Bioinformatics, 37(23), 2021, 4593–4596
doi: 10.1093/bioinformatics/btab665
Advance Access Publication Date: 2 October 2021
Applications Note

Databases and ontologies

MungeSumstats: a Bioconductor package for the standardization and quality control of many GWAS summary statistics

Alan E. Murphy [1,2]*, Brian M. Schilder [1,2] and Nathan G. Skene[1,2]*

[1]UK Dementia Research Institute at Imperial College London, London W12 0BZ, UK and [2]Department of Brain Sciences, Imperial College London, London W12 0BZ, UK

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

Received on June 22, 2021; revised on August 12, 2021; editorial decision on August 29, 2021; accepted on August 29, 2021

Abstract

Motivation: Genome-wide association studies (GWAS) summary statistics have popularized and accelerated genetic research. However, a lack of standardization of the file formats used has proven problematic when running secondary analysis tools or performing meta-analysis studies.

Results: To address this issue, we have developed MungeSumstats, a Bioconductor R package for the standardization and quality control of GWAS summary statistics. MungeSumstats can handle the most common summary statistic formats, including variant call format (VCF) producing a reformatted, standardized, tabular summary statistic file, VCF or R native data object.

Availability and implementation: MungeSumstats is available on Bioconductor (v 3.13) and can also be found on Github at: https://neurogenomics.github.io/MungeSumstats.

Contact: a.murphy@imperial.ac.uk or n.skene@imperial.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) summary statistics are used to distribute the most important outputs of GWASs in a manner which does not require the transfer of individual-level personally identifiable information from participants. Summary statistics from past studies tend to become more valuable over time as it becomes possible to meta-analyze and integrate them with new annotation information through approaches such as Linkage Disequilibrium Score Regression (LDSC) (Bulik-Sullivan *et al.*, 2015), Generalized Gene-Set Analysis of GWAS Data, MAGMA (de Leeuw *et al.*, 2015) and multi-phenotype investigations (Aguirre *et al.*, 2021; Tanigawa *et al.*, 2019). Summary statistics are also commonly integrated for use in the meta-analysis of GWAS. However, these tools and this integration require a standardized data format which was historically lacking from the field. The diversity of data formats in summary statistics has been a result of the phenotypes in question, for example disease-control or quantitative trait, the software used to perform the analysis, such as PLINK (Purcell *et al.*, 2007) and GCTA (Yang *et al.*, 2011) or just the preference of the consortium in question.

There have been movements to standardize the summary statistic file format such as the NHGRI-EBI GWAS Catalogue standardized format (Buniello *et al.*, 2019) and the SMR Tool binary format (Zhu *et al.*, 2016). More recently, the variant call format to store GWAS summary statistics (GWAS-VCF) (Lyon *et al.*, 2021) has been developed which has manually converted over 10 000 GWAS to this format. While GWAS-VCF offers a standardized format that future GWAS consortium may adopt, there are still a multitude of past, publicly available GWAS which have not been standardized (Jansen *et al.*, 2019; Lin *et al.*, 2018; Luciano *et al.*, 2021; McCormack *et al.*, 2018). For instance, although their summary statistics are publicly available, the GWAS for Cerebral small vessel disease (Sargurupremraj *et al.*, 2020) is not yet available in VCF format via IEU GWAS. Furthermore, as VCF is not yet the standard for sharing files between geneticists, unpublished GWAS shared internally within genetics consortia or provided by personal genetics companies are still found in a variety of summary statistic formats. As such, there is a need for tools to move between the various formats in which summary statistics are stored.

The standardization of GWAS summary statistics also requires quality control to ensure cohesive integration. For example, checking if the non-effect allele from the summary statistics matches the reference sequence from a reference genome to ensure consistent directionality of allelic effects across GWAS. In addition, downstream analysis tools often require a degree of quality control which, in the case of meta-analysis, must be applied across all GWAS. One such example is the removal of all non-biallelic SNPs is a common requirement of all downstream analysis (Lyon *et al.*, 2021).

To address these issues, we introduce MungeSumstats a Bioconductor R package for the rapid standardization and quality

control of many GWAS summary statistics. MungeSumstats can handle the most common summary statistic formats as well as GWAS-VCFs to enable the integrative meta-analysis of diverse GWAS. MungeSumstats also offers a comprehensive and tuneable quality control protocol with defaults for common, best-practice approaches. MungeSumstats capitalizes on R's familiar interface, is readily accessible through Bioconductor and utilizes an intuitive approach, running with a single line of input code.

2 Heterogeneity in GWAS formats

To demonstrate the diversity in summary statistics across GWAS, we analyzed a public repository of over 200 publicly available GWAS (Gloudemans, 2021). From this, the most common summary statistics were derived (see Fig. 1 for the 12 most common file header formats).

A total of 327 summary statistic files were derived from the analysis which corresponded to 127 unique formats. Thus, on average, every 2.5 summary statistic files had a unique format, showing the clear disparity across GWAS. The 12 most common formats, shown in Figure 1, accounted for approximately 47% all summary statistics. MungeSumstats has been tested on these 12 most common formats and is able to standardize their summary statistics.

3 Implementation

MungeSumstats was implemented using the R programming language (v 4.0) and Bioconductor S4 data infrastructure (v 3.13) enabling the full analysis of summary statistics within the R environment. The package removes the need for external software to perform the standardization and quality control steps.

MungeSumstats' implementation ensures both memory and speed efficiency through the use of R data.table (v.1.14.0) (Dowle and Srinivasan, 2021), which can take advantage of multi-core parallelization. Moreover, MungeSumstats benefits from Bioconductor's infrastructure for efficient representation of full genomes and their SNPs, using BSgenome (v 1.59.2) SNP reference genomes (Pagès, 2021). Either Ensembl's GRCh37 or GRCh38 are queried dependent on the build for the particular GWAS. Numerous of MungeSumstats' quality control steps for summary statistics require the use of a reference genome. For example, an allele flipping test is run (see Table 1) to ensure consistent directionality of allelic effect and frequency variables. The effect or alternative allele is always assumed to be the second allele (A2), in line with the approach for GWAS-VCF (Lyon *et al.*, 2021). Moreover, MungeSumstats can impute any missing, essential information like SNP ID, base-pair position and effect/non-effect allele.

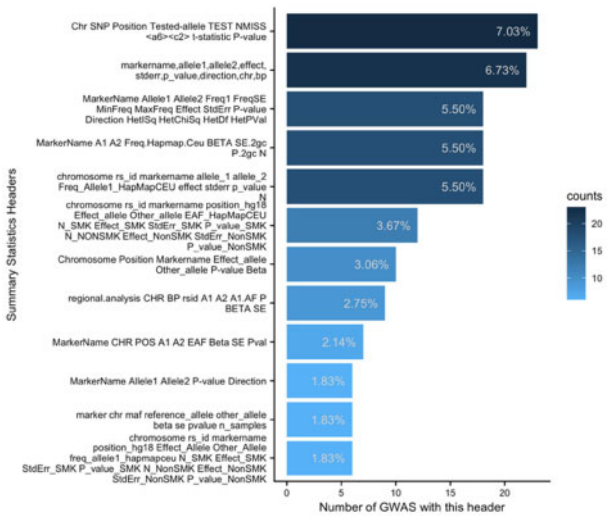

Fig. 1. Most common summary statistic formats show the most common summary statistic formats from a repository of over 200 publicly available GWAS (Gloudemans, 2021). Note that, a GWAS can have more than 1 summary statistics file and '<a6><c2>' is the symbol '¶¬' read into R

Using these two infrastructures, MungeSumstats conducts more than 30 checks on the inputted summary statistics file (see Table 1 for a description of their use). MungeSumstats is also written to ensure the ease of addition of further checks so if users have summary statistics which can't currently be handled in MungeSumstats, these can be incorporated easily in future releases. Finally, MungeSumstats returns a reformatted, tabular summary statistics file, a VCF or an R native data object (data.table, VRanges or GRanges) with standardized columns for the information necessary for downstream analysis.

The quality control and standardization checks conducted. Most checks are optional and can be set by the user. Here, CHR is chromosome, BP is Base-pair position, A1 is the non-effect allele, A2 is the effect allele, N is the sample size, INFO is imputation information score, FRQ is the minor allele frequency (MAF) of the SNP, SNP ID is the single nucleotide polymorphism reference ID, P is the unadjusted *P*-value, Z is z-score, OR is odds ratio, LOG_ODDS is the log odds ratio, BETA is the effect size estimate relative to the alternative allele and SIGNED_SUMSTAT is the directional effect size estimate for the summary statistics.

4 Usage

Once MungeSumstats is installed, usage involves a single line of code or one function call (*format_sumstats*) with the path to the summary statistics file of interest. Then, the path to the reformatted, standardized summary statistic file is returned. MungeSumstats also offers adjustable parameters to manage the quality control steps. These include options to adjust the imputation information score (INFO) cut-off threshold, the number of samples (*N*) outliers cut-off threshold and whether to remove mitochondrial SNPs or SNPs on the X or Y chromosome (see Table 1). Quality control steps which use a reference genome can also be adjusted such as whether to filter SNPs based on their RS ID's presence on the reference genome, whether to check for allele flipping and whether to remove multi-allelic or strand-ambiguous SNPs. These parameters ensure MungeSumstats can be adjusted to the user's analysis pipelines.

5 Conclusion

Here, we presented MungeSumstats, a Bioconductor package for the standardization and quality control of GWAS summary statistics. This package enables integration of summary statistics of vastly different formats, simplifying meta-analysis and summary statistics use in other secondary research applications. The package provides an efficient, user-friendly R-native approach, returning a standardized, tabular format file, VCF or R native data object. This ensures that the summary statistics are accessible to the average user. Moreover, MungeSumstats is written to permit future development of additional standardization steps if users encounter issues with their specific GWAS.

Acknowledgements

We thank Alexandru Voda (@alexandruioanvoda) for contributing via github in making an early version of this package work across operating systems. We also thank Mike Gloudeman collated repository of publicly available GWAS summary statistics at: https://github.com/mikegloudemans/gwas-download.

Funding

This work was supported by a UKDRI Future Leaders Fellowship [MR/T04327X/1] and the UK Dementia Research Institute which receives its

Table 1. MungeSumstats implemented checks

|S|	MungeSumstats check	Description
1	Check VCF format	If the input file is in variant call format (VCF), if so import
2	Check tab, space or comma delimited	If input is space or comma delimited convert to tab delimited. Can handle .tsv, .txt, .csv, .tsv.gz, .txt.gz, .csv.gz, .tsv.bgz, .txt.bgz, .csv.bgz, .vcf, .vcf.gz, .vcf.bgz files.
3	Check for header name synonyms	If any alternative names are found for SNP, BP, CHR, A1, A2, P, Z, OR, BETA, LOG_ODDS, SIGNED_SUMSTAT, N, N_CAS, N_CON, NSTUDY, INFO or FRQ convert them to a standard name. Robust conversion approach with 176 unique mappings
4	Check for multiple models or traits in GWAS	If multiple, user must specify one to analyze
5	Check for uniformity in SNP ID	Ensure no mix of RS ID, missing 'rs' prefix and/or CHR: BP
6	Check for CHR: BP: A2: A1 all in one column	Split into separate columns if found
7	Check for CHR: BP in one column	Split into separate columns if found
8	Check for A1/A2 in one column	Split into separate columns if found
9	Check if CHR and/or BP is missing	If so, infer from the chosen reference genome
10	Check if SNP ID is missing	If so, infer from the chosen reference genome
11	Check if A1 and/or A2 are missing	If so, infer from the chosen reference genome
12	Check that vital columns are present	Check for the necessary columns; SNP, CHR, BP, P, A1, A2
13	Check for one signed/effect column	Effect columns Z, OR, BETA, LOG_ODDS, SIGNED_SUMSTAT
14	Check for missing data	If data is missing from any entry, remove the SNP
15	Check for duplicated columns	If there are any remove one
16	Check for *P*-values lower than 5e-324	These are not recognized in R and cause issues with downstream analysis software like LDSC/MAGMA. User can convert to 0.
17	Check N column	Ensure it is an integer and check if the sample size for a SNP isn't greater than mean multiplied by five times the standard deviation. Removes SNPs that have substantial more samples than the rest.
18	Check SNPs are RS ID's	Checks validity of SNP IDs as RS IDs, other IDs can still be used
19	Check for duplicated rows, based on SNP ID	Duplicates are removed
21	Check for duplicated rows, based on base-pair position	Duplicates are removed
22	Check for SNPs on reference genome	Correct any missing from reference genome using BP and CHR
23	Check INFO score	Remove SNPs with imputation score less than 0.9
24	Check for strand-ambiguous SNPs	Remove strand-ambiguous SNPs if found
25	Check for non-biallelic SNPs (infer from reference genome)	Infer from chosen reference genome and remove any if found
26	Check for allele flipping	The effect/alternative/minor allele is assumed to be A2. The allele flipping function checks A1 against a reference genome. For a given SNP, if A1 doesn't match the reference genome sequence (i.e. it is the alternative allele, not the reference allele for example), A1 and A2 along with the effect and frequency columns are flipped, creating consistent directionality of allelic effects across GWAS.
27	Check for SNPs on chromosome X, Y and mitochondrial SNPs (MT)	If any are found these are removed.
28	Check output format is LDSC ready	Standardized file can be passed to LDSC without pre-processing
29	Check effect column values	Ensure effect columns (like BETA) aren't equal to 0
30	Check Standard Error	Ensure standard error (SE) is positive
31	Check dropped and imputed values	Return indicators of the imputed values for a SNP and return the SNPs and the reason for exclusion because of QC.

funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.

Conflict of Interest: none declared.

References

Aguirre,M. *et al.* (2021) Polygenic risk modeling with latent trait-related genetic components. *Eur. J. Hum. Genet.*, **29**, 1071–1081.

Bulik-Sullivan,B. *et al.*; Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium. (2015) An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, **47**, 1236–1241.

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

de Leeuw,C.A. *et al.* (2015) MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.*, **11**, e1004219.

Dowle,M. and Srinivasan,A. (2021) data.table: Extension of 'data.frame'. https://CRAN.R-project.org/package=data.table (20 April 2021, date last accessed).

Gloudemans,M. (2021) mikegloudemans/gwas-download. https://github.com/mikegloudemans/gwas-download (20 April 2021, date last accessed).

Jansen,P.R. *et al.*; 23andMe Research Team. (2019) Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat. Genet.*, **51**, 394–403.

Lin,H. *et al.* (2018) Common and rare coding genetic variation underlying the electrocardiographic PR interval. *Circ. Genomic Precis Med.*, **11**, e002037.

Luciano,M. *et al.* (2021) The influence of X chromosome variants on trait neuroticism. *Mol. Psychiatry*, **26**, 483–491.

Lyon,M.S. *et al.* (2021) The variant call format provides efficient and robust storage of GWAS summary statistics. *Genome Biol.*, **22**, 32.

McCormack,M. *et al.*; for the EPIGEN Consortium. (2018) Genetic variation in CFH predicts phenytoin-induced maculopapular exanthema in European-descent patients. *Neurology*, **90**, e332–e341.

Pagès,H. (2021) BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. https://bioconductor.org/packages/BSgenome (20 April 2021, date last accessed).

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Sargurupremraj,M. *et al.*; International Network against Thrombosis (INVENT) Consortium. (2020) Cerebral small vessel disease genomics and its implications across the lifespan. *Nat. Commun.*, **11**, 6285.

Tanigawa,Y. *et al.* (2019) Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. *Nat. Commun.*, **10**, 4064.

Yang,J. *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis. *Am J. Hum. Genet.*, **88**, 76–82.

Zhu,Z. *et al.* (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, **48**, 481–487.