MODELLING RIBOSOMAL DNA COPY NUMBER DYNAMICS

Ivan Alejandro Hidalgo Castellanos

Department of Biology
The University of Auckland
Supervisors: Dr. Austen Ganley and Dr. Nobuto Takeuchi.

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Biological Sciences, The University of Auckland, 2022.

Abstract

The ribosomal DNA genes (rDNA) form one of the genome's most conserved families of genes. They are organised in clusters of tandemly repeated units, and this repetitive nature and its high transcription rates makes the rDNA cluster a region of instability. In particular, the high transcription rate makes the cluster vulnerable to collisions between the transcription and replication machineries. The resulting instability is regulated by a mechanism that can increase recombination between rDNA units. However, recombination can occur between misaligned units, with this unequal recombination producing copy number variation. Such copy number variation has been reported across species, organisms, populations and even tissues. In this project, I wanted to evaluate how rDNA copy number distribution is influenced by the dynamics of selective pressure on copy number, the type of recombination event, and the length of misalignment produced during unequal recombination. I chose to evaluate this in haploid *S. cerevisiae* as rDNA dynamics are best characterized in this species. I developed a discrete-generation model with two steps to model these dynamics. The steps are selection and unequal non-reciprocal recombination. Probability distribution functions that describe misalignment probabilities for duplications and deletion events during non-reciprocal recombination were estimated using data from previous studies and then used in the model. Tests were made to ensure the model implementation works as expected. This model was supplemented by experimental determination of rDNA copy number distribution in a haploid *S. cerevisiae* population using ddPCR quantification. The measured distribution was then used to fit the model and compare different model versions. Overall, the assessed model had a reasonable fit to the experimental data, although improvements in the functions that describe selection and misalignment during deletion events may be able to provide a better fit. Surprisingly, under the evaluated conditions there were no improvements in models that allow recombination rate to change with different copy numbers even though this has been experimentally observed to occur. Further testing will be required to establish the cause of this result.

Acknowledgements

I would like to express my special thanks to everyone who made this research project: my parents, Ivan Dario and Nydia Marcela, my brother German, my friends, and my lab mates. Austen and Sylvie.

Table of Contents

List of Figures

List of Tables

List of Equations

Chapter 1. GENERAL INTRODUCTION

1.1. RIBOSOMAL DNA (rDNA)

The ribosomal DNA genes (rDNA) form one of the most conserved families of genes. These genes encode ribosomal RNAs, an important structural and catalytic component of the ribosome. rDNA loci are organised in arrays of tandemly repeated copies of genes that can be in different chromosomal locations. Despite there being multiple copies, they show a high degree of similarity between each other. Each rDNA gene is composed of coding and non-coding sequences. Coding sequences are separated by non-coding region called intergenic spacers (IGS). The two coding sequences in the rDNA of *S. cerevisiae* are the 35S and the 5S, each separated by different IGS regions. RNA polymerase I transcribes the 35S rRNA coding region, which is processed to produce the mature rRNAs 18S, 5.8S and 26S rRNA. The 5S rRNA gene is transcribed independently by RNA polymerase III. The IGS includes functional sequences such as a replication origin (autonomously replicating sequence; rARS) and a replication fork barrier (RFB). The RFB is a sequence that is bound by proteins to form a complex that prevents collision between the transcription and replication machineries (Kobayashi, 2011) (Fig. 1.1.1).

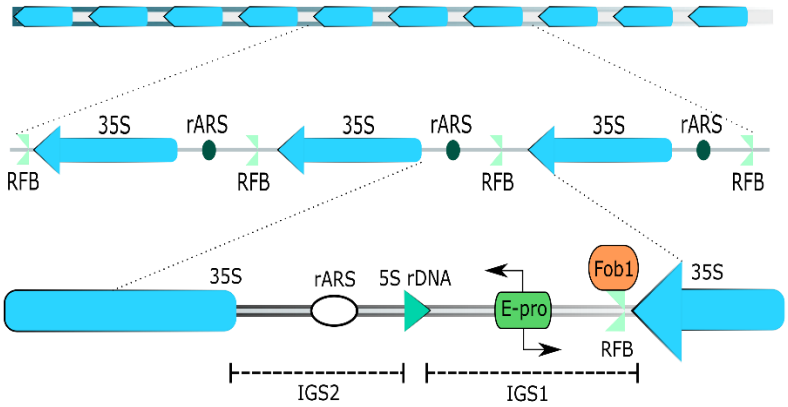


Figure 1.1.1. Schematic of the rDNA repeats in *S. cerevisiae*. The top part shows a head-to-tail tandem organization of the rDNA repeats represented by blue arrows. The middle part shows a zoomed in version of the repeats, with the 35S coding region and non-coding sequences. The bottom part shows a more detailed view of the non-coding region. Each 35S rRNA region (blue) is separated by two intergenic space regions (IGS1 and IGS2). In IGS1, a replication fork barrier site (RFB) forms a complex with the Fob1 protein. IGS1 also contains a bidirectional promotor E-pro involved in rDNA repeat number regulation. IGS2 contains an origin of replication, rARS.

1.2. rDNA COPY NUMBER VARIATION

Ribosomal DNA loci have high copy number variability. Copy number variations have been reported across multiple organisms, with copies ranging from 28 to 26,048 (Lofgren et al., 2019; Prokopowich et al., 2003). Evidence suggests that each species has a characteristic copy number (Rosato et al., 2017; West et al., 2014), and a correlation between genome size and rDNA copy number has been found (Prokopowich et al., 2003). Copy number variation has also been between individuals in a population (Lofgren et al., 2019; Porokhovnik & Lyapunova, 2019; Rosato et al., 2017; West et al., 2014). Variation in copy number can even occur within an organism, with humans and mice reported to have variation across tissues (Wang & Lemos, 2017; Xu et al., 2017).

Phenotypic and physiological consequences have been associated with rDNA copy number variation. In *S. cerevisiae*, low copy numbers and instability in the rDNA cluster have been associated with a shorter lifespan (Ganley & Kobayashi, 2014; Lindstrom et al., 2011; Saka et al., 2013). In *Drosophila*, a reduction of rDNA copies during ageing has been reported (Lu et al., 2018), and copy number variation during ageing has also been found in humans, with negative correlations between ageing and copy number in some human tissues and for some

rDNA sequences (Zafiropoulos et al., 2005). Conversely, no significant difference in the mean rDNA copy number between young and elderly individuals was reported (Malinovskaya et al., 2018). Copy number variation has also been linked to cancer in humans, with a high number of 5.8S and 18S rRNA gene copies associated with an increased risk of lung cancer (Hosgood et al., 2019), and in breast cancer, both a low and a high rDNA copy number has been reported (Valori et al., 2020; Wang & Lemos, 2017). Furthermore, in mice, evidence has linked low rDNA copy number with leukaemia (Xu et al., 2017). Neurological disorders also have been linked to copy number variation, including that some individuals with schizophrenia showed a higher number of rDNA copies than healthy controls (Chestkov et al., 2018).

1.3. RIBOSOMAL DNA IS A REGION OF GENOME INSTABILITY AND MECHANISM THAT MAINTAIN STABILITY

Ribosomal DNA loci form one of the most unstable regions in the genome. High rRNA transcription rates increase the probability of collision between the replication and transcription machineries (Brambati et al., 2015). Moreover, the high transcriptions rates are associated with chemical and structural changes in the DNA (Kim & Jinks-Robertson, 2012) that increase the mutation rates in the rDNA loci. As a countermeasure, a significant percentage of the rDNA copies are not transcribed (McStay and Grummt 2008). Studies have shown that areas with low transcription rates in the rDNA loci are important to maintain stability in the region and prevent damage to inactive copies (Kobayashi, 2014). The RFB is also involved in preventing instability in the rDNA cluster as it forms a DNA-protein complex that prevents collision between the replication and the transcription machineries (Brewer et al., 1992; Kobayashi et al., 1998). In the budding yeast *Saccharomyces cerevisiae* the RFB-associated protein Fob1 binds to the RFB to prevent collision between the transcription and replication machineries, with Fob1 mutants being shown to have more collisions between the transcription and replication machineries when rDNA copy number is low (Takeuchi et al., 2003). However, a side effect of Fob1 is that it causes DNA double-strand breaks (DSB) that are repaired by homologous recombination. This means that Fob1 is crucial for copy number rectification, with studies suggesting that this protein is part of a mechanism by which rDNA copy number is restored to proper levels by homologous recombination (Kobayashi et al., 1998; Kobayashi & Ganley, 2005). consistent with this, *fob1-* mutants are not able to restore their copy number (Kobayashi et al., 1998).

1.4. RECOMBINATION IS A FORCE THAT PRODUCES COPY NUMBER VARIATION

The repetitive structure of rDNA loci makes them prone to different types of homologous recombination. A high frequency of recombination events in the rDNA loci has been reported in different organisms (Ganley & Kobayashi, 2011; McTaggart et al., 2007; Stults et al., 2008). rDNA loci undergo equal and unequal homologous recombination, where the rDNA copy number does not change and produces rDNA copy number variation, respectively (Brown & Wensink, 1972; Kobayashi, 2014, p. 2; Naidoo et al., 2013).

Unequal recombination has been reported during mitosis and meiosis in different organisms (Naidoo et al., 2013). Unequal recombination is characterised by unequal alignments that cause an uneven exchange, leading to gains and loss of rDNA copies. There are four types of unequal crossing over. The first occurs by a sister chromatid recombination. The second depends on the exchange of two chromosomes (interchromosomal exchange), the third one is recombination that takes place inside a chromatid (intrachromatid exchange), and the last one is gene conversion (Eickbush & Eickbush, 2007). Sister-chromatid recombination and intrachromosomal recombination involve a reciprocal exchange of the rDNA copies between chromatids (Fig. 1.4.1A and 1.4.1B). Unequal reciprocal exchange always creates a cell that will lose copies while the other cell will gain the copies lost by the other, but total rDNA copy number remains unchanged.

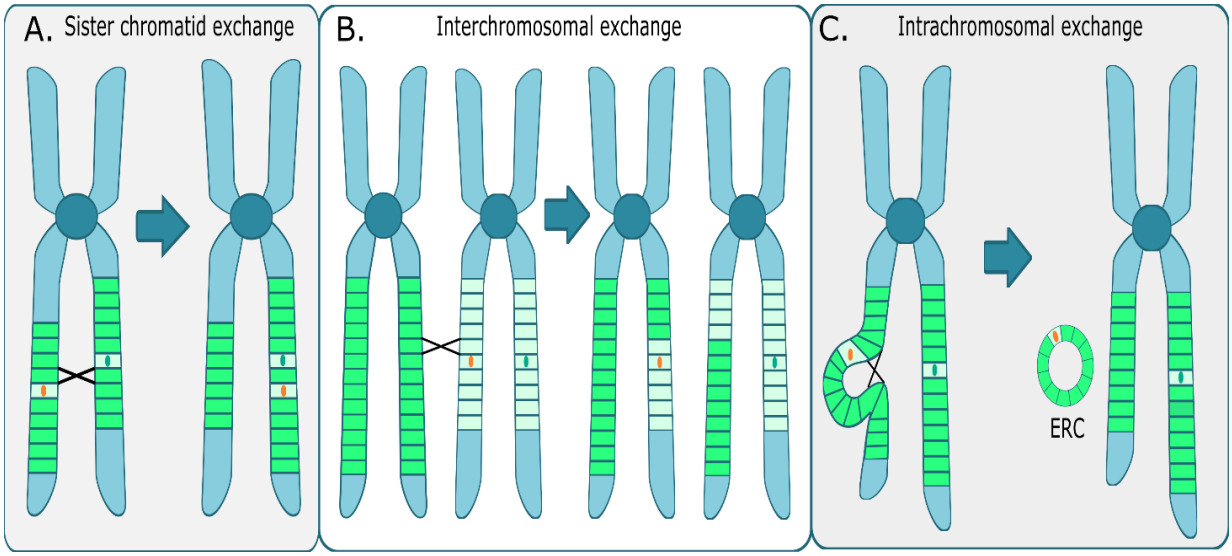


Figure 1.4.1. Schematic of the three classical types of unequal crossing over. The figure shows the three classic ways of unequal recombination. (A) Sister chromatid exchange between two chromatids reduces the copy number of one of the chromatids, while in the other, the copy number increase. (B) Meiotic recombination between two chromosomes. In this type of recombination, the copy number is altered in both chromosomes. (C) Recombination inside a chromatid. During this type of unequal crossing over, the chromatid will lose copies, and copies between the crossing points will be excised from the chromatid in extrachromosomal rDNA circle (ERC) structures.

Intrachromatid exchange occurs when two copies on the same chromatid are aligned and recombined. As a result, all intervening copies are lost. The lost copies are excised from the chromosome and form structures called extrachromosomal ribosomal DNA circles (ERC, Fig. 2C) (Ganley et al., 2009; Kobayashi, 2006; Sinclair & Guarente, 1997). This type of recombination is associated with the loss of the rDNA copies. Increased activity of Fob1 enhances the production of these ERCs, and eliminating this protein reduces ERC production and increases lifespan (Defossez et al., 1999; Lindstrom et al., 2011). ERCs have been reported to be reinserted in the rDNA cluster, and in this way they may be involved in rDNA copy number recovery (Mansisidor et al., 2018). However, reinsertion frequency seems low (Mansisidor et al., 2018); thus, they are probably not a common mechanism of copy number restoration.

Gene conversion is when a sequence from a homologous template is copied to a recipient sequence, which produces a non-reciprocal exchange (Kobayashi, 1992). In this type of

exchange, loss or gain of rDNA copies occurs in the chromatid in which the sequence is copied, but the chromatid that acts as the template remains unaltered (Kobayashi et al., 1998). Reciprocal (sister chromatid exchange and interchromosomal exchange) and non-reciprocal (gene conversion) recombination occur in different organisms and are involved in the high rDNA copy number variability. However, Gangloff et al. (1996) found that, at least for *S. cerevisiae*, non-reciprocal recombination is the main mechanism of expansion and contraction of the rDNA cluster. The frequencies of duplications and deletions events and the length of those events have been estimated. Unequal recombination events occur approximately once every two to three cell divisions (Ganley & Kobayashi, 2011), with the loss of a tagged rDNA copy estimated to occur between 5×10^{-4} to 2.5×10^{-3} events per generation (Gangloff et al., 1996; Ganley & Kobayashi, 2011; Kaeberlein et al., 1999; Prakash & Taillon-Miller, 1981; Szostak & Wu, 1980; Zou & Rothstein, 1997).

In *S. cerevisiae*, reciprocal and non-reciprocal recombination are involved in the repair of rDNA after a double-strand break (DSB). Different pathways control the activation of one mechanism or the other. The classical reciprocal recombination is activated by an RFB-independent pathway, while non-reciprocal recombination is RFB-dependant. The molecular mechanism has been extensively described in the *S. cerevisiae*. In the RFB-dependant pathway, Fob1 binds to the RFB, and a double-strand break is induced to initiate the recombination (Kobayashi, 2014). Two outcomes are possible and are regulated by the histone deacetylase, Sir2. In the first case recombination occurs when the copy number is at a wild-type level. Under these conditions, the Sir2 represses activity of a non-coding promotor in the IGS1, E-pro (Fig. 1.1.1) (Kobayashi & Ganley, 2005). Repression of E-pro leads to equal sister-chromatid recombination. During this recombination, the cognate rDNA unit on the sister-chromatid is used as the template for repair, thus there is no change in the copy number (Kobayashi & Ganley, 2005). The second case occurs when the copies decrease below the wild-type copy number. Under these conditions, Sir2 release the activity of E-pro. Active E-pro favours the dissociation of cohesin, releasing the sister-chromatid, and misalignments are produced. When misalignments are produced at the left side of the DSB on replicated copies, this non-reciprocal recombination increases the number of copies (Duplication event) (Fig 1.4.2). However, when the misalignment is produced in the right sector of the DSB on non-replicated copies, there is a loss of copies (Kobayashi, 2014).

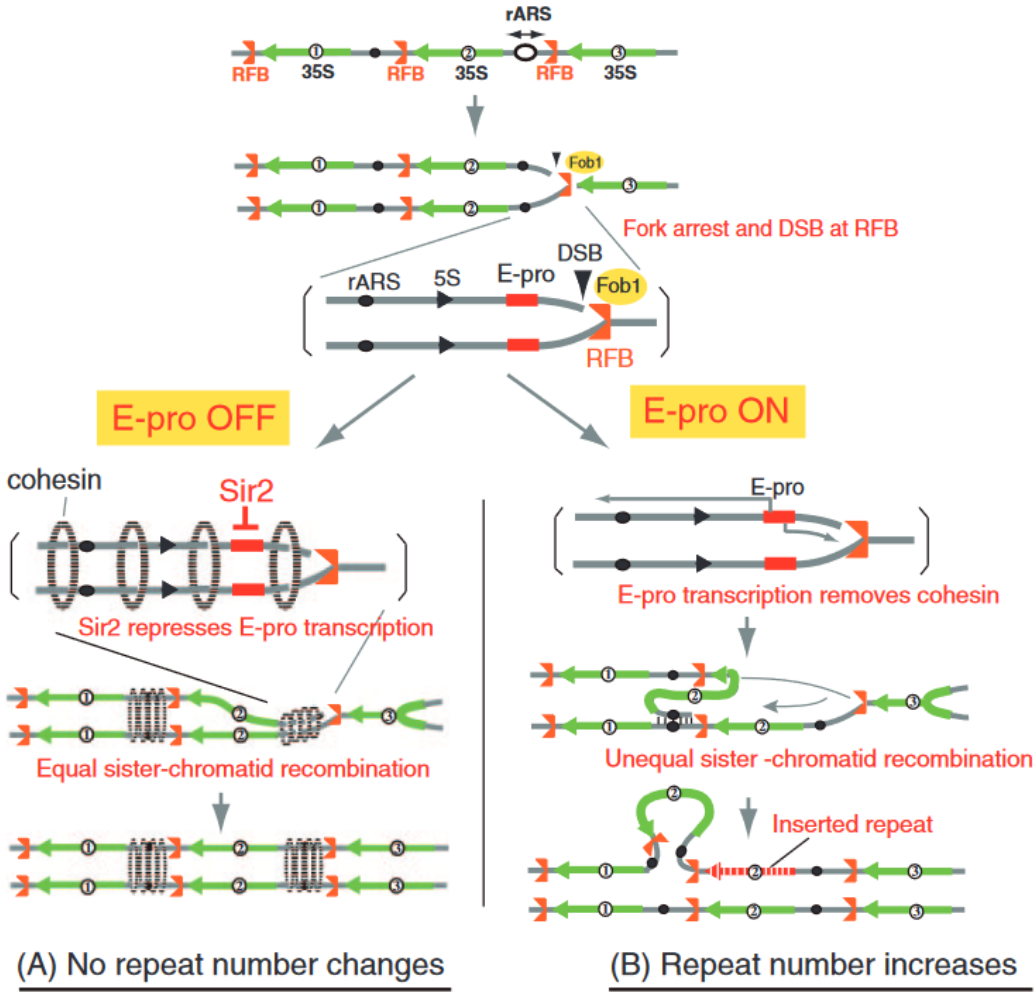


Figure 1.4.1. Nonreciprocal recombination pathway RFB-dependant. The figure shows the non-reciprocal recombination pathway dependent on RFB. On the left, equal reciprocal recombination occurs when Sir2 regulates the E-pro promoter. Cohesin molecules are involved in the DSB created by the interaction of RDB/Fob1. On the right, under low copy number, Sir2 releases E-pro which removes cohesin and leads to misalignments and unequal recombination. The gene is amplified, but the copies of the template chromatid remain unchanged. The figure is taken from Kobayashi (2014).

1.5. rDNA COPY NUMBER MAINTENANCE MECHANISMS

Cells seem to maintain a homeostatic level of rDNA copies despite the high copy number variability produced by their high recombination rates. Iida and Kobayashi (2019) described a mechanism in *S. cerevisiae* that seems to maintain rDNA copy number at these homeostatic

levels. The mechanism involves the interaction of the *SIR2* gene. and UAF (upstream activator factor), which is a transcription factor of polymerase I. UAF can associate with the *SIR2* gene and the rDNA locus. Under normal rDNA copy number conditions, the UAF factor predominantly associates with the rDNA because of its affinity for the rDNA locus. Under a low copy number context, a lack of rDNA copies releases UAF, whose production is independent of the copy number, and UAF is free to bind and repress the *SIR2* gene. The repressed Sir2 cannot repress E-pro, which thus initiates unequal recombination to repair Fob1-dependant DSBs (Iida & Kobayashi, 2019). The copy number increases to homeostatic levels because of the copy number changes produced by this recombination. As the copies start to increase, UAF will bind to the newly restored copies, and the *SIR2* gene repression will decrease, increasing the production of Sir2 and switching on the repression of E-pro. The repression of E-pro then stops the unequal non-reciprocal recombination, and DSBs are repaired by equal sister-chromatid recombination, which does not alter the copy number (Iida et al., 2019).

1.6. CONCERTED EVOLUTION

Last century, kinetic experiments exposed the presence of repetitive sequences in the DNA (Britten & Kohne, 1968). Since then, different types of repetitive sequences have been described, and studies have shown that those sequences present a variety of functions and are abundant in the genomes of several organisms (Ganley & Kobayashi, 2011). The repetitive sequences also evolve differently compared with unique sequences. Early analysis of repetitive tandem rDNA sequences in frogs showed that the sequences within species have a high homogeneity, but between species have higher variation (Brown & Wensink, 1972). This phenomenon, now called concerted evolution, was observed in other repetitive sequences (Ganley & Kobayashi, 2011). Unequal crossing over was proposed as a mechanism for concerted evolution (Eickbush & Eickbush, 2007).

1.7. AIMS AND OBJECTIVES

My research aimed to evaluate how copy number distribution is influenced by the dynamics of rDNA, specifically with the selective pressure of copy number, the type of recombination, and the level of misalignment in the recombination process.

Objective 1: Develop and validate a model that can assess the impacts of selection and non-reciprocal recombination on the distribution of copy numbers of a* Saccharomyces cerevisiae *population.

Developing and validating a model that can calculate rDNA copy number distribution and modelling the effects of selection and non-reciprocal recombination will be useful for understanding how those forces act at the population level. The work in objective one will introduce some of the new findings in rDNA copy number dynamics to update previous models. The model will use previous parameter estimations and will determine by simulation the parameters that do not require experimental data.

Objective 2: Experimentally estimate the copy number distribution of a haploid* Saccharomyces cerevisiae *population.

I will experimentally estimate the copy number distribution to generate data that can be used in conjunction with the model developed in objective one. To achieve this, I will grow a wild-type copy number strain over ~ 60 generations to allow the cells to stabilise the rDNA cluster copy number. Then, using ddPCR, I will estimate the rDNA copy number of multiple colonies in this population.

Objective 3: Assess the model using the experimental data and testing the different parameters to give insight into the dynamics of rDNA copy number distribution

Experimental data from objective two will be used to fit the model from objective one to make comparisons between different model configurations. This comparison aims to produce insights that allow an understanding of the effects of selection and non-reciprocal recombination on the copy number.

Objective 4: Construction of a plasmid with fluorescent proteins that can be inserted in the HO locus to analyse the fitness effect of* S. cerevisiae *strains.

Constructing a plasmid including a fluorescent protein gene that can recombine with the *HO* locus will be useful for future experiments that want to estimate the fitness effects of different copy numbers. The insertion of the fluorescent protein gene in the *HO* locus will not interfere with the normal growth of the cells, which is crucial for measuring effects on fitness. The plasmid will be constructed using a backbone with two sequences from the *HO* locus and cloning the fluorescent proteins between these two sequences.

Chapter 2. MODEL CREATION AND DESCRIPTION FOR MODELING POPULATION DYNAMICS

2.1. GENERAL DESCRIPTION OF THE MODEL

I want to develop a model that can calculate the copy number distributions in a *Saccharomyces cerevisiae* population, with a view to then determining the impact of selection and non-reciprocal recombination on the distribution of copy numbers. Then, to do this, I decided to base my model on the existing model of Lyckegaard and Clark 1991. This model described the ribosomal DNA copy number distributions of a *Drosophila melanogaster* diploid population with sexual reproduction and the individual distribution per sexual chromosome. The model established in this research was designed to describe the equilibrium copy number distributions of a haploid *Saccharomyces cerevisiae* population that divide asexually.

Two key processes affect rDNA copy number distribution: selection and unequal recombination. Selection is a key process that affects rDNA copy numbers because there are fitness differences between genotypes depending on their rDNA copy number. Therefore, cells with different copies may be affected by a selection pressure that is suggested to strongly affect individuals with fewer copies in *Saccharomyces cerevisiae* (Ide et al., 2010). Moreover, Simulations using *Drosophila melanogaster* as a diploid model suggested that selection is necessary to maintain the copy number around the values shown by wild-type populations (Lyckegaard & Clark, 1991). Theoretically, selection pressure can also affect large amounts of rDNA copies. There is evidence that this pressure can prevent genes with multiple copies increase their copies infinitely (Stephan & Cho, 1994; Walsh, 1987).

Selection and unequal recombination processes were included in the model established in this research to model rDNA copy number variation. However, in this project, only the effect of selection at the lower bound was modelled for simplicity because deletions caused by an unequal recombination process can prevent an infinite increase of copies. The rDNA copy number distributions are calculated using a discrete-generation process with two steps corresponding to the two key processes (Selection and unequal recombination). After a selection process, the first step estimates the changes in the proportions of genotypes with a specific copy number. The second step estimates the changes produced by a non-reciprocal

recombination process during which deletions and duplications of rDNA copies occur. This process is iterated over n generations to reach a quasi-equilibrium state, and the final distribution of cells is obtained. The quasi-equilibrium state was used instead of a full equilibrium state because, in some cases, equilibrium conditions can take considerable time, complicating further analysis. After an unequal recombination process, daughter cells could end with more, equal, or fewer copies than their progenitor. This copy number variation results from duplications or deletions of rDNA gene copies during unequal recombination. As described above, there are four types of unequal recombination (Section 1.4). In the model established in this research, I focused on non-reciprocal recombination. Previous evidence suggests duplication and deletions can occur at different rates (Ganley & Kobayashi, 2011). Therefore, we use different values for the duplication rate, denoted as α, and the deletion rate denoted as β

Non-reciprocal recombination processes modelled in this project are subjected to three biological constraints. The constraints are that deletion cannot go below one copy, that duplication can only occur when copy number is greater than one and that the number of duplicated copies cannot be higher than double the length of the rDNA copy minus one. The last constraint is set because at least one copy is required to produce an alignment between chromatids. Therefore, because the other chromatid is used as a template, the duplication of copies can be higher than the number of copies present in the template minus the one used in the aligned. Those constraints imply cells with one just one copy cannot increase or decrease their copy numbers. Then the domain of the functions starts at 0 and ends in the percentage value that does not break such constraints, which I defined as the maximum theoretical percentage of change m.

2.2. MODELING SELECTION EFFECT ON rDNA COPY NUMBER

In this model, I included a fitness function that alters the proportions of cells with i rDNA copies. This is a ramp function with genotypes with fitness values ranging from 0 to 1 depending on the copy number. In the function, two rDNA copy number values are defined as limits. The value w_1 is the lower limit of the fitness function, all the genotypes with i rDNA copies that are equal or under this limit have a fitness of 0. The value w_2 corresponds to the upper limit. All genotypes with rDNA copies equal to or above this limit have a fitness of 1. Finally, all the genotypes with copy numbers between those limits have a fitness value described by a linear function (Equation 2.2.1).

$$W = \begin{cases} x < w_1 = 0 \\ x > w_2 = 1 \\ w_1 < x < w_2 = mx + a \end{cases}$$

Equation 2.2.1 Linear fitness function between w_1 (Copy number at which fitness became zero) and w_2 (The copy number at which fitness becomes one)

To calculate the population's proportions of cells with i copies after selection N_i', the value obtained for the fitness with a specific copy number is multiplied by the previous generation's proportions N_i. This estimation is done for all the copy numbers, i , evaluated in the simulation (Equation 2.2.2).

$$N_i' = W_i \cdot N_i$$

Equation 2.2.2 Step 1: calculation of population's proportions of cells after selection

2.3. MODELING rDNA COPY NUMBER CHANGE THROUGH NON-RECIPROCAL UNEQUAL RECOMBINATION

As described about in the model established in this research, I focused on non-reciprocal recombination because this type of recombination has duplication and deletion recombination rates. The duplication rate is denoted as α, and the deletion rate is denoted as β. Estimations of their values for a wild-type copy number strain are α = 0.00354 and β = 0.00458 (Ganley & Kobayashi, 2011). These estimations were used in the model.

The magnitude of copy number change from a recombination event depends on the length of misalignment. I decided to employ probability distribution functions to determine the length of the misalignment produced during recombination. The maximum misalignment length is dependent on copy number. Encoding the misalignment length in percentages makes the same distribution possible regardless of copy number. Therefore, the percentage of total copy number was used as the unit of misalignment length.

I defined the probability distribution functions Γ for misalignment that ends in deletion (Table 2.6.1) and H for misalignments that end in duplication (Table 2.6.1) to model these two recombination events separately. Γ and H take as parameters the percentage of misalignment to retrieve the probability that a cell group with j copies end with i. Both functions are conditioned to the biological constraint that one copy should be remained intact to allow an anchor point necessary for the recombination. Then the domain of the functions starts at 0 and ends in the percentage value that allows at least one copy as an anchor, defined as the maximum

theoretical percentage of change m. This maximum theoretical imply that cells with just one copy cannot increase or decrease their copy number.

I predicted that the upper limit of the domain of the function does not necessarily have to be the maximum theoretical percentage but rather a value in the range of 1 to that maximum. To test this hypothesis, I set a limit in the percentage of changes b as a parameter in duplication and deletion functions. The probabilities of misalignments' percentages higher than those limits are zero. Using the deletion limit, we calculate value t, which refers to the maximum copy number whose percentage of misalignment to produce i copies is not higher than the maximum percentage of change for deletion. Similarly, we estimate the value l that represents the minimum copy number whose percentage of misalignment to produce i copies is not higher than the maximum percentage of change for duplication.

To determine the new proportion of genotypes with i number of copies after non-reciprocal recombination N_i'' (the proportion following calculation of selection, N_i') the proportion of cells that remained at copy number i and the proportion of cells that were j rDNA copies but reached i copies were calculated. The proportion of cells that keep i rDNA copies is N_i' (the proportion following calculation of selection) multiplied by the proportions of genotypes that do not undergo a deletion or duplication event (first part of Equation 2.3.1). This last proportion of genotypes is calculated by subtracting the duplication (α_i) and deletion (β_{i}) recombination rates of a genotype of i rDNA copies to 1 (The total proportion).

The proportion of cells that reach a i copy number is obtained by calculating all the changes produced in the populations due to duplications and deletions that generate phenotypes with i copy numbers. The changes due to deletions are calculated with a recursion that adds all the values of N_{j}' (the proportion following calculation of selection for the j rDNA copy number) multiplied by the rate of recombination β_{j} and multiplied by the probability of deletion misalignment of length $\left(\frac{i-j}{j}\right)$ $\Gamma_{(i,j)}$ (second part of Equation 2.3.1). The range of this recursion includes all the genotypes of j rDNA copies that are higher than i copy numbers to those with t copies. In a similar way, the changes due to duplication are estimated by adding all values of N_{j}' multiplied by the rate of recombination α_{j} multiplied by the probability of duplication misalignment of length $\left(\frac{i-j}{j}\right)$ $\mathrm{H}_{(\mathrm{i,j})}$ (third part of Equation 2.3.1). Because there is the biological constraint that the number of duplicated copies cannot be higher than double the length of the

rDNA copy minus one, the range of this recursion includes all the genotypes of j copies in the range of l to i - 1 copies.

$$N_i'' = N_i' (1 - \beta_{\mathrm{i}} - \alpha_i) + \sum_{j=\mathrm{i}+1}^{\mathrm{j}=\mathrm{t}} \left(N_j' * \beta_{\mathrm{j}} * \Gamma_{\left(\frac{j-i}{j}\right)} \right) + \sum_{j=l}^{i-1} \left(N_j' * \alpha_{\mathrm{j}} * \mathrm{H}_{\left(\frac{i-j}{j}\right)} \right)$$

Equation 2.3.1 Step 2: calculation of population's proportions of cells after unequal recombination

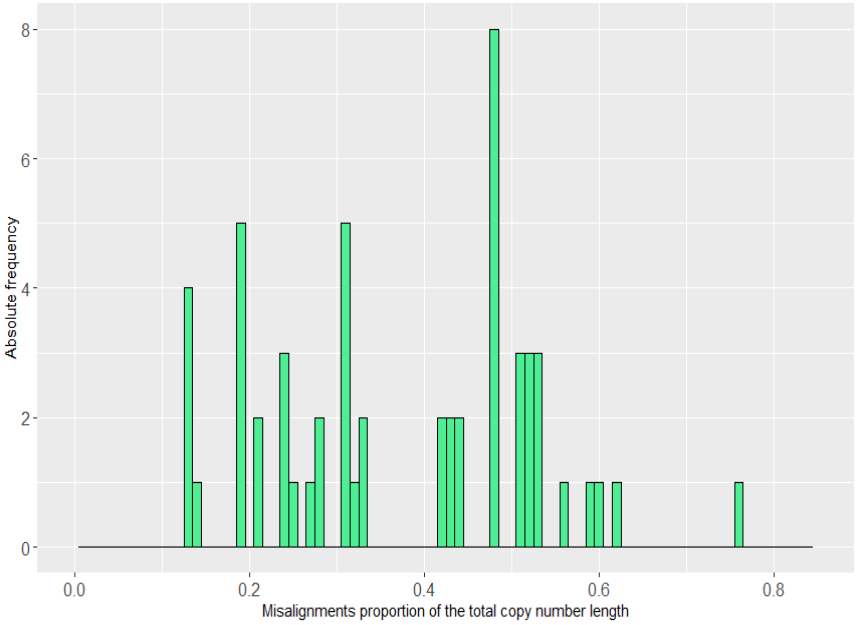
2.4. FITTING THE MISALIGNMENT PROBABILITY DISTRIBUTION FUNCTIONS

The only available published experimental measurements of rDNA misalignment length frequencies come from *S. cerevisiae* (Kobayashi & Ganley, 2011). The data describe the number of deletion and duplication events detected by PFGE and the percentage of change in copy number after these events. There were available 55 instances of deletion estimates and 34 instances of duplication. These deletion and duplication distribution data (Fig. 2.4.1.1) were used to establish the probability distribution functions Γ and H with the best fit.

2.4.1 FITTING THE DELETION DISTRIBUTION TO MODEL DELETION MISALIGNMENTS

Inspection of the data (Fig. 2.4.1.1A) suggested that a uniform distribution can approximate function Γ. Parameter *b* (maximum misalignment percentage) was taken as the maximum value in the dataset, which was 0.7576923. This distribution has an initial parameter *'a'* that was set to 0. Using this, the probability P_0 was estimated (Fig. 2.4.1.2). Misalignment percentages outside the range *a – b* will obtain probabilities of 0 (Fig. 2.4.1.2). In some cases, when the copy number is low, value *b* is higher than the maximum theoretical percentage of change *m*. In those cases, the value *m* is set as *b*.

A) Histogram of proportions of deletion misalignment



B) Histogram of proportions of duplication misalignment

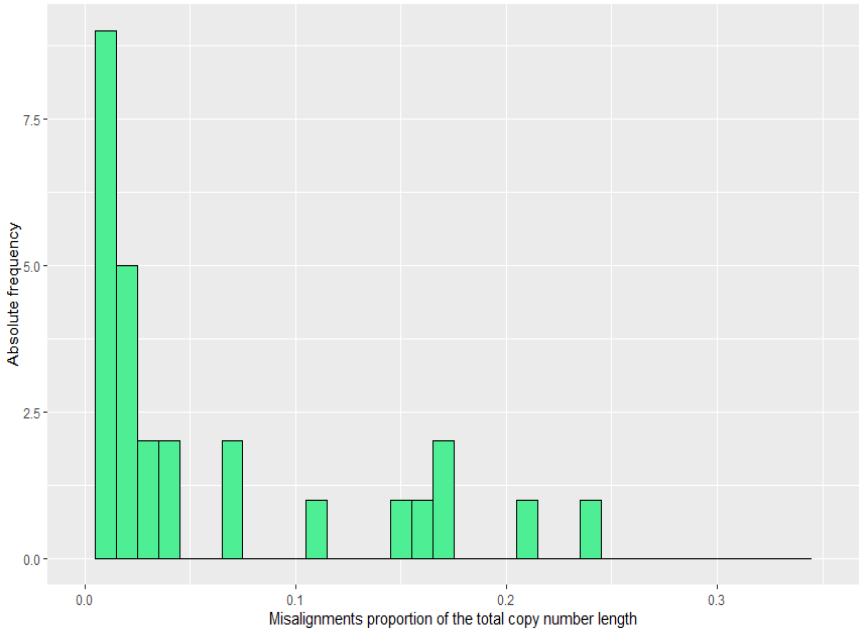


Figure 2.4.1.1 Absolute frequency of deletion and duplication misalignment proportions of the total copy number length. Histogram showing the misalignment proportions of the total copy number length vs their absolute frequencies for A) deletion and B) duplication. Data was taken from (Ganley & Kobayashi, 2011).

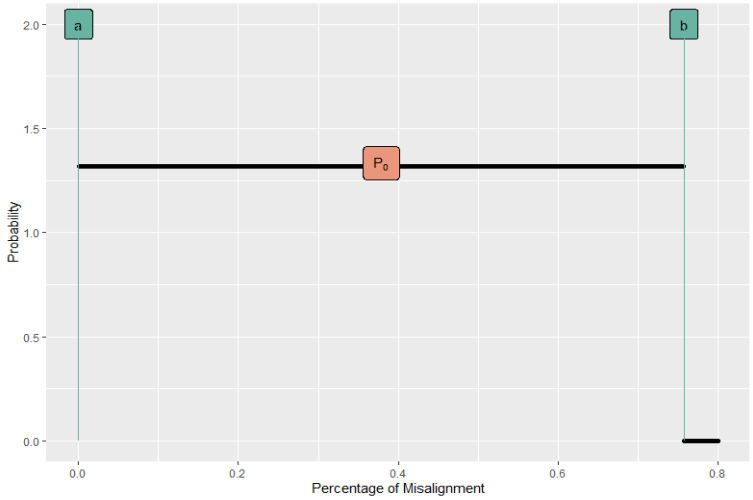


Figure 2.4.1.2 Uniform distribution probability for deletions. The plot shows the proportion of misalignment with respect to the total vs the height of the probability. Value P_0 represents the height of probability density between percentage *a* to percentage *b*. The probabilities of values lower than *a* and higher than *b* were set to 0.

2.4.2 FITTING DISTRIBUTIONS FOR MISALIGNMENT THAT ENDS IN DUPLICATIONS

The duplication data from Ganley and Kobayashi (2011) did not resemble a simple uniform distribution (Fig. 2.4.1.1B). Therefore, we decided to test various distribution types to determine which one fits best. We used common distributions in R, namely the Beta, Exponential, Gamma, Log-Normal, Logistic, Uniform, and Weibull distributions. The distributions were fitted to the data in Fig. 2.4.1.1B using a custom R-script based on the package Fitdistrplus (Appendix A. Script 1), which can calculate the maximum likelihood and Akaike information criterion (AIC) values for common distributions defined in R. In addition, an inspection of the data suggests that it may fit a double uniform distribution model. This double uniform distribution model is constructed using two uniform distributions joined. The first distribution is associated with small misalignments, while the second describes the probability of longer misalignments. Under this model, the first uniform distribution has a height P_0 going from the value *a* to *Xm* (Equation 2.4.2.1; Fig. 2.4.2), while the second uniform distribution has a height P_1 going from value *Xm* to *b* (Equation 2.4.2.1; Fig. 2.4.2). Misalignment's percentages higher than *b* (maximum percentage of change) were set to a probability of zero (P_2) (Fig. 2.4.2). The function that describes the double uniform distribution, its range and its probability was defined as follows:

$$P(x) = \begin{cases} P_0 = (a < x \leq Xm) \\ P_1 = (Xm < x \leq b) \\ P_2 = x > b \end{cases}$$

Equation 2.4.2.1. Double uniform distribution probability

Xm = Percentage limit between the two-uniform distribution.
a = Minimum percentage observed value.
b = Maximum misalignment's percentage of change.

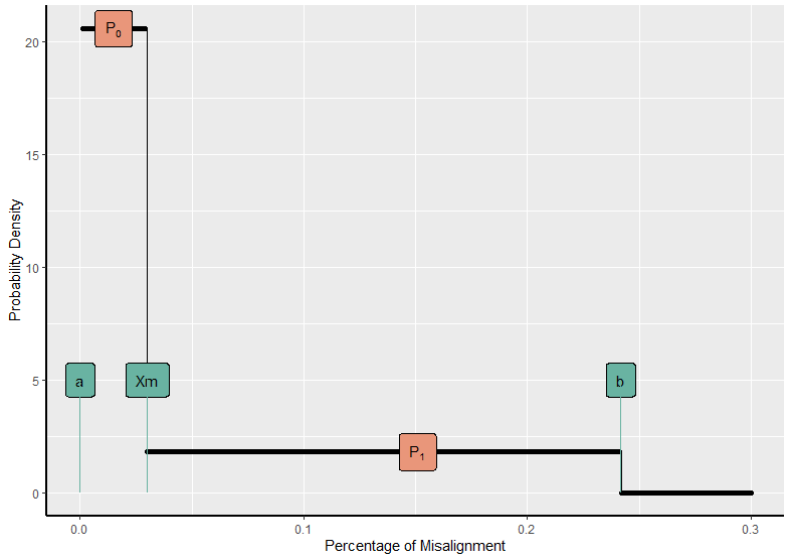


Figure 2.4.2 Double uniform distribution probability for duplications. The plot shows the proportion of misalignment with respect to the total vs the height of the probability. Value P_0 represents the probability density of the values between the minimum percentage observed *a* to *Xm.* The range describes small percentages of misalignments that are more frequent. Value P_1 represents the probability density of the values between *Xm* to *b*, which corresponds to the probability of longer misalignments. The probabilities of values lower than *a* and higher than *b* were set to 0.

The double uniform distribution probability has not been defined as a common distribution function in R. Therefore, the package Fitdistrplus could not calculate the maximum likelihood and AIC values for this double uniform distribution. Thus, maximum likelihood and AIC values were estimated manually using a custom R script (Appendix A Script 2). The script uses equations 2.4.2.2, 2.4.2.3, and 2.4.2.4 to obtain equation 2.4.2.5, which was used to calculate the value of only parameter *Xm* with the maximum likelihood. The AIC value for the distribution was estimated using the maximum log likelihood. Parameter *a* was set to 0, and

parameter *b* was 0.2417, corresponding to the maximum value reported in the data set. The script estimated the *Xm* value to be 0.03. The lowest AIC score across all eight distributions was obtained from the double uniform distribution (Table 2.4.2). Therefore, this distribution was used in the model as the distribution function for duplication misalignments H. Then when it was required to calculate the probability of a given misalignment's percentage, the function was used.

$$Xm * P_0 = n_0 / n_T$$

Equation 2.4.2.2. Probability area of first uniform distribution.

$$(b - Xm) * P_1 = (n_T - n_0) / n_T$$

Equation 2.4.2.3. Probability area of the second uniform distribution

$$\text{Likelihood} = P_0^{n_0} * P_1^{n_T - n_0}$$

Equation 2.4.2.4. The likelihood function for double uniform distribution

$$\text{Log Likelihood} = n_0 \log \frac{n_0}{n_T * Xm} + (n_T - n_0) * \log \frac{n_T - n_0}{(b - Xm) * n_T}$$

Equation 2.5.2.5. The Log-likelihood function for double uniform distribution

P_0 = *probability of shorter misalignments.*
P_1 = *probability of longer misalignments.*
n_0 = *the number of values lower than Xm (Percentage limit between the two-uniform distribution) in the dataset.*
n_T = *the total number of values in the data.*
b = Maximum percentage observed value.

Table 2.4.2. Statistics from fitting experimental duplication results to various distributions

Distribution type	Max. Log-Likelihood	AIC
Weibull	67.46446	-130.92891
Gamma	66.89325	-129.78649
Uniform	48.98434	-93.96867
Log normal	70.68213	-137.36427
Exponential	66.27147	-130.54294
Beta	66.62392	-129.24785
Logistic	45.57896	-87.15792
Double uniform distribution	76.32320	-144.6464

2.5. ADJUSTMENT OF CONTINUOUS DISTRIBUTION TO DISCRETE VALUES

The misalignment functions produce copy number change values as percentages of existing copy numbers. However, these continuous values need adjusting to discrete values because copy numbers (in the model and biology) are discrete. To achieve this, the distribution function for deletion and duplication misalignments (Γ and H) used a factor, Δx, calculated by dividing one by the copy number. Scaling factor Δx was applied to each evaluated copy number. Values x_0 and x_1 were defined as the lower and upper limits of the interval used to calculate the total probability (area under the curve) of a misalignment's percentage transformed to discrete values. x_1 also represent the misalignment's percentage that is required by a genotype of j rDNA copies to become another of i rDNA copies obtained from the percentage of change evaluated by the function. Δx was used to calculate the value x_0 using equation 2.5.1. The x_0 and x_1 values should be in the range of a (Minimum misalignment's percentage in Γ and H functions) and b (Maximum misalignment's percentage in Γ and H functions). Probabilities lower than a have no biological sense, and probabilities higher than b can violate the biological constraints defined above (Deletion cannot go below one copy, duplication can only occur when copy number is greater than one, and the number of duplicated copies cannot be higher than the double of the length of rDNA copy minus one). Therefore, when the x_0 and x_1 values were higher than b or lower than a the probability is zero. However, it is possible to calculate the probability (area under the curve) of the range $x_0 - b$, when at least the x_0 value is lower than b.

With x_0 and the value x_1 then was possible to calculate the probability of discrete units of misalignment's percentage (Fig. 2.5.1) using equation 2.5.2 for the case of Γ (The distribution function for deletion misalignments events).

In the case of H, the misalignment probability was calculated similarly, but with the difference that the probabilities P_0 and P_1 had to be calculated and then added to obtain the total probability for a duplication misalignment's percentage. The equations 2.5.3 and 2.5.4 were used to calculate the probabilities P_0 and P_1. The value 34 in equation 2.5.4 corresponds to the number of instances of duplication obtained in the original data set (Ganley & Kobayashi, 2011) or the value n_T (the total number of values in the data).

$$x_0 = x_1 - \Delta x$$

Equation 2.5.1. Equation to calculate x_0

$$P_0 = \frac{1}{b-a} \cdot (x_1 - x_0)$$

Equation 2.5.2. Calculation of P_0 for Γ

$$P_0 = \begin{cases} \frac{n_0}{x_m \cdot n_t} \cdot (x_1 - x_0), x_1 < X_m \\ \frac{n_0}{x_m \cdot n_t} \cdot (Xm - x_0), x_0 < X_m < x_1 \\ 0, Xm < x_0, x_1 \end{cases}$$

Equation 2.5.3. Equations to calculate P_0 for H

$$P_1 = \begin{cases} \frac{34 - n_0}{(n_t - Xm) \cdot 34} \cdot (x_1 - Xm), x_0 < Xm < x_1 \\ \frac{34 - n_0}{(n_t - Xm) \cdot 34} \cdot (x_1 - x_0), Xm < x_0 < x_1 \\ \frac{34 - n_0}{(n_t - Xm) \cdot 34} \cdot (b - x_0), x_0 < b < x_1 \\ 0, b < x_0, x_1 \end{cases}$$

Equation 2.5.4. Equations to calculate P_1 for H**. The**

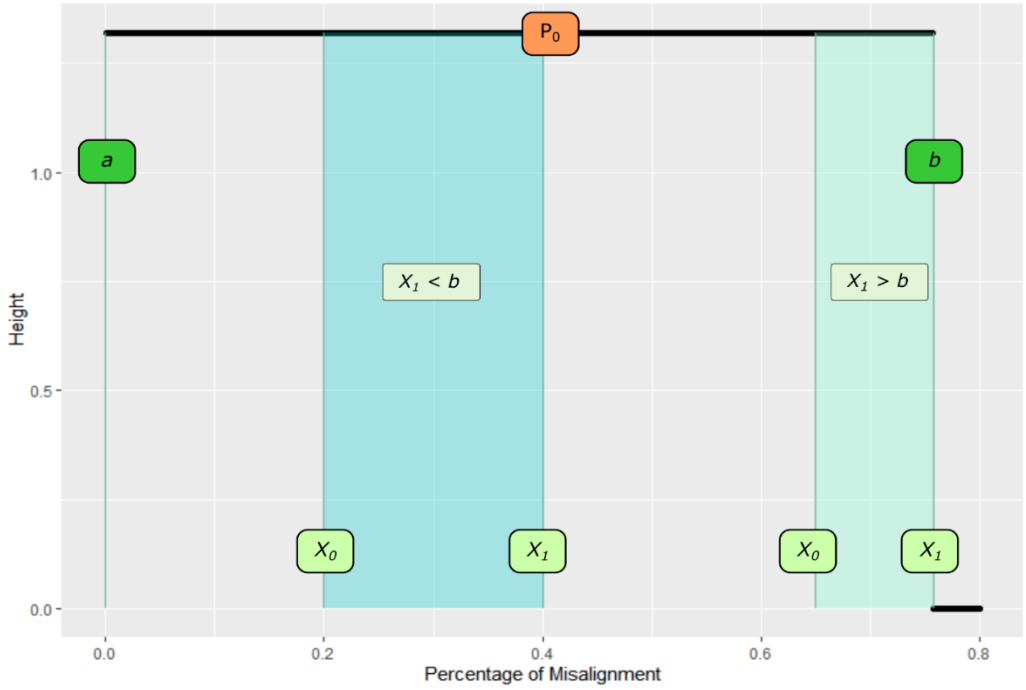


Figure 2.5.1 Examples of calculation of probabilities of change using a uniform distribution. x_1 represents the percentage of change that a genotype with j rDNA copies to a group of copies i. Values x_0 and x_1 are the lower and upper limits of the intervals used to calculate the probability (area under the curve) of a given misalignment's percentage. x_0 is estimated using the scale factor Δx, which depends on the copy number. When $x_1 < b$ the probability is estimated in the interval x_0 - x_1. When $x_1 > b$ the probability is estimated in the range x_0-b. The probability is 0 when both values are not in the range a – b.

2.6. TESTING INDIVIDUAL COMPONENTS OF THE MODEL

The model was implemented in C++ to perform the simulations faster. Different components of the model were tested to ensure they were working properly. All these simulations were performed with the same generation time of 100000 to allow a quasi-equilibrium for all the components. To compare both components, the same recombination rates were defined for duplication and deletion. The parameters used were b of $\Gamma = 1$, $w_1 = 58$, $w_2 = 137$, $\alpha = 0.0045$ and $\beta = 0.0045$ when the individual component was tested. For all tests, 400 was set as the maximum copy number evaluated. Two initial starting copy number distributions were used to assess if the result depended on the initial conditions. One initial distribution was uniform across all the evaluated copies and the other with 100 % of the population at 150 copies.

I ran a model with the selection step only to test if the selection function was working properly. This test gave two different distributions that were associated with the starting conditions. At uniform starting distribution, all the phenotypes that have a copy number higher than 137 (w_2) had the same proportion of approximated 0.003 while values under this value obtain proportions of 0. That result was expected because, over generations, the genotypes with fitness 1 will outcompete the other genotypes. The second distribution obtained in this test also was expected because the fitness of a genotype with 150 copies is 1, and there is no other process that alters the initial proportions. Therefore, for all the generations, the phenotype will keep its initial proportions (Figure 2.6.1).

Non-reciprocal recombination with just deletion without selection (Deletion only) was run to test the deletion process. The distribution obtained for this result was that all of the population ends having one copy. This result was expected because, in the absence of a force that compensates for deletion, all of the population will have one copy after several generations. Testing the interaction between selection and deletion was done by running a model of non-reciprocal recombination with just deletion and selection (Selection and deletion). The distribution obtained from this model push all the population toward 137 copies (w_2). Simulations with deletion plus selection show that selection has an important effect in maintaining a higher mean copy number in the proposed model. That was expected because the selection force will eliminate the genotypes with copies with fitness lower than one, but the deletion will make the genotypes with the highest copies lose their copies. The same tests of non-reciprocal recombination but including selection and full non-reciprocal recombination, including duplication and deletion without selection.

A model with non-reciprocal recombination with just duplication without selection (Duplication only) was run to test if the process was modelled properly. The model shows a distribution in which most of the population reaches higher copy numbers and converges toward the highest value. That is what was expected because when there is just a duplication as a process that alters the copy number, the average copy number is expected to increase its value. Therefore, after running the model for many generations, the proportions of the model's genotype with the highest copy number will reach 1. To test if duplication and selection were interacting properly. A model with non-reciprocal recombination with just duplication and selection was run (Selection and duplication). The distribution obtained from this test was similar to the duplication only. This result was also expected because, in this version, no process decreases the copy number. Therefore, the average copy number will increase and end

with the result of the Duplication-only model. In a similar way to selection and deletion, it was observed that selection favours the convergence toward high copy number values. Comparison between Deletion only or Duplication only exposes an asymmetry between the two processes. Both models were run with equal recombination rates and equal generation numbers. However, deletion converges faster toward lower values than duplication toward higher values. That was evidenced in the models in which duplication was run, but a full convergence was not achieved.

The duplication and deletion processes interaction was tested with a model with full non-reciprocal recombination without selection (Duplication and deletion). The distribution for this model was similar to Deletion only model, with all the population having one copy. This result is expected of this comparison despite the duplication as a force that increases the average copy number. The explanation for this result comes from the behaviour of copy number one. Because genotypes with one copy cannot increase or decrease their copies, all the genotypes that reach this copy number by deletion events will get stock, and duplication will not be able to change the proportion of those genotypes. Therefore, the population will end with one rDNA copy after multiple generations.

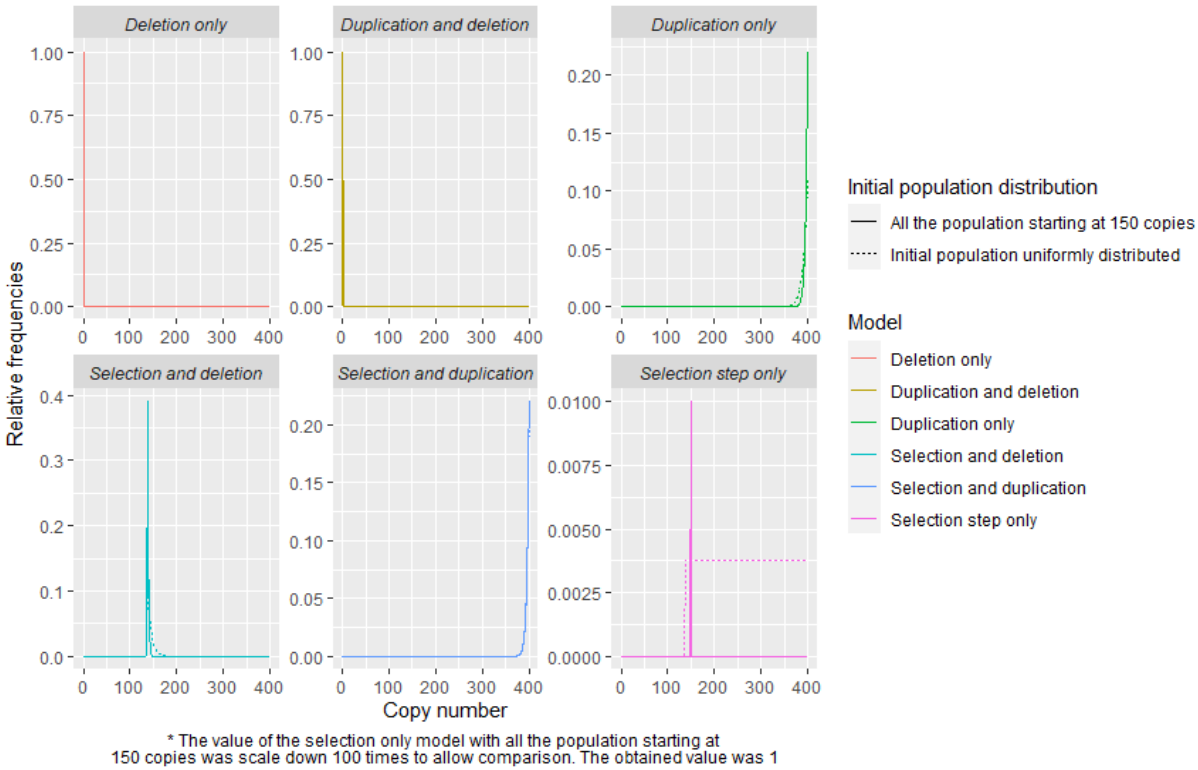


Figure 2.6.1 Individual model's component simulations. Test of the individual components. X-axes represent copy numbers, and y-axes relative frequencies. In the legend, components that were not mentioned are not used. Simulations that included deletion or duplication were

run with a duplication or deletion rate of zero, respectively. The selection step was omitted in simulations that do not include it.

2.7. COMPARING THE MODEL WITH THE PUBLISHED MODEL OF LYCKEGAARD AND CLARK

I wanted to compare the model developed here with that presented by Lyckegaard and Clark (1991) to evaluate if they give similar results. One dataset from the Lyckegaard and Clark model uses Y chromosome sister chromatid exchange to create variation in rDNA copy number and intrachromatid recombination that deletes copies. Thus, it keeps copy number bounded but does not feature interchromosomal exchange (unlike other datasets in their model, as *Drosophila* has rDNA arrays on both the X and Y chromosomes). To reproduce their parameters in my model, I set the parameters $w_1 = 20$, $w_2 = 70$, $\alpha = 0.0045$ and $\beta = 0.0227$ to match the parameters used in Lyckegaard and Clark model. The model was run over 20000 generations, and the range of copy numbers allowed was constrained to 1 - 250 copies. The starting conditions were the ones used in Lyckegaard and Clark, 1991 which are the results of the rDNA copy number frequencies obtained per chromosome. In the simulation, intrachromatid recombination and sister chromatid exchange were replaced by non-reciprocal recombination deletion and duplication but using the same parameters. For both cases, simple uniform distributions were used. However, the two models gave different copy number distributions (Figure 2.7.1.1).

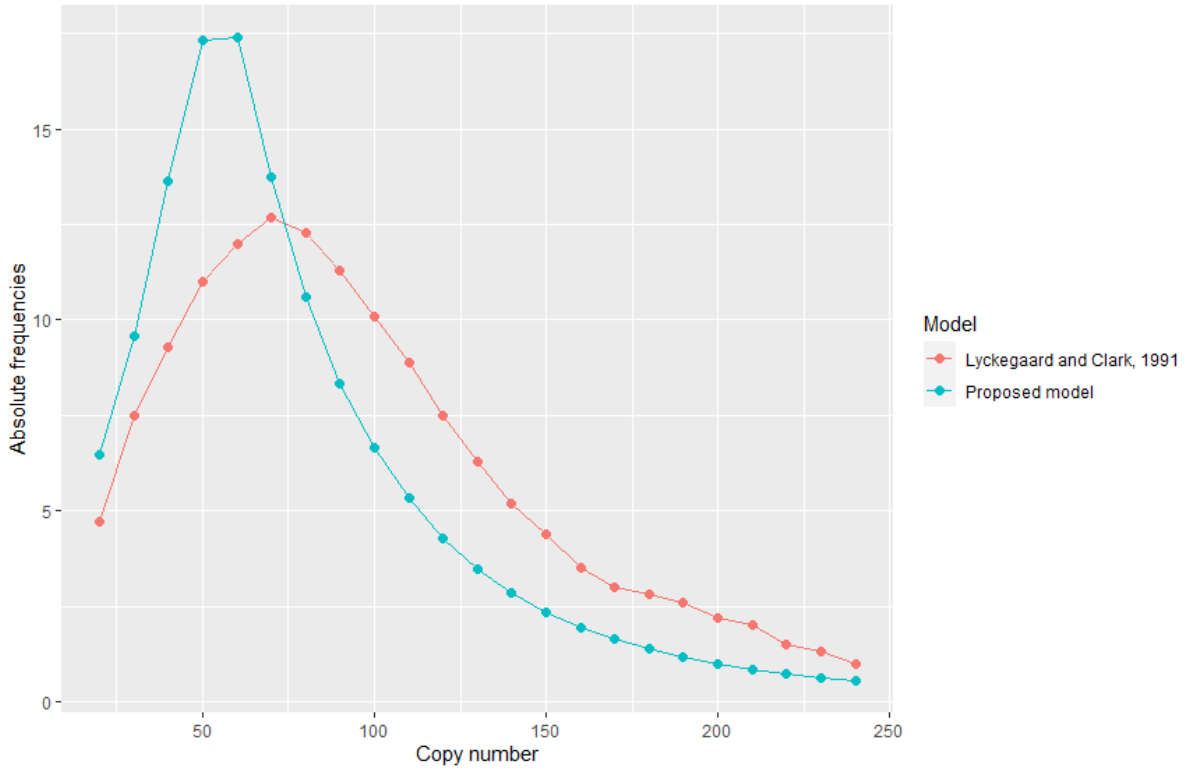


Figure 2.7.1.1 Comparison of copy number distributions generated from the Lyckegaard and Clark and this study's models. Data obtained from Lyckegaard and Clark (1991) is plotted alongside the simulation output from my model. Copy numbers from my model were placed into bins of ten to match the Lyckegaard and Clark data. The relative frequencies generated by my model were transformed to the absolute total value multiplied by the result of the binned data times 144 (This value corresponds to the total number of chromosomes extracted in Lyckegaard and Clark, 1991). The model was run for 20000 generations, and the parameters used for the simulation were $w_1 = 20$, $w_2 = 70$, $\alpha = 0.0045$ and $\beta = 0.0227$, which were the parameters used in Lyckegaard and Clark, 1991. The range of copy numbers allowed was 1 - 250 copies. The starting conditions were the ones used in Lyckegaard and Clark, 1991. The values of these starting conditions come from rDNA copy number frequencies obtained per chromosome in their experiments.

I also compared my model to an alternative, simplified version of this model that has been developed (unpublished results, Nobuto Takeuchi, University of Auckland). These gave numerically identical results when the same parameters were used (Figure 2.7.1.2). As the Lyckegaard and Clark models are not available for inspection or trial, it is difficult to know why there is a difference in distributions. For example, there may be an underlying difference in their model that is not evident from the text, or their model may contain a bug. Given the

consistency in the second model comparison and the results of tests of the components of my model, I conclude that my model is operating correctly.

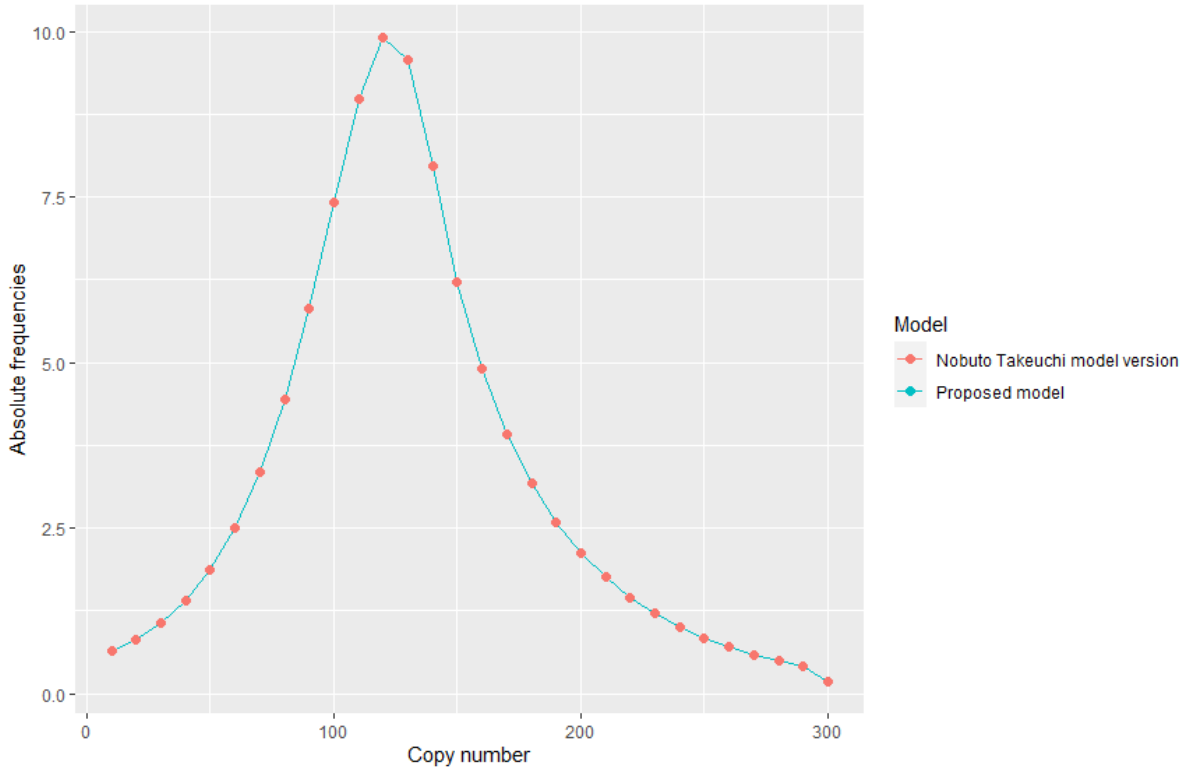


Figure 2.7.1.2 Comparison of copy number distributions generated from the Nobuto Takeuchi and this study's models. Data obtained from unpublished results, Nobuto Takeuchi, University of Auckland, is plotted alongside the output of a simulation from my model. Copy numbers from models were placed into bins of ten. Relative frequencies in both models were multiplied by 100 to obtain absolute frequencies. Models were run for 10000 generations with a range of 1 - 250 copies. The parameters used for the simulations were w_1 = 20, w_2 = 70, α = 0.0045 and β = 0.0227. The starting conditions were uniform distribution in all copy numbers in both models.

2.8. LIST OF SYMBOLS

Table 2.6.1. List of symbols used

Symbol	Description

α	Non-reciprocal duplication rate
β	Non-reciprocal deletion rate
Γ	The distribution function for deletion misalignments events
H	The distribution function for duplication misalignments events
$\boldsymbol{N_j}, \boldsymbol{N_i}$	Genotype's proportions with j or i copies.
$\boldsymbol{N_j}', \boldsymbol{N_i}'$	Genotype's proportions after selection step
$\boldsymbol{N_j}'', \boldsymbol{N_i}''$	Genotype's proportions after non-reciprocal recombination
a	Minimal misalignment's percentage in Γ and H functions
b	Maximum misalignment's percentage in Γ and H functions
m	Minimal theoretical misalignment's percentage in Γ and H functions
t	Maximum copy number whose percentage of misalignment to produce i copies is not higher than b
l	Minimum copy number whose percentage of misalignment to produce i copies is not higher than b
W_i	Fitness function for the i copy
w_1	The lower limit for fitness function
w_2	The upper limit for fitness function
Xm	The percentage limit between the two-uniform distribution.
P_0	Probability for Γ and first uniform distribution probability for H
P_1	The second uniform distribution probability for H.
n_0	The number of values lower than Xm in the dataset for estimating H.
n_T	The total number of values in the dataset for estimating H.
x_0	The lower limit of the range to estimate the probability of misalignment's percentage
x_1	Misalignment percentage of change
Δx	Scale factor to convert continuous values to discrete ones.

Chapter 3. MATERIALS AND METHODS

3.1. MEDIA AND COMMON SOLUTIONS

All media and solutions were prepared using sterile water and then autoclaved at 121 °C for 15 minutes unless another procedure was indicated.

3.1.1 **LURIA BROTH (LB) MEDIA**

LB media contained 5% (w/v) LB-Broth Miller (Formedium #LMM0102) in sterile water. LB Agar was prepared by adding agar (Formedium #AGA03) to a final concentration of 2% (w/v). LB-amp Agar was prepared by adding the antibiotic ampicillin to a final concentration of 100 µg/mL to LB Agar before pouring the plates. LB supplemented with glucose (LB + GLU) was prepared by adding glucose (Formedium #GLU04) to a final concentration of 20 mM.

3.1.2 **YEAST EXTRACT-PEPTONE-DEXTROSE (YPD) MEDIA**

YPD media contained 1% (w/v) Yeast Extract (Formedium #YEM03), 2% (w/v) D+ glucose (Formedium #GLU04) and 2% (w/v) Peptone (Formedium #PEP03) in sterile water. YPD Agar was prepared by adding agar (Formedium #AGA03) to a final concentration of 2% (w/v). G418 (Formedium #G4185) was added to a final concentration of 300 µg/mL to the YPD media when required to create YPD Agar G418.

3.1.3 **YEAST NITROGEN-BASE (YNB) MEDIA**

YNB media contained 1% (w/v) Yeast nitrogen base Extract (Formedium #CYN0410), 2% (w/v) D + glucose (Formedium #GLU04) and in sterile water. YNB Agar was prepared by adding Agar (Formedium #AGA03) to a final concentration of 2% (w/v). Selective media was prepared by adding a mix with all amino acids to the base, excluding either Leucine or Histidine. The exact compositions for the media are shown in Table 3.1.3.

Table 3.1.3. Selective amino acid supplements mix.

Supplement mix	Amino Acid	Weight in mix	Quantity
YNB - Leu	Adenine	2 g	100 mg/L
	Tryptophan	2 g	
	Histidine	2 g	
	Uracil	2 g	
YNB - His	Tryptophan	1 g	120 mg/L
	Uracil	1 g	
	Adenine	1 g	
	Leucine	5 g	

3.1.4 **COMMON BUFFERS AND SOLUTIONS**

Transformation buffer I

Transformation buffer I contain 30 mM potassium acetate (Sigma #P1190-100G), 100 mM RbCl, 10 mM $CaCl_2$ $2H_2O$, 50 mM $MnCl_2$, and 15% Glycerol (Univar #242-2.5L) in sterile water. The pH of the buffer was adjusted to 5.8 by adding acetic acid. The buffer was sterilised by filtration (0.45 µm).

Transformation buffer II

Transformation buffer II contain 10 mM MOPS, 75 mM $CaCl_2$ $2H_2O$, 10 mM RbCl, and 15% Glycerol in sterile water. The pH of the buffer was adjusted to 6.5 by adding NaOH. The buffer was sterilised by filtration (0.22 µm Millex-GV PVDF MERCK #SLGV033RS).

DNA ladder

DNA ladder was prepared by adding 160 µL DNA Gel loading dye (6x) (Thermo Scientific, #R0611), 100 µL GeneRuler DNA Ladder Mix (500 ng/µL) (Thermo Scientific, #SM0331), and 740 µL water, the mixture was vortexed.

EDTA-Na_2 Solutions

EDTA-Na_2-2H2O (Neofroxx #LC-5658) was used to prepare 0.5 M (pH 8.0) and 1 M (pH 7.6) EDTA-Na2 stock solution. pH was adjusted with NaOH (Merck #1.06498.0500).

Genomic DNA extraction solutions for *S. cerevisiae*

Buffer I contains 1 M Sorbitol (Formedium #SOR02) and 0.1 M EDTA-Na_2, pH 7.5. The buffer was sterilised by filtration (0.45 µm Millex-HV PVDF MERCK #SLHV033RS). Zymolyase

(MP-Biomedicals #320921) at 6.3 units/µL were added to Buffer I immediately before use to reach a final concentration of 0.26 units/µL.

The Buffer II contains 50 mM Tris-HCl 50mM, 20 mM EDTA-Na_2 and 0.35 M SDS (Serva #20760), pH 7.4). The buffer was sterilised by filtration (0.22 µm Millex-GV PVDF MERCK #SLGV033RS).

Loading dye 6X

Loading dye was prepared by adding 10 µL of 10 mg/mL Ethidium Bromide (Invitrogen) in 1 mL DNA Gel loading dye (6x) (Thermo Scientific, #R0611).

Salmon Testis (DNA carrier for *S. cerevisiae* transformations)

Salmon sperm carrier DNA used was in a concentration of 10 mg/mL.

SB Buffers

20x Stock solution contain 0.2 M NaOH (Merck #1.06498.0500) and 0.76 M Boric Acid (Merck #A0724365507) in water (pH 8.2). SB buffer was then used at 1x concentration after dilution in water, with final concentrations of 10mM NaOH and 38mM Boric acid at pH 8.2.

TE (10/1) buffer

10x stock solution of TE (10/1) buffer contain 100 mM of Tris-HCl (pH 8.0) and 10 mM of EDTA (pH 8.0). TE (1/10) buffer contained 10mM of Tris-HCL, and 1mM EDTA was prepared from stock. The TE solution used for transformations was sterilised by autoclave while the TE solution for gDNA extraction was filtrated (0.45 µm Millex-HV PVDF MERCK #SLHV033RS). TE (10/1) buffer containing RNase was prepared by adding 0.1 volumes of TE (10/1) solution, 0.01 volumes of 10mg/mL RNase (Invitrogen #12091-021) and adding PCR grade water to prepare the desired volume. The final concentration of RNase was 100 µg/mL.

TE, LiOAc and PEG solution

TE (10/1) – LiOAc 100mM solution was prepared on the day by adding 1 mL of 10x stock solutions and sterile water to 10 mL. TE (10/1) – LiOAc 100mM – PEG 40% solution was prepared on the day by adding 1 mL from 10x stock solutions and freshly made PEG 50%.

Zymolyase

Lyophilised enzyme was dissolved in 20mM Tris-HCl 50% glycerol solution to a final concentration of 30 mg/mL.

Simple solutions

Simple solutions are given in Table 3.1.4.

Table 3.1.4. Simple solutions.

Solution name	Amount / 100 mL	Final concentration	Product/Brand	Sterilisation method
PEG 50% solution *	50 g	50 %	PEG4000 powder (Merck #8.07490.1000)	Filtered (0.45 µm Millex-HV PVDF MERCK #SLHV033RS)
Potassium Acetate *	49.7 g	5 M	Potassium acetate (Sigma #P1190-100G)	Filtered (0.45 µm Millex-HV PVDF MERCK #SLHV033RS)
Lithium Acetate (LiOAc) *	10.2 g	1 M	Lithium acetate di-hydrate (Sigma #L6883-250G)	Autoclaved
Sorbitol *	36.4 g	2 M	Sorbitol (Formedium #SOR02)	Filtered (0.45 µm Millex-HV PVDF MERCK #SLHV033RS)

***The solution was prepared in sterile water

3.2. STRAINS, PLASMIDS AND GROWTH CONDITIONS

3.2.1 FUNGAL AND BACTERIAL STRAINS, PLASMIDS, PRIMERS, AND SYNTHETIC DNA

Fungal and bacterial strains, plasmids and synthetic DNA used are listed in Table 3.2.1. Primers are listed in Table 3.2.2. All primers were obtained from IGT.

Table 3.2.1. A list of the strains and plasmids that were used.

Strain or Plasmid	Relevant Characteristics	Source or reference	Lab code
Bacterial Strains:			
Escherichia coli (DH5α)	DH5α Competent Cells have been prepared by a proprietary modification of the procedure of Hanahan (1983) F- Φ80lacZΔM15 Δ(lacZYA-argF) U169 recA1 endA1 hsdR17 (rk-, mk+) phoA supE44 λ-thi-1 gyrA96 relA1.	MAX Efficiency™ DH5α Competent *E. coli* Cells (Invitrogen, #18258012)	N/A

***Saccharomyces cerevisiae*:**			
Wild-type copy number strain mating type alfa	Laboratory haploid strain derived from NOY398, *MATα ade2-1/ade2-1 ura3-1/ura3-1 his3-11/his3-11 trpl-1/trpl-1 leu2-3,112/leu2-3,112 cani-100lcani-100 rpal35::LEU2/ RPAJ35*	Nogi *et al.*, 1991	YAG135
Wild-type copy number strain fob1-::His+ (150 copies)	Laboratory strain NOY408-1bf same as NOY408-1b except *fob1-::His+*	Kobayashi *et al.*, 1998	YAG92
30 copies strain fob1-::His+	Laboratory strain same as NOY408-1b except *fob1-::His+* and 30 rDNA gene copies. Made using Hygromycin treatment to generate low rDNA copy strain.	Derived from Ide, *et al.*, 2010	YAG80
40 copies strain fob1-::His+	Laboratory strain same as NOY408-1b except *fob1-::His+* and 40 rDNA gene copies. Made using Hygromycin treatment to generate low rDNA copy strain.	Ide, *et al.*, 2010	YAG95
80 copies strain fob1-::His+	Laboratory strain same as NOY408-1b except *fob1-::His+* and 80 rDNA gene copies.	Ide, *et al.*, 2010	YAG98
Plasmids:			
HO-Poly-KanMX4-HO	Yeast plasmid for integration of target sequence at HO-locus with a Kanamycin resistance gene that allows selection with G418.	Addgene #51662	N/A
Synthetic DNA:			
mRuby2 construct	*mRuby2* red fluorescent protein gene derived from *Entacmaea quadricolor* flanked with rDNA sequence.	Austen Ganley Lab	N/A
mTagBFP2 construct	*mTagBFP2* blue fluorescent protein gene derived from *Entacmaea quadricolor* flanked with rDNA sequence.	Austen Ganley Lab	N/A
EGFP construct	*EGFP* green fluorescent protein gene derived from *Aequorea victoria* flanked with rDNA sequence.	Austen Ganley Lab	N/A

Table 3.2.2. Primers' names, their sequences, and their respective reactions.

Primer name	5'-3' primer sequence	Lab code
HO cloning F *Bam*HI	TTTGGATCCTGCACTCTTCTTCTGAAGAGTT Forward primer for synthetic DNA constructs (Table 3.2.1) is designed to amplify the fluorescent protein gene sequence. It binds a few base pairs before the fluorescent protein gene. Include a BamHI active site.	N/A
HO cloning R *Bgl*II	AAAAGATCTATCAAGTAGTAGCAACCCAATG Reverse primer for synthetic DNA constructs (Table 3.2.1) is designed to amplify the fluorescent protein gene sequence. It binds a few base pairs after the fluorescent protein gene. Include a *Bgl*II active site.	N/A
pFA6:KanMX2/ 345 S r	GATGTGAGAACTGTATCCTAGC A reverse primer that binds in the middle of the Kanamycin resistance gene presents the HO-Poly-KanMX4-HO plasmid.	PAG292
HO-Internal-F	TGGCAAAGAAATCGATGCATACC A forward primer that binds in a sequence upstream of the insertion area of the fluorescent protein genes in the HO-Poly-KanMX4-HO plasmid.	N/A
rDNAScSp_F2	ATCTCTTGGTTCTCGCATCG A forward primer that anneals to *rDNA* gene.	PAG672
rDNAScSp_R2	GGAAATGACGCTCAAACAGG A reverse primer that anneals to *rDNA* gene.	PAG673
RPS3ScSp_F2	CACTCCAACCAAGACCGAAG A forward primer that anneals to the *RSP3* gene.	PAG669
RPS3ScSp_R2	GACAAACCACGGTCTTGAAC A reverse primer that anneals to the *RSP3* gene.	PAG668

3.2.2 **INITIAL GROWING CONDITIONS**

Escherichia coli and *Saccharomyces cerevisiae* strains were refreshed from freezer stocks held at -80°C. Part of the stock was scrapped off and spread onto LB Agar and YPD agar, respectively. *E. coli* strains were incubated for 24 hrs at 37 °C and then stored at 4 °C. *S. cerevisiae* strains were incubated for 72 hrs at 30 °C and then stored at 4 °C.

3.2.3 ***ESCHERICHIA COLI* GROWING CONDITIONS**

Escherichia coli cultures were grown at 37° C overnight on LB agar plates, LB broth or LB agar and ampicillin plates (Section 3.1.1), shaking at 180 rpm.

3.2.4 **SACCHAROMYCES CEREVISIAE GROWING CONDITIONS**

General Growing conditions

Saccharomyces cerevisiae cultures were grown at 30° C on YPD agar plates, YPD broth (Section 0), YNB -HIS or YNB -LEU (Section 0). Liquid cultures were shaking at 180 rpm. The *fob1*- mutated strains (Table 3.2.1) were plated in selective YNB -LEU or -HIS to isolate proper individual colonies.

Growing conditions for rRNA gene copy number distribution estimation

Cultures used to estimate the rRNA copy number distribution were propagated over five days (≥60 generations). The propagation uses the following procedure:

Step 1: A starting Wild-type copy number strain *MATα* NOY398 culture was used to inoculate 50 mL YPD for 24 hr.
Step 2: Then, When the culture was at 2 OD 600nm (Section 3.5), five microliters of culture were used to inoculate 50 mL of fresh YPD and grown for a further 24 hr.
Step 3: Step 2 was repeated for the other four days.

On the final day, the culture was diluted using 1:100, 1:100 and 1:3 serial dilution and 100 µL of the last dilution were used to plate the culture in YPD agar at 30 °C for 72 hrs (Section 3.2.4). The dilution and culture allowed the isolation of individual colonies for copy number estimation. The colonies were kept at 4 °C until they were used to isolate the gDNA as described in Section 3.4.2.

Growing conditions for testing sensitivity to G418 of *S. cerevisiae* strains.

A starting culture used to estimate the sensitivity to G418 with colonies of the refreshed strains was growth in YPD as described in the general growing conditions. The cultures were then diluted three times using 1:10 dilutions. Then, 5 mL of all dilutions and the undiluted culture from all the strains were platted in YPD agar plates as described in the general growing conditions. The various strains were spotted at different dilutions onto YPD with different G418 concentrations (200 µg/mL, 250 µg/mL, 400 µg/mL, and 500 µg/mL), and it was assessed whether they grew or not.

3.3. OPTICAL DENSITY MEASUREMENTS

Optical density was measured at 600 nm (Biowave CO8000 Cell Density Meter) to monitor the growth of *E. coli* and *S. cerevisiae* cultures. When the OD_{600} was higher than 2, 1:10 dilution was done to ensure the proper measurement.

3.4. DNA ISOLATION

3.4.1 EXTRACTION OF PLASMID DNA FROM BACTERIAL CULTURES

Plasmid DNA was isolated from transformed using the GeneJET Plasmid Miniprep Kit (Thermo Scientific #K0503). The starting culture was cultivated in LB-amp media overnight (Section 3.2.3), and the cells from 5 mL overnight culture in 50 ml falcon tubes were pelleted at 3184 rcf, and the manufacturer's procedure was followed.

3.4.2 ISOLATION OF *S. CEREVISIAE* GENOMIC DNA

Isolation of *S. cerevisiae* genomic DNA was performed using a modification of the Drumonde-Neves et al. (2013) procedure in 96 well plates. Individual colonies were each inoculated into 100 µL of YPD (Section 3.1.4) in a 96 well plate and grown at 30 °C for 18 hrs. Cells were pelleted by centrifugation at 2129 rcf for 2 mins. Half the supernatant was discarded, and 10 µL of Buffer I (Section 3.1.4) were added. The pellet was resuspended by vortexing and incubated at 37 °C for 30 mins. 10 µL of solution buffer II (Section 3.1.4) was added, and then the plate was vortexed, spun down and incubated at 65 °C for 5 mins. Then the DNA was precipitated as described in Section 3.9.2.

3.5. DNA QUANTIFICATION

3.5.1 DNA QUANTIFICATION BY ETHIDIUM BROMIDE STAINING AND STANDARDS

Samples were loaded on a gel alongside GeneRuler DNA Ladder Mix (500 ng/µL) standards (Thermo Scientific, #SM0331) (Section 3.1.4). The gel was photographed (Section 3.10), and the fluorescent intensity of standards and samples were compared manually to estimate DNA concentration.

3.5.2 **SPECTROPHOTOMETRIC QUANTIFICATION**

Spectrophotometric quantification was performed for routine quantification of DNA concentration using a NanoPhotometer N60/N50 spectrophotometer (Implen) at 260 nm and 280 nm. The absorbance ratio, A_{260}/A_{280}, was used to estimate the DNA purity.

3.6. DEPHOSPHORYLATION

The digested plasmid was dephosphorylated by mixing 1 µg of plasmid with 1 µL rAPid Alkaline Phosphatase (Roche #04898133001), 2 µL 10x rAPid buffer (Roche #04898133001) and sterile water in a final volume of 20 µL. The mixture was incubated at 37°C for 10 mins, vortexed, and then the phosphatase was heat-inactivated by incubating at 75°C for 2 mins.

3.7. DIGESTION OF DNA USING RESTRICTION ENZYMES

DNA was digested with restriction enzyme following the manufacturer's instructions unless a different condition was specified. The restriction enzymes used are listed in Table 3.7.1.

Table 3.7.1. Used restriction enzymes.

Name	Cut site	Company	Buffer
*Bgl*II	5' – A|GATCT – 3' 3' – TCTAG|A – 5'	NEB	NEB 3.1 (#B72035)
*Bam*HI	5' – G|GATCC – 3' 3' – CCTAG|G – 5'	NEB	NEB 3.1 (#B72035)
*Xba*I	5' – T|CTAGA – 3' 3' – AGATC|T – 5'	NEB	NEB CutSmart (#B72045)
*Pac*I	5' – TTAAT|TAA – 3' 3' – AAT|TAATT – 5'	NEB	NEB CutSmart (#B72045)
*Spe*I	5' – A|CTAGT – 3' 3' – TGATC|A – 5'	NEB	NEB CutSmart (#B72045)

3.8. LIGATION

Ligation was performed using a ratio of 18:1 insert-plasmid ratio. The difference in length of both the plasmid and the insert is about 6 times. Therefore, to achieve the ratio of 3 µL of digested (Section 3.7**Error! Reference source not found.**) and dephosphorylated (Section REF _Ref99546843 \r \h * MERGEFORMAT 3.6) plasmid, 9 µL of the insert at ~20 ng/µL, 1 µL T4 DNA Ligase (Thermo Scientific #EL001), and 2 µL 10x T4 DNA Ligase Buffer (Thermo Scientific #EL001), made to a final volume of 20 µL sterile water. The ligation was

incubated at 16 °C for 24 hrs. The ligase was then heat-inactivated at 65 °C for 10 mins, and the ligation was stored at -20 °C.

3.9. DNA PRECIPITATION

3.9.1 **ROUTINE PRECIPITATIONS**

DNA to be precipitated was mixed with 2 µL of Glycogen (Thermo Scientific #R0561), 0.1 volumes of 3M sodium acetate (pH 7.0) 3 volumes of 95% ethanol. This mixture was left either at -20 °C for 2 hrs or -80 °C for 1 hr. The DNA was pellet by centrifugation at 13000 g for 15 mins in a microcentrifuge. Then, the pellet was rinsed with 70% ethanol 2 times using the following procedure:

Step 1: 100 µL of 70% ethanol were added to the pellet.
Step 2: The mix was centrifugated at 13000 g for 5 mins.
Step 3: The supernatant was discarded.

Finally, the pellet was dried and resuspended in either water or TE (1/10) buffer (Section 3.1.4).

3.9.2 ***SACCHAROMYCES CEREVISIAE*** **GENOMIC DNA PRECIPITATION**

S. cerevisiae Genomic DNA precipitation was done in 96 well plates using a modification of the procedure of Drumonde-Neves et al. (2013). 8 µL of potassium acetate 5M were added to each sample of isolated gDNA, mixed by pipetting, and incubated at -20 °C for 8 mins. The 96 well plates were centrifugated (room temperature) at 2129 rcf for 15 mins, the supernatant was transferred to a new plate, 25 µL of isopropanol was added to each sample, and this was incubated at room temperature for 10 mins. The DNA was pelleted by centrifugation at 2129 rcf for 10 mins, the supernatant was discarded, and the pellets were rinsed twice with 50 µL 70% ethanol using the same procedure as above (Section 3.9.13.9.1) but at 2129 rcf. Then the pellets were air dry, resuspended in 50uL of (1/10) TE buffer + RNase (Section 3.1.4), incubated at 37 °C for 1 hr, and stored at -20 °C.

3.10. POLYMERASE CHAIN REACTION (PCR)

All PCR reactions were performed using the KAPA2G Robust HotStart PCR Kit (Roche, #KR0380) and PCR grade water (Solis Biodyne #water-100) according to the manufacturer's instructions. The volumes and proportions of the reagents used are shown in Table 3.10.1. The

reagents were mixed in a mastermix and then dispensed into individual samples in a PCR hood (Airstream PCR cabinet Laminar flow ESCO).

Table 3.10.1. PCR reaction mixes.

Reagent/Template	25 µL reaction	50 µl reaction
PCR-grade water	Up to 25 µL	Up to 50 µL
5X KAPA2G GC Buffer	5.0 µL	10 µL
10 mM KAPA dNTP mix	0.5 µL	1 µL
10 µM Forward Primer	1.25 µL	2.5 µL
10 µM Reverse Primer	1.25 µL	2.5 µL
5 U/µL KAPA2G Robust HotStart DNA Polymerase	0.1 µL	0.2 µL
Template DNA	As recommended	As recommended

The annealing temperature for the primers was assessed by gradient PCR in a Veriti™ 96-Well Fast Thermal Cycler (Applied Biosystems™ # 4375305). An annealing temperature of 55 ˚C was used for all primer sets. The thermocycling conditions applied for all PCR reactions are shown in Table 3.10.2. A 9700 thermocycler GeneAmp PCR system (Applied Biosystems™) was used for routine PCR reactions.

Table 3.10.2. Thermal cycling protocol.

Step	Temperature	Duration	Cycles
Hotstart activation	95 °C	3 min	1
Denaturation	95 °C	15 sec	30
Annealing	55 °C	20 sec	
Extension	72 °C	2 min	
Final Extension	72 °C	2 min	1

3.10.1 **COLONY PCR**

Bacterial DNA was extracted by inoculating individual colonies in 30 µL of sterilised water. The mix was incubated at 95 °C for 5 mins and then placed immediately on ice. Then a PCR reaction was performed as described in Section 3.10.

3.11. AGAROSE GEL ELECTROPHORESIS

3.11.1 **AGAROSE GELS**

1% and 0.8% agaroses were made using 1x SB buffer (Section 3.1.4). The agarose was melted in a microwave and, after a cold down period, was poured. Horizontal agarose gels were run in MupidEXu Submarine Electrophoresis System (Takara Bio) at 100 V for 30 mins in 1x SB buffer (Section 3.1.4).

3.11.2 **DNA VISUALISATION**

DNA was visualised using a Gel Doc XR+ Gel Documentation System (Biorad), with ethidium bromide present in the loading dye (Section 3.1.4). The gel was stained in an ethidium bromide bath (1 µg/ml in sterilised water) after electrophoresis for 20 mins and washed in water for 10 mins if the DNA concentration was low.

3.12. TRANSFORMATIONS

3.12.1 **PREPARING COMPETENT *E. COLI* CELLS**

Preparing pre-starter and starter culture

The pre-starter culture was grown in 10 mL of LB media (Section 3.1.1) aliquoted in a 100 mL conical flask. A single colony was taken from the stock culture of *E. coli* (Section 3.2.2) and used to inoculate the 10 mL of media using a sterile toothpick. The culture was incubated for 24 hrs. After incubation, samples were taken as described in Section 0 to measure the $OD_{600.}$ Starter culture was grown in 250 mL of LB media (Section 3.1.1) aliquoted in a 1 L conical flask. 2.5 mL of the pre-starter culture was inoculated in 250 mL of media. The flask was incubated (Section 3.2.2) and $OD_{600.}$ The OD was measured after 1:30 hrs and then every hr. When the culture reached an OD_{600} higher than 0.5, it was placed on ice for 15 mins.

Making cells competent

Starter culture was harvested by dividing the 250 mL culture equally into five 50 mL Falcon tubes and centrifuging at 4,000 g, for 20 mins, at 4ºC. The supernatant was discarded, and the pellet was gently resuspended by pipetting in 1 mL of Transformation buffer I (Section 3.1.4). Further, 49 mL of Transformation buffer was added. The five tubes were covered with ice and placed on a seesaw shaker for 1 hr. Cells were pellet again as previously, the supernatant was

discarded, and the pellet was resuspended in 2 mL of Transformation buffer II (Section 3.1.4). The tubes were placed on ice. Cells were then consolidated into one falcon tube, and 400 mL of the mixture was aliquoted into pre-chilled 1.7 mL microcentrifuge tubes. The tubes were immediately dropped into a container of liquid nitrogen. Frozen competent cells were stored at -80ºC.

3.12.2 *E. COLI* TRANSFORMATIONS

A vial of existing *Escherichia coli* competent cells was taken from the -80 ˚C freezer, placed on ice, and allowed to thaw slowly. 3 µL of plasmid vectors or water were dispensed into pre-chilled 1.7 microcentrifuge tubes, and 100 µL of competent cells were dispensed per tube. The mixture was mixed by flicking and then returned to the ice for 45 minutes. The cells were then heat-shocked by incubation at 37 °C for 1.5 minutes, returned to the ice for 2 minutes, then resuspended in 900 µL of LB + glucose (20 mM) medium (Section 3.1.4), and incubated at 37 °C for 1 hour. Cells were pelleted at 16.100 rcf for 30 seconds, 900 µL of the supernatant was discarded, and cells were resuspended in the remaining 100 µL of solution. Finally, the 100 µL cells were plated onto LB-amp Agar plates using sterile glass beads, the plates were dried, the beads were removed, and the plates incubated overnight (Section 3.2.2).

3.12.3 PREPARING COMPETENT *SACCHAROMYCES CEREVISIAE* STRAINS

A *S. cerevisiae* colony was inoculated into 10 mL of YPD medium and grown overnight (Section 3.1.3). The OD_{600nm} was measured, and sufficient culture was used to inoculate 10 mL of fresh prewarmed YPD medium at 30 °C to a final OD_{600} of 0.1. The new culture was growing as described before (Section 3.1.3), and the OD was measured (Section 0) after 1:30 hrs and then every hr. When the culture reached an OD_{600} between 0.6 and 0.8, the 10 mL were spun at room temperature for 5 mins at 3000 rpm. The supernatant was discarded in a sterile condition, and the pellet was kept. Pellet was resuspended in 5 mL of sterile water. Then the pellet was vortexed, and the cells were spun down as previously. Water was discarded, and the cell pellet was resuspended in 2.5 mL of TE (1/10) – LiOAc 100 mM buffer (Section 3.1.4). Cells were vortex and spun down as previously. The buffer was discarded, and all residual drops were removed. Then cells were resuspended in 100 µL buffer. The cell suspension was transferred to 1.5 mL tubes.

3.12.4 **SACCHAROMYCES CEREVISIAE TRANSFORMATIONS**

Transformations of *S. cerevisiae* were done the same day the cells became competent. Before the transformation, the Salmon Testis DNA (DNA carrier) (Section 3.1.4) was denatured for 10 mins at 95°C and then kept on ice.

3 µg of plasmid vector DNA (maximum volume 10 µl) and 5 µL of the denatured carrier DNA were added to the 100 µL of competent cell suspension. The suspension was mixed by flicking the tube gently. Negative control was done using just the carrier DNA without cells. Then, the cell suspensions were added to 700 µL of the TE (1/10) – LiOAc 100 mM – PEG 40% solution. The suspension was vortexed and then incubated for 15 mins at 42 °C in a thermomixer (Eppendorf Thermomixer C). After the heat shock, the cell suspension was centrifuged for 1 min at full speed, and the supernatant was discarded by pipetting. The cell pellet was resuspended in 150 µL of sterile water by pipetting, and cells were spread with sterile glass beads on one Petri dish with selective YPD-agar G418 (Section 3.1.2). When plates were dried, the beads were discarded, and the plate was incubated upside down for 4 days (Section 3.2.4).

3.13. DIGITAL DROPLET PCR

The digested *S. cerevisiae* genomic DNA was digested with *Xba*I as described (Section 3.7), but with an incubation time of 24 hrs and adding 0.5 µL of the enzyme. The incubation time was long to ensure complete digestion of the genomic DNA that can be affected by impurities after the isolation. The digested gDNA were diluted until they reached a concentration of 2 pg/µL by serial dilution. Droplet generation and endpoint PCR were performed following the manufacturer's instructions (BioRad). Samples readings were done using a QX200 droplet reader and quantification using QuantaSoft Analysis Pro (v. 1.0.596). rDNA copy number was determined by calculating the (*rDNA* copy/µL) / (*RPS3* copy/µL) ratio.

Chapter 4. MODELLING rDNA COPY NUMBER DISTRIBUTIONS IN A HAPLOID *S. CEREVISIAE* POPULATION USING AN EXPERIMENTALLY ESTIMATED DISTRIBUTION.

4.1. GROWING *S. CEREVISIAE* WILD-TYPE COPY NUMBER STRAIN *MATα* CULTURES AND ISOLATION OF gDNA FOR COPY NUMBER DISTRIBUTION ESTIMATION

This project wanted to estimate the rDNA copy number distribution of a wild-type haploid *S. cerevisiae* population so that this distribution could be fitted by the model I developed. To do this, I used strain *MATα* NOY398 (Section 3.2.1). I wanted the copy number distribution to represent that of a "natural" population. Therefore, I started from a single colony and grew this colony in liquid culture. After 24 hours of growth, 5 µL were transferred to 50 mL of fresh culture. Thus, 11 generations/day was reached. This procedure was repeated for 6 days, meaning the population was grown for ~66 generations. After the ~66 generations, the population was diluted and plated as described in Section 3.2.4 for copy number estimation. Genomic DNA was isolated from 79 colonies (Section 3.4.2).

4.2. ESTIMATING rDNA COPY NUMBER USING ddPCR

I next wanted to estimate the rDNA copy number for each of the individual 79 colonies to obtain a copy number distribution. To do this, I decided to use digital droplet PCR because it is a highly sensitive technique that offers an absolute quantification of target nucleic acid copies per sample volume. Estimating copy numbers using ddPCR requires an accurate estimation of the gDNA concentration. I measured the gDNA concentration by spectrophotometer readings (Section 3.5.2). The gDNA samples' concentrations were in the range of 50 to 350 ng/µL (Appendix B. Table 1). Estimating the rDNA copy number using ddPCR also requires complete digestion to ensure the rDNA copies are separated into individual fragments and can be quantified as a single gene in each droplet. Moreover, a huge fragment of multiple copies could not fit into a single droplet. To achieve this, the gDNA from each colony was digested using

*Xba*I (Section 3.13). After digestion, an aliquot was checked by gel electrophoresis (Section 3.10) to assess the digestion (Fig. 4.2.1).

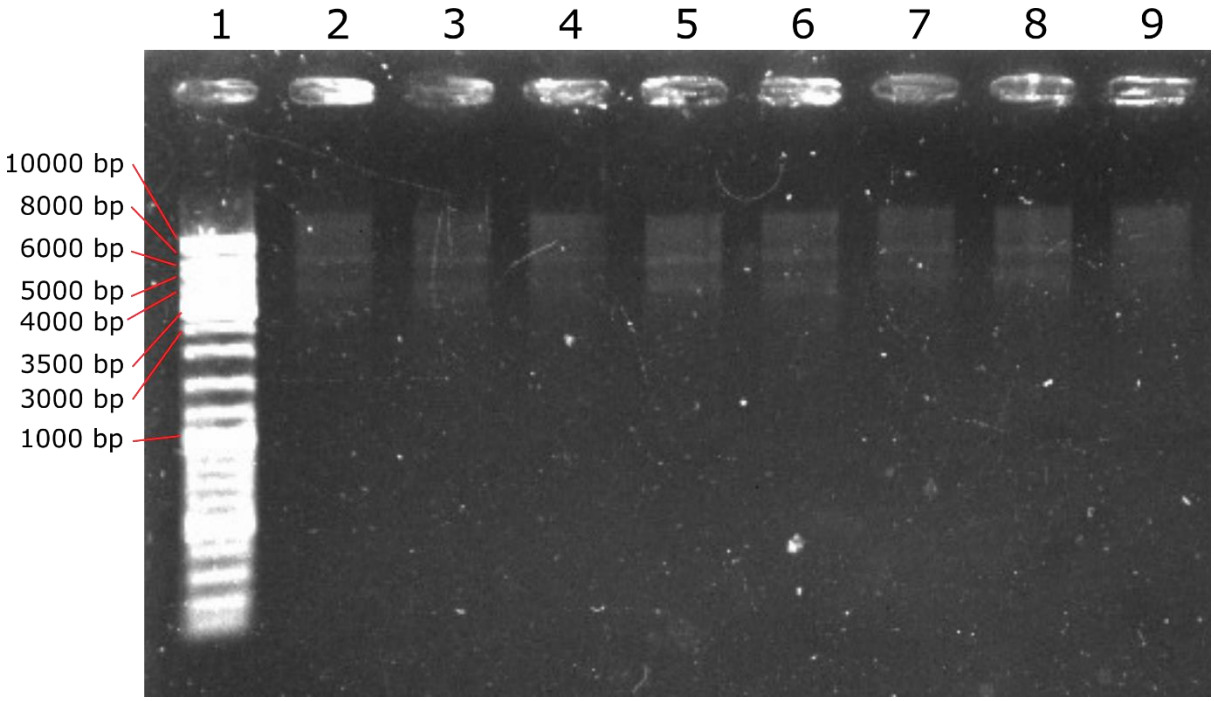


Figure 4.2.1 Representative *S. cerevisiae* genomic DNA digestions with XbaI. Lane 1 is GeneRuler DNA Ladder Mix (Thermo Scientific, #SM0331). Lanes 2-9 correspond to gDNA extractions from colonies 32-40. The smears indicate successful *Xba*I digestion.

The copy number estimation by ddPCR uses absolute quantification of the copies/µL of target genes. A reference gene with one copy of RPS3 (Single copy gene) was used to estimate the copies of the rDNA. This gene was chosen because the previous test was used to estimate the rDNA, and it is known that it just has a copy in the genome. Then the estimation of copies/µL was normalised by dividing the rDNA copy number estimation by the estimation of the *RPS3* (*rDNA* copy/µL) / (*RPS3* copy/µL) ratio. Reliable estimations of rDNA copy number by ddPCR require an appropriate dilution that makes the rDNA copies diluted enough to be quantified but not too diluted, which makes difficult the estimation of copies/µL of the single copy gene. Therefore, an initial test with seven samples was done to assess the optimal gDNA concentration for copy number estimation for target genes. Serial dilutions were done to reach a concentration of 2 pg/µL. Two set of plasmids, one per gene (primers rDNAScSp_F2, rDNAScSp_R2, RPS3ScSp_F2 and RPS3ScSp_R2) were used to estimate the copies/µL of each gene by ddPCR (Section 3.13). Copy number estimates for both genes were performed in independent wells in a plate of 96 wells, with two replicates per sample. An aliquot of 2µL of

the Diluted gDNA was dispensed in each well. A reliable estimation of the copies/µL could not be obtained for both target genes (data not shown). Comparison with previous data from (Sharma, 2021), in which copies were estimated using both targets and positive droplet counts lower than 15, suggested a sub-estimation by 100-fold of the copy number. Possible explanations could be incomplete digestion or samples being too dilute. Incomplete digestion usually manifests as two different groups of positive intensities in the droplet readings, which was not seen in my results. Therefore, too high dilution of the samples is the likely cause, and to test this, sample gDNA concentrations were diluted to 100 pg/µL, and the experiment was repeated with the seven samples. There was an improvement in the positive droplet counts, and the estimations were closer to previous experiments (data not shown). Hence, the gDNA for samples was diluted to 100 pg/µL, and the rDNA copies were estimated as described in Section 3.13. The mean copy number was 231 *SD*= 65.69, the minimum copy number was 104, and the maximum was 394. The copy number values were grouped in bins of 10 (Fig. 4.2.2) to keep the distribution structure the same as that used for developing the model, and the most frequent copy numbers correspond to the values in the range 185 - 195 (Fig. 4.2.2).

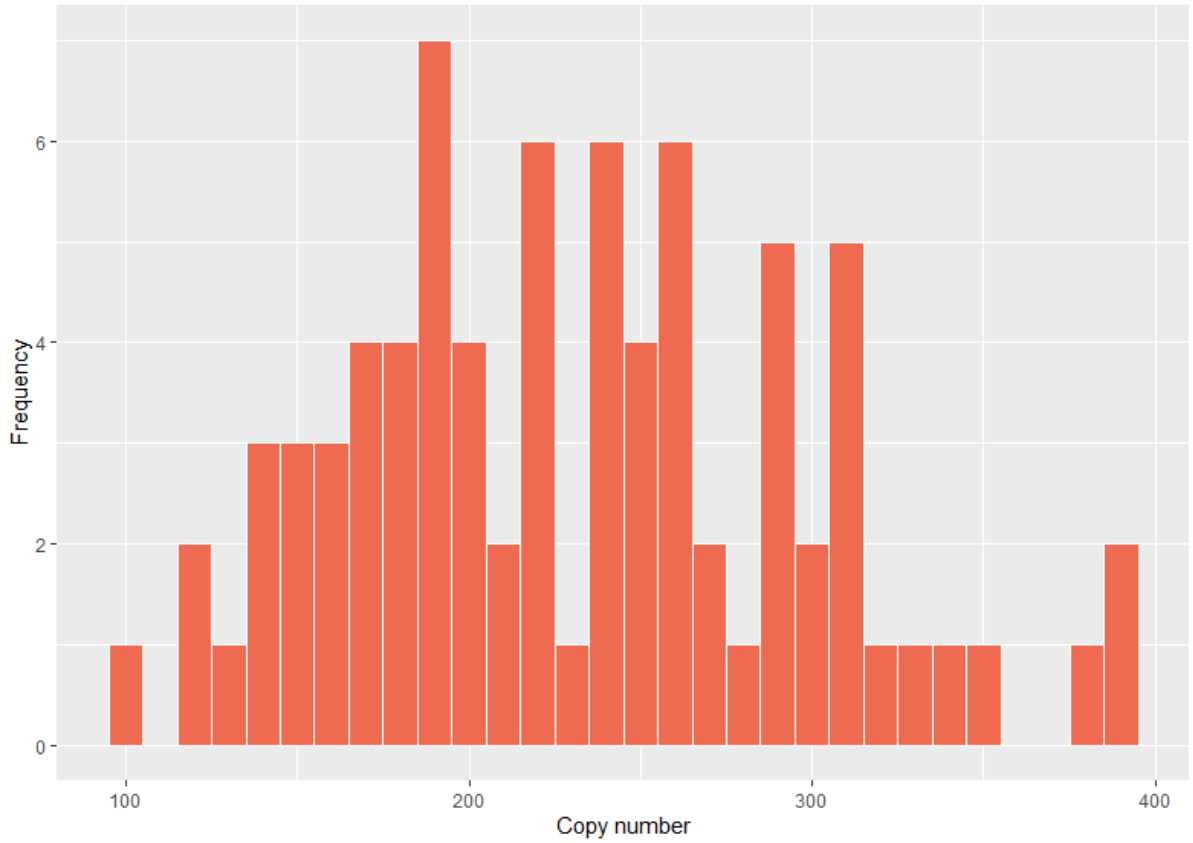


Figure 4.2.2 Observed rDNA copy number distribution. Distribution of rDNA copy numbers for 79 *S. cerevisiae* colonies. The values obtained from the ddPCR readings were

approximated to the nearest integer, then grouped into bins of 10, and plotted as a frequency histogram.

4.3. FITTING THE MODEL TO THE DISTRIBUTION DATA

After obtaining the experimental data, I wanted to use them to fit the model described in section 2 to estimate the parameters that best represent the experimental data. Parameters that were tested were b of Γ (maximum misalignment's percentage for deletion), and w_1 (Lower limit fitness function) and w_2 (Upper limit fitness function), duplication recombination rate α, deletion recombination β, *sDup* and *sDel* (duplication and deletion slopes that allow recombination rates vary in function of the copy number, respectively). To perform the fitting, an implementation of the model was written in C++, and that implementation was loaded as an R-function using RCPP to create an interface between the two languages (The code is available in the GitHub project https://github.com/ivanhc1993/rDNADynamics.git). Model fitting was done using a custom R-script (Appendix A Script 2) based on the package "Optimx", which includes different optimisation methods. I used the quasi-Newton method "L-BFGS-B" optimisation method, as this allows parameter constraints to be set. The optimization algorithm adjusts the parameters to minimise the mean squared error (*MSE*) (Equation 4.3.1), and the observed y_i (Values than were obtain from the copy number distributions) and predicted $\hat{y}_i$ (Values that are estimated using the model) values were binned in groups of 10.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Equation 4.3.1 Mean squared error (*MSE*)

I first tested whether the model estimates the same maximum misalignment percentage for deletion (75%) as that observed in Ganley/Kobayashi. To do this, the model was run with the full dataset (79 values) for 1000 generations with the recombination rates fixed to α= 0.00354, β=0.00458, the *Xm*= 0.03, *b* of *H* = 24% (0.2417582), n_0= 23, n_T=34, 400 as the maximum copy number evaluated, and the starting distribution of the population was uniform. The

starting distribution of the population was uniform. Only b of Γ (maximum misalignment's percentage for deletion), and w_1 (Lower limit fitness function) and w_2 (Upper limit fitness function) left as free parameters (Section 2.8). The starting values for the free parameters were 0.7, 100, and 140, respectively, and their values were constrained to b of Γ (0.01-0.99), w_1 (0-399) and w_2 (0-400). After fitting the function, the parameters obtained were b of Γ = 0.99, w_1 = 60, and w_2 = 148, and the *MSE* was 8.08×10^{-4} . Figure 4.3.1 shows the observed rDNA copy number distribution and the fitted model. The model shows a good fit to the data in the range of 200 – 400 copies, but the left part is not. A similar peak of frequencies with the observed data was achieved for the model. A different estimate was found for parameter b of Γ compared with the one in Ganley/Kobayashi.

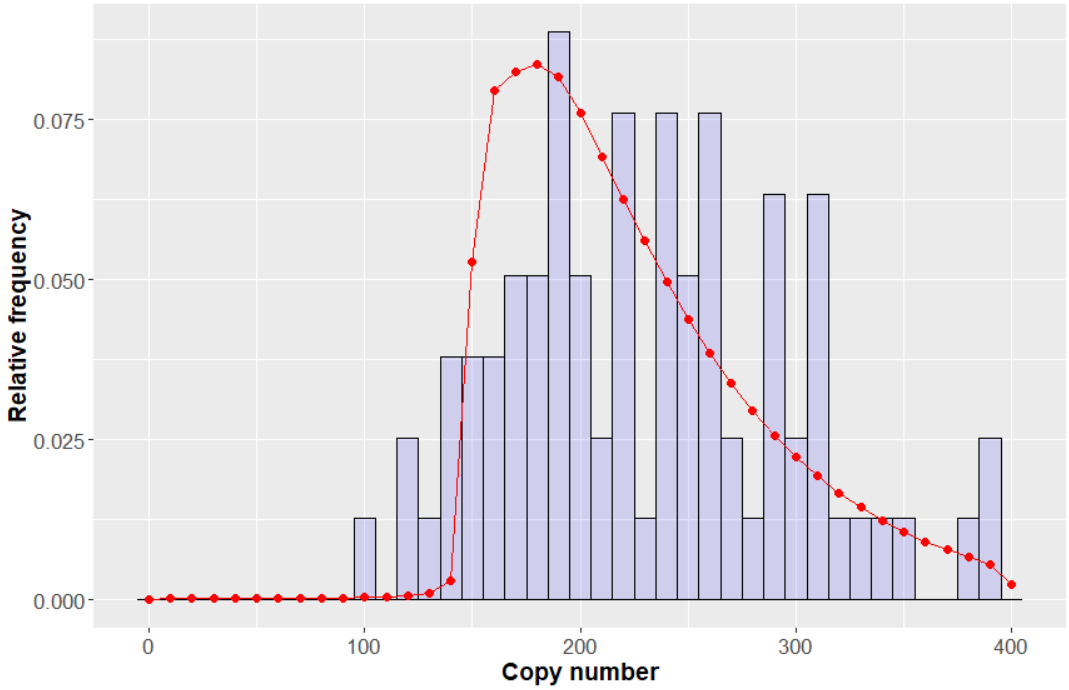


Figure 4.3.1 Frequency distributions of observed rDNA copy number and the best-fitted model prediction. The predicted values of the model and the model outputs were grouped in bins of 10. The lines approximate the continuous predictions.

4.3.1 **DETERMINING EQUILIBRIUM CONDITIONS FOR THE SIMULATION**

wanted to model the copy number distribution in equilibrium or a quasi-equilibrium state (When there is a small variation between estimation at different generations but small). Therefore, I tested whether the simulation reached equilibrium after 1000, 4000, 5000 and 1000

generations. For this test, I used the same parameter values as above, including the values of b of Γ, w_1, and w_2 obtained from the fitting. The obtained copy number distributions were then compared to see if increasing generations produced any change in the distribution (Fig. 4.3.1.1). This shows no change in distribution after 4000 generations, suggesting a quasi-equilibrium state at least is reached sometime before 4000 generations. I also wanted to see if there are differences in the parameter estimates between generation times. To do this, I ran the model with 1000 and 5000 generations using a bootstrapping analysis (Appendix B Script 2) to evaluate the parameter distributions. To perform the bootstrapping, the model was run with 70% of the experimental rDNA copy number data points to fit the model, and the other 30% were used as testing data to estimate the *MSE*. The total of bootstrap iterations was 190 (Bootstrap iterations correspond to estimation with a subset of the data set sampled randomly). More dispersed estimations of the parameters were observed in the model run for 1000 generations, but the mean estimates for parameters b and w_1 were similar between models (figure 4.3.1.2). In contrast, the mean estimate of the parameter w_2 differs between models, with the model run for 1000 generations having the lowest average. The differences in w_2 values, the high variance in parameter estimates and the different distribution achieved for the simulations run for 1000 generations suggest that the model should be run with a higher number of generations than 1000. Therefore, for all subsequent simulations, the number of generations was set to 5000, which likely represents a quasi-equilibrium state but reduces the computational load of running the model compared to using more generations.

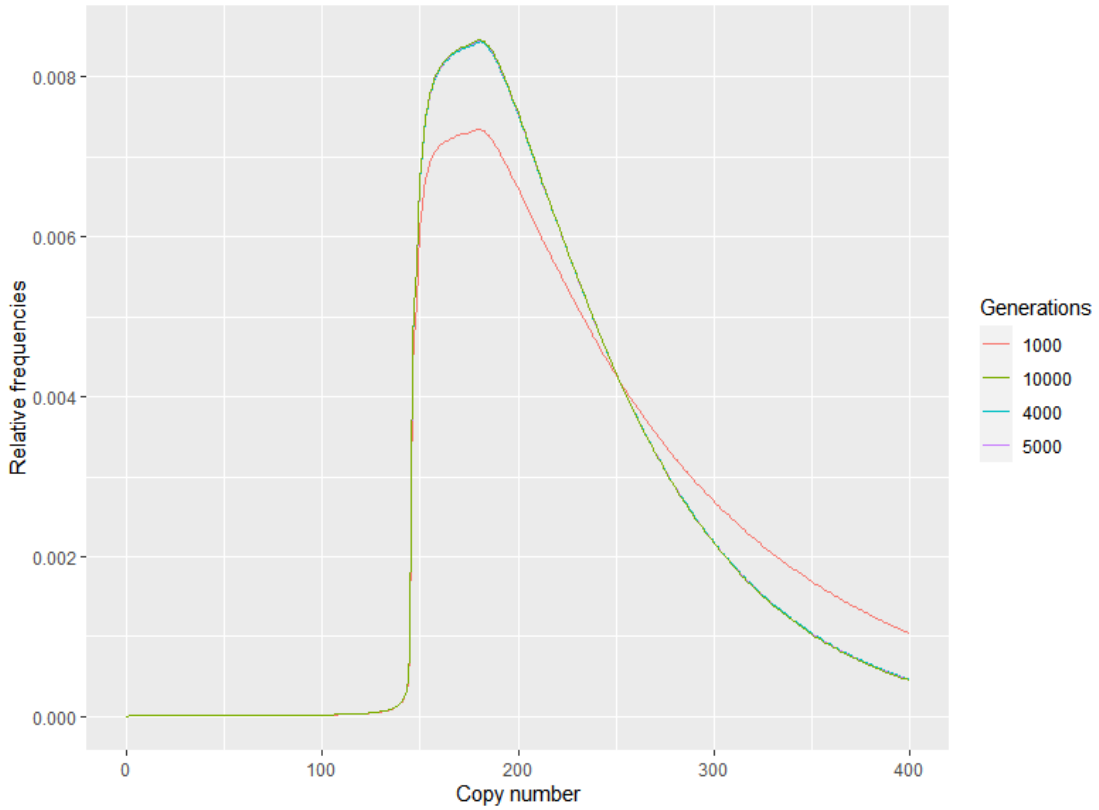


Figure 4.3.1.1 Copy number distributions generated from different generations numbers. Copy number distributions after different generations times (1000, 4000, 5000, 10000) with the same parameters, the x-axis corresponds to the copy number, and the y-axis corresponds to the relative frequencies. Similar distributions are obtained for generation numbers higher than or equal to 4000. This comparison suggests the model achieves a quasi-equilibrium state with at least 4000 generations. For this test, I used the same parameter values of Section 4.3, including the values of b of Γ, w_1, and w_2 obtained from the fitting.

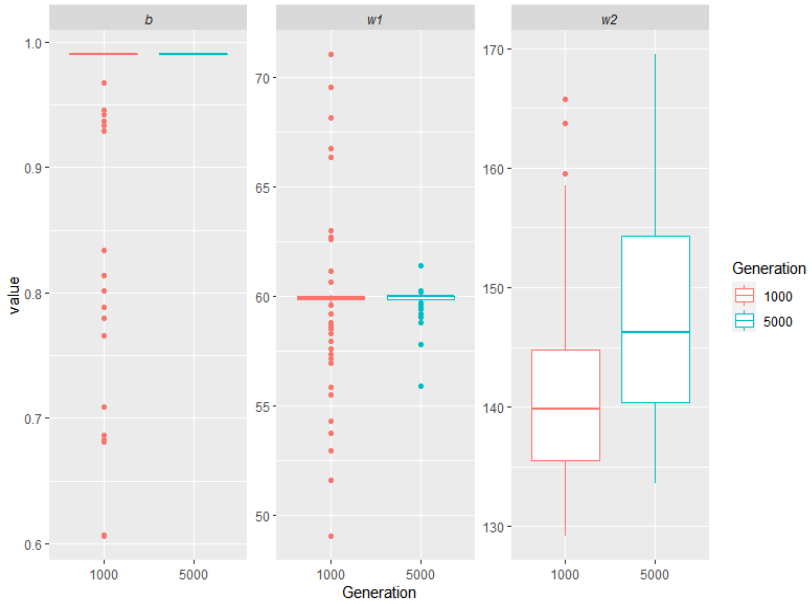


Figure 4.3.1.2 Parameters distributions for 1000 and 5000 generations simulations. Distribution of the parameters estimated from bootstrapping analysis at different generations times (1000, 5000) with the same parameters, the x-axis corresponds to the copy number, and the y-axis corresponds to the relative frequencies. Parameters b and w_1 have a similar value between the two generation runs, but the 1000 generation model run has a higher variance. There was a difference in the parameter w_2 estimates between the two generation run. For this test, I used the same parameter values of Section 4.3 that were not fixed.

4.3.2 **COMPARING DIFFERENT MODEL VERSIONS.**

I wanted to answer three questions: (1) When recombination rates are fixed, are there significant differences between the model with a fixed b of Γ to 75% and the model with this parameter as free? **(2)** if the observed distribution can be better by making the recombination rates free parameters that can be fitted rather than using the values calculated from a previ ous study Section 2.3; and **(3)** if allowing the recombination rates to vary as a function of the copy number improves the model fitting, given that *S. cerevisiae* has a mechanism to amplify rDNA copy number when it is low (Iida and Kobayashi, 2019).

Five models were compared using the bootstrapping approach outlined in section 4.3.1 to address these questions. For all the comparison 5000 generations, with the recombination rates fixed to α= 0.00354, β=0.00458 (Only in models were not free parameters), the Xm= 0.03, b of H = 24% (0.2417582), n_0= 23, n_T=34, 400 as the maximum copy number evaluated, and the starting distribution of the population was uniform. These parameters were used for all the

simulations if they were not free. The first model was a minimal version where parameter *b* of Γ (maximum misalignment percentage for the deletion) is fixed while w_1 (lower limit of the fitness function) and w_2 (upper limit of the fitness function, values higher than that have a fitness of 1) are free parameters. The second model is the same, but parameter *b* is also a free parameter. Model 3 is a variation of model 2, where α (the duplication rate) and β (the deletion rate) are also free parameters. Models 4 and 5 were used to test the third hypothesis and differ from model 3 by including two linear functions that allow the duplication and deletion rates to vary based on the copy number (where the slopes represent how the rates vary depending on copy number). In model 4, because it is thought that the recombination rate increases when copy number is low (Iida et al., 2019), the slopes *sDup* and *sDel* (duplication and deletion, respectively) were set to a range of -1 to 0. Model 5 was a modification of model 4 but with the slopes set to a range of 0 to 1. An estimation of 0 in the slopes means that the same rate of recombination is applied regardless of the copy number.

To reduce the time taken to do the model fitting and bootstrapping, I developed a custom C++ script that uses the C optimisation library Nlopt replaces the R script used in section 4.3. (The code is available in the GitHub project https://github.com/ivanhc1993/rDNADynamics.git). The optimization algorithm used to compare models was "BOBYQA", a derivative-free algorithm that allows setting parameter constraints similar to how "BFGS-B" works. This method was used because the Nlopt library does not provide a version of "BFGS-B" with parameter constraints. Bootstrapping was done as described in section 4.3.1, with 100 bootstrap iterations per model. *MSE* values for the testing and fitting dataset were used to compare the models. The initial conditions for the models are reported in Tables 4.3.2.1 and 4.3.2.2. The results are shown in figure 4.3.2.1. Overall, all training data sets fit better than the testing data set. The data shows two groups one group includes models 1 and 2, and the other models 3, 4 and 5.

Table 4.3.2.1. Initial parameters for the five model comparisons.

Model	*b* of Γ	w_1	w_2	α	β	*sDel*	*sDup*
Model 1	NA	60	136	NA	NA	NA	NA
Model 2	0.75	60	136	NA	NA	NA	NA
Model 3	0.75	60	136	0.00354	0.00458	NA	NA
Model 4	0.75	60	136	0.00354	0.00458	-0.0000005	-0.0000005
Model 5	0.75	60	136	0.00354	0.00458	0.0000005	0.0000005

Table 4.3.2.2. Lower and upper limits for parameters in the five model comparisons.

Model	*b* of Γ	w_1	w_2	α	β	*sDel*	*sDup*
Model 1	NA	0 - 399	0 - 400	NA	NA	NA	NA
Model 2	0.01 - 0.99	0 - 399	0 - 400	NA	NA	NA	NA
Model 3	0.01 - 0.99	0 - 399	0 - 400	0 - 1	0 - 1	NA	NA
Model 4	0.01 - 0.99	0 - 399	0 - 400	0 - 1	0 - 1	(-1) - 0	(-1) - 0
Model 5	0.01 - 0.99	0 - 399	0 - 400	0 - 1	0 - 1	0 - 1	0 - 1

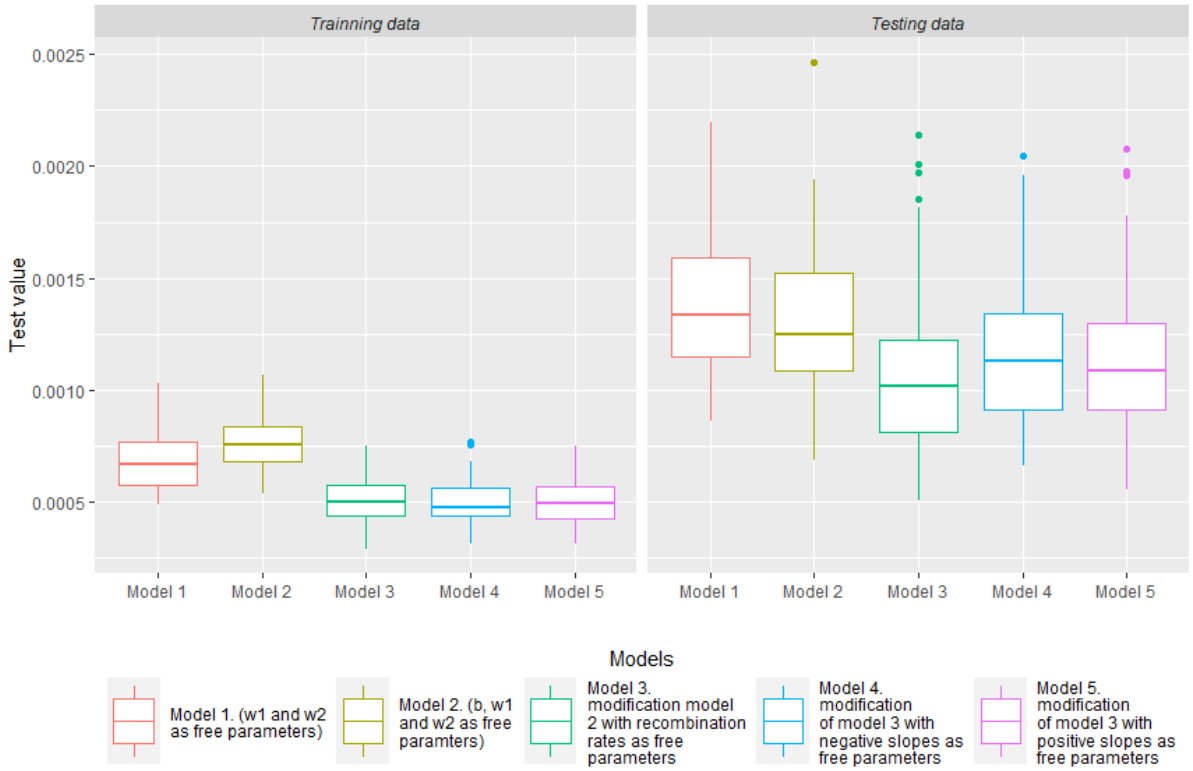


Figure 4.3.2.1 MSE values obtained from bootstrapping analysis for all tested models. MSEs for the training (left) and test (right) data sets from all five models are plotted. The models where recombination rates are free parameters (Models 3, 4 and 5) obtain better fitting with the experimental data, while model 1 and 2 fittings results are similar. *MSEs* are lower for the training data set than for the testing data set. This result can be due either to a higher number of data points or overfitting of the model to the data.

To test if data have equal variance and follow a normal distribution, Levene's and Shapiro-Wilk's tests were performed. Levene's test result for the fitting data set was $F(4) = 4.1019$, $p < 0.01$, and for the testing data set was $F(4) = 0.4025$, $p > 0.05$, which indicates that the fitting data set has equal variance while testing data set not. Shapiro-Wilk's test results are shown in

Table 4.3.2.3. These results show that data does not follow a normal distribution at least in one data set for all models. Because not all the groups in the data sets follow the assumptions of equal variance and normal distribution required to perform a valid parametric test, non-parametric tests were performed to compare the mean *MSE* values between models.

Table 4.3.2.3. Shapiro-Wilk's tests for the five models and the two data sets.

Model	*Data set*	W value	P
Model 1	Training	0.94972	0.0007952***
	Testing	0.95322	0.00136**
Model 2	Training	0.97156	0.3519
	Testing	0.98562	0.02913*
Model 3	Training	0.9908	0.7289
	Testing	0.92937	4.585e-05**
Model 4	Training	0.96707	0.01325*
	Testing	0.96832	0.01647*
Model 5	Training	0.98614	0.3825
	Testing	0.96504	0.009345**

* p < .05, ** p <.01, *** p < .001

A Kruskal-Wallis test was used to assess whether the mean *MSE* significantly differed between the five models. The MSEs for both the training dataset (X^2 (4, N = 100) = 259.01, p = 2.2e-16) and the testing dataset (X^2 (4, N = 100) = 63.194, p = 6.178e-13) had significant differences. Therefore, pairwise comparisons using Wilcox's test were performed to identify where those differences came from (Tables 4.3.2.4 and 4.3.2.5). The results indicate that two groups have significant differences: one contains models 1 and 2, while the other contains models 3, 4 and 5. Significant differences were found in pairwise comparisons between models of both groups for both the training and testing datasets. This result answers the second question and implies there are no differences between models 1 and 2 and the different parameter w_1 does not affect how the model fits to the date to some extent. Which answers the first question. For the case of the other group (Models 3, 4 and 5), significant differences between them were not found, which implies that making recombination rates change in function of the copy number does not improve the model, which answers the third question.

Table 4.3.2.4. Pairwise comparisons between model MSEs for the testing datasets

	Model 1	Model 2	Model 3	Model 4

Model 2	0.584	-	-	-
Model 3	8.5e-10***	4.5e-06***	-	-
Model 4	5.6e-06***	0.012*	0.227	-
Model 5	7.0e-07***	0.002**	1.0	1.0

* p < .05, ** p <.01, *** p < .001

Table 4.3.2.5. Pairwise comparisons between model MSEs for the training datasets

	Model 1	**Model 2**	**Model 3**	**Model 4**
Model 2	3.2e-05***	-	-	-
Model 3	< 2e-16***	< 2e-16***	-	-
Model 4	< 2e-16***	< 2e-16***	1.0	-
Model 5	< 2e-16***	< 2e-16***	1.0	1.0

* p < .05, ** p <.01, *** p < .001

I wanted then to compare the estimation of the parameters to assess if there were important patterns between them and whether their distribution was different from 0 to evaluate if they can omit or not. Parameter estimates per bootstrap analysis iterations were plotted, and their distributions were compared between models (Fig. 4.3.2.2). Almost all parameters obtain distributions that differ from zero, indicating that they have an important contribution to the model. Estimates of w_1 were similar across models. For the parameter w_2 a similar result was obtained, except that a higher estimate was obtained for model 1 compared to all the other models. A higher estimate of parameter *b* was obtained for model 2 compared to the other models where the recombination rates were free parameters. In the models where recombination rates were free to be fitted (models 3, 4 and 5), both duplication and deletion rates were higher than the estimates in section 2.3. The combination of a lower *b* for deletion misalignment value with higher recombination rates could imply that the fits from these models were obtained via more deletion events with shorter misalignment lengths than in models where recombination rates were set. However, more tests should perform to support these claims. The estimates of β (deletion rate) and α (duplication rate) were higher than those obtained from Ganley and Kobayashi (2011), suggesting that deletion events are more frequent than duplication events. However, the opposite was found in the models where both those parameters were free (Fig. 4.3.2.2). This effect was most pronounced when using a positive slope (model 5; Fig. 4.3.2.2).

Another antagonist pattern was also found in the slopes of models 4 and 5. *sDup* estimation tended to have lower values than *sDel*. Data hints that *sDel* could have positive values and

sDup negative. This pattern can imply that duplication events are more frequent at low copy number levels, and deletion events are more frequent at wild-type copy number levels. It is important to highly that both slopes are not particularly large.

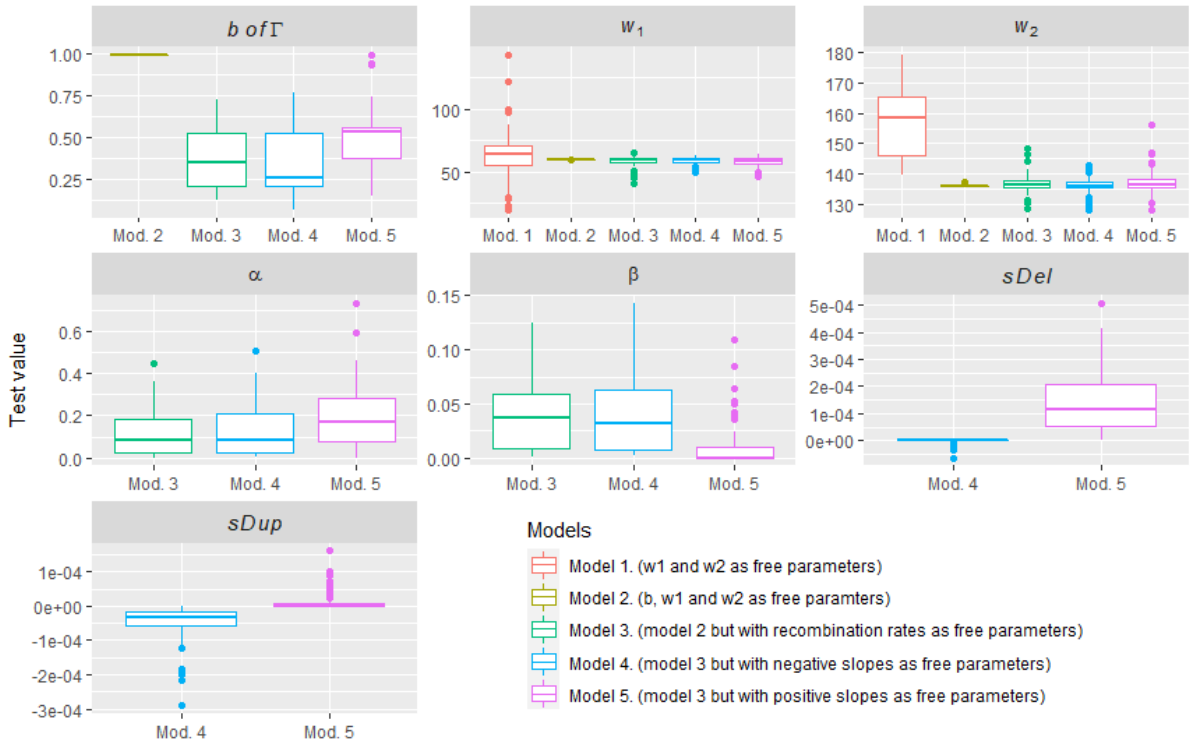


Figure 4.3.2.2 Comparisons of parameter values between the five models following fitting. Boxplots of the distributions of the parameters when they were free to be fitted. Lines in the middle of the box represent the median, lines represent the maximum and minimum values in the estimation, and the dots represent outliers. Overall w_1 (lower limit of the fitness function) estimates are similar in all models, as is w_2 (upper limit of the fitness function) except for model 1, where the estimates are higher. *b* of Γ (maximum deletion misalignment percentage) value estimations were lower when the recombination rates were free parameters. Estimations of the slopes describing how duplication and deletion rates change with copy number, *sDup* and *sDel*, do not deviate much from zero, with which one of the two is estimated closer to zero depending on the model (i.e., what ranges the slopes can take).

The parameter estimations per model were averaged, and simulations using these values as the input parameters were run for each of the five models to compare the resulting distribution. The resulting copy number distribution data were put in bins of 10 and plotted (Figure 4.3.2.3). The *MSE* values for each model with those parameters were 0.000487014 (Model 1), 0.000638977 (Model 2), 0.000392985 (Model 3), 0.000433914 (Model 4) and 0.000388795

(Model 5). Models without free recombination rates produce similar distributions but differ from models where recombination rates are free parameters that fit better. In the first case, models expose poor fitting in the lower part of the copy number distribution compared to the others. The region where models 1 and 2 had a bad fitting contains all the copy numbers that are below the w_2 (upper limit of the fitness function) value.

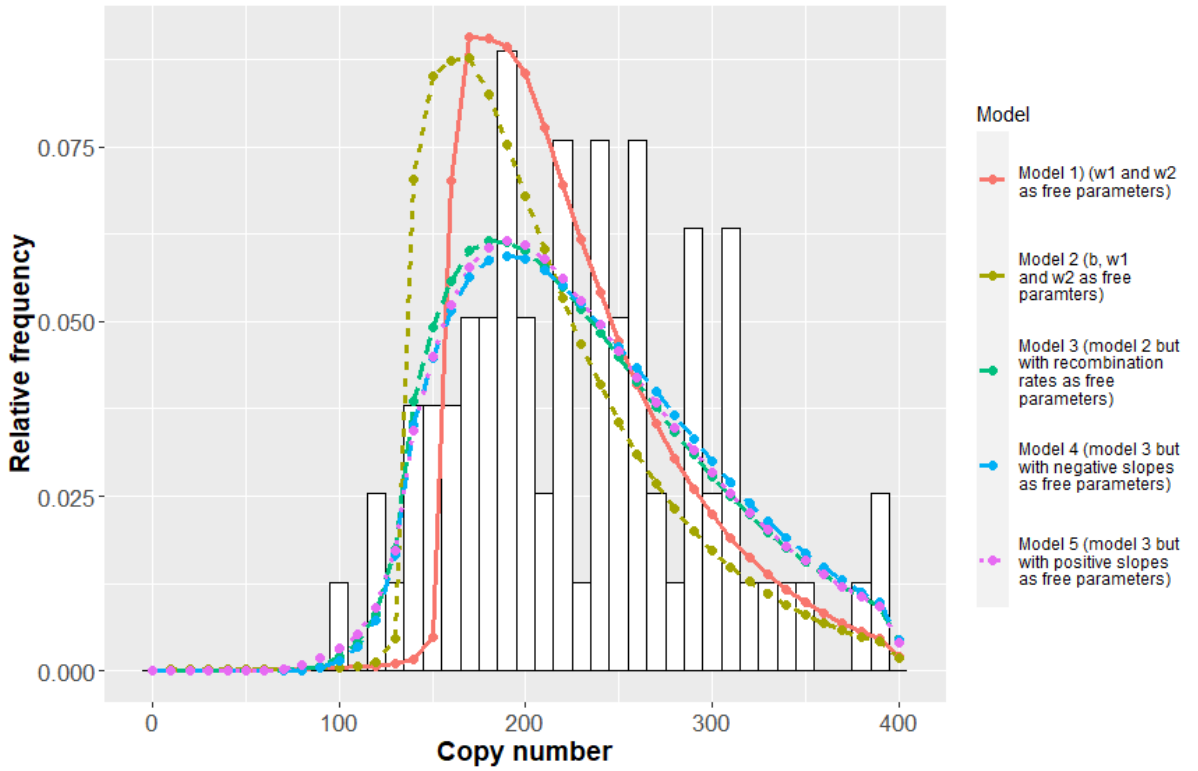


Figure 4.3.2.3 Comparison of the evaluated models' predictions vs the frequency histogram of the observed data. Distribution generated for each model compared with the observed distribution. Each model was run using the averages of the parameters estimated from the fitting to the experimental data. The copy number frequencies from each model were grouped in bins of 10, and the lines in the plot are an approximation of the continuous distribution after the binning. The experimental data are the same as shown in Figure 4.3.1. Models 1 and 2 (*MSE* = 0.000487014, 0.000638977) have more similar predictions, while models 3, 4 and 5 (*MSE* = 0.000392985, 0.000433914, 0.000388795) produce similar predictions.

Chapter 5. CREATION OF PLASMID INCLUDING FLUORESCENT PROTEIN GENES FOR INTEGRATION INTO *HO LOCUS*

The model proposed in this project adapted the fitness function from Lyckegaard & Clark, 1991. However, I wanted to evaluate if experimental data that provides measurements of the effect of selection on different copy numbers could improve the model predictions and fitting. However, the selective effects of different rDNA copy numbers have not been determined. Therefore, I designed competition experiments using strains with different rDNA copy numbers to measure those selection effects. The competition experiments involve inoculating media with different copy numbers and wild-type copy number strains. In yeast, wild-type copy numbers range between 100-250 copies. Because the strains were inoculated in equal proportions, it is expected that their proportion does not change if there is not a selective effect associated with the copy number. Changes in the proportions can indicate that there is a selective effect. The competition experiments require fluorescent protein as markers that allow differentiation of the strains when mixed during competitions. The fluorescent proteins were selected because they can be used in flow cytometers. Those systems can count several cells in liquid culture, provide robust estimates of the proportions of cells per strain competing and discriminate the strains by the associated fluorescent protein signal.

In order to introduce the marker for these competition experiments, the first step is to create a plasmid that has a sequence of the fluorescent protein gene and uses it to transform *S. cerevisiae* strains. The plasmid was designed to contain homologous sequences of the *HO* locus (Fig. 5.1), which allow the insertion of genes in this locus by homologous recombination. The *HO* locus is involved in the interconversion of the mating types and the sexual reproduction of the *S. cerevisiae* cells. I chose the locus because it is reported that it is not involved in cellular growth; therefore, modification in this locus will not significantly affect the growth of the strains. This feature is important to evaluate the fitness effects of the different copies. The plasmid contains *AmpR* resistance gene that can be used in *E. coli* for creating copies of the plasmid (Fig. 5.1). In addition, the plasmid contains the resistance gene *KanMX,* which allows the selection of *S. cerevisiae.*

The strains that will be used in the experiments have different copy numbers. Because the rDNA copies vary from generation due to recombination and their copy number can be restored to wild-type levels, the strains have a mutation in the gene *FOB1* that impair the copy number variation mediated by this protein. Therefore, the strains will retain their initial values during all the experiments and allow measuring the selection effects on the copy numbers. The initial strains are *fob1*-::*His*$^{+}$ strains of 30, 40, and 80 copies and wild-type copy numbers (150 copies), but the experiments can use other strains with other copy numbers. To initiate these experiments, the construction of the plasmids with the fluorescent proteins cloned into them was started.

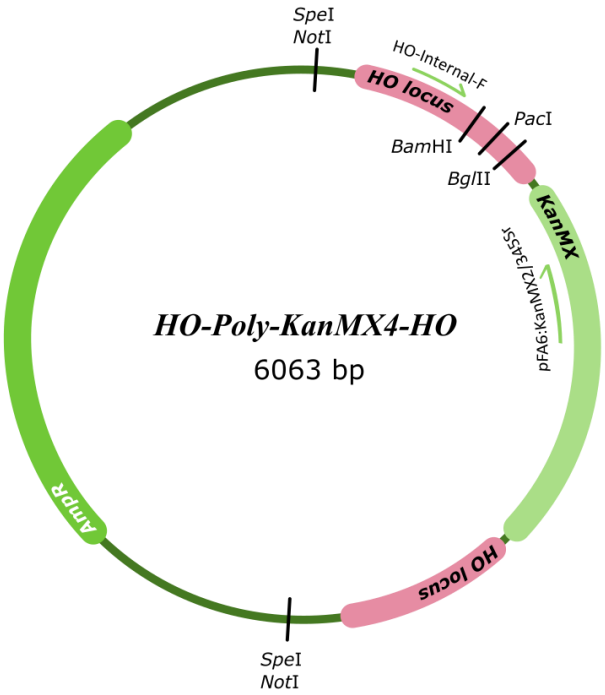


Figure 5.1. *HO-Poly-KanMX4-HO* plasmid map. Plasmid map showing *HO* locus sequences flanking the kanamycin (KanMX) gene. *Spe*I and *Not*I restriction enzyme sites that allow the liberation of the target sequence for transformations *of S. cerevisiae* are indicated, as are *Bam*HI, *Pac*I, and *Bgl*II restriction sites located in one of the HO locus arms. The annealing sites of primers HO-Internal-F and pFA6:KanMX2/345Sr are indicated (not to scale), as is the ampicillin resistance gene (*AmpR*).

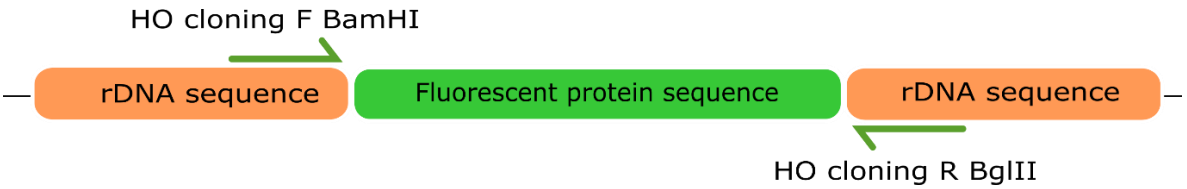
5.1. CLONING FLUORESCENT PROTEINS GENES INTO PLASMIDS

5.1.1 CONSTRUCT CREATION FOR CELLS TRANSFORMATIONS

Synthetic DNA constructs containing fluorescent protein genes *mTagBFP2*, *EGFP*, and *mRuby2* flanked by rDNA gene sequences were already available in our laboratory for a different project. To facilitate the insertion of each fluorescent protein into the *HO* locus, I decided to remove most of the flanking rDNA gene sequences to minimize the chance of the

construct inserting into the rDNA locus rather than the *HO* locus. To remove those sequences, I designed two primers (HO cloning F *Bam*HI and HO cloning R *Bgl*II Table 3.2.2 to anneal close to the edges of the fluorescent protein genes (Fig. 5.1.1 A.). The HO cloning F *Bam*HI includes a *Bam*HI restriction site, and HO cloning R *Bgl*II includes a *Bgl*II. Those restriction sites were designed to produce compatible ends in the fluorescent protein gene constructs that can be used to ligate the constructs to a linearised plasmid with the same compatible ends. PCR was performed to amplify the genes, and the amplification products were checked on a gel (Fig 5.1.1 B.). The amplifications worked properly as intense bands were observed when the PCR products were in the gel. The size of the fragments is about 1.25 kb for the *mTagBFP2* construct and 1.15 kb for *EGFP* and *mRuby2*. That corresponds with the sizes of the observed bands in the gel of the PCR amplificated products, with *mTagBFP2* having a size of ~1250 bp, while *EGFP* and *mRuby2* having ~1159 kb.

A. Schematic of fluorescent proteins gene constructs



B. Gel image of the amplified fluorescent protein genes.

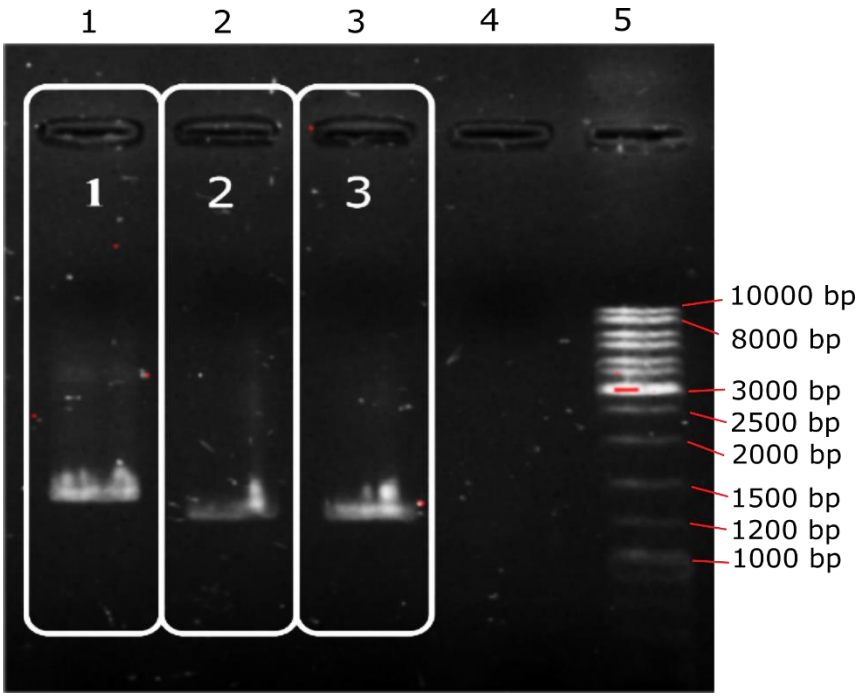


Figure 5.1.1 Schematic of a fluorescent protein gene construct and a gel of the amplified fluorescent protein genes. A. Schematic of a fluorescent protein gene construct. The positions of two primers, HO cloning F *Ba*mHI and HO cloning R *Bgl*II (Table 3.2.2), in the rDNA flanking regions near the edges of the fluorescent protein gene are indicated schematically. The diagram is not to scale. **B.** Gel of the amplified fluorescent proteins genes. Lane 4 corresponds to a water control; Lane 5 is the ladder (GeneRuler DNA Ladder Mix, Thermo Scientific, #SM0331). Lane 1 corresponds to the *mTagBFP2* construct, lane 2 to the *EGFP* construct and lane 3 to the *mRuby2* construct. The size of the fragments is about 1.25 kb for the *mTagBFP2* construct and 1.15 kb for the other two.

5.1.2 CLONING FLUORESCENT PROTEIN GENES INTO *HO-POLY-KANMX4-HO* PLASMID AND TRANSFORMATION OF *E. COLI* CELLS

The *HO-Poly-KanMX4-HO* plasmid (Section 3.2.1, Fig 5.1) and the *EGFP*, *mRuby2*, and *mTagBFP2* PCR-amplified constructs were double digested with *Bam*HI and *Bgl*II (Section 3.7). The plasmid was precipitated as described in Section 3.9 and then digested with *Pac*I (Section 3.7) to increase the transformation efficiency by digesting the small *Bam*HI-*Bgl*II fragment released from the plasmid (Fig. 5.1). The plasmid digests were checked by gel electrophoresis. The gel shows two different band sizes for the linearised plasmid of size 6063 bp (Fig. 5.1) and the undigested plasmid size 8000 bp, which indicates a successfully digested. Then, digested plasmids were dephosphorylated (Section 3.6) to reduce the chances of plasmid re-circularisation. In other to clone the fluorescent protein genes into *E. coli, t*he dephosphorylated plasmid and *mTagBFP2, EGFP* or *mRuby2 constructs* were ligated as described in section 3.8. Different insert-plasmid ratios of DNA amounts (7:1, 8:1, 10:1, 18:1, 20:1 and 50:1) were used.

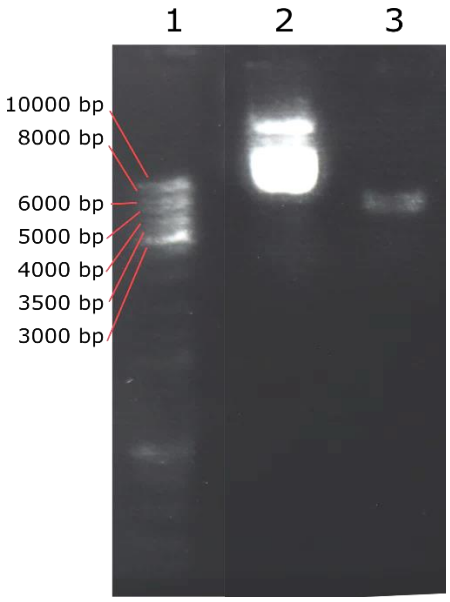


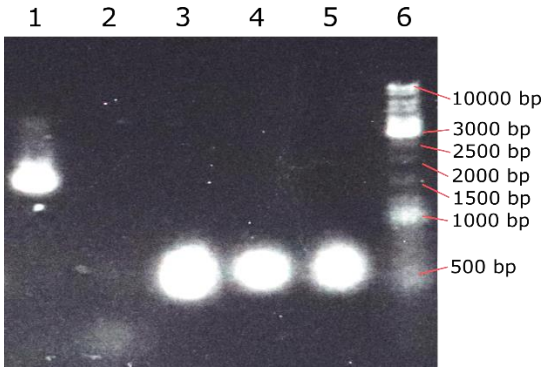
Figure 5.1.2.1. Representative gel of *HO-Poly-KanMX4-HO* plasmid digestion with *Bam*HI and *Bgl*II on the gel to confirm complete digestion. Lane 1 is GeneRuler DNA Ladder Mix (Thermo Scientific, #SM0331). Lane 2 is an undigested plasmid; lane 3 is *Bam*HI and *Bgl*II double-digested plasmid. The linearised plasmid size is 6063 bp.

Plasmids ligated with the three constructs were used to transform *E. coli* competent cells and to clone the plasmid. The transformation protocol described in Section 3.12 was followed, and *E. coli* cells were plated and incubated overnight. When the culture was ready, a part of the colonies obtained from the plate was cultured on LB-amp Agar to create a master plate containing the isolated colonies in a matrix to store the colonies for further tests. The other part was resuspended in water for colony PCR (Section 3.10.1). The PCR reaction was performed as described in Section 3.10. Primers pFA6:KanMX2/345Sr and HO-Internal-F were used as the screening primers as they flank the fluorescent protein gene insertion site (Fig. 5.1). The primers are expected to generate a PCR product of about 500 bp when there is no insert, one of ~1650 bp when *EGFP* and *mRuby2* are present, and one of ~1750 when the *mTagBFP2* construct is present. PCR products were run on a gel (Section 3.11) to confirm which colonies have the construct and which do not. Cloning efficiency was estimated to be two colonies presenting the insert every 200 colonies. The low cloning efficiency made obtaining colonies with plasmids carrying one of the fluorescent protein constructs hard.

To perform a broader screening and because the cloning efficiency was low, colonies were pulled in groups of five in 100 µL of water. About 350 colonies were screened per fluorescent protein construct. Representative gels of colonies containing *mRuby2, EGFP* and *mTagBFP2*

constructs are shown in figure 5.1.2.2. The gel shows colonies with plasmids without the construct have a band of ~500 bp (Fig. 5.1.2.2 A. and B.). Fragments with a size of ~1650 were obtained, which confirm the presence of the *mRuby2* construct (Lane 1, Fig. 5.1.2.2 A.) and *EGFP* constructs (Lane 5, Fig. 5.1.2.2 B.). A fragment with a size of ~1750 confirms the presence of the *mTagBFP2* construct (Lane 3, Fig. 5.1.2.2 B.). Colonies containing plasmids with *mTagBFP2*, *mRuby2* or *EGFP* constructs were used to replicate the plasmids.

A. A gel of a colony PCR screen for colonies transformed with *HO-Poly-KanMX4-HO* plasmid with the *mRuby2*



B. A gel of a colony PCR screen for colonies transformed with *HO-Poly-KanMX4-HO* plasmid with the *mTagBFP2* or *EGFP*

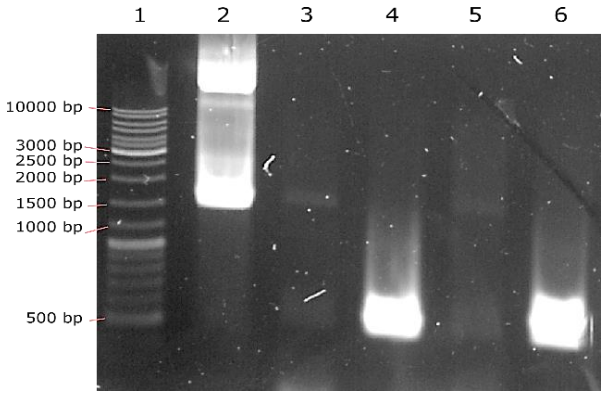


Figure 5.1.2.2 Representative gels of colony PCR screening to determine what colonies contain the HO-Poly-KanMX4-HO plasmid with the constructs cloned into it. A. A gel of a colony PCR screen to determine colonies containing the plasmid with *mRuby2* construct cloned into it. Lane 1 is the PCR product of a colony with a plasmid that contains the *mRuby2* construct. Lane 6 is the GeneRuler DNA Ladder Mix (Thermo Scientific, #SM0331) and Lane 2 is a negative control with water. Lanes 3-5 are colonies that show an amplified fragment of ~500 bp corresponding to an empty plasmid. **B.** Gel of a colony PCR screen to determine colonies containing plasmids with *mTagBFP2* and *EGFP* construct cloned into it. Lane 2 corresponds to a plasmid containing a *mRuby2* construct cloning into it. An amplified fragment of ~1650 bp corresponding to a plasmid containing the *mRuby2* construct (1150 bp) is shown. Lanes 4 and 6 show an amplified fragment of ~500 bp that corresponds to an empty plasmid, and lanes 3 and 5 are plasmids that show amplification of about 1650-1750 bp that corresponds to plasmids, including *mTagBFP2* and *EGFP* construct cloned into them, respectively.

5.1.3 **CONFIRMING THE PRESENCE OF THE CONSTRUCTS IN THE *HO-POLY-KANMX4-HO***

Digestions with restriction enzymes were done to confirm that the right inserts were in the plasmid. A small culture of the transformed colonies, including plasmids with constructs, was required to isolate the plasmid. Therefore, A small piece of those colonies was taken from the master plate with a toothpick and was used to inoculate individual 15 ml falcon tubes. The isolation and purification of the plasmid were done as described in Section 3.4.1. *Spe*I (Section 3.7) is an enzyme that cuts the *HO-Poly-KanMX4-HO* plasmid in two pieces of almost ~3031 bp when the plasmid is empty. When a plasmid contains one of the constructs, the size of one of the fragments is larger. The bigger part will be ~4281 bp for the *mTagBFP2* construct and ~4181 bp for *mRuby2* and *EGFP* constructs. Therefore, the purified *HO-Poly-KanMX4-HO* plasmid was digested with *Spe*I to confirm the presence of the fluorescent protein constructs. Digestions were checked by gel electrophoresis (Fig. 5.1.3.1). Two bands about the size of ~4181 and ~3031 bp confirm the presence of the *mRuby2* construct (Fig. 5.1.3.1). Plasmids containing *mTagBFP2* and *EGFP* could not be obtained.

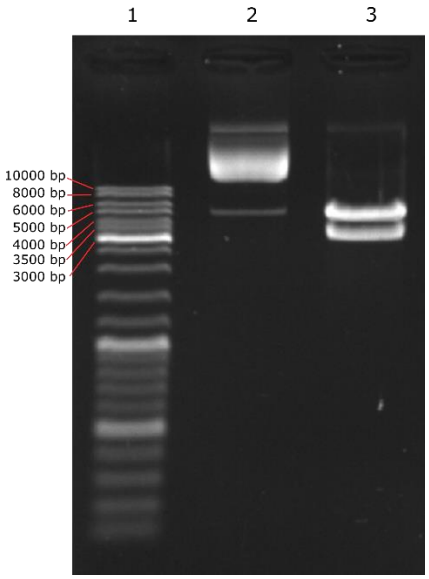


Figure 5.1.3.1. Representative gels of digested purified HO-Poly-KanMX4-HO plasmids to confirm that the fluorescent protein genes constructs were cloned into them. Lane 1 is the GeneRuler DNA Ladder Mix (Thermo Scientific, #SM0331), and lane 2 is the undigested *HO-Poly-KanMX4-HO* plasmid. Lane 3 corresponds to a digested plasmid with *Spe*I to confirm that the *mRuby2* construct is cloned into the plasmid. Two bands of ~4181 bp and ~3031 bp confirm the presence of the *mRuby2* construct.

5.2. INITIAL TEST FOR PREPARING *SACCHAROMYCES CEREVISIAE* TRANSFORMATIONS

I wanted to prepare the *S. cerevisiae* strains that will be used in the competition experiments to determine the fitness effect of the cells with lower copies than the wild-type strain. As described above, those cells have *fob1-* mutations that block the copy number amplification; therefore, their copy number is fixed. The transformation with *mRuby2*, *mTagBFP2* or *EGFP* will allow the discrimination of *Saccharomyces cerevisiae* strains by flow cytometry analysis. Some of the strains that will be transformed are *fob1-*::*His*$^{+}$ strains of 30, 40, and 80 copies and wild-type copy numbers (Section 3.2.1). Initial tests were done to prepare the strains to be transformed with the plasmids that have the fluorescent protein constructs cloned into them. The initial test was refreshing the strains from stock kept at -80 °C. The strains were first refreshed on YNB -His selective plates to ensure the strains are pure (Section 0). Then to have a working stock on plates that can be used for the transformations, individual colonies were taken from YNB -His plate and then were plated onto YPD-agar. from Section 5.1.

Transformation of the fluorescent protein constructs into *S. cerevisiae* involves selection on the antibiotic G418, using a digested *HO-Poly-KanMX4-HO* plasmid. The digested plasmid allows the insertion of the resistance *KanMX* gene to G418 by homologous recombination of the *HO* locus sequences. It has been reported that strains vary in their level of sensitivity (REF). Therefore, I tested different concentrations (200 µg/mL, 250 µg/mL, 400 µg/mL, and 500 µg/mL) of G418 across all the *S. cerevisiae* strains to transform to find a concentration that kills all cells that do not have the resistance gene (Section 3.2.4). After incubation, there was no growth on plates with a G418 concentration higher than 250 µg/mL for any of the dilutions spotted (Fig. 5.2.1.1). On the basis of these results, a G418 concentration of 300 µg/mL is recommended for the *S. cerevisiae* transformations.

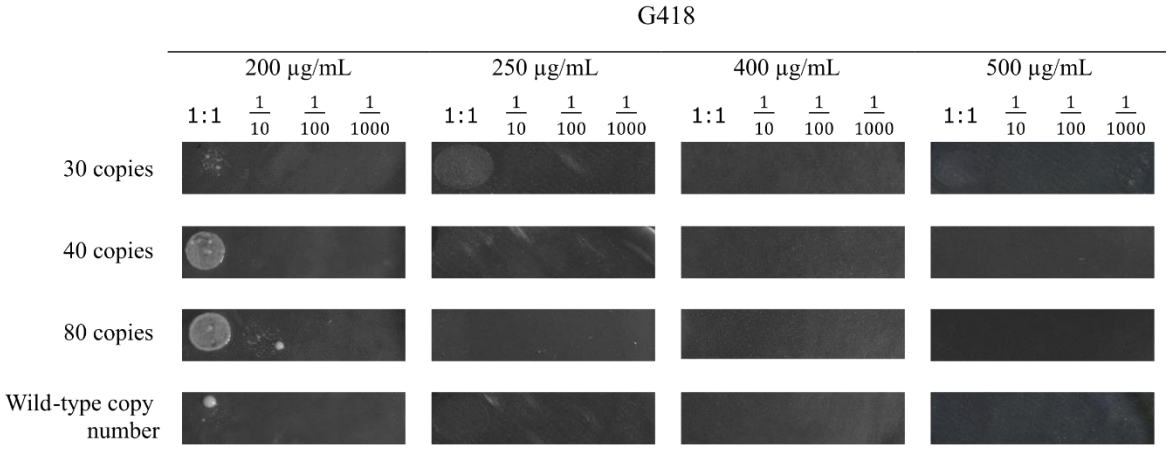


Figure 5.2.1.1 Testing sensitive levels of *S. cerevisiae* strains to G418. The various fob1-::His+ rDNA copy number strains were spotted at different dilutions onto YPD-G418 plates with the concentrations of G418 indicated. Growth was only observed for two strains when the sample was a culture without dilution 1:1 on the 200 µg/mL G418 plate.

Then to evaluate if the concentration of G418 was not too high to kill all the cells even with the resistance gene cloned into them. A testing transformation was performed using the four strains (are *fob1*-::*His*$^{+}$ strains of 30, 40, 80 copies and wild-type copy numbers) intact plasmid and the constructed plasmid, including the *mRuby2* construct. The transformation was done as described in Section 3.12. The cells were grown for 3 days, and then their growth was checked. The cells with the resistance gene could grow as expected. With this initial test done, all the initial tests are done for the strains. Future work will involve the construction of another plasmid that contains either the *mTagBFP2* or *EGFP* constructs. Then the cells can be transformed, and the competition experiment described above in this chapter can be performed.

Chapter 6. DISCUSSION, CONCLUSIONS AND FUTURE DIRECTIONS

6.1. DISCUSSION

In this research project, I developed a model that generates copy number distributions for haploid *Saccharomyces cerevisiae* populations. This model was used to fit selection and non-reciprocal recombination parameters to a copy number distribution estimated from a wild-type copy number *S. cerevisiae* strain. Five models with different parameters were set up to make comparisons that allow assessment of the effects. Models 1 and 2 had fixed recombination, while models 3, 4 and 5 had free recombination rates. Models 4 and 5 had a slope that allowed the recombination rate to change as a function of the copy number. The major finding from this study was that allowing the duplication and deletion recombination rates to vary depending on the copy number provided no clear improvement to the resulting distributions, compared to when the recombination rates were the same regardless of copy number.

Duplication misalignment fits a double uniform probability, suggesting rDNA may have two mechanisms to increase copy number.

A double uniform probability distribution was the best-fitted function to describe the rDNA misalignment length frequencies of duplication events reported in Ganley & Kobayashi (2011). Therefore this probability distribution function was used in the model. The double uniform probability is compatible with the idea that two duplication mechanisms operate in *S. cerevisiae* to alter the copy number, where both mechanisms increase copy number but act at different scales. One produces small misalignments, thus producing small duplications, at a relatively higher frequency, while the other produces a large range of misalignments, thus producing a large range of duplication sizes, but at a relatively lower frequency. One of these mechanisms may represent the duplication events produced by reciprocal recombination, while the other may represent the duplications that result from non-reciprocal recombination. Non-reciprocal recombination is reported to be the predominant recombination under unequal sister chromatid exchange (Gangloff et al., 1996), therefore the mechanism that produces small copy changes but at a higher frequency may be this non-reciprocal recombination. Future work should be done to determine if the two mechanisms coexist and if this confirms what the frequencies are.

The mean rDNA copy number determined by ddPCR is higher than previous estimates

Estimating the copy number distribution was crucial for fitting the model to experimental data. The mean copy number for the *S. cerevisiae* population I assayed using ddPCR was 231 *SD*= 66, with the most frequent copy number being in the range of 185 to 195 copies. This mean value is slightly higher than what is reported in previous estimates of copy number (around 100 - 250 copies) for laboratory strains related to the strain I used (James et al., 2009; Kobayashi, 2006; Kobayashi et al., 1998; Petes, 1979; Schweizer et al., 1969; West et al., 2014). As previously reported, laboratory strains' copy number are substantially higher than the copy numbers of wild *S. cerevisiae* strains (Sharma, 2021). Therefore, differences in the mean copy number can be partially attributed to the differences in this strain's basal copy number compared with the wild strains, but further analyses are required to support this claim. Despite the high mean copy number, the estimations of the copies in the 79 colonies are all within the copy number range of 60 – 511 previously estimated using ddPCR or whole genome sequencing (James et al., 2009; Sharma, 2021; West et al., 2014).

The higher copy numbers I found here could simply be a consequence of sampling. I used 79 samples, which may be insufficient to obtain a fully representative picture of the population. The lack of colonies having rDNA copy numbers lower than 100, which are expected, although at lower frequencies, might indicate insufficient sampling. Another explanation for the higher copy number is that ddPCR overestimates the copy number. This explanation is supported by a previous study that found rDNA copy number estimates in *Aspergillus fumigatus* were significantly higher when ddPCR was used compared to using qPCR (Alanio et al., 2016).

The estimates of the parameter w_1 (the lower limit of the selection function) were around 60 copies. This parameter represents the copy number beneath which genotypes are not viable (fitness = 0). However, evidence from several studies has shown that *S. cerevisiae* strains with lower copy numbers than this are viable and can grow in different media (French et al., 2003; Ide et al., 2010; Iida & Kobayashi, 2019; Kobayashi & Ganley, 2005; Quintana, 2016; Sharma, 2021; Takeuchi et al., 2003). Therefore, this evidence suggests a lower w_1 value. Because the w_1 value in this research was estimated from the estimated copy number distribution, these higher-than-expected values of w_1 can be a consequence of the potential copy number overestimation by ddPCR or insufficient sampling to capture colonies with lower copies.

Estimated recombination rates with higher values than previously estimated may indicate the model requires improvements.

Models with recombination rates as free parameters (Models 3, 4, 5) estimated higher recombination rates than those measured in Ganley and Kobayashi (2011), which were the values used in models 1 and 2. However, these high model-generated recombination rates seem to coincide with previous estimations for deletions (Szostak & Wu, 1980). These models had better fits than models 1 and 2, suggesting that models 1 and 2 are missing an important component to explain copy number distribution. Because the better fitting is associated in part with better fitting to the lower parts of the distribution, the missing component might be a mechanism that has a strong effect when individuals have low copy numbers, such as selection. Experimental measurement of the effect of copy number on fitness could provide evidence to support this prediction.

Recombination rates that vary depending on copy number do not improve model fitting.

Previous research suggests unequal recombination rates increase when the rDNA copy number is low (Iida & Kobayashi, 2019). To see if this feature would improve the fitting of the model, I added a parameter that modifies the non-reciprocal recombination as a linear function of the copy number, as described by the slope of this function, and ran two versions of the model where the slopes are free parameters (Models 4 and 5; Section 4.3.2). However, allowing recombination rates to vary (i.e., with a slope other than zero) in those models does not improve the model fitting, with similar mean *MSEs* obtained compared to model 3 (model without slopes but where the recombination rates were free). Hence, at least in how the model was set up, recombination rate dependence on copy number is not necessary to explain copy number distribution.

This result seems to contradict the model proposed by Iida & Kobayashi, 2019, where in a low copy number context, the UAF protein represses Sir2, thus inducing increased rDNA unequal recombination. One possible explanation for these contradictory results is that the simple linear relationship used in this work is not correct and that if a function that better describes the relationship between copy number and recombination rate is used, perhaps a non-linear function that is dependent on the free concentration of UAF, this would provide better fitting. Another possible explanation for these contradictory results is the higher recombination rates produced in my models compared to those inferred from published experimental data. Models 4 and 5 have the recombination rates as the intercepts of the linear functions that describe the variation of the recombination. Therefore, the high values of these parameters could affect the modelling of relationship between copy number and recombination rates. In addition, the slopes

sDup and *sDel* having the same parameter space, Model 4 just negative values from -1 to 0, Model 5 with positive values from 0 to 1, could reduce the model's performance. The estimations of *sDup* values suggest that negative values are more probable because when the model was constrained to be in a positive range, the *sDup*'s estimates were closer to zero than when constrained to be in a negative range. The opposite trend was found for the parameter *sDel*. Therefore, a model that includes a broader parameter space allowing negative and positive values, namely -1 to 1, for both parameters may improve the model fitting.

6.2. FUTURE DIRECTIONS

6.2.1 **Assess to what extent there is an overestimation by ddPCR of the copy numbers**

The result of the copy number distribution suggests data may be biased to a higher copy number due to an overestimation of the ddPCR or insufficient sampling. To discern whether the results obtained are due to an overestimation or to the detection of individuals with a higher copy number, it would be worth doing more copy number measurements over more colonies by ddPCR while comparing with other techniques such as PFGE and qPCR. The comparison with other techniques would make it possible to adjust the overestimations if they are present and better estimate the parameters. However, all the analyses will also benefit from increasing the number of samples used to estimate the copy number distribution.

6.2.2 **Obtain experimental data that measures the recombination rates and misalignment probability functions in strains with different copy number**

It is important to mention that the technique used to measure the length of duplication misalignment length was PFGE, which does not have a high resolution when the duplication sizes are large. Therefore, the frequencies of large duplication values might be underestimated. Obtaining additional experimental data and performing computational experiments could confirm if misalignments resulting in large duplications are underestimated. To obtain additional experimental data, measurements of misalignment lengths and their frequencies with higher resolution techniques could provide more robust data. For the computational experiments, exploring different configurations of the model that allow the misalignment duplication function parameters to be free-fitted could bring some insight into the effects of different parameter values on the calculated copy number distribution.

The percentage of misalignment was used as the unit of change to make possible the comparison between different individuals with different copy numbers. This was done because there are no available data measuring misalignments from strains with copy numbers different to wild-type. For future experiments, it would be interesting to see what effect changing the way this misalignment distribution scales with copy number has in the model. It would also be interesting to determine if there is variation in the misalignment distributions using strains with different homeostatic copy numbers or that are set at a specific copy number due to deletion of *FOB1*. This last part could be difficult because *FOB1* mutation will change the recombinational activity. However, measurements of recombination rates in strains with different copy numbers is important to assesses if there different recombination rates are associated with different copy numbers.

6.2.3 **Fitting the deletion misalignment distribution probability function could improve the performance of the model**

The distribution function used to model the deletion misalignment was a uniform distribution. Other models have used uniform distributions for modelling recombination processes that produced deletion (Lyckegaard & Clark, 1991; Zhang et al., 2008) to reduce the complexity of their models. In this work, I used uniform because an inspection of the data shows that a uniform distribution can approximate it, and this distribution reduced the model's complexity. However, it would be worth fitting the rDNA misalignment length frequencies data to explore if another distribution can better fit the observed data.

6.2.1 **Competition experiments to evaluate the selective effect of low numbers of rDNA copies.**

I constructed a plasmid containing the mRuby2 construct in this project, but obtaining a plasmid for the other two fluorescent proteins was not achieved. Constructing at least one other plasmid that includes a different fluorescent protein will be required to perform the competition experiments. Then, both plasmids can be used to transform the strains and measure the selective effect of copy number using competition experiments described in Section 5.1. These competition experiments will enable estimates of the selective parameters, which will then enable testing to determine whether the fitness function used in the model can be improved to better fit the other experimental data.

6.3. CONCLUSIONS

This research sought to assess the effects of selection, probability of recombination and the length of misalignment on rDNA copy number distribution in a haploid *Saccharomyces cerevisiae* population. To achieve this, I developed a discrete generation model to calculate the copy number distribution at equilibrium for a haploid *S. cerevisiae* population. The main components of the model are selection and unequal non-reciprocal recombination, which have been reported as the main components of copy number variation at the population level. I also experimentally measured the rDNA copy number distribution of a haploid *S. cerevisiae* strain with wild-type rDNA copy number for use in the model. A high mean of rDNA copies was found compared with previous publications that may result from insufficient sampling, different strain's basal copy numbers or overestimated copy numbers by ddPCR. More tests are required to assess why this high mean was obtained. Comparisons of different models fitted to the measured rDNA copy number distribution were then used to evaluate the effect of unequal chromatid recombination and selection.

The different model comparisons showed a reasonable fit of the models to the experimental data. However, it is possible that improving the function that describes the fitness effects of copy number by using experimentally generated data might improve how the model describes the copy number distribution, particularly at the lower end of the distribution. To help progress this goal, I successfully constructed a plasmid with a fluorescent protein marker that can be used to transform different strains with different copy numbers. These marked strains could then be used in competition experiments to assess the selective effects of different copy numbers.

The model comparisons showed no improvement in the fitting to experimental data when including recombination rates dependent on copy number, despite recent studies demonstrating that recombination rates change in this way. Future tests to help resolve why we saw no improvement could include assessing different probability distributions for the function that describes the misalignment percentages of deletion events. In particular, it might be that a non-linear function can better describe the variation of recombination rates based on copy number.

APPENDIX A

SCRIPT USED TO PERFORM SIMULATIONS

Appendix A. Script 1 Fitting distributions

```
duplication_data <- read_excel(excelFilePath, sheet = "Duplication")

duplication_data

hist(duplication_data$Value, probability = TRUE,

    main = "Deletions",

    xlab = "Deletion Percentage",

    breaks = 100)

box()

grid()

lines(density(duplication_data$Value), col = "darkorange")

qqplot(x = quantile.density(density(duplication_data$Value),
ppoints(duplication_data$Value)),

      y = duplication_data$Value,

      main = "QQ-Plot: Cocaine Potency, KDE",

      xlab = "Theoretical Quantiles, Kernel Density Estimate",

      ylab = "Sample Quantiles, Cocaine Price")

abline(a = 0, b = 1, col = "dodgerblue", lwd = 2)

grid()

dev.off()

plotdist(duplication_data$Value, histo = TRUE, demp = TRUE)

descdist(duplication_data$Value, boot = 1000)

fw <- fitdist(duplication_data$Value, "weibull")

fg <- fitdist(duplication_data$Value, "gamma")
```

```
fu <- fitdist(duplication_data$Value, "unif")

fn <- fitdist(duplication_data$Value, "lnorm")

fe <- fitdist(duplication_data$Value, "exp")

fb <- fitdist(duplication_data$Value, "beta")

flogistic <- fitdist(duplication_data$Value, "logis")

plotdist(duplication_data$Value, "weibull", para = list(shape=0.82776712, scale=0.04664586))

par(mfrow = c(2, 2))

plot.legend <- c("w", "logN")

denscomp(list(fw, fn), legendtext = plot.legend)

qqcomp(list(fw, fn), legendtext = plot.legend)

cdfcomp(list(fw, fn), legendtext = plot.legend)

ppcomp(list(fw, fn), legendtext = plot.legend)

distributionsNames = c("Weibull", "Gamma", "Uniform", "Log normal", "Exponential", "Beta", "Logistic")

distributionsLogLike <- c(fw$loglik, fg$loglik, fu$loglik, fn$loglik, fe$loglik, fb$loglik, flogistic$loglik)

distResult <- gofstat(list(fw, fg, fu, fn, fe, fb, flogistic), fitnames = c("Weibull", "Gamma", "Uniform", "Log normal", "Exponential", "Beta", "Logistic"))

distResult
```

The annotated version is available in in the GitHub project https://github.com/ivanhc1993/rDNADynamics.git

Appendix A. Script 2 Fitting custom distribution

```
duplication_data <- read_excel(excelFilePath, sheet = "Duplication")

xms <- seq(min(duplication_data$Value), max(duplication_data$Value), by=0.001)

n0E <- function(xm, data){
    count = 0
      for (value in data){
        if(value <= xm){
          count = count + 1
        }
      }
    return(count)
}

logLi2Uniform <- function(xm, data){
    n0 = n0E(xm, data) # n0 number of values < than xm
    return(n0 * log(n0 / (34 * xm)) + ((34 - n0) * log( ( (34 - n0) / ((( max(data) - xm ) * 34
))))))
}


maxLogFunction <- function(vector){
    temp <- data.frame()
    for (i in 1:length(vector)){
```

```
        temp <- rbind(temp, data.frame(vector[i], logLi2Uniform(vector[i], duplication_data$Value)))
    }
    return(temp)
}


dupLog <- maxLogFunction(xms)
dupLog
dupss <- dupLog[!sapply(dupLog, is.nan)]
maxLog <- max(dupss)
maxLog
maxX <- max(duplication_data$Value)
maxX
logResult <- logLi2Uniform(0.03, duplication_data$Value)
logResult
aicResult <- (2 * 4) - (2 * (maxLog))
aicResult
```

The annotated version is available in in the GitHub project https://github.com/ivanhc1993/rDNADynamics.git

APPENDIX B

COMPLEMENTARY RESULTS CHAPTER 4

Table 1. Measurements of the concentrations by spectrophotometer for 79 gDNA samples.

Sample	Conc.	Units
1	150.15	ng/ul
2	142.85	ng/ul
3	350.20	ng/ul
4	286.65	ng/ul
5	119.10	ng/ul
6	114.30	ng/ul
7	170.75	ng/ul
8	129.40	ng/ul
9	55.200	ng/ul
10	98.900	ng/ul
11	184.05	ng/ul
12	106.40	ng/ul
13	80.450	ng/ul
14	116.20	ng/ul
15	93.250	ng/ul
16	55.600	ng/ul
17	130.35	ng/ul
18	156.90	ng/ul
19	104.70	ng/ul
20	111.15	ng/ul
21	84.900	ng/ul
22	194.30	ng/ul
23	126.75	ng/ul
24	97.850	ng/ul
25	163.85	ng/ul
26	146.65	ng/ul
27	284.80	ng/ul
28	326.00	ng/ul
29	239.60	ng/ul
30	319.00	ng/ul
31	214.30	ng/ul
32	328.00	ng/ul
33	85.600	ng/ul
34	110.75	ng/ul
35	184.10	ng/ul

36	146.45	ng/ul
37	189.95	ng/ul
38	195.50	ng/ul
39	188.90	ng/ul
40	103.60	ng/ul
41	91.850	ng/ul
42	83.700	ng/ul
43	184.20	ng/ul
44	140.40	ng/ul
45	132.15	ng/ul
46	103.05	ng/ul
47	150.90	ng/ul
48	137.10	ng/ul
49	119.00	ng/ul
50	155.75	ng/ul
51	148.65	ng/ul
52	201.30	ng/ul
53	212.30	ng/ul
54	298.30	ng/ul
55	11.650	ng/ul
56	313.30	ng/ul
57	104.35	ng/ul
58	156.30	ng/ul
59	198.30	ng/ul
60	91.850	ng/ul
61	83.700	ng/ul
62	184.20	ng/ul
63	140.40	ng/ul
64	132.15	ng/ul
65	103.05	ng/ul
66	150.90	ng/ul
67	137.10	ng/ul
68	119.00	ng/ul
69	155.75	ng/ul
70	148.65	ng/ul
71	201.30	ng/ul
72	212.30	ng/ul
73	298.30	ng/ul
74	191.85	ng/ul
75	254.30	ng/ul
76	161.90	ng/ul
77	186.65	ng/ul
78	199.25	ng/ul
79	255.05	ng/ul

REFERENCES

Alanio, A., Sturny-Leclère, A., Benabou, M., Guigue, N., & Bretagne, S. (2016). Variation in copy number of the 28S rDNA of Aspergillus fumigatus measured by droplet digital PCR and analog quantitative real-time PCR. *Journal of Microbiological Methods*, *127*, 160–163. https://doi.org/10.1016/j.mimet.2016.06.015

Brambati, A., Colosio, A., Zardoni, L., Galanti, L., & Liberi, G. (2015). Replication and transcription on a collision course: Eukaryotic regulation mechanisms and implications for DNA stability. *Frontiers in Genetics*, *6*:166. https://doi.org/10.3389/fgene.2015.00166

Brewer, B. J., Lockshon, D., & Fangman, W. L. (1992). The arrest of replication forks in the rDNA of yeast occurs independently of transcription. *Cell*, *71*(2), 267–276. https://doi.org/10.1016/0092-8674(92)90355-G

Britten, R. J., & Kohne, D. E. (1968). Repeated Sequences in DNA. *Science*, *161*(3841), 529–540. https://doi.org/10.1126/science.161.3841.529

Brown, D., & Wensink, C. (1972). A Comparison of the Ribosomal DNA's of Xenopus Zaevis and Xenopus mulleri: The Evolution of Tandem Genes. *Journal of Molecular Biology*, *63*(1), 57–64.

Chestkov, I. V., Jestkova, E. M., Ershova, E. S., Golimbet, V. E., Lezheiko, T. V., Kolesina, N. Y., Porokhovnik, L. N., Lyapunova, N. A., Izhevskaya, V. L., Kutsev, S. I., Veiko, N. N., & Kostyuk, S. V. (2018). Abundance of ribosomal RNA gene copies in the genomes of schizophrenia patients. *Schizophrenia Research*, *197*, 305–314. https://doi.org/10.1016/j.schres.2018.01.001

Defossez, P.-A., Prusty, R., Kaeberlein, M., Lin, S.-J., Ferrigno, P., Silver, P. A., Keil, R. L., & Guarente, L. (1999). Elimination of Replication Block Protein Fob1 Extends the Life Span of Yeast Mother Cells. *Molecular Cell*, *3*(4), 447–455. https://doi.org/10.1016/S1097-2765(00)80472-4

Drumonde-Neves, J., Vieira, E., Lima, M. T., Araujo, I., Casal, M., & Schuller, D. (2013). An easy, quick and cheap high-throughput method for yeast DNA extraction from microwell plates. *Journal of Microbiological Methods*, *93*(3), 206–208. https://doi.org/10.1016/j.mimet.2013.03.016

Eickbush, T. H., & Eickbush, D. G. (2007). Finely Orchestrated Movements: Evolution of the Ribosomal RNA Genes. *Genetics*, *175*(2), 477–485. https://doi.org/10.1534/genetics.107.071399

French, S. L., Osheim, Y. N., Cioci, F., Nomura, M., & Beyer, A. L. (2003). In Exponentially Growing Saccharomyces cerevisiae Cells, rRNA Synthesis Is Determined by the Summed RNA Polymerase I Loading Rate Rather than by the Number of Active Genes. *Molecular and Cellular Biology*, *23*(5), 1558–1568. https://doi.org/10.1128/MCB.23.5.1558-1568.2003

Gangloff, S., Zou, H., & Rothstein, R. (1996). Gene conversion plays the major role in controlling the stability of large tandem repeats in yeast. *The EMBO Journal*, *15*(7), 1715–1725. https://doi.org/10.1002/j.1460-2075.1996.tb00517.x

Ganley, A. R. D., Ide, S., Saka, K., & Kobayashi, T. (2009). The Effect of Replication Initiation on Gene Amplification in the rDNA and Its Relationship to Aging. *Molecular Cell*, *35*(5), 683–693. https://doi.org/10.1016/j.molcel.2009.07.012

Ganley, A. R. D., & Kobayashi, T. (2011). Monitoring the Rate and Dynamics of Concerted Evolution in the Ribosomal DNA Repeats of Saccharomyces cerevisiae Using

Experimental Evolution. *Molecular Biology and Evolution*, *28*(10), 2883–2891. https://doi.org/10.1093/molbev/msr117

Ganley, A. R. D., & Kobayashi, T. (2014). Ribosomal DNA and cellular senescence: New evidence supporting the connection between rDNA and aging. *FEMS Yeast Research*, *14*(1), 49–59. https://doi.org/10.1111/1567-1364.12133

Hosgood, H. D., Hu, W., Rothman, N., Klugman, M., Weinstein, S. J., Virtamo, J. R., Albanes, D., Cawthon, R., & Lan, Q. (2019). Variation in ribosomal DNA copy number is associated with lung cancer risk in a prospective cohort study. *Carcinogenesis*, *40*(8), 975–978. https://doi.org/10.1093/carcin/bgz052

Ide, S., Miyazaki, T., Maki, H., & Kobayashi, T. (2010). Abundance of Ribosomal RNA Gene Copies Maintains Genome Integrity. *Science*, *327*(5966), 693–696. https://doi.org/10.1126/science.1179044

Iida, T., & Kobayashi, T. (2019). RNA Polymerase I Activators Count and Adjust Ribosomal RNA Gene Copy Number. *Molecular Cell*, *73*(4), 645-654.e13. https://doi.org/10.1016/j.molcel.2018.11.029

Iida, T., Kobayashi, T., & Link to external site, this link will open in a new window. (2019). How do cells count multi-copy genes?: "Musical Chair" model for preserving the number of rDNA copies. *Current Genetics; Berlin*, *65*(4), 883–885. http://dx.doi.org.ezproxy.auckland.ac.nz/10.1007/s00294-019-00956-0

James, S. A., O'Kelly, M. J. T., Carter, D. M., Davey, R. P., van Oudenaarden, A., & Roberts, I. N. (2009). Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces cerevisiae* as revealed by whole-genome resequencing. *Genome Research*, *19*(4), 626–635. https://doi.org/10.1101/gr.084517.108

Kaeberlein, M., McVey, M., & Guarente, L. (1999). The SIR2/3/4 complex and SIR2 alone promote longevity in Saccharomyces cerevisiae by two different mechanisms. *Genes & Development*, *13*(19), 2570–2580.

Kim, N., & Jinks-Robertson, S. (2012). Transcription as a source of genome instability. *Nature Reviews Genetics*, *13*(3), 204–214. https://doi.org/10.1038/nrg3152

Kobayashi, I. (1992). Mechanisms for gene conversion and homologous recombination: The double-strand break repair model and the successive half crossing-over model. *Advances in Biophysics*, *28*, 81–133.

Kobayashi, T. (2006). Strategies to maintain the stability of the ribosomal RNA gene repeats –Collaboration of recombination, cohesion, and condensation. *Genes & Genetic Systems*, *81*(3), 155–161. https://doi.org/10.1266/ggs.81.155

Kobayashi, T. (2011). Regulation of ribosomal RNA gene copy number and its role in modulating genome integrity and evolutionary adaptability in yeast. *Cellular and Molecular Life Sciences*, *68*(8), 1395–1403. https://doi.org/10.1007/s00018-010-0613-2

Kobayashi, T. (2014). Ribosomal RNA gene repeats, their stability and cellular senescence. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, *90*(4), 119–129. https://doi.org/10.2183/pjab.90.119

Kobayashi, T., & Ganley, A. R. D. (2005). Recombination regulation by transcription-induced cohesin dissociation in rDNA repeats. *Science (New York, N.Y.)*, *309*(5740), 1581–1584. https://doi.org/10.1126/science.1116102

Kobayashi, T., Heck, D. J., Nomura, M., & Horiuchi, T. (1998). Expansion and contraction of ribosomal DNA repeats in Saccharomyces cerevisiae: Requirement of replication fork blocking (Fob1) protein and the role of RNA polymerase I. *Genes & Development*, *12*(24), 3821–3830.

Lindstrom, D. L., Leverich, C. K., Henderson, K. A., & Gottschling, D. E. (2011). Replicative Age Induces Mitotic Recombination in the Ribosomal RNA Gene Cluster of Saccharomyces cerevisiae. *PLOS Genetics*, *7*(3), e1002015. https://doi.org/10.1371/journal.pgen.1002015

Lofgren, L. A., Uehling, J. K., Branco, S., Bruns, T. D., Martin, F., & Kennedy, P. G. (2019). Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles. *Molecular Ecology*, *28*(4), 721–730. https://doi.org/10.1111/mec.14995

Lu, K. L., Nelson, J. O., Watase, G. J., Warsinger-Pepe, N., & Yamashita, Y. M. (2018). Transgenerational dynamics of rDNA copy number in Drosophila male germline stem cells. *ELife*, *7*, e32421. https://doi.org/10.7554/eLife.32421

Lyckegaard, E., & Clark, A. (1991). Evolution of ribosomal RNA gene copy number on the sex chromosomes of Drosophila melanogaster. *Molecular Biology and Evolution*, *8*(4), 458–474. https://doi.org/10.1093/oxfordjournals.molbev.a040664

Malinovskaya, E. M., Ershova, E. S., Golimbet, V. E., Porokhovnik, L. N., Lyapunova, N. A., Kutsev, S. I., Veiko, N. N., & Kostyuk, S. V. (2018). Copy Number of Human Ribosomal Genes With Aging: Unchanged Mean, but Narrowed Range and Decreased Variance in Elderly Group. *Frontiers in Genetics*, *9*, 306. https://doi.org/10.3389/fgene.2018.00306

Mansisidor, A., Molinar, T., Srivastava, P., Dartis, D. D., Pino Delgado, A., Blitzblau, H. G., Klein, H., & Hochwagen, A. (2018). Genomic Copy-Number Loss Is Rescued by Self-Limiting Production of DNA Circles. *Molecular Cell*, *72*(3), 583-593.e4. https://doi.org/10.1016/j.molcel.2018.08.036

McTaggart, S. J., Dudycha, J. L., Omilian, A., & Crease, T. J. (2007). Rates of recombination in the ribosomal DNA of apomictically propagated Daphnia obtusa lines. *Genetics*, *175*(1), 311–320. https://doi.org/10.1534/genetics.105.050229

Naidoo, K., Steenkamp, E. T., Coetzee, M. P. A., Wingfield, M. J., & Wingfield, B. D. (2013). Concerted Evolution in the Ribosomal RNA Cistron. *PLoS ONE*, *8*(3), e59355–e59355. https://doi.org/10.1371/journal.pone.0059355

Petes, T. D. (1979). Yeast ribosomal DNA genes are located on chromosome XII. *Proceedings of the National Academy of Sciences*, *76*(1), 410–414. https://doi.org/10.1073/pnas.76.1.410

Porokhovnik, L. N., & Lyapunova, N. A. (2019). Dosage effects of human ribosomal genes (rDNA) in health and disease. *Chromosome Research; Dordrecht*, *27*(1–2), 5–17. http://dx.doi.org.ezproxy.auckland.ac.nz/10.1007/s10577-018-9587-y

Prakash, L., & Taillon-Miller, P. (1981). Effects of the rad52 gene on sister chromatid recombination in Saccharomyces cerevisiae. *Current Genetics*, *3*(3), 247–250. https://doi.org/10.1007/BF00429828

Prokopowich, C. D., Gregory, T. R., & Crease, T. J. (2003). The correlation between rDNA copy number and genome size in eukaryotes. *Genome; Ottawa*, *46*(1), 48–50.

Quintana, D. (2016). *Role of the ribosomal DNA repeats on chromosome segregation of Saccharomyces cerevisiae* [PhD thesis]. Massey University.

Rosato, M., Álvarez, I., Feliner, G. N., & Rosselló, J. A. (2017). High and uneven levels of 45S rDNA site-number variation across wild populations of a diploid plant genus (Anacyclus, Asteraceae). *PLOS ONE*, *12*(10), e0187131. https://doi.org/10.1371/journal.pone.0187131

Saka, K., Ide, S., Ganley, A. R. D., & Kobayashi, T. (2013). Cellular Senescence in Yeast Is Regulated by rDNA Noncoding Transcription. *Current Biology*, *23*(18), 1794–1798. https://doi.org/10.1016/j.cub.2013.07.048

Schweizer, E., Mackechnie, C., & Halvorson, O. (1969). The Redundancy of Ribosomal and Transfer RNA Genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, *40*(2), 261–277.

Sharma, D. (2021). *A new method for investigating ribosomal RNA gene copy number dynamics* [PhD thesis]. University of Auckland.

Sinclair, D. A., & Guarente, L. (1997). Extrachromosomal rDNA Circles—A Cause of Aging in Yeast. *Cell*, *91*(7), 1033–1042. https://doi.org/10.1016/S0092-8674(00)80493-6

Stephan, W., & Cho, S. (1994). Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. *Genetics*, *136*(1), 333–341. https://doi.org/10.1093/genetics/136.1.333

Stults, D. M., Killen, M. W., Pierce, H. H., & Pierce, A. J. (2008). Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Research*, *18*(1), 13–18. https://doi.org/10.1101/gr.6858507

Szostak, J. W., & Wu, R. (1980). Unequal crossing over in the ribosomal DNA of Saccharomyces cerevisiae. *Nature*, *284*(5755), 426–430. https://doi.org/10.1038/284426a0

Takeuchi, Y., Horiuchi, T., & Kobayashi, T. (2003). Transcription-dependent recombination and the role of fork collision in yeast rDNA. *Genes & Development*, *17*(12), 1497–1506. https://doi.org/10.1101/gad.1085403

Valori, V., Tus, K., Laukaitis, C., Harris, D. T., LeBeau, L., & Maggert, K. A. (2020). Human rDNA copy number is unstable in metastatic breast cancers. *Epigenetics*, *15*(1–2), 85–106. https://doi.org/10.1080/15592294.2019.1649930

Walsh, J. B. (1987). Persistence of Tandem Arrays: Implications for Satellite and Simple-Sequence DNAs. *Genetics*, *115*(3), 553–567. https://doi.org/10.1093/genetics/115.3.553

Wang, M., & Lemos, B. (2017). Ribosomal DNA copy number amplification and loss in human cancers is linked to tumor genetic context, nucleolus activity, and proliferation. *PLoS Genetics*, *13*(9):e1006994. https://doi.org/10.1371/journal.pgen.1006994

West, C., James, S. A., Davey, R. P., Dicks, J., & Roberts, I. N. (2014). Ribosomal DNA Sequence Heterogeneity Reflects Intraspecies Phylogenies and Predicts Genome Structure in Two Contrasting Yeast Species. *Systematic Biology*, *63*(4), 543–554. https://doi.org/10.1093/sysbio/syu019

Xu, B., Li, H., Perry, J. M., Singh, V. P., Unruh, J., Yu, Z., Zakari, M., McDowell, W., Li, L., & Gerton, J. L. (2017). Ribosomal DNA copy number loss and sequence variation in cancer. *PLoS Genetics*, *13*(6), e1006771. https://doi.org/10.1371/journal.pgen.1006771

Zafiropoulos, A., Tsentelierou, E., Linardakis, M., Kafatos, A., & Spandidos, D. A. (2005). Preferential loss of 5S and 28S rDNA genes in human adipose tissue during ageing. *The International Journal of Biochemistry & Cell Biology*, *37*(2), 409–415. https://doi.org/10.1016/j.biocel.2004.07.007

Zhang, X., Eickbush, M. T., & Eickbush, T. H. (2008). Role of Recombination in the Long-Term Retention of Transposable Elements in rRNA Gene Loci. *Genetics*, *180*(3), 1617–1626. https://doi.org/10.1534/genetics.108.093716

Zou, H., & Rothstein, R. (1997). Holliday Junctions Accumulate in Replication Mutants via a RecA Homolog-Independent Mechanism. *Cell*, *90*(1), 87–96. https://doi.org/10.1016/S0092-8674(00)80316-5