Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Cell Genom. Author manuscript; available in PMC 2021 December 02.

Published in final edited form as:

Cell Genom. 2021 October 13; 1(1): . doi:10.1016/j.xgen.2021.100005.

Sequencing-based genome-wide association studies reporting standards

Aoife McMahon[1,4,*], **Elizabeth Lewis**[1], **Annalisa Buniello**[1], **Maria Cerezo**[1], **Peggy Hall**[2], **Elliot Sollis**[1], **Helen Parkinson**[1,*], **Lucia A. Hindorff**[2], **Laura W. Harris**[1], **Jacqueline A.L. MacArthur**[1,3]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK

[2]Division of Genomic Medicine, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA

[3]BHF Data Science Centre, Health Data Research UK, London, UK

[4]Lead contact

SUMMARY

Genome sequencing has recently become a viable genotyping technology for use in genome-wide association studies (GWASs), offering the potential to analyze a broader range of genome-wide variation, including rare variants. To survey current standards, we assessed the content and quality of reporting of statistical methods, analyses, results, and datasets in 167 exome- or genome-wide-sequencing-based GWAS publications published from 2014 to 2020; 81% of publications included tests of aggregate association across multiple variants, with multiple test models frequently used. We observed a lack of standardized terms and incomplete reporting of datasets, particularly for variants analyzed in aggregate tests. We also find a lower frequency of sharing of summary statistics compared with array-based GWASs. Reporting standards and increased data sharing are required to ensure sequencing-based association study data are findable, interoperable, accessible, and reusable (FAIR). To support that, we recommend adopting the standard terminology of sequencing-based GWAS (seqGWAS). Further, we recommend that single-variant analyses be reported following the same standards and conventions as standard array-based GWASs and be

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

*Correspondence: aoifem@ebi.ac.uk (A.M.), parkinson@ebi.ac.uk (H.P.).

AUTHOR CONTRIBUTIONS

Conceptualization, A.M., H.P., L.A.H., and J.A.L.M.; methodology, A.M.; formal analysis, A.M.; investigation, A.M. and J.A.L.M.; data curation (GWAS Catalog), A.M., E.L., A.B., M.C., P.H., E.S., L.W.H., and J.A.L.M.; data curation (sequencing papers), A.M. and E.L.; writing – original draft, A.M. and J.A.L.M.; writing – review & editing, A.M., J.A.L.M., L.H., and L.W.H.; visualization, A.M.; supervision, J.A.L.M. and L.W.H.; project administration, H.P., J.A.L.M., and L.W.H.; funding acquisition, H.P.

DECLARATION OF INTERESTS

An immediate family member of J.A.L.M. is an employee and shareholder of Illumina.

WEB RESOURCES

GWAS Catalog eligibility criteria, https://www.ebi.ac.uk/gwas/docs/methods/criteria

Update to NIH management of genomic summary results access, https://datascience.nih.gov/foa/update-nih-management-genomic-summary-results-access

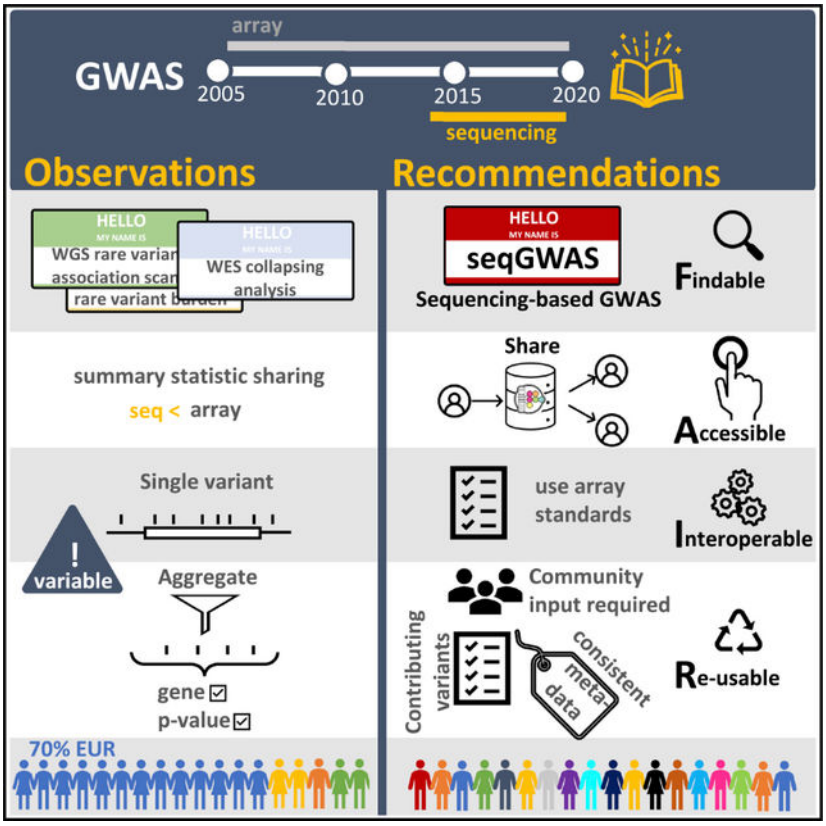
SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2021.100005.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

shared in the GWAS Catalog. We also provide initial recommended standards for aggregate analyses metadata and summary statistics.

Graphical Abstract



In brief

McMahon et al. report an analysis of the sequencing-based GWAS literature, finding a lack of standardized language and incomplete reporting, along with less-frequent sharing of summary statistics compared with that of array-based GWASs. We provide recommendations for the reporting and sharing of sequencing-based GWASs to increase FAIRness of these valuable datasets.

INTRODUCTION

Huge advances in the field of human genetics can be attributed to the advent of genome-wide association studies (GWASs) more than 15 years ago.[1,2] In recent years, decreasing costs and advances in analytic methods have made high-throughput whole-genome sequencing (WGS) and whole-exome sequencing (WES) feasible alternatives to array-based genotyping in GWASs.[3,4] Sequencing offers a significant advantage over array-based methods, with the potential to detect and genotype all variants present in a sample, not only those present on an array or imputation reference panel. Most arrays are designed to assay common variants (minor allele frequency [MAF] > 5%), omitting rare (MAF <

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1%) and low-frequency (MAF 1%–5%) variants. The analysis of these rarer variants could explain additional disease risk or trait variability and help overcome the problem of "missing heritability."[5,6] In addition, most arrays have historically been biased toward coverage of variation in European populations.[7] The fact that sequencing potentially provides an unbiased assessment of variants within the population studied is particularly important for studies of non-European populations.[8,9]

There are challenges with analyzing many more and rarer variants. Single-variant tests, used as the standard in array-based GWASs, are typically underpowered when applied to low-frequency or rare variants, unless sample sizes or effects are very large. There are also issues with correcting for multiple testing when the number of statistical tests is very large. To address those issues, statistical methods have been designed specifically for rare-variant-association testing, which evaluate aggregate association over multiple variants in a genomic region (referred to here as "aggregate tests").[10] Variants are typically aggregated across biologically functional regions (e.g., a gene) with variants enriched for those likely to have larger effect sizes based on annotated or predicted functional effect (e.g., located in a splice junction or a predicted loss of function). The power of a particular aggregate test to detect an association will depend on how closely the model's assumptions and contributing variants represent the true disease mechanism at each locus.

Repositories of scientific data have been indispensable in supporting research and in facilitating discoverability and integration across datasets through standard formats. The National Human Genome Research Institute-European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog[11] is the preeminent data resource of large-scale genetic-association studies, enabling research to identify causal variants, to understand disease mechanisms, and to establish targets for novel therapies.[12] The GWAS Catalog infrastructure, data content, and standard formats have been designed to support array-based GWASs. Attempts to expand the scope of the Catalog to include sequencing-based association studies have been hindered by the need to develop new standards for the differences in methods, the metadata required to represent them, and the format of the results, particularly for aggregate analyses.

Here, we analyze the current landscape of published sequencing-based association studies to determine requirements for hosting and sharing those datasets in the GWAS Catalog and recommend best practices for reporting. First, we comprehensively reviewed publications reporting sequencing-based association studies, assessing the range of experimental designs and statistical methods, as well as the content and quality of reporting for analyses, methods, and datasets included in publications. We hope that this review will form a rallying point for building community consensus on standards. This work has also informed the development of the GWAS Catalog infrastructure and data-representation schema to support inclusion of sequencing-based association studies, which are now accepted for submission at the GWAS Catalog. Our work at the GWAS Catalog is focused on enabling broad data sharing and defining standards to ensure sequencing-based association study data are findable, interoperable, accessible, and reusable (FAIR).[13]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

RESULTS

Finding sequencing-based association studies

In our review of research publications (STAR Methods), we observed that a wide range of terms are used to describe sequencing-based genome or exome-wide association studies. The term “GWAS” is rarely used, and we have not seen an equivalent standard term emerge (Figure S1). Combinations of terminology were used, related to (1) analysis of associations (e.g., rare variant association analysis, rare variant aggregate association analysis, association test, and genome-wide significant associations), (2) the allele frequency of the variants analyzed (e.g., common variant and rare variant), (3) the analysis type, either single variant (e.g., single variant and variant level) or aggregate with multiple variants (e.g., gene-based, region-based, aggregate, gene burden, collapsing analysis, gene-level association, gene-level signal, and collapsed-variant tests).

We identified 167 publications reporting genome-wide sequencing-based association analyses meeting our selection criteria (STAR Methods; Tables S1 and S2). The first study was published in 2014, with the number of publications increasing year after year to 2020 (Figure 1A). Because no standard terminology has been adopted for these studies, we were not able to search discriminately for sequencing-based association studies meeting our criteria, and permissive searches (e.g., for “WGS OR WES association”) yield too many results to feasibly review manually (Figure S2); therefore, we expect this to be an underestimate of publications reporting sequencing-based GWASs (seqGWAS). Most publications analyzed WES data only (68%), approximately one-third analyzed WGS data (30%), and some publications included both coverage types (2%) (Figure 1A). Many publications that used WES and WGS sequencing data limited their analyses to pre-specified regions of interest; those targeted analyses are not the focus of this work and were, therefore, excluded from the analysis.

Association tests and qualifying variants

We surveyed the types of association tests included in these publications. Most frequent was the inclusion of both single-variant and aggregate analyses (48%), followed by aggregate analysis only (33%), and a minority of publications (19%) included single-variant analyses only (Figure 1B). Of the publications including aggregate tests, a wide range of statistical models and tools were used, with publications commonly using multiple models. For example, of publications that used one of the three most-common aggregation methods[10] (burden/collapsing, variance-component [SKAT], and combined burden and variance-component [SKAT-O] tests), 40% (n = 65) used at least two of those methods (Figure 2A). The language used to describe those methods is varied; for example, SKAT is referred to variously as kernel based, dispersion based, or variance-component based (Figure S3).

We also examined variant-filtering or “masking” approaches. Minor allele frequency thresholds were reported in 72% of single-variant and 84% of aggregate-analysis publications, with the remainder either not reporting any MAF threshold or using all variants (26% of single variant/16% of aggregate) (Figure S4). “Greater than” thresholds were

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

typically used for single-variant analysis, with 57% of analyses employing a MAF threshold of 0.01 or greater, limiting those analyses to the common variant space (Figure 2B) (n = 30/53 thresholded analyses from 51 publications). In contrast, aggregate analyses typically employed “less than” thresholds, to include only low-frequency (<0.05), rare (<0.005), or ultra-rare variants. Most aggregate analyses used <0.01 or <0.05 thresholds (78%, n = 67/86 thresholded analyses from 77 publications).

Many publications (63%, n = 75/120) also performed analyses on variants with predicted biological effect. Authors filtered for predicted functional effect based on transcript annotation (e.g., using the Variant Effect Predictor[14]) or protein structure (e.g., using Sorting Intolerant from Tolerant [SIFT],[15] Polymorphism Phenotyping v2 [PolyPhen][16] and combined annotation-dependent depletion [CADD][17]) or based on measures of evolutionary conservation or variation intolerance.[18,19] An analysis of the text used to describe the filtering process highlights that the most commonly used terms were “splice,” “missense,” “protein,” “frameshift,” “stop gain,” “loss of function” (LoF), and “protein-truncating variant” (PTV), but a wide range of terms were used (Figure S5). Variants were often filtered by both annotation/predicted effect and MAF thresholds, with multiple different filtering criteria used per publication (examples are provided in Table S3).

The number of variants analyzed in WES single-variant analyses is considerably less than those typically analyzed in array-based GWASs (median, 158,091; versus 5,554,549), whereas, in WGS single-variant analyses, the number is greater (median, 12,210,410) (Table 1). The median number of statistical tests performed in aggregate analyses was 18,360, approximating the number of protein-coding genes with a consensus CDS (19,033; coding DNA sequence)[20] because the most-common unit over which variants are aggregated is the protein-coding gene. The analyses in which the number of tests was greater than the inter-quartile range were those in which the unit of analysis was non-genic. The most-common non-genic aggregation units we observed were regulatory regions[18,19,21,22] or agnostic sliding windows.[23–26] Authors also aggregated across evolutionary conserved regions or pathways.[19,27]

The outcome of the various variant filters or “masks,” i.e., a list of the qualifying variants included in each analysis, was not provided in any of the 167 publications we analyzed. However, some publications did specify the number of qualifying variants included per unit of aggregation.[28,29]

Sample characteristics

We next surveyed the characteristics of samples (sample size, ancestry, and traits) studied in seqGWAS. We compared the sample sizes of the seqGWAS, because that is a key determinant of statistical power. We classified publications into bins based on the number of individuals in the publication (Figure S6). The most-common sample size bin was 300–3,000 individuals (43% of publications), but in the past few years, there has been a near-even distribution across bins from small to large sample sizes. In 2019, both the smallest (<300 individuals) and the largest (>10,000) sample-size bins were used in approximately a quarter of publications each (23% and 26%, respectively; Figure S6). The number of cases is also a component of statistical power, and unbalanced case/control ratios can inflate type

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1 errors.[30] We observed 10 publications (6%) with unbalanced case/control ratios (cases ≤ 15% of samples), most of those (n = 7, 4%) being highly unbalanced (cases ≤ 4% of samples) (Table S4).[31–33]

The inclusion of diverse ancestral backgrounds in genomics studies is recognized as important,[34,35] but analysis of array-based GWASs has highlighted the extreme bias toward samples of European origin.[36,37] We assessed and compared ancestry in seqGWAS. Following the GWAS Catalog ancestry framework (a standard methodology for representing ancestry),[36] we extracted publication-level, broad ancestral categories of samples. Mirroring what has been seen elsewhere with array-based GWASs, 71% of all publications (n = 85/120) included European ancestry individuals, with 40% not including any other ancestry (n = 48/120) (Figure 3A; Table S5). The second most commonly examined ancestral group was African American (28% of publications, n = 33/120), and most of those publications (21%) also included other ancestries (Figures 3B and S7). This profile may, in part, be due to the presence of large, trans-ancestry consortia, such as the Trans-Omics for Precision Medicine (TOPMed) program, which is the most commonly occurring consortium or cohort mentioned (Table S7).

We also examined the number of traits analyzed within the reported association study. Most publications examined one or two traits (76%, n = 89), whereas a few (4%, n = 5) examined 55–75 traits as part of larger-scale studies.[18,22,39–41] More recently (2019–2020), very-large-scale studies using the UK Biobank have included 791–4,262 traits[42–44] (Figure S8). Non-UK-Biobank publications analyzing multiple traits were mostly focused on quantitative biomarker or metabolite-level-type traits,[18,21,41,45] such as inflammatory biomarkers, blood metabolite levels, blood protein levels. Studies analyzing fewer traits were more likely to be case/control studies.[46–49] A full list of publication-level trait names (analogous to the GWAS Catalog “reported trait”) and corresponding mapped Experimental Factor Ontology (EFO) terms are provided in Table S4.

Data availability

The public availability of full summary statistics from GWASs has great potential to extend the power of initial studies by enabling the community to re-analyze, meta-analyze, and perform follow-up analyses, with minimal risk to participants.[11,50] We assessed whether summary statistics, in addition to individual-level genotyping results, were reported in these publications as available without restriction in a public repository. Sharing of sequencing-based single-variant summary statistics was much lower (5% of publications, n = 4/79, 2014–2019) than the proportion of array-based publications in the GWAS Catalog in the same period (12% of publications, n = 300/2,571, 2014–2019) (Table 1). Sharing of array-GWAS summary statistics is greater in recent years (19% of 2019 GWAS Catalog publications, n = 101/527), but seqGWAS summary statistics still lag (9%, n = 3/32). A further 2.5% of sequencing publications (n = 3/120, 2014–2019) deposited summary statistics in a controlled-access public repository (the Database of Genotypes and Phenotypes [dbGAP]). In contrast, 24% of publications (n = 29/120) deposited individual-level sequencing data in controlled access repositories (dbGAP or European

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Genome-Phenome Archive [EGA]) (Table S6) and, for some summary-level data, may have been co-submitted or bundled with those data but not specifically stated by the authors.

The data content of single-variant summary statistics for seqGWAS is comparable with that for standard-array GWASs and can conform to emerging standards.[11,50] However, summary statistics for aggregate analysis in seqGWAS are commonly composed only of a gene name (or other range specifying chromosomal coordinates), p value, and often the number of contributing variants, sometimes separated by cases/controls. Crucially, we did not observe any publications that reported the list of variants included in each aggregate unit, which is key to interpretation of the data, either in the main text or in accompanying material.

DISCUSSION

Recommended standards

Based on our review and analyses, we recommend standards to improve the reporting and accessibility of seqGWAS. First, to increase transparency when referring to study design and facilitate identification, we recommend that the community adopt the name of "sequencing-based GWAS," abbreviated as "seqGWAS" (Box 1, recommendation 1). Second, to enable accurate interpretation and comparison of results across studies and loci, it is essential that detailed information describing each association test (including statistical tests and contributing variants) are consistently reported (Box 1, recommendations 2 and 3). These recommendations are based upon, and are designed to address, our observations of the state of the field.

Observations

The sequencing-based association studies in the publications we analyzed included either single or aggregate multi-variant analyses. The restriction of single-variant analyses to common variants renders those studies largely comparable with array-based GWASs (Figure 2), with similar implications for data content and reporting (Box 1, recommendation 2) and similar utility for re-use, for example, in the derivation of polygenic scores or in Mendelian randomization. In comparison, studies performing tests of aggregate association across multiple variants, which appear in most (81%) publications, focus on "low-frequency," "rare," and "ultra-rare" variants. Multiple statistical models of aggregate association are frequently used in the same publication because the power of each test depends on how closely the assumptions of the model match the true disease etiology at each locus. Therefore, there is no best model (including statistical tests and variant filtering strategies) across loci and traits, and there is no best model necessarily knowable *a priori*. To enable accurate interpretation and comparison of results across studies and loci, it is, therefore, essential that detailed information describing each association test (including statistical tests and contributing variants) is consistently reported (Box 1, recommendations 2 and 3).

It is in the performance and, therefore, reporting of aggregate association tests that sequencing-based association studies differ most from standard array-based GWASs. We observed that the experimental information provided for aggregate tests was not sufficient to facilitate thorough examination or replication. Variants are filtered (typically by MAF

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and functional annotation/predicted consequence) and combined in different units of aggregation. Crucially, the list of variants contributing to each test is not provided by these publications. Availability of these data would facilitate attempts at replication and enable further analysis and functional investigation[51] (Box 1, recommendation 3b).

Given the rarity of these variants, privacy concerns regarding de-identification may be a barrier to their sharing. We suggest that the community look to the field of rare-variant clinical genomics, in which it is becoming increasingly accepted that the potential benefits of sharing far outweigh the perceived risks.[52] This is illustrated by the number of clinical-laboratory-derived variants in ClinVar more than doubling since 2018.[53,54] We note that individual genetic variants, even very rare ones, are not uniquely identifying and would require in-depth knowledge of an individual's genotype to connect an individual to a phenotype.

Theoretically, lists of qualifying variants could be recapitulated, but filtering information provided by authors is again diverse and often vague and, overall, insufficient to independently derive those lists. The community should consider standardized ways to communicate variant filters or masks (for example, using the sequence ontology to describe functional annotation/predicted functional effect filters[55]). The unit of aggregation, which encompasses the variants included in each test (typically gene), must be clearly defined. This should include the coordinates of the region and the genome assembly or annotation release, along with any additional variant-filtering information (Box 1, recommendation 3a).

We observed that a smaller proportion of full-summary statistics are publicly available from seqGWAS (5%) compared with array-based GWASs (12%). That percentage is low for both types of studies despite guidance and growing community consensus supporting sharing (web resources).[50] There are a number of reasons why full and public data sharing may be less for sequencing than array-based studies. There may be additional perceived privacy concerns regarding the rare variants present in sequencing-based summary statistics. It is also possible that summary statistics may be bundled with the individual-level genotyping data that 24% of publications deposited in controlled-access repositories (dbGAP/EGA). Single-variant summary statistics can conform to the proposed array-based standards (Box 1, recommendation 2)[11] and can already be submitted to the GWAS Catalog. However, aggregate-analysis summary statistics, when they are shared, are typically only a gene name and a p value (sometimes with the number of qualifying variants included). These files are not large or cumbersome, given that the number of human genes is only approximately 20,000 and are easy to share, for example, as a supplementary table. As described above, we recommend authors supply full lists of qualifying variants that contribute to each test (Box 1, recommendation 3b). We hope that the development and adoption of these standards will simplify and encourage the sharing of seqGWAS summary statistics.

The ability of sequencing to genotype all variants present in the cohort offers a significant opportunity to overcome the biases inherent in array-based genotyping, with the potential to reduce disparities among ancestry groups. Despite that, the bias toward European-ancestry populations observed in array-based GWASs (49% European only and 74% including European) remains in sequencing publications (40% European only and 71% including

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

European). Furthermore, we note that the percentage of European sequencing-based analyses is likely to be greater; publications containing multiple GWASs are more likely to be from large cohorts with deep phenotyping data, which are predominantly European (e.g., UK Biobank). Given the advantages of sequencing in analyzing non-Europeans, we question why it is not being further used. There are many possible reasons for this, including increased cost, the lack of diversity in legacy cohorts, pre-existing consent agreements, privacy concerns associated with rare-variant analysis, and analysis methods being complex. The GWAS Catalog reiterates its stance in encouraging analysis of diverse populations and encourages researchers to take advantage of the opportunities offered by sequencing technologies in enabling unbiased genotyping across ancestries (Box 1, recommendation 4).

Limitations of the study

The lack of standardized terms to refer to seqGWAS creates challenges for the reliable identification of these publications using term-based literature-search methods. The 167 publications we identified are, therefore, certainly an underestimate of the number of publications, and we do not claim that this work is a comprehensive analysis of all published seqGWAS. To maintain consistency and enable comparability across studies, we decided to limit our analysis to publications carrying out an unbiased, genome-wide or exome-wide assessment of loci associated with traits, equivalent to the GWAS Catalog's inclusion criteria (web resources). Many of the publications we screened and deemed ineligible were targeted analyses based on prior knowledge, for example, to specific loci, genes, or pathways and are scientifically valid studies but are out of the scope of this manuscript. In our recommendation of the term "seqGWAS" (Box 1, recommendation 1), we note that some may feel the use of "GWAS" is inappropriate, primarily because WES-based analyses are necessarily targeted to expressed regions. However, we observe that the term "GWAS" is commonly used to refer to both genome-wide and exome-wide array-based association studies. Our motivation for suggesting a unique nomenclature (sequencing-based GWAS/seqGWAS) is to facilitate the "find-ability" of these study types (large-scale association studies that analyze variants spread across the genome (e.g., with coverage across all autosomal chromosomes) in the scientific literature.

A necessary limitation of this work is its restriction to a specific time period (2014–2020), and as such, it serves as a snapshot of the state of the field. It is anticipated that the field will grow significantly in the immediate future, and the ratio of WES and WGS studies may change. However, the findings of our work, in terms of how studies are described and reported, are unaffected by whether or not they are WES or WGS or the total number of studies. The recommendations similarly apply to both coverage types. Furthermore, we believe this is an appropriate time to publish a study such as ours so that standards can be established sooner, thus enabling future publications to adhere to the FAIR principles.

Ensuring seqGWAS are FAIR

The maximum benefit of scientific research can only be realized if data are FAIR (findable, accessible, interoperable, and reusable), as described by the FAIR guiding principles for good scientific data management.[13] Our analysis highlights several obstacles to implementation of these principles for seqGWAS, including lack of an appropriate

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

resource or repository to store and disseminate the data, consistency of metadata reporting without the use of structured vocabularies, clarity on metadata indexing that needs to support searching, and a community standard for summary statistics. The GWAS Catalog's primary aim is to provide a comprehensive resource and repository of all large-scale genomic association studies and, as such, has extended its scope to include seqGWAS, initially focusing on single-variant analyses. We will support the community to reach consensus on the reporting of aggregate seqGWAS, including the creation of standards for metadata and summary format and content.[50] The development and adoption of reporting standards will increase the availability, accessibility, and utility of seqGWAS. We include a summary of our recommendations (Box 1) and welcome further input from the community.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Aoife McMahon (aoifem@ebi.ac.uk).

Materials availability—This study did not generate new unique reagents.

Data and code availability—Data underlying analyses in this paper are curated from the literature and are presented in Table S4.

This paper does not report original code.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

To enable direct comparability with array-based GWAS we defined sequencing-based association studies as studies that analyze associations between a trait and a genome-wide distribution of genetic variants from either whole-genome or whole-exome sequencing. This does not include targeted sequencing studies that are limited to specific genomic regions or subsets of genes (e.g., publications[57–59]). From these, we selected studies with population-based association analyses, and did not include studies that used family structure/linkage (e.g., publications[60–62]) or were aimed at diagnostic discovery of pathogenic variants (e.g publications[63–65]). We also included family-based association studies, but only if they performed standard association analysis with relatedness accounted for in the model (e.g., publications[39,66]). Studies that combine array and sequencing-based genotyping, such as partially array-genotyped, or array genotyped with sequencing data used as an imputation panel, were not included in our analyses.

Sequencing-based association publications meeting these inclusion criteria were identified by several routes: Pubmed and EuropePMC literature searches, the GWAS Catalog machine learning-based literature search,[56] examination of grants, cohort and project websites, social

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

media, conference talks, references in publications and personal communications (Table S1). The source of initial identification of each sequencing publication was recorded. Publication level metadata relating to study design, sample description, traits examined and data availability were extracted (Tables S2 and S4). Publication triage, eligibility assessment and extraction of metadata were performed by experienced GWAS Catalog curators. Analysis of study eligibility, genomic coverage and analysis type was performed for 2020 publications. More detailed analysis of the sample, trait, data sharing and statistical tests was available to the end of 2019.

QUANTIFICATION AND STATISTICAL ANALYSIS

For analysis of text related to variant types, curators extracted sentences describing variant selection and relevant terms were identified using the text analysis tool MonkeyLearn (https://monkeylearn.com/word-cloud/). The output was examined by expert curators and non-relevant terms excluded, terms collapsed and missed relevant terms were added and counted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award no. U41HG007823 and EMBL-EBI Core Funds. L.A.H. and P.H. are employees of the National Human Genome Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. In addition, we acknowledge funding from the European Molecular Biology Laboratory. We thank Kalliope Panoutsopoulou and Aimee Deaton for comments on the manuscript.

REFERENCES

1. Klein RJ, Xu X, Mukherjee S, Willis J, and Hayes J (2010). Successes of genome-wide association studies. Cell 142, 350–351, author reply 353–355. [PubMed: 20691890]
2. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678. [PubMed: 17554300]
3. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet 43, 491–498. [PubMed: 21478889]
4. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, Gupta N, Neale BM, Daly MJ, Sklar P, et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet 44, 631–635. [PubMed: 22610117]
5. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, and Nadeau JH (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet 11, 446–450. [PubMed: 20479774]
6. Zuk O, Hechter E, Sunyaev SR, and Lander ES (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA 109, 1193–1198. [PubMed: 22223662]
7. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, and Kenny EE (2017). Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet 100, 635–649. [PubMed: 28366442]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

8. Lachance J, and Tishkoff SA (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. BioEssays 35, 780–786. [PubMed: 23836388]

9. Kim MS, Patel KP, Teng AK, Berens AJ, and Lachance J (2018). Genetic disease risks can be misestimated across global populations. Genome Biol. 19, 179. [PubMed: 30424772]

10. Lee S, Abecasis GR, Boehnke M, and Lin X (2014). Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet 95, 5–23. [PubMed: 24995866]

11. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. (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. [PubMed: 30445434]

12. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, and Yang J (2017). 10 Years of GWAS discovery: Biology, function, and translation. Am. J. Hum. Genet 101, 5–22. [PubMed: 28686856]

13. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. [PubMed: 26978244]

14. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, and Cunningham F (2016). The Ensembl Variant Effect Predictor. Genome Biol. 17, 122. [PubMed: 27268795]

15. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, and Ng PC (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 40, W452–W457. [PubMed: 22689647]

16. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR (2010). A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249. [PubMed: 20354512]

17. Rentzsch P, Witten D, Cooper GM, Shendure J, and Kircher M (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 47 (D1), D886–D894. [PubMed: 30371827]

18. Yu B, de Vries PS, Metcalf GA, Wang Z, Feofanova EV, Liu X, Muzny DM, Wagenknecht LE, Gibbs RA, Morrison AC, and Boerwinkle E (2016). Whole genome sequence analysis of serum amino acid levels. Genome Biol. 17, 237. [PubMed: 27884205]

19. Kim D, Basile AO, Bang L, Horgusluoglu E, Lee S, Ritchie MD, Saykin AJ, and Nho K (2017). Knowledge-driven binning approach for rare variant association analysis: application to neuroimaging biomarkers in Alzheimer's disease. BMC Med. Inform. Decis. Mak 17 (suppl 1), 61. [PubMed: 28539126]

20. Pujar S, O'Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C, Girón CG, Diekhans M, Barnes I, Bennett R, et al. (2018). Consensus coding sequence (CCDS) database: A standardized set of human and mouse protein-coding regions supported by expert curation. Nucleic Acids Res. 46 (D1), D221–D228. [PubMed: 29126148]

21. de Vries PS, Yu B, Feofanova EV, Metcalf GA, Brown MR, Zeighami AL, Liu X, Muzny DM, Gibbs RA, Boerwinkle E, and Morrison AC (2017). Whole-genome sequencing study of serum peptide levels: The Atherosclerosis Risk in Communities study. Hum. Mol. Genet 26, 3442–3450. [PubMed: 28854705]

22. Gilly A, Suveges D, Kuchenbaecker K, Pollard M, Southam L, Hatzikotoulas K, Farmaki AE, Bjornland T, Waples R, Appel EVR, et al. (2018). Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits. Nat. Commun 9, 4674. [PubMed: 30405126]

23. He Z, Xu B, Buxbaum J, and Ionita-Laza I (2019). A genome-wide scan statistic framework for whole-genome sequence data analysis. Nat. Commun 10, 3018. [PubMed: 31289270]

24. Sarnowski C, Satizabal CL, DeCarli C, Pitsillides AN, Cupples LA, Vasan RS, Wilson JG, Bis JC, Fornage M, Beiser AS, et al. ; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; TOPMed Neurocognitive Working Group (2018). Whole genome sequence analyses of brain imaging measures in the Framingham Study. Neurology 90, e188–e196. [PubMed: 29282330]

25. Mak ACY, White MJ, Eckalbar WL, Szpiech ZA, Oh SS, PinoYanes M, Hu D, Goddard P, Huntsman S, Galanter J, et al. ; NHLBI Trans-Omics for Precision Medicine (TOPMed)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Consortium (2018). Whole-genome sequencing of pharmacogenetic drug response in racially diverse children with asthma. Am. J. Respir. Crit. Care Med 197, 1552–1564. [PubMed: 29509491]

26. Sapkota Y, Cheung YT, Moon W, Shelton K, Wilson CL, Wang Z, Mulrooney DA, Zhang J, Armstrong GT, Hudson MM, et al. (2019). Whole-genome sequencing of childhood cancer survivors treated with cranial radiation therapy identifies 5p15.33 locus for stroke: A report from the St. Jude Lifetime Cohort Study. Clin. Cancer Res 25, 6700–6708. [PubMed: 31462438]
27. Monson ET, Pirooznia M, Parla J, Kramer M, Goes FS, Gaine ME, Gaynor SC, de Klerk K, Jancic D, Karchin R, et al. (2017). Assessment of whole-exome sequence data in attempted suicide within a bipolar disorder cohort. Mol. Neuropsychiatry 3, 1–11. [PubMed: 28879196]
28. Gratten J, Zhao Q, Benyamin B, Garton F, He J, Leo PJ, Mangelsdorf M, Anderson L, Zhang ZH, Chen L, et al. (2017). Whole-exome sequencing in amyotrophic lateral sclerosis suggests NEK1 is a risk gene in Chinese. Genome Med. 9, 97. [PubMed: 29149916]
29. Scott WK, Medie FM, Ruffin F, Sharma-Kuinkel BK, Cyr DD, Guo S, Dykxhoorn DM, Skov RL, Bruun NE, Dahl A, et al. (2018). Human genetic variation in GLS2 is associated with development of complicated *Staphylococcus aureus* bacteremia. PLoS Genet. 14, e1007667. [PubMed: 30289878]
30. Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, and Goldstein DB (2019). Rare-variant collapsing analyses for complex traits: Guidelines and applications. Nat. Rev. Genet 20, 747–759. [PubMed: 31605095]
31. Udagawa C, Horinouchi H, Shiraishi K, Kohno T, Okusaka T, Ueno H, Tamura K, Ohe Y, and Zembutsu H (2019). Whole genome sequencing to identify predictive markers for the risk of drug-induced interstitial lung disease. PLoS ONE 14, e0223371. [PubMed: 31584970]
32. Wolock CJ, Stong N, Ma CJ, Nagasaki T, Lee W, Tsang SH, Kamalakaran S, Goldstein DB, and Allikmets R (2019). A case-control collapsing analysis identifies retinal dystrophy genes associated with ophthalmic disease in patients with no pathogenic ABCA4 variants. Genet. Med 21, 2336–2344. [PubMed: 30926958]
33. Alkelai A, Greenbaum L, Heinzen EL, Baugh EH, Teitelbaum A, Zhu X, Strous RD, Tatarskyy P, Zai CC, Tiwari AK, et al. (2019). New insights into tardive dyskinesia genetics: Implementation of whole-exome sequencing approach. Prog. Neuropsychopharmacol. Biol. Psychiatry 94, 109659. [PubMed: 31153890]
34. Popejoy AB, and Fullerton SM (2016). Genomics is failing on diversity. Nature 538, 161–164. [PubMed: 27734877]
35. Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, Lam M, Iyegbe C, Strawbridge RJ, Brick L, et al. (2019). Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. Cell 179, 589–603. [PubMed: 31607513]
36. Morales J, Welter D, Bowler EH, Cerezo M, Harris LW, McMahon AC, Hall P, Junkins HA, Milano A, Hastings E, et al. (2018). A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biol. 19, 21. [PubMed: 29448949]
37. Mills MC, and Rahal C (2019). A scientometric review of genome-wide association studies. Commun. Biol 2, 9. [PubMed: 30623105]
38. Hulsen T, de Vlieg J, and Alkema W (2008). BioVenn—A web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics 9, 488. [PubMed: 18925949]
39. Sabo A, Mishra P, Dugan-Perez S, Voruganti VS, Kent JW Jr., Kalra D, Cole SA, Comuzzie AG, Muzny DM, Gibbs RA, and Butte NF (2017). Exome sequencing reveals novel genetic loci influencing obesity-related traits in Hispanic children. Obesity (Silver Spring) 25, 1270–1276. [PubMed: 28508493]
40. Locke AE, Steinberg KM, Chiang CWK, Service SK, Havulinna AS, Stell L, Pirinen M, Abel HJ, Chiang CC, Fulton RS, et al. ; FinnGen Project (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. Nature 572, 323–328. [PubMed: 31367044]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

41. Höglund J, Rafati N, Rask-Andersen M, Enroth S, Karlsson T, Ek WE, and Johansson Å (2019). Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. Sci. Rep 9, 16844. [PubMed: 31727947]
42. Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM, and Yang J (2019). A resource-efficient tool for mixed model association analysis of large-scale data. Nat. Genet 51, 1749–1755. [PubMed: 31768069]
43. Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, Fath DM, Sandoval E, Isaksson M, Schlauch KA, et al. (2020). Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat. Commun 11, 542. [PubMed: 31992710]
44. Zhao Z, Bi W, Zhou W, VandeHaar P, Fritsche LG, and Lee S (2020). UK Biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. Am. J. Hum. Genet 106, 3–12. [PubMed: 31866045]
45. Long T, Hicks M, Yu HC, Biggs WH, Kirkness EF, Menni C, Zierer J, Small KS, Mangino M, Messier H, et al. (2017). Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. Nat. Genet 49, 568–578. [PubMed: 28263315]
46. Hammer MF, Ishii A, Johnstone L, Tchourbanov A, Lau B, Sprissler R, Hallmark B, Zhang M, Zhou J, Watkins J, and Hirose S (2017). Rare variants of small effect size in neuronal excitability genes influence clinical outcome in Japanese cases of SCN1A truncation-positive Dravet syndrome. PLoS ONE 12, e0180485. [PubMed: 28686619]
47. Grant RC, Denroche RE, Borgida A, Virtanen C, Cook N, Smith AL, Connor AA, Wilson JM, Peterson G, Roberts NJ, et al. (2018). Exome-wide association study of pancreatic cancer risk. Gastroenterology 154, 719–722.e3. [PubMed: 29074453]
48. Kwak SH, Chae J, Lee S, Choi S, Koo BK, Yoon JW, Park JH, Cho B, Moon MK, Lim S, et al. (2018). Nonsynonymous variants in *PAX4* and *GLP1R* are associated with type 2 diabetes in an East Asian population. Diabetes 67, 1892–1902. [PubMed: 29941447]
49. Sveinbjornsson G, Olafsdottir EF, Thorolfsdottir RB, Davidsson OB, Helgadottir A, Jonasdottir A, Jonasdottir A, Bjornsson E, Jensson BO, Arnadottir GA, et al. (2018). Variants in NKX2–5 and FLNC Cause Dilated Cardiomyopathy and Sudden Cardiac Death. Circ. Genom. Precis. Med 11, e002151. [PubMed: 30354339]
50. MacArthur JAL, Buniello A, Harris LW, Hayhurst J, McMahon A, Sollis E, Cerezo M, Hall P, Lewis E, Whetzel PL, et al. (2021). Workshop proceedings—GWAS summary statistics standards and sharing. Cell Genomics 18, 100004-1–100004-8.
51. Lappalainen T, Scott AJ, Brandt M, and Hall IM (2019). Genomic analysis in the age of human genome sequencing. Cell 177, 70–84. [PubMed: 30901550]
52. Wright CF, Ware JS, Lucassen AM, Hall A, Middleton A, Rahman N, Ellard S, and Firth HV (2019). Genomic variant sharing: A position statement. Wellcome Open Res. 4, 22. [PubMed: 31886409]
53. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 46 (D1), D1062–D1067. [PubMed: 29165669]
54. Pérez-Palma E, Gramm M, Nürnberg P, May P, and Lal D (2019). Simple ClinVar: An interactive web server to explore and retrieve gene and disease variants aggregated in ClinVar database. Nucleic Acids Res. 47 (W1), W99–W105. [PubMed: 31114901]
55. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, and Ashburner M (2005). The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 6, R44. [PubMed: 15892872]
56. Lee K, Famiglietti ML, McMahon A, Wei CH, MacArthur JAL, Poux S, Breuza L, Bridge A, Cunningham F, Xenarios I, and Lu Z (2018). Scaling up data curation using deep learning: An application to literature triage in genomic variation resources. PLoS Comput. Biol 14, e1006390. [PubMed: 30102703]
57. Asanomi Y, Shigemizu D, Miyashita A, Mitsumori R, Mori T, Hara N, Ito K, Niida S, Ikeuchi T, and Ozaki K (2019). A rare functional variant of SHARPIN attenuates the inflammatory response and associates with increased risk of late-onset Alzheimer's disease. Mol. Med 25, 20. [PubMed: 31216982]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

58. Moore C, Blumhagen RZ, Yang IV, Walts A, Powers J, Walker T, Bishop M, Russell P, Vestal B, Cardwell J, et al. (2019). Resequencing study confirms that host defense and cell senescence gene variants contribute to the risk of idiopathic pulmonary fibrosis. Am. J. Respir. Crit. Care Med 200, 199–208. [PubMed: 31034279]

59. Miller JE, Metpally RP, Person TN, Krishnamurthy S, Dasari VR, Shivakumar M, Lavage DR, Cook AM, Carey DJ, Ritchie MD, et al. ; DiscovEHR collaboration (2019). Systematic characterization of germline variants from the DiscovEHR study endometrial carcinoma population. BMC Med. Genomics 12, 59. [PubMed: 31053132]

60. Jiang X, Zhang B, Zhao J, Xu Y, Han H, Su K, Tao J, Fan R, Zhao X, Li L, and Li MD (2019). Identification and characterization of SEC24D as a susceptibility gene for hepatitis B virus infection. Sci. Rep 9, 13425. [PubMed: 31530870]

61. Lieberman S, Beeri R, Walsh T, Schechter M, Keret D, Half E, Gulsuner S, Tomer A, Jacob H, Cohen S, et al. (2019). Variable features of juvenile polyposis syndrome with gastric involvement among patients with a large genomic deletion of BMPR1A. Clin. Transl. Gastroenterol 10, e00054. [PubMed: 31259752]

62. Musolf AM, Ho WSC, Long KA, Zhuang Z, Argersinger DP, Sun H, Moiz BA, Simpson CL, Mendelevich EG, Bogdanov EI, et al. (2019). Small posterior fossa in Chiari I malformation affected families is significantly linked to 1q43–44 and 12q23–24.11 using whole exome sequencing. Eur. J. Hum. Genet 27, 1599–1610. [PubMed: 31227808]

63. Moawia A, Shaheen R, Rasool S, Waseem SS, Ewida N, Budde B, Kawalia A, Motameny S, Khan K, Fatima A, et al. (2017). Mutations of KIF14 cause primary microcephaly by impairing cytokinesis. Ann. Neurol 82, 562–577. [PubMed: 28892560]

64. Dinckan N, Du R, Petty LE, Coban-Akdemir Z, Jhangiani SN, Paine I, Baugh EH, Erdem AP, Kayserili H, Doddapaneni H, et al. (2018). Whole-exome sequencing identifies novel variants for tooth agenesis. J. Dent. Res 97, 49–59. [PubMed: 28813618]

65. Di Rocco M, Rusmini M, Caroli F, Madeo A, Bertamino M, Marre-Brunenghi G, and Ceccherini I (2018). Novel spondyloepimetaphyseal dysplasia due to *UFSP2* gene mutation. Clin. Genet 93, 671–674. [PubMed: 28892125]

66. Dapas M, Sisk R, Legro RS, Urbanek M, Dunaif A, and Hayes MG (2019). Family-based quantitative trait meta-analysis implicates rare noncoding variants in DENND1A in polycystic ovary syndrome. J. Clin. Endocrinol. Metab 104, 3835–3850.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Box 1.

Recommendations for sequencing-based GWAS reporting standards

Our recommendations for the development and adoption of reporting standards to increase the availability, accessibility, and utility of sequencing-based GWASs. The GWAS Catalog will support deposition of these datasets and promote adoption of these standards as well as continued discussions to reach consensus on the reporting of aggregate analyses.

1. WGS and WES association studies be referred to as “sequencing-based GWASs” (seqGWAS)
2. Single-variant analysis summary statistics be
   - **a.** Reported using the same standards as proposed for single-variant array-based GWASs[11,50]
   - **b.** Shared openly by submission to the GWAS Catalog
3. Aggregate analyses:
   - **a.** Metadata be reported to enable interpretation and aid reproducibility including
     - **i.** Sufficient details of the statistical test to allow replication of results
     - **ii.** Minor allele frequency thresholds used
     - **iii.** Details of tools used for functional annotation/consequence prediction (e.g., VEP release 103) and ontology terms used to describe the consequence (e.g., Sequence Ontology)
   - **b.** Community reaches consensus for standard content and format for reporting of aggregate seqGWAS summary statistics. This should include
     - **i.** The full list of qualifying variants contributing to each test
     - **ii.** Chromosomal coordinates of aggregation units (including genome assembly builds or gene annotation release version, e.g., GENCODE release 37, GRCh38)
     - **iii.** A standard identifier for the aggregation unit, e.g., HGNC gene name or symbol (if applicable)
     - **iv.** p value
4. SeqGWAS studies be conducted in populations that include more diverse ancestries

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Highlights

- Recommendations for increasing FAIRness of sequencing-based GWASs
- To be findable, we recommend standard terminology of sequencing-based GWAS (seqGWAS)
- To improve access and standards, the GWAS Catalog will support deposition of seqGWAS
- To improve utility, we recommend reporting standards for single and aggregate analyses

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

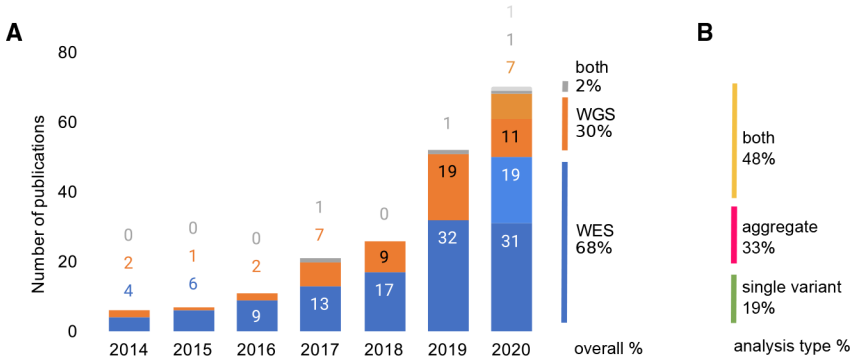


Figure 1. Sequencing-based GWAS publications, numbers, sequencing coverage, and analysis types

(A) Number of sequencing-based association publications identified per year from 2014 to September both 2020, n = 167. Only genome-wide (and not limited to specific regions or subsets of genes) and population-based studies are included (see STAR Methods for more information). The final quarter of 2020 is projected based on the rate of growth in the final quarter of 2019 (projected data are presented in the light shade of each color)

(B) The analysis types included in those publications. “Aggregate” refers to multi-variant analyses.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

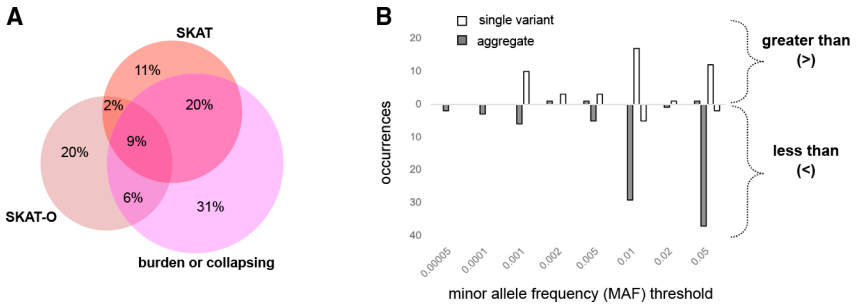


Figure 2. Statistical analysis methods used in sequencing-based GWAS publications

(A) Overlap among methods used in aggregate-analysis publications. Of 65 publications that use either SKAT, SKAT-O, or a burden test, 40% use at least two methods. Text related to study design was extracted by experienced curators and searched for the terms “SKAT,” “SKAT-O,” and “burden” or “collaps*” (where * refers to a wildcard for searching).

(B) Minor allele frequency thresholds used in single-variant and aggregate analyses. “Greater than or equal to” thresholds are displayed above the x axis; “less than or equal to” thresholds are displayed below the x axis. Thresholds were extracted from publications in which one or two thresholds were provided (single variant: n = 53 thresholds from 51 publications; aggregate: n = 86 thresholds from 77 publications). See Figure S4 for additional details on MAF-threshold reporting.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

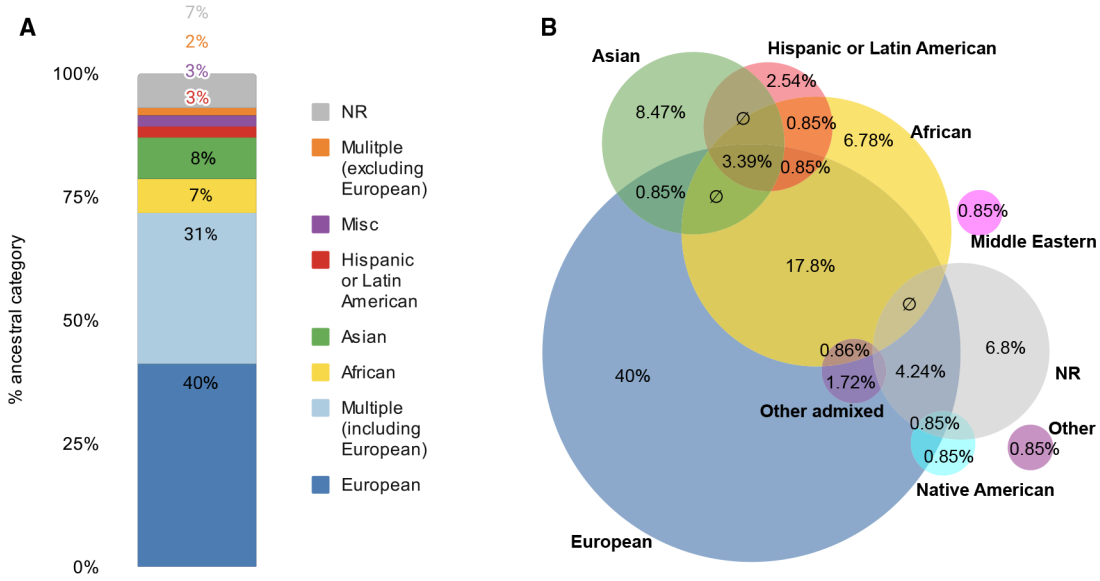


Figure 3. Ancestry of individuals used in sequencing-based GWAS publications

Publication-level breakdown of the broad ancestry categories, defined per the GWAS Catalog ancestry framework.[36] Some categories are collapsed for ease of display, analysis is based on 2014–2019 publications, n = 120.

(A) Overview of the percentage of publications that included only one or multiple ancestral categories.

(B) The proportion of publications that included the specified broad ancestral category. Overlaps indicate multiple ancestries were included in one publication; ∅ indicates an empty set. Venn diagram was created using DeepVenn.[38] Note that Venn diagrams of this size cannot be fully proportional (see Figure S7 and Table S5 for full data).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1.

Availability of summary statistics and number of statistical tests performed in sequencing versus array-based GWASs

	Single-variant array, % (n)	Single-variant sequencing, % (n)	Aggregate sequencing, % (n)
Summary statistics available without restriction	12 (300)	5 (4)	7 (7)
Number of tests (reporting)			
Reported	91 (5,817)	74 (61)	81 (84)
Not reported	9 (610)	26 (21)	19 (20)
Number of tests (distribution)	overall	overall	overall
Minimum	12,033	26,011	339
Q1	899,892	144,477	16,788
Median	5,554,549	548,889	18,665
Q3	9,334,585	8,752,596	20,843
Maximum	90,000,000	32,503,121	129,820,320
		WES only	WES only
Minimum	–	26,011	735
Q1	–	81,843	16,751
Median	–	158,091	18,360
Q3	–	235,133	20,000
Maximum	–	1,810,198	88,183
		WGS only	WGS only
Minimum	–	658,234	339
Q1	–	7,666,134	19,903
Median	–	12,210,410	32,316
Q3	–	29,880,479	1,082,577
Maximum	–	32,503,121	129,820,320

Publications that state that they share summary statistics openly (not including those provided with restricted access). Reported/not reported refers to whether the number of statistical tests performed was detailed in the publication. The number of statistical tests performed in sequencing-based studies is based on publications that provide one “number of statistical tests” (n = 51 of 79 for single-variant analysis, n = 56 of 101 for aggregate analysis). Publications that provide a range of statistical test numbers performed are included in the “reported” category but are not included in the distribution. The data for array-based GWAS were obtained from 2014–2019 studies in the GWAS Catalog (December 2, 2020 release) (see STAR Methods).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Text analysis tool (MonkeyLearn)		https://monkeylearn.com/word-cloud/
GWAS Catalog machine learning-based literature search	Lee et al.[56]	N/A
Literature search engine, EuropePMC		http://europepmc.org
PubMed		https://pubmed.ncbi.nlm.nih.gov
Other		
Literature (primary research journal articles)	Peer reviewed journals	PubMed IDs listed in Table S4
Publicly available curated meta-data	NHGRI-EBI GWAS Catalog	https://www.ebi.ac.uk/gwas/