File S1

SUPPLEMENTARY INFORMATION

Wavelet transform method

Wavelets can be thought of as *localized waves* or oscillations, where localization in this context refers to a region of SNPs along the genome (see **Figure S5**). Our implementation utilizes families of discrete non-decimated wavelets $\{\psi_{j,k}\}$, where $j = 1, \ldots, J$ denotes the scale of the wavelet (related to Fourier frequency) and k denotes the location (i.e., SNP number). For detailed introductions to wavelets and their use in statistics, see Nason (2008) and Vidakovic (2009), or the review by Liò (2003) for an introduction to the use of wavelets in biostatistics.

The importance of localization can be appreciated by contrasting wavelets to the sinusoids (big waves) used in classical Fourier analysis. Sinusoids are associated with a particular frequency, but do not have the location component provided by wavelets. The fact that each wavelet is associated with a particular small genomic region means that they can capture the structure of the data, especially where the admixture tracts are not uniformly distributed over the chromosome.

The wavelet periodogram of a signal is given by the square of the raw wavelet coefficients

$$I_{j,k}^{(i)} = |\sum_{t=0}^{T-1} X_t \psi_{j,k}(t)|^2 \quad (7)$$

The wavelet periodogram for one individual is shown in **Figure S6A**. The wavelet transform provides a decomposition of the data in terms of location along the genome on the *x*-axis and wavelet scale on the *y*-axis.

For the simulated data with 13,000 SNPs, the maximum number of wavelet scales in the decomposition is 13 ($J \leq log_2(13000)$). Scale 1 captures the highest frequency, very local information. Increasing the scale index provides successively coarser, or lower frequency information, zooming out of the signal until we reach the level of the entire chromosome. For an individual chromosome, the information is *nonstationary* in that the width of the admixture tracts can vary over the chromosome. At the population level, the wavelet transforms show greater evidence of stationarity, as demonstrated in **Figure S6B**. The population average periodogram has a smoother appearance when examined from left to right (i.e., along the genome) within each scale. The information can therefore be conveniently summarized by summing the wavelet coefficients within each scale to give the wavelet variance.

The discrete wavelet transform (DWT) has also been used in a similar context. Our implementation makes use of the Maximal Overlap Discrete Wavelet Transform (MODWT), which has several benefits over the DWT. First, there is no restriction that the data need be a power of two, which means it can be applied directly to the available genetic data without first windowing the signal or down-sampling the data.

The resulting wavelet coefficients are translation-equivariant, meaning that circularly shifting the data results in the same shifting of the coefficients. With the DWT, shifting the data could lead to a different decomposition. We note that other localized decompositions, such as the short time Fourier transform or continuous wavelet transform, could also be applied.

Pre-windowing of the admixture signal and visualization

Both *PCAdmix* (Brisbin *et al.* 2012) and *StepPCO* (Pugach *et al.* 2011) compute an averaged admixture signal in predefined localized windows along the genome. Our approach instead uses the SNP level information and offers several advantages. It avoids the subjective choice of properties of the signal (window width and number of bins), and ensures that the information is considered at the most detailed level possible. Subsequent wavelet analysis then considers the data in localized windows, the width of the window increasing as we zoom out to coarser scales. Whether information at a particular window size is informative is determined by reference to the variation observed in the ancestral populations. The informative variation is therefore extracted in an objective, data driven manner.

One possible advantage of pre-windowing the signals is in visualizing local ancestry. Windowing reduces high frequency noise and produces signals that are more easily related to ancestry by eye. For example, applying a window of W SNPs, the signals can be computed as

$$\tilde{Y}_s^i = \frac{1}{W}\sum_{n\in w_n} X_{n,i}^{'} v_{1,n} \tag{8}$$

$$Y_s^i = \begin{cases} \dfrac{2\tilde{Y}_s^i - (\overline{Y}_s^B + \overline{Y}_s^A)}{(\overline{Y}_s^B - \overline{Y}_s^A)}, & |\overline{Y}_s^B - \overline{Y}_s^A| \geq \epsilon \\ 0, & |\overline{Y}_s^B - \overline{Y}_s^A| < \epsilon \end{cases} \tag{9}$$

where $\overline{Y}_s^G = \frac{1}{n_G}\sum_{i\in P_G} Y_s^i$ for $G = A, B$. Subsequent wavelet analysis of the pre-windowed signals would have the same interpretation, as illustrated in **Figure S7**.

Choice of measurement scale

The raw admixture signals are estimated at each SNP location (see **equation 2**) or for each SNP window if pre-windowing is implemented (**equations 8 and 9**). It is also possible to construct the signals in terms of genetic distance along the chromosome (as opposed to physical distance). Both options are implemented in the *adwave* software.

Threshold choices

To extract the informative variation in cases where high levels of noise are present in the signals (e.g., at very low admixture proportions), a higher threshold for μ could be selected. In setting a tougher criterion, this ensures that the raw wavelet variance must be larger before we are willing to accept that it is informative about the admixture process rather than simply being noise. Choice of threshold is a balance

between two extremes: too high a threshold may remove informative variation along with the noise, while a weak threshold may result in noise contamination and potentially biased summary measures, such as the ABS metric. The choice of threshold is necessarily data dependent, but we advocate altering the default value only for rare cases that exhibit evidence of high noise.

The effects of varying μ are illustrated in **Figure S8** for two simulated data sets; one with low levels of noise in the resulting admixture signals, and the other with high levels. In this example, a low admixture proportion from one of the ancestral populations is used to mimic the effect of "high noise" in the admixture signals, but the results are also applicable to other sources of noise, such as short divergence times between the ancestral populations (see Discussion in the main text). In low noise situations, the ABS metrics are strongly robust to the choice of μ, while for high noise situations, a larger value of μ is necessary to avoid bias in the summary measures. The recommended procedure for selecting μ is to produce initial results using the default value, and then increase μ only if there is evidence of low-scale noise. An automatic method for selecting μ may be considered in future work.

Also note that any bias due to non-optimal choice of μ is avoided in the ABC dating procedure by ensuring that the same value is used for both the simulated and sampled data.

Sensitivity to method options

The default method options are to estimate the raw admixture signals for each SNP location, constructed according to physical distance (as opposed to genetic distance) without windowing the signals, and using Daubechies' Least Asymmetric wavelet number 8.

Other options are also implemented in the *adwave* software, providing flexibility that may be required for different applications. Sensitivity to the different options was considered by mimicking the results of the admixture time example for variations on the default options. A summary of these results is presented in **Table S2** and **Figure S9**.

In this instance, results using the Haar wavelet (condition 9) are very similar to the MOWDT default. The slight variation in the ABS metrics is expected since different wavelets cover slightly different frequency ranges, although the effect on the results is insubstantial.

The effect of pre-windowing the signals is illustrated for two window sizes: 130 SNPs (condition 10) and 65 SNPs (condition 11). Choice of window size clearly modifies the relationship between admixture time and the resulting ABS metrics. As illustrated by condition 10, if the window size is too large, it will not be able to capture the small admixture blocks characteristic of ancient admixture. Lack of windowing as the default approach is a major point of advantage of *adwave* over *StepPCO*.

When constructing signals in terms of genetic distance, it is necessary to specify the number of bins for the signal and the size of the analyzing window. The example represented by condition 12 utilizes 13,000 bins (i.e., one per SNP), and small windows of 13 SNPs so that the resulting decomposition is over the same number of wavelet scales and the effect of windowing is minimized. This choice of options provides results that are consistent with the default.

The default options provide ease of implementation, avoiding the subjective choice of properties of the signal (window width and number of bins), and ensuring that the signal information is considered at the most detailed level possible. Nevertheless, experienced users are free to vary these parameters.

Method comparison: a demonstration for one population

Using *StepPCO*, formation of the localized admixture signals requires specification of the number of bins in the signal and a tolerance for the window size. Pugach *et al.* (2011) recommend that the number of bins should be chosen so that the windows span the entire chromosome, leaving no gaps in between. For their wavelet analysis, it is a strict requirement that the number of bins is a power of two.

Window size is allowed to vary along the chromosome and is specified via an automatic method, for which it is necessary to set a tolerance λ. Starting with a small window of SNPs, window size is increased until the mean PCA coordinates of the ancestral populations are separated by λ standard deviations. Pugach *et al.* (2011) use

$K = 1024$ and $\lambda = 3$ for their implementation. For our demonstration, we have used a stronger window criterion of $\lambda = 5$ to ensure that the localized windows cover the entire chromosome.

To produce the wavelet summaries, *StepPCO* uses a three stage filtering procedure:

1. Coefficients smaller than a specified threshold are set to zero, to remove low amplitude oscillations. This parameter was set to 0.1 for our application, following advice stated in the accompanying software manual.
2. Wavelet scales that correspond to high frequencies are deemed characteristic of noise and removed completely. For guidance on setting this option, the manual states that it depends on the length of the chromosome and suggests a maximum scale of 7, 6 and 5 for chromosomes 1-5, 6-20 and 21-22, respectively. For our example, we truncate at 6 scales, since the number of SNPs in the example is comparable to chromosomes 6-10 in the *StepPCO* paper.
3. A scale dependent threshold is then applied. The threshold is computed by averaging the wavelet coefficients across each scale, and subtracting the maximum value observed in the ancestral populations.

Stage 3 in this procedure is similar to the *adwave* thresholding process described by Equation 5, but in *adwave*, this correction is based on population averages rather than individual-level values. *StepPCO* therefore uses a stronger threshold than the *adwave* procedure (i.e., it removes more of the raw information).

With *adwave*, it is not necessary to pre-window the signal. A method demonstration is shown in **Figure 1**, using the raw SNP-level data and default threshold value of μ = 1. However, in order to provide a closer comparison with *StepPCO*, we also provide results using options similar to those applied by Pugach *et al.* (2011). The localized admixture signals were formed using $N = 1024$ points along the chromosome, sampled according to genetic distance with a fixed window size of 13,000 × 0.0025 = 37 SNPs (chosen to mimic the mean window size obtained by *StepPCO*).

A comparison of both methods for one simulated population with $T = 160$ is provided in **Figure S3**. The admixture signals produced by *StepPCO* have a variable window size of 2 to 195 SNPs with mean 39.2 and median 29. The variable window size can sometimes lead to instability in the signals (shown by the 'spikes' in **Figure S3A**, which correspond to windows with small numbers of SNPs). It is possible to set upper and lower bounds for the number of SNPs per window, but this requires more user choice of runtime settings.

The raw *StepPCO* wavelet summaries presented in **Figure S3B** are similar to those obtained by *adwave* (**Figure S3E**), but exhibit a larger amount of high-frequency noise, as is apparent in all three populations. The final *StepPCO* wavelet summaries (**Figure S3C**) look similar to the final informative wavelet variance of *adwave* (**Figure S3F**), but without the four highest-frequency scales. Truncation of these high-frequency scales

will have a particularly large influence for older admixture events, an issue that is mentioned in the Pugach *et al.* (2011) paper.

LITERATURE CITED

Brisbin, A., K. Bryc, J. Byrnes, F. Zakharia, L. Omberg *et al.*, 2012 PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum. Biol. 84: 343–364.

Cox, M. P., A. E. Woerner, J. D. Wall, and M. F. Hammer, 2008 Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. BMC Genetics 9: 76.

Liò, P., 2003 Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics 19: 2–9.

Nason, G., 2008 *Wavelet Methods in Statistics with R*. Springer, New York ; London.

Pugach, I., R. Matveyev, A. Wollstein, M. Kayser, and M. Stoneking, 2011 Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biology 12: R19.

Vidakovic, B., 2009 *Statistical Modeling by Wavelets*. John Wiley & Sons: New York.