Title
Implications of admixture in the Americas for asthma and ancestry

Author
Gignoux, Christopher

Publication Date
2013

Peer reviewed|Thesis/dissertation

Implications of admixture in the Americas for asthma and ancestry

by

Christopher Raymond Gignoux

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

[redacted]

ACKNOWLEDGEMENTS

Human genetics at UCSF has proven to be an extremely welcoming, nurturing and open-minded community. It has been a true honor to be surrounded by so many exciting researchers doing such interdisciplinary and revolutionary research in these fast-paced times. My time at UCSF has been devoted to a number of different projects allowing me to interface broadly with health care professionals, anthropologists, computer scientists, statisticians, and other researchers both old and young. I truly respect that people treated me as a peer even as an incoming graduate student, and I feel like I have collected knowledge worthy of multiple doctorates during my time at UCSF. I cannot imagine anywhere else where I could have developed expertise in statistical genetics and genetic epidemiology along with furthering my interests in evolutionary and population genetics. This would not have happened without the PSPG program taking a chance on an evolutionary geneticist, so first and foremost I'd like to thank Frank Szoka, Debbie Acoba, and Deanna Kroetz for running the PSPG program and creating a diverse and open research environment starting day one at UCSF.

Of course none of this would be possible without Esteban. He took a chance on me as a population geneticist and I on him as an asthma researcher. I have learned so much by placing myself in a lab with a diverse group of scientists, trying to think of projects that take advantage of the "sweet spot" that only happens when scientists from different disciplines collaborate. We truly have been a team and I am proud to have helped shape the direction of the lab. We have had a good ride over the years and I hope we will continue to work together in the future. Esteban has also allowed me to continue other research pursuits, particularly field work and teaching in sub-Saharan Africa, and I am indebted to him for his support on those pursuits.

Within the Burchard Lab, I need to especially thank Celeste for being such an amazing and supportive lab manager and her tireless efforts to keep the wet lab moving. I need to thank Dara and Josh especially, I hope as we all mature and grow in our own research avenues we will continue to find opportunities for collaboration. I would also like to thank everyone in the Burchard Lab over the years whom I have worked with: Marc, Melinda, Kat, Scott, Donglei, Lindsey, Elizabeth, Sandra, Sam, Katie, Maria, Neeta, Cristina, and everyone else. Also Laura Fejerman, who to me has been an honorary labmate. It has been a truly wonderful place to have a desk and be surrounded by a large number of exciting ideas.

With the Burchard Lab I also have had the opportunity to work with several large consortia. I would like to thank the EVE Asthma Genetics Consortium who have provided me the opportunity to work with primary genetic data from nine different labs across the country. I would like to thank Saunak Sen for his assistance with the novel statistical tests developed for the EVE admixture mapping meta-analysis, and as a statistical resource over the years.

I would also like to thank the GALA II investigators, particularly Dr. Jose Rodriguez-Santana for his incredible generosity over the years in hosting our field trips to Puerto Rico. I would also like to thank Taras Oleksyk and Juan Carlos Martinez Crusado for inviting us to teach in Puerto Rico and continuing to collaborate on large-scale human genetics research. I hope I have the opportunity to continue collaborations in Puerto Rico. It has been a wonderful place for field work.

I would like to thank my co-mentor, Jeff, for being such a positive and knowledgeable resource. Jeff has been a great supporter for me from even before I started at UCSF. His innate understanding of genetics continues to astound me. Jeff has been quite supportive over the years

and I hope to continue our dialog and collaborations over the years. I have been supported and had great experiences across multiple continents with members of the Wall Lab, especially Sung and Laurie. I also need to thank Elad who has been a fantastic collaborator and a highly creative scientist, a fountain of interesting ideas. The regular center meetings continue to provide opportunities for new thoughts and shape my research goals. I have to say an additional thank you to Elad for connecting us to the Latin American Cancer Epidemiology consortium, giving us the opportunity to develop and test our panel of ancestry informative markers now used broadly across multiple arrays and consortia.

A special thank you needs to go to the evolutionary genetics lab groups (the Pollard, Wall, Hernandez, and Zaitlen labs) for being an amazing group to learn from and be exposed to a diverse group of ideas. I have looked forward to that lab meeting every week and will truly miss it after I leave. It has been a source of far more knowledge than any course I have taken.

I'd like to thank people outside of UCSF that have made my life as a geneticist possible. First, Joanna Mountain, for sparking my interest in the field of human genetics and providing me opportunities to continue that research even after my "lost weekend" playing music for several years. Joanna was my first true mentor in science and I am honored that we continue to be close. Second, Brenna Henn, who has included me in her scientific pursuits throughout our careers at Stanford, 23andMe, and now across multiple continents and in multiple disciplines, and has become one of my closest collaborators and friends. Through our South African project I need to also thank our collaborators, including Eileen Hoal and Paul Van Helden and their labs, plus Nicola Mulder for providing the teaching opportunities in Africa. I also need to acknowledge the gracious welcome I received from the lab of Carlos Bustamante and his generosity in providing me projects and opportunities for collaboration over the years, especially with Andres Moreno

and Karla Sandoval, who have not only become fantastic collaborators, but also close friends. Our joint project on Mexican genomic diversity (chapters 4 and 5) certainly would not have happened without the close level of collaboration we have enjoyed over the past several years.

Of course I could not have done any of this without my family. My mother has been an incredible source of support and encouragement over the past decade. Also, my sisters Alexandra and Suzanne and their families. We all miss our late father and I hope he would be happy with my inheritance of his fascination with math, science, and the natural world. I also need to thank my close friends who have really become family over the years, including Ryan, Tara, Dan, and the whole extended Parsonage crew. A special acknowledgement must go to Peder for all our carpooling to UCSF over the years – we have logged tens of thousands of miles together.

Finally I'd like to thank my wife Meredith for being my rock throughout everything. Meredith, having known you for the past fifteen years I can honestly say that I do not know where I would be now without you. It was with your encouragement that I first went down the population genetics route and that turn in my academic career has led me throughout my life and work experience to here and beyond. I look forward to continuing our adventures visiting new parts of the world, extending our suburban farm, trying new food and wine, and experiencing all that life can offer.

Implications of admixture in the Americas for asthma and ancestry

Christopher Raymond Gignoux

ABSTRACT

Diverse forces have shaped the genomes of individuals throughout the world. It is crucial to understand those historical processes to study the genetics of individuals alive today. Nowhere else is this more important than in the study of admixed populations. A majority of individuals across the Americas are admixed, having received ancestry from sub-Saharan Africans, Europeans, and Indigenous Americans. However, to this day, admixed populations remain understudied, particularly because harnessing all information from their genomes requires in-depth population genetic analysis. This is not typically part of standard practice in genome-wide association studies. In this work I focus on two important aspects of understanding the history of admixed populations of the Americas to identify important associations with medical traits not possible using standard genetic analysis techniques. This work consists of two primary parts:

1) I develop a framework for genome-wide admixture mapping meta-analysis from high density SNP genotyping data. I use it to identify a novel, heritable risk factor for asthma in over 7,000 Latinos at the *SMAD2* locus that could not be discovered using standard genotype association techniques. I then demonstrate the downstream use of blood-based expression of *SMAD2* as a biomarker for both risk of exacerbation and poor response to bronchodilators in people with asthma.

2) Along with collaborators I developed the first fine-scale genetic map of indigenous and admixed populations across the country of Mexico, to determine how fine-scale differentiation of indigenous populations impacts the local communities of mixed ancestry. Using novel extensions

of Principal Component Analysis we identify striking geographic correlations with the indigenous component of ancestry in admixed individuals, and use this data to identify for the first time a clinically meaningful association between indigenous American origins and lung function.

This work also includes an Introduction and Best Practices recommendations on local ancestry estimation and methods for admixture mapping, and an Appendix on generating ancestry informative marker panels. Relevant code and important functions for running admixture mapping and meta-analyzing output will be made publicly available online.

TABLE OF CONTENTS

List of Tables

Chapter 2

Chapter 3

Chapter 5

List of Figures

Introduction

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendix 1

Leveraging Mixed Ancestry in Complex Trait Genetics

Population structure is a common issue in genome-wide association studies (GWAS), that, when ignored, can lead to false positives as well as decreased power. These concerns have caused researchers to identify methods to adjust out the effects of population structure or simply focus on populations with less structure. In either case choosing to ignore population structure and genetic ancestry results in an incomplete picture of underlying genetic variation important in the study of complex traits. Genetic ancestry captures important aspects of the haplotype patterns observed in real data. In this way ancestry provides important information that can be leveraged for novel discoveries GWAS. Here, I present some background on how to harness the genetic information from ancestry along the genome to map complex traits in populations of mixed ancestry. This set of techniques, also known as admixture mapping, is known to be powerful in situations where disease prevalence differs between racial/ethnic groups. We also demonstrate other important aspects of admixture mapping: we can achieve better coverage across the genome than genotyping and imputation alone since we can estimate ancestry in admixed populations accurately across the entire genome. Finally, as the field migrates from a focus on common variants to rare variants, we argue that patterns of rare variation that contribute to complex traits are likely to be captured in the ancestry of admixed individuals, making admixture mapping a particularly exciting tool to identify regions for resequencing. We will also release a suite of scripts and tools to

assist in performing admixture mapping from genome-wide SNP genotype data, called MIXOMATIC, available at: https://code.google.com/p/mixomatic/

Background

The past ten years have seen a revolution in genetics both in terms of new data and new discoveries. In particular, the advent of genome-wide association studies (GWAS) has led to a wealth of new discoveries in the field of complex trait genetics, uncovering new pathways and unexpected biological drivers of human physiology. However, the field has been plagued with the criticism of so-called “missing heritability”: variants identified via GWAS do not explain the heritable portion of a complex trait. Fewer genome-wide significant variants have been found than initially expected and these loci have explained less of the variance in disease (however notable exceptions include Crohn’s disease and age-related macular degeneration). Over the years, novel methods to capture additional information from the genome-wide array data, including imputation and CNV analysis, have continued to be popular methods for identifying additional heritable associated genetic markers from array data.

An important additional source of variation derived from genotype data is ancestry: the complex demographic and selective events that have affected our genes over the course of human history have shaped the frequencies of variants across the genome, thereby affecting the null distribution used in genetic association testing. This is known to be an issue in genetic association studies, and multiple solutions are readily available

for accounting for potential confounding due to ancestry. Most of these methods involve measuring ancestry-related genetic differences, whether using PCA[1] or another dimensional reduction technique such as clustering[2-4], or inferring an expected level of confounding with summary measures of population differentiation such as F_{st}[5]. These are then incorporated into tests of association in a way to adjust out the effects of potential population stratification. This can work to reduce inflation but it can be underpowered when ancestry is known to itself be associated with the trait of interest. In addition, best practices to account for ancestry in resequencing studies and pooled variant analyses remain open problems. This is especially a concern for populations of mixed ancestry (referred to as "admixed" populations), who have segments of their genome inherited from multiple ancestral populations.

Instead of summarizing ancestry to simply "correct" for its effects, here we discuss the possibilities of harnessing direct measures of ancestry along the genome to discover associations with complex traits. We will focus primarily on populations with recent mixed ancestry (admixed, see Figure 1) as they are most likely to benefit from this kind of analysis. Most individuals across the Americas are admixed, including African Americans and Hispanic/Latinos. These populations are also understudied in genetic epidemiology, since the vast majority of GWAS findings have been identified in populations of European descent[6]. We will discuss the background on how current algorithms estimate segments of ancestry in admixed populations (referred to as local ancestry estimation, or admixture deconvolution) with genome-wide SNP array data,

several methods for using ancestry estimates in trait mapping studies, as well as techniques required for combining admixture mapping findings across multiple studies in a meta-analysis. At the end we will discuss briefly more sophisticated tests that incorporate multiple lines of genetic evidence, such as joint ancestry and genotype association tests.

This is designed to be an introduction to relevant algorithms and analyses for admixture mapping for individuals interested in genome-wide association studies. Details of available methods, particularly computational implementations, to estimate local ancestry can be found in the original papers describing the algorithms. Here we will focus on how each method can be applied to datasets commonly used in the human genetic community. Scripts and tools will be available at the MIXOMATIC website.

Local Ancestry Estimation from Genome-wide Genotype Data

Human history spans tens of thousands of generations back to our origins in sub-Saharan Africa. Most of this history takes place in limited regions of sub-Saharan Africa. This long period of time has allowed for populations to differentiate from each other, particular at the extremes of continents. This is due to the primary method of settlement across the world known as the serial founder effect model[8] that approximates the initial settlement of new habitats as an expansion from a subset of the population that existed before. In this way, populations further along in the range expansion have a subset of

the genetic diversity of populations that are closer in proximity. While this is a very simplistic model, if we ignore admixed populations the serial expansion model fits patterns of genetic diversity across the world quite well, as supported both globally and within multiple continents[8-10]. After time this results in a pattern of isolation-by-distance, where subsequent migration has been geographically restricted, yielding more highly differentiated populations at geographic extremes.

In admixed populations, recent migration has caused at least two previously isolated populations to come into contact and interbreed. This process results in a sharing of alleles from both ancestries in subsequent generations. These alleles are not inherited at random but rather dictated by patterns of recombination, which introduce only a small number of breakpoints between maternal and paternal chromosomes in each generation. With a small number of generations of admixture compared to the previous number of generations of continental population isolation, ancestry will tend to be homogeneous across long tracts of the chromosome as there have not been enough generations to homogenize alleles. **In other words, ancestry LD in admixed populations is much higher than genotypic LD.** Population genetic theory demonstrates that contiguous tracts of ancestry in an admixed population can get into the tens of megabases[11]. This also implies that switchpoints between ancestries along a chromosome will tend to be sparse, yielding an expectation of long, contiguous blocks of ancestry along a chromosome. This is the primary assumption that algorithms to estimate local ancestry use in order to estimate locus-specific ancestry. In Figure 2 we

plot the example output of local ancestry estimation for a Puerto Rican individual, for context, showing the long tracts of ancestry.

Genome-wide Genotype Arrays and the Concept of Coverage in the Whole Genome Sequence Era

Traditional genome-wide association studies involve SNP genotyping arrays containing hundreds of thousands to millions of markers. These markers are enriched for common variants with the hope of "tagging" unobserved variation based on underlying patterns of linkage disequilibrium (LD) in the neighborhood of each SNP. Coverage is then typically evaluated with a fixed pairwise R^2 threshold from the array variants to a set of known variants, whether from HapMap or the 1000 Genomes Project. Manufacturers can only measure coverage for populations that are adequately sampled. In the case of European-descent populations estimates of coverage are likely realistic, however for the vast majority of populations these estimates are less accurate. In particular, moving beyond populations covered or related to those in the 1000 Genomes Project our knowledge of patterns of variation is far more limited.

On the other hand, by virtue of the high degree of ancestry LD in populations of mixed ancestry, genome-wide local ancestry can be estimated from any high density SNP genotyping array. Even if local ancestry estimation is less accurate than direct genotyping of tag SNPs, ancestry can capture genetic variation in regions with even limited genotype coverage or a high degree of ascertainment bias.

In addition, genotype coverage by virtue of correlation structure, is enriched for common variation. On the other hand, there is a growing interest in identifying the contributions of rare variation in disease risk. Traditionally these are identified through direct resequencing. However, rare variants tend to be population specific, which is consistent with population genetic theory and has been demonstrated in large sequencing studies such as the 1000 Genomes Project[7]. Therefore, rare variants are likely better captured through ancestry-LD as compared to genotype-LD, making admixture mapping the ideal tool to identify rare variants that contribute to complex disease. In addition, admixture mapping can combine heterogeneous effects at a single locus by virtue of capturing a larger region. With accurate local ancestry estimation available, admixture mapping is a complimentary approach to standard GWAS that will maximize the potential for novel discoveries from genetic association studies.

Approaches to Local Ancestry Estimation with Genome-wide Data

A large amount of ancestry estimation was performed prior to the widespread availability of genome-wide SNP genotype data. Earlier methods relied on a smaller set of unlinked markers and weakly linked ancestry across sites. Typical algorithms include those found in STRUCTURE[12], ANCESTRYMAP[13], and ADMIXMAP[14]. These have been successful for performing admixture mapping with a small AIMs panels (sized in the hundreds to thousands of SNPs), and typically use a Bayesian approach to integrate over the error in local ancestry estimation and biases in reference data. This, then, allows for mapping

of traits even with imperfect local ancestry estimation given a small number of markers. These algorithms were not designed to handle dense SNP genotype data and thus will not be discussed here. Instead, we will discuss more recent methods designed for similar marker densities as found in GWAS, and that provide higher levels of accuracy (>95% for Latinos, and >99% for African Americans).

As ancestry is a more precise characteristic of haplotypes rather than diploid genotypes, modern local ancestry methods either take as input pre-phased haplotypes or are incorporated into diploid measures using phase-aware methods. Current methods are typically *supervised*, requiring haplotypes representative of the ancestral populations fed into the algorithm (although there are some exceptions). However with more and more data becoming publicly available, identifying reference individuals for training local ancestry algorithms will become less challenging over time.

Local ancestry algorithms tend to fall into two categories: those that are based on fixed windows and those that are not. The fixed window heuristic uses the explicit assumption of ancestry LD: window size is set such that ancestry can reliably assumed to be constant across the window. By virtue of fixing the window size these models are inherently simpler and more computationally efficient. The likelihood of ancestry coming from one of K parental populations can be evaluated within each window. Transitions between windows of ancestry then can be modeled using overlapping sliding windows or a Hidden Markov Model (HMM). For example, the original version of

LAMP/WINPOP[15,16] used an approximate joint likelihood of unlinked genotypes (with the linked high-density SNP data thinned) across overlapping sliding windows of fixed ancestry. This will generate a simple window-based estimator that can be aggregated across multiple windows. A similar method was employed by Wall et al.[17], who used a composite likelihood rather than explicit thinning. As an alternative method of likelihood generation, PCADMIX[18,19], uses loadings from a chromosome-wide PCA to project individuals into PC space in fixed, non-overlapping windows, estimating likelihoods for each window haplotype using Gaussian discriminant analysis fitted to the clusters of reference individuals. For both LAMP and PCADMIX, an HMM is run after the classification to evaluate the most probable ancestry path for each individual.

A recent addition to the fixed-windows approach can be found in RFMIX[20], which approximates a probability using the ensemble of bootstrapped classifiers known as random forests. This method also can use random forests clustering to recruit ancestral haplotypes within the admixed individuals to boost accuracy, particularly in situations with imperfect reference data available. It also evaluates ancestry between switchpoints as a conditional random field, modeling ancestry switches along the chromosome as a discriminative process rather than the generative HMM. Another recent algorithm, LAMP-LD[21], uses a phasing-like approach similar to fastPHASE[22] to evaluate the likelihood of generating the observed haplotypes from ancestral haplotypes. A higher-level HMM is used to estimate ancestry switches between windows. As a secondary step some of these methods can identify novel switchpoints (for example, a local search

around a break in ancestry between windows for a switchpoint with a better fit to the observed alleles). WINPOP and LAMP-LD both include this second step.

In contrast, other methods attempt to jointly model ancestry and switchpoints at each site along the genome. These methods are based on HMMs estimating ancestry as a hidden state from allele to allele generated from differences between ancestral groups, using sequential Markovian processes designed to approximate population genetic theory expectations under admixture. This genotype-level estimation has the potential to localize real switchpoints more accurately, typically at the expense of computational efficiency and robustness. The most commonly used method with this kind of approach is HAPMIX,[23] which uses a similar phasing-inspired approach as LAMP-LD but is limited to two ancestral populations while still achieving high accuracy. Similarly, MULTIMIX[24] extended a similar approach to more complex admixtures and included more methods for parameter estimation depending on the user's interests.

The latest methods as tested by the Thousand Genomes Project[25] (e.g., LAMP-LD, RFMIX, MULTIMIX, and a 3-way version of HAPMIX) all give robust estimates across multiple admixture scenarios, allowing for highly accurate (e.g. >99%) estimation for 2-way admixture as present in African Americans, and >95% accuracy for 3-way admixture as present in Latinos as determined via simulation. Importantly for the user, methods are consistently biased in the same genomic regions, suggesting the specific choice of algorithm is unlikely to change local ancestry estimation much, nor greatly affect admixture mapping results. In MIXOMATIC we provide utilities to translate data

to/from both LAMP-LD and RFMIX, but any of the newer generation of local ancestry algorithms would be expected to give comparable results that can be used reliably for admixture mapping.

Using Local Ancestry to Map Traits on the Genome

Under a neutral scenario, an admixed population will be expected to have admixture proportions drawn from a multinomial distribution defined by the overall (or global) ancestry proportions[26]. In the scenario where ancestry at a locus is harboring causal variants for a certain disease, ancestry at that locus only is expected to be enriched in cases. This intuition brought up the simplest of admixture mapping tests, which does not require the recruitment of any controls for analysis, reducing cost and simplifying recruitment. By comparing the distribution of ancestry at any locus to the global ancestry patterns, each individual can serve essentially as their own control (assuming that the locus driving ancestry differences is small enough to negligibly affect the overall average genomic ancestries). This locus-specific deviation is typically measured using a z-score hypothesis test: measuring the standard deviation either empirically across the genome or using the parametric estimation directly from the multinomial distribution. This case-only test, while simplistic, has been shown to be the most powerful test for admixture mapping, even when the study design includes controls (e.g. [26,27]).

This may be true in ideal scenarios (and has shown some success in two-way admixture scenarios as with African Americans) but this test can be difficult to implement in practice. There are several reasons for this relating to both the assumptions of the test and the imperfect nature of local ancestry estimation. We outline some of the major reasons below:

1) Deviations in ancestry can be caused by other genomic forces, such as positive selection[28]. In a case-only analysis any positively selected local ancestry would appear to be associated.
2) Case-only analyses by definition ignore controls. Regions of the genome that could harbor protective alleles will go unnoticed.
3) Case-only analyses cannot take into account any other known predictors, which can lead to confounding, particularly with multiple correlated phenotypes.
4) Perhaps most important: local ancestry estimation is imperfect, and this process is more accurate in certain parts of the genome than others [29]. Regions with biased local ancestry estimation, whether through inaccurate algorithms or imperfect reference panels can appear to be significant loci [29-31].

Because of these reasons, I argue that it is important to incorporate evidence from both cases and controls in our admixture mapping. Regions with ancestry deviation in all individuals, whether from a history of positive selection or biased local ancestry estimation, would no longer appear significant as the trend would be observed in both cases and controls. Including controls at associated loci decreases power somewhat as

the control ancestry is also drawn from a distribution (one can think of this as an analogous process as adding a degree of freedom), but including them dramatically reduces false positives (see an example in figure 3).

A simple case-control test then can be formulated using a generalized linear model (GLM), modeling the association between a trait and the number of chromosomes of ancestry, incorporating known covariates where appropriate. These covariates can be genetic (e.g., accounting for potential ancestry stratification via inclusion of global ancestry), or environmental (e.g. fertility measures in the study of breast cancer[32]). GLMs are flexible, interpretable and included in numerous statistical and genetic analysis packages including PLINK. This can allow a geneticist to perform admixture mapping and interpret the output similar to standard genetic analyses.

Other study designs lend themselves to admixture mapping as well. A trio-based design with two parents and an affected proband is typically analyzed using a transmission-disequilibrium test (TDT[33]). This same TDT framework is applicable to admixture mapping[34]. Here it is crucial to phase the trios together to estimate the transmitted/untransmitted haplotypes. This will remove any potential for Mendelian errors that would bias the TDT[35]. For more complicated family relationships, the GLM framework can be extended to include variance components that can account for kinship or family relationships. In this way admixture mapping can be performed in

complex pedigrees or populations with a high degree of endogamy using linear mixed models such as EMMAX[36] or GEMMA[37].

Note that the contrasts between case-only and case-control analyses are only applicable to binary traits. Quantitative traits can be tested for association in a standard linear correlation/regression framework similar to that performed with genotypes, or categorized into a binary trait if appropriate.

Omnibus Admixture Mapping Tests

In the case of 2-way admixture (such as African-Americans), the results of admixture mapping can be captured by only looking at one ancestry. The effect of the other ancestry by definition must have the opposite effect. In contrast, populations with more complex histories (such as Hispanic/Latinos) have multiple ancestries that must be analyzed together to understand fully the patterns of ancestry at any given locus. This requires the development of slightly more complicated omnibus tests that can accommodate evidence from multiple ancestries.

To test for an association in the presence of K ancestries, a flexible approach for admixture mapping is to use nested GLMs to perform a likelihood ratio test. With case-control data, the likelihoods are calculated from a full model including K-1 local ancestry terms plus all other relevant covariates, and the restricted model that omits the local ancestry terms. The result then follows a χ^2 distribution with K-1 degrees of freedom.

The extra degrees of freedom are needed as one is combining evidence across all observed ancestries, but the tradeoff is that the model can combine evidence from all ancestries and can identify the regions with the most significant admixture mapping values given any combination of ancestries. A similar method can be extended to linear mixed models for admixture mapping.

In contrast to the other scenarios, the TDT itself does not calculate a likelihood, so to create a multi-allelic TDT we model counts of transmitted/untransmitted ancestries via another GLM in the form of Poisson regression. Here the full regression model counts of the transmitted/untransmitted pairs of each set of *K* ancestries are used, stratified by pairs of ancestry terms in the data. The resulting set of observations will be based on the total evidence given by all *K*! pairs of transmitted/untransmitted ancestries. This GLM is then compared to a null situation, where each of the counts is only modeled by an intercept. By virtue of the a single ancestry in the regression being entirely determined by the others, the result of the likelihood ratio test then again follows a χ^2 distribution with *K*-1 degrees of freedom. This is similar to a McNemar's test of matrix similarity with the removal of a degree of freedom. A more in-depth discussion of these is available in Gignoux et al.[38], and can be seen in the omnibus admixture mapping functions available in MIXOMATIC.

Estimating Local Ancestry at Untyped Sites

Typical GWAS meta-analyses require imputation to create a consensus set of marker data even if samples are typed across various platforms. For admixture

mapping, another form of imputation of missing data is required, but the approach is much simpler than typical genotype imputation algorithms. Given the high level of ancestry LD, local ancestry at an untyped site can be estimated via linear interpolation between the neighboring sites. Essentially this is a distance-weighted average for untyped sites, accounting for the possibility of recombination on either side of the untyped site (Figure 4). Given the high levels of ancestry LD, most sites on the genome will be within a block of ancestry, but the linear interpolation method captures some of the uncertainty we observe around switchpoints. This way any data platform can imputed up to a reference dataset such as HapMap or 1000 Genomes. In MIXOMATIC we provide a function for generating imputed local ancestry calls.

Depending on the goal, it may be possible, particularly with dense imputation (such as 1000 Genomes Project data) to use linear interpolation to impute the test statistic directly. This method has not been tested extensively but only requires one round of imputation, rather than imputing each individual, and so could provide massive computational efficiency. The scripts provided can easily be modified to do this if appropriate for a researcher's specific study.

Meta-analysis of Admixture Mapping

Admixture mapping results in a similar set of statistics used for GWAS meta-analysis, including odds ratios and standard errors for single ancestry associations, and z-scores. These can be combined using standard meta-analysis techniques. Many programs exist for combining p-values; we include Fisher's method in MIXOMATIC and

encourage users to try meta-analysis with covariate testing using METAFOR[39], as well as some of the powerful random effects models available in METASOFT[40].

Estimating the multiple testing burden

By virtue of ancestry LD extending across broader regions, the Bonferroni-based GWAS threshold of $p \leq 5 \times 10^{-8}$, assuming one million independent tests across the genome, is overly stringent. Yet with the observed level of correlation in admixture mapping it can be difficult to evaluate what constitutes genome-wide significance. The gold standard for this is to perform permutations, however these are extremely computationally expensive, particularly for a meta-analysis. Approximations of the data exist, but these still use correlations from the raw data rather than summary statistics shared in a meta-analysis. However, multiple methods exist that attempt to estimate the data directly from genotype correlations, including SLIDE[41] and spectral methods[42], but these become difficult to consolidate across studies using multiple arrays.

An efficient approximation of the multiple testing threshold first proposed by Shriner et al.[43] involves serial autoregression along the chromosome. Given that local ancestry is typically modeled under Markovian assumptions, this method is well suited to modeling correlation structure from site to site. Similar to the local ancestry imputation, this measure of correlation can be done either on the local ancestry calls themselves or the summary of the data. In our meta-analysis of admixture mapping for asthma [38], we used autoregression of the effect size estimates across the meta-analysis

resulting in empirical estimates of the multiple testing burden. These are very close to estimates given from permutations[13,44], while remaining extremely efficient.

We provide a function using a similar method to Shriner et al.[43] to calculate the multiple testing burden, although we encourage readers to investigate multiple methods or try permutations with a high performance computing cluster.

Joint Genotype/Ancestry Tests

In some situations, particularly given the indirect associations observed in GWAS[45], it is advantageous to combine evidence from both traditional GWAS and admixture mapping at a single locus. In its most basic form, this can also be captured by a likelihood ratio test estimated by GLMs, where the difference between the full and restricted models includes all the evidence at a locus: both local ancestry and genotype, while the restricted model only includes the other covariates. While penalized by multiple degrees of freedom, it should have appropriate false positive rates and is not based on heuristics or modeling assumptions of associated regions. This method can also be approximated by combining p-values from admixture mapping and a GWAS adjusting for local ancestry, using a meta-analysis method (for example, Fisher's method included in MIXOMATIC).

The inclusion of the high ancestry LD in the model should serve to decrease the multiple testing burden compared to GWAS. For example, using permutations from the Galanter

et al. data[44] we found that the joint admixture/genotype likelihood ratio in the GALA II study of Latinos had a multiple testing burden of $3x10^{-7}$ as compared to the standard GWAS threshold of $5x10^{-8}$. Here the result of the likelihood ratio test would follow a χ^2 distribution with *K* degrees of freedom (including an extra degree of freedom for genotype). This extra degree of freedom is a penalty however multiple groups have demonstrated that for many realistic scenarios incorporating both lines of evidence has more power than either GWAS or admixture mapping alone [46,47].

Several groups have published on ways of getting around the extra degree of freedom penalty. Each has its benefits and disadvantages. Pasaniuc et al.[27] used a model of a causal allele driving associations in both admixture mapping and GWAS. Their MIX statistic combines evidence from both a genotype association adjusting for local ancestry and the expected admixture mapping value at that locus driven by the causal allele frequencies in the ancestral populations. Because the likelihood is only driven by the single variant's contribution to both genotypic association and local ancestry differences, this remains a 1 degree of freedom test, providing increased power in the right scenario, particularly for markers with high F_{st} between ancestral populations. However by design their model can only identify associations that fit the expectations of their causal model. In practice this may not always fit the data, particularly if the genotypic association is indirect. In addition, their standard admixture mapping values are given by case-only analysis and can be sensitive to all the potential biases given

previously, although case-control admixture mapping values can easily be integrated into the MIX framework.

In contrast, Shriner et al.[43] used a Bayesian framework called BMIX to combine evidence from both association techniques in a two step process. First, admixture mapping was performed assuming a uniform prior across the genome. The posterior was then approximated using p-value-based likelihoods calculated from relevant central and non-central χ^2 distributions (with parameters estimated from multiple testing and expected power values). That posterior was then used to update the prior for the GWAS, stratifying out local ancestry, run as a separate GLM. A GWAS posterior is then calculated that combines evidence from both admixture mapping in the form of the locus-specific prior and the genotypic evidence beyond that from local ancestry. This test then is assumed to be significant then when a posterior value (incorporating the multiple testing burden) is above 50%, indicating that the model supports association. The BMIX method is far more flexible than MIX as it places no restrictions on causality or effect direction, however is more approximate than either the full-df methods or MIX. Both MIX and BMIX represent novel breakthroughs to the field, and demonstrate increased power by leveraging the rich population history of admixed populations to discover new traits. Local ancestry functions in MIXOMATIC will dovetail with both, particularly BMIX as it is coded in R and can be incorporated in standard MIXOMATIC analysis.

Fine Mapping

Unfortunately the downside of high coverage of Ancestry LD means significant ancestry peaks tend to be broad, potentially on the order of megabases (similar to linkage peaks). This can make it difficult to identify the gene driving the association. One must resort to additional association techniques, whether genotypic or gene expression association, to identify the specific variation driving the association with ancestry. First, identifying the borders of the peaks can be a challenge. Researchers can use a fixed threshold of 1 LOD score (or approximated by a change of 1 power term in the p-value), a change in likelihood, or choose a fixed p-value threshold to define the bounds of an admixture mapping peak around the significant maximum. Regardless of the approach it is important to note that these are approximate boundaries. Certainly, identifying the gene of interest within the genome-wide significant portion of the admixture mapping peak is the ideal scenario, and therefore region deserves further focus.

I will focus on several strategies for genotype fine mapping as gene expression validation is similar whether following up on GWAS or admixture mapping hits. With genotype/imputation-based fine mapping, the goal is to identify the genotype driving the association within the admixture mapping peak. This can happen one of two ways. The first is by finding a genotype associated with the outcome. The idea is to identify a genotype associated beyond the admixture mapping signal. Typically this will require adjustment by local ancestry and not just global ancestry (to account for the known local

ancestry effect). Here joint genotype and local ancestry testing can also be used to boost results.

The other approach for fine mapping is based on the notion that the local ancestry estimates are driven by the genotypes themselves, and so causal genotypes should contribute to the local ancestry association. Intuitively markers with a high amount of ancestry information should be driving the local ancestry association. An example of this can be seen in Fejerman et al.[32], where the authors used a greedy search to identify the subset of SNPs that best explained their admixture association (ie, when included in the model, the local ancestry term went from being genome-wide significant to >0.05), as a suggestive list of top variant candidates within the admixture mapping peak.

A Pipeline for Admixture Mapping

Retrieving reference data: For African Americans primarily receiving ancestry from western African and European ancestry[20], the CEU and YRI populations from HapMap/1000 Genomes are publicly available. Native American samples from Mexico and South America are publicly available either on the Affymetrix 6.0 platform[48], or the Illumina 650Y as part of the HGDP[49].

Ensure you have high quality genotype data: low-quality genotypes can cause ancestral misclassification[15] and confound admixture mapping results. Ensure you remove any sites with any high levels of missingness or extreme deviations from Hardy-Weinberg equilibrium. Remove monomorphic sites if required by the local ancestry algorithm. If possible, remove sites C/G and A/T SNPs as these can possibly have ambiguous stranding. The find_cg_at.py script in MIXOMATIC will return a list of markers for PLINK to filter out given a .bim file. Importantly, given the high levels of ancestry LD you do not need the full complement of SNPs to perform accurate local ancestry estimation. Certainly a greater number of SNPs should perform better, but as a general rule several hundred thousand markers should suffice for accurate local ancestry estimation from modern array data[20,25]. This, then, allows for local ancestry estimation utilizing disparate ancestral data from multiple platforms with varying overlap in the SNPs genotyped. As an example, local ancestry tracts for 1000 Genomes CLM, MXL, and PUR individuals were estimated using Native American reference haplotypes from the Affymetrix 6.0 with <200,000 variants overlapping.

Create a data freeze of high quality SNPs from your admixed population and then intersect with your reference data. Ensure your data is internally consistent and free of obvious QC problems using PCA or ADMIXTURE[4] to ensure that the reference populations are identifying expected ancestry proportions without biases.

Phase and run local ancestry estimation: modern algorithms typically require haplotypes, and will be more accurate when haplotypes are used as input[17,19]. This will require parsing scripts, and MIXOMATIC includes parsing scripts to format data both into and out of beagle format. In contrast shapeIT uses PLINK binary files as input, but a parsing script is provided to send shapeIT output to RFMix. Here if you have individuals in trios you will need to ensure that your phaser of choice is expecting family data. Run the data through your local ancestry algorithm. This step can be long and particularly memory intensive. Once the program finishes, the output will be one value for each ancestry per SNP or window, which can then be used for admixture mapping.

Admixture Mapping: Output will be one of K ancestries for each site in each haplotype. The output can be used to calculate K ancestry-specific matrices. This will recapitulate a biallelic SNP (e.g. whether the ancestry itself is major/minor, and its use in effect sizes). Basic GLM functions in MIXOMATIC are provided for logistic and linear models incorporating other covariates in R. In addition, we provide a TDT test for trio data formatted from beagle including the transmitted/untransmitted haplotypes.

Multi-way admixed populations such as Latinos have multiple ancestries at any given locus, necessitating a more complex test of omnibus ancestry. Here we provide GLM-based likelihood ratio tests for omnibus admixture mapping. Similarly we also provide joint genotype/ancestry tests.

Ancestry Imputation: imputation of untyped sites via linear interpolation is needed to consolidate results across multiple platforms as is common in meta-analyses. Here in MIXOMATIC we provide a function in R for interpolation during admixture mapping and a faster text-based interpolation using Python. These can interpolate up to Hapmap II, 1000 Genomes, or any other data set relevant to the researcher.

Results from this pipeline can then be used in standard meta-analysis frameworks.

Figures

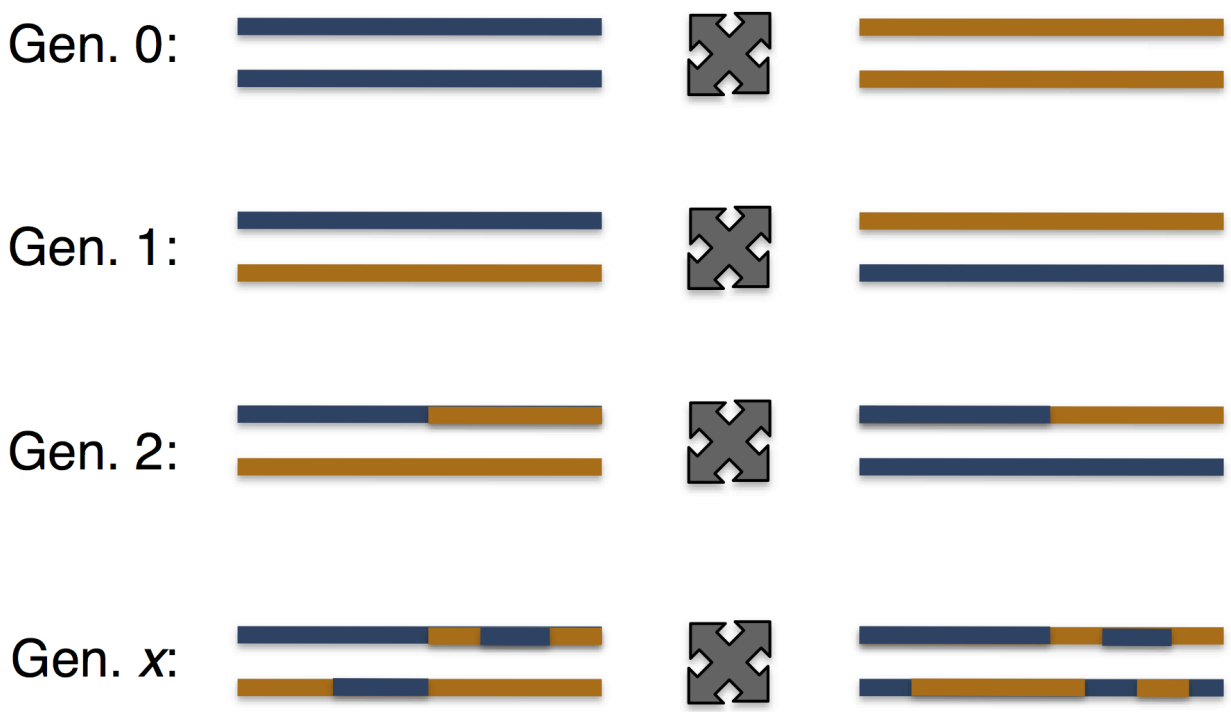


Figure 1. The process of admixture. At generation 0 the two ancestral populations remain distinct as given by the two distinct colors. After one generation, individuals have heterozygous ancestry. From then recombination breaks down the ancestry into smaller and smaller tracts, yielding the mosaics observed today from high density genotype or sequence data.

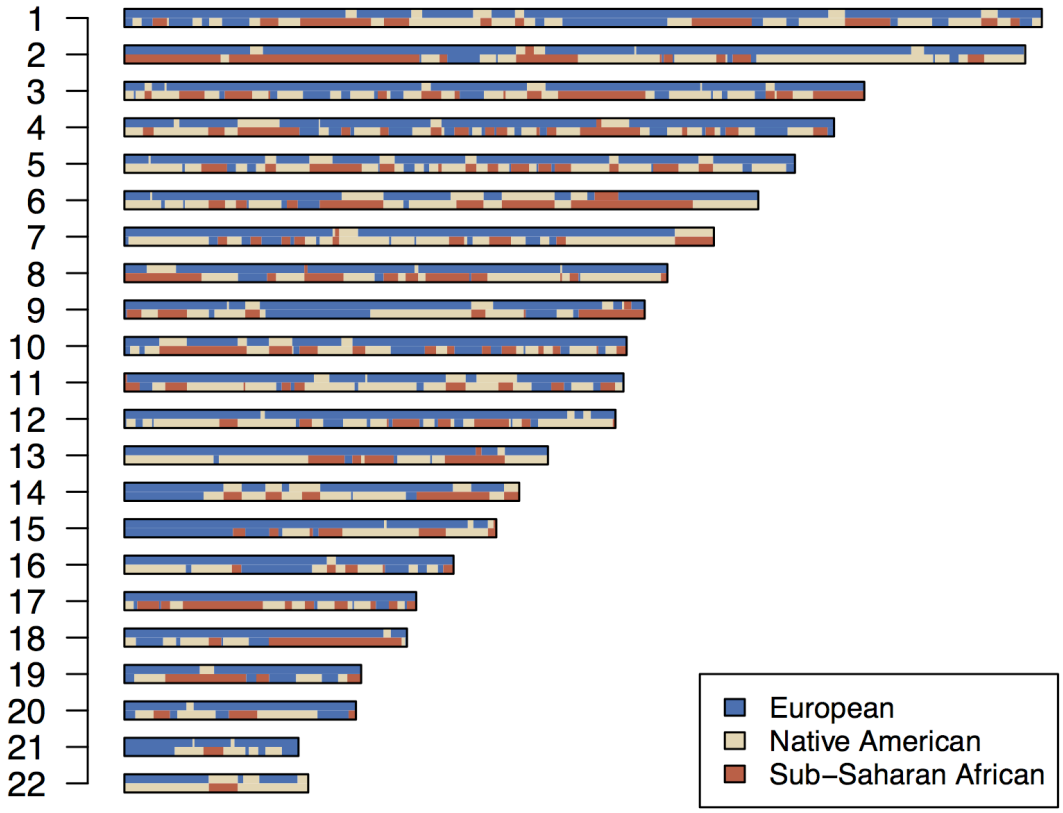


Figure 2. Ancestry karyogram for a Puerto Rican individual in GALA I, as estimated using LAMP/WINPOP. The postcolonial process of admixture results in a mosaic of ancestry, where individuals tend to have tracts of ancestry 0.5 cM-50cM long. These ancestries can be estimated from high-density genotyping data to high accuracy.

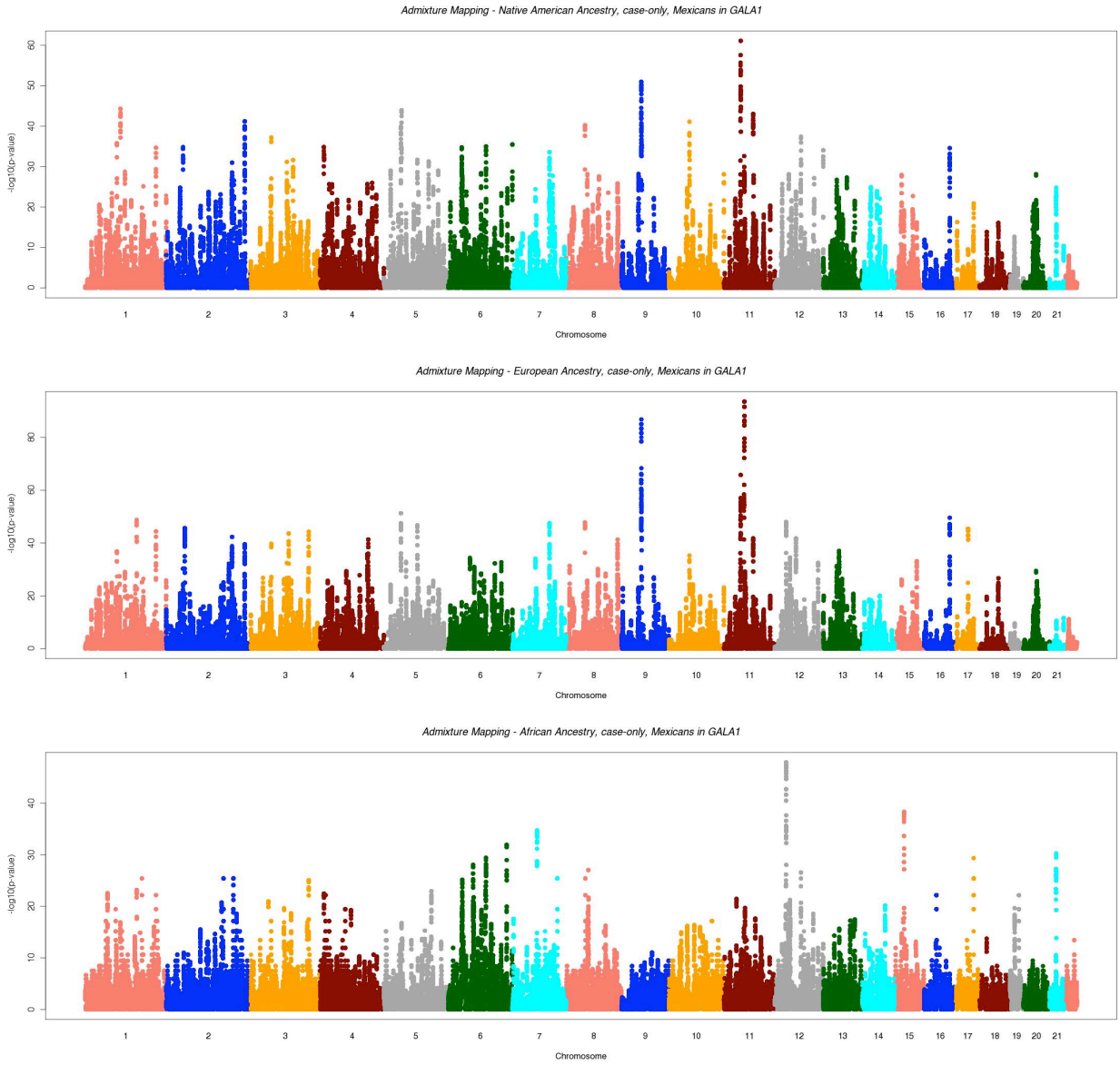


Figure 3. Case-only analysis in Mexicans can lead to extreme inflation, primarily from imperfect reference panels and biased local ancestry estimation. Data is from Mexican asthma cases included in Torgerson et al.[50]. For each ancestry there is a high degree of inflation in contrast to the appropriate type 1 error rates in the case-control analyses using the same data (see [50]).

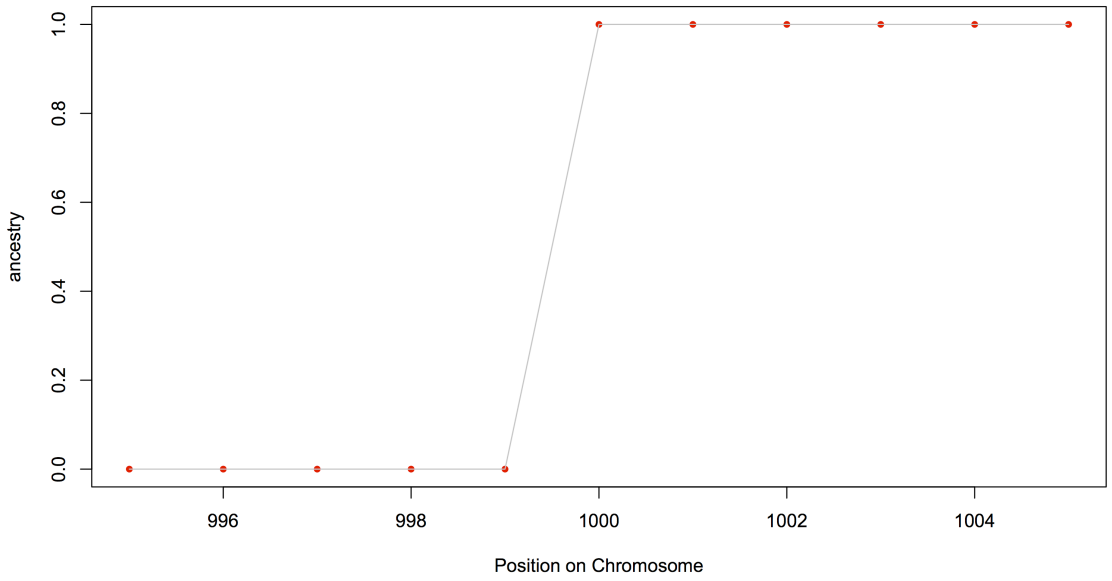


Figure 4. Description of ancestry interpolation/imputation scheme. Given observed ancestries in red from local ancestry estimation, our best guess of ancestry at unobserved sites is given from a linear interpolation of genetic position (in Morgans), here shown as a gray line. Importantly, most interpolated ancestry estimates on the genome will be identical to the flanking site.

Sources

1. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 2006;38:904-9.

2. Pritchard J, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945-59.

3. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genetic Epidemiology 2005;28:289-301.

4. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19:1655-64.

5. Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;55:997-1004.

6. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature 2011;475:163-5.

7. Gravel S, Henn BM, Gutenkunst RN, et al. Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences of the United States of America 2011;108:11983-8.

8. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America 2005;102:15942-7.

9. Wang C, Zollner S, Rosenberg NA. A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet 2012;8:e1002886.

10. Henn BM, Gignoux CR, Jobin M, et al. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proceedings of the National Academy of Sciences 2011;108:5154-62.

11. Gravel S. Population genetics models of local ancestry. Genetics 2012;191:607-19.

12. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003;164:1567-87.

13. Patterson N, Hattangadi N, Lane B, et al. Methods for high-density admixture mapping of disease genes. Am J Hum Genet 2004;74:979-1000.

14. Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. Design and analysis of admixture mapping studies. American journal of human genetics 2004;74:965-78.

15. Pasaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific ancestry in closely related populations. Bioinformatics (Oxford, England) 2009.

16. Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. American journal of human genetics 2008;82:290-303.

17. Wall JD, Jiang R, Gignoux C, et al. Genetic variation in Native Americans, inferred from Latino SNP and resequencing data. Molecular biology and evolution 2011;28:2231-7.

18. Henn BM, Botigue LR, Gravel S, et al. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet 2012;8:e1002397.

19. Brisbin A, Bryc K, Byrnes J, et al. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum Biol 2012;84:343-64.

20. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. Am J Hum Genet 2013.

21. Baran Y, Paşaniuc B, Sankararaman S, et al. Fast and accurate inference of local ancestry in Latino populations. Bioinformatics (Oxford, England) 2012.

22. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American journal of human genetics 2006;78:629-44.

23. Price AL, Tandon A, Patterson N, et al. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLoS genetics 2009;5:e1000519.

24. Churchhouse C, Marchini J. Multiway Admixture Deconvolution Using Phased or Unphased Ancestral Panels. Genet Epidemiol 2012.

25. Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56-65.

26. Montana G, Pritchard JK. Statistical tests for admixture mapping with case-control and cases-only data. Am J Hum Genet 2004;75:771-89.

27. Pasaniuc B, Zaitlen N, Lettre G, et al. Enhanced Statistical Tests for GWAS in Admixed Populations: Assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS genetics 2011;7:e1001371.

28. Tang H, Choudhry S, Mei R, et al. Recent genetic selection in the ancestral admixture of Puerto Ricans. American journal of human genetics 2007;81:626-33.

29. Pasaniuc B, Sankararaman S, Torgerson DG, et al. Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. Bioinformatics 2013;29:1407-15.

30. Moreno A, Gignoux C, al. e. The Genetic History and Structure of Mexican Populations. in prep 2013.

31. Moreno-Estrada A, Gravel S, Zakharia F, et al. Reconstructing the Population Genetic History of the Caribbean. arXiv:13060558 2013.

32. Fejerman L, Chen GK, Eng C, et al. Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas. Hum Mol Genet 2012;21:1907-17.

33. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506-16.

34. Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 1995;57:455-64.

35. Gordon D, Haynes C, Johnnidis C, Patel SB, Bowcock AM, Ott J. A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. European Journal of Human Genetics 2004;12:752-61.

36. Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 2010;42:348-54.

37. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics 2012;44:821-4.

38. Gignoux CT, Torgerson DG; Galanter, JM et al. Meta-analysis of admixture mapping using existing genome-wide association data implicates SMAD2 as a novel asthma-associated locus in Latinos. Submitted 2013.

39. Viechtbauer W. metafor: Meta-Analysis Package for R. R package version 2010;2010:1-0.

40. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. American journal of human genetics 2011;88:586-98.

41. Han B, Kang HM, Eskin E. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. PLoS Genet 2009;5:e1000456.

42. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb) 2005;95:221-7.

43. Shriner D, Adeyemo A, Rotimi CN. Joint ancestry and association testing in admixed individuals. PLoS Computational Biology 2011;7:e1002325.

44. Galanter JM, Gignoux CR, Torgerson DG, et al. GWAS and admixture mapping identify asthma-associated loci in Latinos: The GALA II Study. J Allergy Clin Immunol 2013;accepted.

45. Spencer C, Su Z, Donnelly P, Marchini J. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. PLoS genetics 2009;5:e1000477.

46. Tang H, Siegmund DO, Johnson NA, Romieu I, London SJ. Joint testing of genotype and ancestry association in admixed families. Genetic Epidemiology 2010;34:783-91.

47. Liu J, Lewinger JP, Gilliland FD, Gauderman WJ, Conti DV. Confounding and heterogeneity in genetic association studies with admixed populations. Am J Epidemiol 2013;177:351-60.

48. Bigham A, Bauchet M, Pinto D, et al. Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. PLoS genetics 2010;6:e1001116.

49. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science (New York, NY) 2008;319:1100-4.

50. Torgerson DG, Gignoux CR, Galanter JM, et al. Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. J Allergy Clin Immunol 2012;130:76-82 e12.

Chapter 2. An admixture mapping meta-analysis identifies an ancestry-specific risk factor and potential biomarker for asthma

Abstract

Background

Asthma is a common but complex disease with significant racial/ethnic disparities in prevalence, morbidity, and response to therapies. Analysis of genetic ancestry can help to explain these differences.

Methods

We leveraged the mixed ancestry in 7,008 Latinos and African Americans in the EVE Asthma Genetics Consortium to perform an admixture mapping meta-analysis for asthma. We replicated associations in GALA II, an independent study of 3,774 Latinos. We measured gene expression in the whole blood of 161 Puerto Ricans from our replication sample to identify potential biomarkers for lung function, bronchodilator drug response, and exacerbations.

Results

We identified a genome-wide significant admixture mapping peak centered on *SMAD2* in Latinos ($p=6.8 \times 10^{-6}$), where Native American ancestry was associated with increased risk of asthma (OR=1.20, 95% CI=1.07-1.34, p=0.002) and European ancestry with decreased risk (OR=0.86, 95% CI=0.77-0.96, p=0.008). Our findings replicated in GALA II ($p=5.3 \times 10^{-3}$, overall meta-analysis $p=2.6 \times 10^{-7}$). Asthma cases had 28% lower whole blood expression of *SMAD2* compared with controls (95% CI:12–37%, p<0.001), corresponding to a best-fit OR

of 8.57 (95% CI=2.57-25.19, p<0.001). Lower *SMAD2* expression was also associated with decreased albuterol response and increased numbers of exacerbations.

Conclusions

We identified a Latino-specific association between local ancestry at *SMAD2* and asthma, and found that decreased *SMAD2* expression in the blood was strongly associated with increased asthma risk and severity. Our findings may help explain differences in asthma prevalence and morbidity between racial/ethnic groups, and identified *SMAD2* expression in blood as a potential biomarker for asthma.

Word Count: 246

Introduction (word count 587)

Asthma prevalence varies dramatically between racial and ethnic groups. In the United States, childhood asthma prevalence is highest among Puerto Ricans (24.8%), intermediate among African Americans (16.3%) and lowest among European Americans (7.8%) and Mexican Americans (7.8%).[1] These racial/ethnic disparities extend to asthma mortality, which is four-fold higher in Puerto Ricans and African Americans than in Mexican Americans.[2] Substantial evidence supports a genetic basis for asthma, with estimates of heritability as high as 75%.[3] Genome-wide association studies (GWAS) have identified >25 novel genetic risk factors for asthma.[4] Nonetheless, known genetic associations account for only a small proportion of the genetic basis of asthma, and have provided limited insight into racial disparities in its prevalence and severity. This is partially due to the limited number of GWAS studies in non-European populations.[5-7] Many asthma-associated variants identified in European Americans demonstrate significant heterogeneity or simply have failed to replicate in non-European groups.[8,9] In addition, rare, population-specific genetic risk factors are likely to play a role. These unexplored genetic factors may contribute to disparities in asthma prevalence and severity across populations.

While exome and whole-genome sequencing may identify novel variants associated with complex disease,[10] such approaches are costly in large population samples and present numerous analytic challenges. One alternative is to re-mine existing GWAS data through

admixture mapping to identify novel disease-associated loci in diverse populations using a less expensive[11] and more flexible[12,13] methodology. Latinos are primarily descendants of a three-way admixture of Native American, European, and sub-Saharan African ancestors,[14,15] and African Americans are primarily admixed descendants of sub-Saharan African and European ancestors.[16] This wide variation in genetic ancestry, along with socioeconomic and environmental differences at both individual and population levels, can be leveraged to explore the underpinnings of disparities in asthma prevalence and severity.[17] We previously demonstrated that variation in overall genetic ancestry was associated with asthma,[18] lung function,[19] and bronchodilator responsiveness.[20] If the frequencies of patterns of disease-causing genetic variation are different between the ancestral populations of admixed individuals, the frequency of genetic ancestry at that locus will be also be different: these loci can be identified through admixture mapping. We hypothesize that the dramatic differences in prevalence between racial and ethnic groups make asthma an ideal candidate for this technique. Indeed, we have previously demonstrated the utility of locus-specific genetic ancestry estimated from genome-wide association data to identify novel genetic risk factors in both African Americans and Latinos[21,22].

Although admixture mapping can identify a locus in the genome, further characterization is needed to identify the relevant gene. In several instances, measures of gene expression have augmented GWAS studies in the search for genes that contribute to complex disease.[23,24] Evaluation of gene expression can provide insight as to the functional effect of causal genetic variation driving the observed association, and characterize downstream effects in relevant tissues. In the case of blood and other easily collected tissue, evaluating

gene expression signals also permits identification of novel biomarkers for disease and severity.

We hypothesized that admixture mapping could identify novel, potentially population-specific risk factors for asthma in Latinos and/or African Americans. Our prior meta-analysis using traditional GWAS methods for asthma in three racial/ethnic populations in the U.S. replicated a number of known associated regions, and identified an African American-specific association at *PYHIN1.*[25] Here, we extend these studies by performing an admixture mapping meta-analysis for 7,008 Latino and African American subjects included in the EVE Asthma Genetics Consortium (www.eve.uchicago.edu), with the goal of identifying novel, and potentially population-specific risk factors for asthma that are captured by ancestry from existing genome-wide genotype data.

Methods: (word count 596)

We outline the study approach in brief in Figure 1.

Study Subjects

Discovery Population

We included data from self-identified Latino and African American subjects from nine independent studies included in the EVE Asthma Consortium in our admixture mapping meta-analysis. Detailed descriptions of all studies are published elsewhere.[25] EVE is a large, multi-ethnic assembly of asthma studies with existing genome-wide SNP genotypes from nine different U.S. institutions. All autosomal genotypes passing quality control standards

were included in the current study, including 3,902 Latinos and 3,106 African Americans (Table 1).

Replication Population

We tested genome-wide significant associations from the discovery population in the Genes-environments & Admixture in Latino Americans (GALA II) Study,[26] a large, multi-center case-control study of Latino children between the ages of 8-21 years with and without asthma (Table 1 and Supplementary Material). Local institutional review boards approved the studies and all subjects and legal guardians provided written informed assent/consent. A total of 4,041 children (1,976 participants with asthma and 2,065 healthy controls) were recruited from five centers (Chicago, Bronx, Houston, San Francisco Bay Area, and Puerto Rico) using a combination of community- and clinic-based recruitment. Participants with asthma self-reported a physician diagnosis of asthma and reported at least two symptoms (shortness of breath, wheezing, or cough not associated with upper respiratory illness) or chronic use of controller medication (inhaled corticosteroids, leukotriene modifying agents, theophylline or oral steroids) in the two years preceding recruitment.

All individuals in GALA II were subject to extensive phenotype characterization, including pulmonary function testing in accordance with ATS criteria. Subjects with asthma were evaluated for bronchodilator response. A subset of 3,774 GALA II subjects were genotyped on the Affymetrix Axiom® Genome-Wide LAT1 Array[27] and passed manufacturer-recommended standard quality control measures, yielding 747,129 SNPs.

Gene Expression Analyses

Whole blood RNA was extracted from a random subset of 161 individuals (107 cases, 54 controls) of Puerto Ricans from our replication sample (GALA II). Gene expression levels were measured using quantitative PCR and normalized using the housekeeping gene *GUS*. Expression levels were calibrated using the delta Ct transformation. All samples had an RNA integrity value > 7.

Statistical Analyses

Statistical analyses were performed using R, Python, and PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/). Local ancestry was estimated individually for each study using one of two LAMP algorithms: LAMP[28] for case-control studies, and LAMP-HAP[29] for family-based studies to preserve transmitted/untransmitted haplotype status.

We used a 2-degrees of freedom likelihood ratio test to jointly evaluate the local effect of the three ancestral populations in Latinos (Supplementary Material). Case-control studies were analyzed using a logistic regression model, while trio-based studies were analyzed with a Poisson regression model of the counts of transmitted and untransmitted alleles. To establish a study-specific significance criterion, we employed an empirical autoregression framework, using the *coda* package in R.[30] We combined the p-values of the likelihood ratio test using custom Python scripts.

Gene expression levels between cases and controls were compared using linear regression, adjusting for age, sex, and recruitment center and ancestry as appropriate. We evaluated the goodness-of-fit for varying cut points of high- versus low-expression as a predictor of asthma status using Bayes Factors calculated from generalized linear models. We estimated the prediction power of multiple models using self-reported exacerbation scores (combining information on hospitalizations, emergency department visits, and oral steroid usage), and performed model selection in generalized linear models by estimating the AUC of ROC curves and the Aikake Information Criterion (AIC). We also compared estimates of population attributable risk (PAR) across genetic and environmental factors.

Further details on methodology, including admixture mapping methods, imputation, and predictive modeling are presented in the Supplementary Material.

Results (word count 774)

Admixture Mapping Meta-Analysis and Replication

We performed local ancestry estimation, ancestry interpolation, and admixture mapping independently in five different studies, comprising a total of 3,902 Latino individuals from the EVE Consortium. A meta-analysis using Fisher's method produced highly concordant results to Stouffer-Liptak weighted Z-scores; accordingly, we present only the results from Fisher's method and coefficients estimated from a fixed effects meta-analysis.

We identified a genome-wide significant admixture mapping peak that was specific to Latinos at 18q21 ($p=6.8 \times 10^{-6}$, Figure 2A and B, significance threshold $p< 4.1 \times 10^{-5}$, see

Supplementary Material). The ancestry association in Latinos was primarily driven by differences in European (OR=0.86, 95% CI=0.77-0.96, p=0.0084) and Native American (OR=1.20, 95% CI=1.07-1.34, p=0.0016) ancestry between cases and controls, whereas African ancestry did not appear to play a significant role (p=0.42) (Figure 2C). The admixture peak overlapped two genes: *SMAD2* and *ZBTB7C* (Figure 2B), with ancestry at *SMAD2* having the strongest association with asthma.

We replicated the association between asthma and ancestry at 18q21 in an independent sample of 3,774 Latinos from the GALA II Study ($p=5.3x10^{-3}$, Table S2). The direction of the effect in both the discovery and replication populations were homogeneous for both European and Native American ancestries (OR=0.87, 95% CI 0.78-0.96, p<0.01 and OR=1.09, 95% CI 1.02-1.16, p<0.01 respectively, see Table S2). Applying the same admixture mapping approach in the 3,106 African Americans in EVE, there was no ancestry association at 18q21 (Figure S1, p=0.7). In addition we found no significant genotype associations in the 4,531 European Americans in EVE.

Gene Expression Associations with Asthma and Secondary Phenotypes

We measured the expression of *SMAD2*, *SMAD3* (the cystolic hetero-dimeric partner of *SMAD2*), and *ZBTB7C* via rt-PCR from RNA isolated from whole blood in a random subset of 161 Puerto Ricans in the replication study (GALA II). *SMAD2* expression was significantly negatively associated with asthma. Cases had 28% lower mean levels of *SMAD2* expression than healthy controls (95% CI=12-37%, p<0.001, Figure 3A). Neither *SMAD3* nor *ZBTB7C* showed any difference in expression between asthma cases and controls (p= 0.8 and 0.9

respectively), and neither gene showed a significant correlation with global or local ancestry. We determined, via Bayes Factors, the best-fit cutpoint at 66% of mean expression in controls for partitioning low-vs.-high expressors. We found that low *SMAD2* expression using this cutpoint was associated with an 8-fold increased odds of asthma (OR 8.05, 95% CI 2.57-25.19, p<0.001). In addition, Puerto Rican cases recruited in Puerto Rico had 39% lower *SMAD2* expression as compared with Puerto Rican cases recruited in mainland U.S. (95% CI 22-56%, cutpoint OR 6.79 (1.99-23.19), p<0.001); we observed no significant difference in controls. The association between *SMAD2* expression and asthma remained significant adjusting for island-vs.-mainland or by study center.

After adjustment for known anthropometric predictors of lung function (e.g., age, sex, and height[2]), *SMAD2* expression was not significantly associated with baseline lung function across four standard measures (FEV_1, FVC, FEF_{25-75}, and PEFR). However, low *SMAD2* expression was significantly associated with decreased bronchodilator drug response and increased asthma exacerbation. Specifically, we found that a 10% decrease in *SMAD2* expression corresponded to a 1.7% decrease in bronchodilator drug response (ΔFEF_{25-75}, 95% CI -2.6- -0.7%, p<0.01, Table 2). *SMAD2* expression was also associated with ΔFVC, however this was not significant after adjusting for ΔFEF_{25-75}. The correlation between ΔFEF_{25-75} and *SMAD2* expression remained significant after adjusting for ΔFVC, supporting the primary association between *SMAD2* expression and ΔFEF_{25-75}. Low *SMAD2* expression was also associated with increased asthma exacerbation score (ordered logistic regression per 10% decrease in expression, OR=1.16, 95% CI=1.01-1.35, p=0.02, Table 2).

We then built logistic regression models to test the ability of *SMAD2* expression to explain asthma exacerbations. Individuals were dichotomized into categories of low risk (score of ≤1, i.e., no more than one exacerbation) and high risk (score ≥2, i.e., multiple or severe exacerbations). We applied three prediction models: use of controller medication (any long-term asthma medication), response to albuterol, and *SMAD2* expression. We limited the analysis to 79 GALA II cases residing in Puerto Rico to minimize confounding. Incorporating all three classes of predictors in the model had the highest ROC curve AUC (81%, Figure 3B), while minimizing the AIC, thus providing good predictive power beyond standard clinical measurements.

Transforming these observations to population attributable risks (PARs) in the context of other genetic and environmental risk factors, low *SMAD2* expression has a PAR of 40% (95% CI: 17-60). In contrast, established risk factors such as obesity, air pollution, and well-replicated genotypic risk factors at 17q21 have a more limited role in asthma (Figure 2C), with the total PAR of these risk factors (38%, 95% CI 18-37) being lower than that of *SMAD2* expression by itself.

Discussion (word count 835)

In this novel investigation of admixture mapping and asthma, we identified a genome-wide significant association between ancestry at 18q21, centered on the *SMAD2* gene, and asthma in a meta-analysis including 3,902 Latinos from the EVE Asthma Genetics Consortium. We replicated this finding among 3,774 individuals in the GALA II study.

Further analysis revealed the clinically important finding that low *SMAD2* expression is associated with reduced bronchodilator response and increased asthma exacerbations. Absent our admixture mapping follow-up to a large, consortium-based traditional GWAS meta-analysis,[25] this locus would not have been discovered, as there were no individual genotypic associations within the admixture peak with a $p<10^{-4}$ (Figure S5). Notably, there was no evidence for an ancestry or allelic association in African and European Americans, reinforcing the population-specific nature of the association at 18q21. An important and unique contribution offered by admixture mapping is its increased coverage of genetic variation due to increased ancestry linkage disequilibrium (LD) as compared with genotypic LD.[31] Indeed, the top locus-wide significant imputed SNP in GALA II within the peak (Figures S5&S6) appears to be at low frequency in Europe and Africa, but is common on Native American haplotypes in Latinos, consistent with the admixture signal. This is important because prior estimates of the coverage of commercial genotyping arrays have proved overly optimistic in non-European and admixed populations.[27,32]

However, increased ancestry LD results in larger blocks of the genome being associated with the outcome, rendering identification of the specific gene more challenging than with traditional GWAS. Here, the 674kb genome-wide significant peak overlapped *SMAD2* and *ZBTB7C*. *ZBTB7C* has no known role in asthma pathophysiology and limited functional characterization. In contrast *SMAD2* is a well-characterized cofactor involved in *TGF-β* signaling. In asthma, the *TGF-β* pathway has been implicated in negative regulation of allergic airway inflammation,[33] in airway remodeling,[34] and in drug response.[35] In the *TGF-β* signaling pathway, ligation of *TGF-β* receptors activates the proximal transcription factors

SMAD2 and *SMAD3*; these associate with *SMAD4* and translocate to the nucleus to regulate transcription of several hundred target genes along with a complex of DNA binding cofactors. Lower levels of *SMAD2* are correlated with lower levels of *TGF-β* -mediated signaling effect.[36]

Although *TGF-β* pathway genes are known to play a functional role in asthma, they have rarely been identified via GWAS, and to our knowledge *SMAD2* has not been previously associated with asthma in any genetic association study. However, it has been associated with several other immune system-mediated phenotypes, including a GWAS of placental abruption[37]. Two previous meta-analyses[25,38] identified an association between *SMAD3*, the cytosolic hetero-dimeric partner of *SMAD2*, and asthma in Europeans and European Americans. Here, variation in *SMAD3* was not significantly associated with asthma through admixture mapping or traditional GWAS in either Latinos or African Americans, nor was *SMAD3* expression significantly associated with asthma in GALA II.

Our findings support the role of differential regulation of *SMAD2* in asthma cases, and highlight its potential use as a biomarker to identify individuals with low bronchodilator drug response and increased risk of exacerbation. In GALA II, the population attributable risk of low *SMAD2* expression is a highly important component of asthma, with a higher PAR than many known risk factors for asthma, including obesity, NO_2 exposure, 17q21 genotypes, and *in utero* smoking, or even all these risk factors combined (Figure 3C).

Measuring *SMAD2* expression in whole blood is an attractive biomarker candidate due to its relative ease of collection, as compared with lung tissue. Including *SMAD2* expression levels improved the explanatory power of statistical models of asthma exacerbations beyond the use of traditional variables collected in the clinic. In addition low *SMAD2* expression is associated with low bronchodilator response as measured by ΔFEF_{25-75}. Incidentally we found no association between *SMAD2* expression and ΔFEV_1, the typical spirometric measurement used for assessing drug response. Growing evidence suggests that measures of FEV_1 may underestimate asthma severity in children.[39] FEF_{25-75} better measures small airway obstruction, and has demonstrated sensitivity as a measure of airway obstruction among children and adolescents with asthma,[40] even those with normal FEV_1. Prospective studies in diverse populations are required to definitively test whether measuring *SMAD2* expression can identify children at high risk for asthma exacerbation, and therefore those who will benefit from more intensive or targeted intervention.

Beyond investigation of asthma, admixture mapping is extensible to any genome-wide analysis of disease prevalence and severity in an ancestrally mixed population. It offers superior economic and analytic efficiencies by mining previously generated genomic data. Furthermore, population-specific findings identified by admixture mapping can have clinical relevance to disparities in disease prevalence and severity, as illustrated by the associations of *SMAD2* expression patterns with asthma prevalence, poor bronchodilator drug response, and increased risk of asthma exacerbations. Our findings reinforce that alternative mapping strategies, such as admixture mapping, may capture novel and population-specific findings that traditional GWAS approaches alone cannot uncover. As

our appreciation increases for the heterogeneous genetic ancestry of numerous populations world-wide, more nuanced understanding of disease burden and treatment targets can be uncovered and developed by incorporating this technique.

Acknowledgments

The authors acknowledge the families and patients for their participation, and thank the numerous health care providers and community clinics for their support and participation in all the EVE Consortium and replication studies. In particular, the authors thank GALA I, GALA II and SAGE study coordinator Sandra Salazar; the recruiters who obtained the data: Duanny Alva, MD, Gaby Ayala-Rodriguez, Lisa Caine, Elizabeth Castellanos, Jaime Colon, Denise DeJesus, Blanca Lopez , Brenda Lopez, MD, Louis Martos, Vivian Medina, Juana Olivo, Mario Peralta, Esther Pomares, MD, Jihan Quraishi, Johanna Rodriguez, Shahdad Saeedi, Dean Soto, Emmanuel Viera, Ana Taveras. Some computations were performed using the UCSF Biostatistics High Performance Computing System. The authors would like to thank Dean Sheppard, MD, David Erle, MD, and Amy J. Markowitz for helpful edits, comments and advice on this manuscript.

Declaration of funding

This research was supported in part by National Institutes of Health (AI061774, AI077439, AI077439, AI079139, CA113710, DK064695, ES015794, ES015794, HL078885, HL079055, HL087699, HL088133, HL088133, HL104608, M01-RR00188, MD006902); ARRA grant

RC2 HL101651; Flight Attendant Medical Research Institute (FAMRI); UCSF Chancellor's Research Fellowship, Dissertation Year Fellowship, and in part by NIH Training Grant T32 GM007175 (to CRG); RWJF Amos Medical Faculty Development Award (to EGB); the Sandler Foundation; the American Asthma Foundation (to EGB); KCB was supported in part by the Mary Beryl Patch Turnbull Scholar Program; RAM was supported in part by the MOSAIC initiative of Johns Hopkins University; Ernest S. Bazley Grant (to PCA); NHLBI K23 to RK (K23HL093023); GCRC M01 RR00188 to HJF; NHLBI K23 (K23HL111636) and NCATS KL2 (KL2TR000143) to JMG. This publication [or project] was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through UCSF-CTSI Grant Number KL2TR000143. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN26120080001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

Tables

Table 1. Baseline characteristics of subjects in the meta-analysis. EVE Asthma Genetics consortium Latino individuals were used for discovery, GALA II was used for replication

along with the African American individuals in EVE for the initial admixture mapping. Imputation-based fine mapping was performed in all studies shown. Other measures from these studies can be found in Torgerson et al.[25].

Table 2. Associations of *SMAD2* gene expression with asthma and morbidity outcomes. Dichotomous outcomes use a best-fit cutpoint of 66% of healthy control *SMAD2* expression to estimate the odds ratio. Measures of drug response and exacerbations are per 10% increase in *SMAD2* expression. Static spirometric measures pre-/post-albuterol administration were not significant on their own.

Figure Legends

Figure 1. Study Approach. We began with the Latino studies in the EVE Asthma Genetics Consortium, along with reference individuals to perform local ancestry estimation. We performed ancestry imputation via linear interpolation to create a consistent set of sites across studies. We performed admixture mapping at these interpolated sites via likelihood ratio tests, then combined values across studies at all sites via meta-analysis. Genome-wide significance was measured empirically via autoregression.[30] We replicated the genome-wide signal at 18q21 in GALA II using similar methods. We then used gene expression to characterize associations with genes in 18q21 and known interactors of those genes. *SMAD2* expression was strongly associated with asthma (none of the others were), and so we investigated associations with additional phenotypes including spirometry, bronchodilator response, and exacerbations, as reported in the Results.

Figure 2. Admixture mapping meta-analysis results.

2A: Manhattan plot of the genome-wide results of admixture mapping for asthma in 3,902 Latinos from the EVE asthma genetics consortium. Genome-wide significance threshold accounting for ancestry correlation indicated via dashed line. The single genome-wide significant peak in the 2-df likelihood ratio test is found on chromosome 18.

2B: Summary of the 18q21 locus. LocusZoom[41] plot for sites in the most significant Hispanic/Latino admixture mapping association on 18q21, showing the relative position of the genes closest to the top of the peak. $-\log_{10}$ p-values are shown for the 2-df likelihood ratio test for differences across all three ancestral populations.

2C: Forest plots for admixture mapping across each of three ancestries at the top site in 2B (African, European, and Native American ancestry, respectively). Each study's odds ratio is displayed as a square and corresponding confidence interval with size inversely proportional to the standard error. Meta-analysis estimates via fixed effects models are given as diamonds. No ancestry shows no evidence of significant study heterogeneity at 18q21, where European ancestry at *SMAD2* confers protection from asthma; Native American ancestry confers increased risk, as presented in Table S1.

Figure 3. Whole blood *SMAD2* expression analyses.

3A:Scatterplots displaying *SMAD2* expression in whole blood measured by q-PCR in GALA II Puerto Rican cases (n=107) and controls (n=54). Expression was calibrated to the housekeeping gene *GUS* to create relative-fold values, including means and 95% confidence intervals. On average, cases have 25% lower expression of *SMAD2* than do controls ($p=1.2\times10^{-4}$). *SMAD3*, previously associated with asthma in Europeans, did not show expression differences between GALA II cases and controls (p=0.81, data not shown). 3B: Prediction of self-reported exacerbations using clinical variables and *SMAD2* gene

expression as measured with ROC curves from logistic regression. We only looked at individuals from the island of Puerto Rico to minimize confounding. The AUCs for each model are displayed in the legend. CM=Controller Medication, Spiro=4 Bronchodilator Response variables discussed in the main text. The model incorporating *SMAD2* expression predicts best according to the AIC.

3C: Population Attributable Risk of *SMAD2* expression and ancestry in context with other genetic and environmental risk factors in GALA II, color-coded by type. Black is expression, blue is genotype, green is ancestry, and red is environmental exposure.

Sources

1. National Health Interview Survey (NHIS) Data. 2011. (Accessed at http://www.cdc.gov/asthma/nhis/2011/table4-1.htm.)

2. Homa DM, Mannino DM, Lara M. Asthma mortality in U.S. Hispanics of Mexican, Puerto Rican, and Cuban heritage, 1990-1995. Am J Respir Crit Care Med 2000;161:504-9.

3. Willemsen G, van Beijsterveldt TCEM, van Baal CGCM, Postma D, Boomsma DI. Heritability of self-reported asthma and allergy: A study in adult Dutch twins, siblings and parents. Twin Res Hum Genet 2008;11:132-42.

4. A Catalog of Published Genome-Wide Association Studies, Accessed 03/2013., 2013. (Accessed at http://www.genome.gov/gwastudies.)

5. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature 2011;475:163-5.

6. Ober C, Hoffjan S. Asthma genetics 2006: the long and winding road to gene discovery. Genes Immun 2006;7:95-100.

7. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. Trends Genet 2009;25:489-94.

8. Galanter JM, Torgerson D, Gignoux CR, et al. Cosmopolitan and ethnic-specific replication of genetic risk factors for asthma in 2 Latino populations. The Journal of allergy and clinical immunology 2011;128:37-43.e12.

9. Wu H, Romieu I, Shi M, et al. Evaluation of candidate genes in a genome-wide association study of childhood asthma in Mexicans. J Allergy Clin Immunol 2010;125:321-7 e13.

10. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.

11. Choudhry S, Taub M, Mei R, et al. Genome-wide screen for asthma in Puerto Ricans: evidence for association with 5q23 region. Human Genetics 2008;123:455-68.

12. Cheng C-Y, Kao WHL, Patterson N, et al. Admixture mapping of 15,280 African Americans identifies obesity susceptibility loci on chromosomes 5 and X. PLoS genetics 2009;5:e1000490.

13. Freedman ML, Haiman CA, Patterson N, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proceedings of the National Academy of Sciences of the United States of America 2006;103:14068-73.

14. Bryc K, Velez C, Karafet T, et al. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proceedings of

the National Academy of Sciences of the United States of America 2010;107 Suppl 2:8954-61.

15. Choudhry S, Coyle NE, Tang H, et al. Population stratification confounds genetic association studies among Latinos. Human Genetics 2006;118:652-64.

16. Bryc K, Auton A, Nelson MR, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proceedings of the National Academy of Sciences of the United States of America 2010;107:786-91.

17. Gonzalez Burchard E, Borrell LN, Choudhry S, et al. Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. Am J Public Health 2005;95:2161-8.

18. Salari K, Choudhry S, Tang H, et al. Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. Genet Epidemiol 2005;29:76-86.

19. Kumar R, Seibold MA, Aldrich MC, et al. Genetic ancestry in lung-function predictions. New England Journal of Medicine 2010;363:321-30.

20. Corvol H, De Giacomo A, Eng C, et al. Genetic ancestry modifies pharmacogenetic gene-gene interaction for asthma. Pharmacogenet Genomics 2009;19:489-96.

21. Torgerson DG, Capurso D, Ampleford EJ, et al. Genome-wide ancestry association testing identifies a common European variant on 6q14.1 as a risk factor for asthma in African American subjects. J Allergy Clin Immunol 2012;130:622-9 e9.

22. Torgerson DG, Gignoux CR, Galanter JM, et al. Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. J Allergy Clin Immunol 2012;130:76-82 e12.

23. Cusanovich DA, Billstrand C, Zhou X, et al. The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. Hum Mol Genet 2012;21:2111-23.

24. Cui J, Stahl EA, Saevarsdottir S, et al. Genome-Wide Association Study and Gene Expression Analysis Identifies CD84 as a Predictor of Response to Etanercept Therapy in Rheumatoid Arthritis. PLoS Genet 2013;9:e1003394.

25. Torgerson DG, Ampleford EJ, Chiu GY, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. Nature Genetics 2011;43:887-92.

26. Galanter JM, Gignoux CR, Torgerson DG, et al. GWAS and admixture mapping identify asthma-associated loci in Latinos: The GALA II Study. submitted 2013.

27. Hoffmann TJ, Zhan Y, Kvale MN, et al. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics 2011;98:422-30.

28. Pasaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific ancestry in closely related populations. Bioinformatics 2009;25:i213-21.

29. Baran Y, Pasaniuc B, Sankararaman S, et al. Fast and accurate inference of local ancestry in Latino populations. Bioinformatics 2012;28:1359-67.

30. Shriner D, Adeyemo A, Rotimi CN. Joint ancestry and association testing in admixed individuals. PLoS Computational Biology 2011;7:e1002325.

31. Brisbin A, Bryc K, Byrnes J, et al. PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. Hum Biol 2012;84:343-64.

32. Chanock SJ. A twist on admixture mapping. Nat Genet 2011;43:178-9.

33. Hansen G, McIntire JJ, Yeung VP, et al. CD4(+) T helper cells engineered to produce latent TGF-beta1 reverse allergen-induced airway hyperreactivity and inflammation. The Journal of clinical investigation 2000;105:61-70.

34. Sagara H, Okada T, Okumura K, et al. Activation of TGF-beta/Smad2 signaling is associated with airway remodeling in asthma. J Allergy Clin Immunol 2002;110:249-54.

35. Burchard EG, Avila PC, Nazario S, et al. Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. Am J Respir Crit Care Med 2004;169:386-92.

36. Hough C, Radu M, Dore JJ. Tgf-beta induced Erk phosphorylation of smad linker region regulates smad signaling. PLoS ONE 2012;7:e42513.

37. Moore A, Enquobahrie DA, Sanchez SE, Ananth CV, Pacora PN, Williams MA. A genome-wide association study of variations in maternal cardiometabolic genes and risk of placental abruption. International journal of molecular epidemiology and genetics 2012;3:305-13.

38. Moffatt MF, Gut IG, Demenais F, et al. A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 2010;363:1211-21.

39. Paull K, Covar R, Jain N, Gelfand EW, Spahn JD. Do NHLBI lung function criteria apply to children? A cross-sectional evaluation of childhood asthma at National Jewish Medical and Research Center, 1999-2002. Pediatr Pulmonol 2005;39:311-7.

40. Rao DR, Gaffin JM, Baxi SN, Sheehan WJ, Hoffman EB, Phipatanakul W. The utility of forced expiratory flow between 25% and 75% of vital capacity in predicting childhood asthma morbidity and severity. The Journal of asthma : official journal of the Association for the Care of Asthma 2012;49:586-92.

41. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 2010;26:2336-7.

Table 1. Basic characteristics of studies used in meta-analysis.

Study Name	Genotyping platform	Study Type	Cases/Controls/Probands	Average Age of Onset (s.d.)
EVE Hispanic/Latino Discovery				
CARE	Affymetrix 6.0	Trios	42	1.4 (1.6)
CHS	Illumina 550K, 610K	Case/Control	606/792	6.8 (4.8)
GALA I Mexicans	Affymetrix 6.0	Case/Control	252/151	8.3 (7.7)
GALA I Puerto Ricans	Affymetrix 6.0	Case/Control	277/191	3.4 (4.8)
MCCAS	Illumina 550K	Trios	492	NA
GALA II Hispanic/Latino Replication				
GALA II Mexicans	Affymetrix Axiom LAT1	Case/Control	596/661	5.3 (3.7)
GALA II Puerto Ricans	Affymetrix Axiom LAT1	Case/Control	894/894	2.6 (2.9)
GALA II Mixed/Other	Affymetrix Axiom LAT1	Case/Control	403/326	4.4 (3.9)
EVE African American and African Caribbeans				
Barbados	Illumina 650Y	Pedigrees	382	8.2 (10.6)
CAG/CSGA/SARP	Illumina 1Mv1	Case/Control	541/451	9.8 (12.4)
GRAAD	Illumina 650Y	Case/Control	464/471	11.9 (13.2)
SAPPHIRE	Affymetrix 6.0	Case/Control	149/132	10.5 (11.6)
EVE European Americans				
CAG/CSGA/SARP	Illumina 1Mv1	Case/Control	742/381	13.1 (13.7)
CARE	Affymetrix 6.0	Trios	217	2.1 (2.4)
CAMP	Illumina 550K	Trios	385	3.1 (2.5)
CHS	Illumina 550K, 610K	Case/Control	643/959	7.0 (5.0)

Case-Control Phenotypes	**OR (95% CI)**	**p-value**
Asthma	8.05 (2.57-25.19)	**4.7×10^{-5}**
Island vs Mainland[a]	6.79 (1.99-23.19)	**0.0022**
Bronchodilator Response [b]	**b (95% CI)**	**p-value**
ΔFEF_{25-75}	-1.7 (-2.6 - -0.7)	**0.0013**
ΔFEV_1	-0.1 (-0.2 - 0.4)	0.40
ΔFVC[c]	-0.05 (-0.09 - -0.02)	**0.0068**
$\Delta PEFR$	-0.6 (-1.44 - 0.3)	0.22
Exacerbation Score	**OR (95% CI)**	**p-value**
Full Ordered Model[d]	0.37 (0.13 - 1.03)	**0.023**
More than 1	0.22 (0.05-0.90)	**0.022**

a. adjusting for case-control status, age, gender and admixture proportions

b. measured as (post - pre) / pre for all variables, all adjusted for Center, age, sex and height[2]

c. association not significant when D(FEF.25.75) included in the model

d. Levels 0-5, ordered logistic regression, p-value from likelihood ratio test

Figure 1.

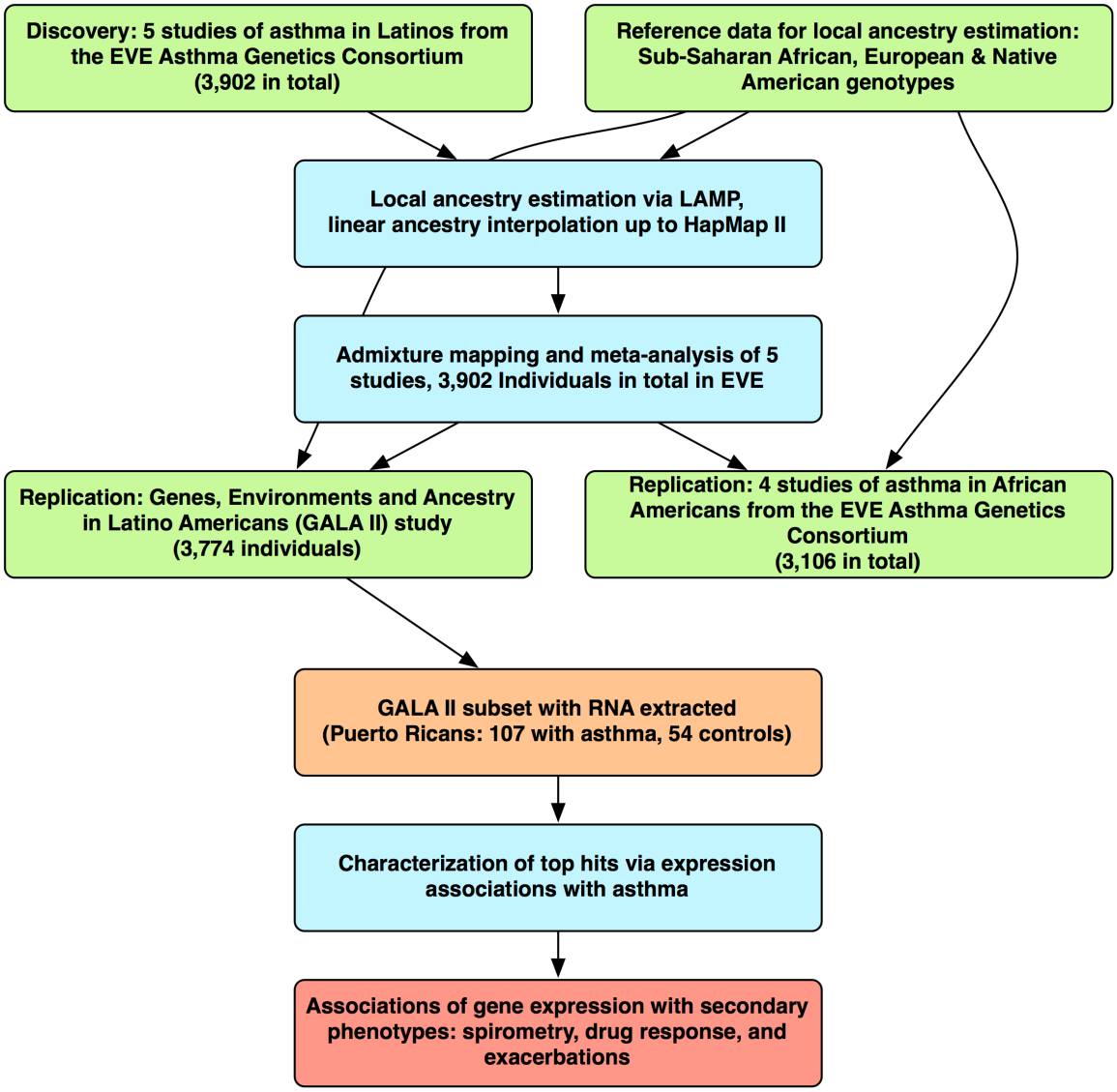

Discovery: 5 studies of asthma in Latinos from the EVE Asthma Genetics Consortium (3,902 in total)
Reference data for local ancestry estimation: Sub-Saharan African, European & Native American genotypes
Local ancestry estimation via LAMP, linear ancestry interpolation up to HapMap II
Admixture mapping and meta-analysis of 5 studies, 3,902 Individuals in total in EVE
Replication: Genes, Environments and Ancestry in Latino Americans (GALA II) study (3,774 individuals)
Replication: 4 studies of asthma in African Americans from the EVE Asthma Genetics Consortium (3,106 in total)
GALA II subset with RNA extracted (Puerto Ricans: 107 with asthma, 54 controls)
Characterization of top hits via expression associations with asthma
Associations of gene expression with secondary phenotypes: spirometry, drug response, and exacerbations

Figure 2.

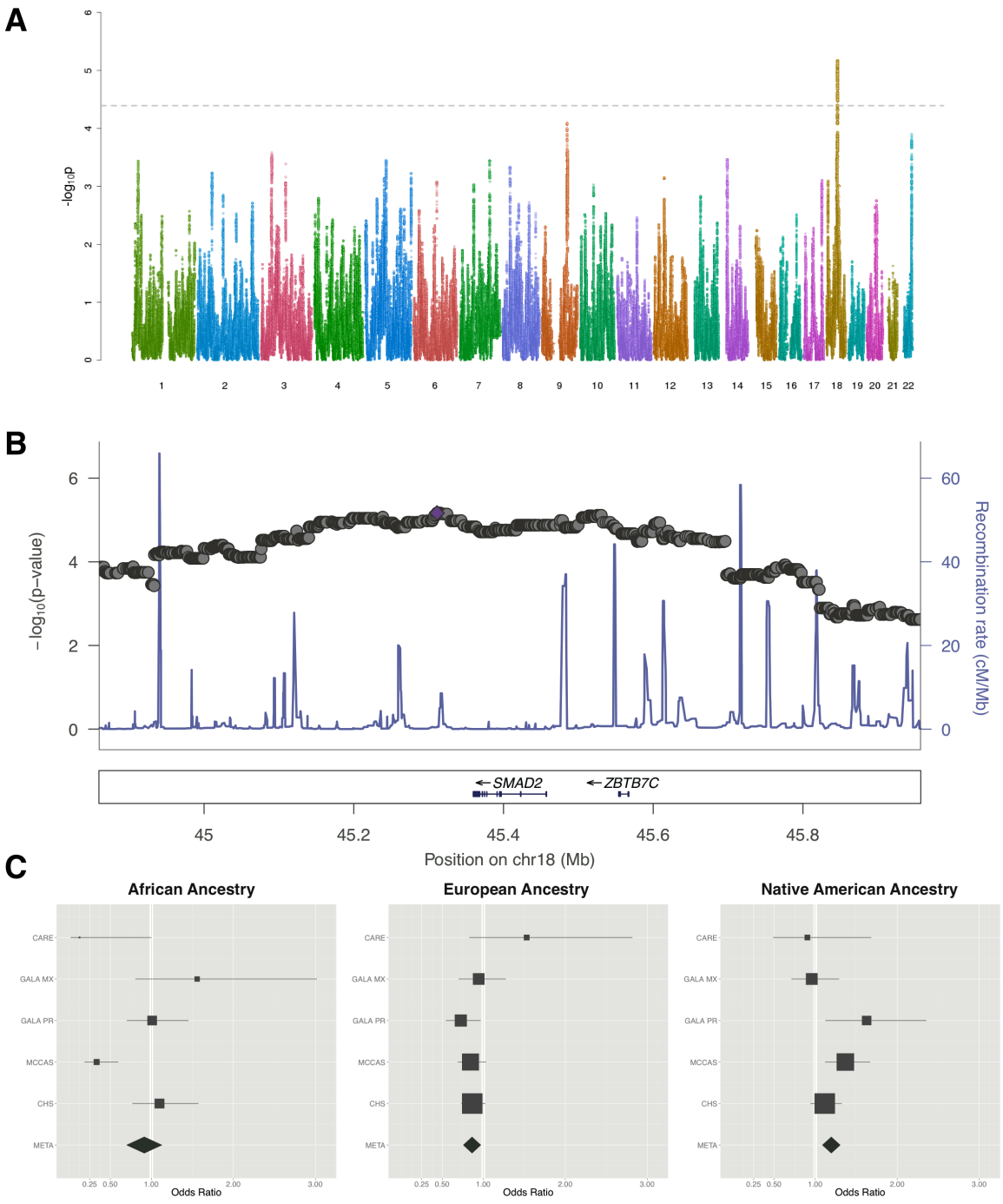
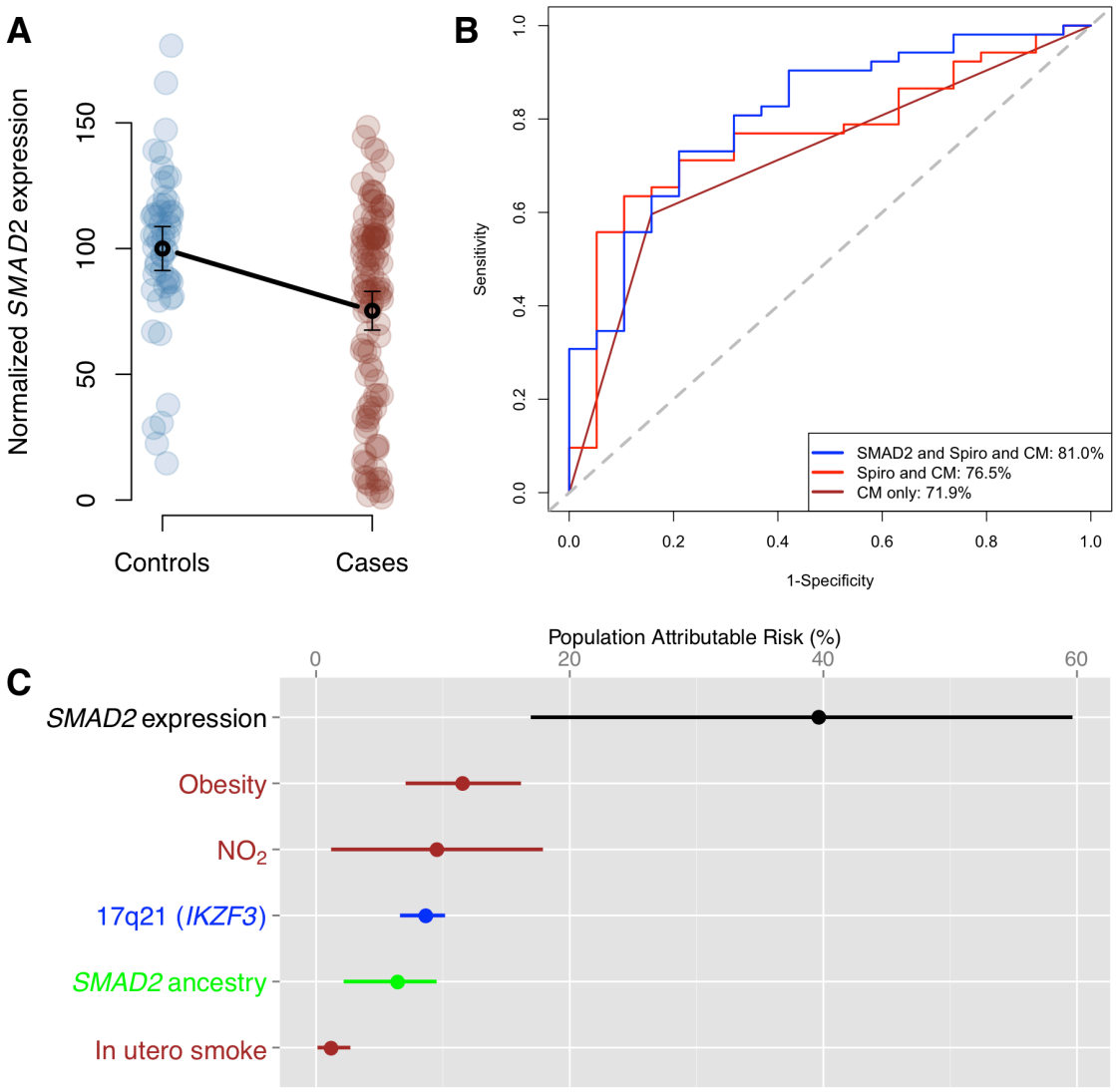

A
-log10p
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
B
– log10(p–value)
Recombination rate (cM/Mb)
SMAD2
ZBTB7C
45 45.2 45.4 45.6 45.8
Position on chr18 (Mb)
C
African Ancestry
European Ancestry
Native American Ancestry
CARE
GALA MX
GALA PR
MCCAS
CHS
META
0.25 0.50 1.00 2.00 3.00
Odds Ratio

Figure 3.


A
Normalized SMAD2 expression
0
50
100
150
Controls
Cases
B
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
SMAD2 and Spiro and CM: 81.0%
Spiro and CM: 76.5%
CM only: 71.9%
C
Population Attributable Risk (%)
0
20
40
60
SMAD2 expression
Obesity
NO2
17q21 (IKZF3)
SMAD2 ancestry
In utero smoke

Chapter 3. Supplementary Material

Admixture mapping meta-analysis identifies an ancestry-specific risk factor and potential biomarker for asthma

Supplementary Methods

Study Subjects

Discovery Population

Self-identified Latino and African American participants from nine independent study populations were included in the initial meta-analysis of genome-wide data (discovery population). The detailed methods for the EVE Asthma Consortium (www.eve.uchicago.edu) meta-analysis were previously published and described in detail.[1] EVE is a large, multi-ethnic assembly of asthma studies with existing genome-wide SNP genotypes from nine different institutions in the U.S. We used the autosomal genotypes passing the original quality control standards and incorporated in the original imputation. 3,902 Latinos and 3,106 African-Americans from EVE (Table 1) were included in the present analysis.

Replication Population

We tested our most significant associations from the discovery population in the Genes-environments & Admixture in Latino Americans (GALA II) Study population, a large, multi-

center case-control study of Latino children between the ages of 8-21 with and without asthma (Table 1). A total of 4,041 children (1,976 participants with asthma and 2,065 healthy controls) were recruited from five centers (Chicago, Bronx, Houston, San Francisco Bay Area, and Puerto Rico) using a combination of community and clinic-based recruitment. Participants with asthma self-reported a physician diagnosis of asthma and reported at least two symptoms (shortness of breath, wheezing, or cough not associated with upper respiratory illness) or chronic use of controller medication (inhaled corticosteroids, leukotriene modifying agents, theophylline or oral steroids) in the two years preceding recruitment. The mean (±SD) age of these subjects was 12.5 (3.3) years, 55.3% were boys. Eligible controls had no reported history of asthma, lung disease or chronic illness, and no reported symptoms of coughing, wheezing, or shortness of breath in the past two years. Controls were 1:1 frequency matched within each region by catchment area and age (within 1 year) and recruited from the same hospitals or community clinics as the cases. The mean (±SD) age of control subjects was 13.7 (3.5), 43.7% were boys. All participants who met criteria for enrollment completed in-person questionnaires related to their medical, asthma, allergic, social, environmental and demographic histories. To be eligible for participation, each participant or parent must have identified all four grandparents as Latino. All participants with asthma underwent spirometry to measure baseline lung function and maximal bronchodilator drug response to albuterol. Local institutional review boards approved the studies and all subjects and legal guardians provided written informed assent/consent.

Enrollment Criteria and Clinical Phenotyping

From July 2006 through June 2011, when genotyping began, a total of 4,041 children (1,976 participants with asthma and 2,065 healthy controls) were recruited from five centers (Chicago, Bronx, Houston, San Francisco Bay Area, and Puerto Rico) using a combination of community and clinic-based recruitment.

Asthma was defined as self-reported physician diagnosis, symptoms and medication use within the last 2 years. Eligible controls had no reported history of asthma, lung disease or chronic illness, and no reported symptoms of coughing, wheezing or shortness of breath in the past two years. Controls were 1:1 frequency matched within each region by catchment area and age (within 1 year) and recruited from the same hospitals or community clinics as the cases. Participants were excluded if they reported any of the following: (1) 10 or more pack-years of smoking; (2) any smoking within 1 year of recruitment date; (3) history of lung diseases other than asthma (cases) or chronic illness (cases and controls); or (4) pregnancy in the third trimester.

All participants who met criteria for enrollment completed in-person questionnaires related to their medical, asthma, allergic, social, environmental and demographic histories. Each participant or parent was also required to identify all four grandparents as Latino. Based on four-grandparent information, we partitioned the study into three major categories: Mexicans, Puerto Ricans, and Other (either individuals from other countries or of mixed Latino ancestry). In addition, all participants provided blood for genetic analysis. All participants with asthma underwent spirometry to measure baseline lung function and maximal bronchodilator drug responsiveness to albuterol. Local institutional review

boards approved the studies and all subjects and legal guardians provided written informed assent/consent.

Medication Use

Subjects were asked to list their asthma prescriptions and responses were grouped into five different treatment categories: (1) no medications, (2) rescue inhalers/short-acting beta agonists (SABA) only, (3) controller monotherapy, (4) combination therapy, and (5) oral corticosteroids (OCS). Monotherapy included subjects who were taking an inhaled corticosteroid (ICS), leukotriene receptor antagonist (LTRA), or theophylline to control their asthma. The combination therapy group included subjects using two or more controller medications with or without long-acting beta agonists (LABA).

Exacerbations

Information regarding self-reported asthma-related hospitalizations, emergency department visit, and oral steroid use over the 12 months prior to recruitment was collected through the administered questionnaire. With these data, a self-reported exacerbation score was assigned to each subject based on the American Thoracic Society and European Respiratory Society consensus statement from 2009.[2] One point was given for each report of hospitalization or emergency department visit in the last 12 months. For reported history of oral steroid use, one point was assigned for a report of any oral steroid use in the last 12 months and two points were assigned if the subject reported greater than two continuous weeks of oral steroid use over the 12 months prior to recruitment.

Pulmonary Function Testing

Subjects with asthma were instructed to withhold their bronchodilator medications for at least 8 hours before lung function testing. Spirometry was performed according to the American Thoracic Society standards (1995). Standard measurements of airway obstruction included Forced Expiratory Volume in one second (FEV_1), Forced Expiratory Flow between 25-75% of vital capacity (FEF_{25-75}) and Forced Vital Capacity (FVC). FEV_1/FVC, and FEF_{25-75} are all adjusted for age, sex and height[2] as covariates, as percent predicted equations are not available for Puerto Ricans.

Maximal bronchodilator drug response (BDR) was calculated as the percent change in baseline lung function (FEV_1 and FEF_{25-75}) before and after administering albuterol, with a 15-minute waiting period following each dose. Albuterol was administered using an extension tube connected to a standard metered dose inhaler. A total of six (if <16 years of age) to eight (if ≥16 years of age) total puffs of albuterol were administered.

Genotyping

Participants were genotyped at 818,154 SNPs on the Affymetrix Axiom® Genome-Wide LAT1 Array (World Array IV)[3], an array optimized for imputation-based association studies of Latinos. Details of individual and SNP quality control procedures are described in [4]. We employed standard quality control procedures as recommended by Affymetrix.

We removed single nucleotide polymorphisms (SNPs) with >5% missing data and failing platform specific cluster quality criteria (n=63,328), along with those out of Hardy-Weinberg equilibrium (n=1845; $p<10^{-6}$) within their respective populations (Puerto Rican,

Mexican, and other Latino). Subjects were filtered based on 97% call rates and gender discrepancies, and run through a pairwise identity by descent (IBD) scan to identify related individuals to remove. The total number of subjects passing QC was 3,774 (1,893 cases, 1,881 controls), and the total number of SNPs passing QC was 747,129.

Admixture Mapping Methods

Admixture mapping required estimating local ancestry at every SNP in each study. Local ancestry was estimated separately for each study via one of two LAMP algorithms: LAMP[5] for case-control studies, and a family-based algorithm available in LAMP-HAP[6] for trio studies that preserves transmitted/untransmitted haplotypes. For African-Americans we used a 2-way admixture model in LAMP in an unsupervised fashion as accuracy does not increase by adding ancestral populations[7]. For Latinos we ran LAMP assuming 3-way admixture using the CEU and YRI from HapMap[8], Indigenous Mexican individuals from HGDP[9] that were genotyped on the Illumina 650Y, and Pima and Maya individuals kindly provided by Drs. Mark Shriver and Abigail Bigham that were genotyped on the Affymetrix 6.0[10,11]. For each study we used the intersection of markers with available ancestral data: ancestral allele frequencies for LAMP, and ancestral haplotypes for LAMP-HAP, phased using Beagle[12].

Ancestry interpolation: As local ancestry estimation can be quite sensitive to genotyping errors we did not want to estimate local ancestry from imputed genotypes. Therefore, to create a consensus set of sites across different genotyping platforms we inferred ancestry

at untyped sites using linear (genomic distance) interpolation. In this framework, an untyped site is assigned the average of its neighboring genotyped sites, weighted by the genetic distance in bp from each. We used this framework to interpolate local ancestry for each study at 3,192,437 HapMap II SNPs. Given that admixture in the Americas is a recent phenomenon we observed few ancestry switches per chromosome per individual, and thus most of the untyped sites sat within blocks of contiguous ancestry.

Admixture mapping: The effect of ancestry at each locus was estimated in one of two ways. For single ancestry testing, we coded ancestry at each position as a biallelic state (e.g., African vs non-African). We then used logistic regression for case-control studies R,[13] and the transmission disequilibrium test (TDT)[14] for trio studies using custom Python[15] scripts. For the complex pedigrees in the Barbados study we used MQLS[16] for association testing, and estimated an odds ratio and standard errors from the ancestry counts in cases and controls (as estimated by MQLS).

Importantly, in Latinos we wanted to perform admixture mapping combining evidence across the three ancestries. To estimate the combined effect of all ancestries in Latinos, we used a 2-degrees of freedom (df) likelihood ratio test comparing regression models with and without local ancestry terms. We used logistic regression for case-control studies and Poisson regression models of counts of transmitted/untransmitted ancestry states for trio studies. All logistic regression models were adjusted for genomic ancestry as determined using ADMIXTURE[17] on the full autosomal data. Our likelihood ratio test for case-control analyses consisted of comparing the likelihoods of two nested generalized linear models:

$$full : \log\left(\frac{p}{1-p}\right) \sim \beta_0 + \beta_1 local_1 + \beta_2 local_2 + \beta_3 global_1 + \beta_4 global_2 + \varepsilon$$

$$restricted : \log\left(\frac{p}{1-p}\right) \sim \beta_0 + \beta_1 global_1 + \beta_2 global_2 + \varepsilon$$

The statistic $2\log\left(\frac{likelihood(full)}{likelihood(restricted)}\right)$ then follows a χ^2 distribution with 2df (equal to the number of additional terms in the full model).

Similarly, for the trio studies we compared likelihoods of two generalized linear models, although the framework is more complicated. We modeled counts of transmitted/untransmitted ancestry pairs at each locus using a mixture of Poisson regression terms for each ancestry (e.g., transmitted African/untransmitted Native American, six different combinations from three different ancestries in total). Each specific ancestry contributes to the counts of the observed transmitted/untransmitted pairs using an indicator function:

$$\Phi_{local} = \{1 \in transmitted, -1 \in untransmitted, 0 \in otherwise\}$$

Each Φ_{local} is then a six-term vector. To account for the investigation of multiple transmitted/untransmitted ancestries at every locus, we adjusted our analyses using a 6-term factor Ξ that stratified counts into corresponding pairs of transmitted/untransmitted ancestries. Then, the joint effects of incorporating multiple ancestry terms were modeled by comparing the likelihoods of two nested generalized linear models:

$$full : \log(counts) \sim \beta_0 + \beta_1 \Phi_{African} + \beta_2 \Phi_{European} + \beta_3 \Phi_{NativeAmerican} + \beta_4 \Xi + \varepsilon$$
$$restricted : \log(counts) \sim \beta_0 + \beta_1 \Xi + \varepsilon$$

The statistic $2\log\left(\frac{likelihood(full)}{likelihood(restricted)}\right)$ then follows a χ^2 distribution with 2df as the 3 ancestries are collinear (and one ancestry term is dropped). This approach is similar to a McNemar's test of symmetry, however our implementation saves a degree of freedom given the inherent correlation structure of the 3 ancestries. Scripts and functions were written in R and Python and are available upon request.

Multiple test correction: The traditional Bonferroni-based GWAS significance threshold of $5x10^{-8}$ is overly stringent given the increased ancestry linkage disequilibrium (LD) in admixture mapping studies. To determine a study-specific significance criterion we employed an empirical autoregression framework. We determined the "effective" number of tests on the genome by fitting an autoregressive model to the summary statistic data, where overall correlation patterns were determined by estimating the correlation sequentially along the chromosome. We implemented our autoregression using the *coda* package in R, similar to Shriner et al. [18] for both odds ratios and p-values. While similar, p-values were slightly more conservative, so we chose to use those. Our criterion for genome-wide significance was then 0.05 divided by the total number of effective tests across the genome.

Meta-analyses: Single ancestry tests were combined using fixed effects models in PLINK[19] to get combined estimates of significance, overall magnitude of effect and heterogeneity level. Sites with an I^2 value of heterogeneity > 50% were inspected to determine whether a random effects model was warranted (to incorporate between-study heterogeneity). For

the joint ancestry analyses, we combined p-values using both Fisher's method and combined Z-scores weighted by the square root of the number of cases as a proxy for the variance, consistent with the prior GWAS. [1] In the present study we report results from Fisher's method as the methods yielded highly concordant results.

Replication samples and genotyping: We tested our most significant associations in GALA II, a large, multi-center cross-sectional study of participants with and without asthma[20]. All individuals in GALA II self-identified as "Hispanic" or "Latino," and reported ethnicity information for all four grandparents. Based on four-grandparent information, we partitioned the study into three major categories: Mexicans, Puerto Ricans, and Other Latino (including individuals from countries other than Mexico or Puerto Rico, and those of mixed Latino ancestry). A total of 3,774 individuals passed quality control (QC) on genotypes obtained from the Affymetrix Axiom® Genome-Wide LAT1 Array (World Array IV, Affymetrix)[3], an array optimized for imputation-based association studies of Latinos. SNPs were filtered based on standard quality control procedures as recommended by Affymetrix. After merging genotypes with available CEU/YRI genotype data from HapMap and the 1000 Genomes[21], and Native Mexican individuals typed on the Axiom LAT1 array, we ended up with 568,037 SNPs for reference. Local ancestry was estimated on transmitted/untransmitted haplotypes in trios using LAMP-LD[6]. Admixture mapping was performed using the same methods as used in the discovery studies, with an additional correction for self-reported ethnicity (Puerto Rican, Mexican, or Other).

Imputation and genotype association: We imputed candidate regions using IMPUTE2[22] using the phase 1 1000 Genomes haplotypes, after phasing our data using SHAPEIT.[23] We used the same criteria for source genotypes as the admixture mapping and previous meta-analysis. Imputation was carried out using the default and recommended settings in IMPUTE2 for prephased data across a ~5Mb region around *SMAD2*. Imputed genotypes with information scores >0.3 were used for downstream analysis. We analyzed each study using a similar framework as described above (i.e., logistic regression, TDT, and pedigree association using EMMAX[24]). A meta-analysis was them performed using a fixed effects meta-analysis in PLINK, and a random effects meta-analysis using the R package *metafor.*[25]

Gene Expression: We measured the expression of *SMAD2*, *SMAD3,* and *ZBTB7C* using TaqMan® RT-PCR assays in a total of 107 cases and 54 controls selected from the Puerto Ricans in GALA II. Total RNA was isolated from PAXgene™ Blood RNA tubes, and RNA integrity was assessed with Aligent's BioAnalyzer. Samples with RNA integrity < 7 were excluded from further analysis. We normalized gene expression of each target gene to the housekeeping gene *GUS*. We transformed fluorescent values to estimate relative-fold expression as 2^(-deltaCT) for downstream analyses and investigated associations with linear, logistic, and ordered logistic regressions in R. We performed preliminary expression associations using the Wilcoxon rank sum tests.

Gene Expression Model Selection: Given that gene expression is continuous, we wanted to find a cutpoint that would best determine high-vs-low gene expression to evaluate its association with asthma. We estimated the maximum *a posteriori* value for a cutpoint of

gene expression by calculating Bayes Factors across the continuum of *SMAD2* gene expression levels. We used a logistic regression model including relevant covariates as a generative model for likelihoods, and then tested the hypothesis that expression affects the odds of disease, vs. the null hypothesis of no expression effect. In modeling both hypotheses, Bayes Factors provide an evidence-based rationale for determining the model that best fits the model of association. For each percentage point in our scale of normalized gene expression, we evaluated the odds ratio, confidence interval, and Bayes Factor. We chose the best-fitting model to differentiate high vs low expression as the cutpoint with the maximum Bayes Factor. Using the AIC or another likelihood-based statistic is expected to give analogous results.

Population Attributable Risk (PAR): GALA II includes a large number of genetic and environmental measures, allowing for the comparison of multiple types of risk factors. In this study we compared the PAR for *SMAD2* expression and ancestry to previously identified significant risk factors that were also identified in GALA II. For ease of comparison, we dichotomized all risk factors to estimate odds ratios, and converted these into risk ratios based on a disease prevalence of 20%. While this is a single point estimate, this represents a compromise between Mexican, Puerto Rican, and Other Latino prevalence estimates. Varying prevalence would be expected to slightly change the overall estimates, but less so the proportional differences. Prevalence of the risk factor was measured based on the observed values in cases and controls in GALA II and the prevalence numbers. Obesity was categorized based on BMI,[26] and NO_2 exposure was measured from monitoring towers and residential history from the first three years of life (and given a cutpoint at the

WHO level of acceptable exposure[27]). We also compared genotypes at the 17q21 locus, which represents the strongest and best replicated GWAS hit for asthma. The region of interest in 17q21 spans multiple genes, and the top genome-wide significant marker changes by study, but often includes *ORMDL3*, *GSDMA*, *GSDMB*, or as in Galanter et al.,[4] a marker in *IKZF3*. Plotted confidence intervals were derived from the confidence intervals in the odds ratios, holding other measurements constant.

Supplementary Results.

Single ancestry admixture mapping: In addition to the likelihood ratio test (see Online Methods), we performed a meta-analysis of single ancestry admixture mapping across Latino individuals and identified two genome-wide significant peaks for European ancestry (the ancestry with the most power across all Latino groups to identify associations) at 9q22 and 12p12 (see Supplementary Table 1). Both of these peaks failed to replicate in GALA II (lowest p-value = 0.17 and 0.21, respectively). P-values for single ancestries approached genome-wide significance at 18q12, particularly for Native American ancestry.

***In Silico* Fine Mapping:** We imputed the full set of Phase I 1000 Genomes within a 5Mb region centered on the 18q21 locus in all study populations separately (12,870 total individuals: 7,606 Latino American, 3,102 African American, 2,088 European American) using 1000 Genomes haplotypes[28] with IMPUTE2.[29] There were no genome-wide

significant SNP associations with asthma within the region, nor any locus-wide significant variants in the EVE meta-analysis. However, rs59002988, a SNP 40Kb upstream of *SMAD2*, met locus-wide significance in our replication study (GALA II, OR 1.67, 95% CI 1.32-2.1, $p=1x10^{-5}$). The T allele of this SNP is rare in Europeans (~2%), elevated in eastern Asians (16%), and common (>10%) in GALA II individuals who are homozygous for Native American ancestry at this SNP. The variant appears to be more common on Native American haplotypes of Puerto Ricans (minor allele frequency=15%, Supplementary Figure 7), and to have an increased effect size (OR 2.2, 95% CI 1.46-3.29, $p=2x10^{-4}$).

Table S1. Ancestry associations at 18q21, centered on *SMAD2*. Summary characteristics of admixture mapping findings at the top hit in the chr18q21 region. Meta-analysis of discovery and replication panels was performed with fixed effects assumptions for effect size estimates, and Fisher's method was used for the overall likelihood ratio test meta-analysis.

18q21	**EVE**	**GALA II (average)**	**Combined**
African *p*	0.42	0.16	0.27
African OR	0.91 (0.73-1.14)	1.05 (0.98-1.13)	1.04 (0.97-1.11)
European *p*	8.35×10^{-3}	5.83×10^{-3}	1.36×10^{-4}
European OR	0.86 (0.77-0.96)	0.87 (0.78-0.96)	0.86 (0.80-0.93)
Native American *p*	1.63×10^{-3}	6.26×10^{-3}	9.15×10^{-5}
Native American OR	1.20 (1.07-1.34)	1.09 (1.02-1.16)	1.11 (1.05-1.17)
Overall *p*	**6.80×10^{-6}**	**0.017 (min 0.0053)**	**2.6×10^{-7}**

Table S2. Suggestive regions in the EVE Latino admixture mapping study, as defined by a minimum p-value < 0.001, out to 0.01 on each side. Joint refers to the 2-df likelihood ratio test for all ancestries. If a region was identified in multiple ancestry scans, the ancestry with the smallest p-value was used and the other ancestry information is given in parentheses. Coordinates are in hg18.

Associated Ancestry	Chr	Start	End	minimum p	Genes in Peak
Native	1	20716967	21885179	3.48x10-4	*USP48, HP1BP3, SH2D5, CDA, EIF4G3, RAP1GAP, NBPF3, KIF17, ECE1, ALPL, PINK1, LOC100506801, DDOST, FAM43B*
Joint	1	23046750	24398722	3.67x10-4	*None*
Native (European)	1	218900613	219027538	8.97x10-4	*MARC2, MARK1, C1orf115, MARC1*
European	1	222843773	223116557	9.93x10-4	*CNIH3*
European	1	226684711	227897382	6.67x10-4	*DUSP5P, ACTA1, NUP133, RAB4A, TAF5L, RNF187, CCSAP, RHOU, ABCB10, HIST3H2BB, URB2, MIR4666A, HIST3H2A, SPHAR*
Joint	2	57596338	58186261	5.85x10-4	*None*
Native	3	4015247	4556887	5.48x10-4	*ITPR1, SUMF1, SETMAR*
Joint	3	40060302	41086194	2.62x10-4	*None*
Joint	3	95606242	96292801	4.09x10-4	*None*
European	4	23534331	23968123	3.19x10-4	*None*
African	4	48077300	52634909	4.05x10-4	*OCIAD2, OCIAD1, SLC10A4, ZAR1, SPATA18, SGCB, CWH43, SLAIN2, FRYL, DCUN1D4, LRRC66*
African	5	55678536	57344829	1.60x10-4	*SETD9, GPBP1, MAP3K1, ACTBL2, MIER3*
Native	5	60547184	61124131	1.94x10-4	*ZSWIM6, C5orf64*

Native	5	62620986	63502012	6.70x10-4	*HTR1A, RNF180*
Joint	5	73710967	74034961	7.35x10-4	*None*
Joint (Native)	5	80682142	81813860	3.57x10-4	*ATG10, ACOT12, SSBP2, RPS23, ATP6AP1L (CKMT2, LOC100131067, RASGRF2, ZCCHC9)*
African (Joint)	5	178998248	180340825	4.43x10-4	*FLT4, CANX, C5orf60, MAML1, OR2Y1, LOC729678, RNF130, LTC4S, LOC100859930, MAPK9, CNOT6, RASGEF1C, MGAT1, MIR1229, C5orf45, CBY3, SQSTM1, BTNL8, SCGB3A1, ZFP62, TBC1D9B, MGAT4B, MIR340, GFPT2*
Joint	6	90202813	90311362	8.35x10-4	*None*
Joint	7	55129505	55621885	9.39x10-4	*None*
Joint	7	117211513	117400523	3.64x10-4	*None*
Joint	7	118344280	118830576	7.12x10-4	*None*
Joint	8	28677088	29253448	4.64x10-4	*None*
African	8	56410052	56871124	7.63x10-4	*XKR4, TGS1, SBF1P1, TMEM68*
Native	8	75215069	75733856	4.27x10-4	*FLJ39080, MIR5681A, MIR5681B, JPH1, GDAP1*
European	9	88694862	90508804	6.04x10-4	*LOC392364, LOC286238, CTSL1P8, LOC100506834, FAM75C2, LOC494127, FAM75C1, C9orf170, LOC440173, CTSL1, CTSL3, CDK20, GAS1, NXNL2, SPIN1, FAM75E1, DAPK1*

European (African, Joint)	9	96446960	100419073	1.67x10-6	*LOC100507346, LOC100499484, TMOD1, NCBP1, FOXE1, MIR27B, HABP4, MIR23B, C9orf3, MIR2278, LOC158434, LOC158435, TSTD2, GABBR2, TRIM14, HIATL2, ZNF782, FANCC, LOC340508, ZNF510, HSD17B3, LOC286359, C9orf174, LINC00092, XPA, TDRD7, ZNF367, LOC441454, HEMGN, AAED1, LOC441455, CORO2A, NANS, ANP32B, MIR24, LOC100132781, FAM22G, CDC14B, MIR3074, C9orf156, ERCC6L2, LINC00476, CTSL2, SLC35D2, PTCH1, TBC1D2 (ANKS6, COL15A1, FBP1, FBP2, GALNT12)*
African	9	108352743	108738041	9.81x10-4	*ZNF462*
Joint	10	51824523	52200679	9.38x10-4	*None*
African	10	83972038	84362276	3.10x10-4	*NRG3*
Native	10	100396273	101144430	7.93x10-4	*CNNM1, HPSE2*
European (Native)	10	102072780	102969197	7.40x10-4	*MIR608, PDZD7, KAZALD1, NDUFB8, SCD, PAX2, C10orf2, WNT8B, TLX1, MRPL43, LINC00263, SFXN3, PKD2L1, HIF1AN, TLX1NB, LZTS2, SEC31B, FAM178A, SEMA4G (ACTR1A, ARL3, BTRC, C10orf76, C10orf95, CUEDC2, DPCD, ELOVL3, FBXL15, FBXW4, FGF8, FLJ41350, GBF1, HPS6, KCNIP2, LBX1, LDB1, LOC100289509, LOC100505761,*

					MGEA5, MIR146B, MIR3158, NFKB2, NOLC1, NPM3, PITX3, POLL, PPRC1, PSD, SUFU, TMEM180, TRIM8)
European (Native)	12	21952123	25112841	1.67x10-5	*MIR920, C12orf77, BCAT1, ST8SIA1, CMAS, ABCC9, ETNK1, LINC00477, LRMP, KIAA0528, SOX5*
Joint	12	40400256	41129450	7.16x10-4	*None*
African	13	67676298	68122400	7.40x10-4	*None*
Joint	14	22233894	23109399	3.47x10-4	*None*
Joint	14	23958088	24164696	8.49x10-4	*None*
European (Native)	16	15398985	16537826	3.98x10-4	*MYH11, KIAA0430, MPV17L, NOMO3, NDE1, FOPNL, MIR484, C16orf45, PKD1P1, ABCC6, MIR3179, ABCC1, MIR3180*
Joint	17	69510566	71397400	7.96x10-4	*None*
Joint	18	6230094	6746661	7.96x10-4	*None*
Native	18	38783085	39434024	4.92x10-4	*SYT4, RIT2*
Joint (European, Native)[1]	18	40530823	44316091	1.71x10-4	*C18orf25, SLC14A2, SLC14A1, ATP5A1, HAUS1, LOXHD1, RNF165, MIR4319, SIGLEC15, SETBP1, PSTPIP2, EPG5, HDHD2, SMAD2, TCEB3C, TCEB3B, ZBTB7C, TCEB3CL, IER3IP1, KATNAL2, ST8SIA5, LOC100506888, PIAS2*
Joint	18	49677502	49903778	9.83x10-4	*None*
Joint	22	46376748	46640808	1.26x10-4	*None*

[1] Peak had a gap in joint analysis from chr18:42207149-42218476, where the p-values were above 0.01 to a maximum of 0.0128. However given how close the two peaks were, we combined them into a single entry in the table.

Figure S1. Omnibus tests of admixture mapping in both the Latino (left, 2-df) and African-American (right, 1-df) studies in the discovery sample in EVE at the top hit on 18q21. The meta-analysis p-value in Latinos is 6.8×10^{-6} while in African-Americans it is 0.7.

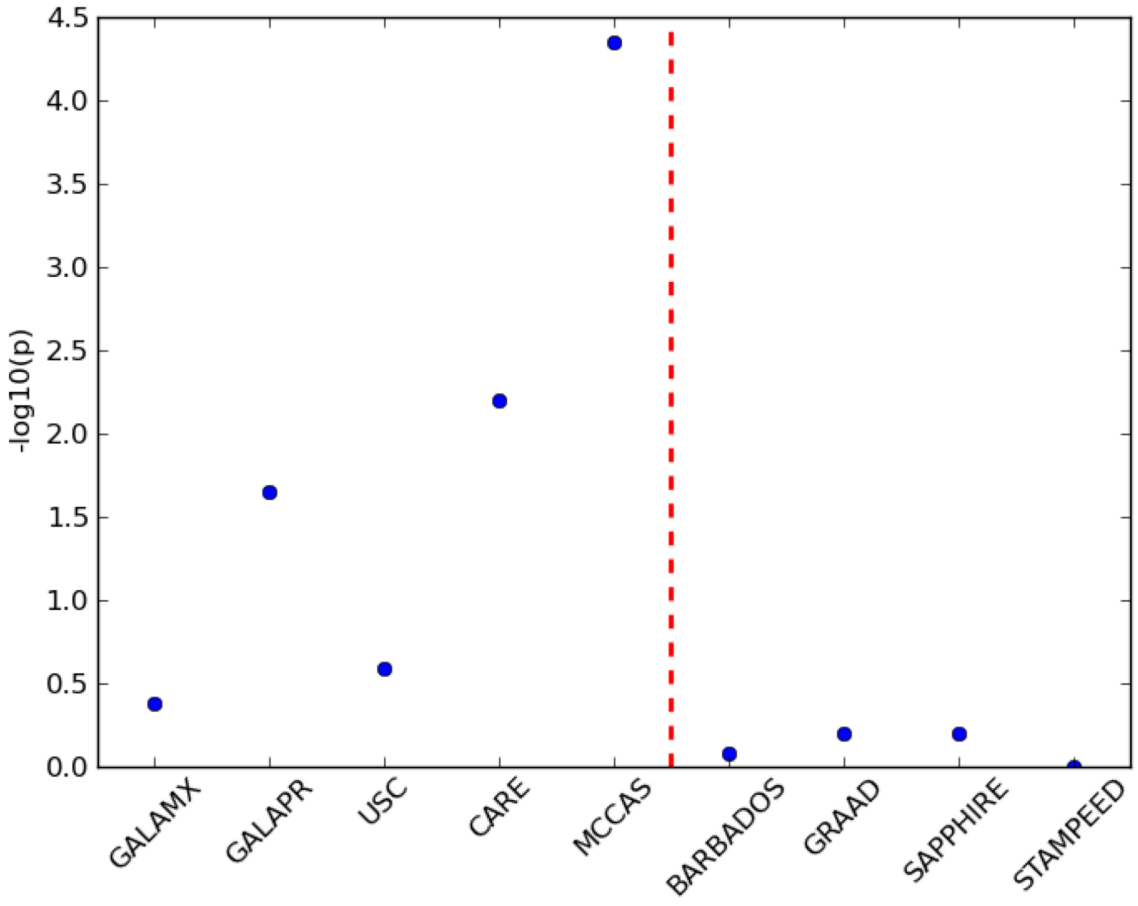


Figure S2 (next page). Manhattan plots of single-ancestry admixture mapping for African, European, and Native American ancestry, respectively. Peaks on chr9p31 and chr12p12 show up as genome-wide significant although they do not replicate in GALA II. Peaks encompassing *SMAD2* approach genome-wide significance in both the European- and Native American-specific admixture mapping meta-analyses.

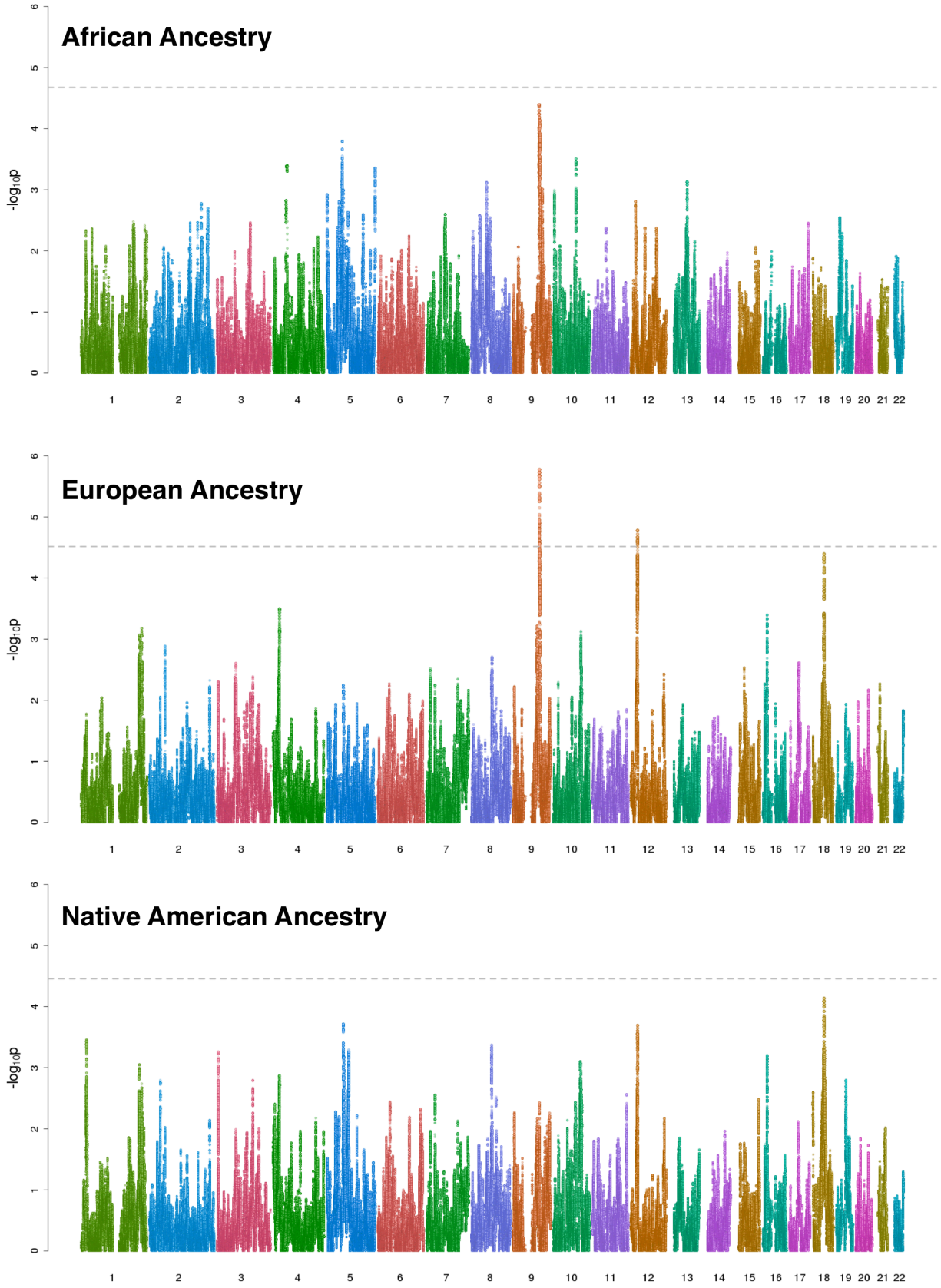

African Ancestry
European Ancestry
Native American Ancestry
-log10p
0
1
2
3
4
5
6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

Figure S2

Figure S3. Qqplots of allelic association in the *SMAD2* region following imputation using 1000 Genomes Phase I haplotypes. Top panel shows the qqplot for GALA II only, bottom left shows GALA II and EVE Latinos, and bottom right, shows the QQplot for all of EVE (including the African American and European American studies). Associations were performed using logistic regression, TDT, or mixed model analysis (EMMAX) depending on study design. Association testing was done ignoring the effects of local ancestry.

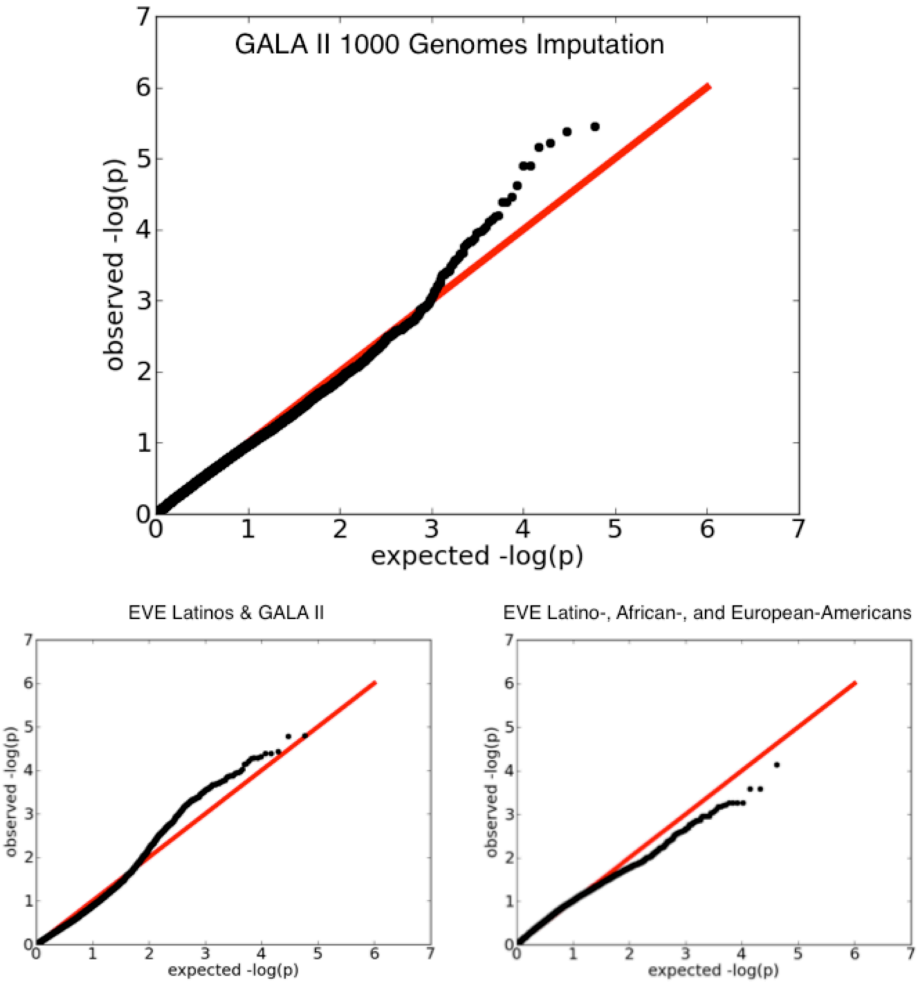


Figure S4. Comparisons of coefficients, standard errors, and p-values (respectively) from imputed allelic associations in GALA II with and without local ancestry. In each comparison, the estimates for variants including both local and global ancestry are plotted on the x-axis and with only global ancestry on the y-axis. The most significant sites are ranked similarly in either model.

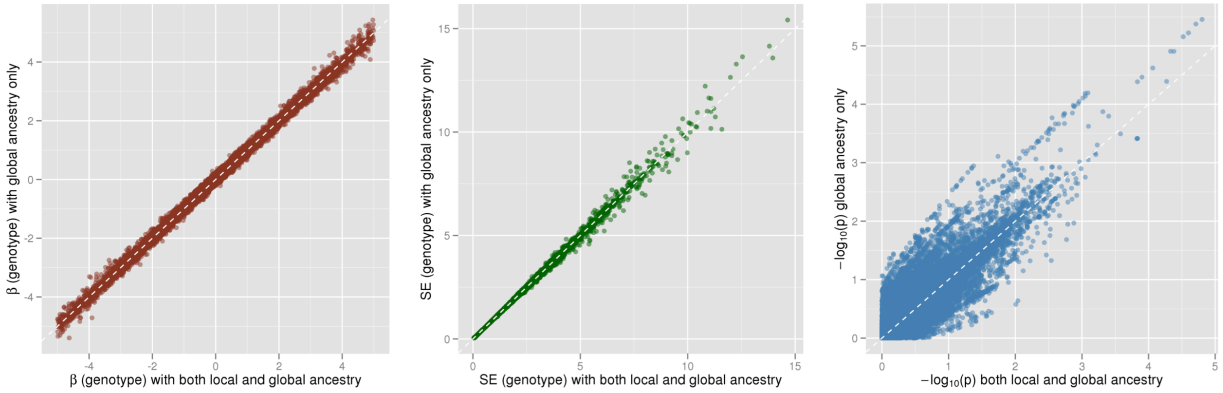


Figure S5. Locuszoom plot of GALA II fine mapping using 1000 Genomes haplotypes for imputation in the neighborhood of the 18q21 admixture mapping peak. Of all the study types, only GALA II provided a region-wide significant allelic association at rs59002988.

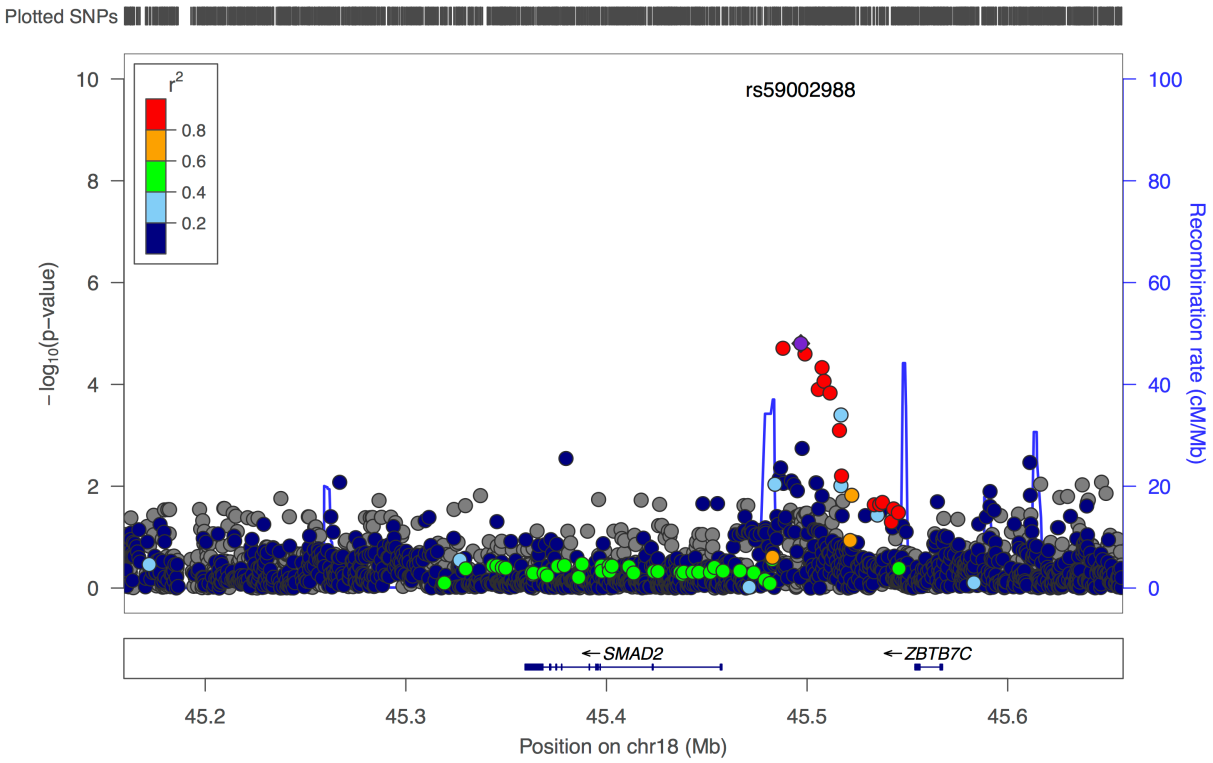


Figure S6. Minor allele frequencies of our fine mapping top hit, rs59002988, as estimated from all individuals in GALA II with homozygous ancestry at the locus. Lines represent 95% confidence intervals. Consistent with observations that the SNP has elevated allele frequencies in eastern Asians in 1000 Genomes, we observe significantly higher allele frequencies in the Native American haplotypes than in the other two.

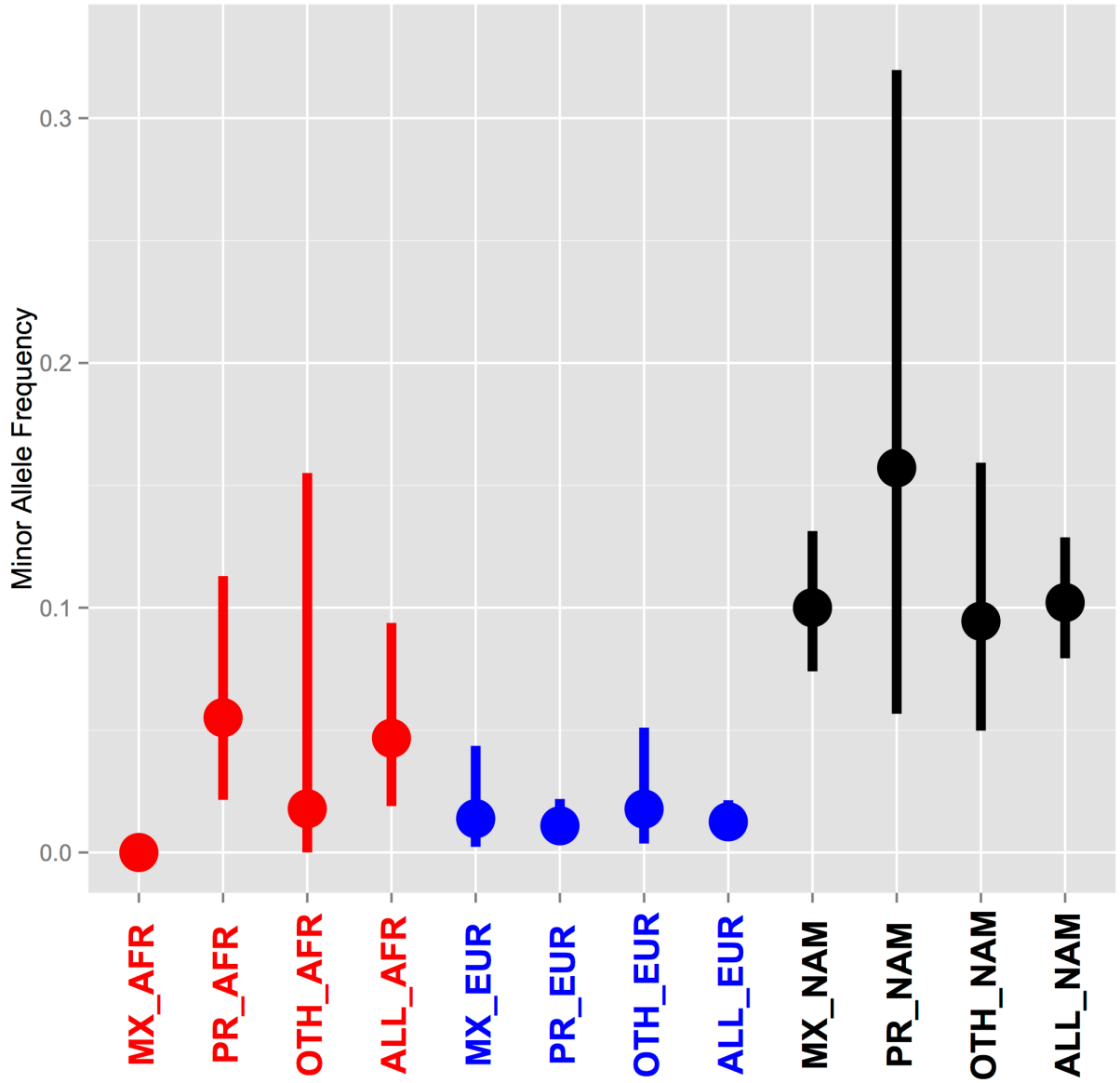


Figure S7. qqplots from the original EVE allelic associations for the chr18 region around *SMAD2* (+/- 1,000 SNPs, although may be missing) split by population. Unlike the admixture results, there is limited evidence from the genotypes themselves for variants associated with asthma, aside from some moderate inflation in African-Americans. No allelic association in any of the populations has a p-value lower than 10^{-4}.

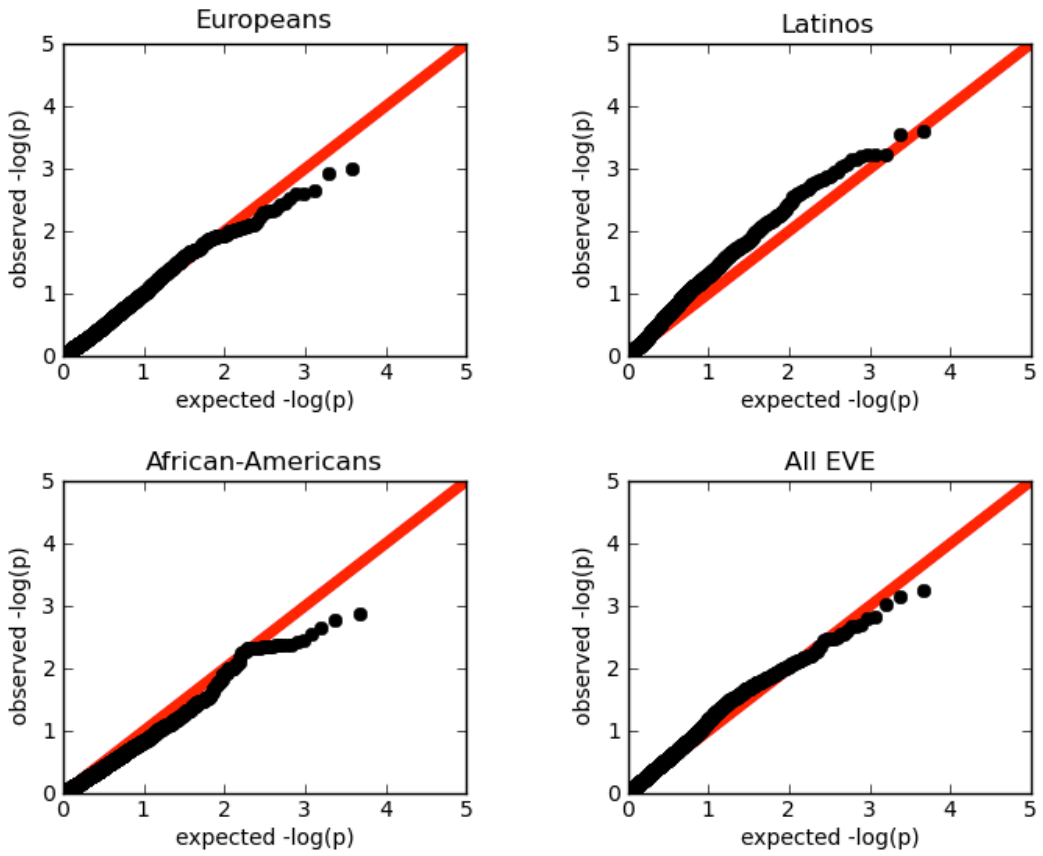


Figure S8. Distribution of self-reported exacerbation scores calculated in the individuals in GALA II with measured *SMAD2* expression.

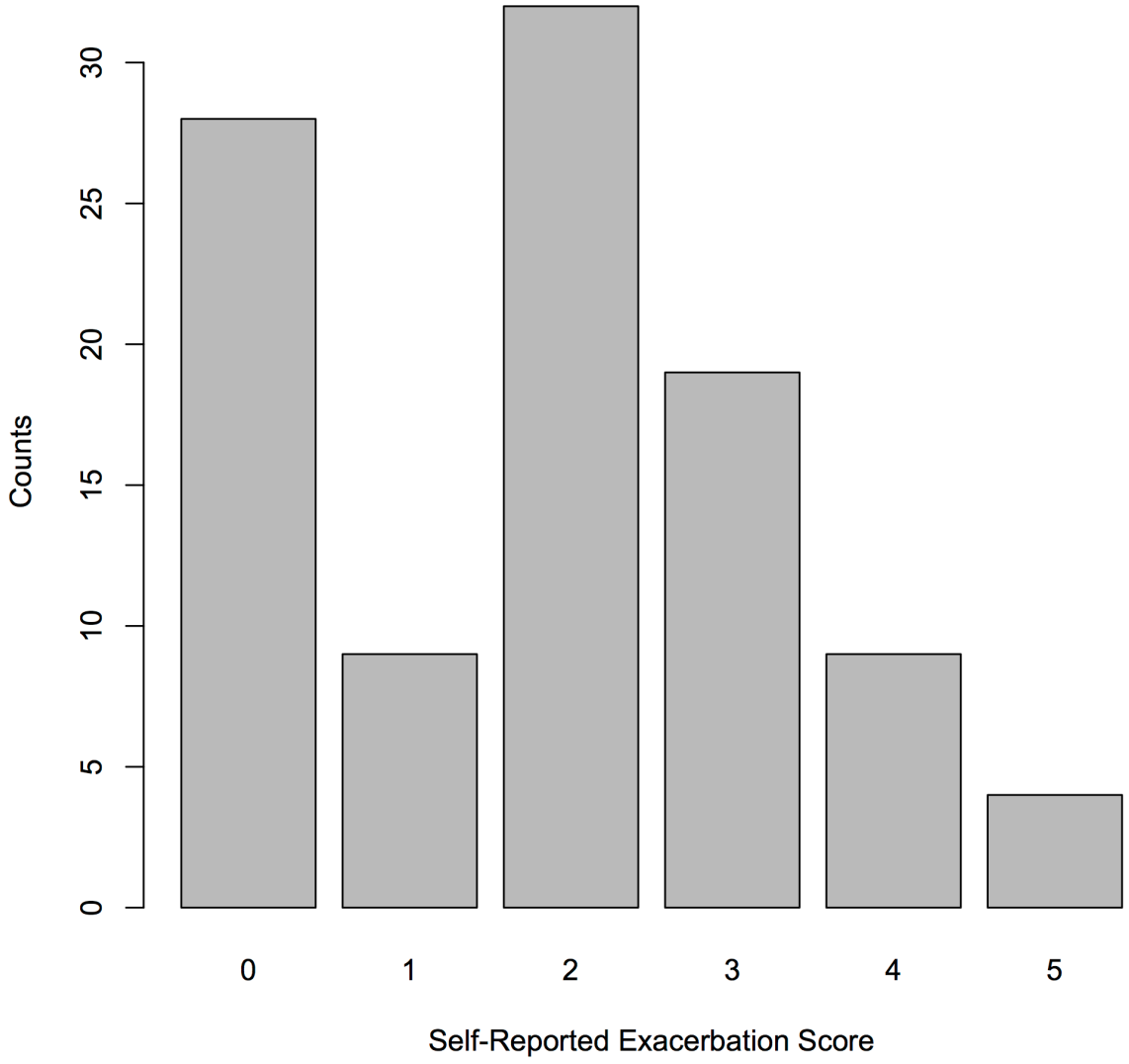
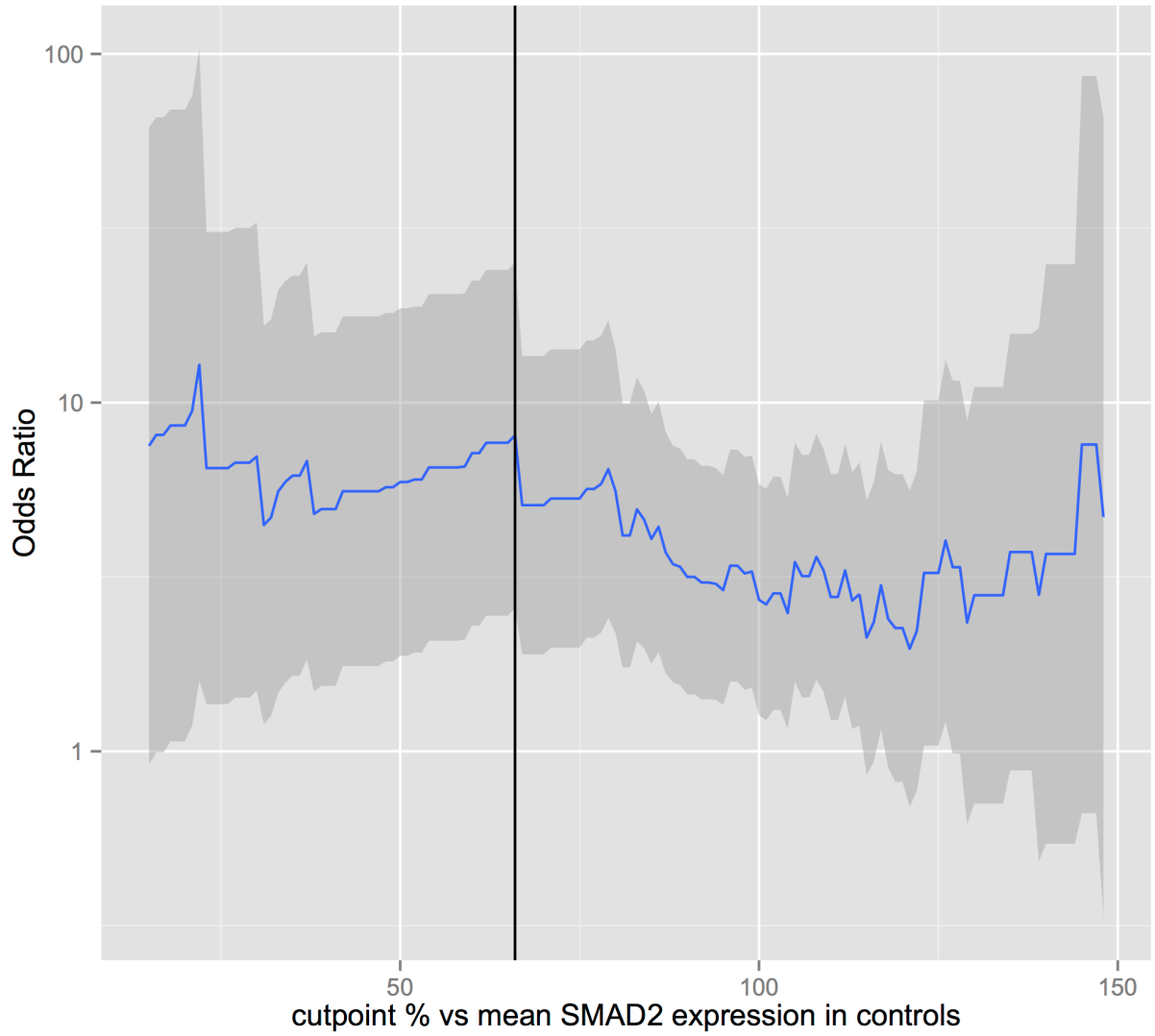


Figure S9. Evaluation of Odds Ratios and cutpoints along the spectrum of possible values for *SMAD2* expression. Best-fit cutpoint as discussed in the main text corresponds to the black line.



Sources

1. Torgerson DG, Ampleford EJ, Chiu GY, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. Nature Genetics 2011;43:887-92.

2. Reddel HK, Taylor DR, Bateman ED, et al. An official American Thoracic Society/European Respiratory Society statement: asthma control and exacerbations: standardizing endpoints for clinical asthma trials and clinical practice. Am J Respir Crit Care Med 2009;180:59-99.

3. Chanock SJ. A twist on admixture mapping. Nat Genet 2011;43:178-9.

4. Galanter JM, Gignoux CR, Torgerson DG, et al. GWAS and admixture mapping identify asthma-associated loci in Latinos: The GALA II Study. submitted 2013.

5. Pasaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific ancestry in closely related populations. Bioinformatics (Oxford, England) 2009.

6. Baran Y, Paşaniuc B, Sankararaman S, et al. Fast and accurate inference of local ancestry in Latino populations. Bioinformatics (Oxford, England) 2012.

7. Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. American journal of human genetics 2008;82:290-303.

8. Consortium IH, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851-61.

9. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science 2008;319:1100-4.

10. Bigham A, Bauchet M, Pinto D, et al. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet 2010;6.

11. Bigham A, Bauchet M, Pinto D, et al. Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. PLoS genetics 2010;6:e1001116.

12. Browning S, Browning B. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies …. The American Journal of Human Genetics 2007.

13. R Core Team. R: A language and environment for statistical computing. In. R Foundation for Statistical Computing, Vienna, Austria; 2012.

14. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506-16.

15. Python Software Foundation. http://www.python.org/. In; 2012.

16. Thornton T, McPeek MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet 2007;81:321-37.

17. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19:1655-64.

18. Shriner D, Adeyemo A, Rotimi CN. Joint ancestry and association testing in admixed individuals. PLoS Computational Biology 2011;7:e1002325.

19. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75.

20. Oh SS, Tcheurekdjian H, Roth LA, et al. Effect of secondhand smoke on asthma control among black and Latino children. J Allergy Clin Immunol 2012;129:1478-83 e7.

21. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061-73.

22. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009;5:e1000529.

23. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods 2012;9:179-81.

24. Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 2010;42:348-54.

25. Schwartz AG, Wenzlaff AS, Bock CH, et al. Admixture mapping of lung cancer in 1812 African-Americans. Carcinogenesis 2011;32:312-7.

26. Borrell LN, Nguyen EA, Roth LA, et al. Childhood Obesity and Asthma Control in the GALA II and SAGE II Studies. Am J Respir Crit Care Med 2013;187:697-702.

27. Nishimura KK, Galanter JM, Roth LA, et al. Early Life Air Pollution and Asthma Risk in Minority Children: The GALA II & SAGE II Studies. . submitted 2013.

28. Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56-65.

29. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 (Bethesda, Md) 2011;1:457-70.

Chapter 4: The Genetic History and Structure of Mexican Populations

(joint work with Dr. Andres Moreno-Estrada, Stanford University)

Abstract:

Mexico harbors one of the most culturally and ethnically diverse populations of the Americas, yet fine-scale patterns of genome-wide variation remain understudied. Here we present genomic data for 511 individuals from 20 indigenous populations, combined with 420 mestizo individuals from 11 cosmopolitan populations throughout Mexico. We found three major genetic components geographically restricted to Northern, Central/Southern, and Southeastern populations, with gene flow from Mayans in the Yucatan peninsula to Central Mexico, likely through a coastal route along the gulf. We implemented a novel ancestry-specific PCA analysis (ASPCA) to investigate sub-continental ancestry for genomic segments of inferred European and Native American origin derived from admixed genomes. We identified a hidden correlation with geography revealed in the indigenous segments of admixed Mexicans resembling a map of Mexico. We evaluate the biomedical implications of this hidden population structure on measures of lung function in Mexican and Mexican American children with asthma. We identified a significant association between ASPCA scores and lung function. Understanding fine-scale ancestry patterns is critical for the next generation of medical and population genetic studies.

One Sentence Summary:

Indigenous and cosmopolitan Mexican populations are highly structured and genomic patterns of variation mirror geography within Mexico, informing future medical genomic studies.

Main Text:

Understanding local patterns of human population structure is crucial to evaluate the geographic stratification of genetic variants. Recent studies have shown that the majority of human genomic variable sites are rare and exhibit little sharing among diverged populations (*1*). Moreover, because rare variants tend to be enriched for potentially functional mutations, their characterization is likely to lead to novel disease associations affecting local populations. Previous genome-wide surveys have provided insight into global (*2, 3*) and continental patterns of population structure across Africa (*4*), Europe (*5*), and the Americas (*6*), among others. However, regional and local genomic surveys are needed as a first step towards the discovery of geographically restricted variation, especially in those regions where populations are likely to be highly structured (*7*). In the Americas, the founding population size was likely very small (perhaps as few as several hundred people (*8*)) and, therefore, indigenous Americans show very low genetic diversity within groups (the lowest of any continental population) yet high divergence among groups (*9*). As a result, present day indigenous populations (and individuals with some indigenous ancestry) may harbor local private alleles rare or absent elsewhere, including functional and medically relevant variants (*10, 11*).

Here we report local patterns of variation for 511 Native Mexican individuals from 20 indigenous groups covering most geographic regions across Mexico based on nearly 1 million genome-wide autosomal SNPs. By combining with genotype data from 500 additional mestizo individuals sampled in cosmopolitan areas of 11 different Mexican states as well as Mexican Americans, we evaluate the impact of sub-continental ancestry into the admixed genomes of cosmopolitan populations within Mexico and US-based Mexican communities. We also demonstrate the biomedical implications of this fine-scale geographic structure by identifying an

association between values of sub-continental ancestry and estimates of lung function in 456 Mexican child-parent trios from the Mexico City Childhood Asthma Study (MCCAS)(*12*) and see consistent effects in an independent study of 219 child-parent trios from Mexico City and the San Francisco Bay Area, which were part of the Genetics of Asthma in Latino Americans (GALA I) Study (*13*).

Native Mexican Diversity

Recent continent-wide surveys of Native American genetic diversity have described a genetic continuity from Mesoamerica southwards (*6, 14, 15*), pointing to present day Mexico as a geographic area of transition where a major breakpoint of diversity likely took place during the settlement of the Americas. Native Mexicans show closer genetic distances from the ancestral population of indigenous Americans and larger effective population sizes compared to South American natives (*14*), suggesting that they hold one of the major sources of diversity at a continental scale. Since the pioneering work by Lisker and others using classical markers (*16, 17*), significant efforts have been made to characterize native Mexican diversity, mostly analyzing either single-locus markers of uniparental transmission(*15, 18*), or limited autosomal loci (*19, 20*). By increasing both marker density across the genome and population sampling, we are able to get a much finer resolution of population relationships across indigenous Mexican groups.

We used principal components analysis (PCA) to summarize the major axes of genetic variation in Mexicans after removing individuals with >10% of European admixture. As expected, PC1 and PC2 separate Africans and Europeans from Native Mexicans, but 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 3 times as much as the variation accounted by the north-south axis of differentiation within Europe (0.30%, according to (*5*)). The northernmost (Seri) and southernmost (Lacandon) populations define the extremes of the distribution within sampled Native Mexicans. Higher PCs show well-defined population clusters, indicating high levels of divergence between groups (Fig. S1).

An important feature of Native American population history is the strong bottleneck associated with the peopling of the continent, followed by population expansions. To evaluate whether this translates into different signatures among contemporary Native Mexicans, we compared observed cumulative runs of homozygosity (cROH) along chromosome 1 against simulated data using a rejection algorithm framework in REJECTOR (*21*) (see Methods), allowing us to estimate effective population sizes during bottleneck and current Ne (Fig. S2 and S3). For instance, we estimate that as few as 71 individuals accounted for the deme size of the Seri population during the bottleneck, while its current Ne is about 1200 individuals. The Seri constitute one of the most historically isolated groups in present day Mexico. In contrast, larger ethnic groups, such as the Maya, have expanded from a couple of hundred to more than 3,500 individuals (Fig. 1B and S2). Interestingly, the estimated Ne during the bottleneck is comparable across all studied populations and rather low: 178 on average, consistent with previous estimates on the number of founders of the Americas (*8*).

To measure population differentiation among extant groups we computed overall pairwise F_{ST} combining all autosomal sites (Fig. 1C). The highest value was observed between Seri and Lacandon (0.14), followed by Tojolabal (0.12) and Triqui (0.10). Both Seri and Lacandon also showed elevated F_{ST} values across all other populations, while lowest F_{ST} values were observed among groups from central Mexico and within the Yucatan peninsula (Fig. 1C). To evaluate the

impact of population isolation in genetic similarity, we measured the total length of segments inferred to be identical by descent (IBD) among all possible pairs of individuals using GERMLINE (*22*) with a minimum threshold of 5cM (see Methods). We visualized both between- and within-population connections binned into nine levels of relatedness (Figure S4). Figure 1D shows the approximate location of sampled populations and their connections among individuals sharing segments of total IBD above 20cM (corresponding to the genomic equivalent of 3rd cousins or closer relatives). We observed high within-population IBD levels compared to between-populations, indicating that after splitting, indigenous populations have largely remained isolated. Some exceptions include either Nahua (e.g., NAJ, NXP, NAG) or Mayan (e.g., MYA.C, MYA.Q, MYA.Y) populations, both of which are some of the most populous indigenous groups in Mexico, resulting in a lower probability of observing within-population connections in our sample. Two groups of closely related populations show higher number of between-population connections: Totonac and Nahua from Puebla (NXP and NFM), and Tzotzil, Tojolabal, and Lacandon from Chiapas (Fig. 1D).

In order to formally evaluate the probability of gene flow between populations after splitting, we used TreeMix (*23*) to construct a maximum likelihood tree allowing for a fixed number of migration events between populations. Figure 1E shows the splitting pattern without migration, which recapitulates the north-south gradient of differentiation observed in our previous analyses with Seri and Lacandon showing the highest levels of drift from the ancestral population, followed by Tojolabal. Shared clades denote clear regional relationships, such as all northern populations branching out from the same initial split at the root, followed by individual population splits and two major clades: one grouping all populations from the southern states of Guerrero and Oaxaca (Triqui, Zapotec south, Zapotec north, Mazatec, and Nahua Guerrero), and

the other all six Mayan speaking populations from the state of Chiapas and the Yucatan peninsula (Tzotzil, Tojolabal, Lacandon, Maya Campeche, Maya Quintana Roo, and Maya Yucatan). When running TreeMix allowing for migration edges in the tree, the matrix of residuals is used to infer pairs of populations with the poorest fit, thus becoming candidates for testing a better fit involving migration between them. Recent admixture can bias these estimations so we removed all indigenous samples with more than 2% of European ancestry as inferred by ADMIXTURE (*24*). We focused on the maximum likelihood trees for the top three events of migration (m=1 to 3) inferred from the data (Fig. S5). Interestingly, the first migration inference (m=1) involves gene flow from the Maya in Yucatan (MYA.Y) to the node of the Totonac (TOT), whose ancestors are believed to have built the large pre-Columbian city of El Tajin, located near the coast of the Gulf of Mexico, revealing a possible coastal corridor of gene flux between the Yucatan Peninsula and Central/Northern Mexico. The strongest migration rate (consistently greater than 50%) was detected between two closely related Nahua populations (NXP and NFM) both at m=2 and m=3. In the latter case an additional gene flow event was inferred from the Totonac to the neighboring Nahua in Puebla (NXP), consistent with the IBD patterns observed in Fig. 1D.

It is noteworthy that the different Nahua groups, while unified by historically speaking the same language, stem from different nodes in the tree. For example, NAJ from Jalisco is separated from the node giving rise to NXP and NFM (both from Puebla); and NAG from Guerrero is grouped together with Zapotec and other groups from southern Mexico. This translates into a lack of a single ancestry relating all the studied Nahua groups (as opposed to the Mayan groups, for instance), suggesting that current groups identified as Nahua are likely the result of linguistic and

cultural assimilation over genetically distinct groups, probably as a result of the extended domination of the Nahua-speaking Aztec empire in pre-Columbian times.

To identify loci showing extreme allele frequency gradients given the geographic origin of the individuals, we applied the SPA method(*25*) to the combined set of Native Mexican populations using their known coordinates (see Methods). Incorporating a model to infer the logistic slope of allele frequencies as a function of geographic positioning, SPA was used to scan the genome for SNPs that show steep allele frequency changes, which can result from the impact of recent positive selection. A total of 50 candidate regions were identified within the top 0.1% of the SPA score distribution (Table S3). SNPs in the MHC region have the most extreme allele frequency slopes, a region known to have been targeted by selection (*26-28*). Other immunity genes outside the MHC region are also among the top regions, including PSMD9 and TNFAIP3, followed by PEAK1 and PTPRD, involved in cell growth. Extreme values were also observed in the MFN2 gene region, which may play a role in the pathophysiology of obesity (*29*), as well as in HBS1L, which has been identified as a quantitative trait locus (QTL) controlling fetal hemoglobin level(*30*). When looking at the geographic distribution of the genotypes from the best SNP in the HBS1L region (rs1014021, see Fig. S6), we observed that the gradient is driven by southern Mexican populations showing high derived allele frequencies (60% on average). The full scan of SPA scores is available in Figure S7, where SNPs with extreme values correspond to potential regions under selection.

Mexican population substructure

In order to characterize the structure of indigenous populations and its impact in the admixture patterns of cosmopolitan Mexican samples we used ADMIXTURE, an unsupervised mixture

model algorithm, to analyze the combined dataset of continental source populations (including our 20 native Mexican populations, 16 European populations, and 50 West African Yorubas), and 420 admixed individuals from 11 Mexican states as well as 49 Mexican Americans from the Los Angeles area (Fig. 2A and 2B). At K=3, each set of reference parental groups gets its own cluster, with the exception of some Native Mexican groups such as Nahua and Maya, previously documented to have considerable proportions of European admixture (*9, 31*). Across the Mexican cosmopolitan samples we observe a clear gradient of increase Native American, and decreasing European, ancestry moving southwards, consistent with previous genome-wide reports of Mexican admixture patterns (*32*). African ancestry proportions are low on average (4.9%) and remain similar across most regions with the exception of the coastal states of Veracruz and Guerrero. Both states are known to have had increased slave trade activity (*33*), and some individuals from these states today show considerably higher proportions of African ancestry (up to 34%), also consistent with previous analyses of a subset of these samples at K=3 (*32*). However more in-depth analyses of ancestry were not possible in such initial screening as a single Native Mexican group, the Zapotec, was used as potential source population, precluding any further detection of sub-continental ancestry.

With a larger reference panel of 20 native populations we observe more detailed substructure at higher K values. We explored clustering patterns from K=2 through 20 (Fig. S8) and focus on K=9 for showing the lowest cross-validation error across runs (Fig. S9). At this level the Native cluster breaks down into six separate Native American components (Fig. 2B). Three of them are restricted to isolated populations (Seri, Lacandon, and Tojolabal), showing little sharing with neighboring indigenous groups. The other three show a wider but geographically well-defined distribution. First, there is a northern component represented by Tarahumara, Tepehuano, and

Huichol, which gradually decreases southwards until is virtually absent in Oaxaca and beyond. The second one is represented by southern populations from Oaxaca including Triqui, Zapotec, and Mazatec, reaching 99.9% in most Triqui individuals, and gradually decreasing northwards. In contrast there is a sudden disruption moving towards the Yucatan peninsula, where this southern component is limited to account for an average of 20% of the genome as it is mostly replaced by a local Mayan component, the third major component observed (Fig. 2B, bottom panel). Interestingly, this Mayan component is also present at ~10-20% in central native populations, but not in southern Oaxaca, supporting the hypothesis of a coastal or maritime route of gene flow between the Yucatan peninsula and central Mexico bypassing the mountain range of the Tehuantepec isthmus.

When looking at the distribution of these native components in the admixed genomes of cosmopolitan samples we observed a striking correlation with the patterns described before. Sonora and neighboring northern states show the highest average proportions (15%) of the northern native component (light blue in Fig 2B, bottom), while 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 northwards. Cosmopolitan samples from the Yucatan peninsula are the only ones whose Native American fraction of the genome is dominated by the Mayan component, while all other states show smaller and decreasing proportions northwards. Likewise, Mayan-related local components, Tojolabal and Lacandon, are detected above 1% exclusively among individuals from the neighboring states of the Yucatan peninsula. In contrast to population samples from particular states, Mexican Americans sampled in Los Angeles (MXL) do not share a homogeneous pattern, denoting their diverse array of origins within Mexico. Additionally, we detected substructure

within the European component at K=9 with a clear gradient of differentiation between northern European and southern Mediterranean populations, in agreement with previous analyses (*5, 34*). In all Mexican samples, the majority of its European ancestry comes from the southern Mediterranean component, consistent with historical records about the admixture process between Spanish Europeans and native Mexicans. The map in Figure 2A summarizes individual admixture proportions into population averages for each continental ancestry at K=3 and each native component at K=9. For instance, Oaxaca and Campeche share similar continental patterns, showing the highest averages of native ancestry at K=3 (85% and 80%, respectively). However, when broken down at K=9, we unveil that their native proportion is composed by completely different profiles, dominated by their corresponding local native components.

In order to formally test whether a correlation exists between the admixture proportions of each native component and geographic distance between samples, we ran a linear regression using individual values against their sampling location along a 45^{o} NW-SE axis along the length of the country. Figure 2C shows the geographic distribution of the six Native American components and the correlation with geographic location of cosmopolitan samples, all of which were highly associated with geography (joint Kruskal-Wallis test for latitude and longitude, all components $p < 10^{-6}$). Using Kriging interpolation we have also estimated the continuous geographic distribution of each native component across the full set of cosmopolitan populations throughout Mexico (Fig. S10).

Sub-continental origin of haplotypes measured from admixed genomes

The level of resolution that can be achieved in assessing admixed genomes' ancestry is largely dependent of the reference panel used to define potential source haplotypes. Most genomic

studies involving Mexican admixed genomes have made use of continental-level ancestral populations (*32*) while more recent ones have explored sub-continental ancestry to a limited extent, such as (*35*), or (*36*), who used three Mexican and five South American indigenous populations to evaluate a large cohort of Mexican samples. The Native components of the Mexican individuals all clustered as a single group next to the native Mexican reference populations.

We used our extensive reference panel of ancestral populations and novel statistical methods to explore the ancestral components of admixed Mexican genomes at a finer scale. First, we estimate local ancestry along the genome for each individual using PCADMIX, a PCA-based method supporting phased haplotype data and three-way admixture deconvolution (*37, 38*). Then we consider only those sites within genomic segments of inferred European, African or Native American ancestry and mask the rest of the genome to perform PCA with sub-continental reference panels (Fig. S11) In order to handle the large amount of missing data resulting from masking ancestry-specific segments across the genome, we implemented a novel Ancestry-Specific PCA (ASPCA) by adapting the subspace PCA algorithm introduced by (*39*) to handle phased haplotype data (see Methods). Previous implementations have adapted the same algorithm to genotype data (*36*), thus limiting the analysis to loci of homozygous ancestry. We applied ASPCA to admixed individuals with more than 25% of Native American or European ancestry (due to the lower amount of data from African segments we did not run ASPCA for the African component). Figure 3A shows the ASPCA of each Mexican individual's European haplotypes in the context of source European populations (data from (*40, 41*)), where they overlay over Southwest European samples, mostly from the Iberian Peninsula. The distribution of ASPCA values extends to a few outliers closer to Central European and Italian samples.

Notably, no European haplotypes from Mexican individuals fall within the Basque cluster, who group separately from the rest of Iberian samples. The Mexican population as a whole primarily has received ancestry then from Iberians, consistent with the primary Spanish colonization in Mexico. Figure 3B in contrast shows Mexican individuals' Native American haplotypes analyzed together with the Native Mexican reference panel. PCA space is dominated by the highly endogamous Native populations noted in prior analyses, but when plotting the ASPCA values for the admixed individuals only, we discover a strong correlation between Native ancestry and geography within Mexico (Fig. 3C). Here ASPC1 represents a geographic gradient from west to east and ASPC2 one from north to south, where the distribution of haplotypes highly resembles a geographic map of Mexico. Three main clusters are identified: that of individuals sampled in northern states, the one from central/southern states, and the one composed by individuals from the Yucatan peninsula. There is a gradual overlap between the first and second cluster of haplotypes, while the separation between the second and the third is much more abrupt, in agreement with the observed distribution of the Native components as described above. These results demonstrate that the structure in the admixed individuals is largely determined by fine-scale Native American ancestry. The correlation between ASPC values and geography is striking and is remarkable that it is uncovered only from the Native segments of cosmopolitan Mexican individuals.

To validate our results, we ran a supervised clustering analysis of Native segments from the admixed Mexican genomes using FRAPPE (*42*) at K=6 (Fig. S12) and confirmed that, on average, Mexicans sampled from different regions of Mexico derive differential ancestral contributions from each of the Native American components (see Methods).

Biomedical implications of sub-continental ancestry

We investigated whether the hidden population structure unveiled with ASPCA would also have biomedical relevance via investigating associations with physiological phenotypes. We focused on lung function testing via spirometry. Presently, lung function testing is one of the few clinical applications where self-reported race/ethnicity is used in interpreting a "normal" range and classifying disease and severity (*43, 44*).

We used physiologic measures of lung function among Mexican and Mexican American children with asthma from two independent studies: The Genetics of Asthma in Latino Americans (GALA I) study (*13, 45*) here comprising 68 probands from Mexico City (MX) and 120 Mexican American probands from the San Francisco Bay Area (SF), genotyped on the Affymetrix 6.0 array. The Mexico City Children's Asthma Study (*12, 46*) (MCCAS) comprised 492 probands all from Mexico City and was genotyped on the Illumina 550K. We focused on these two studies as they are trio-based ensuring accurate long-range haplotypic inference. We performed ASPCA separately for each study using our Native Mexican reference panel given the heterogeneity between the two genotyping platforms and to minimize potential distortions in principal components from unequal population sizes (*47*) (Fig S13). Given that ASPCA is unitless we then normalized each set of ASPC scores for comparison across studies, and used fixed effects meta-analysis where appropriate to estimate effect sizes and confidence intervals for the two studies combined.

First, as GALA I included individuals from two sampling locations we tested for detectable substructure in the ASPCA values to see if we could predict recruitment location merely from ASPCA values. Figure S13 (bottom) shows the ROC curve for the logistic regression classifying MX vs. SF cases based on their ASPCA values of Native American ancestry, with an AUC of

80%. Incorporating these values into a fuller model adjusting for overall global ancestry proportions (here both African and Native American), both ASPCs were significant: ASPC1 OR per SD: 0.44 (95% CI 0.22-0.68), $p=3.8x10^{-4}$, ASPC2 OR per SD: 0.52 (95% CI 1.03-2.75), p=0.039. The ASPCs defined similar axes as in the population structure analyses (Fig. S13). We observe that the region of Native ancestry most associated with immigration to the San Francisco Bay Area is in the Northwest of Mexico (joint ASPC likelihood ratio test $p=6.4x10^{-5}$), closest to the border with the USA, and independent of overall continental ancestry proportions.

With only proband cases in both studies we looked for associations between ASPC values and measures of lung function. We focused on forced expiratory volume in the first second (FEV_1), a standard measure of lung function used in clinical settings, as it is known to have ethnic heterogeneity [Hankinson 1996] and has previously been associated with ancestry (*43*). We used robust linear models (see Methods) to be less sensitive to outliers in our ASPC projections. We stratified by study and looked for associations with percent predicted values (*44*) to account for age, sex and height, while separately adjusting for overall ancestry proportions to minimize confounding. However it is important to note that these values are specific to children with asthma as neither study measured lung function in healthy controls.

We observed a significant association between FEV_1 and ASPC1, with a combined p-value of 0.0045 (-2.2% decrease per 1 SD, 95% CI (-3.74 - -0.69)), corresponding to the East-West component observed previously. MCCAS was significant on its own, while GALA I had a p=0.06, albeit with a much reduced sample size. The normalized association was remarkably homogeneous between GALA I and MCCAS given the differences in genotype platform, sampling locations, timing, and recruitment criteria (Fig. 4A). ASPC2, on the other hand, did not have a significant association with FEV_1. The combined results here indicate that sub-continental

ancestry as measured by ASPCA is important for characterizing clinical measurements, even independent of the overall admixture proportions.

To put the results in geographic context, we used the association with ASPC1 to infer expected values of FEV_1 across the mestizo samples from different states to estimate the expected change in lung function moving west to east across Mexico. Given the relationship with observed ASPCA values in GALA I we used extrapolated normalized values by state to infer the expected amount of change in FEV_1 for children with asthma in each state. We plot the means by state and predicted confidence intervals in Figure 4B based on the association observed in GALA I and MCCAS. Consistent with studies involving children with asthma we see expected values slightly below 100%. While each has fairly wide confidence intervals, the overall association results in an expected 7.3% average decrease in lung function between Sonora to the west and Yucatan to the east. This can have high downstream effects when diseases like asthma and chronic obstructive pulmonary disease (COPD) are partially diagnosed based on specific spirometric values.

A similar, significant association was previously demonstrated with African ancestry in African Americans (*43*). Using that same model the observed decrease of 7.3% in FEV_1 would be associated with a 33% increase in African ancestry in African Americans. In addition, lung function and FEV_1 values are known to decline with age. The 7.3% change is similar to that of a 30 year old Mexican American individual of average height aging 10.3 years if male or 11.8 years if female (*44*). Given that specific percent predicted thresholds are used as part of the diagnostic criteria of diseases such as asthma and COPD, individual sub-continental ancestry can potentially influence diagnoses despite population-specific reference equations.

Haplotype structure and haplotype sharing

Different population genetic profiles are known to influence the outcome of genetic association studies and the replication of significant GWAS hits across worldwide populations. Part of that variation is explained by the ancestry composition of each individual and the geographic stratification of the population. Therefore, the use of catalogs of human genetic variation ascertained in certain continental reference populations, may not be sufficient when the target population's ancestry is not fully represented in such panels. To assess to which extent continental populations from publicly available panels are capturing the haplotype variation found in cosmopolitan Mexican populations, we performed a genome-wide haplotype sharing analysis based on 100 Kb sliding windows. Figure 4C shows the proportion of haplotypes shared between the combined set of mestizo samples and different combinations of HapMap continental populations before and after including a combined set of Native American samples (see Methods). Any of the continental source populations alone (YRI, CEU, NAT) shares a limited proportion of haplotypes with mestizo samples (21.6%, 59.3%, and 78.6%, respectively). Although Mexican-American samples (MXL) were included in both the HapMap and 1000 Genomes catalogs, their average sharing only goes up to 81.2% and to 90.5% when combining MXL with all continental HapMap populations. It is only after adding Native American samples to this previous combination that nearly 100% of haplotypes are shared, maximizing the chances of capturing most of the variation using our catalogue of Mexican-specific variation.

Continental ancestry also varies across the genome and the relative proportion of African, European, and Native American ancestry at a given locus may affect the replication success of associations reported in any of the ancestral populations. By providing a local ancestry map averaging the proportions of European versus Native ancestry in the combined Mexican sample

(Fig. S14) we scan the genome for local ancestry fluctuations that may affect power in genetic studies. While we observed no genome-wide significant deviations in local ancestry patterns, the natural fluctuations in local ancestry can impact medical studies. For instance, one of the genes associated with age-related macular degeneration in populations of European descent (*ARMS2*) is located in one of the strongest peaks of Native American ancestry enrichment (Fig. S14-15), where up to 66% of the sampled Mexican haplotypes occur in a Native American background. Early age-related macular degeneration has been reported to have higher prevalence among Hispanics (*48*), but its local ancestry profile may limit the possibility of replicating the associated variants reported in European individuals while simultaneously increasing the possibility for discovering new population-specific risk variants.

Much effort has been invested in detecting common genetic variants associated with complex disease and replicating associations across populations. But functional and medically relevant variation may be rare and, thus, population-specific so without detailed knowledge about the geographic stratification of genetic variation, false-positive associations and lack of replication are likely to dominate the outcome of genetic studies in uncharacterized populations. Population structure as determined by cryptic relatedness is expected to be elevated in the populations sampled here, potentially complicating genetic association studies. However it also suggests that methods directly harnessing that structure, such as identity-by-descent mapping (*22, 49, 50*) may prove fruitful.

Conclusions

Here we have reported hitherto undetected fine-scale patterns of population substructure within Mexico and refined the genetic picture of relationships among indigenous groups. We

demonstrate that such structure has been shaped by extreme isolation between ancestral populations and that it directly impacts the genetic composition of admixed individuals from the same regions. Furthermore, our work demonstrates that fine-scale population structure going back centuries is not merely a property of isolated or rural indigenous communities. Rather, individuals from large cosmopolitan cities reflect the underlying genetic ancestry of local native populations, arguing for a strong relationship between the indigenous and the Mexican mestizo population and, therefore, against any social segregation between them. Most importantly, this has relevant biomedical implications both within Mexico and U.S.-based Mexican communities, as the observed association between genes, geography, and physiological phenotypes indicates the importance of understanding not just overall ethnicity but also the role of fine-scale patterns of ancestry in complex traits and disease diagnosis.

References and Notes:

1. S. Gravel *et al.*, Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 11983 (Jul 19, 2011).

2. M. Jakobsson *et al.*, Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998 (Mar 21, 2008).

3. N. Rosenberg *et al.*, Genetic structure of human populations. *Science (New York, NY)* **298**, 2381 (Dec 20, 2002).

4. S. Tishkoff *et al.*, The Genetic Structure and History of Africans and African Americans. *Science (New York, NY)*, (May 30, 2009).

5. J. Novembre *et al.*, Genes mirror geography within Europe. *Nature* **456**, 98 (Nov 06, 2008).

6. S. Wang *et al.*, Genetic variation and population structure in native Americans. *PLoS Genet* **3**, e185 (Nov, 2007).

7. B. M. Henn, S. Gravel, A. Moreno-Estrada, S. Acevedo-Acevedo, C. D. Bustamante, Fine-scale population structure and the era of next-generation sequencing. *Hum Mol Genet* **19**, R221 (Oct 15, 2010).

8. J. Hey, On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biology* **3**, e193 (Jul 01, 2005).

9. J. Z. Li *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. *Science (New York, NY)* **319**, 1100 (Mar 22, 2008).

10. V. Acuna-Alonzo *et al.*, A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. *Hum Mol Genet* **19**, 2877 (Jul 15, 2010).

11. E. E. Kenny *et al.*, Melanesian blond hair is caused by an amino acid change in TYRP1. *Science* **336**, 554 (May 4, 2012).

12. D. B. Hancock *et al.*, Genome-wide association study implicates chromosome 9q21.31 as a susceptibility locus for asthma in mexican children. *PLoS genetics* **5**, e1000623 (Aug 01, 2009).

13. E. G. Burchard *et al.*, Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. *Am J Respir Crit Care Med* **169**, 386 (Feb 1, 2004).

14. D. Reich *et al.*, Reconstructing Native American population history. *Nature*, (Jul 11, 2012).

15. K. Sandoval *et al.*, Y-chromosome diversity in Native Mexicans reveals continental transition of genetic structure in the Americas. *American journal of physical anthropology* **148**, 395 (Jul, 2012).

16. H. M. Cann, K. K. Kidd, R. Lisker, R. Radvany, R. Payne, Genetic structure of the HL-A system in a Nahua Indian population in Mexico. *Tissue Antigens* **3**, 364 (1973).

17. R. Lisker, E. Ramirez, V. Babinsky, Genetic structure of autochthonous populations of Meso-America: Mexico. *Hum Biol* **68**, 395 (Jun, 1996).

18. K. Sandoval *et al.*, Linguistic and maternal genetic diversity are not correlated in Native Mexicans. *Hum Genet* **126**, 521 (Oct, 2009).

19. A. Gonzalez-Martin *et al.*, Analyzing the genetic structure of the Tepehua in relation to other neighbouring Mesoamerican populations. A study based on allele frequencies of STR markers. *Am J Hum Biol* **20**, 605 (Sep-Oct, 2008).

20. R. Rubi-Castellanos *et al.*, Pre-Hispanic Mesoamerican demography approximates the present-day ancestry of Mestizos throughout the territory of Mexico. *Am J Phys Anthropol* **139**, 284 (Jul, 2009).

21. M. Jobin, J. Mountain, REJECTOR: Software for Population History Inference from Genetic Data via a Rejection Algorithm. *Bioinformatics (Oxford, England)*, (Oct 20, 2008).

22. A. Gusev, J. Lowe, M. Stoffel, M. Daly, D. Altshuler, Whole population, genomewide mapping of hidden relatedness. *Genome research*, (2008).

23. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967 (2012).

24. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655 (Sep 01, 2009).

25. W. Y. Yang, J. Novembre, E. Eskin, E. Halperin, A model-based approach for analysis of spatial structure in genetic data. *Nat Genet* **44**, 725 (Jun, 2012).

26. A. Albrechtsen, I. Moltke, R. Nielsen, Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* **186**, 295 (Sep, 2010).

27. X. Liu *et al.*, Detecting and Characterizing Genomic Signatures of Positive Selection in Global Populations. *Am J Hum Genet*, (May 22, 2013).

28. J. Pickrell *et al.*, Signals of recent positive selection in a worldwide sample of human populations. *Genome research*, (Apr 23, 2009).

29. A. Zorzano, D. Sebastian, J. Segales, M. Palacin, The molecular machinery of mitochondrial fusion and fission: An opportunity for drug discovery? *Current opinion in drug discovery & development* **12**, 597 (Sep, 2009).

30. S. L. Thein, S. Menzel, M. Lathrop, C. Garner, Control of fetal hemoglobin: new insights emerging from genomics and clinical implications. *Hum Mol Genet* **18**, R216 (Oct 15, 2009).

31. X. Mao *et al.*, A genomewide admixture mapping panel for Hispanic/Latino populations. *Am J Hum Genet* **80**, 1171 (Jun, 2007).

32. I. Silva-Zolezzi *et al.*, Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proceedings of the National Academy of Sciences of the United States of America*, (Jun 11, 2009).

33. G. A. Beltrán, The Slave Trade in Mexico. *The Hispanic American Historical Review* **24**, 412 (1944).

34. O. Lao *et al.*, Correlation between genetic and geographic structure in Europe. *Curr Biol* **18**, 1241 (Aug 26, 2008).

35. K. Bryc *et al.*, Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America* **107 Suppl 2**, 8954 (Jun 11, 2010).

36. N. A. Johnson *et al.*, Ancestral components of admixed genomes in a mexican cohort. *PLoS genetics* **7**, e1002410 (Dec, 2011).

37. B. M. Henn *et al.*, Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* **8**, e1002397 (Jan, 2012).

38. A. Brisbin *et al.*, PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Hum Biol* **84**, 343 (Aug, 2012).

39. T. Raiko, A. Ilin, J. Karhunen, Principal component analysis for large scale problems with lots of missing values. *Machine Learning: ECML 2007*, 691 (2007).

40. M. R. Nelson *et al.*, The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *American journal of human genetics* **83**, 347 (Sep 01, 2008).

41. L. R. Botigue *et al.*, Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc Natl Acad Sci U S A*, (Jun 3, 2013).

42. H. Tang, J. Peng, P. Wang, N. J. Risch, Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology* **28**, 289 (Jun, 2005).

43. R. Kumar *et al.*, Genetic ancestry in lung-function predictions. *New England Journal of Medicine* **363**, 321 (Jul 22, 2010).

44. J. L. Hankinson, J. R. Odencrantz, K. B. Fedan, Spirometric reference values from a sample of the general U.S. population. *Am J Respir Crit Care Med* **159**, 179 (Jan, 1999).

45. D. G. Torgerson *et al.*, Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. *J Allergy Clin Immunol* **130**, 76 (Jul, 2012).

46. H. Wu *et al.*, Evaluation of candidate genes in a genome-wide association study of childhood asthma in Mexicans. *J Allergy Clin Immunol* **125**, 321 (Feb, 2010).

47. G. McVean, A genealogical interpretation of principal components analysis. *PLoS Genet* **5**, e1000686 (Oct, 2009).

48. B. Munoz, R. Klein, J. Rodriguez, R. Snyder, S. K. West, Prevalence of age-related macular degeneration in a population-based sample of Hispanic people in Arizona: Proyecto VER. *Archives of ophthalmology* **123**, 1575 (Nov, 2005).

49. A. Albrechtsen *et al.*, Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology* **33**, 266 (May 01, 2009).

50. S. R. Browning, E. A. Thompson, Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics* **190**, 1521 (Apr, 2012).

Acknowledgments: We thank study participants for generously donating DNA samples, Brenna M. Henn and Simon Gravel for helpful discussions and comments on earlier versions of the manuscript. This project was supported by NIH grant (to CDB), as well as the Mexican government. This project was also supported in part by the George Rosenkranz Prize for Health Care Research in Developing Countries awarded to AM-E; UCSF Chancellor's Research Fellowship, Dissertation Year Fellowship, and NIH Training Grant T32 GM007175 (to CRG); the RWJF Amos Medical Faculty Development Award; the Sandler Foundation; the American Asthma Foundation (to EGB); This research was supported in part by National Institutes of Health (ES015794, GM007546, GM061390, HL004464, HL078885, HL088133, MD006902, RR000083).The collections and methods for the Population Reference Sample (POPRES) are described by Nelson et al. (2008). The datasets used for the analyses described in this manuscript

were obtained from dbGaP at

http://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000145.v1.p1 through dbGaP accession number phs000145.v1.p1.

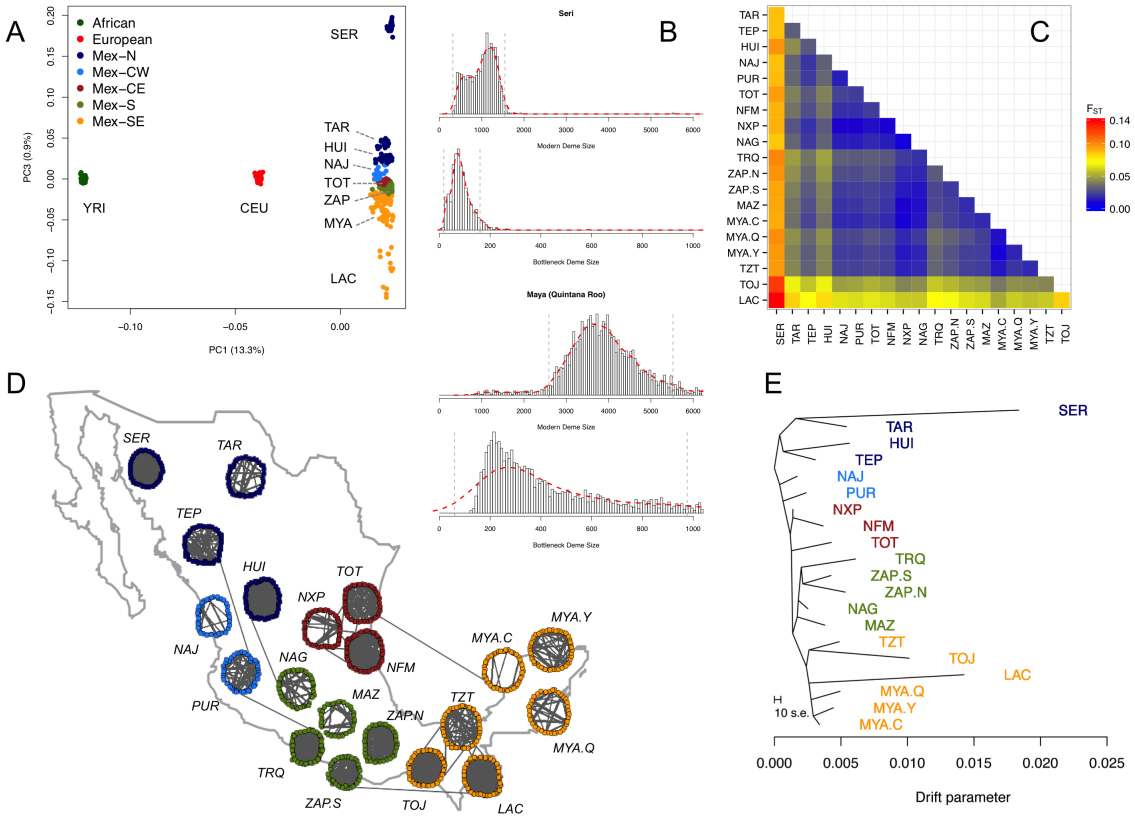


Fig. 1. Genetic structure of Native Mexican populations. (**A**) Principal component analysis of Native Mexicans with HapMap YRI and CEU samples color coded by geographic regions. Population labels as detailed in Table S1. (**B**) Simulated posterior distribution of effective population sizes in the Seri and the Maya based on cumulative runs of homozygosity (cROH), generated by sampling from a uniform distribution of Ne and keeping simulated parameters within 20% of the observed cROH with REJECTOR. Estimates are given for the contemporary deme size and for that during the bottleneck of Native Americans. Parameters for the other studied populations are available in Fig. S2 and S3. (**C**) Pairwise F_{ST} values among Native Mexican populations ordered geographically. (**D**) Pairwise matches between individuals sharing more than 20 cM of the genome as measured by the total of segments identical-by-descent

(IBD). Each line denotes a connection between two individuals and each dot represents one individual, with positions on the map indicating approximate sampling locations. The pattern across different populations shows high within-population sharing compared to between-populations. Results from the full range of IBD thresholds are shown in Fig. S4. (**E**) 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. S5), but a maximum likelihood tree with only Native Mexicans is shown in order to get a closer view at their drift parameter differences.

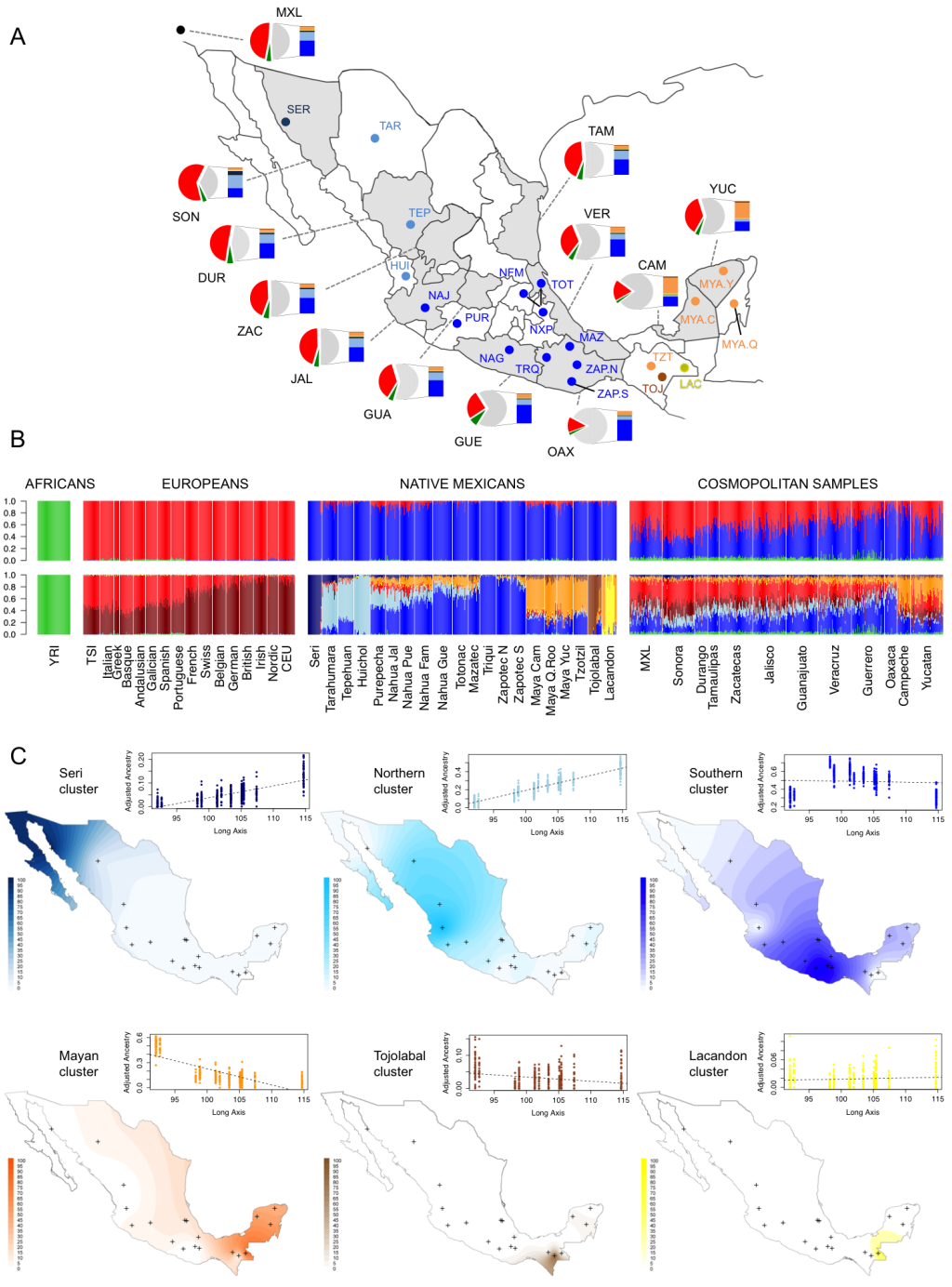


Fig. 2. Mexican population structure. (**A**) Map showing geographic locations of sampled populations and admixture average proportions. Population codes are detailed in Table S1. Dots

correspond to Native Mexican populations color-coded according to K=9 clusters identified in B (bottom), and shaded areas are 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). For each state, 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 for the combined dataset of 1,282 individuals including African, European, Native Mexican, and cosmopolitan Mexican samples (detailed in Table S1). 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. For each native cluster, scatter plots with linear fits show ADMIXTURE values observed in cosmopolitan samples versus a distance metric summarizing latitude and longitude (long axis) for the eleven sampled states. Within each plot 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 (see Methods for details).

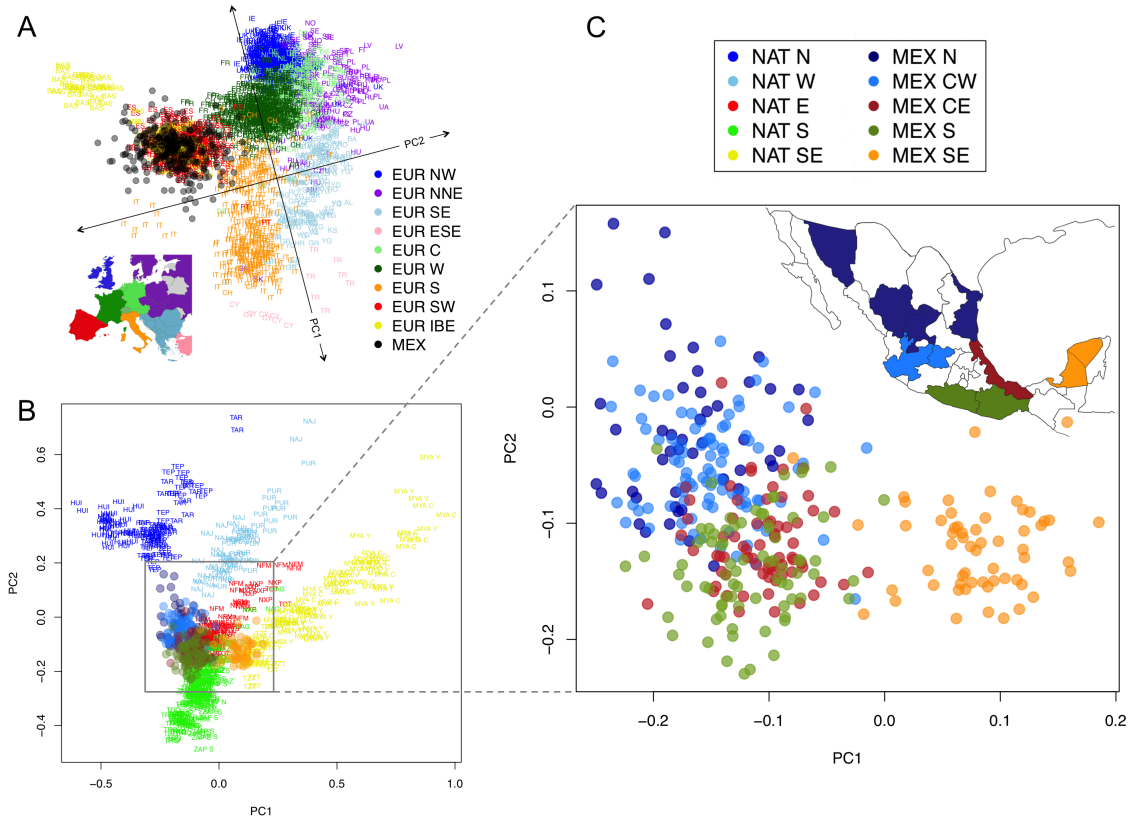


Fig. 3. Sub-continental ancestry of admixed Mexican genomes. (**A**) Ancestry-specific PCA (ASPCA) of European segments from cosmopolitan Mexican samples (black circles) together with our reference panel of 1,387 European individuals from POPRES (labeled by country code) plus 55 additional samples from Spain (yellow labels). Each black circle represents the combined set of Mexican haplotypes called European along the haploid genome of each sample with >25% of European ancestry. Axes were rotated 16 degrees counterclockwise to approximate the geographic orientation of population samples over Europe. Inset map shows POPRES countries of origin color-coded by region (areas not sampled in gray and Switzerland in intermediate shade of green to denote shared membership with EUR W, EUR C, and EUR S). Population codes and regions within Europe are detailed in Table S1. (**B**) ASPCA analysis of Native American segments from Mexican cosmopolitan samples (colored circles) together with our dataset of 20

indigenous Mexican populations (labeled by population code). Samples with >10% of non-native admixture were excluded from the reference panel as well as population outliers such as Seri, Lacandon, and Tojolabal. (**C**) 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 sub-continental 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. Inset map shows the geographic origin of cosmopolitan samples per state color-coded by region. All participants were required to have 4 grandparents born in the same state (see Methods for details).

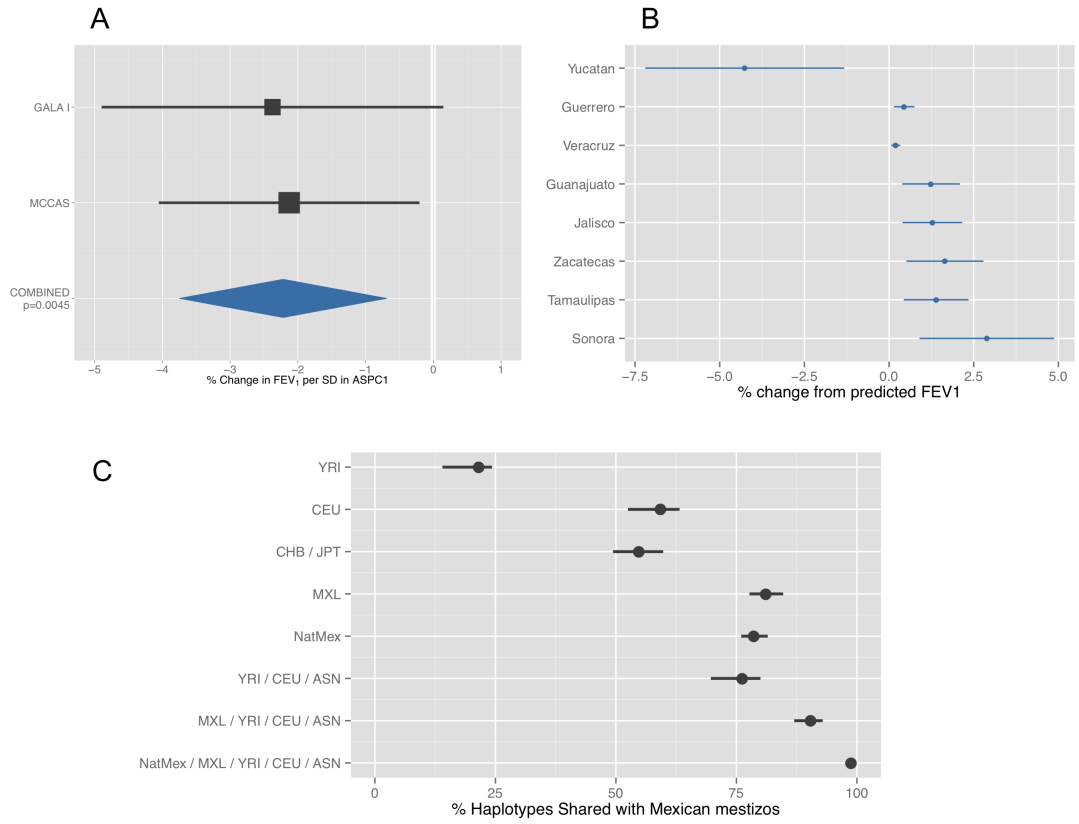


Fig. 4. Biomedical implications of the genetic substructure of Mexican populations. (**A**) Coefficients and 95% confidence intervals for associations between ASPC1 and lung function measures (FEV_1) from Mexican participants of the Genetics of Asthma in Latino Americans (GALA I) study, and the Mexico City Childhood Asthma Study (MCCAS), as well as both studies combined (see Methods for details). (**B**) Extrapolations based on normalized ASPC1 values of estimated FEV_1 values by state, using the regression model in Fig. S13 (**C**) Genome-wide proportions of haplotypes shared between the combined sample of Mexican mestizo populations and different continental populations in autosomal chromosomes. The "NatMex" panel here consists of 71 individuals from 3 indigenous groups, one from each of the major genetic components identified in Fig. 2. Haplotype sharing analysis was performed using the subset of Mexican samples with the largest intersection of genotyped SNPs (785,663) with

HapMap3 populations, which included 312 Mexican mestizos from diverse cosmopolitan populations (see Methods for details).

Chapter 5: Supplementary Materials for

The Genetic History and Structure of Mexican Populations

Materials and Methods

Sample collection and genotyping

Institutional review board (IRB) approval for this project was obtained from Stanford University (File: NOT03H02) for obtaining and analyzing de-identified DNA specimens from participating institutions. Written informed consent was obtained from all participants and research/ethics approval and permits were obtained from the following institutions: the University of Guadalajara, the National Institute of Medical Sciences and Nutrition Salvador Zubirán (INNSZ), and the National Institute of Genomic Medicine (INMEGEN). Samples were collected over several years by researchers from these institutions under protocols consistent with biomedical and/or population genetics studies aimed at characterizing the genetic diversity of Mexican populations. Sampling locations and summary data for the populations included in the study are detailed in **Table S1**. A total of 362 samples from 15 indigenous populations were genotyped at the University of California, San Francisco (UCSF) by using Affymetrix 6.0 arrays and 466 samples were genotyped at the National Institute of Genomic Medicine (INMEGEN) by

using a combination of Affymetrix 500K and Illumina 550 arrays. Samples genotyped at INMEGEN include 370 cosmopolitan samples from 10 different Mexican states and 96 samples from three indigenous populations, which were collected as part of the Mexican Genomic Diversity Project (MGDP)(*1*). All participants were required to have 4 grandparents born in the same state. Overall, this combined genotyping effort generated SNP array data for 828 samples from 28 different Mexican populations. All samples were genotyped from genomic DNA extracted from blood.

Data curation

Curation of Native Mexican samples: a total of 458 samples were initially genotyped (362 by using Affymetrix 6.0 arrays and 96 by using Affymetrix 500K arrays). The number of markers included in the Affymetrix 6.0 SNP array determined our starting SNP density before intersecting with data from additional arrays. A total of 909,622 SNPs were successfully genotyped. We removed 2,919 SNPs with duplicate marker names, 1,217 SNPs with no physical position in the NCBI Build 36.1 human reference sequence (hg18 assembly), and 8,087 SNPs failing Hardy-Weinberg equilibrium at $1x10^{-5}$. We restricted to autosomal SNPs and samples with more than 90% of genotyping rate. We removed 3 samples due to evidence of being duplicates of another sample. As part of the recruiting strategy, 40 trios and 6 duos were included to improve phasing accuracy of haplotype-based analyses and ancestral reference panels for admixture deconvolution (see below). One trio showed an excess of Mendel errors and was thus excluded from trio phasing. Subsequently, the 46 individuals constituting the offspring of all trios and duos were removed from most of the analyses. We did not systematically filter for second-degree or lower relatives as part of our initial curation given that some of the subsequent

analyses make use of IBD information to describe within- and between-population connections among pairs of individuals across Native Mexican populations (see sections below). We then excluded 8 individuals due to a high proportion (>30%) of non-Native ancestry, as these are likely to correspond to sampling exceptions rather than being part of the population's admixture pattern. This was confirmed by PCA analysis where these samples appeared to be outliers relative to others from the same population. Since the scope of the study is to assess the population structure, including the characterization of recent admixture events among Native Mexicans, we did not initially filter genomic segments or individuals with some degree of non-Native ancestry. However, more stringent filters were applied as needed for particular analyses as detailed in the subsequent sections below. After data curation, the number of Native Mexican samples genotyped for this study was 401 (**Table S1**).

Curation of Cosmopolitan Mexican samples: Out of the 370 cosmopolitan samples genotyped at INMEGEN, 313 were genotyped by using both Affymetrix 500K arrays and Illumina 550K arrays (covering 7 Mexican states), and 57 samples were genotyped by using Illumina 550K arrays only (covering 3 additional Mexican states). For the subset of cosmopolitan samples genotyped with both arrays, genotype data for nearly 1 Million SNPs were available for analyses.

Data integration

To combine our dataset with additional preexisting data and assembly continental reference panels of potential ancestral populations relevant to the Mexican admixture process, our data were integrated with previously genotyped datasets from various sources. Additional Mexican data included Affymetrix 500K genotypes for 53 Native individuals from 2 Mexican indigenous populations (*2*), Affymetrix 6.0 genotypes from 49 Mexican-Americans (MXL) sampled in Los

Angeles, California as part of the International HapMap project phase 3, and Affymetrix 500K genotypes for 50 Mexicans of admixed origin sampled in Guadalajara, Jalisco included in the Population Reference Sample (POPRES) data set. European data were obtained from a selected subset of 204 European samples from POPRES to be included as part of the reference panel of ancestral populations. Inclusion criteria were based on maximizing geographic representation of regions within Europe and equalizing sample sizes to those available for the Native Mexican populations (i.e., around 20, see **Table S1**). The collections and methods for the POPRES Sample are described by Nelson et al. (*3*). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v1.p1 through dbGaP accession number phs000145.v1.p1. Additional European populations from Spain (n=55) included Basque, Andalusian, and Galician (*4*), and additional HapMap samples included 25 Tuscans (TSI) and 25 Utah residents of Northern European descent (CEU). Finally, 50 Yorubans from Ibadan, Nigeria (YRI) from HapMap were included as reference panel for West African ancestry. A total of 511 additional samples were integrated from previously generated datasets. The dataset analyzed here is the result of merging autosomal SNP array data from these different sources and consists of up to 1,282 samples, including 454 Native Mexicans from 20 indigenous populations, 469 cosmopolitan Mexican samples from 12 locations, and 359 ancestral European and West African populations.

Three main working datasets with variable SNP densities were constructed after merging multiple datasets and reapplying data quality control filters (now raised to 95% of call rate for SNPs and samples and excluding SNPs with ambivalent strandedness). Namely, one dataset was constructed considering the intersection of Affymetrix and Illumina data (71,581 SNPs), one

consisting of Affymetrix data only (372, 692 SNPs), and one combining the union of both Affymetrix and Illumina arrays (785,663 SNPs).

Table S2 describes the details of these three datasets. Most of the analyses presented here are based on the Affymetrix dataset (including data from 500K and 6.0 arrays) as this combination offered the best balance between SNP density and number of populations included (both indigenous and cosmopolitan). Nonetheless, we also used the combined dataset of Affymetrix and Illumina arrays in those analyses that were more robust to lower marker densities and where maximizing the number of populations was essential. Likewise we used the union of these platforms in those analyses requiring the densest dataset though across a limited number of populations.

The steps described above correspond to our initial data curation and the resulting datasets (listed in **Table S2**) constituted the base of all population structure analyses. However further filters were applied to exclude additional samples or integrate additional data for particular analyses as described below.

Population structure of Native Mexicans

We used the Affymetrix dataset (372, 692 SNPs) in all the analyses focused on Native Mexican populations. We restricted to individuals having >90% of Native American ancestry (average proportion of Native American ancestry among remaining individuals was 97.26%). Therefore, in addition to the initial data curation steps described above, the following samples were removed from the Affymetrix dataset:

Samples filtered due to <90% NAT ancestry

Population	N (pre-filter)	90% filter	N (post-filter)
SER	21	2	19
TAR	24	6	18
TEP	23	3	20
HUI	24	0	24
NAJ	20	8	12
PUR	23	8	15
TOT	23	3	20
NXP	22	15	7
NFM	27	5	22
NAG	29	1	28
TRQ	24	0	24
ZAP.N	21	0	21
ZAP.S	23	0	23
MAZ	17	0	17
TZT	21	0	21
TOJ	21	1	20
LAC	22	1	21
MYA.Q	18	4	14
MYA.C	27	13	14
MYA.Y	24	8	16
TOTALS	454	78	376

Principal component analysis (PCA) and population differentiation: We used EIGENSOFT(*5*) to perform PCA and R package was used to generate the plots. Pairwise F_{ST} values for each population comparison were calculated using the estimator of Weir & Cockerham (*6, 7*), and ggplot2(*8*) was used to create the plots.

Rejection algorithm and demographic estimation: To infer basic parameter estimates about each Native Mexican population, in particular, bottleneck strength and current N_e values, we implemented a demographic estimation method using approximate Bayesian computation via rejection algorithm as built into REJECTOR2 (*9*). We focused on a tract length statistic sensitive to bottlenecks known as cumulative Runs Of Homozygosity (cROH). We assigned ROH based on sliding windows, a minimum of 50 SNPs in a tract, and allowing for no more than one heterozygous SNP per 500kb window. This set of criteria similar to other researchers who

identified ROH in humans (see (*10-12*)), except that we used imputed genotypes from BEAGLE to avoid missing data issues. Given that these tracts are length-based and not dependent on the site frequency spectrum these are unlikely to be highly affected by ascertainment bias. Indeed previous simulations indicate accurate recovery of demographic parameters is possible using cROH statistics calculated from array genotypes (*12*).

We generated a set of simulations similar to Henn et al.: moving forward in time, we begin with a fixed large population size, then the population experiences a bottleneck and subsequent recovery to modern day deme size, with demographic parameters drawn from uniform priors. We used the computationally efficient approximate coalescent simulator MaCS (*13*) for simulation, and a tolerance (alpha level) of 20% between the observed and simulated sequences to accept or reject simulations. To make simulations tractable, we only investigated ROH on chromosome 1, and to use the maximum density of genotyped SNPs, we restricted to Native Mexican populations for which Affymetrix 6.0 array data was available (see Table S1). For each population we generated 100,000 simulated data sets. Acceptance rates varied between ~1-3%. For estimating final parameters, we employed a density-based smoothing in R over each histogram of accepted runs to estimate modes and 95% confidence intervals of each parameter of interest based on the profile approximate likelihoods. We then created plots with both the real histograms and the smoothed density values, plotting the informative portion of the accepted runs, both in summary form (Fig. S2) and the individual profiles (Fig. S3).

Identity-by-descent (IBD) analysis: Genotype data were phased using BEAGLE (*14, 15*) with available duos and trios used as training sets. We estimate the amount of DNA shared identically by descent (IBD) using the GERMLINE software (*16*), with a 5 cM threshold to eliminate false

positive IBD matches. All 5 cM or greater segments shared IBD between pairs of individuals were summed, and binned into 9 categories as detailed below. We then used the graph visualization software ShareViz [http://www.cs.columbia.edu/~itsik/sharevizWeb/shareviz.html] to visualize within- and between-population relationships of pairs of individuals (Fig. 1D and S4).

Estimated degrees of relatedness and IBD binning

% shared	cM	Relation°	IBD range (cM)	binning
100	3000	Self	-	-
50	1500	1°	> 1300	9
25	750	2°	650 - 850	8
12.5	375	3°	325 - 425	7
6.75	188	4°	163 - 213	6
3.37	94	5°	80 - 110	5
1.69	47	6°	40 - 53	4
0.85	23.5	7°	20 - 27	3
0.42	11.75	8°	10 - 13	2
0.21	5.875	9°	5 - 7	1

Population Tree analysis: trees have been widely used in population genetics to visualize the relationships among populations. While providing a valuable initial assessment of population relationships, a bifurcation tree might be a simplistic representation of human population history as it assumes population splits with no further gene flow between them. To overcome this problem, new methods have been recently developed allowing for the inclusion of gene flow between edges and representing population relationships by means of a reticulated graph rather than a strict bifurcation tree. Here we used TreeMix v1.0 (*17*) to infer patterns of population splitting and mixing from genome-wide allele frequency data. It estimates the maximum likelihood tree for a given set of populations given a Gaussian approximation to allele

frequencies, and then attempts to infer a number of admixture events. Before adding migration, we run TreeMix with our set of 20 Native Mexican populations and HapMap continental populations (YRI, CEU, and ASN) as outliers to help us set the root of the tree in subsequent runs (Fig. S5). Although not representing a perfect fit to the data, we used the maximum likelihood tree without migration to evaluate the general topology and the extent of population drift in terms of allele frequency shift from an ancestral population. We then used the residuals matrix to identify pairs of populations showing poor fits in the initial tree. These are then considered as candidates around which we add migration edges and try new rearrangements of the tree now accounting for *n* number of migration events. As a test run, we first used our previous panel (Native Mexicans plus CEU and YRI) adding MXL from HapMap as a population with known recent admixture. The resulting graph with allowed migration events showed that the strongest signal of gene flow comes from CEU to MXL, consistent with known historical records of these populations. Given that recent admixture can bias the signals detected by TreeMix, we restricted further runs with migrations to individuals with ≥98% of Native American ancestry in order to infer historical admixture events among Native Mexican populations. This filter removed the following samples in addition to the ones removed by the 90% filter:

Samples filtered due to <98% NAT ancestry

Population	N (pre-filter)	98% filter	N (post-filter)
SER	19	0	19
TAR	18	7	11
TEP	20	10	10
HUI	24	2	22
NAJ	12	10	2
PUR	15	15	0
TOT	20	5	15
NXP	7	5	2
NFM	22	14	8
NAG	28	4	24
TRQ	24	0	24
ZAP.N	21	0	21
ZAP.S	23	2	21
MAZ	17	6	11
TZT	21	2	19
TOJ	20	6	14
LAC	21	4	17
MYA.Q	14	14	0
MYA.C	14	12	2
MYA.Y	16	12	4
TOTALS	376	130	246

Scan for extreme allele frequency gradients: We used the spatial ancestry analysis (SPA) method (*18*) to identify SNPs with steep allele frequency gradients in Native Mexicans. A supervised analysis was performed using known latitude and longitude coordinates of sampling locations for the combined set of indigenous populations from our Affymetrix global dataset (see Table S2). Seri and Lacandon were subsequently removed to avoid possible bias due to their extreme isolation as revealed in previous analyses. Empirical p-values for each SNP were obtained by rank transformation of the raw SPA scores. Candidate regions were then defined by selecting the top 0.1% of SNPs of the empirical distribution, and subsequently merging SNPs separated by less than 500 kB into a single region. In order to avoid spurious outliers, we required that candidate regions spanned at least 1kB (i.e., a minimum of two outlier SNPs per region). A total of 50 candidate regions were identified within the top 0.1% of the score distribution (summarized in Table S3), and the genotypes for the most extreme SNP within each region are plotted in Figure S6. The full genome scan of SPA scores is available in Figure S7.

Population structure of cosmopolitan samples

We used the combined Affymetrix + Illumina dataset (71,581 SNPs) to run cluster-based analysis and PCA on the full set of samples listed in Table S1. This allowed us to include the maximum number of cosmopolitan samples to evaluate the impact of Native American substructure in the composition of admixed Mexican genomes.

Structure analysis: We used the block relaxation algorithm implemented in ADMIXTURE (*19*) to estimate individual ancestry proportions given *K* ancestral populations. We initially run from *k*=2 through 20 using the global dataset with the maximum number of available individuals to explore general clustering patterns. We then filtered first- and second-degree relatives and selected subsets of HapMap and POPRES individuals to roughly equalize sample sizes to those available for Native Mexican populations (**Table S1**). We found extensive substructure not only among the ensemble of recently admixed cosmopolitan Mexican samples, but also among the different ancestral populations. This was true not only for Native Mexican populations, but also for Europeans showing varying proportions from different clusters within Europe (fig. S8). Therefore, rather than using *reference* individuals as supervised training samples (which are assumed to have 100% ancestry from some ancestral population), we ran an unsupervised analysis to let ADMIXTURE estimate ancestry values across all samples. We used the default setting (folds=5) to perform ADMIXTURE's cross-validation procedure for evaluating fit of different values of K. Figure S9 shows the cross-validation error for each run, where *k*=9 showed the lowest error estimates (0.49798), indicating that sub-continental clustering levels are a sensible modeling choice for Mexican populations rather limiting to the usual continental-level structure of *k*=3. Additionally, we found constantly increasing Log likelihood values for all runs from *k*=2 to *k*=10 (fig. S9), where *k*=9 showed the maximum number of population-level

clusters among Mexicans. An additional European sub-continental component was detected at k=10 and found to be restricted to the Basque population and shared to a limited extent with other Iberian populations (fig. S8). At k=11, a group of 3 MXL samples clustered apart showing full membership to their own component, reflecting possible cryptic relatedness among them. Due to their shared ancestry with other Mexican cosmopolitan samples, residual proportions of this "MXL component" were also assigned to most of the remaining individuals, which is probably not the best description of their actual ancestral components given the observed patterns at earlier ks. This is also reflected in the subtle drop of the Log likelihood increasing curve when compared to all other runs. This component remained stable across higher ks, while other population-specific components appeared among Native Mexicans from k=12 through 20, but with less clear contribution into the admixed Mexican genomes (fig. S8). Likewise, all clusters detected at k=9 remained constant throughout the rest of runs up to k=20. In conclusion, as a result of the observations detailed above, we found k=9 to be the most informative run for purposes of characterizing sub-continental ancestry of Mexican populations, and therefore, several subsequent analyses described below were based on ADMXTURE proportions at k=9. In order to check for possible convergence variation, we performed 10 additional runs using different random seeds per run and the program converged after detecting the same clusters previously observed in all cases. We also estimated parameter standard errors using 200 bootstrap replicates per run. In general, standard errors were lower for individuals showing complete membership to highly divergent populations, such as Yoruba, Seri, Triqui, Tojolabal, and Lacandon (average error <0.01). In contrast, the two components accounting for most of the error at k=9 were Northern versus Southern European (standard error =0.029). The average error across all individuals and components was 0.016. The number of markers used is also known to

affect the performance of cluster-based algorithms. According to the ADMIXTURE guidelines (*19*), 10,000 markers suffice for continental-level distinction, while numbers closer to 100,000 are recommended for within-continent separation, assuming for instance European populations (i.e. $F_{ST} < 0.01$). Given that we are using more than 71,000 markers (using our global Affymetrix + Illumina dataset) and that all ancestral populations involved have $F_{ST} > 0.02$, we expect our ancestry estimates to be reasonably accurate. Nonetheless, we also ran k=2 through k=20 using the global Affymetrix dataset (>370,000 markers) using the same settings described above and there were no significant differences in parameter estimates for individuals represented in both datasets.

Correlation of cluster membership and geographic coordinates: From the clustering patterns observed across Mexican states in the ADMIXTURE analysis, a clear correlation can be appreciated between the geographic location of samples and their membership to the six main Native Mexican clusters. To formally test for significance with Latitude and Longitude we performed a linear regression for each component. We transformed latitude and longitude to create estimates across the "long axis" of Mexico, running NW-SE to better summarize the geography of Mexico in a single distance rather than latitude or longitude alone. Because the southern component decreases both northwards and towards the Yucatan peninsula, the correlation is less pronounced when Campeche and Yucatan samples are included.

Admixture maps: We used Kriging methods to interpolate ADMIXTURE proportion values for displaying the six native components identified at K=9 across both Native Mexican and cosmopolitan samples (Fig. S10). ADMIXTURE values from cosmopolitan samples (which usually show varying proportions of non-native admixture) were adjusted so that the sum of

ancestry proportions coming from Native American components equals 1. Contour maps were created using MapViewer (Golden Software).

Local ancestry estimation

We used a PCA-based admixture deconvolution method (PCAdmix, (*20*)) to estimate local ancestry across the genome. This method uses phased genotype data to estimate posterior probabilities of ancestry for windows along each chromosome. First, ancestral populations are thinned for SNPs with r2<0.8 in order to remove highly linked alleles from different populations, which can overfit and lead to spurious ancestry transitions. Second, chromosomes for each individual in a population are artificially strung together to create two extended chromosomal haplotypes; this step allows us to use the full genome for PCA, and it is of special relevance when masking ancestry-specific portions of the genome (see below). Then, PCA on a number k≤3 of ancestral populations is performed and the admixed population is projected into the determined k≤3 PCA space. PC loadings are used as weights in a weighted average of the allele values in a window of 40 SNPs. These haploid window scores are then used as observed values in a Hidden Markov Model (HMM) to assign posterior probabilities to the ancestry in each window (where chromosome were considered separately). Two complementary algorithms, Viterbi and forward-backward are used to compute posterior probabilities for each window. PCAdmix was implemented in C++ and is available at https://sites.google.com/site/pcadmix/. Additional performance testing and details of the implementation for this approach are available in (*20*-*22*).

The choice of k=3 ancestral populations for running PCAdmix was informed by ADMIXTURE results and is consistent with other investigations of ancestry in Latinos (Fig. 2B). Although

continental-level ancestral populations are a good model at k=3, we observed that PCAdmix performance was improved when including reference panels representing a diverse set of haplotypes. In Mexicans, we expect most of the ancestry variation to come from the Native American (NAT) component rather than the European (EUR) or African (AFR) components. To empirically test the performance of different NAT reference panels in our Mexican dataset, we run PCAdmix in a subset of 30 random samples using separately the different populations for which we had available trio data: Tepehuano (TEP), Nahua (NAH), and Maya (MYA). We limited to available trio data as PCAdmix takes phased data as input. When comparing the 3 different possible NAT ancestral populations we observed that comparable results were obtained when run separately. However, the proportion of windows called "unknown" was lower when using all three NAT populations combined. Therefore we constructed our reference panel by combining five trios from each NAT population (those five showing the highest proportions of NAT global ancestry, 15 trios total), plus 15 CEU, and 15 YRI trios as continental reference samples. We then separately run PCAdmix in two sets of admixed Mexican samples, the 23 complete MXL trios from HapMap3, and the 362 unrelated cosmopolitan samples from MGDP (N=312) and POPRES (N=50). The former set was trio phased using BEAGLE whereas the latter was population phased using phased MXL haplotypes as training set. Figure S11 shows a schematic diagram of the workflow to assign local ancestry and further analyze ancestry-specific fractions of the genome.

Local ancestry scan: We plotted Viterbi posterior probabilities per window against physical distance along autosomal chromosomes to identify peaks of ancestry enrichment across the genome. We limited to EUR and NAT ancestries since AFR ancestry values were based on much

lower number of counts, making deviations from the mean incomparable. The R package ggplot2 was used to visualize normalized ancestry proportions (Fig. S14).

Ancestry-specific PCA (ASPCA)

We implemented a modified version of the subspace PCA (ssPCA) method originally described by Raiko et al. (*23*) to handle the large amount of missing data resulting from masking ancestry-specific segments across the genome of multiple individuals. Previous implementations have adapted the same algorithm to genotype data (*24*). However, no method is currently available for applying subspace PCA to haplotype data. To project ancestry-specific haplotypes derived from the admixed genomes of Mexican cosmopolitan samples we restricted to individuals with more than 25% of their genomes inferred from each continental ancestry. Continental reference panels were constructed to project Native American and European blocks separately. Three populations (Seri, Lacandon, and Tojolabal) were excluded from the Native American panel due to evidence of extreme divergence compared to the rest of populations (and no NAT segments from admixed genomes were projected onto those clusters). The final panel consisted of 17 Native American parental populations. Our European reference panel included 1,387 POPRES individuals from throughout Europe with 4 grandparents from the same country (*3, 25*) plus 55 additional samples from Spain (*4*). We did not project AFR segments due to the low number of haplotypes across the population sample. To validate the consistency of our ASPCA results we performed a supervised structure analysis using *frappe* (*26*) and observed clustering patterns in agreement with our ancestry-specific distribution in PCA space. Our implementation of the method is described in what follows.

Overview of the ASPCA method (subspace learning algorithm): The method we describe here is a close adaptation of the *subspace learning algorithm* described in (*23*) to haplotype data. This implementation can be found in the software PCAmask, and mathematical details of the implementation can be found at http://arxiv.org/abs/1306.0558.

Ancestry-specific clustering analysis

We implemented a modified version of the *frappe* clustering algorithm (*26*) in order to accommodate partial missing resulting from masking specific sites of the genome. Our analyses of ancestry-specific segments of the genomes in the Mexican individuals rely on haplotype data. This leads to the generation of heterozygous missing sites at SNPs inferred to be heterozygous for the desired ancestry. Since the original *frappe* method developed by Tang et al. cannot process partially missing genotypes, we adapted the algorithm to process haplotype data. The algorithm relies on an EM algorithm to jointly infer overall ancestry proportions in admixed individuals and the ancestral allele frequencies at all sites used in the panel. While the standard *frappe* implementation integrates over the two observed alleles at every genotype, this integration is eliminated for haplotype data. Specifically, in the **M** step, an estimate for the ancestral allele frequencies is obtained from the best guess for ancestry proportions using the modified equation:

$$p_{mk}^{n+1} = \frac{\sum_{i \in O} h_{im} E_{imk}^{n}}{\sum_{i \in O} E_{imk}^{n}}$$

where p_{mk} is the allele frequency for ancestral population k at marker m, h_{im} is the observed allele on haplotype i (0/1-based), and O is the set of all haplotypes carrying the

desired ancestry at marker *m*. E_{imk} is a computational device indicating the expected ancestral contribution of ancestor *k* at haplotype *I* on marker *m*. Similarly, an estimate for the overall ancestral contribution q_{ik} of ancestral population *k* at haplotype *i* is obtained from:

$$q_{ik}^{n+1} = \frac{\sum_{m=1}^{M} E_{imk}^{n}}{\sum_{m=1}^{M}\sum_{i \in O} 1}$$

where the denominator simply corresponds to the total number of unmasked sites across all haplotypes used in the analysis. Finally, in the **E** step of the EM algorithm the quantity E_{imk} is updated based on the new estimates for overall ancestry proportion and estimated allele frequencies:

$$E_{imk}^{n+1} = \frac{p_{mk}^{n+1} q_{ik}^{n+1}}{\sum_{k'=1}^{K} p_{mk'}^{n+1} q_{ik'}^{n+1}}$$

This step is identical to the original version of the algorithm.

Biomedical associations with ASPCA values

We leveraged two studies of childhood asthma in Mexicans and Mexican Americans to determine important pulmonary associations with ancestry-specific PCA values. In particular, we focused on lung function as measured via spirometry using standard clinical measurements as ancestry has been shown previously to affect lung function (*27*). Both studies were trio-based ensuring long-range phase determination in the probands and were all of affected children. For continuous lung function measurements, we transformed raw spirometric values into percent

predicted values, which are already adjusted for typical anthropometric measurements (e.g. age, sex, and height) (*28*). Informed consent was obtained from all individuals at the study sites prior to sample collection. Both studies have been described in detail elsewhere. The genotypes included the same thresholds for quality control filtering, as described in (*29*). We briefly describe each study below.

The Genetics of Asthma in Latino Americans (GALA I) study is a trio-based study of Latinos (*30*) that was genotyped on the Affymetrix 6.0 array (*31, 32*). For this study we filtered to individuals sampled in Mexico City and the San Francisco Bay Area with 4 grandparents that all identified as Mexican or Mexican American. We used PCAdmix for local ancestry estimation with the same reference ancestral haplotypes as before, combined with global admixture modeling via ADMIXTURE (*33*). After filtering for individuals with spirometry data and adequate levels of Native American ancestry for use with ASPCA we were left with 68 individuals from Mexico City and 120 from the Bay Area.

The Mexico City Childhood Asthma Study (MCCAS) consists of trio-based sampling of individuals with asthma along with their parents, genotyped on the Illumina 550 platform (*34, 35*). All sampling was performed at a single site within Mexico City. As these samples were generated on an Illumina platform, we used the Native Mexican samples from the Human Genome Diversity Panel (*36*) combined with CEU and YRI genotypes, for local ancestry estimation using PCAdmix. We used global ancestry estimates from *frappe* (*26*) estimated previously (*34*). After filtering for individuals with spirometry and adequate levels of Native American ancestry we included 341 individuals in downstream analysis.

As the two datasets involved different numbers of SNPs and different numbers of individuals, we applied ASPCA independently to each dataset to minimize distorting the ASPCA values of the

reference individuals. To simplify comparisons across the two datasets, we used the normalized values of ASPCs 1 and 2, along with global ancestry covariates, to test for associations with population structure and lung function. As PCA, or ASPCA for that matter, is unitless, normalizing provides a standard for comparing across multiple ASPCA runs.

First, as GALA I includes individuals both from Mexico City and the San Francisco Bay Area, we wanted to investigate whether ASPCA values were associated with recruitment center. To do this, we used a likelihood ratio test of two different logistic regression models: a full model with ASPC1 & 2 along with global ancestry covariates; and a restricted model with simply the global ancestry terms. The statistic 2*log (likelihood ratio) then follows a 2-degree of freedom chi-squared distribution (one for each ASPC). We performed marginal tests for each ASPC using t-tests. We also estimated the raw AUC for a ROC curve including the two ASPCs using the *epicalc* package in R.

Next, for each study, we ran a separate robust linear model (rlm via *MASS* in R) to predict forced expiratory volume in the first second (FEV_1), using the ASPC values and adjusting for global ancestry covariates. We used robust linear models rather than OLS as PCA can have outliers that can potentially bias OLS estimation. Given normalized ASPC1 & 2 values, the regressions took the form:

$$\%(predicted)FEV_1 \sim \beta_0 + \beta_1 z(ASPC1) + \beta_2 z(ASPC2) + \beta_3 African + \beta_4 Native + \varepsilon$$

Where age, sex and height are incorporated in the percent predicted values to be able to compare effects across the entire growth curve. Global ancestry terms are used to adjust for any residual

population stratification, and to ensure that overall levels of Native American ancestry do not confound potential associations with ASPCs 1 and 2.

We performed these regressions separately for GALA I and MCCAS, then combined the effect sizes for ASPC1 and 2 via fixed effects meta-analysis in the R package *metafor*. These values were then used for p-value testing as they represented the largest combined sample and were independent replication with different recruiters, study designs, and genotyping arrays. We extrapolated based on the ASPCA values including GALA I to the data from 8 states to determine the change in FEV_1 due to differences in the origin of Native American ancestry. For context then we compared our observed results with that explained by change in lung function due to age (*28*) and African ancestry levels in African Americans (*27*).

We repeated these same analyses for two other values of lung function: forced vital capacity (FVC) and the FEV_1/FVC ratio, however, neither of these values were significantly associated with either ASPC1 or ASPC2 in any marginal test or meta-analysis and were not investigated further.

Haplotype sharing analysis

We used the densest dataset (785,663 SNPs) consisting of 674 unrelated samples genotyped on both Affymetrix 500K and Illumina 550K SNP arrays. This included a combined group of 71 Native Mexicans (Tepehuano n=20, Zapotec n=21, and Maya n=30), as well as another combined group of 312 Mexican cosmopolitan samples from the states of Guerrero (n=50), Guanajuato (n=48), Sonora (n=48), Tamaulipas (n=17), Veracruz (n=50), Yucatan (n=49), and Zacatecas (n=50). Sampling locations are reported in **Table S1**. To evaluate the level of

haplotype sharing with diverse populations from other regions of the world we also included a subset of HapMap continental reference samples. Namely, CEU (n=62), YRI (n=100), MXL (n=44), and CHB+JPT (n=85). Merged and curated genotype data were phased using BEAGLE software (*14, 15*). To phase the Mexican mestizo samples, we used the 22 MXL trios from HapMap3 as training set, whereas the Tepehuano and Maya trios were used to improve phasing of the Zapotec. Tepehuano (n=10 trios) and Maya (n=15 trios) were trio phased separately. HapMap populations with available trio data (CEU n=31 trios, YRI n=50 trios, and MXL n=22 trios) were also trio phased, whereas for CHB+JPT (n=85 unrelated individuals) we performed population phasing.

Genome-wide haplotype sharing (GWHS)**:** To determine the potential use of Mestizo and Native population data as reference for the genetic analysis of candidate regions and GWAS in Mexicans, we performed GWHS analysis using all available SNP genotypes within 100Kb fragments of the genome. We used BEAGLE phased genotype data and then estimated all plausible haplotypes within each segment across populations using PHASE (*37, 38*). GWHS was assessed by comparing the number of common haplotypes (with frequency >5% across populations) shared between Mexican Mestizos and the different HapMap populations as well as Native Mexicans (Fig. 4C).

The proportions shared between Mexicans and HapMap populations were comparable (SD from 1.4 to 3.0) across chromosomes. On average, Mexicans shared 21.6% with YRI, 54.8% with CHB+JPT, 59.3% with CEU, 78.6% with Natives and 81.2% with MXL. The proportion of shared haplotypes with CEU+CHB+JPT was 76.2%, and this was increased to 90.5% when the MXL group was added, and finally to 98.8% when Mexican Natives were included as reference (Fig. 4C). These results indicate more sharing than those previously reported (*1*) due to a higher

density of markers included in the analysis capturing more LD, and the availability of data from Native Mexicans.

Tag SNP selection efficiency in candidate regions: To determine the potential use of Mestizo and Native Mexican tagSNPs for targeted studies, 10 gene candidate regions were selected for containing SNPs previously associated to diseases or traits of clinical interest including, non-alcoholic fatty acid disease (PNPLA3), dyslipidemias (ABCA1), age-related macular degeneration (ARMS2), response to hepatitis C treatment (DDRGK1), Crohn's disease (NOD2), asthma (PTGDR, NOTCH4 and GC), metabolic syndrome (ApoB) and systemic lupus erythematosus (IKZF1). All genes are included in the Catalog of Published Genome-Wide Association Studies (http://www.genome.gov/gwastudies), two of them, ABCA1 (*39*) and PNPLA3 (*40*) house genetic variants that have been identified in Mexicans or Hispanic populations.

Across all populations analyzed we identified tag SNPs in these 10 candidate gene regions using Tagger, the tag SNP selection algorithm from Haploview software (*41*), with SNPs of frequency >5%, considering pairwise tagging only and r2 threshold of 0.8. We evaluated the performance of tag SNPs and their underlying coverage by estimating coverage from tagSNPs to the rest of the SNPs available in each gene using a pairwise r2 approach. In a similar fashion to the GWHS analysis, we evaluated the mean best r2 coverage based on the tag SNPs determined using various reference panels. Out of the 10 candidate loci, 2 had fewer than 10 SNPs and were dropped for this analysis, resulting in 8 genes evaluated using multiple reference ancestral groups. While the individual results vary from gene to gene, using the whole reference panel of Mexican Mestizos resulted in the best tagging performance overall, better than using the MXL population from HapMap3 (Fig. S15). The results of this analysis underline the importance of

using reference datasets of populations with the same LD structure for a better analysis of genetic variation in recently admixed populations such as Mexicans.

To search for a potential relationship between the enrichment in a particular ancestral component in the region with the haplotype sharing and tagging results, we analyzed the local ancestry estimations for each of the 10 regions included in this analysis. We did not find any clear relationship between local ancestry and proportion of shared haplotypes. Looking at more detail in the haplotype diversity observed in these regions we could identify that in those regions with the highest European or Native American ancestral contribution, corresponding respectively to *ABCA1* and *ARMS2*, this differential ancestry is not related with differences in haplotype diversity or tagging performance. In both cases, ancestral contribution differences are clearly related to differences in the frequency of specific haplotypes, that even if shared with all other populations, show distinct frequency differences in ancestral groups. The previous is shown in *ARMS2* for which all common haplotypes (>1%) present in either Mestizo and Native Mexican groups are shared in at least one HapMap group, but in which two of them are enriched in Native Mexicans (87%) and Mestizo (72%), compared to CEU (50%).

The results of the genome-wide and candidate region haplotype diversity showed that Mexican Native and Mestizo groups show a haplotype structure not fully represented in continental groups of the HapMap3 reference population set, which is comparable to other publicly available resources such as 1000 Genomes in terms of the Mexican diversity represented. Even including the closely related MXL population as reference, does not achieve the same effect than using the combined Mexican groups, the later most probably due to the fact that Mexican-Americans included in the MXL sample have a heterogeneous origin and thus a genetic structure of limited representation when compared to a comprehensive sample across the country. These results

support the fact that a deep genetic characterization and inclusion in association studies of recently admixed populations such as Mexicans represent a great opportunity to discover new genetic variation of relevance for biological traits and disease.

The selection of tag SNPs in candidate regions is of critical relevance for the improvement of genetic studies in Latin America, as this approach would enable the selection of small sets of SNPs for cost-effective study designs in candidate regions derived from GWAS or WGAS in other populations, with the aim of looking for new variants or haplotypes contributing to the genetic structure of biological traits or disease risk. Our results show that using the Mexican dataset generated here as reference population translates into a better haplotype capture than using SNP sets based on the use of combinations of population groups from currently available catalogs of variation.

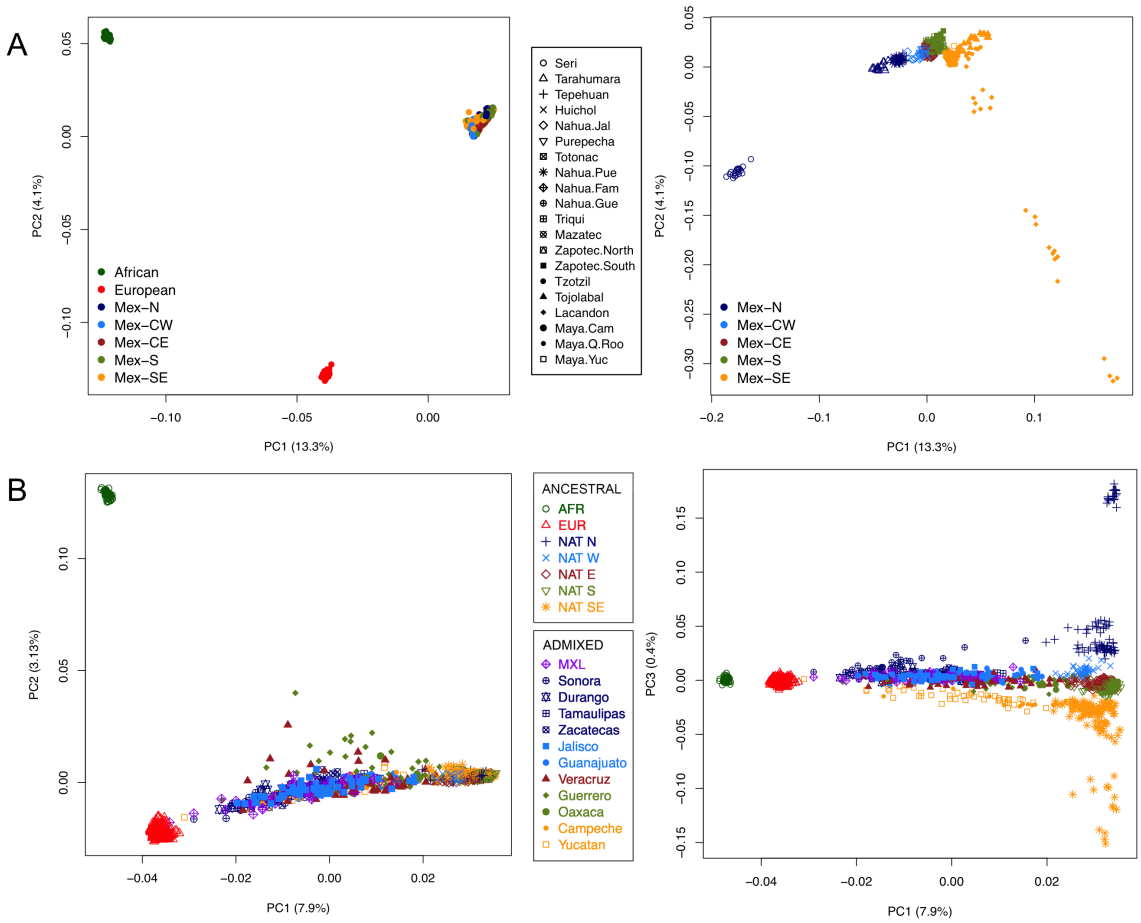


Figure S1: Principal component analyses based on the global dataset of ancestral and admixed Mexican populations. (**A**) *Left*: Global dataset of Native Mexicans combined with HapMap3 YRI African and CEU European samples. *Right*: Global dataset of Native Mexicans alone. (**B**) Combined dataset of ancestral reference samples (African, European, and Native Mexican) and admixed Mexican samples from cosmopolitan populations throughout Mexico and Mexican-Americans in the Los Angeles area. Populations are color-coded by geographic regions as follows: North (N), Central west (CW), Central east (CE), South (S), and Southeast (SE). *Left*: we observe a continuous dispersion of admixed individuals between the European and native Mexican cluster along PC1, reflecting their genome-wide average of native ancestry. PC2 separates a few individuals with higher African ancestry, predominantly from the coastal states of Veracruz and Guerrero. *Right*: along PC3, cosmopolitan samples from different states tend to

be separated by the different native clusters in a north-to-south direction. For example, Yucatan and Campeche individuals form an elongated cluster that is clearly pulled in the direction of the Mayan individuals. Likewise, those Sonora individuals with higher native proportions fall closer to northern native clusters. However, the separation is much more subtle among states from central Mexico, probably because standard PCA methods rely on genome-wide averaged signals from diploid genomes, making it difficult to ascertain finer scale patterns of differentiation.

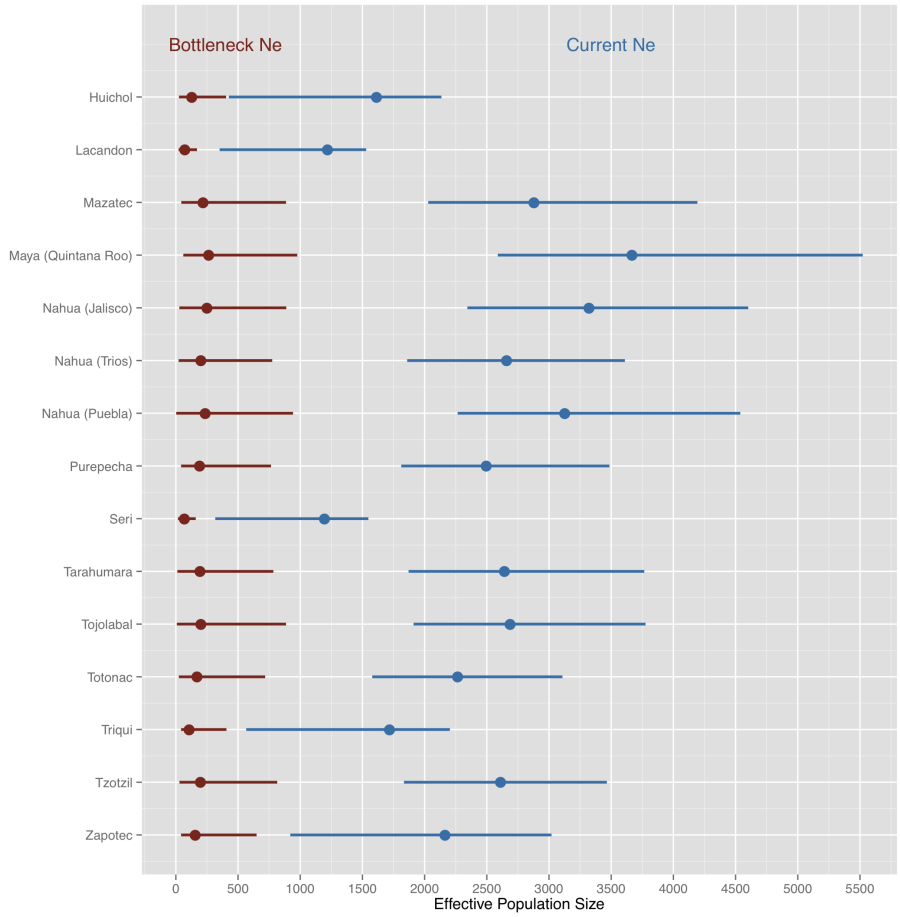


Figure S2: Summary of parameter estimates for the effective population size in different Native Mexican population samples. Estimated deme size during bottleneck and current Ne are given per population showing 95% confidence intervals. Parameters were estimated from cumulative runs of homozygosity (cROH) on chromosome 1 via a rejection algorithm comparing observed and simulated data with REJECTOR (see Methods for details). In order to use the maximum density of genotyped SNPs along chromosome 1, we restricted to Native Mexican populations for which Affymetrix 6.0 array data was available (see Table S1).

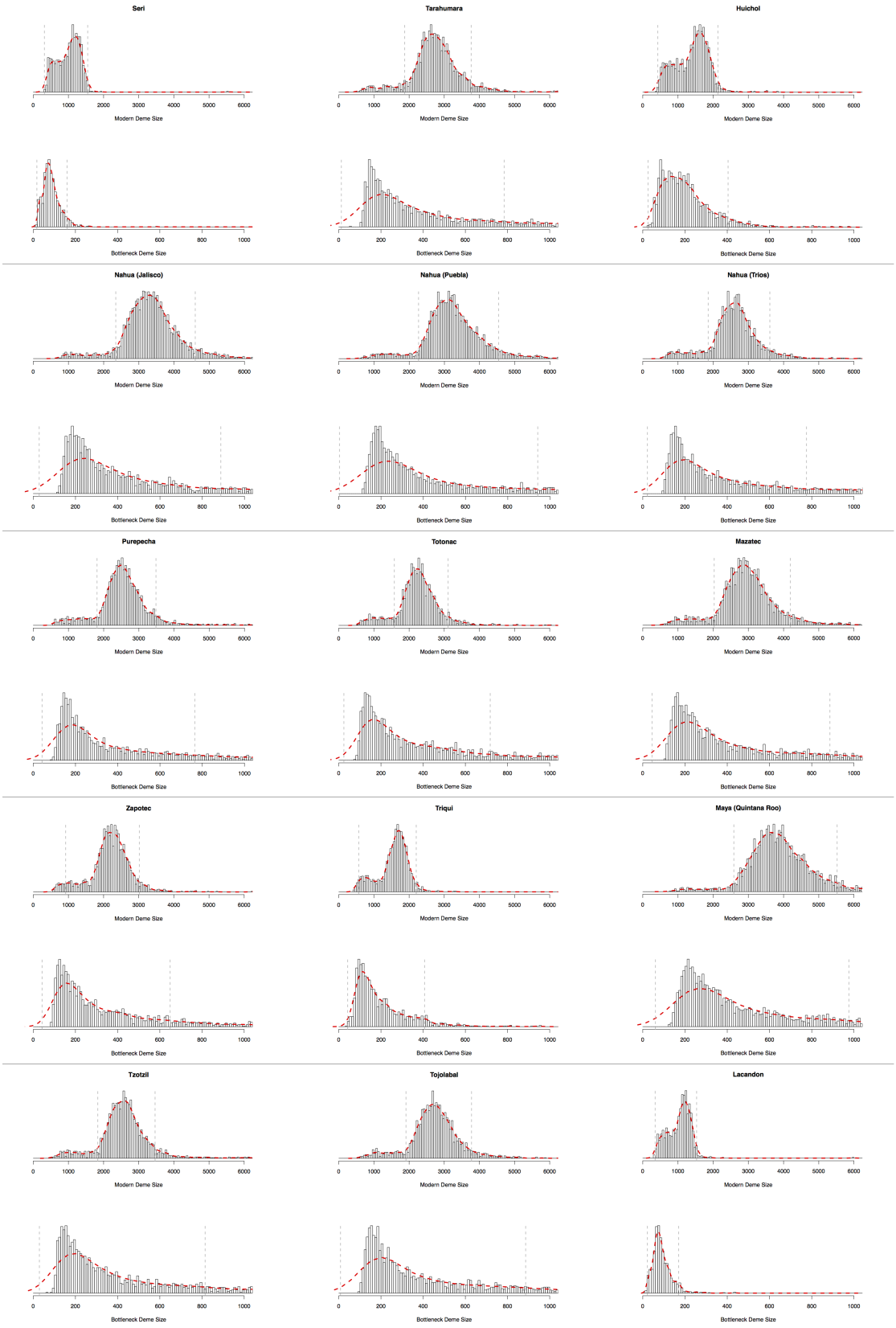

Seri
Tarahumara
Huichol
Modern Deme Size
Bottleneck Deme Size
Nahua (Jalisco)
Nahua (Puebla)
Nahua (Trios)
Purepecha
Totonac
Mazatec
Zapotec
Triqui
Maya (Quintana Roo)
Tzotzil
Tojolabal
Lacandon

Figure S3: Individual population profiles of the simulated posterior distribution of effective population sizes in different Native Mexican samples. For each population, contemporary Ne (top histogram) and bottleneck strength (bottom histogram), were estimated by sampling from a uniform distribution of Ne and keeping simulated parameters within 20% of the observed cROH with REJECTOR (see Methods). Each histogram shows the frequency of accepted simulations and the smoothed density values used for estimating the final parameters shown in Fig. S2.

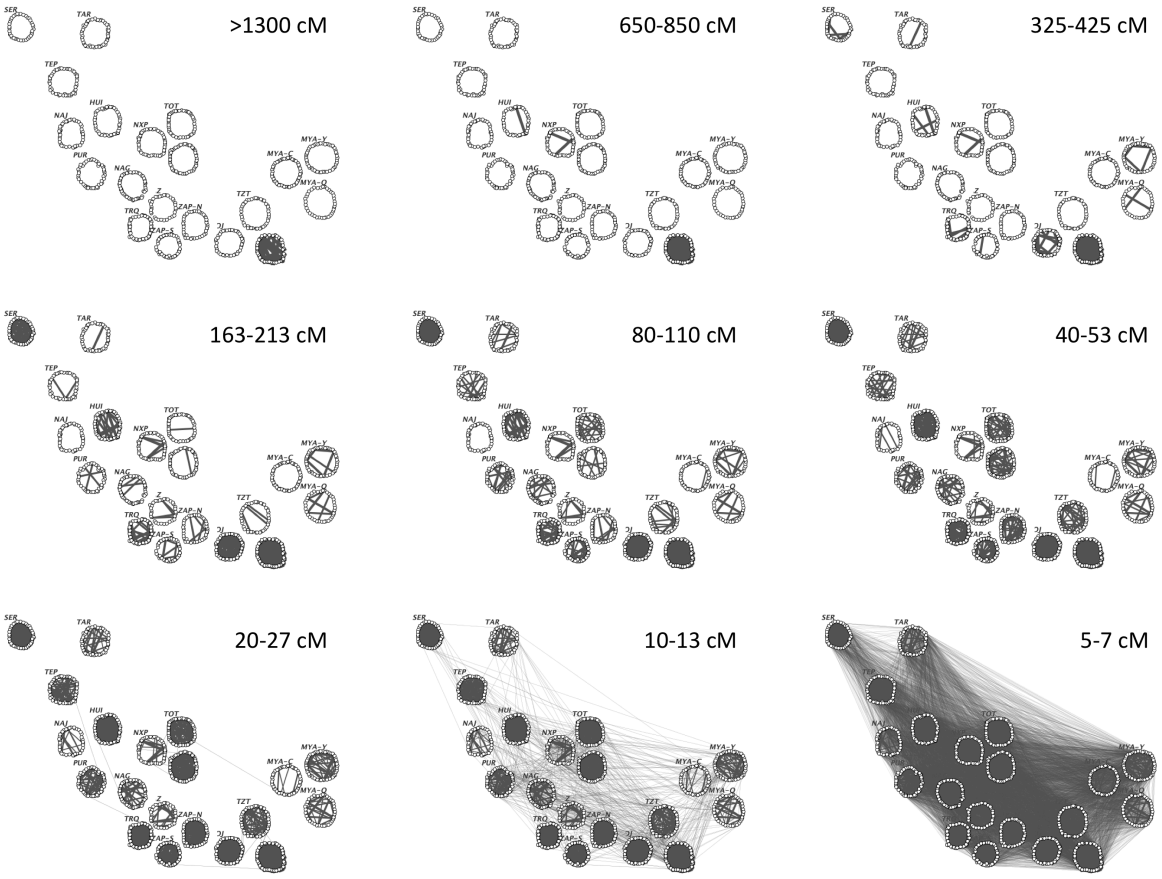


Figure S4: Patterns of relatedness within and between Native Mexican populations as measured by the total amount of segments identical-by-descent (IBD) shared between pairs of individuals. Each dot represents one individual and each line denotes a pairwise match between two individuals sharing more than a given amount of total IBD. Values of total IBD (in cM) were binned into consecutive categories corresponding to the following proportions of the genome: 50% and above, 25%, 12.5%, 6.75%, 3.37%, 1.69%, 0.85%, 0.42%, and 0.21%, which intend to reflect the first 9 degrees of relatedness. Each plot shows the network of connections resulting from each of these IBD thresholds. Specific bin ranges are indicated in cM next to each plot. In order to provide geographic context, individuals were arbitrarily placed in positions that

approximate the location of the sampled populations. The pattern across different populations shows high within-population sharing compared to between-populations for bins above 20 cM.

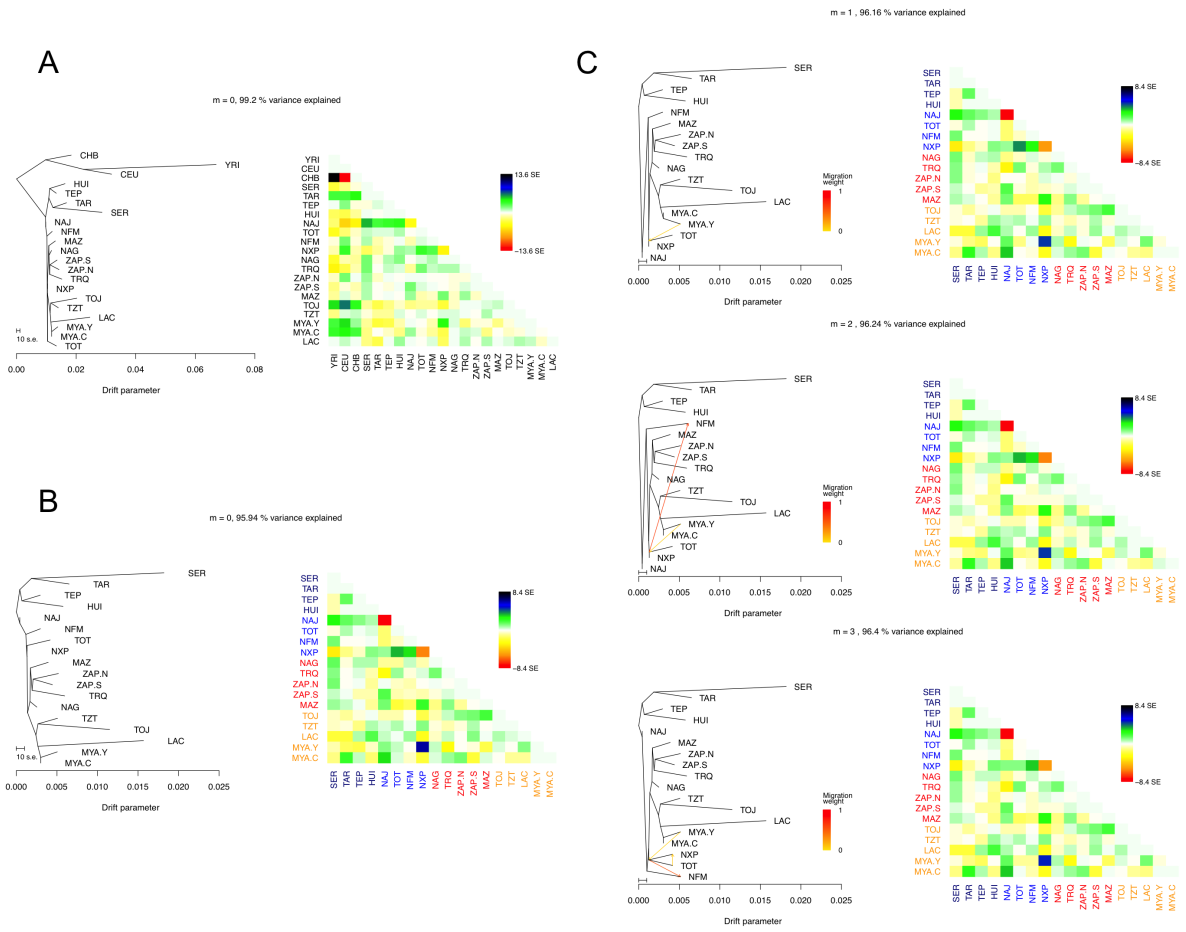


Figure S5: Maximum likelihood trees as inferred by TreeMix representing splitting patterns of Native Mexican populations and inferred migration events. (A) Graph depicting the relationships between Native Mexican populations along with three continental outgroups (HapMap YRI, CEU, and CHB). The length of the branches is proportional to the drift of each population. The resulting topology informed the position of the root in subsequent analyses (i.e., between all four Northern native populations and the rest). (B) TreeMix graph of Native Mexican populations alone without allowing for migration. The matrix next to each graph summarizes the residuals from the fit of the model to the data, where extreme values indicate populations that could be better modeled when adding migration to the model. (C) Models allowing for 1 to 3 events of

migration (m = 1 through 3). Trees were constructed using the known topology from B and including samples with more than 98% of Native American ancestry. Arrows indicate migration edges and directionality of gene flow. Color intensity is proportional to the inferred amount of gene flow according to the migration weight bar. Residuals for each model are presented in pairwise matrices next to each graph.

Figure S6: Genotype maps for the SNPs with the highest SPA score within each of the top 50 regions showing the most extreme allele frequency gradients across Native Mexican populations. Each circle represents one sample color-coded by its genotype. Each cluster represents one population with positions based on known latitude (y-axis) and longitude (x-axis) coordinates.

Some scattering was added to the position of individuals within each cluster to avoid overlap of samples sharing the same coordinates.

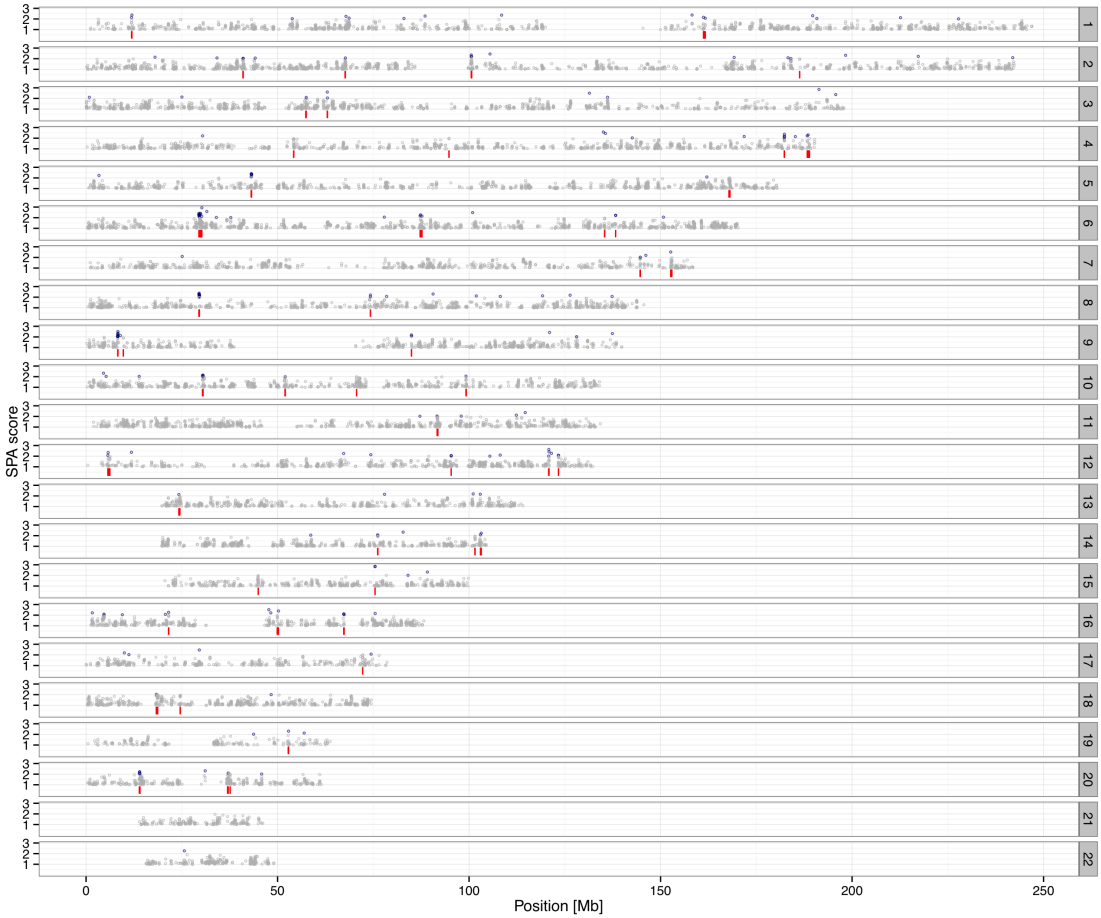


Figure S7: Genome scan for SNPs with steep allele frequency changes given the known location of Native Mexican populations. The y-axis represents the steepness of the slope for each SNP as measured by SPA scores, with values > 2 highlighted in blue. Candidate regions (red blocks) were identified by selecting the top 0.1% of SNPs of the empirical distribution, and subsequently merging SNPs separated by less than 500 kB into a single region. In order to avoid spurious outliers, we required that candidate regions have at least two outlier SNPs. The genomic annotation of the 50 candidate regions identified is summarized in Table S3.

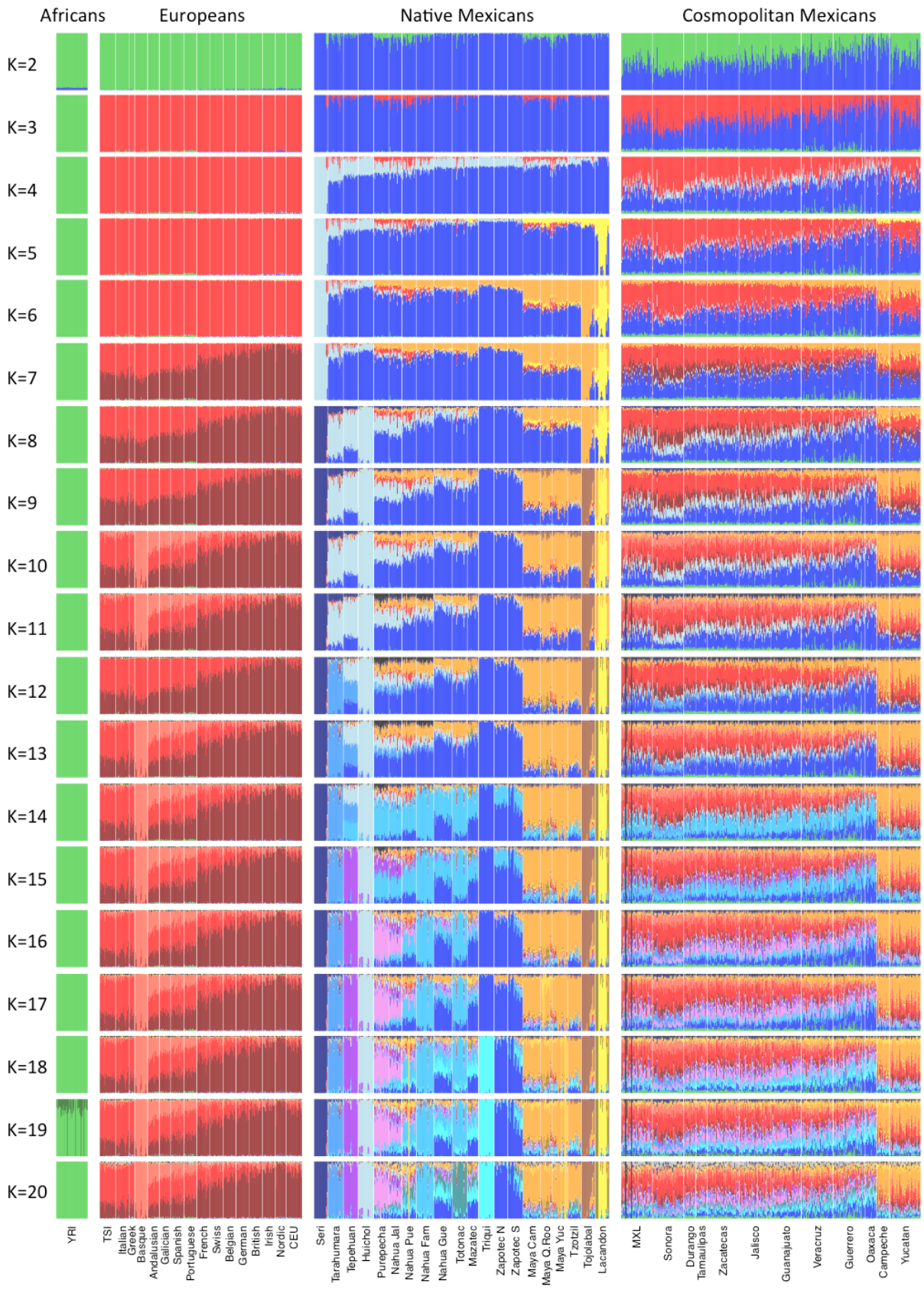

Africans
Europeans
Native Mexicans
Cosmopolitan Mexicans
K=2
K=3
K=4
K=5
K=6
K=7
K=8
K=9
K=10
K=11
K=12
K=13
K=14
K=15
K=16
K=17
K=18
K=19
K=20
YRI
TSI
Italian
Greek
Basque
Andalusian
Galician
Spanish
Portuguese
French
Swiss
Belgian
German
British
Irish
Nordic
CEU
Seri
Tarahumara
Tepehuan
Huichol
Purepecha
Nahua Jal
Nahua Pue
Nahua Fam
Nahua Gue
Totonac
Mazatec
Triqui
Zapotec N
Zapotec S
Maya Cam
Maya Q.Roo
Maya Yuc
Tzotzil
Tojolabal
Lacandon
MXL
Sonora
Durango
Tamaulipas
Zacatecas
Jalisco
Guanajuato
Veracruz
Guerrero
Oaxaca
Campeche
Yucatan

Figure S8: Unsupervised ADMIXTURE results from K=2 through 20 based on the intersection of Affymetrix and Illumina data (71,581 SNPs) from 1,282 samples (454 Native Mexicans, 469 Mexican mestizos, 309 Europeans, and 50 Yorubas). Each vertical bar represents an individual and the y-axis the proportion of the genome assigned to each of the ancestral clusters. Substantial substructure dominates the Native American component of both indigenous and cosmopolitan Mexican samples. European substructure is mainly driven by two sub-continental components following a North-South gradient, with the Basque clustering apart from the rest at K=10 and higher. We limited the representation of West Africans to a subset of HapMap YRI samples due to the study's emphasis on Native American diversity (see Table S1 for details).

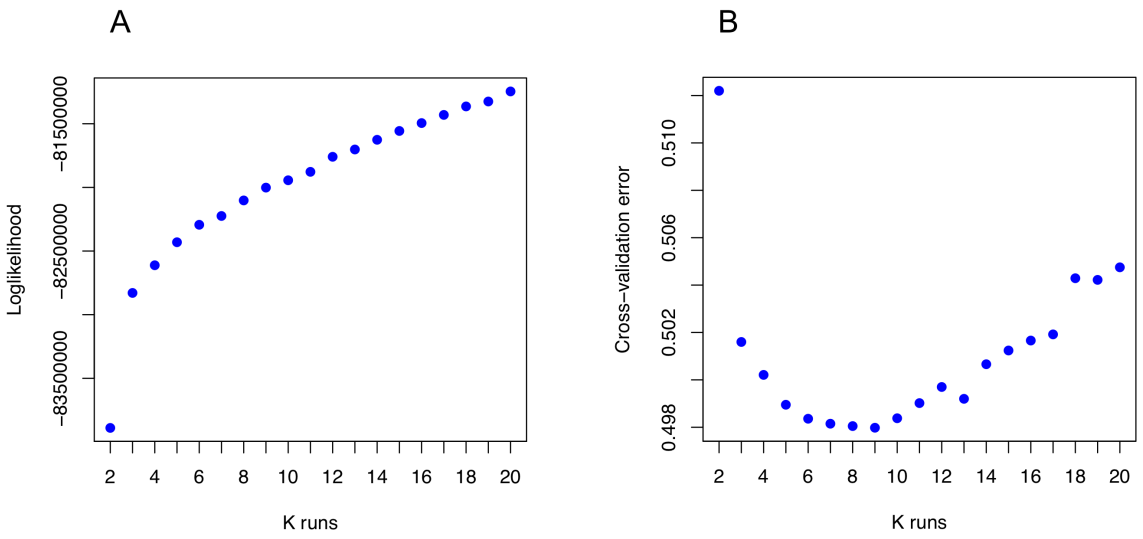


Figure S9: ADMIXTURE metrics at increasing K values based on Log-likelihoods (A) and cross-validation errors (B) for results shown in Fig. S8. While increasing clustering levels were associated with a continuous increase of likelihood values (*left*), K=9 showed the lowest error after cross validation (*right*).

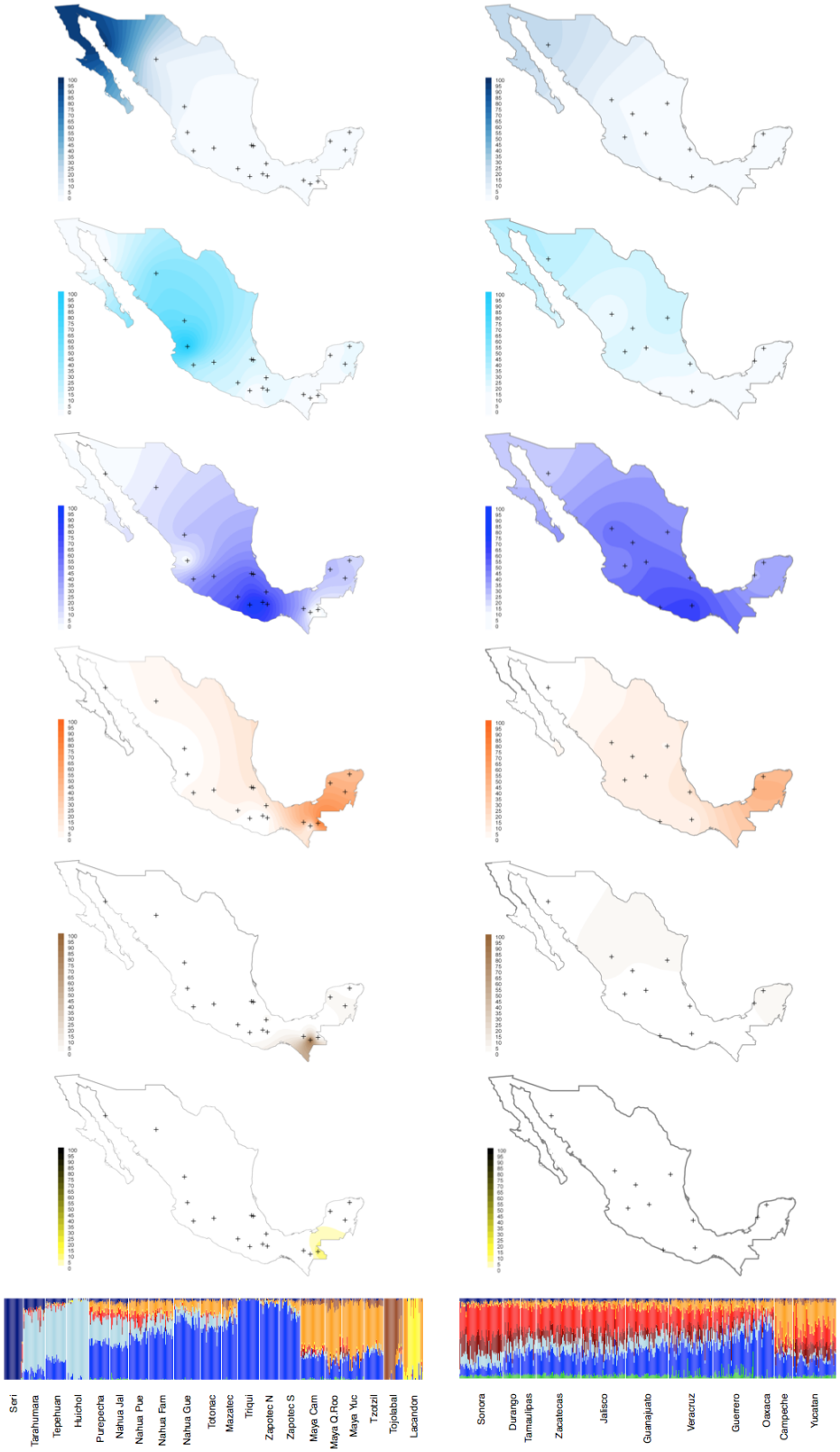


Figure S10: Spatial distribution of the major Native American components across Mexican

populations. Interpolation maps are shown for ADMIXTURE values at K=9 observed among indigenous (left column) and cosmopolitan (right column) samples. Black crosses on the maps of each column indicate sampling locations of indigenous and cosmopolitan populations, respectively. From top to bottom the six pairs of maps correspond to the six Native American components identified at K=9 (shown at the bottom and in Fig. S8). Contour maps were generated using Kriging interpolation methods, where intensities are proportional to ADMIXTURE values. For the group of cosmopolitan samples (thus with higher non-native admixture proportions), values were adjusted relative to the total Native American ancestry of each individual (see Methods for details).

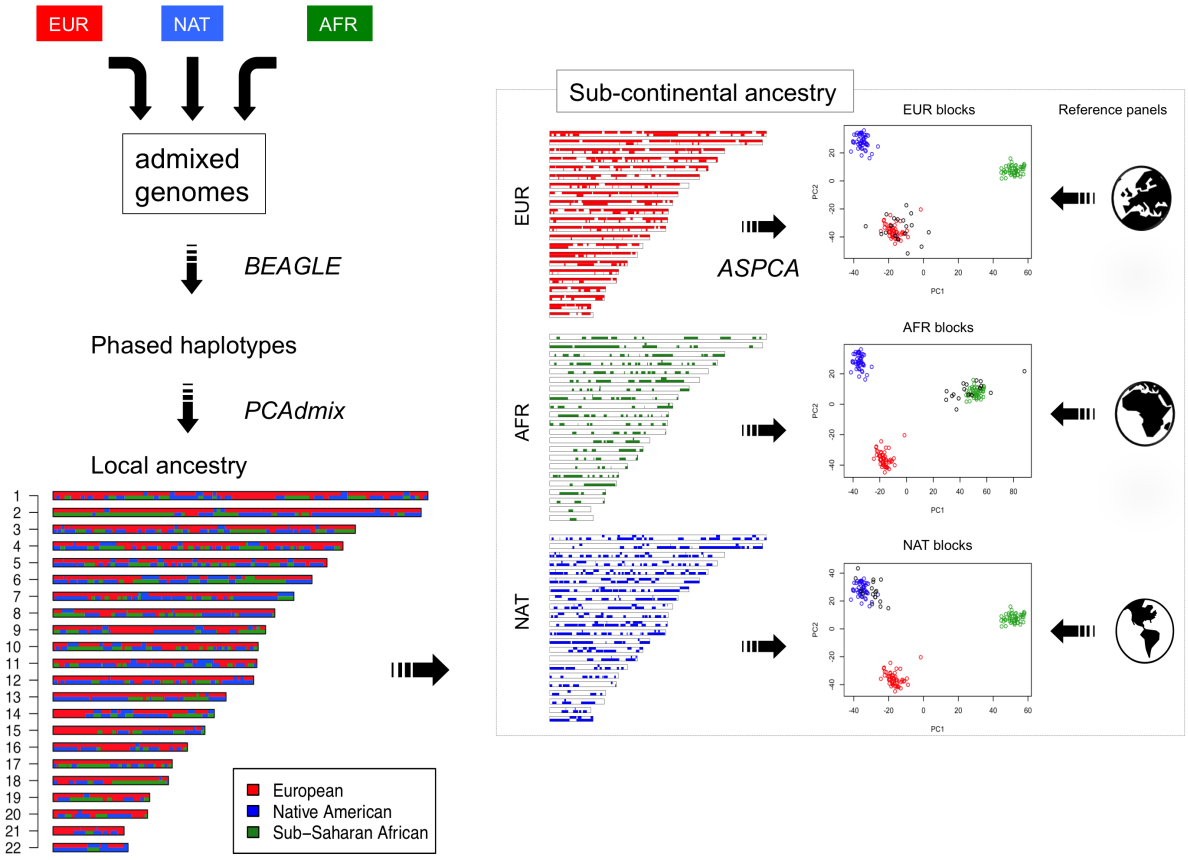


Figure S11: Diagram of the analytical strategy used for inferring sub-continental ancestry in admixed genomes. The starting point consists of genome-wide SNP data from admixed Mexican individuals. Unrelated individuals and family trios are population phased and trio phased, respectively, using BEAGLE. Next, phased haplotypes are used to estimate local ancestry along the genome using PCAdmix and continental reference samples. Then, taking Viterbi calls at each locus, ancestry-specific regions of the genome are masked to separately analyze European, African, and Native American haplotypes in a PCA framework together with large sub-continental reference panels of putative ancestral populations (see Methods for details). We refer to this methodology as ancestry-specific PCA (ASPCA) and the code is packaged into the software PCAmask. Additional details available at Moreno-Estrada et al. (arxiv.org/abs/1306.0558).

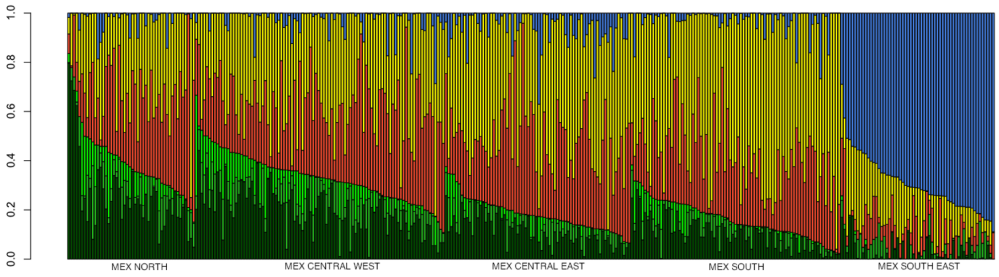


Figure S12: Supervised ancestry-specific clustering analysis of Native American haplotypes derived from admixed cosmopolitan Mexican genomes. On the x-axis bars represent haploid genomes for all admixed individuals with >25% of global Native American ancestry, that is, one individual is usually represented by two bars. The y-axis indicate native ancestry proportions at K=6 using our reference panel of Native Mexican populations (see Table S1). Given the low overall contribution of isolated native components into the mestizo population (as identified in Fig. 2), we excluded Seri, Lacandon, and Tojolabal from the reference panel. Since our ancestry-specific approach relies on haplotype data, we used a modified version of the FRAPPE algorithm to estimate admixture proportions in the presence of missing sites at SNPs inferred to be heterozygous for the desired ancestry (see Methods). Individuals are grouped into regions as

described in Table S1. Because we required more than 25% of Native American ancestry to be included in the analysis, some regions are represented by less individuals that the actual sample size, such as mestizo individuals from Northern states of Mexico, where overall proportions of Native American ancestry are considerably lower than in the rest of the territory. The six clusters identified to run the algorithm on supervised mode were: Northern Native Mexicans, Huichol (which clustered on their own in previous analyses), Native Mexicans from Central West, Central East, South, and Southeast Mexico (excluding Seri, Lacandon, and Tojolabal). Overall, the results replicate the observations from our ASPCA analysis: on average, Mexicans sampled from different regions of Mexico derive differential ancestral contributions from each of the Native American components.

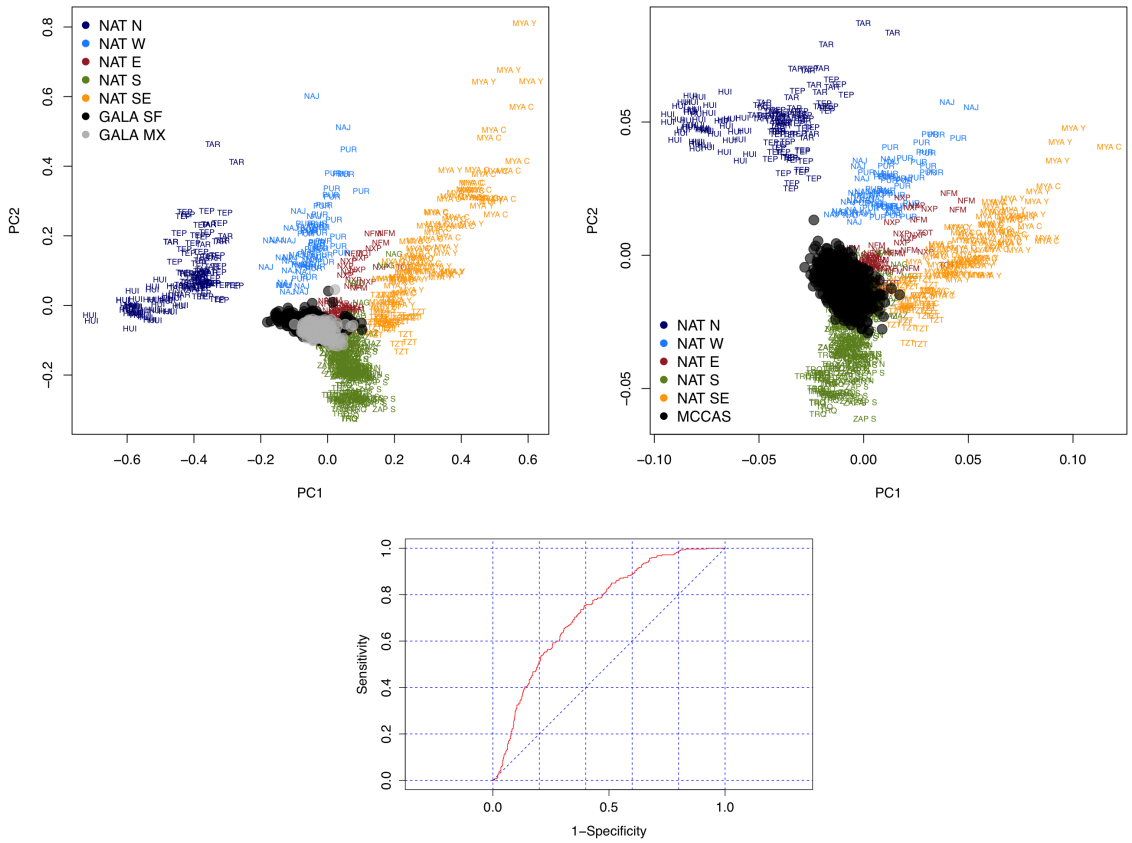


Figure S13: ASPCA analysis of Native American segments from Mexican participants of the GALA I study (*left*) sampled in Mexico City (GALA MX, gray circles) and the San Francisco bay area (GALA SF, black circles), and participants of the MCCAS study (*right*) sampled in Mexico City (black circles), analyzed together with our dataset of 20 indigenous Mexican populations (labeled by population identifier and color-coded by region of origin). Samples with >10% of non-native admixture were excluded from the reference panel as well as population outliers such as Seri, Lacandon, and Tojolabal. Here, a total of 803 phased haploid genomes (280 MX and 523 SF) represent the GALA Mexican sample and 1900 the MCCAS cohort. Bottom: ROC curve for the logistic regression of ASPCA values separating Mexico City (MX) versus San Francisco (SF) cases from the GALA I study (see main text for details).

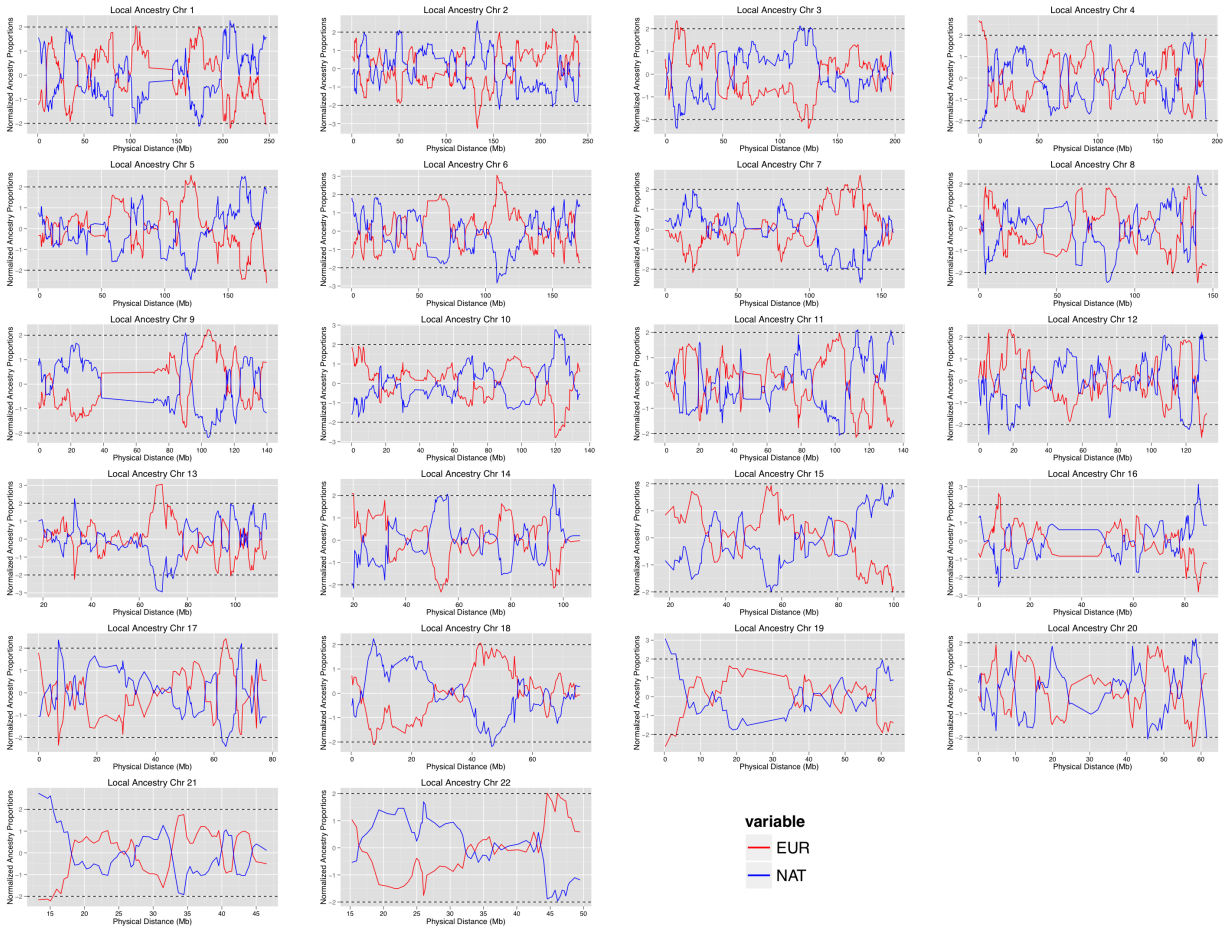


Figure S14: Local ancestry scan in the combined set of cosmopolitan Mexican samples showing normalized Z scores of Native American versus European ancestry proportions along autosomal chromosomes. African ancestry not shown due to the small sample size of African haplotypes across individuals. Local ancestry calls were estimated using PCAdmix and counts were scaled to the total sample size. Dashed lines indicate two standard deviations away from the mean. Results are based on 372,692 SNPs and 362 samples with available Affymetrix data (see Table S1).

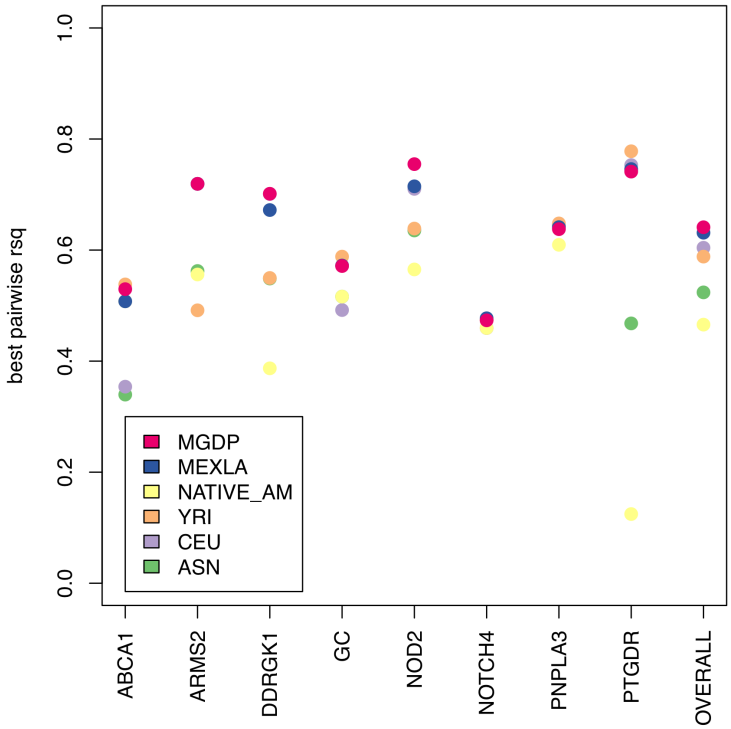


Figure S15: Tagging efficiency using Mexican Mestizos or HapMap Populations as reference. The mean best R2 coverage based on the tag SNPs determined using various reference panels was evaluated in a subset of candidate gene regions of biomedical interest. While the individual results vary from gene to gene, using the whole reference panel of Mexican Mestizos from resulted in the best tagging performance overall, notably, better than using the MXL population from HapMap3.

Table S1. Summary data for 32 Mexican populations and continental reference panels included

Population	Pop ID	N (initial)	Filtered	N (final)	Region*	Latitude	Longitude	Linguistic Family	Study/source	Data
NATIVE MEXICANS										
Seri	SER	25	4	21	North Mex	29.00	-112.15	Serian	This (Stanford/UCSF)	Affymetrix 6.0
Tarahumara	TAR	25	1	24	North Mex	27.75	-107.17	UtoAztecan	This (Stanford/UCSF)	Affymetrix 6.0
Tepehuano	TEP	30	7	23	North Mex	23.48	-104.39	UtoAztecan	This (INMEGEN)	Affymetrix 500K/Illumina 550K
Huichol	HUI	24	0	24	North Mex	21.17	-104.08	UtoAztecan	This (Stanford/UCSF)	Affymetrix 6.0
Nahua (Jalisco)	NAJ	23	3	20	Central-West Mex	19.50	-103.50	UtoAztecan	This (Stanford/UCSF)	Affymetrix 6.0
Purepecha	PUR	23	0	23	Central-West Mex	19.75	-101.50	Tarascan	This (Stanford/UCSF)	Affymetrix 6.0
Totonac	TOT	25	2	23	Central-East Mex	20.00	-97.80	Totonacan	This (Stanford/UCSF)	Affymetrix 6.0
Nahua (Puebla)	NXP	25	3	22	Central-East Mex	19.97	-97.62	UtoAztecan	This (Stanford/UCSF)	Affymetrix 6.0
Nahua (Trios)	NFM	41	14	27	Central-East Mex	19.93	-97.62	UtoAztecan	This (Stanford/UCSF)	Affymetrix 6.0
Nahua (Guerrero)	NAG	29	0	29	South Mex	17.89	-99.13	UtoAztecan	Mao et al. 2007	Affymetrix 500K
Triqui	TRQ	25	1	24	South Mex	17.18	-97.95	Otomanguean	This (Stanford/UCSF)	Affymetrix 6.0
Zapotec (North)	ZAP.N	21	0	21	South Mex	17.41	-96.69	Otomanguean	This (INMEGEN)	Affymetrix 500K/Illumina 550K
Zapotec (South)	ZAP.S	24	1	23	South Mex	17.23	-96.23	Otomanguean	This (Stanford/UCSF)	Affymetrix 6.0
Mazatec	MAZ	17	0	17	South Mex	18.33	-96.33	Otomanguean	This (Stanford/UCSF)	Affymetrix 6.0
Tzotzil	TZT	22	1	21	Southeast Mex	16.83	-92.67	Mayan	This (Stanford/UCSF)	Affymetrix 6.0
Tojolabal	TOJ	22	1	21	Southeast Mex	16.50	-92.00	Mayan	This (Stanford/UCSF)	Affymetrix 6.0
Lacandon	LAC	22	0	22	Southeast Mex	16.75	-91.25	Mayan	This (Stanford/UCSF)	Affymetrix 6.0
Maya (Quintana Roo)	MYA.Q	19	1	18	Southeast Mex	19.58	-88.58	Mayan	This (Stanford/UCSF)	Affymetrix 6.0
Maya (Campeche)	MYA.C	45	18	27	Southeast Mex	20.37	-90.05	Mayan	This (INMEGEN)	Affymetrix 500K/Illumina 550K
Maya (Yucatan)	MYA.Y	24	0	24	Southeast Mex	21.17	-88.14	Mayan	Mao et al. 2007	Affymetrix 500K
TOTAL NAT	20 groups	511	57	454						
COSMOPOLITAN MEXICANS										
Mexican-Americans	MXL	80	31	49	LA, California	34.08	-118.17	-	HapMap3	Affymetrix 6.0
Mexican from Sonora	SON	49	0	49	North Mex	29.07	-110.94	-	This (INMEGEN)	Affymetrix 500K/Illumina 550K
Mexican from Durango	DUR	19	0	19	North Mex	24.06	-104.66	-	This (INMEGEN)	Illumina 550K
Mexican from Tamaulipas	TAM	17	0	17	North Mex	23.74	-99.14	-	This (INMEGEN)	Affymetrix 500K/Illumina 550K
Mexican from Zacatecas	ZAC	50	0	50	North Mex	22.79	-102.59	-	This (INMEGEN)	Affymetrix 500K/Illumina 550K
Mexican from Jalisco	JAL	50	0	50	Central-West Mex	20.67	-103.35	-	POPRES	Affymetrix 500K
Mexican from Guanajuato	GUA	48	0	48	Central-West Mex	21.01	-101.26	-	This (INMEGEN)	Affymetrix 500K/Illumina 550K
Mexican from Veracruz	VER	50	0	50	Central-East Mex	19.57	-96.90	-	This (INMEGEN)	Affymetrix 500K/Illumina 550K
Mexican from Guerrero	GUE	50	0	50	South Mex	16.88	-99.87	-	This (INMEGEN)	Affymetrix 500K/Illumina 550K
Mexican from Oaxaca	OAX	18	0	18	South Mex	17.06	-96.72	-	This (INMEGEN)	Illumina 550K
Mexican from Campeche	CAM	20	0	20	Southeast Mex	19.84	-90.53	-	This (INMEGEN)	Illumina 550K
Mexican from Yucatan	YUC	49	0	49	Southeast Mex	20.98	-89.63	-	This (INMEGEN)	Affymetrix 500K/Illumina 550K
TOTAL MEX	12 groups	500	31	469						
EUROPEANS										
European-Americans	CEU	25	0	25	Northern Europe	-	-	-	HapMap3	Affymetrix 6.0
Tuscan	TSI	25	0	25	Italy (Tuscany)	-	-	-	HapMap3	Affymetrix 6.0
Andalusian	AND	20	2	18	Spain (Andalusia)	-	-	-	Rodriguez et al. 2012	Affymetrix 6.0
Galician	GAL	17	0	17	Spain (Galicia)	-	-	-	Rodriguez et al. 2012	Affymetrix 6.0
Basque	BAS	20	0	20	Spain (Basque Country)	-	-	-	Rodriguez et al. 2012	Affymetrix 6.0
Portuguese	PT	20	0	20	Europe SW	-	-	-	POPRES	Affymetrix 500K
Spanish	ES	20	0	20	Europe SW	-	-	-	POPRES	Affymetrix 500K
Italian	IT	20	0	20	Europe S	-	-	-	POPRES	Affymetrix 500K
Greek	GR	8	0	8	Europe SE	-	-	-	POPRES	Affymetrix 500K
French	FR	20	0	20	Europe W	-	-	-	POPRES	Affymetrix 500K
Swiss	CH	20	0	20	Europe W	-	-	-	POPRES	Affymetrix 500K
Belgian	BE	20	0	20	Europe W	-	-	-	POPRES	Affymetrix 500K
German	DE	20	0	20	Europe C	-	-	-	POPRES	Affymetrix 500K
British	GB	20	0	20	Europe NW	-	-	-	POPRES	Affymetrix 500K
Irish	IE	20	0	20	Europe NW	-	-	-	POPRES	Affymetrix 500K
Scandinavian**	SC	16	0	16	Europe NNE	-	-	-	POPRES	Affymetrix 500K
TOTAL EUR	16 groups	311	2	309						
AFRICANS										
Yoruba	YRI	50	0	50	West Africa	-	-	-	HapMap3	Affymetrix 6.0
TOTAL SUM	49 groups	1372	90	1282						

* Mexicans were grouped according to major geographic areas whereas Europeans from POPRES according to the classification in Auton et al. 2009
** Includes individuals from Sweden, Norway, Denmark, and Finland (n=11, 3, 1, 1, respectively)

Table S2. Three working datasets generated for this study

Name	Samples	SNPs	Average call rate	Notes
global.illu.affy.unrel	1,282	71,581	98.93%	All samples as reported in Table S1
global.affy.unrel	1,224	372,692	98.85%	All samples with available Affymetrix data
mex.hapmap.unrel	674	785,663	99.30%	Samples with both Affymetrix and Illumina data

Note: The *unrel* suffix denotes that all individuals being part of the offspring of trios or duos have been removed.

Table S3. Top 0.1% values of the distribution of SPA scores from the combined dataset of Native Mexican populations

Chr	Start[a]	End	Width	Best SNP[b]	Min p-value	Genes	Best gene[c]	Dist. to best gene
1	11872557	11968650	96094	rs2336384	8.52E-05	KIAA2013, PLOD1, MFN2	MFN2	0
1	161201354	161675875	474522	rs12035861	0.000296765	RGS4, RGS5, NUF2	C1orf110	96125
2	40965587	41032954	67368	rs17026994	0.000505381		SLC8A1	439875
2	67630022	67696704	66683	rs6546319	0.00044074		ETAA1	205667
2	100567880	100688625	120746	rs7594990	0.000102839		PDCL3	8468
2	186314170	186316584	2415	rs9646802	0.00072869	FSIP2	FSIP2	0
3	57369806	57483052	113247	rs7612786	0.000370221	DNAH12	DNAH12	0
3	62970589	63023776	53188	rs2367592	2.06E-05		LOC285401	92815
4	54185643	54240061	54419	rs17083248	0.000781578		LNX1	60000
4	94741372	94769749	28378	rs1435481	0.000658171	GRID2	GRID2	0
4	182295342	182339600	44259	rs10021323	7.64E-05	LINC00290	LINC00290	22304
4	188288936	188887227	598292	rs10007836	0.000138098	LOC339975	LOC339975	0
5	43106913	43208653	101741	rs782978	6.76E-05	ZNF131	ZNF131	0
5	167854709	168124390	269682	rs2305729	0.000634665	RARS, FBLL1, PANK3, MIR103A1, MIR103B1, SLIT3	RARS	0
6	29467091	30222934	755844	rs757262	5.88E-06	OR12D2, OR11A1, OR10C1, OR2H1, MAS1L, LOC100507362, UBD, SNORD32B, OR2H2, GABBR1, MOG, ZFP57, HLA-F, HLA-F-AS1, IFITM4P, HCG4, LOC554223, HLA-G, HLA-H, HCG4B, HLA-A, HCG9, ZNRD1-AS1, HLA-J, ZNRD1, PPP1R11, RNF39, TRIM31, TRIM40	TRIM40	0
6	87217610	87706172	488563	rs33993504	0.000167481	HTR1E	MIR548AD	236583
6	135376293	135416925	40633	rs1014021	0.000808023	HBS1L	HBS1L	0
6	138271512	138272800	1289	rs6570193	0.00020274		TNFAIP3	26658
7	144649450	144715640	66191	rs700260	0.000564147		TPK1	551561
7	152649615	152947294	297680	rs367566	3.23E-05		ACTR3B	466219
8	29469355	29592864	123510	rs6558130	7.93E-05		C8orf75	109135
8	74233577	74249632	16056	rs4415303	0.000232123		LOC100130301	66581
9	8256584	8323100	66517	rs10815823	3.53E-05	PTPRD	PTPRD	0
9	9716233	9741606	25374	rs1768892	0.000937306	PTPRD	PTPRD	0

9	84954051	84977207	23157	rs17086018	0.000252691		RASEF	86188
10	30395199	30537672	142474	rs639390	0.000308518		MTPAP	120140
10	51939492	52023192	83701	rs3011707	0.000611159	SGMS1	SGMS1	0
10	70630454	70635003	4550	rs7091695	0.000664048	SUPV3L1	SUPV3L1	0
10	99194516	99269150	74635	rs7081796	0.000475999	EXOSC1, ZDHHC16, MMS19, UBTD1	EXOSC1	0
11	91664434	91856736	192303	rs922006	0.000528888	FAT3	FAT3	59399
12	5690939	6215341	524403	rs3782652	0.000105778	ANO2, VWF, CD9	ANO2	0
12	95329431	95333800	4370	rs11108495	0.000467184		CDK17	13551
12	120774822	120869596	94775	rs895960	1.76E-05	HPD, PSMD9, WDR66	PSMD9	0
12	123332538	123356470	23933	rs2342468	0.000402542	ZNF664-FAM101A, FAM101A	ZNF664-FAM101A	0
13	24226680	24496118	269439	rs12019525	0.000332024	RNF17, CENPJ, TPTE2P1	RNF17	9621
14	76137790	76187720	49931	rs1642822	0.000431925		ESRRB	99857
14	101526284	101565047	38764	rs4906175	0.000775702	DYNC1H1	DYNC1H1	0
14	102971177	103183235	212059	rs11160756	0.000193925	MARK3, CKB, TRMT61A, BAG5, APOPT1, KLC1	KLC1	0
15	44942157	44943871	1715	rs7177307	0.000699307	MIR548A3	MIR548A3	0
15	75444737	75457456	12720	rs1565757	8.81E-06	PEAK1	PEAK1	0
16	21538199	21573398	35200	rs16972765	0.000179234	METTL9, IGSF6	METTL9	0
16	49909748	50235764	326017	rs7204626	6.46E-05		LOC388276	382001
16	67284570	67372449	87880	rs6499193	0.000356999	CDH3, CDH1	CDH3	0
17	72212760	72229750	16991	rs11868472	0.000628789	MXRA7,JMJD6	MXRA7	0
18	18338350	18707615	369266	rs607660	0.000546517		CTAGE1	131796
18	24563233	24611667	48435	rs356925	0.000899109		CDH2	570163
19	52785167	52869070	83904	rs2098718	0.000141037	GLTSCR1	GLTSCR1	0
20	13911728	14073728	162001	rs6105211	0.000223308	SEL1L2,MACROD2	MACROD2	0
20	36944868	37107381	162514	rs6129158	0.000376098	PPP1R16B,FAM83D,DHX35	DHX35	5601
20	37613066	37625204	12139	rs6129366	0.000661109		LOC339568	338399

[a]Distances are given in base pairs (bp) and positions map to the human genome build hg18
[b]Best SNP is the SNP with the highest SPA score within each region
[c]Best gene is the closest gene to the SNP with the highest SPA score within each region

References

1. I. Silva-Zolezzi *et al.*, Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proceedings of the National Academy of Sciences of the United States of America*, (Jun 11, 2009).

2. X. Mao *et al.*, A genomewide admixture mapping panel for Hispanic/Latino populations. *Am J Hum Genet* **80**, 1171 (Jun, 2007).

3. M. R. Nelson *et al.*, The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *American journal of human genetics* **83**, 347 (Sep 01, 2008).

4. L. R. Botigue *et al.*, Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc Natl Acad Sci U S A*, (Jun 3, 2013).

5. A. L. Price *et al.*, Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904 (Aug 01, 2006).

6. B. S. Weir, C. C. Cockerham, Estimating F-Statistics for the Analysis of Population-Structure. *Evolution* **38**, 1358 (1984).

7. B. S. Weir, W. G. Hill, Estimating F-statistics. *Annual Review of Genetics* **36**, 721 (2002).

8. H. Wickham, in *Use R!* (Springer,, New York, 2009), pp. viii, 212 p.

9. M. Jobin, J. Mountain, REJECTOR: Software for Population History Inference from Genetic Data via a Rejection Algorithm. *Bioinformatics (Oxford, England)*, (Oct 20, 2008).

10. A. Auton *et al.*, Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome research*, 1 (Mar 13, 2009).

11. M. A. Nalls *et al.*, Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. *PLoS genetics* **5**, e1000415 (Apr 01, 2009).

12. B. M. Henn *et al.*, Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences* **108**, 5154 (Apr 29, 2011).

13. G. K. Chen, P. Marjoram, J. D. Wall, Fast and flexible simulation of DNA sequence data. *Genome Res* **19**, 136 (Jan, 2009).

14. S. Browning, B. Browning, Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies …. *The American Journal of Human Genetics*, (2007).

15. S. R. Browning, Missing data imputation and haplotype phase inference for genome-wide association studies. *Human Genetics* **124**, 439 (Dec 01, 2008).

16. A. Gusev *et al.*, Whole population, genome-wide mapping of hidden relatedness. *Genome research* **19**, 318 (Mar 01, 2009).

17. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967 (2012).

18. W. Y. Yang, J. Novembre, E. Eskin, E. Halperin, A model-based approach for analysis of spatial structure in genetic data. *Nat Genet* **44**, 725 (Jun, 2012).

19. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655 (Sep 01, 2009).

20. A. Brisbin *et al.*, PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Hum Biol* **84**, 343 (Aug, 2012).

21. B. M. Henn *et al.*, Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* **8**, e1002397 (Jan, 2012).

22. J. M. Kidd *et al.*, Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *Am J Hum Genet* **91**, 660 (Oct 5, 2012).

23. T. Raiko, A. Ilin, J. Karhunen, Principal component analysis for large scale problems with lots of missing values. *Machine Learning: ECML 2007*, 691 (2007).

24. N. A. Johnson *et al.*, Ancestral components of admixed genomes in a mexican cohort. *PLoS genetics* **7**, e1002410 (Dec, 2011).

25. J. Novembre *et al.*, Genes mirror geography within Europe. *Nature* **456**, 98 (Nov 06, 2008).

26. H. Tang, J. Peng, P. Wang, N. J. Risch, Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology* **28**, 289 (Jun, 2005).

27. R. Kumar *et al.*, Genetic ancestry in lung-function predictions. *New England Journal of Medicine* **363**, 321 (Jul 22, 2010).

28. J. L. Hankinson, J. R. Odencrantz, K. B. Fedan, Spirometric reference values from a sample of the general U.S. population. *Am J Respir Crit Care Med* **159**, 179 (Jan, 1999).

29. D. G. Torgerson *et al.*, Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nature Genetics* **43**, 887 (Sep, 2011).

30. E. G. Burchard *et al.*, Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. *Am J Respir Crit Care Med* **169**, 386 (Feb 1, 2004).

31. J. M. Galanter *et al.*, Cosmopolitan and ethnic-specific replication of genetic risk factors for asthma in 2 Latino populations. *The Journal of allergy and clinical immunology* **128**, 37 (Jul, 2011).

32. D. G. Torgerson *et al.*, Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. *J Allergy Clin Immunol* **130**, 76 (Jul, 2012).

33. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655 (Sep, 2009).

34. D. B. Hancock *et al.*, Genome-wide association study implicates chromosome 9q21.31 as a susceptibility locus for asthma in mexican children. *PLoS genetics* **5**, e1000623 (Aug 01, 2009).

35. H. Wu *et al.*, Evaluation of candidate genes in a genome-wide association study of childhood asthma in Mexicans. *J Allergy Clin Immunol* **125**, 321 (Feb, 2010).

36. J. Z. Li *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. *Science (New York, NY)* **319**, 1100 (Mar 22, 2008).

37. M. Stephens, P. Scheet, Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American journal of human genetics* **76**, 449 (Apr 01, 2005).

38. M. Stephens, N. J. Smith, P. Donnelly, A new statistical method for haplotype reconstruction from population data. *American journal of human genetics* **68**, 978 (May 01, 2001).

39. M. T. Villarreal-Molina *et al.*, Association of the ATP-binding cassette transporter A1 R230C variant with early-onset type 2 diabetes in a Mexican population. *Diabetes* **57**, 509 (Feb, 2008).

40. S. Romeo *et al.*, Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* **40**, 1461 (Dec, 2008).

41. J. C. Barrett, B. Fry, J. Maller, M. J. Daly, Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263 (Jan 15, 2005).

Appendix 1

An Algorithm for Identifying Ancestrally Informative Markers (AIMs)

A common approach to low-cost ancestry estimation is to genotype a small number of markers (approximately 50-500) that provide a high degree of ancestry information, rather than go through the expense of genome-wide array genotyping. These ancestry estimates can then be used for candidate gene analyses or overall characterizations of admixture patterns for various geographic regions or medical studies, or can be incorporated in the design of smaller custom arrays to adjust for potential population stratification and thereby reduce confounding in genetic association testing. Several panels previously existed (e.g. (*1, 2*)), but these panels typically were not based on a specific portable algorithm and therefore are less extensible to novel situations.

In partnership with the Latin American Cancer Epidemiology Consortium(*3*), we were tasked with developing a panel of markers that could be used for Latino populations across the Americas (with ancestry primarily from Native Americans, Europeans, and sub-Saharan Africans). These should be broadly applicable for ancestry estimation in all Northern, Southern, and Caribbean populations identifying as Hispanic or Latino, therefore having the goal of modeling the admixture process with consistently representative AIMs from all three ancestral populations. Here I will present the basic algorithm used in Galanter et al. and, following it, the Python code that is also available at https://code.google.com/p/aims-project/.

Intuitively, we can think of an ideal AIM as being a SNP where ancestry is perfectly correlated to allele state. With these markers there would be a one-to-one correspondence between observed allele and ancestry at that locus. While rare, the marker for the Duffy Null blood type (rs12075) approaches this level of ancestry information, and so do certain markers on the Y chromosome and mitochondrial genome, for example, those that track the Out-of-Africa migration. Alternatively, a SNP with poor ancestry information would have similar allele frequencies in all populations studied, making ancestry impossible to infer from that marker alone. Most markers on the genome fall along the spectrum somewhere in the middle, with imperfect association between ancestry and genotype.

In AIMS_GENERATOR we use a statistic for capturing this spectrum of ancestry information originally developed by Rosenberg et al. (*4*) called the I_n statistic. By virtue of being developed expressly to determine ancestry from information theoretic principles, this statistic is ideal for use in identifying AIMs from genome-wide genotype data. The statistic is related to the better-known F_{st} across a wide range of minor allele frequencies for pairwise comparisons of population allele frequencies. In our implementation, we calculate pairwise I_n values for all 3 pairwise comparisons of 3 ancestral groups, then use those metrics to create an unrooted tree for each population at each SNP. This turns pairwise I_n values into locus-specific branch lengths (LSBL, See Figure 1). SNPs with maximal ancestry informativeness then have the longest branch lengths. The ideal AIMs panel then is one that balances maximal ancestry informativeness for each population with independent

inheritance (therby prioritizing markers that are unlinked), to ensure that each marker chosen contributes novel ancestry information.

Pre-processing genotype data

The program uses summary files generated by plink(*5*) for the ancestral populations of interest. In particular, we chose to calculate allele frequencies, Hardy-Weinberg Equilibrium p-values, and linkage disequilibrium values as measured by R^2. We begin with high-quality genotypes using standard genotype QC methods typical for genome-wide association studies (e.g. filtering for SNPs with low missingness, good genotype clustering, etc.). We then prune markers that are highly divergent Hardy Weinberg in the ancestral populations (as this is an assumption of typical ancestry estimation methods). After that, we prune markers that show substantial heterogeneity within continents (via a chi-squared test of allele frequencies) to ensure generalizability of findings for each ancestral group. Populations representative of ancestry from the same region are then merged in plink to calculate minor allele frequencies. Our version of the code checks for minor/major allele flips (common in AIMs, given the large discrepancy in allele frequencies) between reference populations. LD is calculated based on the sliding window in plink, and the user has options for both a minimum physical distance and correlation level that is allowable in the AIMs panel.

Running AIMS_GENERATOR

An outline of the algorithm is given in Figure 2. The program is written in Python and should run on any platform after installing the Numpy/Scipy Python libraries. It will use its

own interactive command prompt and will ask for the relevant paths to each set of relevant files generated from plink. It will then calculate pairwise I_n values and the branch lengths for each population that correspond to each SNP, and store that output in a text file for later use. The program uses a greedy approach to identify SNPs, beginning with the best AIM in the data. After that first AIM is included, the population with the smallest total LSBL (the least amount of ancestry information) is targeted and the best AIM left for that population is found. Provided that this new AIM is independent of the first one (via the physical and correlation buffers given by the user), it is included, the total LSBLs are recalculated, and the algorithm cycles back to finding the population with the smallest total LSBL. By checking at each step to the total LSBL, the program adapts to varying degrees of population structure and does not give a fixed number of AIMs relevant to each population.

The output of AIMS_GENERATOR is a ranked list of AIMs identified in the genome. We provide in Galanter et al. accuracy metrics for various sized AIMs panels as compared to genome-wide array data as a gold standard, and certainly larger panels provide the most accurate results. However, it is important to note that even small panels of AIMs (ie fewer than 50 markers) still allow individuals to account for population stratification in their data. While the ancestry estimates themselves may have increased noise, they still provide enough information to be used as a covariate to capture potential population stratification in genetic association studies.

AIMs panel developed by AIMS_GENERATOR

The first version of the algorithm was used to develop a panel of 446 markers for Sequenom assays. After filtering out variants that were poor candidates for the genotyping platform, the resulting panel was able to estimate ancestry highly accurately: <0.1 RMSE for all ancestries across diverse populations of mixed ancestry. By virtue of being broadly relevant to the populations of mixed ancestry in the Americas, the AIMS_GENERATOR results were used to pick AIMs in the design of both the Illumina and Affymetrix exome arrays.

Multi-population extension to AIMS_GENERATOR

Our original implementation was specifically designed for 3 ancestral populations. As such the calculation of LSBL is simple as the unrooted tree topology is fixed. With additional ancestral populations, the values become more complicated as the tree topologies can change, and root-tip branch lengths are less interpretable. Therefore as an extension to AIMS_GENERATOR we have chosen to balance pairwise I_n values, rather than branch lengths, with K>3 ancestral populations. The algorithm iterates similarly to before in a greedy fashion. Comparisons to genome-wide genotype data for an admixed population with 5 ancestries (*6*) indicate that this forthcoming version of AIMS_GENERATOR provides more accurate AIMs panels than other methods, including an algorithm based on PCA correlations (*7*) and a genetic algorithm using allele frequency differentiation (*8*).

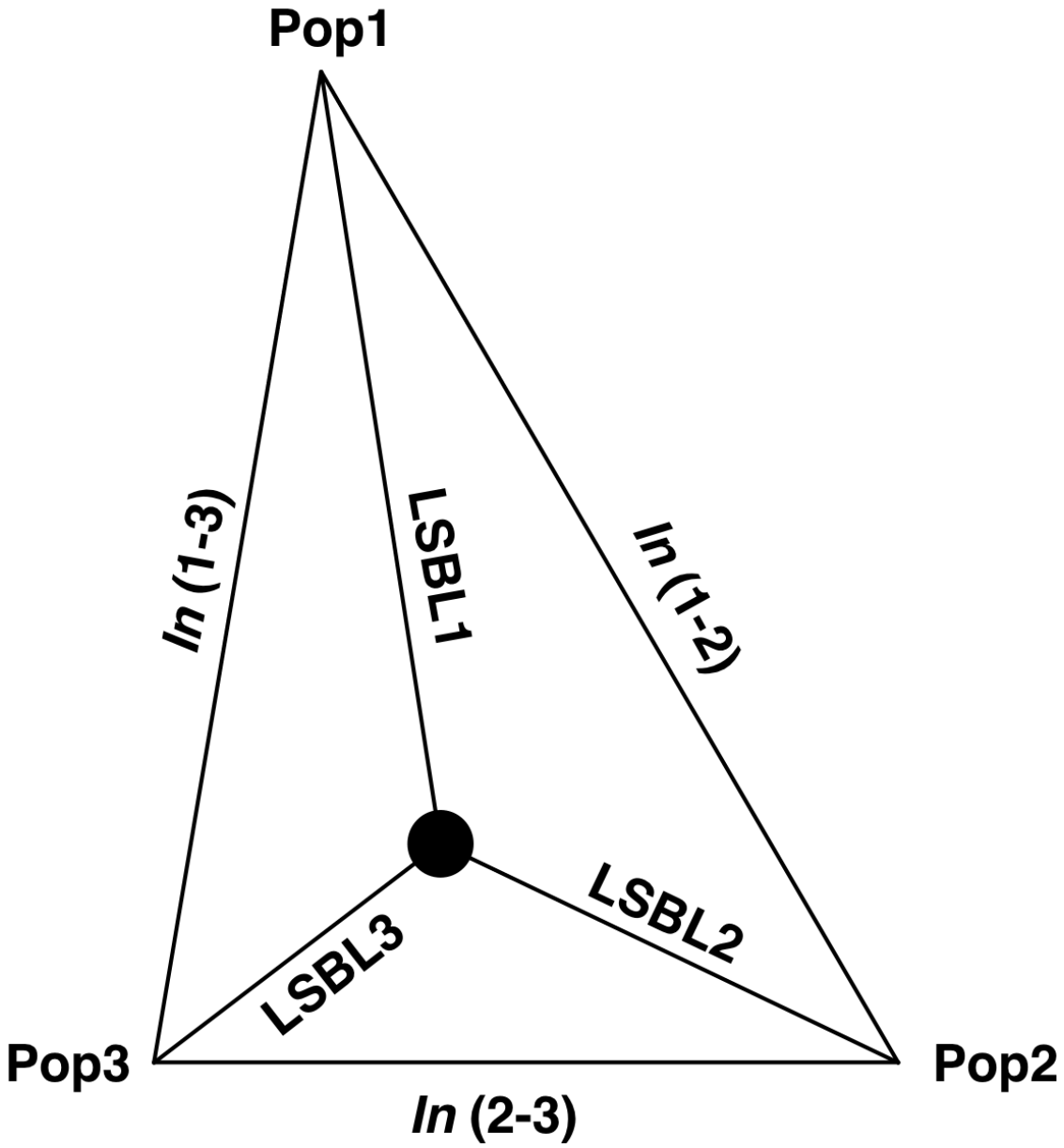


Figure 1. A diagram demonstrating the locus-specific branch length statistic. For each SNP, pairwise I_n values are calculated from allele frequencies. These are transformed into an unrooted tree, where the distance from the centroid to each population is the LSBL, the statistic used by AIMS_GENERATOR.

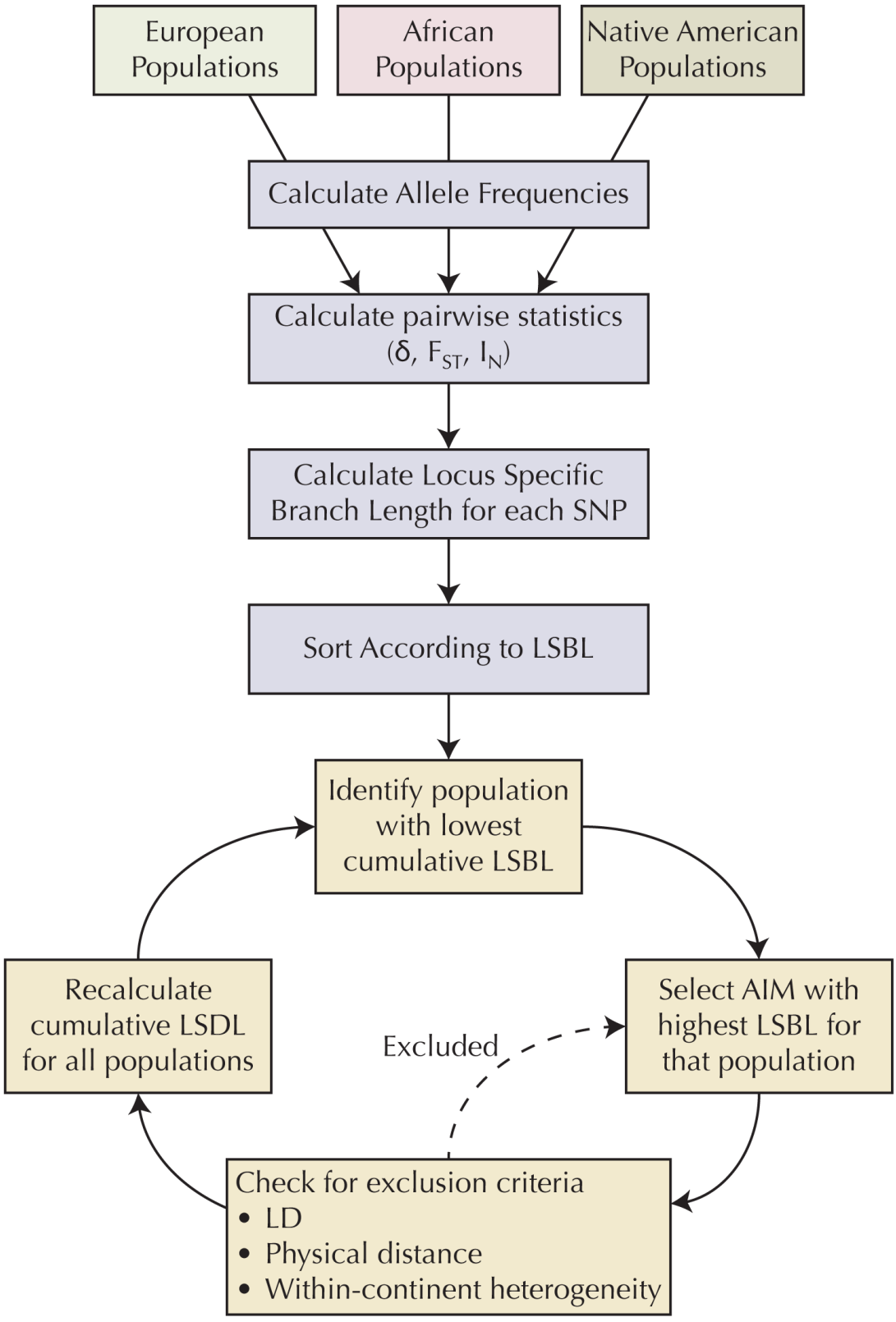


Figure 2. Graphical description of the AIMS_GENERATOR algorithm from Galanter et al.

Sources

1. S. Choudhry *et al.*, Population stratification confounds genetic association studies among Latinos. *Hum Genet* **118**, 652 (Jan, 2006).
2. R. Kosoy *et al.*, Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Human mutation* **30**, 69 (2009).
3. J. M. Galanter *et al.*, Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS genetics* **8**, e1002554 (Apr, 2012).
4. N. A. Rosenberg, L. M. Li, R. Ward, J. K. Pritchard, Informativeness of genetic markers for inference of ancestry. *American journal of human genetics* **73**, 1402 (Dec 01, 2003).
5. S. Purcell *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559 (Sep 01, 2007).
6. M. Daya *et al.*, A panel of ancestry informative markers for the complex five-way admixed South African Coloured population. *submitted*, (2013).
7. P. Paschou *et al.*, PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS genetics* **3**, 1672 (Sep 01, 2007).
8. O. Lao *et al.*, Correlation between genetic and geographic structure in Europe. *Curr Biol* **18**, 1241 (Aug 26, 2008).

Appendix 2

AIMS_GENERATOR python code

```
'''
AIMs generator
Joshua Galanter and Chris Gignoux

This program will take genotype files and allele frequency files from plink as well
as a pairwise LD file and will generate sets of AIMs based on informativeness.
Also available at: https://code.google.com/p/aims-project
'''

import sys
from sys import stdout
import os
from math import log
from numpy import *
from scipy import stats



def calc_sigma(af1, af2):
	return af1 + af2



def calc_delta(af1, af2):
	return abs(af1 - af2)




def calc_Fst(af1,af2):
	'''Calculates pairwise Fst given allele frequencies'''
	if (af1 == 0 and af2 == 0) or (af1 == 1 and af2 == 1):
		return 0
```

```
	return calc_delta(af1, af2) ** 2 / (calc_sigma(af1, af2) * (2 - calc_sigma(af1, af2)))



def calc_In(af1, af2):
	'''calculate Rosenberg et al.'s In statistic for pairs of populations'''
	if (af1 == 0 and af2 == 0) or (af1 == 1 and af2 == 1):
		return 0
	if af1 == 1:
		af1 = 0
		af2 = 1 - af2
	elif af2 == 1:
		af1 = 1 - af1
		af2 = 0
	sigma = calc_sigma(af1, af2)
	delta = calc_delta(af1, af2)
	a_exp = -0.5 * log(sigma ** sigma * (2 - sigma) ** (2 - sigma))
	b_exp = 0.25 * log( (sigma + delta) ** (sigma + delta) * (2 - sigma - delta) ** (2 - sigma - delta) * (sigma - delta) ** (sigma - delta) * (2 - sigma + delta) ** (2 - sigma + delta) )
	# return sum
	return a_exp + b_exp



def calc_lsbl(ab, ac, bc):
	'''Calculates locus specific branch length given three pairwise statistics (either In or Fst)'''
	a = (ab + ac - bc)/2
	b = (ab + bc - ac)/2
	c = (ac + bc - ab)/2
	return (a, b, c)
```

```
def calculate_ld(ldfile, rsq_threshold = 0.1):
	'''read in LD and position calculated in PLINK for all snps in LD, given a
window sizepassed as a parameter.'''
	lddict = {}
	while True:
		try:
			for total, line in enumerate(file(ldfile)):
				pass
			for i, line in enumerate(file(ldfile)):
				if i > 0:
					if i % (total/20) == 0:
						print '%sImporting LD file:  [%s%s]'
%('\b'*50, '=' * (i * 20 / total), ' ' * (19 - (i * 20 / total))),
						stdout.flush()
					line = line.strip().split()
					if float(line[-1]) > rsq_threshold:
						#print line
						snpA = line[2]
						snpB = line[5]
						if snpA in lddict:
							lddict[snpA].append(snpB)
						else:
							lddict[snpA] = [snpB]
						if snpB in lddict:
							lddict[snpB].append(snpA)
						else:
							lddict[snpB] = [snpA]
			print
			return lddict
		except:
```

```
                print 'File error; LD file is not in the correct format.'
                ldfile = getFile(prompt, 'LD')
                continue



def getFile(prompt, type, verbose = False):
    '''
    This function will get a filename entered by the user.
    '''
    print 'Please enter a filename for a %s file.' %(type.replace('_', ' '))
    print 'You can also enter \'ls\' to list directory contents or \'pwd\' to print
the current directory.'
    while True:
        sys.stdout.write(prompt)
        line = sys.stdin.readline()
        filename = line.strip().split()[0]
        if filename == 'pwd':
            print os.getcwd()
        elif filename == 'ls':
            filelist = os.listdir(os.getcwd())
            for i in filelist:
                print i
        else:
            try:
                f = open(filename)
                f.close()
                return filename
            except:
                print 'That was not a valid filename.'
                return getFile(prompt, type, verbose)
```

```
def get_LD(prompt = '> '):
	'''
	This function will read in a plink "ld" file to get pairwise LD between SNPs.
	It will return a SNP - list of SNP's dictionary, where the list of SNPs
consists
	of all the SNPs in LD above the LD threshold with the index SNP.
	'''
	filename = getFile(prompt, 'LD')
	while True:
		print 'Enter LD threshold, between 0 and 1'
		sys.stdout.write(prompt)
		line = sys.stdin.readline()
		try:
			ld_threshold = float(line.strip().split()[0])
			if ld_threshold > 0 and ld_threshold < 1:
				return calculate_ld(filename, ld_threshold)
			else:
				print 'LD threshold must be a number between 0 and 1.'
				continue
		except:
			print 'LD threshold must be a number between 0 and 1.'
			continue



def get_bim(prompt = '> ', filename = ''):
	'''
	This function will read in a plink "bim" file to get the chromosomal positions.
	It will output a SNP - (chromosome, position) dictionary, as well as a
dictionary of
	alleles (minor, major).
	'''
```

```
	if filename == '':
		filename = getFile(prompt, 'plink bim (position)')
	try:
		print 'Reading in position data...'
		posdict = dict([[line.strip().split()[1],(line.strip().split()[0],
line.strip().split()[3])] for line in file(filename)])
		alleledict = dict([[line.strip().split()[1],(line.strip().split()[4],
line.strip().split()[5])] for line in file(filename)])
		return (posdict, alleledict)
	except:
		print 'File error; bim file is not in the correct format.'
		return get_bim(prompt)



def get_freq(population, nsnps = -1, prompt = '> ', filename = ''):
	'''
	This function will read in a plink "frq" file to get minor allele frequencies.
	'''
	if filename == '':
		filename = getFile(prompt, population + ' allele frequency')
	try:
		print 'Reading allele frequency data...'
		freq = file(filename)
		freq.readline()
		if nsnps == -1:
			return [line.strip().split() for line in freq.readlines()]
		else:
			snps = []
			for i, line in enumerate(freq.readlines()):
				if i > nsnps:
					print 'Error, there are more SNPs in the frequency
file than in the bim file'
```

```
                                return get_freq(population, nsnps, prompt)
                        if i % (nsnps/20) == 0:
                                print '%sImporting frequency file:  [%s%s]'
%('\b'*50, '=' * (i * 20 / nsnps), ' ' * (19 - (i * 20 / nsnps))),
                                stdout.flush()
                        snps.append(line.strip().split())
                return snps
        except:
                print 'File error; allele frequency file is not in the correct format.'
                return get_freq(population, nsnps, prompt)



def correct_freq(freq, mafdict):
        n = 0
        for i, snp in enumerate(freq):

                if snp[2] == '0':
                        if snp[3] == mafdict[snp[1]][1]:
                                freq[i][2] = mafdict[snp[1]][0]
                        else:
                                freq[i][4] = '1'
                                freq[i][3] = mafdict[snp[1]][1]
                                freq[i][2] = mafdict[snp[1]][0]
                elif snp[2] == mafdict[snp[1]][1]:
                        n += 1
                        freq[i][4] = str(1-float(snp[4]))
                        freq[i][2] = mafdict[snp[1]][0]
                        freq[i][3] = mafdict[snp[1]][1]
                #if snp[1] == 'rs548824':
                #       print '\nIdentified the SNP'
                #       print freq[i], mafdict[snp[1]][0]
        print '\nFixed %s out of %s minor allele frequency flips' %(n, i)
```

```
	return freq


def sort_snps(snps, positions):
	print 'Sorting snps...'
	return sorted(snps, key = lambda order: (positions[order[1]][0],
positions[order[1]][1]))


def calc_pairwise_aims(positions, pop1_frq, pop2_frq):
	'''
	Calculates the pairwise aims statistics for two populations given their allele
frequencies.
	Returns a list of aims, that include SNP, chromosome, position,
	allele frequency in population 1, allele frequency in population 2,
	sigma (the sum of allele frequencies), delta (the difference in allele
frequencies),
	pairwise Fst, and pairwise In
	'''
	aims = []
	ignored = 0
	for i, snp in enumerate(pop1_frq):
		#if snp[1] == 'rs548824':
		#	print 'Identified the SNP'
		#	print snp
		#	print pop2_frq[i]
		if int(snp[0]) > 22:
			# ignore non-autosomal SNPs
			ignored += 1
			continue
		else:
			if i % (len(pop1_frq)/20) == 0:
```

```
                        print '%sCalculating pairwise AIMs statistics:  [%s%s]'
%('\b'*61, '=' * (i * 20 / len(pop1_frq)), ' ' * (19 - (i * 20 / len(pop1_frq)))),
                        stdout.flush()
                    af1 = float(pop1_frq[i][4])
                    af2 = float(pop2_frq[i][4])
                    #if snp[1] == 'rs548824':
                    #     print '\nIdentified the SNP'
                    #     print
'snp\tchr\tposition\t%s_allele_freq\t%s_allele_freq\tsigma\tdelta\tFst\tIn\n'
                    #     print snp[1], positions[snp[1]][0], positions[snp[1]][1],
af1, af2, calc_sigma(af1, af2), calc_delta(af1, af2), calc_Fst(af1,af2),
calc_In(af1,af2)
                    #     print
                    aims.append([snp[1], positions[snp[1]][0], positions[snp[1]][1],
af1, af2, calc_sigma(af1, af2), calc_delta(af1, af2), calc_Fst(af1,af2),
calc_In(af1,af2)])
      return aims



def sort_pairwise_aims(unsorted_aims, stat = 'In'):
      '''
      Sorts the pairwise aims statistics given a set of aims.  It can sort on the
basis
      of In or Fst.  Defaults to In.
      '''
      if stat == 'In':
            print '\nsorting output...'
            aims = sorted(unsorted_aims, key = lambda order: order[8])
            aims.reverse()
            return aims
      elif stat == 'Fst':
            print '\nsorting output...'
```

```
                aims = sorted(unsorted_aims, key = lambda order: order[7])
                aims.reverse()
                return aims
        else:
                print 'Invalid sort command, returning unsorted aims.'
                return aims



def output_pairwise_aims(aims, pop1, pop2, outfilename, n = -1):
        '''
        Writes the AIMs statistics to a file, given the filename.  n is an optional
        parameter that specifies how many AIMs to output.  It defaults to -1 (all)
        '''
        print 'Writing pairwise AIMs statistics for %s/%s populations to file %s'
%(pop1, pop2, outfilename)
        outfile = file(outfilename, 'w')
        outfile.write('snp\tchr\tposition\t%s_allele_freq\t%s_allele_freq\tsigma\tdelta
\tFst\tIn\n' %(pop1, pop2))
        lines = 0
        for line in aims:
                if lines == n:
                        return
                else:
                        outfile.write('%s\n' % ('\t'.join([str(val) for val in line])))
                        lines += 1
        return



def calc_pop_aims(positions, pop1_frq, pop2_frq, pop3_frq, AB_pairwise_aims,
AC_pairwise_aims, BC_pairwise_aims, populations):
        '''
```

```
	Calculates the locus specific brach length statistics for three populations, given the
	pairwise AIMs statistics of three populations.  It returns a tuple containing three lists,
	one for each of the three populations.  The lists contain variables for:
	rsID, chromosome, position, allele frequency, locus specific branch length (LSBL) by Fst, and LSBL by In
	'''
	pop1_aims = []
	pop2_aims = []
	pop3_aims = []
	#print 'The length of the NAM/AFR pairwise AIMS is', len(AB_pairwise_aims)
	#print 'The length of the NAM/EUR pairwise AIMS is', len(BC_pairwise_aims)
	#print 'The length of the EUR/AFR pairwise AIMS is', len(AC_pairwise_aims)
	#print 'The length of the NAM pop frequency file is', len(pop1_frq)
	n = 0
	for i, snp in enumerate(pop1_frq):
		if int(snp[0]) > 22:
			# ignore non-autosomal SNPs
			continue
		else:
			if i % (len(pop1_frq)/20) == 0:
				print '%sCalculating branch length AIMs statistics: [%s%s]' %('\b'*66, '=' * (i * 20 / len(pop1_frq)), ' ' * (19 - (i * 20 / len(pop1_frq)))),
				stdout.flush()
			threeway_Fst = calc_lsbl(AB_pairwise_aims[n][7], AC_pairwise_aims[n][7], BC_pairwise_aims[n][7])
			threeway_In = calc_lsbl(AB_pairwise_aims[n][8], AC_pairwise_aims[n][8], BC_pairwise_aims[n][8])
			pop1_aims.append([snp[1], positions[snp[1]][0], positions[snp[1]][1], float(pop1_frq[i][4]), threeway_Fst[0], threeway_In[0], populations[0]])
```

```
                    pop2_aims.append([snp[1], positions[snp[1]][0],
positions[snp[1]][1], float(pop2_frq[i][4]), threeway_Fst[1], threeway_In[1],
populations[1]])
                    pop3_aims.append([snp[1], positions[snp[1]][0],
positions[snp[1]][1], float(pop3_frq[i][4]), threeway_Fst[2], threeway_In[2],
populations[2]])
                    n += 1
      print '\n'
      return (pop1_aims, pop2_aims, pop3_aims)



def sort_pop_aims(unsorted_aims, stat = 'In'):
      '''
      Sorts the locus specific branch length aims statistics given a set of aims.  It
can sort on the basis
      of In or Fst.  Defaults to In.
      '''
      if stat == 'In':
            print 'sorting output...'
            aims = sorted(unsorted_aims, key = lambda order: order[5])
            aims.reverse()
            return aims
      elif stat == 'Fst':
            print 'sorting output...'
            aims = sorted(unsorted_aims, key = lambda order: order[4])
            aims.reverse()
            return aims
      else:
            print 'Invalid sort command, returning unsorted aims.'
            return aims
```

```
def output_pop_aims(aims, population, outfilename, n = -1):
	'''
	Writes the AIMs statistics to a file, given the filename.  n is an optional
	parameter that specifies how many AIMs to output.  It defaults to -1 (all)
	'''
	print 'Writing AIMs statistics for %s population to file %s' %(population,
outfilename)
	outfile = file(outfilename, 'w')
	outfile.write('snp\tchr\tposition\tallele_frequency\tLSBL(Fst)\tLSBL(In)\n')
	lines = 0
	for line in aims:
		if lines == n:
			return
		else:
			outfile.write('%s\n' % ('\t'.join([str(val) for val in
line[0:6]])))
			lines += 1
	return



def calc_all_aims(positions, pop1_frq, pop2_frq, pop3_frq, pops, n = -1, outstem =
'aimsfile'):
	'''
	Calculates all aims statistics, given three lists of population frequencies.
It will calculate
	and save to files all three pairwise aims statistics, as well as the population
specific locus
	specific branch length statistics.  It will return a tuple of LSBLs.
	'''
	AB_aims = calc_pairwise_aims(positions, pop1_frq, pop2_frq)
	sorted_AB_aims = sort_pairwise_aims(AB_aims)
```

```
        outfile = outstem + '_' + pops[0] + '_' + pops[1] + '.aims'
        output_pairwise_aims(sorted_AB_aims, pops[0], pops[1], outfile, n)

        AC_aims = calc_pairwise_aims(positions, pop1_frq, pop3_frq)
        sorted_AC_aims = sort_pairwise_aims(AC_aims)
        outfile = outstem + '_' + pops[0] + '_' + pops[2] + '.aims'
        output_pairwise_aims(sorted_AC_aims, pops[0], pops[2], outfile, n)

        BC_aims = calc_pairwise_aims(positions, pop2_frq, pop3_frq)
        sorted_BC_aims = sort_pairwise_aims(BC_aims)
        outfile = outstem + '_' + pops[1] + '_' + pops[2] + '.aims'
        output_pairwise_aims(sorted_BC_aims, pops[1], pops[2], outfile, n)


        pop1_aims, pop2_aims, pop3_aims = calc_pop_aims(positions, pop1_frq, pop2_frq,
pop3_frq, AB_aims, AC_aims, BC_aims, pops)


        sorted_pop1_aims = sort_pop_aims(pop1_aims, 'In')
        outfile = outstem + '_' + pops[0] + '.aims'
        output_pop_aims(sorted_pop1_aims, pops[0], outfile, n)

        sorted_pop2_aims = sort_pop_aims(pop2_aims, 'In')
        outfile = outstem + '_' + pops[1] + '.aims'
        output_pop_aims(sorted_pop2_aims, pops[1], outfile, n)

        sorted_pop3_aims = sort_pop_aims(pop3_aims, 'In')
        outfile = outstem + '_' + pops[2] + '.aims'
        output_pop_aims(sorted_pop3_aims, pops[2], outfile, n)

        return (sorted_pop1_aims, sorted_pop2_aims, sorted_pop3_aims)
```

```
def too_close(index_snp, snp_list, positions, population, distance = 100000):
	if snp_list == [None]:
		print 'returning none'
		return False
	else:
		for snp in snp_list:
			if snp == None:
				print snp
				print 'snp is empty'
				return False
			else:
				if snp[6] == population:
					if snp[1] == index_snp[1]:
						if abs(int(snp[2]) - int(index_snp[2])) <=
distance:
							# print snp[0:3], 'and',
index_snp[0:3], 'are too close'
							return True
	return False



def get_aims(positions, lddict, alleles, populations, pop1aims, pop2aims, pop3aims,
excluded = set(), distance = 100000, n = 500):
	aimslist = []
	print 'Generating informativeness dictionary for %s' %(populations[0])
	pop1stat = dict([[aims[0],aims[5]] for aims in pop1aims])
	print 'Generating informativeness dictionary for %s' %(populations[1])
	pop2stat = dict([[aims[0],aims[5]] for aims in pop2aims])
	print 'Generating informativeness dictionary for %s' %(populations[2])
	pop3stat = dict([[aims[0],aims[5]] for aims in pop3aims])
	print 'Initializing statistics'
```

```
	pop1info = 0.
	pop2info = 0.
	pop3info = 0.
	pop1pos = 0
	pop2pos = 0
	pop3pos = 0
	pop1_numaims = 0
	pop2_numaims = 0
	pop3_numaims = 0
	ldex = set()
	pop1ex = set()
	pop2ex = set()
	pop3ex = set()
	n_aims = 0
	n_het_excluded = 0
	while n_aims < n:
		found = False
		if (pop1info < pop2info) and (pop1info < pop3info):
			aim = pop1aims[pop1pos]
			#print 'Selected aim %s for evaluation' %(aim[0])
			pop1pos += 1
			if (aim[0] not in (ldex | excluded)) and not too_close(aim,
aimslist, positions, populations[0], distance):
				print 'Found an AIM for %s, %s; %s AIMs found so far'
%(populations[0], aim[0], len(aimslist) + 1)
				found = True
				aimslist.append(aim)
				pop1_numaims += 1
		elif pop2info < pop3info:
			aim = pop2aims[pop2pos]
			pop2pos += 1
			#print 'Selected aim %s for evaluation' %(aim[0])
```

```
                    if (aim[0] not in (ldex | excluded)) and not too_close(aim,
aimslist, positions, populations[1], distance):
                            print 'Found an AIM for %s, %s; %s AIMs found so far'
%(populations[1], aim[0], len(aimslist) + 1)
                            found = True
                            aimslist.append(aim)
                            pop2_numaims += 1
             else:
                    aim = pop3aims[pop3pos]
                    pop3pos += 1
                    #print 'Selected aim %s for evaluation' %(aim[0])
                    if (aim[0] not in (ldex | excluded)) and not too_close(aim,
aimslist, positions, populations[2], distance):
                            print 'Found an AIM for %s, %s; %s AIMs found so far'
%(populations[2], aim[0], len(aimslist) + 1)
                            found = True
                            aimslist.append(aim)
                            pop3_numaims += 1
             if found:
                    n_aims += 1
                    #print 'Found %s aims so far.' %(n_aims)
                    try:
                            ldex = ldex | set(lddict[aim[0]])
                    except:
                            print 'snp %s is not in the ld dictionary.' %(aim[0])
                    pop1info += pop1stat[aim[0]]
                    pop2info += pop2stat[aim[0]]
                    pop3info += pop3stat[aim[0]]
             else:
                    if aim[0] in excluded:
                            #print 'Excluded an AIM for heterogeneity'
                            n_het_excluded += 1
      print 'The total locus specific In for the three populations are:'
```

```
	print 'For population %s, found %s aims out of %s evaluated, for a total LSBL
of %s' %(populations[0], pop1_numaims, pop1pos, pop1info)
	print 'For population %s, found %s aims out of %s evaluated, for a total LSBL
of %s' %(populations[1], pop2_numaims, pop2pos, pop2info)
	print 'For population %s, found %s aims out of %s evaluated, for a total LSBL
of %s' %(populations[2], pop3_numaims, pop3pos, pop3info)
	print 'A total of %s AIMs were excluded due to heterogeneity' %(n_het_excluded)
	return aimslist



def print_aims(aims, filename, pop1frq, pop2frq, pop3frq, populations):
	pop1dict = dict([[snp[1],snp[4]] for snp in pop1frq])
	pop2dict = dict([[snp[1],snp[4]] for snp in pop2frq])
	pop3dict = dict([[snp[1],snp[4]] for snp in pop3frq])
	outfile = file(filename, 'w')
	outfile.write('snp\tchr\tposition\t%s_AF\t%s_AF\t%s_AF\tpopulation\tLSBL(Fst)\t
LSBL(In)\n' %(populations[0], populations[1], populations[2]))
	for i, aim in enumerate(aims):
		outfile.write('%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\n' % (aim[0], aim[1],
aim[2], pop1dict[aim[0]], pop2dict[aim[0]], pop3dict[aim[0]], aim[6], aim[4], aim[5]))



def chi(table):
	observed = array(table)
	rowsum = observed.sum(axis = 1)
	colsum = observed.sum(axis = 0)
	expected = (rowsum[:, newaxis] * colsum) / sum(rowsum)
	return stats.chisquare(expected.reshape(-1), observed.reshape(-1))
```

```
def calc_het(populations, filenames, pop_freq, mafdict, posdict, prompt = '>',
threshold = 0.01):
	print 'Calculating heterogeneity for the following populations:',
	for i in populations:
		print i,
	print
	allele_freqs = []
	heterogeneity = []
	for i, filename in enumerate(filenames):
		allele_freqs.append(sort_snps(correct_freq(get_freq(populations[i],
len(posdict), prompt, filename), mafdict), posdict))
	for i, snp in enumerate(pop_freq):
		table = []
		for j, filename in enumerate(filenames):
			minor_allele_count = float(allele_freqs[j][i][4]) *
float(allele_freqs[j][i][5])
			major_allele_count = (1. - float(allele_freqs[j][i][4])) *
float(allele_freqs[j][i][5])
			table.append([minor_allele_count, major_allele_count])
		heterogeneity.append(chi(table))
	return heterogeneity



def exclude_het(hetfile, frq, threshold = 0.01):
	print 'Finding SNPs to exclude on the basis of heterogeneity...'
	exclude = set()
	for i, snp in enumerate(frq):
		if hetfile[i][1] < threshold:
			exclude.add(frq[i][1])
	print 'Excluding %s SNPs on the basis of heterogeneity...' %(len(exclude))
	return exclude
```

```
if __name__ == '__main__':
	interactive = False
	test = True
	dev = True
	verbose = dev
	prompt = '> '
	if interactive:
		print 'What are the ancestral groups?  (enter them separated by enter)'
		sys.stdout.write(prompt)
		pop1 = sys.stdin.readline().strip().split()[0]
		sys.stdout.write(prompt)
		pop2 = sys.stdin.readline().strip().split()[0]
		sys.stdout.write(prompt)
		pop3 = sys.stdin.readline().strip().split()[0]

		populations = (pop1, pop2, pop3)
		posdict, alleledict = get_bim(prompt)

		print 'Now we will need allele frequency files for the three ancestral
groups.'
		nam_freq = correct_freq(get_freq(populations[0], len(posdict), prompt),
alleledict)
		afr_freq = correct_freq(get_freq(populations[1], len(posdict), prompt),
alleledict)
		eur_freq = correct_freq(get_freq(populations[2], len(posdict), prompt),
alleledict)
		print 'Now we will calculate all the AIMs stats.'
		print 'Please enter a file prefix for the output files.'
		sys.stdout.write(prompt)
		outfile = sys.stdin.readline().strip().split()[0]
		print 'How many AIMs do you want calculated? (-1 if you want all SNPs
included)'
```

```
                sys.stdout.write(prompt)
                try:
                        n_aims = int(sys.stdin.readline().strip().split()[0])
                except:
                        print 'Sorry, you did not enter an integer, defaulting to all
SNPs'
                        n_aims = -1

                lddict = get_LD(prompt)

                print 'What distance do you want between AIMS? (defaults to 100Kb)'
                sys.stdout.write(prompt)
                try:
                        distance = int(sys.stdin.readline().strip().split()[0])
                except:
                        print 'Sorry, you did not enter an integer, defaulting to 100Kb'
                        distance = 100000
                print 'How many AIMs do you want? (defaults to 500)'
                sys.stdout.write(prompt)
                try:
                        n = int(sys.stdin.readline().strip().split()[0])
                except:
                        print 'Sorry, you did not enter an integer, defaulting to 500'
                        n = 500
        else:
                populations = ('NAM', 'AFR', 'EUR')
                ldfile = 'NAM.ld'
                lddict = calculate_ld(ldfile, 0.1)
                posfile = 'pos.bim'
                posdict, alleledict = get_bim(prompt, posfile)
                nam_freq_file = 'NAM.frq'
                eur_freq_file = 'EUR.frq'
                afr_freq_file = 'AFR.frq'
```

```
            nam_freq = sort_snps(correct_freq(get_freq(populations[0], len(posdict),
prompt, nam_freq_file), alleledict), posdict)
            afr_freq = sort_snps(correct_freq(get_freq(populations[1], len(posdict),
prompt, afr_freq_file), alleledict), posdict)
            eur_freq = sort_snps(correct_freq(get_freq(populations[2], len(posdict),
prompt, eur_freq_file), alleledict), posdict)

            threshold = 0.01
            eur_pops = ('SPA', 'TSI', 'CEU')
            eur_files = ('SPA.frq', 'TSI.frq', 'CEU.frq')
            eur_het = calc_het(eur_pops, eur_files, eur_freq, alleledict, posdict,
prompt = '>')

            afr_pops = ('YRI', 'LWK')
            afr_files = ('YRI.frq', 'LWK.frq')
            afr_het = calc_het(afr_pops, afr_files, afr_freq, alleledict, posdict,
prompt = '>')

            nam_pops = ('MAY', 'TEP', 'ZAP', 'NAH', 'QUE', 'AYM')
            nam_files = ('MAY.frq', 'TEP.frq', 'ZAP.frq', 'NAH.frq', 'QUE.frq',
'AYM.frq')
            nam_het = calc_het(nam_pops, nam_files, nam_freq, alleledict, posdict,
prompt = '>')


            exclude = exclude_het(eur_het, eur_freq, threshold) |
exclude_het(afr_het, afr_freq, threshold) | exclude_het(nam_het, nam_freq, threshold)
            print 'A total of %s AIMs have significant heterogeneity' %(len(exclude))

      n_aims = -1
      dist = ['100k', '250k', '500k', '1m']
      distances = [100000, 250000, 500000, 1000000]
      number = ['500', '1000']
```

```
	for i, distance in enumerate(distances):
		for j, n in enumerate([500, 1000]):
			outfile = 'LACE_aims_' + dist[i] + '_' + number[j]
			print outfile
			nam_aims, afr_aims, eur_aims = calc_all_aims(posdict, nam_freq,
afr_freq, eur_freq, populations, n_aims, outfile)
			my_aims = get_aims(posdict, lddict, alleledict, populations,
nam_aims, afr_aims, eur_aims, exclude, distance, n)
			print_aims(my_aims, outfile + '.aims', nam_freq, afr_freq,
eur_freq, populations)
```

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

9/5/13

Date