Adaptive evolution with aneuploidy and mutation

Ilia Kohanovski[1,*], Martin Pontz[2,*], Avihu H. Yona[3], and Yoav Ram[1,2,†]

[1]School of Computer Science, Reichman University, Herzliya, Israel

[2]School of Zoology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

[3]Institute of Biochemistry, Food Science and Nutrition, Robert H. Smith Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, Israel

[*]These authors contributed equally to this work

[†]Corresponding author: yoav@yoavram.com

January 24, 2022

Abstract

Aneuploidy is common in eukaryotes, often leading to decreased cell growth and fitness. However, evidence from yeast and fungi, as well as human tumour cells, suggests that aneuploidy can be beneficial under stressful conditions and lead to elevated growth rates and adaptation. Importantly, aneuploidy differs from point mutations in rate, fitness effect, and reversibility. Here, we develop evolutionary models for adaptive evolution with both mutation and aneuploidy. These models are used within an approximate Bayesian computation framework to estimate the formation rate and fitness effect of aneuploidy and mutation from results of evolutionary experiments in which *Saccharomyces cerevisiae* adapted to heat stress: the experimental populations first acquired chromosome duplications, only to later revert back to a euploid state. We also analyze our models to estimate the effect of the aneuploidy and mutation rates on the expected adaptation time and the probability for adaptation via aneuploidy. Our results suggest that aneuploidy can be a transient adaptive solution, which can decelerate adaptation in a non-intuitive manner. By creating an evolutionary conflict between the individual and the population, aneuploidy further complicates the process of adaptation in cell populations.

Introduction

Aneuploidy is common in eukaryotes. Aneuploidy is an imbalance in the number of chromosomes in the cell: an incorrect karyotype. Evidence suggests aneuploidy is very common in eukaryotes, e.g. animals (**???**), and fungi (**????**). Aneuploidy has been implicated in cancer formation and progression (**??**): 90% of solid tumours and 50% of blood cancers are aneuploid (**?**). Aneuploidy is also linked to the emergence of drug resistance (**?**) and virulence (**?**) in fungal pathogens, which are under-studied (**?**) despite infecting close to a billion people per year, causing serious infections and significant morbidity in >150 million people per year and killing >1.5 million people per year (**??**). In addition, aneuploidy is common in protozoan pathogens of the Leishmania genus, a major global health concern (**?**).

Aneuploidy is generally deleterious. The molecular and genetic mechanisms involved in aneuploidy have been explored (**??????**). Experiments with human and mouse embryos found that aneuploidy is usually lethal. It is also associated with developmental defects and lethality in other multicellular organisms (**?**). For example, aneuploid mouse embryonic cells grow slower than euploid cells (**?**). Similarly, in unicellular eukaryotes growing in benign conditions, aneuploidy usually leads to slower growth and decreased overall fitness (**?????**), in part due to proteotoxic stress caused by increased expression in aneuploid cells (**???**) and hypo-osmotic-like stress (**?**).

Aneuploidy can lead to adaptation. However, aneuploidy can be beneficial under stressful conditions due to the wide range of phenotypes it can produce, some of which are advantageous (**?**). Thus, aneuploidy can lead to rapid adaptation in unicellular eukaryotes (**????**), as well as to rapid growth of somatic tumour cells (**??**). For example, aneuploidy in *S. cerevisiae* facilitates adaptation to a variety of stressful conditions like heat and pH (**?**), copper (**??**), salt (**?**), and nutrient limitation (**??**). Importantly, aneuploidy can also lead to drug resistance in pathogenic fungi such as *Candida albicans* (**???**) and *Cryptococcus neoformans* (**?**), which cause candidiasis and meningoencephalitis, respectively.

Transient adaptive solution. Aneuploidy differs from mutation due to its distinct properties. Chromosome duplication usually occurs more often than mutation and on average produces larger fitness effects. Yet, because it affects many genes on a whole chromosome or a chromosome fragment, aneuploidy also carries fitness costs. Thus, aneuploidy can be a *transient adaptive solution*: it can rapidly occur and fix in the population under stressful conditions, and can be rapidly lost when the cost outweighs the benefit—when stress is removed or after beneficial mutations occur. Experimental

evidence of such a transient role of aneuploidy was demonstrated by **?**. They evolved populations of *S. cerevisiae* under strong heat or pH stress. The populations adapted to the heat and pH stress within 450 and 150 generations, and this adaptation was determined to be due to chromosome duplications. Much later, after more than 1500 and 750 generations, for the heat and pH stress, respectively, the populations reverted back to an euploid state, while remaining adapted to the stress and accumulating multiple mutations. However, under gradual heat stress, aneuploidy was not observed. **?** concluded that aneuploidy serves as a transient adaptive solution, or a "quick fix", which is expected to facilitate adaptation.

The present study. Here, we develop evolutionary-genetic models that include the effects of natural selection, genetic drift, aneuploidy, and mutation to examine the role of aneuploidy in adaptive evolution. These models follow a population of cells characterised by both their ploidy and their genotype. We fit these models to the experimental results of **?** using an *approximate Bayesian computation* framework (**?**) to infer model parameters, including selection coefficients and rates of aneuploidy and mutation, and to perform model selection between different models, thereby testing hypotheses about the evolutionary process. We analyze these evolutionary-genetic models to estimate the effects of parameters on the adaptation time and the probability for adaptation via aneuploidy. We find that **TODO**

Models and Methods

Evolutionary Models. We model the evolution of a population of cells using two models: a single-locus model and a multi-locus model. Both models are based on the Wright-Fisher model (**?**), assuming non-overlapping generations and including the effects of natural selection, genetic drift, aneuploidy, and mutation. We focus on beneficial mutations, neglecting the effects of deleterious and neutral mutations. Both models allow for a single aneuploid karyotype (e.g., chromosome III duplication). While the single-locus model allows for only a single mutation to accumulate in the genotype, the multi-locus model allows for multiple mutations to accumulate (Figure 1), as well as for a fluctuating population size.

Single-locus model. This model assumes a constant effective population size N and follows four genotypes (Figure 1A): euploid wild-type, $2n$, the initial genotype; euploid mutant, $2n*$, with the standard karyotype and a single beneficial mutation; aneuploid wild-type, $2n+1$, with an extra chromosome, i.e., following chromosome duplication; and aneuploid mutant, $2n+1*$, with and extra chromosome and a beneficial mutation.

Transitions between the genotypes occur as follows (Figure 1A): Beneficial mutations from $2n$ to $2n*$ and from $2n+1$ to $2n+1*$ occur with probability μ, the mutation rate. We neglect back-mutations (i.e., from $2n*$ to $2n$ and from $2n+1*$ to $2n+1$). Aneuploidy is formed by chromosome mis-segregation, so that cells transition from $2n$ to $2n+1$ and from $2n+1*$ to $2n*$ with probability δ, the aneuploidy

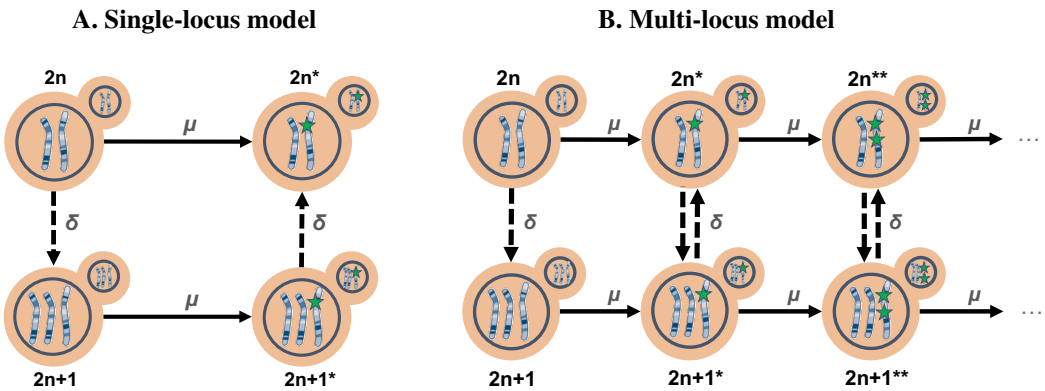

Figure 1: Model illustrations. **(A)** In the single-locus model, the four genotypes are: euploid wild-type, $2n$; euploid mutant, $2n*$; aneuploid wild-type, $2n+1$; and aneuploid mutant, $2n+1*$. Overall there are two possible trajectories from $2n$ to $2n*$. **(B)** In the multi-locus model, each genotype is characterized by its karyotype, $2n$ or $2n+1$, and the number of accumulated beneficial mutations, denoted by stars. In both panels arrows denote transitions between genotypes, with transitions rates: μ, beneficial mutation rate; δ, aneuploidy rate.

rate. That is, we assume chromosomes are gained and lost at the same rate, and we neglect events that form a less-fit genotype (i.e., *2n+1* to *2n* and *2n** to *2n+1**). The fitness values of the four genotypes are given by Table 1.

Table 1: Single-locus model fitness values.

Genotype i	$2n$	$2n+1$	$2n+1^*$	$2n^*$
Fitness w_i	1	$1-c+b$	$(1-c)(1+s)+b$	$1+s$

$s \geq 0$ is the selection coefficient of a beneficial mutation; $0 \leq c \leq 1$ is the fitness cost of aneuploidy; and $b \geq c$ is the selection coefficient, or fitness benefit, of aneuploidy.

The initial population has N cells with genotype *2n*. The effect of natural selection on the frequency f_i of genotype $i = 2n, 2n+1, 2n+1^*$, or $2n^*$ is given by

$$f_i^s = \frac{f_i w_i}{\bar{w}} \,, \tag{1}$$

where the fitness values w_i are given in Table 1 and $\bar{w} = \sum_j f_j w_j$ is the population mean fitness. The effect of mutation and aneuploidy on genotype frequencies is given by

$$\begin{aligned} f_{2n}^m &= (1-\delta-\mu) f_{2n}^s \,, \\ f_{2n+1}^m &= \delta f_{2n}^s + (1-\mu) f_{2n+1}^s \,, \\ f_{2n+1^*}^m &= \mu f_{2n+1}^s + (1-\delta) f_{2n+1^*}^s \,, \\ f_{2n^*}^m &= \mu f_{2n}^s + \delta f_{2n+1}^s + f_{2n^*}^s \,. \end{aligned} \tag{2}$$

Finally, random genetic drift is modeled using a multinomial distribution (**?**),

$$\mathbf{f'} \sim \frac{1}{N} \cdot Mult(N,\ \mathbf{f^m}) \,, \tag{3}$$

where $\mathbf{f^m} = (f_{2n}^m\ ,\ f_{2n+1}^m\ ,\ f_{2n+1^*}^m\ ,\ f_{2n^*}^m)$ are the frequencies of the genotypes after mutation and aneuploidy, $\mathbf{f'}$ are the genotype frequencies in the next generation, and $Mult(N,\ \mathbf{f})$ is a multinomial distribution parameterized by the population size N and the genotype frequencies $\mathbf{f}$. Overall, the change in genotype frequencies from one generation to the next is given by the transformation $f_i \rightarrow f_i'$.

Multi-locus model. This model expands the single-locus model by allowing for (i) the accumulation of beneficial mutations, and (ii) a fluctuating population size.

A genotype is characterized by its karyotype, *2n* or *2n+1*, and the number of accumulated beneficial mutations, which can be zero or more. The selection coefficient of the i-th accumulated mutation in each individual, s_i, is drawn from an exponential distribution with expected value s: $s_i \sim Exp(s)$. The

rest of the parameters (N, μ, δ, b, c) are the same as in the single-locus model. However, since the multi-locus model allows several mutations with smaller fitness effects to accumulate, we expect the mutation rate to be higher compared to the single-locus model, which focuses on a single, large-effect mutation.

The fitness of the different genotypes is the same as in the single-locus model (Table 1), except that the fitness contribution of k beneficial mutations is the product of their independent effects, $\prod_{i=1}^{k}(1+s_i)$, instead of the contribution of the single mutation allowed in the single-locus model, $(1+s)$, see Table 2. Therefore, aneuploidy loss would be favored by selection only if there are enough beneficial mutations and/or the selection coefficients s_i are large enough. The intuition is that when the benefit of the accumulated beneficial mutations is small, ten the benefit of aneuploidy has a large effect; when the benefit of the accumulated beneficial mutations is large, then aneuploidy is no longer beneficial due of its fitness cost.

In contrast to the single-locus model, in the multi-locus model the population size fluctuates to model serial-transfer experimental protocol (**?**): the population is serially diluted by transferring a fraction of the population (1/120) to a fresh medium approximately every seven generations. Thus, the population initial size is $N_0 = N$, and the population size is doubled every generation, $N_1 = 2N, N_2 = 4N, \ldots$, and diluted back to N after seven generations such that $N_8 = N$.

The change in frequencies due to selection is exactly the same as in the single-locus model (Equation 1), only applied using the fitness values in Table 2. The change due to random genetic drift is also the same as in Equation 3, except that the frequencies vector is $\mathbf{f} = (f_{2n}\,,\ f_{2n+1}\,,\ f_{2n^*}\,,\ f_{2n+1^*}\,,\ f_{2n^{**}}\,,\ f_{2n+1^{**}}\,,\ \ldots)$ and that the population size changes between generations, as described above.

The effects of mutation and aneuploidy on genotype frequencies is more elaborate than in the single-locus model. Genotype i is classified according to its karyotype (*2n* or *2n+1*), the number of accumulated beneficial mutations ($k \geq 0$), and their fitness (w_i). Each offspring cell inherits these properties from its mother cell. Then, a new beneficial mutation is accumulated with probability μ, such that the number of mutations is $k+1$, and its effect s_{k+1} is drawn from an exponential distribution with expected value s, such that the contribution of the mutations to the fitness is $\prod_{j=0}^{k+1}(1+s_j)$. Next, euploid offspring become aneuploid with probability δ, and aneuploid offspring become euploid with probability δ_L.

Empirical evidence. We use the results of evolutionary experiments reported by **?**. In their heat-stress experiment, four populations of *S. cerevisiae* evolved under 39 °C. Aneuploidy fixed in all four

Table 2: Multi-locus model fitness values.

Genotype, g	$2n$	$2n+1$	$2n+1^{*k}$	$2n^{*k}$
Fitness, w_g	1	$1-c+b$	$(1-c)\prod_{j=1}^{k}(1+s_i)+b$	$\prod_{j=1}^{k}(1+s_i)$

k is the number of accumulated beneficial mutations; $s \geq 0$ is the selection coefficient of a beneficial mutation; $0 \leq c \leq 1$ is the fitness cost of aneuploidy; and $b \geq c$ is the selection coefficient, or fitness benefit, of aneuploidy.

population in the first 450 generations (hereafter, fixation or elimination of a genotype *by generation* t means that more than 95% or less than 5% of the population carry the genotype at generation t, and possibly earlier). From unpublished results, aneuploidy did not fix before at least 200 generations elapsed. The experiment continued with two populations, in which aneuploidy was eliminated by generation 1,700 and 2,350.

Likelihood function. Because our model, just like the Wright-Fisher model, is non-linear and stochastic, computing the distribution of fixation time $T(g)$ of genotype g for use in the likelihood function is intractable (it is even hard to use a diffusion-equation approximation due to the model having multiple genotypes, rather than just two). We overcome this problem by approximating the likelihood using simulations. We simulate 1,000 experiments per parameter vector $\theta = (\mu, \delta, s, b, c)$, resulting in a set of simulated observations $\tilde{\mathbf{X}} = \{\tilde{X}_i\}_{i=1}^{1000}$. We then compute the approximate likelihood,

$$\begin{aligned}\mathcal{L}(\theta) = P^4(200 \leq T(2n+1) \leq 450) \cdot \Big[1- \\ P^4_{\tilde{\mathbf{X}}}\big(!\{T(2n^*) < 1700\} \mid 200 \leq T(2n+1) \leq 450\big)- \\ P^4_{\tilde{\mathbf{X}}}\big(!\{1700 < T(2n^*) < 2350\} \mid 200 \leq T(2n+1) \leq 450\big)+ \\ P^4_{\tilde{\mathbf{X}}}\big(!\{T(2n^*) < 1700\} \wedge !\{1700 < T(2n^*) < 2350\} \mid 200 \leq T(2n+1) \leq 450\big)\Big] ,\end{aligned} \tag{4}$$

where $!\{\ldots\}$ is the "logical not" operator, $P^4(\ldots)$ is the 4th power of $P(\ldots)$, and all probabilities $P_{\tilde{\mathbf{X}}}(\ldots)$ are approximated from the results of the simulations $\tilde{\mathbf{X}}$. For example, $P_{\tilde{\mathbf{X}}}\big(!\{T(2n^*) < 1700\} \mid 200 \leq T(2n+1) \leq 450\big)$ is approximated by taking simulations in which *2n+1* fixed before generation 450 but not before generation 200, and computing the fraction of such simulations in which *2n** did not fix by generation 1,700, and hence aneuploidy did not extinct before generation 1,700. Figure S4 compares results with less and more simulated experiments, demonstrating that 1,000 simulations are likely enough.

For a model without aneuploidy (that is, when the aneuploidy rate is fixed at zero, $\delta = 0$), the likelihood

is similarly approximated by

$$\begin{aligned}\mathcal{L}_!(\theta) = 1 - P^4_{\tilde{\mathbf{X}}}\big(!\{T(2n^*) < 1700\}\big) - \\ P^4_{\tilde{\mathbf{X}}}\big(!\{1700 < T(2n^*) < 2350\}\big) + \\ P^4_{\tilde{\mathbf{X}}}\big(!\{T(2n^*) < 1700\} \wedge !\{1700 < T(2n^*) < 2350\}\big)\ .\end{aligned} \tag{5}$$

Parameter inference. To infer model parameters, we use approximate Bayesian computation with a sequential Monte-Carlo scheme, or ABC-SMC (**?**), implemented in the `pyABC` Python package (**?**, pyabc.readthedocs.io). This approach uses numerical stochastic simulations of the model to infer a posterior distribution over the model parameters. It is a method of likelihood-free, simulation-based inference (**?**), that is, for estimating a posterior distribution when a likelihood function cannot be directly computed. It is therefore suitable in our case, in which the likelihood function can only be approximated from simulations, and cannot be directly computed.

The ABC-SMC algorithm employs sequential importance sampling over multiple iterations (**???**). In iteration t of the algorithm, a set of parameter vectors, $\{\theta_{i,t}\}_{i=1}^{n_t}$, also called *particles*, are constructed in the following way. A proposal particle, θ^*, is sampled from a proposal distribution, and is either accepted or rejected, until n_t particles are accepted. The number of particles, n_t, is adapted at every iteration t using the adaptive population strategy (**?**, pyabc.readthedocs.io). For $t = 0$, the proposal particle is sampled from the prior distribution, $p(\theta)$. For $t > 0$, the proposal particle is sampled from the particles accepted in the previous iteration, $\{\theta_{i,t-1}\}_{i=1}^{n_{t-1}}$, each with a probability relative to its weight $W_{t-1}(\theta_{i,t-1})$ (see below). The proposal particle is then perturbed using a kernel perturbation kernel, $K_t(\theta^* \mid \theta)$ where θ is the sample from the previous iteration. Then, a set of synthetic observations $\tilde{\mathbf{X}}^*$ is simulated, and the proposal particle θ^* is accepted if its approximate likelihood (Equation 4) is high enough, $\mathcal{L}(\theta^*) > 1 - \epsilon_t$ (or more commonly, if $1 - \mathcal{L}(\theta^*) < \epsilon_t$), where $\epsilon_t > 0$ is the *acceptance threshold*, as higher values of ϵ_t allow more particles to be accepted. The acceptance threshold ϵ_t is chosen as the median of the $1 - \mathcal{L}(\theta)$ of the particles accepted in the previous iteration, $t - 1$, and $\epsilon_0 = 0.01$. For each accepted particle $\theta_{i,t}$ a weight $W_t(\theta_{i,t})$ is assigned: for $t = 0$, $W_0(\theta_{i,0}) = 1$, and for $t > 0$, $W_t(\theta_{i,t}) = p(\theta_{i,t}) / \sum_{i=1}^{n_{t-1}} W_{t-1}(\theta_{i,t-1}) K_t(\theta_{i,t}, \theta_{i,t-1})$, where $p(\theta)$ is the prior density of θ and $K_t(\theta', \theta)$ is the probability of a perturbation from θ to θ'. $K_t(\theta' \mid \theta)$ is a multivariate normal distribution, fitted at iteration t to the particles from the previous iteration, $\{\theta_{i,t-1}\}_{i=1} n_{t-1}$, and their weights, $\{W(\theta_{i,t-1})\}_{i=1}^{n_{t-1}}$.

Acceptance is determined according to the approximate likelihood (Equation 4), which has a maximum value of 0.875. Thus, we terminated the inference when $\epsilon \leq 0.13$ after six iterations, with $n_6 = 982$ accepted parameter vectors and effective sample size ESS=651 (Figure S3). Running the

inference algorithm with different initialization seeds and less or more simulations for approximating the likelihood produced similar posterior distributions (Figure S4).

After producing a set of weighted particles from the the posterior distribution using the above ABC-SMC algorithm, we approximate the posterior using kernel density estimation (KDE) with Gaussian kernels, from which we find the MAP (maximum a posteriori) estimate as the maximum of the KDE function. We then draw 50, 000 samples from the posterior KDE to compute the HDI (highest density interval) and visualize the posterior distribution with histograms.

Model comparison. We perform model selection using WAIC, the widely applicable information criterion (**?**),

$$WAIC(\theta) \;=\; -2\log \mathbb{E}[\mathcal{L}(\theta)] \;+\; 2\mathbb{V}[\log \mathcal{L}(\theta)] \tag{6}$$

where θ is a parameter vector, and $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ are the expectation and variance taken over the posterior distribution. WAIC values are scaled as a deviance measure: lower values imply higher predictive accuracy (**?**).

Prior distributions. We used informative prior distributions for $w_{2n+1} = 1 - c + b$, $w_{2n+1^*} = (1 + s)(1 - c) + b$ and $w_{2n^*} = 1 + s$, which we estimated from growth curves data from mono-culture growth experiments previously reported by **?**, Figs. 3C, 4A, and S2. We used `Curveball`, a method for predicting results of competition experiments from growth curve data (**?**, curveball.yoavram.com). Briefly, `Curveball` takes growth curves of two strains growing separately in mono-culture and predicts how they would grow in a mixed culture, that is, it predicts the results of a competition assay. From these predictions, relative fitness values can be computed. Because `Curveball` uses a maximum-likelihood approach to estimate model parameters, we were able to estimate a distribution of relative fitness values by sampling from a truncated multivariate normal distribution defined by the maximum-likelihood covariance matrix. We sampled 10,000 samples to use as a prior distribution (Figure S1). We used growth curves of *2n* and *2n+1* in 39 °C to estimate a prior distribution for w_{2n+1}. In lieu of a better prior, we used the same prior for w_{2n+1^*} and w_{2n^*}. To increase computational efficiency, we also assumed $w(2n*) > w(2n + 1*) > w(2n + 1) > w(2n)$; running the inference without this assumption produced similar results.

Compared to other priors we tested, this prior produced lower WAIC, better posterior prediction plots, and more stable parameter estimates, as follows. We tried to use additional growth curves. We used growth curves of *2n** (*refined* strain from **?**) and *2n+1* in 39 °C to estimate w_{2n*}/w_{2n+1}. The same prior was used for w_{2n*}/w_{2n+1*}. This prior resulted in WAIC 0.32, compared to 0.27 with the above

informative prior. We also tried to use growth curves of *2n+1* and *2n* in 30 °C to estimate $1 - c$. This estimation assumes that the cost of aneuploidy is the same in 39 °C and 30 °C; this might be incorrect, but we only assumed this to generate a prior distribution for the fitness values. The prior for b was taken the same as for c. This prior did not converge: no parameter sets were found with an approximate likelihood greater than zero. Finally, we also tested a uninformative uniform prior with $U(1, 6)$, for (i) all w_{2n+1}, w_{2n+1^*}, w_{2n^*} and (ii) only for w_{2n+1^*}, w_{2n^*}, using the above informative prior for w_{2n+1}. In both cases the algorithm failed to converge.

For the mutation rate, μ, and aneuploidy rate, δ, we used uninformative uniform priors, $\mu \sim U(10^{-9}\ ,\ 10^{-5})$ and $\delta \sim U(10^{-6}, 10^{-2})$. For a model without aneuploidy ($\delta = 0$), we used a wider uniform prior for the mutation rate, $\mu \sim U(10^{-10}, 10^{-5})$.

Results

Inference: single-locus model

Parameter estimation. We used ABC-SMC to infer the posterior distribution of the parameters of the single-locus model. The MAP (maximum a posteriori) and 50% HDI (highest density interval) of the parameters posterior are (Figure 2): mutation rate, $\mu = 2.93^{+1.122}_{-2.824} \times 10^{-6}$, aneuploidy rate, $\delta = 1.716^{+0.974}_{-0.384} \times 10^{-3}$, and fitness values, $w_{2n+1} = 1.022^{+0.001}_{-0.002}$, $w_{2n+1^*} = 1.025^{+0.001}_{-0.001}$, $w_{2n^*} = 1.028^{+0.001}_{-0.001}$, all relative to the fitness of *2n*, which is set to $w_{2n} = 1$. This estimated aneuploidy rate agrees with previous estimates. The mutation rate corresponds to a mutation target size of 10^4, assuming the mutation rate per base pair is roughly $2 \cdot 10^{-10}$ (**?**).

We also tried to infer parameters of the single-locus model when the mutation rate in aneuploid cells, compared to euploid cells, is increased by a factor τ=1, 33/32 (due to an additional chromosome), 2, 5, 10 or 100 (due to genetic instability). We found that for τ =33/32, 2, and 5 the posterior was not significantly different from the model with $\tau = 1$, while for $\tau = 10$ the estimated mutation rate was slightly lower, $\mu = 1.654^{+0.879}_{-1.775} \times 10^{-6}$, and for $\tau = 100$ the estimated mutation rate was significantly lower and the aneuploidy rate had higher variability. The MAP and 50% HDI with $\tau = 100$ were: $\mu = 5.538^{+1.611}_{-5.311} \times 10^{-7}$, $\delta = 8.292^{+8.988}_{-4.96} \times 10^{-4}$, $w_{2n+1} = 1.023^{+0.001}_{-0.002}$, $w_{2n+1^*} = 1.024^{+0.001}_{-0.001}$, $w_{2n^*} = 1.028^{+0.001}_{-0.002}$. We thus continued our analysis under the assumption that $\tau = 1$, that is, that the mutation rate in euploid and aneuploid cells is the same.

Model checking and comparison. The single-locus model fits the data well: in simulations using the best-fit parameters (MAP estimate) *2n** fixed in 61% of simulations by generation 1,700 and in

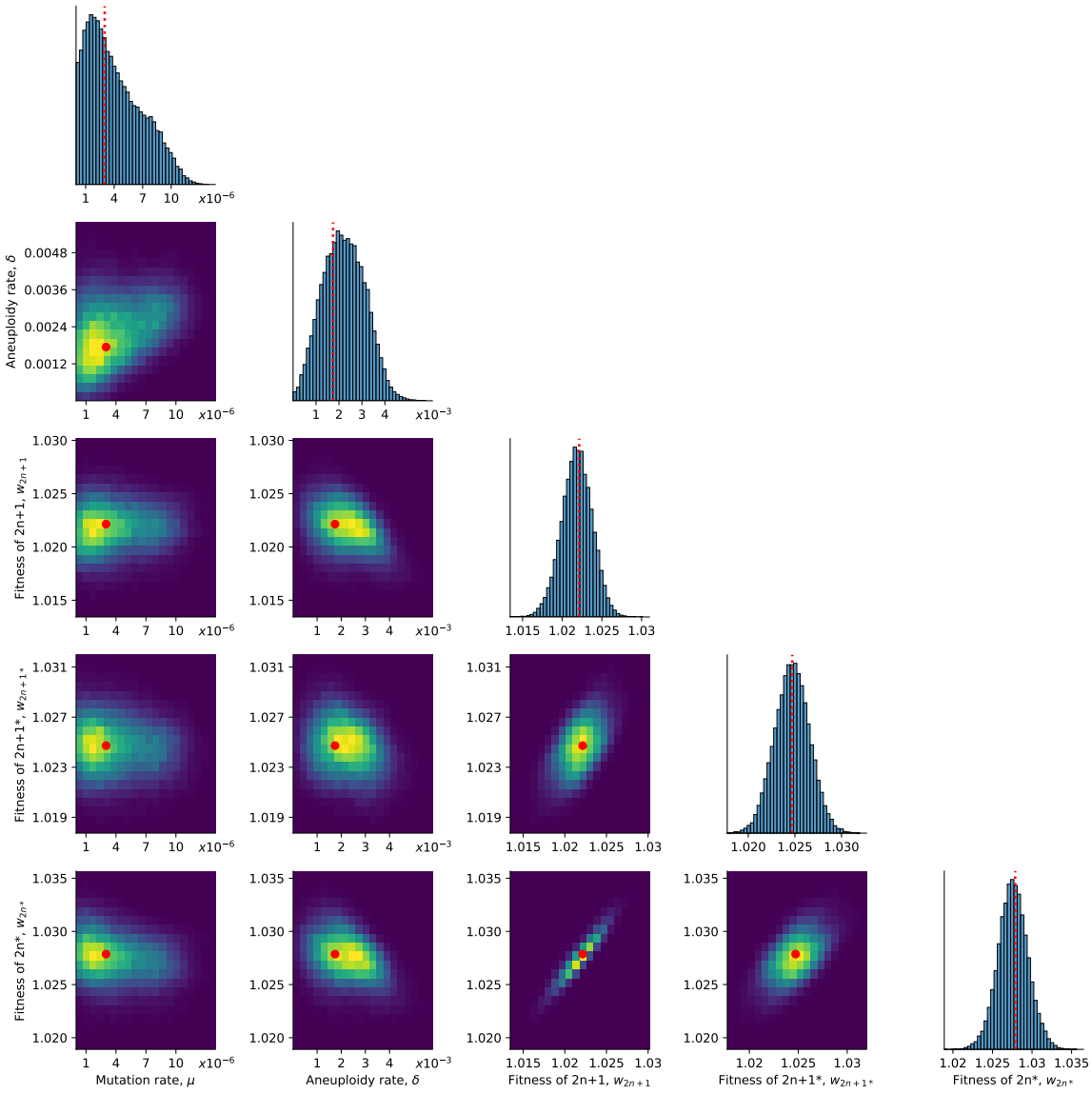

Figure 2: Posterior distribution of single-locus model. On the diagonal, the posterior kernel density estimate for each parameter. Below the diagonal, joint posterior density of two parameters (dark purple and bright yellow for low and high density, respectively). Red markers and orange lines for the joint MAP estimates (which may differ from the marginal maximum density, as the marginal distribution integrates over all other parameters).

100% of simulations by generation 2,350 (Figure 3-B) and *2n+1* fixed in roughly 300 generations on average (green lines in Figure 4), which agrees with experimental results. Interestingly, the genotype frequency dynamics in these simulations demonstrate that *2n+1** never reaches substantial frequency (Figure 4). Furthermore, sensitivity analysis shows that changing the parameter values from the MAP estimate reduces the model fit to the experimental results (Figure S2).

However, a model without aneuploidy where the aneuploidy rate is fixed at zero, $\delta = 0$, cannot explain the experimental observations (Figure 3). The estimated parameter values without aneuploidy are: mutation rate, $\mu = 7.837^{+0.078}_{-0.123} \times 10^{-9}$, fitness of mutant, $w_{2n^*} = 1.013^{+0.000}_{-0.000}$. We can see that the

mutation rate estimate in this model is much lower than in the model with aneuploidy. Higher mutation rate would cause quicker fixation of $2n*$ than in the experimental observations (Figure 3). Thus, even disregarding the appearance of trisomic cells in the experiment, the evidence supports the inclusion of aneuploidy in the model.

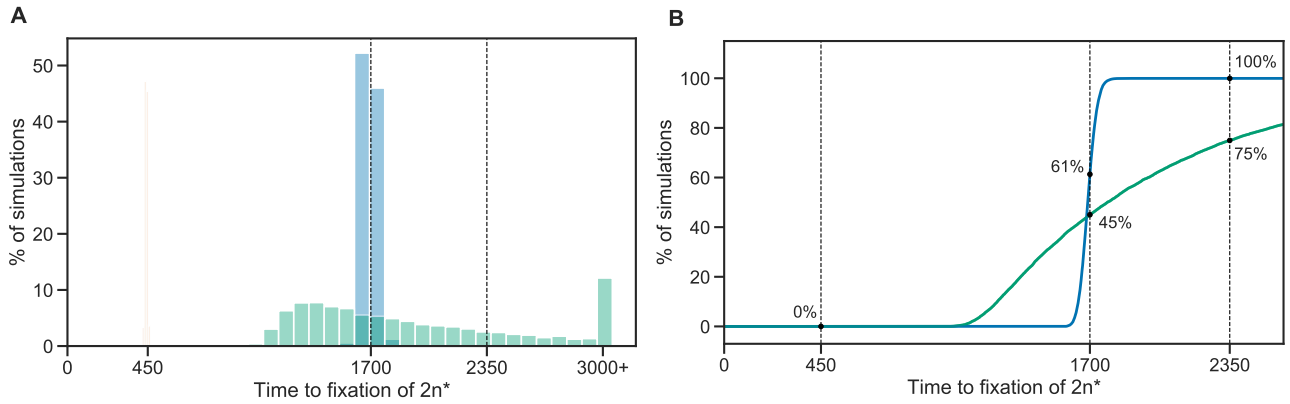

Figure 3: Single-locus model fit with and without aneuploidy. **(A)** The distribution of time to fixation of $2n*$ (i.e., adaptation time) in 10, 000 simulations of the single-locus model with aneuploidy (blue; MAP parameters) compared to two models without aneuploidy: a model with the same parameter values except $\delta = 0$ (orange), and a model fitted to the data assuming $\delta = 0$ (green). In the experiment by **?**, one population lost aneuploidy by generation 1,700 and another by generation 2,350 (dashed lines) but not before generation 450. Thus, the blue distribution is a better fit compared to the green, and the yellow histogram has a very poor fit. The last bin contains all the simulations with time equal or greater than 3,000. **(B)** Cumulative distribution of the time to fixation of $2n*$ in 10,000 simulations using the MAP estimate with and without aneuploidy in blue and green, respectively, and corresponding to the blue and green bars in panel A. The MAP likelihood (Equation 4) is 0.84 and 0.67 with and without aneuploidy, respectively.

Analysis: single-locus model. Here, we aim to analyse the timing of the single-locus model and the relative importance of the various parameters. As a stochastic model, there are two events that can occur from the wild type population. One is mutation that gives the adapted population state, or a missegregation event. The probability with which these events happen is hugely different. We are interested in the average waiting times until one of the two events occurs and lead to fixation. Let T_1 and T_2 denote the random variables for the waiting time for the events mutation and missegregation respectively. Due to discrete generation, they are geometrically distributed with parameter compound of the probability that at least one individual in this generation changes its state and goes to fixation:

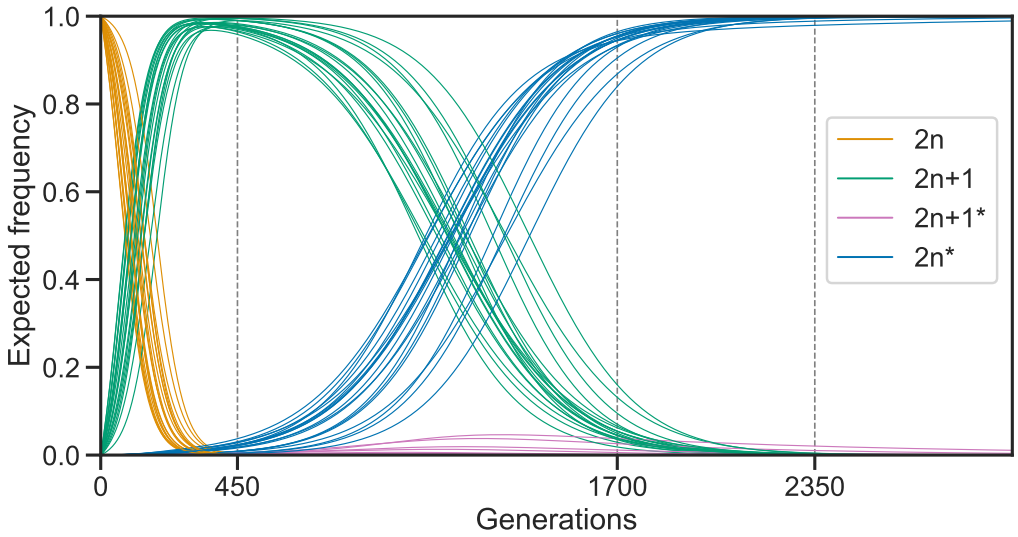

Figure 4: Single-locus model posterior genotype frequency dynamics. The posterior predicted frequency of the four genotypes. Each of 20 curves is the average of 10,000 simulations using parameters drawn from the posterior distribution (Figure 2).

$$T_{2n^*} \sim \text{Geo}(1-(1-2\mu\Delta w)^N) = \text{Geo}(p_{2n^*}) \tag{7a}$$

$$T_{2n+1} \sim \text{Geo}(1-(1-2\delta\Delta w)^N) = \text{Geo}(p_{2n+1}) \tag{7b}$$

Here, Δw denotes the relative difference between the fitness of the newly introduced type and the population mean fitness. However, since we have competing events, we need to find the minimum of these two processes. The minimum of two geometrically distributed processes with parameters p_{2n^*} and p_{2n+1} is again geometrically distributed with parameter $1-(1-p_{2n^*})(1-p_{2n+1})$. The average waiting time until the first event happens is the inverse of the paramter of the geometric distribution. The probability that aneuploidy will be the first event to happen, is given by

$$p_A = \frac{p_{2n+1}}{p_{2n^*}+p_{2n+1}} = \frac{1-(1-2(w_3-w_1)\delta)^N}{2-(1-2(w_3-w_1)\delta)^N-(1-2(w_2-w_1)\mu)^N}. \tag{8}$$

Since we are interested in the average time to fixation of $2n^*$, the steps that are needed differ depending on the actual event that happens first. If mutation happens first, then $2n^*$ is already reached and everything is fine. On the other hand, if aneuploidy occurs first, the population is in state $2n+1$ and several other steps have to occur, until $2n^*$ is reached. Thus, in this aneuploid trajectory, we also need to add the average waiting times to reach $2n+1^*$ and finally $2n^*$ The formulas are written down straightforward.

In general, we assume, that the waiting time until a new type occurs is much longer than the time the population needs to go to fixation for a certain type, but nonetheless, these times need to be included

for a rigorous analysis. There are two ways, how to compute them and they differ in the assumption of the underlying process.

If the population evolves deterministically, then one can derive the formula

$$t_{\text{fix, i}} = 2\frac{\log(N)}{\log(1+s)} \tag{9}$$

from the classical one-locus two-allele equations. As we work with simulations, we assume fixation if it reached 95% of the total population size. This is significantly different from full fixation. Thus we need a slightely different formula

$$t_{\text{fix, } f_c\text{, i}} = \frac{\log\left(\frac{f_c N}{1-f_c}\right)}{\log(1+s)}. \tag{10}$$

Here, f_c denotes the fraction for which we assume fixation.

In the second approach, the population is assumed to evolve stochastically. It is based on the diffusion equation and given in Kimura and Ohta 1969 (eq. 17). We denote it by $T_{\text{fix, i}}$.

Thus, the average time to $2n^*$ over the aneuploidy trajectory is given by

$$T_{\text{aneu}} = (1-(1-p_{2n^*})(1-p_{2n+1}))^{-1} + p_{2n+1^*}^{-1} + p_{2n^*}^{-1} + T_{\text{fix, 2n+1}} + T_{\text{fix, }2n+1^*} + T_{\text{fix, }2n^*}. \tag{11a}$$

The average time to $2n^*$ over the mutant trajectory is given by

$$T_{\text{mut}} = (1-(1-p_{2n^*})(1-p_{2n+1}))^{-1} + T_{\text{fix, }2n^*}. \tag{11b}$$

[MP: $T_{\text{fix, }2n^*}$ is not the same in (11) and (12). Need a better notation.]

In summary, the average time to $2n^*$ is given by

$$T = p_A T_{\text{aneu}} + (1-p_A)T_{\text{mut}}. \tag{12}$$

Remark 1. *There are several implicit assumptions that the equations hold like this.*

1. *The two processes fixation and occurrence of a new type are sequential and do not interfere with each other.*
2. *There is no clonal interference.*
3. *The fixation process starts from exactly one individual.*

There are certainly combinations of the mutation rate, the missegregation rate, the selection strength and the population size for which at least one of these assumptions is violated. Thus, (11) *and* (12) *may not yield a good approximation to the actual time to the adapted state.*

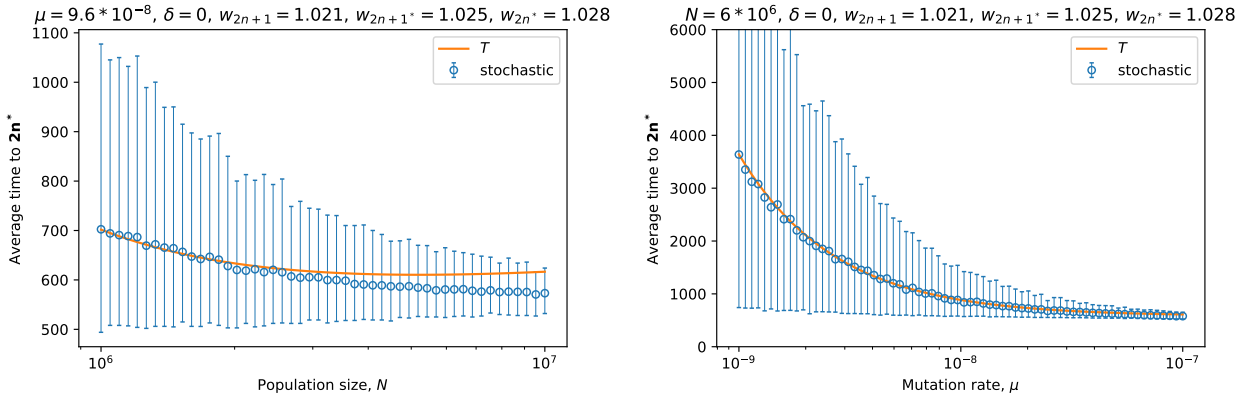

Figure 5: The average time to $2n^*$ is shown as a function of the population size N in the left panel and as a function of the mutation rate in the right panel. The blue circles are the means of the simulation runs and the lines denote the span between the 5th and the 95th percentile. The orange line is the respective analytic approximation. In these two plots, the aneuploidy rate is zero, so only the mutational trajectory is available.

To get a feeling of the limitations of (12), we plot simulations with $\delta = 0$ in Fig. 5. The other paramters are similar to those infered above. Even there, thereare slight deviations for high paramter values. Clearly visible in the left panel and not so good to see in the right. We conjecture, that in this parameter range more than one individual mutates in one generation and our assumption for the fixation time, which starts from exactly one mutant, does not hold.

Inference: multi-locus model.

Discussion

Aneuploidy is not just another type of mutation. The published data indicate that, like mutation, aneuploidy can be both deleterious and beneficial (**??**). Nevertheless, there are important and fundamental differences between adaptation by aneuploidy and adaptation by beneficial mutations (**?**), which make aneuploidy a unique mechanism for generating genetic variation. First, the aneuploidy rate (i.e. the frequency of mis-segregation events) is significantly higher than the mutation rate (**?**). Thus, everything else being equal, adaptation by aneuploidy will be faster and more frequent. Second, fitness effects of aneuploidy are larger than those of the majority of mutations, on average, and are rarely neutral (**???**), allowing selection to quickly sort deleterious and beneficial genotypes. Third, the number of different karyotypes is considerably smaller than the number of different genotypes, and different karyotypes are likely to have different phenotypes (**?**). Therefore, exploration of the phenotype space by aneuploidy requires smaller populations and a shorter time span. Fourth, aneuploidy is a reversible state, as the rate of chromosome loss is high and the cost of aneuploidy is significant (**?**). Indeed, aneuploidy often provides a transient solution: under short-term stress conditions, aneuploidy reverts (chromosome number returns to normal) when the stress subsides; under long-term stress conditions, aneuploidy reverts when refined solutions, generated by beneficial mutations, take over (**?**). Finally, aneuploidy results in increased genome instability, potentially increasing genetic variation by a positive feedback loop (**???**), while also increasing its own transience.

Evolutionary theory of aneuploidy. The role of aneuploidy in adaptation has only recently been observed (**???**), and is largely missing from the literature on evolution and adaptation: the introductory textbook *Evolution* by **?** does not mention the word aneuploidy, and the graduate-level book *Mutation-Driven Evolution* by **?** only briefly mentions aneuploidy in the context of speciation, but not adaptation. In recent reviews of the literature, aneuploidy is suggested to play an important role in fungal adaptation (**??**) and cancer evolution (**???**), yet these reviews cite no theoretical studies nor any quantitative models. Indeed, evolutionary, ecological, and epidemiological studies mostly assume adaptation occurs via beneficial mutations, recombination, and sex. Therefore, there is a critical need to develop an evolutionary theory of aneuploidy like the evolutionary theories of other mechanisms for generation of genetic variation, e.g. mutation (**?**), recombination (**?**), and sex (**?**). An evolutionary theory of aneuploidy will be central to the interpretation of experimental and clinical observations and design of new hypotheses, experiments, and treatments (**?**). For example, despite the lack of

theoretical models, aneuploidy has been invoked in a new strategy to combat pathogens and tumour cells by setting "evolutionary traps" (**??**), in which a condition that predictably leads to emergence of aneuploidy is applied, followed by a condition that specifically selects against aneuploid cells.

Acknowledgements

We thank Yitzhak Pilpel, Orna Dahan, Lilach Hadany, Judith Berman, David Gresham, Shay Covo, Martin Kupiec, and Tal Simon for discussions and comments. This work was supported in part by the Israel Science Foundation (YR 552/19) and Minerva Stiftung Center for Lab Evolution (YR).

Supplementary Material

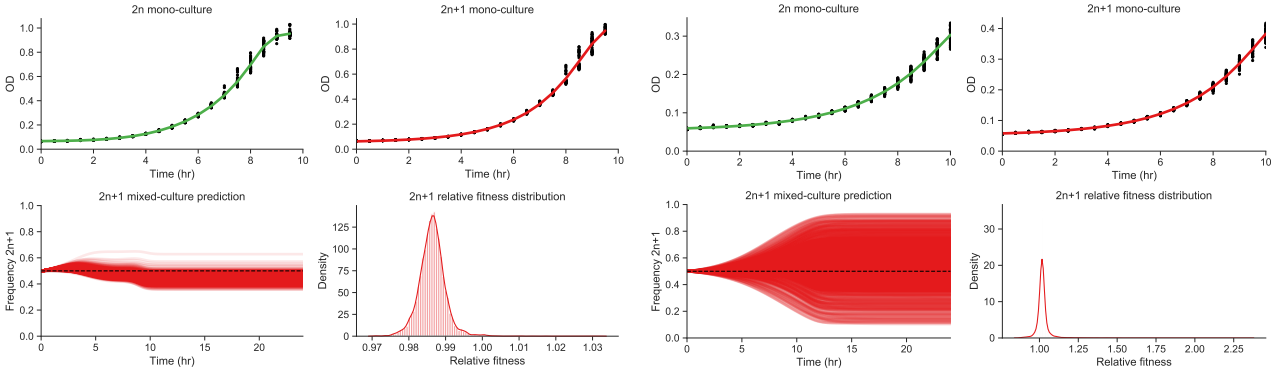

Figure S1: Fitness estimation from growth curves. **(A)** Fitness estimation from growth curves collected at 30°C **(B)** Fitness estimation from growth curves collected at 39°C Growth curves previously described in **?**, Figs. 3C, 4A, and S2. Fitness estimated from growth curves using `Curveball`, a method for predicting results of competition experiments from growth curve data (**?**, curveball.yoavram.com). See *Models and Methods, Prior distributions* for more details.

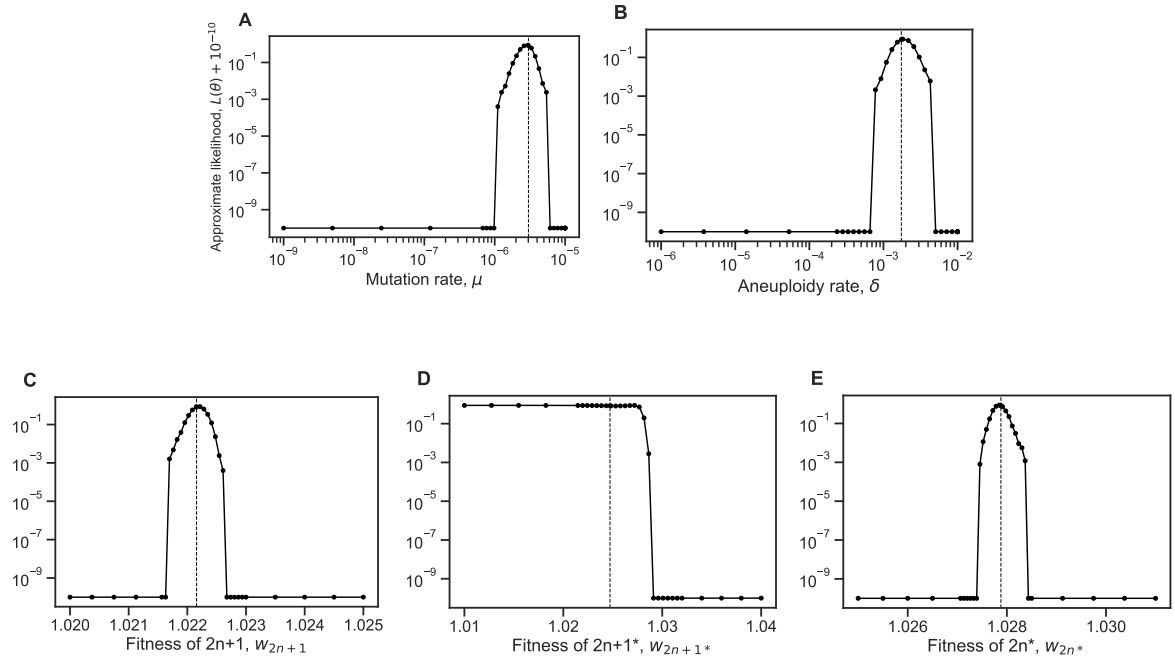

Figure S2: Likelihood profiles. Sensitivity of the approximate likelihood, $\mathcal{L}(\theta)$, of the single-locus model to changing a single parameter while the other parameters remain fixed at their MAP estimates. Dashed vertical line represents the MAP value.

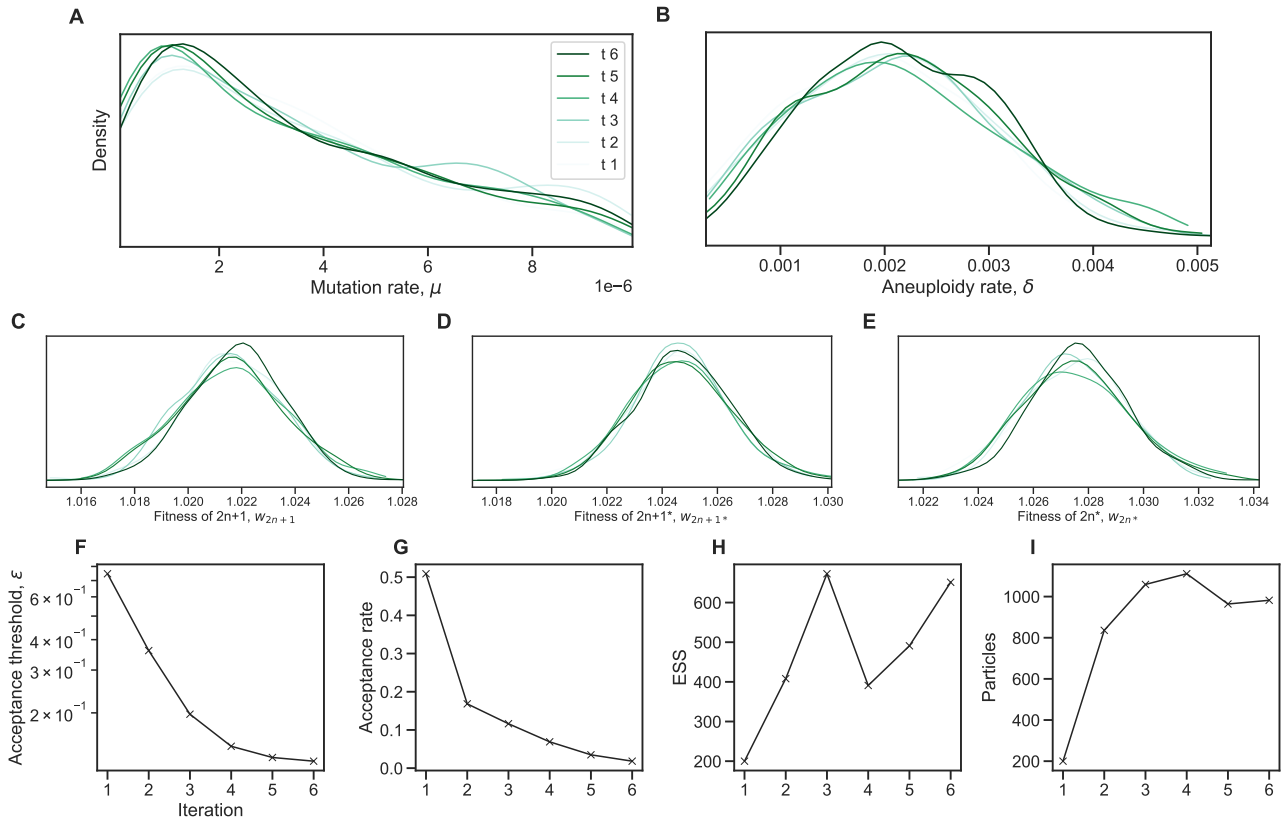

Figure S3: Single-locus model inference convergence. The ABC-SMC algorithm was used to infer the parameters of the single-locus model. **(A-E)** The approximate posterior distributions of model parameters at each iteration of the ABC-SMC algorithm demonstrates convergence, as the posterior did not significantly change after the first iteration, $t = 1$. **(F-I)** ABC-SMC measures of convergence. After iteration number 6, the acceptance threshold was $\epsilon = 0.13$, the acceptance rate was 0.018, the number of particles was 982, and the effective sample size ESS=651.

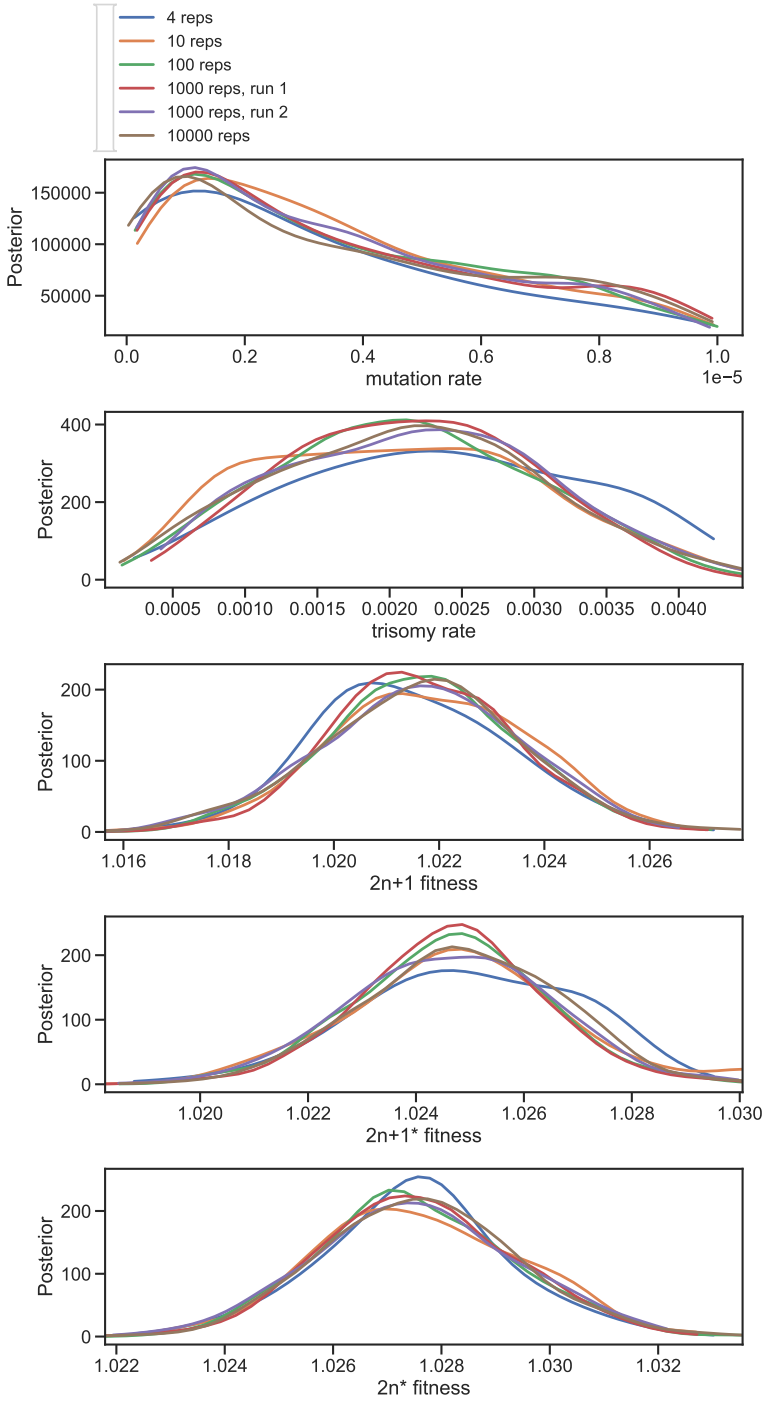

Figure S4: Posterior distribution validation. The posterior distributions of the single-locus model parameters are roughly the same regardless of the number of simulations (10-1,000) used to approximate the likelihood.