Measurement and phenotypic consequences of ribosomal DNA copy number variation

Ashley Nicole Hall

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Christine Queitsch, Chair

Bonita Brewer

Robert Waterston

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Measurement and phenotypic consequences of ribosomal DNA copy number variation

Ashley Nicole Hall

Chair of the Supervisory Committee:

Christine Queitsch

Department of Genome Sciences

The ribosomal DNA (rDNA) encodes the ribosomal RNAs and is present in tens to thousands of tandemly repeated copies in eukaryotic genomes. rDNA copy number varies among individual humans and within model organisms. Having too few rDNA copies restricts ribosome biogenesis and leads to reduced growth rate or death. Variation within the naturally occurring range of rDNA copy numbers likely leaves ribosome biogenesis intact but still confers cellular phenotypes, such as altering the global transcriptome in humans. The extent to which natural rDNA copy number variation impacts whole-organism phenotypes such as fitness and lifespan is poorly understood.

A prerequisite to understanding the impact of rDNA copy number on phenotype is accurate copy number measurement. The gold standard for rDNA copy number measurement is pulsed-field gel electrophoresis and Southern blotting. Because this method is low-throughput and ill-suited for human rDNA measurement, many studies rely on short-read sequencing data. However, we found that sequencing-based rDNA copy number estimates are highly error-prone. This high technical error rate directly impacts conclusions: A study assessing 168 low coverage sequencing samples concluded that the 5S and 45S rDNA arrays co-vary in copy number in humans. With thousands of samples from newer, higher quality datasets, I demonstrated that there is no meaningful co-variation between the 5S and 45S array copy numbers in humans.

I further generated resources of *Caenorhabditis elegans* with rDNA copy numbers from the high and the low end of the worm's natural range, including a set of recombinant inbred lines. Changing rDNA copy number did not result in a detectable change in rRNA abundance, consistent with these copy numbers supporting functional ribosome biogenesis. I found that in the naturally occurring rDNA copy number range of *C. elegans*, copy number differences confer no aging or fitness defects. In addition, this rDNA copy number variation confers few differences in global gene expression. These results suggest that any phenotypic consequences of rDNA copy number variation in the naturally occurring range are subtle, at least in *C. elegans*, and to achieve substantial phenotypic consequence, rDNA copy number must be pushed to extremely high or low levels.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I have many people who have helped me along the way through my PhD. First, I need to thank my adviser, Christine Queitsch. Christine took me into her lab as a 3rd year graduate student and gave me ample room to take a portion of the rDNA project in my own direction. For this, I am grateful.

All members of the Queitsch lab have been incredibly friendly, welcoming, and helpful. I have learned a lot from each person. I would like to specifically thank Elizabeth Morton, who has been my "science buddy" throughout this project. We have co-contributed to many of the same projects, and she has been instrumental in my learning to be a worm biologist.

I would also like to thank my undergraduate research advisor, Daniel Kearns. I was very lucky to be placed in the Kearns lab in my undergraduate research program, and that experience has greatly shaped my research career. Dan has continued to support me throughout graduate school, especially at the times I needed it the most. Also from the Kearns lab, I would like to thank Rebecca Calvo, who was my graduate student mentor. She taught me how to pipette.

My parents have been very supportive of my graduate school journey. They may not understand the details of what I work on but have always been encouraging in discussing what I'm working on.

My husband Jeremy has been incredibly supportive throughout my graduate school career. He always encourages me to keep my end goals in mind. He has been the constant throughout my graduate school journey, and I can't imagine having completed this without him.

To end, I must mention my cats, Tootsie and Rosie. They were a great support system when I had to do more working from home during the pandemic, and they were a popular visitor in Zoom lab meetings. While it is physically more difficult to write with a cat across my arms, it is still a much more pleasant experience than doing it alone.

DEDICATION

I dedicate this dissertation to all of the other graduate students who choose to dive down the rabbit hole. If you have landed here, I wish you luck.

Chapter 1. INTRODUCTION

The ribosomal DNA encodes the rRNAs, essential structural and catalytic components of the ribosome. The rDNA itself is highly repetitive and highly conserved, found as one or many tandem arrays in most eukaryotic genomes. There are tens to thousands of rDNA repeat units per genome, often covering megabases of DNA. The number of rDNA copies per genome, or "rDNA copy number" varies between strains of model organisms such as yeast and worms and between individual humans [1–4]. rDNA copy number can be difficult to measure accurately, due to its repetitive nature and immense size [5]. Recent studies have identified changes in rDNA copy number in cancer and aging, but few causal relationships between rDNA copy number and disease have been described. Differences in rDNA copy number also correlate with differences in global gene expression in cell culture and flies, which suggests that rDNA copy number differences have consequences for cellular physiology [2,6]. How these cellular changes translate to whole-organism traits remains understudied. This introduction chapter will describe the structure and function of the ribosomal DNA locus and relevant human health consequences to variations in the rDNA. To give a more holistic view of the rDNA locus, I will first discuss the key features of rRNA expression, followed by structural and sequence variation at the rDNA. Finally, I will discuss rDNA copy number variation, known phenotypic consequences, and relevance to human health, aging, and disease.

1.1 Regulation and Expression of rRNAs

The core function of the rDNA is to be the template for rRNA expression. Rates of ribosome biogenesis are largely dictated by rates of rRNA production [7,8]. Expression of the rRNAs is a tightly regulated, multistep process that forms the nucleolus. As such, rDNA arrays are also known as nucleolar organizer regions, or NORs [9]. Eukaryotic genomes encode the 18S, 5.8S, and 28S rRNAs in a co-transcribed operon, referred to as the 45S in humans [10]. Throughout the introduction, I will refer to these rRNAs by the human nomenclature. The 45S pre-rRNA is transcribed by RNA Polymerase I (RNA Pol I) in the nucleolus and is then processed into the mature 18S, 5.8S, and 28S rRNAs, including removal of external and internal transcribed spacers and application of various base modifications [11–13]. In contrast, the 5S is transcribed by RNA Polymerase III, often outside of the nucleolus, and is then imported into the nucleolus for ribosome assembly [14,15]. Despite being transcribed by different machineries, there are similarities in the regulation of the 45S and 5S loci [16]. Though genomes can have hundreds to thousands of rDNA copies, typically 50% or fewer are actively transcribed [17–19]. The number of rRNA genes that are actively transcribed is regulated in many ways, including transcription factors and chromatin remodeling proteins [20].

rDNA repeat units can broadly be defined as active or inactive based on their chromatin state [17]. Most rRNA transcriptional regulation comes from modulation of open chromatin repeats, which include actively transcribed and transcriptionally poised repeat units. In mammals, actively transcribed repeat units are associated with the transcription factors Upstream Binding Factor (UBF) and selectivity factor 1 (SL1), as well as RNA Pol I [21]. Poised rDNA units are not actively transcribed but can be activated upon need for increased ribosome

biogenesis, such as changes in nutrient state [22,23]. On the other hand, inactive rDNA repeats can be silenced by heterochromatin and/or DNA methylation. Silencing by CpG methylation is found in mammals, and the repeats silenced by CpG methylation are likely also heterochromatic [21,24]. DNA methylation is absent in both yeast and worms, but rDNA repeats are still silenced with repressive chromatin in these organisms [24–26]. A further level of rDNA transcriptional control is at the level of whole rDNA arrays. In organisms with multiple rDNA arrays, some arrays are completely silenced and reside outside of the nucleolus; these silencing patterns can be heritable [20,27,28].

The most well-known example of whole rDNA array silencing is nucleolar dominance. Nucleolar dominance was first identified in interspecific hybrids of plants, in which the rRNA genes from one parent are expressed and those of the other are entirely silenced [29,30]. Since it was initially discovered, nucleolar dominance in interspecific hybrids has been described in many plants and animals [31,32]. Nucleolar dominance can also be found within a species, including, for example, *A. thaliana* and *D. melanogaster*. The mechanism by which nucleolar dominance is established is not completely understood, though DNA methylation, histone methylation, histone deacetylation, and siRNAs have all been shown to play a role [30]. In addition, nucleolar dominance is established during development: early embryos of *Arabidopsis* and *Drosophila* express rRNAs from all NORs but transition to expression of rRNAs from specific NORs over time [33–35].

Beyond silencing whole arrays, rRNA expression levels are regulated developmentally, in a tissue-specific manner, and in response to stress. One example of increased rRNA expression in a developmental stage is during oogenesis in *Xenopus* species [36]. To meet ribosome

biogenesis demands, extrachromosomal rDNA circles are amplified to provide a higher number of rDNA templates, a process found in many cold-blooded animal species [36–38]. A different mechanism to increase rRNA production in oogenesis is to increase expression of rDNA transcription factors, as observed in mice [39]. In somatic tissues, some processes, such as cardiac hypertrophy, have increased rRNA expression while others, such as osteoblast differentiation, have lower rRNA expression [40–42]. However, steady-state levels of rRNA of different organs such as the liver, brain, and lung, do not differ in mice [43]. Finally, rRNA expression changes in response to stresses including heat shock, nutritional stress, hypoxia, and DNA damage [44–48]. The importance of appropriate regulation of rRNA expression is noted in the many ways by which it is controlled -- and the consequences that arise when it is not.

Ribosomopathies are diseases characterized by defects in ribosome biogenesis with diverse pathologies [49]. These defects can be mutations in ribosomal proteins or deficiencies in ribosome assembly. However, many are rRNA-centric, including mutations in rRNA genes, rRNA transcription factors, or rRNA processing factors [49,50]. One example is Diamond-Blackfan anemia, a congenital bone marrow failure that can be caused by ribosomal protein mutations that not only affect the ribosomal protein, but also disrupt pre-rRNA processing [51–54]. Another example is cartilage hair hypoplasia, caused by a mutation in RMRP, a long-noncoding RNA that is involved in pre-rRNA processing, preventing appropriate maturation of the 5.8S and 18S rRNAs [55,56]. Finally, X-linked dyskeratosis congenita is a bone marrow failure caused by a mutation in the DKC1 enzyme, which performs rRNA pseudouridylation [57]. While this list of diseases with defects in rRNA production or processing is not exhaustive, these examples demonstrate the breadth of rRNA-related processes that, when perturbed, cause a disease phenotype.

Many ribosomopathies are associated with an increased cancer risk -- and aberrations in rRNA expression, modification, or processing are common in cancers [58,59]. One of the first identified hallmarks of cancer is an increase in nucleolar size, which corresponds to an increase in rRNA transcription [60–63]. Having larger nucleoli, and thus more rRNA transcription, is a bad prognosis for the cancer outcome [64]. Even model tumors, such as the proximal germline tumor in *C. elegans*, show aberrations in ribosome biogenesis [65]. Because high levels of rRNA transcription are common in cancer, inhibitors of RNA Pol I are being investigated as cancer therapeutics [66,67]. Overall, appropriate regulation of rRNA biogenesis is important for development and viability of an organism, as well as for maintaining health. Beyond rRNA expression, however, variation at the rDNA locus itself is abundant and also implicated in human health and disease.

1.2 RIBOSOMAL DNA LOCUS STRUCTURE AND ORGANIZATION

The structure of the rDNA loci is important for understanding rDNA biology. The rRNA genes are arranged in tandem arrays in most eukaryotes, present on one or more chromosomes and in tens to thousands of copies. In some species, such as the yeast *Saccharomyces cerevisiae*, the 5S rRNA is encoded in the same repeat unit as the 45S, an arrangement termed the "linked" or "L-type" rDNA structure [68,69] (**Figure 1.1**). More common, however, is the "separate" or "S-type" rDNA, in which the 5S rRNA gene is present as its own tandem array. The S-type arrangement is found in many organisms, including humans, flies, and worms [70] (**Figure 1.1**). Meanwhile, the L-type arrangement is observed in <5% of plant species and a few arthropods and crustaceans [71–73]. In plants, multiple rearrangements have occurred, and each S- and L-

type rDNA have evolved multiple times [74–77]. Similarly in fungi, many rearrangements of the position, orientation, or location of the 5S genes with respect to the 45S genes have occurred [78]. Whether there are benefits to encoding the 5S in the same or a different tandem repeat with respect to the 45S is not known.

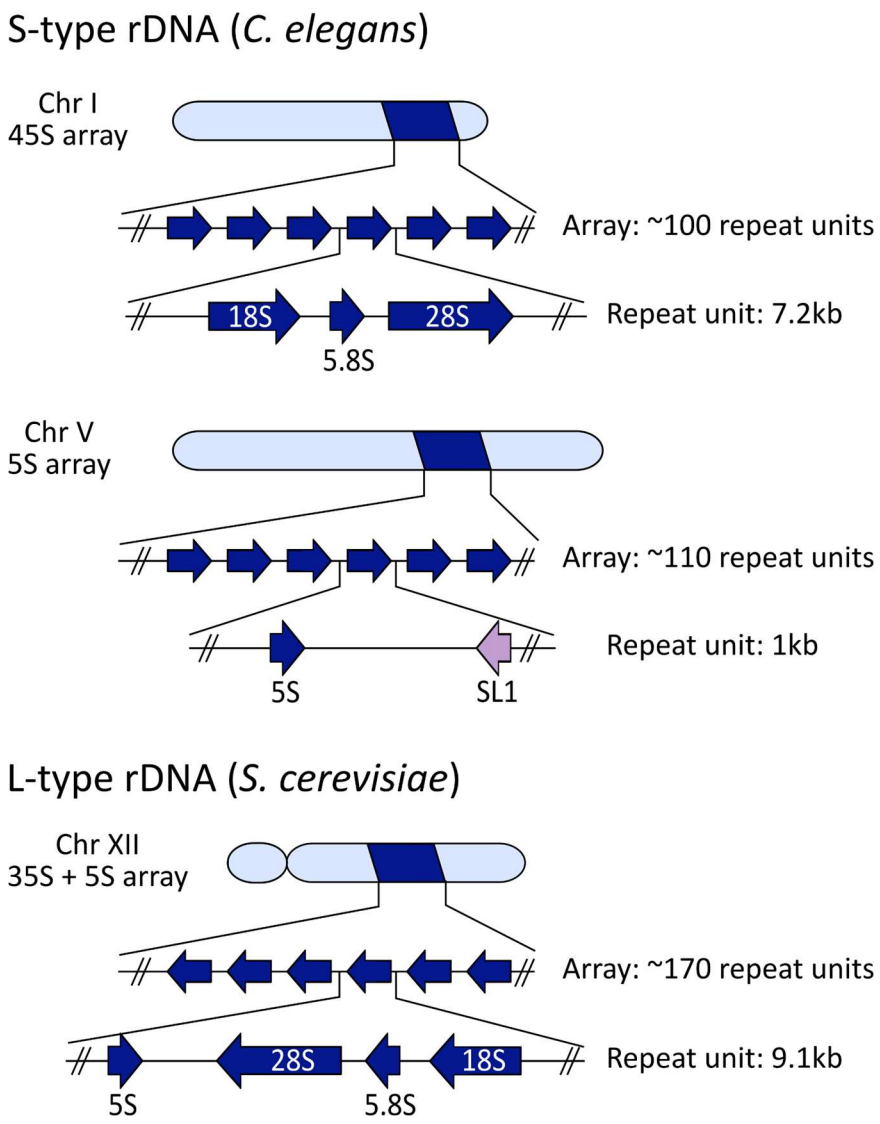

Figure 1.1: Organization of the rDNA arrays in *C. elegans* and *S. cerevisiae*.
Top: Schematic of the rDNA arrangement in *C. elegans*, which has S-type rDNA, encoding the 45S as a single array on chromosome I and the 5S as a single array on chromosome V. The *C. elegans* 5S repeat unit also encodes SL1, a trans-spliced leader transcript. Bottom: Schematic of the rDNA arrangement in *S. cerevisiae*, which has L-type rDNA. In *S. cerevisiae*, the 5S is transcribed in the opposite orientation as the 35S from an array on chromosome XII.

In addition to the arrangement of the 45S and 5S arrays, the number of chromosomes with rDNA loci and their respective positions vary. Positionally, rDNA can be located in terminal (subtelomeric), pericentromeric, or interstitial positions within the chromosome. Number and position of rDNA arrays across eukaryotes has been cataloged in two databases: The plant rRNA database and the animal rRNA database [73,79]. The number of chromosomes with rDNA does not necessarily correlate with the total number of chromosomes in a genome, as genomes with many chromosomes can have few rDNA loci and vice versa [79]. In animals, the 45S is most often in a terminal position, whereas the 5S shows less of a bias in positioning [79]. Even closely related species may differ in rDNA array location and number. In the *Mus* subgenus *Mus*, for example, all strains have 40 chromosomes per diploid genome, but have 6-40 rDNA sites, including a species that carries rDNA on every chromosome [80]. Variation in the number and position of rDNA loci has also been observed in fish, where increased numbers of rDNA loci occur in fish grown in polluted waters [81]. Even in humans, where 45S rDNA arrays are present on the five acrocentric chromosomes (chromosomes 13, 14, 15, 21, and 22) and thus 10 loci per diploid genome, up to four of these loci lack detectable rDNA sequence in some samples [82]. The prevalence of acrocentric chromosomes lacking rDNA in humans is unknown.

Within the context of the rDNA array organization, aberrations can be found. Both palindromic and inverted repeat units are present in human rDNA arrays, disrupting the pure tandem array structure [83,84]. It was initially estimated that up to 33% of rDNA repeat units in a healthy human cell line are in a non-tandem orientation [83]. More recently, methods incorporating long-read sequencing demonstrated that most human rDNA arrays are in the same orientation [85,86]. Whether these discrepancies are due to differences in technologies between

the earlier and later studies, or reflect genuine differences in the cell lines analyzed, is unknown. Beyond these large structural differences between rDNA arrays in organisms, sequence variation of the rDNA arrays is commonly found both within and between species.

1.3 SEQUENCE AND CONTENT VARIATION IN THE rDNA

The ribosome is a molecular machine conserved across all kingdoms of life, and the rRNAs that compose the ribosome have high structural and sequence conservation [87–89]. Eukaryotic rRNAs contain a number of expansions in the 28S and 18S as compared to the respective 23S and 16S rRNAs of prokaryotes [90]. Within and between individuals of a species, the sequence of rDNA repeat units is highly conserved, even in the spacer regions [91]. Concerted evolution maintains rDNA array homogeneity and is driven by a combination of intrachromosomal homologous recombination and gene conversion [92,93]. Despite the concerted evolution process, variation in rRNA gene sequence is still present - even within individuals. Rounds of mitotic cell division can result in *de novo* variation in the rDNA in humans [94,95]. In mice, humans, and plants, rRNA coding variants display tissue-specific expression or are differentially expressed under stress conditions, which suggests that the variants affect ribosome function [96–99]. The specific implications of condition- and tissue- specific variants in rRNAs have been recently reviewed in the context of rRNA evolution [100], and a growing interest in specialized ribosomes is bringing the notion of rRNA coding variants more into view [101,102].

Moving up in scale, retrotransposons that disrupt rRNA genes are found in many organisms [103]. The R1 and R2 retrotransposons are non-long terminal repeat (non-LTR) retrotransposons that insert site-specifically into the 28S rRNA in many species of arthropods

[104,105]. The fraction of rDNA repeat units containing rRNAs disrupted by R1 or R2 retrotransposons in *Drosophila* varies between species and between strains but can be less than 5% of repeat units to more than 80% [106–109]. Some nematode species have R4 retrotransposons that disrupt the 28S, though these are absent in *C. elegans* [110]. *De novo* insertion of retrotransposons can change how many rDNA units are disrupted, which can vary between individuals in a population [108,111]. The disrupted rRNAs in *Drosophila* are not transcribed, indicating that retrotransposon-disrupted repeat units are nonfunctional [112]. Recently, however, these retrotransposons have been found to have functional importance outside of affecting rRNA transcription: In *D. melanogaster*, expression of R2 is important for rDNA copy number maintenance [113]. Overall, these retrotransposons complicate rDNA analysis, as disrupted and intact repeat units appear to have different biological functions.

Variation in the noncoding portion of the rDNA repeat unit is considerably more abundant than variation in the rRNA genes. The intergenic spacer (IGS) is the nontranscribed portion of the rDNA repeat unit, and its content varies greatly between species. The size discrepancy between rDNA repeats of different eukaryotes -- such as the 7.2kb repeat unit of *C. elegans* compared to the ~43kb repeat unit of humans -- is largely due to differences in the IGS. Mammals typically have long rDNA repeat units with large intergenic spacers, with humans and mice each having approximately 30kb of IGS [114,115]. While primates have similarly sized intergenic spacers to humans (24-30kb), content of the IGS differs [116]. For example, humans and apes encode a *cdc27* pseudogene in the IGS, which is absent in monkeys [116,117]. Eukaryotic IGS sequences often carry enhancers or promoters for the rRNA genes in many copies. Length variants in the IGS sequences of closely related organisms often arise from differing numbers of enhancer

elements, and longer IGS sequences are associated with increased rRNA production [118–120]. Even within a species, IGS sequences vary. For example, in *Arabidopsis thaliana*, 18 unique rDNA length variants are found between related plants of the same wild type genotype [121]. Further, in humans, a 2kb length variant in the IGS has been observed and may stratify by population [86], though the functional consequence of this variant has not yet been determined.

Much of this section has focused on variants in the 45S, but there is also variation in the 5S in organisms with S-type rDNA. The 5S array is typically smaller than the 45S due to a shorter repeat unit length: in humans the 5S repeat unit is 2.3kb, compared to the 43kb of the 45S repeat unit [122]. Despite this smaller size, other features, including histone genes and small RNAs, are encoded in the 5S repeat unit for some organisms [72,123,124]. The shorter length of the 5S repeat unit makes it more amenable to long-read sequencing characterization, so differences in individual repeat units across a single array are more readily characterized. Long-read sequencing has been used to assemble across the 5S array in *C. elegans* and *C. briggsae*. Twenty-two unique 5S repeat units were identified in the laboratory strain of *C. elegans*, and their orders and orientations were determined in multiple strains [125]. Further, the shorter length of the 5S makes it amenable to copy number estimation by pulsed-field gel electrophoresis in most organisms, whereas the 45S arrays can only be resolved by pulsed-field gel electrophoresis when total array lengths are less than 6Mb -- a size that is often exceeded in human 45S rDNA arrays [126].

1.4 rDNA COPY NUMBER VARIATION AND MAINTENANCE

Although rDNA copy number variation has been characterized only in a limited number of species, it has been found in each species in which it has been assessed. Numbers of rDNA copies span the tens to thousands of repeat units per genome, depending on the organism. While it was reported that rDNA copy number positively correlates with genome size, this correlation fails to take into account intra-species rDNA copy number variation [127]. Among natural isolates of a species, rDNA copy number can vary: In *S. cerevisiae*, there are 54 to 511 rDNA copies per haploid genome, in *C. elegans*, between 68 and 418 rDNA copies per haploid genome, and in maize, between 1,061 to 17,347 rDNA copies per haploid genome [1,128,129]. Many studies have estimated rDNA copy number variation between individual humans, with ~200-600 rDNA copies per haploid genome being the most consistent range identified [2,4,97,130]. Beyond intra-individual rDNA copy number variation, there is interest in whether rDNA copy number varies among tissues in humans. Thus far, no substantial differences in rDNA copy number have been observed among tissues in mice or chickens, suggesting that it is unlikely that there will be substantial differences among tissues in humans [94,131].

Many mechanisms generate rDNA copy number variation. In sexually reproducing species, one component to inter-individual rDNA copy number variation comes from the combination of parental arrays of differing sizes to make the offspring. Further, meiotic recombination within the arrays can change rDNA copy number, occurring at a frequency of about 10% per meiosis in humans. This frequency typically results in one rDNA array in a child that does not have the same size as any of the parental arrays [132]. Combined, these processes result in the rDNA copy number of the offspring being close to the average of the rDNA copy

numbers of the two parents [4]. Outside of meiosis, mitotic unequal sister chromatid exchange occurs in the rDNA, observed in both yeast and humans [132,133]. In DNA repair processes, double strand break repair in the rDNA typically causes copy loss [134]. Finally, cellular stressors including heat shock and heavy metal exposure also cause rDNA loss [135–138]. Whether environmental stressors are a common and prevalent mechanism by which rDNA copy number variation arises in natural populations is not known.

Despite the many ways by which rDNA copy number can change, rDNA copy number is also quite stable. With inbreeding populations such as *C. elegans*, the rDNA copy numbers of wild strains are stable through laboratory propagation [5]. This stability is due in part to suppression of meiotic recombination within the rDNA, which causes the rDNA array to be transmitted as a single Mendelian locus [139]. Further, unequal mitotic sister chromatid exchange is suppressed at the rDNA by binding of cohesin and Sir2 [140,141]. If rDNA copies are lost, copies can be regained through rDNA amplification, found in *S. cerevisiae*, and rDNA magnification, found in *D. melanogaster* [142,143]. When a critical number of rDNA repeats have been lost, yeast cells appear to amplify extrachromosomal rDNA circles and reinsert them into the genome, increasing rDNA array size over generations [143–146]. In *D. melanogaster*, recovery of rDNA copies occurs quickly over fewer generations through unequal sister chromatid exchange [142,147,148]. Whether similar mechanisms maintain rDNA copy number in human cells is not known.

The question of whether rDNA copy number is controlled by other genetic loci has been posited, in part due to the variation and stability of rDNA copy number among wild isolates of organisms. In the laboratory yeast *S. cerevisiae*, a number of genes that, when mutated, result in rDNA copy number changes or rDNA instability have been identified [143,149–155]. However, in

the cases of some screens, interpretation of the interaction of rDNA copy number with gene knockouts is confounded by the fact that lithium acetate exposure, a common method for transforming yeast, causes spontaneous rDNA copy number changes [156]. Therefore, secondary verification of a gene's effect on rDNA copy number or stability is often necessary. Altogether, mutations that disrupt rDNA replication or transcription are those with the most evidence for increasing or decreasing rDNA copy number [143,149,150,152,154,155,157]. Beyond replication and transcription, modulation of mTOR activity induces rDNA copy number variation in yeast and flies. In *S. cerevisiae*, inhibition of mTOR with rapamycin or caloric restriction represses rDNA amplification [158]. In *Drosophila melanogaster*, overfeeding flies causes a heritable loss of rDNA copies, which is dependent on mTOR signaling [148]. Even without a full understanding of the causes of rDNA copy number variation, many consequences of its variation are being explored.

1.5 KNOWN PHENOTYPIC CONSEQUENCES OF rDNA COPY NUMBER VARIATION

A key question that I address in this dissertation is whether rDNA copy number variation in the naturally occurring range has phenotypic consequences. Many studies have demonstrated consequences of reducing rDNA copy number below what is observed in nature - most commonly in yeast and flies. **Figure 1.2** illustrates both known phenotypes that associate with different rDNA copy number ranges, and my hypothesis regarding how rDNA copy number variation affects phenotypes.

At the lowest rDNA copy numbers, ribosome biogenesis is restricted. At the most extreme, these low levels of ribosome biogenesis cause death during development or embryogenesis of multicellular organisms [159–162] and slow growth in yeast [163]. With

enough rDNA copies to evade death, restricted ribosome biogenesis causes the bobbed phenotype in *D. melanogaster*, characterized by slower development, a shorter abdomen, and shorter scutellar bristles [164]. In plants, rDNA copy number reduced to approximately 10% of wild type results in transcriptional signals of rRNA expression compensation, but no gross morphological defects [165]. Less severe reductions of rDNA copy number have been engineered that leave ribosome biogenesis intact but represent a theoretical minimal copy number required for wild-type levels of ribosome biogenesis [18]. Such reductions have been shown to affect other cellular phenotypes. For example, in *S. cerevisiae*, reduction from ~170 rDNA copies to ~30 rDNA copies causes early replication of the rDNA and a delay in replication of other genomic regions [166].

Modest changes in rDNA copy number, which are more likely to be encountered in naturally occurring populations, are associated with differences in cellular phenotypes. In humans, rDNA copy number inversely varies with mtDNA abundance [2] and in *D. melanogaster*, mtDNA gene expression is lower in organisms with higher rDNA copy number [6]. rDNA copy number in the natural range is also reported to associate with global changes in the transcriptome in humans [2]. Finally, moderate reductions in rDNA copy number affect chromatin silencing, observed in position effect variegation in *D. melanogaster* and Sir2-mediated silencing in *S. cerevisiae* [167,168].

The rDNA is also involved in some components of nuclear organization. The nucleolus is the most readily identified feature of nuclear architecture, and changes in nucleolar structure can affect genome structure [169]. Certain regions of chromosomes associate with nucleoli [170,171] and the rDNA tends to contact regions of repressed chromatin [172–174].

Chromosomal associations with the nucleoli are largely stable, even through cellular senescence [175]. Changes in rDNA copy number would naturally change the total amount of genomic real estate that can interact in the nucleolus and could therefore affect nuclear organization as a whole.

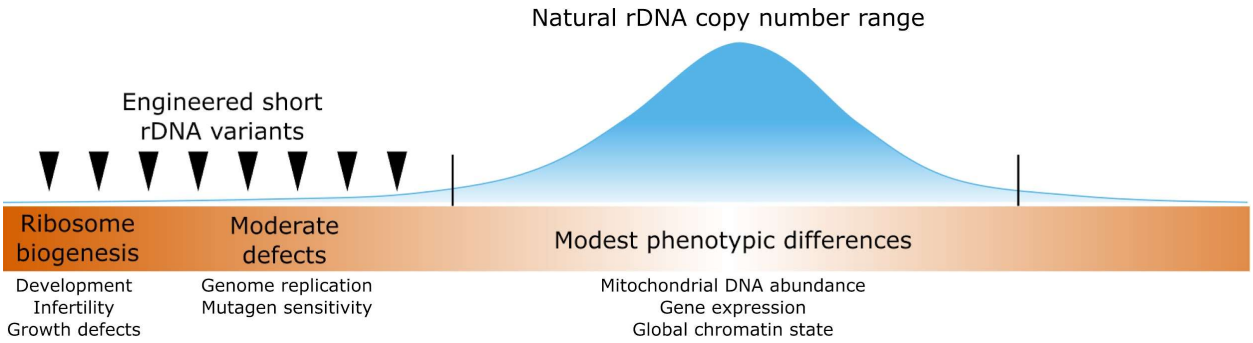

Figure 1.2: Known consequences of rDNA copy number variation on phenotype.
Modest differences in phenotype have been associated with rDNA copy number variation in the natural range. Below what is naturally occurring, engineered short rDNA variants have been used to demonstrate effects on ribosome biogenesis and genome replication, among other traits.

1.6 THE rDNA LOCUS AND AGING

The rDNA locus is tightly tied to aging. At the level of rRNA transcription, reductions in ribosome biogenesis are a universal method for lifespan extension [176–178]. Nucleolar size, which correlates with levels of rRNA transcription, is a molecular marker of longevity [47,177]. Indeed, gene disruptions that extend lifespan in model organisms also reduce nucleolar size, even if these are not mutants classically associated with reduced protein synthesis [179]. Conversely, premature aging disorders such as Hutchinson-Gilford progeria syndrome have increases in rRNA expression [180]. The rDNA, being the template for rRNA expression, could feasibly contribute to differences in aging.

Changes in rRNA expression may occur during the normal aging process. Nucleoli get larger with increased passaging of human cell lines, and large nucleoli are observed in progeria [180,181]. Some studies report changes in rRNA levels with age in healthy humans or wild type laboratory organisms [182,183]. Others, however, find no difference in rRNA expression with organismal age or cellular senescence [184,185]. While typical rRNA expression may or may not change with age, ribosome biogenesis defects and deregulation of protein synthesis occur in aging and premature aging disorders [186,187]. In line with possible changes in rRNA expression, epigenetic changes occur at the rDNA during aging. Methylation of the rDNA, including the promoter region, increases with age [188–192]. The methylation of the rDNA is elevated compared to the rest of the genome, and has been used to develop an epigenetic clock to predict biological age [193–195].

rDNA instability is a hallmark of aging in yeast and is found in some aging mammalian tissues. Instability manifests in two different ways: with the formation of extrachromosomal

rDNA circles (ERCs) [196,197], or through the loss of rDNA copies from breaks in the rDNA [134]. Age-associated ERC accumulation is found only in yeast; in *D. melanogaster* and humans, rDNA circles are present but do not accumulate with age [198,199]. In mammals, age-associated rDNA instability presents as rDNA copy loss. How fundamental this rDNA copy loss is to aging is under debate, as contrasting studies have shown both loss [200–203], no loss [204], and recently, even gains [205] in various tissues with age. Importantly, none of these studies are longitudinal, so rDNA copy numbers are not measured in the same individuals early and late in life. Beyond somatic rDNA loss, old male flies exhibit heritable loss of rDNA copies in the germline [206], further evidence of the possible age-dependence of rDNA copy number variation. Whether or not similar loss of rDNA copies may be found in human germlines has not been determined. Despite the many published connections between the rDNA locus and aging, it has never been clearly established whether changing the rDNA copy number of a multicellular organism causes a difference in lifespan.

1.7 HUMAN HEALTH RELEVANCE OF rDNA COPY NUMBER VARIATION

Many diseases have incompletely understood genetic underpinnings. Repetitive DNA regions such as the rDNA are underexplored sources of heritable variation with the potential to contribute to disease risk [207–209]. Of great interest is if inherent rDNA copy number can predispose an individual to disease. Associations between rDNA copy number and complex diseases such as schizophrenia and Alzheimer's disease implicate differences in copy number and ribosome biogenesis in disease state [210]. A recent study determined that higher rDNA copy number increases lung cancer risk in smokers [211]. This study was small and limited to

individuals who smoke but opens the possibility for future studies of rDNA copy number and cancer risk.

In cancer, expansion of nucleoli and misregulation of ribosome biogenesis has long been associated with aggression and a negative prognosis [60,66]. Recently, rDNA copy number changes have come into view as a cancer-associated trait. rDNA copy number losses and gains have been observed in various cancers, with loss being more prevalent [94,136,212–217]. AIDS-related lymphoma, adenocarcinomas, and osteosarcomas are among the cancers with reported decreases in 45S rDNA copy number [94,215], whereas gastric cancers have reported increases in rDNA copy number [136]. It is not fully understood whether these changes in rDNA copy number are a cause or a consequence of cancer proliferation. In breast cancer, both gains and losses of rDNA copy number are found, so the changes are argued to be a consequence of increased genome instability [217]. Along these lines, it is speculated that cancer driver mutations cause a loss of heterochromatin, leaving the rDNA susceptible to damage and instability [216,217]. Others have proposed that reductions in rDNA array size may facilitate rapid replication during cancer cell propagation [94]. Another model proposes that when rDNA is lost in cancers with elevated rRNA synthesis levels, the loss is an anti-cancer adaptive process [49].

To understand the origins of rDNA copy number variation in cancer, other associated genetic changes have been assessed. Changes in rDNA copy number are not accounted for by aneuploidy of rDNA-bearing chromosomes [94,215]. However, copy number increases in DNA damage response and metabolic genes are associated with rDNA copy loss [94]. Further, inactivation of certain tumor suppressors, such as TP53, PTEN, and ATRX associates with rDNA copy loss in humans and promotes rDNA copy loss in mouse tumor models [94,215,216].

Mutation of BLM helicase increases cancer risk, and human cell lines carrying this mutation have high levels of rDNA instability [218–220]. In contrast, P53 is required for rDNA copy gain in gastric cancers [136].

It is possible that a combined understanding of rDNA copy number changes and tumor suppressor mutations could be informative for predicting cancer risk or for developing cancer treatments. Indeed, the changes of rDNA copy number in many cancers have positioned rDNA copy number assessment as an avenue to inform cancer therapy. The effectiveness of RNA Pol I inhibitor CX-5461 as a cancer treatment depends on rDNA chromatin state, which is in part connected to changes in rDNA copy number [221]. A greater understanding of how rDNA copy number variation impacts the efficacy of RNA Pol I inhibitors could therefore be a valuable foundation for specialized cancer treatments.

While cancer is probably the best-explored human disease with respect to rDNA copy number, other complex human diseases have been associated with rDNA copy number variation or with variation in the number of active rDNA copies. Individuals with schizophrenia display both increased ribosome biogenesis and an elevated rDNA copy number [222–224]. Aberrant regulation of rRNA biogenesis and rDNA amplification have been observed in patients with intellectual disability [225]. Individuals with Down syndrome, who have an extra rDNA array due to the additional copy of chromosome 21, often have more active rDNA copies [226]. In contrast, rDNA copy number between children affected by an autism spectrum disorder and a paired unaffected sibling do not differ [4]. A caveat, however, is that the studies on Down syndrome, schizophrenia, and intellectual disability frequently had restricted sample sizes, and measurement of rDNA copy number was often performed with methods of limited resolution.

Thus far, no direct causative relationship between rDNA copy number and a complex human disease has been identified.

1.8 Overview of the dissertation

In this introductory chapter, I discussed the characteristics of the rDNA and known roles of the rDNA in both model organism phenotypes and human disease. From here, I will focus on rDNA copy number variation, first discussing rDNA copy number measurement, then discussing how variation in rDNA copy number affects phenotypes in *C. elegans*.

In Chapter 2, I assess the accuracy of rDNA copy number estimation methods. Using the model organisms *S. cerevisiae* and *C. elegans*, we demonstrate that rDNA copy number estimates from short read sequencing are error-prone and highly sensitive to batch effects. The take-home message is that pulsed-field gel electrophoresis remains the gold standard for copy number estimation, and that caution needs to be taken when analyzing short-read sequencing data.

In Chapter 3, I demonstrate how different sequencing datasets can directly affect conclusions drawn about rDNA copy numbers. Specifically, I demonstrate a lack of meaningful co-variation between the copy numbers of the 5S and 45S rDNA arrays in humans. I use thousands of short read sequencing samples produced with uniform library preparation methods and compare these to an earlier dataset. This result contradicts the previously published claim that the 5S and 45S rDNA arrays display “concerted copy number variation”, which was built on a small number of samples with sequencing libraries prepared with a variety of methods. Together Chapters 2 and 3 make the compelling argument that we must use caution when drawing conclusions from sequencing-based rDNA copy number estimates.

In Chapter 4, I describe the strain resources I made to study rDNA copy number variation in *C. elegans*. These include both a set of recombinant inbred lines and a set of near isogenic lines, with differing rDNA copy numbers from wild isolate backgrounds.

In Chapter 5, I apply the strains described in Chapter 4 to answer the question of whether rDNA copy number variation in the naturally occurring range has a measurable phenotypic effect. This effort focuses on fitness and aging traits, and I demonstrate that differences in rDNA copy number do not produce differences in the phenotypes assessed.

Availability of chapters:

Chapter 2 is part of the following published manuscript:

Morton, E. A., **Hall, A. N.**, Kwan E., Mok, C., Queitsch, K., Nandakumar, V., Stamatoyannopoulos J., Brewer, B. J., Waterston, R., and Queitsch, C, 2019. Challenges and Approaches to Genotyping Repetitive DNA. G3 Genes Genomes Genet.

Chapter 3 is published as the following manuscript:

Hall, A. N., T. N. Turner, and C. Queitsch, 2021. Thousands of high-quality sequencing samples fail to show meaningful correlation between 5S and 45S ribosomal DNA arrays in humans. Sci. Rep. 11: 449.

For both Chapters 2 and 3, some changes have been made between the above-indicated publications and their presentation in this manuscript, mainly to incorporate information with other publications that have since been published. Chapters 4 and 5 are not published; components of these chapters are included in a manuscript that is in preparation.

CHAPTER 2: MEASUREMENT OF rDNA COPY NUMBER: METHODS AND CAUTIONS

2.1 SUMMARY

Individuals within a species can exhibit vast variation in copy number of repetitive DNA elements. This variation may contribute to complex traits such as aging and disease, yet it is infrequently considered in genotype-phenotype associations. The possible importance of copy number variation is widely recognized, but accurate copy number quantification remains a technological hurdle. Here, we assess the technical reproducibility of several major methods for copy number estimation as they apply to the large repetitive ribosomal DNA array (rDNA). rDNA encodes the ribosomal RNAs and is a tandem gene array in nearly all eukaryotes. Repeat units of rDNA are kilobases in size, often with several hundred units comprising the array, making rDNA particularly intractable to common quantification techniques. We evaluate pulsed-field gel electrophoresis, droplet digital PCR, and Nextera-based whole genome sequencing as approaches to copy number estimation, comparing techniques across model organisms and spanning wide ranges of copy numbers. Nextera-based whole genome sequencing, though commonly used in recent literature, produced high error. We furthermore explore possible causes for this error and provide recommendations for best practices in rDNA copy number estimation. We present a resource of high-confidence rDNA copy number estimates for a set of *S. cerevisiae* and *C. elegans* strains for future use.

2.2 INTRODUCTION

Efforts to understand the genetic basis of complex traits and diseases have almost exclusively focused on the role of single nucleotide variants [209]. However, vast genomic

variation exists beyond the single nucleotide level, in the form of short tandem repeats, long repetitive regions, and transposable elements [209]. This variation remains poorly characterized, not just in humans but even in the most tractable model organisms. The copy number of repetitive DNA elements changes frequently through expansion and contraction, and as a consequence linkage of specific repeat numbers to surrounding single nucleotide variation is markedly reduced [227–230]. Hence, the power to use nearby common single nucleotide variants to tag repetitive DNA genotypes in genome-wide association studies is limited [230,231].

Recent developments in long-read technology, such as PacBio and Oxford Nanopore, have made the assessment of some types of repetitive DNA possible, including complex tandem repeats and short tandem repeats [232,233]. However, long read technologies still suffer from high error rates and do not yet produce reads sufficiently long to resolve long tandem repeats such as the rDNA [233–237]. Despite these limitations, assembly across some larger repetitive regions, such as the centromere of human chromosome 8 or the 5S ribosomal DNA repeat of *C. elegans*, have been accomplished with long read sequencing [125,238]. Short tandem repeats (2-10 base pair repeat units) can also be genotyped with capture-based sequencing methods or with sophisticated computational tools applied to short read sequencing [230,239–243]. Copy number variation of short tandem repeats can have significant impact on phenotype; examples include the well-known polyglutamine expansion disorders such as Huntington's in humans, various examples of incompatibility among closely related species or ecotypes, and altered environmental responses and other complex traits [208,230,244]. In humans, short tandem repeat copy number variation in promoter and enhancer regions is implicated in gene expression

variation of ~3000 loci, some of which could drive signal in previously published GWAS studies for height and schizophrenia [231,245,246].

Short tandem repeats represent just one class of repetitive DNA that is often ignored in genotype-to-phenotype association studies. The characterization of repetitive DNA elements is a critical step in fully understanding human health and disease [207]. However, despite advances in sequencing technology, some types of repetitive DNA remain recalcitrant to genotyping, in no small part due to their extreme lengths, which can traverse tens to hundreds of kilobases. For example, PacBio and Nanopore sequencing of *C. elegans* genomes failed to determine repeat copy numbers for several repetitive DNA loci, among them the rDNA [125,233]. This problem will certainly be exaggerated in humans, in which individual rDNA arrays can scale to more than 6 megabases [132]. Indeed, even the most recent effort to complete the human genome only successfully assembled two of five rDNA arrays [85].

Nearly all eukaryotes maintain their rRNA genes in large, tandemly-repeated arrays known as the rDNA. There are two different pre-rRNA transcripts: The large PolI transcript, variably referred to as the 37S-45S, and the small PolIII transcript known as the 5S. These two transcripts may be encoded as separate arrays, as is the case in humans and *C. elegans*, or encoded as a single array, as is the case in *S. cerevisiae*. A single repeat unit is 9.1kb in *S. cerevisiae*, 7.2kb in *C. elegans*, and 43kb in humans, meaning that in the most extreme cases the rDNA may represent up to 15%, 2.8%, or 1% of these genomes, respectively [1,70,97,128]. In humans, the short arms of the five acrocentric chromosomes (13, 14, 15, 21, and 22) are almost entirely composed of rDNA [70].

Despite the fact that the rDNA arrays are a substantial portion of genomic content, most reference genomes include only a single repeat unit of the rDNA. The actual number of repeat units is remarkably variable [1,2,97,127,154,247–249]. rDNA copy number among natural isolates of the same species frequently varies as much as 10-fold; in *S. cerevisiae* reported numbers range from 54 to 511 copies (per haploid genome) [128], in maize, from 1,061 to 17,347 copies [129], and from 61 to 1,590 copies in humans [97]. rDNA copy number variation in *A. thaliana* largely accounts for the variation in genome size among strains [250].

Studies that have examined the relationship between rDNA copy number variation and phenotype are becoming increasingly common as short read sequencing is becoming increasingly available, and is a seemingly easy and palatable option for rDNA copy number estimation. Such studies have identified effects of rDNA copy number on mitochondrial abundance, global gene expression, position effect variegation in *D. melanogaster*, and susceptibility to diseases including schizophrenia and lung cancer [2,6,168,211,224,251]. A change in rDNA copy number has also been reported in certain cancers: this is typically rDNA loss, but occasional gain of rDNA copies has been observed in breast cancer [94,215–217]. The ability to fully understand the relationship between rDNA copy number and phenotype is, however, ultimately dependent on our ability to accurately quantify that copy number.

This chapter compares major copy number estimation methods available for rDNA, with the goal that principles learned from the rDNA will apply to other tandemly repeated loci of large (kilobase-scale) units such as satellite DNA [235]. These approaches include pulsed-field gel electrophoresis, droplet digital PCR (ddPCR), and short-read whole genome sequencing (WGS). Of these, pulsed-field gel electrophoresis using a contour-clamped homogeneous electric field

(CHEF), followed by rDNA-specific Southern blotting and hybridization, remains a gold standard for rDNA genotyping. The directional switching of the electric field resolves DNA bands in the megabase range, allowing for direct comparison of array size to standardized ladders. However, the method is laborious and the sheer size and multi-locus structure of the rDNA in organisms such as humans limits the utility of this technique [132]. Of PCR-based approaches, quantitative real-time PCR (qPCR) and its derivatives remain popular methods [129,168,206,217,252–254]. However, qPCR is only capable of detecting large changes in copy number and qPCR measurements are often presented as relative, rather than absolute, copy numbers [255]. Recently, ddPCR has gained attention as a more precise alternative to qPCR [94,154,256,257], and offers an estimation of the absolute number of starting target molecules in the reaction. WGS is commonly used to estimate copy number through reads aligned to the repetitive region relative to the rest of the genome [1,2,94,97,123,129,215,258]. While accuracy of WGS in single nucleotide polymorphism detection has been extensively assessed [259,260], its accuracy in large repeat copy number estimation has not been evaluated.

To assess the reliability and reproducibility of these methods, we use a series of test strains with different rDNA copy numbers, in two different model organisms: *C. elegans* and *S. cerevisiae*. These test strains belong to larger collections for which variation in rDNA copy number among strains has been previously reported [1,3]. We find that the CHEF and ddPCR methods can provide reproducible copy number estimates; however, the two methods do not yield the same absolute copy number estimates. In contrast, WGS produces high error in copy number estimation even for technical replicates. This work provides a basis for considering the quality of WGS data in the use for copy number estimation, laying the foundation for future work.

2.3 RESULTS

2.3.1 Short-read sequencing copy number estimates are substantially affected by library preparation condition and are prone to batch effects

In *C. elegans*, the 45S locus encodes the 18S, 5.8S, and 26S rRNAs and is located on the right arm of chromosome I (**Figure 2.1A**). Wild isolates of *C. elegans* were reported to have 68 to 418 copies of the 45S locus per haploid genome [1]. These initial estimates come from a study of 40 *C. elegans* wild isolates, sequenced to high depth with short-read sequencing. To examine the accuracy of these estimates, we pursued further characterization of 8 wild isolates spanning the range of reported rDNA copy numbers with CHEF gel electrophoresis and short-read sequencing. To compare identical biological samples for each strain, a large population of worms was grown and split into separate tubes for parallel genomic DNA extraction and CHEF gel analysis.

For CHEF gel analysis, we embedded worms in agar before proteinase digestion to preserve chromosome integrity. A restriction enzyme was used to digest most of chromosome I, leaving only an intact rDNA array. We ran the genomic fragments on a CHEF gel along with yeast chromosome ladders as size markers and then Southern blotted and hybridized with an rDNA probe (**Figure 2.1B**). Copy number per genome was calculated by dividing the total size of the array by the known size of a single repeat unit (7.2kb). Replicating the CHEF gel results thrice revealed high reproducibility in copy number determination **(Table 2.1, Figure 2.1D)**. A strain with reported rDNA size of 418 copies (MY1) ranged in copy number from 396 to 431, and a strain with reported rDNA size of 70 copies (JU775) ranged in copy number from 63 to 81. Among all strains, the variation in copy number estimates within technical replicates ranged from 3% CV to 15%, with a median of 6% (CV=(standard deviation/mean)*100).

A key benefit to CHEF gel analysis is the comparison of samples to DNA molecules of known size, something other techniques lack, providing confidence in the resulting copy number estimates. Therefore, for the subsequent comparisons and quality assessments, we use the average value of the three replicate CHEF gels as the 'true' value of rDNA copy number in our samples and gauge our accuracy with respect to these values. In addition, the CHEF approach also reveals minor bands, if present, possibly representing copy number differences in a subset of the population (**Fig 2.1B**, ED3040 lane). Pooled amplification measurements reflect only average rDNA array size, therefore information about the potential existence of subpopulations or heterozygosity is lost in sequencing or PCR-based approaches.

To assess the accuracy of WGS with respect to CHEF gels, we used genomic DNA extracted from the worm samples to generate barcoded libraries by Nextera tagmentation and amplification. This method differed from the published study [1], which used sonication and adapter ligation. Two different DNA input amounts were tested: 10ng and 50ng genomic DNA. Three independent replicate libraries of each input amount were prepared from the same tube of genomic DNA. We sequenced these libraries with an Illumina NextSeq 500 and used read coverage of rDNA sequence relative to the whole genome to estimate rDNA copy number, an approach used elsewhere [1,2,94,215].

By this initial estimation method, we observed surprising variability in the rDNA copy number calculated for each library, despite their identical source material. Libraries of the strain JU775 (CHEF-based copy number of 70) reported copy numbers from 78 to 139. Libraries from the strain MY1 (CHEF-based copy number of 412) reported copy numbers from 303 to 603, a much less precise range than that produced by CHEF gel (**Table 2.1, Figure 2.1C**). The 50ng input

DNA samples were particularly distorted, with the percent variation from the CHEF estimate as much as 100% (calculated as 100 * absolute value of (estimated number – CHEF number)/CHEF number)). Even the 10ng input DNA samples varied from 0.6% to 52% (**Table A1.1**). On average, the 10ng and 50ng input amounts varied 18% and 58% in their copy number estimates, respectively (**Table A1.2**), an unacceptably high error for many applications.

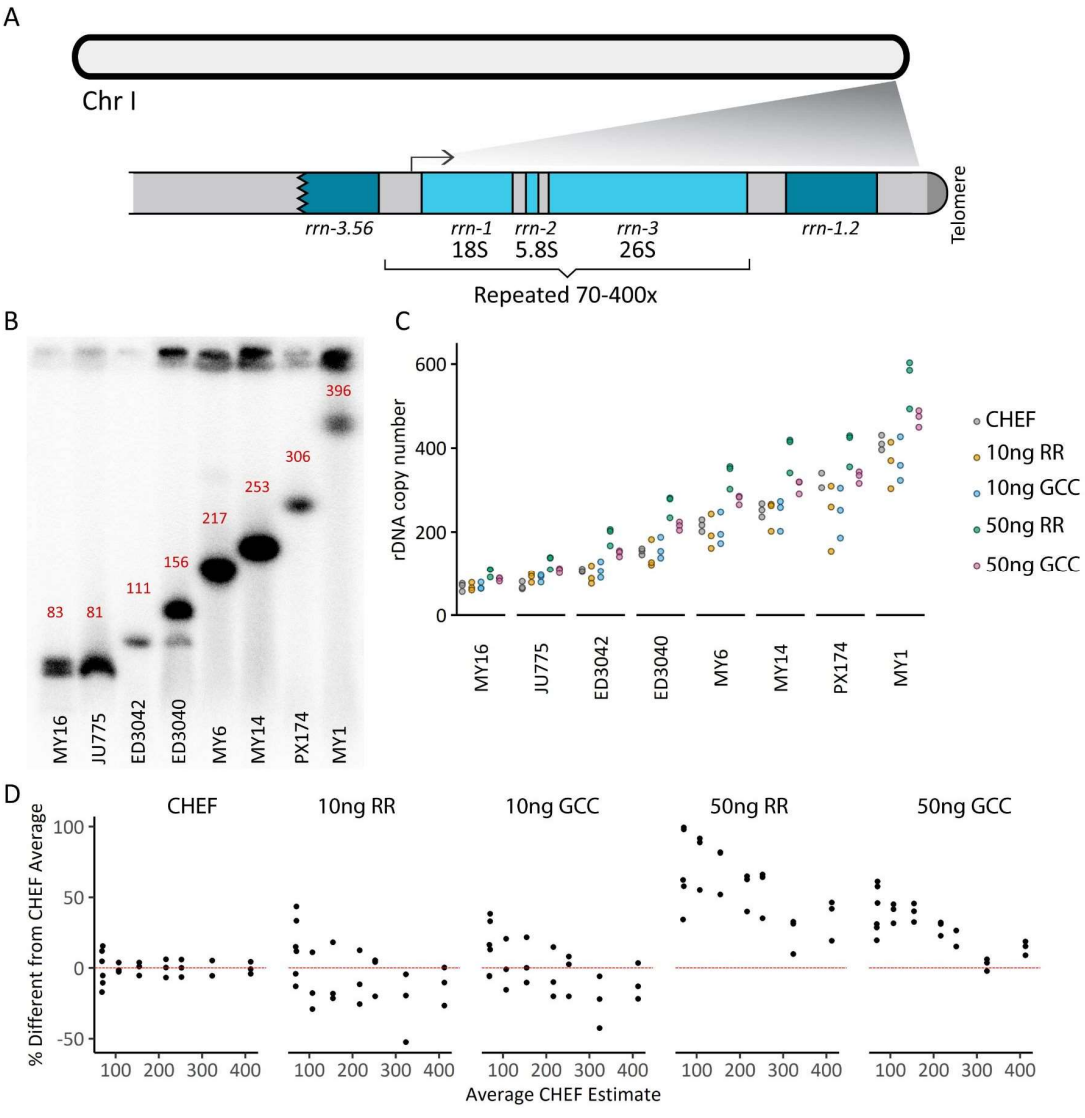

Figure 2.1: rDNA copy number estimation by CHEF gel and short read sequencing in *C. elegans*.

A: Schematic of the *C. elegans* rDNA locus, at the right arm of chromosome I. The 18S, 5.8S, and 26S rRNAs are transcribed as one unit and later processed into the three species. One repeat unit is 7.2kb in length and is tandemly repeated approximately 70 to 400 times, depending on the strain. The array is flanked by a partial copy of the 26S rRNA gene (*rrn-3.56*) and an additional copy of the 18S (*rrn-1.2*), which ends approximately 1kb upstream of the telomere. **B:** Southern blot against rDNA from a CHEF (contour-clamped homogeneous electric field) gel reveals rDNA copy number for eight wild isolates of *C. elegans*. Band size was measured relative to yeast chromosomal ladders visualized by ethidium bromide staining, and copy numbers calculated from the band size for each primary band are listed in red (also in Table 2.1). MY16 displays two bands of similar intensity, copy numbers 83 and 69; only the upper number is listed in the image. **C:** rDNA copy number (per haploid genome) was estimated for eight *C. elegans* wild isolates in five different ways: CHEF gel followed by Southern blot, 10ng input Nextera-based whole genome sequencing followed by relative read count coverage of the rDNA without or with GC content correction ("RR" and "GCC," respectively), and 50ng input whole genome sequencing with RR and GCC. Three replicates for each are plotted. **D:** Data from **C** is plotted as a percentage of the average CHEF-based values ((estimated number – CHEF number)/CHEF number)). The red line indicates 0% error.

Table 2.1. rDNA copy number estimates of *C. elegans* strains using CHEF gel or WGS.

		CHEF gel				Nextera Whole Genome Sequencing											
						10ng input						50ng input					
Strain	WGS[a]	Rep1	Rep2	Rep3	**Ave**	RR[c]	GCC[d]	RR	GCC	RR	GCC	RR	GCC	RR	GCC	RR	GCC
MY16	68	56	76[b]	71[b]	**68**	65	64	60	64	78	79	110	87	110	89	91	81
JU775	70	63	81	66	**70**	100	93	78	79	93	96	139	112	138	110	110	102
ED3042	105	104	111	106	**107**	88	106	76	90	119	129	206	155	202	151	167	141
ED3040	149	146	156	160	**154**	121	154	127	138	182	187	281	224	279	216	234	204
MY6	193	201	217	229	**216**	191	194	161	173	243	248	356	285	351	283	302	265
MY14	237	236	253	267	**252**	266	259	202	202	263	273	419	319	415	319	341	291
PX174	298	-	306	340	**323**	260	252	154	186	309	304	429	343	425	335	355	316
MY1	418	410	396	431	**412**	370	359	303	323	414	427	603	489	585	475	493	449

a rDNA copy number reported by WGS with sonication and adapter ligation in Thompson *et al.* 2013, *Genome Res.*

b Average of two bands

c Relative read count coverage based rDNA estimate

d GC-content corrected rDNA estimate

2.3.2 Computational correction can improve WGS-based copy number estimates

We asked what computational measures could be taken to improve accuracy of the WGS data. We examined factors including GC content bias, total read coverage, input amount, and single copy region coverage (**Table A1.1**).

First, we attempted to correct for tagmentation enzyme bias by implementing a maximum likelihood estimation GC-content correction method [97,261]. This approach corrects for GC content bias and read coverage in a sample-specific manner. Implementation of this method improved copy number estimates, but error ranges remained high: 0.2% to 42% (average 15%) in the case of 10ng input, and 2% to 61% (average 29%) in the case of 50ng input (**Table A1.1, A1.2, Figure 2.1CD**). Of the conditions tested here, lower input amount (10ng) and GC-content correction brought copy number estimates generally closest to CHEF-based values (**Table 2.1, Figure 2.1D**). Although exhibiting greater deviation from the CHEF-based copy number, the combination of higher input amount (50ng) and GC-content correction produced the highest degree of reproducibility (**Figure 2.1D, Table 2.1**). This observation suggests a potential vehicle to computationally correct WGS-based copy number estimates, assuming benchmark samples of CHEF-based copy number have been included. The high reproducibility observed in the 50ng input condition could also represent an artifact of closely spaced library preparation, as two of the three samples under the 50ng input condition were prepared on sequential days (date of library preparation provided in Table A1.1).

Insufficient read coverage seemed another highly plausible source for inaccuracy of WGS-based copy number estimates. However, we did not observe a correlation between the total reads a sample received and its accuracy in copy number call. Pearson's correlation between total

aligned reads and the error of the WGS-based rDNA estimates was 0.06 and 0.12 for the original and GC-corrected copy number estimates, respectively, both non-significant (**Table A1.1**). We furthermore performed an in silico downsampling experiment with published data from strains MY1 and JU775 [1]. Repeated downsampling to a randomly drawn 5% of the initial reads introduced small differences in copy number calls of up to 0.9% or 2.4%, for MY1 and JU775, respectively (**Table A1.3**). This amount of variation was far below the variation we see among library preps of comparable coverage, suggesting coverage is not the major contributor to error in copy number calls.

Changes to other WGS processing measures failed to improve copy number estimates. While still employing GC-content correction, we compared WGS-based estimates using different alignment algorithms, to little effect. BWA-MEM alignment gave an average error of 15% and 23% for 10ng and 50ng input, respectively. This degree of error compares to the Bowtie 2 alignment used above, with average error of 15% and 29% (**Table A1.4**). Similarly, comparing alignment with and without read adapter trimming gave very similar estimates: 15% and 27% average error with adapter trimming compared to 15% and 29% without, for 10ng and 50ng input, respectively (**Table A1.4**).

We asked if samples that were miscalling rDNA copy number were also distorted in their copy number estimates of other regions. To do so, we examined twenty-nine 7.2kb regions of the genome, with representatives from all six chromosomes that should be present at single copy (**Table A1.5**). We estimated the copy number for each of these regions as the rDNA locus copy number was calculated, using read coverage of each region relative to the whole genome (**Table A1.6**). The average of these estimates for each strain ranged from 0.81 to 1.44 and was correlated

with WGS-based rDNA copy number (Pearson's value of 0.93), indicating that some samples were indeed prone to inflated copy number calls of not just the rDNA locus (**Table A1.1**).

2.3.3 A panel of yeast strains confirms high error in WGS-based copy number estimates

The yeast *S. cerevisiae* has arguably the best characterized rDNA locus of any eukaryote [91,262–264]. Unlike many plants and animals, *S. cerevisiae* contains the genes for all four rRNA species in one array, on chromosome XII (**Figure 2.2A**). As in other eukaryotes, this array is highly repetitive, with the laboratory strain BY4741 containing approximately 150 copies of the 9.1 kb repeat unit [156]. With a tractably-sized genome of ~12Mb and a single rDNA locus, *S. cerevisiae* presents another excellent choice for WGS-based copy number estimation.

WGS data is available for many wild isolates of *S. cerevisiae*, from which rDNA copy numbers have been reported [3]. We selected a panel of 28 of these strains with reported rDNA copy numbers ranging from 17.5 to 277 per haploid genome (**Table 2.2, Table A1.7, Table A1.8**)—16 of these were annotated haploid strains and 12 diploid. We quantified rDNA copy number in these strains by CHEF gel analysis as well as by whole genome sequencing on a 14-strain subset of the haploids.

CHEF gel analysis revealed copy numbers wildly discrepant from the published WGS-based copy numbers (**Figure 2.2B, C, Figure A1.1**). CHEF-based rDNA copy numbers varied as much as 143 copies and as little as 1 copy from the reported WGS-based numbers (**Table 2.2**). As before, we ran triplicate CHEF gels to assess the reproducibility of this technique. Variation in copy number calls among CHEF technical replicates ranged from 0.6% CV to 5.6% (average of 1.8%), consistent with the known high precision of this method.

We furthermore performed CHEF analysis on multiple single colony isolates of the same strain, to examine the possibility that colony-to-colony variability explains differences between WGS- and CHEF-based copy number estimates (**Figure A1.2**). We performed CHEF gel analysis of three separate colony isolates as well as a mixed population for 10 haploid strains and 8 diploid strains. Most strains with discrepant CHEF/WGS estimations displayed no rDNA heterogeneity. Of the four haploid strains that did exhibit different rDNA copy numbers between colonies, two strains had a single colony whose CHEF-based copy number estimate was more similar to the WGS. For the other two strains, the "population" CHEF-based copy number estimate was in agreement with the WGS-based estimate. Although we did not comprehensively study rDNA copy number heterogeneity among colonies, the presented data do not implicate such heterogeneities in the large discrepancies between CHEF- and WGS-based estimates.

Fourteen of the above strains were further assessed by whole genome sequencing, with two goals: 1) to verify the genotype of the given strains to eliminate the possibility that disparate copy numbers could be due to strain mislabeling, and 2) to assess how well WGS-based copy number estimates of yeast strains compare to CHEF-based results when both are performed in our lab. We generated yeast genomic DNA libraries of these 14 strains by Nextera tagmentation and amplification. WGS SNP analysis confirmed that genotype labels were correct for 11 of these 14 strains (**Table A1.9, Figure 2.2C**). BY4741 and B30 were included as known copy number controls. The resulting rDNA copy number estimates showed closer agreement with our CHEF-based results than the previously reported sequencing data, but were still highly divergent. Deviation from the CHEF-based copy numbers ranged from 2% to 38% (average 18%) when

corrected for GC-content (**Table 2.2**). Overall, our yeast results reproduce the unreliability of rDNA copy number calls by Nextera WGS that we observed in *C. elegans*.

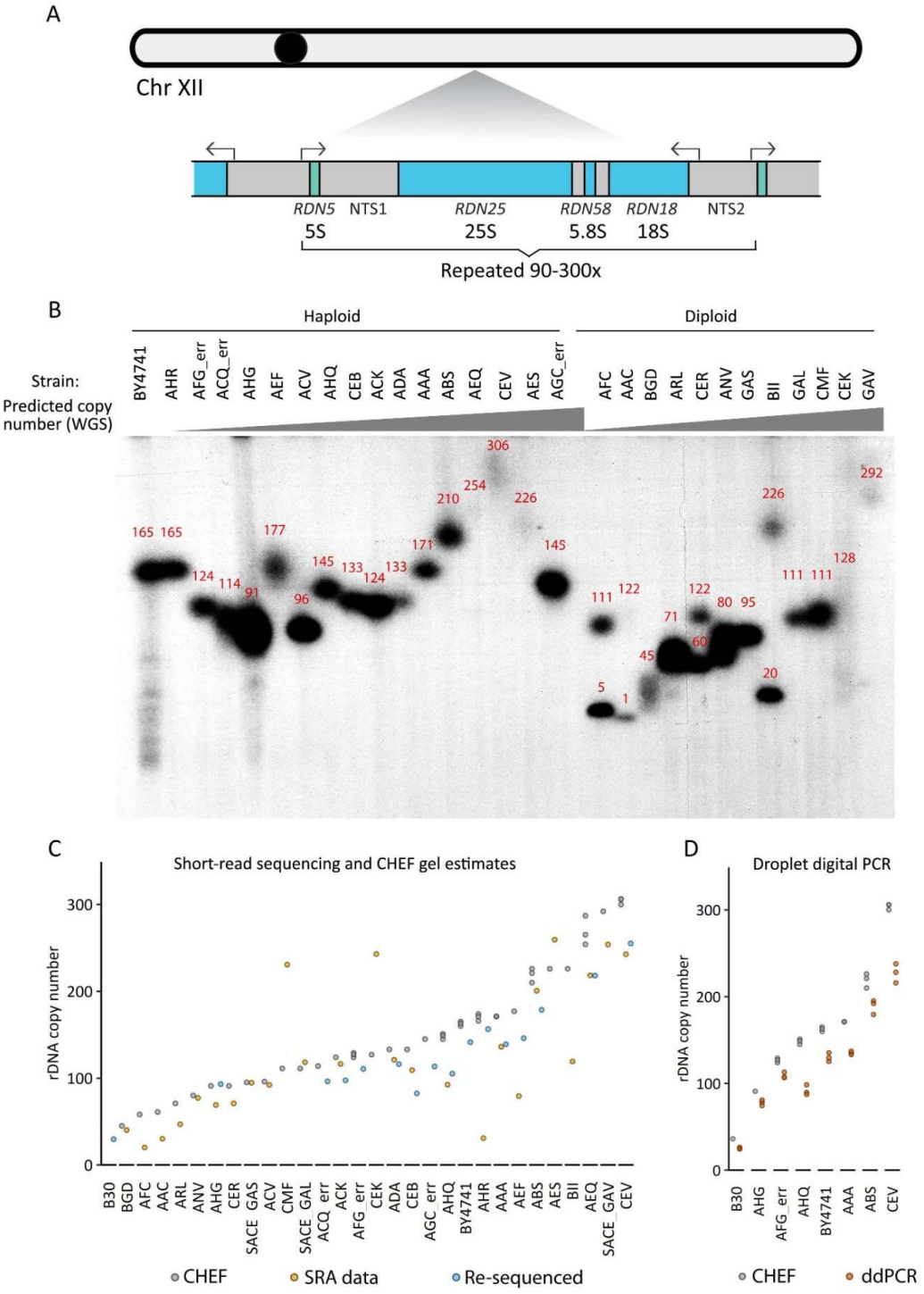
A
Chr XII
RDN5
NTS1
RDN25
RDN58
RDN18
NTS2
5S
25S
5.8S
18S
Repeated 90-300x
B
Haploid
Diploid
Strain:
Predicted copy number (WGS)
BY4741
AHR
AFG_err
ACQ_err
AHG
AEF
ACV
AHQ
CEB
ACK
ADA
AAA
ABS
AEQ
CEV
AES
AGC_err
AFC
AAC
BGD
ARL
CER
ANV
GAS
BII
GAL
CMF
CEK
GAV
C
Short-read sequencing and CHEF gel estimates
rDNA copy number
CHEF
SRA data
Re-sequenced
D
Droplet digital PCR
rDNA copy number
CHEF
ddPCR

Figure 2.2: rDNA copy number estimation by CHEF gel, short read sequencing, and ddPCR in *S. cerevisiae*.

A: Schematic of the *S. cerevisiae* rDNA locus on chromosome XII. The 18S, 5.8S, and 25S are transcribed as one unit by PolI. 5S rRNA is transcribed as a separate unit by PolIII. The repeating unit is 9.1kb in length and in haploid strains is tandemly repeated approximately 90 to 300 times. **B:** Southern blot for chromosome XII from a CHEF gel reveals size variation of *S. cerevisiae* strains from the 1002 genomes project [3]. rDNA copy numbers measured from each band are noted in red, calculated from band size relative to yeast chromosomal ladders. Haploid and diploid strains are noted. Wild haploid yeast strains show copy numbers ranging from 91-306. Wild diploid strains show individual bands with copy numbers ranging from 45-292 (note that two strains, AFC and AAC had bands at the lower limit of our ability to quantify, estimated as 5 and 1 copies, respectively). From left to right, strains were arranged in order of the expected rDNA copy number, based on previously reported whole genome sequencing estimates. **C:** rDNA copy numbers for 30 strains of *S. cerevisiae* isolates are plotted, reflecting CHEF-based rDNA copy number estimates ("CHEF", dark blue), previously reported data ("SRA data", green), or our own Nextera whole genome sequencing-based estimate ("Re-sequenced", mauve, done for 14 strains plus B30 and BY4741 controls). The annotation "err" indicates strains whose re-sequenced genotype did not match SRA library genotypes. The SRA-based estimate has therefore been omitted from the graph for these strains. **D:** Droplet digital PCR estimations (in triplicate) of rDNA copy number for eight yeast strains are plotted next to their CHEF-based rDNA copy numbers.

Table 2.2. rDNA copy number estimates of *S. cerevisiae* haploid strains.

			CHEF gel				ddPCR		
Strain	SRA Number	WGS[a]	Rep1	Rep2	Rep3	Re-sequenced[b]	Rep1	Rep2	Rep3
BY4741	-	-	165	160	163	141	126	135	129
B30	-	-				30	25	26	25
AHR	ERR1308770	31	166	174	171	157	-	-	-
AFG_err[c]	ERR1309145	50	124	129	127	111	113	107	107
ACQ_err	ERR1308618	74	114	-	-	96	-	-	-
AHG	ERR1308781	69	91	-	-	93	75	78	81
AEF	ERR1308873	79	177	-	-	146	-	-	-
ACV	ERR1308596	92	96	-	-	-	-	-	-
AHQ	ERR1309019	93	145	151	149	105	90	98	87
CEB	ERR1309017	109	133	-	-	82	-	-	-
ACK	ERR1308893	116	124	-	-	97	-	-	-
ADA	ERR1309427	121	133	-	-	116	-	-	-
AAA	ERR1309487	136	171	169	171	139	135	133	137
ABS	ERR1309033	201	210	226	221	179	195	192	180
AEQ	ERR1309512	218	254	287	265	218	-	-	-
CEV	ERR1308745	243	306	306	300	255	238	216	228
AES	ERR1309368	259	226	-	-	-	-	-	-
AGC_err	ERR1309000	275	145	-	-	114	-	-	-

[a] rDNA copy number estimated from GCC-analyzed WGS data from Peter *et al.* 2018 *Nature*

[b] rDNA copy number estimated from GCC-analyzed WGS data from this study

[c] err notation: re-sequencing data indicates strain genotype did not match SRA data genotype

2.3.4 Droplet digital PCR of yeast isolates

ddPCR has been proposed as a more precise alternative to qPCR [257]. ddPCR involves independent amplification of single molecules of target DNA, informing a Poisson distribution-based calculation of the absolute number of starting molecules [257,265]. ddPCR has been employed in rDNA copy number quantification with reported high success and has recently been used to estimate rDNA copy number in both human and yeast samples [94,154,266]. We wanted to assess ddPCR reproducibility and determine if ddPCR rDNA copy number estimates agree with CHEF-based estimates for a subset of *S. cerevisiae* wild isolate strains. We used ddPCR to quantify rDNA copy number in six strains (**Table 2.2, Table A1.10**), as well as the laboratory strain BY4741 (170 rDNA copies) and a low rDNA copy number strain, B30 (35 rDNA copies). BY4741 and B30 both have CHEF-based rDNA copy number estimates (**Figure 2.2, Figure A1.2**). Using published primer and probe sequences to target the rDNA [154], we optimized ddPCR conditions by testing and altering annealing temperature (**Figure A1.1**). We quantified both the number of rDNA copies and a single copy gene, *TUB1*, in the same ddPCR reaction and estimated copy number using Quantasoft software.

Estimates of rDNA copy number by ddPCR differed from CHEF-based estimates by 11% to 41% (average 22%)(**Figure 2.2D, Table 2.2**). For all samples, ddPCR underestimated the copy number. The ddPCR calculation provides confidence intervals based on the Poisson distribution. For the three replicates of a given strain, the estimates did not always fall inside each other's confidence intervals. Overall, ddPCR was nearly as technically reproducible as CHEF gel analysis; however, it yielded different, lower copy number estimates (CHEF- and ddPCR-based estimates were significantly different by t-test for all samples, ranging in p-value from 0.0006 to 0.011).

2.4 DISCUSSION

All scientific measurements have associated technical error. The strengths and weaknesses of different methods make them more or less suitable for a given task. The alarming discrepancy among multiple WGS-based copy number calls on a single genomic sample suggests that copy number data obtained from Nextera-enabled whole genome sequencing may not be accurate. Nevertheless, accurate knowledge of rDNA copy number variation has the potential to advance our understanding of complex traits including aging and cancer.

From our observations presented here, we propose several best practices to obtain the highest accuracy for rDNA copy number estimates. 1) If possible in a system, pulsed-field gel electrophoresis techniques should be employed to validate WGS-based copy numbers. 2) When applying whole genome sequencing, samples of validated copy number should be included in library preparations, to control for batch variation. 3) When applying whole genome sequencing or ddPCR, samples should only be compared that have been prepared and run together. 4) In the estimation of copy number from WGS data, GC-content computational correction should be performed [261]. To this end, our study presents a resource of yeast and worm strains with CHEF-gel verified rDNA copy number.

Historically, many methods not evaluated here have been used to quantify rDNA copy number, including 1) quantitative rRNA hybridization to total DNA [267], to microarrays [268,269], or *in situ* to chromosomes [70]; 2) quantitative Southern blot [143,157]; and 3) cytogenetic observations [270]. Similarly, we did not discuss or explore more recent methods such as optical mapping [235]. However, much of recent reporting of rDNA copy number variation has relied on estimation from short-read whole genome sequencing

[1,2,94,97,215,258]. WGS-based copy number estimation is an attractive approach because of the vastness of existing data sets and the ease with which new data can be generated. Our observations, however, suggest that caution should be exercised when making conclusions from WGS data, especially when comparing libraries prepared at different times or by different facilities, even if using the same library preparation methods. Indeed, others have also observed library preparation batch drastically affecting copy number estimation [129,271].

In the future, new technologies for estimating rDNA copy number may be available, including long-read sequencing such as developed by Oxford Nanopore, which can infrequently obtain read lengths of over 1Mb, with a record reported length of 2.27Mb [272]. It should be noted that this read was reconstructed from eleven reads generated consecutively that mapped to a 2.27Mb locus [272], a gambit that may or may not be helpful for the case of long repetitive DNA, depending on the amount of sequence variation present among repeat units. Although researchers are working on extending the Nanopore readable length to make rDNA array size estimation practical, they face the likely obstacle of finding DNA extraction techniques gentle enough to recover fragments large enough to span the rDNA [273]. Recent long-read sequencing of *C. elegans* using both PacBio and Nanopore failed to determine copy number of either the 45S or 5S rDNA repeats [233]. For a *C. elegans* strain of 400 rDNA copies, covering the entire array would require reads of nearly 3Mb. Application in humans, with a single unit size of 43kb, will likely need read lengths upwards of 6Mb [132].

From the biomedical perspective, human rDNA copy number is of substantial interest, and error in rDNA copy number calls impairs our ability to incorporate this potentially valuable source of variation into association studies. It is likely that the problem of human rDNA copy

number determination is even more challenging than our results here suggest, as the model organisms presented have only a single rDNA locus and are thus the simplest cases. Human 45S rDNA loci are spread across five chromosomes, making validation by CHEF gel less feasible due to the difficulty of interpreting multiple gel bands, some of which may be too large to easily resolve on a pulsed-field gel [132]. Development of novel single-cell approaches such as FISH may prove an alternative way to quantify copy number in humans, with the benefit of potentially providing cell and tissue level information.

One caveat of our study is the necessity of designating a 'true' value for rDNA copy number. Herein, we have used CHEF-based estimates as our measure for comparisons. We do, however, acknowledge the possibility that CHEF-based measurements may be subject to electrophoretic artifacts. However, previous reports find either no or only minimal effect of GC content or repetitiveness on mobility in an agarose gel [274–277], giving us confidence in the accuracy of CHEF-based copy number estimates.

The observed difficulties in copy number estimation of the rDNA locus likely extend to other repetitive genomic loci, making this problem of relevance not just to rDNA but to the genome as a whole. It is our hope that the recommendations we outline will advance our ability to characterize this type of variation and promote its eventual incorporation into our understanding of biology.

2.5 METHODS

Strains and growth conditions

C. elegans

Wild isolates of *C. elegans* were kindly provided by the Moerman lab (strains MY1, PX174, JU775, MY16, MY14, MY6, ED3040, and ED3042). Worms were maintained on Nematode Growth Medium seeded with OP50 bacteria for standard maintenance.

S. cerevisiae

Strains from the "1,011 *S. cerevisiae* isolates" collection [3] were generously provided by the Dunham Lab. The laboratory strain BY4741 and the strain with 35 rDNA copies were obtained from the Brewer Lab.

Growth for paired CHEF gel and genomic DNA preparation

C. elegans

C. elegans strains were grown to starvation on 15cm high-peptone (20g/L) NGM NA22 plates, enriching for arrested L1s. Worms were washed from the plates with 15mL M9 and centrifuged 450xg 2min. Supernatant was removed and the pellet was washed in 8mL autoclaved, glass-distilled water and spun again 450xg 2min. Approximately 75µL of worm pellet was added to 200µL ATL buffer (Qiagen DNeasy kit 69504) in a 1.5mL tube and stored at -20°C for eventual genomic DNA extraction and whole genome sequencing (see below). 80µL of the same pellet was embedded in agarose plugs for CHEF gel application (see below).

S. cerevisiae

Strains were streaked out on YPD plates and incubated for 2 days at 30°C. Cells from the patch population and 3 individual colonies were then separately inoculated in 5mL synthetic

complete liquid media buffered with succinic acid and grown for 2 days at 30°C. Cells were then collected into 1.5mL tubes, pelleted, washed once with 50mM EDTA, and stored as dry pellets at -20°C until either preparation of CHEF gel plugs or genomic DNA extraction.

CHEF gel sample preparation and run conditions

C. elegans

720µL of melted 42°C 0.5% SeaPlaque GTG agarose was added to an 80µL worm pellet described above. ~80µL of this suspension was placed in agarose plug molds (Bio-Rad #1703713), on ice, and allowed to solidify at 4°C for at least 30min. Plugs were extracted into 2mL tubes, to which was added 300µL TEL [9mM Tris, 90mM EDTA pH 8, 10mM levamisole], followed by incubation on ice 30min. The levamisole was found to be necessary to prevent the worms from crawling out of the plug during lysis. The supernatant was removed from the tube, being careful to avoid damaging the plugs, and replaced with 300µL lysis buffer [1% SDS, 1mg/mL Proteinase K (Sigma-Aldrich P4850), 8mM Tris, 80mM EDTA pH 8, 1mM levamisole]. Tubes were put at 50°C for ~24 hours to allow lysis of the worms. L1 stage worms are reported to have a weak enough cuticle for in-plug digestion. After digest, plugs were decanted into 24-well plates (two plugs per well). Supernatant was removed, the plugs were rinsed once with 300µL TE [10mM Tris, 1mM EDTA], new TE was added and the plugs were rocked at room temperature 2hr. Supernatant was removed and replaced with fresh TE and rocked overnight at 4°C. In total, eight of these TE washes were performed, with at least one taking place at 4°C overnight. After the last wash, plates were stored in TE and placed at 4°C until use.

Plugs were prepared for CHEF gel by digestion with an enzyme that cuts in chromosome I but not within the rDNA. We found SwaI to be the best restriction enzyme for this purpose; SwaI

cuts 3927bp upstream of rrn-3.56. Plugs were soaked in 1X NEB 3.1 buffer for 30min, buffer was replaced with fresh 1X NEB 3.1, and plugs were soaked another 1hr. This incubation was done without rocking, with the plate on ice. Plugs were then transferred out of buffer to a parafilm-wrapped slide and 4µL of SwaI was added to the top of the plug. The plugs were placed in a container with a wet paper towel and incubated at 25°C for 4hr before CHEF gel loading. To load, plugs were transferred to the teeth of a gel comb, and 1% agarose in 0.5X TBE was poured around them, careful not to dislodge. Once the gel solidified, it was placed in a CHEF gel box (Bio-Rad CHEF DRII), where the gel was run at 100V for 68hr, 14°C, switch times = 300-900 seconds. Replicates 2 and 3 were digested with SwaI, but replicate 1 used MfeI (digests 1057bp upstream of rrn-3.56); we found MfeI to have more off-target cutting (more degradation of rDNA) than SwaI. Ladders of two different ranges were included: *S. cerevisiae* chromosomes were used as one ladder (maximum size 2.5 Mb), and *Hansenula wingei* chromosomes as another ladder (maximum size 3.13 Mb, Bio-Rad 170-3667). After the gel had been run, it was soaked in ethidium bromide (0.3µg/mL in 0.5X TBE) to visualize the ladders.

See Appendix 1 for Southern blotting and probe conditions.

S. cerevisiae

S. cerevisiae genomic DNA plugs were prepared as previously published [278]. 1% low-melt agarose (Lonza SeaPlaque GTG agarose) in 50mM EDTA was melted and cooled in a 45°C water bath for 10 minutes. Approximately 10^8 cells (0.5mL of 2-day culture, frozen), were transferred to a 1.5mL tube and resuspended in 100µL 50mL EDTA pH 8.0. Cells were then mixed with 100µL 1% low-melt agarose, transferred to agarose plug molds (Bio-Rad #1703713, 2 plugs generated for each sample), and incubated at 4°C for 15 minutes to solidify agarose. Plugs from

a single strain were then transferred to a single well in a 24-well plate containing 1mL spheroplasting solution (1M Sorbitol, 20mM EDTA, 10mM Tris-HCl pH 7.5, 14mM β-mercaptoethanol, 0.5mg/mL Zymolyase 20-T), and incubated for 4 hours in a 37°C incubator with periodic agitation. Spheroplasting solution was then removed, plugs were washed 1x15 minutes with 1mL LDS solution (1% lithium dodecyl sulfate, 100mM EDTA, 10mM Tris-HCl, pH 8.0), and then incubated overnight in a 37°C incubator with 1mL fresh LDS solution. In the morning, plugs were washed 3 times for 20 minutes with 0.2X NDS solution (1X NDS: 0.5M EDTA, 10mM Tris base, 1% Sarkosyl, pH 9.5), and then 5+ times with TE pH 8.0. Plugs were then stored in TE pH 8.0 until use.

For CHEF gels examining the sizes of Chromosome XII, which contains the rDNA: 200mL of 0.8% LE agarose in 0.5X TBE was melted and then cooled in a 50°C water bath for 10 minutes. A slice of an agarose plug (~2mm) from each strain was transferred to a separate tooth on a gel-comb and excess moisture was wicked away using a Kimwipe. An *H. wingei* standard ladder (Bio-Rad 170-3667) sample was included on the comb on a separate tooth. The gel-comb with plug slices was then positioned in a Bio-Rad CHEF gel-casting tray and embedded in the prepared 0.8% LE agarose. Once the gel was solidified, the gel-comb was removed and the gel was then transferred to a CHEF gel module (Bio-Rad CHEF-DRII) containing 2.3L 0.5X TBE continuously cooled to 14°C. Yeast Chromosome XII CHEF gels were run for 66 hours at 100V, switch times = 300-900 seconds.

For CHEF gels examining the size of *S. cerevisiae* rDNA arrays excised from Chromosome XII: The *S. cerevisiae* rDNA array contains no BamHI restriction sites; the nearest BamHI sites are 8.8kb from the centromere proximal edge and 30.9kb from the telomere proximal edge. To digest

the rDNA array away from Chromosome XII, agarose plugs were washed for 20 minutes three times in 1mL 1X NEB Buffer 4 + 1X BSA. Two 2mm slices from an agarose plug were then transferred to a dry 24-well plate, plug slices covered with 50µL BamHI-HF restriction digest reaction (1X NEB Buffer 4, 1X BSA, 1.3µL BamHI-HF), and incubated in a 37°C incubator for 5 hours. 1mL TE pH 8.0 was then added to each well to facilitate plug slice handling. Each of the 2 plug slices were run on two separate CHEF gels: the "High Molecular Weight" protocol described above (100V, 66 hours, 300-900 seconds), which resolves fragments between 1Mb-3.13 Mb, and the "Low Molecular Weight" (165V, 66 hours, 47-170 seconds), which resolves fragments between 225kb-1.1Mb. Both BamHI-digested sample CHEF gels were cast as described above for the uncut CHEF gel in 0.8% LE agarose with the appropriate standard ladder: the Bio-Rad *H. wingei* for the "High Molecular Weight" CHEF gel run conditions and the NEB Yeast Chromosome PFG Marker for the "yeast full chromosome" run.

Southern blotting was performed as described for the *C. elegans* samples (see Supplemental Data), using yeast-specific probes. For the uncut CHEF gels, a single copy sequence on Chromosome XII (*CDC45*) was used as a probe for Chromosome XII location. For BamHI-digested CHEF gel samples, the *S. cerevisiae* rDNA NTS2 sequence was used as a probe for the location of the excised rDNA array.

Droplet digital PCR

S. cerevisiae genomic DNA was diluted to 0.05 ng/ul in low-bind tubes. Each 20µl reaction consisted of 10 µl 2X ddPCRTM Supermix for Probes (Bio-rad), 0.125µl EcoRI-HF (NEB, 20,000 U/mL), 1.8µl of 10µM Primer Mix (containing 10µM each rDNA F and R primers and Tub1 F and R primers), 1µl of 5µM Probe mix (containing 5µM each rDNA and Tub1 probes), and 1µl of DNA

at 0.005 ng/µl. The mixture was incubated for 15 minutes for DNA digestion to occur, followed by droplet generation on a QX200 Droplet Generator (Bio-rad). Amplification was performed for 50 cycles with a 57°C annealing temperature, and droplet reading was performed on a Bio-rad QX200 Droplet Reader. Optimal annealing temperature was determined by a temperature gradient of 56-62°C with BY4741 DNA (**Figure A1.2**).

2.6 DATA AVAILABILITY

FASTQs of whole genome sequencing are available at SRA (PRJNA565452 for *C. elegans* data, PRJNA573925 for *S. cerevisiae* data). Supplemental Data and large Supplemental Tables are enumerated in the supplemental materials section and available as separate files.

2.7 PROJECT ACKNOWLEDGEMENTS

We would like to thank the Moerman lab for kindly providing wild worm isolates, and the Dunham lab for kindly providing yeast strains. We would like to acknowledge members of the Queitsch, Waterston, and Brewer/Raghuraman labs for helpful discussion.

2.8 PROJECT CONTRIBUTIONS

This study was initiated and designed by (authors). The specific roles of ANH were as follows: WGS data analysis (read depth and GC correction methods of rDNA copy number estimation, and yeast genotyping), preparation of yeast WGS libraries, ddPCR. Additional data are included in the publication that are not included in this dissertation as ANH was not the major contributor.

CHAPTER 3. THOUSANDS OF HIGH-QUALITY SEQUENCING SAMPLES FAIL TO SHOW MEANINGFUL CORRELATION BETWEEN 5S AND 45S RIBOSOMAL DNA ARRAYS IN HUMANS

3.1 SUMMARY

The ribosomal DNA genes (rDNA) are tandemly arrayed in most eukaryotes and exhibit vast copy number variation. There is growing interest in integrating this variation into genotype-phenotype associations. Here, we explored a possible association of rDNA copy number variation with autism spectrum disorder and found no difference between probands and unaffected siblings. Because short-read sequencing estimates of rDNA copy number are error prone, we sought to validate our 45S estimates. Previous studies reported tightly correlated, concerted copy number variation between the 45S and 5S arrays, which should enable the validation of 45S copy number estimates with pulsed-field gel-verified 5S copy numbers. Here, we show that the previously reported strong concerted copy number variation may be an artifact of variable data quality in the earlier published 1000 Genomes Project sequences. We failed to detect a meaningful correlation between 45S and 5S copy numbers in thousands of samples from the high-coverage Simons Simplex Collection dataset as well as in the recent high-coverage 1000 Genomes Project sequences. Our findings illustrate the challenge of genotyping repetitive DNA regions accurately and call into question the accuracy of recently published studies of rDNA copy number variation in cancer that relied on diverse publicly available resources for sequence data.

3.2 INTRODUCTION

The genes encoding the ribosomal RNAs are present in long tandem arrays in most eukaryotes, often referred to as the ribosomal DNA (rDNA). Most eukaryotic genomes contain two types of rDNA arrays: the 45S, encoding the 18S, 5.8S, and 28S rRNAs, and the 5S, encoding the 5S rRNA. Because of their repetitive nature, both rDNA arrays are susceptible to expansion and contraction, which can lead to vast copy number differences among individuals [2,97]. The phenotypic consequences of natural variation in rDNA copy number remain largely unexplored [209], although a number of human disease phenotypes have been linked to rDNA copy number.

Changes in rDNA copy number have been identified in some cancer and aging studies. The 45S and 5S copy numbers change in some human cancers, in which cancerous tissue has a higher or lower rDNA copy number than a matched control [94,213,215–217]. In some tissues such as the brain, heart, and adipose tissue, 45S rDNA copy loss has been observed with age [202,203,279]. However, one study argues that rDNA instability is only observed in brains of individuals affected by dementia, and is absent in aging brains of healthy individuals [252]. In contrast, a recent study on aging mice found that 45S rDNA copy number increases in the blood with age in mice [205]. Additional studies in mice and rat cell lines report a lack of change in rDNA copy number with age [185,204,280]. In both cancer and aging studies, not all cancers and not all aging tissues or organisms appear to display changes in rDNA copy number [2,94,212,281]. In short, there is no universal agreement on whether rDNA copy number changes with age or cancer.

In addition, an individual's inherited rDNA copy number may be of interest for GWAS studies. In humans, rDNA copy number is associated with differences in global gene expression

and mitochondrial abundance [2]. These differences may modify the effects of trait-associated variation. There are a few studies that report potential effects of rDNA copy number on disease: Higher 45S rDNA copy numbers are found in people with schizophrenia and Alzheimer's disease [222,224,282,283], and higher copy number of the 18S and 5.8S rDNA genes from peripheral blood predicts increased severity of lung cancer. If accurate rDNA copy number estimates were more readily available, integrating rDNA copy number into future genotype-phenotype analyses would be of great interest.

Moreover, the 45S and 5S arrays are reported to covary in copy number in mouse and human, an effect termed concerted copy number variation [258]. This finding was interpreted as evidence for natural selection maintaining balanced 45S and 5S copy numbers to ensure proper rRNA dosage. The reported strong concerted copy number variation implies that the repetitive rDNA arrays undergo compensatory contractions or expansions across several distant genomic loci through yet undiscovered molecular mechanisms.

Although the mechanisms underlying the balanced dosage of the 45S and 5S rRNAs are not fully understood, it is well appreciated that rRNA expression is not primarily a function of rDNA copy number [18]. In exponentially growing human and yeast cells, about half of the 45S rDNA copies are epigenetically silenced [284,285]. It is unknown whether a similar proportion of 5S copies are silenced in mammals; however, 5S silencing is demonstrated in *A. thaliana* and *Xenopus laevis* [286]. In yeast, where the 45S and 5S are encoded in the same array, severe reduction of rDNA copy number results in all rDNA copies being expressed [18]. Similarly, mutations interfering with epigenetic silencing cause the expression of additional 45S rDNA copies in mammals (or of the singular repeat in yeast) [287,288]. Together, these data suggest

that rRNA dosage is largely controlled through transcriptional regulation rather than by changing rDNA copy number.

However, in addition to being a potentially interesting biological phenomenon, concerted copy number variation holds promise for confirming the quality of rDNA copy number estimates. A key limit to incorporating rDNA copy number into genotype-phenotype associations is the ability to accurately quantify these genotypes. Because rDNA copy number estimates by short-read sequencing are error-prone due to batch effects [5], sequencing-based estimates should be validated by methods such as contour-clamped homogeneous electric field gel electrophoresis (CHEF gels), which can separate megabase-sized DNA fragments. The human 45S arrays are too large and too numerous to be accurately quantified by CHEF gels, however the 5S can be readily measured [126]. Combining 5S estimates by CHEF gels with a linear model relating 5S and 45S array copy numbers from sequencing data could permit the assessment of the accuracy of 45S rDNA copy number genotypes. Here, we sought to develop and use this method in order to facilitate the inclusion of 45S rDNA copy numbers in genotype-phenotype association studies for autism spectrum disorder.

We analyzed concerted copy number variation between the 45S and 5S rDNA arrays in multiple datasets covering several thousand samples. The first data set encompasses the 163 samples of the low-coverage 1000 Genomes Project in which concerted copy number variation was previously reported. We expand on this dataset by including two far larger datasets of newly generated high quality, high-coverage sequencing data for both the 1000 Genomes Project and the Simons Simplex Collection. All sequencing samples for the high-coverage 1000 Genomes Project and the Simons Simplex Collection were generated by the New York Genome Center with

a single sequencing method. In contrast, the original, low-coverage 1000 Genomes Project data were generated in multiple genome centers using various methods. We estimated rDNA copy number by short read sequencing read depth in all three datasets.

Unlike previously reported, we observe only weak correlations between the 45S and 5S rDNA arrays in the Simons Simplex Collection and high-coverage 1000 Genomes Project collection. We confirm that our analysis pipeline identifies the previously reported strong concerted copy number variation in the low-coverage 1000 Genomes Project data, and we show that concerted copy number variation is far weaker in the high-coverage data for the same samples. Furthermore, we show that copy number estimates between high and low-coverage data correlate poorly. The weak correlation between the 45S and 5S rDNA arrays is not an artifact of cell passaging between the initial 1000 Genomes Project sampling and the recent high-coverage resequencing effort because we observed the same result in the Simons Simplex Collection samples derived from blood. We recently reported that rDNA copy number estimation from whole-genome short-read sequencing data is sensitive to even subtle variation in sample processing and coverage in technical replicates [5]. Our results on the lack of concerted copy number variation call into question several recently published associations of rDNA copy number with cancer and aging.

3.3 RESULTS

3.3.1 Meaningful concerted copy number variation is not present in the Simons Simplex Collection

Our initial goal was to estimate rDNA copy number in the Simons Simplex Collection to determine if rDNA copy number is associated with autism spectrum disorder. Autism spectrum disorders associate with variants in hundreds of genes, including single nucleotide, tandem repeat, and copy number variants [289–291]. rDNA copy number variation has not been assessed in autism spectrum disorder; however, it has been hypothesized that higher rDNA copy number associates with a more severe intellectual disability due to the increased potential for rRNA transcription [251]. The Simons Simplex Collection has sequenced hundreds of families with a child affected by an autism spectrum disorder [292–296]. We estimated rDNA copy number in 7,268 individuals from families in which both parents, the proband, and an unaffected sibling were sequenced. We detected no difference in rDNA copy number based on autism status (**Figure 3.1a**), using read coverage estimates (Student's paired t-test, p=0.9365) [258]. We asked whether within probands, rDNA copy number associates with degree of intellectual disability by comparing individuals with an IQ of <=50 to those with an IQ of >=100. We detected a statistically significant difference (nominal p=0.03, Welch two-sample t-test) of individuals with an IQ<=50 having on average eleven more rDNA copies than those with IQ>=100, consistent with the published hypothesis (**Figure 3.1b**). However, eleven additional rDNA copies seem unlikely to be biologically relevant, given an average copy number of ~250 and epigenetic silencing of many copies. Testing additional cutoffs for IQ - including assessing individuals with IQ<70 versus those with IQ>=70, the cutoff for severe versus not severe intellectual disability, yielded similar results

(**Figure A2.1**). With more stringent IQ cutoffs, nominal significance was abolished. Because we previously reported that rDNA copy number estimates based on short-read whole genome sequencing data can be error prone, with eleven copies certainly being within the range of error [5], we sought to validate rDNA copy number in a subset of samples by alternate methods.

Our preferred alternate method to estimate rDNA copy number is the CHEF gel. CHEF gels are the gold standard to estimate rDNA copy number, but they have limitations. In humans, the 45S rDNA array makes up the short arm of each of the 5 acrocentric chromosomes, so when assessing rDNA copy number by CHEF gel ten distinct bands should be observed. No previous study has ever observed ten distinct bands in a 45S CHEF gel, possibly because some rDNA loci are too large to be resolved [126]. In contrast, the 5S rDNA, residing in a single locus, can be readily measured by CHEF gels. Given the previously reported concerted copy number variation between the 5S and 45S rDNA arrays [258], we planned to validate the 5S rDNA copy number estimates by CHEF gels to infer the accuracy of the sequencing-based 45S rDNA copy number estimates.

To this end, we first estimated 5S rDNA copy number in the Simons Simplex Collection to determine the strength of concerted copy number variation. We found weaker correlation than what was reported previously: The 18S and 5S copy numbers correlate with a Spearman coefficient of 0.24, while in the initial study using 1000 Genomes Project data the Spearman coefficient was 0.61 (**Figure 3.1c, Table 3.1**). As we were primarily interested in predicting 45S rDNA copy numbers from 5S rDNA copy numbers, we tested a linear model relating the copy numbers. This linear model is not predictive; the R^2 value is 0.061. As expected, we observe

similar trends when analyzing the 28S and 5.8S copy numbers in relation to the 5S number (**Figure 3.1 de, Table 3.1**).

Table 3.1: Correlations between 45S and 5S rDNA copy numbers in the Simons Simplex Collection.

SSC (n=7210)					
y	x	Spearman	Spearman p-value	Linear model	Multiple R-squared
5S	18S	0.244	< 2.2e-16	y=0.127x + 213	0.061
5S	5.8S	0.297	< 2.2e-16	y=0.134x + 214	0.094
5S	28S	0.292	< 2.2e-16	y=0.155x + 215	0.093

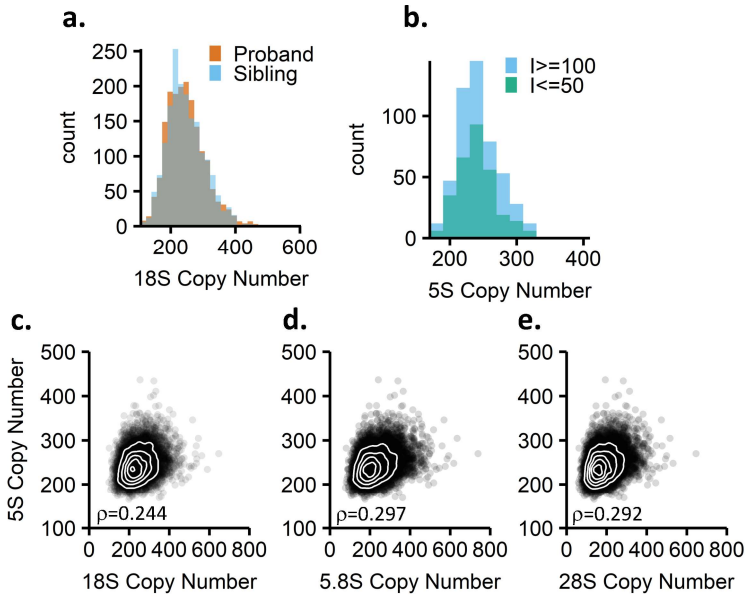

Figure 3.1: rDNA copy number estimates and correlations of 45S and 5S rDNA regions in the Simons Simplex Collection.

a: rDNA copy number distributions for probands (n=1,774) and unaffected siblings (n=1,774) in the Simons Simplex Collection. Paired t-test p-value: 0.937. **b**: Comparison of 18S copy number in probands with low (I<=50, n=298) and high (I>=100, n=500) IQ scores. Welch Two Sample t-test p-value=0.01533, average difference between groups is 9 copies. **c-e**: Correlations of each the 18S, 5.8S, and 28S regions of the 45S rDNA array to the 5S rDNA array with Spearman's rho indicated (n=7,210).

3.3.2 Concerted copy number variation in high-coverage 1000 Genomes data is weak

We next tested if weak concerted copy number variation was specific to the Simons Simplex Collection dataset. The 1000 Genomes Project recently released a new dataset of higher coverage sequencing data for approximately 2,500 samples (Michael Zody, personal communications). The sequencing data from the high-coverage dataset were generated by a single sequencing center with a single library preparation and sequencing method and a target genome coverage of ~30X. This sequencing center also generated the Simons Simplex Collection dataset. We estimated rDNA copy number in 2,419 of the high-coverage 1000 Genomes Project samples which displayed normal karyotypes. We observe a weak but significant correlation between each the 18S, 5.8S, and 28S copy numbers with the 5S copy number: The Spearman coefficients are 0.084, 0.111, and 0.118, respectively (**Figure 3.2a, Table 3.2, and Figure A2.1**). Despite the significance of the correlation, there is no predictive power to the relationship between the 45S and the 5S copy numbers: For example, a linear model relating the 18S and 5S copy numbers has an R^2 value of 0.005 (**Table 2.2**).

The weak concerted copy number variation signal in the high-coverage 1000 Genomes Project dataset is not simply due to an increased number of samples. If we exclusively analyze the high-coverage samples that were a part of the low-coverage study (n=163), we still observe a much weaker correlation than previously reported: the 18S and 5S correlate with a Spearman coefficient of 0.194 (**Figure 3.2b, Table 2.3, and Figure A2.2**). Our results raise the question whether differences in analysis methods or differences in data quality are responsible for the observed discrepancy with the previously reported findings [258].

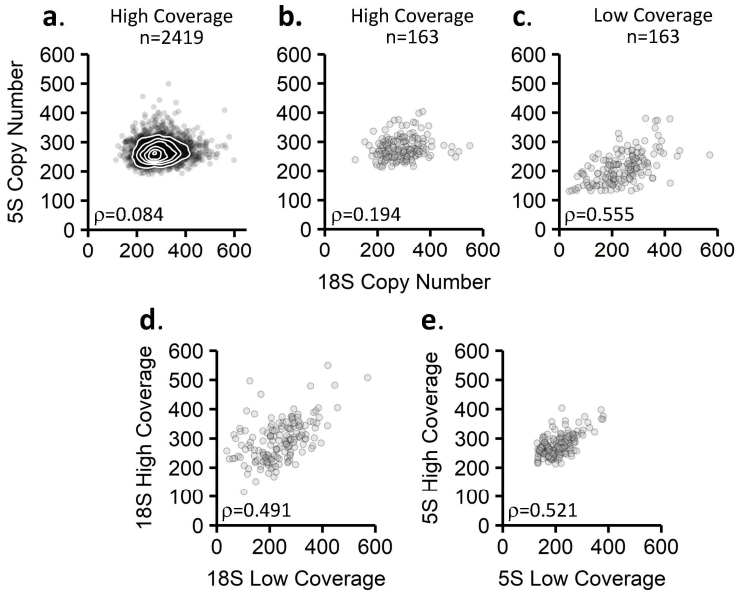

Figure 3.2: Correlations of the 5S and 45S rDNA copy numbers in 1000 Genomes Project data. **a**: Correlation of the 18S to the 5S in the high-coverage 1000 Genomes Project data (n=2,419). **b**: Correlation of the 18S to the 5S in the subset of 1000 Genomes Project data samples also analyzed in the low-coverage dataset (n=163). **c**: Correlation of the 18S to the 5S in the low-coverage 1000 Genomes Project data (n=163). **d-e**: Comparison of rDNA copy number estimates for the same cell lines sequenced separately in the high-coverage and low-coverage 1000 Genomes datasets for the 18S locus (**d**) and the 5S locus (**e**). Spearman's rho is indicated in each panel, n=163.

Table 3.2: Correlations between 45S and 5S rDNA copy numbers in the high-coverage 1000 Genomes Project dataset.

High coverage data (n=2419)					
y	x	Spearman	Spearman p-value	Linear model	Multiple R-squared
5S	18S	0.084	3.69E-05	y = 0.037x + 258	0.005
5S	5.8S	0.111	4.90E-08	y = 0.044x + 255	0.010
5S	28S	0.118	6.97E-09	y = 0.061x + 255	0.011

Table 3.3: Correlations between 45S and 5S rDNA copy numbers in the subset of high-coverage 1000 Genomes Project data also analyzed in the low-coverage dataset.

High coverage data (n=163)					
y	x	Spearman	Spearman p-value	Linear model	Multiple R-squared
5S	18S	0.194	1.36E-02	y = 0.091x + 251	0.029
5S	5.8S	0.217	5.73E-03	y = 0.091x + 251	0.041
5S	28S	0.232	3.08E-03	y = 0.121x + 251	0.042

3.3.3 Our pipeline reproduces concerted copy number variation observed in low-coverage 1000 Genomes Project data

A key difference between our analysis and the original study in which concerted copy number variation was reported lies in the alignment pipelines and post-alignment corrections. We used BWA for alignment [297], while bowtie2 was used in the original study [258,298]. Additionally, the original study used a correction for pseudogene content, which we did not perform because no significant differences in concerted copy number variation were observed with or without corrections [258]. To ensure that our analysis pipeline identifies the previously reported concerted copy number variation, we applied it to 163 of the 168 low-coverage 1000 Genomes Project samples previously studied.

Consistent with the original study, our analysis pipeline detected strong concerted copy number variation between the 45S and 5S loci in the low-coverage 1000 Genomes Project data. The 18S, 5.8S, and 28S rRNA gene copy numbers correlate to the 5S copy number with Spearman coefficients of 0.56, 0.79, and 0.69 (**Table 3.4, Figure 3.2c, Figure A2.2**). These coefficients are similar to those previously published, which are 0.61, 0.80, and 0.73, respectively [258]. We conclude that the failure to detect strong concerted copy number variation does not arise from differences in copy number estimation methods but is likely due to differences in the datasets used for analysis.

Indeed, we find that the low-coverage and high-coverage data yield different rDNA copy number estimates for the same cell lines. In comparing 18S estimates of the same cell lines in the two different datasets, the rDNA copy numbers only correlate with Spearman coefficients of 0.49. The 5S locus shows a Spearman correlation of 0.52 between the two datasets. These values

are far lower than would be expected when analyzing the same cell lines (**Figure 3.2de, Figure A2.3**). The scenario that rDNA copy numbers have changed between samplings for low- and high-coverage data generation is unlikely as there are several reports documenting that rDNA copy number is largely stable in cell lines [204,219]. Taken together, our results are consistent with previous reports that different library preparations of the same samples often yield different rDNA copy number estimates [5].

Table 3.4: Correlations between 45S and 5S rDNA copy numbers in the low-coverage 1000 Genomes Project dataset.

Low Coverage Data (n=163)					
y	x	Spearman	Spearman p-value	Linear model	Multiple R-squared
5S	18S	0.555	< 2.2e-16	y = 0.332x + 132	0.308
5S	5.8S	0.790	< 2.2e-16	y = 0.355x + 136	0.601
5S	28S	0.693	< 2.2e-16	y = 0.366x + 150	0.453

3.3.4 Low-coverage 1000 Genomes Project sequencing data come from multiple sources

We wanted to further explore which differences in the low- and high-coverage datasets lead to different rDNA copy number estimates and the lack of meaningful concerted copy number variation. One difference between the datasets is that the low-coverage 1000 Genomes sequencing data were produced by any of seven sequencing centers, while the high-coverage data were produced by a single center. The seven sequencing centers used various library preparation methods [299], and library preparation methods and batch effects can influence rDNA copy number estimates [5]. Moreover, some samples included reads from multiple library preparations and/or sequencing centers, turning the low-coverage copy number estimates into composite estimates from different libraries.

For each of the 163 low-coverage 1000 Genomes Project samples, we split the sequencing files by library preparation ID. Library depths varied by nearly two magnitudes: chromosome 1 coverage for individual libraries ranged from 0.24X to 14.4X, with an average of 5.4X. When we analyze rDNA copy number estimates from individual libraries by sequencing center, we find that some centers produced sequence data that biased rDNA copy estimates toward higher or lower values (**Figure A2.4**). As it is unlikely that certain centers were assigned samples with abnormally high or low rDNA copy number, the observed bias likely arose through differences in library preparation methods. This bias was not observed in the high-coverage data for the same samples, which supports the notion that sample assignment was not biased.

Of the 163 low-coverage samples, rDNA copy number estimates for 54 samples were based on sequencing data generated from multiple libraries (**Figure 3.3**). Of these 54 samples, some samples, such as NA12154 and NA0700, contained multiple libraries that yielded 18S

estimates very similar to both each other and our high-coverage data estimate. Others, such as NA11892 or NA12778, contained one library with an estimate very similar to our high-coverage data estimate but also other libraries with estimates near zero copies. The libraries with severe underestimates of rDNA copy number tended to have lower coverage, suggesting that read depth may influence rDNA copy number estimates.

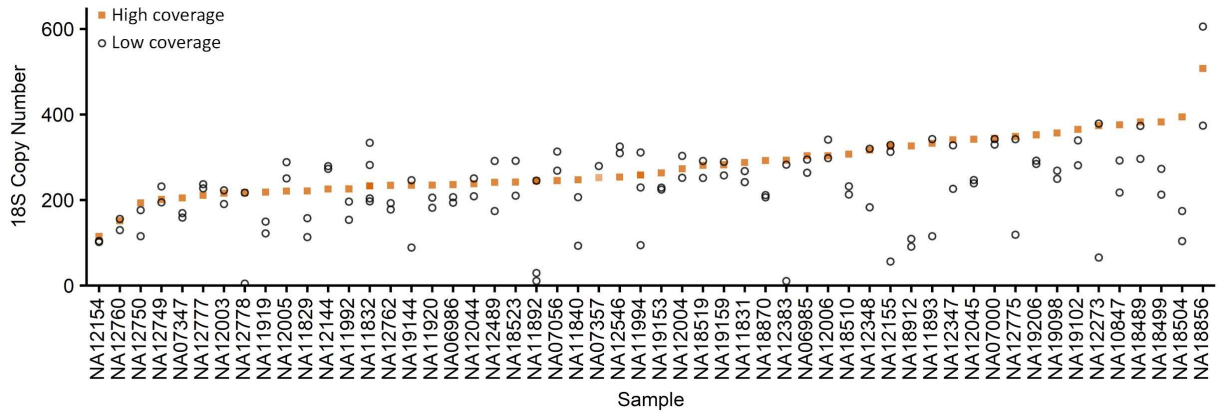

Figure 3.3: Comparison of 18S copy number estimates for different libraries made from the same cell lines.

Of the low-coverage 1000 Genomes Project samples, 54 contained data from multiple sequencing libraries. Orange squares denote the high-coverage 1000 Genomes Project estimate. Open black circles indicate individual library estimates for each sample. Samples are ordered by high-coverage 18S estimate.

3.3.5 Sequencing coverage does not account for the magnitude of rDNA copy number differences observed between the high and low-coverage 1000 Genomes Project datasets

We next investigated if depth of read coverage drives the differences in rDNA copy number estimation between the high- and low-coverage 1000 Genomes Project datasets. We performed downsampling experiments on four samples spanning the range of rDNA copy numbers estimates from the high-coverage 1000 Genomes Project dataset. Randomly downsampling the high-coverage data to 400, 300, 200, 100, 50, and 10 million reads ten times each reveals that reducing coverage does affect rDNA copy number estimates to a small degree. Reassuringly, the average of ten independent downsamplings for a sample was close to the copy number estimate of the full dataset (**Figure 3.4**). For example, NA07357 had an estimated copy number of 252 in the high-coverage dataset. Its average copy number from downsampling ten times to 10 million reads was 249. Nevertheless, the range of copy number estimates increased with decreased coverage. At 10 million reads, the range of 18S estimates for NA07357 varied from 222 to 272, while at 100 million reads it varied from 244 to 260.

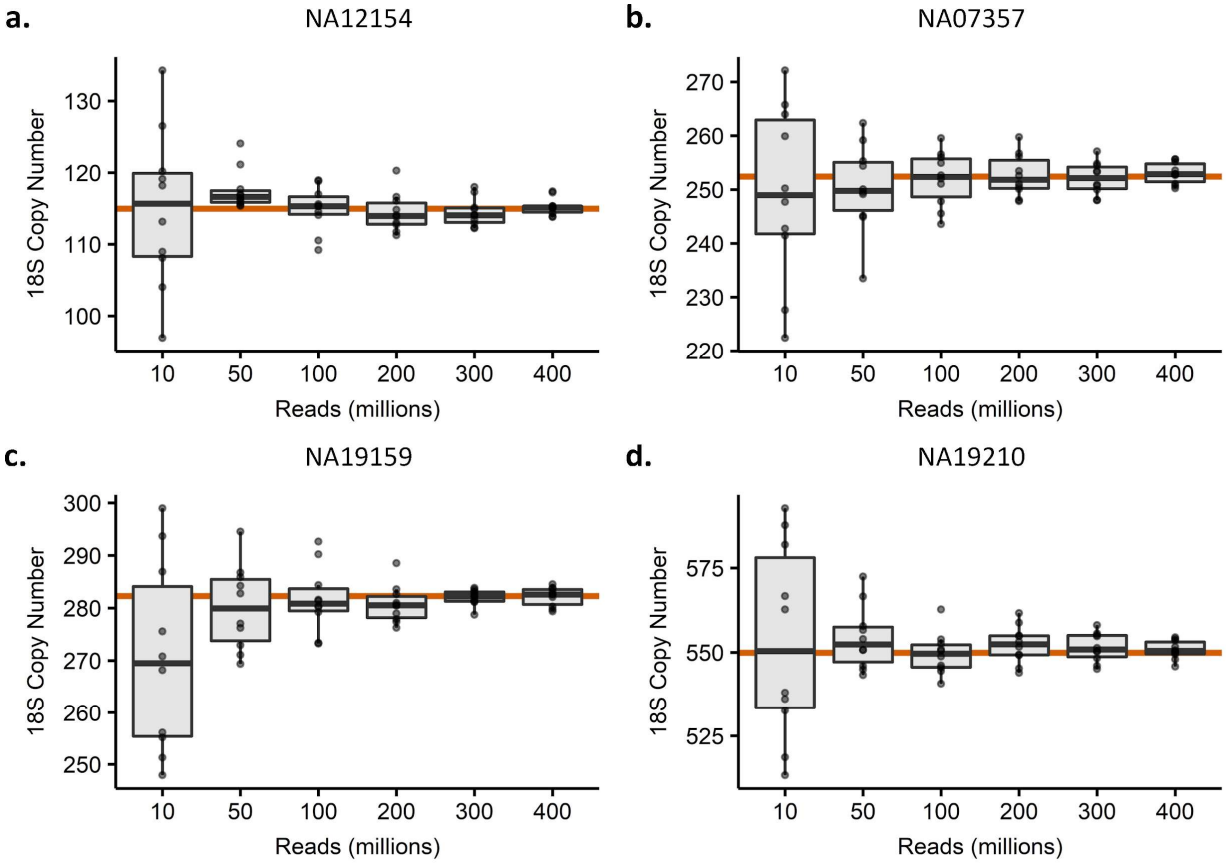

Figure 3.4: Read coverage downsampling of four high-coverage 1000 Genomes Project samples. **a**: NA12154 has an 18S copy number of 115 in the full dataset. **b**: NA07357 has an 18S copy number of 252 in the full dataset. **c**: has an 18S copy number of 282 in the full dataset. **d**: NA19210 has an 18S copy number of 550 in the full dataset. Note different Y axis scales for each graph. Ten independent downsamplings were performed per sample.

For comparison, the individual libraries for the low-coverage 1000 Genomes Project samples have a mean coverage of 5.4X for chromosome 1. The comparable samples in the downsampling experiment would be those with 100 million reads, which have a chromosome 1 coverage of approximately 4.5X. For the four samples analyzed, at ~4.5X coverage, the copy number estimates are at most +/- 13 copies of the full dataset estimate. Even at 10 million reads (~0.45X coverage), copy number estimates are at most +/- 43 copies of the full dataset estimate. Meanwhile, comparing the high- and low-coverage estimates, the low-coverage libraries underestimated the 18S by an average of 57 copies. The different libraries ranged from underestimating by 371 copies to overestimating by 110 copies as compared to the high-coverage data. These differences are much larger than what can be explained by differences in sequencing coverage alone. We conclude that the vast discrepancies in copy number estimates between datasets are not due to sequencing coverage, and hypothesize that the discrepancies may have arisen through batch effects or technical differences in library preparation.

3.3.6 Simons Simplex Collection and high-coverage 1000 Genomes Project data are likely higher quality than the low-coverage 1000 Genomes Project data

To test the above hypothesis, we analyzed the correlation of copy number estimates of the three 45S components to each other, an approach previously used as a quality metric [2]. Because these components are part of the same array, they should correlate highly. As expected, the high-coverage data showed somewhat higher intra-45S correlations than the low-coverage data. For example, the 18S and 28S copy numbers in the high-coverage dataset have a Spearman correlation of 0.97 (**Figure 3.5a, Table 3.5**). In the 163 samples analyzed in both the high- and low-coverage data, the 18S and 28S copy numbers in the high-coverage dataset showed a

Spearman correlation of 0.97 but has a Spearman correlation of 0.92 in the low-coverage dataset (**Figure 3.5bc, Table 3.6, Table 3.7**). This trend is reinforced by analysis of a linear model relating the two copy numbers. For the 163 samples analyzed in both 1000 Genomes Project datasets, the R^2 value for the high-coverage dataset was 0.95 while it was 0.85 for the low-coverage dataset (**Tables 3.6 and Table 3.7**). Analysis of the correlation of the 18S to the 5.8S and the 28S to the 5.8S show the same trends between the three datasets (**Figure A2.5**) The improved intra-45S correlations suggest that the rDNA copy number estimates in the high-coverage 1000 Genomes Project dataset are of higher quality.

There are three different data quality metrics that can be analyzed with the Simons Simplex Collection data. As with the 1000 Genomes data, we first assessed intra-45S correlations, finding similarly high Spearman correlations: The 18S and 28S copy numbers showed a Spearman correlation of 0.943, and a linear model relating the two had an R^2 value of 0.90 (**Figure 3.5d**). The family structure of the Simons Simplex Collection permits analysis of rDNA copy number heritability. We found high heritability of both 45S and 5S rDNA array components (h^2=0.94 for the 18S, h^2=0.92 for the 5S) (**Figure 3.5ef, Figure A2.6**).

Unique to this study, we can also assess reproducibility of rDNA copy number estimates through use of monozygotic twins and duplicate samples. The Simons Simplex Collection includes four pairs of monozygotic twins that showed perfect correlation of 18S copy number. Additionally, some individuals in the Simons Simplex Collection enrolled in other studies and were therefore sequenced twice, albeit by the same sequencing facility. These duplicate samples were identified from shared SNVs and would be expected to share rDNA copy numbers. Although duplicate samples can show considerable deviation from one another due to technical issues [5],

we found reasonably high correlation between the 18S copy number estimates arising from the duplicated samples (Spearman correlation 0.81, **Figure 3.5g, Table 3.8**). The 5S copy number estimates also showed high correlation between the twin and duplicated samples (Spearman correlation 0.902, **Figure 3.5h**). Similar to the comparisons between the high- and low- coverage 1000 Genomes Project data, the correlation of 28S and 5.8S estimates in the twin and duplicate samples of the SSC was lower than that of the 18S (**Figure A2.6**). These values are still substantially higher than the correlation values between the same regions in the high- and low-coverage 1000 Genomes Project, indicating that the replicate sequencing events in the Simons Simplex Collection produce more similar rDNA copy number estimates. Together, the heritability data, intra-45S correlations, and duplicated samples give confidence in the Simons Simplex Collection rDNA copy number estimates.

We conclude that co-variation between the 45S and 5S rDNA arrays in both high-coverage datasets is weak and does not allow prediction of the copy number at one array based on the copy number at the other. This lack of meaningful concerted copy number variation holds true regardless of whether data were generated using lymphoblastoid cell lines or whole blood samples. The previously observed concerted copy number variation in the low-coverage dataset appears to be an artifact of lower data quality.

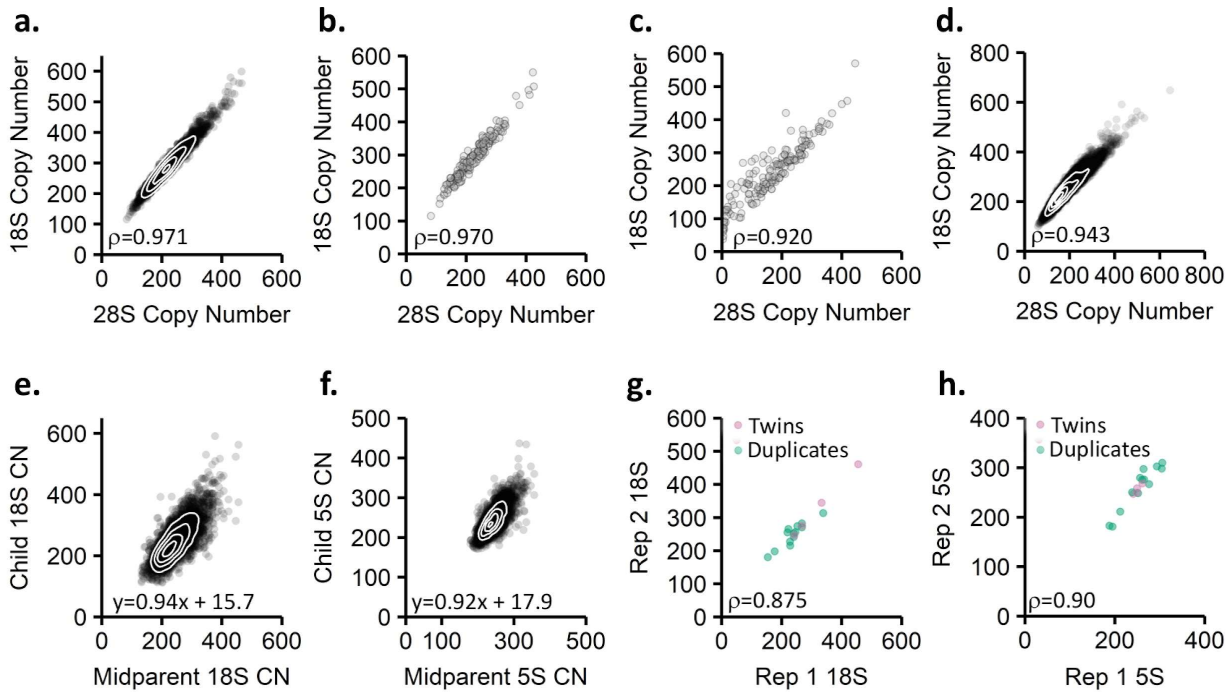

Figure 3.5: Data quality metrics for rDNA copy number estimates.

a-d: Correlations between the 18S and 28S regions of the 45S rDNA repeat unit for the (**a**) high-coverage 1000 Genomes Project data (n=2,419), (**b**) subset of high-coverage 1000 Genomes Project data also analyzed in the low-coverage dataset (n=163), (**c**) low-coverage 1000 Genomes Project data (n=163), and (**d**) Simons Simplex Collection data (n=7,210). **e**: Heritability of the 18S copy number in the Simons Simplex Collection (n=3,548). **f**: Heritability of the 5S rDNA copy number in the Simons Simplex Collection. **g-h**: Comparison of 18S (**g**) and 5S (**h**) rDNA copy number estimates for either monozygotic twins (n=4 pairs) or for individuals sequenced twice in the Simons Simplex Collection (n=13). Spearman correlation indicated is for monozygotic twins and duplicates analyzed together.

Table 3.5: Correlations between the three rRNA genes encoded in the 45S repeat unit in the high-coverage 1000 Genomes Project dataset.

High coverage data (n=2419)					
y	x	Spearman	Spearman p-value	Linear model	Multiple R-squared
5.8S	18S	0.973	< 2.2e-16	y = 1.126x + -33	0.955
28S	18S	0.971	< 2.2e-16	y = 0.862x + -27	0.951
28S	5.8S	0.990	< 2.2e-16	y = 0.761x + -0.19	0.983

Table 3.6: Correlations between the three rRNA genes encoded in the 45S repeat unit in the subset of high-coverage 1000 Genomes Project data also analyzed in the low-coverage dataset.

High coverage data (n=163)					
y	x	Spearman	Spearman p-value	Linear model	Multiple R-squared
5.8S	18S	0.972	< 2.2e-16	y = 1.157x + -44	0.951
28S	18S	0.970	< 2.2e-16	y = 0.886x + -33	0.951
28S	5.8S	0.990	< 2.2e-16	y = 0.760x + 1.96	0.986

Table 3.7: Correlations between the three rRNA genes encoded in the 45S repeat unit in the low-coverage 1000 Genomes Project dataset.

Low Coverage Data (n=163)					
y	x	Spearman	Spearman p-value	Linear model	Multiple R-squared
5.8S	18S	0.834	< 2.2e-16	y = 1.086x + -49	0.690
28S	18S	0.920	< 2.2e-16	y = 1.017x + -75	0.853
28S	5.8S	0.930	< 2.2e-16	y = 0.779x + 3.6	0.857

Table 3.8: Correlations between the three rRNA genes encoded in the 45S repeat unit in the Simons Simplex Collection.

SSC (n=7210)					
y	x	Spearman	Spearman p-value	Linear model	Multiple R-squared
5.8S	18S	0.962	< 2.2e-16	y=1.131x + -49	0.928
28S	18S	0.943	< 2.2e-16	y=0.956x + -48	0.956
28S	5.8S	0.988	< 2.2e-16	y=0.849x + -7.5	0.981

3.4 DISCUSSION

In this study, we sought to further explore the relationship between the copy numbers of the 45S and 5S rDNA arrays. We have previously reported that short-read sequencing estimates of rDNA copy number genotypes are error-prone [5]. Given the previously published concerted copy number variation of the 5S and 45S rDNA arrays in humans, we thought this co-variation may provide a useful metric to predict 45S rDNA copy number from 5S copy number, the latter of which can be readily obtained by pulsed-field gel electrophoresis.

Working with two new, high-coverage sequencing datasets, we found weak concerted copy number variation that was not predictive. Sequencing coverage alone did not explain the discrepancy in copy number estimates or concerted copy number variation between samples sequenced in both the original low-coverage 1000 Genomes Project dataset and the newer high-coverage 1000 Genomes Project dataset. Available sequence data for many samples of the low-coverage dataset were derived from multiple library preparations from multiple sequencing centers. We suspect that differences in library preparation methods lead to these considerable discrepancies in rDNA copy number estimates. Because we were able to recapitulate the previously observed concerted copy number variation from the low coverage 1000 Genomes Project samples, the differences in results between the datasets stem from sample preparation methods, not analysis methods.

In addition to the promise of a predictive model, the previously reported concerted copy number variation had fascinating implications for biology, in particular for genome maintenance and evolution. Selection for strongly concerted copy number variation suggests the existence of mechanisms that “count” and adjust copy numbers accordingly, operating across several

genomic loci on separate chromosomes. It is not fully understood how quantities of the rRNA products of the 45S and 5S arrays are balanced to facilitate ribosome biogenesis. Concerted copy number variation could have provided a possible mechanism of maintaining proper rRNA dosage by means of balancing their genomic templates. However, vast stretches of 45S rDNA repeats are typically silenced, resulting in the total number of 45S rDNA copies not necessarily reflecting the number of transcribed copies. Less is known about 5S regulation in mammals, though 5S silencing is prevalent in *A. thaliana* and *Xenopus* species [286,300–303]. The abundance of 45S silencing suggests that maintaining concerted copy number variation of the rRNA genes is not crucial for balancing rRNA levels.

In support of our findings, this lack of a meaningful correlation between 45S and 5S arrays was previously observed in model organism studies. Co-variation between 45S and 5S rDNA copy numbers was weak in a large mutant collection of over 2000 strains and 40 natural isolates of *C. elegans*, which estimated rDNA copy number from high-coverage sequence data [1]. Meaningful covariation between these arrays was also found to be lacking in a study of various ecotypes of *A. thaliana*, and separately in several species of fish [304,305]. In short, our finding that copy numbers at the 45S and 5S arrays show little correlation is biologically plausible and supported by studies of model organism rDNA.

Some may argue that the observed differences in rDNA copy number between the low- and high-coverage 1000 Genomes Project datasets stem from biological differences; *i.e.* that the studied samples have acquired changes in rDNA copy number. We consider this an unlikely scenario. The DNA used for each sequencing effort was extracted from lymphoblastoid cell lines, which were propagated for an unknown number of generations from a common stock between

each study. However, rDNA copy number has been reported as stable both in cell lines [204,219] and multicellular organisms such as *C. elegans* [1,5], except for when rDNA copy number is reduced to a level that causes fitness defects such as in *S. cerevisiae* or *D. melanogaster* [143,306].

We present our results as a cautionary tale about the challenges of genotyping repetitive DNA. While our results suggest that the high-coverage uniform sequencing performed by the New York Genome Center for the updated 1000 Genomes Project and the Simons Simplex Collection likely yielded more accurate rDNA copy number estimates, there is no certainty without validation through alternate methods. We are reassured by the fact that the observed lack of meaningful covariation between 45S and 5S rDNA copy numbers is consistent with current biological knowledge. We therefore trust our finding that there is no association of rDNA copy number with autism spectrum disorder. However, in light of our findings, we posit that care should be taken when drawing biological conclusions based on rDNA copy number estimates from short-read whole genome sequencing data. Changes in rDNA copy number have been reported for some cancers and in aging, prompting speculation about their role as drivers or essential players in both processes. A subset of these findings relies on rDNA copy number estimates from short read sequencing data [2,216,258]. However, even studies that develop their own rDNA copy number estimation methods still compare to short read sequencing data to comment on their accuracy [94,205]. Some of these findings may need to be re-evaluated by applying multiple data quality metrics to the analyzed sequence data, by conducting uniform, high-coverage re-sequencing, or by validating rDNA copy number through alternative approaches. Going forward, we hope that these results will caution researchers about drawing

firm conclusions from rDNA copy number estimates based on short-read sequencing data, as public data repositories often contain samples generated with various methods and by different sources.

3.5 METHODS

Sequence analysis

Samtools version 1.9 and bwa version 0.7.15 were used for all analyses [297,307].

Reference sequences for the 45S ribosomal DNA (U13369.1), 5S ribosomal DNA (X12811.1), mtDNA, and chromosome 1 of the human genome (GRCh38 reference) were downloaded from the NCBI nucleotide database with GenBank IDs as indicated in parenthesis. CRAM files were converted to fastq by Samtools fastq and aligned to the appropriate reference sequence by bwa mem with default parameters and converted to CRAM files with samtools view. Per-base read depth was calculated with Samtools depth outputting all positions, with the –d 0 flag to eliminate a maximum read depth cutoff.

To estimate ribosomal DNA copy number, average read depth across the whole 5S or 5.8S coding sequence was divided by the average chromosome 1 read depth. These regions correspond to positions 271-391 of X12811.1 for the 5S and positions 6623-6779 of U13369.1 for the 5.8S For the 18S and 28S subunits, segments of these genes previously used for concerted copy number variation analysis were used, and the average read depth at these regions was divided by the chromosome 1 depth [258]. These are positions 3,841-3,985 of U13369.1 for the 18S, and 8,049-8,198 of U13369.1 for the 28S.

Downsampling

Four samples across the rDNA copy number range were used for downsampling: NA03757, NA19210, NA12154, and NA19159. The number of reads in a given CRAM file were determined with samtools view -c. The percent of the total reads that 10, 50, 100, 200, 300, and 400 million reads consisted of was calculated. Downsampling without replacement was performed with samtools view -s. Ten unique seeds (569, 94, 124, 406, 588, 4859, 7734, 3234, 223, and 82) were used to initiate each downsampling, so that downsamplings could be replicated if necessary.

Statistical analysis

Analyses were performed in Rstudio version 1.0.153, with R version 3.5.1.

3.6 AVAILABILITY OF DATA AND MATERIALS

High-coverage 1000 Genomes Project data are available at the following website: https://www.ebi.ac.uk/ena/data/view/PRJEB31736

The following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: [NA06984, NA06985, NA06986, NA06989, NA06994, NA07000, NA07037, NA07048, NA07051, NA07056, NA07347, NA07357, NA10847, NA10851, NA11829, NA11830, NA11831, NA11832, NA11840, NA11843, NA11881, NA11892, NA11893, NA11894, NA11918, NA11919, NA11920, NA11930, NA11931, NA11932, NA11933, NA11992, NA11994, NA11995, NA12003, NA12004, NA12005, NA12006, NA12043, NA12044, NA12045, NA12046, NA12058, NA12144, NA12154, NA12155, NA12156, NA12234, NA12249, NA12272, NA12273, NA12275, NA12282, NA12283, NA12286, NA12287, NA12340,

NA12341, NA12342, NA12347, NA12348, NA12383, NA12399, NA12400, NA12413,, NA12414, NA12489, NA12546, NA12716, NA12717, NA12718, NA12748, NA12749, NA12750, NA12751, NA12760, NA12761, NA12762, NA12763, NA12775, NA12776, NA12777, NA12778, NA12812, NA12813, NA12814, NA12815, NA12827, NA12828, NA12829, NA12830, NA12842, NA12843, NA12872, NA12873, NA12874, NA12878, NA12889, NA12890]. These data were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1.

Sequencing data for the SSC is available through the Simons Foundation for Autism Research Initiative (SFARI) and is available to approved researchers at SFARI base (http://base.sfari.org, accession IDs: SFARI_SSC_WGS_p, SFARI_SSC_WGS_1, and SFARI_SSC_WGS_2).

Copy number estimates used in this study are provided as supplementary data files.

3.7 Acknowledgements and declarations

We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). We appreciate obtaining access to genetic data on SFARI Base. Approved researchers can obtain the SSC population dataset described in this study SFARI_SSC_WGS_p, SFARI_SSC_WGS_1, and SFARI_SSC_WGS_2 by applying at https://base.sfari.org.

Project Contributions

CQ, TNT, and ANH conceived of the study. TNT estimated rDNA copy number from data available on 1000 Genomes and SFARI base. TNT obtained access to SSC data as an approved researcher through SFARI base. ANH performed all downstream analyses on rDNA copy number, as well as estimating rDNA copy numbers from downsampling experiments. ANH wrote the initial draft of the manuscript and made the figures. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported, in part, by grants from the US National Institutes of Health (R00 MH117165 to T.N.T., F31 AG063450 to A.N.H, NIGMS grant R01 GM122088 to C.Q., and NHGRI grant 1RM1HG010461 to C.Q)

CHAPTER 4. *CAENORHABDITIS ELEGANS* RESOURCES TO STUDY rDNA COPY NUMBER

4.1 SUMMARY

Few resources to study rDNA copy number variation in a controlled genetic background exist. Most rDNA copy number resources are in *S. cerevisiae* -- a unicellular yeast. Studies of rDNA copy number in yeast have provided vast insight into rDNA biology, including rDNA replication, transcription, and stability. In multicellular eukaryotes, there are lines of *D. melanogaster* with reduced rDNA copy number that have been used to demonstrate effects of rDNA copy number on global chromatin regulation and gene expression. Despite these sparse resources, rDNA copy number variation in diverse genetic backgrounds is abundant in wild isolates of laboratory organisms. For example, wild isolates of *C. elegans* vary in rDNA copy number from approximately 70-420 rDNA copies in addition to differing from the laboratory strain by thousands of SNPs and INDELs. To address how rDNA copy number affects phenotype, rDNA copy number variation must be isolated from other genetic variation. In this chapter, I describe the development of *C. elegans* recombinant inbred lines (RILs) and near isogenic lines (NILs) to study how rDNA copy number affects fitness and aging traits.

4.2 INTRODUCTION

Rationale of model system

We selected *C. elegans* as our model organism for studying rDNA copy number variation. *C. elegans* is a self-fertilizing hermaphroditic nematode and an established model for aging. From a pragmatic laboratory perspective, *C. elegans* permits rapid strain development with a 3-day life cycle, preservation of genotypes by freezing, and approximate 3-week lifespan that facilitates

aging studies. *C. elegans* has the dual capacity of hermaphroditic self-propagation, as well as male-hermaphroditic cross propagation [308]. With regard to the rDNA, *C. elegans* has a 7.2-kb 45S rDNA repeat unit next to the telomere on chromosome I, and the 5S internally on chromosome V [124,309] (**Figure 1.1**). The 5S locus also encodes SL1, a 22 base pair leader RNA that is trans-spliced onto approximately 60% of *C. elegans* mRNAs [310]. Forty wild isolates of *C. elegans* have been characterized for their rDNA copy numbers and encode 70-420 copies of the 45S locus and 80-339 of the 5S locus [1,311]. Unlike *D. melanogaster*, *C. elegans* does not encode any rRNA-disrupting retrotransposons, so only intact rDNA copies are present. Having a single 45S rDNA array could make rDNA copy number manipulation simpler than manipulating the multiple arrays that would be found in human cell lines.

Generation of rDNA copy number variation in eukaryotic laboratory organisms

Diverse methods have been used to change rDNA copy number in model organisms. Targeted cutting of rDNA arrays or the induction of genome instability can cause a loss of rDNA copies [165,312]. In cases where directly manipulating rDNA copy number is infeasible, phenotypic consequences of rDNA copy number have been studied by incorporating natural variation into association studies [2,129,313]. Others still have identified changes in rDNA copy number in laboratory organisms, either through extended propagation in mutation accumulation or mutagenesis experiments [1,137,314]. In *C. elegans* there are some strains with varying rDNA copy numbers, either recombinant inbred lines that used a parental strain with high or low rDNA copy number, or duplications of part of chromosome I (**Appendix 3**). These strains are limited, however, and often have other non-rDNA variation as well.

The most straightforward method of changing rDNA copy number is to directly induce breaks in the rDNA, which commonly results in rDNA array repair with a reduced rDNA copy number [134]. The homing endonucleases I-PpoI and CreI cut in the rDNA, but rarely elsewhere in the genome, and they have been used to induce double-strand breaks in human rDNA and rDNA copy number reduction in *D. melanogaster* [312,315]. Targeting the rDNA with CRISPR-Cas9 can also change rDNA copy number, which has been used to generate lines of *A. thaliana* with <20% of the rDNA copies of wild type [165]. In *S. cerevisiae*, application of CRISPR editing has been used to both reduce rDNA copy number and introduce point mutations using template-mediated repair [163]. A major limitation to these approaches is that rDNA copy loss is much more likely to occur than rDNA copy gain. Therefore, we are limited in our ability to study increases in rDNA copy number.

Similar to targeted cutting of the rDNA, inducing rDNA instability can cause rDNA loss. One way in which rDNA loss has been induced in this way is through perturbation of chromatin remodelers. In human cells, deletion of SIRT7, a histone deacetylase that acts at the rDNA, leads to rDNA instability and copy number loss [316]. In *A. thaliana*, the FASCIATA (FAS) genes encode components of the chromatin assembly factor complex that deposits histones H3 and H4 after DNA replication [317]. Mutation of the FAS genes reduces rDNA copy numbers down to ~10% that of wild type [318–320]. Importantly, these rDNA copy number reductions are dependent on maintaining the FAS mutation: Re-introduction of the FAS genes allows rDNA copy number to increase over generations [321]. It is therefore unlikely that unique strains or cell lines in a completely wild type background with reduced rDNA copy number could be made by mutating chromatin remodelers and reintroducing the gene after rDNA loss.

It may not always be possible to change rDNA copy number in an organism with directed methods. An alternate way to study rDNA copy number is to harness natural rDNA copy number variation through one of two methods: Use of rDNA copy number in GWAS studies, or by crossing rDNA arrays from wild strains into a laboratory strain background. Examples of the former are limited, likely because rDNA copy number variation has been difficult to ascertain. While GWAS studies have identified the rDNA as a contributing locus for some phenotypes in plants, flies, and yeast, most of these associations are due to rDNA sequence content and not rDNA copy number [119,157]. The phenotypically relevant rDNA-linked genetic variation is often a difference in the length of the intergenic spacer that changes levels of rRNA transcription [119]. Beyond their potential as a contributing locus in GWAS studies, the arrays of differing lengths in wild populations of laboratory organisms are a reservoir of rDNA copy number variation that could be introduced into laboratory strains by crossing. This final technique is one I will use: Introgression of rDNA arrays from wild isolates of *C. elegans* into the laboratory strain background.

rDNA copy number stability

Once rDNA copy number changes are made, a key question is how stable the changed copy numbers are. Wild isolates of *C. elegans* maintain their rDNA copy numbers over laboratory propagation, which could indicate selection for a strain-specific copy number [1,5]. With respect to typical laboratory strains, rDNA copy number has been monitored in mutation accumulation experiments in *C. elegans* and *Daphnia pulex*. Wild type *D. pulex* maintains a stable rDNA copy number over 100 generations [137]. Meanwhile, wild type *C. elegans* propagated for 400 generations under laboratory conditions reportedly increased their rDNA copy number by approximately 2-fold [322]. However, the *C. elegans* mutation accumulation lines propagated for

the longest were sequenced with a different technology than all other lines -- 454 sequencing as compared to Illumina. Differences in copy numbers due to sequencing technology or batch effects could account for some of the large copy number changes observed [5].

In organisms with drastically reduced rDNA copy numbers, vast changes in rDNA copy number can occur over a short period of time [142,143]. Classic rDNA magnification in flies may occur in a single generation -- resulting in a fly with a bobbed phenotype producing wild type progeny [148]. Further, old male flies produce offspring with reduced rDNA copy number, but those offspring regain rDNA copies and produce progeny with normal rDNA levels [206]. It is unknown if worms undergo rDNA magnification or whether increasing rDNA copy number too high will cause spontaneous loss. If, for example, having too many rDNA copies was unfavorable for growth or reproduction, we may observe sudden loss of rDNA copies, similar to how rDNA magnification causes a sudden gain in rDNA copies.

In this chapter, I present two different resources in *C. elegans* to study differences in rDNA copy number. The first is a panel of 118 Recombinant Inbred Lines (RILs) between a wild isolate with ~417 rDNA copies per haploid genome (MY1) and a derivative of the laboratory strain with ~130 rDNA copies per haploid genome (SEA51). When propagated for 20 generations, rDNA copy number is largely stable in these RILs. The second resource is a series of Near Isogenic Lines (NILs) with increased or decreased rDNA copy number in a set genetic background. These resources are the first RIL and NIL populations in *C. elegans* designed around rDNA copy number differences. I will apply these RIL and NIL populations in Chapter 5, to determine if rDNA copy number variation in the naturally occurring range affects fitness or aging traits in *C. elegans*.

4.3 DESCRIPTION OF STRAIN RESOURCES

4.3.1 Recombinant Inbred Lines

To understand the effects of rDNA copy number on health-related traits, we made a set of RILs in which half have high rDNA copy number and half have wild-type rDNA copy number. These RILs can be used to identify whether rDNA copy number contributes additively or as an interacting locus to a trait. In addition, they permit the identification of non-rDNA loci that affect a trait. We selected the MY1 wild isolate, which has 417 rDNA copies per haploid genome and differs from the laboratory strain N2 by approximately 90,000 SNVs and 30,000 small INDELs and crossed to a laboratory strain derivative, SEA51. SEA51 carries the GFP-containing *mIs13* transgene on the right arm of chromosome I, proximal to the rDNA[1] (**Figure 1.1**). The *mIs13* transgene acts as a visual marker which can be used to select for high rDNA copy number (no GFP) or low rDNA copy number (GFP+). The *mIs13* transgene is approximately ~615kb in size encoding ~102 copies of GFP, expressed constitutively in the pharynx, intestine, and germline [323]. To make the RILs, we performed one cross between MY1 and SEA51 and selected 120 F1s to propagate independently to homozygose the genome (**Figure 4.1**).

We genotyped the RILs by short read sequencing after the genotypes were homozygosed by ten generations of single-worm propagation and estimated rDNA copy number from these data. Selection on copy number through use of the *mIs13* transgene was perfect: Every strain that was selected as homozygous for GFP had approximately 130 rDNA copies per haploid genome. Most chromosomes of a given RIL have between one and four large haplotype blocks,

[1] The precise location of *mIs13* is unknown, but given the genotypes of RILs that carry MY1 DNA near the rDNA but are GFP+ (carry *mIs13*), *mIs13* is within 1-2Mb of the rDNA locus.

as expected from the cross strategy (**Figure A3.1**). In addition, some genetic regions display strong, but expected, non-random segregation distortion in the RILs (**Figure 4.2**). The first is the mitochondrial DNA: All mitochondrial genomes in the RILs come from the MY1 genetic background, due to the design of the cross (MY1 as the hermaphrodite, SEA51 as the male). In addition we observe strong segregation distortion on chromosome I. The locus driving the chromosome I segregation distortion is the *zeel-1*/*peel-1* toxin-antitoxin locus. MY1 is *zeel-1*/*peel-1* null, and is therefore incompatible with the SEA51 *zeel-1*/*peel-1* wild type locus [324]. This genotype distribution is compounded by selection on the rDNA as a part of the cross scheme. The right arm of chromosome I is not affected by this *zeel-1*/*peel-1*-driven segregation distortion, because as a part of the cross strategy to make the RILs, we selected half of the strains to be homozygous for MY1 at the right end of chromosome I and half to be SEA51.

Because lines with differing rDNA copy numbers had not previously been made in *C. elegans*, we wondered if the rDNA copy numbers of the RILs would be stable. We propagated the RILs for 20 generations as populations and estimated rDNA copy number from short read sequencing (**Figures 4.1 and 4.2**). While sequencing is error-prone, by including the parental strains as controls, we can assess whether each RIL is within the 130- or 420-copy number range. Overall, the rDNA copy numbers are similar between the RILs and propagated RILs, and the differences are in line with our previous estimates of sequencing-based technical error [5]. I validated the rDNA copy numbers of a subset of RILs by CHEF gel. As expected, the absolute sequencing-based rDNA copy numbers did not precisely match that of the CHEF gel (**Figure A3.2, Table A3.2**). However, the range estimates are appropriate: The strains with the *mIs13* transgene do have ~110-140 rDNA copies per haploid genome, and none have an rDNA copy number in the

range of ~420 copies per haploid genome. Importantly, we see no evidence of drastic rDNA loss or magnification over the 20 generations of laboratory propagation.

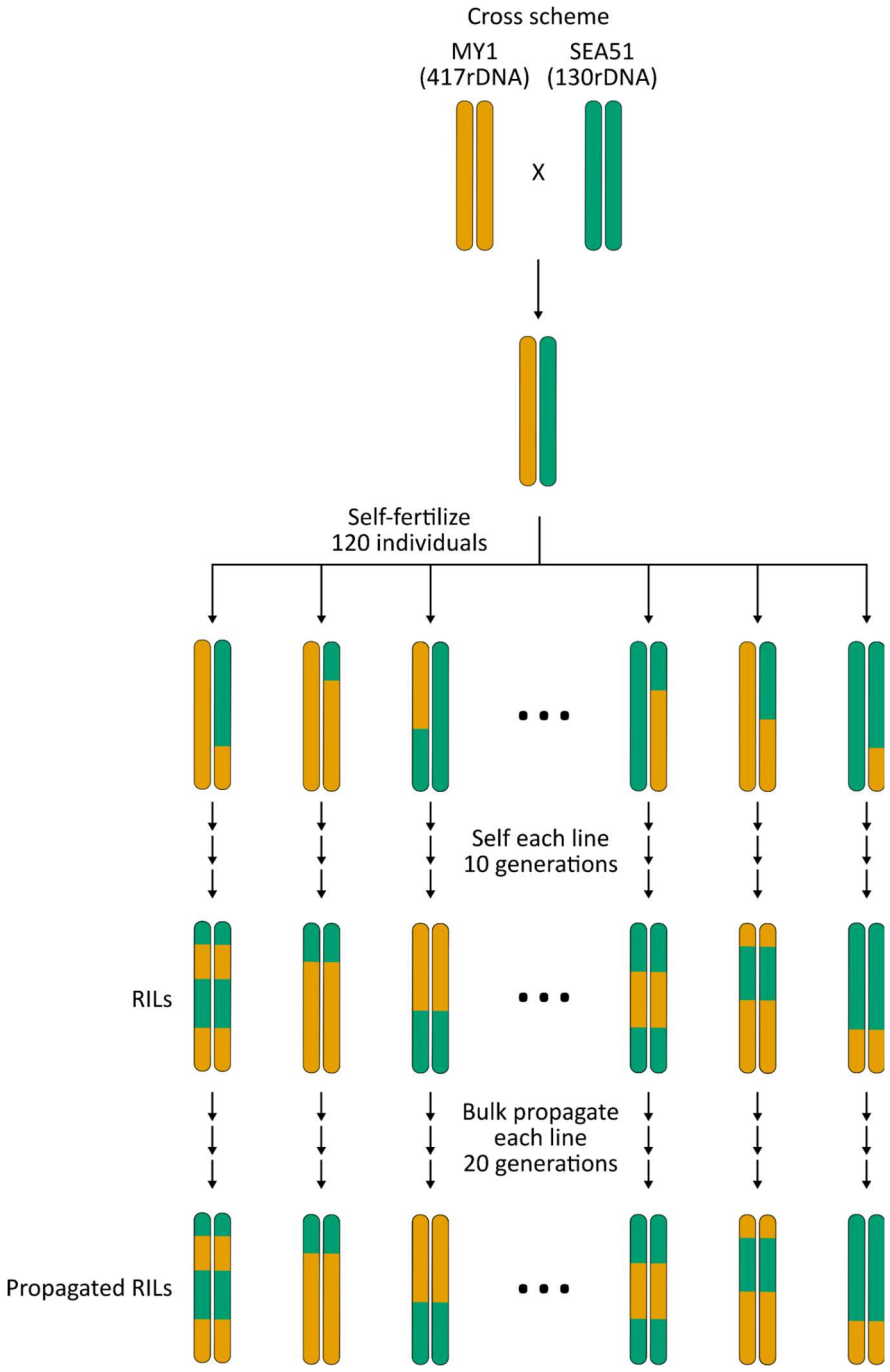

Figure 4.1: Schematic of the construction of MY1xSEA51 recombinant inbred lines.
Parental strains MY1 (hermaphrodite) and SEA51 (male) were crossed and F1s selfed to initiate production of 120 RILs. At the F2 generation, 60 GFP+ (130-rDNA) and 60 GFP- (417-rDNA) worms were selected to self for 10 generations. After 10 generations of single-worm selfing, worm strains were frozen down and designated as RILs. The RILs were then propagated an additional 20 generations in bulk, and the final propagated worms were frozen down and designated as the "Propagated RILs".

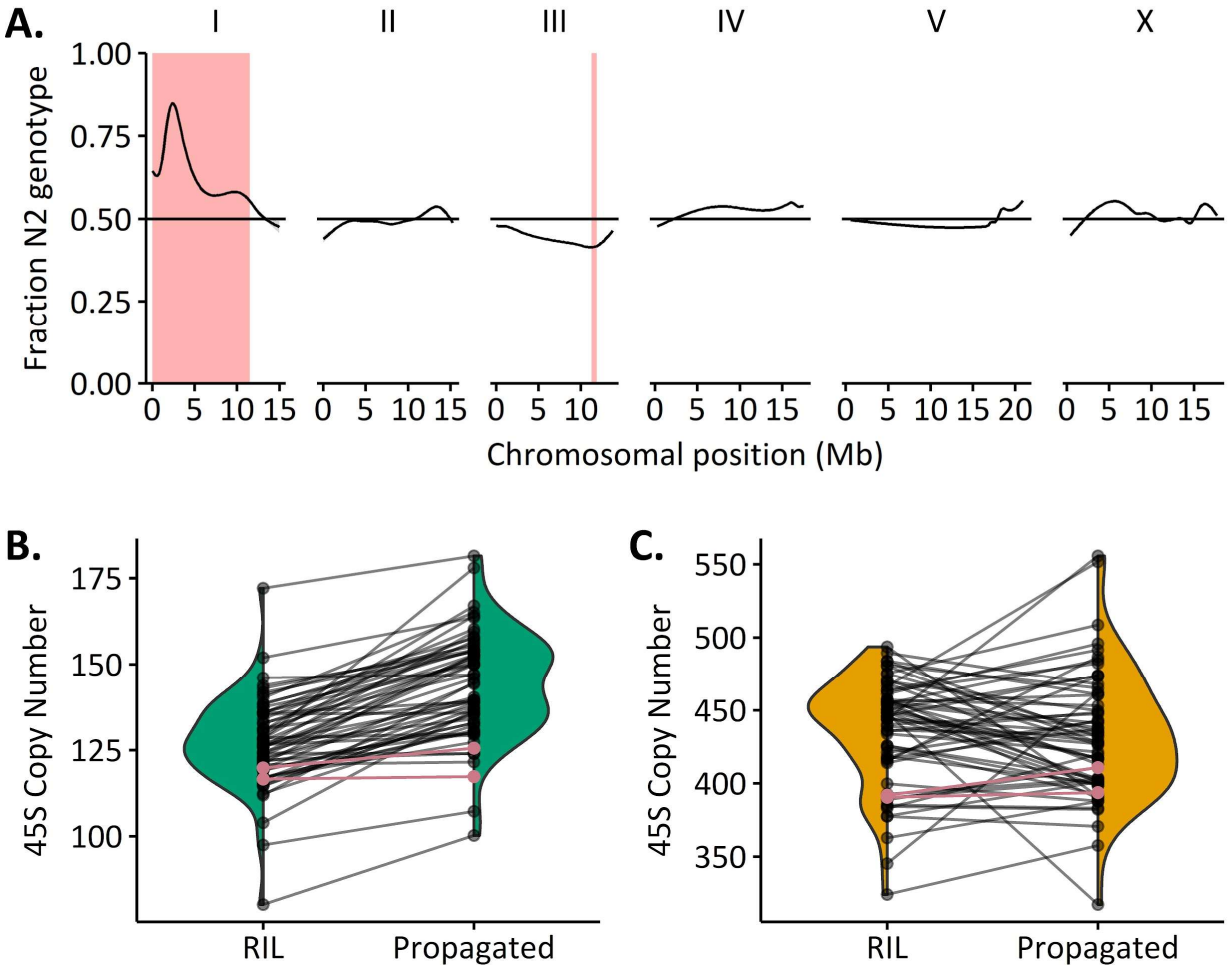

Figure 4.2: Genotyping and rDNA copy number estimation of MY1xSEA51 recombinant inbred lines.

A: Proportion of RILs with N2 genotype at each position along the *C. elegans* genome. Regions highlighted in red show significant distortion from a 50:50 ratio with a Bonferroni corrected p-value of 0.05 or lower. The peak on chromosome I at approximately 2.3 Mb corresponds to the *zeel-1*/*peel-1* toxin-antitoxin locus. **B:** rDNA copy number estimates from short read sequencing for RILs and propagated RILs linked to *mIs13*. Pink lines indicate copy number estimates of the SEA51 parental strain. **C:** rDNA copy number estimates from short read sequencing for RILs and propagated RILs not linked to *mIs13*. Pink lines indicate copy number estimates of the MY1 parental strain.

4.3.2 Near isogenic lines in the MY1 background

RILs are a valuable resource for testing genotype-phenotype interactions. However, determining if rDNA copy number affects a trait using RILs requires phenotyping a large number of strains. In addition, unless we want to test for epistasis with the rDNA, RILs may not be the best application for one-dimensional QTL analysis because we already have a locus of interest: the rDNA. A better resource to screen for phenotypes influenced by rDNA copy number would be strains that differ only in rDNA copy number. Our lab previously tried to induce rDNA copy loss by multiple methods, including application of stresses and targeted cutting of the rDNA array, but was unsuccessful. Instead, I harnessed the rDNA arrays of variable size that already exist among strains of *C. elegans* by repeated backcrossing to make introgressed strains with differing rDNA copy numbers in a set genetic background. Using the MY1xSEA51 RILs as a starting point, I performed additional backcrosses to introgress the *mIs13*-linked rDNA array into the MY1 genetic background, thus producing lines with both high and wild type rDNA copy number in a set genetic background, here termed the "MY1 NILs".

I validated rDNA copy number in the MY1 NILs by CHEF gel and found one NIL that had spontaneously reduced rDNA copy number to 64 copies, and a line with the expected 130 rDNA copies (**Figure 4.3, Figure A3.3**) (Morton *et al* 2021, unpublished). I genotyped the NILs by short-read sequencing, and in addition to the *mIs13* transgene, these NILs have approximately 1 Mb of the laboratory strain N2 genotype linked to the rDNA (**Table 4.1**). These MY1-background NILs provide the opportunity to discern whether a strain adapted to a high rDNA copy number shows fitness defects with a lower number of rDNA copies. We currently do not know why *C. elegans* wild isolates have differing rDNA copy numbers. These copy numbers may be adaptive to the

genetic background or may have arisen by chance. The wild isolates we work with in the lab were bottlenecked to a single worm at least once prior to being frozen and assigned a wild isolate name, so it is possible the copy numbers are a chance event [311]. A disadvantage of these NILs is that in addition to the change in rDNA copy number, the NILs have the *mIs13* transgene, providing a confounding variable. Nevertheless, these NILs provide a resource to test how reducing rDNA copy number 3- to 6-fold in a set genetic background affects fitness and aging traits.

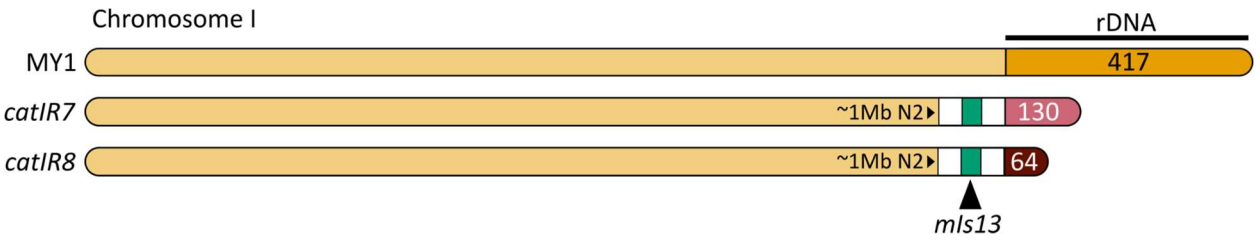

Figure 4.3: Schematic representation of chromosome I genotype of MY1-background NILs. Allele designations for the introgressed region of each NIL are indicated to the left, and the relative amount of N2 genotype linked to the rDNA is indicated for each NIL in white. The GFP-containing *mIs13* transgene is indicated in green. The parental MY1 wild isolate is indicated for reference, with approximately 417 rDNA copies. The *catIR7* and *catIR8* alleles were produced from the same backcross and the strains are siblings; the lower rDNA copy number of *catIR8* arose from a chance loss event.

4.3.3 Near isogenic lines in the N2 background

The ideal set of strains to study rDNA copy number variation would be multiple strains with high or low rDNA copy number in the N2 laboratory strain background. NILs in the N2 background are advantageous for two reasons: N2 is the most well-studied *C. elegans* strain, and with a native copy number of ~100, permits the assessment of both increasing and decreasing rDNA copy number. I independently introgressed the rDNA arrays of four wild isolates into N2 (**Table 4.1**). The high rDNA copy number arrays are donated from wild isolates MY1 and RC301, which have 417 and 420 rDNA copies, respectively. The low rDNA copy number arrays are donated from wild isolates JU775 and MY16, which have 81 and 73 rDNA copies, respectively. For three of the four rDNA arrays, I first generated strains with >1Mb wild isolate genome sequence that remained linked to the rDNA. I further refined these strains through additional backcrosses, screening SNPs that differ between N2 and each wild isolate at sites proximal to the rDNA to select recombinants of interest. Important to note is that there are no differences in rDNA sequence between any of the wild isolates used and the laboratory strain, so all differences in SNPs and INDELs described occur in non-rDNA genomic sequence. **Figure 4.4** and **Table 4.1** detail the rDNA copy numbers and remaining linked wild isolate genotypes of each NIL. The rDNA copy numbers were verified by CHEF gel (**Figure A3.4**). There is some fluctuation in estimates between different CHEF gel runs [5]. For consistency, I will refer to the rDNA copy numbers of the NILs that they or their parents have been assigned in previous work, barring drastic gains or losses of copies that cannot be accounted for by technical error (Morton *et al* 2021 unpublished and [5]). In contrast to the MY1 NILs, these NILs do not have the *mIs13* transgene. The only

confounding variables in the N2 NILs are the regions of wild isolate DNA linked to the rDNA, which can have up to 4,000 SNVs and small INDELs.

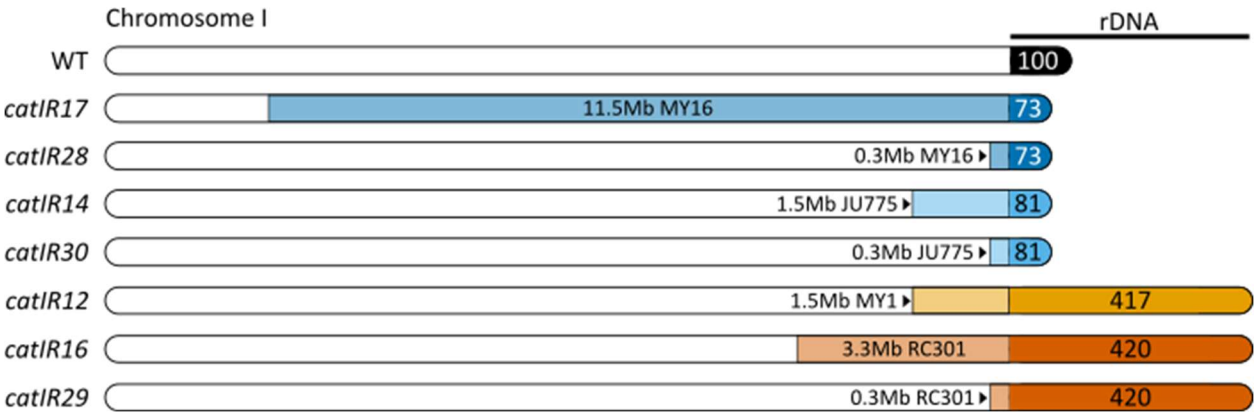

Figure 4.4: Schematic representation of chromosome I genotype of N2-background NILs. Allele designations for the introgressed region of each NIL are indicated to the left, and the relative amount of wild isolate DNA and the rDNA copy numbers are indicated for each NIL. Wild type N2 (WT) is indicated for reference, with no linked wild isolate DNA and approximately 100 rDNA copies. While the rDNA is colored to indicate the wild isolate parent of origin, the consensus sequence of the rDNA repeat unit is identical among all wild isolates and the laboratory strain.

Table 4.1: Near Isogenic Lines of *C. elegans*.

Strain	Genotype	rDNA copy number	Remaining linked SNVs*	rDNA designation
SEA295	*catIR7* [MY1 *mIs13* 64-rDNA; chrI:14170793-end N2]	64	27	MY1 64-rDNA
SEA296	*catIR8* [MY1 *mIs13* 130-rDNA; chrI:14170793-end N2]	130	27	MY1 130-rDNA
SEA300	*catIR12* [MY1 chrI:13527418-end]	417	51	417-rDNA
SEA302	*catIR14* [JU775 I:13529383-end]	81	2,485	81-rDNA
SEA304	*catIR16* [RC301 I:11737213-end]	420	87	420-rDNA
SEA305	*catIR17* [MY16 I:3576802-end]	73	4,192	73-rDNA
SEA328	*catIR28* I [MY16 I:14764185-end]	73	12	73-rDNA
SEA329	*catIR29* I [RC301 I:14989978-end]	420	0	420-rDNA
SEA330	*catIR30* I [JU775 I:14775350-end]	81	220	81-rDNA

*Number of variants from GATK pipeline, ignoring blacklisted regions. While this result shows no variants in SEA329, RC301 genotype was confirmed at position 14989978 as part of the strain construction process. For MY1 NILs, listed is the number of N2 variants linked to the rDNA. Values are approximate. SNVs enumerated are in non-rDNA genomic sequence; no SNVs are found in the consensus sequence of the rDNA repeat unit.

4.4 METHODS

Worm husbandry and strain construction

C. elegans were grown at 20°C on NGM seeded with OP50 bacteria unless otherwise specified. Male stocks of worms were made by heat shocking L4 hermaphrodite worms at 30°C for four hours, then returning plates to 20°C for growth and reproduction. Three or four days later, the heat shocked plates were screened for male progeny. For crosses, 10 male worms were mated with 3 L4 hermaphrodite worms on 3cm NGM plates spotted with OP50. Successful mating was ascertained by the presence of approximately 50% male progeny in the F1 generation. In cases where a GFP marker was used, F1 cross progeny were selected based on the presence of the GFP marker. For construction of NILs, when the GFP marker was not used, an rDNA-proximal restriction fragment length polymorphism or INDEL that differed between N2 and the wild isolate of interest was used as a proxy for rDNA copy number (**Tables A3.3, A4.11**).

RIL construction and propagation

Male SEA51 worms were crossed to hermaphrodite MY1 worms and cross progeny were selected by presence of GFP (*mIs13[myo-2p::GFP + pes-10p::GFP + F22B7.9p::GFP]* transgene). The F1s were self-fertilized and 60 GFP+ (130-rDNA) and 60 GFP- (417-rDNA) F2 worms were selected and propagated for 10 generations by single-worm descent, independently, to homozygose the genomes. After 10 generations of single-worm descent, worm populations were bulked to be frozen down as the RILs and for sequencing. The RILs were then propagated an additional 20 generations in bulk. For propagation, worms were grown to gravid adulthood until many embryos were laid. Adult worms were washed off the plates with 1XM9, leaving behind the embryos. A 1cmx1cm chunk of agar (at least 100 embryos) was cut out and propagated to a

new 6cm NGM+OP50 plate. This process was repeated every three days for 20 generations. After 20 generations, worms were propagated to 10cm NGM+OP50 plates and allowed to starve out to freeze down and harvest for genomic DNA preparation.

Genomic DNA Isolation

The Qiagen DNeasy kit (69504) was used for genomic DNA extraction. Worm pellets frozen in ATL buffer were freeze-thawed three times at -20°C and 37°C. 20µL proteinase K was added and the samples were incubated at 56°C for 3hr with occasional vortexing. 4µL 100mg/mL RNase A (Qiagen 19101) was added to each sample and incubated at room temperature for 5 minutes. 200µL AL buffer was added, and the DNA extraction continued as described in the kit protocol. Final DNA was eluted in a total volume of 100µL. DNA concentration was determined by Qubit assay for sequencing library preparation.

Short-read sequencing library preparation and rDNA copy number estimation

Libraries were prepared as in Chapter 2 and as previously published using 10ng DNA as input [5]. For RILs, mixed stage starved larvae were used for gDNA preparation and sequencing libraries were prepared in two batches for each the RILs and propagated RILs. With each library preparation batch, a control sample of SEA51 DNA and MY1 DNA was prepared in the same batch. For MY1 NILs, genotyping was performed from bam files generated to estimate mtDNA abundance in Chapter 5. For N2 NILs, libraries of mixed stage worms were prepared for genotyping of each NIL.

CHEF plug sample preparation, run conditions, and Southern blotting

CHEF plugs were prepared in a similar manner to Chapter 2, with the following changes: Worms were grown up either by placing ten L4 or day 1 adult worms on a 10cm NGM+OP50 plate

and growing to starvation (~5-6 days at 20°C) or a single L4 or day 1 adult worm was placed on a 6cm NGM+OP50 plate and growing to starvation (~5-6 days at 20°C). Starved worms were washed from plates into 15mL conical tubes or 1.5mL Eppendorf tubes, pelleted, and washed several times in 1X M9 buffer. Worms were washed one time in autoclaved double glass distilled water, and the volume of worms and water after removing the supernatant after washing was estimated. An equal volume of molten 1% SeaPlaque GTG agarose (Lonza) at 42°C was added to the worm solution, and approximately 80µl of the solution was pipetted into agarose plug molds (Bio-Rad #1703713), on ice, and allowed to solidify at 4°C for at least 30 minutes. Plug preparation, digestion, and washing then proceeded as in Chapter 2.

For CHEF gel electrophoresis as follows: Plugs were equilibrated in 1X NEB 3.1 buffer either overnight at 4°C, or on the day of digestion by soaking plugs for 1 hour in 1X NEB 3.1 buffer then replacing with fresh 1X NEB 3.1 buffer in a 24-well plate on ice. Approximately ¼ of the plug was cut with a razor blade and transferred to a parafilm-wrapped slide. The rDNA was cleaved from the rest of chromosome I using ~4µl SwaI per plug, which cuts 3927bp upstream of rrn-3.56 and not at all within the rDNA. Plugs were placed in a small humid chamber in a 25°C incubator and incubated for 4 hours before loading into the CHEF gel. CHEF gels were prepared by placing digested plugs and agarose-embedded ladders onto the teeth of gel combs. The following ladders were used: Yeast Chromosome PFG Marker (*S. cerevisiae*) (NEB #N0345; discontinued); *Hansenula wingei* chromosomes (maximum size 3.13 Mb, Bio-Rad 170-3667). 0.8% agarose prepared in 0.5X TBE at 55°C was poured around the plugs and solidified for at least 30 minutes. Gel was transferred to a CHEF gel box (Bio-Rad CHEF DRII) and run with the following conditions: For long rDNA (total size ~1Mb-3.13Mb): 100V for 68hr, 14°C, switch times = 300-900s. For

medium-length rDNA (total size ~225kb-1.1Mb): 165V for 66 hours, 14°C, switch times = 47-170s. After run completion, gel was stained by soaking in ethidium bromide (0.3ug/mL in 0.5X TBE) to visualize the ladders and imaged on a Bio-Rad GelDoc XR+. Southern blotting was performed as described previously in Chapter 2, see Appendix 1 for details. Band measurement and copy number estimate was performed as in Chapter 2, see Appendix 1 for details.

4.5 PROJECT CONTRIBUTIONS

Near isogenic line construction, genomic preparation, and library construction was performed by Ashley Hall. Recombinant inbred line construction, genomic preparation, and library construction was equally shared between Ashley Hall and Elizabeth Morton. Genotyping and data analysis was performed by Ashley Hall. Size verification of rDNA arrays by CHEF gel was performed by Ashley Hall (through Southern blotting step) and Elizabeth Kwan (blot hybridization).

CHAPTER 5. IDENTIFYING A ROLE FOR rDNA COPY NUMBER IN *C. ELEGANS* PHYSIOLOGY

5.1 SUMMARY

A specific role for rDNA copy number in fitness or aging in a multicellular eukaryote has been elusive to study, due to previous resource limitations. The RILs and NILs described in Chapter 4 are a resource of worm strains ideal for testing hypotheses about the effects of rDNA copy number variation on such traits. Of particular interest are fitness traits, due to previous work showing that too few rDNA copies slows growth, and aging traits, due to previously reported rDNA instability with increased age. In addition, reduction of ribosome biogenesis increases lifespan in multiple model organisms. I asked whether the rDNA copy number itself contributed to lifespan, fitness, and gene expression in *C. elegans*. We used the MY1xSEA51 RILs with high or wild type rDNA copy number to map lifespan-affecting loci and identified loci on chromosomes II and IV that contribute to lifespan. rDNA copy number, however, does not detectably contribute to lifespan. To further characterize the role of rDNA copy number on *C. elegans* physiology, we used the NILs with low, wild type, and high rDNA copy numbers to test for an effect on competitive fitness, early life fertility, and the global transcriptome, again finding no effect due to rDNA copy number.

5.2 INTRODUCTION

Repetitive DNA contributes substantially to genome variability but is understudied when compared to single nucleotide variation [208,209]. A major type of repetitive DNA is the ribosomal DNA (rDNA), encoding the rRNA genes. The rRNAs are the most abundant transcripts

in eukaryotic cells, and to provide enough template to sustain levels of ribosome biogenesis, the rDNA is encoded as a long multicopy tandem array [325]. The repetitive nature of the rDNA leaves it prone to copy number variation [228]. rDNA copy number varies both between individual humans and between strains of laboratory organisms including yeast and worms [1,3,4]. Differences in the global transcriptome and mtDNA abundance have been reported to correlate with differences in rDNA copy number in humans [2]. Beyond these studies, however, few studies consider rDNA copy number variation within the natural range and instead focus on severe reductions in rDNA copy number.

Severe reductions in rDNA copy number restrict ribosome biogenesis, with detrimental effects on fitness, growth, and viability [159–163]. In *D. melanogaster*, restriction of ribosome biogenesis results in the bobbed phenotype, in which flies have many defects, among them a smaller body, slow growth, and reduced fertility [160,164]. With an extreme enough rDNA copy number reduction that ribosome biogenesis is severely restricted, the short rDNA genotype is lethal [159,160,326]. In flies with rDNA reductions that do not restrict ribosome biogenesis, global levels of heterochromatin are reduced and gene expression differences in metabolism genes are observed [6,168]. In *S. cerevisiae*, reduction of rDNA copy number from the wild type ~170 copies to ~35 copies does not negatively affect ribosome biogenesis, but alters genome replication timing and increases sensitivity to DNA mutagens [166]. A less severe reduction, and a copy number represented in wild isolates of *S. cerevisiae*, alters Sir2-mediated silencing at the rDNA, telomeres, and mating-type loci [167]. Beyond these phenotypes, little is known about how moderate changes in rDNA copy number affect an organism.

Changes in the rDNA, including copy number changes, are observed in aging. With regard to ribosome biogenesis, premature aging disorders often have increased ribosome biogenesis, and reducing ribosome biogenesis increases lifespan in many model organisms [8,178,180]. Epigenetic changes at the rDNA locus also occur during aging, and rDNA methylation has recently begun to be used as an epigenetic clock [188,189,191,192,195]. Whether epigenetic changes are causative of aging or a byproduct of the aging process is unknown. With regard to rDNA copy number, rDNA instability is observed in some aging mammalian tissues [200–203]. Further, in *D. melanogaster*, rDNA copies are lost in the male germline stem cells with age, producing heritable rDNA copy number reductions [206]. Whether the inherent rDNA copy number of an organism contributes to aging is unknown. It is possible that having fewer rDNA copies could extend lifespan through restricted ribosome biogenesis, or that having more rDNA copies could extend lifespan through increasing genome stability.

In addition to aging, rDNA copy number differences have been associated with various diseases in humans. rDNA copy number differences in humans must not be so extreme as to severely restrict ribosome biogenesis, or they would not be observed in the population. However, with a range of approximately 200-600 rDNA copies per haploid genome, effects of rDNA copy number on gene expression, chromatin state, or penetrance of disease-associated mutations could be important. Differences in rDNA copy number or in the number of active rRNA genes have been implicated in many complex human diseases, including schizophrenia, intellectual disability, and Alzheimer's disease, among others [210,222,224,226,251,283]. Further, changes in rDNA copy number are found in many types of cancer [94,136,211,213,215–217]. In addition, higher rDNA copy number predicts an increased risk of lung cancer in smokers [211]. There is

overall not a direct understanding of if and how a difference in rDNA copy number directly contributes to phenotype or disease risk in a set genetic background, however. Ascertaining how rDNA copy number affects phenotype in the model organism *C. elegans* will lay the foundation for understanding how rDNA copy number could play a role in human health and disease.

In this chapter, I use the RILs and NILs that I developed in Chapter 4 to ask whether rDNA copy number variation in the natural range affects fitness and aging traits in *C. elegans*. First, I demonstrate that rDNA copy number does not affect steady-state rRNA levels. Then, I demonstrate with the NILs that rDNA copy number does not affect lifespan, competitive fitness, or early life fertility. Minimal changes in the global transcriptome are found with changes in rDNA copy number. Although the NIL data demonstrated that rDNA copy number does not affect lifespan, we were able to identify two lifespan-affecting QTL with our RIL population. Together, these data suggest that rDNA copy number differences within the natural range of *C. elegans* do not have strong phenotypic effects under conditions of laboratory growth.

5.3 RESULTS

5.3.1 There are no large, detectable, statistically significant differences in rRNA levels found between worm strains of differing rDNA copy numbers

Often the first phenotype to be considered when studying rDNA copy number is ribosome biogenesis or rRNA abundance. The simple hypothesis is that by increasing the number of rDNA copies present, there are more templates available for rRNA production, so increasing rDNA copy number will increase rRNA abundance. However, over the minimal threshold of rDNA copies

required to maintain ribosome biogenesis, differences in rDNA copy number do not correspond to differences in rRNA levels in humans [19].

To test whether rDNA copy number differences in the NILs affect rRNA abundance, I measured rRNA levels in synchronized day 1 adult worms by multiple methods. First, I measured the total mature 18S and 28S rRNAs in equal quantities of RNA by TapeStation capillary electrophoresis. The TapeStation method measures a calibrated concentration in ng/µl for each the 18S and 28S rRNAs based on the integrated area of the bands corresponding to each rRNA species as compared to a molecular weight marker. By normalizing total RNA prior to TapeStation analysis, the concentrations of each rRNA species normalized to the molecular weight marker are comparable between samples. By the TapeStation method, there is no detectable, statistically significant difference in 18S or 28S rRNA concentration between any NIL and the laboratory strain (**Figure 5.1BC, Figure A4.1, Table A4.1**). As a secondary measure of 18S and 28S rRNA concentrations, I performed standard RNA gel electrophoresis on equal quantities of RNA for each NIL and laboratory strain sample. I calculated the percentage of each sample that is 18S or 28S rRNA by integrating the sample intensity for each rRNA species and comparing it to the total integrated sample intensity. There are no detectable, statistically significant differences in the percent of total RNA that is 18S or 28S rRNA among the NILs and N2 (**Figure A4.2, Table A4.2**).

In the absence of a substantial difference in mature 18S or 28S transcripts, aberrations in rRNA processing can be detected by measuring the abundance of pre-rRNA processing intermediates [327]. The overabundance of a pre-processed rRNA intermediate could indicate a ribosome biogenesis defect [327]. I quantified the abundance of transcripts that traverse splicing boundaries of the pre-rRNA with RT-qPCR, normalizing to the highly transcribed actin mRNA

(**Figure A4.3**). Transcripts from the 45S pre-rRNA also do not differ in a copy number-dependent manner, but some do fluctuate from strain-to-strain (**Figure A4.3**). Overall, no pre-rRNA regions differ significantly between any NIL and wild type. Finally, I measured levels of the 5S rRNA. While the 5S is encoded separately from the 45S locus, 5S transcripts must still be incorporated into the ribosome in equimolar amounts with respect to the 45S transcripts. There are no differences in 5S rRNA levels among strains tested (**Figure 5.1D**). Together, these results show that there are no measurable differences in rRNA levels, either from the 45S or 5S locus, between NILs with high or low rDNA copy number and the laboratory strain.

While the methods I used to measure rRNA abundance have been used to identify differences of approximately 2-fold in rRNA levels, they likely cannot detect small differences in rRNA abundance because rRNA is the majority of the RNA pool. Indeed, worms with substantial reductions in rRNA abundance have other markers of reduced ribosome biogenesis, including smaller nucleoli and reduced ribosomal protein levels [179]. To rule out an effect of rDNA copy number among the NILs on ribosome biogenesis, thorough analysis of other ribosome biogenesis traits would be required. In addition, measurements of rRNA abundance that have finer resolution, such as measuring rRNA quantities per worm or per cell, could be used to determine if there are any small differences in rRNA abundance among the strains. However, the data presented here show that there are no large, detectable, statistically significant differences in rRNA levels between strains with high- or low- rDNA copy number and the laboratory strain. Therefore, if any further phenotypic differences are identified between the NILs and laboratory strain, the mechanism by which rDNA copy number affects phenotype could be through

mechanisms other than affecting rRNA levels, such as by modulating global chromatin state or genome stability [168,328].

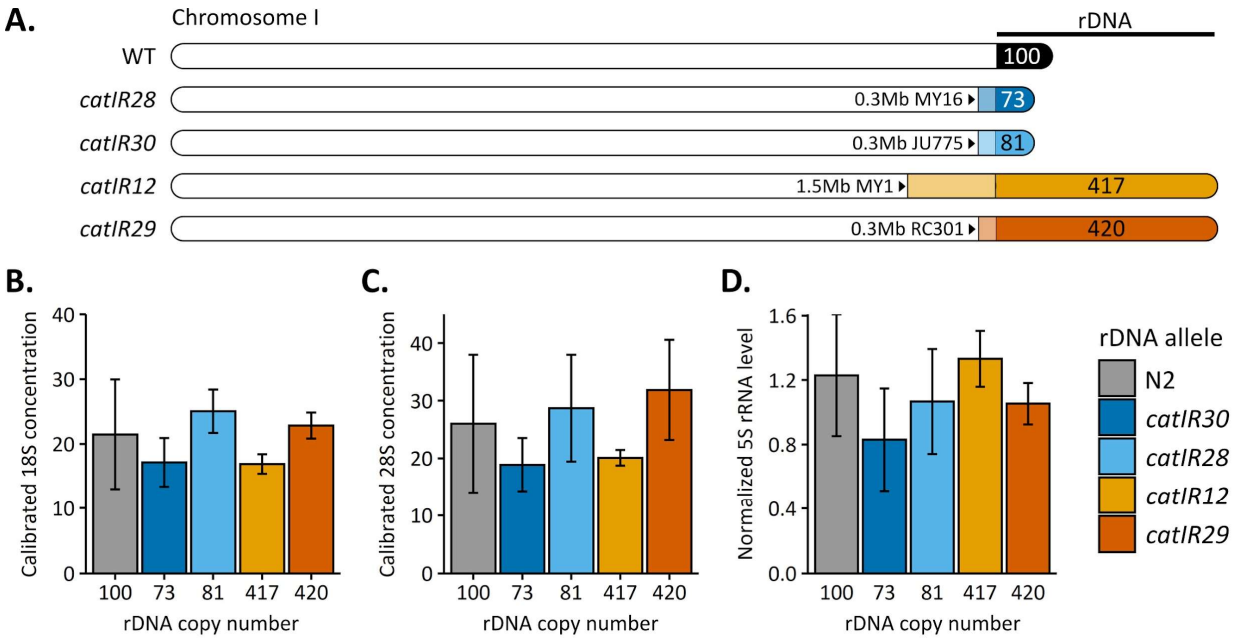

Figure 5.1: Steady-state rRNA levels are equal among NILs with high- and low-rDNA copy number.

A: Schematic of chromosome I of NILs used in this figure. **B**: 18S rRNA levels in NILs measured by TapeStation. No significant differences in 18S rRNA levels between any strains as measured by ANOVA and Tukey's HSD. **C**: 28S rRNA levels in NILs measured by TapeStation. Data are not normally distributed as determined by the Shapiro-Wilk normality test. No significant differences in 28S rRNA levels between any strains as measured by pairwise Wilcoxon test and Benjamini-Hochberg significance adjustment. **D**: 5S rRNA levels normalized to actin in NILs measured by RT-qPCR. No significant differences in 5S rRNA levels between any strains as measured by ANOVA and Tukey's HSD. Legend at the right indicates the rDNA allele for each strain in B-D.

5.3.2 Competitive fitness, likely through an early life fertility defect, is reduced in one strain with high rDNA copy number

When rDNA copy number is reduced, growth defects can occur. In yeast and flies, these growth defects arise when ribosome biogenesis is restricted. In addition, in bacteria, maximal growth rate is determined by rDNA copy number and deletion of rDNA copies compromises competitive fitness [329]. Competitive fitness assays allow for small fitness defects to compound over generations and can be used to ascertain whether any differences in growth, development, or reproduction exist between strains. We performed pairwise competitive fitness tests of NILs and the GFP-marked SEA51 strain. Because low rDNA copy numbers have preferentially been studied in the literature, we focused on increases in rDNA copy number. The *mIs13* transgene that we use to distinguish high- and wild type-rDNA copy number strains causes a fitness defect: on average, the SEA51 strain decreases from 50% to 27.5% of the population when competed against N2 after ~11 generations of propagation (**Figure 5.2, Figure A4.4**). Therefore, we must assess relative competitive fitness to account for this defect.

Overall, we find that increasing rDNA copy number to >400 copies does not confer a consistent fitness defect: The fitness defects are strain-specific. The 417-rDNA strain derived from MY1 has a severe fitness defect: SEA51 outcompetes the 417-rDNA GFP- strain (**Figure 5.2BD, Figure A4.4**). Therefore, the fitness defect associated with this 417-rDNA allele is more severe than the fitness defect associated with the *mIs13* transgene. However, other strains with 420 rDNA copies do not have as severe of a fitness defect. Two different alleles with 420 rDNA copies from wild isolate RC301 compete either near-equally with the GFP control strain or outcompete it to the same extent that N2 does (**Figure 5.2CD, Figure A4.4**). The strains with rDNA

from RC301 differ in the amount of wild isolate DNA linked to the rDNA: *catIR16* has more linked wild isolate DNA and competes similarly to SEA51. It is likely that the reduced competitive fitness of *catIR16* (as compared to N2) comes from a non-rDNA SNV or INDEL native to the RC301 background. Of all of the strains with high rDNA copy number, the 420-rDNA *catIR29* allele performs most similarly to N2, and it also has the fewest SNVs and INDELs linked to the rDNA. Together, these data suggest that competitive fitness defects are strain-specific, and likely caused by non-rDNA variation present in the strain background.

Increasing rDNA copy number does not confer a fitness defect, so we wondered whether decreasing rDNA copy number would compromise fitness. Decreasing rDNA copy number to critically low levels that interfere with ribosome biogenesis slows growth in yeast and flies [163]. The reduced rDNA copy number in the N2 NILs does not change steady-state rRNA levels, and 71 rDNA copies is only about a 30% reduction in rDNA copy number as compared to wild type. To assess whether reduced rDNA copy number affects competitive fitness, we assessed MY1-background NILs with 130 and 64 rDNA copies, which correspond to a reduction of 69% and 85%, respectively. Both MY1-background NILs with 130 and 64 rDNA copies are outcompeted by the wild type MY1 strain to similar levels (**Figure 5.3**). The competitive fitness defect may, however, be due to the presence of the *mIs13* transgene, which causes a fitness defect in the N2 background. Further study of reduced rDNA copy number that can account for the confounding effects of the *mIs13* transgene is necessary to determine if having fewer rDNA copies specifically confers a fitness defect.

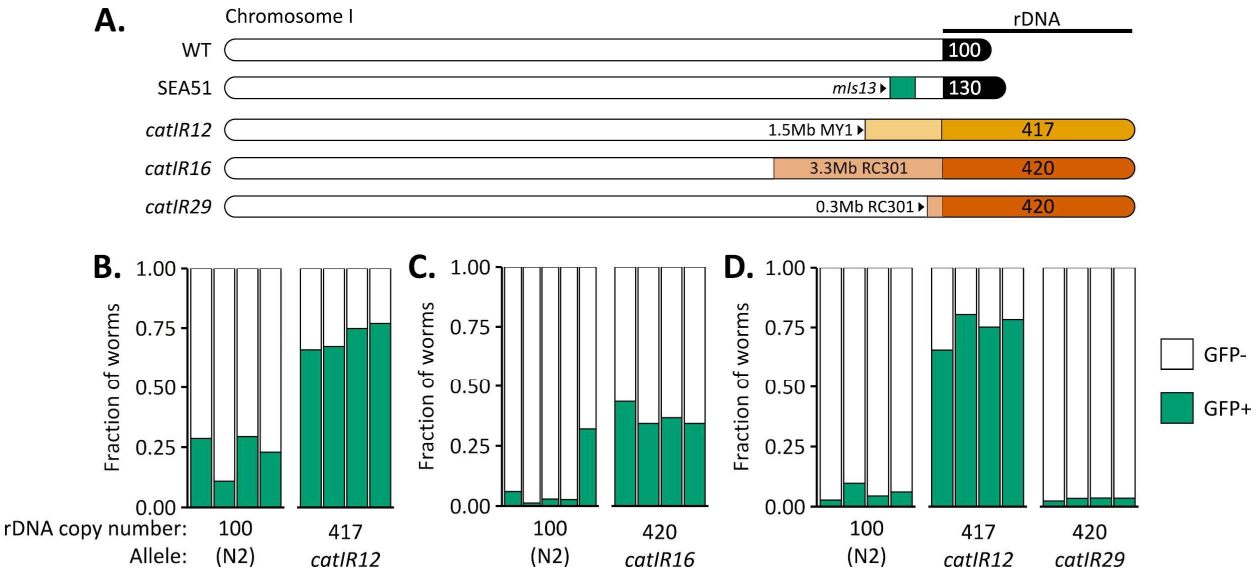

Figure 5.2: Some, but not all, strains with increased rDNA copy number have reduced competitive fitness.

Competitions of high rDNA copy number strains (GFP-; alleles indicated at bottom) against SEA51 (GFP+; 130 rDNA copies). **A:** Schematic of chromosome I of NILs used in this assay. **B:** Four replicates of N2 competed against SEA51 and the 417-rDNA NIL (allele *catIR12*) competed against SEA51, propagated at the same time. **C:** Five replicates of N2 competed against SEA51 and four replicates of the 420-rDNA NIL (allele *catIR16*) competed against SEA51, propagated at the same time. **D:** Four replicates each of SEA51 competed against N2, the 417rDNA NIL (allele *catIR12*), and 420rDNA NIL (allele *catIR29*), propagated at the same time. For all panels, each bar shows the relative proportion of worms after ~11 generations that are GFP+ (green) or GFP- (white). At least 1000 worms were quantified to determine the proportion of each bar.

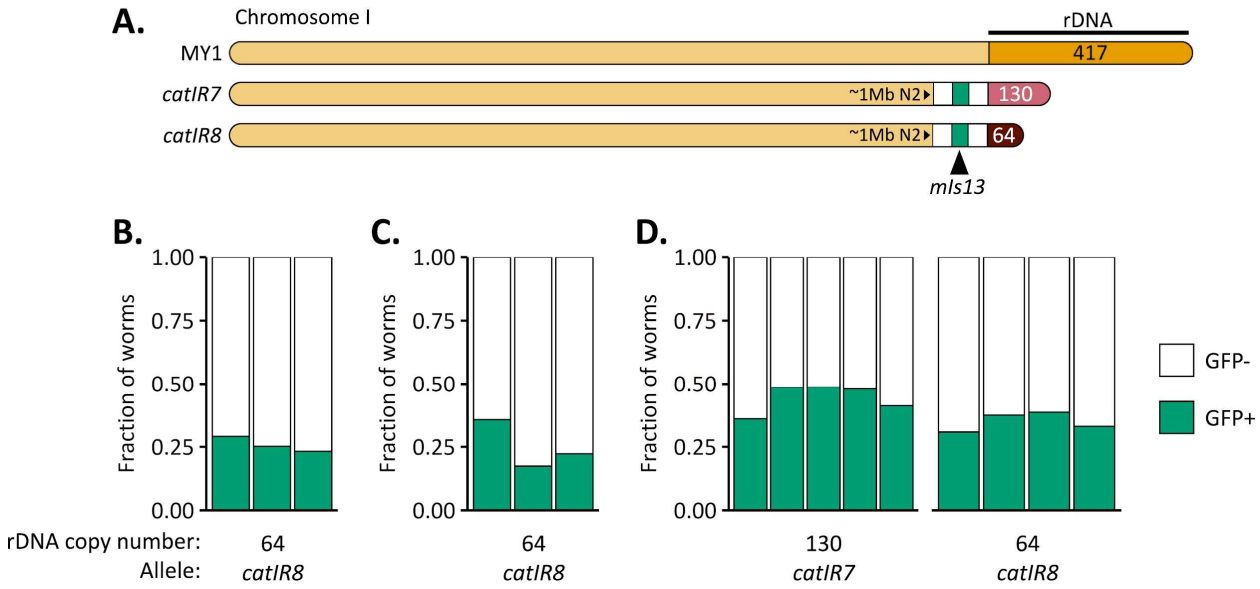

Figure 5.3: In the MY1 background, strains with both lower rDNA copy number and the *mIs13* transgene have lower competitive fitness.

A: Schematic of chromosome I of NILs used in this assay. **B** and **C**: Three replicate plates of MY1 competed against MY1 64-rDNA were propagated simultaneously for each panel. **D:** Five replicate plates of MY1 competed against MY1 130-rDNA and four replicate plates of MY1 competed against MY1 64-rDNA were propagated simultaneously. For all panels, each bar shows the relative proportion of worms after 11 generations that are GFP+ (green) or GFP- (white). At least 1000 worms were quantified to determine the proportion of each bar.

Many different traits can contribute to competitive fitness, including but not limited to rate of development, early life fertility, response to crowding, and recovery from starvation. Due to fluctuations in incubator conditions, for competitions of different strains to the reference strain to be comparable, the different pairwise competitions should be performed at the same time. The competitive fitness assays take nearly two months to complete and are resource intensive, which limits the number of NILs that can be assayed simultaneously. Therefore, I sought an alternative measure for fitness that could be performed on many NILs at once. The trait I selected was early life fertility, a trait that combines both the time it takes to develop to fertile adulthood and the rate of egg production in early adulthood.

In the competition assays, the strains starve out before the full lifetime brood of the adult worms can be laid. The worms are able to lay eggs for approximately one day before they starve. In contrast, under optimal growth conditions, adult worms will lay eggs for approximately three days. I measured early life fertility (approximately 24-30 hours of egg laying) for each NIL, for both the strains with more linked wild isolate DNA and with less (see section 4.3.3). The 417-rDNA NIL (allele *catIR12*) shows a consistent reduction in early life fertility in all five replicates (**Figure 5.4**). However, this fertility defect is likely not due to the rDNA. rDNA alleles *catIR16* and *catIR29* have the same rDNA copy number from the same source strain but differed in competitive fitness. *catIR16*, the 420-rDNA NIL with more linked wild isolate DNA, produces fewer early progeny than N2 (**Figure 5.4D**). Meanwhile, *catIR29*, the 420-rDNA NIL with less linked wild isolate DNA, only produces fewer early progeny than N2 in one of three replicates (**Figure 5.4B**). These differences in early life fertility are consistent with the competition data: The strain with lower relative fitness more consistently produces fewer progeny early in adulthood.

These data suggest that a non-rDNA variant in the chromosome I ~11.5-14.9 Mb region is responsible for the early fertility defect of the *catIR16* 420-rDNA NIL.

With regard to strains with reduced rDNA copy number, there are also no consistent early life fertility defects attributable to the rDNA. *catIR14*, with 81 rDNA copies from JU775, consistently lays fewer progeny than N2 (**Figure 5.4D**). However, *catIR30*, with the same rDNA copy number from the same wild isolate, is not consistently different from N2: In one replicate, it lays more progeny, in a second, it lays fewer progeny, and in a third, it lays the same amount (**Figure 5.4B**). Finally, the strains with 73 rDNA copies from MY16 are the least different from N2: For each allele, there is one replicate in which the 73-rDNA NIL produced fewer offspring than N2, and one or two replicates where it produced the same number of offspring as N2 (**Figure 5.4BD**). Overall, low rDNA copy number does not associate with a change in early life fertility, and fertility defects of strains with reduced rDNA are likely due to linked non-rDNA variation.

Beyond a defect in early fertility, we asked if changing rDNA copy number changed total brood size. The consistent early fertility defect of the *catIR12* 417-rDNA allele could be due to it producing fewer offspring overall. Alternatively, the strain could be slower to develop or it could lay eggs at a lower rate. In the latter two cases, no change in total lifetime fertility would be expected. When measuring total lifetime fertility, we find no difference in brood size between the wild type strain and the 417-rDNA strain (**Figure 5.5**). We also asked whether reducing rDNA copy number in the MY1 background, which also reduces competitive fitness (in the presence of transgene *mIs13*), changed total brood size, and found no difference (**Figure 5.6**). Therefore, the early life fertility defect of the 417-rDNA NIL is most likely caused by either a lower rate of egg laying or by a slower developmental timing.

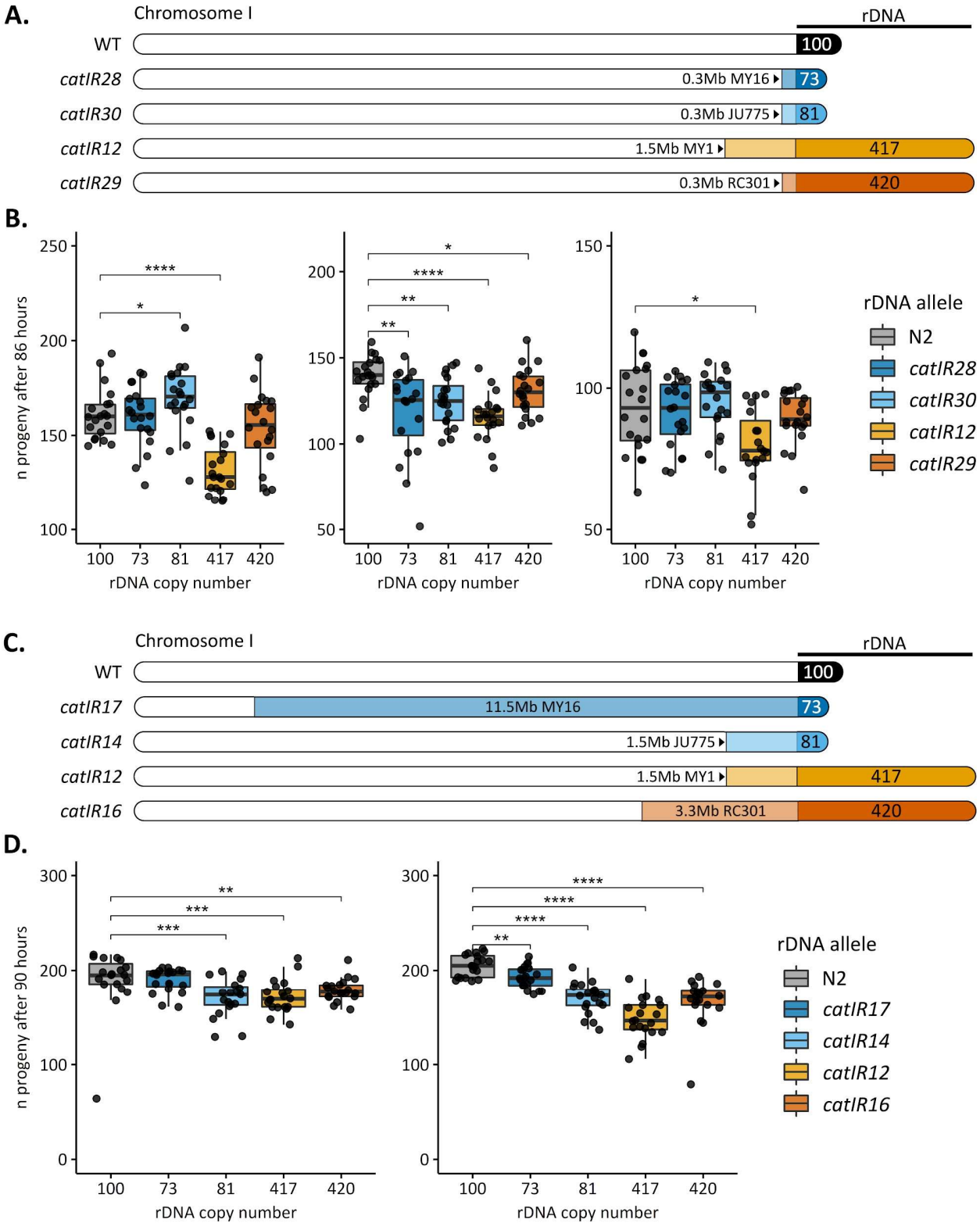
A.
Chromosome I
rDNA
WT
catIR28
catIR30
catIR12
catIR29
100
0.3Mb MY16
73
0.3Mb JU775
81
1.5Mb MY1
417
0.3Mb RC301
420
B.
n progeny after 86 hours
rDNA copy number
100 73 81 417 420
rDNA allele
N2
catIR28
catIR30
catIR12
catIR29
C.
Chromosome I
rDNA
WT
catIR17
catIR14
catIR12
catIR16
100
11.5Mb MY16
73
1.5Mb JU775
81
1.5Mb MY1
417
3.3Mb RC301
420
D.
n progeny after 90 hours
rDNA copy number
100 73 81 417 420
rDNA allele
N2
catIR17
catIR14
catIR12
catIR16

Figure 5.4: Defects in early life fertility do not associate with differences in rDNA copy number. **A:** Schematic of chromosome I of the NILs used for assay in panel **B**. These strains, on the whole, have a lower amount of wild isolate linked to the rDNA than the strains used in panel **D**. **B:** Three replicates of early life fertility assay for NILs diagrammed in panel **A**, n=20 individual worms per strain per replicate. The data fail the Shapiro-Wilk normality test and are not normally distributed. A nonparametric Scheirer Ray Hare test of Progeny by Strain and Replicate shows a significant effect of Strain ($p=0.00137$) and Replicate ($p<0.00001$) across the three replicates of the NILs with less linked wild isolate DNA. Due to the lack of normality and the significant effect of replicate, strain-by-strain comparisons were performed separately for each replicate. A summary of the output of the statistical tests can be found in Appendix 2. **C:** Schematic of chromosome I of the NILs used for assay in panel **D**. These strains, on the whole, have a higher amount of wild isolate linked to the rDNA than the strains used in panel **B**. **D:** Two replicates of early life fertility assay for NILs diagrammed in panel **C**, n=20 individual worms per strain per replicate. The data fail the Shapiro-Wilk normality test and are not normally distributed. For both panels **A** and **B**, statistical tests represented in the figure are Pairwise Wilcoxon tests with Benjamini-Hochberg were performed to compare strains. * $p<0.1$, ** $p<0.05$, *** $p<0.01$, **** $p< 0.001$.

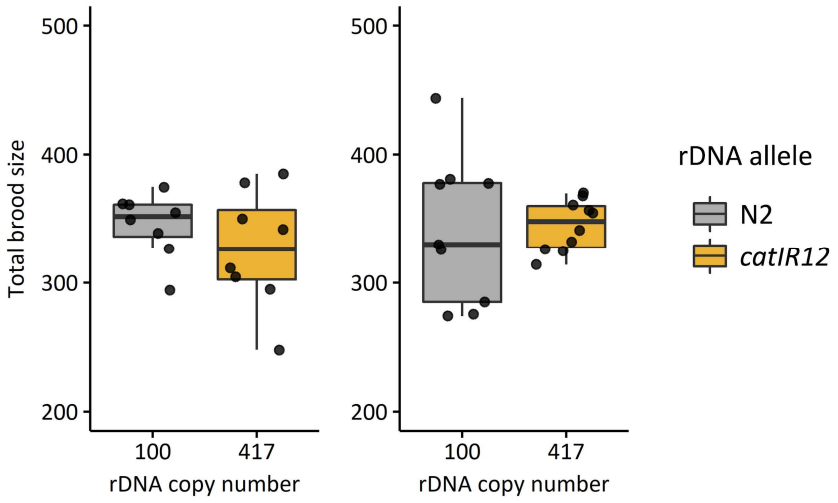

Figure 5.5: Total brood size is unchanged in the 417-rDNA NIL.
Two independent replicates of total fertility measured in N2 and 417-rDNA NIL (allele *catIR12*). Each replicate started with n=10 worms per strain, worms that died or were lost before the end of fertility were censored. No significant difference in total fertility is observed between the two strains with student t-test.

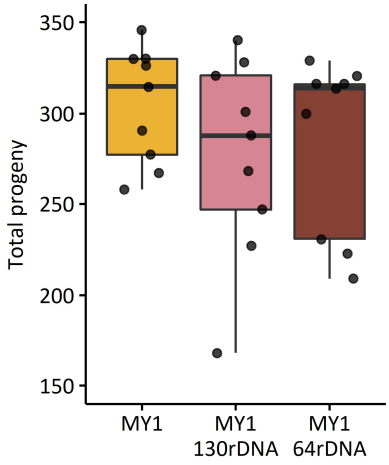

Figure 5.6: Total brood size does not differ between MY1 and MY1-background NILs with reduced rDNA copy number and the *mIs13* transgene.
Total fertility was measured in MY1 (417 rDNA copies per haploid genome) and MY1 NILs (130 and 64 rDNA copies per haploid genome and *mIs13* transgene), with a starting n=10 worms for each NIL. Data are not normally distributed (by Shapiro test), and there are no significant differences between total brood sizes of any strains (Pairwise Wilcox Test with Benjamini Hochberg significance adjustment).

5.3.3 Loci on chromosomes II and IV, but not the rDNA, affect lifespan

Lifespan is a highly multigenic trait. Changes in the rDNA have been tied to aging, leading us to ask whether rDNA copy number itself contributes to lifespan. While our NILs do not differ in rRNA expression, the differences in rDNA copy number could affect lifespan through mechanisms other than ribosome biogenesis. For example, higher rDNA copy number may promote genome stability [328,330]. For lifespan, the combined use of RILs and NILs allows us to test whether rDNA copy number affects lifespan as a single additive locus, or if it interacts with other lifespan-affecting loci as a modifier. In addition, with the RILs, we have the ability to identify non-rDNA loci that are associated with lifespan differences.

To perform lifespan analysis of 118 worm strains, we used the WormBot, an automated lifespan robot [331]. Thirty worms are placed in each well of a 12-well plate, and each well is imaged every 10 minutes for the duration of the lifespan assay (set to 30 days). Loss of spontaneous movement is used as the proxy for time of death. We tested the WormBot on the parental strains SEA51 and MY1, as well as two worm mutants with a known reduction in lifespan (*hsf-1*(*sy441*)) and a known increase in lifespan (*daf-2*(*e1370*)) [332]. Measurable lifespan differences are observed in these control strains when using WormBot (**Figure 5.7**), lending confidence to our ability to use this technology to identify lifespan differences in the RILs. In addition, we confirmed that the *mIs13* transgene does not confer a lifespan defect (**Figure A4.5**). Across the RILs, lifespan varies from 14 to 28.6 days, compared to average median lifespans of 21.5 and 24.8 days for the parental strains SEA51 and MY1, respectively.

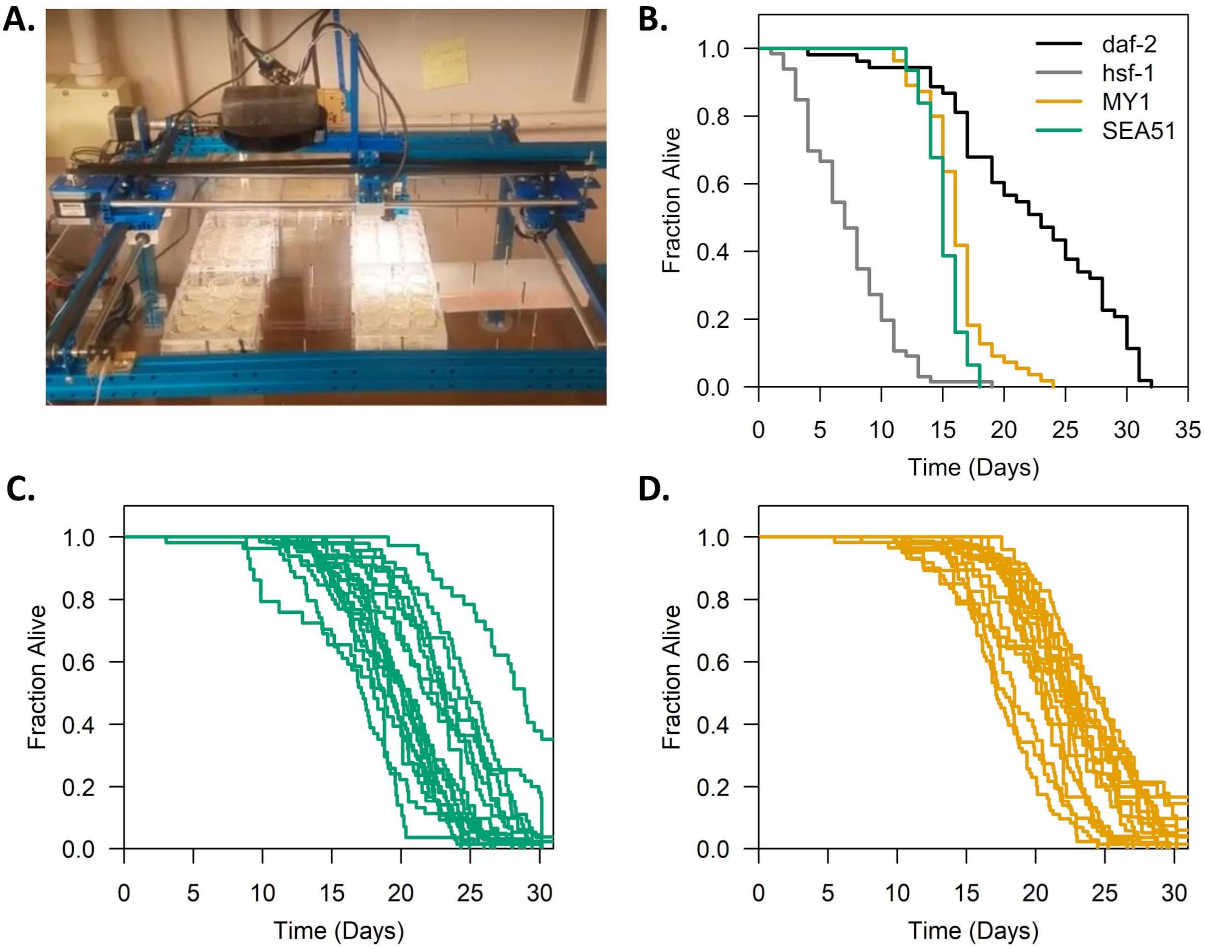

Figure 5.7: Lifespan analysis using WormBot.

A: A photograph of the WormBot holding four 12-well plates. **B:** Trial lifespans using WormBot to verify that known lifespan differences in known mutants with reduced lifespan (*hsf-1(sy441)*) and increased lifespan (*daf-2(e1370)*) could be identified. This assay was performed at room temperature. **C** and **D:** Example lifespans of RILs with 130 rDNA copies (**C**) and 417 rDNA copies (**D**) all performed in the same "Flight" of WormBot lifespan analysis. Median lifespan ranges from 17.2-28.9 days with 22-73 worms scored per strain for **C** and 17.1-24.5 days with 17-81 worms scored per strain for **D**. Lifespan measurements in C and D were performed in a 20-degree temperature controlled room and experiment was terminated at 30 days.

I performed a one-dimensional QTL analysis on the median lifespans of the RILs and identified loci on chromosomes II and IV that significantly associate with median lifespan (**Figure 5.8**), with the MY1 genotype at these loci being associated with longer lifespan. The rDNA locus on the right end of chromosome I is not significantly associated with median lifespan. To investigate whether either the chromosome II or IV QTL may epistatically interact with the rDNA, I performed separate one-dimensional QTL scans on the RILs with high and low rDNA copy number separately (**Figure 5.8B**). In this analysis, the chromosome II QTL is more pronounced among RILs with high rDNA copy number. I then tested for interaction with the rDNA locus by performing a nonparametric epistasis test, which revealed that neither the chromosome II nor the chromosome IV QTL significantly interact with the rDNA (**Figure 5.8C**). To further investigate whether rDNA copy number affects lifespan, we measured lifespan in the N2 and MY1 background NILs. No NILs tested, regardless of strain background or rDNA copy number, differ from wild type in lifespan (**Figures 5.9 and 5.10**). The combined analysis of the RILs and NILs clearly demonstrate that rDNA copy number differences do not affect lifespan in *C. elegans*.

To further investigate the potential of the chromosome II and IV lifespan QTL, we manually measured lifespan on a subset of the RILs for validation. Unfortunately, the manual RIL lifespans were often inconsistent between replicates and were also not consistent with the WormBot lifespans (**Figure A4.5, Table A4.3**). It is possible that the RILs do not differ in lifespan but differ in their duration of youthful adulthood in which they can move freely -- the trait measured by the WormBot. In addition, in the WormBot experiments, many wells had worms burrowing into the agar or were affected by desiccation. Wells severely affected by these problems were manually censored (see **Appendix 4**), but minor problems may still have been

present in the WormBot experiments. Due to time and resources constraints, as well as the inconsistencies between the WormBot and manual lifespans, I did not pursue further validation or fine-mapping of the chrII and chrIV QTL. However, I did analyze the SNVs and INDELs MY1 is reported to have in the chrII and chrIV QTL with the GenAge database (**Table A4.4**) [333]. The genes with the most interesting variants include *hpa-1* ("high performance in advanced age"), *set-26*, *jmjd-2* (jumonji transcription factor domain protein), and *mrpl-37* (mitochondrial ribosomal protein, large). These genes have in-frame or frameshift INDELs that are likely to disrupt function in the MY1 background. Beyond the GenAge database, an additional gene in the chrIV QTL was recently identified as lifespan-promoting in *C. elegans*. *clec-196* encodes a C-type lectin, is located at position ~17.28Mb on chromosome IV, and has four variants in MY1, including two missense mutations and two complex changes in coding exons [1,334]. These genes would be a good starting point as candidates for contributors to lifespan differences between the RILs for future study.

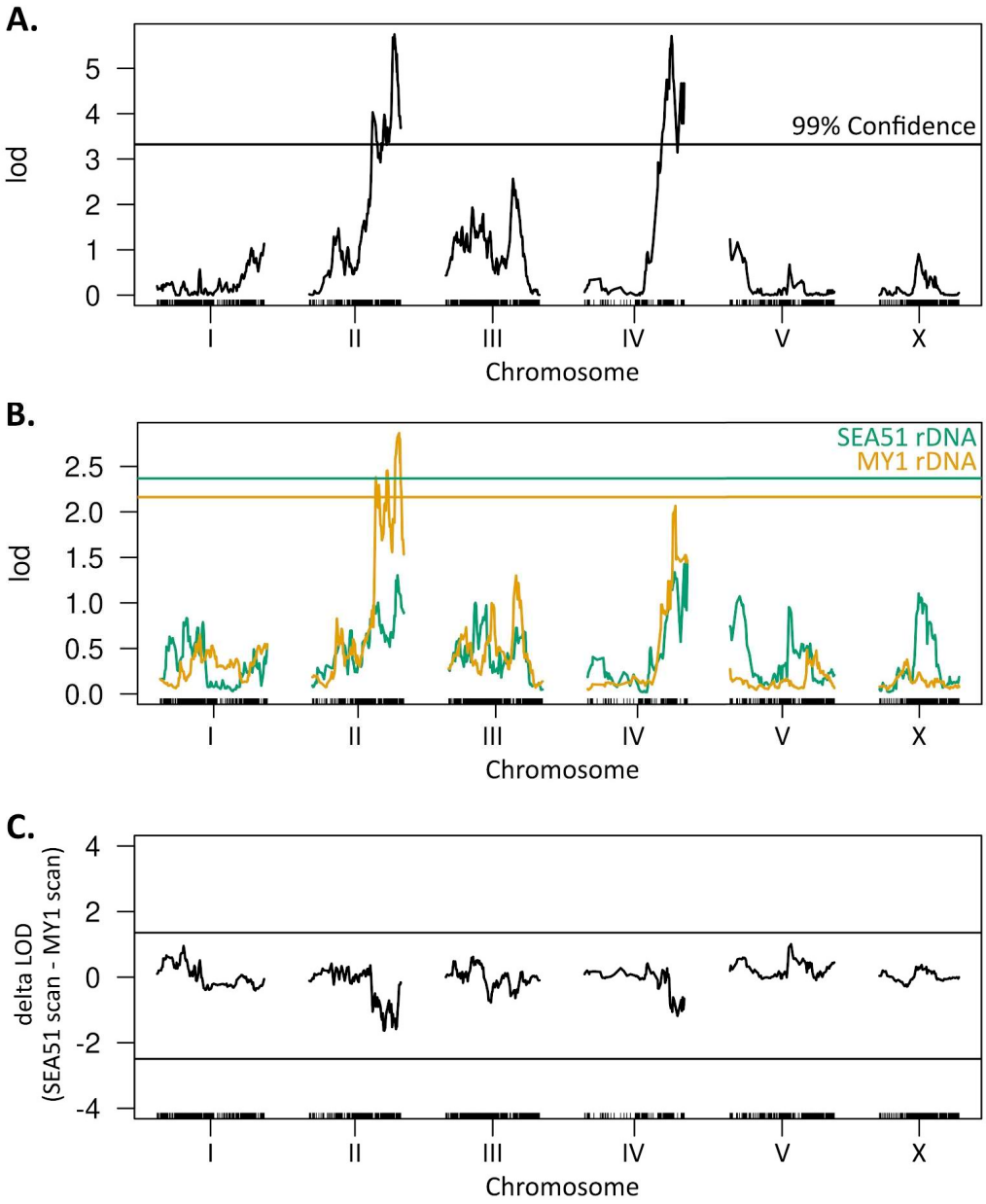

Figure 5.8: Loci on chromosomes II and IV, but not the rDNA, significantly affect median lifespan in the MY1xSEA51 RILs.

A: One-dimensional QTL scan of all genotyped RILs for which median lifespan could be calculated. The horizontal line indicates the 99% significance threshold. **B:** QTL scan stratified by rDNA genotype. Lines with MY1 rDNA (~417 copies) are in goldenrod, lines with SEA51 rDNA (~130 copies) are in green. Colored horizontal lines indicate the 99% significance threshold for each scan. **C:** A nonparametric epistasis test between the rDNA and other loci using the QTL scans shown in **B**. The rDNA is located at the right end of chromosome I. Horizontal lines indicate the 95% significance threshold. All significance thresholds are based on 1000 permutations.

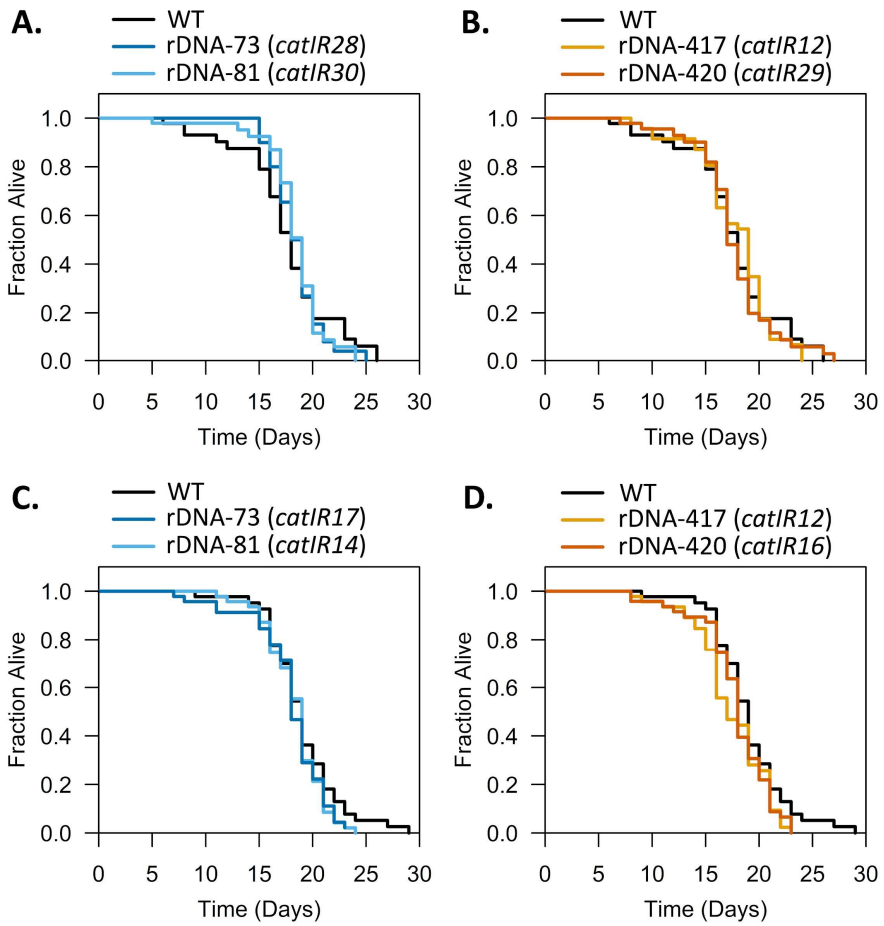

Figure 5.9: Increasing or decreasing rDNA copy number in the N2 background has no effect on lifespan.

A and **B**: Lifespan analysis on N2 and NILs with less linked wild isolate DNA (see Figure 5.4A for schematic). **C** and **D**: Lifespan analysis on N2 and NILs with more linked wild isolate DNA (see Figure 5.4C for schematic). Lifespans for all five strains in **A** and **B** were performed simultaneously (N2 data are the same between the plots) and lifespans for all five strains in **C** and **D** were performed simultaneously (N2 data are the same between the plots).

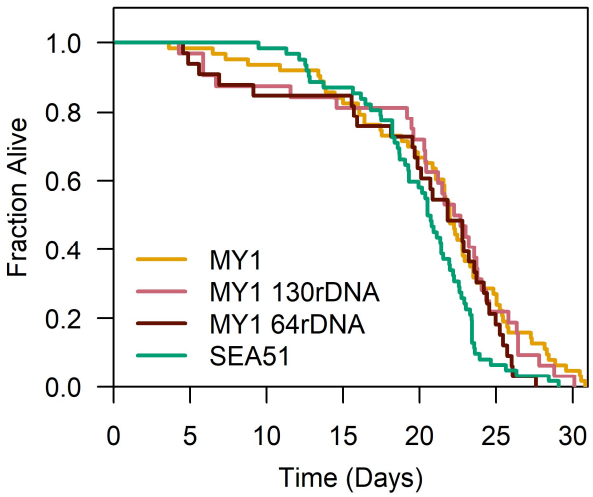

Figure 5.10: Decreasing rDNA copy number in the MY1 background has no effect on lifespan. Lifespan was performed on WormBot at 20°C on NGM with FuDR and Nystatin, seeded with OP50. Median lifespans are as follows: MY1: 22 days, MY1 130-rDNA: 22.5 days, MY1 64-rDNA: 21.8 days, and SEA51: 20.6 days. No significant difference between either MY1 130-rDNA or MY1 64-rDNA and the parental MY1 strain (Log rank test, Bonferroni adjusted P-value). SEA51 and MY1 differ in this replicate, with a Bonferroni adjusted P-value of 0.0277 with a Log rank test.

5.3.4 Mitochondrial DNA abundance and function are not affected by rDNA copy number

The mitochondria and the ribosomes are two of the biggest cellular players in energy metabolism. In humans, it was published that mtDNA abundance inversely correlates with rDNA copy number [2]. In analyzing newer, higher-quality sequencing data (from Chapter 3), I was unable to identify a significant inverse relationship between rDNA copy number and mtDNA abundance (**Figure A5.1**). However, I was also unable to recapitulate the previously reported strong inverse relationship using the same low-coverage data (**Figure A5.1A**). Despite these negative results, the Queitsch and Brewer labs have identified an inverse relationship between mtDNA abundance and rDNA copy number in *S. cerevisiae* (Elizabeth Kwan, unpublished). With these contrasting results, we wondered whether a connection between the mitochondria and rDNA is present in metazoans. I measured mtDNA abundance by sequencing read depth in the MY1 NILs, carrying 417, 64, and 130 rDNA copies in the MY1 background. In *C. elegans*, mtDNA abundance changes with developmental stage: From embryo through L3, mtDNA abundance is constant, but when the worm enters L4 and germline development begins, mtDNA copies per worm increase, leveling off early in adulthood [335]. I measured mtDNA abundance in L1s, L4s, and young adults in the MY1 NILs. I observed the expected increase in mtDNA abundance with developmental stages (**Figure 5.11**). However, I observe no appreciable difference in mtDNA abundance between the strains with differing rDNA copy numbers in any given developmental stage (**Figure 5.11**).

Even without a difference in mtDNA abundance, worms with differing rDNA copy numbers may differ in mitochondrial function. Ethidium bromide (EtBr) inhibits mitochondrial

genome replication, with little if any effect on nuclear genome replication [336]. Growth of worms on plates containing EtBr inhibits development, preventing them from reaching fertile adulthood [335,337]. Despite reports that treatment with EtBr causes *C. elegans* to arrest in the L3 larval stage, I observed severe and variable phenotypes when treating worms with EtBr. As done in previous publications, I binarized the phenotype of the worms on the EtBr plates: Worms that appeared to reach adulthood or were healthy-looking worms in the L4-adult range were scored as "adults", whereas worms that appeared to arrest before reaching L4 or had gross morphological defects were scored as "L3 arrested". On 100 µg/mL EtBr, no worms of any genotype reached adulthood. On 50 µg/mL EtBr, nearly all worms arrested prior to adulthood, while on 25 µg/mL EtBr, 70-80% of worms arrested prior to adulthood, regardless of genotype (**Figure 5.11**). No significant differences between genotypes were observed. Overall, decreasing rDNA copy number in the MY1 background neither affects mtDNA abundance nor sensitivity to mtDNA replication inhibition.

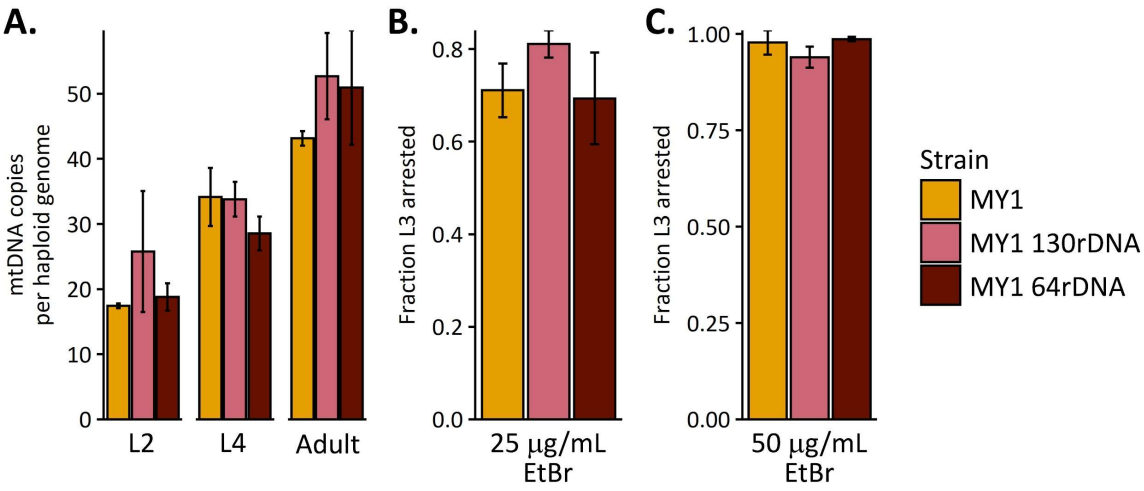

Figure 5.11: Mitochondrial DNA abundance and mitochondrial replication stress survival are not affected by a reduction in rDNA copy number.

A: mtDNA abundance in three developmental stages in the MY1 NILs, calculated from short read sequencing read depth. **B** and **C**: NILs were exposed to ethidium bromide, and 2-3 plates of worms per strain per condition were scored for developmental stage (L3 or earlier = "L3 arrested", indicating impaired mitochondrial function). Error bars represent variation between plates, for each plate, 67-124 worms were scored.

5.3.5 There are few changes in the global transcriptome between worms of differing rDNA copy number

Global transcriptome changes between humans and flies with differing rDNA copy numbers have previously been reported [2,6]. In *D. melanogaster*, these claims were supported by RNA-sequencing data of flies with either wild type or reduced rDNA copy number, with little non-rDNA variation present in the strains [6]. In humans, these claims were supported by RNA-seq of a small number of human cell lines from the 1000 Genomes Project; the rDNA copy numbers of which have since been found to be inaccurate [2,4]. Lines of *A. thaliana* with reduced rDNA copy number also show changes in gene expression, including effects on metabolism and ribosome biogenesis [165]. How rDNA copy number differences affect the transcriptome in *C. elegans* has not been explored, nor has how increasing rDNA copy number in a set genetic background affects the transcriptome. I performed RNA-sequencing on synchronized day 1 adult worms of NILs in the N2 background. I used two NILs with high rDNA copy number and two with low rDNA copy number, selecting strains with the least amount of linked wild isolate DNA (**Figure 4.3.3**). Overall, all worm samples have similar transcriptomes as determined by PCA analysis, and strains with similar rDNA copy numbers are not more similar to one another than they are to strains with different rDNA copy numbers (**Figure 5.12**). In addition, few genes are differentially expressed when comparing any individual NIL to wild type (**Figure 5.13, Tables A4.5-A4.8**).

The strain with the most genes differentially expressed as compared to wild type is the 73-rDNA strain. I performed gene ontology analysis on the genes differentially expressed in the 73-rDNA strain as compared to wild type (**Tables A4.9-A4.11**). Genes involved in cuticle biosynthesis are enriched. In line with the gene ontology analysis, genes associated with

paralyzed, dumpy, movement, and molting phenotypes are enriched (**Table A4.6**), as are genes known to be involved in the epithelial system (**Table A4.7**). There are two possible interpretations for the enrichment of cuticle biosynthesis genes: 1: That there is a difference in cuticle integrity between the 73-rDNA NIL and wild type, or 2: There is a difference in developmental timing of the *in-utero* embryos of the 73-rDNA NIL and wild type. The latter is consistent with the known expression patterns of the genes: the genes differentially expressed between the strains are not expressed in adult worms but are expressed in embryos. We find no difference in cuticle permeability between N2 and the 73-rDNA NIL, supporting the notion that the difference in cuticle gene expression is not corresponding to a difference in adult cuticle integrity (**Figure A4.7**).

Overall, we find that under standard laboratory conditions, rDNA copy number does not globally affect gene expression in day 1 adult worms. We did identify some line-specific patterns, such as the cuticle gene expression in the 73-rDNA NIL. Because these differences were not also identified in the 81-rDNA NIL, it is likely that these differences are influenced by the linked wild isolate DNA, and not the rDNA copy number. Alternatively, the downshift from 81 to 73 rDNA copies could cause the difference in cuticle gene expression, but we consider this unlikely, due to how similarly the strains behave in all other assays.

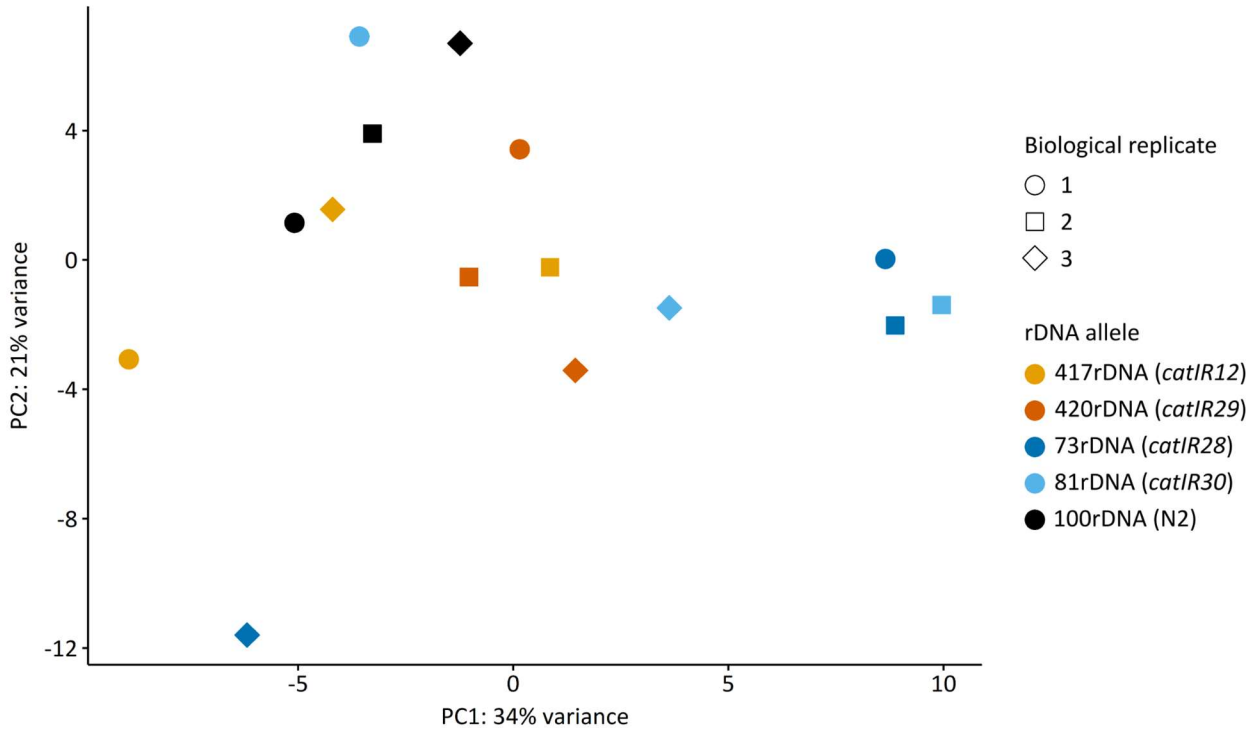

Figure 5.12: Principal component analysis of N2 NIL RNAseq datasets.
The rlog transformed DEseq dataset was analyzed by principal component analysis in the DEseq package. Samples are colored according to the strain (rDNA allele) and the biological replicate is indicated by the shape.

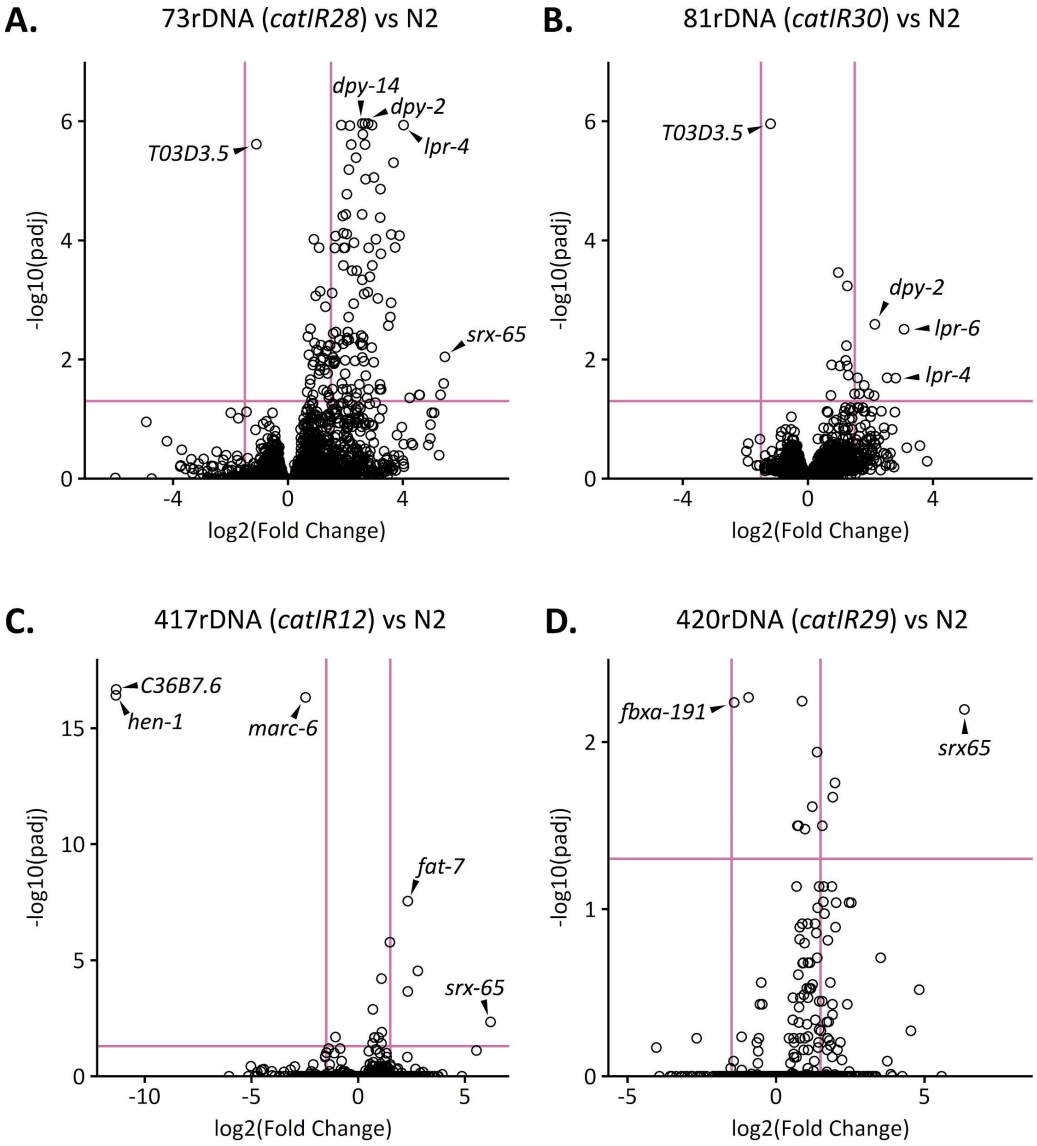

Figure 5.13: Few genes are differentially expressed in individual NILs as compared to N2.
A-D: Three biological replicates for each NIL and N2 were used to calculate differential expression of all genes using DESeq2. All worms assayed were synchronized day 1 adult worms. Horizontal pink line indicates an adjusted p value cutoff of 0.05, vertical pink lines indicate log2 fold changes of 1.5. Genes whose expression differences correspond to a large fold-change or that are highly significant are identified.

5.4 DISCUSSION

5.4.1 In *C. elegans*, variation in the natural range of rDNA copy numbers can largely be ignored when worms are grown under standard laboratory conditions

The data presented in this chapter demonstrate no measurable phenotypic consequences to changing rDNA copy number in the laboratory strain background, among the assays performed. We show that rDNA copy number does not affect steady-state rRNA levels, competitive fitness, early life fertility, or mtDNA abundance. In addition, in contrast to previous studies, we show negligible effects of rDNA copy number on the global transcriptome [2,6]. Further, we show that rDNA copy number does not affect lifespan, and instead identify loci on chromosomes II and IV that do. While we did not test all possible rDNA copy numbers or all possible phenotypes, our data support the conclusion that under standard laboratory conditions, rDNA copy number variation largely does not affect *C. elegans* phenotypes.

We restricted our analysis of rDNA copy numbers to those naturally occurring in *C. elegans* wild isolates. However, the wild isolate strains we used were from a small set of strains of which we verified their rDNA copy numbers by CHEF gel electrophoresis. There are hundreds of *C. elegans* wild isolates for which we have not estimated rDNA copy numbers by CHEF gel, some of which are predicted to have rDNA copy numbers of down to 35 by short-read sequencing [311]. It is therefore possible that *C. elegans* wild isolates with higher or lower rDNA copy numbers than we tested exist. Furthermore, it is also possible that rDNA copy number must be pushed to extremely high or extremely low levels before phenotypic consequences arise. If so, engineering the rDNA loci in *C. elegans* may be necessary to achieve higher or lower rDNA copy numbers than described here.

Beyond expanding the range of rDNA copy numbers, it is also possible that the repertoire of phenotypes considered needs to be expanded to find one affected by rDNA copy number. Recent literature has explored the consequences of accumulation of antisense ribosomal siRNAs (risiRNAs) and the interaction of this accumulation with rDNA copy number. In *C. elegans*, risiRNAs promote the degradation of erroneous rRNAs and the reduction of pre-rRNA levels [338–341]. Further, risiRNAs inhibit RNA Pol I transcription through the nuclear RNAi defective (NRDE) pathway [341]. Because risiRNAs regulate rRNA expression, it is possible that equal levels of steady-state rRNAs are achieved in the NILs with high and low rDNA copy number by differing levels of risiRNAs. Furthermore, risiRNAs accumulate in a *C. elegans prg-1* mutant [342]. *prg-1* is the *C. elegans* Piwi homolog, and mutation of *prg-1* eliminates piRNA production and results in a transgenerational loss of fertility [343]. The *prg-1* mutant can be partially rescued by increasing rDNA copy number, perhaps due to a greater capacity to produce rRNA [342].

In addition to the risiRNAs, rDNA copy number may be connected to the RNAi pathway. Mutations in the RNAi pathway partially rescue the transgenerational sterility defect of the *prg-1* mutant, and these mutations often co-occur with increases in rDNA copy number [342]. Interestingly, in fungi, protists, and flies, mutation of Dicer, the primary endoribonuclease involved in RNAi, results in loss of rDNA copies or rDNA instability [344–347]. In *S. pombe*, the Dicer activity required for rDNA copy number maintenance is independent of its role in producing short dsRNAs for RNAi [346]. Overall, it is thought that the rDNA instability found in RNAi mutants is caused by increased genome instability [348]. Interestingly, in *C. elegans*, a Dicer mutation did not associate with a change in rDNA copy number, but increasing rDNA copy number in the presence of a Dicer mutation led to deleterious effects [342]. Whether the deleterious interaction

of Dicer mutation and increased rDNA copy number is part of the reason that other organisms lose rDNA copies in a Dicer mutant background is unknown.

5.4.2 Differences observed between the 417-rDNA NIL and wild type

When consistent differences were observed between a NIL and N2, those differences were observed in the 417-rDNA NIL (allele *catIR12*), which has rDNA from the wild isolate MY1 with approximately 1.5Mb of linked wild isolate DNA. The 417-rDNA NIL displayed consistent defects in early life fertility and competitive fitness, but no defect in total brood size. The specificity of the fertility defect to early adulthood suggests that this strain may take slightly longer than N2 to develop to fertile adulthood or may lay eggs at a slightly lower rate than N2. If these defects were specifically due to having a higher rDNA copy number than wild type, the other NILs with high rDNA copy number should have displayed the same phenotypes.

The defects observed in the 417-rDNA NIL are therefore more likely due to non-rDNA variants that are specific to that strain. There are four unique missense variants in the 1.5 Mb region proximal to the rDNA that are present in the MY1 genotype (the parent of the 417-rDNA NIL) but not the RC301 genotype (the parent of the 420-rDNA NILs) (**Table 5.1**). The most interesting variant is in the DMAP1 domain of *ekl-4*, an ortholog of human DMAP 1 (DNA methyltransferase 1 associated protein 1). A knockout of *ekl-4* produces a partially penetrant lethal phenotype, whereas RNAi of *ekl-4* causes multiple phenotypes including slow growth and lethality [349]. To determine if this variant causes the fitness defect of the 417-rDNA NIL, we could repair this mutation to wild type in the NIL background and introduce this mutation in the N2 background. This approach would allow us to both determine if returning the variant to wild

type in the NIL background rescues competitive fitness, and if introducing the variant into N2 induces a competitive fitness defect.

Table 5.1: Variants with predicted high impact present in MY1 and RC301 in the ~1.5 Mb proximal to the rDNA.

Gene	Amino Acid	Base pair*	Wild Isolate
W04A4.6	79V>79A	13687336A>G	RC301
Y105E8A.32	50R>50L	14378647C>A	MY1
ekl-4	233R>233K	14470756G>A	MY1
Y105E8A.20	305S>305L	14488002G>A	MY1
Y105E8A.20	423S>423L	14488002G>A	MY1
tag-4	213R>213W	14989978A>T	MY1, RC301
F31C3.3	377R>377Q	15049340C>T	MY1

*Base pair positions correspond to the WS276 genome and variants and their effects were identified with the CeNDR variant effect browser [311].

5.4.3 Future applications of the RILs

While the NILs are a more manageable resource for testing many phenotypes in the future, the RILs can also provide a future resource. The RILs can be used for both identification of variants that differ between MY1 and N2 that cause trait variation, as well as for epistatic testing of how rDNA copy number influences that variation. If, for example, rDNA copy number alone does not affect rRNA abundance, perhaps rDNA copy number combined with a variant in an RNA polymerase I subunit would affect rRNA abundance. Further, in assessing lifespan of the RILs, we noticed a number of wells that produced progeny even in the presence of FuDR (see **Appendix 4**). Eggs made before application of FuDR can still hatch, so worms that produced escaper progeny were likely further along developmentally [350]. In addition, MY1 develops to

adulthood more quickly than N2 does [351]. So, the RILs could be a useful resource to identify loci that contribute to differences in developmental timing in MY1 as compared to N2.

5.5 METHODS

Worm husbandry

C. elegans were grown at 20°C on NGM seeded with OP50 bacteria unless otherwise specified.

Worm synchronization by hypochlorite treatment ("bleaching")

Gravid adult worms were washed from plates into 15mL conical tubes in 1X M9 and pelleted either by gravity settling or by centrifugation. Worms were lysed with a hypochlorite solution (0.5N NaOH, ~0.8% NaOCl) for no more than seven minutes and embryos harvested by centrifugation (1500 rpm for 1 minute, with slow ramp). Embryos were washed 3x in 10-15mL 1X M9. Embryos were then either plated on food or, for stage synchronization, on unseeded NGM plates overlaid with 1X M9. For stage synchronization, embryos were allowed to hatch in 1X M9 overnight and starve out as L1 larvae. Larvae were harvested by centrifugation (1500 rpm for 1 minute) and washed once in 1X M9 before being plated on food.

Worm assays

For all assays, prior to assay initiation, worms were maintained on NGM + OP50 plates unless otherwise specified. All assays were performed on worms that were at least three generations removed from starvation, freezing, or contamination (unless otherwise specified).

Competitive fitness assays

Ten L4 worms of each of the two genotypes to be competed were picked to a 15cm NGM High Peptone plate (20g/L peptone) seeded NA22. Per trial, three to five 15cm plates were set

up in this way for a given competition set. Paired trials involved 4-5 plates of one competition set propagated on the same days and same media batches as 4-5 plates of another competition set (ex. Four plates of N2 competed against SEA51 alongside four plates of SEA300 competed against SEA51). The replicate plates were maintained as separate independent propagations. The initial plates were allowed to grow to starvation (~6-7 days), after which a chunk of agar approximately 3cmx4cm was cut out and propagated onto a new NGM High Peptone + NA22 plate. This was done for seven plates subsequent to the first one, with plates reaching starvation between each propagation (note: occasional plates were propagated when only near starvation, if the others in the trial had starved). A final plate propagation was done through liquid transfer: starved worms from the eighth plate were washed off in M9, spun 1500rpm 1 min, washed once in M9, spun 1500rpm 1 min, and brought down to a volume of ~3mL. L1 density was determined and ~12,000 L1 were plated on a final NGM High Peptone + NA22 plate. We approximate the initial plate to starve after 2-3 generations, and each subsequent plate after ~1 generation, thus equally ~11 generations of competition.

Each competition included the test strain SEA51, an N2 background strain with the *mIs13[myo-2p::GFP + pes-10p::GFP + F22B7.9p::GFP]* transgene on Chromosome I conferring GFP fluorescence. We quantified population proportions of GFP positive and negative worms using a COPAS Biosort (Union Biometrica). Two days after L1s were seeded on the final plate, worms were washed off with M9, spun 1500rpm 1 min, washed once in M9, spun 1500rpm 1 min, and brought up to a ~10mL volume with M9. Worms suspended in M9 were quantified for size and fluorescence by flowing through the COPAS, collecting data for at least 1000 worms per plate. Objects were gated for size (time of flight) to enrich for L4/adult. Thresholds of green

fluorescence peak height were used to call GFP presence or absence: a green fluorescence peak height greater than 50000 indicated GFP presence, a green fluorescence peak height of less than 20000 indicated GFP absence. Controls of the parental strains were also quantified by COPAS to confirm accurate GFP detection. Most paired competitions also included COPAS quantification at an early time point (directly after the first plate propagation).

Early life fertility

Worms were synchronized by pulse-laying 10 Day 1 Adult worms for 1 hour onto NGM + OP50. Two days later, worms were singled to 3cm plates. For some assays, 20 individual worms were singled to 3cm plates and all plates were scored (**Figure 5.4D**). For others, 30 individual worms were singled to 3cm plates and 20 plates per strain were randomly selected to be scored (**Figure 5.4B**). The latter implementation was intended to reduce bias in the exact stage of worm selected, as some variation in larval size was present on each pulse-lay plate, despite the short pulse-lay duration. Either 90 hours (**Figure 5.4D**) or 86 hours (**Figure 5.4B**) after initiation of the pulse-lay, adults were removed from the 3cm plates. Progeny were allowed to grow for 48-72 hours before counting. Plates were stored at 4°C when counting over multiple days, to prevent interference from the next generation of worms.

Total hermaphrodite brood size

Worms were transferred to new 3cms every 24 hours, until progeny production ceased. After transfer of the adult worm, progeny were allowed to grow to L4 or adulthood before counting. Plates were stored at 4°C when counting over multiple days, to prevent interference from the next generation of worms. Plates where the adult either died, crawled up the side of the wall, or crawled under the agar were censored.

Cuticle permeability

Cuticle permeability assay was based on two previous publications [352,353]. Briefly, approximately 30 day 1 adult worms were picked to a watch glass slide containing 150µl 1X M9 containing 100mM Levamisole and 10µg/mL Hoechst 33258 dye (Sigma) to simultaneously stain and anesthetize the worms. Worms were incubated in the stain for 30 minutes statically in the dark, then washed 5x with 100µl 1X M9 in the watch glass slide. Worms were picked into a spot of ~10µl 1X M9 on a 2% agarose pad on a microscope slide with an eyelash pick, recovering approximately 20-25 worms per strain. Hoechst stain was visualized with the DAPI filter on a Zeiss Apotome, and the number of worms with nuclear staining of hypodermal cells was quantified.

WormBot lifespan

For WormBot lifespans, RILs were not propagated for 3 generations to remove from starvation or freezing. Instead, RILs were freshly thawed and allowed to starve out. If thawed worms were contaminated, the thaw plate was not allowed to starve and was instead bleached when enough gravid adult worms were present. After the thaw plate starved, worms were propagated to 10cm NGM+OP50 and grown to starvation (4 days later). Starved worms were washed off of the 10cm plate and plated onto 15cm NGM High Peptone+NA22. Worms on 15cm plates were grown to gravid adulthood (3 days), then bleached and embryos allowed to hatch and starve out as L1 larvae on unseeded NGM plates overlaid with 1X M9 overnight. Starved L1s were then harvested in 15mL conical tubes, washed once in 1X M9, then plated on 10cm NGM plates seeded with OP50. Two days later worms were in the L4 stage. For each strain, 30 L4 worms were picked by hand into each of 3 wells of 12-well plates containing NGM + FuDR (50µM) and Nystatin (25µM) seeded with OP50 bacteria.

The arrangement of strains in the 12-well plates was pseudo-random; “seating charts” for the worm strains were made to prevent the same strain from always occupying the same position of the 12-well plate with respect to the edge or middle of the plate. WormBot assay was started by placing all 12-well plates on the WormBot in a 20°C temperature-controlled room.

After the 30 days, each well was manually scored to determine the time of death of each worm by scrolling through images from day 30 through day 1 for each well and identifying the point at which a nonmoving worm began to move again. As such, WormBot lifespan uses loss of spontaneous movement as a proxy for death. WormBot produces individual files for the lifespan for each well to indicate the time of death for each worm. These files were manually compiled in Microsoft Excel, then annotated with additional information regarding the experiment for filtering.

Manual lifespan

For manual lifespans, all worm strains were propagated such that the worms analyzed were 3 generations away from any starvation, freezing, or contamination. 10 L4 worms per strain were picked to each of 5x 3cm plates seeded with OP50, for a total of 50 worms per strain. Assays were performed with variable plate composition: some use NGM, some NGM + FuDR, and some with NGM + FuDR + Nystatin. Specific conditions are indicated in figure legends. Plates were blinded by a lab member who would not be scoring worms for viability. Worms were transferred to new 3cm plates daily until egg-laying ceased. To determine viability, plates were inspected visibly to count the number of moving worms. As aging progressed, movement in worms was stimulated by plate tapping and gently touching the worm with a wire pick. Lack of response to

repeated prodding by a wire pick was used as the marker for death. Analysis of lifespan data was performed in R.

mtDNA stress survival assay

Day 1 adult worms were pulse-laid onto 3cm NGM+OP50 that contained either 0, 25, 50, or 100µg/mL ethidium bromide (final concentration; added to the agarose before plate pouring). The laid eggs were allowed to hatch and grow for three days before counting. When pulse-laying worms on plates containing ethidium bromide, fewer worms were laid on plates with higher concentrations of ethidium bromide because the adult worms tended to crawl up the walls of the plate.

QTL Analysis

QTL analysis was performed on the set of positions that vary between the MY1 and SEA51 genotypes. A genomic position was manually added to represent the rDNA genotype at position I15062083 and does not represent a specific SNV, rather is used to indicate whether the RIL has MY1 rDNA or SEA51 rDNA. rDNA genotypes were treated as binary: Either MY1 (high) or SEA51 (low). The cross object was generated with the read.cross function from the R/qtl package and converted to a recombinant inbred line type of cross with the convert2riself function. Allele designations were given as “S” or “M” to represent strain names “SEA51” or “MY1”. The genetic map was estimated using the Haldane map function with an error probability of 0.1 and a maximum number of iterations of 1000000. One-dimensional QTL scans were performed with the scanone function in R/QTL, and 1000 permutations were performed to determine the 99% confidence interval for the log odds scores. An interaction with rDNA copy number was tested using a nonparametric epistasis test with the R/qtl package, based on a method previously used

in our lab, using the prespecified I15062083 position to represent the rDNA [354,355]. An empirical null distribution for the difference in log odds scores between RILs was defined, using 1000 iterations. This null distribution was compared to the actual difference in log odds scores observed between RILs with 420 rDNA copies and RILs with 130 rDNA copies.

RNA-sequencing

Preparation of worms

Large pools of *C. elegans* were grown by bleaching an asynchronous worm population and allowing embryos to hatch and starve out for 20-24 hours statically on 10cm unseeded NGM plates overlaid with 10mL 1X M9. Starved L1 worms were transferred to 15mL conical tubes and washed once in 10mL 1X M9. 15,000 worms per strain were plated on individual 15cm high peptone NGM+NA22 plates and grown to the first day of adulthood (72 hours post-plating). Worms were harvested into 15mL conical tubes by washing the 15cms with 15mL 1X M9. Worms were allowed to settle for 3 minutes, supernatant was removed, and washed 4X in 12mL 1X M9. 500ul Trizol was added to the worm pellet, mixture was transferred to a 1.5mL Eppendorf tube, flash frozen in liquid nitrogen, and stored at -80°C.

RNA isolation

RNA was isolated with a Trizol-chloroform extraction protocol. The stored worm/Trizol mixture was freeze-thawed 3X between 37°C and a liquid nitrogen bath. 200µl more Trizol was added to the worm/Trizol mixture and mixed by pipetting up and down, then incubated for 5 minutes at room temperature. 140µl chloroform was added, and samples mixed by shaking vigorously for 20 seconds. Mixture was incubated for 2 minutes at room temperature, then spun 15 minutes at 12,000xg at 4°C. Supernatant was transferred to a clean Eppendorf tube containing

140µl chloroform, and tubes were mixed by shaking vigorously for 20 seconds and allowed to sit for 2 minutes at room temperature, then spun 15 minutes at 12,000xg at 4°C. Supernatant was transferred to a Lo-bind tube containing 300µl of isopropanol, and tubes were inverted gently to mix. Samples were incubated 10 minutes at room temperature, then spun 10 minutes at 12,000xg at 4°C. Supernatant was removed, and pellet washed with 600µl cold 75% ethanol. Pellets were allowed to dry in hood 5-10 minutes, then resuspended in 38µl Invitrogen RNase-free water and resuspended at room temperature.

The 38µl RNA was treated with TURBO DNase, adding 2µl TURBO DNase, 5µl 10X TURBO buffer, and 5µl 100mM $MnCl_2$. DNase treatment was performed for 1 hour at 37°C. RNA was then purified with an ethanol precipitation, and the final RNA pellet was resuspended in 100ul Invitrogen RNase-free water. RNA quality was assessed with TapeStation [Agilent], and RNA concentration was determined with nanodrop. All RNA samples were stored at -80°C.

Library preparation

The library preparation was derived from a previously-published bulk sci-RNA-seq protocol with an added RNase H-based rRNA depletion step [356–358]. For rRNA depletion, 1ug RNA was combined with 1ug rRNA-complementary oligoPool (IDT) and brought to a volume of 5µl with 1X hybridization buffer (200 mM NaCl, 100 mM Tris pH7.4). Mixture was incubated at 95°C for 2 minutes, then brought to 45°C with a ramp speed of -0.1°C/s. 5µl RNase H mixture (5U Hybridase Thermostable RNase H (Epicentre), 0.05µmole Tris HCl pH 7.5, 1µmole NaCl, and 0.2µmole $MgCl_2$) preheated to 45°C was added to the hybridized RNA/oligo mixture and mixed by pipetting. Digestion was performed for 45 minutes at 45°C.

rRNA-depleted RNA was purified with a 2.2X RNAClean SPRI bead treatment (Beckman Coulter). Purified, rRNA-depleted RNA was treated with Turbo DNase (Invitrogen) to remove the rRNA-hybridizing DNA oligos: 0.1 volumes of 10X TURBO DNase Buffer was added to the RNA, followed by 2µl TURBO DNase. Mixture was incubated at 37°C for 45 minutes. DNase-treated RNA was cleaned with 2.2X RNA SPRI beads. Final elution volume was 12µl, of which 9µl was recovered and used directly in the next step (oligo-dT incubation).

To the 9µl final rRNA-depleted RNA, 1µl 10mM dNTPs and 2µl 25µM oligo-dT(VN) were added and mixture was incubated 5 minutes at 65°C. Reverse transcription was performed by adding 4µl SuperScript IV Buffer, 2µl 0.1M DTT, 1µl SuperScript IV Reverse Transcriptase, and 1µl SUPERASEIN and incubating at 42°C for 50 minutes then 70°C for 15 minutes. Second strand synthesis was performed with the NEBNext® Ultra™ II Non-Directional RNA Second Strand Synthesis Kit. 2µl reverse transcription product was combined with 6.5µl Invitrogen RNase Free water, 1µl NEB Second Strand Synthesis Buffer, and 0.25µl NEB Second Strand Synthesis Enzyme Cocktail and incubated at 16°C for 150 minutes then 75°C for 20 minutes. The resulting cDNA was then either used immediately, or stored at 4°C overnight for use the next day.

Libraries were sequenced using a NextSeq 550 with a 75 Hi kit (Illumina). Read lengths used were: Index 1: 10bp, Index 2: 10bp, Read 1: 18bp, Read 2: 52bp.

Data analysis

Reads were demultiplexed with bcl2fastq (version 2.20). UMI information from Read 1 was merged into the Read 2 file for use in later steps. Poly-A tails were trimmed with Trim Galore (version 0.6.6), with accessory modules Perl (version 5.24.0), python (version 2.7.13) and cutadapt (version 1.18). Trimmed reads were aligned to the WS260 *C. elegans* reference genome

using STAR version 2.6.1c. Reads were filtered for ambiguously-mapping reads and sorted with samtools (version 1.9). PCR duplicate reads were identified with the UMI from the oligo-dT and removed with a custom script [356]. Each of the previous steps was performed independently for each sequencing run performed. To merge reads from multiple sequencing runs, Samtools was used to merge duplicate-removed bam files. Duplicate removal was then repeated. The number of reads aligning to the rRNA was counted with bedtools (version 2.29.2).

A file of counts per gene region of interest was produced with HTSeq (version 0.12.4) and accessory modules python (version 3.7.7), numpy (version 1.19.2), samtools (version 1.10), and pysam (version 0.16.0.1). The commands htseq-count -m union -i gene_id -r pos -a 10 --stranded=yes were used, with the WS260 canonical geneset gtf file used to provide gene coordinates, and final counts merged and output into a counts file. The counts file was imported into R version 4.0.4 for analysis of differential gene expression with DESeq2. Gene Set Enrichment Analysis of RNA-seq data was performed with the WormBase tool with a q value threshold of 0.1 [359].

Determination of mtDNA abundance

Stage-synchronized populations of worms were made by bleaching 10cm NGM+OP50 plates with gravid adult worms and allowing embryos to hatch and starve as L1s in 1XM9 overlayed on unseeded NGM plates. L1 worms were quantified and plated on 10cm NGM+OP50 plates. Worms were harvested for genomic DNA preparation at each of the following time points: 24 hours (for L2), 44 hours (for L4), and 67.5 hours (for gravid adults). DNA was prepared from each sample with the Qiagen genomic preparation kit following the previously described protocol (Appendix 1 and [5]).

Libraries were prepared with Nextera protocol as previously described, using 10ng DNA as input [5]. Target sequencing depth was 10 million reads per sample. Actual sequencing depth ranged from ~1.5 million to ~18.5 million reads; samples with less than 3 million reads were censored from analysis. The number of reads aligning to the mitochondrial genome and the nuclear genome were counted with a custom Perl script, as previously described [1].

rRNA quantification

TapeStation: The same sample of RNA that was extracted for the RNA-seq analysis described above was used for the 18S and 28S rRNA quantification by TapeStation. 100ng RNA was used for each sample and the standard Agilent RNA ScreenTape® protocol was used. Concentrations of rRNA were determined by integrating the intensity of the bands corresponding to the 18S or 28S rRNA and normalizing to the integrated intensity of the 25nt lower marker.

RNA gel: The same sample of RNA that was extracted for the RNA-seq analysis described above was used for the RNA gel. RNA was diluted to 500ng/µl and 5.5µl RNA was denatured at 65°C for 5 minutes with 10µl 96% Formamide in a 20µl reaction containing 3.5µl 6X DNA loading dye and 1µl 10mg/ml Ethidium Bromide. After denaturing, the sample was immediately put on ice for 5 minutes, then 18µl of the sample was loaded into a 1.2% agarose gel prepared in fresh 1X TAE buffer. Gel electrophoresis was performed for 7 hours at 80V, and the resulting gel was imaged on a ChemiDoc™ MP (Bio-rad). Band intensities were quantified in ImageJ by integrating the pixel intensity under the 18S and 28S bands. Normalization was performed by dividing the integrated intensity of the 18S or 28S band by the integrated intensity of the entire sample lane and multiplied by 100 to produce a total percent of RNA belonging to the 18S or 28S rRNA species.

RT-qPCR: 200ng RNA was used for reverse transcription in a 20µl reaction with RevertAid Reverse Transcriptase kit (Thermo Scientific™). Reverse transcription was performed at 25°C for 5 min, 42°C for 60 min, and 70°C for 5 min, and the samples were stored at -80°C until use in qPCR reactions. Because rRNA is highly abundant, the reverse transcription product was diluted for the qPCR reactions to an equivalent RNA amount of 0.2ng/µl, and 5µl of this 0.2ng/µl RNA equivalent was used in each qPCR reaction. For each set of primers used in each qPCR plate, a standard curve was set up, diluting the original RT product 1:5 a total of six times. For the standard curve, 5µl of each dilution was used in the qPCR reactions. The qPCR reactions were performed with Roche LightCycler® 480 SYBR Green I Master Mix in 20µl reactions, with 5µl 10µM of each primer per reaction. qPCR reactions were performed in a BioRad CFX Connect with the following conditions: 95°C 10 min, *95°C 10s, 60°C 15s, 72°C 30s, Image, Repeat from * a total of 40X. Primers used are in Table A4.12. qPCR analysis was performed in Microsoft Excel and plotted in R with ggplot2.

Statistics and data visualization

All statistical tests described were performed in R [360]. Data were visualized in R with the base plotting system or with ggplot2 [360,361]. The Color Universal Design palette [362] was used for some visualizations. R libraries used in this chapter include R/DESeq2, dplyr, extrafont, ggplot2, ggpubr, ggsignif, plotrix, qtl, rstatix, splines, survival, and tidyverse [361,363–369].

5.6 PROJECT ACKNOWLEDGEMENTS

We would like to thank the Moerman lab and Teotonio lab for kindly providing wild worm isolates. Some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). We would like to acknowledge members of the

Queitsch, Waterston, Brewer/Raghuraman, and Kaeberlein labs for helpful discussion. We would like to thank Jason Pitt for assistance in setup and troubleshooting the WormBot. This work was supported by funding from NIGMS (grant 5R01GM122088 to C.Q.), NHGRI (grant 1RM1HG010461 to C.Q) and NIA (F31 AG063450 to A.N.H.).

5.7 PROJECT CONTRIBUTIONS

Ashley Hall, Elizabeth Morton, and Christine Queitsch conceived of the study. Ashley Hall and Elizabeth Morton performed experiments. Data analysis and statistical analysis were performed by Ashley Hall. For the WormBot experiments, an undergraduate student, Young Woo Kim, performed the manual lifespan determination from the WormBot images.

CHAPTER 6: DISCUSSION AND FUTURE DIRECTIONS

6.1 FUTURE DIRECTIONS FOR rDNA COPY NUMBER VARIATION IN *C. ELEGANS*

In Chapters 4 and 5, I demonstrated that rDNA copy number variation in the naturally occurring range in *C. elegans* has no impact on any tested phenotypes. However, the phenotypes we tested were limited: It is possible that traits exist in *C. elegans* that are affected by rDNA copy number, but we did not measure them. Similarly, it is possible that differing rDNA copy numbers affect ribosome biogenesis or fitness, but only under certain conditions such as stress or nutrient limitation. Another angle that could be assessed is a full molecular profile of the rDNA arrays of different sizes. The number of rDNA copies transcribed and the chromosome contacts of the rDNA arrays have been studied in other organisms [17–19,172,173,370]. Fully characterizing these traits in *C. elegans* could provide a lens into how low, wild-type, and high rDNA levels affect rDNA chromatin state and nuclear architecture.

Why do worms with differing rDNA copy numbers have the same steady-state rRNA levels? Comparing worms with 420 rDNA copies to 100 rDNA copies, to have the same rRNA levels, I would expect that one of the following is happening: 1) In both worm strains, the same number of rDNA genes are active and transcribed to the same levels. 2) In both worm strains, the same proportion of rDNA genes are active (i.e., 50% of the genes are transcribed), but the total number of transcription initiation events is the same between the strains. 3) The strains differ in basal transcription level of the rRNA genes, but post-transcriptional regulation modulates the final rRNA levels. In *C. elegans*, rRNAs are regulated by antisense ribosomal siRNAs (risiRNAs) through degradation of erroneous rRNAs, reduction of pre-rRNA levels, and inhibition of RNA Pol

I transcription [338–341]. Therefore, it is possible that a combination of gene activation, gene silencing, and post-transcriptional regulation involving risiRNAs maintain equivalent steady-state rRNA levels between *C. elegans* strains of differing rDNA copy numbers. Whether risiRNAs accumulate more in strains with higher rDNA copy numbers is testable, using either small RNA sequencing or a risiRNA biosensor [338,342].

In line with exploring how the same steady-state levels of rRNA are maintained by worms of differing rDNA copy numbers, a complementary question can be asked: What conditions induce differential levels of rRNA expression in strains with differing rDNA copy numbers? In humans, differences in ribosome biogenesis in people with higher rDNA copy number are evident under conditions of resistance exercise but not under rest conditions [371,372]. Stresses including heat shock and nutritional stress change rRNA expression levels in a range of organisms, but how these conditions differentially affect individuals with differing rDNA copy numbers has not been tested [44–48]. Applying these considerations to *C. elegans*, we could in combination test whether differing rDNA copy number affects stress tolerance, as well as how each stress affects rRNA transcription.

Moving beyond a strict focus on ribosome biogenesis, our *C. elegans* NILs could provide a platform to study how rDNA copy number affects cancer risk. Changes in rDNA copy number are commonly observed in cancer, with one study demonstrating an increased lung cancer risk in smokers with high rDNA copy number [211]. Currently, no model organism studies have been performed to determine how rDNA copy number contributes to cancer risk. *C. elegans* forms proximal germline tumors when proliferating germ cells fail to transition from mitotic to meiotic division [373–375]. Mutants in *pro-1* have an increased incidence of proximal germline tumors

and are defective in pre-rRNA processing [65]. Mutation in *ncl-1*, which increases 45S pre-rRNA expression, suppresses the *pro-1* phenotype [65]. Similarly, a weak gain-of function in *glp-1* (*glp-1*(*ar202*)) increases penetrance of proximal germline tumors, a phenotype which can be enhanced by knocking down genes involved in translation [376]. These genes include a subunit of RNA Pol I (*rpoa-2*) and multiple genes involved in rRNA modification, implicating a failure of ribosome biogenesis in tumor formation.

This proximal germline tumor phenotype could provide a model tumor system in which to study whether rDNA copy number differences affect cancer risk. Both baseline incidence of proximal germline tumor formation and the incidence in a sensitized background could be assessed. In the context of the proximal germline tumor, having more rDNA copies could confer the potential to produce more pre-rRNA and could suppress the phenotype in a manner similar to *ncl-1* mutation. This would be the opposite of the proposed effect of rDNA copy number on lung cancer risk [211]. However, cancer is a highly diverse class of diseases, so it is possible that increased rDNA copy number could increase risk for some cancer types while decreasing it for others. Further, proximal germline tumor formation is exacerbated by slow or insufficient germline proliferation in larval stages, which could be affected by reduced ribosome biogenesis, [376]. If increased rDNA copy number reduced the risk of proximal germline tumor formation, the mechanism could be through increased ribosome biogenesis in precursor cells.

Beyond our *C. elegans* NILs, studies of rDNA copy number variation in an isogenic background have been largely limited to reductions of rDNA copy number. Does rDNA copy number variation in the naturally occurring range confer a phenotype in organisms such as yeast, flies, and humans? In humans, there are publications that claim that rDNA copy number

associates with global changes in gene expression [2]. However, in Chapter 3, I demonstrate that the copy number estimates of the individuals used in this study are likely inaccurate. The associations with global transcriptome differences may not be maintained if the updated rDNA copy number estimates were used to re-analyze the association. In *S. cerevisiae*, while most studies have focused on drastic reductions in rDNA copy number, chromatin silencing differs in yeast with 250 rDNA copies versus 100 rDNA copies. With this modest reduction in rDNA copy number, silencing at the telomeres and mating type loci increases, due to increased activity of the histone deacetylase Sir2 [167]. In *D. melanogaster*, studies that found differences in position effect variegation and expression of mitochondrial genes used flies of modestly reduced rDNA copy number [6,168].

With expanded assessment of the naturally occurring range of rDNA copy number variation, insight into human health and disease could be gained. It is possible, however, that there are few effects of rDNA copy number in the naturally occurring range. This would make sense: if there is a strong fitness consequence to having a particularly high or low rDNA copy number, that rDNA copy number would likely not be maintained. In this case, it could be beneficial to explore the consequences of rDNA copy number variation outside of the naturally occurring range.

6.2 EXPANDING THE RANGE OF rDNA COPY NUMBERS ASSESSED IN *C. ELEGANS*

The rDNA copy number variation assessed in this thesis was restricted to that of the naturally occurring range among *C. elegans* wild isolates. In the section above, I propose other conditions and phenotypes that could be affected by these rDNA copy number differences.

However, it is entirely possible that rDNA copy number must be pushed outside of the natural range to induce strong phenotypic differences. *C. elegans* with zero rDNA copies can complete embryogenesis with their supply of maternal ribosomes, but developmentally arrest in the first larval stage [162]. Recently, our lab showed that reducing rDNA copy number to approximately 5% of wild type levels permits development to adulthood, with variable penetrance (Morton *et al* 2021; unpublished). All worms with this severely reduced rDNA copy number, however, are sterile, and display severe and variable morphological defects. Reducing rDNA copy number to approximately 33% of wild type enables superficially normal development to fertile adulthood but confers a subtle delay in developmental timing (Morton *et al* 2021; unpublished).

The worms with rDNA copy number reductions that our lab used to study copy numbers below the naturally occurring range, however, are heterozygotes. The strain with ~5% of wild type levels has the genotype *eDf24*/Δ*rDNA*, which is 11/0 rDNA copies. The strain with ~33% of wild type levels has the genotype *catIR17*/Δ*rDNA*, which is 73/0 rDNA copies - or half of the lowest number of rDNA copies found among wild isolates. Producing enough worms with the heterozygous genotype for assays that require many worms, or the propagation of progeny, is nontrivial. For example, *eDf24*/Δ*rDNA* heterozygous worms must be produced by crossing the two parental genotypes of interest for each assay because the *eDf24*/Δ*rDNA* worms are sterile. This poses some throughput limitations: For example, assays such as RNA-seq that require a large, synchronized population of heterozygous worms would be difficult to accomplish. Further, assays that require multigenerational propagation, such as the competition experiments, are not feasible with these genotypes. This is not to say that continued study of these heterozygous strains is not possible: Plenty of traits - such as lifespan - are readily assayed on a small population

of heterozygotes. Further, for the *catIR17*/Δ*rDNA* genotype, the rDNA genotypes could be paired with genetic balancers such that heterozygosity is obligated [377]. However, it would be ideal to have a strain resource that is homozygous for two short rDNA arrays for further studies.

One way that we could generate a series of strains with reduced rDNA copy number is through CRISPR. Prior to my joining the lab, our lab unsuccessfully tried to reduce rDNA copy number with this method. However, substantial advances have been made in gene editing in *C. elegans* with CRISPR systems. These advances include co-CRISPR strategies and specific preparation conditions for the repair template [378–380]. These improvements could make engineering rDNA deletions in *C. elegans* more feasible. The goal would be to generate a series of *C. elegans* strains with a range of low rDNA copy numbers, for example, 10, 20, 30, 40, and 50 rDNA copies per haploid genome. Any lines with copy number losses obtained by CRISPR will be reductions that are still sufficient for development and reproduction. Care would need to be taken to screen both worms that grow well and those with apparent defects to identify a variety of rDNA reductions.

Mechanisms to increase rDNA copy number are less obvious. Methods of cutting and repairing the rDNA are more likely to cause copy loss than gain [165,312]. There are examples in plants, however, of small rDNA copy gains with a CRISPR approach, but these gains were small (less than 2-fold increase in copy number), rare, and not able to be maintained over multiple generations [165]. In *C. elegans*, it may be more feasible to increase rDNA content of a worm by adding more rDNA arrays, rather than by increasing the size of the array on chromosome I. There are some translocations and duplications of the end of chromosome I, including the rDNA, that are viable (see **Appendix 3** for details). These chromosomes could be put into the NIL

backgrounds with 420 rDNA copies, increasing rDNA copy number to ~520-680 per haploid genome [342,381]. However, these chromosomal variants also duplicate megabases of chromosome I sequence, and the duplicated region may not be benign [381]. Appropriate control strains of lower rDNA copy number in conjunction with the chromosome I region duplications would need to be used to differentiate rDNA-specific and non-rDNA effects. Another option to increase rDNA copy number could be to inject worms with plasmids containing an rDNA repeat unit, to produce an extrachromosomal array that contains extra rDNA copies. Many limitations apply to the array approach, including that there would be non-rDNA plasmid sequence included, the orientations of plasmids incorporated is random, and extrachromosomal arrays are often silenced, so the rDNA copies may not be expressed [382,383]. While neither of these latter two options is ideal, they could provide a starting point for expanding the rDNA content of the worm genome.

There are many phenotypes that could be assessed if strains with rDNA copy numbers above or below the naturally occurring range were made. These include both the phenotypes that I measured in Chapter 5, and the possible phenotypes of interest for future study described above. The three phenotypes that I would prioritize are rDNA copy number stability, competitive fitness, and developmental timing. For rDNA stability, I would propose propagating strains with increased or decreased rDNA copy number for 100 generations in bulk, taking samples every 10 generations for CHEF gel analysis. While in the RILs we did not observe rDNA copy number changes, the starting rDNA copy numbers for all strains were still within the naturally occurring range, and we only propagated the RILs for 20 generations. Strains with extreme rDNA copy number reductions would be more likely to undergo copy number changes, as seen in *D.*

melanogaster and *S. cerevisiae* [143,306]. For competitive fitness, I would improve our assay by using a GFP-expressing transgene that is not linked to the rDNA, such that reciprocal strains can be made and transgene-dependent effects accounted for [384,385]. This assay would be prioritized for the same reasons as in Chapter 5: It can detect the cumulative effects of small fitness defects. Finally, I would select developmental timing because we have evidence that reducing rDNA copy number to ~35 copies per haploid genome slows development, from the heterozygous 73/0 strain. Altogether, expanding our repertoire of copy number variants to those markedly shorter and longer than in *C. elegans* wild isolates could identify phenotypic consequences of rDNA copy number variation. These phenotypes could then be investigated in strains with copy numbers in the naturally occurring range, with the expectation that the effect sizes would be smaller.

6.3 Short read sequencing estimates of rDNA copy number: Where do we go from here?

Through the introduction chapter of this thesis and in the data chapters, I have largely presented results dependent on rDNA copy number estimates as they are published. However, in Chapters 2 and 3 I demonstrate that rDNA copy number estimates based on short read sequencing are often unreliable. Studies that have described changes in rDNA copy number in cancer have relied on short read sequencing [2,94,97,216,258]. Others still have developed their own technologies for rDNA copy number estimation, but still compare them to short read sequencing as if it is the gold standard [94,205]. Here I will discuss some ideas and best practices for rDNA copy number estimation, if sequencing must be used.

Experimentally, the best practice for measuring rDNA copy number with WGS should use a library preparation that minimizes the introduction of biases. Much of sequencing bias comes from uneven coverage of genetic regions with extremely high or low GC content [386–388]. In most genomes, the rDNA has a higher GC content than the rest of the genome - a consideration that could be important, given the GC biases in sequencing. While all steps of the sequencing process can introduce biases, including tagmentation and PCR amplification in library preparation, cluster generation, and the sequencing itself, library preparation is considered the largest bias-introducing step [389–391]. In fragmenting DNA, bias could be reduced by choosing sonication over enzymatic tagmentation, which has known sequence preferences [391,392]. PCR amplification is another step that exacerbates GC biases, so a PCR-free approach would be preferred [389,393]. Ideally, all libraries should be prepared at the same time, with the same batch of reagents, and by the same person. Furthermore, it would be ideal to prepare replicate libraries for each sample, to estimate the technical error, and to include control samples of established rDNA copy number.

Each step to reduce bias and improve rDNA copy number estimates, however, imposes limitations on what can be studied. For example, a recent study in *C. elegans* estimated rDNA copy number from DNA extracted from single worms [342]. When sequencing libraries for *C. elegans* have been prepared with sonication to shear the DNA, micrograms of DNA were used as input [1]. A single worm only contains approximately 1.3 picograms of DNA, which drastically limits the options for library preparation [394]. The study therefore prepared libraries on multiple individual worms per condition, to determine the measurement error [342]. In the case of a small number of *C. elegans* strains, it is feasible to prepare and sequence multiple libraries per

condition. However, in other cases, such as human genome sequencing, preparing and sequencing multiple libraries per sample could be cost prohibitive. Even if ideal sample preparation methods could be achieved, one final limitation to fixing rDNA copy number estimation at the sample preparation stage is that it cannot improve the copy number estimates for already existing data.

The ideal path to improving rDNA copy number estimates from short-read sequencing would be to develop new post-sequencing correction factors or to apply different copy number estimation methods. A computational approach is ideal because it can be applied to the thousands of sequencing datasets that have already been produced. In Chapter 2, we applied two methods for rDNA copy number estimation for the same samples: A method based on read depth, and a method based on a GC fragmentation bias correction algorithm [1,97,261]. We found that the GC correction was insufficient to fix our rDNA copy number estimates, although it was better than the read depth method. One method that we did not test is a k-mer counting approach, which has been applied to rDNA copy number estimation in maize [395,396]. Many computational tools exist for copy number variant detection, and often apply methods for resolving GC-content-based biases, which could be tested for their ability to quantify rDNA copy number [397–399]. However, these detection methods largely focus on the identification of single deletions, duplications, or polyploidies, which is arguably a different problem than accurately quantifying the tens to hundreds of rDNA copies a eukaryotic genome has. While we did not come up with a computational model that could correct our rDNA copy number estimates, we did find that copy number estimates of single-copy regions also vary between

library preparation methods. It is possible that the copy number variation of these single-copy regions could inform a metric for correcting rDNA copy number estimates.

6.4 LONG-READ SEQUENCING AND THE PROMISE OF INCORPORATING rDNA INTO REFERENCE GENOMES

Since I started studying rDNA for my dissertation, long-read sequencing technologies have substantially improved. High quality, and more complete, reference genomes have been published for many organisms [233,400,401]. In most genomes, rDNA arrays are typically represented in reference genomes either with a single rDNA reference copy or on an unassembled scaffold. Long-read sequencing holds unique promise for understanding location, sequence variation, structural variation, and copy number of the rDNA.

Currently, there are many challenges to resolving the rDNA with long reads. Its highly repetitive nature makes tiling across the array with multiple reads difficult unless there are enough unique variants present to anchor specific reads. Reads in the rDNA are often shorter than those of the rest of the genome, potentially due to secondary structures that could form in the genic regions and disrupt the sequencing process [125]. The sheer length of the rDNA also makes finding reads long enough to span an entire array rare, if not impossible [85,125,402]. Nevertheless, some success has been achieved in a few model organisms. In *C. elegans*, the 5S rDNA array was recently assembled with long-read sequencing, allowing the discovery of variation in 5S composition across isolates of the reference strain maintained in different labs [125]. The same study, however, failed to assemble across the 45S due to insufficiently long reads and insufficient sequence variation between 45S rDNA copies. In *A. thaliana*, a combined

approach of BAC cloning, short read, and long read sequencing was successful in assembling one of the two rDNA arrays in the genome [99].

Human rDNA sequencing poses even more of a problem than *C. elegans*, having multiple rDNA loci and larger repeat sizes. Nevertheless, if each rDNA in the human genome could be assembled with long reads, the dual tasks of determining both the total number of rDNA repeat units per genome, as well as the number of rDNA repeat units in each individual array could be accomplished. Recent efforts to do this for a functionally haploid human cell line have had commendable success, with two of the five arrays definitively assembled [85]. Reads that span multiple rDNA repeat units have begun to be used to overturn the theory that approximately a third of rDNA units in an array are in an inverted or palindromic orientation [83–86]. Finally, chromosome-specific rDNA sequence variants were identified in the functionally haploid human cell line [85]. Combining the distribution of rDNA sequence variants by chromosome with analyses of which NORs are active could determine whether certain rDNA sequence variants are expressed or not.

Complete characterization of rDNA arrays has many fascinating applications for rDNA biology, including distinguishing between rDNA copy number variation caused by specific arrays or extrachromosomal variation, or cataloguing R1 and R2 retrotransposon distribution in *D. melanogaster* rDNA. Some rDNA arrays are completely silenced, and activity of rDNA arrays is heritable, which could be important in predicting the number of active rDNA genes and the number of them with coding variants. Even the process of assembling the rDNA from long reads is producing novel computational methods for aligning these highly homogenous, repetitive sequences [85]. Altogether, the fewer genomic unknowns that there are, the better we will be

able to understand human health and disease. Complete resolution of the rDNA arrays will help to provide more definitive answers to whether rDNA sequence or copy number variation contributes to human diseases or aging.

APPENDIX 1: SUPPLEMENTAL DATA, ANALYSIS, AND INFORMATION FROM CHAPTER 2

Appendix 1 contains supplemental data, analysis, methods, and other information from Chapter 2. These are also published as a part of the supplement for Morton 2019 [5].

A1.1 SUPPLEMENTAL FIGURES

All supplemental figures and tables for the published manuscript are available on figshare: https://doi.org/10.25387/g3.10406519. One supplemental figure, relevant for the included material in Chapter 2, is included here:

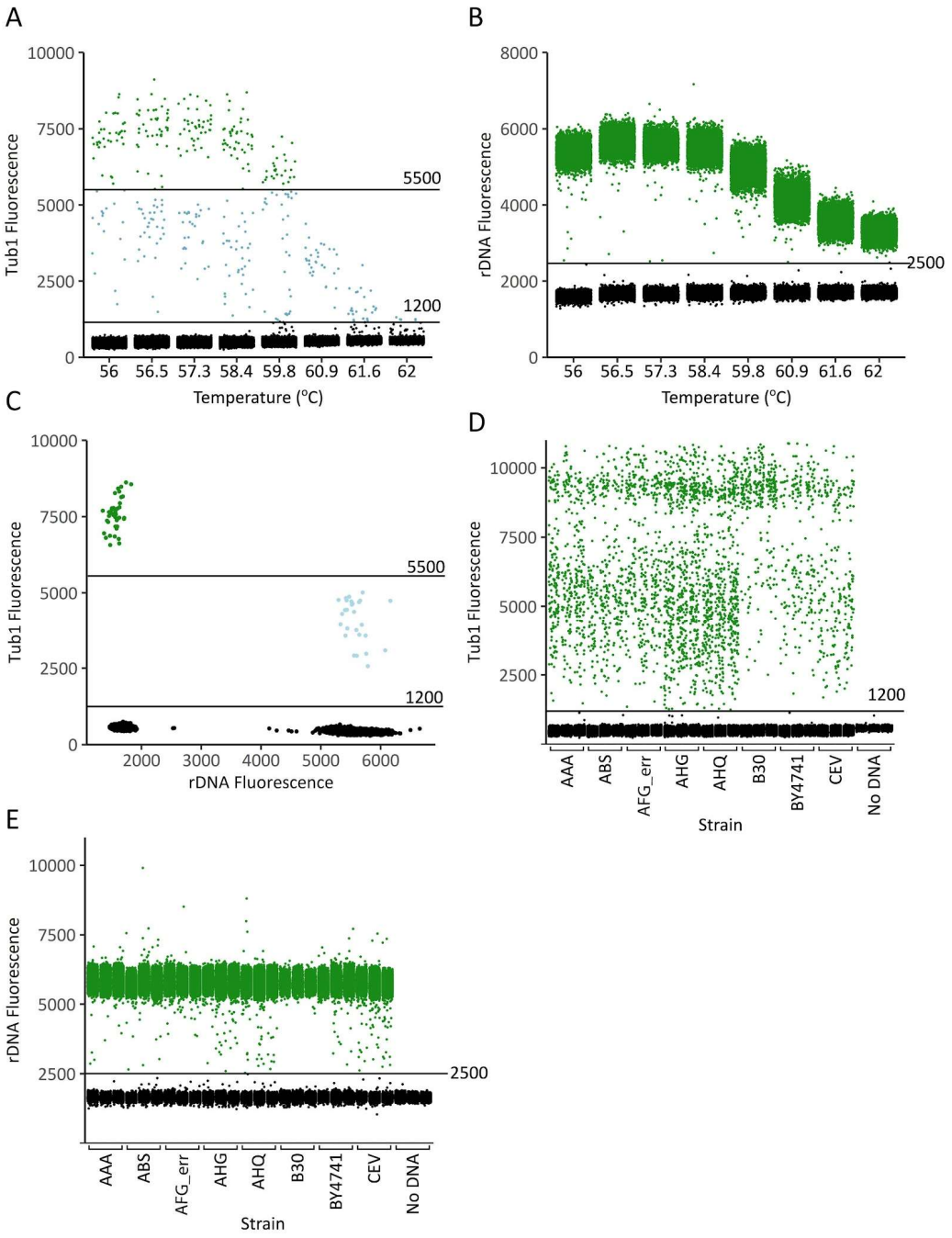

Figure A1.1: Optimization of conditions for ddPCR quantification of rDNA copy number.

A-B: An annealing temperature gradient was performed with BY4741 gDNA to identify a temperature at which both the single-copy gene Tub1 (**A**) and the rDNA (**B**) are efficiently amplified and detected. The final selected annealing temperature for the assay was 57°C. **C:** 2-D plot of rDNA vs Tb1 fluorescence annealed at 57.3°C. Tub1 fluorescence is lower in droplets positive for both Tub1 and rDNA, but clearly segregates from the Tub1 negative droplets at temperatures of 58.4°C and below. **D-E**: 1-D data for Tub1 (**D**) and rDNA (**E**) fluorescence from the first ddPCR replicate with strains BY4741, B30, AAA, ABS, AFG_err, AHG, AHQ, CEV, and a no

DNA negative control. 'Rain droplets' (droplets of intermediate fluorescence) are generally considered positive, because no rain droplets are present in the no DNA negative controls. Thresholds used in our analysis are indicated; 1200 units for Tub1 fluorescence, and 2500 units for rDNA fluorescence.

Figure A1.2: Southern blots and ethidium bromide staining of CHEF gels.
rDNA copy numbers in wild *C. elegans* and *S. cerevisiae* strains were assessed using multiple replicate CHEF gel runs. Ethidium bromide stains of CHEF gels are shown next to the rDNA-probed Southern blots of each gel. The rDNA copy number estimated from the band size is indicated (omitted if accurate assessment was not possible in a given CHEF gel replicate). **A:** Three replicates of *C. elegans* rDNA CHEF gels and Southern blots are shown. The Southern blot of replicate 2 is presented in Figure 1B. Replicate 1 used MfeI restriction digest to excise the rDNA array; replicates 2 and 3 used SwaI restriction digests. An 850bp sequence within the *C. elegans* rDNA was used as a probe to visualize location of the rDNA array on Southern blots (See Methods). Ladders are chromosomes from the indicated yeast species: *S. cerevisiae* or *H. wingei*. **B:** Shown is the *S. cerevisiae* ethidium bromide-stained CHEF gel and corresponding Southern blot presented in Figure 2B. This gel represents undigested, intact *S. cerevisiae* chromosomes. A probe to CDC45 (single copy gene on Chr. XII) was used to visualize the location of chromosome XII, which contains the rDNA. **C:** Three CHEF gel replicates of BamHI-excised *S. cerevisiae* rDNA arrays are shown using conditions such that sizes between 1Mb-3.13Mb are resolved (See Methods, "High Molecular Weight" conditions). rDNA arrays were visualized on Southern blots using an rDNA-specific probe sequence. 16 haploid and 12 diploid "1,002 genomes" strains were examined for replicate 1. Replicates 2 and 3 include 7 haploid "1,002 genomes" and the strain BY4741. **D:** A subset of strains from C were run again under CHEF gel conditions to resolve sizes between 225kb-1.1Mb ("Low Molecular Weight" conditions). The set of replicate 1 strains were split between two gels (part 1 and 2). Replicates 2 and 3 include three haploid "1,002 genomes" strains (AHG, ACV, ACQ) and the B30 strain (35 rDNA copies). **E:** Individual *S. cerevisiae* colonies were examined for rDNA copy number heterogeneity. CHEF gel plugs were generated from a mixed population sample ("P") and 3 single colonies ("1-3") from 10 haploid strains and 8 diploid strains. Undigested samples were resolved with CHEF gels ("High Molecular Weight" protocol) and visualized as in **B**.

This figure is available at Figshare (https://doi.org/10.25387/g3.10406519) as supplemental figure 1.

A1.2 Supplemental Tables

The following supplemental tables are associated with this appendix. All tables are uploaded to Figshare as a part of the published manuscript [5]; those that are small enough are also reprinted here.

Table	**Description**	**Location**
Table A1.1	Worm WGS Metrics	Figshare: TableS1
Table A1.2	Worm Percent Error	Included in Appendix 1
Table A1.3	Worm Downsampling WGS	Figshare: TableS3
Table A1.4	Aligner comparison	Included in Appendix 1
Table A1.5	Worm Single Copy Region Coordinates	Included in Appendix 1
Table A1.6	Worm Single Copy Region Counts	Figshare: TableS6
Table A1.7	Yeast Strain SRA Numbers	Included in Appendix 1
Table A1.8	Yeast Diploid Strain rDNA Estimates	Figshare: TableS8
Table A1.9	Yeast Strain SNV Genotyping	Included in Appendix 1
Table A1.10	ddPCR Raw Data	Figshare: TableS10
Table A1.11	Oligo Sequences	Included in Appendix 1
Table A1.12	CHEF Measurements	Figshare: TableS15

Table A1.2. *C. elegans* percent error of rDNA estimates by different methods compared to published or CHEF averages.

	Compared to Thompson 2013					Compared to CHEF gel average				
Strain	CHEF gel	10ng Read Count	10ng GCC	50ng Read Count	50ng GCC	CHEF gel	10ng Read Count	10ng GCC	50ng Read Count	50ng GCC
MY16	10.9% ± 6.9%	10.4% ± 10.2%	9.4% ± 5.2%	51.4% ± 15.7%	25.4% ± 6.0%	11.3% ± 6.1%	10.2% ± 5.7%	9.2% ± 6.3%	52.7% ± 15.9%	26.5% ± 6.0%
JU775	10.4% ± 5.1%	29.3% ± 16.3%	28.3% ± 13.3%	85.5% ± 23.9%	55.0% ± 8.0%	10.4% ± 5.2%	29.3% ± 16.3%	28.3% ± 13.3%	85.5% ± 12.9%	55.0% ± 8.0%
ED3042	2.2% ± 2.8%	19.0% ± 7.5%	12.4% ± 11.0%	81.7% ± 20.5%	41.6% ± 7.0%	2.6% ± 1.3%	19.2% ± 8.6%	12.3% ± 10.3%	79.1% ± 20.2%	39.5% ± 6.9%
ED3040	4.7% ± 2.7%	18.8% ± 3.6%	12.2% ± 11.9%	77.8% ± 17.7%	44.0% ± 6.7%	3.5% ± 2.1%	19.2% ± 2.1%	10.7% ± 10.8%	72.2% ± 17.1%	39.5% ± 6.5%
MY6	12.0% ± 7.2%	14.5% ± 12.6%	13.3% ± 14.1%	74.7% ± 15.6%	44.3% ± 5.8%	4.4% ± 3.5%	16.5% ± 7.6%	14.9% ± 5.0%	55.9% ± 13.9%	28.8% ± 5.1%
MY14	6.7% ± 6.4%	12.7% ± 1.8%	13.1% ± 3.2%	65.5% ± 18.5%	30.9% ± 7.0%	4.2% ± 3.4%	9.8% ± 8.7%	10.2% ± 8.8%	55.4% ± 17.4%	22.8% ± 6.5%
PX174	5.5% ± 7.5%	21.6% ± 23.7%	18.4% ± 18.0%	35.1% ± 14.0%	11.2% ± 4.7%	5.4% ± 0.0%	25.4% ± 24.5%	23.4% ± 18.4%	24.7% ± 12.9%	4.1% ± 2.1%
MY1	3.4% ± 1.7%	13.2% ± 13.5%	13.0% ± 10.3%	34.2% ± 14.2%	12.9% ± 4.8%	3.0% ± 2.1%	12.4% ± 13.1%	12.7% ± 9.0%	36.0% ± 14.4%	14.4% ± 4.9%

Table A1.5. Worm single copy region coordinates.

Region ID	Chromosome	Coordinates
seq21023	IV	8260386-8267581
seq23058	IV	9548643-9555838
seq27031	IV	11973902-11981097
seq7706	III	2667496-2674691
seq13292	III	4525240-4532435
seq17053	III	6746305-6753500
seq17147	III	6838722-6845917
seq8238	X	3228527-3235722
seq9303	X	3747072-3754267
seq12013	X	4984354-4991549
seq12874	X	5425754-5432949
seq16636	X	7493006-7500201
seq20808	X	9883763-9890958
seq23256	X	11491102-11498297
seq12962	I	4058337-4065532
seq21553	I	8485243-8492438
seq22300	I	8904740-8911935
seq37612	I	14881976-14889171
seq7336	V	2784811-2792006
seq17239	V	7574063-7581225
seq18187	V	8227435-8234630
seq19091	V	8838286-8845481
seq19134	V	8880380-8887575
seq25090	V	13038659-13045854
seq28101	V	14838863-14846058
seq29213	V	15498607-15505802
seq32276	V	17133464-17140659
seq15762	II	7731311-7738506
seq17297	II	8881839-8889034

Table A1.7. SRA numbers of yeast strain data used for re-analysis.

STD_name	Isolate Name	SRA number	Aneuploidies (1002 Genomes)
AAA	RM11-1a	ERR1309487	euploid
AAC	CBS2165a	ERR1309038	euploid
ABS	DBVPG3591_1b	ERR1309033	euploid
ACK	Y12_1b	ERR1308893	euploid
ACQ	CLIB219_2b	ERR1308618	euploid
ACV	WE372_1b	ERR1308596	aneu;+1*1;
ADA	Y55_1b	ERR1309427	euploid
AEF	CBS6414	ERR1308873	euploid
AEQ	CBS1782	ERR1309512	euploid
AES	CBS4054	ERR1309368	aneu;+1*2;
AFC	CBS7764	ERR1309327	aneu;+1*7;-1*14;
AFG	CBS1387	ERR1309145	euploid
AGC_1	CBS2444	ERR1309000	aneu;+1*1;+1*5;+1*10;+1*11;
AHG	CBS1586	ERR1308781	euploid
AHQ	CBS440	ERR1309019	euploid
AHR	CBS1227	ERR1308770	euploid
ANV	RIB0004	ERR1308740	aneu;+1*1;+1*8;
ARL	CBS2087	ERR1309170	aneu;+1*1;
BGD	CLIB552	ERR1309432	euploid
BII	DBVPG1106	ERR1308846	euploid
CEB	DBVPG1619-4B	ERR1309017	euploid
CEK	JCM_2985-4B	ERR1309167	euploid
CER	N34:2-4(a)	ERR1308780	euploid
CEV	IFO_0289:4-3(b)	ERR1308745	euploid
CMF	RIB6001	ERR1308700	aneu;+1*3;
SACE_GAL	SK1	ERR023701	euploid
SACE_GAS	UWOPS83-787_3	ERR049929	euploid
SACE_GAV	W303	ERR023702	euploid

Table A1.8. Yeast diploid strain rDNA estimates.

Strain	SRA Number	WGS*	CHEF gel
AFC	ERR1309327	20	58
AAC	ERR1309038	30	61
BGD	ERR1309432	40	45
ARL	ERR1309170	47	71
CER	ERR1308780	71	91
ANV	ERR1308740	77	80
SACE_GAS	ERR049929	95	95
BII	ERR1308846	119	123
SACE_GAL	ERR023701	118	111
CMF	ERR1308700	231	111
CEK	ERR1309167	243	127
SACE_GAV	ERR023702	254	292

*GCC-analyzed from WGS data [3]

Table A1.9: Yeast strain identity in re-sequencing was confirmed by percent shared SNVs with reported genotypes.

Strain	SRA Only SNVs	Re-sequencing Only SNVs	Shared SNVs	%Reseq SNVs In SRA	%SRA SNVs in Reseq
AAA	3264	1033	43165	97.7%	93.0%
ABS	3511	851	42522	98.0%	92.4%
ACK	3841	916	61537	98.5%	94.1%
ACQ	13029	17471	69387	79.9%	84.2%
ADA	4952	1140	62215	98.2%	92.6%
AEF	4367	1225	62171	98.1%	93.4%
AEQ	4309	1012	48811	98.0%	91.9%
AFG	32627	29086	28861	49.8%	46.9%
AGC	57015	196	64	24.6%	0.1%
AHG	3214	1004	42752	97.7%	93.0%
AHQ	3968	1189	61332	98.1%	93.9%
AHR	3605	833	42647	98.1%	92.2%
CEB	3009	585	43506	98.7%	93.5%
CEV	5625	1085	60567	98.2%	91.5%

Table A1.11: Oligo Sequences.

Primer name	Purpose	Sequence
EM50	*C. elegans* rDNA Southern probe PCR (forward, 550bp upstream of start of rrn-1.1)	cgaggtctccagagagacg
EM51	*C. elegans* rDNA Southern probe PCR (reverse, 300bp down from start of rrn-1.1)	agttgaaagggcagacaccc
5NTS2	*S. cerevisiae* rDNA Southern probe	CTGGTAGATATGGCCGCAACC
3NTS2	*S. cerevisiae* rDNA Southern probe	GTCTTCAACTGCTTTCGCAT
CDC45 probe F	*S. cerevisiae* single copy chr. XII gene Southern probe (CDC45)	ATCTATGCTGGCAAGCACCA
CDC45 probe R	*S. cerevisiae* single copy chr. XII gene Southern probe (CDC45)	TTTTGGGTTAAGTGGCCGT
TUB1_Probe_ddPCR	TUB-1 (single-copy) ddPCR probe	/56-FAM/TCCATGAGT/ZEN/CCAACTCTGTGTCA/3IABkFQ/
25S_rDNA_Probe_ddPCR	rDNA ddPCR probe	/5HEX/AACATAGACAAGGAACGGCCC/3BHQ_1/
TUB1_F_ddPCR	TUB-1 ddPCR primer	CCAGTCTTATCCAAATCAAAGG
TUB1_R_ddPCR	TUB-1 ddPCR primer	GGATCACACTTGACCATCT
rDNA_25S_F_ddPCR	rDNA ddPCR primer	TACCTTCGGTGCCCGAGTTGTAAT
rDNA_25S_R_ddPCR	rDNA ddPCR primer	ACCCTCTATGACGTCCTGTTCCAA

A1.3: SUPPLEMENTAL METHODS

rDNA Southern blotting

C. elegans

For Southern blotting, CHEF gels were prepared and probed using protocols outlined in Tsuchiyama et al. 2013 [278]. In short, each gel was washed twice for 10min each in 0.25N HCl to nick and depurinate DNA, followed by two washes for 15min each in 0.5N NaOH, 1M NaCl. These incubations were followed by two washes for 15min each in 0.5M Tris, 3M NaCl. DNA was then transferred from the gel to a nylon membrane (Perkin Elmer GeneScreen Hybridization Transfer Membrane) and crosslinked using a Stratagene Stratalinker UV Crosslinker in preparation for radioactive probe hybridization.

The probe for Southern blotting *C. elegans* rDNA was created from an 850bp PCR product that overlaps the first 300bp of *rrn-1* (18S) amplified with primer pair EM50+EM51, purified with the Zymo Clean & Concentrator kit (D4013) before radioactive labeling. Please note that we observed another rDNA probe to behave anomalously during standard gel electrophoresis. Our initial PCR product intended for Southern blotting was designed for an 806bp region within the 26S gene and ran at the correct size following PCR amplification, but ran incorrectly (at ~500bp) following purification and elution in water. Addition of buffer (Invitrogen Y02028) returned the product to an 800bp run size. The 850bp PCR product behaved more consistently and thus was the one used for probing in this study.

Band measurement

To determine rDNA copy number from CHEF gel followed by Southern blot, the distance between the bottom of the well and the middle of a sample band was measured manually for

the presented Southern blots. The distance between the bottom of the well and each of the ladder bands was measured from the ethidium bromide stain of the gel before it was transferred to the blot. The relationship between band size and gel distance was plotted and used to determine band size for each rDNA band (**Table A1.12**). There were two distinct, approximately equal-intensity bands for MY16 in two of the CHEF gels (**Table A1.12**). For CHEF average calculation, the two bands were averaged first to produce that CHEF replicate's rDNA estimation for MY16. ED3040 and MY6 displayed minor bands; for these strains, only the major band was considered.

Genomic DNA extraction and whole genome sequencing sample preparation

C. elegans

The Qiagen DNeasy kit (69504) was used for worm genomic DNA extraction. Wild isolate worm pellets frozen in ATL were freeze-thawed three times between -20°C and 37°C. 20µL proteinase K was added and the samples were incubated at 56°C for 3hr with occasional vortexing. 4µL 100mg/mL RNase A (Qiagen 19101) was added to each sample and incubated at room temperature 5min. 200µL AL buffer was added, and the DNA extraction continued as described in the kit protocol. Final DNA was eluted in a total volume of 150µL. For each worm wild isolate, a single genomic DNA sample was stored at 4°C and used for all of the described library preparations for that strain. The dates of each library preparation are provided in Table S1.

Sequencing of libraries was performed using an Illumina Nextera DNA Sample Preparation kit (FC-121-1030). All worm library preparations were performed by the same person. For 10ng input: 10ng of input gDNA was brought up to 9µL with water. 10µL tagmentation buffer and 1µL

tagmentation enzyme were added and mixed by pipetting up and down. Samples were incubated 55C for 8min, then the reaction was halted by addition of 10µL 5M guanidine thiocyanate, mixed by pipetting, and incubated at room temperature for 3min. AMPure XP beads (Beckman Coulter A63881) were used for DNA purification: 15µL AMPure beads and 25µL binding buffer [20% PEG8000, 2.5M NaCl] were added to the reaction and mixed by pipetting. The reaction was allowed to sit at room temperature 10 min before being placed on a magnet stand to attract the beads. After >2min on the magnet, the supernatant was removed from the beads and 150µL 70% ethanol was immediately added, allowed to sit for >30sec, removed, and another 150µL 70% ethanol added. After >30sec, the ethanol was removed completely and the bead pellet was allowed to dry ~30sec. The tubes were removed from the magnet stand and the pellet was resuspended in 20µL elution buffer (Qiagen EB 1014608). After 2min incubation at room temperature off of the magnet, the tubes were returned to the magnet for >2min, after which the DNA-containing liquid was transferred to a new tube. The 50ng input library preparations were performed similarly, with the following volume adjustments: 50ng of gDNA was brought up to 20µL volume with water, 25µL tagmentation buffer was used, 5µL tagmentation enzyme, 25µL 5M guanidine thiocyanate, 20µL AMPure beads, and 80µL binding buffer. Wash and elution volumes were the same.

PCR amplification of the libraries was done using Illumina NPM master mix (part of kit FC-121-1030). 10µL of the above tagmented DNA was put into each reaction, along with NPM and Illumina barcode index primers (FC-121-1012). Dual barcoding was used. PCR conditions were 72°C 3 min, 98°C 30 sec, 98°C 10 sec, 63°C 30 sec, 72°C 40 sec, with these latter three steps cycled 6 times. For trials 1 and 2 of both input amounts, SYBR green dye (Thermo Fisher S7563) was also

added to the reaction, to visually assess amplification. Post-PCR, libraries were again purified by addition of 30µL AMPure XP beads to the PCR reaction, incubation at room temperature for 5 min, incubation on the magnet for 2min, liquid removal and addition of 200µL 80% ethanol to wash followed by a second wash of the same, removal of all ethanol, letting dry for 30 sec, removal from magnet and resuspension of bead pellet in 32.5µL resuspension buffer (Illumina FC-121-1030), incubation at room temperature 2min, incubation on the magnet >2min, and recovery of purified suspended DNA into a new tube.

Final libraries were quantified by Qubit high sensitivity assessment (Invitrogen Q32854) and diluted to 2nM. Denaturation and dilution of libraries for sequencing was done as described (NextSeq Denature and Dilute Libraries Guide 15048776 Rev. D). Sequencing was done using 75bp-paired end NextSeq 500/550 High Output v2 150 Cycle kits (FC-404-2002).

S. cerevisiae

S. cerevisiae genomic DNA was extracted using a phenol:chloroform "Smash and grab" protocol. To the 1.5mL tube containing the frozen pellet of ~3 x 108 cells (1.5mL of 2-day culture), we added 0.1mL of acid-washed glass beads, 0.2mL of lysis buffer (10 mM Tris, pH 8.0, 1 mM EDTA, 100 mM NaCl, 1% SDS, 2% Triton X-100), and 0.2mL of 25:24:1 phenol:chloroform:isoamyl alcohol and vortexed for 2min. 0.2mL of TE (10mM Tris, 1mM EDTA, pH 8.0) was added, the tube was inverted to mix and centrifuged at 14,000 rpm for 5 minutes to separate the phases. The DNA-containing aqueous phase was transferred to a new 1.5mL tube containing 1mL 0.5M potassium acetate in 100% EtOH and centrifuged at 14,000 rpm for 5 minutes to precipitate the DNA. DNA was resuspended in 150µL 10mM Tris pH 8.0 + 0.1ng/mL RNase A, incubated at 37°C for 20 minutes to degrade RNA. To each RNaseA-treated sample, 150µL

phenol:chloroform:isoamyl (25:24:1) was then added, vortexed, and centrifuged again at 14,000 rpm for 5 minutes, aqueous phase transferred to a new tube. For ddPCR and sequencing library generation, 50µL of the DNA-containing aqueous phase was further purified using a Zymo Research Clean & Concentrator column (D4013) as per manufacturer instructions and resuspended in 50µL 10mM Tris pH 8.0. Libraries were prepared on 10ng yeast gDNA as for worms (above) with the exception that library amplification was done using Kapa HiFi Readymix (Kapa Biosystems KK2602).

Droplet digital PCR

S. cerevisiae genomic DNA was diluted to 0.05 ng/ul in low-bind tubes. Each 20µl reaction consisted of 10 µl 2X ddPCRTM Supermix for Probes (Bio-rad), 0.125µl EcoRI-HF (NEB, 20,000 U/mL), 1.8µl of 10µM Primer Mix (containing 10µM each rDNA F and R primers and Tub1 F and R primers), 1µl of 5µM Probe mix (containing 5µM each rDNA and Tub1 probes), and 1µl of DNA at 0.005 ng/µl. The mixture was incubated for 15 minutes for DNA digestion to occur, followed by droplet generation on a QX200 Droplet Generator (Bio-rad). Amplification was performed for 50 cycles with a 57°C annealing temperature, and droplet reading was performed on a Bio-rad QX200 Droplet Reader. Optimal annealing temperature was determined by a temperature gradient of 56-62°C with BY4741 DNA (**Figure A1.1**).

Analysis software and version

bowtie2/2.2.3 [298,403]

BWA-MEM bwa/0.7.15 [297]

samtools/1.4 and samtools/0.1.18 [307]

picard/2.14.0 [404]

trim_galore/0.4.1 [405]

cutadapt/1.8.3 [406]

fastqc/0.11.7 [407]

java/8u25

wget

python/2.7.3

R version 3.5.1 [360]

Data were visualized with ggplot2 [361]. The colorspace package [408] and Color Universal Design palette [362] were used for some visualizations.

Statistics

Pearson's correlations and p-values were calculated using the R cor.test function. To assess if ddPCR and CHEF values were significantly different, t-tests (two-tailed, unequal variance) were performed on the rDNA copy number of the six yeast strains for which we had three replicates each for the two methods.

Relative read coverage-based rDNA copy number estimation

C. elegans

Methods used as described in Thompson 2013 [1]. Reads were demultiplexed from the NextSeq and FASTQ files were aligned to the unmasked WS235 genome with bowtie2/2.2.3 to generate .sam and .bam files [298,403]. Reads mapping to multiple locations were randomly assigned to one. A custom Perl script [1] was used to count the total number of mapped reads in the .bam, and the total number of reads mapping to the rDNA coordinates, using samtools/0.1.18 [307]. rDNA coordinates (including *rrn-3.56* and *rrn-1.2*) used for WS235 were ChrI 15060299-

15071033. Copy number of rDNA was calculated by the ratio of these two counts, corrected for the length of the rDNA (7197bp) and the length of the genome (100286070bp), with the equation:

(rDNA_counts*100286070)/(total_counts*7197) = rDNA copy number

Each line of the bam, and thus each end of a paired end read, was counted independently. Read duplication removal was not used because the repetitive nature of the rDNA engenders a situation in which reads with identical starts and ends nevertheless represent true independently-generated reads and should not be removed.

S. cerevisiae

Methods were modeled after [1] and adapted for the *S. cerevisiae* genome. FASTQ files were downloaded from SRA with wget and split into forward and reverse paired read files. SRR numbers are indicated in Table S6. For in-house sequencing, reads were split as above. Split, paired FASTQ files were aligned to the unmasked S. cerevisiae S288C R64 genome. A custom Perl script was used to count the total number of mapped reads and the total number of rDNA-mapping reads.

Single copy region copy number estimation

Twenty-nine 7.2kb regions of the *C. elegans* genome were selected for use in library quality control analysis (**Table A1.5**). An original list of 32 regions was generated by extracting 7.2kb regions of non-masked sequence from the masked version of the *C. elegans* genome. One of the regions on this list was eliminated for multiple alignment. Another two were eliminated for absence in one or more of our wild isolates. The remaining 29 were analyzed for estimated copy number by read counting, in the same manner described above for rDNA copy number

estimation. A custom Perl script counted how many reads aligned to each of the 29 regions. Copy number for each region was calculated by the ratio of reads aligning for that region to total aligned reads.

Maximum likelihood estimation GC content correction (GCC) method of rDNA copy number determination

Method is based on Parks and Blanchard 2018 and Benjamini and Speed 2012 [97,261].

C. elegans

Demultiplexed, paired FASTQ files were aligned to the unmasked WS230 genome ("all reads") and to a single copy 45S rDNA sequence ("rDNA reads") with bowtie2/2.2.3 [298,403], retaining only mapped reads. Resulting all-reads .bam files were sorted with samtools/1.4 [307]. Median fragment length of all reads was determined with Picard CollectInsertSizeMetrics (picard/2.14.0, java/8u25) [404]. The .bam files were reduced to text files containing the chromosome mapped to and the leftmost mapping position for each properly aligned read on the forward strand for both all reads and rDNA reads. Maximum likelihood estimates of rDNA copy number were determined with the following equations using a custom python script. Briefly, for each sample the GC content-specific fragmentation rate (λG) for a fragment of the median fragment length was calculated based on methods described by Benjamini and Speed [261]. λG was calculated with all properly aligned reads, excluding reads that mapped to the 5S or 45S rDNA or the telomeres.

GCC Equation (based on Parks and Blanchard 2018 [97]):

Copy Number= (# Fragments mapping to rDNA)/($\sum_{i=1}^{n} \lambda_{G(p_i)}/2$)

Where the region of interest is defined by positions [pi,…,pn].

S. cerevisiae

Split, paired FASTQ files were aligned to the unmasked S. cerevisiae S288C R57-1-1 genome ("all reads") and to a single copy 45S rDNA sequence ("rDNA reads") with bowtie2/2.2.3, retaining only mapped reads [298,403]. .bam files were processed as for C. elegans. Maximum likelihood estimates of rDNA copy number were determined with a custom python script. Reads mapping to or overlapping telomeres, centromeres, Ty elements, and long terminal repeats were excluded from the analysis [409]. Reads mapping to the rDNA were excluded from the GC fragmentation calculation.

Alignment with BWA-MEM

For the data represented in **Table A1.4**, the demultiplexed *C. elegans* reads described above were aligned to the WS230 and single copy 45S rDNA sequence with BWA-MEM [297]. Sequential analysis was performed with the GCC metric as described above. Reads were not trimmed for this analysis.

Trimming

Reads for the sequencing data in the main text were not trimmed. For the trimming analysis in **Table A1.4**, the demultiplexed *C. elegans* reads described above were trimmed (maintaining read pairs) with Trim Galore, with the settings --paired -phred33 -q 20 [405]. Trimmed reads were aligned with bowtie2 analyzed with the GCC metric as described above.

Downsampling

Whole genome sequencing data from MY1 and JU775 .bam files from Thompson et al 2013 aligned to WS230 were downsampled to 90%, 50% or 5% of the original bam, using samtools/1.4 view –b –s [307]. Downsampling was done five times for each strain and each

percent, using a different seed for each. rDNA copy number of the 45S (7197bp repeat unit, in WS230 coordinates ChrI 15060288-15071022), as well as the 5S (976bp repeat unit, coordinates ChrV 17115879-17131432) was calculated by relative read coverage method:

(rDNA_counts*100286070)/(total_counts*7197) = 45S rDNA copy number

Mitochondrial genome copies were also counted as number of reads aligning to the mitochondrial genome relative to average total genome coverage:

(mtDNA_counts*100286070)/(total_counts*13794) = mitochondrial copy number

APPENDIX 2: SUPPLEMENTAL DATA, ANALYSIS, AND INFORMATION FROM CHAPTER 3

This appendix contains the supplemental file relevant to Chapter 3, as published in: Thousands of high-quality sequencing samples fail to show meaningful correlation between 5S and 45S ribosomal DNA arrays in humans

Ashley N. Hall, Tychele N. Turner*, Christine Queitsch*

*Co-corresponding authors

A2.1: SUPPLEMENTAL FIGURES

The following supplemental figures are included in this Appendix:

Figure A2.1	Additional comparisons of rDNA copy number of probands with differing IQ
Figure A2.2	Correlations of rDNA copy numbers in 1000 Genomes Project data
Figure A2.3	Replicability of 5.8S and 28S rDNA copy number estimates between the high- and low- coverage 1000 Genomes Project data
Figure A2.4	Distribution of 18S copy number estimates by sequencing center
Figure A2.5	Comparison of copy number estimates of regions of the 45S rDNA repeat to each other
Figure A2.6	Data quality metrics for the Simons Simplex Collection

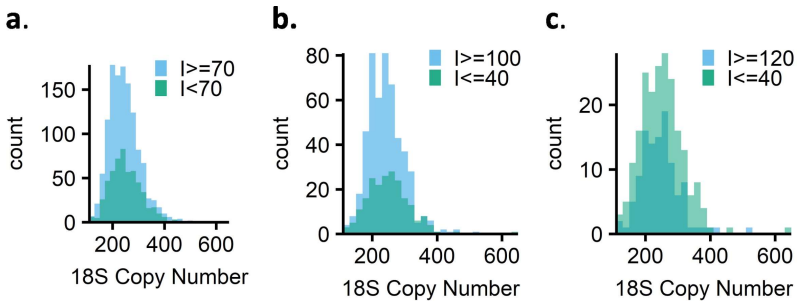

Figure A2.1: Additional comparisons of rDNA copy number of probands with differing IQ.
a: Comparison of rDNA copy number in probands at the cutoff of pathological IQ (I<70, n=548 compared to I>=70, n=1,244). Welch Two Sample t-test p-value: 0.03. **b:** Comparison of rDNA copy number of probands with IQ<=40 (n=210) and IQ>=100 (n=500). Welch Two Sample t-test p-value: 0.1103. **c:** Comparison of rDNA copy number of probands with a more stringent IQ cutoff: I<=40 (n=210) compared to I>=120 (n=112). Welch Two Sample t-test p-value: 0.4129.

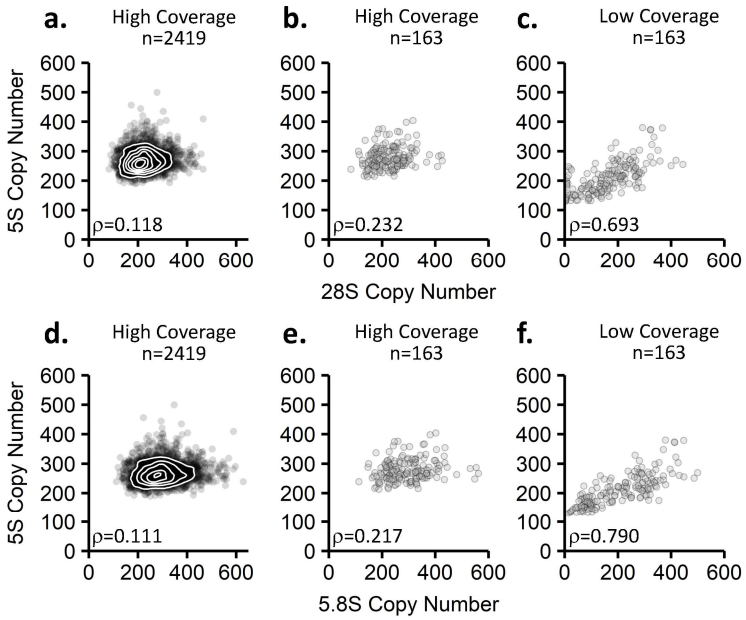

Figure A2.2: Correlations of rDNA copy numbers in 1000 Genomes Project data.

a-c: Comparison of the 28S to 5S rDNA copy numbers in the **a:** high-coverage 1000 Genomes Project data (n=2,419), **b:** subset of high-coverage 1000 Genomes Project data also analyzed in the low-coverage dataset (n=163), and **c:** low-coverage 1000 Genomes Project data (n=163). Bottom: Comparison of the 5.8S to the 5S rDNA copy numbers in the **d:** high-coverage 1000 Genomes Project data (n=2,419), **e:** subset of high-coverage 1000 Genomes Project data also analyzed in the low-coverage dataset (n=163), and **f:** low-coverage 1000 Genomes Project data (n=163).

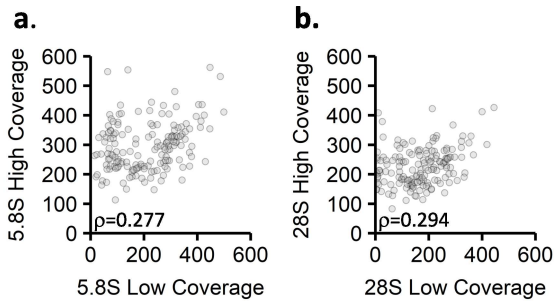

Figure A2.3: Replicability of 5.8S and 28S rDNA copy number estimates between the high- and low- coverage 1000 Genomes Project data.

a: Comparison of 5.8S estimates between high- and low- coverage datasets (n=163). **b**: Comparison of 28S estimates between high- and low- coverage datasets (n=163).

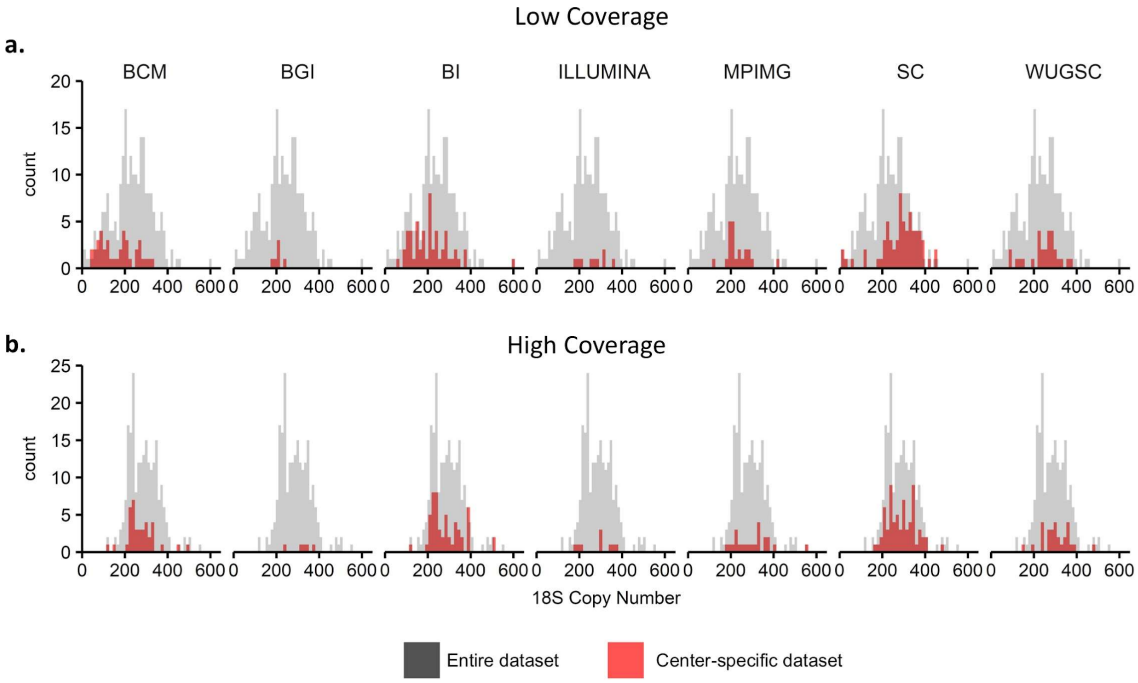

Figure A2.4: Distribution of 18S copy number estimates by sequencing center.

a: Distribution of 18S copy number estimates from individual libraries sequenced by each of 7 sequencing centers (gray) for the 163 samples analyzed in the low coverage data (n=222 distinct libraries). Copy number estimates for samples produced by a given center in the low-coverage sequencing are highlighted in red. **b**: Distribution of 18S copy number estimates from the high-coverage 1000 Genomes Project data for the 163 samples analyzed in the low coverage data. The high-coverage data were not generated by any of these 7 centers: '**b**' serves to demonstrate whether a center was assigned samples with a higher or lower copy number distribution. Samples that were composed of multiple distinct libraries in the low-coverage data are represented multiple times in the high-coverage plot.

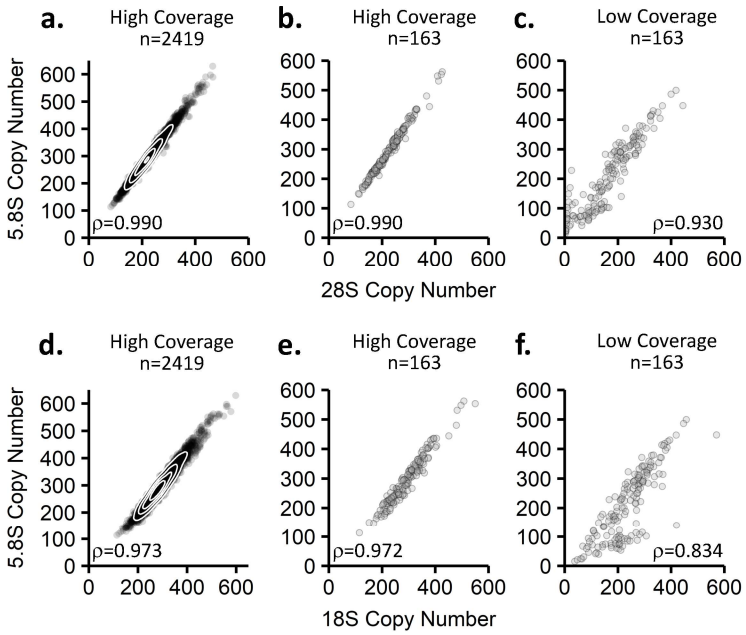

Figure A2.5: Comparison of copy number estimates of regions of the 45S rDNA repeat to each other.

a-c: Comparison of the 28S to 5.8S rDNA copy numbers in the 1000 Genomes Project datasets. **a**: High-coverage 1000 Genomes Project data (n=2,419). **b**: Subset of high-coverage 1000 Genomes Project data also analyzed in the low-coverage dataset (n=163). **c**: Low-coverage 1000 Genomes Project data (n=163). **d-f**: Comparison of the 18S to the 5.8S rDNA copy numbers in the 1000 Genomes Project datasets. **d**: High-coverage 1000 Genomes Project data (n=2,419). **e**: Subset of high-coverage 1000 Genomes Project data also analyzed in the low-coverage dataset (n=163). **f**: Low-coverage 1000 Genomes Project data (n=163)

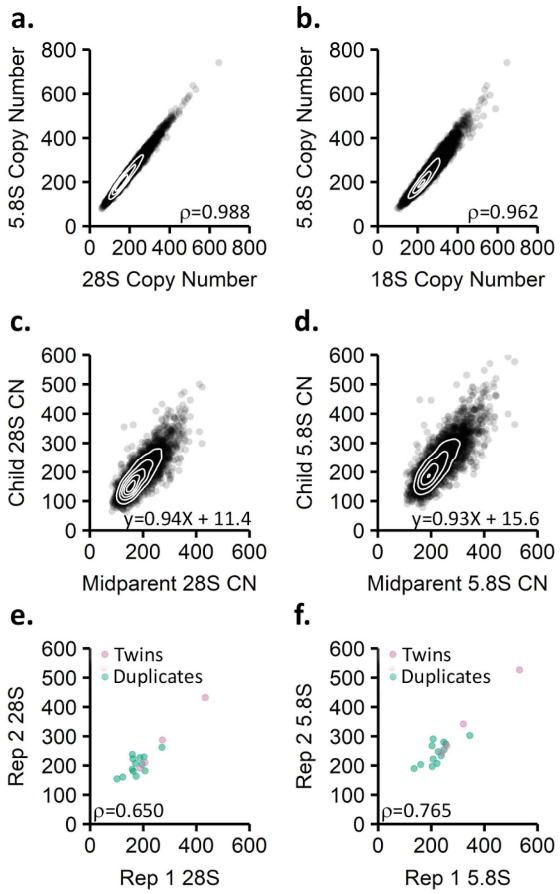

Figure A2.6: Data quality metrics for the Simons Simplex Collection.

a: Comparison of the 28S to 5.8S rDNA copy numbers (n=7,210). **b**: Comparison of the 18S to 5.8S rDNA copy numbers (n=7,210). **c**: Heritability estimate of the 28S rDNA region (n=3,548). **d**: Heritability estimate of the 5.8S rDNA region (n=3,548). **e-f**: Comparison of 5.8S (**e**) and 28S (**f**) copy number estimates for either monozygotic twins (n=4 pairs) or for individuals sequenced twice in the Simons Simplex Collection (n=13). Spearman correlation indicated is for monozygotic twins and duplicates analyzed together.

APPENDIX 3: SUPPLEMENTAL INFORMATION AND DISCUSSION RELEVANT TO CHAPTER 4

A3.1: *C. ELEGANS* RESOURCES WITH DIFFERING RDNA COPY NUMBER - FROM THE LITERATURE

In Chapter 4, I detail the strain resources that I made to study rDNA copy number variation in *C. elegans*. I discuss the presence of rDNA copy number variation in wild isolates of *C. elegans*, but there are other studies that include rDNA copy number variation, even if it was not specifically analyzed. In this appendix, I will discuss examples of this sometimes-hidden rDNA copy number variation. This list is not exhaustive, as many duplications and deficiencies in *C. elegans* are not completely mapped, and we only have CHEF-verified rDNA copy numbers for a small number of wild isolates [5].

A3.1.1: Existing recombinant inbred lines or QTL analyses in *C. elegans*

Most RILs that have been generated in *C. elegans* use the Hawaiian strain CB4856, which has a similar rDNA copy number to the laboratory strain N2 [1,410,411]. Aside from the MY1xSEA51 RILs presented in this dissertation, no RILs have been produced with differences in rDNA copy number as the central focus. However, some do include parental strains that differ in rDNA copy number, and all resources have the potential to incorporate rDNA copy number variation. There are multiple resources summarizing QTL or GWAS studies in *C. elegans* and include a tool through *C. elegans* Natural Diversity Resource (CeNDR) and the WormQTL2 tool [311,412]. Estimating rDNA copy number in these resources could make consideration of rDNA copy number more accessible to many researchers.

Of existing studies on RILs, some include wild isolate parents with high or low rDNA copy numbers. A RIL panel of N2 and MY16 (73 rDNA copies) was used to investigate the regulation of endoderm development [413]. RC301, with 420 rDNA copies, was used to map lifespan-affecting loci in a cross with the Bergerac strain, a strain with a high number of Tc1 transposons [414,415]. Neither study identified the right end of chromosome I (and thus, the rDNA) as a contributing locus. In the case of a cross between a wild type and a single wild isolate, SNVs proximal to the rDNA can be used to tag which parent the rDNA locus came from. This works because we observe a low frequency of spontaneous rDNA copy number change in crossing. Analysis of linked SNVs is, in fact, how I produced some of the NILs discussed in Chapters 4 and 5, which were then validated by CHEF gel to confirm that selection on the linked SNV selected for the rDNA copy number of interest.

Multiparental RILs, such as the *C. elegans* Multiparental Experimental Evolution (CeMEE) panel and others carry rDNA copy number variants, as introduced by the parents [416–418]. The CeMEE panel includes wild isolates MY1 and RC301, with the highest rDNA copy numbers of wild strains, as well as MY16, with one of the lowest rDNA copy numbers. With these multiparental RILs, linked SNVs are not useful in tagging specific rDNA copy numbers, because the SNVs proximal to the rDNA array are shared among many wild isolates [311]. I estimated the rDNA copy numbers of the CeMEE RILs from their short read sequencing data, but these data were produced with multiple library preparation methods and are likely not accurate [416]. Therefore, I consider the copy numbers of the CeMEE RILs to be undetermined. Whether each final line carries a specific rDNA array from a single parent or may have an array that is a recombined variant, is unknown.

A3.1.2: Differing numbers of rDNA-bearing chromosomes in *C. elegans*

A diploid *C. elegans* nucleus contains two 45S rDNA arrays: One on each copy of chromosome I. With various duplications and deficiencies, the number of rDNA arrays in a worm can be manipulated from zero to four arrays per diploid nucleus (**Table A3.1**). Starting at the fewest number of rDNA arrays possible a complete deletion of the rDNA locus was engineered by the Fire lab [162]. The complete rDNA deletion can be maintained as a viable balanced heterozygote and two alleles of this deletion exist, termed *ccDf2620* and *ccDf2621*. The total rDNA deletion was used to show that even with zero rDNA copies, worms can complete embryogenesis using their maternally-inherited ribosomes and then arrest in the first larval stage [162]. The deletion also permits analysis of worms carrying only one rDNA array per diploid genome with heterozygotes. Through use of strains with higher rDNA copy number balancing the deletion, one could feasibly dissect phenotypes caused by having a single rDNA locus per diploid genome versus due to a copy number difference by comparing, for example, worms with 100/100 and 200/Δ*rDNA* copy number genotypes.

Beyond this deletion, rDNA duplications exist. The free duplication *sDp1* covers the right half of chromosome I, including at least some of the rDNA locus [381]. *sDp1* is maintained in one copy; so worms with wild type chromosome I and *sDp1* have three rDNA arrays per diploid cell. *eDp20* duplicates part of chromosome I - including the whole rDNA array - onto the end of chromosome II [381]. Worms with a wild type chromosome I and a chromosome II carrying *eDp20* are viable and have four rDNA arrays per diploid cell. In *S. cerevisiae*, the position of the rDNA on chromosome XII is important for its function [419]. Through use of the *eDp20* duplication, one could start to investigate whether carrying the rDNA on chromosome I is necessary for wild type

worm phenotype. For example, one could make worms that are homozygous for the chromosome I rDNA deletion and for *eDp20*. While the duplication of the non-rDNA portion of chromosome I could be a confounding variable, control strains could be constructed to account for the duplicated non-rDNA sequence.

Table A3.1: Strains with duplication and deletion alleles of the rDNA in *C. elegans*

Strain	Genotype	rDNA	Citation	Available at the CGC
BC159	*dpy-5*(*e61*) *unc-13*(*e51*) I; *sDp1* (I;f)	*sDp1* has rDNA, so a strain with *sDp1* will have 3 rDNA arrays	[381]	Yes
PD2620	*ccDf2620*/*unc-54*(*e1152*)*	Balanced CRISPR deletion of the rDNA	[162]	No
PD2621	*ccDf2620*/*unc-54*(*e1152*)*	Balanced CRISPR deletion of the rDNA	[162]	No
CB3740	*eDf24* I; *eDp20* (I;II); *mnT12* (IV;X).	*eDf24* deletes part of the rDNA, leaving ~10 copies. *eDp20* duplicates the right arm of chrI onto chrII, thus duplicating the rDNA array.	[381]	Yes

*The allele is annotated as *unc-54*(*e1152*), however, we have since determined that these strains and the CGC stock of *unc-54*(*e1152*) do not contain the dominant GK(852,853) → RM mutation, but instead a Q816 → stop mutation (Morton 2021, unpublished).

A3.2 Supplemental Figures and Tables

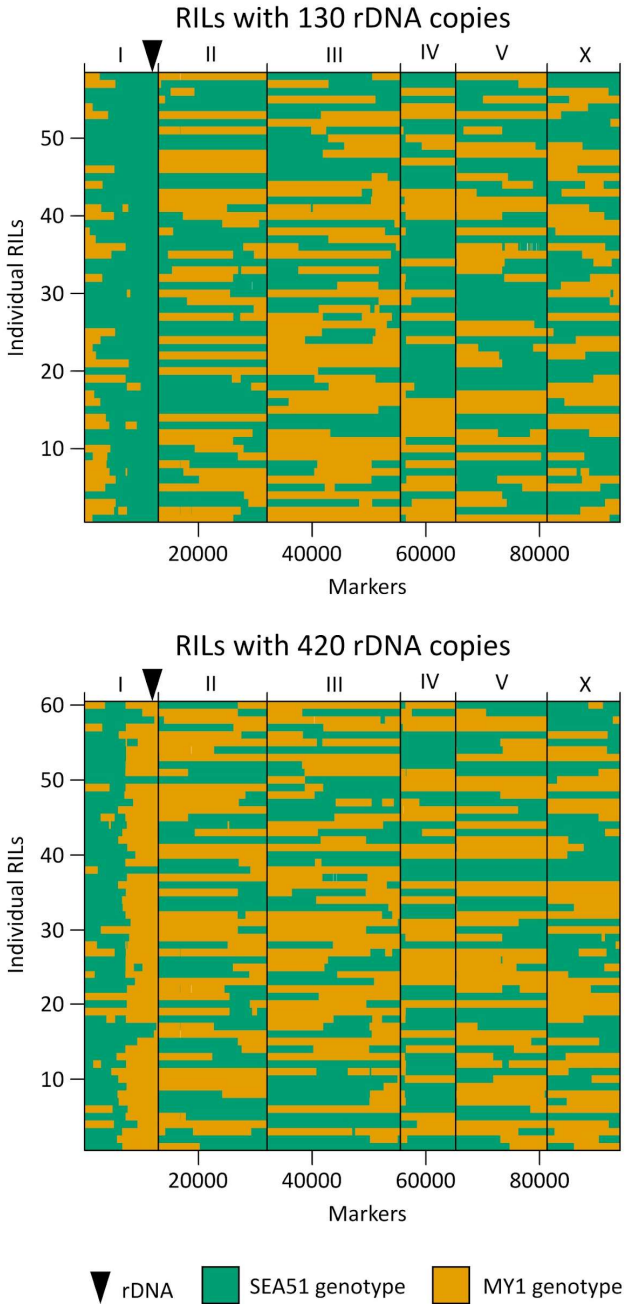

Figure A3.1: Haplotype blocks of RILs.

Top: The genotypes of the 58 RILs with 130 rDNA copies are presented. Bottom: The genotypes of the 60 RILs with 420 rDNA copies are presented. The black carat indicates the rDNA locus at the end of chromosome I. Genotypes were determined with GATK HaplotypeCaller and the map was filled with the max marginal method.

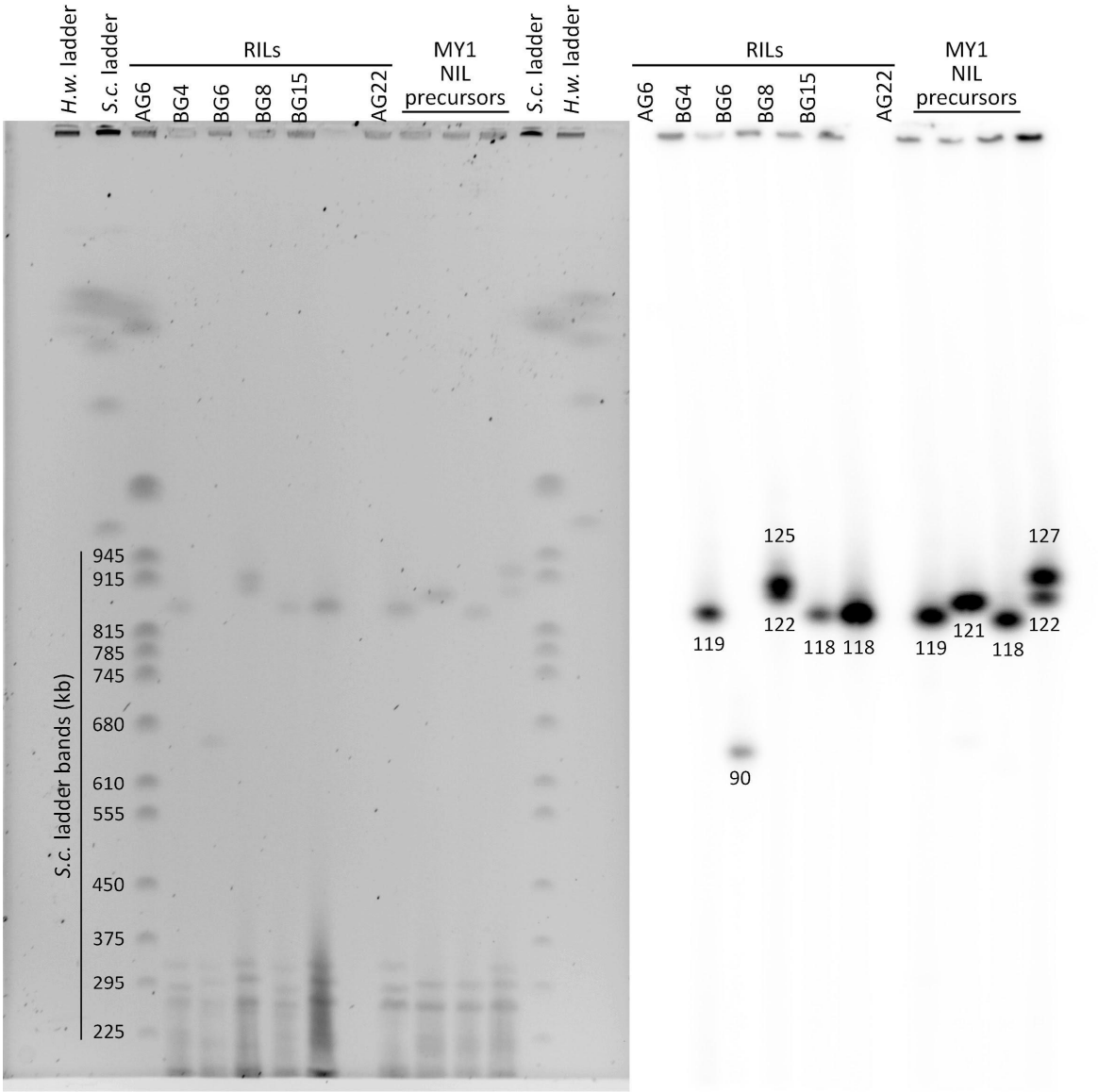

Figure A3.2: CHEF gel validation of rDNA copy number of some RILs with the *mIs13* transgene. Southern blot of six RIL strains and three MY1 NILs (precursors to those shown in Figure 4.3 and Figure A3.3) that were run on a CHEF gel under conditions to separate arrays near 100 copies in length. Copy numbers were calculated from an *S. cerevisiae* chromosome ladder that was visualized by ethidium bromide staining (left) prior to Southern blotting with an rDNA-specific probe (right). All strains in this figure also have the *mIs13* transgene. *S.c.* ladder = *S. cerevisiae* chromosomes ladder, *H.w.* ladder = *H. wingei* chromosomes ladder. Allele designation for each strain is indicated at the top. Ladder sizes on EtBr stain are indicated in kilobases, rDNA band sizes on Southern blot are indicated in number of repeat units (Band size in kb divided by 7.2kb repeat unit size).

Table A3.2: Comparison of rDNA copy number estimates of RILs from WGS and CHEF gel.

RIL	WGS	CHEF
AG6	116.6	119
BG4	97.5	90
BG6	151.9	123.5*
BG8	134	118
BG15	143	118
AG22	127	119

*Average of two bands.

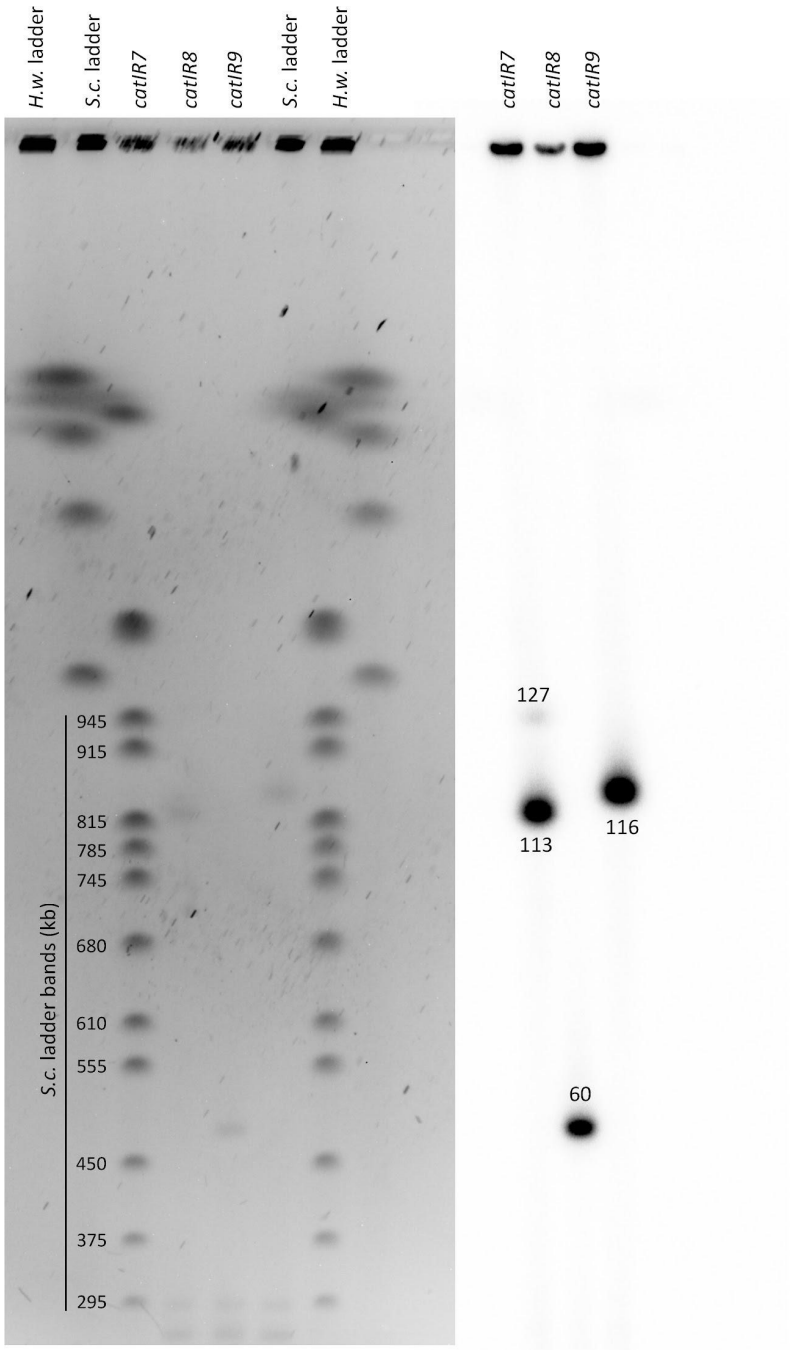

Figure A3.3: rDNA copy number estimation by CHEF gel in MY1 NILs

MY1-background NILs separated with conditions that resolve arrays near 100 rDNA copies. Copy numbers were calculated from an *S. cerevisiae* chromosome ladder that was visualized by ethidium bromide staining (left) prior to Southern blotting with an rDNA-specific probe (right). *S.c.* ladder = *S. cerevisiae* chromosomes ladder, *H.w.* ladder = *H. wingei* chromosomes ladder. Allele designation for each strain is indicated at the top. Ladder sizes on EtBr stain are indicated in kilobases, rDNA band sizes on Southern blot are indicated in number of repeat units (Band size in kb divided by 7.2kb repeat unit size).

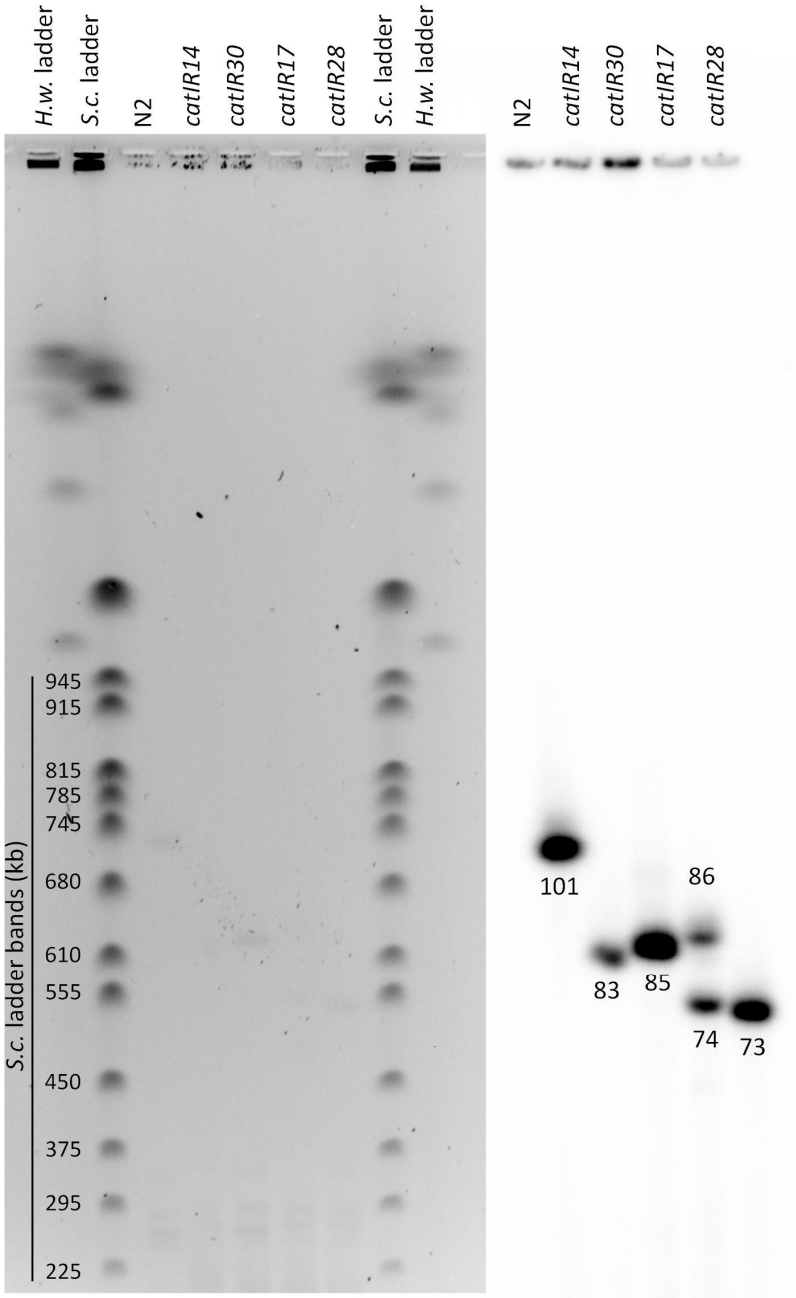

Figure A3.4: rDNA copy number estimation by CHEF gel in N2 NILs with lower copy numbers. N2 and NILs were separated with conditions that resolve arrays near 100 rDNA copies. Copy numbers were calculated from an *S. cerevisiae* chromosome ladder that was visualized by ethidium bromide staining (left) prior to Southern blotting with an rDNA-specific probe (right). *S.c.* ladder = *S. cerevisiae* chromosomes ladder, *H.w.* ladder = *H. wingei* chromosomes ladder. Allele designation for each strain is indicated at the top. Ladder sizes on EtBr stain are indicated in kilobases, rDNA band sizes on Southern blot are indicated in number of repeat units (Band size in kb divided by 7.2kb repeat unit size).

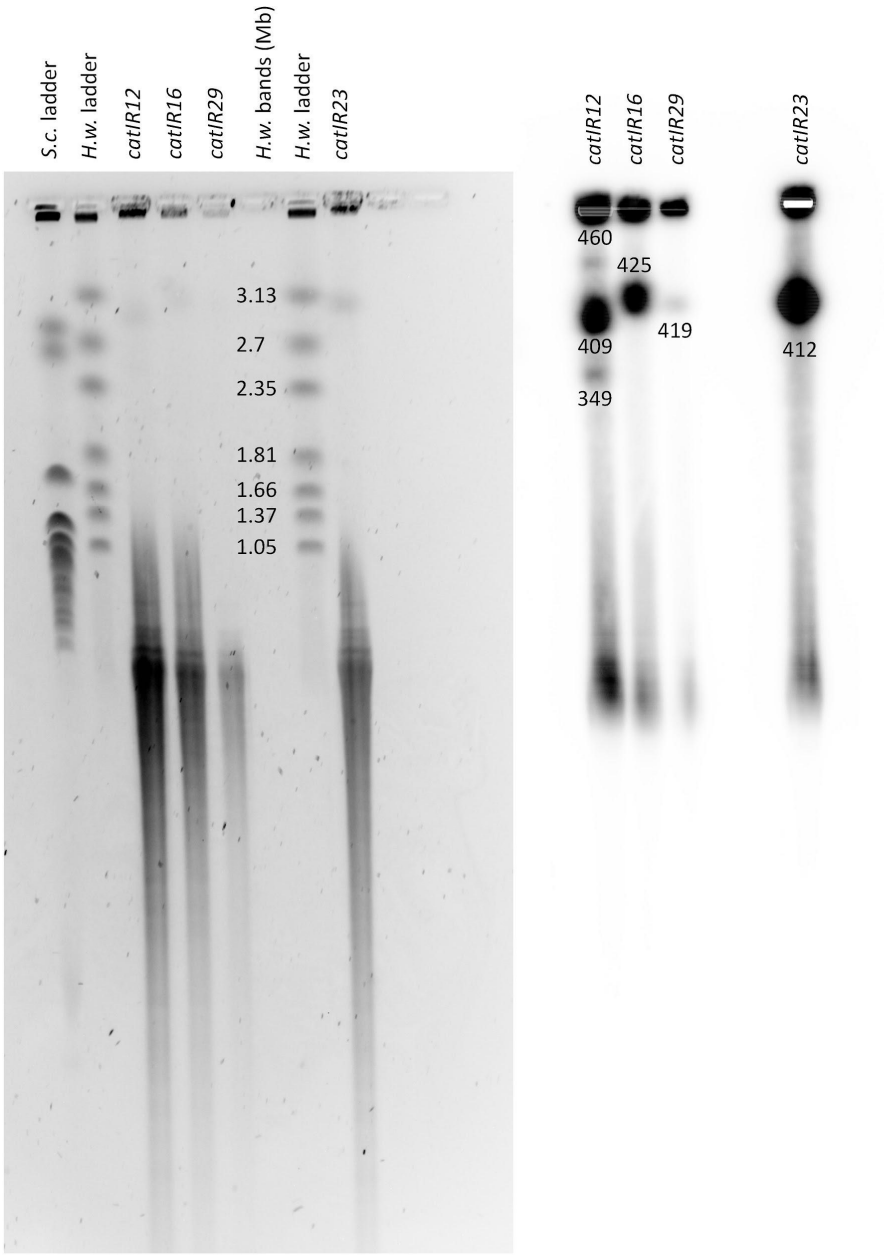

Figure A3.5: rDNA copy number estimation by CHEF gel in N2 NILs with higher copy numbers. NILs with high rDNA copy number separated with conditions that resolve larger arrays. Copy numbers were calculated from an *H. wingei* chromosome ladder that was visualized by ethidium bromide staining prior to Southern blotting with an rDNA-specific probe. *S.c.* ladder = *S. cerevisiae* chromosomes ladder, *H.w.* ladder = *H. wingei* chromosomes ladder. Allele designation for each strain is indicated at the top. Ladder sizes on EtBr stain are indicated in megabases, rDNA band sizes on Southern blot are indicated in number of repeat units (Band size in kb divided by 7.2kb repeat unit size).

A3.3 Supplemental methods: Strain construction details of NILs

Different NILs were made with slightly different cross strategies, which will be detailed here. NILs produced with the same cross strategy are grouped together. Table A3.3 details the various markers used to genotype NILs during construction.

SEA295 and SEA296

First, strain SEA290 was constructed by crossing RIL AG22 (SEA102) into MY1 6x then singled 6x. In singling, homozygous GFP+ worms were selected to have SEA51 homozygous rDNA genotype. Then, males of SEA290 were backcrossed to MY1 6x and progeny of the sixth cross singled 6x, selecting homozygous GFP+ worms at the second generation of single worm propagation. The resulting strains have 130 rDNA copies (SEA295, the copy number that SEA51 is estimated to have) and 64 rDNA copies (SEA296), the latter of which is a chance reduction that occurred during the crossing and propagation process and was only identified after strain construction concluded.

SEA300

First, strain SEA292 was constructed by crossing RIL BO20 (hermaphrodite; MY1 rDNA, SEA159) to SEA51 (male), selecting GFP+ hermaphrodite progeny in the L4 stage, and crossing those L4 progeny to SEA51 males again. This cross to SEA51 males was repeated a total of six times, and individual hermaphrodite progeny from the sixth cross were singled for an additional six generations. To ensure MY1 genotype rDNA, in the second generation of single worm propagation, worms were visually assessed with a fluorescent dissecting microscope and GFP- worms were selected for continued propagation. SEA292 has MY1 genotype mtDNA. To generate SEA300, which has N2 mtDNA, males of SEA292 were crossed to SEA51 hermaphrodites. Progeny

with dim GFP expression (heterozygous GFP+/-) were singled and a GFP- F2 progeny was selected to grow up to freeze down as strain SEA300. Two additional strains were saved from these same crosses (but from separate GFP- F2s); these are strains SEA298 and SEA299 which remain in the Queitsch lab *C. elegans* library but were not used in this dissertation.

SEA302, SEA304, and SEA305

Wild isolates used to make these strains are JU775, RC301, and MY16, respectively. Male SEA51 was crossed with hermaphrodites of the given wild isolate. GFP+ F1 progeny were selected, self-fertilized, and GFP- F2 progeny were crossed to SEA51 males. Crossing to SEA51 was performed a total of 8 times, with the last backcross being with a SEA51 hermaphrodite to restore wild-type mitochondrial DNA. After the final cross, progeny were propagated by single worm propagation for six generations, selecting for homozygous GFP- worms in the second generation.

SEA328, SEA329, and SEA330

These NILs were made by further backcrosses of SEA305, SEA304, and SEA302, respectively. After crossing, specific positions proximal to the rDNA were genotyped either by genotyping PCR with primers specific to a SNV present in N2 or the wild isolate, by restriction digest of a PCR amplified fragment of a region carrying a restriction site polymorphism in N2 as compared to the wild isolate, or by amplification of a region with a large (>100bp) insertion or deletion in N2 as compared to the wild isolate (**Table A3.3, Table A4.11**). Once candidate worms were identified with the desired rDNA-proximal and rDNA-distal genotypes, these worms were selfed and their progeny screened for homozygotes.

Table A3.3: Genotyping loci for NIL construction

Shorthand	Description	Differentiates between
GFP	*mIs13* presence vs absence	SEA51 or *mIs13*-linked rDNA and non-*mIs13*-linked rDNA
12.55 Mb	HhaI RFLP amplified by AHC13+AHC14	RC301 and MY16 cut with HhaI; N2 does not
14.5 Mb	Deletion amplified with primers AHC1+AHC2	JU775 has 300bp deletion as compared to N2
14.68 Mb	SNV detected with genotyping primers	AHC37+AHC48 amplify N2; AHC37+AHC49 amplify MY1, RC301, JU775, and MY16
14.99 Mb	MnlI RFLP amplified with primers AHC32+AHC35	MY1, RC301, JU775, and MY16 do not cut; N2 cuts

A3.3 Anecdotal stability of rDNA arrays in NILs

A key question when working with our *C. elegans* strains with differing rDNA copy numbers is whether those copy numbers are maintained in the NILs. In Chapter 4, we propagated the RILs with 130 and 420 rDNA copies for 20 generations and found no evidence of gross instability of the rDNA loci. I did not perform a formal propagation experiment for the NILs but measured rDNA copy numbers of worms that had been propagated for varying amounts of time in the lab to ensure large copy number changes had not arisen. Specifically, I analyzed the MY1 64-rDNA NIL (allele *catIR8*). The populations analyzed include worms propagated for approximately 40 generations and 80 generations from the initial stock of the NIL, and approximately 40 generations from a fresh thaw of the NIL freezer stock. These estimates of numbers of generations come from the assumption that the NIL was propagated every three days, and therefore represent a maximal number of generations. For each sample, I prepared 5 CHEF plugs: One plug from a population of worms grown for two generations from 10 starting

worms, and four plugs from independent populations of worms grown for two generations from a single starting worm each.

From a given starting population of worms, absolute copy number estimates vary slightly from plug-to-plug (**Figure A3.6**). The samples propagated for 80 generations and the sample from the saved freezer stock have the most similar rDNA copy numbers, with all samples having approximately 70-80 rDNA copies. The initial stock that was propagated for 40 generations is estimated to have a lower rDNA copy number, of 65-68. Overall, the copy numbers are similar even with the many generations of propagation: There is no evidence of rDNA magnification that would quickly restore the reduced rDNA copy numbers to the native MY1 copy number of 417 copies. Interestingly, even in the samples where a single worm was used to initiate a population for CHEF analysis, two rDNA bands of near-equal intensity can be observed (**Figure A3.6**, 12th and 15th lanes). These may be instances of a heterozygous worm having been picked to initiate the population.

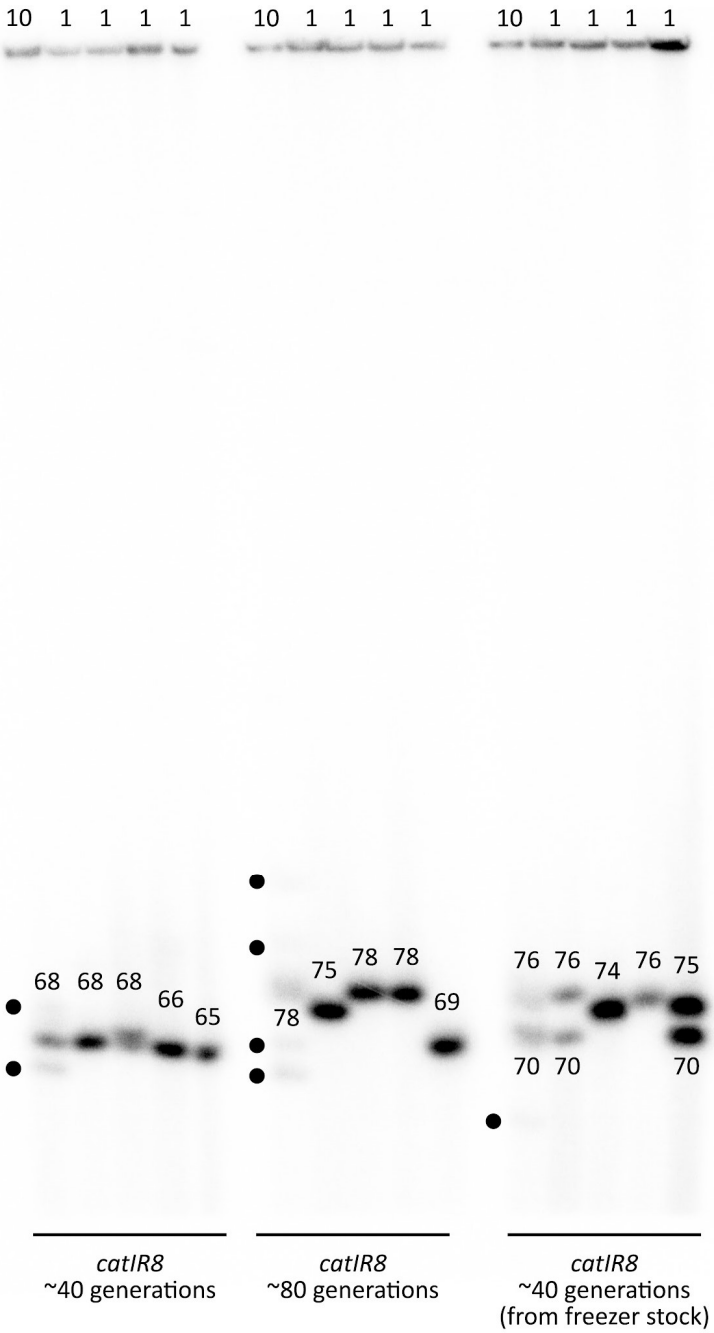

Figure A3.6: rDNA copy numbers of the MY1 64-rDNA NIL are relatively stable after rounds of propagation. Three sets of *catIR8* samples were separated with conditions that resolve arrays near 100 rDNA copies. Copy numbers were calculated from an *S. cerevisiae* chromosome ladder that was visualized by ethidium bromide staining prior to Southern blotting. For each sample, one plug was made with a population initiated with 10 starting worms, and four were made with populations initiated from single starting worms (indicated at top). Black dots to the left of each 10-worm plug indicate faint minor bands that were not quantified for copy number.

APPENDIX 4: SUPPLEMENTAL DATA, ANALYSIS, AND INFORMATION FROM CHAPTER 5

Appendix 4 contains both additional analyses and experiments relevant to Chapter 5, as well as any supplemental figures and tables referenced in Chapter 5.

A4.1 SUPPLEMENTAL FIGURES AND TABLES:

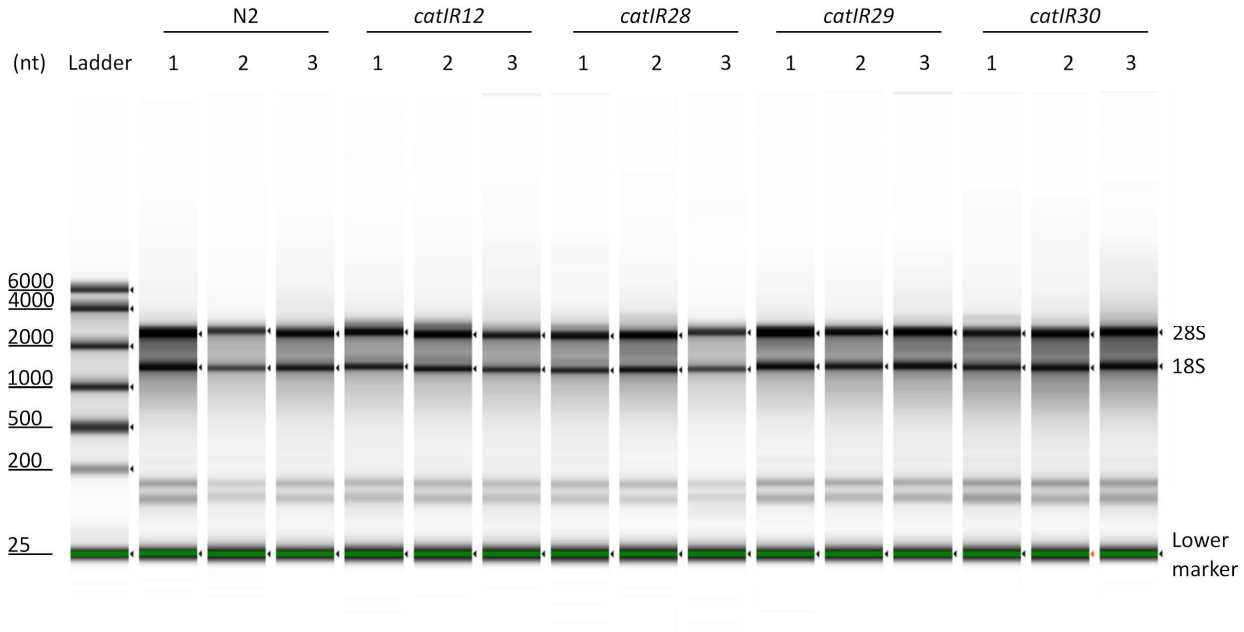

Figure A4.1: Tapestation RNA gel image used for 18S and 28S rRNA quantification. Equal quantities of RNA (100ng) extracted from day 1 adult worms were run using the Agilent RNA ScreenTape®.

Table A4.1: rRNA quantities determined from Tapestation RNA gel.

rDNA allele	**18S rRNA (ng/µl)**			**28S rRNA (ng/µl)**		
	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
N2	30.1	13.1	21.2	38.1	14.1	25.8
catIR12	18.6	16.2	15.8	19.6	21.6	19
catIR28	19.8	18.8	12.8	18.2	23.8	14.6
catIR29	23.4	20.6	24.5	37.9	21.9	35.8
catIR30	21.8	24.9	28.5	20.9	26.2	39

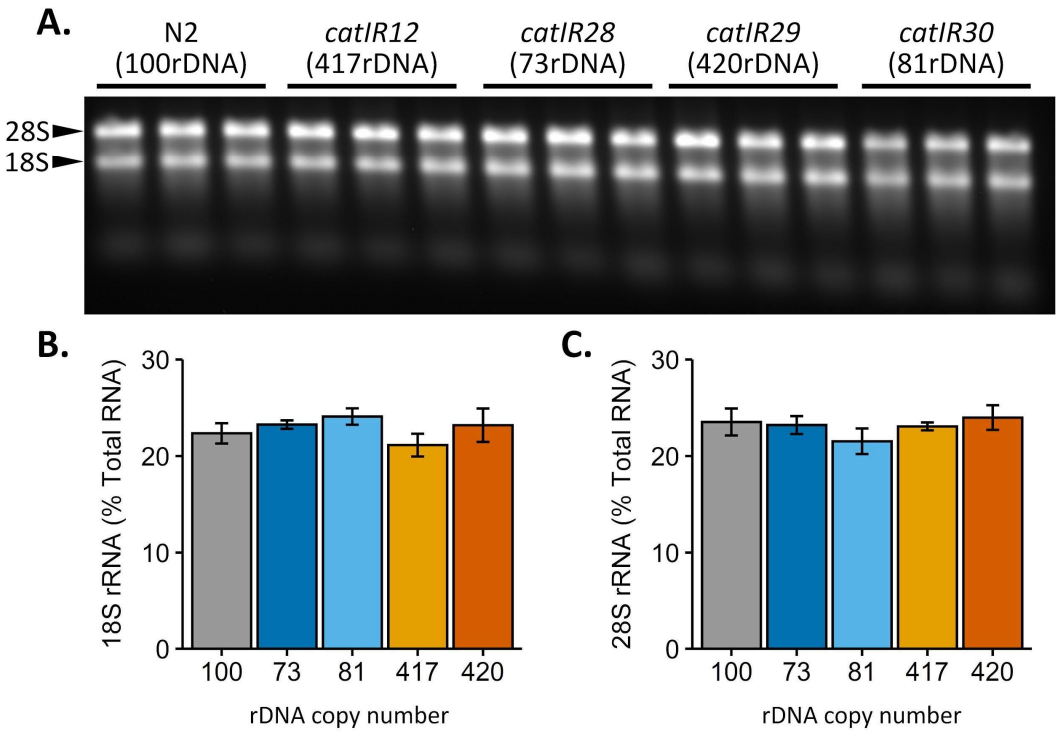

Figure A4.2: rRNA levels in N2 NILs do not differ from N2 in a copy-number-dependent manner. **A:** Equal quantities of RNA extracted from day 1 adult worms were run denatured and run on an agarose gel for each sample. Each lane is a separate biological replicate. **B** and **C:** Integrated intensity of 28S and 18S bands from **A** was quantified with ImageJ and normalized to the total integrated intensity of the sample. No samples were statistically significantly different from each other; all had p>0.05 by ANOVA and Tukey's HSD.

Table A4.2: Percent of total rRNA attributable to 18S or 28S rRNA in N2 NILs.

rDNA allele	**18S rRNA**			**28S rRNA**		
	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
N2	25.1%	22.8%	22.7%	22.9%	23.0%	21.1%
catIR12	23.5%	23.0%	22.7%	20.8%	20.2%	22.4%
catIR28	23.5%	23.9%	22.2%	23.0%	23.0%	23.8%
catIR29	25.4%	23.1%	23.4%	21.2%	24.1%	24.3%
catIR30	20.1%	21.8%	22.7%	23.4%	23.9%	25.0%

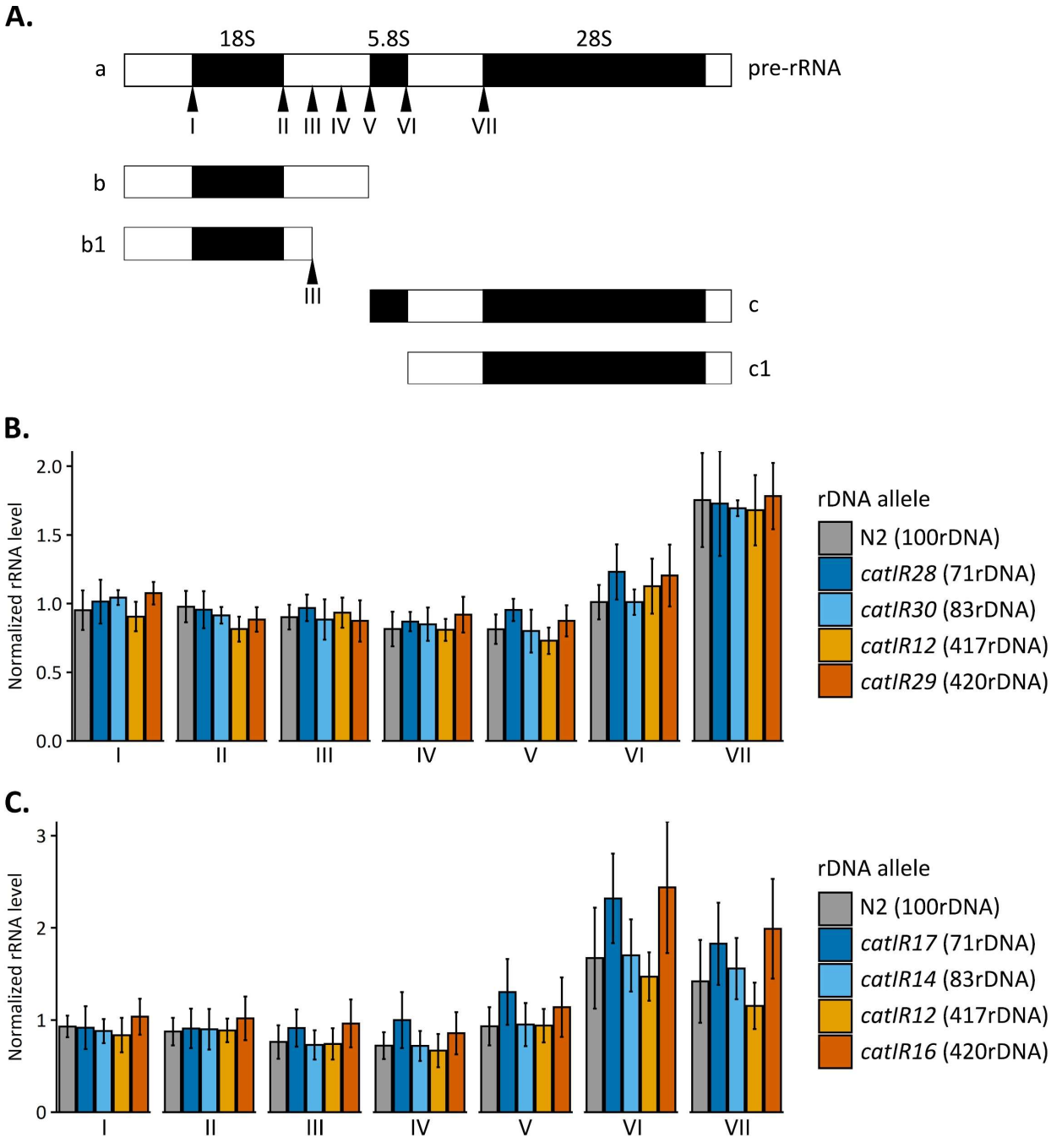
A.
18S
5.8S
28S
a
pre-rRNA
I
II
III
IV
V
VI
VII
b
b1
III
c
c1
B.
Normalized rRNA level
2.0
1.5
1.0
0.5
0.0
I
II
III
IV
V
VI
VII
rDNA allele
N2 (100rDNA)
catIR28 (71rDNA)
catIR30 (83rDNA)
catIR12 (417rDNA)
catIR29 (420rDNA)
C.
Normalized rRNA level
3
2
1
0
I
II
III
IV
V
VI
VII
rDNA allele
N2 (100rDNA)
catIR17 (71rDNA)
catIR14 (83rDNA)
catIR12 (417rDNA)
catIR16 (420rDNA)

Figure A4.3: Steady-state levels of 45S pre-rRNA do not differ in NILs with differing rDNA copy numbers.

A: Diagram of 45S pre-rRNA processing in *C. elegans*, adapted from Wu 2018 [327]. Primer pairs are as indicated in **Table A4.8**. **B-C:** pre-rRNA levels in NILs with less linked wild isolate DNA (**B**) and more linked wild isolate DNA (**C**). rRNA levels are normalized to actin. Error bars are mean ± standard deviation. No significant differences are present between any NILs and N2 (ANOVA with Tukey Honest Significant Difference test). Additional probes in the 5'ETS and at the border of the 28S and 3'ETS were used in the Wu 2018 paper but are not presented here because the CT values were too close to those of the no RT control.

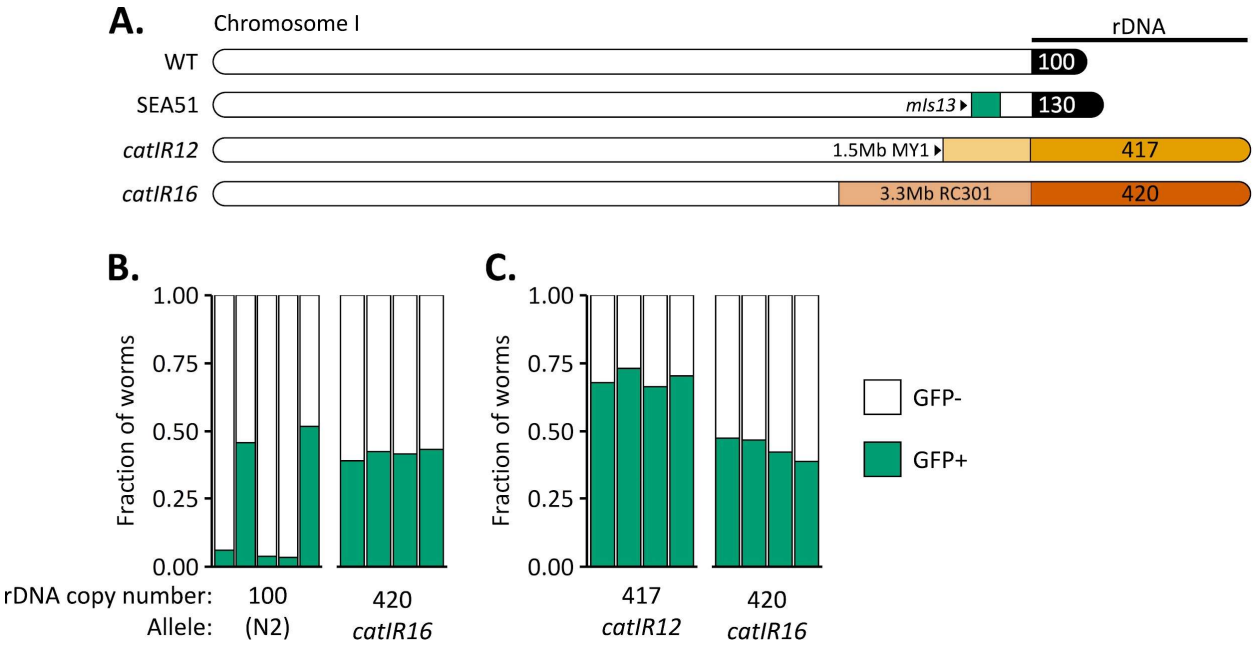

Figure A4.4 The competitive fitness defect of the 420-rDNA strain (allele *catIR16*) is not as severe as that of the 417-rDNA strain.

Competitions of high rDNA copy number strains (GFP-; alleles indicated at bottom) against SEA51 (GFP+; 130 rDNA copies). **A:** Schematic of chromosome I of NILs used in this assay. **B:** Five replicates of N2 competed against SEA51 and four replicates of the 420-rDNA NIL (allele *catIR16*) competed against SEA51, propagated at the same time. **C:** Competitions of two strains with high rDNA copy number against SEA51, with four replicate plates each, propagated at the same time. Strains used are the 417rDNA NIL (allele *catIR12*) and 420rDNA NIL (allele *catIR16*). For all panels, each bar shows the relative proportion of worms after ~11 generations that are GFP+ (green) or GFP- (white). At least 1000 worms were quantified to determine the proportion of each bar.

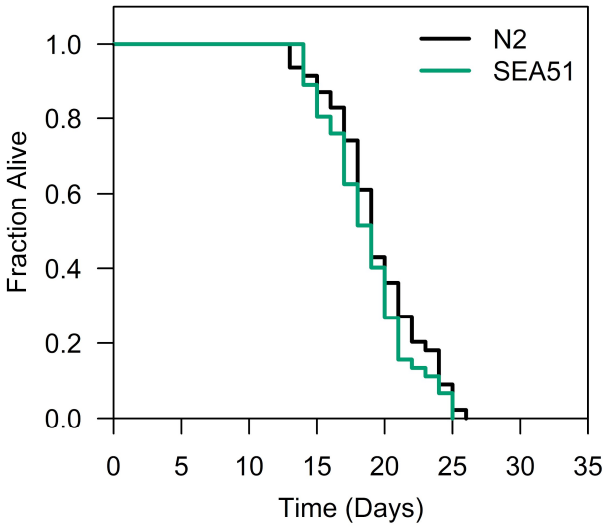

Figure A4.5: The *mIs13* transgene in SEA51 does not confer a lifespan defect.
Manual lifespan of wild type N2 and the *mIs13* transgene-bearing SEA51 strain. Median lifespan of both strains is 19 days, and there is no significant difference between lifespans ($p>0.05$, log-rank test). Assay was performed at 20°C and FuDR was used in plates to prevent progeny production and nystatin was included.

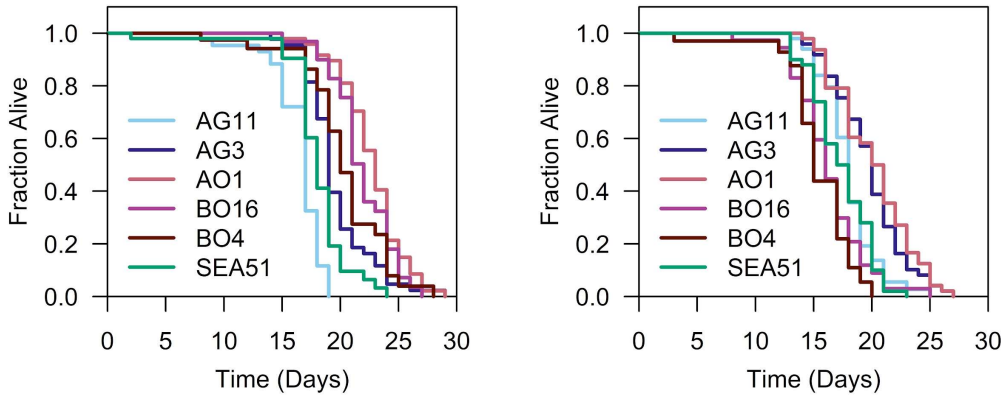

Figure A4.6: Manual validation of lifespans for five RIL strains.
Left and right panels are two separate replicates of manual lifespans for five RILs and the SEA51 control strain. Table A4.1 tabulates median lifespans for each RIL from these manual lifespans and from the WormBot lifespans. Assay was performed at 20°C and FuDR was used in plates to prevent progeny production. In the first replicate (left), nystatin was also included in plates.

Table A4.3: Comparison of three lifespan measurements of five RILs.

Strain	WormBot Lifespan	Manual Lifespan 1	Manual Lifespan 2
AG11	19.9	17	18
AG3	18	19	20
AO1	22	23	20.5
BO16	28.6	22	16
BO4	16	20	15
SEA51*	20.6, 19.8	18	17.5

*SEA51 was run multiple times on the WormBot. Because strains from WormBot Flights 1 and 2 are represented in the table, both values for SEA51 are shown.

Table A4.4: Genes from GenAge Database located in the ChrII and ChrIV QTL with variants in MY1.

Gene	Chr	Position	Mutation	Effect	Type	Lifespan association
mnk-1	II	10553571	T->C	T18A	missense	RNA interference shortened wild-type (20%) and *ife-2* mutant (15%) maximum lifespan.
mpz-1	II	10740321	G->A	P2305S	missense	Wild type animals treated with *mpz-1* RNAi exhibited a 31% increase in average lifespan compared to untreated wild type animals.
	II	10749016	A->T	S1368T	missense	
	II	10754079	C->A	G1037C	missense	
	II	10757491	A->T	D->E	missense	
raga-1	II	11301875	G->A	A153T	missense	RNAi of *raga-1* in adults results in a 27% lifespan extension.
din-1	II	11618653	G->T	T->N	missense	RNAi had little effect on lifespan, but suppressed *daf-9* life-extension *A variant includes a start_lost for some splice forms
	II	11618807	G->T	L->M	missense	
	II	11624957	T->A	H1458L	missense	
	II	11627041	G->C	P829A	missense	
C47D12.2	II	11677458	G->A	E293K	missense	Post-developmental RNAi of C47D12.2 extends lifespan of wild type worms; amount is indicated as 8.7% of WT lifespan
tars-1	II	11693160	A->T	H329Q	missense	RNAi in adulthood extended mean lifespan by 12%
npp-3	II	11872440	C->A	L648F	missense	RNAi decreased median lifespan by 14% in wild type animals, 36% in a *daf-2* mutant background and 11% in *daf-2*/*daf-16* double mutants.
rsr-2	II	12226193	G->C	E230Q	missense	RNAi decreased median lifespan by 34% in a *daf-2* mutant background and 20% in *daf-2*/*daf-16* double mutants.
sinh-1	II	12229812	T->A	M170K	missense	RNAi extends lifespan.
		12231575	G->T	E305D	missense	

mecr-1	II	13186107	C->T	D336N	missense	RNAi extends lifespan.
mrpl-37	II	13571744	G->T	T338Q	missense	Knockdown of *mrpl-37* increased lifespan by 47%. Has a splice_donor_variant with high heterozygosity, and a frameshift_variant
		13571745	T->G	T338Q	missense	
		13571868	G->C	Q297E	missense	
		13571895	C->G	V288L	missense	
		13571970	C->A	A263S	missense	
		13573592	A->T	C262S	missense	
		13574758	**2N->5N**		**coding exon, complex change**	
daf-5	II	14037704	G->A	S4L	missense	Mutation in adults reduced mean and maximum lifespan by 20%.
jmjd-2	II	14237640	**TCGCCGC->T**	**GGD613D**	**coding exon, deletion**	RNAi of *jmjd-2* increases lifespan up to 13%, depending on temperature.
mes-2	II	14384362	A->T	E129D	missense	RNAi of *mes-2* increases lifespan up to 6%, depending on temperature.
unc-52	II	14649893	G->T	T2797K	missense	RNAi in adulthood extended mean lifespan by 11%
		14657383	C->T	R1825K	missense	
		14657701	A->C	L1768V	missense	
		14658311	C->T	A1604T	missense	
		14663593	A->G	Y657H	missense	

		14664884	A->T	D366E	missense	
hpa-1	IV	16230659	C->T	E468K	missense	30% increase in median lifespan. This gene appears to be in a huge cluster of 21u-rnas. The deletion itself does not appear to be in a u-rna
		16231194	G->A	R322C	missense	
		16232102	C->T	E293K	missense	
		16232300	**C+28 N->C**		**deletion, frame shift**	
		16233050	T->A	I27F	missense	
set-26	IV	16648494	G->T	S1452Y	missense	RNAi of *set-26* results in an increased lifespan up to 22%, depending on temperature. CeNDR suggests this is a stop_gained
		16652041	**T+18 N->T**	**LVLRPH H608H**	**coding exon, deletion**	
ifta-2	IV	17301985	C->T	E17K	missense	Knockouts have an extended lifespan phenotype (40-50%) and are defective in dauer formation due to defects in *daf-2* receptor signaling pathway in ciliated sensory neurons. There was no additive effect of the *ifta-2* mutation on the *daf-2* longevity phenotype.
		17302009	A->G	W9R	missense	

Table A4.5: Genes differentially expressed in the 417-rDNA NIL (*catIR12*) as compared to N2 with padj<0.05.

Public Name	Gene ID	Base Mean	Log2 FoldChange	Lfc SE	pvalue	padj
hen-1	WBGene00001841	329.09	-11.36	1.21	4.64E-21	3.84E-17
C36B7.6	WBGene00016469	732.95	-11.34	1.19	1.27E-21	2.11E-17
marc-6	WBGene00018847	1129.87	-2.47	0.26	8.52E-21	4.70E-17
F42A9.6	WBGene00018335	2903.13	-1.06	0.25	1.49E-05	0.0205
Y43C5A.7	WBGene00044922	495.16	0.67	0.16	4.11E-05	0.0378
Y51H4A.7	WBGene00013103	414.38	0.69	0.14	7.11E-07	0.0013
F33E2.5	WBGene00009361	291.16	0.76	0.18	1.94E-05	0.0214
D2023.1	WBGene00014300	523.49	0.78	0.18	1.80E-05	0.0214
svop-1	WBGene00014021	105.82	0.95	0.22	2.07E-05	0.0214
nlp-20	WBGene00003758	639.16	0.96	0.22	1.89E-05	0.0214
Y102A5C.36	WBGene00044213	175.31	1.09	0.20	2.63E-08	6.20E-05
ins-30	WBGene00002113	59.64	1.09	0.27	3.96E-05	0.0378
ZK84.1	WBGene00022649	65.22	1.10	0.25	8.32E-06	0.0125
Y105E8B.9	WBGene00013693	176.23	1.49	0.24	5.07E-10	1.68E-06
fat-7	WBGene00001399	3118.54	2.32	0.34	6.84E-12	2.83E-08
C55C3.3	WBGene00016953	24.12	2.32	0.44	1.07E-07	0.0002
dyf-3	WBGene00001119	29.01	2.79	0.49	1.04E-08	2.87E-05
srx-65	WBGene00005956	6.96	6.20	1.32	2.73E-06	0.0045

Table A4.6: Genes differentially expressed in the 420-rDNA NIL (*catIR29*) as compared to N2 with padj<0.05.

Public Name	Gene ID	Base Mean	Log2 FoldChange	lfc SE	pvalue	padj
fbxa-191	WBGene00010209	54.99	-1.41	0.29	1.05E-06	0.0058
fbxa-192	WBGene00010212	406.02	-0.93	0.18	3.26E-07	0.0054
cyp-13A5	WBGene00011672	1001.70	0.72	0.17	1.88E-05	0.0317
F33E2.5	WBGene00009361	291.16	0.75	0.18	2.11E-05	0.0317
cnc-4	WBGene00000558	300.76	0.88	0.18	6.88E-07	0.0057
T19C9.8	WBGene00011844	657.88	0.97	0.23	2.41E-05	0.0332
C25F9.11	WBGene00045411	135.36	1.22	0.28	1.18E-05	0.0244
dao-2	WBGene00000928	777.34	1.38	0.30	3.47E-06	0.0115
dpy-17	WBGene00001076	451.89	1.56	0.37	1.98E-05	0.0317
dpy-2	WBGene00001064	108.43	1.91	0.43	9.05E-06	0.0214
C55C3.3	WBGene00016953	24.12	1.98	0.44	6.38E-06	0.0176
srx-65	WBGene00005956	6.96	6.34	1.32	1.54E-06	0.0064

Table A4.7: Genes differentially expressed in the 73-rDNA NIL (*catIR28*) as compared to N2 with padj<0.05.

Public Name	Gene ID	Base Mean	Log2 FoldChange	lfc SE	pvalue	padj
T03D3.5	WBGene00020183	288.97	-1.10	0.18	1.27E-09	2.42E-06
nlp-29	WBGene00003767	558.69	0.67	0.18	2.48E-04	3.50E-02
C46F11.6	WBGene00008121	247.08	0.69	0.16	1.49E-05	4.11E-03
cnc-4	WBGene00000558	300.76	0.73	0.18	3.99E-05	8.34E-03
fbxc-51	WBGene00021576	391.04	0.77	0.20	1.64E-04	2.58E-02
cki-1	WBGene00000516	197.57	0.78	0.20	1.34E-04	2.17E-02
F33E2.5	WBGene00009361	291.16	0.78	0.18	9.99E-06	3.05E-03
hil-3	WBGene00001854	579.32	0.82	0.23	3.80E-04	4.86E-02
clec-266	WBGene00016088	1196.14	0.86	0.22	7.07E-05	1.24E-02
Y7A9D.1	WBGene00012420	123.85	0.87	0.24	2.85E-04	3.91E-02
tbb-6	WBGene00006539	297.43	0.90	0.22	6.02E-05	1.10E-02
Y41G9A.10	WBGene00044901	429.87	0.90	0.17	1.62E-07	9.56E-05
his-24	WBGene00001898	3067.02	0.93	0.22	2.13E-05	5.28E-03
C26F1.1	WBGene00016146	157.24	0.97	0.21	2.38E-06	8.50E-04
C06E7.4	WBGene00015541	84.76	1.00	0.24	3.16E-05	6.86E-03
zip-7	WBGene00013100	87.94	1.06	0.26	3.46E-05	7.32E-03
K10D6.2	WBGene00010742	269.05	1.08	0.28	1.05E-04	1.73E-02
cls-3	WBGene00013847	322.63	1.09	0.21	2.47E-07	1.32E-04
ncam-1	WBGene00017184	165.87	1.10	0.28	8.98E-05	1.52E-02
C08F1.10	WBGene00015613	190.93	1.12	0.24	1.84E-06	7.15E-04
T20D4.7	WBGene00020613	139.22	1.13	0.28	7.52E-05	1.30E-02

F53B3.5	WBGene00018743	418.87	1.14	0.28	5.15E-05	1.01E-02
Y38H6C.16	WBGene00012628	85.39	1.14	0.27	2.51E-05	5.75E-03
Y17D7C.4	WBGene00050903	67.27	1.22	0.30	4.72E-05	9.39E-03
col-104	WBGene00000678	35.86	1.23	0.33	2.26E-04	3.26E-02
C16D9.5	WBGene00015860	142.09	1.28	0.33	8.69E-05	1.49E-02
nsy-4	WBGene00021415	104.71	1.31	0.28	3.94E-06	1.30E-03
C05A9.2	WBGene00007317	98.70	1.33	0.32	3.35E-05	7.16E-03
tsp-1	WBGene00006627	164.52	1.36	0.37	2.47E-04	3.50E-02
msp-64	WBGene00003457	43.97	1.38	0.36	1.02E-04	1.71E-02
Y41G9A.5	WBGene00021529	35.74	1.41	0.33	2.63E-05	5.92E-03
B0393.5	WBGene00007170	44.81	1.42	0.40	3.38E-04	4.42E-02
che-14	WBGene00000493	30.03	1.42	0.39	2.88E-04	3.91E-02
dex-1	WBGene00017028	52.08	1.46	0.36	5.74E-05	1.07E-02
fkb-4	WBGene00001429	33.88	1.47	0.35	2.52E-05	5.75E-03
dsl-3	WBGene00001105	87.50	1.54	0.32	2.06E-06	7.67E-04
zag-1	WBGene00006970	22.88	1.55	0.42	2.06E-04	3.14E-02
fbn-1	WBGene00022816	108.43	1.57	0.39	4.60E-05	9.26E-03
cpg-24	WBGene00021525	62.28	1.58	0.36	1.28E-05	3.65E-03
C26B9.3	WBGene00016133	63.37	1.59	0.40	6.23E-05	1.11E-02
fmi-1	WBGene00001475	61.65	1.60	0.40	5.49E-05	1.04E-02
spp-13	WBGene00004998	30.42	1.61	0.40	6.65E-05	1.17E-02
cuti-1	WBGene00022591	72.20	1.63	0.32	2.82E-07	1.34E-04
hbl-1	WBGene00001824	199.19	1.66	0.41	5.43E-05	1.04E-02

dpy-10	WBGene00001072	204.37	1.66	0.31	1.34E-07	8.47E-05
M03D4.4	WBGene00019751	58.32	1.67	0.38	1.19E-05	3.46E-03
F13E9.11	WBGene00008760	93.09	1.68	0.40	2.83E-05	6.29E-03
sym-1	WBGene00006366	196.73	1.79	0.42	1.82E-05	4.66E-03
dao-2	WBGene00000928	777.34	1.86	0.30	4.06E-10	1.16E-06
dyf-3	WBGene00001119	29.01	1.88	0.50	1.59E-04	2.52E-02
let-4	WBGene00002282	79.13	1.91	0.35	4.77E-08	3.89E-05
	WBGene00018737	49.93	1.91	0.52	2.15E-04	3.16E-02
F43D9.1	WBGene00009653	58.94	1.93	0.45	2.26E-05	5.45E-03
mlt-8	WBGene00021095	129.60	1.93	0.36	1.02E-07	7.57E-05
noah-1	WBGene00016422	337.22	1.93	0.39	5.93E-07	2.63E-04
T02E9.5	WBGene00011383	479.66	1.95	0.38	2.58E-07	1.33E-04
E04D5.4	WBGene00008483	52.36	1.99	0.39	2.72E-07	1.33E-04
F58H1.2	WBGene00010285	37.81	1.99	0.51	1.05E-04	1.73E-02
K10D3.4	WBGene00010738	47.19	1.99	0.49	5.30E-05	1.03E-02
wrt-10	WBGene00006956	196.74	2.01	0.37	4.27E-08	3.66E-05
F26F12.4	WBGene00017835	17.50	2.04	0.55	2.13E-04	3.16E-02
Y110A2AL.4	WBGene00022441	130.88	2.05	0.39	1.09E-07	7.81E-05
C13C12.2	WBGene00007552	12.40	2.05	0.57	3.24E-04	4.34E-02
ces-2	WBGene00000469	157.73	2.05	0.36	1.76E-08	1.68E-05
R03H10.2	WBGene00019855	40.91	2.05	0.48	1.92E-05	4.83E-03
F44E2.4	WBGene00018418	45.09	2.06	0.48	1.65E-05	4.34E-03
grl-15	WBGene00001724	36.15	2.08	0.47	1.18E-05	3.46E-03

C02E7.7	WBGene00015340	33.41	2.09	0.48	1.18E-05	3.46E-03
clec-78	WBGene00018547	21.31	2.09	0.59	3.80E-04	4.86E-02
spp-20	WBGene00005005	31.33	2.11	0.47	6.02E-06	1.93E-03
atf-2	WBGene00000220	34.25	2.12	0.49	1.82E-05	4.66E-03
noah-2	WBGene00009926	206.32	2.12	0.36	5.30E-09	6.48E-06
cutl-2	WBGene00013145	79.77	2.16	0.35	4.84E-10	1.18E-06
dpy-17	WBGene00001076	451.89	2.20	0.36	1.58E-09	2.46E-06
fbxa-163	WBGene00015598	13.19	2.21	0.60	2.14E-04	3.16E-02
col-121	WBGene00000695	61.34	2.24	0.45	7.70E-07	3.22E-04
tts-2	WBGene00006651	94.31	2.29	0.49	3.43E-06	1.15E-03
egl-46	WBGene00001210	30.00	2.30	0.44	1.92E-07	1.09E-04
fipr-24	WBGene00007992	25.58	2.32	0.54	1.65E-05	4.34E-03
F15B9.8	WBGene00008851	34.45	2.34	0.58	5.62E-05	1.06E-02
rml-3	WBGene00015781	85.79	2.37	0.40	2.86E-09	4.08E-06
R09E10.5	WBGene00011175	11.38	2.38	0.66	3.28E-04	4.35E-02
nhr-127	WBGene00003717	12.47	2.39	0.65	2.61E-04	3.66E-02
M03B6.3	WBGene00010835	25.00	2.40	0.48	7.56E-07	3.22E-04
atf-8	WBGene00017535	26.72	2.49	0.66	1.68E-04	2.61E-02
ifa-3	WBGene00002051	21.25	2.53	0.60	2.16E-05	5.28E-03
ZK180.5	WBGene00022679	35.68	2.55	0.59	1.41E-05	3.96E-03
F11E6.9	WBGene00008712	15.70	2.57	0.61	2.47E-05	5.75E-03
T20F5.4	WBGene00020626	40.75	2.58	0.47	4.18E-08	3.66E-05
K02E10.4	WBGene00019318	34.61	2.59	0.53	1.15E-06	4.59E-04

dpy-14	WBGene00001075	583.96	2.59	0.41	1.91E-10	1.09E-06
K04H4.2	WBGene00010573	14.04	2.59	0.64	4.60E-05	9.26E-03
T04C12.1	WBGene00011427	31.48	2.61	0.62	2.36E-05	5.62E-03
dpy-7	WBGene00001069	73.75	2.61	0.42	7.71E-10	1.65E-06
lpr-5	WBGene00012256	36.63	2.62	0.61	1.57E-05	4.28E-03
ztf-30	WBGene00015523	12.17	2.65	0.66	5.82E-05	1.07E-02
dsl-6	WBGene00001108	32.81	2.66	0.56	2.17E-06	7.91E-04
T05H10.3	WBGene00011508	118.61	2.68	0.42	1.34E-10	1.09E-06
hch-1	WBGene00001828	127.13	2.68	0.44	1.49E-09	2.46E-06
dpy-3	WBGene00001065	53.72	2.71	0.47	8.78E-09	9.40E-06
C06E7.88	WBGene00189995	10.85	2.71	0.70	1.14E-04	1.87E-02
ZC123.1	WBGene00022517	12.06	2.75	0.77	3.41E-04	4.42E-02
Y41D4B.6	WBGene00021513	31.19	2.78	0.58	1.95E-06	7.41E-04
dpy-2	WBGene00001064	108.43	2.79	0.43	6.88E-11	1.09E-06
R02F11.1	WBGene00019839	16.55	2.81	0.76	2.16E-04	3.16E-02
sqt-3	WBGene00005018	829.59	2.81	0.55	2.72E-07	1.33E-04
wrt-1	WBGene00006947	51.03	2.85	0.58	1.00E-06	4.08E-04
Y37A1B.7	WBGene00012540	25.75	2.92	0.70	2.87E-05	6.29E-03
lpr-3	WBGene00012261	84.21	2.93	0.46	2.85E-10	1.16E-06
hil-7	WBGene00001858	200.96	2.94	0.59	5.98E-07	2.63E-04
K08B12.1	WBGene00019520	22.57	2.99	0.75	6.21E-05	1.11E-02
mlt-11	WBGene00012186	63.17	3.00	0.52	7.68E-09	8.76E-06
C35A5.11	WBGene00044024	44.81	3.06	0.58	1.60E-07	9.56E-05

C02E7.6	WBGene00015339	103.62	3.13	0.67	2.70E-06	9.43E-04
R12E2.14	WBGene00020039	8.28	3.19	0.86	2.18E-04	3.17E-02
cut-3	WBGene00009041	44.54	3.20	0.85	1.69E-04	2.61E-02
H42K12.3	WBGene00019272	36.89	3.21	0.59	5.33E-08	4.15E-05
ptr-4	WBGene00004219	33.09	3.22	0.57	1.36E-08	1.37E-05
glf-1	WBGene00019154	34.11	3.24	0.64	3.64E-07	1.68E-04
fipr-4	WBGene00007541	8.86	3.25	0.88	2.13E-04	3.16E-02
cut-2	WBGene00009983	151.94	3.49	0.79	8.65E-06	2.69E-03
col-74	WBGene00000650	13.47	3.56	0.79	6.10E-06	1.93E-03
C14A4.9	WBGene00007560	23.93	3.59	0.77	3.26E-06	1.12E-03
grl-7	WBGene00001716	26.37	3.60	0.68	1.16E-07	7.93E-05
lpr-6	WBGene00012255	38.55	3.68	0.62	3.77E-09	4.96E-06
col-165	WBGene00000738	35.78	3.73	0.72	2.38E-07	1.31E-04
T03G6.1	WBGene00020194	30.73	3.88	0.73	1.26E-07	8.32E-05
lpr-4	WBGene00012257	24.99	4.03	0.64	3.68E-10	1.16E-06
F18C5.5	WBGene00017560	6.65	4.24	1.18	3.40E-04	4.42E-02
grl-5	WBGene00001714	5.35	4.58	1.26	2.84E-04	3.91E-02
ilys-3	WBGene00016670	5.75	4.59	1.27	2.92E-04	3.93E-02
T19C3.2	WBGene00020560	3.54	5.31	1.46	2.88E-04	3.91E-02
Y26D4A.24	WBGene00200760	4.01	5.42	1.44	1.58E-04	2.52E-02
srx-65	WBGene00005956	6.96	5.47	1.34	4.38E-05	9.05E-03

Table A4.8: Genes differentially expressed in the 81-rDNA NIL (*catIR30*) as compared to N2 with padj<0.05.

Public Name	Gene ID	Base Mean	Log2 FoldChange	lfc SE	pvalue	padj
T03D3.5	WBGene00020183	288.97	-1.20	0.18	7.80E-11	1.11E-06
nlp-29	WBGene00003767	558.69	0.74	0.18	5.37E-05	4.01E-02
cyp-13A5	WBGene00011672	1001.70	0.76	0.17	6.93E-06	1.23E-02
cnc-4	WBGene00000558	300.76	0.97	0.18	4.89E-08	3.46E-04
T19C9.8	WBGene00011844	657.88	1.02	0.23	8.96E-06	1.28E-02
F33H12.7	WBGene00045457	558.97	1.21	0.27	5.11E-06	1.03E-02
pho-9	WBGene00022181	902.81	1.23	0.26	2.48E-06	5.85E-03
clec-169	WBGene00009526	118.68	1.26	0.28	9.03E-06	1.28E-02
ZC204.14	WBGene00022564	263.13	1.26	0.24	1.23E-07	5.83E-04
dao-2	WBGene00000928	777.34	1.30	0.30	1.42E-05	1.83E-02
dpy-17	WBGene00001076	451.89	1.49	0.37	4.51E-05	3.76E-02
tsp-1	WBGene00006627	164.52	1.58	0.37	1.86E-05	2.03E-02
rml-3	WBGene00015781	85.79	1.65	0.41	4.78E-05	3.76E-02
dpy-7	WBGene00001069	73.75	1.80	0.43	2.88E-05	2.72E-02
T20F5.4	WBGene00020626	40.75	1.95	0.48	4.76E-05	3.76E-02
mlt-11	WBGene00012186	63.17	2.11	0.53	5.71E-05	4.04E-02
dpy-2	WBGene00001064	108.43	2.14	0.43	7.22E-07	2.56E-03
C35A5.11	WBGene00044024	44.81	2.53	0.59	1.78E-05	2.03E-02
lpr-4	WBGene00012257	24.99	2.81	0.66	2.03E-05	2.05E-02
lpr-6	WBGene00012255	38.55	3.07	0.63	1.10E-06	3.10E-03

Table A4.9: Gene Ontology Enrichment Analysis of genes differentially expressed in the 73-rDNA (allele *catIR28*) NIL as compared to N2.

Term	Expected	Observed	Enrichment Fold Change	P value	Q value
molting cycle GO:0042303	0.47	11	24	6.50E-14	2.00E-11
structural constituent of cuticle GO:0042302	0.74	12	16	9.20E-13	1.40E-10
cuticle development GO:0042335	0.2	6	30	1.30E-09	1.30E-07
collagen trimer GO:0005581	0.24	5	21	1.60E-07	1.20E-05
extracellular space GO:0005615	1.1	6	5.4	0.00013	0.0083
negative regulation of transcription by RNA polymerase II GO:0000122	0.65	4	6.1	0.00054	0.027
DNA-binding transcription factor activity GO:0003700	2.7	8	2.9	0.002	0.086

Table A4.10: Phenotype Enrichment Analysis of genes differentially expressed in the 73-rDNA (allele *catIR28*) NIL as compared to N2.

Term	Expected	Observed	Enrichment Fold Change	P value	Q value
paralyzed WBPhenotype:0000644	1.2	9	7.7	3.40E-07	8.00E-05
dumpy WBPhenotype:0000583	2.5	13	5.1	3.70E-07	8.00E-05
pericellular component development variant WBPhenotype:0000200	0.68	7	10	4.90E-07	8.00E-05
molt variant WBPhenotype:0002041	1.8	9	5	1.60E-05	0.00096
movement variant WBPhenotype:0001206	13	26	2	0.00026	0.012
age associated fluorescence increased WBPhenotype:0000467	0.57	4	7	0.00029	0.012
breaks in alae WBPhenotype:0000280	0.32	3	9.3	0.00031	0.012
stress induced lethality variant WBPhenotype:0000139	0.42	3	7.2	0.00083	0.025
intestinal vacuole WBPhenotype:0001428	0.42	3	7.2	0.00083	0.025
protein degradation variant WBPhenotype:0001645	1.2	5	4.3	0.0012	0.028
male mating efficiency reduced WBPhenotype:0000843	0.46	3	6.5	0.0012	0.028
body morphology variant WBPhenotype:0000072	8.9	18	2	0.0018	0.037
body region phenotype WBPhenotype:0002557	9.9	19	1.9	0.0027	0.05
antihelmintic response variant WBPhenotype:0001852	0.32	2	6.2	0.0042	0.071
organism hypertonic lethality increased WBPhenotype:0001751	0.33	2	6.1	0.0044	0.071

Table A4.11: Tissue Enrichment Analysis of genes differentially expressed in the 73-rDNA (allele *catIR28*) NIL as compared to N2.

Term	Expected	Observed	Enrichment Fold Change	P value	Q value
epithelial system WBbt:0005730	28	72	2.6	7.70E-14	2.30E-11
excretory duct cell WBbt:0004540	0.25	5	20	2.00E-07	3.00E-05
outer labial sensillum WBbt:0005501	16	38	2.4	3.40E-07	3.40E-05
PVD WBbt:0006831	15	37	2.4	5.80E-07	4.30E-05
hyp6 WBbt:0004679	0.31	4	13	1.70E-05	0.001
touch receptor neuron WBbt:0005237	11	22	2.1	0.00051	0.026
HSN WBbt:0006830	0.69	4	5.8	0.00071	0.03
P5 WBbt:0006774	0.18	2	11	0.00073	0.03
P8 WBbt:0006777	0.18	2	11	0.00073	0.03
P1 WBbt:0006770	0.18	2	11	0.00073	0.03
P2 WBbt:0006771	0.18	2	11	0.0008	0.03
P6 WBbt:0006775	0.18	2	11	0.0008	0.03
P9 WBbt:0006778	0.19	2	11	0.00087	0.03
P7 WBbt:0006776	0.19	2	11	0.00087	0.03
P4 WBbt:0006773	0.19	2	11	0.00087	0.03
P10 WBbt:0006779	0.21	2	9.4	0.0013	0.03
P12 WBbt:0004409	0.26	2	7.8	0.0022	0.039
hyp5 WBbt:0004685	0.28	2	7.2	0.0028	0.046
hyp4 WBbt:0004687	0.29	2	6.8	0.0032	0.051
P11 WBbt:0004410	0.31	2	6.5	0.0038	0.056
male WBbt:0007850	14	24	1.7	0.0046	0.066
PVQ WBbt:0006976	0.34	2	5.9	0.005	0.068

For the P1-P12 tissue enrichments that show up in the TEA, the two genes that underlie the enrichment are the same for each: *hbl-1* and *dpy-7.* These P1-P12 terms are a set of twelve postembryonic blast cells.

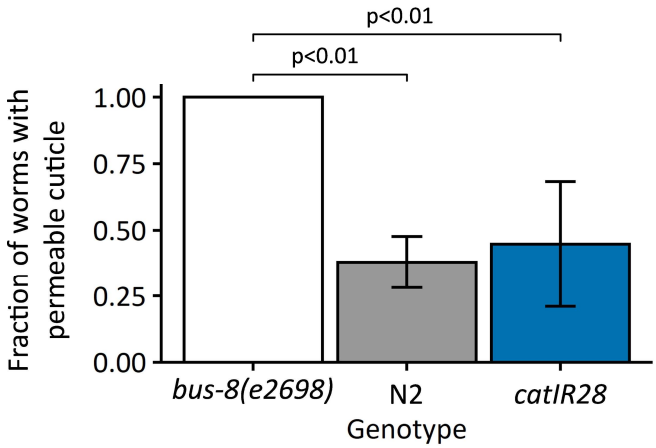

Figure A4.7: Cuticle permeability does not differ between N2 and the 73-rDNA NIL.
Worms were stained with Hoechst, which does not penetrate intact worm cuticles. The number of worms with stained hypodermal nuclei were counted as worms with permeable cuticles. At least 13 worms were quantified for each replicate, three replicates per genotype. The *bus-8*(*e2698*) genotype is a positive control with known reduced cuticle integrity [420]; in all three replicates all worms had permeable cuticles. Both N2 and the 73-rDNA NIL (*catIR28*) have fewer worms with permeable cuticles, and do not differ from one another in cuticle permeability, as measured by ANOVA and Tukey's HSD.

Table A4.12: Strains used in this study.

Strain	Genotype	Source
VC2010	N2 (WT)	Waterston lab
MY1	Wild isolate MY1	Moerman lab
JU775	Wild isolate JU775	Moerman lab
MY16	Wild isolate MY16	Moerman lab
RC301	Wild isolate RC301	Teotonio lab
CB1370	*daf-2*(*e1370*) III	CGC
PS3551	*hsf-1*(*sy441*) I	CGC
SEA51	*mIs13*[myo-2p::GFP + pes-10p::GFP + F22B7.9p::GFP] I	This study
SEA295	*catIR7* [MY1 *mIs13* 64-rDNA; chrI:14170793-end N2]	This study
SEA296	*catIR8* [MY1 *mIs13* 130-rDNA; chrI:14170793-end N2]	This study
SEA300	*catIR12* [MY1 chrI:13527418-end]	This study
SEA302	*catIR14* [JU775 I:13529383-end]	This study
SEA304	*catIR16* [RC301 I:11737213-end]	This study
SEA305	*catIR17* [MY16 I:3576802-end]	This study
SEA328	*catIR28* I [MY16 I:14764185-end]	This study
SEA329	*catIR29* I [RC301 I:14989978-end]	This study
SEA330	*catIR30* I [JU775 I:14775350-end]	This study
RB1562	D1086.4(*ok1896*) V	CGC
SEA333	D1086.4(*ok1896*) V	This study
SEA340	*catIR12 him-5*(*ok1896*)	This study
SEA344	*catIR28 him-5*(*ok1896*)	This study
SEA345	*catIR29 him-5*(*ok1896*)	This study
SEA346	*catIR30 him-5*(*ok1896*)	This study

*RIL strains are listed in a separate table: See Supplemental File.

Table A4.13: Primers used in this study.

Primer name	Primer sequence	Description	Source
AHC1	GGTCAAAATCGAAAGTGGAAAC	ChrI 14.5 Mb F	This study
AHC2	GTGAAAGTGTAGAAAAAGTCTTTAGAAATAG	ChrI 14.5 Mb R	This study
AHC14	GACAAAACATGATCAAATGAGAAC	ChrI 12.5 Mb R	This study
AHC24	GCACCATTTGGAGCTATG	ChrI 12.5 Mb F	This study
AHC32	CCATTTTTGCACCATTGGAG	ChrI 14.99 Mb F	This study
AHC35	CAAGCGCCTTTATTGAAAAGC	ChrI 14.99 Mb R	This study
AHC36	GTACCTATTTCTTCCCACCTG	ChrI 14.68 Mb F	This study
AHC55	ACCGACGGTAACTAAGATTT	ChrI 14.68 Mb R1	This study
AHC56	CGACGGTAACTAAGATCC	ChrI 14.68 Mb R2	This study
AHC66	GTGTCCCATCTCACGATTAG	Probe I F	[327]
AHC67	GTGATATCTGCTCTAATGAG	Probe I R	[327]
AHC68	AACGACTTCGTTGTTGCGG	Probe II F	[327]
AHC69	TTCGACACTCAACTGACCG	Probe II R	[327]
AHC70	TCAACGTTCCAGTTGAGATG	Probe III F	[327]
AHC71	CGATCATCAAGACTATCGTC	Probe III R	[327]
AHC72	TGGCTATATGCGTCTAGGC	Probe IV F	[327]
AHC73	ATCACCGCATGTCCGTGAAG	Probe IV R	[327]
AHC74	CTTCACGGACATGCGGTGAT	Probe V F	[327]
AHC75	AGTTGGTGCTATGCGTTCG	Probe V R	[327]
AHC76	CGAACGCATAGCACCAACT	Probe VI F	[327]
AHC77	TGTGATGCTTCTGGACTAGG	Probe VI R	[327]
AHC78	TCGAATACTGGGATTCGTC	Probe VII F	[327]
AHC79	AGCAGCCAAAGACTGATCG	Probe VII R	[327]
AHC84	TCGGTATGGGACAGAAGGAC	Actin F	[421]
AHC85	CATCCCAGTTGGTGACGATA	Actin R	[421]
AHC88	GCTTACGACCATATCACGTTGAATG	5S_qPCR_F	This study
AHC89	CTTACAACATCCAGGATTCCCAG	5S_qPCR_R	This study

A4.2 SUPPLEMENTAL EXPERIMENTS:

Meiotic nondisjunction frequency is not changed in worms with higher or lower rDNA copy numbers.

Having sufficient rDNA copies is important for maintaining proper genome segregation. In *S. cerevisiae*, transcription levels of the rDNA control levels of condensin binding such that more condensin binds to the rDNA when fewer repeats are transcribed [422]. When there are too few rDNA copies, condensin binds poorly to the rDNA and sister chromatids separate prematurely during the cell cycle [423]. Chromosome segregation defects in *S. cerevisiae* strains with reduced rDNA copy number affect both the rDNA-containing chromosome XII as well as non-rDNA-containing chromosomes [424]. To test for chromosome missegregation in *C. elegans*, the frequency of male progeny is typically assessed. Males arise by nondisjunction of the X chromosome: Hermaphrodites have two copies of the X chromosome while males have only one. In wild type worms, the frequency of male progeny is approximately 1 in 1000. Heat shock induces a higher rate of nondisjunction, and therefore a higher frequency of male progeny [425]. In my hands, the consistency of heat-shock-induced male progeny was too variable to be used for strain comparison assays.

An alternative to inducing higher levels of male progeny by heat shock is to assess frequency of nondisjunction in a sensitized strain background. Various nondisjunction mutants with a high incidence of male progeny exist in *C. elegans* [426]. One mutant, *him-5*, also has a high incidence of embryonic lethality, caused by autosomal nondisjunction. The *him-5* gene is involved in meiotic break positioning and is a paralog of *rec-1* [427]. I introduced the null allele *him-5*(*ok1896*) into the N2 NIL backgrounds and measured the incidence of each type of

nondisjunction progeny this allele produces: Dead embryos (presumed autosomal nondisjunction), Male worms (OX), Dumpy worms (XXX), and those characterized as having "other defects" (typically slow growing). If there was an increase in nondisjunction of the X chromosome but not of the autosomes, I would expect a higher male frequency. If changing rDNA copy number increases the nondisjunction of the rDNA-containing chromosome I, I would expect a higher frequency of dead embryos as compared to wild type. There are no replicable, statistically significant differences between any NILs and wild type carrying *him-5*(*ok1896*) (**Figure A4.7**). Therefore, neither increasing nor decreasing rDNA copy number changes the frequency of meiotic nondisjunction events.

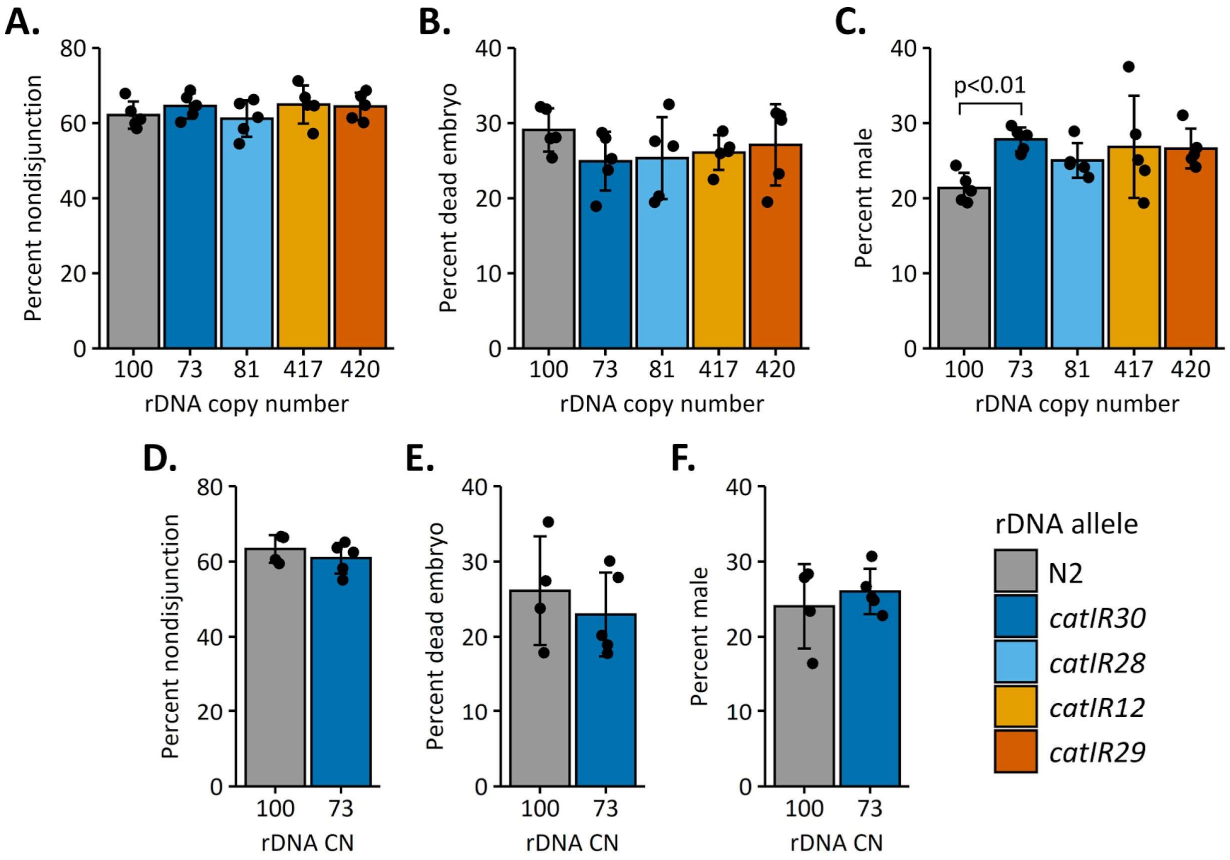

Figure A4.8: Penetrance of mutant progeny in NILs carrying *him-5*(*ok1986*) allele.
A: Percent of progeny with any nondisjunction event (any defect) in NILs with *him-5*(*ok1986*). No significant differences between any strains by t-test and Bonferroni correction. **B:** Percent of progeny that fail to hatch (dead embryos) in NILs with *him-5*(*ok1986*). Dead embryos are assumed to arise from autosomal nondisjunction events. No significant differences between any strains by t-test and Bonferroni correction. **C:** Percent of progeny that are male (X-chromosome nondisjunction) in NILs with *him-5*(*ok1986*). The 73-rDNA NIL and wild type significantly differ in the percent of progeny produced that are male, $p<0.01$ by t-test and Bonferroni correction. Data from A-C come from the same experiments but are plotted to highlight different partially penetrant phenotypes. **D-F:** As **A-C**, but a second replicate with only N2 and the 73-rDNA strain to confirm whether there is a difference between the strains in male frequency. No significant differences are present between these strains for male frequency, percent of worms that fail to hatch, or total worms that had a nondisjunction event (student's T-test).

A4.3 Supplemental Methods

WormBot lifespan data curation

For each WormBot well, I hand-annotated whether any of the following problems were present in the well: Contamination, suspicion of progeny (>30 worms present in the well), or desiccation (early vs late in the experiment). The final median lifespans were calculated from wells that had no progeny or desiccation. Some strains were run on the WormBot in two different Flights. In these cases, I only used one Flight to calculate median lifespan. The Flight selected is the one which contained at least two wells that agreed in lifespan (+/- 2 days when calculating median lifespan for each well independently). If both Flights met the criteria, then the Flight with a higher n was selected. To calculate the difference in median lifespan as compared to the SEA51 control, the average SEA51 median lifespan was subtracted from the average RIL median lifespan from the same WormBot Flight.

Male frequency assay

Worms carrying the *him-5* allele were synchronized by pulse-laying day 1 adults. Two days later, 5 healthy L4 individuals per strain were singled to 3cm NGM plates seeded with OP50. Adult worms were transferred to new 3cm NGM + OP50 plates every 24 hours for 3 days (3 total transfers). For each plate, 24 hours after the adult was removed, plates were scored for the presence of unhatched (dead) embryos. Then, 48-72 hours after the adult was removed, hatched progeny were counted and scored as: Healthy hermaphrodites, male, developmentally delayed (too young to discriminate between males and hermaphrodites), or dumpy hermaphrodites. Males, developmentally delayed worms, dumpy hermaphrodites, and dead embryos are all considered to have arisen from nondisjunction events.

APPENDIX 5: SHORT SUPPLEMENTAL STUDIES

A5.1: ABSENCE OF EVIDENCE FOR INVERSE CORRELATION BETWEEN rDNA COPY NUMBER AND MITOCHONDRIAL DNA ABUNDANCE IN HUMANS.

In humans, rDNA copy number and mitochondrial DNA abundance were reported to inversely correlate [2]. In my follow-up study on 1000 Genomes Project data presented in Chapter 3, I showed that conclusions previously drawn from these datasets are not supported by newer, higher-quality data. As new data show an absence of evidence for meaningful co-variation between the 5S and 45S rDNA arrays, I suspected that the previously observed negative correlation between the 45S rDNA copy number and mtDNA abundance would also not be maintained when analyzing the higher coverage dataset. I found that mtDNA abundance and rDNA copy number do not meaningfully co-vary in the newer 1000 Genomes Project dataset (**Figure A5.1**). When using the lower coverage 1000 Genomes dataset, I recapitulate a weak negative correlation between rDNA copy number and mtDNA abundance (Pearson's correlation=-0.142, p=0.0702), though this is weaker than the initially reported inverse correlation (Pearson's correlation=-0.29, p=0.00013) [2]. In the high coverage 1000 Genomes Project dataset, the correlation between rDNA copy number and mtDNA abundance is still weak, but it is a positive correlation.

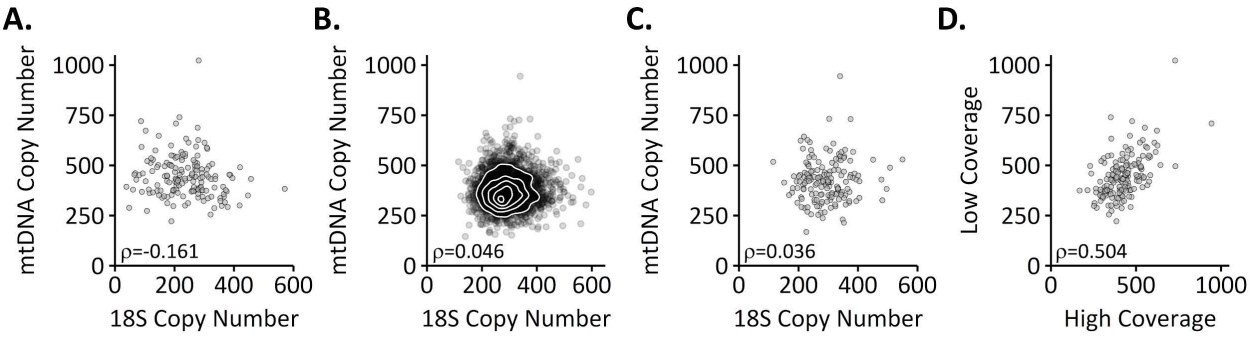

Figure A5.1: Mitochondrial DNA estimates and correlations to rDNA copy numbers from the 1000 Genomes Project.

A: Correlation of 18S rDNA copy number to mtDNA abundance in the low coverage 1000 genomes project, n=163. **B:** Correlation of 18S rDNA copy number to mtDNA abundance in the high coverage 1000 Genomes Project, n=2,419. **C:** Correlation of 18S rDNA copy number to mtDNA abundance in the high coverage 1000 Genomes Project, in the subset of samples analyzed from the low coverage 1000 Genomes Project dataset, n=163. **D:** Correlation of mtDNA copy number estimates in the 163 samples measured in both the low- and high-coverage 1000 Genomes Project datasets. Spearman correlations are indicated in each panel.

A5.2: EXTRACHROMOSOMAL rDNA CIRCLES IN *C. ELEGANS*

Extrachromosomal rDNA circles (ERCs) accumulate with age in *S. cerevisiae* and were long thought to be a cause of aging [196]. Extrachromosomal circular DNA of many types have been found in multicellular organisms including humans, flies, and worms, some of which does contain rDNA sequence [198,199,428]. In *D. melanogaster*, ERCs do not accumulate with age, but levels differ at different developmental stages [198]. I wondered whether ERCs accumulate with age in *C. elegans* and if they are present during the larval stages. I measured ERC levels in worms to determine if ERC accumulation is likely yeast-specific or if *D. melanogaster* may be an outlier and that ERC accumulation can occur in multicellular organisms with age.

To identify ERCs in developing worms, I grew synchronized worm populations and harvested samples at each larval stage (starved L1s, then fed (plated) L1s, L2, L3, L4, and Gravid Adults). I identified ERCs by 2D gel electrophoresis and rDNA-specific Southern blotting. ERCs are barely detectable in any developmental stage (**Figure A5.2**), but they are more prevalent in L2-adulthood than they are in L1 worms. There are two possible explanations for a lower abundance of ERCs in L1 worms. The first is that ERCs are made during germline production, similar to rDNA amplification in *Xenopus* species (though to a much lesser extent) [36]. In this scenario, ERCs would start to be formed when germline proliferation begins. The second is that ERC formation is suppressed during starvation, which is used to synchronize the worms.

Next, I asked whether ERCs accumulate in aging adult worms. I grew synchronized N2 worms and aged populations to days 1, 5, 10, and 14 of adulthood and performed 2D gel electrophoresis and Southern blotting to detect circular rDNA molecules. ERCs are present in day 1 adults, and their abundance diminishes over time: in day 5 of adulthood, they remain

detectable, but in days 10 and 14 are either barely visible or completely absent (**Figure A5.2 B** and **C**). In addition, the ethidium staining of the first dimension for the aging series showed the day 1 adult had less genomic DNA loaded overall (despite equal quantities being loaded as assessed by nanodrop). This indicates that the ERC abundance in day 1 adult animals is likely even greater than shown in Figure A5.2 as compared to the other samples. It is clear, however, that ERCs do not accumulate during aging. My interpretation of the absence of ERC accumulation in aging adult worms is as follows: In yeast, ERCs accumulate with replicative age, and each ERC has an origin of replication. Not only do spontaneous ERCs form during aging, but each ERC replicates with each cell division and ERCs are preferentially sequestered in the aging mother cell [196]. In *C. elegans*, aging worms are post-mitotic; the only actively dividing cells in an adult worm are in the germline. The fact that ERCs are only detectable in worms in which there is an actively replicating germline fits the hypothesis that ERC accumulation relies on active DNA replication.

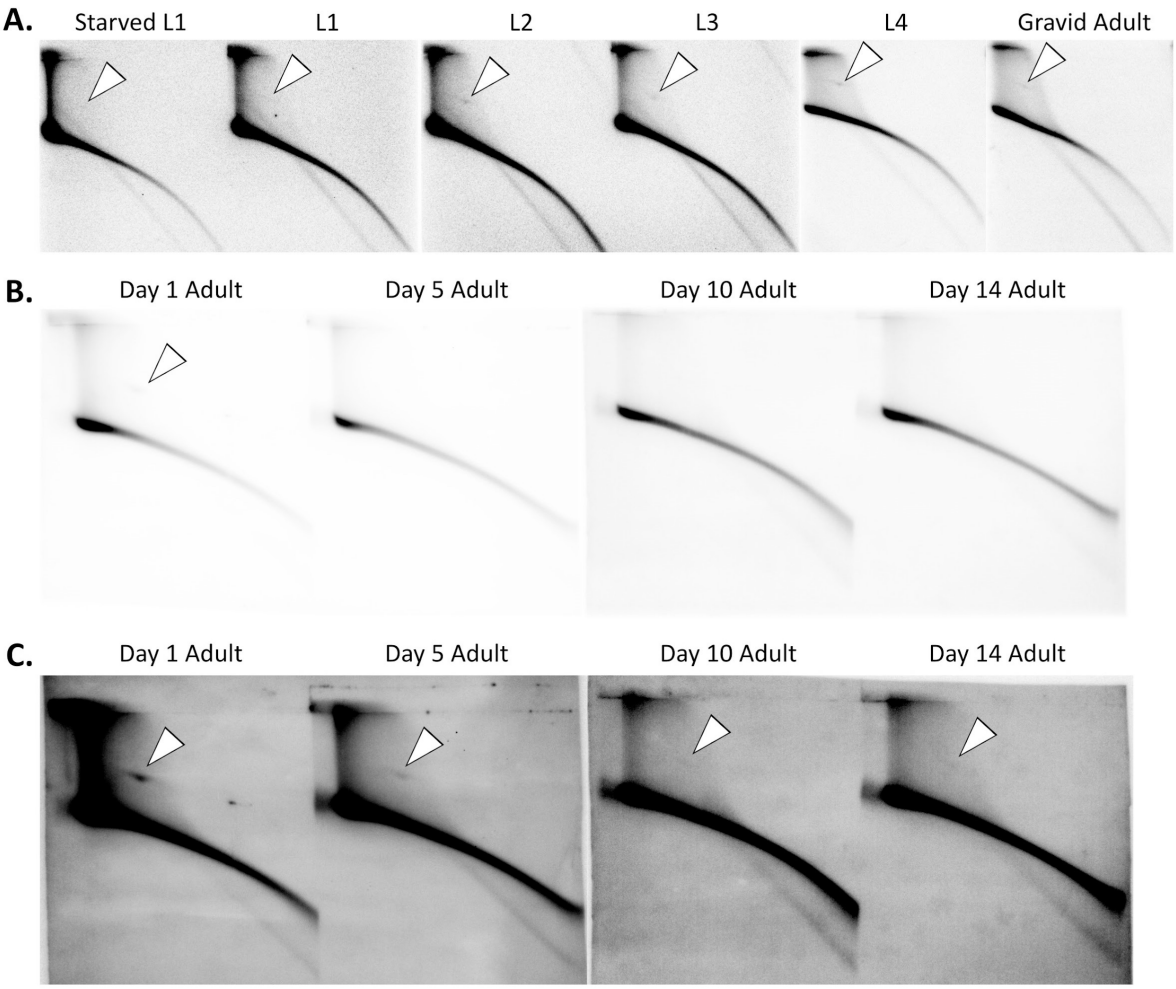

Figure A5.2: Extrachromosomal rDNA circles do not accumulate to high levels in *C. elegans*. Two-dimensional gel electrophoresis followed by rDNA-specific Southern blotting was performed in a developmental time series (**A**) and an aging time series (**B** and **C**) of N2 worms. The white carat indicates the relative position of the spot corresponding to circular rDNA. **C** is a brightness and contrast adjusted version of **B**.

A5.3: Supplemental Methods

Determination of mtDNA abundance

mtDNA abundance was determined similarly to how rDNA copy number was determined in Chapter 2. Reference sequences for the mtDNA, and chromosome 1 of the human genome (GRCh38 reference) were downloaded from the NCBI nucleotide database. CRAM files were converted to fastq by Samtools fastq and aligned to the appropriate reference sequence by bwa mem with default parameters and converted to CRAM files with samtools view. Per-base read depth was calculated with Samtools depth outputting all positions, with the –d 0 flag to eliminate a maximum read depth cutoff. Depth from positions 727-15873 of the mtDNA were used for copy number estimation, to exclude artifacts from the ends and R-loop region.

Two-dimensional gel electrophoresis of *C. elegans* DNA

Worm husbandry and sample preparation

For the developmental series, N2 worms were grown asynchronously on 10cm NGM+OP50 plates to gravid adulthood at 20°C. Gravid adults were bleached (see "hypochlorite treatment"), and embryos plated on 15cm high peptone NGM+NA22 plates to further bulk at 20°C. Three days later, gravid adult worms were again bleached, and embryos allowed to hatch in 1X M9 overlaid on an unseeded 10cm NGM plate, so that hatched L1 larvae starve and stage-synchronize. Approximately 24 hours later, starved L1s were harvested and quantified. ~20,000 L1 worms were plated on each of five 15cm high peptone NGM+NA22 plates, and an additional sample of starved L1s were pelleted and saved in ATL buffer for genomic preparation. Plated worms were grown at 25°C, and one plate harvested at each of the following time points (for genomic preparation): 8, 22-, 28-, 34-, and 50-hours post-plating, equivalent to L1, L2, L3, L4, and

gravid adult stages. DNA was prepared with Qiagen genomic preparation kit; following protocol as previously described (Appendix 1 and [5]).

For the aging series, N2 worms were grown asynchronously on 15cm high peptone (3% agarose) NGM + NA22 plates. The higher agarose concentration is to prevent burrowing. When gravid worms were abundant, bleach stage-synchronized worms. Starved L1s were harvested the next day and washed 1X in 1X M9, and 30,000 worms were plated on each of 3x 15cm high peptone (3% agarose) NGM + NA22 plates. Three days later, day one adult (D1A) worms were washed from plates in 1X M9 onto a pluriStrainer with 40µm mesh (pluriSelect). Worms on mesh were rinsed with 1X M9 to remove laid embryos or hatched larvae. Strainer was inverted onto a clean 50mL conical tube, and worms were harvested by rinsing off of mesh with 1X M9. Worms from all three plates were harvested in this manner, then re-plated onto the same type of media. An aliquot of the D1A worms was saved in TEN buffer for genomic prep. This washing and re-plating process was repeated on each day from D2A-D8A, harvesting an additional sample at D5A. Washing and re-plating was then performed on D10A, D12A, and D14A, with a sample harvested on both D10A and D14A. DNA was prepared from these samples with phenol-chloroform extraction (see below).

Phenol-chloroform extraction of DNA

Worms in TEN buffer were stored at -20°C until genomic preparation. For genomic preparation, worms in TEN buffer were thawed and 200µl of sample was transferred to a 1.5mL tube, spun down, and supernatant removed. 5 volumes of worm gDNA lysis buffer (200mM NaCl, 100mM Tris-HCl pH 8.5, 50mM EDTA pH 8.0, 0.5% SDS) with 0.1mg/mL proteinase K were added. Samples were vortexed, then incubated for 4 hours at 65°C. Proteinase K was inactivated by

incubating 30min at 95°C, then 1µl RNase A (100mg/mL from Qiagen) was added, and samples incubated for an hour at 37°C. Phenol-chloroform extraction was performed by adding 1 volume Phenol/Chloroform/Isoamyl Alcohol (25:24:1), vigorously shaking, and spinning down at 4000rpm. Aqueous layer was transferred to a new tube, and Phenol/Chloroform/Isoamyl Alcohol extraction was repeated two additional times. The final aqueous layer was transferred to a new tube, and 0.1 volumes 3M NaOAc and 1mL absolute ethanol were added to precipitate DNA. Tubes were inverted, then incubated overnight at -20°C. DNA was pelleted for 15min at 14,000rpm and pellets washed 5X in 70% EtOH. Pellets were air dried, then resuspended in 100µl molecular grade water and DNA concentration determined by nanodrop.

Gel electrophoresis

2D gel electrophoresis was performed as previously described [278,429]. For the first dimension, 100ng (developmental series) or 1µg (aging series) of each sample was loaded into a 0.55% LE agarose gel prepared in 1X TBE buffer (0.1M Tris, 0.09M boric acid, 1mM EDTA). Gel was run at 30V for 17 hours in 1X TBE, then stained with ethidium bromide. The second dimension was run at 275V at 4°C in 1% agarose containing 0.3ug/mL ethidium bromide for approximately 4 hours. Gel was transferred to a nylon membrane (Perkin Elmer GeneScreen Hybridization Transfer Membrane), as described in Appendix 1. The probe for detecting rDNA in hybridization was the same as used in Appendix 1, amplified from *C. elegans* genomic DNA with primers EMS50 and EMS51.

REFERENCES

1 Thompson, O. *et al.* (2013) The million mutation project: A new approach to genetics in Caenorhabditis elegans. *Genome Res.* 23, 1749–1762

2 Gibbons, J.G. *et al.* (2014) Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans. *Nat. Commun. Lond.* 5, 4850

3 Peter, J. *et al.* (2018) Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* 556, 339–344

4 Hall, A.N. *et al.* (2021) Thousands of high-quality sequencing samples fail to show meaningful correlation between 5S and 45S ribosomal DNA arrays in humans. *Sci. Rep.* 11, 449

5 Morton, E.A. *et al.* (2019) Challenges and Approaches to Genotyping Repetitive DNA. *G3 Genes Genomes Genet.* DOI: 10.1534/g3.119.400771

6 Paredes, S. *et al.* (2011) Ribosomal DNA Deletions Modulate Genome-Wide Gene Expression: "rDNA–Sensitive" Genes and Natural Variation. *PLOS Genet.* 7, e1001376

7 Grummt, I. (2010) Wisely chosen paths--regulation of rRNA synthesis: delivered on 30 June 2010 at the 35th FEBS Congress in Gothenburg, Sweden. *FEBS J.* 277, 4626–4639

8 Turi, Z. *et al.* (2019) Impaired ribosome biogenesis: mechanisms and relevance to cancer and aging. *Aging* 11, 2512–2540

9 McClintock, B. (1934) The relation of a particular chromosomal element to the development of the nucleoli in Zea mays. *Z. Für Zellforsch. Mikrosk. Anat.* 21, 294–326

10 Sylvester, J.E. *et al.* (1986) The human ribosomal RNA genes: structure and organization of the complete repeating unit. *Hum. Genet.* 73, 193–198

11 Borovjagin, A.V. and Gerbi, S.A. (1999) U3 small nucleolar RNA is essential for cleavage at sites 1, 2 and 3 in pre-rRNA and determines which rRNA processing pathway is taken in Xenopus oocytes. *J. Mol. Biol.* 286, 1347–1363

12 Gerbi, S.A. and Borovjagin, A.V. (2013) *Pre-Ribosomal RNA Processing in Multicellular Organisms*, Landes Bioscience.

13 Goodfellow, S.J. and Zomerdijk, J.C.B.M. (2013) Basic mechanisms in RNA polymerase I transcription of the ribosomal RNA genes. *Subcell. Biochem.* 61, 211–236

14 Geiduschek, E.P. and Tocchini-Valentini, G.P. (1988) Transcription by RNA polymerase III. *Annu. Rev. Biochem.* 57, 873–914

15 Ciganda, M. and Williams, N. (2011) Eukaryotic 5S rRNA biogenesis. *Wiley Interdiscip. Rev. RNA* 2, 523–533

16 Layat, E. *et al.* (2012) Regulation of Pol I-Transcribed 45S rDNA and Pol III-Transcribed 5S rDNA in Arabidopsis. *Plant Cell Physiol.* 53, 267–276

17 Conconi, A. *et al.* (1989) Two different chromatin structures coexist in ribosomal RNA genes throughout the cell cycle. *Cell* 57, 753–761

18 French, S.L. *et al.* (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. *Mol. Cell. Biol.* 23, 1558–1568

19 Gagnon-Kugler, T. *et al.* (2009) Loss of human ribosomal gene CpG methylation enhances cryptic RNA polymerase II transcription and disrupts ribosomal RNA processing. *Mol. Cell*

35, 414–425
20 McStay, B. and Grummt, I. (2008) The epigenetics of rRNA genes: from molecular to chromosome biology. *Annu. Rev. Cell Dev. Biol.* 24, 131–157
21 Moss, T. *et al.* (2019) The chromatin landscape of the ribosomal RNA genes in mouse and human. *Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol.* 27, 31–40
22 Xie, W. *et al.* (2012) The chromatin remodeling complex NuRD establishes the poised state of rRNA genes characterized by bivalent histone modifications and altered nucleosome positions. *Proc. Natl. Acad. Sci.* 109, 8161–8166
23 Salifou, K. *et al.* (2016) The histone demethylase JMJD2A/KDM4A links ribosomal RNA transcription to nutrients and growth factors availability. *Nat. Commun.* 7, 10174
24 Grummt, I. and Pikaard, C.S. (2003) Epigenetic silencing of RNA polymerase I transcription. *Nat. Rev. Mol. Cell Biol.* 4, 641–649
25 Proffitt, J.H. *et al.* (1984) 5-Methylcytosine is not detectable in Saccharomyces cerevisiae DNA. *Mol. Cell. Biol.* 4, 985–988
26 Simpson, V.J. *et al.* (1986) Caenorhabditis elegans DNA does not contain 5-methylcytosine at any time during development or aging. *Nucleic Acids Res.* 14, 6711–6719
27 Kurihara, Y. *et al.* (1994) Chromosomal locations of Ag-NORs and clusters of ribosomal DNA in laboratory strains of mice. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* 5, 225–228
28 Héliot, L. *et al.* (2000) Nonrandom distribution of metaphase AgNOR staining patterns on human acrocentric chromosomes. *J. Histochem. Cytochem. Off. J. Histochem. Soc.* 48, 13–20
29 Navashin, M. (1934) Chromosome Alterations Caused by Hybridization and Their Bearing upon Certain General Genetic Problems. *Cytologia (Tokyo)* 5, 169–203
30 Pikaard, C.S. (2000) The epigenetics of nucleolar dominance. *Trends Genet.* 16, 495–500
31 Pikaard, C.S. (2000) Nucleolar dominance: uniparental gene silencing on a multi-megabase scale in genetic hybrids. In *Plant Gene Silencing* (Matzke, M. A. and Matzke, A. J. M., eds), pp. 43–57, Springer Netherlands
32 Ge, X.-H. *et al.* (2013) Nucleolar dominance and different genome behaviors in hybrids and allopolyploids. *Plant Cell Rep.* 32, 1661–1673
33 Pontes, O. *et al.* (2007) Postembryonic Establishment of Megabase-Scale Gene Silencing in Nucleolar Dominance. *PLOS ONE* 2, e1157
34 Earley, K.W. *et al.* (2010) Mechanisms of HDA6-mediated rRNA gene silencing: suppression of intergenic Pol II transcription and differential effects on maintenance versus siRNA-directed cytosine methylation. *Genes Dev.* 24, 1119–1132
35 Warsinger-Pepe, N. *et al.* (2020) Regulation of Nucleolar Dominance in Drosophila melanogaster. *Genetics* 214, 991–1004
36 Scheer, U. *et al.* (1976) Regulation of transcription of genes of ribosomal rna during amphibian oogenesis. A biochemical and morphological study. *J. Cell Biol.* 69, 465–489
37 VAN GANSEN, P. and SCHRAM, A. (1972) Evolution of the Nucleoli During Oogenesis in Xenopus Laevis Studied by Electron Microscopy. *J. Cell Sci.* 10, 339–367
38 Davidian, A. *et al.* (2021) On some structural and evolutionary aspects of rDNA amplification in oogenesis of Trachemys scripta turtles. *Cell Tissue Res.* 383, 853–864
39 Tian, Q. *et al.* (2001) Function of basonuclin in increasing transcription of the ribosomal

RNA genes during mouse oogenesis. *Development* 128, 407–416
40 Zahradkal, P. *et al.* (1991) Regulation of ribosome biogenesis in differentiated rat myotubes. *Mol. Cell. Biochem.* 104, 189–194
41 Brandenburger, Y. *et al.* (2003) Cardiac hypertrophy in vivo is associated with increased expression of the ribosomal gene transcription factor UBF. *FEBS Lett.* 548, 79–84
42 Neben, C.L. *et al.* (2017) Ribosome biogenesis is dynamically regulated during osteoblast differentiation. *Gene* 612, 29–35
43 Bhat, N.K. *et al.* (1987) Temporal and tissue-specific expression of mouse ets genes. *Proc. Natl. Acad. Sci.* 84, 3161–3165
44 Christensen, A.H. and Quail, P.H. (1989) Sequence analysis and transcriptional regulation by heat shock of polyubiquitin transcripts from maize. *Plant Mol. Biol.* 12, 619–632
45 Rubbi, C.P. and Milner, J. (2003) Disruption of the nucleolus mediates stabilization of p53 in response to DNA damage and other stresses. *EMBO J.* 22, 6068–6077
46 Kruhlak, M. *et al.* (2007) The ATM repair pathway inhibits RNA polymerase I transcription in response to chromosome breaks. *Nature* 447, 730–734
47 Boulon, S. *et al.* (2010) The Nucleolus under Stress. *Mol. Cell* 40, 216–227
48 Samarrai, W. *et al.* (2011) Differential Responses of Bacillus subtilis rRNA Promoters to Nutritional Stress. *J. Bacteriol.* 193, 723–733
49 Kampen, K.R. *et al.* (2020) Hallmarks of ribosomopathies. *Nucleic Acids Res.* 48, 1013–1028
50 De Keersmaecker, K. *et al.* (2015) Ribosomopathies and the paradox of cellular hypo- to hyperproliferation. *Blood* 125, 1377–1382
51 Tyagi, A. *et al.* A Review of Diamond-Blackfan Anemia: Current Evidence on Involved Genes and Treatment Modalities. *Cureus* 12, e10019
52 Flygare, J. and Karlsson, S. (2007) Diamond-Blackfan anemia: erythropoiesis lost in translation. *Blood* 109, 3152–3154
53 Gazda, H.T. *et al.* (2008) Ribosomal protein L5 and L11 mutations are associated with cleft palate and abnormal thumbs in Diamond-Blackfan anemia patients. *Am. J. Hum. Genet.* 83, 769–780
54 Quarello, P. *et al.* (2016) Ribosomal RNA analysis in the diagnosis of Diamond-Blackfan Anaemia. *Br. J. Haematol.* 172, 782–785
55 Thiel, C.T. *et al.* (2005) Severely Incapacitating Mutations in Patients with Extreme Short Stature Identify RNA-Processing Endoribonuclease RMRP as an Essential Cell Growth Regulator. *Am. J. Hum. Genet.* 77, 795–806
56 Goldfarb, K.C. and Cech, T.R. (2017) Targeted CRISPR disruption reveals a role for RNase MRP RNA in human preribosomal RNA processing. *Genes Dev.* DOI: 10.1101/gad.286963.116
57 Mason, P.J. and Bessler, M. (2011) The genetics of dyskeratosis congenita. *Cancer Genet.* 204, 635–645
58 Taskinen, M. *et al.* (2008) Extended follow-up of the Finnish cartilage-hair hypoplasia cohort confirms high incidence of non-Hodgkin lymphoma and basal cell carcinoma. *Am. J. Med. Genet. A.* 146A, 2370–2375
59 Vlachos, A. *et al.* (2018) Increased risk of colon cancer and osteogenic sarcoma in Diamond-Blackfan anemia. *Blood* 132, 2205–2208
60 Pianese, G. (1896) *Beitrag zur Histologie und Aetiologie des Carcinoms*, G. Fischer.

61 Derenzini, M. *et al.* (1998) Nucleolar function and size in cancer cells. *Am. J. Pathol.* 152, 1291–1297
62 Derenzini, M. *et al.* (2017) Ribosome biogenesis and cancer. *Acta Histochem.* 119, 190–197
63 Montanaro, L. *et al.* (2008) Nucleolus, Ribosomes, and Cancer. *Am. J. Pathol.* 173, 301–310
64 Derenzini, M. *et al.* (2009) What the nucleolus says to a tumour pathologist. *Histopathology* 54, 753–762
65 Voutev, R. *et al.* (2006) Alterations in ribosome biogenesis cause specific defects in C. elegans hermaphrodite gonadogenesis. *Dev. Biol.* 298, 45–58
66 Hein, N. *et al.* (2013) The nucleolus: an emerging target for cancer therapy. *Trends Mol. Med.* 19, 643–654
67 Quin, J.E. *et al.* (2014) Targeting the nucleolus for cancer intervention. *Biochim. Biophys. Acta-Mol. Basis Dis.* 1842, 802–816
68 Philippsen, P. *et al.* (1978) Unique arrangement of coding sequences for 5 S, 5.8 S, 18 S and 25 S ribosomal RNA in Saccharomyces cerevisiae as determined by R-loop and hybridization analysis. *J. Mol. Biol.* 123, 387–404
69 Petes, T.D. (1979) Yeast ribosomal DNA genes are located on chromosome XII. *Proc. Natl. Acad. Sci.* 76, 410–414
70 Henderson, A.S. *et al.* (1972) Location of Ribosomal DNA in the Human Chromosome Complement. *Proc. Natl. Acad. Sci. U. S. A.* 69, 3394–3398
71 Drouin, G. *et al.* (1992) Variable arrangement of 5S ribosomal genes within the ribosomal DNA repeats of arthropods. *Mol. Biol. Evol.* 9, 826–835
72 Drouin, G. and de Sá, M.M. (1995) The concerted evolution of 5S ribosomal genes linked to the repeat units of other multigene families. *Mol. Biol. Evol.* 12, 481–493
73 Garcia, S. *et al.* (2017) Cytogenetic features of rRNA genes across land plants: analysis of the Plant rDNA database. *Plant J.* 89, 1020–1030
74 Wicke, S. *et al.* (2011) Restless 5S: The re-arrangement(s) and evolution of the nuclear ribosomal DNA in land plants. *Mol. Phylogenet. Evol.* 61, 321–332
75 Silva, A.E.B. e *et al.* (2013) Linked 5S and 45S rDNA Sites Are Highly Conserved through the Subfamily Aurantioideae (Rutaceae). *Cytogenet. Genome Res.* 140, 62–69
76 Garcia, S. and Kovařík, A. (2013) Dancing together and separate again: gymnosperms exhibit frequent changes of fundamental 5S and 35S rRNA gene (rDNA) organisation. *Heredity* 111, 23–33
77 Sousa, A. *et al.* (2020) Different from tracheophytes, liverworts commonly have mixed 35S and 5S arrays. *Ann. Bot.* 125, 1057–1064
78 Bergeron, J. and Drouin, G. (2008) The evolution of 5S ribosomal RNA genes linked to the rDNA units of fungal species. *Curr. Genet.* 54, 123–131
79 Sochorová, J. *et al.* (2018) Evolutionary trends in animal ribosomal DNA loci: introduction to a new online database. *Chromosoma* 127, 141–150
80 Cazaux, B. *et al.* (2011) Are ribosomal DNA clusters rearrangement hotspots? A case study in the genus Mus (Rodentia, Muridae). *BMC Evol. Biol.* 11, 124
81 Araújo da Silva, F. *et al.* (2019) Effects of environmental pollution on the rDNAomics of Amazonian fish. at <https://pubag.nal.usda.gov/catalog/6455095>
82 Sluis, M. van *et al.* (2020) NORs on human acrocentric chromosome p-arms are active by default and can associate with nucleoli independently of rDNA. *Proc. Natl. Acad. Sci.* 117,

10368–10377
83 Caburet, S. *et al.* (2005) Human ribosomal RNA gene arrays display a broad range of palindromic structures. *Genome Res.* 15, 1079–1085
84 Kim, J.-H. *et al.* (2018) Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read sequencing. *Nucleic Acids Res.* 46, 6712–6725
85 Nurk, S. *et al.* (2021) The complete sequence of a human genome. *bioRxiv* DOI: 10.1101/2021.05.26.445798
86 Hori, Y. *et al.* (2021) The human ribosomal RNA gene is composed of highly homogenized tandem clusters. *bioRxiv* DOI: 10.1101/2021.06.02.446762
87 Hassouna, N. *et al.* (1984) The complete nucleotide sequence of mouse 28S rRNA gene. Implications for the process of size increase of the large subunit rRNA in higher eukaryotes. *Nucleic Acids Res.* 12, 3563–3583
88 Michot, B. *et al.* (1984) Secondary structure of mouse 28S rRNA and general model for the folding of the large rRNA in eukaryotes. *Nucleic Acids Res.* 12, 4259–4279
89 Michot, B. and Bachellerie, J.-P. (1987) Comparisons of large subunit rRNAs reveal some eukaryote-specific elements of secondary structure. *Biochimie* 69, 11–23
90 Raué, H.A. *et al.* (1988) Evolutionary conservation of structure and function of high molecular weight ribosomal RNA. *Prog. Biophys. Mol. Biol.* 51, 77–129
91 Eickbush, T.H. and Eickbush, D.G. (2007) Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* 175, 477–485
92 Zimmer, E.A. *et al.* (1980) Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc. Natl. Acad. Sci. U. S. A.* 77, 2158–2162
93 Dover, G. (1982) Molecular drive: a cohesive mode of species evolution. *Nature* 299, 111–117
94 Xu, B. *et al.* (2017) Ribosomal DNA copy number loss and sequence variation in cancer. *PLOS Genet.* 13, e1006771
95 Ohashi, R. *et al.* (2020) Frequent Germline and Somatic Single Nucleotide Variants in the Promoter Region of the Ribosomal RNA Gene in Japanese Lung Adenocarcinoma Patients. *Cells* 9, 2409
96 Tseng, H. *et al.* (2008) Mouse Ribosomal RNA Genes Contain Multiple Differentially Regulated Variants. *PLOS ONE* 3, e1843
97 Parks, M.M. *et al.* (2018) Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. *Sci. Adv.* 4, eaao0665
98 Kurylo, C.M. *et al.* (2018) Endogenous rRNA Sequence Variation Can Regulate Stress Response Gene Expression and Phenotype. *Cell Rep.* 25, 236-248.e6
99 Sims, J. *et al.* (2021) Sequencing of the Arabidopsis NOR2 reveals its distinct organization and tissue-specific rRNA ribosomal variants. *Nat. Commun.* 12, 387
100 Parks, M.M. *et al.* (2019) Implications of sequence variation on the evolution of rRNA. *Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol.* 27, 89–93
101 Xue, S. and Barna, M. (2012) Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nat. Rev. Mol. Cell Biol.* 13, 355–369
102 Haag, E.S. and Dinman, J.D. (2019) Still Searching for Specialized Ribosomes. *Dev. Cell* 48, 744–746
103 Fujiwara, H. (2015) Site-specific non-LTR retrotransposons. In *Mobile DNA III* pp. 1147–

1163, John Wiley & Sons, Ltd
104 Roiha, H. *et al.* (1981) Arrangements and rearrangements of sequences flanking the two types of rDNA insertion in D. melanogaster. *Nature* 290, 749–753
105 Burke, W.D. *et al.* (1998) Are retrotransposons long-term hitchhikers? *Nature* 392, 141–142
106 Jakubczak, J.L. *et al.* (1992) Turnover of R1 (type I) and R2 (type II) retrotransposable elements in the ribosomal DNA of Drosophila melanogaster. *Genetics* 131, 129–142
107 Malik, H.S. and Eickbush, T.H. (1999) Retrotransposable elements R1 and R2 in the rDNA units of Drosophila mercatorum: abnormal abdomen revisited. *Genetics* 151, 653–665
108 Pérez-González, C.E. and Eickbush, T.H. (2001) Dynamics of R1 and R2 elements in the rDNA locus of Drosophila simulans. *Genetics* 158, 1557–1567
109 Raje, H.S. *et al.* (2018) R1 retrotransposons in the nucleolar organizers of Drosophila melanogaster are transcribed by RNA polymerase I upon heat shock. *Transcription* 9, 273–285
110 Burke, W.D. *et al.* (1995) R4, a non-LTR retrotransponson specific to the large subunit rRNA genes of nematodes. *Nucleic Acids Res.* 23, 4628–4634
111 Pérez-González, C.E. and Eickbush, T.H. (2002) Rates of R1 and R2 retrotransposition and elimination from the rDNA locus of Drosophila melanogaster. *Genetics* 162, 799–811
112 Eickbush, T.H. and Eickbush, D.G. (2015) Integration, regulation, and long-term stability of R2 retrotransposons. *Microbiol. Spectr.* 3, 10.1128/microbiolspec.MDNA3-0011–2014
113 Nelson, J.O. *et al.* (2021) The retrotransposon R2 maintains Drosophila ribosomal DNA repeats. *bioRxiv* DOI: 10.1101/2021.07.12.451825
114 Gonzalez, I.L. and Sylvester, J.E. (1995) Complete Sequence of the 43-kb Human Ribosomal DNA Repeat: Analysis of the Intergenic Spacer. *Genomics* 27, 320–328
115 Grozdanov, P. *et al.* (2003) Complete sequence of the 45-kb mouse ribosomal DNA repeat: analysis of the intergenic spacer☆. *Genomics* 82, 637–643
116 Agrawal, S. and Ganley, A.R.D. (2018) The conservation landscape of the human ribosomal RNA gene repeats. *PLOS ONE* 13, e0207531
117 Gonzalez, I.L. *et al.* (1993) Fixation times of retroposons in the ribosomal DNA spacer of human and other primates. *Genomics* 18, 29–36
118 Gorokhova, E. *et al.* (2002) Functional and ecological significance of rDNA intergenic spacer variation in a clonal organism under divergent selection for production rate. *Proc. R. Soc. Lond. B Biol. Sci.* 269, 2373–2379
119 Weider, L.J. *et al.* (2005) The Functional Significance of Ribosomal (r)DNA Variation: Impacts on the Evolutionary Ecology of Organisms. *Annu. Rev. Ecol. Evol. Syst.* 36, 219–242
120 Mateos, M. and Markow, T.A. (2005) Ribosomal intergenic spacer (IGS) length variation across the Drosophilinae (Diptera: Drosophilidae). *BMC Evol. Biol.* 5, 46
121 Havlová, K. *et al.* (2016) Variation of 45S rDNA intergenic spacers in Arabidopsis thaliana. *Plant Mol. Biol.* 92, 457–471
122 Sørensen, P.D. and Frederiksen, S. (1991) Characterization of human 5S rRNA genes. *Nucleic Acids Res.* 19, 4147–4151
123 McElroy, K.E. *et al.* (2021) Asexuality associated with marked genomic expansion of tandemly repeated rRNA and histone genes. *Mol. Biol. Evol.* DOI: 10.1093/molbev/msab121

124 Nelson, D.W. and Honda, B.M. (1985) Genes coding for 5S ribosomal RNA of the nematode Caenorhabditis elegans. *Gene* 38, 245–251
125 Ding, Q. *et al.* (2021) Genomic architecture of 5S rDNA cluster and its variations within and between species. *bioRxiv* DOI: 10.1101/2021.02.17.431734
126 Stults, D.M. *et al.* (2008) Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res.* 18, 13–18
127 Prokopowich, C.D. *et al.* (2003) The correlation between rDNA copy number and genome size in eukaryotes. *Genome* 46, 48–50
128 James, S.A. *et al.* (2009) Repetitive sequence variation and dynamics in the ribosomal DNA array of Saccharomyces cerevisiae as revealed by whole-genome resequencing. *Genome Res.* 19, 626–635
129 Li, B. *et al.* (2018) Co-regulation of ribosomal RNA with hundreds of genes contributes to phenotypic variations. *Genome Res.* DOI: 10.1101/gr.229716.117
130 Gaubatz, J. *et al.* (1976) Ribosomal RNA gene dosage as a function of tissue and age for mouse and human. *Biochim. Biophys. Acta* 418, 358–375
131 Ritossa, F.M. *et al.* (1966) On the chromosomal distribution of DNA complementary to ribosomal and soluble RNA. *Natl. Cancer Inst. Monogr.* 23, 449–472
132 Stults, D.M. *et al.* (2008) Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res.* 18, 13–18
133 Szostak, J.W. and Wu, R. (1980) Unequal crossing over in the ribosomal DNA of Saccharomyces cerevisiae. *Nature* 284, 426–430
134 Warmerdam, D.O. *et al.* (2016) Breaks in the 45S rDNA Lead to Recombination-Mediated Loss of Repeats. *Cell Rep.* 14, 2519–2527
135 Waters, E.R. and Schaal, B.A. (1996) Heat shock induces a loss of rRNA-encoding DNA repeats in Brassica nigra. *Proc. Natl. Acad. Sci.* 93, 1449–1452
136 Feng, L. *et al.* (2020) Ribosomal DNA copy number is associated with P53 status and levels of heavy metals in gastrectomy specimens from gastric cancer patients. *Environ. Int.* 138, 105593
137 Harvey, E.F. *et al.* (2020) Metal exposure causes rDNA copy number to fluctuate in mutation accumulation lines of Daphnia pulex. *Aquat. Toxicol.* 226, 105556
138 Lou, J. *et al.* (2021) Environmentally induced ribosomal DNA (rDNA) instability in human cells and populations exposed to hexavalent chromium [Cr (VI)]. *Environ. Int.* 153, 106525
139 Brewer, B.J. *et al.* (1980) Replication and meiotic transmission of yeast ribosomal RNA genes. *Proc. Natl. Acad. Sci.* 77, 6739–6743
140 Kobayashi, T. *et al.* (2004) SIR2 Regulates Recombination between Different rDNA Repeats, but Not Recombination within Individual rRNA Genes in Yeast. *Cell* 117, 441–453
141 Kobayashi, T. and Ganley, A.R.D. (2005) Recombination Regulation by Transcription-Induced Cohesin Dissociation in rDNA Repeats. *Science* 309, 1581–1584
142 Ritossa, F.M. (1968) Unstable redundancy of genes for ribosomal RNA. *Proc. Natl. Acad. Sci. U. S. A.* 60, 509–516
143 Kobayashi, T. *et al.* (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 Dev.* 12, 3821–3830
144 Mansisidor, A. *et al.* (2018) Genomic Copy-Number Loss Is Rescued by Self-Limiting

Production of DNA Circles. *Mol. Cell* 72, 583-593.e4
145 Iida, T. and Kobayashi, T. (2019) RNA Polymerase I Activators Count and Adjust Ribosomal RNA Gene Copy Number. *Mol. Cell* 73, 645-654.e13
146 Iida, T. and Kobayashi, T. (2019) How do cells count multi-copy genes?: "Musical Chair" model for preserving the number of rDNA copies. *Curr. Genet.* 65, 883–885
147 Tartof, K.D. (1974) Unequal Mitotic Sister Chromatid Exchange as the Mechanism of Ribosomal RNA Gene Magnification. *Proc. Natl. Acad. Sci.* 71, 1272–1276
148 Aldrich, J.C. and Maggert, K.A. (2015) Transgenerational Inheritance of Diet-Induced Genome Rearrangements in Drosophila. *PLOS Genet.* 11, e1005148
149 Brewer, B.J. *et al.* (1992) The arrest of replication forks in the rDNA of yeast occurs independently of transcription. *Cell* 71, 267–276
150 Ide, S. *et al.* (2007) Abnormality in initiation program of DNA replication is monitored by the highly repetitive rRNA gene array on chromosome XII in budding yeast. *Mol. Cell. Biol.* 27, 568–578
151 Saka, K. *et al.* (2016) More than 10% of yeast genes are related to genome stability and influence cellular senescence via rDNA maintenance. *Nucleic Acids Res.* 44, 4211–4221
152 Sanchez, J.C. *et al.* (2017) Defective replication initiation results in locus specific chromosome breakage and a ribosomal RNA deficiency in yeast. *PLoS Genet.* 13,
153 Kobayashi, T. and Sasaki, M. (2017) Ribosomal DNA stability is supported by many 'buffer genes'—introduction to the Yeast rDNA Stability Database. *FEMS Yeast Res.* 17,
154 Salim, D. *et al.* (2017) DNA replication stress restricts ribosomal DNA copy number. *PLoS Genet.* 13, e1007006
155 Lynch, K.L. *et al.* (2019) The effects of manipulating levels of replication initiation factors on origin firing efficiency in yeast. *PLOS Genet.* 15, e1008430
156 Kwan, E.X. *et al.* (2016) rDNA Copy Number Variants Are Frequent Passenger Mutations in Saccharomyces cerevisiae Deletion Collections and de Novo Transformants. *G3 Genes Genomes Genet.* 6, 2829–2838
157 Kwan, E.X. *et al.* (2013) A Natural Polymorphism in rDNA Replication Origins Links Origin Activation with Calorie Restriction and Lifespan. *PLoS Genet.* 9,
158 Jack, C.V. *et al.* (2015) Regulation of ribosomal DNA amplification by the TOR pathway. *Proc. Natl. Acad. Sci.* 112, 9674–9679
159 Spencer, W.P. (1944) ISO-ALLELES AT THE BOBBED LOCUS IN DROSOPHILA HYDEI POPULATIONS. *Genetics* 29, 520–536
160 Ritossa, F.M. *et al.* (1966) A Molecular Explanation of the Bobbed Mutants of Drosophila as Partial Deficiencies of "Ribosomal" DNA. *Genetics* 54, 819–834
161 Delany, M.E. *et al.* (1994) Effects of rRNA Gene Copy Number and Nucleolar Variation on Early Development: Inhibition of Gastrulation in rDNA-Deficient Chick Embryo. *J. Hered.* 85, 211–217
162 Cenik, E.S. *et al.* (2019) Maternal Ribosomes Are Sufficient for Tissue Diversification during Embryonic Development in C. elegans. *Dev. Cell* 48, 811-826.e6
163 Sanchez, J.C. *et al.* (2019) Phenotypic and Genotypic Consequences of CRISPR/Cas9 Editing of the Replication Origins in the rDNA of Saccharomyces cerevisiae. *Genetics* 213, 229–249
164 Mohan, J. and Ritossa, F.M. (1970) Regulation of ribosomal RNA synthesis and its bearing on the bobbed phenotype in Drosophila melanogaster. *Dev. Biol.* 22, 495–512

165 Lopez, F.B. *et al.* (2021) Gene dosage compensation of rRNA transcript levels in Arabidopsis thaliana lines with reduced ribosomal gene copy number. *Plant Cell* 33, 1135–1150
166 Kwan, E.X. *et al.* (2021) Coordination of genome replication and anaphase entry by rDNA copy number in S. cerevisiae. *bioRxiv* DOI: 10.1101/2021.02.25.432950
167 Michel, A.H. *et al.* (2005) Spontaneous rDNA copy number variation modulates Sir2 levels and epigenetic gene silencing. *Genes Dev.* 19, 1199–1210
168 Paredes, S. and Maggert, K.A. (2009) Ribosomal DNA contributes to global chromatin regulation. *Proc. Natl. Acad. Sci. U. S. A.* 106, 17829–17834
169 O'Sullivan, J.M. *et al.* (2013) The nucleolus: a raft adrift in the nuclear sea or the keystone in nuclear structure? *Biomol. Concepts* 4, 277–286
170 Németh, A. *et al.* (2010) Initial Genomics of the Human Nucleolus. *PLOS Genet.* 6, e1000889
171 van Koningsbruggen, S. *et al.* (2010) High-Resolution Whole-Genome Sequencing Reveals That Specific Chromatin Domains from Most Human Chromosomes Associate with Nucleoli. *Mol. Biol. Cell* 21, 3735–3748
172 Yu, S. and Lemos, B. (2016) A Portrait of Ribosomal DNA Contacts with Hi-C Reveals 5S and 45S rDNA Anchoring Points in the Folded Human Genome. *Genome Biol. Evol.* 8, 3545–3558
173 Yu, S. and Lemos, B. (2018) The long-range interaction map of ribosomal DNA arrays. *PLOS Genet.* 14, e1007258
174 Diesch, J. *et al.* (2019) Changes in long-range rDNA-genomic interactions associate with altered RNA polymerase II gene programs during malignant transformation. *Commun. Biol.* 2, 1–14
175 Dillinger, S. *et al.* (2017) Nucleolus association of chromosomal domains is largely maintained in cellular senescence despite massive nuclear reorganisation. *PLOS ONE* 12, e0178821
176 Guarente, L. and Kenyon, C. 09-Nov-(2000) , Genetic pathways that regulate ageing in model organisms. , *Nature*. [Online]. Available: https://www.nature.com/articles/35041700. [Accessed: 09-Apr-2018]
177 Tiku, V. and Antebi, A. (2018) Nucleolar Function in Lifespan Regulation. *Trends Cell Biol.* 0,
178 Hansen, M. *et al.* Lifespan extension by conditions that inhibit translation in Caenorhabditis elegans. *Aging Cell* 6, 95–110
179 Tiku, V. *et al.* (2017) Small nucleoli are a cellular hallmark of longevity. *Nat. Commun.* 8, ncomms16083
180 Buchwalter, A. and Hetzer, M.W. (2017) Nucleolar expansion and elevated protein translation in premature aging. *Nat. Commun.* 8, 328
181 Bemiller, P.M. and Lee, L.H. (1978) Nucleolar changes in senescing WI-38 cells. *Mech. Ageing Dev.* 8, 417–427
182 Buys, C.H.C.M. *et al.* (1979) Age-dependent variability of ribosomal RNA-gene activity in man as determined from frequencies of silver staining nucleolus organizing regions on metaphase chromosomes of lymphocytes and fibroblasts. *Mech. Ageing Dev.* 11, 55–75
183 Thomas, S. and Mukherjee, A.B. (1996) A longitudinal study of human age-related ribosomal RNA gene activity as detected by silver-stained NORs. *Mech. Ageing Dev.* 92,

101–109
184 Fabian, T.J. and Johnson, T.E. (1995) Total RNA, rRNA and poly(A)+ RNA abundances during aging in Caenorhabditis elegans. *Mech. Ageing Dev.* 83, 155–170
185 Halle, J.P. *et al.* (1997) Copy number, epigenetic state and expression of the rRNA genes in young and senescent rat embryo fibroblasts. *Eur. J. Cell Biol.* 74, 281–288
186 Lessard, F. *et al.* (2018) Senescence-associated ribosome biogenesis defects contributes to cell cycle arrest through the Rb pathway. *Nat. Cell Biol.* 20, 789–799
187 Anisimova, A.S. *et al.* (2020) Multifaceted deregulation of gene expression and protein synthesis with age. *Proc. Natl. Acad. Sci.* 117, 15581–15590
188 Swisshelm, K. *et al.* (1990) Age-related increase in methylation of ribosomal genes and inactivation of chromosome-specific rRNA gene clusters in mouse. *Mutat. Res.* 237, 131–146
189 Machwe, A. *et al.* (2000) Accelerated methylation of ribosomal RNA genes during the cellular senescence of Werner syndrome fibroblasts. *FASEB J.* 14, 1715–1724
190 Oakes, C.C. *et al.* (2003) Aging results in hypermethylation of ribosomal DNA in sperm and liver of male rats. *Proc. Natl. Acad. Sci.* 100, 1775–1780
191 D'Aquila, P. *et al.* (2017) Methylation of the ribosomal RNA gene promoter is associated with aging and age-related decline. *Aging Cell* 16, 966–975
192 Faria, T.C. *et al.* (2020) Characterization of Cerebellum-Specific Ribosomal DNA Epigenetic Modifications in Alzheimer's Disease: Should the Cerebellum Serve as a Control Tissue After All? *Mol. Neurobiol.* DOI: 10.1007/s12035-020-01902-9
193 Wang, M. and Lemos, B. (2019) Ribosomal DNA harbors an evolutionarily conserved clock of biological aging. *Genome Res.* 29, 325–333
194 Kerepesi, C. *et al.* (2021) Epigenetic clocks reveal a rejuvenation event during embryogenesis followed by aging. *bioRxiv* DOI: 10.1101/2021.03.11.435028
195 Fairfield, E.A. *et al.* Ageing European lobsters (Homarus gammarus) using DNA methylation of evolutionarily-conserved ribosomal DNA. *Evol. Appl.* n/a,
196 Sinclair, D.A. and Guarente, L. (1997) Extrachromosomal rDNA Circles— A Cause of Aging in Yeast. *Cell* 91, 1033–1042
197 Kobayashi, T. (2008) A new role of the rDNA and nucleolus in the nucleus—rDNA instability maintains genome integrity. *BioEssays* 30, 267–272
198 Cohen, S. *et al.* (2003) Extrachromosomal Circular DNA of Tandemly Repeated Genomic Sequences in Drosophila. *Genome Res.* 13, 1133–1145
199 Shoura, M.J. *et al.* (2017) Intricate and Cell Type-Specific Populations of Endogenous Circular DNA (eccDNA) in Caenorhabditis elegans and Homo sapiens. *G3 Genes Genomes Genet.* 7, 3295–3303
200 Johnson, R. and Strehler, B.L. (1972) Loss of Genes coding for Ribosomal RNA in Ageing Brain Cells. *Nature* 240, 412–414
201 Gaubatz, J.W. and Cutler, R.G. (1978) Age-related differences in the number of ribosomal RNA genes of mouse tissues. *Gerontology* 24, 179–207
202 Strehler, B.L. *et al.* (1979) Loss of hybridizable ribosomal DNA from human post-mitotic tissues during aging: I. Age-dependent loss in human myocardium. *Mech. Ageing Dev.* 11, 371–378
203 Zafiropoulos, A. *et al.* (2005) Preferential loss of 5S and 28S rDNA genes in human adipose

tissue during ageing. *Int. J. Biochem. Cell Biol.* 37, 409–415
204 Peterson, C.R.D. *et al.* (1984) Constancy of ribosomal RNA genes during aging of mouse heart cells and during serial passage of WI-38 cells. *Arch. Gerontol. Geriatr.* 3, 115–125
205 Watada, E. *et al.* (2020) Age-dependent ribosomal DNA variations in mice. *Mol. Cell. Biol.* DOI: 10.1128/MCB.00368-20
206 Lu, K.L. *et al.* (2018) Transgenerational dynamics of rDNA copy number in Drosophila male germline stem cells. *eLife* 7, e32421
207 Eichler, E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* 11, 446–450
208 Press, M.O. *et al.* (2014) The overdue promise of short tandem repeat variation for heritability. *Trends Genet.* 30, 504–512
209 Press, M.O. *et al.* (2019) Substitutions Are Boring: Some Arguments about Parallel Mutations and High Mutation Rates. *Trends Genet.* 35, 253–264
210 Porokhovnik, L.N. and Lyapunova, N.A. (2019) Dosage effects of human ribosomal genes (rDNA) in health and disease. *Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol.* 27, 5–17
211 Hosgood, H.D., III *et al.* (2019) Variation in ribosomal DNA copy number is associated with lung cancer risk in a prospective cohort study. *Carcinogenesis* 40, 975–978
212 Romão-Corrêa, R.F. *et al.* (2004) Ribosomal DNA exhibits few alterations in human skin cancers. *J. Dermatol. Sci.* 34, 109–111
213 Stults, D.M. *et al.* (2009) Human rRNA Gene Clusters Are Recombinational Hotspots in Cancer. *Cancer Res.* 69, 9096–9104
214 Stults, D.M. *et al.* (2011) Recombination phenotypes of the NCI-60 collection of human cancer cells. *BMC Mol. Biol.* 12, 23
215 Wang, M. and Lemos, B. (2017) Ribosomal DNA copy number amplification and loss in human cancers is linked to tumor genetic context, nucleolus activity, and proliferation. *PLoS Genet.* 13, e1006994
216 Udugama, M. *et al.* (2018) Ribosomal DNA copy loss and repeat instability in ATRX-mutated cancers. *Proc. Natl. Acad. Sci.* DOI: 10.1073/pnas.1720391115
217 Valori, V. *et al.* (2019) Human rDNA copy number is unstable in metastatic breast cancers. *Epigenetics* DOI: 10.1080/15592294.2019.1649930
218 German, J. (1997) Bloom's syndrome. XX. The first 100 cancers. *Cancer Genet. Cytogenet.* 93, 100–106
219 Killen, M.W. *et al.* (2009) Loss of Bloom syndrome protein destabilizes human gene cluster architecture. *Hum. Mol. Genet.* 18, 3417–3428
220 Cunniff, C. *et al.* (2017) Bloom's Syndrome: Clinical Spectrum, Molecular Pathogenesis, and Cancer Predisposition. *Mol. Syndromol.* 8, 4–23
221 Son, J. *et al.* (2020) rDNA Chromatin Activity Status as a Biomarker of Sensitivity to the RNA Polymerase I Transcription Inhibitor CX-5461. *Front. Cell Dev. Biol.* 8,
222 Veiko, N.N. *et al.* (2003) Quantitation of Repetitive Sequences in Human Genomic DNA and Detection of an Elevated Ribosomal Repeat Copy Number in Schizophrenia: The Results of Molecular and Cytogenetic Analyses. *Mol. Biol.* 37, 349–357
223 Krzyżanowska, M. *et al.* (2015) Ribosomal DNA transcription in the dorsal raphe nucleus is increased in residual but not in paranoid schizophrenia. *Eur. Arch. Psychiatry Clin.*

Neurosci. 265, 117–126
224 Chestkov, I.V. *et al.* Abundance of ribosomal RNA gene copies in the genomes of schizophrenia patients. *Schizophr. Res.* DOI: 10.1016/j.schres.2018.01.001
225 Kolesnikova, I.S. *et al.* (2018) Alteration of rRNA gene copy number and expression in patients with intellectual disability and heteromorphic acrocentric chromosomes. *Egypt. J. Med. Hum. Genet.* 19, 129–134
226 Lyapunova, N.A. *et al.* (2017) Viability of carriers of chromosomal abnormalities depends on genomic dosage of active ribosomal genes (rRNA genes). *Russ. J. Genet.* 53, 703–711
227 Willems, T. *et al.* (2014) The landscape of human STR variation. *Genome Res.* 24, 1894–1904
228 Rabanal, F.A. *et al.* (2017) Unstable Inheritance of 45S rRNA Genes in Arabidopsis thaliana. *G3 Genes Genomes Genet.* 7, 1201–1209
229 Rabanal, F.A. *et al.* (2017) Epistatic and allelic interactions control expression of ribosomal RNA gene clusters in Arabidopsis thaliana. *Genome Biol.* 18, 75
230 Press, M.O. *et al.* (2018) Massive variation of short tandem repeats with functional consequences across strains of Arabidopsis thaliana. *Genome Res.* 28, 1169–1178
231 Fotsing, S.F. *et al.* (2019) Multi-tissue analysis reveals short tandem repeats as ubiquitous regulators of gene expression and complex traits. *bioRxiv* DOI: 10.1101/495226
232 Ummat, A. and Bashir, A. (2014) Resolving complex tandem repeats with long reads. *Bioinformatics* 30, 3491–3498
233 Yoshimura, J. *et al.* (2019) Recompleting the Caenorhabditis elegans genome. *Genome Res.* DOI: 10.1101/gr.244830.118
234 Rhoads, A. and Au, K.F. (2015) PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* 13, 278–289
235 Lower, S.S. *et al.* (2018) Satellite DNA evolution: old ideas, new approaches. *Curr. Opin. Genet. Dev.* 49, 70–78
236 Michael, T.P. *et al.* (2018) High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat. Commun.* 9, 541
237 Tyler, A.D. *et al.* (2018) Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Sci. Rep.* 8, 1–12
238 Logsdon, G.A. *et al.* (2021) The structure, function and evolution of a complete human chromosome 8. *Nature* DOI: 10.1038/s41586-021-03420-7
239 Carlson, K.D. *et al.* (2015) MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Res.* 25, 750–761
240 Willems, T. *et al.* (2017) Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* 14, 590–592
241 Bakhtiari, M. *et al.* (2018) Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.* 28, 1709–1719
242 Saini, S. *et al.* (2018) A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nat. Commun.* 9, 4397
243 Matyášek, R. *et al.* (2019) Intragenomic heterogeneity of intergenic ribosomal DNA spacers in Cucurbita moschata is determined by DNA minisatellites with variable potential to form non-canonical DNA conformations. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* 26, 273–286

244 Undurraga, S.F. *et al.* (2012) Background-dependent effects of polyglutamine variation in the Arabidopsis thaliana gene ELF3. *Proc. Natl. Acad. Sci. U. S. A.* 109, 19363–19367
245 Quilez, J. *et al.* (2016) Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.* 44, 3750–3762
246 Gymrek, M. *et al.* (2016) Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* 48, 22–29
247 Rogers, S.O. and Bendich, A.J. (1987) Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. *Plant Mol. Biol.* 9, 509–520
248 Lyckegaard, E.M. and Clark, A.G. (1989) Ribosomal DNA and Stellate gene copy number variation on the Y chromosome of Drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.* 86, 1944–1948
249 Bughio, F. and Maggert, K.A. (2019) The peculiar genetics of the ribosomal DNA blurs the boundaries of transgenerational epigenetic inheritance. *Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol.* 27, 19–30
250 Long, Q. *et al.* (2013) Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat. Genet.* 45, 884–890
251 Porokhovnik, L. (2019) Individual Copy Number of Ribosomal Genes as a Factor of Mental Retardation and Autism Risk and Severity. *Cells* 8, 1151
252 Hallgren, J. *et al.* (2014) Neurodegeneration-associated instability of ribosomal DNA. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* 1842, 860–868
253 Rubio-Piña, J. *et al.* (2016) A quantitative PCR approach for determining the ribosomal DNA copy number in the genome of Agave tequila weber. *Electron. J. Biotechnol.* 22, 9–15
254 Ligerman, C.B. *et al.* (2018) Instability in 18S and 5.8S rDNA copy numbers and 18S:5.8S ratios in longitudinally-collected human DNA samples detected by monochrome multiplex qPCR. *bioRxiv* DOI: 10.1101/361840
255 Weaver, S. *et al.* (2010) Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods San Diego Calif* 50, 271–276
256 Hindson, B.J. *et al.* (2011) High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* 83, 8604–8610
257 Hindson, C.M. *et al.* (2013) Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nat. Methods* 10, 1003–1005
258 Gibbons, J.G. *et al.* (2015) Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *Proc. Natl. Acad. Sci.* 112, 2485–2490
259 Glenn, T.C. (2011) Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* 11, 759–769
260 Guo, Y. *et al.* (2012) The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* 13, 666
261 Benjamini, Y. and Speed, T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 40, e72–e72
262 Kobayashi, T. *et al.* (2001) Identification of DNA cis Elements Essential for Expansion of Ribosomal DNA Repeats inSaccharomyces cerevisiae. *Mol. Cell. Biol.* 21, 136–147
263 Nomura, M. *et al.* (2013) *Transcription of rDNA in the Yeast Saccharomyces cerevisiae*,

Landes Bioscience.
264 Woolford, J.L. and Baserga, S.J. (2013) Ribosome Biogenesis in the Yeast Saccharomyces cerevisiae. *Genetics* 195, 643–681
265 Vogelstein, B. and Kinzler, K.W. (1999) Digital PCR. *Proc. Natl. Acad. Sci.* 96, 9236–9241
266 Alanio, A. *et al.* (2016) Variation in copy number of the 28S rDNA of Aspergillus fumigatus measured by droplet digital PCR and analog quantitative real-time PCR. *J. Microbiol. Methods* 127, 160–163
267 Schmickel, R.D. (1973) Quantitation of Human Ribosomal DNA: Hybridization of Human DNA with Ribosomal RNA for Quantitation and Fractionation. *Pediatr. Res.* 7, 5–12
268 Robyr, D. *et al.* (2002) Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell* 109, 437–446
269 Carter, N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* 39, S16-21
270 Verma, R.S. *et al.* (1977) Size variation polymorphisms of the short arm of human acrocentric chromosomes determined by R-banding by fluorescence using acridine orange (RFA). *Hum. Genet.* 38, 231–234
271 Lofgren, L.A. *et al.* (2019) Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles. *Mol. Ecol.* 28, 721–730
272 Payne, A. *et al.* (2019) BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinforma. Oxf. Engl.* 35, 2193–2198
273 Tyson, J.R. *et al.* (2018) MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Res.* 28, 266–274
274 Schaffer, H.E. and Sederoff, R.R. (1981) Improved estimation of DNA fragment lengths from agarose gels. *Anal. Biochem.* 115, 113–122
275 Elder, J.K. and Southern, E.M. (1983) Measurement of DNA length by gel electrophoresis II: Comparison of methods for relating mobility to fragment length. *Anal. Biochem.* 128, 227–231
276 Pourcel, C. *et al.* (2011) Identification of variable-number tandem-repeat (VNTR) sequences in Acinetobacter baumannii and interlaboratory validation of an optimized multiple-locus VNTR analysis typing scheme. *J. Clin. Microbiol.* 49, 539–548
277 De Bustos, A. *et al.* (2016) Sequencing of long stretches of repetitive DNA. *Sci. Rep.* 6, 36665
278 Tsuchiyama, S. *et al.* (2013) Sirtuins in yeast: phenotypes and tools. *Methods Mol. Biol. Clifton NJ* 1077, 11–37
279 Johnson, L.K. *et al.* (1975) Cardiac hypertrophy, aging and changes in cardiac ribosomal RNA gene dosage in man. *J. Mol. Cell. Cardiol.* 7, 125–133
280 Ono, T. *et al.* (1985) Absence of gross change in primary DNA sequence during aging process of mice. *Mech. Ageing Dev.* 32, 227–234
281 Malinovskaya, E.M. *et al.* (2018) Copy number of human ribosomal genes with aging: unchanged mean, but narrowed range and decreased variance in elderly group. *Front. Genet.* 9,
282 Pietrzak, M. *et al.* (2011) Epigenetic Silencing of Nucleolar rRNA Genes in Alzheimer's Disease. *PLOS ONE* 6, e22585
283 Ershova, E.S. *et al.* (2020) Copy number variations of satellite III (1q12) and ribosomal

repeats in health and schizophrenia. *Schizophr. Res.* DOI: 10.1016/j.schres.2020.07.022
284 Dammann, R. *et al.* (1993) Chromatin structures and transcription of rDNA in yeast Saccharomyces cerevisiae. *Nucleic Acids Res.* 21, 2331–2338
285 Jackson, D.A. *et al.* (1998) Numbers and Organization of RNA Polymerases, Nascent Transcripts, and Transcription Units in HeLa Nuclei. *Mol. Biol. Cell* 9, 1523–1536
286 Douet, J. and Tourmente, S. (2007) Transcription of the 5S rRNA heterochromatic genes is epigenetically controlled in Arabidopsis thaliana and Xenopus laevis. *Heredity* 99, 5–13
287 Smith, J.S. and Boeke, J.D. (1997) An unusual form of transcriptional silencing in yeast ribosomal DNA. *Genes Dev.* 11, 241–254
288 Ford, E. *et al.* (2006) Mammalian Sir2 homolog SIRT7 is an activator of RNA polymerase I transcription. *Genes Dev.* 20, 1075–1080
289 Iossifov, I. *et al.* (2014) The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216–221
290 Sanders, S.J. *et al.* (2015) Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* 87, 1215–1233
291 Coe, B.P. *et al.* (2019) Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* 51, 106–116
292 Turner, T.N. *et al.* (2017) Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* 171, 710-722.e12
293 An, J.-Y. *et al.* (2018) Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* 362,
294 Brandler, W.M. *et al.* (2018) Paternally inherited cis-regulatory structural variants are associated with autism. *Science* 360, 327–331
295 Werling, D.M. *et al.* (2018) An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* 50, 727–736
296 Zhou, J. *et al.* (2019) Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.* 51, 973–980
297 Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arxiv.org* at <https://arxiv.org/abs/1303.3997>
298 Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359
299 Consortium, T. 1000 G.P. (2015) A global reference for human genetic variation. *Nature* 526, 68–74
300 Douet, J. *et al.* (2009) A Pol V–Mediated Silencing, Independent of RNA–Directed DNA Methylation, Applies to 5S rDNA. *PLOS Genet.* 5, e1000690
301 Douet, J. *et al.* (2008) Interplay of RNA Pol IV and ROS1 During Post-Embryonic 5S rDNA Chromatin Remodeling. *Plant Cell Physiol.* 49, 1783–1791
302 Blevins, T. *et al.* (2009) Heterochromatic siRNAs and DDM1 Independently Silence Aberrant 5S rDNA Transcripts in Arabidopsis. *PLOS ONE* 4, e5932
303 Peterson, R.C. *et al.* (1980) Characterization of two xenopus somatic 5S DNAs and one minor oocyte-specific 5S DNA. *Cell* 20, 131–141
304 Simon, L. *et al.* (2018) Genetic and epigenetic variation in 5S ribosomal RNA genes reveals genome dynamics in Arabidopsis thaliana. *Nucleic Acids Res.* 46, 3019–3033

305 Sochorová, J. *et al.* (2018) Evolutionary trends in animal ribosomal DNA loci: introduction to a new online database. *Chromosoma* 127, 141–150
306 Boncinelli, E. *et al.* (1972) rDNA magnification at the bobbed locus of the Y chromosome in Drosophila melanogaster. *Cell Differ.* 1, 133–142
307 Li, H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079
308 Brenner, S. (1974) The Genetics of Caenorhabditis Elegans. *Genetics* 77, 71–94
309 Ellis, R.E. *et al.* (1986) The rDNA of C. elegans : sequence and structure. *Nucleic Acids Res.* 14, 2345–2364
310 Zorio, D.A. *et al.* (1994) Operons as a common form of chromosomal organization in C. elegans. *Nature* 372, 270–272
311 Cook, D.E. *et al.* (2017) CeNDR, the Caenorhabditis elegans natural diversity resource. *Nucleic Acids Res.* 45, D650–D657
312 Paredes, S. and Maggert, K.A. (2009) Expression of I-CreI Endonuclease Generates Deletions Within the rDNA of Drosophila. *Genetics* 181, 1661–1671
313 Zhang, Q.F. *et al.* (1990) Effects on adaptedness of variations in ribosomal DNA copy number in populations of wild barley (Hordeum vulgare ssp. spontaneum). *Proc. Natl. Acad. Sci.* 87, 8741–8745
314 Konrad, A. *et al.* (2018) Mutational and transcriptional landscape of spontaneous gene duplications and deletions in Caenorhabditis elegans. *Proc. Natl. Acad. Sci.* 115, 7386–7391
315 Sluis, M. van and McStay, B. (2015) A localized nucleolar DNA damage response facilitates recruitment of the homology-directed repair machinery independent of cell cycle stage. *Genes Dev.* 29, 1151–1163
316 Paredes, S. *et al.* (2018) The epigenetic regulator SIRT7 guards against mammalian cellular senescence induced by ribosomal DNA instability. *J. Biol. Chem.* 293, 11242–11250
317 Ramirez-Parra, E. and Gutierrez, C. (2007) The many faces of chromatin assembly factor 1. *Trends Plant Sci.* 12, 570–576
318 Mozgová, I. *et al.* (2010) Dysfunction of Chromatin Assembly Factor 1 Induces Shortening of Telomeres and Loss of 45S rDNA in Arabidopsis thaliana. *Plant Cell* 22, 2768–2780
319 Varas, J. *et al.* (2017) The Absence of the Arabidopsis Chaperone Complex CAF-1 Produces Mitotic Chromosome Abnormalities and Changes in the Expression Profiles of Genes Involved in DNA Repair. *Front. Plant Sci.* 8, 525
320 Picart-Picolo, A. *et al.* (2020) Large tandem duplications affect gene expression, 3D organization, and plant–pathogen response. *Genome Res.* DOI: 10.1101/gr.261586.120
321 Pavlištová, V. *et al.* (2016) Phenotypic reversion in fas mutants of Arabidopsis thaliana by reintroduction of FAS genes: variable recovery of telomeres with major spatial rearrangements and transcriptional reprogramming of 45S rDNA genes. *Plant J. Cell Mol. Biol.* 88, 411–424
322 Bik, H.M. *et al.* (2013) Intra-Genomic Variation in the Ribosomal Repeats of Nematodes. *PLOS ONE* 8, e78230
323 Wang, J. *et al.* (2010) Chromosome Size Differences May Affect Meiosis and Genome Size. *Science* 329, 293–293
324 Seidel, H.S. *et al.* (2008) Widespread Genetic Incompatibility in C. Elegans Maintained by

Balancing Selection. *Science* 319, 589–594
325 Warner, J.R. (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* 24, 437–440
326 Stern, C. (1927) Ein genetischer und zytologischer Beweis für Vererbung im Y-Chromosom vonDrosophila melanogaster. *Z. Für Indukt. Abstamm.- Vererbungslehre* 44, 187–231
327 Wu, J. *et al.* (2018) PHA-4/FoxA senses nucleolar stress to regulate lipid accumulation in Caenorhabditis elegans. *Nat. Commun.* 9, 1195
328 Ide, S. *et al.* (2010) Abundance of Ribosomal RNA Gene Copies Maintains Genome Integrity. *Science* 327, 693–696
329 Stevenson, B.S. and Schmidt, T.M. (2004) Life History Implications of rRNA Gene Copy Number in Escherichia coli. *Appl. Environ. Microbiol.* 70, 6670–6677
330 Kobayashi, T. (2011) How does genome instability affect lifespan? *Genes Cells* 16, 617–624
331 Pitt, J.N. *et al.* (2019) WormBot, an open-source robotics platform for survival and behavior analysis in C. elegans. *GeroScience* 41, 961–973
332 Kenyon, C.J. 24-Mar-(2010) , The genetics of ageing. , *Nature*. [Online]. Available: https://www.nature.com/articles/nature08980. [Accessed: 09-Apr-2018]
333 Tacutu, R. *et al.* (2018) Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res.* 46, D1083–D1090
334 Townes, F.W. *et al.* (2020) Identifying longevity associated genes by integrating gene expression and curated annotations. *PLOS Comput. Biol.* 16, e1008429
335 Tsang, W.Y. and Lemire, B.D. (2002) Mitochondrial Genome Content Is Regulated during Nematode Development. *Biochem. Biophys. Res. Commun.* 291, 8–16
336 Nass, M.M. (1972) Differential effects of ethidium bromide on mitochondrial and nuclear DNA synthesis in vivo in cultured mammalian cells. *Exp. Cell Res.* 72, 211–222
337 Kamal, M. *et al.* (2016) Loss of hif-1 promotes resistance to the exogenous mitochondrial stressor ethidium bromide in Caenorhabditis elegans. *BMC Cell Biol.* 17, 34
338 Zhou, X. *et al.* (2017) RdRP-synthesized antisense ribosomal siRNAs silence pre-rRNA via the nuclear RNAi pathway. *Nat. Struct. Mol. Biol.* 24, 258–269
339 Zhu, C. *et al.* (2018) Erroneous ribosomal RNAs promote the generation of antisense ribosomal siRNA. *Proc. Natl. Acad. Sci.* 115, 10082–10087
340 Wang, Y. *et al.* (2020) CDE-1 suppresses the production of risiRNA by coupling polyuridylation and degradation of rRNA. *BMC Biol.* 18, 115
341 Liao, S. *et al.* (2021) Antisense ribosomal siRNAs inhibit RNA polymerase I-directed transcription in C. elegans. *Nucleic Acids Res.* DOI: 10.1093/nar/gkab662
342 Wahba, L. *et al.* (2021) An essential role for the piRNA pathway in regulating the ribosomal RNA pool in C. elegans. *Dev. Cell* 56, 2295-2312.e6
343 Simon, M. *et al.* (2014) Reduced Insulin/IGF-1 Signaling Restores Germ Cell Immortality to Caenorhabditis elegans Piwi Mutants. *Cell Rep.* 7, 762–773
344 Peng, J.C. and Karpen, G.H. (2007) H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. *Nat. Cell Biol.* 9, 25–35
345 Cecere, G. and Cogoni, C. (2009) Quelling targets the rDNA locus and functions in rDNA copy number control. *BMC Microbiol.* 9, 44
346 Castel, S.E. *et al.* (2014) Dicer Promotes Transcription Termination at Sites of Replication Stress to Maintain Genome Stability. *Cell* 159, 572–583

347 Khurana, J.S. *et al.* (2018) Small RNA-mediated regulation of DNA dosage in the ciliate Oxytricha. *RNA* 24, 18–29
348 Gutbrod, M.J. and Martienssen, R.A. (2020) Conserved chromosomal functions of RNA interference. *Nat. Rev. Genet.* 21, 311–331
349 Fraser, A.G. *et al.* (2000) Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature* 408, 325–330
350 Mitchell, D.H. *et al.* (1979) Synchronous Growth and Aging of Caenorhabditis elegans in the Presence of Fluorodeoxyuridine1. *J. Gerontol.* 34, 28–36
351 Lee, Y. *et al.* (2016) Inverse correlation between longevity and developmental rate among wild C. elegans strains. *Aging* 8, 986–994
352 Sandhu, A. *et al.* (2021) Specific collagens maintain the cuticle permeability barrier in Caenorhabditis elegans. *Genetics* 217,
353 Moribe, H. *et al.* (2004) Tetraspanin protein (TSP-15) is required for epidermal integrity in Caenorhabditis elegans. *J. Cell Sci.* 117, 5209–5220
354 Sangster, T.A. *et al.* (2008) HSP90-buffered genetic variation is common in Arabidopsis thaliana. *Proc. Natl. Acad. Sci.* 105, 2969–2974
355 Press, M.O. and Queitsch, C. (2017) Variability in a Short Tandem Repeat Mediates Complex Epistatic Interactions in Arabidopsis thaliana. *Genetics* 205, 455–464
356 Cao, J. *et al.* (2017) Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661–667
357 Diag, A. *et al.* (2018) Spatiotemporal m(i)RNA Architecture and 3' UTR Regulation in the C. elegans Germline. *Dev. Cell* 47, 785-800.e8
358 Adiconis, X. *et al.* (2013) Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* 10, 623–629
359 Angeles-Albores, D. *et al.* (2018) Two new functions in the WormBase Enrichment Suite. *MicroPublication Biol.* 2018,
360 R Core Team (2018) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing.
361 Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag.
362 Okabe, M. and Ito, K. (2008) , Color Universal Design (CUD): How to Make Figures and Presentations That Are Friendly to Colorblind People. . [Online]. Available: http://jfly.uni-koeln.de/color/. [Accessed: 07-Nov-2019]
363 Broman, K.W. *et al.* (2003) R/qtl: QTL mapping in experimental crosses. *Bioinforma. Oxf. Engl.* 19, 889–890
364 Wickham, H. *et al.* (2019) Welcome to the Tidyverse. *J. Open Source Softw.* 4, 1686
365 Wickham, H. *et al.* (2021) *dplyr: A Grammar of Data Manipulation*,
366 Kassambara, A. (2020) *ggpubr: "ggplot2" Based Publication Ready Plots*,
367 Kassambara, A. (2021) *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*,
368 Ahlmann-Eltze, C. and Patil, I. (2021) *ggsignif: Significance Brackets for "ggplot2,"*
369 Therneau, T.M. *et al.* (2021) *survival: Survival Analysis*,
370 Cerqueira, A.V. and Lemos, B. (2019) Ribosomal DNA and the Nucleolus as Keystones of Nuclear Architecture, Organization, and Function. *Trends Genet.* 35, 710–723
371 Figueiredo, V.C. and McCarthy, J.J. (2018) Regulation of Ribosome Biogenesis in Skeletal Muscle Hypertrophy. *Physiology* 34, 30–42

372 Figueiredo, V.C. *et al.* (2021) Genetic and epigenetic regulation of skeletal muscle ribosome biogenesis with exercise. *J. Physiol.* DOI: 10.1113/JP281244
373 Berry, L.W. *et al.* (1997) Germ-line tumor formation caused by activation of glp-1, a Caenorhabditis elegans member of the Notch family of receptors. *Development* 124, 925–936
374 Killian, D.J. and Hubbard, E.J.A. (2005) Caenorhabditis elegans germline patterning requires coordinated development of the somatic gonadal sheath and the germ line. *Dev. Biol.* 279, 322–335
375 Kirienko, N.V. *et al.* (2010) Cancer models in C. elegans. *Dev. Dyn. Off. Publ. Am. Assoc. Anat.* 239, 1413–1448
376 Dalfó, D. *et al.* (2020) A Genome-Wide RNAi Screen for Enhancers of a Germline Tumor Phenotype Caused by Elevated GLP-1/Notch Signaling in Caenorhabditis elegans. *G3 GenesGenomesGenetics* 10, 4323–4334
377 Edgley, M.L. *et al.* (2018) *Genetic balancers*, WormBook.
378 Kim, H. *et al.* (2014) A Co-CRISPR Strategy for Efficient Genome Editing in Caenorhabditis elegans. *Genetics* 197, 1069–1080
379 Dokshin, G.A. *et al.* (2018) Robust Genome Editing with Short Single-Stranded and Long, Partially Single-Stranded DNA Donors in Caenorhabditis elegans. *Genetics* 210, 781–787
380 Ghanta, K.S. and Mello, C.C. (2020) Melting dsDNA Donor Molecules Greatly Improves Precision Genome Editing in Caenorhabditis elegans. *Genetics* 216, 643–650
381 Albertson, D.G. (1984) Localization of the ribosomal genes in Caenorhabditis elegans chromosomes by in situ hybridization using biotin-labeled probes. *EMBO J.* 3, 1227–1234
382 Stinchcomb, D.T. *et al.* (1985) Extrachromosomal DNA transformation of Caenorhabditis elegans. *Mol. Cell. Biol.* 5, 3484–3496
383 Aljohani, M.D. *et al.* (2020) Engineering rules that minimize germline silencing of transgenes in simple extrachromosomal arrays in C. elegans. *Nat. Commun.* 11, 6300
384 Duveau, F. *et al.* (2018) Fitness effects of altering gene expression noise in Saccharomyces cerevisiae. *eLife* 7, e37272
385 Crombie, T.A. *et al.* (2018) Head-to-head comparison of three experimental methods of quantifying competitive fitness in C. elegans. *PLOS ONE* 13, e0201507
386 Hillier, L.W. *et al.* (2008) Whole-genome sequencing and variant discovery in C. elegans. *Nat. Methods* 5, 183–188
387 Quail, M.A. *et al.* (2008) A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* 5, 1005–1010
388 Kozarewa, I. *et al.* (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* 6, 291–295
389 Aird, D. *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12, R18
390 Oyola, S.O. *et al.* (2012) Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. *BMC Genomics* 13, 1
391 Sato, M.P. *et al.* (2019) Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *DNA Res.* 26, 391–398

392 Lan, J.H. *et al.* (2015) Impact of three Illumina library construction methods on GC bias and HLA genotype calling. *Hum. Immunol.* 76, 166–175
393 Kebschull, J.M. and Zador, A.M. (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* 43, e143–e143
394 Sulston, J.E. and Brenner, S. (1974) The DNA of Caenorhabditis elegans. *Genetics* 77, 95–104
395 Liu, S. *et al.* (2017) Unbiased K-mer Analysis Reveals Changes in Copy Number of Highly Repetitive Sequences During Maize Domestication and Improvement. *Sci. Rep.* 7, 42444
396 Shen, F. and Kidd, J.M. (2020) Rapid, Paralog-Sensitive CNV Analysis of 2457 Human Genomes Using QuicK-mer2. *Genes* 11, 141
397 Smith, S.D. *et al.* (2015) GROM-RD: resolving genomic biases to improve read depth detection of copy number variants. *PeerJ* 3, e836
398 Roller, E. *et al.* (2016) Canvas: versatile and scalable detection of copy number variants. *Bioinformatics* 32, 2375–2377
399 Zhang, L. *et al.* (2019) Comprehensively benchmarking applications for detecting copy number variation. *PLOS Comput. Biol.* 15, e1007069
400 Rhie, A. *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592, 737–746
401 Beyter, D. *et al.* (2021) Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* DOI: 10.1038/s41588-021-00865-4
402 Heasley, L.R. and Argueso, J.L. (2021) *Genomic characterization of a pathogenic isolate of Saccharomyces cerevisiae reveals an extensive and dynamic landscape of structural variation*,
403 Langmead, B. *et al.* (2019) Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* 35, 421–432
404 Picard Tools - By Broad Institute. . [Online]. Available: http://broadinstitute.github.io/picard/. [Accessed: 30-Aug-2019]
405 Krueger, F. (2015) , Trim Galore. . [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. [Accessed: 07-Nov-2019]
406 Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10–12
407 Andrews, S. (2010) , FastQC A Quality Control tool for High Throughput Sequence Data. . [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [Accessed: 07-Nov-2019]
408 Zeileis, A. *et al.* (2019) colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes. at <https://arxiv.org/abs/1903.06490v1>
409 Rienzi, S.C.D. *et al.* (2012) Maintaining replication origins in the face of genomic change. *Genome Res.* 22, 1940–1952
410 Andersen, E.C. *et al.* (2015) A Powerful New Quantitative Genetics Platform, Combining Caenorhabditis elegans High-Throughput Fitness Assays with a Large Collection of Recombinant Strains. *G3 Genes Genomes Genet.* 5, 911–920
411 RECOMBINANT INBRED LINE PANELS (or other useful strain panels) here - Nematode

Evolution Community. . [Online]. Available: https://evolution.wormbase.org/index.php/RECOMBINANT_INBRED_LINE_PANELS_(or_other_useful_strain_panels)_here. [Accessed: 05-Sep-2021]
412 Snoek, B.L. *et al.* (2020) WormQTL2: an interactive platform for systems genetics in Caenorhabditis elegans. *Database* 2020,
413 Torres Cleuren, Y.N. *et al.* Extensive intraspecies cryptic variation in an ancient embryonic gene regulatory network. *eLife* 8, e48220
414 Moerman, D.G. and Waterston, R.H. (1984) Spontaneous unstable unc-22 IV mutations in C. elegans var. Bergerac. *Genetics* 108, 859–877
415 Ayyadevara, S. *et al.* (2001) Genetic Mapping of Quantitative Trait Loci Governing Longevity of Caenorhabditis elegans in Recombinant-Inbred Progeny of a Bergerac-BO × RC301 Interstrain Cross. *Genetics* 157, 655–666
416 Noble, L.M. *et al.* (2017) Polygenicity and Epistasis Underlie Fitness-Proximal Traits in the Caenorhabditis elegans Multiparental Experimental Evolution (CeMEE) Panel. *Genetics* 207, 1663–1685
417 Noble, L.M. *et al.* (2021) Gene-level quantitative trait mapping in Caenorhabditis elegans. *G3 GenesGenomesGenetics* 11,
418 Snoek, B.L. *et al.* (2019) A multi-parent recombinant inbred line population of C. elegans allows identification of novel QTLs for complex life history traits. *BMC Biol.* 17, 24
419 Kim, Y.-H. *et al.* (2006) Chromosome XII context is important for rDNA function in yeast. *Nucleic Acids Res.* 34, 2914–2924
420 Partridge, F.A. *et al.* (2008) The C. elegans glycosyltransferase BUS-8 has two distinct and essential roles in epidermal morphogenesis. *Dev. Biol.* 317, 549–559
421 Lee, C.-C. *et al.* (2014) Mutation of a Nopp140 gene dao-5 alters rDNA transcription and increases germ cell apoptosis in C. elegans. *Cell Death Dis.* 5, e1158
422 Wang, B.-D. *et al.* (2006) Condensin Function in Mitotic Nucleolar Segregation is Regulated by rDNA Transcription. *Cell Cycle* 5, 2260–2267
423 Johzuka, K. and Horiuchi, T. (2002) Replication fork block protein, Fob1, acts as an rDNA region specific recombinator in S. cerevisiae. *Genes Cells* 7, 99–113
424 Rincon, Q. and Maria, D. (2016) , Role of the ribosomal DNA repeats on chromosome segregation of Saccharomyces cerevisiae : a dissertation presented in fulfillment of the requirements for the degree of Doctor of Philosophy in Genetics at Massey University, Albany, New Zealand. , Thesis, Massey University
425 Hodgkin, J. (1983) Male Phenotypes and Mating Efficiency in CAENORHABDITIS ELEGANS. *Genetics* 103, 43–64
426 Hodgkin, J. *et al.* (1979) Nondisjunction Mutants of the Nematode CAENORHABDITIS ELEGANS. *Genetics* 91, 67–94
427 Chung, G. *et al.* (2015) REC-1 and HIM-5 distribute meiotic crossovers and function redundantly in meiotic double-strand break formation in Caenorhabditis elegans. *Genes Dev.* 29, 1969–1979
428 Cohen, S. *et al.* (2010) Extrachromosomal circles of satellite repeats and 5S ribosomal DNA in human cells. *Mob. DNA* 1, 11
429 Brewer, B.J. and Fangman, W.L. (1987) The localization of replication origins on ARS plasmids in S. cerevisiae. *Cell* 51, 463–471

VITA

Ashley Hall completed her Bachelor of Science in Microbiology at Indiana University in 2015. There she participated in two undergraduate research programs - Science, Technology, and Research Scholars (STARS) and Integrated Freshman Learning Experience (IFLE), which allowed her to start doing research in 2011 and continue for all four years of her degree. At IU, she worked in the Kearns lab, characterizing the gene *swrD* for its role in swarming motility in *Bacillus subtilis*. At the University of Washington, Ashley first rotated through three microbiology labs before ultimately deciding to complete her dissertation in the Queitsch lab. While she has enjoyed getting to know eukaryotic genetics, she will now be returning to bacterial genetics as a postdoc at the University of Wisconsin and Great Lakes Bioenergy Research Center.