

Internship report

Modeling ecDNA copy number dynamics in cancer cell populations

Herbert and Florence Irving Institute for Cancer Dynamics (IICD), Columbia University

Tess Breton, X2021, École polytechnique

Supervisor: Khanh Ngoc Dinh, IICD



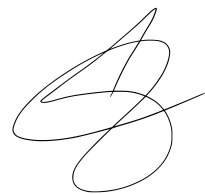
Declaration of Academic Integrity

I, Tess Breton, hereby declare that this research project report is my own work and that all sources and materials used in its preparation have been acknowledged.

I understand the principles of academic integrity and affirm that this work is free from any form of plagiarism.

Tess Breton

July 31st, 2024

A handwritten signature in black ink, consisting of a large, stylized 'T' and 'B' intertwined.

Acknowledgement

I am first grateful to my supervisor, Dr. Khanh Dinh, for his mentorship throughout my internship. His ideas and advice were always very helpful, and I greatly appreciated his availability.

I also wish to express my gratitude to Dr. Simon Tavaré, director of the IICD, for his weekly supervision and for always making sure we had a good time in New York City.

I wish to thank Zijin Xiang, a Master's student in the Statistics Department at Columbia University, for her help on using CINner and the insomnia cluster, and for being always cheerful and welcoming.

I am also grateful to Lorenza Favrot for inviting us Alliance interns to several events at the IICD, along with the interns from the Summer Research Program.

Finally, I wish to express my gratitude to everyone at the IICD for their warm welcome.

Abstract:

Extrachromosomal DNA (ecDNA) is a circular DNA entity often found within cancer cells, recently identified as playing a potentially crucial role in oncogene amplification. Unlike chromosomal DNA, ecDNA does not follow equal segregation during cell division, leading to highly variable copy numbers within a single population. My research project explores ways of simulating the dynamics of these copy numbers, based on reference data. After testing different models, we focused on a type of Moran process, adapted to incorporate the random segregation of ecDNA during cell division. To account for the biological properties of ecDNA, we introduced different fitness functions based on ecDNA copy number. All these functions were parameterized by a single selection parameter s , which we estimated using Approximate Bayesian Computation (ABC). We achieved great similarity between some simulations and our reference data, but the synthetic testing made us realize that our model's high stochasticity was a problem for accurate inference.

Code:

Python code available at: <https://github.com/tessbreton/ecDNA>

Contents

Introduction	1
1 Literature review	2
1.1 Biological features of ecDNA in human cancer	2
1.2 Random segregation of ecDNA during cell division	2
1.3 A first model for ecDNA dynamics	3
2 Reference data	4
2.1 Overview	4
2.2 Comments	4
3 Mathematical modeling	5
3.1 Assumptions and general considerations	5
3.2 Exploring different models	5
4 Moran death-birth process with random segregation	8
4.1 Fundamentals of Moran processes	8
4.2 Model description	8
4.3 Mathematical notations and formulas	9
4.4 Fitness functions	10
4.5 Simulation set-up	11
5 Parameter inference using ABC	13
5.1 Introduction	13
5.2 ABC set-up for selection parameter inference	13
5.3 Results	15
5.4 Synthetic testing	17
5.5 Double inference	18
Conclusion	19
References	20

Introduction

Extrachromosomal DNA (ecDNA) has recently gained significant attention in cancer research, due to its potential impact on tumor biology and therapy outcomes. Over the past few years, studies have unveiled the role of these circular DNA molecules in amplifying oncogenes, driving tumor heterogeneity, and conferring adaptability to cancer cells.

Given these findings, there is a growing interest in developing mathematical models to better understand and predict the dynamics of ecDNA in cancer cell populations. Since the role of ecDNA was discovered only recently, few previous works have focused on modeling its copy number dynamics. Most of the existing mathematical research on the subject remains relatively simple, and we believe that there is room for improvement regarding the impact of copy number on selection. More information on the ecDNA literature, both biological and mathematical, is provided in Section 1.

This research project was based on a single dataset made of the distributions of ecDNA copy number in a cell line at two different time points, presented in Section 2. Our goal was to produce simulations that matched this reference data as closely as possible. Specifically, we aimed to infer a single selection parameter that quantifies the selective advantage conferred by ecDNA on cell fitness.

To do so, we first explored several models to select the most suitable one. Section 3 outlines our approach to model ecDNA dynamics, presenting our main assumptions and discussing the three different models that we considered. Following these preliminary experiments, we focused on one single model inspired by Moran processes, detailed in Section 4. Our final goal was then to infer the selection parameter in our data, which we did using an ABC algorithm. The inference method and results are presented in Section 5.

Lab context: The initial goal of my project was to incorporate ecDNA dynamics into a larger model called CINner^[8], developed by my supervisor and colleagues at the IICD. However, after reviewing the literature on ecDNA, it became clear that achieving this within reasonable computation time would not be feasible. Indeed, CINner tracks population evolution by creating new clones whenever a copy number aberration occurs. Given the random segregation of ecDNA, this approach would require creating a new clone at almost every cell division, which undermines the clonal optimization. As a result, we decided that I should focus exclusively on ecDNA dynamics. I started my own project, which also proved to be more convenient.

1 Literature review

1.1 Biological features of ecDNA in human cancer

Most of the literature on ecDNA focuses on its biological features and impact on cancer growth. Several papers^{[11],[14]} review its known properties, focusing on its role in tumor heterogeneity and oncogene¹ amplification. Yet, since ecDNA is a very recent topic², many of its properties remain unknown. This is particularly true regarding its formation, for which a few models have been proposed without experimental confirmation. While we initially wanted to incorporate the formation of ecDNA into our model, we soon realized that it would not be feasible nor necessary.

Indeed, most ecDNAs impose a metabolic load on the cell^[9], leading to its rapid loss. However, in rare events, an ecDNA can be formed with a proliferative element such as an oncogene, providing a selective advantage to the cell. The goal of this project is to model the dynamics of such ecDNAs, which could have a significant impact on population evolution in the long term.

Despite the limited understanding of ecDNA, there is a consensus that it plays a potentially significant role in cancer evolution. First, ecDNA has been identified as a major carrier of amplified oncogenes^[14] and is more accessible than chromosomal DNA, resulting in increased transcriptional activity. It has also been found in very high copy numbers, leading to significant overexpression of the genes it carries. Finally, ecDNA molecules often form clusters or hubs^[6], further driving oncogene expression.

All these findings suggest that cells with higher ecDNA copy numbers might have a greater fitness, meaning a selective advantage over other cells in the population. As to how cells can reach such high ecDNA counts, recent discoveries indicate that ecDNA copies are randomly segregated among daughter cells during cell division, leading to high copy number variability.

1.2 Random segregation of ecDNA during cell division

One of the key studies investigating the mathematics behind ecDNA dynamics is presented in [9]. The main point of this paper was to prove the random segregation of ecDNA in human cancer cells, with equal probabilities for daughter cells to inherit each copy. Figure 1 shows the possible outcomes of such segregation, which can result in high copy number heterogeneity.

This randomness can lead to exceptionally high ecDNA counts, which are not reachable through chromosomal segregation. Given that ecDNA is a major carrier of amplified oncogenes, this may significantly impact the fitness of the cells, further contributing to cancer progression.

¹An oncogene is a gene that has the potential to cause cancer when mutated or overexpressed.

²ecDNA was first known as "double minutes", see [2]

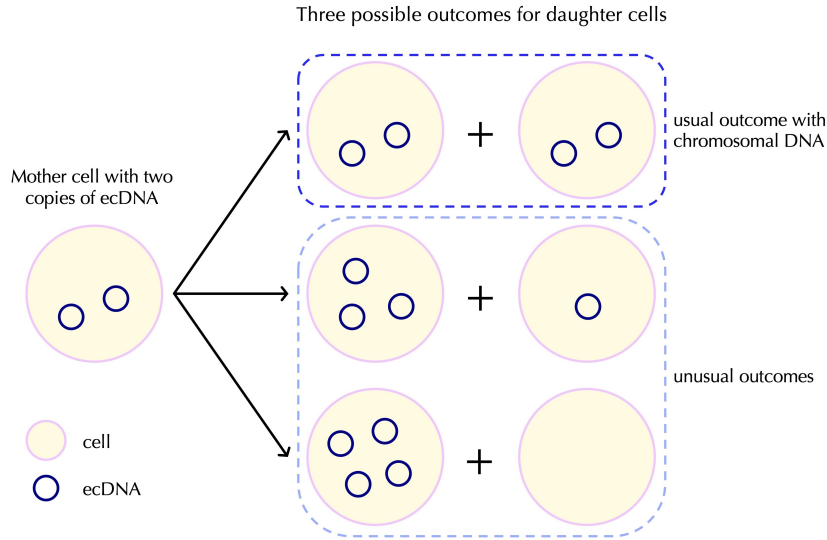


Figure 1: Possible outcomes of random ecDNA segregation

1.3 A first model for ecDNA dynamics

The first mathematical model found in the literature on ecDNA dynamics is a baseline model presented in the Supplementary of [9]. The goal of the authors was to make predictions to determine whether ecDNA is under neutral or positive selection, based on comparisons with patient and cell line data.

The authors build a baseline agent-based model in an exponentially growing population, starting from one single cell with one copy of ecDNA. In this model, cells divide, but never die without giving birth to two daughters. At every division, the copies of ecDNA are doubled and shared randomly among daughter cells, following a Binomial trial with probability $p = 0.5$. The simulation methodology is given below:

The cell to divide is chosen using a Gillespie^[5] algorithm, with different rates for cells with ecDNA and without. More precisely, with N_+ (resp. N_-) representing the number of cells with ecDNA (resp. without), we draw two independent random numbers ξ_+ and ξ_- from $\mathcal{U}([0, 1])$ and compute the corresponding reaction times :

$$\tau_+ = -\frac{1}{sN_+} \ln(\xi_+) \quad \text{and} \quad \tau_- = -\frac{1}{N_-} \ln(\xi_-), \quad \text{where } s \text{ is the selection parameter.}$$

The smallest reaction time determines which type of cell is chosen for division, and the cell to divide is picked uniformly within this category. This process is iterated until the population reaches a given input size $N = N_+ + N_-$.

Given this model, the first step of my project was to run simulations and compare the results to our real data, presented in the next section.

2 Reference data

2.1 Overview

Since ecDNA is a relatively new topic, very limited data was available on ecDNA counts. Fortunately, Ben Wesley, a PhD student at IICD, had recently worked on ecDNA for [1] and was able to share a small dataset³ with us. This dataset was built using single-cell DNA sequencing, and it includes the copy number distribution of the organoid CAM277 at two different time points, referred to as "passages". These distributions are shown in Figure 2.

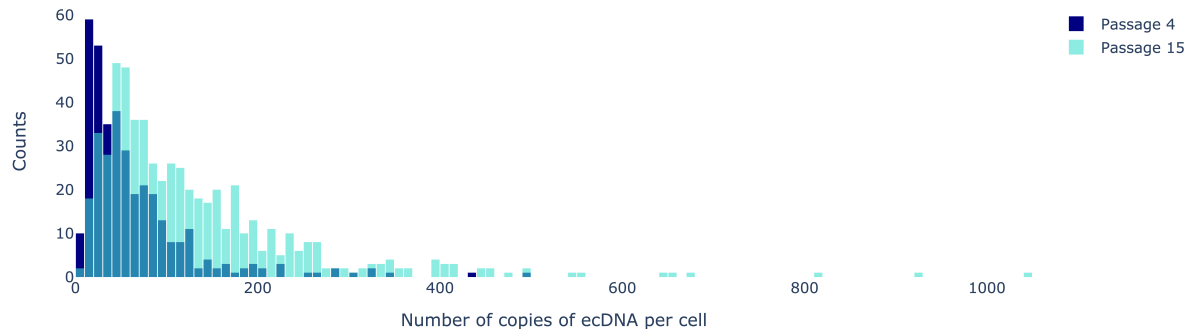


Figure 2: Histograms of ecDNA copy number distributions at passages 4 and 15

The authors of [13] provided limited experimental details, but we uncovered some valuable information with Ben's assistance. The cells were cultured over approximately 6 months for up to 16 passages, averaging about 11 days per passage. Looking at esophageal adenocarcinoma cell lines, Ben found that they typically divide every 30 hours. Using this information, we estimated that a single cell undergoes around 9 divisions per passage.

2.2 Comments

Still, several significant questions remained. First, we do not know when the first cell with ecDNA appeared in the population. Ben told us to assume that it emerged around passage -5, which we did. But we soon realized that the starting point was a critical parameter that should ideally also be inferred from the data. Moreover, we have no idea what the size of the population was at each time point. We only know that the number of cells for which we have ecDNA counts is 355 at passage 4, and 581 at passage 15. Finally, we do not know much about the sequencing method and its uncertainties. The sequencing detects ecDNA copies, so naturally we do not have any detection for cells without any copy. But for very small copy numbers (less than 10), we are not sure whether we can trust the data. This is why we decided, for distance computations later on, to consider only copy numbers larger⁴ than 10.

³The original data comes from [13]. It was processed to obtain ecDNA copy number distributions for [1].

⁴We filter all distributions once the simulations are over, removing copy numbers smaller than 10.

3 Mathematical modeling

3.1 Assumptions and general considerations

In every model, we assume that all copies of ecDNA are segregated randomly among daughter cells, and that ecDNA replicates at the same rate as chromosomal DNA during the cell cycle. We also assume that the first copy of ecDNA appeared during a single rare event, so that a cell without ecDNA will never gain new copies. For simplicity purposes, we focus on one single type of ecDNA, but the modeling could be adapted to track different kinds of ecDNA in a population.

Given that cancer sample statistics depend on particular tumor growth characteristics, it was hard to decide which dynamic to use for our cell populations. We tried different models with constant size, exponential growth and logistic growth. We also had to decide when to end our simulations, which took us a long time to figure out: we tried reaching a given cell count (for growing populations), waiting up to a given time, and finally chose to stop after a given number of cell divisions.

Since we wanted to study the distributions of ecDNA counts, we had to condition all our simulations on non-extinction of ecDNA. Starting with one single cell containing one copy, the ecDNA could get lost early in many simulations. So we decided to run another simulation every time one failed to keep ecDNA in the population.

3.2 Exploring different models

After reviewing the literature, the next step was to implement different models in order to identify the most suitable one for our problem. We considered three models, for which we conducted multiple simulations with various parameters, and visualized the results to gain insights into their respective behaviors.

3.2.1 Baseline model from [9]

We initially explored the model presented in 1.3, which we implemented to run simulations and visualize output distributions. We realized that these distributions' shapes, one of which is shown in Figure 3, did not align with our reference data. The range of ecDNA copy numbers was consistently too narrow. Even when simulating up to a million cells with a higher selection parameter, the maximum copy numbers achieved were significantly lower than those observed in the data. This observation suggests that the model might have been too simple to account for variations in fitness based on ecDNA copy number.

Indeed, this model does not take into account the fact that cells with higher ecDNA counts are thought to be fitter. It distinguishes only cells with and without ecDNA, no matter the copy number. We could try to adapt this model by drawing one uniform for every copy number, but it would be too time-consuming. For all these reasons, we concluded that we should build a new model in which ecDNA counts would have a greater impact on the evolution of the

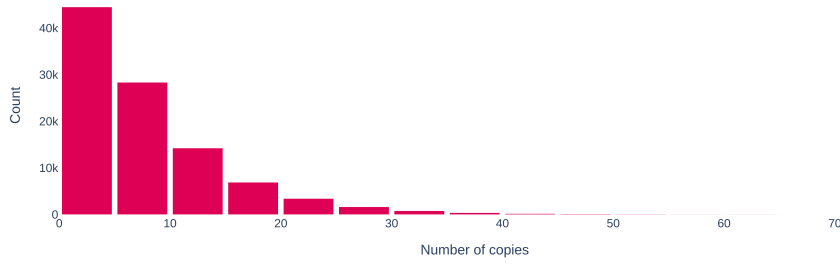


Figure 3: Histogram of ecDNA counts on 10^4 cells with $s = 3$

population.

3.2.2 Feedback loop model

The second model that we tested was inspired from CINner^[8], using a feedback loop to make the population follow a given input dynamic. In this model, we used the same event rate for all cells, but different division probabilities depending on ecDNA counts: at every step, a random cell is picked, and divides or dies with a probability determined by its ecDNA copy number.

In our first simulations, we used only two different division probabilities: with and without ecDNA, similarly to the previous model. Yet we could easily add complexity to incorporate the whole range of ecDNA counts, now without the computation time issues that we had with the previous model. The simulation methodology is given below:

At every step, we compute the time until the next event using the total cell count, and the cell to undergo the event is picked randomly. Then, the chosen cell divides with probability

$$p^{\text{div}}(t) = \begin{cases} g(t) \times s_+(t) & \text{if the cell has ecDNA} \\ g(t) \times s_-(t) & \text{otherwise} \end{cases}$$

where g is a negative feedback loop ensuring that the total cell count $P(t)$ follows a given input dynamic $P^*(t)$:

$$g(t) = \frac{P^*(t)}{P^*(t) + P(t)}$$

and s_+ , s_- model the selection for the fittest cell, with f (resp. 1) the fitness of cells with (resp. without) ecDNA:

$$s_+(t) = \frac{f}{f_{\text{avg}}(t)}, \quad s_-(t) = \frac{1}{f_{\text{avg}}(t)}, \quad \text{with } f_{\text{avg}}(t) = \frac{P_+(t) \times f + P_-(t) \times 1}{P(t)}$$

This model was more flexible than the previous one thanks to the input dynamic, and the results showed some improvement. Yet we were still far from reaching copy numbers as large as those from the reference data, and simulations could be time-consuming depending on what

condition we chose to terminate them. We also encountered some issues when the range of fitness values in the population was too wide, which was the case for functions that would grow quickly. We did not try simulating with different division probabilities for each copy number, as we were already satisfied with another model that we tested in parallel.

3.2.3 Moran process

The last model that we considered is inspired by Moran processes. A key difference from the other two models is that it maintains a constant total population size N . Simulations begin with one cell containing a single copy of ecDNA, while the remaining $N - 1$ cells are ecDNA-free. The population size remains constant through paired events of death and division, and the selective advantage conferred by ecDNA is modeled using a fitness function impacting cell division. This model turned out to be our best option, with enough modeling flexibility, reasonable computation time, and results that were a great fit with our data. Further details and simulation outputs are provided in Section 4.

4 Moran death-birth process with random segregation

4.1 Fundamentals of Moran processes

The Moran process is a stochastic model used to simulate genetic drift, devised at first for populations with two alleles A and a . Initially, each of the N individuals carries one of the two alleles. Then at every generation, one individual is chosen at random to reproduce, producing an offspring that replaces another individual chosen randomly in the population.

While Moran processes can be neutral, they can also model selective pressures in reproduction. In such cases, an individual's likelihood of reproduction is determined by its relative fitness in the population, referring to its ability to proliferate. Such models have been extensively studied, leading to well-known theoretical results^[4], including the probability of fixation of an allele.

Beyond the basic two-allele scenario, Moran processes have been generalized to multi-type Moran models, accommodating more than two possible genetic traits within a population. This extension finds applications in various fields, including modeling cell populations where different cell types interact and compete over generations.

Furthermore, in the standard Moran process, both the reproducing individual and the one being replaced are chosen randomly through two independent draws with replacement. This allows for the same individual to be selected for both events. However, an alternative approach consists in defining a Moran process without replacement, where the order in which reproduction (birth) and replacement (death) events occur becomes significant. We decided to use a Moran death-birth process^{[3],[10]}, in which we pick the cell to die before selecting the cell to divide.

To sum up, without considering the random distribution of ecDNA copies among daughter cells, the model follows the pattern shown in Figure 4 below:

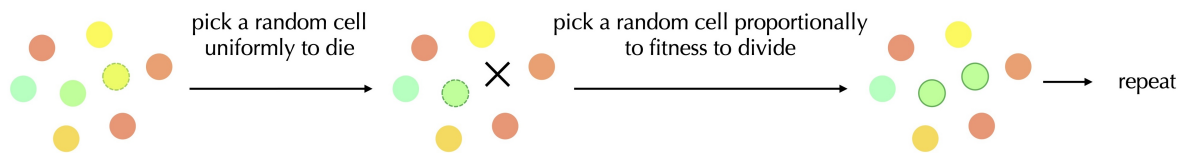


Figure 4: Basic multi-type Moran process

Although not implemented in our model, it is worth mentioning that Moran processes have been extended to populations with spatial structures, where individuals are represented as vertices of a graph^[4]. In this framework, each individual can replace only one of its neighbors, leading to fixation probabilities that heavily rely on the graph's structure.

4.2 Model description

In our Moran model for ecDNA dynamics, we consider a population of cells with a constant size N , where each cell carries a certain number of copies of ecDNA. The model is presented

schematically in Figure 5, and formal mathematical definition is given in 4.3. The key distinction from typical Moran processes is that ecDNA copies are randomly distributed between daughter cells. As a result, in most cases, the daughters will not share the same profile as their mother. Our model proceeds in discrete time steps, and each step is composed of one cell death followed by one cell division:

- **Cell death:** A cell is selected for death uniformly at random, and removed from the population.
- **Cell division:** A cell is picked for division at random among the remaining cells, proportionally to fitness (see 4.4 for considerations on fitness functions). During division, the ecDNA copies of the mother cell are doubled and shared randomly between its daughter cells.

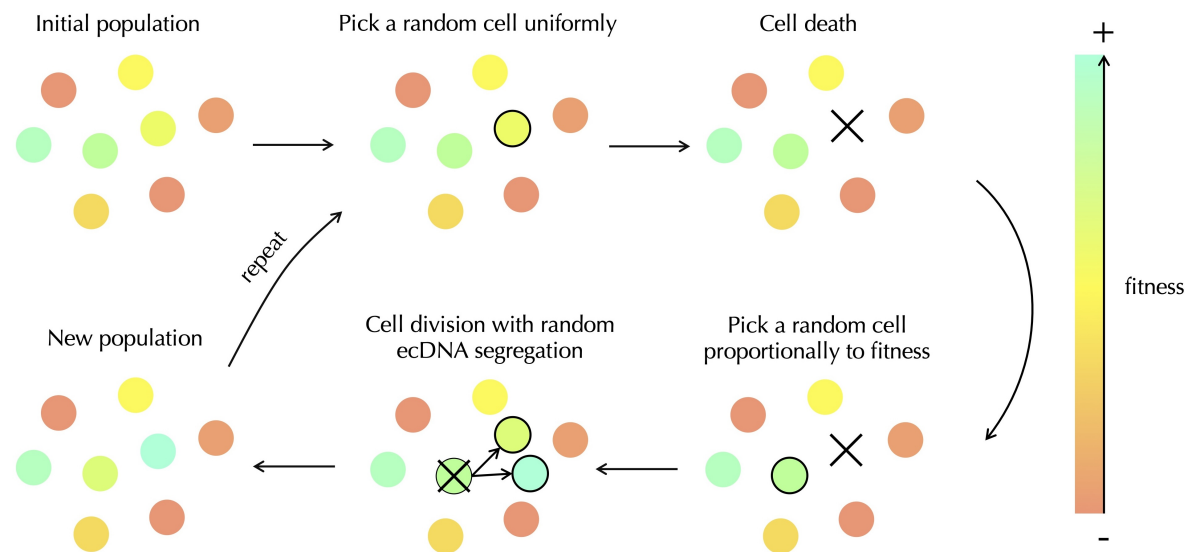


Figure 5: Moran death-birth process with random ecDNA segregation

4.3 Mathematical notations and formulas

Notations: N denotes the total population size, and N_k^i the number of cells containing k copies of ecDNA at time step i . Therefore, at every time step i , we have $N = \sum_k N_k^i$ and the population is fully described by $(N_k^i)_{k \geq 0}$. We also denote by D_i the number of copies of ecDNA in the cell picked to die at time step i , and similarly B_i for the cell picked to divide. Finally, f_k denotes the fitness of a cell containing k copies of ecDNA.

Initialization: Simulations start with one cell containing one copy of ecDNA, and all the others containing no copy. With the previous notations: $N_1^0 = 1$ and $N_0^0 = N - 1$.

Simulation methodology:

We repeat the death-birth steps a given number of times, using the following probabilities:

Since the dying cell is picked uniformly in the population:

$$\mathbb{P}(\text{death of a cell with } k \text{ copies at time step } i) = \mathbb{P}(D_i = k) = \frac{N_k^i}{N}.$$

After the death of the chosen cell, the population is updated accordingly:

$$N_{D_i}^{i+} = N_{D_i}^i - 1 \text{ and for } k \neq D_i, N_k^{i+} = N_k^i.$$

Then, the dividing cell is picked uniformly in the population, proportionally to fitness:

$$\mathbb{P}(\text{division of a cell with } k \text{ copies at time step } i) = \mathbb{P}(B_i = k) = \frac{N_k^{i+} f_k}{\sum_j N_j^{i+} f_j}.$$

Once the dividing cell is chosen, the copies of ecDNA are doubled and shared randomly among daughter cells. The number of copies C_i given to the first daughter cell is drawn from the binomial distribution $\mathcal{B}(2B_i, 0.5)$. The other daughter cell gets the remaining copies: $2B_i - C_i$.

After the division of the chosen cell, the population is updated accordingly for the next step:

- First, for every k , $N_k^{i+1} = N_k^{i+}$
- Then $N_{B_i}^{i+1} \leftarrow N_{B_i}^{i+} - 1$ (mother cell)
- And for $j \in \{C_i, 2B_i - C_i\}$, $N_j^{i+1} \leftarrow N_j^{i+} + 1$ (daughter cells)

4.4 Fitness functions

A crucial aspect of our modeling was the selection of the fitness function, which plays a role in determining the cell chosen for reproduction at each step. For the sake of simplicity and interpretability, we aimed for a function that remained straightforward and depended on a single selection parameter s .

We began by running many simulations using different simple fitness functions. Our goal was to examine the resulting output distributions, and identify those whose overall shape matched best our reference data. With f_k the fitness of a cell with k copies of ecDNA, we first considered the following functions, which seemed the most intuitive:

- Binary fitness: $f_0 = 1$, and $f_k = s$ for $k \geq 1$, which distinguishes only cells with and without ecDNA, regardless of copy number.
- Linear fitness: $f_k = 1 + s k$, with s a selection parameter.

- Power fitness: $f_k = (1 + s)^k$, with s a selection parameter.

As expected, we quickly realized that binary fitness did not provide enough selective advantage to cells with ecDNA to reach the high copy numbers observed in our data. On the other hand, both linear and power fitness grew too rapidly, resulting in the complete loss of cells with low copy numbers and a strong shift toward very high counts, even with a small selection parameter. We concluded that we needed an intermediate function that would allow for large ecDNA counts, while maintaining some cells with lower copy numbers. This led us to try with a logarithmic fitness function, which proved to be the best among those we considered. Therefore, we first decided to use the following fitness function for our simulations:

$$f_k = 1 + \ln(1 + sk)$$

Figure 6 provides insight into why logarithmic fitness outperforms linear fitness on our data: it increases rapidly for small copy numbers, promoting ecDNA proliferation, but its slope decreases significantly as the copy number grows. This also results in simulations aligning quite well with biological observations, which indicate that ecDNA counts cannot grow indefinitely due to a biological upper bound on the number of copies a cell can carry.

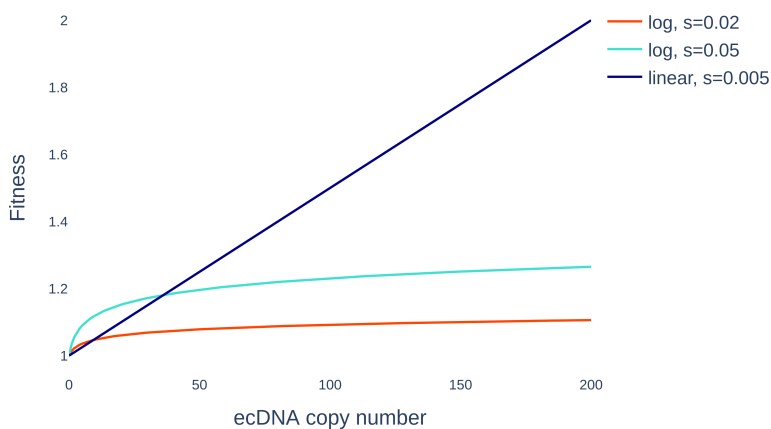


Figure 6: Logarithmic and linear fitness functions

4.5 Simulation set-up

One of the main parameters we had to choose to run simulations was the total population size N . The reference data does not provide any information on the total size in the cell line, only indicating that 355 cells were sampled at passage 4 and 581 at passage 15. Therefore, we decided to use $N = 1000$ and kept this value throughout the project.

Another important consideration was determining when to stop the simulation. We eventually decided to run our simulations up to a specified number of cell divisions, determined through

an approximation based on cell properties. Based on our estimations from 2.1, we decided that one passage would correspond to $9 * N = 9000$ cell divisions.

To study the distribution of ecDNA copy numbers, we needed simulations in which ecDNA was still present at the end. Since our model starts with a single cell containing one copy and includes cell death, it was possible for all ecDNA to be lost during the simulation. This led us to introduce a non-extinction condition: if at some point no cell had any ecDNA left, then the simulation would be terminated and restarted. Frequently, especially when the selection parameter was small, ecDNA was lost early in the simulations, requiring several hundred runs to obtain one that preserved some ecDNA.

5 Parameter inference using ABC

5.1 Introduction

5.1.1 Overview of ABC methodology

Approximate Bayesian Computation (ABC) is a computational method used for parameter inference when the likelihood function is intractable or difficult to compute. In that case, usual techniques such as Maximum Likelihood Estimation cannot be implemented. Then, instead of relying on explicit likelihood calculations, ABC starts from a prior distribution of the parameters and generates a posterior distribution through repeated simulations. In each simulation, parameters are sampled from the prior distribution (often a uniform distribution on a range likely to contain the optimal value), and simulated data is generated using these parameters. The simulated data is then compared to the reference data using a distance function. Parameters that produce synthetic data sufficiently "close" to the observed data are accepted, and used to approximate the posterior distribution.

By running a large number of simulations and accepting parameters that result in close matches, ABC effectively builds a posterior distribution that reflects the parameters most likely to explain the reference data.

5.1.2 Rationale for using ABC with ecDNA

One reason why ABC seems to be a good method for parameter inference in our model is that computing an explicit likelihood function is too complicated. While some theoretical results are available on basic Moran processes, the random segregation of ecDNA adds another layer of complexity to our model.

5.2 ABC set-up for selection parameter inference

Our first goal was to run an ABC algorithm to infer the selection parameter s , assuming that the starting passage was known and equal to -5 (see 2.2 for explanation). This single parameter inference is the main point of the project. Yet we also tried to run a double inference to infer both the selection parameter and the starting passage, presented in 5.5.

We first explain the algorithm overall, then give more detailed information on prior distribution and distance function.

5.2.1 Our ABC algorithm

In Algorithm 1, we outline the method used to infer the selection parameter s . This approach follows a standard ABC framework, except for the definition of the posterior distribution. Instead of employing a fixed distance threshold, as in many ABC methods, we decided to keep the top 5% of our simulations. This approach offered flexibility, given our uncertainty about the magnitude of the distances encountered. It also remains as adaptable as a threshold-based

method, since we can adjust the percentage as needed. We could as well have later defined a distance threshold, based on observed values, but we decided to keep the 5% threshold that seemed to work fine.

Algorithm 1 ABC Algorithm for Selection Parameter Inference

- 1: **Input:** Reference data \mathcal{R} . Prior distribution of s . Number of samples N_{samples} . Distance function d .
 - 2: **Output:** Posterior distribution of s .
 - 3: Sample N_{samples} values of s from the prior distribution.
 - 4: **for** each sampled s **do**
 - 5: Run one simulation \mathcal{S} with selection parameter s .
 - 6: Compute the distance $d(\mathcal{S}, \mathcal{R})$ between simulated and reference data.
 - 7: **end for**
 - 8: Find the simulations that gave the 5% smallest distances $d(\mathcal{S}, \mathcal{R})$.
 - 9: **Return** the values of s corresponding to these 5% best simulations.
-

A simplified schematic version of the algorithm is also presented in Figure 7.

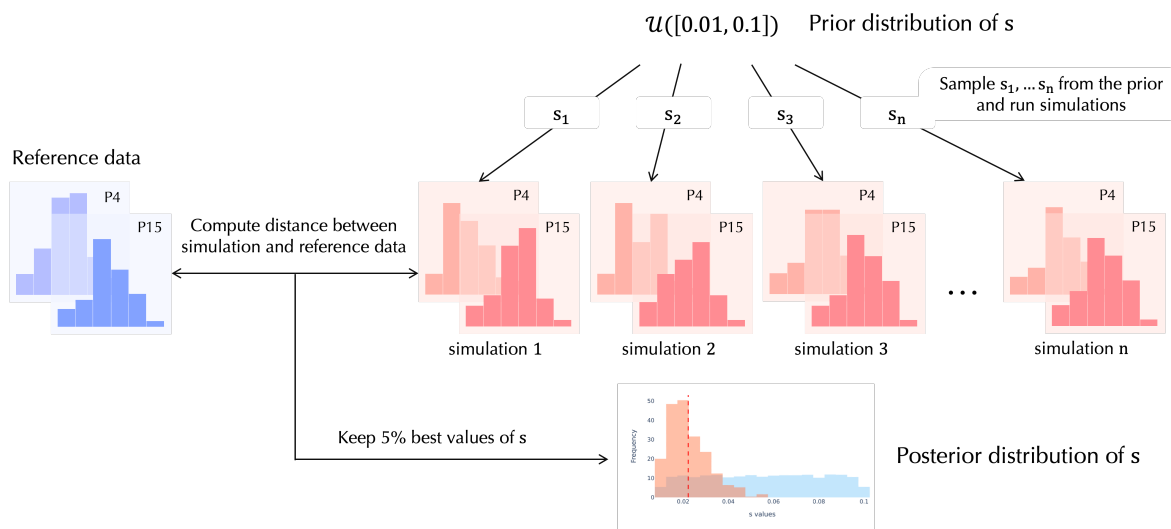


Figure 7: ABC algorithm for selection parameter inference

5.2.2 Prior distribution

Based on the outcomes of preliminary simulations, we decided to set our prior distribution for s as uniform over the range $[0.01, 0.1]^5$. Values of s below 0.01 were impractical because ecDNA frequently disappeared due to insufficient selective pressure, often resulting in thousands of unsuccessful simulations. As for the upper bound, the average copy numbers reached with

⁵Figure 6 shows that the fitness functions are significantly different within this range.

$s = 0.1$ were significantly larger than those from our reference data. This ensured that the posterior would not be constrained to the upper bound of the range.

5.2.3 Distance

Another major parameter of our ABC setup was the choice of the distance metric used to compare simulated and reference data. We opted for the Wasserstein-1 distance^[12], also known as Earth Mover's Distance (EMD). Intuitively, this metric quantifies the minimum "cost" required to transform one distribution into another. We believed it could perform better than other traditional metrics (such as L1 or L2) in capturing the overall shape of the distribution. Our primary concern was the overall distribution rather than specific individual copy numbers, especially given the measurement uncertainties in the reference data. The Wasserstein-1 distance is well-suited for this purpose, as it remains robust to small variations within the distributions.

In the one-dimensional case⁶, the Wasserstein-1 distance between probability distributions μ and ν with cumulative distribution functions (CDFs) F_μ and F_ν is given by:

$$W_1(\mu, \nu) = \int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)| dx \quad (1)$$

In other words, Equation (1) means that the Wasserstein-1 distance between μ and ν is the area between their respective CDFs.

We used both time points to compute the distance, defined as a weighted sum of the Wasserstein-1 distances at passages 4 and 15 with coefficients $(2, 0.5)$ ⁷. The distance $d(\mathcal{S}, \mathcal{R})$ between the simulation \mathcal{S} and the reference \mathcal{R} is then given by:

$$d(\mathcal{S}, \mathcal{R}) = 2 * W_1(\mathcal{S}_{P4}, \mathcal{R}_{P4}) + 0.5 * W_1(\mathcal{S}_{P15}, \mathcal{R}_{P15}) \quad (2)$$

5.3 Results

Based on the previous explanations, we implemented our ABC algorithm with the following configuration:

- Prior distribution for s : $\mathcal{U}([0.01, 0.1])$
- Distance: d defined in Equation (2)
- Number of samples: $N_{\text{samples}} = 10^4$
- Starting passage: $P = -5$
- Number of cells in the population: $N = 1000$
- Number of events per simulation: $n_{\text{events}} = (15 - (-5)) * 9 * N = 180\,000$

The posterior distribution obtained with this set-up is given in Figure 8, centered around a mean of 0.022. There is a notable concentration of the distribution around this mean,

⁶See [12] for the general formula.

⁷We chose these values based on the Wasserstein-1 distances observed at P4 and P15 on all 10^4 samples.

indicating a satisfactory fit. However, given that this average is near the lower boundary of the prior distribution, the posterior might be constrained. But as discussed earlier in 5.2.2, we decided not to run simulations with smaller values because they would fail too often.

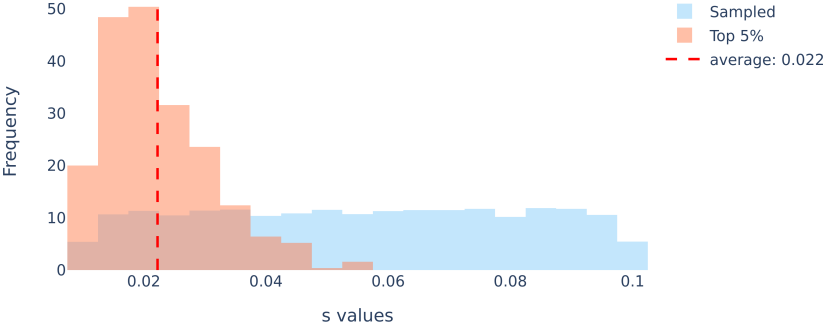


Figure 8: Posterior distribution of s , over 10^4 samples starting at P-5

To confirm the alignment of our top simulations with the reference data, we plotted the reference and simulated distributions of a few top simulations at passages 4 and 15. The distributions closely matched the reference both in shape and values, confirming the relevance of our modeling. Figure 9 shows the distributions at passages 4 and 15 of the best simulation based on the distance d . Additionally, to provide an overview of the top 5% simulations, the average histograms and confidence intervals are shown in Figure 10.

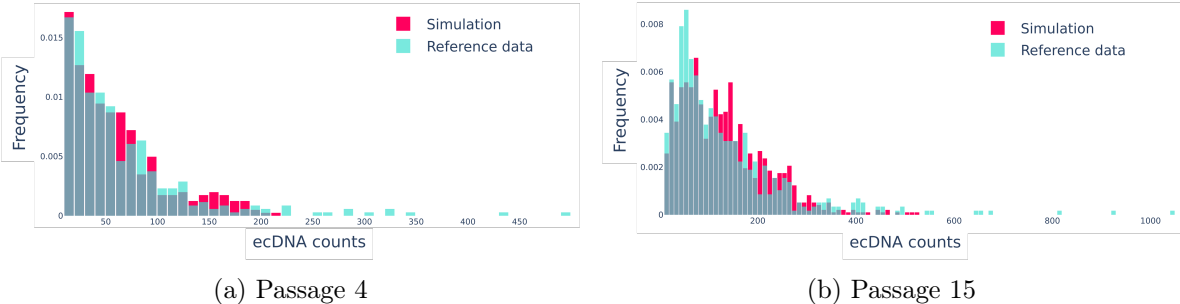


Figure 9: ecDNA counts of reference data and best simulation over 10^4 samples

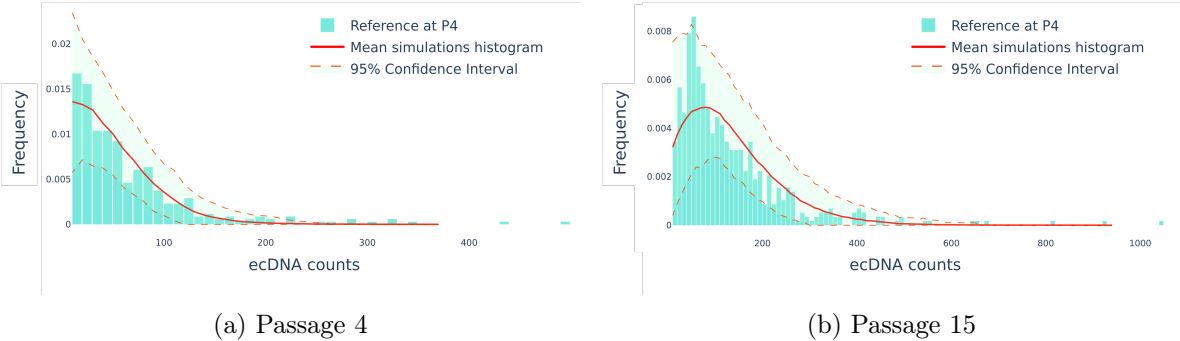


Figure 10: ecDNA counts of reference data and top 5% best simulations over 10^4 samples

5.4 Synthetic testing

To evaluate our method, we conducted synthetic testing. This consists in generating synthetic reference data using predefined parameters, and then applying our ABC algorithm to infer these parameters. By comparing the inferred posterior distributions with the actual parameter values, we can assess the accuracy and reliability of our method. The inference should ideally return posterior distributions that are centered around the true parameter values used to generate the synthetic data.

To prevent biases around the borders of the prior range $[0.01, 0.1]$, we sampled 500 s values uniformly in $[0.03, 0.08]$. For each of these values, we ran one simulation and gave it as reference to our ABC algorithm. Then, we took as estimated value the average of the posterior distribution obtained. A scatter plot of the estimated and true values is shown in Figure 11.

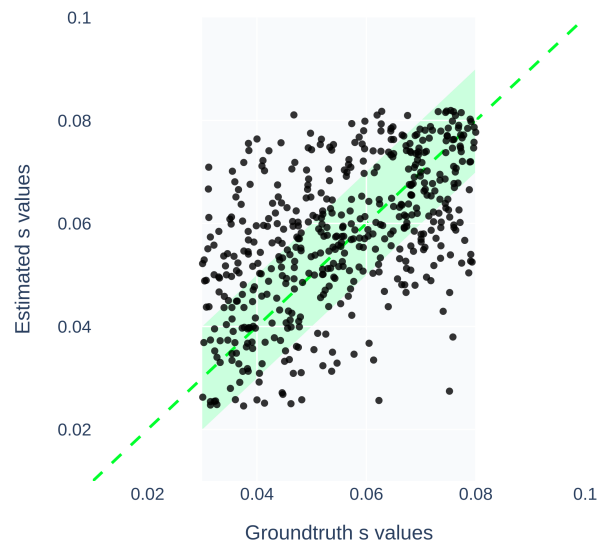


Figure 11: Ground truth vs. estimated values of s

If the inference worked perfectly, then all points would be on the green diagonal. Here, it is obviously not the case, and some points are even very far from the diagonal. Yet the average error is around 0.01, which appears to be quite reasonable. After seeing this plot, we tried to gain a deeper understanding of what was happening, and why the inferred value could be so far from the ground truth in some cases. We visualized more distributions and ended up concluding that our simulations' high stochasticity posed a great challenge for parameter inference.

Indeed, our stochastic model often yields quite different final copy number distributions, even with identical input parameter sets (see Figure 12). The final distribution depends heavily on when the ecDNA counts start exploding.

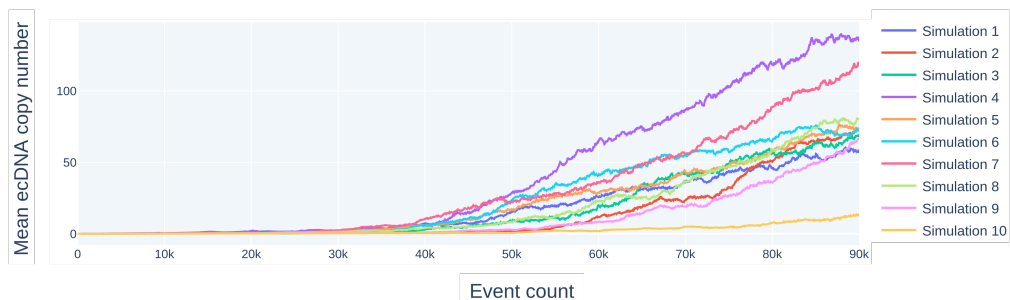


Figure 12: Mean ecDNA copy number in the cell population over time on 10 simulations with $s = 0.04$

5.5 Double inference

Finally, we also tried to infer both the selection parameter s and the starting passage P (see 2.2) at the same time. The method is similar in all points to the previous one, except that the prior for the parameters (s, P) is now uniform on $[0.01, 0.1] \times \{-9, -8, \dots, 0\}$. We proceed in the exact same way as before for the ABC inference, and obtain the joint posterior distribution shown in Figure 13.

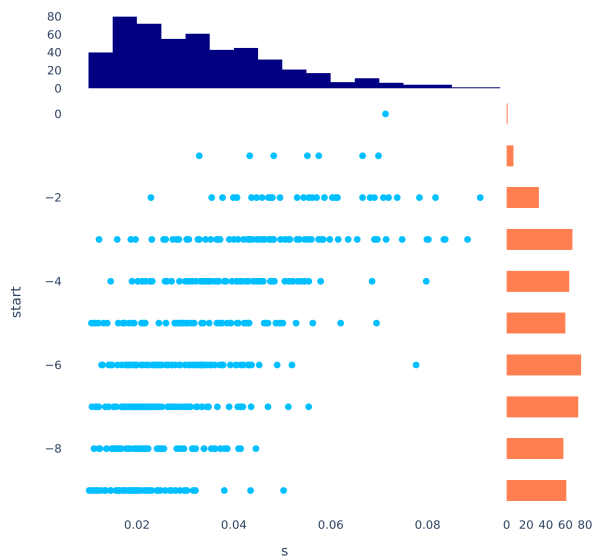


Figure 13: Scatter plot of 5% best (s, P) parameters over 10^4 samples

This plot does not seem to show any strong concentration around a single point, but more of a concentration along the diagonal. Our interpretation is that there might be a compensation between the selection parameter and the starting passage, making it difficult to infer both of them with only two time points. For some given (s, P) , similar output distributions could be reached either starting earlier with a larger s (strong selective advantage), or later with a smaller s .

Conclusion

Discussion

Overall, this research project has produced promising results. We notably reached a strong alignment between simulations and reference data, with similar copy numbers and distribution shapes. This is a significant improvement over previous models from the literature, which showed limited similarity and used simpler methods to model the influence of ecDNA counts on cell fitness. Additionally, our ABC inference method shows posterior distributions that tend to center around specific values, supporting the significance of our selection parameter and the relevance of our model.

However, a notable challenge lies in the limitations observed during synthetic testing. The high stochasticity of our simulations, particularly the unpredictable timing of ecDNA count explosion, complicates the inference. While the selection parameter significantly impacts long-term outcomes, its influence is minimal during the initial stages of the simulations. Although the selection of the dividing cell is made proportionally to fitness, there are initially so many cells without ecDNA that it takes a long time before the ecDNA starts spreading. This results in output distributions that may appear similar despite substantial variations in selection parameters, posing challenges for accurate inference.

One potential solution could involve initializing simulations from a known initial state, where ecDNA propagation is already underway. This could help mitigate the stochastic variability associated with the timing of ecDNA counts explosion. Thus, we tried running simulations from passage 4 to passage 15, to identify the optimal selection parameter to transition between these two states. However, we did not know how to handle the constant size of our population: while we have data on the copy numbers of some cells, the number of cells lacking ecDNA remains unknown. Addressing this requires determining how many cells without ecDNA to include, which we do not know. Due to time constraints at the end of the internship, we did not look into this any further, but we consider it a promising direction for future research.

Next steps

While we have already obtained some interesting results from only two time points on one single organoid, we think that more robust conclusions could emerge from wider data. We believe that future research could benefit from incorporating additional time points and more comprehensive data. This includes information such as total population size, the timing of ecDNA formation, and estimates of the number of cells without ecDNA or with few copies that may be missed during sequencing.

In conclusion, while our current model provides valuable insights, there are several paths for further exploration to enhance the accuracy and reliability of our modeling and inference.

References

- [1] Ng A.W.T., McClurg D.P. and Wesley B. et al. *Disentangling oncogenic amplicons in esophageal adenocarcinoma*. 2024. DOI: [10.1038/s41467-024-47619-4](https://doi.org/10.1038/s41467-024-47619-4).
- [2] Spriggs AI, Boddington MM and Clarke CM. *Chromosomes of human cancer cells*. 1962. DOI: [10.1136/bmj.2.5317.1431](https://doi.org/10.1136/bmj.2.5317.1431).
- [3] Khanh N. Dinh and Monika K. Kurpas. *Comparison of Tug-of-War Models Assuming Moran versus Branching Process Population Dynamics*. 2024. DOI: [10.1101/2023.10.20.563302](https://doi.org/10.1101/2023.10.20.563302).
- [4] Lieberman E., Hauert C. and Nowak M. *Evolutionary dynamics on graphs*. 2005. DOI: [10.1038/nature03204](https://doi.org/10.1038/nature03204).
- [5] Daniel T. Gillespie. *Exact Stochastic Simulation of Coupled Chemical Reactions*. 1977. URL: <https://www.cmor-faculty.rice.edu/~cox/gillespie.pdf>.
- [6] King L. Hung, Kathryn E. Yost and Liangqi Xie. *ecDNA hubs drive cooperative intermolecular oncogene expression*. 2021. DOI: [10.1038/s41586-021-04116-8](https://doi.org/10.1038/s41586-021-04116-8).
- [7] Marek Kimmel. *Quasistationarity in a branching model of division-within-division*. 1997. URL: https://link.springer.com/chapter/10.1007/978-1-4612-1862-3_11.
- [8] Dinh KN, Vázquez-García I, Chan A, Malhotra R, Weiner A, McPherson AW and Tavaré S. *CINner: modeling and simulation of chromosomal instability in cancer at single-cell resolution*. Preprint. 2024. DOI: [10.1101/2024.04.03.587939](https://doi.org/10.1101/2024.04.03.587939).
- [9] Joshua T. Lange, John C. Rose, Celine Y. Chen and Yuriy Pichugin. *The evolutionary dynamics of extrachromosomal DNA in human cancers*. 2022. DOI: [10.1038/s41588-022-01177-x](https://doi.org/10.1038/s41588-022-01177-x).
- [10] Jakub Svoboda, Soham Joshi, Josef Tkadlec and Krishnendu Chatterjee. *Amplifiers of selection for the Moran process with both Birth-death and death-Birth updating*. 2024. DOI: [10.1371/journal.pcbi.1012008](https://doi.org/10.1371/journal.pcbi.1012008).
- [11] Tianyi Wang, Haijian Zhang, Youlang Zhou and Jiahai Shi. *Extrachromosomal circular DNA: a new potential role in cancer progressions*. 2021. DOI: [10.1186/s12967-021-02927-x](https://doi.org/10.1186/s12967-021-02927-x).
- [12] Wikipedia. *Wasserstein metric*. Consulted on 07/15/2024. URL: https://en.wikipedia.org/wiki/Wasserstein_metric.
- [13] Li X., Francies H.E. and Secrier M. et al. *Organoid cultures recapitulate esophageal adenocarcinoma heterogeneity providing a model for clonality studies and precision therapeutics*. 2018. DOI: [10.1038/s41467-018-05190-9](https://doi.org/10.1038/s41467-018-05190-9).
- [14] Dong Yucheng, He Qi, Chen Xinyu, Yang Fan, He Li and Zheng Yongchang. *Extrachromosomal DNA (ecDNA) in cancer: mechanisms, functions, and clinical implications*. 2023. DOI: [10.3389/fonc.2023.1194405](https://doi.org/10.3389/fonc.2023.1194405).