

# The GWAS Catalog

## Accessing GWAS Catalog summary statistics

**Elliot Sollis**

Senior Curator, GWAS Catalog

I have no conflict of interest to declare

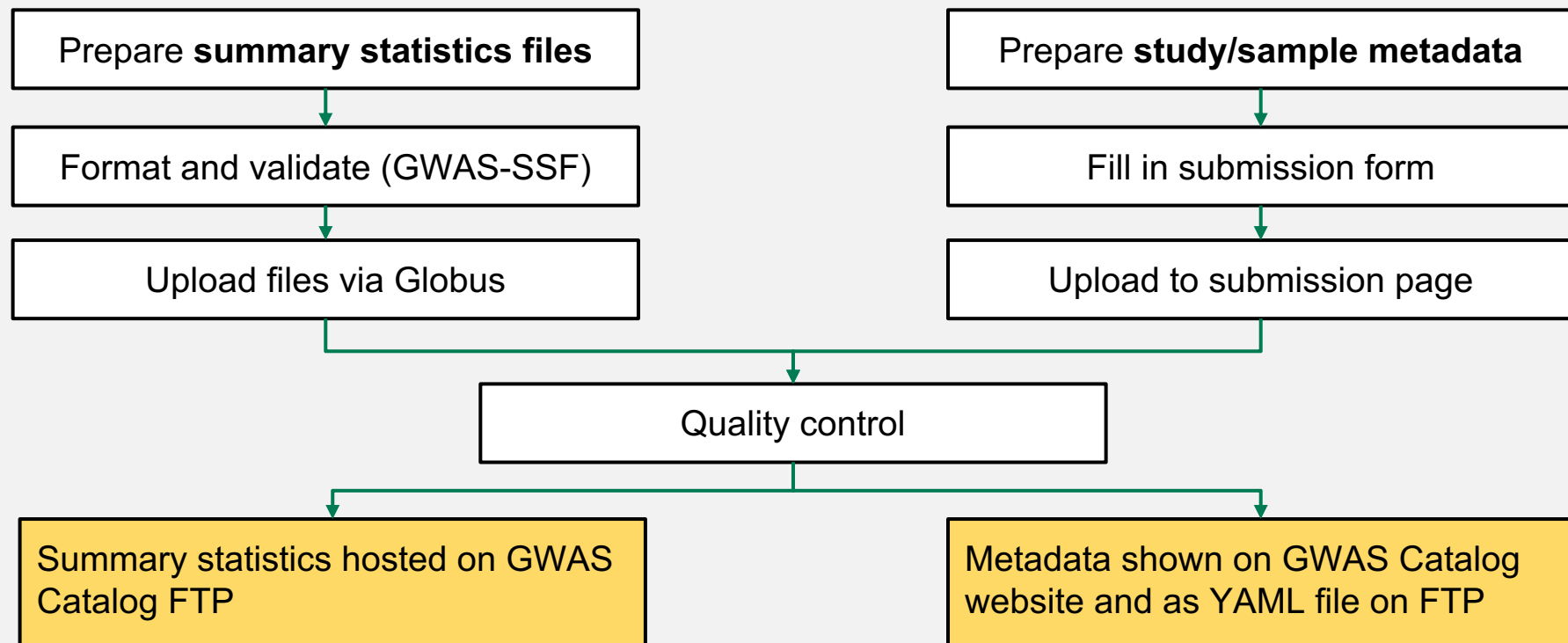
EMBL-EBI



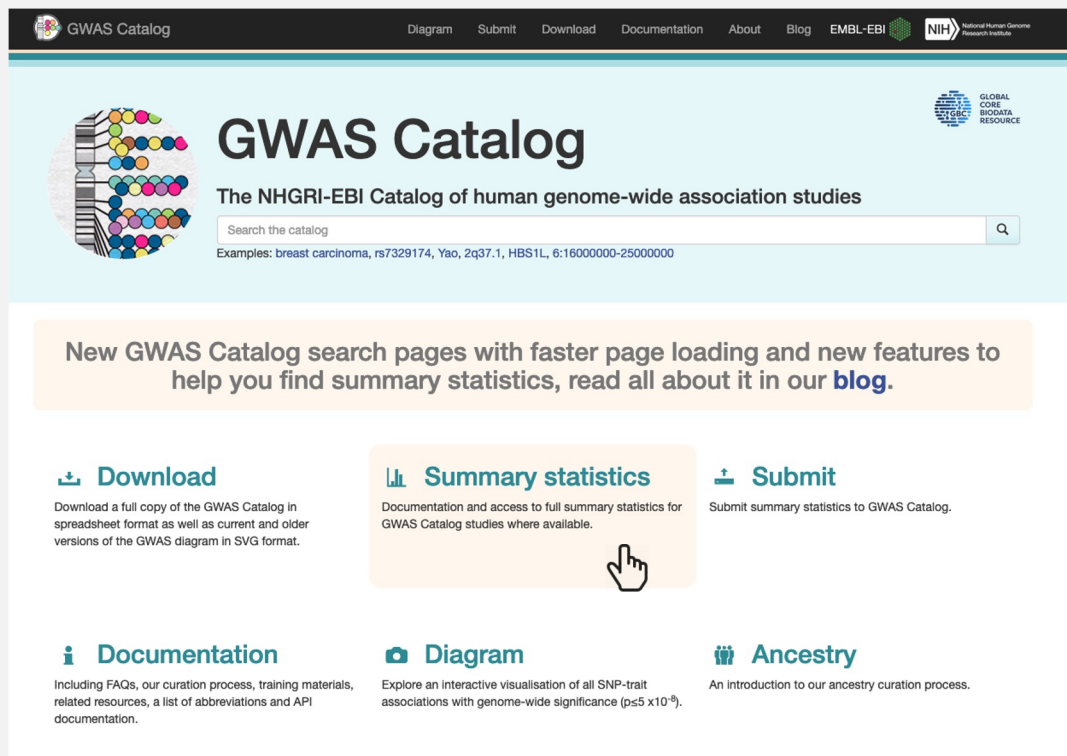
# Outline

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

# Submission process



# How to access summary statistics



The screenshot shows the GWAS Catalog homepage. At the top is a dark navigation bar with links: Diagram, Submit, Download, Documentation, About, Blog, EMBL-EBI, and NIH. The main header features the GWAS Catalog logo (a colorful circular genome map) and the text 'The NHGRI-EBI Catalog of human genome-wide association studies'. Below this is a search bar with the placeholder 'Search the catalog' and a magnifying glass icon. Examples of search terms are provided: 'breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:160000000-250000000'. A light orange banner below the search bar reads: 'New GWAS Catalog search pages with faster page loading and new features to help you find summary statistics, read all about it in our [blog](#).' Below the banner are six sections arranged in a 2x3 grid. The first row contains 'Download' (with a download icon), 'Summary statistics' (with a bar chart icon and a hand cursor pointing to it), and 'Submit' (with an upload icon). The second row contains 'Documentation' (with an information icon), 'Diagram' (with a camera icon), and 'Ancestry' (with a family tree icon). Each section has a brief description of its content.

GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:160000000-250000000

New GWAS Catalog search pages with faster page loading and new features to help you find summary statistics, read all about it in our [blog](#).

**Download**

Download a full copy of the GWAS Catalog in spreadsheet format as well as current and older versions of the GWAS diagram in SVG format.

**Summary statistics**

Documentation and access to full summary statistics for GWAS Catalog studies where available.

**Submit**

Submit summary statistics to GWAS Catalog.

**Documentation**

Including FAQs, our curation process, training materials, related resources, a list of abbreviations and API documentation.

**Diagram**

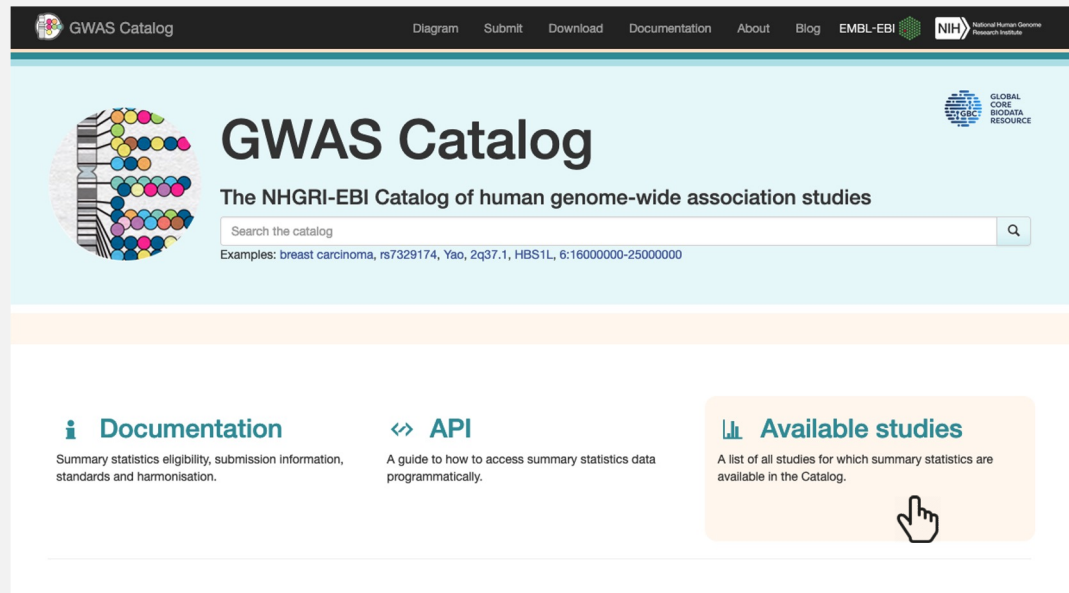
Explore an interactive visualisation of all SNP-trait associations with genome-wide significance ( $p \leq 5 \times 10^{-8}$ ).

**Ancestry**

An introduction to our ancestry curation process.

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

# How to access summary statistics



The screenshot shows the GWAS Catalog website. At the top is a dark navigation bar with links: Diagram, Submit, Download, Documentation, About, Blog, EMBL-EBI, and NIH National Human Genome Research Institute. The main header area is light blue and features the GWAS Catalog logo (a colorful circular plot) and the text "GWAS Catalog" and "The NHGRI-EBI Catalog of human genome-wide association studies". A search bar is present with the placeholder "Search the catalog" and a magnifying glass icon. Below the search bar, example search terms are listed: "Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000". The main content area is white and contains three sections: "Documentation" with an information icon and text about eligibility and standards; "API" with a code icon and text about programmatic access; and "Available studies" with a bar chart icon and text about a list of studies, accompanied by a hand cursor icon pointing at it.

GWAS Catalog

Diagram Submit Download Documentation About Blog EMBL-EBI NIH National Human Genome Research Institute

GLOBAL CORE BIODATA RESOURCE

## GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

### Documentation

Summary statistics eligibility, submission information, standards and harmonisation.

### API

A guide to how to access summary statistics data programmatically.

### Available studies

A list of all studies for which summary statistics are available in the Catalog.

# All published studies with summary statistics

## List of published studies with summary statistics

Studies **58828**

Show  entries

Column visibility Export Clear search




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

Showing 1 to 5 of 58,828 entries

« 1 2 3 4 5 ... 11766 »

# Unpublished studies with summary statistics

## List of prepublished/unpublished studies with summary statistics

<div>Search</div>								<div>  </div>
First Author	Date Submitted	Study Accession	Title	Reported trait	Ancestry Category	No of Individuals	FTP Path	
Krystyna Taylor	2020-07-09	<a href="#">GCST90002217</a>	Analysis of Genetic Host Response Risk Factors in Severe COVID-19 Patients	Severe response to COVID-19 (Sepsis controls)	European	2083	<a href="#">Download</a>	
Krystyna Taylor	2020-07-09	<a href="#">GCST90002218</a>	Analysis of Genetic Host Response Risk Factors in Severe COVID-19 Patients	Severe response to COVID-19 (Sepsis controls)	European	2083	<a href="#">Download</a>	
Sayoni Das	2020-07-09	<a href="#">GCST90002219</a>	Identification and Analysis of Shared Risk Factors in Sepsis and High Mortality Risk COVID-19 Patients	Sepsis	European	12696	<a href="#">Download</a>	
Sayoni Das	2020-07-09	<a href="#">GCST90002220</a>	Identification and Analysis of Shared Risk Factors in Sepsis and High Mortality Risk COVID-19 Patients	Sepsis	European	12696	<a href="#">Download</a>	
Julia M Sealock	2020-11-18	<a href="#">GCST90011993</a>	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	<a href="#">Download</a>	

Showing 1 to 5 of 6880 rows 

5 rows per page

1

2

3

4

5

...

1376

# Search by Publication, Trait etc.

## GWAS Catalog



The NHGRI-EBI Catalog of human genome-wide association studies



Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

### Search results for *hepatitis C virus infection*



**hepatitis C virus infection**

EFO\_0003047

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 b... [Show more >](#)

Associations **70** Studies **25**



# 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

# 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 visibilityExportClear search

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

Showing 1 to 4 of 4 entries

« 1 »

# FTP folder structure

Studies grouped in  
batches of 1000







**Index of /pub/databases/gwas/summary\_statistics  
/GCST90018001-GCST90019000/GCST90018995**

Folders named by  
study accession

Main summary  
statistics file

Metadata file

Subfolder for  
harmonised sumstats

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

# Metadata YAML files

# Standard metadata format

## Metadata YAML file

Basic file information

Genotyping information

Sample information

Trait information

```
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

```
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

<u>Name</u>	
 <a href="#">Parent Directory</a>	
 <a href="#">GCST90086099_buildGRCh37.tsv</a>	
 <a href="#">GCST90086099_buildGRCh37.tsv-meta.yaml</a>	
 <a href="#">harmonised/</a>	
 <a href="#">md5sum.txt</a>	

Index of /pub/databases/gwas/summary\_statistics/  
GCST90086001-GCST90087000/GCST90086099/  
harmonised

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 <a href="#">Parent Directory</a>		-	
 <a href="#">GCST90086099.h.tsv.gz</a>	2024-02-08 14:51	551M	
 <a href="#">GCST90086099.h.tsv.gz-meta.yaml</a>	2024-02-08 14:51	674	
 <a href="#">GCST90086099.h.tsv.gz.tbi</a>	2024-02-08 14:51	1.5M	
 <a href="#">GCST90086099.running.log</a>	2024-02-08 14:51	2.3K	
 <a href="#">md5sum.txt</a>	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](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