# Evidence for selection using human GWAS data 

Naomi R Wray<br>naomi.wray@uq.edu.au

## PERSPECTIVES

## Polygenic adaptation: a unifying framework to understand positive selection

Neda Barghi( $\mathbb{D}$, Joachim Hermisson (D) and Christian Schlötterer (D)

## 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 pr 66 kg per cow per year. The phenotypic and genetic SD of milk yield in 1957 were $\sim 1200$ and $\sim 600 \mathrm{~kg}$. 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)

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

$$
\begin{gathered}
A_{\text {child }}=\frac{1}{2} A_{\text {dad }}+\frac{1}{2} A_{\text {mum }}+A_{\text {seg }} \\
V\left(A_{\text {seg }}\right)=\frac{1}{2} V(A)
\end{gathered}
$$

Half the genetic variation in a population is
Can more be learned from selection experiments of value in animal breeding programmes? Or is it time for an obituary? w.g. Hill generated by the sampling of genetic material within families


## Common diseases are polygenic



248 risk loci identified at genome-wide significance level
We predict thousands are associated


## Many polygenic genetic architectures




Type 2 Diabetes


Schizophrenia


Crohn's Disease


Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using Bayesian Mixture Model


Mixture component

| $10^{-4} \times \sigma_{g}^{2}$ | small |
| :--- | :--- |
| $10^{-3} \times \sigma_{g}^{2}$ | middle |
| $10^{-2} \times \sigma_{g}^{2}$ | big |

Very,very small Very small small


## Common diseases are complex

Adaptive immunity


HLA locus, PRDM1, BACH2, IL2RA, RASGRP1,

D28, FCGR2A, IFNG, PTPN22, STAT4
e Cell-stress pathways


-
d Inflammation and fibrosis

OSMR, SMAD3, TAB2, IL6ST




MEFV, NLRP7, NLRC4, IL18RAP





Review
Pathway paradigms revealed from the genetics of inflammatory bowel disease
 Received: 5 August 2019
australi
.
(
$\qquad$
r


## All traits are polygenic

The University OF QUEENSLAND

Lessons from Huntington's Disease


https://www.newswise.com/images/uploads/2011/04/14/retrieve.cfm.jpg


Gusella \& MacDonald (2009)

9000 HD cases
~10 GWS Ioci

Article
CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset
( as expected)


## The Ǎtu Jork Eimes



## Observational data for selection in humans


#### Abstract

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 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.

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.

## Signatures of negative selection



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.


## Active area of research

| nature ${ }^{\text {genetics }}$ |  |
| :---: | :---: |
| Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection <br> Antonio F. Pardiñas ${ }^{\bullet}{ }^{1}$, Peter Holmans ${ }^{1}$, Andrew J. Pocklington¹, Valentina Escott-Price ${ }^{1}$, |  |


| ANALYSIS <br> mtery/doiory | nature <br> genetics |
| :---: | :---: |
| Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations |  |
|  |  |
|  |  |
| Steven Gazal ${ }^{1,2 \pi}$, Po-Ru Loh ${ }^{2,3}$, Hilary K. Finucane ${ }^{2,4}$, Andrea Ganna ${ }^{2,5,6}$, Armin Schoech ${ }^{1,2,7}$, ShamilSunyaev ${ }^{2,3,8}$ and Alkes L. Price $\odot^{1,2,7 \star}$ |  |


| ARTICLE |
| :--- |
| Extreme Polygenicity of Complex Traits <br> Is Explained by Negative Selection <br> Luke J. O'Connor <br> and Alkes L. Price ${ }^{1,2, *, *}$${\text { Armin P. Schoech, }{ }^{1} \text { Farhad Hormozdiari, }{ }^{1} \text { Steven Gazal, }{ }^{1} \text { Nick Patterson, }{ }^{3}}$ |



Research
Genome-wide signals of positive selection in human evolution

David Enard, ${ }^{1}$ Philipp W. Messer, and Dmitri A. Petrov ${ }^{1}$
2014


| PERSPECTIVE <br> https://doi.org/10.1038/s41588-019-0383-1 | $\begin{aligned} & \text { nature } \\ & \text { genetics } \end{aligned}$ |
| :---: | :---: |
| Measuring intolerance to mutation in human |  |
| genetics | 2 |
| Zachary L. Fuller® ${ }^{1 *}$, Jeremy J.Be |  |



- Bayesian random regression (BayesS)

$$
\beta_{j}\left\{\begin{array}{lr}
\sim N\left(0,\left[2 p_{j} q_{j}\right]^{S} \sigma_{\beta}^{2}\right), & \pi \\
=0, & 1-\pi
\end{array} \quad\right. \text { Polygenicity }
$$



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


## Simulation based on WGS data

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




## biobank

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 $\mathrm{N}=126 \mathrm{k}$ (unrelated Europeans) for the interim release
~500k Affymetrix SNPs (MAF > 1\%) after QC

Estimated genetic architecture: height vs. BMI
Genetic architecture of 28 traits





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

Use SLIM v2.3 to investigate selection models:

- 10-Mb region
- mutation rate $1.65 \times 10^{-8}$
- new mutations probability
> 0.95 neutral
$>0.05$ 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



## GWAS summary statistics

```
SNP ID
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


## Summary-data-based model

The University OF QUEENSLAND
australia

Consider an individual-data model with a standardised genotype matrix $\mathbf{X}$ :

$$
\mathbf{y}=\mathbf{X} \boldsymbol{\beta}+\mathbf{e}
$$

Multiply both sides by $\frac{1}{N} \mathbf{X}^{\prime}$ gives


- Prior for SNP effect:

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

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

$$
h_{S N P}^{2}=\boldsymbol{\beta}^{\prime} \mathbf{R} \boldsymbol{\beta} / V_{P}
$$

Jian Zeng

## Benchmark SBayesS with BayesS

UKB GWAS=100k, ref=50k, 1.1M HM3 SNPs, sparse LDM, chisq 10, estimate PS, inc MHC

SNP-based heritability


S


-     - Basal metabolic rate
- Educational attainment - - Hand grip strength right
- Neuroticism score $\quad-$ - WCadjBMI
-     - BMI
- Forced expiratory volume - HCadjBMI
-- Peak expiratory flow - Weight
- Body fat percentage
- Forced vital capacity - - Height
-     - Pulse rate
-- WHRadjBMI
- Diastolic blood pressure - - Hand grip strength left - - Mean time to correctly identify matches - - Systolic blood pressure


## Genetic architecłure of 44 traits

- Full release of UKB + public GWAS summary data ( $\max \mathrm{n}=547 \mathrm{k}$ ).
- $1.8 \%$ 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.
- 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.



## 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 le-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


Key parameters varied:

- Mean selection coefficient
- Proportion of mutations that can have a causal effect
- Proportion of phenotypic variance attributed to causal mutations ${ }^{28}$


## 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 h², 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 $\sim 1 \%$ on average across traits $=\sim 30$ million base pairs of the human genome were mutational targets for a complex trait.

## 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
$>$ 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.


## Thanks



NIH \ National Institutes of Health

## FTEHT MNIE. IT TAKES PEOPLE <br> 路该 <br> AutismCRC SALSA Systems Genomics Consortium MINDAUS PARTNERSHIP

## PGC

 Pispl=The University of Queensland

## Program in Complex Jian Zeng

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

