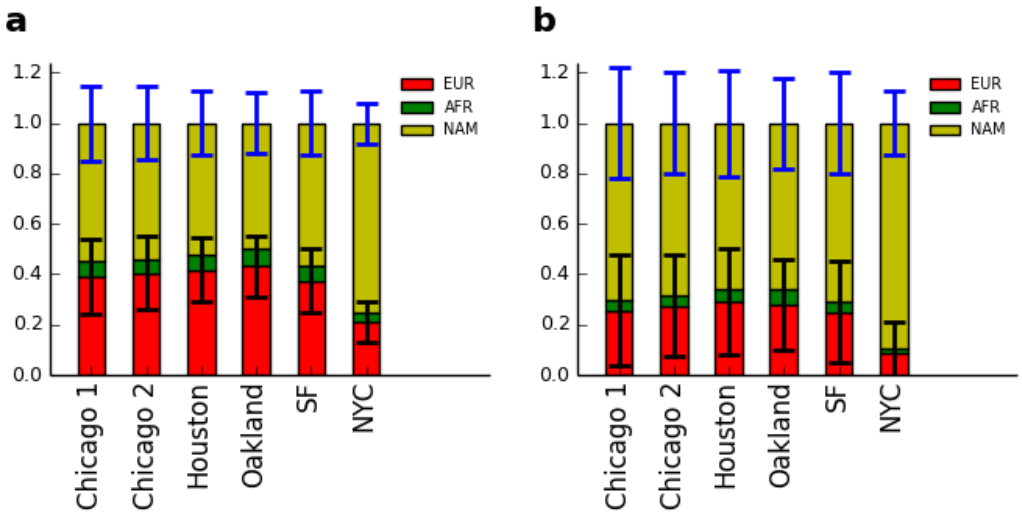
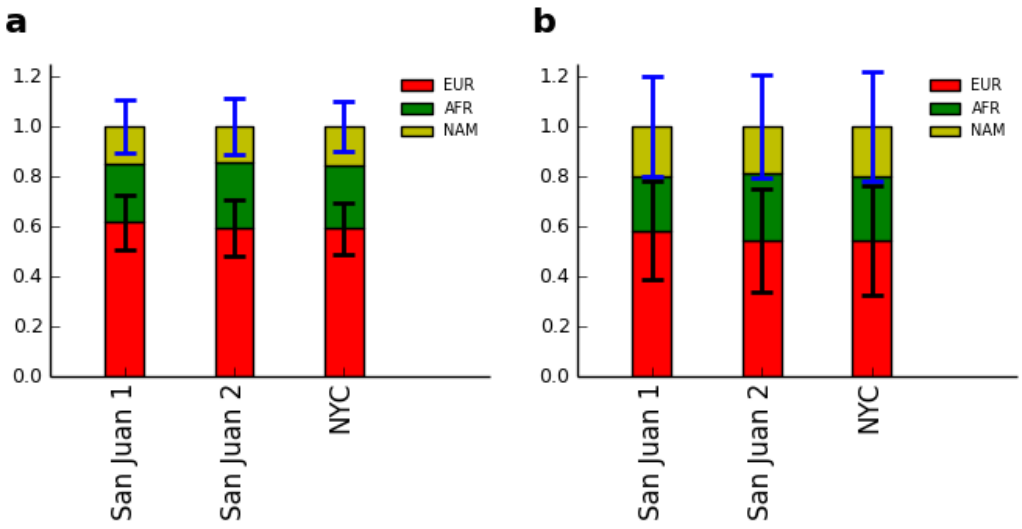
Supplementary Information

Supplementary Figure 1



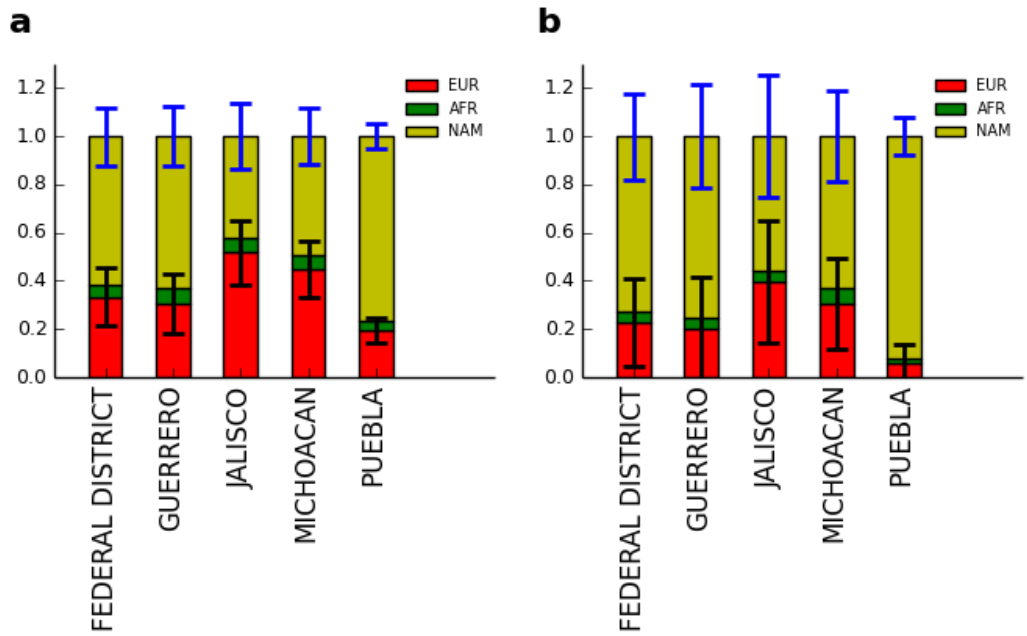
Global ancestries of Mexican individuals. We computed the global genomic ancestry for each individual, which is defined to be the fraction of the genome that comes from European (EUR), African (AFR) or Native American (NAM) ancestries. The GALAII Mexican participants were grouped by their current location—two places in Chicago, Houston, Oakland, San Francisco and New York City. Bar-plots show the average genomic ancestries on the autosomes (a) and the X chromosome (b). The error bars correspond to one standard deviation in the European (black) and Native American ancestries (blue). The error bars for African ancestries were omitted for clarity.

Supplementary Figure 2



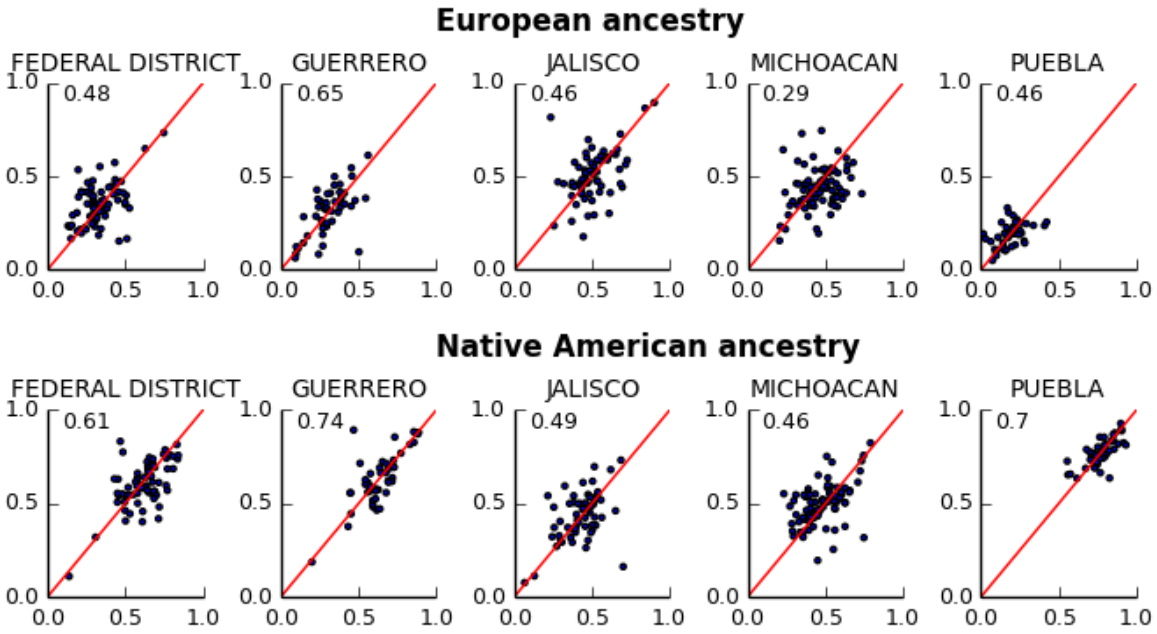
Global ancestries of Puerto Rican individuals. We computed the global ancestries of each Puerto Rican individual. The GALAII Puerto Rican participants were grouped by their current location—two places in San Juan, PR, and New York City. Bar-plots show the average genomic ancestries on the autosomes (a) and the X chromosome (b). The error bars correspond to one standard deviation in the European (black) and Native American ancestries (blue). The error bars for African ancestries were omitted for clarity.

Supplementary Figure 3



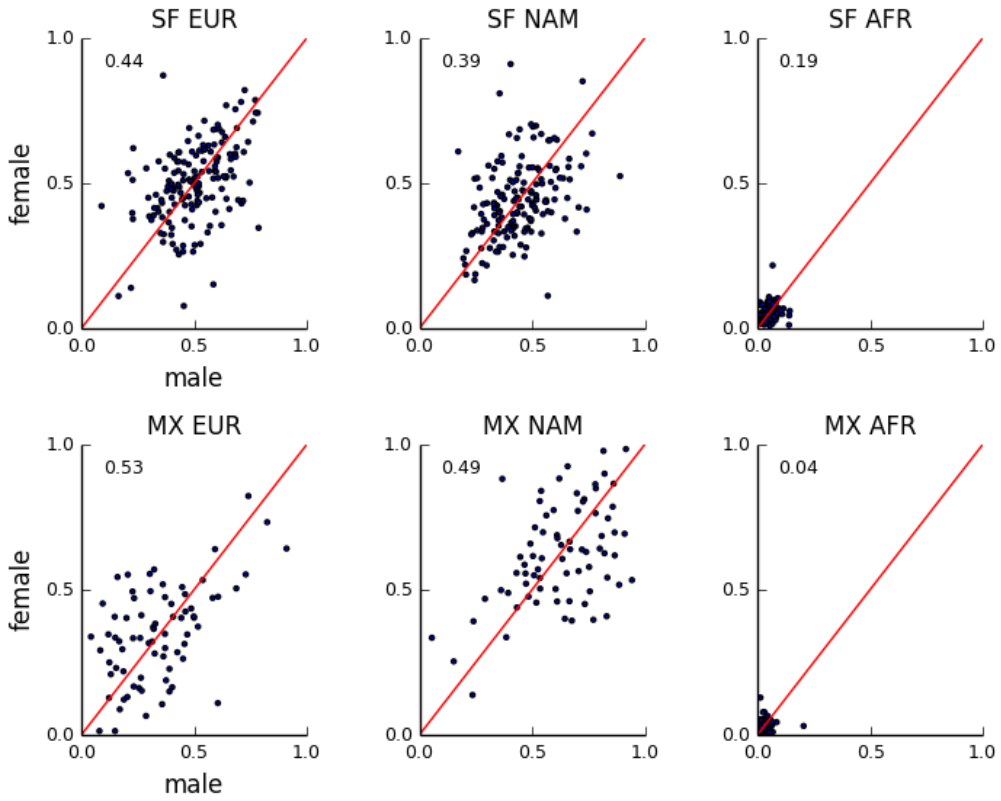
Global ancestries of Mexican individuals grouped by state of birth. The GALAII Mexican participants were grouped by their state of birth. Bar-plots show the average genomic ancestries on the autosomes (a) and the X chromosome (b). The error bars correspond to one standard deviation in the European (black) and Native American ancestries (blue). The error bars for African ancestries were omitted for clarity.

Supplementary Figure 4



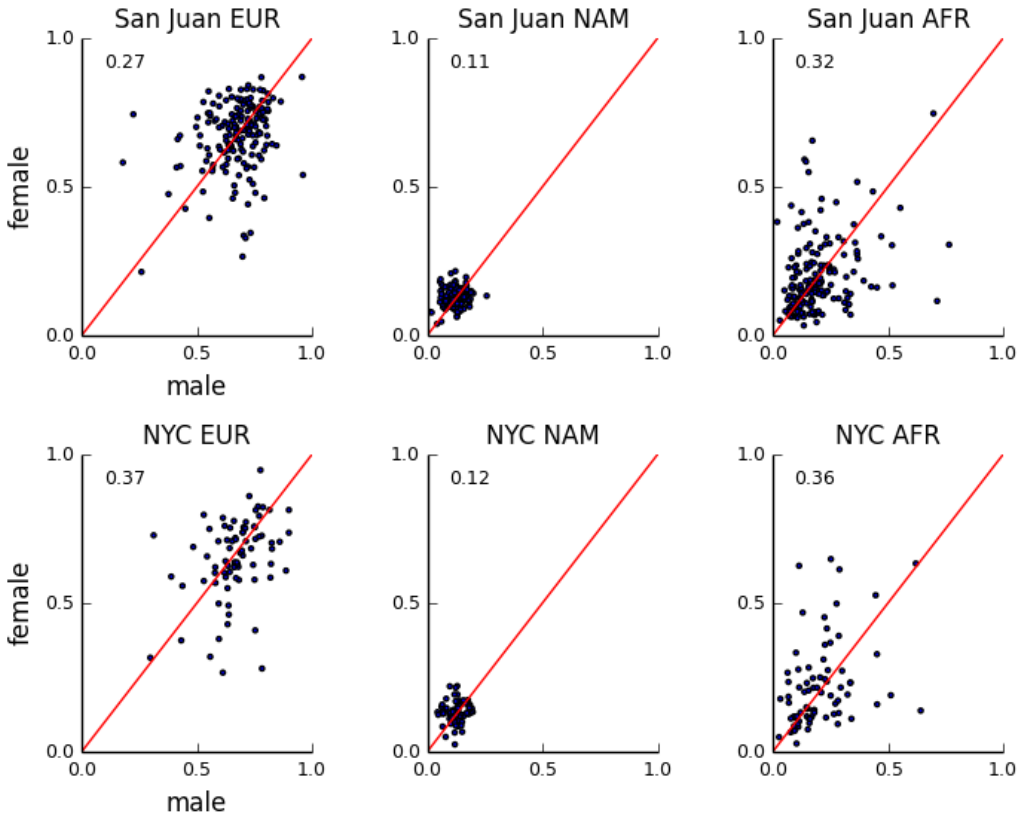
Correlation of genomic ancestries in GALAII Mexican couples grouped by their state of birth. Each dot corresponds to one Mexican couple, and the axes correspond to the ancestries of the two partners. Here we only considered couples in which both partners were born in the same state. The top row shows scatterplots of European ancestry and the bottom row shows scatterplots of Native American ancestry. The Pearson correlation value is shown on each plot.

Supplementary Figure 5



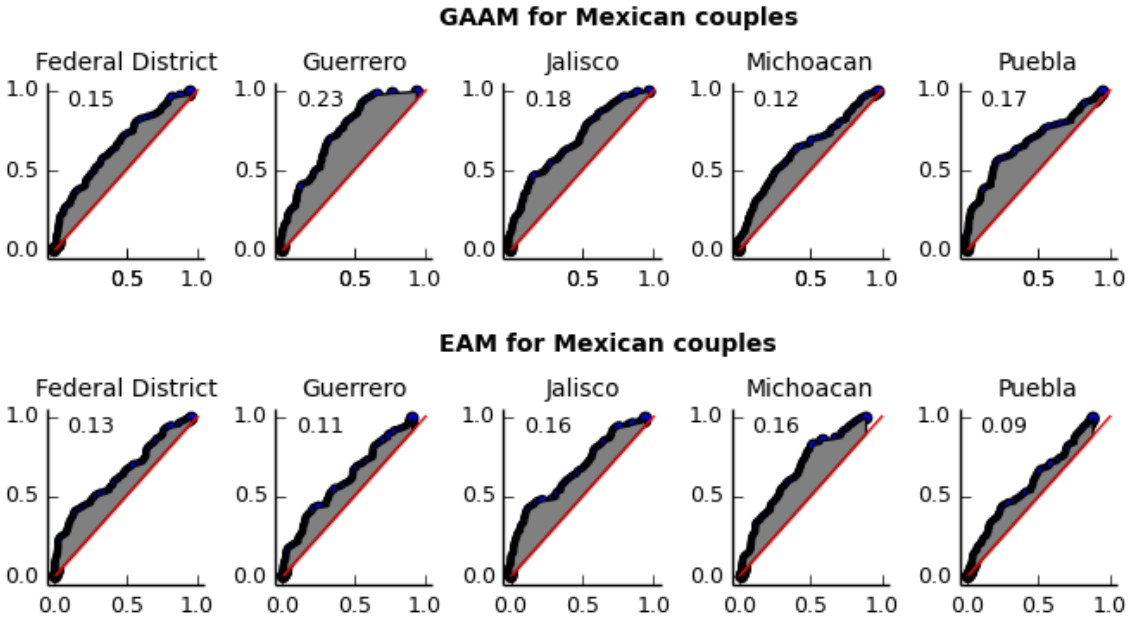
Correlation of genomic ancestries in GALAI Mexican couples. Each dot corresponds to one Mexican couple. The X-axis is the male's ancestry and the Y-axis is the female's ancestry. GALAI Mexican participants from San Francisco (SF) are shown in the top row and participants from Mexico City (MX) are shown in the bottom row. The left, middle and right panels are the scatterplots of the European (EUR), Native American (NAM) and African (AFR) ancestries, respectively. The Pearson correlation value is shown on each plot.

Supplementary Figure 6



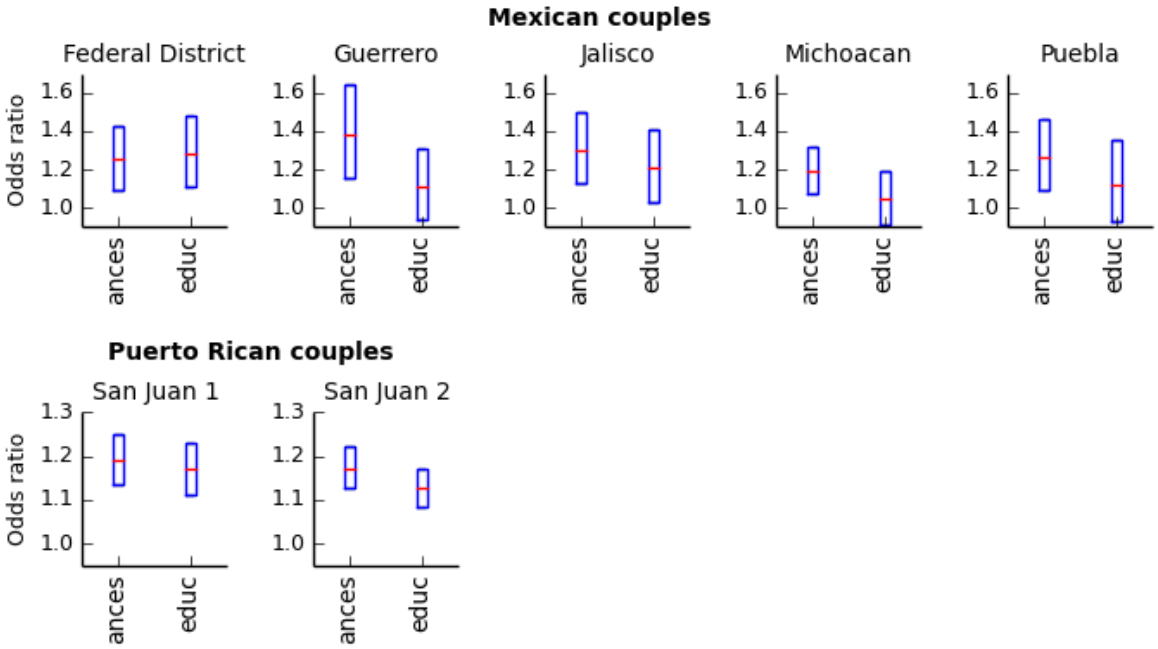
Correlation of genomic ancestries in GALAI Puerto Rican couples. Each dot corresponds to one Puerto Rican couple. The X-axis is the male's ancestry and the Y-axis is the female's ancestry. GALAI Puerto Rican participants from San Juan are shown in the top row and participants from New York City (NYC) are shown in the bottom row. The left, middle and right panels are the scatterplots of the European (EUR), Native American (NAM) and African (AFR) ancestries, respectively. The Pearson correlation value is shown on each plot.

Supplementary Figure 7



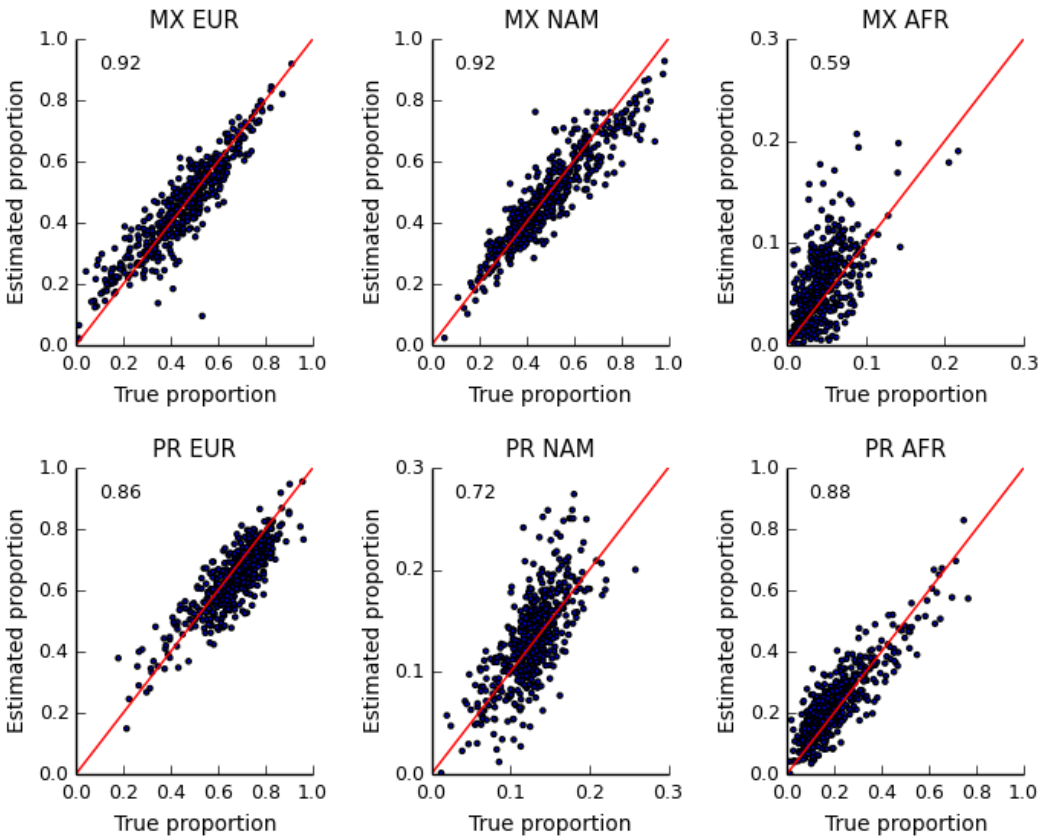
Genomic ancestry assortment (GAAM) scores compared with education assortment scores (EAM) in Mexican couples, stratified by state of birth. The X-axis is the percentile of the actual couples' similarity scores and the Y-axis is the percentile of the random pairs' similarity scores. In each plot, the black curve shows how the distribution of similarity scores in the real couples match with the distribution in random couples. Couples are more similar than expected by random mating when the black curve is above the 45° diagonal (red line). The assortment score, defined as the area between the black curve and the diagonal, is shown for each plot. 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 state of birth.

Supplementary Figure 8



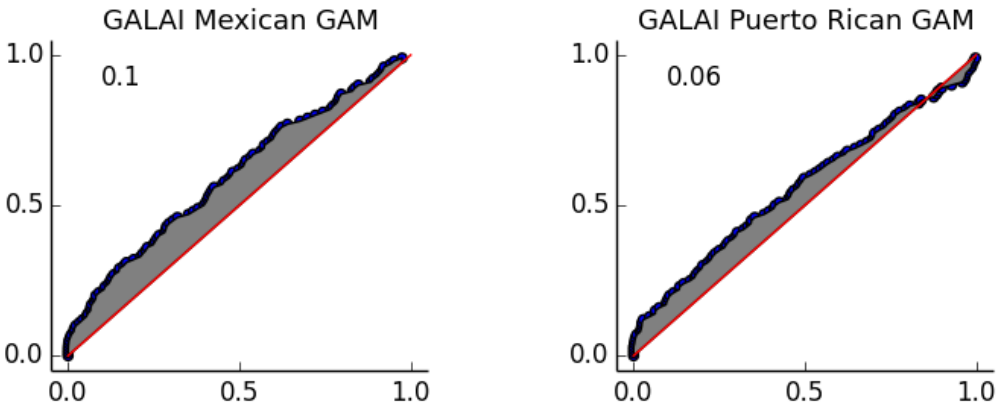
The odds ratios of ancestry similarity and education similarity. Odds ratios from the multiple logistic regression are shown in the box plots. The red bars show the maximum likelihood estimates of the odds ratios and the blue boxes indicate the 95th confidence interval. The couples are stratified by their state of birth. Results for GALAII Mexican couples are shown in the top row, and results for GALAII Puerto Rican couples are shown in the bottom row.

Supplementary Figure 9



Validation of ANCESTOR-estimated genomic ancestries. Each dot is an individual. The X-axis corresponds to the true genomic ancestry and the Y-axis is the genomic ancestry as estimated by ANCESTOR. GALAI Mexican participants are shown in the top row and GALAI Puerto Rican participants are shown in the bottom row. The left, middle and right panels are the scatterplots of the European (EUR), Native American (NAM) and African (AFR) ancestries, respectively. The Pearson correlation value is shown on each plot.

Supplementary Figure 10



Genetic assortment scores (GAM) for GALAI Mexican and Puerto Rican populations. The X-axis is the percentile of the actual couples' genetic kinship scores and the Y-axis is the percentile of the random pairs' genetic kinship scores. In each plot, the black curve shows how the distribution of similarity scores in the real couples match with the distribution in random couples. Couples are more similar than expected by random mating when the black curve is above the 45° diagonal (red line). The assortment score, defined as the area between the black curve and the diagonal, is shown for each plot.

Supplementary Table 1

Dataset	Location	Collection site	Number of participants
Mexican trios (GALAI)	San Francisco, CA	SF General Hospital	157
	Mexico City, Mexico	National Institute of Pulmonary Diseases	74
Mexican unrelated (GALAII)	Chicago, IL 1	Children's Memorial Hospital	291
	Chicago, IL 2	Northwestern University	158
	Houston, TX	Texas Children's Hospital	281
	New York, NY	Jacobi Medical	116
	Oakland, CA	La Clinica	163
	San Francisco, CA	SF General Hospital	142
	Other		95
Puerto Rican trios (GALAI)	Puerto Rico	University of Puerto Rico	170
	New York, NY	Columbia University	77
Puerto Rican unrelated (GALAII)	San Juan, PR 1	Hato Rey	686
	San Juan, PR 2	San Juan	615
	New York, NY	Jacobi Medical	122
	Other		88

Collection sites for the Mexican and Puerto Rican participants. Location indicates the location of the individual at the time of sample collection.

Supplementary Table 2

Ethnicity	Location	Ancestry	r^2	Coefficient	CI lower	CI upper	P-value
Mexican	Chicago 1	EUR	0.04	3.44	1.48	5.39	0.01
		NAM	0.05	-3.34	-5.12	-1.55	0.01
		AFR	0.02	8.93	0.84	17.01	0.03
	Chicago 2	EUR	0.04	-2.72	-4.84	-0.59	0.01
		NAM	0.03	2.41	0.33	4.48	0.02
		AFR	0.01	8.76	-6.51	24.03	0.26
	Houston	EUR	0.05	3.7	1.73	5.66	0.01
		NAM	0.04	-3.29	-5.19	-1.38	0.01
		AFR	0	-6.03	-17.93	5.88	0.32
Puerto Rican	San Juan 1	EUR	0.01	1.37	-0.02	2.76	0.05
		NAM	0.01	-7.76	-12.94	-2.58	0.01
		AFR	0	-0.74	-2.07	0.58	0.27
	San Juan 2	EUR	0.03	2.83	1.53	4.12	0.01
		NAM	0	0.41	-4.16	4.98	0.86
		AFR	0.03	-2.66	-3.91	-1.41	0.01
	NYC	EUR	0.07	5.73	1.59	9.88	0.01
		NAM	0	3.17	-6.1	12.45	0.50
		AFR	0.08	-5.77	-9.68	-1.86	0.01

Correlation of genomic ancestry with education. Results of the linear regression with the couples' average ancestries as the explanatory variable and the couples' average education as the response. The 95th confidence interval (CI) for the coefficient is listed in CI lower and CI upper. Locations where results did not reach statistical significance are not shown.

Supplementary Table 3

Ethnicity	Location	Ancestry	r^2	Coefficient	CI lower	CI upper	P-value
Mexican	Oakland	EUR	0.04	2.89	0.54	5.25	0.02
		NAM	0.03	-2.49	-4.78	-0.2	0.03
		AFR	0.01	-6.99	-19.78	5.8	0.28
	San Francisco	EUR	0.04	2.84	0.12	5.56	0.04
		NAM	0.04	-2.56	-5.01	-0.12	0.04
		AFR	0.01	6.32	-5.76	18.41	0.30
Puerto Rican	San Juan 1	EUR	0.01	1.9	0.42	3.39	0.01
		NAM	0	-1.13	-6.83	4.57	0.70
		AFR	0.01	-1.64	-3.05	-0.23	0.02
	San Juan 2	EUR	0.08	5.41	3.75	7.06	0.01
		NAM	0.01	-5.47	-11.16	0.22	0.06
		AFR	0.06	-4.53	-6.14	-2.92	0.01
	NYC	EUR	0.04	4.62	0.01	9.22	0.05
		NAM	0	-0.4	-7.66	6.86	0.91
		AFR	0.04	-5.15	-10.09	-0.21	0.04

Correlation of genomic ancestry with income. Results of the linear regression with the couples' average genomic ancestry as the explanatory variable and the family income as the response. The 95th confidence interval (CI) for the coefficient is listed in CI lower and CI upper. Location where results did not reach statistical significance are not shown.

Supplementary Table 4

Ethnicity	Location	r^2	Coefficient	CI lower	CI upper	P-value
Mexican	Houston	0.02	0.14	0.02	0.25	0.02
	Oakland	0.06	0.19	0.05	0.33	0.01
Puerto Rican	San Juan 1	0.14	0.38	0.29	0.47	0.01
	San Juan 2	0.24	0.58	0.49	0.67	0.01
	NYC	0.05	0.28	0.01	0.55	0.05

Correlation of education with income. Results of the linear regression with parents' average education as the explanatory variable and the family income as the response. The 95th confidence interval (CI) for the coefficient is listed in CI lower and CI upper. Locations where the results are not statistically significant are not shown.

Socio-economic variables.

For each of the GALAII Mexican and Puerto Rican trios, several socio-economic attributes were collected from each member of the family (father, mother and offspring).

Location. Location at the time of data collection.

Employment status. A binary status encoding whether or not the participant was employed at the time of data collection.

Ethnicity. Self-identified ethnicity of the participant.

Race. Self-identified race of the participant.

Education level. The participant's education level was encoded as an integer from 1 to 14. Level 1 through 12 correspond to the highest grade that the participant has completed in primary or secondary school. Level 12 encodes the final year of high school or a high school graduate. Level 13 corresponds to some college but no BA degree. Level 14 corresponds to college graduate or post-graduate degree.

Country and state of birth. If the participant was not born in the U.S., the number of years he/she has lived in the U.S. was also recorded.

For each family, the total annual income and the number of children were recorded.

Family income. The combined family annual income was encoded as an integer from 1 to 9. The unit of currency is USD.

1= Less than $5000; 2 = $5000 to $11999; 3 = $12000 to $15999;

4 = $16000 to $24999; 5 = $25000 to $34999; 6 = $35000 to $49999;

7 = $50000 to $74999; 8 = $75000 to $99999; 9 = larger than $100000.

Number of offspring. Total number of children in the family at the time of data collection.

Sample collection.

Participants were recruited from major metropolitan regions in the U.S., Mexico and Puerto Rico, and as such the study population was broadly representative of the Latino/Hispanic community. Study participants were recruited from schools and community-based clinics. Trained bilingual interviewers (English-Spanish) administered questionnaires to remove language barriers and to reach more deeply into the community.

Local and global ancestry inference.

For GALAI, the reference panels were created as previously described in Pasaniuc et al (1). For GALAII, we used reference panels composed of European (CEU) and Yoruba (YRI) individuals from HAPMAP (2), and Human Genome Diversity Project (HGDP) Pima and Mayan individuals (3). LAMP-LD (4) was then run on GALAI and GALAII samples using a window size of 50 and 15 hidden Markov model states for each chromosome. The output of LAMP-LD gives the biallelic local ancestry states for each region of the individual's genome. Each SNP allele was assigned to one of three ancestral states: European, Native American or African. To estimate the global genomic ancestry of the individual, we added the local ancestry estimates to obtain the fraction of the genome assigned to European, Native American or African states.

We also ran ADMIXTURE (5) on both GALAI and GALAII using the same reference panels in a supervised mode with the parameter for the number of parental populations, K=3. Before running ADMIXTURE, datasets were first LD pruned using the plink option --indep-pairwise 50 10 0.1 as recommended by the authors. The global genomic ancestry estimates from ADMIXTURE agreed very closely with that of LAMP-LD ($r^2 > 0.99$).

Inference of parental genomic ancestries.

To infer the genomic ancestries of the parents from the GALAII dataset, we first ran LAMP-LD on the offspring genotype to infer his/her local ancestry blocks. The local ancestry blocks were then used as input to ANCESTOR (6), a method we recently developed that uses a pooled semi-Markov Process to model the local ancestry blocks of the transmitted alleles from parents to offspring. ANCESTOR uses Markov Chain Monte Carlo to infer the genomic ancestries of each parent.

To establish that the significant, positive correlation between parental ancestries was not an artifact of our ancestry inference method, we used the actual ancestries of the GALAI trios as a gold-standard validation. We applied ANCESTOR to the offspring genotypes, and comparison of the inferred parental genomic ancestries with the actual genomic ancestries showed excellent agreement [Supp. Fig 9]. In addition, we simulated uniform random mating by creating 100 random male-female pairs from the GALAI samples. We simulated the diploid offspring genotype by generating transmitted alleles from the male and female and merging the alleles. Applying ANCESTOR to the random mating offspring showed no significant correlations in the inferred genomic ancestries of the parents ($P > 0.3$ for the Pearson correlation for each of the European, Native American and African ancestries).

Computation of education assortative mating (EAM), genomic-ancestry assortative mating (GAAM) and genetic assortative mating (GAM) scores. .

EAM, GAAM and GAM scores were computed as previously defined (7). We took the actual couples of each ethnic group in a region and generated 5000 random male-female pairs to

simulate the background distribution under uniform random mating. For each couple (real and simulated), the education similarity was defined as one minus the absolute difference in the number of years of schooling between the partners. The genomic similarity was defined as one minus the Euclidean distance between the partners' ancestry vectors; so if (f_E, f_A, f_N) and (F_E, F_A, F_N) are the global ancestries of two individuals, the genomic ancestry similarity is $1 - (f_E - F_E)^2 - (f_A - F_A)^2 - (f_N - F_N)^2$. The genetic similarity was defined to be the genotype kinship computed using REAP (8). In the EAM, GAAM and GAM plots [Fig 2, Supp. Fig 7, Supp. Fig 10], the X-axis is the percentile of the actual couples' similarity scores and the Y-axis is the percentile that the scores would correspond to in the randomly paired couples. For example, the value of the black curve at X = 0.5 is computed by taking the median similarity score among the actual couples and finding the fraction of simulated couples with a lower score. The shaded area between the black curve and the diagonal measures the strength of assortment. The maximal score for EAM, GAAM and GAM is 0.5. In principle, if an individual is more likely to partner with someone that's more different from him/herself, the score can be negative up to -0.5. A score of -0.5 indicates that each individual is partnered with the person in the population that is the most different from him/herself. In each of regions, we found significant positive EAM, GAAM and GAM scores ($P < 0.01$, permutation test), indicating individuals are more likely to partner with someone with more similar education, genomic ancestry and genotype. .

Logistic regression model of genomic ancestry and education assortment.

We used multiple logistic regression to jointly analyze the assortment of genomic ancestry and education in couples. The individuals were stratified by their location of birth and a separate analysis was performed on each region. In each region, we generated the same number of

random male-female pairs as the number of actual couples. The real couples were encoded as 1 and the simulated pairs were encoded as 0, and this was the response variable. The education and genomic ancestry similarities between spouses were computed in the same way as described above for EAM and GAAM. These scores were then normalized as the fraction of random pairs that had a similarity score lower than the observed score. This percentile-normalized education and genomic ancestry similarities were the covariates in the logistic regression. For both Mexicans and Puerto Ricans, in every location, we found ancestry and education similarities to be significant factors [$P < 0.01$, Wald test]. In all the regions, the effect size of ancestry was comparable to or stronger than that of education [Supp. Fig 8].

Ancestry-adjusted genotype correlation.

In admixed individuals, such as Latinos, we need to account for the different local ancestry that individuals may have at a SNP. Due to the possibility of dramatically different allele frequencies between ancestral populations, a correlation or anti-correlation may be observed between couples' genotypes due to ancestry and not assortative mating.

Consider a group of n couples and a SNP s. Let f_i denote the female member of couple i and m_i denote the male member of couple i. Let g_{sfi} and g_{smi} denote the counts of the minor allele in the sample for the female and male, respectively. To control for ancestry, we normalize the genotype of each individual, j, according to his/her local ancestry, a_{sj} at SNP s. This approach is similar to the global ancestry adjustment method developed by Thornton et.al (8) for estimating kinship in admixed individuals. We use LAMP-LD to infer the local ancestry state, a_{sj}. Let $p(a_{sj})$ be the ancestral population's allele frequency. Then the ancestry-normalized genotype is given by

$b_{sj} = (g_{sj} - 2p(a_{sj})) / \text{sqrt}(2\ p(a_{sj})\ (1-\ p(a_{sj}))$.

We define b_{sf} and b_{sm} to be vectors of ancestry-adjusted genotypes of the females and males, where each entry of the vector corresponds to one couple. The 1 degree of freedom chi-squared test statistic of SNP *s*, H_s, is defined as $n^*\text{corr}(b_{sf}, b_{sm})^2$.

Controls for facial development gene set analysis.

To ensure that the enriched correlation in facial development genes is not due to unusual genotype diversity, we computed the B-statistic to quantify genotype diversity (9). The B-statistic distribution in the facial development gene set was not different from the B-statistic distribution of random sets of genes used in our permutation control (P = 0.7 Kolmogorov-Smirnov test). As an additional precaution, we also checked that the distribution of Fst statistics for the facial development genes was not different from the distribution in random sets of genes (P = 0.5 Kolmogorov-Smirnov test).

In addition, we simulated synthetic Puerto Rican male-female pairs using GALAI genotypes, with genetic relatedness matching that of the true Puerto Rican couples. On these synthetic pairs, we performed the exact same genotype correlation analysis as before; i.e., we computed the ancestry-adjusted genotype correlations across the synthetic pairs and analyzed whether the facial development genes showed higher correlation compared to random gene sets. This analysis controls for potential differences in genotyping density and quality at the facial genes. We found no difference between the correlation signal at the facial genes and randomly selected genes (P = 0.82) in these synthetic couples. This finding suggests that the enriched correlation

we observed in the actual couples is a true signal and not an artifact of unusual genotyping at these genes.

Pigmentation genes.

Our pigmentation gene set consists of the following eight genes: MC1R, KITLG, ASIP, SLC24A5, SLC45A2, TYR, TYRP1, OCA2 (10).

Deviation from Hardy-Weinberg.

We used the exact test of Hardy-Weinberg equilibrium (11) to test for deviations from Hardy-Weinberg SNP-by-SNP. We also tested for deviations from HW in the local ancestry states by running the same test on the LAMP-LD-inferred local ancestry states at each SNP.

Runs of homozygosity.

We inferred ROH in each individual using PLINK (12). We used PLINK to prune SNPs in LD with each other ($r^2 > 0.8$). We then ran PLINK autozygosity analysis with the requirement that each ROH have at least 50 SNPs and be longer than 100 kb. As in Pemberton et al. (13), we use mclust to cluster the ROH segments of each population into three groups, corresponding to short, medium and long ROHs. We ran mclust on the Puerto Rican and Mexican GALAII datasets separately. For Mexicans, short ROH were less than 370.716 kb, medium ROH were between 370.716 kb and 1187.125 kb and long ROH were longer than 1187.125 kb. For Puerto Ricans, short ROH were less than 357.116 kb, medium ROH were 357.116 kb to 1169.05 kb and long ROH were longer than 1169.05 kb. This is consistent with the cluster boundaries measured in Pemberton et al. (13). Short ROH reflect ancient haplotype sharing, and intermediate ROH likely

result from population-level background sharing (13). Long ROH result from more recent parental relatedness (13), though not necessarily consanguinity.

IBD sharing.

At a genetic locus, two haplotypes are identical-by-descent (IBD) if the haplotypes have identical sequences and were inherited from a common ancestor. For each set of individuals in the GALAI and GALAII datasets, we used BEAGLE 4 to phase the genotypes and call pairwise IBD (14). All pairwise IBD calls were thresholded at a length of 0.1 centimorgans and at an LOD score of 0.1. We then ran a multiway method, PIGS (15), on each unique connected component generated by the pairwise IBD calls from BEAGLE. All IBD calls made by PIGS were thresholded at a probability of 0.99 and a length of 0.5 centimorgans. Any overlapping segments on the same haplotype were merged.

Ancestry-matched permutations.

As a control set for the IBD analysis, we generated random male-female pairings for each ethnic group and within each collection site. Random pairings were generated to reflect the observed global ancestry correlations seen in real couples by fixing the mothers and permuting the fathers within 5% bins of European.

Analysis of a model of assortative mating and disease prevalence.

We analyzed a simple mathematical model in order to quantitatively illustrate the effect of ancestry-based assortative mating on disease prevalence. Suppose there are two populations, A and B, that admix. Let α denote the fraction of individuals that comes from A; $1-\alpha$ is the

fraction of individuals that comes from B. For an individual, let θ denote the fraction of his/her genome with ancestry state A. The population size is assumed to be very large so we can effectively use the infinite population model. During each generation, pairs of individuals are randomly selected such that the correlation between their ancestries is a constant ρ. The offspring of two individuals with ancestries θ^F and θ^M has ancestry $\theta = (\theta^F + \theta^M)/2$. Note that under this model, the population's average A ancestry, $E[\theta] = \alpha$, is constant across generations.

We assume that there is one recessive risk allele with frequency f in A and frequency 0 in B. Under random mating, after one generation Hardy-Weinberg equilibrium is reached and the homozygosity at this allele is $\alpha^2 f^2$. The prevalence of this recessive disease in the population is $c\alpha^2 f^2$ where c is the penetrance of the allele. Next we compute the homozygosity of the risk allele after t generations of assortative mating to be:

$$\begin{aligned}\int P(\theta^F, \theta^M)\theta^F\theta^M f^2 d\theta^F\theta^M &= E[\theta^F\theta^M]f^2 \\ &= f^2(Cov[\theta^F, \theta^M] + \alpha^2) \\ &= f^2(\alpha^2 + V_t\rho)\end{aligned}$$

where V_t is the variance of θ after t generations. We can compute V_t recursively,

$$\begin{aligned}V_{t+1} &= E[\theta_{t+1}^2] - \alpha^2 \\ &= E[(\theta_t^M + \theta_t^F)(\theta_t^M + \theta_t^F)/4] - \alpha^2 \\ &= \frac{1}{4}(2E[(\theta_t^2] + 2E[\theta_t^F\theta_t^M]) - \alpha^2 \\ &= \frac{1}{2}(V_t + \alpha^2 + V_t\rho + \alpha^2) - \alpha^2 \\ &= \frac{V_t(1+\rho)}{2} = V_0\frac{(1+\rho)^{t+1}}{2^{t+1}} = \alpha(1-\alpha)\frac{(1+\rho)^{t+1}}{2^{t+1}}.\end{aligned}$$

Plugging this recursion into the expression for homozygosity, we have that the homozygosity of the risk allele after t generations of assortative mating is $f^2\alpha^2 + f^2\rho\alpha(1-\alpha)\frac{(1+\rho)^t}{2^t}$. Compared to random mating, the homozygosity of the risk allele under assortative mating is inflated by $\alpha(1-\alpha)f^2\rho\frac{(1+\rho)^t}{2^t}$, and hence the prevalence of the disease is also increased by $\alpha(1-\alpha)f^2\rho\frac{(1+\rho)^t}{2^t}$ multiplied by the penetrance of the allele. As the correlation in the ancestries of the partners increases, ρ increases and the disease prevalence increases as well.

As an illustration, suppose the correlation in the genomic ancestries of the partners is $\rho = 0.6$, which is in the range that we quantified for Mexicans and Puerto Ricans. Moreover assume that the two founder populations had the same number of individuals initially, i.e. $\alpha = 0.5$. After $t = 10$ generations of admixture, the ratio of the risk allele homozygosity under assortative mating to the homozygosity under random mating is 1.06. This corresponds to a 6% increase in the prevalence of this recessive disease.

References

1. Pasaniuc B, et al. (2013) Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics* 29(11):1407–15. Available at: http://bioinformatics.oxfordjournals.org/content/29/11/1407.abstract?sid=bc64ae15-8461-4c6e-aeb8-f06f6d58637f [Accessed December 23, 2014].

2. Frazer KA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–61. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2689609&tool=pmcentrez&rendertype=abstract [Accessed July 11, 2014].

3. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319(5866):1100–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18292342 [Accessed July 11, 2014].

4. Baran Y, et al. (2012) Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28(10):1359–67. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3348558&tool=pmcentrez&rendertype=abstract [Accessed December 3, 2014].

5. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–64. Available at: http://genome.cshlp.org/content/early/2009/07/31/gr.094052.109.abstract [Accessed July 9, 2014].

6. Zou J, Halperin E, Burchard EG, Sankararaman S (2015) Inferring parental genomic ancestries using pooled semi-Markov processes. *Bioinformatics*:1–10.

7. Domingue BW, Fletcher J, Conley D, Boardman JD (2014) Genetic and educational assortative mating among US adults. *Proc Natl Acad Sci U S A* 111(22):7996–8000. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24843128.

8. Thornton T, et al. (2012) Estimating kinship in admixed populations. *Am J Hum Genet* 91(1):122–38. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3397261&tool=pmcentrez&rendertype=abstract [Accessed November 19, 2014].

9. McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* 5(5):e1000471. Available at: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000471 [Accessed March 30, 2015].

10. Welter D, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42(Database issue):D1001–6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965119&tool=pmcentrez&rendertype=abstract [Accessed July 16, 2014].

11. Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 76(5):887–93. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1199378&tool=pmcentrez&rendertype=abstract [Accessed November 18, 2014].

12. Purcell S, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–75. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1950838&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

13. Pemberton TJ, et al. (2012) Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* 91(2):275–292. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3415543&tool=pmcentrez&rendertype=abstract [Accessed November 17, 2014].

14. 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):1084–97. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2265661&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

15. Park DS, et al. PIGS : Improved estimates of identity-by-descent probabilities by Probabilistic IBD Graph Sampling. *BMC*:1–20.