The SHAPE of logistic growth shows that timing does matter

Jonathan Dench[1,*],

1 Department of Biology, University of Ottawa, Ottawa, ON, Canada

¤Current Address: Department of Biology, University of Ottawa, Ottawa, ON, Canada
* jdenc017@gmail.com

Abstract

Experimental evolution is a powerful tool for studying evolutionary questions and the use of *in silico* systems is growing as they offer complete parameter control and more fine grained tracking of dynamics. However, existing software are limited by the models implemented, output obtainable, or their lack of interpretability within biological systems. Here I introduce SHAPE (Simulated Haploid Asexual Population Evolution) a forward time simulation tool closely replicating microbial experimental evolution that offers the flexibility to implement alternative models and study detailed output. I first validate SHAPE by replicating the evolutionary outcome expected by several theoretical works. I then extend theory by contrasting how serial passaging affects the evolutionary outcome for microbial communities undergoing exponential and logistic growth.

Author summary

The forward time simulation tool SHAPE (Simulated Haploid Asexual Population Evolution) was designed to compliment both theoretical and empirical study of evolutionary dynamics. As experimental evolution often studies the dynamics of *de novo* mutants, I validated SHAPE by comparing its ability to replicate theoretical expectations concerning the fixation of mutants. I have demonstrated that SHAPE closely replicates analytical approximations of the considered theoretical models. I then applied SHAPE to study how the laboratory protocol of serial passaging, common in microbial experimental evolution, affects the fixation probability of *de novo* mutants. This study extends related theory [1] by considering populations which grow logistically rather than exponentially. In contrast to established theory concerning exponential growth, under logistic growth the probability of a mutant arising and eventually fixing is not uniform throughout the growth phase.

Introduction

The study of evolutionary dynamics remains a challenge that should be addressed through the combined efforts of both evolutionary theory and empirical study [2]. Numerous theoretical works have developed models to describe the probability of fixation for new (*i.e. de novo*) mutants [1, 3–6], delineate parameter ranges between types of evolutionary response [7, 8], and to quantify unknown evolutionary parameters such as the fitness landscape [9–12]. However, models are necessarily simplifications of a system such as when their analytical assessment requires simplifying assumptions to

make the mathematics tractable. While we can study evolutionary response in more complex systems such as those of microbial experimental evolution, the power of these studies is limited both by our lack of control - or even precise knowledge of - underlying factors, and physical resources. It would benefit both theoretical and empirical work if we could gain an improved understanding of evolutionary response across a broad range of parameters and conditions. To achieve this we need tools to study more complex systems while controlling parameters and offering detailed tracking of dynamics.

Theoretical and empirical work can be complimented by simulation tools as they allow complex models to be implemented, provide control of parameters, and can be designed output highly detailed results. There exists tools to simulate evolution while focusing on either population dynamics [13] or specific genetic changes [14–18], though most of these implement a single theoretical framework and provide only specific output. The AVIDA [19] and AEVOL [20] tools are remarkable simulation platforms for studying genetic changes but while their genomic abstraction is practical for implementation, it makes it unclear to what extend their results apply to living systems. A powerful tool for simulating microbial experimental evolution is the Haploid Evolutionary Constructor (HEC) [21], though it simplifies evolution of the genome by mapping each loci to a single quantitative trait value and assumes mutations have no pleiotropic effects. There exists no single tool which simulates experimental evolution while tracking both population demographics and the genetics of large genomes, and provides a single framework in which to study various models and parameter ranges. Without a single tool it can be hard to study different models and parameter ranges as the output of different tools may not be comparable. To address this I developed SHAPE (Simulated Haploid Asexual Population Evolution).

The simulation tool SHAPE was designed to simulate microbial experimental evolution and offers a flexible implementation which can handle various models. I modelled SHAPE after microbial evolution experiments as they are a common type of living system used for the manipulative study of evolutionary dynamics. Microbial evolution experiments are often used to study how factors influence the emergence and persistence of *de novo* mutants arising from an initially isogenic population. The design of microbial evolution experiments often includes the practice of serial passaging by which a fraction of stationary phase communities are transferred to fresh growth media to increase the number of generations in the experiment. Theory suggests that the probability of a mutant arising during exponential growth, and subsequently surviving many rounds of serial transfer, is roughly uniform [1]. However, it remains unclear how the laboratory practice of serial passaging affects the dynamics of *de novo* mutants when growth is logistic.

In this paper I begin by explaining the framework of SHAPE. I then report on my validation of SHAPE achieved through its comparison to the theoretically expected fixation probability of a *de novo* mutant under various conditions. I then apply SHAPE to extend theory on how serial passaging affects the probability of a *de novo* mutant, growing logistically, eventually fixing.

Design and Implementation

The SHAPE tool is a forward time, discrete step, simulator of population demographics and genotype evolution. SHAPE records all population demographic and genotype fitness information using SQL databases. The framework of SHAPE has a modular design to simplify the addition of new models or functionality. In its current form, populations simulated with SHAPE are finite and evolve in a single environment which is homogenous throughout time and space. Further, genotypes are tracked in binary states where the coding of an unmutated site is 0 and 1 is used for sites which have

mutated. A simplified flow-chart of SHAPE is presented in Fig. 1.

SHAPE tracks the number of individuals for each of the n unique living genotypes which together are referred to as the community. Runs of SHAPE record the step-wise changes in community demographics (*i.e.* deaths, births, and mutation events - DBMs) at discrete time steps which for convenience I refer to as generations though depending on the chosen parameters this term may not fit a conventional meaning. Once a genotype i (where $i \epsilon \{1, 2, ..., n\}$) exists within the community, SHAPE will track the number of individuals (N_i), their fitness (w_i), and their DBMs. The number of DBMs are scaled by their respective per generation probability variables P_d, P_b, and P_m respectively. By changing the birth and death probabilities, population dynamics may simulate discrete or overlapping generations. Growth is calculated with a function that accepts numerous arguments including at least the current community size N_i, the probabilities P_b and P_d, and the model used to calculate growth (g_{mod} - *i.e.* growth model). When growth of the community is bounded, and the maximum carrying capacity K is reached, evolution will only continue provided individuals within the community die because $P_d > 0$ or the number of individuals is reduced by a disturbance event. Disturbance events instantly reduce the number of individuals N_i for each genotype i in a stochastic, but proportional, manner. These events can be parameterised to replicate situations such as stochastic catastrophic events, chemostat growth conditions, or the common laboratory protocol of serial passaging. Evolution proceeds as new genotypes arise through mutation and effects of selection are expressed via the relative fitness values w_i calculated for each genotype. Once a mutant genotype is encountered in SHAPE, its fitness value is recorded in a reference database to ensure constancy throughout the simulation and so that particular genotype fitness mappings can be used in later simulations.

The SHAPE tool was written for, and implemented in, R [22]. Nearly all functions used in SHAPE were either part of the base R package or self written. The exceptions being the SQL interface provided by *DBI* [23] and *RSQLite* [24], the probability generating functions provided by *evd* [25], *VGAM* [26] and *sn* [27], the parallelisation functions of *foreach* [?], *doMC* [?] and *doSNOW* [?], the data concatenation function of *abind* [?], and the plotting utilities of *lattice* [?], *plot3D* [?], *colorRamps* [?], and *RColorBrewer* [?].

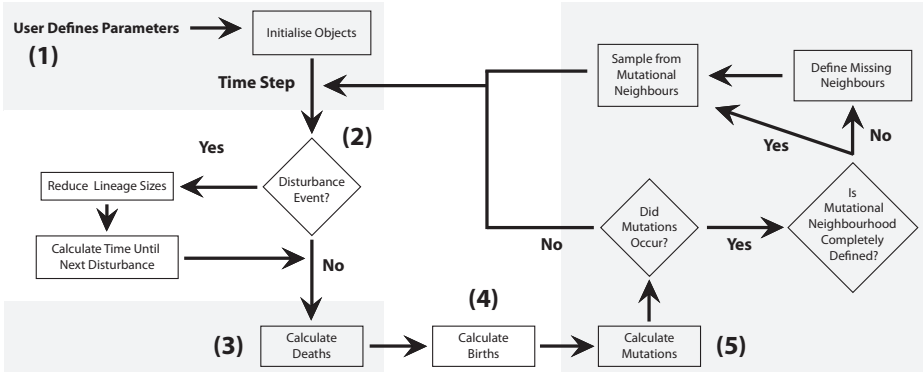

Fig 1. SHAPE tool workflow chart. The alternating shaded boxes highlight the sequential functions called in each generation. The calculation of disturbance events, deaths, births and mutations are performed for each of the genotypes with at least one living member prior to the function call. Deaths and births are treated as occurring simultaneously such that the same vector N_i is used in both calculations. The explored genotype fitness mapping is stored in an SQL database ensuring that genotypes are uniquely defined. At the end of each time step, SHAPE records the current state of the community and changes which occurred.

Validation

Like many microbial evolution experiments, a run of SHAPE starts with a population of wild-type (WT) individuals from which many *de novo* mutants will arise and some proportion will fix. To validate my software, I parameterised SHAPE using the conditions outlined by theoretical works and estimated the probability of fixation for *de novo* mutants. The theoretical works were selected as they studied the probability of fixation under the different models for growth implemented in SHAPE, as well as effect of serial passaging.

Constant number of individuals

The seminal theoretical work of Haldane [3] approximates the probability of fixation for a single mutant as being $\approx 2s$ (where s is the selection coefficient). Haldane's work makes the assumption that the WT population is very large, that the number of individuals is constant, that s is small ($s \ll 1$), that generations do not overlap, and that mutation only generates a single mutant. A large population is relative so I simulated a constant number of 10^3 and 10^7 individuals but as the findings are similar I only discuss the results for the larger number in the main text (see SI Fig. 6 for both) . SHAPE implements two constant number of individual growth models, *Poisson* or *Constant*. The first is called the *Poisson* form and is identical to the method used by Haldane [3] and is based on a Galton-Watson branching process [28]. This model of growth assumes non-overlapping generations (*i.e.* $P_d = 1$) during which an average individual gives birth to one offspring. The exact number of offspring is stochastic and is drawn from a Poisson distribution (hence the name) with location parameter λ given by the product of P_b and an individual's relative fitness $w_i = 1 + s$ ($\lambda = w_i\, P_b$). The *Constant* form is similar to *Poisson* except it scales the number of births to be exactly the same as the number of deaths. Using a range of selective coefficients, $s\epsilon\{0.001, 0.005, 0.01, 0.03, ..., 0.09, 0.1, 0.15, ..., 0.3\}$, I calculated the fixation probability from 1,000,000 replicate runs of SHAPE. The comparison of fixation probabilities calculated with eq. 1 (from eq. 1 of Haldane's theoretical work [3]) and SHAPE are presented in Fig. 2.

$$s = \sum_{i=2}^{\infty} \frac{prob_{fix}^{i-1}}{i} \tag{1}$$

where $prob_{fix}$ is the probability of fixation. The classic $prob_{fix} \approx 2s$ comes from considering only the first term of eq. 1 but a more accurate value for $prob_{fix}$ given s can be estimated in a two step process. The first step simply rearranges the un-simplified form of eq. 1 of Haldane's work to give

$$s = \frac{-ln(1 - prob_{fix})}{prob_{fix}} - 1 \tag{2}$$

then using $prob_{fix} \approx 2s$ as a starting point and testing nearby values ($prob_{fix} \pm \delta x$), the exact $prob_{fix}$ which solves for s can be found. When comparing SHAPE to Haldane's work I used this two step method to calculate the true fixation probability and use $2s$ as the analytical approximation (Fig. 2 A). The fixation probability estimated with SHAPE matches the true fixation probability while the $2s$ analytical approximation noticeably over estimates the fixation probability when $s \geq 0.05$ (Fig. 2 B). The two constant growth models provide nearly identical results except when the selection coefficient is quite large ($s \geq 0.25$) at which point the *Constant* form underestimates the fixation probability by $\sim 1\%$. When comparing the exact fixation probability to the analytical approximation $2s$ (Fig. 2 B), the two differ by about 6%

when $s = 0.05$. In fact only when $s < 0.01$ will the two differ by less than 1.3% which suggests that s be considered small, for theory derived similarly to Haldane's, only when $s \leq 0.01$. Regardless, I have shown that for a mutant arising in an environment with a constant number of individuals, SHAPE can accurately estimate the exact fixation probability but SHAPE underestimates the analytical approximation of $2s$.

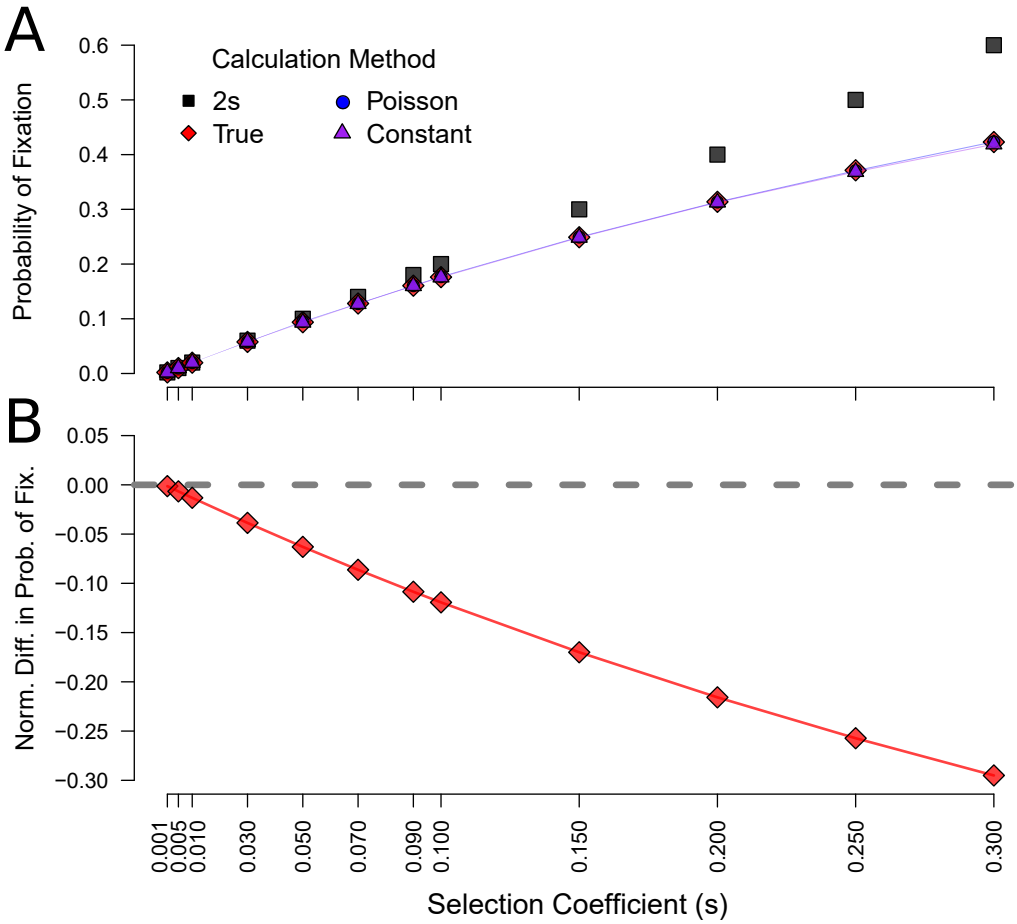

Fig 2. Comparison between the probability of a single mutant fixing, in an environment with a constant number of individuals, as calculated with SHAPE or using a theoretical model [3]. Panel A shows the fixation probability, dependent on the selection coefficient s, where coloured circles represent the values calculated using different constant growth models (*Poisson* - blue; *Constant* - purple) implemented in SHAPE, red diamonds represent the true fixation probability and black squares represent the analytical approximation of $2s$. In B, I present the normalised difference between the true fixation probability and the analytical approximation of $2s$. Negative values reflect that the true fixation probability is lower than the analytical approximation of $2s$.

Logistic growth

I next used SHAPE to estimate the fixation probability of a single mutant when growth is, like bacteria, logistic. The related theoretical works [4, 5] suggest that fixation probability is $2s$ if the mutant appears during an infinite stationary phase but is greater if it appears during growth. The theory was first worked out by Ewens [4] but later put into analytical form by Otto and Whitlock [5]. The assumptions are similar to Haldane's work except the analytical approximations assume that growth rate is small. In the model, the fixation probability is dependent on the total number of individuals

($\sum_{i=1}^{n} N_i$), relative to the carrying capacity K, that exist when the mutant is generated. Simulations began with a number of individuals equal to one hundredth the carrying capacity ($\frac{K}{100}$). Logistic growth for each genotype is calculated given the number of individuals N_i, their fitness $w_i = 1 + s$, the total number of individuals $\sum_{i=1}^{n} N_i$, the carrying capacity K and the intrinsic growth rate r. I simulated growth for small and large carrying capacities $K\epsilon 10^{(4,8)}$ but discuss only the results for when $K = 10^8$ as the results are qualitatively similar (see SI Fig. 7 for $K = 10^4$). The range of s and r are identical to the range of s used in my comparison to Haldane's work and cover the range of positive values used in the work of Otto and Whitlock [5]. To calculate the theoretical fixation probability, I used the analytical approximation shown in eq. 3 (eq. 11 from Otto and Whitlock [5])

$$prob_{fix}(t) \approx \frac{2sK(s+r)}{sK + r\sum_{i=1}^{n} N_i(t)} \tag{3}$$

where $prob_{fix}(t)$ is the fixation probability given the mutant arose at time t when there were $\sum_{i=1}^{n} N_i$ individuals in the environment with carrying capacity K. Estimates with SHAPE were calculated from 1,000,000 replicates. A comparison between SHAPE and the analytical approximation is shown in Fig. 3.

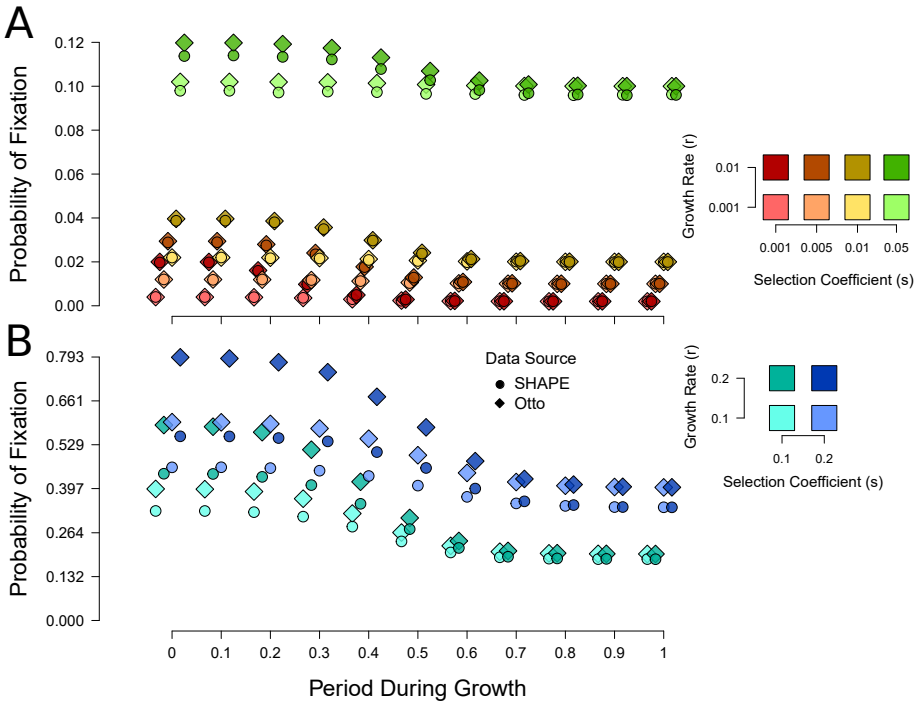

Fig 3. Comparison of the theoretical and estimated fixation probability for a mutant growing logistically. Diamond shapes represent the analytical approximation [5] derived similarly to Haldane's $\approx 2s$ while circles are for estimates using SHAPE. The colour used to fill points reflects the selection coefficient s and darker colours represent higher intrinsic growth rates r. The period during the growth phase is scaled from the start of growth until the point where the number of individuals reaches carrying capacity. Panel A represent the range of parameters originally studied in the theoretical work whereas B is for large parameter values.

The fixation probability estimated by SHAPE is quite similar to the analytical approximation when both s and r are within the range of values presented in the work of Otto and Whitlock [5] (Fig. 3 A). When s is small (*i.e.* $s \leq 0.01$) SHAPE replicates the analytical approximation. When either are large (Fig. 3 B), SHAPE underestimates

the analytical approximation by an amount that decreases until K is reached. Once at carrying capacity, the difference between SHAPE and the analytical approximation equals the difference between the true fixation probability and the analytical approximation of $2s$ calculated in the previous section.

Serial Passaging

The laboratory practice of serial passaging, common to microbial experimental evolution, involves transferring a small volume of liquid media, containing individuals at stationary phase, into new media to allow continued growth. The ratio of new to old volume is commonly expressed as the dilution factor D. As long as the sample was perfectly mixed prior to transfer, the inoculum will proportionally represent the number of each genotype present prior to transfer. These repeated bottlenecks result in rare genotypes being lost [1] since on average any genotype with less than D individuals is not expected to survive a transfer. To validate the population disturbance function used to simulate serial passing in SHAPE, I measured the probability that at least one focal genotype individual survived an event. I compared the probability estimated with SHAPE to the theoretical expectation of a binomial process [1] (Fig. 4). I present results for two representative cases in the main text, for all results see SI Fig. 8. The probability was measured with 10,000 replicate disturbance events for each of a range of the number of individuals prior to transfer ($\sum_1^n N_i = 10^{2,3,...,9}$) and number of focal genotype individuals ($N_{focal} \epsilon \{0, 5, 10, ..., min(\sum N_i, 750)\}$), and for the commonly used dilution factor $D = 100$.

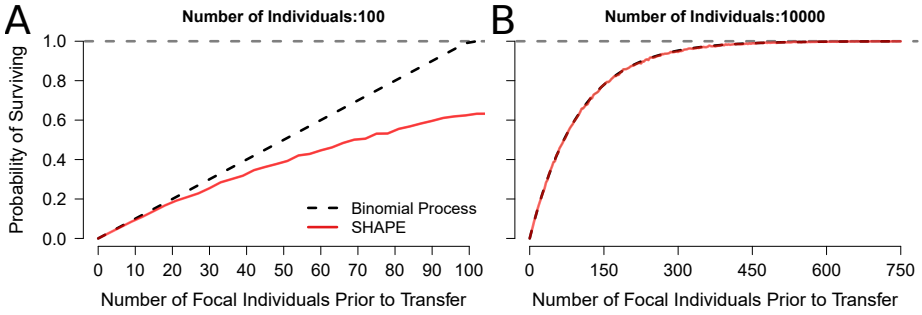

Fig 4. The probability that at least one individual of a focal genotype survives a disturbance event (serial passaging) reducing the total number of individuals 100 fold ($D = 100$). The solid red line shows the 95% CI of survival probability calculated with SHAPE, while the black dashed line shows the deterministic expectation calculated from a binomial process. The dashed grey line highlights when probability reaches unity. Panels A and B differ in the total number of individuals and number of focal individuals prior to transfer.

The smallest total number of individuals, where $D = \sum_{i=1}^n N_i$, represents the case where estimates with SHAPE differ from the deterministic expectation of a binomial process (Fig. 4 A). When the focal genotype is fixed within the community and the dilution factor equals the total number of individuals (*i.e.* $N_{focal} = D = \sum_1^n N_i$) the probability of transfer reaches unity with a binomial process but not with estimates using SHAPE. This is because SHAPE is a stochastic process, whereas calculations using a binomial process are deterministic. Consider a microbial evolution experiment where a bacterial culture has reached stationary phase at 100 individuals in 1mL. To continue the experiment, serial passaging with $D = 100$ is used whereby 0.01 mL of old media is transferred to 0.99 mL of new media. If mixing is perfect and all bacteria are equally distributed throughout the media, then we expect exactly one individual to be transfered. However, there is always some error with sampling so that even from a pure

culture it is always possible to sample no individuals. Regardless, in microbial experimental evolution and likely for most simulations run with SHAPE, the carrying capacity is not expected to be of the same order of magnitude as D in which case SHAPE samples similarly to the expectation of a binomial process (Fig. 4 B).

Extension to theory

Theory suggests that the probability of a *de novo* mutant arising during the exponential growth of a community, and subsequently surviving many serial transfers, is roughly uniform [1]. So while rare genotypes may be lost due to serial passaging, the protocol should not bias evolutionary outcome unless the mutations affect particular growth parameters reviewed by another study [6]. The theoretical work considered here assumes effectively exponential growth between serial passaging events, that there is no competition, and that deaths can be ignored. Using the parameters of the exponential and nutrient limited growth models from Wahl *et al.* [1] to parameterise exponential and logistic growth with SHAPE, I estimated the probability of long term survival for a single mutant arising in a community undergoing regular serial passaging. I compare the results for exponential growth to the theoretical expectation, whereas the results of logistic growth provide an extension to theory. Simulations were run for up to 1,000,000 growth phases of τ generations which were reset after a serial passaging event of dilution factor D. In the first round of growth, a single mutant ($s = 0.1$) could arise during any generation $t\epsilon\{1, .., \tau\}$ and the probability of survival was counted as the proportion of 1,000,000 replicates in which the mutant was not lost during transfer. As per the assumptions of the theoretical work, deaths were ignored ($P_d = 0$) so serial transfer was the only means by which the mutant could be lost. Other parameters for the exponential (logistic) growth models were as follow: dilution factor was set so that each growth phase started with a total number of $10^5(10^7)$ individuals, growth rate was $r = ln(2)$ ($r = 2$), mutation rate $\mu = 5 \times 10^{-5}(4 \times 10^{-9})$, and the length of each growth phase was $\tau = 7$ ($\tau = 6$) generations. To compare the estimate of SHAPE with theoretical expectation, I used the analytical approximation of survival which was calculated as $1 - P_{extinction}$, where the extinction probability ($P_{extinction}$) is given by eq. 4 [1]

$$P_{extinction} \approx 1 - 2se^{-rt}r\tau \tag{4}$$

where τ is the number of generations that occur during each growth phase and t is the generation during which the mutant arises. The probability being considered is the joint probability that the mutant arose at time t and subsequently survives. The probability of a mutant arising during generation t was calculated as the product of the number of births and the value calculated with an exponential probability density function with rate $\alpha = 100$ (personal communications with Dr. Wahl) for the mutant selection coefficient $s = 0.1$. So, the analytical approximation of the joint probability that a mutant arose and surived was calculated given eq. 5 [1]

$$P_{survival} \approx births(t)\, \mu\alpha e^{-\alpha s}(1 - P_{extinction}) \tag{5}$$

where the number of births during generation t is the difference in number of individuals between generations t and $t - 1$. The estimated joint probabilities of a mutant arising and surviving calculated with SHAPE, as well as comparison to the analytical approximation, are shown in Fig. 5. When growth is exponential, SHAPE estimates a roughly uniform joint probability throughout the growth phase though the exact value is lower than the analyitical approximation (Fig. 5 A). By looking at the independent probabilities of arrival and fixation (Fig. 5 C-D), we see this difference is due to SHAPE estimating a lower probability of fixation. This difference is not

surprising as the selection coefficient studied ($s = 0.1$) is relatively large, the analytical approximation of eq. 4 derives from Haldane's $2s$, and I've previously shown that under constant growth there is an 11.93% difference between the $2s$ approximation and true fixation probability.

For logistic growth (Fig. 5 B), SHAPE estimated a joint probability that is not uniform and is two orders of magnitude lower than for exponential growth. The joint probability is roughly uniform for the first half of the growth phase but declines sharply thereafter. These two halves are separated by the point of inflection in the logistic growth curve. To understand why the joint probability is two orders of magnitude lower for logistic growth, I plotted the independent probabilities of arrival and fixation (Fig. 5 C-D) and found that while both are lower for logistic growth, the probability of arrival is two orders of magnitude lower. The difference in probability of arrival reflects the that the parameters chosen for logistic growth would lead to a mutation supply rate roughly two orders of magnitude lower. The difference in probability of fixation reflects that during any given growth phase the mutant lineage will produce less individuals growing logistically compared to exponential (Fig. 5 E). This result suggests, similar to other theoretical work [6], that the protocol of serial passaging can affect which mutations are fixed during an experiment. Because microbial populations used in experimental evolution tend to grow logistically, the practice of serial passaging will bias the fixation of mutations arising early during growth. This biased fixation probability will increase stochasticity in response between replicate populations and reduce the confidence in estimates of selection coefficients based on either the repeatability of fixation or time to fixation estimated by sampling at the end of growth phases. If the time to fixation is long, due to the selective advantage of any mutation being small and the number of individuals large, then the bias of serial passaging could be ignored since we would expect maximal clonal interference dynamics whereby all possible mutants compete for fixation and the probability of fixation is driven by selection coefficient. However, previous work suggests that the dynamics of microbial experimental evolution is more likely to reflect intermediate clonal interference [29] whereby some, but not all, possible mutants compete for fixation and so the order (*i.e.* timing) in which they appear matters. Other studies have shown how the order in which mutations appear can affect evolutionary outcome [30–32].

Availability and future directions

All the scripts required to run SHAPE, as well as the scripts used to compare SHAPE with analytical theory, are available from GitHub github.com/JDench/. While the scripts are highly commented to help a more advanced user make changes, I provide a Readme file to describe the basics of how to run SHAPE. This software is made freely available under the GNU General Public License v3.0.

Future developmental priorities for SHAPE are threefold. First it would expand the applicability of SHAPE if I included functions for different habitat patches, partitioning of the genome, and temporal heterogeneity to selection. Second I would like to add additional fitness landscape models such Fisher's Geometric Model [33] and an eggbox model. Lastly, I want to translate SHAPE from the language of R to C in order to drastically improve runtime through compiling and use of POSIX thread libraries. This last would benefit from collaboration with someone who has a strong background in C programming.

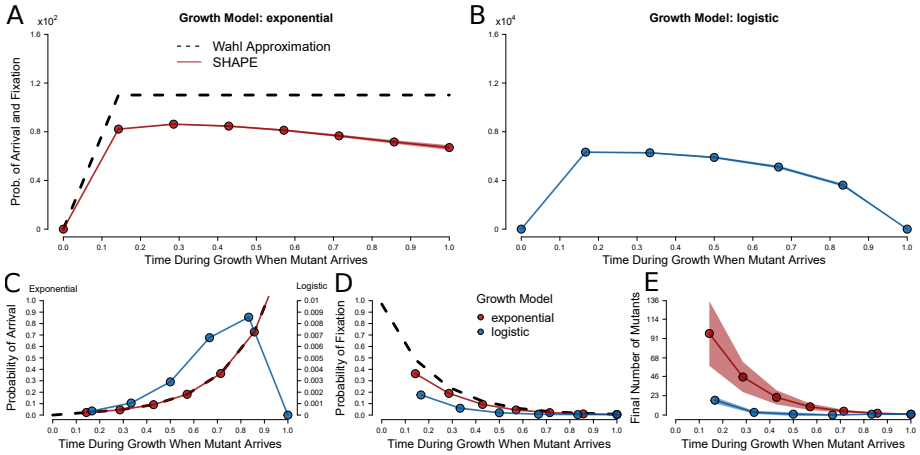

Fig 5. The joint probability of a single mutant arising during growth and then surviving many rounds of community bottlenecks. The parameters used are similar to those of Wahl *et al.* [1] such that the mutation has a selection coefficient of 0.1, bottleneck strength was set to 1:100, and other parameters varied based on the model of growth. Panel A compares the exponential growth analytical approximations of Wahl *et al.* [1] (dashed line presents the discrete form, solid line shows continuous time) and the probability calculated with SHAPE. Panel B shows the probability calculated from simulations with SHAPE when growth was logistic and parameters were similar to the nutrient growth conditions of Wahl *et al.* [1].

Conclusions

Simulations allow models to be implemented without the simplifying assumptions required by analytical approaches. I designed SHAPE to provide a unified framework for testing biological questions that compare different models and assumptions. Herein, I have shown that SHAPE is able to replicate evolutionary dynamics consistent with evolutionary theory. By expanding on existing theory concerning the joint probability of a mutant arising during growth and surviving the common laboratory practice of serial passaging, I provide evidence that this practice can affect the interpretation of microbial experimental evolution. My results echo the warnings of previous theoretical works which suggest that serial passaging will increase stochasticity [1] as well as bias phenotypic selection [6]. I suggest that the sensitivity of parameter estimates, based on the results of microbial experimental evolution, could be improved if we could quantify the proportion of variance attributable to the practice of serial passaging. SHAPE is one tool which could address just such a question.

Acknowledgments

I would like to thank the Center for Advanced computing for access to their computational resources. I would also like to thank the National Science and Energy Research Council of Canada for funding support accessed through the grants of my supervisors Drs. St'ephane Aris-Brosou and Rees Kassen.

References

1. Wahl LM, Gerrish PJ, Saika-Voivod I. Evaluating the impact of population bottlenecks in experimental evolution. Genetics. 2002;162(2):961–971.

2. de Visser JAGM, Krug J. Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet. 2014;15(7):480–90. doi:10.1038/nrg3744.

3. Haldane JBS. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. Proceedings of the Cambridge Philosophical Society. 1927;23:838. doi:10.1017/S0305004100015644.

4. Ewens W. The probability of survival of a new mutant in a fluctuating environment. Heredity. 1967;22(3):438–443.

5. Otto SP, Whitlock MC. The Probability of Fixation in Populations of Changing Size. Genetics. 1997;146(2):723–733.

6. Wahl LM, Zhu AD. Survival Probability of Beneficial Mutations in Bacterial Batch Culture. Genetics. 2015;200(1):309–320. doi:10.1534/genetics.114.172890.

7. Desai MM, Fisher DS. Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics. 2007;176(3):1759–98. doi:10.1534/genetics.106.067678.

8. Weissman DB, Desai MM, Fisher DS, Feldman MW. The rate at which asexual populations cross fitness valleys. Theor Popul Biol. 2009;75(4):286–300. doi:10.1016/j.tpb.2009.02.006.

9. Kauffman S, Levin S. Towards a general theory of adaptive walks on rugged landscapes. Journal of Theoretical Biology. 1987;128(1):11 – 45. doi:http://dx.doi.org/10.1016/S0022-5193(87)80029-2.

10. Kauffman SA, Weinberger ED. The NK model of rugged fitness landscapes and its application to maturation of the immune response. Journal of Theoretical Biology. 1989;141(2):211 – 245. doi:http://dx.doi.org/10.1016/S0022-5193(89)80019-0.

11. Aita T, Uchiyama H, Inaoka T, Nakajima M, Kokubo T, Husimi Y. Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: Application to prolyl endopeptidase and thermolysin. Biopolymers. 2000;54(1):64–79. doi:10.1002/(SICI)1097-0282(200007)54:1¡64::AID-BIP70¿3.0.CO;2-R.

12. Neidhart J, Szendro IG, Krug J. Adaptation in tunably rugged fitness landscapes: the rough Mount Fuji model. Genetics. 2014;198(2):699–721. doi:10.1534/genetics.114.167668.

13. Guillaume F, Rougemont J. Nemo: an evolutionary and population genetics programming framework. Bioinformatics. 2006;22(20):2556–2557. doi:10.1093/bioinformatics/btl415.

14. Lambert BW, Terwilliger JD, Weiss KM. ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth. Bioinformatics. 2008;24(16):1821–1822. doi:10.1093/bioinformatics/btn317.

15. Dalquen DA, Anisimova M, Gonnet GH, Dessimoz C. ALF—A Simulation Framework for Genome Evolution. Molecular Biology and Evolution. 2012;29(4):1115–1123. doi:10.1093/molbev/msr268.

16. Zanini F, Neher RA. FFPopSim: an efficient forward simulation package for the evolution of large populations. Bioinformatics. 2012;28(24):3332–3333. doi:10.1093/bioinformatics/bts633.

17. Matthey-Doret R, Whitlock M. SimBit - Simulating Individual Markers with Bits; 2017. Available from: https://github.com/RemiMattheyDoret/SimBit [cited 08-02-2018].

18. Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M, et al. Weak Epistasis May Drive Adaptation in Recombining Bacteria. Genetics. 2018;doi:10.1534/genetics.117.300662.

19. Ofria C, Bryson DM, Wagner AP. Avida Digital Evolution Platform; 2017.

20. Beslon G, Knibbe C, Parsons D, Batut B, Misevic D, Frénoy A, et al.. AEVOL: An *in silico* experimental evolution platform; 2018. Available from: http://www.aevol.fr/ [cited 01-02-2018].

21. Lashin S, Matushkin YG, Klimenko A, Suslov VV, Kolchanov NA. Evolution, Biodiversity and Ecology in Microbial Communities: Mathematical Modeling and Simulation with the "Haploid Evolutionary Constructor"Software Tool. IntechOpen; 2014.

22. R Core Team. R: A Language and Environment for Statistical Computing; 2016. Available from: https://www.R-project.org/.

23. R Special Interest Group on Databases (R-SIG-DB), Wickham H, Müller K. DBI: R Database Interface; 2016. Available from: https://CRAN.R-project.org/package=DBI.

24. Müller K, Wickham H, James DA, Falcon S. RSQLite: 'SQLite' Interface for R; 2017. Available from: https://CRAN.R-project.org/package=RSQLite.

25. Stephenson AG. evd: Extreme Value Distributions. R News. 2002;2(2):0.

26. Yee TW, Wild CJ. Vector Generalized Additive Models. Journal of Royal Statistical Society, Series B. 1996;58(3):481–493.

27. Azzalini A. The R package sn: The Skew-Normal and Skew-t distributions (version 1.4-0); 2016. Available from: http://azzalini.stat.unipd.it/SN.

28. Kimmel M, Axelrod DE. In: The Galton-Watson Process. New York, NY: Springer New York; 2002. p. 33–63. Available from: https://doi.org/10.1007/0-387-21639-1_3.

29. Bailey SF, Blanquart F, Bataillon T, Kassen R. What drives parallel evolution? BioEssays. 2016;39(1):e201600176. doi:10.1002/bies.201600176.

30. Weinreich DM, Chao L. Rapid Evolutionary Escape by Large Populations from Local Fitness Peaks Is Likely in Nature. Evolution. 2005;59(6):1175–1182.

31. Kvitek DJ, Sherlock G. Reciprocal Sign Epistasis between Frequently Experimentally Evolved Adaptive Mutations Causes a Rugged Fitness Landscape. PLoS Genet. 2011;7(4):e1002056. doi:10.1371/journal.pgen.1002056.

32. Sackman AM, Rokyta DR. Additive Phenotypes Underlie Epistasis of Fitness Effects. Genetics. 2017;doi:10.1534/genetics.117.300451.

33. Fisher RA. The Genetical Theory of Natural Selection. 2nd ed. New York: Dover Publications; 1958.

Supplementary Information

Detailed workflow of SHAPE

The sections of this detailed workflow follow the same numbering as used in the SHAPE flow chart in the main text (Fig. 1).

1. General Parameters

The parameters described here are not an exhaustive list but instead those most likely to be of general interest for a SHAPE user. A key parameter for any evolution experiment is its length which SHAPE controls through the number of generations T to be simulated. In each generation, SHAPE sequentially calls the population disturbance, death, birth, and mutation (DBM) functions. Each of the DBM functions has an associated probability parameter ($P_{d,b,m}$ respectively) which controls the per generation probability of any individual having an associated event. The current growth models implemented in SHAPE can allow exponential or logistic growth as well as a constant number of individuals. Calculations of growth are performed with the growth function which is controlled by the growth model. The user also defines a focal number of individuals N_f that has different interpretations given the growth model. For the two constant number of individual growth models in SHAPE, named *Poisson* and *Constant*, the focal number of individuals defines the starting number of individuals. The *Poisson* model comes from the Galton-Watson branching process [28] and assumes $P_d = P_b$ and that the intrinsic growth rate r is two. This model will result in the population having a roughly constant size over many replicates, but in any one replicate this number is likely to deviate by an increasing amount proportional to T. I developed the *Constant* model to relax these assumptions and enforce a strictly constant number of individuals. The *Constant* model calculates the per genotype proportional number of births and multiplies these proportions to the number of deaths. It a constant growth model is selected the population disturbance function will never be called. For exponential growth, N_f is the starting number of individuals as well as the target number of individuals for the population disturbance function. Under logistic growth, N_f represents the environmental carrying capacity (*i.e.* K).

Throughout a run of SHAPE, mutants arise with probability P_m which is multiplied with the number of births or as one theoretical paper suggests to the total number of individuals [7]. The genomes of individuals have a constant length L of binary state sites where 0 is the wild-type (WT) state and 1 the mutant. The user may choose if revertant mutations are permitted (*i.e.* if 1 can mutate to 0). The fitness w_i of a genotype i is calculated given the fitness landscape model and parameterised random effect distribution. The fitness landscape models currently implemented are: Additive, Fixed, House of Cards (HoC) [9], Kauffman's NK (NK) [10], or Rough Mount Fuji (RMF) [11, 12]. The additive model initially calculates the effect (*i.e.* selective coefficient) of mutations from the random effect distribution and calculates w_i as the sum of the mutation effects. The fixed model is only practical for small genotypes as it requires that the user supply a matrix defining the fitness value for each genotype. The fixed model is included to allow users to test questions using a specific fitness landscape. A full description of the HoC, NK and RMF models are beyond the scope of this work but in brief each calculates w_i using the values drawn from the random effect distribution, and some require additional constants that can be defined in SHAPE. At present, SHAPE implements the following probability generating functions for the random effect distribution: Beta, χ^2, Exponential, Frétchet, Gamma, General Extreme Value Distribution [25], Normal, Reverse Weibull, Skew Normal [27] and Uniform.

Once initial parameters are defined and a run starts, each generation will begin with

a call to the population disturbance function.

2. Population Disturbance

This function is used to simulate disturbance events that stochastically reduce the total number of individuals. Disturbance events occur based on a schedule defined by setting a fixed number of generations between events or so that the expected number of births between events restore the number of individuals lost. The expected number of individuals lost for each genotype i is proportional to their frequency. Using the disturbance factor D, the new expected number of individuals of each genotype i is calculated as $\frac{N_i}{D}$. This process is stochastic because the expectation is used as a location parameter for a draw from the Poisson distribution which is returned as the remaining number of individuals. For simulations using exponential growth, the value D is controlled by N_f such that:

$$D = \sum_{i=1}^{n} N_i / N_f \tag{6}$$

where N_i is the number of individuals of genotype i. This choice prevents the total number of individuals from growing indefinitely since they are reset to the focal number after an event. As constant growth models do not allow population disturbance, the user can only define D when growth is logistic. Users can set D to be a constant value or drawn from a parameterised normal distribution. Any value $D < 1$ will automatically be set to 1.

Next SHAPE will calculate the number of deaths.

3. Death Events

The probability of death P_d controls the proportion of individuals that die in a generation. The user can choose if this probability is a constant value or if it is density dependent. When P_d is constant, the number of deaths for a genotype i is given by:

$$deaths_i = N_i \, P_d \tag{7}$$

where N_i is the number of individuals of genotype i. If deaths are density dependent the number is calculated by:

$$deaths_i = N_i \, P_d \, (\frac{\sum_{i=1}^{n} N_i}{K_d})_d^c \tag{8}$$

where K_d is the population size at which P_d is 100% its defined value. The exponent c_d is used to scale how the ratio of community size and K_d affects P_d where larger c_d lead to the effect of P_d being very small until the population is close to K_d. Please note that K_d is defined separately from the logistic growth carrying capacity K.

This function is called prior to births but does not directly affect the N_i used when determining births. The number of deaths is first calculated so that births can be scaled to deaths such as under conditions of the *Constant* growth model.

4. Birth Events

The per generation probability of any individual giving birth is controlled by P_b. The number of offspring generated by an individual is calculated given the growth model, the intrinsic growth rate r, and the genotype's fitness w_i. For the *Constant* growth model, the number of births for each genotype is proportional to their size N_i and is calculated in two steps. The first is to calculate the birth potential of each genotype by

$$births_{potential_i} = N_i\,(1 + w_i - \bar{w})\,P_b \tag{9}$$

where the term $(1 + w_i - \bar{w})$ calculates relative fitness centred around 1 using the mean fitness $\bar{w}$. This method of calculating relative fitness accounts for when $\bar{w} = 0$ but is sensitive to the magnitude of $\bar{w}$. The user can choose relative fitness to be calculated using the more traditional $\frac{w_i}{\bar{w}}$, but this will be overridden if $\bar{w} = 0$. The second step in calculating the enforced constant growth uses the potential births $births_{potential_i}$ calculated in eq. 9 as weightings in the final calculation:

$$births_i = \frac{births_{potential_i}}{\sum_{i=1}^{n} births_{potential_i}} \sum_{i=1}^{n} deaths_i \tag{10}$$

where the number of births is scaled to the sum of deaths and each genotype contributes proportional to their size and fitness. If no potential births occurred then the second step is skipped and $births_i$ is returned as a vector of zeroes. If the *Poisson* growth model was chosen, then $births_i$ is obtained through draws from a Poisson distribution where the location parameter is given by $N_i\,w_i\,P_b$. This approach was derived from theoretical work (*e.g.* Haldane [3]) and assumes a large population and $P_b = P_d$. When growth is either exponential or logistic growth, then the intrinsic growth rate r is used to calculate the number of offspring from a birth event. As an example, to simulate a bacterium undergoing binary fission you might set $r = 2$. For exponential growth the expected number of births for each genotype are calculated as

$$births_i = N_i\,(e^{ln(r)\,w_i\,P_b} - 1) \tag{11}$$

where $ln(r)$ is the natural logarithm of the intrinsic growth rate and w_i is the fitness of a genotype. This equation is derived from the exponential growth model but I subtract 1 from the growth term so that births are calculated. Note that if $e^{ln(r)\,w_i\,P_b} l 1$ then it is set as one. When growth is logistic, the expected number of births for each genotype is calculated in two steps. First, a density dependent growth term dg_i for each genotype i is calculated using the logistic equation

$$dg_i = N_i + (w_i\,r\,P_b)\frac{K - \sum_i^n N_i}{K} \tag{12}$$

where K represents the carrying capacity. This density dependent term dg_i represents the amount of growth expected for genotype i given the current total number of individuals. Using dg_i, the number of births for each genotype is calculated as

$$births_i = N_i\,(\frac{dg_i}{\sum_i^n N_i} - 1) \tag{13}$$

where similar as to with exponential growth, I subtract 1 to calculate births. Recall that while deaths are calculated prior to births they do not affect N_i used in these calculations. Users may set the number of births calculated to be scaled by deaths which will cause the vector $births_i$ to be added to the vector output from a nested call of the growth function with the *Constant* growth model. All of these calculations are deterministic and so to make growth calculated by SHAPE a stochastic process, users can toggle that the births are recalculated by passing the vector $births_i$ as location parameters to draws from the Poisson distribution.

Once the number of births are calculated, SHAPE will determine if mutants are generated.

5. Mutation Events

The last step of each generation is to calculate if there have been mutation events. The number of mutations is controlled principally by the per genome, per generation, mutation rate μ. Classically the number of mutants is a product of μ and the number of replication (*i.e.* birth) events, but more recent theoretical work has suggested using all living individuals [7]. SHAPE allows users to choose either. The number of mutants generated from replication events is calculated as:

$$mutants_i = \mu \, births_i \, \frac{r}{r-1} \tag{14}$$

where recall r is the number of offspring expected from a single birth event and the term $\frac{r}{r-1}$ ensures that the number of mutants considers not just offspring but also the parental individuals. If mutants can arise from any individual in the population, SHAPE adds the following vector:

$$\mu \, (N_i - deaths_i - \frac{births_i}{r-1}) \tag{15}$$

to the values calculated in eq. 14. The first time any genotype generates mutants, SHAPE will calculate, and record, the fitness value of all genotypes in the unexplored neighbouring mutational space. The fitness for a genotype is calculated based on the chosen fitness landscape model (see *General Parameters* above). For each genotype i with at least one mutant, SHAPE draws $mutant_i$ times (with replacement) from the list of genotypes in the neighbouring mutational space. Recall users may control if revertant mutants are considered.

SI Figures

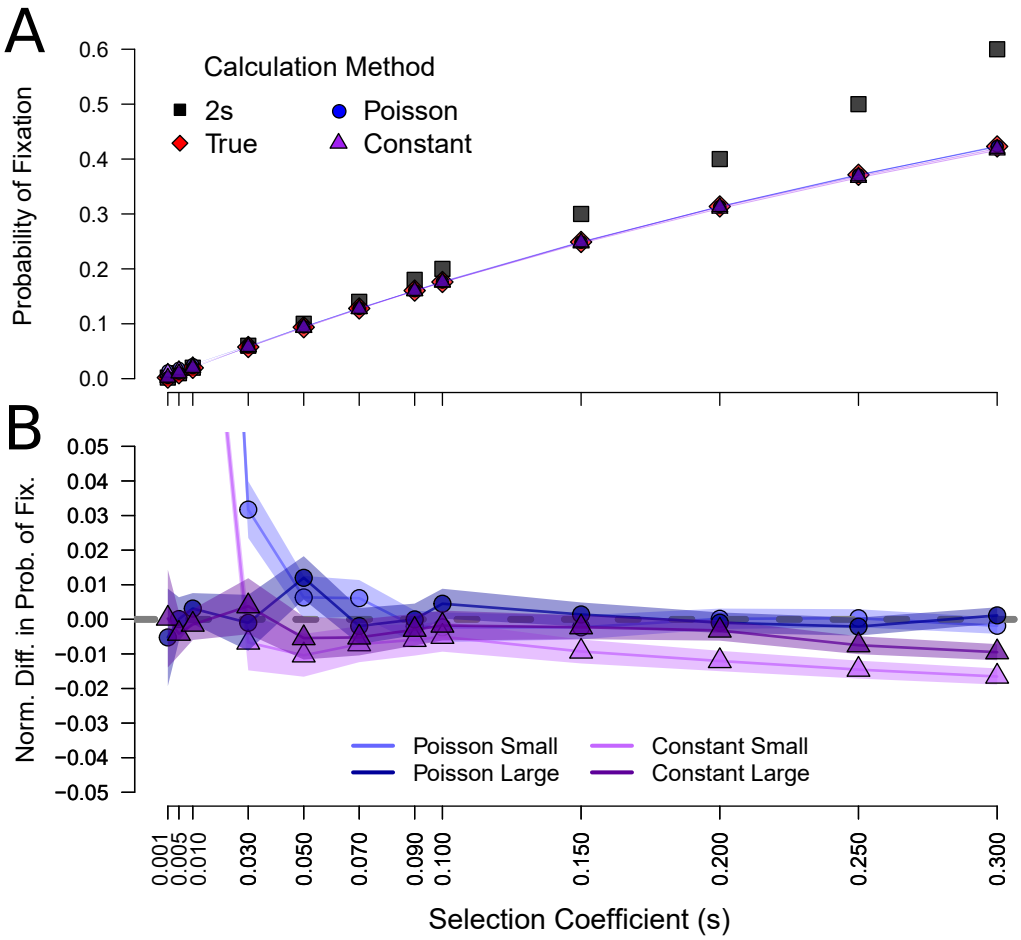

Fig 6. Comparison between the probability of a single mutant fixing, in an environment with a constant number of individuals, as calculated with SHAPE or using a theoretical model [3]. Panel A shows the fixation probability, dependent on the selection coefficient s, where coloured circles represent the values calculated using different constant growth models (*Poisson* - blue; *Constant* - purple) implemented in SHAPE, red diamonds represent the true fixation probability and black squares represent the analyitcal approximation of $2s$. For values calculated with SHAPE, lighter colours were based on a smaller (1000) number of initial individuals and darker colours for the larger number (10^7). In B, I present the normalised difference between the true fixation probability and the values calculated with SHAPE. Negative values reflect when the values calculated with SHAPE are lower than the true fixation probability.

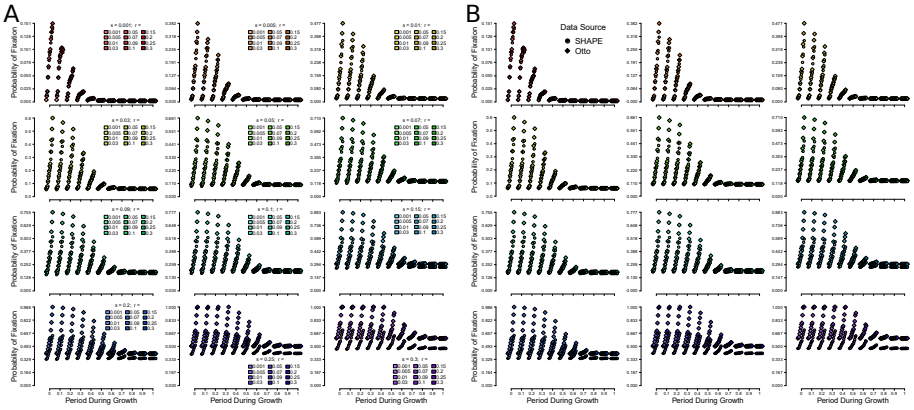

Fig 7. Comparison of the theoretical and estimated fixation probability for a mutant growing logistically. Diamond shapes represent the analytical approximation [5] derived similarly to Haldane's $\approx 2s$ while circles are for estimates using SHAPE. The colour used to fill points reflects the selection coefficient s and darker colours represent higher intrinstic growth rates r. The period during the growth phase is scaled from the start of growth until the point where the number of individuals reaches carrying capacity. The range of parameters is similar between panels but A is for when the initial number of individuals was 100 whereas B is for an initial 10^4 individuals.

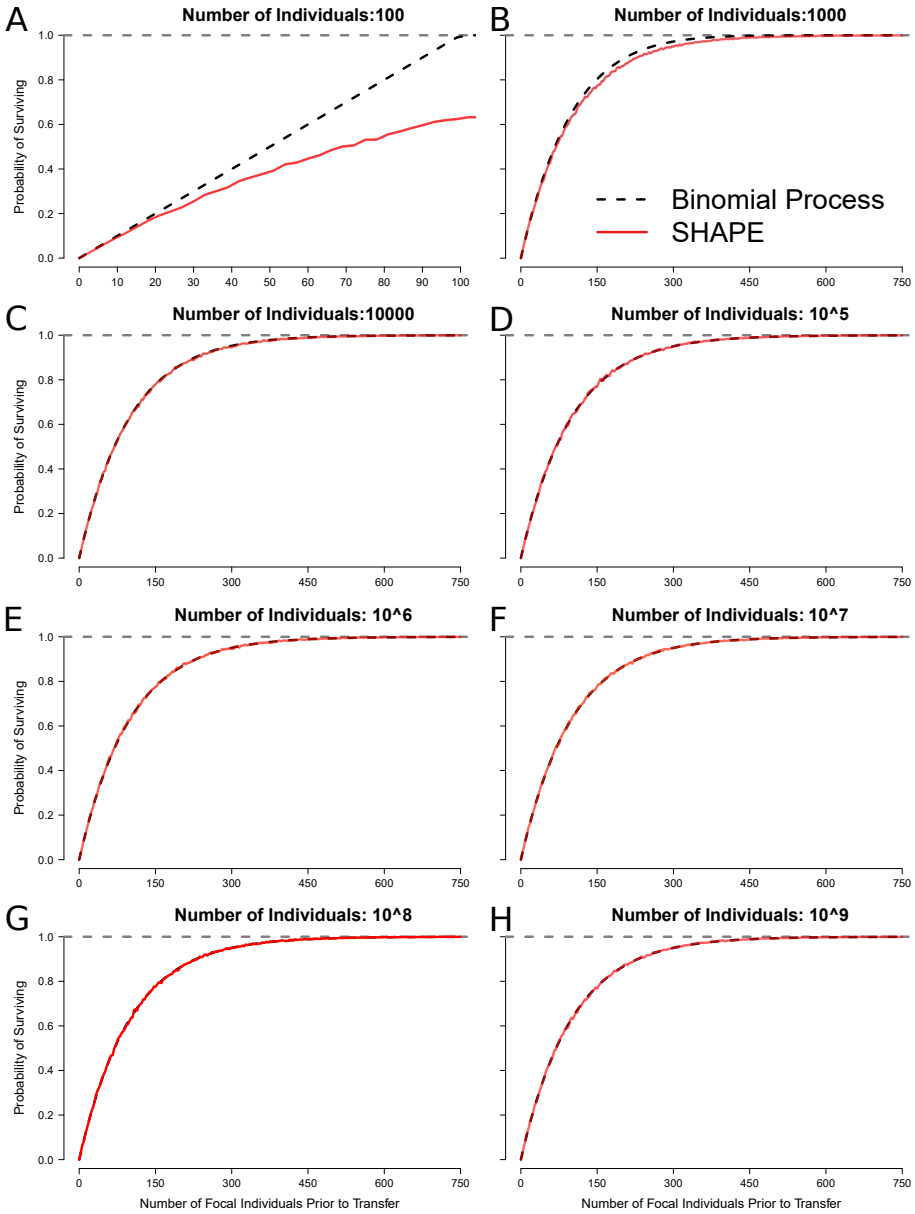

Fig 8. The probability that at least one individual of a focal genotype survives a disturbance event (serial passaging) reducing the total number of individuals 100 fold ($D = 100$). The solid red line shows the 95% CI of survival probability calculated with SHAPE, while the black dashed line shows the deterministic expectation calculated from a binomial process. The dashed grey line highlights when probability reaches unity. Panels differ in the total number of individuals and number of focal individuals prior to transfer.