IEU GWAS Database

Biogen presentation: Thursday 13th February 2020

Denis Baird, Yi Liu and *Ben Elsworth*

https://github.com/MRCIEU/ieu-gwas-db-demo
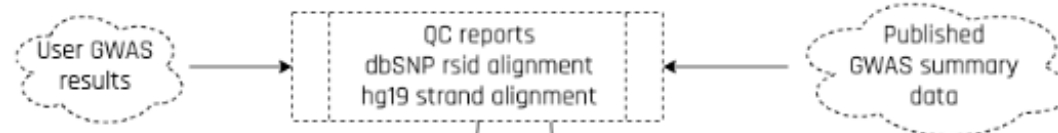
Overview of talk

- Introduction to IEU GWAS Database

- Worked examples:
  - Lookup selected SNPs for selected traits (for example to run MR)
  - Lookup SNPs within a genomic region (for example to run coloc)
  - Lookup association across all traits for one SNP (PheWas)
  - Download copy of full GWAS summary statistics

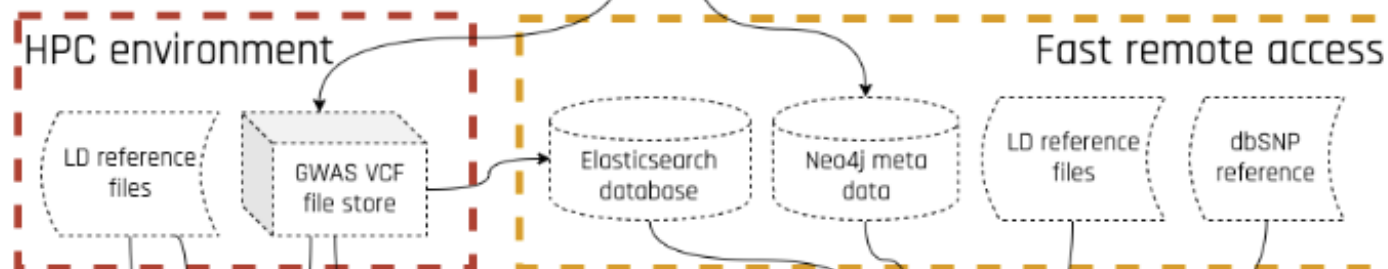- Questions

Description of database and packages

- Large collection of curated GWAS summary statistics

- Three ways to access IEU GWAS Database:
  1. Web interface (https://gwas.mrcieu.ac.uk/)
  2. Database can be queried using ieugwasr R package (https://github.com/mrcieu/ieugwasr)
  3. RESTful API (http://gwas-api.mrcieu.ac.uk/docs/)

- Additional functions to help integrate output with other software
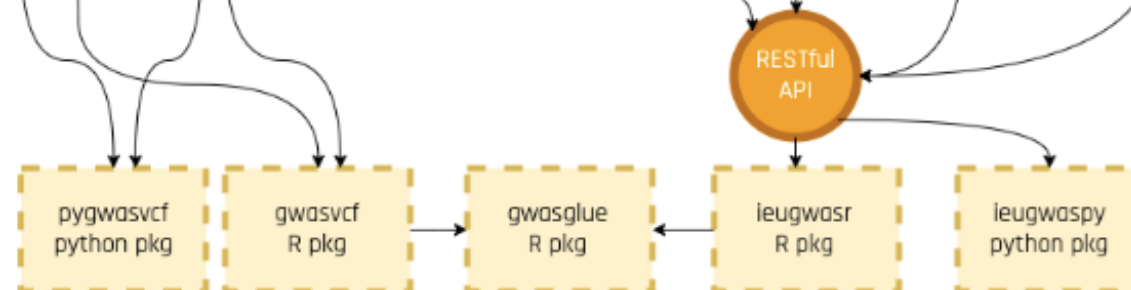 - GWASglue R package (currently under development)
    - https://github.com/mrcieu/gwasglue

GWAS summary data ecosystem

Data ingestion

User GWAS results → QC reports / dbSNP rsid alignment / hg19 strand alignment ← Published GWAS summary data

Data storage

HPC environment
- LD reference files
- GWAS VCF file store

Fast remote access
- Elasticsearch database
- Neo4j meta data
- LD reference files
- dbSNP reference

RESTful API

Data retrieval

- pygwasvcf python pkg
- gwasvcf R pkg
- gwasglue R pkg
- leugwasr R pkg
- leugwaspy python pkg

Analytical tools

Genetic architecture	Mendelian randomization	Finemapping	Colocalisation
LDSC	TwoSampleMR	FINEMAP	coloc
PRSice	MendelianRandomization	PAINTOR	HEIDI
	RadialMR	CAVIAR	eCAVIAR
	TRYX	GCTA-COJO	S-Predixcan
	MRMix	ABF	
	MR-PRESSO	JAM	
	GSMR	SuSiE	
	mr.raps		

Text mining

Vectology

MELODI

IEU GWAS DB

EpiGraphDB

MRC IEU GWAS Pipeline

Genetic

MR-EvE

PART_OF

Vectology:
http://vectology.mrcieu.ac.uk/

MELODI:
http://melodi.biocompute.org.uk/

EpigraphDB:
http://epigraphdb.org/

IEU GWAS DB:
http://gwas.mrcieu.ac.uk/

GWAS Pipeline:
https://data.bris.ac.uk/data/dataset/pnoat8cxo0u52p6ynfaekeigi

MR-EvE:
https://www.biorxiv.org/content/10.1101/173682v2

Database content

- >11,000 traits

- >200 diseases

- Much more to come

Data overview

The GWAS summary datasets are currently organised into the following 7 batches:

Batch	Description	Count
ebi-a	Datasets that satisfy minimum requirements imported from the EBI database of complete GWAS summary data	293
ieu-a	GWAS summary datasets generated by many different consortia that have been manually collected and curated, initially developed for MR-Base	1,193
prot-a	Complete GWAS summary data on protein levels as described by Sun et al 2018	3,282
prot-b	Complete GWAS summary data on protein levels as described by Folkersen et al 2017	83
ubm-a	Complete GWAS summary data on brain region volumes as described by Elliott et al 2018	3,143
ukb-a	Neale lab analysis of UK Biobank phenotypes, round 1	596
ukb-b	IEU analysis of UK Biobank phenotypes	2,514

Major changes to the IEU GWAS resources for 2020
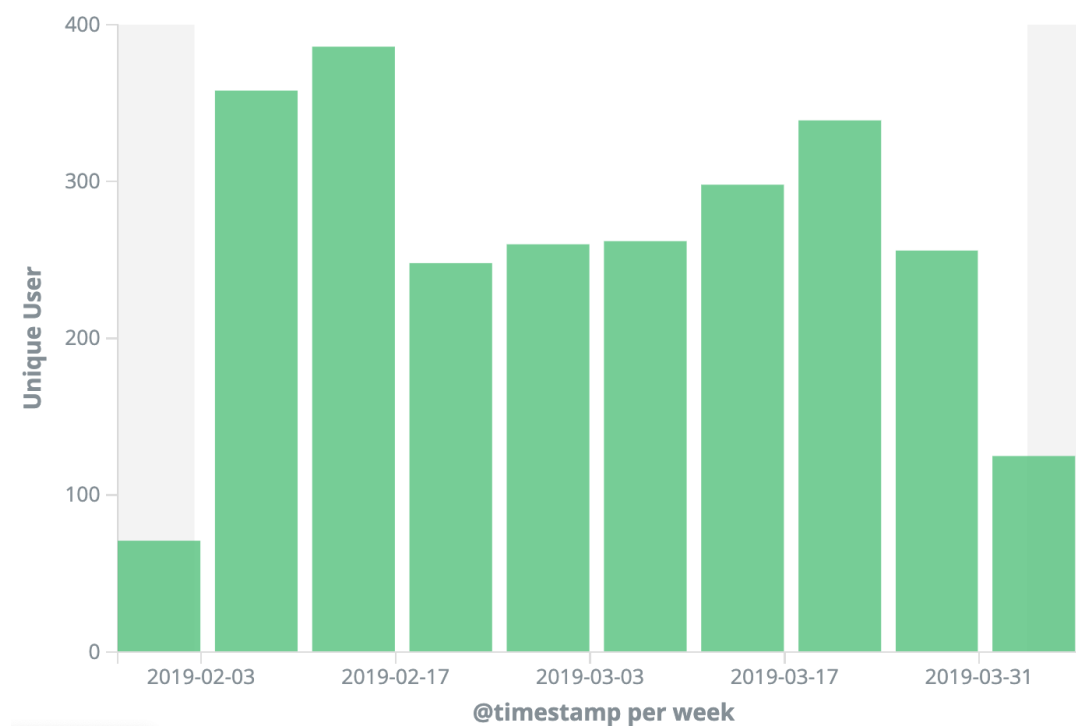
https://mrcieu.github.io/TwoSampleMR/articles/gwas2020.html
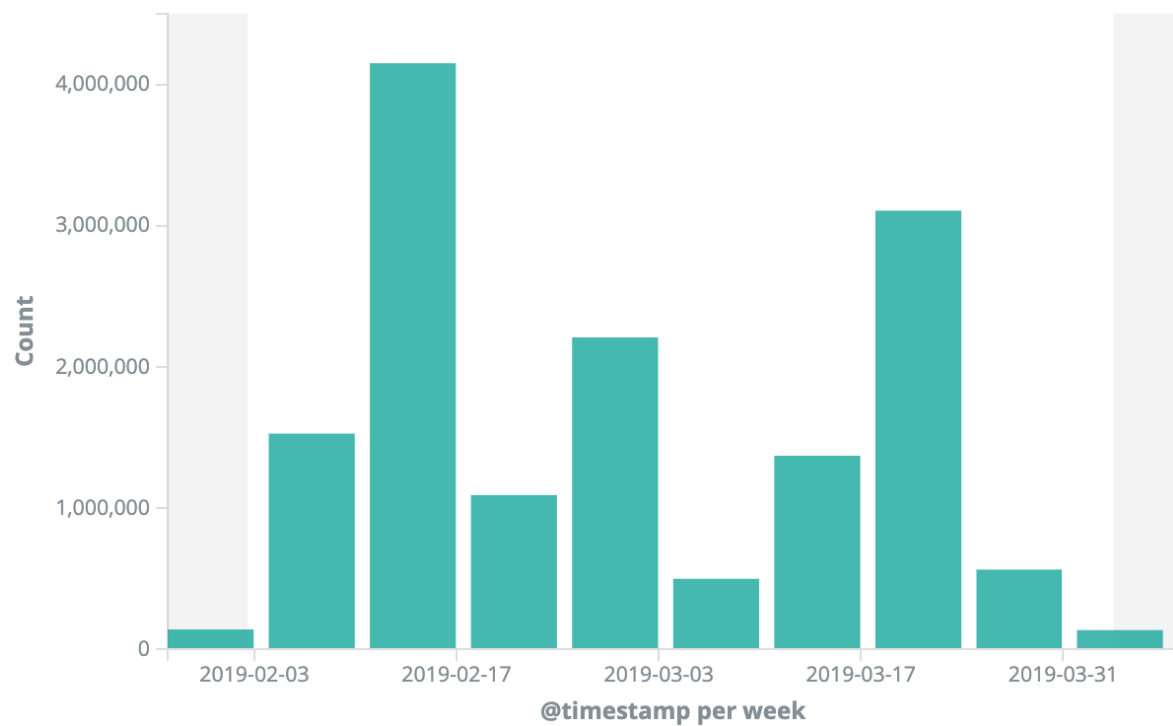
User statistics

Worked example one

Find associations for selected SNPs and traits:

Mendelian Randomization between protein expression loci and Parkinson's Disease

Steps to conduct MR

1. Find the database ids for the exposure (pQTL) and outcome (PD) GWAS.

2. Extract the top hits from the exposure GWAS.

3. LD clump the top hits to obtain instruments.

4. Extract the instruments from the outcome GWAS.

5. Harmonize the exposure and outcome SNP effects.

6. Run the MR on the harmonized effects.

Step	Package	Function
1. Find the exposure and outcome GWAS info.	**Web** OR ieugwasr	gwasinfo
2. Extract SNPs from exposure GWAS.	ieugwasr	**tophits (also LD clumps SNPs so can skip step 3)** OR associations
3. LD clump SNPs.	ieugwasr OR TwoSampleMR	ld_clump clump_data
4. Extract SNPs from outcome GWAS.	ieugwasr	associations
5. Harmonize SNP effects.	TwoSampleMR	harmonise_data
6. Run MR.	TwoSampleMR	mr, **mr_single_snp**

See https://mrcieu.github.io/ieugwasr/articles/guide.html for full documentation on ieugwasr package.

See https://mrcieu.github.io/TwoSampleMR/ for full documentation on TwoSampleMR package.

Find the exposure GWAS in database

https://gwas.mrcieu.ac.uk/

https://gwas.mrcieu.ac.uk/datasets/

GWAS ID	Year	Trait	Consortium	Sample size	Number of SNPs

Clear filters/sorting

Data overview

The GWAS summary datasets are currently organised into the following 7 batches:

Batch	Description	Count
ebi-a	Datasets that satisfy minimum requirements imported from the EBI database of complete GWAS summary data	293
ieu-a	GWAS summary datasets generated by many different consortia that have been manually collected and curated, initially developed for MR-Base	1,193
prot-a	Complete GWAS summary data on protein levels as described by Sun et al 2018	3,282
prot-b	Complete GWAS summary data on protein levels as described by Folkersen et al 2017	83
ubm-a	Complete GWAS summary data on brain region volumes as described by Elliott et al 2018	3,143
ukb-a	Neale lab analysis of UK Biobank phenotypes, round 1	596
ukb-b	IEU analysis of UK Biobank phenotypes	2,514

Click here to navigate to publication

Genomic atlas of the human plasma proteome.

Sun BB[1], Maranville JC[2,3], Peters JE[1,4], Stacey D[1], Staley JR[1], Blackshaw J[1], Burgess S[1,5], Jiang T[1], Paige E[1,6], Surendran P[1], Oliver-Williams C[1,7], Kamat MA[1], Prins BP[1], Wilcox SK[8], Zimmerman ES[8], Chi A[2], Bansal N[1,9], Spain SL[10], Wood AM[1], Morrell NW[4,11], Bradley JR[12], Janjic N[8], Roberts DJ[13,14], Ouwehand WH[4,15,16,17,18], Todd JA[19], Soranzo N[4,15,17,18], Suhre K[20], Paul DS[1], Fox CS[2], Plenge RM[2,3], Danesh J[21,22,23,24], Runz H[2,25], Butterworth AS[26,27].

⊕ Author information

Abstract

Although plasma proteins have important roles in biological processes and are the direct targets of many drugs, the genetic factors that control inter-individual variation in plasma protein levels are not well understood. Here we characterize the genetic architecture of the human plasma proteome in healthy blood donors from the INTERVAL study. We identify 1,927 genetic associations with 1,478 proteins, a fourfold increase on existing knowledge, including trans associations for 1,104 proteins. To understand the consequences of perturbations in plasma protein levels, we apply an integrated approach that links genetic variation with biological pathway, disease, and drug databases. We show that protein quantitative trait loci overlap with gene expression quantitative trait loci, as well as with disease-associated loci, and find evidence that protein biomarkers have causal roles in disease using Mendelian randomization analysis. By linking genetic factors to diseases via specific proteins, our analyses highlight potential therapeutic targets, opportunities for matching existing drugs with new disease indications, and potential safety concerns for drugs under development.

Step 1. Find the outcome GWAS.

Type outcome here. NB: need to hit clear filters button beforehand.

Search

GWAS ID: [   ]   Trait contains: [Parkinson   ]   [Filter]   Total of 11,016 records.

Year: [   ]   Consortium contains: [   ]

GWAS ID	Year	Trait	Consortium	Sample size	Number of SNPs
ebi-a-GCST006414	2018	Atrial fibrillation	NA	1,030,836	33,519,037
ieu-a-1239	2018	Years of schooling	SSGAC	766,345	10,101,242
ebi-a-GCST006867	2018	Type 2 diabetes	NA	655,666	5,030,727
ebi-a-GCST005195	2017	Coronary artery disease	NA	547,261	7,934,254
ebi-a-GCST006061	2018	Atrial fibrillation	NA	537,409	12,095,506

1 2 3 4 5 6 7 8 ... 2204 next »

GWAS ID:		Trait contains:		Filter	Filtered to 5 records
Year:		Consortium contains:			of a total of 11,016.

GWAS ID	Year	Trait	Consortium	Sample size	Number of SNPs
ukb-b-6548	2018	Illnesses of mother: Parkinson's disease	MRC-IEU	422,464	9,851,867
ukb-b-956	2018	Illnesses of father: Parkinson's disease	MRC-IEU	399,089	9,851,867
ukb-b-16943	2018	Illnesses of siblings: Parkinson's disease	MRC-IEU	361,199	9,851,867
ieu-a-812	2009	Parkinson's disease	—	5,691	453,218
ieu-a-818	2011	Parkinson's disease	—	1,672	318,553

Clear filters/sorting

Illnesses of Mother: Parkinson's Disease

Year	2018
Category	Binary
Sub category	NA
Population	European
Sex	Males and Females
ncase	6,998
ncontrol	415,466
Sample size	422,464
Number of SNPs	9,851,867
Unit	SD
Priority	1
Author	Ben Elsworth
Consortium	MRC-IEU
Note	20110#11: Output from GWAS pipeline using Phesant derived variables from UKBiobank

Top 30 related datasets ℹ️

ukb-b-956: Illnesses of father: Parkinson's disease

ukb-b-16943: Illnesses of siblings: Parkinson's disease

ukb-a-210: Illnesses of mother: Alzheimer's disease/dementia

ukb-b-14699: Illnesses of mother: Alzheimer's disease/dementia

ukb-a-214: Illnesses of mother: Chronic bronchitis/emphysema

ukb-b-12018: Illnesses of mother: Chronic bronchitis/emphysema

ukb-a-209: Illnesses of mother: Heart disease

ukb-b-12477: Illnesses of mother: Heart disease

ieu-a-812: Parkinson's disease

ieu-a-818: Parkinson's disease

ukb-b-10807: Illnesses of mother: Severe depression

ukb-b-323: Illnesses of father: Alzheimer's disease/dementia

ukb-b-4024: Illnesses of mother: Stroke

ukb-a-216: Illnesses of mother: Diabetes

Worked example two

Find associations within a genomic region surrounding a SNP

Steps to lookup region around a SNP

1. Find genomic coordinates (hg19) of SNP.
   - variants_rsid in ieugwasr package

2. Define region around SNP (chrom and hg19 pos +- 500kb)

3. Lookup this region
   1. Associations in ieugwasr package

Web-Based Genome-Wide Association Study Identifies Two Novel Loci and a Substantial Genetic Component for Parkinson's Disease

Chuong B. Do ✉, Joyce Y. Tung, Elizabeth Dorfman, Amy K. Kiefer, Emily M. Drabant, Uta Francke, Joanna L. Mountain,
Samuel M. Goldman, Caroline M. Tanner, J. William Langston, Anne Wojcicki, Nicholas Eriksson ✉

Article	Authors	Metrics	Comments	Media Coverage

Abstract

Author Summary

Introduction

Results

Discussion

Materials and Methods

Supporting Information

Acknowledgments

Author Contributions

References

Reader Comments (2)

Media Coverage (2)

Figures

Abstract

Although the causes of Parkinson's disease (PD) are thought to be primarily environmental, recent studies suggest that a number of genes influence susceptibility. Using targeted case recruitment and online survey instruments, we conducted the largest case-control genome-wide association study (GWAS) of PD based on a single collection of individuals to date (3,426 cases and 29,624 controls). We discovered two novel, genome-wide significant associations with PD–rs6812193 near $SCARB2$ ($p = 7.6 \times 10^{-10}$, OR $= 0.84$) and rs11868035 near $SREBF1/RAI1$ ($p = 5.6 \times 10^{-8}$, OR $= 0.85$)—both replicated in an independent cohort. We also replicated 20 previously discovered genetic associations (including $LRRK2$, GBA, $SNCA$, $MAPT$, GAK, and the HLA region), providing support for our novel study design. Relying on a recently proposed method based on genome-wide sharing estimates between distantly related individuals, we estimated the heritability of PD to be at least 0.27. Finally, using sparse regression techniques, we constructed predictive models that account for 6%–7% of the total variance in liability and that suggest the presence of true associations just beyond genome-wide significance, as confirmed through both internal and external cross-validation. These results indicate a substantial, but by no means total, contribution of genetics underlying susceptibility to both early-onset and late-onset PD, suggesting that, despite the novel associations discovered here and elsewhere, the majority of the genetic component for Parkinson's disease remains to be discovered.

Associated SNPs from PD GWAS:
rs6812193
rs11868035

Worked example three

Find associations across the phenome for a SNP of interest

GWAS summary data.

e.g. Body mass index, rs1000940

A database of 111,217,925,970 genetic associations from 11,104 GWAS summary datasets, for querying or download.

See the API page for fast programmatic options to query the data, including R, python and HPC environments.

Use the gwasglue R package to apply the data to Mendelian randomization, fine mapping, colocalisation, etc.

PheWAS

Info

Search

Examples

Info

Perform a Phenome-wide association study (PheWAS), which entails searching for the effects of a genetic variant across all publicly available datasets. Enter an rs ID for a variant of interest in the search box below.

```
rs6812193
```
Search

Enter the rsid and hit return key.

Examples

Some example variants taken from SNPedia.

- rs53576 in the oxytocin receptor influences social behavior and personality
- rs1815739 muscle performance
- rs7412 and rs429358 can raise the risk of Alzheimer's disease by more than 10x
- rs6152 can influence baldness
- rs333 resistance to HIV

Search: [   ]

ID	Trait	Position	P	SE	N	Beta	CHR	EA	NEA	EAF
ieu-a-1105	Serum creatinine (eGFRcrea)	77198986	4.10015e-13	0.00092	133723	0.0066	4	T	C	0.398
ieu-a-1104	Serum creatinine (eGFRcrea)	77198986	1.10002e-12	0.00092	118373	0.0066	4	T	C	0.398
ebi-a-GCST004603	Platelet count	77198986	0.00000156498	0.00376057	166066	-0.018061	4	T	C	0.3756
ebi-a-GCST004607	Plateletcrit	77198986	0.00000349704	0.00377232	164339	-0.0175006	4	T	C	0.3755
ukb-b-16056	Diagnoses - secondary ICD10: F10.1 Harmful use	77198986	0.00000969996	0.000133663	463010	0.00059132	4	T	C	0.375243
ieu-a-850	Creatinine	77198986	0.0000143001	0.00999	22583	-0.043452	4	T	C	0.349173
prot-a-449	CD44 antigen	77198986	0.000047863	0.0253	3301	0.1027	4	T	C	0.36663
ieu-a-812	Parkinson's disease	77198986	0.0000666807	0.0431662	5691	0.173045	4	T	C	
ukb-b-12301	Treatment/medication code: salbutamol	77198986	0.000129999	0.000232824	462933	0.000891984	4	T	C	0.37525
ukb-b-12828	Leg fat-free mass (right)	77198986	0.00016	0.00134226	454835	0.00505964	4	T	C	0.375218

Showing 1 to 10 of 39 entries

Previous | 1 | 2 | 3 | 4 | Next

Human GRCh38/hg38 [4:76,177,833..76,377,833]

200bp | 2kb | 50kb | 500kb

✖ Genome
76,180,000 | 76,200,000 | 76,220,000 | 76,240,000 | 76,260,000 | 76,280,000 | 76,300,000 | 76,320,000 | 76,340,000 | 76,360,000

✖ GENCODE ⊕
<SCARB2   >FAM47E   >FAM47E-STBD1   <CCDC158   >SNX5P1 <RNU6-1000P

✖ Repeats
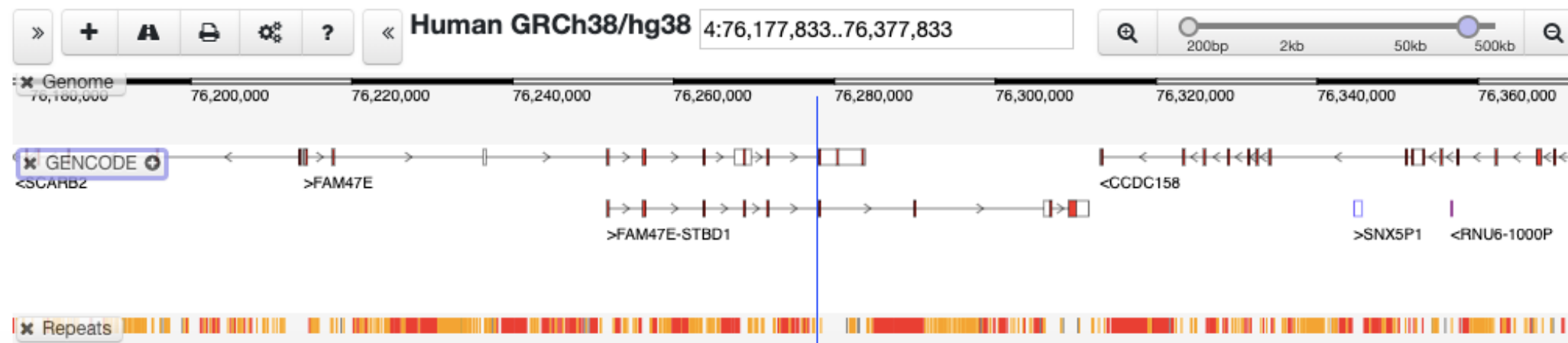
rs11868035 SNP lookup

ID	Trait	Position	P	SE	N	Beta	CHR	EA	NEA	EAF
ukb-b-14068	Impedance of leg (left)	17715101	6.79986e-11	0.00200404	454857	0.0130785	17	A	G	0.293525
ebi-a-GCST004627	Lymphocyte counts	17715101	1.167e-9	0.00398072	171643	0.0242215	17	A	G	0.29
ukb-b-7376	Impedance of leg (right)	17715101	1.09999e-8	0.00198726	454863	0.0113478	17	A	G	0.293524
ukb-b-5776	Daytime dozing / sleeping (narcolepsy)	17715101	3.79997e-8	0.00113542	460913	-0.00624353	17	A	G	0.293419
ukb-b-5237	Coffee intake	17715101	4.79999e-8	0.00177861	428860	-0.00970978	17	A	G	0.293248
ukb-a-271	Impedance of leg (left)	17715101	1.0466e-7	0.00244328	331296	-0.0129947	17	A	G	0.708054
ebi-a-GCST006250	Intelligence	17715101	4.41896e-7	0.00294305	269867	-0.0148624	17	A	G	
ukb-a-199	Mean time to correctly identify matches	17715101	7.82817e-7	0.00264841	335139	-0.0130823	17	A	G	0.708054
ukb-b-19373	Duration to first press of snap-button in each round	17715101	9.29994e-7	0.00224511	459281	0.011013	17	A	G	0.293422
ebi-a-GCST004625	Monocyte count	17715101	0.000001191	0.00394808	170721	0.0191763	17	A	G	0.2901

Showing 1 to 10 of 94 entries

Previous | 1 | 2 | 3 | 4 | 5 | ... | 10 | Next

Human GRCh38/hg38 | 17:17,711,787..17,911,787
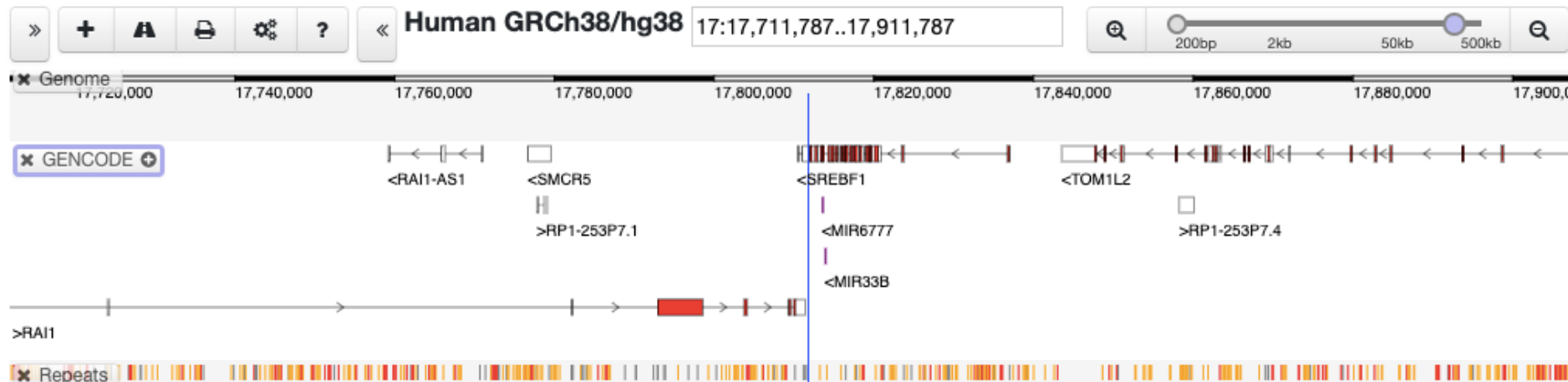
Worked example four

Download full summary statistics for a GWAS from database

Search

GWAS ID: [_____] Trait contains: [_____] **Filter** Filtered to 1 records of a total of 11,016.

Year: [_____] Consortium contains: [_____]

GWAS ID	Year	Trait	Consortium	Sample size	Number of SNPs
ukb-b-6548	2018	Illnesses of mother: Parkinson's disease	MRC-IEU	422,464	9,851,867

Clear filters/sorting

Click here

Click here

Illnesses of mother: Parkinson's disease
Dataset: ukb-b-6548

Download VCF   Download index

Year	2018
Category	Binary
Sub category	NA
Population	European
Sex	Males and Females
ncase	6,998

Top 30 related datasets ⓘ

ukb-b-956: Illnesses of father: Parkinson's disease

ukb-b-16943: Illnesses of siblings: Parkinson's disease

ukb-a-210: Illnesses of mother: Alzheimer's disease/dementia

ukb-b-14699: Illnesses of mother: Alzheimer's disease/dementia

ukb-a-214: Illnesses of mother: Chronic bronchitis/emphysema

See (https://github.com/mrcieu/gwasvcf) for guidance on manipulation of VCF format file.

Acknowledgements

MRC-IEU

- Ben Elsworth
- Denis Baird
- Chris Zheng
- Gibran Hemani
- Matt Lyon
- Peter Matthews
- Philip Haycock
- Valeriia Haberland
- Yi Liu
- Tom Gaunt

University of Bristol Research IT

- Tessa Alexander
- Jon Hallett