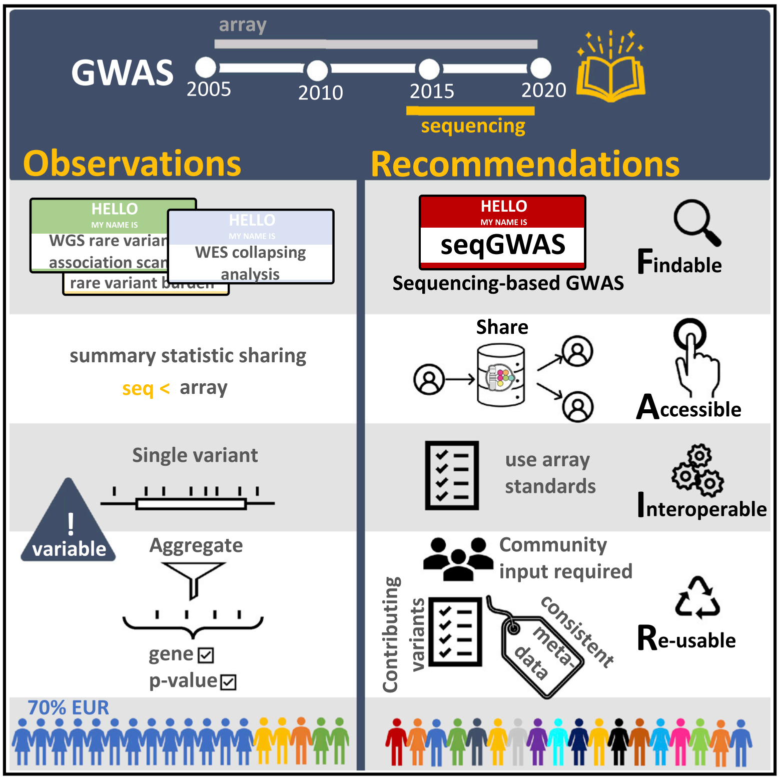
Cell Genomics

Short article

Sequencing-based genome-wide association studies reporting standards

Graphical abstract



Authors

Aoife McMahon, Elizabeth Lewis,
Annalisa Buniello, ..., Lucia A. Hindorff,
Laura W. Harris,
Jacqueline A.L. MacArthur

Correspondence

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

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.

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

McMahon et al., 2021, Cell Genomics *1*, 100005
October 13, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.xgen.2021.100005

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Short article

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

https://doi.org/10.1016/j.xgen.2021.100005

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

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 < 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

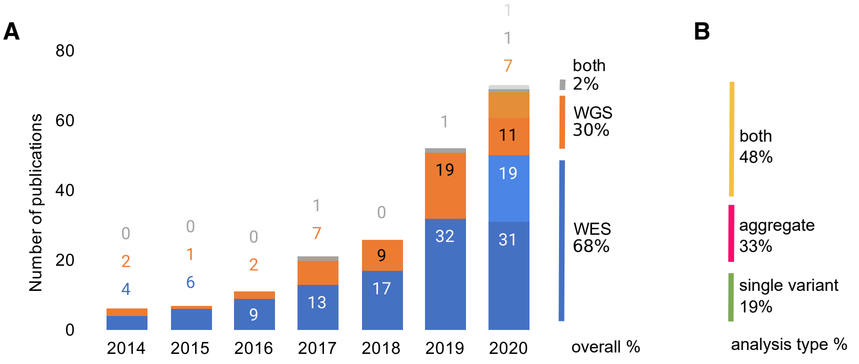


Figure 1. Sequencing-based GWAS publications, numbers, sequencing coverage, and analysis types

(A) Number of sequencing-based association publications identified per year from 2014 to September 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.

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]

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 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

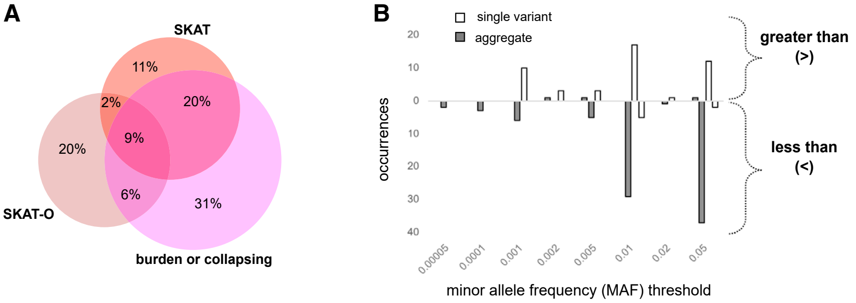


Figure 2. Statistical analysis methods used in sequencing-based GWAS publications

(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.

"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 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

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).

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 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

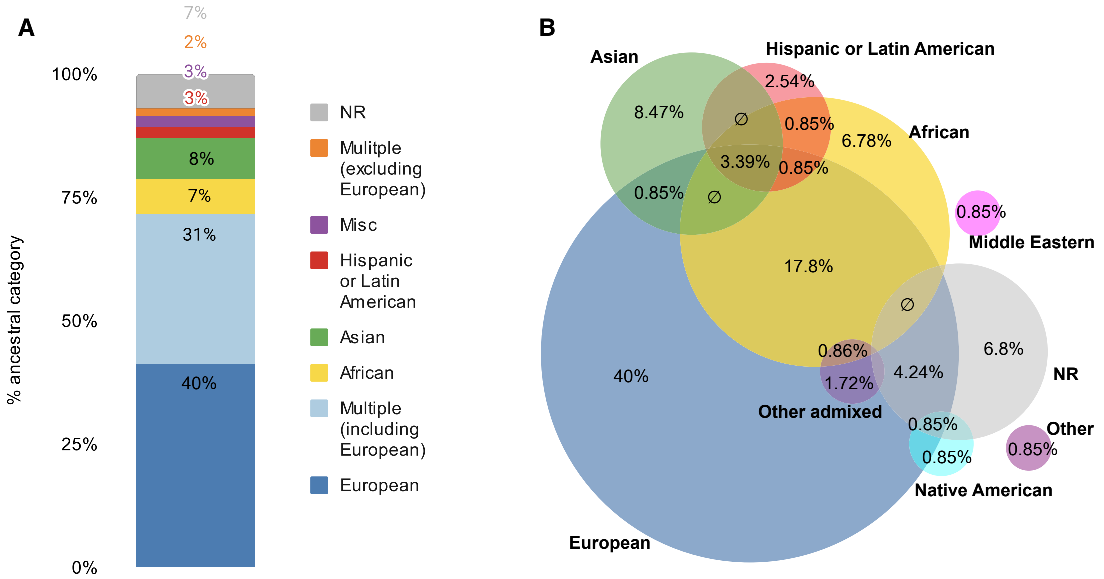


Figure 3. Ancestry of individuals used in sequencing-based GWAS publications

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).

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 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

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

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 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 "findability" 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 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:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

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

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.

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.

Received: March 24, 2021
Revised: June 28, 2021
Accepted: July 28, 2021
Published: October 13, 2021

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

REFERENCES

1. Klein, R.J., Xu, X., Mukherjee, S., Willis, J., and Hayes, J. (2010). Successes of genome-wide association studies. Cell *142*, 350–351, author reply 353–355.
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.
3. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.
4. Pasaniuc, B., Rohland, N., McLaren, P.J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B.M., Daly, M.J., Sklar, P., et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet. *44*, 631–635.
5. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.
6. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA *109*, 1193–1198.
7. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet. *100*, 635–649.
8. Lachance, J., and Tishkoff, S.A. (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. BioEssays *35*, 780–786.
9. Kim, M.S., Patel, K.P., Teng, A.K., Berens, A.J., and Lachance, J. (2018). Genetic disease risks can be misestimated across global populations. Genome Biol. *19*, 179.
10. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. *95*, 5–23.
11. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., 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.
12. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS discovery: Biology, function, and translation. Am. J. Hum. Genet. *101*, 5–22.
13. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E.,

et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data *3*, 160018.

14. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

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

16. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

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

18. Yu, B., de Vries, P.S., Metcalf, G.A., Wang, Z., Feofanova, E.V., Liu, X., Muzny, D.M., Wagenknecht, L.E., Gibbs, R.A., Morrison, A.C., and Boerwinkle, E. (2016). Whole genome sequence analysis of serum amino acid levels. Genome Biol. *17*, 237.

19. Kim, D., Basile, A.O., Bang, L., Horgusluoglu, E., Lee, S., Ritchie, M.D., Saykin, A.J., 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.

20. Pujar, S., O'Leary, N.A., Farrell, C.M., Loveland, J.E., Mudge, J.M., Wallin, C., Girón, C.G., 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.

21. de Vries, P.S., Yu, B., Feofanova, E.V., Metcalf, G.A., Brown, M.R., Zeighami, A.L., Liu, X., Muzny, D.M., Gibbs, R.A., Boerwinkle, E., and Morrison, A.C. (2017). Whole-genome sequencing study of serum peptide levels: The Atherosclerosis Risk in Communities study. Hum. Mol. Genet. *26*, 3442–3450.

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

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.

24. Sarnowski, C., Satizabal, C.L., DeCarli, C., Pitsillides, A.N., Cupples, L.A., Vasan, R.S., Wilson, J.G., Bis, J.C., Fornage, M., Beiser, A.S., 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.

25. Mak, A.C.Y., White, M.J., Eckalbar, W.L., Szpiech, Z.A., Oh, S.S., Pino-Yanes, M., Hu, D., Goddard, P., Huntsman, S., Galanter, J., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2018). Whole-genome sequencing of pharmacogenetic drug response in racially diverse children with asthma. Am. J. Respir. Crit. Care Med. *197*, 1552–1564.

26. Sapkota, Y., Cheung, Y.T., Moon, W., Shelton, K., Wilson, C.L., Wang, Z., Mulrooney, D.A., Zhang, J., Armstrong, G.T., Hudson, M.M., 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.

27. Monson, E.T., Pirooznia, M., Parla, J., Kramer, M., Goes, F.S., Gaine, M.E., Gaynor, S.C., 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.

28. Gratten, J., Zhao, Q., Benyamin, B., Garton, F., He, J., Leo, P.J., Mangelsdorf, M., Anderson, L., Zhang, Z.H., Chen, L., et al. (2017). Whole-exome sequencing in amyotrophic lateral sclerosis suggests NEK1 is a risk gene in Chinese. Genome Med. *9*, 97.

29. Scott, W.K., Medie, F.M., Ruffin, F., Sharma-Kuinkel, B.K., Cyr, D.D., Guo, S., Dykxhoorn, D.M., Skov, R.L., Bruun, N.E., Dahl, A., et al. (2018). Human genetic variation in GLS2 is associated with development of complicated *Staphylococcus aureus* bacteremia. PLoS Genet. *14*, e1007667.

30. Povysil, G., Petrovski, S., Hostyk, J., Aggarwal, V., Allen, A.S., and Goldstein, D.B. (2019). Rare-variant collapsing analyses for complex traits: Guidelines and applications. Nat. Rev. Genet. *20*, 747–759.

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.

32. Wolock, C.J., Stong, N., Ma, C.J., Nagasaki, T., Lee, W., Tsang, S.H., Kamalakaran, S., Goldstein, D.B., 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.

33. Alkelai, A., Greenbaum, L., Heinzen, E.L., Baugh, E.H., Teitelbaum, A., Zhu, X., Strous, R.D., Tatarskyy, P., Zai, C.C., Tiwari, A.K., et al. (2019). New insights into tardive dyskinesia genetics: Implementation of whole-exome sequencing approach. Prog. Neuropsychopharmacol. Biol. Psychiatry *94*, 109659.

34. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.

35. Peterson, R.E., Kuchenbaecker, K., Walters, R.K., Chen, C.Y., Popejoy, A.B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R.J., Brick, L., et al. (2019). Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. Cell *179*, 589–603.

36. Morales, J., Welter, D., Bowler, E.H., Cerezo, M., Harris, L.W., McMahon, A.C., Hall, P., Junkins, H.A., 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.

37. Mills, M.C., and Rahal, C. (2019). A scientometric review of genome-wide association studies. Commun. Biol. *2*, 9.

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.

39. Sabo, A., Mishra, P., Dugan-Perez, S., Voruganti, V.S., Kent, J.W., Jr., Kalra, D., Cole, S.A., Comuzzie, A.G., Muzny, D.M., Gibbs, R.A., and Butte, N.F. (2017). Exome sequencing reveals novel genetic loci influencing obesity-related traits in Hispanic children. Obesity (Silver Spring) *25*, 1270–1276.

40. Locke, A.E., Steinberg, K.M., Chiang, C.W.K., Service, S.K., Havulinna, A.S., Stell, L., Pirinen, M., Abel, H.J., Chiang, C.C., Fulton, R.S., et al.; FinnGen Project (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. Nature *572*, 323–328.

41. Höglund, J., Rafati, N., Rask-Andersen, M., Enroth, S., Karlsson, T., Ek, W.E., 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.

42. Jiang, L., Zheng, Z., Qi, T., Kemper, K.E., Wray, N.R., Visscher, P.M., and Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. Nat. Genet. *51*, 1749–1755.

43. Cirulli, E.T., White, S., Read, R.W., Elhanan, G., Metcalf, W.J., Tanudjaja, F., Fath, D.M., Sandoval, E., Isaksson, M., Schlauch, K.A., et al. (2020). Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat. Commun. *11*, 542.

44. Zhao, Z., Bi, W., Zhou, W., VandeHaar, P., Fritsche, L.G., 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.

45. Long, T., Hicks, M., Yu, H.C., Biggs, W.H., Kirkness, E.F., Menni, C., Zierer, J., Small, K.S., 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.

46. Hammer, M.F., 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.

47. Grant, R.C., Denroche, R.E., Borgida, A., Virtanen, C., Cook, N., Smith, A.L., Connor, A.A., Wilson, J.M., Peterson, G., Roberts, N.J., et al. (2018). Exome-wide association study of pancreatic cancer risk. Gastroenterology *154*, 719–722.e3.

48. Kwak, S.H., Chae, J., Lee, S., Choi, S., Koo, B.K., Yoon, J.W., Park, J.H., Cho, B., Moon, M.K., 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.

49. Sveinbjornsson, G., Olafsdottir, E.F., Thorolfsdottir, R.B., Davidsson, O.B., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., Bjornsson, E., Jensson, B.O., Arnadottir, G.A., et al. (2018). Variants in NKX2-5 and FLNC Cause Dilated Cardiomyopathy and Sudden Cardiac Death. Circ. Genom. Precis. Med. *11*, e002151.

50. MacArthur, J.A.L., Buniello, A., Harris, L.W., Hayhurst, J., McMahon, A., Sollis, E., Cerezo, M., Hall, P., Lewis, E., Whetzel, P.L., et al. (2021). Workshop proceedings—GWAS summary statistics standards and sharing. Cell Genomics *18*, 100004-1–100004-8.

51. Lappalainen, T., Scott, A.J., Brandt, M., and Hall, I.M. (2019). Genomic analysis in the age of human genome sequencing. Cell *177*, 70–84.

52. Wright, C.F., Ware, J.S., Lucassen, A.M., Hall, A., Middleton, A., Rahman, N., Ellard, S., and Firth, H.V. (2019). Genomic variant sharing: A position statement. Wellcome Open Res. *4*, 22.

53. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., 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.

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.

55. Eilbeck, K., Lewis, S.E., Mungall, C.J., 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.

56. Lee, K., Famiglietti, M.L., McMahon, A., Wei, C.H., MacArthur, J.A.L., 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.

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.

58. Moore, C., Blumhagen, R.Z., Yang, I.V., 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.

59. Miller, J.E., Metpally, R.P., Person, T.N., Krishnamurthy, S., Dasari, V.R., Shivakumar, M., Lavage, D.R., Cook, A.M., Carey, D.J., Ritchie, M.D., et al.; DiscovEHR collaboration (2019). Systematic characterization of germline variants from the DiscovEHR study endometrial carcinoma population. BMC Med. Genomics *12*, 59.

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

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.

62. Musolf, A.M., Ho, W.S.C., Long, K.A., Zhuang, Z., Argersinger, D.P., Sun, H., Moiz, B.A., Simpson, C.L., Mendelevich, E.G., Bogdanov, E.I., 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.

63. Moawia, A., Shaheen, R., Rasool, S., Waseem, S.S., 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.

64. Dinckan, N., Du, R., Petty, L.E., Coban-Akdemir, Z., Jhangiani, S.N., Paine, I., Baugh, E.H., Erdem, A.P., Kayserili, H., Doddapaneni, H., et al. (2018). Whole-exome sequencing identifies novel variants for tooth agenesis. J. Dent. Res. *97*, 49–59.

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.

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

STAR★METHODS

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/

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 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.

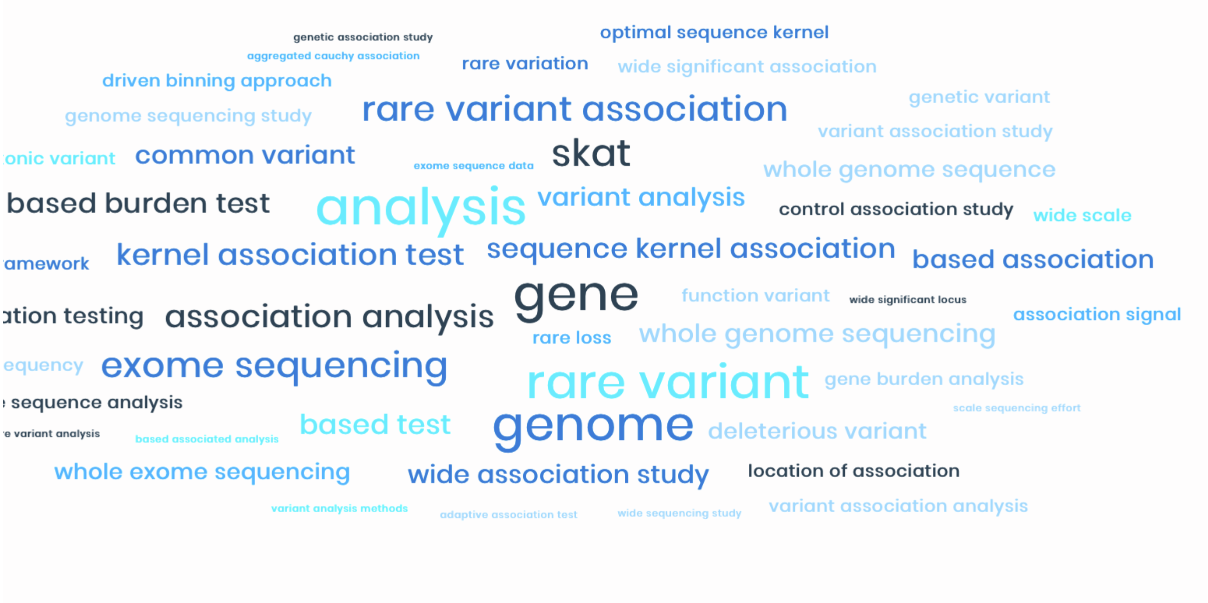
Cell Genomics, Volume *1*

Supplemental information

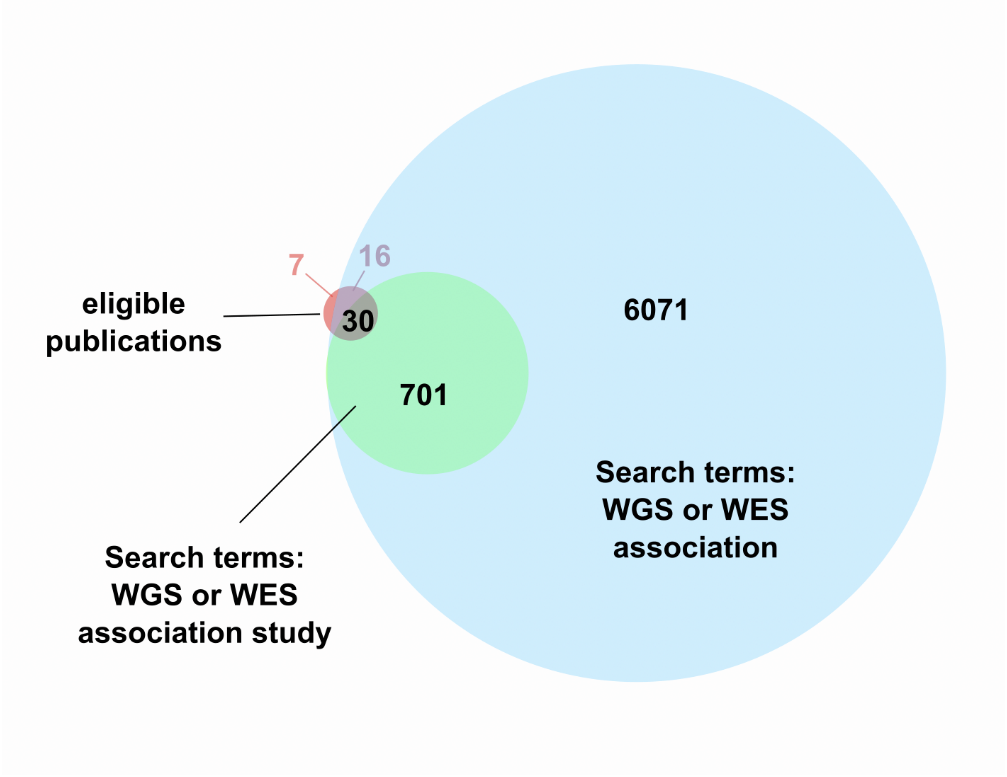
Sequencing-based genome-wide association studies reporting standards

Aoife McMahon, Elizabeth Lewis, Annalisa Buniello, Maria Cerezo, Peggy Hall, Elliot Sollis, Helen Parkinson, Lucia A. Hindorff, Laura W. Harris, and Jacqueline A.L. MacArthur

Supplemental Figures



Supplemental Figure 1
Word cloud from abstract and title text related to the study design of publications (including aggregate and single analysis) (related to Figure 1).
Word size corresponds to frequency x relevance metric (inverse document frequency of the term in an unrelated corpus of text), the top 50 enriched terms are displayed.

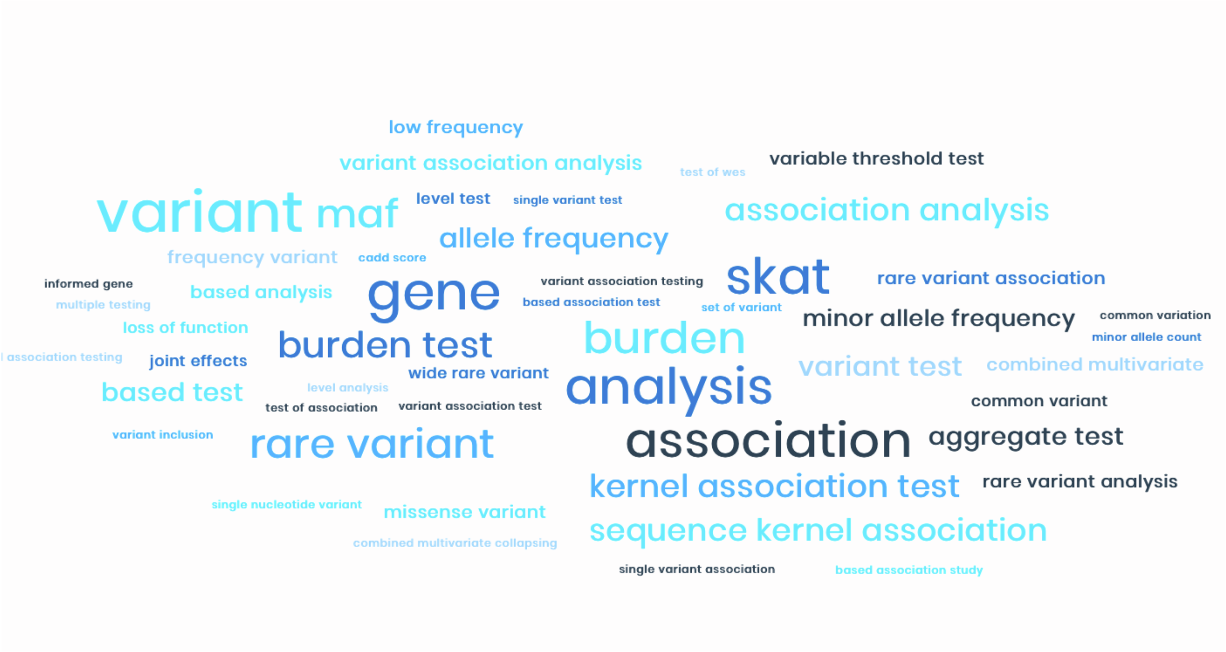

7
16
eligible publications
30
6071
701
Search terms: WGS or WES association
Search terms: WGS or WES association study

Supplemental Figure 2

An illustration of the difficulty in ascertainment of sequencing-based GWAS publications (Related to STAR methods).

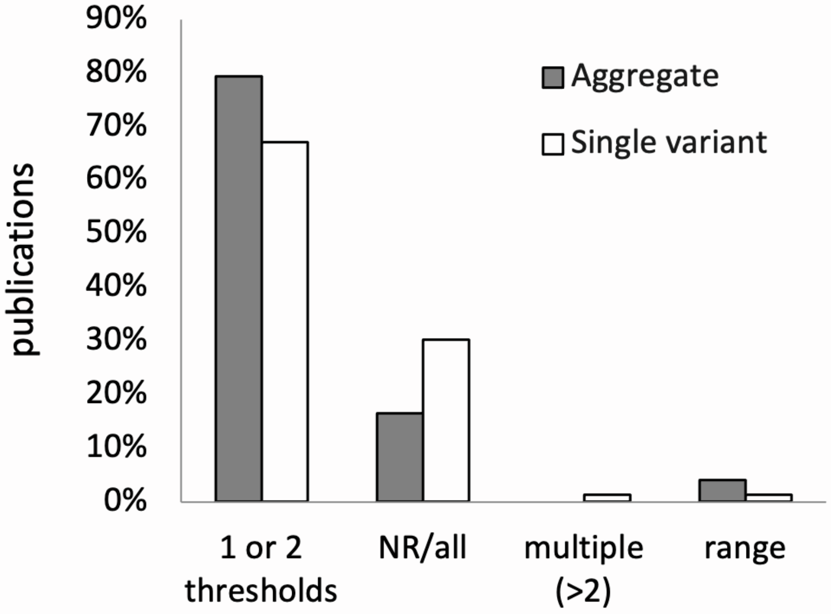
The overlap between the list of eligible 2019 publications with the search results of permissive query searches conducted using a literature search engine (EuropePMC). This analysis is limited to 2019. The labels shown on the diagram represent specific search terms used in EuropePMC. Label (WGS or WES association) = Query (("WGS" AND "association") OR ("whole genome sequencing" AND "association") OR ("WES" AND "association") OR ("whole exome sequencing" AND "association")) AND (FIRST_PDATE:[2019-01-01 TO 2019-12-31])

Label (WGS or WES association study) = Query (("WGS" AND "association study") OR ("whole genome sequencing" AND "association study") OR ("WES" AND "association study") OR ("whole exome sequencing" AND "association study")) AND (FIRST_PDATE:[2019-01-01 TO 2019-12-31])



Supplemental Figure 3

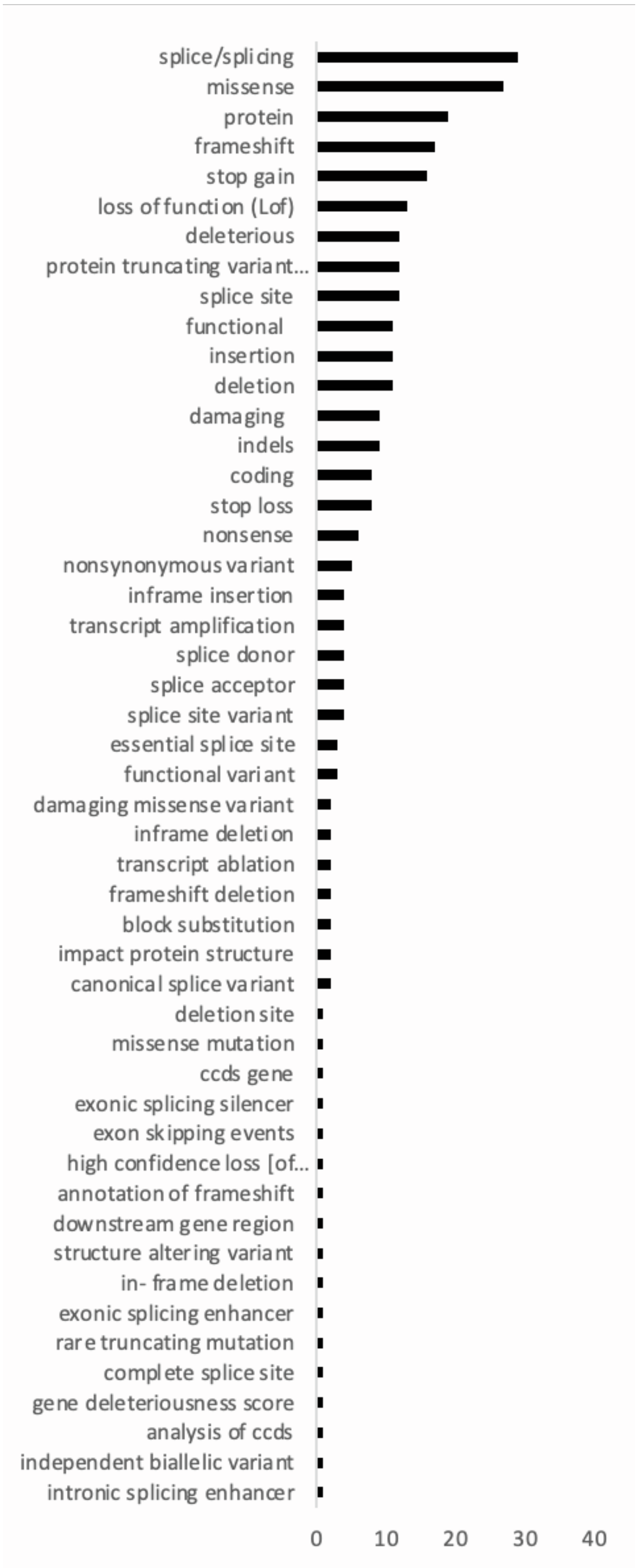
Word cloud from text related to the study design of publications which perform aggregate analysis, from sections other than abstract and title (related to Figure 2). Word size corresponds to frequency x relevance metric (frequency-inverse document frequency (TD-IDF), the top 50 enriched terms displayed.



Supplemental Figure 4

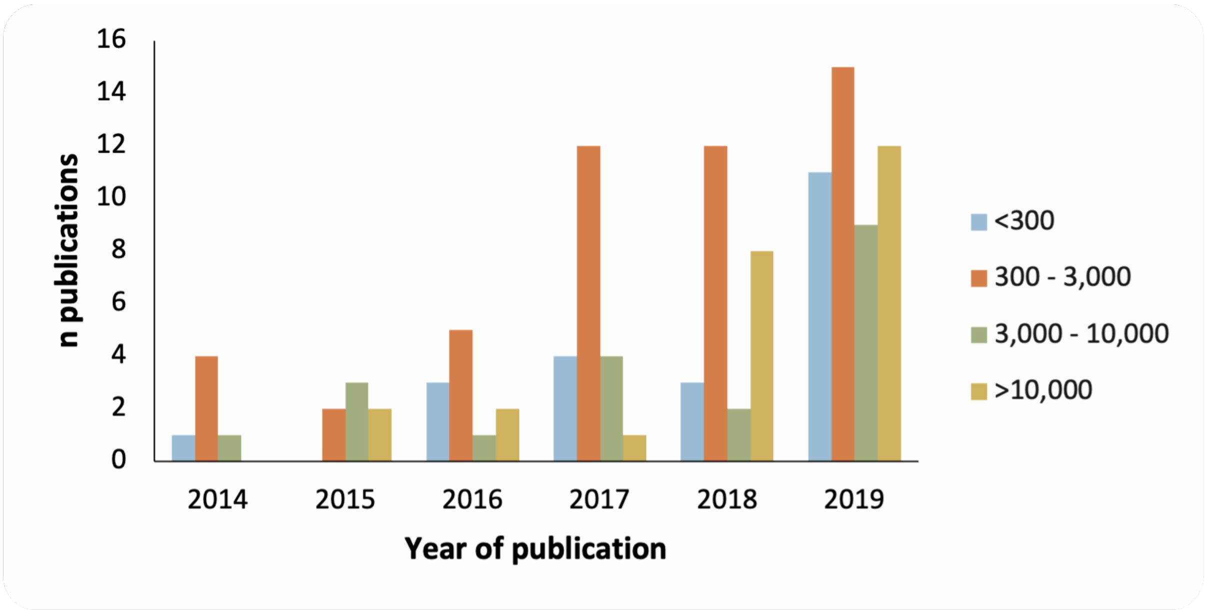
Reporting of minor allele frequency thresholds in single variant and aggregate analyses (related to Figure 2).

This figure shows how MAF was reported in publications (2014-2019) (single variant: n publications = 97, aggregate: n publications = 76). Data in Figure 2 are derived from those publications that report one or two thresholds. 'NR/all' includes publications that provided no information on thresholds as well as publications which implied that all variants were included.' 'Range' represents publications that included variants within a specific range of MAFs e.g. 1-5%. NR = not reported.



Occurrences of terms in text related to variant selection

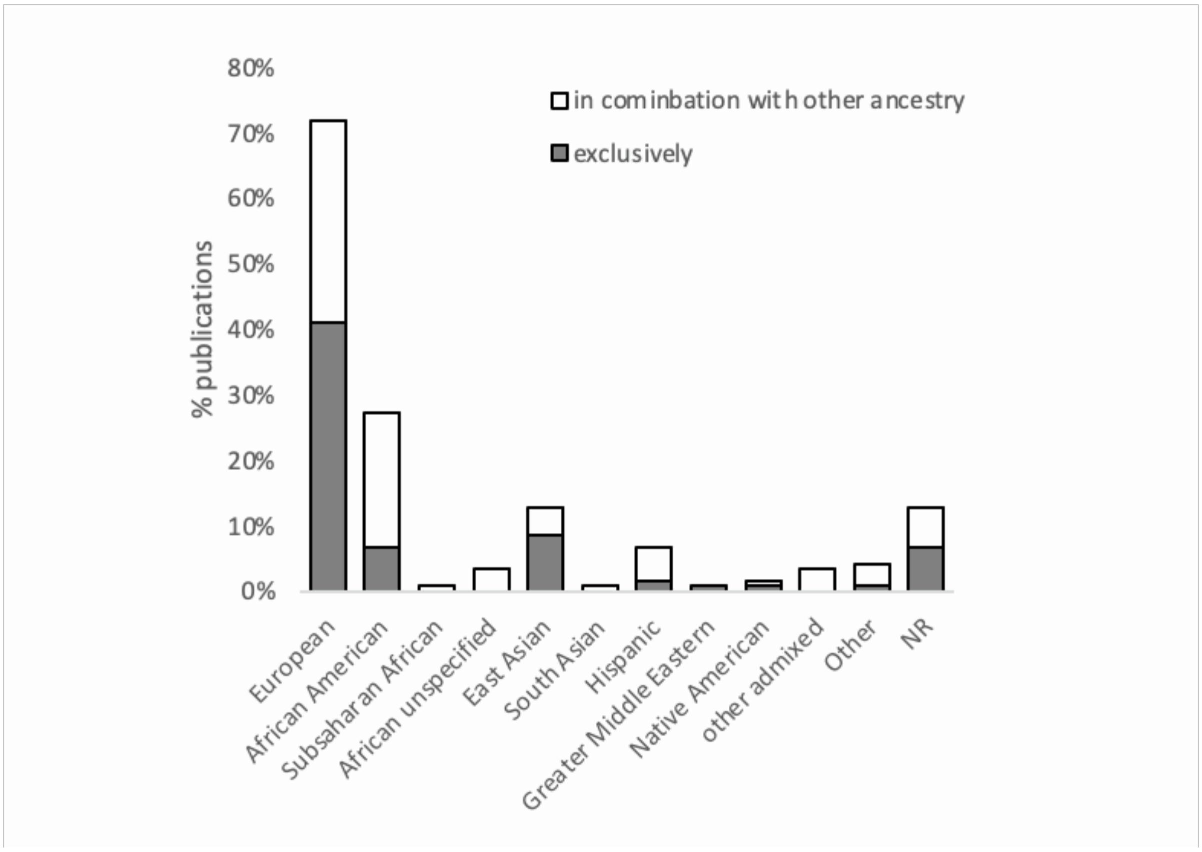
Supplemental Figure 5
Language describing variant types (related to Figure 2)



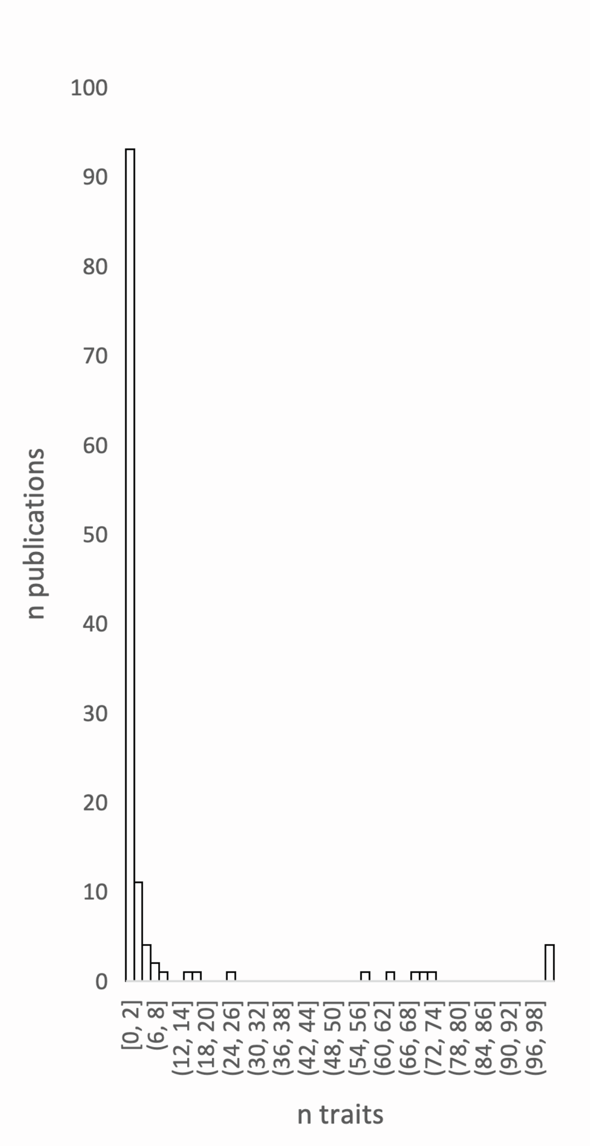
Supplemental Figure 6

Sample size bins in sequencing-based association studies (related to Figure 3).

Publication level sample sizes were classified into brackets of <300, 300-3000, 3000-10,000 or >10,000 individuals. Number of publications in each sample size bracket, by year.



Supplemental Figure 7
Detail on percentage of publications including ancestries exclusively or in combination with other ancestries (related to Figure 3).



Supplemental Figure 8
Number of traits analysed in sequence-based association publications (related to STAR Methods).
Distribution of number of traits analysed per publication. The final overflow bar represents >100 traits (publications of 644, 791 and 2048 traits).

SUPPLEMENTAL TABLES

Publication identification sources
GWAS Catalog machine learning search:
see Lee et al[3]
Pubmed and EuropePMC text query searches:
sequencing association genome OR exome
sequencing association genetic
rare variant whole genome whole exome
rare variant association analysis
gwas sequencing population
genetic association studies"[MeSH Terms] AND "high-throughput nucleotide sequencing"[MeSH Terms] NOT Review[ptyp]
("whole genome sequencing" OR "whole exome sequencing") AND (METHODS:"skat-o" OR METHODS:"gene-based" OR METHODS:"single variant" OR METHODS:"burden")
("whole exome sequencing"[MeSH Terms] OR ("whole"[All Fields] AND "exome"[All Fields] AND "sequencing"[All Fields]) OR "whole exome sequencing"[All Fields]) AND ("association"[MeSH Terms] OR "association"[All Fields])
("rare variant association" AND "whole genome sequencing" OR "whole exome sequencing")
("gene-based" OR "collapsing analysis" AND "whole exome sequencing" OR "whole genome sequencing")
"skat-o" AND "sequencing"
"rare variant" sequencing association whole-genome
Cohort/project website:
TopMEd publications list (https://www.nhlbiwgs.org/publications)
NIH project reporter (https://projectreporter.nih.gov)
UKBiobank list (in house)
Open Targets list (in house)
References
Twitter
Conferences
GWAS Catalog author summary statistics submission
Personal communication

Supplemental Table 1
Sources where sequencing-based GWAS were identified (related to STAR Methods)

	Curated meta-data	**2014-2019**	**2020 + 2019/2020 preprints**
	n:	120	40 + 7
Study design:	*Coverage (WGS/WES)*	✓	✓
	Analysis type (single/aggregate)	✓	✓
	Number of statistical tests (range or number)	✓	✗
	Minor allele frequency thresholds (if >2 provided extracted as 'multiple', if 2 extracted both)	✓	✗
	Reference genome	✓	✗
	Terms related to study design (abstract/title, elsewhere)	✓	✗
	Terms related to qualifying variants	✓	✗
Sample:	*Sample size category (<300=0, <3000=1, <10,000=2, >10,000=3)*	✓	✗
	Broad ancestral category	✓	✗
	Additional ancestry descriptor	✓	✗
	Country of recruitment (if ancestral category NR)	✓	✗
	Consortium/Cohort	✓	✗
Traits:	*Number analysed*	✓	✗
	Reported trait	✓	✗
	Mapped trait EFO name	✓	✗
	Mapped trait EFO ID	✓	✗
	Background trait EFO name	✓	✗
	Background trait EFO ID	✓	✗
Data availability:	**Summary statistics**		
	Single variant/aggregate; freely, restricted, partial or no.	✓	✗
	Location	✓	✗
	Sequence data		
	In restricted repository or no	✓	✗
	Location	✓	✗
	Accession ID	✓	✗

Supplemental Table 2 Overview of publication meta-data extracted (related to STAR Methods). All curated meta-data is included in Supp. Table 4.

Selected examples of variant filtering descriptions (annotation/function)
non-synonymous
putative damaging
nonsynonymous and otherwise presumed functional
variants that are most likely to affect a protein's function, that is, non-synonymous, stop gain, stop loss, frameshift deletions and insertions, and splice site variants.;
nonsynonymous and splice-site variants
“qualifying” variants; 1) all non-synonymous and canonical splice variants (coding model), 2) all non- synonymous coding variants except those predicted by PolyPhen-2 HumVar(13) to be benign (not benign model), and 3) only stop gain, frameshift and canonical splice variants (loss-of-function [LoF] model).
All aggregation tests utilized only variants that were rare (defined as MAF<5% in the population set) and either truncating (frameshift, essential splice site, nonsense) or missense and predicted to be deleterious (by at least one of Polyphen, SIFT, or Condel) as annotated by Variant Effect Predictor (VEP) release 74. The analysis of rare truncating mutations, however, only included variants annotated as nonsense (SNVs only), essential splice site (SNVs/indels), or frameshift (indels only). (multi)
ultrarare, deleterious, nonsynonymous variants ; qualifying variants were restricted to indels and single-nucleotide variants annotated as having either a loss-of-function (LoF) effect, an in-frame indel, or a “probably damaging” missense prediction by Polymorphism Phenotyping version 2 (PolyPhen, HumDiv; http://genetics.bwh.harvard.edu/pph2/) (16). These analyses relied on the predicted effects of the LoF and missense annotated variants whose functions have not been individually confirmed in the laboratory. We subsequently performed analyses of CCDS genes using six alternative qualifying variant models as defined in Table E4, including an autosomal recessive model and a synonymous variant negative control model.
We defined qualifying variants in four ways (Table 1); ultra-rare variants; loss-of-function, inframe insertion or deletion, or a “probably damaging” missense effect by PolyPhen-2 (HumDiv); Three secondary analyses were performed to evaluate the contribution to epilepsy risk from: rare loss-of-function variants with an internal and external population MAF up to 0.1%; rare non-synonymous variation in the general population with an internal and external MAF up to 0.1%; and a presumed neutral model that imposed similar MAF thresholds as our primary analysis, but focused specifically on protein-coding variants predicted to have a synonymous effect.
deleterious - predicted by variant effect predictor (VEP) to have “HIGH” impact, cause protein loss-of-function (stop-gain, frameshift insertion and deletion [indel], etc.), or were missense mutations with a combined annotation dependent depletion (CADD)26 score >25
we considered six functional annotations, CADD [7], RegulomeDB [18], FunSeq [19], Funseq2 [20], GERP++ [21] and GenoSkyline [8]
loss of function (LoF) variants defined as follows were used for further analysis: stop gain/loss, coding INDELs, splice-site acceptors, and splice-site donors. We also included variants predicted as damaging according to their SIFT [23] score and a CAD [24] score of > 20.; gene score (a gene deleteriousness score) quantified the impact of damage of a gene, and was defined as the geometric mean of the SIFT scores for the multitude of deleterious variants in a gene.
Two sets of analyses were performed: The first included only frameshift (insertion/deletion/block substitution), stopgain, stoploss and splicing SNVs (jointly defined as loss-of-function (LOF) variants), while the sec- ond included all variants captured in the first analysis as well as non-synonymous SNVs and non-frameshift indels or block substitutions that were predicted to be probably dam- aging by Polyphen 2 and deleterious by SIFT [1, 62].
(1) PTVs at any allele frequency with VEP annotations: frameshift_variant, initiator_codon_variant, splice_acceptor_variant, splice_donor_variant, stop_lost, stop_gained; (2) PTVs included in (1) plus missense variants with MAF<0.1% scored as “damaging” or “deleterious” by all five functional prediction algorithms; (3) PTVs included in (1) plus missense variants with MAF<0.5% scored as “damaging” or “deleterious” by all five algorithms. (multi)

Supplemental Table 3

Examples of variant filtering descriptions provided by authors (related to Figure 2).

Terms in text related to variant selection.

Supplemental Table 4
Full curated meta-data from publications included in this analysis

Supplied as separate .xlsx file

Broad ancestral category	Overall % (n)		Exclusively % (n)		In combination with other ancestry % (n)	
European	71%	(85)	40%	(48)	31%	(37)
African American	28%	(33)	7%	(8)	21%	(25)
Subsaharan African	1%	(1)	0%	(0)	1%	(1)
African unspecified	3%	(4)	0%	(0)	3%	(4)
East Asian	13%	(15)	8%	(10)	4%	(5)
South Asian	1%	(1)	0%	(0)	1%	(1)
Hispanic	8%	(9)	3%	(3)	5%	(6)
Greater Middle Eastern	1%	(1)	1%	(1)	0%	(0)
Native American	2%	(2)	1%	(1)	1%	(1)
Other admixed	3%	(4)	0%	(0)	3%	(4)
Other	4%	(5)	1%	(1)	3%	(4)
NR	13%	(15)	7%	(8)	6%	(7)

Supplemental Table 5
Publication level breakdown of the broad ancestral categories of individuals, defined per the GWAS Catalog ancestry framework (related to Figure 3).
Overall = percentage of all publications that include an ancestry, either exclusively or in combination with other ancestries.

Individual level sequence data availability	%	N publications
Controlled access repository (accession ID provided)	19%	23
Controlled access repository (no ID provided)	2%	2
Partial dataset in repository	2%	2
Partial dataset in repository, partial available upon request	2%	2
Available upon request	3%	4
None	73%	90

Supplemental Table 6
Individual level sequence data availability (related to Table 1).
Analysis of author statements regarding individual level sequence data.

Cohort/consortium	Count
NR	24
TOPMed	15
ARIC	8
NHLBI GO ESP	8
JHS	6
Alzheimer Disease Sequencing Project (ADSP)	6
TwinsUK	6
UK10K	5
UKBiobank	5
FHS	4
FINRISK	3
ADNI	2
CHARGE	2
Estonian Biobank	2
GenTAC	2
HELIC-MANOLIS	2
IGM	2
Epi4K	2
Old Order Amish Study	2
ALSPAC	1
ARC	1
ARRA	1
AURORA	1
BDR	1
Boston Early-Onset COPD Study (EOCOPD)	1
CASPMI	1
CHS	1
CONVERGE	1
COPDGene	1
CUMC	1
deCODE	1
DiscovEHR	1
EGD	1
Emory	1
ENGAGE	1
EPGP	1
EPIC Potsdam	1
Epilepsy Phenome/Genome Project	1
Familial dyslipidemia	1
FinMetSeq	1
FinnDiane	1
Genetic Epidemiology of Asthma in Costa Rica	1
Genomic Translation for ALS Care (GTAC study)	1
Georgia Centenarian Study (GCS)	1
GOLDN	1
Health 2000	1
Healthy Nevada Project	1
iJGVD (controls)	1
International FTLD-TDP WGS Consortium	1
INTERVAL	1
IRASFS	1

IRCCS	1
KARE	1
MESA	1
METSIM	1
Minnesota Twin Family Study (MTFS)	1
Nottingham Smokers cohort	1
NSPHS	1
OPCS	1
PACA-AU	1
PAH biobank	1
PanCuRx	1
PDAY	1
PEACH	1
PREDICTION-ADR Consortium and EUDRAGENE	1
PROP	1
QPCS	1
RISK	1
ROSMAP	1
RS	1
SABG	1
SDR	1
SJLIFE	1
Steno Diabetes Center	1
T2D-GENES	1
TCGA	1

Supplemental Table 7
A count of the occurrence of cohort and consortium/project names in sequencing-based GWAS publications (related to Figure 3). This table does not distinguish between cohorts (e.g. Old Order Amish Study) or consortia/projects (e.g. TOPMed) because this distinction is typically not made by authors. All instances were extracted, for example ‘the JHS cohort sequencing by the TOPMed program’, is represented as one instance of JHS, and one instance of TOPMed.

Supplemental References

1. Morales, J. *et al.* A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 21 (2018).
2. Hulsen, T., de Vlieg, J. & Alkema, W. BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **9**, 488 (2008).
3. Lee, K. *et al.* Scaling up data curation using deep learning: An application to literature triage in genomic variation resources. *PLoS Comput. Biol.* **14**, e1006390 (2018).