

The GWAS Catalog

Formatting and submitting summary statistics

Elliot Sollis

Senior Curator, GWAS Catalog

I have no conflict of interest to declare

Outline

- Introduction to summary statistics in the GWAS Catalog
 - Standard format (GWAS-SSF)
 - GWAS SumStats Tools
- Practical exercise:
 - Formatting and validating summary statistics using SSF-morph
- How to submit your summary statistics

Introduction

Summary statistics in the GWAS Catalog



>65,000

full GWAS summary statistics datasets
in the GWAS Catalog



Submitted by authors through our
submission portal:

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

Standard format: GWAS-SSF

- Standard format since 2023
- Enables interoperability between different summary statistics datasets
- Mandatory columns in a fixed order

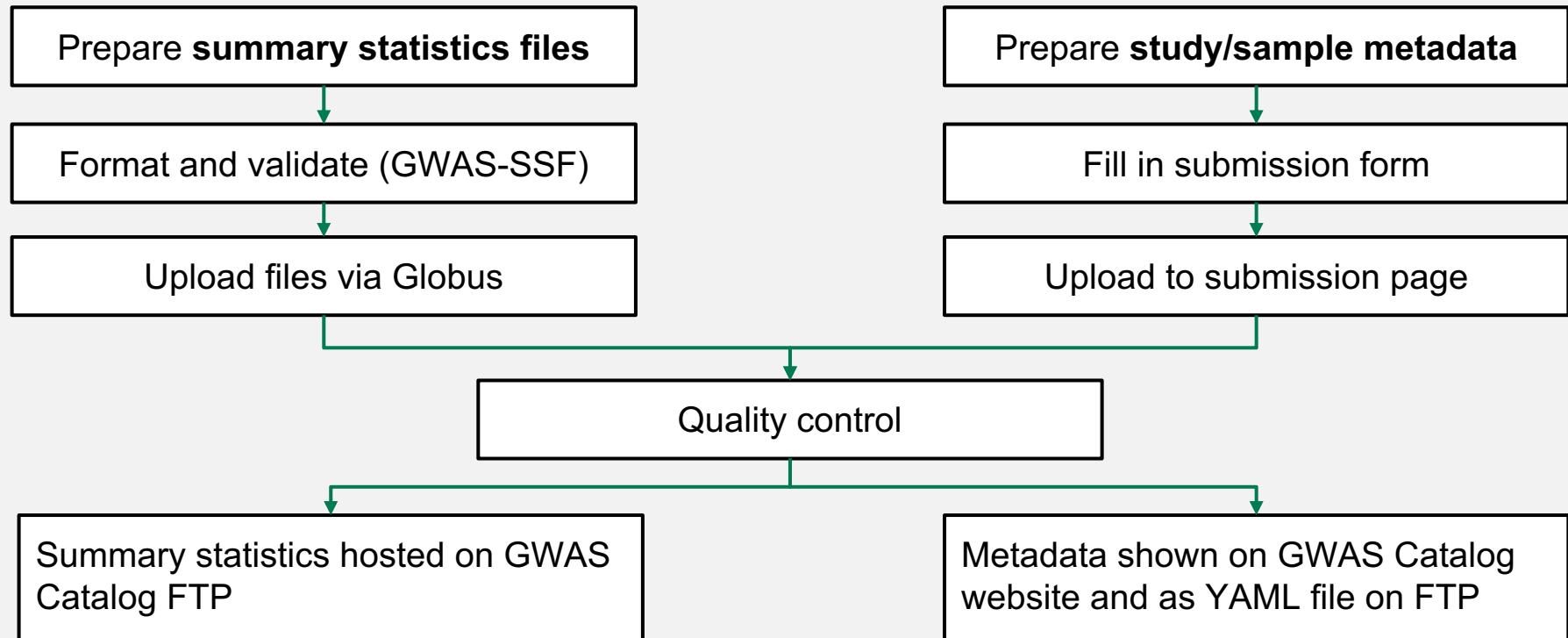
Column	Header
0	chromosome
1	base_pair_location
2	effect_allele
3	other_allele
4	beta / odds_ratio / hazard_ratio
5	standard_error
6	effect_allele_frequency
7	p_value / neg_log_10_p_value

Standard format: GWAS-SSF

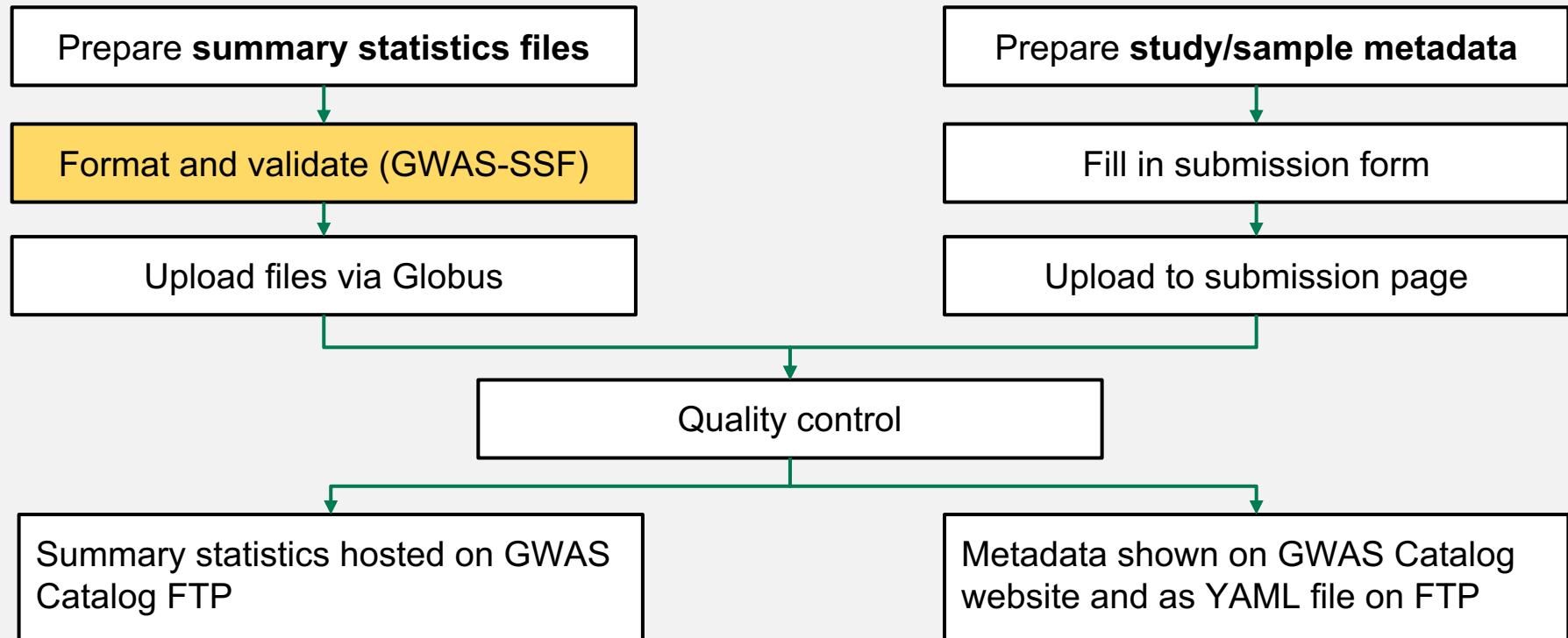
- Standard format since 2023
- Enables interoperability between different summary statistics datasets
- Mandatory columns in a fixed order
- Encouraged fields with standard headers

Column	Header
Encouraged	ci_upper
	ci_lower
	rsid
	variant_id
	info
	ref_allele
	n
Optional	...

Submission process



Submission process



GWAS SumStats Tools

Formatter

- Helps convert your summary statistics to GWAS-SSF format
 - Split and rename columns
 - Arrange columns in correct order
 - Edit data in bulk
- Built-in format conversions for popular analysis software:
 - REGENIE, BOLT-LMM, SNPtest, SAIGE

Validator

- Check if your file meets the GWAS-SSF requirements
- View error messages to find out what needs to be corrected

GWAS Sumstats Tools

Web interface (SSF-morph)	Command line tool
No installation required	Requires Python and gwas-sumstats-tools library
Requires Chromium-based browser (e.g. Google Chrome)	Can be run offline
Format or validate one file at a time	Format or validate multiple files
File size limit of 2GB	No file size limit

Practical example

Plan for the practical

[Next section](#)

1. Download an example summary statistics file: `eshg2024_example_sumstats.tsv`
2. Open **SSF-morph**: <https://ebispot.github.io/gwas-sumstats-tools-ssf-morph/>
3. Run the **validator** to check for errors
4. Create a **configuration file** to format the summary statistics file
5. **Edit and test** the configuration to correct errors
6. **Apply** the configuration file to the summary statistics file
7. Run the **validator** again to confirm that the errors have been corrected

1. Download example file

<https://github.com/EBISPORT/eshg-2024-workshop/tree/main/2-gwas-sumstat-demo>

Screenshot of a GitHub repository page for 'eshg-2024-workshop' showing the 'Code' tab selected. The repository is public and has 1 issue, 0 pull requests, 0 actions, 0 projects, 0 security, and 0 insights.

The left sidebar shows the repository structure:

- main
- 1-gwas-intro
- 2-gwas-sumstat-demo (selected)
- README
- eshg2024_example_sumstats....

The main area displays the commit history for the '2-gwas-sumstat-demo' branch:

Name	Last commit message	Last commit date
..		
README	reorganise	yesterday
eshg2024_example_sumstats.tsv	add example sumstats file for workshop	3 hours ago

The file 'eshg2024_example_sumstats.tsv' is highlighted with an orange border.

2. Open SSF-morph

<https://ebispot.github.io/gwas-sumstats-tools-ssf-morph/>

SSF-morph: Simplifying GWAS Sumstats Formatting and Validating

SSF-morph is an online tool designed to streamline the formatting and [validation](#) process for GWAS summary statistics files, specifically tailored for submission to the GWAS Catalog.

2. Run the validator

Select the folder containing the example file

STEP 5: Validate your formatted result

5.1 If you have not granted permission to read and edit files from your local directory in STEP 1, please click the **Grant permission** . Otherwise, you can skip this step and select the file you want to validate

Select to validate

Nice, you have granted the permission to the local directory formatter-testing



Grant permission

5.2 In your file, encountering fewer than 100,000 variants (rows) or zero p_values can lead to validation failure. To bypass the minimum row number requirement, and enable you to validate the rest of the file, please set the specific data validation requirements: the minimum number of rows should be and if you have zero-pvalues

For submission to the GWAS Catalog, lower row numbers may be permissible under certain circumstances (please contact gwas-subs@ebi.ac.uk to request an eligibility review). For files containing zero pvalues, analysis software type must be provided in the metadata template (see [GWAS Catalog submission documentation](#) for more details)

5.3 Click **Validate** to validate the selected file

2. Run the validator

Select the example file

STEP 5: Validate your formatted result

5.1 If you have not granted permission to read and edit files from your local directory in STEP 1, please click the [Grant permission](#). Otherwise, you can skip this step and select the file you want to validate

[Select to validate](#)

Nice, you have granted the permission to the local directory formatter-testing

You have selected the file eshg2024_example_sumstats.tsv for validation

5.2 In your file, encountering fewer than 100,000 variants (rows) or zero p_values can lead to validation failure. To bypass the minimum row number requirement, and enable you to validate the rest of the file, please set the specific data validation requirements: the minimum number of rows should be and if you have zero-pvalues

For submission to the GWAS Catalog, lower row numbers may be permissible under certain circumstances (please contact gwas-subs@ebi.ac.uk to request an eligibility review). For files containing zero pvalues, analysis software type must be provided in the metadata template (see [GWAS Catalog submission documentation](#) for more details)

5.3 Click [Validate](#) to validate the selected file

Set the minimum row number to **100** for this example file

2. Run the validator

5.3 Click **Validate the selected file** to validate the selected file

The validation result is:False.

Reason:Data table is invalid

```
error_preview:+-----+-----+-----+-----+-----+
| schema_context | column   | check      | check_number | failure_case | index |
+=====+=====+=====+=====+=====+
| Column      | chromosome | coerce_dtype('int64') | None        | chr1       | 0  |
+-----+-----+-----+-----+-----+
| Column      | chromosome | coerce_dtype('int64') | None        | chr14      | 64 |
+-----+-----+-----+-----+-----+
| Column      | chromosome | coerce_dtype('int64') | None        | chr16      | 74 |
+-----+-----+-----+-----+-----+
| Column      | chromosome | coerce_dtype('int64') | None        | chr16      | 73 |
+-----+-----+-----+-----+-----+
| Column      | chromosome | coerce_dtype('int64') | None        | chr16      | 72 |
+-----+-----+-----+-----+-----+
...
primary_error_type:data
```

The example file fails validation because the **chromosome** column contains an incorrect data type.

It should contain only **integers**.

Therefore we need to remove the “**chr**” prefixes from the chromosome numbers.

4. Create a configuration file

STEP 1: Select the input file

1.1 Please grant the permission to read and edit files in one of your local directories [Grant permission](#)

Nice, you have granted the permission to the local directory formatter-testing X

1.2 Please select your input file [Select input file](#)

You have selected the file eshg2024_example_sumstats.tsv X

STEP 2: Prepare the configuration file

2.1 Please specify the delimiter in the input file and please specify the first character in the line which indicates that this line is a comment: . (Note: please leave any fields empty if they are not applicable.)

2.2 Please select the appropriate analysis software from the dropdown menu:

[Generate](#)

The example file is **tab** delimited

5. Edit and test the configuration file

STEP 3: Edit and test your configuration

Preview how your configuration changes the input file using the 'Test' button, and then edit the configuration

Configuration:

```
{  
  "fileConfig": {  
    "outFileSuffix": "formatted_ ",  
    "fieldSeparator": "\t",  
    "naValue": null,  
    "convertNegLog10Pvalue": false,  
    "removeComments": ""  
  },  
  "columnConfig": {  
    "split": [  
      {  
        "field": "chromosome",  
        "separator": null,  
        "capture": null,  
        "new_field": null,  
        "include_original": null  
      },  
      {  
        "field": "base_pair_location",  
        "separator": null,  
        "capture": null,  
        "new_field": null,  
        "include_original": null  
      },  
      {  
        "field": "effect_allele",  
        "separator": null,  
        "capture": null,  
        "new_field": null,  
        "include_original": null  
      }  
    ]  
  }  
}
```

Show Example data

Show Your Input data

chromosome	base_pair_location	effect_allele	other_allele	odds_ratio	stan
chr1	751756	C	T	1.024	0.05
chr1	1040472	T	C	1.007	0.04
chr1	1048571	G	A	1.004	0.04
chr1	1059269	T	C	0.9952	0.05
chr1	1157631	A	G	1.031	0.05

Showing 1 to 5 of 5 entries

Your output:

Test

- If you are happy with your configure file, you can also download the configure file here [download configure file](#)

See a preview of
your input file

5. Edit and test the configuration file

STEP 3: Edit and test your configuration

Preview how your configuration changes the input file using the 'Test' button, and then edit the config

Configuration:

```
'  
  "edit": [  
    {  
      "field": "chromosome",  
      "rename": "chromosome",  
      "find": "chr",  
      "replace": "",  
      "extract": null  
    },  
    {  
      "field": "base_pair_location",  
      "rename": "base_pair_location",  
      "find": null,  
      "replace": null,  
      "extract": null  
    },  
    {  
      "field": "effect_allele",  
      "rename": "effect_allele",  
      "find": null,  
      "replace": null,  
      "extract": null  
    },  
    {  
      "field": "other_allele",  
      "rename": "other_allele",  
      "find": null,  
      "replace": null,  
      "extract": null  
    }  
  ]  
'
```

Test

- If you are happy with your configure file, you can also download the configure file here

[download configure file](#)

We need to find all instances of "chr" in the **chromosome** column and delete them, i.e. replace with ""

Show Your Input data

chromosome	base_pair_location	effect_allele	other_allele	odds_ratio	stan
chr1	751756	C	T	1.024	0.05
chr1	1040472	T	C	1.007	0.04
chr1	1048571	G	A	1.004	0.04
chr1	1059269	T	C	0.9952	0.05
chr1	1157631	A	G	1.031	0.05

Showing 1 to 5 of 5 entries

Your output:

5. Edit and test the configuration file

STEP 3: Edit and test your configuration

Preview how your configuration changes the input file using the 'Test' button, and then edit the configuration

Configuration:

```
,  
],  
"edit": [  
{  
    "field": "chromosome",  
    "rename": "chromosome",  
    "find": "chr",  
    "replace": "",  
    "extract": null  
,  
{  
    "field": "base_pair_location",  
    "rename": "base_pair_location",  
    "find": null,  
    "replace": null,  
    "extract": null  
,  
{  
    "field": "effect_allele",  
    "rename": "effect_allele",  
    "find": null,  
    "replace": null,  
    "extract": null  
,  
{  
    "field": "other_allele",  
    "rename": "other_allele",  
    "find": null,  
    "replace": null,  
    "extract": null
```

Show Example data

Show Your Input data

chromosome	base_pair_location	effect_allele	other_allele	odds_ratio	stan
chr1	751756	C	T	1.024	0.05
chr1	1040472	T	C	1.007	0.04
chr1	1048571	G	A	1.004	0.04
chr1	1059269	T	C	0.9952	0.05
chr1	1157631	A	G	1.031	0.05

Showing 1 to 5 of 5 entries

Your output:

formatting result
example

chromosome	base_pair_location	effect_allele	other_allele	odds_ratio	stan
1	751756	C	T	1.024	0.05
1	1040472	T	C	1.007	0.04
1	1048571	G	A	1.004	0.04
1	1059269	T	C	0.9952	0.05
1	1157631	A	G	1.031	0.05

Showing 1 to 5 of 5 entries

Test

- If you are happy with your configuration file, you can also download the configuration file here

[download configuration file](#)

Preview shows
“chr” has been
removed from the
output file

6. Apply the configuration file

STEP 4: Applying the configuration to your selected input file (1 file)

It's time to apply your configuration to your input file. The formatted result will then be downloaded to the folder you have permitted.

we are applying the configure to eshg2024_example_sumstats.tsv

Apply

Choose a file name for your formatted file

7. Re-run the validator

STEP 5: Validate your formatted result

5.1 If you have not granted permission to read and edit files from your local directory in STEP 1, please click the [Grant permission](#). Otherwise, you can skip this step and select the file you want to validate

[Select to validate](#)

You have selected the file eshg2024_example_sumstats_formatted.tsvfor validation X

5.2 In your file, encountering fewer than 100,000 variants (rows) or zero p_values can lead to validation failure. To bypass the minimum row number requirement, and enable you to validate the rest of the file, please set the specific data validation requirements: the minimum number of rows should be and if you have zero-pvalues False

For submission to the GWAS Catalog, lower row numbers may be permissible under certain circumstances (please contact gwas-subs@ebi.ac.uk to request an eligibility review). For files containing zero pvalues, analysis software type must be provided in the metadata template (see [GWAS Catalog submission documentation](#) for more details)

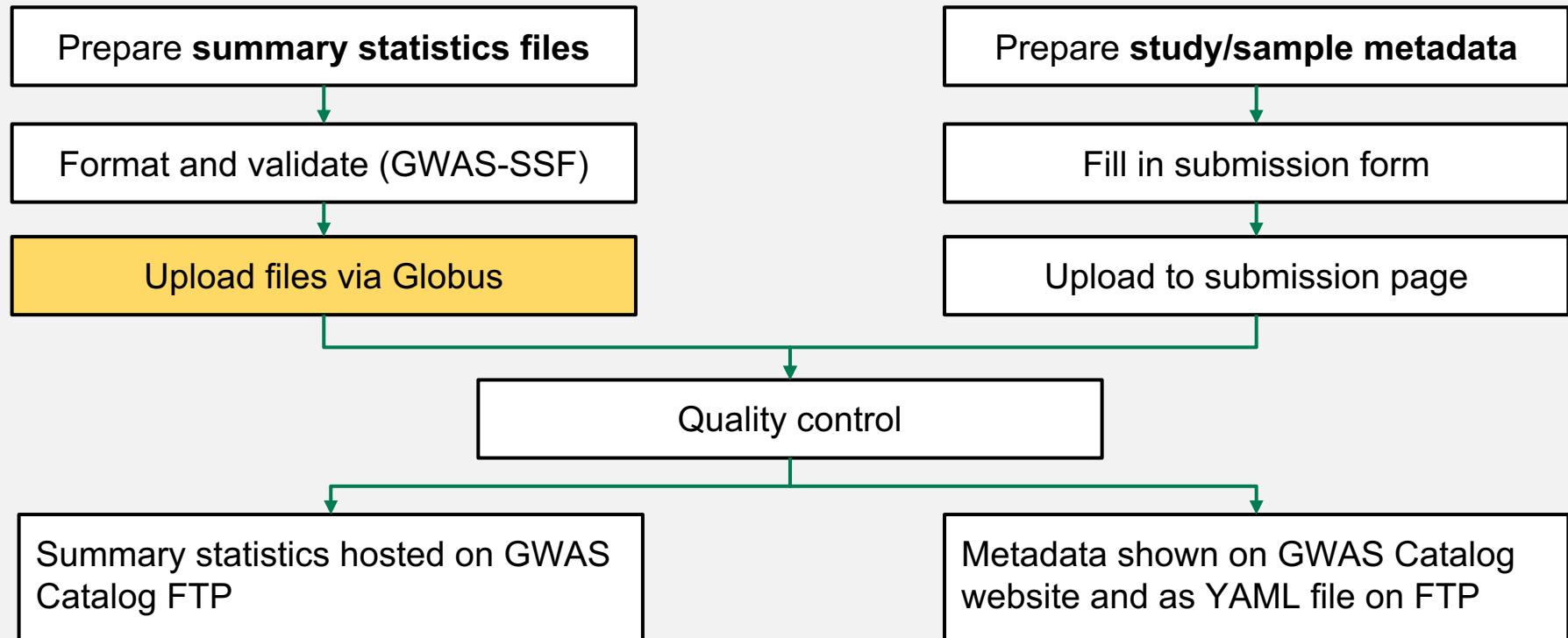
5.3 Click [Validate the selected file](#) to validate the selected file

The validation result is:True.
Reason:Data table is valid.
error_preview:None
primary_error_type:None

The formatted file is now valid!

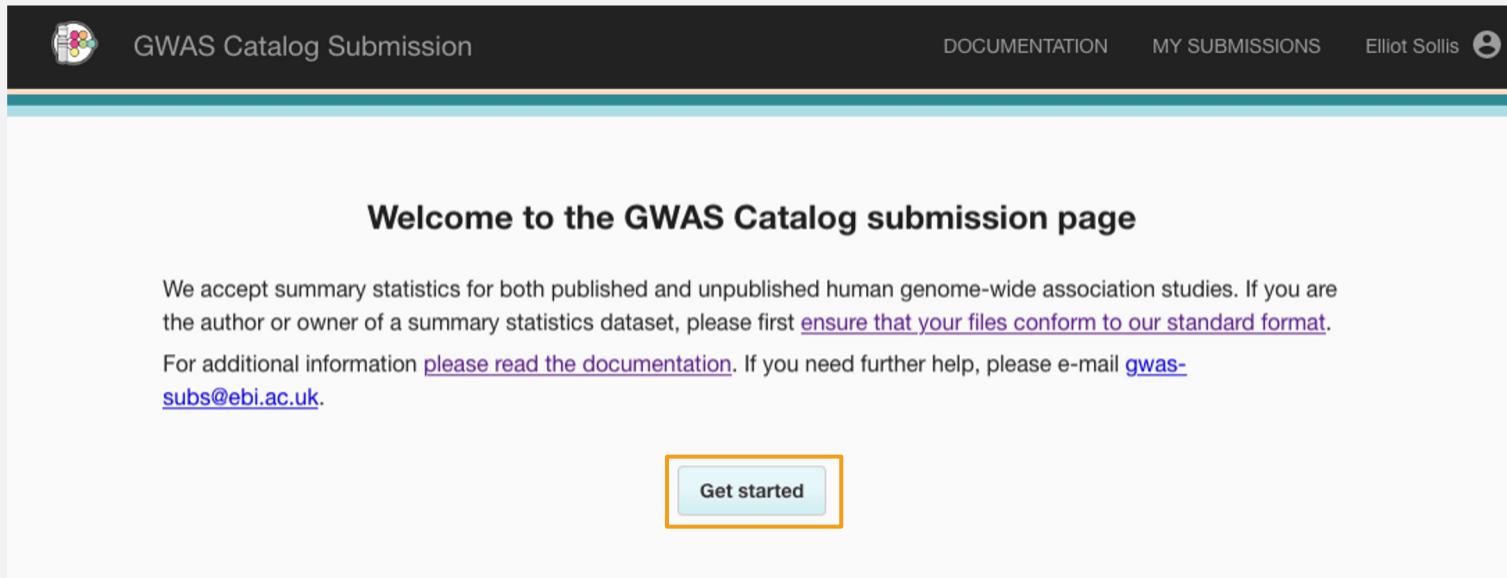
How to submit

Submission process



Submission portal

Go to the GWAS Catalog Submission portal and follow the prompts to **create a submission**



The screenshot shows the 'GWAS Catalog Submission' website. At the top, there is a dark header bar with the site's name on the left, 'DOCUMENTATION' and 'MY SUBMISSIONS' in the center, and a user profile for 'Elliot Sollis' on the right. Below the header is a teal navigation bar. The main content area has a light gray background. In the center, the text 'Welcome to the GWAS Catalog submission page' is displayed in bold. Below this, two paragraphs provide instructions: one about accepting summary statistics and another about contacting support. At the bottom of the main content area is a light blue button with the text 'Get started' in white, which is highlighted with an orange rectangular border.

GWAS Catalog Submission

DOCUMENTATION MY SUBMISSIONS Elliot Sollis

Welcome to the GWAS Catalog submission page

We accept summary statistics for both published and unpublished human genome-wide association studies. If you are the author or owner of a summary statistics dataset, please first [ensure that your files conform to our standard format](#).

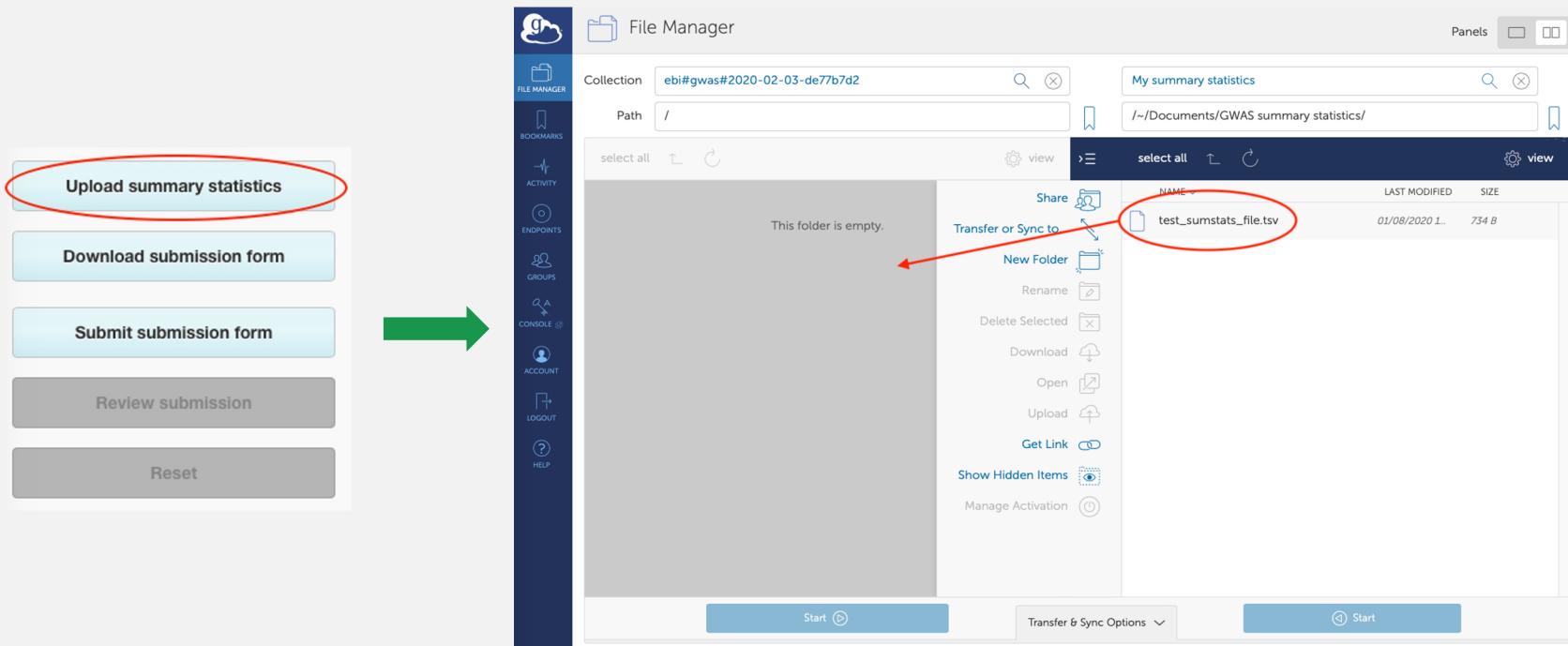
For additional information [please read the documentation](#). If you need further help, please e-mail gwas-subs@ebi.ac.uk.

Get started

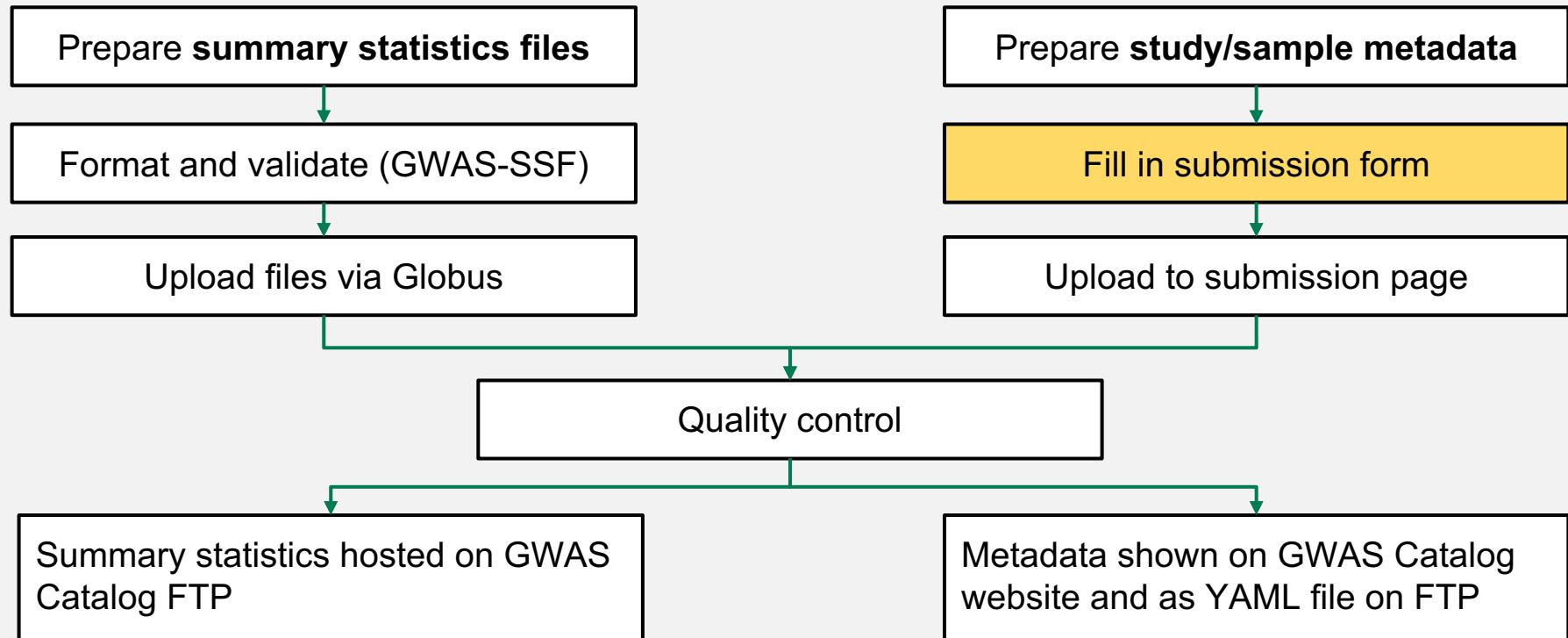
<https://www.ebi.ac.uk/gwas/deposition>

Upload summary statistics

Use Globus to **upload summary statistics** from your computer to the GWAS Catalog file system



Submission process



Fill in metadata submission form

Download the submission form to [enter your metadata](#)

Upload summary statistics

Download submission form

Submit submission form

Review submission

Reset

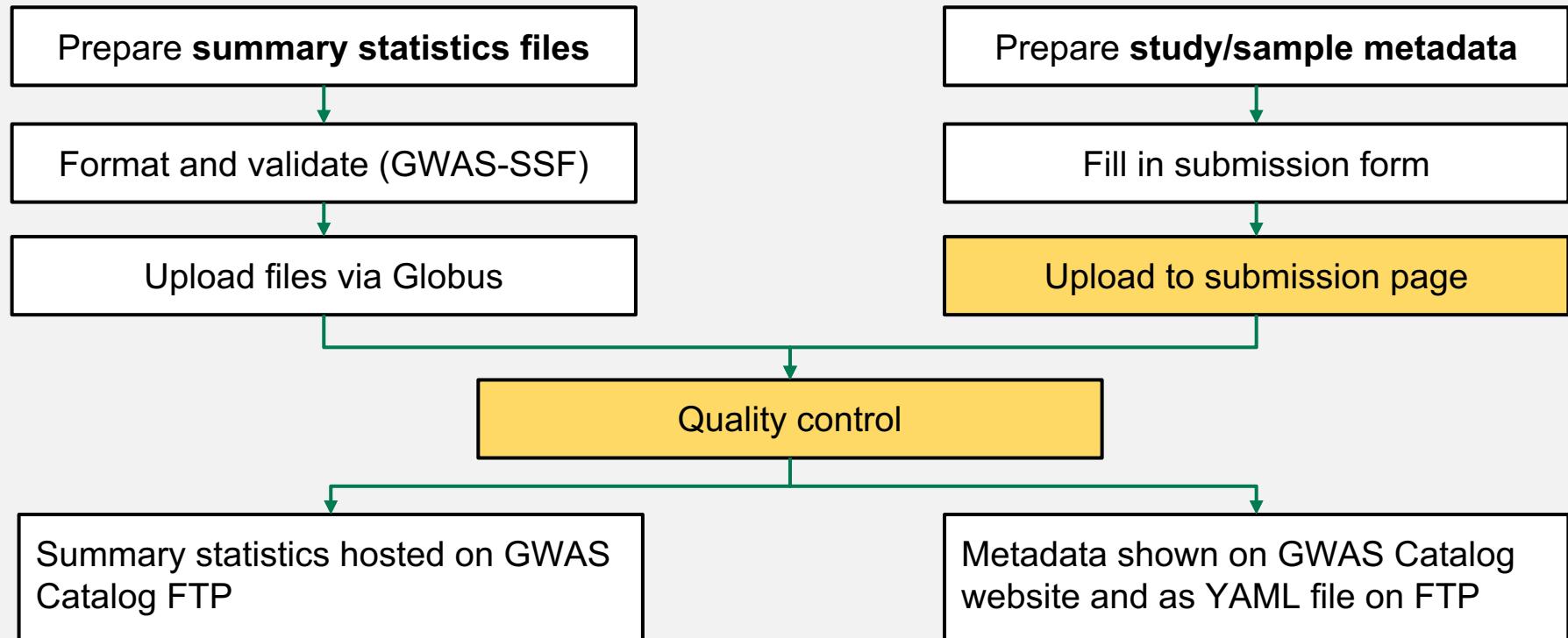


submission_form_example

A	B	C	D	E	F
A unique free-text label for each genome-wide association study in the publication	The method used to genotype variants in the discovery stage. Multiple values can be listed separated by ' '. Example: illumina	Manufacturer of the genotyping array used for the discovery stage. Example: illumina	Additional information about the genotyping array. Example: immunochip	Were SNPs imputed for the discovery GWAS? Example: yes, no	The number of variants analysed in the discovery stage (after QC)
Study tag	Genotyping technology	Array manufacturer	Array information	Imputation	Variant count
Add your data below this line					
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					

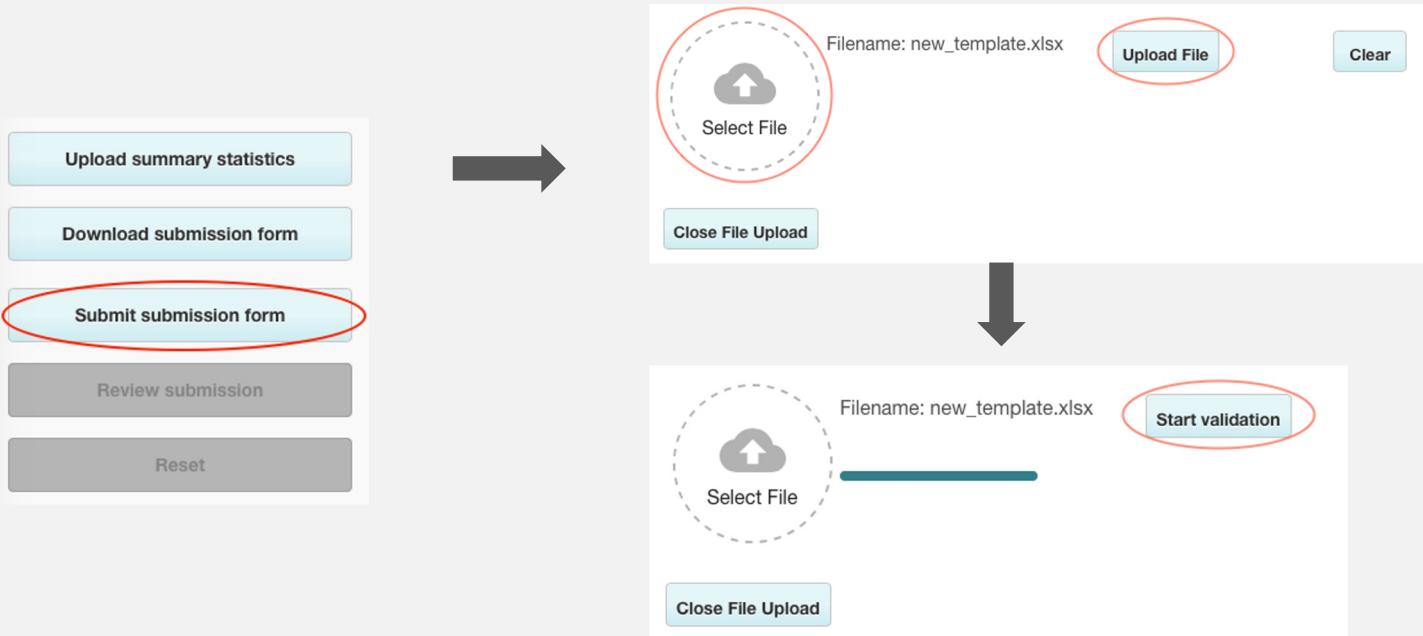
See documentation for detailed instructions:
<https://www.ebi.ac.uk/gwas/docs/submission>

Submission process



Upload submission form

Upload the **submission form** and **start validation**



Useful links

- GWAS Sumstats Tools documentation
 - <https://ebispot.github.io/gwas-sumstats-tools-documentation>
- Run SSF-morph
 - <https://ebispot.github.io/gwas-sumstats-tools-ssf-morph>
- General submission documentation:
 - <https://www.ebi.ac.uk/gwas/docs/submission>
- Submission portal
 - <https://www.ebi.ac.uk/gwas/deposition>
- Help with your submission
 - Email gwas-subs@ebi.ac.uk