Contents

1 Introduction **2**

2 Parameter estimation from growth curves **3**
2.1 Replication rates estimation from growth curves . . . 3
2.2 Solution to the model equations . . . 4
2.3 Parameter inference . . . 6
2.3.1 Bayesian Inference for the two state model . . . 7
2.3.2 Bayesian Inference for the three state model . . . 8
2.4 Parameter space exploration . . . 10

3 Methods **12**
3.1 Determining convergence of the experiment . . . 12
3.2 Pending tasks . . . 13

1 Introduction

To describe the evolution experiment, we proposed the mathematical model below based on the population of each type of bacterial strain. It is important to clarify that this models is a simplification and by no means it is assumed to be complete, in the sense that there can be additional mutations that are not yet included. Nonetheless, it is expressive enough to capture two big types of mutations which are relevant for the context of this experiment

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \left(r_F\left(1 - \frac{\mu_{F\to D}}{\ln 2} - \frac{\mu_{F\to S}}{\ln 2}\right)F + r_D\frac{\mu_{D\to F}}{\ln 2}D + r_S\frac{\mu_{S\to F}}{\ln 2}S\right)\left(1 - \frac{F+D+S}{K}\right) \tag{1a}$$

$$\frac{\mathrm{d}D}{\mathrm{d}t} = \left(r_F\frac{\mu_{F\to D}}{\ln 2}F + r_D\left(1 - \frac{\mu_{D\to F}}{\ln 2}\right)D\right)\left(1 - \frac{F+D+S}{K}\right) \tag{1b}$$

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \left(r_F\frac{\mu_{F\to S}}{\ln 2}F + r_S\left(1 - \frac{\mu_{S\to F}}{\ln 2}\right)S\right)\left(1 - \frac{F+D+S}{K}\right) \tag{1c}$$

Where F is the founder strain population, D is the population of bacteria carrying a duplication, and S is the population of bacteria with an additional stabilizing SNP mutation. The model parameters are the replication rates (r_F, r_D, r_S) mutation/loss rates $(\mu_{F\to D}, \mu_{F\to S}, \mu_{D\to F}, \mu_{D\to S})$ and carrying capacity K. These equations assume that there is no direct transition between D and S, and mutations can only happen upon replication. Additionally, the finite resources is introduced through the carrying capacity term in the equations.

An evolution experiment carried out through 100 days provides measurements of the fraction of large and small bacterial colonies at the end of every day. This data is used for the inference of the model's parameters. However, this inference task is not straightforward due to the inability of the experimental data to distinguish between the different types of mutations. Duplications and other mutations arise as a way to stabilize the genome after gene deletion, therefore it is plausible to assume that bacteria carrying mutations will have greater fitness and therefore they will be found in colonies of larger size. This consideration allow us to define a two state model that could be compared directly to the experimental data The two state toy model is described by the following system of equations

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \left(r_F\left(1 - \tilde{\mu}_{FM}\right)F + r_M\tilde{\mu_{MF}}M\right)\left(1 - \frac{F+M}{K}\right) \tag{2a}$$

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \left(r_F\tilde{\mu_{FM}}F + r_M\left(1 - \tilde{\mu}_{MF}\right)M\right)\left(1 - \frac{F+M}{K}\right) \tag{2b}$$

F, M are the populations of founder and mutant strains, each with replication rate r_F, r_M respectively. Mutations gain/loss occur with rate $\tilde{\mu}_{FM} = \frac{\mu_{FM}}{\ln 2}$, $\tilde{\mu}_{MF} = \frac{\mu_{MF}}{\ln 2}$.

It is important to note that even though Eqs 1 and 2 are used to describe the same phenomena, it is not possible to define an invertible linear transformation between the variables and parameters of both; therefore, it is necessary to search for an additional way to relate the set of parameters that could be determined independently for each model

To gain some understanding on the range of the model's parameters, we used values reported in [include reference] as well as parameters inferred from growth curves for individual strains which we detail below. The founder strain is referred to as *delserCGA* and the reference mutant is *M2lop*

Parameter	Value
r_f	4.0603×10^{-2} min^{-1}
r_M	5.4478×10^{-2} min^{-1}
μ_{FM}	4.25×10^{-9} $\frac{\text{mutation}}{\text{generation}}$
K	10^{10}

Table 1: Initial parameters for the toy model

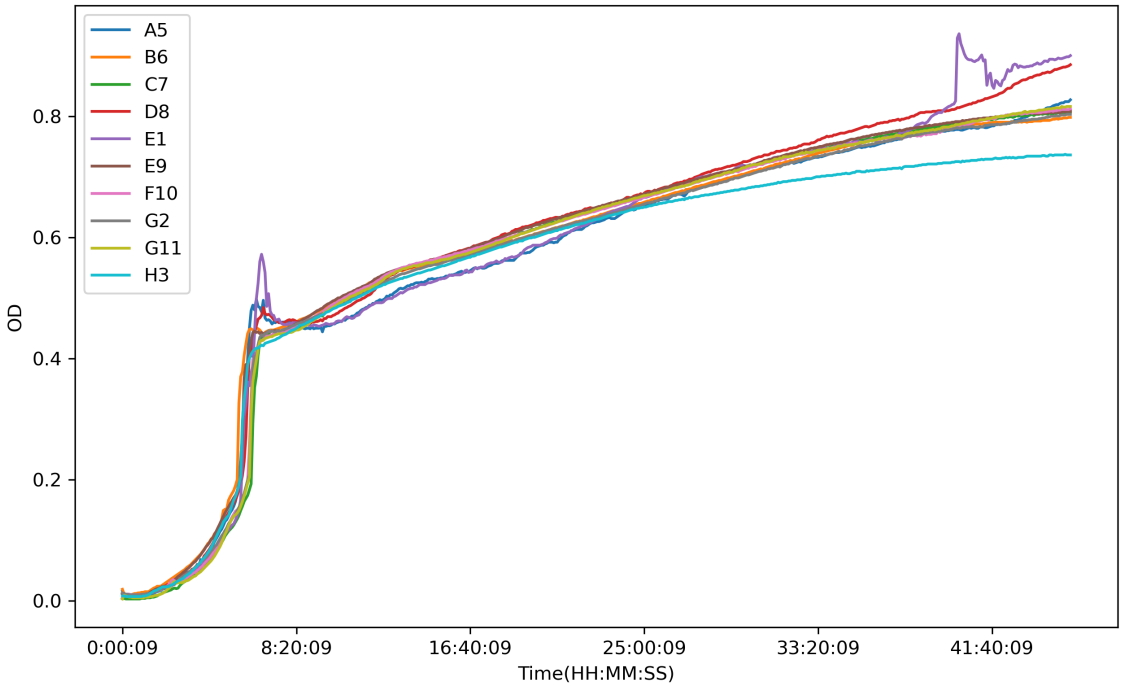

Figure 1: Growth curves for the wild type strain. Each curve represents a position in the 96 well plate

2 Parameter estimation from growth curves

Before numerically solving the system of equations for the toy model, we re-write it in a dimensionless way by defining $\tau = r_F t$ and $\alpha = \frac{r_M}{r_F}$, thus obtaining

$$\frac{\mathrm{d}F}{\mathrm{d}\tau} = \left(\left(1 - \tilde{\mu}_{FM}\right) F + \alpha \tilde{\mu}_{MF} M\right) \left(1 - \frac{F+M}{K}\right) \tag{3a}$$

$$\frac{\mathrm{d}M}{\mathrm{d}\tau} = \left(\tilde{\mu}_{FM} F + \alpha \left(1 - \tilde{\mu}_{MF}\right) M\right) \left(1 - \frac{F+M}{K}\right) \tag{3b}$$

Values for the replication rates can be inferred using data from the growth curves. Figure 1 shows the typical shape of this curves. A distinct feature of these plots is the near linearly-increasing portion that happens after the initial exponential growth, as opposed to the expected plateau due to finite resources. This behaviour has been previously observed [**?**] and it is attributed to the effects of multi-scattering arising in samples with high bacterial concentration. To make sure that we stayed in the single scattering regime where the number of bacteria is proportional to OD values, we choose to work with measurements below an OD threshold of 0.4.

2.1 Replication rates estimation from growth curves

Determining growth rates from OD data is a process trickier than it looks; Ghenu and some of the subtleties are discussed by Ghenu [**?**, **?**].

Based on all the previous considerations, we fit a generalized logistic function (Eq 4) to OD data for each plate's position. The fit was done with Scipy's optimize function, and further verified with lmfit package in python.

$$OD(t) = \frac{A}{\left(C + Q \exp\{-Bx\}\right)^{\frac{1}{\nu}}} \tag{4}$$

It is possible to make a parallel with the parameters in 4 and the Logistic and Richardson models as proposed in [**?**]. Table 2 shows the dependence between parameters of the different models

Logistic	Richards
$\mu = B$	$\mu = -\frac{B}{\nu}$
$K = \frac{K}{K-1}Q$	$K = A$
$N_0 = \frac{1}{k-1}Q$	$N_0 = (Q+1)^{\frac{-1}{B}}$
$\nu = 1$	$\nu = -B$
$C = 1$	$C = 1$

Table 2: Correspondence between the parameters for the Logistic and Richards models in [**?**] and the generalized logistic model (Eq 4) used to fit OD curves

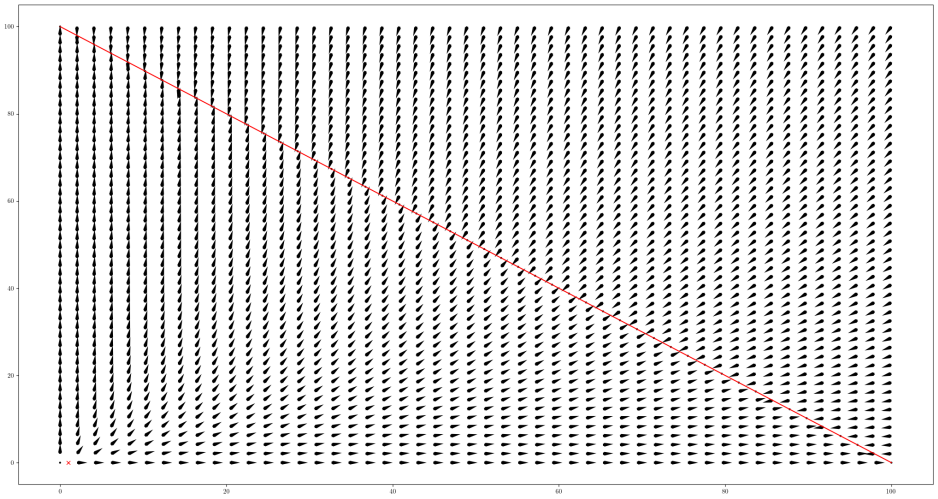

Figure 2: Vector field plot for Eq 3

It is important to point out that this parallel holds only under the assumption that in the single scattering regime, OD values are proportional to the number of cells, N. In general, the relation between OD values and the number of cells is not well studied, it often requires calibration from the equipment used for each bacterial strain [**?**]. Alternative experimental methods for estimating bacterial population from OD values have been suggested [**?**]; however, for the purposes of this project, we are not concerned with this problem directly. It is enough to restrict ourselves to the the single scattering regime, where we assume a linear relation between OD and N.

2.2 Solution to the model equations

So far we haven't found an analytical solution to the equations; thus, we opted to analyze their stability and asymptotic behaviour. The first thing we looked at was the vector field in Fig 2 I found that there are only two equilibrium points, a unstable one at the origin ($F = M = 0$) and a "degenerate attractor" along the line $F = K - M$. This "atractor" arises due to the carrying capacity term in the model. The trajectory way in which the system approaches this "attractor" depends on the initial conditions, as it can be see in Figure 3 at the initial stages of the experiment, when the mutant population is zero, the systems approaches horizontally to the line; as the days pass and the starting mutant population increases, the system approaches the line more diagonally and eventually almost straight vertically. Furthermore, we observed that once the system reaches the line, it doesn't fluctuate and the only way for it to reset is by intervention through the dilutions.

Figure 3 shows the trajectory in the phase space for a set of parameters with an appropriate order of magnitude inferred by comparison with the experimental data. The darker colors represent the first days of the experiment, here the population growth is mostly for the founder. Consecutive dilutions

shift the initial point of the new trajectories, and cause a rapid increase for the mutant population, which after around 15 days takes over the system. Nonetheless, the founder strain does not extinguish and there is a remnant in agreement with the experiment.

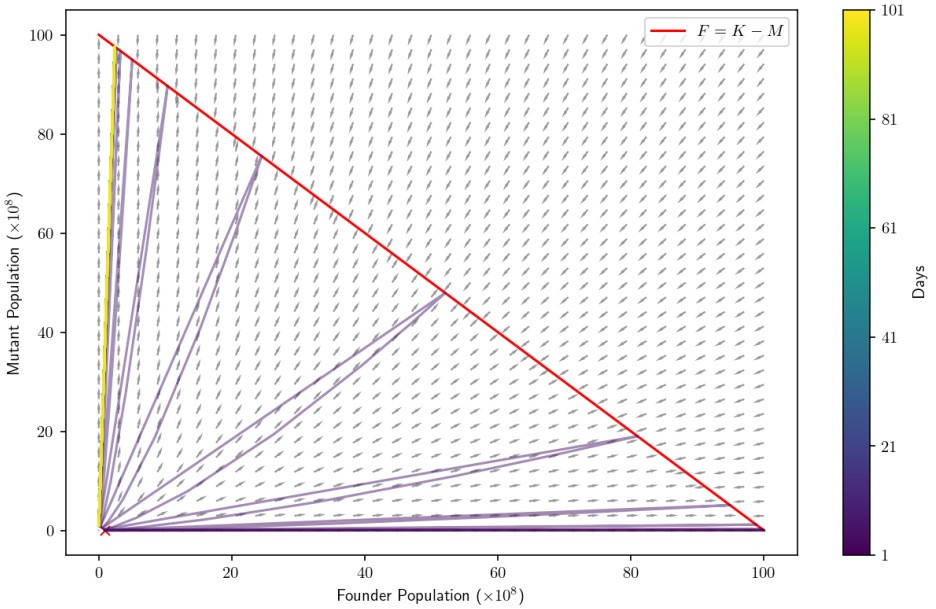

Figure 3: Phase space trajectory for the two state system. The colors indicate the day of the experiment, darker colors indicate the earlier days and brighter the latter ones. The system rapidly goes from the initial state with mostly founders to the final state where there is coexistence between mutant and founder, with a higher abundance of the former.

Now, in Fig 4 I look at the solution for the two state system[1]. The sharp drops in the proportions of founder and mutant are captured, as well as the remnant population in the equilibrium state

[1]Even though we illustrate the behavior with replicate 1, the discussion is still valid for the other replicates as their values almost the same

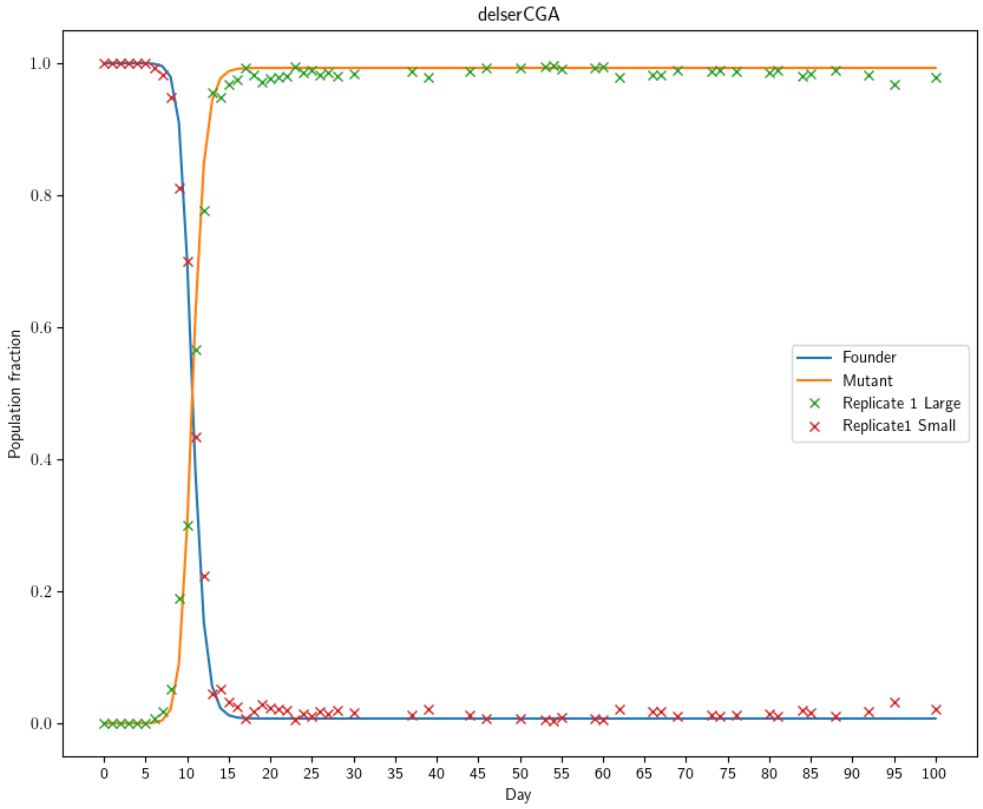

Figure 4: Population fractions for the two state system. The parameters used were $K = 10^2$, $\alpha = 1.34187, \tilde{mu}_{FM} = 0.0039\times10^{-6}, \tilde{mu}_{MF} = 1.26611\times10^{-3}$. The solid lines are the theoretical predictions for the populations through the duration of the experiment. The x indicate the measured values of the populations proportions for small and large bacteria in Replicate 1. Large colonies are interpreted as encompassing all types of mutations, while small bacteria are understood as founder colonies.

The evolution experiment is carried out with the following conditions:

- The carrying capacity is set to $K = 10^{10}$cells
- The estimated initial cell density is $N_0 = 10^9$ cells/mL in 100 μL of solution
- After each day 1% of the solution is used to start the next culture
- The founder strain is the named *delserCGA* and the reference duplication mutant is *M2lop*

To solve Eq 3 we need a proxy for the values of the transition rates $(\tilde{\mu}_{FM}, \tilde{\mu}_{MF})$ and the mutant replication rates

2.3 Parameter inference

To gain some understanding on the behaviour of the solutions to Eq. 3 with respect to the parameters, I initially solved the two state system for different values of $(\tilde{\mu}_{FM}, \tilde{\mu}_{MF})$, the value of α was initially fixed as the growth curve data allowed the determination of the growth rates. This naive search provided some useful observations. I found that varying $\tilde{\mu}_{FM}$ shifts the day at which the final populations of mutant and founder are the same, without noticeably changing the ratio of their populations at the end of the experiment, meaning that this transition rate determines how fast the founder population will disappear. Qualitatively, this makes sense for me, considering that at the beginning of the experiment,

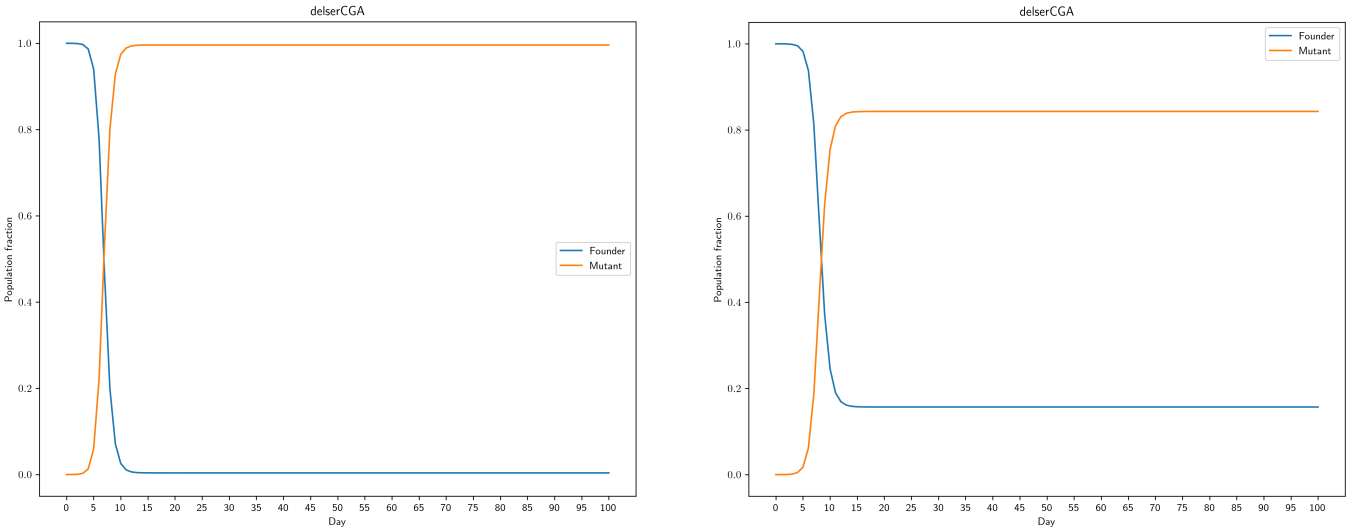

Figure 5: Effect of varying $\tilde{\mu}_{MF}$ on the population fraction behavior in the evolution experiment. Both panels have fixed values for $\tilde{\mu}_{FM}, \alpha$, $\tilde{\mu}_{MF} = 0.001 * \ln 2$ and bottom panel $\tilde{\mu}_{MF} = 0.4 * \ln 2$

there are no mutants, therefore the transition rate from F to M is an important contributor to the quickly appearance of a mutant population

On the other hand, varying $\tilde{\mu}_{MF}$ seems to strongly determine the final fraction of population at the end of the experiment. Figure 5 illustrates this behaviour, for a small $\tilde{\mu}_{MF}$ the ratio of founder in the final state is considerably higher than the ratio of mutants. As $\tilde{\mu}_{MF}$ increases, the final founder ratio increases as well. This might indicate that this parameter is a key determinant for the fate of each strain, and in consequence of the proposed set of hypotheses.

An important observation here is the difference in the order of magnitude for the transition rates, $\tilde{\mu}_{MF}$ has a stronger influence around 10^{-4}, whereas $\tilde{\mu}_{MF}$ does it around 10^{-9}. Nonetheless, this behaviour is similar to the one described in [**?**]

Now, when α is allowed to vary it exhibits a rather interesting behavior. In the limit where $\alpha \to 1$ there is no notorious shift in the population from founder and mutants; they remain close to their initial values, at least in a time frame suitable for an experiment. This behavior completely changes as α increases. As soon as $\alpha \to 2$ the mutant population quickly takes over the system and remains like that for the rest of the experiment.

2.3.1 Bayesian Inference for the two state model

My initial approach to simplify the number of parameters to determine was to assume that the replication rate of the mutant was known, for this I used the value of the reported rate for *M2lop* ($r_M = 0.05448$ replication/min). By playing with the parameters I had some intuition about a range where the solution fit the experimental results (4). I chose to use Affine Invariant MCMC to determine the transition rates. With an uniform prior between 0 and 1 , the posterior distribution fails to reproduce similar experimental measurements.

Given that the initial flavor of MCMC was not successful at fitting the parameters to the experimental data, we switch to an Approximate Bayesian Calculation (ABC). This new approach yield satisfactory results shown in Figure 6. This parameter distribution supports the values encountered previously for α, and the difference between the orders of magnitude of mutation and loss rates is close to the behavior described in [**?**]

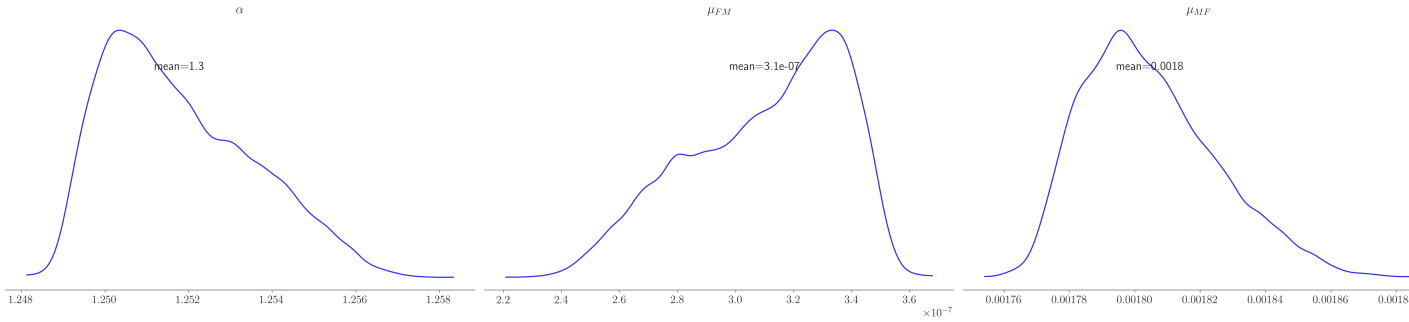

Figure 6: Posterior distribution for the parameters of the two state model inferred with ABC

Figure 7 shows the experimental points along with the theoretical predictions (solid line) obtained by solving the model with the mean of the parameters distribution in Figure 6. As we can see, the predictions reproduce the measured behavior of the populations, capturing the fast increase in mutants and the remnant founder population after 100 days.

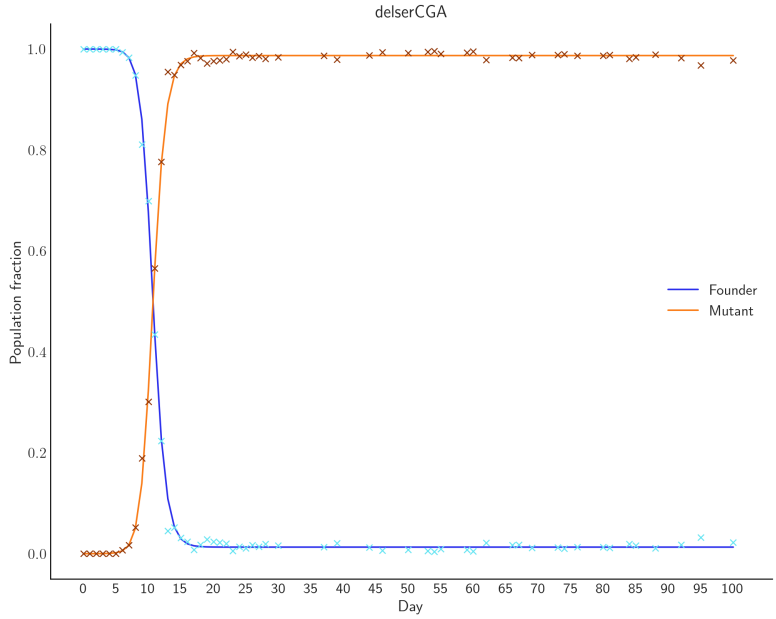

Figure 7: The solid lines represent the theoretical evolution of the population fraction using the two state model. The x are the experimental measurements of the ratios of small and large colonies.

2.3.2 Bayesian Inference for the three state model

Let's start by writing the adimensional expression of 1, where time has been scaled by a factor of r_F

$$\frac{\mathrm{d}F}{\mathrm{d}\tau} = \left(\left(1-\tilde{\mu}_{FD}-\tilde{\mu}_{FS}\right)F + \alpha\tilde{\mu}_{DF}D + \beta\tilde{\mu}_{SF}S\right)\left(1-\frac{F+D+S}{K}\right) \tag{5a}$$

$$\frac{\mathrm{d}D}{\mathrm{d}\tau} = \left(\tilde{\mu}_{FD}F + \alpha\left(1-\tilde{\mu}_{DF}\right)D\right)\left(1-\frac{F+D+S}{K}\right) \tag{5b}$$

$$\frac{\mathrm{d}S}{\mathrm{d}\tau} = \left(\tilde{\mu}_{FS}F + \beta\left(1-\tilde{\mu}_{SF}\right)S\right)\left(1-\frac{F+D+S}{K}\right) \tag{5c}$$

with $\tau = r_F t$, $\alpha = \frac{r_D}{r_F}, \beta = \frac{r_S}{r_F}$, $\tilde{\mu}_{FD} = \frac{\mu_{F\to D}}{\ln 2}, \tilde{\mu}_{FS} = \frac{\mu_{F\to S}}{\ln 2}, \tilde{\mu}_{DF} = \frac{\mu_{D\to F}}{\ln 2}, \tilde{\mu}_{SF} = \frac{\mu_{S\to F}}{\ln 2}$. This model involves more unknown parameters than Eq 3, thus MCMC computation time will be significantly larger as it scales with the dimensionality of the parameter space. Running an Approximate Bayesian Computation algorithm is a more suitable option, given that it doesn't involve calculating a likelihood

function and the implementation chosen is well optimized. From solving Eqs 5 for different combinations of parameters and previous work with the two state model, I already had prior knowledge of at least some orders of magnitude for the parameters. In this way I define the prior probability as

$$\begin{aligned}
&\alpha, \beta \sim \mathcal{U}_{[1,2]} \\
&\tilde{\mu}_{FD} \sim \mathcal{N}(0, 10^{-5}), \quad \tilde{\mu}_{FD} > 0 \\
&\tilde{\mu}_{FS} \sim \mathcal{N}(0, 10^{-5}), \quad \tilde{\mu}_{FS} > 0 \\
&\tilde{\mu}_{DF} \sim \mathcal{N}(0, 10^{-5}), \quad \tilde{\mu}_{DF} > 0 \\
&\tilde{\mu}_{SF} \sim \mathcal{N}(0, 10^{-5}), \quad \tilde{\mu}_{SF} > 0
\end{aligned}$$

The ABC calculations yield the following posterior distribution for the parameters. converge to the following parameter distribution, Fig 8

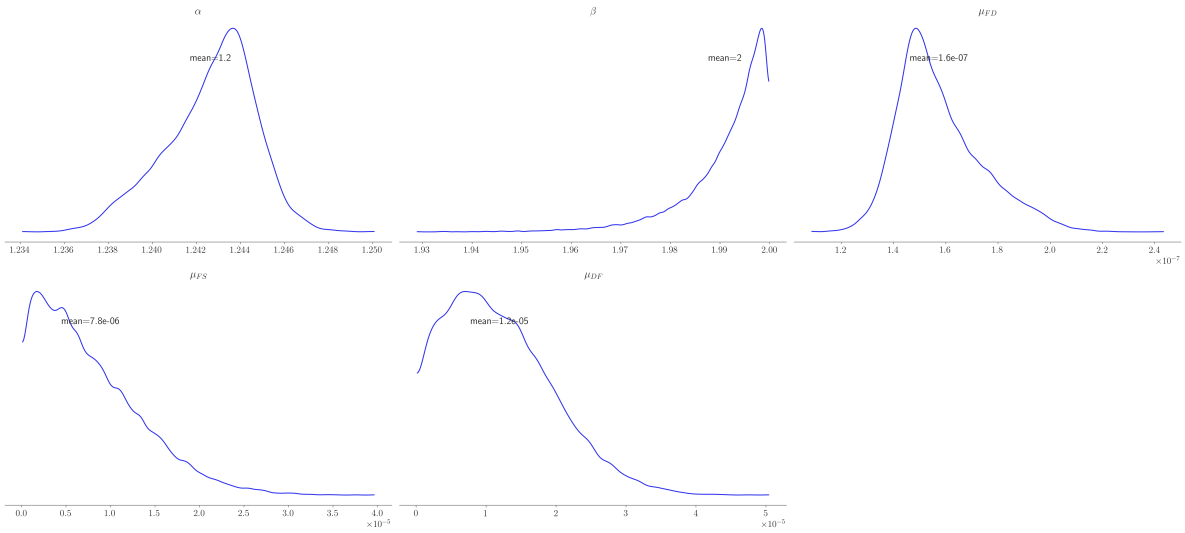

Figure 8: Posterior distribution for parameters in Eq 5

Even though we had a value for the replication rate of the *M2lop* mutant, we set it to vary in the model. Table 3 summarizes the mean and std of each distribution.

	mean	**std**
α	1.2427	2.0×10^{-3}
β	1.9911	8.5×10^{-3}
μ_{FD}	1.576×10^{-7}	1.634×10^{-8}
μ_{FS}	7.811×10^{-6}	5.804×10^{-6}
μ_{DF}	1.199×10^{-5}	7.598×10^{-6}

Table 3: Statistics summary of the parameter distribution for Eq 5 from ABC results

I choose the mean of each parameter as a representative sample of the algorithm results, and use them solve the experiment. To compare with the experimental data, I sum the population fractions from Duplication and SNP strains, this should correspond to the fraction of observed large colonies. Visually inspecting the population fractions in Fig 9 we can see that the prediction with the inferred parameters is close to the experimental data. Furthermore, the parameters seem to reproduce the long-time behaviour of the experiment as there is the founder population is not completely extinguished, instead there is a small remnant of it.

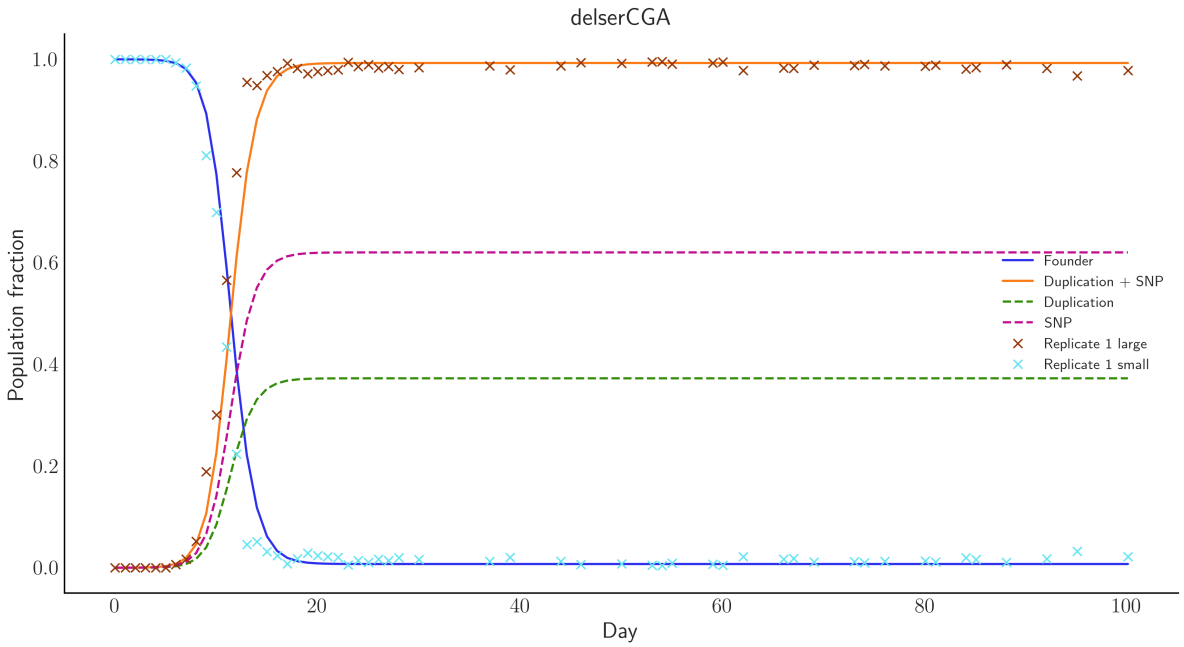

Figure 9: Sample of the experiment results using the mean of the parameters distribution

Furthermore, this plot shows that in the end, there is a stable equilibrium between SNP, Duplication and Founder strains with SNP being the dominant one. Moreover, this result does not completely support any of the hypotheses, perhaps signalling that it is necessary to review and update them.

Nonetheless, these are exciting results as now we seem to be able to estimate a reasonable value from the parameters. We can use this information to impose constraints in the parameter space when carrying out calculations for the time to reach the equilibrium.

2.4 Parameter space exploration

We explore the parameter space of our models to analyze the time to the steady state, and the fate of the different strains. We start by looking at the two state model; based on our previous work we know that the two parameters which determine the dynamics of the system are the ratio of the replication rates ($\alpha = r_M/r_F$) and the mutation loss rate μ_{MF}. Using the results of our inference task we restrict the search interval to $10^{-3} \leq \mu_{MF} \leq 2 \times 10^{-1}$ and $1 \leq \alpha \leq 2$

We say that the system has reached the steady state[2] when the variation of the population for three consecutive days is smaller than a set threshold of $\epsilon = 10^{-5}$

[2]The simulation ran for a maximum of 1000 days, in case that convergence was not achieved prior to that point, we show the max number of steps as the convergence time

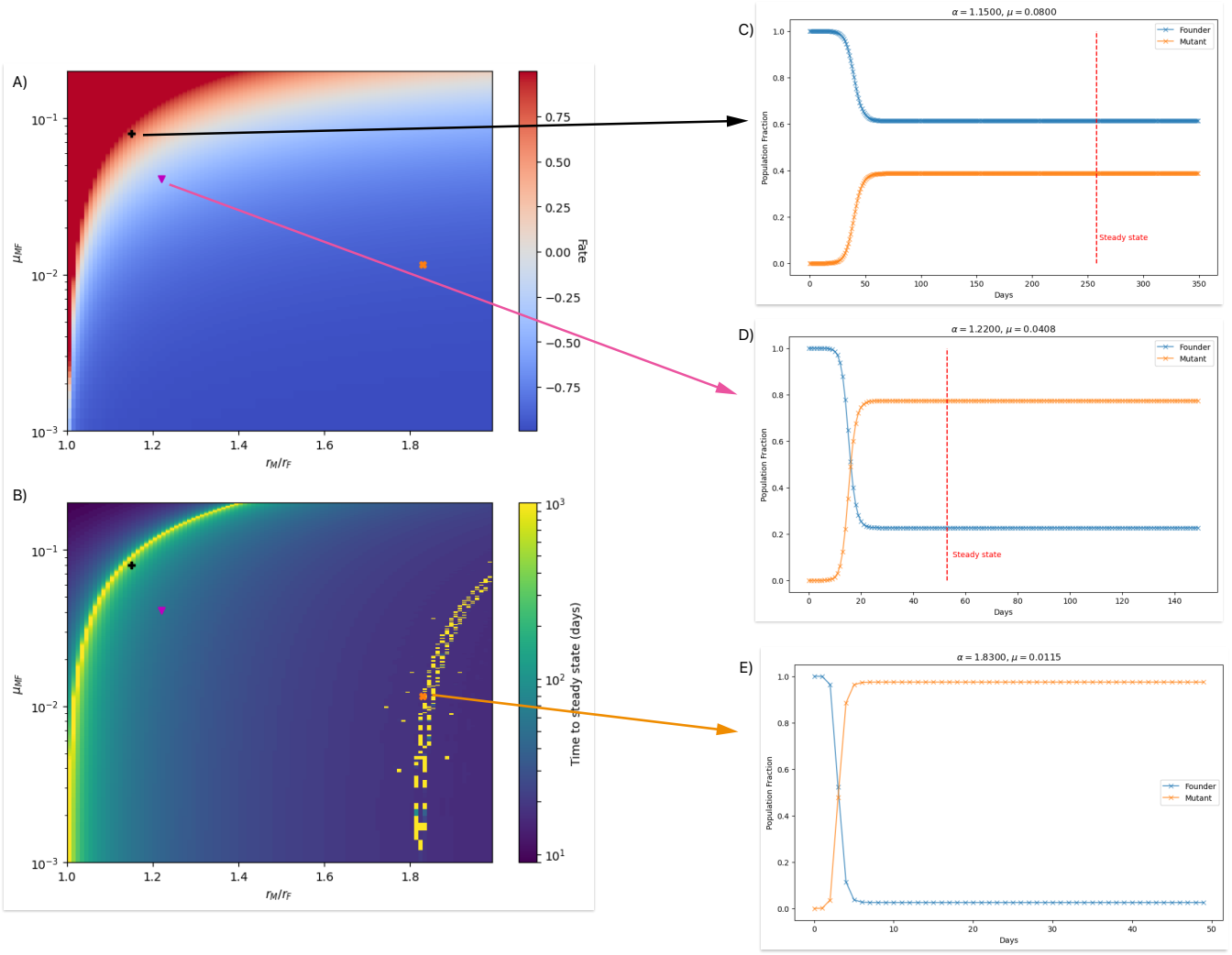

Figure 10: Parameter space for the two state system. The left panel shows the difference between the bacterial abundance at the end of the experiment (top), and the time to reach the steady state (bottom). In the former, the intensity of the color indicates a higher abundance of founder (red) or mutant (blue). In the latter, the color grading represents the number of days to reach the steady state. The right panel shows the evolution of population fraction at the end of every day for areas of interest in the parameter space.

The top panel in Figure 10 displays the difference between the proportions of founder and mutant at the steady state. The red tones indicate a higher abundance of founder, whereas the blue tones higher abundance of mutants. The red areas support the intuition that a higher mutation loss rate combined with a low mutant replication rate result in a great abundance of founders, therefore there will be no coexistence of mutants and founders. A similar affirmation is true for

All the lighter tones represent coexistence between founders and mutants, with a slight preference towards either population depending on the specific tone. We can see that coexistence is possible for most values of α but it only happens for $\mu_{MF} > 10^{-2}$

In addition, we can relate the behavior in the population abundance with the time to steady state. The transition between bold and light red areas in panel A matches the yellower curve in panel B, implying that the transition between a state of coexistence to a state where the founder takes over the population causes the system to reach the steady state at a later time.

The right panel in Figure 10 exemplifies of the evolution of the population fraction for different regions of the parameter space. C and D show two points at the coexisting region The regions where

one of the populations is dominant are characterized for small times to the steady state. These times get larger as we approach the regions where coexistence is possible. It is interesting to note as well, that this transition is seemingly smooth for the mutants and quite sharp for the founders.

3 Methods

To solve the ODE systems, I used the `numbalsoda` library in python

I used the MCMC implemented in python's `emcee` package. This choice is purely based on familiarity with the package and the simplicity of its implementation

I used the ABC implementation in PyMC v5.10.4 with the `sample_smc` function. I chosed 8 chains running on 4 cores of the M2 chip, the distance function is set to `laplace` and the tolerance is 0.001 The figure below shows the distribution of the chains for each parameter

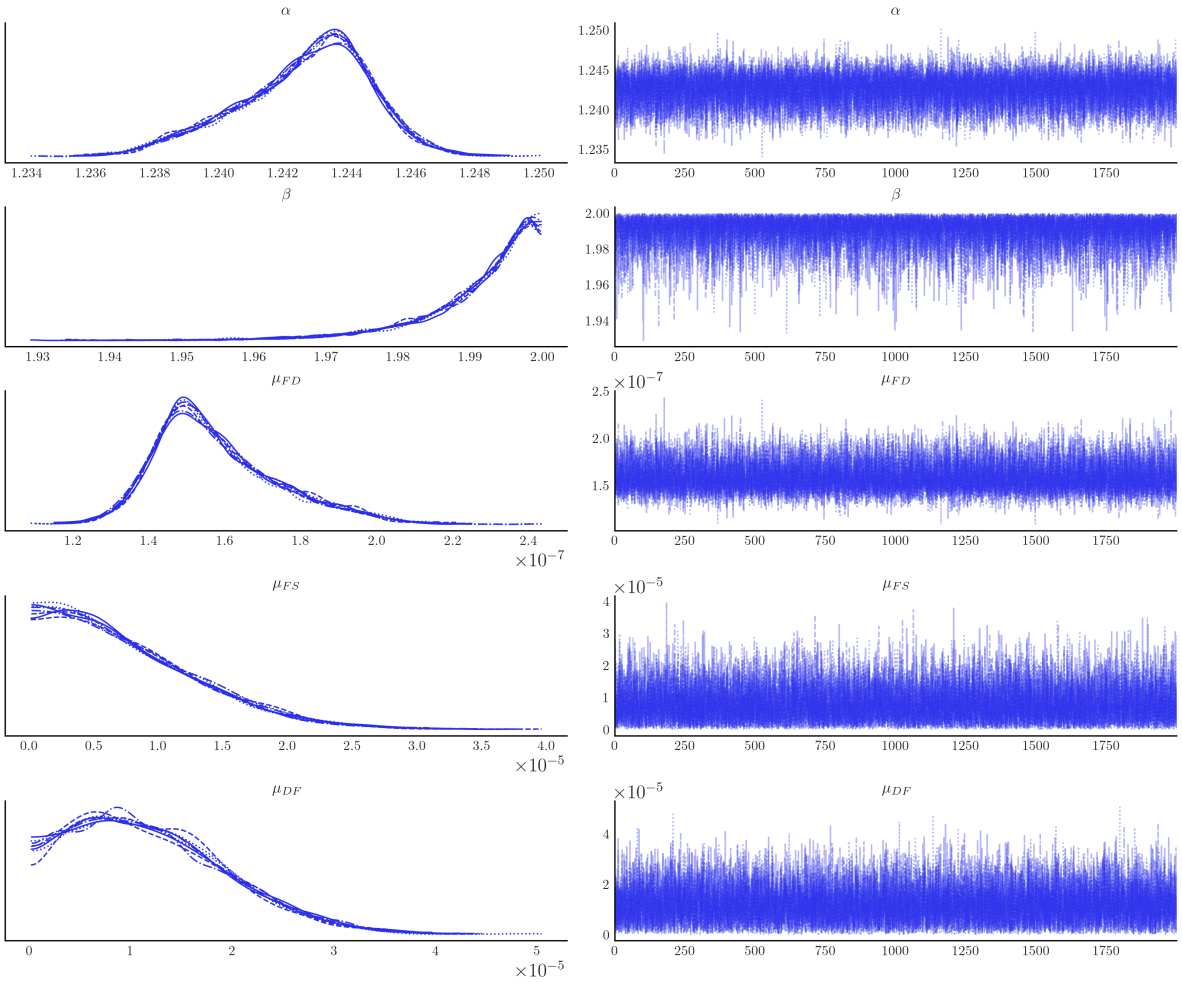

Figure 11: Distribution of the chains for ABC results in Fig 8

3.1 Determining convergence of the experiment

To determine the convergence of the experiment we calculate the area under the population curves using the Simpson's method implementation in `scipy.integrate.simpson`. In more detail, the procedure above is described by

$$\vec{A}_n = \int_n \{F_n(t), M_n(t)\} \mathrm{d}t, \tag{7}$$

where n represents a day in the experiment

$$\|\vec{A}_n - \vec{A}_{n-1}\| < \epsilon \tag{8}$$

where $\| \bullet \|$ is the euclidean distance. When this condition is satisfied for three consecutive days $(n, n+1, n+2)$ we say that the system has reached the steady state

3.2 Pending tasks

Here I want to summarize some tasks that are still pending without any particular order

- Implement a measure for the goodness of fits from the parameters obtained by bayesian inference. So far I am only visually inspecting the results and there is no quantitative measurement of it
- So far I have been running all the inference algorithms for replicate 1 of the experimental data. Even though upon visual inspection all the replicates look indistinguishable, it is necessary to run the algorithms for the remaining replicates to compare the results and test the robustness of the predictions. Nonetheless, I don't expect the results to change significantly.
- We discussed at some point in the past weeks about the apparent symmetry of the equations. Running the experiment for the 2 state system interchanging all the parameters between Founder and Mutant will be useful to test that independence of the predictions
- I still need to address the time that it'll take the experiment to converge for different experimental conditions. Since so far I haven't found an analytical solution, the easiest way to do it will be to explore the parameter space using the knowledge gained from the bayesian inference results.