The Spatial Mixing of Genomes in Secondary Contact Zones

Alisa Sedghifar,*[,1] Yaniv Brandvain,† Peter Ralph,‡ and Graham Coop*
*Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, California 95616,
†Department of Plant Biology, University of Minnesota, St. Paul, Minnesota 55108, and ‡Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089

ABSTRACT Recent genomic studies have highlighted the important role of admixture in shaping genome-wide patterns of diversity. Past admixture leaves a population genomic signature of linkage disequilibrium (LD), reflecting the mixing of parental chromosomes by segregation and recombination. These patterns of LD can be used to infer the timing of admixture, but the results of inference can depend strongly on the assumed demographic model. Here, we introduce a theoretical framework for modeling patterns of LD in a geographic contact zone where two differentiated populations have come into contact and are mixing by diffusive local migration. Assuming that this secondary contact is recent enough that genetic drift can be ignored, we derive expressions for the expected LD and admixture tract lengths across geographic space as a function of the age of the contact zone and the dispersal distance of individuals. We develop an approach to infer age of contact zones, using population genomic data from multiple spatially sampled populations by fitting our model to the decay of LD with recombination distance. To demonstrate an application of our model, we use our approach to explore the fit of a geographic contact zone model to three human genomic data sets from populations in Indonesia, Central Asia, and India and compare our results to inference under different demographic models. We obtain substantially different results from those of the commonly used model of panmictic admixture, highlighting the sensitivity of admixture timing results to the choice of demographic model.

KEYWORDS contact zones; admixture; linkage disequilibrium

POPULATIONS frequently undergo periods of relative isolation that are followed by secondary contact. During isolation, the evolutionary processes of genetic drift, mutation, and selection act to differentiate populations at many markers throughout the genome. When these populations come back into contact, the restoration of gene flow generates admixed populations, which start as an assemblage of differentiated parental genomes that are broken up every generation by segregation and recombination between chromosomes.

Under this process, linked alleles of the same ancestry will tend to be co-inherited until separated by recombination. Because the parental populations are differentiated with respect to each other, this co-inheritance leads to a nonrandom association of alleles, referred to as linkage disequilibrium (LD). This admixture-induced LD (or admixture LD) is the resulting covariance between loci and initially extends over a much larger genomic scale than LD does in either parental population and is a signature of relatively recent admixture (Cavalli-Sforza and Bodmer 1971; Chakraborty and Weiss 1988). One can also think of this signature as the persistence of parental haplotypes in admixed populations that, rather than being measured directly, is measured as the extent of co-occurrence along a chromosome of alleles that are diagnostic of parental origin. Recombination acts every generation to gradually break apart long tracts of ancestry into smaller tracts, and so the association between nearby alleles lasts many generations. The physical scale over which admixture LD breaks down is determined by the timescale over which parental populations have been interbreeding; the conservation of many ancestral haplotypes over large physical distances would imply very recent admixture, whereas a longer

doi: 10.1534/genetics.115.179838
Manuscript received June 25, 2015; accepted for publication July 18, 2015; published Early Online July 22, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179838/-/DC1.
[1] Corresponding author: Department of Evolution and Ecology, One Shields Ave., Davis, CA 95616. E-mail: sedghifar@ucdavis.edu

history of admixture produces many smaller parental tracts. We assume that population differentiation within the parental populations is weak relative to that between them and so consider only admixture LD (and not mixture LD that is covariance between loci induced by substructure) in the focal populations.

Data from many (potentially weakly) differentiated markers allow for the identification and quantification of admixture in individuals (*e.g.*, Pritchard *et al.* 2000) and the inference of the ancestral origin of a given chromosomal region (*e.g.*, Falush *et al.* 2003; Price *et al.* 2009; Hellenthal *et al.* 2014). The continued mixing of differentiated genotypes, as described above, produces predictable population genomic patterns that change through time, and these signals can be used to not only detect past admixture in an extant population, but also learn about the timing and history of these admixture events (*e.g.*, Harris and Nielsen 2013; Loh *et al.* 2013; Hellenthal *et al.* 2014). Such inferences have been used to reconstruct historical population movements, highlighting the importance of admixture in shaping patterns of diversity in human populations (Reich *et al.* 2009; Patterson *et al.* 2012; Loh *et al.* 2013; Moorjani *et al.* 2013; Hellenthal *et al.* 2014). These studies have utilized powerful methods that first identify stretches of chromosome inherited from a particular parental population [admixture tracts (Gravel 2012; Hellenthal *et al.* 2014)] or measure the covariance, over spatial scales, of variants that are diagnostic of parental populations [admixture LD (Patterson *et al.* 2012; Loh *et al.* 2013)] and then infer the genetic scale over which this measured coancestry decays. Commonly this is done by assuming a model of admixture in which one isolated population is formed by a single admixture event in time, with subsequent random mating. Under this simple model, the distribution of admixture tract lengths and the decay of admixture LD with respect to genetic distance are approximately exponential, with the rate parameter corresponding to the time in generations since admixture. However, violations of the assumptions of the single-pulse model can result in substantial departure between expected and observed rates of decay of coancestry with respect to time.

Models incorporating multiple admixture times, or sustained migration (Pool and Nielsen 2009; Gravel 2012; Hellenthal *et al.* 2014; Liang and Nielsen 2014b), have been built to address more complex admixture scenarios in single populations. However, these do not incorporate the fact that admixture often occurs in a geographic context—beginning at a given point in time, then spreading across space. Most current models treat each admixed population as an independent event, not accounting for this spatial context, even when admixture in spatially distributed populations is potentially attributable to a single historical event.

In this article we build an alternative model of diffusion of ancestry across geography in time. Specifically, we consider a scenario in which two populations spread back into contact, generating a gradient of admixture across space with the greatest variance in ancestry at the point of initial contact. We refer to this mixture across space, where migration is sustained through both time and space, as a contact zone. This geographic mixing leads to departures from a simple model of exponential decay of admixture LD as there is exchange of migrants between neighboring populations with different admixture proportions. We describe the expected covariance in ancestry (ancestry LD) in contact zones, accounting for migration in continuous space. By assuming a large population that is not affected by genetic drift, and therefore ignoring coalescence, we are able to derive an analytic expression for LD in contact zones. This model provides a framework to simultaneously examine admixture patterns over a set of geographically distributed populations and a potential geographic null model for studying historical movements of populations. Inference under this model provides a means to estimate both the time at which populations spread back into contact and some measures of dispersal. We analyze several potential human contact zones under our model and show that simpler "point" models of admixture can infer unreasonably recent admixture dates. In addition to human admixture, LD has also been used to characterize hybrid zones (*e.g.*, Szymura and Barton 1986; Mallet *et al.* 1990; Wang *et al.* 2011), and so the model presented here also has applications in the study of such secondary contact between incipient species. This could potentially complement earlier investigations of coancestry along hybrid zones in the presence of selection (Barton 1979, 1983; Barton and Bengtsson 1986).

Materials and Methods

Outline of neutral model

We consider two differentiated populations along a transect in space, formerly separated by a barrier that completely prevented migration (at position $x = 0$) that was removed τ generations ago (Figure 1). We imagine the barrier as a physical obstruction to migration; however, in practice the two previously isolated populations could come into contact through a variety of means. We use a continuous-space limit of randomly mating (Wright–Fisher) populations on a line, made formal in, *e.g.*, Nagylaki (1975) and Shiga (1980), which can be described informally as follows.

Since time τ, individuals have moved without restriction, in such a way that the distribution of displacements between an ancestor and a descendant separated by t generations is Gaussian with mean zero and variance $\sigma^2 t$. The Gaussian assumption is appropriate since, over many generations, the sum of many steps under a finite-variance dispersal kernel will converge to a Gaussian distribution. This forms a gradient of admixed populations across space, whose degree of admixture depends on the time that has passed and the distance to the point of initial contact. Over time, genotypes of different ancestries diffuse across the entire range, and recombination breaks down tracts of continuous ancestry. We aim to describe this diffusion of ancestry throughout time and space.

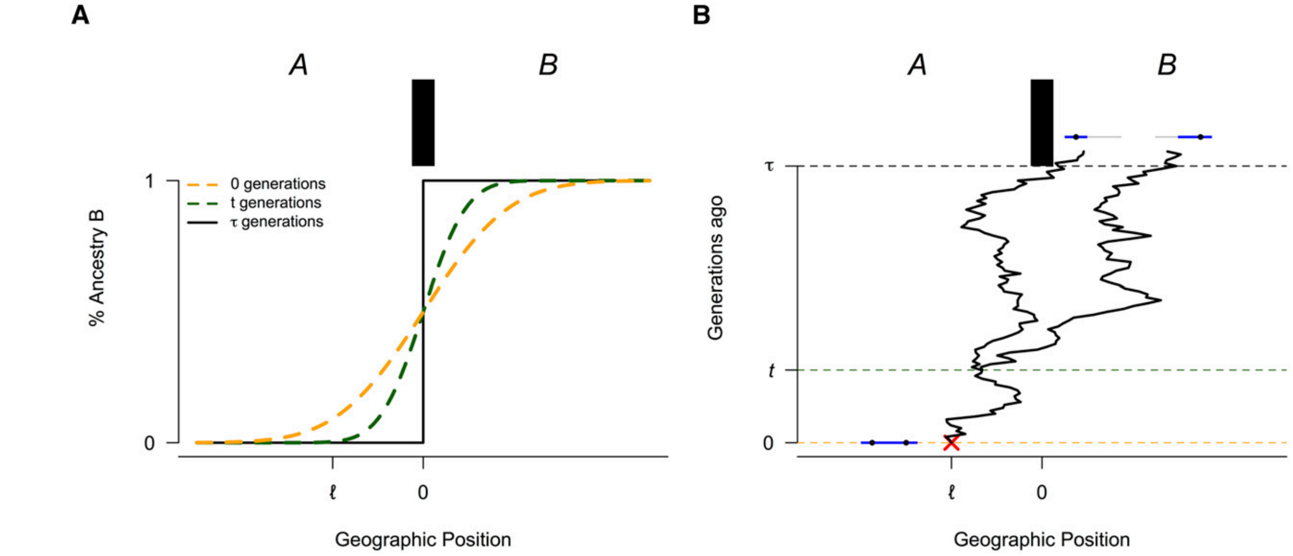


Figure 1 A schematic of the model in space and time. In A and B, the black rectangle represents a barrier to dispersal that is removed at a time τ generations in the past, after which there is unrestricted gene flow. (A) Gene flow resumes between two initially isolated populations τ generations ago, with a dispersal kernel of variance σ^2. The expected cline in ancestry proportion is a function of time since initial contact, and the observed cline in the present day (0 generations ago) is a function of τ. (B) We follow backward in time the Brownian motion paths of two initially linked lineages, represented here by two black circles located on a blue chromosome. The paths of the two lineages are identical until the first recombination event between them at time *t*, after which they follow independent Brownian paths. The red cross indicates the position, relative to the center of the zone, where the chromosome was sampled in the present day. In this example, both alleles are of ancestry *B*, since they are on the same side of the barrier to dispersal at time τ.

To determine the typical degree of admixture at a location, we follow the lineage of a sampled individual back through time, tracing the spatial location of the ancestor of today's sample back to the initiation of secondary contact. The ancestral type of today's sample is determined by the geographic position of its ancestor τ generations ago: we say that a sampled individual whose lineage falls to the left of the barrier (*i.e.*, whose ancestor τ generations ago lived at a location $x < 0$) is of ancestry *A* and is otherwise of ancestry *B*. This represents the alleles belonging to ancestral population *A* or *B* before the initiation of secondary contact. We treat time and space as continuous variables, and the time-reversible properties of Brownian motion allow us to model the movement of lineages as a continuous Brownian process. The framework we present here is explicitly defined in a one-dimensional geographic transect, but also applies, unchanged, in two dimensions due to the absence of drift.

Behavior of a single locus

We start by describing the marginal profile of admixture proportions. Suppose that we sample a randomly chosen individual today from position ℓ relative to the center of the contact zone and define $\mathcal{A}$ to be that individual's ancestry at a randomly chosen locus (*i.e.*, $\mathcal{A}$ is equal to *B* if their ancestor at the time of secondary contact τ generations ago lived to the right of the barrier). Since we assume that the displacement between parents and offspring is Gaussian with variance σ^2, we can describe the movement of the lineage as a Brownian motion, and so the probability that $\mathcal{A}$ is of ancestry *B* is equal to the probability that a Brownian motion that begins at ℓ is to the right of zero after τ generations; *i.e.*,

$$\mathbb{E}[\mathbf{1}_B(\mathcal{A})] = \int_{-(\ell/\sigma\sqrt{\tau})}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = \Phi\left(\frac{\ell}{\sigma\sqrt{\tau}}\right). \quad (1)$$

Here $\mathbf{1}_B(\mathcal{A})$ is the indicator function:

$$\mathbf{1}_B(\mathcal{A}) = \begin{cases} 1 & \mathcal{A} \text{ has ancestry } B \\ 0 & \mathcal{A} \text{ has ancestry } A. \end{cases}$$

In other words, the probability that an individual sampled at geographic position ℓ inherits at a given locus from ancestral population *B* is the probability that $x_\tau > 0$, where x_τ is Gaussian with mean ℓ and variance $\tau\sigma^2$. This is also the expected frequency of ancestry *B* at position ℓ, τ generations after contact, and therefore provides an expectation of the cline in ancestry proportion (Figure 1). Although it is convenient to imagine the motion of a lineage as a Brownian motion in continuous time, this expression also holds for discrete generations since the distribution of parent–offspring dispersal is Gaussian with variance σ^2, and then the total displacement across τ generations is also Gaussian, with variance $\tau\sigma^2$.

Under this model, we expect the zone of significant admixture, where admixture LD is observable, to extend over distance $\sim 2\sigma\sqrt{\tau}$ in either direction from the center of the zone. Therefore, to fit our model using the inference framework we describe below, we will need samples on this spatial scale. We note that while this may be a good model for small τ, the prediction under this model that admixture proportions homogenize as τ becomes very large can be unrealistic as barriers to intermixing may persist over time and therefore may not be an appropriate model for very old contact zones.

Ancestry LD between linked loci

In our model, all chromosomes begin as unbroken tracts of ancestry prior to initial contact. As time progresses, recombination between haplotypes of different ancestry breaks down these associations. To model this effect, we consider two linked loci separated by a recombination fraction r, on a single chromosome sampled at geographic position ℓ, τ generations after secondary contact (see Figure 1 and legend). The ancestries of a sampled individual at these two loci are denoted $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. If there is no recombination between these two loci, then both lineages trace the same path back in time, and $\mathbf{1}_B(\mathcal{A}_1) = \mathbf{1}_B(\mathcal{A}_2)$. The recombination fraction between the loci is the per generation probability of observing a recombinant haplotype as the product of meiosis. For close pairs of markers it may suffice to use the genetic distance d in morgans that separates markers, but for more distant markers we can use the probability of an *observed* recombination event, which is the probability of an odd number of recombination events between focal loci, accounting for interference when possible. Assuming no interference (*i.e.*, a Poisson model), the relationship between d and r is given by Haldane's mapping function, $r = (1 - e^{-2d})/2$ (Haldane 1919).

We measure ancestry LD as the covariance in ancestry between the alleles at the two loci,

$$\begin{aligned}&\operatorname{Cov}(\mathbf{1}_B(\mathcal{A}_1), \mathbf{1}_B(\mathcal{A}_2))\\ &\quad = \mathbb{E}[\mathbf{1}_B(\mathcal{A}_1)\mathbf{1}_B(\mathcal{A}_2)] - \mathbb{E}\big[\mathbf{1}_B(\mathcal{A}_1)\big]\mathbb{E}\big[\mathbf{1}_B(\mathcal{A}_2)\big].\end{aligned} \tag{2}$$

Since $\mathcal{A}_1$ and $\mathcal{A}_2$ have the same distribution, the second term is simply $\mathbb{E}[\mathbf{1}_B(\mathcal{A}_1)]^2$, which by Equation 1 is $\Phi(\ell/\sigma\sqrt{\tau})^2$.

The first term of Equation 2 is the probability that both $\mathcal{A}_1$ and $\mathcal{A}_2$ are of ancestry B, which we can compute by considering the joint distribution of the movement of the two lineages over the last τ generations. At the time of sampling, and until the first recombination event between the two loci, the two lineages follow an identical path back through time. We assume that after the first recombination event the two lineages never coalesce back onto the same chromosome and therefore pursue independent Brownian paths for the remainder of the τ generations since secondary contact (Figure 1). This assumption ignores drift since secondary contact and therefore does not account for the possibility of the two lineages coming back onto the same background to once again assume identical paths. This is a reasonable assumption for large populations and recent contact zones where the probability of coalescing back onto the same background is small, but neglects some additional covariance in smaller populations.

This assumption of no drift will be good if $\sqrt{\tau}$ is much smaller than the number of individuals falling in a circle (or interval) of radius σ, proportional to Wright's neighborhood size N_σ (Wright 1946). This is because in one dimension, assuming Gaussian dispersal, the number of generations that two randomly moving lineages that start in the same place spend within distance σ of each other across τ generations is of order $\sqrt{\tau}$; the chance that they coalesce each time they are nearby is proportional to $1/N_\sigma$, and so the chance of coalescence is negligible if $\sqrt{\tau}/N_\sigma \ll 1$. Since coalescence is less likely in two dimensions than in one, this gives a bound in the two-dimensional case as well; for more discussion, see Nagylaki (1978) and Barton *et al.* (2002).

To find an expression for covariance in ancestry, observe that the random number of generations T since the most recent recombination event between the two loci is geometrically distributed; continuing with the continuous time model, we can take T to be exponentially distributed with rate parameter r. Given that the most recent recombination along this lineage occurred T generations ago, with $T < \tau$, the spatial displacements of the two lineages from which $\mathcal{A}_1$ and $\mathcal{A}_2$ derive at τ generations in the past are distributed as a bivariate Gaussian with covariance $T\sigma^2$ and variance $\tau\sigma^2$, the probability density of which we denote $f_T(x_1, x_2)$.

The probability that both lineages are to the right of zero τ generations ago, and hence are both of ancestry B, is therefore given by

$$\begin{aligned}\mathbb{E}\big[\mathbf{1}_B(\mathcal{A}_1)\mathbf{1}_B(\mathcal{A}_2)\big] &= e^{-r\tau}\Phi\left(\frac{\ell}{\sigma\sqrt{\tau}}\right)\\ &\quad + \int_0^\tau re^{-rt}\int_{-\ell}^{\infty}\int_{-\ell}^{\infty} f_t(x_1, x_2)\, dx_1 dx_2\, dt.\end{aligned} \tag{3}$$

The first term of Equation 3 corresponds to the probability that there has been no recombination for the last τ generations, multiplied by the probability that the path of our single ancestral lineage is on the right side of the barrier when the barrier was removed. The second term integrates the probability that two lineages that recombined t generations ago are both to the right of the barrier at time τ, *i.e.*, the bivariate Gaussian density integrated over the quadrant $x_1 > 0$ and $x_2 > 0$, over all possible times of first recombination. Rescaling t so that $u = t/\tau$, Equations 2 and 3 come together to give

$$\begin{aligned}&\operatorname{Cov}[\mathbf{1}_B(\mathcal{A}_1), \mathbf{1}_B(\mathcal{A}_2)]\\ &\quad = \int_0^1 e^{-ru\tau}\frac{1}{2\pi\sqrt{1-u^2}}\exp\left(-\frac{\ell^2}{\tau\sigma^2(1+u)}\right) du\\ &\quad =: D(r, \ell, \tau, \sigma).\end{aligned} \tag{4}$$

To obtain this expression, we integrate by parts, make use of the identity in Equation A3, and rescale (0, τ) onto (0, 1) (see *Appendix A* for more detail). We denote this covariance as a function $D(r, \ell, \tau, \sigma)$, which expresses the expected covariance in ancestries of two loci in a randomly sampled individual from a given geographic location (ℓ) as a function of recombination fraction (r) between the loci, time since admixture (τ), and rate of dispersal (σ). In *Appendix B* we also develop analogous results for arbitrary migration schemes in discretized space, for both continuous and discrete time.

These functional forms give us a way to relate observed patterns of LD in admixed populations to the parameters of the demographic model generating admixture. We later use this to develop an inference method to estimate these parameters in a contact zone. Before doing this, we explore strategies to obtain the full distribution of admixture block lengths in a contact zone model.

Admixture block lengths

An extension to the above approach for describing admixture LD between two loci is to consider how ancestry along the chromosome is partitioned into unbroken genomic tracts of ancestry drawn from one parental population. This is a natural way to think about coancestry in admixed populations (Fisher 1954; Barton 1983; Ungerer *et al.* 1998; Gravel 2012), and the genome-wide distribution of ancestry tract length will contain information about admixture and is a richer source of information than pairwise LD alone.

We again examine a chromosome drawn at random at geographic position ℓ, this time considering the probability that between physical positions P and Q, separated by genetic distance d, the chromosome inherits entirely from ancestry B. As above, we assume that once linkage is broken by recombination, the lineages from which the products of recombination are descended move independently of each other. This again assumes that τ is small relative to the timescale of coalescence (genetic drift). Further, it ignores the correlation structure imposed by the pedigree (Wakeley *et al.* 2012; Liang and Nielsen 2014a), the impact of which we return to in the *Discussion*.

We note that our measure of recombination rate d will differ from the earlier definition of recombination fraction (r) as we will be tracking all recombination events between P and Q. We now assume that recombination events occur as a Poisson process with rate d, which reflects genetic distance on the genetic map between our two endpoint loci, and assume no interference.

If there have been K recombination events that occurred along the tract of the chromosome over the last τ generations, then this region has $K+1$ genetic ancestors from τ generations ago. Denote the spatial locations at time τ of these ancestors $\mathbf{X} = (X_1,\cdots,X_{K+1})$. As our assumption of infinite population size neglects coalescence, these ancestors are assumed to be distinct. The segment contains only ancestry from population B if all $X_i > 0$ (*i.e.*, all $K+1$ ancestors lived at locations to the right of 0 at time τ; see Figure 2 for an example of $K=3$). We denote the probability of this segment containing only ancestry from population B as

$$U_d(\tau,\ell) = \mathbb{E}\left[\prod_{i=1}^{K} \mathbf{1}_B(\mathcal{A}_i)\right], \tag{5}$$

where, as before, $\mathcal{A}_i$ is B if $X_i > 0$. This expected value averages over both the number and timing of recombination events and the locations of the ancestral lineages at time τ ago. We now outline one approach to obtain an expression for $U_d(\tau,\ell)$ by conditioning on the number of recombination events, and give a complementary approach in *Appendix D*.

Since we assume no coalescence, the branching order of the ancestral lineages via recombination specifies a labeled tree structure with $K+1$ tips, given K recombinations, meaning that a modern individual at location ℓ has $K+1$ distinct ancestors from τ generations ago. Since, looking backward in time, each lineage moves as an independent Brownian motion once it has split from the others, $\mathbf{X}$ has a $(K+1)$-dimensional multivariate Gaussian distribution with mean $(\ell,\cdots,\ell)$ and variance–covariance matrix $\boldsymbol{\Sigma}$. The entries $\boldsymbol{\Sigma}_{i,j}$ are determined by the amount of time that the lineages leading to tips i and j spend in linkage. If we let $t_{i,j}$ be the time, in generations, from the present to the recombination that separates tip i from tip j, then $\boldsymbol{\Sigma}_{i,j} = \sigma^2 t_{i,j}$, and the diagonal entries are $\boldsymbol{\Sigma}_{i,i} = \sigma^2\tau$.

Conditioning on $K=k$ recombinations and $\boldsymbol{\Sigma}$, the probability that all $k+1$ tips are of ancestry B is given by the integral of the $k+1$-dimensional Gaussian density over the space for which all $X_i > 0$:

$$U(\tau,\ell|\boldsymbol{\Sigma}) = \int_{-\ell}^{\infty}\cdots\int_{-\ell}^{\infty} \frac{\exp\left(-(1/2)\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} dx_1\cdots dx_{k+1}. \tag{6}$$

The above integrand is the density for the multivariate Gaussian whose covariance matrix is determined by the timing and ordering along the chromosome of recombination events. As an example, Figure 2 presents the two different unlabeled topologies that can be obtained for $K=3$. The topology of Figure 2A would produce a multivariate Gaussian with covariance matrix

$$\sigma^2\begin{bmatrix} \tau & t_1 & t_1 & t_1 \\ t_1 & \tau & t_2 & t_2 \\ t_1 & t_2 & \tau & t_3 \\ t_1 & t_2 & t_3 & \tau \end{bmatrix}$$

and the topology of Figure 2B would be represented by the covariance matrix

$$\sigma^2\begin{bmatrix} \tau & t_2 & t_1 & t_1 \\ t_2 & \tau & t_1 & t_1 \\ t_1 & t_1 & \tau & t_3 \\ t_1 & t_1 & t_3 & \tau \end{bmatrix}.$$

Obtaining the unconditioned value of $U(\tau,\ell)$ from the conditioned version in Equation 6 requires averaging over possible trees; to do this we sum over possible tree topologies and for each tree topology integrate over possible split times (*i.e.*, t_1, t_2, and t_3 in the case of the three-recombination trees shown in Figure 2).

For a given unlabeled tree topology $\mathcal{T}$, we therefore need to integrate Equation 6 over the possible split times of the

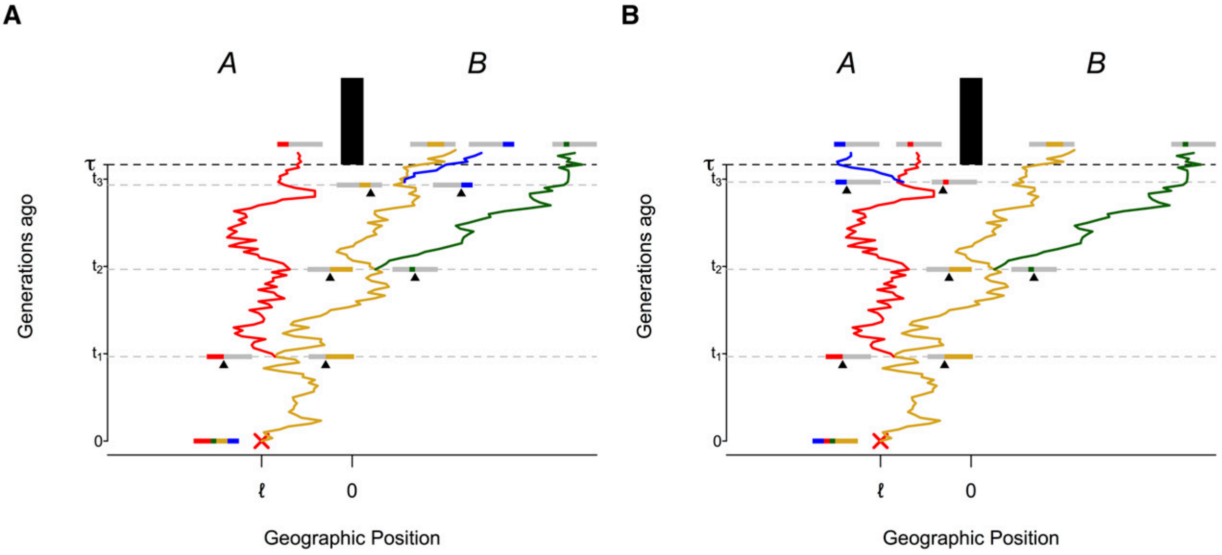


Figure 2 Brownian motion paths of a tract of chromosome. As in Figure 1B, the paths along chromosomal fragments are identical until recombination breaks the fragments up. Here, the position of each chromosome fragment at time τ is shown. For the entire portion of chromosome to be of uniform ancestry, all products of recombination must be on the same side of the barrier to dispersal at time τ. In A and B, both tracts of chromosome have experienced three recombination events at times t_1, t_2, and t_3, but have different topologies due to the different orderings of the events along the chromosome: (t_1, t_2, t_3) in A and (t_1, t_3, t_2) in B. This results in different covariance matrices for the positions of fragments, as described in the main text. The yellow, blue, and green fragments in A constitute an unbroken tract of *B* ancestry, and in B, the yellow and green fragments make up an unbroken tract of *B* ancestry.

tree. First note that we can reduce to the case of trees with height 1 by scaling time. We define $\Sigma' = \Sigma/(\sigma^2\tau)$ so that by Gaussian scaling, $U(\tau,\ell|\Sigma) = U(1,\ell/(\sigma\sqrt{\tau})|\Sigma')$.

We then obtain the following term:

$$U(\tau,\ell|\mathcal{T}) = \int_{\mathbf{t}'} U\left(1,\frac{\ell}{\sigma\sqrt{\tau}}\middle|\Sigma'\right) V\left(t_1',\cdots t'_{k+1}\right) dt_1'\cdots dt'_{k+1}. \tag{7}$$

Here, $V(t_1',\cdots t'_{k+1})$ is the probability of the branching times given the topology. V accounts for the fact that the $\mathbf{\Sigma}$ we integrate over represents a topology $\mathcal{T}$ and so the possible entries of $\mathbf{\Sigma}$ are constrained by $\mathcal{T}$. Thus, the set of possible times, $\mathbf{t}'$, over which we integrate depends on the tree topology, and correspondingly, each topology has a probability given k recombinations. [See *Appendix C* for a further description of $\mathbf{t}'$ and $V(\mathbf{t}')$.]

Finally, we sum across k and, for each k, all $k+1$-tipped unlabeled topologies. Recalling that the probability of the number of recombination events is Poisson distributed,

$$U_d(\tau,\ell) = \sum_{k=0}^{\infty} \frac{(d\tau)^k e^{-d\tau}}{k!} \sum_{\mathcal{T}_i^k \in \mathcal{T}^k} \Pr\left(\mathcal{T}_i^k\right) U_r\left(\tau,\ell\middle|\mathcal{T}_i^k\right), \tag{8}$$

where $\Pr(\mathcal{T}_i^k)$ is the probability of the ith unlabeled topology given that there are $k+1$ tips [we describe the calculation of $\Pr(\mathcal{T}_i^k)$ in *Appendix C*]. We note that Equation 8 is a Wild sum expansion for $U_d(\tau,\ell)$ (Etheridge 2000). We outline an approach using differential equations to obtain an equivalent expression in *Appendix D*.

In practice, we can approximate this sum by conditioning on k^* or fewer recombination events in τ generations:

$$\begin{aligned} U_d^{k^*}(\tau,\ell) &= \frac{1}{\Pr(k \le k^*|d\tau)} \sum_{k=0}^{k^*} \frac{(d\tau)^k e^{-d\tau}}{k!} \\ &\quad \times \sum_{\mathcal{T}_i^k \in \mathcal{T}^k} \Pr\left(\mathcal{T}_i^k\right) U_r\left(\tau,\ell\middle|\mathcal{T}_i^k\right). \end{aligned} \tag{9}$$

In the *Results* section below, we briefly explore the convergence of this sum to distributions obtained by simulation. Summing over the large number of topologies for large k^* is computationally expensive, but terms in the sum can be reused over some parameter values. Given reliable measurements of block-length distributions from genomic data, the above estimate of U provides a means by which timing and migration in contact zones can be inferred (see Gravel 2012 for a recent application of such an approach). However, we do not implement this strategy here, concentrating instead on fitting models to admixture LD.

Simulations

We developed two classes of simulations to (1) evaluate the accuracy of our analytic results and (2) explore the consequences of realistic violations of our model that likely occur under the specified biological process. Specifically, we are concerned with the assumption that movement of alleles is independent following recombination that follows from the assumption of infinite population size as well as the assumption of continuous time, rather than discrete generations.

For the first class of simulations, *simulations under the model*, we consider chromosomes moving in continuous space and time, with recombination modeled as a Poisson process through continuous time and independent movement of all products of recombination. Chromosomes were simulated by generating a vector of recombination times under a Poisson process and then uniformly placing the recombination events on a chromosome. Geographic positions through time were assigned to chromosomal segments at each recombination event such that adjacent segments followed the same path until the time at which a recombination event occurred between them, after which they followed independent paths. The positions were assigned by drawing the geographic displacement that occurred between recombination events from a Gaussian distribution with mean zero and variance equal to σ^2 multiplied by the amount of elapsed time. This is an explicit simulation of the model described above. We simulated 10,000 chromosomes under the model.

The second class of simulations, *simulations under the process*, are forward-time, discrete-generation simulations of a grid of discrete populations in continuous time. In these simulations we record the complete recombination history of each chromosome. As such simulations allow genetic drift, enforce a pedigree structure onto local ancestry, and occur in discrete time and space, these simulations under the process present a biologically realistic challenge to many of our major modeling assumptions. We consider 200,000 diploids (400,000 chromosomes) evenly spread across 20 demes. Demes are connected through nearest-neighbor migration with a per-generation, per-individual probability m of migration (this migration rate is reduced to $m/2$ on demes at the edges of one-dimensional space). We sample the number of recombination events each generation from a Poisson distribution with mean of one, corresponding to a 1 Morgan chromosome, and recombination events are uniformly placed along a chromosome (*i.e.*, no recombinational interference). Every generation, migration, random mating, and recombination take place, and we record for each piece of chromosome the population from which it was inherited (*i.e.*, its ancestry). After τ generations we sample chromosomes and assign ancestry along each individual's chromosome based on whether ancestors originated in populations 1–10 (ancestry B) or in populations 11–20 (ancestry A).

Inference of parameters in human admixture data

We now use our theory to infer parameters in a demographic model, using real data. To do this, we can use either ancestry LD (Equation 4) or ancestral block length distributions (derivable from Equation 5). While the distribution of continuous-ancestry tracts necessarily contains more information than LD alone, there are limits to the precision of the measurement of tract length over short recombination distances (which would reflect old events). This, combined with the relative ease of obtaining LD measurements from genomic data, motivates the use of LD in our analysis of human admixture contact zones. A variety of methods, including ALDER (Loh *et al.* 2013) and Globetrotter (Hellenthal *et al.* 2014), estimate some measure of admixture LD that is an estimate of ancestry LD. We use the weighted LD curves generated by ALDER, which computes the statistic

$$a(r) = \frac{1}{|S(r)|} \sum_{(M,N)\in S(r)} \widehat{\mathrm{Cov}}(M,N)(p_A(M) - p_B(M)) \times (p_A(N) - p_B(N)), \quad (10)$$

where the sum is over a set of pairs of autosomal loci, $S(r)$, each of which is r apart (in practice, this method uses d in place of r, and for analysis using ALDER output, we do the same). After Loh *et al.* (2013), (M,N) is a pair of loci, $p_A(\cdot)$ and $p_B(\cdot)$ are sample allele frequencies in the parental populations A and B, and $\widehat{\mathrm{Cov}}(M,N)$ is the sample covariance between alleles at the two loci within the target population. If r is large enough that background LD in the ancestral populations can be ignored, and the allele frequencies in the parental populations are known, then $\mathbb{E}[a(r)] = 2\alpha(1-\alpha)F_2(A;B)^2\mathrm{Cov}(\mathbf{1}_B(\mathcal{A}_1), \mathbf{1}_B(\mathcal{A}_2)|r)$, where $\mathrm{Cov}(\mathbf{1}_B(\mathcal{A}_1), \mathbf{1}_B(\mathcal{A}_2)|r)$ is the expected covariance in ancestry between pairs of loci a recombination fraction r apart, α is the ancestry proportion of population A in the admixed population, and the constant $F_2(A;B)^2$ measures differentiation of allele frequencies between the two parental populations. Often, the true parental populations no longer exist, or are not sampled, and instead proxy parental populations are designated. In these cases, $F_2(A;B)^2$ is a measure of the shared differentiation between the true parental populations and the proxy populations (Loh *et al.* 2013).

Admixture at a single time point: Under a basic model of admixture, decay in ancestry LD can be described by the parameters F, t, and G in the exponential model

$$\mathbb{E}[a(r)] = Fe^{-rt} + G \quad (11)$$

corresponding to a single pulse of admixture t generations ago. This reflects the exponential decay of admixture LD with time due to recombination between two loci separated by recombination fraction r and is the model used by ALDER (Loh *et al.* 2013) and Globetrotter (Hellenthal *et al.* 2014) to estimate admixture timing in a single population.

The term G represents LD between unlinked markers due to substructure in the sampled individuals with respect to their ancestry proportions. Under the model of Loh *et al.* (2013), the value $F + G$ corresponds to $F_2(A;B)^2(2\alpha(1-\alpha) + 2\mathrm{Var}(\alpha))$, where α is the admixture proportion and therefore is a compound parameter reflecting both admixture proportion and differentiation between parental populations (Loh *et al.* 2013). This is the expected variance in allele frequency at a single locus, which is a function of the differences in allele frequency between the parental populations, the proportion of ancestry from each parental population, and the covariance that arises from nonrandom mating with respect to ancestry in the admixed population.

Fitting to a geographic contact zone: Under our model, we take a set of admixed samples drawn from n populations, who fall at positions $\ell_1, \cdots, \ell_n$ along a linear geographic transect. The geographic location of the center of the zone along this transect is C, such that sample 1 is a distance $\ell_1 - C$ from the zone. We specify a pair of proxy parental populations A and B to represent the end points of the contact zone. Using ALDER, we generate the statistic $a_j(r_i)$ for the jth population sample for each genetic distance bin (i), giving us a set, $\mathbf{a}$, of weighted-LD decay curves (as defined in Equation 10). We use the minimum inter-SNP distance determined by ALDER based on LD in the parental populations.

To assess the uncertainty in $\mathbf{a}$, we estimate the variance in ALDER's statistics, using the jackknife (which is an output of ALDER). For each of the $c = 22$ iterations, one chromosome is removed before recalculating $\mathbf{a}$ for the remaining 21 chromosomes. We use this to calculate the variance $V_{i,j} = \mathrm{Var}\left(a_j(r_i)\right)\left((c-1)/c\right)$. We then conduct a least-squares fit of the ALDER output to our prediction given by Equation 4 for values of τ, σ, $\mathcal{F}$ [which accounts for differentiation between the parental populations in the same way that $F_2(A;B)$ does], and C. We fit all n populations simultaneously, minimizing

$$\mathcal{L}(\mathbf{a};\tau,\sigma,C,\mathcal{F}) = \sum_{i=1}^{n}\sum_{j}\frac{1}{V_{i,j}}(a_i(r_j) - D(r_j,\ell_i - C,\tau,\sigma)\mathcal{F})^2. \tag{12}$$

Our choice of $\mathcal{L}()$ would be the negative log-likelihood of our parameters if our $a_j(r_i)$ were Gaussian [a reasonable approximation given the large number of pairs of markers contributing to each value of $a_i(r_i)$] and independent. We refer to $\mathcal{L}()$ as the log-likelihood, and because we are mainly interested in τ and σ, we generate profile surfaces of $\mathcal{L}$ across combinations of τ and σ. Specifically, we set a value for C based on a fit of Equation 1 to ancestry proportion and generate a likelihood surface over a grid of $\tau \times \sigma \times \mathcal{F}$ and for each combination of τ and σ we define the profile log-likelihood as the maximum log-likelihood across all of our corresponding $\mathcal{F}$ grid points. The grid of $\mathcal{F}$ values that we fit over is informed by the strength of differentiation between the parental populations.

We note that, although Equation 11 includes an affine term to account for LD that could be generated by an unspecified model of population substructure, our model does not. This is because a source of long-range LD is incorporated into our model via gene flow from neighboring populations with different admixture proportions.

Data availability

Using Equation 12, we fit our model to genomic data from populations that potentially represent admixture contact zones. These data were obtained from previously published studies, as detailed in Table S1, and are available publicly or upon request from their respective authors.

Custom scripts for simulations and fitting of data are available as Supporting Information, File S2 and File S3, and at https://github.com/asedghifar/NeutralZones.

Results

Simulation results and comparison to exponential model

Figure 3 shows the decay in LD at various points in time and space and shows the exact correspondence between the analytic expression of Equation 4 and the output of simulations under the model. As expected, the rate of decay increases with age, and LD is greatest at the center of the zone. While LD decays from a higher point in populations closer to the center of the zone, the rate of decay is greater in populations farther from the center of the zone. Dispersal is measured in the same units as distance from the zone center, and so the impact of dispersal on curves can be measured simply by rescaling the distance parameter.

To evaluate the consequences of fitting a single-pulse model to data generated by our spatial model of continuous admixture, we fitted the exponential decay of Equation 11 to a set of simulated populations from a 50-generation-old contact zone under the model. The comparison, shown in Supporting Information, Figure S1, of best-fit parameters indicates that the simple exponential tends to underestimate the age of the admixed population, presumably because of the continuous introduction of migrants bearing long ancestral haplotypes. In other words, the poor fit of the single-pulse model to these LD decay curves, especially close to the center of the contact zone, is due to the heterogeneous mixture of recombination times. Consistent with this interpretation, the effect diminishes in populations far from the center of the zone, as the difference in ancestry composition between neighboring populations decreases as the distance to the center increases.

To demonstrate our inference method as described above, we fitted our model (Equation 4) to the curves generated under the process. Because we simulated single chromosomes, we could not use the jackknife estimator of variance and therefore modified Equation 12 by removing the denominator. We removed populations with no detectable admixture from the fit, limiting our analysis to populations close to the center of the contact zone. The profile likelihoods of these surfaces are shown in Figure 4. The maximum-likelihood estimates of τ and σ are $(2, 0.17)$, $(38, 0.12)$, and $(93, 0.11)$ for zones simulated under $\tau = 5, \tau = 50$, and $\tau = 100$, respectively, all with $\sigma = 0.1$. We further explored the performance of our inference framework under different values of σ and τ (Figure S2). The method generally performs well. The accuracy of inference decreases with smaller values of σ and τ, presumably due to the increased noise from the small number of migrants and recombination events and perhaps also due to increased discrepancies between the discrete time and space simulations compared to the continuous time model.

Compared to the true values we use to simulate under the process our inference method tends to slightly underestimate

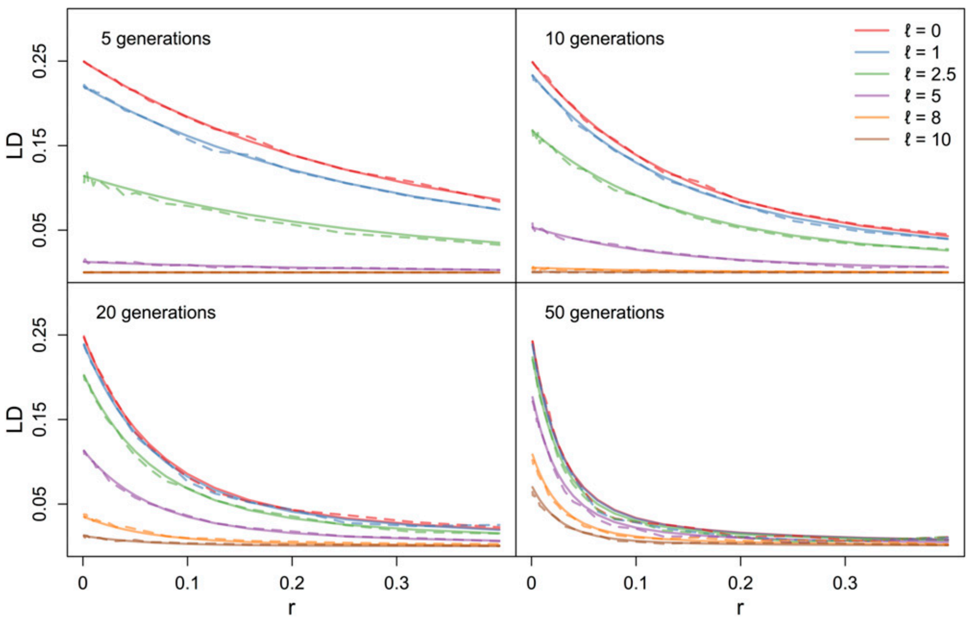


Figure 3 LD decay curves for populations of increasing distance ℓ from the zone center and increasing age of contact zone. Solid lines represent analytic predictions and dotted lines represent the output of simulations under the model as described in *Materials and Methods*.

the age of the contact zone. We expect that this is in part due to the discrete nature of the simulation. These estimates are closer to the true simulated ages than those obtained using the same method to fit an exponential (Equation 11), which estimated $1.9 < \hat{\tau} < 4.2$ for $\tau = 5$, estimated $20.4 < \hat{\tau} < 25.6$ for $\tau = 50$, and estimated $40.0 < \hat{\tau} < 59.5$ for $\tau = 100$ (compare to the results of our method on the same data above).

Finally, we compared our estimates of continuous length distribution to simulated tract lengths under the model. Figure 5 shows the distribution of tract lengths in simulated populations along contact zones of two different ages as well as the convergence of the sum in Equation 9 as k^* (the maximum number of recombinations) is increased. The approximation using k^* works best for young contact zones and for the distribution for short tract lengths. For older zones and longer tracts, summing over k^* is computationally intensive, and the numerical approximations using the differential equation approach (*Appendix D*) may be more tractable.

Application to human data sets

We applied our model to three independent sets of populations that potentially represent admixture in a spatial context: populations along the Indonesian archipelago and New Guinea, populations in Central Asia, and populations in India (Table S1). Genetic distances between SNPs were inferred using sex-averaged recombination rates from deCODE (Kong *et al.* 2010). Each of these data sets has been previously analyzed for admixture times, using the single-pulse admixture model (Xu *et al.* 2012; Moorjani *et al.* 2013; Hellenthal *et al.* 2014), and our aim was to compare results and goodness-of-fit of our geographic admixture model to the single-pulse ("exponential") model.

Indonesian archipelago: Populations along the Indonesian archipelago and New Guinea show a longitudinal cline of admixture between East Asian and Papuan autosomal ancestry (HUGO Pan-Asian SNP Consortium 2009; Xu *et al.* 2012; Lipson *et al.* 2014). The decrease in proportion of Asian ancestry with longitude has been interpreted as evidence of the Austronesian expansion from the west through Indonesia. Xu *et al.* (2012) fitted simple admixture models independently to each of the populations to infer admixture times of 120–200 generations, with populations with higher levels of Papuan ancestry having more recent admixture times. A more recent analysis using ALDER estimated single admixture dates for populations in the region in the range of 30–60 generations, suggesting that this in part is the result of subsequent waves of gene flow from populations with varying levels of Asian ancestry (Lipson *et al.* 2014).

We obtained the genotypes for seven population samples in Indonesia (shown in Figure 6 and Table S1) from the HUGO Pan-Asian SNP Consortium (2009) and a Papuan population from the Human Genome Diversity Project (HGDP) (Li *et al.* 2008). The combined data set yielded 17,057 shared SNPs. We first ran STRUCTURE (Pritchard *et al.* 2000) with $k = 2$ on these nine samples. The admixture proportions obtained from STRUCTURE confirm the east-to-west cline (shown in Figure 6). We then used least squares to fit Equation 1 to these admixture proportions, which estimated the cline center at $C = 124°9'\mathrm{E}$ and $\sigma^2\tau = 50.9$. Based on ancestry proportions, we chose the Mentawai population and the Papua New Guinean population as proxy source populations to generate ALDER curves. Simultaneously fitting our model to the six admixed populations, we generated the profile-log-likelihood surface shown in Figure 6. The parameters that minimized Equation 12 were an approximate contact time of ~200 generations ago [or 5800 years, given a generation time of 29 years (Fenner 2005)], $\sigma = 0.63°$ per generation (~66 km per generation), and $F = 0.0045$. Our estimate of σ seems reasonable; using differences in estimates of

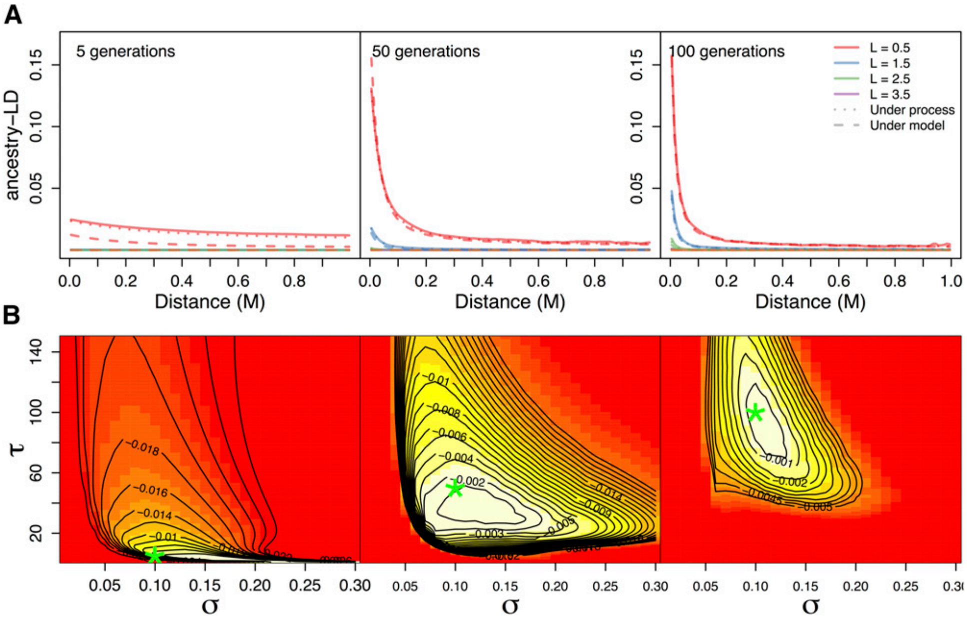


Figure 4 Analysis of simulations under the process run with parameters $\tau = 5$, $\tau = 50$, or $\tau = 100$ and $m = 0.01$ under nearest-neighbor migration, corresponding to $\sigma^2 = 0.01$ in the continuous model. (A) Output of simulations (solid lines), compared to the continuous time and space model of Equation 4 (dashed lines) and a discrete time and space expression from Equation B2 (dotted lines). (B) Profile-likelihood surfaces describing the fit of our continuous model to simulations under the process. Green asterisks indicate simulated values.

admixture dates in each of these populations, Xu *et al.* (2012) propose a mean dispersal of 22.5 km per generation in these populations. The fit to LD decay curves under these estimates is shown in Figure 6 and Figure S3.

We also explored the fit to LD decay curves of the single-pulse model, fitting Equation 11 by least squares (weighted by jackknife variance as in Equation 12). Unsurprisingly, the fit of our model is not as good as that of a model in which all admixed populations are considered as having a single admixture time but allowed different values of F ($\mathcal{L} = 100,370$ compared to $\mathcal{L} = 94,147$) since independently fitting the y intercept to each population allows for many more parameters while these intercepts in our model are constrained by geographic distances between the populations. The fits to each population are presented in Table S2 and are in good accordance with those found by Lipson *et al.* (2014), using similar methods. With this approach, the mean timing among the admixed populations is 60.8 generations (we ignore the Javanese population that has little admixture and an estimated admixture time of 665 generations as this is far older than all the other populations).

Additionally, we considered fitting all populations simultaneously for a single time under the exponential model (Equation 11), allowing each population to choose its own F parameter to account for differences in admixture proportions. Under this model we obtain an estimated age of $\tau \approx 63$ generations ($\mathcal{L} = 98,706$). Given that the truth is likely more complex than both the exponential and contact zone models, this better fit is not surprising given that we are allowing each population to fit its own intercept.

Linguistic evidence suggests that the Austronesian expansion through Indonesia dates to ~4000 years ago (Gray *et al.* 2009). As noted by Lipson *et al.* (2014), these single-pulse dates (Table S2) are too recent to reflect this, consistent with our earlier observation that admixture times may be underestimated by a simple exponential model if admixture has been ongoing. Our estimate of timing based on fitting a geographic contact zone (5800 years ago) is much older than dates estimated by single-pulse models, but is also considerably older than the Austronesian expansion. Considering that it is constrained by having to fit all populations simultaneously, our model provides a good fit to these genomic data. One possible explanation for our overestimate of admixture time is the assumption of a continuous rate of diffusion after initial contact. Despite this, our model may be a more realistic depiction of ongoing gene flow than a single-pulse model.

India: Population structure in India is complex and multilayered. While the precise history of human movement in this region is unclear, work by Moorjani *et al.* (2013) and Reich *et al.* (2009) suggests that many modern Indian populations are descendants of an admixture event between differentiated ancestral North Indian (ANI) and ancestral South Indian (ASI) populations, with a cline in the extent of ANI ancestry from north to south across the subcontinent, shown in Figure 7. While it is difficult to identify modern proxies of the parental populations, the ANI population appears to be most closely related to Western Eurasian populations (such as Georgia) and the Onge population of the Andaman Islands seems to draw much of its ancestry from the ASI population. Moorjani *et al.* (2013) broadly grouped their samples into Indo-European or Dravidian samples and, under this classification, found that the decay in ancestry LD in their samples was consistent with two historical admixture events, the first ~108 generations ago, giving rise to the Dravidian populations, and a second wave of admixture from the north taking place 36 generations later that contributed to the ancestry of Indo-European populations.

We obtained the genomic data used in Moorjani *et al.* (2013) (from Li *et al.* 2008, Reich *et al.* 2009, Metspalu

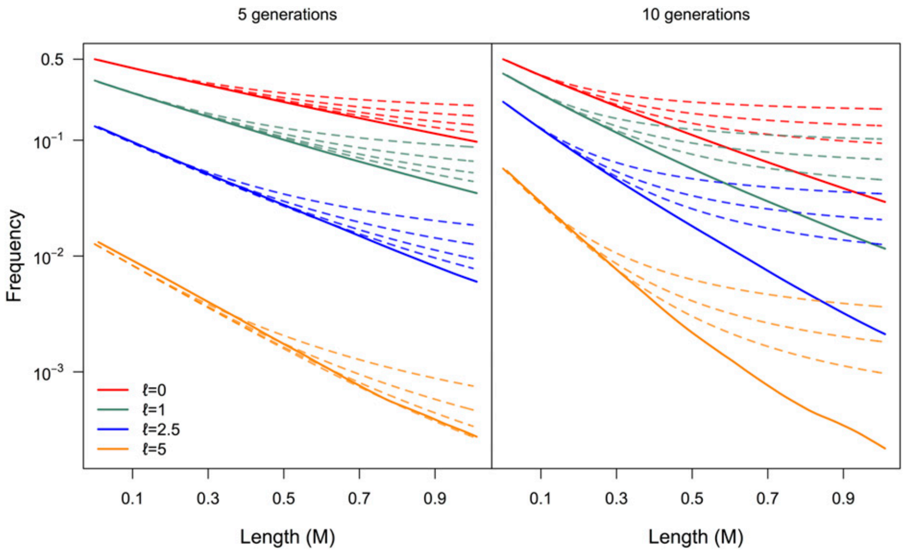


Figure 5 Distribution of tract lengths, expressed as the frequency of tracts that are at least a given length (*i.e.*, the complementary cumulative distribution of tract lengths). The following shows the distribution for populations *L* units away from the center of a contact zone. The solid lines represent the output of a simulated contact zone with no drift (simulated under the model, described in *Materials and Methods*). For the 5-generation contact zone the four dotted lines per geographic position represent the predicted distribution under approximations conditioning on at most three, four, five, or six recombination events. For the 10-generation contact zone, the three dotted lines represent approximations conditioning on at most three, four, or five recombination events.

et al. 2011, and Moorjani *et al.* 2013), yielding $\sim 83,000$ shared SNPs, and used only the populations represented in table 1 of Moorjani *et al.* (2013) (see Table S1). We then fit our model to LD curves generated by ALDER, as described above. Fitting a model in which all Indo-European and Dravidian populations are the outcome of a single admixture contact zone yielded an age of ~220 generations since contact (Figure 7, Figure S4). (Additional details of the fitting procedure for the Indian populations can be found in File S1.)

Several aspects of the data indicate potential misestimation of dates. Some populations, presumably the oldest, have very little admixture LD, which may prevent an accurate fit to the decay. Second, it is possible that the sampling of populations in space does not span a broad enough distance to obtain an accurate fit. Substructure within populations, due to practices such as endogamy, may also influence ancestry LD within a population and cause a deviation from expectations under a null model of locally random mating. We take these challenges, and the uncertainty in our results, as an indication that the complicated demographic histories of these populations are poorly described by a simplistic model of the sort we consider here. These challenges also likely apply to other analyses of these data, and caution is warranted in judging the age of this zone.

Central Asia: Populations in central Eurasia show varying levels of East Asian ancestry. In a global analysis, Hellenthal *et al.* (2014) identified a signal of admixture, using Mongolians and Iranians as proxy source samples, in Turkish, Uzbek, Hazara, and Uygur samples. The proportion of Mongolian ancestry decreases with longitudinal distance from Mongolia, with the Turkish populations harboring the lowest proportion of Mongolian ancestry. The estimated admixture dates for these populations of 20–30 generations in the past found by Hellenthal *et al.* (2014) are consistent with the timing of the westward military movement of Mongolians during the 13th century.

We took the genomic data for the four admixed populations and the two proxy source populations from the data set of Hellenthal *et al.* (2014) (~500,000 SNPs). A STRUCTURE analysis of these populations, with $k = 2$, is consistent with a gradient in Mongolian ancestry across Central Asia (Figure 8). We used ALDER to generate weighted covariance curves, using the Mongolian and Iranian samples as the two proxy source populations. For the four admixed populations, the best fit under our simple contact zone model is ~49 generations or 1421 years ago (29 years per generation), with $\sigma = (3.7° \approx 300\ \text{km})$ per generation (see Figure 8 for the profile-likelihood surface, computed over 20 values of $\mathcal{F}$ between 0.001 and 0.01). This admixture date predates the Mongolian invasion of Central Asia that took place ~800 years ago. However, it is known that human movement in Central Asia was complex and preceded the Mongolian invasions by centuries, and it is possible that our estimated date is capturing a signal of these earlier migrations. This is supported by recent analyses of Central Asian populations by Yunusbayev *et al.* (2015). Our estimated parameters under the exponential model can be found in Table S3.

ALDER identified extensive long-range LD in the Hazara population, possibly due to population substructure within this sample with respect to Mongolian ancestry. Because this could potentially influence our inference, we refitted the LD curves to the set of admixed populations, excluding the Hazara. This produced an estimate of 37 generations (see Figure S6 and Figure S7 for LD curves in Central Asian populations).

One consideration in our applications is our assumption that the populations spread back into contact and then simply passively diffused into each other. This is obviously likely a poor description of the movement of Mongolian genotypes across Asia during the 13th century invasions, which could result in a discrepancy between expected and predicted decay in ancestry LD. We therefore proposed an alternate model that allows for an initial fast pulse of Mongolian migration into central Asia, followed by diffusion through local geographic

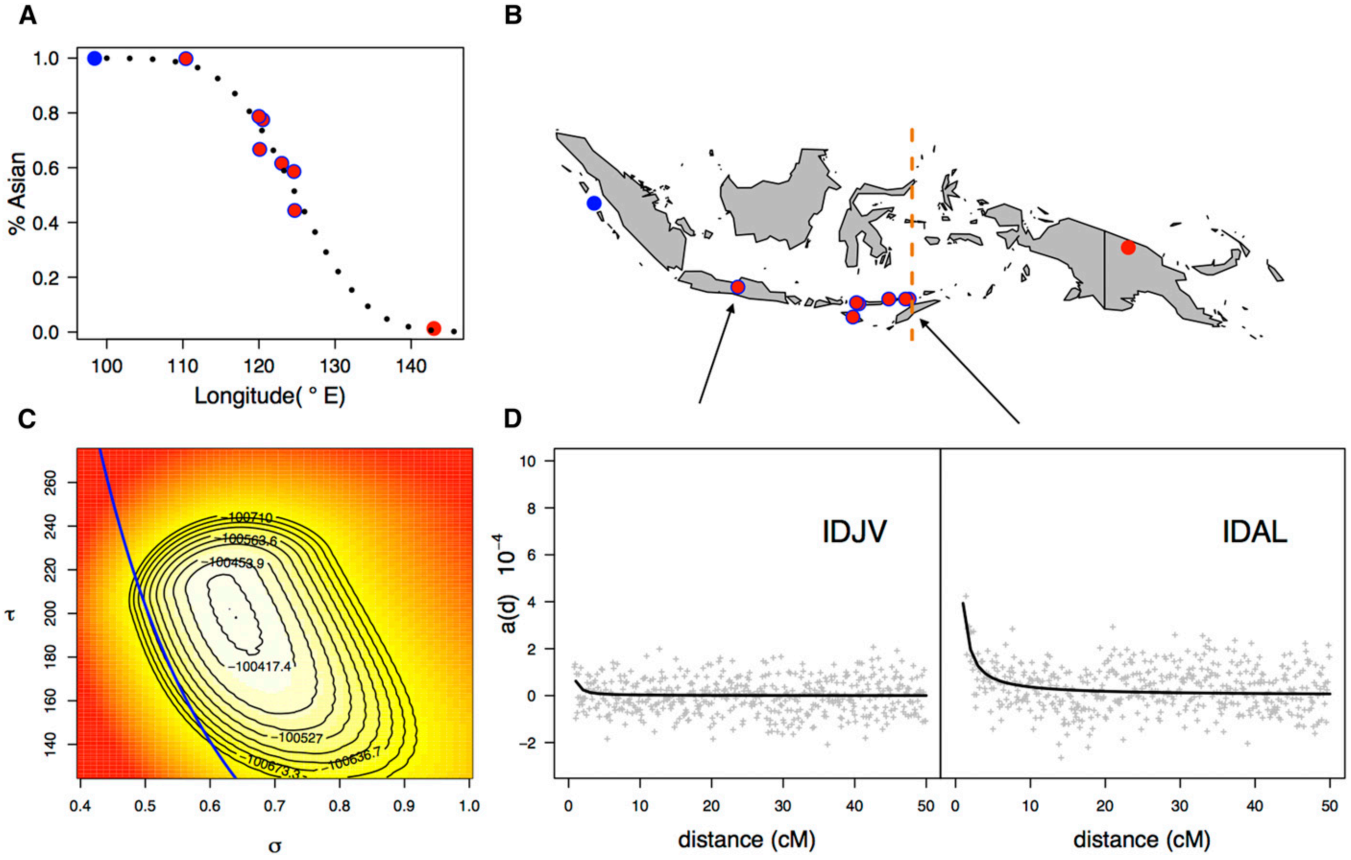


Figure 6 (A) Longitudinal cline in Asian ancestry. Black dotted line shows best fit to Equation 1. (B) Sampling locations of Indonesian populations. Blue circle denotes the representative Asian ancestral population and red circle the representative Papuan population. Vertical yellow line shows location of the inferred cline center. (C) Profile likelihood surface (over a grid of 30 $\mathcal{F}$ values between 0.002 and 0.005) for τ and σ under Equation 12 for all admixed Indonesian populations. The blue line represents the curve $50.9 = \sigma^2\tau$, corresponding to the value of this compound parameter that is obtained by fitting to admixture proportions alone as shown in A. (D) Weighted-LD curves for two populations, Java (IDJV) and Alorese (IDAL), that are different distances away from the center of the cline. Gray points represents estimates of LD generated by ALDER, and black curves are expected LD under the estimated parameters.

dispersal (*i.e.*, our Brownian motion). Explicitly, we construct a model that defines two new parameters: C_1, a point in space to the east of which some proportion, ψ, of the population is replaced by Mongolian genotypes τ generations ago (see *Appendix E* for mathematical details). In specifying this model, we are trying to capture a scenario in which, at least initially, unadmixed Mongolian genotypes rapidly spread westward. However, we acknowledge that this is at best a very crude approximation to the true history. Note that while C_1 is analogous to the contact zone center estimated in the original model, which was estimated using admixture proportion, when fitting the weighted-LD curves to the modified model, we are fitting two additional parameters, C_1 and ψ.

While this alternate model provides a better fit to admixture proportions (Figure 8 shows the fit with $\psi = 0.55$ and $C_1 = 62.7$), given the few populations, this good fit may reflect overparameterization of the model. Furthermore, a search for the best fit to the LD decay curves returned parameters that were effectively identical to the initial basic model proposed ($\psi \approx 1$, cline center $\sim$71°E), indicating that this is not a likely alternative model (profile-likelihood curves for each fitted parameter are shown in Figure S8). Given the early estimated admixture date, it is possible that admixture across Central Asia is not a product of a single event as our models, and those of others (Hellenthal *et al.* 2014), assume, but rather a result of complex human migrations throughout time. Despite the limitations imposed on inference of parameters by the small number of populations, broad patterns of ancestry LD across space are nevertheless somewhat consistent with our proposed model of ancestry LD decay across space along an admixture gradient.

Discussion

The generation and subsequent decay of admixture LD as an outcome of interbreeding between differentiated populations leaves a population genetic signature that is a valuable tool for understanding the nature and timing of admixture. Existing methods for modeling decay in admixture LD consider the expected rate of decay in one population at a time and often assume a simple one-time "pulse" of admixture without subsequent gene flow from neighboring admixed populations.

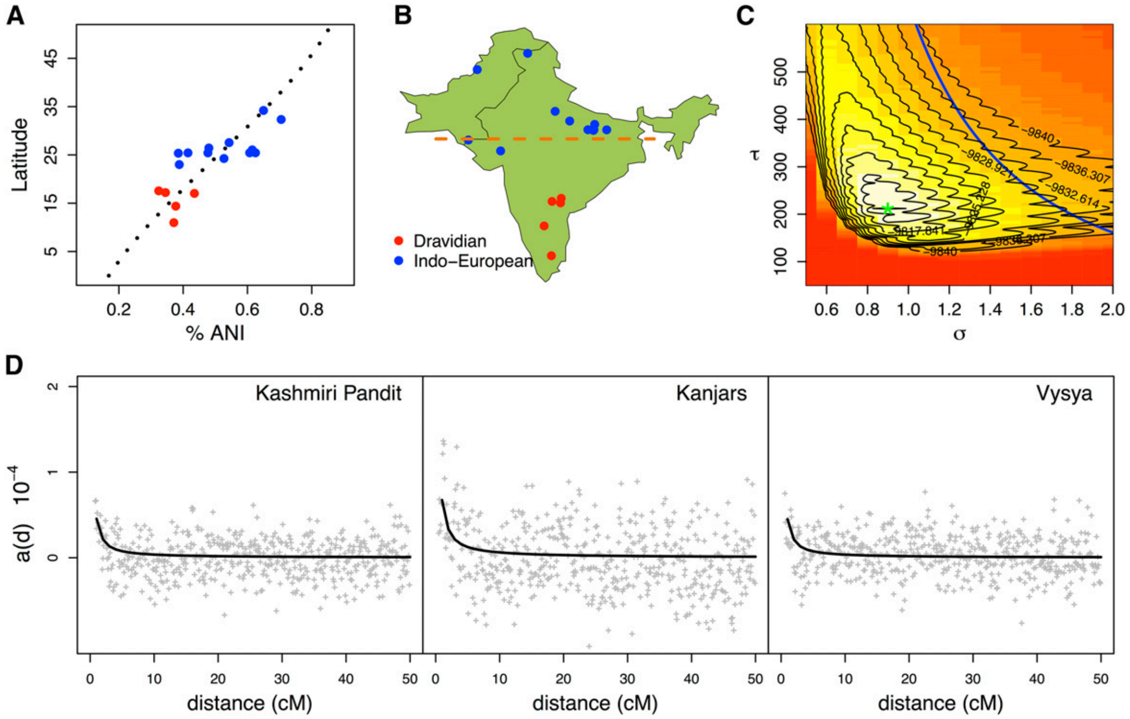


Figure 7 (A) Latitudinal cline in ANI ancestry. (B) Locations of Indian populations used in the analysis. Yellow line indicates location of inferred cline center. (C) Profile-likelihood surface for τ and σ under Equation 12. Blue line represents the relationship $\sigma\sqrt{\tau} = 25.4$, as obtained from the cline in ancestry proportion. Asterisk denotes values providing best fit. (D) Weighted LD curves estimated by ALDER, for northwest (Kashmiri Pandit), southern (Vysya), and northeast (Kanjars) populations. Gray points are estimates generated by ALDER, and black curves are expected LD under the estimated parameters.

Here, we have described a neutral model under which individuals diffuse across space after secondary contact. Based on this model, we derive an analytic expression for the expected decay in ancestry LD as a function of time since contact and a population's position in space. We consider this an alternate model to one in which admixed populations are independently formed by a single-pulse event with potential subsequent gene flow from parental populations. This model is suitable for recent secondary contact, when the genomic signatures of admixture are detectable and the timescale of admixture is smaller than that of drift.

In contrast to previous analyses of spatial admixture that treated populations as independent admixture events (*e.g.*, Xu *et al.* 2012), we consider data from all sampled populations simultaneously to build a model that incorporates all available information and accounts for the movement of individuals between populations. Compared to the expression for ancestry LD derived here, a simple exponential model tends to underestimate the time since admixture, as it does not account for the introduction of long ancestral haplotypes from neighboring populations.

Additional sources of covariance

In developing tractable approximations to spatial admixture contact zones we have ignored genetic drift and the genealogical structure imposed by the pedigree.

Genetic drift is not problematic if population densities, and dispersal rates, are high enough that coalescence between geographically close lineages is unlikely over the time since coalescence (as is likely the case in our human applications). Otherwise, a theoretical approach incorporating coalescence will be needed (see Barton *et al.* 2013 for recent progress). However, in that case, background LD and admixture LD will be on comparable genomic scales, making the job of separating the two much more challenging.

The other form of correlation structure that we have ignored is that imposed by the genealogy (Wakeley *et al.* 2012; Liang and Nielsen 2014a). When multiple crossovers during meiosis segment the stretch of chromosome we are considering, odd-numbered recombinant segments come from one parent, and even number segments come from the other parent; the result is that nonadjacent segments are found in the same parent and are hence nonindependent. This additional covariance from the pedigree structure does not affect our pairwise model of ancestry LD if r is strictly defined as a recombination fraction, as an odd number of recombinations between our pair of loci means that the two alleles are present in different parents in the preceding generation and thereafter follow independent trajectories back in time. Our block length calculations ignore this form of covariance, as we assume that fragments follow independent spatial paths backward in time after recombination events. This assumption will be problematic only for long regions (where more than one recombination can happen per generation) and for short time intervals (*i.e.*, small τ). However, in such cases, ignoring genetic interference may present

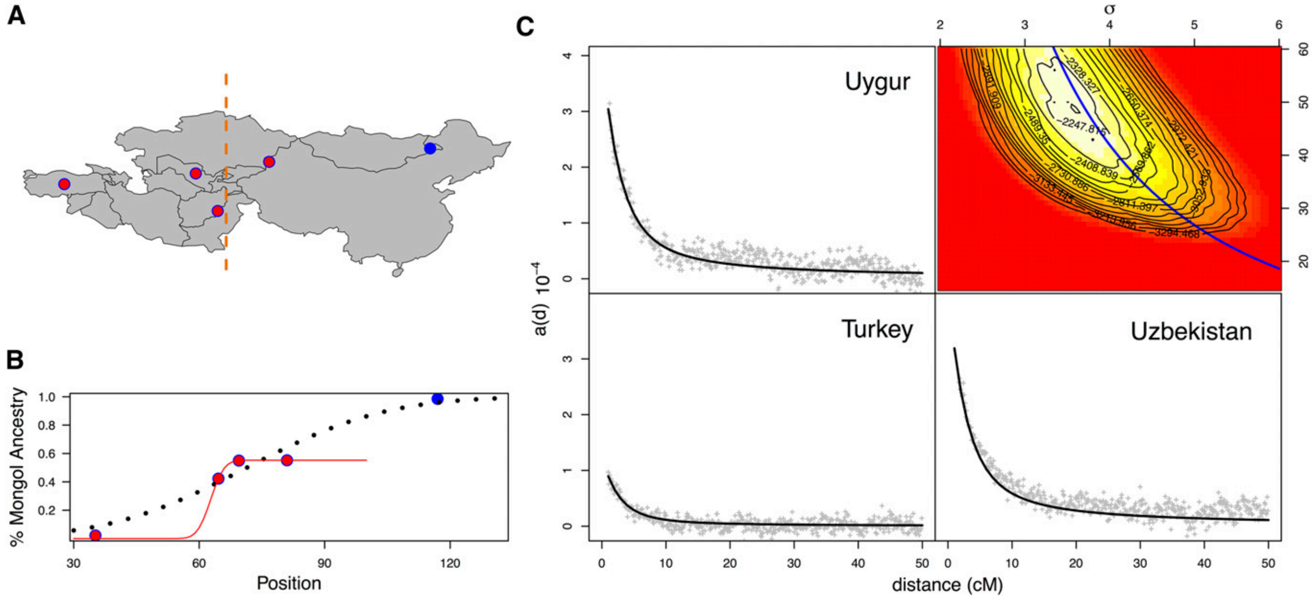


Figure 8 (A) Geographic location of Mongol–Iranian admixed populations used in the analysis. (B) Ancestry proportions, with best fit under the basic Brownian model (dotted, thick line) and under the pulse model (solid thin line). (C) Best fit under our model to LD-decay curves (Hazara not shown) and profile-likelihood surface to the set of all four populations (top right). Blue line indicates $4.2 = \sigma^2\tau$, the compound parameter estimated by fitting to admixture proportions.

a greater source of error than ignoring this additional source of covariance.

Application of the model to human admixture data

We used our model to estimate ages of admixture events and dispersal distances, using genomic data from admixed human populations. To do this we fitted our expressions for ancestry LD to the output of weighted LD from ALDER, but similar information about ancestry LD can be obtained from other methods such as Chromopainter (Lawson *et al.* 2012).

Our spatial model provided a good fit to admixed populations along the Indonesian archipelago, consistent with a relatively straightforward history of admixture across space. Our estimated time of initial contact is somewhat consistent with the work of Xu *et al.* (2012) and is older than that reported by Lipson *et al.* (2014). Our deeper admixture time estimate likely reflects the fact that inference under single-population admixture models will produce estimates of timing of initial admixture that is more recent than estimates under our contact zone model. Our estimate of ~6000 years ago is older than estimates obtained from linguistic analysis (Gray *et al.* 2009). This could be in part due to the simplifying assumptions of our model, which requires dispersal to be constant in time and space. One could imagine, for example, that if there were pulses of human movement followed by a slowing down of dispersal, this would affect our estimate. Finally, we note that the analysis on the Indonesian population was carried out on a relatively small number of SNPs (~17,000). While increasing the number of SNPs would likely improve the analysis, we believe that this demonstrates the utility of this approach even on smaller data sets. It should be noted that the density of SNPs required will depend on the scale of LD present in the populations.

Our spatial model provided a poor fit to the Indian and Central Asian populations. This is likely due, in part, to deviations from a simple model of instantaneous removal of a barrier to contact and continuous diffusion thereafter. A better fit to the data is possible using separate "single-pulse" models for each population; this is unsurprising, given the number of additional parameters such a model uses.

In India, a complex population structure, a caste system, and potentially two waves of contact may have all contributed to difficulties in finding parameters that fit under our model. In particular, the need to separately estimate the intercept meant that there was relatively little information in the decay curves about the timing and mode of admixture. This is especially problematic for older admixture (particularly in the Dravidians), as there is relatively little admixture LD over larger scales and consequently much of our information relies on LD over short genetic distances (<1 cM). Given this paucity of information, it is likely that many, and quite different, admixture models would fit these data nearly equally well. As such, our fit and estimate of timing, and indeed the estimates under alternate models, should be interpreted with caution.

The limited number of populations in Central Asia with signals of Iranian and Mongolian admixture places a limit on the confidence for the fit to the data under any dispersal model. Furthermore, it is known that human movement in the region spans many centuries and is unlikely to be simple. While earlier attempts to date admixture in these populations estimate admixture times of ~30 generations, corresponding to the Mongolian invasions (Hellenthal *et al.* 2014), our estimated time is much older, at $\sim$50 generations. It is unlikely that our demographic model is a good approximation to historical human movement in the area, and this is likely to have affected our inference. However, it is possible that our estimate of earlier admixture is in part reflecting older human movements in the region, and this is in part supported by the findings of Yunusbayev *et al.* (2015).

While our model may not provide a great fit to these human contact zones, our results highlight the sensitivity of the age of admixture estimates to the model used. The continuous and potentially complicated mixing of individuals in a contact zone, especially close to the center of the zone, means that the decay of admixture curves can be much slower than the age of the zone would suggest. Therefore, in many cases it may be difficult to know exactly when admixture began.

Extensions of the simple neutral model and other applications

Our relatively simple expressions describing ancestry LD depend on assuming Brownian movement and on ignoring genetic drift and pedigree structure. The examples of human admixture zones provided above indicate, however, that alternative models may be needed to describe patterns of LD, given different demographic scenarios. Because of the simplicity of our model, modifications can be made with relative ease to describe different geographic scenarios. For example, we were able to apply a model in which the movement of Mongolian genotypes began as a pulse of migrants, followed by diffusion. In a similar vein, one could modify movement to contain a Brownian drift parameter to account for directional migration, although this would require some consideration of how the dispersal kernel of an admixed individual is determined. Discrete deme models could also be used (as we develop in *Appendix B*) to model complex histories of populations in geographic and temporal heterogeneity. However, in practice there is not enough information in admixture LD decay curves to infer detailed population histories with many parameters.

We have demonstrated that inference of admixture parameters can be greatly influenced by the choice of demographic model. This highlights the need for more admixture models to be developed to test with population genomic data and for careful consideration of which model is appropriate for a given biological scenario. The model presented here makes some progress toward addressing the movement of admixed individuals and presents a potential framework for future development of dispersal models. As a final point, we note that all (to our knowledge) admixture models to date, including ours, assume that populations undergo differentiation in relative isolation prior to secondary contact. Under this assumption, there is a strong appeal to fit pulse models (such as a wave of secondary contact) to human admixture data, with the goal of estimating the timing of a pulse and relating it to particular historical events. It seems that perhaps a more appropriate null model in these scenarios would be one in which gene flow has been ongoing between populations, but at a rate slow enough to allow some differentiation to occur. Testing for patterns of LD under this isolation-by-distance (or isolation–migration) model would be a first step toward understanding the demographic history of spatially distributed populations, and the development of such a null model seems an important step in creating future tools for population genomic inference.

As mentioned above, LD has been used to characterize hybrid zones (*e.g.*, Szymura and Barton 1986; Mallet *et al.* 1990; Wang *et al.* 2011), and we see our framework as a potential null model for spatial models of secondary contact, whereby incipient species come back into contact. Although tension zones can maintain distinct species, reproductive isolation is often weak enough to allow diverged populations to exchange alleles. In such scenarios, patterns of diversity that depart from expected ancestry LD could be used to detect potential targets of selection relevant to speciation or local adaptation. It should be noted, however, that good estimates of decay in ancestry LD require reliable genetic maps, as overestimates of genetic distance may give the appearance of a slower rate of decay by inflating LD; this may be a limiting factor in many systems.

The LD induced by the admixing of two differentiated populations provides a powerful population genetic signal that has been measured with genome-wide data to inform the timing of historical admixture events. Building on models that use the patterns of LD to infer admixture dates under scenarios with discretized migration events, we have developed a novel framework that accounts for continuous movements of haplotypes through time and space. We believe that this can serve as a reasonable basic model for understanding patterns of diversity in contact zones. Furthermore, we see potential for this model to be further developed and tailored to fit a range of demographic scenarios, including those that incorporate selection.

Acknowledgments

We thank P. Moorjani, G. Hellenthal, and M. Metspalu for access to data and Simon Aeschbacher, Alison Etheridge, Jeremy Berg, Gideon Bradburd, and Ivan Juric for helpful conversations and comments. We thank Nick Barton, Nick Patterson, and an anonymous reviewer for comments on an earlier draft of this article. This work was supported by the National Science Foundation (NSF) Graduate Research Fellowship under Grant No. (1148897), by the NSF under grant 1262645 to P. Ralph and G. Coop, and by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under award nos. NIH RO1GM83098 and RO1GM107374 awarded to G. Coop.

Literature Cited

Baird, S., N. Barton, and M. Etheridge, 2003 The distribution of surviving blocks of an ancestral genome. Theor. Popul. Biol. 64: 451–471.

Barton, N., and B. Bengtsson, 1986 The barrier to genetic exchange between hybridising populations. Heredity 56: 357–376.

Barton, N. H., 1979 Gene flow past a cline. Heredity 43: 333–339.

Barton, N. H., 1983 Multilocus clines. Evolution 37: 454–471.

Barton, N. H., F. Depaulis, and A. M. Etheridge, 2002 Neutral evolution in spatially continuous populations. Theor. Popul. Biol. 61: 31–48.

Barton, N. H., A. M. Etheridge, and J. Kelleher, and A. Véber, 2013 Inference in two dimensions: allele frequencies *vs.* lengths of shared sequence blocks. Theor. Popul. Biol. 87: 105–119.

Cavalli-Sforza, L., and W. Bodmer, 1971 *The Genetics of Human Populations*, Ed. 1. W. H. Freeman, San Francisco.

Chakraborty, R., and K. Weiss, 1988 Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. USA 85: 9119–9123.

Etheridge, A. M., 2000 *An Introduction to Superprocesses*. American Mathematical Society, Providence, RI.

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.

Fisher, R., 1954 A fuller theory of "junctions" in inbreeding. Heredity 8: 187–197.

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

Gray, M. M., J. M. Granka, C. D. Bustamante, N. B. Sutter, A. R. Boyko *et al.*, 2009 Linkage disequilibrium and demographic history of wild and domestic canids. Genetics 181: 1493–1505.

Haldane, J., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. J. Genet. 8: 299–309.

Harris, K., and R. Nielsen, 2013 Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 9: e1003521.

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.

HUGO Pan-Asian SNP Consortium, 2009 Mapping human genetic diversity in Asia. Science 326: 1541–1545.

Kong, A., G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson *et al.*, 2010 Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467: 1099–1103.

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush, 2012 Inference of population structure using dense haplotype data. PLoS Genet. 8: e1002453.

Li, J., D. Absher, H. Tang, A. Southwick, A. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science 25: 1100–1105.

Liang, M., and R. Nielsen, 2014a The lengths of admixture tracts. Genetics 197: 953–967.

Liang, M., and R. Nielsen, 2014b Understanding admixture fractions. bioRxiv doi: http://dx.doi.org/10.1101/008078.

Lipson, M., P.-R. Loh, N. Patterson, P. Moorjani, Y.-C. Ko *et al.*, 2014 Reconstructing Austronesian population history in Island Southeast Asia. Nat. Commun. 5: 4689.

Loh, P., 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.

Mallet, J., N. Barton, G. Lamas, J. Santisteban, M. Muedas *et al.*, 1990 Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in heliconius hybrid zones. Genetics 124: 921–936.

McKean, H., 1975 Application of Brownian motion to the equation of Kolmogorov-Petrovskii-Piskunov. Commun. Pure Appl. Math. 28: 323–331.

Metspalu, M., I. Romero, B. Yunusbayev, G. Chaubey, C. Mallick *et al.*, 2011 Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. Am. J. Hum. Genet. 89: 731–744.

Moorjani, P., K. Thangaraj, and N. Patterson, 2013 Genetic evidence for recent population mixture in India. Am. J. Hum. Genet. 93: 422–438.

Nagylaki, T., 1975 Conditions for the existence of clines. Genetics 80: 595–615.

Nagylaki, T., 1978 Random genetic drift in a cline. Proc. Natl. Acad. Sci. USA 75: 423–426.

Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland *et al.*, 2012 Ancient admixture in human history. Genetics 192: 1065–1093.

Pearson, K., 1901 Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. Philos. Trans. R. Soc. Lond. Ser. A 195: 1–47, 405.

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.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh, 2009 Reconstructing Indian population history. Nature 461: 489–494.

Shiga, T., 1980 An interacting system in population genetics. J. Math. Kyoto Univ. 2: 213–242.

Szymura, J., and N. Barton, 1986 Genetic analysis of a hybrid zone between the fire-bellied toads, Bombina bombina and B. variegata, near Cracow in southern Poland. Evolution 40: 1141–1159.

Ungerer, M. C., S. J. Baird, J. Pan, and L. H. Rieseberg, 1998 Rapid hybrid speciation in wild sunflowers. Proc. Natl. Acad. Sci. USA 95: 11757–11762.

Wakeley, J., L. King, B. S. Low, and S. Ramachandran, 2012 Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190: 1433–1445.

Wang, L., K. Luzynski, J. E. Pool, V. Janoušek, P. Dufková *et al.*, 2011 Measures of linkage disequilibrium among neighbouring SNPs indicate asymmetries across the house mouse hybrid zone. Mol. Ecol. 20: 2985–3000.

Wright, S., 1946 Isolation by distance under diverse systems of mating. Genetics 31: 39–59.

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.

Yunusbayev, B., M. Metspalu, E. Metspalu, A. Valeev, S. Litvinov *et al.*, 2015 The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. PLoS Genet. 11: e1005068.

Communicating editor: N. H. Barton

Appendix A: Covariance in Ancestry

By integration by parts, Equation 3 becomes

$$\mathbb{E}[\mathbf{1}_B(\mathcal{A}_1)\mathbf{1}_B(\mathcal{A}_2)] = e^{-r\tau}\Phi\left(\frac{\ell}{\sigma\sqrt{\tau}}\right) + \left[-e^{-rt}\int_\ell^\infty\int_\ell^\infty f_t(y,z)\;dydz\right]_{t=0}^{t=\tau} + \int_0^\tau e^{-rt}\int_\ell^\infty\int_\ell^\infty \frac{\partial f_t}{\partial t}\;dydz\;dt, \tag{A1}$$

where $f_t(y,z)$ is the bivariate Gaussian density for jointly distributed (Y,Z) with correlation t. The second term of (A1) is

$$\left[-e^{-rt}\int_\ell^\infty\int_\ell^\infty f_t(y,z)\;dydz\right]_{t=0}^{t=\tau} = \Phi\left(\frac{\ell}{\sigma\sqrt{\tau}}\right)^2 - e^{-r\tau}\Phi\left(\frac{\ell}{\sigma\sqrt{\tau}}\right). \tag{A2}$$

For the third term of (A1), we can utilize the useful identity that for a bivariate Gaussian with variances 1 and correlation t (Pearson 1901),

$$\frac{\partial}{\partial t}f_t(y,z) = \frac{\partial^2}{\partial y \partial z}f_t(y,z), \tag{A3}$$

and so the last term of (A1) becomes

$$\int_0^\tau e^{-rt}\int_\ell^\infty\int_\ell^\infty \frac{\partial f_t}{\partial t}\;dydz = \int_0^\tau e^{-rt}f_t(\ell,\ell)\;dt. \tag{A4}$$

Combining Equation 2 and (A1), (A2), and (A4) therefore leaves us with

$$\mathrm{Cov}(\mathcal{A}_1,\mathcal{A}_2) = \int_0^\tau e^{-rt}\frac{1}{2\pi\tau\sigma^2\sqrt{1-(t/\tau)^2}}\exp\left(-\frac{\ell^2(1-t/\tau)}{\tau\sigma^2\left(1-(t/\tau)^2\right)}\right)dt. \tag{A5}$$

Appendix B: Island Model

In a discretized time and space model, with n islands of equal, constant size and per-generation migration rates defined by the $n \times n$ matrix M, the motion of a single lineage is described by a discrete-time Markov chain, and so the expected frequency of loci inherited from ancestry B in population X is

$$\mathbb{E}\left[\mathbf{1}_B(\mathcal{A})\right] = \sum_{j\in S} M^\tau_{X,j}, \tag{B1}$$

where τ is the number of generations since admixture began, S is the set of demes that are defined as being ancestry B at the time of contact, and $M^\tau_{X,j}$ is the X,jth element of the τth matrix power of M. The covariance is derived by summing over possible recombination times and the location of the allele at the time of recombination (N is the set of all locations):

$$\mathrm{Cov}(\mathcal{A}_1,\mathcal{A}_2) = (1-r)^\tau \sum_{i\in S} M^\tau_{X,i} + \sum_{t=0}^{\tau-1}(1-r)^t r\left(\sum_{j\in N} M^t_{X,j}\left(\sum_{a\in S} M^{\tau-t}_{j,a}\right)^2\right) - \left(\sum_{i\in S} M^\tau_{X,i}\right)^2. \tag{B2}$$

Note that r is the probability of any odd number of recombinations occurring, *i.e.*, the probability that a Poisson random variable with mean d is odd.

Appendix C: Unlabeled Rooted Topologies and Their Probabilities

Given that we condition on the number of recombination events within a time interval, the events are uniformly distributed both through time and along a chromosome of unit length. The most recent recombination back in time is therefore the minimum time for k uniformly distributed events. This first recombination event (in time) splits the chromosome into two products, which become two subtrees, with j and $k-1-j$ recombinations, respectively. The two subtrees are independent and have similar properties to those of the whole chromosome, so the process of bifurcation can be iterated. The first recombination splits the remaining ones uniformly, such that the probability of generating one subtree of size j is $2^s/k$, where $s=0$ when $2j=k-1$ and $s=1$ otherwise.

Thus, for the set of a_k unlabeled topologies with $k+1$ leaves, $\mathcal{T}^k=\{\mathcal{T}^k_1\dots\mathcal{T}^k_{a_k}\}$, a topology $\mathcal{T}^k_i$ is generated by joining $\mathcal{T}^j_m\in\mathcal{T}^j$ with $\mathcal{T}^{k-j-1}_n\in\mathcal{T}^{k-j-1}$ at the root (m and n are arbitrary). Because each subtree is independent, the probability, $P(\mathcal{T}^k_i)$, of topology $\mathcal{T}^k_i$, conditioning on k recombination events, can be obtained by the product of the probabilities of each subtree and the probability of generating two subtrees of sizes j and $(k-1-j)$,

$$P\left(\mathcal{T}^k_i\right)=\frac{2^s}{k}P(\mathcal{T}^j_m)P\left(\mathcal{T}^{k-j-1}_n\right), \tag{C1}$$

where $\mathcal{T}^k_i$ is the topology made by joining topologies $\mathcal{T}^j_m$ and $\mathcal{T}^{k-1-j}_n$ at the root. Here s indicates the symmetry of $\mathcal{T}^k_i$, such that $s=0$ if the tree is symmetric (*i.e.*, the two subtrees $\mathcal{T}^j_m$ and $\mathcal{T}^{k-j-1}_n$ are the same) and $s=1$ otherwise, so that $2^s/k$ gives the probability that the first recombination event in $\mathcal{T}^k_i$ produces two subtrees of the required sizes.

When integrating over the set of all trees conditioning on a topology, we need to multiply by the probability density $V(t_1,\cdots,t_k)$ of the recombination times given the topology. We start by considering the probability of the first recombination event t_1, at the root of a $(k+1)$-tipped tree, which is the first order statistic of k independent and uniformly distributed events over the interval $(0,\tau)$. The probability density of this first recombination given k total recombinations is therefore

$$\frac{k}{\tau}\frac{(\tau-t_1)^{(k-1)}}{\tau^{k-1}}.$$

Similarly, the timing of the jth node is the first-order statistic within the subset of recombination events that generate the subtree of which it is the root. The jth node, from which M_j nodes are descended, which itself is directly descended from the parent node p_j, thus has the probability density

$$v\left(t_j,t_{p_j},M_j\right)=(M_j+1)\frac{(\tau-t_j)^{M_j}}{\left(\tau-t_{p_j}\right)^{M_j+1}}. \tag{C2}$$

The independence of each subbranch allows us to compute the probability density of all recombination events as

$$V(t_1,\cdots,t_k)=\prod_{j=1}^{k}v\left(t_j,t_{p_j},M_j\right).$$

Note that Equation C2 is the Beta distribution $B(1,M_j+1)$, scaled to lie on the interval $[t_i,\tau]$.

Appendix D: Obtaining Block Length Distributions by a Branching Brownian Motion

Here we describe an alternative approach to finding an expression for the probability that an entire region of length d is of ancestry B, $U_d(\tau,\ell)$ of Equation 5, without conditioning on the number of recombination events. The process of recombination and dispersal described above is analogous to a branching Brownian motion (BBM), where recombination is represented by a splitting event. In standard BBM, each lineage has the same rate of splitting, but here the total length of the chromosome is constant, and so we have conservation of the total rate of splitting d. The rate of splitting on a lineage decreases with each recombination event, as both products of recombination are shorter (and therefore have a smaller probability of recombination).

Below, we derive an integro-differential equation satisified by U, similar to the classic analysis of branching Brownian motion by McKean (1975). Starting in the present, we follow a single lineage backward in continuous time. The movement of this lineage is Brownian with variance σ^2. We model recombination events between the two loci as a Poisson process with rate d. At the first recombination event, we generate a uniform random variable, $r_1\in[0,d)$ to represent the genomic position of the recombination event. We then split the sequence into left and right fragments—$[0,r_1)$ and $[r_1,d)$, respectively. Following this, the two linages move independently backward in time with respective recombination (splitting) rates of r_1 and $d-r_1$. This process is iterated over the time period τ.

We consider moving back a very short time interval Δt from the present and take the expectation over the random events that could have occurred in that time interval. (In other words, we are writing down the infinitesimal generator of this Markov process.)

With probability $1 - d\Delta t + O(\Delta t^2)$ there is no recombination during the interval Δt and conditioning on this, we have only to take the expectation over the small random change Δx in spatial location during this time,

$$U_d(\tau, \ell | \text{no rec.}) = \mathbb{E}_{\Delta x}\left[U_d(\tau - \Delta t, \ell + \Delta x)\right], \tag{D1}$$

where $\mathbb{E}_{\Delta x}$ is the expectation over all changes in position X.

A recombination event occurs in the interval Δt with probability $d\Delta t$. Conditioning on recombination occurring at time t_{rec} at position $\ell + \Delta x'$, producing two recombinants of length d_1 and $d - r_1$,

$$U_d(\tau, \ell | \text{rec.}) = \int_0^d \int_0^{\Delta t} \mathbb{E}_{\Delta x'}\left[U_{r_1}\left(\tau - t_{\text{rec}}, \ell + \Delta x'\right) U_{d-r_1}\left(\tau - t_{\text{rec}}, \ell + \Delta x'\right)\right] \, dr_1 dt_{\text{rec}}, \tag{D2}$$

where U_{r_1} is the probability that all subsequent recombinants along the chromosomal fragment of length r_1 are of ancestry type B.

Combining (D1) and (D2) and taking $\Delta t \to 0$ obtains

$$\frac{\partial U_d}{\partial t}(\tau, x) = \frac{\sigma^2}{2} \frac{\partial^2 U_d}{\partial x^2}(\tau, x) + \int_0^d U_{r_1}(\tau, x) U_{d-r_1}(\tau, x) - U_d(\tau, x) \, dr_1, \tag{D3}$$

with boundary conditions $U_d(0, x) = 1$ for $x > 0$ and $U_d(0, x) = 0$ for $x \leq 0$. This differential equation is solved by $U_d(t, x)$, defined in Equation 5, and is the probability that at time τ in the past, the leftmost branch of this branching process initiated at position x_0 is at a position $x > 0$. This differential equation is related to that presented by Baird *et al.* (2003) to describe the survival of genomic blocks within a panmictic population (but the latter does not have a spatial diffusion term). The equation is similar to the Fisher-KPP equation, with differences arising from the nonconstant splitting rate. The first term of Equation D3 reflects the spatial diffusion of lineages, and the second term reflects the splitting of blocks of length d by recombination into two shorter blocks (of size $d - r_1$ and r_1) that each must be of type B.

Appendix E: Invasion Pulse

Because it is unlikely that Mongolian movement during the 13th century was Brownian, we construct an alternative model for Mongolian movement, in which individuals of Mongolian ancestry (ancestry B) initially invade and displace some proportion ψ of the resident population over some geographic space $[C_1, \infty]$. We assume that most of the distance between C_1 and the source Mongolian population has been invaded in this way. This invasion occurs instantaneously at time τ, such that the frequency of ancestry B at time τ is

$$g(B) = \begin{cases} 0 & x < C_1 \\ \psi & x > C_1. \end{cases}$$

Following the model presented in *Materials and Methods*, we assume that a lineage that is found at position $x > C_1$ at time τ has probability ψ of having ancestry B. The probability that a single lineage $\mathcal{A}$, sampled at location ℓ at $t = 0$, is given by

$$\mathbb{E}\left[\mathbf{1}_B(\mathcal{A})\right] = \psi \cdot \Phi\left(\frac{\ell - C_1}{\sigma\sqrt{\tau}}\right).$$

For a lineage that has recombined, for both lineages to have ancestry B, both lineages have to be found in $x > C_1$ at time τ, at which point there is a ψ^2 chance of both having ancestry B. Using the approach taken in *Appendix A*, we can obtain the following expression for ancestry LD at position ℓ:

$$(1 - \psi)\psi e^{-r\tau} \Phi\left(\frac{(\ell - C_1)}{\sigma\sqrt{\tau}}\right) + \psi^2 \int_0^1 e^{-rt\tau} \frac{1}{2\pi\sigma^2\sqrt{1 - t^2}} \exp\left(-\frac{(\ell - C_1)^2}{\tau\sigma^2(1 + t)}\right) \, dt. \tag{E1}$$

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179838/-/DC1

The Spatial Mixing of Genomes in Secondary Contact Zones

Alisa Sedghifar, Yaniv Brandvain, Peter Ralph, and Graham Coop

DOI: 10.1534/genetics.115.179838

Supplemental tables

Table S1: Human populations used in analyses, with the geographic locations that were used.

Population	latitude	longitude	sample size
IDMT (Mentawai)[1]	0.3S	98.4E	15
IDJV (Javanese)[1]	7.3S	110.4E	19
IDSB (Kambera)[1]	9.8S	120.0E	20
IDSO (Manggarai)[1]	8.6S	120.1E	19
IDRA (Manggarai)[1]	8.7S	120.5E	17
IDLA (Lamaholot)[1]	8.3S	123.0E	20
IDLE (Lembata)[1]	8.3S	124.6E	19
IDAL (Alorese)[1]	8.3S	124.7E	19
Papuan[2]	4.0S	143.0E	17
Mongola[3]	48.0N	119.0E	10
Hazara[3]	33.0N	69.5E	22
Turkish[3]	39.0N	35.2E	17
Uygur[3]	44.0N	81.0E	10
Uzbekistani[3]	41.4N	64.6E	15
Iran[3]	32.4N	53.7E	20
Kashmiri Pandit[4]	34.22N	75.5E	20
Pathan[2]	32.35N	69.72E	23
Kshatriya[5]	27.56N	78.65E	27
Kanjar[5]	26.45N	80.32E	8
Brahmin (UP)[5]	26.02N	83.18E	15
Brahmin[4]	25.45N	82.41E	8
Kshatriya (UP)[5]	24.45N	82.41E	7
Kshatriya[4]	27.56N	78.65E	20
Dharkar[5]	25.44N	83.10E	12
Chamar[5]	25.37N	83.04E	10
Sindhi[2]	24.27N	68.70E	25
Bhil[4]	23.02N	72.40E	17
Madiga[4]	17.58N	79.35E	19
Mala[4]	17.22N	78.29E	18
Velama[6]	17.05N	79.27E	4
Vysya[4]	14.41N	77.39E	20
Kallar[5]	10.99N	78.22E	7
Onge[5]	11.6N	92.7E	9
Basque[2]	43N	0	24

[1] The HUGO Pan-Asian SNP Consortium (2009);
[2] Li *et al.* (2008);
[3] Hellenthal *et al.* (2014);
[4] Metspalu and Romero (2011);
[5] Moorjani *et al.* (2013);
[6] Reich *et al.* (2009)

Table S2: Estimated parameters, under the exponential model (Eq. 11) for the Indonesian populations used in our analysis (sum of squares fit). Here, each population has been fit independently

Population	% Asian	Timing	Constant(Multiplicative)	Constant (Additive)	$\mathcal{L}$
IDAL	44.4	29.1	3.01e-04	2.47e-06	13732.60
IDJV	99.8	665.2	2.68e-2	5.80e-29	12138.46
IDLA	61.6	60.9	7.85e-04	2.016e-08	12505.03
IDLE	58.6	40.7	5.18e-04	6.22e-07	14097.24
IDRA	77.5	106.0	1.45e-03	2.80e-07	11188.83
IDSB	78.7	94.9	1.22e-03	8.16e-13	14042.64
IDSO	66.7	33.5	4.57e-04	5.81e-06	16442.91

Table S3: Estimated parameters, under the exponential model (Eq. 11) for the Central Asian populations used in our analysis (sum of squares fit). Here, each population has been fit independently

Population	% Mongola	Timing (gens)	Constant (Mult.)	Constant (Add.)	$\mathcal{L}$
Hazara	55.0	25	3.8e-04	1.7e-05	347.1
Turkey	2.2	30	1.2e-04	1.8e-06	436.6
Uygur	55.2	24	3.3e-04	2.3e-05	509.8
Uzbekistan	42.3	20	3.1e-04	2.8e-05	393.5

Supplemental figures

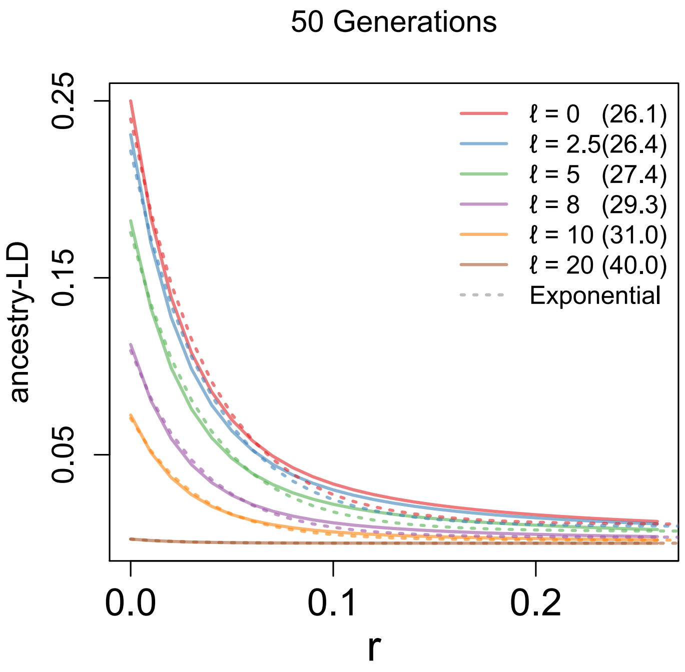


Figure S1: Exponential fits (Eq. 11) to ancestry-LD in populations sampled at locations (L) from a 50-generation old contact zone. Solid lines represent the output of simulations *under the model*, and dashed lines the best exponential fit. The estimated timing for each population is shown in parentheses.

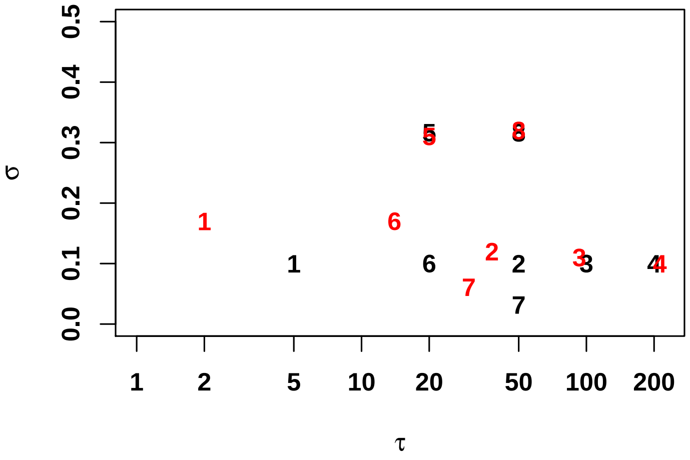


Figure S2: Simulations run under different combinations of τ and σ (in black), and the inferred combinations of parameters for each simulated dataset (red). All simulations here were run *under the process*.

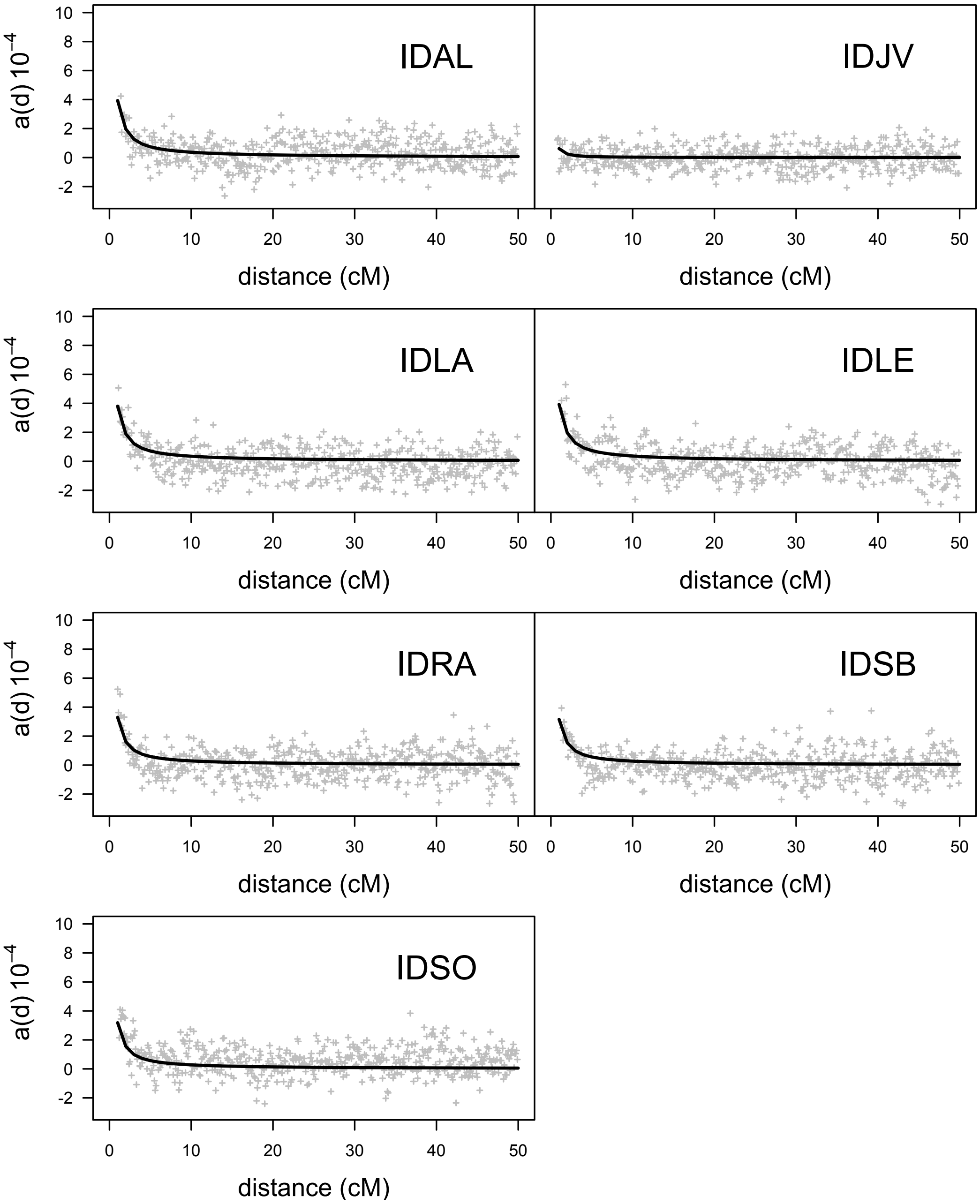


Figure S3: Fits to decay for all Indonesian populations used in analysis, described in Table S1 using the best fit parameters as described in the main text. Grey points are estimates generated by ALDER, and black curves are expected LD under the estimated parameters.

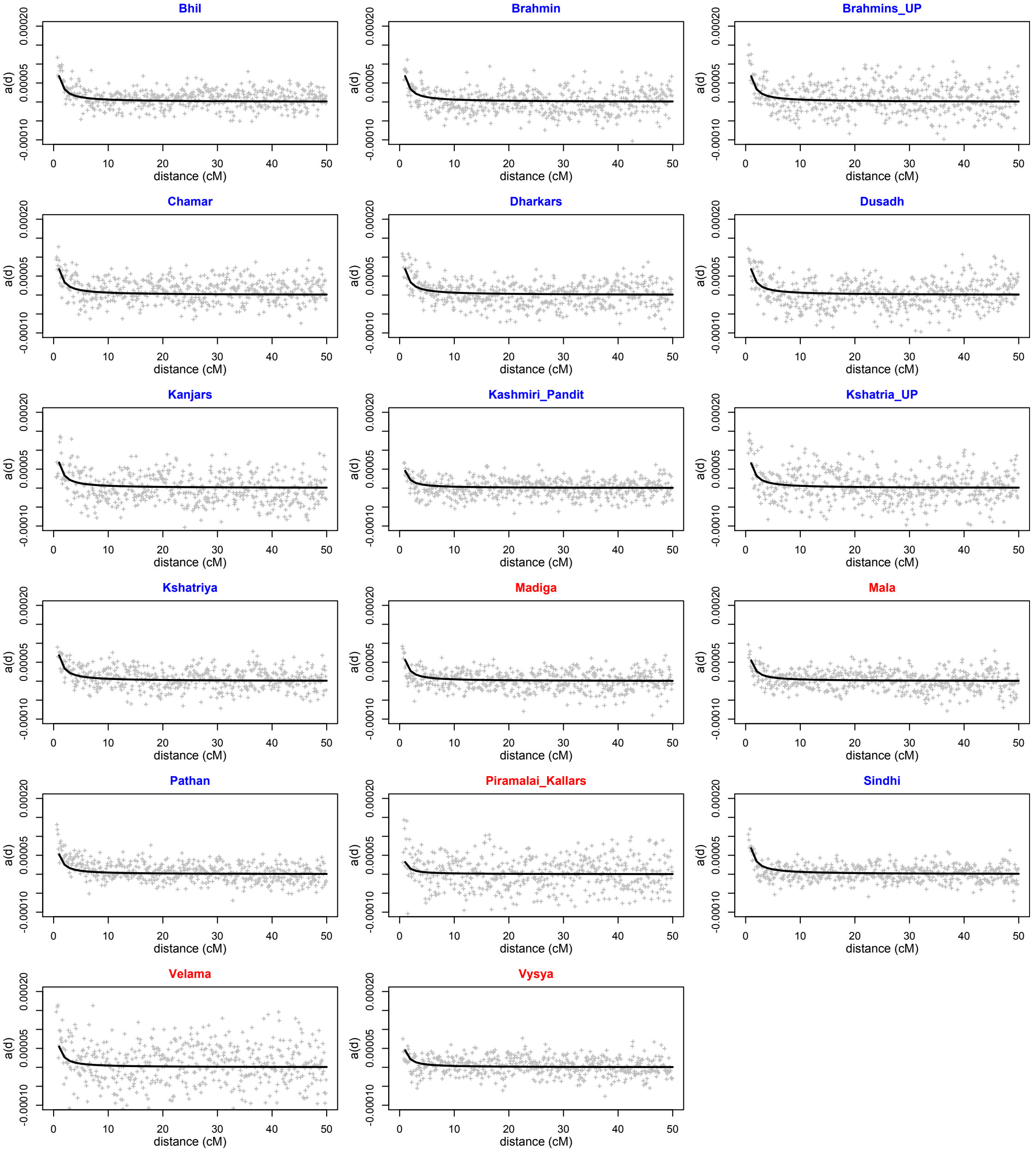


Figure S4: Fits to decay for all Indian populations used in analysis, described in Table S1 using the best fit parameters as described in the main text. Grey points are estimates generated by ALDER, and black curves are expected LD under the estimated parameters. Blue names indicate Indo-European populations, and red labels Dravidian.

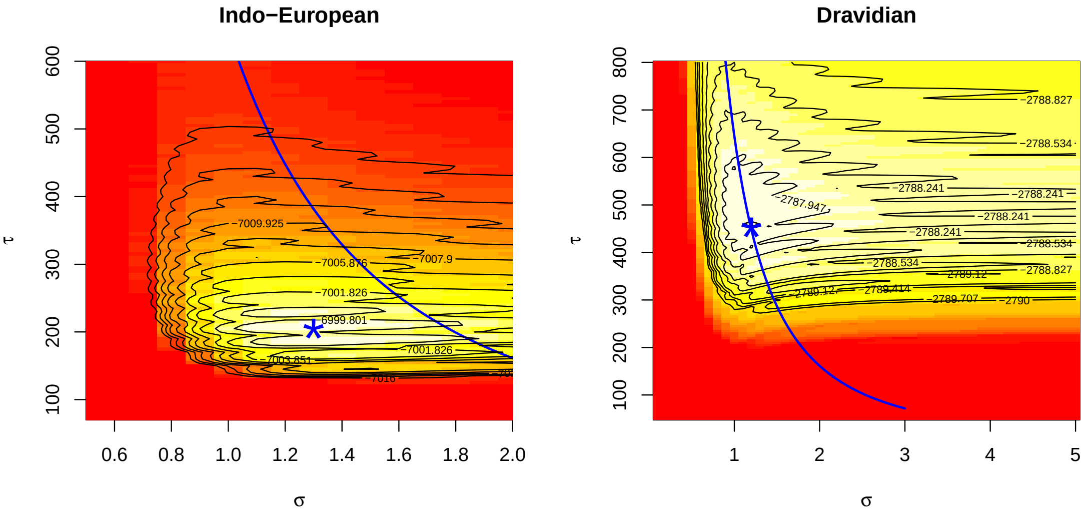


Figure S5: Profile likelihood surfaces for fits to the Indo-European and Dravidian subsets of the population. Blue asterisk indicates parameters giving best fit.

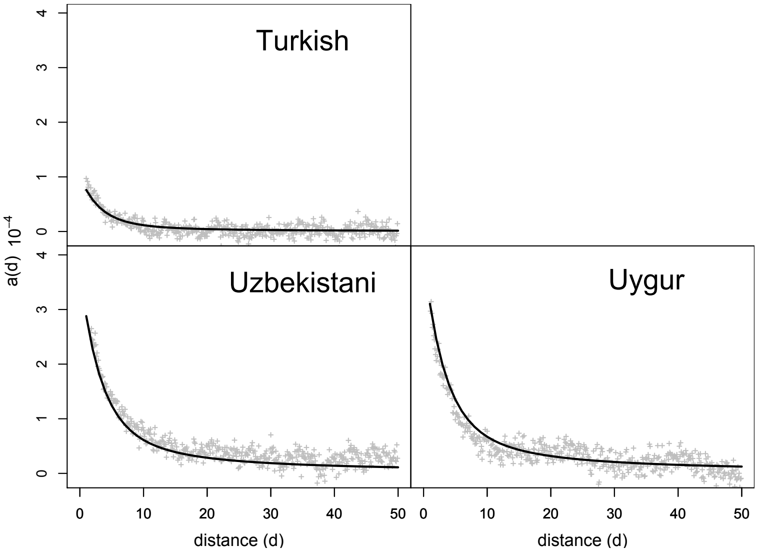
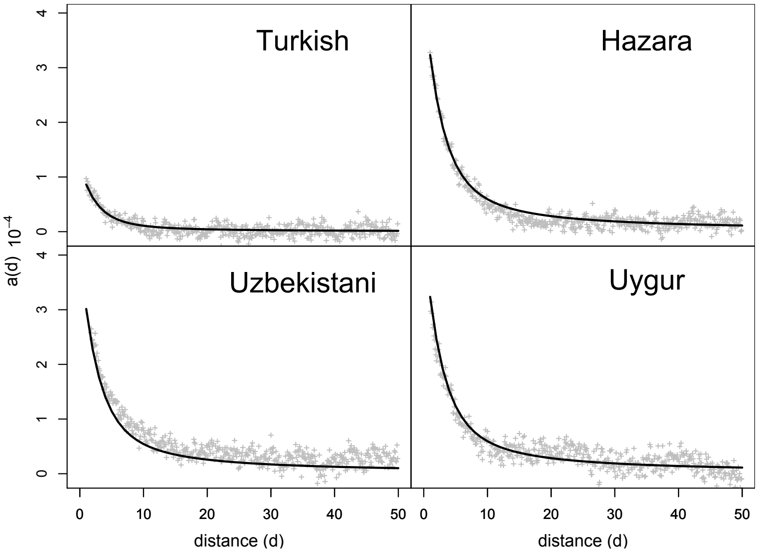


Figure S6: Best-fit curves for each population when fit is made to the set of three Asian populations used in our analysis (Hazara omitted). Grey points are estimates generated by ALDER, and black curves are expected LD under the estimated parameters.

Figure S7: Best-fit curves for each population when the fit is made to the set of the four Central Asian populations used in our analysis. Grey points are estimates generated by ALDER, and black curves are expected LD under the estimated parameters.



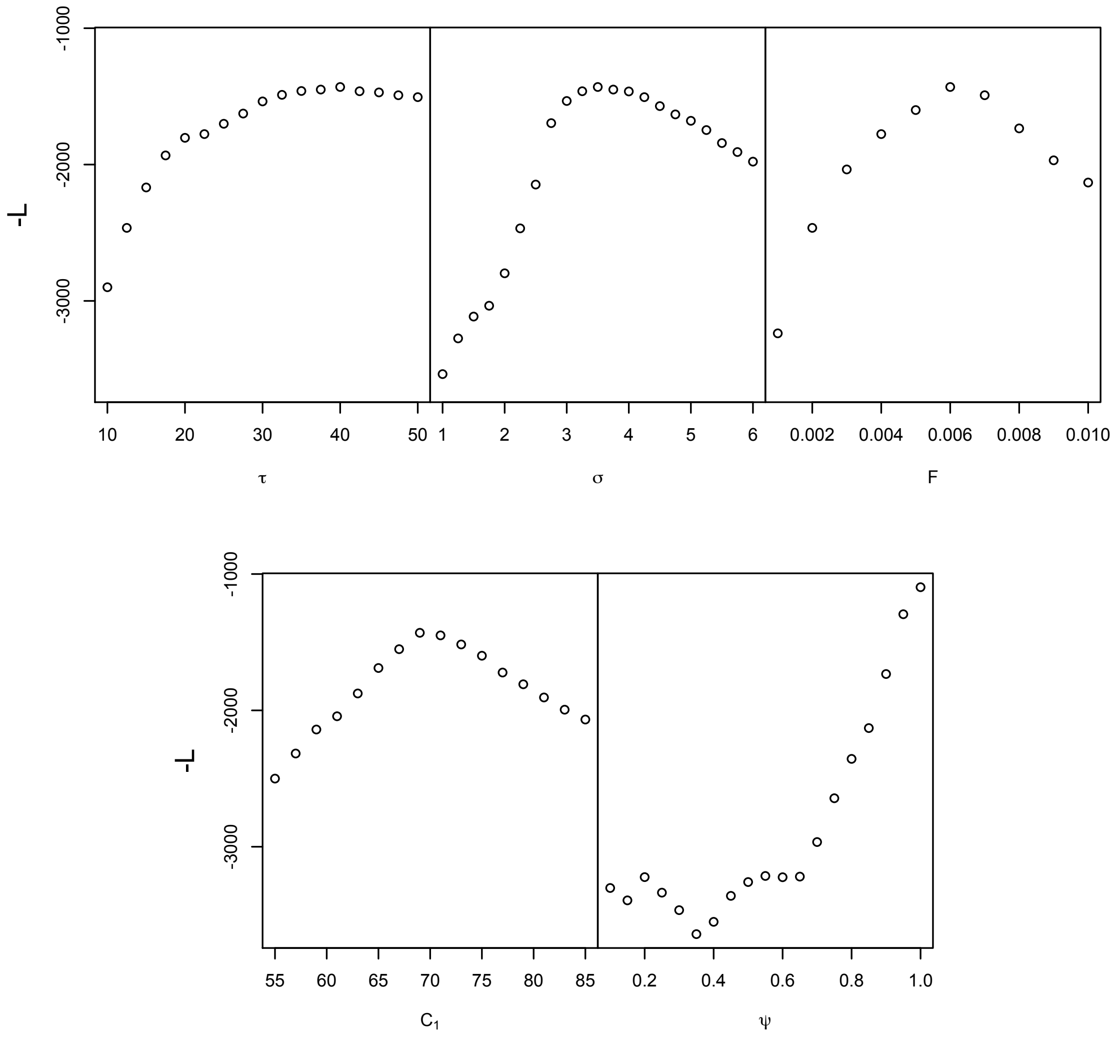


Figure S8: Profile likelihood curves for the five parameters fitted to the Central Asian populations under the invasion-pulse model, showing that the best fit is to a model with cline center at approximately 67°E and $\psi = 1$. This is roughly equivalent to the original model of secondary contact.

File S1: Analysis of Indian populations

Following Moorjani *et al.* (2013), we ran the F_4 ratio tool in the ADMIXTOOLS package (Patterson *et al.* 2012) on Georgian, Basque, Yoruba, Onge and the focal Indian population to estimate ANI ancestry proportions in these populations (Fig. 5). We fit a latitudinal cline to these ancestry proportions (Eq. 1) returning a cline center at $24°4'N$ and $\sigma\sqrt{\tau} = 25.4$. Because the gradient of ancestry could run along any geographic axis, we also tried to fit ancestry proportion clines to various transects using linear combinations of latitude and longitude. Since these did not produce substantially better fits than latitude alone, we chose to use latitude as our geographic axis (results not shown). Through this analysis, we aimed to closely follow the procedure outlined in (Moorjani *et al.* 2013) to generate LD curves and improve model-fitting.

We then generated co-ancestry decay curves in ALDER for each of these samples, using weightings from Basque and Onge parental populations as proxies for the ANI and ASI populations (see Moorjani *et al.* (2013)). We consider three possible contact zone scenarios under our geographic model: One in which all population samples form a contact zone and, based on the findings of earlier studies, one that comprises only the Indo-European and one that comprises only the Dravidian populations. We initially attempted to fit the τ, σ and F parameters in Eq. 12 simultaneously, but faced some difficulty as there appears to be limited information about F. This results in wide range of values fitting the data equally well, but give rise to very different surfaces for σ and τ. We attributed this to a deficit of information in the curves, leading to non-identifiability, due to relative low levels of differentiation and relatively rapid decay of ancestry-LD. The difficulty in estimating the intercept of admixture-LD curves had been noted before (Loh *et al.* 2013), and can reflect the fact that very close pairs of markers are discarded to remove the effects of LD in the ancestral populations. This results in the fitted curve being relatively unconstrained near $r = 0$. To remedy this, we estimated $\mathcal{F}$ using an approach similar to that taken by Moorjani *et al.* (2013). Using MIXMAPPER (Lipson *et al.* 2013), we estimated the value of $\mathcal{F}$ as $2F_2(ANI; ASI)^2$ using the Onge and Basque populations as present day proxies. We then fit values of σ and τ under the range of F_2 values computed by MIXMAPPER ($(0.015, 0.042)$). The profile likelihood surface was generated over 20 values of $\mathcal{F}$. We also use the value estimated above as the cline center for all three fits.

We first fit our LD curves to all populations under a model in which all Indo-European and Dravidian populations are the outcome of a single admixture contact zone. The best fit was approximately 220 generations since contact with $\sigma = (0.9\,\text{degrees} \approx 100\,\text{km})/\text{generation}$ (Fig. 5). Fits to the subset of populations classified as Indo-European yielded a contact zone age of approximately 200 generations, and $\sigma = (1.3\,\text{degrees} \approx 144\,\text{km})/\text{generation}$ (Fig. S5). Finally, we fit the subset of Dravidian populations (Fig. S5), which found a best fit of 460 generations with $\sigma = (1.2\,\text{degrees} \approx 133\,\text{km})/\text{generation}$ on a relatively flat surface. This is likely because there is very little information in the decay of LD in this subset given there are so few Dravidian populations, and that the LD curves are relatively flat. The profile likelihood surface was generated over 40 evenly distributed values of $\mathcal{F}$ spanning the values inferred above using MIXMAPPER. For all three groups of populations we used our earlier estimate for cline center.

References

Hellenthal, G., G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers, 2014 A genetic atlas of human admixture history. Science **343**: 747–51.

Li, R., Y. Li, K. Kristiansen, and J. Wang, 2008 SOAP: short oligonucleotide alignment program. Bioinformatics (Oxford, England) **24**: 713–4.

Lipson, M., P.-R. Loh, A. Levin, D. Reich, N. Patterson, and B. Berger, 2013 Efficient moment-based inference of admixture parameters and sources of gene flow. Molecular biology and evolution **30**: 1788–802.

Loh, P., M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell, D. Reich, and B. Berger, 2013 Inferring admixture histories of human populations using linkage disequilibrium. Genetics **193**: 1233–1254.

Metspalu, M., and I. Romero, 2011 Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. The American Journal of . . . : 731–744.

Moorjani, P., K. Thangaraj, and N. Patterson, 2013 Genetic evidence for recent population mixture in India. The American Journal of Human Genetics : 422–438.

Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich, 2012 Ancient admixture in human history. Genetics **192**: 1065–1093.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh, 2009 Reconstructing Indian population history. Nature **461**: 489–94.

The HUGO Pan-Asian SNP Consortium, 2009 Mapping human genetic diversity in Asia. Science (New York, N.Y.) **326**: 1541–5.

Files S2-S3

Available for download at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179838/-/DC1

File S2 The scripts in this file provide a very basic framework for fitting a spatial diffusion model of admixture to weighted LD.

File S3 This R script was written to simulate recombining chromosomes in finite populations. This works by tracing the ancestry of each chromosome portion in an admixed population to the initial population. Each generation consists of SELECTION, RANDOM MATING (WITH RECOMBINATION) and MIGRATION and therefore accounts for population genealogy and drift. Generations are discrete.