



Published in final edited form as:

*Mol Genet Metab.* 2022 November ; 137(3): 292–300. doi:10.1016/j.ymgme.2022.10.002.

## Metabolic diversity in human populations and correlation with genetic and ancestral geographic distances

Gang Peng<sup>a,b</sup>, Andrew J. Pakstis<sup>a</sup>, Neeru Gandotra<sup>a</sup>, Tina M. Cowan<sup>c</sup>, Hongyu Zhao<sup>a,b</sup>, Kenneth K. Kidd<sup>a</sup>, Curt Scharfe<sup>a,\*</sup>

<sup>a</sup>Department of Genetics, Yale University School of Medicine, New Haven, CT, USA

<sup>b</sup>Department of Biostatistics, Yale University School of Public Health, New Haven, CT, USA

<sup>c</sup>Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA

### Abstract

DNA polymorphic markers and self-defined ethnicity groupings are used to group individuals with shared ancient geographic ancestry. Here we studied whether ancestral relationships between individuals could be identified from metabolic screening data reported by the California newborn screening (NBS) program. NBS data includes 41 blood metabolites measured by tandem mass spectrometry from singleton babies in 17 parent-reported ethnicity groupings. Ethnicity-associated differences identified for 71% of NBS metabolites (29 of 41, Cohen's  $d > 0.5$ ) showed larger differences in blood levels of acylcarnitines than of amino acids ( $P < 1e-4$ ). A metabolic distance measure, developed to compare ethnic groupings based on metabolic differences, showed low positive correlation with genetic and ancient geographic distances between the groups' ancestral world populations. Several outlier group pairs were identified with larger genetic and smaller metabolic distances (Black versus White) or with smaller genetic and larger metabolic distances (Chinese versus Japanese) indicating the influence of genetic and of environmental factors on metabolism. Using machine learning, comparison of metabolic profiles between all pairs of ethnic groupings distinguished individuals with larger genetic distance (Black versus Chinese, AUC = 0.96), while genetically more similar individuals could not be separated metabolically (Hispanic versus Native American, AUC = 0.51). Additionally, we identified metabolites informative for inferring metabolic ancestry in individuals from genetically similar populations, which included biomarkers for inborn metabolic disorders (C10:1, C12:1, C3, C5OH, Leucine-Isoleucine). This work sheds new light on metabolic differences in healthy newborns in diverse populations, which could have implications for improving genetic disease screening.

### Keywords

Newborn screening; Public health; Inborn metabolic disorders; Population genetics; Biochemical genetics; Metabolism

---

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

\*Corresponding author at: 333 Cedar Street, SHM I-325B, USA. [curt.scharfe@yale.edu](mailto:curt.scharfe@yale.edu) (C. Scharfe).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ymgme.2022.10.002>.

## 1. Introduction

An individual's ancestry can be predicted through genotyping of a set of single-nucleotide polymorphisms (SNPs) [1-6]. Indeed, while millions of SNPs have been recorded in human genomes, selecting only a small set of highly informative SNPs is sufficient to accurately identify ancestry [7]. Grouping individuals based on genetic similarities has also been found to correspond with their self-defined ethnicity. A study of four racial/ethnic groupings (African American, East Asian, Hispanic, White) at sites in the United States and Taiwan found high correlation between self-identified ethnicity categorization and genetic cluster categories but only modest genetic differentiation between current geographical location within each ethnic group [8]. The results suggested that ancient geographic ancestry, which highly correlated with self-defined ethnicity, has been the major determinant of genetic structure in the US population.

Collecting and interpreting ancestral relationships in human populations could be utilized to minimize one source of confounding in genetic disease screening. For example, one potential reason for false positive results in newborn screening (NBS) are blood metabolite markers of disease that vary with the ethnicity status of the parents. Ethnicity-related differences in metabolite levels have been identified for several conditions including cystic fibrosis, congenital hypothyroidism, and glutaric acidemia type 1 [9-12]. Information on a newborns' ethnicity status, together with other confounding variables such as gestational age, sex, birth weight, age at blood collection, season of birth, or nutritional therapy, are increasingly being considered in genetic disease screening [13-16].

Here we studied whether ancient ancestral relationships could be uncovered from newborn metabolic screening data. To identify metabolic differences, we analyzed tandem mass spectrometry (MS/MS) data from a large and ethnically diverse cohort of screen-negative singleton babies reported by the California NBS program. Using this data we performed unique analysis to examine metabolic differences in human populations, to develop metabolic distance measures between populations, and to compare findings to genetic distances [17] and to ancestral geographic distances in world populations. Results identify novel relationships between metabolism and genetic ancestry – shedding new light on the breadth of metabolic diversity in healthy individuals [18]. Many studies have addressed how common genetic variation influence metabolite levels [19-21], while we still have limited knowledge on metabolic differences in non-European populations [22,23]. In addition, we developed an approach to capture a newborns' metabolic ancestry based on computational learning of metabolic screening data, with implications for improving genetic disease screening.

## 2. Materials and methods

### 2.1. Study population and data summary

This is a cross-sectional study of a cohort of screen-negative singleton newborns ( $n = 503,935$ ) who underwent population-based genetic disease screening in the California NBS program between 2013 and 2017. Screen-negatives receive no additional testing. Data included 41 metabolites detected using MS/MS screening from dried blood spots and 6

covariates including birth weight (BW), gestational age (GA), age at blood collection (AaBC), sex, parent-reported ethnicity, and total parenteral nutrition (TPN). Infants with unknown AaBC or collection before 12 h or after 168 h were removed as were infants with BW below 1000 or above 5000 g, or with GA under 28 weeks or after 42 weeks. We also removed infants with positive or unknown TPN status. Of the remaining 489,258 infants, 392,303 (80.2%) belonged to only one of 17 racial/ethnic groups (Asian East Indian, Black, Cambodian, Chinese, Filipino, Guamanian, Hawaiian, Hispanic, Japanese, Korean, Laos, Middle Eastern, Native American, Other Southeast Asian, Samoan, Vietnamese, White). Infants with multiple ethnicity categories (17.8%,  $n = 87,097$ ) and those with unknown ethnicity (2.0%,  $n = 9858$ ) were removed (Table 1). This study used the STROBE cross sectional reporting guidelines [24], and was overseen by institutional review boards at Yale University, Stanford University and the State of California Committee for the Protection of Human Subjects.

## 2.2. Metabolic distance analysis

Blood levels of 41 NBS metabolites were compared between the 17 parent-reported newborn groupings. A 41 by 17 data matrix  $M$  was created with  $M_{i,j}$  indicating the mean level of metabolite  $i$  in ethnic group  $j$  (S1 Table). Due to differences in scale between metabolites, each row in the data matrix was standardized to have a mean of 0 and standard deviation of 1. Hierarchical clustering based on Manhattan distance was used to explore relationships between metabolites and groups. To compare metabolite level differences between all pairs of groups ( $n = 136$  pairs), a 17 by 17 data matrix  $D$  was computed with  $D_{i,j}$  indicating the Manhattan distance of all 41 metabolites between two groups  $i$  and  $j$  ( $D_{ij} = \sum_{k=1}^{41} |M_{ki} - M_{kj}|$ ) (S2 Table). Phylogenetic tree analysis (PTA) and multidimensional scaling (MDS) was used to analyze and compare the metabolic distances between group pairs.

PTA employing the Neighbor-Joining method [25] was used to analyze metabolic distance data. This method assumes an additive tree and generates a tree that is a good approximate least squares solution for the input data, but the tree is not necessarily the best solution considering all possible tree structures nor is it an exact solution for the tree structure. Under the standard assumptions of genetic drift, analyzing the Tau genetic distance matrix should be additive and yield a tree structure in which the branch lengths represent additive components in units of  $t/2N_e$ . Manhattan distances were employed since they should be additive. Although it is impossible to examine every possible tree, it is possible to derive an exact least squares solution for a particular tree structure, because each structure corresponds to a different set of linear equations. Each tree structure was represented by a set of linear equations with each pairwise distance set equal to the sum of the lengths of the connecting branches.

## 2.3. Comparing metabolic, genetic and biogeographic distances

Manhattan metabolic distances between group pairs were compared to both genetic and biogeographic distances. To measure genetic distance, we used Tau as proposed by Kidd and Cavalli-Sforza [17]. Tau was calculated using 55 ancestry informative SNPs [26] that were

either good matches or reasonable proxies for 15 of the 17 groups in this study (Filipino and Hawaiian not available, S3, S4, and S5 Table). To measure geographic distance between ancestors of the 17 groups, we used the “Measure distance” function in Google maps. For each group, a single geographic point was defined (S6 Table) and distance (in km) between these points was recorded for the 136 group pairs (S4 Table). For groups synonymous with a current country name that country's capital was used as geo point. For example, the distance between Cambodian (Phnom Penh) and Guamanian (Hagåtña) was recorded at 4324 km. For several groups (Black, Hispanic, Middle Eastern, Native American, Other Southeast Asian), defining a geo point was based on literature searches and could only present a rough approximation of the geographic and historic origins of these populations. For Chinese infants, Guangzhou was selected as the geo point instead of Beijing due to significant numbers of Chinese immigrant populations to California from the southern provinces of China [27]. Correlation coefficients were calculated between metabolic, genetic and geographic distances, and to investigate outlier group pairs with larger variation between distance measures (S7 Table).

#### 2.4. Predicting metabolic ancestry and associated metabolites

Association to one of two ethnic groupings was estimated for individual newborns in a group pair ( $n = 136$  pairs) using supervised machine learning of metabolic screening data. For each group pair, newborns in each of the two ethnic groups were randomly divided into two equal-size sets with one half of the pair used for training and the other for testing. Ten-fold cross validation was used in the training set to select tuning parameter for the logistic regression model with L1 penalty (Lasso method) [28]. The tuning parameter was chosen as the largest value such that the error was within 1 standard error of the minimum. The performance of the model for predicting group association in the testing set (e.g., half of samples in a group pair blinded for analysis) was estimated using the area under the receiver operating characteristic curve (AUC). This process was repeated 20 times and the average AUC was recorded for each group pair (S8 Table). Metabolites selected using Lasso in the 20 replications were further analyzed with a focus on group pairs with smaller genetic ( $\tau = 0.05$ ) and larger metabolic differences ( $\text{AUC} > 0.7$ ).

#### 2.5. Statistical analysis

Statistical analyses were performed in R 4.0.3 (R-Core-Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Core Team; 2020 <https://www.R-project.org/>). Cohen's  $d$  was calculated with R package *effsize* [29]. Logistic regression with L1 penalty was performed with R package *glmnet* [30]. Figures were prepared with R packages *ggplot2* [31], *ggpubr* [32], and *ComplexHeatmap* [33].

### 3. Results

#### 3.1. Mapping metabolic distances between populations

Hierarchical clustering applied to compare 17 parent-reported ethnic groups based on differences in blood metabolite levels identified three large clusters (Fig. 1). Japanese, Filipino, Korean, Chinese and the Vietnamese groups emerged as one cluster, while other Asian groups (Cambodian, Laos, Other Southeast Asian) clustered with Black infants.

Samoan and Asian East Indian clustered separately. A third cluster included Native American, Hawaiian, Middle Eastern, White, Guamanian and Hispanic infants.

Higher medium-chain acylcarnitine levels (C8, C10, C10:1, C12, C12:1, C14:1, C14:2, C14OH, C16OH, C18OH) were found in the Japanese, Filipino and Korean groups. Chinese, Vietnamese, and Korean groups had lower levels of short-chain acylcarnitine (C2, C3, C3DC, C4, C5, C5DC, C5OH) and amino acid (arginine, glycine, leucine/isoleucine, ornithine, tyrosine, valine). The Samoan group had higher levels for 33 of 41 metabolites compared to average levels in all groups ( $P < 0.001$ , one side binomial test after Bonferroni correction). Overall, ethnicity-related differences were identified for 70.7% of the NBS metabolites (29 of 41, Cohen's  $d > 0.5$  in at least one group pair) (S1 Fig). The largest metabolite level differences were found for C3 between Japanese and Samoan infants (Cohen's  $d = 1.02$ ), for C2 between Chinese and Asian East Indian infants (Cohen's  $d = 0.96$ ), and for C5OH between White and Black infants (Cohen's  $d = 0.95$ ). Acylcarnitine levels showed higher variability between ethnic groups than amino acid levels ( $P < 1e-4$ , 10,000 times of permutation test of running-sum statistic) (S1 Fig). While the blood levels of most amino acids were higher than of acylcarnitines (Table S1), the Coefficient of Variation (standard deviation divided by mean, CV) was higher for most acylcarnitines compared to amino acids (S2 Fig).

Multidimensional scaling (MDS) was used to compare groups based on their metabolic data on a two-dimensional plane, such that the distances between points match the metabolic distances (Fig. 2A). The largest metabolic distances were seen between groups with ancestries from different continents and little admixture (Black and Japanese, Asian East Indian and Hispanic). In comparison, groups with some shared ancestry were metabolically more similar (Middle Eastern and White). Asian-ancestry groups (Cambodian, Chinese, Japanese, Korean, Laos, Other Southeast Asian, Vietnamese) were metabolically separate from most other groups, although the metabolic distances between some of these groups was large (Japanese and Vietnamese). Asian East Indian group was closer to Middle Eastern and White than to other Asian-ancestry groups. MDS-based analysis of genetic distances (Fig. 2B) showed similar associations but also differences such as overall smaller distances between the Asian-ancestry groups, and larger distance between Black and White groups to other groups.

Phylogenetic tree analysis (PTA) of Manhattan metabolic and Tau genetic distances between 15 matched ethnic groupings revealed both similarities and differences. Fig. 2C shows the best fit of a metabolic distance tree using overall length of the tree structure based on absolute values of the segment lengths. Using arbitrarily specified, random, and/or neighbor joining tree structures to start, we employed a search algorithm shown to improve the fit of an additive tree structure when applied to the Manhattan distances [34]. Eight Asian populations and Samoan infants were grouped together metabolically. Within the Asian groups, Chinese infants were closer to Vietnamese compared to Korean and Japanese. White and Middle Eastern, and Hispanic and Native American grouped at first and then the two group pairs were clustered. In comparison, genetic distance tree analysis (Fig. 2D) grouped seven Asian populations, Samoan and Guamanian infants, while the Asian East Indian group was closer to Middle Eastern and White. Notably, the genetic distance between the

Black and White newborn groups, and their genetic distances to most other groups, were much larger compared to the smaller metabolic distances between them. Finally, comparing metabolic distances for the 15 groups (Fig. 2A and C) to all 17 groups available (including Filipino and Hawaiian) showed similar associations, except of the Hispanic group which grouped differently (S3 Fig).

### 3.2. Correlation of metabolic, genetic and geographic distances

Pairwise comparison of metabolic, genetic, and geographic distances showed positive linear correlations (Fig. 3 and S7 Table). However, variation between distances was large with low correlation coefficients between metabolic and genetic distances ( $Cor = 0.29$ ), and between metabolic and geographic distances ( $Cor = 0.27$ ). To identify outlier group pairs that may have contributed to reduced correlations, a robust method based on first and third quartile of the data was used (S4 Fig). Two types of outliers were identified when comparing metabolic and genetic distances (Fig. 3A), which included group pairs with smaller genetic and larger metabolic distance (top left), and group pairs with larger genetic and smaller metabolic distance (bottom right). Most outliers in the top-left area involved the Japanese group, while most outliers in the bottom-right area involved the Black and White infant groups. Similarly, two types of outliers were identified when comparing metabolic and ancient geographic distances (Fig. 3B). Most group pairs with larger metabolic and smaller geographic distances were related to Japanese, Chinese and Asian East Indian groups, while outlier pairs with large ancient geographic and smaller metabolic distance involved the Hawaiian group. In comparison, the genetic and ancient geographic distances had a higher correlation coefficient ( $Cor = 0.54$ ). Outlier pairs with smaller geographic and larger genetic distances involved the White groups, while outliers with larger geographic and smaller genetic distances were associated with the Hispanic group (Fig. 3C).

### 3.3. Prediction of metabolic ancestry for ethnic group pairs

Performance of our machine learning model to infer association to one of two ethnic groups based on a newborns' metabolic profile was estimated using the area under the receiver operating characteristic curve (AUC). Higher AUC values indicated larger metabolic differences between two ethnic groups, which correlated with the model's ability to infer group association. The Guamanian group had overall lower AUC values (S8 Table) due to smaller sample size ( $n = 43$ , Table 1) and was removed from this analysis. Fig. 4 shows the model's performance for predicting ethnic group association for individuals in 16 groups ( $n = 120$  group pairs). Lower AUC for inferring group association was seen for metabolically more similar populations, such as Cambodian, Laos, and Other Southeast Asian groups, or between the Hawaiian, Filipino, Native American, and Hispanic groups. In contrast, high performance was seen for separating populations with larger ancestral geographic distances such as Black and Chinese ( $AUC = 0.96$ ), or Korean and Native American ( $AUC = 0.91$ ).

### 3.4. Identifying metabolite differences between ethnic groups

The machine learning model was run 20 times for all group pairs to find metabolites informative for inferring group association. S5 Fig shows the number of times a metabolite was selected for a group pair, and which metabolites were consistently selected or omitted across all pairs. A larger number of metabolites were repeatedly selected for group pairs

with larger genetic and ancient geographic distances and larger AUC (e.g., Black and Asian-ancestry pairs). In contrast, for metabolically more similar pairs (lower AUC), the model selected only a few metabolites and at fewer times (<10 times of 20 repeats) indicating lower predictive power. Additionally, we investigated which metabolites were selected by the model for group pairs with smaller genetic distances (Tau = 0.05) and large AUC (AUC = 0.7). This analysis was based on the hypothesis that metabolic differences in genetically similar populations located in the same geographic regions (e.g., babies born in California) may more likely be influenced by environmental variables. In most of the eight group pairs that met above thresholds, C10:1 was selected in all 20 repeats of the model. Leucine-Isoleucine, C5OH, C3, and C12:1 were selected in > 10 of 20 repeats in six out of the eight group pairs (S6 Fig).

#### 4. Discussion

Previous studies have shown that individuals tend to cluster genetically with others of the same ancestral geographic origins, and that genotyping a small number of SNPs is sufficient to accurately identify ancestry [7,8]. Here we studied whether ancestral relationships in 17 parent-reported ethnicity groupings could be identified using newborn metabolic screening data. Ethnicity-related differences were detected for 71% of NBS metabolites (29 of 41, Cohen's  $d > 0.5$ ) (S1 Fig) confirming results from a smaller cohort study [12]. Most amino acids had higher blood levels compared to acylcarnitines, while acylcarnitine levels showed higher variability compared to amino acids ( $P < 1e-4$ ) (S1-2 Fig). Samoan-ancestry newborns had the highest mean metabolite levels ( $P < 0.001$ ) compared to others, which may correlate with the prevalence of pregnancy obesity and excessive gestational weight gain among American Samoan women [35].

A metabolic distance measure was developed to compare the 17 ethnic groupings based on their metabolite level differences, and to correlate results to genetic distances available for 15 matched world populations (Fig. 2). Pairwise comparison of all groups showed a low positive correlation coefficient for metabolic and genetic distances. Inspection of individual group pairs revealed several outliers including pairs with larger genetic and smaller metabolic distances, or smaller genetic and large metabolic distances (Fig. 3A). We also searched for outlier pairs with lower correlation between metabolic and ancient geographic distances (Fig. 3B). The lower correlations between the three distance measures could be due to environmental variables with the following examples provided:

First, both Black and White groupings were outliers with larger genetic and smaller metabolic distances between each other and to others. Notably, all babies studied were born in California and it is possible that genetically distant groups have become more similar metabolically as they have interacted over a period of time resulting in smaller metabolic differences. Larger genetic and smaller metabolic distance were also found between Native American and White, which may be, in part, attributed to shifts in Native American diets and lifestyles after European colonization [36,37].

Second, the Chinese-ancestry group had short genetic distances to Korean (0.01), Japanese (0.014) and the Vietnamese (0.016) ancestry groups (S7 Table). In contrast, metabolic

distance between Chinese and Vietnamese ( $M = 23.24$ ) was only about two-third that of Chinese and Korean ( $M = 31.7$ ) and one-third that of Chinese and Japanese ( $M = 58.36$ ). Notably, the Chinese-ancestry and the Other Southeast Asian group had much larger genetic ( $\text{Tau} = 0.13$ ) and smaller metabolic ( $M = 38.2$ ) distances. The shorter metabolic distances between infants of Chinese, Vietnamese and Other Southeast Asian ancestry suggest that environmental factors (f.e., dietary preferences) could contribute to metabolic similarities between these groupings. Although Chinese people arrived at different times and from different regions in China, these findings correlate with reports that Chinese-ancestry populations living in California mostly speak Cantonese, which is the home language of Hong Kong and the southern coastal region of Guangdong in China and in other Asian countries [27,38].

Third, the Asian East Indian ancestry group was an outlier with larger metabolic distances and relative shorter ancient geographic distances to five Asian-ancestry groups (Chinese, Japanese, Korean, Laos, Vietnamese). The larger metabolic distances correlated with large genetic distances between these groupings. Notably, the Asian East Indian group had much shorter genetic and metabolic distances to Middle Eastern, Other Southeast Asian and White groups. The Asian East Indian group had also shorter metabolic distances to the Hawaiian group ( $\text{Tau}$  not available for Hawaiian). Although the Hawaiian group had one of the largest geographic distances, it was metabolically similar to several other groups indicating a complex population structure with influences from several other groupings [39].

Fourth, the group of Hispanic infants was an outlier with small genetic distances but large ancient geographic distance to several Asian-ancestry groups (e.g., Asian East Indian, Middle Eastern, Other Southeast Asian) when Mexico City was used as the ancient geographic point for this population (Fig. 3C). While genetic similarity generally decays with geographic distance, the relationship is often subtly distorted [40]. Although ancestors of many Hispanic infants born in California came from Mexico and neighboring Central American countries or territories, many have 60% or higher European ancestry [41], which could increase genetic distances to other groupings.

Following the study of population-level metabolic differences, we investigated whether ethnic group affiliation could be inferred from an individual's metabolic profile. This approach is similar in concept to the use of SNP genotypes when predicting an individual's ancestry [7,41]. To obtain ancestry estimates from metabolic profiles, we employed machine learning, which showed the best performance (higher AUC) for predicting one of two ethnicities for individuals in metabolically more distant group pairs (Fig. 4). For example, the model's ability for inferring metabolic ancestry was higher for Black or Chinese-ancestry infants ( $\text{AUC} = 0.96$ ) as compared to newborns with either Laotian or Cambodian ancestry ( $\text{AUC} = 0.52$ ).

Individual-level ancestry predictions correlated well with the population-level analysis. One exception was the Hawaiian-ancestry group with high metabolic similarity (low AUC) to newborns of Filipino, Hispanic, and Other Southeast Asian ancestry (Fig. 4), which was in contrast to their shorter metabolic distances (e.g., high metabolic similarity) to Native American, Middle Eastern and White populations (Figs. 1 and 2). We considered

the following factors contributing to these differences. First, metabolic distance between populations was estimated based on differences in mean metabolite levels between populations. Mean metabolite levels for a population was computed from metabolite level data across all individuals in that population. In contrast, metabolic ancestry was estimated based on an individual newborn's metabolic profile. Second, a potential source of bias for estimating metabolic distance (population-level) and metabolic ancestry (individual-level) could be related to population admixture. Based on recent census data, nearly a fourth (23.7%) of Hawaii's population identified as multiracial [42] with the five largest groups self-identifying as White (43.0%), Filipino (25.0%), Japanese (22.1%), Native Hawaiian (21.3%), and Chinese (14.1%). Although infants reported with multiple ethnicity categories were removed from our analysis, it is possible that individuals in the Hawaiian-ancestry grouping are admixed with Asian-ancestry populations. This could in part explain the metabolic similarities seen between Hawaiian and Filipino-ancestry individuals (AUC = 0.51), and the slightly lower metabolic ancestry predictions between Hawaiian and Japanese (0.78) compared to Hawaiian and Chinese (0.81) and Hawaiian and Korean (0.82) [43].

The machine learning approach developed here enables selection of candidate metabolites that could be informative for predicting ethnic group association. Our model selected several metabolites that could differentiate between individuals from group pairs with larger genetic and ancestral geographic distances (S5 Fig). More distant populations had larger differences in blood levels of specific metabolites but also in the overall number of metabolites with such differences. To uncover potential environmental influences on newborn metabolism, we searched for metabolites in group pairs with larger metabolic (AUC > 0.7) and smaller genetic distances (Tau < 0.05) (S6 Fig). Notably, the top-ranked metabolites identified in this analysis include primary NBS markers for the detection of Medium-chain Acyl-CoA Dehydrogenase Deficiency (C10:1), Very Long Chain Acyl CoA Dehydrogenase Deficiency (C12:1), 3-Methylcrotonyl CoA carboxylase deficiency (C5OH), Methylmalonic acidemia (C3), and Maple Syrup Urine Disease (Leucine-Isoleucine). We hypothesize that variable blood levels of these disease biomarkers in relation to ancestry could be associated with false-positive newborn screens in the respective populations [12].

Our study has several limitations. First, analysis was limited to 17 parent-reported ethnicity categories recorded by a public NBS program. Many families do not identify themselves belonging to any of these predefined categories and/or may affiliate with more than one of these or other categories. Additionally, population admixture is often unknown and combining individuals from diverse populations into a few broad categories (e.g., "Hispanic") disregards a multitude of cultural and ancestral identities [44]. Second, estimates for metabolic and for genetic distances were obtained from different samples, which could limit comparison between the two measures. Future studies should generate data concurrently for both measures. Third, selecting a single geographic point for ancestral locations has major limitations, particularly for larger geographic areas. Methods for higher biogeographic resolution of population origins [40,45] could improve analysis of the complex relationship with genetic and metabolic distances. Fourth, Manhattan metabolic distance between groups was estimated using 41 NBS metabolites and with each metabolite having the same weight. Future studies estimating metabolic distances between populations could include additional metabolites and evaluate weighted metabolite analysis.

## 5. Conclusions

Our work demonstrates the utility of population-level metabolomics discovery for capturing metabolic differences at birth. To estimate metabolic differences between ethnicity grouping, a metabolic distances measure was developed which is shown to be under the influence of genetic and of environmental factors. Such approaches provide new tools to study associations between metabolic phenotypes and genotypes, improve our understanding of metabolic diversity in human populations, and incorporate knowledge on ethnicity-related differences for blood metabolic biomarkers to minimize confounding factors in genetic disease screening.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

We thank Robin Cooley, Hao Tang, Steve Graham, Stanley Sciortino, Lisa Feuchtbaum and Robert Currier at California's Genetic Disease Screening Program (GDSP) for support of this project. We thank Else Roosevelt for help with mapping geographic distances. The data used in this study were obtained from the California Biobank Program (CBP) under SIS request 886. The California Department of Public Health is not responsible for the results or conclusions drawn by the authors of this publication. The data can be obtained by others after submitting a new request to the CBP coordinator. Requests to access these datasets should be directed to CaliforniaBiobank@cdph.ca.gov.

## Funding

This work was in parts funded by a grant from the National Institute of Child Health and Human Development (R01HD102537).

## Abbreviations:

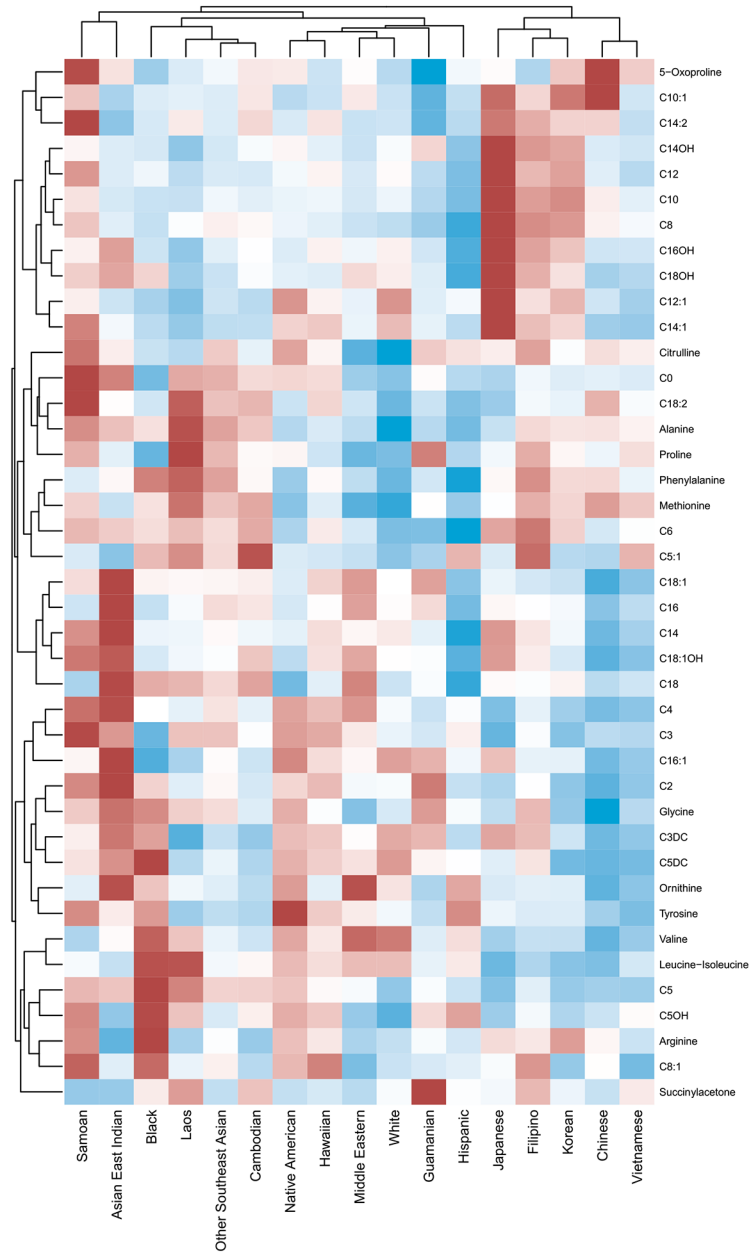
<b>NBS</b>	newborn screening
<b>MS/MS</b>	tandem mass spectrometry
<b>AUC</b>	area under the receiver operating characteristic curve

## References

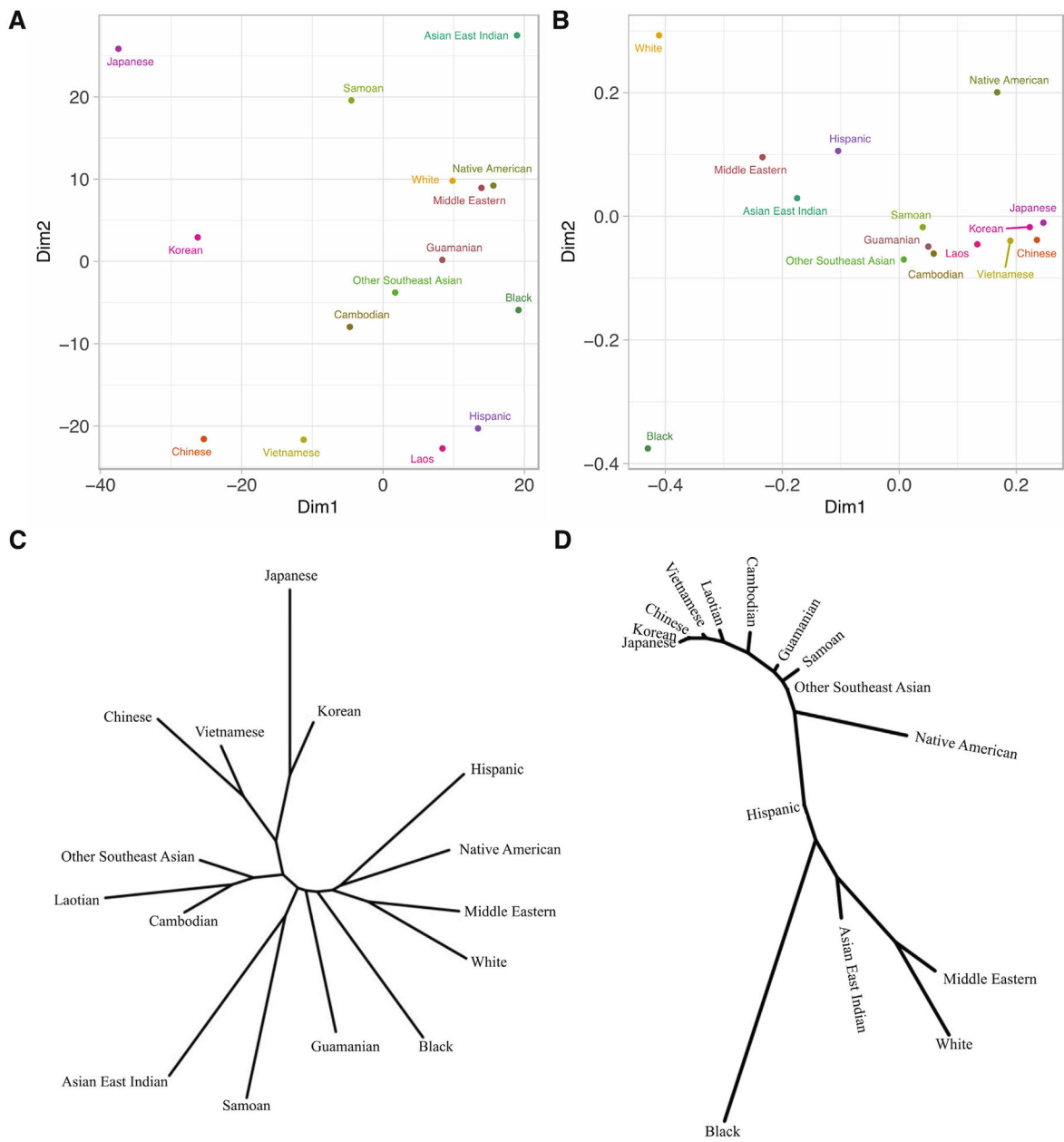
- [1]. Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, et al. , Ethnic-affiliation estimation by use of population-specific DNA markers, *Am. J. Hum. Genet* 60 (4) (1997) 957–964 , Epub 1997/04/01. [PubMed: 9106543]
- [2]. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. , Genetic structure of human populations, *Science* 298 (5602) (2002) 2381–2385 Epub 2002/12/21. [PubMed: 12493913]
- [3]. Jorde LB, Wooding SP, Genetic variation, classification and 'race', *Nat. Genet* 36 (11 Suppl) (2004) S28–S33 Epub 2004/10/28. [PubMed: 15508000]
- [4]. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG, Measures of human population structure show heterogeneity among genomic regions, *Genome Res.* 15 (11) (2005) 1468–1476 Epub 2005/10/28. [PubMed: 16251456]
- [5]. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, et al. , PCA-correlated SNPs for structure identification in worldwide human populations, *PLoS Genet.* 3 (9) (2007) 1672–1686 Epub 2007/09/26. [PubMed: 17892327]

- [6]. Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, et al. , Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies, *Am. J. Hum. Genet* 83 (4) (2008) 445–456 Epub 2008/09/27. [PubMed: 18817904]
- [7]. Sampson JN, Kidd KK, Kidd JR, Zhao H, Selecting SNPs to identify ancestry, *Ann. Hum. Genet* 75 (4) (2011) 539–553 Epub 2011/06/15. [PubMed: 21668909]
- [8]. Tang H, Quertermous T, Rodriguez B, Kardias SL, Zhu X, Brown A, et al. , Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies, *Am. J. Hum. Genet* 76 (2) (2005) 268–275 Epub 2004/12/31. [PubMed: 15625622]
- [9]. Cheillan D, Vercherat M, Chevalier-Porst F, Charcosset M, Rolland MO, Dorche C, False-positive results in neonatal screening for cystic fibrosis based on a three-stage protocol (IRT/DNA/IRT): should we adjust IRT cut-off to ethnic origin? *J. Inherit. Metab. Dis* 28 (6) (2005) 813–818 Epub 2006/01/26. [PubMed: 16435172]
- [10]. Giusti R, Elevated IRT levels in African-American infants: implications for newborn screening in an ethnically diverse population, *Pediatr. Pulmonol* 43 (7) (2008) 638–641 Epub 2008/05/27. [PubMed: 18500736]
- [11]. Peters C, Brooke I, Heales S, Ifederu A, Langham S, Hindmarsh P, et al. , Defining the newborn blood spot screening reference interval for TSH: impact of ethnicity, *J. Clin. Endocrinol. Metab* 101 (9) (2016) 3445–3449 Epub 2016/07/12. [PubMed: 27399348]
- [12]. Peng G, Tang Y, Gandotra N, Enns GM, Cowan TM, Zhao H, et al. , Ethnic variability in newborn metabolic screening markers associated with false-positive outcomes, *J. Inherit. Metab. Dis* 43 (5) (2020) 934–943 Epub 2020/03/28. [PubMed: 32216101]
- [13]. Ryckman KK, Berberich SL, Shchelochkov OA, Cook DE, Murray JC, Clinical and environmental influences on metabolic biomarkers collected for newborn screening, *Clin. Biochem* 46 (1–2) (2013) 133–138 Epub 2012/09/27. [PubMed: 23010448]
- [14]. Hall PL, Marquardt G, McHugh DMS, Currier RJ, Tang H, Stoway SD, et al. , Postanalytical tools improve performance of newborn screening by tandem mass spectrometry, *Genet. Med* 16 (12) (2014) 889–895 Epub 2014/05/29. [PubMed: 24875301]
- [15]. Clark RH, Kelleher AS, Chace DH, Spitzer AR, Gestational age and age at sampling influence metabolic profiles in premature infants, *Pediatrics* 134 (1) (2014) e37–e46 Epub 2014/06/11. [PubMed: 24913786]
- [16]. Peng G, Tang Y, Cowan TM, Enns GM, Zhao H, Scharfe C, Reducing false-positive results in newborn screening using machine learning, *Int. J. Neonatal Screen* 6 (1) (2020) 10.3390/ijns6010016, Epub 2020/03/20.
- [17]. Kidd KK, Cavalli-Sforza LL, The role of genetic drift in the differentiation of Icelandic and Norwegian cattle, *Evolution* 28 (3) (1974) 381–395 Epub 1974/09/01. [PubMed: 28564858]
- [18]. DeBerardinis RJ, Thompson CB, Cellular metabolism and disease: what do metabolic outliers teach us? *Cell* 148 (6) (2012) 1132–1144 Epub 2012/03/20. [PubMed: 22424225]
- [19]. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, et al. , An atlas of genetic influences on human blood metabolites, *Nat. Genet* 46 (6) (2014) 543–550 Epub 2014/05/13. [PubMed: 24816252]
- [20]. Kettunen J, Demirkan A, Wurtz P, Draisma HH, Haller T, Rawal R, et al. , Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA, *Nat. Commun* 7 (2016) 11122 Epub 2016/03/24. [PubMed: 27005778]
- [21]. Lotta LA, Pietzner M, Stewart ID, Wittemans LBL, Li C, Bonelli R, et al. , A cross-platform approach identifies genetic regulators of human metabolism and health, *Nat. Genet* 53 (1) (2021) 54–64 Epub 2021/01/09. [PubMed: 33414548]
- [22]. Kim YJ, Go MJ, Hu C, Hong CB, Kim YK, Lee JY, et al. , Large-scale genome-wide association studies in east Asians identify new genetic loci influencing metabolic traits, *Nat. Genet* 43 (10) (2011) 990–995 Epub 2011/09/13. [PubMed: 21909109]
- [23]. Hebbar P, Abubaker JA, Abu-Farha M, Alsmadi O, Elkum N, Alkayal F, et al. , Genome-wide landscape establishes novel association signals for metabolic traits in the Arab population, *Hum. Genet* 140 (3) (2021) 505–528 Epub 2020/09/10. [PubMed: 32902719]

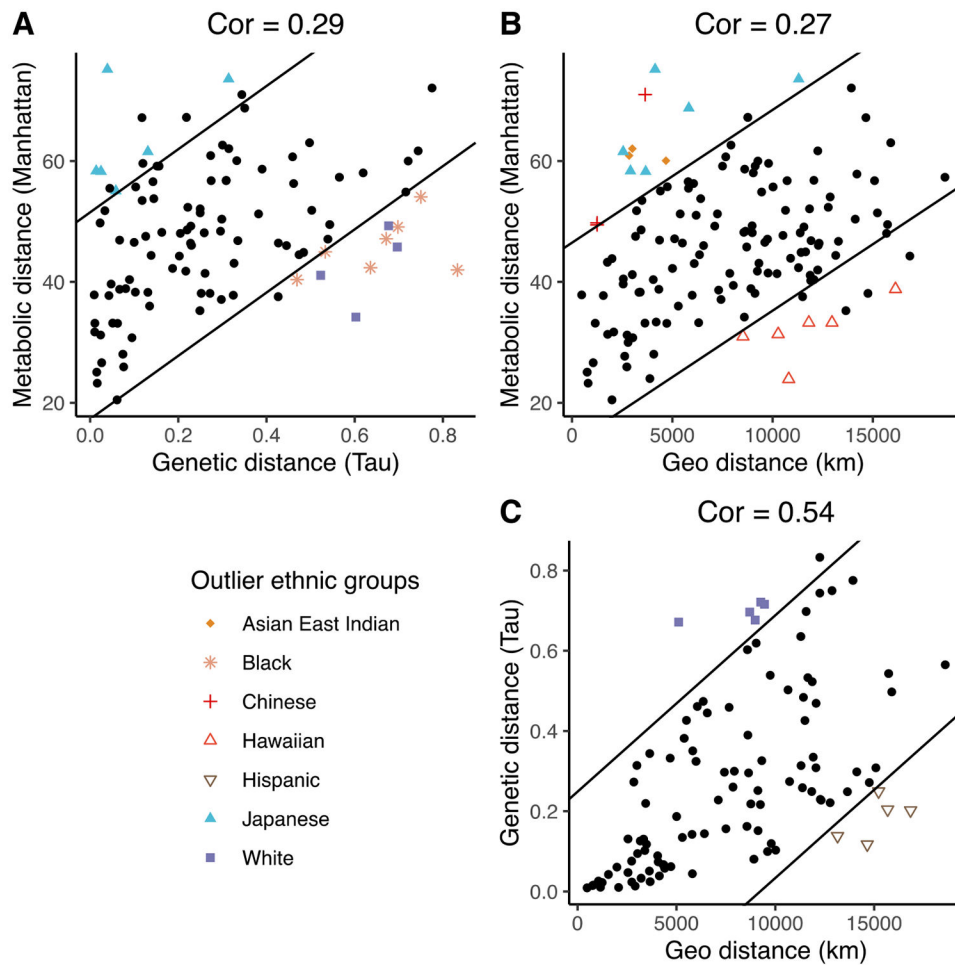
- [24]. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. . The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies, *J. Clin. Epidemiol* 61 (4) (2008) 344–349 Epub 2008/03/04. [PubMed: 18313558]
- [25]. Saitou N, Nei M, The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol* 4 (4) (1987) 406–425 Epub 1987/07/01. [PubMed: 3447015]
- [26]. Kidd KK, Soundararajan U, Rajeevan H, Pakstis AJ, Moore KN, Ropero-Miller JD, The redesigned Forensic Research/Reference on Genetics-knowledge base, FROG-kb, *Forensic Sci. Int. Genet* 33 (2018) 33–37 Epub 2017/11/28. [PubMed: 29175726]
- [27]. North Cooc GL, Who are “Chinese” speakers in the United States?: examining differences in socioeconomic outcomes and language identities, *AAPINexus Policy Pract. Community* 15 (2017) 137–164.
- [28]. Tibshirani R, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B Methodol* 58 (1) (1996) 267–288.
- [29]. Torchiano M, effsize: Efficient Effect Size Computation, 2020 10.5281/zenodo.1480624.
- [30]. Friedman J, Hastie T, Tibshirani R, Regularization paths for generalized linear models via coordinate descent, *J. Stat. Softw* 33 (1) (2010) 1–22 , Epub 2010/09/03. [PubMed: 20808728]
- [31]. Wickham H, ggplot2: Elegant Graphics for Data Analysis, Springer, 2016.
- [32]. Kassambara A, ggpubr: 'ggplot2' Based Publication Ready Plots, 2020.
- [33]. Gu Z, Eils R, Schlesner M, Complex heatmaps reveal patterns and correlations in multidimensional genomic data, *Bioinformatics* 32 (18) (2016) 2847–2849 Epub 2016/05/22. [PubMed: 27207943]
- [34]. Kidd KK, Sgaramella-Zonta LA, Phylogenetic analysis: concepts and methods, *Am. J. Hum. Genet* 23 (3) (1971) 235–252, Epub 1971/05/01. [PubMed: 5089842]
- [35]. Hawley NL, Johnson W, Hart CN, Triche EW, Ah Ching J, Muasau-Howard B, et al. , Gestational weight gain among American Samoan women and its impact on delivery and infant outcomes, *BMC Pregnancy Childbirth* 15 (2015) 10 Epub 2015/02/04. [PubMed: 25643752]
- [36]. FRANK LE, History on a Plate: How Native American Diets Shifted after European Colonization, 2020.
- [37]. Schulz LO, Bennett PH, Ravussin E, Kidd JR, Kidd KK, Esparza J, et al. , Effects of traditional and western environments on prevalence of type 2 diabetes in Pima Indians in Mexico and the U.S, *Diabetes Care* 29 (8) (2006) 1866–1871 Epub 2006/07/29. [PubMed: 16873794]
- [38]. Chinese Immersion School (CIS) at De Avila. Why Cantonese First? <https://www.sfusd.edu/school/chinese-immersion-school-cis-de-avila/about/why-cantonese-first>
- [39]. United States Census Bureau. QuickFacts Provides Statistics for All States and Counties, and for Cities and Towns with a Population of 5,000 or More. 2019 [cited 2021 August 16]. Available from: <https://www.census.gov/quickfacts/HI>.
- [40]. Peter BM, Petkova D, Novembre J, Genetic landscapes reveal how human genetic diversity aligns with geography, *Mol. Biol. Evol* 37 (4) (2020) 943–951 Epub 2019/11/30. [PubMed: 31778174]
- [41]. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL, The genetic ancestry of African Americans, Latinos, and European Americans across the United States, *Am.J. Hum. Genet* 96 (1) (2015) 37–53 Epub 2014/12/23. [PubMed: 25529636]
- [42]. Research and Economic Analysis Division DoB, Economic Development and Tourism, STATE OF HAWAII, Demographic, Social, Economic, and Housing Characteristics for Selected Race Groups in Hawaii, 2018.
- [43]. State of Hawaii, Demographic, Social, Economic, and Housing Characteristics for Selected Race Groups in Hawaii, The Department of Business EDaT, editor 2018.
- [44]. Popejoy AB, Too many scientists still say Caucasian, *Nature* 596 (7873) (2021) 463 Epub 2021/08/26.
- [45]. Elhaik E, Tatarinova T, Chebotarev D, Piras IS, Maria Calo C, De Montis A, et al. . Geographic population structure analysis of worldwide human populations infers their biogeographical origins, *Nat. Commun* 5 (2014) 3513 Epub 2014/05/02. [PubMed: 24781250]



**Fig. 1.** Association between ethnic groups based on metabolic screening data. Hierarchical clustering of differences in mean metabolite levels ( $n = 41$ ) between ethnic groupings ( $n = 17$ ) is based on Manhattan distance matrix (S1 Table). Heatmap colors indicate metabolite levels ranging from lower (blue) to higher (red) for the respective groups. Several larger population clusters are visible including a cluster of Japanese, Filipino, Korean, Chinese and Vietnamese-ancestry groupings.



**Fig. 2.** Mapping metabolic and genetic distances between ethnic groups. Multidimensional scaling (MDS) and Phylogenetic tree analysis (PTA) was used to visualize proximity between ethnic groupings based on metabolic and genetic distances (S4 Table). MDS plots show (A) Manhattan metabolic distance and (B) Tau genetic distance, and PTA visualizes (C) Manhattan metabolic distance and (D) Tau genetic distance.



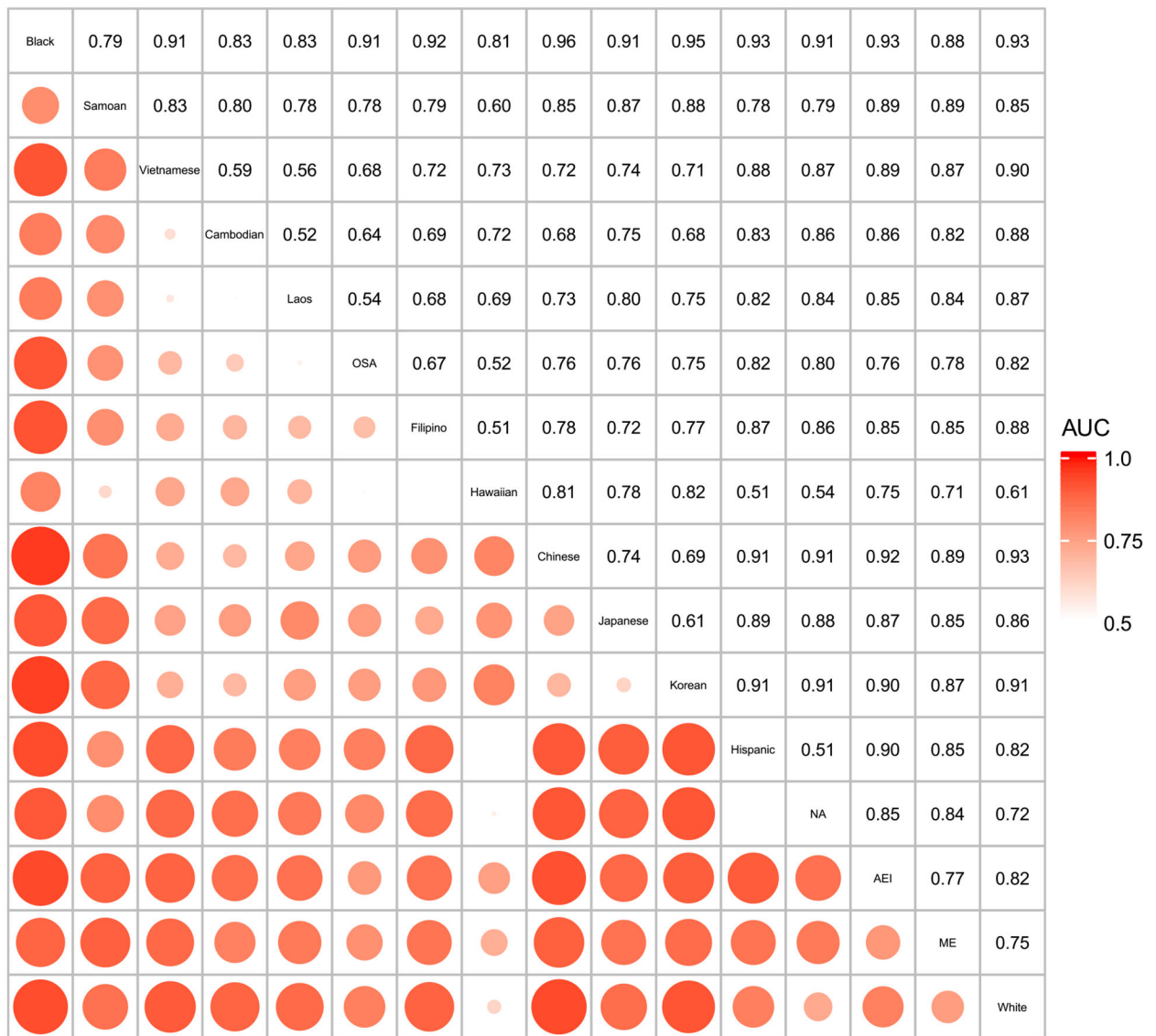
**Fig. 3.** Correlation of metabolic, genetic and ancient geographic distances between ethnic group pairs. Pairwise correlation of ethnic groups (e.g., each dot represents a group pair) based on (A) Manhattan metabolic and Tau genetic distance, (B) Manhattan metabolic and ancient geographic distance, and (C) Tau genetic and geographic distance (S7 Table). Color labels indicate seven ethnic groups that are more commonly found among outlier group pairs located outside of the two solid lines.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



**Fig. 4.** Prediction of ethnic group association for individual metabolic profiles. A machine learning classifier was used to infer one of two ethnicities for individuals in each group pair based on metabolic screening data. Higher AUC values (top right triangle) indicate larger metabolic differences between a group pair with higher accuracy for inferring group affiliation (S8 Table). The color and size of circles correspond to AUC values. AEI: Asian East Indian. ME: Middle Eastern. NA: Native American. OSA: Other Southeast Asian.

**Table 1**

Participants and size of ethnic groupings.

<b>Ethnicity</b>	<b>Sample size</b>	<b>Percentage</b>
Asian East Indian	10,361	2.64%
Black	24,524	6.25%
Cambodian	915	0.23%
Chinese	22,162	5.65%
Filipino	8147	2.08%
Guamanian	43	0.01%
Hawaiian	244	0.06%
Hispanic	188,833	48.13%
Japanese	1014	0.26%
Korean	3452	0.88%
Laos	413	0.11%
Middle Eastern	5479	1.40%
Native American	556	0.14%
Other Southeast Asian	1871	0.48%
Samoan	399	0.10%
Vietnamese	5042	1.29%
White	118,848	30.29%

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript