# Standardised and efficient querying of GWAS summary statistics by adaptaing the VCF format

Matthew Lyon1,2\*, Shea J Andrews3\*, Ben Elsworth2, Tom Gaunt1,2, Gibran Hemani2¥, Edoardo Marcora3¥

\* Joint first author

¥ Joint last author

Correspondence to Matthew Lyon (matt.lyon@bristol.ac.uk)

1. National Institute for Health Research Bristol Biomedical Research Centre
2. Medical Research Council (MRC) Integrative Epidemiology Unit (IEU), Bristol Medical School (Population Health Sciences), University of Bristol, Bristol, UK
3. Ronald M. Loeb Center for Alzheimer’s disease, Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA

## Abstract

TO DO

## Introduction

The genome-wide association study (GWAS) is a powerful tool for identifying loci associated with traits, diseases and molecular phenotypes such as gene expression and biomarker concentration[1]. Sharing non-identifiable test summary statistics (i.e. variant, effect size, standard error, p-value etc) has enabled a range of important secondary research applications including gene prioritization[2], causal inference (Mendelian randomization; MR)[3], risk prediction[1], genetic correlation[4] and heritability estimation[5]. However, existing tools have a wide variety of input formats making analyses cumbersome and error prone, and making it difficult to automate analyses across multiple traits. Additionally, data requirements for downstream analysis vary widely, for example gene prioritization can be undertaken with loci and P value alone while MR additionally demands allele and standard error which is often unavailable. Storing appropriate meta-data with the summary data itself is often difficult also, which can lead to errors if it is not clear what the genome build is, which allele represents the effect allele, or what the trait is or the units of effect sizes are.

Every GWAS analysis tool outputs results in a different way (e.g. plink, GCTA, Bolt-LMM, GEMMA etc). GWAS meta-analysis tools also output varying results formats. Typical problems with these formats are that it is not obvious which allele relates to the effect size estimate, they are slow to query, and they each store different types of information with different field names. Some proposals have been made for standard formats. The EBI-NHGRI GWAS catalog proposed a text format with standardized column names. The SMR tool proposed a way to store GWAS summary data in a binary format for rapid query of eQTL summary data. Learning from these examples we have identified a set of requirements for a suitable universal format:

1. Human readable and easy to parse
2. Unambiguous interpretation of the data
3. Unambiguous representation of bi-allelic, multi-allelic and INDEL variants
4. Genomic information can be validated
5. Flexibility on which GWAS fields are recorded
6. Stores meta-data about the study or studies
7. Allows multiple studies to be stored together
8. Rapid querying by rsid, position ranges or gwas summary data values
9. Compressed
10. Readable by existing tools
11. Amenable to cloud based streaming and database storage

We determined that adapting the widely adopted variant call format (VCF) [8], [9] was a convenient and constructive approach to meet these requirements. Here we describe the implementation, explain how it meets these requirements, describe existing and new software that creates and connects to the data format, and show results of querying times. Over 10000 complete GWAS summary datasets have been converted to this format as part of the IEU GWAS database, and are available for download here: <https://gwas.mrcieu.ac.uk>.

## Implementation

The VCF is a specialized text format for robust storage of genetic variants and metadata which is sorted karyotypically and indexed to enable rapid queries by genomic position. The file body contains one locus per row with fields for user defined content where we propose to store the association effect size, standard error and P value. Required fields and variable type are set in the file header and enforced with each read/write to guarantee data integrity. Using this approach alternative allele(s) are always ‘effect allele(s)’ allowing consistency between studies for ease of comparison. Multiple traits can be stored in a single file allowing distribution of related phenotypes or individually as desired. The full specification provides detailed information including reserved keys (link below).

## How this format meets the specification

* Explain what can be done in terms of storing meta data, and specifying what the fields of data are
* Multiple studies can be stored
* Tools exist to validate the data
* Libraries have been designed for rapid indexing
  + Mention that later in the paper we demonstrate querying speed

## Open source tools that use the format

* Creating the format, and interchanging between the EBI and VCF format
  + Python library that is dockerised
  + Online web app
* Querying or reading completely into R or python
* Connects to multiple other tools through the gwasglue package
* Use it natively with LD score regression
* A table would be good to summarise this – a list of tools, what they do, links

We have developed open-source libraries for mapping to (gwas2vcf) and reading from VCF (R and python packages: gwasvcf & pygwasvcf) according to our specification. These tools implement the HTSLIB library[10] which provides routines to handle complex multi-allelic and insertion-deletion variants that are frequently discarded in current analyses. File validity can be assessed using GATK[11] ValidateVariants function.

We encourage users to provide feedback via the issue pages.

## Query performance

To assess query performance for extracting a single variant using chromosome position from either unprocessed tab-separated text or VCF (Figure 1), we obtained GWAS summary statistics from a large metanalysis of body mass index containing 2,336,269 variants [12]. Through each of the 100 repetitions we randomly chose a variant to query and then measured retrieval time using UNIX awk or bcftools[13] (v1.10).

The mean query time using gzip compressed VCF was 0.07 seconds (95% CI 0.07, 0.07) compared with 0.91 seconds for compressed text using gzip and awk (95% CI 0.91, 0.91) or 0.75 seconds for uncompressed text using awk directly (95% CI 0.74, 0.75). These tests suggest query time with VCF is around ten times quicker than using awk.

Can you also compare querying by

* Rsid
* Chrom position ranges
* P-value

## Conclusion

The VCF format is a robust solution to storing and distributing GWAS summary statistics. We find the query performance to be around ten times faster than reading the whole dataset with standard UNIX tools. To facilitate adoption, we have developed tools for mapping data from tabular data to VCF and reading using R or Python.

**Specification**

Available from: <https://github.com/MRCIEU/gwas_vcf_spec>

**Code availability**

Open-source Python3 script for converting GWAS summary statistics to VCF available from <https://github.com/MRCIEU/gwas2vcf>

Open-source Python3 and R libraries for reading GWAS in VCF format available from <https://github.com/MRCIEU/pygwasvcf> and <https://github.com/MRCIEU/gwasvcf>

**Acknowledgments**

This study was funded by the NIHR Biomedical Research Centre at University Hospitals Bristol National Health Service Foundation Trust and the University of Bristol. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

M.L., B.E., T.R.G. work in the Medical Research Council Integrative Epidemiology Unit at the University of Bristol, which is supported by the Medical Research Council and the University of Bristol (MC\_UU\_00011/4). G.H. is supported by the Wellcome Trust and Royal Society [208806/Z/17/Z].

**References**

[1] P. M. Visscher *et al.*, “10 Years of GWAS Discovery: Biology, Function, and Translation,” *American Journal of Human Genetics*, vol. 101, no. 1. Cell Press, pp. 5–22, 06-Jul-2017.

[2] L. Hou and H. Zhao, “A review of post-GWAS prioritization approaches,” *Front. Genet.*, vol. 4, no. DEC, p. 280, Dec. 2013.

[3] G. D. Smith and S. Ebrahim, “‘Mendelian randomization’: Can genetic epidemiology contribute to understanding environmental determinants of disease?,” *International Journal of Epidemiology*. 2003.

[4] B. Bulik-Sullivan *et al.*, “LD score regression distinguishes confounding from polygenicity in genome-wide association studies,” *Nat. Genet.*, 2015.

[5] J. Yang, J. Zeng, M. E. Goddard, N. R. Wray, and P. M. Visscher, “Concepts, estimation and interpretation of SNP-based heritability,” *Nature Genetics*, vol. 49, no. 9. Nature Publishing Group, pp. 1304–1310, 01-Sep-2017.

[6] A. Buniello *et al.*, “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1005–D1012, Jan. 2019.

[7] S. Purcell *et al.*, “PLINK: A tool set for whole-genome association and population-based linkage analyses,” *Am. J. Hum. Genet.*, 2007.

[8] “The Variant Call Format (VCF) Version 4.2 Specification,” 2019.

[9] P. Danecek *et al.*, “The variant call format and VCFtools,” *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, Aug. 2011.

[10] “Samtools.” [Online]. Available: http://www.htslib.org/. [Accessed: 20-Feb-2020].

[11] A. McKenna *et al.*, “The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data,” *Genome Res.*, 2010.

[12] L. Yengo *et al.*, “Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry,” *Hum. Mol. Genet.*, vol. 27, no. 20, pp. 3641–3649, 2018.

[13] H. Li, “A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.,” *Bioinformatics*, vol. 27, no. 21, pp. 2987–93, Nov. 2011.

Figure 1. Performance comparison for querying GWAS summary statistics in plain text and VCF format using chromosome position

A screenshot of a cell phone

Description automatically generated

Query time for extracting a single variant using chromosome position with tab-separated text or VCF format. GWAS summary statistics were obtained from a large metanalysis of body mass index containing 2,336,269 variants [12]. The simulation was performed on Ubuntu v18.04 running an Intel Xeon(R) 2.0 Ghz processor with 100 repetitions. During each repetition a variant was chosen at random and queried using awk or bcftools[13] v1.10. There were robust differences (paired t-test VCF vs compressed text mean query time [sec] P < 2.2 x 10-16) in mean query time between compressed VCF (0.07s [95% CI 0.07, 0.07]) and compressed (0.91s [95% CI 0.91, 0.91]) or uncompressed text (0.75s [95% CI [0.74, 0.75]).