A precise estimate of allele and haplotype polymorphism is of great interest for theoretical population genetics, but also practical issues, such as bone marrow registries. Allele polymorphism is driven mainly by point mutations, while haplotype polymorphism is also affected by recombination events. Even in the simple case of two loci in a haploid individual, there is currently no good estimate of the number of haplotypes as a function of the mutation and recombination rates. We here propose such an estimate and show that the common approximation that recombination can be treated as mutations is limited to recombination rates of the same order as the mutation rate. Beyond this regime, the total number of haplotypes is much lower than expected from the approximation above. Moreover, in contrast with mutations, the number of haplotypes does not grow linearly with the population size. We apply this new estimate to very large-scale human haplotype frequencies from human populations to show that the current estimated haplotype recombination rate in the HLA region is underestimated. This high recombination rate may be the source of HLA haplotype extreme polymorphism.

Multiple genetics models relate allele frequencies to their populations’ dynamics \([1,2]\). Such models typically include the processes of mutation, genetic drift, selection, and migration between sub-populations \([0,9]\). For the haplotype frequencies, another essential process to consider is recombination \([10,12]\). Meiotic recombination is a process starting with the formation of DNA double-strand breaks (DSB) \([13]\) by which chromosomes are recombined to produce new allele combinations during the cell division occurring in sexually reproducing eukaryotes. In this process, crossovers can result in the exchange of genetic material between the maternal and paternal homologous chromosomes \([14,16]\). As a result, offspring can have different combinations of genes than their parents on the same chromosome, and new haplotypes are created from the combination of existing alleles. Determining the recombination rate in the genealogical history of a sample of haplotypes is crucial in evolutionary biology and medical population genetics \([17,18]\). It also has important implications for transplant donors registry management \([19,20]\). Recombination events can lead to the formation of new haplotypes and the resulting increase in genetic variability. In most species, recombination is concentrated in narrow regions known as hot spots, 1 to 2 kilobases in length, flanked by large zones with low recombination rates or cold regions \([21,22]\).

Two approaches have been proposed to build recombination maps and estimate the recombination rate. The first approach, referred to as the direct approach, is strictly experimental and consists of sperm genotyping. In this method, crossover sites on DNA with nucleotide resolution can be detected. However, it cannot be used for genome-wide analyses \([23]\). It requires a sufficient quantity of cells, therefore, human male semen cells are usually used. From those cells, one can either observe locations where the chromosome broke and recombination occurred using, for instance, fluorescence \([24]\), illumina sequencing \([25]\), or identifying binding proteins \([26]\) and their binding sites \([27,28]\).

The second approach is an indirect method that uses genetic linkage (co-inheritance of markers in families) to produce recombination maps for chromosome segments \([30]\). These segments can then be linked to provide estimates of recombination frequencies for specific chromosomes. The most detailed maps of human recombination have been generated through computational inference of recombination rates from patterns of linkage disequilibrium (LD) in the human population \([31]\). Such maps do not provide information about recombination rates in individuals. Recombination rates can be obtained experimentally by performing whole genome sequencing, or single nucleotide polymorphism (SNP) typing of pedigrees of at least two to three generations leading to a resolution in the order of tens of kilobases \([32]\), or inferring recombination hot spots through patterns of linkage disequilibrium (LD) \([33]\). The latter approach provides genome-wide historical recombination events in both males and females inferred from polymorphism characterized in many individuals within a population. A technical difference between the two approaches is that the recombination rate computed in the direct approach is per cell division, whereas the rate computed in the second is per generation.

The indirect approach uses tools such as maximum likelihood estimation (MLE) \([34,35]\) usually based on a coalescent tree model \([36,37]\). It is commonly assumed that recombination behaves like mutations. In a single gene, fixed population, neutral model, the mutation rate has been related to the number of alleles through \(\theta = 4N_e \mu\) (the overall mutation rate for the population, where \(N_e\) is the effective population as defined by Kimura \([2]\), and \(\mu\) is the individual mutation rate). This formula
was first derived by Watterson to describe mutations [3]. 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 \( \rho = 4N_e r \).

The validity of this approach is limited by the need to determine the effective population \( N_e \). [40][41]. Since this estimate is not reliable, studies either display the ratio between recombination and mutations \([42]\) or simply compute \( \rho \) and \( \theta \) instead of \( \mu \) and \( r \), separately or jointly, using MLE \([47][48]\). Other studies use samples for which the origin of the population is known \([49]\). Assuming that the effective population is the same in recombination and mutations, one can estimate \( N_e \) by first determining the number of alleles and the mutation rate (using statistics on the DNA) and then inverse Watterson’s estimator. Note that the number of alleles differs from Watterson’s estimate. Indeed, multiple corrections were proposed \([50][51][52]\). Another limitation, which is typically less of a problem, is that, to estimate \( \theta \) or \( \rho \) using the number of different alleles or haplotypes, the sample has to be relatively small compared with the effective population. The last and most significant limitation is that the assumption that recombination behaves like mutations in the infinite site model fails when the recombination rate is large enough. Specifically, given the fat tail of the allele distributions, recombination often reproduces the same combinations of alleles, as such the recombination rate expected from Watterson’s formula is largely underestimated.

One way to remedy those caveats is to use statistical models on the observed alleles and haplotype distributions and infer the mutation and recombination rates. We here propose that one can relate the number of different two-locus haplotypes in a population (the haplotype polymorphism), with the recombination rate within this haplotype in a generation (i.e. the probability that offspring would inherit a recombinant haplotype), using a revised relationship between the number of alleles and haplotypes and the mutation rate and the recombination rate in each of the two loci. To this end, we use a birth and death stochastic process \([40]\) with mutations and recombination, where one can study the frequencies of alleles and haplotypes simultaneously, for large populations and over a long period. We identify three regimes depending on the order of magnitude of the recombination rate compared with the mutation rate. In the first regime, for recombination rates close to the mutation rate, recombination behaves like mutations and can be treated as such in the stochastic process. For very high recombination rates, the infinite sites assumption does not hold, and the number of haplotypes reaches an upper bound. In between those two regimes, one observes an intermediary phase. We use this model to analyze large-scale populations and detailed information on their haplotype frequency distributions to infer the recombination rate in the HLA locus. We show that the recombination rate is much higher than previously estimated.

**Single Locus and two loci 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 \([40]\) 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 \( \mu \) as the mutation rate and \( P_k \) as the probability for an allele to have a population of size \( k \) and get \( P_k = \frac{e^{-\mu k}}{k!} \ln \mu \) (see Supplemental Material for derivations). The resulting expected total number of different alleles (richness or first moment of the distribution \( m_0 \)) is given by:

\[
m_0 = -N \mu \ln \mu.
\]  

Consider now a pair of loci A and B, and alleles in each locus. We first 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 (again, 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 observe on Fig. 1(a) and (b) that simulations fit our results for the number of haplotypes and the marginal distributions for alleles in A and B.

Now introduce a recombination rate \( r \) in each generation. When the recombination rate is of the order of the mutation rate and both are low, the infinite site assumption still holds, the two loci are almost coupled, and one can treat recombination as another type of mutation. This yields a total number of haplotypes equal to:

\[
m_0(r) = -N(\mu_A + \mu_B + r) \ln (\mu_A + \mu_B + r).
\]  

In 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 \([47]\). On the other hand, infinite recombination can happen, for example between two loci on different chromosomes. Nevertheless, since recombination occurs between already existing haplotypes (unlike mutations that create new alleles), the creation of new combinations and, therefore, also the number of haplotypes would be limited and dependent on the number of alleles and the mutation rates \( \mu_A \) and \( \mu_B \). The allele equilibrium distribution in both loci is not affected by recombination. To compute this upper bound, we now recombine the entire population
At equilibrium, we assume creations and extinctions are equal (see the Supplemental Material for derivations). We approximate that the creation of new families of types and the second to the marginal distributions for the alleles in A and B.

Without mutation as if there were infinite recombination \((r \approx 1)\). 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 Supplemental Material for derivations). We obtain:

\[
m_0(1) = N^2 \mu_A \mu_B \left[ \sum_{k=1}^{N} e^{-\mu_A k} \frac{k}{\mu_B N} \right]. \tag{3}
\]

Again, as seen in Fig. 1, simulations confirm our results for the total number of haplotypes and the marginal distributions. The expected number of haplotypes can, therefore, follow one of three regimes. For \(r\) in the order of \(\mu\), the first moment is given by Eq. (2). When \(r\) tends to 1, the first moment tends to the upper bound in Eq. (3). When \(r\) is greater than \(\mu\) but still smaller than 1, we observe an intermediary regime where the total number of alleles parts from the model where recombination behaves like mutations and tends to the upper bound. One can expect a mix between the two extreme regimes. We thus propose an interpolation between these two regimes with a ratio linked to the recombination rate \(r\). We approximate that the creation of new families of size 1 comes from two sources: either the no recombination regime or the infinite recombination regime. The extinction of types of size 1 comes from the mixed regime.

At equilibrium, we assume creations and extinctions are equal (see the Supplemental Material for derivations).

\[
m_0(r)_{\text{int1}} = \frac{rm_0(1)P_1(1) + (1-r)m_0(0)P_1(0)}{P_1(r)} \tag{4}
\]

As mentioned above, the allele distribution has a fat tail and, therefore, not all recombination event leads to the creation of a new type but rather to an already existing type, especially as \(r\) increases (i.e. the infinite sites assumption does not hold). Therefore, this interpolation underestimates the first moment. We thus performed another interpolation where we simply compute a log regression with respect to \(r\) between the value of the first moment for \(r = \mu\) and the value of the first moment at the upper bound from Eq. 3.

\[
m_0(r)_{\text{int2}} = \begin{cases} 
-N(\mu_A + \mu_B + r) \ln(\mu_A + \mu_B + r) & r \leq \mu \\
\frac{\ln r}{\ln \mu} m_0(\mu) + \left(1 - \frac{\ln r}{\ln \mu}\right) m_0(1) & r > \mu
\end{cases} \tag{5}
\]

The second interpolation provides an upper estimate as seen in Fig. 2. As a result we compute the first moment in the intermediary regime as the average between those two interpolations:

\[
m_0(r) = \frac{m_0(r)_{\text{int1}} + m_0(r)_{\text{int2}}}{2} \tag{6}
\]

We confirm our hypothesis against simulations (see Fig. 2). In those simulations, the population is comprised 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 on allele A or B or both, hence creating a new type of size 1, or finally, recombination where the offspring gets its allele A from one parent and its allele B from another. 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. For more details on this methodology, see [48]. We run those simulations for a number of steps sufficiently large so that we achieve a steady state. In conclusion, given the first moment of the alleles (obtained from the marginal frequencies) and haplotypes frequencies, one can estimate \(r\), as shall be further discussed.

Recombination in the HLA Complex.—The most polymorphic genes in the human genomes are the ones coding the major histocompatibility complex (MHC). This locus is denoted as HLA (Human Leukocyte Antigen) in humans, on chromosome 6. It is of interest since a match
between donors and recipients is crucial for recipient survival following solid organ or bone marrow transplants [49]. Given its importance, large-scale donor registries were developed [50–52]. The purpose of these registries is to find fully matched bone marrow donors to any recipient. The population coverage of such registries depends on the haplotype coverage [53]. However, if the recombination rate is high, new haplotypes may be created too fast to allow full coverage of the population at current recruitment levels, or ever (unless obviously a large fraction of the entire human population is recruited as donors).

The HLA locus has been for a long time the hallmark of balancing selection, where the high HLA polymorphism has been argued to result from different mechanisms favoring heterozygosity. However, recent results, based on detailed haplotype frequencies, suggest that an alternative explanation, more consistent with current observations would be a high haplotype creation rate [20, 54, 55]. Nevertheless, this contradicts current estimates of recombination rates of ≈0.008 ± 0.004 base pair [56]. We then invert Eq. 1 and obtain the effective population (see Supplemental Material). The recombination rate is then computed by inverting Eq. 6. Plot (a) in Fig. 3 shows the sum of the recombination rates for the four pairs of genes (since the rates are small, the overall rate can be approximated by the sum). Plot (b) shows the average rates across populations for all four gene pairs for all populations. The values obtained for the recombination rates are on average more than ten times higher than the rates estimated with an infinite site model as seen in Table I.

One relevant feature from our model is that the recombination rate is not linear with respect to the effective population. Instead it has a square term as demonstrated in Eq. 3. This is in contrast with the mutation rate, which is linear with respect to the effective population. One can thus expect that, while the ratio between allele frequencies in different loci should be fixed among populations, the ratio between haplotypes and

![FIG. 2. Plots of the number of haplotypes as a function of the recombination rate for mutation rates 10^{-4} and 10^{-5} for alleles A and B. The dots correspond to the simulations and the full line to our model.](image)

![FIG. 3. Plot (a) - Recombination rates across populations in the HLA region computed as the sum of the four recombination rates for each pair of adjacent genes. The dotted line corresponds to the rates computed from an infinite site model and extracting r solely from Eq. 2. The full line corresponds to the rates computed from Eq. 6. We observe high variability across populations. Plot (b) - Average recombination rates between adjacent genes. The grey rectangles correspond to the rates computed from an infinite site model and extracting r solely from Eq. 2. The black rectangles correspond to the rates computed from Eq. 6 with much higher values. Plot (c) - Ratio of the number of haplotypes for a given pair of genes and the number of alleles with respect to the effective population. We observe a linear relation.](image)
alleles should be linear. To test that, we computed the ratio between the number of alleles pairs (for instance, the ratio between the number of alleles in A and C) and the ratio of the number of haplotypes and alleles (for instance, the number of AC haplotypes divided by the number of C alleles), for each pair of HLA genes in the different populations. Plot (c) in Fig. 3 shows that the latter ratio varies (linearly) with respect to the effective population, confirming that recombination does not behave like mutations (see Supplemental Material for a table with all regression coefficients between pairs).

Conclusions.—The amount of genetic data and the number of 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. This is especially true 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-based models and Watterson’s estimator.

We have here proposed a new estimate based on the difference between recombination and mutations in the infinite site model. Recombination draws from a pool of existing alleles, some very frequent, and, as such, quite often, it 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. We applied this estimator to the HLA locus and obtained a ten-time higher recombination rate than currently believed. This high haplotype creation rate is in agreement with recent results [20], and it implies that, unless huge surveys are conducted, genome registries will seldom approach an exhaustive list of existing haplotypes.

One major caveat in our approach is the computation of the mutation rate for each allele. Indeed, we use a neutral model with infinite sites as is most standard in genetic research, but the data clearly show that the distribution of alleles among populations deviates from a neutral model [54]. This could be due, as is classically argued, to balancing selection [57], or might be the results of other mechanisms, such as catastrophes [18]. There is, therefore, a disparity between the computed and actual distribution disabling us from computing the mutation rate (see the Supplemental Material). In order to overcome this issue, we used the mutation rate from external sources, and use this mutation rate to estimate the effective population in each population.

### Table I

|                | New model  | Infinite site model |
|----------------|------------|---------------------|
| A - C          | -2.20 ± 0.15 | -3.91 ± 0.01 |
| C - B          | -2.05 ± 0.16 | -3.87 ± 0.02 |
| B - DR         | -1.51 ± 0.14 | -3.73 ± 0.03 |
| DR - DQ        | -3.22 ± 0.09 | -4.17 ± 0.05 |
| Total          | -1.27 ± 0.13 | -3.28 ± 0.02 |

* Corresponding author: louzouy@math.biu.ac.il

[1] P. Moran, Random processes in genetics, Math. Proc. Camb. Philo. 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: Vol. 2. The theory of gene frequencies, (1969).
[4] W. Ewens, The sampling theory of selectively neutral alleles, Theo. Pop. Bio. 3, 87 (1972).
[5] G. Watterson, On the number of segregating sites in genetical models without recombination, Theo. Pop. Bio. 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 M, and C. Neuhaiser, Ancestral processes with selection, Theo. Pop. Bio. 51, 210 (1997).
[8] J. Kingman On the genealogy of large populations, J. App. Prob. 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. Nat. Acad. Sci. 72, 358 (1975).
[11] N. Barton, A general model for the evolution of recombination, Gen. Res. 65, 123 (1995).
[12] R. Hudson, Properties of a neutral allele model with intragenic recombination, Theo.Pop.Bio. 23, 183 (1983).
[13] J. Szostak, T. Orr-Weaver, R. Rothstein, and F. Stahl, The double-strand-break repair model for recombination, Cell 33, 25 (1983).
[14] M. Neale and S. Keeney, Clarifying the mechanics of DNA strand exchange in meiotic recombination, Nature 442, 153 (2006).
[15] J. San Filippo, Joseph, P. Sung, and H. Klein, Mechanism of eukaryotic homologous recombination, Annu. Rev. Biochem. 77, 229 (2008).
[16] W. Heyer, K. Ehmsen, and J. Liu, Regulation of homologous recombination in eukaryotes, Annu. Rev. Genet. 44, 113 (2010).
[17] 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).
[18] G. Coop and M. Przeworski, An evolutionary view of human recombination, Nat. Rev. Genet. 8, 23 (2007).
[19] E. Yunis and D. Amos, Three closely linked genetic systems relevant to transplantation, Proc. National Acad. Sci. 68, 3031 (1971).
[20] 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, PNAS 116, 14098 (2019).

[21] Jeffreys, Alec J and Kauppi, Liisa and Neumann, Rita, Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex, Nature genetics 29, 217 (2001).

[22] A. Lynn, T. Ashley, and T. Hassold, Variation in human meiotic recombination, Annual review of genomics and human genetics 5, 317 (2004).

[23] F. Pratto, K. Brick, P. Khil et al., Recombination initiation maps of individual human genomes, Science 346, (2014).

[24] F. Sun, M. Oliver-Bonet et al., Human male recombination maps for individual chromosomes, Am. J. Hum. Gen. 74, 521 (2004).

[25] Genome-wide recombination map construction from single individuals using linked-read sequencing, Dréau, Andreea and Venu, Vrina and Avdeievich, Elena and Gaspar, Ludmila and Jones, Felicity C, Nat. Comm. 10, 1 (2019).

[26] A. Hinch, G. Zhang, P. Becker, D. Moralli, R. Hinch, B. Davies, R. Bowden, and P. Donnelly, Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm, Science 363, (2019).

[27] F. ´Ubeda, T. Russell, W. Timothy, and V. Jansen, PRDM9 and the evolution of recombination hotspots, Theo. Pop. Bio. 126, 19 (2019).

[28] F. Ali, M. Arif, Muhammad, Z. Khan et al., SDBP-Pred: Prediction of single-stranded and double-stranded DNA-binding proteins by extending consensus sequence and k-segmentation strategies into PSSM, Analytical biochemistry, 589, 113-94 (2020).

[29] A. Patel, J. Horton, G. Wilson, X. Zhang, and X. Cheng, Structural basis for human PRDM9 action at recombination hot spots, Genes & development 30, 257 (2016).

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

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

[32] P. Hermann, A. Heissl, I. Tiemann-Boege, and A. Futschik, LDJump: Estimating variable recombination rates from population genetic data, Molecular ecology resources, 19, 623 (2019).

[33] G. Clarke, and L. Cardon, Disentangling linkage disequilibrium and linkage from dense single-nucleotide polymorphism trio data, Genetics 171, 2085 (2005).

[34] R. Hudson, Estimating the recombination parameter of a finite population model without selection, Genetics Research 50, 245 (1987).

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

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

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

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

[39] M. Stumpf and G. McVean, Estimating recombination rates from population-genetic data, Na. Rev. Gen. 4, 959 (2003).

[40] 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 research 17, 520 (2007).

[41] P. Leberg, Genetic approaches for estimating the effective size of populations, The Journal of Wildlife Management 69, 1385 (2005).

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

[43] A. Blancher, A. Aarnink, N. Savy, 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: Genomes— Genomics— Genetics textbf2, 123 (2012).

[44] 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).

[45] J. Felsenstein, Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci?, Molecular biology and evolution 23, 691 (2006).

[46] Shem-Tov, Yahav and Danino, Matan and Shnerb, Nadav M, Solution of the spatial neutral model yields new bounds on the amazonian species richness, Sci. Rep. 7, 1 (2017).

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

[48] A. Melka, N. Dori, and Y. Louzoun, Invasion rate versus diversity in population dynamics with catastrophes, Phys. Rev. Lett. 124, 158301 (2020).

[49] B. Bradly, The role of HLA matching in transplantation, Immuno. Lett. 29, 55 (1991).

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

[51] S. Spellman, Stephen, M. Setterholm, M. Maiers, Martin 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, Bio.Blood Marrow Transplant. 14, 37 (2008).

[52] 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 immunology 74, 1313 (2013).

[53] 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).

[54] 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 Comp. Bio. 13, e1005693 (2017).

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

[56] V. Narasimhan, R. Rahbari, A. Scally et al., Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes, Nat. Comm. 8, 7 (2017).

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