The GWAS Catalog

Formatting and submitting summary statistics

Elliot Sollis

Senior Curator, GWAS Catalog

I have no conflict of interest to declare

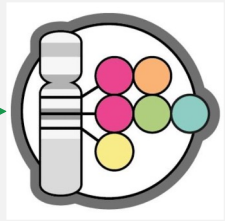
EMBL-EBI

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**

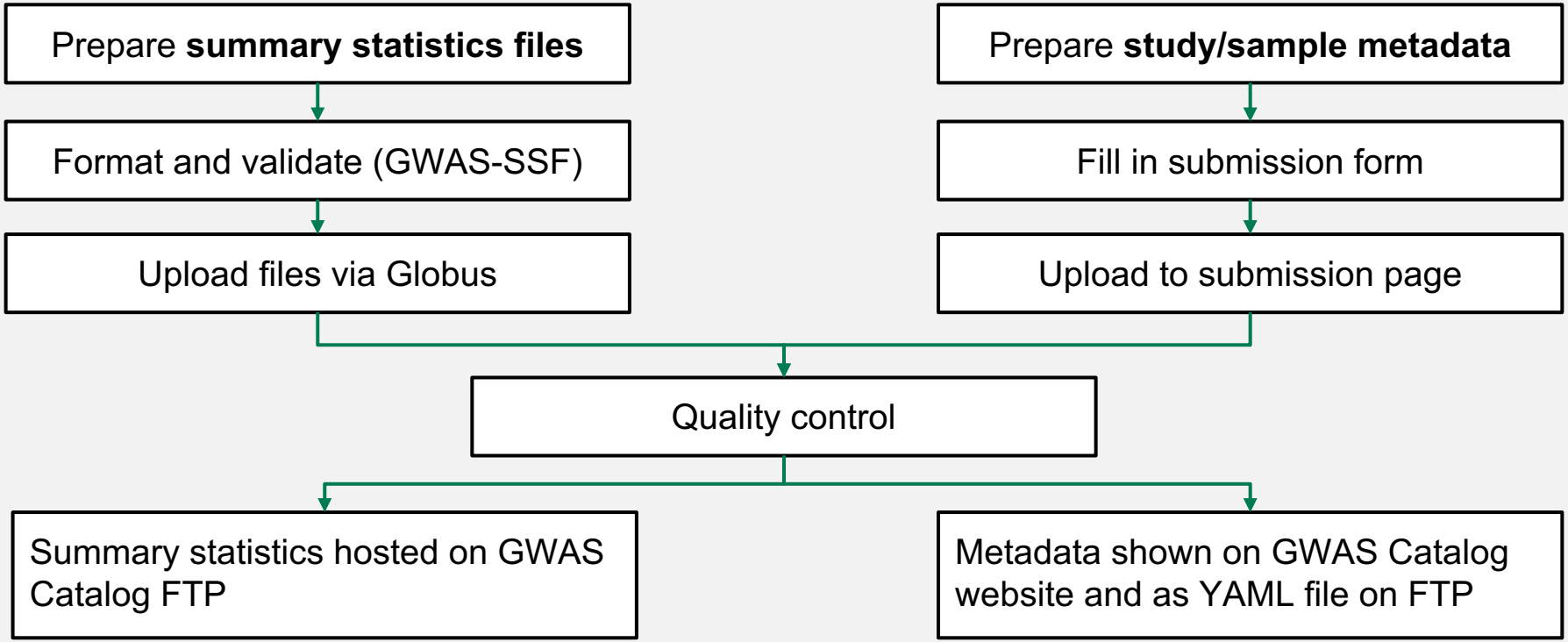
Standard format: GWAS-SSF

- Standard format since 2023
- Enables interoperability between different summary statistics datasets

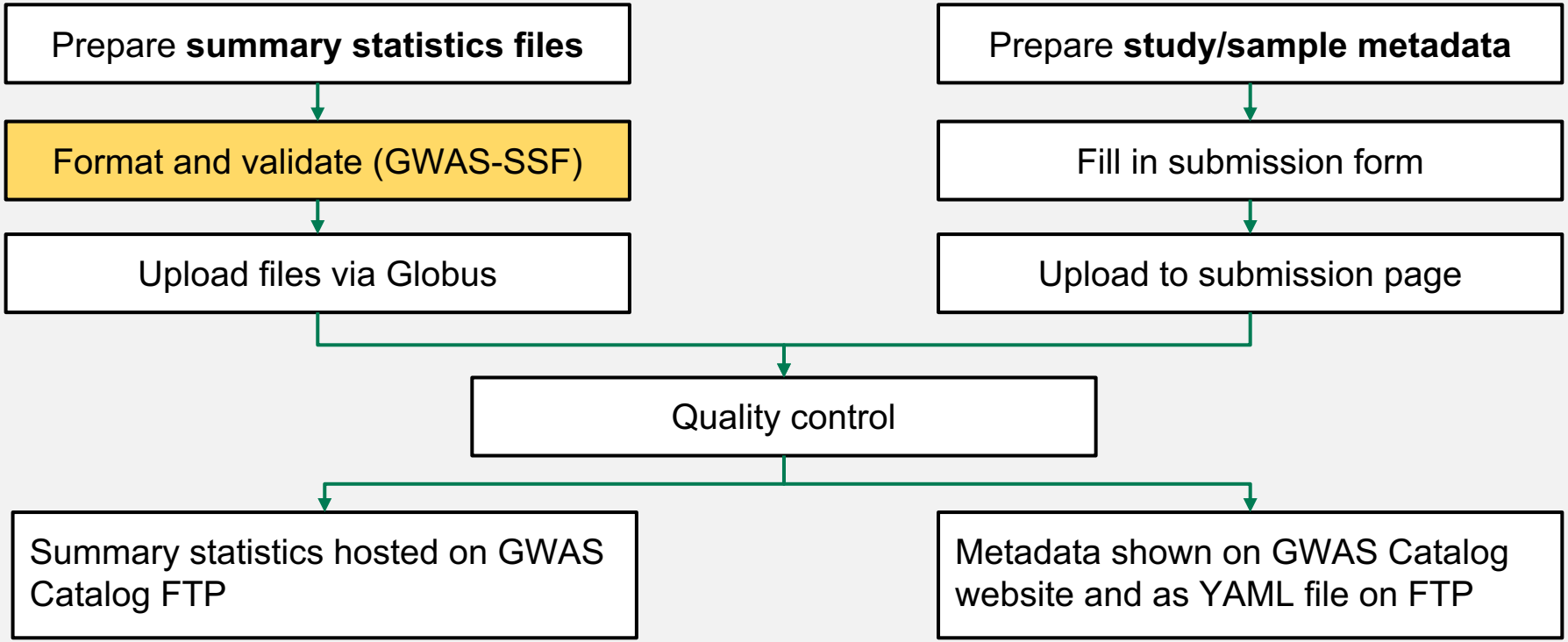
- Mandatory columns in a fixed order
- Encouraged fields with standard headers

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

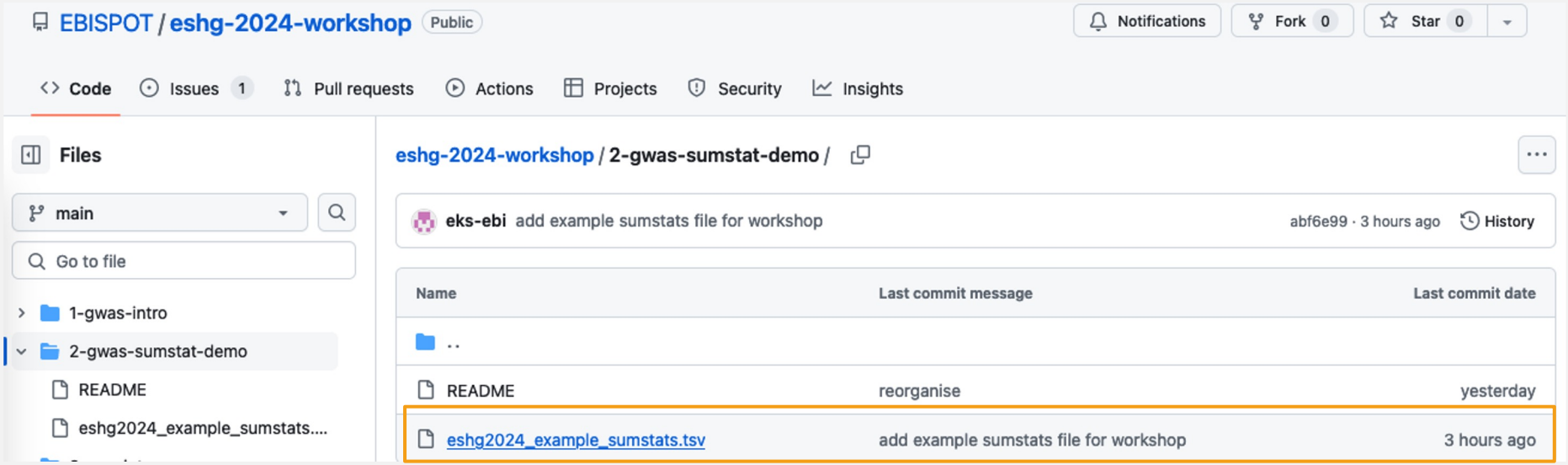
Plan for the practical

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

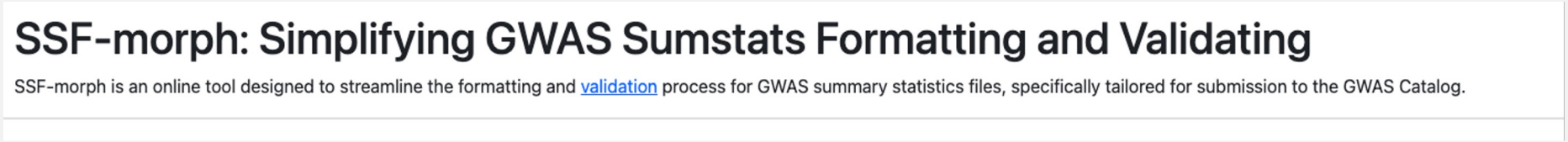
1. Download example file

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

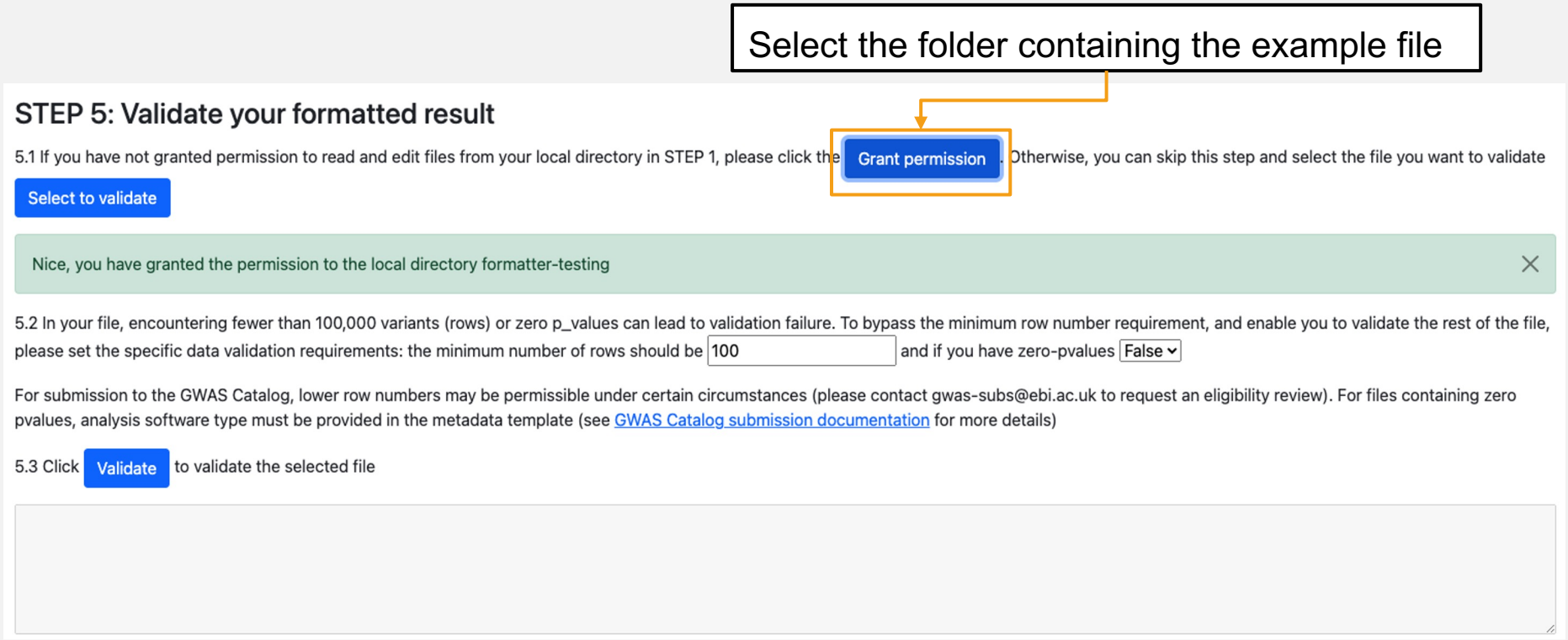


2. Open SSF-morph

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

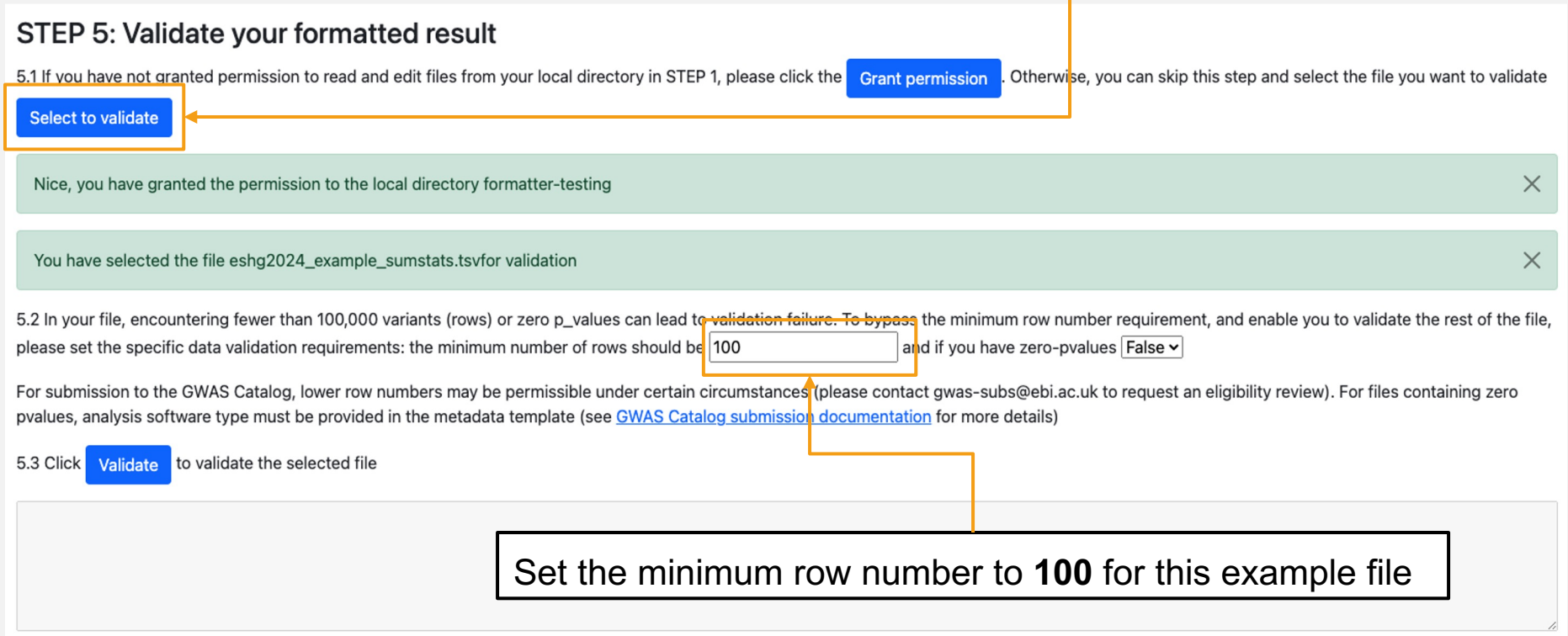


2. Run the validator



2. Run the validator

Select the example file



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

EMBL-EBI

2. Run the validator

5.3 Click [Validate the selected file] to validate the selected file

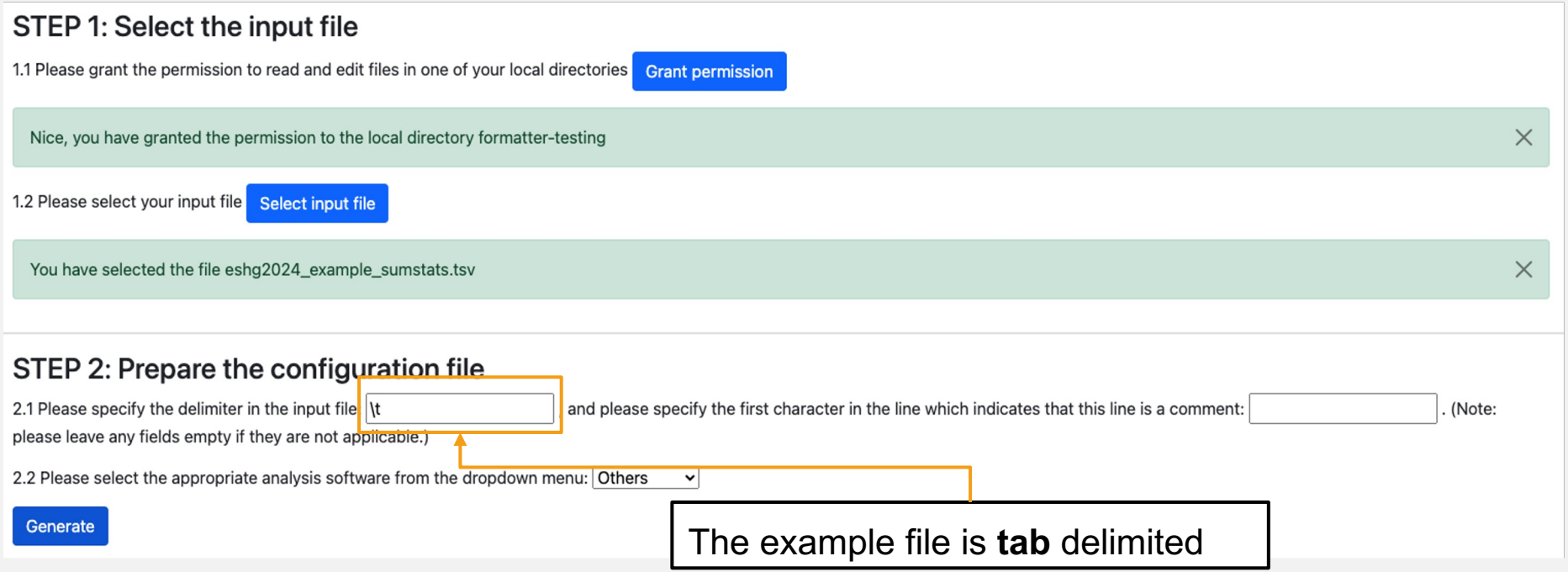
```
The validation result is:False.
Reason:Data table is invalid
error_preview:+----------------+------------+----------------------+--------------+--------------+-------+
| schema_context | column    | check              | check_number | failure_case | index |
+================+============+======================+==============+==============+=======+
| 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.

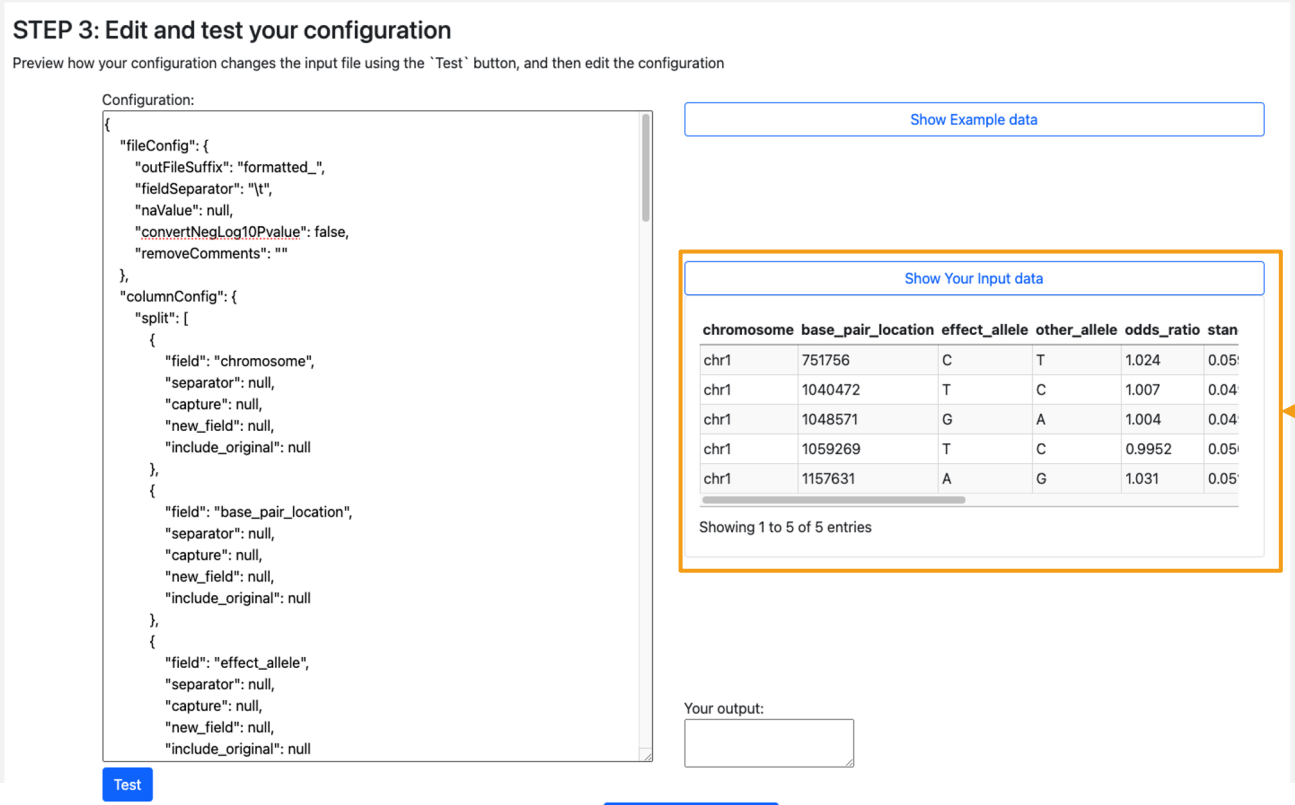
It should contain only **integers**.

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

4. Create a configuration file



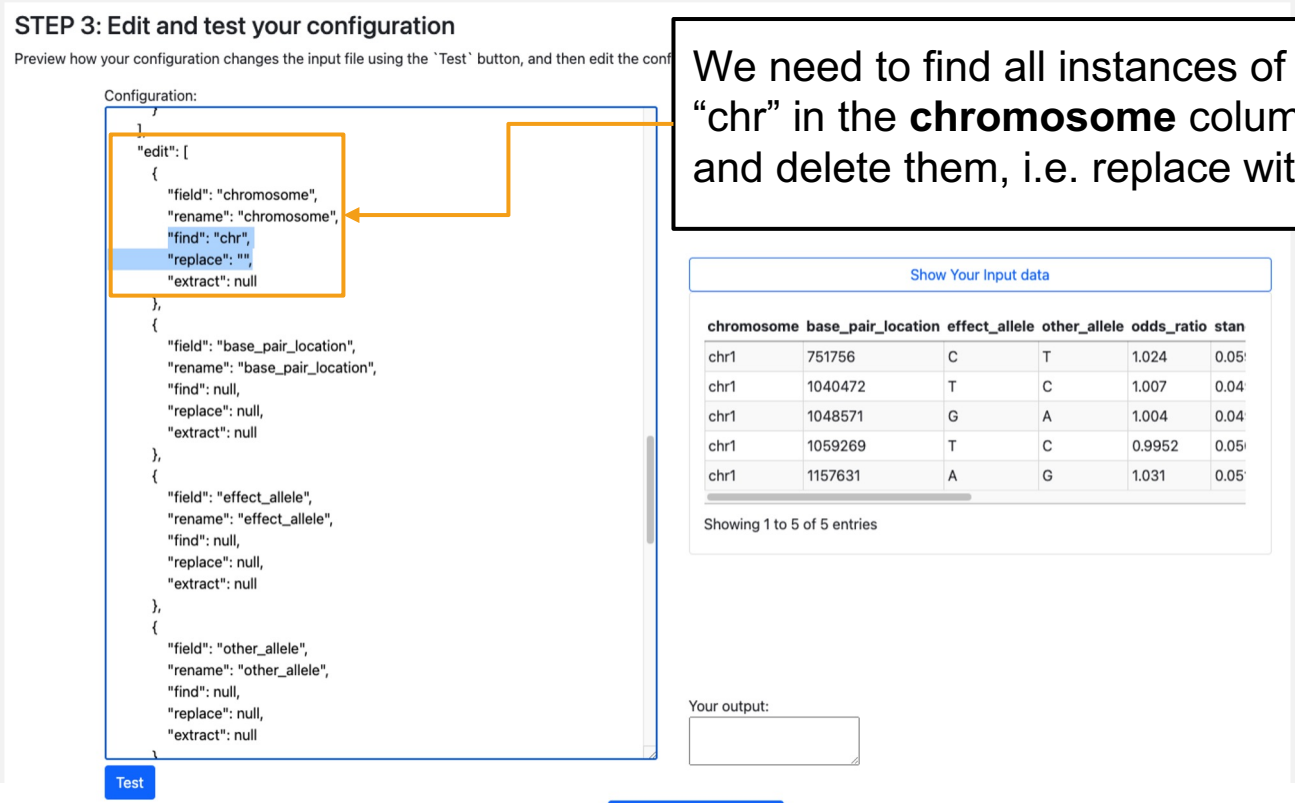
5. Edit and test the configuration file



See a preview of your input file

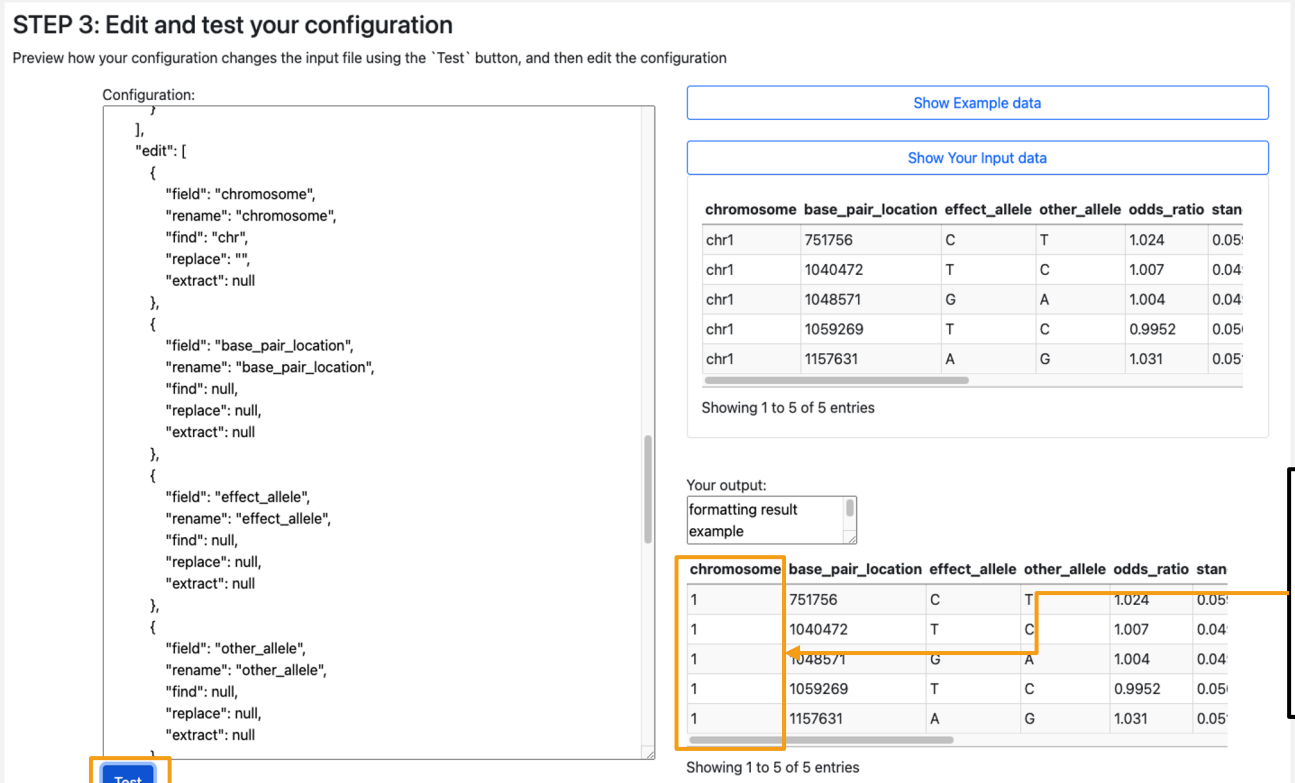
EMBL-EBI

5. Edit and test the configuration file



EMBL-EBI

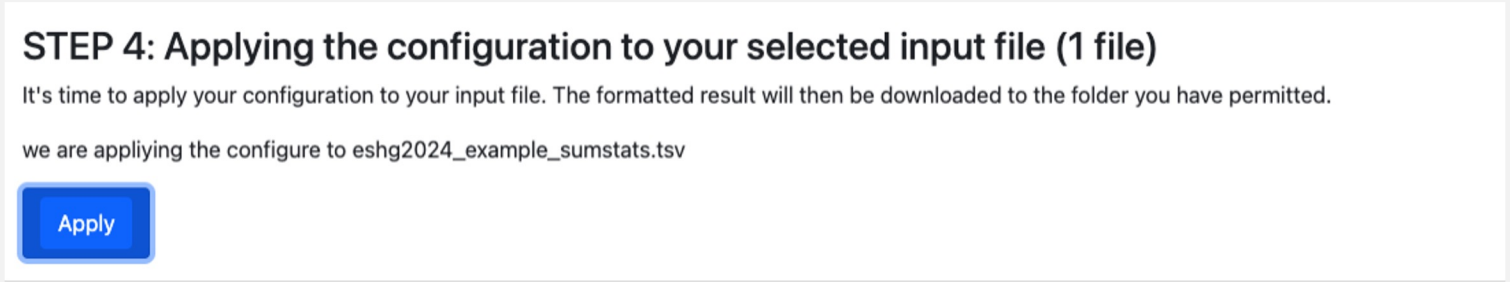
5. Edit and test the configuration file



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

EMBL-EBI

6. Apply the configuration file



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

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 100 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

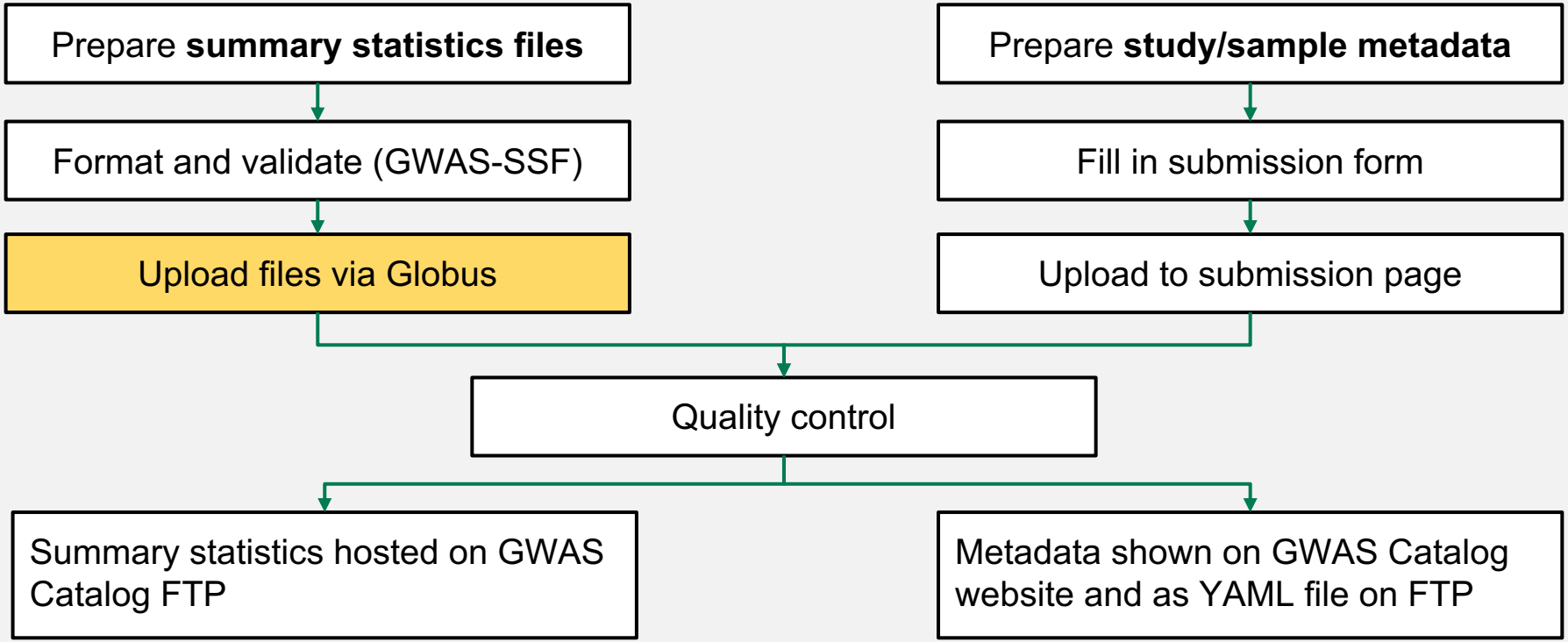
The validation result is:True.
Reason:Data table is valid.
error_preview:None
primary_error_type:None

The formatted file is now valid!

EMBL-EBI

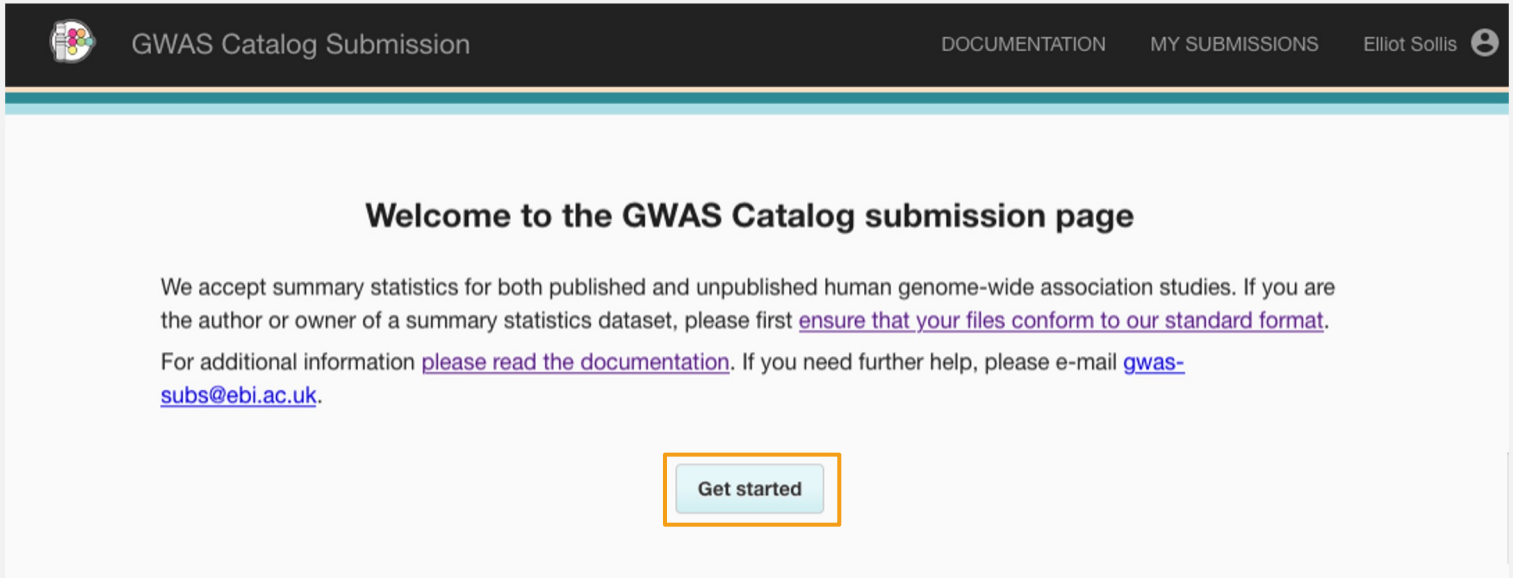
How to submit

Submission process



Submission portal

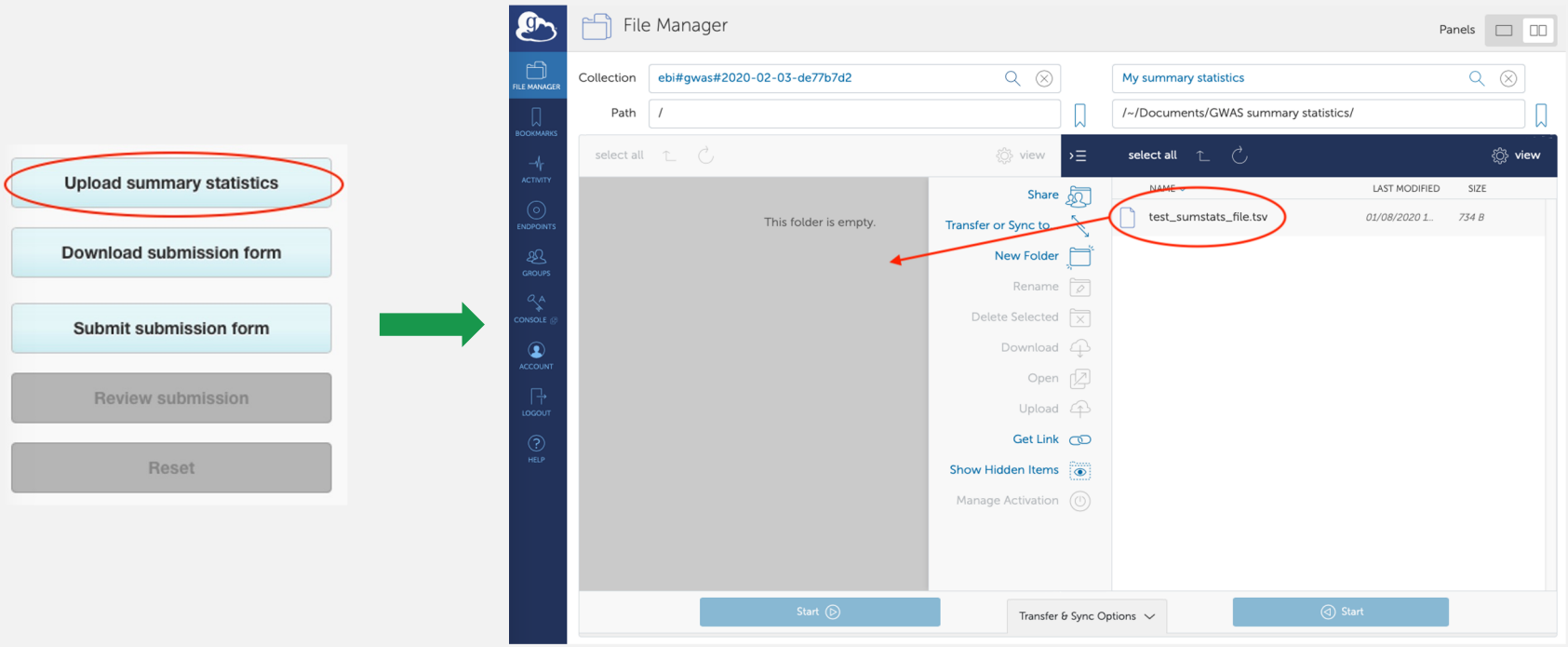
Go to the GWAS Catalog Submission portal and follow the prompts to **create a submission**



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

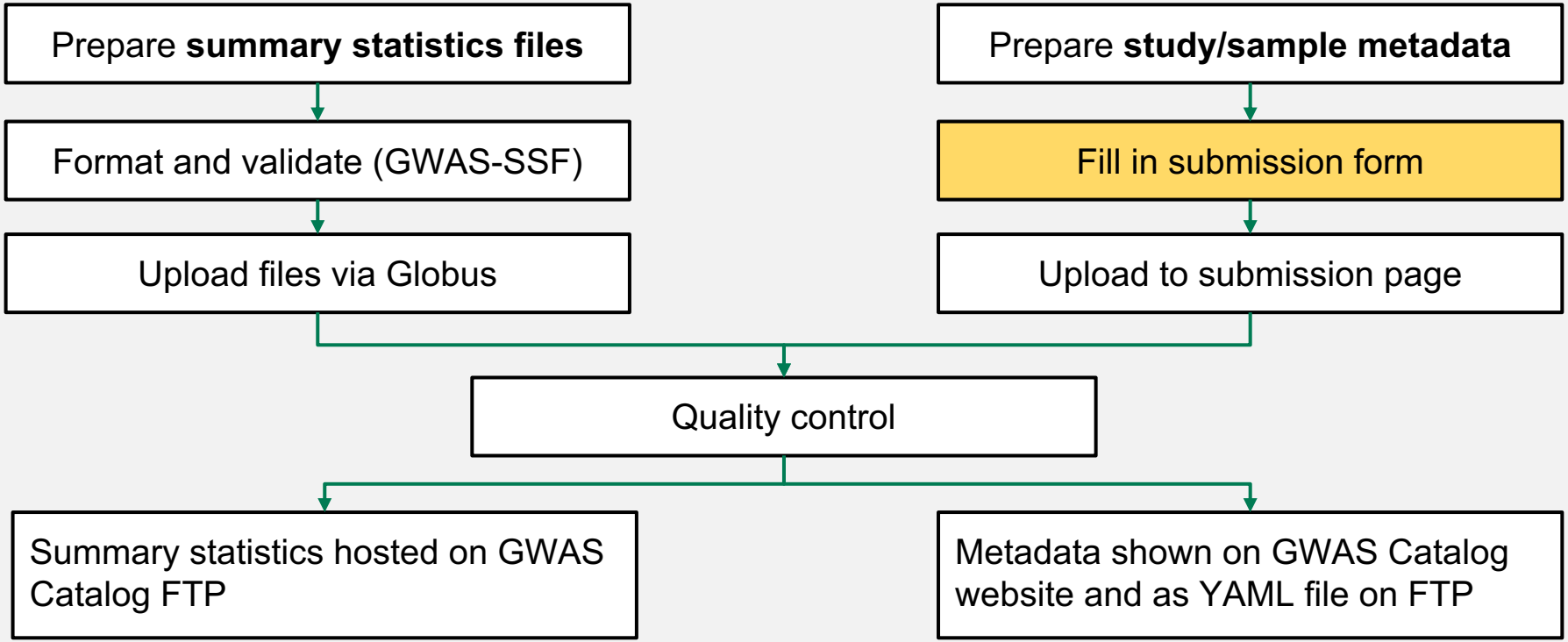
Upload summary statistics

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



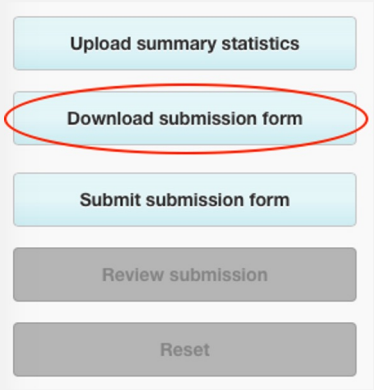
EMBL-EBI

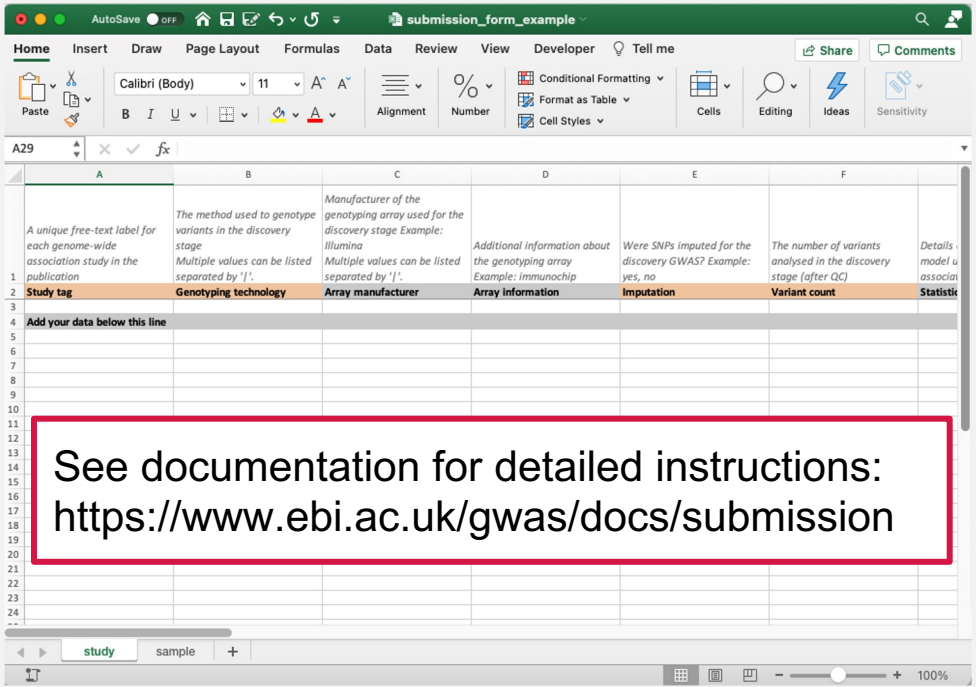
Submission process



Fill in metadata submission form

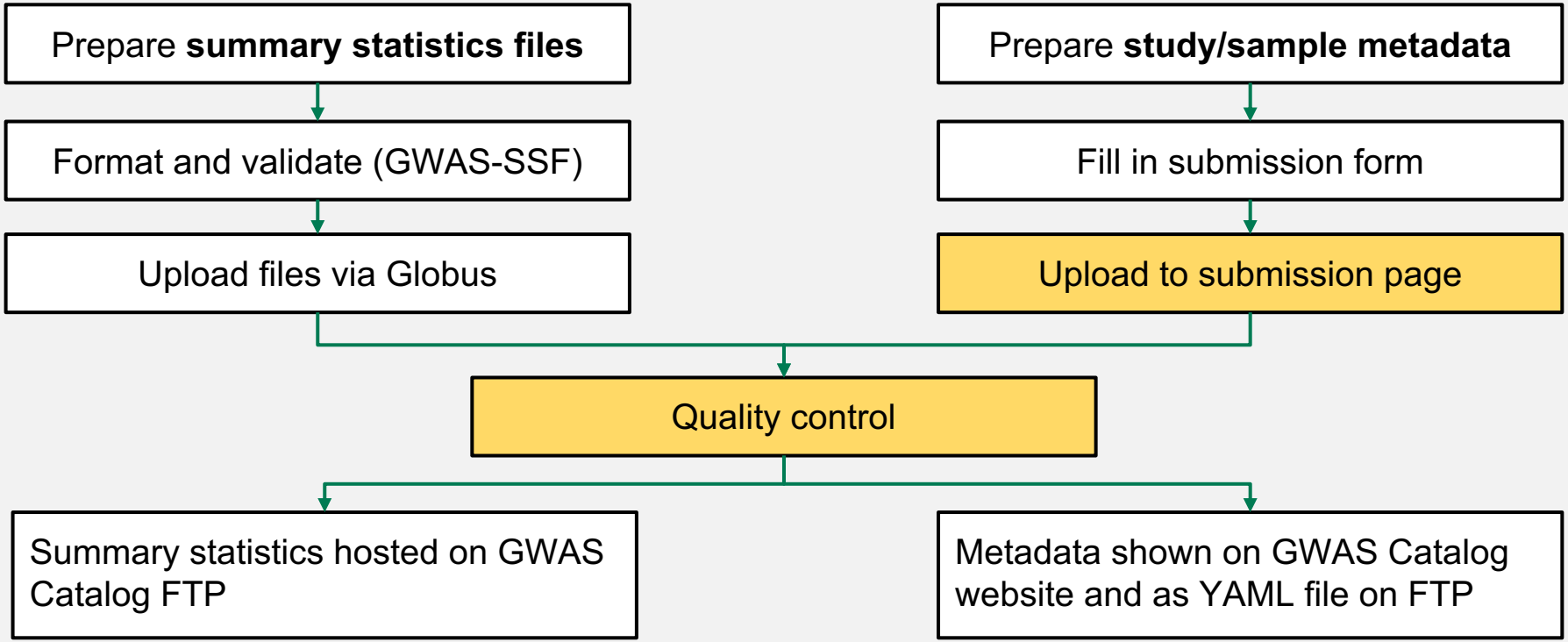
Download the submission form to **enter your metadata**





EMBL-EBI

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