Published paper

Sanderson, J., Sudoyo, H., Karafet, T.M., Hammer, M.F. and Cox, M.P. (2015) *Reconstructing past admixture processes from local genomic ancestry using wavelet transformation.* Genetics, 200 (2). 469 – 481

10.1534/genetics.115.176842

Reconstructing Past Admixture Processes from Local Genomic Ancestry Using Wavelet Transformation

Jean Sanderson,*,[1] Herawati Sudoyo,[†] Tatiana M. Karafet,[‡] Michael F. Hammer,[‡,§] and Murray P. Cox*,[2]
*Statistics and Bioinformatics Group, Institute of Fundamental Sciences, Massey University, Palmerston North 4442, New Zealand, [†]Eijkman Institute for Molecular Biology, Jakarta, Indonesia, [‡]Division of Biotechnology, Arizona Research Laboratories and [§]Department of Anthropology, University of Arizona, Tucson, Arizona 85721

ABSTRACT Admixture between long-separated populations is a defining feature of the genomes of many species. The mosaic block structure of admixed genomes can provide information about past contact events, including the time and extent of admixture. Here, we describe an improved wavelet-based technique that better characterizes ancestry block structure from observed genomic patterns. principal components analysis is first applied to genomic data to identify the primary population structure, followed by wavelet decomposition to develop a new characterization of local ancestry information along the chromosomes. For testing purposes, this method is applied to human genome-wide genotype data from Indonesia, as well as virtual genetic data generated using genome-scale sequential coalescent simulations under a wide range of admixture scenarios. Time of admixture is inferred using an approximate Bayesian computation framework, providing robust estimates of both admixture times and their associated levels of uncertainty. Crucially, we demonstrate that this revised wavelet approach, which we have released as the R package *adwave*, provides improved statistical power over existing wavelet-based techniques and can be used to address a broad range of admixture questions.

KEYWORDS wavelets; principal component analysis (PCA); admixture; local ancestry; dating

ADMIXTURE occurs when previously separated populations interact and merge. This process has been instrumental in human history, with most global groups showing at least some signals of population merger (Hellenthal *et al.* 2014). The admixture process produces "mosaic" genomes with alternating blocks of DNA from each ancestral population. Over time, recombination decreases the length of these ancestry blocks, and therefore the distribution of block sizes is informative about the time of admixture. However, the extent to which these patterns can provide additional information about historic admixture processes is still a young area of exploration.

A range of methods have been developed to partition the genome of an admixed individual into ancestry blocks based on raw genomic data (Falush *et al.* 2003; Price *et al.* 2009). Some methods assign ancestry directly. For instance, *HAPMIX* uses a hidden Markov model to estimate the break points of ancestry blocks, while other approaches define ancestry blocks using simple empirical criteria, such as strings of shared *vs.* nonshared polymorphisms (Pool and Nielsen 2009) or the differential presence of population-specific variants (Brown and Pasaniuc 2014). Another set of methods is more indirect. *ROLLOFF* (Moorjani *et al.* 2011), *LAMP* (Baran *et al.* 2012), and *ALDER* (Loh *et al.* 2013) all search for rapid changes in linkage disequilibrium to define the borders of ancestry blocks, while other approaches assign ancestry for predefined genomic windows using conditional random fields (Maples *et al.* 2013) or principal component analysis (PCA) (Gravel 2012).

These methods vary in their effectiveness. Simple empirical criteria perform surprisingly well for admixture between species (as for the mouse admixture zone studied by Pool and Nielsen 2009). Similarly, most of these methods tend to be highly accurate for recent admixture between well-separated human groups (such as African Americans or American Latinos). Indeed, in these settings, subtleties such as multiple waves of admixture have even be detected (Gravel 2012).

doi: 10.1534/genetics.115.176842
Manuscript received October 29, 2014; accepted for publication April 3, 2015; published Early Online April 7, 2015.
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176842/-/DC1.
[1]Present address: School for Health and Related Research, University of Sheffield, Sheffield, United Kingdom.
[2]Corresponding author: Institute of Fundamental Sciences, Massey University, Private Bag 11 222, Palmerston North 4442, New Zealand. E-mail: m.p.cox@massey.ac.nz.

However, reconstructing complex demographic features for much older admixture events (*i.e.*, thousands rather than hundreds of years in the past) remains extremely challenging (Moorjani *et al.* 2011). While methods have in principle been proposed to detect multiple waves of ancient admixture, in many realistic settings they are still restricted to single admixture events (Loh *et al.* 2013), although some evidence for multiple ancient admixture events has been presented for several Indian populations (Moorjani *et al.* 2011).

Other indirect methods look increasingly promising in this "old admixture" space. Approaches based on principal components analysis and wavelets have been employed with some success. PCA is a nonparametric data-reduction technique, which has been used widely to identify patterns of population structure in genetic data (Patterson *et al.* 2006; Novembre and Stephens 2008; McVean 2009; Bryc *et al.* 2010; Ma and Amos 2012). Dispersion of admixed individuals along the first principal component connecting ancestral populations can be used as a diagnostic for two-way admixture (Patterson *et al.* 2006; Mcvean 2009). For instance, *PCAdmix* employs PCA to assign ancestry to localized windows along the genome for each individual (Brisbin *et al.* 2012). Pugach *et al.* (2011) also use PCA, but do not directly assign ancestry to genomic regions, instead applying a wavelet transform to obtain an indirect measure of the average admixture block length. While this approach has been shown to be powerful for dating old admixture events, there remains considerable scope for (i) the development of more sophisticated wavelet constructions, (ii) examining the resulting wavelet decompositions in greater detail (particularly to identify aspects of non-time-related information in the transformed data), and (iii) to provide a more user-friendly software solution for wavelet analysis.

Wavelet techniques themselves are an active and evolving area, with much potential for novel application in population genetics, as highlighted in the review article by Liò (2003). Wavelets can be thought of as localized, oscillatory functions and are particularly useful for representing data that has local features such as sharp changes and discontinuities. In the context of genome-wide single nucleotide polymorphism (SNP) data, wavelets can be used to represent the mosaic pattern of ancestry blocks. A wavelet decomposition of the data provides information on the size of the ancestry blocks and, importantly, how they are distributed along the chromosomes. Summary measures of the wavelet decomposition allow aspects of the admixture process to be reconstructed, such as the time of admixture and admixture proportions.

Here, we present a substantially revised wavelet-based approach to describe population admixture that builds on the work of Pugach *et al.* (2011). This new method has significantly fewer model assumptions and allows us to identify more complex demographic processes, such as multiple admixture events. As with previous methods, PCA is first employed to describe the population structure. The maximal overlap discrete wavelet transform (MODWT) is then applied directly to the SNP-level data, without the need to compute averages over localized genomic windows as implemented in related procedures (Pugach *et al.* 2011; Brisbin *et al.* 2012). Instead, windowing is performed naturally and objectively as part of the wavelet decomposition procedure. We show that this new method provides robust estimates of admixture time (including improved control of uncertainty estimates), as well as recognizing other aspects of admixture processes that previous wavelet-based methods have not been able to identify with any accuracy.

Methods

General framework

Initially, we consider a simple admixture scenario where two ancestral populations P_A and P_B merged T generations ago to form the admixed population P_C. The ancestral populations contribute to the admixed population with probabilities p and $1-p$. The sizes of the populations, the admixture time, and the admixture proportions are free to vary.

To quantify patterns of genomic block size variation, a three-step analysis procedure was used: (i) PCA was applied to the genomic data to describe population structure; (ii) the wavelet variance was computed to provide a scale-by-scale decomposition of the variance for each population; and (iii) the portion of this measure that is informative for admixture processes was extracted relative to background levels observed in the ancestral populations.

Data simulation

Genome-wide SNP data were simulated using the sequential coalescent simulator *MaCS* (Chen *et al.* 2009). Because our primary interest is in the admixture history of Island Southeast Asia (see *Real genomic data* section below), we chose parameter settings that produce genomic data that broadly fit observed patterns of genetic diversity in this study region (Cox *et al.* 2008). The demographic model, parameters, and information sources are described in more detail in the Supporting Information (Figure S1). We emphasize, however, that the method we describe is general and can be applied to most admixed genomic systems.

Data setup

Given an admixed population P_C derived from two ancestral populations P_A and P_B, the number of individuals in the analysis (*i.e.*, present day samples) is $n = n_A = n_B + n_C$. For each individual i, we observe a collection of T SNPs along a chromosome. Thus the raw data matrix $\boldsymbol{X}$ is a $T \times n$ matrix with T genotype counts in columns and n individuals in rows. The SNPs s are ordered by their physical positions along the chromosome, with the cells of the data matrix $X_{s,i}$ taking the value 0 if heterozygous, and arbitrarily -1 or 1 for the alternative homozygous states. Prior to principal components analysis, the data matrix is centered such that the column mean with respect to the ancestral reference populations is zero, giving

$$X'_{s,i} = X_{s,i} - \frac{1}{n_A + n_B} \sum_{i \in P_E, P_B} X_{s,i}.$$

Principal components analysis

PCA is performed using only individuals from the ancestral populations. Rather than performing PCA on all samples combined, this approach has the advantage that other features of the admixed sample (such as admixture from additional ancestral populations) will not influence the projection (McVean 2009). The first eigenvector v_1 reflects the primary population structure. Projection of individuals onto this axis of variation is given by

$$y_1^i = \sum_{s=1}^{T} X'_{s,i} v_{1,s}. \quad (1)$$

The proportion of ancestry inherited from population P_A can be estimated for each individual (or population) using the distance from the centroids of the ancestral populations; that is, $p_i = (c_B - y_1^i)/(c_B - c_A)$, where $c_A = (1/n_A)\sum_{i \in P_E} y_1^i$ and $c_B = (1/n_B)\sum_{i \in P_B} y_1^i$ are the centroids of the ancestral populations along the first principal axis (Bryc *et al.* 2010). Note that variation between individuals within a population is represented by the smaller eigenvalues and corresponding eigenvectors (Ma and Amos 2010).

This representation of admixed individuals in PCA space, as shown in Figure 1A, provides a genome-wide estimate of average ancestry, but does not indicate how admixture tracts are distributed along the chromosomes. To obtain localized estimates, the projection is performed at the SNP level rather than summing over the length of the genome as in Equation 1. The raw SNP-level admixture signals are given by

$$Y_s^i = \begin{cases} \dfrac{2X'_{s,i} v_{1,s} - (\overline{Y}_s^B + \overline{Y}_s^A)}{(\overline{Y}_s^B - \overline{Y}_s^A)}, & \left|\overline{Y}_s^B - \overline{Y}_s^A\right| \geq \varepsilon \\ 0, & \left|\overline{Y}_s^B - \overline{Y}_s^A\right| < \varepsilon \end{cases}, \quad (2)$$

where $\overline{Y}_s^G = (1/n_G)\sum_{i \in P_G} X'_{s,i} v_{1,s}$ for $G \in A,\ B$. The additional terms in Equation 2 ensure that the signals are normalized such that the mean of the ancestral populations are arbitrarily 1 and -1. This normalization step makes the measure robust to uneven sample sizes, which can affect the structure of the PCA (Novembre and Stephens 2008; Mcvean 2009). Stability of the signals is maintained by specifying a tolerance $\mathcal{E}$ for separating the ancestral populations at a given SNP. This ensures that SNPs with poor discrimination are treated as uninformative in the next step of the analysis.

Wavelet transform

The resulting SNP-level admixture signals indicate how ancestry varies along the genome, but they invariably exhibit a high noise-to-information ratio. To interpret the signal, its frequency content can be described using the wavelet variance (Percival 1995). The wavelet variance S_j for scales $j = 1, \ldots, J$ provides a scale-by-scale decomposition of the variance of the signal. The first scale ($j = 1$) captures the highest frequency patterns, representing very local information. Increasing the scale index provides successively coarser, or lower frequency information, equivalent to "zooming out" on the signal until the level of the entire chromosome is reached. A plot of S_j *vs.* j indicates which scales are important contributors to the process variance and indirectly provides information about the distribution of admixture tracts. For example, recent admixture produces a peak in the wavelet variance at a large wavelet scale, reflecting long admixture tracts, while more ancient admixture events produce peaks at lower wavelet scales, reflecting shorter admixture tracts.

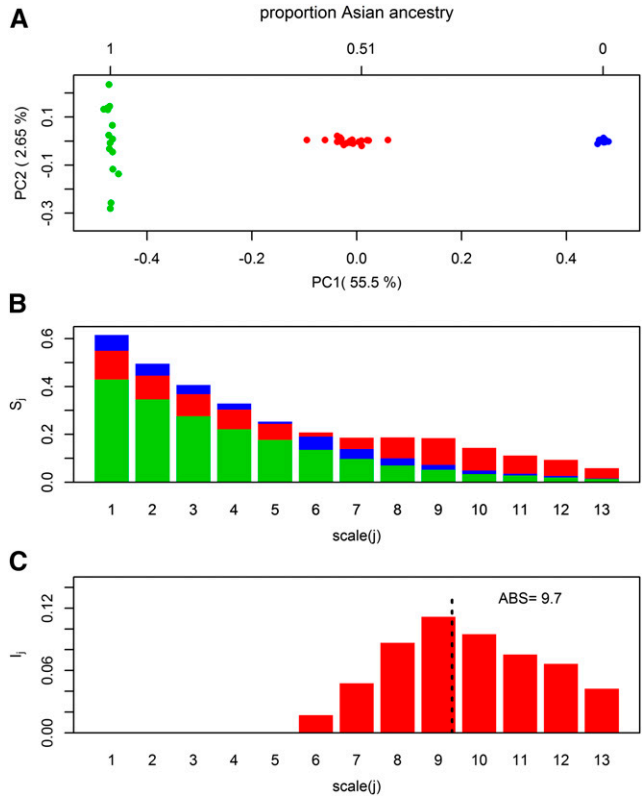


Figure 1 Simulated example with 13,000 SNPs, 15 diploid individuals in ancestral populations (P_A, P_B), and 20 diploid individuals in the admixed population (P_C). Populations are shown in green (P_A), blue (P_B), and red (P_C). (A) PCA is used to describe the primary population structure; (B) raw wavelet variance for each population illustrates high frequency noise; (C) informative variation in the admixed population after standard correction for noise estimated from the ancestral populations. Note that this example uses the default threshold $\mu = 1$.

The wavelet variance for an individual i is given by

$$S_j^i = \frac{1}{T} \sum_{k=1}^{T} |d_{j,k}^i|^2, \quad (3)$$

where $d_{j,k}^i = \sum_s Y_s^i \psi_{j,s-k}$ are the wavelet coefficients for the signal Y^i constructed using the wavelet system ψ. To

appreciate the methodology, it is sufficient to understand that the wavelet variance reflects the frequency content of the signal, but more detailed background material is provided in the Supporting Information (Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, and File S1). Our implementation employs Daubechies' least asymmetric wavelet number 8 in the *waveslim* (Whitcher 2013) package of the statistical software R (R Development Core Team 2014). We emphasize, however, that the methods proposed here are robust to other choices of analyzing wavelet (see Supporting Information, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, and File S1).

Population averages are computed as

$$S_j^{\mathrm{C}} = \frac{1}{n_{\mathrm{C}}} \sum_{i \in P_{\mathrm{C}}} S_j^i$$

(and similarly for populations P_{A} and P_{B}). An example of the average wavelet variance for each population is shown in Figure 1B. The wavelet variance is highest at fine scales, but as the ancestral populations also show this pattern, it should be considered background noise. It is intuitive that the very finest wavelet scales are uninformative because small numbers of SNPs should be insufficient to differentiate between populations. The raw wavelet variance is therefore considered as a combination of informative variation and background noise

$$S_j^{\mathrm{C}} = I_j^{\mathrm{C}} + N_j. \tag{4}$$

To extract the informative variance I_J^{C}, we subtract the proportion that can be attributed to noise. This is estimated from the variation observed in the ancestral populations; $\hat{N}_j = \mu \cdot \max(\overline{S}_j^{\mathrm{A}}, \overline{S}_j^{\mathrm{B}})$, where μ is a multiplicative factor that allows the degree of thresholding to be controlled. Under almost all conditions, a default value of $\mu = 1$ may be assumed, and this threshold should be raised only if the admixture signals exhibit high levels of noise (see Supporting Information, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, and File S1 for details). Population characteristics that influence noise levels in the admixture signals are explored in the next section. The final measure of the informative variance is given by

$$I_J^{\mathrm{C}} = \max\left(S_J^{\mathrm{C}} - \hat{N}_j, 0\right), \tag{5}$$

which describes the frequency content that is unique to the admixed population (in contrast to the ancestral populations).

Real genomic data

To illustrate that our method performs well in real-world situations, it was applied to a SNP genotyping chip data set of 394 individuals from 16 communities spread across the Indonesian archipelago (Table 1). Equivalent SNP data from Southern Han Chinese and Papua New Guinea Highlanders were used as proxies for the ancestral populations. Permission to conduct research in Indonesia was granted by the Indonesian Institute of Sciences. Blood samples or buccal swabs were collected from consenting, closely unrelated, and seemingly healthy individuals by J. Stephen Lansing (University of Arizona) and Herawati Sudoyo (Eijkman Institute for Molecular Biology, Indonesia), with the assistance of Indonesian Public Health clinic staff. All sample collection followed protocols for the protection of human subjects established by both the Eijkman Institute and the University of Arizona institutional review boards. Participant interviews confirmed local residence for at least two generations into the past. Samples were genotyped with the Affymetrix Axiom chip, yielding 548,994 SNPs across the autosomes. (Sex-linked markers were excluded from the analysis.) The SNP data were cleaned using standard protocols in PLINK v. 1.07 (Purcell *et al.* 2007; Purcell 2009) and the wavelet transform performed as described above.

The approximate Bayesian computation analysis employed 1000 data sets with sample sizes and SNP numbers set to those of the real data. These data sets were simulated by drawing from a uniform prior of admixture times between 10 and 300 generations. The admixture proportion for the Bena population (used as our primary test case) was set to 0.6, as estimated previously from the real data. The ABS metric was calculated for each simulation, and the multiple chromosome structure of the data were mimicked by sampling each individual repeatedly with different data densities.

Results

As proof of concept, we first applied our wavelet method to simulated data. A range of admixture scenarios was explored by varying parameters of the demographic model, particularly the time of admixture, admixture proportion, and single *vs.* multiple admixture events. Fifty simulations were performed for each scenario with modest (but therefore realistic) ancestral sample sizes of $n_{\mathrm{A}} = n_{\mathrm{B}} = 15$ and an admixed sample size $n_{\mathrm{C}} = 20$.

Admixture time

Because the ability of wavelet methods to calculate the time of admixture is well known from earlier work (Pugach *et al.* 2011), we explored this feature first. Simulations were performed for admixture times ranging from 10 to 320 generations (*i.e.*, from the recent past to ~10,000 years ago, using a generation interval of 30 years; Fenner 2005). Admixture at 10 generations shows the highest informative wavelet variance at scale 13, reflecting relatively few, long admixture blocks (Figure 2). As the time of admixture occurs further back in the past, the peak in wavelet variance shifts toward successively lower wavelet scales, reflecting ever-smaller admixture blocks driven by cumulative recombination along the chromosome. The average frequency content can be characterized by the average block size metric ABS, termed the "wavelet center" by Pugach *et al.* (2011), which as shown later, can be used to date the admixture event

Table 1 Summary of case study populations describing sample size (*n*), proportion of Asian ancestry as inferred by PCA (*p*), and the average block size metric (ABS, for admixed populations only)

Population	*n*	*p*	Average block size metric (ABS)
Southern Han Chinese	13	1.00	–
Nias	28	0.87	4.26
Mentawai	29	0.87	4.30
Java	21	0.84	4.41
Sumatra	30	0.83	4.49
Bali	19	0.83	4.90
Sulawesi	21	0.80	6.57
Sumba, Wunga	30	0.67	7.90
Sumba, Anakalang	30	0.66	7.83
Flores, Rampasasa	12	0.66	8.05
Flores, Bena	30	0.57	8.14
Flores, Bama	30	0.55	8.34
Timor, Umanen Lawalu	17	0.55	8.44
Timor, Kamanasa	19	0.53	8.42
Lembata	28	0.53	8.39
Pantar	27	0.45	8.47
Alor	23	0.42	8.46
Papua New Guinea Highlands	13	0.00	–

$$\mathrm{ABS} = \frac{\sum_j j \cdot \bar{I}_j^{\mathrm{AB}}}{\sum_j \bar{I}_j^{\mathrm{AB}}}. \qquad (6)$$

Admixture proportion

Admixture proportions were varied between 0.5 (equal ancestry from P_A and P_B) and 0.025 (ancestry predominately from P_A). For this analysis, the time of admixture was fixed at 160 generations. As the proportion of admixture decreases, the raw wavelet variance exhibits increasing levels of noise relative to informative variation. This is shown by the reduced magnitude of the informative wavelet variance (Figure 3) and emphasizes that, as expected, it is increasingly difficult to extract informative variation at low admixture proportions (small p) even where the signal is technically present. In this example, informative estimates were obtained for admixture proportions as low as 2.5%, although in general, the range of p for which this method is applicable will also depend on other characteristics of the data, such as the SNP density and sample size, as considered in the next section.

Sensitivity analyses

The sensitivity of the method to a wide range of data characteristics was considered by repeating the results of the admixture time example with a large number of simulated data sets. Results are summarized in Table 2 and Figure 4.

Condition 1 shows the original results, exactly as described above. New simulations were then performed to mimic realistic linkage disequilibrium (LD) (condition 2). To do so as accurately as possible, we applied the real recombination rates observed along the first 100 Mb of chromosome 1, as recombination rates for chromosome 1 are near the average of all chromosome-level recombination rates (Figure S2). The effect of lower sample size (condition 3) was investigated by reducing the number of individuals sampled from each population by 5, thus yielding sample sizes that would be smaller than almost any published population genetics study (n_A, $n_B = 10$, $n_C = 15$). The effect of more recent divergence between the ancestral populations (condition 4) was investigated by decreasing $T_{\mathrm{Ancestral}}$ from 2000 to 1200 generations ago (50,000–30,000 years ago). The effect of using a misrepresentative modern population as a proxy for an ancestral population (condition 5) was investigated by studying ancestral populations with mixed (rather than "pure") ancestry. Rather than using samples from the true ancestral population P_A, an admixed ancestral population P_A^* was employed instead ($p = 0.1$). A wide range of parameters was applied for sensitivity testing, but for clarity, only results for single parameter values are shown on Figure 4. These examples are representative of all the tests that were run.

Variation in summary measures between simulations was compared by computing the relative standard deviation (RSD) at each admixture time. For all of the error conditions above, the computed ABS metrics are consistent with the reference case (condition 1), but with slightly larger relative standard deviations. Only for one case (condition 4; reduced divergence between the ancestral populations and admixture at 320 generations) are the ABS metrics biased, with the mean falling outside the range of values observed for the reference example. We emphasize that this is expected: admixture should be more difficult to detect when it occurs between two ancestral populations that diverged only recently. Stability of the ABS metrics in this particular scenario could be improved by applying a higher level of thresholding. However, the default value of $\mu = 1$ was retained here to provide consistency across scenarios, to demonstrate the deterioration in resolution, and to illustrate that the thresholding parameter can be ignored for all but the most extreme admixture cases.

In all of these examples, including the standard reference case, the localized admixture signals provide a noisy indication of how ancestry varies along the chromosome. Indeed, the inherent stochasticity of the block structure is the primary reason why other sources of variance, such as the cases discussed above, have relatively little additional effect on the overall results. This noise is addressed using wavelets to capture the distribution of block sizes, coupled with a correction based on the ancestral populations to distinguish informative signals from background variation. The cases considered above all slightly increase noise levels relative to informative variation, which, as demonstrated by the admixture proportion example in Figure 3, reduces the magnitude of the extracted informative wavelet variance. As noise increases, it naturally becomes more difficult to extract informative variation. However, this increase in noise levels is minimal for all but the most extreme confounds, thus allowing the technique to be applied robustly to a very wide range of scenarios.

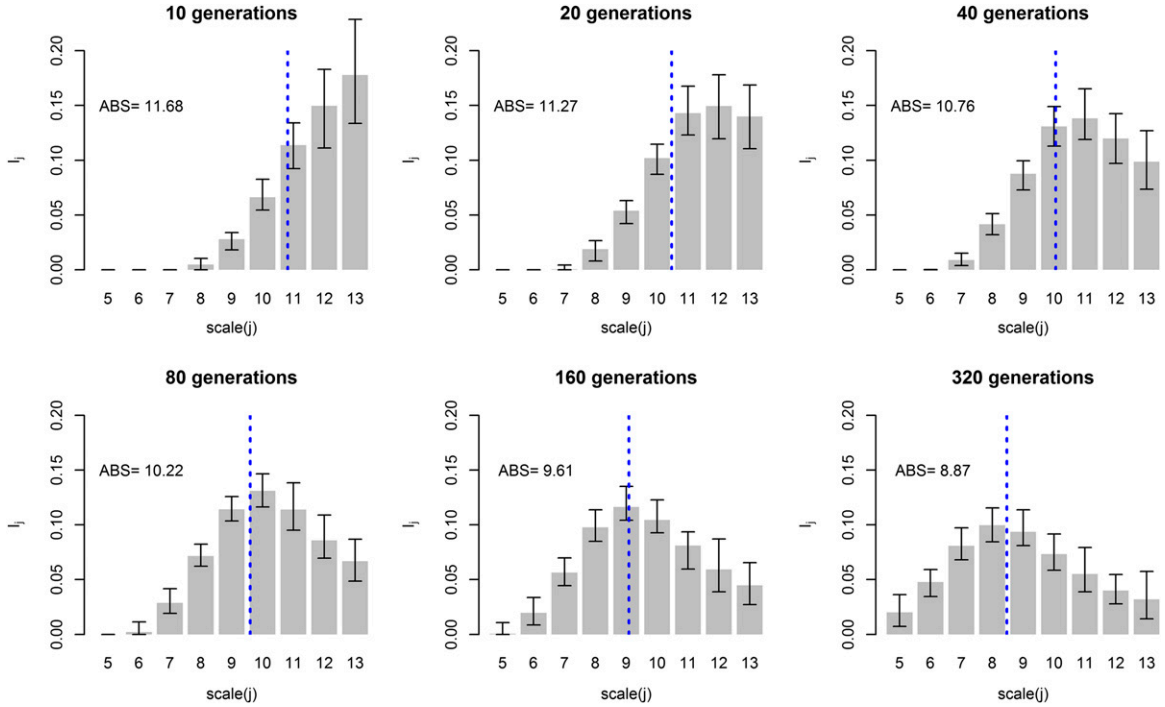


Figure 2 Informative wavelet variance for each time of admixture (10–320 generations using default thresholding $\mu = 1$). Shaded bars represent the average over 50 simulations at each admixture time; black bars represent the range across individual simulations. The average block size metric for each scenario is indicated by a dotted blue line.

The effect of SNP density (which is always a known variable) is demonstrated by down sampling the data (conditions 6–8). The original density of 4306 SNPs (condition 6) was chosen to correspond to the size of our real chromosome 22 data set. Reducing the SNP density of this data set means that the resulting wavelet decomposition is given over 11 wavelet scales rather than the earlier 13, and so as expected, the computed mean ABS metrics are correspondingly much smaller. However, this has no effect on the inference, as the data size is always known and simulations are simply run to match the size of the observed data. Further reductions in SNP density to 3250 SNPs (condition 7) and 1625 SNPs (condition 8) are also shown. Note that although the absolute values of the ABS metrics are shifted, the trend with admixture time remains consistent.

Method comparison

The original *StepPCO* method (Pugach *et al.* 2011) has already been tested extensively against other admixture detection methods, particularly *HAPMIX* (Price *et al.* 2009). We therefore focus here on comparing our improved wavelet method against the *StepPCO* procedure. Figure 5 shows that the summary measure (wavelet center) used in *StepPCO* is comparable to the *adwave* ABS metrics, as both exhibit a strong trend with time of admixture. However, the dispersion is consistently smaller for the *adwave* ABS metrics. For example, the wavelet centers (*StepPCO*) computed for $T = 320$ and $T = 160$ show substantial overlap, while the ABS metrics (*adwave*) for the same populations show only minimal overlap. This illustrates that *adwave* offers increased power to differentiate between older admixture scenarios, with substantially reduced uncertainty in dating.

We also emphasize that *adwave* requires far fewer user specifications with regard to runtime options. The only variable for *adwave* is the thresholding parameter, and as shown above, the default value of $\mu = 1$ should be used for almost all admixture scenarios. In contrast, the *StepPCO* results required a signal length parameter ($K = 1024$), a window size parameter ($\lambda = 5$), and two thresholding parameters (threshold = 0.1, maxlevel = 6) (all notations from Pugach *et al.* 2011). A detailed demonstration of this method comparison, with explanation of the settings chosen for *StepPCO*, is provided in Figure S3.

Admixture in Indonesian populations

Populations across Indonesia show genomic admixture between Asian and Melanesian ancestral sources (Cox *et al.* 2010), which has been dated using other methods to an admixture event ~4000 years ago (~130 generations) (Xu *et al.* 2012). We calculated wavelet summary measures for 16 communities across the Indonesian archipelago using 548,994 autosomal SNPs screened in 394 individuals (Table 1). Equivalent data from Southern Han Chinese and Papua New Guinea Highlanders was used as proxies for ancestral populations, as described in Cox *et al.* (2010).

The PCA for all individuals, where only the ancestral populations were used to define the axes, is shown in Figure 6. Admixed individuals dispersed along the first principal

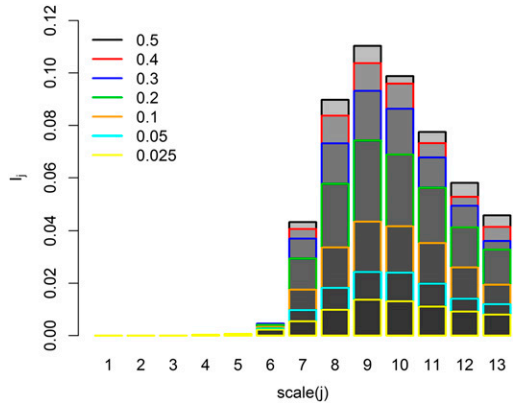


Figure 3 Relationship between proportion of admixture and informative wavelet variance. For this example only, a nondefault value for the threshold $\mu = 1.1$ was used to account for increased noise in the admixture signals due to low proportions of admixture, as described in the text. The magnitude of the wavelet variance decreases with the admixture proportion, shown as colored bars from black ($P = 0.50$) to yellow ($P = 0.025$).

component illustrate the primary genomic signal, a strong gradient in Asian-Melanesian ancestry that has previously been observed across the region (Cox *et al.* 2010). The informative wavelet variance was computed separately for each chromosome and individual and subsequently combined to provide a single measure for each population (Figure S4). To combine information across chromosomes, which vary considerably in size, the raw admixture signals were windowed: all signals were reduced to the size of the smallest chromosome (importantly without discarding any data) by computing averages over a window of SNPs (details of the windowing procedure are provided in Supporting Information, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, and File S1). The SNP density and window size for each chromosome are shown in Table S1. This windowing procedure is used only to standardize chromosomes to the same length and utilizes very short windows of SNPs (unlike the approach of Pugach *et al.* 2011).

The average block size metrics calculated for each population are shown in Table 1. The first six Indonesian populations (Nias, Mentawai, Java, Sumatra, Bali, and Sulawesi) exhibit predominantly Asian ancestry, with high-frequency noise in the signals causing some bias in the computed ABS metrics (Figure S4). The remaining Indonesian populations exhibit less extreme Asian ancestry proportions (42–67%), with the resulting ABS metrics appearing broadly similar between populations.

Under the assumption of a single admixture time (relaxed in later sections), the average block size metric can be used to date the time of admixture using approximate Bayesian computation (ABC). A general introduction to ABC can be found in Csilléry *et al.* (2010) and Sunnåker *et al.* (2013), while ABC in the context of parameter estimation for population admixture has been considered by Sousa *et al.* (2009) and Robinson *et al.* (2014).

The ABC inference procedure allows us to capture uncertainty in admixture time estimates more robustly than earlier wavelet dating approaches (Pugach *et al.* 2011; Xu *et al.* 2012). To illustrate this process, dating was performed on the Bena population of Flores in eastern Indonesia, resulting in an estimated median admixture time of 147 generations (95% credible region: 122–178 generations), or 4410 years before present (95% CR: 3660–5340 years BP). This almost exactly matches earlier point estimates of the admixture time (Xu *et al.* 2012) and is consistent with our current understanding of Island Southeast Asian prehistory (Bellwood 2007).

The relationship between time of admixture and the ABS metric across all simulations is illustrated in Figure 7A. ABC was implemented using the R package *abc* (Csilléry *et al.* 2012), and the posterior distribution of admixture time was computed using a local linear regression (Beaumont *et al.* 2002) with a tolerance rate of 0.2. Cross validation was used to evaluate the accuracy of this estimate: the prediction error was low (0.038) and insensitive to the exact tolerance value. For future research focusing on parameter inference, this procedure could be modified to use a larger number of simulated data sets and a lower tolerance rate. However, this simple example clearly illustrates that the *adwave* method has good statistical power to date admixture using a relatively small number of simulations.

Multiple admixture events

Another aim of this work is to show that our improved wavelet approach can be used to study other features of the admixture process beyond the well-explored question of admixture time. In the examples covered thus far, it has been assumed that admixture occurred as a single event. However, additional waves of admixture will result in the introduction of new ancestry tracts, replacing a proportion of older, shorter ancestry blocks with newer, longer ones. Pugach *et al.* (2011) briefly considered the effects of continuous admixture within a wavelet setting, showing that this leads to underestimated admixture times in their original methodological framework. In contrast, we instead consider scenarios with two distinct admixture events. We show that this process creates distinctive patterns in the observed informative variation, which can be used to reconstruct more complex demographic processes (as opposed to being treated solely as a potential source of bias).

In the following dual-admixture scenarios, the first admixture event always occurs at 160 generations. To investigate the effect of separation between admixture events, the second admixture event varies between 10 and 80 generations. In the extreme case of admixture at 160 and 10 generations ago, the localized admixture signals contain two dominant frequencies. Single admixture events at 160 and 10 generations lead to peaks in the informative wavelet variance at wavelet scales of 9 and 13, respectively. When two admixture events occur, the informative wavelet variance is instead spread between these scales (Figure 8A). As the admixture events occur closer

Table 2 Sensitivity of the *adwave* method to a range of data limitations

Data limitations		Admixture time (generations)					
Condition	Description	10	20	40	80	160	320
1	Reference	11.55 (0.69)	11.14 (0.58)	10.65 (0.7)	10.13 (0.66)	9.54 (1.12)	8.84 (1.72)
2	Realistic LD	11.71 (0.84)	11.31 (0.88)	10.85 (1.08)	10.32 (1.04)	9.82 (1.23)	9.09 (2.28)
3	Reduced sample size	11.66 (0.90)	11.25 (0.63)	10.74 (0.88)	10.20 (0.85)	9.58 (1.19)	8.85 (1.78)
4	Reduced divergence between ancestral populations	11.64 (0.83)	11.24 (0.69)	10.76 (0.82)	10.21 (1.02)	9.49 (1.35)	8.32 (3.59)
5	Non-representative ancestral populations	11.73 (1.34)	11.23 (1.42)	10.77 (1.51)	10.21 (1.45)	9.58 (1.94)	8.60 (3.06)
6	SNP density $T = 4036$	10.47 (0.91)	10.04 (0.73)	9.55 (0.92)	8.98 (1.22)	8.37 (1.62)	7.60 (3.13)
7	SNP density $T = 3250$	9.96 (0.76)	9.60 (0.60)	9.19 (0.76)	8.68 (1.08)	8.12 (1.41)	7.37 (2.6)
8	SNP density $T = 1625$	8.99 (1.23)	8.63 (1.09)	8.20 (1.50)	7.68 (2.29)	7.10 (3.15)	6.33 (5.86)

Mean average block size values (relative standard deviation in parentheses) are shown for each admixture time. Reference data were simulated with $T = 13{,}000$ SNPs, populations sizes of n_A, $n_B = 15$, $n_C = 15$, and divergence between the ancestral populations at $T_{Ancestral} = 2000$ generations ago.

together, this spread in the observed informative wavelet variance decreases (Figures 8, B–D).

For one admixture event, a single dominant peak is observed in the informative wavelet variance, and the ABS metric therefore provides a convenient summary measure. For multiple admixture events, the ABS metric describes the average admixture time, but provides no information about the duration over which admixture occurred. In contrast, the informative wavelet variance should provide additional information about the peak dispersion. To explore the potential for identifying more complex admixture scenarios, a simple classification rule was implemented. An admixed population P_C is assigned to one of two groups $G_1,\ G_2$, which are characterized by the summary measures $M_1,\ M_2$. This scheme is described with abstract choice of summary measure, but below, we consider how different summary measures (taking $M_1,\ M_2$ to be either the ABS metric or wavelet informative variance) affect the success of classification.

The classification rule is implemented as follows:

1. The "true" summary measures $M_1,\ M_2$ are computed for each group using values obtained from the first 25 simulations.
2. For each of the remaining 25 trial data sets ($s = 1,\ldots,25$), estimated summary statistics $\widehat{M_s}$ are calculated. The divergence measures are defined as

$$D_i = \sum_{s=1}^{S} \left| \widehat{M_s} - M_i \right|, \qquad (7)$$

for $i = 1, 2$.

3. If $D_1 < D_2$, classify to G_1; otherwise classify to G_2.

The classification rates are shown in Table 3 for scenario 1 (a single admixture event at 60 generations; mean ABS 10.47, range 10.30–10.64) and scenario 2 (two admixture events at 160 and 10 generations; mean ABS 10.57, range 10.35–10.90). With a sample size of just 10 individuals for the admixed population, perfect classification is achieved using the informative wavelet variance, while the ABS metric correctly classifies only 60% of cases. For real multiple admixture situations, this classification framework could be extended to a more complex inferential setting (such as ABC), but this simple example demonstrates the potential for reconstructing complex admixture scenarios from the full wavelet variance profile.

Discussion

Wavelet techniques provide information on the ancestry block structure of admixed genomes and hence can be used to reconstruct the processes involved in past admixture events. Ancestry blocks are strictly unobservable and can be inferred only from the data. Wavelets provide indirect information on the block structure, thus providing an alternative over methods that assign ancestry directly (Sankararaman *et al.* 2008; Price *et al.* 2009). A growing body of methods now assign ancestry indirectly using various unrelated approaches (Moorjani *et al.* 2011; Baran *et al.* 2012; Gravel 2012; Loh *et al.* 2013; Maples *et al.* 2013; Brown and Pasaniuc 2014), but here we extend the use of wavelet techniques as introduced by Pugach *et al.* (2011). Importantly, our implementation differs markedly from the original *StepPCO* program, with the main differences at each stage of the analysis highlighted below:

- Localized admixture signal formation: StepPCO (Pugach *et al.* 2011) uses large windows of SNPs to produce an averaged admixture signal in localized windows along the genome. Our work demonstrates that wavelet methods are equally applicable to the raw unwindowed signals, with the windowing procedure performed intrinsically as part of the wavelet analysis, and therefore not requiring arbitrary *a priori* decisions on window size.
- Wavelet analysis: The wavelet methods we describe are based on the MODWT, which offers more flexibility in its application since there is no restriction on the length of the signals. Conversely, StepPCO employs the discrete wavelet transform (DWT), which has the strict requirement that

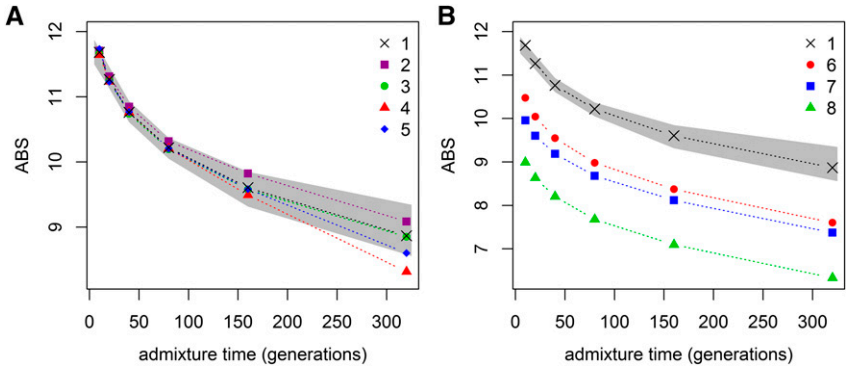


Figure 4 Sensitivity to a range of realistic data limitations. Comparison to reference data (condition 1) simulated with T = 13,000 SNPs, populations sizes $n_A, n_B = 15$, $n_C = 15$, and ancestral population divergence at $T_{Ancestral} = 2000$ generations. The gray area shows the range of ABS metrics observed under the standard reference condition. (A) Potential sources of error (conditions 2–5); (B) varying SNP densities (conditions 6–8). Note that the decline in absolute values of the ABS metrics in B is expected; these are easily accounted for in an inference setting because the SNP density is always a known variable. Condition descriptions and numeric values are presented in Table 2.

signals be of length $2n$. Data must therefore be windowed, or discarded, to meet the restrictive length requirements of the DWT framework. Another advantage of the MODWT is that the resulting wavelet coefficients are translation equivariant, meaning that circularly shifting the data results in the same shifting of the coefficients. Said differently, changing the starting point—for instance, to avoid a poor quality SNP—does not affect the resulting wavelet coefficients, whereas this is not true under the DWT framework. This property is particularly important if the results are to be used for specific localized genomic regions (as discussed briefly below) and thus provides a solid statistical foundation for future work.

- Extraction of relevant information: The portion of the resulting wavelet decomposition that is informative about the admixture process is extracted in a simple procedure with reference to the ancestral populations, offering greater simplicity and objectivity than the multistage thresholding procedure described by Pugach *et al.* (2011).
- Software: The adwave software, which implements the method described in this article, is an official package in the R project (http://cran.r-project.org/web/packages/adwave/index.html). This allows extremely easy installation and use, as well as providing a series of simple worked examples as a learning exercise. The adwave package is also faster than the existing StepPCO code and offers more flexibility in the choice of analyzing wavelet (unlike StepPCO, which employs only the simplest "square-shaped" Haar wavelet).

The work presented here also makes several other advances. The average block size metric has previously been shown to capture the time of admixture. Here, we have implemented a more formal dating procedure using ABC under the assumption of a single admixture event. In reality, populations may have experienced multiple admixture events leading to complex patterns of genetic variation. We have shown that the wavelet variance contains additional information to identify these more complex admixture scenarios. This highlights the potential of wavelet-based techniques to be coupled with formal statistical inference procedures to robustly distinguish between the range of scenarios that could have resulted in any observed genetic pattern.

Method performance for the *StepPCO* procedure has already been tested against other admixture detection methods, most extensively with *HAPMIX* (Price *et al.* 2009), with favorable results. This is especially true for older admixture events (Pugach *et al.* 2011). While an in-depth comparison with other local ancestry detection methods would be of great interest (Moorjani *et al.* 2011; Baran *et al.* 2012; Gravel 2012; Loh *et al.* 2013; Maples *et al.* 2013; Brown and Pasaniuc 2014), such an analysis is beyond the scope of this manuscript. We have therefore focused instead on showing how *adwave* markedly improves on the original wavelet method implemented in *StepPCO*. As shown above, *adwave* offers improved statistical power to differentiate between admixture scenarios, offers much reduced uncertainty in model parameter estimates, and importantly, is far easier to use than *StepPCO*, especially by requiring far fewer user-specified runtime parameters.

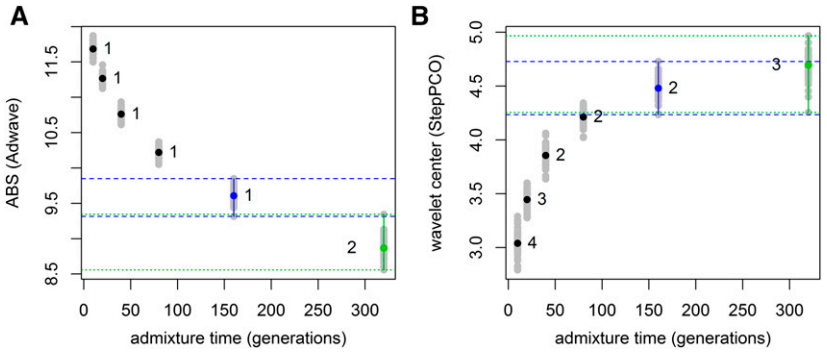


Figure 5 Comparing *StepPCO* and *adwave* showing the relationship between wavelet transform summaries and time of admixture. (A) *Adwave* using $\mu = 1$; (B) *StepPCO* using $K = 1024$, $\lambda = 5$, threshold = 0.1, and maxlevel = 6. Numbers indicate the relative standard deviation (RSD, %) for each admixture time. Note the difference in discrimination power between the two methods for older admixture events (95% confidence intervals as dashed blue and green horizontal lines).

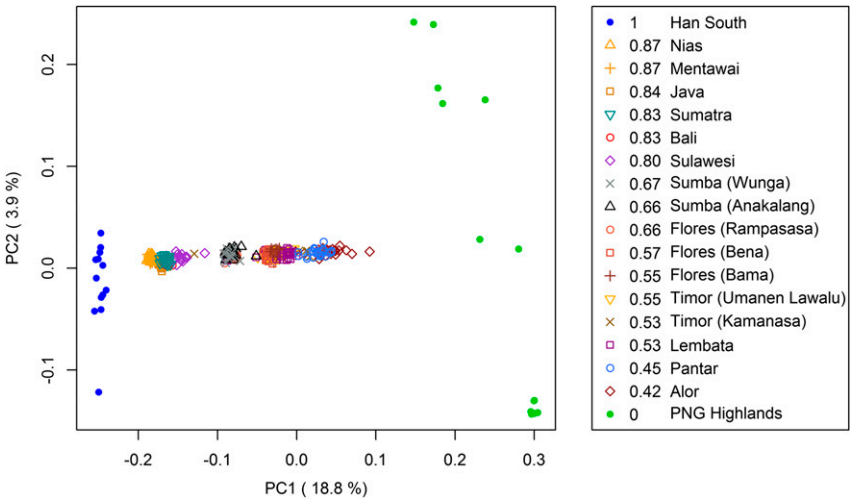


Figure 6 PCA of autosomal SNP data from Indonesian populations, with Southern Han Chinese (blue circles) and Papua New Guinea Highlanders (green circles) employed as proxy ancestral populations. Numbers give calculated admixture proportions.

In the future, considering the full wavelet periodogram, rather than the genome-wide summary measures used in both *adwave* and *StepPCO*, may yield promising results wherever the distribution of ancestry tracts along the genome is substantially nonstationary. Bryc *et al.* (2010) use their formulation of localized admixture signals to address whether regions of the genome show predominant ancestry from a given population. Wavelets are well suited to distinguishing local features in data and could be helpful in this regard, identifying features that may not be easily detected by considering the localized admixture signals in their raw form.

Other prospective areas for further work include the extension of these methods to the more general case of multipopulation admixture. Ma and Amos (2012) describe the use of PCA as a diagnostic in this setting, and PCA has been used to assign multipopulation ancestry in the software *PCAdmix* (Brisbin *et al.* 2012). The wavelet methods described here could be extended in a similar way by considering pairwise combinations of any number of ancestral populations.

In contrast, key restrictions that determine our ability to reconstruct admixture events include the degree of differentiation between the ancestral populations and the representativeness of samples used as surrogate ancestral groups. As the ancestral populations become more similar or the surrogate populations become more different from the true ancestral populations, the localized admixture signals become increasingly noisy. Although this ultimately leads to a loss of identifiability in extreme cases, the method is remarkably robust to moderate deviations from these assumptions. As shown above for low admixture proportions, through judicial choice of the thresholding parameter even extremely noisy data can still provide meaningful estimates (the only situation in which we encourage deviation from the default setting).

Sample size (both in terms of SNPs and individuals) is also important and affects the PCA step of the procedure. The purpose of the PCA step is to summarize the overall variability among individuals, which includes both between-population and within-population variability. In reconstructing population ancestry, we aim to describe between-population variation, while ignoring within-population variation. This is achieved by selecting the first principal component, as long as the sample sizes are sufficiently large. Within-population fluctuations of individual coordinates on the PCA scatterplot can be caused by subtle population substructure. Assuming that no such substructure is present, these fluctuations decrease as the total sample size increases, and an asymptotically stable pattern of the eigenvector plot results (Ma and

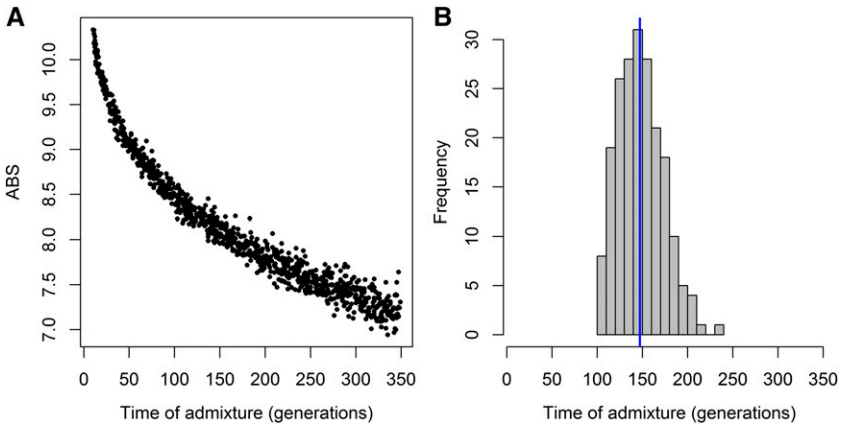


Figure 7 Dating time of admixture for Bena (Flores, eastern Indonesia) using approximate Bayesian computation. (A) Relationship between admixture time and average block size metric for all simulations; (B) weighted posterior distribution of admixture time. Median estimated time of admixture, indicated by the blue line, is 147 generations (95% credible region: 122–178 generations).

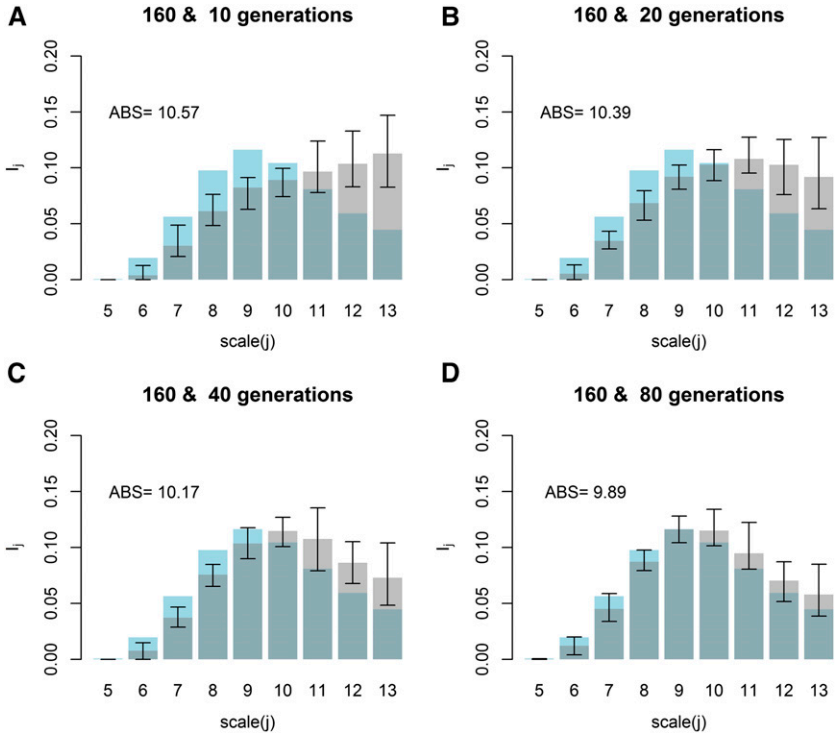


Figure 8 Dual admixture events at 160 and 10–80 generations. Gray bars represent the average over 50 simulations for each scenario; black bars represent the range for individual simulations. Blue bars show the average informative wavelet variance for a single admixture event at 160 generations, providing a reference point for comparison.

Amos 2010). When the number of individuals is large, variation between individuals from the same population is small compared to that of the different populations, so that the first eigenvector describes the primary population structure of the data. However, as the sample size decreases, individual variation carries more weight, which may be addressed in more than the first principal component. Note that the use of methods other than PCA may be helpful in this regard. For example, Jombart *et al.* (2010) introduced discriminant analysis of principal components to achieve separation of individuals into predefined groups. In practice, as long as the sizes of the ancestral population samples is sufficiently large, discriminant analysis provides the same result as PCA (unpublished data). The two methods may, however, perform differently for small sample sizes.

How far back in time admixture processes can be reliably identified is strongly influenced by the number of genotyped SNPs. The relationship between the number of admixture blocks, time of admixture and wavelet scale is summarized in Table 4. The shaded column indicates the findings described in the *Results*, using simulated data sets of 13,000 SNPs (chosen for a region ~100 Mb in length, comparable to the SNP content of our 100 Mb chromosome 15 data set). For admixture at 10 generations, the informative wavelet variance is highest at scale 13, reflecting a small number of large admixture blocks. As the time of admixture increases, the peak shifts toward lower scales, reflecting a larger number of smaller admixture blocks. This pattern is illustrated for admixture up to 320 generations (~10,000 years), but importantly, it is possible to reconstruct even older admixture events. The highest frequency (relating to the smallest admixture blocks) that can be detected, as determined purely by the data density, is termed the Nyquist frequency (Chatfield 2003). However, resolution power is likely to deteriorate well before this point and will be strongly influenced by the degree of differentiation between the ancestral populations. The more closely related the ancestral populations, the less well they can be discriminated using only a small number of SNPs. Increasing the SNP density allows detection of higher frequency information, relating to shorter (more ancient) admixture tracts. To illustrate this, the mapping to wavelet scale is illustrated for a hypothetical twofold and fourfold increase in the number of genotyped SNPs (26,000 and 104,000 SNPs, respectively). As genetic data sets improve (particularly through whole-genome sequencing), wavelet methods will therefore

Table 3 Classification rate for the summary measures average block size and informative wavelet variance with increasing sample size ($1 \leq n_C \leq 10$)

	Correct classification (%)	
Sample size (individuals)	Wavelet variance	Average block size
1	76	56
2	84	58
3	89	59
4	92	60
5	94	61
6	96	61
7	97	62
8	98	62
9	99	61
10	100	60

Table 4 Relationship between the number of admixture blocks, time of admixture, and wavelet scale

		Wavelet scale (no. of SNPs)		
Admixture blocks	Time of admixture (generations)	13,000	26,000	104,000
8,192–16,384	163,840	—	—	1
4,096–8,192	81,920	—	1	2
2,048–4,096	40,960	1	2	3
1,024–2,048	20,480	2	3	4
512–1,024	10,240	3	4	5
256–512	5,120	4	5	6
128–256	2,560	5	6	7
64–128	1,280	6	7	8
32–64	640	7	8	9
16–32	320	8	9	10
8–16	160	9	10	11
4–8	80	10	11	12
2–4	40	11	12	13
1–2	20	12	13	14
0–1	10	13	14	15

The dominant admixture block size decreases with time since admixture, while conversely, the number of admixture blocks increases. Underlined numbers are from the example presented in the *Results* section (13,000 SNPs from a genomic region ~100 Mb in length, comparable to the data set for chromosome 15). Columns to the right show how mapping to wavelet scale depends heavily on SNP density: increasing the number of SNPs two- and fourfold allows higher frequency information to be detected, which in turn informs about shorter (more ancient) admixture tracts.

gain substantial resolution. It seems entirely feasible that wavelet approaches will have sufficient statistical power to reconstruct admixture events far deeper in time than those currently studied. Advances in wavelet methods therefore offer exciting potential for future research, particularly for ancient and complex human admixture processes.

Software

Software for the analyses described here has been released in the form of an R package, *adwave*, which is freely available from the R project's central package repository: http://cran.r-project.org/web/packages/adwave/index.html

Acknowledgments

We gratefully acknowledge assistance with sample collection by Agustini Leonita and Alida Harahap (Eijkman Institute for Molecular Biology, Jakarta, Indonesia) and J. Stephen Lansing (University of Arizona), as well as data processing by Olga Savina (University of Arizona). We also thank Matthew Nunes (University of Lancaster, United Kingdom) and Martin Hazelton (Massey University, New Zealand) for their valuable comments.This research was supported by the Royal Society of New Zealand through a Rutherford Fellowship (RDF-10-MAU-001) and Marsden Grant (11-MAU-007) to M.P.C., and by funding from the Allan Wilson Center for Molecular Biology and Evolution.

Literature Cited

Baran, Y., B. Pasaniuc, S. Sankararaman, D. G. Torgerson, C. Gignoux *et al.*, 2012 Fast and accurate inference of local ancestry in Latino populations. Bioinformatics 28: 1359–1367.

Beaumont, M. A., W. Zhang, and D. J. Balding, 2002 Approximate Bayesian Computation in population genetics. Genetics 162: 2025–2035.

Bellwood, P., 2007 *Prehistory of the Indo-Malaysian Archipelago*. ANU E Press, Canberra, Australia.

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.

Brown, R., and B. Pasaniuc, 2014 Enhanced methods for local ancestry assignment in sequenced admixed individuals. PLOS Comput. Biol. 10: e1003555.

Bryc, K., A. Auton, M. R. Nelson, J. R. Oksenberg, S. L. Hauser *et al.*, 2010 Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. USA 107: 786–791.

Chatfield, C., 2003 *The Analysis of Time Series: An Introduction*, 6th Ed. Chapman & Hall/CRC, Boca Raton, FL.

Chen, G. K., P. Marjoram, and J. D. Wall, 2009 Fast and flexible simulation of DNA sequence data. Genome Res. 19: 136–142.

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 Genet. 9: 76.

Csilléry, K., M. G. B. Blum, O. E. Gaggiotti, and O. François, 2010 Approximate Bayesian Computation (ABC) in practice. Trends Ecol. Evol. 25: 410–418.

Csilléry, K., O. François, and M. G. B. Blum, 2012 abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. 3: 475–479.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Fenner, J. N., 2005 Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. 128: 415–423.

Gravel, S., 2012 Population genetics models of local ancestry. Genetics 191: 607–619.

Hellenthal, G., G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli *et al.*, 2014 A genetic atlas of human admixture history. Science 343: 747–751.

Jombart, T., S. Devillard, and F. Balloux, 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11: 94.
Liò, P., 2003 Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics 19: 2–9.
Loh, P.-R., M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell *et al.*, 2013 Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193: 1233–1254.
Ma, J., and C. I. Amos, 2010 Theoretical formulation of principal components analysis to detect and correct for population stratification. PLoS ONE 5: e12510.
Ma, J., and C. I. Amos, 2012 Principal components analysis of population admixture. PLoS ONE 7: e40115.
Maples, B. K., S. Gravel, E. E. Kenny, and C. D. Bustamante, 2013 RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. 93: 278–288.
McVean, G., 2009 A genealogical interpretation of principal components analysis. PLoS Genet. 5: e1000686.
Moorjani, P., N. Patterson, J. N. Hirschhorn, A. Keinan, L. Hao *et al.*, 2011 The history of African gene flow into southern Europeans, Levantines, and Jews. PLoS Genet. 7: e1001373.
Novembre, J., and M. Stephens, 2008 Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. 40: 646–649.
Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. 2: e190.
Percival, D. P., 1995 On estimation of the wavelet variance. Biometrika 82: 619–631.
Pool, J. E., and R. Nielsen, 2009 Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181: 711–719.
Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5: e1000519.
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 Biol. 12: R19.
Purcell, S., 2009 PLINK, http://pngu.mgh.harvard.edu/purcell/plink/.
Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.
R Development Core Team, 2014 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Robinson, J. D., L. Bunnefeld, J. Hearn, G. N. Stone, and M. J. Hickerson, 2014 ABC inference of multi-population divergence with admixture from unphased population genomic data. Mol. Ecol. 23: 4458–4471.
Sankararaman, S., G. Kimmel, E. Halperin, and M. I. Jordan, 2008 On the inference of ancestries in admixed populations. Genome Res. 18: 668–675.
Sousa, V. C., M. Fritz, M. A. Beaumont, and L. Chikhi, 2009 Approximate Bayesian Computation without summary statistics: the case of admixture. Genetics 181: 1507–1519.
Sunnåker, M., A. G. Busetto, E. Numminen, J. Corander, M. Foll *et al.*, 2013 Approximate Bayesian computation. PLOS Comput. Biol. 9: e1002803.
Whitcher, B., 2013 waveslim: Basic wavelet routines for one-, two- and three-dimensional signal processing. http://cran.r-project.org/web/packages/waveslim/index.html.
Xu, S., I. Pugach, M. Stoneking, M. Kayser, L. Jin *et al.*, 2012 Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. Proc. Natl. Acad. Sci. USA 109: 4574–4579.

Communicating editor: K. M. Roeder

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176842/-/DC1

Reconstructing Past Admixture Processes from Local Genomic Ancestry Using Wavelet Transformation

Jean Sanderson, Herawati Sudoyo, Tatiana M. Karafet, Michael F. Hammer, and Murray P. Cox

DOI: 10.1534/genetics.115.176842

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.

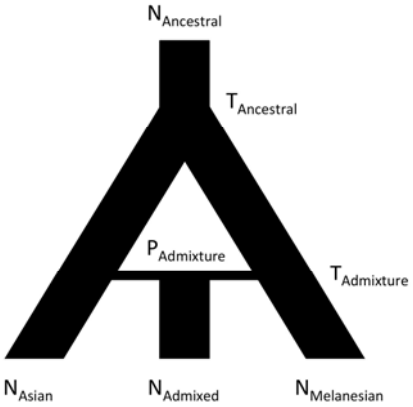
Cox, M. P., A. E. Woerner, J. D. Wall, and M. F. Hammer, 2008 Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. BMC Genetics 9: 76.

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.



Parameter	Value	Reason
$N_{Ancestral}$	10,500	Average N_A from Cox *et al.* (2008)
N_{Asian}	2,050	Average N_e for Han Chinese from Cox *et al.* (2008)
$N_{Melanesian}$	800	Average N_e for Melanesians from Cox *et al.* (2008)
$N_{Admixed}$	1,425	Average of N_{Asian} and $N_{Melanesian}$
$T_{Admixture}$	160 gen	~4,000 years ago; starting value, varied in simulations
$T_{Ancestral}$	2,000 gen	~50,000 years ago from Cox *et al.* (2008)
$P_{Admixure}$	0.5	Starting value; varied in simulations between (0,1)

Figure S1 Demographic model and parameters.

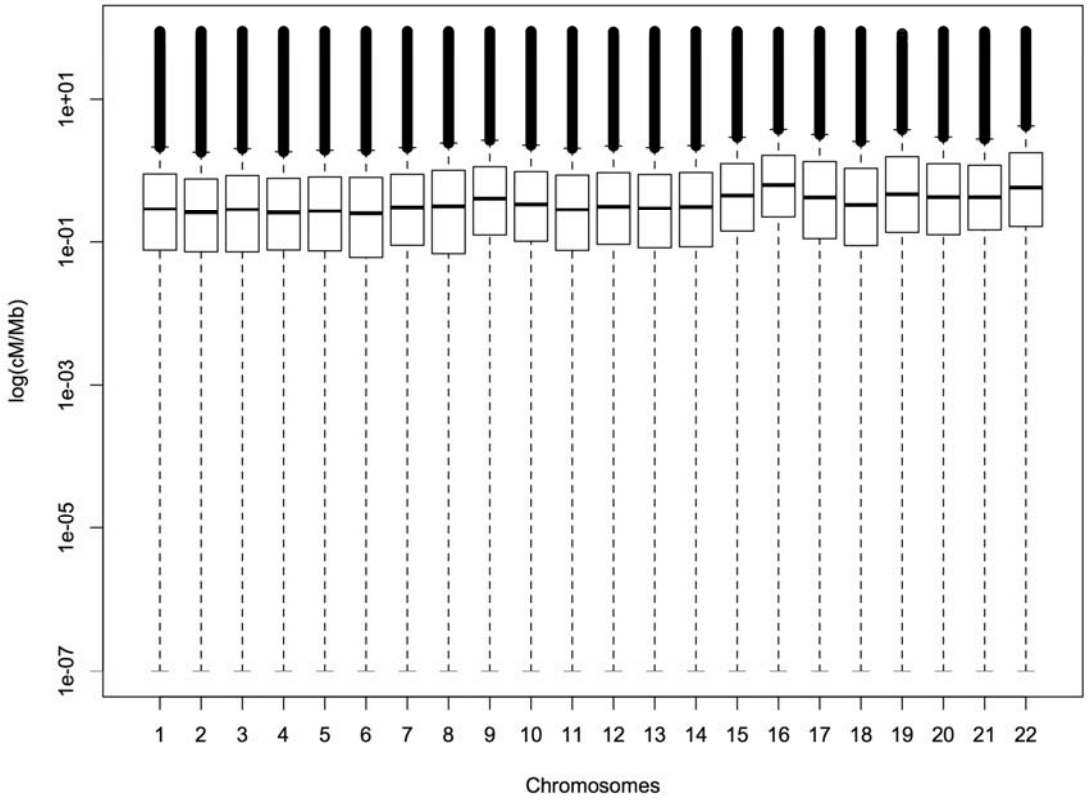


Figure S2 Distribution of recombination rates (cM/Mb) across the 22 human autosomal chromosomes. Note that chromosome 1 is representative of the other chromosomes.

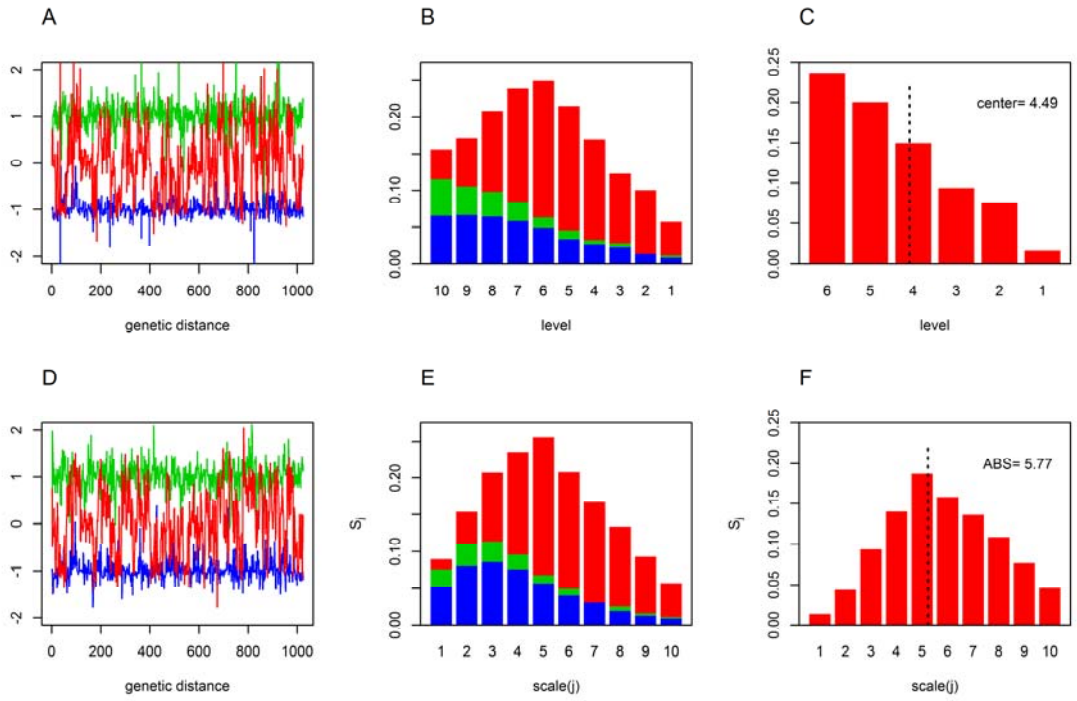


Figure S3 Method demonstration for (top) *StepPCO* and (bottom) *adwave*.

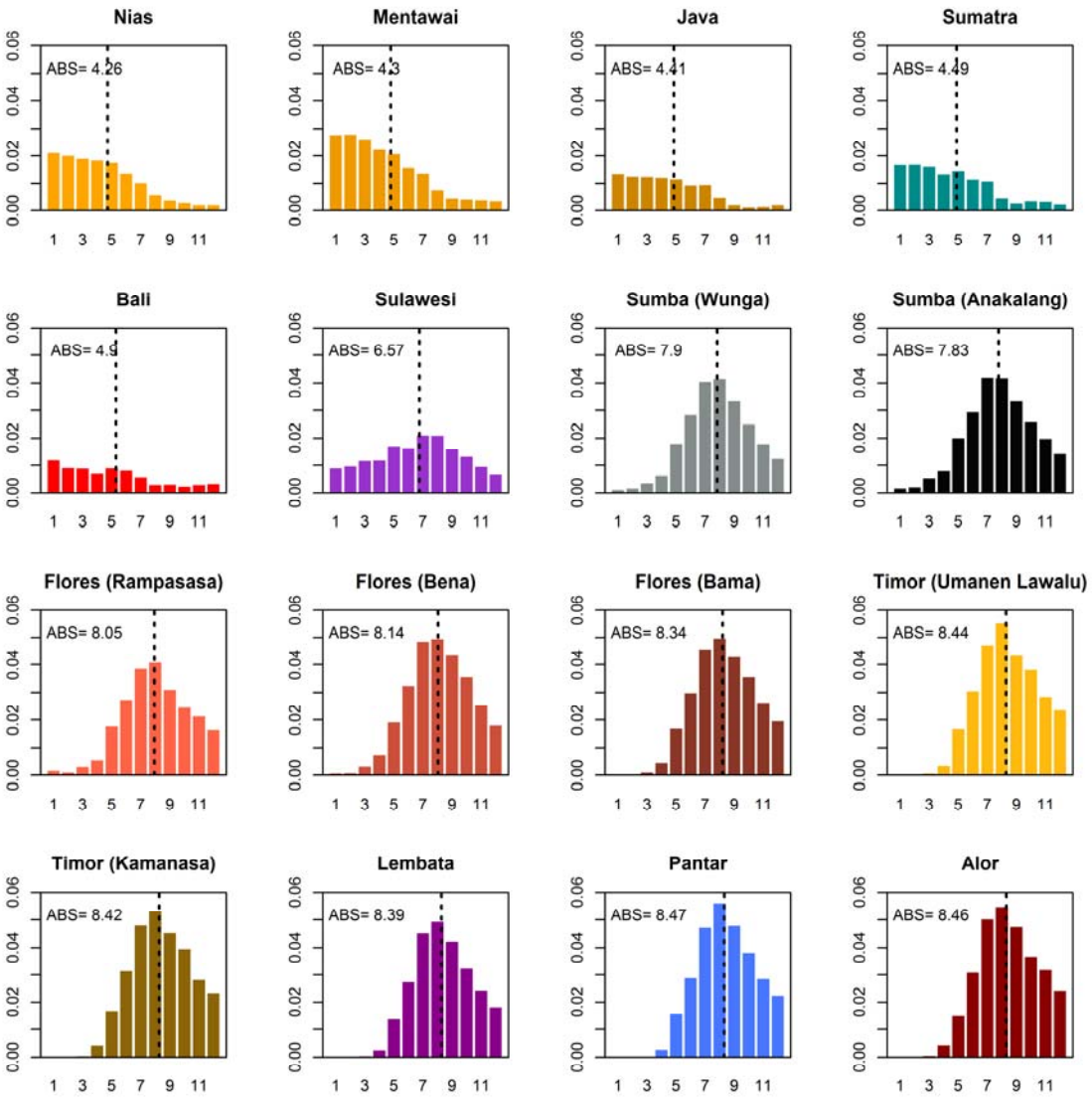


Figure S4 Informative wavelet variance for 16 Indonesian populations. The average block size (ABS) metric is indicated by a vertical line.

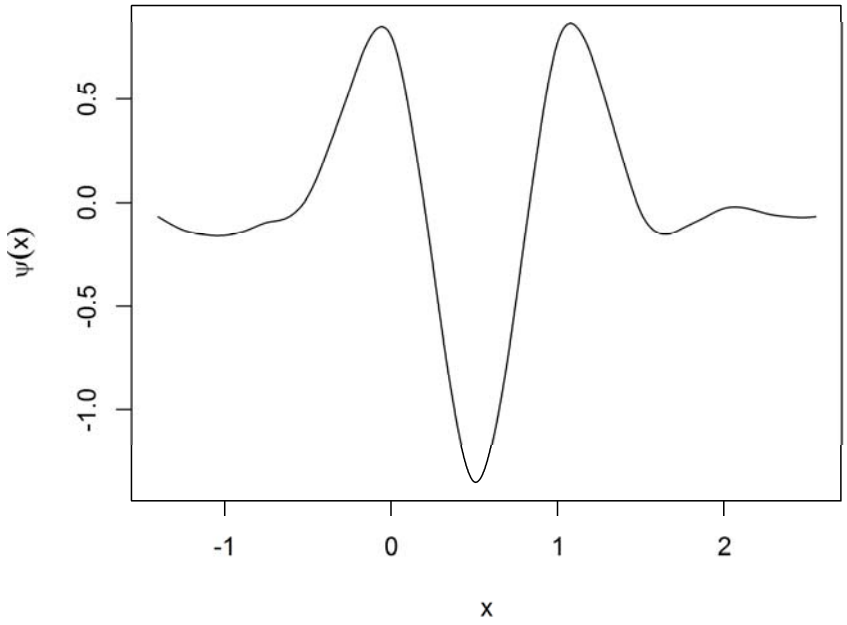


Figure S5 Daubechies' Least Asymmetric wavelet with eight vanishing moments.

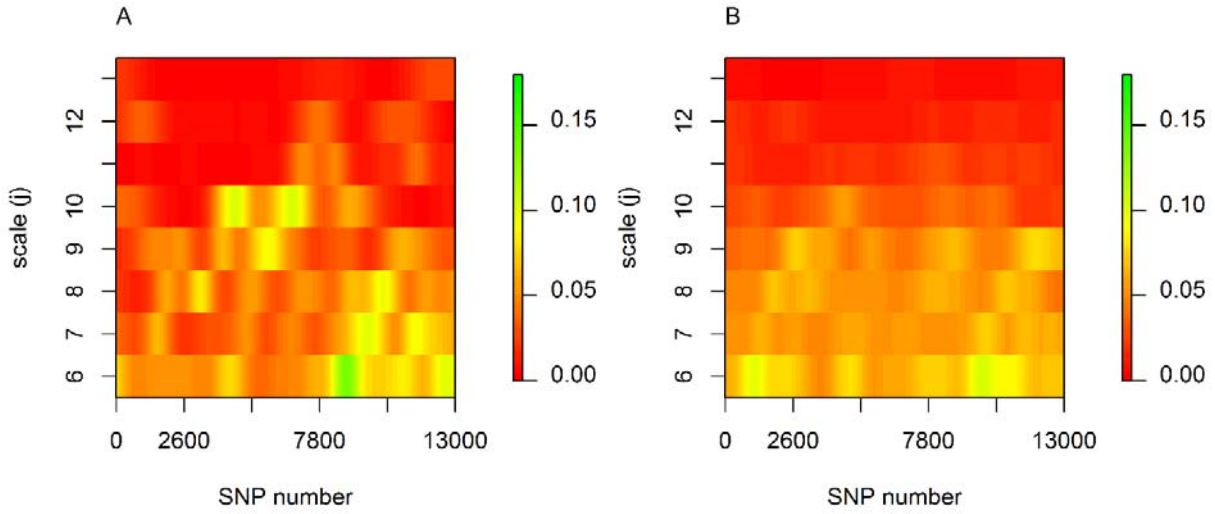


Figure S6 Wavelet periodograms A) for one individual and B) for the population average. Periodograms have been smoothed using a simple density smoother.

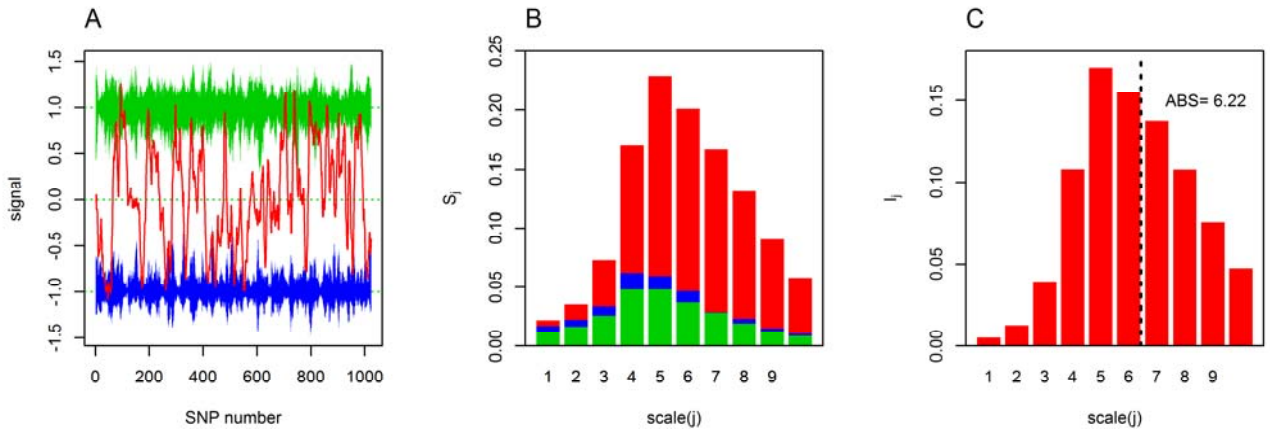


Figure S7 Visualization of ancestral block structure by windowing the admixture signal, using the same data as in **Figure 1**, but pre-windowing signals with W = 130. A) Admixture signal for one individual; B) raw wavelet variance for each population showing less high frequency noise than the non-windowed version in **Figure 1B**; C) informative variation after correcting for noise estimated from the ancestral populations.

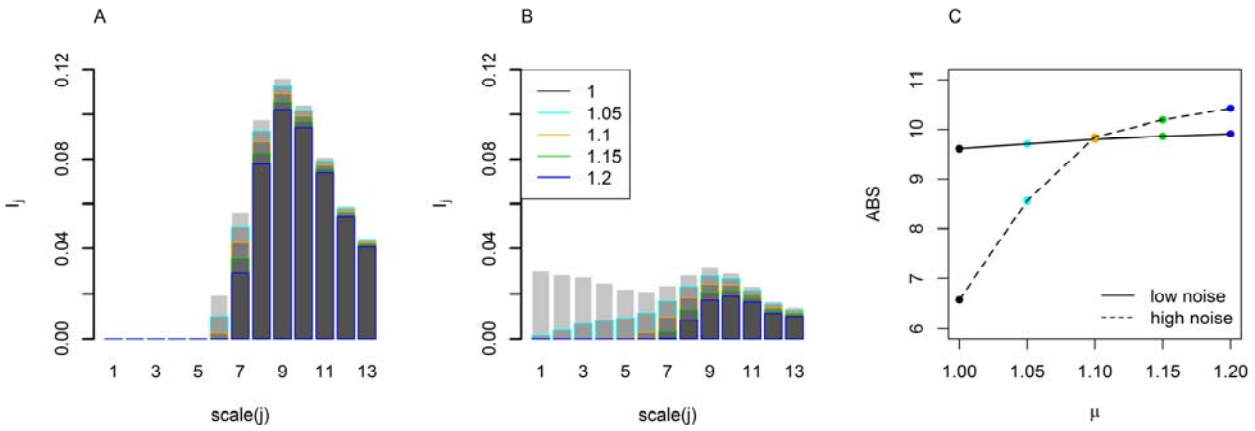


Figure S8 Relationship between choice of thresholding parameter $\boldsymbol{\mu}$ and the resulting informative wavelet variance and ABS metrics for two simulated examples. A) For admixture signals with low noise levels ($\boldsymbol{p = 0.5}$), increasing $\boldsymbol{\mu}$ results in a decrease in the magnitude of the extracted informative wavelet variance, but the location of the peak remains unchanged. B) For admixture signals with high noise levels ($\boldsymbol{p = 0.05}$), increasing $\boldsymbol{\mu}$ successfully eliminates the noise observed at low scales, while maintaining the peak in the informative wavelet variance that is attributed to the admixture process. C) The resulting ABS metrics for both the low and high noise examples. For low levels of noise, the ABS metrics are robust to choice of $\boldsymbol{\mu}$, while for high levels of noise, a larger value is necessary to avoid bias.

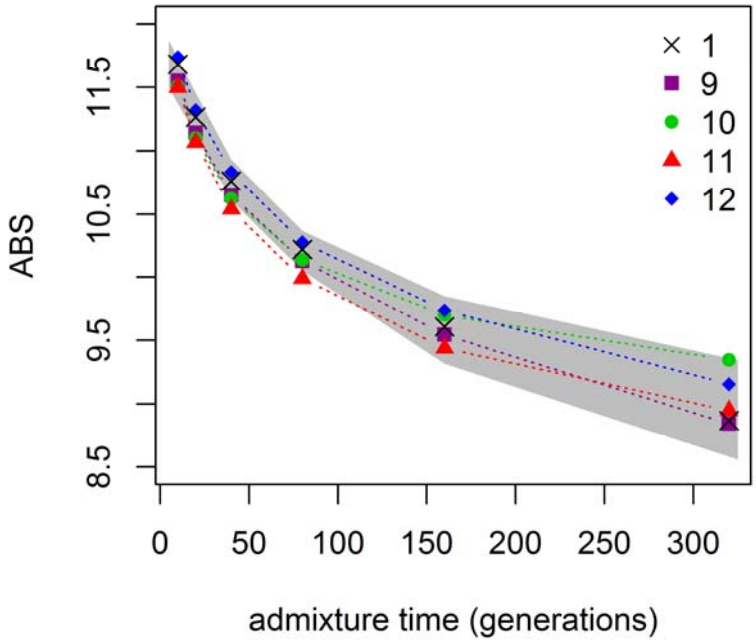


Figure S9 Sensitivity to different method options. The grey area reflects the range of ABS values observed under the default method options (condition 1). Condition descriptions and numeric values are presented in **Table S2**.

Table S1 Chromosome level information. The reported number of SNPs for each chromosome reflects SNPs that are not fixed in both of the ancestral reference populations; percentage missing data due to failed genotyping; size of analyzing window used to combine information across chromosomes.

Chromosome	Number of SNPs	Missing data (%)	Analyzing window
1	32,417	0.75	7.52
2	35,221	0.69	8.17
3	32,005	0.62	7.43
4	29,117	0.65	6.76
5	28,179	0.70	6.54
6	33,862	0.72	7.86
7	24,676	0.73	5.73
8	25,025	0.67	5.81
9	20,914	0.65	4.85
10	21,755	0.76	5.05
11	20,494	0.77	4.76
12	21,398	0.74	4.97
13	17,564	0.63	4.08
14	14,552	0.77	3.38
15	13,856	0.75	3.22
16	13,402	0.81	3.11
17	9,348	0.90	2.17
18	14,246	0.74	3.31
19	5,733	1.12	1.33
20	10,553	0.87	2.45
21	6,403	0.74	1.49
22	4,309	1.16	1.00

Table S2 Sensitivity to different method options. The mean ABS (relative standard deviation in parentheses) is given for each admixture time.

Method option		Admixture time (generations)					
Condition	**Description**	**10**	**20**	**40**	**80**	**160**	**320**
1	Reference	11.68(0.69)	11.27 (0.58)	10.76(0.70)	10.22(0.66)	9.61(1.12)	8.82 (1.72)
9	Haar wavelet	11.55 (0.65)	11.14 (0.56)	10.65 (0.70)	10.13 (0.68)	9.54 (1.08)	8.84 (1.61)
10	Pre-windowing signal (130 SNPs)	11.51 (0.73)	11.10 (0.56)	10.62 (0.58)	10.14 (0.57)	9.70 (0.79)	9.34 (0.99)
11	Pre-windowing signal (65 SNPs)	11.50 (0.78)	11.07(0.60)	10.54 (0.61)	9.99 (0.64)	9.44 (0.85)	8.95(1.10)
12	Genetic distance	11.74 (0.66)	11.32 (0.58)	10.83(0.64)	10.28(0.65)	9.74 (0.98)	9.16 (1.33)