IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

gnomAD Allele Frequency Variant Curation: GroupMax FAF Guide

> 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 · 2026-11-14 · gnomad allele frequency · variant curation · groupmax faf · filtering allele frequency · gnomad grpmax · popmax vs grpmax · acmg ba1 bs1 · gnomad coverage · gnomad v4.1 · population frequency evidence

Original article: https://intuitionlabs.ai/articles/gnomad-frequency-evidence-groupmax-faf-variant-curation



In this article

1. Executive Summary
2. Introduction and Background
3. Definitions: AC, AN, AF, FAF, GroupMax FAF and Coverage
4. Disease-Specific Frequency Policy: Why No Universal Threshold Applies
5. Absent, Uncallable and Discordant: Reading Missing and Conflicting Data
6. Evidence Capture, Versioning and Policy Tests
7. Data Analysis and Evidence
8. Implications and Future Directions
9. Frequently Asked Questions (FAQs)
10. Conclusion

Executive Summary

This report explains which gnomAD frequency fields belong in a research variant-exclusion policy, what each field can and cannot support, and when a pipeline should escalate to a human curator instead of excluding a candidate. It is a research-process reference, not individual clinical advice. As of November 2026, the frequency data in gnomAD are the v4.1 family: v4.1.1 was released on March 30, 2026 (gnomad.broadinstitute.org), and the later v4.1.2 is a constraint-only update in which all variant frequency data and filtering status are unchanged from v4.1.1 (gnomad.broadinstitute.org).

Three distinctions carry most of the weight. Observed allele frequency (AF) is allele count (AC) divided by allele number (AN), so it reflects confidently sequenced haplotypes (onlinelibrary.wiley.com). The filtering allele frequency (FAF) is a different quantity: gnomAD defines it as the maximum credible genetic ancestry group AF, for example the lower bound of the 95% confidence interval (gnomad.broadinstitute.org). The browser's GroupMax FAF comes from the group with the highest FAF, not the group with the highest AF (gnomad.broadinstitute.org). The ClinGen Variant Curation Expert Panel (VCEP) Review Committee's June 23, 2025 communication uses the same definition and names GroupMax FAF as the frequency to apply for BA1 and BS1 (www.clinicalgenome.org).

No universal rarity threshold exists. The 2015 ACMG/AMP standards set BA1 at an allele frequency above 5% (pmc.ncbi.nlm.nih.gov), but ClinGen's 2018 refinement lets disease experts define gene-specific values numerically smaller than 0.05. In published 2018 specifications, BA1 was 0.1% for one cardiomyopathy gene (pmc.ncbi.nlm.nih.gov) and 0.05% for RASopathy genes (pmc.ncbi.nlm.nih.gov). The June 2025 communication recommends avoiding strict absence when specifying BA1, BS1 and PM2 thresholds (www.clinicalgenome.org).

"Not seen" and "not callable" are different states. Since v4.1, gnomAD computes AN across all callable sites (gnomad.broadinstitute.org), and a locus missing from the all-sites AN files means zero gnomAD samples had any genotype call there (gnomad.broadinstitute.org). Absence at a well-covered site is weak evidence: ClinGen's Sequence Variant Interpretation (SVI) group moved PM2 from Moderate to Supporting strength

(clinicalgenome.org). Exome and genome frequencies can also disagree: gnomAD reports that about 2.5% of variants (2,230,151 of 91,177,483) have statistically different frequencies between data types (gnomad.broadinstitute.org), a release-specific comparison and not an error rate.

The practical recommendation is a versioned evidence row for every frequency lookup (build, normalized allele, release, AC, AN, AF, group, FAF semantics, filters, flags, review status) plus acceptance tests that force escalation when AN is low, the site is flagged, datasets disagree, or no disease-specific policy exists. No current variant-count audit or real reclassification was performed for this report.

Introduction and Background

Research teams that triage candidate variants in rare-disease programs routinely use population frequency to remove variants that are too common to explain a penetrant disorder. The decision looks mechanical, yet it combines several ambiguities: which release and data type were queried, which ancestry group supplied the number, whether the site was sequenced in enough individuals, and whether a threshold derived for one disease is being applied to another.

This report addresses data stewards, curation leads and translational genomics platform owners who must write that policy down. It treats the gnomAD (Genome Aggregation Database) frequency fields as an evidence interface with documented semantics and separates three questions that are often collapsed. What does the number mean? Is it trustworthy at this site? Is it comparable with the threshold being applied?

The exclusion decision and the human-review boundary

A frequency rule in a research pipeline has two defensible outputs: exclude with a recorded rationale, or escalate to a curator. The design principle proposed here is that automatic exclusion should require every prerequisite to be machine-checkable, and any missing prerequisite should route to review. This is an editorial recommendation for software teams, not a ClinGen requirement. The June 2025 ClinGen communication gives no numeric thresholds, coverage requirements or minimum allele counts for VCEPs generally; it sets an approach, and those choices remain in each disease-specific specification.

Escalation is the safer default because frequency evidence is probabilistic. The committee notes that although gnomAD removed severe pediatric disease cohorts, the database still contains individuals with rare disease (www.clinicalgenome.org). A variant wrongly excluded on a flawed number disappears silently, whereas a false candidate is usually rejected later by a curator.

The process point here is that frequency lookups are versioned data products, and a policy that does not record release, field semantics and review state cannot be audited. The factual claims below rely only on independent public sources, and 2018 specifications are shown purely as examples of disease-to-disease variation; their current registry values were not rechecked.

Definitions: AC, AN, AF, FAF, GroupMax FAF and Coverage

The basic fields

Allele count (AC) is the number of observed alternate alleles. Allele number (AN) is the number of alleles with a genotype call at the site, which the VCF specification defines as the total number of alleles in called genotypes (samtools.github.io). In a VCF, AC carries one value per ALT allele in the same order as the ALT listing (samtools.github.io). gnomAD's AF is AC divided by AN, and AN can differ substantially between positions because coverage and quality vary (onlinelibrary.wiley.com). The VCF specification's own AF field is defined differently, as an estimate from primary data rather than called genotypes (samtools.github.io), so a generic VCF AF should not be assumed to equal gnomAD's AC/AN.

AN is a site-specific quantity, not a function of cohort size. Since version 4.1, gnomAD computes AN across all callable sites, including sites with no observed variation, and reports AN for both data types if the variant was called in either data type (gnomAD v4.1 release notes). Pipelines should never assume AN equals twice the number of samples.

Filtering allele frequency (FAF)

FAF answers a different question from AF. Rather than "what was observed?", it asks "what is the highest true population frequency still consistent with this observation, given the sample size?" Whiffin and colleagues illustrate why: a variant observed only once in a sample of 10,000 chromosomes is more likely to have a true frequency below 1 in 10,000 (www.nature.com), and the ExAC authors likewise report that frequency estimates from low allele counts are upward-biased and imprecise (www.nature.com). Hail's documentation defines FAF as the highest true population allele frequency for which the upper bound of the confidence interval of allele count, under a Poisson distribution, is still below the observed count given the allele number (hail.is). The method originates with Whiffin et al., who computed a maximum tolerated allele count at the upper bound of a one-tailed 95% Poisson interval (www.nature.com) and then reversed the process to obtain a filtering AF (www.nature.com). gnomAD introduced the annotation in v2.1, computed with Poisson 95% and 99% intervals per population (gnomad.broadinstitute.org).

The 95 versus 99 distinction is a confidence level, not a population: gnomAD lists FAFs for each genetic ancestry group separately at both levels (gnomad.broadinstitute.org), and the gnomad_methods library defaults to thresholds of 0.95 and 0.99 (raw.githubusercontent.com). Hail is explicit that its filtering AF is not disease-specific (hail.is); it must be compared with a disease-specific maximum credible AF, and a variant whose filtering AF is at or above that value is filtered (hail.is).

GroupMax FAF, grpmax and popmax

Three similar names are easy to confuse.

- **Popmax (older releases):** in v2.1, the allele frequency information for the non-bottlenecked population with the highest frequency (gnomad.broadinstitute.org). The ExAC authors found that filtering on the highest single-population AF was more powerful than filtering on global AF (www.nature.com).

- **Grpmax (v4 vocabulary):** gnomAD's 2023 update replaced "population" with "genetic ancestry group" to signal that groups are inferred, artificially created clusters (gnomad.broadinstitute.org). The grpmax annotation holds AC, AN, AF and homozygote count for the group with the highest AF (gnomad.broadinstitute.org) and is distinct from filtering allele frequency (gnomad.broadinstitute.org).
- **GroupMax FAF (browser label "Grpmax Filtering AF"):** the FAF of the group with the highest FAF, which can name a different group than grpmax.

The ClinGen committee gives the curation definition: the lower-bound AF of the 95% CI from the genetic ancestry group with the highest FAF (www.clinicalgenome.org). In the VCF, staff report the value as fafmax_faf95_max in the separate exome and genome files, with _exomes, _genomes and _joint suffixes in the joint file (discuss.gnomad.broadinstitute.org). In the gnomad_methods library, the maximum FAF and its ancestry group are retrieved per threshold (broadinstitute.github.io).

Which groups are excluded

FAF calculations include only non-bottlenecked groups (gnomad.broadinstitute.org). The ClinGen communication states that gnomAD intentionally does not include the Amish, Ashkenazi Jewish and European Finnish groups (and remaining individuals) (www.clinicalgenome.org); the library documents the excluded groups as typically bottlenecked or consanguineous (broadinstitute.github.io). Because the group set can differ by data type, a policy must record which data type (exome, genome or joint) supplied the value.

Coverage and callability

Exome AN varies by position because coverage can differ with exon capture between sequencing platforms (onlinelibrary.wiley.com), which is why a coverage check belongs beside every frequency lookup.

Table 1 summarizes the fields a policy engine is most likely to ingest, with denominator, scope and main failure mode for each.

Field	Numerator / denominator	Population scope	Confidence meaning	Allowed interpretation	Failure mode
AC	Alternate alleles after genotype filters	One data type and stratum (exome, genome or joint)	None (a count)	Input to AF and FAF	Reading AC as carrier count on non-PAR X or Y without checking ploidy handling
AN	Called alleles at the site (samtools.github.io)	Same stratum as AC	None (a denominator)	Direct measure of how many alleles were callable at this site	Assuming AN is constant or twice the cohort size
AF	AC / AN	Stratum-specific	None: a point estimate	Observed frequency in that stratum	Treating a small-AN estimate as stable
Ancestry-group AF	AC / AN within one inferred group	One genetic ancestry group	None	Describing group-level observation	Reading groups as race, ethnicity or the case's ancestry

Field	Numerator / denominator	Population scope	Confidence meaning	Allowed interpretation	Failure mode
FAF95 / FAF99	Poisson-based bound given AC and AN (hail.is)	Global and per non-bottlenecked group	Conservative bound at 95% or 99%	Comparison with a disease-specific maximum credible AF	Substituting AF for FAF, or mixing 95 and 99
GroupMax FAF	Maximum FAF across included groups	Non-bottlenecked groups only	Lower bound at the stated level	BA1/BS1 input named by the ClinGen committee	Missing value read as zero; grpmax AF read as GroupMax FAF
Coverage / AN percent	Fraction of samples above a depth, or AN relative to maximum	Per position and data type	Not a confidence interval	Judging whether absence is informative	Treating a coverage plot as per-site callability proof

Every row has a different denominator and population scope, so a single "frequency" column in a curation database is under-specified. FAF and GroupMax FAF are comparison inputs for a threshold test, whereas AC and AN are evidence about how far any frequency can be trusted.

Disease-Specific Frequency Policy: Why No Universal Threshold Applies

What the standards say

The ACMG/AMP standards define BA1 as an allele frequency above 5% in the Exome Sequencing Project, 1000 Genomes or ExAC (pmc.ncbi.nlm.nih.gov) and BS1 as an allele frequency greater than expected for the disorder (pmc.ncbi.nlm.nih.gov). The United Kingdom's Association for Clinical Genomic Science (ACGS) 2024 guidelines describe the 5% value as very high for most rare diseases and point to gene-specific calculation (www.genomicseducation.hee.nhs.uk).

ClinGen's BA1 refinement requires at least 2,000 observed alleles at the site (pmc.ncbi.nlm.nih.gov) and lets disease experts set gene-specific values below 0.05 (pmc.ncbi.nlm.nih.gov). For PM2, the SVI recommendation moved the strength from Moderate to Supporting (clinicalgenome.org), noting among other reasons that 54% of high-quality ExAC variants were seen only once (clinicalgenome.org). The ACGS guidelines add that the PM2 frequency or count should be at least a factor of 10 below the BS1 threshold (www.genomicseducation.hee.nhs.uk), and allow PM2 for recessive disorders when no homozygotes or hemizygotes appear in the relevant gnomAD datasets (www.genomicseducation.hee.nhs.uk). Two national or consortium frameworks therefore express PM2 differently, which is another reason to store the framework with the evidence.

The June 2025 communication moves ClinGen practice further from strict absence. It recommends avoiding strict absence for BA1, BS1 and PM2 thresholds (www.clinicalgenome.org) and treats a change to approved values as a major change requiring piloting (www.clinicalgenome.org). For VCEPs whose thresholds were already calculated frequencies, a lift-over to gnomAD v4 without modification is generally a minor change that does not require piloting (www.clinicalgenome.org). These statements concern VCEP specifications and are not a blanket migration approval for any other pipeline.

The derivation transfers, the number does not

Whiffin and colleagues' principle is that a fully penetrant variant cannot be present in the general population more frequently than the disease it causes (www.nature.com). Their hypertrophic cardiomyopathy example combines prevalence, the maximum allelic contribution of any single variant and penetrance to reach a maximum credible population AF of 4.0 x 10^-5 (pmc.ncbi.nlm.nih.gov). The inputs vary by disease, so the outputs do too. The ACGS guidelines advise conservative inputs (highest likely prevalence, lowest likely penetrance) when applying that approach to gnomAD for BS1 (www.genomicseducation.hee.nhs.uk), and a Genetics in Medicine review adds that population size and MAF estimate error matter when setting BS1 thresholds (www.nature.com).

Table 2 shows published 2018 expert-panel specifications as illustrations. They are historical examples calibrated to ExAC-era data; this report did not verify whether current Criteria Specification Registry versions still carry these values, and none should be applied outside the named disease.

Disease context	BA1	BS1	Derivation note
ACMG/AMP 2015 generic standard	Above 5%	Greater than expected for the disorder	Qualitative; designed to be refined per disease
Inherited cardiomyopathy (MYH7), 2018	0.1% (pmc.ncbi.nlm.nih.gov)	Not tabulated here; see the cited paper	Prevalence and penetrance assumptions stated in the paper
Hearing loss, 2018	0.5% recessive, 0.1% dominant (pmc.ncbi.nlm.nih.gov)	Lower than BA1, by inheritance mode	Prevalence-based derivation stated in the paper
RASopathy genes, 2018	0.05% (pmc.ncbi.nlm.nih.gov)	Not tabulated here; see the cited paper	PM2 specified as complete absence from population databases (pmc.ncbi.nlm.nih.gov)

BA1 spans two orders of magnitude in these examples, from 5% down to 0.05%, before accounting for the inheritance-mode split in hearing loss. The final row also shows why the 2025 communication matters: a strict-absence PM2 rule written for an earlier database generation is the kind of element VCEPs are asked to reassess for v4.

Implications for a research exclusion policy

A policy should store a threshold only together with its derivation record: disease, inheritance mode, prevalence, penetrance, heterogeneity assumption, FAF confidence level, and the author and date of the specification. Generic prioritization tools use frequency cutoffs too. Genomics England's Exomiser implementation retains variants below a minor allele frequency of 0.1% (or 2% for compound heterozygotes) (pipeline-rd-help.genomicsengland.co.uk), while the Exomiser protocol authors caution that 0.1% may be too low for recessive diseases with higher carrier prevalence (pmc.ncbi.nlm.nih.gov). Neither a tool default nor a 2018 panel value substitutes for a specification written for the disease at hand.

Ancestry handling is a separate prerequisite. A variant common in any one group can rule out pathogenicity despite a low global frequency, and gnomAD describes the groups as inferred and distinct from race or ethnicity (gnomad.broadinstitute.org). A pipeline should not translate a patient's self-described background into

a gnomAD group label. ACGS also cautions that some sub-populations are poorly represented in gnomAD (www.genomicseducation.hee.nhs.uk).

Absent, Uncallable and Discordant: Reading Missing and Conflicting Data

"Not seen" versus "not callable"

A variant missing from gnomAD can be in one of several states. The site may have been callable in many samples with no carrier. Coverage may have been limited, so absence says little. Or the locus may lie outside the calling intervals. The all-sites AN files separate the last case: a locus absent from them means no sample had any genotype call there. ACMG/AMP guidance makes the same point in curator language: a dataset must have coverage data to determine whether a variant is absent (www.nature.com), and read depth should be sufficient for an accurate call (pmc.ncbi.nlm.nih.gov).

For a position with no variants, the gnomAD FAQ advises inspecting coverage: a region search still displays coverage plots for zero-variant sites (gnomad.broadinstitute.org).

Some expert panels write explicit numbers. The ClinGen PALB2 specification includes a coverage requirement of more than 30X at the locus (erepo.clinicalgenome.org), and the BRCA1 specification attaches read-depth conditions to PM2 and to BA1 and BS1 (erepo.clinicalgenome.org). Those figures belong to named genes and specification versions and should not be reused for another gene. Absence is also release-dependent: in one study of dystonia-associated variants, 192 of 247 (77.7%) were still absent from v4.1 (movementdisorders.onlinelibrary.wiley.com), while 27 of 219 (12.3%) variants absent from v2.1.1 appeared for the first time in v4.1 (movementdisorders.onlinelibrary.wiley.com). That is one disease-specific study and not a general rate.

Low coverage and mappability

In v4.1.1, gnomAD added gene-level flags and noted that allele frequency and constraint metrics in low-coverage or low-mappability regions may not be reliable (gnomad.broadinstitute.org). The release's other change was a recalculation of gene constraint, and the later v4.1.2 update left frequency data unchanged. Teams should therefore not describe v4.1.1 as having changed how AF is calculated.

An earlier review notes that the browser highlights variants covered in fewer than 50% of individuals because their frequencies may be inaccurate (onlinelibrary.wiley.com), and that presence in v2 genomes only is the most common reason for low allele number (onlinelibrary.wiley.com). The ACGS guidelines advise using the latest gnomAD (v4) for variants with quality flags in v2 (www.genomicseducation.hee.nhs.uk), and a Genetics in Medicine review says frequency databases should be evaluated for sufficient depth of coverage to ensure accurate MAF estimates (www.nature.com).

Exome versus genome discordance

Version 4.1 introduced a warning for variants whose frequencies are highly discordant between exomes and genomes (gnomad.broadinstitute.org). The release post reports about 2.5% of variants (2,230,151 of 91,177,483) with statistically different frequencies in the two data types. This is a test count at gnomAD's

documented threshold in the v4.1 comparison, not a count of erroneous variants and not a count for later data.

The post also states that determining which frequency to apply downstream is often highly context-dependent (gnomad.broadinstitute.org). A preprint comparing 946 individuals sequenced both ways found that some variants passing filters have systematically different genotype calls across technologies (www.biorxiv.org) and that discordant sites are largely shared across ancestry groups (www.biorxiv.org); the published version could not be retrieved, so its wording may differ. Tool defaults add a further wrinkle: Ensembl's Variant Effect Predictor reports gnomAD exome and genome frequencies through separate options, each including only its own data (www.ensembl.org). For policy, a discordance flag should be an escalation trigger, and the record should state which data type the rule used.

Multi-allelic sites, normalization and ploidy

Staff explain that when gnomAD splits multi-allelic sites, different alternate alleles at one site can show slightly different AN because of genotype downcoding (discuss.gnomad.broadinstitute.org). A pipeline that joins on position alone, or compares an unnormalized representation, can attach the wrong record. Normalization has precise definitions: a VCF entry is normalized when it is left-aligned and parsimonious (academic.oup.com); bcftools norm does this only when a reference FASTA is supplied (samtools.github.io) and can split multiallelic sites into rows (samtools.github.io); HGVS assigns the most 3' position possible, a different convention (hgvs-nomenclature.org). The GA4GH Variation Representation Specification (VRS) states that normalization removes an ambiguity that impedes comparison across systems (vrs.ga4gh.org).

For sex chromosomes, the gnomad_methods export code sets chrY AC, AN and homozygote counts for XX samples to missing (raw.githubusercontent.com). No official statement of per-site AN arithmetic on non-PAR regions was located, so teams should verify AN behavior in their own extracts.

Decision table

Table 3 summarizes proposed handling of the main edge cases. The routing column is an editorial recommendation for a research-exclusion pipeline, not a ClinGen rule, and every row assumes a disease-specific frequency policy exists.

Situation	What is known	Evidence to capture	Default routing
Adequate AN, GroupMax FAF above the disease threshold, no flags	Frequency is inconsistent with the stated maximum credible AF	Release, group, FAF level, AC, AN, threshold ID	Eligible for automatic exclusion with an audit row
Adequate AN, GroupMax FAF missing	Missing when no non-bottlenecked group has a positive FAF (discuss.gnomad.broadinstitute.org); the maximum of zeros is treated as missing	AC and AN per group, reason for missing	Do not exclude on frequency; missing is neither zero nor rare
Variant absent, site callable	Not observed in sampled alleles; PM2 is Supporting at most	All-sites AN for the locus, release	Retain; absence alone supports neither exclusion nor pathogenicity (pmc.ncbi.nlm.nih.gov)

Situation	What is known	Evidence to capture	Default routing
Variant absent, no genotype calls at the locus	Not callable in gnomAD	All-sites AN file status	Escalate; no frequency evidence exists
Low-coverage or low-mappability flag, or coverage below the specification's requirement	Frequency may be unreliable	Flag text, coverage metric, AN	Escalate
Exome and genome frequencies discordant	Data types disagree by the documented test	Both values, flag, AN per data type	Escalate; do not silently pick one
High frequency only in a bottlenecked group	Excluded from FAF	Group label, group AC and AN	Escalate; pathogenic alleles can rise to high frequency in such groups (pmc.ncbi.nlm.nih.gov)
Multi-allelic site	Alleles split; AN can differ per allele	Normalized allele, source record key	Match on normalized allele, never position alone
Disease threshold unavailable or from another disease	Derivation does not transfer	Policy gap	Escalate to a domain expert

The pattern is that automatic exclusion is the narrowest cell. Every other row lacks either the information needed to trust the frequency or the policy needed to interpret it, and review handles both better than a default does.

Evidence Capture, Versioning and Policy Tests

Source-version ownership

Frequency evidence depends on release and build. gnomAD v4 is mapped to GRCh38 (gnomad.broadinstitute.org) and comprises 807,162 individuals (gnomad.broadinstitute.org), described in a 2025 peer-reviewed paper as 76,215 genomes and 730,947 exomes (pmc.ncbi.nlm.nih.gov); v2 held 125,748 exomes and 15,708 genomes (www.nature.com). One passage of the v4.1 post gives a slightly different exome figure from the v4.0 post; this report uses 730,947. Because variants can appear or disappear between releases, as the dystonia study above shows, "absent" needs a release identifier.

A reasonable ownership model assigns one steward per frequency source, who owns the release pin, the field mapping and the refresh calendar. As of November 2026, a frequency-only integration pinned to v4.1.1 frequency files is current because v4.1.2 is constraint-only.

The evidence row

A coverage and provenance row should be written for every lookup, whether or not the variant is excluded.

- **Genome build and normalized allele:** build, contig, position, reference and alternate after normalization. VRS computed identifiers are deterministic, so conforming implementations generate the same identifier for the same object (vrs.ga4gh.org).
- **Source and version:** dataset, release (for example v4.1.1), file or endpoint, retrieval date. ClinGen's curation interface now sources gnomAD v4.1 data from ClinGen for newly added variants and requires a fetch for earlier ones (vci-gci-docs.clinicalgenome.org).
- **Data type, group and counts:** exome, genome or joint; the ancestry group label; and AC, AN and AF exactly as released. The GA4GH Variant Annotation specification models these as a focus count, a locus count and a frequency (va-spec.ga4gh.org), with ancestry and sex strata as nested sub-cohorts (va-spec.ga4gh.org).
- **FAF semantics:** level (95 or 99), method (Poisson), group set, and whether the value is global, per-group or GroupMax. The same specification anticipates a group-max FAF as a derived result (va-spec.ga4gh.org).
- **Filters and flags:** the filter set, plus discrepancy and coverage flags.
- **Coverage evidence:** the coverage fraction or AN relative to maximum and the all-sites AN status; the specification names a sequencing coverage metric as an example quality measure (va-spec.ga4gh.org).
- **Policy reference and review status:** threshold ID, derivation version, source document, and automatic, escalated or curator-decided status with reviewer and date.

External assertions should stay separate: ClinVar is an archive of classifications made by its submitters (www.ncbi.nlm.nih.gov), so a ClinVar assertion is its own evidence item and not part of the frequency row. Tool behavior should be recorded too. Exomiser, for example, fails a variant that exceeds the cutoff in any configured frequency source (exomiser.readthedocs.io), and without its frequency filter step variants carry no frequency data (exomiser.readthedocs.io).

Synthetic arithmetic: AF is not FAF

The following calculation is synthetic and illustrates definitions only. It uses the Poisson-based definition documented by Hail (raw.githubusercontent.com); it was computed with a short bisection routine and does not reproduce gnomAD's production code, so a real pipeline's numbers may differ in later digits.

(Hypothetical Example) Suppose one site in one group has AC = 2 and AN = 1,000, so observed AF = 2 / 1,000 = 0.002. The 95% FAF under the Poisson definition is about 0.000355 and the 99% FAF about 0.000149. A second site with AC = 20 and AN = 10,000 has the same observed AF of 0.002, but its 95% FAF is about 0.001325 and its 99% FAF about 0.001108. Against an illustrative policy threshold of 0.001 (not a real disease value), the first site would not pass a FAF-based exclusion test and the second would, although observed AFs are identical. A pipeline must therefore never compute or substitute FAF as AF, and must store AN beside AC.

Acceptance tests

Each test uses synthetic records and expects a routing, not a clinical result.

1. **Threshold present, adequate AN, FAF above threshold, no flags:** automatic exclusion with a complete evidence row.

2. **Same AF, different AN:** with the two sites above and a threshold of 0.001, expect different routings.
3. **GroupMax FAF null with AC above zero only in an excluded group:** no exclusion, reason code "FAF missing".
4. **Absent variant, locus present in all-sites AN with high AN:** retention; PM2 no stronger than Supporting.
5. **Absent variant, locus missing from all-sites AN:** escalation with reason "not callable".
6. **Discrepant-frequency flag set:** escalation with both data-type values stored.
7. **Low-coverage gene flag set:** escalation.
8. **Multi-allelic site with differing AN across alleles:** correct per-allele record by normalized key.
9. **Threshold record from a different disease:** escalation, not reuse.
10. **Release change (v2.1.1 to v4.1.1) for a stored variant:** re-evaluation and a new evidence row, not an overwrite.

Data Analysis and Evidence

The quantitative material falls into four groups, each with a caveat about what it measures.

Release sizes. gnomAD v2.1 had 125,748 exomes and 15,708 genomes; the v3.1.2 genome set had 76,156 samples (www.nature.com); v4 reports 807,162 individuals, roughly ten exomes per genome. Larger denominators sharpen rare-variant frequencies, but sampling is uneven: the v2 paper notes a near-complete absence of representation from the Middle East, central and southeast Asia (www.nature.com) and that sample sizes remained far from capturing complete mutational saturation of the exome (www.nature.com).

Threshold ranges. Published BA1 values range from 5% generic to 0.05% for RASopathy genes, a hundredfold difference, and from 0.5% to 0.1% within hearing loss by inheritance mode. ClinGen's generic BA1 refinement sets a minimum of 2,000 alleles, and panel coverage cutoffs such as more than 30X in the PALB2 specification are further disease-specific parameters.

Discordance. About 2.5% of variants (2,230,151 of 91,177,483) are flagged as statistically discordant between data types at gnomAD's documented threshold. If queried variants resemble the whole set, which is an assumption and not a measured rate, a policy that escalates every flagged variant needs reviewer capacity for roughly one variant in forty.

Absence. In the dystonia study, 77.7% of 247 variants remained absent from v4.1 while 12.3% of 219 previously absent variants appeared; the SVI cited ExAC data in which 54% of high-quality variants were observed once. Together these figures show absence is common, release-dependent and weakly specific. For hypertrophic cardiomyopathy, Whiffin and colleagues obtained a maximum credible AF of 4.0 x 10^-5, a prevalence and penetrance assumption that cannot be transferred to another disease.

No variant-count audit of any pipeline and no reclassification of real variants was performed. Counts are those published by the cited sources at their stated dates, and gnomAD pages may change after the November 14, 2026 retrieval date.

Implications and Future Directions

For data architecture

Frequency evidence is best modeled as an append-only, versioned fact table keyed by normalized allele, release and stratum, not as a single "gnomAD AF" column. Field names change across releases and tools: v2 used population labels and popmax, v4 uses genetic ancestry groups and grpmax, and the VCF exposes GroupMax FAF under fafmax-prefixed names. A quarterly source review and a refresh at each gnomAD release keep the pin honest. Because specification changes may require piloting (www.clinicalgenome.org), thresholds derived from ClinGen specifications should be re-verified at the specification source rather than copied from an earlier pipeline.

For scientific AI assurance

Any model or rule engine that consumes frequency evidence inherits these distinctions. An assistant that summarizes a variant as "absent from gnomAD" without the callability state is making an unverifiable claim. The acceptance tests above can be run on each release and each policy edit, with failures blocking promotion of the new configuration.

Open questions

- **Specification coverage:** many diseases lack a gnomAD v4-era expert specification, so a policy must decide whether to escalate all of them.
- **Flag stability:** the discrepant-frequency flag is new in v4.1, so its definition may change.
- **Implementation detail:** the documented FAF definition is Poisson-based (hail.is), but this report did not verify the numerical implementation of Hail's function.

Frequently Asked Questions (FAQs)

How should a curator interpret a gnomAD allele frequency?

Read it as AC divided by AN for a specific data type and group, and check AN first. The same AF at different AN supports different conclusions.

What is the gnomAD filtering allele frequency?

FAF is the maximum credible genetic ancestry group AF, for example the lower bound of the 95% confidence interval, meant for comparison with a disease-specific maximum credible frequency (hail.is).

What is GroupMax FAF, and how does it relate to grpmax FAF in v4?

GroupMax FAF is the FAF of the group with the highest FAF, not the FAF of the group with the highest AF. The v4 browser label is "Grpmax Filtering AF" in the ClinGen description (www.clinicalgenome.org), and in the VCF it appears as the fafmax FAF95 fields.

What is the difference between popmax and grpmax?

Popmax in v2.1 was the highest frequency among non-bottlenecked populations. In v4 the vocabulary became genetic ancestry groups, and grpmax reports AC, AN, AF and homozygote count for the group with the highest AF. The Human Mutation review recommended popmax for filtering common variants (onlinelibrary.wiley.com). Neither popmax nor grpmax is a filtering allele frequency.

What does gnomAD v4 FAF95 mean versus FAF99?

They are the 95% and 99% confidence levels of the Poisson bound, both listed per group. A policy must record which level a threshold was derived for; the ClinGen communication describes the 95% version.

What are the ACMG BA1 and BS1 allele frequency thresholds?

The 2015 standards give BA1 as above 5% and BS1 as greater than expected for the disorder (pmc.ncbi.nlm.nih.gov). Disease-specific panels set lower values, for example 0.05% for RASopathy genes in a 2018 specification. No single number applies across diseases.

Does absence from gnomAD support pathogenicity?

By itself it is weak. PM2 was set to Supporting strength, and a review of gnomAD interpretation states that absence is consistent with, but far from sufficient evidence for, involvement in disease (pmc.ncbi.nlm.nih.gov).

How should coverage and callable sites affect interpretation, and how should low-coverage variants be curated?

Check whether the site was callable before reading absence as rarity: all-sites AN files record callable sites, and curators are asked to confirm sufficient read depth (pmc.ncbi.nlm.nih.gov). Treat low-coverage cases as escalations, since gnomAD states that frequency metrics in low-coverage or low-mappability regions may not be reliable.

How does gnomAD population frequency evidence enter variant classification?

It supplies BA1, BS1 and PM2 evidence, with thresholds and strengths set by the applicable specification and committee guidance. This report concerns research exclusion policy; clinical classification remains the responsibility of qualified professionals applying the relevant specifications.

Conclusion

A frequency number is only as meaningful as its definition, denominator and policy. AC and AN describe what was observed and how much was callable. FAF and GroupMax FAF convert that observation into a conservative statement about the true frequency in the most frequent non-bottlenecked ancestry group. A disease-specific maximum credible frequency, derived from prevalence, penetrance and heterogeneity, gives the comparison its meaning, and the published examples show those derivations differ widely between diseases.

The most useful design distinctions are between absent and uncallable, and between a frequency that is missing and one that is zero. Both are recoverable if AN files, flags and field semantics are captured at lookup time, and both are lost if only an AF column is stored. Escalation is the default whenever a prerequisite is unmet: no disease-specific policy, low coverage, discordant data types, bottlenecked groups, or a missing FAF.

Organizations adopting these practices should recheck the ClinGen and gnomAD documents at each refresh, pilot specification changes with domain experts, and avoid inferring blanket approval from any single committee communication. Specific diseases, thresholds and cases belong with qualified experts.

Source links

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

1. https://gnomad.broadinstitute.org/news/2026-03-gnomad-v4-1-1
2. https://gnomad.broadinstitute.org/news/changelog
3. https://onlinelibrary.wiley.com/doi/full/10.1002/humu.24309
4. https://gnomad.broadinstitute.org/help/faf
5. https://www.clinicalgenome.org/site/assets/files/9445/june_2025_communication_to_clingen_vceps_from_clingen_vcep_review_committee.pdf
6. https://pmc.ncbi.nlm.nih.gov/articles/PMC4544753
7. https://pmc.ncbi.nlm.nih.gov/articles/PMC5876064
8. https://pmc.ncbi.nlm.nih.gov/articles/PMC6119537
9. https://gnomad.broadinstitute.org/news/2024-04-gnomad-v4-1
10. https://clinicalgenome.org/site/assets/files/5182/pm2_-_svi_recommendation_-_approved_sept2020.pdf
11. https://samtools.github.io/hts-specs/VCFv4.5.pdf
12. https://www.nature.com/articles/gim201726
13. https://www.nature.com/articles/nature19057
14. https://hail.is/docs/0.2/experimental/index.html
15. https://gnomad.broadinstitute.org/news/2018-10-gnomad-v2-1
16. https://raw.githubusercontent.com/broadinstitute/gnomad_methods/main/gnomad/utils/annotations.py
17. https://gnomad.broadinstitute.org/news/2023-11-genetic-ancestry
18. https://gnomad.broadinstitute.org/help/grpmax
19. https://discuss.gnomad.broadinstitute.org/t/grpmax-faf-value-in-gnomad-vcf-files/759
20. https://broadinstitute.github.io/gnomad_methods/api_reference/utils/annotations.html
21. https://www.genomicseducation.hee.nhs.uk/wp-content/uploads/2024/08/ACGS-2024_UK-practice-guidelines-for-variant-classification.pdf
22. https://pmc.ncbi.nlm.nih.gov/articles/PMC6188666
23. https://pmc.ncbi.nlm.nih.gov/articles/PMC5563454
24. https://www.nature.com/articles/s41436-018-0100-y
25. https://pmc.ncbi.nlm.nih.gov/articles/PMC6188673

26. https://pipeline-rd-help.genomicsengland.co.uk/latest/variant-prioritisation-approaches/exomiser/exomiser-implementation
27. https://pmc.ncbi.nlm.nih.gov/articles/PMC5467691
28. https://www.nature.com/articles/gim201530
29. https://gnomad.broadinstitute.org/help
30. https://erepo.clinicalgenome.org/cspec/ui/svi/doc/1526222356
31. https://erepo.clinicalgenome.org/cspec/ui/svi/doc/1530970787
32. https://movementdisorders.onlinelibrary.wiley.com/doi/10.1002/mds.30066
33. https://www.biorxiv.org/content/10.1101/2022.03.24.485707v2.full
34. https://www.ensembl.org/info/docs/tools/vep/script/vep_options.html
35. https://discuss.gnomad.broadinstitute.org/t/discrepancies-in-an-values-for-multiallelic-sites/796
36. https://academic.oup.com/bioinformatics/article/31/13/2202/196142
37. https://samtools.github.io/bcftools/bcftools.html
38. https://hgvs-nomenclature.org/stable/recommendations/general
39. https://vrs.ga4gh.org/en/2.0/conventions/normalization.html
40. https://raw.githubusercontent.com/broadinstitute/gnomad_methods/main/gnomad/utils/vcf.py
41. https://discuss.gnomad.broadinstitute.org/t/groupmax-values-missing-for-a-lot-of-variants/250/6
42. https://pmc.ncbi.nlm.nih.gov/articles/PMC9160216
43. https://gnomad.broadinstitute.org/downloads
44. https://gnomad.broadinstitute.org/news/2023-11-gnomad-v4-0
45. https://pmc.ncbi.nlm.nih.gov/articles/PMC12579199
46. https://www.nature.com/articles/s41586-020-2308-7
47. https://vrs.ga4gh.org/en/2.0/conventions/computed_identifiers.html
48. https://vci-gci-docs.clinicalgenome.org/vci-gci-docs/gci-help/evidence-collection-genetic-evidence/curating-variant-information
49. https://va-spec.ga4gh.org/en/latest/va-standard-profiles/base-profiles/study-result-profiles.html
50. https://www.ncbi.nlm.nih.gov/clinvar/docs/faq
51. https://exomiser.readthedocs.io/en/stable/advanced_analysis.html
52. https://raw.githubusercontent.com/hail-is/hail/main/hail/python/hail/experimental/filtering_allele_frequency.py
53. https://www.nature.com/articles/s41586-023-06045-0

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.