Evolutionary Innovation by Polyploidy

Tetsuhiro S. Hatakeyama*

Earth-Life Science Institute (ELSI),

Tokyo Institute of Technology, 2-12-1-IE-1 Ookayama,

Meguro-ku, Tokyo 152-8550, Japan

Ryudo Ohbayashi

Department of Biological Sciences, Tokyo Metropolitan University,

1-1 Minami-Osawa, Hachioji-shi, Tokyo 192-0397, Japan

(Dated: April 15, 2024)

Abstract

The preferred conditions for evolutionary innovation is a fundamental question, but little is known, in part because the question involves rare events. We focused on the potential role of polyploidy in the evolution of novel traits. There are two hypotheses regarding the effects of polyploidy on evolution: Polyploidy reduces the effect of a single mutation and slows evolution. In contrast, the gene redundancy introduced by polyploidy will promote neofunctionalization and accelerate evolution. Does polyploidy speed up or slow down evolution? In this study, we proposed a simple model of polyploid cells and showed that the evolutionary rate of polyploids is similar to or much slower than that of haploids under neutral selection or during gradual evolution. However, on a fitness landscape where cells should jump over a lethal valley to increase their fitness, the probability of evolution in polyploidy could be drastically increased, and the optimal number of chromosomes was identified. We theoretically discussed the existence of this optimal chromosome numbers from the large deviation theory. Furthermore, we proposed that the optimization for achieving evolutionary innovation could determine the range of chromosome number in polyploid bacteria.

I. INTRODUCTION

How novel traits in organisms have emerged is one of the most critical questions in biology. Through decades of theoretical development in population genetics [1–3] and recent large-scale data analysis and laboratory evolutionary experiments [4–6], we have learned a great deal about the quantitative evolution of traits. However, the mechanism of evolutionary innovation is challenging to investigate because we must consider the probability of rare events occurring. In this aim, we consider the conditions under which novel traits are likely to evolve from existing organisms believed to have been more likely to acquire evolutionary innovation.

We first focus on polyploid bacteria, particularly cyanobacteria. Cyanobacteria are more diverse regarding their morphology and habitats than most other bacteria [7]. They range from unicellular cocci and rods to bead-like and branching multicellular forms that exhibit cell differentiation, and their habitats range from underwater to in the soil and above the

* hatakeyama@elsi.jp

ground. This indicates that cyanobacteria are more likely to achieve evolutionary innovation than most other bacteria. One of the significant differences between cyanobacteria and other bacteria, other than photosynthesis, is their polyploid nature [8]. Interestingly, most marine cyanobacteria, although photosynthetic, are known to be haploid and possess a simple unicellular morphology [8]. Such polyploidy has also been reported in halophilic archaea living in extreme environments [9]. Although why these species have been able to diversify and migrate to extreme environments is unclear, the diversity of polyploid bacteria and archaea implies the potential impact of polyploidy on the evolution of novel traits. Although several studies examined the physiological significance of polyploidy in bacteria [10, 11], the evolutionary impact of polyploidy has rarely been studied.

The impact of polyploidy on evolutionary innovation has also been demonstrated in medicine in recent years. For example, cell wall-deficient/defective bacteria, i.e., the L-form of bacteria, which are known to be polyploid [12], were reported to easily develop resistance to multiple drugs [13, 14]. Furthermore, in the last decade, it has been intensively reported that tumor cells become large and polyploid through endoreplication or cell fusion under stress conditions [15, 16]. Such cells are termed polyploid giant cancer cells (PGCCs). These cells are recognized to be important for the development of resistance to various cancer therapies, e.g., chemotherapy, hormone therapy, immunotherapy, and systemic radiation therapy, and for metastasis, i.e., for achieving novel traits. Although the need to understand the influence of polyploidy on evolution is increasing even in the medical field, few results have been obtained.

How can polyploidy actually promote the evolutionary innovation? There are two conflicting hypotheses. In polyploidy, even if favorable or unfavorable mutations are introduced into one genome, the effect will be weakened by the presence of a large number of unmutated genomes [17], and the resulting phenotypic changes will be smaller than those in haploidy. According to Fisher's fundamental theorem of natural selection [1], the evolutionary rate should be proportional to the genetic variance of the fitness, and the evolutionary rate should be slower in polyploid organisms because the genetic variance of the phenotype is small in such organisms. A similar point was raised theoretically in a previous study as the minority control of genetic information [18]. By contrast, polyploidy is the same situation as when whole-genome duplication occurs. Such whole-genome duplication was postulated by Susumu Ohno as a source of neofunctionalization attributable to the redundancy of dupli-

cated genes [19]. According to this idea, the evolution of novel traits appears more likely to occur in polyploidy. Is one of these hypotheses correct, or can both apparently contradictory hypotheses coexist?

In this paper, we first proposed one of the simplest models of the evolution of polyploid organisms. We demonstrated that the evolutionary rate in polyploidy is at best unchanged or is much slower than that in haploidy under neutral selection or on the smooth fitness landscape. This is because, as mentioned previously, the genetic variance of the phenotype is lower in polyploids than in haploids. However, the evolutionary dynamics of the same model were very different in the multimodal fitness landscape, where cells should jump over a lethal valley to increase their fitness. Although polyploidy still reduced the genetic variance of the phenotype, the probability of evolving a novel trait increased with increasing ploidy levels, and an optimal number of chromosomes existed. We demonstrated with the aid of the large deviation theory that the evolution of such novel traits could be attributed to a bias in genetic information on chromosomes in polyploidy. Furthermore, we hypothesized that the range of genome number in various cyanobacterial species results from the optimization of chromosome bias and hence the evolutionary innovation.

II. MODEL

We considered a simple case in which phenotype depends on a translation product from a single locus on chromosomes. That product is assumed to be transcribed and translated from all chromosomes at an equal rate, as experimentally observed [10, 20], and it has activity represented by a single scalar quantity, namely $x_{i,j}(t)$ for a jth chromosome in an ith cell at tth generation. Alternatively, we can consider the case in which various genes on the chromosome contribute to a single trait, and the sum of their contributions is represented as a single scalar quantity. Even in that case, it does not affect the conclusion. Every cell has N homologous chromosomes that carry somewhat different copies of the same genes; and thus, the genotype is given as a set of $x_{i,j}(t)$, and the phenotype $y_i(t)$ is given as a function of that set. The fitness of ith cell is represented as $f(y_i(t))$ as a function of the phenotype. Inherited chromosomes are mutated every generation by adding a random value as $x_{i,j}(t+1) = x_{i',j'}(t) + \mathcal{N}(0, \sigma^2)$, where $\mathcal{N}(0, \sigma^2)$ is a random number sampled from the Gaussian distribution with a mean of 0 and variance of σ^2. We set σ as one unless otherwise

noted. For the sake of simplicity, we assumed that the phenotype is determined by the averaged activity of the product, i.e., the fitness is given by $y_i = \langle x_{i,j} \rangle_{\mathrm{c}}$, where $\langle z \rangle_{\mathrm{c}}$ is the average of z over chromosomes given by $\langle z \rangle_{\mathrm{c}} = \Sigma_j z_j / N$. We also defined $\langle z \rangle_{\mathrm{p}}$ as the average of z over the population given by $\langle z \rangle_{\mathrm{p}} = \Sigma_i z_i / M$, where M is the total population and set as 1000 in simulations. In addition, $\langle z \rangle_{\mathrm{ens}}$ indicated the ensemble average among different samples. We set the number of samples as 10,000 with a different random seed.

We investigated the extent to inheritance modes affect the evolutionary rate. We defined two different modes of inheritance for polyploidy to investigate how such differences affect evolution. We introduced two extreme modes: set inheritance and random inheritance modes (Fig. 1A). The set inheritance mode is similar to the chromosome segregation in eukaryotes in that daughter cells inherit a set of chromosomes from a mother cell. All of the chromosomes in the mother cell are precisely duplicated individually and inherited by daughter cells in the same manner. Thus, duplication of the same chromosome does not occur. Conversely, in the random inheritance mode, daughter cells randomly inherit the mother cell's chromosomes. Each chromosome is randomly chosen and inherited until the number of chromosomes reaches the same number as that in the ancestor. Thus, multiple copies of some chromosomes may be inherited, whereas other chromosomes may not be inherited at all. Organisms lacking have sophisticated DNA replication control mechanisms, such as bacteria, are believed to have inherit chromosomes via the random inheritance mode. At least, it was experimentally observed that polyploid bacteria replicate chromosomes in a one-by-one manner [10] opposed to simultaneously as observed in eukaryotic cells.

Each cell is selected according to fitness, whereas each chromosome is neutrally selected in each cell because it usually has no fitness. Thus, chromosomes in the population are only selected through selection at the cell level.

III. RESULTS

A. Evolutionary rate drastically depends on the mode of inheritance in polyploidy

First, we considered that each cell is neutrally selected, and therefore, each chromosome is also neutrally selected. In this situation, evolutionary speed can be simply defined as the rate of phenotype changes, i.e., the diffusion constant in the phenotypic space (D). Because

the phenotype exhibits Brownian motion in the phenotypic space, the diffusion constant is proportional to the variance of $y_i(t)$. Thus, we must consider the dependence of the variance of phenotype on the mode of inheritance.

In the set inheritance mode, each chromosome is independently mutated in the same manner, and thus, $x_{i,j}(t)$ has no correlation for different j. In other words, each $x_{i,j}(t)$ is an independent and identically distributed random variable (i.i.d.) from the probability distribution obeyed by a single genotype at time t. Thus, the variance of $\sum_j x_{i,j}$ is proportional to N, and then the variance of y_i is proportional to $1/N$ because it is given by the summation of $x_{i,j}$ divided by N. Therefore, the evolutionary rate should decrease if the number of chromosomes increases in the set inheritance mode. Intuitively, it corresponds that an effect of a mutation in a single chromosome is divided by the number of chromosomes. Indeed, in the numerical calculation, the variance of phenotype was proportional to N^{-1} (Fig. 1B), and inverse of the first passage time, which is proportional to $D^{-3/2}$ [21], was proportional to $N^{-3/2}$ (Fig. 1C). Additionally, even if we consider evolution on any fitness landscape instead of neutral evolution, the genetic variance of fitness should be smaller in polyploid organisms with the set inheritance mode because of the decrease in the phenotypic variance. Hence, the evolutionary rate should be slower in any situation (Fig. 1E).

Contrarily, in the random inheritance mode, because a single chromosome may be duplicated multiple times, $x_{i,j}(t)$ may have a correlation for different values of j. In this situation, the number of chromosomes is fixed, and each chromosome is neutrally selected. This situation is similar to the gene fixation process through the neutral selection in a finite population, in which the single mutation is fixed within the same generation as the number of individuals in the population, on average [22]. In the context of polyploidy, this indicates that all chromosomes are duplicated from one chromosome N generations ago on average. In other words, the number of generations in which all chromosomes are replaced by those replicated from a single chromosome equals the number of chromosomes possessed by the cell. Thus, a set of chromosomes accumulates N generations of mutations, and the variance of chromosomes is given as $N\sigma^2$ because an independent mutation with the variance σ^2 is applied to a chromosome in each generation (Fig. 1D). Since randomly chosen chromosomes from the mother cell construct the chromosome set of the daughter cell, each $x_{i,j}$ is an i.i.d. from the probability distribution given by the mother cell's chromosome set. Thus, the variance of $\sum_j x_{i,j}$ is proportional to $N^2\sigma^2$, and then the variance of y_i is proportional to σ^2

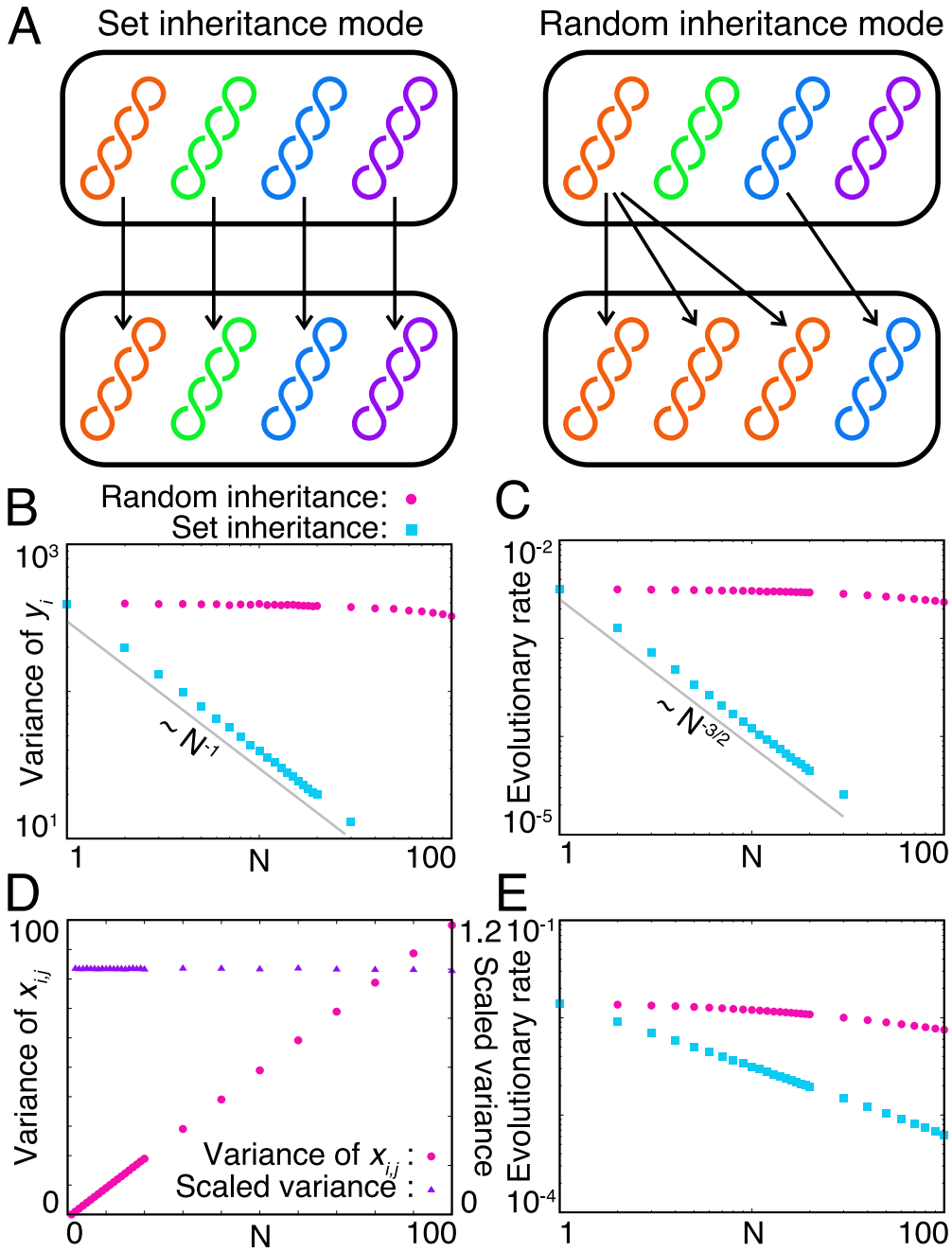


FIG. 1. Dependence of evolution on modes of chromosome inheritance. A) Scheme of modes of chromosome inheritance. B and C) Dependence of (B) the variance of phenotype and (C) the evolutionary rate on the number of chromosomes. The variance of the phenotype was calculated at time $t = 500$. The evolutionary rate was calculated as the inverse time when an absolute value of the phenotype of a cell exceeded the threshold $y_{\mathrm{th}} = 50$. Each point represents an average of 10,000 samples. Magenta circles and cyan squares represent data for the random inheritance and set inheritance modes, respectively. D) Dependence of the population average of variance among chromosomes with the random inheritance mode on the number of chromosomes. Purple triangles represent the variance among chromosomes scaled by dividing by $N - 1$. E) Evolutionary rate on the smooth landscape ($f(y) = -2\cos(x/15)$). We set $x_{i,j}(0) = 0$ as the initial condition and

(Fig. 1B), which is independent of N. Indeed, in the numerical calculation, the phenotypic variance and inverse of the first passage time were almost constant against changes in N (Fig. 1C). Note that the numerically calculated evolutionary rate was slightly slower, but this was because we set $x_{i,j}(0) = 0$ for all i and j and it takes N generations for the variance of $x_{i,j}$ to relax.

Because the set and random inheritance modes represent the extreme cases, an organism with an imperfect DNA replication mechanism will exhibit an intermediate trait between two extreme modes. This suggests that from the strong constraint of the inheritance manner, the genetic variance of the phenotype of polyploid organisms cannot exceed that of haploid organisms. Indeed, the evolutionary rate of polyploid organisms on a smooth fitness landscape, which corresponds to the gradual evolution of quantitative traits, did not exceed that of the haploid organism with either the random or set inheritance mode (Fig. 1E). Thus, polyploidy does not confer an advantage regarding evolution speed during neutral selection or gradual evolution.

B. Evolution of novel traits can be accelerated in polyploid organisms

Next, to investigate whether polyploidy is not also advantageous outside of neutral selection and gradual evolution, i.e., for the evolution of novel traits, we introduced a simplified multimodal fitness landscape with global and local maximal peaks (Fig. 2A). This is one of the simplest models for evolutionary innovation. On the high dimensional fitness landscape, if one peak is connected to a second peak by the gradual ridgeline, on which the fitness slightly decreases and selection is almost neutral, then cells will evolve their traits through that line from one phenotype to another. Then, continuous quantitative traits will be observed between the two peaks, and no one will consider this situation as an appearance of evolutionary innovation. Conversely, when deep valleys surround one peak in all directions, under which situation the fitness decreases lethally, cells must jump over a valley to reach another peak. Then, the intermediate phenotype between two peaks is difficult to observe, and it corresponds well with the situation in which evolutionary innovation appears. The multimodal fitness landscape presented in Fig. 2A is regarded as the projection of a path connecting two peaks of a high-dimensional fitness landscape onto one dimension. The distance between the peaks is much larger than the mutational rate, which is the unique

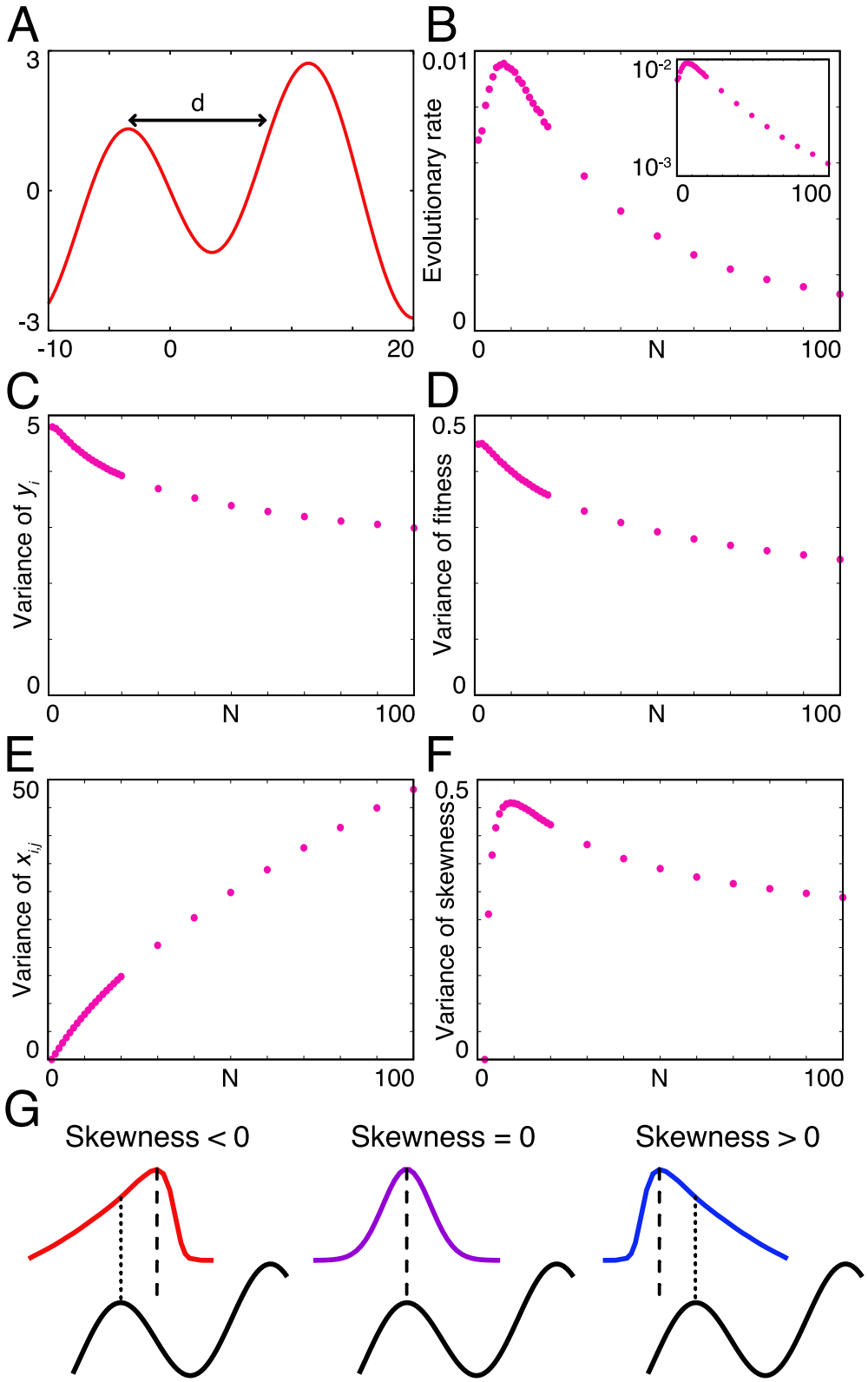


FIG. 2. Evolution of a novel trait in polyploid organisms. A) The multimodal fitness landscape given by $f(y) = \sin(y/5) - 2\sin(2y/5)$. The local maximum is $y = -3.441212$, and the global minimum is $y = 11.374583$. d is the distance between a local maximal point and a distant point displaying the same fitness as the local maximum. B) Evolutionary rate ($1/\langle\tau_t\rangle_{\text{ens}}$) on the multimodal fitness landscape. (Inset) Semi-log plot of the evolutionary rate. C) Variance of phenotype

characteristic length in the phenotypic space determining the behavior of the model, and rare and drastic changes in genotype are required for a jump from one peak to another.

We assumed that all cells initially exhibited a locally but not globally optimal trait, i.e., $x_{i,j}(0)$ and thus, $y_i(0)$ was set as the value at the local maximum for all i and j. Then, we numerically calculated the time needed to jump to the maximal peak if we changed the number of chromosomes carried by the cells. We first considered that all cells exhibited the random inheritance mode because the set inheritance mode of polyploidy decelerates the evolutionary rate regardless of the shape of the fitness landscape, as previously mentioned (Fig. 1).

When observing a single sample, the fitness of the fittest cell maintains the local maximum for some time and then suddenly jumps to the global maximum. We defined the transition time τ_t when the average fitness of the population crosses a threshold (we set it as 1.5). The jump from the local maximum to the global maximum follows a Poisson process, and the distribution of τ_t follows an exponential distribution. Hence, the characteristic timescale for transition is described by the average of τ_t, i.e., the evolutionary rate is represented by $1/\langle\tau_t\rangle_{\mathrm{ens}}$.

The evolutionary rate exhibited non-monotonic dependence on the number of chromosomes (Fig. 2B), as observed in the experiments. When N increased, the evolutionary rate increased, peaking around $N = 10$. Then, the rate began to decline as an exponential of $-N$ (see the inset of Fig. 2B). Thus, the existence of an optimal number of chromosomes for the evolutionary rate of novel traits was demonstrated in experiments and simulations.

To explain the existence of an optimal number of chromosomes for the evolutionary rate, we measured characteristics in steady state around the local maximum by setting the fitness of cells across the valley as zero. If the number of chromosomes increases, the variance of $x_{i,j}$ almost linearly increases for small values of N, and the phenotypic variance and genetic variance of fitness monotonically decrease (Fig. 2C, D, and E). Note that the variance of $x_{i,j}$ for large values of N increases only sublinearly, in contrast to the case of neutral selection. This may be because a cell with a genotype with excessively higher variance tends to fall into the valley and dies for large values of N. This suggests that changes in the genetic variance of fitness are not sufficient to explain the dependence of the evolutionary rate on the number of chromosomes: We cannot apply Fisher's fundamental theorem for natural selection to the evolution of novel traits in polyploidy organisms.

C. Evolutionary innovation is described by the large deviation theory and it depends on the third-order moment of chromosomes

The major difference between evolution under neutral selection and that of novel traits is the rarity of evolutionary events. The former is obeyed by mundane mutations that accumulate over several generations, whereas the latter is obeyed by rare mutations introduced over a few generations. Hence, we applied the large deviation theory to describe a rare event in which a daughter cell exhibits a largely different phenotype from the mother cell.

We considered the probability that a phenotype of a daughter cell crosses the valley between two phenotypes at the local and global maxima on the fitness landscape. The chromosome set of a daughter cell was constructed as i.i.d. from the distribution of chromosomes in a mother cell, and a phenotype was given as the average of chromosomes. In this analysis, we gave the distribution of the chromosome set of a mother cell with a mutation to the next generation as a cumulant-generating function $\log M(s;N)$, where N is the number of chromosomes carried by the mother cell. We denoted the distance in the phenotypic space between the local minimum and a distant point exhibiting the same fitness as d. For the fixation of different phenotypes from the local maximum, a daughter cell must display higher fitness than that observed at the local minimum in the limit of high selection pressure. Then, the daughter cell's phenotype must at least differ more than d from the mother cell's phenotype. Such a probability of a rare event is known to follow the large deviation theory [23] as follows:

$$P\left(\sum_{j=1}^{N'} \frac{x_{i,j}}{N'} \geq d\right) \simeq \exp\left(-N' I(d;N)\right), \tag{1}$$

where $I(p;N)$ is a rate function as given by the Legendre transformation of a cumulant-generating function as $I(p;N) = \sup_s \left(sp - \log M(s;N)\right)$, where N' is the number of chromosomes in the daughter cell. Although the probability on the left-hand side converges to the value on the right-hand side when N is infinite, the large deviation theory will give a valid approximation if N is sufficiently large. Therefore, if a cumulant-generating function is given, then the probability of the evolution of a novel trait can be calculated.

Because it is difficult to analytically derive the cumulant-generating function of a mother cell, we first assumed that it is given as the Gaussian distribution, which only has the first- and second-order cumulants. The first-order cumulant is the average and the same

as the phenotype of the mother cell. Thus, the first-order cumulant is zero. The second-order cumulant is the variance and a function of N, $f(N)$. Then, the cumulant-generating function is given as

$$\log M(s;N) = \frac{f(N)}{2!}s^2, \tag{2}$$

and the rate function is

$$\begin{aligned} I(p;N) &= \sup_s \left(sp - \frac{f(N)}{2!}s^2 \right) \\ &= \frac{p^2}{2f(N)}. \end{aligned} \tag{3}$$

Hence, the probability that the daughter cell achieves a novel trait is

$$P\left(\sum_{j=1}^{N'} \frac{x_{i,j}}{N'} \geq d \right) \simeq \exp\left(-N' \frac{d^2}{2f(N)} \right). \tag{4}$$

In this analysis, we considered the best situation in which the daughter cell can evolve a novel trait with the highest probability, where the variance $f(N)$ is proportional to N, as in the case of neutral selection, and it be given as $\sigma^2 N$. Then, the aforementioned probability is given as

$$P\left(\sum_{j=1}^{N'} \frac{x_{i,j}}{N'} \geq d \right) \simeq \exp\left(-N' \frac{d^2}{2N\sigma^2} \right). \tag{5}$$

Because there is a constraint that mother and daughter cells carry the same number of chromosomes, i.e., N' is equal to N, the aforementioned probability is calculated as

$$P\left(\sum_{j=1}^{N} \frac{x_{i,j}}{N} \geq d \right) \simeq \exp\left(-\frac{d^2}{2\sigma^2} \right), \tag{6}$$

and it is independent of N. Thus, the evolutionary rate of novel traits is independent of the number of chromosomes. Furthermore, $f(N)$ is sublinear of N for large values of N under selection. The aforementioned probability decreases with increasing N as $\exp(-cN)$, consistent with the numerical result (the inset of Fig. 2B). Therefore, if the distribution is Gaussian, that is, if we consider up to the second-order moment, then we never explain the acceleration of evolution in polyploid organisms.

We then considered the effect of the third-order moment. We represented the skewness of distribution, the third-order cumulant normalized by 3/2 power of variance, as a function of N as $g(N)$. If the contribution of the third-order cumulant is smaller than that of the

second-order one, then the probability that the daughter cell achieves a novel trait (see Supplementary Note 1 for derivation) is

$$P\left(\sum_{j=1}^{N'} \frac{x_{i,j}}{N'} \geq d\right) \simeq \exp\left\{-N'\left(\frac{d^2}{2f(N)} + \frac{8d^3 g(N)}{3f(N)^{\frac{3}{2}}}\right)\right\}.. \tag{7}$$

Then, when the sign of skewness differs from that of d, i.e., the mode of $x_{i,j}$ is closer to a peak corresponding to the novel trait than to the mean, then the probability of the rare event to across the valley increases (Fig. 2G). Intuitively, it is interpreted that if a large bias of genetic information exists between chromosomes, then cells can hold a high frequency of mutations for novel traits without reducing their fitness.

To confirm the effect of skewness, we numerically calculated higher-order moment in steady state. The population average of skewness was almost zero because the fitness landscape around the peak was near symmetry. Then, we measured the variance of the skewness of chromosomes in the population to estimate the probability that a cell with high skewness emerges. The result revealed non-monotonic dependency on N similarly as observed for the evolutionary rate. The variance of the skewness increased as N increased for small values of N and began to decline after the peak around $N \simeq 10$ (Fig. 2F). From Eq. (7), if the variance $f(N)$ is proportional to N, then the probability that the daughter cell achieve a novel trait is

$$P\left(\sum_{j=1}^{N} \frac{x_{i,j}}{N} \geq d\right) \simeq \exp\left(-\frac{d^2}{2\sigma^2} - \frac{8d^3 g(N)}{3\sigma^3\sqrt{N}}\right). \tag{8}$$

Thus, when the dependence of skewness on N is a higher order than $\mathcal{O}(N^{\frac{1}{2}})$, the probability of achieving a novel trait increases with N, where $\mathcal{O}$ is the Landau symbol. Although the variance of skewness is only a statistical amount of the whole population and not an amount for individual cells, an increase in variance indicates an increase in the number of cells with a large absolute value of skewness. Hence, the agreement between the peaks of variance of skewness and the evolutionary rate implies that an increase in skewness in mother cells associated with an increase in the number of chromosomes accelerates the evolution of novel traits.

Why does skewness exhibit non-monotonic dependency on the number of chromosomes? Intuitively, in regions with small values of N, there are insufficient numbers of chromosomes for the distribution to be biased. By contrast, in regions with large values of N, the distribution will be close to the Gaussian distribution because of the central limit theorem, and

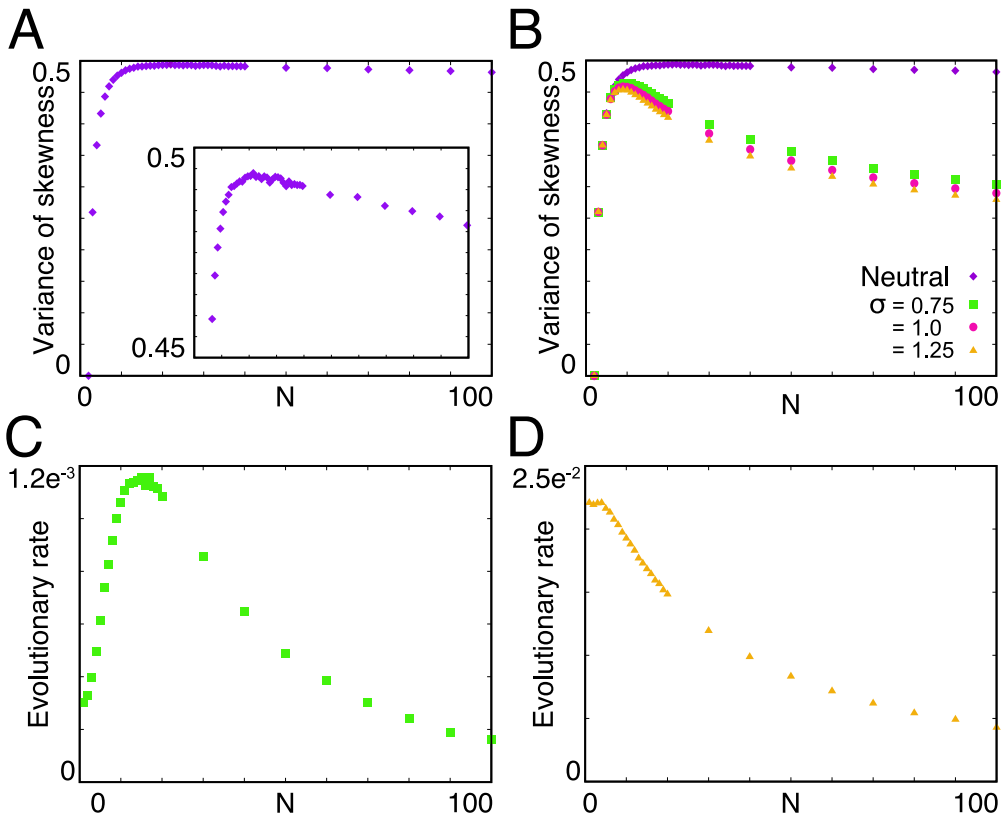


FIG. 3. Dependence of the skewness of chromosomes on selection pressure. A) Variance of skewness of $x_{i,j}$ at the time $t = 500$ under neutral selection. (Inset) Enlarged plot of the variance of skewness. B) Variance of skewness of cells with various mutation rates. Purple diamonds: Neutral selection. Green squares: $\sigma = 0.75$. Red circles: $\sigma = 1.0$. Yellow triangles: $\sigma = 1.25$. C, D) Evolutionary rate of cells with mutation rates of (C) $\sigma = 0.75$ and (D) $\sigma = 1.25$.

thus, the skewness must be small. Indeed, even if the selection is neutral, the variance of the skewness displayed a maximal value of approximately $N \simeq 20$–30 (Fig. 3A).

If there is a selection process depending on the fitness, then a cell with a large N and large skewness of $x_{i,j}$ will tend to fall into the valley and die. Then, the skewness for large values of N will be suppressed. This implies that if the range of phenotypes, in which a cell is under nearly-neutral selection, is wider, then the optimal number of chromosomes exhibiting maximum skewness and evolutionary rates will become larger. Indeed, when we set σ as a smaller value, it corresponded to a wider range of nearly-neutral selection, both maximum skewness and the evolutionary rate became larger (Fig. 3B and C). Although the value of N at the peaks of the variance of skewness and the evolutionary rate slightly differed,

this may be explained by the contribution of other moments. For the case with a smaller mutation rate, a region in which the variance of $x_{i,j}$ linearly increases expanded for larger values of N (see Supplementary Fig. 1), and thus, the peak of the evolutionary rate shifted to a larger N. Conversely, if the mutation rate is excessively large, then the evolution of novel traits is no longer a rare event, and thus, the large deviation theory cannot be applied. Consequently, the haploid again has an advantage regarding the speed of evolution, as it exhibits the highest variance in phenotype (Fig. 3B and D). Note that the normalized fourth-order moment kurtosis did not exhibit non-monotonic dependence on the number of chromosomes, and it was almost independent of the mutation rate (Supplementary Fig. 2). Hence, only skewness displayed non-monotonic dependence on the number of chromosomes at least up to fourth-order moment.

IV. DISCUSSION

In this study, we demonstrated that polyploidy confers the advantage of evolving novel traits even though the genetic variance of the phenotype does not increase with polyploidy. To elucidate this seemingly contradictory problem to Fisher's fundamental theorem of natural selection, we theoretically analyzed rare evolutionary events by introducing the large deviation theory. We found that the skewness of the distribution of genotypes coded in each chromosome is essential to determining the probability of evolving new traits. Because skewness peaked with a finite number of chromosomes (Fig. 2F), there is an optimal number of chromosomes for evolving novel traits (Fig. 2B).

Although the optimal number of chromosomes depends on the shape of the fitness landscape and the mutation rate, the upper limit of the optimal number can be estimated as approximately $20-30$ because the maximal skewness is approximately $N = 20$–30 in the case of neutral selection, and it decreases under selection. Interestingly, such a range of the number of chromosomes agrees well with the actual number of chromosomes in various species of cyanobacteria (see Table I). This suggests that the number of chromosomes in cyanobacteria might result from optimization for evolutionary innovation. Moreover, we found that the optimal number of chromosomes is correlated with the range of the fitness landscape in which a cell can change its phenotype under nearly-neutral selection. Therefore, further cooperative studies involving theoretical, experimental, and bioinformatic analyses will help

TABLE I. Typical number of chromosomes in each cyanobacterial species

Species	Number of chromosomes	Ref.
Synechococcus elongatus PCC 7942	2–6	[10, 20, 24, 25]
Synechococcus sp. PCC 6301	2–8	[26]
Anabaena variabilis	8–9	[27]
Anabaena sp. PCC 7120	8.2	[28]
Microcystis sp.	1–10	[29]
Synechococcus sp. PCC 7002	5–11	[30, 31]
Synechocystis sp. PCC 6803	9.7–22.2	[32]
Anabaena cylindrica	25	[33]
Cyanobacterium aponinum PCC 10605	16–32	[31]
Geminocystis sp. NIES-3708	17–34	[31]

us infer the fitness landscape of polyploid organisms based on the number of chromosomes.

Because our model and settings are general, our theory will be universally applicable to other polyploid organisms than cyanobacteria. In particular, our theory will be useful for organisms without sophisticated DNA replication control mechanisms. PGCCs and L-form bacteria represent promising examples of the successful application of this theory. Polyploidy in PGCCs is attributable to errors in the chromosome replication and segregation mechanism, and thus, random chromosome replication and segregation are expected to occur opposed to normal replication and segregation. Similarly, random chromosome inheritance is expected to occur in L-form bacteria, which develop polyploidy through cell fusion. Because our theory is applicable regardless of the details of the systems as long as chromosomes are inherited in random mode, we believe that our theory can unravel the mechanisms of achieving novel drug resistance and metastatic potential in cancer cells and multidrug-resistant bacteria.

In addition, we expect that our results can be applied to cases of gene conversion [34] and multi-copy plasmids [35], both of which were previously reported to be capable of facilitating the evolutionary innovation. In particular, the typical number of native plasmids ranges from approximately 10 to several dozen [36], which corresponds to the maximal skewness observed in the present study (Fig. 3A). This suggests that the copy number of native plasmids might

result from optimization for evolutionary innovation. We expect that applying our theory to plasmid systems and experiments in such systems will uncover the evolutionary advantage of the regulation of plasmid copy numbers.

Furthermore, our study has provided a new theoretical perspective on neofunctionalization by whole-genome duplication as described by Susumu Ohno [19]. Considering the cells immediately after whole-genome duplication, they would not have had sophisticated duplication mechanisms. Thus, our theory may be applicable to whole-genome duplication events, and the skewness of chromosomes may provide the theoretical representation of the early phase of neofunctionalization. This study paves the way for the understanding of evolution via whole-genome duplication, and the mechanism by which that early biased segregation of genes, as represented by skewness, is fixed is an important future problem to understand the late phase of neofunctionalization. We expect that additional theoretical studies with more complex models, bioinformatic models, and further constructive experimental studies would perfectly reveal the mechanism of neofunctionalization in the future.

ACKNOWLEDGMENTS

We thank Kunihiko Kaneko and Chikara Furusawa for a critical reading of the manuscript, and Tomoko Ohta for fruitful discussion.

Appendix A: Derivation of the rate function with third-order cumulant.

A cumulant generating function up to the third-order term is given as

$$\log M(s) = \mu s + \frac{\sigma^2}{2!}s^2 + \frac{c_3}{3!}s^3, \tag{A1}$$

where μ, σ^2, and c_3 are the average, variance, and the third-order cumulant, respectively. Here, we set μ as zero for simplicity without loss of generality. The rate function for the above cumulant generating function is

$$\begin{aligned} I(p) &= \sup_s \left(sp - \log M(s)\right) \\ &= \sup_s \left(sp - \frac{\sigma^2}{2!}s^2 - \frac{c_3}{3!}s^3\right). \end{aligned} \tag{A2}$$

Since the first derivative of $sp - \log M(s)$ is

$$\frac{d\left(sp - \log M(s)\right)}{ds} = p - \sigma^2 s + \frac{c_3}{2}s^2, \tag{A3}$$

and s for the extrema is

$$s = \frac{-\sigma^2 \pm \sqrt{\sigma^4 + 2pc_3}}{c_3}. \tag{A4}$$

The second derivative of $sp - \log M(s)$ is

$$\begin{aligned}\frac{d^2\left(sp - \log M(s)\right)}{ds^2} &= -\sigma^2 - c_3 s \\ &= -\sigma^2 + \sigma^2 \mp \sqrt{\sigma^4 + 2pc_3} \\ &= \mp\sqrt{\sigma^4 + 2pc_3},\end{aligned} \tag{A5}$$

and then,

$$s^* = \frac{-\sigma^2 + \sqrt{\sigma^4 + 2pc_3}}{c_3} \tag{A6}$$

is s for the maximal value when $\sigma^4 + 2pc_3$ is ≥ 0. Hence, the rate function is given as

$$\begin{aligned}I(p) &= s^* p - \frac{\sigma^2}{2!}s^{*2} - \frac{c_3}{3!}s^{*3} \\ &= \frac{2p^2}{\sigma^2 + \sqrt{\sigma^4 + 2pc_3}} - \frac{2p^2\sigma^2}{\left(\sigma^2 + \sqrt{\sigma^4 + 2pc_3}\right)^2} - \frac{4c_3p^3}{3\left(\sigma^2 + \sqrt{\sigma^4 + 2pc_3}\right)^3} \\ &= \frac{2p^2}{\sigma^2}\frac{1}{\left(1 + \sqrt{1 + \frac{2pc_3}{\sigma^4}}\right)^2} + \frac{8}{3}\frac{c_3p^3}{\sigma^6}\frac{1}{\left(1 + \sqrt{1 + \frac{2pc_3}{\sigma^4}}\right)^3}.\end{aligned} \tag{A7}$$

In the limit of $c_3 \to 0$, the above equation is

$$I(p) = \frac{p^2}{2\sigma^2}, \tag{A8}$$

which coincides with Eq. (3) in the main text, a case of Gaussian distribution. When the squared variance is much larger than the third-order cumulant, $|c_3| \ll \sigma^4$, Eq. (A7) is

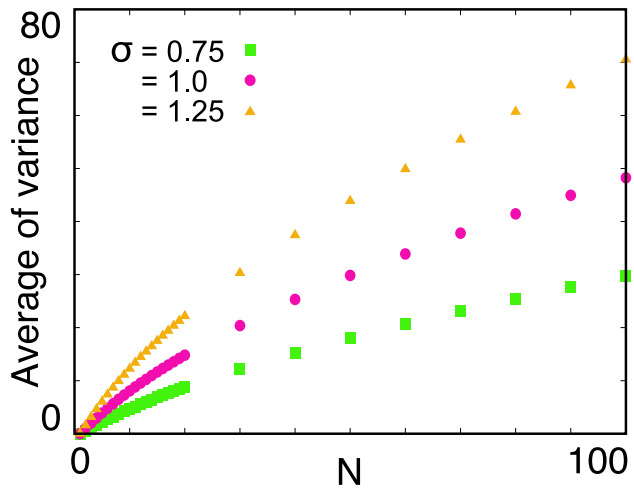
$$\begin{aligned}I(p) &\simeq \frac{2p^2}{\sigma^2} + \frac{8}{3}\frac{c_3p^3}{\sigma^6} \\ &= \frac{2p^2}{\sigma^2} + \frac{8}{3}\beta\frac{p^3}{\sigma^3},\end{aligned} \tag{A9}$$

where β is skewness, given as $\beta = \frac{c_3}{\sigma^3}$. When the variance and skewness depend on the number of chromosomes N, and are given as functions of N, $\sigma^2 = f(N)$ and $\beta = g(N)$, respectively, the rate equation is

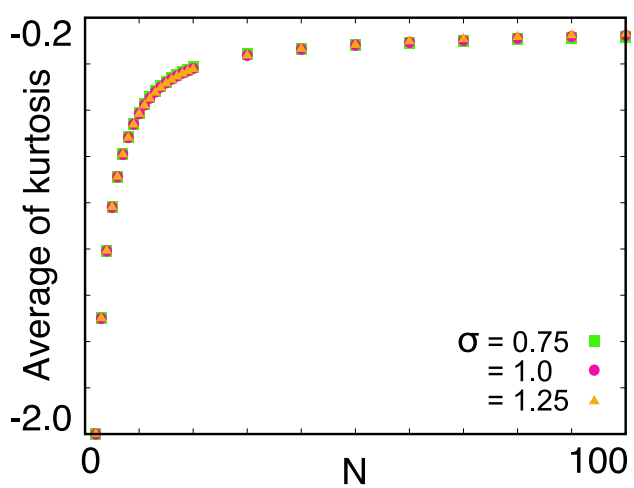
$$I(p; N) = \frac{2p^2}{f(N)} + \frac{8}{3}g(N)\frac{p^3}{f(N)^{3/2}}, \tag{A10}$$

and the probability that the daughter cell achieves a novel trait is

$$P\left(\sum_{j=1}^{N'}\frac{x_{i,j}}{N'}\geq d\right)\simeq\exp\left\{-N'\left(\frac{d^2}{2f(N)}+\frac{8d^3g(N)}{3f(N)^{\frac{3}{2}}}\right)\right\}. \tag{A11}$$



Supplementary FIG. 1. Dependence of the variance of chromosomes on the mutation rate. Average of variance of each cell over the population is plotted. Green squares: $\sigma = 0.75$. Red circles: $\sigma = 1.0$. Yellow triangles: $\sigma = 1.25$.



Supplementary FIG. 2. Dependence of the kurtosis of chromosomes on the mutation rate. The average of kurtosis of each cell over the population is plotted. Green squares: $\sigma = 0.75$. Red circles: $\sigma = 1.0$. Yellow triangles: $\sigma = 1.25$.

[1] R. Fisher, *The genetical theory of natural selection.* (The Clarendon Press, Oxford, United Kingdom, 1930).

[2] J. F. Crow, M. Kimura, *introduction to population genetics theory.* (Harper and Row, New York, NY, 1970).

[3] D. L. Hartl, A. G. Clark, *Principles of population genetics.* (Sinauer associates, Sunderland, MA, 1997).

[4] T. M. Conrad, N. E. Lewis, B. Ø. Palsson, Microbial laboratory evolution in the era of genome-scale science. *Mol. Syst. Biol.*, **7**(1), 509 (2011).

[5] A. C. Palmer, R. Kishony, Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. *Nat. Rev. Genet.*, **14**(4), 243-248 (2013).

[6] T. Maeda, J. Iwasawa, H. Kotani, N. Sakata, M. Kawada, T. Horinouchi, *et al.* High-throughput laboratory evolution reveals evolutionary constraints in *Escherichia coli. Nat. Commun.*, **11**(1), 1-13 (2020).

[7] F. G. Flores. *The cyanobacteria: molecular biology, genomics, and evolution.* (Horizon Scientific Press, Poole, UK, 2008).

[8] M. Griese, C. Lange, J. Soppa, Ploidy in cyanobacteria. *FEMS Microbiol. Lett.*, **323**, 124-131 (2011).

[9] K. Zerulla, J. Soppa, Polyploidy in haloarchaea: advantages for growth and survival. *Front. Microbiol.*, **5**, 274 (2014).

[10] R. Ohbayashi, A. Nakamachi, T. S. Hatakeyama, S. Watanabe, Y. Kanesaki, T. Chibazakura, H. Yoshikawa, S.-y. Miyagishima, Coordination of polyploid chromosome replication with cell size and growth in a cyanobacterium. *mBio*, **10**, e00510-00519 (2019).

[11] J. Paijmans, M. Bosman, P. R. Ten Wolde, D. K. Lubensky, Discrete gene replication events drive coupling between the cell cycle and circadian clocks. *Proc. Natl. Acad. Sci. U.S.A.*, **113**(15), 4063-4068 (2016).

[12] Y. Briers, T. Staubli, M. C. Schmid, M. Wagner, M. Schuppler, M. J. Loessner, Intracellular vesicles as reproduction elements in cell wall-deficient L-form bacteria. *PLOS ONE*, **7**, e38514 (2012).

[13] G. J. Domingue, H. B. Woody, Bacterial persistence and expression of disease. *Clin. Microbiol.*

Rev., **10**, 320-344 (1997).

[14] M. Leaver, P. Domínguez-Cuevas, J. M. Coxhead, R. A. Daniel, J. Errington, Life without a wall or division machine in *Bacillus subtilis*. *Nature*, **457**, 849-853 (2009).

[15] S. Zhang, I. Mercado-Uribe, Z. Xing, B. Sun, J. Kuang, J. Liu, Generation of cancer stem-like cells through the formation of polyploid giant cancer cells. *Oncogene*, **33**(1), 116-128 (2014).

[16] Y. Song, Y. Zhao, Z. Deng, R. Zhao, Q. Huang, Stress-induced polyploid giant cancer cells: unique way of formation and non-negligible characteristics. *Front. Oncol.*, **11**, 3390 (2021).

[17] A. L. Koch, Evolution vs the number of gene copies per primitive cell. *J. Mol. Evol.*, **20**(1), 71-76 (1984).

[18] K. Kaneko, T. Yomo, On a kinetic origin of heredity: minority control in a replicating system with mutually catalytic molecules. *J. Theor. Biol.*, **214**, 563-76 (2002).

[19] S. Ohno, *Evolution by gene duplication.* (Springer, Berlin, Germany, 1970).

[20] X.-y. Zheng, E. K. O'Shea, Cyanobacteria maintain constant protein concentration despite genome copy-number variation. *Cell Rep.*, **19**, 497-504 (2017).

[21] C. W. Gardiner, *Handbook of stochastic methods.* (Springer, Berlin, Germany, 1985).

[22] M. Kimura, T. Ohta, The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61**(3), 763 (1969).

[23] S. S. Varadhan, *Large deviations and applications.* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1984).

[24] A. H. Chen, B. Afonso, P. A. Silver, D. F. Savage, Spatial and temporal organization of chromosome duplication and segregation in the cyanobacterium *Synechococcus elongatus* PCC 7942. *PLOS ONE*, **7**(10), e47837 (2012).

[25] I. H. Jain, V. Vijayan, E. K. O'Shea, Spatial ordering of chromosomes enhances the fidelity of chromosome partitioning in cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **109**(34), 13638-13643 (2012).

[26] B. J. Binder, S. W. Chisholm, Relationship between DNA cycle and growth rate in *Synechococcus* sp. strain PCC 6301. *J. Bacteriol.*, **172**(5), 2313-2319 (1990).

[27] R. D. Simon, DNA content of heterocysts and spores of the filamentous cyanobacterium *Anabaena variabilis*. *FEMS Microbiol. Lett.*, **8**(4), 241-245 (1980).

[28] B. Hu, G. Yang, W. Zhao, Y. Zhang, J. Zhao, MreB is important for cell shape but not for chromosome segregation of the filamentous cyanobacterium *Anabaena* sp. PCC 7120. *Mol.*

Microbiol., **63**(6), 1640-1652 (2007).

[29] R. Kurmayer, T. Kutzenberger, Application of real-time PCR for quantification of microcystin genotypes in a population of the toxic cyanobacterium *Microcystis* sp. *Appl. Environ. Microbiol.*, **69**(11), 6723-6730 (2003).

[30] K. A. Moore, J. W. Tay, J. C. Cameron, Multi-generational analysis and manipulation of chromosomes in a polyploid cyanobacterium. *BioRxiv*, 661256 (2019).

[31] R. Ohbayashi, S. Hirooka, R. Onuma, Y. Kanesaki, Y. Hirose, Y. Kobayashi, T. Fujiwara, C. Furusawa, S.-y. Miyagishima, Evolutionary changes in DnaA-dependent chromosomal replication in cyanobacteria. *Front. Microbiol.*, **11**, 786 (2020).

[32] K. Zerulla, K. Ludt, J. Söppa, The ploidy level of *Synechocystis* sp. PCC 6803 is highly variable and is influenced by growth phase and by chemical and physical external parameters. *Microbiology*, **162**, 730-739 (2016).

[33] R. D. Simon, Macromolecular composition of spores from the filamentous cyanobacterium *Anabaena cylindrica*. *J. Bacteriol.*, **129**(2), 1154-1155 (1977).

[34] P. Bittihn, L. S. Tsimring, Gene conversion facilitates adaptive evolution on rugged fitness landscapes. *Genetics*, **207**(4), 1577-1589 (2017).

[35] J. Rodriguez-Beltran, J. C. R. Hernandez-Beltran, J. DelaFuente, J. A. Escudero, A. Fuentes-Hernandez, R. C. MacLean, *et al.* Multicopy plasmids allow bacteria to escape from fitness trade-offs during evolutionary innovation. *Nat. Ecol. Evol.*, **2**(5), 873-881 (2018).

[36] H. Yano, M. Shintani, M. Tomita, H. Suzuki, T. Oshima, Reconsidering plasmid maintenance factors for computational plasmid design. *Comput. Struct. Biotechnol. J.*, **17**, 70-81 (2019).