High fraction of silent recombination in a finite-population two-locus neutral birth-death-mutation model

A. B. Melka[1] and Y. Louzoun[1,2,*]

[1]*Department of Mathematics, Bar-Ilan University, Ramat Gan 52900, Israel*
[2]*Gonda Brain Research Center, Bar-Ilan University, Ramat Gan 52900, Israel*

A precise estimate of allele and haplotype polymorphism is of great interest in theoretical population genetics, but also has practical applications, such as bone marrow registries management. Allele polymorphism is driven mainly by point mutations, while haplotype polymorphism is also affected by recombination. Current estimates treat recombination as mutations in an infinite site model. We here show that even in the simple case of two loci in a haploid individual, for a finite population, most recombination events produce existing haplotypes, and as such are silent. Silent recombination considerably reduces the total number of haplotypes expected from the infinite site model for populations that are not much larger than one over the mutation rate. Moreover, in contrast with mutations, the number of haplotypes does not grow linearly with the population size. We hence propose a more accurate estimate of the total number of haplotypes that takes into account silent recombination. We study large-scale human leukocyte antigen (HLA) haplotype frequencies from human populations to show that the current estimated recombination rate in the HLA region is underestimated.

I. INTRODUCTION

Multiple genetics models relate allele frequencies to their populations' dynamics [1–5], typically including processes such as mutations, genetic drift, selection, or migration between subpopulations [6–9]. For the haplotype frequencies, another essential process to consider is recombination [10–12]. During cell division in sexual reproduction, crossovers can occur between the maternal and paternal homologous chromosomes and result in the exchange of genetic material [13–15]. Therefore, offspring may have different combinations of genes than either of their parents on the same chromosome, leading to the creation of new haplotypes and increasing genetic variability. Determining the recombination rate (i.e., the probability for those crossovers to occur) is crucial in evolutionary biology and medical population genetics [16,17]. It also has important implications for transplant donors registry management [18,19].

Two approaches have been proposed to build recombination maps and estimate the recombination rate. The first one, referred to as the direct approach, is strictly experimental and consists of sperm genotyping [20]. The second approach is an indirect method that uses genetic linkage (coinheritance of markers in families) to produce recombination maps for chromosome segments [21]. These maps describe the distance between genes, or markers, as a function of their probability to recombine. If two genes are on two different chromosomes or very distant, they are considered uncorrelated and the distribution of the haplotypes reflects this independence. On the other hand, two adjacent genes will have a high probability of

being inherited together. This probability is a direct function of the recombination rate between those genes and the distance between them [22]. These segments can then be linked to provide estimates of recombination frequencies for specific chromosomes [23], typically using maximum likelihood estimation (MLE) [24,25] based on a coalescent tree model [26–29]. The coalescent model goes backward to estimate the time for two individuals to reach their most recent common ancestor. However, this model does not take into consideration silent recombination producing twice the same offspring. In the direct approach, recombination events are observed over multiple 10^6 of sperm cells and, therefore, the computed rate is per cell division. In the indirect approach, recombination events are observed from one generation to the other and the rate computed is per generation.

In a single-gene, fixed population, neutral model, the mutation rate has been related to the number of alleles through $\theta = 4N_e\mu$ (where θ is the overall number of mutations for the population, N_e is the effective population as defined by Kimura [2], and μ is the individual mutation rate per generation). This estimator was first derived by Watterson to describe mutations [5]. It is based on an infinite site model (i.e., each mutation creates a new allele). It has often been assumed that recombination behaves like mutations and the same concept was extended to multiple estimates of the recombination rate, where the number of alleles was simply replaced by the number of haplotypes with $\rho = 4N_e r$ (where ρ is the overall number of recombination for the population and r is the individual recombination rate per generation) [26–28]. The usage of this estimator is limited by the need to determine the effective population N_e [30,31]. As such, one needs to simultaneously estimate N_e and μ (or r). Therefore, studies often either display the ratio between recombination

and mutations [32] or simply compute ρ and θ instead of μ and r, separately or jointly [27,28]. Other studies use samples for which the origin of the population is known [33]. Note that the number of alleles actually differs from Watterson's estimator. Indeed, multiple corrections were proposed [25,34,35]. The most significant limitation of this estimator is the assumed equivalence between recombination and mutations. As mentioned above, unlike mutations, recombination is drawn from a finite existing pool of alleles. Specifically, given the fat tail of the haplotype size distribution, there is a nonzero probability of reproducing the same combination of haplotypes (silent recombination). As such, the recombination rate expected from Watterson's formula is largely underestimated. We show that for average size populations, the fraction of silent recombination is close to 1.

We here use a statistical model on the observed alleles and haplotypes distributions and infer the mutation and recombination rates. We use a birth-and-death process [36] rather than a coalescent tree since it allows for the inclusion of silent recombination. We identify two regimes depending on the size of the effective population. For $N_e \gg 1/\mu$, the pool of alleles is large, and the typical type size for each allele is small so that almost all recombination events create a new haplotype and the infinite site assumption holds for recombination. However, for $N_e \leqslant 1/\mu$, the number of potential recombination creating new haplotypes is limited by an upper bound, induced by the high probability of sampling very frequent alleles. In this regime, almost all recombination events are silent. The fraction of silent recombination grows with the recombination rate. Moreover, the population size where the transition from one regime to the other occurs increases with the recombination rate. Using this insight, we compute the number of different two-locus haplotypes in a population (the haplotype polymorphism), using a revised relationship between the number of alleles and haplotypes and the mutation and recombination rates. As an application, we analyze the distributions of alleles and haplotypes in the HLA locus for human populations and show that the recombination rate is underestimated.

II. SINGLE-LOCUS AND TWO-LOCUS MODELS

As a preliminary step, we focus on a single locus and estimate the number of alleles with respect to the mutation rate. We follow Ref. [36] and assume a neutral infinite site Moran model with equal birth and death rates (so that the total population is maintained fairly constant) that can be arbitrarily set equal to 1 (up to a time scaling). We define μ as the per generation mutation rate and the probability for an allele to have a population of size k is given by Fisher log series [37] (see the Appendix for a simplified derivation). Accordingly, the expected total number of different alleles (richness or first moment of the distribution) is given by $m_0 = -N\mu \ln \mu$.

Consider now a pair of loci A and B, and alleles in each locus. We assume that the mutation rates for each gene are low enough so that repeated mutations are rare (i.e., the infinite site model). When combining the two loci (no recombination occurs for now), they would simply behave like one long locus with mutation rate $\mu = \mu_A + \mu_B$ and, therefore, the first moment is simply $m_0(0) = -N(\mu_A + \mu_B) \ln (\mu_A + \mu_B)$. We
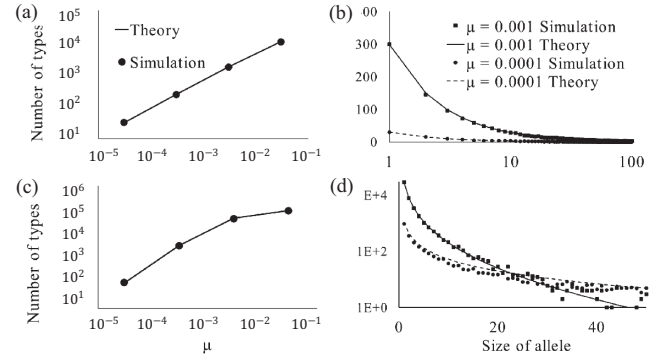

FIG. 1. Comparison between the richness obtained from the simulations and the analytical model for different values of the mutation rate μ. The first row (a) and (b) corresponds to the infinite site model with no recombination and the second row (c) and (d) to the upper bound for two independent loci from Eq. (2). The left column corresponds to the total number of haplotypes and the second to the marginal distributions for the A and B alleles.

observe on Figs. 1(a) and 1(b) that simulations fit these results for the number of haplotypes and the marginal distributions for the A and B alleles.

Let us now introduce a per generation recombination rate r. For a large enough population, the infinite site assumption holds, and one can consider recombination as another type of mutation. This yields an expected number of haplotypes with a maximum value of:

$$m_0(r)_\infty = -N(\mu_A + \mu_B + r) \ln (\mu_A + \mu_B + r). \quad (1)$$

On the opposite extreme case of independent loci, the number of haplotypes in equilibrium can be computed too. It is essential to note that if we were to have infinite mutations, the total number of alleles would be equal to the population since each allele would be distinct. In practice, such mutation rates do not happen, since an error threshold would occur at a finite mutation rate [38]. On the other hand, very high recombination rates can happen (up to 0.5 in the extreme case where one chooses randomly between two loci on different chromosomes). However, since recombination occurs between already existing haplotypes (unlike mutations that create new alleles), new combinations and, therefore, the number of haplotypes would be limited and dependent on the number of alleles and the mutation rates μ_A and μ_B. The allele equilibrium distribution in both loci is not affected by recombination. To compute this upper bound, we now recombine the entire population without mutation as if there were infinite recombination. In practice, this corresponds to a Wright-Fisher process: we randomly choose two individuals and create their offspring with the allele A from the first parent and the allele B from the second parent. The two parents may have the same alleles (see the Appendix for derivations). We obtain:

$$m_{0\text{bound}} = N^2 \mu_A \mu_B \left[\sum_{k=1}^{N} \frac{e^{-\mu_A k}}{k} \ln \left(1 + \frac{k}{\mu_B N} \right) \right]. \quad (2)$$
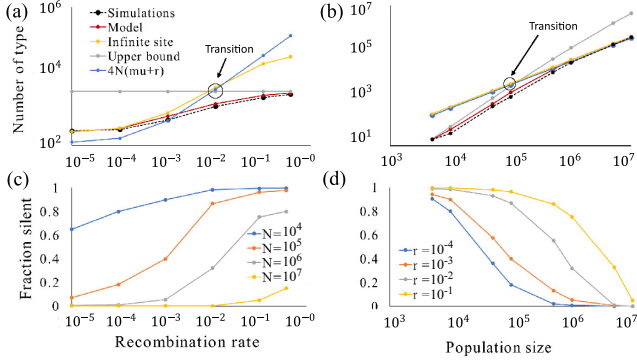
FIG. 2. Plots of the number of haplotypes as a function of (a) the population size and (b) the recombination rate and plots of the ratio of silent recombination as a function of (c) the population size and (d) the recombination rate. The mutation rate is 10^{-4} for allele A and 2×10^{-4} for allele B.

Again, simulations in Figs. 1(c) and 1(d) confirm our results for the expected number of haplotypes and the marginal distributions for the A and B alleles.

The number of haplotypes will always be bounded by the lower of the two extreme cases. We denote the first regime as the infinite site regime and the second one as the bounded regime. For $N \gg 1/\mu$ (or tends to ∞) and r small enough, the number of haplotypes from Eq. (2) is higher than the one given by Eq. (1) and, therefore, the maximum value for the number of haplotypes is given by the infinite site assumption with Eq. (1). For $N \leqslant 1/\mu$ and a high r, the opposite occurs and the maximum value is given by Eq. (2). In the infinite site regime, there are almost no silent recombination whereas, in the bounded regime, the ratio of silent recombination tends to 1 (depending on r) although never reaching 1 as observed in Figs. 2(c) and 2(d).

For a given recombination rate, as N increases, the fraction of silent recombination decreases from almost 1 to almost 0, and the number of haplotype shifts from the bounded regime to the infinite site regime. The higher the recombination rate, the larger N needs to be for this transition to occur [Fig. 2(d)]. For a given population size N, as r increases, most haplotypes are created by recombination. As such, the fraction of silent recombination increases, and the number of haplotypes goes from the infinite site regime to the bounded regime [Fig. 2(c)].

Therefore, for intermediary values of N (between $1/\mu$ and $1000/\mu$) and intermediary values of r (between μ and 0.5), an intermediary regime emerges, and one can expect a mix between the two extreme regimes. We thus performed an interpolation where we compute a log regression with respect to r between the value of the first moment for $r = \mu$ in Eq. (1) and the value of the first moment at the upper bound from Eq. (2).

$$m_0(r)_{\text{interp}} = \begin{cases} -N(\mu_A + \mu_B + r)\ln(\mu_A + \mu_B + r) & r \leqslant \mu \\ \frac{\ln r}{\ln \mu}m_0(\mu) + \left(1 - \frac{\ln r}{\ln \mu}\right)m_{0\text{bound}} & r > \mu \end{cases}.$$
(3)

The interpolation in Eq. (3) slightly overestimates m_0 in the bounded regime but it is much tighter to the simulations than the infinite site estimate as seen in Figs. 2(a) and 2(b) (see the Appendix and Ref. [39] for a description of the simulations).

Finally, since the upper bound from Eq. (2) can be higher than the infinite site model for very large values of N (after the transition), we need to take the minimum of this interpolation and the infinite site model in Eq. (1):

$$m_0(r) = \min\left(m_0(r)_{\text{interp}}, m_0(r)_\infty\right).$$
(4)

In conclusion, given the first moment of the alleles (obtained from the marginal frequencies) and haplotypes frequencies, one can estimate r, as will be further discussed. It already emerges that, for a given value of m_0 and an intermediary value of the population size, the recombination rate obtained from the infinite site model or Watterson's estimator is smaller than the actual one since it does not take into account the silent recombination.

III. RECOMBINATION IN THE HLA COMPLEX

To test the applicability of the boundary above to a real-life system, we analyzed the most polymorphic genes in the human genomes: the major histocompatibility complex (MHC). This locus is denoted as human leukocyte antigen (HLA) in humans, on chromosome 6. This region is of interest since an HLA allele match between donors and recipients is crucial for recipient survival following solid organ or bone marrow transplants [40]. Given its importance, large-scale HLA typing of donors is performed by registries [41–43]. The fraction of patients in a population that can find appropriate donors depends on the frequency of their haplotypes [44]. However, if the recombination rate is high, new haplotypes may be created too fast to allow full coverage.

The HLA gene complex contains the A, C, B, DR, and DQ genes, which together account for over 15000 distinct alleles, and over 10^6 haplotypes. Recent results suggest that a high haplotype creation rate could explain the observed haplotype polymorphism [19,45,46], in contrast with the low current recombination rates estimates [33]. However, current estimates do not incorporate silent recombination and, as such, may be underestimates.

To determine if one can expect a large number of silent recombination in the HLA locus, we analyzed the haplotype frequencies of 6.59×10^6 donors from the National Marrow Donor Program registry, divided into 21 subpopulations [47]. One can estimate N_e, using the marginal distributions of alleles, assuming an infinite site model for the number of alleles in each population. With a mutation rate of 1.45×10^{-8} per base pair [48], yielding an overall mutation rate of $\approx8\times10^{-6}$ per gene (for a length of ≈550 nucleotides per gene), we compute the effective population by inverting the richness formula from the Fisher log series. Then, assuming an equal effective population for recombination and mutations, the recombination rate is computed by inverting Eq. (4) as described in Fig. 3(a). Figure 3(c) shows the recombination rates for the pair of genes A-C across populations (recombination rates for other pairs of genes may be found in the Appendix). Our estimate is 10–1000 times higher than the ones obtained from the infinite site model or Watterson's estimator, which is consistent with the presence of a large fraction of silent recombination.

Another relevant feature from our model is that the recombination rate is not linear with respect to the effective
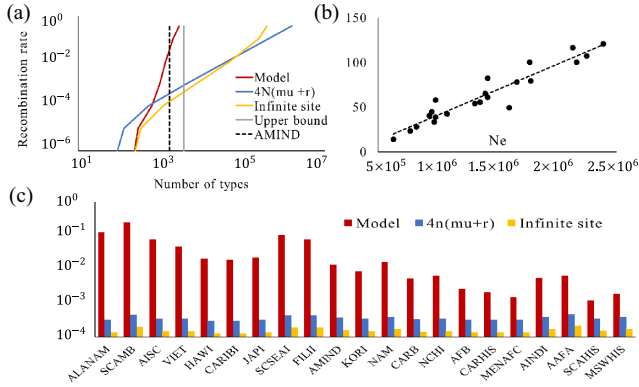
FIG. 3. (a) Recombination rate as a function of the number of haplotypes with different estimators. The red line corresponds to the number of haplotypes for AMIND (Amerindian population). We observe that our estimator yields a higher rate of recombination than the infinite site model or Watterson's estimator. (b) Ratio of the number of haplotypes for the A-C pair and the number of C alleles with respect to the effective population. We observe a linear relation. (c) Recombination rates across populations in the HLA region for the A-C pairs computed with our estimate, the infinite site model, and Watterson's estimator. Our estimator is from 10–1000 times higher.

population. Instead, it has a square term as demonstrated in Eq. (2) (obviously up to some bound). As the population grows, the total number of alleles and haplotypes grows. However, in addition to that, the fraction of nonsilent recombination also increases with the population size inducing a second-order correction term. This is in contrast with the number of mutations, which is linear with respect to the effective population (per definition in the current analysis). One can thus expect that, while the ratio between allele frequencies in different loci should be fixed among populations, the ratio between haplotypes and alleles should be linear in the effective population size.

To validate this claim, we computed the ratio between the number of alleles (for instance, the ratio between the number of alleles in A over the number of alleles in C) and we computed the ratio between the number of haplotypes pairs and the number of alleles (for instance, the number of A-C pairs over the number of C alleles) as a function of the effective population size [Fig. 3(b)]. The latter ratio varies linearly with respect to the effective population, confirming that recombination does not behave like mutations (see the Appendix for a table with all regression coefficients between pairs).

IV. CONCLUSION

The amount of genetic data and detailed haplotype samples have rapidly grown over the last few years. Nevertheless, precise methods to use such samples in order to estimate the recombination rate within haplotypes are still lacking.

We have here proposed a new estimate of the number of haplotypes that incorporates the difference between recombination and mutations. Recombination draws from a pool of existing alleles, some very frequent, and, as such, quite

often, reproduces existing haplotypes. The resulting number of haplotypes is bounded at a level much lower than the total population even for a very high recombination rate. To the best of our knowledge, silent recombination was seldom considered when estimating the recombination rate.

Nevertheless, our estimator, although more accurate than existing models, suffers a few caveats. For the computation of the mutation rate for each allele, we use a neutral infinite site model as is most standard in genetic research, but this may not be the case in all loci [45]. Such a deviation could be due, as is classically argued, to selection [49], or might be the results of other mechanisms, such as catastrophes [39]. For example, in the HLA locus studied here, there is a disparity between the computed and actual distribution (see the Appendix). Besides, we assumed that allele distribution was at equilibrium in order to compute our upper bound on the number of haplotypes, but a very long time is required to reach such an equilibrium, especially for low mutation rates. Finally, the flattening slope of our estimate close to the upper bound might result in large differences in the recombination rate even for small differences in the number of haplotypes. This problem is aggravated if the allele equilibrium is not achieved.

We have here studied a purely neutral model with the type size distribution of a Moran model. Obviously, in the presence of selection, the allele size distribution would differ, and, accordingly, the recombination rate estimate. Such effects will be even more important with epistatic selection, where recombination and selection interfere [50–52]. Still, in the presence of selective sweeps, the fraction of individuals in very large families will actually increase, and, with it, the fraction of silent recombination.

Recombination rates estimates are of interest in the HLA locus, where, although haplotype frequencies are estimated over very large populations, the within haplotype recombination rate is still debated. Most current recombination rate estimates use coalescent models and Watterson's estimator. We analyzed data from the HLA locus and obtained a 10–1000 times higher recombination rate than currently estimated. This difference and the nonlinearity of the number of haplotypes with respect to the effective population size are evidence that recombination cannot be treated as another type of mutation due to the presence of silent recombination. This high haplotype creation rate is in agreement with recent results [19], and it implies that, unless huge surveys are conducted, genome registries will seldom approach an exhaustive list of existing haplotypes.

APPENDIX A: RESULTS DERIVATIONS

1. Mutation model

In a population with different alleles, we define as P_k, the probability for an allele to be of size k. We define α as the birth and death rate (which we assume equal so that the population stays constant) and can be set equal to 1 up to a time scaling. μ is the per generation mutation rate. A type of size k can endure a death at rate α or a birth at rate $\alpha(1 - \mu)$ and not be of size k anymore. A type of size $k - 1$ can have a birth at rate $\alpha(1 - \mu)$ and become a type of size k. Finally, a type of

size $k + 1$ can die at rate α and become a type of size k. The dynamics of P_k is

$$
\frac{dP_k}{dt}
$$
$$
= \alpha[-(2-\mu)kP_k + (1-\mu)(k-1)P_{k-1} + (k+1)P_{k+1}].
$$
(A1)

In steady state, $\frac{dP_k}{dt} = 0$. Also, we assume P_k to be a smooth enough function (in the infinite site model), so, if we denote $P_k = P$, we get:

$$
P_{k+1} = P + P' + 1/2P'' \quad P_{k-1} = P - P' + 1/2P'' \quad \text{(A2)}
$$

$$
0 = -2kP + k\mu P + (k-1)(1-\mu)(P - P' + 1/2P'')
$$
$$
+ (k+1)(P + P' + 1/2P'')
$$
$$
0 = \underbrace{kP'' + 2P'}_{(kP)''} + \underbrace{\mu kP' + \mu P}_{\mu(kP)'} \underbrace{-\mu P' - 1/2\mu(k-1)P''}_{\text{negligible}}
$$
$$
0 = (kP)'' + \mu(kP)'. \quad \text{(A3)}
$$

We define $Q = kP$ and get:

$$
Q'' + \mu Q' = 0 \quad Q' + \mu Q = B \quad (Qe^{\mu k})' = Be^{\mu k}
$$

$$
Qe^{\mu k} = \frac{B}{\mu}e^{\mu k} + A \quad Q = \frac{B}{\mu} + Ae^{-\mu k} \quad Q = Ae^{-\mu k}, \quad \text{(A4)}
$$

$B = 0$ because of the limit condition $\lim_{k \to +\infty} P_k = 0$. So, we finally get:

$$
P_k = \frac{A}{k}e^{-\mu k}. \quad \text{(A5)}
$$

This result is actually the Fisher log series obtained in a simplified way. We define as $\bar{n}$ the average size of a family, N the total population and therefore the average number of types or first moment $m_0 = N/\bar{n}$. μ is considered small in order to make approximations.

$$
\bar{n} = \sum_{k=1}^{\infty} kP_k = A \sum_{k=1}^{\infty} e^{-\mu k}
$$
$$
= \frac{Ae^{-\mu}}{1 - e^{-\mu}} \approx \frac{A(1-\mu)}{\mu} \approx \frac{A}{\mu} \quad \text{(A6)}
$$

$$
1 = \sum_{k=1}^{\infty} P_k = A \sum_{k=1}^{\infty} \frac{e^{-\mu k}}{k}
$$

$$
1 = A \sum_{k=1}^{\infty} \left[\int_{\mu}^{\infty} e^{-\theta k} d\theta \right] = A \int_{\mu}^{\infty} \left[\sum_{k=1}^{\infty} e^{-\theta k} \right] d\theta
$$

$$
1 = A \int_{\mu}^{\infty} \frac{e^{-\theta}}{1 - e^{-\theta}} d\theta = A \int_{e^{-\mu}}^{0} \frac{-x}{1-x} \frac{dx}{x} = A \int_{0}^{e^{-\mu}} \frac{1}{1-x} dx
$$

$$
1 = -A[\ln(1-x)]_0^{e^{-\mu}} = -A\ln(1 - e^{-\mu}) \approx -A \ln \mu \quad \text{(A7)}
$$

$$
\Rightarrow m_0 = \frac{N}{\bar{n}} \approx -N\mu \ln \mu \quad \text{(A8)}
$$

$$
P_k = -\frac{e^{-\mu k}}{k \ln \mu}, \quad N_k = m_0 P_k = \frac{N\mu e^{-\mu k}}{k}. \quad \text{(A9)}
$$

2. Two-locus model with no recombination

In this regime, we assume asexual reproduction and an infinite site model (i.e., each mutation leads to the creation of a new type). The mutation rate in gene A is μ_A and in gene B is μ_B. A birth event leads to four possible outcomes:

(1) No mutation with probability $(1 - \mu_A)(1 - \mu_B)$;
(2) A mutation on gene A with probability $\mu_A(1 - \mu_B)$;
(3) A mutation on gene B with probability $\mu_B(1 - \mu_A)$;
(4) A mutation on both genes with probability $\mu_A\mu_B$.

One individual is randomly selected. In case of no mutation, the corresponding type size is merely increased by 1. In the case of mutation on gene A, a new type is created with size 1. Its A allele is new, and its B allele is the same as the original individual, and vice versa for mutations in B. In the case of a double mutation, both alleles are new. Since genes A and B are independent, we can treat their combination as a single gene, and the mutation rate is, therefore, $\mu = \mu_A + \mu_B$. We already solved this model and obtained the probability P_k for a type to have a size k is given by:

$$
P_k = -\frac{e^{-(\mu_A+\mu_B)k}}{k \ln (\mu_A + \mu_B)}, \quad \text{(A10)}
$$

leading to a total number of types

$$
m_0 \approx -N(\mu_A + \mu_B) \ln(\mu_A + \mu_B). \quad \text{(A11)}
$$

3. Infinite recombination model

In this regime, we assume sexual reproduction and the allele distribution for the two genes are at equilibrium according to the previous model. We randomly select two existing individuals (they can belong to the same type). the newly created individual has the same A allele as the first selected individual and the same B allele as the second. The probability of choosing a type (A_i, B_j) with size $N_{i,j}$ is $p_{i,j} = \frac{N_{i,j}}{N}$. The probability of choosing an individual with a given allele A_i is $p_i = \sum_j \frac{N_{i,j}}{N}$. Similarly, the probability of choosing an individual with allele B_j is $q_j = \sum_i \frac{N_{i,j}}{N}$. There are therefore Np_i individuals with allele A_i and out of those, Np_iq_j individuals of type (A_i, B_j). This implies that $N_{i,j} = Np_iq_j$ and $p_{i,j} = p_iq_j$. We assume that steady state is achieved from the first regime, and A and B alleles are distributed accordingly. This assumption is consistent as seen in Fig. 4. Indeed, simulations where we wait for the alleles to achieve equilibrium or start the recombination process from the beginning yield the same number of haplotypes. We then determine, for each individual, if their A allele is i and if their B allele is j. $\mathbb{E}[N(k)] = \sum_{i,j} \mathbb{P}(N_{i,j} = k)$ is the expected number of types

with size k with $\mathbb{P}(N_{i,j} = k) = \binom{N}{k} p_{i,j}^k (1 - p_{i,j})^{N-k}$,

$$
\mathbb{E}[N(k)] = \binom{N}{k} \sum_{i,j} (p_i q_j)^k (1 - p_i q_j)^{N-k} \approx \binom{N}{k} \sum_{k_A, k_B} N(k_A) N(k_B) \left(\frac{k_A}{N} \frac{k_B}{N} \right)^k \left(1 - \frac{k_A}{N} \frac{k_B}{N} \right)^{N-k}
$$

$$
= \binom{N}{k} m_{0_A} m_{0_B} \sum_{k_A, k_B} P_{k_A} P_{k_B} \left(\frac{k_A}{N} \frac{k_B}{N} \right)^k \left(1 - \frac{k_A}{N} \frac{k_B}{N} \right)^{N-k}
$$

$$
= \binom{N}{k} (N\mu_A \ln(\mu_A))(N\mu_B \ln(\mu_B)) \sum_{k_A, k_B} \frac{e^{-\mu_A k_A}}{k_A \ln(\mu_A)} \frac{e^{-\mu_B k_B}}{k_B \ln(\mu_B)} \left(\frac{k_A}{N} \frac{k_B}{N} \right)^k \left(1 - \frac{k_A}{N} \frac{k_B}{N} \right)^{N-k}
$$

$$
= \binom{N}{k} N^2 \mu_A \mu_B \sum_{k_A, k_B} \frac{e^{-\mu_A k_A}}{k_A} \frac{e^{-\mu_B k_B}}{k_B} \left(\frac{k_A}{N} \frac{k_B}{N} \right)^k \left(1 - \frac{k_A}{N} \frac{k_B}{N} \right)^{N-k} \tag{A12}
$$

$$
\mathbb{E}[N(1)] = N N^2 \mu_A \mu_B \sum_{k_A, k_B} \frac{e^{-\mu_A k_A}}{k_A} \frac{e^{-\mu_B k_B}}{k_B} \left(\frac{k_A}{N} \frac{k_B}{N} \right)^1 \left(1 - \frac{k_A}{N} \frac{k_B}{N} \right)^{N-1} = N\mu_A \mu_B \sum_{k_A, k_B} e^{-\mu_A k_A} e^{-\mu_B k_B} \underbrace{\left(1 - \frac{k_A}{N} \frac{k_B}{N} \right)^{N-1}}_{\approx e^{-\frac{k_A k_B}{N}}}
$$

$$
= N\mu_A \mu_B \sum_{k_A} e^{-\mu_A k_A} \sum_{k_B} e^{-k_B \left(\mu_B + \frac{k_A}{N} \right)} = N\mu_A \mu_B \sum_k \frac{e^{-\mu_B - k \left(\mu_A + \frac{1}{N} \right)}}{1 - e^{-\left(\mu_B + \frac{k}{N} \right)}} \tag{A13}
$$

$$
m_0 = \sum_k \mathbb{E}[N(k)] = \sum_k \binom{N}{k} N^2 \mu_A \mu_B \sum_{k_A, k_B} \frac{e^{-\mu_A k_A}}{k_A} \frac{e^{-\mu_B k_B}}{k_B} \left(\frac{k_A}{N} \frac{k_B}{N} \right)^k \left(1 - \frac{k_A}{N} \frac{k_B}{N} \right)^{N-k}
$$

$$
= N^2 \mu_A \mu_B \sum_{k_A, k_B} \frac{e^{-\mu_A k_A}}{k_A} \frac{e^{-\mu_B k_B}}{k_B} \underbrace{\left[\sum_k \binom{N}{k} \left(\frac{k_A}{N} \frac{k_B}{N} \right)^k \left(1 - \frac{k_A}{N} \frac{k_B}{N} \right)^{N-k} \right]}_{\underbrace{1^N - \left(1 - \frac{k_A}{N} \frac{k_B}{N} \right)^N}_{\approx 1 - e^{-\frac{k_A k_B}{N}}}}
$$

$$
= N^2 \mu_A \mu_B \left[\sum_{k_A} \sum_{k_B} \left[\frac{e^{-\mu_A k_A}}{k_A} \frac{e^{-\mu_B k_B}}{k_B} - \frac{e^{-\mu_A k_A}}{k_A} \frac{e^{-\mu_B k_B}}{k_B} e^{-\frac{k_A k_B}{N}} \right] \right]
$$

$$
= N^2 \mu_A \mu_B \left[\sum_{k_A} \sum_{k_B} \frac{e^{-\mu_A k_A}}{k_A} \frac{e^{-\mu_B k_B}}{k_B} - \sum_{k_A} \sum_{k_B} \frac{e^{-\mu_A k_A}}{k_A} \frac{e^{-k_B \left(\mu_B + \frac{k_A}{N} \right)}}{k_B} \right]
$$

$$
= N^2 \mu_A \mu_B \left[\underbrace{\left(\sum_{k_A} \frac{e^{-\mu_A k_A}}{k_A} \right)}_{\approx -\ln \mu_A} \underbrace{\left(\sum_{k_B} \frac{e^{-\mu_B k_B}}{k_B} \right)}_{\approx -\ln \mu_B} - \sum_{k_A} \frac{e^{-\mu_A k_A}}{k_A} \underbrace{\left(\sum_{k_B} \frac{e^{-k_B \left(\mu_B + \frac{k_A}{N} \right)}}{k_B} \right)}_{\approx -\ln \left(\mu_B + \frac{k_A}{N} \right)} \right]
$$

$$
= N^2 \mu_A \mu_B \left[\ln \mu_A \ln \mu_B + \sum_{k_A} \frac{e^{-\mu_A k_A}}{k_A} \ln \left(\mu_B + \frac{k_A}{N} \right) \right]
$$

$$
= N^2 \mu_A \mu_B \left[\ln \mu_A \ln \mu_B + \sum_{k_A} \frac{e^{-\mu_A k_A}}{k_A} \left[\ln \mu_B + \ln \left(1 + \frac{k_A}{\mu_B N} \right) \right] \right]
$$

$$
= N^2 \mu_A \mu_B \left[\ln \mu_A \ln \mu_B + \ln \mu_B \underbrace{\left(\sum_{k_A} \frac{e^{-\mu_A k_A}}{k_A} \right)}_{\approx -\ln \mu_A} + \sum_{k_A} \frac{e^{-\mu_A k_A}}{k_A} \ln \left(1 + \frac{k_A}{\mu_B N} \right) \right] = N^2 \mu_A \mu_B \left[\sum_k \frac{e^{-\mu_A k}}{k} \ln \left(1 + \frac{k}{\mu_B N} \right) \right]. \tag{A14}
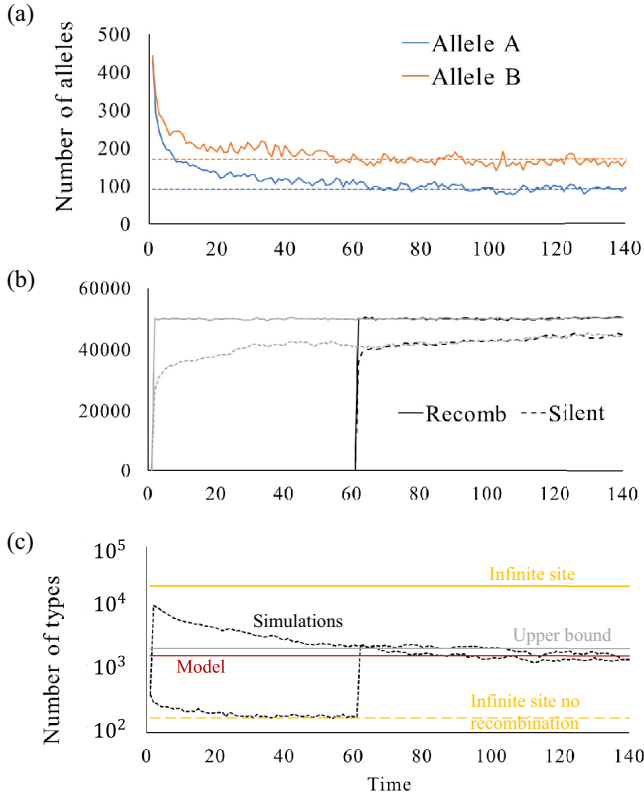$$

(a)

(b)

(c)

FIG. 4. Time series of the simulations. The final equilibrium is the same whether the recombination process started from time 0 or after a lag. The mutation rate is 10^{-4} for the A alleles and 2×10^{-4} for the B alleles. The number of steps is 1.4×10^9. The recombination rate is 0.1. (a) represents the time series for the number of alleles for genes A and B. (b) represents the time series of the number of recombination and silent recombination. (c) represents time series of the simulations of the number of types.

In Eq. (A14), the expected number of haplotypes is actually different from $N^2 \mu_A \mu_B \ln(\mu_A) \ln(\mu_B)$, which represents the product of all combinations of A and B alleles, since, even in the steady state, not all possible pairs of alleles are present. Just as in Eq. (A7), the finite sum can be approximated by the logarithm since $\mu \ll 1$ and N is large so the remainder of the sum is negligible. Although Eq. (A14) does not seem to be symmetric with respect to gene A and B due to the approximations we performed in the computation, inverting μ_A and μ_B would yield almost identical results as can be observed in Fig. 5.

4. Intermediary regime

a. Lower bound interpolation

We tested whether a regular birth-and-death process with mutations and recombination could be modeled as a combination of the two regimes described above and if the ratio between those two regimes is linked to the recombination rate r. We approximate that the creation of new types
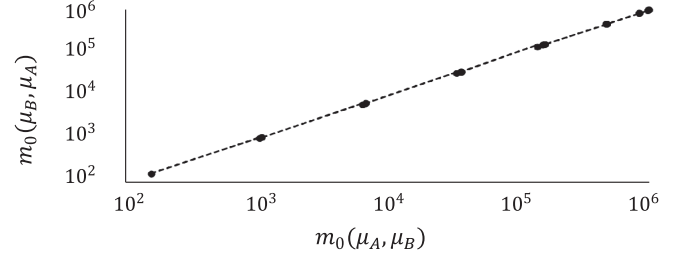

FIG. 5. Computation of the expected number of haplotypes in the infinite recombination model for different values of μ_A and μ_B (dots). The x axis represents m_0 from Eq. (A14) and the y axis is obtained when inverting μ_A and μ_B. The dots are on the first diagonal confirming that Eq. (A14) is symmetric.

(with a size of 1) comes from two sources: either the recombination regime or the no recombination regime. The extinction of types of size 1 comes from the mixed regime. At equilibrium, we assume creations and extinctions are equal. We denote by 0 the no-recombination regime, by 1 the infinite-recombination regime, and by r the mixed regime. P_1 is the probability for a type of size 1, m_0 is the number of types, and m_1 is the total population. We obtain:

$$m_0(0) = -N(\mu_A + \mu_B) \ln(\mu_A + \mu_B) \qquad (A15)$$

$$P_1(0) = -\frac{e^{-(\mu_A + \mu_B)}}{\ln(\mu_A + \mu_B)} \qquad (A16)$$

$$m_0(1) = N^2 \mu_A \mu_B \left[\sum_{k=1}^{\infty} \frac{e^{-\mu_A k}}{k} \ln\left(1 + \frac{k}{\mu_B N}\right) \right] \qquad (A17)$$

$$P_1(1) = \frac{\mathbb{E}[N(1)]}{m_0(1)}. \qquad (A18)$$

The number of creations of new types is given by:

$$r \frac{m_0(1) P_1(1)}{m_1} + (1 - r) \frac{m_0(0) P_1(0)}{m_1}. \qquad (A19)$$

The number of extinctions of types of size 1 is given by:

$$\frac{P_1(r) m_0(r)}{m_1}. \qquad (A20)$$

At equilibrium, creations and extinctions rates are equal. This yields:

$$m_0(r)_{int1} = \frac{r m_0(1) P_1(1) + (1 - r) m_0(0) P_1(0)}{P_1(r)}. \qquad (A21)$$

We also assume that $P_1(r)$ is a linear combination of $P_1(0)$ and $P_1(1)$:

$$P_1(r) = (1 - r) P_1(0) + r P_1(1). \qquad (A22)$$

b. Upper bound interpolation

We also developed another interpolation between the two extreme regimes of no recombination and infinite recombination. In this version, we simply compute a log regression between the value of m_0 for $r = \mu$ given by $m_0(\mu) = -2N(\mu_A + \mu_B) \ln[2(\mu_A + \mu_B)]$ and $m_0(1)$ from Eq. (A14). This yields:

$$m_0(r)_{int2} = \begin{cases} -N(\mu_A + \mu_B + r)\ln(\mu_A + \mu_B + r) & r \leqslant \mu \\ \frac{\ln r}{\ln \mu} m_0(\mu) + \left(1 - \frac{\ln r}{\ln \mu}\right)m_0(1) & r > \mu \end{cases}.$$
(A23)

c. Mixed model

The upper interpolation is quite a good fit to our simulation. Nevertheless, in some cases it slightly overestimates the actual number of types m_0. Since the first interpolation slightly underestimates m_0, we took the average of those two and achieve an even better fit from our simulations. This average, although heuristic, provides a more accurate estimate of the first moment since both aforementioned interpolations are actually close to each other compared to the upper bounds. An intermediate estimate is therefore, a natural choice and the average is the simplest one. We tested other ratios and yield similar results.

$$m_0(r) = \frac{m_0(r)_{int1} + m_0(r)_{int2}}{2}.$$
(A24)

APPENDIX B: SIMULATIONS

In the simulations, the population is composed of several haplotypes resulting from the combinations of the alleles from genes A and B. We assume birth and death rates to be equal so that the population is constant. For simplicity, at each step, a birth and a death event occur. Each birth results either in a regular birth where a haplotype simply increases its size by 1, or a mutation in A or B alleles or in both, hence creating a new type of size 1, or finally, recombination where the offspring gets its A allele from one parent and its B allele from the other. We compute the normalized probabilities for all those events and randomly choose which will occur. For the sake of efficiency, all the initial haplotypes (and thereafter all haplotypes) are plugged into a tree, in order to keep track of each haplotype size. Each leaf corresponds to a haplotype and the number associated with this entry is the haplotype size. The value of each internal node in the tree is the sum of its two sons, the tree root being the size of the total population. We run those simulations for a number of steps sufficiently large so that we achieve a steady state.

APPENDIX C: EFFECTIVE POPULATION

Knowing the actual number of haplotypes m_0 for a population, we can numerically invert Eq. (A24) and compute r.
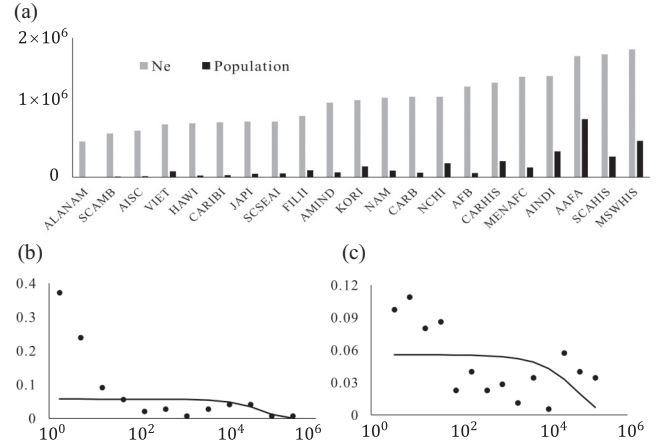

FIG. 6. (a) represents the respective population and effective population. Distribution of the alleles (dots). The full line represents the theoretical distribution with $\mu = 8 \times 10^{-6}$. (b) is for the gene C in the HIS population. (c) is for the gene B in the KORI population.

We first need to compute the effective population N_e by inverting Eq. (A8). Therefore, we also need μ independently of N_e. One way to do it is to use the probabilities P_k from Eq. (A9).

We assume an infinite site model, P_k is continuous. We design bins around k with length Δ_k. Therefore, the number of alleles N_k with a size comprised in that bin is:

$$N_k = P_k \Delta_k m_0.$$
(C1)

Taking the log on both sides and using Eq. (A5), we get:

$$-\ln\left[\frac{kN_k}{\Delta_k m_0}\right] = \mu k + \ln[|\ln(\mu)|].$$
(C2)

Since for each population, we have the frequencies of all the haplotypes, we determine the marginal distribution for a given gene and compute N_k for each bin. We then perform a linear regression to get μ.

One problem we encountered is that the actual distributions have fat tails (power-law distributions with an exponential cutoff) and, therefore, deviate from our estimate as in Figs. 6(b) and 6(c). Thus, we use the mutation rate from an external source computed using statistics on the genes. We get the effective populations in Fig. 6(a).

APPENDIX D: HLA COMPLEX

We demonstrated that recombination can be assumed to behave like mutations only for low rates and large populations. For mutations, the number of alleles is linear with respect to

TABLE I. Linear regression coefficients between gene pair and haplotype ratios.

Ratio	Slope	p value
m_{0_A}/m_{0_C}	0.2286	0.1236
m_{0_C}/m_{0_B}	-0.3717	7.1953×10^{-2}
$m_{0_B}/m_{0_{DR}}$	-0.2187	0.2255
$m_{0_{DR}}/m_{0_{DQ}}$	0.74615	1.9447×10^{-6}
$m_{0_{A,C}}/m_{0_C}$	0.4270	2.8047×10^{-2}
$m_{0_{C,B}}/m_{0_B}$	0.5715	1.1130×10^{-3}
$m_{0_{B,DR}}/m_{0_{DR}}$	0.6226	1.2045×10^{-5}
$m_{0_{DR,DQ}}/m_{0_{DQ}}$	0.9334	4.3806×10^{-14}

the effective population. If we assume that the effective population is the same for each gene, then the ratio between the number of alleles in two loci should be constant. On the other hand, according to our estimate, the number of haplotypes is not linear in N_e. Therefore, the ratio of the number of haplotypes and alleles varies among populations. Table I represents the slope coefficients for linear regression and their p values. The slopes for the ratio of the number of alleles are close to 0 with high p values, confirming that those ratios do not depend on N_e, whereas the slopes for the ratio of haplotypes over alleles are different from 0 with very small p-values. Figure 7 shows the recombination rates across populations in the HLA region computed with our estimate, the infinite site model, and Watterson's estimator, for the pair of adjacent genes C-B, B-DR, and DR-DQ.
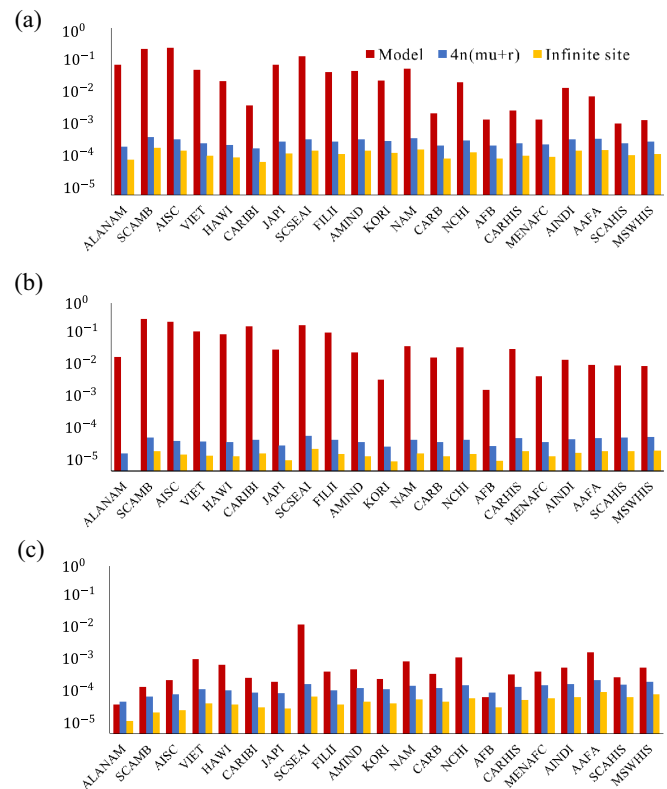

FIG. 7. Recombination rates across populations in the HLA region computed with our estimate, the infinite site model, and Watterson's estimator, for the pair of adjacent genes (a) C-B, (b) B-DR, and (c) DR-DQ (c).

[1] P. Moran, Random processes in genetics, Math. Proc. Camb. Philos. Soc. **54**, 60 (1958).

[2] M. Kimura and J. Crow, The number of alleles that can be maintained in a finite population, Genetics **49**, 725 (1964).

[3] S. Wright, *Evolution and the Genetics of Populations: Theory of Gene Frequencies* (University of Chicago Press, 1984), Vol. 2.

[4] W. Ewens, The sampling theory of selectively neutral alleles, Theor. Popul. Biol. **3**, 87 (1972).

[5] G. Watterson, On the number of segregating sites in genetical models without recombination, Theor. Popul. Biol. **7**, 256 (1975).

[6] M. Kimura, On the probability of fixation of mutant genes in a population, Genetics **47**, 713 (1962).

[7] S. Krone, M. Stephen, and C. Neuhauser, Ancestral processes with selection, Theor. Popul. Biol. **51**, 210 (1997).

[8] J. F. C. Kingman, On the genealogy of large populations, J. Appl. Prob. **19**, 27 (1982).

[9] G. Constable and A. McKane, Stationary solutions for metapopulation moran models with mutation and selection, Phys. Rev. E **91**, 032711 (2015).

[10] M. Meselson and C. Radding, A general model for genetic recombination, Proc. Natl. Acad. Sci. USA **72**, 358 (1975).

[11] N. Barton, A general model for the evolution of recombination, Genet. Res. **65**, 123 (1995).

[12] R. Hudson, Properties of a neutral allele model with intragenic recombination, Theor. Popul. Biol. **23**, 183 (1983).

[13] M. Neale and S. Keeney, Clarifying the mechanics of DNA strand exchange in meiotic recombination, Nature (London) **442**, 153 (2006).

[14] J. San Filippo, P. Sung, and H. Klein, Mechanism of eukaryotic homologous recombination, Annu. Rev. Biochem. **77**, 229 (2008).

[15] W. Heyer, K. Ehmsen, and J. Liu, Regulation of homologous recombination in eukaryotes, Annu. Rev. Genet. **44**, 113 (2010).

[16] A. Aarnink, E. Mee, N. Savy, N. Congy-Jolivet, N. Rose, and A. Blancher, Deleterious impact of feto-maternal MHC compatibility on the success of pregnancy in a macaque model, Immunogenetics **66**, 105 (2014).

[17] G. Coop and M. Przeworski, An evolutionary view of human recombination, Nat. Rev. Genet. **8**, 23 (2007).

[18] E. Yunis and D. Amos, Three closely linked genetic systems relevant to transplantation, Proc. Natl. Acad. Sci. USA **68**, 3031 (1971).

[19] A. Lobkovsky, L. Levi, Y. Wolf, I. Alter, L. Gragert, M. Martin, Y. Louzoun, and E. Koonin, Multiplicative fitness, rapid haplotype discovery, and fitness decay explain evolution of human MHC, Proc. Natl. Acad. Sci. USA **116**, 14098 (2019).

[20] M. Cullen, S. Perfetto, W. Klitz, G. Nelson, and M. Carrington, High-resolution patterns of meiotic recombination across the

human major histocompatibility complex, Am. J. Hum. Genet. **71**, 759 (2002).

[21] G. McVean, S. Myers, S. Hunt, P. Deloukas, D. Bentley, and P. Donelley, The fine-scale structure of recombination rate variation in the human genome, Science **304**, 581 (2004).

[22] T. Morgan, Developmental mechanics, Science **7**, 156 (1898).

[23] A. Begovich, G. McClure, V. Suraj, R. Helmuth, N. Fildes, T. Bugawan, H. Erlich, and W. Klitz, Polymorphism, recombination, and linkage disequilibrium within the HLA class II region, J. Immunol. **148**, 249 (1992).

[24] R. Hudson, Estimating the recombination parameter of a finite population model without selection, Genet. Res. **50**, 245 (1987).

[25] Y. Fu and W. Li, Maximum likelihood estimation of population parameters, Genetics **134**, 1261 (1993).

[26] R. Hudson and N. Kaplan, The coalescent process in models with selection and recombination, Genetics **120**, 831 (1988).

[27] G. McVean, P. Awadalla, and P. Fearnhead, A coalescent-based method for detecting and estimating recombination from gene sequences, Genetics **160**, 1231 (2002).

[28] P. Fearnhead and P. Donnelly, Estimating recombination rates from population genetic data, Genetics **159**, 1299 (2001).

[29] M. Stumpf and G. McVean, Estimating recombination rates from population-genetic data, Nat. Rev. Genet. **4**, 959 (2003).

[30] A. Tenesa, P. Navarro, B. Hayes, D. Duffy, G. Clarke, M. Goddard, and P. Visscher, Recent human effective population size estimated from linkage disequilibrium, Genome Res. **17**, 520 (2007).

[31] P. Leberg, Genetic approaches for estimating the effective size of populations, J. Wildl. Manage. **69**, 1385 (2005).

[32] M. Kuhner, J. Yamato, and J. Felsenstein, Maximum likelihood estimation of recombination rates from population data, Genetics **156**, 1393 (2000).

[33] A. Blancher, A. Aarnink, N. Savy, and N. Takahata, Use of cumulative Poisson probability distribution as an estimator of the recombination rate in an expanding population example of the Macaca fascicularis major histocompatibility complex, G3: Genes Genom. Genet. **2**, 123 (2012).

[34] A. Ramírez-Soriano and R. Nielsen, Correcting estimators of θ and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process, Genetics **181**, 701 (2009).

[35] J. Felsenstein, Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci?, Mol. Biol. Evol. **23**, 691 (2006).

[36] Y. Shem-Tov, M. Danino, and N. M. Shnerb, Solution of the spatial neutral model yields new bounds on the amazonian species richness, Sci. Rep. **7**, 42415 (2017).

[37] R. A. Fisher, A. S. Corbet, and C. B. Williams, The relation between the number of species and the number of individuals

in a random sample of an animal population, J. Anim. Ecol. **12**, 42 (1943).

[38] J. Summers and S. Litwin, Examining the theory of error catastrophe, J. Virol. **80**, 20 (2006).

[39] A. Melka, N. Dori, and Y. Louzoun, Invasion Rate Versus Diversity in Population Dynamics with Catastrophes, Phys. Rev. Lett. **124**, 158301 (2020).

[40] B. Bradley, The role of HLA matching in transplantation, Immunol. Lett. **29**, 55 (1991).

[41] P. G. Beatty, M. Mori, and E. Milford, Impact of racial genetic polymorphism on the probability of finding an HLA-matched donor, Transplantation **60**, 778 (1995).

[42] S. Spellman *et al.*, Advances in the selection of HLA-compatible donors: Refinements in HLA typing and matching over the first 20 years of the National Marrow Donor Program Registry, Biol. Blood Marrow Transplant. **14**, 37 (2008).

[43] L. Gragert, A. Madbouly, J. Freeman, and M. Maiers, Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry, Human Immunol. **74**, 1313 (2013).

[44] N. Slater, Y. Louzoun, L. Gragert, M. Maiers, A. Chatterjee, and M. Albrecht, Power laws for heavy tailed distributions: Modeling allele and haplotype diversity for the national marrow donor program, PLoS Comput. Biol. **11**, e1004204 (2015).

[45] I. Alter, L. Gragert, S. Figerson, M. Martin, and Y. Louzoun, HLA class I haplotype diversity is consistent with selection for frequent existing haplotypes, PLoS Comput. Biol. **13**, e1005693 (2017).

[46] A. Simanovsky, A. Madbouly, M. Halagan, M. Maiers, and Y. Louzoun, Single haplotype admixture models using large scale HLA genotype frequencies to reproduce human admixture, Immunogenetics **71**, 589 (2019).

[47] S. Israeli, L. Gragert, M. Maiers, and Y. Louzoun, HLA haplotype frequency estimation for heterogeneous populations using a graph-based imputation algorithm, Human Immunol. **82**, 746 (2021).

[48] V. M. Narasimhan, R. Rahbari, A. Scally *et al.*, Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes, Nat. Commun. **8**, 303 (2017).

[49] P. Hedrick and G. Thomson, Evidence for balancing selection at HLA, Genetics **104**, 449 (1983).

[50] R. Neher and B. Shraiman, Competition between recombination and epistasis can cause a transition from allele to genotype selection, Proc. Natl. Acad. Sci. USA **106**, 6866 (2009).

[51] D. B. Weissman and N. H. Barton, Limits to the rate of adaptive substitution in sexual populations, PLoS Genet. **8**, e1002740 (2012).

[52] T. Held, D. Klemmer, and M. Lässig, Survival of the simplest in microbial evolution, Nat. Commun. **10**, 2472 (2019).