Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ancestry-stratified frequency assessment of 50,634 ClinVar variants in 32 kidney-disease genes: a targeted West African comparison

Achab-Awelana Ernest

achabawelana@gmail.com

Kwame Nkrumah University of Science and Technology https://orcid.org/0000-0001-8411-7895

Lucy Binnah Owusu Ansah

Kwame Nkrumah University of Science and Technology

Research Article

Keywords: ClinVar, gnomAD v4.1, 1000 Genomes Project, kidney disease genetics, filtering allele frequency, ancestry-stratified analysis, variant classification, continental West Africa

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11762334/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Ancestry-stratified frequency assessment of 50,634 ClinVar variants in 32 kidney-disease genes: a targeted West African comparison

Research article | Secondary computational analysis of public human-genomic resources

Ernest Achab-Awelana[1],* | Lucy Binnah Owusu Ansah[1]

[1] Thrivus University for Biomedical Science and Technology, Ghana

* Correspondence: Achab-Awelana Ernest, Thrivus University for Biomedical Science and Technology, Ghana. E-mail: achabawelana@gmail.com

Abstract

Background: Ancestry-stratified population frequencies can inform clinical interpretation of kidney-disease variants, but reference-genome representation differs across ancestry groups. We assessed 50,634 ClinVar classifications in 32 kidney-disease-associated genes using gnomAD v4.1, model-derived gene-specific frequency ceilings, submission-level records, and selected variants from 421 continental West African reference genomes (Nigeria, Sierra Leone, and The Gambia).

Results: Among 29,214 ClinVar variants observed in gnomAD, five of 2,339 pathogenic/likely pathogenic variants (0.21%) and 846 of 3,211 conflicting variants (26.3%) exceeded the primary frequency ceiling in at least one reference group. The observed within-variant comparison among 3,208 conflicting variants was not significant (African/African American-only flags 240; non-Finnish European-only flags 231; exact McNemar P=0.712). Following 500 simulated allele-number equalizations, median discordant counts were 253 versus 160; this sensitivity result does not replace the observed comparison. Five flagged pathogenic/likely pathogenic variants materially influenced estimated allele load (modelled frequency ratio 7.24; 95% Poisson-based confidence interval 5.11–10.27). Expected rare unresolved coding-variant loads were 1.51 and 1.02 alleles per reference individual in the respective groups. In the selected West African subset, 90 of 99 flagged conflicting variants had also been flagged in gnomAD v4.1 African/African American data.

Conclusions: Ancestry-stratified frequencies highlight ClinVar kidney-gene classifications warranting condition-specific expert review. Findings depend on reference-group coverage and assumed disease models; selected West African genomes provide cross-resource concordance, not a representative population replication. Frequency flags do not independently establish pathogenicity, misclassification or clinical diagnostic risk.

Keywords: ClinVar; gnomAD v4.1; 1000 Genomes Project; kidney disease genetics; filtering allele frequency; ancestry-stratified analysis; variant classification; continental West Africa

Background

Genetic diagnosis is becoming part of kidney care. In a cohort of more than 3,000 patients with chronic kidney disease, exome sequencing identified a monogenic diagnosis in 9.3% [1], and gene panels for cystic, glomerular, and tubulointerstitial disease are now offered commercially worldwide. The value of a genetic result depends on how its variants are classified. In practice, laboratories classify variants using the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) framework [2] and consult ClinVar, the public archive of laboratory classifications [3]. Population allele frequency is the first and most powerful filter in that framework: a variant that is too common in healthy people cannot be a fully penetrant cause of a rare disease [2,4,5].

That filter is only as good as the reference populations behind it. People of African ancestry carry the greatest human genetic diversity yet remain under-represented in reference and clinical datasets [6]. In hypertrophic cardiomyopathy, variants once reported as pathogenic were later shown to be common benign variants in Black Americans, and patients of African ancestry had received misdiagnoses [7]. Similar ancestry-linked misclassification has been described for inborn errors of metabolism [8] and cancer genes [9], and for a cystinosis variant whose high frequency in African-ancestry groups would make it benign under ACMG frequency rules [10].

Kidney disease adds a specific tension. A series of 2026 population-sequencing studies from one group estimated the frequency of predicted pathogenic variants in kidney disease genes and repeatedly found the highest frequencies in African or African American participants: for example, 1 in 149 for congenital anomalies of the kidney and urinary tract (CAKUT) genes [11], and 1 in 43 versus 1 in 187 in Europeans for genes causing polycystic liver disease with kidney cysts [12]. These estimates were read mainly as evidence of undiagnosed disease [11–13]. An alternative explanation has not been tested: variants classified on European-ancestry evidence may be over-called as pathogenic in African genomes, while genuinely pathogenic African alleles remain uncatalogued. The two errors push in opposite directions and could mask each other.

Kidney-disease genetic interpretation is particularly relevant to populations historically underrepresented in reference datasets. Chronic kidney disease is a major global health challenge [14], but disease burden alone cannot establish the validity of variant classifications. This analysis used population reference datasets and ClinVar records without recruiting patients or assessing clinical outcomes; gnomAD AFR represents the African/African American ancestry category, and APOL1 risk genotypes were outside the Mendelian-variant screening framework applied here [15,16].

Our objectives were to identify aggregate ClinVar classifications that exceed specified disease-modelled frequency ceilings; compare the same variants across AFR and NFE with and without statistical sample-size equalization; estimate unresolved-allele loads in those reference groups; review available SCV submissions; and examine concordance for selected candidates in 1000 Genomes

continental West African reference samples. We did not estimate national kidney-disease prevalence, country-specific genetic risk, or real-world diagnostic accuracy.

Methods

Study design and data sources

This was a cross-sectional computational analysis of public ClinVar and gnomAD resources, with a targeted 1000 Genomes population-reference comparison; no participants were recruited. We studied 32 genes commonly included in kidney diagnostic panels and associated with cystic kidney disease, glomerulopathies, complement-mediated disease, tubulointerstitial disease and CAKUT (Supplementary Table S1). *MUC1* was excluded because its common disease-causing frameshift lies in a variable-number tandem repeat that short-read sequencing does not resolve.

For each gene, we retrieved all variants and all ClinVar variants on GRCh38 from the gnomAD browser application programming interface (gnomAD v4.1; accessed 15 November 2026). gnomAD v4.1 provides joint exome-plus-genome allele counts (AC) and allele numbers (AN) for each genetic ancestry group. The browser integrates ClinVar; the integrated release was dated 11 October 2026. We used the aggregate (variant-level) germline classification, review status and number of review stars. For conflicting variants, we retrieved the full ClinVar variation archive records through the National Center for Biotechnology Information E-utilities (accessed 15–16 November 2026) to obtain the classification from each individual submission (SCV).

To test whether findings based on gnomAD's African/African American group (predominantly African American) hold in continental West Africans, we queried the high-coverage 1000 Genomes Project genomes [17] distributed with gnomAD v3.1.2 as part of the harmonised HGDP+1KG resource [18] (GRCh38). Four West African populations were pooled as a continental West African group (WAF): Yoruba in Ibadan and Esan in Nigeria (YRI, ESN), Mende in Sierra Leone (MSL) and Gambian in Western Division (GWD). Luhya in Webuye, Kenya (LWK) served as an East African comparator, and African Caribbean (ACB) and African American (ASW) samples as diaspora comparators. Because the gnomAD interface returns these subsets only through single-variant queries, we queried (i) every P/LP and conflicting variant observed in gnomAD v4.1 and (ii) a seeded random sample of uncertain (VUS) variants, restricted in both cases to variants observed at least once in the gnomAD v4.1 African/African American group. In total, 4,130 variants were queried, including 489 of the 3,589 VUS present in that group. This restriction was checked empirically: none of 333 P/LP variants absent from gnomAD v4.1 AFR and queried before applying this restriction was detected in the selected continental West African genomes. This preliminary observation does not establish the absence of West African-specific variation outside the selected variant list. For computational comparison, missing subset frequencies were imputed as zero counts at the median allele number; this assumption

may misclassify uncalled sites as true non-observations and is a limitation of the exploratory analysis. VUS contributions to per-person estimates were weighted by the inverse of the sampling fraction.

We analysed eight gnomAD genetic ancestry groups: African/African American (AFR), Admixed American (AMR), Ashkenazi Jewish (ASJ), East Asian (EAS), Finnish (FIN), Middle Eastern (MID), non-Finnish European (NFE) and South Asian (SAS). The "remaining" and Amish groups were excluded. Following gnomAD practice for filtering allele frequencies, frequency-based flags used only the five non-founder groups (AFR, AMR, EAS, NFE and SAS).

Variant classes and quality control

Aggregate classifications were grouped as P/LP (Pathogenic, Likely pathogenic, or both); uncertain significance (VUS); conflicting classifications of pathogenicity; benign/likely benign (B/LB); and other (for example, drug response). P/LP entries labelled "low penetrance", "risk allele" or "risk factor" (n=155) were excluded from the primary P/LP analyses because frequency ceilings for fully penetrant disease do not apply to them; the same caveat applies to alleles whose effect depends on the allele in trans, such as *NPHS2* p.Arg229Gln [19]. Frequencies were used only for variants passing all gnomAD joint quality filters. They were used in a given group only when that group's AN was at least 50% of its median AN across passing variants, which removes poorly covered sites. Coding variants were defined as missense, protein-truncating, in-frame, start- or stop-altering, or splice-region variants.

Maximum credible allele frequency

We adapted the maximum credible allele-frequency framework [4] for one modelled disease per gene, with an alternative recessive model where relevant (Supplementary Table S1). Prevalence, penetrance and maximum allelic contribution were prespecified in the supplied analysis code and varied in sensitivity analyses. The resulting cutoffs are intentionally permissive screening assumptions, not validated ACMG/AMP BA1/BS1 thresholds for every variant–condition pair. Frequency exceedance identifies a candidate for re-evaluation, not an independently demonstrated error in a submitted clinical classification [2,5,24,25].

- dominant: (prevalence ÷ 2) × maximum allelic contribution ÷ penetrance;
- recessive: √(prevalence ÷ penetrance) × maximum allelic contribution;
- X-linked: prevalence × maximum allelic contribution ÷ penetrance.

Genetic heterogeneity was set to one, making each gene potentially responsible for all cases in the relevant model. Generic rare-disease upper prevalence was 1/10,000 at penetrance 0.5, with disease-specific exceptions documented in Supplementary Table S1. For several complement-related susceptibility genes penetrance 0.1 was assumed; such susceptibility and hypomorphic alleles should not be interpreted as fully penetrant Mendelian causes. Seven genes were additionally evaluated using a recessive model, and the larger of the dominant and recessive ceilings was used where applicable. Published population-sequencing studies of cystic kidney disorders and collagen IV-associated disease

informed selected prevalence assumptions [20,21]. Primary maximum allelic contribution was 0.1, with a permissive sensitivity setting of 1.0. These are model choices and should not be interpreted as measured gene-specific penetrances or allele contributions.

Frequency flags

For each variant and group, we computed the filtering allele frequency: the one-sided 95% lower Poisson confidence bound of the allele frequency, calculated as $\chi^2_{0.05}(2 \cdot AC)/2$ divided by AN. A variant was flagged as frequency-implausible in a group when this lower bound exceeded the gene's ceiling. Because the bound is computed within each group, a smaller group (such as AFR, AN about 75,000) needs proportionally stronger evidence than a large group (NFE, AN about 1.18 million) to be flagged, which biases comparisons against detecting excess flags in AFR. To remove any residual effect of sample size, we repeated the NFE analysis 500 times after binomially downsampling NFE allele counts to the median AFR AN.

Submission-level analysis of conflicting variants

For every variant with an aggregate conflicting classification, and for every P/LP variant observed in gnomAD, we retrieved each underlying submission's germline classification, review status, date last evaluated (or, if absent, date of creation), submitter, collection method and asserted condition. For conflicting variants we counted submissions by classification. We recorded whether at least one laboratory had submitted P or LP, whether at least one had submitted B or LB, and whether both had (a direct pathogenic-versus-benign conflict). We then cross-tabulated these submission profiles against the frequency flags. We compared the dates, review status and collection method of P/LP submissions on frequency-implausible variants with those of P/LP submissions on frequency-compatible P/LP variants, and compiled a case-level table of every frequency-implausible variant carrying at least one P/LP submission. Submission retrieval and flag computation for conflicting variants were carried out in a second, independent implementation (JavaScript in the browser), which reproduced every frequency-flag count obtained in the primary Python pipeline.

Ancestry predominance and classification outcome

Among rare (frequency below 1% in every non-founder group), coding, classified variants observed in gnomAD, each variant was assigned to the group with the highest allele frequency if that frequency was at least five times the next-highest. Otherwise it was labelled "shared". We modelled the odds that a variant remained unresolved (VUS or conflicting) rather than resolved (P/LP or B/LB) using logistic regression with gene fixed effects and variant consequence (missense, protein-truncating, other). Models were fitted without a frequency covariate, with $\log_{10}$ global frequency, and with $\log_{10}$ maximum group frequency. Because the much larger European sample detects far rarer variants, we refitted the global-frequency model 100 times after downsampling NFE counts to the AFR sample size before assigning predominance.

Per-person burden and carrier frequency

For each genetic ancestry group, we estimated the expected number of rare (group frequency below 1%) classified coding-variant alleles per reference individual by summing 2×AC/AN across sites, separately for P/LP, VUS, conflicting and B/LB classes. We report this quantity as a modelled allele load, not the observed number of test results in clinical patients. Approximate P(at least one allele)=1−exp(−load) assumes a Poisson distribution; correlation between variants, phasing, diploidy, coverage differences, and selection of source genotypes are not fully represented. Dominant- and recessive-gene summaries describe expected alleles carrying catalogued P/LP labels, not monogenic disease prevalence. Poisson-based confidence intervals quantify counting variability but do not capture disease-model, variant-calling, and reference-population uncertainty. We also compared modelled unresolved load before and after the hypothetical exclusion of frequency-flagged variants; this is not equivalent to their clinical reclassification.

Statistical analysis

Three principal descriptive estimands were defined in the revised analysis: (P1) within-variant AFR-versus-NFE discordance in model-derived frequency flags, (P2) modelled P/LP-labelled allele-load ratios before and after excluding flagged variants, and (P3) expected rare unresolved-allele load. This secondary analysis was not preregistered; all significance tests are exploratory and not multiplicity-adjusted. The observed paired analysis used 3,208 conflicting variants with adequate allele numbers in both groups and an exact McNemar test. To explore unequal detection power, NFE allele counts were binomially downsampled to the median AFR allele number in 500 independent replicates; medians, intervals and per-draw paired tests describe this simulated counterfactual and must not be interpreted as independent population observations. The sign test for per-gene direction used downsampled median discordances and is likewise sensitivity evidence. Comparisons of summed allele counts used approximate Poisson-rate methods; these do not account for correlated variants or individual-level linkage. No clinical misdiagnosis endpoint was measured.

Sensitivity analyses and software

Maximum allelic contribution was varied (0.05, 0.1, 0.25, 0.5, 1.0) together with prevalence multipliers of 1 or 10, yielding ten sensitivity scenarios. Further analyses omitted complement susceptibility genes and NPHS2. A subset of the core variant-level results and all 500 NFE downsampling replicates were independently recalculated from the supplied raw export for this revision; the corrected reconciliation is provided in the reproducibility addendum. The original package reports use of Python 3.13, pandas 3.0, SciPy 1.18 and matplotlib 3.11, alongside JavaScript retrieval. ClinVar uses condition-specific SCV/RCV and variant-level VCV aggregates, which may differ [24,25]. STROBE principles were followed where applicable [22]. Gene-specific models and all reported supplementary analyses are supplied in Additional file 1 (Tables S1–S11); reusable scripts, derived variant-level data and independent audit files are provided in Additional file 2.

Generative-AI tools were used during drafting and analysis support: Anthropic Claude assisted with data retrieval, analysis-code preparation, figure generation and drafting, and OpenAI ChatGPT assisted with reproducibility auditing, methodological clarification and language editing. These systems were not authors; the named authors retain responsibility for all analyses, citations, interpretations and the final manuscript.

Ethics

The analyses used public, de-identified, summary genomic data and ClinVar assertions; no patients were recruited and no individual-level clinical outcomes were obtained. Applicable institutional and journal requirements for a determination of ethics-exemption status should be confirmed by the authors.

Results

Dataset

The 32 genes carried 50,634 ClinVar variants: 11,603 P/LP (including 155 low-penetrance or risk entries), 19,343 VUS, 3,734 conflicting, 15,874 B/LB and 80 other (Supplementary Table S1). Most aggregate classifications had one review star (36,516; 72.1%). There were 11,576 with two stars and 2,542 with none; no variant in these genes carried an expert-panel (three-star) or practice-guideline classification. Overall, 29,214 variants (57.7%) passed gnomAD quality filters and were observed at least once. The median allele number was 74,914 in AFR (about 37,500 individuals) and 1,179,532 in NFE (about 590,000 individuals) (Figure 1).

Rare P/LP frequency flags and limits of the ancestry comparison

Among 2,339 observed P/LP variants after excluding 155 low-penetrance/risk-labelled entries, five (0.21%) exceeded the principal gene-level ceiling in at least one group (Table 2); four had zero or one ClinVar review star. The observed group-specific flag proportions were 4/369 in AFR and 3/1,692 in NFE, but these are overlapping variant sets and should not be treated as independent samples. Across 2,339 P/LP variants assessable in both groups, the raw paired discordance was one AFR-only versus zero NFE-only flags (exact P=1.00). After NFE downsampling, median discordances were three versus zero, with similarly limited information. WT1 p.His474Tyr remained flagged under the most permissive model, warranting detailed gene-, phenotype- and allele-specific evaluation rather than automatic reclassification.

Table 1. Frequency audit and modelled allele-load estimates by gnomAD genetic ancestry group

Group	Median allele number	P/LP frequency-implausible / observed (%)	Conflicting frequency-implausible / observed (%)	Conflicting implausible, maximally lenient model (%)	Rare unresolved variants per person	Modelled dominant-gene P/LP carrier equivalent	Modelled recessive-gene P/LP carrier equivalent
African/African American	74,914	4/369 (1.08%)	336/1,802 (18.6%)	80 (4.4%)	1.51	1 in 126	1 in 70
Admixed American	59,984	1/280 (0.36%)	243/1,649 (14.7%)	75 (4.5%)	1.13	1 in 137	1 in 74
East Asian	44,860	3/221 (1.36%)	163/1,025 (15.9%)	50 (4.9%)	1.00	1 in 92	1 in 71
Non-Finnish European	1,179,532	3/1,692 (0.18%)	328/2,607 (12.6%)	76 (2.9%)	1.02	1 in 136	1 in 44
South Asian	91,038	2/360 (0.56%)	222/1,423 (15.6%)	50 (3.5%)	1.16	1 in 144	1 in 69

Primary model: maximum allelic contribution 0.1; maximally lenient model: 1.0. P/LP excludes low-penetrance and risk entries. Unresolved = uncertain significance or conflicting classifications with frequency below 1% in the individual's own group, summed over 32 genes. Modelled P/LP carrier equivalents include all P/LP variants (see Figure 3 and Supplementary Table S3 for restricted sets and confidence intervals).

Table 2. ClinVar pathogenic/likely pathogenic variants whose frequency exceeds the maximum credible allele frequency

Gene	Variant	ClinVar classification (review stars)	AFR AF	NFE AF	Highest other AF	Ceiling	Implausible in	Also at maximally lenient model
ACTN4	c.2338-5G>A	Likely pathogenic (1★)	5.3e-05	2.1e-05	4.4e-05 (SAS)	1e-05	AFR+NFE+SAS	No
LMX1B	c.919+13_919+51del	Likely pathogenic (0★)	6.2e-05	4.5e-05	7.7e-05 (EAS)	1e-05	AFR+EAS+NFE	No
PAX2	c.685+2T>C	Pathogenic/Likely pathogenic (2★)	0.0e+00	2.9e-06	2.3e-04 (EAS)	2e-05	EAS	No
TRPC6	c.2534T>G (p.Leu845Arg)	Likely pathogenic (0★)	1.3e-04	0.0e+00	0.0e+00 (AMR)	1e-05	AFR	No
WT1	c.1420C>T (p.His474Tyr)	Likely pathogenic (1★)	4.3e-04	3.4e-05	2.1e-04 (EAS)	1e-05	AFR+AMR+EAS+NFE+SAS	Yes

AF, allele frequency (gnomAD v4.1 joint). Ceiling, maximum credible allele frequency under the primary model. AFR, African/African American; AMR, Admixed American; EAS, East Asian; NFE, non-Finnish European; SAS, South Asian. Variant nomenclature follows the canonical transcript reported by gnomAD.

Modelled P/LP allele loads are sensitive to a small number of flagged variants

For dominant-disease genes, the estimated P/LP-labelled allele load corresponded to an approximate modelled carrier equivalent of one per 126 AFR and 136 NFE reference individuals (approximate ratio 1.07; 95% Poisson-based CI 0.95–1.21; Figure 3A). Five flagged variants contributed 46 of 291 counted AFR alleles and 100 of 4,299 NFE alleles (15.8% and 2.3%). Excluding them changed the estimated AFR:NFE ratio to 0.92 (0.81–1.04); requiring at least two review stars yielded 0.89 (0.76–1.03). The flagged-allele load ratio itself was 7.24 (5.11–10.27), based on 46 versus 100 allele counts with very different reference denominators. These ratios are sensitive to variant-specific observations and are not estimates of clinical carrier status or disease risk in any population. For WT1 and TRPC6, a single flagged variant drove the within-gene difference (Supplementary Table S2).

In genes assigned recessive disease models, the expected load of catalogued P/LP-labelled alleles was lower in AFR than NFE (approximately one per 70 versus one per 44; ratio 0.62, 95% Poisson-based CI 0.57–0.68; Figure 3B). This observation may reflect ancestry-specific allele spectra, ascertainment and catalogue composition; it does not establish that true pathogenic recessive variants are less

prevalent in AFR or that variants absent from ClinVar are pathogenic. The most frequent NFE alleles, including NPHS2 p.Arg138Gln and PKHD1 p.Thr36Met, accounted for 26% of the estimated recessive-labelled load. X-linked COL4A5 analyses require additional sex-aware interpretation and are secondary. See Supplementary Table S3.

Conflicting classifications and ancestry-specific frequency evidence

Of 3,211 conflicting variants observed in gnomAD, 846 (26.3%) exceeded the principal frequency ceiling in at least one group, and 248 (7.7%) remained flagged under the maximally permissive model. Observed group counts were 336/1,802 in AFR and 328/2,607 in NFE, but variants overlap between groups. Among 3,208 variants assessable in both groups, unmodified paired discordances were AFR-only 240 and NFE-only 231 (exact McNemar P=0.712). After binomially downsampling NFE to the median AFR allele number in 500 replicates, median discordances were 253 versus 160 (approximate 95% simulation ranges AFR-only 248–258; NFE-only 149–170). The exact paired P-value was below 0.001 in each of these modelled replicates, although these are simulation-conditional tests rather than a primary inference about underlying biological differences. At the gene level, the original data favored AFR in 16 genes and NFE in nine (seven ties), whereas downsampled median discordances favored AFR in 23 and NFE in six (two ties; Supplementary Table S4). The distinction is essential: the stronger imbalance appears only under a sample-size-equalization model. Excluding complement-related genes and NPHS2 gave descriptive flag proportions of 18.3% in AFR and 12.5% in NFE. AFR-only flags across the five examined non-founder ancestry groups numbered 178, including 53 under the most permissive model. All represent curation candidates, not automatically resolved variants (Supplementary Table S5).

ClinVar submissions underlying the 3,734 variant-level conflicting aggregates (16,538 SCV records) were predominantly VUS-versus-B/LB (2,601 aggregates) or P/LP-versus-VUS (1,079), with 50 direct P/LP-versus-B/LB aggregate conflicts. Among assessable subsets, frequency ceilings were exceeded in 818/2,525 VUS-versus-B/LB, 8/634 P/LP-versus-VUS and 20/49 direct P/LP-versus-B/LB conflicts (Figure 4; Supplementary Table S6). Twenty-eight flagged conflicting aggregates retained at least one P/LP submission (76 such submissions). NPHS2 p.Arg229Gln exemplifies why conditional pathogenicity must be reviewed: the relevant disease phenotype can depend on the allele in trans [19]. Of the 178 AFR-only flag candidates, 175 had no P/LP submission. These records predominantly motivate re-evaluation of uncertainty; an ACMG/AMP benign classification requires the full clinical, phenotypic and allele-specific evidence assessment [2,5].

Submission-level characteristics of frequency-flagged P/LP calls

Thirty-three flagged variants retained at least one P/LP submission (five aggregate P/LP and 28 conflicting); there were 82 associated P/LP submission records, compared with 7,944 submissions on frequency-compatible aggregate P/LP variants (Supplementary Table S7). The flagged group had more pre-2018 assertions (14.6% versus 8.2%) and more submissions lacking assertion criteria (20.7%

versus 12.3%); these unadjusted comparisons are descriptive because multiple SCVs can pertain to the same variant and are not independent. Moreover, 85% of flagged submissions were dated 2018 or later. Case review identified distinct disease assertions and methods, including literature-only records and non-renal conditions; this reinforces the need to audit VCV-level versus condition-specific RCV/SCV conclusions before any clinical action. NPHS2 p.Arg229Gln (19 P/LP submissions) and PKHD1 p.Val836Ala (15 submissions) were notable examples. No variant was adjudicated pathogenic or benign by this analysis.

Robustness to model assumptions

The AFR excess of frequency-implausible conflicting classifications was present in all 10 parameter scenarios (Figure 5; Supplementary Table S8). It ranged from 23.4% vs 17.4% at a maximum allelic contribution of 0.05 to 0.8% vs 0.3% when prevalence bounds were inflated tenfold and a single variant was assumed to explain every case. For P/LP variants, AFR flags were at least as frequent as NFE flags in every scenario, and NFE flags fell to zero in 7 of 10 scenarios whereas AFR flags fell to zero in only 2.

Modelled burden of unresolved variant classifications

Among 17,502 rare coding variants considered for predominance analyses, AFR-predominant variants showed more unresolved aggregate labels before accounting for unequal detection power. The gene- and consequence-adjusted association did not persist in the NFE-downsampling sensitivity analysis (odds ratio 0.98, simulation interval 0.89–1.06; Supplementary Table S9), indicating ascertainment sensitivity. We do not treat ancestry predominance as an intrinsic cause of classification uncertainty.

A separate reference-frequency model estimated 1.51 rare unresolved coding-variant alleles per AFR reference individual (0.41 VUS and 1.10 conflicting) versus 1.02 in NFE (ratio 1.48; Figure 6; Supplementary Table S10). These totals are sums of allele frequencies, not observed patient-level genetic test yields. Under the explicitly hypothetical Poisson approximation, the probability of carrying one or more such alleles was approximately 78% versus 64%. Removing model-flagged variants from the sums decreased the expected loads to approximately 0.72 and 0.47, respectively. This is a sensitivity bound, not evidence that the excluded variants could be clinically reclassified or that reanalysis would achieve these diagnostic reductions.

These outputs depend on the set of variants catalogued and detected in each source dataset, quality and allele-number restrictions, the specified rare-frequency cutoff, and assumed independence. Per-person loads are not necessarily directly comparable with a prospective kidney-disease patient cohort. The observed proportion of rare coding variant alleles with no ClinVar record was similar in the referenced groups, but that alone cannot establish equal clinical interpretability or catalogue completeness (Supplementary Table S10).

Table 3 presents illustrative expected values per 1,000 reference genomes based on aggregate allele-frequency models, not observed clinical testing among patients in any country. The estimates depend on the panel, population sampling, variant coverage, classification rules and Poisson assumptions.

Exploratory comparison with continental West African reference genomes

The supplemental resource contained 421 high-coverage continental West African genomes drawn from Yoruba, Esan, Mende and Gambian populations (Supplementary Table S11). This was a targeted follow-up of variants initially selected in part on their presence in gnomAD v4.1 AFR, rather than an independently ascertained population-wide survey. Among 244 conflicting variants observed in this selected subset, 99 exceeded the modelled WAF ceiling; 90 of those had also been flagged in v4.1 AFR. Because detection and site inclusion differed between releases, this is best understood as cross-resource concordance for selected variants, not independent replication of a population-level rate.

The approximate expected unresolved-allele load in the selected WAF analysis was 1.32 (variant-resampling interval 1.08–1.55), compared with 1.44 (1.29–1.60) estimated using the same v3.1.2 method in its broader AFR reference group. The WAF VUS component required inverse-probability weighting from 489 sampled VUS variants, while unlisted variants were treated as zero at representative allele number. The resulting uncertainty interval does not fully represent selection, callability, genomic structure or weighting uncertainty. The WAF recessive P/LP allele-load comparison with v4.1 NFE was imprecise (rate ratio 0.41, 95% CI 0.16–1.10). A single P/LP-labelled variant flagged in WAF (PAX2 c.44-2A>G, observed once) illustrates the limited ability of n=421 to verify rare P/LP signals.

These selective population-reference comparisons do not measure individual genotypes, kidney-disease incidence, diagnostic error or panel performance in clinical populations. They show that some variants meeting frequency-flag criteria in gnomAD AFR also meet those criteria in a small, selected set of continental West African genomes. Representative, phenotype-linked datasets are still needed to test clinical relevance.

Table 3. Illustrative modelled outcomes per 1,000 reference genomes for a 32-gene panel; not observed patient test results or expected clinical diagnoses

Measure	gnomAD v4.1 African/African American	Continental West African (1000 Genomes; gnomAD v3.1.2)	gnomAD v4.1 non-Finnish European
Individuals (median)	37,457	421	589,766
Conflicting variants implausible / observed in group	336 / 1,802 (18.6%)	99 / 244 (40.6%)	328 / 2,607 (12.6%)
Rare unresolved variants per person	1.51	1.32 (1.08–1.55)	1.02
Expected per 1,000: ≥1 rare unresolved variant	779	733	640
Expected per 1,000: ≥1 unresolved variant after frequency-based resolution	512	511	375
Expected per 1,000: carrier of a frequency-implausible P/LP variant	1.4	— (too few carriers)	0.2
Expected per 1,000: dominant-gene P/LP carriers	7.9	7.1 (1.5–20.8)	7.4
Expected per 1,000: recessive-gene P/LP carriers	14.3	9.5 (2.6–24.3)	22.9

Continental West African: Yoruba and Esan (Nigeria), Mende (Sierra Leone), Gambian (The Gambia). West African estimates use all P/LP and conflicting variants and a random sample of uncertain variants (weighted), restricted to variants observed in gnomAD African/African American individuals; 95% intervals are bootstrap (burden) or exact Poisson (carriers). "Expected per 1,000 reference individuals" values are modelled from allele counts and assume Poisson counts ($1 - e^{-\lambda}$). Unresolved = uncertain or conflicting, coding, frequency below 1% in the group. Frequency-based resolution removes variants exceeding the maximum credible allele frequency in the same group. Proportions of implausible conflicting variants are not directly comparable across columns, because a small sample detects mainly the more common variants.

Discussion

Principal findings

This study links two forms of uncertainty in kidney-variant interpretation: model-derived frequency exceedance of aggregate ClinVar labels and ancestry-dependent coverage in population reference data. The raw paired AFR–NFE analysis of conflicting variants did not show convincing flag asymmetry (P=0.712), while a 500-replicate sample-size-equalization experiment consistently produced larger AFR-only discordance. A small number of frequency-flagged P/LP variants materially changed estimated P/LP-labelled allele loads, but that does not validate any particular clinical reclassification. The WAF follow-up showed partial concordance for selected variants only. These findings support structured expert re-review and improved African genomic representation, not evidence that ClinVar data establish clinical misdiagnosis or its prevalence in any population.

What this means for population-frequency estimates of genetic kidney disease

Population-frequency screens are often used to prioritize variant evidence under disease-specific assumptions [4,5]. In the present panel, removing five P/LP-labelled variants was sufficient to change the direction of a modelled dominant-gene AFR:NFE allele-load comparison; the underlying estimates nevertheless had overlapping confidence intervals. The recessive-labelled allele-load contrast could reflect differences in allele spectra, reference ascertainment and the variants included in ClinVar; under-cataloguing remains a plausible hypothesis rather than a quantified fact. Disease-level inference

requires pairing each genotype with the specific phenotype, inheritance model, penetrance and allelic configuration. The gnomAD v4.1 and v3.1.2 comparisons are not interchangeable population samples.

Frequency flags warrant curation, not automatic resolution

The 846 conflicting variant-level aggregates crossing the modelled ceiling are opportunities for systematic reassessment of evidence, particularly when disagreements are predominantly between VUS and benign/likely benign submissions. A threshold computed from a general gene model does not identify the correct clinical significance for every asserted condition, and the original evidence can involve reduced penetrance, risk alleles, nonrenal conditions or pathogenicity conditional on a second allele. ClinVar distinguishes variant-level VCV records from variant–condition RCV records and individual SCV submissions [24], and published studies have examined diverse sources of disagreement among clinical classifications [23]. ClinGen recommends ancestry-aware, disease-specific population-frequency evaluation with explicit quality safeguards [5,25]. We therefore describe the 178 AFR-only candidates as flagged for review, not as 178 classifications already resolved. Case-level curation should prioritize variants with low review status and marked frequency elevation while retaining phenotype and functional evidence.

Implications for ancestry-aware kidney-variant interpretation

The findings are relevant to laboratories interpreting kidney-disease variants for patients whose ancestry is incompletely represented in population-reference resources. The gnomAD AFR category contains substantial African-diaspora representation, while the 421 continental West African genomes analysed here came from populations in Nigeria, Sierra Leone and The Gambia; neither resource represents the complete diversity of continental West African populations. Frequency thresholds should be applied with disease-specific inheritance and penetrance assumptions, adequate ancestry-stratified allele numbers, and review of the exact condition asserted in ClinVar. A negative or uncertain panel result cannot be converted into an ancestry-specific clinical prediction from these reference data alone. APOL1 G1/G2 risk genotypes require a different interpretive framework from the rare Mendelian frequency ceilings applied here [15].

Strengths and limitations

The main strengths are a variant-level dataset of 50,634 ClinVar entries, transparent disease-model thresholds, SCV-level case review, analysis of overlapping variant sets, a ten-scenario sensitivity grid, and an available derived-data/code package. An independent audit of the supplied extraction reproduced the 50,634-row/32-gene denominator, the 29,214 observed variants, and the five P/LP and 846 conflicting frequency flags. Re-execution of the 500-draw downsampling exercise produced the reported direction and median counts, although it cannot resolve structural limitations of the study design.

Important limitations remain. First, no patient-level kidney phenotypes or clinical outcomes were available; the targeted continental West African follow-up comprised reference genomes from Nigeria, Sierra Leone and The Gambia and only a selected set of variants. Second, aggregate VCV labels can merge classifications made for different conditions; disease-gene model parameters are approximations, not variant-specific evidence adjudications. Third, the stronger paired asymmetry arises in a downsampling counterfactual, not in the observed paired data; simulation P-values are conditional on the chosen resampling model, and the analysis was not preregistered. Fourth, rare-variant allele-load estimates use sums of AC/AN and Poisson assumptions that do not capture phasing, linkage, clinical ascertainment, sex-specific X-linked effects or individual penetrance. Fifth, the WAF sample was small, variants were preselected on v4.1 AFR presence, VUS were subsampled and missing v3.1.2 subset sites were treated as zero; this does not constitute an independent representative population replication. Sixth, source releases, callability and structural/complex regions (including PKD1) differ. Seventh, clustered submission records and exploratory multiple testing limit inferential claims. Finally, frequency flags identify candidates for review rather than proven errors, and the reported confidence intervals understate uncertainty from gene models and data selection.

Practical considerations for clinical genetics laboratories

- Use high-quality ancestry-stratified filtering allele frequencies together with explicit coverage, allele count, reference release and quality-control documentation; do not substitute a broad ancestry label for an individual clinical assessment.
- Apply disease-specific ClinGen/VCEP criteria where available, distinguishing dominant, recessive, X-linked, susceptibility, reduced-penetrance and allele-dependent mechanisms. Treat this study's gene-level thresholds as screening hypotheses rather than clinical classification rules.
- Prioritize expert review of P/LP-labelled and conflicting variants exceeding a plausible ceiling, checking the asserted RCV condition, raw SCV evidence, allele configuration, phenotype, segregation, functional support and previous laboratory assessments.
- When reporting an uncertain classification, state the population-reference limitations and provide a policy for periodic reassessment; do not imply that model-based exclusion is equivalent to clinical resolution.
- Support representative, consented reference-genomic and phenotype data generation in underrepresented populations, including continental West African communities; document fine-scale ancestry and encourage expert-reviewed ClinVar submissions when scientifically justified.
- Treat APOL1 risk testing separately from rare monogenic variant classification, with a validated indication, consent framework and interpretation appropriate to the available evidence.

Conclusions

In this computational assessment of 50,634 ClinVar classifications across 32 kidney-disease-associated genes, five observed P/LP variants and a larger subset of conflicting classifications crossed permissive gene-modelled population-frequency ceilings. The unmodified paired gnomAD AFR–NFE comparison did not show statistically convincing asymmetry; equal-allele-number simulation suggested differences requiring cautious interpretation. The expected unresolved-allele and P/LP-labelled allele loads differed across reference groups, but these are modelled quantities, not observed clinical outcomes or national prevalence estimates. A targeted 421-genome 1000 Genomes comparison drawn from Nigeria, Sierra Leone and The Gambia supported selected-variant flag concordance without establishing representative West African population estimates. Expert condition-specific curation and more representative reference data are required before clinical translation.

List of abbreviations

AC, allele count; ACMG/AMP, American College of Medical Genetics and Genomics/Association for Molecular Pathology; AFR, African/African American; AN, allele number; B/LB, benign/likely benign; ClinVar, clinical variant archive; FAF, filtering allele frequency; gnomAD, Genome Aggregation Database; HGDP, Human Genome Diversity Project; NFE, non-Finnish European; P/LP, pathogenic/likely pathogenic; RCV, variant–condition classification record; SCV, submitted classification record; VCV, variant-level aggregate classification; VUS, variant of uncertain significance; WAF, continental West African reference group.

Declarations

Ethics approval and consent to participate

Not applicable. The study involved secondary analysis of publicly available, de-identified aggregate genomic reference data and public variant classifications; no participants were recruited or identifiable records accessed.

Consent for publication

Not applicable. No identifiable individual participant information is presented.

Availability of data and materials

The datasets supporting the findings are included in the article and its additional files. The publicly available source resources are ClinVar (snapshot integrated into the gnomAD browser dated 11

October 2026), gnomAD v4.1, and the gnomAD v3.1.2 HGDP+1KG subset. Supplementary Tables S1–S11 are supplied as Additional file 1 (xlsx). Derived data, code, and audit outputs are supplied as Additional file 2 (zip). The underlying public resources and their release details are cited in the Methods and reference list.

Competing interests

The authors declare that they have no competing interests.

Funding

No external funding was received for this study.

Authors' contributions

AAE conceived and designed the study, developed the methods and software, conducted the investigation and formal analysis, validated findings, curated data, prepared visualizations, managed the project, drafted the manuscript, and performed critical review and editing. LBOA contributed to data curation, literature searching, evidence review, and critical revision of the manuscript.

Acknowledgements

The authors acknowledge ClinVar submitters, the gnomAD consortium and the 1000 Genomes Project contributors for making public genomic data available.

References

1. 1. Groopman EE, Marasa M, Cameron-Christie S, et al. Diagnostic utility of exome sequencing for kidney disease. N Engl J Med. 2019;380:142–151. doi:10.1056/NEJMoa1806891
2. 2. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17:405–424. doi:10.1038/gim.2015.30
3. 3. Landrum MJ, Chitipiralla S, Kaur K, et al. ClinVar: updates to support classifications of both germline and somatic variants. Nucleic Acids Res. 2025;53:D1313–D1321. doi:10.1093/nar/gkae1090
4. 4. Whiffin N, Minikel E, Walsh R, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med. 2017;19:1151–1158. doi:10.1038/gim.2017.26

5. 5. Ghosh R, Harrison SM, Rehm HL, Plon SE, Biesecker LG. Updated recommendation for the benign stand-alone ACMG/AMP criterion. Hum Mutat. 2018;39:1525–1530. doi:10.1002/humu.23642

6. 6. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538:161–164. doi:10.1038/538161a

7. 7. Manrai AK, Funke BH, Rehm HL, et al. Genetic misdiagnoses and the potential for health disparities. N Engl J Med. 2016;375:655–665. doi:10.1056/NEJMsa1507092

8. 8. Sharo AG, Zou Y, Adhikari AN, Brenner SE. ClinVar and HGMD genomic variant classification accuracy has improved over time, as measured by implied disease burden. Genome Med. 2023;15:51. doi:10.1186/s13073-023-01199-y

9. 9. Pollard RD, Wilkerson MD, Rajagopal PS. Identification of germline population variants misclassified as cancer-associated somatic variants. Front Med (Lausanne). 2024;11:1361317. doi:10.3389/fmed.2024.1361317

10. 10. Wu CW, Tomaszewski A, Stark L, et al. Perspectives from cystinosis: access to healthcare may be a confounding factor for variant classification. Front Genet. 2024;15:1402667. doi:10.3389/fgene.2024.1402667

11. 11. Huang M, Savige J. The population frequency of predicted pathogenic genetic variants in commonly affected CAKUT genes in the general population. Pediatr Nephrol. 2026;41:3519–3530. doi:10.1007/s00467-026-07324-4

12. 12. Varughese S, Huang M, Savige J. The population frequency of predicted pathogenic variants in the genes associated with autosomal dominant polycystic liver disease (ADPLD) and kidney cysts. PLoS One. 2026;21:e0348109. doi:10.1371/journal.pone.0348109

13. 13. Varughese S, Huang M, Savige J. Typical and atypical ADPKD: predicted pathogenic genetic variants and population frequencies. Nephrol Dial Transplant. 2026;41:286–296. doi:10.1093/ndt/gfaf158

14. 14. GBD 2023 Chronic Kidney Disease Collaborators. Global, regional, and national burden of chronic kidney disease in adults, 1990–2023, and its attributable risk factors. Lancet. 2025;406:2461–2482. doi:10.1016/S0140-6736(25)01853-7

15. 15. Gbadegesin RA, Ulasi I, Ajayi S, et al. APOL1 bi- and monoallelic variants and chronic kidney disease in West Africans. N Engl J Med. 2025;392:228–238. doi:10.1056/NEJMoa2404211

16. 16. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–443. doi:10.1038/s41586-020-2308-7

17. 17. Byrska-Bishop M, Evani US, Zhao X, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell. 2022;185:3426–3440.e19. doi:10.1016/j.cell.2022.08.004

18. 18. Koenig Z, Yohannes MT, Nkambule LL, et al. A harmonized public resource of deeply sequenced diverse human genomes. Genome Res. 2024;34:796–809. doi:10.1101/gr.278378.123

19. 19. Tory K, Menyhárd DK, Woerner S, et al. Mutation-dependent recessive inheritance of NPHS2-associated steroid-resistant nephrotic syndrome. Nat Genet. 2014;46:299–304. doi:10.1038/ng.2898

20. 20. Lanktree MB, Haghighi A, Guiard E, et al. Prevalence estimates of polycystic kidney and liver disease by population sequencing. J Am Soc Nephrol. 2018;29:2593–2600. doi:10.1681/ASN.2018050493

21. 21. Gibson J, Fieldhouse R, Chan MMY, et al. Prevalence estimates of predicted pathogenic COL4A3–COL4A5 variants in a population sequencing database and their implications for Alport syndrome. J Am Soc Nephrol. 2021;32:2273–2290. doi:10.1681/ASN.2020071065

22. 22. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med. 2007;147:573–577.

23. 23. Yang S, Lincoln SE, Kobayashi Y, Nykamp K, Nussbaum RL, Topper S. Sources of discordance among germ-line variant classifications in ClinVar. Genet Med. 2017;19:1118–1126. doi:10.1038/gim.2017.60

24. National Center for Biotechnology Information. ClinVar's data model; representation of classifications in ClinVar. Available at: https://www.ncbi.nlm.nih.gov/clinvar/docs/data_model/ and https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/ (accessed 16 November 2026).

25. ClinGen Sequence Variant Interpretation / VCEP Review Committee. Guidance regarding the use of gnomAD v4, March 2024. Available at: https://www.clinicalgenome.org/site/assets/files/9445/clingen_guidance_to_vceps_regarding_the_use_of_gnomad_v4_march_2024.pdf (accessed 16 November 2026).

Figure legends

Figure 1. Study workflow and data sources. Workflow linking variant-level ClinVar classifications in 32 kidney-disease-associated genes with gnomAD v4.1 and SCV records. The continental West African follow-up is a selected-variant comparison, not independent whole-panel prevalence replication. AC, allele count; AN, allele number; VUS, variant of uncertain significance.

Figure 2. Ancestry-stratified frequency-threshold flags. Frequency-threshold exceedance among observed P/LP (A) and conflicting (B) ClinVar variants, stratified by reference genetic ancestry. Variants overlap across groups; unpaired bar percentages do not justify between-group independent tests. NFE downsampling to AFR allele numbers is a separate sensitivity analysis.

Figure 3. Modelled allele loads in dominant and recessive kidney-disease genes. Modelled expected frequency of P/LP-labelled alleles in genes assigned dominant (A) and recessive (B) disease models, before and after filtering. These are reference allele-load proxies rather than clinically observed carriers, diagnosed monogenic disease or risk estimates.

Figure 4. Submission-level categories of ClinVar classification disagreement. SCV submission-level patterns underlying variant-level conflicting ClinVar aggregates, and the proportion exceeding the modeled allele-frequency ceiling in at least one reference group. Distinct clinical conditions may be asserted for the same variant; flags are reassessment candidates.

Figure 5. Frequency-screen sensitivity to disease-model assumptions. Sensitivity of frequency flags to assumptions on maximum allelic contribution and prevalence. The colored curves are model-derived screens, not observed misclassification rates.

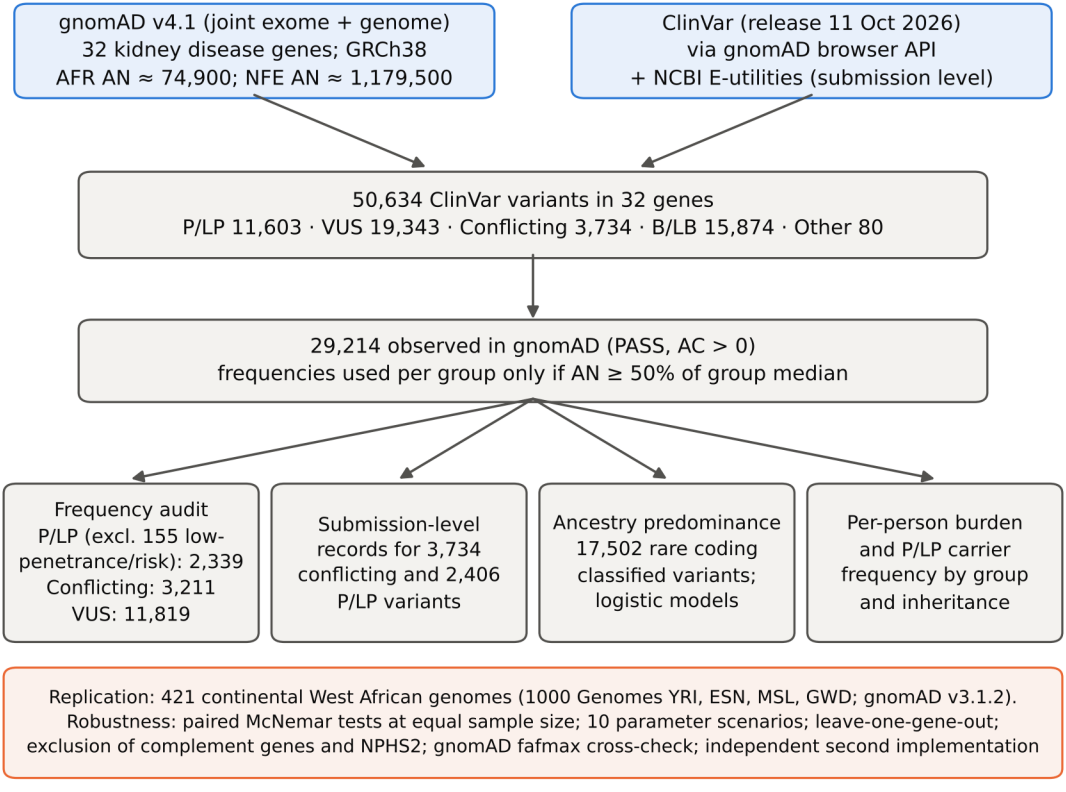
Figure 6. Expected unresolved coding-variant allele loads by ancestry. Modelled expected number of rare, coding, ClinVar-classified unresolved variant alleles per reference individual across 32 genes. These are frequency-based sums, not observed clinical test results. AFR, African/African American; NFE, non-Finnish European.

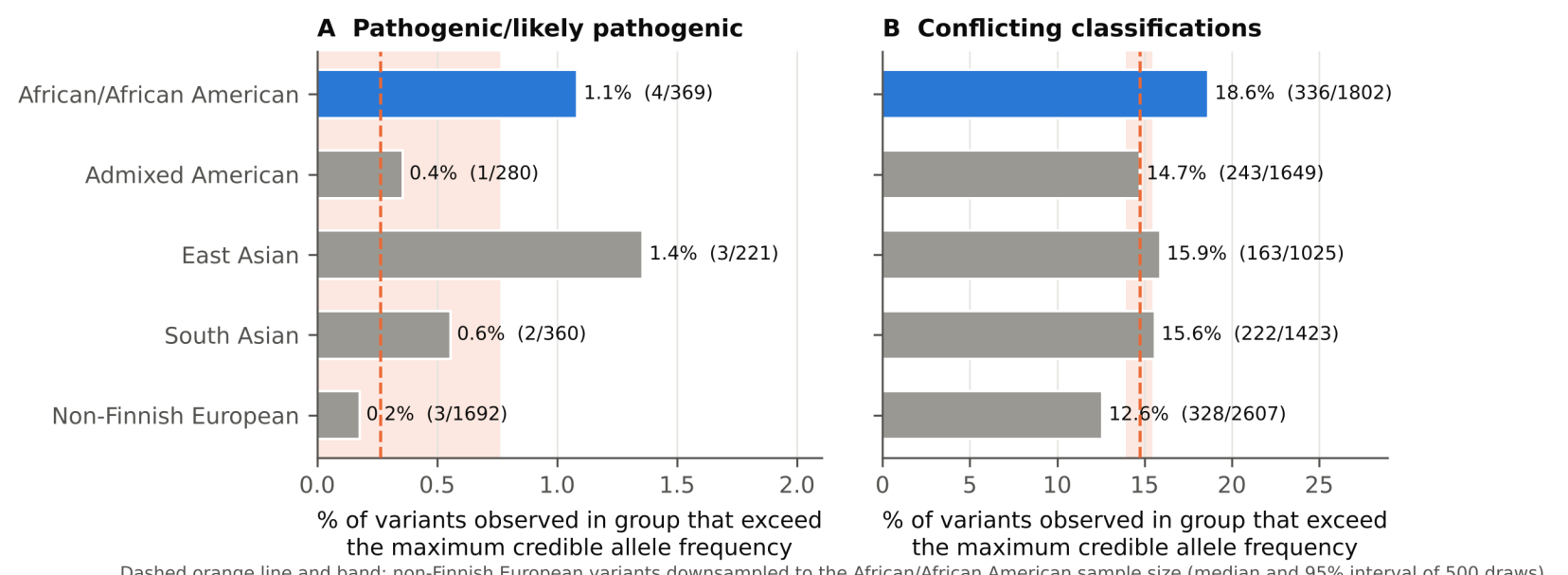
Figure 7. Concordance of selected frequency flags in continental West African genomes. Targeted concordance of model-derived frequency flags in selected continental West African reference genomes (1000 Genomes YRI and ESN, Nigeria; MSL, Sierra Leone; GWD, The Gambia; gnomAD v3.1.2 HGDP+1KG subset; n=421). Variant selection depended on prior presence in gnomAD v4.1 AFR. This is a targeted, nonrepresentative comparison rather than population-wide replication.

Additional files

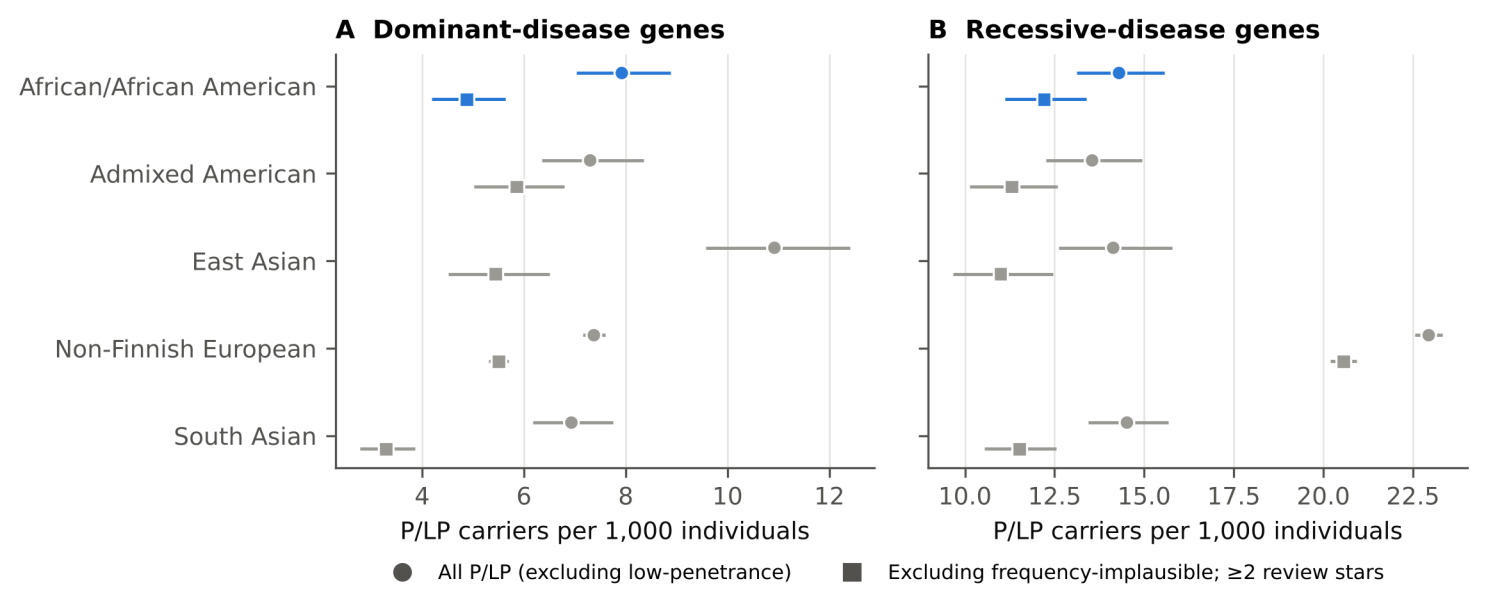
Additional file 1 (XLSX). Supplementary Tables S1–S11. Gene-specific disease models, frequency and classification summaries, sensitivity analyses, submission-level reviews and the targeted continental West African data.

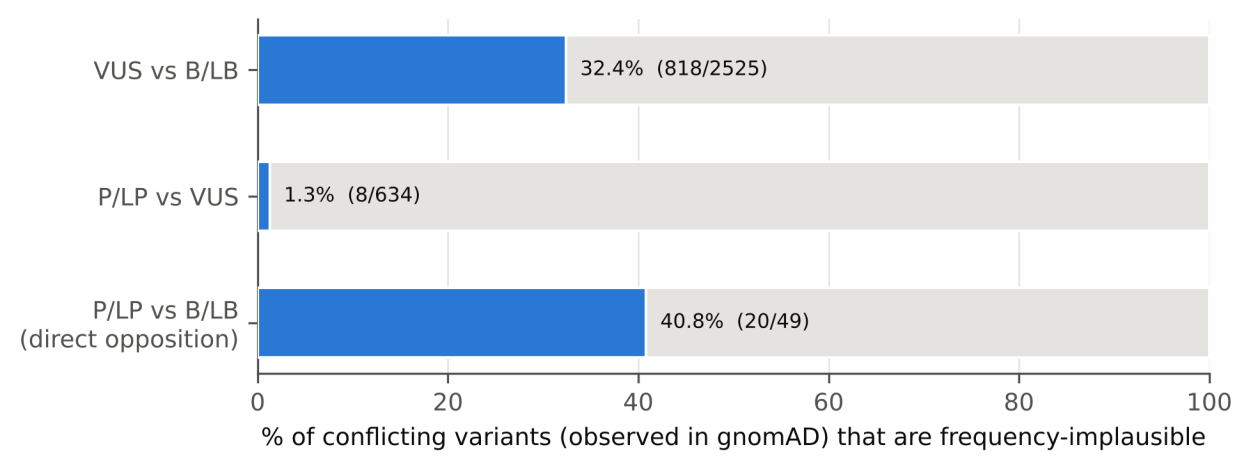
Additional file 2 (ZIP). Reproducibility package. Derived ClinVar/gnomAD data tables, analysis scripts, 500-draw audit files and supplementary source code. The seven main TIFF figures are uploaded separately.

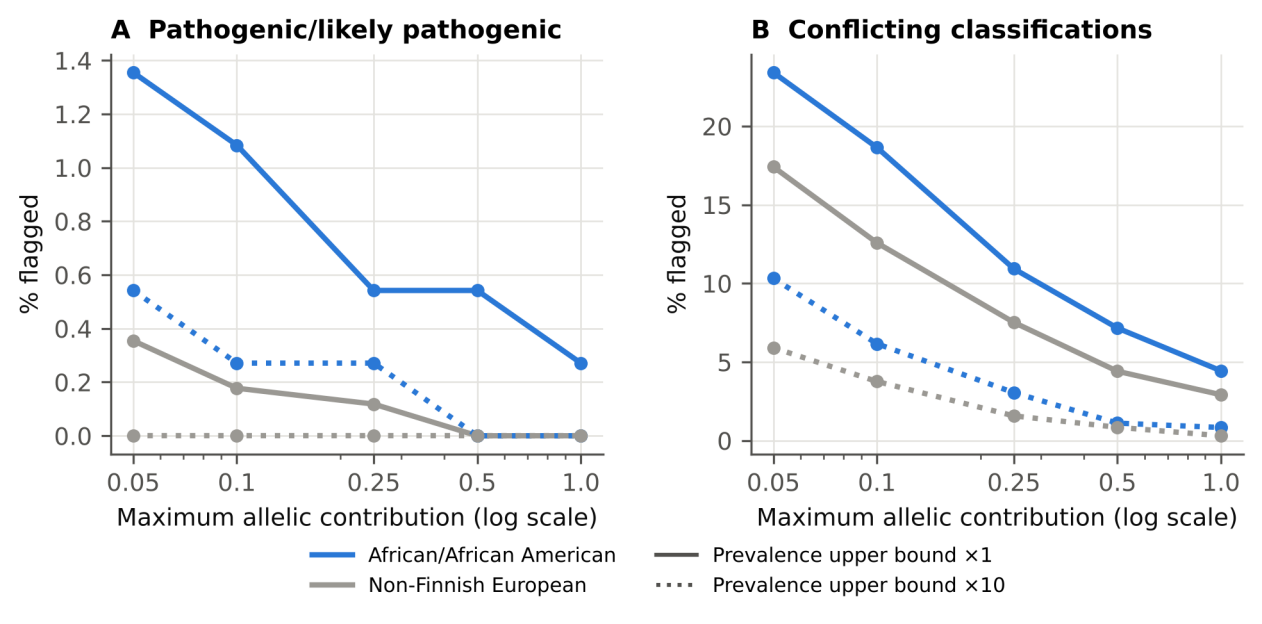

gnomAD v4.1 (joint exome + genome)
32 kidney disease genes; GRCh38
AFR AN ≈ 74,900; NFE AN ≈ 1,179,500
ClinVar (release 11 Oct 2026)
via gnomAD browser API
+ NCBI E-utilities (submission level)
50,634 ClinVar variants in 32 genes
P/LP 11,603 · VUS 19,343 · Conflicting 3,734 · B/LB 15,874 · Other 80
29,214 observed in gnomAD (PASS, AC > 0)
frequencies used per group only if AN ≥ 50% of group median
Frequency audit
P/LP (excl. 155 low-penetrance/risk): 2,339
Conflicting: 3,211
VUS: 11,819
Submission-level records for 3,734 conflicting and 2,406 P/LP variants
Ancestry predominance
17,502 rare coding classified variants; logistic models
Per-person burden and P/LP carrier frequency by group and inheritance
Replication: 421 continental West African genomes (1000 Genomes YRI, ESN, MSL, GWD; gnomAD v3.1.2).
Robustness: paired McNemar tests at equal sample size; 10 parameter scenarios; leave-one-gene-out;
exclusion of complement genes and NPHS2; gnomAD fafmax cross-check; independent second implementation

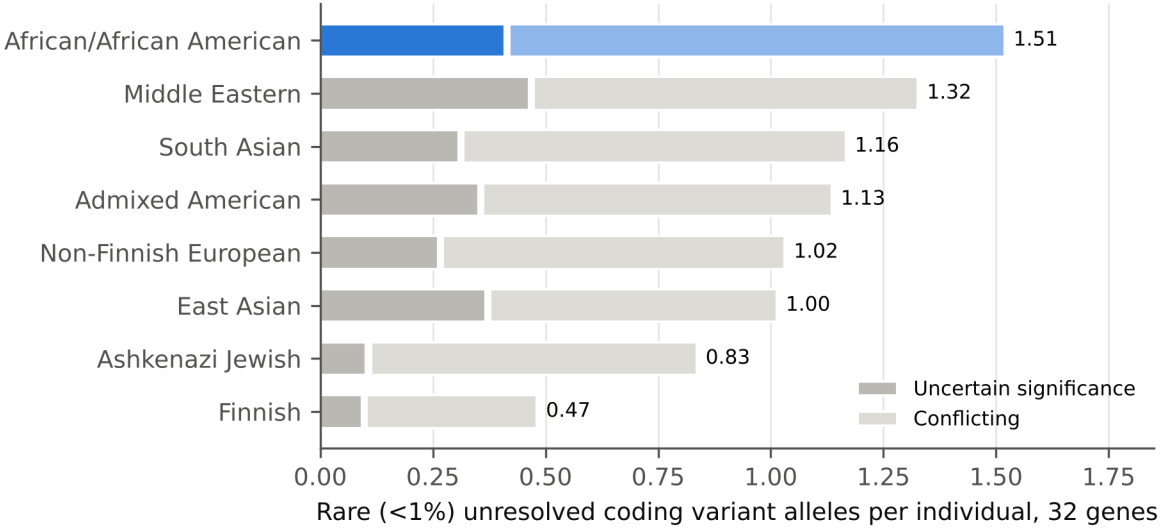


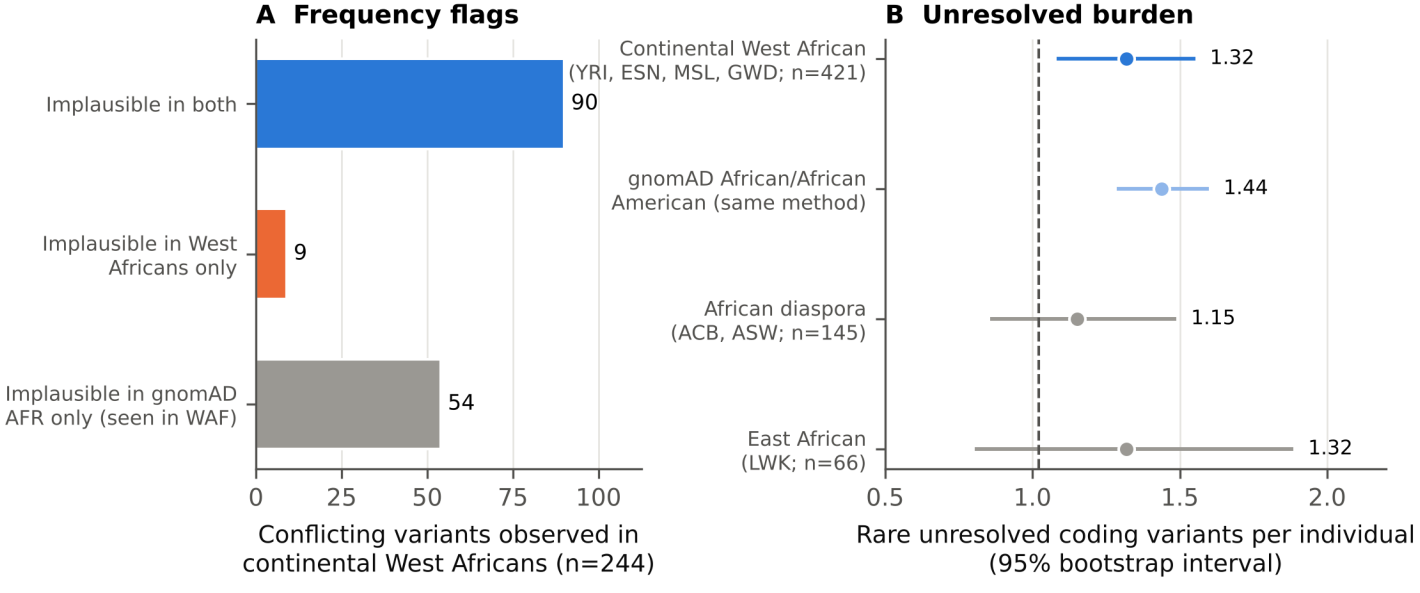
Dashed orange line and band: non-Finnish European variants downsampled to the African/African American sample size (median and 95% interval of 500 draws).


A Dominant-disease genes
B Recessive-disease genes
African/African American
Admixed American
East Asian
Non-Finnish European
South Asian
4
6
8
10
12
10.0
12.5
15.0
17.5
20.0
22.5
P/LP carriers per 1,000 individuals
P/LP carriers per 1,000 individuals
All P/LP (excluding low-penetrance)
Excluding frequency-implausible; ≥2 review stars


VUS vs B/LB
32.4% (818/2525)
P/LP vs VUS
1.3% (8/634)
P/LP vs B/LB
(direct opposition)
40.8% (20/49)
0
20
40
60
80
100
% of conflicting variants (observed in gnomAD) that are frequency-implausible


A Pathogenic/likely pathogenic
% flagged
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.05
0.1
0.25
0.5
1.0
Maximum allelic contribution (log scale)
B Conflicting classifications
% flagged
20
15
10
5
0
0.05
0.1
0.25
0.5
1.0
Maximum allelic contribution (log scale)
African/African American
Non-Finnish European
Prevalence upper bound ×1
Prevalence upper bound ×10


African/African American
Middle Eastern
South Asian
Admixed American
Non-Finnish European
East Asian
Ashkenazi Jewish
Finnish
1.51
1.32
1.16
1.13
1.02
1.00
0.83
0.47
Uncertain significance
Conflicting
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Rare (<1%) unresolved coding variant alleles per individual, 32 genes


A Frequency flags
Implausible in both
90
Implausible in West Africans only
9
Implausible in gnomAD AFR only (seen in WAF)
54
0
25
50
75
100
Conflicting variants observed in continental West Africans (n=244)
B Unresolved burden
Continental West African (YRI, ESN, MSL, GWD; n=421)
1.32
gnomAD African/African American (same method)
1.44
African diaspora (ACB, ASW; n=145)
1.15
East African (LWK; n=66)
1.32
0.5
1.0
1.5
2.0
Rare unresolved coding variants per individual (95% bootstrap interval)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesS1S11.xlsx