Published in final edited form as:

Pharmacogenomics. 2009 April ; 10(4): 565–577. doi:10.2217/pgs.09.5.

Physiogenomic analysis of the Puerto Rican population

Gualberto Ruaño^{1,†}, Jorge Duconge², Andreas Windemuth¹, Carmen L Cadilla², Mohan Kocherla¹, David Villagra¹, Jessica Renta², Theodore Holford³, and Pedro J Santiago-Borrero²

¹ Genomas Inc., 67 Jefferson Street, Hartford, CT 06106, USA, Tel.: +1 860 545 3773, Fax: +1 860 545 4575, g.ruano@genomas.net

² University of Puerto Rico, San Juan, Puerto Rico

³ Yale University School of Medicine, CT, USA

Abstract

Aims—Admixture in the population of the island of Puerto Rico is of general interest with regards to pharmacogenetics to develop comprehensive strategies for personalized healthcare in Latin Americans. This research was aimed at determining the frequencies of SNPs in key physiological, pharmacological and biochemical genes to infer population structure and ancestry in the Puerto Rican population.

Materials & methods—A noninterventional, cross-sectional, retrospective study design was implemented following a controlled, stratified-by-region, random sampling protocol. The sample was based on birthrates in each region of the island of Puerto Rico, according to the 2004 National Birth Registry. Genomic DNA samples from 100 newborns were obtained from the Puerto Rico Newborn Screening Program in dried-blood spot cards. Genotyping using a physiogenomic array was performed for 332 SNPs from 196 cardiometabolic and neuroendocrine genes. Population structure was examined using a Bayesian clustering approach as well as by allelic dissimilarity as a measure of allele sharing.

Results—The Puerto Rican sample was found to be broadly heterogeneous. We observed three main clusters in the population, which we hypothesize to reflect the historical admixture in the Puerto Rican population from Amerindian, African and European ancestors. We present evidence for this interpretation by comparing allele frequencies for the three clusters with those for the same SNPs available from the International HapMap project for Asian, African and European populations.

†Author for correspondence.

For reprint orders, please contact: reprints@futuremedicine.com

Online supplement

The current study's database of genes, SNPs and allele frequencies is provided as an online supplement [103].

Financial & competing interests disclosure

Gualberto Ruaño, Andreas Windemuth, Mohan Kocherla and David Villagra are full time employees of Genomas Inc. This investigation was supported, in part, by a Research Centers in Minority Institutions Award, G12RR-03051, from the National Center for Research Resources, NIH; a Clinical Research Center Infrastructure Initiative Pilot Projects Award (RCRII) Grant No. 5P20RR011126; and by the Puerto Rico Newborn Screening Program and Genomas internal research and development funds. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Conclusion—Our results demonstrate that population analysis can be performed with a physiogenomic array of cardiometabolic and neuroendocrine genes to facilitate the translation of genome diversity into personalized medicine.

Keywords

ancestry; admixture; personalized medicine; physiogenomics; population genetics; Puerto Rico; SNP

Pharmacogenetics and physiogenomics are rapidly growing fields of research shaping the panorama of genetic research and its potential clinical implications. Pharmacogenetics is based on the concept that inherited, innate pharmacokinetic and dynamic attributes of the individual account for many of the differences in how people respond to drugs [1]. Physiogenomics is a medical application of sensitivity analysis and systems engineering in which the relationship of gene variation to physiology sheds light on the role of specific genes in environmental response. Sensitivity analysis is the study of the dependence of a system on changes in its components [2]. Both disciplines have gained tremendous momentum from recent research on human genome diversity. These advances have been possible by novel DNA typing technologies for gene polymorphisms in pharmacokinetic (drug metabolism and transport), pharmacodynamic (target receptors and enzymes) and physiological networks affecting human pharmacological response.

Heterogeneity in how people respond to medications has confounded the prescription of modern medicines, with detrimental consequences for the safety, efficacy and patient compliance of potent drugs. A major advance in healthcare would be a transition from the current empirical approaches and trial employed in drug therapy to a genetically predictive framework for determining the individual patient's response to medicines. Solidly founded on human evolutionary history, physiogenomics transcends population averages to develop predictive genetic markers of physiological response for the individual. The use of physiogenomic markers, instead of anthropological ancestry markers, may provide a direct venue to personalized medicine without the interference of confounding surrogates such as race and ethnicity.

Both quantitative (allele frequencies) and qualitative (allelic DNA sequences) genomic differences have been observed among different population groups [3–6]. Ethnogeographic origin has become an integral part of research in pharmacogenetics. The database Pharmacogenomics in Admixed Populations was conceived to compile data from peoples of Africa, America, Asia and Oceania, where admixture and population stratification occurs in distinct patterns [3]. Admixture is a common type of gene flow in human population, and occurs when individuals from two or more parental populations form a new hybrid population. Population stratification, the presence in one population of subgroups that differ in allele frequencies, is of great relevance to public health and to clinical application of pharmacogenetics.

Mainland Puerto Ricans currently represent 1.2% of the US population and 9.6% of the Hispanic population in the USA [101]. Similar to others in Latin America, the Puerto Rican population originated as a result of admixture between Amerindians, whose ancestors had migrated from the Amazon Basin and arrived in Puerto Rico 2200 years before present, and Spaniard and West-African individuals. The island of Puerto Rico thus is endowed with a distinctive population in terms of its gene flow. There are growing numbers of Puerto Ricans in the USA, as Puerto Rico has been a US territory or commonwealth since 1898. Admixture studies in Puerto Ricans, either in the island or the continental USA, have been scarce.

Endeavors, such as the International HapMap Project, have provided insights into the variability of polymorphism frequency between worldwide populations. Yet frequencies of most described functional polymorphisms have not yet been assessed in multiple populations comprehensively. Most studies addressing genetic population differentiation and its applications in ancestral and demographic extrapolations have found that there are significant differences in genotype, haplotype and copy-number variation between worldwide populations [4–7]. Future studies that document worldwide allele frequencies, such as the Allele Frequency Database (ALFRED), Pharmacogenetics for Every Nation Initiative (PGENI) and 1000 Genomes Project are essential to establish global profiles of functional polymorphisms [8,9]. Furthermore, several studies have shown recent natural selection continues to alter the context of genetic population differentiation, further underscoring the need for continued worldwide genomic analyses, particularly in admixed populations such as that of Puerto Rico [10–12].

The purpose of this study was to determine distribution frequencies of 332 SNPs from 196 cardiometabolic and neuroendocrine genes in the Puerto Rican population. Another goal of this work was to apply these physiogenomic markers for inferring the Puerto Rican population structure, following a Bayesian clustering algorithm.

Methods

Design overview & DNA samples

This research represents a noninterventional, retrospective, cross-sectional study of 100 genomic DNA samples provided by the Puerto Rico Newborn Screening Program (PRNSP). The PRNSP has been in operation for more than 8 years under a Puerto Rico law (number 84, as of July 2, 1987), which mandates that all newborns be screened for various inborn errors of metabolism and hemoglobinopathies. The PRNSP sample repository and research resource is maintained at the Pediatrics Hospital of the University of Puerto Rico Medical Sciences Campus. Provided as dried blood spotted in filter cards, the samples were representative of all Puerto Rico hospitals with birthing facilities. The samples were selected following a controlled, stratified-by-region, random sampling protocol adjusted for the regional birthrates of the island of Puerto Rico based on the 2004 National Birth Registry. Figure 1 depicts location and number of samples selected among different municipalities of Puerto Rico.

For each assay, approximately 90 ng/μl (volume 50 μl) of genomic DNA from leukocytes in dried-blood spot was used after being extracted and purified by a Generation DNA Purification kit (QIAGEN Inc., CA, USA). Extracted DNA samples were separated into fractions and duplicates stored at –80°C in TRIS-EDTA (tris(hydroxymethyl)aminomethane ethylenediaminetetraacetic acid [TE]) buffer for retrieval as needed.

Physiogenomic array

The physiogenomic (PG) array is composed of 384 SNPs from 222 genes. It has been developed by Genomas (CT, USA) as a product and service, and tested on nearly 5000 patients from various clinical studies. This PG-array has been successfully applied in cardiovascular and neuropsychiatric research involving responses to cardiovascular, psychotropic and diabetic drugs as well as to dietary, exercise and acoustical interventions. The PG-array has been the cornerstone of physiological genomics research at our laboratory resulting in ten publications to date [201,13–21].

A representative selection of proteins whose genes are included in the PG-array are shown in **Box 1**. The following key physiological and pharmacological pathways were represented: insulin resistance, glucose metabolism, energy homeostasis, adiposity, apolipoproteins and receptors, fatty acid and cholesterol metabolism, lipases, receptors, cell signaling and

transcriptional regulation, growth factors, drug metabolism, blood pressure, vascular signaling, endothelial dysfunction, coagulation and fibrinolysis, vascular inflammation, cytokines, neurotransmitter axes (serotonin, dopamine: cholinergic, histamine and glutamate) and behavior (satiety). We utilized public databases (dbSNP and Ensembl) for validated SNPs with known allele frequencies for mixed or Caucasian populations. We selected SNPs with minor allele frequency consistently between 10 and 30%, while avoiding those with higher allele frequencies, presuming them to be more likely phenotypically neutral. The full listing of genes and SNPs in the PG-array has been published as part of a patent application on the physiogenomics technology [201]. The current analysis was restricted to autosomal genes.

Genotyping technology & assay—Genotyping was performed using the BeadArray™ platform (Illumina, CA, USA) and the GoldenGate™ assay (Illumina) [22,23]. Owing to the high sensitivity of heterozygosity and population frequency comparisons on genotyping quality, careful manual analysis was performed on the alignments underlying the genotype calls using GenCall 6.1.3.24, and 50 SNPs with even a slight degree of uncertainty about calling accuracy were not included in the analysis, leaving 332 SNPs from 196 genes. Similarly, two subjects were eliminated from the genetic distance analysis due to inconsistent genotyping results, leaving 98. Finally, the Bayesian clustering analysis requires complete combinatorial genotyping results for each individual. Therefore, 27 samples were excluded due to partial genotyping results, leaving 71 individuals for the STRUCTURE analysis.

Genetic analysis

allele frequencies—Allele frequencies (f) have been determined in the Puerto Rican population for all 332 loci.

Heterozygosity—Wright's F -statistic (F_{ST}) was calculated for each locus from the observed total heterozygosity $H_T = N_{hz}/N$ and the subpopulation heterozygosity $H_S = 2f \cdot (1-f)$, assumed under Hardy–Weinberg (H–W) equilibrium, as $F_{ST} = (H_T - H_S)/H_T$. A t -test was performed to see whether the average F_{ST} across all loci was different from zero. Since deviations from H–W equilibrium are used in genotyping quality control, a detailed analysis of H–W disequilibrium in individual SNPs was not performed.

STRUCTURE—Individual admixture estimates (proportions) were derived using a Bayesian approach, as implemented in the STRUCTURE version 2.2 software package [102,24,25]. This program is a public domain software package for using multilocus genotype data to investigate population structure, identifying admixed individuals and estimating population allele frequencies in situations where individuals are admixed. STRUCTURE results were obtained using default parameters and an assumed cluster number of three, determined as follows.

We performed a series of three Markov chain Monte Carlo runs for each with the cluster size (K) chosen as $K = 1, 2, 3, 4$ and 5 . A burn-in time of 20,000 iterations was followed by 80,000 further iterations. We noted the log-probability of the data and the log likelihood for each run, in order to estimate the posterior probabilities for the different numbers of clusters. We found that the log-probability of the data was maximal at $K = 2$, while the log-likelihood shows a clear maximum between $K = 3$ and $K = 4$. The log likelihood was remarkably consistent between three different runs ($<1\%$ at $K = 1$; $<10\%$ at $K = 5$, as percentage of variation between different K). The log probability had greater variation ($<10\%$ at $K = 1$; $<100\%$ at $K = 5$), but still varied less between runs than between different values of K . The log-probabilities for one of the runs were -17174 , -17094.3 , -17353.1 , -17732.2 and -17203.9 , respectively, for $K = 1, 2, 3, 4$ and 5 . Similarly, the log likelihoods were -17116 , -16924.5 , -16860.4 , -16808.1 and -16962.6 . Overall, these results point towards a good fit for $K = 3$, but do not exclude $K = 2$.

Box 1. Selected representative proteins whose gene polymorphisms compose the physiogenomics array

Insulin resistance

- Insulin, insulin receptor, insulin-receptor substrate-1, Akt1, Akt2, cholecystokinin and cholecystokinin receptor A and B, resistin, regulatory subunit of PI3-kinase (polypeptide 1, p85 α) and ATP-binding cassette B1, C8 and G5.

Glucose metabolism

- Glucagon, glycogen synthase 1, 2 and 3 β , phosphofructokinase (liver, muscle and platelet), pyruvate kinase (liver, red blood cells and muscle) and phosphoenolpyruvate carboxykinase 1.

Energy homeostasis

- Uncoupling protein 2 and 3, adrenergic receptor α 1A, α 2A, α 2B, β 1, β 2 and β 3, carnitine palmitoyltransferase 1A, 1B and 2, melanocortin receptor 3 and 4, pro-opiomelanocortin, malate dehydrogenase, AMPK, subunit α 1, catalytic AMP-activated protein kinase α -2, noncatalytic AMP-activated protein kinase β 1, β 2, γ 1, 2 and 3, and thioredoxin reductase 2.

Adiposity

- Leptin receptor, ghrelin precursor, adiponectin receptor 1 and 2, and adipocyte.

Lipases

- Hepatic, lipoprotein, hormone-sensitive, lysosomal acid, endothelial and gastric.

Growth factors

- Insulin-like growth factor 1, growth hormone 1, growth hormone releasing hormone and TGF.

Cell signaling

- Catalytic PI3K α , β , γ , δ polypep, class 2 PI3K β , γ polypep, class 3 PI3K, catalytic PI3K and α -polypep.

Vascular signaling

- VEGF-A, VEGF receptor (KDR), angiopoietin 1, 2, TEK tyrosine kinase and adenosine receptors A1, A2a, A2b and A3.

Transcriptional regulation

- Peroxisome proliferation-activated receptor α and γ , hypoxia-inducible factor 1A, sterol regulatory element-binding transcription factor 1, retinoic receptor α , β , γ ; α , β and γ , and WBSCR14.

Adrenal function

- Glucocorticoid receptor, corticotrophin-releasing hormone and receptor 1 and 2.

Apoptosis

- TNF apoptosis, caspase activation inhibitor (AVEN).

Cytochrome P450

- 1A2, 2C19, 2D6, 2B6, 2C9, 3A4, 3A5 and 7.

Fatty acid homeostasis

- Acetyl-CoA-acetyltransferase 1 and 2, acetyl-CoA carboxylase α and β , CETP, lecithin:cholesterol acyltransferase, fatty acid synthase, fatty acid-binding protein 2, microsomal triglyceride transfer protein, paraoxonase 1, choline kinase β , HMG-CoA reductase.

The physiogenomic array consists of 222 genes and 384 SNPs. The published patent application contains a full listing of its genes and SNPs.

Assuming that $K = 3$, each sample represented by a dot in Figure 2 is quantified with a proximity value (between 0 and 1) corresponding to each of the three vertices of the triangle. Closer proximity to any vertex resulted in proportionally higher values. We related each subject to one of the three clusters by determining which of the three proximity values was greatest. We added a categorical criterion requiring that in order for the subject to be assigned to a specific cluster, its highest proximity value must be greater than the two others by at least 0.20. This criterion excludes those samples not clearly positioned towards any vertex and located near the middle of the triangle in Figure 2. The value 0.20 was chosen to optimize discrimination and representation among clusters. Figure 2 shows these demarcated regions by the addition of dashed lines isolating each vertex of the triangle.

Genetic distance

Allelic dissimilarity was defined by the Euclidean metric:

$$d_{ij} = \sqrt{\sum_k^{N_{loc}} (g_{ik} - g_{jk})^2}$$

Involving the genotypes g_{ik} for subject i at locus K , using values of 0 for common homozygotes, 1 for heterozygotes and 2 for variant homozygotes. Allelic dissimilarity was calculated for all pairs of subjects. Agglomerative hierarchical clustering was performed on the dissimilarity matrix, using the complete linkage method [26,27]. Of the other linkage methods commonly used, single linkage was excluded because of its tendency to produce long chains, and the centroid linkage often used in biology was rejected because of its tendency towards reversals [28]. The hierarchical cluster (hclust) procedure from the R Statistical Environment with default parameters was used for the computation [29]. It utilizes a very efficient nearest neighbor chain algorithm [30]. The hclust procedure does not require an assumption of the number of clusters and none was made. The results of this analysis are presented in Figure 3A. With various subjects assigned to one of the three STRUCTURE clusters, we then designated red, green and blue colors to cluster 1, 2 and 3, respectively, and applied the color code to each corresponding sample on the genetic distance dendrogram, comprising Figure 3B.

Ancestry—A database of genotypes available from the International HapMap Project [31] was utilized for ancestry inferences by calculating the allelic dissimilarity for each cluster in the Puerto Rico sample with reference populations. The African reference population consisted of a set of 30 trios (parents and one child) of the Yoruba people from Idaban, Nigeria (YRI). The Asian reference population consisted of 45 unrelated individuals of Han Chinese from Beijing (HCB), China. The European reference population consisted of a set of 30 trios of US residents with ancestry from Northern and Western Europe (CEU). HapMap sample allele frequencies in the African and European reference populations were calculated counting only the 120 independent alleles of the trio parents.

The inferred allele frequencies for the three clusters derived from STRUCTURE analysis were exported and compared with the allele frequencies of the same SNPs in the HapMap project database [31]. The root mean square (RMS) of the allele frequencies over all SNPs was used to measure population dissimilarity r_{ij} :

$$r_{ij} = \sqrt{\sum_k^{N_{loc}} (f_{ik} - f_{jk})^2}$$

Involving the allele frequencies f_{ik} for cluster i at locus k .

Results

In this research, we performed 32,536 genotype assays from 332 SNPs in 196 cardiometabolic and neuroendocrine genes in 98 individuals from the island of Puerto Rico. Genotype frequencies are reported from at least 90 individuals per SNP. Hierarchical clustering analysis utilized 98 individuals. STRUCTURE analysis utilized 71 individuals with complete combinatorial genotypes. The database of genes, SNPs, alleles and genotype frequencies for the Puerto Rican population sample is shown in the online supplement [103]. These genotypes in aggregate served to ascertain various characteristics of the Puerto Rico population sample, including heterozygosity, clustering, stratification and ancestry.

Heterozygosity

The average F_{ST} for the Puerto Rican population is -0.0033 and not significantly different from zero. A positive F_{ST} would be expected if there was a large population of first-generation immigrants, or marked segregation of non-intermixing groups of different heritage. The observation of no excess heterozygosity is not in contradiction with the markedly admixed character of the population, since H-W equilibrium is established after a single generation under random mating conditions.

Clustering

For inference of the major ancestry contributions within the Puerto Rican population, a Bayesian clustering algorithm was used. The analysis of the PG-array derived informative markers data with the STRUCTURE version 2.2 software shows that the samples readily separate into three clusters. In Figure 2, each corner of the triangle corresponds to one of the three clusters, and individuals are plotted as points corresponding to the amount of admixture inferred by the algorithm from each of the assumed three ancestral populations. Given that complete combinatorial genotype results are necessary for accurate STRUCTURE analysis, 71 individual samples meeting that criteria were included in the STRUCTURE analysis seen in Figure 2. Cluster membership, as inferred by STRUCTURE [25], was nearly evenly divided in the population (34, 28 and 38%, respectively). When the categorical criterion of proximity to the vertex was applied to the data, 53 samples were assigned to a particular cluster, and the cluster membership became 40, 19 and 41% in cluster 1, 2 and 3, respectively.

Genetic distance

Figure 3A shows a hierarchical dendrogram to illustrate the population structure within the Puerto Rican population as represented by allelic dissimilarity. This analysis was independent of the one with STRUCTURE. One pair of samples (26 and 92) shows an unusually low allelic dissimilarity, suggesting relatedness. Owing to the anonymous study design, this could not be further investigated. There appear to be three main sectors, from left to right: PR126–PR341,

PR321–PR69 and PR97–PR175. There are also two smaller sectors between the three main sectors.

Figure 3B highlights the same genetic distance dendrogram as Figure 3A, but with the sample identification numbers replaced by color-coded rectangles highlighting which of the three STRUCTURE clusters from Figure 2 the particular sample belongs to. For descriptive purposes, consider Figure 3B to be a map with three sectors, 98 individual lanes and 53 sites (colored rectangles). Some lanes do not have sites because the genotyping data was not complete or because the individual residing in the site did not meet the categorical criterion for assignment to any of the three clusters. Figure 3B demonstrates a clear correlation between the two independent grouping methods. Sector 1 (leftmost in the dendrogram) bears a concentration of samples assigned to cluster 2, the bottom right vertex of the triangle in Figure 2 ($p = 0.083$). Sector 2, in the middle of the genetic distance dendrogram, appears disproportionately blue in color, corresponding to samples from cluster 3 in Figure 2 ($p = 0.017$). Finally, sector 3, the rightmost sector of Figure 3B, is clearly enriched with cluster 1 samples ($p = 2.79 \times 10^{-6}$). Overall, sector 1 is comprised of ten samples, sector 2 of 13 samples, and sector 3 of 16 samples. There are two smaller sectors between these main sectors, which contain a total of 14 samples.

Ancestry

Figure 4 shows the population dissimilarities between the subpopulations inferred by STRUCTURE and the reference populations from the HapMap project database [31]. We calculated the population dissimilarity for each cluster with each of the reference populations. The results show that cluster 1 can be identified with the African population, a set of 30 trios (parents and one child) of the Yoruba people from Idaban, Nigeria. Cluster 2 most resembles the Asian population, consisting of 45 unrelated individuals of Han Chinese from Beijing. Cluster 3 is most closely related with the Caucasian population, a set of 30 trios of US residents with ancestry from CEU. It is likely that these clusters correspond to the three main historical populations of Puerto Rico: West Africans (cluster 1), brought in via the slave trade; Amerindians (cluster 2), the original inhabitants historically related to East Asians; and Western Europeans (cluster 3), who came as colonists predominantly from Spain.

Figure 5 shows the relative ancestry contributions for each of the sampled individuals in Figure 2. Individuals are displayed according to their predominant cluster contribution. Each individual has contributions from each of the clusters, but in widely different proportions. The maximal contribution from a single cluster observed in this sample is approximately 85% for red and blue (clusters 1 and 3) and approximately 70% for green (cluster 2). The contribution of cluster 2 (green color) to the relative ancestry proportion is less pronounced than that of the two other clusters. From Figure 4, clusters 1, 2 and 3 can be identified with the African, Amerindian and European lines of ancestry, respectively.

Discussion

Genome variation and ancestry in the Puerto Rican population were analyzed using 332 SNPs from 196 cardiometabolic and neuroendocrine genes. According to the results, a trichotomous structure with European, Amerindian and West-African contributions can be ascertained (see Figure 2 and Figure 3B). As a comparison, in Brazil, genetic evidence for a highly admixed population with Amerindian, European and African ancestral roots, has been ascertained using two pharmacogenetic relevant genes, *NAT2* (*N*-acetyltransferase 2) and *ABCB1* (ATP-binding cassette, sub-family B, member 1) [32,33]. Our data reflect a trichotomous origin of the Puerto Rican population settled in the island, which is consistent with various reports for the ancestry of Hispanics in the USA [3,34–38].

Ancestral contributions to the Puerto Rican population have been estimated by using polymorphic blood group and protein marker data, as 45% European, 37% West African, and 18% Amerindian [39]. Notably, analysis of mitochondrial DNA in Puerto Ricans living in the island revealed a higher Amerindian contribution of nearly 53% [38]. By contrast, both a set of 61 Y chromosome SNPs (Y-SNPs) and 11 core Y-short tandem repeats (STRs) pointed to a much larger European paternal contribution in Puerto Ricans [40]. Despite regional variation, there have been reported low levels of Hispanic Y-STR haplotype heterogeneity in previous surveys, although the higher frequency of African-derived Y chromosomes in the East is consistent with a greater contribution of Puerto Rican and Cuban Hispanics to East Coast US populations [41–43]. The contributions of the three parental populations to contemporary Puerto Ricans have been ascertained using a set of 35 autosomal ancestry informative markers (AIMs) [44]. The analysis provided evidence of a Puerto Rican gene-pool having European, West African and Amerindian origins, with larger European and West African components but an unquestionable Amerindian contribution.

The physiogenomic analysis showed herein supports these earlier reports and suggests that heterogeneity is a major characteristic of the Puerto Rican population (see Figure 2). The PG-array consists of SNPs representative of cardiometabolic and neuroendocrine genes, including various energy and inflammatory relevant pathways, instead of the typical AIMs. We believe the PG-array heralds the extension of genetic ancestry analysis to functional genes. Genotyping technology is rapidly evolving towards total genome arrays and ultimately individual genome sequencing. The utility of AIMs may wane, as their main advantage is for obtaining ancestry information from a limited set of DNA polymorphisms without physiological significance. With markers essentially unlimited, such considerations may be superseded. As a case in point, a recent study on the structure of European populations using total genome arrays has presented an image of unprecedented detail, linking genetic and geographic ancestry in Europe to within a few hundred kilometers [45]. Following this example, we will direct future work on the Puerto Rican population structure towards total genome arrays.

A moderate correlation between skin color (melanin index) and ancestry has been reported using 36 AIMs in a Puerto Rican sample settled in New York city [46]. This relationship was explained as a result of population structure due to admixture stratification in the Puerto Rican sample. The individuals were of primarily European ancestry ($53.3 \pm 2.8\%$) but also had relatively large proportions of West African ($29.1 \pm 2.3\%$) and Amerindian ($17.6 \pm 2.4\%$) ancestry.

In our STRUCTURE results using genotype data from 332 SNPs in 196 cardiometabolic and neuroendocrine genes, we have found three approximately evenly divided clusters in the Puerto Rican population sample. This distribution does not agree with the sizes of the allelic dissimilarity clusters or historical admixture proportions. This is likely because the STRUCTURE algorithm has a bias towards equal sized clusters if the cluster separation is weak. However, note that cluster 2, corresponding to the Amerindian component, is the smallest in Figure 2 and Figure 3B, as would be expected. It will be necessary to obtain a larger sample to encompass the Amerindian component in the Puerto Rican population more distinctively. The sample size of 100 individuals falls within the same order of magnitude as multiple previously published admixture studies of the Puerto Rican population [10,47,48]. Additionally, Figure 1 illustrates the geographic diversity of the sample cohort, contributing to the random and widespread nature of the genotyped population. The selection of 71 individuals included in the STRUCTURE analysis also is valid, given that the larger ($n = 98$) hclust analysis parallels the findings from the STRUCTURE clustering. With a larger sample size it should be possible to detect regional differences in population admixture. Such have been already detected with mitochondrial markers [35]. Further, our analysis utilized autosomal genes and excluded mitochondrial DNA and genes located on the sex chromosomes.

The overlaying of the clustering from Figure 2 onto the genetic distance dendrogram of Figure 3A illustrates that there is indeed a correlation between the two independent methods. The results, represented in Figure 3B, validate the accuracy of the groupings and clusters presented in this study. Other notable findings from Figure 3B include the relatively low proportion of samples assigned to cluster 2. We identified the ancestry of cluster 2 as Amerindian. The Amerindian population had the most distant geographic provenance and therefore would have the least markedly identifiable ancestral genetic markers.

Figure 4 depicts population dissimilarity for each identified cluster with each of the reference populations from the HapMap project database. cluster 2 most resembles the Asian population. It is likely that the inferred cluster 2 corresponds to the ancestral Amerindian population of Puerto Rico. Amerindians have been historically related to East Asians by the 'Bering bridge' theory, according to which the Amerindian populations have a basic Asian Mongoloid configuration, which was secondarily modified by local micro-evolutionary processes in the Americas upon adaptation over several millennia [49,50]. There is general agreement that the first human inhabitants of the Americas originated from small groups of prehistoric immigrants from Northeastern Asia. The Bering Strait is the most likely place for migrating groups to have crossed into the New World. Recent research supports the land-bridge theory [51]. In a study analyzing DNA variation at 678 markers in 29 modern Amerindian populations from North, Central and South America and two Siberian groups it was found that genetic proximity to the Siberian groups increases in populations closer to the Bering Strait [51].

With the advent of AIMs, the possibility of examining the association between disease or trait frequency and individual ancestry estimates has been postulated [37,47,48]. The interaction between individual ancestry and asthma within Puerto Ricans, a population showing the highest prevalence for this condition among Hispanics in the USA, has been demonstrated to be complex using 44 AIMs [52]. Puerto Ricans of lower socioeconomic status (SES) with asthma had less African ancestry and greater European ancestry than healthy control subjects. However, the reverse of the ancestry differential was observed for higher SES Puerto Ricans.

The history of the Americas has been marked by the encounter of populations from different continents. The discovery of the New World began a period defined by human migrations at a much larger scale than in previous history. This movement of people, voluntary or forced, changed profoundly the human landscape. As populations came into contact, admixture followed in varying degrees depending on the circumstances. Today, many people living in the Caribbean islands (Puerto Rico, Dominican Republic, Cuba) can trace their ancestry to more than one continent. The most important genetic contributions came from the Amerindians, Western Europeans and West Africans, although there have also been influences from other regions, such as East Asia and South Asia.

The history of migration and admixture can be reconstructed and interpreted using genetic markers. A very complete perspective can be obtained when analyzing autosomal markers, maternally transmitted mitochondrial DNA markers and paternally transmitted Y-chromosome markers. Genome-wide panels with markers showing high frequency differences between West African and European populations are already available for disease-gene discovery in African-Americans. Such a resource is being developed for Hispanic populations [53]. In this research article, we included autosomal markers in the proposed analysis focusing on physiological and drug-metabolism genes in Puerto Ricans.

Further research is planned in additional cohorts to discover disease and drug-response genetic associations using total-genome arrays in Puerto Ricans. As the entire genome can be screened, a multigene model can be developed where an individual's configuration of various significant SNPs can reliably predict the probability of drug-related adverse events, tolerability and

effectiveness for each patient. As multigene models require integrated genotyping of independently segregating genes, the range of possible allelic combinations in the Puerto Rican populations is considerable, and certain to exceed that in populations without admixture. Generalized clinical use of such diagnostics for DNA-guided medical management could help improve safety and effective use of prescription medicines. The determination of individualized treatment most suitable to each patient, using the personal genome for clinical decision support, and the implementation of drug prescription safeguards, are integral to evidence-based medicine and personalized health [54].

DNA-guided medicine poses great advantages for highly heterogeneous populations. With history and ancestry spanning Amerindians, Africans, and Europeans, the typical Puerto Rican person defies conventional ethnogeographic definitions. Personalized medicine based on pharmacogenetic and physiogenomic research could improve existing drug treatments and dosing regimens in such a highly heterogeneous population. The characterization of gene diversity in Puerto Ricans provides a unique opportunity to apply personalized medicine to individuals of Latin American descent in clinical areas with existing healthcare disparities, including cancer, diabetes, cardiovascular and respiratory disease and mental illness.

Executive summary

- The present research is the first demonstration of the utility of an array of polymorphisms in physiological genes to derive population ancestry. Previously published studies in this area have utilized classical autosomal ancestry informative markers, mitochondrial DNA, Y-SNPs/Y-short tandem repeats and protein marker data.
- Using allelic dissimilarity and Bayesian clustering analysis, the population structure of the Puerto Rican island-wide sample was found to be highly heterogeneous.
- The results from this study provide further evidence of the unique trihybrid model of genetic admixture in the Puerto Rican population. Evidence for the trihybrid model was also obtained by comparing allele frequencies for the three clusters with those for the same SNPs in reference populations from the HapMap project.
- The results demonstrated that population analysis can be performed with a physiogenomic array of 384 SNP from 222 key cardiometabolic and neuroendocrine genes providing a direct venue to personalized medicine.
- Admixture is of great relevance to apply pharmacogenetics and personalized medicine in the Puerto Rican population. As multigene models are developed to predict drug response, the range of possible allelic combinations in the Puerto Rican populations is certain to exceed that in populations without admixture.
- The characterization of gene diversity in Puerto Ricans provides a unique opportunity to apply personalized medicine to individuals of Latin American descent in clinical areas with existing healthcare disparities.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Yolanda Rodríguez for her support in collecting the samples for this survey and Dr Hernando Mattei (UPR School of Public Health) for his drafting of the Puerto Rico map in Figure 1.

Bibliography

Papers of special note have been highlighted as:

of interest

1. Kalow, W. Pharmacogenetics of Drug Metabolism. Elsevier Science Publishing Co; USA: 1992. Very useful and comprehensive overview of pharmacogenetics, it brings together many different facets of polymorphic drug biotransformation and features thorough reviews of drug-metabolizing reactions that show genetic variability between humans
2. Saltelli, A.; Chan, K.; Scott, EM. Sensitivity Analysis. John Wiley and Sons; Chichester, UK: 2000.
3. Suarez-Kurtz G, Pena SD. Pharmacogenomics in the Americas: the impact of genetic admixture. *Curr Drug Targets* 2006;7:1649–1658. Compilation of pharmacogenomics data from peoples of different geographic regions, where admixture and population stratification occur in distinct patterns. [PubMed: 17168840]
4. Jakobsson M, Scholz SW, Scheet P, et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 2008;451:998–1003. [PubMed: 18288195]
5. Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* 2003;73:1402–1422. [PubMed: 14631557]
6. Lao O, van Duijn K, Kersbergen P, de Knijff P, Kayser M. Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *Am J Hum Genet* 2006;78:680–690. [PubMed: 16532397]
7. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008;319:1100–1104. [PubMed: 18292342]
8. Rajeevan H, Cheung KH, Gadagkar R, et al. ALFRED: an Allele Frequency Database for Microevolutionary Studies. *Evolutionary Bioinformatics* 2005;31:270–271. Description of a valuable resource of gene-frequency data on anthropologically defined human populations.
9. Marsh S. Pharmacogenetics: global clinical markers. *Pharmacogenomics* 2008;9:371–373. [PubMed: 18384250]
10. Tang H, Choudhry S, Mei R, et al. Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am J Hum Genet* 2007;81:626–633. Very interesting use of a Puerto Rico asthma population for admixture mapping of potential disease loci. [PubMed: 17701908]
11. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. *Nat Genet* 2008;40:340–345. [PubMed: 18246066]
12. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol* 2006;4:E72. [PubMed: 16494531]
13. Ruano, G.; Windemuth, A.; Holford, T. Physiogenomics: integrating system engineering and nanotechnology for personalized medicine. In: Bronzino, JD., editor. *The Biomedical Engineering Handbook*. CRC Press; London, UK: 2005. Introduction to the fundamentals and applications of physiogenomics for the understanding of disease etiology and treatment and for the advancement of personalized medicine
14. Seip RL, Volek JS, Windemuth A, et al. Physiogenomic comparison of human fat loss in response to diets restrictive of carbohydrate or fat. *Nutr Metab (Lond)* 2008;5:4. [PubMed: 18254975]
15. Ruano G, Thompson PD, Windemuth A, et al. Physiogenomic analysis links serum creatine kinase activities during statin therapy to vascular smooth muscle homeostasis. *Pharmacogenomics* 2005;6:865–872. [PubMed: 16296949]
16. Ruano G, Bernene J, Windemuth A, et al. Physiogenomic comparison of edema and BMI in patients receiving rosiglitazone or pioglitazone. *Clin Chim Acta* 2009;400(1–2):48–55. [PubMed: 18996102]
17. Windemuth A, Calhoun VD, Pearlson GD, Kocherla M, Jagannathan K, Ruano G. Physiogenomic analysis of localized fMRI brain activity in schizophrenia. *Ann Biomed Eng* 2008;36:877–888. [PubMed: 18330705]
18. Liu J, Pearlson G, Windemuth A, Ruano G, Perrone-Bizzozero NI, Calhoun V. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Hum Brain Mapp* 2009;30(1):241–255. [PubMed: 18072279]

19. de Leon J, Correa JC, Ruaño G, Windemuth A, Arranz MJ, Diaz FJ. Exploring genetic variations that may be associated with the direct effects of some antipsychotics on lipid levels. *Schizophr Res* 2008;98:40–46. [PubMed: 18031993]
20. Ruaño G, Thompson PD, Windemuth A, et al. Physiogenomic association of statin-related myalgia to serotonin receptors. *Muscle Nerve* 2007;36:329–335. [PubMed: 17600820]
21. Ruaño G, Goethe JW, Caley C, et al. Physiogenomic comparison of weight profiles of olanzapine- and risperidone- treated patients. *Mol Psychiatry* 2007;12:474–482. [PubMed: 17199131]
22. Oliphant A, Barker DL, Stuelpnagel JR, Chee MS. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques Suppl* 2002;56–58:60–61.
23. Fan JB, Oliphant A, Shen R, et al. Highly parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol* 2003;68:69–78. [PubMed: 15338605]
24. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;155:945–959. [PubMed: 10835412]
25. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* 2007;7:574–578. [PubMed: 18784791]
26. Johnson SC. Hierarchical clustering schemes. *Psychometrika* 1967;32:241–254. [PubMed: 5234703]
27. Hartigan, J. *Clustering Algorithms*. John Wiley & Sons, Inc; NY, USA: 1975.
28. Hair, JF., Jr; Black, WC.; Babin, BJ.; Anderson, RE. *Multivariate Data Analysis*. Prentice Hall; NJ, USA: 1998.
29. Maindonald, J.; Braun, J. *Data Analysis and Graphics Using R*. Cambridge Press; Cambridge, UK: 2003.
30. Murtagh, F. *Multidimensional Clustering Algorithms*. Physica-Verlag; Wuerzburg, Germany: 1985.
31. The International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;437:1299–1320. Most complete haplotype map of the human genome, a widely available public resource to help researchers describe the common patterns of human DNA sequence variation. [PubMed: 16255080]
32. Estrela RC, Ribeiro FS, Carvalho RS, et al. Distribution of *ABCB1* polymorphisms among Brazilians: impact of population admixture. *Pharmacogenomics* 2008;9:267–276. [PubMed: 18303963]
33. Teixeira RL, Miranda AB, Pacheco AG, et al. Genetic profile of the arylamine *N*-acetyltransferase 2 coding gene among individuals from two different regions of Brazil. *Mutat Res* 2007;624:31–40. [PubMed: 17509624]
34. Bertoni B, Budowle B, Sans M, Barton SA, Chakraborty R. Admixture in Hispanics: distribution of ancestral population contributions in the Continental United States. *Hum Biol* 2003;75:1–11. [PubMed: 12713142]
35. Martinez-Cruzado JC, Toro-Labrador G, Viera-Vera J, et al. Reconstructing the population history of Puerto Rico by means of mtDNA phylogeographic analysis. *Am J Phys Anthropol* 2005;128:131–155. Fundamental study that revealed a substantial maternal contribution of Native Americans to the human gene-pool of Puerto Ricans by means of mitochondrial DNA phylogeographic analysis. [PubMed: 15693025]
36. Bonilla C, Parra EJ, Pfaff CL, et al. Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. *Ann Hum Genet* 2004;68:139–153. [PubMed: 15008793]
37. Choudhry S, Coyle NE, Tang H, et al. Population stratification confounds genetic association studies among Latinos. *Hum Genet* 2006;118:652–664. [PubMed: 16283388]
38. Martinez-Cruzado JC, Toro-Labrador G, Ho-Fung V, et al. Mitochondrial DNA analysis reveals substantial Native American ancestry in Puerto Rico. *Hum Biol* 2001;73:491–511. [PubMed: 11512677]
39. Hanis CL, Hewett-Emmett D, Bertin TK, Schull WJ. Origins of US Hispanics. Implications for diabetes. *Diabetes Care* 1991;14:618–627. [PubMed: 1914811]
40. Hammer MF, Chamberlain VF, Kearney VF, et al. Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. *Forensic Sci Int* 2006;164:45–55. [PubMed: 16337103]

41. Kayser M, Brauer S, Schadlich H, et al. Y chromosome STR haplotypes and the genetic structure of US populations of African, European, and Hispanic ancestry. *Genome Res* 2003;13:624–634. [PubMed: 12671003]
42. Redd AJ, Chamberlain VF, Kearney VF, et al. Genetic structure among 38 populations from the United States based on 11 US core Y chromosome STRs. *J Forensic Sci* 2006;51:580–585. [PubMed: 16696705]
43. Budowle B, Adamowicz M, Aranda XG, et al. Twelve short tandem repeat loci Y chromosome haplotypes: genetic analysis on populations residing in North America. *Forensic Sci Int* 2005;150:1–15. [PubMed: 15837004]
44. Bonilla C, Shriver MD, Parra EJ, Jones A, Fernandez JR. Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York city. *Hum Genet* 2004;115:57–68. [PubMed: 15118905]
45. Novembre J, Johnson T, Bryc K, et al. Genes mirror geography within Europe. *Nature* 2008;456:98–101. [PubMed: 18758442]
46. Parra EJ, Kittles RA, Shriver MD. Implications of correlations between skin color and genetic ancestry for biomedical research. *Nat Genet* 2004;36:S54–S60. [PubMed: 15508005]
47. Gonzalez BE, Borrell LN, Choudhry S, et al. Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. *Am J Public Health* 2005;95:2161–2168. Valuable review of the formation of Hispanic populations with some examples of clinical, social, epidemiological and genetic interest to highlight the importance and potential benefits of studying Hispanics. [PubMed: 16257940]
48. Salari K, Choudhry S, Tang H, et al. Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. *Genet Epidemiol* 2005;29:76–86. [PubMed: 15918156]
49. Jennings, J. Across an Arctic Bridge. In: Billard, J., editor. *The World of the American Indian*. National Geographic Society; Washington, DC, USA: 1979.
50. Stewart, TD. *The People of America*. Scribner; New York, NY, USA: 1973.
51. Wang S, Lewis CM, Jakobsson M, et al. Genetic variation and population structure in native Americans. *PLoS Genet* 2007;3:E185. [PubMed: 18039031]
52. Choudhry S, Burchard EG, Borrell LN, et al. Ancestry-environment interactions and asthma risk among Puerto Ricans. *Am J Respir Crit Care Med* 2006;174:1088–1093. [PubMed: 16973984]
53. Mao X, Bigham AW, Mei R, et al. A genomewide admixture mapping panel for Hispanic/Latino populations. *Am J Hum Genet* 2007;80:1171–1178. [PubMed: 17503334]
54. Ruano G. *Quo Vadis* personalized medicine? *Personalized Medicine* 2004;1:1–7.

Websites

101. USA Census Bureau. 2000. www.census.gov
102. The University of Chicago Department of Biological Sciences, Pritchard Lab .
<http://pritch.bsd.uchicago.edu>
103. Online supplement
www.futuremedicine.com/doi/suppl/10.2217/PGS.09.5/suppl_file/RuanoSuppl.pdf

Patent

201. Ruano, G. Physiogenomic method for predicting clinical outcomes of treatments in patients 20060278241. 2006. p. 12-14.

Figure 1. Representative sampling (n = 100) per geographic regions based on percentage of birth at each region around the island of Puerto Rico according to the 2004 National Birth Registry
Colors are employed only to highlight the different municipalities in Puerto Rico.

Figure 2. Population stratification of the Puerto Rico sample as represented by the STRUCTURE version 2.2 triangle plot

The diagram represents 71 samples with complete combinatorial genotype results. Samples between the dashed lines and the respective vertices of the triangle are assigned to a specific cluster, but samples bordering the dashed lines or between the dashed lines and the center are not. The total number of individuals assigned to the three clusters was 53. The cluster-specific assignment subtotals were 21 individuals for cluster 1, ten for cluster 2 and 22 for cluster 3. Some samples overlap in their position in the triangle.

Figure 3. Genetic distance dendrogram for 98 individuals genotyped in the Puerto Rican population sample

(B) Dendrogram of **(A)**, with identical sample identifications, color coded to denote the 53 individuals assigned to one of the three clusters in Figure 2. Cluster 1 samples (n = 21) are depicted in red, cluster 2 samples (n = 10) are depicted in green and cluster 3 samples (n = 22) are depicted in blue. The diagram uses a dashed line to highlight the sectors of the dendrogram with disproportionate representation of samples assigned to a given cluster. The p-values are the results of χ^2 tests comparing the actual number of cluster 2 samples in sector 1, the number of cluster 3 samples in sector 2 and the number of cluster 1 samples in sector 3 to expected values given a random distribution.

Figure 4. Population dissimilarity between the three clusters (cluster 1, cluster 2, cluster 3) found by STRUCTURE in the Puerto Rico sample of 71 individuals with complete combinatorial genotypes compared with three International HapMap reference populations of different ethnicity (CEU, HCB, YRI)

CEU represents individuals of European descent, HCB represents individuals of Asian origin and YRI represents individuals of sub-Saharan African origin. For each of the reference populations, the y-axis plots the relevant dissimilarity in allele frequency based on RMS matches to the individual clusters. The best matching cluster for each reference population is indicated by a circle.

CEU: Northern and Western Europe; HCB: Han Chinese from Beijing; RMS: Root mean square; YRI: Yoruba people from Idaban, Nigeria.

Figure 5. Ancestry contributions calculated using STRUCTURE for each of the 71 individuals from the Puerto Rican sample with complete combinatorial genotypes

Individuals are aligned according to their highest relative contributions of ancestry from cluster 1 (red), samples to the left of diagram; cluster 2 (green), samples to the center; and cluster 3 (blue), samples to the right.