

# Genetic and socioeconomic study of mate choice in Latinos reveals novel assortment patterns

James Y. Zou<sup>a,b,1</sup>, Danny S. Park<sup>c</sup>, Esteban G. Burchard<sup>c,d</sup>, Dara G. Torgerson<sup>c,d</sup>, Maria Pino-Yanes<sup>c,d</sup>, Yun S. Song<sup>e,f,g</sup>,  
Sriram Sankararaman<sup>h,2</sup>, Eran Halperin<sup>i,2</sup>, and Noah Zaitlen<sup>d,2</sup>

<sup>a</sup>Microsoft Research, Cambridge, MA 02142; <sup>b</sup>Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02142; <sup>c</sup>Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94143; <sup>d</sup>Department of Medicine, University of California, San Francisco, CA 94143; <sup>e</sup>Computer Science Division, University of California, Berkeley, CA 94720; <sup>f</sup>Department of Statistics, University of California, Berkeley, CA 94720; <sup>g</sup>Department of Integrative Biology, University of California, Berkeley, CA 94720; <sup>h</sup>Department of Genetics, Harvard Medical School, Boston, MA 02115; and <sup>i</sup>Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69102, Israel

Edited by Andrew G. Clark, Cornell University, Ithaca, NY, and approved June 4, 2015 (received for review January 26, 2015)

**Nonrandom mating in human populations has important implications for genetics and medicine as well as for economics and sociology. In this study, we performed an integrative analysis of a large cohort of Mexican and Puerto Rican couples using detailed socioeconomic attributes and genotypes. We found that in ethnically homogeneous Latino communities, partners are significantly more similar in their genomic ancestries than expected by chance. Consistent with this, we also found that partners are more closely related—equivalent to between third and fourth cousins in Mexicans and Puerto Ricans—than matched random male–female pairs. Our analysis showed that this genomic ancestry similarity cannot be explained by the standard socioeconomic measurables alone. Strikingly, the assortment of genomic ancestry in couples was consistently stronger than even the assortment of education. We found enriched correlation of partners’ genotypes at genes known to be involved in facial development. We replicated our results across multiple geographic locations. We discuss the implications of assortment and assortment-specific loci on disease dynamics and disease mapping methods in Latinos.**

assortative mating | genomic ancestry | homogamy

Mate selection in human populations is a topic of broad general and academic interest, with important implications for economics (1), sociology (2), psychology (3), and genetics (4, 5). Studies from all of these fields have uncovered myriad genetic and socioeconomic factors that are correlated between partners. Assortment by education, for example, is one of the strongest and most widely studied effects, with studies showing that partners tend to have similar numbers of years of schooling (6). Recent collections of phenotype-rich population-scale genomic datasets provide the opportunity to combine genomic analysis with a range of socioeconomic measurements, such as education, on the same set of individuals. By integrating these diverse sources of information, we seek to better understand how genetics relates to each of the socioeconomic factors, as well as how all these factors jointly contribute to and are reflected by human mating patterns.

Recent work from several groups suggests that genetic factors contribute to human mating preferences (5, 7–10). In European populations, there is evidence that individuals tend to be attracted to people with major histocompatibility complex (MHC) regions that are distinct from their own (7, 8). In Latinos, genotypes of ancestry-informative markers were shown to be positively correlated in partners (5). A recent study of non-Hispanic whites in the United States found that spouses were more genetically similar than random pairs of individuals; however, genetic assortment was substantially smaller than educational assortment in the same couples (11).

In this paper, we investigate genetic and socioeconomic patterns in assortative mating within Mexican and Puerto Rican populations in the United States, Mexico, and Puerto Rico. In particular, we explore how genomic ancestry correlates with education, income, location, and mating patterns. These ethnically homogeneous but genetically admixed populations have more diverse physical features compared with genetically homogeneous populations (12).

Moreover, research in anthropology and sociology has documented significant discrimination based on skin color and facial features in these communities (13, 14). In Yucatan, Mexico, for example, Spanish surnames are thought to carry a higher social status than native Mayan surnames, and individuals are more likely to marry someone with a surname of the same ethnic origin (15). Together, these findings suggest that there might be important genetic components that associate with mating patterns.

In contrast to non-Hispanic whites, we find that the assortment of genomic ancestry in Mexican and Puerto Rican couples is significantly greater than the assortment of education in the same couples. This result is consistent across regions in Mexico, Puerto Rico, and the United States and cannot be explained by socioeconomic factors alone. Consistent with this finding, we find that partners are significantly more closely related than matched random male–female pairs from the same locations. The genetic relatedness of partners is equivalent to that between third and fourth cousins in both Mexicans and Puerto Ricans. Further, we find that couples are especially more likely to share similar genotypes at genes involved in facial development. The similarity at these genes is stronger than the global genome-wide correlations, suggesting that similarity of facial features can be an axis of assortment.

Assortment in the genomic ancestries of couples can have significant impact on the propagation of disease burdens through generations. For example, in an assortative mating population, recessive founder effects, such as those hypothesized to underlie the increased prevalence of asthma in Puerto Rico (16), will be

## Significance

**Nonrandom mating in human populations reflects a confluence of socioeconomic and biological factors. In this paper, we integrate population genomics with quantitative social sciences to address fundamental questions about mate selection. In Mexican and Puerto Rican couples, we find that partners share more similar genomic ancestries, which is characterized as the fractions of genome that come from European, Native American, and African ancestries, than random pairs of individuals from the same community. Our analysis shows that this similarity cannot be explained by the individuals’ socioeconomic factors alone. Our mathematical model illustrates how assortative mating can increase the prevalence of recessive diseases in Latino populations.**

Author contributions: J.Y.Z., D.S.P., S.S., E.H., and N.Z. designed research; J.Y.Z., D.S.P., S.S., E.H., and N.Z. performed research; J.Y.Z., D.S.P., E.G.B., D.G.T., M.P.-Y., Y.S.S., S.S., E.H., and N.Z. contributed new reagents/analytic tools; J.Y.Z. and D.S.P. analyzed data; and J.Y.Z., S.S., E.H., and N.Z. wrote the paper.

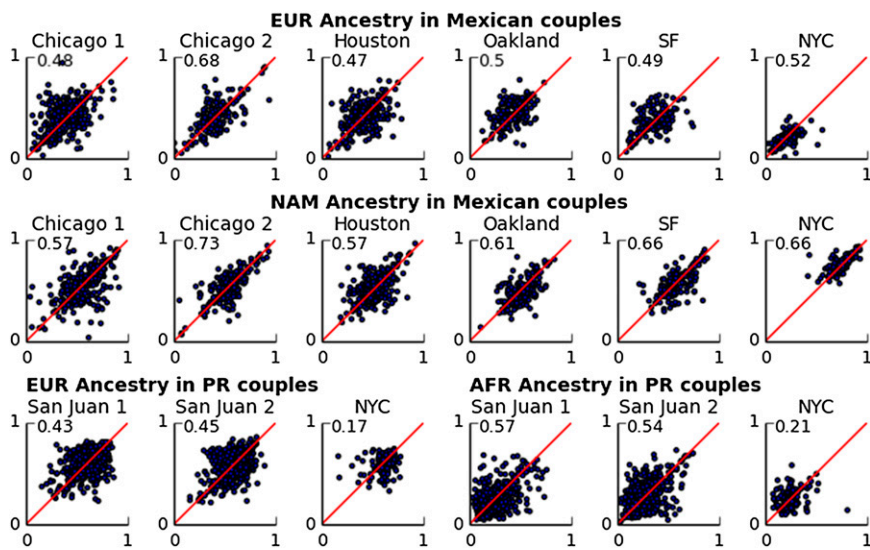
The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

<sup>1</sup>To whom correspondence should be addressed. Email: jazo@microsoft.com.

<sup>2</sup>S.S., E.H., and N.Z. contributed equally to this work.

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1501741112/-DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1501741112/-DCSupplemental).



**Fig. 1.** Correlation of partners' genomic ancestries. Each dot is a couple, and the x and y axes correspond to the genomic ancestries of the two individuals. The couples are grouped by their current location. (*Top*) European (EUR) ancestry in GALAI Mexican couples. (*Middle*) Native American (NAM) ancestry in the same Mexican couples. (*Bottom*) European (*Left*) and African (AFR) (*Right*) ancestries in GALAI Puerto Rican (PR) couples. The Pearson correlation value is shown in each plot. The least common ancestry for each ethnic group (African ancestry for Mexicans and Native American ancestry for Puerto Ricans) was omitted for clarity.

maintained longer and lead to increased disease prevalence than in random mating populations. We analyze a simple mathematical model to quantitatively illustrate this effect.

## Results

We analyzed 1,246 Mexican trios (father–mother–offspring) and 1,511 Puerto Rican trios. Extensive socioeconomic data were collected for each individual as a part of the Genes-Environments & Admixture in Latino Americans II (GALAI) study (17, 18) (*Materials and Methods* and *SI Appendix*, Table S1). All of the offspring were genotyped, and the genomic ancestries of the offspring and each of their parents were computationally inferred. To validate our findings, we also analyzed 231 Mexican trios and 258 Puerto Rican trios from the Genetics of Asthma in Latino Americans I (GALAI) study (19, 20), where each member of the trio was genotyped.

**Genomic Ancestries of Latino Populations.** We first characterized the genomic ancestries of the Mexicans and Puerto Ricans in our datasets. The genomic ancestry of an individual was defined to be the fraction of his/her genome that came from European, Native American, and African ancestries. We computed the global genomic ancestry of each individual by first inferring the local ancestry of each single-nucleotide polymorphism (SNP) using LAMP-LD (21) and then aggregating the results (see *SI Appendix* for details). Mexican participants were primarily of Native American and European ancestries, with a small amount of African ancestry (*SI Appendix*, Fig. S1). Puerto Ricans were primarily of European ancestry, with a moderate amount of African and Native American ancestries (*SI Appendix*, Fig. S2). There was significant variation in the genomic ancestries of the individuals within each sampling location (Mexicans were sampled in six sites from five cities, and Puerto Ricans were sampled in three sites from two cities) but the ancestry distributions were consistent across the cities. Mexicans living in New York City, however, were an outlier population in that they had a significantly higher fraction of Native American ancestry compared with Mexicans in other US cities. This result reflects the fact that the majority of Mexicans in New York City originally came from Puebla, Mexico, a region whose population had significantly higher Native American ancestry and lower European ancestry compared with other Mexican populations ( $P < 0.01$ ,  $t$  test) (*SI Appendix*, Fig. S3).

We also computed the X chromosome ancestry using ADMIXTURE (22). In both the Mexican and Puerto Rican participants, the X chromosomes contained significantly higher Native American ancestry compared with the autosomes ( $P < 10^{-4}$ ,  $t$  test in each collection center). These findings were consistent with the hypothesis that in initial admixture events during colonization,

the females were more likely to be Native Americans and Africans whereas the males were more likely to be Europeans (23).

**Correlations in the Ancestries of Partners.** We next sought to determine whether the genomic ancestries of partners in couples were correlated. The genomic ancestry of each individual in a couple was computationally inferred by applying ANCESTOR (24) to the genotype of the couple's offspring, because the GALAI study only genotyped the offspring, not the parents. Using the GALAI validation dataset (in which both parents and offspring were genotyped) and simulations, we verified that ANCESTOR accurately imputed genomic ancestry and did not introduce false correlation into the ancestries of the partners (*SI Appendix*). We found that the genomic ancestries of partners were substantially correlated in both Mexicans and Puerto Ricans (Fig. 1). This finding was replicated in each of the cities from which we had participants. The correlation was strongest for the primary ancestries (Native American and European ancestries for Mexicans; European and African ancestries for Puerto Ricans) ( $P < 0.01$ , Pearson correlation). For the less common ancestries (African ancestry for Mexicans; Native American ancestry for Puerto Ricans) the correlations were not significant, possibly because the estimates were noisier.

Because many of the Mexican individuals were recent immigrants to the United States, we also grouped the individuals by their state of birth rather than their current city. Focusing on couples in which both partners were born in the same state, we again found that the genomic ancestries of partners were significantly correlated ( $P < 0.01$ , Pearson correlation) (*SI Appendix*, Fig. S4). This result was replicated in all five states where we had sufficient sample size (at least 40 couples). We additionally validated this finding using the GALAI dataset, in which the genomic ancestry was computed directly from each individual's genotype using LAMP-LD. We found significant correlations in the genomic ancestries of Mexican (*SI Appendix*, Fig. S5) and Puerto Rican (*SI Appendix*, Fig. S6) partners ( $P < 0.01$ , Pearson correlation). This result is consistent with previously reported correlations in 104 ancestry-informative markers in the same participants (5). Together, these results demonstrated that the genomic ancestries of partners in couples were significantly correlated in both Mexicans and Puerto Ricans.

**Ancestry and Education Assortment in Couples.** We next sought to precisely quantify the magnitude of genomic ancestry assortment in couples and compare it with the magnitude of education assortment. To quantify the assortment of genomic ancestry while accounting for ancestry distribution at the population level in each location, we computed genomic ancestry assortative mating (GAAM) scores following the procedure proposed by Domingue

et al. (11). This score quantifies the magnitude of assortment above population background and can take a value from  $-0.5$  (each person is paired with the most different partner) to  $0.5$  (complete assortment, where each person is paired with the most similar partner). A score of  $0$  corresponds to uniform random mating. To compute the GAAM scores, we computed the squared difference in genomic ancestry of the observed couples and random pairs. Quantiles of the observed similarities were plotted against quantiles of the similarities that would be expected given random mating. For example, if this curve has a value of  $0.6$  at the midpoint, this means that the median couple has genomic ancestries that are more similar than  $60\%$  of all random pairs. A curve that is higher above the diagonal indicates that couples are more similar than expected. To quantify this magnitude, the GAAM score was defined as the area between the curve and the  $45^\circ$  line.

We grouped couples by their location and computed the GAAM score in each location separately (Fig. 2). The GAAM scores for Mexican couples ranged from  $0.14$  to  $0.19$  across six sites, and the GAAM scores for Puerto Rican couples ranged from  $0.1$  to  $0.13$  across three sites. These scores indicated significant assortment of genomic ancestry in each site ( $P < 0.01$ , permutation test), although the magnitude of assortment was substantially lower in Puerto Ricans compared with Mexicans. To better interpret the GAAM scores, we used a logistic regression model to translate the GAAM scores into couple odds ratios (SI Appendix). A GAAM score of  $0.12$ – $0.23$  corresponds to a couple odds ratio of  $1.2$ – $1.4$  (SI Appendix, Fig. S8). This means that a  $10$  percentile increase in the ancestry similarities of two individuals corresponds to a  $20$ – $40\%$  increase in the odds that they form a couple.

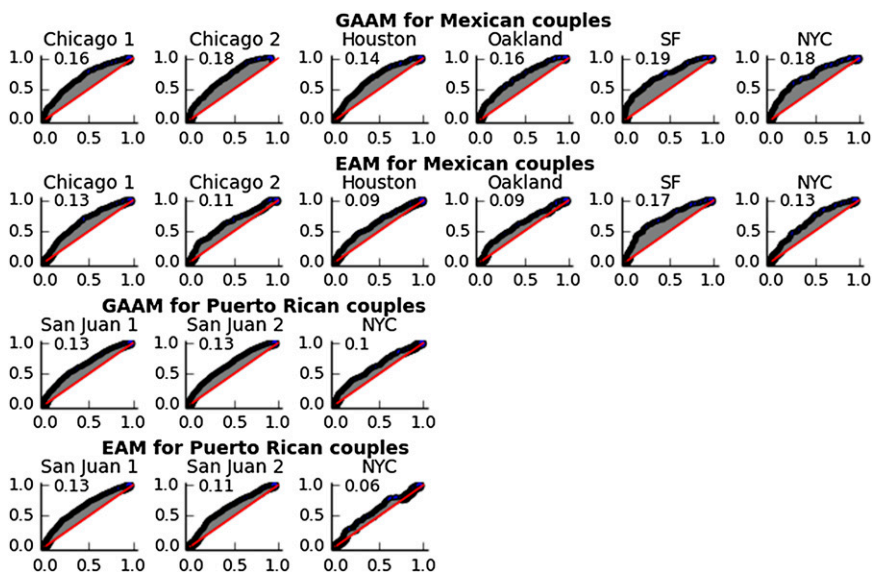
To place the magnitude of genomic ancestry assortment scores into context, we computed education assortment scores in the same populations as a benchmark. Education was a natural comparison, because extensive research has shown that partners tend to have similar education levels (6). In a cohort of non-Hispanic whites in the United States, a recent analysis (11) estimated an education assortative mating (EAM) score of  $0.127$ . Using the same approach, we computed the EAM scores in our populations, where the education level of an individual was quantified as the number of years of schooling. Grouping couples by location, we found EAM scores from  $0.09$  to  $0.17$  in Mexicans and scores from  $0.06$  to  $0.13$  in Puerto Ricans. The magnitude of education assortment in Mexican and Puerto Rican couples was consistent with the education assortment in non-Hispanic whites measured by Domingue et al. (11). Surprisingly, the GAAM scores were higher than the EAM scores in every location. Using bootstrapping, we found the difference between GAAM and EAM to be significant

at the  $P < 0.05$  level in Chicago, Houston, and Oakland for Mexicans and in San Juan for Puerto Ricans. This finding suggests that the assortment of genomic ancestry in Mexicans and Puerto Ricans was stronger than the assortment of education.

As a validation, we stratified individuals by their state of birth rather than by their current location. Across all states, we observed significant assortment of genomic ancestry ( $P < 0.01$ , permutation test), with GAAM scores ranging from  $0.12$  to  $0.23$  (SI Appendix, Fig. S7). We also computed the EAM scores, again stratifying by state of birth, and found EAM scores ranging from  $0.09$  to  $0.16$ . In four of the five Mexican states, the genomic ancestry assortment was higher than education assortment, and the differences in Guerrero and Puebla were significant, with  $P < 0.05$  by bootstrapping.

**Correlations Between Ancestry and Socioeconomic Factors.** To interpret the substantial similarity in the ancestries of partners, it is important to understand how much of this similarity is confounded by socioeconomic factors. In Puerto Ricans, we found that the amount of European ancestry was positively correlated with education level ( $P < 0.05$ ,  $t$  test) (SI Appendix, Table S2). Although the correlation was statistically significant, the variance in education that was explained by ancestry was small ( $r^2 < 0.07$  in all of the locations). In Mexicans, in three of the six sites, we found that European ancestry was also positively correlated with education ( $P < 0.05$ ,  $t$  test), although the variance that was explained was also small ( $r^2 < 0.05$ ) (SI Appendix, Table S2). Applying a similar analysis to income, we found a positive correlation between the amount of European ancestry and family income in several locations for Mexicans and Puerto Ricans ( $P < 0.05$ ,  $t$  test) (SI Appendix, Table S3). Again, however, the variance in income that was explained by ancestry was small ( $r^2 < 0.08$  in all of the locations). Interestingly, European ancestry remained positively correlated with income in these locations even after accounting for education level ( $P < 0.05$ ,  $t$  test). Together, these results demonstrated that although ancestry was correlated with education and income, the effect sizes were small, suggesting that the assortment of education and income cannot fully explain the assortment of genomic ancestry in partners.

To test this hypothesis more directly, we used a linear regression model to predict the genomic ancestry of one member of a couple using the ancestry of their partner as well as socioeconomic and demographic covariates (education, family income, birth location, and current location) of both individuals. When combined, all of the socioeconomic and demographic covariates of individuals explained less than  $10\%$  of the variations in genomic ancestry of their partners. In contrast, in every region, the ancestry of individuals explained



**Fig. 2.** Genomic ancestry assortment scores compared with education assortment scores in Mexican and Puerto Rican couples. The x axis is the percentile of the actual couples' similarity scores, and the y axis is the percentile of the random pairs' similarity scores. In each plot, the black curve shows how the distribution of similarity scores in the real couples matches the distribution in random couples. Couples are more similar than expected by random mating when the black curve is above the  $45^\circ$  diagonal (red line). The assortment score, defined as the area between the black curve and the diagonal, is shown for each plot (gray). The GAAM (EAM) plots for GALAII Mexicans are shown in the first (second) row, and the GAAM (EAM) plots for GALAII Puerto Ricans are shown in the third (fourth) row. Couples are grouped by their current location.



20–50% of the variations in the genomic ancestry of their partners, after adjusting for socioeconomic covariates, and was highly statistically significant ( $P < 10^{-5}$ ). Therefore, the genomic ancestry of individuals is a stronger predictor of their partners' genomic ancestry than are socioeconomic and demographic covariates.

**Positive Correlations in Facial Development Genes.** Because the standard socioeconomic measurables alone did not explain the similarities in the genomic ancestries of partners, we investigated potential biological factors associated with assortative mating. We took a candidate gene-set approach focused on genes involved in facial development. Psychology studies have reported that people find faces that are more similar to their own more attractive (25). In the GALAI dataset, we computed the correlation between partners at each SNP in a set of 49 genes previously curated to be related to human facial development (26). We pooled all 49 genes into one analysis to gain statistical power. To adjust for ancestry, we normalized the SNP's allele frequency by its expected frequency given its local ancestry state. We found that Puerto Rican partners in the GALAI dataset were significantly more likely to have similar genotypes in these facial development genes than expected based on correlations in the rest of the genome ( $P = 0.004$ , permutation test). In Mexican couples, the enrichment was not significant. We performed the same analysis on two other gene sets—86 genes linked to genome-wide association study (GWAS) height loci and 8 genes linked to pigmentation (27)—and did not find a significant deviation from genome-wide correlations in either population. This lack of deviation may be due to the relatively small effects of these loci and our limited understanding of the relevant genes (28).

We validated the similarities in facial development genes between Puerto Rican partners using the Puerto Rican individuals from the GALAI dataset. Because only the offspring, and not the couple, were genotyped in the GALAI study, we assessed the couples' similarity in facial development genes by measuring the homozygosity in the genotype of the offspring at these loci. The magnitude of homozygosity was quantified by deviations from the Hardy–Weinberg equilibrium. The SNPs in the facial development gene set showed significantly higher homozygosity compared with the rest of the genome ( $P = 0.02$ , permutation test). Together, these results suggest that similarities in facial features could be a factor in mate selection in Puerto Ricans.

**Excessive Heterozygosity in MHC.** Previous studies of European-American couples found that partners were significantly more dissimilar in the MHC region compared with random male–female pairs, although it is still unclear to what extent this phenomenon applies in other populations (8). This finding had appealing evolutionary implications, because offspring with more diverse MHC loci can have increased immune recognition and better fitness (29). Consistent with this finding, we found that both Mexicans and Puerto Ricans from the GALAI dataset showed excessive heterozygosity relative to the rest of the genome in the genotypes at the MHC region ( $P = 0.005$  in Mexicans and  $P = 0.003$  in Puerto Ricans, as tested via deviation from Hardy–Weinberg). We also tested for deviations from

Hardy–Weinberg in the local ancestry of the MHC region. In Puerto Ricans, but not Mexicans, local ancestry at the MHC region was more heterozygous than expected ( $P = 0.01$ , permutation test).

**Genetic Relatedness of Couples.** Our finding that partners had similar genomic ancestries suggested that they could also be more recently related. We investigated this hypothesis in couples from the GALAI study, where we had the genotypes of both individuals. To quantify the relatedness of two individuals, we used the standard metric of identical by descent (IBD), which we computed using PIGS (30) (SI Appendix). Two haplotypes are IBD if they are inherited copies from a common ancestor. As a control set, we generated random male–female pairings for each ethnic group at each collection site. Random pairings were generated to reflect the observed global ancestry correlations seen in the actual couples (SI Appendix). In both Mexicans and Puerto Ricans, we found that partners were more closely related than random male–female pairs from the same collection site ( $P < 0.01$ , permutation test) (Table 1). The average IBD sharing between Mexican couples was 0.36% of the genome in San Francisco and 0.35% in Mexico City, compared with 0.19% and 0.26% in ancestry-matched random male–female pairs. The average IBD sharing between Puerto Rican couples was 0.70% of the genome in New York City and 0.91% in San Juan, compared with 0.42% and 0.54% in ancestry-matched random male–female pairs. The higher IBD sharing in Puerto Rican couples compared with Mexican couples may be due to the relatively small island population of Puerto Rico. To place these findings into context, the expected IBD sharing is 0.78% of the genome between third cousins and 0.19% of the genome between fourth cousins. Thus, the IBD sharing of Mexicans and Puerto Ricans fell between that of third and fourth cousins. In addition, both Mexican and Puerto Rican couples shared significantly more IBD blocks as well as longer IBD blocks compared with ancestry-matched random couples ( $P < 0.01$ , permutation test). This result is also consistent with partners sharing more recent relations than random couples.

**Correlations Between Genetic Relatedness and Average Education Level.** We next asked whether the genetic relatedness of couples correlated with socioeconomic factors such as education. We analyzed the genetic relatedness of partners in the GALAI dataset, which included socioeconomic information about the couples but not their genotypes. Because we did not have the partners' genotypes to perform direct IBD analysis, we measured runs of homozygosity (ROHs) in the genomes of the offspring as a proxy. Long ROHs in offspring reflect recent relatedness of the parents (31) (SI Appendix). We asked whether more recent relatedness between partners (measured by the total length of long ROHs in the offspring) correlated with the average education levels of the partners. We included the European, Native American, and African genomic ancestries of the offspring as additional covariates in the analysis to correct for differences in the ROH length due to different ancestries. Among Puerto Ricans in San Juan and Mexicans in Chicago, Houston, and New York City, we found a significant ( $P < 0.05$ ) negative correlation between the partners' average education and

**Table 1. Genetic similarities between partners**

Population	Location	Actual couples			Matched random pairs		
		Mean IBD sharing, %	Mean number of IBD blocks	Mean length of IBD blocks, cM	Mean IBD sharing, %	Mean number of IBD blocks	Mean length of IBD blocks, cM
Mexican	San Francisco	0.36	15.5	1.4	0.19	8.3	1.2
	Mexico City	0.35	15.7	1.4	0.26	13.9	1.3
Puerto Rican	New York City	0.70	20.4	2.1	0.42	17.3	1.8
	San Juan	0.91	25.3	2.2	0.54	18.0	1.9

Identical-by-descent sharing in Mexican and Puerto Rican couples. IBD sharing was computed for GALAI couples, and IBD sharing for ancestry-matched random pairs was also computed as a benchmark. In each location, the differences in IBD sharing, number of blocks, and length of blocks between the actual couples and matched random pairs were significant ( $P < 0.01$ , permutation test).

their recent relatedness. The results in the other locations were not statistically significant, but a majority exhibited negative correlations as well. We reran the analysis for Mexican partners who were born in the same state and found a similar negative correlation in Jalisco, Mexico ( $P < 0.05$ ). The results in the other four states that we tested were not statistically significant, but three of the four also exhibited a negative correlation. In contrast, ancient haplotype sharing, as measured by short ROHs (31), was not significantly correlated with education in either direction in any of the locations. These findings suggest that partners with higher average education levels tend to be genetically more dissimilar.

**Disease Implications.** The ancestry-based assortment that we have characterized in Mexicans and Puerto Ricans can have a significant impact on the population dynamics and propagation of disease burdens. We analyzed a simple mathematical model to illustrate this effect. In our model there are two founder populations, A and B, that admix, with  $\alpha$  proportion of the individuals coming from A. Suppose we have a recessive risk allele present at frequency  $f$  in A and frequency 0 in B; that is, this risk allele is specific to ancestry A. After one generation of random mating, the fraction of the total admixed population at risk for the disease is the homozygosity of this risk allele,  $\alpha^2 f^2$ . With assortative mating, when the correlation between the genomic ancestries of the partners is  $\rho$ , the homozygosity of the risk allele after  $t$  generations of admixture is  $\alpha^2 f^2 + \alpha(1-\alpha)f^2\rho^{\frac{1+\rho}{2}}$  (see *SI Appendix* for a detailed derivation). The second term captures the inflation of the disease risk due to assortative mating. When the correlation between the ancestries of the partners is high,  $\rho$  is large and the homozygosity of the risk allele is substantially larger than under random mating. As an example, suppose that  $\rho$  is between 0.4 and 0.7, which was the range of ancestry correlation that we quantified in Mexican and Puerto Rican couples. Assuming that the founder populations have similar sizes, this range of assortative mating corresponds to a 2–14% increase in the prevalence of the recessive disease after 10 generations of admixture. Although this is a simplified model, it illustrates that in a population with strong assortative mating, recessive diseases associated with specific ancestry can have increased prevalence. This analysis can provide insights into the elevated prevalence for asthma and Hermansky–Pudlak syndrome, a recessive lung disease, in Puerto Ricans (32).

## Discussion

We have presented a rich analysis of assortative mating in Latino populations that integrates population genomics with socioeconomic attributes including education, location, and income. We found that partners are likely to have similar genomic ancestries. Strikingly, the assortment by genomic ancestry was stronger than even the assortment of education level and appeared to be a widespread phenomenon in all of the regions we investigated. We showed that the similarities in partners' ancestries cannot be explained by the standard socioeconomic measurables alone. By quantifying the genotype correlations between partners, we found that Puerto Rican couples were especially similar in genes involved in facial development.

There are multiple models consistent with this finding on facial structure genes. Socioeconomic factors could contribute to this genotypic correlation in a variety of ways. For example, individuals may use certain facial features as markers of social status and cultural background, which factor into their preference for partners. In communities where there is substantial preference for partners of similar status or background, this can be reflected in enriched correlations at facial feature genes. Another model that is consistent with our finding is the narcissism hypothesis, which shows that people find faces that are more similar to their own to be more attractive (25). This preference can also manifest in partners having correlated face genotypes. It is important to note that these two models are not mutually exclusive. That individuals find self-resembling features to be attractive, either consciously or subconsciously, could be due to cultural and socioeconomic as well as evolutionary factors.

We quantified the magnitude of the genomic ancestry assortment in terms of the GAAM scores. Across Mexican and Puerto

Rican populations in multiple locations, the GAAM scores ranged from 0.1 to 0.19. A recent study measuring genetic assortment in non-Hispanic white couples in the United States reported an assortment score of 0.045, which is substantially smaller than the GAAM scores of our populations (11). This difference could be due to the fact that the assortment score for the non-Hispanic whites was computed from genetic similarity rather than ancestry similarity, because all of the participants were of European ancestry (11). To make a more direct comparison, we analyzed the genotypes of the GALAI Mexican and Puerto Rican couples and computed the genetic assortative mating (GAM) scores using the same methodology as was used for non-Hispanic white couples (11). We found GAM scores of 0.103 and 0.063 in GALAI Mexican and Puerto Rican populations (*SI Appendix*, Fig. S10), which are significantly larger than the GAM score of 0.045 reported in non-Hispanic whites (11) ( $P < 10^{-4}$  for Mexicans and  $P = 0.003$  for Puerto Ricans, bootstrapping test).

Our analysis focused on genomic ancestry as a key variable because it is a meaningful statistic both for biology and social sciences. It also has the advantage of being accurately inferred from the genotype of the offspring, thus making it possible to perform our analysis on the GALAI couples where we only had the offspring's genome. We extended our analysis to also quantify the genetic relatedness of Mexican and Puerto Rican couples from the GALAI dataset, which contained the genotype of all three individuals in father–mother–offspring trios. Consistent with the genomic ancestry correlations, we showed that partners were more closely related than expected from random mating. The average relatedness between Mexican and Puerto Rican spouses was equivalent to that between third and fourth cousins. It would be important to better understand the social and demographic implications of this range of genetic relatedness in parents. In the Icelandic population, for example, fecundity is reported to be maximal when the relatedness between spouses is equivalent to fourth cousins (33).

The high level of correlations in the genomic ancestry of partners has significant implications for the prevalence of diseases in Mexicans and Puerto Ricans. In particular, our mathematical model illustrates how recessive risk alleles have inflated homozygosity in populations with strong assortative mating. An important agenda in epidemiology has been to understand the inflated prevalence of asthma and lung diseases such as Hermansky–Pudlak syndrome in Puerto Ricans (17, 32). Our analysis suggests that assortative mating can be a significant factor, and it should be incorporated into epidemiological models of these diseases. Our finding that individual gene sets (MHC and facial morphology genes) are correlated with assortative mating also has implications for disease mapping in these populations. Genes positively correlated with assortment will harbor excessively long haplotypes. Many widely used analysis strategies, such as IBD and homozygosity mapping, can be confounded by this effect, leading to false discoveries of disease-associated loci.

Like all such studies, our results are contingent on the ascertainment scheme. The participants in our dataset were recruited to be diverse in terms of genetics, environmental exposures, and socioeconomic status. We find it reassuring that the main finding—that the genomic ancestry assortment is comparable or stronger than education assortment—was reproduced in all six Mexican cohorts (from five cities) and in all three Puerto Rican cohorts (from two cities). An important area of future research would be to extend our analysis to other populations with different levels of genetic and ethnic diversity. Our datasets of Mexicans and Puerto Ricans occupy an interesting middle ground between ethnic homogeneity and genetic heterogeneity. Each dataset was ethnically homogeneous, because couples were composed of either two self-identified Mexicans or two self-identified Puerto Ricans. On the other hand, within each location, there were substantial variations in the genomic ancestries of the individuals. The participants filled a wide spectrum of distributions of European, Native American, and African ancestries, and no individual was purely of one race. The heterogeneity of genomic ancestries was reflected in the diversity of physical features, which can influence mate selection. Moreover, this influence can be compounded by cultural attitudes, such as ones that associate European-looking

features with higher social status (34). It would be interesting to compare and contrast our findings with analysis of genomic assortment in African Americans, another population that is genetically diverse and has been reported to share similar cultural attitudes (35).

In this paper, we have demonstrated how population genomics can be integrated with quantitative social sciences to address fundamental questions about mate selection. Our findings have important implications for the biology and sociology of mate selection; they also can be applied more broadly to inform models of human evolution. Random mating and Hardy–Weinberg equilibrium are fundamental assumptions in human population genetics and are the basis of popular tools to infer haplotypes (36, 37), detect IBD (38, 39), infer demographic patterns (40), search for signals of selection (41), and simulate coalescence (42). Departure from random mating can lead to deviations in the final biological estimates. In this study, we characterized the highly nonrandom mating structure in Latinos and showed that genomic ancestry is a key determinant of the observed assortment. An important direction of future work will be to develop new models that better capture the realities of nonrandom mate selection by incorporating the assortment patterns characterized here into the tools and lingua franca of population genetics.

## Materials and Methods

### Datasets.

**Genetics of Asthma in Latino Americans I.** The GALAI dataset consisted of 231 Mexican trios (father–mother–offspring) from San Francisco and Mexico City and 258 Puerto Rican trios from New York City and San Juan, Puerto Rico. Each member of the trio was genotyped using the Affymetrix 6.0 GeneChip. Genotypes were quality-controlled as previously described (20). This dataset did not contain any socioeconomic information about the individuals or the families.

- Kalmijn M (2014) Assortative mating by cultural and economic occupational status. *Am J Sociol* 100(2):422–452.
- Schwartz CR (2013) Trends and variation in assortative mating: Causes and consequences. *Annu Rev Sociol* 39:451–470.
- Buston PM, Emlen ST (2003) Cognitive processes underlying human mate choice: The relationship between self-perception and mate preference in Western society. *Proc Natl Acad Sci USA* 100(15):8805–8810.
- Vandenberg SG (1972) Assortative mating, or who marries whom? *Behav Genet* 2(2):127–157.
- Risch N, et al. (2009) Ancestry-related assortative mating in Latino populations. *Genome Biol* 10(11):R132.
- Mare RD (1991) Five decades of educational assortative mating. *Am Sociol Rev* 56(1):15–32.
- Jacob S, McClintock MK, Zelano B, Ober C (2002) Paternally inherited HLA alleles are associated with women's choice of male odor. *Nat Genet* 30(2):175–179.
- Chaix R, Cao C, Donnelly P (2008) Is mate choice in humans MHC-dependent? *PLoS Genet* 4(9):e1000184.
- Sebro R, Hoffman TJ, Lange C, Rogus JJ, Risch NJ (2010) Testing for non-random mating: Evidence for ancestry-related assortative mating in the Framingham Heart Study. *Genet Epidemiol* 34(7):674–679.
- Laurent R, Toupan B, Chaix R (2012) Non-random mate choice in humans: Insights from a genome scan. *Mol Ecol* 21(3):587–596.
- Domingue BW, Fletcher J, Conley D, Boardman JD (2014) Genetic and educational assortative mating among US adults. *Proc Natl Acad Sci USA* 111(22):7996–8000.
- Ruiz-Linares A, et al. (2014) Admixture in Latin America: Geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet* 10(9):e1004572.
- Espino R, Franz MM (2002) Latino phenotypic discrimination revisited: The impact of skin color on occupational status. *Soc Sci Q* 83(2):612–623.
- Telles EE, Murguía E (1988) *Phenotypic Discrimination and Income Differences Among Mexican Americans* (University of Arizona, Tucson, AZ), MASRC Working Paper Series 13.
- McCullough JM, Giles E, Thompson RA (1985) Evidence for assortative mating and selection in surnames: A case from Yucatan, Mexico. *Hum Biol* 57(3):375–386.
- Beckett WS, Belanger K, Gent JF, Holford TR, Leaderer BP (1996) Asthma among Puerto Rican Hispanics: A multi-ethnic comparison study of risk factors. *Am J Respir Crit Care Med* 154(4 Pt 1):894–899.
- Galanter JM, et al. (2014) Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: The Genes-Environments & Admixture in Latino Americans study. *J Allergy Clin Immunol* 134(2):295–305.
- Pino-Yanes M, et al. (2015) Genetic ancestry influences asthma susceptibility and lung function among Latinos. *J Allergy Clin Immunol* 135(1):228–235.
- Torgerson DG, et al. (2012) Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. *J Allergy Clin Immunol* 130(1):76–82.e12.
- Burchard EG, et al.; Genetics of Asthma in Latino Americans (GALA) Study (2004) Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. *Am J Respir Crit Care Med* 169(3):386–392.
- Baran Y, et al. (2012) Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28(10):1359–1367.

**Genes-Environments & Admixture in Latino Americans II.** The GALII dataset consisted of 1,246 unrelated Mexican individuals living in the United States and 1,511 unrelated Puerto Rican individuals living in the United States or Puerto Rico. For each participant in the study, all four grandparents self-identified as Puerto Rican or Mexican. For each participant, genotypes were measured on the Affymetrix Axiom LAT1 array (>800,000 SNPs) designed for Latino populations, with quality control as previously described (17, 18). The GALII dataset also included information about the socioeconomic status of the participants and their parents, as described in detail in *SI Appendix*. Institutional review board (IRB) approval for GALAI and GALII was obtained from the University of California, San Francisco, IRB and each local recruitment site's own IRB department. Consent was obtained from participants at the time of visit before any procedures. Consent was given by the participant or by the participant's parent or legal guardian if the participant was under 18 y of age at recruitment.

**Methods.** All of the computational analyses and methods are explained in detail in *SI Appendix*. We used LAMP-LD (21) to infer the local ancestries from genotype, and ANCESTOR (24) to infer the global ancestries of the parents from an offspring's genotype. IBD mapping was done using PIGS (30), and the run-of-homozygosity analysis was performed using PLINK (43). All other statistical analyses, including computations of the GAAM and EAM scores, were performed using custom software written in Python.

**ACKNOWLEDGMENTS.** We thank Yvonne Yamanaka and Mary Gray for feedback on the manuscript. Y.S.S. is supported by NIH Grant R01-GM094402 and a Packard Fellowship for Science and Engineering. S.S. was partially supported by NIH Grant K99 GM111744. E.H. was partially supported by the Israeli Science Foundation (Grant 1425/13) and by National Science Foundation Grant III-1217615. E.G.B. was supported by National Cancer Institute Grant R25CA113710, National Heart, Lung, and Blood Institute Grant R01HL088133, and National Institute of Allergy and Infectious Diseases U19A1077439.

- Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–1664.
- Bryc K, et al. (2010) Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc Natl Acad Sci USA* 107(Suppl 2):8954–8961.
- Zou JY, Halperin E, Burchard E, Sankararaman S (2015) Inferring parental genomic ancestries using pooled semi-Markov processes. *Bioinformatics* 31(12):i190–i196.
- Alvarez L, Jaffe K (2004) Narcissism guides mate selection: Humans mate assortatively, as revealed by facial resemblance, following an algorithm of “self seeking like.” *Evol Psychol* 2(1):177–194.
- Claes P, et al. (2014) Modeling 3D facial shape from DNA. *PLoS Genet* 10(3):e1004224.
- Welter D, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42(Database issue):D1001–D1006.
- Sturm RA (2009) Molecular genetics of human pigmentation diversity. *Hum Mol Genet* 18(R1):R9–R17.
- Sommer S (2005) The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Front Zool* 2(1):16.
- Park DS, et al. (2015) PIGS: Improved estimates of identity-by-descent probabilities by probabilistic IBD graph sampling. *BMC Bioinformatics* 16(Suppl 5):S9.
- Pemberton TJ, et al. (2012) Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* 91(2):275–292.
- Witkop CJ, et al. (1990) Albinism and Hermansky-Pudlak syndrome in Puerto Rico. *Bol Asoc Med P R* 82(8):333–339.
- Helgason A, Palsson S, Gudbjartsson DF, Kristjánsson T, Stefánsson K (2008) An association between the kinship and fertility of human couples. *Science* 319(5864):813–816.
- Tafoya S (2004) Shades of belonging: Latinos and racial identity. *Harvard Journal of Hispanic Policy* 17:55–78.
- Conrad C, Whitehead J, Mason P, Stewart J (2005) *African Americans in the U.S. Economy* (Rowman & Littlefield, Lanham, MD).
- Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81(5):1084–1097.
- Williams AL, Patterson N, Glessner J, Hakonarson H, Reich D (2012) Phasing of many thousands of genotyped samples. *Am J Hum Genet* 91(2):238–251.
- Browning BL, Browning SR (2011) A fast, powerful method for detecting identity by descent. *Am J Hum Genet* 88(2):173–182.
- Gusev A, et al. (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res* 19(2):318–326.
- Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5(10):e1000695.
- Grossman SR, et al. (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327(5967):883–886.
- Wakeley J (2008) *Coalescent Theory: An Introduction* (Roberts & Company, Greenwood Village, CO).
- Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.