

*This copy is for your personal, non-commercial use only.*

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

***The following resources related to this article are available online at  
www.sciencemag.org (this information is current as of September 2, 2014 ):***

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/content/344/6189/1280.full.html>

Supporting Online Material can be found at:

<http://www.sciencemag.org/content/suppl/2014/06/11/344.6189.1280.DC1.html>

<http://www.sciencemag.org/content/suppl/2014/06/11/344.6189.1280.DC2.html>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/content/344/6189/1280.full.html#related>

This article **cites 60 articles**, 13 of which can be accessed free:

<http://www.sciencemag.org/content/344/6189/1280.full.html#ref-list-1>

This article appears in the following **subject collections**:

Genetics

<http://www.sciencemag.org/cgi/collection/genetics>

## HUMAN GENETICS

# The genetics of Mexico recapitulates Native American substructure and affects biomedical traits

Andrés Moreno-Estrada,<sup>1,\*†</sup> Christopher R. Gignoux,<sup>2,†‡</sup> Juan Carlos Fernández-López,<sup>3,†</sup> Fouad Zakharia,<sup>1</sup> Martín Sikora,<sup>1</sup> Alejandra V. Contreras,<sup>3</sup> Víctor Acuña-Alonso,<sup>4,5</sup> Karla Sandoval,<sup>1</sup> Celeste Eng,<sup>6</sup> Sandra Romero-Hidalgo,<sup>3</sup> Patricia Ortiz-Tello,<sup>1</sup> Victoria Robles,<sup>1</sup> Eimear E. Kenny,<sup>1,§</sup> Ismael Nuño-Arana,<sup>7</sup> Rodrigo Barquera-Lozano,<sup>4</sup> Gastón Macín-Pérez,<sup>4</sup> Julio Granados-Arriola,<sup>8</sup> Scott Huntsman,<sup>6</sup> Joshua M. Galanter,<sup>6,9</sup> Marc Via,<sup>6||</sup> Jean G. Ford,<sup>10</sup> Rocío Chapela,<sup>11</sup> William Rodriguez-Cintrón,<sup>12</sup> Jose R. Rodríguez-Santana,<sup>1,3</sup> Isabelle Romieu,<sup>14</sup> Juan José Sienra-Monge,<sup>15</sup> Blanca del Rio Navarro,<sup>15</sup> Stephanie J. London,<sup>16</sup> Andrés Ruiz-Linares,<sup>5</sup> Rodrigo García-Herrera,<sup>3</sup> Karol Estrada,<sup>3,¶</sup> Alfredo Hidalgo-Miranda,<sup>3</sup> Gerardo Jiménez-Sánchez,<sup>3,‡</sup> Alessandra Carnevale,<sup>3</sup> Xavier Soberón,<sup>3</sup> Samuel Canizales-Quinteros,<sup>3,17</sup> Héctor Rangel-Villalobos,<sup>7</sup> Irma Silva-Zolezzi,<sup>3,\*\*</sup> Esteban Gonzalez Burchard,<sup>6,9\*</sup> Carlos D. Bustamante<sup>1,\*</sup>

populations have experienced high degrees of isolation. Indeed, when autozygosity is inferred using runs of homozygosity (ROH), all populations on average have long homozygous tracts, with the Huichol, Lacandon, and Seri all having on average over 10% of the genome in ROH [figs. S2 and S3 (9)]. These populations are relatively small, increasing the effects of genetic drift and driving some of the high  $F_{ST}$  values. In contrast, the Mayan and Nahuan populations have much smaller proportions of the genome in ROH, consistent with ROH levels found in Near Eastern populations in HGDP (12). These populations are the descendants of large Mesoamerican civilizations, and concordant with large historical populations, have relatively low proportions of ROH. The high degree of variance in ROH among populations is an additional indicator of population substructure and suggests a large variance in historical population sizes. Comparing the observed ROH patterns to those derived from coalescent simulations, we find that Native American groups within Mexico are characterized by small effective population sizes under a model with a strong bottleneck, in agreement with other studies of Native American populations (13). The degree of population size recovery to the current day is consistent with the degree of isolation of the extant populations, ranging from 1196 chromosomes [95% confidence interval (CI) 317 to 1548] for the Seri in the Sonora desert, to 3669 (95% CI 2588 to 5522) for the Mayans from Quintana Roo (figs. S4 to S6; (9)).

**U**nderstanding patterns of human population structure, where regional surveys are key for delineating geographically restricted variation, is important for the design and interpretation of medical genetic studies. In particular, we expect rare genetic variants, including functionally relevant sites, to exhibit little sharing among diverged populations (1). Native Americans display the lowest genetic diversity of any continental group, but there is high divergence among subpopulations (2). As a result, present-day American indigenous populations (and individuals with indigenous ancestry) may harbor local private alleles rare or absent elsewhere, including functional and medically relevant variants (3, 4). Mexico serves as an important focal point for such analyses, because it harbors one of the largest sources of pre-Columbian diversity and has a long history of complex civilizations with varying contributions to the present-day population.

Previous estimates of Native Mexican genetic diversity examined single loci or were limited to a reduced number of populations or small sample sizes (5–8). We examined local patterns of variation from nearly 1 million genome-wide autosomal single-nucleotide polymorphisms (SNPs) for 511 Native Mexican individuals from 20 indigenous groups, covering most geographic

regions across Mexico (table S1). Standard principal components analysis (PCA) summarizes the major axes of genetic variation in the data set [see (9)]. Whereas PC1 and PC2 separate Africans and Europeans from Native Mexicans, PC3 differentiates indigenous populations within Mexico, following a clear northwest-southeast cline (Fig. 1A). A total of 0.89% of the variation is explained by PC3, nearly three times as much as the variation accounted for by the north-south axis of differentiation within Europe [0.30%, in (10)]. The northernmost (Seri) and southernmost (Lacandon) populations define the extremes of the distribution, with very clear clustering of individuals by population, indicating high levels of divergence among groups (fig. S1). Seri and Lacandon show the highest level of population differentiation as measured with Wright's fixation index  $F_{ST}$  (0.136, Fig. 1B and table S4), higher than the  $F_{ST}$  between Europeans and Chinese populations in HapMap3 (0.11) (11). Other populations within Mexico also show extreme  $F_{ST}$  values; for example, the Huichol and Tojolabal have a pairwise  $F_{ST}$  of 0.068, similar to that observed between the Gujarati Indians and the Chinese in HapMap3 (0.076).

The high degree of differentiation between populations measured by  $F_{ST}$  argues that these

<sup>1</sup>Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. <sup>2</sup>Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA. <sup>3</sup>Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, Mexico. <sup>4</sup>Escuela Nacional de Antropología e Historia (ENAH), Mexico City, Mexico. <sup>5</sup>Department of Genetics, Evolution and Environment, University College London, London, UK.

<sup>6</sup>Department of Medicine, University of California, San Francisco, CA, USA. <sup>7</sup>Instituto de Investigación en Genética Molecular, Universidad de Guadalajara, Ocotlán, Mexico. <sup>8</sup>Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico. <sup>9</sup>Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA. <sup>10</sup>The Brooklyn Hospital Center, Brooklyn, NY, USA. <sup>11</sup>Instituto Nacional de Enfermedades Respiratorias (INER), Mexico City, Mexico. <sup>12</sup>Veterans Caribbean Health Care System, San Juan, Puerto Rico.

<sup>13</sup>Centro de Neurología Pediátrica, San Juan, Puerto Rico. <sup>14</sup>International Agency for Research on Cancer, Lyon, France.

<sup>15</sup>Hospital Infantil de México Federico Gomez, Mexico City, Mexico. <sup>16</sup>National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC, USA. <sup>17</sup>Facultad de Química, Universidad Nacional Autónoma de México, Mexico City, Mexico.

\*Corresponding author. E-mail: cd bustam@stanford.edu (C.D.B.); morenoe@stanford.edu (A.M.-E.); esteban.burchard@ucsf.edu (E.G.B.). <sup>†</sup>These authors contributed equally to this work. <sup>‡</sup>Present address: Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. (C.R.G.) <sup>§</sup>Present address: Center for Statistical Genetics, Mount Sinai School of Medicine, New York, USA. (E.E.K.) <sup>||</sup>Present address: Department of Psychiatry and Clinical Psychobiology - IR3C, Universitat de Barcelona, Spain. (M.V.). <sup>¶</sup>Present address: Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, USA. (K.E.) <sup>#</sup>Present address: Harvard School of Public Health and Global Biotech Consulting Group. (G.J.-S.) <sup>\*\*</sup>Present address: Nutrition and Health Department Nestec Ltd, Nestle Research Center, Lausanne, Switzerland. (I.S.-Z.)

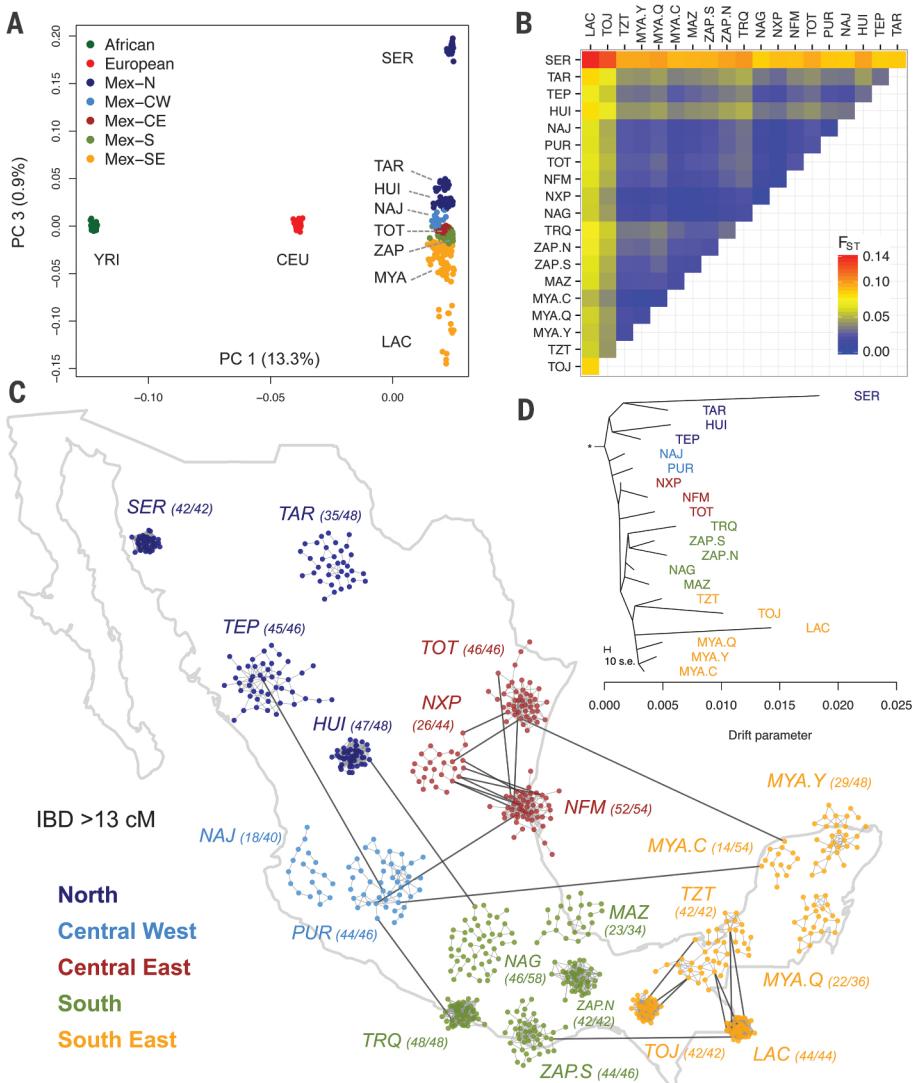
Isolation also correlates with the degree of relatedness within and between ethnic groups, ultimately shaping the pattern of genetic relationships among populations. We built a relatedness graph (Fig. 1C) of individuals sharing >13 cM of the genome identically by descent (IBD) (corresponding to third/fourth cousins or closer relatives). Almost all the connections are within-versus among-population, consistent with the populations being discrete rather than exhibiting large-scale gene flow [figs. S7 and S8 (9)].

As seen with the ROH calculations, the Mayan and Nahuan groups have fewer internal connections. The few between-population connections appear in populations close to each coastline, such as the connections between the Campeche Mayans and populations to the west along the Gulf of Mexico.

The long-tract ROH and IBD analyses are especially relevant to the recent history of isolation of Native American populations. We ran TreeMix (14) to generate a probabilistic model of diver-

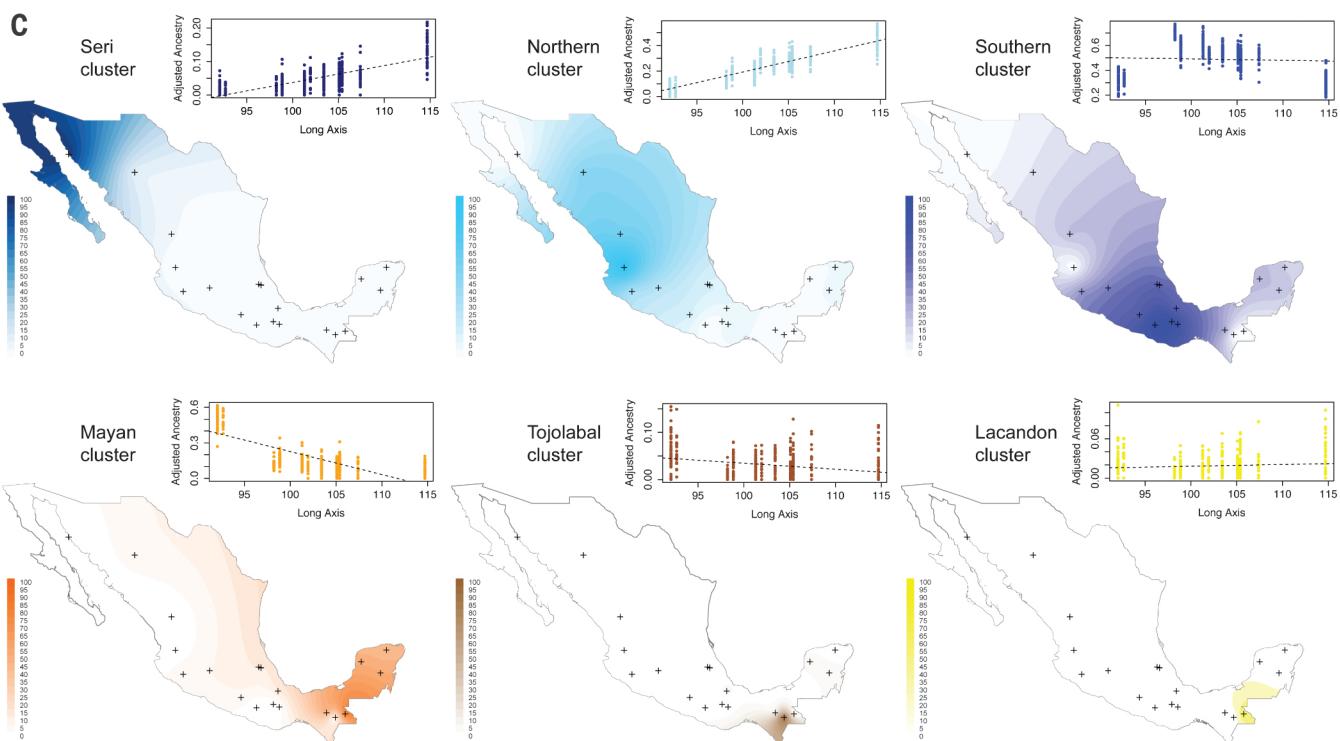
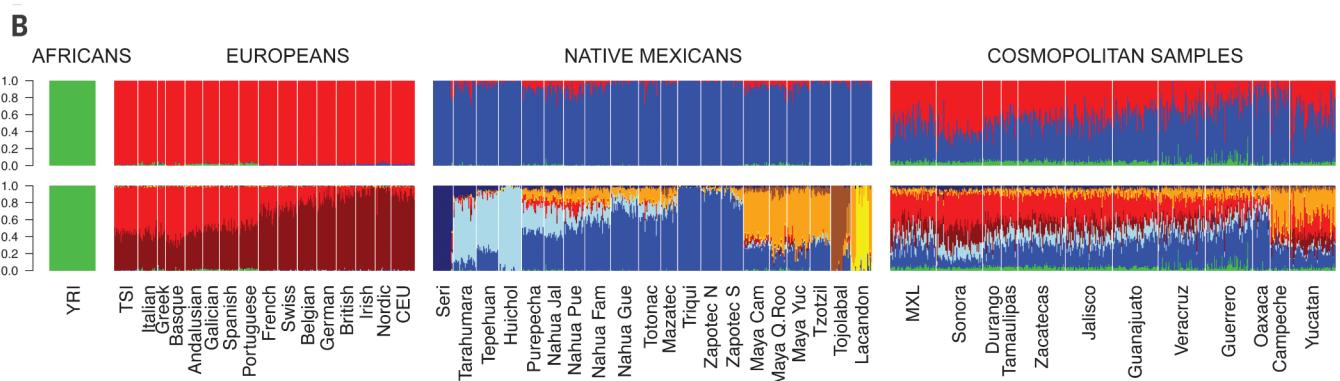
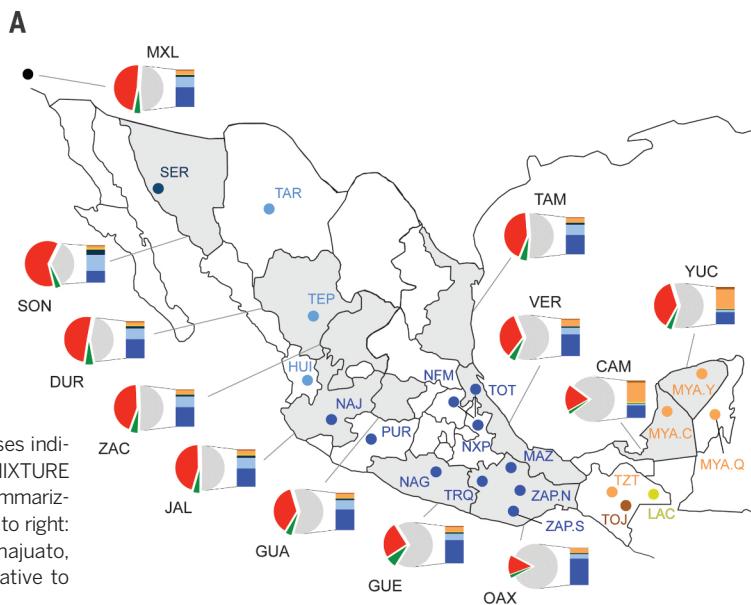
gence and migration among the Native American populations (Fig. 1D). The inferred tree with no migration paths recapitulates the north/south and east/west gradients of differentiation from the PCA and IBD analyses, with populations with high ROH values also exhibiting longer tip branches. The primary branches divide populations by geography. All northern populations (dark blue) branch from the same initial split at the root. We also find two additional major clades: a grouping of populations from the southern states of Guerrero and Oaxaca (green labels) and a “Mayan clade” composed of Mayan-speaking populations from Chiapas and the Yucatan peninsula in the southeast (orange labels). Introducing migratory edges to the model connects the Maya in Yucatan to a branch leading to the Totonac, whose ancestors occupied the large pre-Columbian city of El Tajin in Veracruz (15). This result points to an Atlantic coastal corridor of gene flow between the Yucatan peninsula and central/northern Mexico (fig. S9), consistent with our IBD analysis. Indeed, the only Mayan language outside the Mayan territory is spoken by the Huastec, nearby in northern Veracruz, supporting a shared history (16).

These signals remain today as a legacy of the pre-Columbian diversity of Mexican populations. Over the past 500 years, population dynamics have changed drastically. Today, the majority of Mexicans are admixed and can trace their ancestry back not only to indigenous groups but also to Europe and Africa. To investigate patterns of admixture, we combined data from continental source populations (including the 20 native Mexican groups, 16 European populations, and 50 West African Yorubas) with 500 admixed mestizo individuals from 10 Mexican states recruited by the National Institute of Genomic Medicine (INMEGEN) for this study, Mexicans from Guadalajara in the POPRES collection (17), and individuals of Mexican descent from Los Angeles in the HapMap Phase 3 project (table S1). We ran the unsupervised mixture model algorithm ADMIXTURE (18) to estimate ancestry proportions for individuals in our combined data set (Fig. 2, fig. S10, and table S5). Allowing for three ancestral clusters ( $K = 3$ ), we find that most individuals have a large amount of Native and European ancestry, with a small (typically <5%) amount of African ancestry. At the best-fit model for  $K = 9$ , the Native American cluster breaks down into six separate components (Fig. 2B). Three of these are mostly restricted to isolated populations (Seri, navy blue; Lacandon, yellow; and Tojolabal, brown). The other three show a wider but geographically well-defined distribution: A northern component (light blue) represented by Tarahumara, Tepehuano, and Huichol, gradually decreases southward. Correspondingly, a southern component (blue), which includes Triqui, Zapotec, and Mazatec, gradually decreases northward. In the Yucatan peninsula and the neighboring state of Chiapas, we found what we termed the “Mayan component” (orange in Fig. 2B, bottom panel), found primarily in Mayan-speaking groups. This Mayan component is also



**Fig. 1. Genetic differentiation of Native Mexican populations.** (A) PCA of Native Mexicans with HapMap YRI and CEU samples. Population labels as in table S1. (B) Pairwise  $F_{ST}$  values among Native Mexican populations ordered geographically (see also table S4). (C) Relatedness graph of individuals sharing more than 13 cM of the genome as measured by the total of segments IBD. Each node represents a haploid genome, and edges within clusters attract nodes proportionally to shared IBD. The spread of each cluster is thus indicative of the level of relatedness in each population, as determined by a force-directed algorithm. Only the layout of nodes within each cluster is the result of the algorithm, as populations are localized to their approximate sampling locations to ease interpretation. Parentheses indicate the number of individuals represented out of the total sample size ( $2N$ ). The full range of IBD thresholds are shown in fig. S8. (D) TreeMix graph representing population splitting patterns of the 20 Native Mexican groups studied. The length of the branch is proportional to the drift of each population. African, European, and Asian samples were used as outgroups to root the tree (fig. S9).

**Fig. 2. Mexican population structure.** (A) Map of sampled populations (detailed in table S1) and admixture average proportions (table S5). Dots correspond to Native Mexican populations color-coded according to  $K = 9$  clusters identified in (B) (bottom), and shaded areas denote states in which cosmopolitan populations were sampled. Pie charts summarize per-state average proportions of cosmopolitan samples at  $K = 3$  (European in red, West African in green, and Native American in gray). Bars show the total Native American ancestry decomposed into average proportions of the native subcomponents identified at  $K = 9$ . (B) Global ancestry proportions at  $K = 3$  (top) and  $K = 9$  (bottom) estimated with ADMIXTURE, including African, European, Native Mexican, and cosmopolitan Mexican samples (tables S1 and S2). From left to right, Mexican populations are displayed north-to-south. (C) Interpolation maps showing the spatial distribution of the six native components identified at  $K = 9$ . Contour intensities are proportional to ADMIXTURE values observed in Native Mexican samples, with crosses indicating sampling locations. Scatter plots with linear fits show ADMIXTURE values observed in cosmopolitan samples versus a distance metric summarizing latitude and longitude (long axis) for the sampled states. From left to right: Yucatan, Campeche, Oaxaca, Veracruz, Guerrero, Tamaulipas, Guanajuato, Zacatecas, Jalisco, Durango, and Sonora. Values are adjusted relative to the total Native American ancestry of each individual (9).



present at ~10 to 20% in central Mexican natives, consistent with the IBD and migration edges connecting the regions. This relationship between the Yucatan peninsula and central Mexico, seen in both recent shared IBD and genetic drift-based models of allele frequencies (TreeMix and ADMIXTURE), suggests that gene flow between the two regions has been ongoing for a long time. In contrast, Mayan admixture is not found at appreciable levels in highlanders of the southern state of Oaxaca (Triqui and Zapotec), where mountain ranges may have acted as geographic barriers to gene flow.

Patterns of Native American population substructure are recapitulated in the genomes of Mexican mestizos from cosmopolitan populations throughout Mexico. Sonora and neighboring northern states show the highest average proportions of the northern native component (15%, light blue in Fig. 2B, bottom), whereas only traces are detected in Oaxaca and the Yucatan peninsula. Conversely, the southern native component is the most prevalent across states, reaching maximum values in Oaxaca and decreasing northward. Cosmopolitan samples from the Yucatan peninsula have Native American fractions of the genome dominated by the Mayan component, which diminishes in northward populations. Likewise, Mayan-related local components, Tojolabal and Lacandon, are detected above 1% exclusively among individuals from the states neighboring the Yucatan peninsula. In contrast, Mexican-Americans sampled in Los Angeles (MXL) do not show a homogeneous pattern, consistent with their diverse origins within Mexico. Overall, the continuous geographic distribution of each Native American component across Mexico (fig. S12) demonstrates a high correlation of individual admixture proportions with geography, even in individuals of mixed ancestry (Fig. 2C, NW-SE axis *F*-test for all native clusters,  $P < 10^{-16}$ ).

To further test whether ancestral population structure is recapitulated in the genomes of mestizos, we used an ancestry-specific PCA (ASPCA) approach [fig. S13 (9, 19)]. We estimated local ancestry using PCAdmix (20) to identify segments of the genome belonging to Native American, European, or African ancestries. We focused on the European and Native American components of ancestry, given the low proportions of African ancestry overall. We would expect the history of Spanish occupation and colonization in Mexico to be reflected in the European segments of Mexican mestizos, as has been seen previously (21). ASPCA of the European haplotypes in present-day Mexicans confirms this, as individuals cluster tightly with present-day Iberians even with a dense set of European populations (17, 22) (fig. S14).

In contrast, given the complex demographic history of Native Americans, high isolation, and limited characterization of regional ancestry patterns (23, 24), it remains unknown whether the correlation between genes and geography observed in Europe (10) can be similarly recapitulated within Mexico. We used ASPCA to uncover

hidden population structure within Native American ancestry beyond that found solely in extant indigenous groups (Fig. 3A). Consistent with the previous PCA analyses, we observed the most diverged indigenous populations defining the extremes of the top PCs due to high levels of genetic drift and isolation. However, including all the indigenous groups in the plot masks the signal contained in the indigenous segments of the Mexican mestizos. When plotting the ASPCA values for the admixed individuals only, we discovered a strong correlation between Native American ancestry and geography within Mexico (Fig. 3B), with ASPC1 representing a west-to-east dimension and ASPC2 one from north to south. Both of these correlations are highly significant and linearly predictive of geographic location (Pearson's  $r^2$  of 72% and 38% for ASPC1 and 2, respectively, both  $P$  values  $< 10^{-5}$ ). The correlation is strong enough that the overall distribution of mestizo-derived indigenous haplotypes in ASPCA space resembles a geographic map of Mexico (Fig. 3B and fig. S15). This finding suggests that the genetic composition of present-day Mexicans recapitulates ancient Native American substructure, despite the potential homogenizing effect of postcolonial admixture. Fine-scale population structure going back centuries is not merely a property of isolated or rural indigenous communities. Cosmopolitan populations still reflect the underlying genetic ancestry of local native populations, arguing for a strong relationship between the indigenous and the Mexican mestizo population, albeit without the extreme drift exhibited in some current indigenous groups.

Having found these hidden patterns of ancestry in the native component of Mexican mestizos, we investigated whether this structure could have potential biomedical applications. Over the past decade, genetic ancestry has been associated with numerous clinical endpoints and disease risks in admixed populations, including neutrophil counts (25), creatinine levels (26), and breast cancer susceptibility (27). Similarly, ancestral background is especially important in pulmonary medicine, where different reference equations are used for different ethnicities, defining normative predicted volumes and identifying thresholds for disease diagnosis in standard clinical practice (28). That is, depending on one's ethnic background, the same value of forced expiratory volume in 1 s ( $FEV_1$ , a standard measure of lung function) could be either within the normal range or indicative of pulmonary disease. Previous work has shown that the proportion of African and European ancestry was associated with  $FEV_1$  in African Americans (29) and Mexicans (30), respectively, establishing the importance of genomic ancestry in lung function prediction equations.

To investigate possible associations between ancestral structure in Mexicans and  $FEV_1$ , we applied our ASPCA approach to two studies measuring lung function in Mexican or Mexican-American children: the Mexico City Childhood Asthma Study (MCCAS) (31) and the Genetics of Asthma in Latino Americans (GALA I) Study

(32). Due to differences in protocols and genotyping platforms, we calculated ASPCA values for the two studies independently (fig. S17) using the same reference populations described above, then used fixed-effects meta-analysis to combine the results (9).

First, in GALA I we looked for significant ancestry-specific differences between Mexico City and the San Francisco Bay Area, the two recruitment sites. ASPCA values were associated with recruitment location, with the receiver-operator characteristic curve from the Native American ancestry dimensions resulting in an area under the curve (AUC) of 80% (fig. S17). After we adjusted for overall ancestry proportions (here both African and Native American), both ASPCs were significant in a logistic regression: ASPC1 OR per SD: 0.44 (95% CI 0.22 to 0.68),  $P = 3.8 \times 10^{-4}$ , ASPC2 OR per SD: 1.68 (95% CI 1.03 to 2.76),  $P = 0.039$ . The ASPCs defined similar east-west and north-south axes as in the previous analysis (fig. S17) and show that Mexican-Americans in the San Francisco Bay Area tend to have increased Native American ancestry from northwest Mexico as compared to individuals from Mexico City (joint logistic regression likelihood ratio test  $P = 6.4 \times 10^{-5}$ ).

We then used the ASPCA values for both studies to test for an association with  $FEV_1$  as transformed to percentile of predicted "normal" function via the standard set of reference equations (28) for individuals of Mexican descent. These equations use population-specific demographic characteristics to account for age, sex, and height in estimates of lung function. Adjusting for overall ancestry proportions in linear regressions, we observed a significant association between  $FEV_1$  and the east-west component (ASPC1) in both studies, with a meta-analysis  $P$  value of 0.0045 (2.2% decrease in  $FEV_1$  per 1 SD, 95% CI 0.69 to 3.74). The effect sizes were homogeneous (Fig. 3C and table S6) despite differences in recruitment strategy, geography, and genotyping platform (9). In contrast, ASPC2 showed no association with  $FEV_1$ . Whereas  $FEV_1$  has previously been associated with overall ancestry in several populations, the effect seen here is not correlated with overall admixture proportions, because we adjusted for those in the regression model. The combined results here indicate that subcontinental ancestry as measured by ASPCA is important for characterizing clinical measurements.

To estimate how variation in genetic ancestry within Mexico may affect  $FEV_1$ , we used the results from GALA I and MCCAS to predict trait values by state (Fig. 3D) for the INMEGEN mestizo samples. We found that difference in subcontinental Native American ancestry as measured by ASPC1 results in an expected 7.3% change in  $FEV_1$ , moving from the state of Sonora in the west to the state of Yucatan in the east. These results suggest that fine-scale patterns of native ancestry alone could have significant impacts on clinical measurements of lung function in admixed individuals within Mexico.

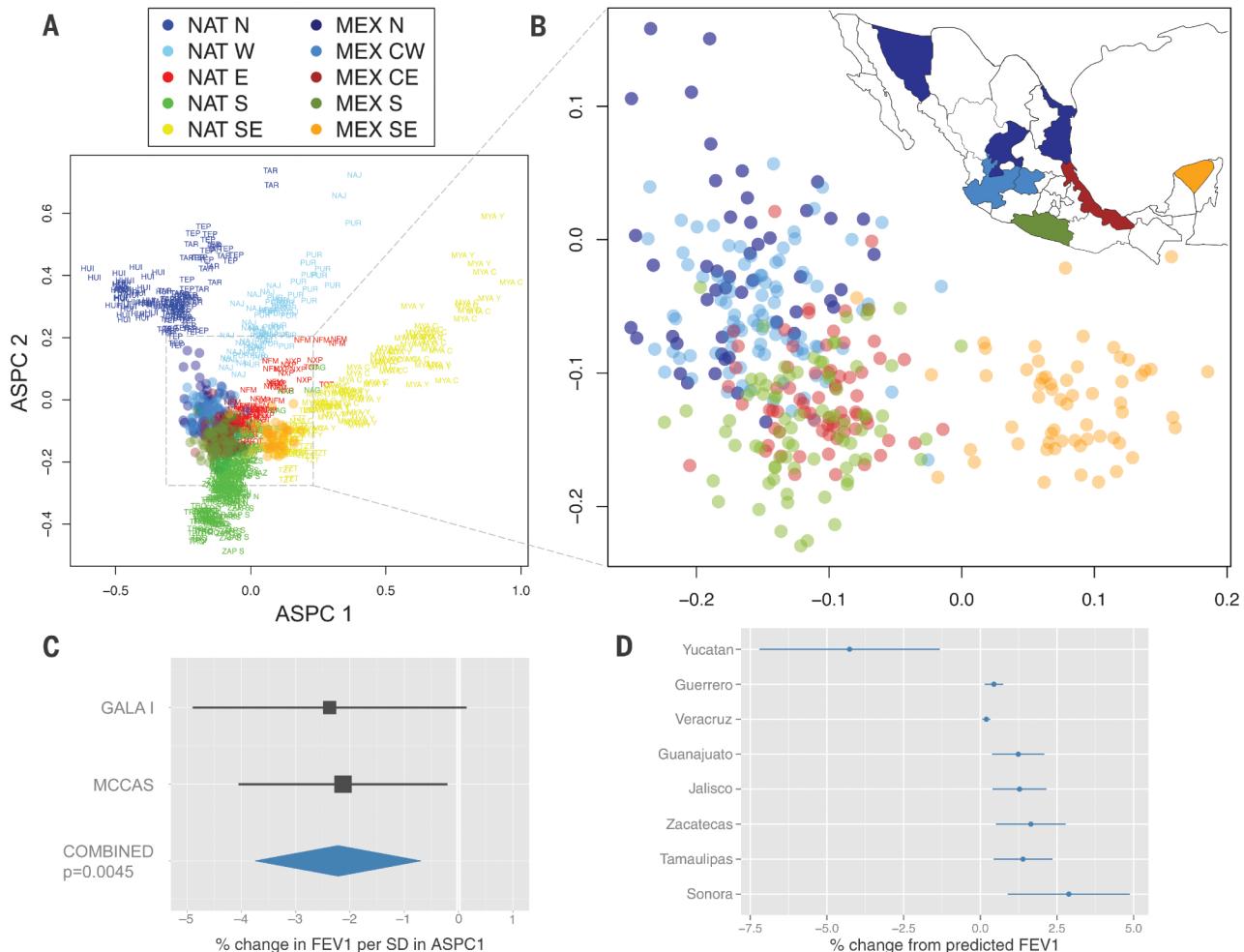
These changes due to ancestry are comparable to other factors affecting lung function. Comparing the expected effect of ancestry across Mexico with the known effects of age in the standard Mexican-American reference equations (28), the inferred 7.3% change in FEV<sub>1</sub> associated with subcontinental ancestry is similar to the decline in FEV<sub>1</sub> that a 30-year-old Mexican-American individual of average height would experience by aging 10.3 years if male and 11.8 years if female. Similarly, comparing our results from the Mexican data with the model incorporating ancestry in African Americans, a difference of 7.3% in FEV<sub>1</sub> would correspond to a 33% difference in African ancestry (29). The association between FEV<sub>1</sub> and ASPC1 is not an indicator of impaired lung function on its own—rather, it contributes to the distribution of FEV<sub>1</sub> values and would modify clinical

thresholds. This finding indicates that diagnoses of diseases such as asthma and chronic obstructive pulmonary disease (COPD) relying on specific lung function thresholds may benefit from taking finer-scale ancestry into consideration.

An important implication of our work is that multi- and transthetic mapping efforts will benefit from including individuals of Mexican ancestry, because the Mexican population harbors rich amounts of genetic variation that may underlie important biomedical phenotypes. A key question in this regard is whether existing catalogs of human genome variation capture the genetic variation present in the samples analyzed here. We performed targeted SNP tagging and genome-wide haplotype sharing analysis within 100-kb sliding windows to assess the degree to which haplotype diversity in the Mexican mestizo samples could be captured

by existing reference panels [figs. S18 to 20 (9)]. Although Mexican-American samples (MXL) were included in both the HapMap and 1000 Genomes catalogs, average haplotype sharing for the INMEGEN mestizo samples is limited to 81.2 and 90.5% when combined with all continental HapMap populations. It is only after including the Native American samples genotyped here that nearly 100% of haplotypes are shared, maximizing the chances of capturing most of the variation in Mexico.

Much effort has been invested in detecting common genetic variants associated with complex disease and replicating associations across populations. However, functional and medically relevant variation may be rare or population-specific, requiring studies of diverse human populations to identify new risk factors (4). Without detailed knowledge of the geographic



**Fig. 3. Subcontinental ancestry of admixed Mexican genomes and biomedical implications.** (A) ASPCA of Native American segments from Mexican cosmopolitan samples (colored circles) together with 20 indigenous Mexican populations (population labels). Samples with >10% of non-native admixture were excluded from the reference panel, as well as population outliers such as Seri, Lacandon, and Tojolabal. (B) Zoomed detail of the distribution of the Native American fraction of cosmopolitan samples throughout Mexico. Native ancestral populations were used to define PCA space (prefixed by NAT) but removed from the background to highlight the

subcontinental origin of admixed genomes (prefixed by MEX). Each circle represents the combined set of haplotypes called Native American along the haploid genome of each sample with >25% of Native American ancestry. The inset map shows the geographic origin of cosmopolitan samples per state, color-coded by region (9). (C) Coefficients and 95% CIs for associations between ASPC1 and lung function (FEV<sub>1</sub>) from Mexican participants in the GALA I study, and the MCCAS, as well as both studies combined (table S6 and fig. S17) (9). (D) Means and CIs of predicted change in FEV<sub>1</sub> by state, extrapolated from the model in (C).

stratification of genetic variation, negative results and lack of replication are likely to dominate the outcome of genetic studies in uncharacterized populations. Here we have demonstrated a high degree of fine-scale genomic structure across Mexico, shaped by pre-Columbian population dynamics and affecting the present-day genomes of Mexican mestizos, which is of both anthropological and biomedical relevance. Studies such as this one are crucial for enabling precision medicine, providing novel data resources, empowering the next generation of genetic studies, and demonstrating the importance of understanding and measuring fine-scale population structure and its associations with biomedical traits.

## REFERENCES AND NOTES

- S. Gravel et al., *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11983–11988 (2011).
- S. Wang et al., *PLOS Genet.* **3**, e185 (2007).
- V. Acuña-Alonso et al., *Hum. Mol. Genet.* **19**, 2877–2885 (2010).
- A. L. Williams et al., *Nature* **506**, 97–101 (2014).
- R. Lisker, E. Ramírez, V. Babinsky, *Hum. Biol.* **68**, 395–404 (1996).
- K. Sandoval et al., *Am. J. Phys. Anthropol.* **148**, 395–405 (2012).
- A. Gorostiza et al., *PLOS ONE* **7**, e44666 (2012).
- D. Reich et al., *Nature* **488**, 370–374 (2012).
- See supplementary materials on Science Online.
- J. Novembre et al., *Nature* **456**, 98–101 (2008).
- D. M. Altshuler et al., *Nature* **467**, 52–58 (2010).
- B. M. Henn et al., *PLOS ONE* **7**, e34267 (2012).
- J. Hey, *PLOS Biol.* **3**, e193 (2005).
- J. K. Pickrell, J. K. Pritchard, *PLOS Genet.* **8**, e1002967 (2012).
- A. Pascual Soto, *El Tajín. En Busca de los Orígenes de una Civilización* (UNAM-INAH, Mexico, 2006).
- L. Campbell, T. Kaufman, *Annu. Rev. Anthropol.* **14**, 187–198 (1985).
- M. R. Nelson et al., *Am. J. Hum. Genet.* **83**, 347–358 (2008).
- D. H. Alexander, J. Novembre, K. Lange, *Genome Res.* **19**, 1655–1664 (2009).
- A. Moreno-Estrada et al., *PLOS Genet.* **9**, e1003925 (2013).
- A. Brisbin et al., *Hum. Biol.* **84**, 343–364 (2012).
- N. A. Johnson et al., *PLOS Genet.* **7**, e1002410 (2011).
- L. R. Botigué et al., *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11791–11796 (2013).
- S. Wang et al., *PLOS Genet.* **4**, e1000037 (2008).
- I. Silva-Zolezzi et al., *Proc. Natl. Acad. Sci. U.S.A.* **21**, 8611–8616 (2009).
- M. A. Nalls et al., *Am. J. Hum. Genet.* **82**, 81–87 (2008).
- C. A. Peralta et al., *Am. J. Nephrol.* **31**, 202–208 (2010).
- L. Fejerman et al., *Cancer Res.* **68**, 9723–9728 (2008).
- J. L. Hankinson, J. R. Odencrantz, K. B. Fedan, *Am. J. Respir. Crit. Care Med.* **159**, 179–187 (1999).
- R. Kumar et al., *N. Engl. J. Med.* **363**, 321–330 (2010).
- K. Salari et al., *Genet. Epidemiol.* **29**, 76–86 (2005).
- D. B. Hancock et al., *PLOS Genet.* **5**, e1000623 (2009).
- D. G. Torgerson et al., *J. Allergy Clin. Immunol.* **130**, 76, e12 (2012).

## ACKNOWLEDGMENTS

We thank all volunteers for generously donating DNA samples and participating in the study. This project was possible with the joint support from multiple institutions in Mexico and the United States. Stanford University supported C.D.B. with funding from the Department of Genetics. INMEGEN received support from the Federal Government of Mexico, particularly the Ministry of Health, the Mexican Health Foundation (FONSAUD), and the Gonzalo Rio Arrechea Foundation. State governments and universities of Durango, Campeche, Guanajuato, Guerrero, Oaxaca, Sonora, Tamaulipas, Veracruz, Yucatan, and Zacatecas contributed significantly to this work. This research was also supported by the George Rosenkranz Prize for Health Care Research in Developing Countries awarded to A.M.-E.; University of California San Francisco (UCSF) Chancellor's Research Fellowship, Dissertation Year Fellowship, and NIH Training Grants T32GM007175 and T32HG000044 (to C.R.G.); the Robert Wood Johnson Foundation Amos Medical Faculty Development Award; the Sandler

Foundation; the American Asthma Foundation (to E.G.B.); CONACYT grant 129693 (to H.R.-V.); BBSRC grant BB/I021213/1 (to A.R.-L.); the National Institutes of Health (NIH) (grants R01GM090087, R01HG003229, ES015794, GM007546, GM061390, HL004464, HL078885, HL088133, HL11636, RR000083, P60MD006902, and ZIA ES49019); and National Science Foundation award DMS-1201234. This work was supported in part by the Intramural Research Program of NIH, National Institute of Environmental Health Sciences (to S.J.L.). Some computations were performed using the UCSF Biostatistics High Performance Computing System. We also thank B. Henn, S. Gravel, and J. Byrnes for helpful discussions; C. Gunter and M. Carpenter for editing the manuscript; and M. Morales for informatics and programming support. C.D.B. is on the advisory board of a project at 23andMe; and on the scientific advisory boards of Personalis, Inc.; InVitae; Etalon, Inc.; and Ancestry.com. The collections and methods for the Population Reference Sample (POPRES) are described by Nelson et al. (2008). The POPRES data sets used for the analyses described here were obtained from

dbGaP through accession number phs000145.v1.p1. Access to the MCCAS data set may be obtained under the terms of a data transfer agreement with the National Institute of Environmental Health Sciences; the contact is S.J.L.. Individual-level genotypes for new data presented in this study are available, through a data access agreement to respect the privacy of the participants for the transfer of genetic data, by contacting C.D.B., A.M.-E., and INMEGEN (<http://www.inmegen.gob.mx/>).

## SUPPLEMENTARY MATERIALS

[www.sciencemag.org/content/344/6189/1280/suppl/DC1](http://www.sciencemag.org/content/344/6189/1280/suppl/DC1)  
Materials and Methods  
Supplementary Text  
Figs. S1 to S20  
Tables S1 to S6  
References (33–64)  
3 February 2014; accepted 21 May 2014  
10.1126/science.1251688

## TRANSCRIPTION

# Interactions between RNA polymerase and the “core recognition element” counteract pausing

Irina O. Vvedenskaya,<sup>1\*</sup> Hanif Vahedian-Movahed,<sup>2\*</sup> Jeremy G. Bird,<sup>1,2\*</sup> Jared G. Knoblauch,<sup>1</sup> Seth R. Goldman,<sup>1</sup> Yu Zhang,<sup>2</sup> Richard H. Ebright,<sup>2†</sup> Bryce E. Nickels<sup>1†</sup>

Transcription elongation is interrupted by sequences that inhibit nucleotide addition and cause RNA polymerase (RNAP) to pause. Here, by use of native elongating transcript sequencing (NET-seq) and a variant of NET-seq that enables analysis of mutant RNAP derivatives in merodiploid cells (mNET-seq), we analyze transcriptional pausing genome-wide *in vivo* in *Escherichia coli*. We identify a consensus pause-inducing sequence element, G<sub>-10</sub>Y<sub>-1</sub>G<sub>+1</sub> (where -1 corresponds to the position of the RNA 3' end). We demonstrate that sequence-specific interactions between RNAP core enzyme and a core recognition element (CRE) that stabilize transcription initiation complexes also occur in transcription elongation complexes and facilitate pause read-through by stabilizing RNAP in a posttranslocated register. Our findings identify key sequence determinants of transcriptional pausing and establish that RNAP-CRE interactions modulate pausing.

**R**egulation of gene expression during transcription elongation often involves sequences in DNA that cause the transcription elongation complex (TEC) to pause. Pausing can affect gene expression by facilitating engagement of regulatory factors, influencing formation of RNA secondary structures, and enabling synchronization of transcription and translation.

Several lines of evidence suggest that pausing involves specific sequence signals that inhibit nucleotide addition (1–11). To define key sequence determinants for pausing, we used native elongating transcript sequencing (NET-seq), which permits occupancies of TECs to be mapped genome-wide with base-pair resolution (12, 13) (fig. S1). The

occupancy of the TEC at a given position is correlated with the tendency of the TEC to pause at the position. Accordingly, NET-seq analysis enables identification of pause sites. To perform NET-seq in *Escherichia coli*, cells carrying a chromosomal *rpoC-3xFLAG* gene, encoding RNAP β' subunit with a C-terminal 3xFLAG tag were grown to midexponential phase; cells were flash-frozen and lysed; 3xFLAG-tagged TECs were immunoprecipitated with an antibody against FLAG; RNAs were extracted from TECs; and RNA 3' ends were converted to cDNAs and analyzed using high-throughput sequencing. We defined pause sites as positions where TEC occupancy exceeded TEC occupancy at each position 25 base pairs (bp) upstream and downstream. We identified 15,553 pause sites, which corresponds to ~19,800 total pause sites, given the estimated ~78% saturation of the analysis (tables S1 to S7). Alignment of pause-site sequences revealed a clear consensus pause element (PE): G<sub>-10</sub>Y<sub>-1</sub>G<sub>+1</sub>, where position -1 corresponds to the position of the RNA 3' end (Fig. 1A and fig. S2). Of the

<sup>1</sup>Department of Genetics and Waksman Institute, Rutgers University, Piscataway, NJ 08854, USA. <sup>2</sup>Department of Chemistry and Waksman Institute, Rutgers University, Piscataway, NJ 08854, USA.

\*These authors contributed equally. †Corresponding author. E-mail: bnickels@waksman.rutgers.edu (B.E.N.); ebright@waksman.rutgers.edu (R.H.E.)