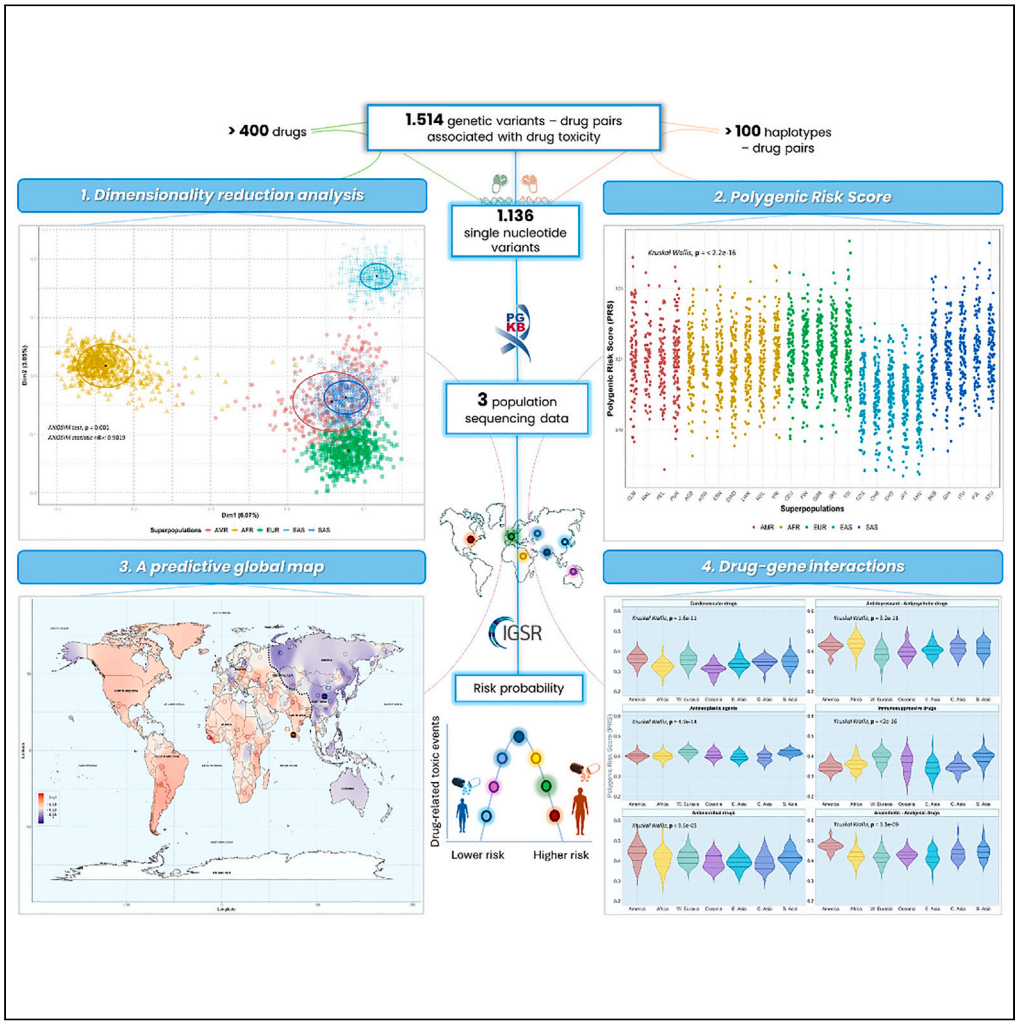
iScience

CellPress
OPEN ACCESS

Article

Genetic ancestry in population pharmacogenomics unravels distinct geographical patterns related to drug toxicity



Kariofyllis Karamperis, Sonja Katz, Federico Melograna, Francesc P. Ganau, Kristel Van Steen, George P. Patrinos, Oscar Lao

k.karamperis@goldenhelix.org (K.K.)
gpatrinos@upatras.gr (G.P.P.)
oscar.lao@ibe.upf-csic.es (O.L.)

Highlights

A predictive global map illustrating the likelihood of drug-related adverse events

Text-mining approach to identify risk alleles associated with drug toxicity

East Asians and Oceanians exhibit a relatively protective genetic profile

Ancestry-based stratified medicine holds promise for reshaping personalized medicine

Karamperis et al., iScience 27, 110916
October 18, 2024
https://doi.org/10.1016/j.isci.2024.110916

iScience

CellPress
OPEN ACCESS

Article

Genetic ancestry in population pharmacogenomics unravels distinct geographical patterns related to drug toxicity

Kariofyllis Karamperis,[1,2,3,*] Sonja Katz,[4,5] Federico Melograna,[6,7] Francesc P. Ganau,[2] Kristel Van Steen,[6,7] George P. Patrinos,[1,8,9,10,*] and Oscar Lao[2,11,*]

SUMMARY

Genetic ancestry plays a major role in pharmacogenomics, and a deeper understanding of the genetic diversity among individuals holds immerse promise for reshaping personalized medicine. In this pivotal study, we have conducted a large-scale genomic analysis of 1,136 pharmacogenomic variants employing machine learning algorithms on 3,714 individuals from publicly available datasets to assess the risk proximity of experiencing drug-related adverse events. Our findings indicate that Admixed Americans and Europeans have demonstrated a higher risk of experiencing drug toxicity, whereas individuals with East Asian ancestry and, to a lesser extent, Oceanians displayed a lower risk proximity. Polygenic risk scores for drug-gene interactions did not necessarily follow similar assumptions, reflecting distinct genetic patterns and population-specific differences that vary depending on the drug class. Overall, our results provide evidence that genetic ancestry is a pivotal factor in population pharmacogenomics and should be further exploited to strengthen even more personalized drug therapy.

INTRODUCTION

Adverse drug reactions (ADRs) encompass a significant public health concern and a leading worldwide cause of morbidity and mortality, ranking among the fourth to sixth most common causes of death.[1,2] By definition, ADRs are characterized as detrimental and unintended responses arising from an intervention with the utilization of a drug leading to a global expenditure.[3] Notably, the pharmacokinetics of chemical compounds exhibit inter-individual variability due to a mixture of genetic, environmental, or other factors (age, ethnicity, gender) harnessing the recommended drug and dosage prediction.[4–6] When these factors collectively lead to drug concentrations surpassing therapeutic thresholds, they can potentially give rise to a spectrum of adverse events, ranging from mild to severe, or even life-threatening conditions.[7] To address this challenge, pharmacogenomics (PGx) holds promise in customizing drug treatments and reducing the risk of drug-related adverse events within the context of personalized medicine by optimizing efficacy while simultaneously minimizing the risk of drug-related adverse events.[8–10]

Presently, the adoption of PGx in medical practice has garnered significant attention, driven by recent advancements in sequencing methods and the growing accessibility of human genomics data. These developments have facilitated the identification and analysis of numerous pharmacogenomic variants, serving as potential biomarkers with a profound impact on drug treatments.[11–13] Up to date, the majority of PGx studies[14–16] reflect the importance of identifying genetic variants modifying the risk of drug response and ADRs in important pharmacogenes encoding drug-metabolizing enzymes (e.g., CYP450),[17,18] drug transporters (such as ATP-binding cassette subfamilies),[19,20] drug receptors, or targets with a key role in pharmacokinetics and pharmacodynamics of drugs.[21] For instance, the CYP1A2, CYP2D6, CYP2C19, and CYP3A4 enzymes belonging to the CYP1, CYP2, and CYP3 families, respectively, are extensively involved in the biotransformation (metabolism) of the majority of foreign compounds, encompassing approximately 70%–80% of all drugs in clinical use.[22,23] Consequently, any genetic variations or alterations in pharmacogenes can significantly impact enzyme function, affecting the metabolism of drugs and potentially leading to different metabolizer status for particular medications to corresponding patients.[21,22] According to most recent

[1]Laboratory of Pharmacogenomics and Individualized Therapy, Department of Pharmacy, School of Health Sciences, University of Patras, Patras, Greece
[2]Group of Algorithms for Population Genomics, Department of Genetics, Institut de Biologia Evolutiva, IBE, (CSIC-Universitat Pompeu Fabra), Barcelona, Spain
[3]The Golden Helix Foundation, London, UK
[4]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, the Netherlands
[5]Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, the Netherlands
[6]Department of Human Genetics, KU Leuven, Leuven, Belgium
[7]GIGA-R Molecular and Computational Biology, University of Liège, Liège, Belgium
[8]Erasmus University Medical Center, Faculty of Medicine and Health Sciences, Department of Pathology, Clinical Bioinformatics Unit, Rotterdam, the Netherlands
[9]United Arab Emirates University, College of Medicine and Health Sciences, Department of Genetics and Genomics, Al-Ain, Abu Dhabi, UAE
[10]United Arab Emirates University, Zayed Center for Health Sciences, Al-Ain, Abu Dhabi, UAE
[11]Lead contact
*Correspondence: k.karamperis@goldenhelix.org (K.K.), gpatrinos@upatras.gr (G.P.P.), oscar.lao@ibe.upf-csic.es (O.L.)
https://doi.org/10.1016/j.isci.2024.110916

studies, a significant number of pharmacogenes (i.e., *CYP2C19*, *CYP2D6*, *CYP2C9*, or *CYP3A4*, among others) encoding the corresponding enzymes and beyond appear to be quite polymorphic not only among individuals but, most importantly, across different populations.[24–26] In the given context, population differentiation might be attributed, in part, to the intricate interplay of evolutionary forces such as natural selection, genetic drift, and gene flow, which contribute to shaping genetic diversity within populations.[27,28]

PGx evidence-based recommendations are continuously updated by regulatory bodies such as the U.S. Food and Drug Administration (FDA)[29] and the European Medicines Agency (EMA)[30] containing drug labeling information for more than 300 and 150 drug-biomarker pairs, respectively.[31–33] This dynamic process reflects the ongoing advancements in the field of PGx and the efforts to incorporate genetic information into clinical practice. However, given the aforementioned considerations, pharmacogenomic recommendations and guidelines may not adequately align with a specific cohort of patients, prompting the need to further stratify the risk individuals and explore alternative avenues or possibilities before clinical implementation.[16,26,34] Among others, the concept of population pharmacogenomics[34–38] has been prioritized.

Contemporary evidence regarding this concept has emphasized the importance of proactively stratifying individuals, considering additional biological information pertaining to population characteristics, such as ethnicity or ancestry.[39–43] Based on previous studies, the safety and effectiveness of drug treatment may differ based on the patient's genetic ancestry, and therefore, population-related factors should have a central role as guidance for personalized drug therapy.[44–46] As a result, ethnic background knowledge could potentially lead to better risk stratification.[47] Recent findings[48,49] in large, diverse populations with complex demographic and adaptation histories suggest that current pharmacogenomic guidelines from regulatory bodies may not fully harness the predictive therapeutic outcomes. This underscores the importance of accurately identifying genetic variants linked to ADRs in cohorts of human populations representing diverse ancestry and sex.[50] Additionally, investigating trends among populations with shared genetic backgrounds in specific drug categories (such as cardiovascular, antidepressant, and antineoplastic agents) could further enhance the stratification of personalized medicine and drug therapy.[51–54]

In essence, there is a strong recommendation to leverage the inherent genetic diversity within populations, in the context of PGx and particularly, within the concept of population pharmacogenomics. This approach can be instrumental in unveiling population-specific variants and providing guidance tailored to particular regions characterized by a prevalence of specific alleles, ultimately improving the stratification of individuals.[38,55,56]

In this study, we employed genomic information from various datasets, encompassing individuals from diverse geographical regions and locations. We developed a text-mining approach to extract risk alleles in genetic variants linked to adverse drug reactions and more specifically to drug toxicity from various databases, enabling us to investigate the genetic patterns related to ADR-associated variants. This allowed us to identify hidden patterns and similarities among populations in various drug classes by measuring the polygenic risk score.

RESULTS

In the current study, we have applied a text-mining approach but also supervised and unsupervised machine learning methods. This multifaceted approach was designed to unravel two major key aspects: first, to delineate the worldwide proximity of risk in the occurrence of ADRs, and second, based on the results to further explore these geographical patterns and assess drug-gene interactions in commonly prescribed medications. We thus developed a text-mining method for characterizing risk alleles from pharmacogenomic guidelines in the PharmGKB database (Data S2/Methods S2). In total, we included 1,136 PGx variants, including haplotypes (star alleles), associated with drug-related toxic events. These variants are distributed across 512 pharmacogenes and span evidence levels 1 to 3 (Data S1/Methods S1; Tables S1 and S2). To assess the validity of the text-mining methodology, we subsequently conducted a comparison between the final reports (semiautomatic vs. manually curated reports) for a subset of single nucleotide variants (SNVs) (n = 200) corresponding to multiple drugs, randomly selected among hundreds of reports. The vast majority of the risk genotypes were successfully parsed (>80%). Apart from that, the developed methodology has shown greater efficiency in comparison to the manual report in parsing risk alleles. In summary, our developed method has shown a series of advantages such as time efficiency in data mining, handling of large data volumes, and most importantly, ensuring the reliability and validity of our results.

Pharmacogenomic profile, variant annotation, and pathogenicity prediction

We first assessed the variant annotation and classification within a total of 1,136 PGx variants of significance. In our dataset, PGx variants are located in pharmacogenes linked to drug targets (63.6%), encoding drug-metabolizing enzymes—phases I and II (15.2%), drug receptors (11.5%), and transporters (9.7%) (Figure 1A). Depending on the variant position and signaling pathway, PGx variants represent alterations in the amino acid and protein structure, mainly resulting in a decreased activity and hence, to alter drug response or increase the risk for adverse drug reactions. In fact, most of these genetic variations are located in exonic regions (45.3%). Interestingly, missense variations represent 55.3% of the considered variants, following to a lesser extent the 3′ prime UTR variants (14%) (Figure 1B). Following this, we proceeded to predict the potential impact of missense variations (n = 285) using the PolyPhen-2 prediction tool (Figure 1C). The percentage of PGx variants for being possibly or even probably damaging in the structure and function of proteins was found to be relatively high (76.1%) with an average PolyPhen-2 score of 0.476 (see Table S3). This could lead to drug-related events and, in particular, an increased susceptibility to drug toxicity. In summary, the aforementioned findings highlight the significance of the included PGx variants, and therefore, a deeper analysis is highly recommended.

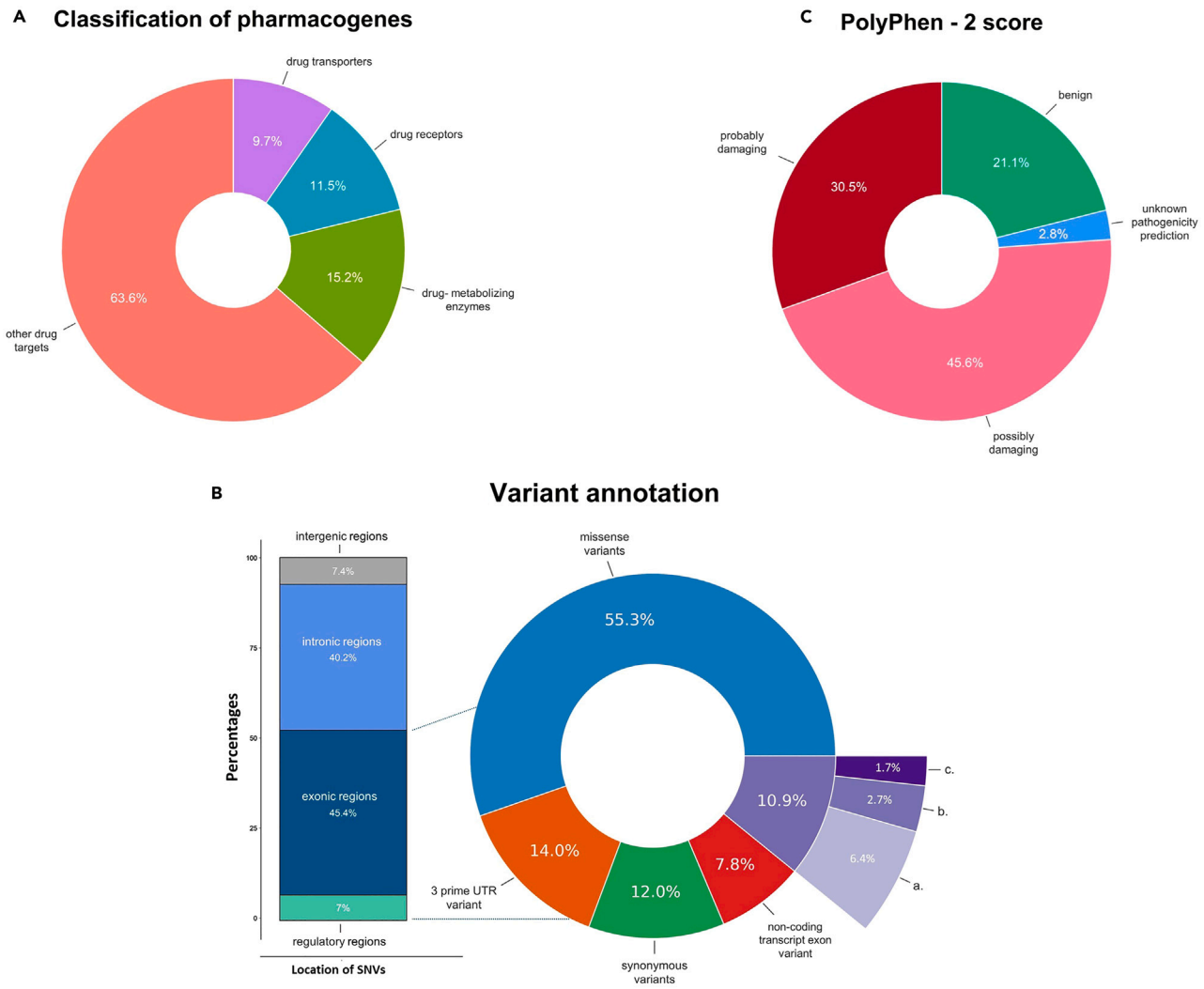


Figure 1. Pharmacogenomic profile, variant annotation, and pathogenicity prediction of the selected 1136 clinically relevant SNVs

(A) Distribution of SNVs located in genes encoding drug-metabolizing enzymes, transporters, receptors, and other drug targets.

(B) Variant annotation and consequences: the bar chart depicts the percentages of the SNVs typically located across four main genomic regions, within or between the genes. The nested pie chart illustrates the distribution of exonic variants, the coding regions, and the different types of consequences based on the genomic location. The category « other SNV consequences » includes the consequences of (a) stop gained, (b) frameshift variants, (c) inframe insertion, (d) inframe deletion, and (e) start lost. Similarly, this analysis was implemented using the Bioconductor and the BiomaRt package.[57,58]

(C) A predictive analysis to assess the damaging effects of SNVs and therefore, the possible impact of amino acid substitutions on the structure and function of proteins. Access to the biological databases was obtained through the Bioconductor, an open-source software, and the analysis was implemented with the PolyPhen-2 score[59] (Polymorphism Phenotyping v2) tool for missense variants using the BiomaRt package.[57,58]

Dimensionality reduction reveals distinct clustering of superpopulations

We first analyzed the genetic relationships of PGx variants associated with ADRs between individuals from worldwide populations. For this purpose, we consider the 1KGP3-ALL dataset due to its large number of individuals across diverse populations. We projected in two dimensions the genetic relationship defined by IBS estimated from a subset of PGx variants (n = 440) between individuals from the 1KGP3-ALL dataset using a classic MDS. The first two dimensions explained 9.12% of the variance present in the IBS distance matrix between pairs of individuals. Typically, a higher percentage of variance explained is recommended to achieve a more precise representation of the original data. However, even with a lower percentage of variance explained, valuable insights into the underlying structure of the data can still emerge. The first dimension (6.07% variance explained) distinguishes individuals from recent Sub-Saharan African ancestry compared to the other populations. The second dimension (3.05% variance explained) distinguishes East Asian populations from the others (Figure 2). Although a similar analysis was performed in the HGDP dataset, findings exhibited a high degree of similarity with the 1KGP3-ALL dataset,

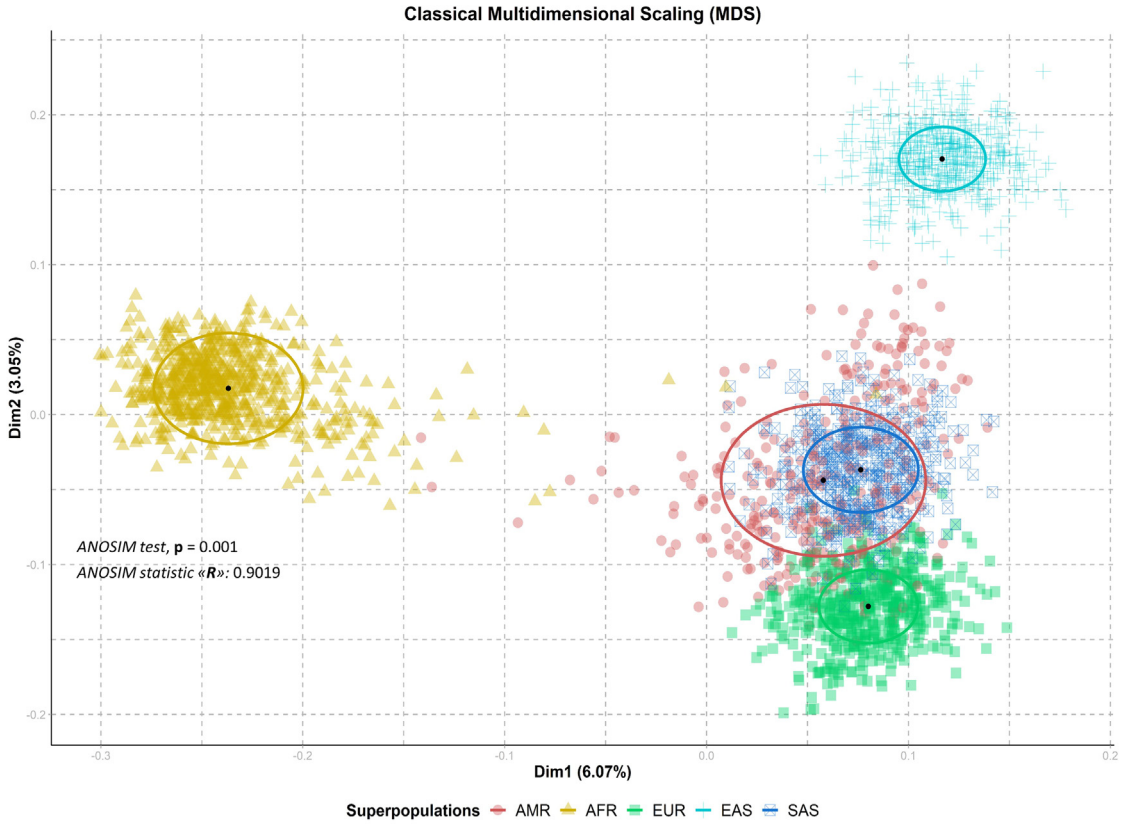


Figure 2. Dimensionality reduction reveals distinct clustering of superpopulations (1KGP3-ALL dataset)

This computational analysis was performed on 2.504 individuals across 26 populations divided into 5 superpopulations derived from the 1KGP3-ALL dataset. Centroids are depicted by black dots, whereas circular clusters (ellipses) are used to illustrate the calculated clusters, each distinguished by group-specific colors. Statistical significance, evaluated using the ANOSIM test[60] (Statistic R: 0.9019, significance: 0.001, number of permutations: 999), underscores the distinct groupings observed within the dataset.

demonstrating consistent patterns and supporting the presence of population structure between East Asian (first dimension, 3.38% variance explained) and African sub-Saharan (second dimension, 2.71% variance explained) groups. However, the MDS analysis of the HGDP dataset unveiled a notable proximity shared among the East Asian, Oceanian, and Native American populations, hinting at a compelling notion of shared genetic ancestry (Figure S2). To further evaluate the differences between predefined superpopulations, we also applied the ANOSIM test to both datasets, calculated from the MDS data. The ANOSIM test yielded an R statistic of 0.92 and a highly significant *p* value of 0.001 for the 1KGP3-ALL dataset, whereas in the HGDP dataset an R statistic of 0.902 and a highly significant *p* value of 0.001 was found, indicating substantial and statistically significant dissimilarities between the superpopulation groups at both datasets.

Polygenic risk score shows significant associations and disparities across diverse populations

Following the previous results, we computed a polygenic risk score (PRS) for each individual from the 1KGP3-ALL dataset using the previous subset of PGx variants. For this analysis, PRS is defined as the average number of risk alleles per individual. We found statistically significant differences in PRS across various populations and superpopulations, as confirmed by performing the Kruskal-Wallis test (Kruskal-Wallis chi-squared = 534.73, df = 25, *p*-value <2.2e-16) in the 1KGP3-ALL dataset. Individuals from East Asian ancestry tend to show a significantly lower PRS compared to other sub-continent groups (Figure 3). Similar findings were robustly replicated when extending the analysis to the HGDP dataset (Kruskal-Wallis chi-squared = 211.21, df = 53, *p*-value <2.2e-16). Interestingly, individuals of East Asian ancestry, and to a lesser extent the Oceanians, have demonstrated a notably lower PRS compared to other superpopulations, indicating reduced genetic susceptibility to ADRs (Figure S3). On the contrary, Americans, Europeans, and South Asians, following to a lesser extent the Africans, represent a higher PRS.

Unrevealing complex geographical patterns measuring the frequency of protective alleles across the globe

Given the previous findings, we further investigated and quantified the prevalence of protective alleles within each superpopulation. Via the logistic regression with each superpopulation as an outcome and the SNVs as input, we identified the SNVs distinctive of each

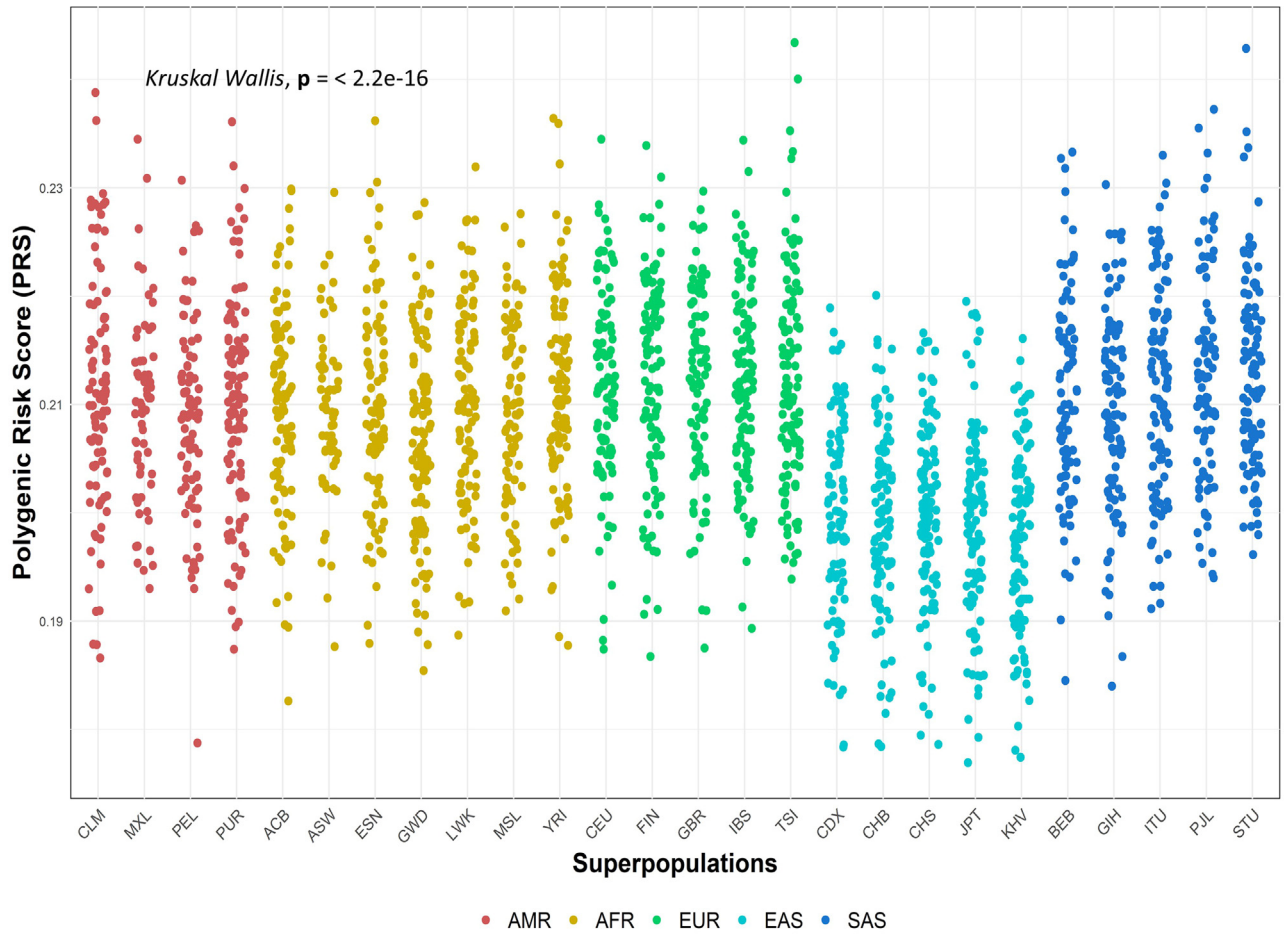


Figure 3. Polygenic risk score shows significant associations and disparities across diverse populations (1KGP3-ALL dataset)

In this scatterplot, we represent the polygenic risk score (PRS) of experiencing adverse drug effects among 26 populations divided into 5 superpopulations derived from the 1KGP3-ALL dataset. Acronyms are the following: Admixed American superpopulation: CLM, Colombians in Medellin, Colombia; MXL, People with Mexican ancestry in Los Angeles, USA; PEL, Peruvians in Lima, Peru; PUR, Puerto Ricans in Puerto Rico. African superpopulation: ACB, African Caribbean in Barbados; ASW, people with African Ancestry in Southwest USA; GWD, Gambian in Western Division; LKW, Luhya in Webuye, Kenya; MSL, Mende in Sierra Leone; YRI, Yoruba in Ibadan, Nigeria. European superpopulation: CEU, Utah residents (CEPH) with Northern and Western European ancestry; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian populations in Spain; TSI, Toscani in Italia. East Asian superpopulation: CDX, Chinese Dai in Xishuangbanna, China; CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese; JPT, Japanese in Tokyo, Japan; KHV, Kinh in Ho Chi Minh City, Vietnam. South Asian superpopulation: BEB, Bengali in Bangladesh; GIH, Gujarati Indians in Houston, TX, USA. ITU, Indian Telugu in the UK; PJL, Punjabi in Lahore, Pakistan; STU, Sri Lankan Tamil in the UK. Statistical analysis was performed with the Kruskal-Wallis test (Kruskal-Wallis chi-squared = 534.73, degrees of freedom (df) = 25, *p* value <2.2e-16).

superpopulation. To illustrate our findings, we utilized a bar plot (Figure 4) and an UpSet plot (Figure S4)—an extended version of the Venn Diagram—depicting the population-specific distribution of protective/risk-associated SNVs, using the 1KGP3-ALL dataset.

East Asians showed the highest number of protective alleles, i.e., low frequency of risk alleles (n = 329 SNVs) followed by Sub-Saharan African ancestry individuals (n = 265 SNVs), Europeans (n = 197 SNVs), South Asians (n = 146 SNVs), and Admixed Americans (n = 111 SNVs). It should be noted that, in almost all cases except East Asians, the average of risk alleles at each individual was almost equal to the protective ones, which constrained the interpretability of the results. Moreover, it was observed that, in particular cases, Admixed Americans, South Asians, and Europeans showed a high level of admixture when compared to the other superpopulations (Figure 4). These findings indicate that these particular superpopulations have a similar genetic background as referred to the selected SNVs of study and most importantly, a high level of admixture and heterogeneity, as previously identified. Of significant interest, individuals with Sub-Saharan African and East Asian ancestry maintain relatively distinct genetic profiles. Nevertheless, Sub-Saharan Africans tend to exhibit a higher proximity to increased risk in the occurrence of ADRs, whereas individuals of East Asian descent tend to show a decreased risk proximity.

Subsequently, our objective is to identify SNVs and evaluate their significance in terms of both protective and risk associations across multiple superpopulations. Notably, these SNVs exhibit distinctive frequencies within various superpopulations, prompting a detailed analysis of

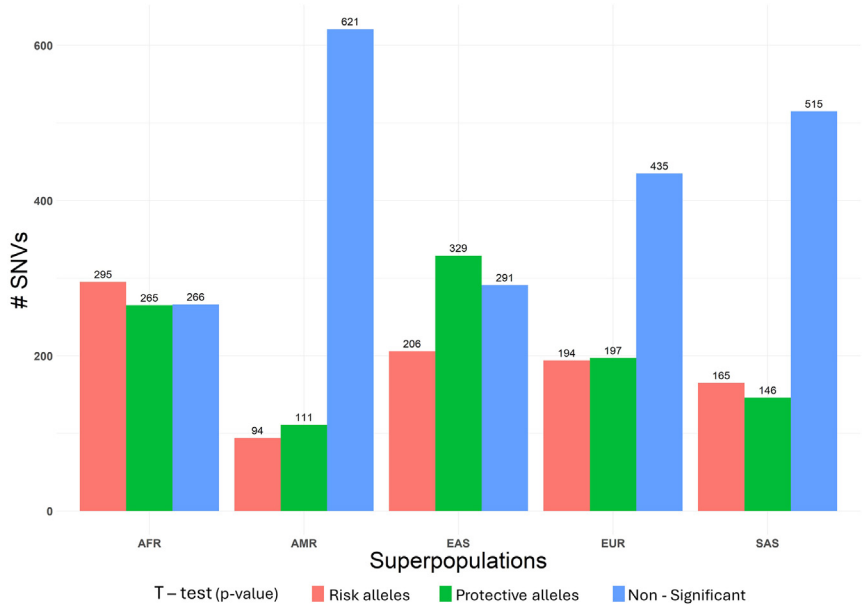


Figure 4. Regression analysis of risk and protective allele frequencies across superpopulations (1KGP3-ALL dataset)

Estimation of the frequency of the protective alleles per superpopulation derived from a subset of PGx variants (n = 884 SNVs). For this analysis, we compared the SNV average value within a superpopulation group with those of the remaining groups. In this bar plot, we represent the superpopulation groups and the number of SNVs (y axis) that were found with a lower (colored green), higher (colored red), or not significantly different (colored turquoise) value than the mean. The *p* value was calculated with a t test, and the results were adjusted for multiple testing with Bonferroni correction. These values with a lower mean suggest a protective effect, whereas higher values indicate a risk effect associated with drug-related toxic events. Non-statistical significance (>0.05) indicates a high admixture and similarity between superpopulation groups. East Asians have the highest number of protective alleles (frequency-beta), whereas Africans have the highest number of risk alleles.

their importance. In particular, 17 out of 826 SNVs have significantly different frequencies in every superpopulation, whereas 83 SNVs have similar frequencies in every superpopulation, i.e., no significant differences. Among all the superpopulations, the AMR seems the most isolated one, whereas Sub-Saharan ancestry African and East Asians share the highest number of predictive SNVs (Figure S4). Overall, these observations reveal population differences among these SNVs. Specifically, these differences primarily involve a small number of SNVs displaying extreme values, contributing to the emergence of a concealed pattern across two distinct dimensions.

A global map of genetic diversity reveals distinct geographical patterns linked to drug-related toxic events through spatial interpolation

Considering the findings mentioned earlier, a logical question may revolve around defining the spatial boundaries of the PRS. Even though the 1KGP3-ALL and HGDP datasets contain a relatively large population sampling size, they are both limited in terms of spatial coverage. Therefore, for spatial analyses, we employed the SGDP dataset, which consists of a relatively small number of samples per site but is more homogeneously geographically distributed. The geographic map of the mean of risk alleles suggests the presence of a sharp geographic discontinuity dividing Central Asia and including all East Asian, Siberian, and Oceanian populations, with a trend toward lower risk probability than the values found among the remaining global populations (Figure 5). To test this visual observation, we applied an unsupervised analysis inspired by K-means to identify the presence of geographic barriers. Following the assumption of two distinct groups, we applied a genetic algorithm (GA) to explore the space of possible solutions of spatial clusters in the SGDP, each group minimizing the difference in the PRS within the group and maximizing it between groups. Given the metaheuristic nature of the algorithm, we conducted multiple replicates, always obtaining the same spatial clustering at two groups. The Wilcoxon signed-rank *p* value between these two groups for the PRS was 1.276e-12. The geographic barrier between the two groups identified by this approach divides East Asian populations from the rest following the visual impression from the IWD-based map (Figure 5), supporting our previous findings with the 1KGP3-ALL and HGDP datasets.

Assessing the influence of discovery bias on PRS across populations

The presence of differences in PRS across populations could reflect biases in the discovery population of these ADRs. In order to test this hypothesis, we estimated the average number of derived alleles in SNVs that are polymorphic (minor allele frequency [MAF] >0.05) in European populations from HGDP and are classified as missense or 3′ prime UTR variants and compared with the expected if no MAF bias was applied. From the 875,428 missense and 3′ prime UTR variants SNVs with genotypes in all the HGDP individuals, 10,011 SNVs with a minimum

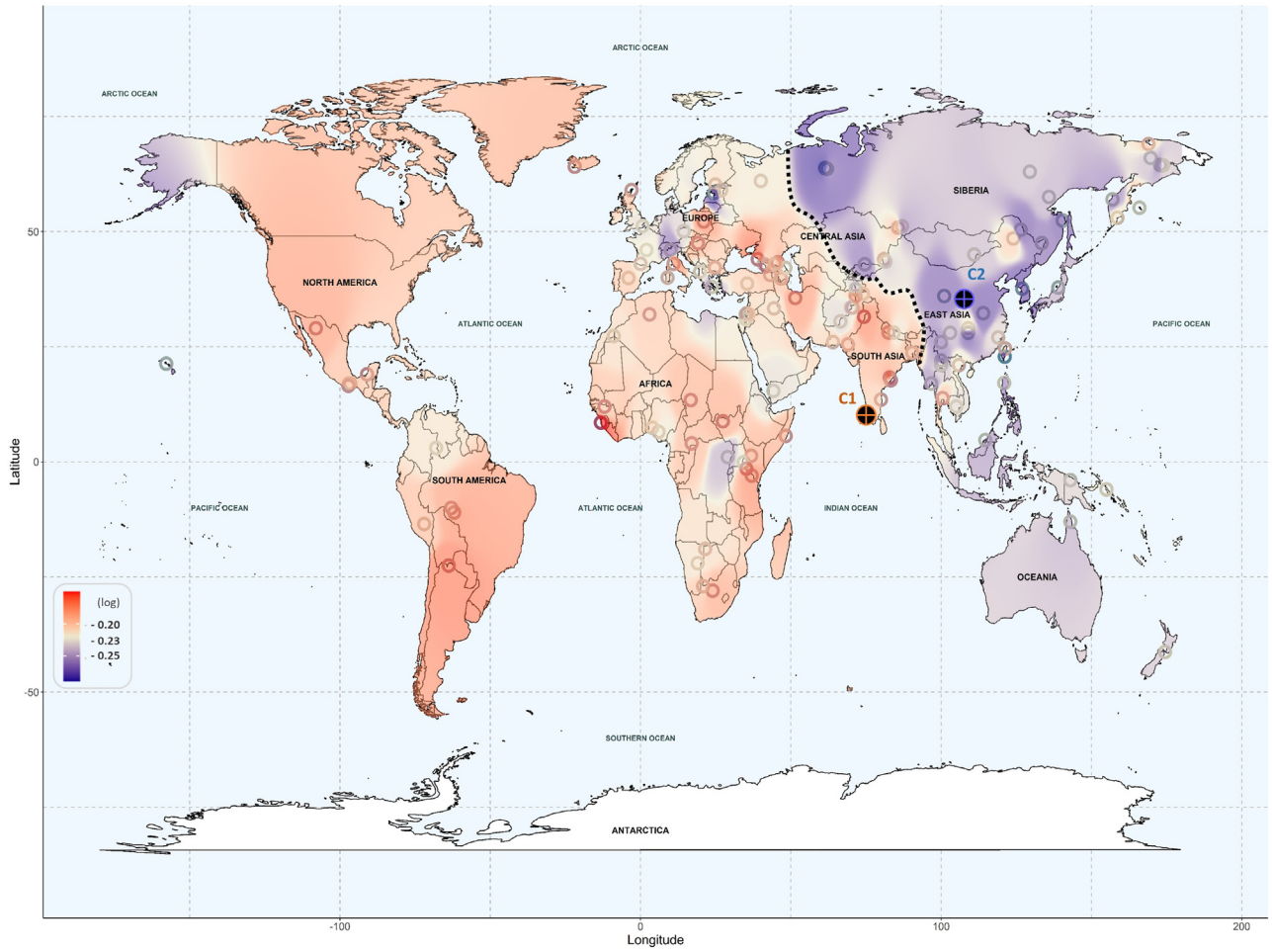


Figure 5. A global map of genetic diversity reveals distinct geographical patterns linked to drug-related toxic events through spatial interpolation

The mapping depicts the risk frequency in a set of pharmacogenomic variants involved in calculating the mean using logarithm (log) transformation in 281 individuals from 127 distinct populations and 7 superpopulations derived from the SGDP dataset. Each sampling region has been represented on the map by circular data points, with the colors (dark blue or light red) denoting the risk estimation within the respective regions. The outliers of other regions were interpolated using geospatial analysis and geostatistical modeling packages within R programming to estimate the interpolation and spatial distribution based on the known surrounding regions. Moreover, we applied a genetic algorithm (GA) to identify two distinct clusters. The centroid (C1) of the first group was lon = 107.17, lat = 26.8 (dark blue circular point featuring a cross at its center.), whereas, for the second centroid (C2), lon = 85.45 and lan = 7.69 (light red circular point featuring a cross at its center). The statistical significance between these groups was p = 1.276e-12, as determined by the Wilcoxon signed-rank test. The black line with square dots represents the geographical boundaries identified through spatial analysis and geostatistical modeling.

allele frequency of 0.1 in Europeans were ascertained. For each of the superpopulations of HGDP, we generated 10,000 PRS. Results, as illustrated in Figure S5A, reveal that PRS computed on alleles derived from missense or 3′ prime UTR variants with high MAF in European ancestry tends to be elevated in these populations. Conversely, PRS in other superpopulations tend to be lower. In contrast, in the absence of MAF bias, a higher frequency of derived alleles in missense and 3′ prime UTR variants is observed in African populations. A similar result is observed when the analysis is conducted only considering the pharmacogenes used in this study (see Figure S5B).

Geographic disparities in drug-gene interactions for commonly prescribed medications

Our previous findings provide compelling evidence for the presence of variance in the PRS across human populations. However, since we calculated a PRS by pooling effects from different drug categories, we conducted further analyses after stratifying the ADR genetic variants by the drug category that was targeted. In total, we analyzed more than 220 drugs linked to approximately 750 SNVs from both HGDP and SGDP datasets and stratified them into six main drug categories (Table S5). Since this a subset of the SNVs used in the pooled analyses, we first tested whether previous results could be replicated with this subset of SNVs. According to the results derived from the SGDP dataset, initial analyses without stratifying each medication category have shown a relatively higher PRS on individuals with American, West Eurasian, and South Asian ancestry while Africans followed to a lesser extent (Figure S6A). The statistical analysis revealed a significant Kruskal-Wallis

chi-squared value of 55,478 (df = 6, *p* value = 3.711e-10), indicating substantial differences among the groups compared. On the other hand, individuals with Central Asian and Siberian, and Oceanian ancestry represent a lower frequency, whereas East Asians have been observed with the lowest PRS, consistent with the previous analyses. Similarly, the HGDP dataset with the same subset of SNVs revealed statistically significant differences in distributions among the superpopulation groups (Kruskal-Wallis chi-squared = 55.478, df = 6, *p* value = 3.711e-10, Figure S6B). In the secondary analysis with the SGDP dataset, population differences have been observed, depending on the medication category. More specifically, Admixed Americans were found with a relatively higher risk in four out of six medication categories (cardiovascular drugs, antidepressant and antipsychotic drugs, antimicrobial drugs, anesthetic and analgesic drugs), following the Western Eurasians (cardiovascular drugs, antineoplastic agents, immunosuppressive and antimicrobial drugs) and, in a lower extent, South Asians (antidepressant and antipsychotic drugs, immunosuppressive drugs, anesthetic and analgesic drugs). Conversely, Oceanians, Central Asians and Siberians, Africans, and East Asians were found to be at a relatively lower risk compared to other populations, although we observed some exceptions that do not follow this trend (Table S6). For example, individuals with Central Asian ancestry in anesthetic and analgesic drugs were found to possess a considerably higher risk compared to others, while Africans tend to have a greater risk in antidepressant and antipsychotic drugs.

In the HGDP dataset, we observed relatively similar geographical patterns, although, compared to the SGDP dataset, some distinctions among superpopulations cannot be captured due to the presence of different sub-continental groups (Figure S7). Summary statistics and statistical significance among superpopulations from both datasets were performed using the Wilcoxon rank-sum test and can be found in Tables S7 and S8.

DISCUSSION

The impact of ethnicity on pharmacogenetics is unquestionably a dynamic area of study. Heritability estimates of SNVs in particular pharmacogenetic studies, predominantly conducted in European cohorts, have shown variations in other ethnic groups.[61] This indicates the possibility of concealed associations linked to ethnic differences, potentially influenced by background genome or environmental effects, that modify the drug response.[62,63]

Prior research examining the genetic diversity of specific genes, notably *CYP2D6*, has indicated a lack of significant geographic structuring in risk alleles associated with ADRs.[64] Conversely, investigations into various PGx genes, including but not limited to *CYP2D6*, *CYP2C19*, *DPYD*, *TPMT*, *NUDT15*, *SLC22A1*, *CFTR*, *HLA-A*, *HLA-B*, and *G6PD*, propose discernible distinctions both within and across continents.[45,65–70] Furthermore, studies concentrating on particular ethnic groups (i.e., Asian population) and specific gene subsets on commonly prescribed drugs provide evidence supporting divergences between these ethnicities and others.[52]

Nevertheless, to the best of our knowledge, a thorough examination of the impact of ancestry on ADRs has not been undertaken. In this investigation, we devised a systematic text-mining process for extracting SNVs associated with ADRs from accessible text databases. Subsequently, we spatially analyzed these variants to identify distinct geographic patterns in adverse drug events across global populations, utilizing diverse publicly available datasets. This text-mining approach enabled us to retrieve ADR data for 1,136 different variants across 512 pharmacogenes, establishing the most extensive dataset studied to date. Our findings indicate that the vast majority of considered ADR SNVs in this study are functional and that their minor alleles are likely to be risk alleles, as Kido et al. (2018) have recently underlined for complex diseases.[71]

From the MDS analysis using all ADR SNVs (Figure 2), we observed that the relative position of individuals at the 1KGP3-ALL and HGDP datasets exhibited similar clustering patterns, primarily based on their continental ancestry. These continental patterns broadly recapitulate the results obtained from analyses conducted using genetic variation from the whole genome.[72,73] Discrepancies observed in individuals from the American ancestry group between both datasets can be explained by the admixed ancestries of American individuals in the 1KGP3-ALL dataset compared to the HGDP dataset.[72,73] In summary, these findings support that pharmacogenetic loci serve as ancestry-informative markers, suggesting potentially significant differences in the genetic structure among diverse populations related to specific groups of drugs.[39]

However, it is important to note that this outcome does not automatically define the existence of genetic differences in susceptibility to ADRs across human populations. Analyses based on the average count of ADR alleles across all ADRs, as summarized by the PRS, reinforce the idea that the distribution of ADR alleles is contingent on the population and specifically at the continental level. In particular, all the spatial analyses in the SGDP dataset reflect distinct genetic profiles, especially among East Asian and Oceanian populations and the rest; individuals from East Asian and Oceania ancestry tend to have a lower PRS compared to the rest of individuals. It has been previously suggested that Asian (considering both East Asian and South-Asian ancestry) populations show different ADRs profiles for particular drugs such as carbamazepine or clopidogrel, among others.[52,74] Our results support such observations but also highlight that clustering South-Asian populations and East Asian populations to describe ADRs into a supra category such as "Asian" should be avoided, as these ancestries reflect different degrees of ADR susceptibilities. The observed disparities between East Asia and South Asia can be related to distinct genetic and ancestral backgrounds, despite being geographically close to each other.[75] South Asians have a unique genetic makeup influenced by historical migrations, interactions, and a mix of ancestral populations while East Asians represent a relatively more homogeneous genetic profile.[76]

Interestingly, discernible genetic disparities quantified with the PRS of a subset of SNVs were identified among populations across various medication categories. These disparities cannot be explained by a bias in the ascertained SNVs for conducting this analysis, as the PRS computed with this subset of SNVs prior to stratification underlined quite a similar distribution (Figures S6A and S6B) to the PRS considering all SNVs. When considering the PRS computed with the SGDP dataset for each drug category (Figure 6), Oceania, East, and Central Asia groups tend to have lower ADR-PRS for antineoplastic agents, immunosuppressive, cardiovascular, or antimicrobial drugs (Table S6).

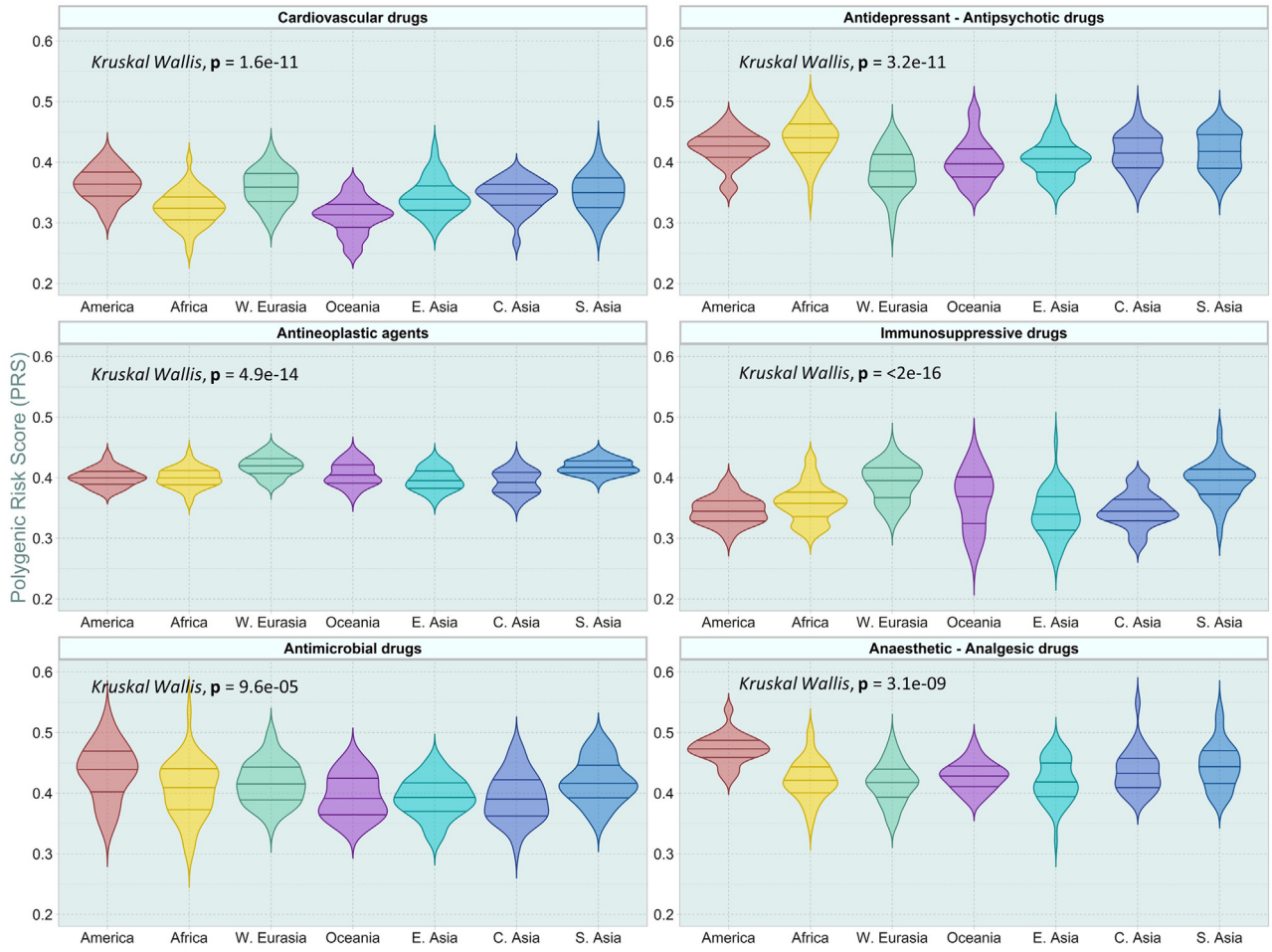


Figure 6. Faceted violin plot representing the polygenic risk scores of drug-gene interactions stratified into six main drug categories (SGDP dataset)

The medication categories refer to (1) cardiovascular drugs, (2) antidepressant and antipsychotic drugs, (3) antineoplastic agents, (4) immunosuppressive drugs, (5) antimicrobial drugs, (6) analgesic and anesthetic drugs. Violin plots representing the distribution of Polygenic Risk Scores per superpopulation, highlighting median values (0.50 quantiles) within each plot. The width of the violins denotes data density, whereas the upper and lower edges of the violins correspond to the 0.25 and 0.75 quantiles, respectively. We also performed the Wilcoxon rank-sum test (Mann-Whitney U test) to analyze all superpopulations. The statistical hypothesis test was used to assess whether two independent groups differ significantly in their distributions or central tendencies. Summary statistics and statistical significance can be found in Tables S7 and S8, respectively. To test whether there are statistically significant differences in the distribution among the superpopulation groups for each medication category, we performed the Kruskal-Wallis test and Bonferroni correction method to adjust the significance threshold. We also performed the Kruskal-Wallis rank-sum test (Kruskal-Wallis chi-squared = 52.845, df = 6, *p* value = 1.261e-09).

Nevertheless, the well-recognized challenge of restricted PRS transferability across ancestries persists in genetic epidemiology. This issue is attributed to various factors, with one of the primary considerations being the ancestral background of the population where the SNV-phenotype association was first identified, along with limited comprehension of shared causal variants.[77] This is particularly important in PGx, where pharmacovigilance efforts and policies differ across countries[78,79], and studies mainly focus on Eurocentric cohorts.[9,10,69] This bias implies a heightened ability to detect genetic variants that are more common in populations of European ancestry compared to other populations, thereby accentuating the PRS within this particular ancestral group, as evidenced by our observations. However, several pieces of evidence indicate that the identified differences in ADR-PRS are not a result of artifacts in the SNV discovery. Firstly, our study primarily focused on functional variants, distinguishing it from PRS derived from GWAS, which typically prioritizes top independently associated SNVs whose functional impact is unknown.[77] Secondly, biases in drug trial studies, predominantly featuring individuals of (North) European ancestry, lead to inadequate representation of individuals from ethnic minority groups.[80] This bias implies a preference for drugs that mitigate ADRs in European ancestries, regardless of their impact on other ancestral groups. Consequently, any SNV at intermediate frequency promoting ADR in Europeans might impede the commercialization of the drug before its release. Furthermore, our analyses of PRS computed on missense and 3′ prime UTR SNVs from the whole genome do not replicate the observed patterns with ADR-associated SNVs. While biasing by high MAF in European populations increases the PRS in this ancestry, all other ancestries tend to exhibit lower PRS values. This contrasts with the observed pattern in ADR-associated SNVs, where only East Asian and Oceanian superpopulations show a lower PRS relative to other

populations (see Figure S5B). On the contrary, in the absence of bias, the trends of increased enrichment of derived alleles in missense and 3′ prime UTR SNVs are more pronounced in African populations. This correspondence is consistent with the expectations outlined in the Out of Africa hypothesis,[81] wherein higher effective population sizes are noted in African populations compared to others, and populations outside Africa harbor a subset of the genetic variation present in Africa.[82] Therefore, our observed discrepancies in ADR-PRS across populations may truly reflect the average risk against ADRs and consequently, might also influence how individuals belonging to that population respond to medications, potentially impacting drug efficacy or susceptibility to adverse drug reactions.[35,83]

The incorporation of ancestry information within a stratified medicine framework holds promising implications for advancing personalized medicine into clinical practice.[44] Given the fact that the inter-population variability is distinctly present, the detection of clinically actionable PGx variants and disparities in the prevalence of risk alleles among populations can lead to the development of population-specific panels and genotype-guided prescriptions for common medications.[84,85] This approach can be applied strategically either to small geographic regions with a distinct genetic background compared to the surrounding areas,[86,87] or to large but relatively homogeneous regions where individuals share a relatively common genetic background.[33] In both cases, our findings indicate that the risk probability prediction could be more efficient, enhancing even more the concept of genome-guided treatment[55] for a corresponding population.

The broad implementation of predictive methodologies utilizing population sequencing data holds significant promise in mitigating challenges associated with the advancement of pharmacogenomics at a global scale.[84,88] This is particularly relevant, especially for low-developing countries where the absence of infrastructure, sequencing technology, and proficient personnel impedes the establishment of pharmacogenomic guidelines by regulatory bodies.[89–91] The application of population pharmacogenomics holds the potential to successfully overcome these challenges and play a pivotal role in advancing the adoption of pharmacogenomic guidelines, particularly in regions where PGx information from the regulatory bodies is either limited or absent.[31,92] Furthermore, it can serve to optimize existing frameworks, underscoring the critical influence of genetic ancestry in this process, prior to an individual's genetic makeup.[39] Overall, our results highlight the necessity of tailoring personalized medicine to consider both the diverse genetic landscape within populations and extend to this, the distinct pharmacological characteristics of various medication categories that might reflect hidden population discrepancies.

Limitations of the study

Our study contains a considerable number of SNVs-drug pairs, as well as, population sequencing data derived from three different publicly available databases, each serving distinct aims and purposes. Our findings support the conclusion that genetic ancestry plays a critical role in pharmacogenomics, although there are some limitations that need to be further discussed. At first, the majority of the SNVs-drug pairs reports were classified as levels of evidence (LoE) of 3, indicating that the strength and reliability of the evidence supporting pharmacogenomic associations are relatively low. Strengthening the evidence with mechanistic studies on those specific pharmacogenetic biomarkers and their functional implications for adverse drug reactions would improve our understanding of their clinical applicability and population-based heterogeneity. Furthermore, this study focused on single nucleotide variants due to the challenges posed by the data structure to parse other types of genetic variants. Improved data availability will allow us to further broaden our analysis beyond single nucleotide variants to include copy-number variations and structural variants. Finally, a deeper analysis, including larger-scale studies with diverse populations, is needed to strengthen the conclusions in order to validate patterns and ensure robustness, enhancing the depiction of risk disparities between populations and addressing gaps in regions with insufficient coverage. Additionally, heterogeneity in sequencing coverage, reference genome assembly, population classification, and sample sizes were noted in the composite dataset. These discrepancies could introduce bias and should be considered in future research endeavors.

RESOURCE AVAILABILITY

Lead contact

Further information and request for resources and reagents should be directed to and will be fulfilled by the lead contact, Oscar Lao (oscar.lao@ibe.upf-csic.es).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes, existing, publicly available data. DOIs are listed in the key resources table. Population sequencing data have been deposited at Mendeley and are publicly available as of the date of publication. The DOI is listed in the key resources table. Data from the text-mining methodology have been deposited at Mendeley and are publicly available as of the date of publication. The DOI is listed in the key resources table. Genetic variants-drug pairs are available in the PharmGKB database (https://www.pharmgkb.org/). Star allele nomenclature is available in the PharmVar database (https://www.pharmvar.org/). Drug or chemical compound information is available in the DruBank database (https://go.drugbank.com/).
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

GP gratefully acknowledges the funding from the European Union's Horizon 2020 research and Innovation Programme under the Marie Sklodowska-Curie Grant Agreements No 813533 (MLFPM) and No 860895 (TranSYS), the FNRS convention PDR T.0294.24 "Expanded PRS embracing pathways and interactions for

increased clinical utility. O.L. gratefully acknowledges the financial support from the Engineering and Physical Sciences Research Council of UK [grant number EP/X025160/1].

AUTHOR CONTRIBUTIONS

Conceptualization, K.K., G.P.P., and O.L.; methodology, K.K. and O.L.; formal analysis, K.K., F.M., S.K., F.P.G., and O.L.; writing, K.K.; writing review and editing, all authors; supervision, K.K., G.P.P., and O.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Data sources and processing methods
  - Computational predictive tools and machine learning algorithms
  - Exploratory and quantitative data analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110916.

Received: February 9, 2024
Revised: May 18, 2024
Accepted: September 6, 2024
Published: September 10, 2024

REFERENCES

1. Le Louët, H., and Pitts, P.J. (2023). Twenty-First Century Global ADR Management: A Need for Clarification, Redesign, and Coordinated Action. Ther. Innov. Regul. Sci. *57*, 100–103. https://doi.org/10.1007/s43441-022-00443-8.
2. Micaglio, E., Locati, E.T., Monasky, M.M., Romani, F., Heilbron, F., and Pappone, C. (2021). Role of Pharmacogenetics in Adverse Drug Reactions: An Update towards Personalized Medicine. Front. Pharmacol. *12*, 651720. https://doi.org/10.3389/FPHAR.2021.651720.
3. Sultana, J., Cutroneo, P., and Trifirò, G. (2013). Clinical and economic burden of adverse drug reactions. J. Pharmacol. Pharmacother. *4*, S73–S77. https://doi.org/10.4103/0976-500X.120957.
4. Božina, N., Vrkić Kirhmajer, M., Šimičević, L., Ganoci, L., Mirošević Skvrce, N., Klarica Domjanović, I., and Merćep, I. (2020). Use of pharmacogenomics in elderly patients treated for cardiovascular diseases. Croat. Med. J. *61*, 147–158. https://doi.org/10.3325/CMJ.2020.61.147.
5. Thummel, K.E., and Lin, Y.S. (2014). Sources of interindividual variability. Methods Mol. Biol. *1113*, 363–415. https://doi.org/10.1007/978-1-62703-758-7_17.
6. Alomar, M.J. (2014). Factors affecting the development of adverse drug reactions (Review article). Saudi Pharmaceut. J. *22*, 83–94. https://doi.org/10.1016/J.JSPS.2013.02.003.
7. Kaniwa, N., and Saito, Y. (2013). Pharmacogenomics of severe cutaneous adverse reactions and drug-induced liver injury. J. Hum. Genet. *58*, 317–326. https://doi.org/10.1038/jhg.2013.37.
8. Skokou, M., Karamperis, K., Koufaki, M.I., Tsermpini, E.E., Pandi, M.T., Siamoglou, S., Ferentinos, P., Bartsakoulia, M., Katsila, T., Mitropoulou, C., et al. (2024). Clinical implementation of preemptive pharmacogenomics in psychiatry. EBioMedicine *101*, 105009. https://doi.org/10.1016/J.EBIOM.2024.105009.
9. van der Wouden, C.H., Cambon-Thomsen, A., Cecchin, E., Cheung, K.C., Dávila-Fajardo, C.L., Deneer, V.H., Dolžan, V., Ingelman-Sundberg, M., Jönsson, S., Karlsson, M.O., et al. (2017). Implementing Pharmacogenomics in Europe: Design and Implementation Strategy of the Ubiquitous Pharmacogenomics Consortium. Clin. Pharmacol. Ther. *101*, 341–358. https://doi.org/10.1002/CPT.602.
10. Swen, J.J., van der Wouden, C.H., Manson, L.E., Abdullah-Koolmees, H., Blagec, K., Blagus, T., Böhringer, S., Cambon-Thomsen, A., Cecchin, E., Cheung, K.C., et al. (2023). A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomised crossover implementation study. Lancet *401*, 347–356. https://doi.org/10.1016/S0140-6736(22)01841-4.
11. Crews, K.R., Hicks, J.K., Pui, C.H., Relling, M.V., and Evans, W.E. (2012). Pharmacogenomics and individualized medicine: translating science into practice. Clin. Pharmacol. Ther. *92*, 467–475. https://doi.org/10.1038/CLPT.2012.120.
12. Lavertu, A., McInnes, G., Daneshjou, R., Whirl-Carrillo, M., Klein, T.E., and Altman, R.B. (2018). Pharmacogenomics and big genomic data: from lab to clinic and back again. Hum. Mol. Genet. *27*, R72–R78. https://doi.org/10.1093/HMG/DDY116.
13. Pirmohamed, M. (2023). Pharmacogenomics: current status and future perspectives. Nat. Rev. Genet. *24*, 350–362. https://doi.org/10.1038/S41576-022-00572-8.
14. Wei, C.Y., Lee, M.T.M., and Chen, Y.T. (2012). Pharmacogenomics of adverse drug reactions: implementing personalized medicine. Hum. Mol. Genet. *21*, R58–R65. https://doi.org/10.1093/HMG/DDS341.
15. Wang, L., McLeod, H.L., and Weinshilboum, R.M. (2011). Genomics and Drug Response. N. Engl. J. Med. *364*, 1144–1153. https://doi.org/10.1056/NEJMRA1010600.
16. Goh, L.L., Lim, C.W., Sim, W.C., Toh, L.X., and Leong, K.P. (2017). Analysis of Genetic Variation in CYP450 Genes for Clinical Implementation. PLoS One *12*, e0169233. https://doi.org/10.1371/JOURNAL.PONE.0169233.
17. Zhao, M., Ma, J., Li, M., Zhang, Y., Jiang, B., Zhao, X., Huai, C., Shen, L., Zhang, N., He, L., and Qin, S. (2021). Cytochrome p450 enzymes and drug metabolism in humans. Int. J. Mol. Sci. *22*, 12808. https://doi.org/10.3390/ijms222312808.
18. Tracy, T.S., Chaudhry, A.S., Prasad, B., Thummel, K.E., Schuetz, E.G., Zhong, X.B., Tien, Y.C., Jeong, H., Pan, X., Shireman, L.M., et al. (2016). Interindividual Variability in Cytochrome P450–Mediated Drug Metabolism. Drug Metab. Dispos. *44*, 343–351. https://doi.org/10.1124/DMD.115.067900.
19. Keogh, J.P. (2012). Membrane transporters in drug development. Adv. Pharmacol. *63*, 1–42. https://doi.org/10.1016/B978-0-12-398339-8.00001-X.
20. Arbitrio, M., Di Martino, M.T., Scionti, F., Barbieri, V., Pensabene, L., and Tagliaferri,

P. (2018). Pharmacogenomic Profiling of ADME Gene Variants: Current Challenges and Validation Perspectives. High. Throughput. *7*, 40. https://doi.org/10.3390/HT7040040.

21. Katara, P., and Yadav, A. (2019). Pharmacogenes (PGx-genes): Current understanding and future directions. Gene *718*, 144050. https://doi.org/10.1016/J.GENE.2019.144050.
22. Zanger, U.M., and Schwab, M. (2013). Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol. Ther. *138*, 103–141. https://doi.org/10.1016/J.PHARMTHERA.2012.12.007.
23. Fischer, A., and Smieško, M. (2021). A Conserved Allosteric Site on Drug-Metabolizing CYPs: A Systematic Computational Assessment. Int. J. Mol. Sci. *22*, 13215. https://doi.org/10.3390/IJMS222413215.
24. Zhou, S.F., Liu, J.P., and Chowbay, B. (2009). Polymorphism of human cytochrome P450 enzymes and its clinical impact. Drug Metab. Rev. *41*, 89–295. https://doi.org/10.1080/03602530902843483.
25. Fujikura, K., Ingelman-Sundberg, M., and Lauschke, V.M. (2015). Genetic variation in the human cytochrome P450 supergene family. Pharmacogenetics Genom. *25*, 584–594. https://doi.org/10.1097/FPC.0000000000000172.
26. Preissner, S.C., Hoffmann, M.F., Preissner, R., Dunkel, M., Gewiess, A., and Preissner, S. (2013). Polymorphic Cytochrome P450 Enzymes (CYPs) and Their Role in Personalized Therapy. PLoS One *8*, 82562. https://doi.org/10.1371/JOURNAL.PONE.0082562.
27. Li, J., Zhang, L., Zhou, H., Stoneking, M., and Tang, K. (2011). Global patterns of genetic diversity and signals of natural selection for human ADME genes. Hum. Mol. Genet. *20*, 528–540. https://doi.org/10.1093/HMG/DDQ498.
28. Auwerx, C., Lepamets, M., Sadler, M.C., Patxot, M., Stojanov, M., Baud, D., Mägi, R., Estonian Biobank Research Team, Porcu, E., Reymond, A., and Kutalik, Z. (2022). The individual and global impact of copy-number variants on complex human traits. Am. J. Hum. Genet. *109*, 647–668. https://doi.org/10.1016/J.AJHG.2022.02.010.
29. US Food and Drug Administration, FDA. https://www.fda.gov.
30. European Medicine Agency, EMA. https://www.ema.europa.eu.
31. Lee, M., Han, J.M., Lee, J., Oh, J.Y., Kim, J.S., Gwak, H.S., and Choi, K.H. (2023). Comparison of pharmacogenomic information for drug approvals provided by the national regulatory agencies in Korea, Europe, Japan, and the United States. Front. Pharmacol. *14*, 1205624. https://doi.org/10.3389/FPHAR.2023.1205624.
32. Ehmann, F., Caneva, L., Prasad, K., Paulmichl, M., Maliepaard, M., Llerena, A., Ingelman-Sundberg, M., and Papaluca-Amati, M. (2015). Pharmacogenomic information in drug labels: European Medicines Agency perspective. Pharmacogenomics J. *15*, 201–210. https://doi.org/10.1038/TPJ.2014.86.
33. Sahana, S., Bhoyar, R.C., Sivadas, A., Jain, A., Imran, M., Rophina, M., Senthivel, V., Kumar Diwakar, M., Sharma, D., Mishra, A., et al. (2022). Pharmacogenomic landscape of Indian population using whole genomes. Clin. Transl. Sci. *15*, 866–877. https://doi.org/10.1111/CTS.13153.
34. Nagar, S.D., Moreno, A.M., Norris, E.T., Rishishwar, L., Conley, A.B., O'Neal, K.L., Vélez-Gómez, S., Montes-Rodríguez, C., Jaraba-Álvarez, W.V., Torres, I., et al. (2019). Population Pharmacogenomics for Precision Public Health in Colombia. Front. Genet. *10*, 241. https://doi.org/10.3389/FGENE.2019.00241.
35. Bachtiar, M., and Lee, C.G.L. (2013). Genetics of Population Differences in Drug Response. Curr. Genet. Med. Rep. 162–170. https://doi.org/10.1007/S40142-013-0017-3.
36. Jordan, I.K., Sharma, S., and Mariño-Ramírez, L. (2023). Population Pharmacogenomics for Health Equity. Genes *14*, 1840. https://doi.org/10.3390/GENES14101840.
37. Ji, X., Ning, B., Liu, J., Roberts, R., Lesko, L., Tong, W., Liu, Z., and Shi, T. (2021). Towards population-specific pharmacogenomics in the era of next-generation sequencing. Drug Discov. Today *26*, 1776–1783. https://doi.org/10.1016/J.DRUDIS.2021.04.015.
38. Lakiotaki, K., Kanterakis, A., Kartsaki, E., Katsila, T., Patrinos, G.P., and Potamias, G. (2017). Exploring public genomics data for population pharmacogenomics. PLoS One *12*, e0182138. https://doi.org/10.1371/JOURNAL.PONE.0182138.
39. Yang, H.C., Chen, C.W., Lin, Y.T., and Chu, S.K. (2021). Genetic ancestry plays a central role in population pharmacogenomics. Commun. Biol. *4*, 171. https://doi.org/10.1038/S42003-021-01681-6.
40. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. Science *298*, 2381–2385. https://doi.org/10.1126/SCIENCE.1078311.
41. Nebert, D.W., and Menon, A.G. (2001). Pharmacogenomics, ethnicity, and susceptibility genes. Pharmacogenomics J. *1*, 19–22. https://doi.org/10.1038/sj.tpj.6500002.
42. Hernandez, W., Danahey, K., Pei, X., Yeo, K.T.J., Leung, E., Volchenboum, S.L., Ratain, M.J., Meltzer, D.O., Stranger, B.E., Perera, M.A., and O'Donnell, P.H. (2020). Pharmacogenomic genotypes define genetic ancestry in patients and enable population-specific genomic implementation. Pharmacogenomics J. *20*, 126–135. https://doi.org/10.1038/S41397-019-0095-Z.
43. Mersha, T.B., and Abebe, T. (2015). Self-reported race/ethnicity in the age of genomic research: Its potential impact on understanding health disparities. Hum. Genom. *9*, 1. https://doi.org/10.1186/S40246-014-0023-x.
44. Krainc, T., and Fuentes, A. (2022). Genetic ancestry in precision medicine is reshaping the race debate. Proc. Natl. Acad. Sci. USA *119*, e2203033119. https://doi.org/10.1073/PNAS.2203033119.
45. Zhou, Y., and Lauschke, V.M. (2022). Population pharmacogenomics: an update on ethnogeographic differences and opportunities for precision public health. Hum. Genet. *141*, 1113–1136. https://doi.org/10.1007/S00439-021-02385-X.
46. Khoury, M.J., Gwinn, M.L., Glasgow, R.E., and Kramer, B.S. (2012). A population approach to precision medicine. Am. J. Prev. Med. *42*, 639–645. https://doi.org/10.1016/J.AMEPRE.2012.02.012.
47. Ramamoorthy, A., Kim, H.H., Shah-Williams, E., and Zhang, L. (2022). Racial and Ethnic Differences in Drug Disposition and Response: Review of New Molecular Entities Approved Between 2014 and 2019. J. Clin. Pharmacol. *62*, 486–493. https://doi.org/10.1002/JCPH.1978.
48. Runcharoen, C., Fukunaga, K., Sensorn, I., Iemwimangsa, N., Klumsathian, S., Tong, H., Vo, N.S., Le, L., Hlaing, T.M., Thant, M., et al. (2021). Prevalence of pharmacogenomic variants in 100 pharmacogenes among Southeast Asian populations under the collaboration of the Southeast Asian Pharmacogenomics Research Network (SEAPharm). Hum. Genome Var. *8*, 7. https://doi.org/10.1038/S41439-021-00135-Z.
49. Suarez-Kurtz, G. (2005). Pharmacogenomics in admixed populations. Trends Pharmacol. Sci. *26*, 196–201. https://doi.org/10.1016/J.TIPS.2005.02.008.
50. Corpas, M., Siddiqui, M.K., Soremekun, O., Mathur, R., Gill, D., and Fatumo, S. (2024). Addressing Ancestry and Sex Bias in Pharmacogenomics. Annu. Rev. Pharmacol. Toxicol. *64*, 53–64. https://doi.org/10.1146/ANNUREV-PHARMTOX-030823-111731.
51. Westervelt, P., Cho, K., Bright, D.R., and Kisor, D.F. (2014). Drug–Gene Interactions: Inherent Variability In Drug Maintenance Dose Requirements. P T *39*, 630–637.
52. Lo, C., Nguyen, S., Yang, C., Witt, L., Wen, A., Liao, T.V., Nguyen, J., Lin, B., Altman, R.B., and Palaniappan, L. (2020). Pharmacogenomics in Asian Subpopulations and Impacts on Commonly Prescribed Medications. Clin. Transl. Sci. *13*, 861–870. https://doi.org/10.1111/CTS.12771.
53. Malki, M.A., and Pearson, E.R. (2019). Drug–drug–gene interactions and adverse drug reactions. Pharmacogenomics J. *20*, 355–366. https://doi.org/10.1038/s41397-019-0122-0.
54. Ortega, V.E., and Meyers, D.A. (2014). Pharmacogenetics: Implications of Race and Ethnicity on Defining Genetic Profiles for Personalized Medicine. J. Allergy Clin. Immunol. *133*, 16–26. https://doi.org/10.1016/J.JACI.2013.10.040.
55. Patrinos, G.P. (2018). Population pharmacogenomics: impact on public health and drug development. Pharmacogenomics *19*, 3–6. https://doi.org/10.2217/PGS-2017-0166.
56. Patrinos, G.P. (2020). Sketching the prevalence of pharmacogenomic biomarkers among populations for clinical pharmacogenomics. Eur. J. Hum. Genet. *28*, 1–3. https://doi.org/10.1038/S41431-019-0499-X.
57. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191. https://doi.org/10.1038/NPROT.2009.97.
58. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440. https://doi.org/10.1093/BIOINFORMATICS/BTI525.
59. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of

human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. *76*, 7–20. https://doi.org/10.1002/0471142905.HG0720S76.

60. Oksanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., Solymos, P., Stevens, M., Szoecs, E., et al. (2022). vegan: Community Ecology Package.
61. Perera, M.A., Gamazon, E., Cavallari, L.H., Patel, S.R., Poindexter, S., Kittles, R.A., Nicolae, D., and Cox, N.J. (2011). The Missing Association: Sequencing-Based Discovery of Novel SNPs in VKORC1 and CYP2C9 That Affect Warfarin Dose in African Americans. Clin. Pharmacol. Ther. *89*, 408–415. https://doi.org/10.1038/CLPT.2010.322.
62. Hatta, F.H.M., Lundblad, M., Ramsjo, M., Kang, J.H., Roh, H.K., Bertilsson, L., Eliasson, E., and Aklillu, E. (2015). Differences in CYP2C9 Genotype and Enzyme Activity Between Swedes and Koreans of Relevance for Personalized Medicine: Role of Ethnicity, Genotype, Smoking, Age, and Sex. OMICS *19*, 346–353. https://doi.org/10.1089/OMI.2015.0022.
63. Shah, R.R., and Gaedigk, A. (2018). Precision medicine: does ethnicity information complement genotype-based prescribing decisions? Ther. Adv. Drug Saf. *9*, 45–62. https://doi.org/10.1177/2042098617743393.
64. Sistonen, J., Sajantila, A., Lao, O., Corander, J., Barbujani, G., and Fuselli, S. (2007). CYP2D6 worldwide genetic variation shows high frequency of altered activity variants and no continental structure. Pharmacogenetics Genom. *17*, 93–101. https://doi.org/10.1097/01.FPC.0000239974.69464.F2.
65. Zhou, Y., Dagli Hernandez, C., and Lauschke, V.M. (2020). Population-scale predictions of DPD and TPMT phenotypes using a quantitative pharmacogene-specific ensemble classifier. Br. J. Cancer *123*, 1782–1789. https://doi.org/10.1038/S41416-020-01084-0.
66. Wright, G.E.B., Carleton, B., Hayden, M.R., and Ross, C.J.D. (2018). The global spectrum of protein-coding pharmacogenomic diversity. Pharmacogenomics J. *18*, 187–195. https://doi.org/10.1038/tpj.2016.77.
67. Zhang, B., and Lauschke, V.M. (2019). Genetic variability and population diversity of the human SLCO (OATP) transporter family. Pharmacol. Res. *139*, 550–559. https://doi.org/10.1016/J.PHRS.2018.10.017.
68. Gaedigk, A., Sangkuhl, K., Whirl-Carrillo, M., Klein, T., and Leeder, J.S. (2017). Prediction of CYP2D6 phenotype from genotype across world populations. Genet. Med. *19*, 69–76. https://doi.org/10.1038/GIM.2016.80.
69. Mizzi, C., Dalabira, E., Kumuthini, J., Dzimiri, N., Balogh, I., Başak, N., Böhm, R., Borg, J., Borgiani, P., Bozina, N., et al. (2016). A European Spectrum of Pharmacogenomic Biomarkers: Implications for Clinical Pharmacogenomics. PLoS One *11*, e0162866. https://doi.org/10.1371/JOURNAL.PONE.0162866.
70. Petrović, J., Pešić, V., and Lauschke, V.M. (2019). Frequencies of clinically important CYP2C19 and CYP2D6 alleles are graded across Europe. Eur. J. Hum. Genet. *28*, 88–94. https://doi.org/10.1038/s41431-019-0480-8.
71. Kido, T., Sikora-Wohlfeld, W., Kawashima, M., Kikuchi, S., Kamatani, N., Patwardhan, A., Chen, R., Sirota, M., Kodama, K., Hadley, D., and Butte, A.J. (2018). Are minor alleles more likely to be risk alleles? BMC Med. Genom. *11*, 3. https://doi.org/10.1186/S12920-018-0322-5.
72. 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/NATURE15393.
73. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*, eaay5012. https://doi.org/10.1126/SCIENCE.AAY5012.
74. Ang, H.X., Chan, S.L., Sani, L.L., Quah, C.B., Brunham, L.R., Tan, B.O.P., and Winther, M.D. (2017). Pharmacogenomics in Asia: a systematic review on current trends and novel discoveries. Pharmacogenomics *18*, 891–910. https://doi.org/10.2217/PGS-2017-0009.
75. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). Reconstructing Indian Population History. Nature *461*, 489–494. https://doi.org/10.1038/NATURE08365.
76. Moorjani, P., Thangaraj, K., Patterson, N., Lipson, M., Loh, P.R., Govindaraj, P., Berger, B., Reich, D., and Singh, L. (2013). Genetic Evidence for Recent Population Mixture in India. Am. J. Hum. Genet. *93*, 422–438. https://doi.org/10.1016/J.AJHG.2013.07.006.
77. Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K.K., Matsuda, K., Murakami, Y., Price, A.L., Kawakami, E., et al. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat. Genet. *52*, 1346–1354. https://doi.org/10.1038/s41588-020-00740-8.
78. Hans, M., and Gupta, S.K. (2018). Comparative evaluation of pharmacovigilance regulation of the United States, United Kingdom, Canada, India and the need for global harmonized practices. Perspect. Clin. Res. *9*, 170–174. https://doi.org/10.4103/PICR.PICR_89_17.
79. Khan, M.A.A., Hamid, S., and Babar, Z.-U.-D. (2023). Pharmacovigilance in High-Income Countries: Current Developments and a Review of Literature. Pharmacy *11*, 10. https://doi.org/10.3390/PHARMACY11010010.
80. Buffenstein, I., Kaneakua, B., Taylor, E., Matsunaga, M., Choi, S.Y., Carrazana, E., Viereck, J., Liow, K.K., and Ghaffari-Rafi, A. (2023). Demographic recruitment bias of adults in United States randomized clinical trials by disease categories between 2008 to 2019: a systematic review and meta-analysis. Sci. Rep. *13*, 42. https://doi.org/10.1038/S41598-022-23664-1.
81. Ashraf, Q., and Galor, O. (2013). The "Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development. Am. Econ. Rev. *103*, 1–46. https://doi.org/10.1257/AER.103.1.1.
82. Subramanian, S. (2019). Population size influences the type of nucleotide variations in humans. BMC Genet. *20*, 93. https://doi.org/10.1186/S12863-019-0798-9.
83. Fuselli, S. (2019). Beyond drugs: the evolution of genes involved in human response to medications. Proc. Biol. Sci. *286*, 20191716. https://doi.org/10.1098/RSPB.2019.1716.
84. Ahn, E., and Park, T. (2017). Analysis of population-specific pharmacogenomic variants using next-generation sequencing data. Sci. Rep. *7*, 8416. https://doi.org/10.1038/s41598-017-08468-y.
85. Verma, S.S., Keat, K., Li, B., Hoffecker, G., Risman, M., Regeneron Genetics Center, Sangkuhl, K., Whirl-Carrillo, M., Dudek, S., Verma, A., et al. (2022). Evaluating the frequency and the impact of pharmacogenetic alleles in an ancestrally diverse Biobank population. J. Transl. Med. *20*, 550. https://doi.org/10.1186/S12967-022-03745-5.
86. Idda, M.L., Zoledziewska, M., Urru, S.A.M., McInnes, G., Bilotta, A., Nuvoli, V., Lodde, V., Orrù, S., Schlessinger, D., Cucca, F., and Floris, M. (2022). Genetic Variation among Pharmacogenes in the Sardinian Population. Int. J. Mol. Sci. *23*, 10058. https://doi.org/10.3390/IJMS231710058.
87. Branco, C.C., Bento, M.S., Gomes, C.T., Cabral, R., Pacheco, P.R., and Mota-Vieira, L. (2008). Azores Islands: genetic origin, gene flow and diversity pattern. Ann. Hum. Biol. *35*, 65–74. https://doi.org/10.1080/03014460701793782.
88. Russell, L.E., Zhou, Y., Almousa, A.A., Sodhi, J.K., Nwabufo, C.K., and Lauschke, V.M. (2021). Pharmacogenomics in the era of next generation sequencing – from byte to bedside. Drug Metab. Rev. *53*, 253–278. https://doi.org/10.1080/03602532.2021.1909613.
89. Lauschke, V.M., and Ingelman-Sundberg, M. (2020). Emerging strategies to bridge the gap between pharmacogenomic research and its clinical implementation. NPJ Genom. Med. *5*, 9. https://doi.org/10.1038/s41525-020-0119-2.
90. Olivier, C., and Williams-Jones, B. (2011). Pharmacogenomic technologies: A necessary "luxury" for better global public health? Glob. Health *7*, 30. https://doi.org/10.1186/1744-8603-7-30.
91. Soko, N.D., Muyambo, S., Dandara, M.T.L., Kampira, E., Blom, D., Jones, E.S.W., Rayner, B., Shamley, D., Sinxadi, P., and Dandara, C. (2023). Towards Evidence-Based Implementation of Pharmacogenomics in Southern Africa: Comorbidities and Polypharmacy Profiles across Diseases. J. Personalized Med. *13*, 1185. https://doi.org/10.3390/JPM13081185.
92. Koutsilieri, S., Tzioufa, F., Sismanoglou, D.C., and Patrinos, G.P. (2020). Unveiling the guidance heterogeneity for genome-informed drug treatment interventions among regulatory bodies and research consortia. Pharmacol. Res. *153*, 104590. https://doi.org/10.1016/J.PHRS.2019.104590.
93. Karamperis, K., Katz, S., Melograna, F., Ganau, F.P., Van Steen, K., Patrinos, G.P., and Lao, O. (2024). Genetic ancestry in Population Pharmacogenomics unravels distinct geographical patterns related to drug toxicity. Mendeley Data *V1*. https://doi.org/10.17632/vtky42nggm.1.
94. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse

populations. Nature *538*, 201–206. https://doi.org/10.1038/NATURE18964.
95. Barbarino, J.M., Whirl-Carrillo, M., Altman, R.B., and Klein, T.E. (2018). PharmGKB: A worldwide resource for pharmacogenomic information. Wiley Interdiscip. Rev. Syst. Biol. Med. *10*, e1417. https://doi.org/10.1002/WSBM.1417.
96. Gaedigk, A., Whirl-Carrillo, M., Pratt, V.M., Miller, N.A., and Klein, T.E. (2020). PharmVar and the Landscape of Pharmacogenetic Resources. Clin. Pharmacol. Ther. *107*, 43–46. https://doi.org/10.1002/CPT.1654.
97. Gaedigk, A., Casey, S.T., Whirl-Carrillo, M., Miller, N.A., and Klein, T.E. (2021). Pharmacogene Variation Consortium: A Global Resource and Repository for Pharmacogene Variation. Clin. Pharmacol. Ther. *110*, 542–545. https://doi.org/10.1002/CPT.2321.
98. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082. https://doi.org/10.1093/NAR/GKX1037.
99. Ensembl REST API. https://rest.ensembl.org.
100. Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods *12*, 115–121. https://doi.org/10.1038/nmeth.3252.
101. Gräler, B., Pebesma, E., and Heuvelink, G. (2016). Spatio-Temporal Interpolation using gstat. Rom. Jahrb. *8*, 204.
102. Pebesma, E.J. (2004). Multivariable geostatistics in S: the gstat package. Comput. Geosci. *30*, 683–691. https://doi.org/10.1016/J.CAGEO.2004.03.012.
103. Pebesma E, B.R. (2005). Classes and methods for spatial data in R. R. News *5*, 9–13.
104. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D.A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., and Kuhn, M. (2019). Welcome to the Tidyverse. J. Open Source Softw. *4*, 1686. https://doi.org/10.21105/JOSS.01686.
105. Hijmans, R. (2023). raster: Geographic Data Analysis and Modeling. R package version *3*, 6–20.
106. Robert, J., van Etten, H., and van Etten, J. (2012). raster: Geographic analysis and modeling with raster data.
107. Kassambara (2023). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. https://CRAN.R-project.org/package=rstatix.Rpackageversion0.7.2.
108. South, A. (2011). rworldmap: A new R package for mapping global data. Rom. Jahrb. *3*. https://doi.org/10.32614/RJ-2011-006.
109. Scrucca, L. (2013). GA: A package for genetic algorithms in R. J. Stat. Software *53*, 1–37. https://doi.org/10.18637/JSS.V053.I04.
110. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006. https://doi.org/10.1101/GR.229102.
111. Ensembl. https://www.ensembl.org.
112. Relling, M.V., and Klein, T.E. (2011). CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. Clin. Pharmacol. Ther. *89*, 464–467. https://doi.org/10.1038/CLPT.2010.279.
113. Dutch Pharmacogenetics Working Group. DPWG. https://www.knmp.nl.
114. Pharmacogenomics Knowledge Base. PharmGKB. https://www.pharmgkb.org.
115. Clinical Pharmacogenetics Implementation Consortium. CPIC. https://cpicpgx.org.
116. Thorn, C.F., Klein, T.E., and Altman, R.B. (2013). PharmGKB: the Pharmacogenomics Knowledge Base. Methods Mol. Biol. *1015*, 311–320. https://doi.org/10.1007/978-1-62703-435-7_20.
117. Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C.F., Whaley, R., and Klein, T.E. (2021). An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. Clin. Pharmacol. Ther. *110*, 563–572. https://doi.org/10.1002/CPT.2350.
118. Kalman, L.V., Agúndez, J.A.G., Appell, M.L., Black, J.L., Bell, G.C., Boukouvala, S., Bruckner, C., Bruford, E., Caudle, K., Coulthard, S.A., et al. (2016). Pharmacogenetic allele nomenclature: International workgroup recommendations for test result reporting. Clin. Pharmacol. Ther. *99*, 172–185. https://doi.org/10.1002/CPT.280.
119. Robarge, J.D., Li, L., Desta, Z., Nguyen, A., and Flockhart, D.A. (2007). The star-allele nomenclature: retooling for translational genomics. Clin. Pharmacol. Ther. *82*, 244–248. https://doi.org/10.1038/SJ.CLPT.6100284.
120. Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., et al. (2023). Ensembl 2023. Nucleic Acids Res. *51*, D933–D941. https://doi.org/10.1093/NAR/GKAC958.
121. Pharmacogene Variation Consortium. Pharmvar. https://www.pharmvar.org.
122. Koromina, M., Pandi, M.T., van der Spek, P.J., Patrinos, G.P., and Lauschke, V.M. (2021). The ethnogeographic variability of genetic factors underlying G6PD deficiency. Pharmacol. Res. *173*, 105904. https://doi.org/10.1016/J.PHRS.2021.105904.
123. Appell, M.L., Berg, J., Duley, J., Evans, W.E., Kennedy, M.A., Lennard, L., Marinaki, T., McLeod, H.L., Relling, M.V., Schaeffeler, E., et al. (2013). Nomenclature for alleles of the thiopurine methyltransferase gene. Pharmacogenetics Genom. *23*, 242–248. https://doi.org/10.1097/FPC.0B013E32835F1CC0.
124. Hein, D.W., and Doll, M.A. (2012). Accuracy of various human NAT2 SNP genotyping panels to infer rapid, intermediate and slow acetylator phenotypes. Pharmacogenomics *13*, 31–41. https://doi.org/10.2217/PGS.11.122.
125. Huddart, R., Fohner, A.E., Whirl-Carrillo, M., Wojcik, G.L., Gignoux, C.R., Popejoy, A.B., Bustamante, C.D., Altman, R.B., and Klein, T.E. (2019). Standardized Biogeographic Grouping System for Annotating Populations in Pharmacogenetic Research. Clin. Pharmacol. Ther. *105*, 1256–1262. https://doi.org/10.1002/CPT.1322.
126. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.
127. McDonald, J., and Lambert, D.G. (2022). Drug–receptor interactions in anaesthesia. BJA Educ. *22*, 20–25. https://doi.org/10.1016/J.BJAE.2021.07.009.
128. Marc, J. (2008). 7. Pharmacogenetics of Drug Receptors. EJIFCC *19*, 48–53.
129. Nigam, S.K. (2015). What do drug transporters really do? Nat. Rev. Drug Discov. *14*, 29–44. https://doi.org/10.1038/NRD4461.
130. Santos, R., Ursu, O., Gaulton, A., Bento, A.P., Donadi, R.S., Bologa, C.G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T.I., and Overington, J.P. (2017). A comprehensive map of molecular drug targets. Nat. Rev. Drug Discov. *16*, 19–34. https://doi.org/10.1038/NRD.2016.230.
131. Crettol, S., Petrovic, N., and Murray, M. (2010). Pharmacogenetics of phase I and phase II drug metabolism. Curr. Pharmaceut. Des. *16*, 204–219. https://doi.org/10.2174/138161210790112674.
132. DrugBank. Database for Drug and Drug Target Info. https://go.drugbank.com..
133. Gower, J.C. (1966). Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. Biometrika *53*, 325. https://doi.org/10.2307/2333639.
134. Mardia, K.V. (1978). Some properties of clasical multi-dimesional scaling. Commun. Stat. Theor. Methods *7*, 1233–1241. https://doi.org/10.1080/03610927808827707.
135. Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938–2940. https://doi.org/10.1093/BIOINFORMATICS/BTX364.
136. MacQueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations, *1*, pp. 281–297.
137. Park, J.H., Gail, M.H., Weinberg, C.R., Carroll, R.J., Chung, C.C., Wang, Z., Chanock, S.J., Fraumeni, J.F., and Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc. Natl. Acad. Sci. USA *108*, 18026–18031. https://doi.org/10.1073/PNAS.1114759108.
138. Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alföldi, J., Watts, N.A., Vittal, C., Gauthier, L.D., et al. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. Nature *625*, 92–100. https://doi.org/10.1038/S41586-023-06045-0.
139. R Core Team (2022). R A Language and Environment for Statistical Computing (R Foundation for Statistical Computing). https://www.scirp.org/reference/referencespapers?referenceid=3456808.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw and analyzed data	This paper; Mendeley Data[93]	https://doi.org/10.17632/vtky42nggm.1
Population Sequencing data	This paper; Mendeley Data	https://doi.org/10.17632/vtky42nggm.1
Data from text-mining methodology	This paper; Mendeley data	https://doi.org/10.17632/vtky42nggm.1
1000 Genomes Project	1000 Genomes Project Consortium[72]	https://www.internationalgenome.org/data-portal/data-collection, https://doi.org/10.1038/nature15393
Human Genome Diversity Project	Bergström et al.[73]	https://www.internationalgenome.org/data-portal/data-collection, https://doi.org/10.1126/science.aay5012
Simons Genome Diversity Project	Mallick et al.[94]	https://www.internationalgenome.org/data-portal/data-collection, https://doi.org/10.1038/nature18964
Human reference genome NCBI build 38, GRCh38	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
PharmGKB database	National Institutes of Health (NIH); Barbarino et al.[95]	https://www.pharmgkb.org/
dbSNP database	National Center for Biotechnology Information	https://www.ncbi.nlm.nih.gov/snp/
PharmVar database (version: 5.1.12)	Gaedigk et al.[96,97]	https://www.pharmvar.org/
DrugBank database (version: 5.1.9)	Wishart et al.[98]	https://go.drugbank.com/
Software and algorithms		
R software (version: 4.2.1)	R Core Team	https://www.r-project.org/
Python software (version: 3.11)	Python Software Foundation	https://www.python.org/
JavaScript software (version: ES2022)	Pluraslight	https://www.javascript.com/
Ensembl API	Ensembl[99]	https://rest.ensembl.org/
Bioconductor	Huber et al.[100]	https://bioconductor.org, https://doi.org/10.1038/nmeth.3252
BiomaRt (version: 2.54.1)	Durinck et al.[57,58]	https://bioconductor.org/packages/release/bioc/html/biomaRt.html/
gstat (version:2.1.1)	Gräler; Pebesma[101–103]	https://github.com/r-spatial/gstat
stats (version: 4.2.2)	R foundation	https://www.R-project.org/
tidyverse (version: 2.0.0)	Wickham et al.[104]	https://www.tidyverse.org/
sp (version 2.1.3)	Pebesma et al.[103]	https://github.com/edzer/sp/
raster (version 3.6.26)	Hijmans et al.[105,106]	https://github.com/rspatial/raster/tree/master/
rstatix (version: 0.7.2)	Kassambara et al.[107]	https://github.com/kassambara/rstatix/
vegan (version 2.6.4)	Oksanen et al.[60]	https://github.com/vegandevs/vegan/
rworldmap (version: 1.3.8)	South et al.[108]	https://github.com/andysouth/rworldmap/
GA (version:3.2.4)	Scrucca[109]	https://github.com/luca-scr/GA/
Other		
UCSC human genome assembly	Kent et al.[110]	https://genome.ucsc.edu/
Ensembl database	Ensembl[111]	https://www.ensembl.org/
CPIC	Relling et al.[112]	https://cpicpgx.org/
DPWG	Dutch Pharmacogenetics Working Group[113]	https://www.knmp.nl/

METHOD DETAILS

Data sources and processing methods

Annotation of PGx variants from the PharmGKB database

A list of genetic variants-drug pairs was obtained from the Pharmacogenomics Knowledgebase (PharmGKB) database.[114] PharmGKB includes thousands of clinical annotations and pharmacogenomic guideline recommendations[95] published by the Clinical Pharmacogenetics Implementation Consortium (CPIC)[112,115] and Dutch Pharmacogenetics Working Group (DPWG)[113] linked to SNVs and haplotype variants for a number of drugs and drug-related phenotypes.[116,117] More specifically, 1,514 out of 5,015 annotations pertaining to PGx variant-drug pairs were selected, following certain criteria (see Figure S1). Any structural variants located in *HLA*, *SLC6A4*, *GSTM1*, and *GSTT1* genes such as gene insertions or deletions were excluded as this study was restricted to SNVs.

Overall, a set of 1,136 pharmacogenomic variants were meticulously curated and subsequently, selected for in-depth analysis. These variants comprised SNVs and haplotype variants described by the star allele nomenclature system; a standardized method used to categorize and designate genetic variations or alleles within pharmacogenes.[118,119] Star alleles are described as the phenotypic outcome of a genetic variant in terms of functionality. Therefore, an additional step was required to convert haplotype variants from PharmGKB that are written in star allele nomenclature system[119] into genetic variants with known dbSNP identifiers. The genomic position in the GRCh38 human genome build of the selected PGx variants was retrieved using the Ensembl REST API,[99] leveraging the Ensembl genomic data repository.[111,120]

Annotation of PGx haplotype variants using the PharmGKB and PharmVar databases

To convert star alleles into genetic variants with known dbSNP identifiers, we acquired data from the PharmVar database[121] (downloaded version 5.1.12). The aforementioned database serves as a comprehensive repository for pharmacogenetic variation, providing valuable insights into haplotype structures and allelic variations.[96,97] In cases where haplotype nomenclature was not available in the PharmVar database, we subsequently used the PharmGKB as an alternative source. On occasion, data were absent in both of the previously mentioned databases, we then conducted an extensive review of the scientific literature to identify haplotype variants.[65,122–124]

Population sequencing data

For this study, genomic data were collected from three main publicly available datasets namely the 1000 Genomes Project Phase 3 (1KGP3-ALL),[72] Human Genome Diversity Project (HGDP),[73] and Simons Genome Diversity Project (SGDP).[94] Given the diversity of populations within these datasets, the term « superpopulation» is used to systematically classify and group multiple geographically related populations. The availability of multiple datasets enables a comprehensive exploration of a distinct set of populations, facilitating a detailed analysis, essential for exploring various aspects of the study. This approach allows us to potentially replicate and reinforce our findings, thereby enhancing the robustness of our conclusions. Related information about the population structure of the sampling individuals and the Superpopulation classification of the mentioned datasets can be found in Table S4.

The 1KGP 3-ALL dataset is one of the most comprehensive and largest data resources of WGS providing a broad representation of human genetic variation. More specifically, it includes the genotype information of 2.504 unrelated individuals from 26 populations and five superpopulations categorized as Admixed American (AMR), African (AFR), European (EUR), East Asian (EAS), and South Asian (SAS), following the suggestion by Huddart et al.[125] The dataset was built on the human reference genome assembly CRCh38 and sequencing coverage of 30× for each individual's genome.[72]

In addition, the HGDP dataset was also used as it provides a relatively large number of samples per population and a more geographic homogeneous distribution across the globe. The HGDP includes the full genomes at high coverage of 929 individuals from 54 populations (CRCh37), which can be classified into seven superpopulations (Native American, African, European, Oceanian, East Asian, Middle Eastern, and Central South Asian).[73]

At last, SGDP provides a well-distributed geographical coverage of 281 individuals from 127 diverse populations (CRCh36) across seven superpopulations (America, Africa, West Eurasia, Oceania, East Asia, Central Asia & Siberia, and South Asia), each individual sequenced at a coverage of 30×.[94] The HGDP and SGDP datasets were lifted over to GRCh38 through the UCSC human genome assembly tool.[110,126]

Computational predictive tools and machine learning algorithms

Deriving risk alleles through a text-mining approach

Parsing risk alleles and genotypes for a specific genetic variant associated with a drug or chemical compound, as well as, drug phenotypes were derived through the PharmGKB database.[95,114,116] However, the current data structure does not allow us to automatically predict whether a particular risk genotype and allele for a corresponding drug is classified as risk. To address this issue, we developed a semi-automatic text-mining approach in order to parse risk genotypes from text-based clinical annotations and subsequently, to calculate the risk allele.

More specifically, the summary descriptions present in PharmGKB for each genotype were scanned for keywords indicating the direction of effect. Indeed, we identified a direction of effects through keywords, enabling us to define the risk proximity for each variant based on text-based clinical annotations, which corresponds to a particular drug or chemical compound. We differentiated between increased, decreased, and unknown risk. Subsequently, the likely risk allele was derived following logic reasoning, which determined that the allele occurring most frequently in all increased risk genotypes was most likely to be the causative allele. The textual descriptions for drug effects in PharmGKB greatly differ in diversity and complexity of phrasing, implying that the use of multiple - also contradictory - keywords may occur. To assess the validity of the

risk genotypes and alleles parsed by our approach, we included a confidence score ranging from 1 to 4, as inspired by the level of evidence score used in PharmGKB. This confidence score describes the complexity of the parsed text by estimating the number of keywords appearing in a passage. A score of 1 indicates the presence of only a single keyword, for which our parsing approach produces high fidelity results. The presence of multiple keywords in a description may lead to lower scores and should therefore be interpreted cautiously. This step was critical in determining the text-mining approach's ability to accurately extract risk genotypes, particularly when a specific SNV might corresponds to different drugs each time and therefore, different PGx recommendations might be accompanied. Subsequently, the risk allele is calculated based on the frequency of the risk genotypes derived from the text-based clinical annotations for each genetic variant - drug pair. This step was implemented with Python programming language. A comprehensive flowchart illustrating the procedure, along with predefined keywords and confidence scores, can be found in Data S2/Methods S2.

Pharmacogenomic profile, variant annotation, and pathogenicity prediction

To further gain insight into the function of the retrieved PGx variants, we predicted the possible impact of amino acids substitutions of the considered mutations on the structure and function of human proteins as described by the PolyPhen −2 score[59]; variant information was retrieved from Bioconductor[100] with the BiomaRt package.[57,58] In addition, we classified the pharmacogenes associated with the PGx variants in drug-metabolizing enzymes, drug transporters, etc., and their functional impact on the corresponding drug targets. Related data were obtained from both literature[21,127–131] and DrugBank database.[98,132]

Exploratory and quantitative data analysis

Dimensionality reduction analysis

In order to describe the genetic relationship between worldwide individuals at PGx SNVS, we computed an identical by-state (IBS) distance matrix between individuals in a subset of SNVs from the 1KGP3-ALL (n = 440) and HGDP datasets (n = 437). In both datasets, our analysis focused on a reduced set of SNVs, following specific parameters: SNVs should be located at least >100 kilobases to minimize the influence of linkage–disequilibrium (LD). AT and CG alleles were excluded due to difficulty distinguishing whether they belong to the reverse or forward strand of DNA. We applied a classical (metric) multidimensional scaling (MDS) approach, implemented using the cmdscale function in R software[133,134] to represent the IBS matrix in two dimensions. Statistical significance to evaluate the dissimilarity matrix was performed using the ANalysis Of SIMilarity (ANOSIM) test.[60]

Quantitative analysis

To further characterize the differential impact of ADRs across human populations, we estimated the mean number of ADR risk alleles for each individual from the different considered datasets. This raw polygenic risk score (PRS) was used as a quantitative measure of an individual's genetic predisposition to the occurrence of drug-related toxic events to commonly prescribed medications. Additionally, from the 1KGP3-ALL dataset, our goal was to discern SNVs uniquely associated with a certain superpopulation. To accomplish this, we modeled the superpopulations as a function of the SNVs via a logistic regression.

In more detail, our objective is to identify, for each superpopulation, the SNVs exhibiting varying frequencies of risk alleles. To achieve this, we utilized regression analysis to classify a superpopulation according to the frequency of risk alleles associated with each SNV. We first implemented a coding scheme where 0 value represented homozygous common (protective alleles), 1 represented heterozygotes (likely risk alleles), and 2 represented homozygous (risk alleles). Subsequently, we iteratively singled out one superpopulation, e.g., EUR, to whom we assigned the label "1", as opposed to all the other superpopulations, with a label of "0". Furthermore, we employed a logistic regression to predict the binary superpopulation label mentioned above for each SNV. We employed the default "binomial" logit function from the glm function in the stats package (R, version 4.2.2). The rationale behind this approach is that SNVs significantly predicting a specific superpopulation exhibit a unique risk frequency profile for that superpopulation. To assess significance, we monitored the *p*-values and test statistics for each SNV.[135] We repeat the analysis for each SNV in the dataset and for each superpopulation as target, i.e., with the label "1". The formula, for an individual *i*, a generic *SNVj* and superpopulation, i.e., EUR in the example, can be expressed as follows:

$$log\left(\frac{isEUR_i}{1 - isEUR_i}\right) = \beta_0 + \beta_1 SNP_{ij} + \varepsilon_i$$

$$where\ is\ EUR_i = \begin{cases} 1, & if\ i = EUR \\ 0, & otherwise \end{cases}$$

In the formula, our objective is to determine whether *SNVj* can predict if an individual belongs to the EUR population. Consequently, for each superpopulation, we identified a collection of SNVs that exhibited a distinct frequency pattern exclusive to that particular superpopulation. These distinctive SNVs, singled out as the ones with a significant *p*-value after Bonferroni correction (with the number of SNVs as a correction) in the previous formula, could exert either a higher, i.e., positive ($Beta_1$) or a lower, i.e., negative $Beta_1$, influence on the superpopulation. The interpretation of a positive significant $Beta_1$ suggests that the SNV displays a uniquely high frequency of risk alleles within that superpopulation, whereas a negative $Beta_1$ indicates that the SNV is associated with a high frequency of protective allele-specific to that superpopulation. We thereby label the significant SNVs as *risk* or *protective* alleles based on the sign of the Beta coefficient. This combined approach provides significant insights into the complex patterns and connections, in terms of genetic variability, among superpopulations.

A predictive global map: Spatial interpolation of the PRS

Expanding upon the preceding analyses related to the identification of population substructure among SNVs associated with ADRs, we further described the spatial relationship of risk disparities associated with ADRs. We computed the PRS, on a subset of PGx variants (n = 805) at each individual from the SGDP dataset. Geographic PRS interpolation for non-sampled points at the SGDP was conducted by the inverse weighted distance (IWD). For a non-sampled point from the dataset, IWD calculates an interpolated estimate as a weighted average considering the values at sampled locations. Essentially, the influence of nearby sampled locations is higher in shaping the interpolated estimate. This geostatistical modeling was performed using a combination of R packages (sp, raster, tidyverse and, gstat).[101–106,108]

Inference of geographic barriers of the PRS

In order to further describe the spatial distribution of the PRS, we inferred geographic barriers following an approach related to the K-Means algorithm.[136] Specifically, when confronted with a predefined number of K spatial clusters, our research goal was to identify geographic clusters such as the distance in the PRS between the individuals within each geographic cluster was as small as possible, and between clusters the largest possible quantified by means of the Wilcoxon signed-rank *p*-value. We employed the GA package[109] implemented in R to explore from a metaheuristic point of view the space of possible cluster configurations. Following the canonical GA, we generated a set of initial solutions, each coding the geographic coordinates for each of the K geographic clusters. Each solution is evaluated by assigning each geographic sampling point of the SGDP dataset to its closest geographic cluster. Wilcoxon signed-rank test is applied to the PRS. The best solutions, quantified in terms of the *p*-value computed for the given configuration in the Wilcoxon signed-rank *p*-value and then, were selected for generating new solutions (mating step) by mixing their geographic positions (recombination step) and updating the new solutions by adding innovations (mutation step). We considered a number of K geographic groups of two (centroid 1 and centroid 2), following the suggestion from the spatial interpolation of PRS, and a population of solutions of 100 in each iteration. The algorithm was run for 1000 generations. Recombination and mutation hyperparameters were set to default.

Generation of the null PRS under the assumption of European ADR discovery bias

Anticipating that a substantial proportion of adverse drug reactions (ADRs) variants were initially identified in individuals of European ancestry, it is plausible that the calculated polygenic risk scores (PRS) may primarily reflect this bias rather than genuine variations across diverse populations. To investigate this potential bias, simulated datasets of PRS were generated, leveraging the HGDP dataset. The methodology involved extracting all missense mutations documented in Ensembl for humans as proxies for functional changes.[111]

From the 875,428 missense and 3′ prime UTR variants SNVs with genotypes in all the HGDP individuals, 31,976 SNVs with a minimum allele frequency (MAF) of 0.05 in Europeans were ascertained out, recognizing that association studies often exhibit increased power in detecting statistical associations at highly polymorphic markers.[137] Adopting the assumption that derived missense changes are potentially detrimental[138] a pseudo-functional-PRS expected for an individual was computed by sampling the number of derived alleles on the frequency of each SNV in the population to which the individual belongs. The ancestral status of each SNV was retrieved from the information of the HGDP (VCF. file). To enhance robustness, 10,000 replicates of these simulated datasets were generated for each superpopulation.

Classification of commonly prescribed medications using the DrugBank database

From the extracted PGx variants - drug pairs list, we further proceeded to refine the stratification of commonly prescribed medications grouped into specific medication categories. Approximately, more than 400 drugs, chemical compounds, or combinations of drugs were assessed and thereafter, divided into six main categories. The final list of the medication categories and drug classes includes 1. cardiovascular drugs: (a. angiotensin drugs, b. anti-arrhythmic drugs, c. statins, d. cardioselective beta blockers, e. anticoagulant drugs, f. antiplatelet drugs and g. diuretics), 2. antidepressant: (a. SSRIs, b. SNRIs, c. SARIs, d. TCAs, e. TeCA) and antipsychotic drugs: (typical and atypical antipsychotics, phenothiazine antipsychotics), 3. antineoplastic agents: (a. antimetabolites, b. alkylating agents c. drug inhibitors, d. immunomodulatory drugs, e. monoclonal antibodies, f. anthracyclines, g. other antineoplastic agents) 4. immunosuppressive drugs: (a. corticosteroids, b. antimetabolites, c. biological response modifiers, d. calcineurin inhibitors), 5. antimicrobial drugs: (a. antiviral drugs, b. antibiotic drugs), and last but not least, 6. anesthetic and analgesic drugs: (a. opioid drugs, b. non-opioid drugs, and c. anesthetic). The grouping criteria of the drugs mostly concern the mechanism of action and the chemical properties of the corresponding medications and most importantly, to have been approved by at least one regulatory body (FDA,[29] EMA,[30] etc.). All the above drug-related information was extracted through DrugBank[98,132] and PharmGKB[95,114,116] databases, publicly available and serve as comprehensive resources giving necessary information on thousands of drugs.

In particular, we analyzed the PRS for commonly prescribed medications as a summary a) without stratifying the medication categories across different superpopulation, (Figures S6A and S6B) either, b) by analyzing the PRS for each medication category between superpopulations (see Figures 6 and S7). It's important to highlight that we observed situations where duplicate data (SNVs – drugs pairs) were identified multiple times within a medication category as they correspond to a different phenotype, each time. During the PRS analysis we excluded them to avoid the repeating values to enhance the reliability and validity of our results. However, if a duplicate single nucleotide variant corresponds to a different drug each time we included it. Similarly, if a single nucleotide variant has been detected across different medication categories corresponding to different drugs, we retained and integrated it into the analysis. Duplicate data were exclusively considered during the PRS analysis. The final drug list contains information about the generic name of a drug or a chemical compound, drug class, and broad

classification, as well as, the PGx variants - drug pairs per medication category, (see Table S5). For this analysis, we used as a primary source the SGDP and as complementary the HGDP dataset, as both have a greater spatial sampling size compared to 1KGP3-ALL in Asian sub-continents in which significant differences were previously observed. To evaluate the significance levels across superpopulations, we calculated the polygenic risk score for each medication category employing the Kruskal-Wallis test with Bonferroni correction (Table S6). Additionally, we conducted pairwise t-tests[107] to examine the differences between superpopulation groups (see Tables S7 and S8). All the above statistical analyses were performed using a combination of packages within the R programming language.[107,139]

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were accomplished using R programming language (version: 4.2.1). To support our findings, a variety of statistical hypothesis tests and algorithms were employed. During the dimensionality reduction analysis, we employed the Analysis of Similarities (ANOSIM), a non-parametric statistical test to compare and assess the dissimilarities between superpopulation groups based on a distance matrix. In this study, Euclidean dissimilarity was utilized as the distance measure, calculated from the multidimensional scaling (MDS) data. The significance value associated with the ANOSIM statistic indicates whether the observed differences between groups are statistically significant. Lower significance values (typically less than 0.05) suggest stronger evidence against the null hypothesis, indicating significant dissimilarities between groups.

In this study, we also measured the Polygenic risk score for each individual, a composite measure of genetic risk based on the average number of risk alleles identified from the text-mining approach we developed. To test whether there are statistically significant differences in the distribution among the superpopulation and population groups, based on the polygenic risk score, we performed the Kruskal-Wallis test and Bonferroni correction method to adjust the significance threshold. Similar statistical analysis was utilized to test the polygenic risk score for each medication category to assess the drug-gene interactions. In the latter case, we also performed the Wilcoxon rank-sum test (Mann-Whitney U test) on all superpopulations. This statistical hypothesis test was used to assess whether two independent groups differ significantly in their distributions or central tendencies. We have used the following convention for symbols indicating statistical significance: ns $\leq$ non-statistical significance; $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$. Polygenic risk scores for each medication category are available in Table S6, whereas tables presenting Summary Statistics and Statistical Significance for Polygenic Risk Scores of drug-gene interactions from the SGDP and HGDP datasets can be found in Tables S7 and S8. On the contrary, for the slope linear regression analysis, we utilized a t-test, a parametric approach and then adjusted for multiple testing with Bonferroni correction to compare the mean differences between the superpopulations.

To depict worldwide risk probability, we employed geospatial analysis and geostatistical modeling packages in the R programming environment. This facilitates the construction of a global map, enabling us to estimate interpolation and spatial distribution patterns based on sampled regions from the SGDP dataset. Our methodology involves the utilization of algorithms such as K-means clustering and genetic algorithms (GA) to delve into population variances and predict risk probability for the unsampled regions. For the GA, the statistical significance between the centroids was performed with Wilcoxon signed-rank *p*-value.