Long-Range LD Can Confound Genome Scans in Admixed Populations

To the Editor: In the September 2007 issue of *The Journal*, Tang et al. analyzed data from 192 Puerto Ricans genotyped at 112,584 autosomal markers and identified three regions with a deficiency in the proportion of European ancestry. They concluded that recent selection occurred at these regions after the admixture of European, African, and Native American ancestors.[1] These signals of selection are very strong: We estimate that they each correspond to selection coefficients of >0.08 per generation, which if confirmed would represent the three most powerful selective adaptations discovered to date in humans. Here, we demonstrate that on the basis of the method the authors applied, these signals of selection could be explained as artifacts of the unusual long-range linkage disequilibrium (LD) that occurs at these regions and that is not specific to Puerto Ricans. We failed to replicate the signal of selection in an independent and larger study of 364 Puerto Rican samples, when we applied a method that is not susceptible to this confounder. Our results highlight a complexity in the analysis of dense genotype data from recently admixed populations; this complexity needs to be taken into account not only in genome-wide screens for selection but also in genome-wide association studies to ensure that false-positive signals are avoided.

The signals of selection were identified with methods described in Tang et al.,[2] which uses an extension of a Hidden Markov Model (HMM) to infer segments of ancestry from dense genotype data. The authors note that the assumptions of an HMM "are violated when the marker map is dense and linkage disequilibrium (LD) exists within an ancestral population"; they partially address this confounder by modeling the LD between consecutive pairs of markers but describe this approach as a "compromise" because they do not account for higher order LD.[2] In light of the phenomenon that nearby sites in a region may be in weak LD, whereas more distant sites may be in much stronger LD, the approach of modeling only LD between consecutive markers is potentially inadequate.[3] As we demonstrate below, local-ancestry estimates in regions where LD is not fully modeled will not only be overconfident but will also be systematically biased, thereby leading to false-positive deficiencies in the population contributing majority ancestry.

In a separate analysis focusing on long-range LD in European populations, we applied principal components analysis (PCA) to several genome-wide data sets and identified 24 autosomal long-range LD regions, each spanning >2 megabases (Mb) (Table 1). The functional basis for these regions is currently being explored. The 24 PCA regions were identified by running the EIGENSOFT software[4,5] on a data set of 327 European Americans genotyped on the Illumina 550K array and identifying all regions where there was significant long-range LD extending >2 Mb that explained one of the top eigenvectors. The regions were independently replicated in 1593 European Americans from the Illumina iControl data set genotyped on the Illumina 550K array and in 1504 + 1500 British samples from the Wellcome Trust Case Control Consortium (1958 Birth Cohort and National Blood Service Cohorts, genotyped on the Affymetrix 500K array), confirming that these regions genuinely harbor long-range LD in European populations.

Strikingly, all three of the signals of selection reported by Tang et al.[1] lie in one of the PCA regions (Table 1). Because the PCA regions comprise <4.7% of the autosomal genome, the hypothesis that the regions discussed in Tang et al.[1] and the PCA regions are independent is violated with a p value of $(0.047)^3 = 0.0001$. As we will show, the presence of long-range LD in populations ancestral to Puerto Ricans could explain both the signals from Tang et al.[1] and the PCA results.

Long-range LD can arise for reasons unrelated to selection. For example, inversions are known to suppress viable recombination, and a known inversion polymorphism at position 8–12 Mb on chromosome 8 has previously been shown to be the cause of long-range LD[6] (also see Table 1). (Interestingly, this inversion polymorphism appears to produce a signal of unusual ancestry in Figure 1 of Tang et al.,[1] in addition to the three regions highlighted in the same paper.) It is important for studies inferring the action of selection to rule out alternative explanations for the observed data. For the regions identified by Tang et al.,[1] long-range LD that arose because of inversion polymorphism or other reasons provides a plausible alternative explanation.

LD that is not properly modeled impacts not only the uncertainty in local-ancestry estimates but also the expected value of these estimates, leading to large systematic biases in regions of long-range LD. To demonstrate this, we consider a hypothetical admixed population with ancestry $\alpha_1 = 80\%$ from ancestral population 1 and $\alpha_2 = 20\%$ from ancestral population 2. We then consider an A/C marker in which the A allele has frequency $p_1 = 25\%$ in population 1 and $p_2 = 75\%$ in population 2, so that its frequency in the admixed population is $p = \alpha_1 p_1 + \alpha_2 p_2 = 35\%$. Let $q_1 = 75\%$, $q_2 = 25\%$, and $q = 65\%$ denote the corresponding frequencies of the C allele. If local ancestry on a single-haploid chromosome is inferred with only information from that marker, we obtain $P(\text{population } 1|A) = \alpha_1 p_1/(\alpha_1 p_1 + \alpha_2 p_2) = 0.57$ and $P(\text{population } 1|C) = \alpha_1 q_1/(\alpha_1 q_1 + \alpha_2 q_2) = 0.92$, so that the expected value of the ancestry estimate is $E(P(\text{population } 1)) = p\ P(\text{population } 1|A) + q\ P(\text{population } 1|C) = 0.80$, which is an unbiased estimate of α_1. Now, we consider a second marker that has identical allele frequencies and that is in perfect LD with the first and suppose that the two markers are used to infer local ancestry, treating them as if they were unlinked (this could happen with the method of Tang et al.[2] if the markers

Table 1. Correspondence between Regions from Tang et al. and Regions of Extended LD in European Populations

Chromosome	SNP at Region Peak, from Tang et al.[1]	SNP Position	Extended LD Region, from PCA Analysis
6	rs169679	29.0 Mb	25.5–33.5 Mb
8	rs896760	113.5 Mb	112–115 Mb
11	rs637249	56.0 Mb	46–57 Mb

For each region reported to be under selection, we list the SNP defining the peak of this region as described in Tang et al.,[1] the physical position of the SNP, and the physical position of the corresponding region of extended LD from PCA analysis. The other autosomal long-range LD regions identified by PCA analysis were chromosome 1: 48–52 Mb, 2: 86–100.5 Mb, 2: 134.5–138 Mb, 2: 183–190 Mb, 3: 47.5–50 Mb, 3: 83.5–87 Mb, 3: 89–97.5 Mb, 5: 44.5–50.5 Mb, 5: 98–100.5 Mb, 5: 129–132 Mb, 5: 135.5–138.5 Mb, 6: 57–64 Mb, 6: 140–142.5 Mb, 7: 55–66 Mb, 8: 8–12 Mb, 8: 43–50 Mb, 10: 37–43 Mb, 11: 87.5–90.5 Mb, 12: 33–40 Mb, 12: 109.5–112 Mb, and 20: 32–34.5 Mb.

are nonconsecutive). The resulting local-ancestry estimates are $P(\text{population } 1|AA) = \alpha_1 p_1^2/(\alpha_1 p_1^2 + \alpha_2 p_2^2) = 0.31$ and $P(\text{population } 1|CC) = \alpha_1 q_1^2/(\alpha_1 q_1^2 + \alpha_2 q_2^2) = 0.97$, so that the expected value of the ancestry estimate is $E(P(\text{population } 1)) = p\ P(\text{population } 1|AA) + q\ P(\text{population } 1|CC) = 0.74$, a downwardly biased estimate of α_1. More generally, when n perfectly linked markers are used to infer ancestry and are treated as unlinked, for large n (e.g., $n \geq 5$), the evidence of ancestry associated to a particular allele becomes overwhelming, and the estimated ancestry proportion will equal the allele frequency: $P(\text{population } 1|A^n) = \alpha_1 p_1^n/(\alpha_1 p_1^n + \alpha_2 p_2^n) \approx 0$ and $P(\text{population } 1|C^n) = \alpha_1 q_1^n/(\alpha_1 q_1^n + \alpha_2 q_2^n) \approx 1$, so that $E(P(\text{population } 1)) = p\ P(\text{population } 1|A^n) + q\ P(\text{population } 1|C^n) = q = 0.65$. The deficiency of 15% local ancestry, compared to genome-wide ancestry of 80%, shows that the bias could produce effects as large as the 14% deficiencies in European ancestry reported by Tang et al.[1]; such deficiencies will persist when local-ancestry estimates are incorporated into an HMM. In a data set of 112,584 markers, the regions of long-range LD listed in Table 1 would be expected to contain at least 100 markers each. As in our example, unmodeled LD could bias ancestry estimates in the direction of allele frequencies, thereby favoring a deficiency of the population contributing majority ancestry—just as reported in Tang et al.[1]

In addition to their analysis of 112,584 markers, Tang et al.[1] report evidence of selection in analyses of individual HLA markers (Table 1 of their paper). These single-marker analyses are immune to the effects of long-range LD but may be affected by their use of inaccurate ancestral populations to model Puerto Rican ancestry. In particular, the Native American ancestry of Puerto Ricans derives from the Taino, a Native South American population that is likely to be highly genetically diverged from the Native North American populations such as the Pima and Maya used by Tang et al.[1] to model Native American ancestry.[7] Frequency differences among Native American populations could explain why Table 1 of Tang et al.[1] reports a 13% increase in Native American ancestry based on allele frequencies of individual markers at the HLA locus, whereas Figure 1 of Tang et al.[1] reports no deviation in Native American ancestry at the same locus when flanking genomic data were used.[2] We note that if single-marker analyses are affected by the use of inaccurate ancestral populations, analyses of individual markers in new samples from the same populations would not provide an independent replication because the genetic drift underlying the inaccuracy occurs at the population level, not at the individual level.

As an independent test for selection at the chromosome 6 locus, we analyzed 364 new Puerto Rican samples, consisting of 170 individuals with Crohn's disease and 194 matched controls recruited at the University of Puerto Rico School of Medicine. We genotyped these samples at 2459 autosomal markers from our published admixture map that were powerful for distinguishing African from non-African ancestry.[8] (Most markers in the map have relatively similar frequencies in Europeans and Native Americans, with very different frequencies in Africans.) Genotyping was performed with the Illumina Golden Gate technology, and standard quality filters were applied.[9] After additional filtering to exclude markers that were highly differentiated between Europeans and Native Americans (so as to ensure an effective two-way African versus non-African admixture analysis in a three-way admixed population[7]) and disallow LD between markers in the ancestral populations,[10] we retained 1438 markers for downstream analysis. We found that these markers were sufficient to generate useful ancestry estimates: Our calculations indicate that we capture 61% of maximum information about African versus non-African ancestry at the chromosome 6 region, so our effective sample size is $(0.61)(364) = 223$, which is larger than the sample size of 192 in Tang et al.[1]

By using the ANCESTRYMAP software[1] to obtain local-ancestry estimates, we failed to replicate the finding of Tang et al.[1] of an increase in African ancestry at chromosome 6 (Figure 1) and did not observe an unusual deviation in ancestry at any region of the genome. (These results do not shed light on selection signals at the chromosome 8 and 11 regions because Tang et al.[1] reported deviations in European and Native American ancestry at these loci, whereas our 1,438 markers only distinguish African versus non-African ancestry.) To test whether our negative result could be a consequence of low power, we simulated a data set of 364 samples from an admixed population that has 18% African ancestry genome wide but 32% at the chromosome 6 region.[1] In detail, we simulated samples by generating ancestry segments and genotypes at the same set of 1438 markers (with the same pattern of missing data as our Puerto Rican samples) assuming 18% African ancestry, 82% European ancestry, and an average of nine generations since admixture (This quantity was inferred from the Puerto Rican data and is similar to values for other Latino populations.[7]). We preferentially selected samples

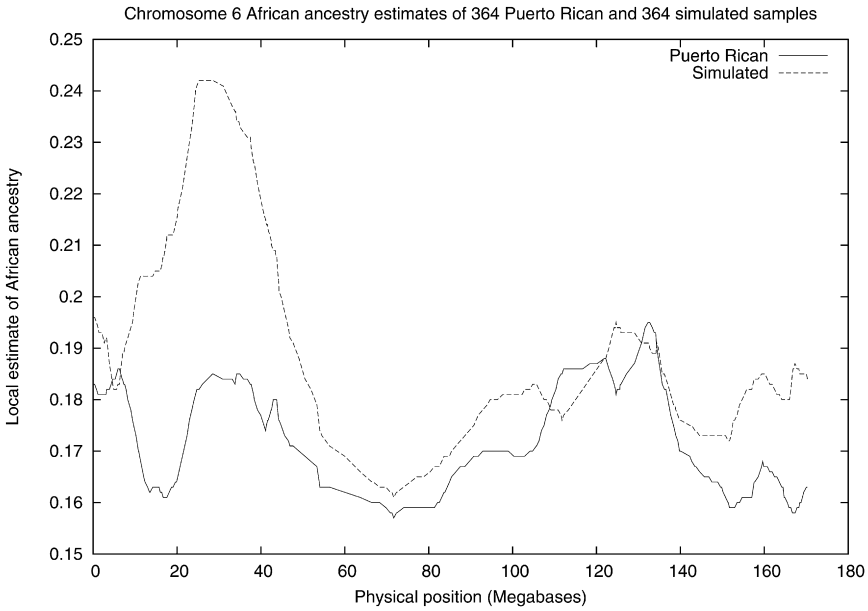

Figure 1. A Replication Study in 364 Puerto Ricans Finds No Significant Rise in African Ancestry at the Chromosome 6 Locus

Local estimates of percent African ancestry on chromosome 6 for 364 Puerto Rican samples and the same number of samples from a hypothetical admixed population simulated to have unusually high African ancestry at the chromosome 6 region centered at position 29.0 Mb as reported in Tang et al.[1] Local ancestry was estimated by ANCESTRYMAP with unlinked markers. We note that the Puerto Rican samples in our study show a slight peak at this region, but this is not significant because there are 41 larger peaks of African ancestry elsewhere in the genome. In contrast, the simulated samples show an excess of African ancestry at this locus, and this is more than twice as large as is observed anywhere else in the genome.

with African ancestry at marker rs451774 (position 28.6 Mb on chromosome 6) so as to achieve 32% African ancestry at this locus. By running ANCESTRYMAP on 364 simulated samples, we detected a large rise in African ancestry at the chromosome 6 region (Figure 1). Although the local estimate of 24% African ancestry at this region is less than the value of 32% used to simulate the data (because ANCESTRYMAP assumes the null model of no unusual deviation in local ancestry and thus imposes a strong prior of 18% African ancestry), the excess of African ancestry is more than twice what is observed anywhere else in the genome. Thus, our failure to identify a rise in African ancestry in Puerto Rican samples on chromosome 6 is not due to a lack of power.

To test the robustness of our negative result, we reran our analysis of the 364 Puerto Rican samples with marker sets chosen to have different thresholds for maximum differentiation between Europeans and Native Americans and reran with all African and European allele-frequency data omitted to ensure that our results were not affected by inaccurate ancestral populations. We also reran with the control individuals only, to ensure that our results were not influenced by the inclusion of Crohn's disease cases. In none of these runs did we observe a signal of a rise in African ancestry at the chromosome 6 locus. The above runs used markers that are not in LD in ancestral populations, as required by ANCESTRYMAP. However, as a demonstration of the pitfalls of not accounting for LD between markers, we reran ANCESTRYMAP on a larger set of 1852 markers in which no constraint was applied to disallow LD in ancestral populations. African-ancestry estimates across the genome varied wildly from 15% to 54%, corresponding to large deficiencies in European ancestry analogous to the signals from Tang et al.[1]

Our analysis demonstrates that the signals of recent selection reported by Tang et al.[1] could theoretically be explained as artifacts caused by regions of long-range LD (with which they strikingly coincide) and inaccurate ancestral populations. Furthermore, we empirically failed to replicate the finding of an unusual deviation in African ancestry at the chromosome 6 region in our analysis of a larger Puerto Rican sample set. We believe that the hypothesis of selection since admixture should therefore be viewed with caution. We note that in a joint analysis of more than 10,000 African American samples that we have scanned in admixture-mapping studies, we have not yet found a single locus at which there is signal of a local-ancestry deviation that is not specific to disease cases. We consider it unlikely that recent selection events could lead to three distinct local-ancestry deviations that are large enough to be detected with only 192 Puerto Rican samples, when we failed to detect any such effect in African Americans using >50-fold more samples.

These results also have methodological significance for genome-wide association studies in admixed populations such as Latinos and African Americans. To have maximum power, such studies need to take advantage of admixture association signals (deviations in local ancestry in disease cases compared to their genome-wide average) as well as case-control association signals. The method of Tang et al.[2] has been shown to accurately infer ancestry in simulated data sets, but our results suggest that it may produce false-positive admixture association signals in regions of long-range LD in admixed populations. In association studies, such errors can be controlled by computation of local-ancestry estimates in both cases and controls. However, case-only admixture association analyses are known to provide higher statistical power.[11] Thus, carrying out robust, fully powered genome-wide association studies in admixed

populations will require methods that rigorously account for the confounding effects of long-range LD.

Alkes L. Price,[1,2] Michael E. Weale,[3] Nick Patterson,[2] Simon R. Myers,[2,4] Anna C. Need,[3] Kevin V. Shianna,[3] Dongliang Ge,[3] Jerome I. Rotter,[5] Esther Torres,[6] Kent D. Taylor,[5] David B. Goldstein,[3] and David Reich[1,2,*]
[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; [2]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [3]Institute for Genome Sciences and Policy, Duke University, Durham, NC 27710, USA; [4]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK; [5]Cedars-Sinai Medical Center, University of California Los Angeles, Los Angeles, CA 90048, USA; [6]University of Puerto Rico School of Medicine San Juan, PR 00936, USA
*Correspondence: reich@genetics.med.harvard.edu

Acknowledgments

A.L.P. is supported by a Ruth Kirschstein National Research Service Award from the NIH. N.P. is supported by a K-01 career development award from the NIH. D.R. is supported by a Burroughs Wellcome Career Development Award in the Biomedical Sciences. This research was also supported by U-01 award HG004168 from the NIH (D.R. and N.P.), by NIDDK grant PO1DK46763 (J.I.R.), and by the Board of Governor's Chair in Medical Genetics at Cedars-Sinai Medical Center (J.I.R.). Genotyping of the Puerto Rican samples was supported in part by grant M01-RR00425 to the Cedars-Sinai GCRC genotyping core (K.D.T.) and by NIH grant DK62413 (K.D.T).

References

1. Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Clinton, W., Burchard, E.G., and Risch, N.J. (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. Am. J. Hum. Genet. *81*, 626–633.
2. Tang, H., Coram, M., Wang, P., Zhu, X., and Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. Am. J. Hum. Genet. *79*, 1–12.
3. Wall, J.D., and Pritchard, J.K. (2003). Linkage disequilibrium in the human genome. Nat. Rev. Genet. *4*, 587–597.
4. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
5. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet *2*, e190.
6. Tian, C., Plenge, R.M., Ransom, M., Lee, A., Villoslada, P., Selmi, C., Klareskog, L., Pulver, A.E., Qi, L., Gregersen, P.K., et al. (2008). Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet *4*, e4.
7. Price, A.L., Patterson, N., Yu, F., Cox, D.R., Waliszewska, A., McDonald, G.J., Tandon, A., Schirmer, C., Neubauer, J., Bedoya, G., et al. (2007). A genomewide admixture map for Latino populations. Am. J. Hum. Genet. *80*, 1024–1036.
8. Smith, M.W., Patterson, N., Lautenberger, J.A., Truelove, A.L., McDonald, G.J., Waliszewska, A., Kessing, B.D., Malasky, M.J., Scafe, C., Le, E., et al. (2004). A high-density admixture map for disease gene discovery in African Americans. Am. J. Hum. Genet. *74*, 1001–1013.
9. Fan, J.B., Oliphant, A., Shen, R., Kermani, B.G., Garcia, F., Gunderson, K.L., Hansen, M., Steemers, F., Butler, S.L., Deloukas, P., et al. (2003). Highly parallel SNP genotyping. Cold Spring Harb. Symp. Quant. Biol. *68*, 69–78.
10. Reich, D., Patterson, N., De Jager, P.L., McDonald, G.J., Waliszewska, A., Tandon, A., Lincoln, R.R., DeLoa, C., Fruhan, S.A., Cabre, P., et al. (2005). A whole-genome admixture scan finds a candidate gene for multiple sclerosis susceptibility. Nat. Genet. *37*, 1113–1118.
11. Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O'Brien, S.J., Altshuler, D., et al. (2004). Methods for high-density admixture mapping of disease genes. Am. J. Hum. Genet. *74*, 979–1000.

DOI 10.1016/j.ajhg.2008.06.005.

Response to Price et al.

To the Editor: In 2006, Tang and colleagues[1] presented a novel statistical method for genetic admixture analysis based on high-density SNP arrays rather than conventional ancestry informative markers (AIMs). The chromosomes of an admixed individual represent a consecutive patchwork of ancestry blocks representing the ancestral populations contributing to the admixed individual. Their approach[1] is based on the probabilistic reconstruction of those chromosomal ancestry blocks within single individuals. From the block reconstructions, estimates of ancestry at any location in the genome can be derived. The authors recognized that high-density SNP arrays could include nearby markers that are in linkage disequilibrium (LD) in the ancestral population and that such LD could contribute noise to the block reconstructions and subsequent locus-specific ancestry estimation. Therefore, they proposed a Markov-Hidden Markov Model (MHMM) that allowed for pairwise dependency between adjacent markers in the ancestral populations in the estimation process and developed a computer program (SABER) to perform these calculations. They showed, through extensive simulations with data derived from the HapMap project,[2] that the method was robust in reconstructing ancestry blocks, even for very dense sets of markers and for an individual with three ancestral components, and when some of the model parameters were misspecified.[1] Subsequently, Tang et al.[3] used the MHMM to reconstruct ancestry blocks from Affymetrix 100K data in a sample of 192 Puerto Ricans from the Genetics of Asthma in Latino Americans (GALA) study[4] and examined the genome-wide distribution of African, European, and Native American ancestry in this sample.