Missing genetic diversity impacts variant prioritisation for rare disorders

Sam Tallman[1]*, Loukas Moutsianas[1], Thuy Nguyen[1], Yoonsu Cho[1,2], Maxine Mackintosh[1,3], Dalia Kasperaviciute[1], Matthew A Brown[1,4], Jamie Ellingford[1,5], Karoline Kuchenbaecker[1,6]* and Matt J Silver[1,7,8]*

1. Genomics England, UK
2. Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK
3. Alan Turing Institute, UK
4. Department of Medical and Molecular Genetics, King's College London, London, UK
5. Division of Evolution, Infection and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK
6. Division of Psychiatry, University College London, UK
7. Department of Population Health, London School of Hygiene & Tropical Medicine, UK
8. Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, The Gambia

* Corresponding authors

Abstract

Whole genome sequencing identifies millions of genetic variants per individual. When applied to rare disease diagnosis, potentially pathogenic variants are prioritised for clinical interpretation, a process that may be influenced by an individual's genetic ancestry. We analysed millions of rare protein-altering variants prioritised in 29,425 participants with rare disease from the UK 100,000 Genomes Project. We observed disparities in the number of variants prioritised across genetic ancestry groups, with an up to 3-fold increase in participants with African compared to European ancestries. Variants prioritised in participants with non-European ancestries were less likely to be assessed as pathogenic. Leveraging a cohort of 34,701 diverse genomes from the UK, we identified thousands of candidate variants that were ultra-rare or unobserved across populations in gnomAD but common among ancestry-matched individuals. Our findings highlight the importance of using reference databases that reflect patient genetic diversity when prioritising variants for rare disease diagnosis.

Main

Over 400 million people worldwide are estimated to be living with a rare disease[1]. While it is thought that more than 80% of rare diseases have a genetic component, most patients do not receive a genetic diagnosis after diagnostic testing[2]. By capturing the majority of genetic variation in an individual, whole genome sequencing (WGS) has improved genetic diagnosis rates[2,3]. However, separating causative pathogenic variants from the millions of benign variants present in the genome remains a key challenge[4].

The rare diseases arm of the 100,000 Genomes Project (100kGP) is the largest study of sequenced rare disease probands and family members to date[5]. Candidate variants derived from WGS and prioritised through automated pipelines as part of the 100kGP have led to the discovery of thousands of novel pathogenic variants in hundreds of previously undiagnosed participants[2]. The 100kGP has provided the foundation for the introduction of WGS testing into the UK National Health Service for patients with rare disease as part of routine clinical care[6].

With some exceptions[7,8], penetrant pathogenic variants causing monogenic diseases or traits are expected to be rare in all human populations[9]. Reference databases such as the Genome Aggregation Database (gnomAD; https://gnomad.broadinstitute.org)[10,11] which aggregate WGS and whole-exome sequencing (WES) data from hundreds of thousands of individuals are therefore used to provide allele frequency estimates to assist in the prioritisation of candidate variants underlying rare diseases[4,12]. Despite these endeavours, a large fraction of global genetic diversity has yet to be surveyed and there continues to be a European bias in genomics[13,14]. Furthermore, reference databases are often stratified into broad, continental ancestry groups (e.g. *Africa*), homogenising the genetic structure that exists within continental regions and between their diaspora communities[15]. As routine genomic sequencing for rare disease diagnosis is incorporated into healthcare systems across the world[6,16,17], it is vital to understand the impact that this uneven representation may have on identifying pathogenic variation in individuals and families.

Here, using data from the 100kGP, we investigate the potential influence of an individual's genetic ancestry on the number of candidate protein-altering variants prioritised for clinical review, and assess the role of ancestry-related ascertainment biases in reference databases on the variant prioritisation process. We also investigate the relationship between a proband's genetic ancestry and their likelihood of receiving a genetic diagnosis with a prioritised variant.

Cohort overview

We analysed WGS data from a cohort of 61,512 individuals (29,425 probands and 32,087 family members) recruited to the 100kGP[1] (Supplementary Information 1-5). Disorders covered a broad spectrum of 112 rare diseases and probands had no genetic diagnosis at the time of recruitment (Table 1).

Table 1. *Cohort characteristics*.

Family structure	
Singleton	11,793
Duo†	4,088
Trio†	9,839
Other	3,705
Total	29,425
Proband karyotype	
XY	15,198
XX	14,227
Median (IQR) age of proband at recruitment (Years)	26 (42)
Median (IQR) autosomal coverage (x)	40.0 (8.4)
Median (IQR) number of small variants	4,898,153 (66,199)
Cases fully or partially solved¶	5,793

† Refers to complete parent-offspring duos or trios.
‡ Multi- and single nucleotide variants and indels <50bp
¶ As indicated by clinical teams in the Genomics England v18 data release (21st December 2023)
IQR = interquartile range.

Representation of self-reported ethnic groups among probands in the 100kGP was broadly similar to that reported for England in the 2021 Office for National Statistics (ONS) Census[18] given the age profile of the cohort (Supplementary Figure 1). 75% of probands whose ethnic group was reported and known self-reported as White British (Methods).

Genetically inferred ancestries and population structure within the 100kGP

Classification of individuals into discrete populations is an inadequate description of human genetic diversity, which is continuous and varies throughout the genome[19]. However, given the reliance of rare variant prioritisation for rare disease diagnosis on population allele frequencies, we first organised 100kGP probands on the basis of their genetic similarity to a set of sub-continental reference population groups previously curated using the UK Biobank[20,21], a reference dataset that reflects the diversity of the UK (Methods; Extended Data 1). We refer to these as genetically inferred ancestry (GIA) groups (Figure 1).

GIA groups aligned with clustering patterns in a low-dimensional topological map (UMAP) generated from the first 16 PCs (Supplementary Figures 2 & 3). 1,712 probands (5.8% of the total) mapping to multiple reference populations were not assigned to any one GIA group (described as '*remaining participants';* Extended Data 1) (Supplementary Figure 4). GIA groups and ancestry coefficient estimates[21] (Methods) broadly corresponded with probands' self-reported ethnicity (Supplementary Figures 5 and 6) and with ancestry group classifications

predicted using reference populations provided by gnomADv3.1[10] (Supplementary Figure 7 and 8), often with added granularity.

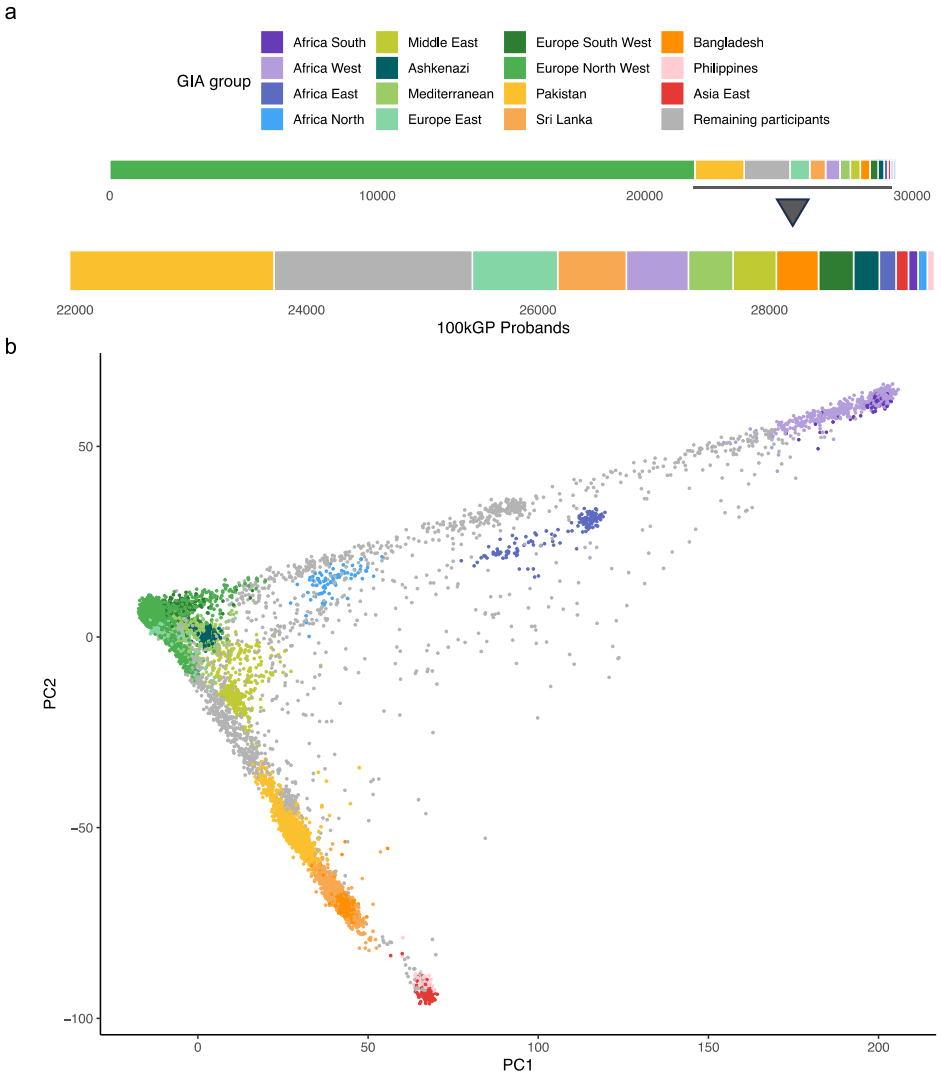

Figure 1. *Genetic structure and GIA groups within the 100kGP* **a:** The numbers of 100kGP rare disease programme probands assigned to sub-continental GIA groups curated using the UK Biobank[21] (Extended Data 1). **b:** PCA of 100kGP probands coloured by assigned GIA. The top 2 genotype principal components are shown. PCs were calculated using all unrelated probands and family members (<3rd degree, Supplementary Information 4) across a genotype matrix of 60,878 high-quality SNVs used to perform the PCA (Methods; Supplementary Information 5). PCs 1-16 are shown in Supplementary Figure 2. *Remaining participants* include those without an assigned GIA or those assigned to groups with less than 50 probands in the 100kGP (Extended Data 1).

Candidate protein-altering variants (cPAVs)

Candidate small variants predicted to be protein-altering (candidate protein-altering variants or cPAVs) potentially linked to rare Mendelian disease were triaged using Genomics England's bioinformatics pipelines (Methods; Supplementary Information 7, Supplementary Figure 9). Briefly, these pipelines applied variant filters based on population allele frequencies, predicted variant consequence (Supplementary Table 1), and (where family data was available) co-segregation with disease. In total, we identified 1,951,659 cPAVs, 30.3% of which were identified in more than one proband (Supplementary Figure 10). 1,865,089 cPAVs (95.6% of the total) were either absent or ultra-rare in every annotated population across all queried reference databases (population maximum allele frequency (popmax AF) <0.1%) including gnomADv2[11], gnomADv3[10], and UK10K[22] (Supplementary Table 2). The remaining 86,570 uncommon cPAVs (popmax AF >0.1%; 4.4% of the total) passed a more lenient frequency threshold (popmax AF <1%) applied exclusively to variants with biallelic modes of inheritance.

An average of 172 (standard deviation (SD): 120, minimum-maximum (min-max): 0-1013) cPAVs were identified per proband (Supplementary Figure 10), with differences predominantly driven by the availability of WGS data from family members and the clinically indicated penetrance mode (Supplementary Figure 11). Probands analysed as part of full parent-offspring trios (mean (SD, min-max) = 23 (21,1-422)) or larger family group types (25 (40,0-404)) under the assumption of complete penetrance had the fewest cPAVs identified on average. Singleton probands (273 (69, 0-1,013)) or probands for whom co-segregation pattern filters were bypassed to account for incomplete penetrance (216 (74,16-760)) had the greatest number of cPAVs on average.

Adjusting for family group type, penetrance mode, sequencing quality metrics, and other covariates predicted to impact variant calling and triaging (Methods; Supplementary Figure 12, Supplementary Table 3) we observed a strong association between GIA and the number of cPAVs identified in the proband (ANOVA Type II $p < 2.20 \times 10^{-308}$). 13 of 14 GIA groups had significantly greater number of cPAVs when compared to the Europe North West group, the largest group in the cohort (Figure 2; Supplementary Table 4a). The Africa East GIA group had the greatest number of cPAVs (Rate ratio (RR) vs Europe North West = 2.93, $p < 2.20 \times 10^{-305}$). Only the Ashkenazi GIA group had significantly fewer cPAVs (RR vs Europe North West = 0.73, $p = 4.40 \times 10^{-51}$).

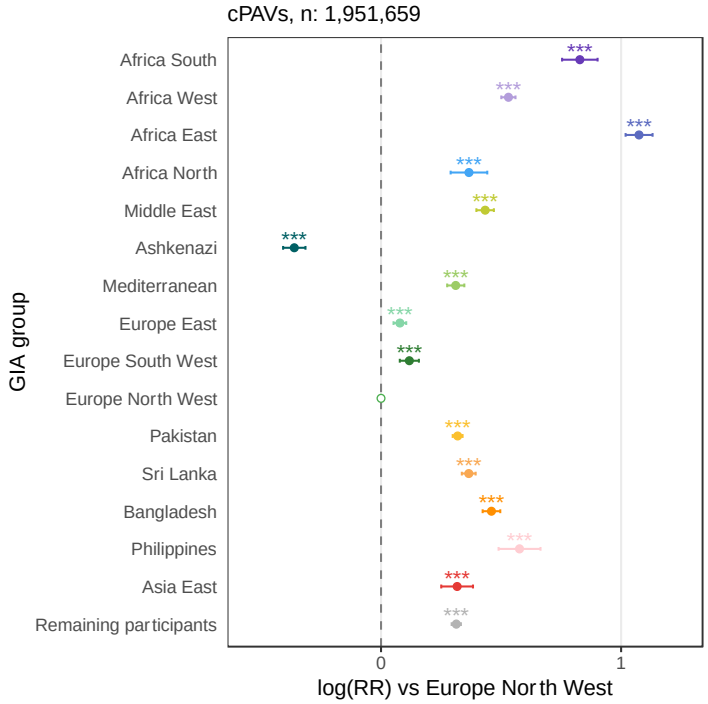

Figure 2. *GIA predicts the number of cPAVs identified in the proband.* Multivariable negative binomial regression model coefficients showing the association between each GIA group and the number of cPAVs identified in the proband vs 21,872 probands assigned to the Europe North West GIA group, adjusted for covariates (Supplementary Tables 4 and 5a). Error bars show 95% confidence intervals. RR = Rate ratio. * 0.05< p-value < 0.01. ** 0.001 <p-value < 0.01. *** p-value < 0.001 after Bonferroni adjustment.

Protein-altering variation missing from gnomAD

gnomAD (v2.1 and v3.1) was the main source of population allele frequencies used to identify cPAVs in the 100kGP (Supplementary Table 2). To explore whether the observed disparity in the numbers of cPAVs across GIA groups may be linked to ancestry-related ascertainment bias in gnomAD, we calculated the total number of small, predicted protein-altering variants (Supplementary Table 1) called in each proband that were missing from gnomADv2.1 (exomes) and gnomADv3.1 (whole genomes). We considered all missing protein-altering variants, irrespective of whether they were triaged as cPAVs, so that ascertainment bias could be assessed independent of any potential influence from the triaging process.

Protein-altering variation in European GIA groups appeared to be well captured by gnomAD. For example, the Europe North West group had fewer missing protein-altering variants per proband (mean (SD, min-max) = 46 (7, 22-102)) than any other GIA group apart from Ashkenazi (33 (6, 21-56)) (Figure 3a). The latter aligns with the recent genetic bottleneck in the demographic history of the Ashkenazi population[23] (Supplementary Figure 13; Supplementary Information 9), resulting in a relative paucity of genetic variation in this group not already captured by the Ashkenazi reference populations annotated in gnomAD.

In marked contrast, the Africa East GIA group had an average of 161 (17, 123-216) protein-altering variants missing from gnomAD per proband, whilst the Africa South GIA group had the greatest intra-group variability in missing protein-altering variants (111 (32, 66-227). These observations are consistent with Africa's immense genetic diversity[24], as well as the over-representation of individuals with ancestries from West Africa (e.g. African Americans) in genomic reference databases relative to other sub-continental regions[15].

Overall, we found that average differences in the number of predicted protein-altering variants missing from gnomAD were highly correlated with rate ratios for the relative increase in the number of triaged cPAVs across GIA groups (Pearson's product of moment (Pr) = 0.98, $p = 1.09 \times 10^{-10}$; Figure 3b). This positive correlation was also observed at the individual level ($p < 2.20 \times 10^{-308}$; Supplementary Table 4b), inclusive of probands without an assigned GIA group. This relationship between the number of triaged cPAVs and predicted protein-altering variants missing from gnomAD was apparent irrespective of family group type, despite the latter's marked effect on cPAV numbers (Figure 3c-e).

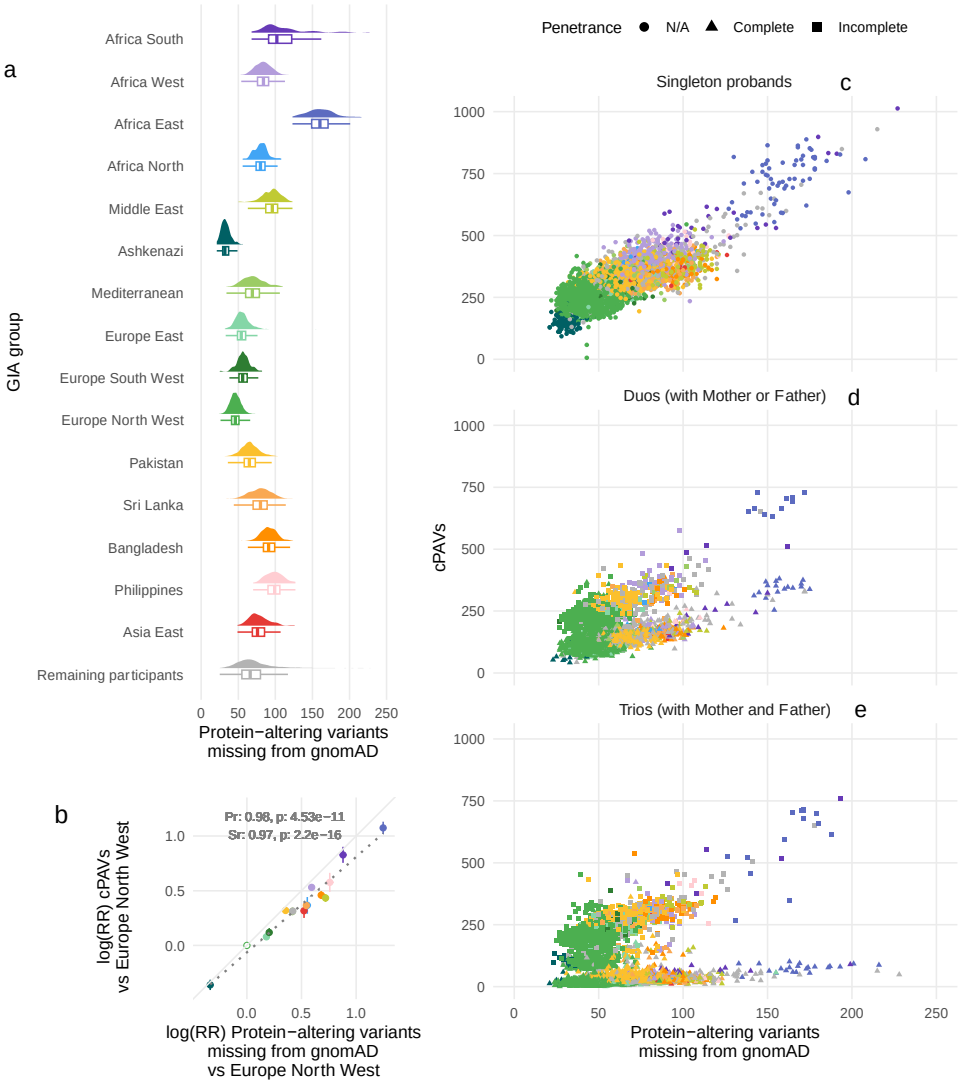

Figure 3. *Protein-altering variation missing from gnomAD and cPAVs across the ancestry continuum*. **a:** Raincloud plot showing the numbers of protein-altering variants missing from gnomADv2.1 (exomes) and gnomADv3.1 (genomes) per proband in the 100kGP, stratified by GIA group. **b:** Rate ratios describing the number of small, predicted protein-altering variants missing from gnomAD (vs Europe North West), plotted against rate ratios from Figure 2 describing the number of cPAVs (also vs Europe North West) identified in the proband. See main text and Methods for details on regression models. Error bars show 95% confidence intervals. RR = Rate ratio. Pr = Pearson's product of moment. Sr = Spearman's rank. Colour-filled points show significant ($p < 0.05$) differences vs Europe North West after Bonferroni adjustment. **c-e:** Number of cPAVs plotted against the number of protein-altering variants missing from gnomAD per proband, coloured by proband GIA, stratified into **c:** Singleton probands **d**: Parent-offspring duos **e**: full parent-offspring trio. Point shapes are determined by the indicated penetrance mode, relevant only when familial co-segregation data is available.

Predicted deleteriousness of ultra-rare cPAVs

Variants with deleterious effects are more likely to be rare or unobserved across all populations due to purifying selection[25]. Conversely, common variants are more likely to be tolerated. We hypothesised that a greater number of common variants may have been misclassified as ultra-rare cPAVs in probands with genetic ancestries that are under-represented across queried reference databases (Supplementary Table 2), resulting in an excess of triaged cPAVs that are predicted to be non-deleterious in these individuals. To test this, we annotated cPAVs using CADD[26] and two missense effect prediction tools: AlphaMissense[27] and PrimateAI-3D[28]. We excluded common (popmax AF >0.1%) cPAVs triaged under biallelic modes of inheritance (Supplementary Figure 10b) and tested the association between GIA and the proportion of cPAVs estimated to be deleterious, adjusted for covariates (Methods; Supplementary Table 3). All three tools predicted that ultra-rare cPAVs identified in European GIA groups contained a significantly higher proportion of deleterious variants, in line with our hypothesis (Figure 4; Supplementary Figure 14, Supplementary Table 5a-c). Odds ratios measuring the association between GIA and the proportion of deleterious, ultra-rare cPAVs showed strong negative correlations with rate ratios measuring the association between GIA and the number of ultra-rare cPAVs identified in the proband (Figure 4d-f). Africa South and Africa East GIA groups were consistently estimated as having the lowest proportion of deleterious cPAVs (Figure 4).

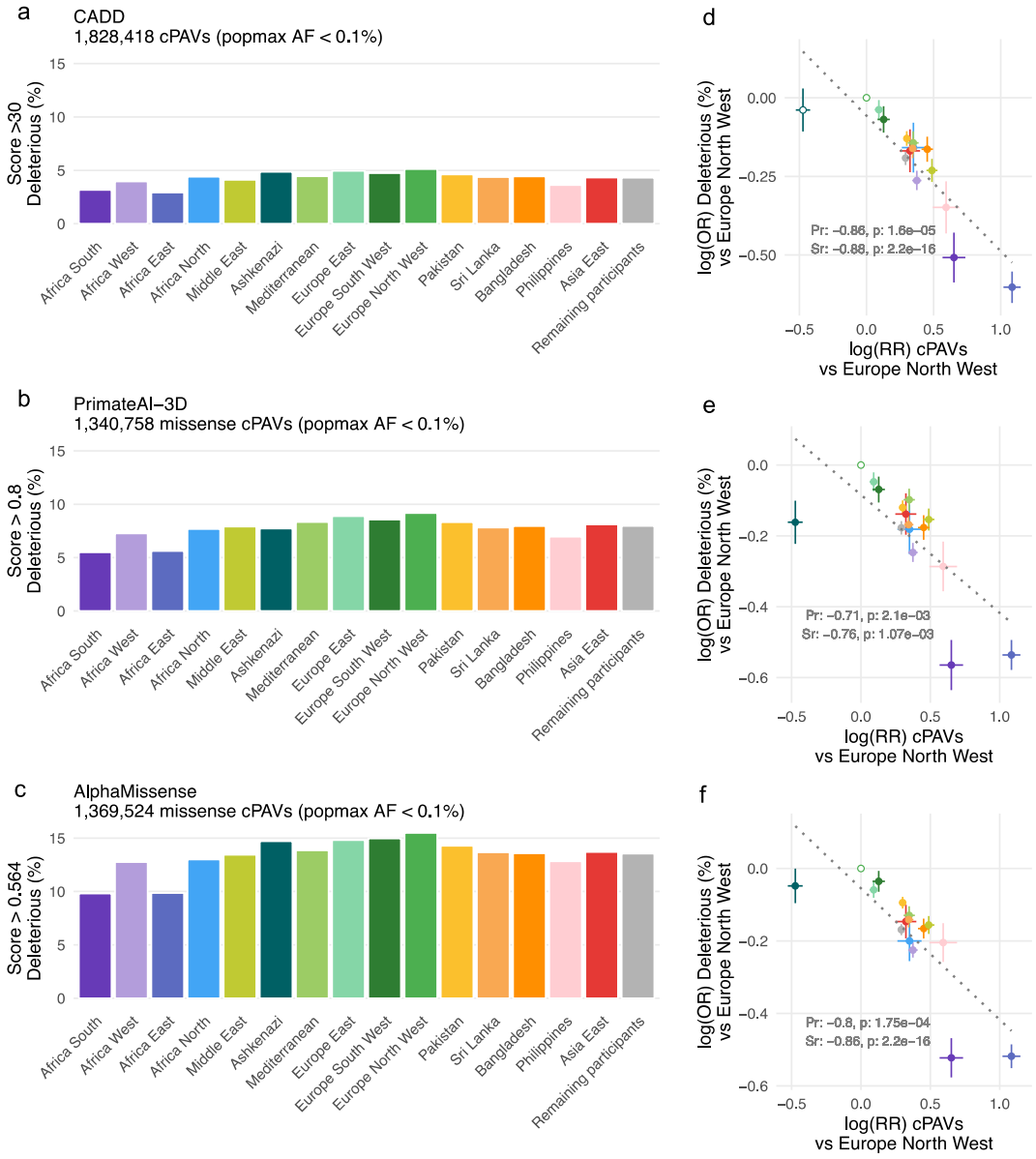

Figure 4. *Predicted deleteriousness of ultra-rare cPAVs by GIA group.* **a-c:** Percentage of ultra-rare cPAVs across GIA groups annotated as deleterious using **a**: CADD (n = 1,822,798); **b**: PrimateAI-3D (n = 1,337,158); and **c**: AlphaMissense (n = 1,365,877). Method-specific default score cut-offs are shown in y-axis labels. **d-f:** Rate ratios (vs Europe North West) describing the number of ultra-rare cPAVs, plotted against odds ratios (also vs Europe North West) describing the proportion of annotated ultra-rare cPAVs predicted to be deleterious by **d:** CADD; **e**: PrimateAI-3D; and **f**: AlphaMissense (Supplementary Tables 5 and 6). cPAVs exclude common variants (popmax AF > 0.1%) triaged under biallelic modes of inheritance. Error bars show 95% confidence intervals. RR = Rate ratio. OR = Odds ratio. Pr = Pearson's product of moment. Sr = Spearman's rank. Colour-filled points show significant ($p < 0.05$) associations vs the Europe North West reference GIA group after Bonferroni adjustment.

cPAVs within applied virtual gene panels

After application of the triaging process used to identify cPAVs in the 100kGP Methods; Supplementary Figure 9), an additional step involves the identification of a subset of cPAVs that appear within genes contained in expert-assessed PanelApp[29] virtual gene panels associated with the proband's condition (Supplementary Information 7). These gene-panel candidate protein-altering variants (gene-panel cPAVs) undergo clinical interpretation with the highest priority. In total, we identified 51,838 unique gene-panel cPAVs using the most recent PanelApp gene panel versions (Extended Data 2). An average of 1.46 (2.37, 0-38; SD, min-max) gene-panel cPAVs were identified per proband.

Despite representing only 2.7% of the total cPAVs in the 100kGP, observed associations between GIA and both the number (Figure 2) and the proportion of deleterious cPAVs identified in the proband (Figure 4) remained when evaluating gene-panel cPAVs alone (Supplementary Figures 15-17; Supplementary Tables 5c-d and 6d-f).

Identifying cPAVs classified as ultra-rare that are common in a diverse reference database from the UK population

To investigate whether a relatively small number of diverse genomes from a GIA-matched control cohort could reveal under-represented common variation classified as ultra-rare cPAVs in the 100kGP, we leveraged WGS data from 34,701 individuals recruited into the UK COVID-19 genomics study[30] (Supplementary Information 2). Individuals in the COVID-19 cohort were organised into GIA groups using the same method as described above for the 100kGP. We found analogous patterns of predicted protein-altering variation missing from gnomAD across GIA groups in the COVID-19 cohort as previously observed in the 100kGP (Supplementary Figure 18). This indicates that the COVID-19 cohort captures genetic diversity that is under-represented in reference databases, such as gnomAD, and could therefore be informative for variant prioritisation and interpretation.

For all 14 GIA groups with n>100 unrelated (<3rd degree) individuals in the COVID-19 cohort (Figure 5a), we calculated allele frequencies for ~500 million small variants across the autosomes and the X-chromosome. We also calculated 95% confidence maximum filtering allele frequency thresholds (FAF95; Methods) to account for the reduced precision and upward-bias with which AFs are estimated in groups with smaller sample sizes (Figure 5b)[31].

In total, 986,893 variants identified as cPAVs in the 100kGP (50.6% of all cPAVs) were observed at least once in the COVID-19 cohort. Of these, we identified 25,420 cPAVs that were ultra-rare across all previously queried reference databases (popmax AF <0.1%; Supplementary Table 2) yet appeared at FAF95 >0.1% in at least one GIA group in the COVID-19 cohort. 1,941 such cPAVs were missing from all previously queried reference databases, 66.0% (n = 1,282 of which were observed at FAF95 >0.1% amongst individuals in the COVID-19 cohort Africa East GIA group (Supplementary Figure 19).

Of note, we identified 2,046 cPAVs previously classified as ultra-rare (popmax AF < 0.1%; Supplementary Table 2) that appeared at least an order of magnitude more frequently (FAF95 >1%) in at least one GIA group in the COVID-19 cohort (Figure 6c), meaning that they are

unlikely to be the cause of a rare genetic disease. 62.7% (n = 1,282) of these common cPAVs appeared at >1% FAF95 in the Africa East GIA group, 34.2% (n = 438) of which were unobserved in any other group. Indeed, overall, 11.7% (n = 8,336) of all cPAVs identified in the 100kGP Africa East GIA group were observed at FAF95 >1% among GIA-matched individuals in the COVID-19 cohort; an average of 58 (SD, min-max; 41,1-159) variants per proband.

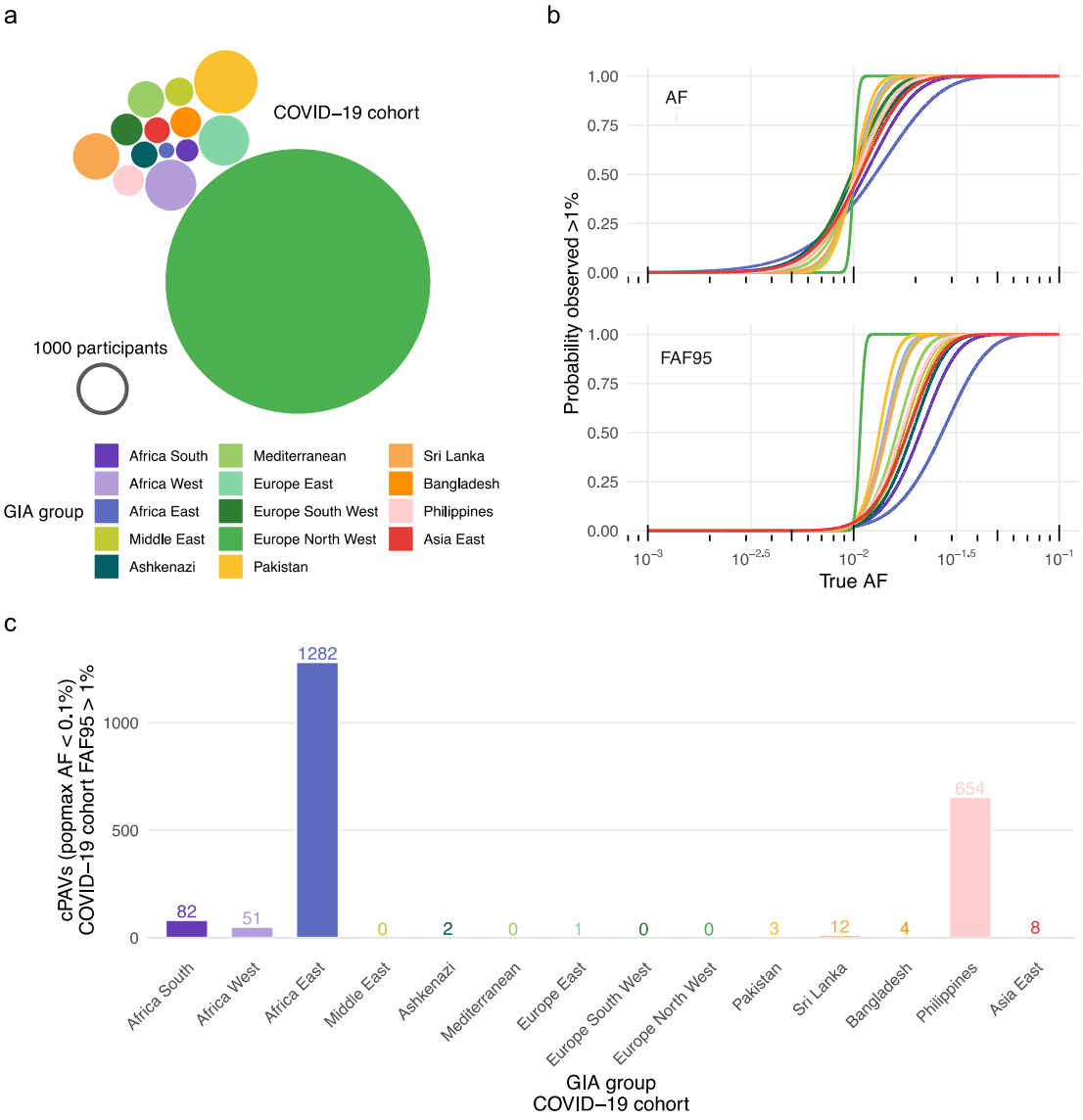

Figure 5. *cPAVs classified as ultra-rare that are common in a diverse reference database from the UK population*. **a:** Circle plot with circle sizes representing the number of individuals from each GIA group in the COVID-19 cohort. Only GIA groups with n >100 individuals genetically unrelated to the 3rd degree (both to others in the COVID-19 cohort and all probands in the 100kGP) are shown. **b:** For each GIA group in the COVID-19 cohort, the probability of observing a variant of a given allele frequency (true AF) at either *(top)* AF >1% or *(bottom)* FAF95 >1% given the GIA group sample size (in number of haplotypes), where FAF95 is the maximum credible GIA group AF (lower bound of the 95% CI). We assumed true allele frequencies in each group followed Hardy-Weinberg equilibrium and calculated probabilities using the binomial distribution. **c:** Number of cPAVs identified in the 100kGP and observed at popmax AF<0.1% across all previously queried reference databases (Supplementary Table 2) observed at FAF95 >1% across GIA groups in the COVID-19 cohort.

Of the 111 ultra-rare gene-panel cPAVs that appeared at FAF95 >1% in at least one GIA group in the COVID-19 cohort (Extended Data 3), none were clinically assessed as pathogenic or likely pathogenic (P/LP) with all either assessed as variants of uncertain significance (VUS) or remaining unclassified. As an example, a heterozygous splice site donor variant (rs1456162375) in the 5'UTR region of the nuclear-encoded mitochondrial mitofusin 2 (*MFN2*) gene was triaged as a monoallelic gene-panel cPAV in a proband with Charcot-Marie Tooth Disease assigned to the Africa East GIA group and was assessed as a VUS. This variant is missing from both gnomADv2.1 (exomes), gnomADv3.1 (genomes) and the COVID-19 cohort except among the Africa East GIA group where it was observed at an allele frequency of >3% (FAF95: 1.5%).

Clinical assessment of gene panel cPAVs

Gene-panel cPAVs have been a major source of novel genetic diagnoses in the 100kGP. Among the 29,425 probands analysed in this study, they represent 75.7% (n = 5,415) of the total disease-associated variants assessed as P/LP using criteria from the American College of Medical Genetics and Genomics and Association of Molecular Pathology (ACMG/AMP)[32,33] (Methods). A further 263 cPAVs were assessed as P/LP but did not appear within applied virtual gene panels.

A principal aim of the 100kGP variant triaging process is to maximise the number of P/LP variants while minimising the number of variants requiring clinical assessment[2]. We explored the influence of the proband's ancestry on this process by estimating the positive predictive value (PPV) of gene-panel cPAVs, defined as the proportion of gene-panel cPAVs assessed as P/LP. After adjusting for covariates (Methods; Supplementary Table 3), we found that gene-panel cPAVs identified in 9 out of 11 non-European GIA groups had a significantly lower PPV than those identified in the Europe North West group (Figure 6, Supplementary Table 6). Odds ratios measuring the association between GIA and the PPV of gene-panel cPAVs were negatively correlated with rate ratios measuring the association between GIA and number of gene-panel cPAVs (Figure 5b).

We next used multivariable logistic regression to examine the association between GIA and the likelihood of receiving a diagnosis with a gene-panel cPAV, with diagnosis defined as a case that was solved or partially solved through the discovery of at least one P/LP variant in the proband (Methods). The full model explained 18.7% of the variance (Nagelkerke's R^2), with the majority (14.5%) attributed to the proband's disease phenotype. A range of additional factors were associated with the likelihood of receiving a diagnosis with a gene-panel cPAV, including the proband's sex, family group type, and whether family members were affected by the disease (Supplementary Figure 20). However, we found no significant association between the proband's GIA and their likelihood of receiving a diagnosis with a gene-panel cPAV (ANOVA Type II p = 0.21; Supplementary Figure 21; Supplementary Table 7a). Whilst these sub-continental GIA groupings enable more granular comparisons, they may limit power to detect broader differences due to small group sample sizes and greater multiple testing burden. We therefore repeated our analysis, stratifying probands using commonly-used, continental reference populations provided by gnomADv3[10] ('continental GIA group'; Methods; Figure

6c) and restricting our comparison to groups with greater than 150 individuals (Non-Finnish European (*nfe*) = 23,838, South Asian (*sas*) = 2,925, African (*afr*) = 1,197, East Asian (eas) = 198). Again, we found no significant association between a proband's continental GIA group and their likelihood of receiving a diagnosis with a gene-panel cPAV (Figure 6d; Supplementary Table 7b).

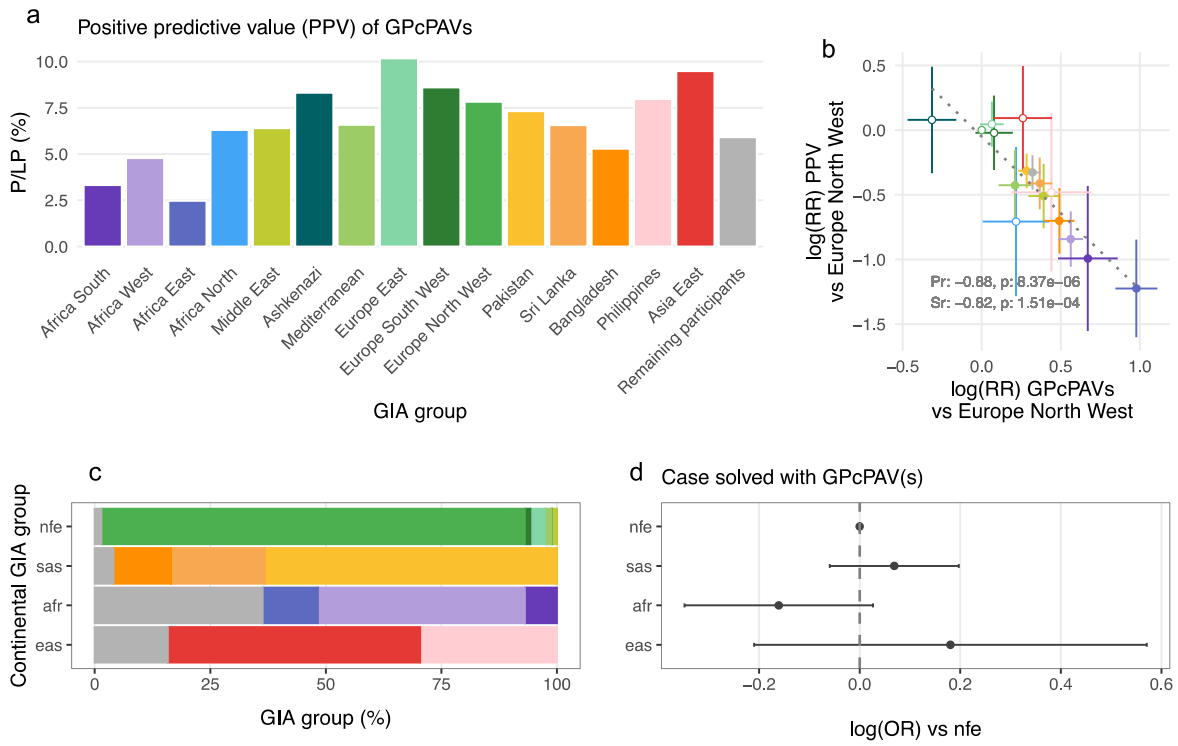

Figure 6. *Clinical assessment of gene-panel cPAVs (GPcPAVs).* **a:** Summarised differences in the proportion of gene-panel cPAVs assessed as P/LP (PPV) across GIA groups. **b:** Rate ratios describing differences in the number of gene-panel cPAVs (x), plotted against the odds ratios describing differences in the PPV of gene-panel cPAVs (y), both compared to Europe North West Group. Filled points show significant ($p < 0.05$) differences vs the Europe North West reference group after Bonferroni adjustment. **c**: The proportion of probands in each GIA group (see x-labels in **a**) by assigned continental GIA group **d**: Model coefficients showing the association between a proband's assigned continental GIA group and the likelihood of a case being solved using at least one gene-panel cPAV. ORs are compared to the reference group (nfe). See Supplementary Tables 4, 9 & 10 for covariates and regression model outputs. Error bars show 95% confidence intervals. P/LP = Pathogenic or Likely Pathogenic RR = Rate Ratio. OR = Odds Ratio. PPV = positive predictive value. Pr = Pearson's product of moment. Sr = Spearman's rank. sas = South Asian, afr = African, eas = East Asian, nfe = Non-Finnish European.

Overall, 90.1% (n = 47,021) of all gene-panel cPAVs were unclassified or assessed as VUS in at least one proband (Supplementary Information 8), including 98.6% (n = 36,672) of gene-panel cPAVs identified in undiagnosed probands. Adjusting for covariates (Methods; Supplementary Table 3), we observed that 10 out of 12 non-European GIA groups had a significant excess of gene-panel cPAVs that were unclassified or assessed as VUS relative to the Europe North West group across both diagnosed and undiagnosed cohorts (Supplementary Figure 22; Supplementary Table 8).

Discussion

With the increasing use of genomic sequencing for the clinical care of rare disease patients, it is vital to understand the effect of genetic ancestry on variant prioritisation for clinical assessment. Here, we investigated rare variant prioritisation in the 100kGP, a unique resource comprising WGS of rare disease probands and family members drawn from the UK population. We found significant differences in both the number and predicted deleteriousness of rare candidate protein-altering variants prioritised for clinical assessment between genetically inferred ancestry groups. Probands with inferred non-European ancestries had up to 3 times more candidate variants within disease-associated gene panels compared to those with European ancestries. By estimating allele frequencies using ancestry-matched controls from an independent UK cohort, we identified thousands of common (FAF95 >1%) candidate variants that were ultra-rare (AF < 0.1%) or unobserved in all reference populations used for variant prioritisation. Candidate variants identified in probands with inferred non-European ancestries were less likely to be clinically assessed as P/LP, with the majority remaining unclassified. Finally, we found no evidence that a proband's genetically inferred ancestry was associated with their likelihood of receiving a genetic diagnosis from candidate variants.

Determination of a variant's frequency in large population reference databases is an essential step in prioritising and interpreting candidate pathogenic variants for rare disease diagnosis[12]. Over time, complex demographic and adaptive histories have resulted in variability in allele frequencies across globally diverse populations[34]. As a result, variants that are common among individuals from ancestral backgrounds that are under-represented in allele frequency resources can appear superficially rare, increasing the likelihood that they are prioritised for clinical assessment[35,36]. Without allele frequency evidence from a large number of ancestry-matched controls available to assist in interpretation of pathogenicity, there is also an increased risk that candidate variants are assessed as VUS[12]. In contexts where VUS are clinically assessed, this can be both time consuming and expensive, and may result in increased uncertainty for clinical teams when considering appropriate treatment or care options[37]. Our finding that higher numbers of superficially rare candidate variants were prioritised in probands with inferred non-European ancestries and were either unclassified or assessed as VUS is therefore a matter of concern.

Difficulties in the ability to pinpoint benign or pathogenic variants in patients from diverse ancestral backgrounds have been widely reported, and have resulted in recommendations that

a patient's genetic ancestry should be a key point of consideration during clinical variant assessment[36,38]. Our findings reiterate the importance of considering genetic ancestry throughout the variant prioritisation process for rare diseases. In particular, our analysis indicates that sub-groups used for allele frequency estimation should reflect the structure of genetic variation in the patient population, and suggests that the use of broad, continental population descriptors[39] may contribute to an excess of candidate variants in some patients. For example, our results are consistent with complex demographic and admixture histories of ancestral lineages in under-represented East African populations[40] resulting in patterns of common genetic variation distinct from the African (*afr*) reference populations annotated in gnomADv2 and gnomADv3. Efforts to include larger, more ancestrally diverse reference data are vital if we are to mitigate the disparities in variant prioritisation observed in our study. Future investigations should consider gnomADv4 (https://gnomad.broadinstitute.org/news/2023-11-gnomad-v4-0/) and other genomic resources such as the Regeneron Genetic Center Million Exomes (RGC-ME) dataset[41], which collectively include a further 335,000 individuals from previously under-represented backgrounds. Integration of data from sequencing initiatives such as those pioneered by H3Africa[24], GenomeAsia100k[42] and All of Us[43] will also help to address the European ancestry bias in existing resources. However, increased investment in the generation of genomic data from additional under-represented regions such as Eastern Africa or Oceania[36] is needed to achieve the goal of globally representative reference databases.

Investigation of the influence of genetic ancestry on diagnostic yield from genomic testing for rare diseases is challenging. Several previous studies have reported a lower diagnostic yield among patients with non-European ancestries likely resulting from greater uncertainty in variant classification[13,44]. Despite the observed disparities in the number of identified candidate variants, including those appearing within disease-associated genes, we found no evidence that a proband's genetic ancestry was associated with their overall likelihood of receiving a genetic diagnosis with a prioritised candidate variant. As such, these results may reflect developments in *in-silico* prediction methods[27,28] and functional genomic screening technologies[45] that are better able to distinguish pathogenic variants from the benign bystanders that likely make up the majority of excess candidate variants identified in probands with non-European ancestries. However, the large variance in diagnostic yield attributed to rare disease phenotypes in the 100kGP suggests that we should be cautious concluding that biases related to genetic ancestry do not exist, since proband numbers with non-European ancestries for many diseases are small. Furthermore, we note that there is increased potential for genetic misdiagnosis to occur among patients from under-represented ancestral backgrounds[36,38.]. With criteria such as absence from reference databases used as supporting evidence of a variant's pathogenicity (ACMG/AMP criteria; PM2[32], the continual reassessment of current diagnoses in the light of novel genetic variation captured by increasingly diverse reference data is warranted. Importantly, this would likely benefit patients from all backgrounds since variants that are shown to be common and benign in any one population are likely to be benign in all populations. Finally, the diagnostic evaluation of participants in the 100kGP is an ongoing process, and pathogenic variants identified through additional research are increasing the number of diagnoses across the cohort.

Further work is required to investigate the potential for ancestry biases in these novel and emerging diagnostic pathways.

Our study has several limitations. Firstly, as highlighted in the Deciphering Developmental Disorders study[46], a broad array of unmodelled factors are likely to influence the diagnostic yield for rare diseases, for example, those linked to the prenatal environment, and any of these may be correlated with reference population-based ancestry groups. Care should therefore be taken when attributing differences (or lack thereof) in diagnostic outcomes to patterns of genetic variation resulting from shared ancestry. Secondly, the large proportion of individuals with European ancestries and the very large variety of diseases phenotypes in the 100kGP cohort makes the detection of potential disease-specific patterns of bias challenging. Future work involving more ancestrally diverse rare disease cohorts with additional linked family and patient data may help elucidate any such patterns. Finally, we focussed on the prioritisation of small protein-altering variants, the majority of which are captured by both WGS and WES technologies. Evidence that both non-coding and large structural variants can contribute to rare disease pathogenesis is continuing to emerge[47,48]. Future investigations should therefore assess the influence of genetic ancestry on the prioritisation and interpretation of all types of genetic variation covered by WGS.

In conclusion, our analysis highlights the continuing impact of sampling biases that are endemic to human genomics. A better understanding of human genetic variation, together with a greater emphasis on the use of more granular or continuous metrics of genetic ancestry[49] will help bring us closer to the goal of equity in the genetic diagnosis of rare disorders.

Online methods

The 100,000 Genomes Project rare disease programme (100kGP) dataset

We analysed data from participants recruited and sequenced by Genomics England as part of the 100kGP rare diseases programme and included in the 100kGP data release (v17). From this release we selected 61,512 participants with high-coverage short-read WGS data aligned to the NCBI GRCh38 reference genome (mean autosomal read depth ~41x) and included in the multi-sample small variant aggregate (aggV2; https://re-docs.genomicsengland.co.uk/aggv2/) (Supplementary Information 1). Genomic data was linked to detailed metadata for 100kGP participants (Supplementary Information 5 and Supplementary Table 4). 29,425 participants were probands recruited with a wide range of 112 rare disease phenotypes and without a molecular diagnosis (described in detail elsewhere[5,250]). The remaining 32,087 individuals comprised both affected and unaffected family members.

Self-reported ethnicity

Self-reported ethnic groups encoded using NHS standard 16+1 data categories (https://www.datadictionary.nhs.uk/attributes/ethnic_category_code_2001.html) were available as linked metadata for 84.4% of probands (n = 24,854), with the remaining recorded as *`Not Known`* or *`Not Stated`*. After harmonising ethnic group labels (Supplementary

Information 6), the proportion of these probands belonging to each ethnic group recruited in England (n = 23,939) and stratified by age, (Supplementary Figures 1) was compared to self-reported ethnicity data from the ONS 2021 Census for England[18].

Principal Components Analysis (PCA)

We performed PCA on genotype matrices composed of a previously generated selection of 60,825 high-quality (HQ) variant sites from the 100kGP dataset (Supplementary Information 3) using the *bed_autoSVD()* function from the *bigsnpr* R package (https://privefl.github.io/bigsnpr/). A subset of 32,459 participants genetically unrelated to the 3rd degree (estimated using *KING*[51]; Supplementary Information 4) were used to construct PCs and the *bed_projectSelfPCA()* function was used to overlay data from all 29,053 related participants. Uniform manifold approximation and projection (UMAP) was subsequently performed using a matrix composed of the top 16 PCs using the *umap* (https://github.com/tkonopka/umap) R package with parameters *n_neighbours = 15* and *min_dist = 0.4*.

Genetically inferred ancestry (GIA)

Participants in the 100kGP dataset were assigned to genetically-inferred ancestry (GIA) groups according to their genetic similarity to an initial set of 21 sub-continental reference populations curated using data from the UK Biobank[20] as described in Privé *et al* 2022[21] (Extended Data 1). Specifically, the *big_prodMat()* function from the *bigsnpr* R package was used to project participant genotypes and reference group allele frequencies (AFs) across 55,706 intersecting HQ sites (Supplementary Information 3) onto the top 16 linkage disequilibrium-scaled principal components (PCs) previously generated using a selection of individuals from the UK Biobank and the 1000 Genomes Project[21]. Convex ancestry coefficients (α) (via an extension of the Summix method[52]) and squared Euclidean distances on the PC space (converted into an approximate F_{ST}[53]) were calculated between each participant and the 21 reference groups. Participants were assigned to the genetically closest reference group when $F_{ST} < 0.002$ or otherwise where $\alpha > 80\%$. A small number of closely related groups were merged (as in Privé *et al* 2022) and groups with fewer than 50 participants in the 100kGP (*South America, Finland, Japan*) (Extended Data 1) were relabelled as '*remaining participants'* alongside all individuals unable to be assigned using the above criteria. This resulted in a final set of 15 GIA groups (plus those labelled remaining participants).

For comparison, participants in the 100kGP were also assigned to continental GIA groups according to their genetic similarity to a set of 9 reference populations released as part of gnomADv3.1. Following the protocol outlines in https://gnomad.broadinstitute.org/news/2021-09-using-the-gnomad-ancestry-principal-components-analysis-loadings-and-random-forest-classifier-on-your-dataset/, the *pc_project* function from the Hail Python package (https://github.com/hail-is/hail) was used to project participant genotypes and reference group AFs across 76,003 sites onto the top 20 gnomADv3.1 PC loadings. The *assign_population_pcs* function from gnomAD Hail utilities (https://github.com/hail-is/hail) was then used to predict group assignments alongside using the

trained ONNX random forest model as the *fit* parameter with a minimum probability of 0.8. Participants with a fitted probability of <0.8 to all reference populations were labelled '*oth*'.

Candidate protein altering variants (cPAVs)

For all 100kGP probands, genome-wide small variants (SNVs and short <50bp indels) previously triaged using Genomics England's automated bioinformatics pipelines (described in detail elsewhere[22]; https://re-docs.genomicsengland.co.uk/tiering/) were included in the 100kGP v17 data release (Supplementary Information 7). In summary, after sample and genotype-level QC, small variants called in the proband and in available family members that were predicted to alter protein coding or splicing (sequence ontology (SO; http://www.sequenceontology.org/) consequence terms; Supplementary Table 1) were initially triaged by the automated bioinformatics pipeline into 'candidate small protein-altering variants' (cPAVs) using mode(s) of inheritance, reference population AF thresholds and appropriate familial co-segregation with the disease phenotype given the clinically indicated penetrance mode (complete or incomplete) (Supplementary Figure 9).

To summarise the number of cPAVs and gene-panel cPAVs per proband in the 100kGP dataset, we started with all cPAVs included in the 100kGP v17 data release as described above, restricted to only include cPAVs identified in families with WGS data aligned to the NCBI GRCh38 reference genome. With reference to the specific panel(s) applied to the proband and the variant's potential mode(s) of inheritance, we classified a subset of cPAVs as 'gene-panel cPAVs' (gene-panel cPAVs) using the most recent (as of January 1st 2024) 'green' genes included in PanelApp[29] (Supplementary Information 1, Supplementary Information 7, and Supplementary Table 3). gene-panel cPAVs identified in multiple virtual gene panels for each proband were recorded only a single time. This database was then filtered to only include cPAVs or gene-panel cPAVs that were either absent or rare across a selection of reference population databases (Supplementary Table 2), comprised of gnomADv2.1 (exomes, gnomAD v3.1 (genomes), the UK10k (exomes), and 6,628 genomes sequenced by Genomics England. Specifically, for reference populations with n>2,000 sequenced genomes or exomes, a filtering threshold of AF <0.1% was applied under monoallelic (dominant) modes of inheritance and <1% under biallelic (recessive) modes of inheritance. To ensure parity with AF filtering initially performed by the rare disease prioritisation pipeline (Supplementary Table 2), we further filtered cPAVs using a threshold of AF <0.2% (monoallelic) or AF<2% (biallelic) in databases with n <2,000 diploid genomes or exomes, with more lenient thresholds used to avoid filtration of rare variants whose population AFs may be inflated owing to the smaller number of sequenced individuals in the sub-group. All variants that passed these filters were defined as cPAVs (and a subset as gene-panel cPAVs) and counts were summarised per proband, with a count treated as singular irrespective of the homozygous or heterozygous state of the genotype.

Protein altering variants missing from gnomAD

To summarise the number of small predicted protein-altering variants missing from gnomAD per participant in the 100kGP dataset, we first extracted all *PASS* variants in aggV2 (Supplementary Information 3) across the autosomes and the X chromosome annotated using

ENSEMBL VEP v105[55] with SO consequence terms predicted to alter protein coding (Supplementary Table 1) in at least one transcript. *bcftools v10.1.2* [5657] *isec* was then used to extract only those protein-altering variants that were not present in either gnomAD v2.1 exomes (lifted over to GRCh38 coordinates) or gnomAD v3.1 genomes. *PLINK2*[58] *–-sample-counts* was used to summarise the number of missing protein-altering variants genotyped per participant (non-missing genotypes, including at multi-allelic sites in aggV2, Supplementary Information 1) with a count treated as singular irrespective of the homozygous or heterozygous state of the genotype.

Predicted deleteriousness of cPAVs

Deleteriousness predictions for cPAVs reported in the 100kGP across autosomes and the X chromosome were annotated where possible with two proteome-wide missense predictors AlphaMissense[27] and PrimateAI-3D[28] and one genome-wide variant effect predictor: CADDv1.6[26]. AlphaMissense scores were extracted from public Google Cloud buckets (https://console.cloud.google.com/storage/browser/dm_alphamissense), whilst PrimateAI-3D scores were downloaded from (https://primad.basespace.illumina.com/download) (both with consent for research-only use). CADD scores (PHRED-scaled) were generated using *ENSEMBL VEP v105* (as previously applied to all variants in aggV2; Supplementary Information 1). Annotations at *ENSEMBL* canonical transcripts were used for all methods and cPAVs were labelled as 'deleterious' according to specific method-specific score cut-offs (see Figure 4).

Clinical assessment information for cPAVs

Clinical assessments of variants, including those triaged as gene-panel cPAVs, provided by NHS Genomics England medical centres (GMCs; https://www.england.nhs.uk/publication/nhs-genomic-england-medicine-centres-map/) using standard ACMG/AMP criteria[59] and following the Association for Clinical Genomic Science best practice guidelines[33] were extracted from the Genomics England 100kGP v18 data release. Probands assessed as fully or partially resolved cases were defined as having received a diagnosis. A partially solved case is one where the causal variant identified was recorded in the v18 release as not fully explaining the patient's phenotype(s). Variants without a recorded ACMG/AMP assessment were taken as unclassified.

Multivariable generalised linear models (GLMs)

Utilising genomic, phenotypic, and pipeline-related data linked to probands in the 100kGP dataset, we used a series of multivariable regression models to assess the effect of predictor variables (X) on response variables (Y), adjusting for relevant covariates (Supplementary Table 4), including an interaction term between family structure and penetrance (β_6).

All models took the base form:

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 \text{proband age at date WGS data was received} + \beta_3 \text{proband karyotype} + \beta_4 \text{family group type} + \beta_5 \text{penetrance mode} + \beta_6 \text{ family structure, penetrance mode} + \beta_7$$

$$\text{disease phenotype} + \beta_8\text{date WGS data was received} + \beta_9\text{cumulative runs of homozygosity (ROH) >1Mb} + \beta_{10}\text{number of ROH >1Mb} + \beta_{11}\text{PanelApp gene panel size (Mb)} + \beta_{12}\text{Number of PanelApp panels applied} + \beta_{13}\text{read mapping error rate} + \beta_{14}\text{percentage of aligned reads} + \beta_{15}\text{mean autosomal read depth}$$

Where for each model, X is either the proband's GIA group as assigned using reference populations from the UK biobank (with *remaining participants* represented as a single additional group), the probands assigned group utilising gnomADv3.1 reference populations, or the number of protein-altering variants called in the proband that were missing from gnomAD (calculated as described above); and Y is one of the following variables each of which also defines the GLM link function as described:

(1) The count of cPAVs or gene-panel cPAVs (all cPAVs or gene-panel cPAVs considered) or the count of rare (population maximum AF < 0.1%) cPAVs or gene-panel cPAVs identified in the proband. This regression was performed using a negative binomial GLM with a log link (as the Poisson was found to be over-dispersed; cPAVs Z: 177.9, $p < 2.2 \times 10^{-16}$) with the glm.nb() function from the *MASS* R package (https://www.stats.ox.ac.uk/pub/MASS4/).

(2) The proportion of rare (population maximum AF < 0.1%, Supplementary Table 2) cPAVs or gene-panel cPAVs identified in the proband predicted to be deleterious (the count of rare cPAVs or gene-panel cPAVs annotated as deleterious divided by the total count of annotated rare cPAVs or gene-panel cPAVs). This regression was performed using a binomial GLM with a logit link with the *glm()* function in R was run separate for each of the variant effect predictors used to annotate deleterious cPAVs (SNPs only; AlphaMissense, PrimateAI-3D, and CADD) as described above. Unannotated variants were not included as counts in the regression.

(3) The proportion of gene-panel cPAVs (or cPAVs) identified in the proband that went on to be assessed as pathogenic or likely pathogenic (P/LP) resulting in a full or partial genetic diagnosis i.e the positive predictive value (PPV) = the count of gene-panel cPAVs assessed as P/LP divided by the total count of gene-panel cPAVs. This regression was performed using a binomial GLM with a logit link with the *glm()* function in R. Here, an additional covariate describing the GMC Trust handling the case was added to account for possible variation in the clinical assessment of gene-panel cPAVs.

(4) A binary variable describing whether at least one disease-associated P/LP cPAV or gene-panel cPAV was identified in the proband. This regression was performed using logistic regression with the *glm()* function from base R. Here, as in (3), the handling GMC Trust was included as an additional covariate in the model.

(5) The count of gene-panel cPAVs identified in the proband that were either unclassified (no ACMG/AMP classification) or were assessed as a variant of uncertain significance (VUS). This regression was performed using a negative binomial GLM with a log link with the glm.nb() function from the *MASS* R package. Here, as in (3), the handling GMC Trust was included as an additional covariate in the model. This analysis was performed after stratifying the dataset into two cohorts depending on whether the proband had received a full or partial genetic diagnosis.

To identify correlations between the effect of GIA group on the number of cPAVs identified in the proband (response variable Y = cPAVs) and the effect of GIA group on the number of protein-altering variants missing from gnomAD (as in Figure 3b) we performed the following additional negative binomial regression using the *glm.nb()* function, adjusting for sequencing quality control metrics which may influence variant calling:

$$E(\text{number of protein altering variants missing from gnomAD}) = \beta_0 + \beta_1 \text{proband GIA group} + \beta_2 \text{read aligned error rate} + \beta_3 \text{percentage of aligned reads} + \beta_4 \text{mean autosomal read depth}$$

GIA specific coefficient estimates (effect sizes on a log scale) across specific regression models were compared using the *cor.test()* R function where described (e.g. as in Figure 3b). The proportion of variance explained by each model was estimated using Nagelkerke's R^2 via the PseudoR2 function from the *DescTools* R package (https://cran.r-project.org/web/packages/DescTools/index.html). Type II ANOVA was performed using the Anova function from the car R package (https://cran.r-project.org/web/packages/car/index.html).

The COVID-19 study cohort

We analysed data from 34,701 participants recruited and sequenced by Genomics England as part of the COVID-19 Genomics Study (the COVID-19 cohort)[30] and included in the aggregated dataset v5 data release (aggCOVIDV5; https://re-docs.genomicsengland.co.uk/covid5/). This included high-coverage short-read WGS sequencing data aligned to the NCBI GRCh38 reference genome (mean autosomal read depth: ~43x) and subject to variant calling, sample and genotype-level QC, and aggregation as previously described[30] (Supplementary Information 2). Following the same method for GIA assignment as outlined for the 100kGP (see above), we selected 32,043 participants who were unrelated to the 3rd degree to participants both in the 100kGP and others within the COVID-19 cohort (as estimated using *KING*; Supplementary Information 5) mapping to one of 14 GIA groups in the COVID-19 cohort with more than 100 participants (see Figure 6).

Filtering allele frequency (FAF) calculation

For 505,993,047 variants in aggCOVIDv5 (including multi-allelic sites; Supplementary Information 2) across the autosome and the X chromosome, (non-missing) allele counts (AC) and allele numbers (AN) for each GIA group were calculated using *PLINK2 –-freq-counts*. After excluding variants with >10% missing genotypes (included masked low-quality sites;

Supplementary Information 2), we selected 986,893 variants in the COVID-19 cohort (mean missingness: 0.09%) identified as cPAVs in one or more probands in the 100kGP. 10 variants with a population maximum AF of <0.1% across all previously queried reference databases (Extended Data 1) and >2% cohort-wide AF across the COVID-19 cohort (mean AF: 19.3%), likely occurring because of differences in processing, variant calling, or QC between aggCOVIDV5 and each of the previously queried reference databases, were excluded from downstream analysis. For each of the 14 GIA groups in the dataset we used the Poisson sampling method described in Whiffin *et al* 2017[31] (https://github.com/ImperialCardioGenetics/frequencyFilter) to calculate AN adjusted, 95% confidence filtering allele frequencies (FAF95), which give the maximum credible GIA group AF (lower bound of the 95% CI) to account for the effect of GIA group sample size variance on AF estimation.

References

1. RARE Disease Facts. *Global Genes* https://globalgenes.org/rare-disease-facts/.
2. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).
3. Souche, E. *et al.* Recommendations for whole genome sequencing in diagnostics for rare diseases. *Eur. J. Hum. Genet.* **30**, 1017–1021 (2022).
4. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18**, 599–612 (2017).
5. Turnbull, C. *et al.* The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* **361**, k1687 (2018).
6. Snape, K., Wedderburn, S. & Barwell, J. The new genomic medicine service and implications for patients. *Clin. Med.* **19**, 273–277 (2019).
7. Inusa, B. P. D. *et al.* Sickle Cell Disease—Genetics, Pathophysiology, Clinical Presentation and Treatment. *Int. J. Neonatal Screen.* **5**, (2019).
8. Xiao, Q. & Lauschke, V. M. The prevalence, genetic complexity and population-specific founder effects of human autosomal recessive disorders. *Npj Genomic Med.* **6**, 1–7 (2021).
9. Quintana-Murci, L. Understanding rare and common diseases in the context of human evolution. *Genome Biol.* **17**, 225 (2016).

10. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
11. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
12. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* **43**, 1012–1030 (2022).
13. Hindorff, L. A. *et al.* Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).
14. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
15. Fatumo, S. & Choudhury, A. African American genomes don't capture Africa's genetic diversity. *Nature* **617**, 35–35 (2023).
16. Stranneheim, H. *et al.* Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. *Genome Med.* **13**, 40 (2021).
17. Marshall, C. R. *et al.* The Medical Genome Initiative: moving whole-genome sequencing for rare disease diagnosis to the clinic. *Genome Med.* **12**, 48 (2020).
18. Population of England and Wales. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/population-of-england-and-wales/latest/ (2022).
19. Mathieson, I. & Scally, A. What is ancestry? *PLoS Genet.* **16**, e1008624 (2020).
20. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
21. Privé, F. Using the UK Biobank as a global reference of worldwide populations: application to measuring ancestry diversity from GWAS summary statistics. *Bioinforma. Oxf. Engl.* **38**, 3477–3480 (2022).
22. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

23. Waldman, S. *et al.* Genome-wide data from medieval German Jews show that the Ashkenazi founder event pre-dated the 14th century. *Cell* **185**, 4703-4716.e16 (2022).
24. Choudhury, A. *et al.* High-depth African genomes inform human migration and health. *Nature* **586**, 741–748 (2020).
25. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
26. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
27. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
28. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).
29. Martin, A. R. *et al.* PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* **51**, 1560–1565 (2019).
30. Kousathanas, A. *et al.* Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature* **607**, 97–103 (2022).
31. Whiffin, N. *et al.* Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet. Med.* **19**, 1151–1158 (2017).
32. Harrison, S. M., Biesecker, L. G. & Rehm, H. L. Overview of specifications to the ACMG/AMP variant interpretation guidelines. *Curr. Protoc. Hum. Genet.* **103**, e93 (2019).
33. Durkie, M. *et al.* ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2024.
34. Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* **20**, 520–535 (2019).
35. Petrovski, S. & Goldstein, D. B. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol.* **17**, 157 (2016).

36. Butters, A. *et al.* A rare splice-site variant in cardiac troponin-T (TNNT2): The need for ancestral diversity in genomic reference datasets. 2024.02.08.24302375 Preprint at https://doi.org/10.1101/2024.02.08.24302375 (2024).
37. Vears, D. F., Niemiec, E., Howard, H. C. & Borry, P. Analysis of VUS reporting, variant reinterpretation and recontact policies in clinical genomic sequencing consent forms. *Eur. J. Hum. Genet.* **26**, 1743–1751 (2018).
38. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).
39. Use of Race Ethnicity and Ancestry as Population Descriptors in Genomics Research | National Academies. https://www.nationalacademies.org/our-work/use-of-race-ethnicity-and-ancestry-as-population-descriptors-in-genomics-research.
40. Pereira, L., Mutesa, L., Tindana, P. & Ramsay, M. African genetic diversity and adaptation inform a precision medicine agenda. *Nat. Rev. Genet.* **22**, 284–306 (2021).
41. Sun, K. Y. *et al.* A deep catalogue of protein-coding variation in 983,578 individuals. *Nature* 1–3 (2024) doi:10.1038/s41586-024-07556-0.
42. Wall, J. D. *et al.* The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
43. Bick, A. G. *et al.* Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
44. Shickh, S., Mighton, C., Uleryk, E., Pechlivanoglou, P. & Bombard, Y. The clinical utility of exome and genome sequencing across clinical indications: a systematic review. *Hum. Genet.* **140**, 1403–1416 (2021).
45. Hartin, S. N., Means, J. C., Alaimo, J. T. & Younger, S. T. Expediting rare disease diagnosis: a call to bridge the gap between clinical and functional genomics. *Mol. Med.* **26**, 117 (2020).
46. Wright Caroline F. *et al.* Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland. *N. Engl. J. Med.* **388**, 1559–1571 (2023).
47. Pagnamenta, A. T. *et al.* The impact of inversions across 33,924 families with rare disease from a national genome sequencing project. *Am. J. Hum. Genet.* **0**, (2024).

48. De novo variants in the non-coding spliceosomal snRNA gene RNU4-2 are a frequent cause of syndromic neurodevelopmental disorders | medRxiv. https://www.medrxiv.org/content/10.1101/2024.04.07.24305438v1.full#T1.
49. Lewis, A. C. F. *et al.* Getting genetic ancestry right for science and society. *Science* **376**, 250–252 (2022).
50. Turro, E. *et al.* Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96–102 (2020).
51. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
52. Arriaga-MacKenzie, I. S. *et al.* Summix: A method for detecting and adjusting for population structure in genetic summary data. *Am. J. Hum. Genet.* **108**, 1270–1282 (2021).
53. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**, 12–23 (2022).
54. Martin, A. R. *et al.* PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* **51**, 1560–1565 (2019).
55. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
56. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
57. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
58. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
59. 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. Off. J. Am. Coll. Med. Genet.* **17**, 405–424 (2015).

Data availability

The data supporting the findings of this study are available within the Genomics England Research Environment. To access genomic and clinical data within the Research Environment, researchers must first apply to become a member of either the Genomics England Research Network (academics/healthcare professionals) or the Discovery Forum (industry) via the Genomics England website (https://www.genomicsengland.co.uk/research/academic/join-research-network).

Code availability

All code used for data generation, analyses, and plotting within the Genomics Research Environment are available at: https://github.com/stallmanGEL/gel-ancestry-variant-prioritisation-publication.

Acknowledgements

This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. We thank the participants in the 100,000 Genomes Project (100kGP), who made this study possible. We also acknowledge the data contributed by the GenOMICC, REACT and ISARIC4C teams and thank the participants from those studies. The COVID-19 (GenOMICC) study was funded by the Department of Health and Social Care, Illumina, LifeArc, the Medical Research Council (MRC), UKRI, Sepsis Research (the Fiona Elizabeth Agnew Trust), the Intensive Care Society, a Wellcome Trust Senior Research Fellowship (223164/Z/21/Z) a BBSRC Institute Program Support Grant to the Roslin Institute (BBS/E/D/20002172, BBS/E/D/10002070 and BBS/E/D/30002275) and UKRI grants MC_ PC_20004, MC_PC_19025, MC_PC_1905 and MRNO2995X/1. We also acknowledge the National Institute for Healthcare Research Clinical Research Network (NIHR CRN) and the Chief Scientist's Office (Scotland), who facilitated recruitment into research studies in NHS hospitals, and to the global ISARIC and InFACT consortia.

Author Contributions

S.T., K.K, and L.M designed the study. S.T. carried out the analysis. S.T. and M.J.S. wrote the first draft of the manuscript. All authors contributed to discussion of the results, and reviewed and edited the manuscript.

Competing Interests

All authors are employees of Genomics England Ltd.