StatGen Workshop
IGSS

Benjamin Neale, Ph.D.
Analytic and Translational Genetics Unit, MGH
Stanley Center for Psychiatric Research & Program in Medical and Population Genetics, Broad Institute

MGH 1811

CENTER FOR GENOMIC MEDICINE

STANLEY CENTER FOR PSYCHIATRIC RESEARCH

VE RI TAS

Analysis of UK Biobank

GWAS of UK Biobank

Download and decryption → Software development → Phenotype wrangling → QC and GWAS

Sam Bryant

Cotton Seed

hail

Andrea Ganna, Duncan Palmer,
Caitlin Carey

Liam Abbott
Dan Howrigan

Also thanks to:

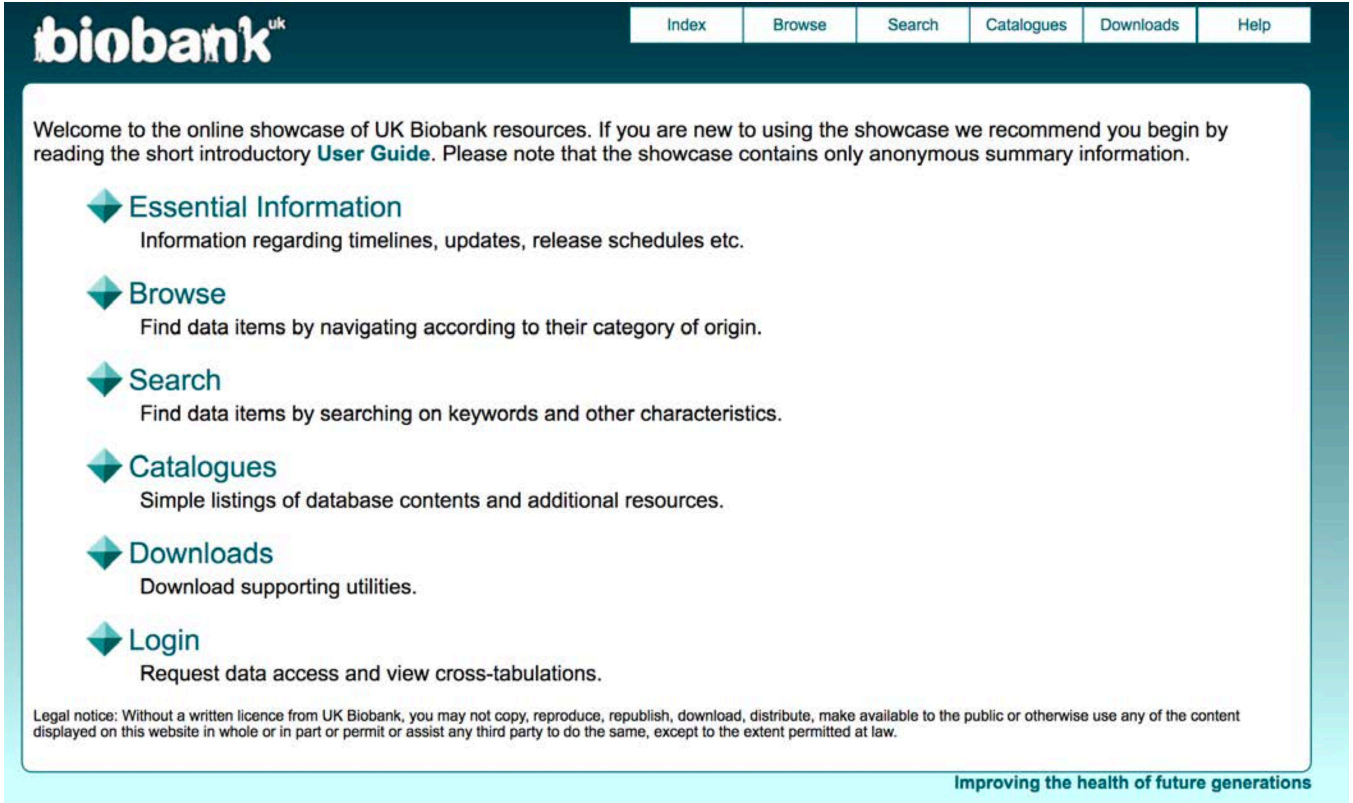
Verneri Anttila
Krishna Aragam
Alex Baumann
Jon Bloom
Joanne Cole
Mark J. Daly
Mark J. Daly
Rob Damien
Steven Gazal
Jackie Goldstein
Mary Haas
Joel Hirschhorn
Eric Jones
Sekar Kathiresan
Dan King
Ruchi Munshi
Tim Poterba
Manuel Rivas
Sailaja Vedantam

biobank uk

Improving the health of future generations

- Follows health and well-being of 500,000 participants
- Genotyped using the Affymetrix Biobank Array
- Lots of phenotypes collected [needs harmonization]
- Lots of opportunity!

Data showcase
http://biobank.ctsu.ox.ac.uk/crystal/



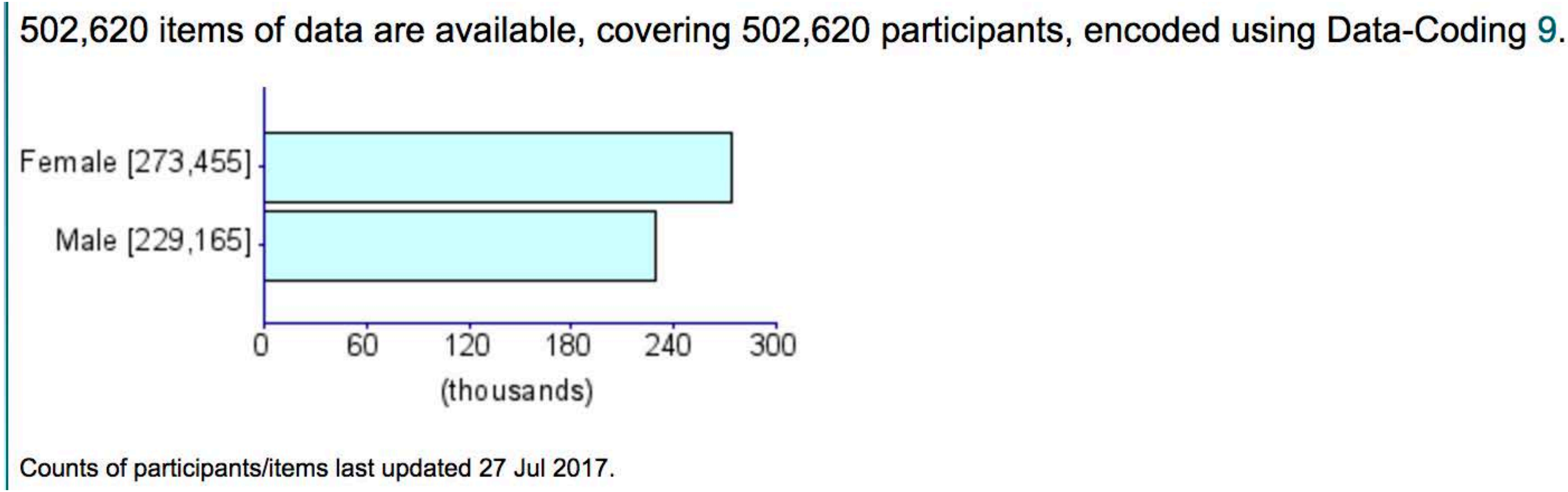
Welcome to the online showcase of UK Biobank resources. If you are new to using the showcase we recommend you begin by reading the short introductory **User Guide**. Please note that the showcase contains only anonymous summary information.

- **Essential Information**
  Information regarding timelines, updates, release schedules etc.
- **Browse**
  Find data items by navigating according to their category of origin.
- **Search**
  Find data items by searching on keywords and other characteristics.
- **Catalogues**
  Simple listings of database contents and additional resources.
- **Downloads**
  Download supporting utilities.
- **Login**
  Request data access and view cross-tabulations.

Legal notice: Without a written licence from UK Biobank, you may not copy, reproduce, republish, download, distribute, make available to the public or otherwise use any of the content displayed on this website in whole or in part or permit or assist any third party to do the same, except to the extent permitted at law.

Improving the health of future generations

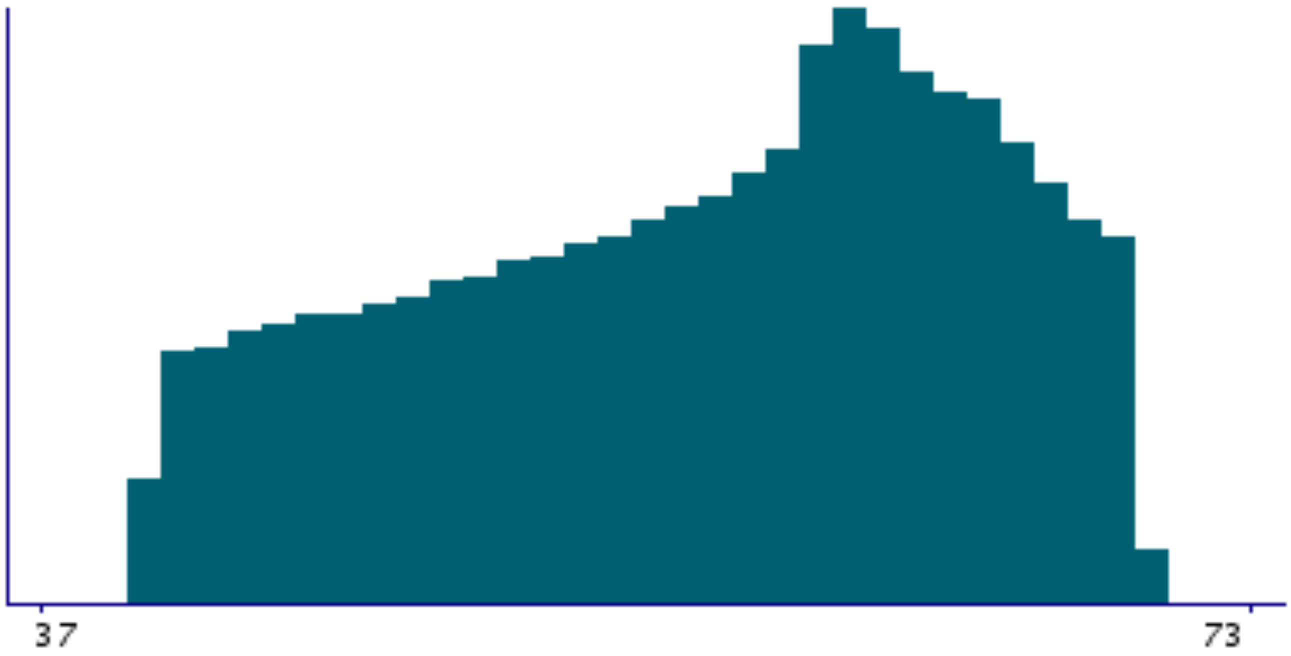
Sex distribution



502,620 items of data are available, covering 502,620 participants, encoded using Data-Coding 9.

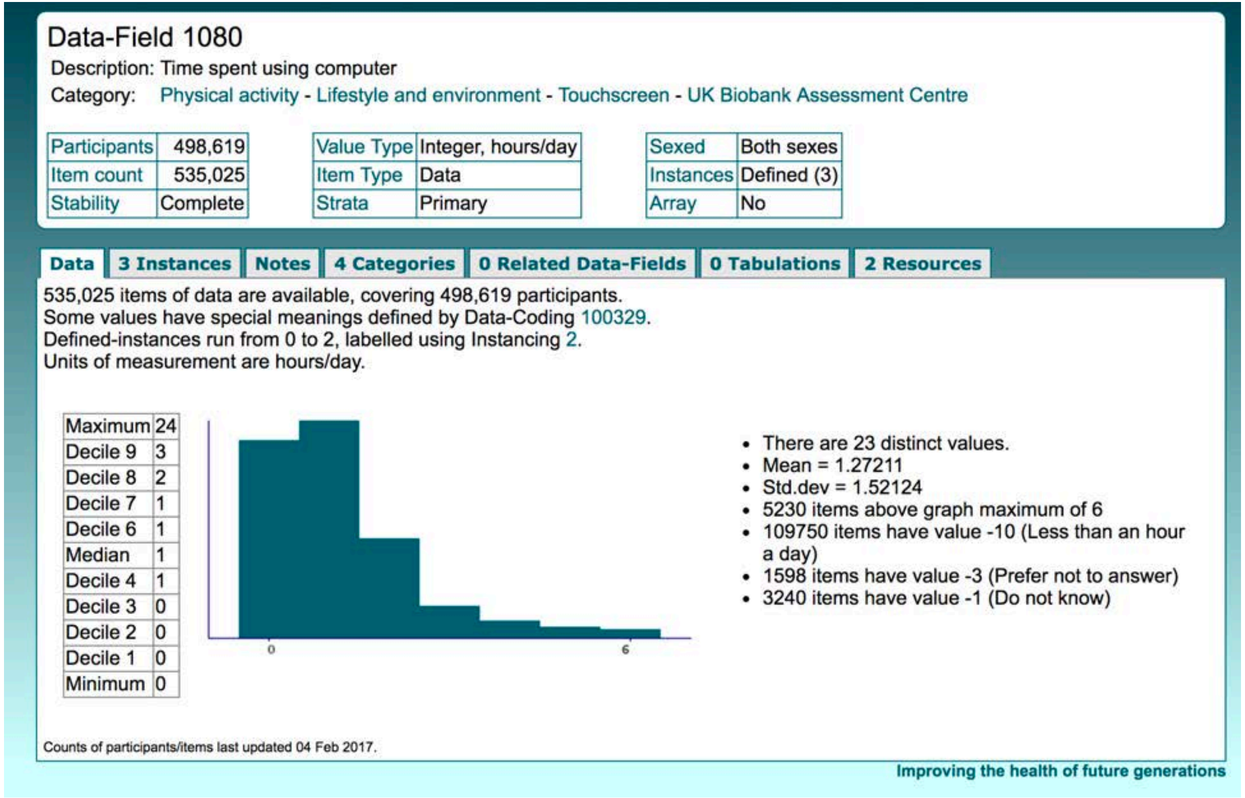
Counts of participants/items last updated 27 Jul 2017.

Age distribution at recruitment



Mean = 56.5286
Std.dev = 8.09516

Example self-report



Data-Field 1080

Description: Time spent using computer

Category: Physical activity - Lifestyle and environment - Touchscreen - UK Biobank Assessment Centre

Participants	498,619
Item count	535,025
Stability	Complete

Value Type	Integer, hours/day
Item Type	Data
Strata	Primary

Sexed	Both sexes
Instances	Defined (3)
Array	No

535,025 items of data are available, covering 498,619 participants.
Some values have special meanings defined by Data-Coding 100329.
Defined-instances run from 0 to 2, labelled using Instancing 2.
Units of measurement are hours/day.

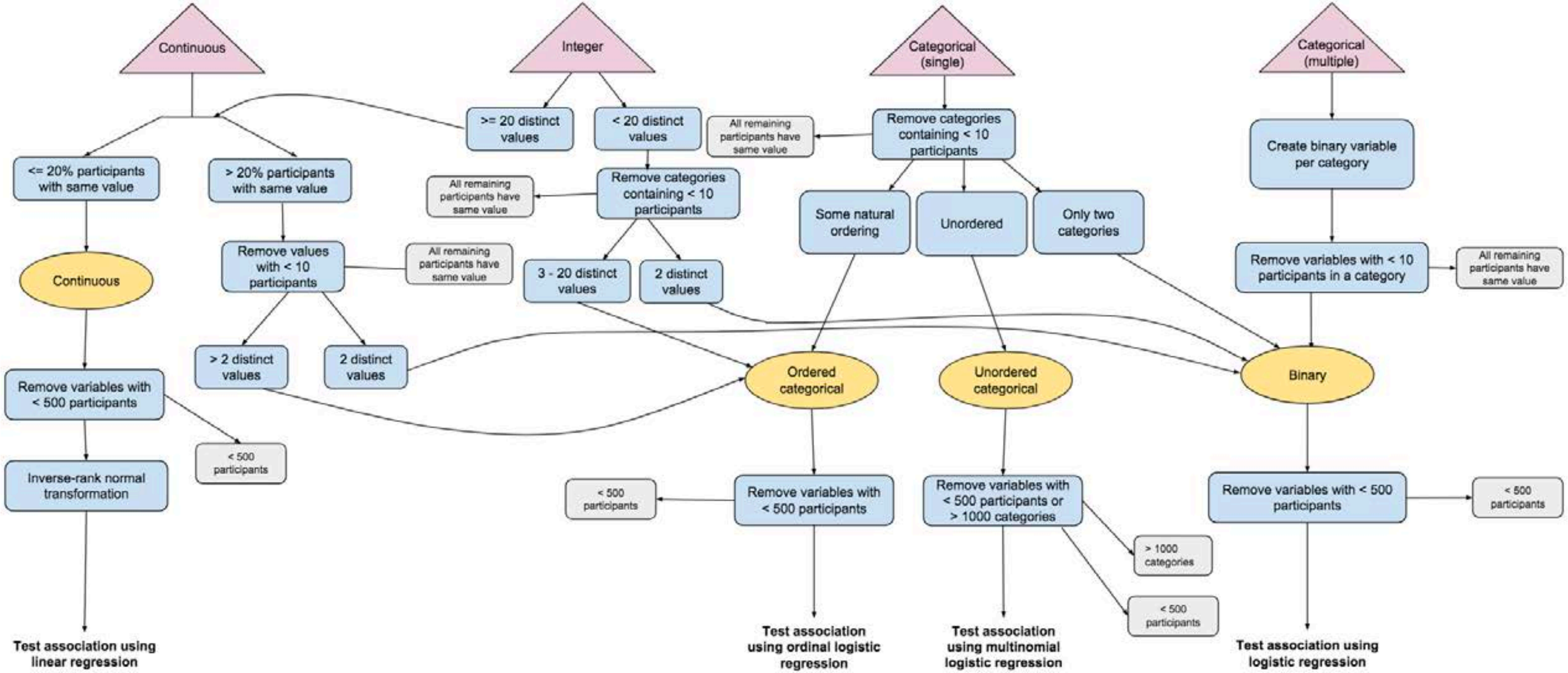
Maximum	24
Decile 9	3
Decile 8	2
Decile 7	1
Decile 6	1
Median	1
Decile 4	1
Decile 3	0
Decile 2	0
Decile 1	0
Minimum	0

- There are 23 distinct values.
- Mean = 1.27211
- Std.dev = 1.52124
- 5230 items above graph maximum of 6
- 109750 items have value -10 (Less than an hour a day)
- 1598 items have value -3 (Prefer not to answer)
- 3240 items have value -1 (Do not know)

Counts of participants/items last updated 04 Feb 2017.

Improving the health of future generations

PHESANT!



Continuous

Integer

Categorical (single)

Categorical (multiple)

>= 20 distinct values

< 20 distinct values

All remaining participants have same value

Remove categories containing < 10 participants

Create binary variable per category

<= 20% participants with same value

> 20% participants with same value

All remaining participants have same value

Remove categories containing < 10 participants

Some natural ordering

Unordered

Only two categories

Continuous

Remove values with < 10 participants

All remaining participants have same value

3 - 20 distinct values

2 distinct values

Remove variables with < 10 participants in a category

All remaining participants have same value

> 2 distinct values

2 distinct values

Ordered categorical

Unordered categorical

Binary

Remove variables with < 500 participants

< 500 participants

Inverse-rank normal transformation

< 500 participants

Remove variables with < 500 participants

Remove variables with < 500 participants or > 1000 categories

Remove variables with < 500 participants

< 500 participants

> 1000 categories

< 500 participants

Test association using linear regression

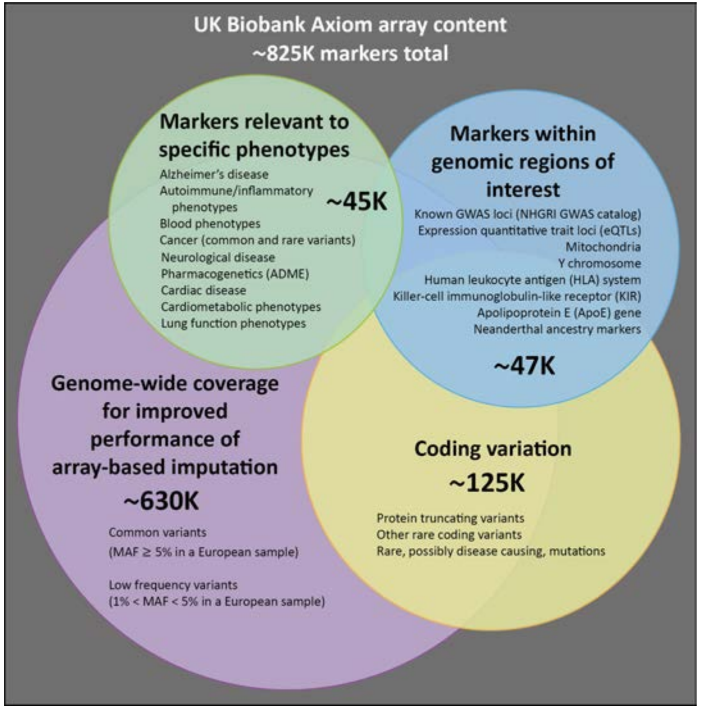
Test association using ordinal logistic regression

Test association using multinomial logistic regression

Test association using logistic regression

Copious thanks to Millard LAC, Davies NM, Gaunt TR, Davey Smith G, Tilling K. PHESANT: a tool for performing automated phenome scans in UK Biobank. bioRxiv (2017)

What’s on the array?



UK Biobank Axiom array content
~825K markers total

Markers relevant to specific phenotypes

~45K

Alzheimer’s disease
Autoimmune/inflammatory phenotypes
Blood phenotypes
Cancer (common and rare variants)
Neurological disease
Pharmacogenetics (ADME)
Cardiac disease
Cardiometabolic phenotypes
Lung function phenotypes

Markers within genomic regions of interest

Known GWAS loci (NHGRI GWAS catalog)
Expression quantitative trait loci (eQTLs)
Mitochondria
Y chromosome
Human leukocyte antigen (HLA) system
Killer-cell immunoglobulin-like receptor (KIR)
Apolipoprotein E (ApoE) gene
Neanderthal ancestry markers

~47K

Genome-wide coverage for improved performance of array-based imputation

~630K

Common variants
(MAF ≥ 5% in a European sample)

Low frequency variants
(1% < MAF < 5% in a European sample)

Coding variation

~125K

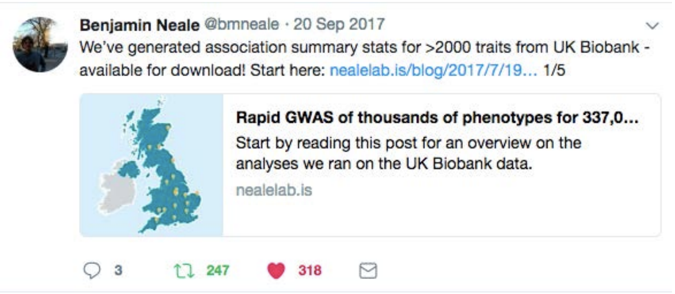
Protein truncating variants
Other rare coding variants
Rare, possibly disease causing, mutations

Imputed to HRC

Round 1 GWAS

- Last fall, the Neale lab…
  - GWASed 2,419 phenotypes
    - Blogged about it
    - Put them on dropbox
      - And people made browsers
  - Estimated h^2 for all of them
  - Made an h^2 browser
    - Blogged about that too

Nealelab.is/blog



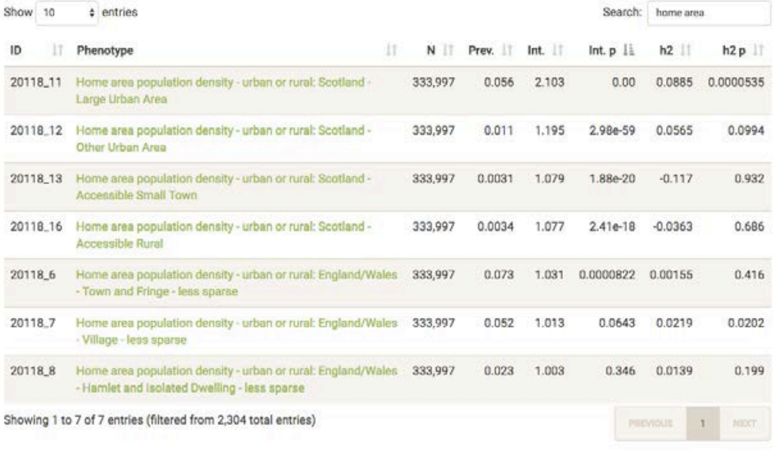
Benjamin Neale @bmneale · 20 Sep 2017

We've generated association summary stats for >2000 traits from UK Biobank - available for download! Start here: nealelab.is/blog/2017/7/19… 1/5

Rapid GWAS of thousands of phenotypes for 337,0…

Start by reading this post for an overview on the analyses we ran on the UK Biobank data.

nealelab.is



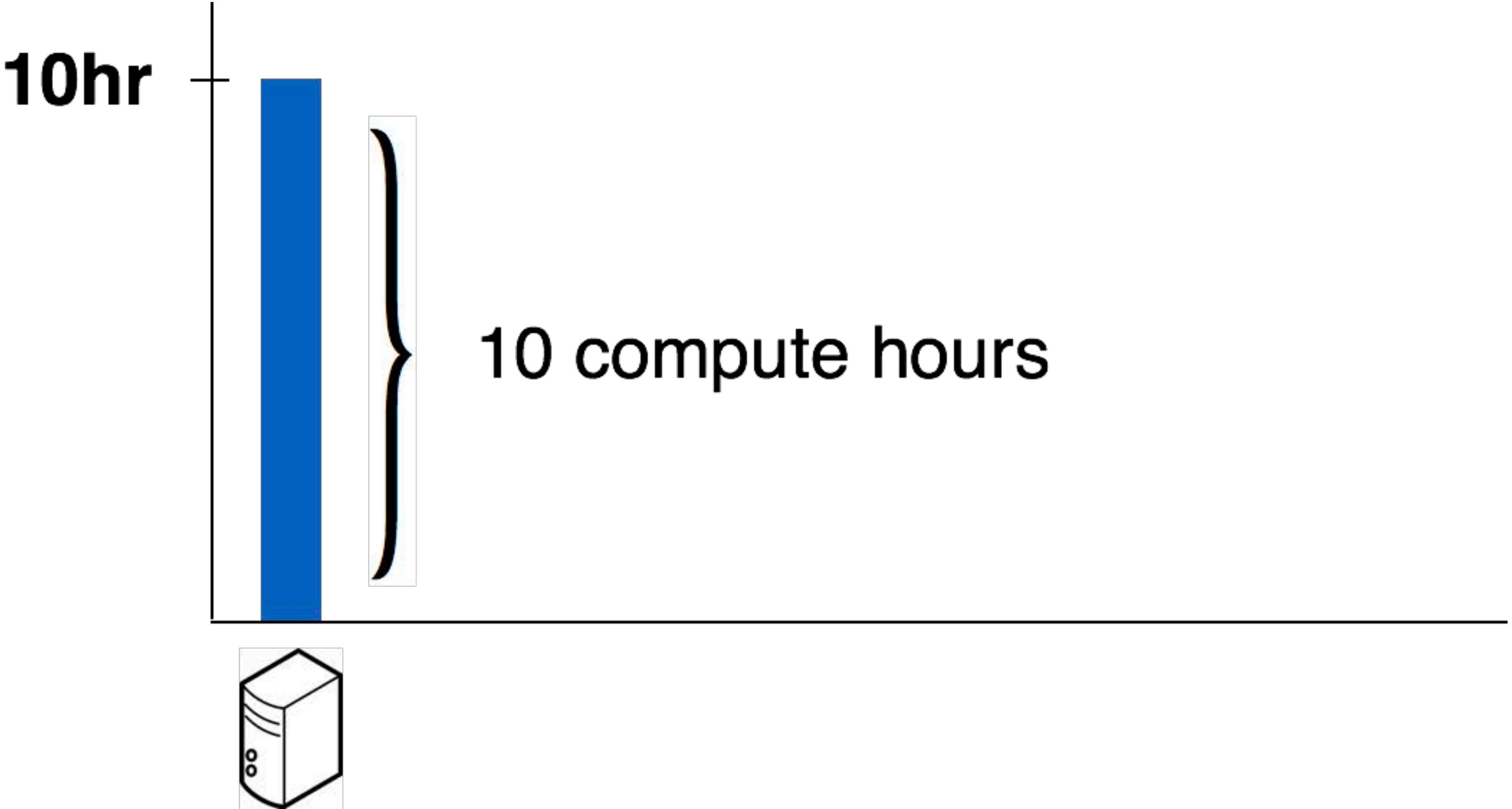
Show 10 entries

Search: home area

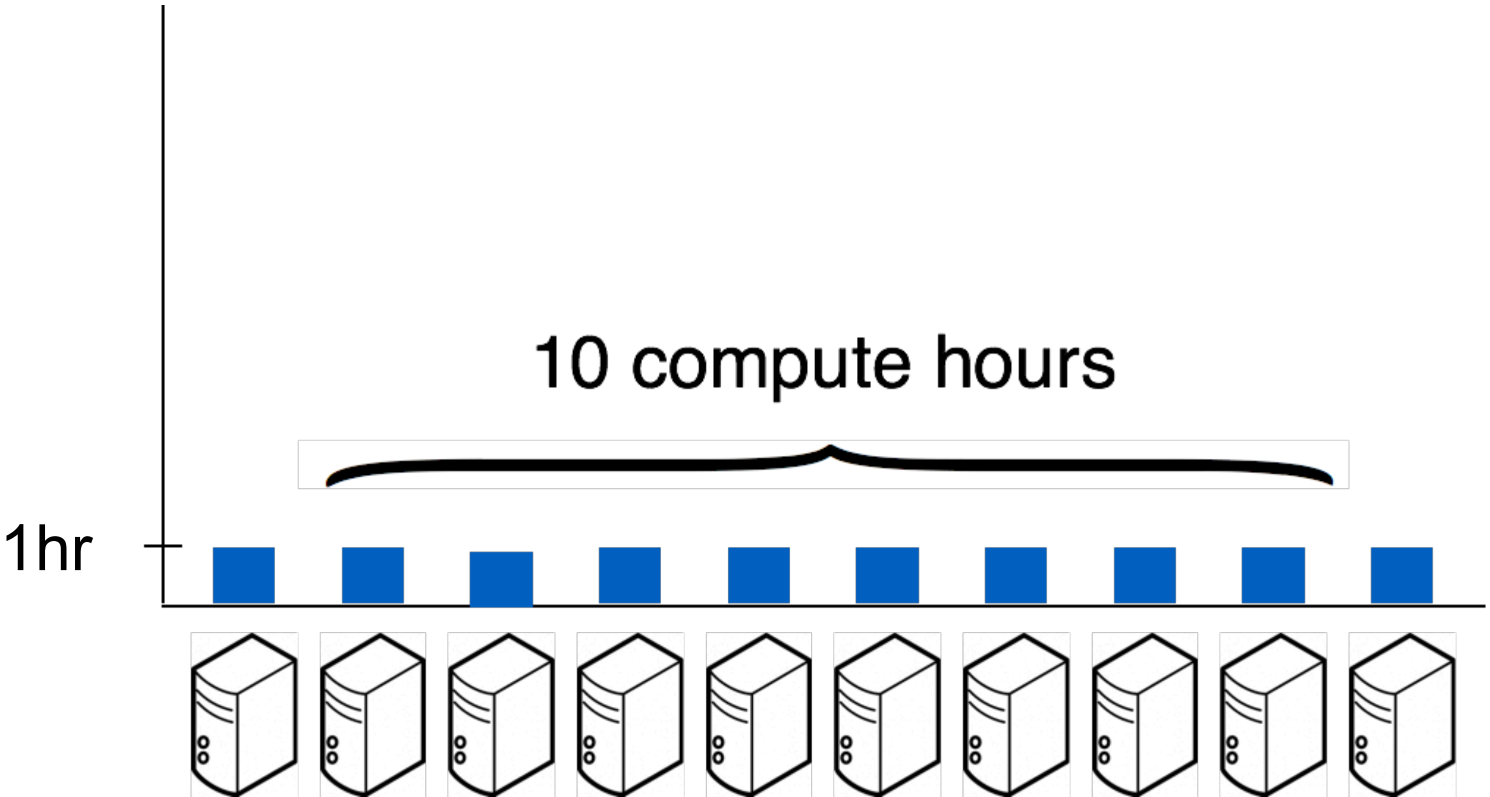
ID	Phenotype	N	Prev.	Int.	Int. p	h2	h2 p
20118_11	Home area population density - urban or rural: Scotland - Large Urban Area	333,997	0.056	2.103	0.00	0.0885	0.0000535
20118_12	Home area population density - urban or rural: Scotland - Other Urban Area	333,997	0.011	1.195	2.98e-59	0.0565	0.0994
20118_13	Home area population density - urban or rural: Scotland - Accessible Small Town	333,997	0.0031	1.079	1.88e-20	-0.117	0.932
20118_16	Home area population density - urban or rural: Scotland - Accessible Rural	333,997	0.0034	1.077	2.41e-18	-0.0363	0.686
20118_6	Home area population density - urban or rural: England/Wales - Town and Fringe - less sparse	333,997	0.073	1.031	0.0000822	0.00155	0.416
20118_7	Home area population density - urban or rural: England/Wales - Village - less sparse	333,997	0.052	1.013	0.0643	0.0219	0.0202
20118_8	Home area population density - urban or rural: England/Wales - Hamlet and Isolated Dwelling - less sparse	333,997	0.023	1.003	0.346	0.0139	0.199

Showing 1 to 7 of 7 entries (filtered from 2,304 total entries)

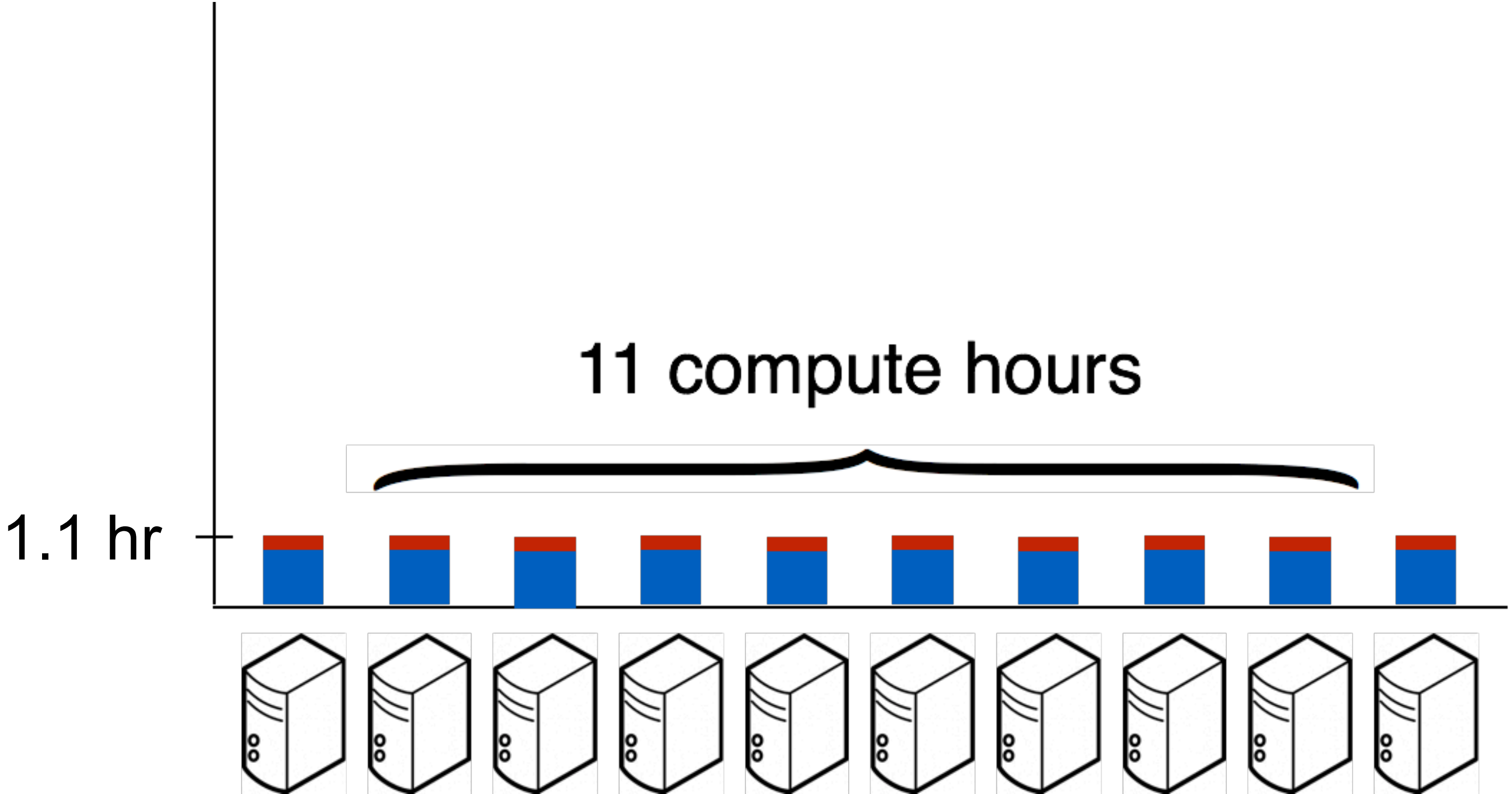
Scalability



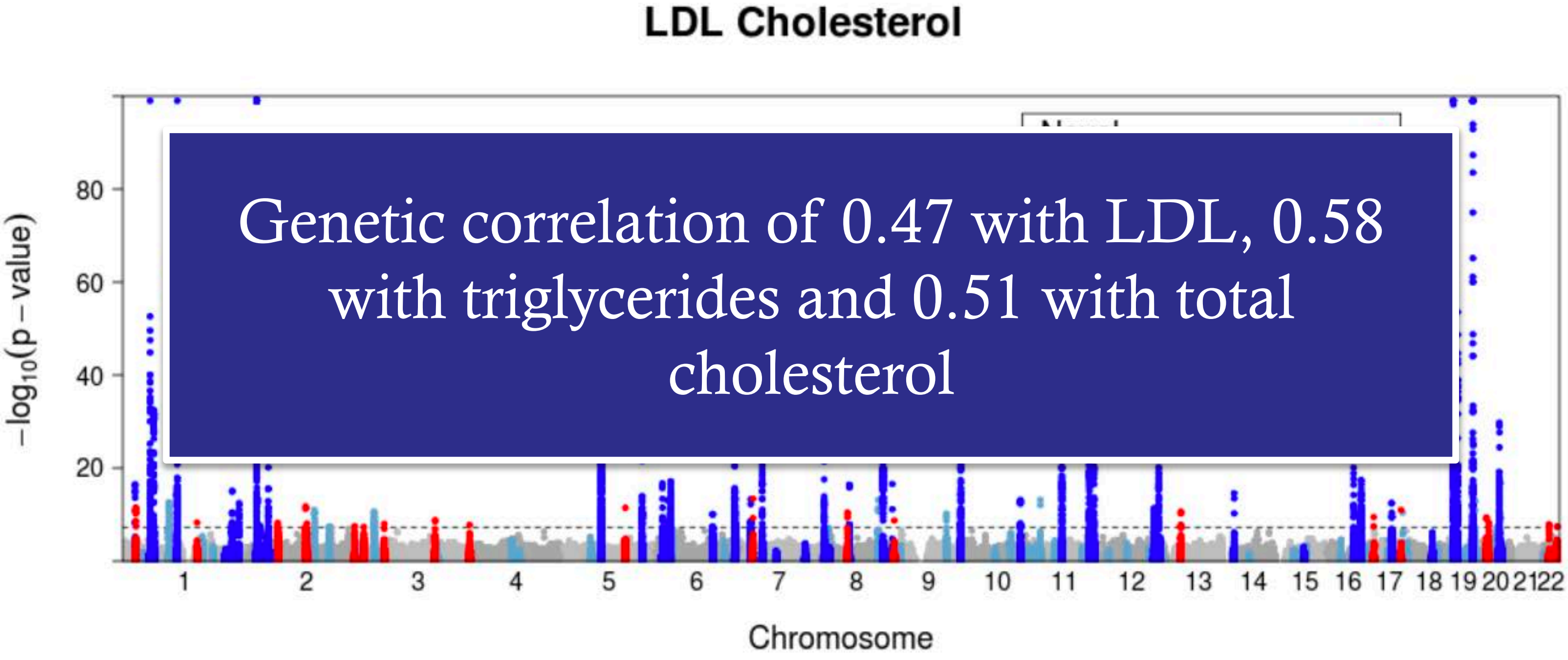
Scalability



Scalability



Association results for many things!
Taking cholesterol lowering meds



Genetic correlation of 0.47 with LDL, 0.58 with triglycerides and 0.51 with total cholesterol

GLGC Nat Genet

6 months later, we did it all again

Why Round 2 of UKB GWAS?

- Missing a batch of imputed SNPs
  - Corrected data released in March

- Hadn’t gotten permissions for *all* the phenotypes
  - Expanded UKB application

- Feedback on improvements for the GWAS
  - Age, sex, stratification

Round 2: QC Updates

- Variant QC:
  - Added the new imputed data
  - Added chrX variants
  - Added VEP missense and PTVs with MAF > 1e-6
  - Net: 3 million more variants
    - 13.8 million total

- Sample QC:
  - Relaxed restriction to “white British” samples

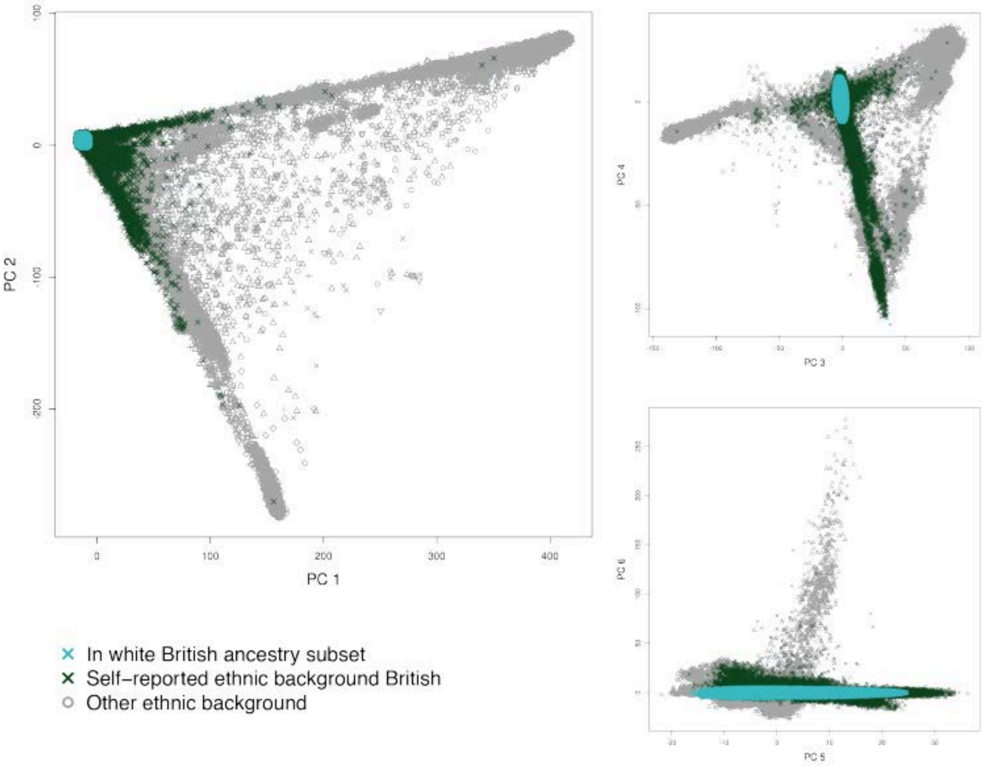
How “White British” is defined

- What is your ethnic group?
  - **White**
  - Mixed
  - Asian or Asian British
  - Black or Black British
  - Chinese
  - Other ethnic group
  - Do not know
  - Prefer not to answer

- What is your ethnic background?
  - **British**
  - Irish
  - Any other white background
  - Prefer not to answer

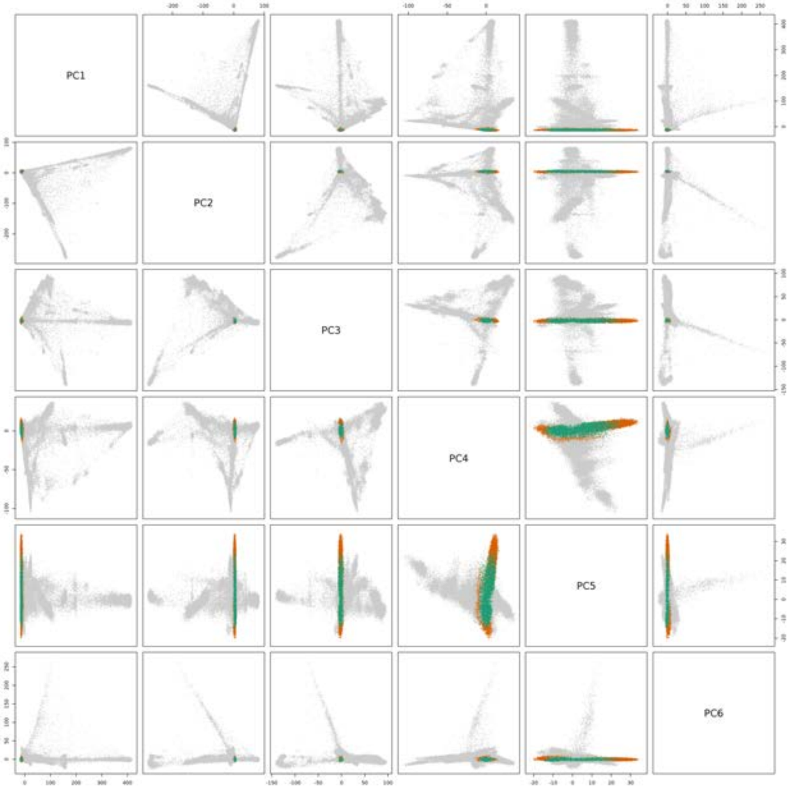
- Don’t be defined as a PCA outlier
  - Bayesian outlier detection algorithm on PCs 1&2, 3&4, and 5&6

How “White British” is defined



Widening out definition of Europeans

- Get mean and SD of top 6 PCs among the “white British”
- Draw ellipse in PCA space with radius of 7 SDs along each PC axis
  - Provides good predictive accuracy for self-reporting “White” vs. other ethnicities
- Discard any self-reported as non-white
- Final N (after QC): 361,194
  - Previously 337,199

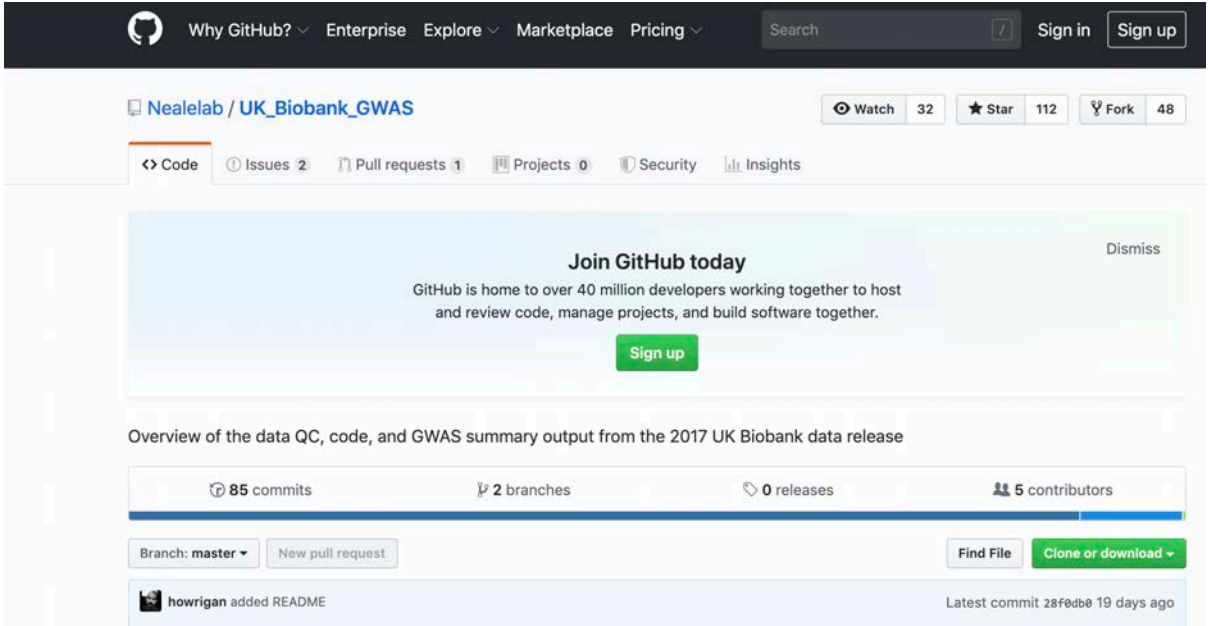


Round 2: GWAS Changes

- Add age, age^2, sex*age, and sex*age^2 as covariates
- Increase number of PC covariates from 10 to 20
- Compute PCs within the GWAS sample rather than using the PCs computed by UKB on the full sample
- In addition to main GWAS, run sex-specific GWAS [withou sex covariates

Let’s go to the code!

https://github.com/Nealelab/UK_Biobank_GWAS



Join GitHub today

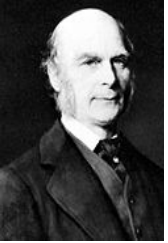
GitHub is home to over 40 million developers working together to host and review code, manage projects, and build software together.

Overview of the data QC, code, and GWAS summary output from the 2017 UK Biobank data release

85 commits · 2 branches · 0 releases · 5 contributors

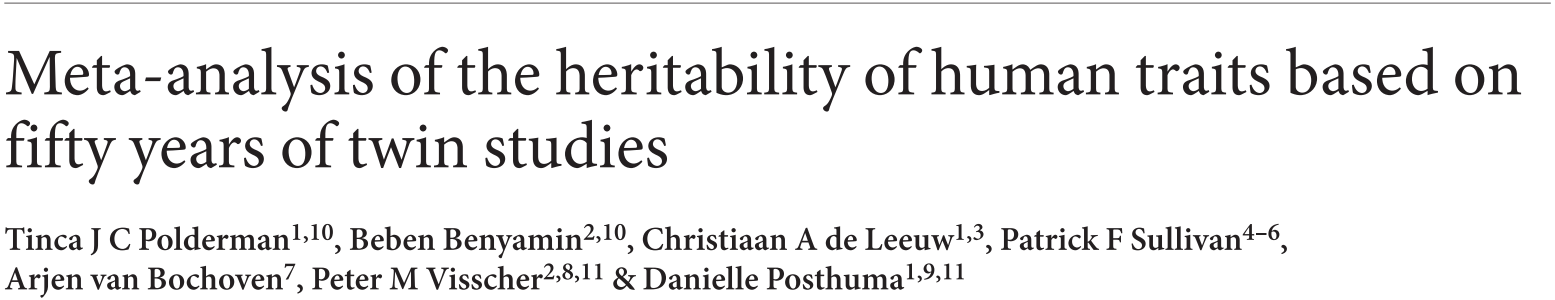
howrigan added README · Latest commit 28f0db0 19 days ago

We’ll start with the readme



Francis Galton
Twin and family studies

- Relatives are more similar

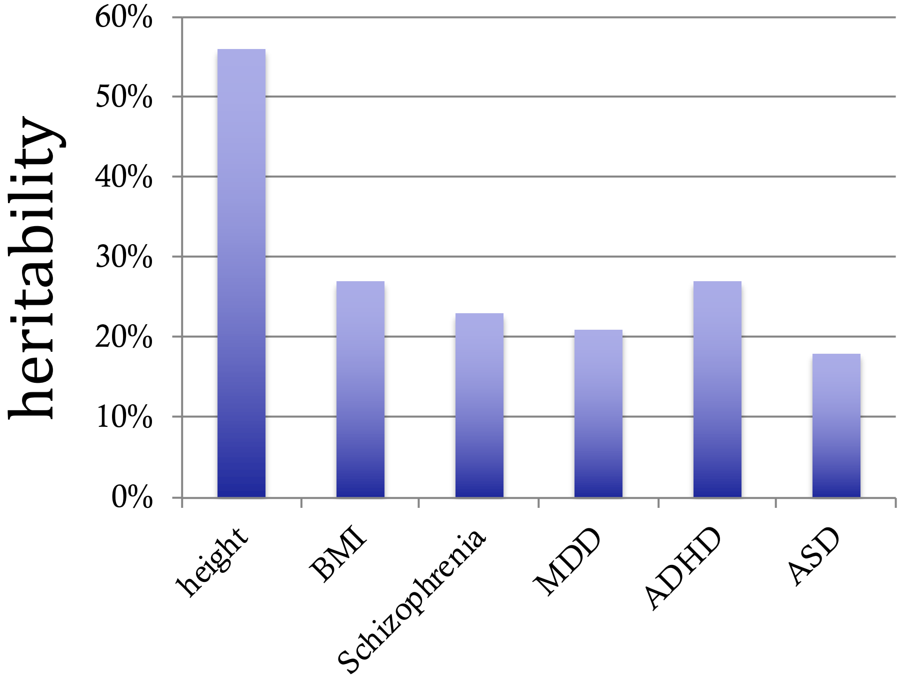


Meta-analysis of the heritability of human traits based on fifty years of twin studies

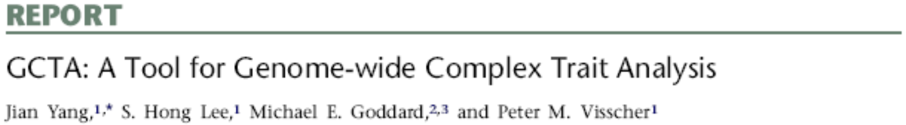
Tinca J C Polderman[1,10], Beben Benyamin[2,10], Christiaan A de Leeuw[1,3], Patrick F Sullivan[4–6], Arjen van Bochoven[7], Peter M Visscher[2,8,11] & Danielle Posthuma[1,9,11]

Average estimate of heritability 49%
69% of twin studies support a purely additive genetic model

GREML/GCTA



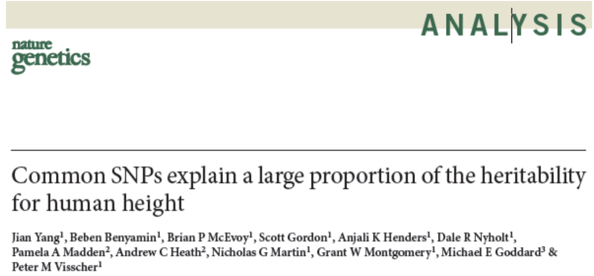
- Use estimated genetic similarity



REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

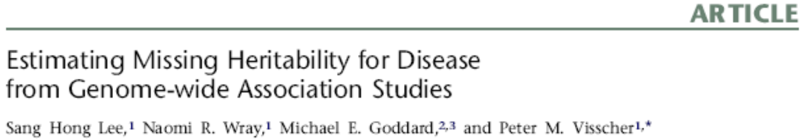


ANALYSIS

nature genetics

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]



ARTICLE

Estimating Missing Heritability for Disease from Genome-wide Association Studies

Sang Hong Lee,[1] Naomi R. Wray,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1,*]

LD Score regression

With thanks

Brendan Bulik-Sullivan

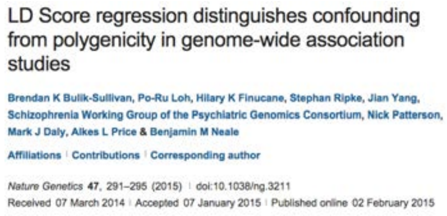
Hilary Finucane

Po-Ru Loh

Mark Daly

Alkes Price

How does LD shape association?



LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics **47**, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015

How does LD shape association?

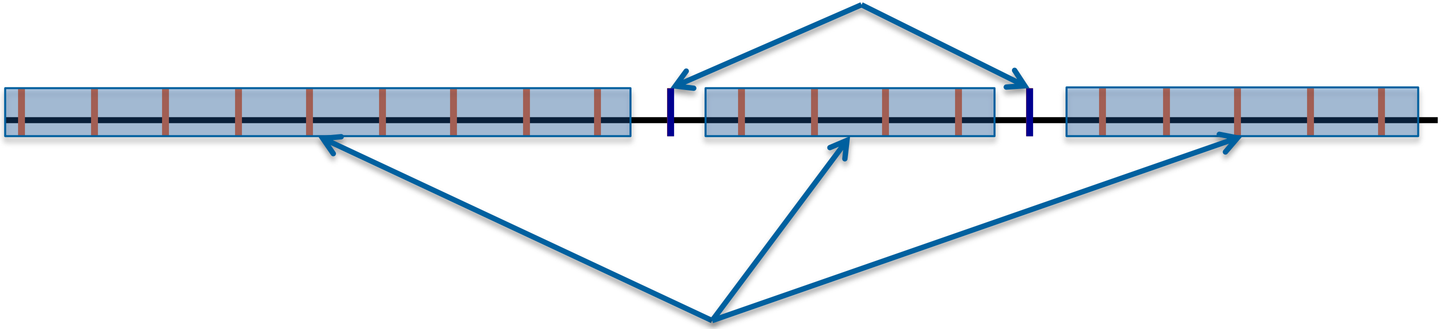
LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics **47**, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015

Lonely SNPs [no LD]



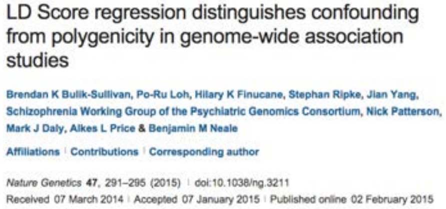
LD blocks

How does LD shape association?

| Lonely SNPs [no LD]

▨ LD blocks

* Causal variants



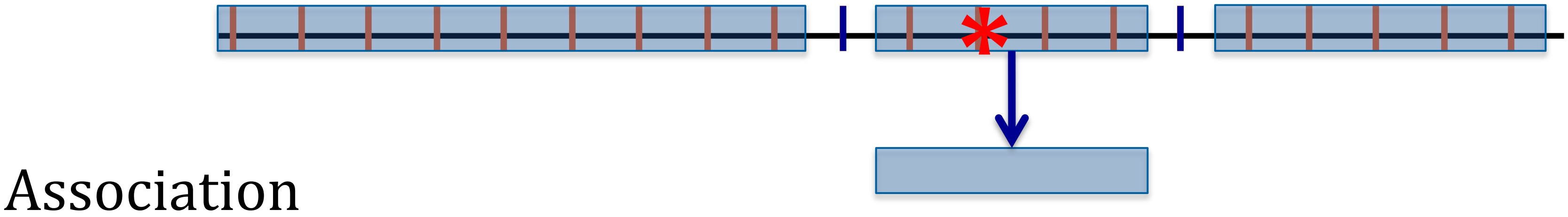
LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics **47**, 291–295 (2015) | doi:10.1038/ng.3211

Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015



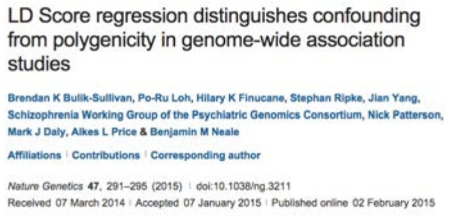
All markers correlated with a causal variant show association

How does LD shape association?

| Lonely SNPs [no LD]

🟦 LD blocks

* Causal variants

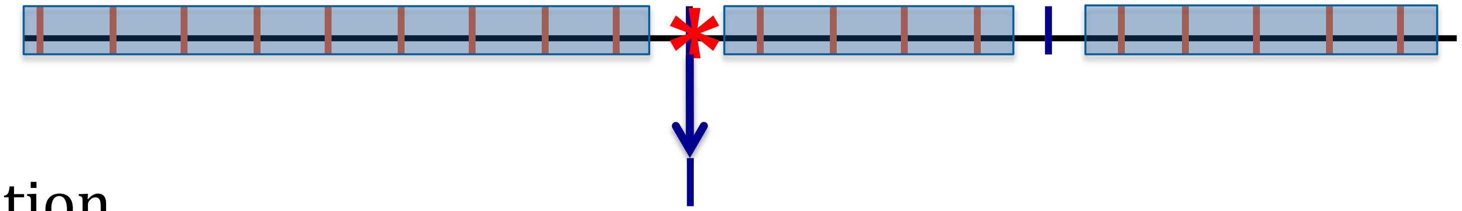


LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics 47, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015



Association

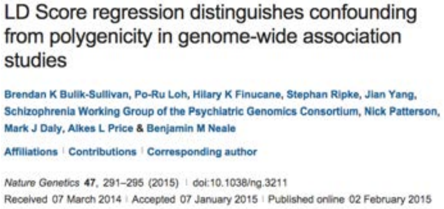
Lonely SNPs only show association if they are causal

What happens under polygenicity?

| Lonely SNPs [no LD]

▨ LD blocks

* Causal variants

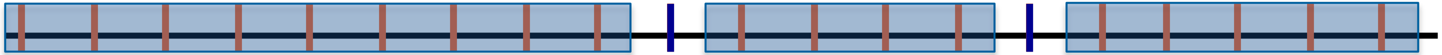


LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics **47**, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015

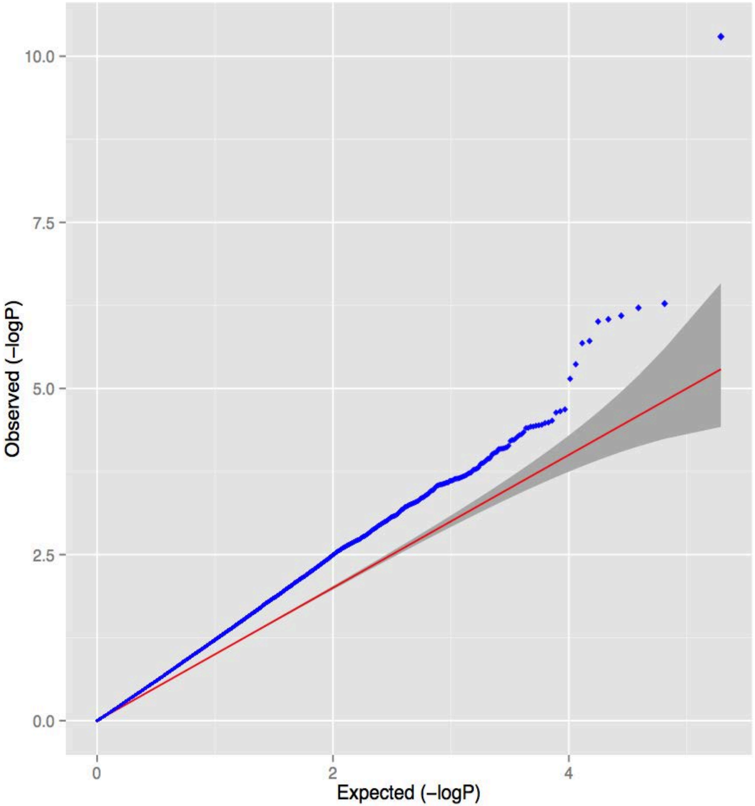


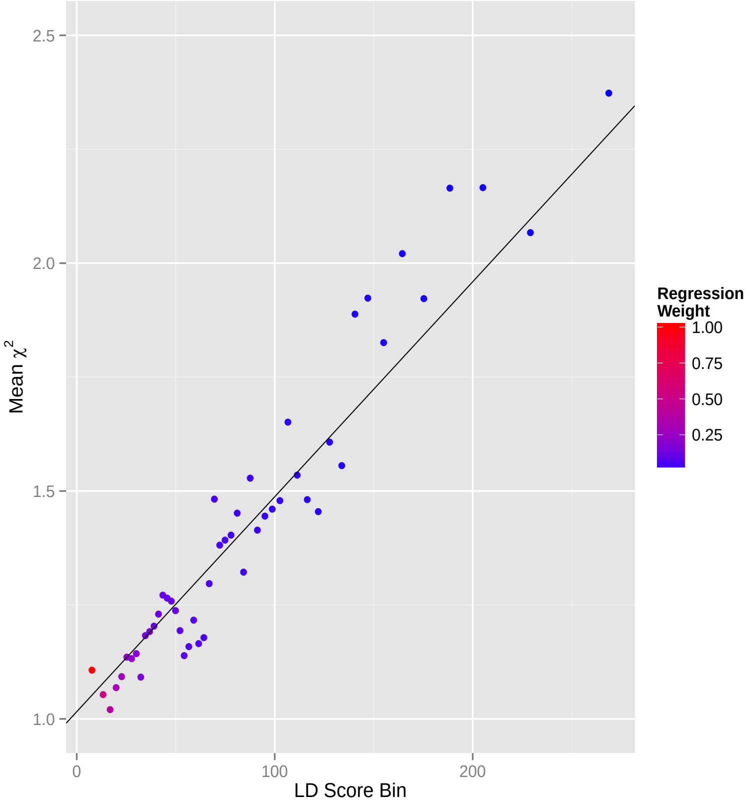
Assuming a uniform prior, we see SNPs with more LD friends showing more association

The more you tag, the more likely you are to tag a causal variant

Simulated polygenic architecture

Lambda = 1.30 LD score intercept = 1.02



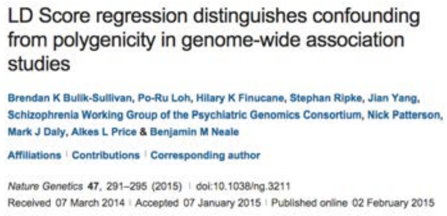


What happens under stratification?

| Lonely SNPs [no LD]

■ LD blocks

* Causal variants

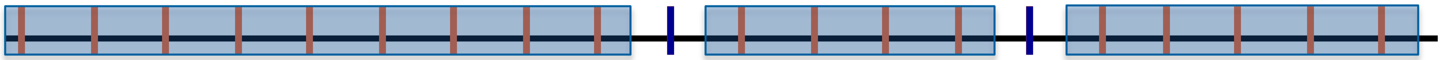


LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

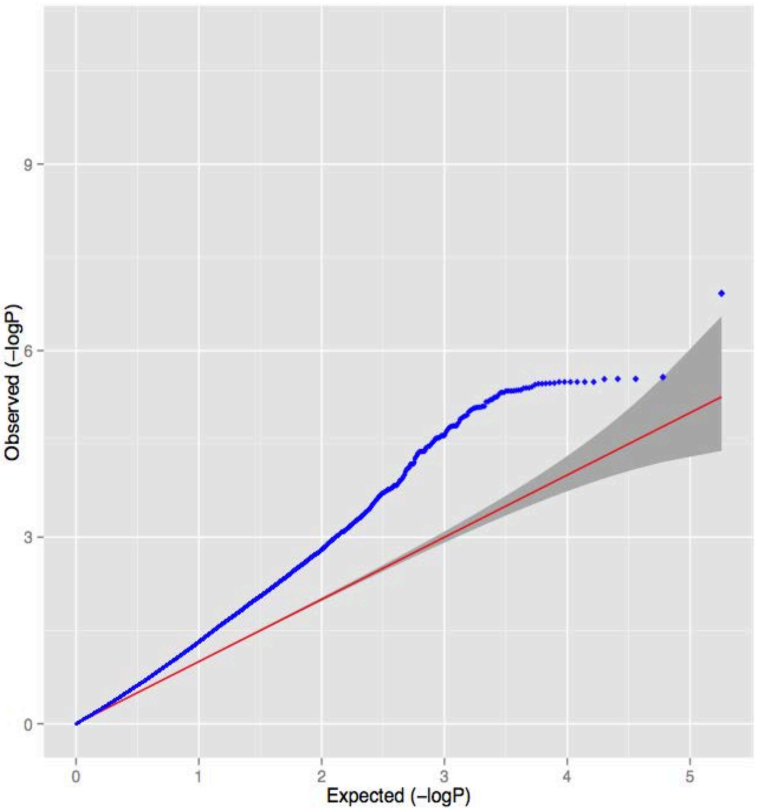
Nature Genetics **47**, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015

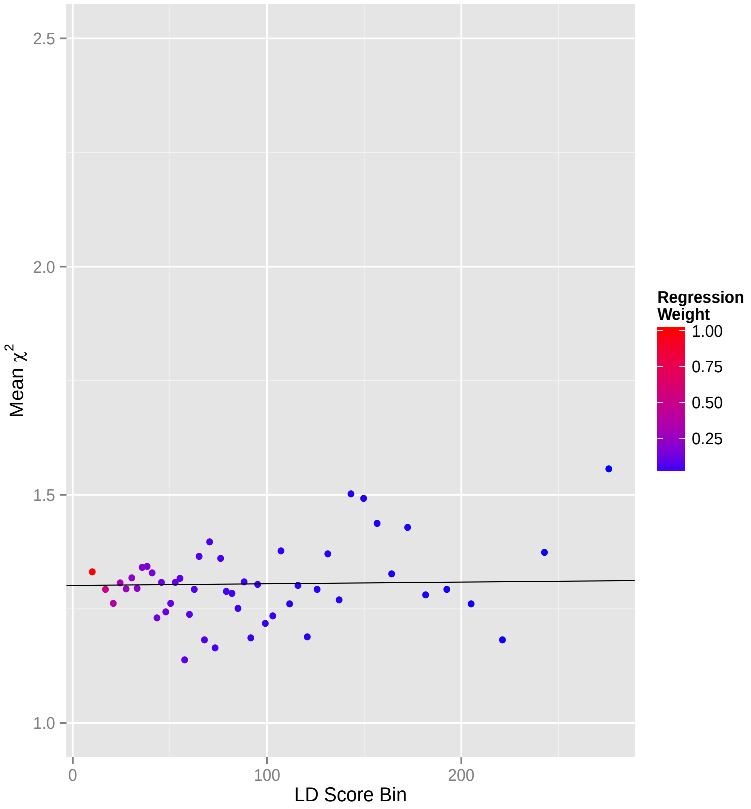


Under pure drift we expect LD to have no relationship to differences in allele frequencies between populations

UK controls versus Sweden controls

Lambda = 1.30 LD score intercept = 1.32





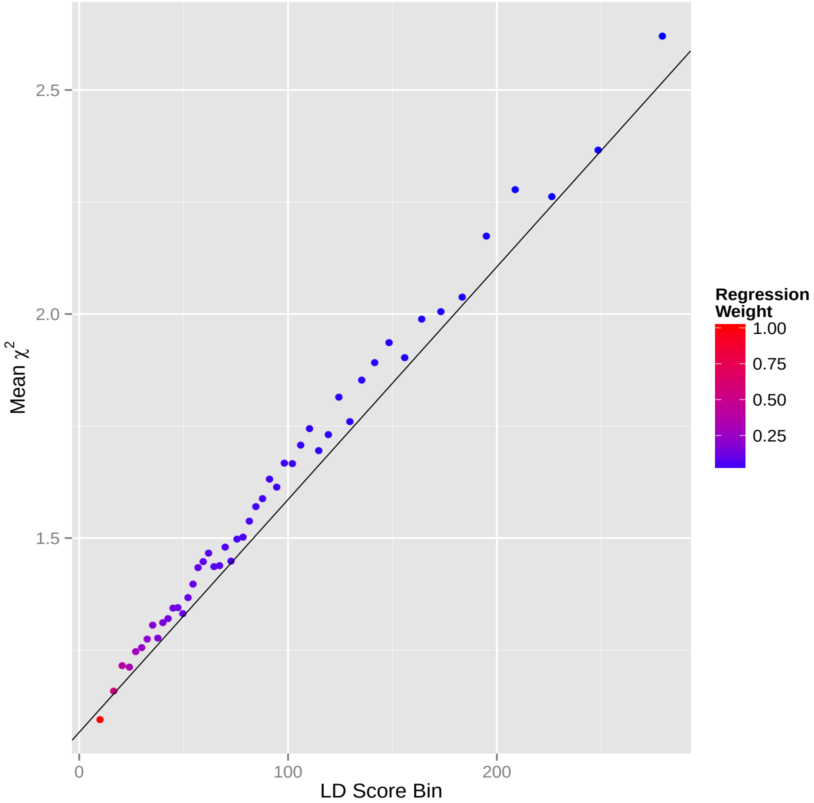
PGC Schizophrenia

Lambda = 1.48

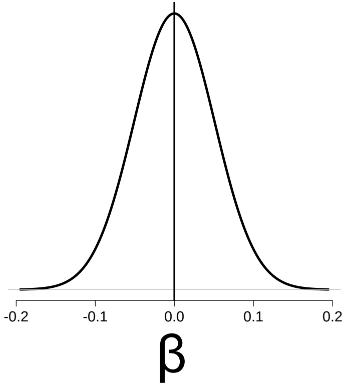
Intercept = 1.06

Slope *p*-value < 10^{-300}

Overwhelming majority of inflation is consistent with polygenic architecture



LD Score regression



Draw polygenic effects from N (0, n/m^2), var =

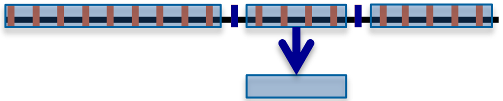
What is the E[χ^2] for variant *j*?

$$E\left[\chi_j^2\right] = 1 + Na + \frac{\boxed{h_g^2}N}{M} l_j$$

New estimator of heritability

where N=sample size, M=# of SNPs, a=inflation due to confounding, h^{2}g is heritability (total obs.) and *l*$_j$ is the *LD Score*

$$l_j = \sum\nolimits_{k \neq j} r_{jk}^2$$



Bulik-Sullivan et al. Nature Genetics 2015
Yang et al. EJHG 2011

9,928 GWAS later… let’s talk *h*^{2} using LD score regression

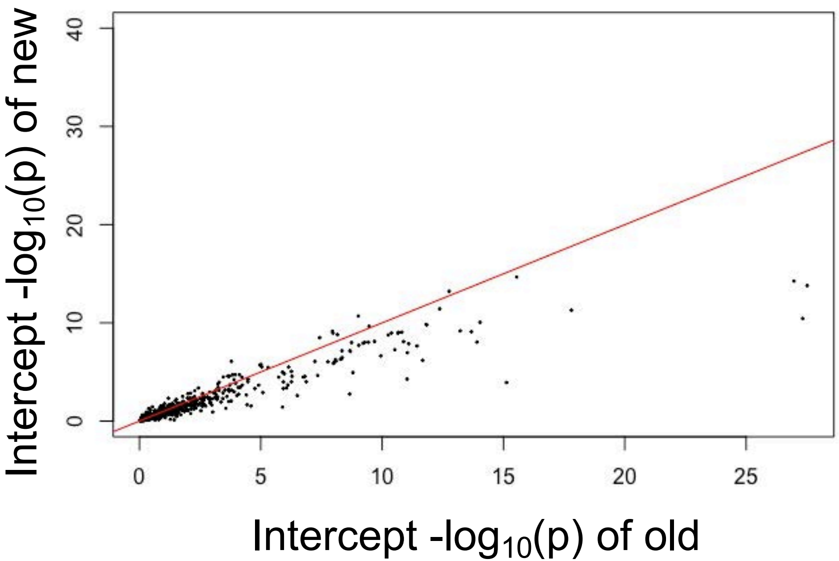
$$E\left[\chi_j^2\right] = 1 + Na + \frac{h_g^2 N}{M} l_j$$

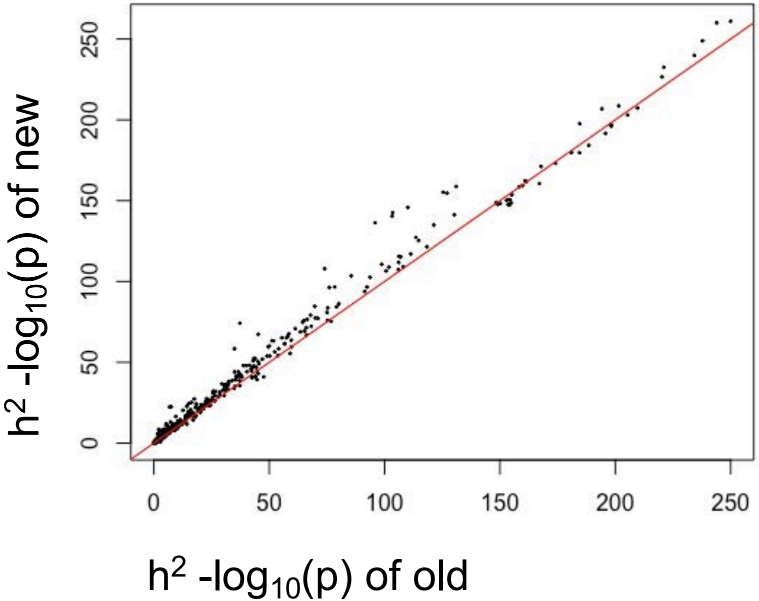
Estimating heritability from GWAS summary statistics

How do round 2 ldsc results compare?

Raymond Walters

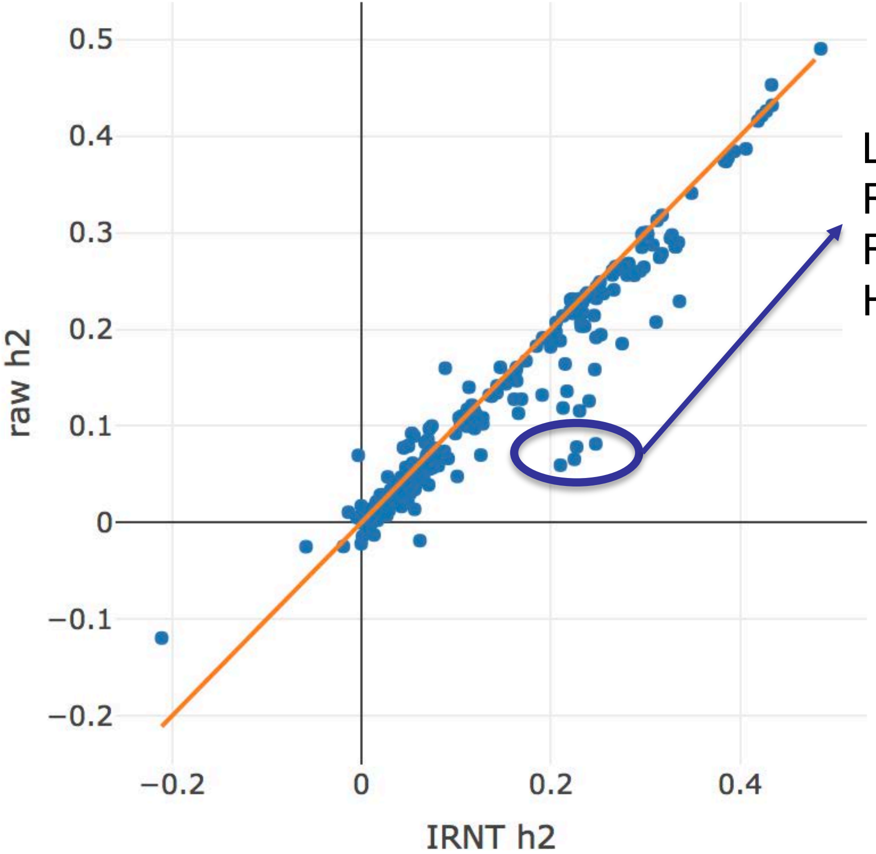
- Intercept less significant
- h2 more significant with stable estimates



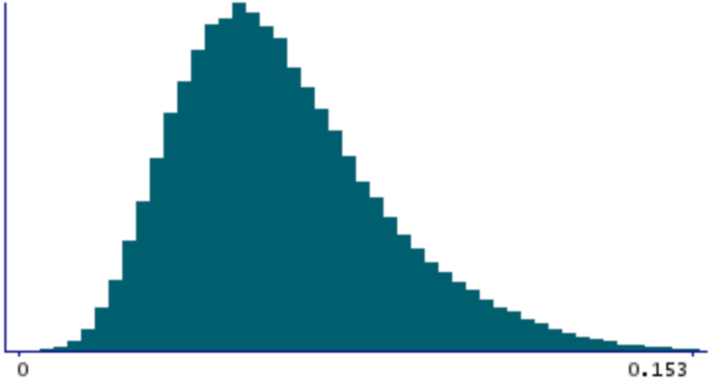


Let’s look at heritability

Raymond Walters



Lymphocyte count
Reticulocyte count
Reticulocyte %
High light scatter reticulocyte %



Reticulocyte count

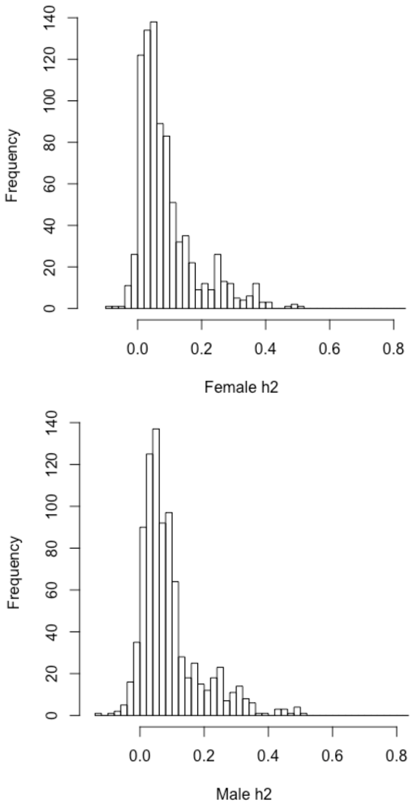
What about sex-specific effects?

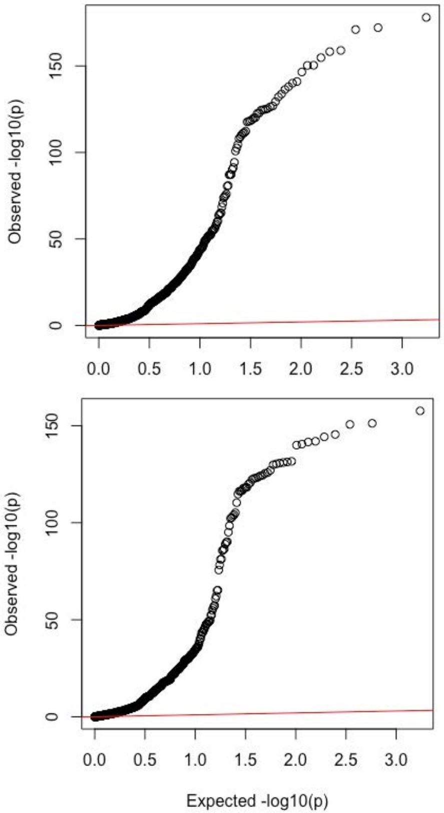
Raymond Walters

- Sex-specific GWAS allow us to scan for:
  - Differences in female vs. male h^2
    - E.g. could indicate differences in variance of environmental effects, measurement differences
  - female vs. male $r_g < 1$
    - E.g. relative effects of different SNPs differ by sex
- Can also test for SNP-level differences
  - Slower and labor intensive, so h^2, r_g can help prioritize
- To start: look at 448 phenotypes with Neff > 10000 in both sexes and z-score of h2 > 4 is at least 1 sex

Strong h^2 observed in both sexes

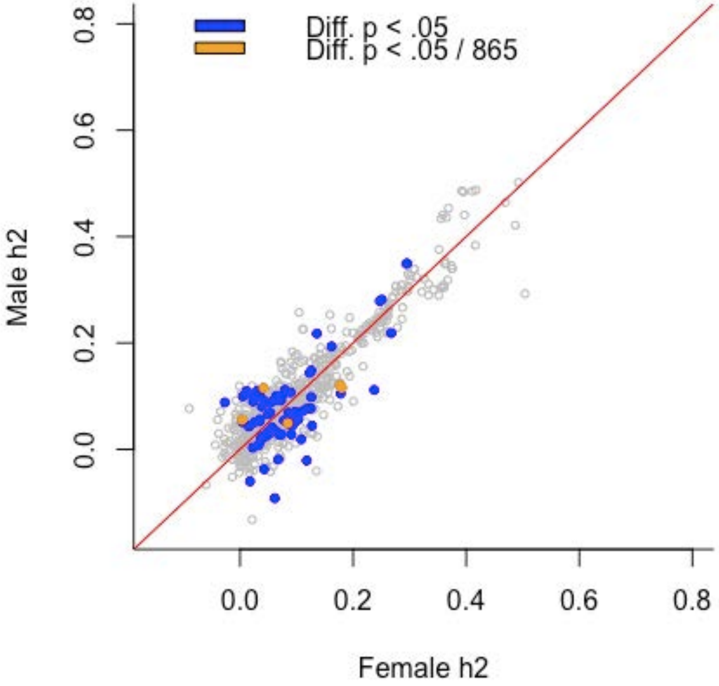
- >70% of traits at least nominally heritable in each sex
  - $P < .05$
- Mean $h^2 \sim .09$
- Consistent with joint analysis of both sexes





Is h^2 equal across sexes?

h^2 strongly correlated across sex



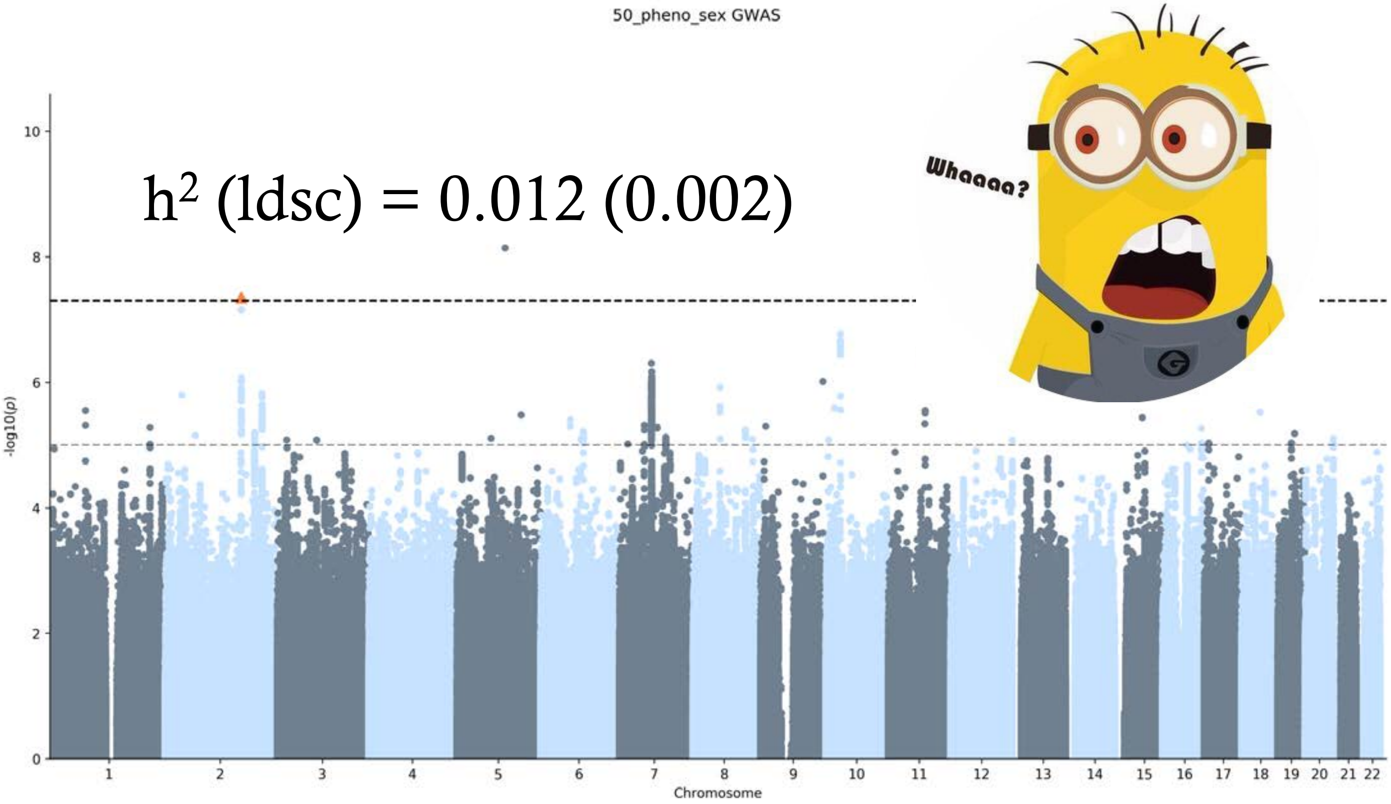
~10% of traits have nominally different h2 between sexes

description	Fem. h2	Male h2	P diff
Average weekly beer plus cider intake	0.0416	0.1152	3.11E-10
Diastolic blood pressure, automated	0.1799	0.1160	1.13E-06
Systolic blood pressure, automated	0.1768	0.1208	1.03E-05
Number of operations, self-reported	0.0845	0.0491	2.53E-05
Duration of vigorous activity	0.0037	0.0555	3.91E-05

Female (1) vs male (0) GWAS

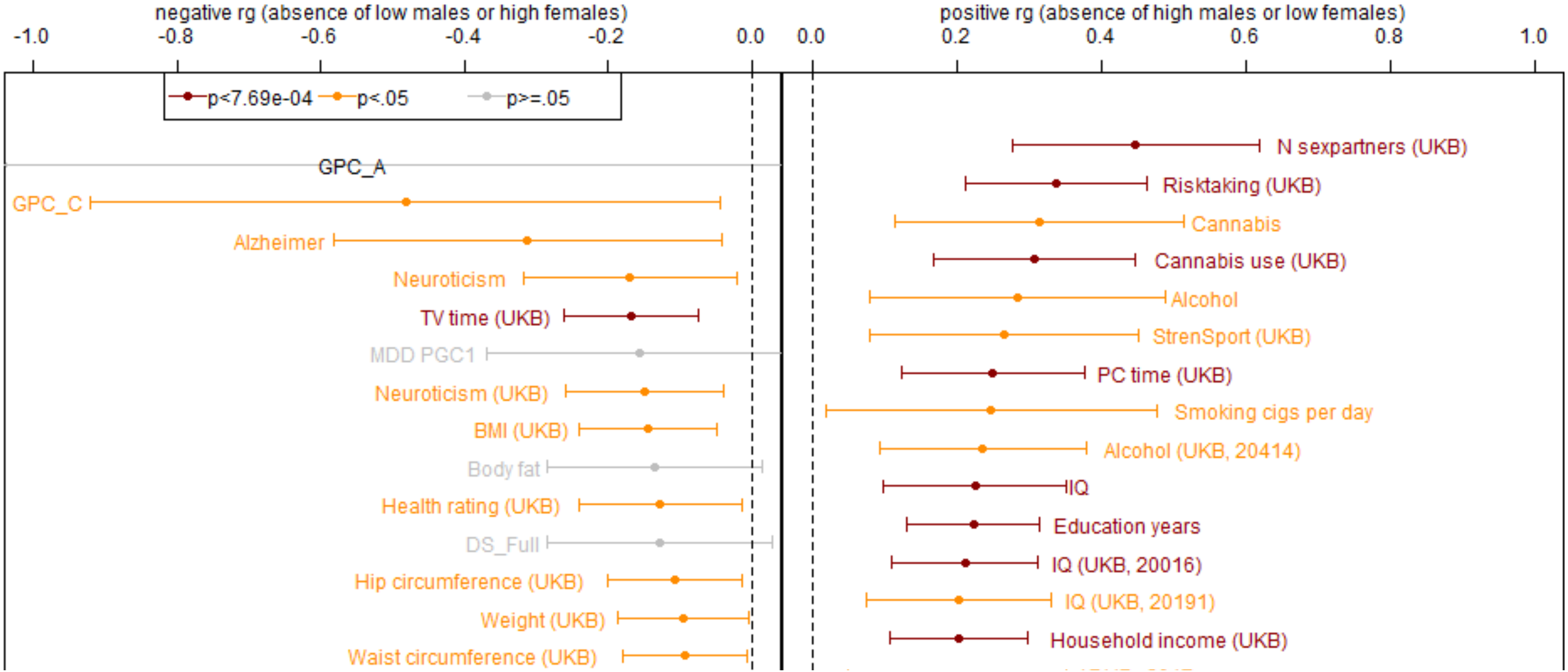
Michel Nivard

Mattijs van der Zee



h^2 (ldsc) = 0.012 (0.002)

Differential ascertainment bias



Male/Female genetic correlation

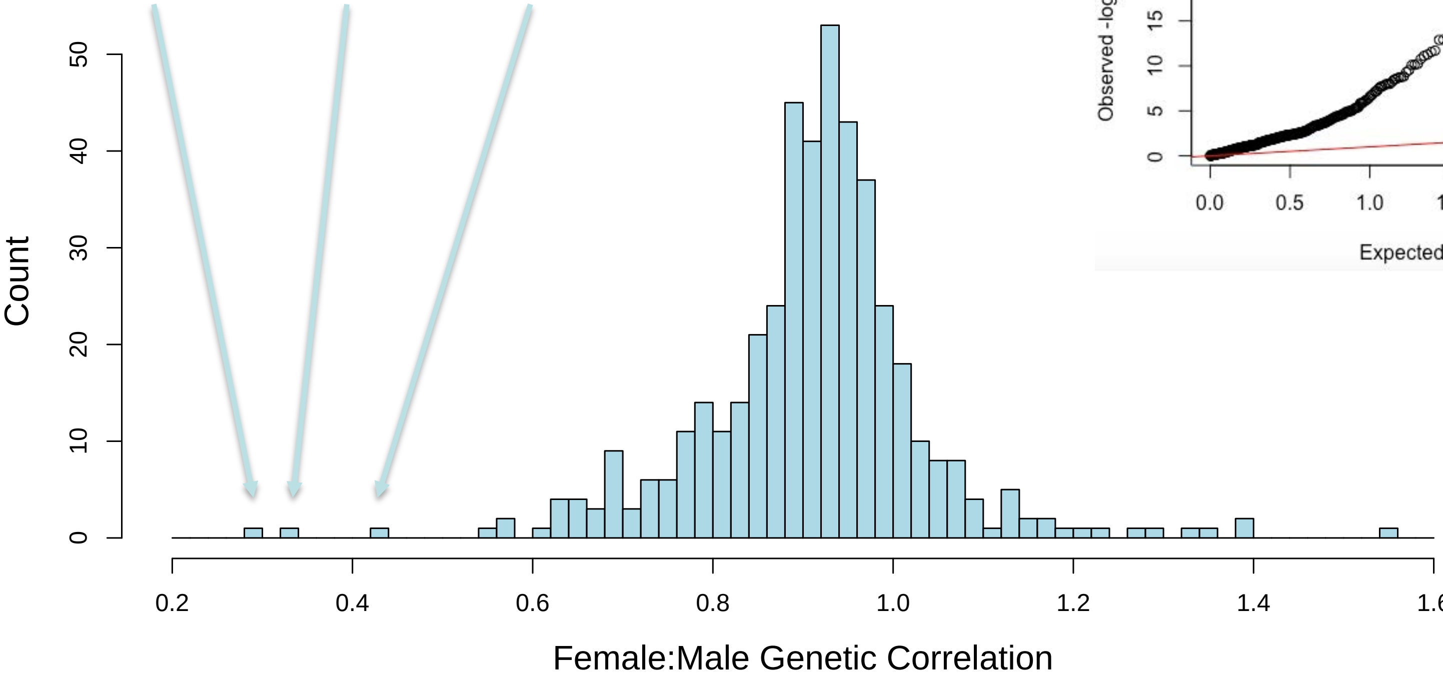
Raymond Walters

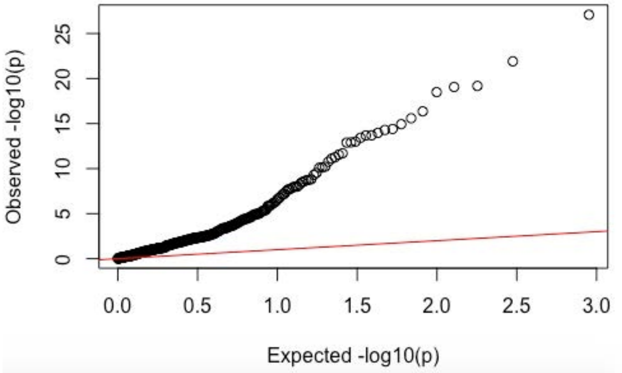
- Next step is to look at genetic correlation between female and male results for each trait
  - Again using LD score regression

- Focus on 448 traits with significant h^2 in at least one sex
  - After Bonferroni correction for 865 traits

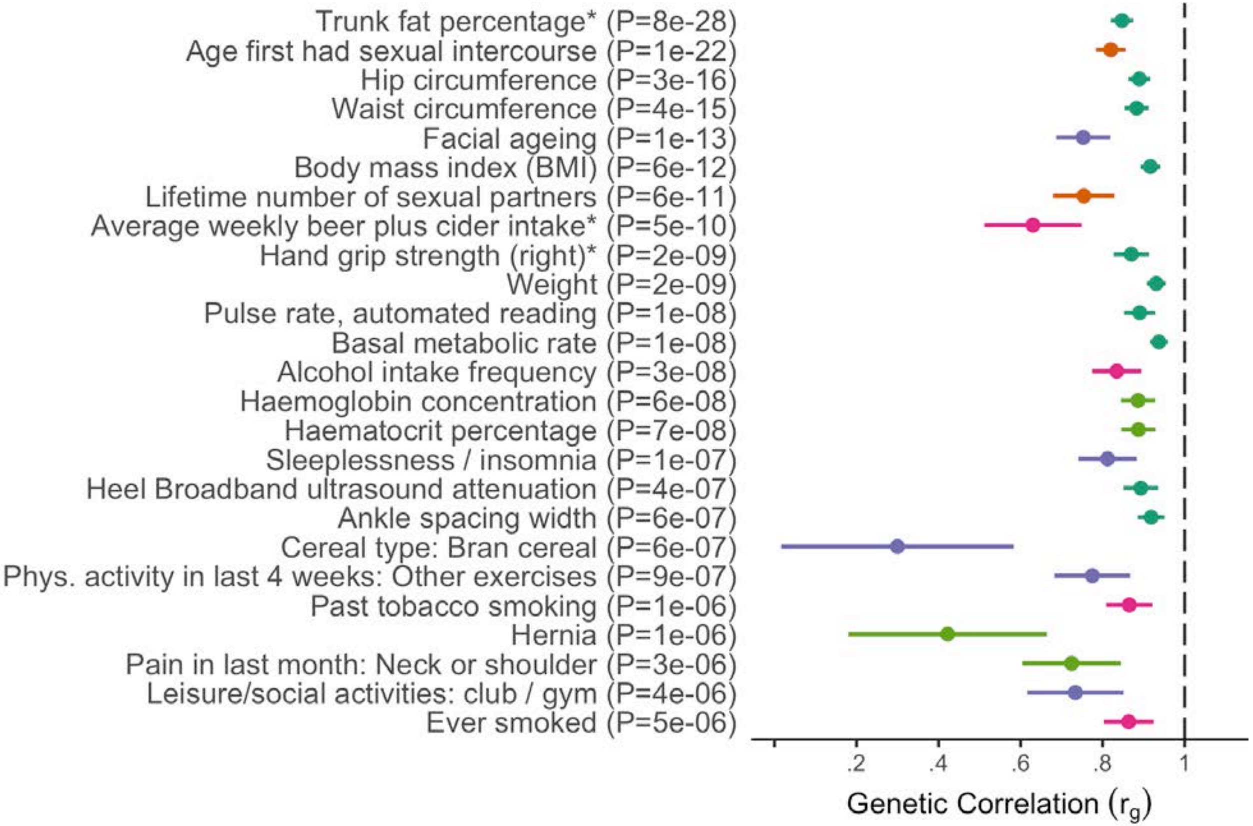
Genetic correlation estimate between females and males

CerDiseatsyepe: oBfraincaortyesxisaem





Phenotypes with male/female rg significantly < 1 (p < 1e-5)

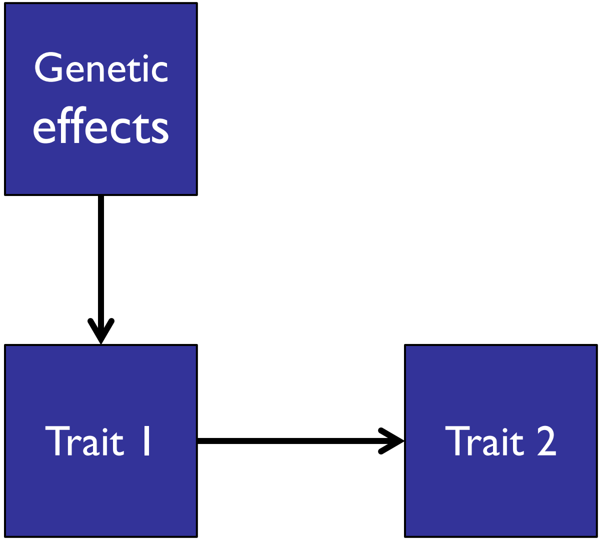


Genetic Correlation Method in:

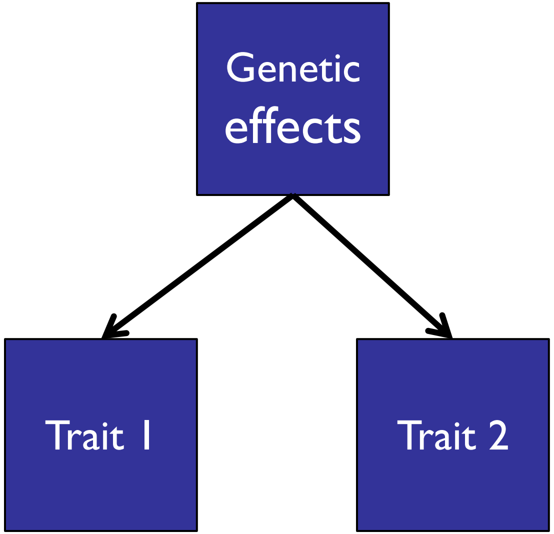
nature genetics

An atlas of genetic correlations across human diseases and traits

Potential sources of genetic correlation

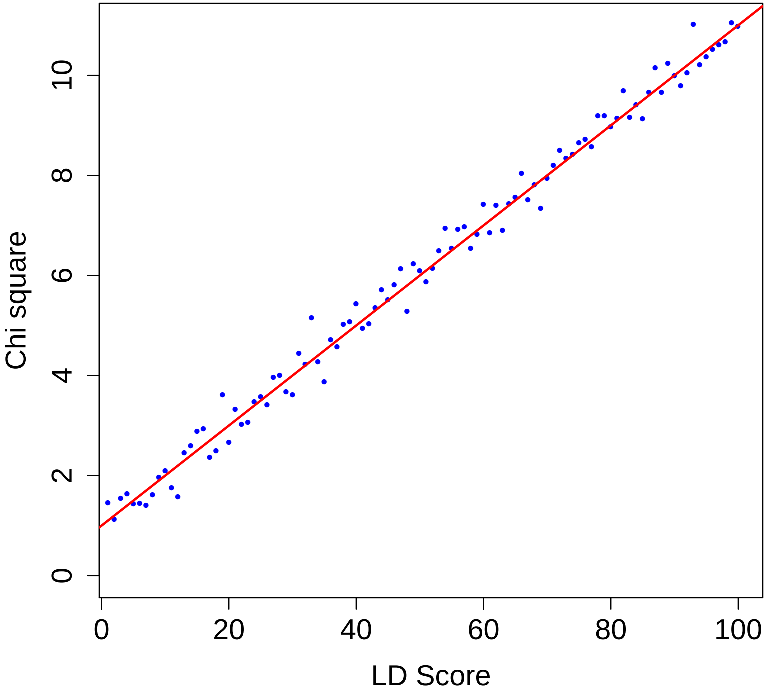


Trait 1 exerts causal effect on Trait 2



Genetic effects influence
Trait 1 and Trait 2

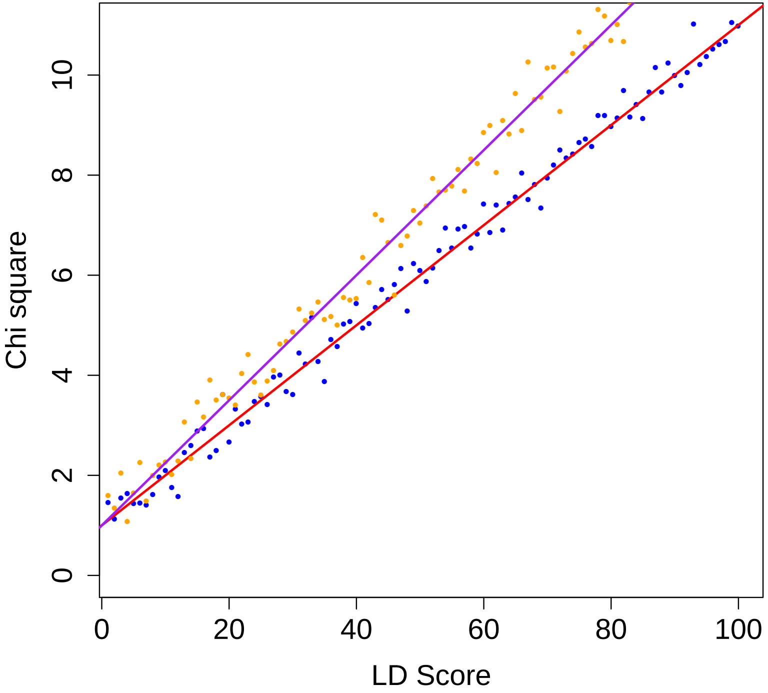
LD Score regression
Genetic correlation



| Trait 1

Slope estimates heritability

LD Score regression
Genetic correlation



🟥 Trait 1

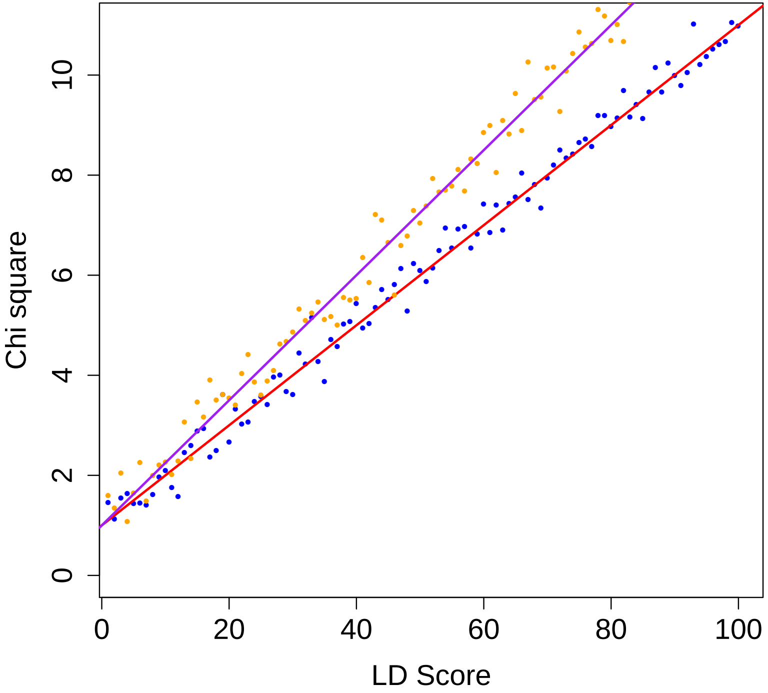
🟪 Trait 2

We can a second trait and obtain two heritability estimates

LD Score regression
Genetic correlation

🟥 Trait 1

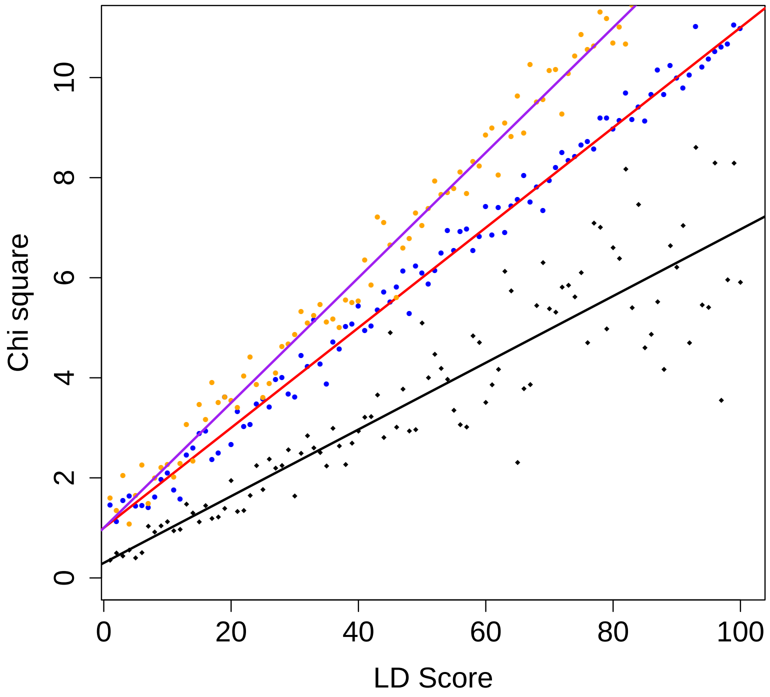
🟪 Trait 2



$Z^*Z = \chi^2$

So we can estimate genetic covariance from the product of the Z-scores

LD Score regression
Genetic correlation



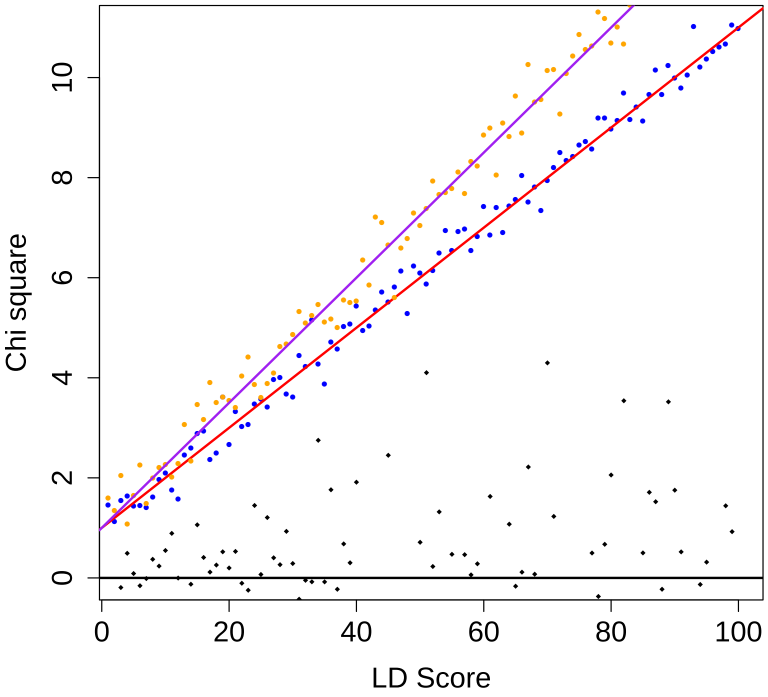
🟥 Trait 1
🟪 Trait 2
⬛ R_G

$Z*Z = \chi^2$

So we can estimate genetic covariance from the product of the Z-scores for the two traits

$R_G = 0.5$

LD Score regression
Genetic correlation



🟥 Trait 1

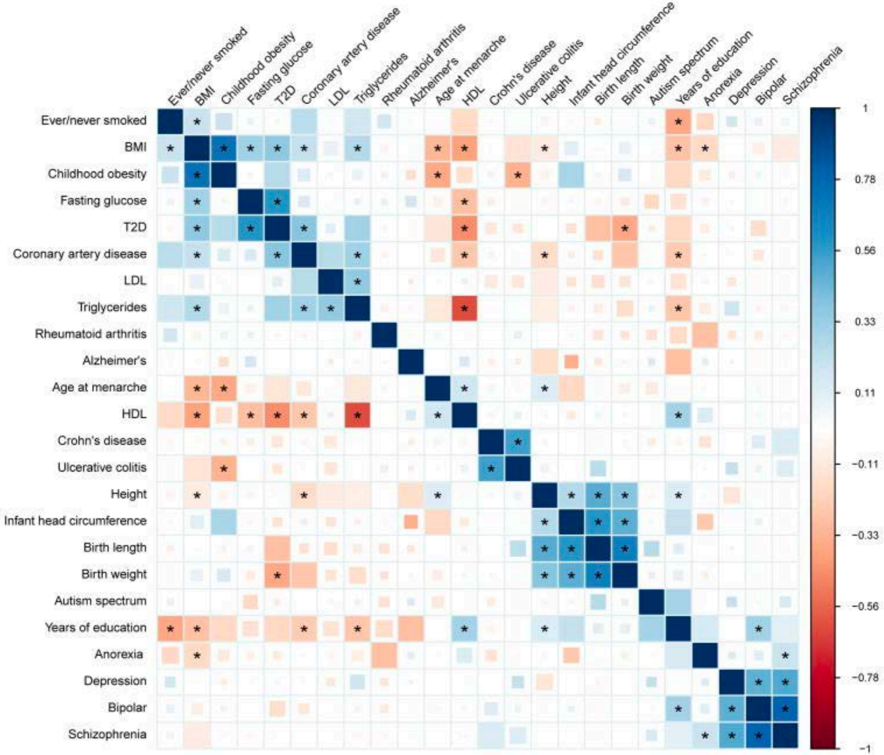
🟪 Trait 2

⬛ R_G

Here $R_G = 0$

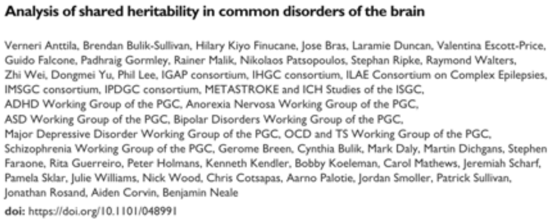
This approach is robust to sample overlap as all variants are equally inflated

Genetic correlations



- Genetic correlation is a widespread phenomenon

Brainstorm Project



Analysis of shared heritability in common disorders of the brain

Verneri Anttila, Brendan Bulik-Sullivan, Hilary Kiyo Finucane, Jose Bras, Laramie Duncan, Valentina Escott-Price, Guido Falcone, Padhraig Gormley, Rainer Malik, Nikolaos Patsopoulos, Stephan Ripke, Raymond Walters, Zhi Wei, Dongmei Yu, Phil Lee, IGAP consortium, IHGC consortium, ILAE Consortium on Complex Epilepsies, IMSGC consortium, IPDGC consortium, METASTROKE and ICH Studies of the ISGC, ADHD Working Group of the PGC, Anorexia Nervosa Working Group of the PGC, ASD Working Group of the PGC, Bipolar Disorders Working Group of the PGC, Major Depressive Disorder Working Group of the PGC, OCD and TS Working Group of the PGC, Schizophrenia Working Group of the PGC, Gerome Breen, Cynthia Bulik, Mark Daly, Martin Dichgans, Stephen Faraone, Rita Guerreiro, Peter Holmans, Kenneth Kendler, Bobby Koeleman, Carol Mathews, Jeremiah Scharf, Pamela Sklar, Julie Williams, Nick Wood, Chris Cotsapas, Aarno Palotie, Jordan Smoller, Patrick Sullivan, Jonathan Rosand, Aiden Corvin, Benjamin Neale

doi: https://doi.org/10.1101/048991

Verneri Anttila

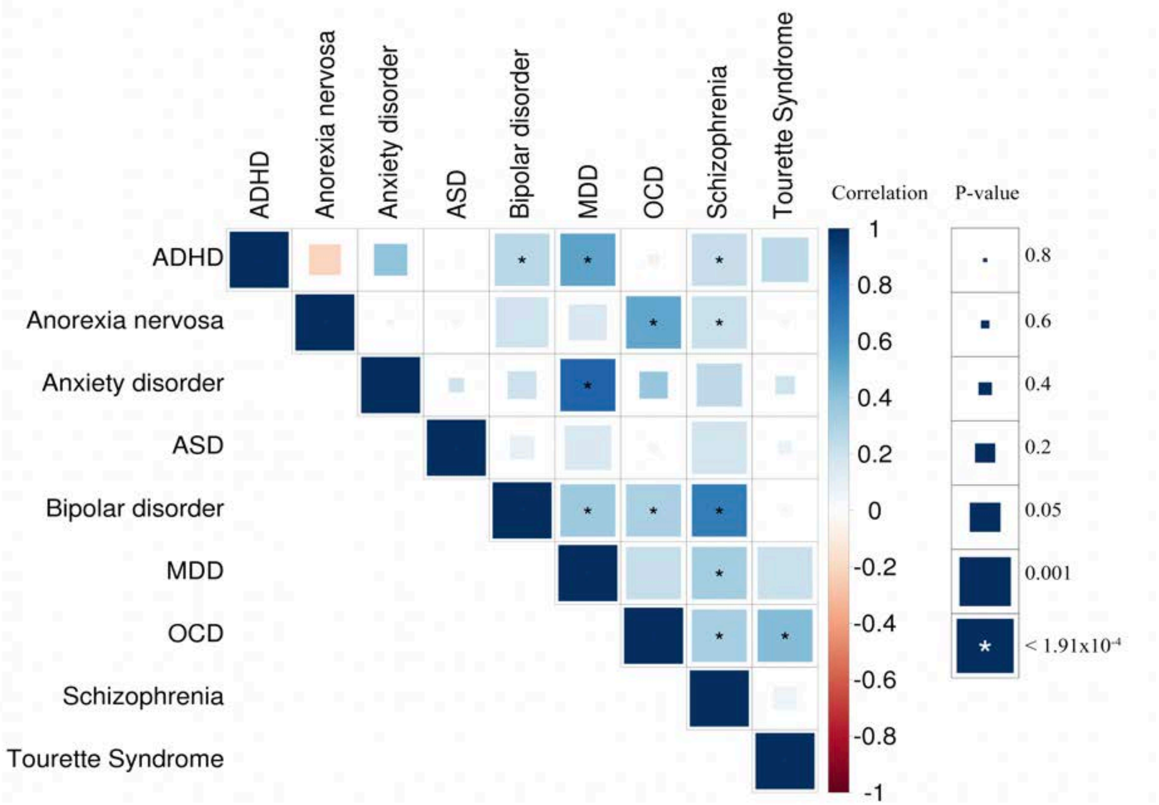
Aiden Corvin

Brendan Bulik-Sullivan
Hilary Finucane
Jonathan Rosand
Aarno Palotie
Mark Daly
Patrick Sullivan
Bobby Koeleman
Nick Wood
Julie Williams

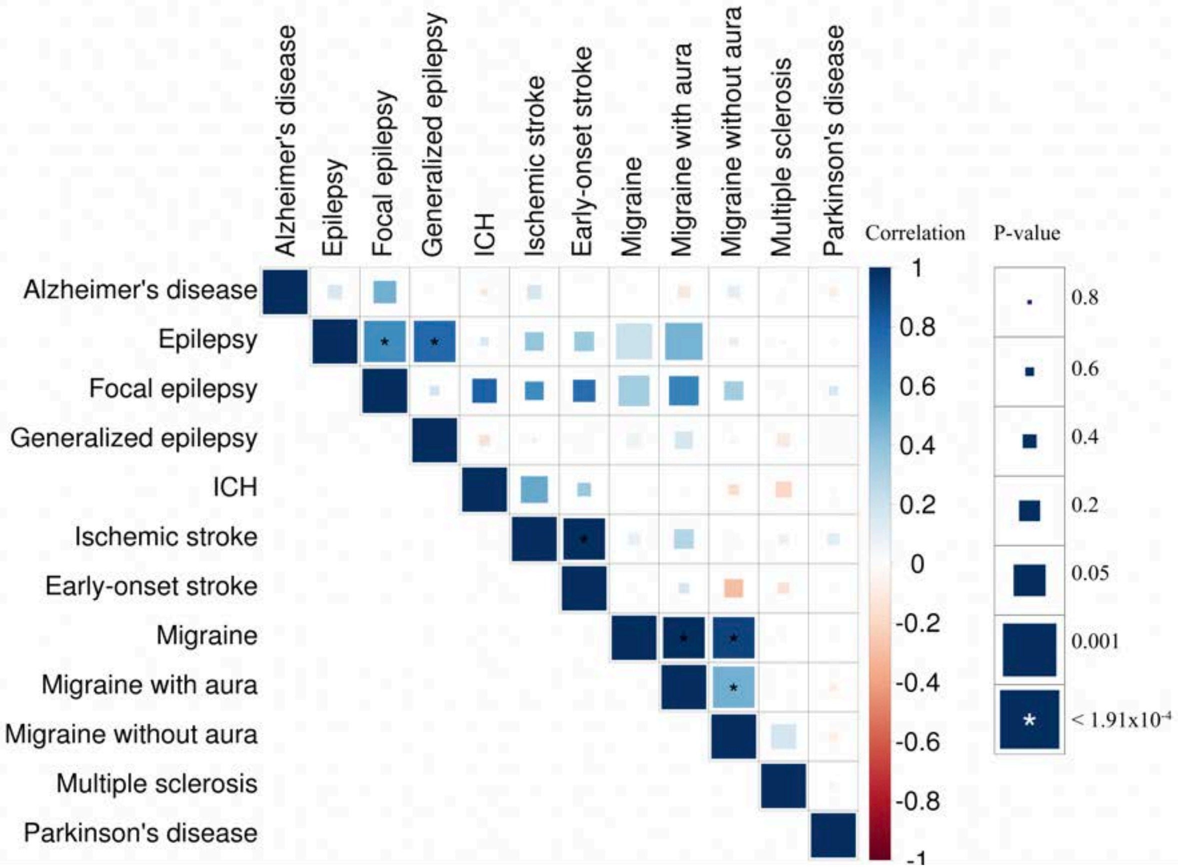
Alessandro Biffi
Jeremiah Scharf
Kenneth Kendler
Stephan Ripke
Alkes Price
Chris Cotsapas
Padhraig Gormley
Zhi Wei
Rainer Malik

Hailiang Huang
Andrea Byrnes
Dongmei Yu
Laramie Duncan
Kai-How Farh
Namrata Gupta
Miriam Raffeld
...and many, many others
in their respective study groups

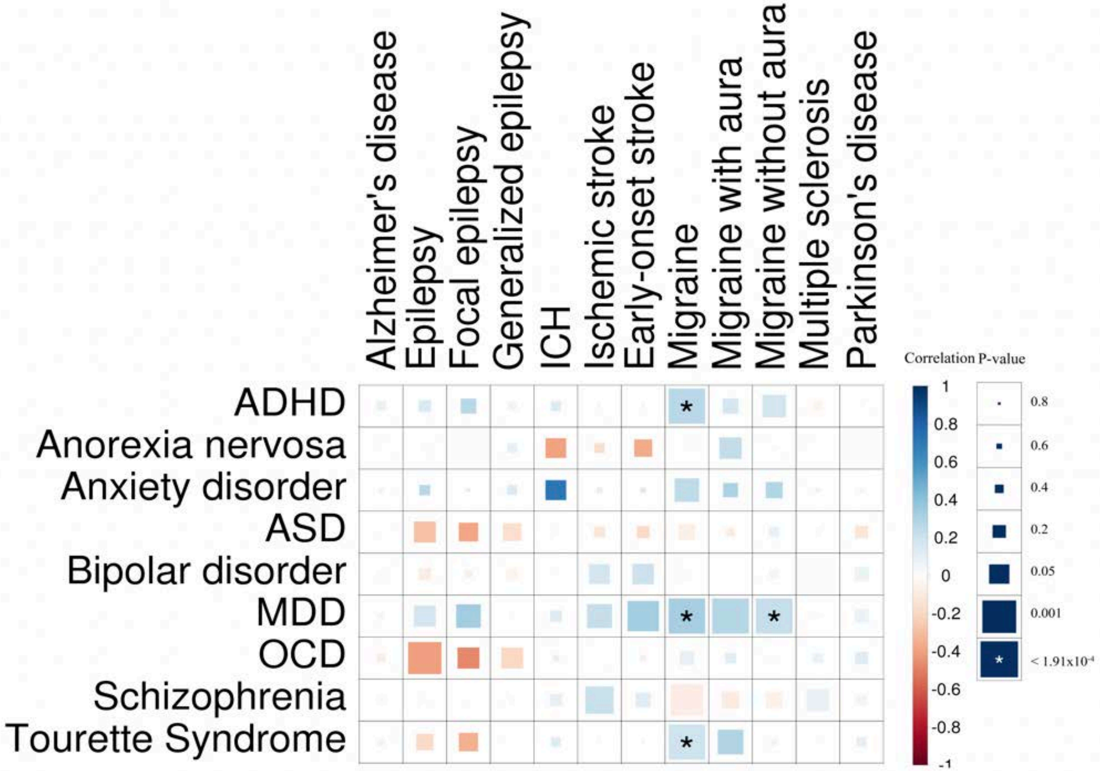
Brainstorm within psychiatry



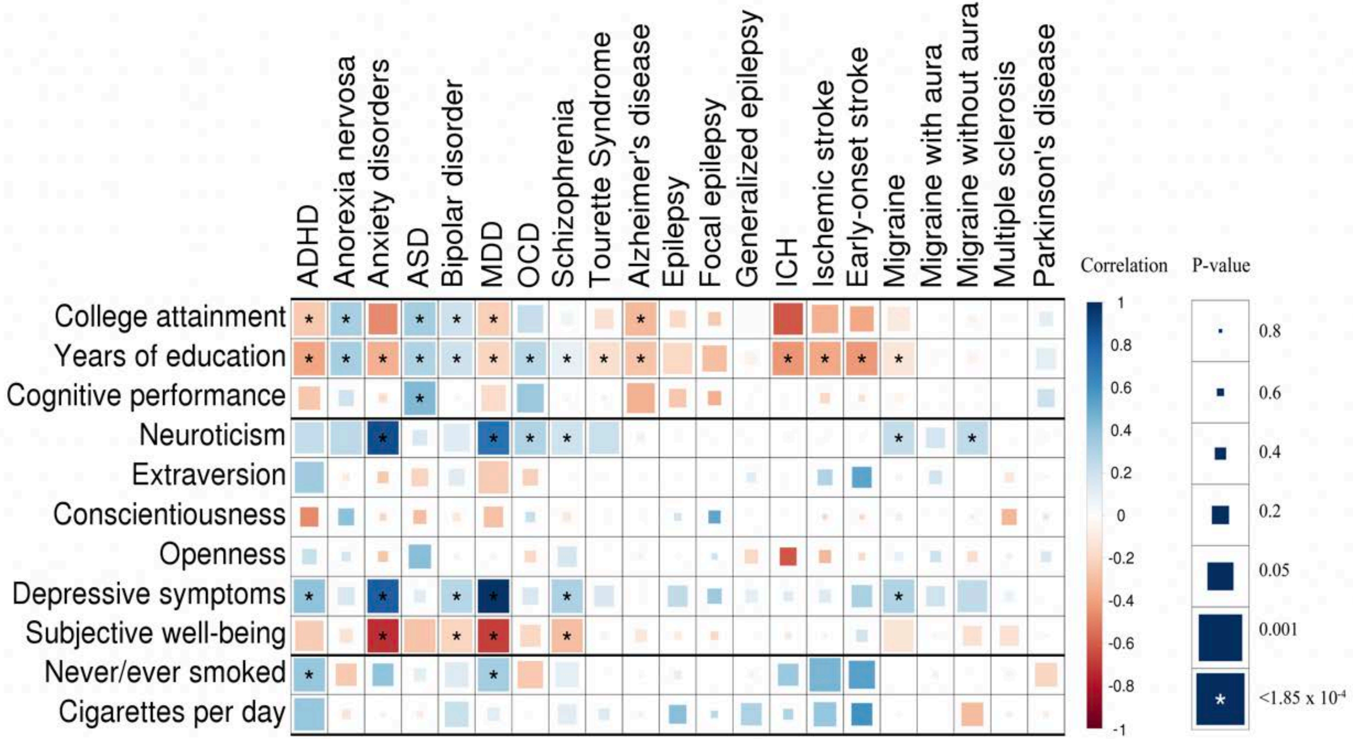
Brainstorm within neurology



Brainstorm – across neurology and psychiatry

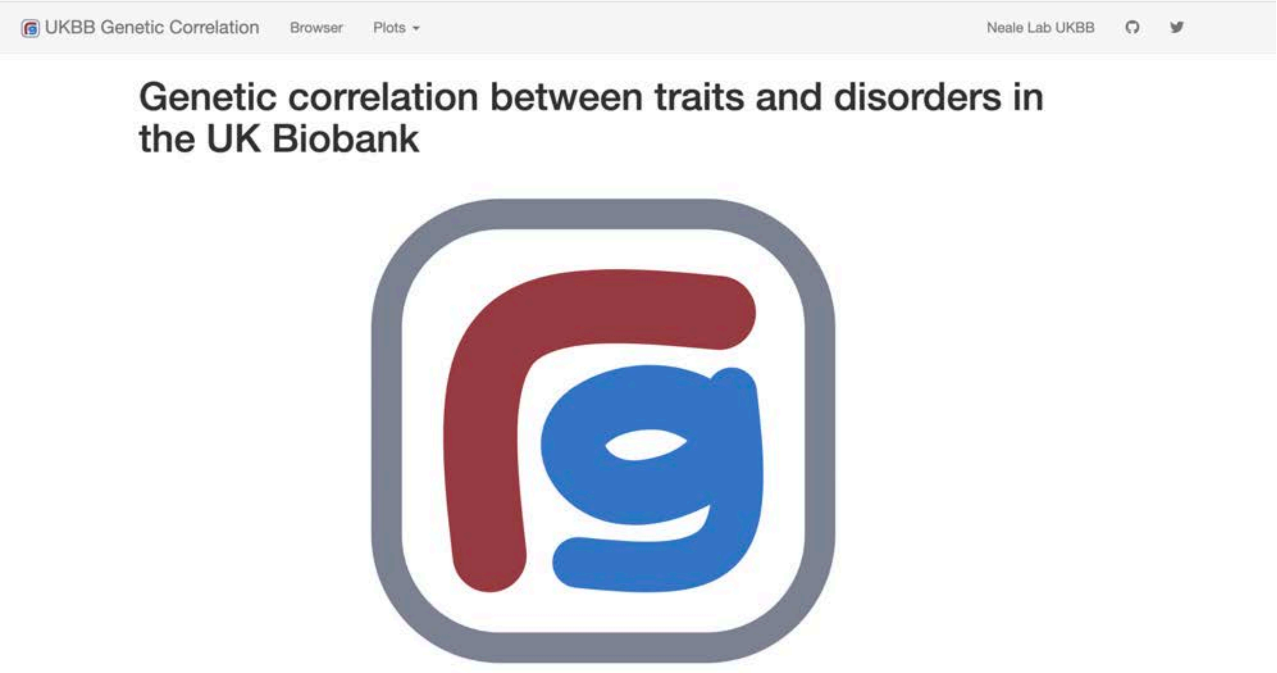


Brainstorm – take it further?



Comprehensive evaluation of genetic correlation

Duncan Palr



Genetic correlation between traits and disorders in the UK Biobank

https://ukbb-rg.hail.is/

https://github.com/astheeggeggs/UKBB_ldsc_r2