Session 1 :- Using summary stats & PheWAS in the UKB

9 – 9:15am; **Kathryn Kemper**, research fellow IMB University of Queensland

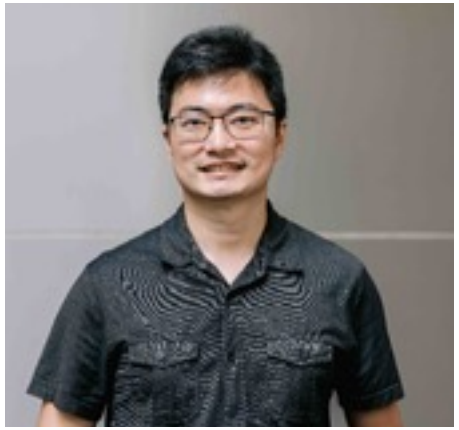
- Welcome & housekeeping
- Introduction to UKB & GWAS

9:15am – 10am; **Jian Zeng**, research fellow IMB University of Queensland

- Using summary statistics
  - *where* to get them, *what* to look for, *why* are the useful?
- Workshop session, summary-based BayesR (sBayesR) for polygenic score prediction

10am – 10:30am; **Isabelle McGrath**, PhD student IMB University of Queensland

- Introduction to health-related outcomes in the UKB
- Mapping phecodes using UKB ICD-10 codes
- Running a PheWAS



Jian Zeng

Isabelle McGrath

biobank uk

Improving the health of future generations

A study of genes, the environment & health

500,000 volunteers

recruited 2006-2010

aged 40-69 years

22 assessment centres

Initial ('baseline') assessment

Broad consent

Lifestyle & diet questionnaire

Physical measurements

Blood, urine & saliva samples

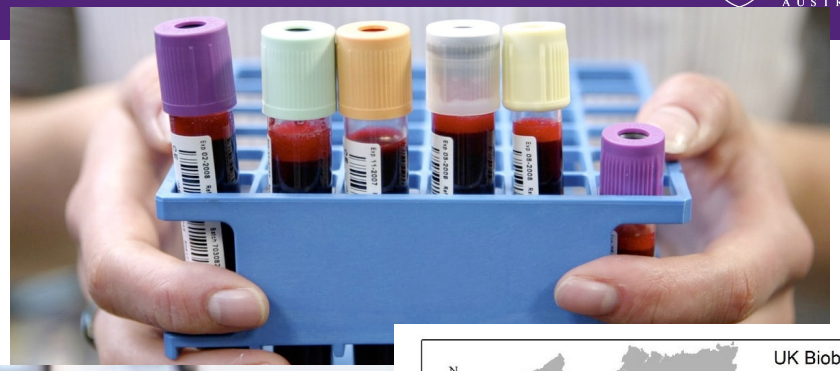


Figure. Spatial locations of 22 UK Biobank assessment centres with number of participants (*Sarkar, Webster & Gallacher, 2014*).

‘New’ data in the UK Biobank

- Imaging: Brain, heart & full body MRI; fully body DEXA scan of bones & joins. Goal of 100,000 participants plus repeat visit
- Genetics: Whole genome sequence & genotyping for all participants, whole exome sequence 470K participants
- Health-linkages: linkage to health-related databases, e.g. death, cancer, hospital and primary care records
- Blood Biomarkers: 30 key biochemistry markers for all participants
- Online questionnaires: for a range of exposures such as diet, work history, pain, cognitive function, and mental health
- COVID-19 antibody data on 260K participants
- ...on-going data collection & releases

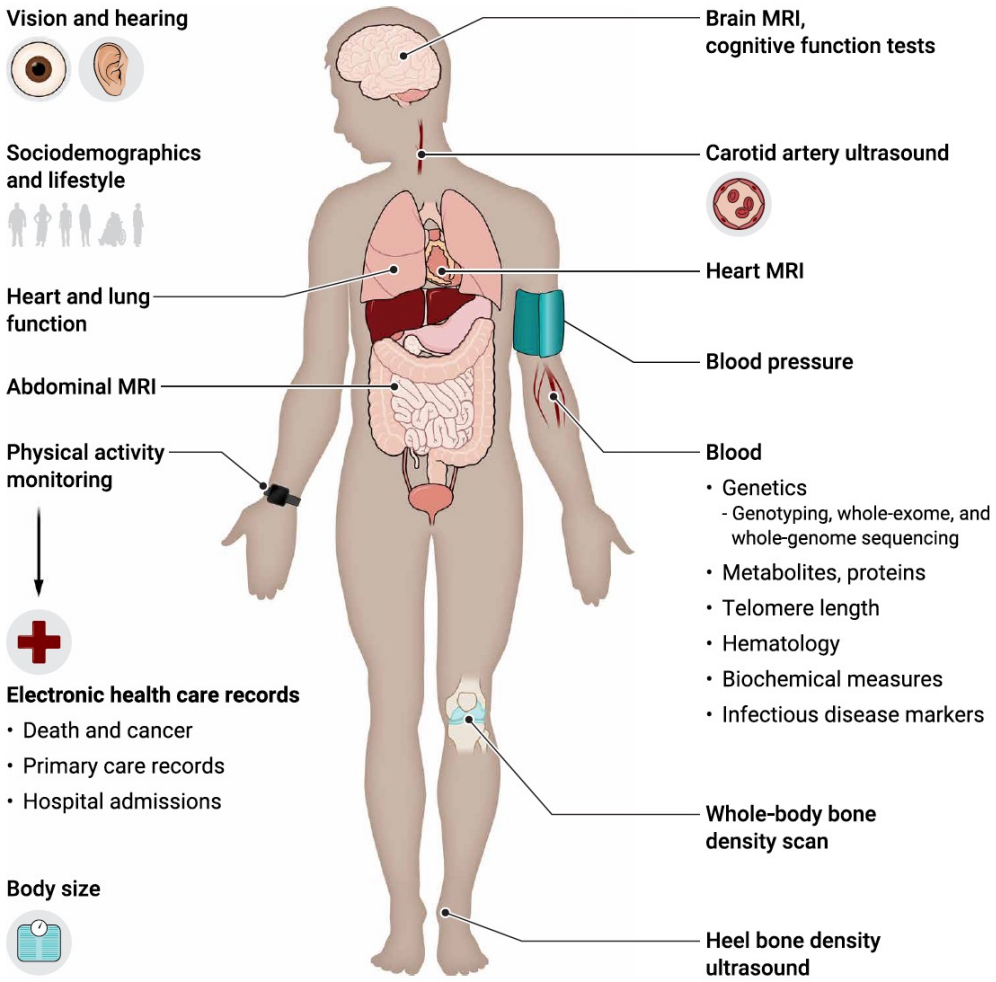


Fig. 1. Types of data in UK Biobank. Shown are the types of data collected in UK Biobank, including data collected at in-person assessments such as lifestyle factors, medical history, blood pressure and other physical measures, and imaging scans. Other data include information from online questionnaires, data generated from biological samples, and data derived from electronic health care records.

Allen *et al.*, *Sci. Transl. Med.* **16**, eadf4428 (2024) 10 January 2024

‘returned’ datasets catalogue

biobank uk

Index | Browse | Search | Catalogues | Downloads | Login | Help

Application 12505

Title: Dissecting the genetic basis of relationships between early-life and later-life events
Lead Institution: University of Queensland
Principal investigator: Professor Peter Visscher

About | 11 Returns | 11 Publications | 1 Category

Return ID	App ID	Description	Archive Date
3583	12505	A unified framework for association and prediction from vertex-wise grey-matter structure	22 Jun 2021
3201	12505	Association Between Population Density and Genetic Risk for Schizophrenia	11 Mar 2021
1831	12505	Detection and quantification of inbreeding depression for complex traits from SNP data	26 Nov 2019
2904	12505	Dissection of genetic variation and evidence for pleiotropy in male pattern baldness	30 Nov 2020
3365	12505	Extreme inbreeding in a European ancestry sample from the contemporary UK population	22 Apr 2021
3393	12505	Genome-wide association study identifies 143 loci associated with 25 hydroxyvitamin D concentration	29 Apr 2021
3301	12505	Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank	9 Apr 2021
3630	12505	Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies	9 Jul 2021
3255	12505	Improved polygenic prediction by Bayesian multiple regression on summary statistics	26 Mar 2021
3447	12505	The effect of X-linked dosage compensation on complex trait variation	25 May 2021
3672	12505	Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations	27 Jul 2021

Enabling scientific discoveries that improve human health

Examples of data types we use in our group:

- SNP genotypes
- brain imaging data
- baseline measurements
  - quantitative traits
  - categorical variables
- self-report disease & hospital records
- family history
- blood biomarkers

collider bias / participation bias / G-E correlation

- Postal invitations were sent to 9.2 million individuals living near an assessment center, but 'only' 5.2% of those people joined the UK Biobank
- Those who joined were not a random sample of the UK population. Commonly referred to as 'healthy volunteer' bias. That is UKB participants tend to be, e.g.
  - from more affluent areas
  - non-smokers
  - use vitamin supplements
  - have lower rates of disease
  - etc.
- Be aware of how selection bias may influence your results! Various ways to address this issue:
  - sensitivity analysis
  - probability weightings
  - covariates
  - simulations

Technical Report | Open access | Published: 13 July 2023

Studying the genetics of participation using footprints left on the ascertained genotypes

Stefania Benonisdottir & Augustine Kong

Nature Genetics **55**, 1413–1420 (2023) | Cite this article

Abstract

The trait of participating in a genetic study probably has a genetic component. Identifying this component is difficult as we cannot compare genetic information of participants with nonparticipants directly, the latter being unavailable. Here, we show that alleles that are more common in participants than nonparticipants

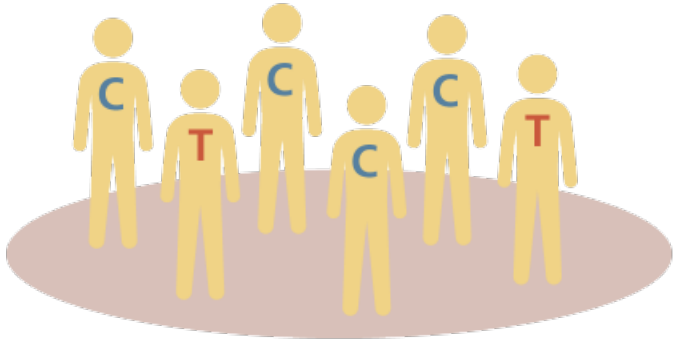
What is a GWAS?

- first large-scale set of analyses done when genetic data was released in 2017/2018; now 7,221 phenotypes across 6 continental ancestry groups

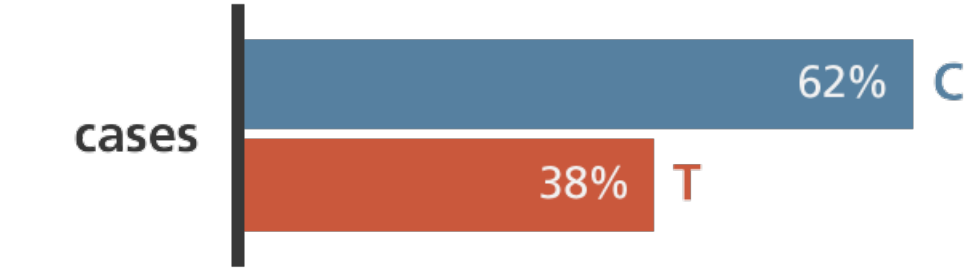
- A **G**enome **W**ide **A**ssociation **S**tudy is a hypothesis-free method for identifying associations between locations in the genome and a trait of interest

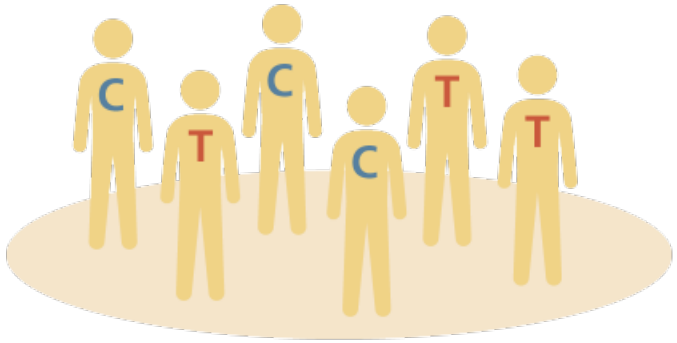
- Three key parts to a GWAS:
  - A trait of interest or phenotype
  - Genetic markers measured across the genome
  - Statistical test of association between markers & phenotype

Example: Binary trait



cases (n=1,000)
people with heart disease





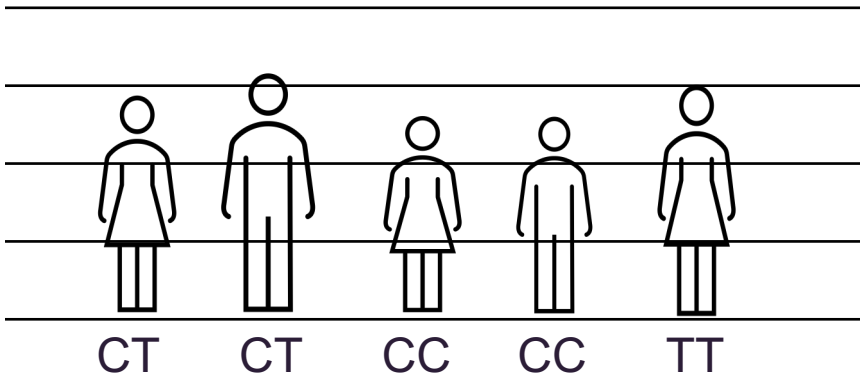
controls (n=1,000)
people without heart disease

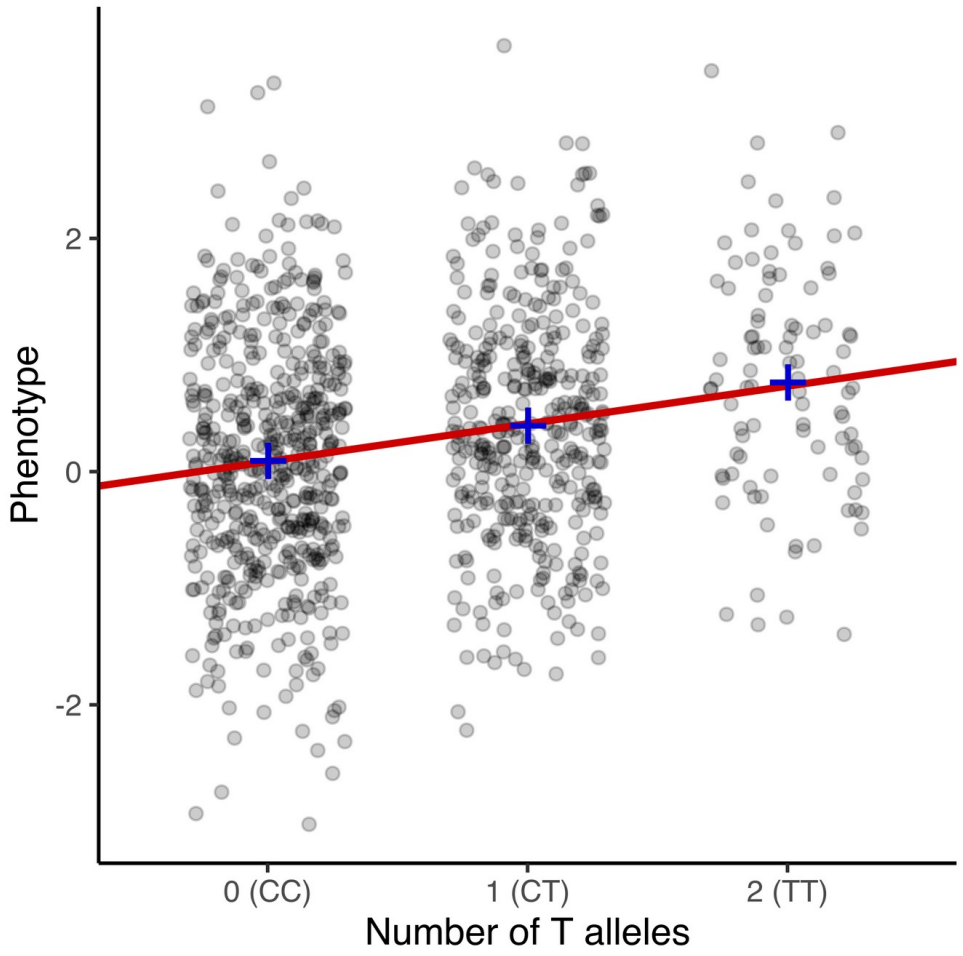
Test of association:

"Is the frequency of the 'C' allele different in cases vs. controls"

P = 0.0012

Example: Quantitative trait





- Linear model:

$$y = \alpha + x\beta + e$$

phenotypes

intercept

genotypes

SNP effect

error

Example: human height

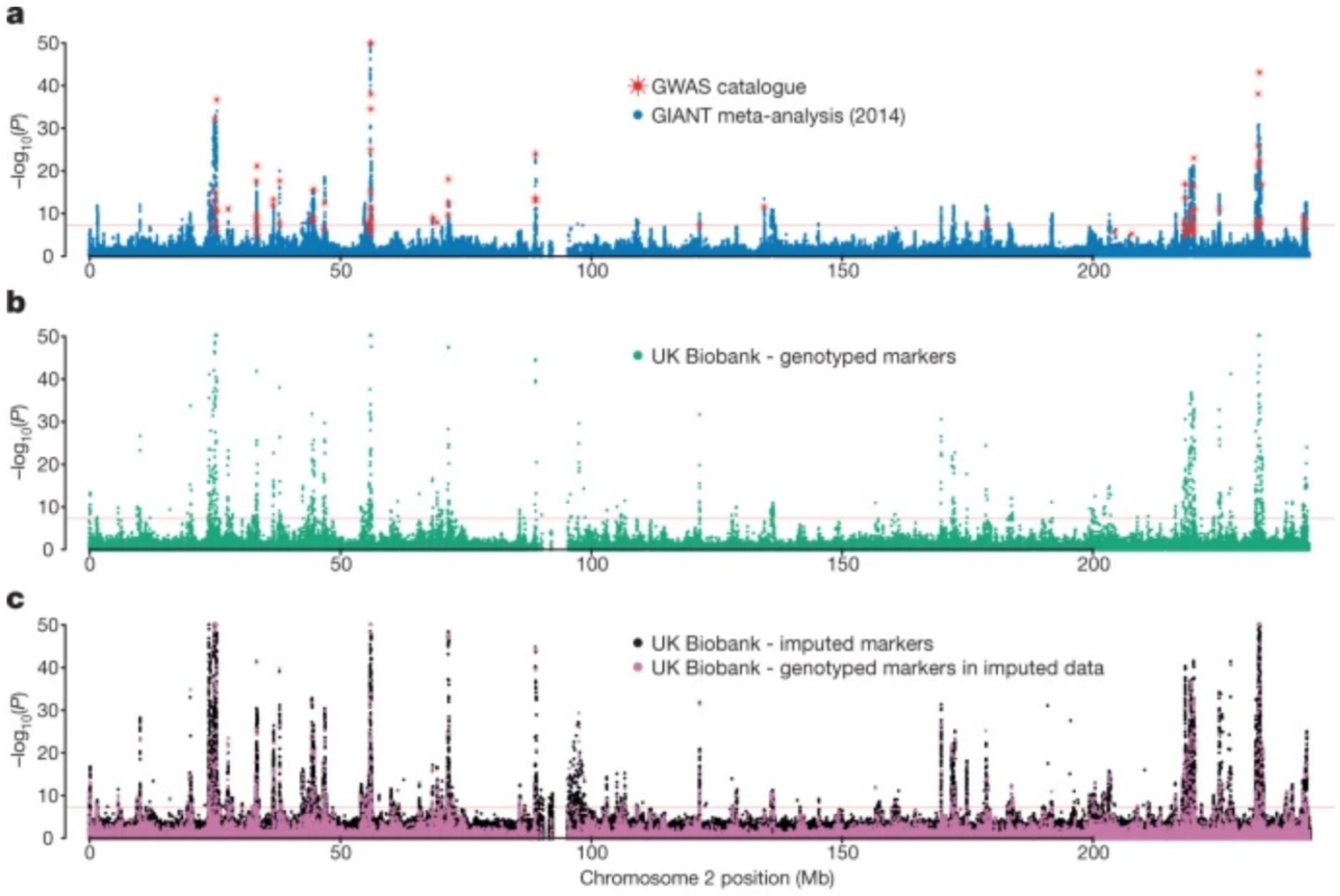
- Results typically visualised as a ‘Manhattan plot’

y-axis: $-\log_{10}P$

x-axis: genome location

- Each test/marker is *not* independent because of linkage disequilibrium (LD)
- We don’t necessarily expect to identify the ‘causal variant’

Fig. 4: Association statistics for human height.



Bycroft et al. (2018)