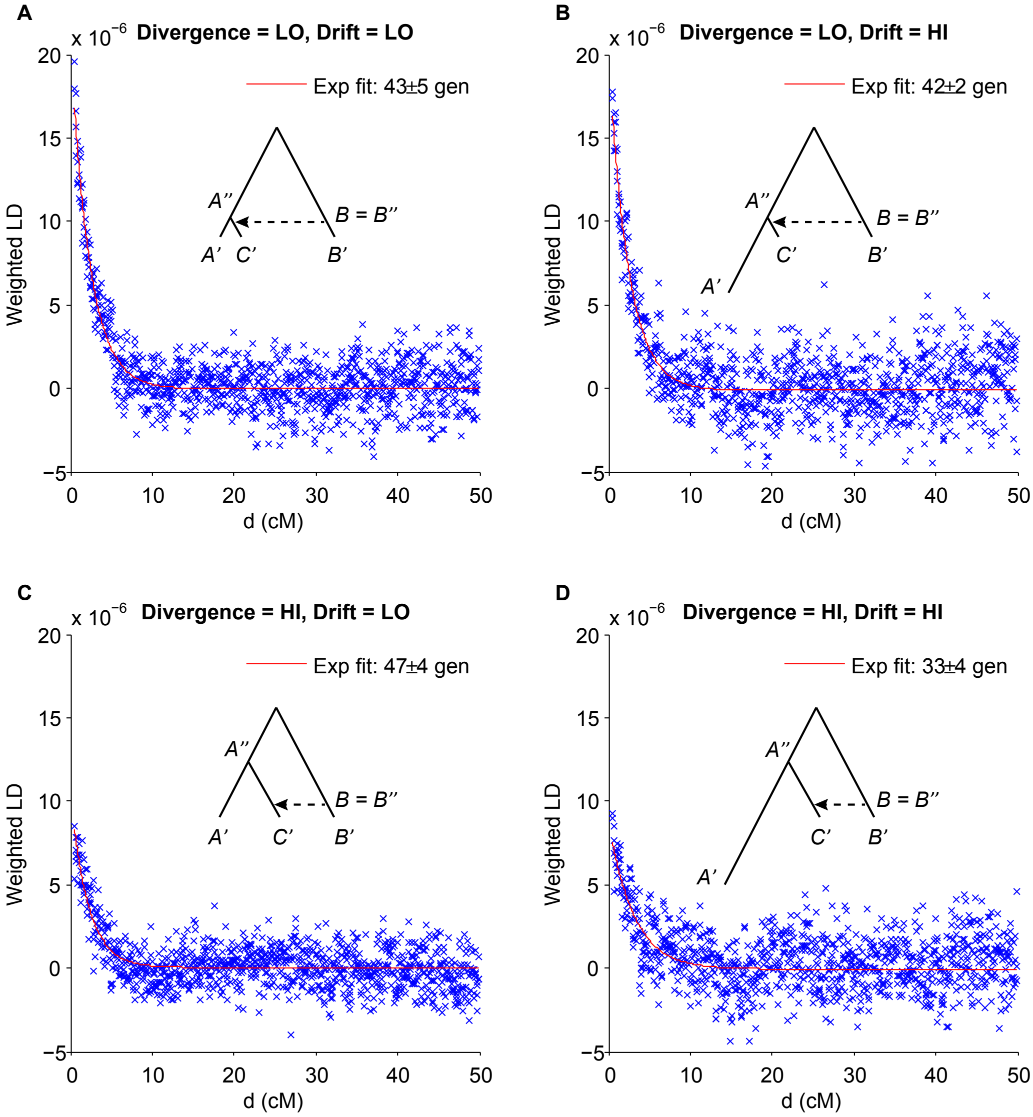

Figure S1. Weighted LD curves from four coalescent simulations of admixture scenarios with varying divergence times and drift between the reference population A' and the true mixing population. In each case, gene flow occurred 40 generations ago. In the low-divergence scenarios, the split point A'' is immediately prior to gene flow, while in the high-divergence scenarios, A'' is halfway up the tree (520 generations ago). The high-drift scenarios are distinguished from the low-drift scenarios by a 20-fold reduction in population size for the past 40 generations. Standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation.

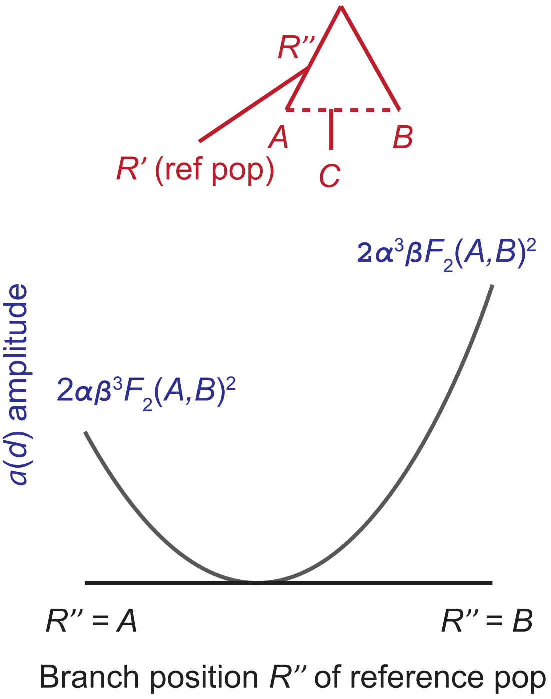
R''
A
B
C
R' (ref pop)
$2\alpha^3\beta F_2(A,B)^2$
$2\alpha\beta^3 F_2(A,B)^2$
$a(d)$ amplitude
R'' = A
R'' = B
Branch position R'' of reference pop

Figure S2. Dependence of single-reference weighted LD amplitude on the reference population. When taking weights as allele frequency differences between the admixed population and a single reference population R', the weighted LD curve $a(d)$ has expected amplitude proportional to $(\alpha F_2(A, R'') - \beta F_2(B, R''))^2$, where R'' is the point along the A--B lineage at which the reference population branches. Note in particular that as R'' varies from A to B, the amplitude traces out a parabola that starts at $2\alpha\beta^3 F_2(A, B)^2$, decreases to a minimum value of 0, and increases to $2\alpha^3\beta F_2(A, B)^2$.

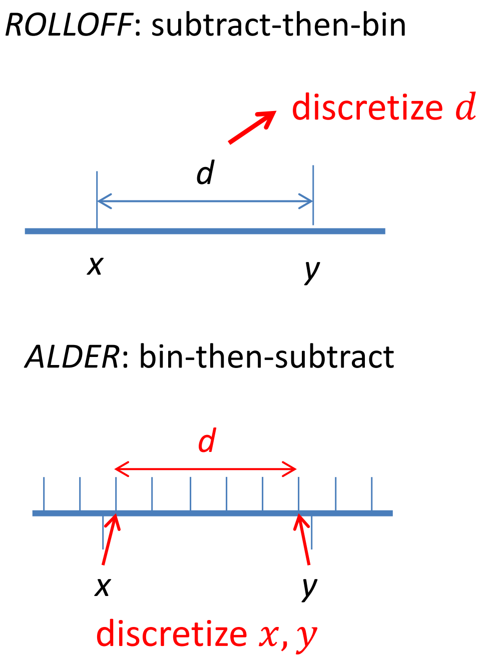

Figure S3. Comparison of binning procedures used by *ROLLOFF* and *ALDER*. Instead of discretizing inter-SNP distances, *ALDER* discretizes the genetic map before subtracting SNP coordinates.

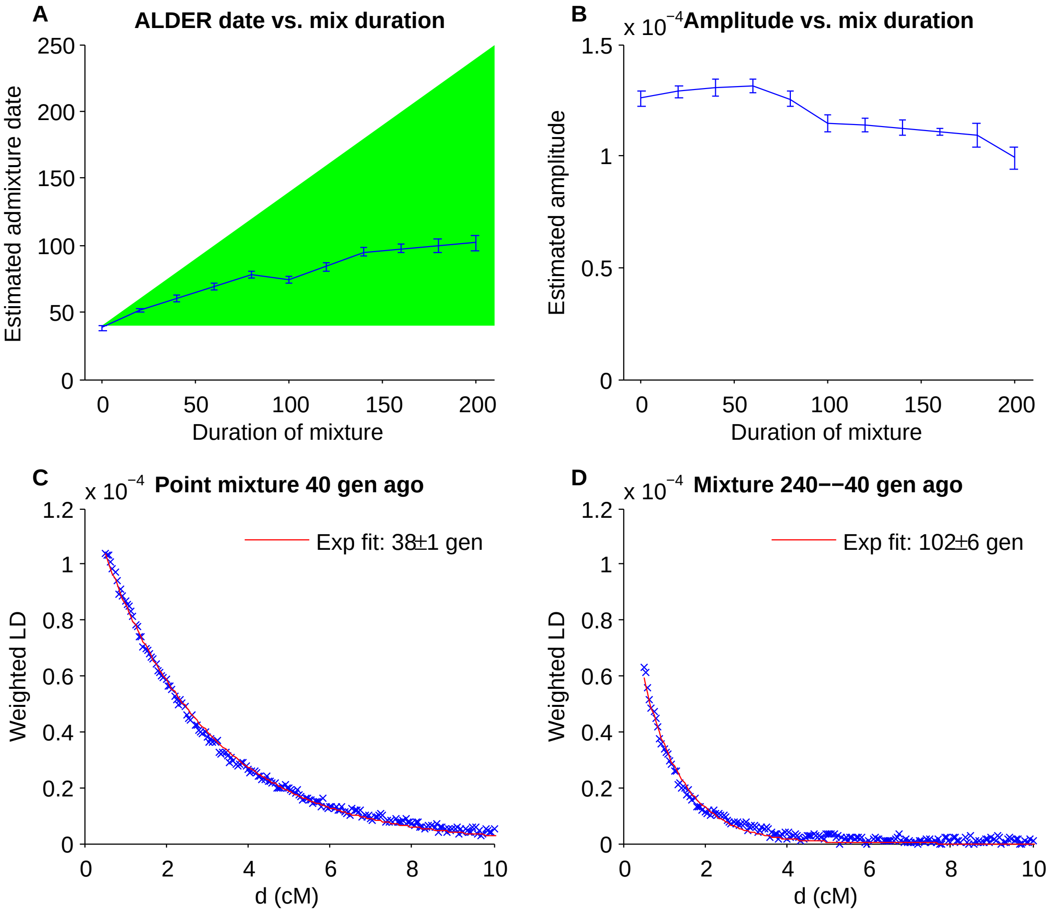

Figure S4. Weighted LD curve parameters from coalescent simulations of continuous admixture. In each simulation the mixed population receives 40% of its ancestry through continuous gene flow over a period of 0--200 generations ending 40 generations ago. Panels (A) and (B) show the admixture dates and weighted LD amplitudes computed by *ALDER* for each of 11 simulations (varying the duration of mixture from 0 to 200 in increments of 20). Panels (C) and (D) show the curves and exponential fits for mixture durations at the two extremes. Standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation.

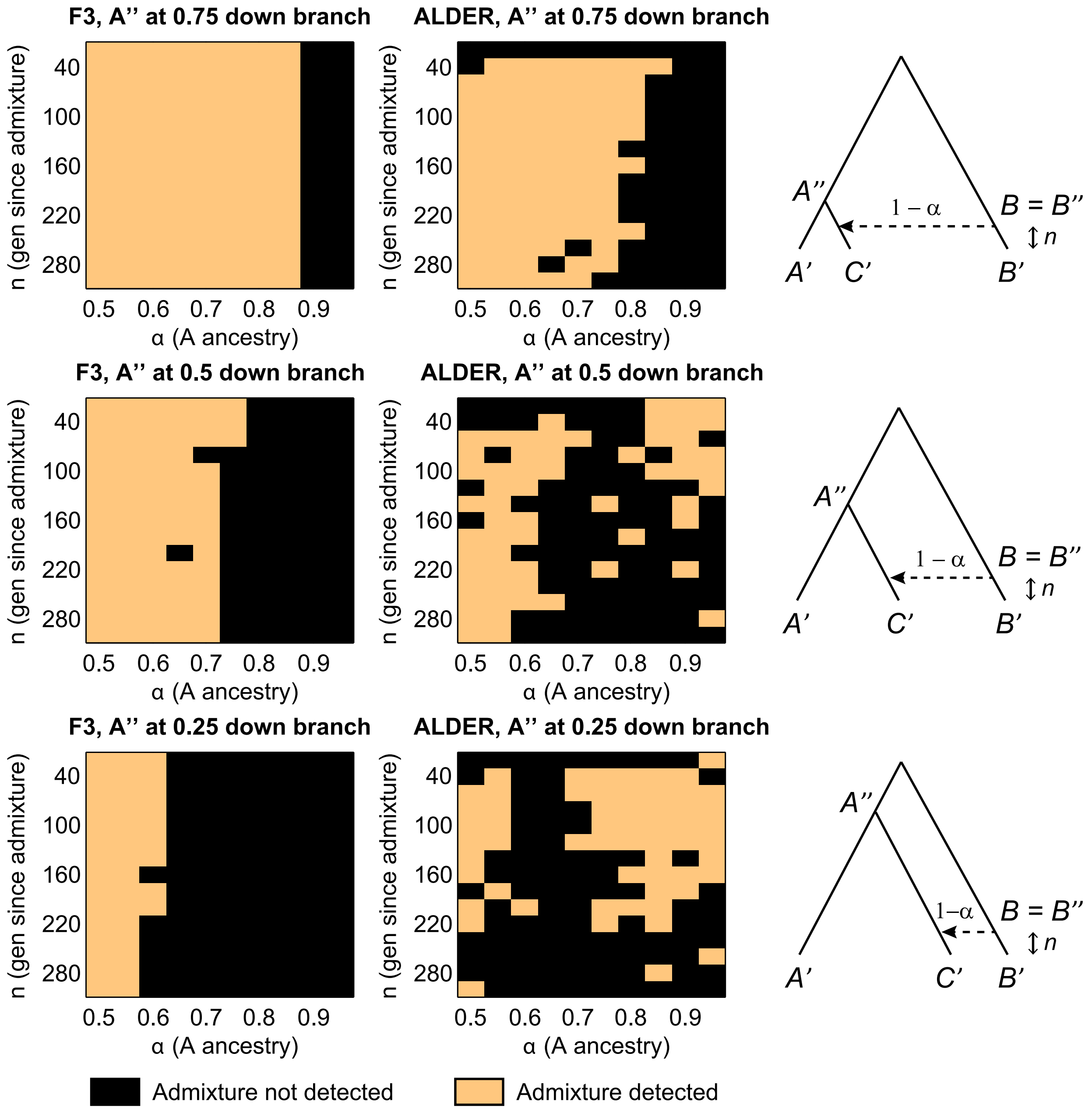

Figure S5. Coalescent simulations comparing the sensitivities of the 3-population moment-based test for admixture (f_3) and the LD-based test implemented in *ALDER*. We varied three parameters: the age of the branch point A'', the date n of gene flow, and the fraction α of A ancestry.

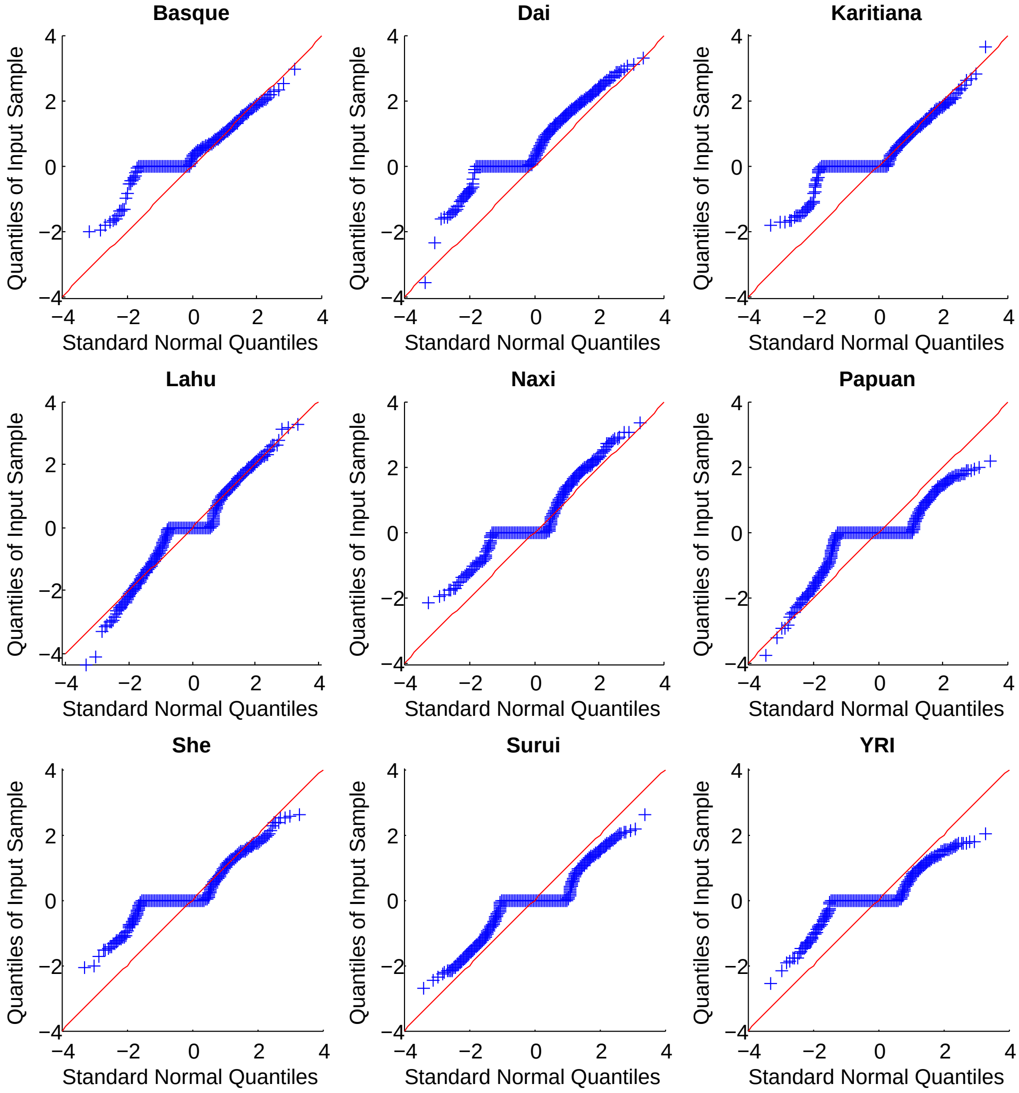

Figure S6. Q-Q plots comparing *ALDER* z-scores to standard normal on null examples. We show results from nine HGDP populations that neither *ALDER* nor the 3-population test found to be admixed. We are interested in values of $z > 0$; the Q-Q plots show that these values follow the standard normal reasonably well, tending to err on the conservative side.

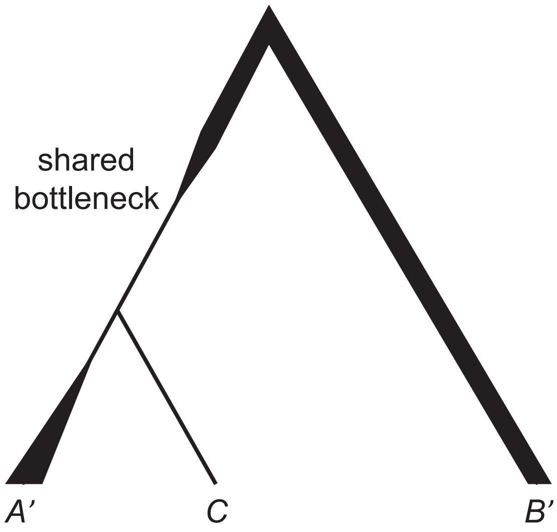

Figure S7. Non-admixture-related demography producing weighted LD curves. The test population is C and references are A' and B'; the common ancestor of A' and C experienced a recent bottleneck from which C has not yet recovered, leaving long-range LD in C that is potentially correlated to all three possible weighting schemes (A'--B', A'--C, and B'--C).

Table S1. Dates of admixture estimated for simulated 75% YRI / 25% CEU mixtures.

Ref 1	Ref 2	10	20	50	100	200
Yoruba	French	9±1	20±1	49±2	107±5	195±9
Yoruba	Han	9±1	21±1	50±2	107±6	191±12
Yoruba	Papuan	9±1	21±1	49±3	118±8	223±23
San	French	9±1	20±1	50±2	109±4	197±15
San	Han	9±0	21±1	51±3	111±4	194±16
San	Papuan	9±1	21±1	51±3	115±6	209±16
Yoruba		9±1	21±1	48±2	107±5	181±17
San		9±1	20±2	56±7	139±22	213±97
French		9±1	20±1	50±2	108±3	194±9
Han		9±0	21±1	52±2	110±6	192±17
Papuan		9±1	21±1	53±3	125±8	217±26

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various references. Rows in which only one reference is listed indicate runs using the admixed population itself as one reference. Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table S2. Dates of admixture estimated for simulated 90% YRI / 10% CEU mixtures.

Ref 1	Ref 2	10	20	50	100	200
Yoruba	French	10±0	21±1	50±2	107±7	193±19
Yoruba	Han	10±0	20±1	51±2	109±10	220±32
Yoruba	Papuan	10±0	22±1	53±3	111±11	233±65
San	French	10±0	21±1	51±2	112±6	223±19
San	Han	10±0	21±1	52±3	121±5	254±40
San	Papuan	11±0	23±1	53±3	126±8	287±56
Yoruba		9±1	20±2	55±7	100±27	363±183
San		98±87	56±28	94±69	2±0	9±5
French		10±0	21±1	51±2	107±5	217±13
Han		11±0	21±1	52±2	111±7	234±25
Papuan		11±0	23±1	56±3	117±8	256±47

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various references. Rows in which only one reference is listed indicate runs using the admixed population itself as one reference. Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table S3. Amplitudes of weighted LD curves (multiplied by 10^6) for simulated 75% YRI / 25% CEU mixtures.

Ref 1	Ref 2	Expected	10 gen	20 gen	50 gen	100 gen	200 gen
Yoruba	French	1173	1139±20	1203±40	1188±54	1283±100	1202±88
Yoruba	Han	693	678±17	717±28	711±43	774±73	716±74
Yoruba	Papuan	602	598±13	631±23	595±34	775±96	835±152
San	French	1017	981±23	1028±34	1044±49	1128±70	1037±130
San	Han	574	556±18	590±24	604±42	667±39	626±65
San	Papuan	491	487±17	514±20	503±34	589±45	574±60
Yoruba		75	77±2	81±4	74±4	83±6	71±13
San		40	40±3	42±3	50±6	66±13	43±34
French		655	626±12	660±21	666±31	721±42	656±49
Han		312	304±10	324±14	332±23	364±25	332±36
Papuan		252	256±9	273±13	267±17	331±34	314±55

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various references. Rows in which only one reference is listed indicate runs using the admixed population itself as one reference. Expected amplitudes were computed according to formulas (10) and (11). Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table S4. Amplitudes of weighted LD curves (multiplied by 10^6) for simulated 90% YRI / 10% CEU mixtures.

Ref 1	Ref 2	Expected	10 gen	20 gen	50 gen	100 gen	200 gen
Yoruba	French	563	587±27	579±26	550±25	600±43	562±96
Yoruba	Han	333	353±20	336±15	339±17	381±49	456±128
Yoruba	Papuan	289	307±19	303±16	309±18	343±54	426±248
San	French	488	522±25	512±22	488±25	519±28	625±89
San	Han	276	305±18	291±12	289±16	338±23	464±132
San	Papuan	236	266±18	262±13	254±12	306±38	486±186
Yoruba		6	6±1	6±1	7±1	7±3	44±89
San		1	16±15	8±3	10±7	-0±0	-1±1
French		454	473±19	471±18	450±19	481±19	566±55
Han		250	268±13	261±10	264±11	288±23	369±68
Papuan		212	231±14	233±13	243±11	276±35	366±125

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various references. Rows in which only one reference is listed indicate runs using the admixed population itself as one reference. Expected amplitudes were computed according to formulas (10) and (11). Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table S5. Mixture fraction lower bounds on simulated 75% YRI / 25% CEU mixtures.

Ref	10	20	50	100	200
French	24.6±0.3	25.7±0.5	25.7±0.7	27.0±1.0	25.2±1.3
Russian	23.8±0.3	24.9±0.5	24.8±0.7	25.6±0.8	25.3±1.0
Sardinian	21.3±0.3	21.9±0.5	22.0±0.6	23.6±0.9	22.3±1.1
Kalash	14.7±0.2	15.5±0.4	15.5±0.5	16.4±0.6	15.6±0.9
Yoruba	73.6±0.7	74.8±0.4	74.0±0.6	76.2±1.3	73.8±3.4
Mandenka	50.5±0.6	51.2±1.0	50.4±1.4	54.9±2.0	60.8±5.6

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various single references. The first four rows are European surrogates and give lower bounds on the amount of CEU ancestry (25%); the last two are African surrogates and give lower bounds on the amount of YRI ancestry (75%). Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table S6. Mixture fraction lower bounds on simulated 90% YRI / 10% CEU mixtures.

Ref	10	20	50	100	200
French	10.5±0.4	10.5±0.3	9.9±0.3	10.6±0.4	12.3±1.0
Russian	10.2±0.3	10.0±0.3	9.7±0.3	10.3±0.5	11.8±0.9
Sardinian	9.3±0.3	9.2±0.3	8.7±0.3	9.5±0.4	10.3±1.2
Kalash	7.2±0.3	7.0±0.3	6.8±0.2	7.4±0.4	8.9±0.8
Yoruba	89.1±1.0	89.1±1.1	90.1±1.5	89.4±3.7	98.5±2.0
Mandenka	18.2±2.3	17.3±2.5	19.5±4.8	63.1±25.5	30.7±220.4

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various single references. The first four rows are European surrogates and give lower bounds on the amount of CEU ancestry (10%); the last two are African surrogates and give lower bounds on the amount of YRI ancestry (90%). Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table S7. Dates of admixture estimated for simulated 75% YRI / 25% CEU mixtures.

Yoruba--French references					
Samples	10 gen	20 gen	50 gen	100 gen	200 gen
5	12±2	18±2	55±3	103±7	258±24
10	10±1	19±2	50±2	105±7	236±24
20	10±1	20±1	52±2	104±5	223±16
50	9±0	20±1	52±1	96±2	186±10
100	10±0	20±0	52±1	101±2	210±9
San--Han references					
Samples	10 gen	20 gen	50 gen	100 gen	200 gen
5	12±2	18±2	58±5	107±11	283±73
10	10±1	19±2	54±3	114±8	219±64
20	10±1	21±1	55±2	115±6	219±46
50	9±0	21±1	54±1	107±5	213±20
100	9±0	21±1	53±1	105±5	216±13

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using varying numbers of admixed samples. Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table S8. Dates of admixture estimated for simulated 90% YRI / 10% CEU mixtures.

Yoruba--French references					
Samples	10 gen	20 gen	50 gen	100 gen	200 gen
5	11±2	21±2	52±6	101±17	253±42
10	11±1	19±1	48±4	94±8	241±46
20	11±1	21±1	48±3	102±8	209±30
50	11±0	21±1	48±2	98±5	202±21
100	10±0	20±1	50±1	99±4	185±15
San--Han references					
Samples	10 gen	20 gen	50 gen	100 gen	200 gen
5	14±2	22±3	63±8	110±30	335±91
10	12±1	20±2	54±4	110±15	265±55
20	12±1	21±1	52±4	131±15	234±33
50	11±0	20±1	53±4	122±8	221±23
100	11±0	20±0	53±3	109±5	219±10

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using varying numbers of admixed samples. Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table S9. Effect of SNP ascertainment on date estimates.

Mixed pop	Ref 1	Ref 2	French asc	Han asc	San asc	Yoruba asc
Burusho	French	Han	47±12	51±13	56±10	41±10
Uygur	French	Han	15±2	14±2	13±2	16±2
Hazara	French	Han	22±2	22±3	23±2	22±3
Melanesian	Dai	Papuan	93±24	62±15	76±13	70±18
Bedouin	French	Yoruba	27±3	23±3	23±3	24±3
MbutiPygmy	San	Yoruba	33±12	33±6	41±14	30±8
BiakaPygmy	San	Yoruba	39±6	50±14	35±6	36±7

We compared dates of admixture estimated by *ALDER* on a variety of test triples from the HGDP using SNPs ascertained as heterozygous in full genome sequences of one French, Han, San, and Yoruba individual (Panels 1, 2, 4, and 5 of the Affymetrix Human Origins Array (PATTERSON *et al.* 2012)). Standard errors are from a jackknife over the 22 autosomes.

Table S10. Effect of SNP ascertainment on weighted LD curve amplitudes (multiplied by 10^6).

Mixed pop	Ref 1	Ref 2	French asc	Han asc	San asc	Yoruba asc
Burusho	French	Han	180±44	171±53	61±11	65±15
Uygur	French	Han	360±28	304±29	102±7	161±19
Hazara	French	Han	442±31	436±48	146±10	203±21
Melanesian	Dai	Papuan	868±277	559±150	207±51	312±91
Bedouin	French	Yoruba	227±32	196±25	104±11	146±13
MbutiPygmy	San	Yoruba	64±23	78±14	83±26	82±18
BiakaPygmy	San	Yoruba	104±19	133±46	90±15	103±22

We compared amplitudes of weighted LD curves fitted on a variety of test triples from the HGDP using SNPs ascertained as heterozygous in full genome sequences of one French, Han, San, and Yoruba individual (Panels 1, 2, 4, and 5 of the Affymetrix Human Origins Array (PATTERSON *et al.* 2012)). Standard errors are from a jackknife over the 22 autosomes.

File S1. Unbiased polyache estimator for weighted LD using the admixed population itself as one reference.

```
In[28]:= MomentConvert[
           MomentConvert[CentralMoment[{1, 1}] (Moment[{1, 0}] - pAx) (Moment[{0, 1}] - pAy),
            "UnbiasedSampleEstimator"], "PowerSymmetricPolynomial"] // TraditionalForm
```

Out[28]//TraditionalForm=

$$
\begin{aligned}
&-\frac{\mathrm{pAx}\,\mathrm{pAy}\,S_{1,0}\,S_{0,1}}{{S_0}^{(2)}}+\frac{\mathrm{pAx}\,\mathrm{pAy}\,S_{1,1}\left({S_0}^{(2)}+S_0\right)}{S_0\,{S_0}^{(2)}}+\frac{\mathrm{pAx}\,S_{1,0}\,{S_{0,1}}^2}{{S_0}^{(3)}}-\frac{\mathrm{pAx}\,S_{1,1}\,S_{0,1}\left(2\,{S_0}^{(2)}+{S_0}^{(3)}\right)}{{S_0}^{(2)}\,{S_0}^{(3)}}-\\
&\frac{\mathrm{pAx}\,S_{0,2}\,S_{1,0}}{{S_0}^{(3)}}+\frac{\mathrm{pAx}\,S_{1,2}\left(2\,{S_0}^{(2)}+{S_0}^{(3)}\right)}{{S_0}^{(2)}\,{S_0}^{(3)}}+\frac{\mathrm{pAy}\,{S_{1,0}}^2\,S_{0,1}}{{S_0}^{(3)}}-\frac{\mathrm{pAy}\,S_{2,0}\,S_{0,1}}{{S_0}^{(3)}}-\frac{\mathrm{pAy}\,S_{1,0}\,S_{1,1}\left(2\,{S_0}^{(2)}+{S_0}^{(3)}\right)}{{S_0}^{(2)}\,{S_0}^{(3)}}+\\
&\frac{\mathrm{pAy}\,S_{2,1}\left(2\,{S_0}^{(2)}+{S_0}^{(3)}\right)}{{S_0}^{(2)}\,{S_0}^{(3)}}-\frac{{S_{1,0}}^2\,{S_{0,1}}^2}{{S_0}^{(4)}}+\frac{S_{2,0}\,{S_{0,1}}^2}{{S_0}^{(4)}}+\frac{S_{1,0}\,S_{1,1}\,S_{0,1}\left(4\,{S_0}^{(3)}+{S_0}^{(4)}\right)}{{S_0}^{(3)}\,{S_0}^{(4)}}+\frac{S_{2,1}\,S_{0,1}\left(-4\,{S_0}^{(3)}-{S_0}^{(4)}\right)}{{S_0}^{(3)}\,{S_0}^{(4)}}+\\
&\frac{S_{0,2}\,{S_{1,0}}^2}{{S_0}^{(4)}}+\frac{{S_{1,1}}^2\left(-2\,{S_0}^{(3)}-{S_0}^{(4)}\right)}{{S_0}^{(3)}\,{S_0}^{(4)}}+\frac{S_{1,0}\,S_{1,2}\left(-4\,{S_0}^{(3)}-{S_0}^{(4)}\right)}{{S_0}^{(3)}\,{S_0}^{(4)}}-\frac{S_{0,2}\,S_{2,0}}{{S_0}^{(4)}}+\frac{2\,S_{2,2}\left(3\,{S_0}^{(3)}+{S_0}^{(4)}\right)}{{S_0}^{(3)}\,{S_0}^{(4)}}
\end{aligned}
$$

Mathematica code and output are shown for computing the polyache statistic that estimates the one-reference weighted LD, $E[(X-\mu_x)(Y-\mu_y)(\mu_x-p_A(x))(\mu_y-p_A(y))]$, where $p_A(\cdot)$ are allele frequencies of the single reference population and μ_x and μ_y denote allele frequencies of the admixed population. In the above, $S_0^{(k)}:=m(m-1)\cdots(m-k+1)$ and $S_{r,s}:=\sum_{i=1}^m X_i^r Y_i^s$, where m is the number of admixed samples and i ranges over the admixed individuals, which have allele counts X_i and Y_i at sites x and y.

File S2. FFT computation of weighted LD.

In this note we describe how to compute weighted LD (aggregated over distance bins) in time

$$O(m(S + B \log B)),$$

where m is the number of admixed individuals, S is the number of SNPs, and B is the number of bins needed to span the chromosomes. In contrast, the direct method of computing pairwise LD for each individual SNP pair requires $O(mS^2)$ time. In practice our approach offers speedups of over 1000x on typical data sets. We further describe a similar algorithm for computing the single-reference weighted LD polyache statistic that runs in time

$$O(m^2(S + B \log B))$$

with the slight trade-off of ignoring SNPs with missing data.

Our method consists of three key steps: (1) split and factorize the weighted LD product; (2) group factored terms by bin; and (3) apply fast Fourier transform (FFT) convolution. As a special case of this approach, the first two ideas alone allow us to efficiently compute the affine term (i.e., horizontal asymptote) of the weighted LD curve using inter-chromosome SNP pairs.

TWO-REFERENCE WEIGHTED LD

We first establish notation. Say we have an $S \times m$ genotype array $\{c_{x,i}\}$ from an admixed population. Assume for now that there are no missing values, i.e.,

$$c_{x,i} \in \{0, 1, 2\}$$

for x indexing SNPs by position on a genetic map and $i = 1, \ldots, m$ indexing individuals. Given a set of weights w_x, one per SNP, we wish to compute weighted LD of SNP pairs aggregated by inter-SNP distance d:

$$R(d) := \sum_{\substack{|x-y| \approx d \\ x<y}} D_2(x, y) w_x w_y = \frac{1}{2} \sum_{|x-y| \approx d} D_2(x, y) w_x w_y$$

where D_2 is the sample covariance between genotypes at x and y, the diploid analog of the usual LD measure D:

$$\begin{aligned} D_2(x, y) \quad &:= \quad \frac{1}{m-1} \sum_{i=1}^{m} c_{x,i} c_{y,i} - \frac{1}{m(m-1)} \sum_{i=1}^{m} c_{x,i} \sum_{j=1}^{m} c_{y,j} \\ &= \quad \frac{1}{m-1} \sum_{i=1}^{m} c_{x,i} c_{y,i} - \frac{1}{m(m-1)} s_x s_y, \end{aligned} \qquad (1)$$

where we have defined

$$s_x := \sum_{i=1}^{m} c_{x,i}.$$

Substituting for $D_2(x, y)$, we have

$$\begin{aligned} R(d) \quad &= \quad \frac{1}{2} \sum_{|x-y| \approx d} \left(\frac{1}{m-1} \sum_{i=1}^{m} c_{x,i} c_{y,i} - \frac{1}{m(m-1)} s_x s_y \right) w_x w_y \\ &= \quad \left(\sum_{i=1}^{m} \frac{1}{2(m-1)} \sum_{|x-y| \approx d} c_{x,i} w_x \cdot c_{y,i} w_y \right) - \frac{1}{2m(m-1)} \sum_{|x-y| \approx d} s_x w_x \cdot s_y w_y. \end{aligned} \qquad (2)$$

We have thus rewritten $R(d)$ as a linear combination of $m+1$ terms of the form

$$\sum_{|x-y|\approx d} f(x)f(y).$$

(The sum over i consists of m such terms, and the final term accounts for one more.)

In general, sums of the form

$$\sum_{|x-y|\approx d} f(x)g(y)$$

can be efficiently computed by convolution if we first discretize the genetic map on which the SNP positions x and y lie. For notational convenience, choose the distance scale such that a unit distance corresponds to the desired bin resolution. We will compute

$$\sum_{\lfloor x\rfloor - \lfloor y\rfloor = d} f(x)g(y). \tag{3}$$

That is, we divide the chromosome into bins of unit distance and aggregate terms $f(x)g(y)$ by the distance between the bin centers of x and y. Note that this procedure does not produce exactly the same result as first subtracting the genetic positions and then binning by $|x-y|$: with our approach, pairs (x,y) that map to a given bin can have actual distances that are off by as much as one full bin width, versus half a bin width with the subtract-then-bin approach. However, we can compensate simply by doubling the bin resolution.

To compute expression (3), we write

$$\begin{aligned}\sum_{\lfloor x\rfloor - \lfloor y\rfloor = d} f(x)g(y) &= \sum_{b=0}^{B} \sum_{\lfloor x\rfloor = b} \sum_{\lfloor y\rfloor = b-d} f(x)g(y) \\ &= \sum_{b=0}^{B} \left(\sum_{\lfloor x\rfloor = b} f(x) \right) \left(\sum_{\lfloor y\rfloor = b-d} g(y) \right). \end{aligned} \tag{4}$$

Writing

$$F(b) := \sum_{\lfloor x\rfloor = b} f(x), \quad G(b) := \sum_{\lfloor x\rfloor = b} g(x),$$

expression (4) becomes

$$\sum_{b=0}^{B} F(b)G(b-d) = (F \star G)(d),$$

a cross-correlation of binned $f(x)$ and $g(y)$ terms.

Computationally, binning f and g to form F and G takes $O(S)$ time, after which the cross-correlation can be performed in $O(B \log B)$ time with a fast Fourier transform. The full computation of the $m+1$ convolutions in equation (2) thus takes $O(m(S + B \log B))$ time. In practice we often have $B \log B < S$, in which case the computation is linear in the data size mS.

One additional detail is that we usually want to compute the average rather than the sum of the weighted LD contributions of the SNP pairs in each bin; this requires normalizing by the number of pairs (x,y) that map to each bin, which can be computed in an analogous manner with one more convolution (setting $f \equiv 1$, $g \equiv 1$). Finally, we note that our factorization and binning approach immediately extends to computing weighted LD on inter-chromosome SNP pairs (by putting all SNPs in a chromosome in the same bin), which allows robust estimation of the horizontal asymptote of the weighted LD curve.

Missing Data The calculations above assumed that the genotype array contained no missing data, but in practice a fraction of the genotype values may be missing. The straightforward non-FFT computation has no difficulty handling missing data, as each pairwise LD term $D_2(x, y)$ can be calculated as a sample covariance over just the individuals successfully genotyped at both x and y. Our algebraic manipulation runs into trouble, however, because if k individuals have a missing value at either x or y, then the sample covariance contains denominators of the form $1/(m-k-1)$ and $1/(m-k)(m-k-1)$---and k varies depending on x and y.

One way to get around this problem is simply to restrict the analysis to sites with no missing values at the cost of slightly reduced power. If a fraction p of the SNPs contain at least one missing value, this workaround reduces the number of SNP pairs available to $(1-p)^2$ of the total, which is probably already acceptable in practice.

We can do better, however: in fact, with a little more algebra (but no additional computational complexity), we can include all pairs of sites (x, y) for which at least one of the SNPs x, y has no missing values, bringing our coverage up to $1-p^2$.

We will need slightly more notation. Adopting `eigenstrat` format, we now let our genotype array consist of values

$$c_{x,i} \in \{0, 1, 2, 9\}$$

where 9 indicates a missing value. (Thus, $\{c_{x,i}\}$ is exactly the data that would be contained in a `.geno` file.) For convenience, we write

$$c_{x,i}^{(0)} := \begin{cases} c_{x,i} & \text{if } c_{x,i} \in \{0, 1, 2\} \\ 0 & \text{otherwise.} \end{cases}$$

That is, $c_{x,i}^{(0)}$ replaces missing values with 0s. As before we set

$$s_x := \sum_{i:c_{x,i} \neq 9} c_{x,i} = \sum_{i=1}^{m} c_{x,i}^{(0)}$$

to be the sum of all non-missing values at x, which also equals the sum of all $c_{x,i}^{(0)}$ because the missing values have been 0-replaced. Finally, define

$$k_x := \#\{i : c_{x,i} = 9\}$$

to be the number of missing values at site x.

We now wish to compute aggregated weighted LD over pairs (x, y) for which at least one of k_x and k_y is 0. Being careful not to double-count, we have:

$$\begin{aligned} R(d) &:= \sum_{\substack{|x-y|\approx d \\ x<y \\ k_x=0 \text{ or } k_y=0}} D_2(x, y) w_x w_y \\ &= \frac{1}{2} \sum_{\substack{|x-y|\approx d \\ k_x=0 \text{ and } k_y=0}} D_2(x, y) w_x w_y + \sum_{\substack{|x-y|\approx d \\ k_x=0 \text{ and } k_y\neq 0}} D_2(x, y) w_x w_y \\ &= \sum_{|x-y|\approx d} \frac{I[k_x = 0]}{1 + I[k_y = 0]} D_2(x, y) w_x w_y, \qquad (5) \end{aligned}$$

where the shorthand $I[\cdot]$ denotes a $\{0,1\}$-indicator.

Now, for a pair of sites (x,y) where x has no missing values and y has k_y missing values,

$$D_2(x,y) = \frac{1}{m-k_y-1}\sum_{i=1}^{m} c_{x,i}c_{y,i}^{(0)} - \frac{1}{(m-k_y)(m-k_y-1)}\left(s_x - \sum_{i=1}^{m} I[c_{y,i}=9]c_{x,i}\right)s_y. \tag{6}$$

Indeed, we claim the above equation is actually just a rewriting of the standard covariance formula (1), appropriately modified now that the covariance is over $m-k_y$ values rather than m:

- In the sum $\sum_{i=1}^{m} c_{x,i}c_{y,i}^{(0)}$, missing values in y have been 0-replaced, so those terms vanish and the sum effectively consists of the desired $m-k_y$ products $c_{x,i}c_{y,i}$.
- Similarly, s_y is equal to the sum of the $m-k_y$ non-missing $c_{y,i}$ values.
- Finally, $s_x - \sum_{i=1}^{m} I[c_{y,i}=9]c_{x,i}$ represents the sum of $c_{x,i}$ over individuals i successfully genotyped at y, written as the sum s_x over all m individuals minus a correction.

Substituting (6) into expression (5) for $R(d)$ and rearranging, we have

$$\begin{aligned} R(d) &= \sum_{|x-y|\approx d} \frac{I[k_x=0]}{1+I[k_y=0]}\left(\frac{1}{m-k_y-1}\sum_{i=1}^{m} c_{x,i}c_{y,i}^{(0)}\right. \\ &\qquad\qquad \left. - \frac{1}{(m-k_y)(m-k_y-1)}\left(s_x - \sum_{i=1}^{m} I[c_{y,i}=9]c_{x,i}\right)s_y\right) w_x w_y \\ &= \sum_{i=1}^{m}\sum_{|x-y|\approx d} \left(I[k_x=0]c_{x,i}w_x\right)\cdot\left(\frac{1}{1+I[k_y=0]}\left(c_{y,i}^{(0)} + \frac{I[c_{y,i}=9]s_y}{m-k_y}\right)\frac{w_y}{m-k_y-1}\right) \\ &\quad - \sum_{|x-y|\approx d}\left(I[k_x=0]s_x w_x\right)\cdot\left(\frac{s_y w_y}{(1+I[k_y=0])(m-k_y)(m-k_y-1)}\right). \end{aligned}$$

The key point is that we once again have a sum of $m+1$ convolutions of the form $\sum_{|x-y|\approx d} f(x)g(y)$ and thus can compute them efficiently as before.

ONE-REFERENCE WEIGHTED LD

When computing weighted LD using the admixed population itself as a reference with one other reference population, a polyache statistic must be used to obtain an unbiased estimator (File S1). The form of the polyache causes complications in our algebraic manipulation; however, if we restrict our attention to SNPs with no missing data, the computation can still be broken into convolutions quite naturally, albeit now requiring $O(m^2)$ FFTs rather than $O(m)$.

As in the two-reference case, the key idea is to split and factorize the weighted LD formula. We treat the terms in the polyache separately and observe that each term takes the form of a constant factor multiplied by a product of sub-terms of the form $S_{r,s}$, $p_A(x)$, or $p_A(y)$. We can use convolution to aggregate the contributions of such a term if we can factor it as a product of two pieces, one depending only on x and the other only on y. Doing so is easy for some terms, namely those that involve only $p_A(x)$, $p_A(y)$, $S_{r,0}$, and $S_{0,s}$, as the latter two sums depend only on x and y, respectively.

The terms involving $S_{r,s}$ with both r and s nonzero are more difficult to deal with but can be written as convolutions by further subdividing them. In fact, we already encountered $S_{1,1} = \sum_{i=1}^{m} c_{x,i}c_{y,i}$ in our two-reference weighted LD computation: the trick there was to split the sum into its m components, one per admixed individual, each of which could then be factored into x-dependent and y-dependent parts and aggregated via convolution.

Exactly the same decomposition works for all of the polyache terms except the one involving $S_{1,1}^2$. For this term, we write

$$S_{1,1}^2 = \sum_{i=1}^{m} c_{x,i}c_{y,i} \sum_{j=1}^{m} c_{x,j}c_{y,j} = \sum_{i=1}^{m}\sum_{j=1}^{m} c_{x,i}c_{x,j} \cdot c_{y,i}c_{y,j},$$

from which we see that splitting the squared sum into m^2 summands allows us to split the x- and y-dependence as desired. The upshot is that at the expense of $O(m^2)$ FFTs (and restricting our analysis to SNPs without missing data), we can also accelerate the one-reference weighted LD computation.