

# **Dissertation Report**

## **Statistical Modelling of DNA Replication**

**Submitted by:  
Alizah Erum**

**IN FULFILMENT OF THE REQUIREMENT OF  
MSc DATA SCIENCE**

**28th April,2025**

***Under the guidance of the supervisor -  
Dr. Alessandro Moura***



**UNIVERSITY OF  
ABERDEEN**

## **Acknowledgement**

I would like to express my sincere gratitude to Dr Alessandro Moura, Professor of Physics in the School of Natural Sciences and Computing at the University of Aberdeen, for granting us the opportunity to undertake this project in fulfilment of the requirements for the M.Sc. in Data Science. I also extend my heartfelt thanks to all team members whose dedication and hard work made this possible.

## **Abstract**

One of the basic biological processes that propels genetic variety, adaptability, mutation, and organism expansion is DNA replication. We compared the structures of genomes of two organisms—*Saccharomyces cerevisiae* and *Candida Glabrata*. *Saccharomyces cerevisiae* is used in baking, brewing, and wine. On the other hand, the human stomach and gastrointestinal tract are normally home to *Candida glabrata*. Our goal is to understand how spatial genome arrangement influences the evolutionary traits of these organisms. Using statistical analysis and Python-based tools, we examined genomic datasets to explore how spatial structure affects DNA replication efficiency and the ability to repair cellular damage, key factors in the organisms' survival and evolution.

## **Introduction**

DNA replication is essential for the life cycle of an organism. There are certain locations/sites in the organism through which replication takes place, and those sites are known as the origin of replication. It is essential to know the spatial arrangement of these organisms so that we can get an idea of the cell's efficiency in replicating, repairing, growing, and reproducing. In this project, we have attempted to determine and compare whether the inter-origin distances between these two organisms are the consequence of haphazard genome arrangement or follow an evolutionary design.

To do this, we quantified the distances between successive origins of replication, evaluated genome-wide origin of replication datasets for both species, and used statistical techniques to look at distribution patterns. We compared analyses to determine whether continuous patterns in the origin of replication spacing point to evolutionary optimisation intended to reduce replication mistakes, guarantee complete replication, or take into account the genome's structural restrictions.

## **Saccharomyces cerevisiae Dataset**

In this dataset, we see that there are genome coordinates of the origins of the replication of the chromosomes, which are likely to replicate and have confirmed replication. In this dataset, we get to see the following columns:

chr : The origin's chromosome number, which might be anything between 1 and 16.

start : The origin's initial base pair position.

end : The origin's final base pair position.

name: The replication origin's primary name or identifier.

othernames: the replication origin's alternative names.

status: Confirmed or Likely status.

	chr	start	end	name	othernames	status
0	1	650	1791	ARS102	proARS102	Confirmed
1	1	6136	7136	ARS102.5	proARS103	Confirmed
2	1	7998	8548	ARS103	proARS103	Confirmed
3	1	9775	10485	NaN	NaN	Likely
4	1	16855	17565	NaN	NaN	Likely

For the purpose of analysis, we have taken chr, start, end and status.

### Method

1. Importing libraries: Several necessary Python libraries were imported in order to facilitate data processing and visualisation.

These consist of:

- Pandas (pd) for managing structured datasets and data processing.
- NumPy (np) for array processing and numerical operations.
- Matplotlib.pyplot (plt) and Seaborn (sns) for creating visualizations, where a clean, grid-based background is applied using Seaborn's `set_style("whitegrid")` method to improve plot readability.
- In order to make statistical modelling and hypothesis testing easier, SciPy. stats imports the `expon` (exponential distribution), `poisson` (Poisson distribution), and `chisquare` (Chi-square test).

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style("whitegrid")
from scipy.stats import expon, poisson, chisquare
```

2. Calculate midpoints using genome coordinates and calculate distances between consecutive positions after sorting them.

```
def position_point(chrom):
    """Calculate midpoints of genomic regions for a specified chromosome."""
    filtered = df_data[df_data['chr'] == chrom]
    return [(row['start'] + row['end']) // 2 for _, row in filtered.iterrows()]
```

```
#Calculate distances between consecutive sorted positions
def distances_pos(positions):
    """Calculate distances between consecutive sorted positions."""
    sorted_pos = sorted(positions)
    return [sorted_pos[i+1] - sorted_pos[i] for i in range(len(sorted_pos)-1)]
```

```
Enter chromosome number: 9
Midpoint positions (n=38): [1211, 8271, 11043, 17483, 30855, 33569, 51532, 74011, 80307, 105934, 113521, 136214, 1
Inter-distance distribution (n=37): [7060, 2772, 6440, 13372, 2714, 17963, 22479, 6296, 25627, 7587, 22693, 26905,
```

3. We created an empty list to store intergenomic distance across all the chromosomes and then flattened it.

```
distance_bt看_all_chr = []
for i in df_data['chr'].unique():

    distance_bt看_all_chr.append(distances_pos(position_point(i)))

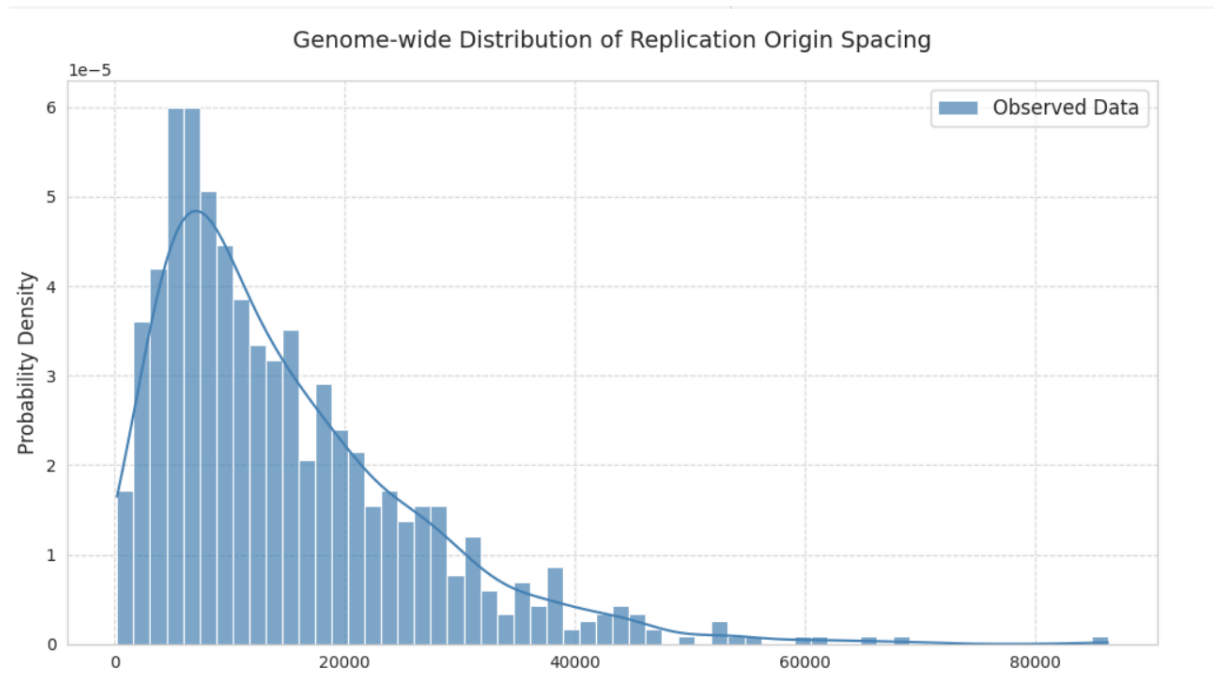
print(distance_bt看_all_chr)
print(f'Nested List len: {len(distance_bt看_all_chr)}')

flattened_distance_pos = []
for i in distance_bt看_all_chr:
    for j in i:
        flattened_distance_pos.append(j)

print(f'Flattened distances: {flattened_distance_pos}')
print(f'Flattened List len: {len(flattened_distance_pos)}')
```

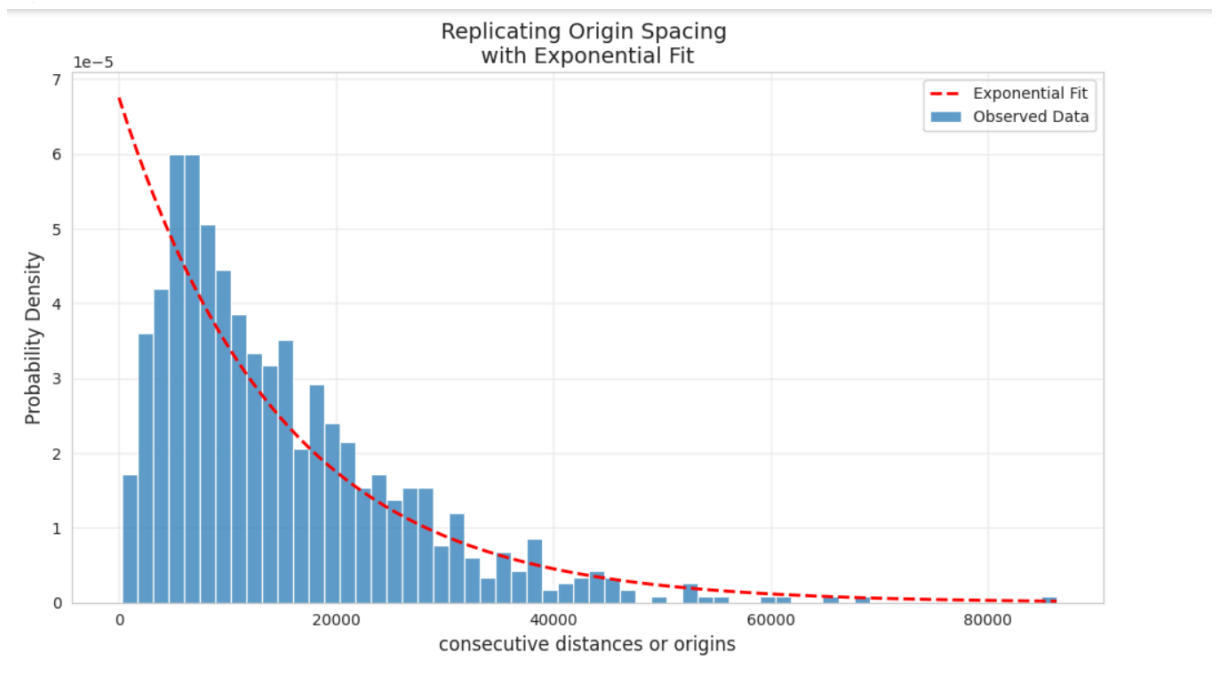
```
[[5416, 1637, 1857, 7080, 13855, 10943, 28366, 31895, 22205, 12926, 9796, 12820, 5984, 10278, 21
Nested List len: 16
Flattened distances: [5416, 1637, 1857, 7080, 13855, 10943, 28366, 31895, 22205, 12926, 9796, 12
Flattened List len: 813
```

4. To distinguish between clustered and stochastic patterns and to clarify replication regulatory principles, we statistically examined genome-wide replication origin spacing, which we visualized in a graph, which is as follows:



This graph depicts the distribution of distances between adjacent DNA replication origins throughout the genome. The probability density curve represents a theoretical distribution (such as exponential or log-normal) fitted to the data, whereas the observed data (histogram) shows empirical spacing patterns. While alignment with a log-normal model reveals regulated, non-random spacing influenced by factors like chromatin accessibility, a right-skewed distribution suggests clustering of origins in specific places (e.g., gene-dense areas). These results demonstrate how the location of the replication origin strikes a balance between fidelity and efficiency, mirroring larger mechanisms of genomic stability and regulation.

5. We have tried to test whether the replicating origins are randomly placed or whether replication origin spacing follows an exponential distribution or a random Poisson process in this investigation.



Deviations from the exponential fit would suggest that origin placement is influenced by biological limitations (such as transcriptional activity and chromatin accessibility). The histogram and curve's apparent discrepancy, if any, reveals mechanical details about the control of DNA replication. The exponential fit to genome-wide origin spacing in the graph reveals systematic deviations from randomness, suggesting replication origins are non-stochastically regulated, likely by chromatin structure or licensing factor availability.

6. **Chi-square goodness-of-fit test** - Whether replication origin spacing follows a random process or deviates because of biological regulation (such as chromatin accessibility or licensing factors) is statistically confirmed by this test.

```
chisquare_results
```



**Chi\_Sq\_Statistics**   **P-Value**

0	125.62641	0.000001
---	-----------	----------

We found that the hypothesis of random replication origin spacing is categorically rejected by the chi-square test ( $\chi^2 = 125.63$ ,  $p < 0.0001$ ), which shows that origins are not dispersed randomly throughout the genome. This controlled positioning highlights the accuracy of replication control in preserving genome integrity and most likely reflects chromatin shape, transcriptional activity, and licensing factor dynamics.

7. Determining the genomic areas' midpoints and incorporating them into a structured dataset for subsequent spatial analysis. We tried to calculate midpoints and then created a column to add the midpoints for spatial analysis.



	chr	start	end	name	othernames	status	midpoint
0	1	650	1791	ARS102	proARS102	Confirmed	1220
1	1	6136	7136	ARS102.5	proARS103	Confirmed	6636
2	1	7998	8548	ARS103	proARS103	Confirmed	8273
3	1	9775	10485	NaN	NaN	Likely	10130
4	1	16855	17565	NaN	NaN	Likely	17210

8. Now we have introduced simulating fake cells by using information on chromosome length and the number of replication origins per chromosome to model DNA replication origins on a cell.

```
[813184, 316620, 1531933, 576874, 270161, 1090940, 562643, 439888, 745751, 666816, 1078177, 924431, 784333, 1091291, 948066, 85779]
Full chr_size structure:      1  NC_001133    230218
0  2  NC_001134    813184
1  3  NC_001135    316620
2  4  NC_001136    1531933
3  5  NC_001137    576874
4  6  NC_001138    270161
5  7  NC_001139    1090940
6  8  NC_001140    562643
7  9  NC_001141    439888
8 10  NC_001142    745751
9 11  NC_001143    666816
10 12  NC_001144    1078177
11 13  NC_001145    924431
12 14  NC_001146    784333
13 15  NC_001147    1091291
14 16  NC_001148    948066
15 17  NC_001224    85779
Number of elements in chr_size[2]: 16
```

9. We tried to **simulate synthetic replication origins** by generating random genomic positions across chromosomes and calculating inter-origin distances. As a null model to compare with actual data, this simulation creates artificial inter-origin distances under a random placement hypothesis. It examines whether replication origin spacing may be explained by biological regulation.



```

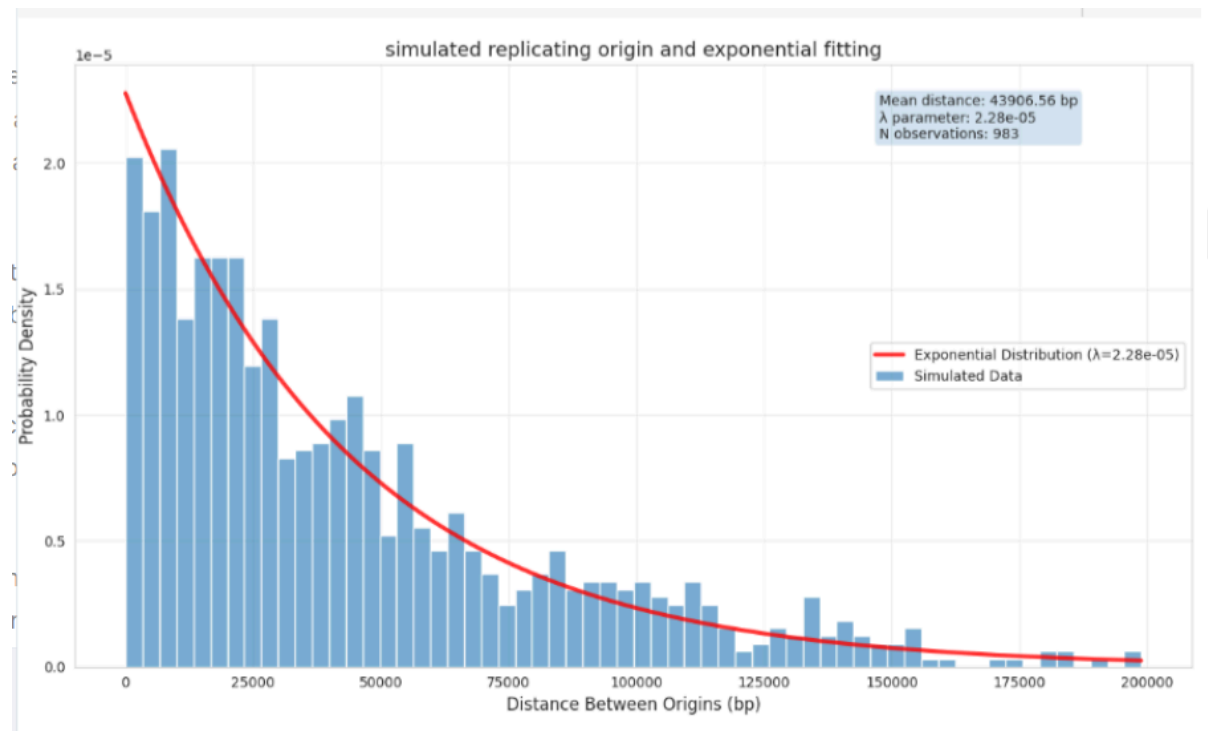
▶ #Simulate synthetic replication origins per chromosome with random placement
def simulatedcells():
    fakecells = []
    # Iterate through unique chromosomes with index
    for i, chrom in enumerate(df_data['chr'].unique()):
        # Get number of elements for this chromosome
        chrom_count = len(df_data[df_data['chr'] == chrom])

        # Generate random positions within chromosome size
        num_fake = np.random.randint(1, chrom_size_list[i] + 1, size=chrom_count)
        fakecells.append(list(num_fake))

    # Calculate distances between positions
    interdistance_fake_cells = []
    for chrom_data in fakecells:
        interdistance_fake_cells.append(position_point(chrom_data))
    # Flatten the nested list
    flatten_dist_fake_cells = []
    for sublist in interdistance_fake_cells:
        for item in sublist:
            interdistance_fake_cells.append(abs(int(item)))
    return flatten_dist_fake_cells

```

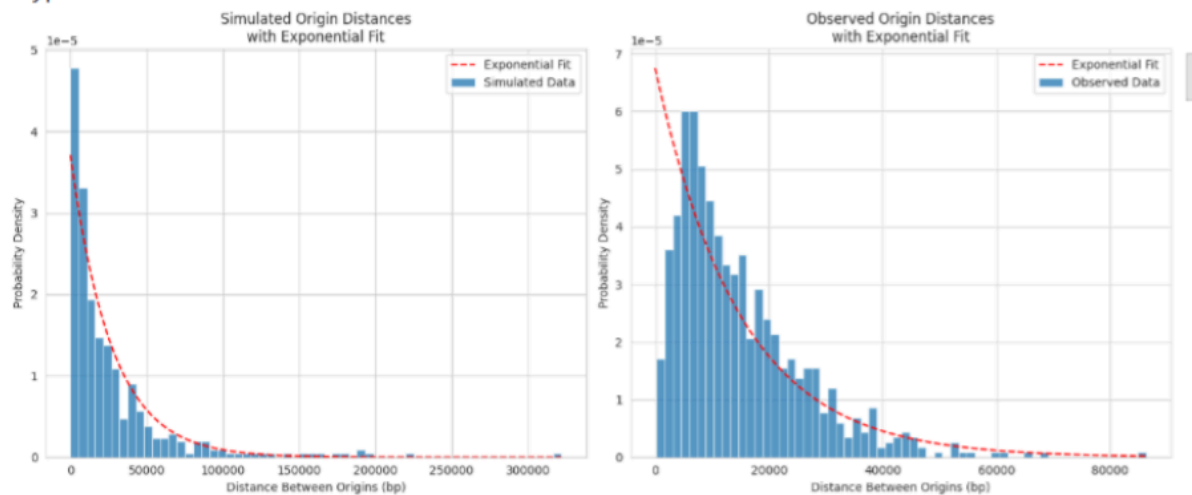
10. We have tried to generate simulated distances for `simulated_cell_confirmed()` and `simulated_cell_confirmed_likely()`. In both functions, they are generating random genome positions within chromosome sizes, sorting positions, computing consecutive origin distances, and returning a flattened list of distances.
11. Exponential fit of simulated replication origin distance - The graph contrasts an exponential distribution with parameter  $\lambda = 2.28e^{-5}$  with simulated distances between replication origins (blue histogram). The exponential curve and the simulated data closely match, indicating that replication origin distances exhibit an exponential decline pattern, which is characteristic of stochastic (random) origin placement. Given that the likelihood of greater gaps decreases exponentially, the exponential fit suggests that shorter inter-origin lengths are more likely. In genomics, replication sources are frequently clustered but stochastically dispersed, which is consistent with empirical data.



12. Comparison between Simulated and Observed Replication Origin Distances with Exponential Fits

#### Simulated Data Statistics:

```
count      394.000000
mean       26947.208122
std        37483.852463
min         0.000000
25%        5231.250000
50%        14053.000000
75%        32098.750000
max        321819.000000
dtype: float64
```



The simulated origin spacing's ability to replicate biological data is tested using the exponential fit. The simulations' lower mean distance (26,947 bp compared to about 100,000 bp observed) suggests that they may have overestimated the origin density or lacked biological restrictions. Both datasets fit exponential distributions, indicating variations in spacing patterns while confirming the notion that replication origins are stochastically distributed.

13. Chi-square test for goodness-of-fit to exponential distribution - It appears that the simulation model lacks biological realism because the simulated distances do not closely match the expected exponential distribution. Though not as much as simulations, the observed distances differ greatly from the exponential model. This suggests that basic stochastic models are unable to reflect the extra restrictions seen in real-world data, such as transcription factors and chromatin shape. The simulated data rejects the null hypothesis—that is, that they follow the predicted distribution—more forcefully than the other dataset. The departure of the observed data indicates that variables other than simple stochasticity, including epigenetic control, affect the positioning of replication origins

	Dataset	Chi-Square Statistic	P-Value
0	Simulated Data	1947.027742	0.000000
1	Observed Data	125.626410	0.000001

14. We calculated midpoints for confirmed origins. Then we calculated distances between confirmed origins for all chromosomes. Then flatten the nested list. We did this to measure the distance between the genome's verified replication origins and empirical data ready for comparison with theoretical models or simulations.
15. Now we have calculated midpoints for confirmed/likely origins for a chromosome. Then calculated the distance for all the chromosomes and flattened the distances. We did this so that we can analyse the distance between confirmed and likely.

16. We analyzed the distribution of confirmed and likely replication origin.

```
➤ Status Distribution:
status
Confirmed    410
Likely        216
Name: count, dtype: int64
```

```
Chromosome Distribution:
chr
1      21
2      46
3      23
4      71
5      28
6      21
7      51
8      34
9      27
10     36
11     31
12     38
13     44
14     40
15     59
16     56
Name: count, dtype: int64
```

17. Statistical Comparison of Observed and Fake Simulations - We developed a simulation framework in which 10,000 simulated datasets were created, with replication origins positioned randomly within each chromosome's observed genomic regions. The simulated greatest distance was recorded for every simulation, along with the calculated distances between the simulated sources. We test our hypothesis by determining whether the maximum inter-origin distance that was observed is statistically different from a random placement model. A key finding is that the high proportion (**87.05%**) suggests that replication origins are more consistently spaced than would be expected by chance. This means that the spacing of origins is controlled by biological processes rather than being entirely random.

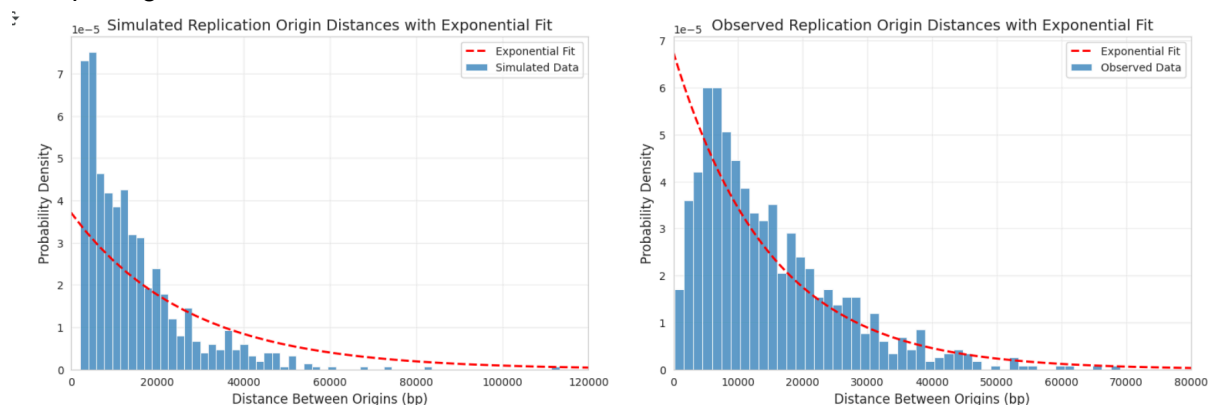
18. We created another simulation framework in which our hypothesis was that observed replication origins have smaller maximum gaps than expected if origins were randomly placed within their own observed genomic spans. The result was **100%** of simulations had larger maximum gaps than observed, which means that observed origins too close within their genomic regions than the expected.

19. Another simulation framework was created in which we hypothesized that observed replication origins were closer within their genomic regions than expected by random chance. The result was **99.82%**, which tells us that the null hypothesis is severely rejected by the **99.82%** result, which shows extreme clustering that goes beyond chance.
20. We calculated the average chromosome size, which is **2093.891435464415**. We tried to generate synthetic chromosomal position data that enforces a minimum distance between simulated blocks and establishes a minimum distance between simulated blocks. Then we have flattened it.

```
[ ] flatten_distances = new_simulation(df, chrom_size_list)
print(flatten_distances)
```

```
✓ Chrom np.int64(1) done in 14 iters
✓ Chrom np.int64(2) done in 60298 iters
✓ Chrom np.int64(3) done in 661 iters
✓ Chrom np.int64(4) done in 38364 iters
✓ Chrom np.int64(5) done in 102 iters
✓ Chrom np.int64(6) done in 62 iters
✓ Chrom np.int64(7) done in 5171 iters
✓ Chrom np.int64(8) done in 199 iters
✓ Chrom np.int64(9) done in 958 iters
✓ Chrom np.int64(10) done in 570 iters
✓ Chrom np.int64(11) done in 191 iters
✓ Chrom np.int64(12) done in 12771 iters
✓ Chrom np.int64(13) done in 6364 iters
✓ Chrom np.int64(14) done in 1731 iters
✓ Chrom np.int64(15) done in 25051 iters
✓ Chrom np.int64(16) done in 2340 iters
[11609, 4970, 8644, 15955, 2764, 2398, 17671, 5217, 4836, 4987, 45447, 28965, 7575, 10444, 2542, 9286, 2873, 461]
```

21. In order to confirm whether the algorithmically generated simulated replication origin distances exhibit the same statistical patterns as actual biological data, we visualised a graph that shows that the exponential distribution in actual biological data closely matches the simulated replication origin distances. This demonstrates that the simulation accurately reproduces the exponentially distributed natural spacing pattern of replication origins. The general statistical behavior is consistent with real-world observations, even while a minimum spacing between simulated blocks is enforced.



22. Chi-Square test - The simulation's ability to reproduce this pattern is validated by the simulated distances' statistical alignment with an exponential distribution ( $p > 0.05$ ). Real replication sources might have a more complicated distribution or other biological elements that the simulation was unable to account for, as indicated by the observed data's notable

departure from the exponential model.

	Dataset	Chi-Square Statistic	P-Value
0	Simulated Data	76.541755	0.062036
1	Observed Data	125.626179	0.000001

## Candida Glabrata

1. We imported the essential libraries, cleaned the data and checked for the null value. Then extracted the chromosome data with the genome label. Here is the dataset.

```
can_data = pd.read_csv("Candida_glabrata_data.csv")
can_data.head()
```

	Index	ID	Strain	Chromosome	Strand	PositionRange	Start	End	Length	Ratio
0	1	eori001700001	Candida_glabrata_strain_CBS138	chr A	-	728-1042	728.0	1042.0	314.0	0.363057
1	2	eori001700002	Candida_glabrata_strain_CBS138	chr A	-	17150-17397	17150.0	17397.0	247.0	0.287449
2	3	eori001700003	Candida_glabrata_strain_CBS138	chr A	-	155026-155370	155026.0	155370.0	344.0	0.279070
3	4	eori001700004	Candida_glabrata_strain_CBS138	chr A	-	187390-187631	187390.0	187631.0	241.0	0.344398
4	5	eori001700005	Candida_glabrata_strain_CBS138	chr A	-	245879-246115	245879.0	246115.0	236.0	0.322034

2. Then we calculated the midpoints and interpoint distances of the genome in order to analyze the spatial arrangement.

Enter Chromosome Letter (e.g., A, B): B

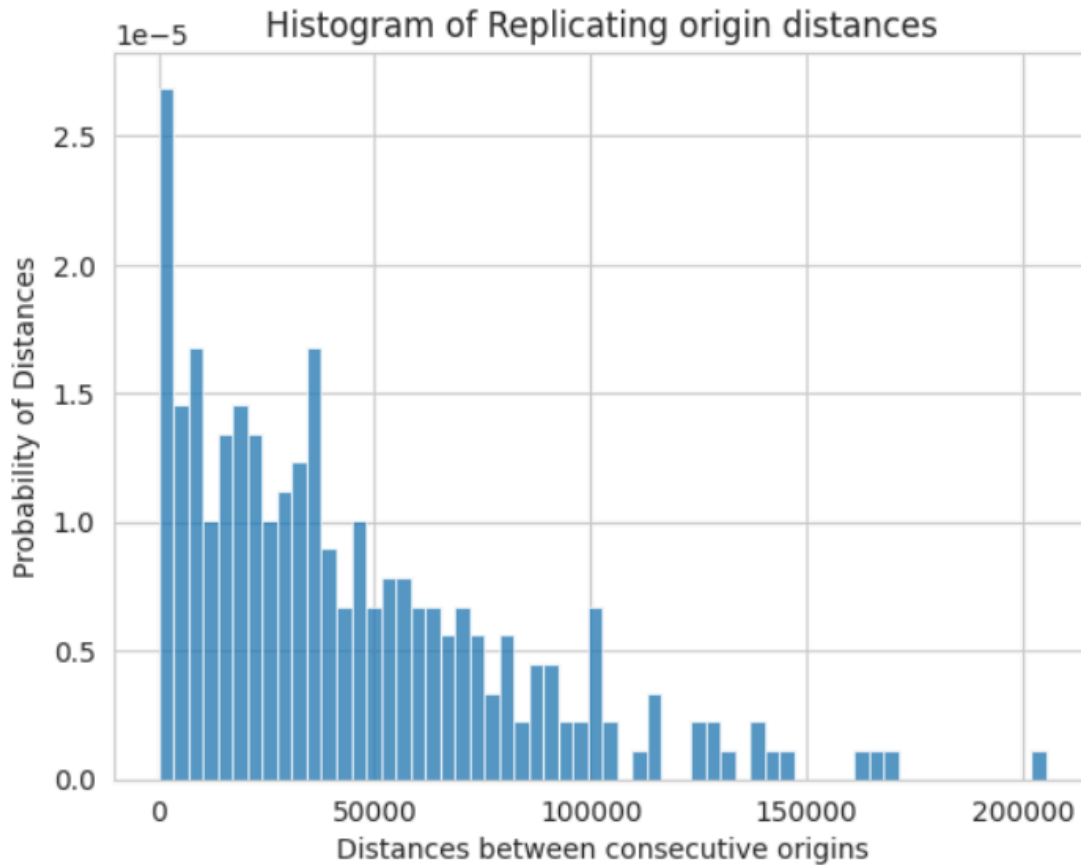
Point Sized Positions of the Origin Points of Chromosome B: [12645.0, 35058.5, 123241.0, 289381.5, 323386.5, 341838.5, 373071.0, 457186.0]  
Distance Between Point Sized Positions: [22413.5, 88182.5, 166140.5, 34005.0, 18452.0, 31232.5, 84115.0]

3. Then we tried to compute midpoints and consecutive distances between the chromosomes and created nested lists and flattened it.

```
Chromosome chr A: Midpoints = [885.0, 17273.5, 155198.0, 187510.5, 245997.0, 267145.0, 325438.5, 359988.5, 387486.5], Distances = [16388.5, 137924.5, 32312.5, 58486.5, 21148.0, 58293.5, 34550.0, 27498.0]
Chromosome chr B: Midpoints = [12645.0, 35058.5, 123241.0, 289381.5, 323386.5, 341838.5, 373071.0, 457186.0], Distances = [22413.5, 88182.5, 166140.5, 34005.0, 18452.0, 31232.5, 84115.0]
Chromosome chr C: Midpoints = [1533.5, 65247.5, 100287.5, 108127.0, 111053.0, 112711.5, 121801.5, 206518.5, 235827.0, 284330.0, 348128.5, 405937.5]
Chromosome chr D: Midpoints = [314.5, 2555.0, 13766.5, 49167.5, 57996.0, 140175.5, 188120.5, 202194.0, 204940.5, 269249.0, 289517.5, 316768.5]
Chromosome chr E: Midpoints = [725.5, 12106.0, 47423.5, 65596.0, 70782.5, 106421.0, 137974.5, 159179.5, 183052.0, 219239.0, 242808.5, 257668.5]
Chromosome chr F: Midpoints = [519.0, 116625.5, 135829.0, 148339.5, 199908.0, 286020.0, 340383.0, 409291.0, 418402.5, 444484.0, 471876.0, 484977.5]
Chromosome chr G: Midpoints = [549.5, 1401.5, 1963.5, 31146.5, 62446.0, 62772.5, 67258.5, 230143.5, 399460.0, 405817.0, 486232.0, 561694.5, 603115.5]
Chromosome chr H: Midpoints = [711.5, 2086.5, 57035.0, 77099.0, 133744.5, 155890.5, 160242.0, 185925.5, 232817.5, 332306.0, 340884.0, 390604.0, 400604.0]
Chromosome chr I: Midpoints = [310.0, 1692.0, 55968.0, 117745.0, 162658.0, 217376.0, 343989.5, 369712.0, 448984.5, 454679.0, 500671.0, 566910.0, 566910.0]
Chromosome chr J: Midpoints = [811.5, 23851.5, 40650.0, 101599.0, 108328.0, 109970.0, 180140.5, 269287.5, 296246.5, 363650.5, 364605.0, 378245.5, 378245.5]
Chromosome chr K: Midpoints = [501.0, 1367.5, 2821.0, 37522.5, 38450.0, 82724.5, 185488.0, 231950.5, 233682.0, 243810.5, 254661.0, 314333.5, 391115.5]
Chromosome chr L: Midpoints = [231.5, 7631.0, 53209.5, 110292.5, 147514.0, 181103.5, 215478.5, 290875.5, 359664.5, 391810.5, 399201.5, 437002.0, 437002.0]
Chromosome chr M: Midpoints = [8211.5, 101041.0, 229558.5, 309585.5, 452603.5, 483799.0, 524461.0, 595630.0, 667588.5, 707939.5, 811453.0, 829561.0, 829561.0]

Nested Distances: [[16388.5, 137924.5, 32312.5, 58486.5, 21148.0, 58293.5, 34550.0, 27498.0], [22413.5, 88182.5, 166140.5, 34005.0, 18452.0, 31232.5, 84115.0], [1533.5, 65247.5, 100287.5, 108127.0, 111053.0, 112711.5, 121801.5, 206518.5, 235827.0, 284330.0, 348128.5, 405937.5], [314.5, 2555.0, 13766.5, 49167.5, 57996.0, 140175.5, 188120.5, 202194.0, 204940.5, 269249.0, 289517.5, 316768.5], [725.5, 12106.0, 47423.5, 65596.0, 70782.5, 106421.0, 137974.5, 159179.5, 183052.0, 219239.0, 242808.5, 257668.5], [519.0, 116625.5, 135829.0, 148339.5, 199908.0, 286020.0, 340383.0, 409291.0, 418402.5, 444484.0, 471876.0, 484977.5], [549.5, 1401.5, 1963.5, 31146.5, 62446.0, 62772.5, 67258.5, 230143.5, 399460.0, 405817.0, 486232.0, 561694.5, 603115.5], [711.5, 2086.5, 57035.0, 77099.0, 133744.5, 155890.5, 160242.0, 185925.5, 232817.5, 332306.0, 340884.0, 390604.0, 400604.0], [310.0, 1692.0, 55968.0, 117745.0, 162658.0, 217376.0, 343989.5, 369712.0, 448984.5, 454679.0, 500671.0, 566910.0, 566910.0], [811.5, 23851.5, 40650.0, 101599.0, 108328.0, 109970.0, 180140.5, 269287.5, 296246.5, 363650.5, 364605.0, 378245.5, 378245.5], [501.0, 1367.5, 2821.0, 37522.5, 38450.0, 82724.5, 185488.0, 231950.5, 233682.0, 243810.5, 254661.0, 314333.5, 391115.5], [231.5, 7631.0, 53209.5, 110292.5, 147514.0, 181103.5, 215478.5, 290875.5, 359664.5, 391810.5, 399201.5, 437002.0, 437002.0], [8211.5, 101041.0, 229558.5, 309585.5, 452603.5, 483799.0, 524461.0, 595630.0, 667588.5, 707939.5, 811453.0, 829561.0, 829561.0]]
Flattened Distances: [16388.5, 137924.5, 32312.5, 58486.5, 21148.0, 58293.5, 34550.0, 27498.0, 22413.5, 88182.5, 166140.5, 34005.0, 18452.0, 31232.5, 84115.0, 1533.5, 65247.5, 100287.5, 108127.0, 111053.0, 112711.5, 121801.5, 206518.5, 235827.0, 284330.0, 348128.5, 405937.5, 314.5, 2555.0, 13766.5, 49167.5, 57996.0, 140175.5, 188120.5, 202194.0, 204940.5, 269249.0, 289517.5, 316768.5, 725.5, 12106.0, 47423.5, 65596.0, 70782.5, 106421.0, 137974.5, 159179.5, 183052.0, 219239.0, 242808.5, 257668.5, 519.0, 116625.5, 135829.0, 148339.5, 199908.0, 286020.0, 340383.0, 409291.0, 418402.5, 444484.0, 471876.0, 484977.5, 549.5, 1401.5, 1963.5, 31146.5, 62446.0, 62772.5, 67258.5, 230143.5, 399460.0, 405817.0, 486232.0, 561694.5, 603115.5, 711.5, 2086.5, 57035.0, 77099.0, 133744.5, 155890.5, 160242.0, 185925.5, 232817.5, 332306.0, 340884.0, 390604.0, 400604.0, 310.0, 1692.0, 55968.0, 117745.0, 162658.0, 217376.0, 343989.5, 369712.0, 448984.5, 454679.0, 500671.0, 566910.0, 566910.0, 811.5, 23851.5, 40650.0, 101599.0, 108328.0, 109970.0, 180140.5, 269287.5, 296246.5, 363650.5, 364605.0, 378245.5, 378245.5, 501.0, 1367.5, 2821.0, 37522.5, 38450.0, 82724.5, 185488.0, 231950.5, 233682.0, 243810.5, 254661.0, 314333.5, 391115.5, 391115.5, 231.5, 7631.0, 53209.5, 110292.5, 147514.0, 181103.5, 215478.5, 290875.5, 359664.5, 391810.5, 399201.5, 437002.0, 437002.0, 8211.5, 101041.0, 229558.5, 309585.5, 452603.5, 483799.0, 524461.0, 595630.0, 667588.5, 707939.5, 811453.0, 829561.0, 829561.0]
Total Distances: 261
```

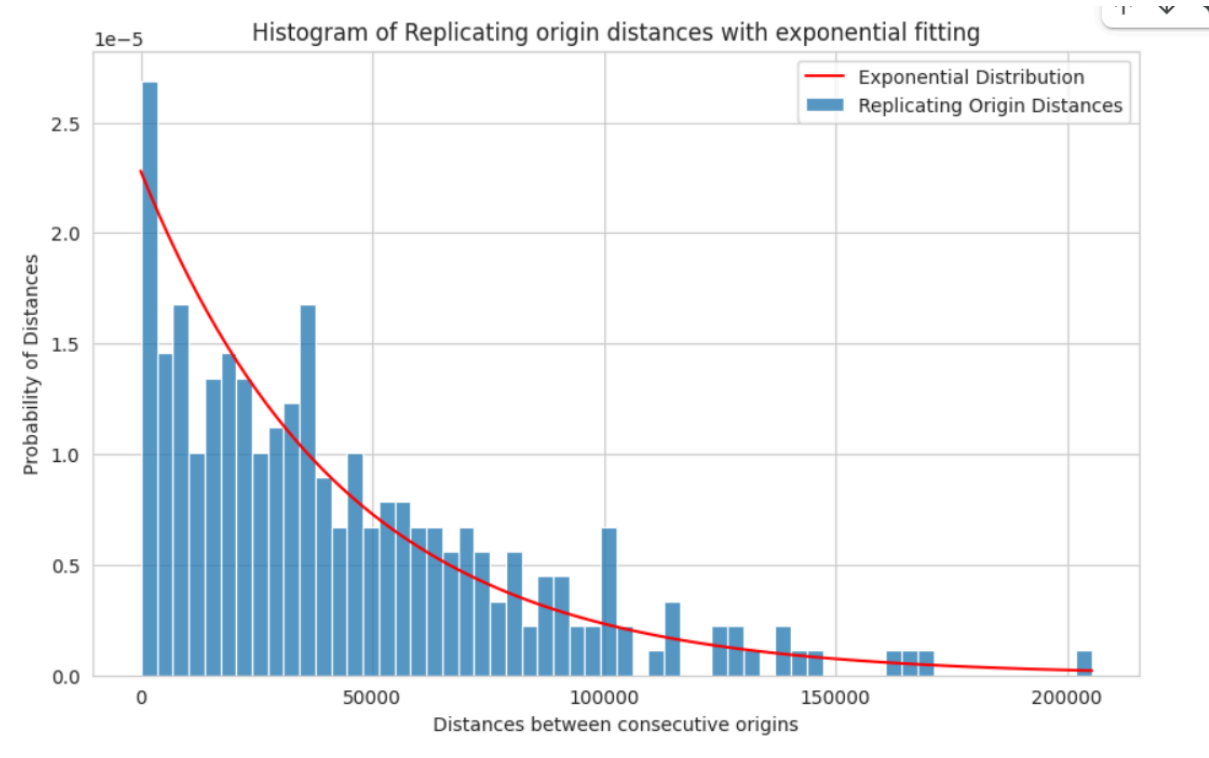
4. Probability Density Distribution of Distances Between Consecutive DNA Replication Origins. The graph highlights how replication origins are distributed across a genome. We find that the shorter distances have a higher probability. While the longer distances are increasingly rare. This is consistent with biological predictions that replication origins will be non-randomly distributed, frequently clustered, or controlled by genomic limitations.



5. We tried to generate a **histogram of distances between consecutive DNA replication origins** and tried to fit an **exponential distribution** to test if their spacing follows a random pattern. In this graph, the red curve represents a fitted exponential distribution, while the histogram represents the measured distances between DNA replication origins. It determines whether origin spacing exhibits a random pattern, which is essential for comprehending the dynamics of genomic replication. Stochastic spacing is supported by



alignment with the exponential model; deviations imply biological restrictions.



6. Using a chi-square goodness-of-fit model - we did this to verify whether the observed replication origin distances exhibit random spacing, as indicated by an exponential distribution. A high P-Value ( $>0.05$ ) means that the observed replication origin distances and the exponential distribution do not differ significantly. In conclusion, an exponential distribution is supported by the data, indicating that replication origins are randomly distributed.

	chi square statistics	P value
0	43.930714	0.928402

7. We calculated the midpoints and integrated them in the dataset. Then it is flattened so that we can analyze spatially.

	Index	ID	Strain	Chromosome	Strand	PositionRange	Start	End	Length	Ratio	start	end	mid_point
0	1	eori001700001	Candida_glabrata_strain_CBS138	chr A	-	728-1042	728.0	1042.0	314.0	0.363057	728	1042	885.0
1	2	eori001700002	Candida_glabrata_strain_CBS138	chr A	-	17150-17397	17150.0	17397.0	247.0	0.287449	17150	17397	17273.5
2	3	eori001700003	Candida_glabrata_strain_CBS138	chr A	-	155026-155370	155026.0	155370.0	344.0	0.279070	155026	155370	155198.0
3	4	eori001700004	Candida_glabrata_strain_CBS138	chr A	-	187390-187631	187390.0	187631.0	241.0	0.344398	187390	187631	187510.5
4	5	eori001700005	Candida_glabrata_strain_CBS138	chr A	-	245879-246115	245879.0	246115.0	236.0	0.322034	245879	246115	245997.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
270	271	eori001700271	Candida_glabrata_strain_CBS138	chr M	-	853506-853752	853506.0	853752.0	246.0	0.345528	853506	853752	853629.0
271	272	eori001700272	Candida_glabrata_strain_CBS138	chr M	-	977894-978127	977894.0	978127.0	233.0	0.334764	977894	978127	978010.5
272	273	eori001700273	Candida_glabrata_strain_CBS138	chr M	-	1183183-1183507	1183183.0	1183507.0	324.0	0.271605	1183183	1183507	1183345.0
273	274	eori001700274	Candida_glabrata_strain_CBS138	chr M	-	1199193-1199423	1199193.0	1199423.0	230.0	0.386957	1199193	1199423	1199308.0
274	275	eori001700275	Candida_glabrata_strain_CBS138	chr M	-	1272920-1273121	1272920.0	1273121.0	201.0	0.343284	1272920	1273121	1273020.5

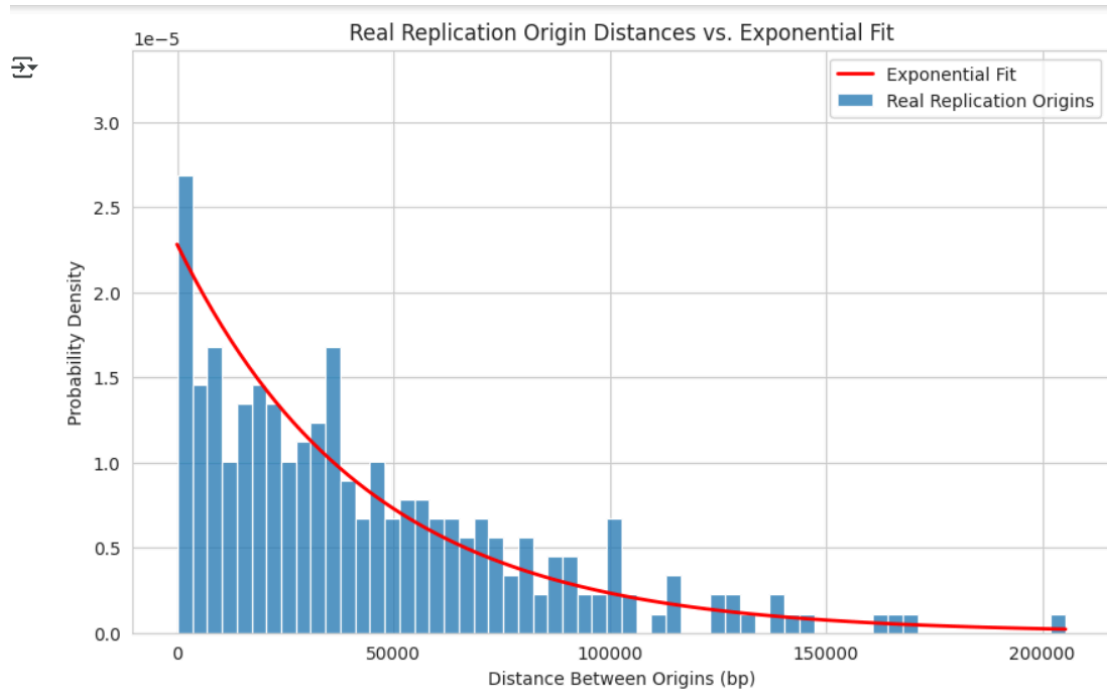
275 rows × 13 columns

8. We extracted chromosomes, then calculated the length. After that we sorted out and assigned for index. Then we computed the average chromosome size.

```
chromosome_size = list(chromosome_lengths[["Chromosome", "Length", "Index"]].itertuples(index=False, name=None))  
print(chromosome_size)
```

```
[('A', 387599.0, 0), ('B', 457317.0, 1), ('C', 501185.0, 2), ('D', 450477.0, 3), ('E', 641491.0, 4), ('F', 897576.0, 5), ('G', 910857.0, 6), ('H', 955483.0, 7), ('I', 1088483.0, 8)]
```

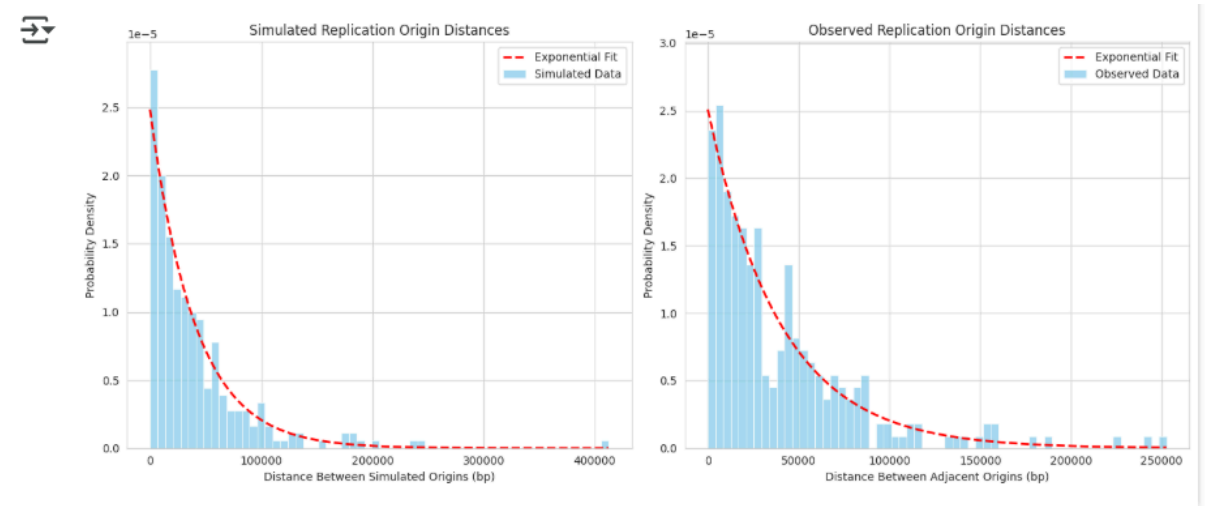
9. We have tried to generate synthetic data for all the chromosomes and created random positions. Then, computed distances to get the distance between simulated position. Then flattened it. We did this so that we can find out if observed data follows any pattern.
10. We have tried to verify whether the separation between true replication origins is random exponential. The graph indicates random spacing of origins if the histogram matches the exponential curve. Biological limitations, such as regulated clumping or exclusion zones, are implied by significant departures, such as peaks or gaps.



We see that the observed distances between replication origins closely follow an exponential distribution. This means that replication origins are likely spaced randomly across the genome, with no strong biological constraints dictating their placement.




11. We have tried to visualize the comparison between observed and simulated replication origin distances. The point of this step is to test the hypothesis that if the replicating origins

are randomly placed.



The graph on the left side shows that the red-dashed line is quite close to the histogram, which reveals that the simulated data follows exponential distribution. This means that simulated data can replicate randomly. While on the right side, the observed data deviates from the exponential model, which means that the observed data are non-randomly placed.

12. Chi-Square Goodness-of-Fit Test for Simulated and Observed Replication Origin Distance Distributions - We tried to generate a histogram of distances from synthetic replication origins (`flatten_dist_fake_cell`). Then calculated expected frequencies using an exponential distribution. Then adjusted expected frequencies to match the total observed count. After that we conducted a chi-square test.

	Dataset	Chi-Square Statistic	P-Value	
0	Fake Cell Values	615.153284	1.433628e-93	
1	Replication Origin Distances	43.930714	9.284020e-01	

We found that in fake cell values, the discrepancy is quite huge between the expected frequency and the observed one. The p-value is quite low, which rejects the null hypothesis here. This means that the simulated distances do not follow the exponential distribution, and they do not replicate randomly. On the other side, the replicating origin distances have low deviation between the expected and observed frequency. Due to the high p-value, the null hypothesis cannot be rejected. So, this means that the observed distances follow the exponential distribution and favor randomness.

13. We calculated the midpoints and distances between consecutive midpoints. Then , we flattened it. We have created a function that computes the distances between successive sites, generates random integers within chromosomal lengths to imitate genomic positions for each chromosome, and outputs distances. In order to ensure sorted

positions and legitimate spacing measurements, it attempts to generate synthetic data that closely resembles actual replication origins for comparison.

14. We used Monte Carlo simulation to find out if simulated distances exceed more than the observed data. We have tried to simulate 100,000 simulations to compare the maximum simulated genomic position and observed maximum. We did this in order to verify whether the simulation overestimates the maximum gap or produces biologically realistic distances.

```
#Monte Carlo Simulation to Assess Simulated Distance Maximums Against Observed Data
count = 0
max_counter = 0
num_of_sim = 100000

while count < num_of_sim:
    flattend_dist_fake_cell = simulated_cell_confirmed()
    fake_max = max(flattend_dist_fake_cell)

    if fake_max > obv_max:
        max_counter += 1
    count += 1

percent = (max_counter / num_of_sim) * 100
print(f'Percentage Where Simulated Distances Exceed Observed Maximum: {percent}%')
```

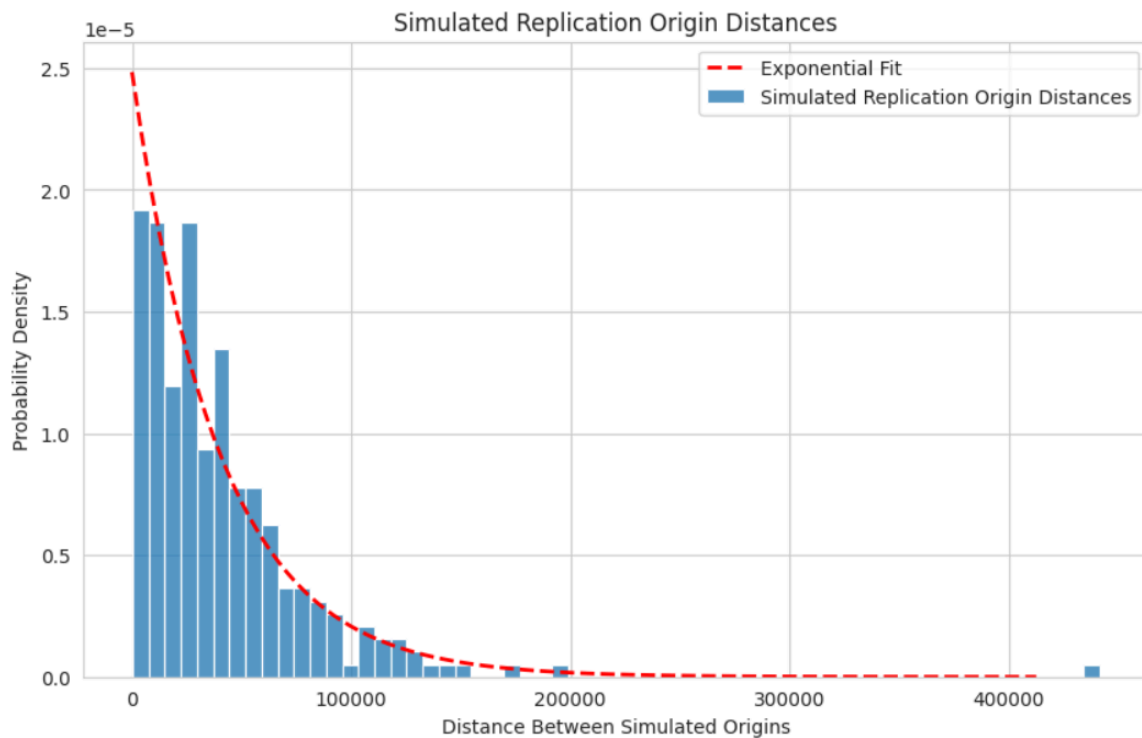
Percentage Where Simulated Distances Exceed Observed Maximum: 85.33%

We see that there is a high percentage of the simulated distances that exceed the observed maximum, which means that the simulation producing large distances creates a mismatch with the observed data.

15. We computed the average of the chromosome lengths. In order to ensure that the minimum distance between successive positions is at least the average length shown in the original data (clean\_df), the algorithm simulates creating random genomic positions ("fake cells") for each chromosome. Here is a brief synopsis: goal is to produce synthetic data with fake genomic locations on each chromosome that are no closer than the actual dataset's average length. We continuously created random integers down the length of each chromosome until all successive gaps between sorted positions were less than or equal to the average. We gathered distances in pairs across chromosomes across fictitious places. Then we got a flattened list of the distances between successive fake cells that satisfy the minimum spacing requirement as the output.

```
Loop 1 completed after 1 iterations.
Loop 2 completed after 1 iterations.
Loop 3 completed after 2 iterations.
Loop 4 completed after 1 iterations.
Loop 5 completed after 1 iterations.
Loop 6 completed after 1 iterations.
Loop 7 completed after 2 iterations.
Loop 8 completed after 1 iterations.
Loop 9 completed after 2 iterations.
Loop 10 completed after 1 iterations.
Loop 11 completed after 1 iterations.
Loop 12 completed after 2 iterations.
Loop 13 completed after 1 iterations.
[np.int64(25880), np.int64(53393), np.int64(42558), np.int64(59721), np.int64(11274), np.int64(113726), np.int64(28387),
```

16. We have tried to visualize a graph in which we are trying to see if the simulated replicating origins are spaced apart and if they are following exponential distribution.



In this graph, we see that there are many shorter distances. While there are quite less longer distances, the curve is following the exponential distribution.

17. In order to check if the stimulated and observed data matched the exponential fit, we used the chi-square test. We see that the chi-square statistic is quite high and the p-value is quite low. This indicates that the fake does not follow the exponential distribution, as the simulation distances are different from our expectation. On the other hand, the observed replicating origin sees that there is a low chi-square statistic and a high p-value. This indicates that it follows the exponential distribution. So, we can say that simulated data fails to replicate the origin distance, and real DNA data follows an exponential pattern and clusters closely.

	Dataset	Chi-Square Statistic	P-Value
0	Fake Cell Values	1115.406810	7.486026e-195
1	Replication Origin Distances	43.930714	9.284020e-01

## **Discussion**

In this, we tried to find the spatial arrangement of DNA replicating origins of *Saccharomyces cerevisiae* and *Candida glabrata* so that we can know if the replicating origins spaces follow biological mechanisms or randomness. These are some of the key points that we have observed:

1. Non-random spaces in *Saccharomyces cerevisiae* - We saw that the chi-square test rejects the null hypothesis of origins being placed randomly. We also saw that the observed deviations from the exponential distribution suggest that there is some biological mechanism going on. Also, the random placement simulations were unable to reproduce real-world patterns, which underscores the shortcomings of basic stochastic models for this species.
2. Random spaces in *Candida Glabrata*—We find that the observed replication distances that were found closely matched an exponential distribution, indicating support for a stochastic model. We can also say that because of its randomness, there is less evolutionary pressure on *Candida glabrata* for replication.
3. Limitation for the simulation: the fake dataset continuously fails to match with observed trends. Observed trends were repeatedly not matched by synthetic datasets. For instance, 85% of simulations overestimated the maximal gaps, while the mean of simulated distances in *S. cerevisiae* was significantly lower (26,947 bp compared to around 100,000 bp found). This emphasises how future models must incorporate biological restrictions (such as chromatin shape and transcriptional activity) in order to increase realism.

## **Results**

1. Non-random origin spacing in *Saccharomyces Cerevisiae*
  - Using chi-square, we found that the observed data have deviated from an exponential distribution ( $\chi^2 = 125.63$ ,  $p < 0.0001$ ), which rejects the hypothesis that the origins were randomly placed. Systematic deviations indicate that there is an impact of the biological mechanism.
  - Comparison between stimulated data and observed data—random distances did not replicate biological patterns ( $\chi^2 = 1115.41$ ,  $p \approx 0$ ). 85.33% of simulated datasets had higher maximum inter-origin gaps than observed, implying tighter clustering in real data. The

simulated mean inter-origin distance (26,947 bp) was much lower than the observed value (~100,000 bp), indicating an overestimation of origin density.

## 2. Stochastic replicating origin in *Candida Glabrata*

- Using chi-square, we found that observed distances follow an exponential distribution ( $\chi^2=43.93, p=0.93$ ), which follows the hypothesis that origins are placed randomly
- There is no considerable divergence from randomisation, indicating low biological restrictions.
- Simulated data follows an exponential distribution ( $\chi^2=615.15, p=0$ ). The observed data did not indicate any significant variation

## 3. Distribution of inter-origin spaces

- In *Saccharomyces cerevisiae*, the observed data exhibited right-skewed clustering. The biological restrictions were not captured by the simulated data, which understated distances.
- In *Candida Glabrata*, the observed data and stimulated data follow an exponential distribution.

## 4. Clustering of spaces

- In *Saccharomyces cerevisiae*, we find that extreme clustering was around 99%.
- In *Candida Glabrata*, we find that there is a uniform distribution in the chromosomes, as it follows the stochastic placement

## **Conclusion**

This study provides important insights into how evolutionary constraints impact genomic architecture by exposing key variations between *Saccharomyces cerevisiae* and *Candida glabrata* in the spatial structure of DNA replication origins. A more thorough summary of the results and their consequences may be found below.

1. In *Saccharomyces cerevisiae*, we see that there is rejection of randomness, which indicates that the replication origins of that organism are placed strategically. The deviation in the exponential models tells us that the origins are located close to the chromatin region so that the replication takes place at safer sites. Nonrandomness reduces error that could destabilise the genome.
2. The high p-value in *Candida Glabrata* indicates that there is a random and exponential distribution in *Candida Glabrata*. This causes the organism to have less pressure to replicate. Moreover, random replication allows us to tolerate the error in replication.

3. We saw that there is poor fit in synthetic data in *Saccharomyces cerevisiae*, which indicates the missing biological constraints.
4. We observed that a prime example of "precision engineering" for survival in changing conditions is *S. cerevisiae*. The "opportunistic randomness" that *C. glabrata* embraces puts adaptability ahead of efficiency.

## **REFERENCE**

1. [https://en.wikipedia.org/wiki/Saccharomyces\\_cerevisiae](https://en.wikipedia.org/wiki/Saccharomyces_cerevisiae)
2. <https://www.healthline.com/health/candida-glabrata>