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.

# Carrier Burden and Birth Prevalence of Recessive Inborn Errors of Metabolism in South Asians: A gnomAD and GenomeIndia Study

**Atanu Dutta**

atanu.biochem@aiimskalyani.edu.in

All India Institute of Medical Sciences, Kalyani https://orcid.org/0000-0002-8413-3454

**Article**

**Keywords:** Autosomal-recessive inborn errors of metabolism, gnomAD v4, GenomeIndia, South Asian population genetics, carrier frequency

**Posted Date:** October 20th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11386338/v1

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

**Additional Declarations:** There is no duality of interest

# Abstract

Inborn errors of metabolism are individually rare but collectively common, often inherited recessively and causing severe early-life illness. Using gnomAD v4 data from 86,254 South Asian individuals across 696 genes, this study estimates their birth prevalence and carrier frequency, with findings validated against GenomeIndia summary statistics. Pathogenic variants were analysed in three groups—ClinVar P/LP, LOFTEE-HC, and predicted-damaging missense variants (AlphaMissense/REVEL)—using strict quality filters and ClinVar-validated predictors. Under random mating, the estimated birth prevalence is 255 per 100,000 people (1 in 392): 40.6 for ClinVar pathogenic or likely pathogenic variants, 41.3 for high-confidence loss-of-function variants, and 62.7 for predicted-damaging missense variants; the remainder involves compound heterozygosity. In the GenomeIndia cohort (n = 9,768; 83 endogamous groups), prevalence was 139.6 per 100,000 (1 in 716), rising to 151.5 (1 in 660) when India-specific ClinVar P/LP alleles absent from gnomAD were included. Consistent per-gene allele loads between cohorts (Spearman ρ = 0.83) yielded a pooled estimate of 189.1 per 100,000 (1 in 529). Because consanguinity varies across South Asia, it was modeled using F = 0.01–0.0625 scenarios rather than a single estimate. Overall, 84.2% of South Asians carry at least one pathogenic allele — the highest collective carrier frequency among eight global ancestries, though closely clustered with other outbred continental populations (82.1%–84.0%), while founder populations exhibited lower rates (63.6%–64.8%). Cross-ancestry comparison places South Asian prevalence mid-distribution and reveals a Eurocentric ascertainment bias in ClinVar. All analyses are reproducible from the accompanying code and data.


# 1. Introduction

Inborn errors of metabolism (IEMs) are rare single-gene disorders that disrupt cellular biochemical pathways. Although individually uncommon, they collectively cause significant illness, intellectual disability, and childhood mortality. Most are autosomal recessive, and early detection through newborn or carrier screening can enable treatments that prevent irreversible complications [1–3, 16–19].

Estimating IEM prevalence in South Asia is difficult because studies vary in design, screening coverage, and case identification. Reported estimates range from 1 in 1,000 to 1 in 2,500 live births, while earlier models estimated approximately 1 in 3,100 in India. These figures may be understated because they often exclude predicted pathogenic variants and do not account for nonrandom mating [1, 2].

High rates of endogamy and consanguinity in South Asian populations increase autozygosity and the likelihood that recessive variants occur in a homozygous state. To improve prevalence estimates, this study analyzed 696 autosomal-recessive IEM genes using gnomAD v4 sequencing data. It incorporated inbreeding corrections, ClinGen gene-validity tiers, structural-variant detection gaps, comparisons across eight global populations, and calibration against newborn-screening registries [1, 2, 4–6, 15–19].

# 2. Methods

## 2.1. Gene Panel Curation

The gene panel was derived from the Genomics England PanelApp “Likely inborn error of metabolism” panel (R98) [15]. Diagnostic-grade (“green”) genes with confirmed autosomal-recessive inheritance yielded 696 AR-IEM genes. Gene symbols and coordinates were mapped to GRCh38, with obsolete HGNC symbols reconciled to gnomAD v4 nomenclature, including MUT to MMUT, GBA to GBA1, and G6PC to G6PC1.

## 2.2. Allele-Frequency Extraction and Quality Control

South Asian (SAS) variant data were obtained from the gnomAD v4 GraphQL API using $gnomad_r4$. PASS-only exome and genome allele counts (AC) and allele numbers (AN) were aggregated, and allele frequency (AF) was calculated as AC/AN. Observed homozygote counts were also recorded.

Two filters were applied:

- Homozygote-only artifact filter: Variants observed in no more than two homozygotes and no heterozygotes were excluded.
- Allele-frequency ceiling: Variants with AF > 0.5% were excluded to remove common hypomorphic or mild variants, such as MTHFR c.677C > T.

## 2.3. Pathogenic Variant Classification

Variants were evaluated in two sets:

- Broad Panel Set: Included ClinVar Pathogenic/Likely Pathogenic (P/LP) variants without conflicting interpretations, LOFTEE high-confidence loss-of-function variants, and missense variants classified as AlphaMissense “likely pathogenic” with REVEL ≥ 0.75. ClinVar Benign/Likely Benign classifications took precedence.
- ClinVar-Only Set: Included P/LP variants supported by an expert panel or clinical submission.

Variant classes were retained separately. Against 20,329 cleanly classified ClinVar missense variants, the combined AlphaMissense–REVEL criterion achieved 98.9% specificity and positive predictive value. Among cleanly classified loss-of-function variants, 99.3% were ClinVar P/LP, supporting the inclusion of LOFTEE high-confidence variants.

## 2.4. Population Genetics and Consanguinity Modeling

For each gene, the aggregate pathogenic allele frequency (q) was calculated as: q = ∑ a_i where a_i is the frequency of the i-th pathogenic allele.

Under Hardy–Weinberg Equilibrium (random mating), genetic prevalence (GP) was calculated using the full allele-pair matrix model accounting for both true homozygotes (∑ a_i²) and compound heterozygotes (∑ 2a_i a_j), which simplifies to: GP = q².

To model inbreeding, Wright's coefficient of inbreeding (F) was incorporated: $GP_F = q^2 + F\,q(1 - q)$

Because South Asia comprises many endogamous sub-populations with different inbreeding levels, a single inbreeding coefficient is not assumed. Instead, F is evaluated as an explicit set of what-if scenarios — outbred (F = 0), low (F = 0.01), moderate (F = 0.02), high (F = 0.03), and a first-cousin reference (F = 0.0625) — and the collective birth prevalence is reported for each scenario separately.

Per-gene carrier frequency (CF) was calculated as: $CF = 1 - (1 - q)^2 - q^2$

Collective carrier frequency and collective genetic birth prevalence across all 696 independent disease loci were calculated using the independence aggregation formula: Collective Burden = $1 - \prod (1 - x_g)$ [from g = 1 to 696] where $x_g$ represents either CF or $GP_F$ for gene g. Non-parametric binomial bootstrap resampling (B = 1,000 iterations, seed 42) was performed on raw allele counts to derive 95% confidence intervals (CIs).

# 2.5. Cross-Ancestry Comparison

The pipeline was repeated for AFR, AMR, ASJ, EAS, FIN, MID, and NFE gnomAD v4 ancestry groups. The 0.5% AF threshold and homozygote-artifact filter were reassessed independently for each group.

# 2.6. Gene Validity, Discovery Gap, and Newborn-Screening Calibration

Genes were stratified by ClinGen validity:

- Tier A: Definitive or Strong (n = 452)
- Tier B: Moderate (n = 75)
- Tier C: Limited, Disputed, Refuted, or No Known Disease Relationship (n = 27)
- Tier U: Uncurated (n = 142)

Tier U genes were cross-referenced with Genomics England PanelApp panels. Curated ClinVar pathogenic variants were classified as SNVs/indels or CNVs to estimate structural-variant discovery gaps. Predicted genetic prevalence for 15 IEM conditions was compared with newborn-screening incidence estimates to derive an empirical clinical-to-genetic scaling factor 16–19]. Analyses used Biomni [20] and Google Antigravity (Google, 2025, [https://antigravity.google/).

# 2.7. Independent Replication in the GenomeIndia Cohort

The estimators were independently applied to GenomeIndia summary allele frequencies [21, 22] from 9768GI_SummaryStats (Apr 2025)`. The dataset included 129.9 million quality-filtered variants from 9,768 unrelated individuals across 83 endogamous Indian populations. Data were aligned to GRCh38 and obtained from the Indian Biological Data Centre.

Two variant sets were queried using exact chromosome-position-reference-alternate matching with minimal indel normalization:

- The 20,859 classified pathogenic panel alleles.
- ClinVar P/LP variants absent from the gnomAD-derived panel (n = 50,429), restricted to GenomeIndia AF ≤ 0.5%.

Absent variants were assigned AF = 0. Per-gene allele loads, collective carrier frequency, prevalence, and consanguinity estimates were then recalculated.

Three sensitivity analyses were performed:

- Non-observation sensitivity: Absent variants were assigned AF = 1/19,536, the single-allele detection limit.
- Pooled estimate:, a two-cohort pooled estimate was derived by pooling allele counts per variant — pooled AF = (AC_SAS + AC_GI)/(AN_SAS + AN_GI), with GenomeIndia allele counts recovered by rounding AF × 19,536 and GI-only ClinVar variants pooled with AC_SAS = 0 at the median SAS allele number — and per-gene loads and collective prevalence were recomputed from pooled allele frequencies using the same estimators.
- Curation of candidate enriched alleles: GenomeIndia AF ≥ 0.1% with at least two-fold enrichment over gnomAD SAS, or GenomeIndia AF ≥ 0.5%) were curated: consequences were annotated on MANE Select/canonical transcripts with Ensembl VEP, population frequencies and homozygote counts were examined across gnomAD v4 ancestry groups, ClinVar records were reviewed, and the literature was searched for founder reports [23, 24].

## 3. Results

# 3.1. Primary Estimates of Carrier Burden and Birth Prevalence

Quality filtering identified 20,859 unique pathogenic alleles across 695 active genes in the South Asian gnomAD sample: 9,225 predicted-damaging missense, 6,777 high-confidence LoF, and 4,857 ClinVar P/LP alleles. NDUFAF8, the only inactive gene in the 696-gene panel, had no qualifying alleles and was excluded from burden calculations.

Overall, 84.2% of individuals carried at least one pathogenic allele (95% CI: 83.8%–84.6%). Assuming random mating, the estimated genetic birth prevalence was 255 per 100,000 live births (1 in 392; 95% CI: 251–269): 40.6 for ClinVar P/LP, 41.3 for high-confidence LoF, and 62.7 for predicted-damaging missense variants (Fig. 1).

The combined estimate is 255.1 per 100,000, reported as 255, exceeding the within-category total of 144.6 because compound heterozygosity contributes an additional 110.5 cases per 100,000—approximately 43% of the total. Under random mating, 82% of predicted affected infants are compound heterozygotes and 18% are homozygotes, highlighting the allelic diversity of IEM loci and the importance of biallelic matrix modelling.

Because inbreeding varies across South Asia, consanguinity is represented through scenarios. Prevalence rises from 255 per 100,000, or 1 in 392, under random mating to 1,168 at F = 0.01, 2,072 at F = 0.02, 2,968 at F = 0.03, and 5,824 per 100,000, or 1 in 17, among offspring of first-cousin unions (F = 0.0625). The appropriate scenario depends on the population; random mating provides the conservative baseline.

Table 1
Collective AR-IEM genetic birth prevalence in South Asians across increasing inbreeding coefficients (F)

| Mating / Inbreeding Model | Collective Prevalence (per 100,000) | Rate | 95% CI (per 100k) |
|---|---|---|---|
| Hardy–Weinberg (Random Mating) | 255 | 1 in 392 | 251–269 |
| F = 0.01 | 1,168 | 1 in 86 | — |
| F = 0.02 (Population Average) | 2,072 | 1 in 48 | 2043–2112 |
| F = 0.03 | 2,968 | 1 in 34 | — |
| F = 0.0625 (First-Cousin Unions) | 5,824 | 1 in 17 | — |

[Note: 95% bootstrap confidence intervals (B = 1,000 iterations) were empirically derived for the random-mating baseline (Hardy–Weinberg) and the representative South Asian population-average consanguinity scenario (F = 0.02); intermediate and extreme consanguinity scenarios (F = 0.01, F = 0.03, and F = 0.0625) are evaluated as deterministic scenario projections across the reported inbreeding range.]

# 3.2. Top Contributing Genes

Table 2
Top 12 AR-IEM genes in South Asians ranked by consanguinity-adjusted (F = 0.02) genetic prevalence

| Gene | Pathogenic Variants (n) | Aggregate q (per 1,000) | Carrier Frequency (%) | HW Prevalence (per 100k) | F = 0.02 Prevalence (per 100k) |
|---|---|---|---|---|---|
| *MAN2B2* | 69 | 10.09 | 2.00% | 10.2 | 30.2 |
| *DPYD* | 98 | 9.72 | 1.93% | 9.4 | 28.7 |
| *PYGL* | 116 | 8.59 | 1.70% | 7.4 | 24.4 |
| *ATP7B* | 150 | 7.91 | 1.57% | 6.3 | 21.9 |
| *PAH* | 82 | 7.42 | 1.47% | 5.5 | 20.2 |
| *MMUT* | 107 | 7.14 | 1.42% | 5.1 | 19.3 |
| *PYGM* | 126 | 7.03 | 1.40% | 4.9 | 18.9 |
| *DHCR24* | 33 | 6.97 | 1.39% | 4.9 | 18.7 |
| *SPG7* | 103 | 6.58 | 1.31% | 4.3 | 17.4 |
| *SACS* | 154 | 6.42 | 1.28% | 4.1 | 16.9 |
| *GLDC* | 109 | 6.34 | 1.26% | 4.0 | 16.6 |
| *MLYCD* | 55 | 6.31 | 1.25% | 4.0 | 16.5 |

No single gene accounts for most of the burden. Major contributors include PAH, ATP7B, MMUT, GLDC, and glycogen-storage-disease genes such as PYGL, PYGM, and GYS2; MAN2B2 and DPYD are also predicted to contribute substantially.

# 3.3. Sensitivity Analyses and Quality-Control Validation

Sensitivity analyses tested alternative variant-filtering thresholds (Supplementary Table S1).

Variant-definition thresholds: Restricting the panel to 4,857 ClinVar-curated P/LP variants across 590 genes reduced prevalence to 41 per 100,000 (1 in 2,461) under random mating. At F = 0.02, prevalence was 551 per 100,000 (1 in 181), with a 40.1% carrier frequency (Supplementary Figure S1).

Impact of the AF ceiling: Removing the 0.5% allele-frequency limit inflated random-mating prevalence approximately 350-fold, to 88,948 per 100,000, largely because of common variants such as MTHFR c.677C > T.

Missense criteria: Estimates were 255 per 100,000 using AlphaMissense plus REVEL ≥ 0.75, 473 per 100,000 using AlphaMissense alone, and 499 per 100,000 using REVEL ≥ 0.75 alone.

Predictor validation against ClinVar: The combined missense criterion achieved 98.9% specificity and positive predictive value, with 59.1% sensitivity across 20,329 cleanly classified ClinVar missense

variants. These results indicate few false positives and suggest that prevalence is conservatively estimated. High-confidence loss-of-function variants were also strongly supported: 99.3% were ClinVar P/LP variants (Fig. 3).

## 3.4. Cross-Ancestry Comparison and Ascertainment Bias

Across eight gnomAD ancestries, South Asians had the highest collective carrier frequency (84.2%), followed by AFR (84.0%), EAS (83.5%), AMR (83.0%), NFE (82.6%), and MID (82.1%). Founder populations had lower frequencies: 64.8% in Finnish and 63.6% in Ashkenazi Jewish populations. South Asians had an estimated Hardy–Weinberg birth prevalence of 255 per 100,000, comparable to European, African, and Latino populations.

ClinVar-only analysis instead placed NFE populations highest for disease prevalence (85 per 100,000; 52.5% carrier frequency), while the South Asian estimate fell to 41 per 100,000. This reversal likely reflects Eurocentric ascertainment bias, given the larger number of curated NFE alleles than SAS alleles. Ancestry-neutral pipelines may reduce this imbalance, while cross-ancestry heatmaps identified regional founder effects involving MAN2B2 and DPYD in South Asians and genes with substantial variation in pathogenic-allele burden.

## 3.5. Gene Validity Tiering, Discovery Gap, and NBS Calibration

ClinGen Gene Curation: The 452 ClinGen Tier A genes yielded a random-mating prevalence of 167 per 100,000; adding 142 validated Tier U genes (Tier A + B+U+, n = 669) increased it to 240 per 100,000.

Structural Variant Discovery Gap: Of 54,554 curated ClinVar alleles, 37.2% (n = 20,310) were structural variants or CNVs undetectable by short-read exome sequencing, while gnomAD SAS captured a median of only 14% of known SNVs/indels per gene; accounting for these gaps raised the estimated upper-bound prevalence to 406 per 100,000 (1 in 247) (Supplementary Figure S3).

Newborn Screening Calibration: Across 15 standard NBS conditions, the median clinical-to-genetic prevalence ratio was 0.65× (IQR: 0.25–1.22), supporting agreement between Hardy–Weinberg predictions and observed incidence; applying this factor yielded a calibrated incidence of 165 per 100,000 live births (1 in 608; 255 × 0.65) (Supplementary Table S2) [16–19].

## 3.6. Plausible Range Synthesis

For transparency, a bounded plausible range is presented under random mating. The primary estimate is the uncorrected random-mating value; the structural-variant and newborn-screening adjustments are heuristic and are shown only as clearly labelled exploratory bounds, not as validated corrections:

- Conservative Lower Bound: 33 per 100,000 (1 in 3,004) [ClinVar-only, ClinGen Tier A].
- ClinGen Tier A Reference: 167 per 100,000 (1 in 598) [Broad panel, Tier A].

- Exploratory bound (not validated) — NBS clinical calibration: 165 per 100,000 (1 in 608) [broad panel × 0.65 clinical-to-genetic factor; calibrated on published NBS cohorts].
- Central Primary Estimate: 255 per 100,000 (1 in 392) [Broad full panel, HW].
- Structural-Corrected Upper Bound: 406 per 100,000 (1 in 247) [Broad panel X 1.59].

# 3.7. Independent Replication in the GenomeIndia Cohort

Of 20,859 panel alleles, 2,657 (12.7%) were observed in GenomeIndia, which had 540 of 50,429 GI-only ClinVar P/LP variants. Estimated birth prevalence under random mating was 139.6 per 100,000 (1 in 716) for shared variants and 151.5 (1 in 660) including GI-only variants, compared with 255.1 (1 in 392) in gnomAD SAS.

High-confidence LoF variants were slightly more prevalent in GenomeIndia than in gnomAD SAS (44.6 vs. 41.3 per 100,000), whereas ClinVar P/LP and damaging missense variants were less common (23.8 vs. 40.6 and 37.8 vs. 62.7). Carrier frequency was also lower (66.1% vs. 84.2%), although per-gene pathogenic allele loads were strongly correlated (Spearman ρ = 0.83 across 608 genes).

Three apparently enriched LoF alleles—in XPNPEP3, RBCK1, and GLDC—were classified as population polymorphisms because their LoF annotations occurred on minor non-canonical transcripts. After excluding them, high-confidence LoF prevalence was 14.6 per 100,000; panel-only and combined prevalences were 108.5 and 119.9 per 100,000, respectively. GI-only variants increased prevalence by 11.9 per 100,000, while consanguinity scenarios raised GenomeIndia estimates to 688–3,459 per 100,000 (F = 0.01–0.0625).

Table 3
Collective AR-IEM birth prevalence in GenomeIndia versus gnomAD South Asian, by variant set, under random mating (per 100,000 births)

| Variant set | gnomAD SAS (per 100,000) | GenomeIndia (per 100,000) | GenomeIndia / SAS |
|---|---|---|---|
| ClinVar P/LP | 40.6 | 23.8 | 0.59 |
| High-confidence LoF (raw) | 41.3 | 44.6 | 1.08 |
| High-confidence LoF (artifact-filtered) | 41.3 | 14.6 | 0.35 |
| Predicted damaging missense | 62.7 | 37.8 | 0.60 |
| Combined (same variant set, raw) | 255.1 | 139.6 | 0.55 |
| Combined (same variant set, artifact-filtered) | 255.1 | 108.5 | 0.43 |
| Combined (+ GI-only ClinVar P/LP, raw) | 255.1 | 151.5 | 0.59 |
| Combined (+ GI-only ClinVar P/LP, artifact-filtered) | 255.1 | 119.9 | 0.47 |

[Note: Artifact-filtered sensitivity estimates recalculate GenomeIndia birth prevalence after excluding three high-frequency loss-of-function variants identified during transcript-level review (Section 3.8) as non-canonical intronic polymorphisms / annotation artefacts: XPNPEP3 22:40860676:A:G (GI AF 1.25%), RBCK1 20:417296:C:T (GI AF 1.03%), and GLDC 9:6622821:C:A (GI AF 0.59%). Removing these false-positive alleles resolves the apparent LoF excess in GenomeIndia (dropping LoF prevalence from 44.6 to 14.6 per 100,000; GI / SAS = 0.35), aligning with the ~ 0.6× capture ratio observed across the other variant classes due to sample size differences.]

# 3.8. Sensitivity Analyses, Pooled Meta-Estimate, and Founder-Variant Review

Assigning undetected panel variants the single-allele detection limit (AF = 1/19,536) increased the panel-only estimate from 139.6 to a maximum of 549.3 per 100,000. The corresponding ranges were 23.8–65.0 for ClinVar P/LP variants, 44.6–83.7 for high-confidence LoF variants, and 37.8–143.2 for predicted damaging missense variants. Thus, the GenomeIndia estimate likely falls between 139.6 and 549.3 per 100,000; its difference from the gnomAD SAS estimate of 255.1 is compatible with ascertainment and sampling differences. The GI-only ClinVar upper bound of 2,956 per 100,000 is a mathematical worst-case scenario, not an empirical estimate.

Pooling the two cohorts yielded a collective birth prevalence of 189.1 per 100,000 (1 in 529) and a carrier frequency of 78.1%. Because gnomAD SAS contributes approximately 4.4 times more alleles, this estimate is weighted toward its value. The pooled LoF prevalence was 28.3 per 100,000, reflecting

dilution of alleles enriched in GenomeIndia and the absence of gnomAD SAS-specific alleles from GenomeIndia. Under consanguinity assumptions, the estimate rose from 943 per 100,000 at F = 0.01 to 4,808 per 100,000 at F = 0.0625. These pooled values are weighted averages, not definitive pan-Indian estimates.

Twenty-one potentially enriched alleles were reviewed (Supplementary Table S3). Transcript review showed that all three tier-1 LoF-annotated alleles are intronic in MANE Select or canonical transcripts, with LoF effects limited to minor non-canonical transcripts. XPNPEP3 and RBCK1 are also common East Asian polymorphisms, making severe, highly penetrant disease unlikely; they were therefore considered annotation artefacts. NAXD showed a similar pattern, while the LDHA splice-region deletion is common worldwide at approximately 0.1%.

Three candidates remained. The canonical DYM stop-gained variant is enriched in South Asians, but no homozygotes were observed, so independent replication is needed. The MMACHC c.848del termination-loss variant occurs at 0.31% in GenomeIndia versus 0.009% elsewhere but remains a variant of uncertain significance. The BTD 3'-UTR variant occurs at 0.154% in GenomeIndia, is nearly absent globally, and has conflicting ClinVar classifications supported by evidence from a single family [24].

The known North Indian cblC founder variant MMACHC R132 (c.394C > T) [23] was detected in both cohorts, at 0.08% in gnomAD SAS and 0.036% in GenomeIndia, supporting the enrichment-based approach. Population-frequency enrichment can help identify Indian founder alleles, but transcript-aware review is essential before making clinical interpretations.

## 4. Discussion

# 4.1. Principal Findings and Biological Significance

This study estimates that 84.2% of South Asians carry at least one pathogenic allele linked to autosomal-recessive metabolic disorders. Under random mating, the expected prevalence is 255 affected births per 100,000, similar to estimates for outbred populations worldwide.

The higher clinical prevalence in some South Asian populations is therefore likely driven mainly by consanguinity and endogamy, which increase the chance that two carriers have an affected child. With reported inbreeding coefficients of F = 0.01–0.03, projected prevalence rises to 1,168–2,968 per 100,000 births—approximately 1 in 86 to 1 in 34. These estimates vary among sub-populations and should not be applied uniformly to all South Asians.

# 4.2. Within-Region Population Heterogeneity and Founder Effects

gnomAD groups diverse South Asian populations into one cohort, although the region contains thousands of endogamous communities with distinct histories of separation and founder bottlenecks

[4–6]. Studies report rare homozygous variants at rates up to 100 times higher than in outbred populations, along with population-specific disease loci [4–7].

Key findings include:

- **Canavan disease**: An *ASPA* founder mutation (c.526G > A, p.Gly176Ser) contributed to juvenile cases in the Telugu Devanga Chettiar community [7].
- **Wilson disease**: *ATP7B* variants show regional distributions and distinct founder origins across India [8–10].
- **Phenylketonuria**: Indian patients have a distinctive *PAH* variant spectrum, including more splice-site, insertion/deletion, and nonsense variants than commonly reported in European populations [11].
- **Tribal populations**: Odisha cohorts show elevated inbreeding ($F \geq 0.0156$), extended *ROH* (> 8 Mb), and potentially pathogenic variants in BTD and hemoglobinopathy-related genes [13].
- **Carrier screening**: Panels based mainly on European populations have reduced sensitivity and higher residual carrier risk in South Asia, supporting population-specific panels [12, 14].

GenomeIndia supports these findings despite having a sample approximately one-ninth the size of the comparator cohort. It showed a similar disease burden (139.6–151.5 versus 255.1 per 100,000) and comparable gene rankings (Spearman $\rho = 0.83$). Its lower estimate likely reflects reduced detection of ultra-rare alleles in smaller cohorts. Unlike gnomAD SAS, GenomeIndia includes 83 Indian endogamous groups.

The apparent excess of loss-of-function (LoF) alleles was mainly attributable to XPNPEP3, RBCK1, and GLDC. Transcript review identified these as non-South-Asian polymorphisms affecting minor, non-canonical transcripts, suggesting annotation artefacts rather than founder-effect disease alleles (Section 3.8).

Three candidates—DYM c.1488T > A, MMACHC c.848del [1:45509213 TG > T], and BTD c.159G > A—require functional and clinical validation. Future studies should assess all 83 groups, validate these variants, and compare newborn-screening yields with GenomeIndia-based predictions.

# 4.3. Ascertainment Disparities and Genomic Equity

ClinVar-only analyses substantially underestimate South Asian carrier frequency (40.1%) and make the Non-Finnish European burden appear highest. This reflects Eurocentric biases in clinical testing and database submissions, leading to under-ascertainment among South Asian patients [14]. Validated predictors such as LOFTEE, AlphaMissense, and REVEL provide a more equitable basis for genomic modelling.

# 4.4. Clinical and Public Health Policy Implications

Newborn Screening (NBS) Programme Expansion:

Expanded Carrier Screening (ECS): North American and European panels may have lower sensitivity in South Asian populations because of differing mutation spectra [12, 14]. Priority loci include DPYD, ATP7B, PAH, SACS, BTD, and PYGL (Supplementary Figure S4).

Targeted Sequencing Priorities: Panels for endogamous populations should balance diagnostic yield with sequencing feasibility. DPYD, ATP7B, and PAH require pipelines capable of detecting copy-number variants (CNVs).

## 4.5. Study Limitations

Genetic prevalence versus clinical incidence: Genetic models count predicted biallelic genotypes, including hypomorphic, late-onset, or incompletely penetrant variants.

Population heterogeneity: The gnomAD SAS dataset pools South Asian samples, although the region contains many endogamous subpopulations with distinct founder effects and inbreeding coefficients (F) [4–6].

Structural and intronic variants: Exome data miss deep-intronic and complex structural variants, which account for up to 37% of known pathogenic alleles in IEM loci.

Consanguinity scenarios and adjustments: Because inbreeding varies across subpopulations, consanguinity estimates are exploratory scenarios rather than a single South Asian estimate. The structural-variant (approximately 1.59×) and clinical-to-genetic (approximately 0.65×) adjustments are heuristic and excluded from the headline estimate.

# 5. Conclusions

Under random mating, the estimated combined birth prevalence of autosomal-recessive inborn errors of metabolism among South Asians is 255 per 100,000 births, or 1 in 392. Under the modelled consanguinity scenarios, it rises to 1,168–2,968 per 100,000 (approximately 1 in 86 to 1 in 34) for F = 0.01–0.03. At F = 0.02, the estimate is 2,072 per 100,000; among offspring of first-cousin unions (F = 0.0625), it reaches 5,824 per 100,000, or 1 in 17. An estimated 84% or more of individuals carry at least one pathogenic allele among the 696 genes studied. These findings may inform genomic policy, carrier screening, and newborn-screening programmes across South Asia.

# Declarations

### Data and Code Availability

The frozen 696-gene panel, the complete variant-level table (20,859 QC-passed variants with allele counts/frequencies, predictor flags, pathogenicity class, curation tier, and feasibility tier), all analysis scripts (in run order), and a README documenting inputs and environment are provided in the accompanying reproducibility bundle (AR_IEM_reproducibility_bundle.zip). Every figure and table is

generated by an included script from these data. External annotation sources: gnomAD v4 (gnomad_r4), ClinVar GRCh38, AlphaMissense hg38, and REVEL v1.3. The GenomeIndia replication (lookup keys, per-variant GenomeIndia allele frequencies, per-type and per-gene estimates, comparison tables, figures, and the four analysis scripts) is provided in the accompanying genome_india_comparison package.

**Data sharing statement**

The data that support the findings of this study are available in gnomAD at https://gnomad.broadinstitute.org/, reference number gnom AD v.4.1.1 & GenomeIndia. These data were derived from the following resources available in the public domain: - gnomAD, https://gnomad.broadinstitute.org/ & GenomeIndia https://ibdc.dbtindia.gov.in/genomeindia/.

**Declaration of generative AI and AI-assisted technologies in the research and writing process**

During the preparation and execution of this work, the author utilized Microsoft Copilot to improve text readability and flow. In addition, computational data retrieval, variant curation queries, and population-genetics pipeline execution were facilitated using the autonomous biomedical research AI agent Biomni [20] & Google Antigravity under the author's direct supervision and validation. Following the use of these tools, the author independently verified, evaluated, and edited all analytical outputs, code, and manuscript text, and takes full responsibility for the content and integrity of the publication.

**Acknowledgments**

The author acknowledges the infrastructural support provided by AIIMS Kalyani for this work.

**Details of the contributions of individual authors**

Atanu Kumar Dutta conceptualized the study, designed and supervised the data extraction and computational analyses (facilitated by the AI agent Biomni [20] & Google Antigravity), verified all analytical workflows, and wrote the manuscript.

**Funding**

The author declares no funding support for this work.

**Ethical Approval**

As this study involved deidentified open-source data from the GnomAD database & GenomeIndia database the Ethics Committee approval was waived off.

**Informed Consent**

As this study involved deidentified open-source data from the GnomAD database & GenomeIndia database the informed consent was not necessary.

**Consent for publication**

As this study involved deidentified open-source data from the GnomAD database & GenomeIndia database the consent for publication was not necessary.

# References


1. Hanany M, Rivolta C, Sharon D. Worldwide carrier frequency and genetic prevalence of autosomal recessive inherited retinal diseases. Proc Natl Acad Sci USA. 2020;117(5):2710–6. doi:10.1073/pnas.1913179117
2. Mondal S, Goswami K, Dutta AK. Estimation of the global burden of autosomal recessive rare inborn errors of metabolism [Internet]. Genetic and Genomic Medicine; 2025 [cited 2026 Aug 11]. Available from: http://medrxiv.org/lookup/doi/10.1101/2025.02.14.25322285 doi:10.1101/2025.02.14.25322285
3. Waters D, Adeloye D, Woolham D, Wastnedge E, Patel S, Rudan I. Global birth prevalence and mortality from inborn errors of metabolism: a systematic analysis of the evidence. Journal of Global Health. 2018;8(2):021102. doi:10.7189/jogh.08.021102
4. Wall JD, Sathirapongsasuti JF, Gupta R, Rasheed A, Venkatesan R, Belsare S, et al. South Asian medical cohorts reveal strong founder effects and high rates of homozygosity. Nat Commun. 2023;14(1):3377. doi:10.1038/s41467-023-38766-1
5. Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, et al. The promise of discovering population-specific disease-associated genes in South Asia. Nat Genet. 2017;49(9):1403–7. doi:10.1038/ng.3917
6. Machha P, Gopalan A, Elangovan Y, Veeravalli SCM, Sowpati DT, Thangaraj K. Endogamy and high prevalence of deleterious mutations in India: evidence from strong founder events. Journal of Genetics and Genomics. 2025;52(4):570–82. doi:10.1016/j.jgg.2025.02.001
7. Kotambail A, Selvam P, Muthusamy K, Thomas M, Sudhakar SV, Ghati C, et al. Clustering of Juvenile Canavan disease in an Indian community due to population bottleneck and isolation: genomic signatures of a founder event. Eur J Hum Genet. 2023;31(1):73–80. doi:10.1038/s41431-022-01198-4
8. Mukherjee S, Dutta S, Majumdar S, Biswas T, Jaiswal P, Sengupta M, et al. Genetic defects in Indian Wilson disease patients and genotype–phenotype correlation. Parkinsonism & Related Disorders. 2014;20(1):75–81. doi:10.1016/j.parkreldis.2013.09.021
9. Aggarwal A, Chandhok G, Todorov T, Parekh S, Tilve S, Zibert A, et al. Wilson Disease Mutation Pattern with Genotype-Phenotype Correlations from Western India: Confirmation of p.C271 as a Common Indian Mutation and Identification of 14 Novel Mutations. Annals of Human Genetics. 2013;77(4):299–307. doi:10.1111/ahg.12024
10. Singh N, Kallollimath P, Shah MH, Kapoor S, Bhat VK, Viswanathan LG, et al. Genetic analysis of ATP7B in 102 south Indian families with Wilson disease. Janecke AR, editor. PLoS ONE.

2019;14(5):e0215779. doi:10.1371/journal.pone.0215779

11. Bashyam MD, Chaudhary AK, Kiran M, Nagarajaram HA, Devi RR, Ranganath P, et al. Splice, Insertion-Deletion and Nonsense Mutations that Perturb the Phenylalanine Hydroxylase Transcript Cause Phenylketonuria in India. J of Cellular Biochemistry. 2014;115(3):566–74. doi:10.1002/jcb.24692
12. Singh K, Bijarnia-Mahay S, Ramprasad VL, Puri RD, Nair S, Sharda S, et al. NGS-based expanded carrier screening for genetic disorders in North Indian population reveals unexpected results – a pilot study. BMC Med Genet. 2020;21(1):216. doi:10.1186/s12881-020-01153-4
13. Datta S, Jena D, Ghosh A, Biswas VK, Jena S, Swain M, et al. Fine-scale genetic structure and population-specific clinically relevant variants in the indigenous tribal populations of India [Internet]. Genetic and Genomic Medicine; 2026 [cited 2026 Aug 11]. Available from: http://medrxiv.org/lookup/doi/10.64898/2026.01.11.26343796 doi:10.64898/2026.01.11.26343796
14. Lee E, Orton K. Cystic fibrosis carrier screening in Australia: comparing sequencing and targeted panels across diverse ancestries. J Med Genet. 2025;62(3):219–26. doi:10.1136/jmg-2024-110365
15. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019;51(11):1560–5. doi:10.1038/s41588-019-0528-2
16. Shoraka HR, Haghdoost AA, Baneshi MR, Bagherinezhad Z, Zolala F. Global prevalence of classic phenylketonuria based on Neonatal Screening Program Data: systematic review and meta-analysis. Clin Exp Pediatr. 2020;63(2):34–43. doi:10.3345/kjp.2019.00465
17. Almási T, Guey LT, Lukacs C, Csetneki K, Vokó Z, Zelei T. Systematic literature review and meta-analysis on the epidemiology of methylmalonic acidemia (MMA) with a focus on MMA caused by methylmalonyl-CoA mutase (mut) deficiency. Orphanet J Rare Dis. 2019;14(1):84. doi:10.1186/s13023-019-1063-z
18. Colburn R, Lapidus D. An analysis of Pompe newborn screening data: a new prevalence at birth, insight and discussion. Front Pediatr. 2024;11:1221140. doi:10.3389/fped.2023.1221140
19. Moorthie S, Cameron L, Sagoo GS, Bonham JR, Burton H. Systematic review and meta-analysis to estimate the birth prevalence of five inherited metabolic diseases. J of Inher Metab Disea. 2014;37(6):889–98. doi:10.1007/s10545-014-9729-0
20. Huang K, Zhang S, Wang H, Qu Y, Lu Y, Li R, et al. Autonomous biomedical research with an artificial intelligence agent. Science. 2026;eadz4351. doi:10.1126/science.adz4351
21. Bhattacharyya C, Subramanian K, Uppili B, Biswas NK, Ramdas S, Tallapaka KB, et al. Mapping genetic diversity with the GenomeIndia project. Nature Genetics. 2025;57:767–773. doi:10.1038/s41588-025-02153-x
22. Mondal D, Bhattacharyya C, Shekhawat DS, Tada NG, Rajial T, Parameswaran AS, et al. Capturing India's phenotypic diversity: Health insights from the GenomeIndia project. medRxiv. 2026. doi:10.64898/2026.04.01.26349926
23. Kaur R, Attri SV, Saini AG, Sankhyan N. A high frequency and geographical distribution of MMACHC R132 mutation in children with cobalamin C defect. Amino Acids. 2021;53:253–264.

doi:10.1007/s00726-021-02942-8

24. Mühl A, Möslinger D, Item CB, Stöckler-Ipsiroglu S. Molecular characterisation of 34 patients with biotinidase deficiency ascertained by newborn screening and family investigation. Eur J Hum Genet. 2001;9(4):237–243. doi:10.1038/sj.ejhg.5200620

# Figures

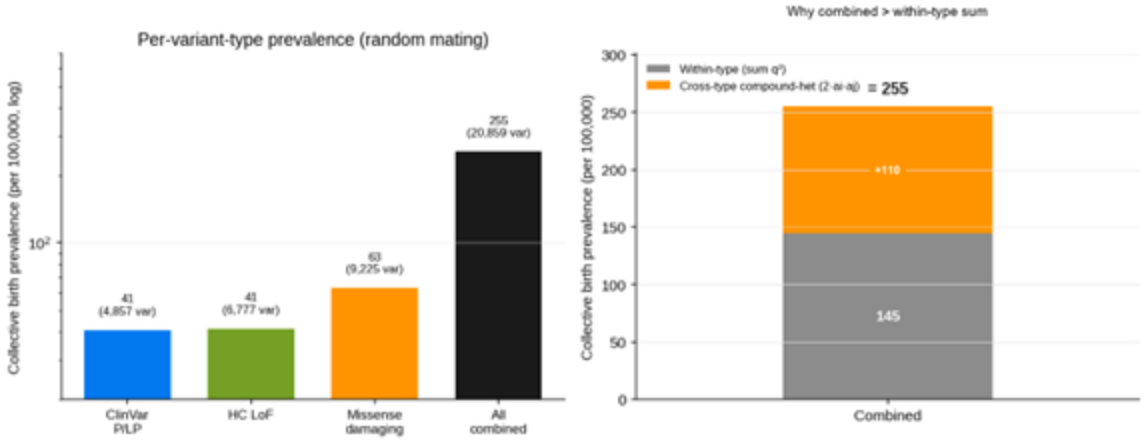


**Figure 1**

Collective AR-IEM birth prevalence by variant type under random mating (left), and decomposition of the combined estimate into within-type and cross-type compound-heterozygous components (right; illustrating why combined prevalence [255] exceeds the within-type sum [145]).

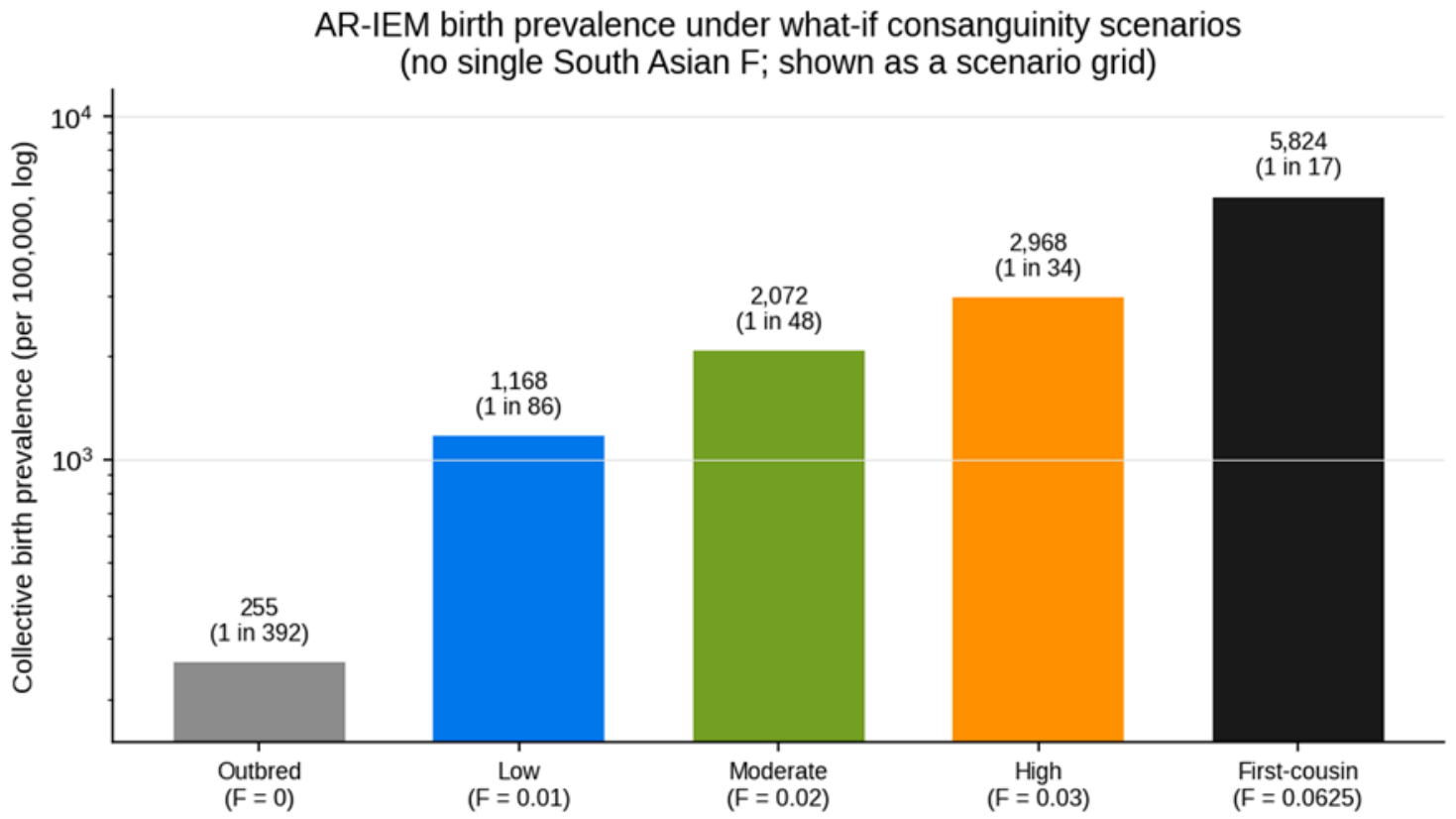


**Figure 2**

Collective AR-IEM birth prevalence under what-if consanguinity scenarios. No single inbreeding coefficient represents all of South Asia; each bar is a labelled scenario.

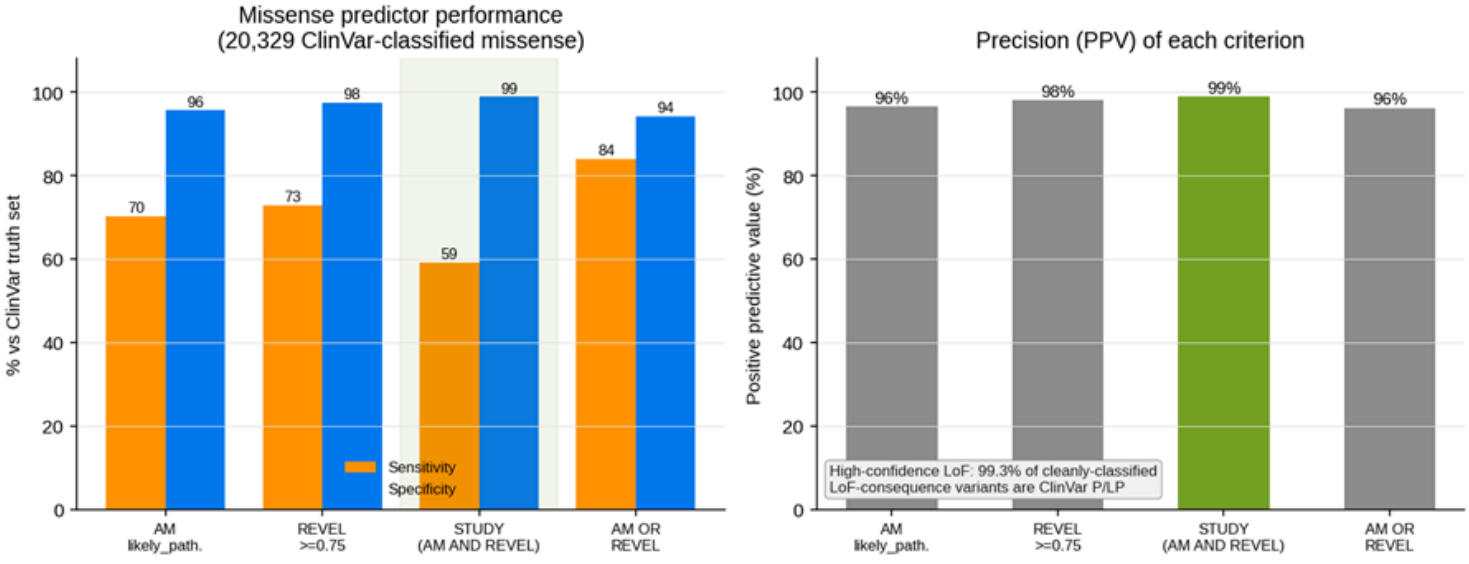


**Figure 3**

Validation of the computational variant classes against ClinVar. Left: sensitivity and specificity of each missense criterion; the study criterion (AlphaMissense intersect REVEL≥0.75) is the most conservative.

Right: positive predictive value; high-confidence LoF is supported by 99.3% ClinVar P/LP enrichment.

## Supplementary Files

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

- genomeindiacomparison.zip
- reproducibility.zip
- Supplementarymaterial.docx