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

24th May 2022

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. 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. We continue to take public feedback on the proposed format via our github repository `https://github.com/EBISPOT/gwas-summary-statistics-standard` or via email to gwas-info@ebi.ac.uk.

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 by the presence of `NA`. 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
1	869388	A	G	-0.016619	0.00806496	0.997221	0.1	1_869388_A_G	NA
1	205811055	C	T	-0.0089589	0.00331941	0.983589	9.7E-03	1_205811055_C_T	rs74143854
2	70478797	T	TG	0.0187528	0.00167685	0.934121	3.5E-30	2_70478797_T_TG	rs142640435
2	27875036	TAAA	T	-0.0184003	0.00101051	0.78451	5.7E-76	2_27875036_TAAA_T	rs774624803
X	24145170	A	G	0.00387762	0.08757958	0.627178	2.3E-08	X_24145170_A_G	rs5949232

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 headers in the order defined by Table 1. *variant_id* and *rsid* are optional and can placed anywhere after the mandatory 8 columns.

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

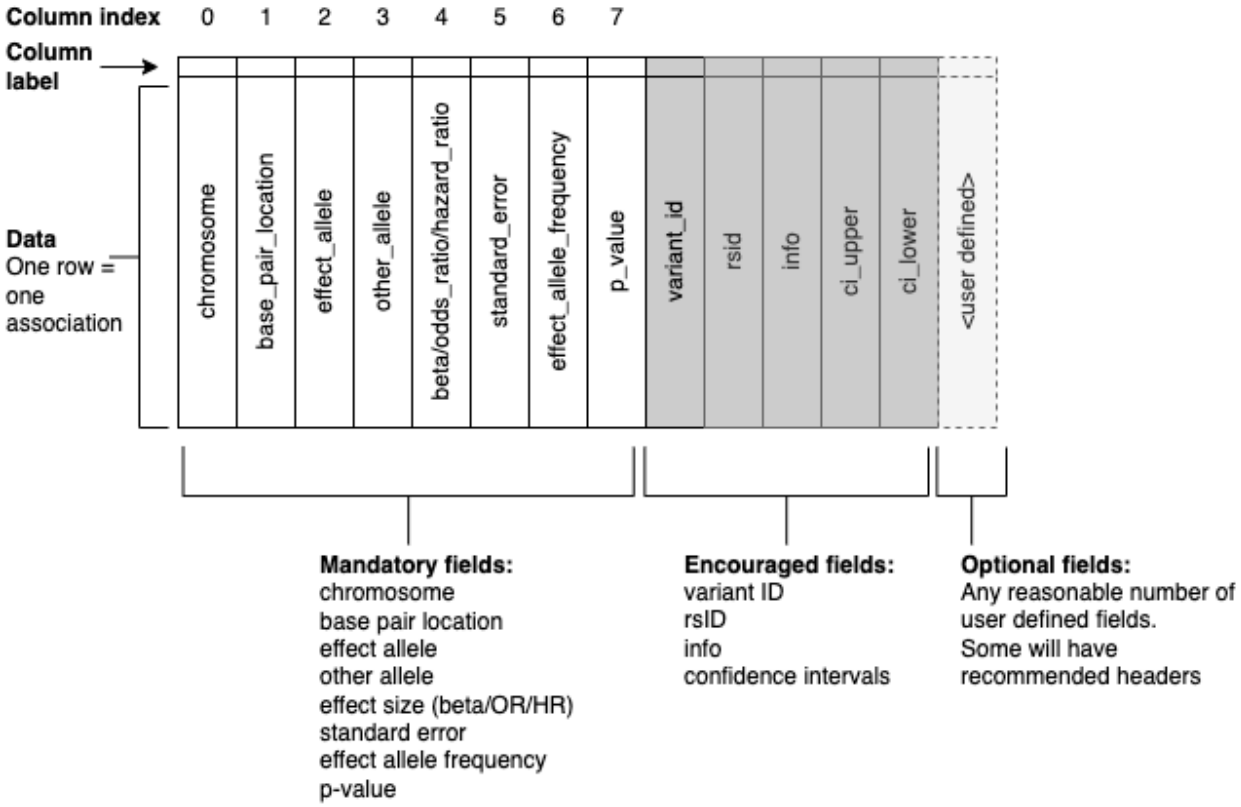


Table 1: Summary statistics field definitions

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, counting on the bases, from 1 (1-based)	$x > 0$	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	$x >= 0$	As above
hazard_ratio	Column 4: Effect size as hazard ratio	$x >= 0$	As above
standard_error	Column 5: Standard error	Numeric	Mandatory
effect_allele_frequency	Column 6: Frequency of the effect allele	$0 =< x <= 1$[b]	Mandatory
p_value	Column 7: P-value of the association statistic	$0 =< x <= 1$ or $x >= 0$ if p_value is in the $-log_{10}$ form[a]	Mandatory
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*, *other_allele* and *effect_allele* with underscores	[1-25]_[0-9]+_([ACGT]+_[ACGT]+|LONG_STRING)[c]	Encouraged
info	Imputation information metric	$0 =< x <= 1$	Encouraged
n	Sample size	Numeric	Encouraged
hm_code	Harmonisation code, which can be looked up in the metadata to determine the transformation	Numeric	Only given in harmonised datasets

[a] If p-value is equal to 0, the precision of the p-value calculation must be given in the accompanying metadata.
[b] Effect allele frequency can be rounded up to a threshold value defined in the metadata.
[c] '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), unambiguously 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 1-based indexing (see Figure 2) to maximise interoperability with variant call format (VCF) (Danecek et al 2011). 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 (see Figure 2). These four fields (*chromosome, base_pair_location, effect_allele, other_allele*) are concatenated to populate the *variant_id* field and rsID can be stored in the *rsid* field, but both fields are optional.

All rows contain the following association statistics: p-value (*p_value*), the effect size (*effect_size*), and the standard error (*standard_error*). Whether the effect size is given as beta or odds ratio is defined in the metadata in the field *effectStatistic*. 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. Alternatively, p-values can be expressed as negative log values, in which case the metadata field *pvalueIsNegLog10* should be set to true. 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 (*effectAlleleFreqLowerLimit* field, see Table 2) so that frequencies below that cutoff are rounded-up to mask their true values. For example, *effectAlleleFreqLowerLimit* = 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 Supplement 1 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 ancestryMethod 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 ontologyMapping 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 2.

3.3.1 Metadata example

```
---
# GWAS Catalog Summary Statistics Metadata
summaryStatisticsMetadata:
  genotypingTechnology:
    - Genome-wide genotyping array
  GWASCatalogStudyAccession: GCST90000123
  sampleSize: 12345
  sampleAncestry:
    - European
  traitDescription:
    - breast carcinoma
  effectAlleleFreqLowerLimit: 0.001
  ancestryMethod:
    - self-reported
    - genetically determined
  caseControlStudy: false
  dataFileName: 0000123.tsv
  fileType: GWAS-SSF v0.1
  md5sum: 5b00c03a568bca2fcd0a09f1bf4f77fa
  harmonised: false
  fileDescription: GWAS summary statistics; author uploaded.
  dateLastModified: 01-08-2021
  genomeAssembly: GRCh37
  sortedByGenomicLocation: false
  effectStatistic: beta
  pvalueIsNegLog10: false
```

4 References

Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-2158. doi:10.1093/bioinformatics/btr330

Morales J, Welter D, Bowler EH, Cerezo M, Harris LW, McMahon AC, Hall P, Junkins HA, Milano A, Hastings E, Malangone C, Buniello A, Burdett T, Flicek P, Parkinson H, Cunningham F, Hindorff LA, MacArthur JAL. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biol. 2018 Feb 15;19(1):21. doi: 10.1186/s13059-018-1396-2. PMID: 29448949; PMCID: PMC5815218.

Table 2: Metadata field definitions

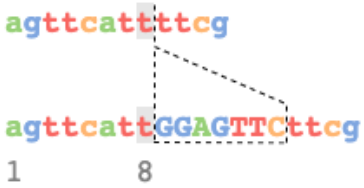
Field	Description	Accepted value	Mandatory
genomeAssembly	Genome assembly	GRCh/NCBI/UCSC value	Yes
traitDescription	Author reported trait description	Text string (multiple possible)	Yes
sampleSize	Sample size	Integer	Yes
caseCount	Number of cases for case/control study	Integer	No, unless caseControlStudy is true
controlCount	Number of controls for case/control study	Integer	No, unless caseControlStudy is true
caseControlStudy	Flag whether the study is a case-control study	Boolean	No (default is false)
sampleAncestry	Sample ancestry	Text string (multiple possible)	Yes
genotypingTechnology	Genotyping technology	Text string (multiple possible)	Yes
analysisSoftware	Software and version used for the association analysis	Text string	Yes if p-values of 0 given
imputationPanel	Imputation panel	Text string	No
imputationSoftware	Software used for imputation	Text string	No
effectAlleleFreqLowerLimit	Lowest possible effect allele frequency	Numeric	No
ancestryMethod	Method used to determine sample ancestry e.g. self-reported/genetically determined	Text string (multiple possible)	No
sortedByGenomicLocation	Flag whether the file is sorted by genomic location	Boolean	Yes
effectStatistic	Indicate whether beta or odds ratio is used	beta/odds ratio/hazard ratio	yes
hmodeDefinition	Description of harmonisation codes	Text string	Only given in harmonised datasets
pvalueIsNegLog10	Flag whether p value is given as $-log_{10}$	Boolean	No (default is false)
adjustedCovariates	Any covariates the GWAS is adjusted for	Text string (multiple possible)	No
ontologyMapping	Short form ontology terms describing the trait	Text string (multiple possible)	No

Figure 2: Illustration of how variants are recorded in the summary statistics table for (a) SNP, (b) insertion, and (c) deletion alleles. Note that for insertions and deletions, the position of the base preceding the indel (the highlighted T at 8) is the position used to index the variant.

a. Single nucleotide polymorphism (effect allele of C at position 8)

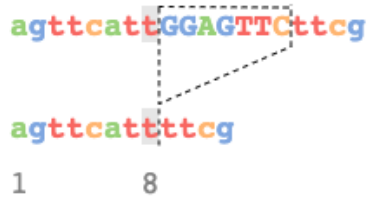
chromosome	base_pair_location	other_allele	effect_allele
11	8	T	C

b. Insertion (effect allele has an insertion of GGAGTTC between positions 8 and 9)



chromosome	base_pair_location	other_allele	effect_allele
11	8	T	TGGAGTTC

c. Deletion (effect allele has a deletion of GGAGTTC from positions 9-15)



chromosome	base_pair_location	other_allele	effect_allele
11	8	TGGAGTTC	T