Software tools for genetic analysis

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

First looks at the CCDG freeze 1 callset

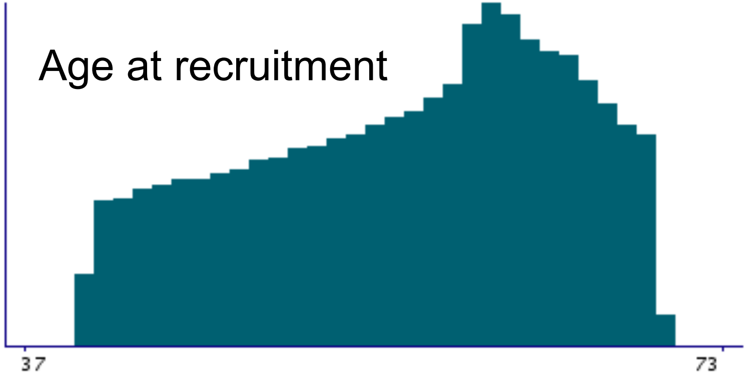
UK Biobank!

The team

- Liam Abbott
- Verneri Anttila
- Krishna Aragam
- Jon Bloom
- Sam Bryant
- Claire Churchhouse
- Joanne Cole
- Mark J. Daly
- Andrea Ganna
- Steven Gazal
- Jackie Goldstein
- Mary Haas
- Joel Hirschhorn
- Daniel Howrigan
- Sekar Kathiresan
- Dan King
- Duncan Palmer
- Tim Poterba
- Manuel Rivas
- Cotton Seed
- Sailaja Vedantam
- Raymond Walters

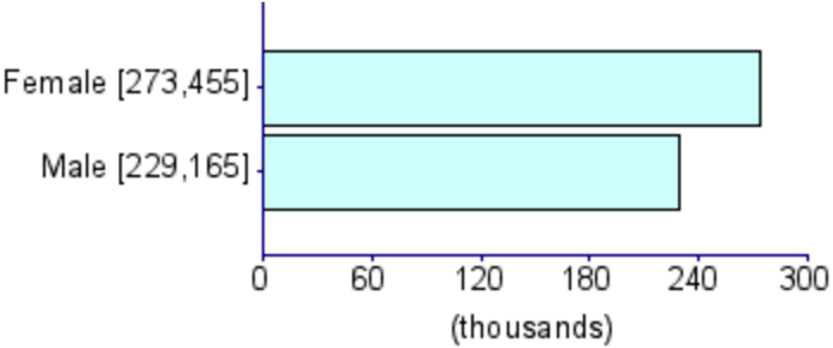
UK Biobank Overview

- Prospective cohort of 502,620 participants
- Recruited across the UK 2006-2010
- Healthy volunteer effect - healthier, leaner, smoke less than population average



Mean 56.5 years (S.D. 8.1)

Participant gender

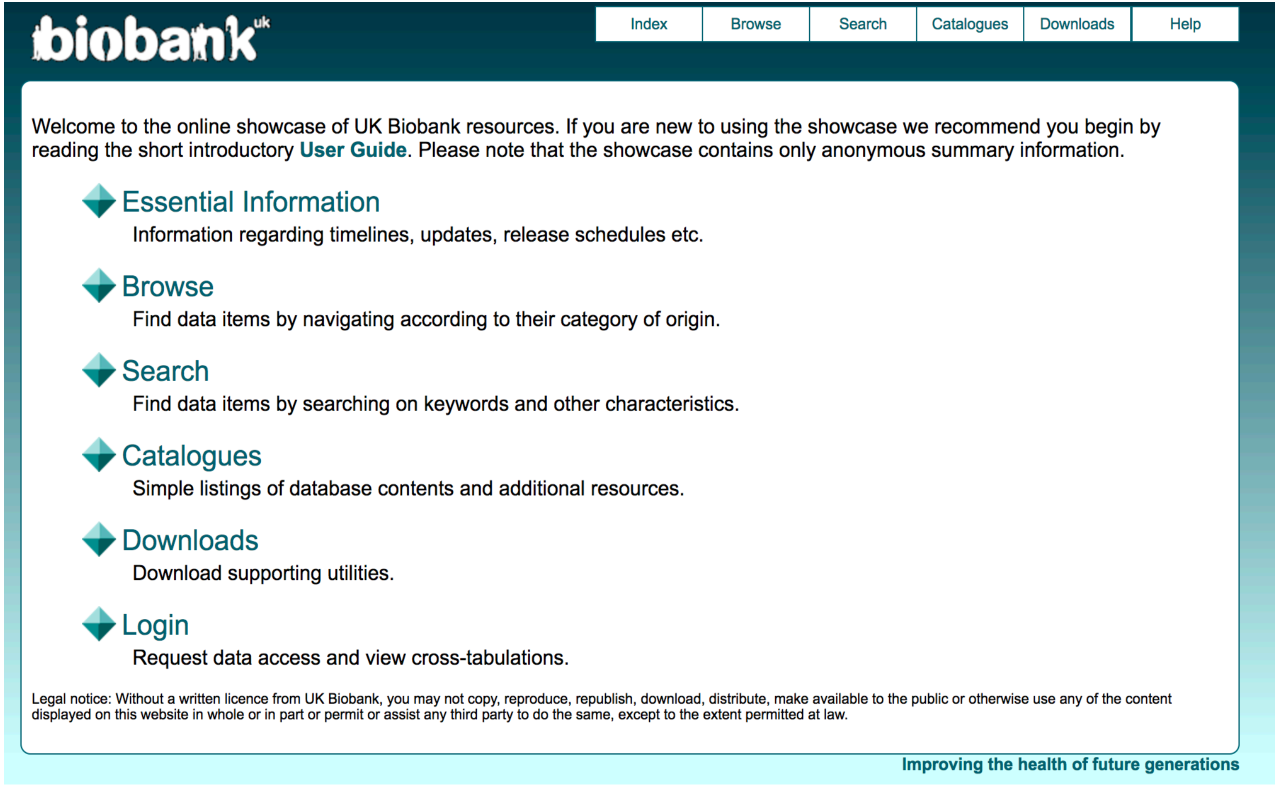


Fry et al, Am J Epi, 21 June 2017

Phenotyping

- Self-reported demographics (500,00), diet and exercise habits (211,000)
- Physical (500,00) and cognitive measurements (190,000)
- Imaging of brain, heart, abdomen (1,000 - 13,700)
- Blood, saliva and urine assays (500,00)
- Medical records (392,00) and cancer registers (80,000)

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



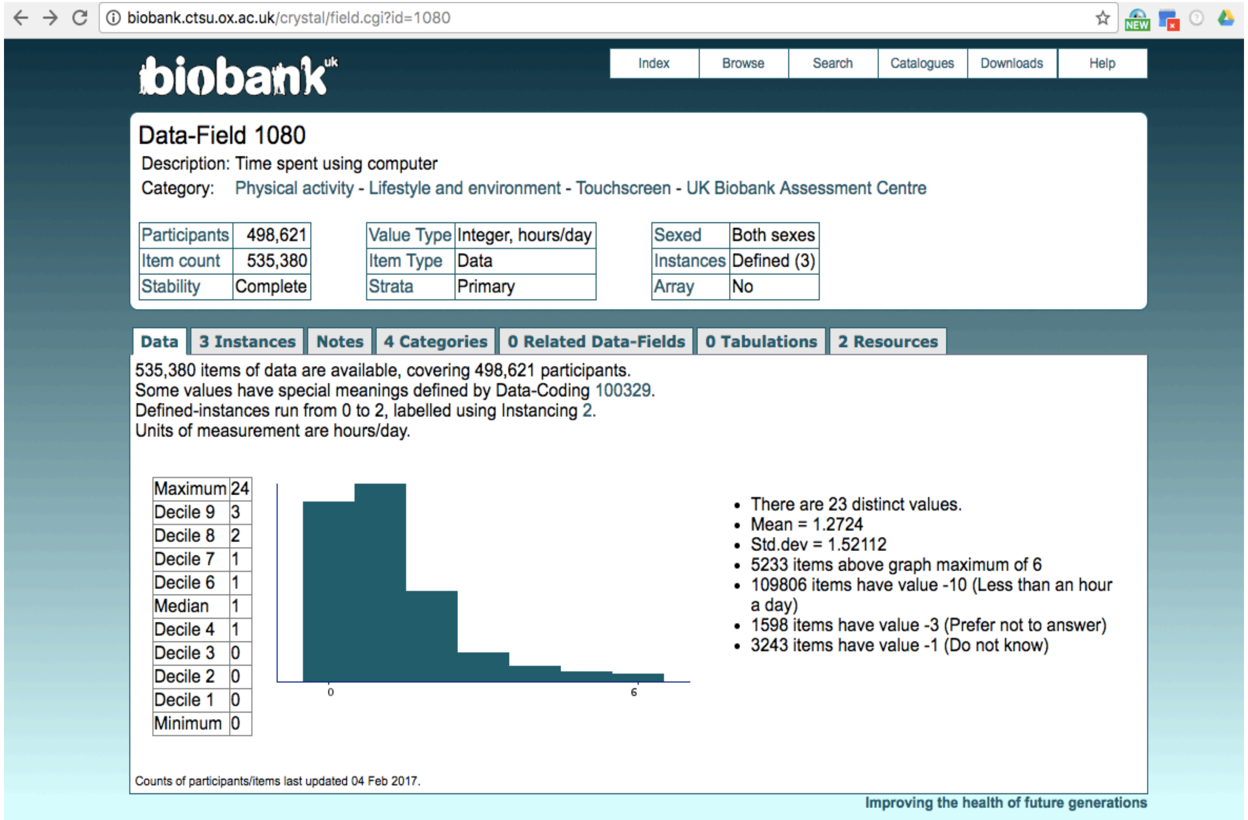
Welcome to the online showcase of UK Biobank resources. If you are new to using the showcase we recommend you begin by reading the short introductory **User Guide**. Please note that the showcase contains only anonymous summary information.

- **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

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



biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=1080

Data-Field 1080

Description: Time spent using computer

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

Participants	498,621
Item count	535,380
Stability	Complete

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

Sexed	Both sexes
Instances	Defined (3)
Array	No

535,380 items of data are available, covering 498,621 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.

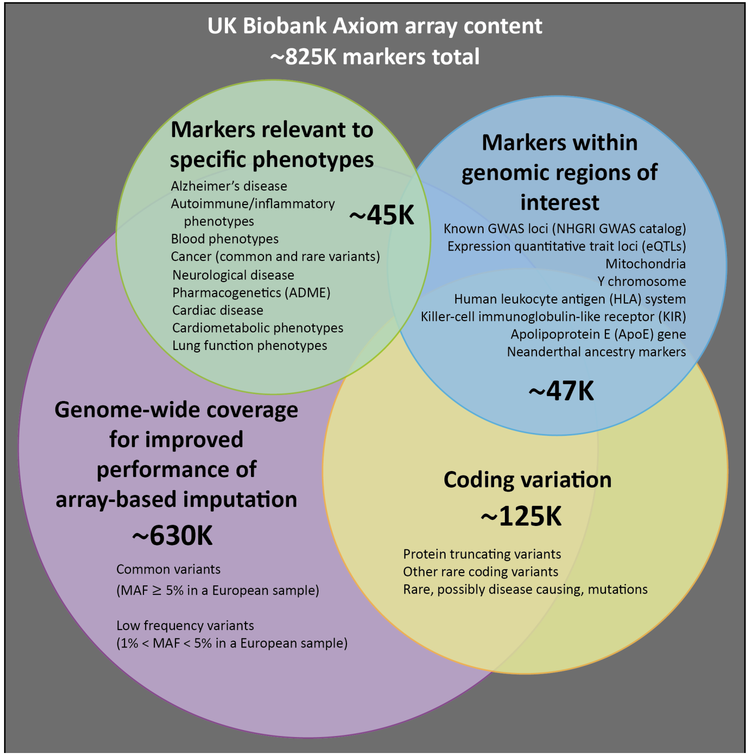
Maximum	24
Decile 9	3
Decile 8	2
Decile 7	1
Decile 6	1
Median	1
Decile 4	1
Decile 3	0
Decile 2	0
Decile 1	0
Minimum	0

- There are 23 distinct values.
- Mean = 1.2724
- Std.dev = 1.52112
- 5233 items above graph maximum of 6
- 109806 items have value -10 (Less than an hour a day)
- 1598 items have value -3 (Prefer not to answer)
- 3243 items have value -1 (Do not know)

Counts of participants/items last updated 04 Feb 2017.

Improving the health of future generations

Genotypic Data



UK Biobank Axiom array content
~825K markers total
Markers relevant to specific phenotypes
Alzheimer's disease
Autoimmune/inflammatory phenotypes
Blood phenotypes
Cancer (common and rare variants)
Neurological disease
Pharmacogenetics (ADME)
Cardiac disease
Cardiometabolic phenotypes
Lung function phenotypes
~45K
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

825,000 markers genotyped

Imputation to HRC
Plus 1KG + UK10K in works
96 million variants

Bycroft et al. on biorxiv

Genome-wide association analysis of 2,400 traits!

How did we do it?

Team hail

Alex Bloemendal Comp Bio

Jon Bloom
Comp Bio

Cotton Seed
Software Eng
Team lead

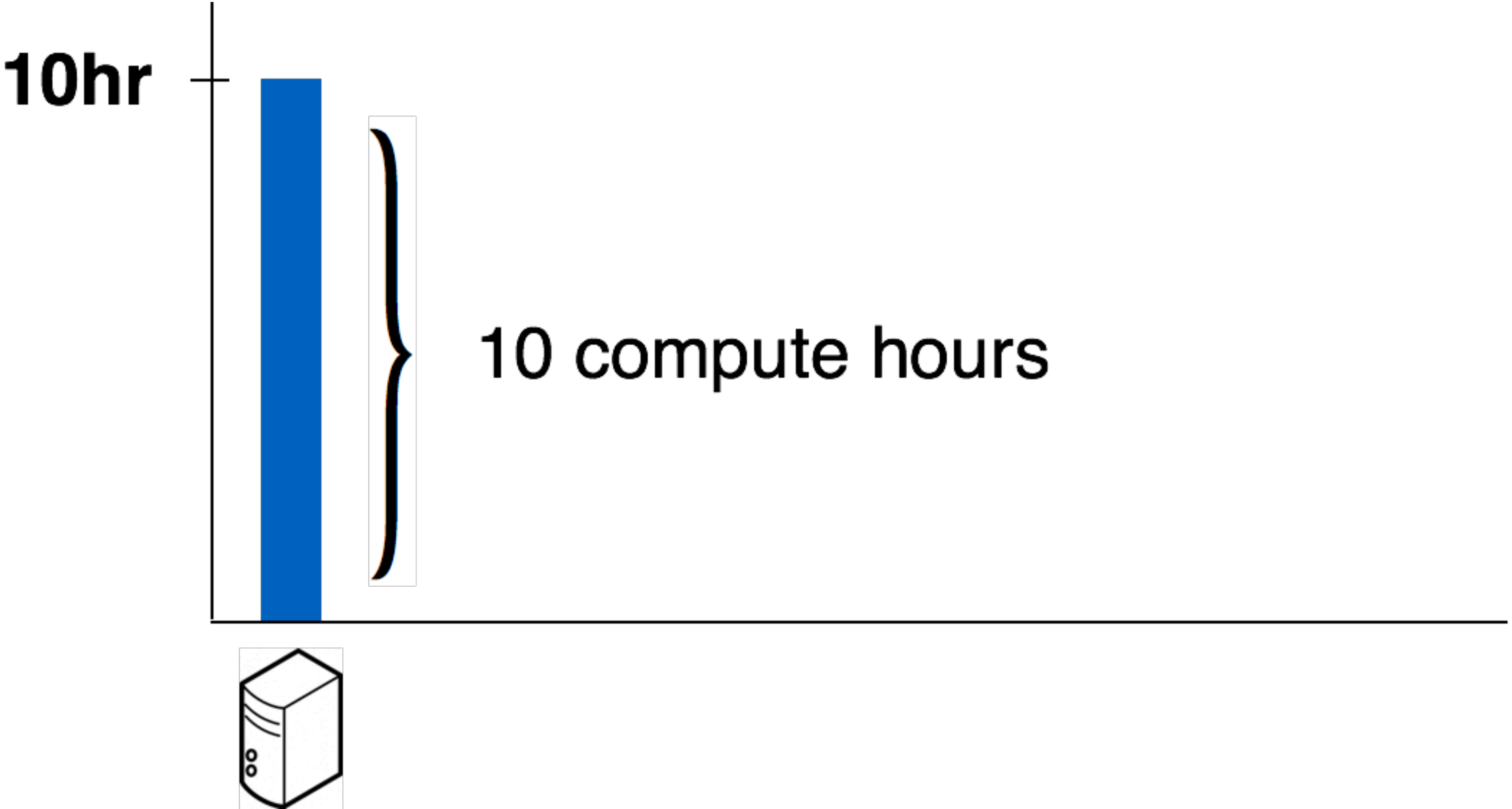
Jackie Goldstein
Software Eng

Amanda Wang
Software Eng

Tim Poterba
Software Eng

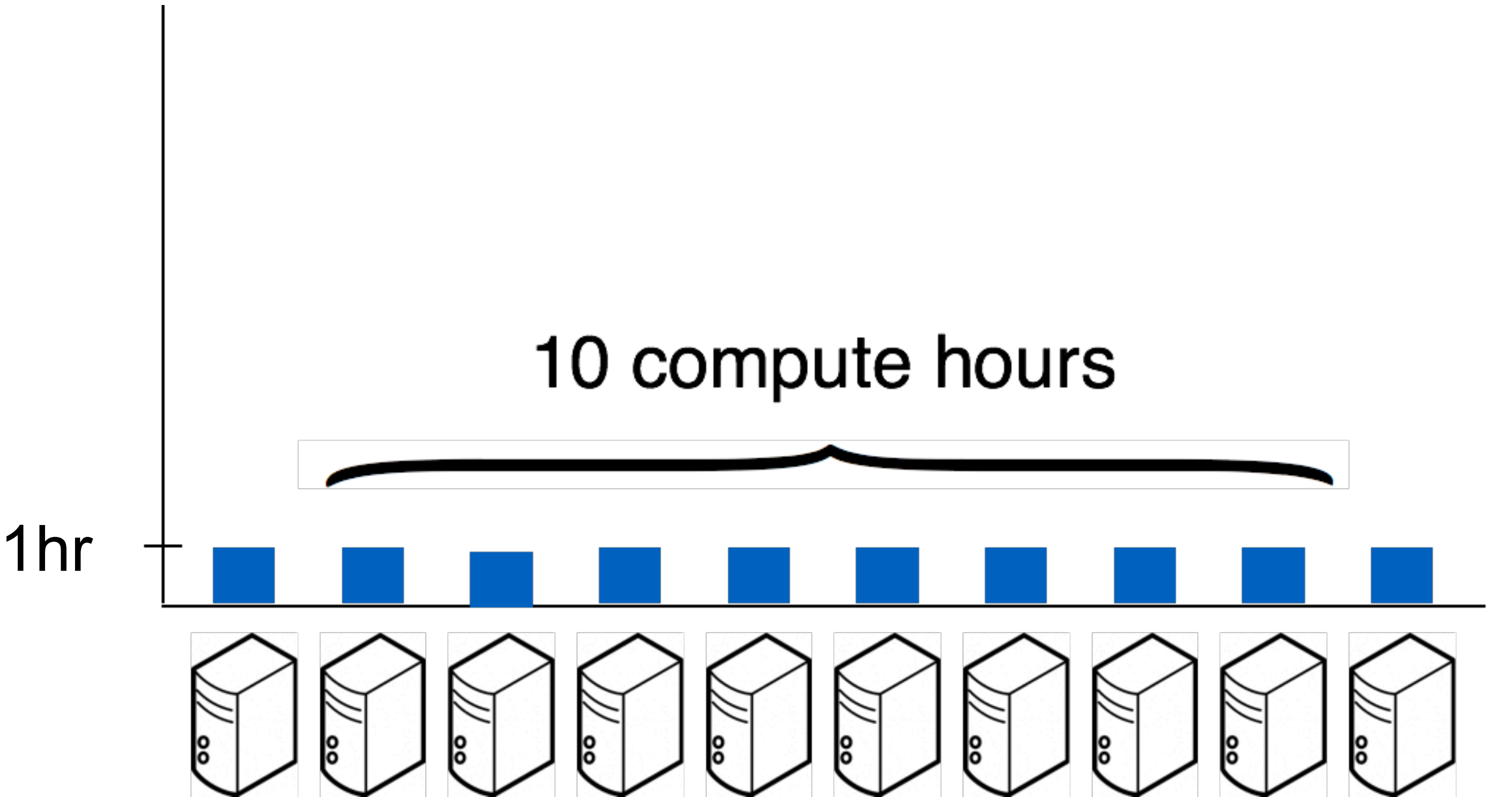
Daniel King
Software Eng

Scalability

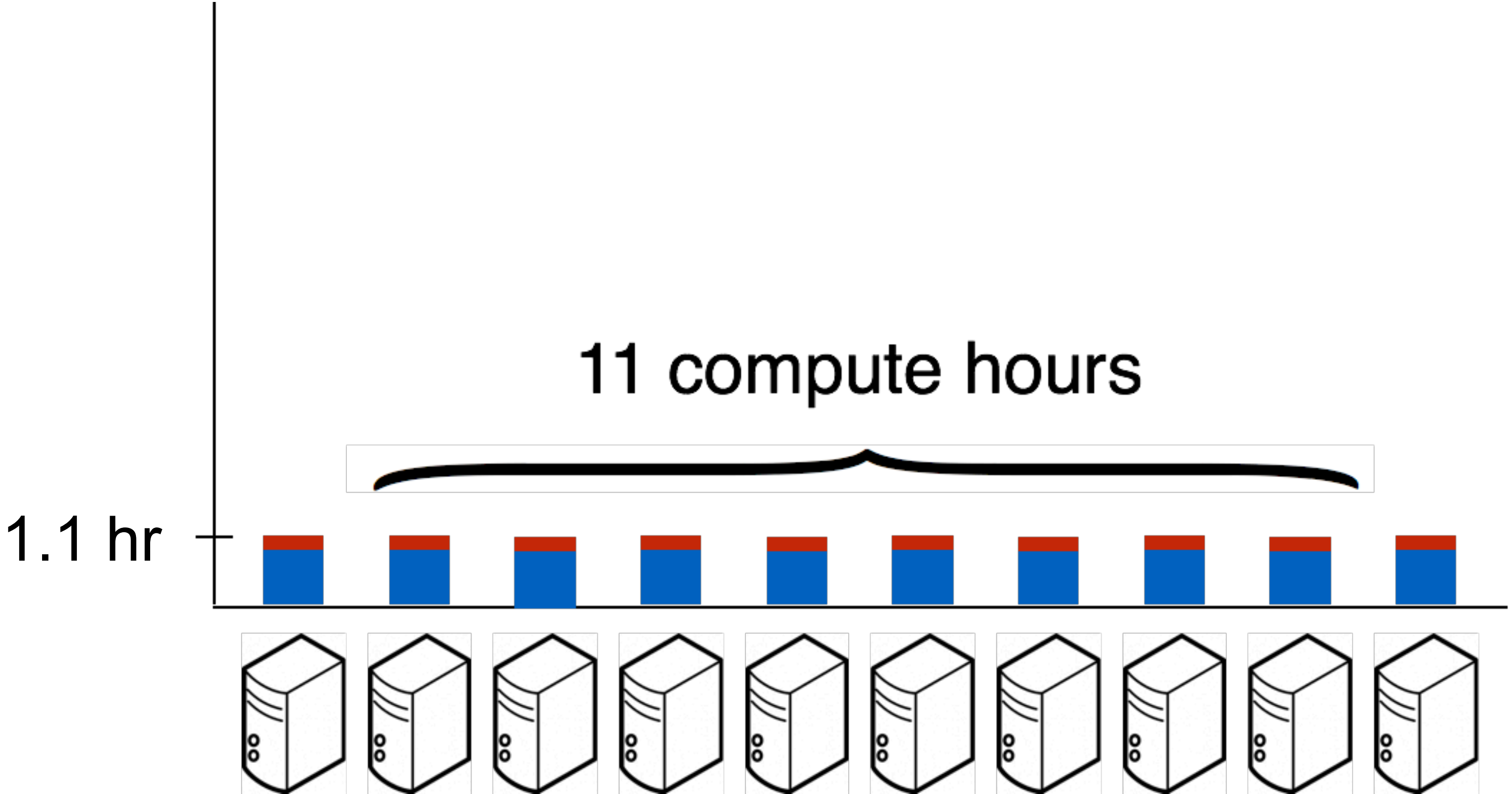


10 compute hours

Scalability

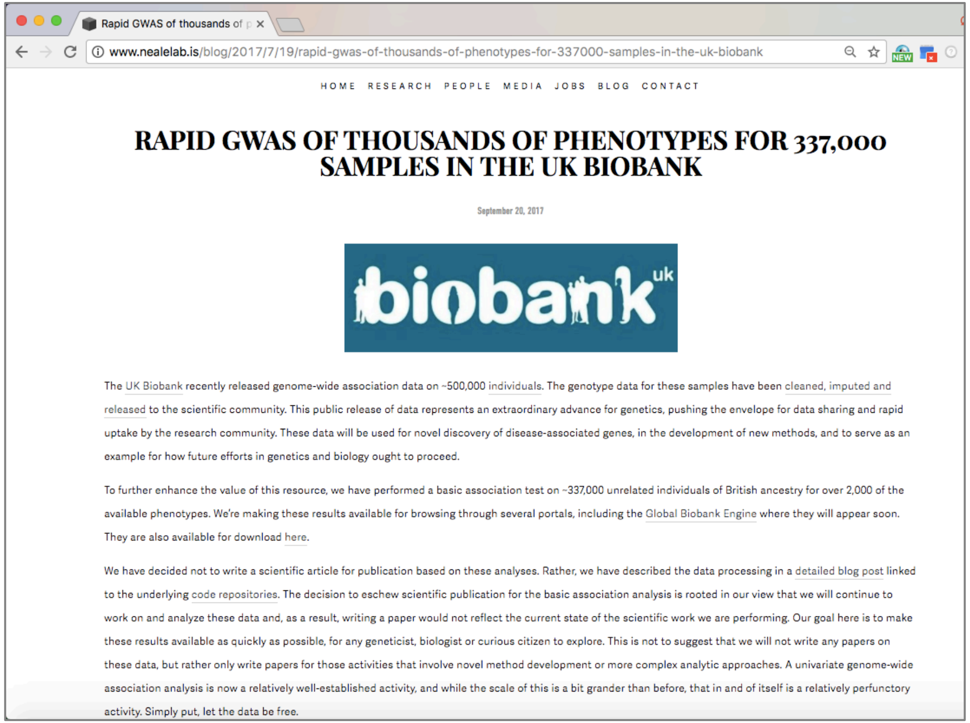


Scalability



Creating a Resource of GWAS Results

- **Enhance** the value of UK Biobank
- **Public** and **easily accessible**
  - www.nealelab.is/blog
- **Blog posts and code** available on GitHub
  - Remove time lag seen with publication
  - Better reflect updates and developments
  - Share how we performed analyses
  - Publish on novel methods and downstream analyses


Rapid GWAS of thousands of p
www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank
HOME RESEARCH PEOPLE MEDIA JOBS BLOG CONTACT
biobank uk

RAPID GWAS OF THOUSANDS OF PHENOTYPES FOR 337,000 SAMPLES IN THE UK BIOBANK

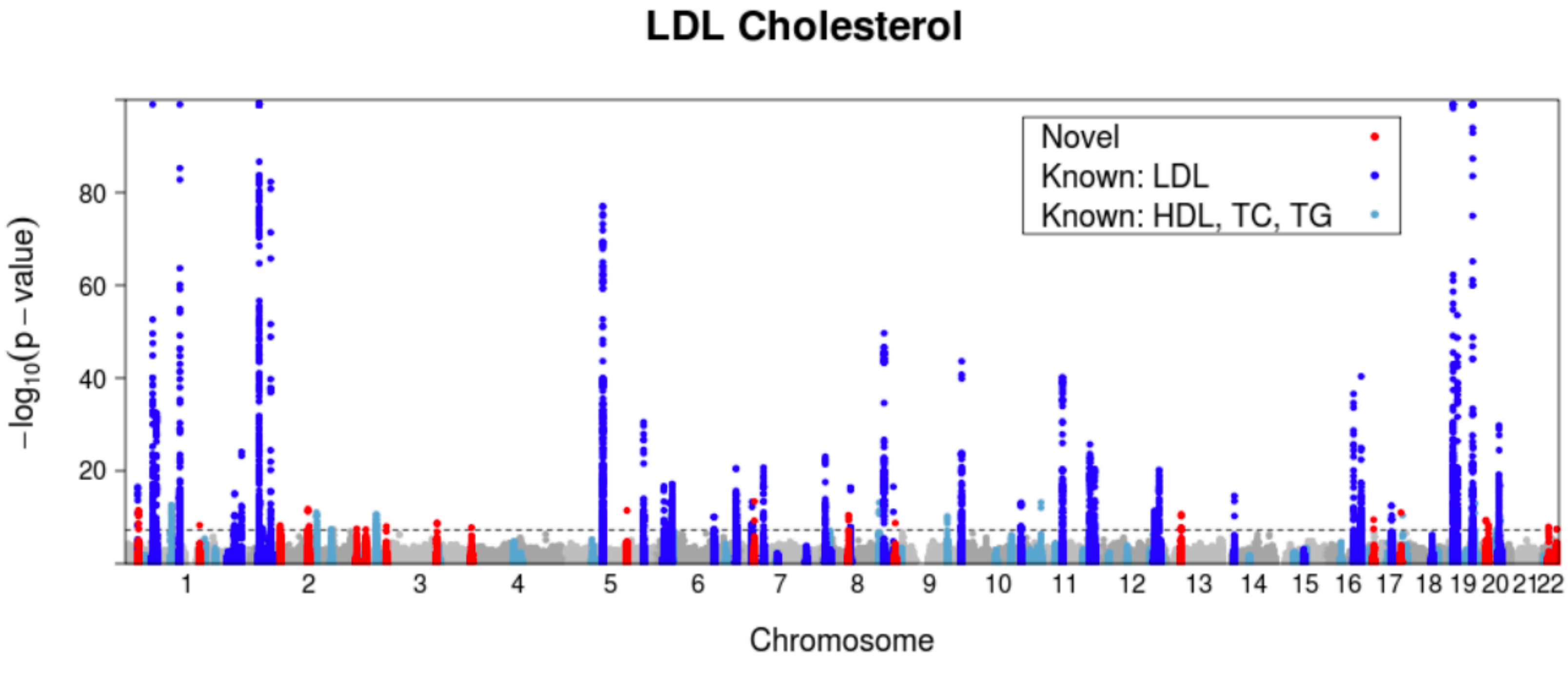
September 20, 2017

The UK Biobank recently released genome-wide association data on ~500,000 individuals. The genotype data for these samples have been cleaned, imputed and released to the scientific community. This public release of data represents an extraordinary advance for genetics, pushing the envelope for data sharing and rapid uptake by the research community. These data will be used for novel discovery of disease-associated genes, in the development of new methods, and to serve as an example for how future efforts in genetics and biology ought to proceed.

To further enhance the value of this resource, we have performed a basic association test on ~337,000 unrelated individuals of British ancestry for over 2,000 of the available phenotypes. We're making these results available for browsing through several portals, including the Global Biobank Engine where they will appear soon. They are also available for download here.

We have decided not to write a scientific article for publication based on these analyses. Rather, we have described the data processing in a detailed blog post linked to the underlying code repositories. The decision to eschew scientific publication for the basic association analysis is rooted in our view that we will continue to work on and analyze these data and, as a result, writing a paper would not reflect the current state of the scientific work we are performing. Our goal here is to make these results available as quickly as possible, for any geneticist, biologist or curious citizen to explore. This is not to suggest that we will not write any papers on these data, but rather only write papers for those activities that involve novel method development or more complex analytic approaches. A univariate genome-wide association analysis is now a relatively well-established activity, and while the scale of this is a bit grander than before, that in and of itself is a relatively perfunctory activity. Simply put, let the data be free.

Association results for taking cholesterol lowering meds



GLGC Nat Genet

Online resources

gnomAD browser beta
About
Downloads
Terms
Contact
Jobs
FAQ

gnomAD browser beta | genome Aggregation Database

Search for a gene or variant or region

Example - Gene: PCSK9, Variant: 1-55516888-G-G

Ab

The Genome Aggregation Database (gnomAD investigators, with the goal of aggregating an from a wide variety of large-scale sequencing scientific community.

The data set provided on this website spans sequences from unrelated individuals sequen genetic studies. The gnomAD Principal Invest current release are listed here.

All data here are released for the benefit of th

Inflammatory Bowel Disease Exomes Browser

Examples - Gene: NOD2, Transcript: ENST00000407236, Variant: 16-50745656-G-A, RS ID: rs104895438, Region: 16:50727514-50766988

Meta-ar

This browser presents IBD case-cont cohorts, each of which contains sam tallies of samples included in the curr

- Ashkenazi Jews (AJ): 2641 IBD
- Non-Finnish Europeans (NFE): 4
- Finland (FINN): 696 IBD (210 CD

A full listing of collaborative partners

Global Biobank Engine (pre-alpha)

Search for a gene or variant or region or phenotype coding (coming soon)

Examples - Gene: F5, Transcript: ENST00000367797, Variant: 1:169519049, RS ID: rs6025, Region: 10:114686614-114786614

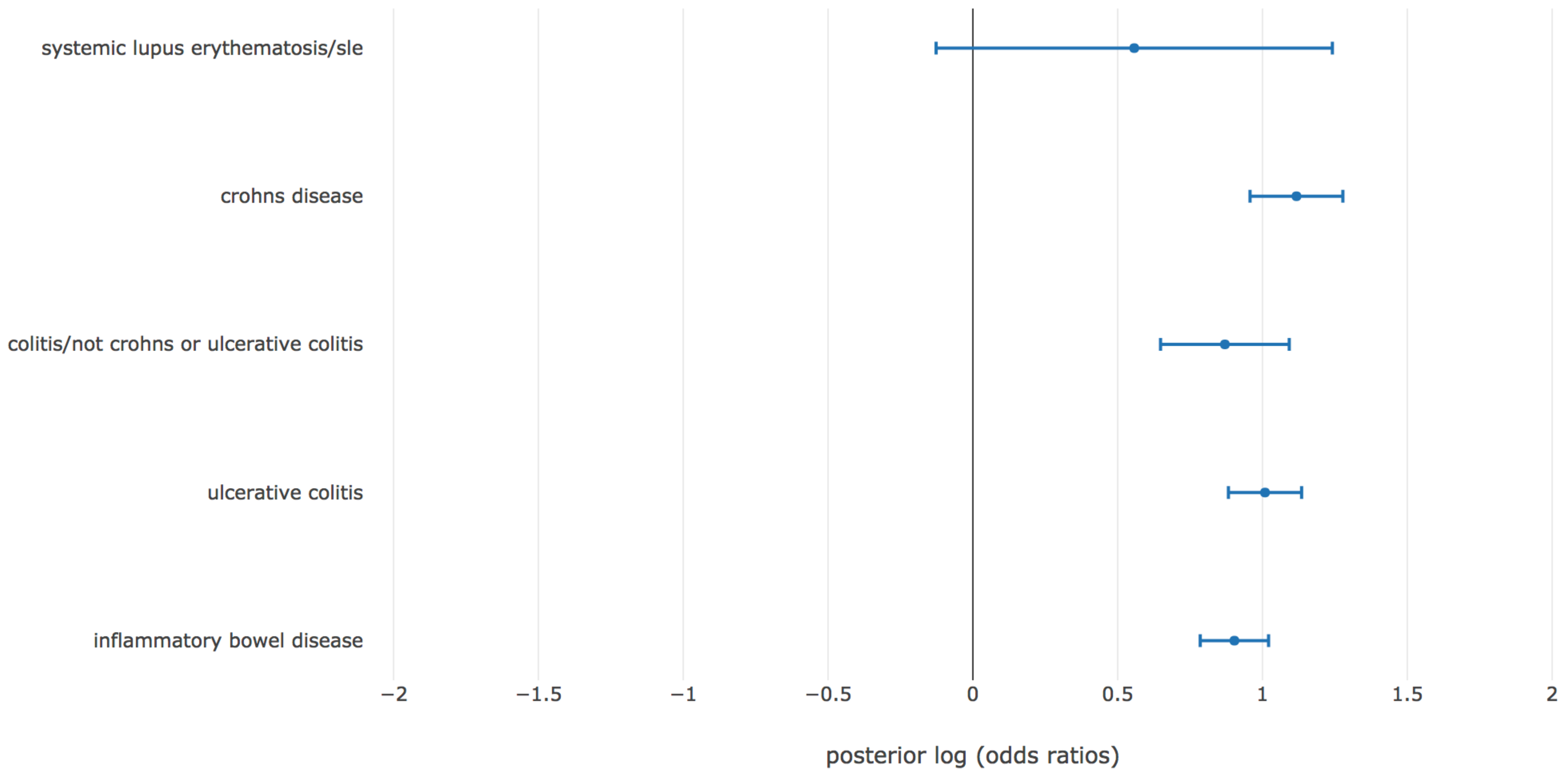
Genetic Association Results

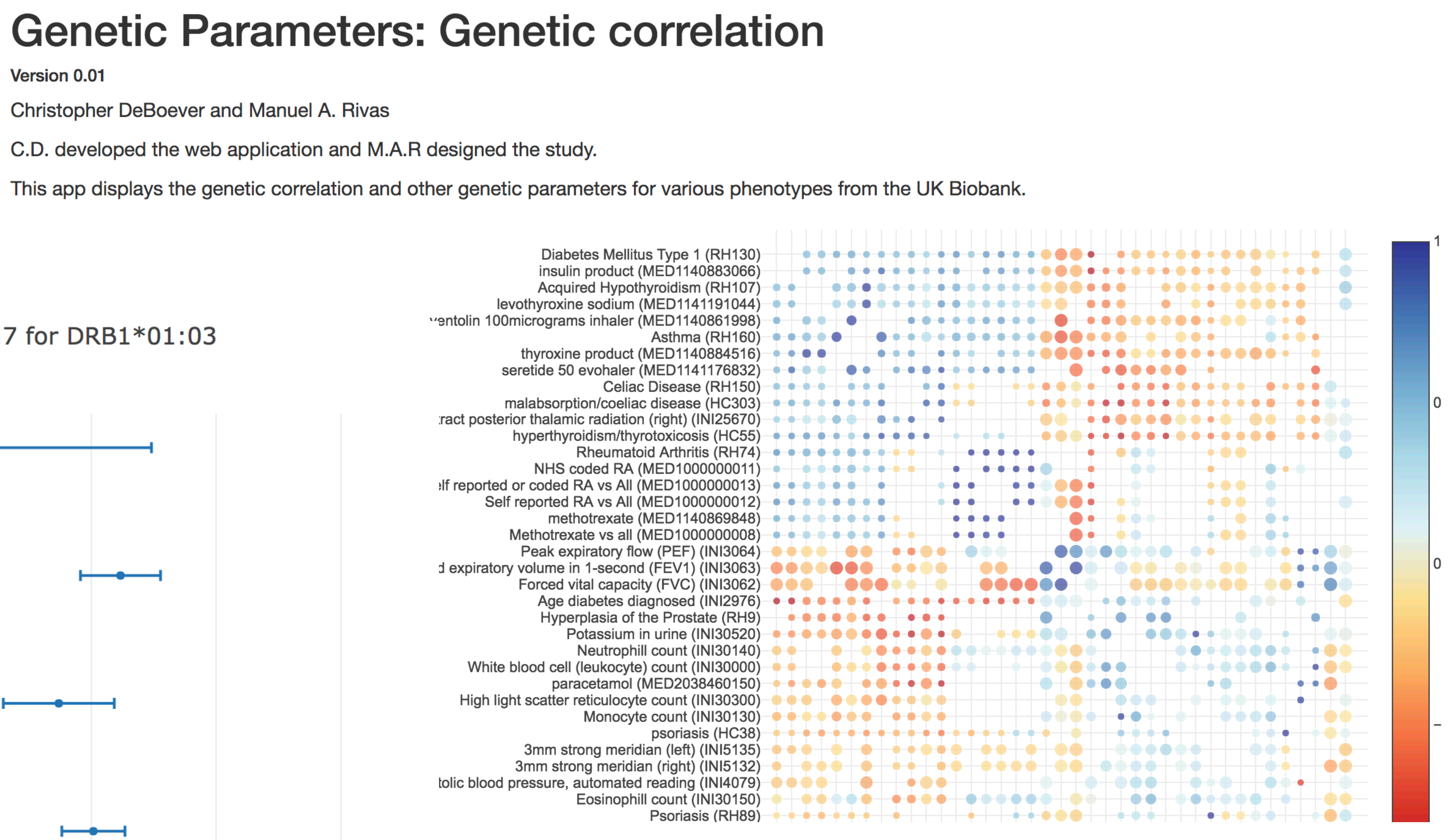
Note: We present summary statistic results from the UK Biobank hospital in-patient health-related outcomes summary information data (Data-Field 41202); computational grouping of phenotypes with

Genetic parameters page

HLA Map

Posterior log(odds ratios) for phenotypes with posterior probability >0.7 for DRB1*01:03





Genetic Parameters: Genetic correlation

Version 0.01

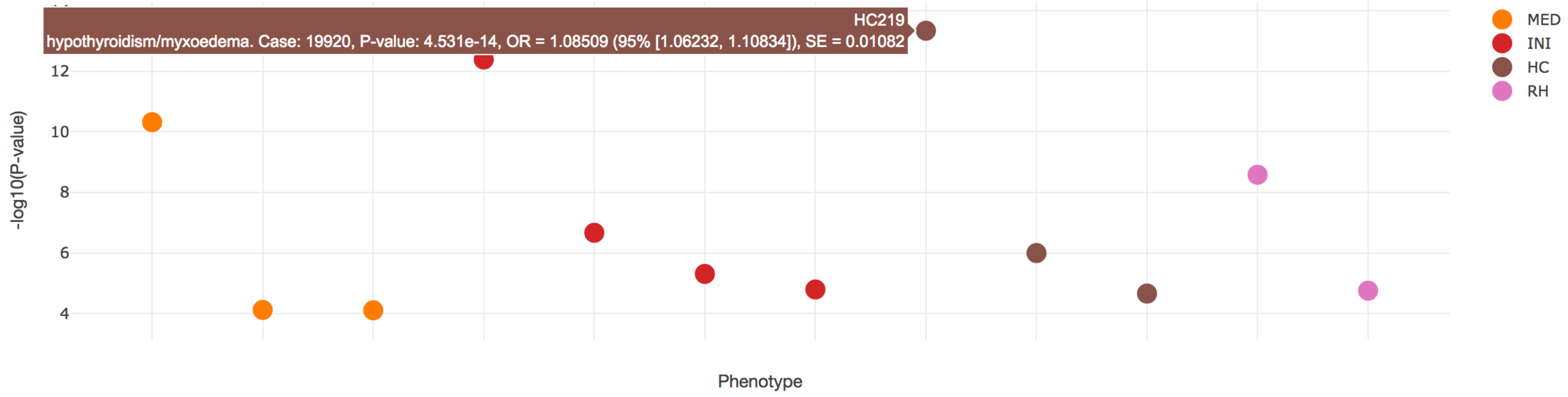
Christopher DeBoever and Manuel A. Rivas

C.D. developed the web application and M.A.R designed the study.

This app displays the genetic correlation and other genetic parameters for various phenotypes from the UK Biobank.

Diabetes Mellitus Type 1 (RH130)
insulin product (MED1140883066)
Acquired Hypothyroidism (RH107)
levothyroxine sodium (MED1141191044)
entolin 100micrograms inhaler (MED1140861998)
Asthma (RH160)
thyroxine product (MED1140884516)
seretide 50 evohaler (MED1141176832)
Celiac Disease (RH150)
malabsorption/coeliac disease (HC303)
ract posterior thalamic radiation (right) (INI25670)
hyperthyroidism/thyrotoxicosis (HC55)
Rheumatoid Arthritis (RH74)
NHS coded RA (MED1000000011)
elf reported or coded RA vs All (MED1000000013)
Self reported RA vs All (MED1000000012)
methotrexate (MED1140869848)
Methotrexate vs all (MED1000000008)
Peak expiratory flow (PEF) (INI3064)
d expiratory volume in 1-second (FEV1) (INI3063)
Forced vital capacity (FVC) (INI3062)
Age diabetes diagnosed (INI2976)
Hyperplasia of the Prostate (RH9)
Potassium in urine (INI30520)
Neutrophill count (INI30140)
White blood cell (leukocyte) count (INI30000)
paracetamol (MED2038460150)
High light scatter reticulocyte count (INI30300)
Monocyte count (INI30130)
psoriasis (HC38)
3mm strong meridian (left) (INI5135)
3mm strong meridian (right) (INI5132)
tolic blood pressure, automated reading (INI4079)
Eosinophill count (INI30150)
Psoriasis (RH89)
1
0
0
–

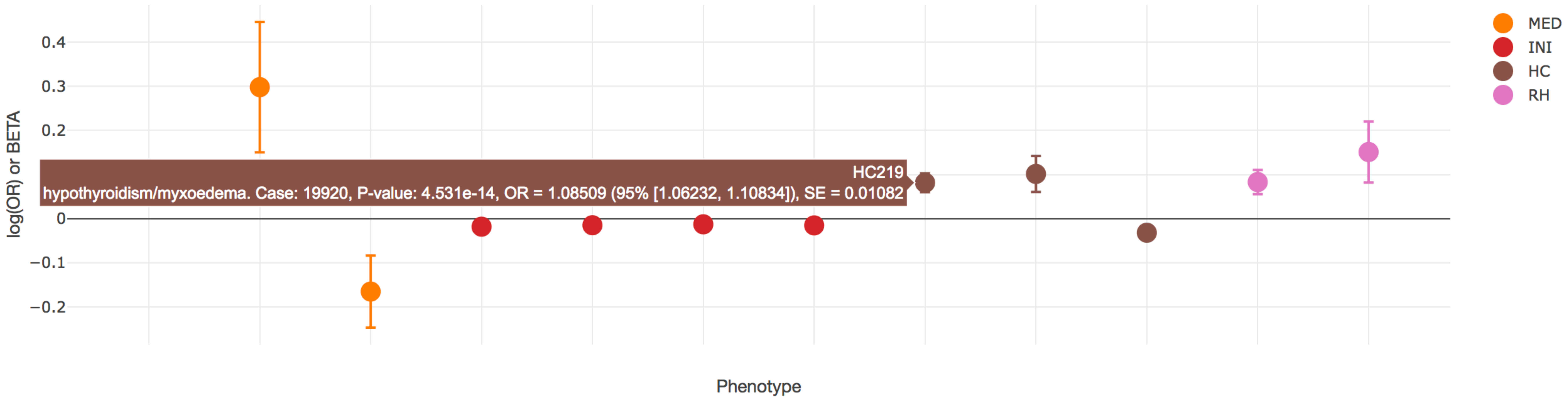
PheWAS page



HC219
hypothyroidism/myxoedema. Case: 19920, P-value: 4.531e-14, OR = 1.08509 (95% [1.06232, 1.10834]), SE = 0.01082

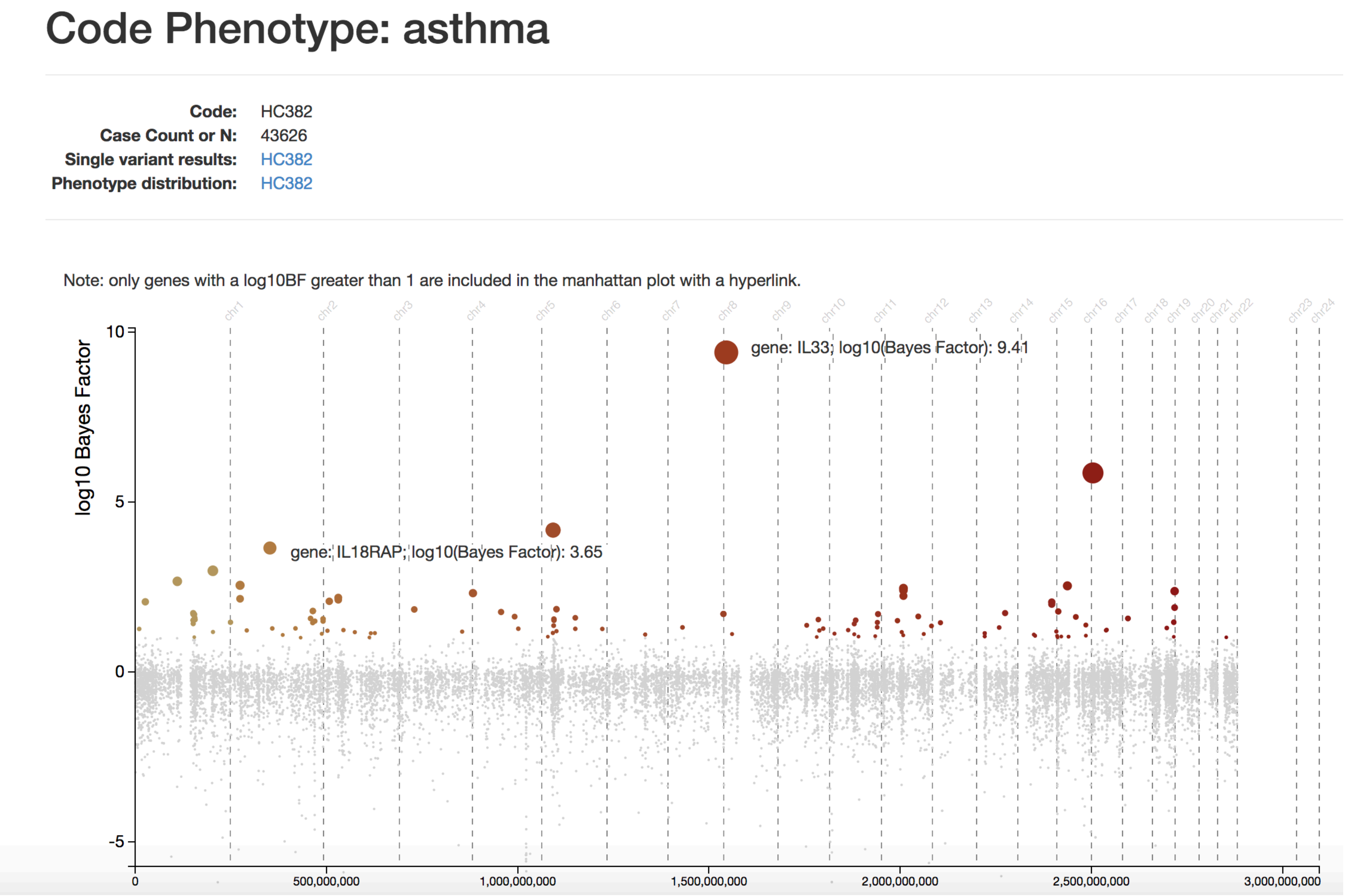
“PheWAS” page

log(OR) or BETA of PheWAS (chr2:163124051)



HC219
hypothyroidism/myxoedema. Case: 19920, P-value: 4.531e-14, OR = 1.08509 (95% [1.06232, 1.10834]), SE = 0.01082

Rare variant aggregate analysis



Code Phenotype: asthma

Code:	HC382
Case Count or N:	43626
Single variant results:	HC382
Phenotype distribution:	HC382

Note: only genes with a log10BF greater than 1 are included in the manhattan plot with a hyperlink.

Rare variant aggregate analysis

Goals of Annotation

- **Annotate the genome comprehensively to**
  - Boost the power of rare variant association analysis
  - Assist with fine mapping to identify possible causal variants
  - Study population genetics
  - Improve the interpretability of whole-genome data.
- **Annotations at different levels: SNV, SV, Indel, gene**
- **Annotate CCDG and CMG data, e.g. CCDG Freeze 1 data (n=24K)**

GSP and TOPMed Collaboration on Annotation

- GSP Annotation WG (ACs, CCDGs, CMGs)
- TOPMed Annotation Interest WG
- Collaboration between GSP and TOPMed , with the goal of consistent annotations across both data sets.
- NHLBI RFI on Strategically Critical Resources or Infrastructures Using R24/U24
- Make annotation resources to the whole GSP community, as well as the general research community, .e.g., the NIH Data Commons Use Cases.

Types of Raw Annotation Useful for Disease Mapping in WGS

- **MAFs**
- **Variant types**
  - VEP annotation using Genecode, e.g., LOF, missense, synonymous, 5”, 3”
  - Promoters using FANTOM5
  - Enhancers using GeneHancer
- **Protein scores (coding variants only),** e.g., Polyphen and SIFT
- **Evolutionary/conservation scores**, e.g, Priphylop, GerpN, GerpS
- **Epigenetic scores** (ENCODE, ROADMAP)(cell specific), e.g. H3AK27ac, H3K4me1
- **ChromHMM States**

Types of Raw Annotation Useful for Disease Mapping in WGS: Continues

- **LD/heritability-related scores**, e.g., bStatistics, recombination rates
- **Higher-order chromatin interactions**, e.g., Hi-C
- **Transcpriptome scores (eQTLs)** (GTEx, n too small for rare variant eQTLs)
- **Methylation scores (mQTLs)**
- **Post-translational/protein (pQTL) scores** (UniProt)
- **Phenotype-specific heritability**
- **Structural Variant and Indel annotation**

Composite Annotation Scores

- Existing (mainly driven by conservation scores for non-coding variants)
  - CADD/EIGEN (similar)
  - LINSIGHT (not defined for coding variants)
  - FATHMM-XF
  - GenoCanyon and GenoSkyline (cell/tissue specific)
- Newer scores
  - PINES (epigenetic-score based)
  - LACE (two dimension scores: epigenetics and conservation)
  - Annotation PCs

Population Genetics annotation

- **Variant level -** additional population and evolutionary scores (i.e. Fst; recurrent mutations; allele age estimates)

- **Locus level -** local ancestry inference; local IBD inference; local recombination rate; haplotype selection scores (i.e. B-statistics)

- **Individual level -** self-reported ancestry, global genetic ancestry, inbreeding coefficient

- **Population level -** global genetic ancestry; IBD community; Founder effects (IBDNe, ROH), admixture;