A community driven GWAS summary statistics standard

James Hayhurst[1,2], Annalisa Buniello[1,2], Laura Harris[1], Abayomi Mosaku[1], Christopher Chang[3], Christopher R. Gignoux[4], Konstantinos Hatzikotoulas[5], Mohd Anisul Karim[2,6], Samuel A. Lambert[7,8,9,] Matt Lyon[10,11], Aoife McMahon[1], Yukinori Okada[12,13,14], Nicola Pirastu[15,16], N. William Rayner[5], Jeremy Schwartzentruber[2,6], Robert Vaughan[17], Shefali Verma[18], Steven P. Wilder[19], Fiona Cunningham[1], Lucia Hindorff[20], Ken Wiley[20], Helen Parkinson[1], and Inês Barroso[21]

1. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
2. Open Targets, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
3. GRAIL, LLC, Menlo Park, California, 94025, USA
4. Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA
5. Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
6. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
7. Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
8. British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
9. Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK
10. National Institute for Health and Care Research (NIHR) Bristol Biomedical Research Centre, University of Bristol, Oakfield House, Bristol, BS8 2BN, UK
11. Medical Research Council (MRC) Integrative Epidemiology Unit (IEU), Bristol Medical School (Population Health Sciences), University of Bristol, Oakfield House, Bristol, BS8 2BN, UK
12. Osaka University Graduate School of Medicine, Suita, 565-0871, Japan
13. The University of Tokyo, Tokyo, Japan
14. RIKEN Center for Integrative Medical Sciences, Yokohama, 230-0045, Japan
15. Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, UK
16. Genomics Research Centre, Human Technopole, Milan, Italy
17. Congenica Ltd, Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, UK.
18. Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
19. Genomics Plc, King Charles House, Oxford, Park End Street, OX1 1JD, UK
20. National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.

21. Exeter Centre of Excellence for Diabetes Research (EXCEED), University of Exeter Medical School, Exeter, UK

Abstract

Summary statistics from genome-wide association studies (GWAS) represent a huge potential for research. A challenge for researchers in this field is the access and sharing of summary statistics data due to a lack of standards for the data content and file format. For this reason, the GWAS Catalog hosted a series of meetings in 2021 with summary statistics stakeholders to guide the development of a standard format. The key requirements from the stakeholders were for a standard that contained key data elements to be able to support a wide range of data analyses, required low bioinformatics skills for file access and generation, to have easily accessible metadata, and unambiguous and interoperable data. Here, we define the specifications for the first version of the GWAS-SSF format, which was developed to meet the requirements discussed with the community. GWAS-SSF consists of a tab-separated data file with well-defined fields and an accompanying metadata file.

Introduction

Summary statistics are defined as the aggregate p-values and association data for every variant analysed in a genome-wide association study (GWAS). The depth of information contained in the summary statistics represents huge potential to extend the power of GWAS and improve disease understanding. In recent years a number of methods have been developed to enable the use of GWAS summary statistics to gain insights into the mechanisms of complex disease, identify new drug targets and evaluate disease risk. Example methods include large meta-analyses (Wheeler E. 2017), trait pleiotropy (Smeland O.B, 2017), prediction using polygenic scores (PGS) (Lambert S, 2019) and Mendelian randomisation (MR) (Paternoster L, 2017). However, a considerable number of summary statistics are still not fully and openly shared with the community, either being made available under controlled access, upon agreement to restrictive terms, with incomplete data, or not shared at all. One of the main challenges associated with sharing full GWAS results is the lack of standards for data content and format, meaning that researchers do not have clear guidelines for appropriate file generation for sharing, and the re-usability of the resulting files can be poor. Typically, each GWAS will produce a single file with a table of summary statistics containing a list of variants with p-values, other statistics and relevant annotations or metadata. Generated by different software packages and made available via different resources, summary statistics can vary in a myriad of ways from one study to the next. A recent analysis of 327 summary statistics files found over 100 unique formats (Murphy et al, 2021). Differences in file formats, header definitions, data types, genetic variant or association data reporting and missing data create challenges for users by reducing data interoperability.

The GWAS Catalog began hosting summary statistics in 2018, and rapidly developed a first minimal data format based on the most commonly included fields in publicly available files (Buniello, 2019), but without community input. In parallel, other summary statistics formats have been defined for specific purposes, e.g. dbGap's Minimum Information Required for Association Data guidelines, designed to fulfil data sharing requirements in dbGaP (https://www.ncbi.nlm.nih.gov/gap/docs/submissionguide/#apha); GWAS-VCF (Lyon, 2021) developed for robustness and performance and to underpin the OpenGWAS platform (Elsworth, 2020) and associated tools. Limitations to more widespread adoption of the GWAS-VCF are that it requires knowledge of bioinformatics and relevant tools to parse and prepare, which present a barrier to data sharing for some users.

The field is advancing rapidly and summary statistics data sharing is quickly becoming more common. More than 70% of GWAS Catalog studies are now linked to freely accessible summary statistics (27,500 studies from 550 publications) with the highest yearly increment observed in 2020-2021, and 77% of summary statistics submissions to the GWAS Catalog in 2021 were made before publication, upon a journal's mandate. These metrics show that GWAS summary statistics have now reached a critical mass, and to maximise the utility of this body of data there is a need for the community to adopt a standard to which users can expect all studies to adhere (MacArthur et al, 2021). A single standard with stricter definitions on the data included will increase the utility of GWAS summary statistics, reduce the risk of misinterpretation of data and enable users to easily analyse and integrate data from different GWAS. A range of mandatory data fields are required to support the major use cases for downstream analysis, such as PGS development, MR, meta-analysis and functional annotation of variants, at scale.

Methods

Following initial discussions with the GWAS community at the 2020 GWAS summary statistics standards and sharing workshop (MacArthur et al, 2021), the GWAS Catalog hosted a series of meetings between June 2021 and September 2021 with invited 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 finalise minimum information elements for data sharing to maximise downstream utility, and to complete a phase of iteration on the proposed standard.

An initial set of standard reporting elements had been proposed in MacArthur et al, based on a public survey and workshop discussion. The first meeting reviewed these findings, assessed currently available formats with their strengths and weaknesses, and then comprised a guided discussion focusing on the details of a potential new format. Topics covered included user considerations, in particular the balance of requirements for data consumers and data generators; data reporting requirements including consistent and unambiguous variant reporting and p-value precision; mandatory and optional fields for data and metadata; storage and access of metadata; scaling considerations. Following the first meeting we circulated a survey amongst meeting attendees to summarise opinions on variant reporting, association reporting, metadata fields and location, and file format. The outcome of the survey was reviewed and discussed at the second meeting with the goal of reaching consensus on requirements. Based on the identified requirements, GWAS Catalog staff developed a proposed format which was presented and iterated at the third meeting in Sept 2021.

Impact assessment

In order to evaluate the impact of the proposed format on data sharing, we assessed the content of existing submissions to the GWAS Catalog. To date, the minimum requirements for submission are a p-value with either rsID or chromosome and base-pair location, with other fields supplied at the submitter's discretion.

Taking all author-submitted single-variant GWAS summary statistics files in the GWAS Catalog (315 submissions, 27845 GWAS) as a sample, we used GNU grep to search all the files for the newly defined mandatory fields from the proposed format. We searched the first row of each file for a) the presence of each mandatory field; and b) the presence of all the mandatory fields (an OR operator was used between beta and odds_ratio).

Results

Requirements

The key requirements for an expanded GWAS summary statistics standard obtained from the stakeholders’ use cases were as follows:

- Consistent representation of data to enable interoperability
- Easily accessible metadata for summary statistics to facilitate data interpretation and re-usability
- Unambiguously reported genetic variants for standard annotation
- 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
- A set of encouraged fields with standard headers, which are strongly recommended but not mandatory

- 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
- 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.

GWAS-SSF, a newly proposed GWAS summary statistics format

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

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, Supplement 1 for example file), 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.

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, 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 (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. 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.

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 2 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 reported 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 o*ntologyMapping* 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.

Impact assessment

Of 27,845 valid summary statistics files obtained by the GWAS Catalog since the release of our submission system in 2020, ~17% were missing at least one of the new mandatory fields (Fig 3(a)). The most commonly omitted field was Effect Allele Frequency, followed by standard error and effect size (beta/OR). Each submission may contain many files, and it is reasonable to assume that all the files within a submission adhere to the same format, being generated as part of the same project. We therefore wished to ascertain the proportion of submissions that were missing data, as this may be more indicative of practices within the data-generating community. More than 50% of submissions omitted at least one of the new mandatory fields, with the most commonly omitted field being Effect Allele Frequency, followed by standard error, other allele and effect size (beta/OR) (Fig 3(b)).

These results show that a substantial portion of GWAS summary statistics shared under minimal requirements is severely limited in usability for downstream purposes. In the case of effect size, the summary statistics will be rendered unusable for the majority of methods that leverage summary statistics. Since more than 50% of submissions omitted at least one of the new mandatory fields, implementation of GWAS-SSF has the potential to double the number of usable datasets for important downstream uses.

Discussion

Community activities have been effective in the development of agreed standards and sharing principles for scientific data (e.g. Brazma et al, 2001). The GWAS summary statistics format (GWAS-SSF) presented here is the result of meetings with the community to make a simple, easy to access standard which promotes cross-dataset consistency and is useful for varied use cases. The mandatory content of the table meets the requirement set by the stakeholders of the working group to perform most analyses e.g. beta and standard error to support MR analyses, effect and non-effect allele to support meta-analysis and generation of polygenic scores. Another requirement from the working groups is a consistent approach to variant reporting and representation which is important for users of the data to be able to easily merge or compare datasets. By adopting the variant reporting standard embodied in VCF (Danecek et al 2011) to define single nucleotide polymorphisms and short indels, consistency and interoperability will be achieved. More complicated variants (e.g. structural variants) and their shorthand notations fall outside the primary scope of GWAS summary statistics. Regarding file type, large numbers of GWAS summary statistics have been stored in the GWAS-VCF format (Lyon et al 2021; Elsworth et al 2020), but less-technical stakeholders preferred a TSV file and we have (in collaboration with representatives from the MRC IEU) codesigned a generic TSV/YAML format and maintained interoperability with VCF.

Metadata for the summary statistics files and study design are available in a separate file. There are advantages to storing metadata within data files, primarily that the metadata and summary statistics cannot become inadvertently decoupled, but this complicates file parsing whereas a generic tabular file format is universally accessible. The metadata is therefore an optional source that can help reduce ambiguity and provide useful information about the datasets.

The GWAS-SSF is designed to represent each GWAS analysis in a separate file, and in this respect differs from the GWAS-VCF which can represent multiple phenotypes in the same file. Although there are some advantages in sharing data between individual users in this way, the number of GWAS per unit is rapidly growing, for example >18K phenotypes in Wang et al 2021, and this may cause usability issues to the average user where large volumes of data are stored in a single file. Data stored in the GWAS-SSF with the required data elements can be easily converted to GWAS-VCF if required using publicly available tools (Lyon et al 2021).

GWAS-SSF includes a number of mandatory fields, and we heard from our working group that many more fields may be important in certain contexts, e.g. imputation info for filtering variants to identify those of high enough quality for downstream analyses, such as fine-mapping, enrichment analyses, MR, or genetic correlation estimation. However, there was an acceptance that these may not be readily available or necessary for all users and their absence should not preclude data sharing and reuse. The standard should promote open data sharing as widely as possible, while providing the essential information for most major downstream uses. We have therefore included additional encouraged fields with standard headers to promote interoperability, and data generators are strongly encouraged to share these data unless they are genuinely unavailable (for example, in the case of historical/legacy data) or there is a scientific or ethical reason not to (e.g. privacy concerns). Furthermore, the list of standard fields is not intended to be exhaustive and data generators are encouraged to share as much additional data as possible.

FAIRification of GWAS results is currently a significant challenge for the genetics community, as thoroughly discussed in our working group meetings and reported in this work. The new community driven GWAS summary statistics format we propose conforms to the FAIR principles (Wilkinson et al 2016) and we believe that its widespread adoption will facilitate sharing and usability of summary statistics in the public domain.

Implementation in the GWAS Catalog and other resources

Support for data submission

Previous discussions with the community suggested that the burden of formatting data should lie with data generators (MacArthur et al, 2021). However, with increasingly large datasets, the effort and associated cost required in preparing files for submission to repositories is a key concern raised by our working group attendees and others (Kozlov 2022), and support is required to mitigate the burden. The GWAS Catalog will provide easy to use tools for formatting, converting to the standard headers and checking the validity of summary statistics files prior to upload to the GWAS Catalog, requiring input of a simple TSV file. Summary statistics submitted to the GWAS Catalog will continue to be accessioned and citable at point of submission, prior to journal publication.

Harmonised datasets

In addition to the author-submitted summary statistics that the GWAS Catalog makes available for download, eligible summary statistics are also made available to users in a harmonised format (Buniello, 2019). The harmonised data have been oriented to the same reference strand and variants normalised (left-aligned and trimmed to the shortest representation) so that the harmonised summary statistics of one study are interoperable with any other harmonised study. Harmonised files will adhere to the GWAS-SFF format described here but will benefit from being sorted and indexed by genomic location and compressed with bgzip*,* allowing fast retrieval of variants of interest by location. An additional field, *hm_code,* is present in the harmonised data files for storing a code indicating the transformation that was applied to harmonise the variant. The codes can be defined for reference in the associated metadata YAML file.

Remaining steps to first implementation of GWAS-SSF

A number of steps are required to fully implement the new standard, and these are under active development, with an estimated release date in late 2022.

1. Updated validator for submitted summary statistics
   The validator runs upon submission of summary statistics to the GWAS Catalog, and must pass in order for data to be successfully submitted. An offline version is provided for users to check the validity of their files prior to upload with detailed feedback provided on failures. The validator will be updated to ensure files adhere to the new format.
2. Generation of a metadata file
   In the GWAS Catalog submission tool, metadata can be entered via a simple Excel-based form. The submissions processing pipeline will be modified to generate a metadata YAML file upon release of summary statistics. The scripts used to do this will be made publicly available under the Apache version 2.0 open source license (https://www.apache.org/licenses/LICENSE-2.0). Metadata files will be generated retrospectively for all pre-existing summary statistics in the Catalog.
3. Updated harmonisation of summary statistics
   Formatted files are processed internally to produce the harmonised version, requiring no further input by the submitter. The harmonisation pipeline (https://github.com/EBISPOT/gwas-sumstats-harmoniser) is publicly available to enable data generators to produce their own harmonised versions. This pipeline will be changed to accommodate field changes, and harmonised files will be sorted and

indexed by genomic location optimised for fast retrieval of variants. The technology that will be used to do this is currently under investigation.

4. Provision of tools for the generation of GWAS-SSF
   PLINK (Purcell 2007), one of the most popular GWAS data analysis tools, has committed to creating an option to generate results files in the standard format, thus removing the need for data generators to further manipulate files after analysis prior to submission to the GWAS Catalog. We also plan to make available a formatting tool to easily convert from the outputs of other analysis softwares such as METAL.
5. Ensuring interoperability with other resources
   As outlined above, we have designed GWAS-SSF to be compatible with GWAS-VCF. dbGaP will accept submissions of GWAS summary statistics in the standard format to ensure flow of data between these two important public resources. We hope that other resources will follow suit to enhance interoperability and maximise the number of datasets that can be available in a central resource.

Acknowledgements

Special thanks to Mike Feolo, working group participants and the wider community for their engagement and contributions. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award no. U41HG007823 and EMBL-EBI Core Funds. In addition, we acknowledge funding from: the European Molecular Biology Laboratory; I.B., “Expanding excellence in England” award from Research England; ML, the MRC Integrative Epidemiology Unit (MC_UU_00011/4), supported by the NIHR Biomedical Research Centre at University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the NHS, the National Institute for Health and Care Research or the Department of Health.

References

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001 Dec;29(4):365-71. doi: 10.1038/ng1201-365. PMID: 11726920.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorff LA, Cunningham F, Parkinson H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D1005-D1012. doi: 10.1093/nar/gky1120. PMID: 30445434; PMCID: PMC6323933.

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

Ben Elsworth, Matthew Lyon, Tessa Alexander, Yi Liu, Peter Matthews, Jon Hallett, Phil Bates, Tom Palmer, Valeriia Haberland, George Davey Smith, Jie Zheng, Philip Haycock, Tom R Gaunt, Gibran Hemani. The MRC IEU OpenGWAS data infrastructure. bioRxiv 2020.08.10.244293v1. doi: 10.1101/2020.08.10.244293

Kozlov M. NIH issues a seismic mandate: share data publicly. Nature. 2022 Feb;602(7898):558-559. doi: 10.1038/d41586-022-00402-1. PMID: 35173323.

Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 2019 Nov 21;28(R2):R133-R142. doi: 10.1093/hmg/ddz187. PMID: 31363735.

Lyon MS, Andrews SJ, Elsworth B, Gaunt TR, Hemani G, Marcora E. The variant call format provides efficient and robust storage of GWAS summary statistics. Genome Biol. 2021 Jan 13;22(1):32. doi: 10.1186/s13059-020-02248-0. PMID: 33441155; PMCID: PMC7805039.

MacArthur et al, 2021 Workshop proceedings: GWAS Catalog standards and sharing. Cell Genomics Vol. 1, Issue 1, 100004.

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.

Murphy AE, Schilder BM, Skene NG. MungeSumstats: A Bioconductor package for the standardisation and quality control of many GWAS summary statistics. Bioinformatics. 2021 Oct 2;37(23):4593–6. doi: 10.1093/bioinformatics/btab665. Epub ahead of print. PMID: 34601555; PMCID: PMC8652100.

Paternoster L, Tilling K, Davey Smith G. Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. PLoS Genet. 2017 Oct 5;13(10):e1006944. doi: 10.1371/journal.pgen.1006944. PMID: 28981501; PMCID: PMC5628782.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559-75. doi: 10.1086/519795. Epub 2007 Jul 25. PMID: 17701901; PMCID: PMC1950838.

Smeland OB, Frei O, Kauppi K, Hill WD, Li W, Wang Y, Krull F, Bettella F, Eriksen JA, Witoelar A, Davies G, Fan CC, Thompson WK, Lam M, Lencz T, Chen CH, Ueland T, Jönsson EG, Djurovic S, Deary IJ, Dale AM, Andreassen OA; NeuroCHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) Cognitive Working Group. Identification of Genetic Loci Jointly Influencing Schizophrenia Risk and the Cognitive Traits of Verbal-Numerical Reasoning, Reaction Time, and General Cognitive Function. JAMA Psychiatry. 2017 Oct 1;74(10):1065-1075. doi: 10.1001/jamapsychiatry.2017.1986. PMID: 28746715; PMCID: PMC5710474.

Wang, Q., Dhindsa, R.S., Carss, K. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* 597, 527–532 (2021). https://doi.org/10.1038/s41586-021-03855-y

Wheeler et al, Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. PLoS Med. 2017 Sep 12;14(9):e1002383. doi: 10.1371/journal.pmed.1002383. PMID: 28898252; PMCID: PMC5595282.

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Figures & Tables

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

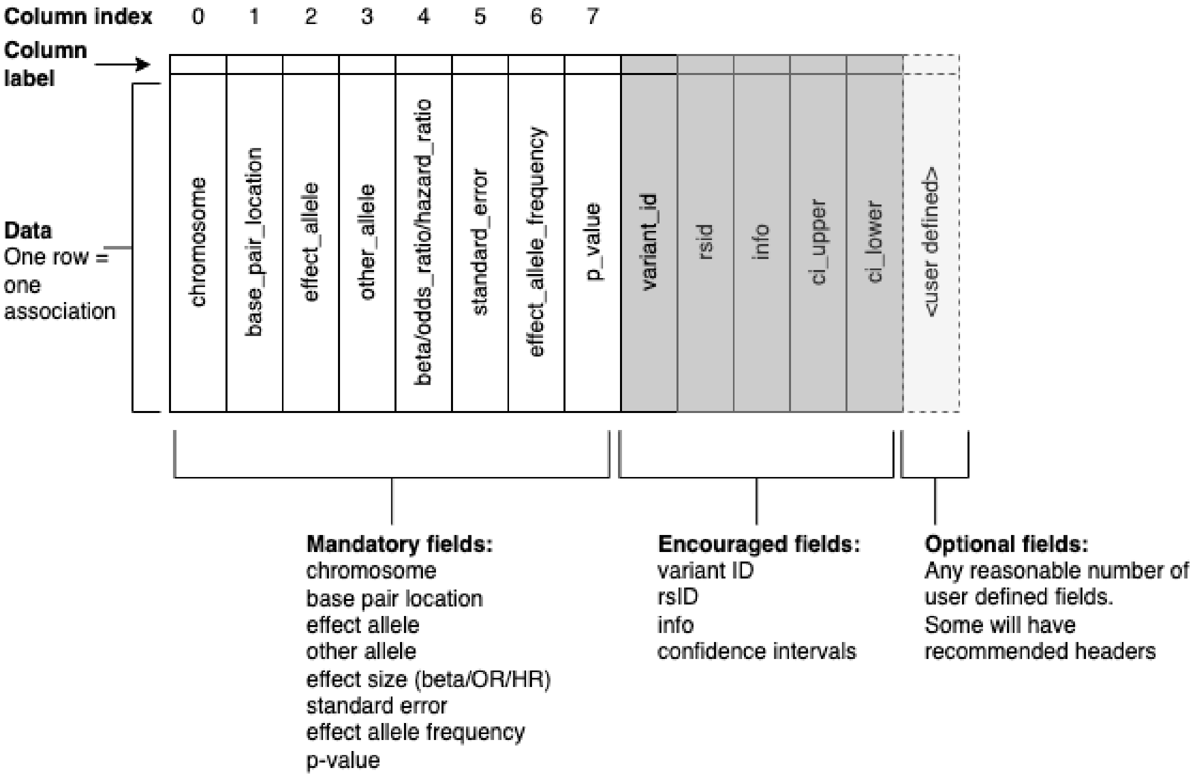


Table 1. Summary statistics field definitions. Note 1: If p-value is equal to 0, the precision of the p-value calculation must be given in the accompanying metadata. Note 2: Effect allele frequency can be rounded up to a threshold value defined in the metadata.

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 beta	Numeric	Mandatory that either beta, odds_ratio or hazard_ratio is given
odds_ratio	Column 4: odds ratio	x >= 0	Mandatory that either beta, odds_ratio or hazard_ratio is given
hazard _ratio	Column 4: hazard ratio	x >= 0	Mandatory that either beta, odds_ratio or hazard_ratio is given

standard_error	Column 5: Standard error	Numeric	Mandatory
effect_allele_frequency	Column 6: Frequency of the effect allele	0=<x<=1	Mandatory
p_value	Column 7: P-value of the association statistic	0=<x<=1 or x >= 0 if p_value is -log10	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 in the form of <chromosome>_<base_pair_location>_<other_allele>_<effect_allele>	[1-25]_[0-9]+_([ACGT]+_[ACGT]+|LONG_STRING)	Encouraged
info	Imputation information metric	0=<x<=1	Encouraged
n	Sample size	integer	Encouraged
hm_code	Harmonisation code, which can be looked up in the metadata to determine the transformation	integer	Only given in harmonised datasets

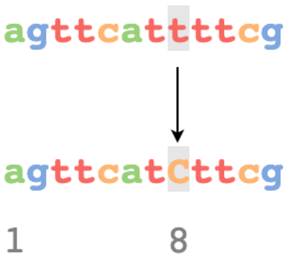
Table 2. Metadata field definitions

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	Boolean	Yes

	genomic location		
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 negative log10	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

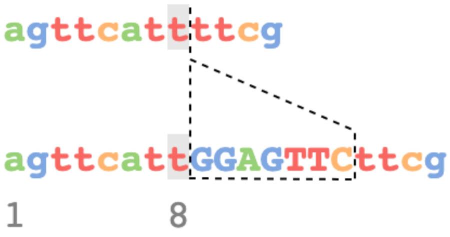
Figure 2. Illustration of how variants are recorded in the summary statistics table for (a) SNP , (b) insertion , and (c) deletion alleles . Note that for insertions and deletions, the position of the base preceding the indel (the highlighted T at 8) is the position used to index the variant.

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



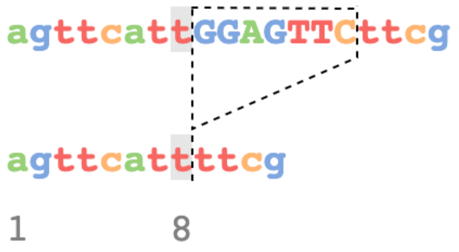
chromosome	base_pair_location	effect_allele	other_allele
11	8	C	T

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



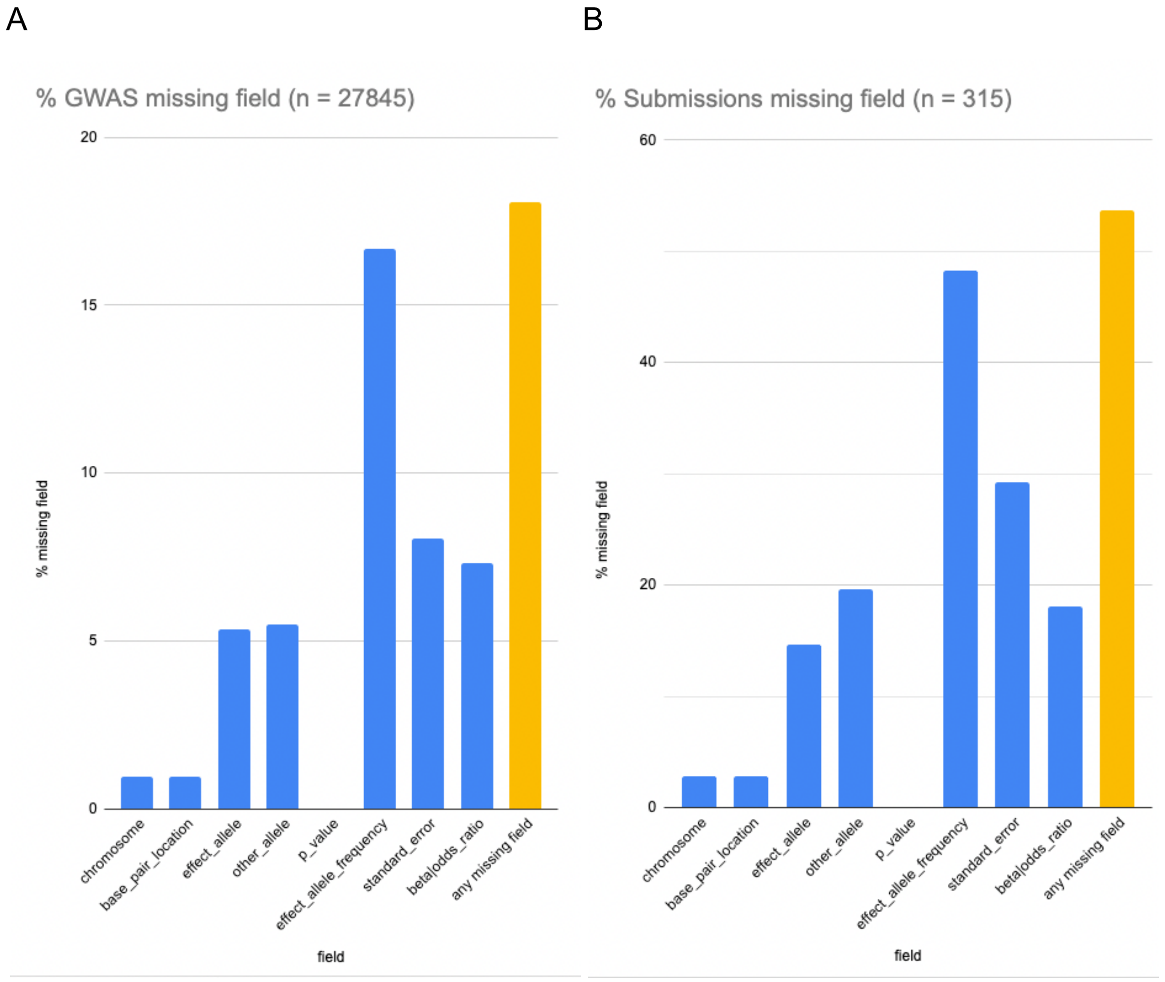
chromosome	base_pair_location	effect_allele	other_allele
11	8	TGGAGTTC	T

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



chromosome	base_pair_location	effect_allele	other_allele
11	8	T	TGGAGTTC

Figure 3: Assessment of missing data in summary statistics shared under minimal requirements, based on (a) individual GWAS datasets and (b) submissions, each typically associated with a single manuscript or project.



Supplement 1. Example of summary statistics TSV data file

```
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
23          24145170            A              G             0.00387762  0.08757958      0.627178                 2.3E-08  23_24145170_A_G    rs5949232
```

Supplement 2. Example of summary statistics metadata YAML file

```
---
# 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
```