scientific reports

Check for updates

OPEN

Long-range linkage effects in adapting sexual populations

Igor M. Rouzine

In sexual populations, closely-situated genes have linked evolutionary fates, while genes spaced far in genome are commonly thought to evolve independently due to recombination. In the case where evolution depends essentially on supply of new mutations, this assumption has been confirmed by mathematical modeling. Here I examine it in the case of pre-existing genetic variation, where mutation is not important. A haploid population with N genomes, L loci, a fixed selection coefficient, and a small initial frequency of beneficial alleles f_0 is simulated by a Monte-Carlo algorithm. When the number of loci, L, is larger than a critical value of $4\log^2(Nf_0)$, simulation demonstrates a host of linkage effects that decrease neither with the distance between loci nor the number of recombination crossovers. Due to clonal interference, the beneficial alleles become extinct at a fraction of loci $1-2\log(Nf_0)/L^{0.5}$. Due to a genetic background effect, the substitution rate varies broadly between loci, with the fastest value exceeding the one-locus limit by the factor of $[L^{0.5}/\log(Ns)]^{0.75}$. Thus, the far-situated parts of a long genome in a sexual population do not evolve as independent blocks. A potential link between these findings and the emergence of new Variants of Concern of SARS-CoV-2 is discussed.

Humans are heterozygous at millions of genomic sites, loci. The average difference between an individual's genome and the consensus genome is estimated at 20 million base pairs, or 0.6% of the total of 3.2 billion base pairs. The invention of the new methods of full-genome DNA sequencing caused the emergence of the field of genomics and proteomics dedicated to the quantitative aspects of genetic diversity and gene expression at a large number of loci. To describe and visualize the genetic complexity, various computational methods have been developed including phylogenetics, the principle-components analysis, the cluster analysis. Among them, mathematical modeling of evolution stands out as a tool of a high predictive power. Modeling allows to connect, in the most direct and reproducible fashion, the assumptions about the dominant factors of evolution to the predictions for the observable parameters of genetic diversity and evolutionary dynamics.

The assumptions and simplifications of models vary broadly depending on the systems studied and the questions asked. Two distinct groups of models and methods have been applied to animal populations and microbial populations. The classical one-locus and two-locus models that neglect interaction with the other loci in genome dominate the way in which many evolutionary biologists think about the evolution of higher organisms. In contrast, monocellular eukaryotes, viruses, and bacteria that are characterized by an extremely high genetic diversity and ultrarapid evolution, are often described by asexual or partly sexual population models that include explicitly large numbers of interacting loci. Analysis of the evolutionary dynamics of multi-locus models is more complex than one-locus and two-locus models and relies either on Monte-Carlo simulation[1,2] or the advanced mathematical methods of statistical physics[3–10].

The heavy mathematical artillery is required, because the evolution of many different loci is coupled. Two kinds of interference effects exist. One kind, not considered in this article, is epistasis arising from biological interaction of different loci, including protein–protein interactions or the interactions of gene regulation network[8,11–14]. The second type of interference, which is the focus of the present article, is the effects originating from the common ancestry of different loci, including clonal intereference and background selection. The effect of competition between clones with beneficial mutations at different sites was first described by Fisher[15] and Muller[16]. Later, Hill and Robertson provided another argument showing that the action of selection at different sites is not independent[17]. Both effects were shown to be equivalent by Felsenstein[18] and will be referred to below as "clonal interference"[3]. Linkage effects slow down adaptation[4,7], increase accumulation of deleterious alleles[19], and change the statistical shape of genealogical tree[5,6,20]. The focus of the present work is on clonal interference and "background selection", the last term referring to the fact that selection acts at the level of whole genomes, and not separate loci.

Sechenov Institute of Evolutionary Physiology and Biochemistry, Russian Academy of Sciences, Saint-Petersburg, Russia 194223. email: igor.rouzine@iephb.ru

In sexually reproducing organisms and viruses with frequent recombination, linkage effects are partly compensated by recombination between parental genomes. A fundamental fact of genetics discovered by Morgan is that frequent recombination destroys allelic associations, so that alleles at far-spaced loci segregate independently. Conventional wisdom tells us that all the other linkage effects between far-situated loci must vanish as well. Models of long-term sexual evolution depending on new mutation events confirm this expectation[21–23]. Assuming that genome consists from independently-evolving blocks and applying the phylogenetic theory of asexual evolution to each block, these authors constructed a scaling argument expressing the length of each block, the lead of the traveling wave, and the average coalescent time in terms of the average adaptation rate. The analytic predictions have been confirmed numerically for two particular models of population in the presence of natural selection and mutation.

In the present work, I investigate linkage effects in a different biological scenario, when natural selection and recombination act on pre-existing beneficial alleles, and new mutations can be neglected. This model is appropriate, for example, when a population migrates to a new environment, or a virus was subjected to rapid mutation for a period of time. Then the fixation of pre-existing beneficial alleles does not depend on mutation de novo.

Results

Model. The evolutionary factors included in the model are directional natural selection, random genetic drift, linkage, and recombination. A sexually reproducing population is comprised of N individual genomes (or $N/2$ diploid genomes without allelic dominance), where each genome has L loci, and L is assumed to be much larger than unit. In the beginning, each locus is assumed to have a fraction f_0 of beneficial alleles, with a fixed fitness benefit s. The initial state of a population is obtained by the generation of random and uniform distribution of alleles across all sites and individual genomes. The value of f_0 is assumed to be in interval $\frac{1}{Ns} \ll f_0 \ll 1$.

The evolution is simulated in MATLAB™ using a Wright-Fisher process, in which the progeny genomes replace the parental genome. The logarithm of the average progeny number (log fitness) is given by the product of the number of beneficial alleles and the selection coefficient s (Eq. 9, Section "Materials and Methods"). The random number of progeny for each genome is generated using a random number generator and the broken-stick algorithm. The total number of genomes N does not change. With some probability r, which is an input parameter of the model, a genome undergoes a number of random crossovers with another, randomly chosen genome. The average crossover number is M. One of the two parents is replaced with the recombinant. Below I assume $r = 1$, which corresponds to fully-sexual reproduction. Parameters r and M can be connected to the average number of crossovers between two sites r_2, which enters 2-locus models, as given by $r_2 = rM/L$. Altogether, the model has 5 input parameters (N, L, s, r, M) and the initial value f_0.

New mutation events are absent. Epistasis and allelic dominance are neglected, and a haploid population is considered. Indeed, as it is well-known in population genetics, a diploid population with $N/2$ genomes and without dominance is effectively haploid, with a double number of genomes N. The details are given in section "Materials and Methods".

Extinction of beneficial alleles depends on a single composite parameter. If the number of loci L is sufficiently large, beneficial alleles at most loci become extinct. The fraction of remaining polymorphous loci, denoted $1 - C_{loss}(t)$, decreases in time from 1 to at a low plateau (Fig. 1, red line).

This result differs strongly from the prediction of the single-locus model, in which multiple lineages per site are expected to reach fixation in the chosen parameter interval, and C_{loss} is exponentially small. Indeed, in the single-locus model, the fixation probability of an allele is s, and the extinction probability is $1 - s$[24]. The probability of the extinction of Nf_0 beneficial alleles present in the beginning is given by $C_{loss}(\infty) = (1 - s)^{Nf_0} \approx e^{-Nf_0 s}$, which is exponentially small. Thus, in the parameter regime investigated, $Nf_0 s \gg 1$, the one-locus model predicts that many alleles are fixed per each locus, and the loss of polymorphous loci due to genetic drift[25] is negligible. Therefore, the observed loss is not a one-locus effect ocurring due to genetic drift, but is caused by competition between the clones of beneficial alleles expanding at different loci[3,15,16,18].

Varying model parameters in simulation, I found out empirically that the fraction of loci with non-extinct alleles, $1 - C_{loss}(\infty)$, depends mostly on a single composite parameter (Fig. 2A–C)

$$1 - C_{loss} = \begin{cases} 2.0\frac{\log(Nf_0)}{\sqrt{L}} & 1 \ll \log(Nf_0) < 0.5\sqrt{L} \\ 1 & \log(Nf_0) > 0.5\sqrt{L} \end{cases} \quad (1)$$

Note the critical point, $\log(Nf_0) = 0.5\sqrt{L}$. If the population size is too large or the number of loci is too small, no significant loss of polymorphism is predicted. Intuitively, the clonal interference effects are expected to increase with the number of intefering loci, i.e., the length of genome L, just as the linkage effects on the adaptation rate increase with L[4]. The reason for a seemingly sharp transition remains to be investigated by analytic methods.

The fastest adaptation rate among loci is much faster than in a single-locus model. Because most loci fail to complete adaptation, the average frequency of beneficial alleles per locus, $f_{av}(t)$, saturates far below 1 (Fig. 1, blue line). The dependence of average heterozygocity on time, $2w^2(t)$, is decreased accordingly (Fig. 1, green). The allele frequency averaged over remaining polymorphic sites, $f_{pol}(t)$, increases in the same general time range as the one-locus prediction. The time of half-fixation of polymorphous sites, t_{50}, is very close to the deterministic one-locus prediction, $t_{50} \approx t_{1loc}$ (Fig. 1)

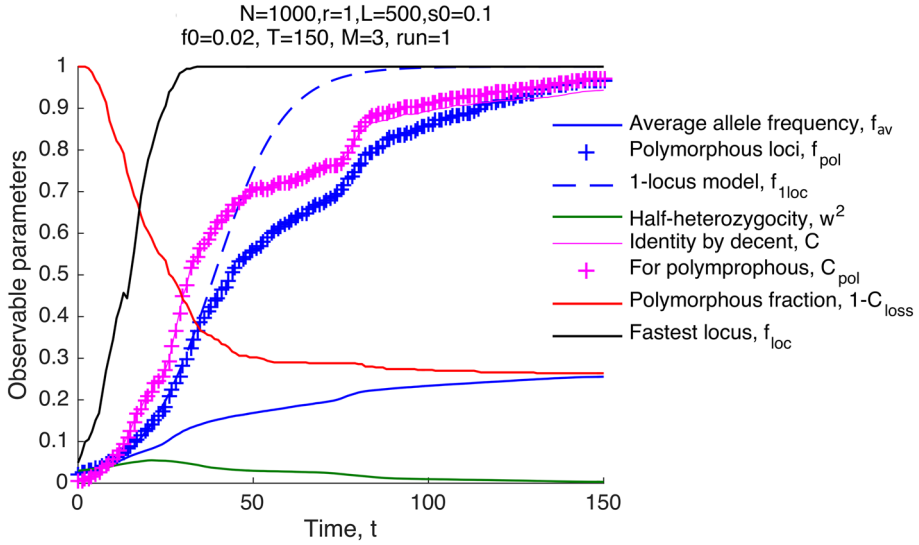

Figure 1. Dynamics of observables in the model with standing variation and the absence of mutation. Beneficial alleles become extinct at most loci. X-axis: Time in generations, t. Y-axis: observable parameters calculated during simulation. The average frequency of beneficial alleles per locus per individual, f_{av}, the same value averaged over polymorphous loci only, f_{pol}, the prediction for f_{av} of the deterministic one-locus model, f_{1loc}, half-heterozygocity $w^2 = \langle f(1-f)\rangle$, the fraction of homologous pairs of loci with a common initial ancestor, C, the same value for polymorphous loci, C_{pol}, the fraction of polymorphous loci, $1 - C_{loss}$, and the largest of allelic frequencies among loci, $\max(f_{loc})$. Parameter values are shown on the top. Parameters are defined in *Methods* and values are shown.

$$t_{1loc} = \frac{1}{s}\log\frac{1}{f_0} \tag{2}$$

In the range of parameters $s = 0.025 - 0.2, L = 200 - 2000, N = 1000 - 10,000$, the relative difference between t_{50} and t_{1loc} is between -0.11 and 0.14. Compared to the one-locus model prediction (blue dashed curve in Fig. 1), the dependence $f_{pol}(t)$, experiences a delay in the late phases of adaptation and has a noticeable random oscillation component (Fig. 1, blue +).

At some loci, alleles accumulate much faster than predicted by the one-locus model (Fig. 1, black line). Indeed, the half-time of adaptation of the fastest locus, $\max(t_{loc})$, is much shorter than t_{1loc} and increases as power ¾ of composite parameter $\frac{\log(Ns)}{\sqrt{L}}$(Fig. 2E). The ultrarapid evolution at some loci implies that the genome segments comprising these loci tend to have unusually high numbers of beneficial alleles. Indeed, the fitness of a genome is set to be proportional to that number. In other words, the broad variation in the evolution rate between loci with an identical selection coefficient demonstrates the existence of a strong background selection effect created by random recombination events. Recombination brings together different numbers of favorable alleles in different segments, and natural selection favors the fittest. The resulting distribution of genomes in fitness forms a traveling wave, well-known for both asexual and sexual populations[24] (Fig. 3A). The new genomes at the high-fitness edge of the wave are born by recombination. They grow much faster in number than the average genome at the wave maximum, causing the wave to move forward[26].

The population has a complex lineage structure that varies between loci. For a given locus, a lineage is determined as the set of individuals that have the same initial ancestor. The lineages all initially consists from a single individual, the founder (Fig. 3B), but their sizes grow in time, at different rates for different loci, and become distributed in a very broad range (Fig. 3C–G). The lineage size distribution between loci shifts in time towards larger lineages eventually occupying almost the entire population. If we take into account only the largest lineage for each locus, their size distribution looks similar but has a low cutoff increasing in time (Fig. 3B–G, column 2). The largest lineages grow to a half of the population at a much earlier time than t_{1loc} in Eq. (2).

Phylogenetic time scale depends only on the same composite parameter. Another quantity affected by linkage effects is the identity by descent, C, defined as the probability of a homologous locus pair to have the same initial ancestor. The average identity by descent averaged over all loci and over only polymorphous loci is almost the same (magenta line and magenta +, Fig. 1). This result, again, differs from the single-locus prediction, where common ancestry is rare, $C(t) < f_{pol}^2(t)$, because each of the pair of loci must fall into the same growing lineage to have the same ancestor, and the size of each lineage relative to the population size is smaller than $f_{pol}(t)$. In contrast, in the present simulation, $C(t)$ is larger than $f_{pol}(t)$, which is larger than $f_{pol}^2(t)$.

At the time point $t = T_2$ such that $C(T_2) = 1 - C_{loss}(T_2)$, both quantities are close to a half, in a broad parameter range, as given by

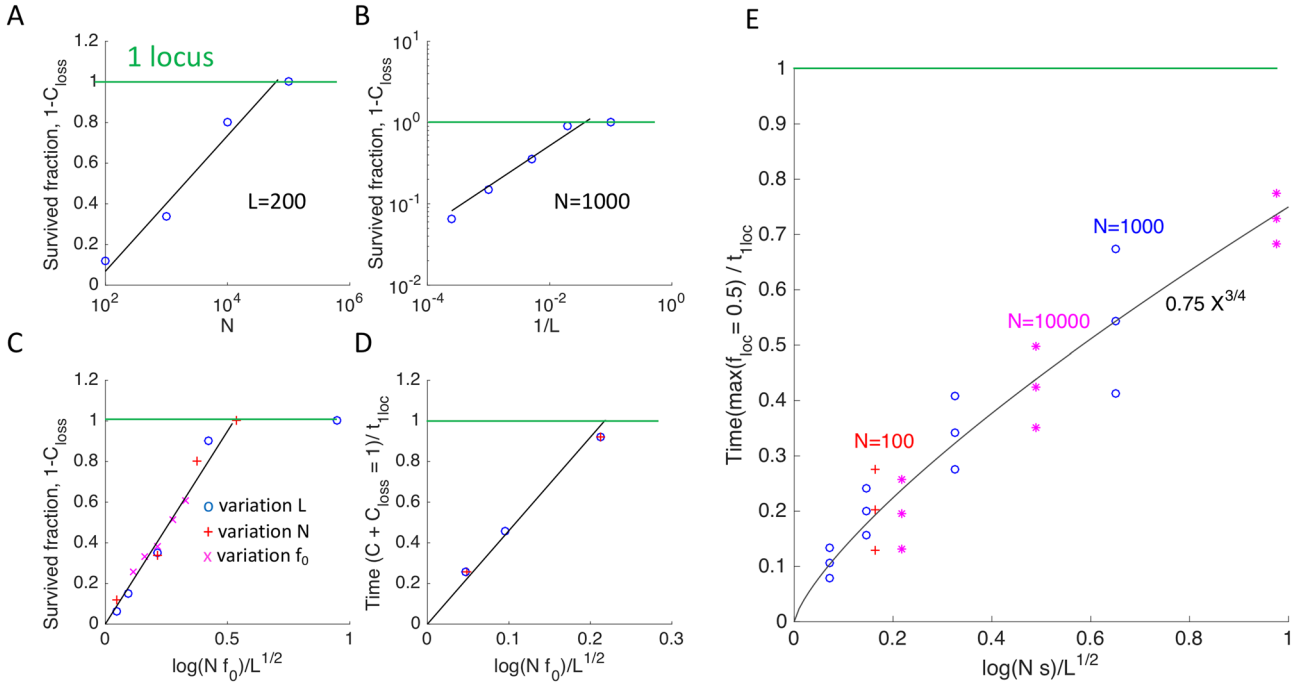

Figure 2. The observables depend mostly on a single composite parameter. (**A–C**). The locus fraction where beneficial alleles have survived and completed adaptation, $1 - C_{loss}(\infty)$, increases linearly with the natural logarithm of the population size, $\log N$, the inverse square root of the locus number, $1/\sqrt{L}$, and a composite parameter, $\log(Nf_0)/\sqrt{L}$. Colored symbols o, +, and x correspond to the variation of model parameters L, N, and f_0, respectively, where $f_0 >> 1/Ns$. The green horizontal line shows the prediction of the one-locus model, $C_{loss} \approx 0$. (**D**) The time, t, when the survived-loci fraction, $1 - C_{loss}(t)$, equals the average identity by descent, $C(t)$, [intersection of red and pink curves in Fig. 1] scales linearly with $\log(Nf_0)/\sqrt{L}$ as well. **(E)** The time when the allelic frequency at the fastest locus reaches 50%, scales as a power ¾ of a similar parameter, $\log(Ns)/\sqrt{L}$. The symbol triplets show the mean and the 95% confidence interval. Colored symbols o, +, and * show different values of N. The sensitivity to the variation of selection coefficient s, crossover number M, and initial allele frequency f_0 is shown in Fig. 4 and Fig. S1. The default parameter values are $N = 1000, L = 200, f_0 = 0.02$ unless shown otherwise. The other parameters are as in Fig. 1.

$$C(T_2) \approx C_{loss}(T_2) \approx 0.5$$

The dependence of T_2 on model parameters can be interpolated by the formula

$$T_2 \approx t_{1loc} \frac{5.0 \log(Nf_0)}{\sqrt{L}} \tag{3}$$

In other words, time T_2 is proportional to the same composite parameter that controls the fraction of successful loci, $1 - C_{loss}(\infty)$, given by Eq. (1) (Fig. 2D). Time T_2 determined by Eq. (3) represents a proxy time scale of the phylogenetic tree. Although, at this time point, a population does not have a single ancestor for an average locus as yet, T_2 approximates the time to the most recent common ancestor by an order of magnitude.

Observables depend weakly on the average number of recombination crossovers. The above results in Figs. 1, 2, and 3 are weakly sensitive to the average crossover number, M. In its entire range of between 1 and L, the fraction of loci that do not lose alleles, $1 - C_{loss}(\infty)$, varies only by the factor of ~ 2 (Fig. 4). The variation of selection coefficient s is rescaling units of time; otherwise, its effect is modest (Fig. 4).

The absence of long-range linkage disequilibrium. A measure of linkage disequilibrium (LD) is Pearson correlator between allelic frequencies at two loci

$$r^2(l_{12}) = \frac{\langle (f_1 - \langle f \rangle)(f_2 - \langle f \rangle) \rangle}{\left\langle (f_1 - \langle f \rangle)^2 \right\rangle} \tag{4}$$

which is averaged over pairs of sufficiently heterozygous loci, $2f_{loc}(1 - f_{loc}) > 0.1$. The value of r^2 defined by Eq. (4) depends on time, as follows. Initially, $r^2 \equiv 0$ due to the random initial distribution of alleles set in simulation. After some time has elapsed, r^2 becomes positive at any distance between loci (Fig. 5, $t = 30$). After passing through a maximum, r^2 decreases in time again. At each next time point, r^2 depends more and more sharply

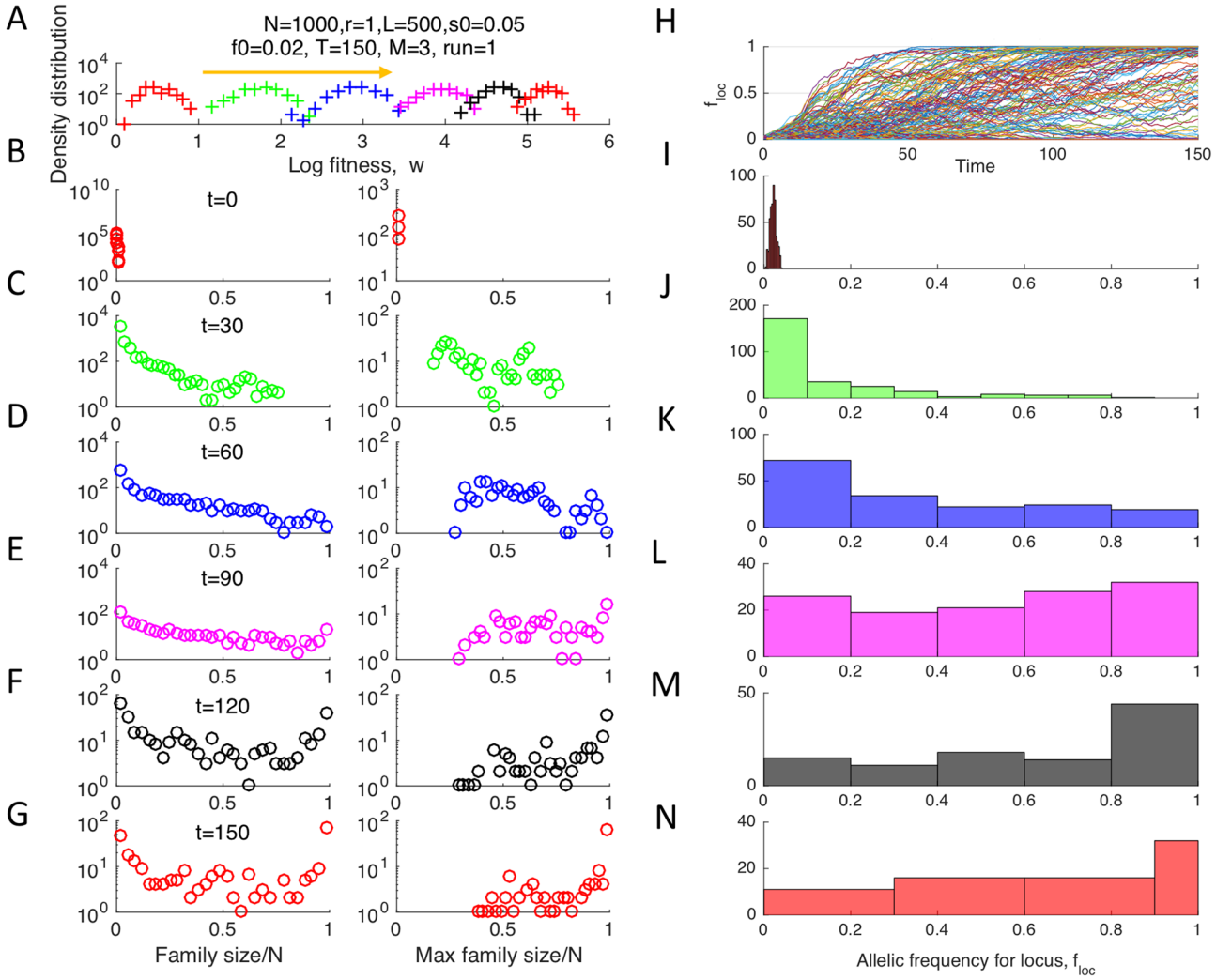

Figure 3. Traveling fitness wave and nonuniform dynamics of separate loci. (**A**) Distribution density of genomes in fitness at different time points shown in (B-G). (**B–G**) First column: Histograms of the family size defined as the number of sequences with the same initial ancestor at a locus. Second column: Only the largest family per locus is taken into account. (**H**) The average allelic frequency for each separate locus, f_{loc}, as a function of time. (**I–N**) Histograms of f_{loc} across loci at different time points (shown). Parameters are as in Fig. 1.

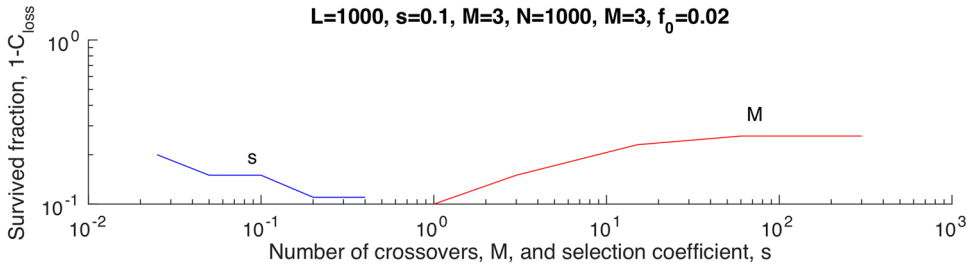

Figure 4. The fraction of loci that are not lost and complete adaptation are weakly senstive to the average crossover number. The default parameter values are shown on the top.

on the distance between loci, l_{12} (Fig. 5). In other words, alleles at far-situated loci segregate independently, as expected in the presence of recombination (Morgan's law).

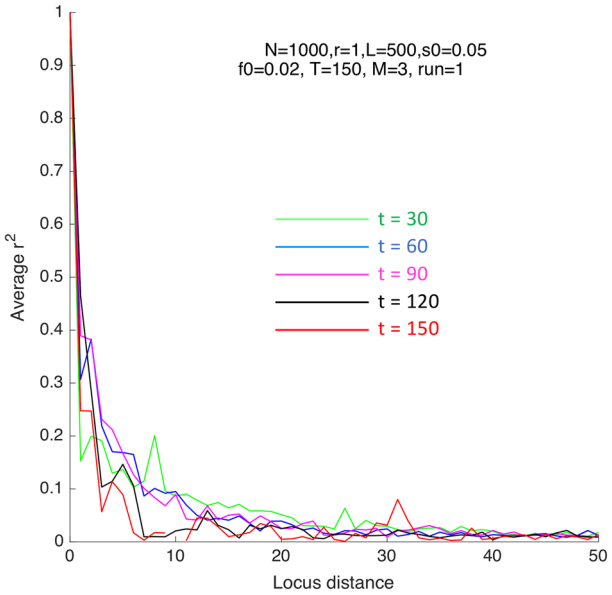

Figure 5. Linkage disequilibrium decays rapidly with the distance between loci in a genome. Y-axis: Pearson's measure r^2, Eq. (4). The time points and parameters (shown) are the same as in Fig. 3. At $t = 0$, linkage disequilibrium is identically zero due to the initially-random distribution of alleles.

Alleles are fixed inter-dependently. The fixation probability of an allele can be calculated as

$$P_{fix} = \frac{\log[1/C_{loss}(\infty)]}{Nf_0} \approx \frac{1 - C_{loss}(\infty)}{Nf_0} \tag{5}$$

In the parameter interval of interest, this value falls far below the 1-locus prediction, $P_{fix}^{1loc} = s$ (Fig. S1). Probability P_{fix}, Eq. (5), plateaus on the value of s in the dilute limit of sufficiently small f_0 [27]. Based on simulation, the transition point to the dilute limit f_0^{dilute} decreases with N and L. One can determine the transition point from condition $P_{fix}(f_0^{dilute}) = s$ and Eqs. (1) and (5), as follows

$$f_0^{dilute} \sim \frac{1}{Ns\sqrt{L}} \log \frac{1}{s\sqrt{L}}, \; s\sqrt{L} \ll 1 \tag{6}$$

The estimate from Eq. (6) agrees with the simulation results (Fig. S1).

Phylogenetic tree varies between loci. In addition to calculating the phylogeny time scale T_2, we constructed the ancestral trajectory of a locus between individuals in real-time by recording the parentage of each individual locus and then tracing its ancestry back in time. Lineage of each locus jumps randomly between individuals due to recombination (Fig. S2A). If we straighten these trajectories and keep only the topology of coalescence and the coalescent times, we arrive at phylogenetic trees for different loci (Fig. S2B–D). As expected, the tree varies strongly between loci due to recombination, and the early branches are much shorter relative to late branches compared to Kingman's coalescent constructed in the absence of selection[28]. The average density of coalescent events averaged over 10 runs and normalized to the prediction of the selectively-neutral model (*Methods*) decreases exponentially with time (Fig. S2E, F). This is because the coalescent event density is proportional to the inverse effective population size[28], which is the size of the growing variant subpopulation. Importantly, the coalescent density is much larger than in the one-locus limit and increases with number of loci L. Thus, in agreement with the previous studies, uncompensated linkage in the presence of selection makes phylogenetic trees denser and changes their shape by making early stems shorter[5,6,20,29] (Fig. S2E,F).

Alleles "surf" between lineages. In addition to the ancestor number trajectory of a locus (Fig. S2A), one can also construct its fitness trajectory, by recording the fitness values of its ancestors (Fig. S2G). The fitness trajectory comprises alternating straight horizontal segments due to the clonal expansion connected to jumps caused by recombination. The jumps occur in both directions, but more often towards a genetic background with a higher fitness (Fig. S2G). This "allelic surfing" behavior was predicted for sexual populations analytically[27,30].

Discussion

In the presence of standing variation, on moderate time scales, rapid evolution can occur in the absence of new mutations, and even much faster than due to mutation[26]. Such evolution based on standing variation and natural selection, with or without recombination, has been observed for poliovirus[31] and VSV[32]. The present study considers a process of adaptation after minute quantities of beneficial alleles are generated in the beginning, for

example, due to the change of external environment, or due to early mutation. If an allele is not lost to random genetic drift, its further accumulation is dominated by natural selection and recombination working together. The process continues, until all alleles are either fixed or lost.

Despite of the lack of observable LD for far-situated loci, simulation predicts the existence of strong long-range linkage effects encompassing the entire genome. The effects include the extinction of beneficial alleles at most loci, due not to random drift but to clonal interference, weak sensitivity of results to the number of crossovers, and ultrarapid evolution at some loci, even faster than in the independent-locus limit. The last observation implies the existence of genomic segments enriched in beneficial alleles over the average. Taken altogether, these results imply that far-situated genomic regions do not evolve independently, and recombination is not strong enough to break down linkage effects caused by selection acting, thus, at the level of a whole genome, as well as at the level of genomic segments.

If the locus number is decreased, or if the population size is increased, a transition to the independent-locus limit is predicted. The predicted dependence of all linkage effects on the population size N is logarithmic (Fig. 2). For a genome of 200 loci and $f_0 = 0.02, s = 0.1$, the transition to the independent-locus regime can be observed already for 100,000 individuals. For a longer genome of 1000 loci, they would evolve independently only for populations of 10^{12} individuals or larger, which is unrealistic for most species. A human or an animal population has millions of variable loci, of which a significant fraction is under natural selection, so that independent-locus models, probably, never work in most animals, except for rare mutations that are under very strong selection pressure.

The results of the present study carried out for the moderate-term evolution are in striking contrast to the previous findings for the long-term evolution driven by mutation, selection, and recombination, where genome was demonstrated to consist from quasi-independent blocks[21–23]. In my notation, the cited result for the average time to the most recent common ancestor has the form [[21], Eq. 5]

$$T_{MRCA} \approx const \frac{M}{v} \log\left(\frac{Nv}{M}\right) \tag{7}$$

where v is the average rate of long-term adaptation, defined as the fitness gain per unit time, *const* is a number on the order of 1, and the logarithm is supposed to be much larger than 1. In the present model, the proxy of T_{MRCA} , by the order of magnitude, is T_2 in Eq. (3), and the adaptation rate is

$$v \approx const \frac{sL(1 - C_{loss})}{t_{50}} \tag{8}$$

As already mentioned, the average time to a half-fixation for the loci that do not lose alleles, t_{50}, is always close to one-locus limit t_{1loc}. Substituting Eqs. (3) and (8) into Eq. (7), we get

$$\frac{M\log(Nv/M)}{s\log^2(Nf_0)} = const$$

The last equality is clearly false, because M, N, f_0 and s are independent parameters. Hence, Eq. (7) does not apply in the present case with pre-existing variation.

Note that the analytic argument in[21] was developed and tested for the stationary long-term adapation. After the medium-term evolution ends with the loss or fixation of all initially-present alleles, further evolution requries a constant supply of new mutations. The derivation in[21] is based on two statements: the assumption that a genome evolves as quasi-independent asexual blocks, and an expression for the time to the most recent common ancestor in terms of the average adaptation rate. The expression was based on the basic concept that the time to most recent common ancestor is the lead of the wave divided by the adaptation rate and was confirmed for various multi-locus models, both sexual and asexual. Therefore, it is likely that the quasi-independence assumption is the cause of the discrepancy.

In other words, in the case of pre-existing variation and moderate-term evolution studied here, the genome does not evolve as a set of quasi-independent segments. That conclusion is indirectly confirmed by the results in Fig. 3 showing that beneficial alleles can form highly-fit genomes whose rapid growth outruns mixing of genomes due to recombination (Fig. 3). A recombinant that decreases fitness is not relevant for future generations. Furthermore, within one realization (Monte Carlo run), the fitness variance of a genomic segment normalized to the genome fitness is not linearly proportional to its length, but shows a complex step-like dependence (Fig. 6).

The results obtained in the present study might be potentially relevant for the viruses with frequent recombination, such as HIV, polio, or SARS-CoV-2. Similar to seasonal human coronaviruses or influenza virus, SARS-CoV-2 is constantly acquiring new mutations in its genome and has hundreds (if not thousands) of observably-diverse sites. Evolution is especially fast in receptor Spike protein, 5 replacements per year[33–36]. Two major reasons account for the high speed of evolution, as follows. Firstly, Spike has receptor-binding motives that affect transmission, and their evolution leads to the emergence of variants with enhanced transmissibility. Secondly, Spike contains epitopes, regions that are important for the immune response because of their involvement in binding of antibodies that can neutralize virus. Mutations in epitopes are a major factor that limits the virus recognition by the immune system and, hence, the durability of protection[37–39].

A puzzle important for devising future vaccination strategies is the origin of the variants of concern (VOC) produced by large groups of new mutations that emerge all together at once[40–44]. Alternative theories of the emergence of VOCs[45] include reverse zoonosis, the evolution within immunocompromised patients[46,47] and the evolution in population pockets not covered by the genetic surveillance. Still another possibility is the fitness

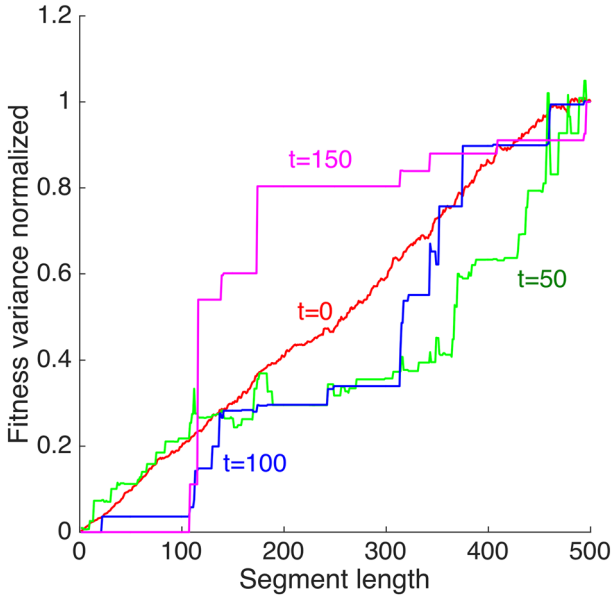

Figure 6. Non-linear dependence of genome segment variance on segment length. X-axis: the length of a genome segment starting from locus 1. Y-axis: Fitness variation between homologous genomic segments divided by the genome fitness variation at the same moment of time. A single Monte-Carlo run is shown. Parameters are, as in Fig. 1.

valley effect, a cascade emergence of compensating mutations following a primary mutation, an effect previously inferred for HIV and influenza[12,48] and studied theoretically[49].

Based on the present study, I may add yet another possible explanation. SARS-CoV-2, with its single-chromosome genome, has observable crossover recombination[50–52]. Hence, the large packages of mutations may emerge due to the combined effects of recombination and natural selection and represent the sequences comprising the fastest loci (Fig. 3H and J).

VOC are characterized by the sudden emergence of sister strains with large groups of mutations, occuring not on the background of the existing strain, but in parallel. This real-life observation can be compared to the tree predicted by the present simulation obtained by neighbor-joining analysis from a sample of 10 simulated sequences, at 3 time points each (Fig. 7). Note that genome samples obtained at different time points form separate subtrees (clades) growing on the same stem, the first two subtrees without apparent relation to each other. Each tree has an excess in the number of beneficial alleles compared to the previous tree. These features bear resemblance to VOC.

To understand the importance of recombination for SARS-CoV-2, we need to know the frequency of co-infected individuals among all the infected, which determines outcrossing probability r, an important input parameter entering the models of sexual populations[1,21,30]. For fully sexual reproduction considered in the present work, by the definition, $r = 1$. The outcrossing number for SARS-CoV-2 is presently unknown, but it could be larger than it seems due to the possibility of a co-infection during superspreading events[53,54]. Methods developed previously to quantify recombination from RNA sequence data for HIV could be re-applied to SARS-CoV-2[1,55]. The potential relevance of the present model for VOC of SARS-CoV-2, as well as a detailed analysis of SARS-CoV-2 in the light of this hypothesis, will be investigated elsewhere.

Assumptions and limitations.

(i) The model considered only variable sites under purifying selection. While a real population has many conserved sites and some selectively neutral sites as well, the existence of these sites has no effect on the evolutionary dynamics of the sites under purifying selection. Hence, these sites are not included into consideration explicitly.

(ii) We have investigated the case of a constant selection coefficient, but the results are expected to apply also for a sufficiently fast decaying distribution of selection coefficients, such as a Gaussian distribution. Distributions with long tails may have different properties, where the traveling wave is replaced by pairwise clonal interference[7]. The border case is the exponential distribution often observed in experiments on pathogens, which fact has been explained[56]. In this case, the present scenario with a fixed effective value of selection coefficient applies at sufficiently large population sizes[7].

(iii) The model assumed homologous recombination for the two reasons, as follows. Firstly, recombination in viruses and organisms is similar in the sense that the vast majority of recombination events occur between homologous templates and do not include insertions or deletions. Secondly, when non-homologous recombination does occur, which effect is more frequent in viruses, the resulting progeny is often defective and, hence, not important for the evolutionary dynamics of a population.

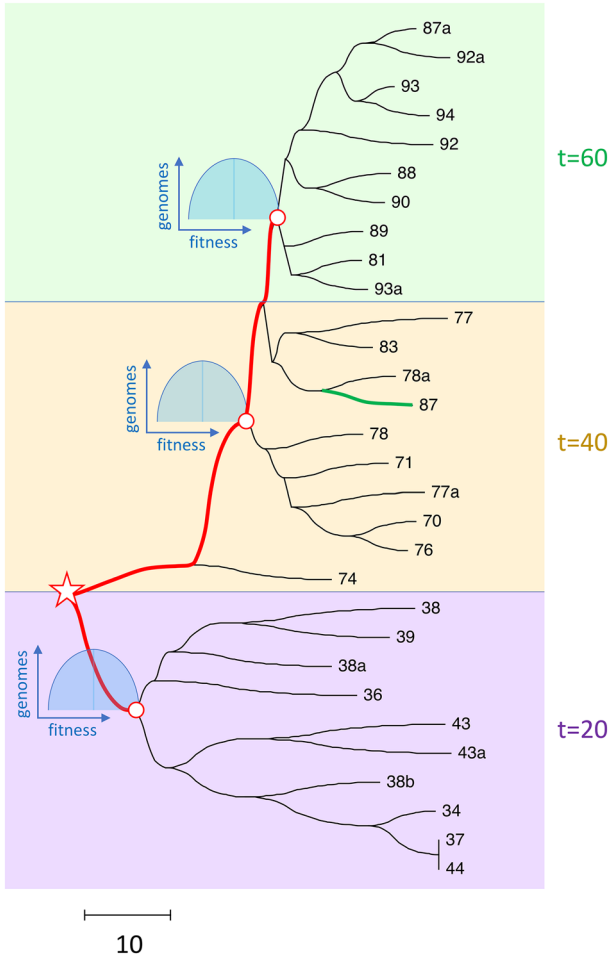

Figure 7. Phylogenetic tree at different time points has a polyphiletic structure with a common stem. Phylogeny of 10 simulated genomes sampled at three time points, $t = 20, 40, 60$, is obtained by neighbor-joining analysis in MEGA11. Fat red line shows the stem connecting the subtrees. Open circles show roots of three subtrees emerging at the high-fitness edge of the genome distribution in fitness (shown). Number of beneficial alleles in a genome is shown in leaves. The scale of genetic distance measured in the number of differences is shown below. Model parameters are as in Fig. 1.

Conclusion

In sexual populations with pre-existing beneficial alleles, in an exponentially broad range of population size, recombination cannot suppress long-range linkage effects, including the excessive loss of beneficial alleles due to clonal interference, independence of observables on the average number of crossovers, and superfast evolution at some loci due to a genetic background effect. A potential link of these findings to the emergence of VOC of SARS-CoV-2 will be investigated elsewhere.

Materials and methods

Consider a fully sexual population with L loci comprised of N individual genomes. Each locus has initially Nf_0 alleles, $1/Ns \ll f_0 \ll 1$, with fitness benefit $s \ll 1$. In each generation step, each genome undergoes random crossovers with another, randomly chosen genome, with average crossover number M, producing a recombinant genome. One of the two parents is replaced with the recombinant. Genome number j with k_j favorable alleles is replaced with a random number of its copies distributed according to the polynomial distribution implemented by "broken stick" method, as follows. N random points are generated uniformly within the interval $[0, N]$ broken into N segments. The length of segment j is proportional to the fitness of the corresponding genome w_j

$$w_j = \frac{\exp\left(sk_j\right)}{\sum_{j=1}^{N} \exp\left(sk_j\right)} \tag{9}$$

The number of random values that fall into segment j are taken to be the number of his progeny in the next generation. Thus, the total number of genomes stays constant. New mutations are neglected, which is

shown to be correct in the short-term in the presence of pre-existing genetic variation, both in simulation and experimentally[31,32]. Epistasis is absent; for epistatic analysis, see[12] and references therein.

Input model parameters are the selection coefficient across loci, $s = s_0$, population size N, outcrossing rate $r = 1$, number of loci L, initial beneficial allele frequency f_0, total simulation time t, average number of recombination crossovers M, and the seed number of the generator of pseudorandom numbers.

Parameter ranges studied are $s = [0.025, 0.4], L = [10, 4000], N = \left[10^2, 10^5\right], M = [1, 300], f_0 = [0.0001, 0.02]$. The main focus is on the interval of f_0 such that $\frac{1}{Ns} \ll f_0 \ll 1$. The transition to dilute limit $Nf_0s \ll 1$ when alleles are fixed independently is shown in Fig. S1.

Data availability

The simulation code is available at https://github.com/irouzine/Strong-linkage-in-sex.

Received: 16 November 2022; Accepted: 25 July 2023
Published online: 01 August 2023

References

1. Batorsky, R. *et al.* Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 5661–5666. https://doi.org/10.1073/pnas.1102036108 (2011).
2. Bedford, T., Rambaut, A. & Pascual, M. Canalization of the evolutionary trajectory of the human influenza virus. *BMC Biol.* **10**, 38. https://doi.org/10.1186/1741-7007-10-38 (2012).
3. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102–103**, 127–144 (1998).
4. Rouzine, I. M., Wakeley, J. & Coffin, J. M. The solitary wave of asexual evolution. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 587–592. https://doi.org/10.1073/pnas.242719299 (2003).
5. Brunet, E., Derrida, B., Mueller, A. H. & Munier, S. Effect of selection on ancestry: an exactly soluble case and its phenomenological generalization. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **76**, 041104. https://doi.org/10.1103/PhysRevE.76.041104 (2007).
6. Desai, M. M., Walczak, A. M. & Fisher, D. S. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics* **193**, 565–585. https://doi.org/10.1534/genetics.112.147157 (2013).
7. Good, B. H., Rouzine, I. M., Balick, D. J., Hallatschek, O. & Desai, M. M. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 4950–4955. https://doi.org/10.1073/pnas.1119910109 (2012).
8. Neher, R. A. & Shraiman, B. I. Statistical genetics and evolution of quantitative traits. *Rev. Mod. Phys.* **83**, 1283 (2011).
9. Pedruzzi, G., Barlukova, A. & Rouzine, I. M. Evolutionary footprint of epistasis. *PLoS Comput. Biol.* **14**, e1006426. https://doi.org/10.1371/journal.pcbi.1006426 (2018).
10. Rouzine, I. M. *Mathematical Modeling of Evolution. Volume 2 Fitness Landscape, Red Queen, Evolutionary Enigmas, and Applications to Virology* (De Gruyter, 2023).
11. Wei, W.-H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15**, 722. https://doi.org/10.1038/nrg3747 (2014).
12. Pedruzzi, G. & Rouzine, I. M. An evolution-based high-fidelity method of epistasis measurement: Theory and application to influenza. *PLoS Pathog.* **17**, e1009669. https://doi.org/10.1371/journal.ppat.1009669 (2021).
13. Jerison, E. R. & Desai, M. M. Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments. *Curr. Opin. Genet. Dev.* **35**, 33–39. https://doi.org/10.1016/j.gde.2015.08.008 (2015).
14. Kryazhimskiy, S., Rice, D. P., Jerison, E. R. & Desai, M. M. Microbial evolution. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science* **344**, 1519–1522. https://doi.org/10.1126/science.1250939 (2014).
15. Fisher, D. S. *The Genetical Theory of Natural Selection*. (Clarendon Press, 1930).
16. Muller, H. Some genetic aspects of sex. *Am. Nat.* **66**, 118 (1932).
17. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
18. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737–756. https://doi.org/10.1093/genetics/78.2.737 (1974).
19. Rouzine, I. M., Brunet, E. & Wilke, C. O. The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. *Theor. Popul. Biol.* **73**, 24–46. https://doi.org/10.1016/j.tpb.2007.10.004 (2008).
20. Neher, R. A. & Hallatschek, O. Genealogies of rapidly adapting populations. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 437–442. https://doi.org/10.1073/pnas.1213113110 (2013).
21. Neher, R. A., Kessinger, T. A. & Shraiman, B. I. Coalescence and genetic diversity in sexual populations under selection. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15836–15841. https://doi.org/10.1073/pnas.1309697110 (2013).
22. Good, B. H., Walczak, A. M., Neher, R. A. & Desai, M. M. Genetic diversity in the interference selection limit. *PLoS Genet.* **10**, e1004222. https://doi.org/10.1371/journal.pgen.1004222 (2014).
23. Weissman, D. B. & Hallatschek, O. The rate of adaptation in large sexual populations with linear chromosomes. *Genetics* **196**, 1167–1183. https://doi.org/10.1534/genetics.113.160705 (2014).
24. Rouzine, I. M. *Mathematical Modeling of Evolution: Volume 1: One-Locus and Multi-Locus Theory and Recombination* (De Gruyter, 2020).
25. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159. https://doi.org/10.1093/genetics/16.2.97 (1931).
26. Rouzine, I. M. & Coffin, J. M. Evolution of human immunodeficiency virus under selection and weak recombination. *Genetics* **170**, 7–18. https://doi.org/10.1534/genetics.104.029926 (2005).
27. Neher, R. A., Shraiman, B. I. & Fisher, D. S. Rate of adaptation in large sexual populations. *Genetics* **184**, 467–481. https://doi.org/10.1534/genetics.109.109009 (2010).
28. Kingman, J. F. C. Origins of the Coalescent: 1974–1982. *Genetics* **156**(4), 1461–1463. https://doi.org/10.1093/genetics/156.4.1461 (2000).
29. Brunet, E., Derrida, B. & Simon, D. Universal tree structures in directed polymers and models of evolving populations. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **78**, 061102. https://doi.org/10.1103/PhysRevE.78.061102 (2008).
30. Rouzine, I. M. & Coffin, J. M. Highly fit ancestors of a partly sexual haploid population. *Theor. Popul. Biol.* **71**, 239–250. https://doi.org/10.1016/j.tpb.2006.09.002 (2007).
31. Xiao, Y. *et al.* RNA recombination enhances adaptability and is required for virus spread and virulence. *Cell Host Microbe* **19**, 493–503. https://doi.org/10.1016/j.chom.2016.03.009 (2016).
32. Dutta, R. N., Rouzine, I. M., Smith, S. D., Wilke, C. O. & Novella, I. S. Rapid adaptive amplification of preexisting variation in an RNA virus. *J. Virol.* **82**, 4354–4362. https://doi.org/10.1128/JVI.02446-07 (2008).
33. Yewdell, J. W. Antigenic drift: Understanding COVID-19. *Immunity* **54**, 2681–2687. https://doi.org/10.1016/j.immuni.2021.11.016 (2021).
34. Eguia, R. T. *et al.* A human coronavirus evolves antigenically to escape antibody immunity. *PLoS Pathog.* **17**, e1009453. https://doi.org/10.1371/journal.ppat.1009453 (2021).

35. Rochman, N. D. *et al.* Ongoing global and regional adaptive evolution of SARS-CoV-2. *Proc. Natl. Acad. Sci. U. S. A.* https://doi.org/10.1073/pnas.2104241118 (2021).
36. Haynes, W. A. *et al.* High-resolution epitope mapping and characterization of SARS-CoV-2 antibodies in large cohorts of subjects with COVID-19. *Commun. Biol.* **4**, 1317. https://doi.org/10.1038/s42003-021-02835-2 (2021).
37. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463-476.e46. https://doi.org/10.1016/j.chom.2021.02.003 (2021).
38. Greaney, A. J. *et al.* Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* **29**, 44-57.e49. https://doi.org/10.1016/j.chom.2020.11.007 (2021).
39. Rouzine, I. M. & Rozhnova, G. Evolutionary implications of SARS-CoV-2 vaccination for the future design of vaccination strategies. *Commun. Med.* **3**, 86. https://doi.org/10.1038/s43856-023-00320-x (2023).
40. SARS-CoV-2 variants of concern as of 27 January 2022. *European Centre for Disease Prevention and Control*, https://www.ecdc.europa.eu/en/covid-19/variants-concern (2022).
41. Martin, D. *et al.* The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. *Cell* **184**, P5189-5200.e5187. https://doi.org/10.1016/j.cell.2021.09.003 (2021).
42. Ghafari, M., Liu, Q., Dhillon, A., Katzourakis, A. & Weissman, D. Investigating the evolutionary origins of the first three SARS-CoV-2 variants of concern. *Front. Virol.* https://doi.org/10.3389/fviro.2022.942555 (2022).
43. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* **21**, 361–379. https://doi.org/10.1038/s41579-023-00878-2 (2023).
44. Tay, J. H., Porter, A. F., Wirth, W. & Duchene, S. The emergence of SARS-CoV-2 variants of concern is driven by acceleration of the substitution rate. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msac013 (2022).
45. Otto, S. P. *et al.* The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Curr. Biol.* **31**, R918–R929. https://doi.org/10.1016/j.cub.2021.06.049 (2021).
46. Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592**, 277–282. https://doi.org/10.1038/s41586-021-03291-y (2021).
47. Corey, L. *et al.* SARS-CoV-2 variants in patients with immunosuppression. *N. Engl. J. Med.* **385**, 562–566. https://doi.org/10.1056/NEJMsb2104756 (2021).
48. Rouzine, I. M. & Coffin, J. M. Search for the mechanism of genetic variation in the pro gene of human immunodeficiency virus. *J. Virol.* **73**, 8167–8178. https://doi.org/10.1128/JVI.73.10.8167-8178.1999 (1999).
49. Weissman, D. B., Desai, M. M., Fisher, D. S. & Feldman, M. W. The rate at which asexual populations cross fitness valleys. *Theor. Popul. Biol.* **75**, 286–300. https://doi.org/10.1016/j.tpb.2009.02.006 (2009).
50. Ignatieva, A., Hein, J. & Jenkins, P. A. Ongoing recombination in SARS-CoV-2 revealed through genealogical reconstruction. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msac028 (2022).
51. Jackson, B. *et al.* Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell* **184**, 5179-5188.e5178. https://doi.org/10.1016/j.cell.2021.08.014 (2021).
52. Yi, H. 2019 novel coronavirus is undergoing active recombination. *Clin. Infect. Dis.* **71**, 884–887. https://doi.org/10.1093/cid/ciaa219 (2020).
53. Lau, M. S. Y. *et al.* Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 22430–22435. https://doi.org/10.1073/pnas.2011802117 (2020).
54. Liu, Y., Eggo, R. M. & Kucharski, A. J. Secondary attack rate and superspreading events for SARS-CoV-2. *Lancet* **395**, e47. https://doi.org/10.1016/S0140-6736(20)30462-1 (2020).
55. Neher, R. A. & Leitner, T. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput. Biol.* **6**, e1000660. https://doi.org/10.1371/journal.pcbi.1000660 (2010).
56. Barlukova, A. & Rouzine, I. M. The evolutionary origin of the universal distribution of mutation fitness effect. *PLoS Comput. Biol.* **17**, e1008822. https://doi.org/10.1371/journal.pcbi.1008822 (2021).

Acknowledgements

The study was carried out within the framework of the state assignment of the Federal Agency for Scientific Organizations (FASO Russia: topic no. AAAA-A18-118012290142–9).

Author contributions

I.M.R. has done the research, wrote and revised the manuscript, and prepared figures.

Competing interests

The author declares no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-39392-z.

Correspondence and requests for materials should be addressed to I.M.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.