Wootton, R. E., & Sallis, H. M. (2020). Let's call it the effect allele: A suggestion for GWAS naming convention. *International Journal of Epidemiology*, [dyaa149]. https://doi.org/10.1093/ije/dyaa149

Peer reviewed version

Let's call it the effect allele: A suggestion for GWAS naming conventions

To the Editor – In recent years, the amount of publicly available summary data from genome-wide association studies (GWAS) has rapidly increased. So too, has the number of researchers accessing and utilising this data. These summary data can be used in many subsequent analyses, including Mendelian randomization, genetic correlation, and polygenic score analysis. It is therefore vital that we can ensure consistency across these datasets to minimise the risk of analytical mistakes due to user error. One such inconsistency is the naming of the effect allele in these datasets. This is of particular concern given the increasing availability of automated software packages for downstream analyses (e.g. MR-Base[1], LDpred[2], LD Hub[3]) which make complex analyses easy to run, even with little understanding of the underlying processes. Data harmonization is a crucial step for many of these analyses. Guidelines for harmonizing datasets to use in two-sample Mendelian randomization have been suggested by Hartwig and colleagues[4]. In order to effectively implement these steps, it is critical to know which is the effect allele. This starting point is assumed to be known in the Hartwig et al. paper, however, in many datasets this is unclear and could lead to substantial errors. This issue could be easily avoided if consistent naming conventions were used.

We have seen them all: A1/A2, A0/A1, effect allele/non-effect allele, effect allele/other allele, reference (REF) allele/alternative (ALT) allele, the list goes on. This difference in terminology can be due to the software used to perform the GWAS (e.g., plink[5]: ALT (effect)/REF (other); SNPTEST[6]: A(other)/B (effect)) or simply due to the naming convention used by the analyst when they create the datasets. If these terms were used consistently it would only be a minor annoyance to learn the different naming conventions. However, in GWAS summary datasets the terms A1/A0/alternative allele do not consistently refer to the effect allele (e.g., Liu et al.[7]: ALT (effect)/ REF (other); Snieckers et al.[8]: REF(effect)/ALT (other)). Sometimes they refer to the non-effect allele, or even to the minor allele (the variant that is less frequently found in the population). The effect allele is the allele to which the effect estimate refers, regardless of whether this estimate is increasing or decreasing and regardless of whether this allele is coding or non-coding. Sometimes, the minor allele is used as the effect allele but the two are not synonymous. The major allele (most frequently found in the population) can also be chosen as the effect allele. Therefore, knowing the minor allele alone is not sufficient for the majority of downstream analyses.

In order to work out which is the effect allele, researchers are reliant on clear documentation from the authors of the GWAS - this is often not available or is lacking in detail. When it does exist, it can be hard to find and is often buried deep in a supplementary table. We have experienced much confusion and many misunderstandings that could be easily avoided if GWAS authors were to use the terms 'effect allele' and 'other allele' instead. Despite the publication of reporting guidelines for genetic association studies (STrengthening the REporting of Genetic Association Studies (STREGA)[9]), there is currently no naming convention for reporting the effect allele. We would encourage this as an addition to point 16 (Main Results) of the current STREGA guidelines.

We are not trying to say that introducing this naming convention would be a panacea, and there are many other pieces of information and metadata that would help with the harmonisation process, as discussed elsewhere[10]. For example, information on effect allele frequency or strand alignment and genomic build that would help with the harmonization of ambiguous variants (e.g., palindromic variants with intermediate allele frequency). However, we believe that introducing a consistent allele naming convention is a simple suggestion that would be straightforward to implement. This simple step could go a long way to making data harmonisation more efficient and less error prone.

References

1. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
2. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
3. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
4. Hartwig, F. P., Davies, N. M., Hemani, G. & Davey Smith, G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int. J. Epidemiol.* **45**, 1717–1726 (2016).
5. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
6. Marchini, J. & Band, G. *SNPTEST v2 Technical Details*. (2010).
7. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
8. Sniekers, S. *et al.* Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
9. Little, J. *et al.* STrengthening the REporting of Genetic Association Studies (STREGA)—an extension of the STROBE statement. *Genet. Epidemiol. Off. Publ. Int. Genet. Epidemiol. Soc.* **33**, 581–598 (2009).
10. Lyon, M. S. *et al.* The variant call format provides efficient and robust storage of GWAS summary statistics. *bioRxiv* 2020.05.29.115824 (2020) doi:10.1101/2020.05.29.115824.