Guidelines for genome-wide association (GWAS) studies and studies involving GWAS data

There is a widely accepted need to improve the robustness of published genetic association findings—a main-stay of which is now driven by genome-wide association studies (GWAS). Where possible and depending on the objective of studies using genetic association study data, we also need to provide the readership of the journal with information that allows a more complete assessment of the biological relevance of the findings reported both in GWAS and in studies using GWAS data. Submissions to Diabetologia should, therefore, pay careful attention to the following fundamental issues. It is not intended that these represent absolute criteria for publication in Diabetologia (we don't want to block otherwise interesting studies that fail to meet one or other of these). However, guidelines set out the main factors that we expect our reviewers and Associate Editors to use in evaluating the quality of the manuscripts we receive.

The points raised below lie in parallel to the development of explicit guidelines concerning the deployment of genetic association study results in applied epidemiological study designs, for example Mendelian randomisation (https://www.strobe-mr.org).

These guidelines have been updated (August 2022) in light of a seminal guide review by Emil Uffelman and colleagues, published in Nature Reviews Methods (Primers) in August 2021 [1]. The extracts below have been adapted from Uffelman et al with permission. We strongly urge that authors, reviewers and handling editors be familiar with this methods text in the process of considering and managing GWAS papers and those involving GWAS data.

Overview points and critical observations from Uffelmann et al [1]

GWAS have come to replace some forms of genetic association studies, in particular candidate gene/variant studies since the early 2000s. They represent a deviation from the linkage approach to tracking segregating recombination events and an extension to the analysis of a limited number of genetic variants directed in a targeted fashion. GWAS look to identify associations of genotype with phenotype by testing for differences in frequency of genetic variant forms between individuals in a given sample and by observed variation in health outcomes or measured intermediaries. GWAS can consider copy-number variants or sequence variations in the human genome, although the most commonly studied genetic variants in GWAS are single-nucleotide polymorphisms (SNPs). GWAS typically report blocks of correlated SNPs that all show a statistically significant association with the trait of interest. The following key concepts should be considered.

Selecting study populations GWAS often require very large sample sizes to identify reproducible associations and—depending on the genetic architecture of the traits in question [2]—suitable sample sizes are needed to resolve what can be relatively small genetic associations. Studies should include sufficient samples to have power to detect effect sizes that are reasonable given current understanding of the genetic architecture of complex traits. Power calculations should be included that make explicit the effect sizes that the study was powered to detect: such power calculations should guide the interpretation of the data. Wherever possible, all available samples should be genotyped—results based on only a portion of a larger sample are of limited interest. Increasingly, there is a push to include more non-European samples in GWAS analyses and it may be important to consider the value of retaining non-Europeans and designing GWAS and their analyses to incorporate different populations.

Genotyping Genotyping of individuals is typically done using microarrays for common variants or next- generation sequencing methods such as whole-exome sequencing (WES) or whole-genome sequencing (WGS) that also include rare variants. Microarray-based genotyping is the most commonly used method for obtaining genotypes for GWAS owing to the current cost of next-generation sequencing. However, the choice of genotyping platform depends on many factors and tends to be guided by the purpose of the GWAS (partly adapted from [1]).

Data processing Input files for a GWAS include anonymised individual ID numbers, coded family relations between individuals, sex, phenotype information, covariates, genotype calls for all called variants and information on the genotyping batch. Following input of the data, generating reliable results from GWAS requires careful quality control. Once sample and variant quality control have been performed on GWAS array data, variants can undergo phasing in order to enable imputation to a common variant set using a sequenced haplotype reference panel (partly adapted from [1]).

Phenotypes and sampling frame Authors should explicitly justify why the samples genotyped are well suited to addressing the particular hypothesis posed—this includes awareness of the possible relationship between sampling frame and true biological or artefactual variation. Care also needs to be taken in the definition of cases using standardised criteria, in the selection of appropriate control samples and in the measurement of intermediate health outcomes. Although current methods can address unaccounted-for population stratification (with moves towards linear mixed models and the use of matrices of relatedness being common approaches), it can still cause spurious or biased associations, particularly in the meta-analyses of multiple cohorts which is increasingly—and appropriately—being promoted [3]. Whilst early GWAS of relatively small size and low power were (to some extent) shielded from the possible impact of population stratification, newer, large (up to and over 1 million participants) studies are able to detect heritable contributions to apparently environmental traits and outcomes (e.g. complex behaviours). This has led to a need for careful reappraisal of structure in genetic data (partly adapted from [1]; for further information see also [4, 5]).

Polygenicity Extreme polygenicity can pose a challenge when attempting to uncover underlying biological mechanisms, particularly in cases where thousands of variants each have a small effect on a trait. WES and WGS studies are increasingly being used to discover rare variants of large effect for which causal mechanisms are generally easier to elucidate, however GWAS are well designed to detect polygenic signals. Rare variants of large effect have yet to be reported for all traits and looking for convergence of the effects of thousands of variants remains the best strategy for traits not linked to rare variants of large effect (partly adapted from [1]).

Testing for associations Typically in GWAS, linear or logistic regression models are used to test for associations, depending on whether phenotypic outcomes of concern are continuous (such as height, blood pressure or body mass index) or binary (such as the presence or absence of disease). Full disclosure of methods used to undertake association testing (relevant to the outcomes of interest) are necessary for any GWAS report (partly adapted from [1]).

Accounting for false discovery Testing millions of associations between individual genetic variants and a phenotype of interest requires a stringent multiple-testing threshold to avoid false positives. As a consequence, manuscripts should feature explicit discussion of the consequences of multiple hypothesis-testing for the interpretation of the findings. The International HapMap Project and other studies have shown that there are approximately 1 million independent common genetic variants across the human genome on average, resulting in a Bonferroni testing threshold of $p < 5 \times 10^{-8}$ (representing a false discovery rate of $0.05/10^{6}$). Despite this, assessments of the significance of the findings should be related to ‘study-wide significance’—a form of penalised genome-wide significance that accounts for the nature of the population sample and linkage disequilibrium (LD) present for the genetic data available and the nature of the measures and tests collected and performed (partly adapted from [1]).

Replication and meta-analysis Replication is highly desirable for all association studies, particularly for studies where extensive multiple testing means that study-wide significance is not clear. However, replication should only be claimed when it addresses the same variant, phenotype and genetic model (all too often other phenotypes or variants within a gene are offered as evidence of replication). However, to increase sample size, GWAS is typically carried out in the context of a consortium such as Genetic Investigation of Anthropometric Traits (GIANT) consortium, where data from multiple cohorts are analysed together using standardised tools (partly adapted from [1]).

Positive/negative studies Well performed association studies that represent important null findings are welcome provided the gene examined has clear relevance to disease pathogenesis (or has been implicated on the basis of prior association data).

Statistical fine-mapping Many non-causal variants are associated with a trait of interest owing to LD and whether these reach the significance threshold depends on their level of correlation with an assumed the causal variant. The output of GWAS therefore represents clusters of ‘risk loci’—sets of correlated variants that all show evidence for association with the trait of interest—and LD typically prevents pinpointing causal variants without further analysis. This has immediate implications for the direct interpretation of observed association signals (partly adapted from [1]).

Functional data Following on from the desire to map association signals and a major motivation for GWAS is to use the identified associations to determine the biological cause of heritable phenotypes and provide a starting point for investigating causal biology. Although GWAS have led to the identification of thousands of complex trait-associated genetic variants, the biological implications of these variants are typically not easily inferred. Functional data (e.g. demonstration that a SNP alters expression) can strengthen association findings but the functional assays must have demonstrable relevance to the phenotype showing the association. Good functional data do not compensate for a poor association study. The same approach can be applied to other molecular phenotypes such as splicing, chromatin accessibility or methylation status. By integrating this information with GWAS results, trait-associated variants can be mapped to the genes they are likely to regulate in specific tissues and the molecular processes mediating these associations (partly adapted from [1]).

Risk prediction So called polygenic risk scores (PRS) can be used not only to generate summaries of the heritable contributions to outcomes and traits captured by GWAS arrays and imputed variants, but also for prediction. In this case, these aggregate scores of alleles weighted by their genetic associations with the phenotypes of interest are used to predict the risk of disease in a target cohort. This process usually uses GWAS summary statistics taken from an independent discovery cohort and predicts in a target collection. Theoretically, PRS can be used to identify individuals at a high risk of disease for clinical interventions and provide additional information over traditional clinical risk scores. They can be constructed from collections of evidence/signal-driven variants or from optimised collections of all variants recorded and have numerous applications with respect to applied genetic epidemiology, prediction and the analysis of ‘chip-based’ heritability.

Sharing results and the use of publicly available data from genome-wide association studies In 2007, two of the type 2 diabetes GWAS (the DGI and WTCCC) made their case–control data available to bone fide researchers. Furthermore, results on ~2.2M directly genotyped and imputed SNPs from a meta-analysis of three high-density GWAS (FUSION, DGI and WTCCC: effective sample size ~9500), which incorporates data from a fourth study (deCODE), were also made publicly available (diagram-consortium.org). One study (the DGI) also made public their data from multiple diabetes-related quantitative traits. This was largely seen as a positive step in the field, although it has not been without cross-examination. In August of 2008, a report by Homer et al [6] showed that, if a knowledgeable investigator had access to a specific DNA sample, s/he could determine with a high degree of certainty whether the owner of that DNA sample had participated in a given GWAS and infer case–control assignment (based on anonymised genotype counts across ~25,000 SNPs). This realisation motivated the withdrawal of similar datasets from public websites whilst appropriate privacy protection measures are implemented.

Whilst subsequent studies now need to consider their findings in light of these findings and any similar results that become available in the future, there is a growing culture of sharing results from well-defined and undertaken analyses which yield information of real value to the research community. We encourage authors of genetic association studies submitted to Diabetologia (whether on candidate genes, regions of previous linkage or the whole genome) to share and to use shared results to allow meta-analyses providing a more definitive answer on whether or not the variation under consideration alters diabetes risk or related quantitative traits. This can be done in datasets downloaded in relation to a specific publication, by requesting the relevant data from GWAS investigators, or by formal collaboration; ideally, attempts to pursue this kind of analysis should be documented in submitted manuscripts.

An example resource illustrating shared summary statistics can be found in the NHGRI-EBI GWAS Catalog.

Ethical challenges GWAS do raise ethical issues relating to consent for future use of samples and data, storage and reuse of samples and data, privacy challenges and sharing data with individual participants. Where these are an issue, there should be a clear report of any special circumstances. However more generally, a full report of the ethical foundations of the project in question (or collections involved in a joint project) should be given in all GWAS reports (partly adapted from [1]).

Recommended technical requirements

The information provided by a manuscript can be improved if certain technical requirements are observed. Many of these are outlined in Uffelman et al [1] and summarised above. They can be seen undertaken in the exemplars at the bottom [7–12].

In addition to this, a series of points to attend to also include:

- Provide a full description of the sampling frame employed in the analysis undertaken—consider STROBE or STREGA guidelines with respect to the description of study numbers and consider: (1) population specificity and epidemiological study design (random sample, case–control, prospective study, etc.), (2) whether the analysis is being undertaken in a family or individual 'non-related' sample base, or (3) whether there are special properties of the sample in question (for example the inclusion of isolated populations with known differences in demographic history).

- Provide rs numbers for all variants reported (these are quite easy to obtain for novel variants). Where these are provided, details of the assay (primer sequences, PCR conditions) can be kept brief. Also provide a build reference (for variant positions).

- Provide explicit details of the measures taken to ensure genotyping/imputation precision/accuracy (including, for example, % successful genotype calls, number of duplicated genotypes, % correspondence, metrics of imputation performance).

- Provide approved HUGO/NCBI gene names in the appropriate case and italics.

- Use standard terminology for variants (see http://varnomen.hgvs.org).

- Describe LD relationships between typed variants.

- Provide information on departures from Hardy–Weinberg equilibrium (HWE), not only as a check for possible genotyping errors, but also because methods assuming HWE may be employed in the downstream association analyses (e.g. haplotype inference using the expectation–maximisation [EM] algorithm/single-point analyses testing the multiplicative model).

- Provide raw genotype frequencies (that is, allele frequencies alone are not sufficient).

- Where appropriate, provide the criteria used to select tagSNPs. In addition, also carry out association analyses consistent with the tagging method employed, e.g. if an aggressive multi-marker tagging approach has been followed, appropriate analyses are required to retrieve all the information captured.

- Denote the boundaries considered when studying a gene of interest (e.g. 5kb upstream of transcription initiation etc.); and indicate which portions of the gene have been examined (e.g. exons and exon/intron boundaries).

- Make code used to derive datasets and run GWAS openly accessible, for example in GitHub.

- Be explicit about any and all software used, including the version used in running the analyses, and if appropriate, the date of download.

Select GWAS published in Diabetologia:

Downie et al (2022) Multi-ethnic GWAS and fine-mapping of glycaemic traits identify novel loci in the PAGE Study [7]

Chen et al (2019) Genome-wide association study of type 2 diabetes in Africa [8]

Sandholm et al (2014) Genome-wide association study of urinary albumin excretion rate in patients with type 1 diabetes [9]

Rich et al (2009) A genome-wide association scan for acute insulin response to glucose in Hispanic-Americans: the Insulin Resistance Atherosclerosis Family Study (IRAS FS) [10]

Select Diabetologia papers using GWAS data:

Liu et al (2022) A genome-wide cross-trait analysis identifies shared loci and causal relationships of type 2 diabetes and glycaemic traits with polycystic ovary syndrome [11]

Zheng et al (2022) Evaluating the efficacy and mechanism of metformin targets on reducing Alzheimer's disease risk in the general population: a Mendelian randomisation study [12]

References

1. Uffelmann E, Huang QQ, Munung NS et al (2021) Genome-wide association studies. Nat Rev Methods Primers 1: 59. https://doi.org/10.1038/s43586-021-00056-9
2. Timpson NJ, Greenwood CMT, Soranza N, Lawson DJ, Richards JB (2018) Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat Rev Genet 19(2): 110–124. DOI: 10.1038/nrg.2017.101
3. Barroso I (2021) The importance of increasing population diversity in genetic studies of type 2 diabetes and related glycaemic traits. Diabetologia 64: 2653–2664. https://doi.org/10.1007/s00125-021-05575-4
4. Haworth S, Mitchell R and Corbin L (2019) Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nat Commun 10: 333. https://doi.org/10.1038/s41467-018-08219-1
5. Tan VY, Timpson NJ (2022) The UK Biobank: a shining example of genome-wide association study science with the power to detect the murky complications of real-world epidemiology. Annu Rev Genom Hum Genet 23:569–589. https://doi.org/10.1146/annurev-genom-121321-093606
6. Homer N, Szelinger S, Redman et al (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PloS Genetics 4:e1000167. https://doi.org/10.1371/journal.pgen.1000167
7. Downie CG, Dimos SF, Bien SA et al (2022) Multi-ethnic GWAS and fine-mapping of glycaemic traits identify novel loci in the PAGE Study. Diabetologia 65:477–489. https://doi.org/10.1007/s00125-021-05635-9
8. Chen J, Sun M, Adeyemo A et al (2019) Genome-wide association study of type 2 diabetes in Africa. Diabetologia 62:1204–1211. https://doi.org/10.1007/s00125-019-4880-7
9. Sandholm N, Forsblom C, Mäkinen V-P et al (2014) Genome-wide association study of urinary albumin excretion rate in patients with type 1 diabetes. Diabetologia 57:1143–1153. https://doi.org/10.1007/s00125-014-3202-3
10. Rich SS, Goodarzi MO, Palmer ND et al (2009) A genome-wide association scan for acute insulin response to glucose in Hispanic-Americans: the Insulin Resistance Atherosclerosis Family Study (IRAS FS). Diabetologia 52:1326–1333. https://doi.org/10.1007/s00125-009-1373-0
11. Liu Q, Tang B, Zhu Z et al (2022) A genome-wide cross-trait analysis identifies shared loci and causal relationships of type 2 diabetes and glycaemic traits with polycystic ovary syndrome. Diabetologia (2022) 65:1483–1494. https://doi.org/10.1007/s00125-022-05746-x
12. Zheng J, Xu M, Walker V et al (2022) Evaluating the efficacy and mechanism of metformin targets on reducing Alzheimer's disease risk in the general population: a Mendelian randomisation study. Diabetologia 65: 1664–1675. https://doi.org/10.1007/s00125-022-05743-0