

RESEARCH ARTICLE

Open Access



Evaluating genetic ancestry and self-reported ethnicity in the context of carrier screening

Roman Shraga¹, Sarah Yarnall², Sonya Elango³, Arun Manoharan², Sally Ann Rodriguez², Sara L. Bristow^{1,10}, Neha Kumar², Mohammad Niknazar¹, David Hoffman⁴, Shahin Ghadir⁵, Rita Vassena⁶, Serena H. Chen⁷, Avner Hershlag⁸, Jamie Grifo⁹ and Oscar Puig^{1*} 

Abstract

Background: Current professional society guidelines recommend genetic carrier screening be offered on the basis of ethnicity, or when using expanded carrier screening panels, they recommend to compute residual risk based on ethnicity. We investigated the reliability of self-reported ethnicity in 9138 subjects referred to carrier screening. Self-reported ethnicity gathered from test requisition forms and during post-test genetic counseling, and genetic ancestry predicted by a statistical model, were compared for concordance.

Results: We identified several discrepancies between the two sources of self-reported ethnicity and genetic ancestry. Only 30.3% of individuals who indicated Mediterranean ancestry during consultation self-reported this on requisition forms. Additionally, the proportion of individuals who reported Southeast Asian but were estimated to have a different genetic ancestry was found to depend on the source of self-report. Finally, individuals who reported Latin American demonstrated a high degree of ancestral admixture. As a result, carrier rates and residual risks provided for patient decision-making are impacted if using self-reported ethnicity.

Conclusion: Our analysis highlights the unreliability of ethnicity classification based on patient self-reports. We recommend the routine use of pan-ethnic carrier screening panels in reproductive medicine. Furthermore, the use of an ancestry model would allow better estimation of carrier rates and residual risks.

Keywords: Ancestry informative markers, Genetic risk, Carrier screening, Ethnicity

Background

High throughput technologies like microarrays and next generation sequencing permit the simultaneous interrogation of multiple disease-causing mutations in many genes, allowing expanded carrier screening for almost any disease to be routinely used in medical practice. In contrast, some professional guidelines still recommend carrier screening for certain mutations based on ethnicity [1–3] and physicians commonly use a patient's self-reported ethnicity to help determine which genetic conditions or mutations to test. When society guidelines recommend expanded carrier screening, they stress that mutation frequencies should be known in the population

being tested, so that residual risk in individuals who test negative can be assessed accurately [4, 5].

Ethnicity is defined as the membership in a specific group sharing cultural, religious, or racial traits [6], while genetic ancestry refers to the variations in genomic structure among different populations, or the genotypes an individual may have as a result of their ancestors [7]. Because the definition of ethnicity is predicated on shared culture, there is a level of self-identification involved, which may obscure genetic ancestry. For example, an individual may not know about or self-identify with a certain group, despite having an ancestor from it. This issue is further highlighted in admixed populations such as Latinos or African Americans. Different African American groups across the United States have been found to have varying proportions of European

* Correspondence: oscar@phosphorus.com

¹Phosphorus, Inc, 25 West 26th St, New York, NY 10010, USA

Full list of author information is available at the end of the article



contribution to their genome with estimates ranging from 7 to 23% [8]. This genetic heterogeneity means that self-reported ethnicity may not give a complete picture of genetic composition [9].

The incidence of several monogenic diseases varies widely across different populations: Asian Americans account for over 50% of thalassemia cases in the US [10], African Americans are more likely to develop sickle cell anemia [11], and 10-40% of Ashkenazi Jews carry recessive mutations for known Mendelian diseases [12]. As a result, carrier screening programs in the past were selectively targeted to certain racial or ethnic groups [13]. Beyond carrier screening and monogenic diseases, multiple associations between genetic ancestry and clinical outcomes have been observed. For example, lung function in African Americans is inversely correlated with the percentage of African ancestry [14, 15], coronary artery calcium and carotid intima media thickness in African Americans and Hispanics vary depending on genetic ancestry and the percentage of European admixture [16], and diabetes among Native Americans is correlated with European admixture [17].

In order to determine the reliability of self-reported ethnicity for clinical decision-making, we compared two sources of self-reported ethnicity, written and verbal, with genetic ancestry as estimated by a statistical model, from a cohort of 9138 patients undergoing genetic carrier screening. Additionally, the relationship between disease carrier status and genetic ancestry within self-reported ethnic groups was studied. Our results indicate that self-reported ethnicity is not a reliable source on which to base clinical decisions.

Methods

Study population and self-reported ethnicity

Our study population included 9138 patients undergoing genetic carrier screening, who were referred from fertility specialists, obstetricians/gynecologists, and genetic counselors. Eighty-six percent of the samples were from US clinics, eighty-six percent from Spanish clinics and the remainder from other countries. Before undergoing in vitro fertilization, couples in fertility clinics were offered CarrierMap (Cooper Genomics, Inc.), a genetic test that determines carrier status for 311 autosomal recessive or X-linked genetic diseases. Informed consent to perform this research was obtained during sample requisition, and confirmed during the genetic counseling session that all patients are offered when CarrierMap results are reported to them. All data presented here is de-identified (HHS 45 CFR part 46.101(b)(4)).

Participants' ethnicities were self-reported at two separate points in the testing process. The initial report was gathered on the test requisition form, where patients were asked to select all ethnicities that apply from the following list of options (chosen by reviewing the literature): African; East

Asian; European; French Canadian; Jewish; Latin American; Mediterranean; Middle Eastern; Native American; South Asian; Southeast Asian; Other (see Additional file 1: for test requisition form). There was also space next to "Other" to write in a response. These responses were mapped to the appropriate category when possible (e.g. Caucasian/White mapped to European). Any participant who selected "Other" without writing in a clarification was excluded from analysis.

The second self-report was made during post-test consultation with a genetic counselor. Standard counseling protocol includes the collection of a complete family history, during which participants were asked to identify their race/ethnicity or the origin of their family. One or more ethnicities from the aforementioned options were then selected based upon the patient's report. This second source of ethnicity was included in order to determine if there is a difference in what is self-reported depending on the collection technique.

In cases where patients opted out of counseling or were unreachable, a "family history" ethnicity was not generated and the patients were not considered in that part of the analysis. These patients were, however, still included in the comparison between "requisition form" ethnicity and genetic ancestry.

Genotyping

DNA was extracted from blood or saliva samples, and genomic data was analyzed via Illumina's Infinium CoreExome-24 v1.0 and v1.1 (catalog ID WG-330-2014 and WG-331-1111, Illumina Inc., San Diego, CA) genotyping platform using standard protocols recommended by the manufacturer.

Determination of genetic ancestry

In order to estimate the genetic ancestry of our samples, we modified the Expectation Maximization (EM) Algorithm used in FRAPPE [18, 19]. We assumed a paradigm where an individual's genome is divided into segments of different ancestral origin from a set of geographic regions. Our goal was to estimate the proportion of ancestral origin from each geographic region for each individual. Instead of uncovering the ancestral populations from our samples directly, we chose to use an independent source of population allele frequencies to predefine geographic regions at the continental level. It should be noted that this model of genetic ancestry is not intended to fully describe the genetic structure of the human species, as there are significant genetic differences within continental groups. Moreover, there are populations that may be ill described by the geographic groups included in the model. Despite these limitations, the genetic ancestry model is useful as a statistical tool to investigate self-reported ethnic labels in the context of clinical decision-making.

Estimation of allele frequencies

We obtained allele frequencies for all populations and SNPs in the ALFRED database [20]. This database groups populations by the geographic region where the samples originate. In order to ensure complete allele frequency information, we removed any population that had allele frequency information for fewer than 90% of the total SNPs in the database. Next, we removed populations known to be heterogeneous (African Americans, Hispanics, etc.) and populations that are rare or that could not be easily classified into a single continental group (Yakut, Hezhe). This process left us with a total of 44 populations (Additional file 2: Table S1). These 44 populations can be further grouped into 8 continental groups: African (AF), Central Asian (CA), East Asian (EA), European (EU), Middle Eastern (ME), Native American (NA), Native Oceanian (NO), and South Asian (SA).

Next, we filtered the list of markers to include only biallelic SNPs existing in Illumina’s Infinium CoreExome-24. We also removed any SNP with a call rate lower than 99.9%. Finally, since the EM algorithm assumes independence of each locus, we used Plink’s pairwise independence command with an R^2 of 0.5 and a window size of 1000 to get a final list of SNPs that are in linkage equilibrium [21]. The final set contained 147,550 SNPs.

Marker selection

In order to reduce the full list of SNPs to only those that are informative of ancestry (Ancestry Informative Markers, AIMS), we used Wright’s F_{st} measure [22]. Instead of considering each population separately, we calculated allele frequencies at the continental level by averaging across populations. Next, we used a two-stage procedure to pick AIMS. First, in order to get a global set of markers, we considered the eight continental groups and picked SNPs that maximize F_{st} across all groups. Next, in order to ensure a balanced set of AIMS and to distinguish between more closely related groups, we enriched our set with extra SNPs that can differentiate specific groups. We did this by finding SNPs with maximum pairwise F_{st} , considering each pair of continental groups separately. This process generated a total of 1142 AIMS (Additional file 2: Tables S2 and S3).

Genetic ancestry prediction model

The model was formalized in a similar manner to FRAPPE and ADMIXTURE [18, 19]. For a large number of individuals (I) and a set of six ancestral populations (K), a vector $Q_i = (q_{i1}, \dots, q_{ik})$ is estimated for each individual (i). Each coordinate in the vector Q_i is an estimate of the probability that a random allele from i originates from population k . Our dataset contains genotypes for all i at the 1142 AIMS we selected (M). Each individual’s genotype is captured in a column vector g_i ,

where each entry g_{im} holds the number of copies of “allele 1” at marker m for individual i . The choice of which allele is “allele 1” is arbitrary as long as it is consistent across samples. Departing from FRAPPE, we also have a computed matrix of allele frequencies across populations (F). In this matrix, element f_{km} holds the allele frequency of marker m in population k .

The log-likelihood of each sample’s vector Q_i is independent and equal to

$$L(Q_i) = \sum_m \left\{ g_{im} \ln \left[\sum_k q_{ik} f_{km} \right] + (2 - g_{im}) \ln \left[\sum_k q_{ik} (1 - f_{km}) \right] \right\}$$

The Expectation Maximization algorithm updates the parameters via the maximization step:

$$q_{ik}^{n+1} = \frac{1}{2M} \sum_m [g_{im} a_{imk}^n + (2 - g_{im}) b_{imk}^n]$$

and the expectation step:

$$a_{imk}^n = \frac{q_{ik}^n f_{km}}{\sum_p q_{ip}^n f_{pm}}, \quad b_{imk}^n = \frac{q_{ik}^n (1 - f_{km})}{\sum_p q_{ip}^n (1 - f_{pm})}$$

These steps are repeated until the estimation converges. Namely, once:

$$L(Q^{n+1}) - L(Q^n) < \epsilon$$

Since we are not estimating the population allele frequencies, we can choose strict stopping criteria ($\epsilon = 0.01$) and still retain an acceptable computation time.

The set of 1142 AIMS was tested on 1000 Genomes Project data by processing genotype data available for 2504 samples (Additional file 2: Table S4) [23] using the algorithm described above. Genotype data for the 9138 subjects in this study is available upon request or at dbGAP (accession number phs001482.v1.p1). Scripts used to select the 1142 AIMS and to run the Genetic Ancestry algorithm as well as the Supplementary Data are available at <https://github.com/P15/phosphorus-public/releases/tag/v1.0.0>.

Results

The goal of this study was to determine the utility of self-reported ethnicity data for clinical decision-making in the context of genetic carrier screening. In the first step, we selected a set of SNPs that could accurately determine continental genetic ancestry in our patient population. We obtained SNP frequencies from the ALFRED database [24], and through an iterative process, we determined a set of SNPs that could separate the continental groups selected (Additional file 2: Tables S1 and S2). Additional file 3: Figure S1 plots the first two principal components of the 44 subpopulations across the 1142 selected SNPs, and shows that six of the eight

continental groups are well separated. The Middle Eastern and Central Asian groups are closely related to the European and South Asian groups, respectively, and require an extra set of markers to properly estimate population divergences [25, 26] (Additional file 3: Figure S1). For this reason, we chose not to consider these two groups as separate ancestral populations and removed them from the final estimation.

Model validation

In order to validate the genetic ancestry model, we applied it on a set of 2504 samples with known origin from the 1000 Genomes Project [23] (Fig. 1 and Additional file 2: Table S5). The 1000 Genomes Project groups samples by geographic origin and insures that all four grandparents of each sample come from the same group. As expected, the maximum ancestral component of each sample matched the known origin. We conclude that this set of 1142 SNPs correctly estimates continental ancestry in the included populations. These results also validate our approach of using pre-computed population allele frequencies.

Ethnicity reported on requisition form vs. family history

We compared self-reported ethnicity from two distinct sources: first, from a requisition form on which patients were asked to check all ethnicities that apply from a provided list and second, from a detailed family history intake performed by a certified genetic counselor. For each ethnic group, we counted the number of patients that selected it on the requisition form, the number of patients that identified it during consults, and the number

of patients that did both (Table 1). We excluded all patients that selected 'Other' on the requisition form.

Some of the groups had very consistent patterns of identification across both self-reported sources. For example, 92.7% of patients that selected African on the requisition form identified having African ancestry during the consult, while 95.5% of patients that identified having African ancestry during the consult also selected African on the requisition form. Other groups showed more complex patterns of self-identification. For example, 89.6% of patients that selected Mediterranean on the requisition form also identified it during the consult, while only 30.3% of patients that reported having some Mediterranean ancestry during the consult actually selected it on the requisition form (p -value = $2.63E-64$ using McNemar's test). Similarly, there were significant discrepancies in the requisition form vs. consult for Native American, Jewish and European ethnicities. These differences suggest that the manner in which self-reported ethnicity is collected affects what is reported.

Self-reported ethnicity vs. genetic ancestry

In order to see the relationship between self-reported and genetic ancestry, we first considered samples that marked only a single ethnicity on the requisition form. We applied our genetic ancestry prediction model to all of these samples and noted the maximum ancestral group (Fig. 2a and Additional file 2: Table S6). We saw a high level of agreement among African and European samples. Of the samples that marked African on the requisition form, 97.2% were predicted to have majority African ancestry by the model. Similarly, of the samples that marked European, 99.3% were predicted to have majority European ancestry by the model. These results are in line with what has been previously reported [27].

The results were quite different when considering Asian continental groups. Of patients that marked only Southeast Asian on the requisition form, 27.5% were predicted to have majority South Asian ancestry by the model, instead of East Asian as expected. At the same time, 8.7% of patients that marked only South Asian on the requisition form were predicted to have majority East Asian ancestry. While this result is consistent with previously published work that shows that the self-reported label "Asian" is concordant with some kind of Asian genetic ancestry, it also suggests that there may be confusion among patients about the distinction between the South, East, and Southeast Asian sub-groups [28].

In order to determine how the source of self-reported ethnicity impacts the genetic ancestry estimate, we repeated this analysis, but this time looked at ethnicity collected during genetic consults. Again, we considered samples that reported only a single ethnicity or country of origin during their consult

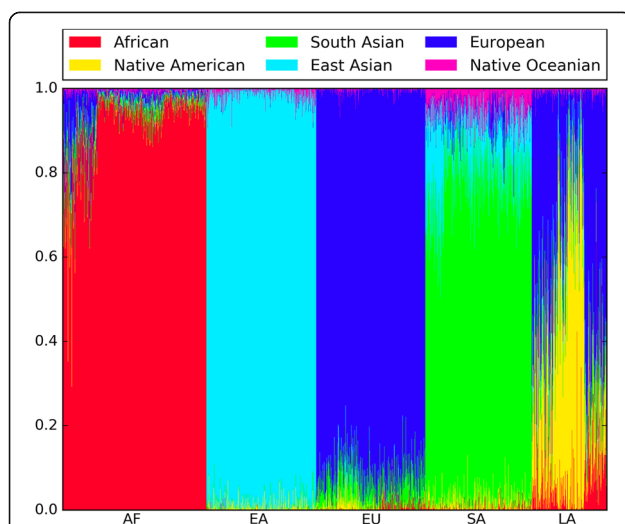


Fig. 1 1000 Genomes Project validation results. Each individual is represented as a thin vertical line, where each color shows the proportion of ancestry predicted from each continental group. Individuals are grouped by population: African (AF), East Asian (EA), European (EU), South Asian (SA), and Latin American (LA)

Table 1 Self-reported ethnicity reported on requisition forms vs. family history discussion

Ethnicity	Form	Consult	Both	% form that identified during consult	% consult that identified on form	<i>p</i> -value
African	206	200	191	92.7%	95.5%	0.307
East Asian	264	260	258	97.7%	99.2%	0.289
European	1815	1854	1708	94.1%	92.1%	1.67E-02
Jewish	403	431	380	94.3%	88.2%	1.52E-03
Latin American	333	332	318	95.5%	95.8%	1
Mediterranean	134	396	120	89.6%	30.3%	2.63E-64
Native American	70	122	64	91.4%	52.5%	9.03E-12
South Asian	139	140	121	87.1%	86.4%	1
South East Asian	128	114	99	77.3%	86.8%	0.049

Each row shows counts of the number of patients that selected each ethnicity on the requisition form, reported it during consult, and did both. Additionally the overlap proportions are shown. *P*-values were computed using McNemar's test to assess if the proportions are significantly different

and noted the maximum ancestral group predicted by the model (Additional file 2: Table S6). Once again, we saw a high level of agreement between self-reported and predicted ancestry among European and African samples. In contrast to the requisition form, the self-reports from consults among Asian continental groups were mostly concordant with the genetic ancestry estimates. Only 5.6% of patients who self-reported Southeast Asian during the consult were predicted to be have majority South Asian ancestry, with the remaining 93.3% having majority East Asian ancestry. This difference in the genetic ancestry estimates between the two sources of self-reported ethnicity suggests that people may alter what they report depending on the manner in which the report is collected.

Genetic ancestry prediction in admixed populations

We investigated the predicted genetic ancestry of populations with known admixture (Table 2). Patients that marked only African on the requisition form had on average 79.3% African ancestry and 12.4% European ancestry. Although our samples may include individuals who immigrated from Africa recently, this admixture proportion is in line with previously reported estimates of African Americans [27]. Notably, we found large variability in the predicted proportion of African ancestry in self-identified Africans: from a high of 99% African ancestry to a low of just 1.5% (Fig. 2b).

The highest variability was found in patients who marked Latin American on the requisition form. On average, these patients had 52.1% European, 24.4% Native American, and 12.0% African ancestry. These numbers

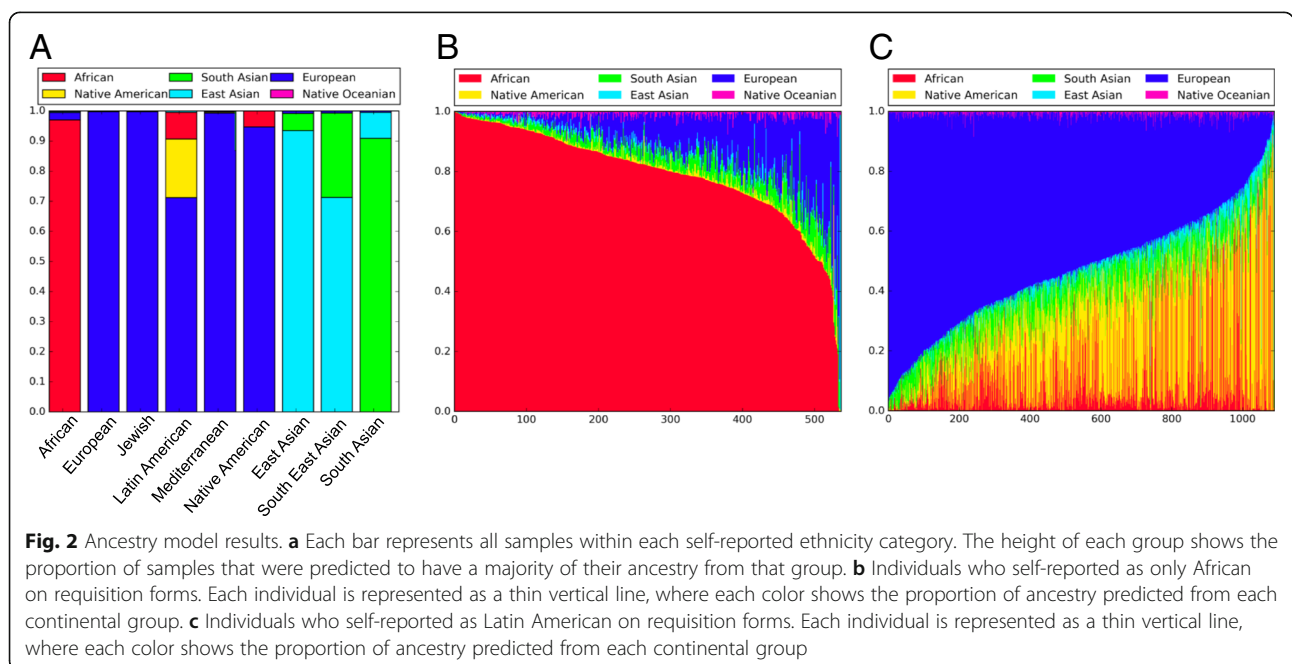


Table 2 Mean genetic ancestry component by self-reported ethnicity on requisition form. Each row shows samples based on self-reported ethnicity on requisition forms. Each column shows the average plus minus two standard errors of predicted ancestry proportion across all samples in that self-reported category

Requisition Form Ethnicity	African	East Asian	European	Native American	Native Oceanian	South Asian
African	79.3% + -1.5%	1.7% + -0.6%	12.4% + -1.1%	1.3% + -0.3%	1.2% + -0.2%	4.4% + -0.4%
East Asian	0.1% + -0.1%	91.7% + -1.7%	1.2% + -0.5%	1.2% + -0.2%	1.2% + -0.2%	4.9% + -1.4%
European	0.8% + -0.1%	0.8% + -0.2%	90.5% + -0.3%	1.1% + -0.1%	0.5% + -0.1%	6.7% + -0.2%
Jewish	3.0% + -0.3%	1.2% + -0.2%	79.9% + -0.7%	1.0% + -0.2%	1.0% + -0.1%	14.1% + -0.6%
Latin American	12.0% + -1.0%	3.3% + -0.4%	52.1% + -1.2%	24.4% + -1.2%	1.0% + -0.1%	7.5% + -0.3%
Mediterranean	2.6% + -0.5%	0.9% + -0.3%	86.4% + -1.0%	1.4% + -0.4%	0.7% + -0.1%	8.2% + -0.6%
Native American	7.3% + -9.2%	0.7% + -0.5%	81.4% + -10.2%	3.8% + -3.5%	0.6% + -0.4%	6.5% + -1.8%
South Asian	0.5% + -0.2%	18.8% + -2.5%	4.8% + -0.8%	2.0% + -0.3%	4.0% + -0.4%	70.2% + -2.3%
South East Asian	0.5% + -0.3%	68.3% + -4.3%	3.0% + -0.8%	1.7% + -0.3%	3.4% + -0.4%	23.4% + -3.9%
African & European	36.4% + -10.2%	1.1% + -0.6%	54.8% + -10.6%	0.8% + -0.4%	1.0% + -0.5%	6.1% + -1.5%
African & Latin American	55.0% + -12.6%	2.4% + -1.1%	26.4% + -8.0%	6.3% + -4.9%	1.5% + -0.8%	8.7% + -2.6%
East Asian & European	0.4% + -0.3%	44.3% + -4.8%	41.8% + -5.1%	3.8% + -2.0%	2.1% + -0.7%	7.9% + -2.5%
European & Latin American	4.2% + -1.3%	1.8% + -0.5%	72.7% + -2.6%	13.4% + -2.4%	0.8% + -0.3%	7.4% + -0.9%

varied widely among individual patients. For example, the Native American component ranged from a high of 90% in some samples to a low of <1% in others (Fig. 2c).

We also investigated samples that marked multiple ethnicities on the requisition form and found a wide range of genetic ancestry estimates (Table 2). For example, on average, patients that marked both European and East Asian on the form were predicted to have 41.8% European and 44.3% East Asian ancestry. Among the individual samples however, the East Asian component ranged from a low of 8.8% to a high of 58%. Finally, we considered samples with at least 20% predicted ancestry from both the European and East Asian continental groups (88 subjects). Of these samples, 29 (33%) reported both European and Asian on the requisition form, while 15 (17%) reported only Asian, 12 (13.6%) reported only European, and 32 (36.4%) reported something else (Additional file 2: Table S7).

Carrier risk in admixed populations

In order to better understand how genetic ancestry estimates within self-reported ethnic groups impact clinical estimates of risk, we investigated the relationship between carrier status of cystic fibrosis, sickle cell anemia, and GJB2-related nonsyndromic hearing loss and predicted ancestry among self-reported African and Latin American patients. These diseases were chosen because they are common and strongly associated with specific ancestral groups. We found significant differences in carrier risk based on genetic ancestry (Additional file 2: Table S8). For example, the mean African ancestry of Latin American patients identified as carriers of sickle cell anemia is 30.3%, while the mean African ancestry of

non-carriers is 11.5% ($P < 0.0001$, Mann-Whitney's U test). Similarly, the mean European ancestry of Latin American patients identified as carriers of cystic fibrosis is 59.4% while the mean European ancestry of non-carriers is 51.9% ($P = 0.03$). Finally, the mean European ancestry of African carriers of GJB2-related nonsyndromic hearing loss is 23.0%, while the mean European ancestry of non-carriers is 12.2% ($P = 0.01$).

In order to see how carrier rates differ based on genetic ancestry, we computed the 80th percentile of European and African ancestry proportion among self-reported Latin Americans and Africans. We then computed carrier rates among individuals above and below this 80th percentile threshold and found significant differences in carrier rates of these diseases (Table 3). For example, the carrier rate of cystic fibrosis among Latin Americans with less than 68.7% European ancestry is 1.6%, while the carrier rate of those with more than 68.7% European ancestry is 3.7%. To highlight the significance of this difference, if we assume a detection rate of 72% (based on the ACMG panel of 23 common CFTR mutations), in the case of a negative screen, the residual risk of actually being a carrier is 1/220 in Latin Americans with less than 68% European ancestry and 1/94 in Latin Americans with a higher European ancestry proportion. A similar situation is observed for Africans with sickle cell anemia, where the carrier rate varies from 6.7 to 15.7%, depending on whether they have less or more than 93% African ancestry. Finally, the carrier rate of GJB2-related nonsyndromic hearing loss among Africans varies from 1.4 to 5.6%, depending on whether they have less or more than 20% European Ancestry.

Table 3 Carrier rates by genetic ancestry

Sickle Cell Anemia				
	carrier rate < threshold	carrier rate > threshold	p-value	African Ancestry Threshold (80th percentile)
African	6.760%	15.741%	0.00605	93.439%
Latin American	1.261%	4.587%	0.00372	18.677%
Cystic Fibrosis				
	carrier rate < threshold	carrier rate > threshold	p-value	European Ancestry Threshold (80th percentile)
African	1.631%	6.482%	0.01095	19.928%
Latin American	1.606%	3.670%	0.06128	68.715%
GJB2-Related Nonsyndromic Hearing Loss				
	carrier rate < threshold	carrier rate > threshold	p-value	European Ancestry Threshold (80th percentile)
African	1.399%	5.556%	0.01870	19.928%
Latin American	3.440%	3.670%	0.83750	68.715%

Differences in carrier rates of sickle cell anemia, cystic fibrosis, and GJB2-related nonsyndromic hearing loss by proportion of African and European ancestry among self-reported Africans and Latin Americans. For each group, an ancestry threshold was chosen by computing the 80th percentile of ancestry proportion. For example the 80th percentile of African ancestry among Latin Americans is 18.68%. Thus, the carrier rate of sickle cell anemia of Latin Americans below this threshold is 1.261% and the carrier rate above it is 4.587%. P-values were computed using Fisher's exact test

Discussion

We compared two sources of self-reported ethnicity to genetic ancestry as predicted by a validated statistical model. First we measured concordance between a written self-report on a requisition form and a verbal report during a genetic counseling session. While for most ethnicities there was concordance between them, for ethnicities like Mediterranean, Native American, and Southeast Asian, responses on the requisition form and during genetic counseling were different. We also observed differences between self-reported ethnicity on the requisition form and genetic ancestry in South Asians and Southeast Asians. Notably, these differences were mitigated when looking at self-reported ethnicity during genetic counseling consults. The discrepancies imply that there is confusion about the meaning of the different labels, impacting our ability to rely on self-reported ethnicity, which is in line with prior reports [29, 30].

In the realm of carrier screening, the main consequence of inaccurate classification is miscalculation of reproductive risk. Our results show that carrier rates and residual risks are dependent on genetic ancestry in admixed populations (Table 3). For example, for Latin Americans, the carrier rate of cystic fibrosis varies from 1.6% to 3.7% depending on the percentage of European ancestry, and the carrier rate of sickle cell anemia varies from 1.3% to 4.6% depending on the percentage of African ancestry. Additionally, we found that in admixed populations, the proportion of genetic ancestry estimated from each continental group as well as what ethnicities are self-reported, varies by individual. As such, it is erroneous to assume that genetic disease risks affect admixed populations in a uniform manner.

Carrier rates and residual risk estimates vary significantly depending on which source is used to account for

population differences. Self-identification does not always paint a complete picture, as there may be uncertainty about family origins, confusion about labels, and identification or lack thereof with a particular group due to personal or cultural reasons. Since mutation allele frequencies are linked to genetic ancestry, using self-identified ethnicity to select panel content or to adjust for risk will lead to errors, which may impact clinical decisions. Despite recommendations from professional societies to base some carrier screening decisions on ethnicity, this study suggests that in order to ensure that carriers of severe genetic disorders are not missed, expanded pan-ethnic carrier screening panels should be utilized. Pan-ethnic panels are able to detect mutations regardless of minor allele frequency (low or high), and include all mutations described for a given disease, present in different populations. Additionally, targeted next generation sequencing can be used to detect novel or poorly described mutations in genes of interest. While, pan-ethnic panels have some disadvantages [31, 32], given the unreliability of self-reported ethnicity and the goal of providing couples with information to optimize pregnancy outcomes, pan-ethnic panels represent the most comprehensive approach.

A second conclusion from this study is that, since carrier rates, detection rates and residual risks vary based on ancestry, in order to accurately counsel patients about their reproductive risks, genetic ancestry should be determined in clinical practice. Prior reports also describe sets of Ancestry Informative Markers (AIMs) that could be used to determine genetic ancestry, although further work is needed to ensure that allele frequency estimates are representative of all populations and that the ancestry estimates are meaningful for all individuals [33–38].

Limitations of the present study include that the data was studied retrospectively using the records of a clinical carrier screening company. Self-reported ethnicity from requisition forms was entered into the database during sample accessioning on a rolling basis. It is possible that some of the records were entered incorrectly or later modified by genetic counselors. To ensure accurate measurement, we manually reviewed several hundred scanned requisition forms: both randomly and in cases where discordance between self-reports and genetic ancestry was found. Reviewing records randomly showed that less than 1% of the self-reported labels were incorrect in the database. Another limitation is that our genetic ancestry model is based on allele frequency estimates from a limited sample size and assumes that grouping individuals by continent provides meaningful estimates of origin. A better way would be to account for subtle differences in genetic ancestry among local subpopulations, however, this would require expanded screening platforms and increased resources. Despite these limitations, we were able to correctly estimate the origin of validation samples from the 1000 Genomes Project and we found a strong relationship between carrier rates and ancestry proportion in admixed populations. This shows that the ancestry model provides meaningful information. Further work with larger cohorts is needed to refine the ancestry model and to measure the relationship between carrier rates and genetic ancestry for more diseases. Additionally, in diseases where environmental factors are known to play a part, further work is needed to understand how much impact ancestry has versus social and cultural influences of an individual's self-identified group.

Conclusions

Estimates of carrier rates and residual risks depend on whether self-reported ethnicity or genetic ancestry is used to account for population differences. Furthermore, self-reports are not reliable for clinical decision-making. In order to mitigate the risk of ordering the wrong testing panel, we recommend the use of expanded pan-ethnic carrier screening panels. Additionally, in order to accurately estimate carrier rates and residual risks, we recommend the use of a genetic ancestry model in clinical genetic testing.

Additional files

Additional file 1: Requisition form used in this study. (PDF 287 kb)

Additional file 2: Supplementary tables. (XLSX 39181 kb)

Additional file 3: Figure S1. Plot of the first and second principal components obtained by Principal Component Analysis on 44 geographic groups (described in Additional file 2: Table S1) and 1142 AIMs. Each geographic group is shown as a point and is colored

according to the continental group to which it belongs. The plot illustrates that the AIMs separate most continental groups well, but the Middle Eastern and Central Asian groups do not form distinct clusters. (PNG 49 kb)

Acknowledgements

We thank Ed O'Neill for editorial comments in the manuscript.

Funding

This study was fully funded by Phosphorus, Inc. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

Genotype data for the 9138 subjects in this study is available upon request or at dbGAP (accession number phs001482.v1.p1). Scripts used to select the 1142 AIMs and to run the Genetic Ancestry algorithm as well as Supplementary Data are available at <https://github.com/P15/phosphorus-public/releases/tag/v1.0.0>.

Authors' contributions

RS, AM and OP participated in study design and data analysis. RS, SY, SE, AM, SAR, SB, NK, MN, DH, SG, RV, SHC, AH, JG and OP participated in data collection, interpretation of results and drafting the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Institutional Review Board to consent for this research was obtained from ASPIRE (IRB-R-004). All data presented here is de-identified (HHS 45 CFR part 46.101(b)(4)). Consent for this study was obtained from each participant verbally during genetic counseling consult, which is part of the standard of care, as approved by clinical protocol (IRB-R-004).

Consent for publication

Not applicable. All data presented here is de-identified (HHS 45 CFR part 46.101(b)(4)).

Competing interests

Authors Shraga, Bristow, Niknazar and Puig are full time employees of Phosphorus.

Authors Rodriguez, Kumar, Yarnall, and Manoharan are full time employees of Cooper Surgical.

Authors Elango, Hoffman, Ghadir, Vassena, Chen, Hershlag and Grifo have no conflicts to report.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Phosphorus, Inc, 25 West 26th St, New York, NY 10010, USA.

²CooperGenomics, Livingston, NJ, USA. ³Sarah Lawrence College, Bronxville, New York, USA. ⁴IVF Florida Reproductive Associates, Margate, FL, USA.

⁵Department of Obstetrics and Gynecology, The David Geffen School of Medicine at UCLA, California, Los Angeles, USA. ⁶Clinica Eugin, Barcelona, Spain. ⁷Institute for Reproductive Medicine and Science, Saint Barnabas Medical Center, Livingston, NJ, USA. ⁸Division of Reproductive Endocrinology, Department of Obstetrics and Gynecology, Hofstra Northwell School of Medicine, Manhasset, NY, USA. ⁹Division of Reproductive Endocrinology & Infertility, Department of Obstetrics and Gynecology, NYU Langone, New York, NY, USA. ¹⁰Northwell Fertility, North Shore University Hospital, Manhasset, NY, USA.

Received: 13 February 2017 Accepted: 16 November 2017

Published online: 28 November 2017

References

1. ACOG. Committee in Genetics. Screen Tay-Sachs Dis. 2005;318:1–2.
2. ACOG. Committee in Genetics. Fam History Risk Assess Tool. 2015;478:1–4.
3. ACOG. Committee in Genetics. Preconception Prenat Carrier Screen Genet Dis Individ Easter Eur Jewisj Descent. 2009;442:1–4.

4. Grody WW, Thompson BH, Gregg AR, Bean LH, Monaghan KG, Schneider A, et al. ACMG position statement on prenatal/preconception expanded carrier screening. *Am Coll Med Genet.* 2013;15:482–3. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23619275>.
5. ACOG. Committee in Genetics. Update Carrier Screen Cystic Fibrosis. 2011; 486:1–4.
6. Cooper RS, Kaufman JS, Ward R. Race and genomics. *N Engl J Med.* 2003; 348:1166–70.
7. Dries DL. Genetic ancestry, population admixture, and the genetic epidemiology of complex disease. *Circ Cardiovasc Genet.* 2009;2:540–3.
8. Yaeger R, Avila-Bront A, Abdul K, Nolan PC, Grann VR, Birchette MG, et al. Comparing genetic ancestry and self-described race in African Americans born in the United States and in Africa. *Cancer Epidemiol Biomark Prev.* 2008;17:1329–38.
9. Mersha TB, Abebe T. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. *Hum Genomics.* 2015;9:1–15. Available from: <http://www.humgenomics.com/content/9/1/1>.
10. Vichinsky EP. Changes in the epidemiology of Thalassemia in North America: a new minority disease. *Pediatrics.* 2005;116:e818–25. Available from: <http://pediatrics.aappublications.org/cgi/doi/10.1542/peds.2005-0843>.
11. Hassell KL. Population estimates of sickle cell disease in the U.S. *Am J Prev Med.* 2010;38:S512–S521. Available from: [doi:10.1016/j.amepre.2009.12.022](https://doi.org/10.1016/j.amepre.2009.12.022).
12. Lazarin GA, Haque IS, Nazareth S, Iori K, Patterson AS, Jacobson JL, et al. An empirical estimate of carrier frequencies for 400+ causal Mendelian variants: results from an ethnically diverse clinical sample of 23,453 individuals. *Genet Med.* 2013;15:178–86. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3908551&tool=pmcentrez&rendertype=abstract>.
13. Mitchell JJ, Capua A, Clow C, Scriver CR. Twenty-year outcome analysis of genetic screening programs for Tay-Sachs and beta-thalassemia disease carriers in high schools. *Am J Hum Genet.* 1996;59:793–8.
14. Kumar R, Seibold MA, Melinda C, Aldrich L, Williams K, Reiner AP, Colangelo L, Galanter J, Gignoux C, Hu D, Saunak Sen SC, Nickerson DA, Fornage M, Enright P, Ziv E, Smith LJ, Liu K, EGB. Genetic ancestry in lung-function predictions. *N Engl J Med.* 2010;363:321–30.
15. Menezes AMB, Wehrmeister FC, Hartwig FP, Perez-Padilla R, Gigante DP, Barros FC, et al. African ancestry, lung function and the effect of genetics. *Eur Respir J.* 2015;45:1582–9. Available from: <http://dx.doi.org/10.1183/09031936.00112114>.
16. Wassel CL, Pankow JS, Peralta CA, Choudhry S, Seldin MF, Arnett DK. Genetic ancestry is associated with subclinical cardiovascular disease in African-Americans and Hispanics from the multi-ethnic study of atherosclerosis. *Circ Cardiovasc Genet.* 2009;2:629–36.
17. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet.* 1988;43:520–6. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1715499/>.
18. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol.* 2005;28:289–301.
19. Alexander DH, Novembre J, Lange K. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res.* 2009;19:1655–64.
20. Rajeevan H, Soundararajan U, Kidd JR, Pakstis AJ, Kidd KK. ALFRED: an allele frequency resource for research and teaching. *Nucleic Acids Res.* 2012;40:1–6.
21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559–75.
22. Wright S. The Genetical structure of populations. *Ann. Eugen.* 1951;15:323–54.
23. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. *Nature.* 2015;526:68–74. Available from: <https://www.nature.com/articles/nature15393>.
24. Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK. ALFRED: an allele frequency database for diverse populations and DNA polymorphisms—an update. *Nucleic Acids Res.* 2001;29:317–9.
25. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovskiy LA, et al. Genetic structure of human populations. *Science.* 2002;298:2381–5. <http://science.sciencemag.org/content/298/5602/2381.full>.
26. Tian C, Kosoy R. European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. *Mol Med.* 2009; 371–83. Available from: http://www.molmed.org/content/pdfstore/09_94_Tian.pdf.
27. 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.* 2015;96:37–53. Available from: [doi:10.1016/j.ajhg.2014.11.010](https://doi.org/10.1016/j.ajhg.2014.11.010).
28. Smith E, Jepsen K, Arias A, Shepard P, Chambers C, Frazer K. Genetic ancestry of participants in the National Children's study. *Genome Biol.* 2014; 15:R22. Available from: <http://genomebiology.com/2014/15/2/R22>.
29. Burnett MS, Strain KJ, Lesnick TG, De Andrade M, Rocca WA, Maraganore DM. Reliability of self-reported ancestry among siblings: implications for genetic association studies. *Am J Epidemiol.* 2006;163:486–92.
30. Hunt L, Megyesi M. The ambiguous meanings of the racial/ethnic categories routinely used in human genetics research. *Soc Sci Med.* 2008;66:349–61.
31. Foulkes WD, Knoppers BM, Turnbull C. Population genetic testing for cancer susceptibility: founder mutations to genomes. *Nat Rev Clin Oncol.* 2016;13: 41–54. Available from: [doi:10.1038/nrclinonc.2015.173](https://doi.org/10.1038/nrclinonc.2015.173).
32. Edwards JG, Feldman G, Goldberg J, Gregg AR, Norton ME, Rose NC, et al. Expanded carrier screening in reproductive medicine—points to consider. *Obstet Gynecol.* 2015;125:653–62.
33. Enoch M-A, Shen P-H, Xu K, Hodgkinson C, Goldman D. Using ancestry-informative markers to define populations and detect population stratification. *J Psychopharmacol.* 2006;20:19–26.
34. Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Hum Mutat.* 2008; 29:648–58.
35. Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, Madbouly A, et al. Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int Genet.* 2014;10:23–32.
36. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat.* 2009;30:69–78.
37. Phillips C, Parson W, Lundsberg B, Santos C, Freire-Aradas A, Torres M, et al. Building a forensic ancestry panel from the ground up: the EUROFORGEN global AIM-SNP set. *Forensic Sci Int Genet.* 2014;11:13–25. Available from: [doi:10.1016/j.fsigen.2014.02.012](https://doi.org/10.1016/j.fsigen.2014.02.012).
38. Nievergelt CM, Maihofer AX, Shekhtman T, Libiger O, Wang X, Kidd KK, et al. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig Genet.* 2013;4:13. Available from: <http://investigativegenetics.biomedcentral.com/articles/10.1186/2041-2223-4-13>.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

