IntuitionLabs

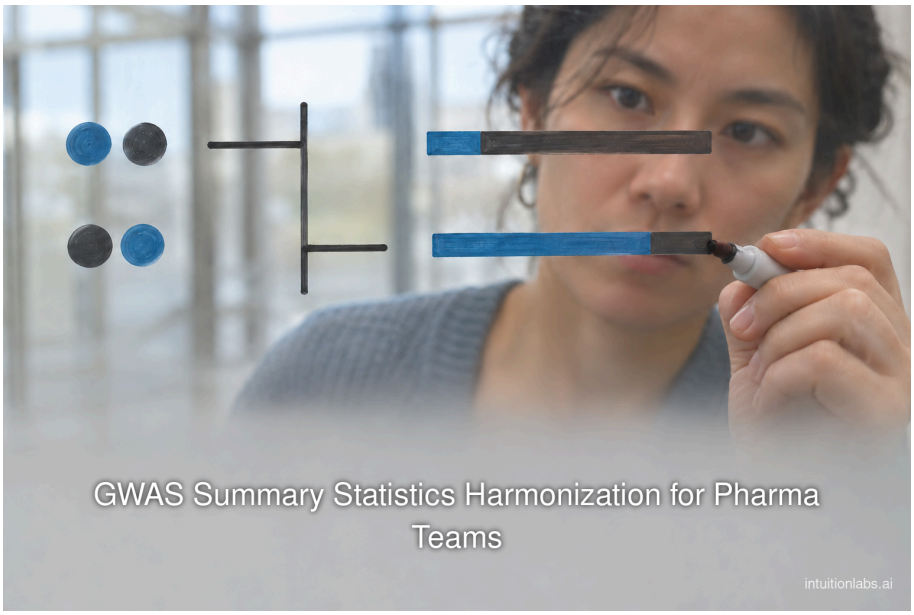
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

GWAS Summary Statistics Harmonization for Pharma Teams

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-11-20 · gwas summary statistics harmonization · gwas summary statistics · gwas-ssf · gwas catalog harmonised summary statistics · effect allele orientation · mendelian randomization drug target · genetic evidence drug discovery · summary statistics quality control

Original article: https://intuitionlabs.ai/articles/gwas-summary-statistics-harmonisation-pharma



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes in the Standards, Pipeline and Access Routes
4. Implementation Considerations and Process Changes
5. Data Analysis and Evidence
6. Implications and Future Directions
7. Frequently Asked Questions (FAQs)
8. Conclusion

Executive Summary

This report is a file-state and statistical-semantics acceptance reference for teams that download public GWAS (genome-wide association study) summary statistics and want to reuse them in colocalisation, Mendelian randomisation (MR) or target prioritisation. It does not teach the methods and it does not claim that a clean file proves causality. The decision it supports is narrower: whether a given file, in a given state, with a given sidecar, can enter an analysis, and which checks an ingest pipeline must preserve.

The GWAS Catalog separates curated, significant loci from full summary statistics, and as of 17 November 2026 it listed 239,508 full summary statistics alongside 1,206,717 top associations (www.ebi.ac.uk). Full summary statistics were downloadable for 66% of journal-published GWAS in the Catalog at the time of its 2024 paper (pmc.ncbi.nlm.nih.gov). Files exist in three states: the original submitted file, an optional legacy "formatted" file whose headers were standardised but whose data were not manipulated, and a harmonised file reported against one genome build with alleles orientated to the forward strand (pmc.ncbi.nlm.nih.gov).

Harmonisation is not neutral. When the effect and other alleles are swapped, the beta, odds ratio, confidence interval and effect allele frequency are adjusted (ebispot.github.io). In the one real study audited here (GCST90270927, harmonised in November 2023), 4,854,469 forward-strand variants (35.14% of the 13,815,944 mapped sites) carried the flipped code (ftp.ebi.ac.uk), and 661,936 mapped sites (4.79%) could not be harmonised, 661,899 of them because no matching reference variant existed (ftp.ebi.ac.uk). That is one file under one pipeline version, not a benchmark. The Catalog's paper gives an average of about 3.5% unharmonisable variants removed (pmc.ncbi.nlm.nih.gov), and the pipeline documentation's 95.29% example success rate is illustrative only (ebispot.github.io).

Three cautions shape the acceptance rules. First, the standard is moving: GWAS-SSF v1.1.0 added a z-score fallback on 26 June 2026 (github.com). Second, copy number variant (CNV) and gene-based data are not SNP data. The methods page says they should follow the beyond-SNPs standards since April 2026 (www.ebi.ac.uk), while the linked landing page calls those standards drafts that are not yet fully integrated into the submission system (www.ebi.ac.uk). Third, metadata checks do not resolve sample overlap, ancestry mismatch or linkage disequilibrium (LD) reference mismatch, which need scientific review (pmc.ncbi.nlm.nih.gov). Genetic support is informative for target selection, with one analysis estimating a 2.6-fold higher probability of success for supported mechanisms (pmc.ncbi.nlm.nih.gov), but that estimate concerns portfolios, not any single file.

Introduction and Background

Pharma statistical-genetics leads and target-discovery teams increasingly pull public GWAS summary statistics into internal evidence stores. The temptation is to treat a downloaded file as a finished input. In practice a GWAS Catalog file passes through several states, each with different guarantees about genome build, allele orientation and the rows that remain. A pipeline that cannot tell these states apart can silently reverse the direction of an effect, or compute a denominator on a file from which variants have been dropped.

The Catalog draws a clear line between its curated associations and the full data. Curated associations are the significant (p<1e-5) and independent loci reported in a paper (www.ebi.ac.uk), whereas summary statistics are the full, aggregate p-values and association data for every variant analysed in a study. Only the second is useful for colocalisation, which needs the whole region, and the coloc documentation asks users to include all SNPs in a region rather than trimming by significance or minor allele frequency (chr1swallace.github.io). Files come from several sources, including author submission, so provenance varies.

Why does this matter commercially? Human genetic evidence is one of the better-studied predictors of drug development outcomes. An early analysis reported that the share of mechanisms with direct genetic support rose from 2.0% at the preclinical stage to 8.2% among approved-drug mechanisms (www.nature.com). A 2024 update estimated that the probability of success is 2.6 times greater for genetically supported mechanisms (pmc.ncbi.nlm.nih.gov), and reported that this ratio is sensitive to confidence in the variant-to-gene mapping. Those are population-level associations across drug programmes. They are not evidence that any particular summary-statistics file is correct, and they depend on the upstream data quality examined here.

IntuitionLabs, a life-sciences and AI consultancy, prepared this report as an adjacent advisor on data architecture and scientific AI assurance. It is not a vendor of GWAS data, harmonisation software or any product discussed. Its perspective, stated in prose only, is that file-state handling belongs in the data platform as an explicit, testable control, with the statistical-genetics team owning scientific sign-off. No client experience is described or implied. Volatile items (software versions, API status, counts) are dated to sources opened in October and November 2026.

The scope is deliberately bounded: file states, build and effect-allele invariants, the harmonisation-code ledger, metadata and licence checks, and separate routes for SNP, CNV and gene-based data. It excludes method tutorials, biobank access and downstream scoring. No dataset-wide failure fraction was measured for this report, and none of the sources read offers a representative benchmark.

Key Changes in the Standards, Pipeline and Access Routes

From pre-GWAS-SSF to GWAS-SSF, and a revised specification in 2026

Before 2023 the Catalog accepted a minimum format called pre-GWAS-SSF. A survey of 327 public summary statistics files found 127 unique formats, one for every 2.5 files (www.biorxiv.org). The GWAS-SSF (GWAS summary statistics format) preprint reported that of 27,845 valid files obtained by the Catalog, about 17% lacked at least one of the new mandatory fields (www.biorxiv.org), and that effect allele frequency was the most commonly omitted (www.biorxiv.org).

GWAS-SSF was finalised and implemented in the Catalog in April 2023 (pmc.ncbi.nlm.nih.gov). It has two parts, a tab-separated data file and a YAML metadata file (raw.githubusercontent.com). The mandatory fields are chromosome, base pair location, effect allele, other allele, one effect size, standard error, effect allele frequency (EAF) and a p-value, and they must appear in a defined order. Submissions per month rose 33% after the standard was published (pmc.ncbi.nlm.nih.gov).

The standard has continued to change. Release v1.1.0, dated 26 June 2026, added z-score fallback support for meta-analysis studies where other effect size and uncertainty measurements are unavailable (github.com). The specification explains that some meta-analysis outputs, such as METAL output, may only provide a z-score (raw.githubusercontent.com). For an ingest pipeline the consequence is concrete: a file can now be valid GWAS-SSF with no usable standard error, so code that assumes beta and standard error together will fail or, worse, treat missing values as zero. The metadata file_type string in the Catalog's example remains "GWAS-SSF v1.0", so the specification revision a file satisfied cannot be inferred from that field alone.

Original, formatted and harmonised: three file states

Table 1 below summarises the states a pipeline will meet in the Catalog's FTP (file transfer protocol) tree. The state determines which invariants can be assumed, so the pipeline should record it as data, not infer it from a filename.

File state	Where it sits	Genome build	Allele orientation	Sidecars and provenance	Accepted use (proposed)
Original	Accession directory, original .tsv	As declared in metadata; GCST90270927 is GRCh37	As submitted. GWAS-SSF recommends effect and other allele fixed to ALT and REF where possible but accepts other orders	Metadata YAML with is_harmonised false; md5 checksum required at submission	Archive and audit; analyse only after your own build and allele checks
Formatted (legacy)	A *.f.tsv file inside harmonised/, for files harmonised before 2023	Unchanged	Unchanged from submission	Older layout; very old files with no GCST prefix have no YAML	Header-mapping reference only
Harmonised	harmonised/ with .h.tsv, .tbi, .log and .yaml	One build; GCST90270927 sidecar states GRCh38 (ftp.ebi.ac.uk)	Forward strand, aligned to the reference	is_harmonised true; reference recorded; running log with counts	Candidate input to colocalisation and MR after the acceptance tests below
Awaiting or non-standard	No harmonised directory	As original	As original	Likely awaiting harmonisation, or flagged "Non-GWAS-SSF"	Quarantine; do not substitute the original silently

The table needs one interpretive caution. The harmonised file is a different file with different rows, not the original with a column added. The pipeline documentation says that since April 2023 only the harmonised results are retained in the final .h.tsv file (ebispot.github.io), while older harmonised files list harmonised values

in separate hm_ prefixed columns (ebispot.github.io). A reader written for one layout will misread the other, so treat the layouts as different parsers and record which one produced each row.

Genome build and coordinate invariants

The Catalog states that its data are mapped to GRCh38.p14 and dbSNP build 156, that genome build has been mandatory in summary statistics metadata for submissions since 2020, and that for older submissions the build may have been inferred during curation. The GWAS-SSF specification places the genome assembly and coordinate system in metadata, with positions following the declared system, either 1-based or 0-based (raw.githubusercontent.com), while the preprint describes the coordinate-system field as optional (www.biorxiv.org), so legacy files may not declare it. Chromosome values are integers from 1 to 25 (raw.githubusercontent.com).

The harmonisation pipeline maps build in an ordered sequence: it first tries to update the position by mapping the rsID against Ensembl, then falls back to liftover to the latest build, and otherwise removes the variant. The pipeline documentation names the UCSC LiftOver tool for the fallback (ebispot.github.io). Two consequences follow. A position in a harmonised file may come from an rsID lookup rather than from the submitted coordinate, so rsID-to-position disagreement should be tested, not assumed away. And variants dropped as unmapped are a separate class from hm_code failures: in the pipeline's example log, 132,485 of 21,727,440 sites were dropped because they could not be mapped (ebispot.github.io).

MungeSumstats can infer the genome build from the data, which offers an independent check on the declared build (bioconductor.org). The standalone harmoniser repository documents support for the Ensembl variants VCF reference (hg38, dbSNP 151) (github.com), while sidecars may name Ensembl release 95, so record both. Other resources must have their build read from each source. FinnGen documents summary statistics on genome build 38 (finngen.gitbook.io), whereas Pan-UK Biobank flat files report positions in GRCh37 coordinates (pan.ukbb.broadinstitute.org). Merging them without liftover would misalign every position. Liftover itself loses records: UCSC liftOver writes failed conversions to a separate "unMapped" output (genome.ucsc.edu), CrossMap writes an .unmap file (crossmap.sourceforge.net), and a ClinVar benchmark associated conversion failure with reference-build gaps and with variant types such as indels and duplications. A pipeline that discards the unmapped file loses the evidence of how many variants were dropped.

Effect-allele orientation and the flip arithmetic

An effect size is only meaningful relative to a named allele. MR guidance states that associations must be expressed per additional copy of the same allele when summarised data from different sources are combined (pmc.ncbi.nlm.nih.gov), and the coloc documentation tells users to check which allele the beta describes and to switch its sign if needed (chr1swallace.github.io). Conventions differ by source. In GWAS-VCF the alternative allele is always the effect allele (pmc.ncbi.nlm.nih.gov), MungeSumstats assumes the second allele, A2, is the effect allele, and FinnGen (finngen.gitbook.io) and Pan-UK Biobank (pan.ukbb.broadinstitute.org) both use the alternate allele. GWAS-SSF itself does not require alignment, and its authors wrote that a mandatory requirement for allele alignment should be considered in future iterations (www.biorxiv.org).

Table 2 gives a clearly synthetic worked example, using separate illustrative rows so that beta and odds ratio are never mixed in one row. The values are invented and are not from any study. The arithmetic follows the documented inversion of effect size, confidence interval and EAF on a swap (ebispot.github.io), and a statistician should confirm it against the specific model before it is encoded as a test.

Illustrative row (synthetic)	Before swap (effect allele A, other allele G)	After swap (effect allele G, other allele A)	Rule applied
Row 1, continuous trait	beta 0.2, standard error 0.05, EAF 0.2, p-value 6.3e-5	beta -0.2, standard error 0.05, EAF 0.8, p-value 6.3e-5	beta changes sign; standard error and p-value unchanged; EAF becomes 1 minus EAF
Row 2, binary trait	OR 2.0, 95% CI 1.50 to 2.67, EAF 0.2	OR 0.5, 95% CI 0.37 to 0.67, EAF 0.8	OR becomes its reciprocal; CI bounds become reciprocals and exchange roles; EAF becomes 1 minus EAF

Rows 1 and 2 are independent examples. In Row 2 the new lower bound is 1 divided by 2.67 (0.37) and the new upper bound is 1 divided by 1.50 (0.67). The two-sided p-value is unchanged because the test statistic changes sign but not magnitude. The invariants are testable on any harmonised file: the sign of beta (or log odds ratio) must move opposite to the swap, EAF before and after must sum to 1 within rounding, and standard error and p-value must be unchanged. Harmonisation does not re-estimate the association statistics, but when the effect and other alleles are swapped it inverts the beta, odds ratio, 95% CI and effect allele frequency, so it does not leave the signs untouched (www.ebi.ac.uk).

Harmonisation codes and the dropped-record ledger

The pipeline's reference page defines hm_code values 1 to 18 and labels codes 9 and 14 to 18 "Not harmonised", for example code 9, "Palindromic; Drop palindromic" (ebispot.github.io), code 15 for no matching variant in the reference (ebispot.github.io), code 16 for multiple matching variants (ebispot.github.io) and code 18 for EAF below a specified minor allele frequency threshold (ebispot.github.io). The numeric threshold behind code 18 was not stated in the pages read, so it must come from the pipeline version that produced the file. The methods page adds that processes involving "Infer strand" (codes 1 to 4) are currently not being used (www.ebi.ac.uk).

Palindromic variants (A/T or G/C) cannot be oriented from alleles alone. The pipeline samples 10% of non-palindromic sites to estimate strand consensus (ebispot.github.io), uses a default threshold of 0.99 (ebispot.github.io), and drops palindromic variants where the evidence is weak (ebispot.github.io). Elsewhere the problem is handled differently: TwoSampleMR's default infers strand from allele frequencies for palindromes, and its more conservative action drops all palindromic SNPs (mrcieu.github.io), while TwoSampleMR also drops variants coded as D/I in one dataset and as equal-length indel alleles in the other (mrcieu.github.io), and MR-Base notes that allele frequency may not indicate strand when it is close to 0.5.

Table 3 proposes a retain or quarantine decision per class. The "documented meaning" column follows the pipeline's reference pages and the methods page. The decision column is this report's proposal and must be verified against the pipeline version recorded in the running log: the example log reports pipeline version 0.1.0 (ftp.ebi.ac.uk), and later standalone releases added features such as z-score harmonisation (github.com).

hm_code	Documented meaning	Proposed decision	Condition to verify
10, 12	Non-palindromic; correct orientation (12 is reverse strand, so alleles are reverse complemented)	Retain	Count matches the log; spot-check sign against source
11, 13	Non-palindromic; flipped alleles	Retain after sign test	Beta sign reversed, EAF equals 1 minus source EAF, OR reciprocal
5, 6	Palindromic; assume forward strand	Retain conditionally	Log shows strand consensus near 1; for MR and coloc, drop where EAF is near 0.5
7, 8	Palindromic; assume reverse strand	Quarantine	Unexpected for a forward-consensus file
1 to 4	Palindromic; infer strand	Quarantine	Not in use per methods page; presence implies a different pipeline version
9, 17, 18	Palindromic dropped, or EAF missing or low	Exclude from effect-direction analyses	Record counts as drop reasons
14, 16	Required fields unknown; multiple reference matches	Exclude	Record counts; multiple matches are ambiguous
15	No matching variant in reference	Exclude	Dominant failure class in both example logs

The documentation pages read here do not state whether rows with failure codes persist in the harmonised file or are removed. The pipeline should therefore reconcile the row count of the .h.tsv.gz against the log (mapped, harmonised and failed counts) and fail the ingest if they disagree. The final quality-control step also removes variants lacking a valid variant ID, chromosome, base pair location or p-value, so filtered counts can differ from mapped counts. Every drop reason needs its own denominator, as the audit below shows.

Metadata, licence, ancestry and sample overlap

Metadata is the contract that tells a pipeline what the numbers mean. The YAML sidecar sits next to each data file with the suffix -meta.yaml. The specification requires analysis software and version in the metadata so that p-values of zero can be interpreted (raw.githubusercontent.com), and the validation tool treats zero p-values as failures unless the analysis software is provided (raw.githubusercontent.com). An underflowed p-value should therefore be carried as a flagged value or as a negative log10 p-value, not coerced to the smallest representable number. The standard also allows EAF to be rounded up to a floor value stated in the metadata (raw.githubusercontent.com), which matters for rare-variant work.

Licence and terms must be stored per study. The methods page says newly submitted summary statistics have been required to be CC0 since 2022 (www.ebi.ac.uk), with special terms in rare cases appearing in an adjacent readme. The 2024 paper states a March 2021 cut-off for CC0 (pmc.ncbi.nlm.nih.gov). Two dates therefore appear in current sources, and the discrepancy should be preserved in a data catalogue as "per study page". In the audited GCST90270927 directory the YAML sidecars carried no licence field, so terms had to come from the study page, not the file.

Ancestry and sample structure are not solved by metadata checks. Coloc's original model assumes the two datasets come from the same ethnic group with identical allele frequencies and LD patterns (pmc.ncbi.nlm.nih.gov), whereas its documentation notes that coloc.abf does not assume similar LD structures provided a dense SNP map exists. Fine-mapping with a reference LD matrix needs a panel of similar ancestry (pmc.ncbi.nlm.nih.gov) and consistent allele encodings. The LD reference therefore must be recorded and compared with sample ancestry.

Sample overlap needs its own gate. In two-sample MR, bias from sample overlap is a linear function of the proportion of overlap: with a null effect and a 10% one-sample relative bias, relative bias is 5% at 50% overlap and 3% at 30% overlap (pmc.ncbi.nlm.nih.gov). A later study described overlap as a modifier of weak instrument bias and winner's curse (onlinelibrary.wiley.com). The sidecar fields documented on the methods page describe sample size and ancestry but not contributing cohorts, so overlap cannot be tested from the sidecar alone and needs a manual or cohort-registry step.

Beyond SNPs: CNV and gene-based data are different objects

The methods page says that since April 2026 CNV and gene-based GWAS data should follow the beyond-SNPs standards, and that previously these data were submitted without a standard schema, which limited downstream reuse (www.ebi.ac.uk). The linked landing page is more cautious: these are draft standards, not yet fully integrated into the submission system, and users are asked to validate data locally (www.ebi.ac.uk). It calls them proposed and evolving. This report preserves that tension. The methods page wording ("should follow") describes expectation, the landing page describes status (draft), and neither should be read as a finalised mandatory specification. The SNP harmoniser should not be assumed to process non-SNP input: the 2024 paper says files not harmonised include non-standard submissions such as CNVs or gene-based analyses (pmc.ncbi.nlm.nih.gov), and the methods page says non-GWAS-SSF files can include gene-based, CNV and multivariate analyses where effect information is not available.

Table 4 compares field rules for the SNP standard and the two draft requirement pages as opened in October 2026. Draft cells should be re-checked at each release.

Concept	SNP (GWAS-SSF)	CNV (draft)	Gene-based (draft)
Position	Chromosome and base pair location; build and 0/1-based in metadata	Chromosome, start and end, "0/1 based"	Chromosome, start and end of the gene
Identifier	Effect and other allele; optional variant ID and rsID	Region coordinates; no alleles	HGNC symbol or Ensembl gene ID, at least one
Effect size	One of beta, OR, HR; z-score fallback in v1.1.0	At least one required: beta, OR, HR or z-score	Optional
Uncertainty	Standard error mandatory (#NA in z-score files)	Required with an effect size; z-score accepts none (www.ebi.ac.uk)	Required if an effect size is given
CI field names	ci_lower, ci_upper (encouraged)	confidence_interval_lower	ci_lower
Model or count	Not required	statistical_model_type required	n_snps optional

Concept	SNP (GWAS-SSF)	CNV (draft)	Gene-based (draft)
Minimum size	More than 100,000 SNPs	At least 10,000 pre-QC variants	At least 10,000 pre-QC genes

One parser cannot serve all three. The CNV draft uses confidence_interval_lower while the gene draft uses ci_lower, and effect size is mandatory for CNV but optional for genes. The drafts also restrict combinations: odds ratio and hazard ratio are mutually exclusive in a file, and with several effect sizes one must be designated primary.

Gene-level statistics are produced differently from variant-level ones. MAGMA aggregates marker data to whole genes and reports a Z-value derived from the gene p-value (ibg.colorado.edu). These are not effect-allele data and cannot enter variant-level colocalisation or MR without a separate, documented method.

Access routes: API status

The methods page says API access to summary statistics has been deprecated and that an improved API is coming soon (www.ebi.ac.uk). The Catalog API documentation states that REST API v2 is the only supported version (www.ebi.ac.uk), that the previous API was frozen in 2021 and retired in 2026, and that there is currently no API for summary statistics (www.ebi.ac.uk). The FAQ says the Catalog is working towards a new database for summary statistics. The wording differs on timing ("coming soon" against "working towards"), and neither page names a date. Bulk retrieval is by FTP, updated nightly, at paths based on the GCST accession, and a nightly list of harmonised studies is published.

Implementation Considerations and Process Changes

An ingest pipeline for public summary statistics should be a set of explicit routes with acceptance tests, not one loader.

Four ingest routes

- **SNP route (GWAS-SSF or pre-GWAS-SSF).** Require a harmonised file and its sidecar. Refuse to substitute the original if harmonisation is pending. Record file name, md5 checksum, source URL, retrieval date, licence, genome assembly, harmonisation reference and pipeline version.
- **CNV route (draft standard).** Parse with a dedicated schema, validate locally as the landing page requests, store region coordinates rather than alleles, and keep these rows away from variant-level methods.
- **Gene-based route (draft standard).** Parse gene identifiers, p-value and optional effect fields. Join to variant-level evidence only through a documented gene-level step, never by position overlap alone.
- **Non-standard route.** Files flagged non-GWAS-SSF go to manual review because they may lack effect information.

Acceptance tests

Each test below is a proposal that is mechanical, produces a count, and can be written as code.

- **State test.** Sidecar says is_harmonised true and the file is a .h.tsv.gz; otherwise reject for effect-direction use.
- **Build test.** Sidecar genome_assembly equals the build of your reference panel (GRCh38 for current Catalog harmonised files).
- **Checksum test.** The md5 of the download matches the sidecar value.
- **Row reconciliation.** Rows in the .h.tsv.gz equal the log's harmonised count plus any retained failure rows.
- **Flip test.** For flipped codes (11, 13, 6), recompute beta and EAF from the original file and compare signs and 1 minus EAF.
- **Palindromic test.** Count codes 5, 6, 9, 17 and 18. For MR and colocalisation, exclude palindromic variants with EAF near 0.5 unless orientation is verified (pmc.ncbi.nlm.nih.gov).
- **Missing-field test.** Count missing standard error and EAF; a missing standard error can be legitimate in z-score files, and #NA is the standard's missing-value token (raw.githubusercontent.com).
- **p-value test.** Flag p-values equal to 0 and read the analysis software from metadata. The validator also fails files with fewer than 100,000 rows by default (raw.githubusercontent.com).
- **Reference test.** Compare harmonised EAF with a reference panel for non-palindromic variants. Choose and document a deviation threshold against the reference population; frequency checks cannot detect problems near 0.5.
- **Duplicate test.** Decide a key (chromosome, position, ref, alt) and fail on duplicates; MungeSumstats checks duplicates by base pair position.
- **Licence test.** Store the licence per study and flag special terms.

Scientific sign-off gates

Metadata quality control cannot decide the following, so the pipeline should emit a review task for each:

- **Sample overlap** between exposure and outcome datasets; the direction and size of bias vary with the degree of overlap.
- **Ancestry and LD reference.** LD patterns can differ between ancestry groups.
- **Colocalisation assumptions.** The original model assumes at most one causal variant per trait in a region and two distinct datasets of unrelated individuals.
- **Instrument strength and LD in drug-target MR.** Report variance explained and the F statistic; unaccounted correlation between instruments can understate standard errors.
- **Gene mapping.** For target prioritisation, relative success estimates vary with confidence in variant-to-gene mapping.

Versioned refresh and source ledger

The sources moved during 2026: the specification reached v1.1.0 on 26 June (github.com), the validation tool published release 2.0.2 in September (github.com), and the harmoniser's latest release carried a missing-value fix (github.com). A refresh policy should therefore be tied to source releases, not calendar dates: re-run the acceptance tests and re-fetch sidecars when the specification, validator or harmoniser changes. Keep a ledger with, for each study, the immutable file names, URLs, retrieval dates, licence, pipeline version and test results.

Data Analysis and Evidence

A bounded real-study audit: GCST90270927

To anchor the checks in a real file, the accession directory GCST90270927 was opened on 19 November 2026. It holds the original data file, its YAML sidecar, a checksum file and a harmonised subdirectory. The harmonised subdirectory holds the .h.tsv.gz data file, its sidecar, a tabix index and a running log. The data file, index and log carry a modified time of 2023-11-03 (ftp.ebi.ac.uk), while the sidecars were modified on 2025-02-09. A pipeline should therefore store checksums and the log's pipeline version, not the sidecar date alone. The original sidecar declares GRCh37 and the harmonised sidecar GRCh38 (ftp.ebi.ac.uk).

Table 5 audits the log with fixed denominators. The release is the November 2023 harmonisation run, the target build is GRCh38 (source GRCh37), the reference is Ensembl release 95, and the pipeline version is 0.1.0. The log's percentages are of carried-forward sites. The input-based figures are calculated here from the log counts.

Stage or class	Count	Share of carried-forward sites (log)	Share of input sites (calculated)
Input sites	13,822,306	n/a	100%
Dropped, could not be mapped	6,362	n/a	0.046%
Carried forward	13,815,944	100%	99.954%
Code 10, forward strand, correct	6,382,570	46.20%	46.18%
Code 11, forward strand, flipped	4,854,469	35.14%	35.12%
Code 5, palindromic, assumed forward, correct	952,827	6.90%	6.89%
Code 6, palindromic, assumed forward, flipped	954,049	6.91%	6.90%
Codes 12 and 13, reverse strand	10,093	0.07%	0.07%
Harmonised in total	13,154,008	95.21%	95.17%
Code 15, no matching reference variant	661,899	4.79%	4.79%
Code 16, multiple matching variants	37	0.00%	0.00%

The log reports a palindromic mode of forward, inferred from a consensus on 10% of sites. Four readings follow. First, the harmonised fraction of input is 95.17% against the log's headline 95.21%, and the gap is the denominator difference between mapped and input sites; reporting only the headline hides the unmapped drops. Second, forward-strand, palindromic and reverse-strand flips total 5,812,386 variants, or 42.07% of carried-forward sites, so sign inversion is the rule, not the exception, for this file. Third, palindromic variants

(codes 5 and 6) make up 13.80% of carried-forward sites and are oriented by assumption, so any MR or colocalisation using this file inherits that assumption. Fourth, failures are dominated by code 15, which accounts for 661,899 of 661,936 failed sites. This is a single file from a 2023 run with pipeline version 0.1.0 and must not be generalised.

Other quantified reference points

Aggregate figures give context without replacing a per-file audit. The Catalog's 2024 paper reports an average of about 3.5% unharmonisable variants removed (pmc.ncbi.nlm.nih.gov), harmonisation of all GWAS-SSF files and 96% of pre-GWAS-SSF files, and 625,113 curated SNP-trait associations at July 2024. It also notes that the Catalog serves as a data feed to resources such as Open Targets Genetics.

Downstream platforms add their own filters. Open Targets ingests Catalog full summary statistics only when they have been harmonised (platform-docs.opentargets.org), filters SNPs with unavailable beta, standard error or p-value (platform-docs.opentargets.org), and flags curated lead variants with palindromic alleles. Its QC flags studies with fewer than 2,000,000 variants and makes them ineligible for SuSiE fine-mapping (platform-docs.opentargets.org). For proteomics data it aligns allele order with gnomAD and flips the effect sign where alleles are reversed (platform-docs.opentargets.org), and its locus-to-gene model is updated in each release, so scores are release-specific (platform-docs.opentargets.org). Each downstream consumer thus repeats some harmonisation checks with its own thresholds, which argues for storing intermediate counts from your own pipeline.

Genetic-evidence magnitudes should be quoted with their limits. The 2015 analysis reported that selecting genetically supported targets could double the success rate in clinical development (www.nature.com). The 2019 re-analysis noted that only 10% of candidate new molecular entities in early clinical trials are eventually approved. The 2024 analysis found relative success largely unaffected by effect size, minor allele frequency or year of discovery, and highest for Online Mendelian Inheritance in Man (OMIM) evidence at 3.7. None of these figures describes the risk that a file has the wrong effect direction.

Implications and Future Directions

For data platform owners, file state is a first-class attribute. A data model should carry state, build, allele convention, harmonisation code, pipeline version and licence for every batch, and allow queries such as "only rows with codes 10 to 13". Harmonisation does not remove downstream checks: allele flips between z-scores and an LD reference are described as a common source of fine-mapping problems, so the check must be repeated when a different LD panel is introduced.

For statistical-genetics leads, the acceptance rules separate mechanical checks from scientific review. The mechanical checks can be automated and logged. The scientific gates (sample overlap, ancestry compatibility, LD reference choice, instrument strength and colocalisation assumptions) remain human decisions, and no metadata field replaces them. A defensible record states which gate was applied and who signed it off.

For tooling, three moving parts deserve watching. The specification is still evolving, as the 2026 z-score fallback shows (github.com). The draft non-SNP standards invite feedback, so a schema registry should version them and tolerate change. And the Catalog is developing a new summary statistics database, so

retrieval code should sit behind an adapter. Alternative formats persist as well: GWAS-VCF stores one variant per row with the alternative allele as the effect allele, so an ingest layer supporting both conventions needs an explicit allele-convention field.

On open questions, no representative benchmark of harmonisation failure across the Catalog was found, and the threshold behind code 18 was not located in the documentation read. Teams that need either should measure them on their own refresh using the reconciliation tests. Scientific interpretation of any test result stays with the study team. Questions about licences and data use should go to the relevant legal professionals, as this report is not legal advice.

Frequently Asked Questions (FAQs)

What are GWAS summary statistics, and how are they used for drug discovery?

They are the full, aggregate association results for every variant analysed in a study. In drug discovery they feed colocalisation, MR and locus-to-gene scoring, where genetic support is associated with a higher probability of success for a mechanism. They support prioritisation, not proof of causality.

What does harmonising GWAS summary statistics mean?

In the Catalog it means reporting positions against the same and latest genome build, inferring missing chromosome, position and variant ID data where possible, and orientating alleles to the forward strand where possible. Variants that cannot be mapped or matched are removed or flagged.

What is the GWAS summary statistics format standard (GWAS-SSF)?

It is a two-file standard, a tab-separated data file and a YAML metadata file (raw.githubusercontent.com), in the Catalog since April 2023 and revised to v1.1.0 in June 2026 (github.com).

What are the Catalog's harmonised summary statistics, and where are they?

They are .h.tsv files in a harmonised subdirectory of each accession folder. A nightly text list of harmonised studies is on the FTP, and a folder with no harmonised file is likely awaiting harmonisation.

What quality control should I run on summary statistics?

Run the acceptance tests above: state, build, checksum, reconciliation, flip, palindromic, missing-field, p-value, reference frequency and duplicate checks. Published tools offer related checks, such as the GWAS-SSF validator's zero p-value rule (raw.githubusercontent.com).

Can I reuse public GWAS summary statistics in pharma analyses?

Often yes, subject to licence, which is per study. Newly submitted files are CC0, with special terms in rare cases, and older files may fall under standard EBI terms.

How does this relate to Mendelian randomisation for drug target validation?

MR requires the same allele for exposure and outcome, careful handling of palindromic variants, and attention to overlap and LD. External LD references are one way to model correlation between cis variants for drug targets.

How should CNV and gene-based GWAS data be handled?

Separately from SNP data. The beyond-SNPs standards are drafts (www.ebi.ac.uk) with different field rules for CNV and genes, and the SNP harmoniser should not be assumed to process them.

Conclusion

Public GWAS summary statistics can enter colocalisation, MR and target prioritisation, but only as files whose state, build, allele convention and exceptions are known. The Catalog documents original, formatted and harmonised files, a metadata sidecar, harmonisation codes and a running log, which together supply the raw material for an acceptance test suite. The most consequential invariant is effect direction. In the one real study audited, about four in ten mapped variants were flipped during harmonisation, so a pipeline that ignores file state would carry the wrong allele for a large share of effects.

Four limits remain. The standard is evolving, and a valid file may now lack a standard error. The CNV and gene-based formats are drafts with their own field rules, and the SNP harmoniser should not be assumed to cover them. Dropped records need their own denominators, and no representative failure benchmark is available, so each team should measure its own. Finally, metadata checks do not resolve sample overlap, ancestry or LD reference mismatch, which remain scientific decisions with named owners.

The practical recommendation is to treat each downloaded file as a versioned artefact with a ledger entry, run the mechanical tests on every refresh, and hold the scientific gates open until a statistical geneticist signs them off. Source status here is as of November 2026 and should be re-checked on each specification, validator or harmoniser release.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.ebi.ac.uk/gwas
2. https://pmc.ncbi.nlm.nih.gov/articles/PMC11701593
3. https://ebispot.github.io/gwas-sumstats-harmoniser-documentation/Introduction/Harmonising-the-variants
4. https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90270001-GCST90271000/GCST90270927/harmonised/GCST90270927.running.log
5. https://ebispot.github.io/gwas-sumstats-harmoniser-documentation/Explanation/output-folder-structure
6. https://github.com/gwas-catalog/summary-statistics-standard/releases
7. https://www.ebi.ac.uk/gwas/docs/methods/summary-statistics
8. https://www.ebi.ac.uk/gwas/apps/beyond-snps
9. https://pmc.ncbi.nlm.nih.gov/articles/PMC7384151

10. https://pmc.ncbi.nlm.nih.gov/articles/PMC11096124
11. https://chr1swallace.github.io/coloc/articles/a02_data.html
12. https://www.nature.com/articles/ng.3314
13. https://www.biorxiv.org/content/10.1101/2022.07.15.500230.full.pdf
14. https://www.biorxiv.org/content/10.1101/2022.07.15.500230.full
15. https://raw.githubusercontent.com/gwas-catalog/summary-statistics-standard/main/gwas-ssf_v1.1.0.tex
16. https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90270001-GCST90271000/GCST90270927/harmonised/GCST90270927.h.tsv.gz-meta.yaml
17. https://ebispot.github.io/gwas-sumstats-harmoniser-documentation/Introduction/Genome-Build-Mapping
18. https://bioconductor.org/packages/release/bioc/vignettes/MungeSumstats/inst/doc/MungeSumstats.html
19. https://github.com/EBISPOT/gwas-sumstats-harmoniser
20. https://finngen.gitbook.io/documentation/data-description
21. https://pan.ukbb.broadinstitute.org/docs/per-phenotype-files
22. https://genome.ucsc.edu/FAQ/FAQdownloads.html
23. https://crossmap.sourceforge.net
24. https://pmc.ncbi.nlm.nih.gov/articles/PMC7805039
25. https://ebispot.github.io/gwas-sumstats-harmoniser-documentation/Reference-guide/Hm_code
26. https://ebispot.github.io/gwas-sumstats-harmoniser-documentation/Introduction/Orientation-of-palindromic-variants
27. https://ebispot.github.io/gwas-sumstats-harmoniser-documentation/Tutorials/Running-the-Pipeline
28. https://mrcieu.github.io/TwoSampleMR/reference/harmonise_data.html
29. https://github.com/EBISPOT/gwas-sumstats-harmoniser/releases
30. https://raw.githubusercontent.com/gwas-catalog/sumstats-tools/main/docs/gwas-sumstats-tools-doc/CLI_validate.md
31. https://pmc.ncbi.nlm.nih.gov/articles/PMC4022491
32. https://pmc.ncbi.nlm.nih.gov/articles/PMC9337707
33. https://pmc.ncbi.nlm.nih.gov/articles/PMC5082560
34. https://onlinelibrary.wiley.com/doi/full/10.1002/gepi.22522
35. https://www.ebi.ac.uk/gwas/apps/beyond-snps/data_requirements/cnv
36. https://ibg.colorado.edu/cdrom2025/4.1.PathwayAnalysis_DougWightman/final/manual_v1.10.pdf
37. https://www.ebi.ac.uk/gwas/docs/api
38. https://github.com/gwas-catalog/sumstats-tools/releases
39. https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90270001-GCST90271000/GCST90270927/harmonised
40. https://platform-docs.opentargets.org/gentropy/data-sources
41. https://platform-docs.opentargets.org/gentropy/locus-to-gene-l2g

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.