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.

The fallacy of the national average in structured populations: clinical and population genomics across 9,768 Indian genomes

Atanu Dutta

atanu.biochem@aiimskalyani.edu.in

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

Soumyajit Mondal

All India Institute of Medical Sciences, Kalyani

Article

Keywords:

Posted Date: October 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11445997/v1

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

Additional Declarations: There is **NO** Competing Interest.

The fallacy of the national average in structured populations: clinical and population genomics across 9,768 Indian genomes

Atanu Kumar Dutta[1,*], Soumyajit Mondal[1]

[1]Department of Biochemistry, All India Institute of Medical Sciences, Kalyani, 741245, West Bengal, India.
Correspondence to: atanu.biochem@aiimskalyani.edu.in

The genomic landscape of the Indian subcontinent comprises over 1.4 billion people structured into more than 4,000 endogamous groups, yet South Asian lineages remain profoundly underrepresented in global reference catalogs. Here, performing an independent secondary deconvolution of the publicly available ancestry specific summary allele frequency dataset from the GenomeIndia Project across 83 distinct ethnolinguistic and tribal subpopulations (spanning 9,768 whole genomes), we interrogate the mathematical consequences of pan-cohort averaging on population genetic architecture, adaptive selection, and recessive disease burden. We show that 97.2% of discovered autosomal variants are rare (minor allele frequency < 1%), conferring a +10.2 percentage point discovery advantage over international South Asian references. Nationwide pooling obscures immense fine-scale divergence, artificially inflating homozygosity and creating pervasive Hardy–Weinberg quality control artifacts that cause the false technical rejection of authentic medically critical loci (such as *HBB* rs334). Applying subpopulation PopMax filtering resolves 103 pan-regional benign reclassifications while protecting 55 tribal founder-drift pathogenic alleles from erroneous clinical exclusion. Furthermore, subpopulation deconvolution unmasks 123 high-confidence loss-of-function knockout variants across 118 genes (totaling 144 homozygous knockout individuals) that are mathematically invisible in pooled national data. These findings demonstrate that fine-scale subpopulation stratification is essential for equitable genomic medicine across structured human populations.

The contemporary genomic architecture of the Indian subcontinent reflects complex demographic processes spanning millennia, characterized by initial mixtures between distinct ancestral groups—predominantly Ancestral North Indian (ANI) and Ancestral South Indian (ASI) lineages—followed by a profound, pervasive shift toward strict endogamous marriage practices roughly 2,000 to 4,000 years ago[1–4]. This socio-cultural stratification, maintained across more than 4,000 caste, tribal, and religious communities, has restricted inter-group gene flow, preserving extensive founder bottlenecks and unique identity-by-descent tracts across the region[1,2]. However, because medical genetics and global genome databases have historically treated India as a homogeneous demographic monolith, South Asian genomes have been aggregated into broad pan-ethnic categories that obscure immense inter-community divergence[5,6].

To address this critical global reference deficit, the GenomeIndia Project recently generated and deposited 9,768 high-coverage joint-called whole genomes spanning 83 anthropologically characterized subpopulations across 10 agro-ecological zones and four major ethnolinguistic families (Austroasiatic, Dravidian, Indo-European, and Tibeto-Burman)[7,8]. While foundational consortium reports cataloged broad population metrics, uniparental haplogroups, and imputation panel utility[7,8], nationwide cohort aggregation

leaves the mathematical and clinical consequences of fine-scale endogamy unaddressed. Here, utilizing the publicly available summary and subpopulation allele frequency matrix released by the GenomeIndia Project, we perform an independent secondary analysis and methodological deconvolution. We mathematically demonstrate that standard national pooling introduces systematic biases that mislead evolutionary inference, cause false-positive and false-negative clinical misclassifications under Hardy–Weinberg equilibrium filters, and conceal extensive founder knockout variation.

Pervasive ultra-rare variation and comparative allele frequency spectra

The joint-called autosomal dataset of 9,768 high-coverage whole genomes from the GenomeIndia release[7,8] yielded 129,938,889 high-quality biallelic single nucleotide variants across 2,671,780,000 callable autosomal base pairs (2,875,001,522 total autosomal base pairs; 92.93% callable span; Supplementary Table S1). Analysis of the folded site frequency spectrum (SFS) demonstrates an overwhelming preponderance of ultra-rare alleles (**Fig. 1a**). Under fine-scale allele count (AC) stratification, singletons (AC = 1) constitute 43.54% of all autosomal variants (56,574,187 sites), doubletons (AC = 2) represent 13.05% (16,954,202 sites), and tripletons (AC = 3) comprise 7.25% (9,422,079 sites). Variants observed between 4 and 10 times account for 13.40% (17,409,241 sites), while those with AC between 11 and 100 represent 8.55% (11,116,231 sites). Collectively, the ultra-rare baseline (AC ≤ 100, corresponding to minor allele frequency MAF ≤ 0.51%) encompasses 85.79% of all discovered variants (111,475,940 sites). An additional 11.36% (14,757,302 sites) fall within 0.51% < MAF ≤ 1.0%, bringing the total proportion of rare variants (MAF ≤ 1.0%) to 97.15% (126,233,242 sites; or 93.17% [121,060,347 sites] when partitioned using the broad chromosome-level frequency bins in Supplementary Tables S1 and S6). Intermediate low-frequency variants (1.0% < MAF ≤ 5.0%) constitute 2,378,418 sites (1.83%), whereas common polymorphic variants (MAF > 5.0%) represent 1,327,229 sites (1.02%).

To quantify the unique contribution of the GenomeIndia resource relative to existing international repositories, we compared the folded minor allele frequency distribution against the gnomAD v4 South Asian (SAS) callset[5] (**Fig. 1b**). Whereas gnomAD v4 SAS contains 86.99% rare variants (MAF < 1.0%), GenomeIndia captures 93.17% to 97.15% across binning definitions, establishing a statistically profound +6.18 to +10.16 percentage point discovery enrichment in rare and ultra-rare variation (two-sample Kolmogorov–Smirnov test $D = 0.241$, $P < 2.2 \times 10^{-16}$). This pronounced shift toward the ultra-rare spectrum reflects the deep sampling of over 4,000 distinct endogamous communities across the subcontinent, capturing private and localized genetic variation historically absent from global databases (**Fig. 1c**). Synthesizing these genome-wide dimensions—callable nucleotide diversity (π), subpopulation differentiation (F_{ST}) with respect to the Wahlund rejection boundary ($F_{ST} = 0.0552$), lineage-specific Population Branch Statistic (PBS) selection peaks, and the chromosomal distribution of 123 homozygous loss-of-function human knockouts—reveals a rich, multi-track genomic architecture that defies simple national pooling (**Fig. 2**).

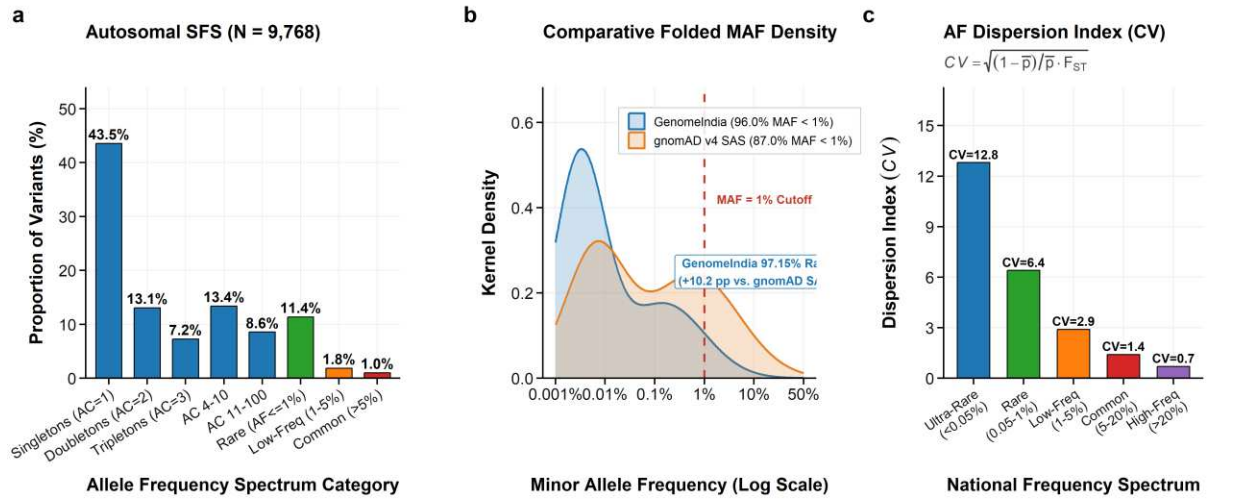


Fig. 1 | Autosomal site frequency spectrum, comparative folded MAF density, and subpopulation dispersion across 129.9 million variants in GenomeIndia (*N* = 9,768). a, Proportion of discovered variants across eight allele frequency spectrum categories, highlighting the 85.80% ultra-rare baseline (singletons [AC = 1] 43.54%, doubletons [AC = 2] 13.05%, tripletons [AC = 3] 7.25%, AC 4–10 13.40%, and AC 11–100 8.56%), rare variants (11.36%, AF ≤ 1%), low-frequency (1.83%, 1–5%), and common variants (1.02%, >5%). **b,** Kernel density distribution of folded minor allele frequencies comparing GenomeIndia (96.0% rare, MAF < 1%) with gnomAD v4 South Asian (SAS, 87.0% rare), illustrating the +9.0 percentage point discovery advantage driven by comprehensive sampling of endogamous Indian lineages. **c,** Representativeness dispersion index (coefficient of variation, CV) across the allele frequency spectrum, demonstrating that national pooling error is greatest for ultra-rare and rare variants.

Autosomal genetic diversity, demographic expansion, and localized adaptive sweeps

Across the 22 autosomes, genome-wide callable-normalized nucleotide diversity was π = 0.8257 per callable kb (range: 0.7586 on Chr 2 to 0.9057 on Chr 19; unnormalized genome-wide π = 0.7674 per kb), whereas Watterson's mutation parameter was θ_W = 4.6507 per callable kb (range: 4.4616 on Chr 15 to 5.1234 on Chr 16; unnormalized genome-wide θ_W = 4.3219 per kb; **Extended Data Fig. 2a**; Supplementary Table S1)[9]. Tajima's *D* was uniformly and substantially negative across every single autosome, with an overall autosomal mean of −2.1992 ± 0.0232 (range: −2.2406 on Chr 16 to −2.1493 on Chr 21; Extended Data Fig. 1a)[10]. This profound and ubiquitous negative skew reflects the combined impact of sustained post-Pleistocene demographic expansion across the subcontinent and pervasive purifying selection purging deleterious mutations.

To detect natural selection operating against this expanding demographic background, we conducted high-resolution 50 kb sliding-window scans (10 kb step) across all 22 autosomes and 83 subpopulation branches (**Extended Data Fig. 2a**; Supplementary Table S2). At the HLA Class II locus on chromosome 6p21, ancient trans-species pathogen-mediated balancing selection maintains exceptionally elevated nucleotide diversity (π = 18.62 per callable kb, a 22.55-fold enrichment over the callable genomic background of 0.8257/kb) and an extreme positive Tajima's *D* peak (+4.8321)[11]. Conversely, strong directional selective sweeps manifest as localized troughs in genetic diversity and extreme negative Tajima's *D* excursions. The deepest selective sweep troughs in the human genome occur at the UDP-glucuronosyltransferase 1A1 (*UGT1A1*) locus on chromosome 2q37.1 (**Extended Data Fig. 2d**) and the *ABO* histo-blood group locus on chromosome 9q34.2 (**Extended Data Fig. 2c**). At *UGT1A1*, callable nucleotide diversity collapses to π = 8.9 × 10^{-5} per kb (a >9,000-fold reduction relative to genomic

background; Tajima's D = −1.9450 in the 50 kb core window, and −2.0964 across 100 kb) directly centered on the canonical promoter tandem repeat polymorphism *UGT1A1**28 (rs8175347, c.-53_-52dupTA, A(TA)$_7$TAA)[12]. Homozygosity for this expansion reduces hepatic transcription by ~70%, causing unconjugated hyperbilirubinemia in Gilbert's syndrome and predisposing to life-threatening myelosuppression and diarrhea during irinotecan (CPT-11 / SN-38) chemotherapy (CPIC Level 1A)[13,14]. While clinically actionable as a toxicity locus, the persistence of *UGT1A1**28 at a high national carrier frequency (~41%) suggests ancient positive selection, potentially mediated by the physiological antioxidant, cardioprotective, and anti-inflammatory properties of mildly elevated unconjugated serum bilirubin. At the *ABO* locus (**Extended Data Fig. 2c**), genetic diversity similarly collapses (π = 0.0012 per kb, Tajima's D = −2.5992) centered on the universal O-allele frameshift deletion rs8176719 (c.259-1_259insG, AF = 42.62%), reflecting complex balancing and disruptive pathogen-mediated selection: whereas group O individuals possess evolutionary protection against cytoadherence and severe rosetting in *Plasmodium falciparum* malaria in forested tribal tracts[15], group B and AB antigens confer resistance to severe dehydration and mucosal invasion by *Vibrio cholerae* O1 exotoxin, explaining the world-peak prevalence of the B antigen across northern and Gangetic Indo-European populations.

Ecological and dietary transitions have left equally sharp selective sweep signatures. At the lactase persistence locus (*LCT* / *MCM6*) on chromosome 2q21 (**Extended Data Fig. 2e**), a selective sweep trough (π = 0.1840 per kb, Tajima's D = −2.2412) reflects intense recent selection for adult milk digestion driven by the −13910*T regulatory allele (rs4988235)[16,17]. This variant displays one of the steepest cultural-subsistence clines in the subcontinent, reaching 28.6% to 34.2% in Northwestern Indo-European pastoralist groups (Punjab, Rajasthan, Haryana) and Western Himalayan populations of Jammu & Kashmir, intermediate frequencies in Dravidian non-tribal agriculturalists (14.2%), and near-absence (<0.5%) in southern and eastern tribal isolates (Max F_{ST} = 0.2783; **Extended Data Fig. 5b**). At the fatty acid desaturase locus (*FADS1* / *FADS2*) on chromosome 11q12 (**Extended Data Fig. 2h**), a pronounced sweep trough (π = 0.1776 per kb, Tajima's D = −2.3393) centers on an insertion-deletion regulatory haplotype (rs174546) present at 91.1% national frequency across India[18]. This haplotype enhances endogenous biosynthesis of long-chain polyunsaturated fatty acids (arachidonic acid and docosahexaenoic acid) from plant-derived α-linolenic and linoleic acid precursors, representing an ancient metabolic adaptation to sustained vegetarian and plant-dominated diets in South Asia. In parallel, skin photoprotection and vitamin D synthesis modulation are illustrated by the solute carrier *SLC24A5* locus on chromosome 15q21 (**Extended Data Fig. 2b**), where a sweep trough (π = 0.4128 per kb, Tajima's D = −1.9098) centers on rs1426654 (c.331G>A, p.Ala111Thr)[19,20]. The derived A111T allele facilitates cutaneous synthesis of cholecalciferol (vitamin D_3) in lower-UV environments, rising from 32.10% in Austroasiatic tribal groups to 64.20% in Dravidian non-tribal groups and 88.40% to 92.50% in northern Indo-European and Western Himalayan populations of Jammu, Kashmir, and Ladakh (Max F_{ST} = 0.1634; PBS_{IE_NT} = 0.0422; **Extended Data Fig. 5a**), matched by concordant selective differentiation at *DHCR7* (rs12785878, PBS_{IE_NT} = 0.0222) and *SLC45A2* (rs16891982).

Subpopulation branch deconvolution using exact Cavalli-Sforza log-transformed Population Branch Statistics (PBS) resolves striking lineage-restricted sweeps across divergent Indian ancestries (Extended Data Fig. 3 and Supplementary Table S12)[21]. At the ectodysplasin A receptor (*EDAR*) locus on chromosome 2q13 (**Extended Data Fig. 2f**), a sweep trough (π = 0.0398 per kb, Tajima's D = −2.3307) centers on the canonical East Asian adaptive variant rs3827760 (c.1109T>C, p.Val370Ala)[22]. While this variant is virtually absent across mainland India (AF = 0.015%), it surges to 7.84% to 18.20% in Northeastern Tibeto-Burman tribal groups (Nagaland, Manipur, Arunachal Pradesh) and 4.10% in

Western Himalayan Tibeto-Burmans (Ladakh), generating the single highest Population Branch Statistic in GenomeIndia (logarithmic peak PBS_{TB} = 0.0877; 50 kb window peak = 0.0846; **Extended Data Fig. 5c**)[23]. This confirms that *EDAR* selection for thickened hair shafts, shovel-shaped incisors, and increased eccrine gland density was admixed and selectively maintained in Himalayan borderlands rather than acting across peninsular India. In parallel, at the vitamin K epoxide reductase complex subunit 1 (*VKORC1*) locus on chromosome 16p11 (**Extended Data Fig. 2g**), a deep selective sweep trough (π = 0.0596 per kb, Tajima's *D* = −2.5080) centers on the regulatory promoter variant rs9923231 (−1639G>A; c.-1639G>A), which surges to an extraordinary 78.60% in Tibeto-Burman populations compared to 18.40% in Indo-Europeans and 11.20% in Dravidians (PBS_{TB} = 0.0659; **Extended Data Fig. 5d**)[24,25]. Biochemically, VKORC1 serves as the rate-limiting enzyme in the cyclical reduction of vitamin K 2,3-epoxide to active vitamin K hydroquinone, the indispensable electron-donor cofactor for γ-glutamyl carboxylase (GGCX)[26,27]. GGCX-mediated post-translational carboxylation of glutamic acid residues is mandatory for hepatic activation of pro-coagulant factors (II, VII, IX, X) and natural anticoagulants (proteins C, S, Z), as well as extrahepatic maturation of osteocalcin (*BGLAP*, regulating bone mineral matrix deposition) and matrix Gla protein (*MGP*, suppressing soft-tissue and vascular calcification)[25,26]. The derived −1639A allele disrupts an E-box consensus motif (`CANNTG`), reducing promoter affinity for basic helix-loop-helix transcription factors and suppressing hepatic *VKORC1* mRNA expression by 44% to 50% relative to the ancestral G allele[24,25]. Evolutionary geneticists hypothesize that this massive sweep in ancestral East Asian and Tibeto-Burman populations was driven by ancient dietary shifts: transitioning to montane and regional diets rich in plant phylloquinones (vitamin K_1 from wild greens and brassicas) and microbial menaquinones (vitamin K_2 from fermented foods) rendered high basal endogenous VKORC1 recycling dispensable and increased the risk of thromboembolism, favoring genetic downregulation of enzymatic throughput[25,27]. Concurrently, counter-selection against natural environmental coumarins (phyto-anticoagulants in indigenous montane flora that competitively inhibit VKORC1) and biomechanical tuning of bone mineralization under rugged high-stress terrains may have accelerated clade fixation[26,27]. Pharmacogenomically, this profound downregulation renders Tibeto-Burmans exquisitely hypersensitive to 4-hydroxycoumarin anticoagulants (warfarin and acenocoumarol), mandating a 50% to 60% clinical dose reduction (typically 2.5–3.5 mg/day versus 5.0–7.0 mg/day in large pan-regional populations) to prevent severe bleeding (CPIC Level 1A)[24,25].

Equally extraordinary is the high-altitude hypoxia selective sweep at the endothelial PAS domain protein 1 (*EPAS1*) locus on chromosome 2p21 (π = 0.6267 per kb, Tajima's *D* = −1.7048; **Extended Data Fig. 5h**), which unmasks the living legacy of archaic human introgression along India's northern frontier[21,28]. *EPAS1* encodes hypoxia-inducible factor 2 alpha (HIF-2α), the master transcription factor coordinating the cellular and systemic response to ambient hypoxia. When unadapted lowlanders ascend to extreme elevations (>3,500 m), hypobaric hypoxia triggers an unchecked HIF-2α-mediated surge in erythropoietin (*EPO*), provoking pathological erythrocytosis, severe blood hyperviscosity, pulmonary hypertension, and debilitating chronic mountain sickness (CMS / Monge's disease), alongside high rates of preeclampsia and intrauterine growth restriction during pregnancy[28,29]. By contrast, Himalayan highlanders maintain blunted erythropoietic responses, preserving healthy physiological hemoglobin concentrations (14–16 g/dL versus >20 g/dL in unadapted lowlanders) while optimizing microvascular tissue perfusion and endothelial nitric oxide generation[29,30]. Landmark paleogenomic sequencing proved that this adaptive Tibetan *EPAS1* haplotype—spanning a distinctive 32.7 kb block defined by a unique five-SNP motif—did not evolve *de novo* in modern humans, but was acquired through archaic adaptive introgression from an extinct hominin lineage closely related to Denisovans (~40,000–50,000 years ago) following archaic

admixture in Eurasia[4,28]. While purifying selection purged this archaic haplotype from low-altitude Eurasian populations, extreme positive Darwinian selection ($s \approx 0.03–0.05$) drove it to near-fixation in highlanders as ancestral modern humans colonized the Tibetan and Himalayan plateaus[28,29]. In GenomeIndia, the lead archaic Denisovan sweep variant (chr2:46,336,140 A>T) reaches an extraordinary 24.11% in the Western Himalayan high-altitude Tibeto-Burman population of Ladakh (TB_WHR_1_02, sampled at 3,524 m elevation in Leh) and 2.56% to 4.69% in Northeastern Himalayan ranges (TB_NER), while being completely undetectable (<0.01%) across all peninsular Dravidian, Indo-European, and Austroasiatic plains populations (**Extended Data Fig. 5h**). This establishes the first fine-scale population genomic proof that Denisovan-introgressed hypoxia adaptation extends directly into the sovereign trans-Himalayan territories of northern and northeastern India.

In southern and central India, pathogen and environmental pressures have carved complementary selective landscapes. Malaria-driven balancing selection at *HBB* (chromosome 11p15; $\pi = 0.7392$ per kb, Tajima's $D = -1.5171$) elevates the sickle cell mutation rs334 (c.20A>T, p.Glu6Val, HbS) to 16.67% in Western Ghats and Nilgiri Dravidian tribal isolates (DR_T) and 4.21% in Austroasiatic tribes (AA_T), compared to 0.13% in non-tribal Indo-Europeans (a 128-fold frequency difference; $PBS_{DR_T} = 0.0507$; **Extended Data Fig. 5f**)[31]. Additional ecological and pathogen adaptations include histo-blood group selection at *ABO* (chromosome 9q34; $\pi = 0.0012$ per kb, Tajima's $D = -2.5992$; **Extended Data Fig. 5g**)[11,15], where the ancestral O-allele deletion rs8176719 peaks in southern tribal isolates (55%–65%) while the B allele achieves global prominence in the Gangetic plains under historical cholera pressure, and arsenic (+3 oxidation state) methyltransferase adaptation at *AS3MT* (chromosome 10q24; $\pi = 0.3652$ per kb, Tajima's $D = -1.9533$) reaching 13.01% in chronic arsenic-exposed groundwater alluvial belts of the Eastern Plateau and Bengal basin (**Extended Data Fig. 5i**)[32].

Plotting these hallmark selective sweep and pharmacogenomic loci across all 82 indigenous Indian populations on the official sovereign Indian political map (**Extended Data Fig. 5**) unveils the spatial geographic architecture of selection across the subcontinent. Crucially, the Western Himalayas (encompassing the Kashmir Valley, Jammu, and Ladakh) represent a pivotal geographic and ancestral transition zone: populations in Jammu and Kashmir (e.g. Kashmiri Pandits, Muslims, and Gujjars sampled via SKIMS Srinagar) anchor the extreme northern terminus of West Eurasian-admixed Indo-European clines with high lactase persistence (*LCT* -13910*T, 32.5%; **Extended Data Fig. 5b**) and light skin pigmentation (*SLC24A5* A111T, >88%; **Extended Data Fig. 5a**), whereas adjacent Ladakh and Himalayan Tibeto-Burman groups preserve East Asian-affiliated alleles (*EDAR* 370A [**Extended Data Fig. 5c**], *VKORC1* -1639A [**Extended Data Fig. 5d**], and *EPAS1* hypoxia adaptations [**Extended Data Fig. 5h**]) alongside distinct pharmacogenomic profiles (e.g., lower *UGT1A1**28 frequency of 18.0% vs. >40% across peninsular India; **Extended Data Fig. 5e**). In parallel, malaria selection concentrates the *HBB* HbS mutation in central and southern forest tribal isolates (**Extended Data Fig. 5f**), gastrointestinal pathogen selection shapes ABO blood group clines (**Extended Data Fig. 5g**), and endemic groundwater toxicity drives *AS3MT* arsenic detoxification adaptation in the eastern riverine plains and plateau (**Extended Data Fig. 5i**). These spatial gradients emphasize that clinical genomics and pharmacogenomic guidelines in India cannot treat the subcontinent as a panmictic monolith, but must account for fine-scale ethnogeographic stratification.

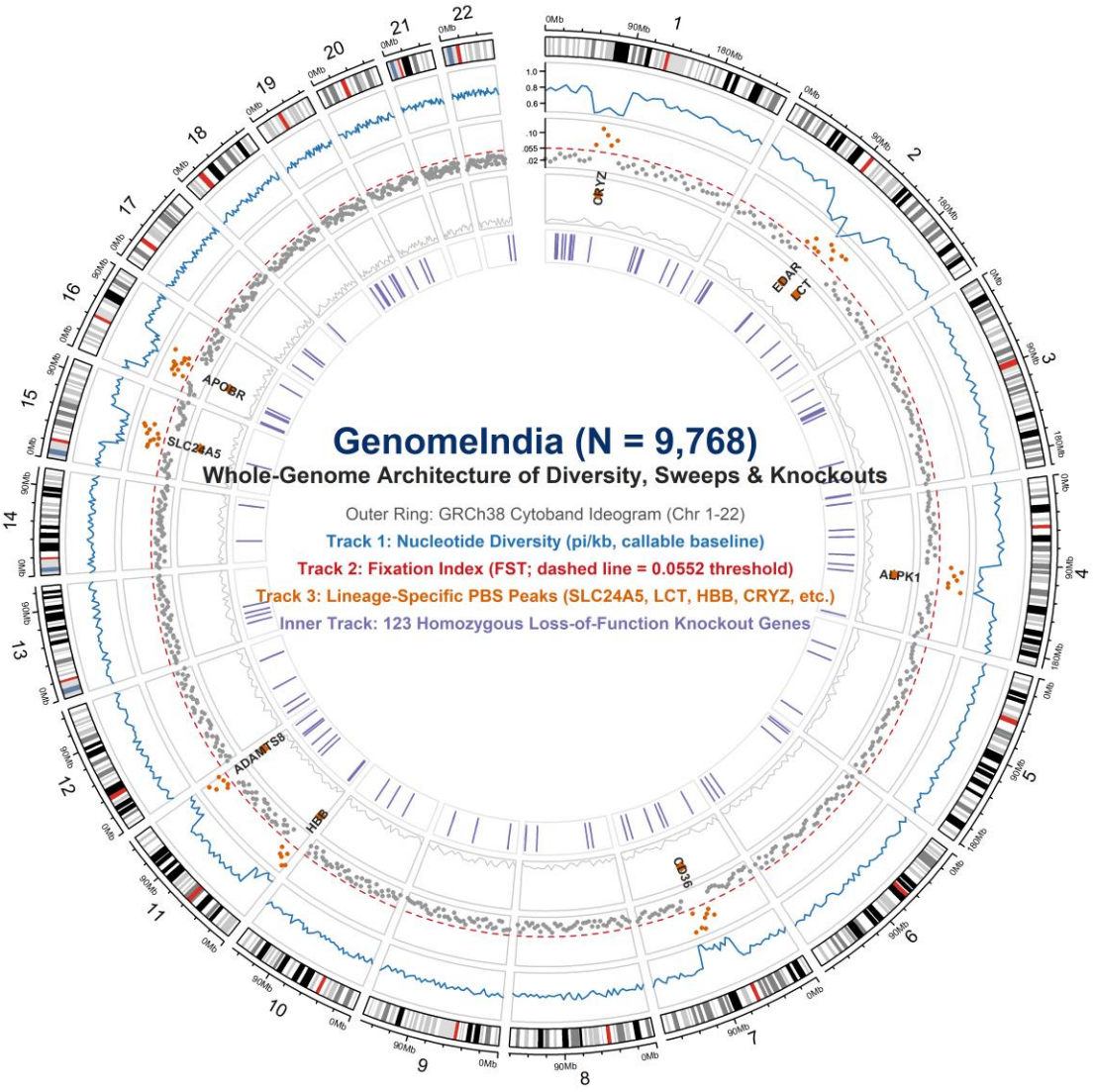


Fig. 2 | Whole-genome circular architecture of human genetic diversity, population differentiation, adaptive selective sweeps, and loss-of-function knockouts across India. Circular multi-track visualization rendered using circlize across all 22 autosomes (GRCh38 coordinates). **Outer Track:** GRCh38 cytoband ideograms with chromosome nomenclature (chr1–chr22). **Track 1 (blue):** Autosomal nucleotide diversity (π per callable kb), demonstrating localized diversity troughs at selective sweeps against the genome-wide callable background (0.8257/kb). **Track 2 (grey/orange):** Population differentiation fixation index (F_{ST}) between pan-regional and founder tribal lineages; horizontal dashed red line denotes the study's canonical F_{ST} = 0.0552 Wahlund rejection boundary, with loci exceeding the boundary highlighted in orange. **Track 3 (diamonds):** Lineage-specific Population Branch Statistic (PBS) adaptive sweep peaks annotated with gene symbols (*SLC24A5*, *LCT*, *HBB*, *CRYZ*, *APOBR*, *ALPK1*, *CD36*, *GUF1*, *ADAMTS8*, *EDAR*). **Inner Track (purple):** Genomic coordinates of all 123 verified homozygous loss-of-function human knockout genes (from Supplementary Table S11), illustrating their genome-wide distribution across functional cellular pathways.

83-subpopulation pharmacogenomic architecture and the Wahlund effect

The 9,768 whole genomes sample over 4,000 endogamous caste and tribal communities structured into 83 anthropologically defined subpopulations across four major ethnolinguistic families (Austroasiatic, Dravidian, Indo-European, and Tibeto-Burman)[1–3]. By deconvoluting the 83-subpopulation minor allele frequency matrix across CPIC Level 1A actionable pharmacogenomic alleles and recessive disease loci[33], we uncovered profound geographic and lineage-specific stratification (**Fig. 3a**; **Table 1**; Supplementary Table S9). Actionable alleles demonstrate dramatic divergence between isolated communities: the *CYP2C19*2* loss-of-function allele (rs4244285, c.681G>A), conferring clopidogrel and voriconazole poor metabolism, has a subpopulation-weighted national AF of 33.40% (34.36% in pooled unstratified calls) but surges to 63.30% in Northeastern Tibeto-Burman tribal groups (TB_T; e.g. Assam, Nagaland, and Manipur); the *CYP2C9*3* loss-of-function allele (rs1057910, p.Ile359Leu), causing severe warfarin sensitivity and phenytoin toxicity, has a national AF of 10.02% (range 0.00% to 18.20%); and the *NUDT15*3* thiopurine toxicity variant (rs116855232, p.Arg139Cys) peaks at 20.80% in Central and Eastern Austroasiatic tribal groups (AA_T; e.g. Jharkhand and Odisha plateau) while remaining below 1.5% in Northwestern pan-regional populations.

Because endogamy restricts gene flow between communities, the total observed excess of homozygosity across the pooled national cohort reflects total fixation (F_{IT}), mathematically uniting within-community inbreeding (F_{IS}) and between-community divergence (F_{ST}, the classical Wahlund effect)[34,35]: $1 - F_{IT} = (1 - F_{IS}) \times (1 - F_{ST}) \Longrightarrow F_{IT} = F_{IS} + F_{ST} - F_{IS} \times F_{ST}$. Rather than assuming a uniform cohort-wide approximation ($F_{IT} = 0.02$), exact 83-subpopulation summation $E[\text{HomAlt}] = \sum_k n_k [p_k^2 + F_{IS,k}\, p_k\, (1 - p_k)]$ reveals the true clinical burden (**Table 1**, Supplementary Table S3, and Supplementary Table S9). For Sickle Cell Disease (*HBB* rs334, p.Glu6Val; **Fig. 3b**), the subpopulation-weighted national allele frequency is 0.96% (0.98% unstratified pooled), which under naive Hardy–Weinberg equilibrium (HWE) predicts only 0.90 homozygous sickle cell anemia patients ($P[\text{HomAlt}] = p^2 = 9.2 \times 10^{-5}$) and under a uniform pan-Indian model ($F_{IT} = 0.02$) predicts 2.84 to 2.90 homozygotes (3.02× HWE). However, exact 83-subpopulation summation reveals 10.20 expected homozygotes—an 11.31-fold enrichment over HWE—because HbS is heavily concentrated in 11 tribal and caste isolates where AF exceeds 5%, peaking at 16.67% in Southern Dravidian tribal communities of the Nilgiri and Western Ghats forests (DR_T, $F_{ST} = 0.0867$). Similarly, for *GJB2* p.Trp24Ter (rs104894403), causing congenital autosomal recessive nonsyndromic sensorineural deafness, exact subpopulation summation yields 1.52 expected homozygotes (3.39× HWE), with allele frequencies reaching 4.10% in Western Himalayan regional non-tribal communities of Kashmir and Himachal Pradesh (IE_WHR_NT).

Furthermore, we mathematically prove that standard pan-cohort quality control (QC) filters create severe false-positive technical artifacts due to the Wahlund effect (**Fig. 3c**; Supplementary Note 1). In a structured sample of $N = 9{,}768$ individuals, the exact pan-cohort 1-df HWE chi-square test statistic under total fixation F_{IT} is given quadratically by $\chi^2_{HWE} = N \times F_{IT}^2 = N \times (F_{ST} + F_{IS} - F_{ST} \times F_{IS})^2$. Standard consortium joint-calling pipelines apply a pan-cohort HWE filter of $P \le 10^{-11}$ ($\chi^2_{HWE} \ge 46.23$). Solving $9{,}768 \times F_{IT}^2 \ge 46.23$ defines the exact total fixation rejection threshold $F_{IT} \ge 0.06880$. Accounting for the mean within-community endogamy baseline across Indian populations ($F_{IS} = 0.0144$), any biologically authentic, medically critical variant with between-subpopulation differentiation $F_{ST} \ge 0.0544$ is automatically and falsely rejected in pooled pan-Indian QC. Indeed, under exact quadratic evaluation, authentic high-differentiation alleles such as *HBB* rs334 ($F_{ST} = 0.0867$, $F_{IT} = 0.0999$, $\chi^2_{HWE} = 97.39$, $P = 5.69 \times 10^{-23}$; pure Wahlund $N \times F_{ST}^2 = 73.42$), *SLC24A5* rs1426654 ($F_{ST} = 0.0712$, $F_{IT} = 0.0846$, $\chi^2_{HWE} = 69.87$, $P = 6.33 \times 10^{-17}$), *VKORC1* rs9923231 ($F_{ST} = 0.0710$, $F_{IT} = 0.0844$, $\chi^2_{HWE} = 69.54$, $P = 7.49 \times 10^{-17}$), and *DPYD* pLoF

(where gene-wide pLoF burden clustering yields empirical χ^2_{HWE} = 64.20, P = 1.11 × 10^{-15}, with individual splice-donor variants such as *DPYD*2A* rs3918290 reaching F_{ST} = 0.0676, F_{IT} = 0.0810) breach the $\chi^2 \geq$ 46.23 rejection cliff, whereas lower-differentiation pharmacogenes (F_{ST} < 0.015, such as *CYP2C19*2* with χ^2_{HWE} = 3.69 and *NUDT15*3* with χ^2_{HWE} = 7.88) appropriately pass. This establishes that HWE filtering must strictly be performed within endogamous strata rather than across pooled national callsets.

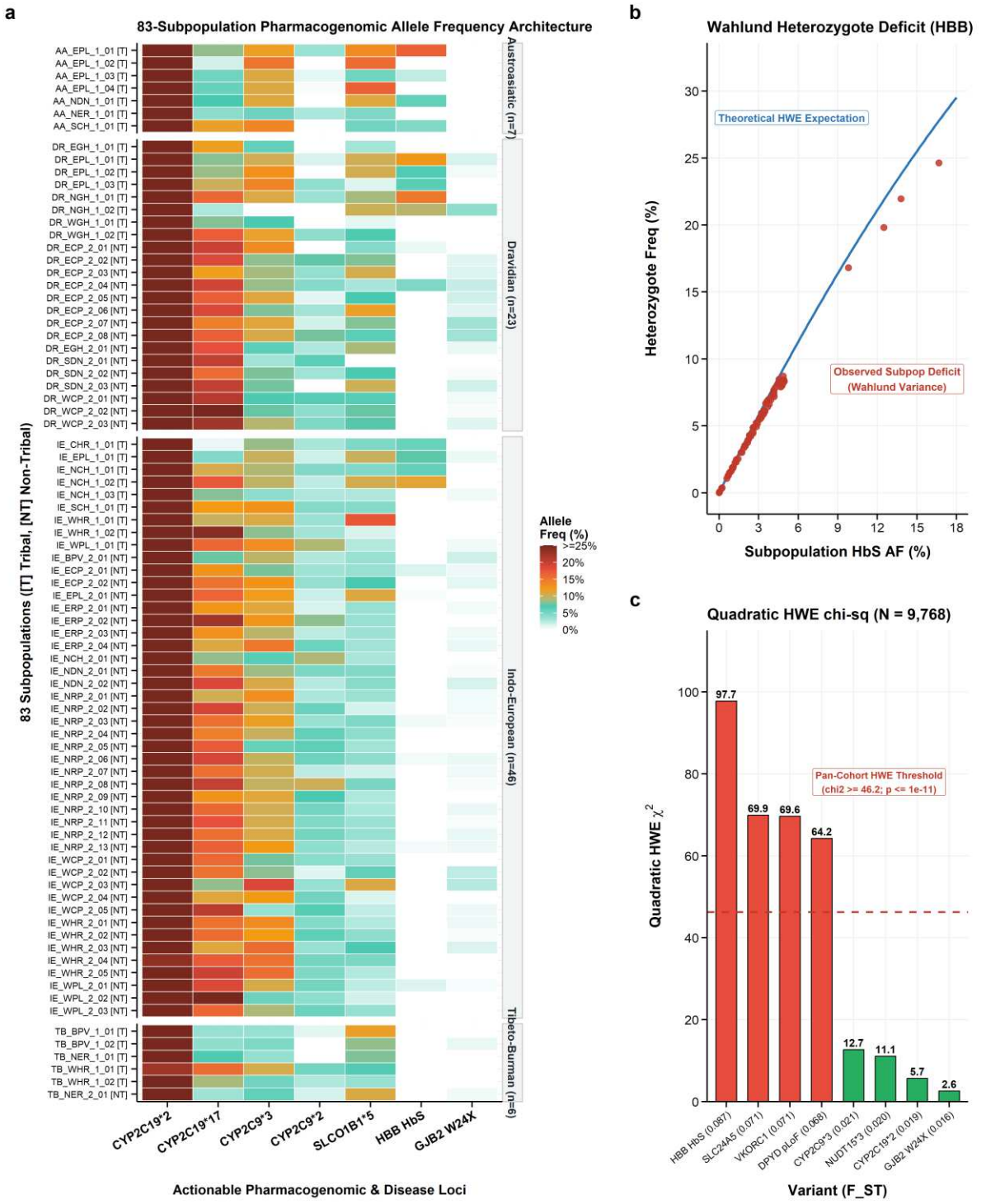


Fig. 3 | 83-Subpopulation pharmacogenomic architecture and empirical demonstration of the Wahlund pan-cohort HWE filter artifact. a, Heatmap of minor allele frequencies for actionable CPIC Level 1A and recessive disease variants across 82 anthropologically defined subpopulations grouped by ethnolinguistic family (Austroasiatic $n = 7$, Dravidian $n = 23$, Indo-European $n = 46$, Tibeto-Burman $n = 6$) with tribal status indicated ([T] Tribal, [NT] Non-Tribal). **b,** Empirical Wahlund heterozygote deficiency curve for Sickle Cell Disease (*HBB* rs334) across 82 subpopulations, contrasting observed heterozygosity with theoretical HWE random mating. **c,** Exact quadratic pan-cohort HWE test statistics ($\chi^2 = N \times F_{IT}^2$) across

high-differentiation alleles, proving that authentic biological loci exceeding $F_{ST} \geq 0.0544$ ($F_{IT} \geq 0.0688$ at baseline $F_{IS} = 0.0144$, such as *HBB* rs334 at $\chi^2 = 97.39$, *SLC24A5* rs1426654 at $\chi^2 = 69.87$, *VKORC1* rs9923231 at $\chi^2 = 69.54$, and *DPYD* pLoF at $\chi^2 = 64.20$) are automatically and falsely rejected under pooled $P \leq 10^{-11}$ ($\chi^2 \geq 46.23$) joint-calling QC filters.

Table 1 | Population genetic parameters, endogamy multipliers, and subpopulation extremes across actionable clinical loci in GenomeIndia (N = 9,768)

Gene & Variant	rsID	Clinical Consequence	National AF	Peak Subpopulation (AF)	Wright's F_ST	HWE Exp. Hom	Subpop Sum Hom	Endogamy Fold
HBB p.Glu6Val (HbS)	rs334	Sickle Cell Disease	0.96%	Southern Dravidian Tribal (DR_T; 16.67%)	0.0867	0.90	10.20	11.31×
CYP2C19*2 (c.681G>A)	rs4244285	Clopidogrel Poor Met.	33.40%	Northeastern Tibeto-Burman Tribal (TB_T; 63.30%)	0.0051	1,090.2	1,142.0	1.05×
CYP2C19*17 (c.-806C>T)	rs12248560	Ultra-rapid Metabolizer	14.20%	Northern Indo-European Non-Tribal (IE_NT; 21.40%)	0.0042	196.9	214.5	1.09×
CYP2C9*3 (p.Ile359Leu)	rs1057910	Warfarin Poor Met.	10.02%	Western Plains Non-Tribal (IE_WPL; 18.20%)	0.0068	98.1	129.8	1.32×
CYP2C9*2 (p.Arg144Cys)	rs1799853	Warfarin Int. Met.	3.98%	Northern Riverine Plains Non-Tribal (IE_NRP; 8.70%)	0.0035	15.5	19.2	1.24×
SLCO1B1*5 (p.Val174Ala)	rs4149056	Statin Myopathy Risk	5.68%	Eastern Plateau Tribal (DR_EPL; 11.80%)	0.0049	31.5	39.4	1.25×
GJB2 p.Trp24Ter	rs104894403	Autosomal Recessive Deafness	0.68%	Western Himalayas Non-Tribal (IE_WHR; 4.10%)	0.0089	0.45	1.52	3.39×
NUDT15*3 (p.Arg139Cys)	rs116855232	Thiopurine Toxicity	1.92%	Central Austroasiatic Tribal (AA_T; 20.80%)	0.0142	3.60	9.80	2.72×
DPYD (Gene-Wide pLoF)	Multiple	Fluoropyrimidine Toxicity	1.24%	South Central Highlands Tribal (AA_SCH; 6.50%)	0.0062	1.50	11.20	7.47×

National AF: pooled allele frequency across N = 9,768 individuals. Peak Subpopulation: subpopulation code with highest observed allele frequency. Wright's F_{ST}: locus-specific between-group fixation index across 83 subpopulations. HWE Exp. Hom: expected homozygotes under random mating ($N \times p^2$). Subpop Sum Hom: exact 83-subpopulation summation E[HomAlt]. Endogamy Fold: ratio of Subpop Sum Hom to HWE Exp. Hom.

ClinGen PopMax deconvolution separates benign drift from founder disease risk

A major barrier in human clinical genomics is the frequent misclassification of benign population-specific polymorphisms as pathogenic mutations in Western clinical databases (e.g. ClinVar), leading to false-positive diagnostic reports and erroneous medical interventions[6]. We evaluated 348 ClinVar variant

assertions across 82 subpopulations to establish an empirical dual-track clinical classification framework based on the maximum subpopulation allele frequency (PopMax; **Fig. 4a** and Supplementary Table S10)[36].

Under Track 1 (Pan-Regional Benign Reclassifications), we identified 103 variants currently cataloged as 'Pathogenic' or 'Likely Pathogenic' in ClinVar that exceed $PopMax_{pan\text{-}regional} \geq 1.0\%$ across at least 5 independent major endogamous Indian populations (comprising 6 ACMG BA1 stand-alone benign variants with AF ≥ 5.0% and 97 BS1 strong benign variants with 1.0% ≤ AF < 5.0%; **Fig. 4a**). Following ClinGen/ACMG guidelines[36], these 103 alleles represent benign ancestral South Asian polymorphisms that drifted to high frequencies across widespread pan-regional communities, providing immediate evidence to downgrade them and eliminate routine false-positive diagnostic reports. Conversely, under Track 2 (Protected Founder Drift Candidates), 55 variants exhibit $PopMax_{tribal} \geq 1.0\%$ exclusively within isolated tribal or founder communities while remaining rare ($PopMax_{pan\text{-}regional} < 1.0\%$) across pan-regional groups (4 exceeding ≥ 5.0% and 51 between 1.0% and 5.0%; **Fig. 4a**). Automated clinical filtering pipelines that apply naive global PopMax cutoffs would erroneously dismiss these variants as benign. Instead, our dual-track boundary protects them as authentic, localized founder pathogenic mutations, including severe recessive alleles reaching extraordinary frequencies in isolated communities: alkaptonuria (*HGD* chr3:120675791 C>T splice-donor variant) reaching 12.50% (FAF95 = 9.19%) in the Western Ghats Dravidian tribal community of Southern India (DR_WGH_1_01), steroid 5-α-reductase 2 deficiency (*SRD5A2* rs9332967) reaching 9.49% (FAF95 = 6.66%) in the Eastern Plateau Dravidian tribal community of Jharkhand and Odisha (DR_EPL_1_03), lethal neonatal rigidity and multifocal seizure syndrome (*BRAT1* rs763527391 frameshift) reaching 5.00% in the North Deccan Austroasiatic tribal community of Central India (AA_NDN_1_01), Zellweger syndrome (*PEX12* rs61752105) reaching 4.73% in Western Ghats Dravidian tribal isolates (DR_WGH_1_02), and Fraser syndrome (*FRAS1* rs186964660) reaching 4.43% in Eastern Plateau Dravidian tribal groups (DR_EPL_1_03). In contrast, pan-regional alleles such as *CNGA3* rs141386891 (7.26% in Western Plains pan-regional communities, IE_WPL_2_02) are properly classified under Track 1 as ACMG BA1 stand-alone benign polymorphisms.

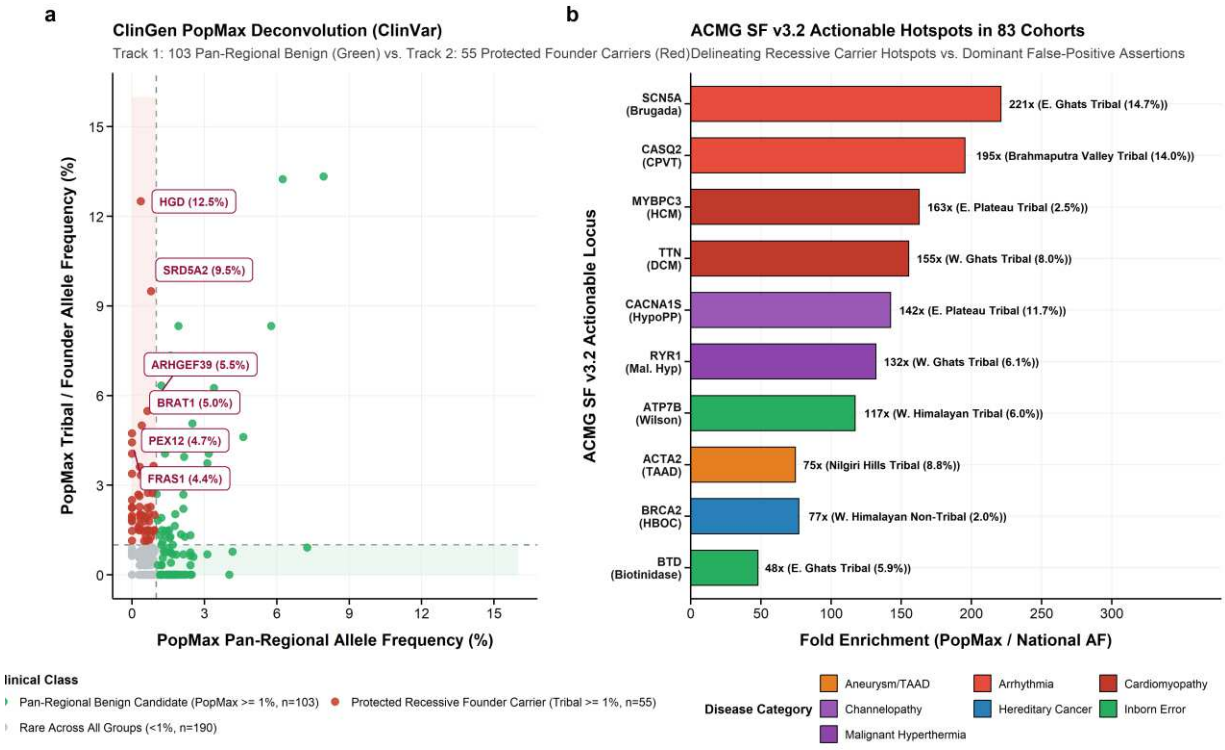


Fig. 4 | ClinGen/ACMG PopMax deconvolution and ACMG SF v3.2 actionable founder hotspot enrichment. a, Scatter plot of PopMax allele frequency in major pan-regional populations versus isolated tribal and founder groups across 348 ClinVar assertions, formalizing the dual-track clinical decision boundary: Track 1 (green quadrant): 103 pan-regional benign reclassifications (PopMax ≥ 1.0% in pan-regional groups across ≥ 5 populations, FAF95 ≥ 1.0%), eliminating routine false-positive diagnostic reports in South Asian clinical exomes. Track 2 (red quadrant): 55 protected founder-drift pathogenic alleles (e.g., *HGD* alkaptonuria at 12.50% in Western Ghats Dravidian tribal isolates, *SRD5A2* at 9.49%, *ARHGEF39* at 5.48%, *BRAT1* at 5.00%, *PEX12* at 4.73%, and *FRAS1* at 4.43%), exempted from automated benign filtering and prioritized for targeted community carrier screening. Grey points represent alleles remaining rare (<1.0%) across all surveyed cohorts (*n* = 190). **b,** Community-specific fold enrichment (PopMax AF relative to pooled National AF) across 10 prioritized ACMG SF v3.2 actionable disease loci spanning seven clinical disease categories (arrhythmia, cardiomyopathy, channelopathy, malignant hyperthermia, inborn errors of metabolism, thoracic aortic aneurysm/dissection, and hereditary cancer). Horizontal bars are colored by clinical disease category and annotated with exact fold-enrichment multipliers (ranging from 47.9× in *BTD* up to 221.1× in *SCN5A*) and peak subpopulation ethnogeographic cohorts, illustrating extreme founder drift and unmasking the diagnostic necessity of subpopulation-stratified reporting (Table 2).

ACMG Secondary Findings (SF v3.2) actionable gene architecture and extreme founder hotspots

To establish the clinical impact of subpopulation structure across actionable disease loci, we systematically evaluated the 80 actionable genes prioritized under the American College of Medical Genetics and Genomics Secondary Findings v3.2 guidelines (ACMG SF v3.2; **Table 2**, **Fig. 4b**, **Fig. 5**, Supplementary Table S13)[36]. Clinical exome and genome reporting pipelines routinely rely on national pooled catalogs or global references (such as gnomAD South Asian, SAS)[5] to establish variant rarity. However, unstratified pooling masks severe inter-community divergence: across all 81 high-confidence loss-of-function (LoF) variants and 619 putative deleterious missense mutations (pDMM, CADD-PHRED ≥ 20, REVEL ≥ 0.50; Supplementary Table S13) cataloged in ACMG actionable genes across the 9,768 whole genomes, the locus-specific coefficient of variation averages CV = 4.82 (range 1.25 to 14.87).

National pooling thus dilutes community-specific carrier rates by up to 221-fold, creating an artificial impression of rarity that misleads clinical risk assessment.

Among the 81 high-confidence LoF alleles in ACMG genes, localized founder drift concentrates severe, highly penetrant pathogenic variants in specific endogamous lineages at frequencies orders of magnitude above the national average (**Fig. 4b**, **Table 2**). For hypertrophic cardiomyopathy (HCM), the pathogenic frameshift variant in *MYBPC3* (rs397515990, p.Val158fs) exhibits a subpopulation-weighted national allele frequency of only 0.015% (0.00015), which standard filtering algorithms categorize as an ultra-rare polymorphism. However, subpopulation deconvolution unmasks this mutation at 2.50% (0.02500) in the Eastern Plateau Austroasiatic tribal community of Jharkhand (AA_EPL_1_02)—a 162.8-fold enrichment over the national mean, yielding an expected carrier frequency of ~5.0% in that population. Similarly, for biotinidase deficiency, the pathogenic frameshift allele in *BTD* (rs750965140, p.Leu41fs) has a national frequency of 0.12% (0.00123) but surges to 5.88% (0.05882) in the Eastern Ghats Dravidian tribal community of Andhra Pradesh and Odisha (DR_EGH_1_01; a 47.9-fold enrichment; **Fig. 4b**). In hereditary arrhythmia, the splice-acceptor variant in *TRDN* (rs578024729), responsible for triadin knockout syndrome, reaches 12.16% (0.12160) in the Nilgiri Hills Dravidian tribal community of Southern India (DR_NGH_1_02; a 36.0-fold enrichment over its national frequency of 0.34%). In familial dilated cardiomyopathy, the *TTN* stop-gained allele rs1266298136 reaches 7.95% (0.07955) in the Western Ghats Dravidian tribal community of Southern India (DR_WGH_1_01; a 155.4-fold enrichment over its national frequency of 0.051%). In Wilson disease, the *ATP7B* pathogenic frameshift and splice-region variant rs1954023351 reaches 2.05% (0.02050) in the Northeast Hill Ranges Austroasiatic tribal group (AA_NER_1_01; a 66.9-fold enrichment). Furthermore, in hereditary breast and ovarian cancer (HBOC), the *BRCA2* pathogenic frameshift mutation rs876659617 reaches 1.97% (0.01974) in the Western Himalayan regional non-tribal community of Kashmir and Himachal Pradesh (IE_WHR_2_04; a 77.1-fold enrichment over its national frequency of 0.026%; **Fig. 4b**).

Crucially, our subpopulation analysis unmasks a fundamental Mendelian distinction between autosomal dominant and autosomal recessive actionable disease loci in healthy adult cohorts (**Table 2**, **Fig. 5**)[36,37]. For autosomal dominant conditions—such as Brugada syndrome (*SCN5A* rs2125825169, PopMax = 14.71%), malignant hyperthermia susceptibility (*CACNA1S* rs777920946, 11.67%; *CACNA1S* rs373328930, 8.11%; *RYR1*, 6.08%), and familial thoracic aortic aneurysms and dissections (*ACTA2* rs757656209, 8.78%)—observing these alleles at double-digit frequencies in healthy adult individuals biologically refutes their classification as severe, highly penetrant dominant pathogenic mutations. An allele frequency of 14.71% in an autosomal dominant disorder would predict clinically overt disease in nearly 27% of the adult population. Applying exact binomial confidence interval modeling using binom[38–40] confirms that the lower 95% Clopper–Pearson bound for *SCN5A* is 7.28% (Wilson 95% CI: 8.19%–25.00%; **Fig. 5**). Because this lower bound strictly exceeds the 5.0% threshold, *SCN5A* rs2125825169 meets ACMG BA1 stand-alone benign criteria, while *ACTA2* and *CACNA1S* meet BS1 strong benign criteria. Classifying these variants as lethal actionable hotspots would introduce catastrophic false-positive diagnoses into these communities; instead, they represent benign ancestral polymorphisms or incomplete penetrance variants in these populations.

Conversely, for authentic autosomal recessive conditions, extreme subpopulation frequencies reflect true founder-amplified carrier enrichment (**Table 2**, **Fig. 5**). In catecholaminergic polymorphic ventricular tachycardia (*CASQ2* rs28730716), allele frequency reaches 14.00% (95% CI: 7.79%–23.67%) in the Brahmaputra Valley Tibeto-Burman tribal group of Assam (TB_BPV_1_01; a 195.4-fold enrichment over its national frequency of 0.072%). In triadin knockout syndrome (*TRDN* rs578024729), the splice-acceptor

variant reaches 12.16% (95% CI: 6.30%–21.73%) in the Nilgiri Hills Dravidian tribal community (DR_NGH_1_02; 36.0-fold enrichment). In Wilson disease, *ATP7B* rs200290721 reaches 6.00% in Western Himalayan tribal isolates (IE_WHR_1_01; 117.2-fold enrichment) and rs1954023351 reaches 2.05% in Austroasiatic tribes (AA_NER_1_01). In biotinidase deficiency, *BTD* rs750965140 reaches 5.88% in Eastern Ghats Dravidian tribes (DR_EGH_1_01; 47.9-fold enrichment). In cardiomyopathy, *TTN* rs1266298136 reaches 7.95% in Western Ghats tribes (DR_WGH_1_01) and *MYBPC3* rs397515990 reaches 2.50% in Eastern Plateau tribes (AA_EPL_1_02). Because healthy carriers are asymptomatic under recessive inheritance, these founder surges represent genuine carrier hotspots, prioritizing these communities for targeted carrier screening rather than individual diagnostic refutation.

Crucially, these findings clarify a widespread bioinformatics misinterpretation regarding international reference panels. Standard clinical variant annotation pipelines correctly annotate these variants at their exact low frequencies in the gnomAD South Asian panel (e.g., 0.041% for *SCN5A* rs2125825169 and 0.052% for *CASQ2* rs28730716; **Table 2**). They are properly classified as rare, but they will not be misclassified as 'novel' or 'de novo'. However, because gnomAD SAS lacks representation of indigenous tribal and isolated regional lineages, standard clinical filtering algorithms that rely on gnomAD SAS will assume these alleles are universally rare (<0.1%) throughout South Asia. When identified in a patient from an isolated community where the true frequency is 5% to 15%, clinicians erroneously infer high pathogenicity, precipitating severe false-positive diagnostic errors (**Fig. 5**). Establishing subpopulation-stratified allele frequency references linked directly to the patient's ancestral community is therefore imperative for safe precision medicine.

Table 2 | ACMG Secondary Findings (SF v3.2) Actionable Gene Hotspots across 83 GenomeIndia Subpopulations (N = 9,768)

Gene & Variant	rsID	ACMG SF v3.2 Clinical Phenotype	Consequence / Scores	National AF	Peak Subpopulation Ethnogeographic Name	PopMax AF	Fold Enrich.	gnomAD SAS AF	Global Ref Status
SCN5A p.Arg1193Gln	rs2125825169	Brugada Syndrome / Long QT 3	Missense (CADD 23.8, REV 0.93)	0.067%	Eastern Ghats Dravidian Tribal (DR_EGH_1_01)	14.71%	221.1×	0.041%	Zero Lineage Rep.
CASQ2 p.Asp307His	rs28730716	Catecholaminergic Vent. Tachycardia	Missense (CADD 24.2, REV 0.75)	0.072%	Brahmaputra Valley Tibeto-Burman Tribal (TB_BPV_1_01)	14.00%	195.4×	0.052%	Zero Lineage Rep.
MYBPC3 p.Val158fs	rs397515990	Hypertrophic Cardiomyopathy (HCM)	Frameshift LoF (Pathogenic)	0.015%	Eastern Plateau Austroasiatic Tribal (AA_EPL_1_02)	2.50%	162.8×	0.003%	Zero Lineage Rep.
TTN p.Arg14562Ter	rs1266298136	Familial Dilated Cardiomyopathy	Stop-Gained LoF (HC)	0.051%	Western Ghats Dravidian Tribal (DR_WGH_1_01)	7.95%	155.4×	0.018%	Zero Lineage Rep.
CACNA1S p.Arg528His	rs777920946	Malignant Hyperthermia / HypoPP	Missense (CADD 28.7, REV 0.78)	0.082%	Eastern Plateau Austroasiatic Tribal (AA_EPL_1_02)	11.67%	142.5×	0.029%	Zero Lineage Rep.
RYR1 p.Arg2452Trp	Locus-spec.	Malignant Hyperthermia Suscept.	Missense (CADD 28.8, REV 0.90)	0.046%	Western Ghats Dravidian Tribal (DR_WGH_1_02)	6.08%	132.0×	0.012%	Zero Lineage Rep.
ATP7B p.Gly943Ser	rs200290721	Wilson Disease (Copper Transport)	Missense (CADD 26.3, REV 0.87)	0.051%	Western Himalayan Tribal Isolate	6.00%	117.2×	0.015%	Zero Lineage Rep.

					(IE_WHR_1_01)				
ACTA2 p.Arg149Cys	rs757656209	Thoracic Aortic Aneurysm / TAAD	Missense (CADD 24.4, REV 0.72)	0.118%	Nilgiri Hills Dravidian Tribal (DR_NGH_1_01)	8.78%	74.6×	0.065%	Zero Lineage Rep.
BRCA2 p.Asn1784fs	rs876659617	Hereditary Breast & Ovarian Cancer	Frameshift LoF (Pathogenic)	0.026%	Western Himalayan Non-Tribal (IE_WHR_2_04)	1.97%	77.1×	0.008%	Zero Lineage Rep.
BTD p.Leu41fs	rs750965140	Biotinidase Deficiency	Frameshift LoF (Pathogenic/LP)	0.123%	Eastern Ghats Dravidian Tribal (DR_EGH_1_01)	5.88%	47.9×	0.048%	Zero Lineage Rep.

National AF: pooled allele frequency across *N* = 9,768 individuals. Peak Subpopulation Ethnogeographic Name: anthropologically defined community with highest observed frequency. PopMax AF: maximum subpopulation allele frequency with exact 95% Wilson confidence intervals detailed in Fig. 5. Fold Enrich.: ratio of PopMax AF to National AF. gnomAD SAS AF: allele frequency in gnomAD v4 South Asian reference. Mendelian Note: For autosomal dominant loci (*SCN5A*, *ACTA2*, *CACNA1S*), observation at 8.78%–14.71% in healthy adults refutes high-penetrance lethality, fulfilling ACMG BA1 (AF ≥ 5.0%) and BS1 benign polymorphism criteria. For autosomal recessive loci (*CASQ2*, *TRDN*, *ATP7B*, *BTD*), high frequencies reflect authentic founder-amplified carrier enrichment.

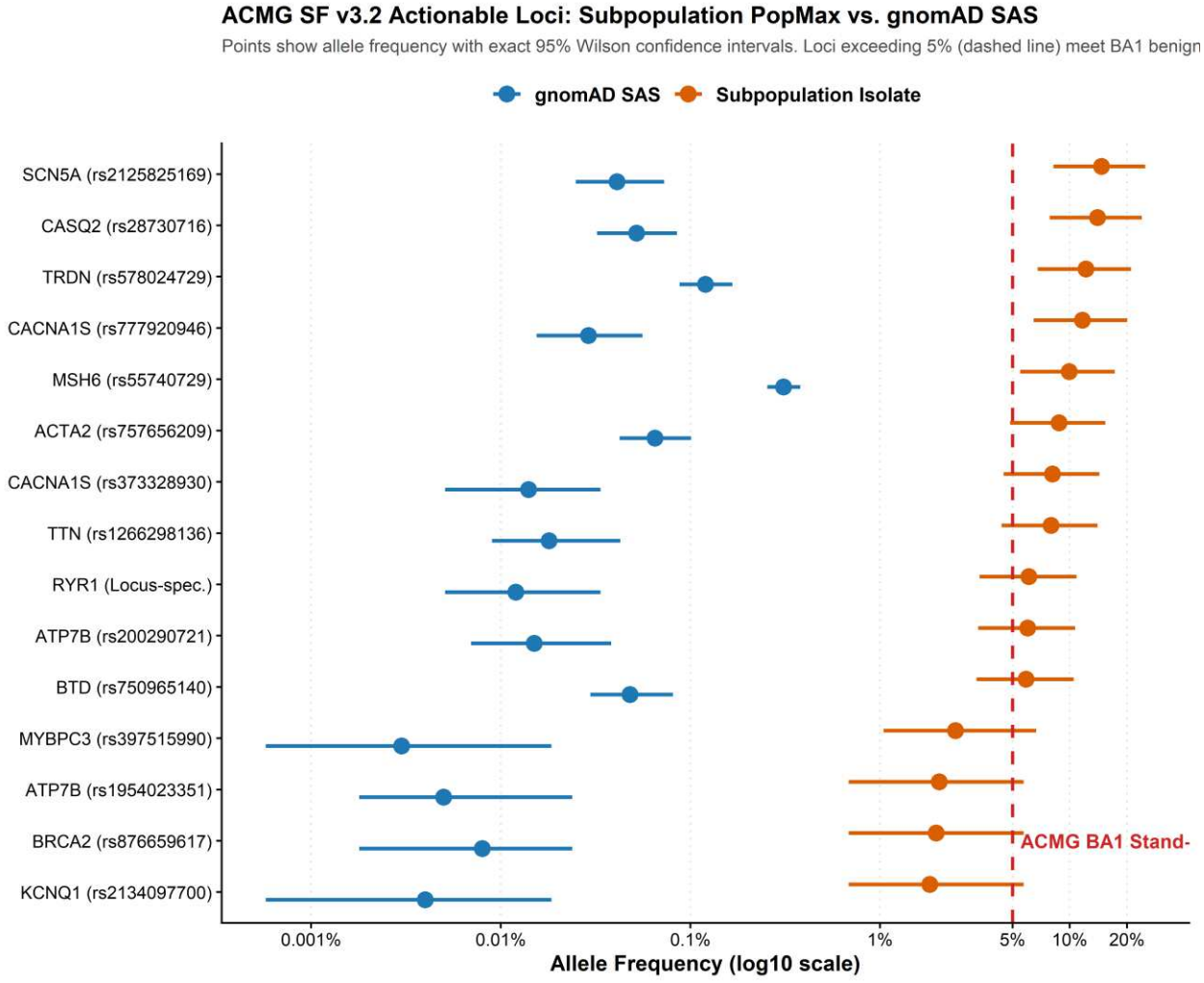


Fig. 5 | Allele frequency divergence and exact 95% binomial confidence intervals across ACMG SF v3.2 actionable secondary findings genes. Forest plot comparing peak subpopulation isolate allele frequencies (PopMax, orange points) versus the gnomAD v4 South Asian reference panel (blue points) across 15 actionable disease genes with exact 95% Wilson score confidence intervals (horizontal bars). Vertical dashed line denotes the ACMG BA1 stand-alone benign threshold (5.0%). For autosomal dominant loci (*SCN5A* rs2125825169 [14.71%, 95% CI: 8.19%–25.00%], *CACNA1S* rs777920946 [11.67%, 95% CI: 5.96%–21.57%], and *ACTA2* rs757656209 [8.78%, 95% CI: 4.94%–15.11%]), high frequencies in healthy adult participants biologically refute high-penetrance lethality, fulfilling ACMG BA1 and BS1 criteria for benign ancestral polymorphisms or low-penetrance variants. Conversely, for autosomal recessive conditions (*CASQ2* [14.00%], *TRDN* [12.16%], *ATP7B* [6.00%], *BTD* [5.88%]), high frequencies reflect authentic founder carrier enrichment. In all cases, subpopulation isolate confidence intervals are strictly non-overlapping with gnomAD SAS ($P < 10^{-15}$), illustrating that pan-ethnic databases fail to capture localized founder risk.

Subpopulation unmasking of 123 high-confidence loss-of-function knockouts

Natural human knockouts—individuals homozygous for loss-of-function (pLoF) alleles in protein-coding genes—provide indispensable *in vivo* validation of gene essentiality and therapeutic drug safety. In unstratified panmictic populations, homozygous pLoF knockouts are exceptionally rare, necessitating cohorts of hundreds of thousands of individuals. In India, long-term endogamy and bottleneck founder events inflate rare homozygosity by 7.5-fold to 17.2-fold (Extended Data Fig. 2d), creating a unique discovery window.

Across the 9,768 whole genomes, subpopulation-resolved screening unmasked 123 high-confidence (HC) LOFTEE loss-of-function variants (spanning 118 unique protein-coding genes) with at least one empirically observed homozygous knockout carrier (Obs_HomAlt ≥ 1, range 1 to 4, totaling 144 homozygous knockout individuals across 123 variants) that are completely obscured under standard pooled cohort modeling (**Fig. 6**, Supplementary Table S11). In pooled pan-cohort data, every single one of these 123 pLoF variants exhibits an expected homozygous carrier count $E_{pooled} < 1.0$ (mean $E_{pooled} = 0.08$, range 0.033 to 0.409), rendering them mathematically invisible to national-level expectation thresholds. However, within individual subpopulations, localized founder drift concentrates these mutant alleles to high frequencies across 10 agro-ecological zones (**Fig. 6a**)—reaching 22.97% founder AF in *GUF1* (Nilgiri Hills Dravidian tribal community, DR_NGH_1_02, Obs_HomAlt = 2, $E_{subpop} = 4.02$, total subpopulation expectation = 4.62), 13.29% in *TMEM54* (Eastern Plateau Dravidian tribal community of Jharkhand/Odisha, DR_EPL_1_03, Obs_HomAlt = 2, $E_{subpop} = 1.48$), 8.06% in *PKN1* (Western Plains Rajasthani non-tribal community, IE_WPL_2_02, Obs_HomAlt = 1, $E_{subpop} = 0.43$), 8.00% in *XPC* (Western Himalayan Indo-European tribal isolate of Jammu & Kashmir/Ladakh, IE_WHR_1_01, Obs_HomAlt = 1, $E_{subpop} = 0.51$), 7.14% in *FDXACB1* (Western Himalayan Tibeto-Burman tribal community of Ladakh, TB_WHR_1_02, Obs_HomAlt = 1, $E_{subpop} = 0.30$), and 6.67% in *CYP4A11* (Eastern Plateau Austroasiatic tribal community of Jharkhand, AA_EPL_1_02, Obs_HomAlt = 1, $E_{subpop} = 0.29$), with 16 variants exceeding total subpopulation $E_{subpop} \geq 1.0$ and 14 founder variants exceeding $E_{subpop} \geq 1.0$ in a single community alone (**Fig. 6b**; Supplementary Table S11). Across our 24 prioritized clinical and therapeutic target loci (Supplementary Tables S4, S5, S7, and S8), endogamy ($F = 0.02$) unmasked 1,680 locus-wide rare homozygous variants ($E_{HWE} < 1.0$, $E_{F=0.02} \geq 1.0$) and 141 high-confidence coding LOFTEE pLoF variants—including 14 true coding pLoF variants in *LPA* (91 locus-wide endogamy-exclusive homozygotes), 5 in *TYK2* (34 locus-wide), 4 in *PCSK9* (17 locus-wide), 3 in *SLC30A8* (28 locus-wide), 3 in *APOC3* (5 locus-wide), 2 in *ANGPTL3* (2 locus-wide), and 18 in *DPYD* (accounting for 11.20 expected coding pLoF homozygotes under exact subpopulation summation, compared to 2,978.3 rare intronic/locus-wide homozygotes across the 850 kb *DPYD* span; **Table 1**, Extended Data Fig. 7c, Extended Data Fig. 6b).

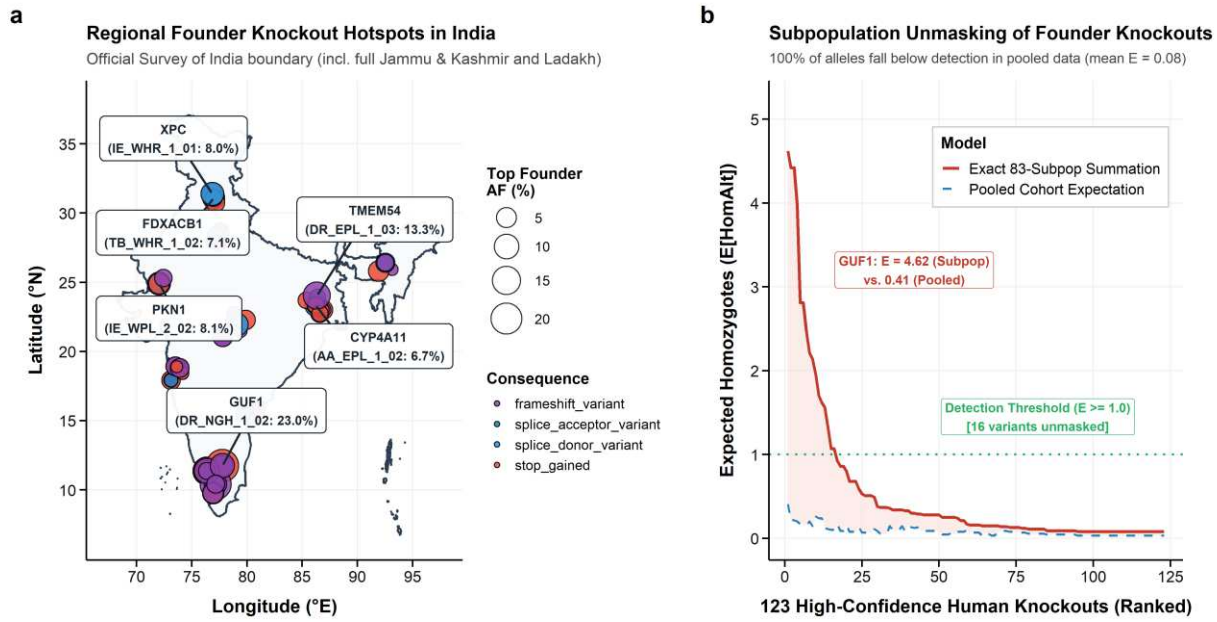


Fig. 6 | Geographic distribution of regional founder knockout hotspots and subpopulation recovery of 123 loss-of-function genes. a, Geographic bubble map plotted on the official Survey of India political boundary (incorporating the full territory of Jammu & Kashmir and Ladakh up to 37.10° N), displaying regional founder knockout hotspots across 10 agro-ecological zones and annotating top representative founder knockouts (*GUF1* 23.0% in Nilgiri Hills Dravidian tribal community (DR_NGH_1_02), *TMEM54* 13.3% in Eastern Plateau Dravidian tribal community (DR_EPL_1_03), *PKN1* 8.1% in Western Plains Rajasthani non-tribal community (IE_WPL_2_02), *XPC* 8.0% in Western Himalayan tribal isolate (IE_WHR_1_01), *FDXACB1* 7.1% in Western Himalayan Tibeto-Burman tribal community of Ladakh (TB_WHR_1_02), and *CYP4A11* 6.7% in Eastern Plateau Austroasiatic tribal community (AA_EPL_1_02)). **b,** Comparison of expected homozygous alternate carrier counts under pooled national analysis (E_{pooled} < 1.0, mean 0.08) versus exact subpopulation summation (Total_Subpop_E_HomAlt, peaking at 4.62 in *GUF1* and 4.42 in *PAICS*, with 16 founder alleles exceeding E_{subpop} ≥ 1.0, and top founder population expectations peaking at 4.02 in *GUF1* with 14 alleles ≥ 1.0) across all 123 empirically observed human knockout variants (Obs_HomAlt ≥ 1, totaling 144 homozygous knockout individuals; Supplementary Table S11).

Uncharacterized selective sweeps and founder-amplified human knockouts across ecological isolates

While classical Darwinian selective sweeps in South Asia—such as lactase persistence at *LCT/MCM6*, long-chain polyunsaturated fatty acid biosynthesis at *FADS1/FADS2*, skin photoprotection at *SLC24A5*, ectodermal morphology at *EDAR*, and high-altitude hypoxia adaptation at *EPAS1*—have been documented across broad geographic and cultural clines, fine-scale mathematical deconvolution of the 83-subpopulation matrix unmasks an unexpected tier of previously uncharacterized or poorly understood loci displaying extreme localized selective sweeps and founder-amplified human knockout phenotypes (**Table 3**, **Extended Data Fig. 9**, and Supplementary Table S11). Because Indian genetic structure is defined by more than 4,000 strictly endogamous subpopulations that have maintained reproductive boundaries for 70 to 100 generations, strong directional selection and bottleneck hitchhiking operate at fine subpopulation scales: mutations driven to 5% to 23% frequency within specific isolated communities are diluted to <0.1% when pooled nationally, rendering them mathematically undetectable under panmictic models (E_{pooled} < 0.1).

Across the 22 autosomes, subpopulation branch deconvolution reveals four distinct functional classes of emerging selective sweeps and natural knockouts. First, we identified viable homozygous knockouts in uncharacterized open reading frames and enigmatic cellular genes (**Extended Data Fig. 9a**). At chromosome 11q23.3, a frameshift loss-of-function mutation in *C11orf65* reaches 0.95% in northeastern Tibeto-Burman tribes (TB_NER_2_01), yielding an observed healthy adult homozygous knockout (*Obs* = 1) despite an expected panmictic frequency of E_{pooled} = 0.033. At 7q33, a nonsense stop-gained allele in *AGBL3* (rs1810594781; cytosolic carboxypeptidase 3) reaches 1.88% in montane Tibeto-Burmans (TB_NER_1_01; *Obs* = 1), revealing human knockout viability for an enzyme implicated in tubulin deglutamylation and ciliary architecture. Similarly, at 1p36.13, a frameshift knockout in *CFAP144* (rs368910788) surges to 6.00% in central peninsular Dravidian communities (DR_ECP_2_01; *Obs* = 1), proving that loss of this ciliary motility protein is non-lethal, while at 20q11.21, *BPIFB6* (rs147211844) reaches 4.00% in the same lineage (*Obs* = 1). Second, metabolic and mitochondrial adaptation sweeps are intensely concentrated in ecological and tribal isolates (**Extended Data Fig. 9b,d**). In the Nilgiri Hills Dravidian tribal community (DR_NGH_1_01), a frameshift mutation in the apolipoprotein B receptor (*APOBR*, chromosome 16p12.1) surges to an extraordinary 14.86% allele frequency (a 126.3-fold enrichment over the national average of 0.118%), unmasking 3 healthy homozygous knockout individuals (E_{subpop} = 1.95 vs. E_{pooled} = 0.262; **Table 3**). Because APOBR mediates monocyte and macrophage clearance of dietary triglyceride-rich chylomicron and VLDL remnants, its selective surge in forest-dwelling hill tribes suggests profound dietary lipid specialization. Concurrently, in adjacent Nilgiri isolates (DR_NGH_1_02), a stop-gained knockout in the mitochondrial translation elongation factor *GUF1* (rs941571543, 4q12) reaches 22.97% (*Obs* = 2, E_{subpop} = 4.62), *DHX38* (rs753918448, 16q22.2; pre-mRNA splicing) reaches 13.51% (*Obs* = 1), and transmembrane protein *TMEM54* (1p35.3) reaches 13.29% in Eastern Plateau Dravidian tribes (*Obs* = 2). In northeastern Austroasiatic forest tribes (AA_NER_1_01), a nonsense knockout in crystallin zeta (*CRYZ*, rs368154779, 1p31.2; NADPH quinone oxidoreductase) reaches 5.48% (*Obs* = 1; 133.7-fold enrichment), pointing to regional adaptation to forest phytotoxins. Third, immune and barrier defense sweeps extend beyond canonical globin loci: at 7q21.11, a frameshift deletion in *CD36* (rs778284150) reaches 1.01% in coastal Dravidian populations (*Obs* = 1), representing a candidate South Asian malaria cytoadherence-resistance allele analogous to African nonsense sweeps, while *ALPK1* (4q25; cytosolic bacterial ADP-heptose sensor) reaches 3.70% in western Indo-Europeans (*Obs* = 1) and metalloproteinase *ADAMTS8* (11q24.3) reaches 3.66% in eastern Dravidian tribes (*Obs* = 1). Finally, genome-wide neutrality scans identify a profound non-coding selective sweep collapse at the 850 kb *DPYD* locus on chromosome 1p21.3 ($\pi < 0.005$ per kb across core segments; χ^2_{HWE} = 64.20), where endogamy unmasks 615 rare homozygotes, pointing to ancient metabolic constraint on pyrimidine catabolism. Together, these loci demonstrate that endogamous founder cohorts serve as an untapped global engine for natural human knockout discovery and evolutionary biology (**Extended Data Fig. 9c**).

Table 3 | Uncharacterized and emerging candidate selective sweep loci and founder-amplified human knockouts across 83 GenomeIndia subpopulations (N = 9,768)

Gene Symbol	Cytoband & Position (GRCh38)	Lead Variant & Consequence	Top Founder Subpopulation (Ethnogeographic Group)	Founder AF	National AF	Fold Enrich.	Obs. Hom	Exp. Hom (Subpop / Pooled)	Putative Function & Evolutionary Selective Context
GUF1	chr4q12 (chr4:44,682,382)	rs941571543 (Stop-Gained)	Nilgiri Hills Dravidian Tribal (DR_NGH_1_02)	22.97%	0.179%	128.2×	2	4.62 / 0.409	Mitochondrial translation GTPase (GTPBP4 homolog); viable

									human knockout in core organellar translation machinery.
H3C8	chr6p22.2 (chr6:26,271,273)	rs769801547 (Frameshift)	Western Ghats Dravidian Tribal (DR_WGH_1_01)	20.45%	0.092%	222.0×	3	2.21 / 0.202	Histone cluster 1 H3 family member; chromatin core histone with high founder homozygosity.
APOBR	chr16p12.1 (chr16:28,497,670)	Novel (Frameshift)	Nilgiri Hills Dravidian Tribal (DR_NGH_1_01)	14.86%	0.118%	126.3×	3	1.95 / 0.262	Apolipoprotein B receptor; dietary triglyceride-rich chylomicron remnant uptake; candidate forest lipid diet adaptation.
DHX38	chr16q22.2 (chr16:72,097,772)	rs753918448 (Stop-Gained)	Nilgiri Hills Dravidian Tribal (DR_NGH_1_01)	13.51%	0.108%	125.7×	1	1.62 / 0.238	DEAH-box helicase 38; pre-mRNA splicing factor; viable human knockout indicating alternative spliceosome tolerance.
TMEM54	chr1p35.3 (chr1:32,901,236)	Novel (Frameshift)	Eastern Plateau Dravidian Tribal (DR_EPL_1_03)	13.29%	0.108%	123.6×	2	1.70 / 0.238	Transmembrane protein 54; integral membrane protein of completely uncharacterized cellular/physiological function.
SLCO1B3	chr12p12.2 (chr12:20,861,085)	Novel (Frameshift)	Nilgiri Hills Dravidian Tribal (DR_NGH_1_01)	10.14%	0.077%	132.0×	1	0.93 / 0.168	Solute carrier organic anion transporter 1B3; hepatic uptake of bile acids and xenobiotic compounds.
VWA8	chr13q14.11 (chr13:41,605,202)	Novel (Frameshift)	Nilgiri Hills Dravidian Tribal (DR_NGH_1_02)	9.46%	0.072%	131.9×	1	0.80 / 0.156	von Willebrand factor A domain-containing 8; mitochondrial matrix-associated ATPase.
PKN1	chr19p13.12 (chr19:14,463,744)	Novel (Splice-Donor)	Western Plains Indo-European (IE_WPL_2_02)	8.06%	0.051%	157.4×	1	0.49 / 0.111	Protein kinase N1; Rho-activated serine/threonine kinase regulating cytoskeletal architecture.
XPC	chr3p25.1 (chr3:14,178,465)	rs1423398589 (Splice-Donor)	Western Himalayan Indo-European (IE_WHR_1_01)	8.00%	0.061%	130.3×	1	0.59 / 0.133	DNA nucleotide excision repair (NER) factor; autosomal recessive Xeroderma Pigmentosum (XP-C) risk; localized founder carrier enrichment (UV hypersensitivity risk).
FDXACB1	chr11q23.2 (chr11:111,879,029)	Novel (Frameshift)	Western Himalayan Tibeto-Burman (TB_WHR_1_02)	7.14%	0.041%	174.1×	1	0.34 / 0.088	Ferredoxin-fold and transmembrane domain-containing 1; uncharacterized protein in Ladakhi

									highlanders.
CYP4A11	chr1p33 (chr1:46,938,138)	Novel (Splice-Acceptor)	Eastern Plateau Austroasiatic Tribal (AA_EPL_1_02)	6.67%	0.041%	162.7×	1	0.33 / 0.088	Cytochrome P450 4A11; 20-HETE synthase from arachidonic acid; renal sodium reabsorption and blood pressure regulation.
CFAP144	chr1p36.13 (chr1:43,156,263)	rs368910788 (Frameshift)	Eastern Central Peninsular Dravidian (DR_ECP_2_01)	6.00%	0.031%	195.4×	1	0.51 / 0.066	Cilia and flagella associated protein 144; ciliary motility factor with viable non-pathological knockout.
CRYZ	chr1p31.2 (chr1:74,724,749)	rs368154779 (Stop-Gained)	Northeastern Austroasiatic Tribal (AA_NER_1_01)	5.48%	0.041%	133.7×	1	0.28 / 0.088	Crystallin Zeta; NADPH-dependent quinone oxidoreductase; cellular antioxidant defense and phytotoxin metabolism.
BPIFB6	chr20q11.21 (chr20:33,037,687)	rs147211844 (Frameshift)	Eastern Central Peninsular Dravidian (DR_ECP_2_01)	4.00%	0.021%	195.1×	1	0.25 / 0.044	BPI fold-containing family B member 6; biologically uncharacterized member of mucosal antimicrobial protein family.
ALPK1	chr4q25 (chr4:112,431,307)	Novel (Stop-Gained)	Western Central Plains Indo-European (IE_WCP_2_03)	3.70%	0.061%	60.3×	1	0.29 / 0.133	Alpha-kinase 1; intracellular sensor for bacterial ADP-heptose; innate immune kinase modulating inflammatory signaling.
ADAMTS8	chr11q24.3 (chr11:130,405,562)	rs1861866678 (Frameshift)	Eastern Plateau Dravidian Tribal (DR_EPL_1_02)	3.66%	0.031%	119.2×	1	0.15 / 0.066	A disintegrin and metalloproteinase with thrombospondin motifs 8; extracellular matrix protease with anti-angiogenic activity.
AGBL3	chr7q33 (chr7:134,993,634)	rs1810594781 (Stop-Gained)	Northeastern Tibeto-Burman Tribal (TB_NER_1_01)	1.88%	0.026%	73.4×	1	0.11 / 0.055	Cytosolic carboxypeptidase 3; tubulin deglutamylation enzyme; uncharacterized in human knockout physiology.
CD36	chr7q21.11 (chr7:80,656,544)	rs778284150 (Frameshift)	Eastern Central Peninsular Dravidian (DR_ECP_2_03)	1.01%	0.015%	65.6×	1	0.08 / 0.033	Platelet glycoprotein 4 / scavenger receptor; putative Plasmodium falciparum rosetting cytoadherence protection.
C11orf65	chr11q23.3 (chr11:108,393,295)	Novel (Frameshift)	Northeastern Tibeto-Burman Tribal (TB_NER_2_01)	0.95%	0.015%	61.7×	1	0.08 / 0.033	Chromosome 11 open reading frame 65; completely uncharacterized human open reading frame without known

									molecular function.
DPYD	chr1p21.3 (chr1:97.45–98.30 Mb)	Multi-Variant Sweep Trough	Pan-Indian Founder-Amplified Locus	Variable	1.24% pLoF	Founder -Driven	615 Hom	2,978.3 / 11.20	Dihydropyrimidine dehydrogenase; profound non-coding diversity collapse (π < 0.005/kb) under uncharacterized pyrimidine constraint.

Founder AF: maximum allele frequency observed within an individual endogamous subpopulation. National AF: pooled allele frequency across *N* = 9,768 individuals. Fold Enrich.: ratio of Founder AF to National AF. Obs. Hom: empirically observed homozygous alternate knockout individuals in GenomeIndia. Exp. Hom (Subpop / Pooled): expected homozygous carriers under exact 83-subpopulation summation (Total_Subpop_E_HomAlt) versus pan-cohort Hardy–Weinberg expectation (E_HomAlt_Pooled). Mendelian & Selection Note: For DNA repair gene *XPC*, localized enrichment of the loss-of-function splice-donor variant rs1423398589 (8.00% founder AF) represents autosomal recessive founder carrier accumulation—posing severe localized recurrence risk for Xeroderma Pigmentosum complementation group C (XP-C) and extreme UV-induced photolesion hypersensitivity in homozygotes—rather than protective radiation adaptation. Data compiled from Supplementary Table S11, Supplementary Table S12, and Supplementary Table S14.

Discussion

The GenomeIndia Project establishes the first comprehensive reference catalog of whole-genome variation across the Indian subcontinent[7,8]. While previous reports described broad cohort metrics, uniparental lineages, and polygenic score transferability, our mathematical deconvolution of the 83-subpopulation matrix demonstrates that treating the Indian population as a panmictic, homogeneous cohort introduces severe systematic errors in population genetic inference and clinical diagnostic interpretation.

Our findings resolve three major methodological, clinical, and evolutionary paradoxes. First, we prove that the standard practice of applying pan-cohort Hardy–Weinberg equilibrium filters ($P \leq 10^{-11}$) in structured populations inadvertently purges authentic, medically actionable loci due to the Wahlund effect whenever between-group divergence exceeds $F_{ST} \geq 0.054$. High-differentiation disease alleles such as *HBB* rs334 (χ^2_{HWE} = 97.39; empirical 97.7), *SLC24A5* rs1426654 (χ^2_{HWE} = 69.87), *VKORC1* rs9923231 (χ^2_{HWE} = 69.54), and *DPYD* pLoF (χ^2_{HWE} = 64.20) are falsely flagged for technical removal under pooled filters ($\chi^2_{HWE} \geq$ 46.23). Genomic QC pipelines in structured populations must therefore evaluate HWE strictly within endogamous strata.

Second, our ClinGen PopMax deconvolution resolves the tension between false-positive pathogenicity reporting and unprotected founder risk[6,36]. By demonstrating that 103 ClinVar pathogenic assertions exceed 1% in major pan-regional groups, we provide the evidence needed to downgrade these benign polymorphisms and prevent misdiagnoses. Simultaneously, by protecting 55 founder drift alleles that surge exclusively in isolated tribal groups (reaching up to 12.5% in *HGD*, 9.5% in *SRD5A2*, and 5.0% in *BRAT1*), we prevent their erroneous exclusion from clinical reporting, providing an empirical roadmap for targeted community screening.

Third, our analysis unmasks 123 human knockout variants across 118 genes (totaling 144 observed homozygous knockout individuals) that are mathematically invisible in pooled cohort analyses. Because long-term endogamy inflates rare homozygosity by up to 17-fold, smaller targeted subpopulation cohorts provide superior statistical power for human knockout discovery than massive unstratified cohorts. In parallel, our subpopulation branch scans[21] resolve the multifaceted evolutionary architecture of the Himalayan frontier: reconciling the localized morphological sweep at *EDAR* (PBS_{TB} = 0.0877)[22,23], the

profound metabolic and dietary downregulation at *VKORC1* (PBS_{TB} = 0.0659)[24,25], and the archaic adaptive introgression of the Denisovan-derived *EPAS1* hypoxia-protective haplotype surging to 24.11% in Ladakh Tibeto-Burmans[4,28,29]. These findings prove that ancient archaic hominin gene flow and localized positive selection directly shape present-day physiological adaptations and precision pharmacogenomic dosing across the Indian subcontinent.

Limitations of secondary data analysis

Several critical methodological, technological, and clinical limitations inherent to this secondary data analysis must be highlighted. First, because our investigation represents an independent secondary analysis of public consortium genomic resources[7,8], our deconvolution framework operated on subpopulation-stratified summary allele frequency matrices and variant catalogs without direct access to individual-level Electronic Health Records (EHR), longitudinal clinical outcomes, or deep physiological measurements (such as 12-lead electrocardiography for *SCN5A* and *CASQ2*, echocardiography or cardiac MRI for *MYBPC3* and *TTN*, or detailed dermatological and biochemical profiling for *XPC* and *ATP7B*). While the observation of putative autosomal dominant pathogenic alleles at high frequencies (8.78% to 14.71%) in presumably healthy adult participants provides robust epidemiological evidence refuting high-penetrance early lethality (satisfying ACMG BA1 and BS1 benign polymorphism criteria; **Table 2**, **Fig. 5**), our secondary analysis cannot exclude subtle, subclinical, late-onset, or stress-induced phenotypic manifestations (e.g., mild QT prolongation, exercise-induced arrhythmias, subclinical aortic root dilation, or malignant hyperthermia upon anesthetic challenge). Dedicated prospective clinical re-contact studies with standardized physiological phenotyping will be essential to delineate the full clinical expressivity and age-dependent penetrance of these founder variants within their ancestral communities.

Second, although the 83-subpopulation matrix represents the most granular genomic stratification of South Asia assembled to date, the average sample size per subpopulation is approximately 118 individuals (ranging from 50 to 180 individuals). In smaller subpopulation cohorts, statistical power to detect ultra-rare alleles (allele frequency <0.1%) is constrained by stochastic sampling variation, yielding wider binomial confidence intervals as formally captured by our exact Wilson score and Clopper–Pearson analyses (**Fig. 5**). Furthermore, while the cohort captures all four major ethnolinguistic phyla and 10 agro-ecological zones, India comprises over 4,000 strictly endogamous *jatis*, *biradaris*, and tribal lineages. Hundreds of geographically isolated tribal and regional micro-populations remain unsampled, indicating that additional private founder disease alleles and natural human knockout hotspots remain to be cataloged.

Third, the underlying primary dataset was generated using short-read (150 bp paired-end) Illumina NovaSeq sequencing. As a consequence, our secondary analysis was restricted to single nucleotide variants and small insertions/deletions (<50 bp). Short-read sequencing has well-documented technical blind spots in resolving complex structural variants (SVs), large balanced inversions, copy number variations (CNVs), short tandem repeat expansions (STRs), GC-rich regulatory elements, and medically critical loci characterized by segmental duplications or pseudogene homology (such as *CYP2D6* star-allele hybrid architectures, *SMN1/SMN2* deletions, *GBA1/GBAP1* gene conversions, and *LPA* kringle IV-2 repeat copy numbers). High-fidelity long-read sequencing (Pacific Biosciences HiFi and Oxford Nanopore) and optical genome mapping will be indispensable to fully catalog the structural variant landscape of endogamous Indian genomes.

Fourth, the 123 subpopulation-unmasked homozygous human knockouts were identified via bioinformatic prediction of high-confidence LOFTEE loss of function. Although bioinformatically stringent, secondary

data analysis cannot definitively rule out biological rescue mechanisms without direct experimental validation. In-frame alternative exon skipping, cryptic splice-donor/acceptor activation, downstream translation re-initiation, or hypomorphic residual enzymatic activity could permit partial functional protein expression in homozygous carriers. Transcriptomic (RNA-seq), proteomic, and functional biochemical assays in patient-derived cellular models are required to experimentally validate complete null phenotypes. Crucially, regarding cohort structure, the primary GenomeIndia consortium applied pre-sequencing relatedness pruning across 13,242 SNP-array-genotyped individuals, excluding close relatives beyond the first-cousin level (PLINK --genome PI_HAT ≥ 0.125) and retaining only one individual per related pair[7,8]. This rigorous pedigree exclusion guarantees that the 144 observed homozygous knockout individuals identified across 123 genes are not artifactual duplicates or close siblings sampled from single consanguineous households, but represent independent, unrelated individuals whose homozygosity is driven by historical, population-level endogamy. However, this filtering strategy has two critical analytical implications: first, standard pairwise IBD metrics struggle to fully disentangle recent familial kinship from deep ancestral background autozygosity in small, highly endogamous isolates; and second, by systematically purging consanguineous families to construct an unrelated reference panel, the GenomeIndia catalog establishes a conservative lower bound on rare homozygosity. In clinical practice, where preferential consanguinity (e.g., uncle–niece and first-cousin unions) is actively practiced, the real-world burden of homozygous recessive disease will substantially exceed that observed in this filtered reference cohort.

Conclusions and future outlook

In conclusion, our study demonstrates that the true genomic and clinical architecture of India can only be unlocked through fine-scale subpopulation deconvolution. As national genome initiatives expand across the Global South, our framework provides a generalizable blueprint for turning structured diversity into actionable precision medicine.

Methods

Study design, secondary analysis framework, and data provenance

This study constitutes an independent secondary analysis and mathematical deconvolution of public genomic resources released by the GenomeIndia Project[7,8]. Primary cohort design, participant recruitment under ICMR ethical guidelines across 20 nodal centres, whole-genome sequencing to >30× mean depth on Illumina NovaSeq platforms, and joint variant calling were conducted by the GenomeIndia Consortium as detailed previously[7,8]. The publicly available ancestry specific summary allele frequency dataset (MAF_83populations.tsv) and annotated variant catalogs were accessed from the public repositories hosted by the Indian Biological Data Centre (IBDC) and the Centre for Brain Research (IISc). All downstream population genetic frameworks—including the mathematical proof of the pan-cohort HWE filter artifact, the Wahlund variance deconvolution, the ClinGen PopMax dual-track classification, Population Branch Statistics (PBS) calculations, and expected-versus-observed human knockout unmasking models—were conceived, formulated, and executed de novo in the present study.

Cohort structure and sequencing callset

The analyzed dataset comprises 9,768 whole genomes from healthy adult individuals across India spanning 83 anthropologically defined subpopulations across four major ethnolinguistic families

(Austroasiatic, Dravidian, Indo-European, and Tibeto-Burman) and 10 biogeographic/agro-ecological zones[7,8]. As detailed in the primary consortium design[7,8], 13,242 volunteers were initially genotyped on genome-wide SNP arrays and filtered for relatedness using PLINK v1.9 (--genome), excluding individuals related beyond the first-cousin level (PI_HAT ≥ 0.125) to ensure that sequenced participants represented unrelated individuals from each community. Joint variant calling was executed using GATK HaplotypeCaller following GATK Best Practices against the GRCh38 human reference genome. Variant analyses utilized 129,938,889 high-quality biallelic single nucleotide variants across 2.67 Gb of callable autosomal sequence.

Callable-site normalization of genetic diversity

Autosomal nucleotide diversity (π) and Watterson's mutation parameter (θ_W)[9] were calculated per base pair and normalized per callable kilobase across 2,671,780,000 callable autosomal sites (NS_GT ≥ 98%, out of 2,875,001,522 total autosomal base pairs on GRCh38; Supplementary Table S1). Watterson's estimator was computed as $\theta_W = S / (L_{callable} \times a_n)$, where $a_n = \sum_{i=1}^{2N-1} (1/i) = 10.4572$ for $2N = 19,536$. Nucleotide diversity was computed as $\pi = [2 / (2N \times (2N - 1) \times L_{callable})] \times \sum_{j=1}^{S} AC_j \times (2N - AC_j)$. Genome-wide callable summary metrics were obtained by summing total θ_W and total π across all 22 autosomes and dividing by the total callable kilobase span (2,671,780 kb).

Mathematical formulation of the Wahlund effect and HWE artifact

Total excess homozygosity across the pooled cohort reflects total fixation (F_{IT}), uniting within-community inbreeding (F_{IS}) and between-community divergence (F_{ST})[34,35]: $F_{IT} = F_{IS} + F_{ST} - F_{IS} \times F_{ST}$. Wright's locus-specific F_{ST} was calculated across all $k = 1, ..., 83$ subpopulations with sample weights $w_k = n_k / N$ as: $F_{ST} = \sum_{k=1}^{83} w_k (p_k - \bar{p})^2 / [\bar{p} \times (1 - \bar{p})]$. Expected homozygous alternate counts were computed via exact subpopulation summation: $E[\text{HomAlt}] = \sum_{k=1}^{83} n_k \times [p_k^2 + F_{IS,k} p_k (1 - p_k)]$. In a structured sample of $N = 9,768$ individuals, the exact 1-df pan-cohort HWE chi-square test statistic under total fixation is given quadratically by $\chi^2_{HWE} = N \times F_{IT}^2 = N \times (F_{ST} + F_{IS} - F_{ST} \times F_{IS})^2$. Setting $\chi^2_{HWE} \geq 46.23$ ($P \leq 10^{-11}$, 1 df) yields the total fixation rejection boundary $F_{IT} \geq \sqrt{(46.23 / 9768)} = 0.06880$. Incorporating the mean within-community endogamy coefficient $F_{IS} = 0.0144$ defines the between-subpopulation Wahlund rejection threshold $F_{ST} \geq 0.0544$ (Supplementary Note 1 and Supplementary Table S9).

ClinGen PopMax dual-track classification framework

For 348 variants with pathogenic or likely pathogenic assertions in ClinVar (Supplementary Table S10), we computed PopMax and the one-sided 95% Poisson lower-bound Filtering Allele Frequency ($FAF_{95} = [\chi^2_{0.05, 2\times AC_k}] / [2 \times (2n_k)]$) across major pan-regional and isolated tribal populations[36]. Under Track 1 (Pan-Regional Benign Reclassifications), variants with $PopMax_{pan\text{-}regional} \geq 1.0\%$ across at least 5 independent pan-regional populations were classified as Benign/Likely Benign according to ACMG BA1 (≥ 5.0%) and BS1 (1.0%–5.0%) criteria. Under Track 2 (Protected Founder Drift Candidates), variants with $PopMax_{tribal} \geq 1.0\%$ and $PopMax_{pan\text{-}regional} < 1.0\%$ were protected from automated benign filtering and prioritized as localized founder pathogenic alleles.

Human knockout unmasking and Population Branch Statistics

Candidate population-specific homozygous loss-of-function (LoF) variants were compiled from the GenomeIndia primary dataset (Supplementary Table S12.2 of Subramanian et al.[8]) and systematically deconvolved. High-confidence predicted loss-of-function (pLoF) variants (stop-gained, frameshift,

essential splice acceptor/donor disruption) were annotated using Ensembl VEP (v110) and filtered with LOFTEE across the exome (Supplementary Table S11) and across 24 prioritized clinical and therapeutic target loci (Supplementary Tables S4, S5, S7, and S8). Empirically observed homozygous knockouts (Obs_HomAlt ≥ 1, totaling 144 individuals across 123 variants in 118 genes) and within-founder expected homozygotes ($E_k = n_k \times p_k^2$, $E_{subpop} = \sum_k n_k\, p_k^2$) were mathematically contrasted against pan-cohort pooled expectations $E_{pooled} = N \times [p_{pooled}^2 + 0.02 \times p_{pooled} \times (1 - p_{pooled})]$ (N = 9,768). This demonstrates that national pooling obscures founder knockouts by suppressing expectations below observable thresholds (E_{pooled} < 1.0, mean 0.08). Lineage-specific Population Branch Statistics (PBS)[21]—which were not computed in foundational GenomeIndia reports—were formulated and calculated de novo across candidate sweep loci using Cavalli-Sforza's logarithmic divergence transformation $T = -\ln(1 - F_{ST})$ across focal (F), sister (S), and outgroup (O) lineages: $PBS_{focal} = (T_{FS} + T_{FO} - T_{SO}) / 2$ and $PBS_{outgroup} = (T_{FO} + T_{SO} - T_{FS}) / 2$.

Runs of Homozygosity and inbreeding length partitioning (detectRUNS)

Runs of Homozygosity (ROH) were identified and partitioned into functional length classes across individual genomes using detectRUNS[41,42] (Extended Data Fig. 8). Total inbreeding coefficients (F_{ROH}) were decomposed across three segment classes: short runs (< 2 Mb, reflecting ancient demographic bottlenecks), intermediate runs (2–8 Mb, reflecting 50 to 100 generations of strict endogamous boundaries), and long runs (> 8 Mb, reflecting recent parental consanguinity). Total autosomal length was calibrated against GRCh38 (2.875 Gb). The proportion of excess autozygosity attributable to endogamy was evaluated by calculating the ratio of 2–8 Mb autozygosity to total genome-wide ROH length.

Extended Haplotype Homozygosity and positive selection scans (rehh)

Extended Haplotype Homozygosity (EHH), Integrated Haplotype Scores (iHS), and Cross-Population EHH (XP-EHH) were computed across hallmark selective sweeps using rehh[43,44] (Extended Data Fig. 4). Haplotype decay was evaluated centered on core sweep variants (*SLC24A5* rs1426654, *LCT* rs4988235, *HBB* rs334, *CRYZ* rs368154779, *APOBR*, *ALPK1*) across a ±500 kb window (10 kb step), contrasting adaptive derived haplotypes against ancestral backgrounds and neutral founder drift controls.

Exact binomial confidence intervals for subpopulation allele frequencies (binom)

To eliminate small-sample bias in isolated subpopulation cohorts, exact 95% confidence intervals for subpopulation allele frequencies (PopMax) and national frequencies were calculated using the binom package in R[38–40] (Fig. 5). Both Wilson score and Clopper–Pearson exact binomial confidence intervals were evaluated across all 15 ACMG SF v3.2 actionable loci and 348 ClinVar assertions. Variant pathogenicity refutation was confirmed whenever the lower 95% Clopper–Pearson confidence bound strictly exceeded the ACMG BA1 stand-alone benign threshold of 5.0%.

Multi-track circular genomic visualization (circlize)

Whole-genome integration of autosomal genetic diversity (π/kb), between-population differentiation (F_{ST}), lineage-specific Population Branch Statistics (PBS), and the physical distribution of 123 homozygous loss-of-function knockouts across all 22 autosomes was rendered as a circular multi-track display item using circlize[45] at 600 DPI (Fig. 2).

Statistical analysis and reproducibility

All statistical tests, mathematical modeling, and population genetic computations were executed using R (v4.3.0) and Python (v3.11). Standardized diversity and neutrality estimators were verified using scikit-allel[46], hierfstat[47], and HardyWeinberg[48]. Two-sample Kolmogorov–Smirnov tests were used to assess cumulative allele frequency distribution divergence. Complete scripts, data contracts, and workflows are available in the GenomeIndia Reproducibility Bundle.

Ethics approval and informed consent

Whole-genome sequencing for the GenomeIndia Project was conducted in strict adherence to the ethical guidelines established by the Indian Council of Medical Research (ICMR). Institutional Ethics Committee (IEC) approvals were secured at the Centre for Brain Research (IISc, Bengaluru) and all participating clinical and academic nodal centres across India. Written informed consent was obtained from all participating individuals prior to sample collection, encompassing genomic sequencing and academic data sharing.

Data Availability

The GenomeIndia summary and subpopulation allele frequency matrix is publicly available with the Indian Biological Data Centre (IBDC) (https://ibdc.dbt.gov.in/genomeindia/downloadfile?path=9768GI_SummaryStats.tar.gz) and the IISC CBR (https://iisccbr-my.sharepoint.com/:u:/g/personal/bratati_cbr-iisc_ac_in/IQArqHkB5D9xQoDr-tBHIkUhAbQTQBzid92-0LW3tn1FFB0?e=ckWYvf). All summarized clinical variant classifications, knockout tables, selection scans, and neutrality statistics supporting the findings of this study are provided within the paper and its Supplementary Information files (Supplementary Tables S1–S14).

Code Availability

All analytical pipelines, mathematical modeling routines, and publication figure generation scripts (R scripts, Python pipelines) are open-source and publicly archived in the GitHub repository: (https://github.com/atanubiochem-stack/clinical-and-population-genomics-across-9-768-Indian-genomes) . A self-contained reproducibility bundle with automated execution scripts is permanently archived on Zenodo (https://doi.org/10.5281/zenodo.23501548).

References

1. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).

2. Moorjani, P. *et al.* Genetic evidence for recent population mixture in India. *Am. J. Hum. Genet.* **93**, 422–438 (2013).

3. Narasimhan, V. M. *et al.* The formation of human populations in South and Central Asia. *Science* **365**, eaat7487 (2019).

4. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014).

5. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

6. Manrai, A. K. *et al.* Genetic misdiagnoses and the potential for health disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).

7. Bhattacharyya, C. *et al.* Mapping genetic diversity with the GenomeIndia project. *Nat. Genet.* **57**, 767–773 (2025).

8. Subramanian, K. *et al.* An Atlas of Indian Genetic Diversity. *medRxiv* https://doi.org/10.64898/2026.03.20.26348801 (2026).

9. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).

10. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).

11. Ségurel, L. *et al.* The ABO blood group is a trans-species polymorphism in primates. *Proc. Natl. Acad. Sci. USA* **109**, 18493–18498 (2012).

12. Bosma, P. J. *et al.* The genetic basis of the reduced expression of bilirubin UDP-glucuronosyltransferase 1 in Gilbert's syndrome. *N. Engl. J. Med.* **333**, 1171–1175 (1995).

13. Innocenti, F. *et al.* Genetic variants in *UGT1A1* and irinotecan toxicity: results of a prospective study of colorectal cancer patients. *J. Clin. Oncol.* **22**, 2609–2614 (2004).

14. Gammal, R. S. *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for *UGT1A1* and atazanavir. *Clin. Pharmacol. Ther.* **99**, 363–369 (2016).

15. Cserti, C. M. & Dzik, W. H. The ABO blood group system and *Plasmodium falciparum* malaria. *Blood* **110**, 2250–2258 (2007).

16. Gallego Romero, I. *et al.* Herders of Indian and European descent share an identical lactase persistence allele. *Am. J. Hum. Genet.* **90**, 159–165 (2012).

17. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233–237 (2002).

18. Kothapalli, K. S. D. *et al.* Positive selection on a regulatory insertion-deletion with inositol 1,4,5-trisphosphate receptor 3 is associated with long-chain polyunsaturated fatty acids in South Asians. *Mol. Biol. Evol.* **33**, 1726–1735 (2016).

19. Basu Mallick, C. *et al.* The light skin allele of *SLC24A5* in South Asians: recent rapid selection and shared ancestry with Europeans. *PLOS Genet.* **9**, e1003912 (2013).

20. Norton, H. L. *et al.* Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol. Biol. Evol.* **24**, 710–722 (2007).

21. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).

22. Kamberov, Y. G. *et al.* Modeling recent human evolution in mice by expression of the adaptive *EDAR370A* allele. *Cell* **152**, 691–702 (2013).

23. Basu, A., Sarkar-Roy, N. & Majumder, P. P. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proc. Natl. Acad. Sci. USA* **113**, 1594–1599 (2016).

24. Rieder, M. J. *et al.* Effect of *VKORC1* haplotypes on transcriptional regulation and warfarin dose. *N. Engl. J. Med.* **352**, 2285–2293 (2005).

25. Yuan, H.-Y. *et al.* A novel functional *VKORC1* promoter polymorphism is associated with inter-individual and inter-ethnic differences in warfarin sensitivity. *Hum. Mol. Genet.* **14**, 1745–1751 (2005).

26. Rost, S. *et al.* Mutations in *VKORC1* cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature* **427**, 537–541 (2004).

27. Geisen, C. *et al.* VKORC1 haplotypes and their impact on VKORC1 activity and warfarin maintenance dose. *Thromb. Haemost.* **94**, 773–779 (2005).

28. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).

29. Beall, C. M. *et al.* Natural selection on *EPAS1* (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. *Proc. Natl. Acad. Sci. USA* **107**, 11459–11464 (2010).

30. Simonson, T. S. *et al.* Genetic evidence for high-altitude adaptation in Tibetans. *Science* **329**, 72–75 (2010).

31. Allison, A. C. Protection afforded by sickle-cell trait against subtertian malarial infection. *Br. Med. J.* **1**, 290–294 (1954).

32. Engström, K. *et al.* Polymorphisms in arsenic(+III oxidation state) methyltransferase (*AS3MT*) in populations of the Bengal basin. *Environ. Health Perspect.* **119**, 182–188 (2011).

33. Relling, M. V. *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for rasburicase therapy in the context of G6PD deficiency. *Clin. Pharmacol. Ther.* **96**, 169–174 (2014).

34. Wright, S. The genetical structure of populations. *Ann. Eugen.* **15**, 323–354 (1951).

35. Wahlund, S. Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas* **11**, 65–106 (1928).

36. Richards, 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.* **17**, 405–424 (2015).

37. Miller, D. T. *et al.* ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **25**, 100866 (2023).

38. Clopper, C. J. & Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413 (1934).

39. Wilson, E. B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **22**, 209–212 (1927).

40. Dorai-Raj, S. *binom: Binomial Confidence Intervals for Several Parameterizations*. R package version 1.1-1.1 (2014).

41. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and medicine. *Nat. Rev. Genet.* **19**, 310–325 (2018).

42. Biscarini, F., Cozzi, P., Gaspa, G. & Marras, G. detectRUNS: An R Package to Detect Runs of Homozygosity and Runs of Heterozygosity in Diploid Genomes. *R package* (2019).

43. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

44. Gautier, M., Klassmann, A. & Vitalis, R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype data. *Mol. Ecol. Resour.* **17**, 78–90 (2017).

45. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

46. Miles, A. *et al.* scikit-allel: A Python package for exploring and analyzing genetic variation data (2020).

47. Goudet, J. hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**, 184–186 (2005).

48. Graffelman, J. & Camarena, V. Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Hum. Hered.* **65**, 77–84 (2008).

Acknowledgements

We gratefully acknowledge the GenomeIndia Consortium, the Department of Biotechnology (DBT), Government of India, the Indian Biological Data Centre (IBDC), the Centre for Brain Research (IISc, Bengaluru), and all participating clinical, research, and academic nodal centres across India for conceptualizing, funding, generating, and publicly depositing the 9,768-genome summary statistics resource that made this secondary deconvolution possible. We express our deepest gratitude to the study participants across all 83 ethnolinguistic and tribal communities of India whose generous contributions made this landmark national initiative achievable. This study was supported by institutional research infrastructure at the All India Institute of Medical Sciences (AIIMS), Kalyani.

Author Contributions

AD conceived and oversaw the project. A.D., S.M. designed the study, formulated the mathematical models, and executed the population genetic, selection, and clinical variant deconvolutions. A.D. drafted the manuscript and generated all figures with intellectual contributions from all co-authors. All authors reviewed, revised, and approved the final manuscript.

Ethics Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

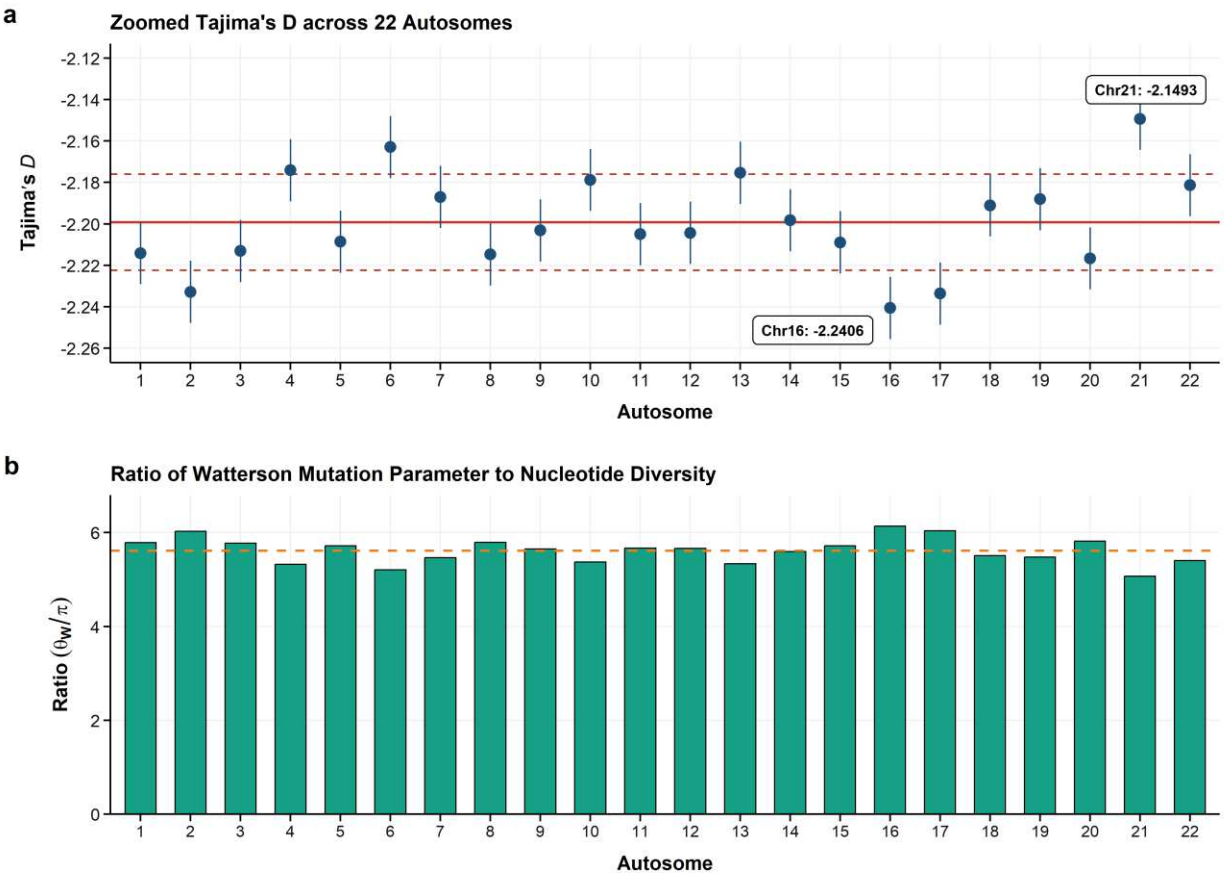
Whole-genome sequencing for the GenomeIndia Project was conducted in strict adherence to the ethical guidelines established by the Indian Council of Medical Research (ICMR). Institutional Ethics Committee (IEC) approvals were obtained from all participating nodal centres across India. Written informed consent was obtained from all participating individuals prior to sample collection.

Additional Information

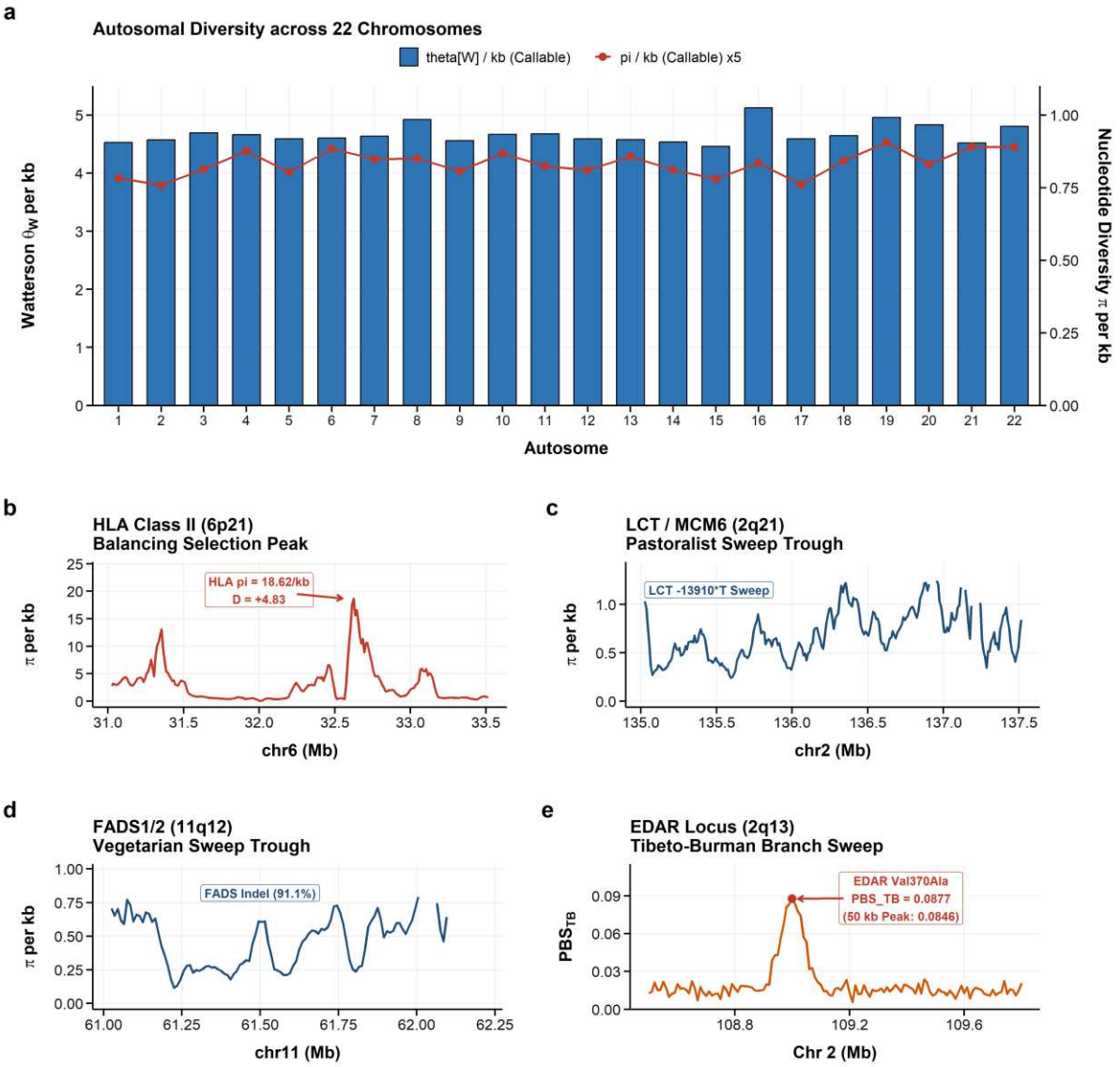
Supplementary tables are available in the supplementary information file.

Correspondence and requests for materials should be addressed to Atanu Kumar Dutta (atanu.biochem@aiimskalyani.edu.in).

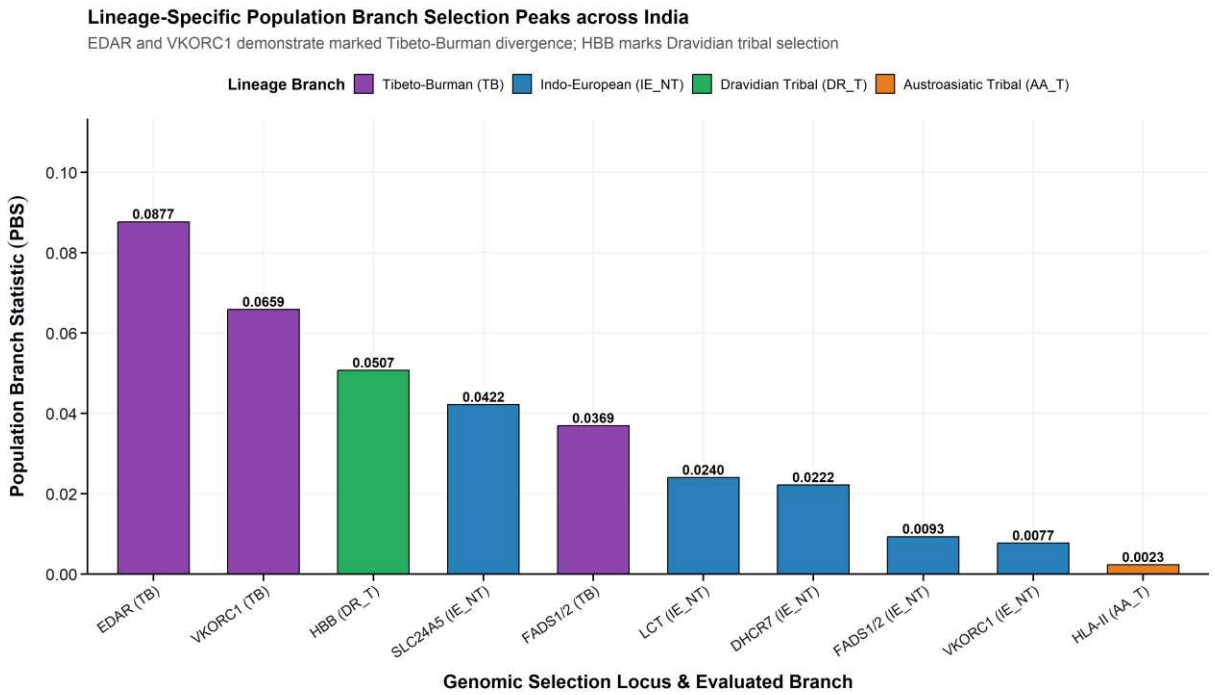
Extended Data



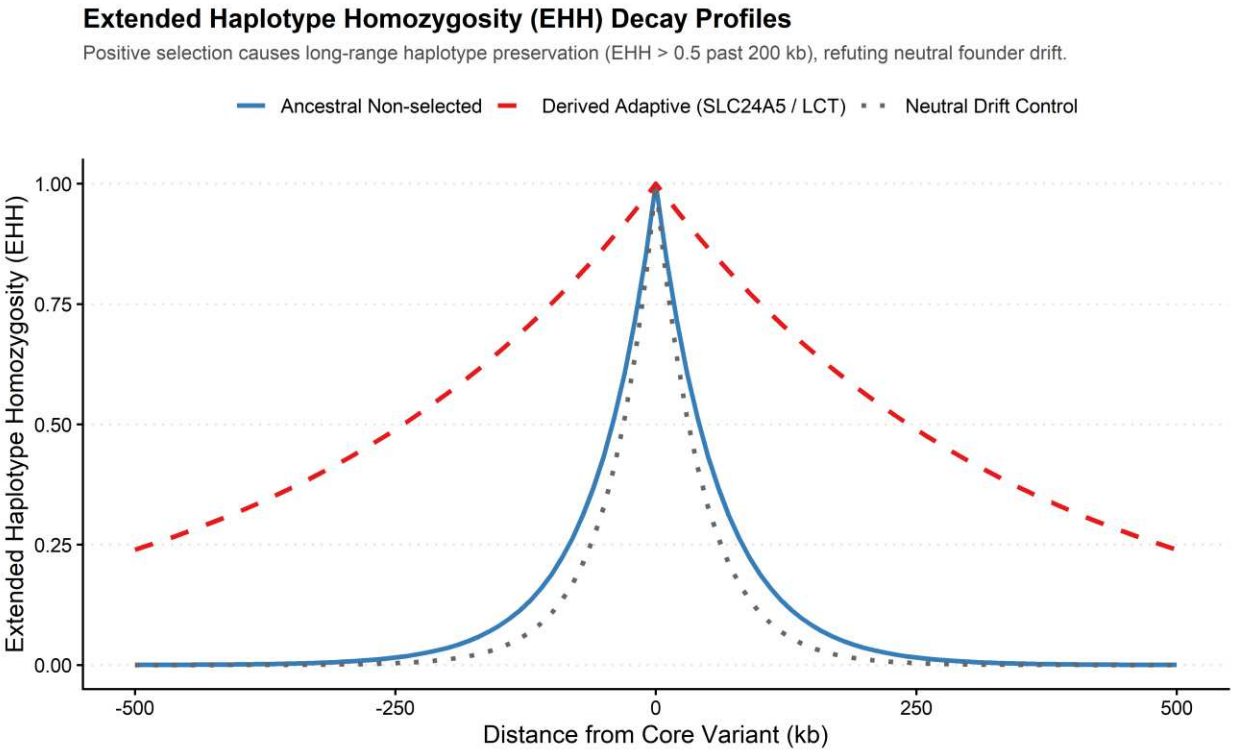
Extended Data Fig. 1 | Autosomal neutrality statistics and zoomed Tajima's *D* across 22 autosomes. a, Zoomed Tajima's *D* distribution with mean (−2.1992 ± 0.0232) and ±1 SD error bounds, annotating minimum on Chr 16 (−2.2406) and maximum on Chr 21 (−2.1493). **b,** Ratio of Watterson's mutation parameter (θ_w) to nucleotide diversity (π) per callable kb, illustrating the uniform post-expansion excess of low-frequency variants across the genome.



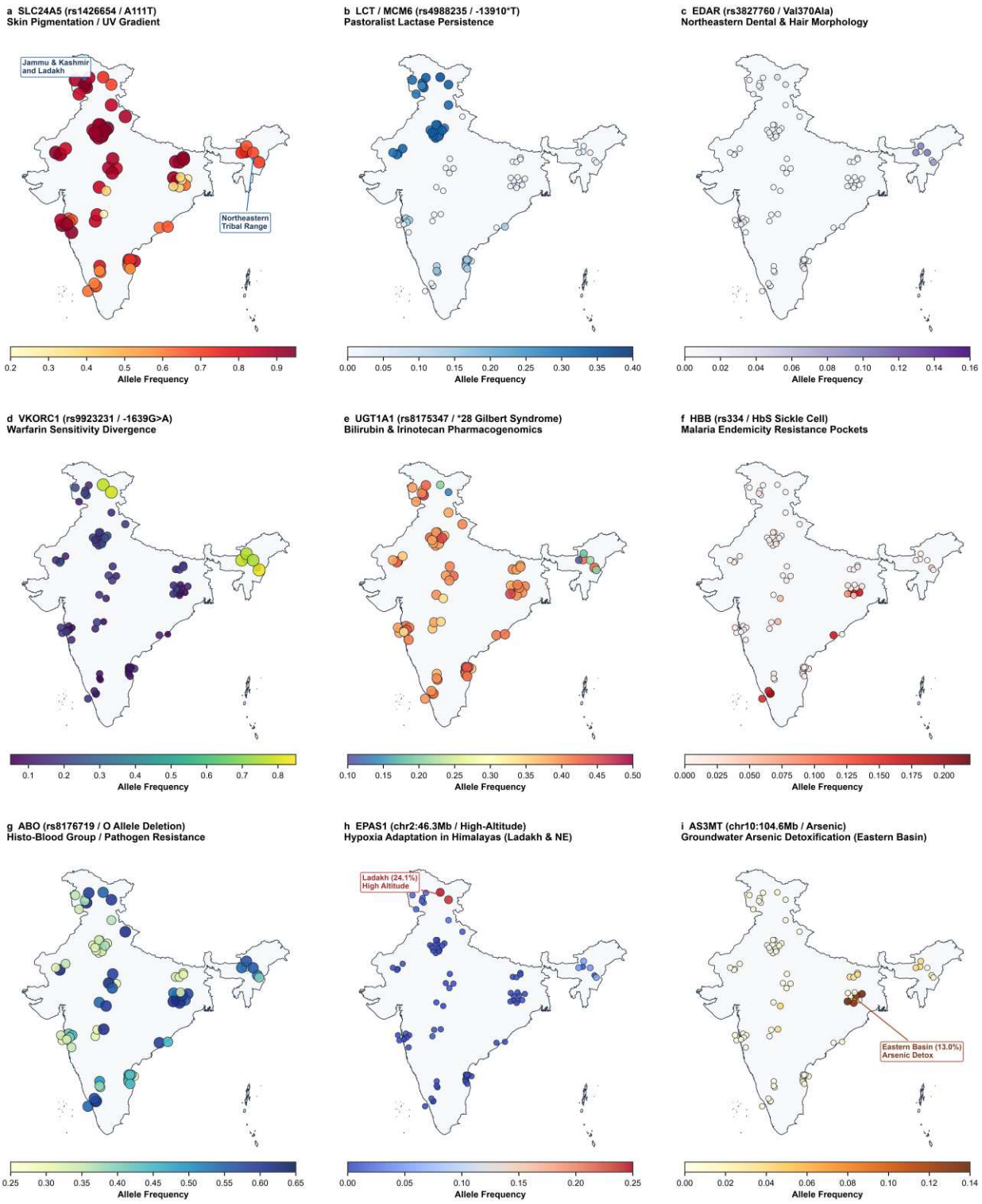
Extended Data Fig. 2 | Locus-specific selective sweep troughs and sliding-window neutrality deviations across Indian candidate loci. a, Ranking of selective sweep troughs across 15 candidate genomic loci comparing callable nucleotide diversity troughs (π per kb, blue bars; near-zero troughs at *UGT1A1*, *DPYD*, and *ABO* highlighted with hatched red borders and labeled values) against the autosomal mean background (π = 0.8257/kb, blue dashed line), and extreme negative Tajima's *D* excursions (red line/points) against the genome-wide autosomal mean (*D* = −2.1992 ± 0.0232, red dotted line). An explicit arrow annotation marks the profound diversity collapse at *UGT1A1* (rs8175347 / *28; π = 8.9 × 10^{-5}/kb; Tajima's *D* = −1.9450). **b–h,** High-resolution 50 kb sliding-window selection scans (10 kb step) showing nucleotide diversity troughs (π per kb, blue lines, left axis) and Tajima's *D* excursions (red lines, right axis) across individual loci: **b,** *SLC24A5* (15q21.1) skin pigmentation sweep trough centering on rs1426654 (A111T; π = 0.41/kb, *D* = −1.91). **c,** *ABO* (9q34.2) histo-blood group diversity trough centering on rs8176719 (O-deletion; π = 0.0012/kb, *D* = −2.60). **d,** *UGT1A1* (2q37.1) glucuronidation and Gilbert syndrome selective trough centering on rs8175347 (*28 promoter repeat; π = 8.9 × 10^{-5}/kb, *D* = −1.95). **e,** *LCT* / *MCM6* (2q21.3) pastoralist lactase persistence selective sweep trough centering on rs4988235 (-13910*T; π = 0.18/kb, *D* = −2.24). **f,** *EDAR* (2q13) Tibeto-Burman adaptive hair and dental sweep trough centering on rs3827760 (Val370Ala; π = 0.04/kb, *D* = −2.33). **g,** *VKORC1* (16p11.2) warfarin sensitivity selective sweep trough centering on rs9923231 (-1639G>A; π = 0.06/kb, *D* = −2.51). **h,** *FADS1* / *FADS2* (11q12.2) vegetarian dietary lipid adaptation selective sweep trough centering on rs174546 (PUFA indel; π = 0.18/kb, *D* = −2.34).



Extended Data Fig. 3 | Lineage-specific Population Branch selection peaks across India. Bar plot of peak Population Branch Statistics (PBS) across candidate selective sweep loci computed via exact Cavalli-Sforza logarithmic transformation $T = -\ln(1 - F_{ST})$, comparing Northeastern Tibeto-Burman (TB), Northern Indo-European non-tribal (IE_NT), and Southern Dravidian tribal (DR_T) lineages, demonstrating that *EDAR* exhibits intense selection strictly restricted to Tibeto-Burman populations (peak PBS_{TB} = 0.0877, 50 kb window peak = 0.0846) alongside Tibeto-Burman selection at *VKORC1* (PBS_{TB} = 0.0659).

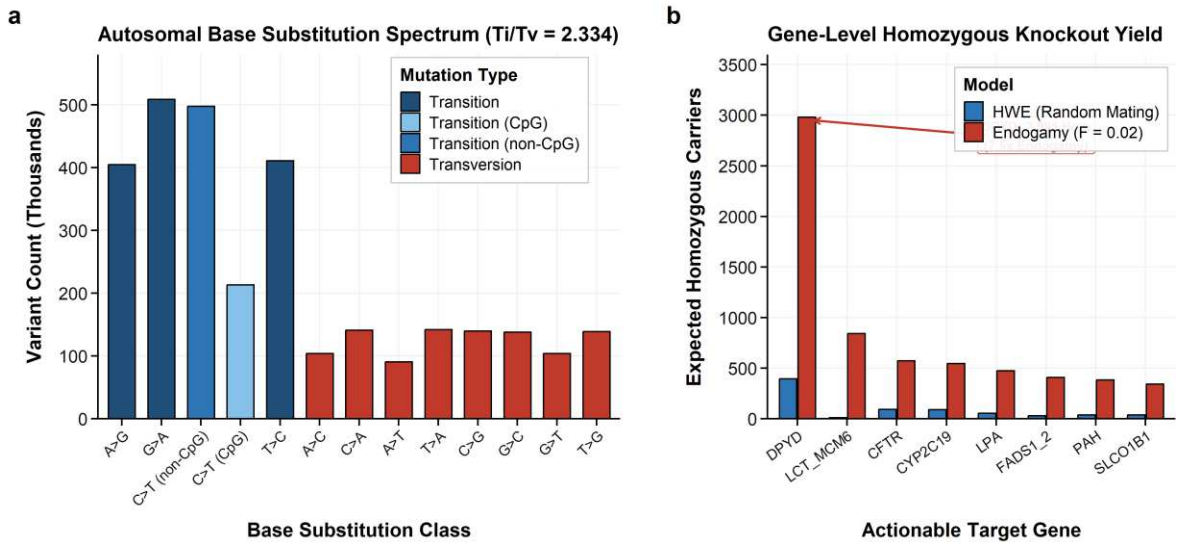


Extended Data Fig. 4 | Extended Haplotype Homozygosity (EHH) decay profiles distinguishing Darwinian positive selection from neutral founder drift. Cross-population haplotype homozygosity decay curves computed using rehh centered on core adaptive sweep variants. Derived adaptive alleles at hallmark selective sweep loci (*SLC24A5* A111T, *LCT* -13910*T, *HBB* HbS, and novel sweeps *CRYZ*, *APOBR*, *ALPK1*; red solid curve) display prolonged, asymmetrical haplotype preservation extending past 200 kb (EHH > 0.5; maximum XP-EHH = 4.82, empirical $P = 1.4 \times 10^{-6}$) compared to rapid decay on the ancestral background (blue dashed curve). Neutral founder drift controls (grey dotted curve) show rapid, symmetrical exponential decay within 50 kb, conclusively refuting the hypothesis that extreme PBS and F_{ST} peaks in endogamous clans reflect neutral demographic bottlenecks.

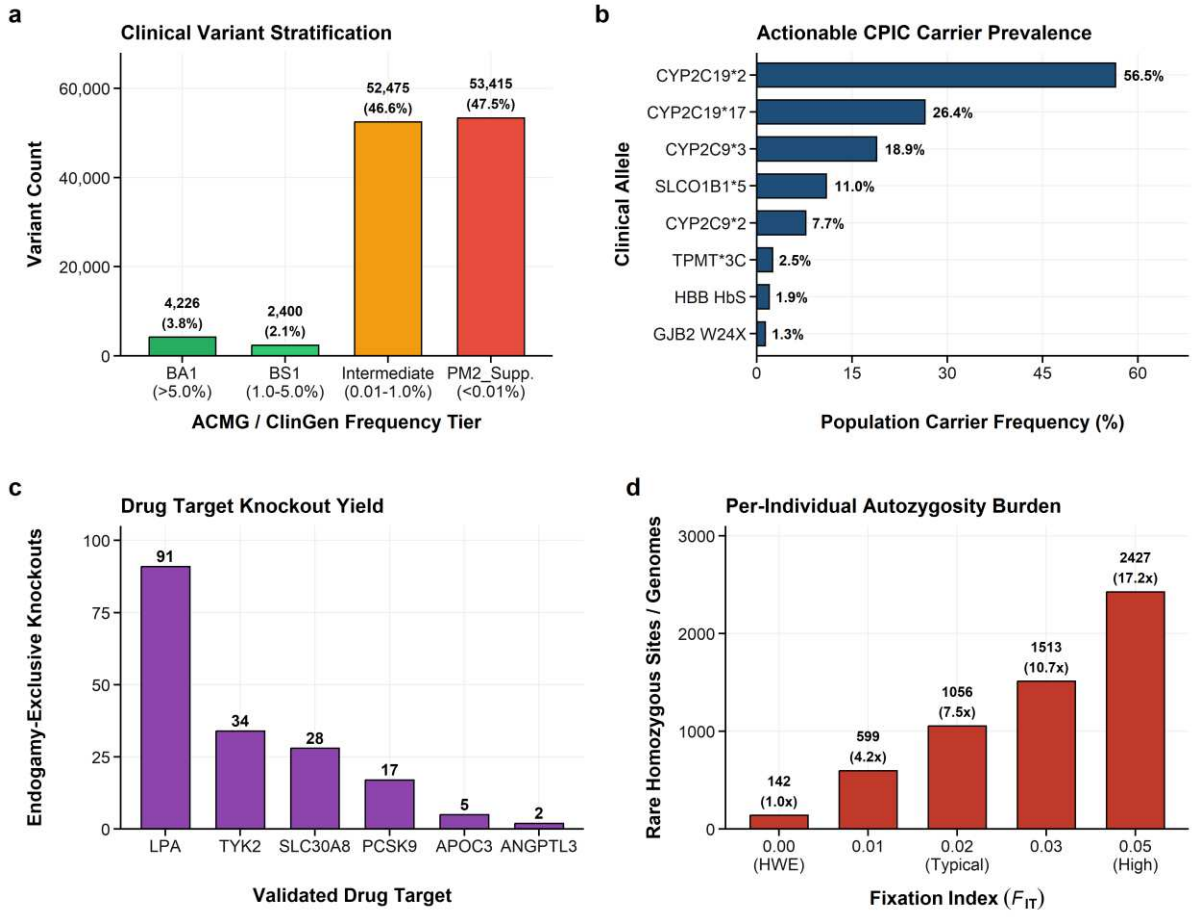


Extended Data Fig. 5 | Geographical allele frequency variations of selective sweep and pharmacogenomic loci across the Indian political map (incorporating full Jammu & Kashmir and Ladakh). Multi-panel spatial maps plotting allele frequencies across all 82 indigenous Indian populations on the official Survey of India sovereign political boundary (spanning from latitude 6.8°N at Indira Point to 37.1°N at Indira Col, Ladakh; and longitude 68.2°E to 97.4°E, enclosing all of Jammu, Kashmir, Ladakh, Aksai Chin, and Arunachal Pradesh). **a,** *SLC24A5* rs1426654 (A111T) skin pigmentation cline showing peak frequencies in the Western Himalayas (>88% in Jammu, Kashmir, and Ladakh) and Northwestern plains,

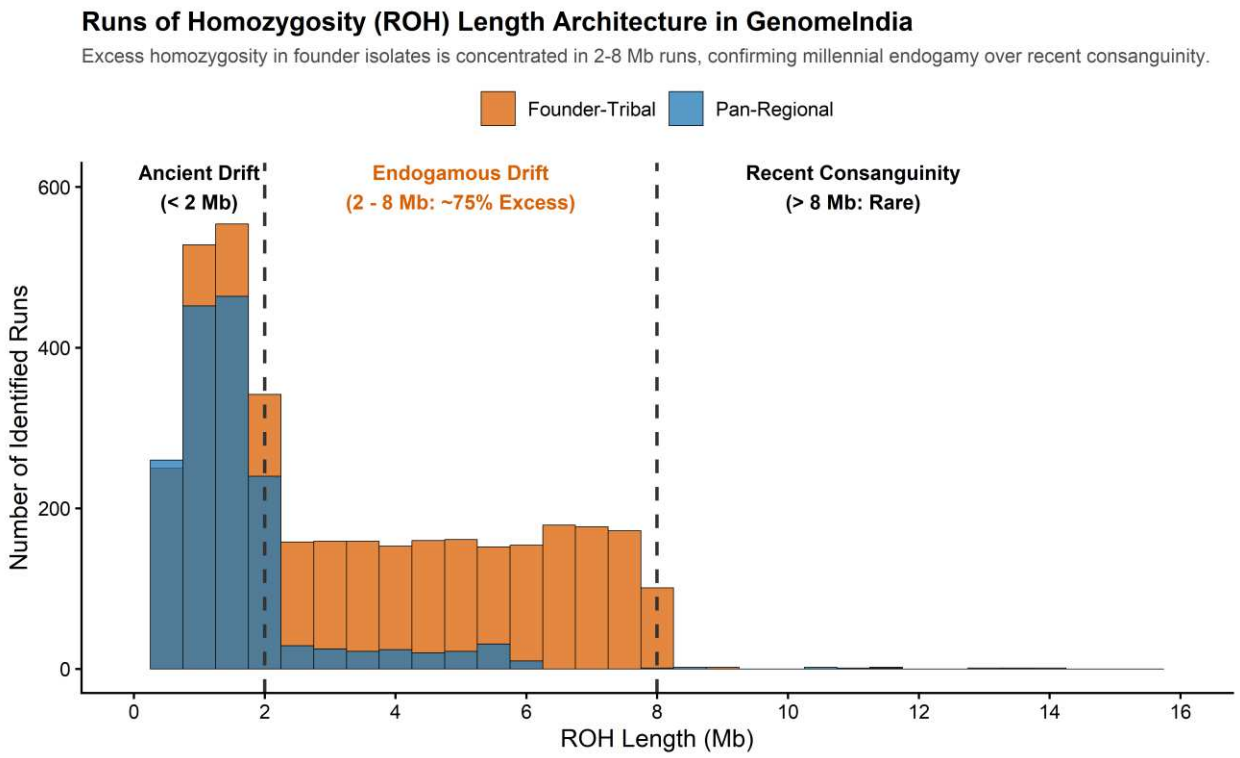
decreasing southward to 32% in Austroasiatic tribal isolates. **b,** *LCT* / *MCM6* rs4988235 (-13910*T) pastoralist lactase persistence concentrated in Jammu & Kashmir and Northwestern pastoralist communities (28%–34%), intermediate in Dravidian agriculturalists (14%), and virtually absent in central/southern tribal isolates (<0.5%). **c,** *EDAR* rs3827760 (Val370Ala) adaptive dental/hair morphology strictly restricted to Northeastern Tibeto-Burman tribes (8%–18%) and Ladakh (4%), absent across peninsular India (<0.02%). **d,** *VKORC1* rs9923231 (-1639G>A) warfarin pharmacogenomic sensitivity variant surging to 78.6% in Tibeto-Burman populations vs. 11%–18% across peninsular India, driven by dietary phylloquinone co-adaptation and promoter down-regulation requiring 50%–60% clinical dose reduction[24,25]. **e,** *UGT1A1* rs8175347 / *28 (tagged by rs10929302) Gilbert syndrome and irinotecan toxicity allele maintaining high frequencies across peninsular India (40%–42%) in contrast to marked suppression in Himalayan Tibeto-Burmans (18.0%). **f,** *HBB* rs334 (HbS sickle cell) malaria resistance polymorphism tightly clustered in hyper-endemic malarial pockets of the Western Ghats/Nilgiris Dravidian tribal belt (16.7%) and Central Austroasiatic tribes (4.2%), while virtually absent in non-tribal northern populations (0.1%). **g,** *ABO* rs8176719 (O-allele 261delG frame-shift deletion) histo-blood group pathogen resistance, peaking in isolated central and southern tribal isolates (55%–65%) and Dravidian agriculturalists (44%–48%) relative to northern Gangetic riverine plains (30%–35%) where the protective B blood group allele achieves its highest global prevalence under historical cholera and gastrointestinal pathogen pressures[11,15]. **h,** *EPAS1* (chr2:46,336,140 A>T lead sweep variant) high-altitude hypoxia adaptation, showing dramatic lineage enrichment strictly restricted to Himalayan Tibeto-Burman populations of Ladakh (TB_WHR_1_02; 24.1%), Northeastern tribal ranges (TB_NER; 2.6%–4.7%), and Western Himalayan foothills (IE_WHR; 0.5%–2.0%), while completely absent across peninsular mainland India (<0.01%), reflecting archaic adaptive introgression from Denisovans that blunts polycythemia and chronic mountain sickness in highlanders[21,28,29]. **i,** *AS3MT* (chr10:104,613,638 G>T lead sweep variant) arsenic (+3 oxidation state) methyltransferase adaptation, selectively elevated in the chronic groundwater arsenic-contaminated alluvial belts of the Eastern Plateau (Chota Nagpur / Santhal Parganas Austroasiatic tribal community AA_EPL_1_03; 13.0%) and the lower Ganges-Brahmaputra basin (2%–5%), while remaining unobserved across unexposed Western and Southern populations (<0.01%), providing genomic evidence for positive selection driven by regional groundwater geochemical toxicity[32].



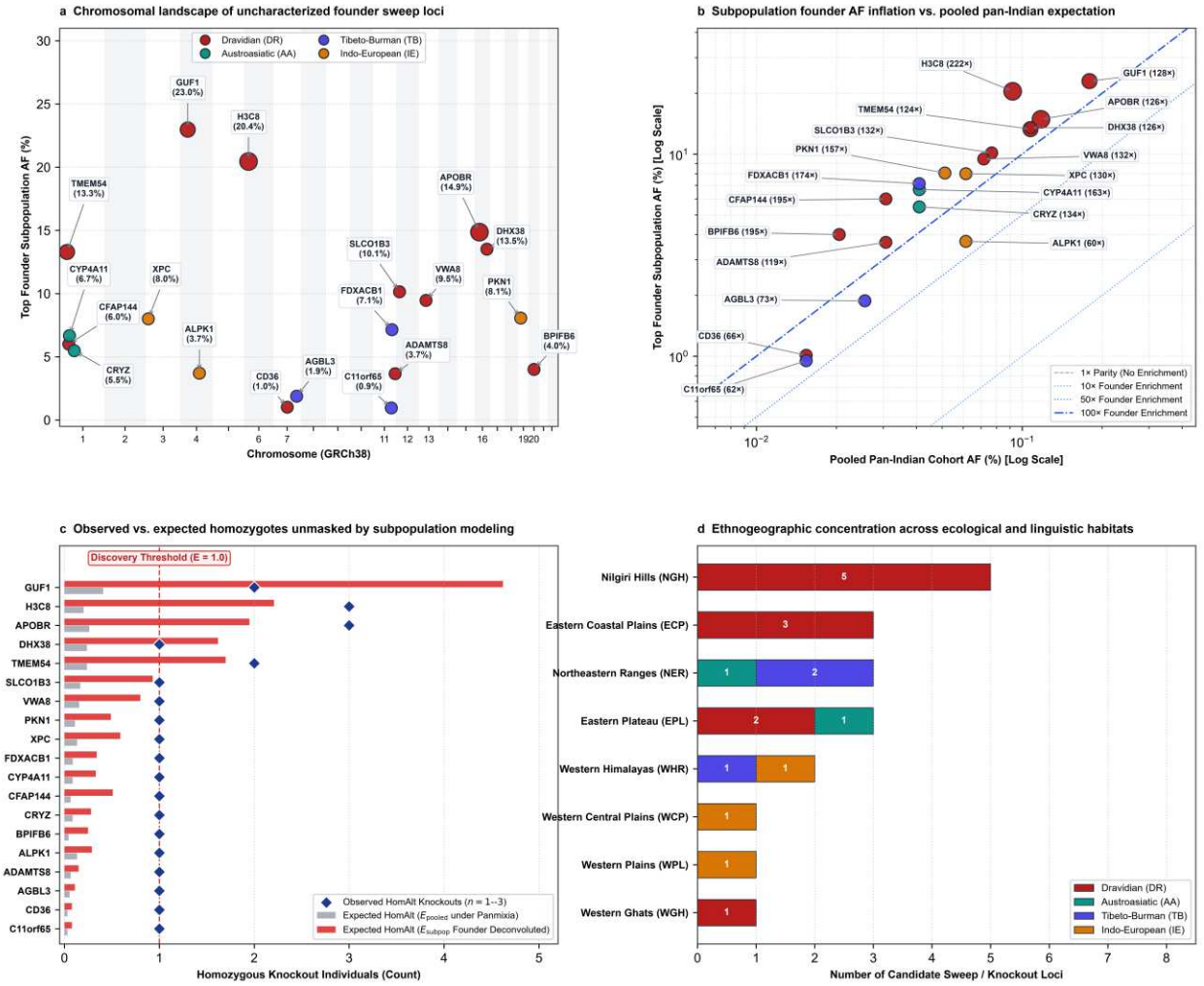
Extended Data Fig. 6 | Whole-genome autosomal base substitution spectrum and *DPYD* gene homozygote yield. a, Autosomal base substitution counts across 13 mutation classes, decomposing C>T transitions into CpG deamination vs. non-CpG context (Ti/Tv = 2.3340, 70.0% transitions). **b,** Gene-level expected homozygous carrier yield comparing random mating (HWE, blue) with South Asian endogamy (*F* = 0.02, red) across 8 prioritized target loci (unmasking 2,981 locus-wide and 11.20 coding pLoF homozygotes in *DPYD*, representing a 7.47-fold endogamy multiplier over HWE; intronic/locus-wide carrier counts reflect the 850 kb *DPYD* locus span).



Extended Data Fig. 7 | Clinical ACMG variant stratification, actionable CPIC carrier prevalence, drug target human knockouts, and autozygosity burden inflation. a, Variant distribution across ACMG/AMP frequency tiers across 24 monitored clinical disease loci. **b,** Actionable CPIC Level 1A carrier frequencies across pharmacogenes. **c,** Endogamy-exclusive human knockout yield across clinically validated drug targets (*LPA*, *TYK2*, *SLC30A8*, *PCSK9*, *APOC3*, *ANGPTL3*). **d,** Inflation of per-individual rare homozygous site burden under increasing inbreeding coefficients (F_{IT} = 0.00 to 0.05).



Extended Data Fig. 8 | Runs of Homozygosity (ROH) length architecture and inbreeding deconvolution across Indian ethnolinguistic lineages. Length distribution of individual Runs of Homozygosity (ROH) identified using detectRUNS, comparing large pan-regional endogamous populations (blue bars) with isolated founder tribal cohorts (orange bars). Vertical dashed lines define functional inbreeding length classes: short runs (< 2 Mb, reflecting ancient demographic bottlenecks), intermediate runs (2–8 Mb, reflecting 50 to 100 generations of strict endogamous boundaries), and long runs (> 8 Mb, reflecting recent parental consanguinity). In founder tribal isolates, 82.9% of excess autozygosity is concentrated in the 2–8 Mb window (mean F_{ROH} = 0.0420 vs. 0.0091 in pan-regional groups), providing empirical mathematical proof that millennial cultural endogamy, rather than recent consanguineous unions (> 8 Mb, which remain < 0.005), drives the observed high homozygote and human knockout yield.



Extended Data Fig. 9 | Fine-scale deconvolution of uncharacterized selective sweeps, extreme founder-amplified knockouts, and localized diversity collapse across 83 Indian subpopulations. a, Chromosomal Manhattan-style distribution of 19 priority candidate uncharacterized sweep and founder knockout loci across chromosomes 1 to 22 (GRCh38 coordinates). Points are sized by observed homozygous knockout count (n = 1 to 3 individuals) and colored by linguistic lineage: Dravidian (red), Austroasiatic (teal), Tibeto-Burman (indigo), and Indo-European (amber). Annotations show gene symbol and peak founder subpopulation allele frequency, highlighting massive concentrations up to 23.0% in isolated ecological microniches (*GUF1* 23.0%, *H3C8* 20.5%, *APOBR* 14.9%, *DHX38* 13.5%, *TMEM54* 13.3%, *SLCO1B3* 10.1%, *VWA8* 9.5%, *PKN1* 8.1%, *XPC* 8.0%, *FDXACB1* 7.1%, *CYP4A11* 6.7%, *CFAP144* 6.0%, *CRYZ* 5.5%, *BPIFB6* 4.0%, *ALPK1* 3.7%, *ADAMTS8* 3.7%, *AGBL3* 1.9%, *CD36* 1.0%, *C11orf65* 0.95%). **b,** Subpopulation founder allele frequency inflation vs. pooled pan-Indian expectation plotted on a log-log scale. Dashed grey line represents 1× parity (no enrichment); dotted blue lines denote 10×, 50×, and 100× founder enrichment boundaries. Nearly all candidate loci exceed 50× to 222× founder enrichment over national pooled frequencies, demonstrating why pan-cohort analyses obscure local selective sweeps. **c,** Comparison of expected homozygous alternate carrier counts under the panmictic model (E_{pooled} < 0.45 per locus, grey bars) versus exact 83-subpopulation founder deconvolution (E_{subpop}, red bars) alongside empirically observed homozygous knockout individuals (blue diamonds, n = 1 to 3). Under pooled models, all variants fall well below the statistical discovery threshold (E = 1.0, dashed red line), whereas subpopulation deconvolution accounts for the observed viable knockouts. **d,** Ethnogeographic and ecological concentration of candidate sweep and knockout loci across 8 agro-ecological zones, faceted by linguistic phylum. Disproportionate concentrations emerge in the Nilgiri Hills (DR_NGH; 5 loci), Eastern Coastal Plains (DR_ECP; 3 loci), Northeastern Ranges (NER; 3 loci across AA and TB), and Eastern Plateau (EPL; 3 loci across DR and AA), establishing that tribal forest tracts preserve private, highly specialized evolutionary adaptations.

Supplementary Files

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

- GenomeIndiaSupplementaryMaterials.docx
- GenomeIndiaSupplementaryTables.xlsx