A community driven GWAS summary statistics standard

James Hayhurst[1,2], **Annalisa Buniello**[1,2], **Laura Harris**[1], **Abayomi Mosaku**[1], **Christopher Chang**[3], **Mike Feolo**[4], **Christopher R. Gignoux**[5], **Konstantinos Hatzikotoulas**[6], **Mohd Anisul Karim**[2,7], **Samuel A. Lambert**[8,9,10], **Matt Lyon**[11,12], **Aoife McMahon**[1], **Yukinori Okada**[13,14,15], **Nicola Pirastu**[16,17], **N. William Rayner**[6], **Jeremy Schwartzentruber**[2,7], **Robert Vaughan**[18], **Shefali Verma**[19], **Steven P. Wilder**[20], **Fiona Cunningham**[1], **Lucia Hindorff**[21], **Ken Wiley**[21], **Helen Parkinson**[1], and **Inês Barroso**[22]

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** National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, 20892-6510, USA **5** Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA **6** Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany **7** Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. **8** Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK **9** British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK **10** Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK **11** National Institute for Health and Care Research (NIHR) Bristol Biomedical Research Centre, University of Bristol, Oakfield House, Bristol, BS8 2BN, UK **12** Medical Research Council (MRC) Integrative Epidemiology Unit (IEU), Bristol Medical School (Population Health Sciences), University of Bristol, Oakfield House, Bristol, BS8 2BN, UK **13** Osaka University Graduate School of Medicine, Suita, 565-0871, Japan **14** The University of Tokyo, Tokyo, Japan **15** RIKEN Center for Integrative Medical Sciences, Yokohama, 230-0045, Japan **16** Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, UK **17** Genomics Research Centre, Human Technopole, Milan, Italy **18** Congenica Ltd, Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, UK. **19** Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA **20** Genomics Plc, King Charles House, Oxford, Park End Street, OX1 1JD, UK **21** National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. **22** Exeter Centre of Excellence for Diabetes Research (EXCEED), University of Exeter Medical School, Exeter, UK

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 et al., 2017), trait pleiotropy (Smeland et al., 2017), prediction using polygenic scores (PGS) (Lambert et al., 2019) and Mendelian randomisation (MR) (Paternoster et al., 2017). However, still a considerable number of summary statistics are 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 et al., 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 et al., 2021) developed for robustness and performance and to underpin the OpenGWAS platform (Elsworth et al., 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.

Following initial discussions with the GWAS community at the 2020 workshop, which provided a defined set of recommendations as an outcome (MacArthur et al., 2021), the GWAS Catalog hosted a series of meetings between June 2021 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. During these meetings a phase of iteration on the proposed standard was completed and the final outcome is detailed here.

Requirements

The key requirements 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 new 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 formt

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, https://github.com/EBISPOT/gwas-summary-statistics-standard/examples/0000123.tsv 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.
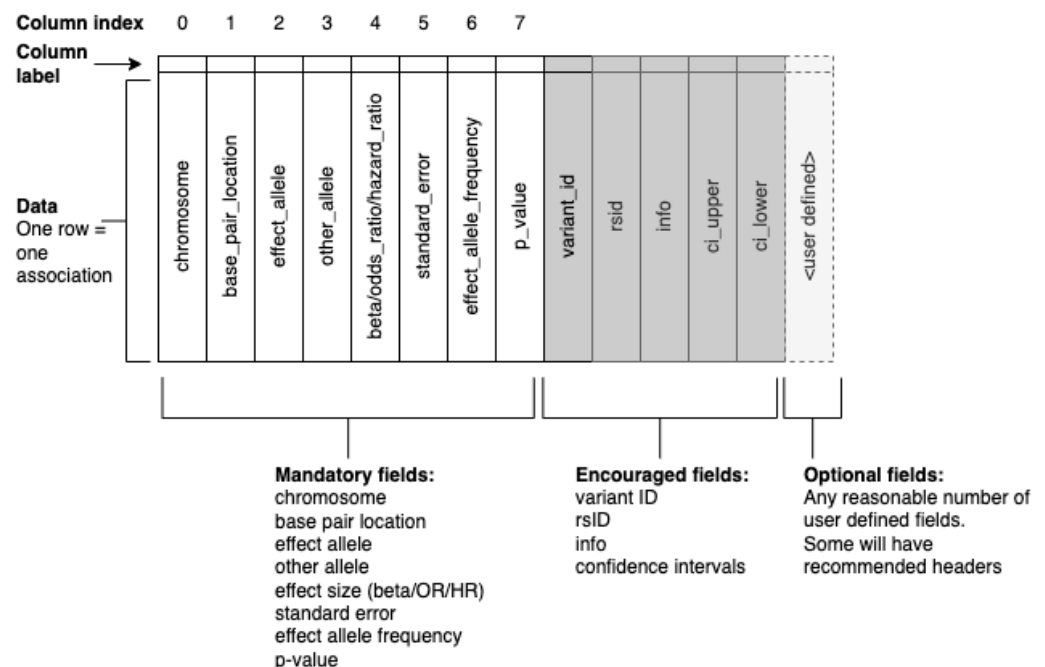

Figure 1: Schematic representation of the summary statistics table. Examples of data content within each specific field are provided in Table 1.

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.

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 beta	Numeric	Mandatory*
odds_ratio	Column 4: odds ratio	$x >= 0$	Mandatory*
hazard _ratio	Column 4: hazard ratio	$x >= 0$	Mandatory*
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	Internal variant identifier by concatenating chromosome, base_pair_locaion n, other_allele and effect_allele with underscores	[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	Harmonised datasets only

* Mandatory that either beta, odds_ratio or hazard_ratio is given
** 'LONG_STRING' can be used where allele string is too long to be represented.

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. $p < 10E\text{-}300$) 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.

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

agttcattttcg

agttcatCttcg

1   8
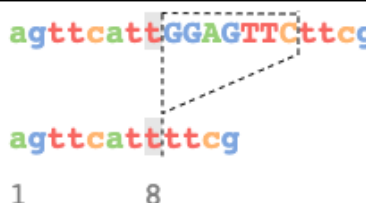
chromosome	base_pair_location	other_allele	effect_allele
11	8	T	C

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

agttcattttcg

agttcattGGAGTTCttcg

1   8

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)

agttcattGGAGTTCttcg

agttcattttcg

1   8

chromosome	base_pair_location	other_allele	effect_allele
11	8	TGGAGTTC	T

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.

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 https://github.com/EBISPOT/gwas-summary-statistics-standard/examples/0000123.yaml 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 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.

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	Yes if case-ControlStudy
controlCount	Number of controls for case/control study	Integer	Yes if case-ControlStudy
caseControlStudy	Flag whether the study is 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	Association analysis software and version	Text string	Yes if p-value of 0
imputationPanel	Imputation panel	Text string	No
imputationSoftware	Imputation software	Text string	No
effectAlleleFreqLowerLimit	Lowest possible effect allele frequency	Numeric	No
ancestryMethod	Method to determine sample ancestry e.g. self-reported or 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 or hazard ratio	yes
hmodeDefinition	Description of harmonisation codes	Text string	Harmonised datasets only

Field	Description	Accepted value	Mandatory
pvalueIsNegLog10	Flag whethe p value is 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

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 et al., 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.

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.

Acknowledgements

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.

For citations of references, we prefer the use of parenthesis, last name and year. If you use a citation manager, Elsevier – Harvard or American Psychological Association (APA) will work. If you are referencing web pages, software or so, please do so in the same way. Whenever possible, add authors and year. We have included a couple of citations along this document for you to get the idea. Please remember to always add DOI whenever available, if not possible, please provide alternative URLs. You will end up with an alphabetical order list by authors' last name.

GitHub repositories

- https://github.com/EBISPOT/gwas-summary-statistics-standard/

References

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., . . . Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, *29*(4), 365–371. https://doi.org/10.1038/ng1201-365

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., . . . Parkinson, H. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Elsworth, B., Lyon, M., Alexander, T., Liu, Y., Matthews, P., Hallett, J., Bates, P., Palmer, T., Haberland, V., Smith, G. D., Zheng, J., Haycock, P., Gaunt, T., & Hemani, G. (2020). *The MRC IEU OpenGWAS data infrastructure*. bioRxiv. https://doi.org/10.1101/2020.08.10.244293

Lambert, S. A., Abraham, G., & Inouye, M. (2019). Towards clinical utility of polygenic risk scores. *Human Molecular Genetics*, *28*(R2), R133—R142. https://doi.org/10.1093/hmg/ddz187

Lyon, M. S., Andrews, S. J., Elsworth, B., Gaunt, T. R., Hemani, G., & Marcora, E. (2021). The variant call format provides efficient and robust storage of GWAS summary statistics. *Genome Biology*, *22*(1), 32. https://doi.org/10.1186/s13059-020-02248-0

MacArthur, J. A. L., Buniello, A., Harris, L. W., Hayhurst, J., McMahon, A., Sollis, E., Cerezo, M., Hall, P., Lewis, E., Whetzel, P. L., Bahcall, O. G., Barroso, I., Carroll, R. J., Inouye, M., Manolio, T. A., Rich, S. S., Hindorff, L. A., Wiley, K., & Parkinson, H. (2021). Workshop proceedings: GWAS summary statistics standards and sharing. *Cell Genomics*, *1*(1), 100004. https://doi.org/https://doi.org/10.1016/j.xgen.2021.100004

Morales, J., Welter, D., Bowler, E. H., Cerezo, M., Harris, L. W., McMahon, A. C., Hall, P., Junkins, H. A., Milano, A., Hastings, E., Malangone, C., Buniello, A., Burdett, T., Flicek, P., Parkinson, H., Cunningham, F., Hindorff, L. A., & MacArthur, J. A. L. (2018). A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS catalog. *Genome Biology*, *19*(1), 21. https://doi.org/10.1186/s13059-018-1396-2

Murphy, A. E., Schilder, B. M., & Skene, N. G. (2021). MungeSumstats: A bioconductor package for the standardisation and quality control of many GWAS summary statistics. *Bioinformatics (Oxford, England)*, btab665. https://doi.org/10.1093/bioinformatics/btab665

Paternoster, L., Tilling, K., & Davey Smith, G. (2017). Genetic epidemiology and mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. *PLoS Genetics*, *13*(10), e1006944. https://doi.org/10.1371/journal.pgen.1006944

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., Bakker, P. I. W. de, Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559—575. https://doi.org/10.1086/519795

Smeland, O. B., Frei, O., Kauppi, K., Hill, W. D., Li, W., Wang, Y., Krull, F., Bettella, F., Eriksen, J. A., Witoelar, A., Davies, G., Fan, C. C., Thompson, W. K., Lam, M., Lencz, T., Chen, C.-H., Ueland, T., Jönsson, E. G., Djurovic, S., . . . NeuroCHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) Cognitive Working Group. (2017). Identification of genetic loci jointly influencing schizophrenia risk and the cognitive traits of verbal-numerical reasoning, reaction time, and general cognitive function. *JAMA Psychiatry*, *74*(10), 1065—1075. https://doi.org/10.1001/jamapsychiatry.2017.1986

Wang, Q., Dhindsa, R. S., Carss, K., Harper, A. R., Nag, A., Tachmazidou, I., Vitsios, D., Deevi, S. V. V., Mackay, A., Muthas, D., Hühn, M., Monkley, S., Olsson, H., AstraZeneca Genomics Initiative, Wasilewski, S., Smith, K. R., March, R., Platt, A., Haefliger, C., & Petrovski, S. (2021). Rare variant contribution to human disease in 281,104 UK biobank exomes. *Nature*, *597*(7877), 527—532. https://doi.org/10.1038/s41586-021-03855-y

Wheeler, E., Leong, A., Liu, C.-T., Hivert, M.-F., Strawbridge, R. J., Podmore, C., Li, M., Yao, J., Sim, X., Hong, J., Chu, A. Y., Zhang, W., Wang, X., Chen, P., Maruthur, N. M., Porneala, B. C., Sharp, S. J., Jia, Y., Kabagambe, E. K., . . . Meigs, J. B. (2017). 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 Medicine*, *14*(9), e1002383. https://doi.org/10.1371/journal.pmed.1002383

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18