The GWAS Catalog

Accessing GWAS Catalog summary statistics

Elliot Sollis

Senior Curator, GWAS Catalog

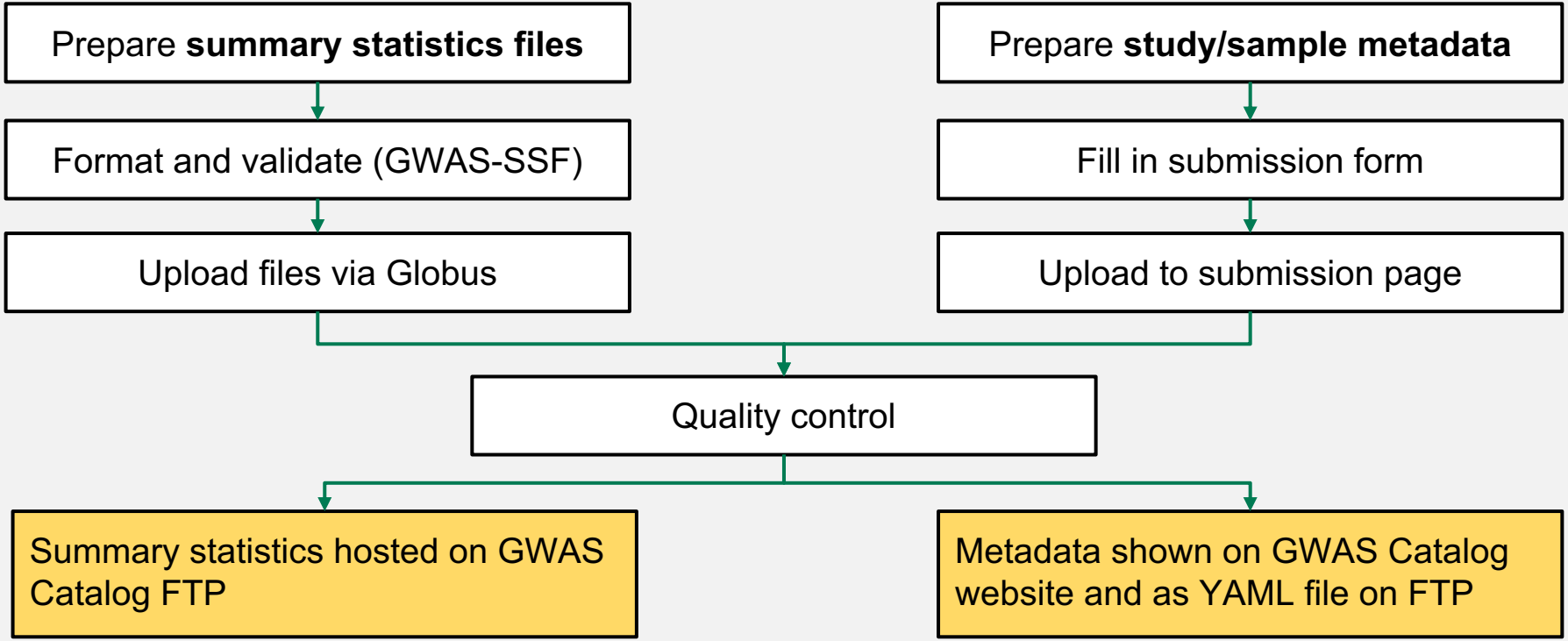
I have no conflict of interest to declare

EMBL-EBI

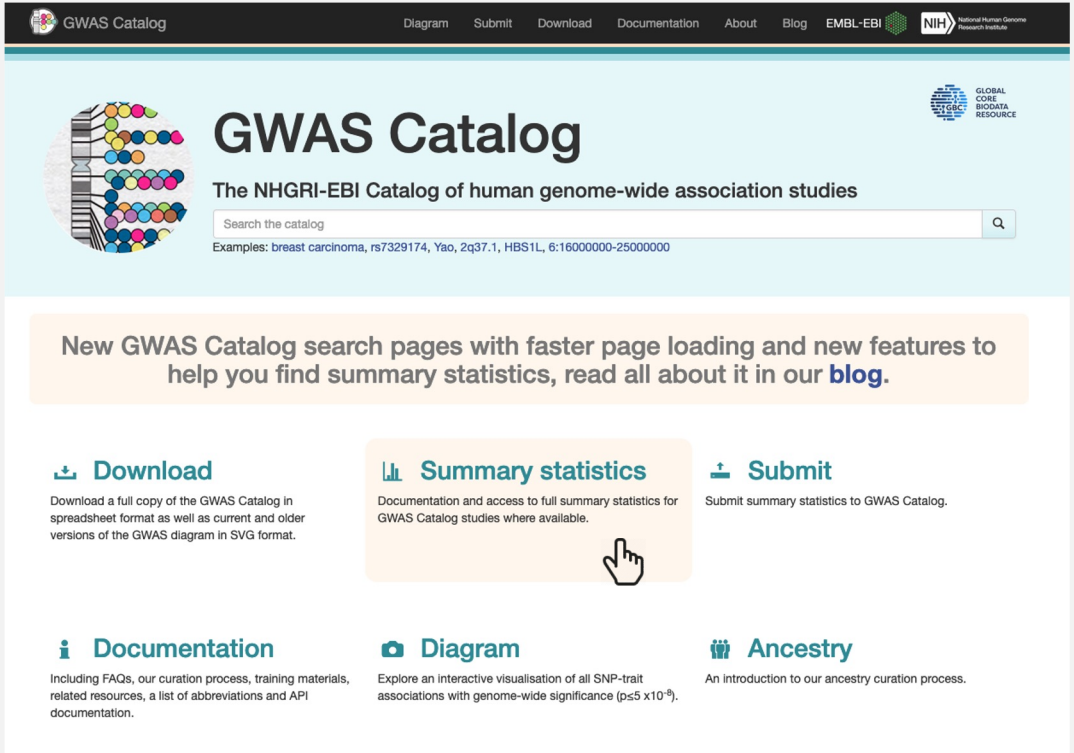
Outline

- How to access summary statistics
- Metadata YAML files
- Summary statistics harmonisation
- Summary statistics API

Submission process

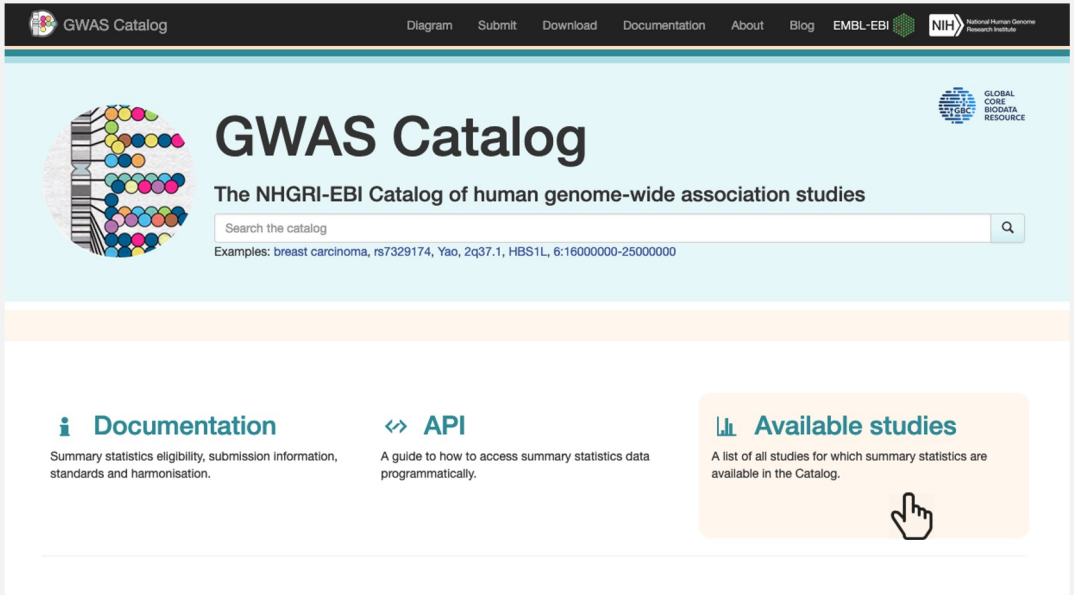


How to access summary statistics



https://www.ebi.ac.uk/gwas

How to access summary statistics



All published studies with summary statistics

List of published studies with summary statistics

Studies 58828

Show 5 entries

Column visibility | Export | Clear search

First author	PubMed ID	Study accession	Pub. date	Journal	Title	Reported trait	Trait(s)	Data access
Soranzo N	20858683	GCST000803	2010-09-21	Diabetes	Common variants at 10 genomic loci influence…	Glycated hemoglobin levels	HbA1c measurement	FTP Download
Barrett JC	19430480	GCST000392	2009-05-10	Nat Genet	Genome-wide association study and meta-analysis…	Type 1 diabetes	type 1 diabetes mellitus	FTP Download
den Hoed M	23583979	GCST001969	2013-04-14	Nat Genet	Identification of heart rate-associated loci and…	Heart rate	heart rate	FTP Download
Saxena R	17463246	GCST000028	2007-04-26	Science	Genome-wide association analysis identifies loci…	Type 2 diabetes	type 2 diabetes mellitus	FTP Download
Ferreira MA	19853236	GCST000510	2009-10-22	Am J Hum Genet	Sequence variants in three loci influence…	Platelet count	platelet count	FTP Download

Showing 1 to 5 of 58,828 entries

« 1 2 3 4 5 … 11766 »

EMBL-EBI

Unpublished studies with summary statistics

List of prepublished/unpublished studies with summary statistics

Search

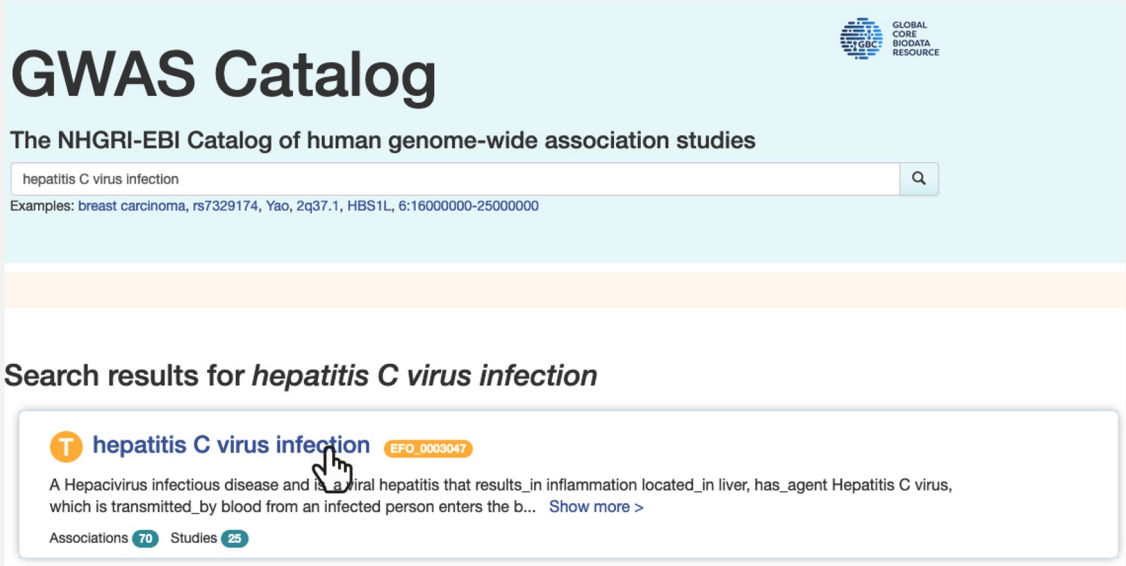
First Author	Date Submitted	Study Accession	Title	Reported trait	Ancestry Category	No of Individuals	FTP Path
Krystyna Taylor	2020-07-09	GCST90002217	Analysis of Genetic Host Response Risk Factors in Severe COVID-19 Patients	Severe response to COVID-19 (Sepsis controls)	European	2083	Download
Krystyna Taylor	2020-07-09	GCST90002218	Analysis of Genetic Host Response Risk Factors in Severe COVID-19 Patients	Severe response to COVID-19 (Sepsis controls)	European	2083	Download
Sayoni Das	2020-07-09	GCST90002219	Identification and Analysis of Shared Risk Factors in Sepsis and High Mortality Risk COVID-19 Patients	Sepsis	European	12696	Download
Sayoni Das	2020-07-09	GCST90002220	Identification and Analysis of Shared Risk Factors in Sepsis and High Mortality Risk COVID-19 Patients	Sepsis	European	12696	Download
Julia M Sealock	2020-11-18	GCST90011993	Clinical laboratory test-wide association scan of polygenic scores identifies biomarkers of complex disease	Alpha 1 antitrypsin [Mass/volume] in Serum or Plasma	European	1756	Download

Showing 1 to 5 of 6880 rows 5 rows per page

‹ 1 2 3 4 5 ... 1376 ›

EMBL-EBI

Search by Publication, Trait etc.



Trait page example

Trait: hepatitis C virus infection

GWAS / Traits / EFO_0003047

Trait information

Trait label	hepatitis C virus infection
EFO ID	EFO_0003047
Synonyms	**17** synonyms +
Mapped terms	**9** mapped terms +
Description	A Hepacivirus infectious disease and is_a viral hepatitis that results_in inflammation located_in liver, has_agent Hepatitis C virus, which is transmitted_by blood from an infected person enters the body of an uninfected person. The infection has_symptom fever, has_symptom fatigue, has_symptom loss of appetite, has_symptom nausea, has_symptom vomiting, has_symptom abdominal pain, has_symptom clay-colored bowel movements, has_symptom joint pain, and has_symptom jaundice. +
Reported Traits	**4** reported traits +
Child traits	**1** child traits +

Trait in **OLS**

Trait in **OXO**

Trait in **Open Targets**

Available data: **Associations** 70 | **Studies** 25 | **Full summary statistics** 4 | **LocusZoom**

Download Associations

☐ **Include background traits data**

☑ **Include child trait data**

EMBL-EBI

Table of studies with summary statistics

Available data: Associations 70 | Studies 25 | Full summary statistics 4 | LocusZoom | Download Associations

☐ Include background traits data
☑ Include child trait data

Studies with summary statistics 4

Show 5 entries | Column visibility | Export | Clear search

First author	Study accession	Pub. date	Journal	Title	Reported trait	Trait(s)	Discovery sample number	Association count	Summary statistics
Jiang L	GCST90041714	2021-11-04	Nat Genet	A generalized linear mixed model association…	Viral hepatitis C (PheCode 70.3)	hepatitis C virus infection	• 456348 European	0	FTP Download
Jiang L	GCST90225539	2022-09-27	BMC Genomics	Genome-wide association analyses of common…	Hepatitis C	hepatitis C virus infection	• 7560 European	0	FTP Download
Sakaue S	GCST90018585	2021-09-30	Nat Genet	A cross-population atlas of genetic associations…	Chronic hepatitis C infection	chronic hepatitis C virus infection	• 176698 East Asian	4	FTP Download
Sakaue S	GCST90018805	2021-09-30	Nat Genet	A cross-population atlas of genetic associations…	Chronic hepatitis C infection	chronic hepatitis C virus infection	• 352013 European • 176698 East Asian	4	FTP Download

Showing 1 to 4 of 4 entries

« 1 »

EMBL-EBI

FTP folder structure

Studies grouped in batches of 1000

Folders named by study accession

Main summary statistics file

Metadata file

Subfolder for harmonised sumstats

Index of /pub/databases/gwas/summary_statistics/GCST90018001-GCST90019000/GCST90018995

Name	Last modified	Size	Description
Parent Directory		-	
GCST90018995_buildGRCh37.tsv.gz	2021-12-16 00:35	296M	
GCST90018995_buildGRCh37.tsv.gz-meta.yaml	2023-07-18 21:48	313	
README.txt	2021-12-16 00:35	177	
harmonised/	2022-01-26 10:38	-	
md5sum.txt	2021-12-16 00:35	109	

Metadata YAML files

Standard metadata format

Metadata YAML file

Basic file information

Genotyping information

Sample information

Trait information

```yaml
coordinate_system: 1-based
data_file_md5sum: d91015c653100b5b2a3c6ee31b80eab2
data_file_name: GCST90274838.tsv.gz
date_last_modified: 2023-07-25
file_type: GWAS-SSF v1.0
genome_assembly: GRCh37
genotyping_technology:
- Genome-wide genotyping array
gwas_id: GCST90274838
samples:
- sample_ancestry:
  - European
  sample_size: 14733
trait_description:
- Protein abundance level
```

Standard metadata format

File type (different formats)

- **GWAS-SSF v1.0**
  - Current standard
- **Pre-GWAS-SSF**
  - Minimal columns requirements
  - May be missing some fields
- **Non-GWAS-SSF**
  - Non-standard summary statistics files, eg. gene-based, CNV


```yaml
coordinate_system: 1-based
data_file_md5sum: d91015c653100b5b2a3c6ee31b80eab2
data_file_name: GCST90274838.tsv.gz
date_last_modified: 2023-07-25
file_type: GWAS-SSF v1.0
genome_assembly: GRCh37
genotyping_technology:
- Genome-wide genotyping array
gwas_id: GCST90274838
samples:
- sample_ancestry:
  - European
  sample_size: 14733
trait_description:
- Protein abundance level
```

Harmonisation

Harmonised summary statistics

To maximise interoperability we also make summary statistics available in a harmonised format wherever possible:

1. All files mapped to the same reference genome (GRCh38)
2. All alleles reported on the reference strand
3. Allele frequency and effect direction matched to reference allele
4. Variants sorted by chromosome and base pair location
5. Quality control to remove variants with missing data after harmonisation

Why use harmonised files?

- Eliminates the need to run your own allele orientation steps
- Useful for various downstream applications:
  - Mendelian randomisation
  - Co-localisation studies
  - Polygenic score development

Accessing harmonised summary statistics

Harmonised files available in subfolder on FTP:

Index of /pub/database
GCST90086001-GCST

Name
Parent Directory
GCST90086099_buildGRCh37.tsv
GCST90086099_buildGRCh37.tsv-meta.yar
harmonised/
md5sum.txt

Index of /pub/databases/gwas/summary_statistics/GCST90086001-GCST90087000/GCST90086099/harmonised

Name	Last modified	Size	Description
Parent Directory		-	
GCST90086099.h.tsv.gz	2024-02-08 14:51	551M	
GCST90086099.h.tsv.gz-meta.yaml	2024-02-08 14:51	674	
GCST90086099.h.tsv.gz.tbi	2024-02-08 14:51	1.5M	
GCST90086099.running.log	2024-02-08 14:51	2.3K	
md5sum.txt	2024-02-08 14:51	116	

Contents of the harmonised folder

a. The **harmonised summary statistics** (*.h.tsv.gz)
b. An **index file** (*.h.tsv.gz.tbi) - used for quick data retrieval
c. A **report file** (report.txt) - summarises harmonisation methods, variants dropped by QC
d. A **metadata YAML file** (*.h.tsv.gz-meta.yaml) for the harmonised data file

Why do some studies not have harmonised files?

- If the file type is GWAS-SSF or pre-GWAS-SSF:
  - Likely awaiting harmonisation, please check back at a later date
- If the file type is non-GWAS-SSF:
  - No harmonised file – file is incompatible with our pipeline

- Current full list of harmonised files: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/harmonised_list.txt

External access to the harmonisation pipeline

- The harmonisation pipeline is also publicly available if you are interested in using it for your own applications.

- Please see the documentation on GitHub:
  - https://github.com/EBISPOT/gwas-sumstats-harmoniser

Summary statistics API

Summary statistics API

- Programmatic method for accessing summary statistics data
  - Contains **~30,000** studies (not all summary statistics in the Catalog)
  - Currently not updated while we plan a redesign

- See full API documentation: https://www.ebi.ac.uk/gwas/summary-statistics/docs/

Summary

- Summary statistics available for >65,000 studies
- View all available summary statistics, or search for specific publications, traits etc.
- FTP includes summary statistics files, metadata YAML files
- Harmonised files may available wherever possible – improves interoperability