Concept Paper

Haplotype-Aware Bayesian Variant Interpretation in Admixed Populations: The Brazilian Case as a Conceptual Stress Test

Marcelo R. S. Briones *, Renata C. Ferreira , Fernando Antoneli

Posted Date: 27 April 2026

doi: 10.20944/preprints202604.1833.v1

Keywords: ACMG/AMP guidelines; human population genetics; pathogenic variants; medical genetics; human genomics

 doi:10.20944/preprints202604.1833.v1

Concept Paper

Haplotype-Aware Bayesian Variant Interpretation in Admixed Populations: The Brazilian Case as a Conceptual Stress Test

Marcelo R. S. Briones *, Renata C. Ferreira and Fernando Antoneli

Center for Medical Bioinformatics, Escola Paulista de Medicina – UNIFESP, Rua Pedro de Toledo, 669, São Paulo, 04039-032, SP, Brazil

* Correspondence: marcelo.briones@unifesp.br

Abstract

Current variant interpretation frameworks, including those proposed by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) and ClinGen, rely on implicit assumptions regarding allele frequencies, linkage disequilibrium (LD), and independence of variables. These assumptions are largely derived from European populations and are not valid in highly admixed populations. Here, we argue that the Brazilian population constitutes a natural stress test for these frameworks, not merely due to the extent of admixture, but due to its qualitative structure, characterized by recent tri-hybrid admixture (European, African, Indigenous American) and extensive recombination generating novel haplotypic configurations. We further highlight the instability of borderline variants of uncertain significance (VUS) under Bayesian classification and propose a reformulation explicitly incorporating haplotype structure and local ancestry.

Keywords: ACMG/AMP guidelines; human population genetics; pathogenic variants; medical genetics; human genomics

Introduction

Bayesian frameworks have become central to clinical variant interpretation, formalizing the integration of heterogeneous evidence into posterior probabilities of pathogenicity (Tavtigian et al., 2018). However, these frameworks depend critically on priors and likelihoods derived from population data that remain heavily biased toward European ancestry (Sirugo et al., 2019).

Admixed populations expose the limitations of this approach. Brazil, in particular, represents one of the most complex admixture scenarios globally, shaped by recent (~500 years) gene flow among European colonizers, enslaved Africans from multiple regions, and diverse Indigenous populations (Pena et al., 2011). This is directly supported by the recent whole-genome analysis of 2,723 Brazilian individuals, which identified over 8 million previously unreported variants absent from major reference databases (Nunes et al., 2025), demonstrating that allele frequency priors derived from non-admixed populations are systematically miscalibrated for this context.

Crucially, this is not simply a matter of degree. Other populations, such as those in India, also exhibit extensive and ancient (4,200 to 1,900 years) admixture (Reich et al., 2009). However, the Brazilian population is characterized not only by a high degree of admixture but by a distinct admixture structure, involving recent, large-scale, tri-continental gene flow that generates mosaic haplotypes with frequent ancestry switching along chromosomes (Pena et al., 2011; Gravel, 2012).

This qualitative difference leads to the emergence of haplotypic combinations that are rare or absent in ancestral populations, challenging the assumptions underlying current variant interpretation frameworks.

 doi:10.20944/preprints202604.1833.v1

The Bayesian Framework and Its Hidden Assumptions

Variant classification can be formalized as (Tavtigian et al., 2018):

$$P(\,Pathogenic \mid D\,) \propto P(\,D \mid Pathogenic\,) \cdot P(Pathogenic) \quad (1)$$

where D denotes the totality of observed evidence used to classify a variant, encompassing functional data, population frequency estimates, computational predictions, segregation data, and clinical observations. Although widely adopted, this formulation embeds several implicit assumptions: (1) allele frequencies are stable across populations, (2) LD structure is predictable and transferable and (3) criteria for significant evidence are conditionally independent. These assumptions are rarely stated explicitly but are fundamental to the validity of the framework. In (Tavtigian et al., 2018) D is not named as such, instead, it is defined as *P(A)* the prior probability of pathogenicity, *P(B)* as the probability of the evidence, *P(A|B)* as the posterior probability, and *P(B|A)* as the probability of the evidence given that the variant is pathogenic. Therefore, D in the present study (Equation (1)) corresponds to what (Tavtigian et al., 2018) calls *P(B)*, the evidence composite.

The ACMG/AMP Evidence is used in an Odds Pipeline where each piece of evidence (a criterion like PS4, PM3, PP1, etc.) is assigned a category of strength, and each category maps to a calibrated odds of pathogenicity (OP). Tavtigian et al. (2018) estimated these as: when evidence strength (ES) is "supporting", the odds of pathogenicity (OP) is ~2.08, when ES is moderate, OP is ~4.33, when ES is strong, OP is ~18.7 and when ES is very strong, OP is ~350. For benign evidence, the reciprocals apply (e.g., supporting benign ≈ 1/2.08). D enters the formula as a full Bayesian calculation:

$$\text{Post P} = (\text{Prior P} \times \text{OP}_{\text{combined}}) / [(\text{Prior P} \times \text{OP}_{\text{combined}}) + (1 - \text{Prior P})]$$

where $\text{OP}_{\text{combined}}$ is the product of all individual odds: $\text{OP}_{\text{combined}} = \text{OP}_1 \times \text{OP}_2 \times \text{OP}_3 \times \text{OP}_n$

This product is the operationalization of D inside Equation (1). Therefore D is not a single input, it enters the formula as the joint likelihood ratio formed by multiplying all individual evidence odds together, under the assumption that they are independent. The standard prior used is *P(Pathogenic)* = 0.10 for a generic variant of uncertain significance, though this varies by gene context.

Admixture as a Source of Model Misspecification

In admixed genomes, chromosomes consist of alternating ancestry tracts, reflecting recombination between distinct ancestral populations (Gravel, 2012). In Brazil, this process is particularly recent, resulting in short ancestry blocks, high recombination density and complex haplotype mosaics. These mosaics generate novel combinations of variants in *cis*, not present in reference populations. This leads to prior miscalibration.

Allele frequency is a cornerstone of prior estimation. However:

$$P(variant) \neq P(\,variant \mid H, A\,) \quad (2)$$

where *H* denotes haplotype and *A* local ancestry. A haplotype is a combination of alleles at multiple loci that are physically linked and co-inherited on the same chromosome, reflecting local linkage disequilibrium and recombination history (Daly et al., 2001). Failure to condition on these variables leads to systematic bias, particularly in admixed populations (Martin et al., 2019).

Likelihood misspecification is observed because evidence criteria, such as segregation (PP1), case–control enrichment (PS4), and trans configuration (PM3) implicitly assume known LD and independence between variants. However, in admixed genomes LD varies locally and unpredictably, and co-occurrence may reflect shared haplotypes. This leads to distorted likelihood estimates.

If variants in LD are treated as independent the evidence is overcounted and the posterior probability is therefore inflated. This issue is amplified in recombined haplotypes, and constitutes the core mechanism of evidence double counting.

Borderline VUS Instability as a Structural Phenomenon

 doi:10.20944/preprints202604.1833.v1

Variants near classification thresholds are inherently sensitive to small perturbations in priors or likelihoods (Tavtigian et al., 2018). In admixed populations, this sensitivity becomes a structural property of the system. As representative examples three should be mentioned: (1) *CFTR* p.Arg117His, a haplotype-dependent penetrance driven by intronic variation (Thauvin-Robinet et al., 2009); (2) *GJB2* p.Met34Thr, a high allele frequency inconsistent with high penetrance (Shen et al., 2019); (3) *BRCA2* p.Lys3326Ter, where reclassification was driven by prior recalibration (Gaudet et al., 2010). These examples illustrate phase transitions in classification, where small changes in model parameters shift variants across decision thresholds.

Haplotype Classes in Brazilian Populations

The Brazilian genomic landscape can be conceptualized as generating three major haplotype classes (**Figure 1**), each with distinct implications for Bayesian inference. Five known examples of relevant haplotype configurations originated by these admixture patterns (**Table 1**).

Imported risk haplotypes (**Figure 1A**) include the *APOL1 G1*/*G2* haplotypes, which originate in African populations and exhibit ancestry-dependent risk when embedded in admixed genomes (Genovese et al., 2010).

Recombined haplotypes (**Figure 1B**) occur in highly polymorphic regions such as HLA and pharmacogenomic loci such as *CYP3A5*, displaying novel haplotypic combinations due to recombination (Kuehl et al., 2001; Lam et al., 2013). These haplotypes exhibit altered LD patterns, ancestry-dependent expression and non-transferable associations.

Ancestry-mixed functional haplotypes, such as *SLC24A5* locus, exemplify recombination between ancestry-specific alleles, generating new functional contexts (Lamason et al., 2005).

Among Founder haplotypes (**Figure 1C**), the *TP53* p.R337H variant represents a regional founder haplotype with elevated frequency in Southern Brazil (Achatz et al., 2007).

 doi:10.20944/preprints202604.1833.v1

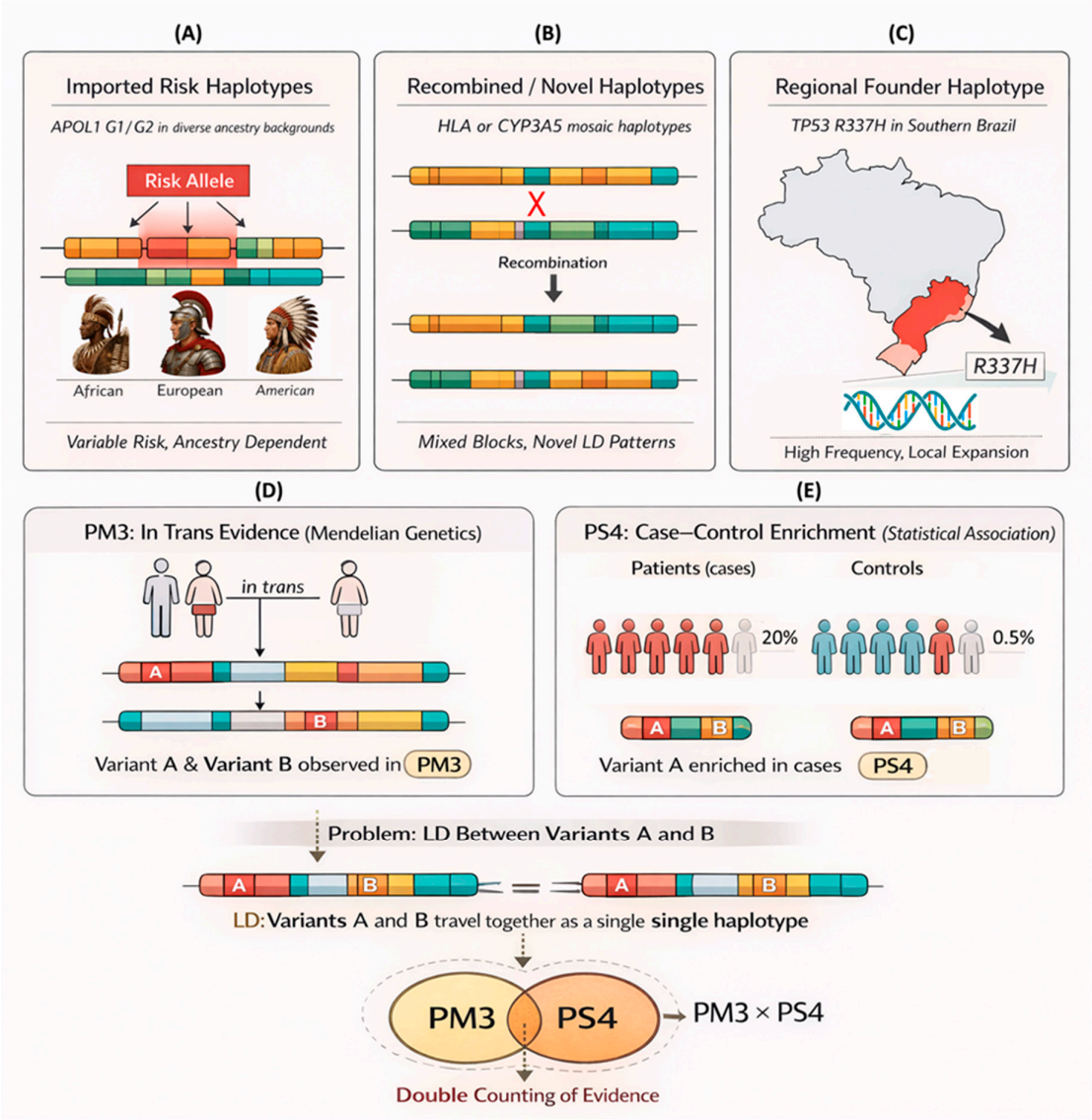

Figure 1. Haplotype classes in admixed populations and their implications for variant interpretation. This figure illustrates three distinct haplotype configurations that arise in admixed populations and their consequences for Bayesian variant interpretation. In (A) (Imported risk haplotypes): Risk alleles originating in specific ancestral populations, such as APOL1 G1 and G2, are introduced into admixed genomes. Although the variant itself is unchanged, its phenotypic effect becomes dependent on the surrounding haplotypic and ancestry context. This leads to variability in penetrance and challenges in estimating prior probabilities of pathogenicity. In (B) Recombined or mosaic haplotypes: Recent admixture generates novel haplotypes through meiotic recombination between chromosomes of distinct ancestry. This process produces mosaic haplotypes composed of alternating ancestry blocks, resulting in local linkage disequilibrium (LD) patterns that differ from those observed in ancestral populations. Consequently, variants that are independent in one population may be in LD in another, leading to misestimation of likelihoods and potential double counting of evidence in Bayesian frameworks. In (C) Regional founder haplotypes: Founder variants, such as TP53 p.R337H in Southern Brazil, arise from a single ancestral mutation that expands within a geographically or culturally defined population. These haplotypes are associated with high local allele frequencies and relatively conserved haplotypic backgrounds, which can strongly influence prior probabilities and lead to population-specific disease risk profiles. In (D) and (E) 1) PM3 "in trans" (basic Mendelian genetics), PM3 is applied when two variants are observed in a recessive gene and they are located on different chromosomes (in trans) one inherited from the father and another inherited from the mother so Chromosome 1 has variant A while Chromosome 2 has variant

 doi:10.20944/preprints202604.1833.v1

B For a recessive disease, two altered copies are required, therefore Variant A plus Variant B in trans can explain the disease phenotype. PM3 summary: The variant contributes to a causal allelic pair, with one variant on each chromosome. PS4 come from case–control enrichment therefore PS4 is based on statistical evidence. This variant is more frequent in affected individuals than in controls. In the example, 20 out of 100 patients carry the variant while in controls 5 out 1000 individuals carry the variant. Therefore, the variant is enriched in cases, and this supports pathogenicity. PS4 summary: The variant occurs more frequently in individuals with the disease than in unaffected controls. The problem arises when both criteria are combined. In this scenario, Variant A (the variant under study) and Variant B (on the same haplotype) are in linkage disequilibrium (LD), in other words, are frequently inherited together. Although Variant A is observed in trans with another pathogenic variant, Variant B may be the true causal allele, and Variant A is simply co-inherited on the same haplotype. In PS4, Variant A is enriched in cases, however, the enrichment may be driven by Variant B and Variant A is merely hitchhiking with it. The error is taking PM3 evidence and PS4 evidence and combining them as independent when, in reality, both may reflect the same underlying haplotype. Linkage disequilibrium can transform apparently independent ACMG/AMP criteria into correlated observations of the same underlying haplotypic signal, leading to double counting of evidence if not properly modeled.

Table 1. Haplotype configurations relevant in Brazilian populations.

Haplotype Class	Gene	Variant / Structure	Origin	Key Property	Bayesian Impact
Imported risk	*APOL1*	G1 / G2	African	Context-dependent risk	Prior distortion
Recombined	*HLA*	Multi-locus haplotypes	Mixed	Novel LD	Likelihood error
Recombined	*CYP3A5*	*1 / *3	African+ European	Expression variability	Prior+ likelihood
Mixed functional	*SLC24A5*	A111T context	European + mixed	Functional recombination	Prior miscalibration
Founder	*TP53*	R337H	Brazilian	Local expansion	Prior shift

Toward a Haplotype-Aware Bayesian Framework

To account for haplotype structure, the classification model should be:

$$P(\mathit{Pathogenic} \mid D, H, A) \tag{3}$$

with key components

Prior conditioning:

$$P(\mathit{Pathogenic} \mid H, A) \tag{4}$$

Likelihood refinement:

$$P(D \mid \mathit{Pathogenic}, H, A) \tag{5}$$

Dependency modeling: explicit incorporation of LD

However, LD is not added as a variable. The factorization of likelihood is changed so independence between pieces of evidence is not assumed.

How LD Breaks the Non-Haplotype Aware Model

 doi:10.20944/preprints202604.1833.v1

In practice, ACMG/Bayesian implementations behave like:

$$P(D \mid Pathogenic) \approx \prod_i P(Di \mid Pathogenic) \quad (6)$$

This assumes independence among evidence components D_i. However, under LD, this is false, because it leads to double counting.

Correct Formulation

Condition on haplotype H:

$$P(Pathogenic \mid D, H) \propto P(D \mid Pathogenic, H)\, P(Pathogenic \mid H) \quad (7)$$

H encodes LD (correlation structure) and evidence is no longer treated as independent. Note that in Equation (7), *A* is not shown explicitly because local ancestry is intrinsic to the haplotype structure encoded in *H*. Once the haplotype is resolved, local ancestry is determined. The full conditioning on *A* is therefore implicit in the conditioning on *H*.

Three Practical Ways to Incorporate LD

Three methods can be used to incorporate LD into Bayesian frameworks for significance estimation of genomic variants and correct for the effect of admixture generated haplotypes: (1) joint likelihood, (2) correlation penalty and (3) hierarchical model. These are detailed below.

Joint Likelihood

Instead of:

$$P(D_1, D_2 \mid Pathogenic) = P(D_1 \mid Pathogenic)P(D_2 \mid Pathogenic) \quad (8)$$

The joint occurrence should be used:

$$P(D_1, D_2 \mid Pathogenic, H) \quad (9)$$

In practice: co-occurrence can be estimated from data or approximated using LD measures (e.g., r^2, D').

Correlation Penalty

If full modeling is not feasible:

$$log\ LR_{total} = \sum_i log\ LR_i - \lambda \cdot corr(D_i, D_j) \quad (10)$$

where LR_{total} is the combined likelihood ratio summarizing the cumulative evidence for pathogenicity, and each LR_i is the individual likelihood ratio contributed by evidence component D_i, that is, the ratio of the probability of observing D_i given pathogenicity versus given benignity. The correction term $\lambda \cdot corr(D_i, D_j)$, with $corr(D_i, D_j) \sim r^2$, penalizes the sum when evidence components are correlated due to linkage disequilibrium, with λ calibrated empirically. This formulation prevents inflation of the posterior probability arising from double counting of correlated evidence.

Hierarchical Model

$$H \sim P(H \mid A)$$
$$D \sim P(D \mid H, \text{Pathogenic})$$

A: local ancestry

H: haplotype

LD emerges naturally through H

Translation to ACMG/AMP Criteria

 doi:10.20944/preprints202604.1833.v1

For example, a scenario with a PM3 (in trans evidence) and PS4 (case–control enrichment) is considered (**Figure 1D,E**). PM3 and PS4 are as defined by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) (Richards et al., 2015) where PM3 means Pathogenic Moderate criterion #3 and PS4 means Pathogenicity Strong criterion #4. In this example PM3 captures allelic configuration in recessive disease, whereas PS4 reflects case–control enrichment. However, when variants are in linkage disequilibrium, both criteria may reflect the same underlying haplotypic signal, leading to double counting of evidence if treated as independent. If variants are in LD they cannot be treated as independent lines of evidence. Either one must be downweighed or they must be modelled jointly. In the presence of linkage disequilibrium, the variables used to estimate the significance of variants cannot be assumed as independent variables, and the likelihood term must be reformulated as a joint distribution conditioned on haplotype structure, i.e., $P(D \mid \text{Pathogenic}, H)$, rather than a product of marginal terms. Therefore, stabilization of VUS classification can be achieved by incorporating haplotypes. Posterior distributions become more robust, threshold crossings decrease and clinical interpretation should stabilize.

The implicit assumption of independence used to calculate $OP_{combined}$ is precisely the vulnerability the present study is targeting. Because $OP_{combined}$ is computed as a simple product of individual terms: $OP_{combined} = OP(PM3) \times OP(PS4) \times OP(PP1) \times ...$, any two criteria that share a common haplotypic signal (i.e., are in LD) will inflate $OP_{combined}$ by being counted twice as independent factors, when in reality they carry the same underlying information. The correct formulation proposed in Equations (7)–(9) above replaces this product with a joint likelihood conditioned on haplotype H, so that correlated evidence items are not treated as multiplicative. This is also why the correlation penalty in Equation (10) subtracts $\lambda \cdot corr(D_i, D_j)$, from the log-likelihood sum, as it works as a practical correction for this inflation.

Conclusions

Current variant interpretation frameworks are formally misspecified for admixed populations. The instability of borderline VUS is a measurable manifestation of this misspecification. We argue that incorporating haplotype structure and local ancestry into Bayesian models is necessary for accuracy, stability, and equity in genomic medicine.

The empirical scale of this challenge is underscored by the recent identification of over 8 million previously unreported variants in Brazilian genomes (Nunes et al., 2025). Variants absent from reference databases do not merely represent gaps in allele frequency tables; they introduce uncharacterized linkage disequilibrium relationships that propagate through every component of the Bayesian framework: distorting priors through miscalibrated frequency estimates, inflating likelihoods through undetected evidence correlation, and rendering independence assumptions untenable in ways that are invisible to current pipelines.

We conjecture that Brazil is not an outlier but a precursor. As global populations become increasingly admixed, the limitations described here will become universal. Ignoring haplotype structure, and the novel LD architectures that admixture continuously generates, risks embedding systematic and inequitable bias into the foundations of genomic medicine.

Author Contributions: MRSB: Writing – original draft, Writing – review and editing. MRSB, RCF, FA: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review and editing.

Funding: The authors declared that financial support was received for this work from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant # 20/08943-5.

Data Availability Statement: This study did not generate datasets or scripts. Additional information can be requested from the corresponding author.

Acknowledgments: The authors declare that generative AI was used in editing, English grammar verification, assistance with image generation and verification of mathematical expressions of this manuscript.

 doi:10.20944/preprints202604.1833.v1

Conflicts of Interest: The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The author MRSB declared that he was an Associate Editor of Frontiers at the time of submission and this had no impact on the peer review process and the final decision.

References

Achatz, M. I. W., Olivier, M., Calvez, F. L., Martel-Planche, G., Lopes, A., Rossi, B. M., et al. (2007). The *TP53* mutation, R337H, is associated with Li-Fraumeni and Li-Fraumeni-like syndromes in Brazilian families. *Cancer Letters* 245, 96–102. doi: 10.1016/j.canlet.2005.12.039

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat Genet* 29, 229–232. doi: 10.1038/ng1001-229

Gaudet, M. M., Kirchhoff, T., Green, T., Vijai, J., Korn, J. M., Guiducci, C., et al. (2010). Common Genetic Variants and Modification of Penetrance of BRCA2-Associated Breast Cancer. *PLOS Genetics* 6, e1001183. doi: 10.1371/journal.pgen.1001183

Genovese, G., Friedman, D. J., Ross, M. D., Lecordier, L., Uzureau, P., Freedman, B. I., et al. (2010). Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. *Science* 329, 841–845. doi: 10.1126/science.1193032

Gravel, S. (2012). Population Genetics Models of Local Ancestry. *Genetics* 191, 607–619. doi: 10.1534/genetics.112.139808

Kuehl, P., Zhang, J., Lin, Y., Lamba, J., Assem, M., Schuetz, J., et al. (2001). Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nat Genet* 27, 383–391. doi: 10.1038/86882

Lam, T. H., Shen, M., Chia, J.-M., Chan, S. H., and Ren, E. C. (2013). Population-specific recombination sites within the human MHC region. *Heredity* 111, 131–138. doi: 10.1038/hdy.2013.27

Lamason, R. L., Mohideen, M.-A. P. K., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., et al. (2005). SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. *Science* 310, 1782–1786. doi: 10.1126/science.1116238

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 51, 584–591. doi: 10.1038/s41588-019-0379-x

Nunes, K., Araújo Castro e Silva, M., Rodrigues, M. R., Lemes, R. B., Pezo-Valderrama, P., Kimura, L., et al. (2025). Admixture's impact on Brazilian population evolution and health. *Science* 388, eadl3564. doi: 10.1126/science.adl3564

Pena, S. D. J., Pietro, G. D., Fuchshuber-Moraes, M., Genro, J. P., Hutz, M. H., Kehdy, F. de S. G., et al. (2011). The Genomic Ancestry of Individuals from Different Geographical Regions of Brazil Is More Uniform Than Expected. *PLOS ONE* 6, e17063. doi: 10.1371/journal.pone.0017063

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009). Reconstructing Indian population history. *Nature* 461, 489–494. doi: 10.1038/nature08365

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). 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, 405–423. doi: 10.1038/gim.2015.30

Shen, J., Oza, A. M., Castillo, I. del, Duzkale, H., Matsunaga, T., Pandya, A., et al. (2019). Consensus interpretation of the p.Met34Thr and p.Val37Ile variants in GJB2 by the ClinGen Hearing Loss Expert Panel. *Genetics in Medicine* 21, 2442–2452. doi: 10.1038/s41436-019-0535-9

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell* 177, 26–31. doi: 10.1016/j.cell.2019.02.048

 doi:10.20944/preprints202604.1833.v1

Tavtigian, S. V., Greenblatt, M. S., Harrison, S. M., Nussbaum, R. L., Prabhu, S. A., Boucher, K. M., et al. (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine* 20, 1054–1060. doi: 10.1038/gim.2017.210

Thauvin-Robinet, C., Munck, A., Huet, F., Génin, E., Bellis, G., Gautier, E., et al. (2009). The very low penetrance of cystic fibrosis for the R117H mutation: a reappraisal for genetic counselling and newborn screening. *Journal of Medical Genetics* 46, 752–758. doi: 10.1136/jmg.2009.067215