The Distribution of Fitness Effects of Beneficial Mutations in the Mutational Landscape Model

Lei Dai*
Department of Physics
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

(Dated: May 13, 2011)

In this paper I present theoretical predictions of the distribution of fitness effects of beneficial mutations in the framework of the mutational landscape model. The fitness effects of new beneficial mutations follow an invariant exponential distribution regardless of how fit the wild type is [1], while the mean selection coefficient s_0 decreases with the wild type fitness. The fitness effects of beneficial mutations that become fixed in the population behave as a Gamma distribution(shape parameter=2, mode=s_0) in the strong-selection, weak-mutation regime. When more than one beneficial mutations are established in the population, clonal interference shifts the mode of the distribution to a larger positive value. I show that in the scenario where the number of established beneficial mutations in the population, K, is large, the distribution of fitness effects of fixed beneficial mutations is asymptotically a Gumbel distribution [2] with its mode approximately equal to $s_0 \ln K$. The theoretical predictions are compared with simulation results and followed by a brief review of their application and limitation in certain biological scenarios.

INTRODUCTION

Adaptation is one of the most interesting yet unresolved problem for evolutionary geneticists. As Ronald A. Fisher pointed out, it is a process that a population moves towards the phenotype that best fits the environment. It is surprising that many important questions about the genetic basis of adaptation is still unanswered. [3] For example, do most adaptations involve single genes of large fitness effects? How large the effects could be? Moreover, what is the distribution of fitness effects of the mutations that are substituted in adaptation?

Fisher's geometric model laid down the theoretical ground of phenotypic adaptation. The progress in the second half of twentieth century focused on the development of models that are sequence-based. John Maynard Smith proposed that adaptation can be viewed as an adaptive walk through sequence space. The wild type samples its one-mutation-step neighbors and moves towards a local optima in the fitness landscape. The mutational landscape model put forward by John Gillespie [4] and recently developed by H.Allen Orr assumes that the fitness of beneficial mutants lies in the tail of the distribution and extreme value theory can be imported into the study of adaptation.

Based on the theoretical framework of mutational landscape model, I will discuss what we can say about the distribution of fitness effects of two sets of beneficial mutations:
1)New beneficial mutations. They are randomly produced in the population without experiencing either selection or genetic drift.
2)Fixed beneficial mutations. They have survived genetic drift and become fixed in the population.

THE DISTRIBUTION OF FITNESS EFFECTS OF NEW BENEFICIAL MUTATIONS

In the mutational landscape model, we consider a gene sequence that is L base pairs long, with 3L single-step neighbors in the sequence space. We assume that each of these 3L mutations rise with equal constant frequency. Starting at a wild type sequence, the gene can adapt by fixing any of the 3L mutants with higher fitness. This process is repeated until the wild type has the highest fitness among all its neighbors, that is, the population reaches a local optima in the fitness landscape.

Our first step is to rank all the 3L mutant sequences from high fitness to low fitness, with the rank of the best allele equal to 1. A reasonable assumption is that the wild type allele has a relatively high rank, say i. Beneficial mutations is the set of mutations with rank $j < i$. The ranked alleles form a frequency distribution. So, we want to know how the distribution of fitness effects of these new beneficial mutations looks like. Although we don't know the exact shape of the distribution, the fact that the wide type lies in the high fitness tail makes the problem mathematically tractable.

The tails of many distributions behave similarly and can be described by extreme value theory(EVT). Using EVT, H.Allen Orr proves out that the distribution of fitness effects of new beneficial mutations is exponential, and surprisingly, that the mean of the distribution is invariant with respect to the wild type fitness W_i [1]. Suppose the wild type allele has rank i, the distribution of fitness effects of new beneficial mutations is then the mixed distribution formed by considering all the possibilities of

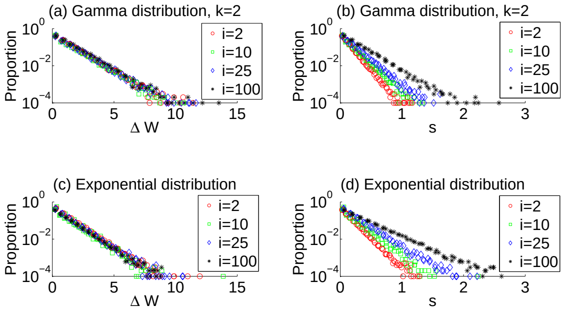

FIG. 1: **The distribution of fitness effects of new beneficial mutations.**
A gene of length L=1000 has 3000 mutants, the wild type has fitness rank i among a total of 3001 genotypes. (a),(b): Gamma distribution with shape parameter $k = 2$; (c),(d) Exponential distribution. In (a),(c), the pdf of ΔW is invariant with different ranks of wild type. In (b),(d), the mean of selection coefficient s is smaller for larger W_i after scaling(See Appendix C for simulation details).

fitness jumps, from rank i to rank $j = 1, 2, ..., i - 1$.

$$f(\Delta W \mid i) = \frac{1}{i-1}\sum_{j=1}^{i-1} f(\Delta W \mid i, j) \qquad (1)$$

where $f(\Delta W \mid i, j)$ is the probability density function(pdf) of fitness effects of a jump from rank i to rank j.

EVT shows that the fitness gaps(spacing) between allele rank j and rank j+1 $\Delta_j = W_j - W_{j+1}$ are asymptotically independent exponential distributions,

$$f(\Delta W \mid j+1, j) = \frac{1}{E[\Delta_j]} e^{-\frac{\Delta W}{E[\Delta_j]}} \qquad (2)$$

where the mean satisfies $E[\Delta_j] = \frac{E[\Delta_1]}{j}$.

Using the moment generating function, H.A.Orr proves that the distribution of fitness effects of new beneficial mutations $f(\Delta W \mid i)$ is independent of the wild type rank i(See Appendix A for an intuitive derivation). In general, for any i

$$f(\Delta W \mid i) = f(\Delta W \mid i = 2, j = 1) = \frac{1}{E[\Delta_1]} e^{-\frac{\Delta W}{E[\Delta_1]}} \quad (3)$$

To test the validity of theory, I did simulation for a gene of length L=1000, with different underlying fitness distributions and different ranks of the wild type allele(or equivalently, different W_i). It is clear that the distribution is independent of i(Figure. 1(a),(c)). ΔW can be scaled to the selection coefficient $s \equiv \frac{\Delta W}{W_i}$ by a simple factor. The distribution is still exponential, however, wide type alleles with higher fitness W_i will lead to smaller selection coefficients (Figure. 1(b),(d)).

THE DISTRIBUTION OF FITNESS EFFECTS OF FIXED BENEFICIAL MUTATIONS: TWO REGIMES

Oftentimes we are more interested to know how fixed beneficial mutations distribute. This is because in a finite population, due to genetic drift, a large fraction of produced beneficial mutations will be lost. Also, if more than one beneficial mutations survive genetic drift, they have to compete until one of them reaches fixation. This scenario, named clonal interference, can occur in a large population with restricted recombination.

In the following paragraphs, I will first focus on deriving the distribution in the mutation-limiting regime with no clonal interference. I then discuss how clonal interference affects the distribution as beneficial mutations become common enough to interfere with each other. In the case of large population size/high beneficial mutation rate, I show that the distribution will asymptotically approach an extreme value distribution under the simple assumption that the best allele always wins. Finally I compare theory with simulation results of a population undergoing Moran process.

I.The strong-selection, weak-mutation regime

We consider a haploid asexual population with effective population size N, thus N copies of genes. U_b is the beneficial mutation rate, s is the selection coefficient defined as the difference in relative growth rate.

I use the concept of established beneficial mutations to denote the mutations that have survived genetic drift(See Appendix B).The time it takes a mutant to get from establishment(or equivalently, survival of genetic drift) to being half of the population is approximately $\frac{1}{s}\ln(Ns)$, while the time between the establishment of successive mutations is $\frac{1}{NU_bs}$ [5]. Thus, if $\ln(Ns) << \frac{1}{NU_b}$, mutations fix much more rapidly than they are established. This is called the strong-selection weak-mutation(SSWM) regime, in which beneficial mutations reach fixation successionally.

The fixation probability of a single beneficial mutant with selection coefficient s is(see Appendix B),

$$P_{fix} = \frac{1 - e^{-s}}{1 - e^{-Ns}} \sim s, Ns >> 1 \qquad (4)$$

From Eq. 3, we know that the probability density of producing a new beneficial mutation with selection coefficient s, f(s), is exponentially distributed with mean s_0,

$$f(s) = \frac{1}{s_0} e^{-\frac{s}{s_0}} \qquad (5)$$

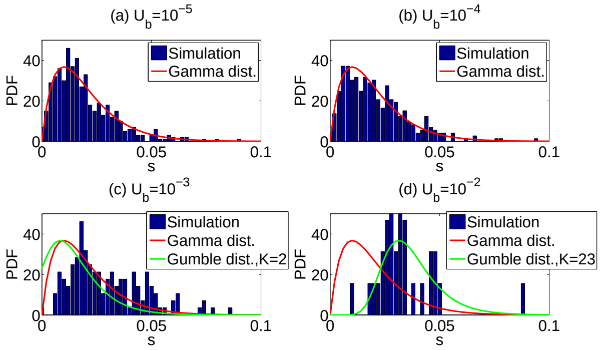

FIG. 2: **The distribution of fitness effects of fixed beneficial mutations.**
The population size is fixed at$N = 1000$, while the beneficial mutation rate in (a) to (d) ranges from 10^{-5} to 10^{-2}. As the beneficial mutation rate increases, the time between the establishment of successive mutations decreases and the criterion for SSWM regime $\ln(Ns) << \frac{1}{NU_b}$ no longer holds (See Appendix C for simulation details).

Combining Eq. 4 and Eq. 5, the probability density of fitness effects for fixed beneficial mutations, g(s), is a Gamma distribution with shape parameter equal to 2 [2],

$$g(s) = \frac{P_{fix}(s)f(s)}{\int_0^\infty P_{fix}(u)f(u)du} = \frac{1}{s_0^2} s e^{-\frac{s}{s_0}} \tag{6}$$

II.The strong-selection, strong-mutation regime

The regime where $\ln(Ns) > \frac{1}{NU_b}$ is named strong-selection, strong-mutation(SSSM). In this regime, several beneficial mutations occur and get established together by chance. They then contend for fixation with other beneficial mutations(clonal interference). I'm not going to discuss the extreme case where $\ln(Ns) >> \frac{1}{NU_b}$, which is named the weak-selection, strong-mutation regime[6].

The distribution of fitness effects of established beneficial mutations

The probability of surviving genetic drift for a single beneficial mutant is proportional to s(See Appendix B). Following the derivation of Eq. 6, we can see that the distribution of established beneficial mutations is a Gamma distribution, g(s), the same as the distribution of fixed beneficial mutations in SSWM case.

$$g(s) = \frac{1}{s_0^2} s e^{-\frac{s}{s_0}} \tag{7}$$

$$G(s) = \int_{-\infty}^{s} g(s)ds = 1 - (1 + \frac{s}{s_0})e^{-\frac{s}{s_0}} \tag{8}$$

where G(s) is the corresponding cumulative density function(cdf).

The distribution of fitness effects of fixed beneficial mutations

If we assume that the fixed beneficial mutation is the best among all the established beneficial mutations, the cdf H(s) for the fitness effects of fixed beneficial mutations is $H(s) = [G(s)]^K$, K being the the number of existing established beneficial mutations. For the case of large population size N or high beneficial mutation rate U_b, K is large, and H(s) turns into the extreme value distribution of Gumbel form(see Appendix A),

$$\begin{aligned} H(s) &= [G(s)]^K \tag{9}\\ &= [1 - \int_s^\infty g(x)dx]^K \tag{10}\\ &\approx \exp[-K \int_s^\infty g(x)dx] \tag{11}\\ &= \exp[-K(1 - G(s)] \tag{12}\end{aligned}$$

The approximation is valid if the integral is small. For large K, s is typically in the tail of g(x), so it is justified. It can be shown that g(x), a Gamma distribution, falls exponentially in its tail. So the distribution of fitness effects of fixed beneficial mutations, h(s), under the assumption that only the best mutants get fixed, is of Gumbel form,

$$h(s) = \lambda \exp[-\lambda(s - \bar{s}) - e^{-\lambda(s-\bar{s})}] \tag{13}$$

where $\bar{s} \approx s_0 \ln K$ is the most likely value of s, $\lambda = \frac{1}{s_0} - \frac{1}{\bar{s}} \approx \frac{1}{s_0}$. They are determined by the Gamma distribution $g(s)$(See Appendix A).

III.Simulation results in different regimes

In Figure. 2, I compare the distribution of fitness effects of fixed beneficial mutations generated by simulation with Eq. 6 and Eq. 13. The population size is fixed at$N = 1000$, while the beneficial mutation rate ranges from 10^{-5} to 10^{-2}. As the beneficial mutation rate increases, the time between the establishment of successive mutations decreases, the condition of SSWM regime $\ln(Ns) << \frac{1}{NU_b}$ no longer holds.

Figure. 2(a)-(b) show that in the SSWM regime($\ln(Ns) \approx 2 << \frac{1}{NU_b}$ =100,10), the distribution of fitness effects of fixed beneficial mutations follows the Gamma distribution in Eq. 6. While in the SSSM regime($\ln(Ns) \approx 2 > \frac{1}{NU_b} = 0.1$) and with a large K(estimated by $K \sim \frac{\ln(Ns_0)}{s_0} U_b N s_0$), Figure. 2(d) suggests that the distribution is asymptotically of the Gumble form. The intermediate regime, Figure. 2(c), $\ln(Ns) \sim \frac{1}{NU_b}$, has a heavier tail than the Gamma distribution and does not fit well with both distributions.

The discrepancy in Figure. 2(d) between simulation results and Gumbel distribution could be due to:

1) The lack of simulation data(sample size is small because simulation is time-consuming).
2)The inaccurate estimate of K. In a real biological scenario, we do not know how many established beneficial mutations are competing for fixation. With smaller s_0, the mean time to fixation is larger, which means more time for subsequent mutations to become established.
3)The assumption that the best mutant always becomes fixed is not valid. In fact, there is a probability that the best of K mutants will be fixed. Philip J.Gerrish and others have done a more rigorous analysis into this problem[2]. For the case of large population/high beneficial mutation rate, they show that the asymptotic expression of $H(s)$ is an extreme value distribution, with some modification of the exponent in Eq. 13.

FURTHER DISCUSSION

Due to the constraint of time, I just briefly list some of my thoughts on limitations of theory, especially in the context of its application to the evolution of antibiotic resistance.

1) The wild type does not always have a high fitness. The extreme value theory is applicable only when beneficial mutations lie in the tail of the whole distribution of mutations. For example, the assumption roughly holds for compensatory adaptation as the fitness of the genetic background is still relatively high. However, this assumption is not always valid. For example, in the presence of antibiotics, the fitness of the wild type(sensitive strain) is low, and the distribution of fitness effects of new beneficial mutations that confer resistance are no longer exponential[7]. In this case, the distribution depends on specific context and could be of any shape.

2) There could be interactions between genes of multiple loci, called epistasis. This means that fitness effects of new beneficial mutations depend on the genetic background. For example, it is found that mutations give rise to multi-drug resistance exhibit positive epistasis[8]. Because of this, the multi-drug resistant strain in an antibiotic free environment could have smaller steps of adaptive walks, meaning that the compensatory adaptation process will be slowed down[9].

In the scenarios that the mutational landscape model does apply, some experiments have been done to test the predictions about the distribution of beneficial mutations[10].

SUMMARY AND CONCLUSIONS

In this paper, I applied theory from population genetics and EVT to derive the distribution of fitness effects of beneficial mutations.

The first conclusion is that the distribution of fitness effects of generated by new beneficial mutation is exponential, and this distribution is invariant with respect to the absolute value of the wild type fitness. The selection coefficient, defined as the difference in relative fitness, also follows an exponential distribution, with its mean decreasing with the wild type fitness.

From an experimental point of view it is more interesting to know how the distribution for beneficial mutations that get fixed in the population looks like, because these mutations constitute the adaptive walk to peaks in fitness landscape. In the strong-selection weak-mutation regime, when beneficial mutations get fixed much faster than they escape from genetic drift, it is demonstrated that the fitness effects follow a Gamma distribution. However, when more mutations get established, they will compete for fixation. In the strong-selection strong-mutation regime, clonal interference shifts the mode of the distribution to a larger positive value. Under a simple assumption that only the fittest mutation becomes fixed, the asymptotic distribution in large population/beneficial mutation rate behaves as an extreme value distribution.

The theoretical prediction of the distributions fit reasonably well with a limited collection of simulation data. Its application and limitation in some biological scenarios is briefly discussed, and its merit needs to be tested in more simulation and experiments.

APPENDIX A: EXTREME VALUE THEORY

Exponential distribution of ΔW is invariant

Instead using the moment generating function, here I derive Eq. 3 for the case i=3 by convolution of two exponentials. It provides some intuition to why the distribution is the same for different fitness ranks of the wild type, i.

If the wild type has fitness rank 2, then the distribution of fitness effect of beneficial mutations, ΔW is just the distribution of the top spacing Δ_1,

$$f(\Delta W \mid i=2) = f(\Delta W \mid i=2, j=1) = \frac{1}{E[\Delta_1]} \exp \frac{-\Delta W}{E[\Delta_1]} \tag{14}$$

For i=3, the distribution $f(\Delta W \mid i=3, j=1)$ is a convolution of the top two spacings, namely $f(\Delta W \mid i=3, j=2)$ and $f(\Delta W \mid i=2, j=1)$. We also know that $f(\Delta W \mid i=3, j=2)$ is also an exponential distribution with mean $\frac{E[\Delta_1]}{2}$. This leads to,

$$f(\Delta W \mid i=3, j=1) \tag{15}$$
$$= f(\Delta W \mid i=3, j=2) * f(\Delta W \mid i=2, j=1) \tag{16}$$
$$= \frac{2}{E[\Delta_1]}[e^{-\frac{\Delta W}{E[\Delta_1]}} - e^{-2\frac{\Delta W}{E[\Delta_1]}}] \tag{17}$$

From Eq. 1 we know,

$$f(\Delta W \mid i=3) \tag{18}$$
$$= \frac{1}{2}[f(\Delta W \mid i=3, j=2) + f(\Delta W \mid i=3, j=1)] \tag{19}$$
$$= \frac{1}{E[\Delta_1]} \exp \frac{-\Delta W}{E[\Delta_1]} \tag{20}$$

So the exponential distribution is invariant for $i = 2$ and $i = 3$.

Gumbel distribution

Let $x = \max(r_1, r_2, ..., r_N)$ be the largest of N independent, identically distributed variables. When N is large, x lies in the tail of the distribution. For many common distributions, e.g.Gaussian, Exponential, Gamma, Half-Normal, etc., extreme value theory shows that the distribution of x is of Gumbel form,

$$h(x) = \lambda \exp[-\lambda(x-\bar{x}) - e^{-\lambda(x-\bar{x})}] \tag{21}$$

$\bar{x}$ is the mostly likely value of x. The pdf is determined by a single parameter λ.

Parameters in Gumbel distribution

In 8.592J Problem Set 3, we derived the Gumbel distribution for the extreme value of independent Gaussian variables and identified the parameters. A similar calculation can be done for variables obeying Gamma distribution(shape parameter k=2),

$$p(r) = \alpha^2 r e^{-\alpha r} \tag{22}$$

here $\alpha = \frac{1}{s_0}$.For $N >> 1$, the most likely value of x, $\bar{x}$ satisfies

$$p'(\bar{x}) + Np(\bar{x})^2 = 0 \tag{23}$$

Considering that $\bar{x} >> \frac{1}{\alpha}$, from Eq. 23 we get an approximate expression for $\bar{x}$

$$\alpha\bar{x}e^{-\alpha\bar{x}} = \frac{1}{N} \Rightarrow \bar{x} \approx \frac{1}{\alpha}\ln N \tag{24}$$

The value of $\lambda \approx \alpha - \frac{1}{\bar{x}}$ can be determined by Taylor expanding the exponent in Eq. 13 at $x = \bar{x}$.

The corresponding parameters in the Gumbel form distribution for exponentially distributed variables $p(r) = \alpha e^{-\alpha r}$ can be derived in a similar way. There the most likely value $\bar{x} = \frac{1}{\alpha}\ln N$, and $\lambda = \alpha$. The extreme value distribution for Gaussian variables $N(\epsilon, \sigma^2)$ has $\bar{x} \approx \epsilon + \sqrt{2\ln N}\sigma$ and $\lambda = \frac{\sqrt{2\ln N}}{\sigma}$.

APPENDIX B: POPULATION GENETICS

Probability of surviving genetic drift

Established beneficial mutations are those that survived the genetic drift. Mathematically, it can be defined as mutations that reach a population threshold $N_e = \frac{1}{s}$. The probability for a mutant species with selection coefficient s and size $n = \frac{1}{s}$ to take over a population of N individuals is,

$$P_{fix}(n = \frac{1}{s}) = \frac{1-e^{-ns}}{1-e^{-Ns}} = 1 - e^{-1} \approx 0.6, Ns >> 1 \tag{25}$$

So mutants that reach a population of N_c have a larger probability to fix than to go extinct, in which case I describe them as having survived genetic drift (establishment). The probability that a single beneficial mutant with small s will reach establishment threshold Nc is proportional to s on the first order,

$$P_{establish} = \frac{1-e^{-s}}{1-e^{-1}} \sim s \tag{26}$$

So most mutations with small fitness effects get lost in the establishment process. In large population/high beneficial mutation rates, more than one beneficial mutants may survive and be present in the population at the same time. They then compete for fixation.

Criterion for different regimes

From Eq. 26, we know that the time between the establishment of beneficial mutation s is $t_{est} = \frac{1}{NU_b s}$. The time from establishment to fixation, in the absence of other competing mutations, is $t_{fix} = \frac{\ln[Ns]}{s}$.

Thus, when $t_{fix} << t_{est}$, or $\ln[Ns] << \frac{1}{NU_b}$, the mutations fix much more rapidly than they get established in the population. This is the criterion for strong-selection, weak-mutation regime. In the scenario of large population or high beneficial mutation rate, the criterion breaks down and we enter the strong-selection, strong-mutation regime $\ln[Ns] > \frac{1}{NU_b}$. The extreme case that $\ln[Ns] >> \frac{1}{NU_b}$, named weak-selection, strong-mutation regime, is not discussed here.

APPENDIX C: SIMULATION

The exponential distribution in Figure 1

Following H.A.Orr's paper[1], I simulate a gene with sequence length $L = 1000$ and has 3000 one-step mutants. The wild type has fitness rank i among a total of 3001 genotypes. The fitness value is drawn randomly from a Gamma or Exponential distribution. ΔW is recorded as the fitness difference between the wild type and a randomly chosen beneficial mutant with a higher rank $j > i$. The distribution is plotted for data from 10,000 runs.

Moran process and the distributions in Figure 2

I start a population with all wild types and normalized fitness $r_0 = 1, N_0 = N$. The population undergoes a Moran process in each generation. In the presence of multiple mutant species with fitness r_i and population size N_i, the probability for an individual of the i_{th} species to be chosen for reproduction is proportional to its fitness, $\frac{r_i}{\sum_i r_i N_i}$; the probability for any individual to be chosen for death is equal, being $\frac{1}{N}$. If an individual is chosen for reproduction, it has a probability U_b to gain a fitness advantage s. The selection coefficient s follows an exponential distribution $p(s) = \frac{1}{s_0} e^{-\frac{s}{s_0}}$. The run restarts after each fixation event($N_i = N, i \neq 0$). In the simulation, I use parameters $N = 1000, U_b = 10^{-5} \sim 10^{-2}, s_0 = 0.01$. For $U_b = 10^{-5}, 10^{-4}$, I collected 500 samples. For $U_b = 10^{-3}, 10^{-2}$, the simulation becomes time-consuming as multiple mutants compete for fixation. Due to the limited sample size, the distributions shown in Figure. 2(c) and (d) are not very smooth. With more time the simulation could produce a histogram with better resolution.

* dailei@mit.edu

[1] H. Orr, Genetics **163**, 15191526 (2003).

[2] D. E. Rozen, J. G. de Visser, and P. J. Gerrish, Current Biology **12**, 1040 (2002).

[3] H. Orr, Nature Reviews Genetics **6**, 119 (2006).

[4] J. Gillespie, *The Causes of Molecular Evolution* (Oxford University Press, 1991).

[5] M. M. Desai, D. S. Fisher, and A. W. Murray, Current Biology **17**, 385 (2007).

[6] P. D. Sniegowski and P. J. Gerrish, Phil. Trans. R. Soc. B **365**, 1255 (2010).

[7] R. MacLean and A. Buckling, PLoS Genetics **5**, e1000406 (2009).

[8] S. Trindade and I. Gordo, PLoS Genetics **5**, e1000578 (2009).

[9] R. MacLean, A. R.Hall, G. G.Perron, and A. Buckling, Nature Reviews Genetics **11**, 405 (2010).

[10] A. J. Betancourt and J. P. Bollback, Current Opinion in Genetics and Development **16**, 618 (2006).