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

June 8, 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 an evolutionary-genetic model that includes the effects of natural selection, genetic drift, aneuploidy, and mutation to examine the role of aneuploidy in adaptive evolution. This model follows a population of cells characterised by both their ploidy and their genotype. We fit this model 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 test hypotheses about the evolutionary process by performing model selection between different versions of the model. We also mathematically analyze this evolutionary-genetic model to estimate the effects of selection, mutation, and aneuploidy on the adaptation time and the probability for adaptation via aneuploidy.

Models and Methods

Evolutionary Models. We model the evolution of a population of cells using a Wright-Fisher model (**?**), assuming a constant effective population size N, non-overlapping generations, and in-

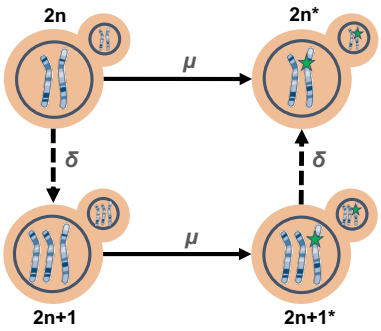

Figure 1: Model illustration. There are four genotypes in our model: euploid wild-type, $2n$; euploid mutant, $2n$*; aneuploid wild-type, $2n+1$; and aneuploid mutant, $2n+1$*. Overall there are two possible trajectories from $2n$ to $2n$*. Arrows denote transitions between genotypes, with transitions rates: μ, beneficial mutation rate; δ, aneuploidy rate.

cluding the effects of natural selection, genetic drift, aneuploidy, and mutation. We focus on beneficial genetic modifications, neglecting the effects of deleterious and neutral mutations or karyotypic changes. The model allows for a single aneuploid karyotype (e.g., chromosome III duplication) and a single mutation to accumulate in the genotype. Thus, the model follows four genotypes (Figure 1): 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 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 $\mathbf{f^m} = (f^m_{2n}\ ,\ f^m_{2n+1}\ ,\ f^m_{2n+1^*}\ ,\ f^m_{2n^*})$ 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 $f_i \rightarrow f_i'$.

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 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,

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

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, $P_{\tilde{\mathbf{X}}}\big(!\{T(2n^*) < 1700\} \mid 200 \le T(2n+1) \le 450\big)$ 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.

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 (eq. (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 (eq. (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 (**?**),

$$WAIC(\theta) \;=\; -2\log \mathbb{E}[\mathcal{L}(\theta)] \;+\; 2\mathbb{V}[\log \mathcal{L}(\theta)] \tag{5}$$

where θ is a parameter vector, and $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ are the expectation and variance taken over the posterior distribution, which in practice are approximated using 50,000 samples from the posterior KDE. 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} (Figure S1-D). In lieu of a more suitable 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. First, we tried to use additional growth curves of *2n** (*refined* strain from **?**) and *2n+1* in 39 °C to estimate w_{2n*}/w_{2n+1} (Figure S1L). The same prior was used for w_{2n*}/w_{2n+1*}. This prior resulted in WAIC 71.91, compared to 67.87 with the above informative prior. We also tested an uninformative uniform prior with $U(1,6)$, for (i) all w_{2n+1}, w_{2n+1^*}, w_{2n^*}, or (ii) only for w_{2n+1^*}, w_{2n^*}, using the above informative prior for w_{2n+1}. In both cases the inference 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 $\delta \sim U(10^{-6}, 10^{-2})$. A wider mutation rate prior, $\mu \sim U(10^{-9}, 10^{-3})$, produced similar results.

Results

Statistical inference

Parameter estimation. We used ABC-SMC to infer the posterior distribution of model parameters (Figure 2). We report parameter estimates using the MAP (maximum a posteriori) and providing the 50% HDI (highest density interval) in square brackets. The estimated aneuploidy rate, $\delta = 1.722 \cdot 10^{-3}$ $[1.394 \cdot 10^{-3} - 2.754 \cdot 10^{-3}]$, agrees with previous estimates. The estimated mutation rate, $\mu = 2.942 \cdot 10^{-6}$ $[2.017 \cdot 10^{-7} - 4.15 \cdot 10^{-6}]$, corresponds to a mutation target size of 10^4, assuming the mutation rate per base pair is roughly $2 \cdot 10^{-10}$ (**?**) or $3.3 \cdot 10^{-10}$ (**?**). The estimated fitness values are $w_{2n+1} = 1.022$ $[1.021 - 1.023]$, $w_{2n+1^*} = 1.025$ $[1.024 - 1.026]$, $w_{2n^*} = 1.028$ $[1.026 - 1.029]$, all relative to the fitness of *2n*, which is set to $w_{2n} = 1$. Thus, we can infer that the benefit and cost of trisomy was $b = 2.5\%$ and $c = 0.3\%$, and the benefit of the beneficial mutation was 2.8% (Table 1).

Model checking and comparison. The model fits the data well: in simulations using the MAP parameter estimates, *2n** fixed in 61% of simulations by generation 1,700 and in 100% of simulations by generation 2,350 (Figure 3B). Interestingly, the genotype frequency dynamics in these simulations demonstrate that *2n+1** never reaches substantial frequency (Figure 4).

We also inferred model parameters under the assumption that the mutation rate increases in aneuploid cells by a factor $\tau = 33/32$ (due to an additional chromosome), 2, 5, 10, or 100 (due to genetic instability). We found that the posterior distribution was similar for τ =1, 33/32, 2, and 5 (Figure S5). With $\tau = 100$, the estimated mutation rate was about 7-8-fold lower compared to $\tau = 1$ ($\mu = 3.81 \cdot 10^{-7}$ $[1.508 \cdot 10^{-7} - 4.995 \cdot 10^{-7}]$) and the aneuploidy rate was about 2-3-fold lower ($\delta = 0.782 \cdot 10^{-3}$ $[0.661 \cdot 10^{-3} - 1.462 \cdot 10^{-3}]$). With $\tau = 10$, the estimated mutation rate was only slightly lower compared to $\tau = 1$ ($\mu = 1.674 \cdot 10^{-6}$ $[2.501 \cdot 10^{-7} - 1.741 \cdot 10^{-6}]$). WAIC is lowest for $\tau = 100$ (Table S1), and if we rule such a strong effect of genetic instability a-priori, the next two lowest WAIC values are for $\tau = 33/32$ and $\tau = 5$. Therefore, we cannot rule out an increased mutation rate in aneuploid cells, but unless the effect is strong ($\tau = 100$), it does not seem to affect our inference results. We also checked the differences in fixation/loss times of *2n*, *2n+1*and *2n** for the models with different τ. We saw that the models can be distinguished using these parameters, especially $\tau = 100$ (Figure 5).

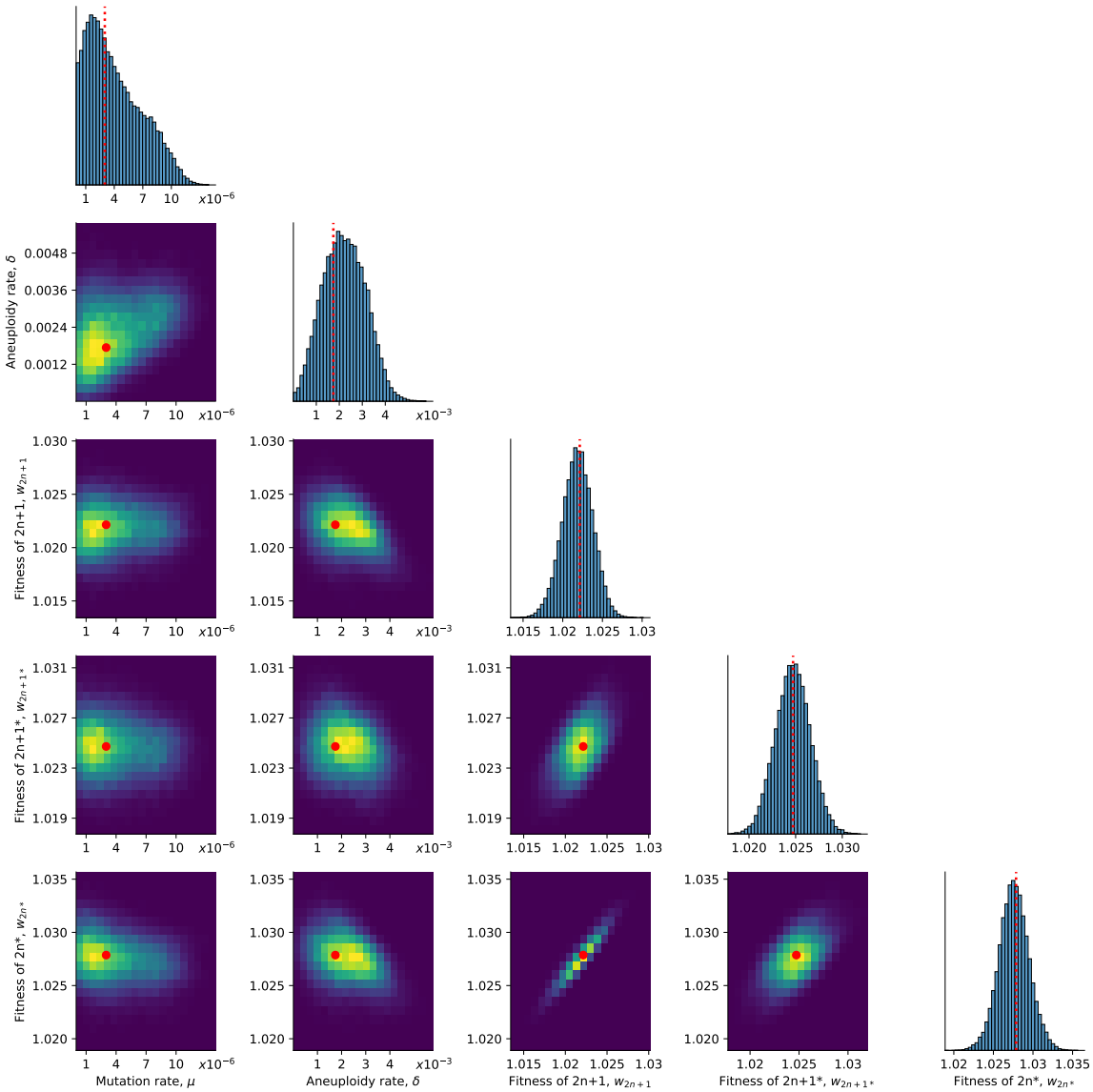

Figure 2: Posterior distribution of model parameters. On the diagonal, the inferred posterior distribution of each model parameter. Below the diagonal, the inferred joint posterior distribution of pairs of model parameters (dark purple and bright yellow for low and high density, respectively). Red markers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

Sensitivity analysis shows that changing a single parameter while keeping the rest fixed at the MAP estimate produces a worse fit to the data (Figure S2). Furthermore, we fitted models with a mutation rate fixed at μ=10^{-5}, 10^{-6} and 10^{-7}. We inferred similar parameters estimates compared to the model with a free μ parameter. WAIC was lower when μ is fixed (Table S1), but this is not surprising, as WAIC attempt to balance between model fit and model complexity, where the latter takes into account the number of model parameters.

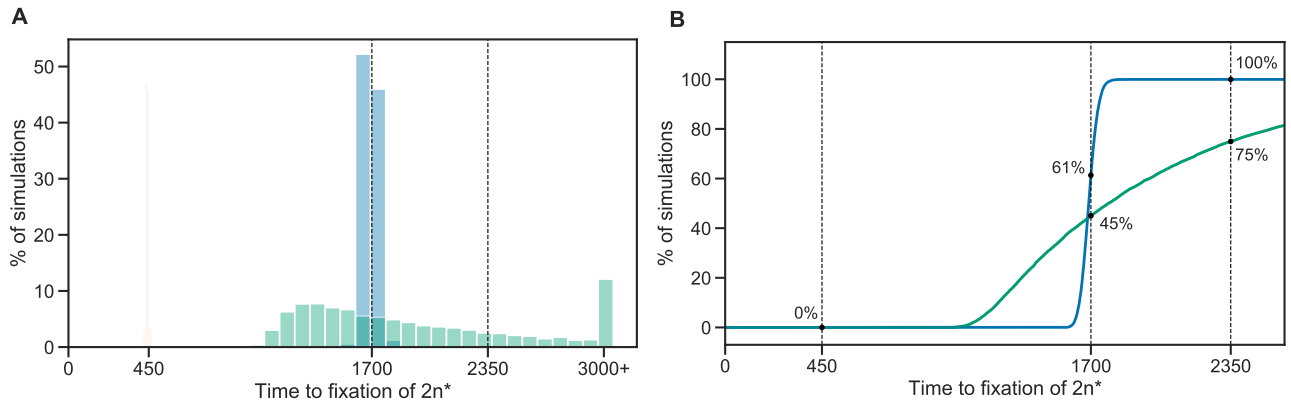

Figure 3: Single-locus model fit with and without aneuploidy. (A) The distribution of time to fixation of $2n*$ (i.e., adaptation time) in 10, 000 simulations of the single-locus model with aneuploidy (blue; MAP parameters) compared to two models without aneuploidy: a model with the same parameter values 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 (eq. (4)) is 0.84 and 0.67 with and without aneuploidy, respectively.

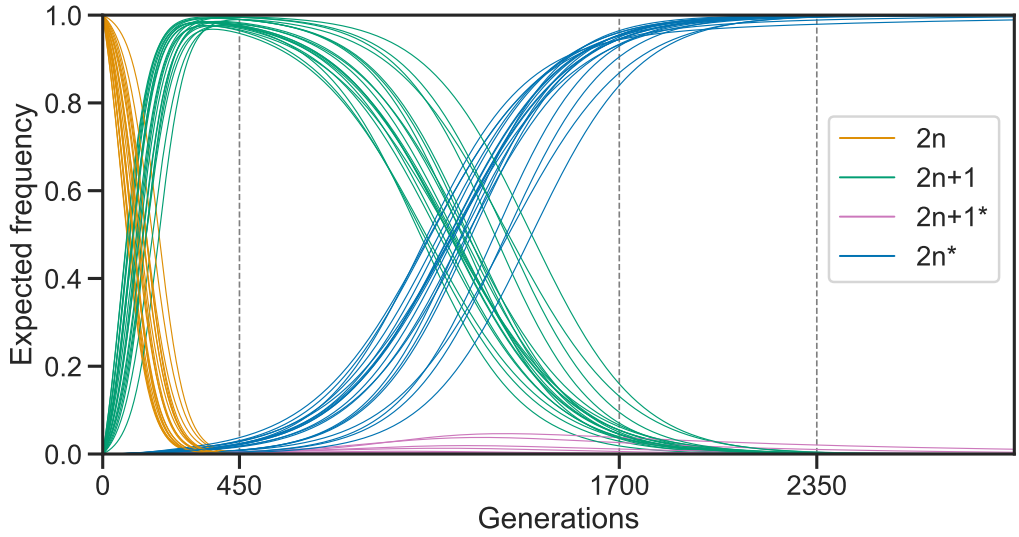

Figure 4: Posterior genotype frequency dynamics for the single-locus model. The posterior prediction for the frequencies of the four genotypes over time. Each of the 20 curves is the average of 10,000 simulations of the single-locus model using parameters drawn from the posterior distribution.

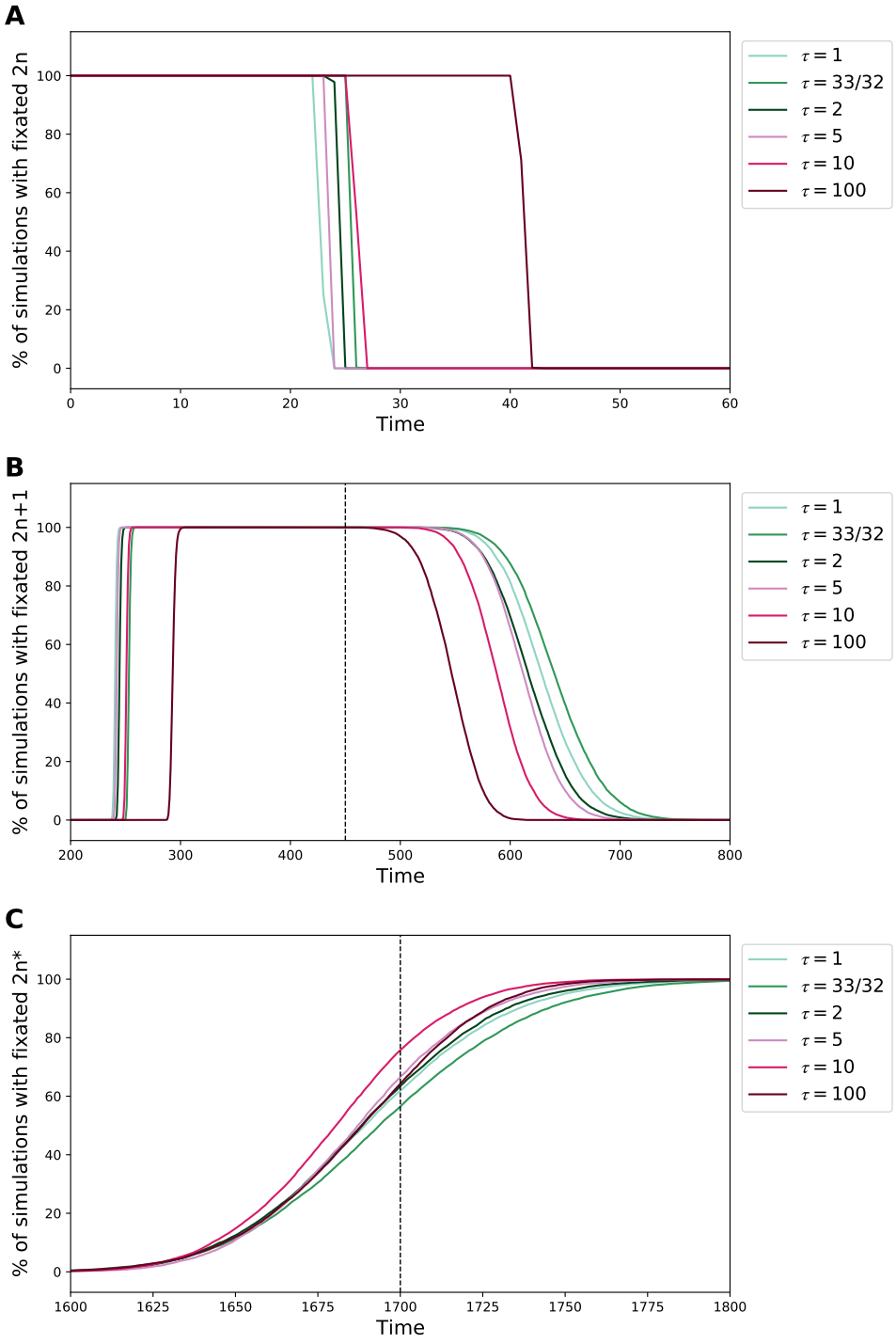

Figure 5: Genotype fixations for models with different τ. We estimated the parameters for different models, each assuming a different value of τ, the fold-increase in mutation rate in aneuploid cells. We then generated 10,000 simulations using the MAP estimate of each model and evaluated the fraction of simulations in which the genotype *2n* (**A**), *2n+1* (**B**), and *2n** (**C**) is fixed (y-axis) at each generation (x-axis). Note that *2n+1** did not fix. We can see that $\tau = 100$ can be distinguished if time to loss of *2n* is known (panel A) or if time to fixation or loss of *2n+1* is known (panel B). It is harder to distinguish between $1 \leq \tau \leq 10$.

Analysis

Approximation. We aim to approximate the adaptation time (i.e., time to fixation of *2n**) and analyse how it is affected by the model parameters. There are two genotypes events that can occur on the *2n* genotype background. The first is a mutation that produces a *2n** genotype. The second is a mis-segregation event that produces a *2n+1* genotype. The rate at which these genotypes appear is significantly different. We assume that if one genotype appears (e.g., *2n**), then it will genotype rapidly fix in the population and therefore the other genotype (e.g., *2n+1*) is unlikely to occur. The expected waiting time until one of the two genotype appears and then leads to adaptation is therefore

$$T = p_a T_a + (1 - p_a) T_m \ , \tag{6}$$

where p_a is the probability that *2n+1* is the first to appear without going to extinction, T_m is the expected time for appearance and fixation of *2n** given it appears before *2n+1*, and T_a is the expected time for appearance of *2n+1*, followed by a mutation to *2n+1**, and loss of aneuploidy to *2n**, given that *2n+1* appears before *2n**. Note that we assume the probability that *2n** is the first to appear without going to extinction is $1 - p_a$.

We assume the waiting time $T_{i,j}$ for appearance of genotype j in a population of genotype i is geometrically distributed with rate $p_{i,j}$, such that

$$p_{2n,2n^*} = 1 - \left(1 - 2\frac{w_{2n^*} - w_{2n}}{w_{2n}}\mu\right)^N , \tag{7a}$$

$$p_{2n,2n+1} = 1 - \left(1 - 2\frac{w_{2n+1} - w_{2n}}{w_{2n}}\delta\right)^N \tag{7b}$$

$$p_{2n+1,2n+1^*} = 1 - \left(1 - 2\frac{w_{2n+1^*} - w_{2n+1}}{w_{2n+1}}\mu\right)^N \tag{7c}$$

$$p_{2n+1^*,2n^*} = 1 - \left(1 - 2\frac{w_{2n^*} - w_{2n+1^*}}{w_{2n+1^*}}\delta\right)^N . \tag{7d}$$

However, if the mis-segragation supply is high enough, $\delta N > 1$, then *2n** may appear before *2n+1** is fixed (Figure 4), and the latter equation is

$$p_{2n+1^*,2n^*} = 1 - \left(1 - 2\frac{w_{2n^*} - w_{2n+1}}{w_{2n+1}}\delta\right)^{zN} , \tag{8}$$

where $z \ll 1$ is the frequency of *2n+1** when *2n** first appears (note that here we use w_{2n+1} rather than w_{2n+1^*}).

We assume that if *2n+1* appears in a population of *2n* and does not go to extinction then *2n** does not appear, and vice versa. Thus, we need to find the minimum of the waiting times for appearance of these genotypes. The minimum of two geometric random variables with rates $p_{2n,2n^*}$ and $p_{2n,2n+1}$ is also a geometric random variable with rate $p_{min} = 1 - (1 - p_{2n,2n^*})(1 - p_{2n,2n+1})$, which is the

inverse of the expected waiting time until the first event occurs, $T_{min} = 1/p_{min}$. Given that one of these genotypes appeared, the probability that it was *2n+1* rather than *2n** is $p_a = \frac{p_{2n,2n+1}}{p_{2n,2n^*}+p_{2n,2n+1}}$.

Given a mutation occurs first, the expected adaptation time via the mutation trajectory is

$$T_m = T_{min} + \tau_{2n,2n^*} \ , \tag{9}$$

where $\tau_{i,j}$ is the waiting time for fixation of genotype j in a population of genotype i, given it starts with a single copy and does not go to extinction by drift. Given that aneuploidy occurs first, the expected adaptation time via the aneuploidy trajectory is approximated by

$$T_a = T_{min} + T_{2n+1,2n+1^*} + T_{2n+1^*,2n^*} + \tau_{2n,2n+1} + \tau_{2n+1,2n+1^*} + \tau_{2n+1^*,2n^*} \ . \tag{10}$$

We assume that the waiting time until a new genotype occurs, $T_{i,j}$, is much longer than the waiting time for its fixation, $\tau_{i,j}$. Nevertheless, these fixation times must be included in a rigorous analysis. There are two ways to compute the fixation times. A deterministic-process approach is given by the classical bi-allelic single-locus equation. It approximates the expected time for genotype j to reach frequency x when starting from frequency y in a population of genotype i by

$$\tau_{i,j}(x,y) \approx \frac{\log\left(\frac{x(1-y)}{(1-x)y}\right)}{\log\left(\frac{w_j}{w_i}\right)} \ , \tag{11}$$

and we use $\tau_{i,j} = \tau_{i,j}(1-1/N, 1/N) \approx 2\log(N)/\log(w_j/w_i)$, except for $\tau_{i,2n^*} = \tau_{i,2n^*}(0.95, 1/N)$ because we stop our simulations when genotype *2n** reaches frequency of 0.95. A stochastic-process approach for approximating $T_{i,j}$ is given by the diffusion-equation approximation (**?**, eq. 17).

Validation. To validate our approach we performed simulations of the full model (eqs. (1) to (3)) without aneuploidy, i.e., $\delta = 0$, and with the other parameters as inferred above. The approximation for the expected adaptation time T (eq. (6)) has a good fit to simulation results except when the effective population size is above $5 \cdot 10^6$ cells (Figure 6A). In that case, the expected number of new mutants per generation is roughly 0.5, and the fixation event is likely to start with more than a single *2n** cell appearing in the same generation, which violates our assumption that a fixation event starts from a single mutant cell (i.e., $y = 1/N$ in eq. (11)).

Next, we performed simulations without mutations from *2n* to *2n** but with mutations from *2n+1* to *2n** (Figure 7). Here, aneuploidy is transient and fixation of *2n+1* or *2n+1** could be outpaced by the arrival of the next genotype (*2n+1** and *2n**, respectively). This would explain the discrepancy between the approximations and simulation results in Figure 7. To asses this further, Figure 8 provides

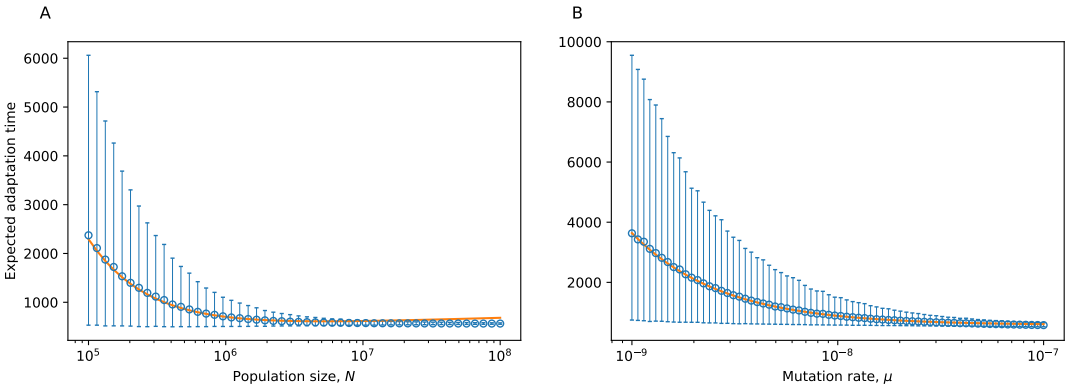

Figure 6: Expected adaptation time without aneuploidy. The expected time to appearance and fixation of $2n*$ as a function of the population size, N (**A**) or as a function of the mutation rate, μ (**B**). Mean and 95% CI from simulations in blue. Analytic approximation (eq. (6)) in orange. Here, $w_{2n+1} = 1.021$, $w_{2n+1*} = 1.025$ and $w_{2n*} = 1.028$. In panel A, $\mu = 9.6 \cdot 10^{-8}$. In panel B, $N = 6 \cdot 10^6$.

a clear picture. At least for intermediate to high values of the respective range of the parameters, the waiting time for appearance, $T_{i,j}$ is far shorter than the time to fixation, $\tau_{i,j}$. Indeed, if we ignore the fixation time of *2n+1* and *2n+1** ($T_{2n,2n+1}$ and $T_{2n+1,2n+1^*}$) in eq. (10), we get a better fit to the data (compare red curves and blue markers in Figure 7). Furthermore, setting $x = 0.017$ for $T_{2n+1^*,2n^*}$ in eq. (11), we get an even better fit in some cases (Figure 7B), but not others (Figure 7A,C). However, we stop simulations when genotype $2n*$ reaches frequency of $0.95 > 0.017$.

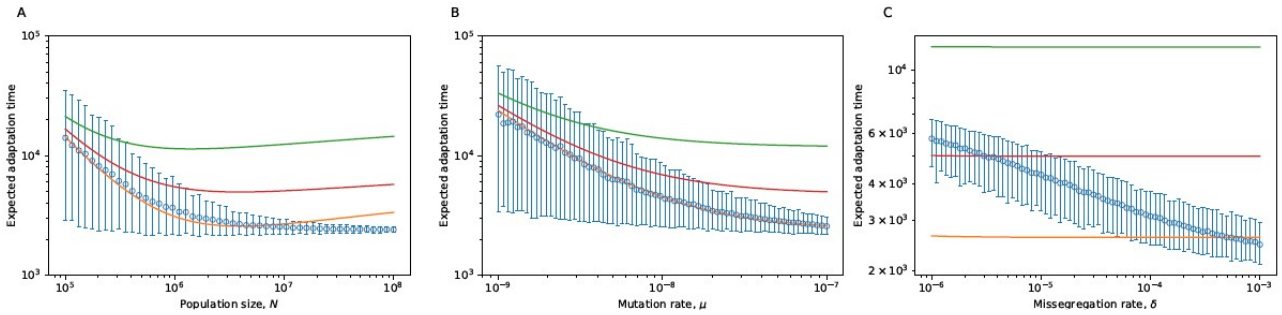

Figure 7: Expected adaptation time without mutations from *2n* to *2n*.* The expected time to appearance and fixation of $2n*$ as a function of the population size, N (**A**) and as a function of the mutation rate, μ (**B**) and as a function of the aneuploidy rate, δ (C). Blue error bars for mean and 95% CI from simulations. Green line for the analytic approximation (eq. (6)). Red curve for the approximation with $T_{2n,2n+1} = T_{2n+1,2n+1*} = 0$. Orange curve for the approximation with $x = 0.017$ in eq. (11) for the fixation time of $2n*$; we found this gives a good fit in panel B. Here, $w_{2n+1} = 1.021$, $w_{2n+1*} = 1.025$ and $w_{2n*} = 1.028$, $N = 6 \cdot 10^6$, $\mu = 9.6 \cdot 10^{-8}$ and $\delta = 5.4 \cdot 10^{-4}$ unless parameter is varied on the x-axis.

In summary, whereas the two possible trajectories to $2n*$ do not compete against each other in both Figure 6 and Figure 7, the approximations are doing well only in restricted parameter ranges. These

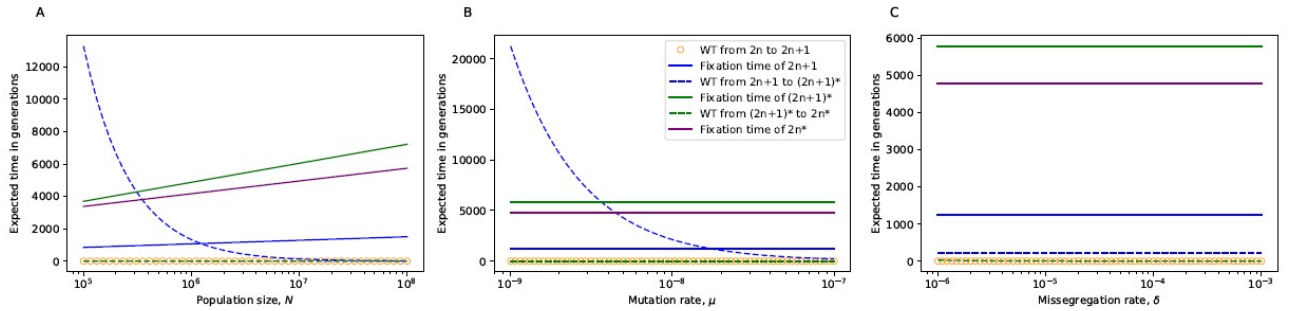

Figure 8: Expected time of fixation and waiting time events on the aneuploid trajectory. The expected time to appearance and fixation of each step seperately in the aneuploid trajectory as a function of the population size, N (panel A) and as a function of the mutation rate, μ (panel B) and as a function of the aneuploidy rate, δ (panel C). Solid curves give the fixation time and dashed curves the waiting time. Curves of the same color indicate the same base population state; e.g. the blue dashed curve is the waiting time in state *2n+1* until a succesful individual of type *2n+1** appears. The corresponding solid curve indicates the time to fixation of the population in state *2n+1*. Here, $w_{2n+1} = 1.021$, $w_{2n+1^*} = 1.025$ and $w_{2n^*} = 1.028$. Parameter values in panels where it is not varied, $N = 6 \cdot 10^6$, $\mu = 9.6 \cdot 10^{-8}$ and $\delta = 5.4 \cdot 10^{-4}$.

tend to be, where the respective parameters are rather small. If they grow bigger, they are violating some assumptions of our approximations.

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

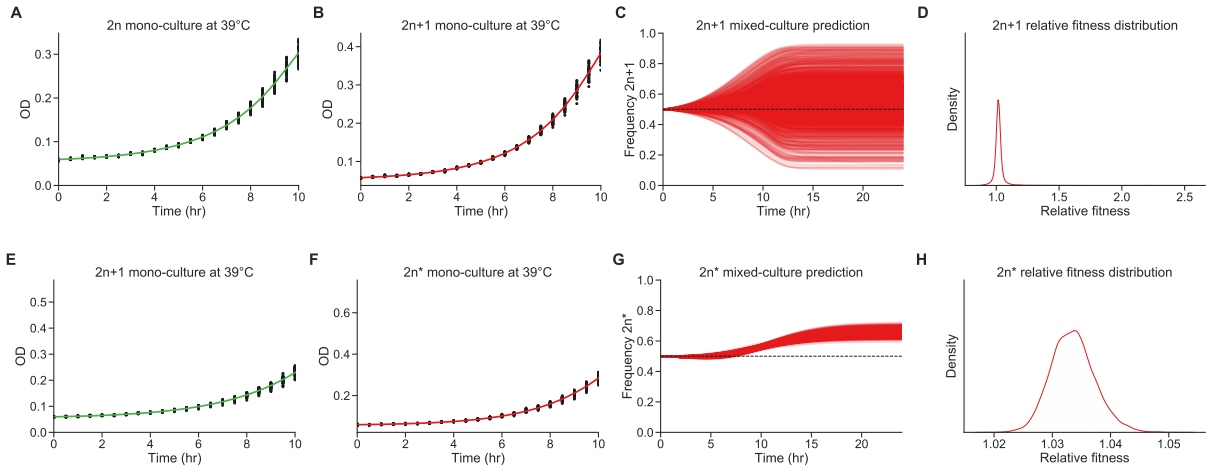

Figure S1: Fitness estimation from growth curves. **(A-D)** Fitness estimation from growth curves of *2n* and *2n+1* at 39 °C. $\hat{w}_{2n+1}/w_{2n}$=1.024 (95% CI: 0.959 - 1.115). **(E-H)** Fitness estimation from growth curves of *2n+1* and *2n** at 39 °C. $\hat{w}_{2n*}/w_{2n+1}$=1.033 (95% CI: 1.027 - 1.041). 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.

Table S1: WAIC values for various model specifications.

id	Model	WAIC
1	Without aneuploidy	0.83
2	Fixed mutation rate, $\mu = 10^{-5}$	0.60
3	Fixed mutation rate, $\mu = 10^{-6}$	33.89
4	Fixed mutation rate, $\mu = 10^{-7}$	29.58
5	Free mutation rate, $\tau = 1$	78.52
6	Free mutation rate, $\tau = 33/32$	65.92
7	Free mutation rate, $\tau = 2$	85.32
8	Free mutation rate, $\tau = 5$	71.89
9	Free mutation rate, $\tau = 10$	77.38
10	Free mutation rate, $\tau = 100$	57.67

TODO caption

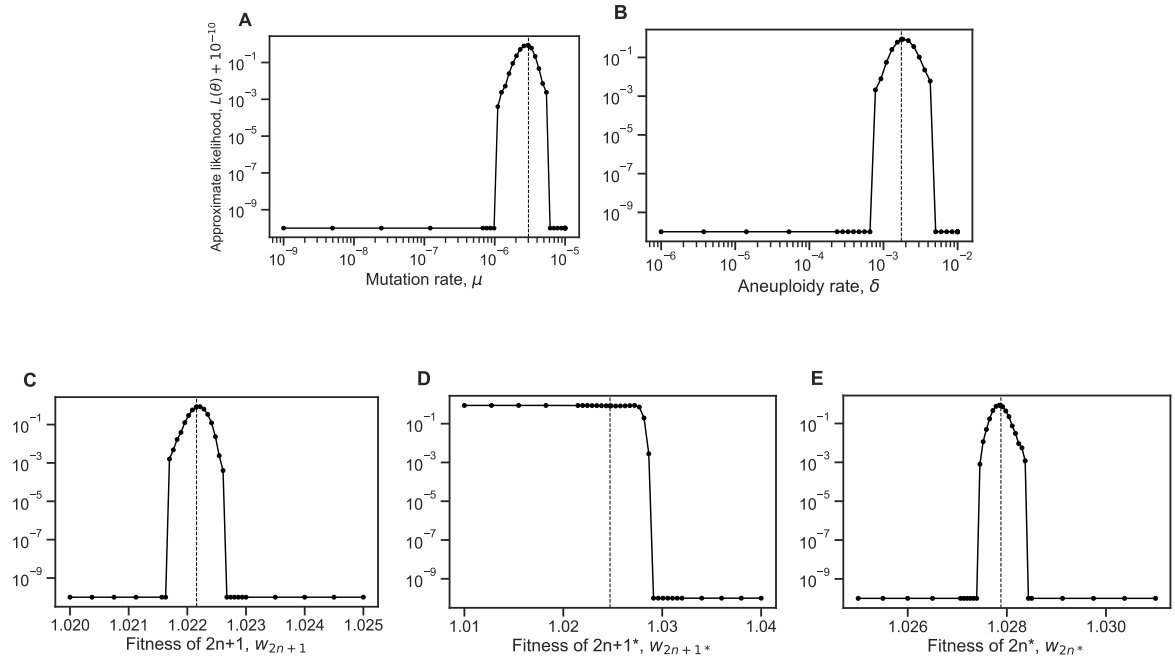

Figure S2: Likelihood profiles. Sensitivity of the model approximate likelihood, $\mathcal{L}(\theta)$, to changing a single parameter while the other parameters remain fixed at their MAP estimates. Dashed vertical line represents the MAP value. The prior distributions for the mutation rate and aneuploidy rate are $\mu \sim U(10^{-9}, 10^{-5})$ and $\delta \sim U(10^{-6}, 10^{-2})$, respectively.

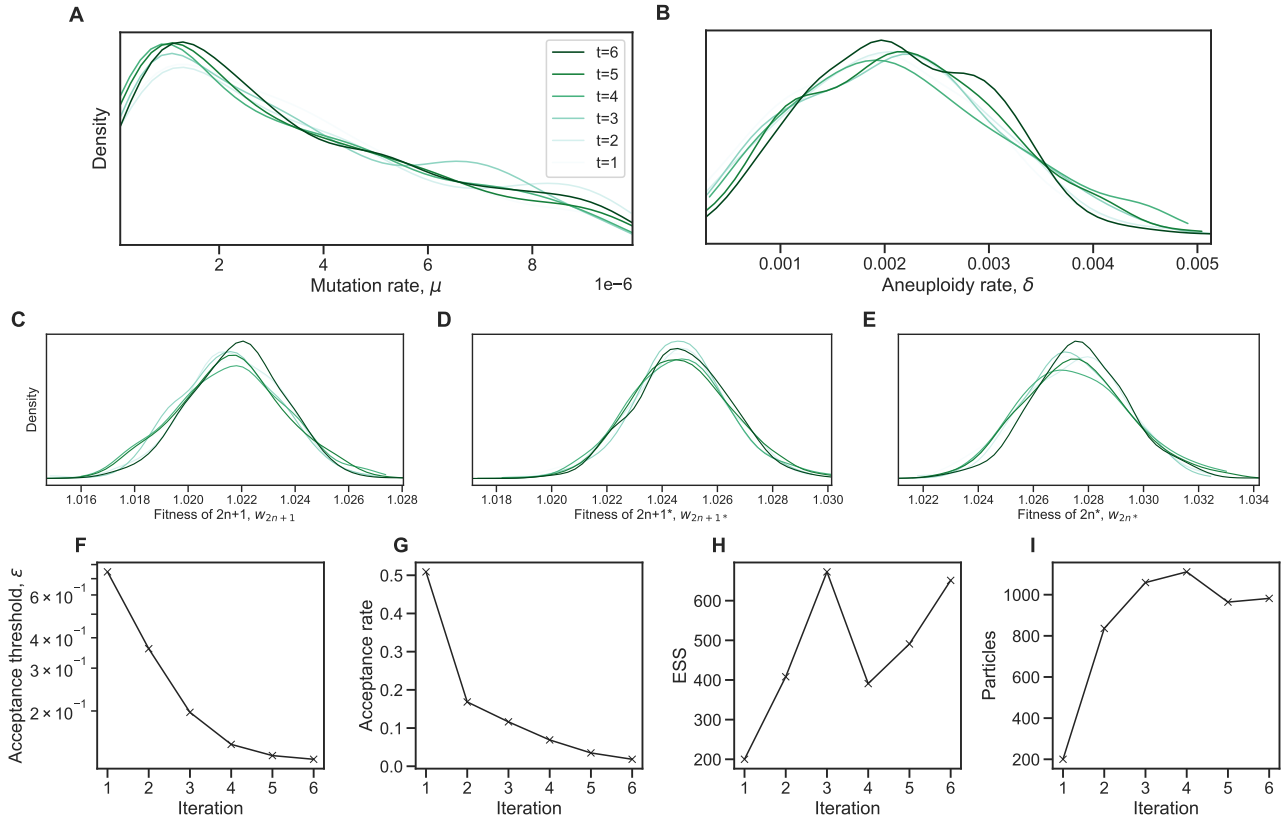

Figure S3: Inference convergence. The ABC-SMC algorithm was used to infer the model parameters. **(A-E)** The approximate posterior distributions of model parameters at each iteration of the ABC-SMC algorithm demonstrates convergence, as the posterior did not significantly change after the first iteration, $t = 1$. **(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.

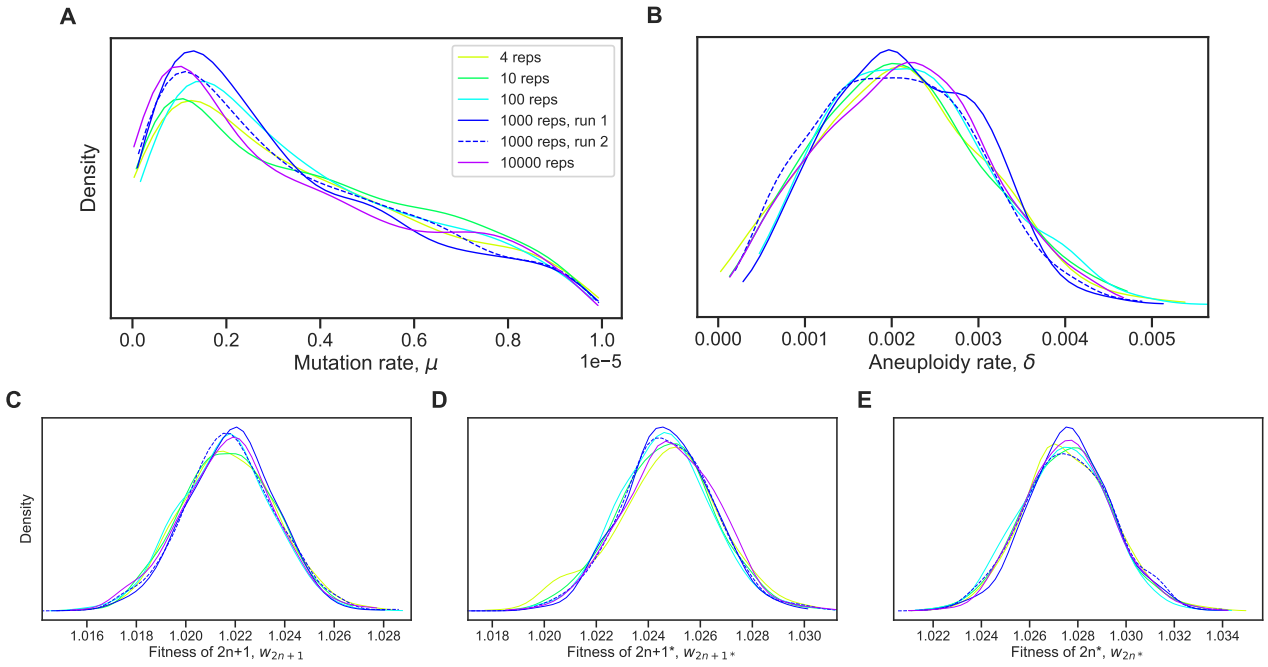

Figure S4: Posterior distribution validation. The posterior distribution of model parameters is roughly the same regardless of the number of simulations (4-10,000 replicates) used to approximate the likelihood (eq. (4)).

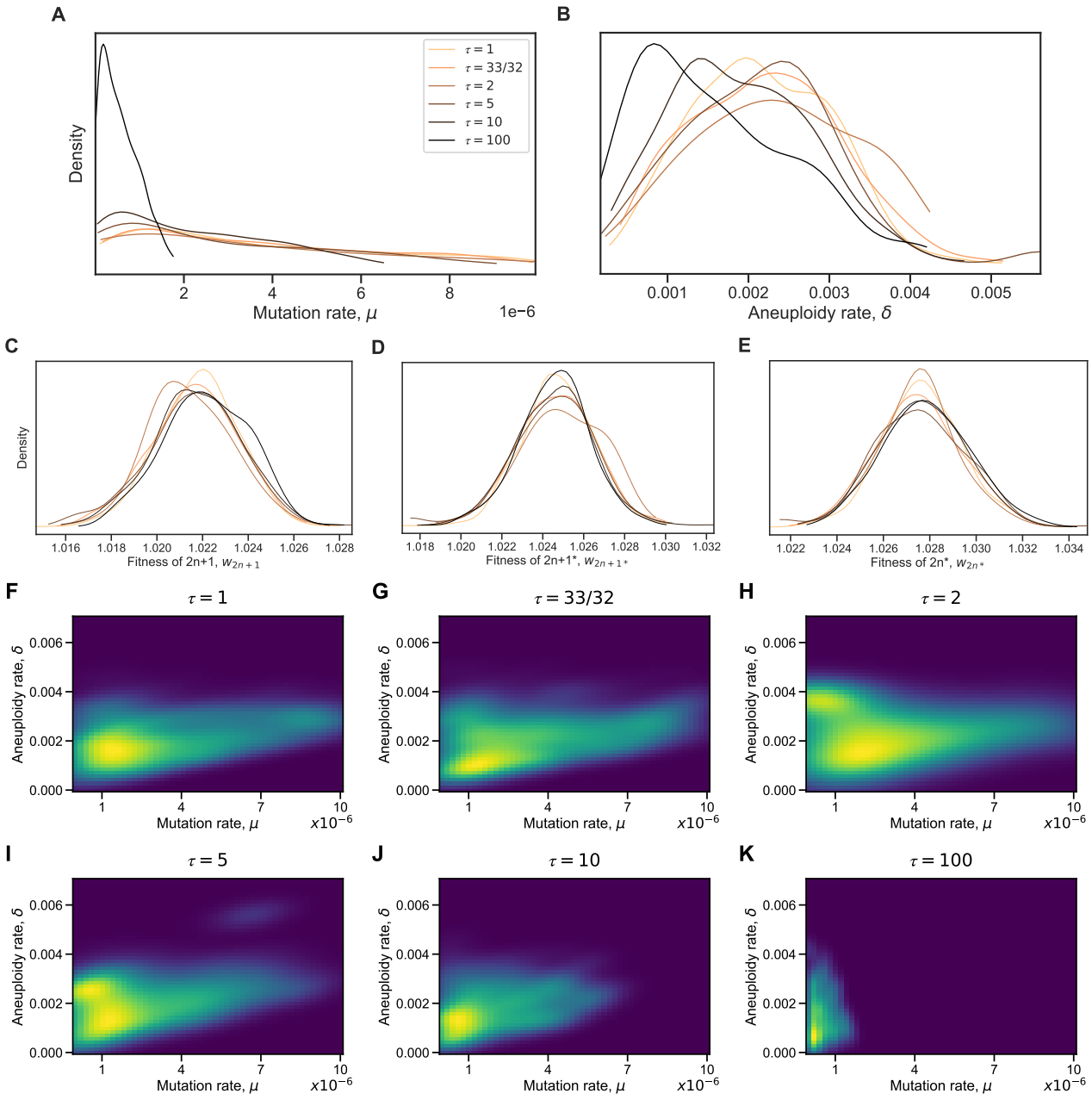

Figure S5: Single-locus model inference with elevated mutation rate in aneuploid cells. (A-E) The inferred posterior distributions of model parameters for the single-locus model with different values of τ, the fold-increase in mutation rate in aneuploid cells (*2n+1* and *2n+1**). When the increase in mutation rate is high, $\tau = 10$ and $\tau = 100$, the inferred mutation (A) and aneuploidy (B) rate tends to be lower. **(F-K)** The inferred joint posterior distribution of mutation rate (μ) and aneuploidy rate (δ) with different τ values (dark purple and bright yellow for low and high density, respectively).