GWAS-SSF: A GWAS Summary Statistics Format (v0.1)

January 16, 2024

Contents

1 Introduction

The GWAS Catalog hosted a workshop and series of meetings between June 2020 and September 2021 with summary statistics stakeholders, including data generators, data users, data managers and bioinformaticians, representing diverse user groups of human GWAS. These meetings gathered requirements and identified challenges. The aim of this process was to set minimum information elements for data sharing to maximise downstream utility.

2 Requirements

The key requirements obtained from the stakeholders' use cases were as follows:

1. Consistent representation of data to enable interoperability
2. Easily accessible metadata for summary statistics to facilitate data interpretation and re-usability
3. Unambiguously reported genetic variants for standard annotation
4. A set of mandatory (i.e. must be present and filled with non-null values) fields, providing the information necessary to enable a wide range of data analyses including MR and PGS development
5. A set of encouraged fields with standard headers, which are strongly recommended but not mandatory
6. A balance between these mandatory and encouraged fields that includes essential data but does not set the bar impossibly high for the community using and implementing the standard
7. A low bioinformatics requirement for data consumers and data producers, reflecting the composition of the user community, to maximise stakeholder uptake

These requirements were used to define the backbone of a format—the GWAS-SSF—which will be implemented within the GWAS Catalog and promoted more widely in the community. The format has been designed to be interoperable with other major formats and resources. Queries can be addressed to gwas-info@ebi.ac.uk or raised as issues here: `https://github.com/EBISPOT/gwas-summary-statistics-standard`.

3 Specification

The GWAS summary statistics format (GWAS-SSF) is composed of two files, the summary statistics data file and accompanying metadata file.

3.1 Summary statistics data format

The GWAS-SSF data file is a TSV flat file of tab-delimited values that can be compressed (see Figure 1 for a schematic representation), reporting data from a single genome-wide analysis. The first line of the file contains the headers to the table. The rows after the header store the variant association data. Where permitted, values can be omitted and we recommend the use of `#NA` for missing value representation. There are no limits to the number of rows or columns that the table can have, however, a set of mandatory fields (defined in Table 1) must be present in a defined order. A file may contain additional columns beyond the set of mandatory fields. Table 1 shows some non-mandatory (encouraged) fields that may be present.

3.1.1 Example file data

chromosome	base_pair_location	effect_allele	other_allele	beta	standard_error	effect_allele_frequency	p_value	variant_id	rsid	ref_allele
1	869388	A	G	-0.016619	0.00806496	0.997221	0.1	1_869388_A_G	#NA	EA
1	205813916	G	C	-0.0089589	0.00331941	0.983589	9.7E-03	1_205813916_G_C	rs74143855	EA
2	70478797	T	TG	0.0187528	0.00167685	0.934121	3.5E-30	2_70478797_T_TG	rs142640435	EA
7	8458030	TC	T	-0.0184003	0.00101051	0.78451	5.7E-76	7_8458030_TC_T	rs774624811	EA
23	24173186	A	C	0.00387762	0.08757958	0.627178	2.3E-08	23_24173186_C_A	rs5949233	OA

In this example, the summary statistics data file (TSV) has been pretty-printed to display the columns more clearly. The first line contains the column labels and every line thereafter are for variant-trait association data. Column labels and column order are in adherence to the definitions in Table 1. *variant_id* and *rsid* are optional (encouraged) they are simply placed anywhere after the mandatory 8 columns. Here the effect statistic is beta, so the column label of the effect size column is *beta*. The first data row represents a variant-trait association for a single-nucleotide polymorphism where the effect allele is an 'A' at the genomic location (genome assembly is given in the accompanying metadata file, see **??**). No rsID was provided for this first variant, so *#NA* was given as the value in the *rsid* column because there must be a value in all columns. The second data row shows an example where the p-value is given in scientific notation and rsID is provided. The third and fourth data rows are examples of deletions and insertions, respectively. The fifth example shows a variant located on the X-chromosome, which is mapped to *23* (Table 1).

Figure 1: Schematic representation of the summary statistics table. Examples of data content within each specific field are provided in Table 1.

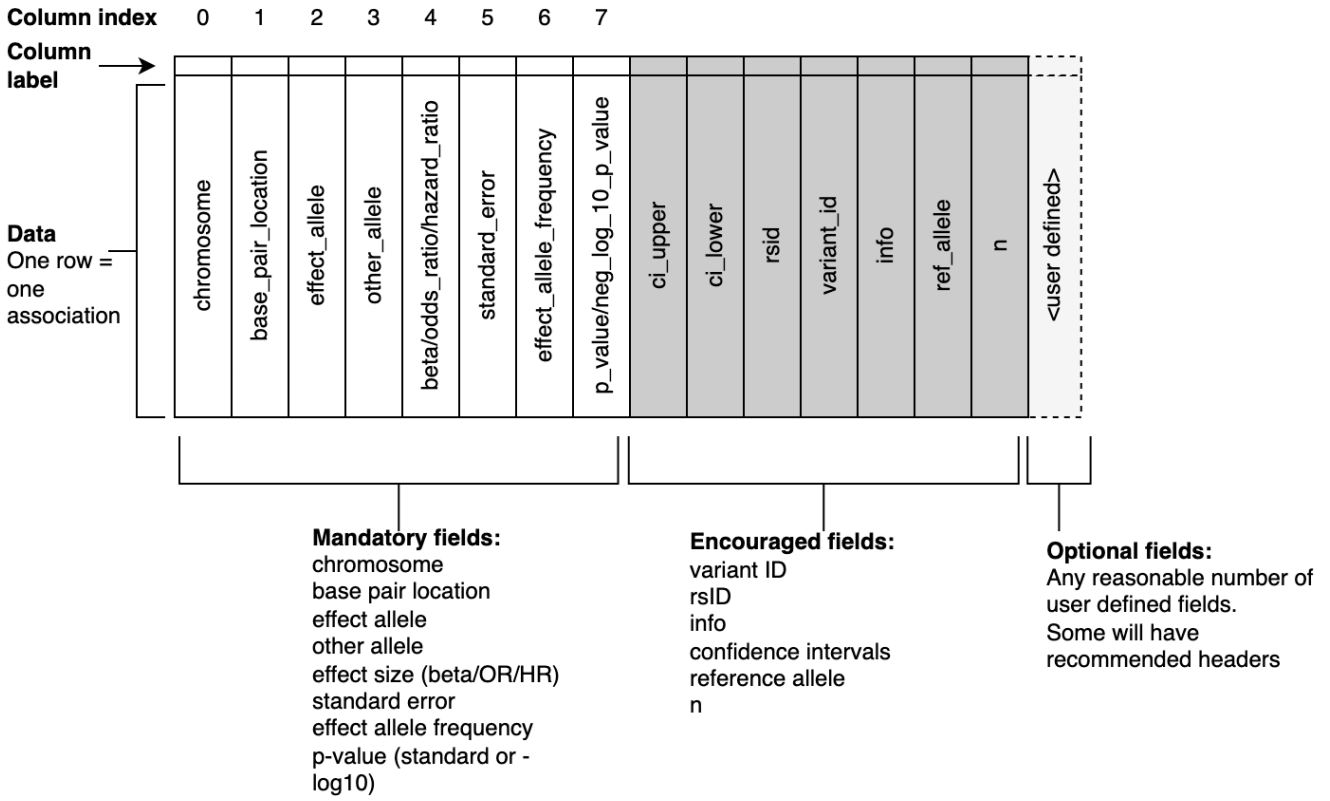

Table 1: Summary statistics field definitions

Field name	Description	Accepted values	Field type
chromosome	Column 0: Chromosome where the variant is located (X=23, Y=24, MT=25)	[1-25]	Mandatory
base_pair_location	Column 1: The first position of the variant in the reference, using the coordinate system declared (*coordinate_system* in metadata)	$x > 0$	Mandatory
effect_allele	Column 2: Allele associated with the effect	[ACGT]+	Mandatory
other_allele	Column 3: The non-effect allele	[ACGT]+	Mandatory
beta	Column 4: Effect size as beta	Numeric	Mandatory that either *beta*, *odds_ratio* or *hazard_ratio* is given
odds_ratio	Column 4: Effect size as odds ratio	$x \geq 0$	As above
hazard_ratio	Column 4: Effect size as hazard ratio	$x \geq 0$	As above
standard_error	Column 5: Standard error	Numeric	Mandatory
effect_allele_frequency	Column 6: Frequency of the effect allele	$0 \leq x \leq 1$b	Mandatory
p_value	Column 7: P-value of the association statistic	$0 \leq x \leq 1$ a	Mandatory that either *p_value* or *neg_log_10_p_value* is given
neg_log_10_p_value	Column 7: $-\log_{10}$ of above	$x \geq 0$ a	As above
ci_upper	Upper confidence interval	Numeric	Encouraged
ci_lower	Lower confidence interval	Numeric	Encouraged
rsid	rsID	^rs[0-9]+$	Encouraged
variant_id	An internal variant identifier formed by concatenating *chromosome*, *base_pair_location*, reference allele and alternate allele with underscores	[1-25]_[0-9]+_([ACGT]+_[ACGT]+|LONG_STRING)c	Encouraged
info	Imputation information metric	$0 \leq x \leq 1$	Encouraged
n	Sample size	Numeric	Encouraged
ref_allele	State which of the alleles is the reference allele	"EA" for *effect_allele*, "OA" for *other_allele* or "#NA" if unknown	Encouraged
hm_code	Harmonisation code, corresponding to transformation that occurred	Numeric	Only given in harmonised datasets

If p-value is equal to 0, the precision of the p-value calculation must be given in the accompanying metadata.

Effect allele frequency can be rounded up to a threshold value defined in the metadata.

'LONG_STRING' can be used where allele string is too long to be represented.

3.2 Summary statistics table contents

Four fields in the summary statistics table, combined with the reference genome assembly provided in a metadata file (see below), define the genetic variants (all field definitions can be found in Table 1). These fields are the chromosome (*chromosome*), the genomic location position on the chromosome (*base_pair_location*), the effect allele (*effect_allele*), and the non-effect allele (*other_allele*). Chromosome values are integers from 1 to 25, with chromosome X mapping to 23, chromosome Y to 24, and mitochondrial to 25. Genomic location is an integer value representing the first position of the variant in the reference genome, using the coordinate system specified in the metadata, either 1-based or 0-based. The *effect_allele* field captures the allele for which the effect is associated with, while the *other_allele* field reports the non-effect allele. Both of the allele fields will contain allele strings, including cases where variants are insertions and deletions. *variant_id* is for storing *chromosome*, *base_pair_location*, reference allele and alternate allele information as a single concatenated (with underscores) field and rsID can be stored in the *rsid* field, but both fields are optional. The reference allele can be declared using the encouraged *ref_allele* field, stating whether the reference allele is the *effect_allele* ("EA") or the *other_allele* ("OA"). All rows contain the following association statistics: p-value (either *p_value* or *neg_log_10_p_value*), the effect size (either *beta*, *odds_ratio* or *hazard_ratio*), and the standard error (*standard_error*). Depending on the precision of software that performed the calculation of association, p-values in GWAS analyses may appear rounded to zero or one. This is particularly problematic where highly significant associations (e.g. $p < 10^{-300}$) are rounded to zero, preventing associations being ranked in order of significance. Calculation of accurate p-values is recommended where possible. Where this is not possible due to limitations of the software used, the GWAS-SSF requires the analysis and genotype imputation software and version to be present in the metadata, to help users of the summary statistics interpret these values. P-values can be expressed as negative log values, in which case the field at index 7 should be labelled *neg_log_10_p_value* instead of *p_value*. Effect allele frequency (*effect_allele_frequency*) is a mandatory field. However, where privacy concerns might otherwise be a barrier to sharing the data, a cutoff may be specified in the metadata (*minor_allele_freq_lower_limit* field, see Table 2) so that frequencies below that cutoff are rounded-up to mask their true values. For example, *minor_allele_freq_lower_limit* = 0.01 in the metadata file would communicate that the lowest possible value for the effect allele frequency in this file is 0.01, and anything below this threshold has been rounded up to 0.01.

3.3 Summary statistics metadata

An additional file accompanies the summary statistics data file containing metadata describing the summary statistics such as the name and md5sum of the summary statistics data file (see Section **??** for example) and the GWAS metadata itself, including sample and experimental metadata (Table 2), thereby ensuring the reusability of the data. The metadata file fields can be expanded as needed in the future, and as with the summary statistics file, additional columns can be included as required. Sample metadata fields include descriptions of the trait under investigation and the sample size and ancestry. An additional field *ancestry_method* can be used to indicate whether the ancestry descriptor is self-reported or genetically defined (encouraged). We recommend that ancestry is described according to the standardised framework guidelines described in Morales et al, 2018 [**?**]. Every effort should be made to explicitly note whether the sample is admixed and the ancestral backgrounds that contribute to admixture. The trait description is free text and should include a clear description of the trait under study, including any relevant background characteristics of the study population, e.g. "lung cancer in asthma patients". Trait ontology terms can be stored in the metadata *ontology_mapping* field. The metadata file is in YAML format, which is "a human-friendly data serialisation language for all programming languages" (`https://yaml.org/`). There are both mandatory and encouraged metadata fields, which are detailed in Table **??**.

3.3.1 Metadata example

```
---
# Study meta-data
gwas_id: GCST90000123
author_notes: Summary statistics of a meta-analysis
gwas_catalog_api: https://www.ebi.ac.uk/gwas/rest/api/studies/GCST90000123
date_metadata_last_modified: 2024-01-16

# Trait Information
trait_description:
  - breast carcinoma
ontology_mapping:
  - EFO_0000305

# Genotyping Information
genome_assembly: GRCh37
coordinate_system: 1-based
genotyping_technology:
  - Genome-wide genotyping array
imputation_panel: 1000 Genomes Phase 3
imputation_software: IMPUTE

# Sample Information
samples:
  - sample_ancestry_category:
      - European
    sample_size: 12345
    ancestry_method:
      - self-reported
      - genetically determined
    case_control_study: true
sex: combined

# Summary Statistic information
data_file_name: GCST90000123.tsv
file_type: GWAS-SSF v1.0
data_file_md5sum: 32ce41c3dca4cd9f463a0ce7351966fd
analysis_software: PLINK 1.9
adjusted_covariates:
  - age
  - sex
minor_allele_freq_lower_limit: 0.001

# Harmonization status
is_harmonised: false
is_sorted: false
```

Table 2: Metadata field definitions

Field	Description	Accepted value	Mandatory
genome_assembly	Genome assembly	GRCh/NCBI/UCSC value	Yes
coordinate_system	Coordinate System	1-based/0-based	Yes
trait_description	Author reported trait description	Text string (multiple possible)	Yes
sample_size	Sample size	Integer	Yes
case_count	Number of cases for case/control study	Integer	No, unless caseControl-Study is true
control_count	Number of controls for case/control study	Integer	No, unless caseControl-Study is true
case_control_study	Flag whether the study is a case-control study	Boolean	No (default is false)
sample_ancestry	Sample ancestry	Text string (multiple possible)	No
sample_ancestry_category	Sample ancestry board category	Text string (multiple possible)	Yes
genotyping_technology	Genotyping technology	Text string (multiple possible)	Yes
analysis_software	Software and version used for the association analysis	Text string	Yes if p-values of 0 given
imputation_panel	Imputation panel	Text string	No
imputation_software	Software used for imputation	Text string	No
minor_allele_freq_lower_limit	Lowest possible minor allele frequency	Numeric	No
ancestry_method	Method used to determine sample ancestry e.g. self-reported/genetically determined	Text string (multiple possible)	No
is_sorted	Flag whether the file is sorted by genomic location	Boolean	Yes
is_harmonised	Flag whether the file is harmonised	Boolean	Yes
hm_code_definition	Description of harmonisation codes	Text string	Only given in harmonised datasets
adjusted_covariates	Any covariates the GWAS is adjusted for	Text string (multiple possible)	No
ontology_mapping	Short form ontology terms describing the trait	Text string (multiple possible)	No
sex	Indicate if and how the study was sex-stratified	"M", "F", "combined" or "#NA"	No