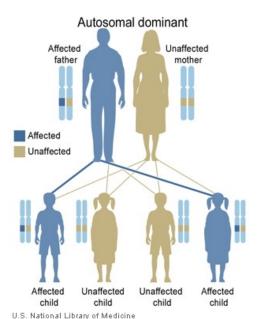


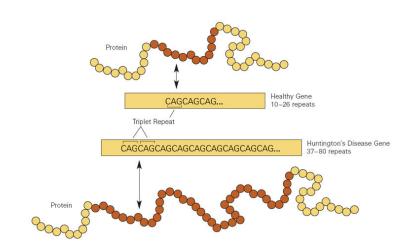
# All traits are polygenic

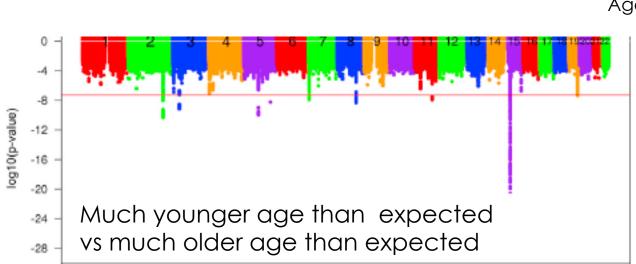


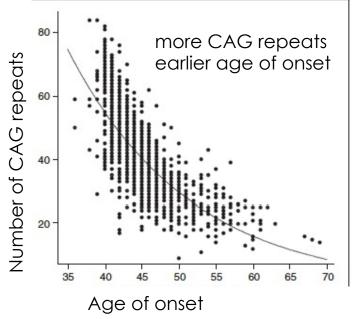
## Lessons from Huntington's Disease











Gusella & MacDonald (2009)

9000 HD cases ~10 GWS loci

Article

CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset

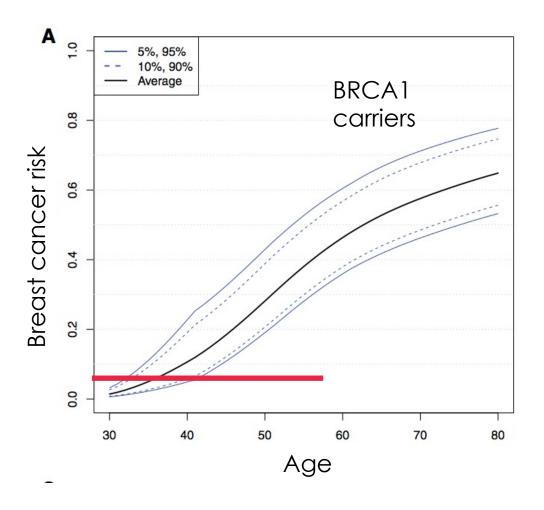
Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium<sup>1</sup>
\*Lead contact (James F. Gusella)

\*Correspondence: gusella@heix.mgh.harvard.edu (James F. Gusella)

## Large effect and common effect variants combine



## (as expected)





## Observational data for selection in humans

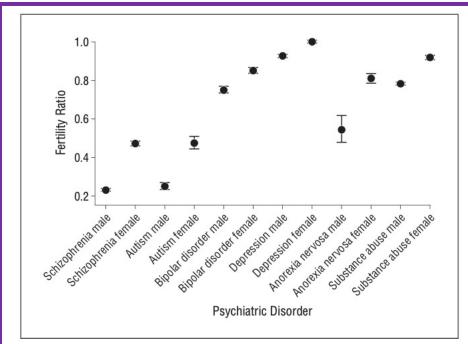


#### **ORIGINAL ARTICLE**

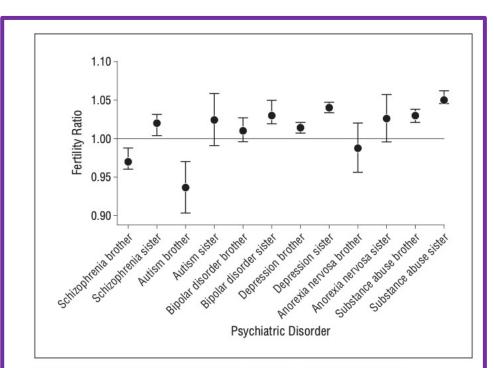
#### **ONLINE FIRST**

Fecundity of Patients With Schizophrenia, Autism, Bipolar Disorder, Depression, Anorexia Nervosa, or Substance Abuse vs Their Unaffected Siblings

Robert A. Power, BSc; Simon Kyaga, MD; Rudolf Uher, MD, PhD, MRCPsych; James H. MacCabe, PhD, MRCPsych; Niklas Långström, MD, PhD; Mikael Landen, MD, PhD; Peter McGuffin, FRCP, FRCPsych, PhD; Cathryn M. Lewis, PhD; Paul Lichtenstein, PhD; Anna C. Svensson, PhD



**Figure 1.** Fertility ratios for individuals with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, and substance abuse. A fertility ratio of 1 (highlighted) represents that of the general population.



**Figure 2.** Fertility ratios for unaffected siblings of individuals with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, and substance abuse. A fertility ratio of 1 (highlighted) represents that of the general population.

# Signatures of negative selection

a

100

50

20

**GRIA3** 

CUL1

GRIN2A

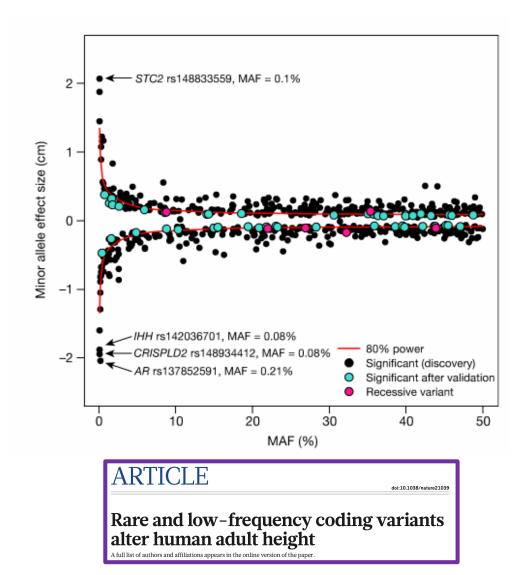
SETD1A



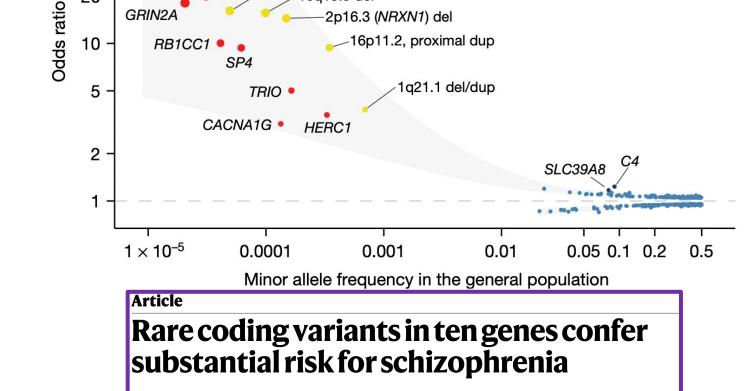
Common variants

Copy number variants

Protein-truncating variants



Marouli et al. 2017 (Nature)



2p16.3 (NRXN1) del

Singh et al. 2022 (Nature)

3q29 del

22q11.21 del

16p11.2, distal del

7q11.23 dup

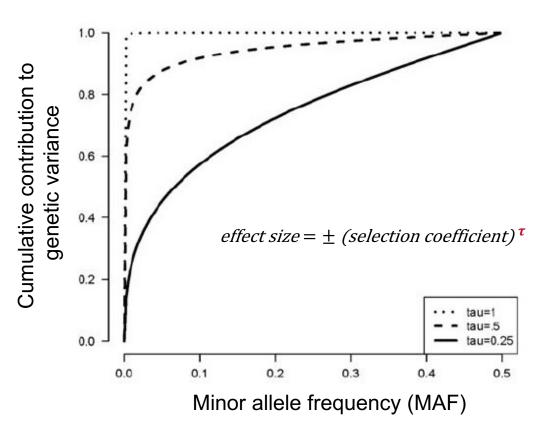
15q13.3 del

# Signatures of negative selection



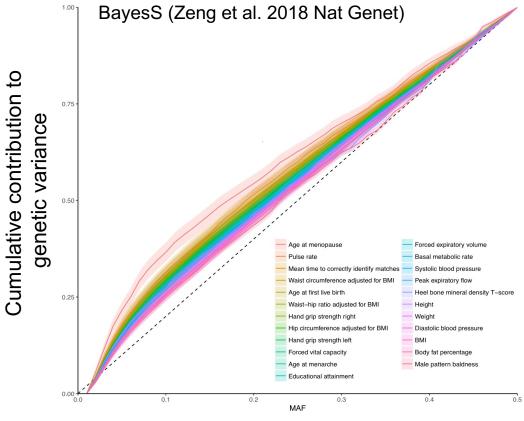
13

#### Theoretical prediction



Eyre-Walker 2010 PNAS Visscher et al 2012 Mol Psych

# 23 out of 28 traits in UK Biobank (max n=120k) have significant signatures of negative selection.



Minor allele frequency (MAF)

### Active area of research



ARTICLES
Representation

Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection

2018

ANALYSIS

Intro-Vidal org./10.1038/v41588-018-0231-8

Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations

Steven Gazal © 1.2\*, Po-Ru Loh²2, Hilary K. Finucane²4, Andrea Ganna²45, Armin Schoech¹27, Shamil Sunyaev²38 and Alkes L. Price © 1.27\*

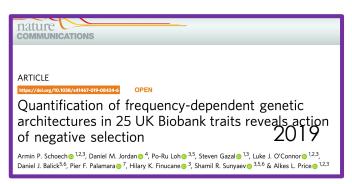
#### **ARTICLE**

Extreme Polygenicity of Complex Traits Is Explained by Negative Selection

Antonio F. Pardiñas 1, Peter Holmans, Andrew J. Pocklington, Valentina Escott-Price

Luke J. O'Connor, 1,2,\* Armin P. Schoech, 1 Farhad Hormozdiari, 1 Steven Gazal, 1 Nick Patterson, 3 and Alkes L. Price 1,3,\*

2019





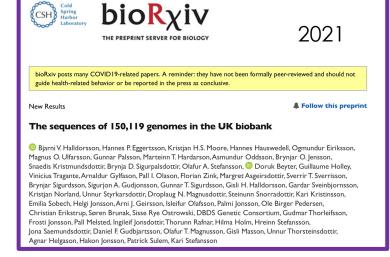
Research

# Genome-wide signals of positive selection in human evolution 2014

David Enard,<sup>1</sup> Philipp W. Messer, and Dmitri A. Petrov<sup>1</sup>

Department of Biology, Stanford University, Stanford, California 94305, USA





# Bayesian random regression: BayesS

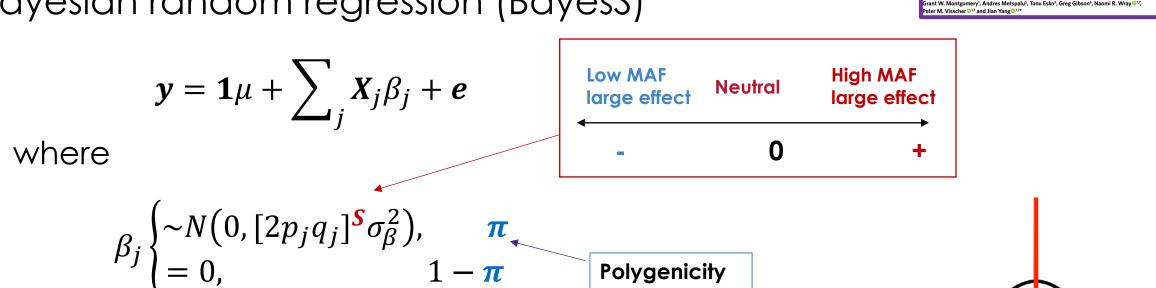


Signatures of negative selection in the genetic architecture of human complex traits

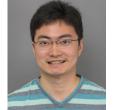
genetics

**ANALYSIS** 

• Bayesian random regression (BayesS)



- SNP-based heritability is estimated based on the variance of genetic values
- Simultaneously estimate SNP effects and model parameters using MCMC



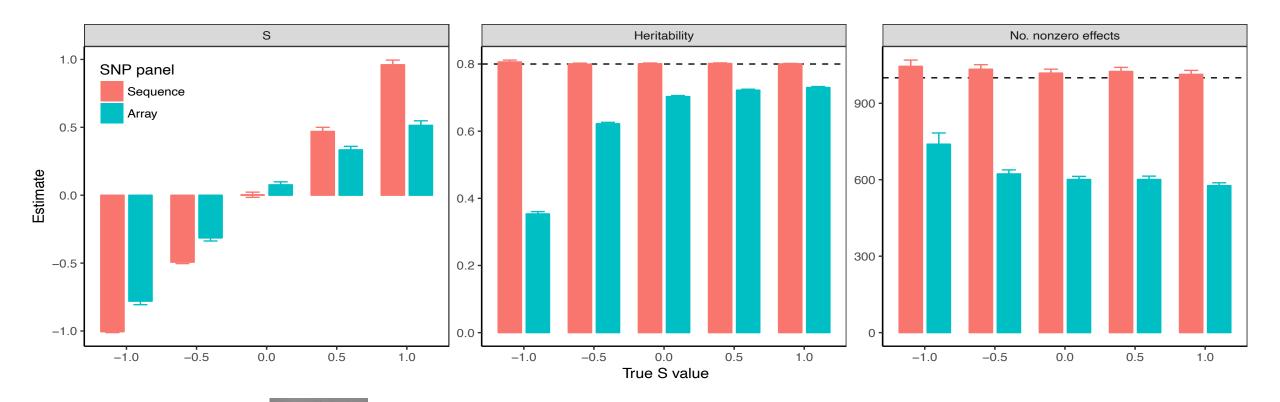
Jian Zeng

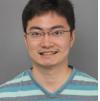


## Simulation based on WGS data



• UK10K sequence data, chr 21 & 22, n = 3,642, m  $\approx$  500k sequence SNPs or 1.5k array SNPs





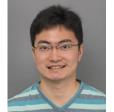
# **UK Biobank analysis**





#### 28 traits

- 24 quantitative: anthropometric, cardiovascular, reproductive
- 2 categorical: male pattern baldness (MPB), educational attainment (EA)
- 2 diseases: type 2 diabetes (T2D), major depressive disorder (MDD) Max N = 126k (unrelated Europeans) for the interim release
- ~500k Affymetrix SNPs (MAF > 1%) after QC



Jian Zeng

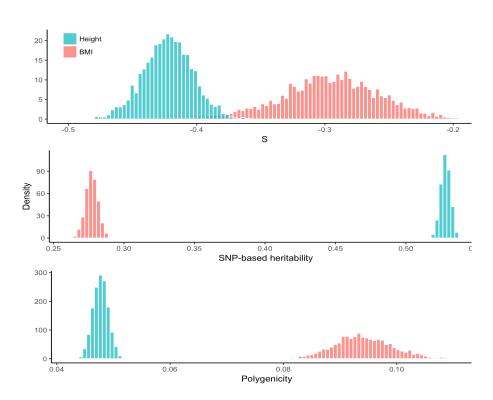
**ANALYSIS** nature genetics Signatures of negative selection in the genetic architecture of human complex traits Jian Zeng¹, Ronald de Vlaming 02.3, Yang Wu¹, Matthew R. Robinson¹.4, Luke R. Lloyd-Jones¹, Loic Yengo¹, Chloe X. Yap 👵¹, Angli Xue¹, Julia Sidorenko¹,⁵, Allan F. McRae¹, Joseph E. Powell¹, Grant W. Montgomery¹, Andres Metspalu⁵, Tonu Esko⁵, Greg Gibson6, Naomi R. Wray <sup>⊚ 1,7</sup>, Peter M. Visscher 17 and Jian Yang 17.7\*

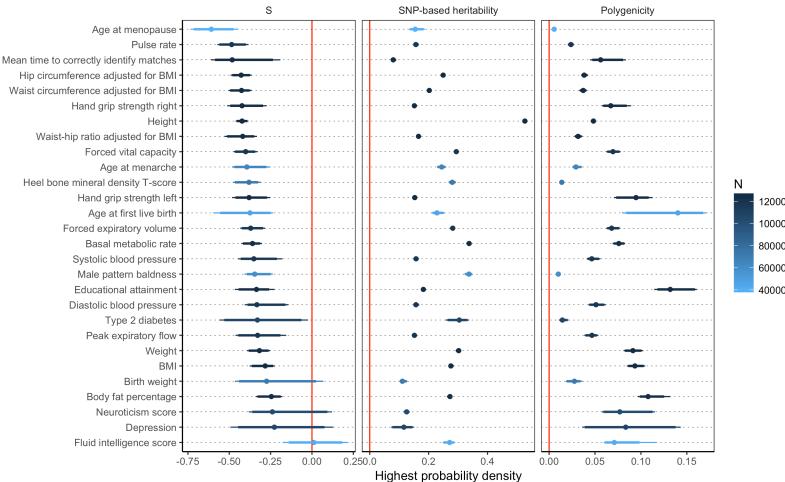
# **UK Biobank analysis**

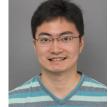


#### Estimated genetic architecture: height vs. BMI

#### Genetic architecture of 28 traits







On average 6% of SNPs explain 22% of phenotypic variance

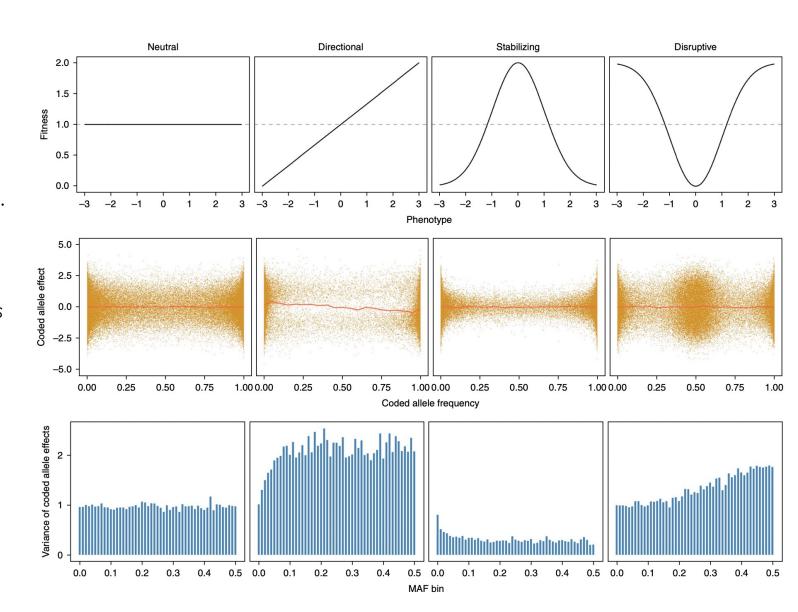
Jian Zeng

## Investigate relationship between MAF and effect size



# Use SLIM v2.3 to investigate selection models:

- 10-Mb region
- mutation rate 1.65 x 10<sup>-8</sup>
- new mutations probability
  - > 0.95 neutral
  - $\triangleright$  0.05 causal effect sampled from N(0,1).
- Phenotype based on the cumulated genotypic values
  - heritability of 0.1 across all causal variants in the current generation.
- Evolution of a population of 1,000 individuals over 10,000 generations
  - (equivalent to 10,000 individuals in a population of 100,000 generations).
- Burn-in 5,000 generations
  - phenotype did not affect fitness
  - all variants under neutral variation.
- Generation 5,001 on
  - standardized phenotype, with mean 0 and variance 1
  - > phenotype related to fitness
- 200 simulation replicates.
- Results robust to demographic model of bottleneck and expansion

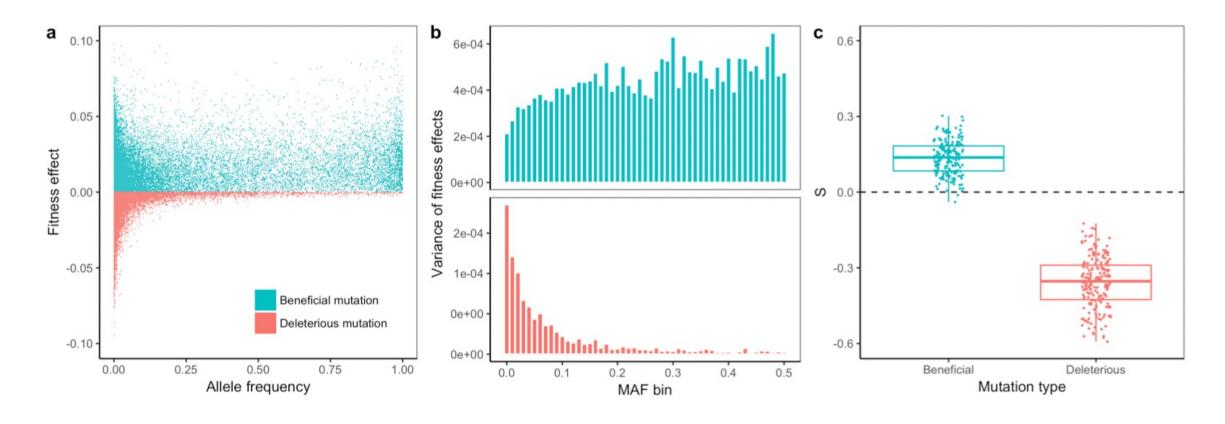


# Estimate of S does reflect s



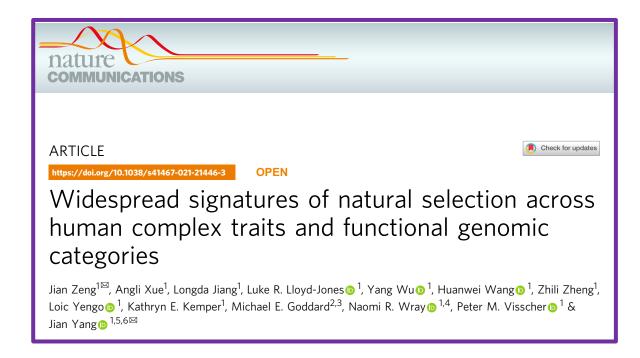
5% of new mutations beneficial wrt fitness OR

5% of new mutation deleterious wrt fitness



## SBayesS Method





## **GWAS** summary statistics

**SNPID** Chromosome Base Pair position Reference Allele Frequency of reference allele Effect size of reference allele Standard error Sample size

- Most GWAS are meta-analyses from multiple cohorts
- Summary statistics more easily shared than individual level data
- Computational efficiency





**Jian Yang** 

## Summary-data-based model



Consider an individual-data model with a standardised genotype matrix X:

$$y = X\beta + e$$

Multiply both sides by  $\frac{1}{N}$  **X**' gives

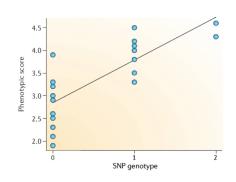
$$\frac{1}{N}\mathbf{X}'\mathbf{y} = \frac{1}{N}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \frac{1}{N}\mathbf{X}'\mathbf{e}$$

 $\mathbf{b} = \mathbf{R} \boldsymbol{\beta} + \boldsymbol{\epsilon}$ 



 $Var(\boldsymbol{\epsilon}) = \frac{1}{N} \mathbf{R} \sigma_e^2$ 

#### **GWAS** marginal SNP effects



#### LD correlation matrix



▶ Prior for SNP effect:

$$\beta_{j} \begin{cases} \sim N\left(0, \left[2p_{j}q_{j}\right]^{s} \sigma_{\beta}^{2}\right), & \pi \\ = 0, & 1 - \pi \end{cases}$$

- ▶ *S* quantifies the relationship between SNP effect size and minor allele frequency (a signature of selection).
- $\blacktriangleright$   $\pi$  quantifies the proportion of SNPs with nonzero effects (polygenicity).
- ► SNP-based heritability:

$$h_{SNP}^2 = eta' \mathbf{R} eta / V_P$$

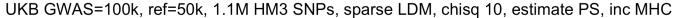
▶ Implemented in *GCTB* (https://cnsgenomics.com/software/gctb).

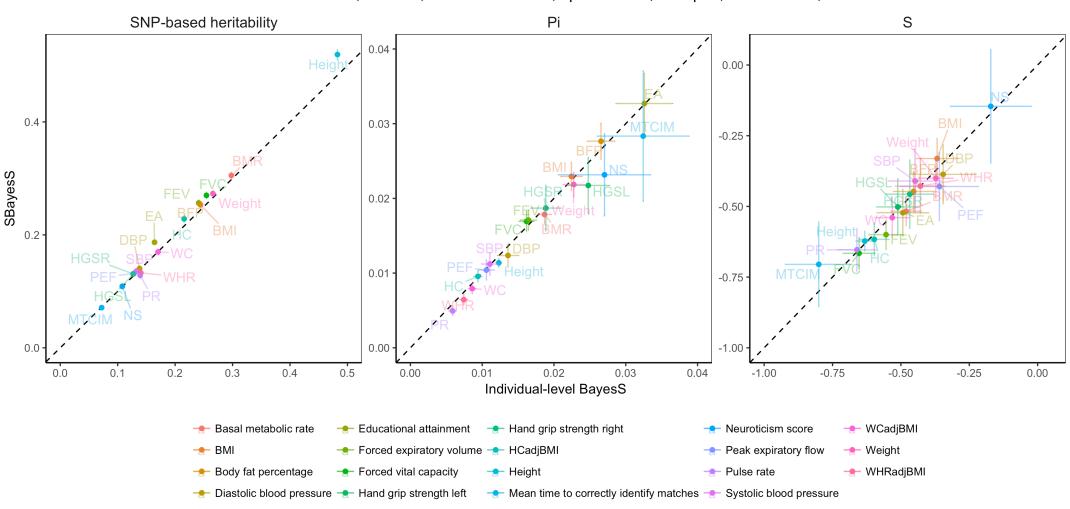


Jian Zeng

## Benchmark SBayesS with BayesS



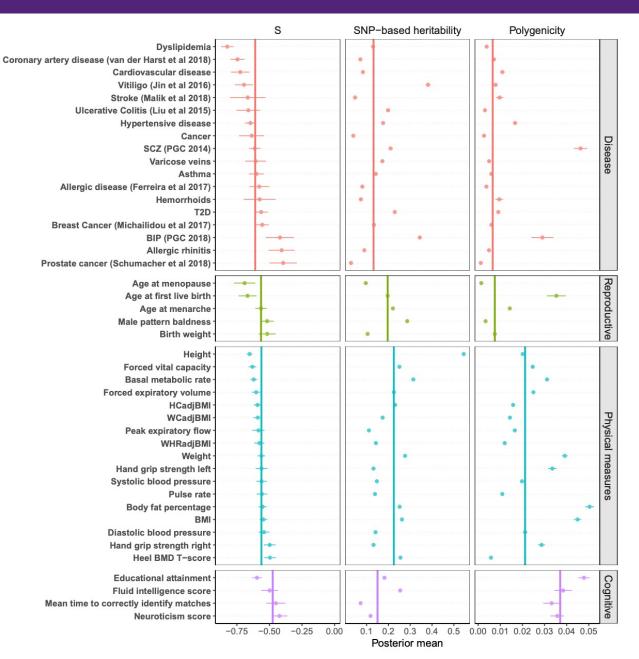


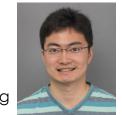


## Genetic architecture of 44 traits



- Full release of UKB + public GWAS summary data (max n = 547k).
- 1.8% of the 1.1 million common HapMap3 SNPs explained 18% of the phenotypic variance.
- The estimate of S was significantly negative (P<0.001) in all the traits analysed.</li>
- Median  $\hat{S} = -0.6$  (SD = 0.1).
- Pervasive action of negative selection on the trait-associated variants.
- Genetic architecture parameters varied across trait categories.





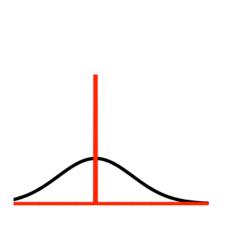
Jian Zeng

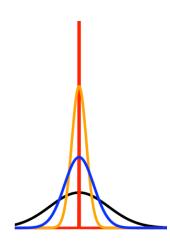
## Estimates depend on..



#### Robust to:

- Robust to LD reference sample and LD filtering (but best to make close in ancestry)
- Robust to Size of LD reference as long as not too small
- Robust to over sampling of cases in case/control studies
- Mostly robust to GWAS sample size (larger sample sizes imply higher polygenicity)
- Mostly robust to modelling of genetic architecture
  - SNP-based heritability is robust
  - S-parameter robust although pattern of differences across traits is changed
  - Polygenicity parameter most sensitive (simulations...)



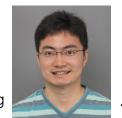


## Parameter relationships under natural selection



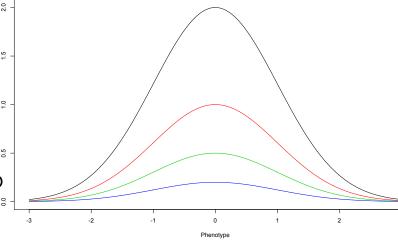
Evolutionary forward simulation (SLiM 3; Haller & Messer 2018)

- 100 MB sequence with a recomb. rate of 1e-5
- Stabilizing selection on phenotypes with different selection strength
  - Selection coefficients sampled from either
    - a normal distribution
    - Mixture of many small and some very large values
- Selection on 10K individuals for 10K generations
- Gravel model for human out-of-Africa evolution (see Ben Haller talk)
- Last generation use two pleiotropic models (Simons or Eyre-Walker) to generate causal effects on focal trait
- Last generation GWAS on unrelated individuals
  - sample size as UKB
  - SNP density same as real data HapMap 3 SNPs, scaled by genome size
- Estimate S, polygenicity  $(\pi)$ , and SNP-based heritability
- Project real traits into the simulation scenarios



**Jian Zeng** 

Individual Fitness =  $\theta *$  dnorm(Phenotype)/dnorm(0), where  $\theta$  is the selection strength



#### Key parameters varied:

- Mean selection coefficient
- Proportion of mutations that can have a causal effect
- Proportion of phenotypic variance attributed to causal mutations

## Simulation results - interdependence

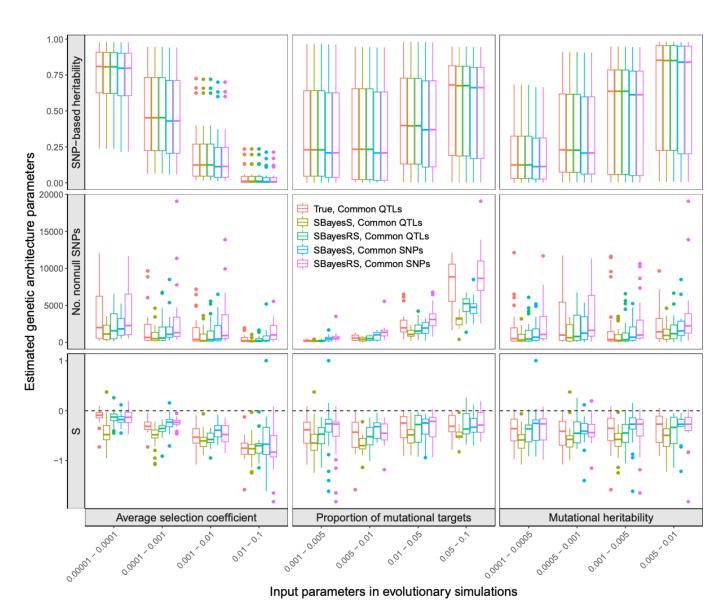


The results were generally consistent regardless of the use of SNPs

- 36k common SNPs or the actual common causal variants
- Genetic architecture estimation method (SBayesS or SBayesRS)
- simulation model (the Simons et al. or Eyre-Walker model),
- underlying distribution of selection coefficients (mixture or normal distribution)

#### Key differences:

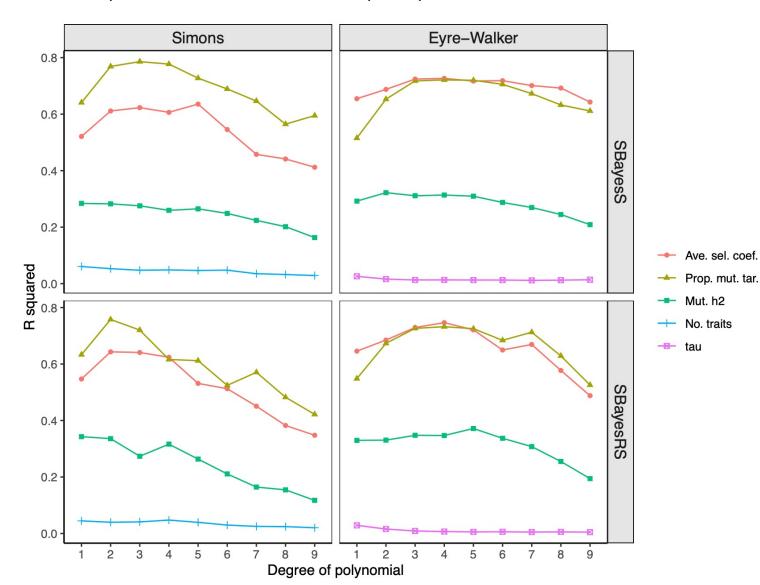
- High s, low h<sup>2</sup>, low pi, low S
- (large effects are purged)



## Prediction – polynomial regression



Can we predict simulation input parameters from the simulation output estimates?



#### Input simulation parameters:

- s-bar
- proportion of genome that are mutational targets
- Mutational heritability

#### Simulation output parameters:

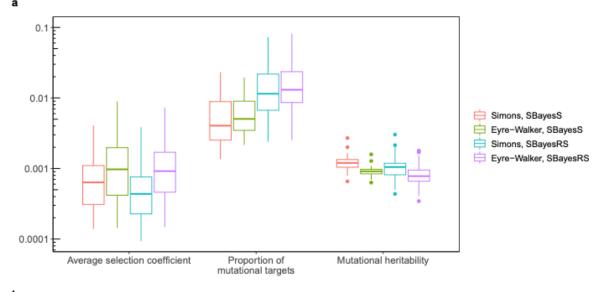
- SNP-based heritability
- Polygenicity parameter
- S coefficient (relationship between allele frequency and effect size

Cross-validation approach

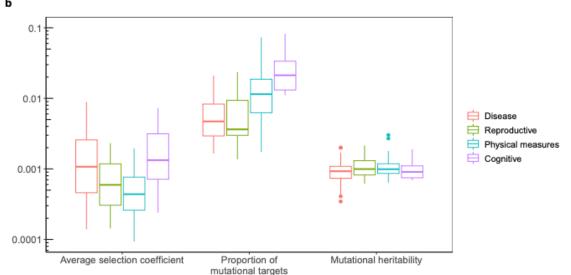
## 44 complex diseases and traits



Apply polynomial prediction equation to parameter values estimated from real data



Reasonably robust to regression polynomial



Stronger selection on disease and cognitive traits?

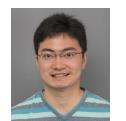
The predicted proportion of mutational targets was ~1% on average across traits = ~30 million base pairs of the human genome were mutational targets for a complex trait.

31

## Conclusions



- The action of negative selection is widespread in the genetic architecture of human complex traits
- The strength of negative selection is relatively strong in most traits
- Interdependence of underlying evolutionary parameter drive estimated parameters
  - Cannot infer selection strength based solely on S; have to also take SNP-based heritability and polygenicity into account
  - $\triangleright$  estimated polygenicity  $\pi$  is driven by the mutational target size and selection strength
    - increased average selection coefficient results in decreased estimated  $\pi$ .
    - negative selection removes causal variants of large effects as well as SNPs in LD with them (i.e.,background selection).
- Cognitive trait associated SNPs are under relatively strong selection
- But selection signals detected in the disease associated SNPs are most likely driven by relatively smaller number of mutational targets
- The large estimates of mutational target size per trait implicate widespread pleiotropy across the genome, another study estimated that 90% of GWAS loci affect multiple traits.





Jian Yang

## **Thanks**





Australian Government

National Health and Medical Research Council









National Institutes of Health

















#### The University of Queensland

Jian Zeng
Jian Yang
Peter Visscher
Luke Lloyd-Jones
Loic Yengo
University of Melbourne
Michael Goddard



