SCIENTIFIC REP🅞RTS

OPEN

The rich phase structure of a mutator model

David B. Saakian[1,2], Tatiana Yakushkina[3] & Chin-Kun Hu[1,4,5]

We propose a modification of the Crow-Kimura and Eigen models of biological molecular evolution to include a mutator gene that causes both an increase in the mutation rate and a change in the fitness landscape. This mutator effect relates to a wide range of biomedical problems. There are three possible phases: mutator phase, mixed phase and non-selective phase. We calculate the phase structure, the mean fitness and the fraction of the mutator allele in the population, which can be applied to describe cancer development and RNA viruses. We find that depending on the genome length, either the normal or the mutator allele dominates in the mixed phase. We analytically solve the model for a general fitness function. We conclude that the random fitness landscape is an appropriate choice for describing the observed mutator phenomenon in the case of a small fraction of mutators. It is shown that the increase in the mutation rates in the regular and the mutator parts of the genome should be set independently; only some combinations of these increases can push the complex biomedical system to the non-selective phase, potentially related to the eradication of tumors.

The concepts of genome instability[1–11] and clonal evolutionary dynamics[12–14] are key ideas in understanding the causes and behavior of cancer. Genome instability is a hallmark of cancer[11], plays a significant role in bacterial evolution[15–19], and has been a focus of physics-community research on oncology[20]. We need more complex non-linear models[12–14,21] than those for the viral evolution[22–25] for the cancer analysis.

Models with relatively simple evolutionary dynamics can be applied to cancers that arise from one or two driver mutations, e.g., inherited retinoblastoma; however, most cancers require three or more. It has been proposed[1,26] that some cancers may be associated with a substantial increase in the mutation rate of the genome. The question here is how the dynamics and phase structure of the evolving clonal population.

There have been some attempts to construct and solve evolutionary dynamics systems with mutator genes[16,17]. The earlier work[16] on this subject, influenced by experimental results on *E. coli*[27], discusses the problem of how selection in natural populations favors more rapid mutation. The authors suggest a simple model with a linear fitness function to identify the conditions that lead to selection for or against mutator states. Some approximate estimates were made for the steady state distribution and the dynamics of the mutator fraction in the population. The later paper[17] used a quasispecies model with a multiplicative fitness function.

Classical quasispecies models[22,23] describe the evolutionary process of an infinite population of binary sequences caused by two forces: mutation and selection; the former is controlled by the mutation rate and the latter by the fitness function. One of the most impressive results for these models is the existence of a phase transition[23,28], known as the "error threshold" in some versions of the fitness landscape. When the mutation rate is small, a selective phase is observed, where sequences organize into a "quasispecies" near the sequence with the maximum fitness value. The non-selective phase observed with higher mutation rates shows dilution throughout the sequence space.

Returning to mutator dynamics, Nagar and Jain[17] first studied the possible phases for these systems. Here, the fraction of mutator genome sequences in the population defines different phases: a normal phase with only normal sequences (a degenerate case when there are no back-mutations in a mutator gene), a mixed phase in which both mutators and non-mutators coexist, and a pure mutator phase with only mutators.

[1]Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan. [2]A.I. Alikhanyan National Science Laboratory (Yerevan Physics Institute) Foundation, Alikhanian Brothers St. 2, Yerevan 375036, Armenia. [3]National Research University Higher School of Economics, Moscow, 101000, Myasnitskaya Street 20, Russia. [4]National Center for Theoretical Sciences, National Tsing Hua University, Hsinchu 30013, Taiwan. [5]Business School, University of Shanghai for Science and Technology, Shanghai 200093, China. Correspondence and requests for materials should be addressed to D.B.S. (email: saakian@yerphi.am) or T.Y. (email: tyakushkina@hse.ru) or C.-K.H. (email: huck@phys.sinica.edu.tw)

Among other findings, it has been shown that the small mutation rates for the mutator gene lead to the mixed phase with a small fraction of the mutator allele, but an increase in these rates turns the system to the mutator phase. The infinite population model with a constant mutation rate has been solved in ref. 29 (see also the recent review[30]).

When we map the evolutionary dynamics to statistical physics, the genome length plays the role of lattice size. In statistical physics, different phases arise as different analytical expressions for free energy at a large lattice size. Similarly, we need the large genome length to define the evolutionary dynamics phases. In this paper, we derive these different expressions for the mean fitness in mixed and mutator phases. The most realistic case, when back mutation for the mutator gene can be neglected due to its small impact during the time of the experiment, gives a very simple situation: in a mutator phase in a steady state, there is a zero fraction of normal sequences, even for small genome length, but there is a finite fraction of normal sequences in the mixed phase.

Several different classes of evolution models form the basis of mutator phenomenon analysis:

Class I   Phenomenological models[5,6]. Such models consider deterministic or stochastic equations for a set of several genotypes. Models in this class are relatively simple but useful for describing the specific types of experiments, e.g., investigations that show how the mutator mechanism competes with the ordinary evolution scheme with a high mutation rate[5,6].

  In ref. 5, a simple infinite population model without selection has been exposed, and it has been assumed that cancer arises after a given number of mutations in oncogenes. The probability of the start of cancer development at a fixed time has been calculated for the mutator mechanism as well as without mutators. The accuracy of the derivations has been well controlled using a small parameter: the ratio of the ordinary mutation rate to the mutation rate due to mutators. A different phenomenological model has been proposed without derivation details[6], which discusses three types of fitness landscapes. This concept is of interest, and we consider one of the fitness landscapes in our work. However, the phenomenological models as a methodology cannot describe the different phases of the evolutionary dynamics accurately.

Class II   Infinite population microscopic models. Models in this class are defined in the binary sequence space with a finite genome length[16,17]. The Crow-Kimura[18,22,31] and Eigen[23–25,32] models, which are the quasi-species models mentioned above, have been widely investigated over recent decades, especially in the context of virus evolution. Mathematically, each of these models is described via a system of nonlinear master equations, which can be mapped to the chain of linear ordinary differential equations with some nonlinear transformation. In the Crow-Kimura setting, mutation and selection are parallel processes, while in the Eigen model, mutation is connected to selection. In smooth fitness landscapes, the steady-state distributions of the Crow-Kimura and Eigen models can be mapped to each other. In the large genome limit, the Crow-Kimura model is equivalent to a discrete-time Eigen model[33] and is very close to branching processes. Moreover, the Wright-Fisher model[34,35] for large populations can be mapped to a discrete-time Eigen model[33].

Class III   Finite population microscopic models[36]. In this case, models are also defined in the binary sequence space[36]. In ref. 4, numerical simulations have been performed for the Moran model by considering the Muller ratchet in the presence of mutators. In ref. 7, the growth population version of the Wright-Fisher model[34,35] for the four-dimensional landscape has been formulated; there are neutral, deleterious and advantageous mutations, as well as a sufficient number of mutations in the part of the genome that brings about the mutator effect. According to these results, the mutator phenotype evolves only at a moderate level of selection for the advantageous mutations.

Class IV   Infinite population and infinite number of genotypes models without backward mutations[29,30]. These models are useful to describe the virus evolutionary dynamics for a short period of time. Compared with the microscopic models, which represent different possible phases and collective effects in evolution, class IV models are less fundamental.

In the current paper, we consider the microscopic infinite population models (class II). The discrete-time version of the Eigen model[23–25] was originally introduced to describe the self-replication of macromolecules at the origin of life (see also[37,38]). In this model, the selection and mutation happen together (mutation directly follows replication), while in the Crow-Kimura model, originally proposed to describe biological speciation, mutation and selection happen in parallel. In both models, the genome is represented by a chain of L letters (genes), which take the values ± 1, similar to the Ising spin[39,40]. There is a mutation rate of μ per letter. We take the sequence with all "+" letters as the reference sequence. All of the genomes with the same total number l of "−" letters (l point mutations from the reference sequence) have fitness $r_l \equiv f(1 - 2l/L)$, where $f(x)$ as a fitness function, $x = 1 - 2l/L$. For the initial symmetric distribution, the model can be described via differential equations of $L + 1$ variables P_l, corresponding to the total probabilities of the sequences with l mutations in the genome. The mean fitness of the Crow-Kimura model has been calculated using algebraic methods[41] and the Hamilton-Jacobi equation (HJE) method[42,43]. The exact dynamics had been derived using the HJE[44]. Here, we construct a mutator gene model on the basis of the Crow-Kimura model[18,22,31] and solve it for the general fitness case, by providing also the solution for the Eigen model version of the mutator phenomena.

Our solutions for both models are consistent with each other.
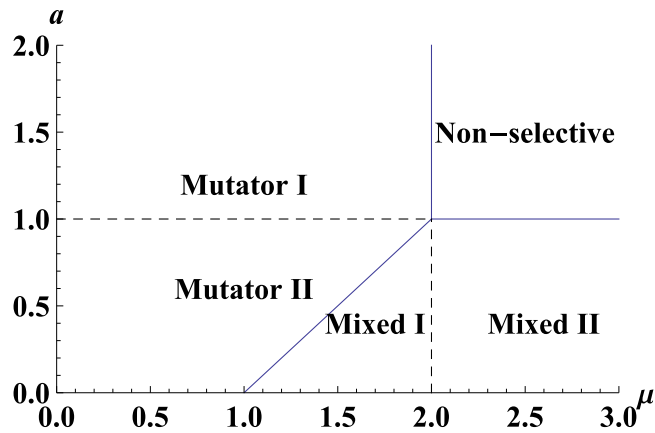
Figure 1. The phase structure of the mutator model with single peak landscape with zero value for any argument except $l = 0$: $f(x = 1) = g(x = 1) = J$. The parameters of the model are $\alpha_1 = a$, $\alpha_2 = 0$, $\mu_1 = 1$, and $\mu_2 = \mu$. There are three phases: mixed phase with $0 < s < 1$, $0 < q < 1$, non-selective phase with $s = 0$, $0 < q \leq 1$ and mutator phase with $0 < s$, $q = 1$. The border between non-selective and mutator phases is given by $\mu = J$, the border between non-selective and mixed phases is given by $a = J - 1$, between mixed and mutator phases is given by $a + 1 = \mu$ line. From the bio-medical perspectives we distinguish the mutator I and mutator II, mixed I and mixed II subphases. From the mutator I, the system transforms to the non-selective phase simply increasing the μ. From the mixed II the system transformers to the non-selective phase simply increasing the $a \equiv \alpha_1$. From the mutator II and mixed I subphases we need change both a and μ to transform the system to the non-selective phase. We have the same picture in case of Eigen model with a fitness A for the peak sequence and fitness 1 for other sequences. We have for the mixed phase $R = e^{-(h+\gamma)}A$, for the mutator phase $R = e^{-\mu\gamma}A$ and for the non-selective phase $R = 1$. The border between non-selective and mutator phases is given by $\mu\gamma = \ln A$, the border between non-selective and mixed phases is given by $h = \ln A - \gamma$, between mixed and mutator phases is given by $h = (\mu - 1)\gamma$ line.

We find the analytical solution for the mean fitness for the general fitness function, including the multidimensional case, and clarify and calculate the order parameters and the phases of the model, which was an open problem until now. Previous work[5,6] has been focused on the comparison between the efficiencies of cancer development via ordinary evolutionary dynamics and the evolutionary dynamics using the mutator mechanism. Our quantitative results of the phase structure (shown in Fig. 1) depict the borders of different phases. The procedure to obtain Fig. 1 is described below.

The phase structure may have biomedical implications. In ref. 9, it was proposed that the transition from the mixed phase to the non-selective phase is associated with the possible eradication of the tumor. We provide the solution for two types of fitness landscapes; a symmetric case, where the fitness landscape depends on the total number of mutations from the reference sequences, and a random fitness landscape.

Key technical terms and notations are listed in Tables 1 and 2, respectively.

Results

The model with a symmetric fitness landscape.
In this section, we consider a modification of the Crow-Kimura model[18,22,31] for evolutionary processes under the influence of a mutator gene. The chain of $(L + 1)$ genes defines the genome, which can be formally subdivided into two parts; a regular part with L genes and a mutator gene, whose mutations change the fitness and the mutation rate in the regular part of the genome. We represent each gene with two alleles by $\xi_\theta = \pm 1$, $\theta = 0, 1, \ldots, L$. The regular set of genes is denoted by the sequences $S_i = (\xi_1, \ldots, \xi_L)$, $i = 1, \ldots, 2^L$. The mutator gene in the state $\xi_0 = +1$ determines the normal dynamics of the regular part with a mutation rate μ_1. This wild-type form of the gene corresponds to the most common phenotype in the population. If the state of the mutator gene changes to $\xi_0 = -1$, then the evolutionary process switches to a different regime with a mutation rate μ_2. The mutant-type gene defines an altered phenotype, which is called a mutator phenotype. We use α_1 and α_2 to designate forward ($\xi_0: +1 \rightarrow -1$) and backward ($\xi_0: -1 \rightarrow +1$) mutation rates of the mutator gene.

We assume that the fitness landscape is symmetric, which means it depends solely on the number of mutations, l, from the reference sequence (without loss of generality, we take the reference sequence to be $S = (+1, \ldots, +1)$). For this purpose, it is natural to use the notion of Hamming distance between S and S_i: $l \equiv d_{1i} = (L - \Sigma_{\theta=1}^{L} \xi_\theta)/2$ because all sequences in the same Hamming class have the same value of fitness. To facilitate further analysis, we introduce the variable for the mean value, $x = \sum_{\theta=1}^{L} \xi_\theta / L$, $x \in [-1, 1]$. The state of the system is represented by two probability distributions. We use $\hat{P}_l(t)$ to denote the relative frequency of normal genome sequences (wild-types) with l mutations (in the l—th Hamming class) at the time moment, t, and $\hat{Q}_l(t)$ for the relative frequency of sequences with an increased mutation rate (mutator-types) with l mutations, and we have a probability balance condition for any time moment, $\sum_l (\hat{P}_l(t) + \hat{Q}_l(t)) = 1$. All admissible transitions between system states can be seen as arrows in Fig. 2 [45,46].

Lethal mutations	A mutation that results in the premature death of the organism carrying it.
Selective phase	A state of evolutionary system of binary sequences, when the majority of the population is concentrated near the peak genome (with the maximum fitness value) in the sequence space.
Non-selective phase	A state of evolutionary system of binary sequences, when the population is diluted in the sequence space.
Solvable cases of fitness landscape	Random fitness landscape and symmetric fitness landscape.
Random fitness landscape	The values of fitness defined on a sequence space are random numbers with some distribution.
Symmetric fitness landscape	The values of fitness defined on a sequence space by a function of the total mutation number.
Mutator phase	The absolute majority in the population has a mutator-type (mutator allele of the mutator gene) after long enough time of evolution.
Mixed phase	There is a finite fraction q of mutator sequences in the population after longtime evolution ($0 < q < 1$).

Table 1. Terms.

To describe the evolution of the probability distribution under consideration, we use the following system of $2(L+1)$ differential equations:

$$
\begin{aligned}
\frac{d\hat{P}_l(t)}{Ldt} &= \alpha_2 \hat{Q}_l + \hat{P}_l(f(x_l) - (\mu_1 + \alpha_1)) + \mu_1\left(\hat{P}_{l-1}\frac{L-l+1}{L} + \hat{P}_{l+1}\frac{l+1}{L}\right) - \hat{P}_l R(t), \\
\frac{d\hat{Q}_l(t)}{Ldt} &= \alpha_1 \hat{P}_l + \hat{Q}_l(g(x_l) - (\mu_2 + \alpha_2)) + \mu_2\left(\hat{Q}_{l-1}\frac{L-l+1}{L} + \hat{Q}_{l+1}\frac{l+1}{L}\right) - \hat{Q}_l R(t), \\
R(t) &= \sum_l (\hat{P}_l(t)f(x_l) + \hat{Q}_l(t)g(x_l)), \\
x_l &= 1 - 2l/L, \ \ 0 \le l \le L.
\end{aligned}
\tag{1}
$$

Here, $f(x_l)$ is a fitness function for the normal genome sequence, and $g(x_l)$ is a fitness function for the mutator sequence. As long as changes in the mutator gene significantly affect both the fitness and mutation rate in the regular part, fitness functions $f(x_l)$ and $g(x_l)$ can be different, and μ_2 is commonly 10–100 times larger than μ_1[2,10]. Note that the rates α_1 and μ_1 also differ because that α_1 is proportional to the number of possible mutagenic loci. The coefficients $(L-l+1)$ and $(l+1)$ appear in the Eq. (1) transition terms according to the combinatorial formulae for Hamming class probabilities[41,47]. To investigate Eq.(1), we consider a nonlinear transformation[48]:

$$
\hat{P}_l = \frac{P_l}{\sum_l (P_l + Q_l)}, \ \hat{Q}_l = \frac{Q_l}{\sum_l (P_l + Q_l)}.
\tag{2}
$$

It is easy to show that P_l and Q_l are the solutions of the following system of linear equations:

$$
\begin{aligned}
\frac{dP_l(t)}{Ldt} &= \alpha_2 Q_l + P_l(f(x_l) - (\mu_1 + \alpha_1)) + \mu_1\left(P_{l-1}\frac{L-l+1}{L} + P_{l+1}\frac{l+1}{N}\right), \\
\frac{dQ_l(t)}{Ldt} &= \alpha_1 P_l + Q_l(g(x_l) - (\mu_2 + \alpha_2)) + \mu_2\left(Q_{l-1}\frac{L-l+1}{L} + Q_{l+1}\frac{l+1}{L}\right).
\end{aligned}
\tag{3}
$$

In Eqs (1) and (2), $\hat{P}_l$ and $\hat{Q}_l$ are the probabilities under the normalization constraint, while P_l and Q_l of Eq. (3) are not normalized.

In this paper, we focus on the following characteristics of the model in a steady state: the mean fitness R, the total surplus s (the expected value of x_l), the surplus for the wild-type part of the population s_1 and for the mutator-type part s_2, and the fraction of the mutator sub-population q:

$$
R = \frac{\sum_l (P_l f(x) + Q_l g(x))}{\sum_l (P_l + Q_l)}, \ s = \frac{\sum_l (P_l + Q_l)x_l}{\sum_l (P_l + Q_l)},
$$
$$
s_1 = \frac{\sum_l P_l x_l}{\sum_l P_l}, \ s_2 = \frac{\sum_l Q_l x_l}{\sum_l Q_l}, \ q = \frac{\sum_l Q_l}{\sum_l (P_l + Q_l)}.
$$

The surplus s shows the average number $\bar{l}$ of mutations in the population, $\bar{l} = L(1-s)/2$. The maximum of the distributions P_l and Q_l is attained at points $L(1-s_1)/2$ and $L(1-s_2)/2$[44]. Therefore, the value of the mean fitness for wild-type sequences is equal to $f(s_1)$, and that for mutator-type sequences is equal to $g(s_2)$, as the distribution is narrow due to the large constant L in the exponent in Eq. (5) below. Assuming a smooth distribution, we obtain the following from Eqs (1) and (2):

$$
R = qg(s_2) + (1-q)f(s_1).
\tag{4}
$$

Calculating the mean fitness in growing populations is crucial for understanding evolutionary dynamics and is the primary concern of this investigation. Depending on the values of the main model parameters, we can obtain different analytical expressions for the mean fitness, which correspond to different phases of the system. The key problem is to first solve the large L limit, and clarify the following items:

L	The genome length.
l	The number of mutations.
$\xi_\theta, \theta = 0,1, \dots, L$	$\xi_0 = \pm 1$ — the state of the mutator gene, $\xi_\theta = \pm 1$ ($1 \le \theta \le L$) — the states of genes in the regular part of the genome.
$S_i, i = 1, \dots, 2^L$	The genome sequence.
r_l	The fitness value after l mutations in the Crow-Kimura model.
P_l, Q_l	The probability of the wild-types and mutator-types with l mutations.
$x_l = 1 - 2l/L$	The average gene state in the genome sequence that corresponds to l mutations from the reference sequence.
$f(x_l), g(x_l)$	The fitness functions of the wild-type and mutator-type genome sequences which assigns the reproductive rate to the average gene state x_l
μ_1, μ_2	The mutation rate for the wild-types and mutator-types.
μ	The substitution used to denote briefly: $\mu_2 = \mu$ for the choice $\mu = 1$.
α_1, α_2	The transition rates from the wild-type to mutator-type and vise versa.
α	The substitution used to denote briefly: $\alpha = \alpha_1$ for the choice $\alpha_2 = 0$.
s	The surplus of the whole population: the average state of the gene.
s_1, s_2	The surpluses for the wild-types and mutator types.
$\bar{l}$	The average number of mutations, which is equal to $(1-s)L/2$.
q	The total probability of the mutator-types.
R	The mean fitness.
$u(x, t)$	The probability of l mutations is $\sim \exp[Lu(x, t)]$, Eq. (5).
$v_1(x, t), v_2(x, t)$	The relative probabilities of the wild and mutator types after l mutations, $x = 1 - 2l/L$, Eq. (5).
$V_\pm(x)$	The potential of evolutionary dynamics, Eq. (26).
k	The coefficient in the linear fitness, $f(x) = kx$.
J	The peak fitness in the single peak fitness landscape for the Crow-Kimura model.
U	The total number of mutations for one generation in population genetics.
h	The transition probability wild-type to mutator-type for one generation in population genetics.
ε	The time period of one generation.
p_i, q_i	The probabilities of the i-th sequences for the wild type and mutator type in the Eigen model.
$Q_{ij}, \hat{Q}_{ij}$	The transition probabilities between the sequences in wild-type and mutator-type in the Eigen model.
$w, \hat{w}$	The probabilities of errorless replication per nucleotide for the wild-type and mutator-type.
$Q, \hat{Q}$	The probabilities of errorless replication for the genome the wild-type and mutator-type.
$d(i, j)$	The Hamming distance between the i-th and j-the genomes
A	The peak fitness in single peak fitness landscape for the Eigen model.
τ	The generation period of the virus.
y_i	The relative length of the i-th piece of genome in multi-dimensional model.

Table 2. Notations.

....

- A. Is it possible to have selective and non-selective phases in the system with a phase transition between them?
- B. Does the average number of mutations depend on the state of the mutator gene or is it the same in both cases ($s_1 = s_2$)?
- C. Is it possible to have a significant fraction of normal genomes in the population: $1 - q > 0$?
- D. Are there large finite L corrections for q at $L \sim 10000$?

These four questions are discussed in this paper; we show that both negative and positive answers exist for each. The second question (B) leads to the most considerable distinction in the analysis. In the Results section, we first investigate the case $s_1 = s_2$; then, in the Methods section, we describe a more complex situation with $s_1 \neq s_2$. Moreover, contrary to the classical evolution models for quasispecies, here, the evolutionary picture depends on the genome length.

The case of forward and backward mutations of the mutator gene. Let us consider a smooth solution of Eq. (3) at the limit $L \to \infty$ after the simplification mentioned in the previous section. We assume the following ansatz[42,44,49]:

$$P_l(t) \equiv P(x, t) = v_1(x, t)e^{Lu(x,t)}, \quad Q_l(t) \equiv Q(x, t) = v_2(x, t)e^{Lu(x,t)}, \qquad (5)$$

where we denoted $x = 1 - 2l/L$ for large values of L. This form of expressions for P_l and Q_l allows the application of the HJE method to the system of equations (5) (see Methods). Putting our ansatz (5) into Eq. (3), we get a system of two equations for the variables v_1, v_2, du/dx and du/dt. Looking at these equations as a system of two linear algebraic equations for v_1 and v_2, we get a zero determinant condition for the non-zero solution of v_1 and v_2, so the final equation contains only du/dt and du/dx. The fact that L disappears in the Hamilton-Jacobi equation in the large L limit shows the correctness of our ansatz. We derive the mean fitness investigating the Hamiltonian without solving it and calculating the function $u(x)$ in the steady state.
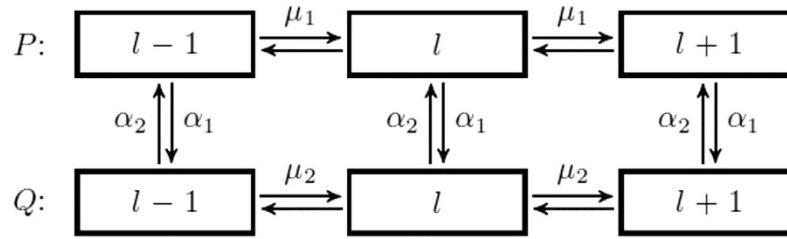
Figure 2. The scheme of available transitions for the system states (arrows denote transitions). The upper chain corresponds to the genome without a mutator allele (wild-type); the lower chain corresponds to the genome with a mutator allele (mutator-type). l is the number of mutations in the regular part of genome.

We obtain the following formulae for the mean fitness R using the potential function $V_\pm$[44]:

$$R = \max_x \left[V_+(x) \right],$$
$$V_\pm(x) = \frac{f(x) + g(x)}{2} - \frac{1 + \mu}{2} - \frac{\alpha_1 + \alpha_2}{2} + \frac{1 + \mu}{2}\sqrt{1 - x^2} \pm \frac{1}{2}\sqrt{A(x)^2 + 4\alpha_1\alpha_2}. \tag{6}$$

Here, we use $\mu_1 = 1$, $\mu_2 = \mu$, and $A(x) = f(x) - g(x) - \alpha_1 + \alpha_2 + (1 - \mu)(\sqrt{1 - x^2} - 1)$. The numerical simulation for the system (1) supports the analytical result (6) well, as seen from Table 3.

Having the expression for the mean fitness R, we can calculate the surplus of distribution s from the following equation (see Methods):

$$R = \frac{f(s) + g(s) - \alpha_1 - \alpha_2}{2} + \frac{1}{2}\sqrt{(f(s) - g(s) - \alpha_1 + \alpha_2)^2 + 4\alpha_1\alpha_2}. \tag{7}$$

At the time limit $t \to \infty$, we denote $v_k(x) \equiv v_k(x, t)$, $k = 1, 2$.

It is important to note that Eq. (7) is valid whether $s_1 = s_2$ or not. However, for $\alpha_1 \cdot \alpha_2 \neq 0$, we have that $s_1 = s_2$. In this case, substituting the solution of R and s, we get the following system of equations for the surplus s and ratio $v_2(s)/v_1(s)$ in the steady state (see Methods):

$$R = f(s) - \alpha_1 + \frac{v_2(s)}{v_1(s)}\alpha_2,$$
$$R = g(s) - \alpha_2 + \frac{v_1(s)}{v_2(s)}\alpha_1. \tag{8}$$

In the limit of large t and L, we have $q = v_2(s)/(v_1(s) + v_2(s))$. For the case $f(x) \equiv g(x)$, we get a simple relationship from (8):

$$q = \frac{\alpha_1}{\alpha_1 + \alpha_2}. \tag{9}$$

The model with uni-directional mutations of the mutator gene. The probability of back mutation (from mutant to wild-type) is small and can be neglected in finite populations. Let us generalize this idea, and consider the infinite population model with $\alpha_2 = 0$, $\alpha_1 = a$ ($\mu_1 = 1$, $\mu_2 = \mu$). In this case, the expressions obtained for potential functions $V_\pm(x)$ are still valid, but it is necessary to take into account both branches of the potential, V_+ and V_-, and define the mean fitness as their maxima.

Taking $V_-(x)$ in Eq. (6), we get the mutator phase (here and below, we use terminology from[17]):

$$R = \max_x [g(x) + \mu(\sqrt{1 - x^2} - 1)],$$
$$g(s) = R. \tag{10}$$

The latter equation for the surplus has the same form as for the ordinary Crow-Kimura model[41]. It can be derived directly from system (1) if we assume that the vast majority of the population has the mutator allele and omit the contribution of P_l in the second equation of Eq. (1). In the mutator phase, the mean fitness is defined by the fitness landscape for the mutator genome. Further, we will consider the case when two fitness landscapes are identical, e.g., $f(x) = g(x)$.

Taking $V_+(x)$ in Eq. (6) under the condition $\alpha_2 = 0$, $\alpha_1 = a$, $\mu_1 = 1$, $\mu_2 = \mu$ leads to the following expression for the mean fitness of the mixed phase (see Methods):

$$R = \max_x [f(x) - a + \sqrt{1 - x^2} - 1],$$
$$f(s) = R. \tag{11}$$

μ	3.5	3.0	2.5	2.0	1.5	1.0	0.5
R_{num}	0.1907	0.2514	0.32405	0.4127	0.5240	0.6684	0.8626
R_{th}	0.1811	0.2436	0.3180	0.4084	0.5212	0.6666	0.8615

Table 3. The comparison of the results for $f(x) = g(x) = 3x^2/2$, $\mu_1 = 1$, $\mu_2 = \mu$, $\alpha_1 = \alpha_2 = 1$, $L = 400$. R_{num} is the numerical result and R_{th} is given by Eq. (6).

In the mixed phase, the mean fitness is defined by the wild-type fitness function and the mutation rate. However, numerical calculations for the linear fitness landscapes $f(x) = kx$ or simple form quadratic fitness landscapes $f(x) = kx^2/2$ show that the vast majority of the population has the mutator allele for a large enough genome length. It is worth pointing out that this result (the majority of population has a mutator-type) is correct for any smooth fitness landscapes for a long enough genome length. We verify that for the single peak fitness function, there is a finite fraction of wild-type sub-population at the large genome limit. Such results will be presented below.

Two phases have been found for the discrete-time Eigen model in ref. 17, where the mean fitness has been calculated for $f(x) = k(x - 1)$, $k \gg 1$. This model can be exactly mapped into the Crow-Kimura model at the large L limit, when the mutation rate per gene is fixed. Figure 3 gives a comparison of our analytical results for the mean fitness R with the numerics. Figure 3 shows that our analytical results are quite reliable.

The single peak model with uni-directional mutations of the mutator gene.
Let us consider a single peak fitness function, $f(x = 1) = g(x = 1) = J$ and $f(x \neq 1) = g(x \neq 1) = 0$. We have three possible phases in a steady state (Fig. 1). We consider either the steady state of Eq. (1), or the asymptotic solution of Eq. (3), replacing $\frac{dQ_l}{Ldt} = RQ_l$, $\frac{dP_l}{Ldt} = RP_l$[25]. We provide the following expressions for the mean fitness in the mutator phase R_{mu}, mixed phase R_{mix} and non-selective phase R_{ns}; see Fig. 1:

$$R_{mix} = J - 1 - a, \quad R_{mu} = J - \mu, \quad R_{ns} = 0. \tag{12}$$

We first calculate P_0 considering the equation for dP_0/dt in Eq. (1) and ignore P_1 term. The fraction of the wild-type sequences in the l-th Hamming class in the population is $P_l = P_0(1/J)^l$, which can be derived from Eq. (55) of ref. 50; therefore, for the fraction of the population with a normal mutator gene, we have $\sum_{l=0}^{L} P_l = \frac{P_0 J}{J - 1}$. Considering analogously the equation for dQ_0/dt and ignoring the Q_1 term, we obtain $Q_0 = -P_0 \frac{a}{J - \mu - R} = P_0 \frac{a}{\mu - 1 - a}$. Conversely, the definition of R gives $P_0 + Q_0 = R_{mix}/J \equiv (J - a - 1)/J$. Thus, we derive for P_0 the following equation:

$$P_0 = \frac{(J - a - 1)(\mu - a - 1)}{J(\mu - 1)}. \tag{13}$$

This expression allows us to calculate the mutator allele probability q using the equivalence $\sum_{l=0}^{L} Q_l = 1 - \sum_{l=0}^{L} P_l$:

$$q = 1 - P_0 \frac{J}{J - 1} \approx \frac{a}{J - 1} + O\left(\frac{1}{\mu}\right) \tag{14}$$

At $L = 5000$, the accuracy of our analytical result is approximately 0.1%, as seen in Fig. 4. Figure 4 and Eq. (14) show that the fraction of the wild allele does not approach to 0 even for a large L.

The model with uni-directional mutations of the mutator gene and a smooth fitness landscape.
As previously mentioned, we have the mixed and mutator phases. In the mutator phase, the numerics give that the mutator sub-population dominates with $q = 1$, even for finite L, so we can ignore the first chain of transitions and reduce the system to the standard Crow-Kimura model with mutation rate μ_2 and fitness function $g(x)$.

Let us now consider the mixed phase. If $\alpha_1 = a$, $\alpha_2 \to 0$, then Eq. (8) gives $q \to 1$ for this case. Assuming an ansatz $P_l(t) \equiv P(x, t) = \exp(Lu(x, t))$, we obtain:

$$R = f(s_1) - a. \tag{15}$$

The smooth curves in Figs 5 and 6 obtained by numerical calculations illustrate that the $1 - q$ strongly depends on the genome length, L. For the small genome length L, Fig. 7 obtained by numerical calculations support the behavior

$$q \sim a. \tag{16}$$

As has been derived in ref. 29, for the large L, we get $1 - q \ll 1$. In the Methods section, we calculate $(1 - q)$ for the system with a general smooth symmetric fitness function and parameters a, μ. For linear fitness functions $f(x) = kx$ and small values of a and $\mu_1/\mu_2 \equiv 1/\mu$, we obtain a simple expression:

$$1 - q = \exp\left[\frac{-La^2}{2(\sqrt{k^2 + 1} - 1)\mu}\right]. \tag{17}$$
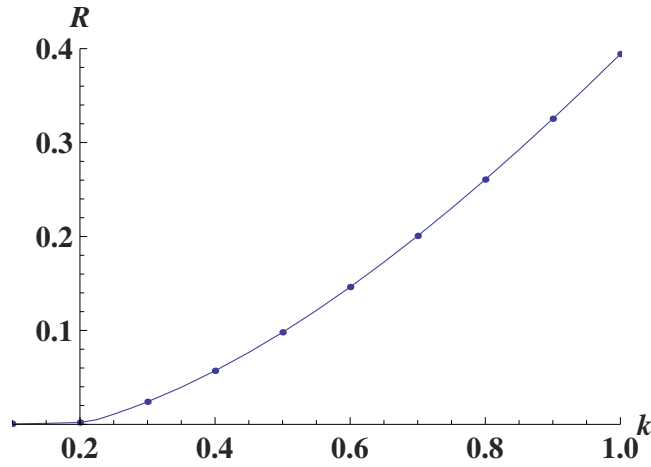
Figure 3. The mean fitness R versus k of the model with linear fitness landscape $f(x) = kx$, $\mu_1 = 1$, $\mu_2 = 10$, $\alpha_2 = 0$, $\alpha_1 = 0.02$. There are two phases in the model for the general values of parameters: mixed phase with $R = \sqrt{k^2 + 1} - \alpha_1 - 1$ and mutator phase with $R = \sqrt{k^2 + (\mu_2)^2} - \mu_2$, $q = 1$. The border between two phases is given by equation $\alpha_1 = \mu_2 - 1 + \sqrt{k^2 + 1} - \sqrt{k^2 + (\mu_2)^2}$. In our case $k_c \approx 0.212$.

This result is well supported by the numerics, as seen in Fig. 5 and Table 4 for $1 - q \ll 1$, when the exponent is a large negative number. The transition between two regimes (16) and (17) is for the values of L where the expression of the exponent in Eq. (17) is ~1.

In population genetics, the fitness difference for one mutation is denoted by S and corresponds to the linear coefficient k; U is the total number of mutations per generation. We use the equalities $U = \varepsilon\mu_1$, $S = \varepsilon 2k/L$ and $h = \varepsilon\alpha_1$, where ε is the discrete time step of one generation and h is the transition probability to the mutator allele during one generation. Applying the discrete-time Eigen model to Crow-Kimura model mapping, we get a condition for the exponent:

$$\frac{L\alpha_1^2}{\mu_2 k} \equiv \frac{2h^2}{U\mu S} \gg 1. \tag{18}$$

Otherwise, the numerics give $q \ll 1$ instead of Eq. (17).

The linear fitness case can be solved exactly. We introduce the generation functions $Q(z) = \sum_l Q_l z^l$ and $P(z) = \sum_l P_l z^l$, so rewriting Eq. (1) with these variables gives a system of equations for the generation functions:

$$L(R - k + a + (1 - z))P = (-2kz + 1 - z^2)P',$$
$$(R - k + \mu(1 - z))Q - aP = (-2kz + \mu(1 - z^2))Q'. \tag{19}$$

It is easy to calculate R from the first equation with the constraint that $P(z)$ is an L-th order polynomial. Another constraint is the probability balance condition, $P(1) + Q(1) = 1$. Solving this system, we can calculate P_l, Q_l with relative accuracy $O(1/\sqrt{L})$. The generation function method has been applied before for investigation of the Crow-Kimura model[51].

The Eigen model with a mutator gene and random fitness landscape. The probability distribution for the Eigen model is defined for genome sequences with the wild and mutator type (p_i and q_i, respectively, $0 \leq i < 2^L$). Consider now the following system of equations

$$\frac{dp_i(t)}{dt} = \sum_j p_j r_j e^{-h} Q_{ij} - p_i \sum_j r_j(p_j + q_j),$$
$$\frac{dq_i(t)}{dt} = \sum_j q_j r_j \hat{Q}_{ij} + \sum_j p_j r_j(1 - e^{-h})Q_{ij} - q_i \sum_j r_j(p_j + q_j), \tag{20}$$

where r_i is the fitness function, Q_{ij} and $\hat{Q}_{ij}$ are probabilities of mutation from S_j to S_i, $Q_{ij} = w^{L-d(i,j)}(1-w)^{d(i,j)}$, w is the probability of errorless replication per nucleotide for sequences with normal mutator gene and $1 - e^{-h} \approx h$ is the transition probability from the wild type to the mutator type. The diagonal terms of the mutation matrix are $Q_{ii} = w^L \equiv Q \equiv e^{-\gamma}$, where $\gamma = -L\ln(w) \approx L(1-w)$ is the mutation parameter in the Eigen model, $\hat{Q}_{ij} = \hat{w}^{L-d(i,j)}(1-\hat{w})^{d(i,j)}$, $\hat{w}$ is the probability of errorless replication per nucleotide for sequences with a mutator allele. The mutation parameter for the mutator type is $\hat{\gamma} = L(1 - \hat{w})$. Here, $d(i,j)$ is the Hamming distance between two sequences, S_i and S_j.
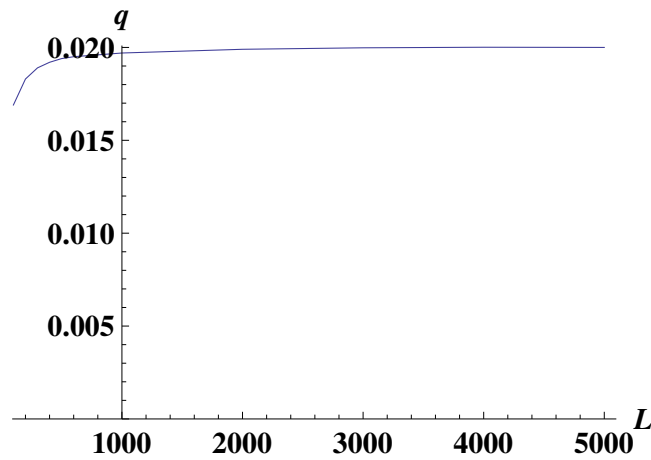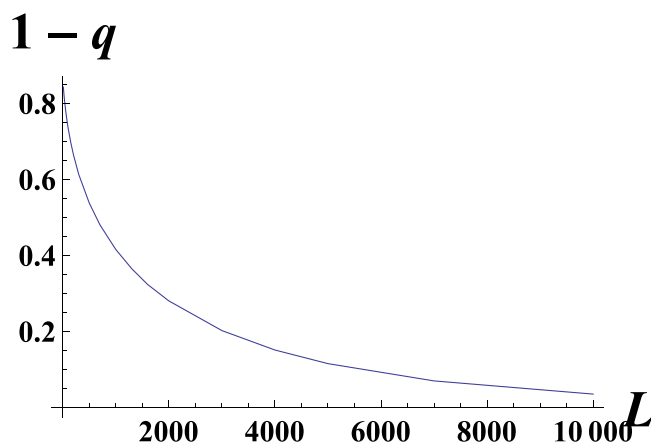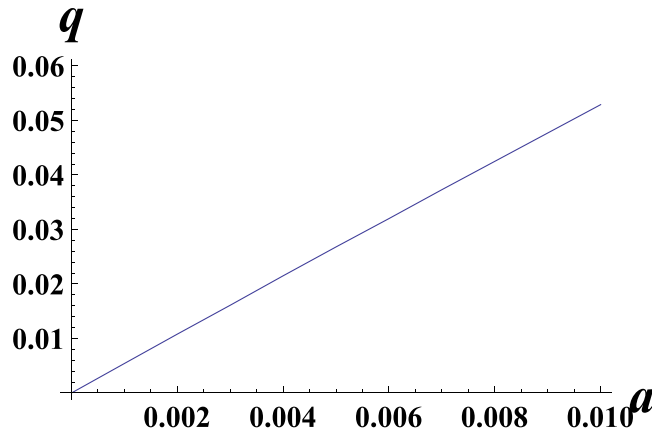
Figure 4. The dependence of the mutator probability q on the genome length L. The the single peak fitness model (smooth line) with $J = 1.05$, $\mu_1 = 1$, $\mu_2 = 10$, $a = \alpha_1 = 0.001$. For the $L = 5000$ the single peak model's numerical result coincides with the analytical result for $L = \infty$ with the relative accuracy about 0.1%.
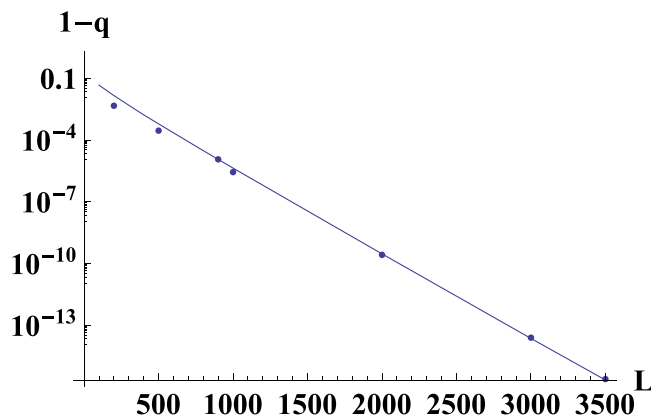
Figure 5. The dependence of the mutator probability q on the genome length L for the linear fitness model: $f(x) = kx$, $k = 1$, $\mu_1 = 1$, $\mu_2 = 10$, $\alpha_2 = 0$, $a = \alpha_1 = 0.3$. The smooth line corresponds to the numerics, the solid dots correspond to our analytical result.

Figure 6. The dependence of the mutator probability q on the genome length L for $10 < L < 10000$. We consider the linear fitness model: $f(x) = kx$, $k = 1$, $\mu_1 = 1$, $\mu_2 = 10$, $\alpha_2 = 0$, $a = \alpha_1 = 0.05$.

Figure 7. The dependence of the mutator probability q on the $a = \alpha_1$ for the linear fitness model: $f(x) = kx$, $k = 1$, $\mu_1 = 1$, $\mu_2 = 10$, $\alpha_2 = 0$, $L = 1000$.

L	1000	1000	1000	1000
k	0.3	0.3	0.3	1
a	0.0001	0.001	0.01	0.3
R_n	0.0439	0.0430	0.0340	0.1142
R	0.0439	0.0430	0.0340	0.1142
$1 - q$	0.9945	0.9460	0.530	$6/10^7$
K	0.999996	0.9994	0.930	$1/10^5$

Table 4. The results for $f(x) = g(x) = kx$, $\mu_1 = 1$, $\mu_2 = \mu$, $\alpha_1 = a$, $\alpha_2 = 0$. $K = \exp\left[L\left(\int_{s_1}^{s_3} u'(x)\,dx - \int_{s_2}^{s_3} \hat{u}'(x)\,dx\right)\right]$. R_n is the numerical result.

For the single peak fitness function with $r_0 = A$ and $r_i = 1$, $i \geq 1$ the mean fitness is similar to the one obtained for the Crow-Kimura model: QAe^{-h} for the mixed phase, $\hat{Q}A$ for the mutator phase and 1 for the non-selective phase. The random fitness landscape is a reasonable approximation for real biological data[52]. However, it was shown that it is almost equivalent to the single peak fitness landscape[33]. If we assume a log-normal distribution of Wrightian fitnesses, r_i:

$$\rho(r_i) \sim \exp\left[-L\frac{(\ln r_i)^2}{c^2}\right],$$

(21)

then, we have the maximal value of fitness $A = c\sqrt{\ln 2}$[33], while for the majority of the population, fitness is equal to 1. Using the formulae for the single peak fitness case from the Methods section, we obtain:

$$q = 1 - \frac{QAe^{-h} - 1}{QA - 1}\frac{1}{1 + \frac{Q(1 - e^{-h})}{Qe^{-h} - \hat{Q}}} \approx \frac{QAh}{QA - 1}$$

(22)

where we dropped the small term, h/μ. The steady state distribution is the same for the continuous-time Eigen model (20) and for the discrete-time Eigen model. For small values of $h \ll 1$, $U \equiv \gamma = L(1 - q) \ll 1$ and $(A - 1) \equiv S \ll 1$, we get:

$$q = \frac{h}{S - U}.$$

(23)

Equation (23) coincides with the Crow-Kimura model result in Eq. (14) after mapping $\tau J = S$, $U = \tau$, $h = a\tau$, where τ is the growing time scale of the replicator, and we identified the Crow-Kimura model dynamics after the time τ with the dynamics of the discrete time Eigen model[33].

It is interesting that there is no μ dependence in our expression. Taking the case of strong selection, $s \gg U$, we obtain the situation when $q \ll h/U$, as in the case of refs 53 and 54. According to ref. 53 (Table 1), we have $U = 0.003$, $q = 0.00001$ and $h/U = 0.0001$, while for ref. 54 (see abstract), we have $q = 3/10^5$, $h = 5/10^6$. The case of large $q = 1/408 = 0.245\%$[55], p. 109, perhaps is related to the smooth fitness landscape.

Discussion

In this work, we have investigated the mutator phenomenon in the evolutionary process in a stable environment for infinite population size. Previous studies have obtained some results for a microscopic model with uni-directional mutation (wild-type → mutator-type) and linear Malthusian fitness in a 1-dimensional landscape. The mutator and mixed phases have been considered theoretically in the literature, so we continued to focus on the phase structure in this research. We have calculated exactly, for the first time, the phase structure of the model for the general fitness function, the analytical expression for the mean fitness for any scheme of mutations between normal and mutator states in the d-dimensional fitness landscape (see Methods). Moreover, we have discussed the important case of the random fitness landscape. We have derived the formula for the mutator-type probability, q, for the large genome length. This mathematical problem is highly non-trivial, as it is non-perturbative via the value of the back mutation rate: see Eqs (8,9) and (17). Equation (17) shows that the mutator-type takes over the entire population for the model with a linear fitness function. This effect has been observed experimentally[27,56].

In ref. 17, an expression for the mutator allele probability has been derived when the mutator-type is the minority and the fitness function is linear. Our results confirm these findings[17]; furthermore, our method provides an analytical solution for the model in this case. When condition (18) is broken, the wild-type takes over the population, which has also been proven experimentally[54]. The value of q depends on the genome length ($q \ll 1$ or $1 - q \ll 1$; see Fig. 5), and for the surplus, both alternatives hold ($s_1 = s_2$ or $s_1 \neq s_2$). Thus, the finite size corrections could be large, even for the large genome length $L \sim 10^4$.

A careful investigation of the evolutionary models is essential for biological applications. Looking at possible phases with order parameters is needed to provide an appropriate level of accuracy. As long as cancer is assumed to be a collective phenomenon, even with some collective intellect[57], it is necessary to use analytical methods for complex systems. Let us compare our results with those by Desai and Fisher[29] who solved a phenomenological model, while we rigorously solved the microscopic model. For "small" genome sizes, the mutator fraction is proportional to the small parameter α, as has been found in ref. 29. What we claim is $q \to 1$ for sufficiently large genome length for any smooth fitness function. The latter case is a non-perturbative phenomenon, and it is not represented in the approach of[29], as they neglected the backward mutation between different classes of wild-types. The approach by Desai and Fisher[29] cannot be applied for the steady state of our model, as it is completely wrong for that case; however, Desai and Fisher obtained[29] a useful (especially for the cancer case) analytical estimate for the short-time dynamics for the involved case of general fitness function.

Theoretical results concerning the error threshold[58] have a practical impact for viruses and cancer, and it has been recently suggested to use an error catastrophe as a new therapeutic strategy for the solid cancer treatment[10]. A phase structure shown in Fig. 1 qualitatively describes the different stages of cancer; the mixed phase $s \neq 0$ characterizes the early non-aggressive version of tumor, and the mutator phase is an aggressive stage of cancer, possibly with metastasis. To eradicate the tumor according to the strategy suggested in ref. 9, it should be transferred to the non-selective phase $s = 0$. As seen in Fig. 1, there are four different situations. In the mixed II subphase, we can push the tumor to the non-selective phase (the eventual goal of error-catastrophe therapy) to increase the forward mutation rate for the mutator gene α_1, which is responsible for the genome stability. In the mutator I subphase, we should increase the mutation rate μ to push the tumor to the non-selective phase. In the mutator II and mixed I subphases, we need to increase both versions of mutation rates. The phase structure has the same form for the random fitness case and a similar form for other fitness landscapes.

Further research should focus on the finite population version of the mutator model, as the population size can drastically affect the evolutionary dynamics[8]. The problem has already attracted attention[19], but it is reasonable first to solve a simpler case of infinite population limit.

To summarize, let us directly reply to the four questions stated in the beginning of the Results section.

- A. There is a phase transition between two phases, selective and non-selective, for the random fitness landscape and for the last two types of fitness landscapes considered in ref. 6 (the fitness landscape is flat until some critical number of mutations, then jumps and later is flat again; the fitness decreases linearly until some number of mutations and then jumps). For the linear fitness landscape, there is no phase transition.
- B. For small enough genome length $s_1 \neq s_2$, see the Methods section for a detailed explanation.
- C. For a small genome length, it is possible to have some non-zero fraction of wild-types (see the Methods section).
- D. There are large finite L corrections for q; see Figs 5 and 6.

Methods

To perform the analytical investigation Eq. (3), we use the HJE approach. Ignoring $O(1/L)$ correction terms and using Eq. (5) and the formulae $P_{l \pm 1} = v_1 e^{Lu \pm 2u'}$ and $Q_{l \pm 1} = v_2 e^{Lu \pm 2u'}$ in Eq. (3), we get:

$$v_1\left(-u'_t + f(x) - \alpha_1 + \mu_1\left(\frac{1+x}{2}e^{2u'} + \frac{1-x}{2}e^{-2u'} - 1\right)\right) + v_2\alpha_2 = 0,$$

$$v_1\alpha_1 + v_2\left(-u'_t + g(x) - \alpha_2 + \mu_2\left(\frac{1+x}{2}e^{2u'} + \frac{1-x}{2}e^{-2u'} - 1\right)\right) = 0.$$

(24)

Here, we denoted $u' = \frac{\partial u(x,t)}{\partial x}$ and $u'_t = \frac{\partial u(x,t)}{\partial t}$. We consider Eq. (24) as a homogeneous system of equations for v_1 and v_2, so its determinant is equal to 0 for the non-zero solution of v_1 and v_2. First, we find u'_t from the latter

condition and write an equation, $u'_t + H_\pm(x, u') = 0$, where $H_\pm$ is a Hamiltonian. Then, we derive the potential function as

$$V_\pm(x) = \min \; [-H_\pm(x, p)]\big|_p \tag{25}$$

For the general case we have the following Hamiltonian:

$$-H_\pm = \frac{f(x) + g(x)}{2} - \frac{\alpha_1 + \alpha_2}{2} + \frac{\mu_1 + \mu_2}{2}\left(-1 + \frac{1 + x}{2}e^{2u'} + \frac{1 - x}{2}e^{-2u'}\right)$$
$$\pm \frac{1}{2}\sqrt{\left(f(x) - g(x) - (\alpha_1 - \alpha_2) + (\mu_1 - \mu_2)\left(-1 + \frac{1 + x}{2}e^{2u'} + \frac{1 - x}{2}e^{-2u'}\right)\right)^2 + 4\alpha_1\alpha_2}. \tag{26}$$

Deriving $V_\pm(x)$ according to Eqs (25) and (26), we get Eq. (6) to calculate the mean fitness R.

At the maximum point $x = s$ of the distribution, we take $u' = 0$, $u'_t = -H_+ = R$ in Eq. (26) to obtain Eq. (7) for the surplus s. Putting to Eq. (24) the expression for the R, and looking the point $x = s$, we derive Eq. (8).

If f(x) = g(x), $\alpha_1 = a$, $\alpha_2 = 0$ ($\mu_1 = 1$, $\mu_2 = \mu$), then, taking into account the ansatz $P_l = e^{Lu(x,t)}$, we derive an equation:

$$R = u'_t = f(x) - a + \frac{1 + x}{2}e^{2u'} + \frac{1 - x}{2}e^{-2u'} - 1, \tag{27}$$

and the second equation for the mixed phase in Eq. (11): f(s) = R.

The mutator-type probability.

Consider the general case $s_1 \neq s_2$ and parameters $\alpha_1 = a$, $\alpha_2 = 0$, $\mu_1 = 1$, $\mu_2 = \mu$ and f(x) = g(x). For $Q_l \sim P_l$, we consider the second equation in (1) with the substitution $P_l = e^{Lu(x,t)}$ in a steady state. We derive an equation for Q_l:

$$RQ_l = a \exp[Lu(x, t)] + Q_l g(x) + \mu\left(Q_{l-1}\frac{L - l + 1}{L} + Q_{l+1}\frac{l + 1}{L} - 1\right). \tag{28}$$

We can write an approximate solution:

$$Q_l \sim \frac{a \exp[Lu(x, t)]}{\mu}. \tag{29}$$

If $P_l \ll Q_l$, then, substituting $Q_l = \exp[L\hat{u}(x, t)]$, we get the following equation for $\hat{u}$:

$$R = f(x) + \mu\left(\frac{1 + x}{2}e^{2\hat{u}'} + \frac{1 - x}{2}e^{-2\hat{u}'} - 1\right). \tag{30}$$

The expression for s_2 can be derived from R = f(s_2). We took the solution (27) at the interval $[s_3, 1]$ (where s_3 is a switching point between s_1 and s_2) and the solution (30) at the interval $[s_1, s_3]$. We use the concatenation of these two solutions, assuming the smoothness of the derivative $u'(s_3) = \hat{u}'(s_3)$, thus we obtain the following equation for s_3:

$$R - f(s_3) + a = \frac{R - f(s_3)}{\mu}. \tag{31}$$

Putting a condition u(s_1) = 0 (which is equivalent to $\sum_l P_l = 1$) and $\sum_l Q_l \sim Q_{l_0}$, $l_0 = (1 + s_2)L/2$, we get:

$$1 - q \sim \frac{a}{\mu} \exp\left[L\left(\int_{s_1}^{s_3} u'(x)\,dx - \int_{s_2}^{s_3} \hat{u}'(x)\,dx\right)\right]. \tag{32}$$

Here, u' and $\hat{u}'$ are calculated using Eqs (27) and (30), respectively. The expression in the exponent in Eq. (32) is exact at the limit $L \to \infty$.

These analytical results allow us to predict, which allele takes over the population. The analysis of Table 4 reveals, that if $K \equiv \exp\left[L\left(\int_{s_1}^{s_3} u'(x)\,dx - \int_{s_2}^{s_3} \hat{u}'(x)\,dx\right)\right] \ll 1$, then the mutator-type dominates. If $1 - K \ll 1$, then the wild-type takes over the population.

The case of small mutation rate for mutator allele *a*.

Let us suppose that forward mutation rate *a* is sufficiently small. This assumption leads to the following equations:

$$s_2 - s_1 = \frac{a}{f'(s_1)}, \; u''(s_1) = \frac{f'(s_1)}{s_1},$$
$$\hat{u}''(s_2) = \frac{f'(s_2)}{\mu s_2}. \tag{33}$$

From (31), we obtain

$$s_1 - s_3 = \frac{\alpha}{\mu f'(s_1)}. \tag{34}$$

We are looking the case of large μ, therefore, we can replace s_3 by s_1:

$$\left(\int_{s_1}^{s_3} u'(x)\,dx - \int_{s_2}^{s_3} \hat{u}'(x)\,dx \right) = \frac{-\hat{u}''(s_2)(s_2 - s_3)^2}{2}. \tag{35}$$

Thus

$$1 - q = \exp\left[\frac{-La^2}{2f'(s_1)s_1\mu} \right]. \tag{36}$$

For the linear fitness $f(x) = kx$, we get Eq. (17):

$$1 - q = \exp\left[\frac{-La^2}{2(\sqrt{k^2 + 1} - 1)\mu} \right]. \tag{37}$$

This solution is correct when the following condition is met:

$$s_1 \gg \frac{\alpha}{\mu k}. \tag{38}$$

The mean fitness for the model with *d*-dimensional fitness landscape and mutator gene. The multidimensional model has been discussed in ref. 19, where two parts of genome were considered; the first part is with lethal mutations and the second part is with deleterious and advantageous mutations. The genome is formally fractured into d parts with the lengths Ly_i. For the wild-type, the fitness is equal to $\sum_i f(x_i)$, and for the mutator-type, the fitness is $\sum_i g(x_i)$, where $x_i = 1 - 2l_i/(Ly_i)$ and l_i is the number of "−" spins (alleles) in the i-th part of genome. We denote mutation rates as μ_i for the wild-type and ν_i for the mutator-type, see ref. 49 for the multidimensional fitness model without a mutator gene. For these modified expressions, we calculate the mean fitness R as a maximum of the potential function $V_\pm$, defined as:

$$V_\pm(x) = \frac{\sum_i (f_i(x_i) + g_i(x_i) - y_i(\mu_i + \nu_i + \alpha_1 + \alpha_2))}{2}$$
$$+ \sum_i y_i \frac{\mu_i + \nu_i}{2}\sqrt{1 - x_i^2} \pm \frac{1}{2}\sqrt{(A - \alpha_1 + \alpha_2)^2 + 4\alpha_1\alpha_2},$$
$$A = \sum_i (f_i(x_i) - g_i(x_i) + y_i(\mu_i - \nu_i)(\sqrt{1 - x_i^2} - 1)). \tag{39}$$

Here α_1, α_2 have the same meaning as in Eq. (1). The model considered in ref. 19 corresponds to the values: $d = 2$, $f_1(1) = g_1(1) = 0$, $f_1(x) = g_1(x) = -\infty$ for $-1 \le x < 1$ and $f_2(x) = g_2(x) = k_2(1 - x)$, for the case of finite population.

The Eigen setting of the mutator model. Consider the first equation in the system of Eq. (20) as an equation for the eigenvalue R. Compared with the ordinary Eigen model, the matrix of the linear system is multiplied by e^{-h}. Thus, we can directly use the results from[58] and write the expression for the mean fitness of the mixed phase for the fitness function $r_i = f(x)$:

$$R = \max_x \left[e^{-h} f(x) e^{\gamma(\sqrt{1 - x^2} - 1)} \right]. \tag{40}$$

For the single peak fitness, we obtain for the selective phase[58]:

$$R = QAe^{-h}. \tag{41}$$

For the non-selective phase, we simply have $R = 1$.

In the mutator phase, we ignore the first chain of equations in Eq. (20) and get R as the mean fitness of the Eigen model with the value:

$$R = \max_x \left[f(x) e^{\hat{\gamma}(\sqrt{1 - x^2} - 1)} \right]. \tag{42}$$

Consider now the distribution of population in the single peak fitness case. Ignoring the back mutations for the equation for q_0 (because they give $O(1/L)$ terms), we obtain:

$$q_0 = \frac{(1 - e^{-h})QA}{A(Qe^{-h} - \hat{Q})} p_0. \tag{43}$$

As all the sequences, besides p_0, q_0, have a fitness value 1, we have an equation $(q_0 + p_0)A + (1 - q_0 - p_0) = R$. Therefore

$$p_0 + q_0 = \frac{QAe^{-h} - 1}{(A - 1)}. \tag{44}$$

We derive eventually:

$$p_0 = \frac{QAe^{-h} - 1}{(A - 1)\left[1 + \frac{(1 - e^{-h})Q}{(Qe^{-h} - \hat{Q})}\right]}. \tag{45}$$

In the ordinary Eigen model

$$\frac{\sum_l p_l}{p_0} = \frac{A - 1}{QA - 1}. \tag{46}$$

This equation is valid in our case as all the equations in the first chain of Eq. (20) are derived from the ordinary model[58] by multiplying on e^{-h}, moreover, we have that $Re^h = QA$. These two equations give together Eq. (22) of the text.

For the discrete-time Eigen model with a time step n, we have the probabilities $p_i(n)$, $q_i(n)$ and the following iteration equations:

$$p_i(n + 1) = \frac{\sum_j p_j(n)e^{-h}Q_{ji}}{\sum_j r_i(p_i(n) + q_i(n))},$$

$$q_i(n + 1) = \frac{\sum_j q_j(n)\hat{Q}_{ji} + \sum_j p_j(n)(1 - e^{-h})Q_{ji}}{\sum_j r_i(p_i(n) + q_i(n))}. \tag{47}$$

References

1. Loeb, L. A., Springgate, C. F. & Battula, N. Errors in DNA replication as a basis of malignant change. *Cancer Res.* **34,** 2311–2321 (1974).
2. Loeb, L. A., Loeb, K. R. & Anderson, J. P. Multiple mutations and cancer. *Proc. Natl. Acad. Sci. USA* **100,** 776–781 (2003).
3. Lynch, M. The cellular, developmental and population-genetic determinants of mutation-rate evolution. *Genetics* **180,** 933 (2008).
4. Soderberg, R. J. & Berg, O. G. Kick-starting the ratchet: the fate of mutators in asexual populations. *Genetics* **187,** 1129 (2011).
5. Beckman, R. A. & Loeb, L. A. Efficiency of carcinogenesis with and without a mutator mutation. *Proc. Natl. Acad. Sci. USA* **103,** 14140–14145 (2006).
6. Beckman, R. A. Mutator Mutations Enhance Tumorigenic Efficiency across Fitness Landscapes. *PLoS One* **4,** e5860 (2009).
7. Datta, R. S. *et al.* Modelling the evolution of genetic instability during tumour progression. *Evolutionary Applications* **6,** 20–33 (2013).
8. Raynes, Y., Gazzara, M. R. & Sniegowski, P. D. Contrasting dynamics of a mutator allele in asexual populations of different size. *Evolution* **66,** 2329–2334 (2012).
9. Fox, E. J. & Loeb, L. A. Lethal mutagenesis: targeting the mutator phenotype in cancer. *Semin Cancer Biol.* **20,** 353–359 (2010).
10. Loeb, L. A. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature Review Cancers* **11,** 450–457 (2011).
11. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57 (2000).
12. Nowell, C. P. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
13. Merlo, L. M. F., Pepper, J. W., Reid B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature Review Cancers* **6,** 924–935 (2006).
14. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–311 (2012).
15. Taddei, F. *et al.* Role of mutator alleles in adaptive evolution. *Nature* **387,** 700–702 (1997).
16. Kessler, D. & Levine, H. Mutator dynamics on a smooth evolutionary landscape. *Phys. Rev. Lett.* **80,** 2012 (1998).
17. Nagar A. & Jain, K. Exact phase diagram of quasispecies model with a mutator rate modifier. *Phys. Rev. Lett.* **102,** 038101 (2009).
18. Baake, E., Baake, M. & Wagner, H. Quantum Chain is Equivalent to a Model of Biological Evolution. *Phys. Rev. Lett.* **78,** 559 (1997).
19. Wylie, C. S., Ghim, C. M., Kessler, D. & Levine, H. The Fixation Probability of Rare Mutators in Finite Asexual Populations. *Genetics* **181,** 1595–1612 (2009).
20. Michor, F., Liphardt, J., Ferrari, M. & Widom, J. What does physics have to do with cancer? *Nature Reviews Cancer* **11,** 657–670 (2011).
21. Sole, R. V. Phase transitions in cancer. in Onofrio, d'. *et al.* eds *New Challenges for Cancer Systems Biomedicine* (Springer-Verlag, Berlin, 2013).
22. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory* (Harper Row, NY, 1970).
23. Eigen, M. Selforganization of matter and the evolution biological macromolecules. *Naturwissenschaften* **5,** 465–523 (1971).
24. Eigen, M. & Shuster, P. *The hypercycle: A principal of natural selforganization* (Springer, Berlin-Heidelberg, 1979).
25. Eigen, M., McCascill, J. & Schuster, P. The Molecular Quasi-Species. *Advances in Chemical Physics* **75,** 149–263 (1989).
26. Ao, P. Orders of magnitude change in phenotype rate caused by mutation, *Cellurar Oncology* **29,** 67–69 (2007).
27. Sniegowski, P. D., Gerrish, P. J. & Lenski, R. E. Evolution of high mutation rates in experimental populations of E. coli. *Nature* **387,** 703–705 (1997).
28. Biebricher, C. K. & Eigen, M. The error threshold. *Virus research* **107(2),** 117–127 (2005).
29. Desai, M. M. & Fisher, D. S. The balance between mutators and nonmutators in asexual populations. *Genetics* **188,** 997–1014 (2011).
30. Kussell, E. & Vucelja, M. Non-equilibrium physics and evolution-adaptation, extinction, and ecology: a Key Issues review. *Rep. Prog. Phys.* **77,** 102602 (2014).
31. Saakian, D. B. & Hu, C.-K. Solvable biological evolution model with parallel mutation-selection scheme, *Phys. Rev. E* **69,** 046121 (2004).
32. Saakian, D. B. & Hu, C.-K. Eigen model as a quantum spin chain: exact dynamics. *Phys. Rev. E* **69,** 021913 (2004).
33. Kirakosyan, Z., Saakian, D. B. & Hu, C.-K. Evolution models with lethal mutations on symmetric or random fitness landscapes. *Phys. Rev. E* **82,** 011904 (2010).
34. Fisher, R. A. *The genetical theory of natural selection* (Claredon press, Oxford, 1930).
35. Wright, S. Evolution in Mendelian populations. *Genetics* **16,** 97 (1931).

36. Davies, P. C. W. & Lineweaver, C. H. Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors. *Phys. Biol.* **8,** 015001 (2011).
37. Saakian, D. B., Biebricher, C. K. & Hu, C.-K. Lethal Mutants and Truncated Selection Together Solve a Paradox of the Origin of Life. *PLoS One* **6,** e21904 (2011).
38. Kirakosyan, Z., Saakian, D. B. & Hu, C.-K. Eigen model with correlated multiple mutations and solution of error catastrophe paradox in the origin of life. *J. Phys. Soc. Jpn.* **81,** 114801 (2012).
39. Brush, S. G. History of the Lenz-Ising Model. *Rev. Mod. Phys.* **39,** 883–893 (1967).
40. Hu, C.-K. Historical Review on Analytic, Monte Carlo, and Renormalization Group Approaches to Critical Phenomena of Some Lattice Models. *Chin. J. Phys.* **52,** 1–76 (2014).
41. Baake E. & Wagner, H. Mutation-selection models solved exactly with methods of statistical mechanics. *Genet. Res.* **78,** 93–117 (2001).
42. Saakian, D. B. A new method for the solution of models of biological evolution: Derivation of exact steady-state distributions. *J. Stat. Phys.* **128,** 781 (2007).
43. Sato, K. & Kaneko, K. Evolution equation of phenotype distribution: General formulation and application to error catastrophe. *Phys. Rev. E* **75,** 061909 (2007).
44. Saakian, D. B., Rozanova, O. & Akmetzhanov, A. Exactly solvable dynamics of the Eigen and the Crow-Kimura model. *Phys. Rev. E* **78,** 041908 (2008).
45. Yakushkina, T., Saakian, D. B., Bratus, A. & Hu, C.-K. Evolutionary games with randomly changing payoff matrices. *J. Phys. Soc. Japan* **84,** 064802 (2015).
46. Yakushkina, T., Saakian, D. B. & Hu, C.-K. Exact Dynamics for a Mutator Gene Model. *Chin. J. Phys.* **53,** 100904 (2015).
47. Woodcock, G. & Higgs, P. G. Population evolution on a multiplicative single-peak fitness. *J. Theor. Biol.* **179,** 61–73 (1996).
48. Thompson, C. J. & McBride, J. L. On The Eigen's theory of selforganization of matter. *Math. Biosci.* **21,** 127 (1974).
49. Saakian, D. B. & Galstyan, V. Dynamics of the Chemical Master Equation, a strip of chains of equations in d-dimensional space. *Phys. Rev. E* **86,** 011125 (2012).
50. Saakian, D. B., Hu, C.-K. & Khachatryan, H. Solvable biological evolution models with general fitness functions and multiple mutations in parallel mutation-selection scheme. *Phys. Rev. E* **70,** 041908 (2004).
51. Bratus, A. S., Novozhilov, A. S. & Semenov, Y. S. Linear algebra of the permutation invariant Crow-Kimura model of prebiotic evolution. *Math. Biosciences* **256,** 42–57 (2014).
52. Sanjuan, R., Moya, A. & Elena, S. F. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. USA* **101,** 8396 (2004).
53. Ninio, J. Transient Mutators: A Semiquantitative Analysis of the Influence of Translation and Transcription Errors on Mutation Rates. *Genetics* **129,** 957 (1991).
54. Boe, L., Danielsen, M., Knudsen, S., Petersen, J. B., Maymann, J. & Jensen P. R. The frequency of mutators in populations of Escherichia coli. *Mutation Research* **448,** 47–55 (2000).
55. Gross, M. D. & Siegel, E. C. Incidence of mutator strains in Escherichia coli and coliforms in nature, *Mutat. Res.* **91,** 107–110 (1981).
56. Chao, L. & Cox, E. C. Competition between high and low mutating strains of Escherichia coli. *Evolution* **37,** 125–134(1983).
57. Tarabichi, M. *et al.* Systems biology of cancer: entropy, disorder, and selection-driven evolution to independence, invasion and swarm intelligence. *Cancer Metastasis Rev*. **32,** 4031 (2013).
58. Saakian D. B. & Hu, C.-K. Exact solution of the Eigen model with general fitness functions and degradation rates. *Proc. Natl. Acad. Sci. USA* **103,** 4935 (2006).

Acknowledgements

Author Contributions

D.B.S. and C.-K.H. designed and performed the research. T.Y. did numerical calculations and participated in the research implementation. All authors wrote and reviewed the manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Saakian, D. B. *et al.* The rich phase structure of a mutator model. *Sci. Rep.* **6**, 34840; doi: 10.1038/srep34840 (2016).