2021

Thousands of high-quality sequencing samples fail to show meaningful correlation between 5S and 45S ribosomal DNA arrays in humans

Ashley N. Hall
University of Washington

Tychele N. Turner
Washington University School of Medicine in St. Louis

Christine Queitsch
University of Washington

scientific reports

Check for updates

OPEN

Thousands of high-quality sequencing samples fail to show meaningful correlation between 5S and 45S ribosomal DNA arrays in humans

Ashley N. Hall[1,2], Tychele N. Turner[3]✉ & Christine Queitsch[1]✉

The ribosomal RNA 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.

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[1,2]. The phenotypic consequences of natural variation in rDNA copy number remain largely unexplored[3], 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[4–8]. In some tissues such as the brain, heart, and adipose tissue, 45S rDNA copy loss has been observed with age[9–11]. 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[12]. In contrast, a recent study on aging mice found that 45S rDNA copy number increases in the blood with age in mice[13]. Additional studies in mice and rat cell lines report a lack of change in rDNA copy number with age[14–16]. In both cancer and aging studies, not all cancers and not all aging tissues or organisms appear to display changes in rDNA copy number[1,4,17,18]. 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[1]. 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

[1]Department of Genome Sciences, University of Washington, 3720 15th Ave NE, Seattle, WA 98195, USA. [2]Molecular and Cellular Biology PhD Program, University of Washington, Seattle, WA 98195, USA. [3]Department of Genetics, Washington University School of Medicine, 4523 Clayton Avenue, Campus, Box 8232, St. Louis, MO 63110, USA. ✉email: tychele@wustl.edu; queitsch@uw.edu

and Alzheimer's disease[19–22], 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[23]. 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[24]. In exponentially growing human and yeast cells, about half of the 45S rDNA copies are epigenetically silenced[25,26]. It is unknown whether a similar proportion of 5S copies are silenced in mammals; however, 5S silencing is demonstrated in *Arabidopsis thaliana* and *Xenopus laevis*[27]. 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[24]. Similarly, mutations interfering with epigenetic silencing cause the expression of additional 45S rDNA copies in mammals (or of the singular repeat in yeast)[28,29]. 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[30], sequencing-based estimates should be validated by methods such as contour-clamped homogenous 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[31]. 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[30]. 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.

Results

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[32–34]. 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 translation[35]. The Simons Simplex Collection has sequenced hundreds of families with a child affected by an autism spectrum disorder[36–40]. We estimated rDNA copy number in 7,268 individuals from families in which both of the parents, the proband, and an unaffected sibling were sequenced. We detected no difference in rDNA copy number based on autism status (Fig. 1a), using read coverage estimates (Student's paired t-test, $p = 0.9365$)[23]. 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 (Fig. 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

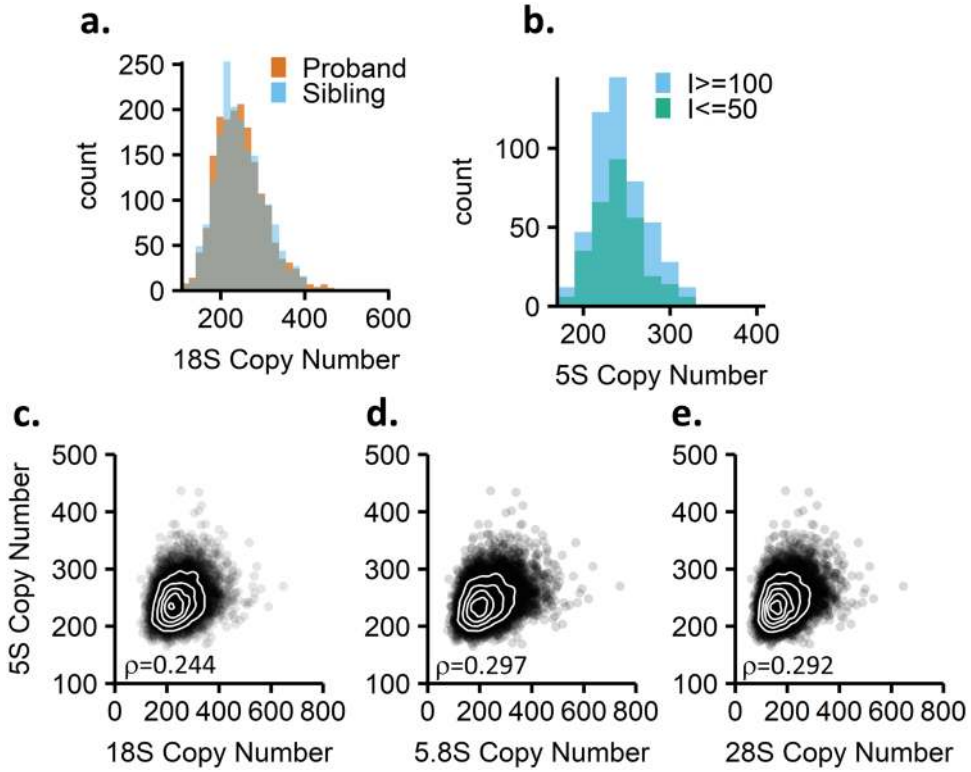

Figure 1. rDNA copy number estimates and correlations of 45S and 5S rDNA regions in the Simons Simplex Collection. (**a**) rDNA copy number distributions for probands (n = 1774) and unaffected siblings (n = 1774) in the Simons Simplex Collection. Paired t-test p-value: 0.937. (**b**) Comparison of 18S 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 = 7210).

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

Table 1. Correlations between 45S and 5S rDNA copy numbers in the Simons simplex collection. The Spearman correlation coefficient and linear models that describe the relationships between the 45S and 5S rDNA copy numbers are shown.

disability, yielded similar results (Supplemental Fig. 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[30], 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[31]. 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[23], 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 (Fig. 1c, Table 1). As we were primarily interested in

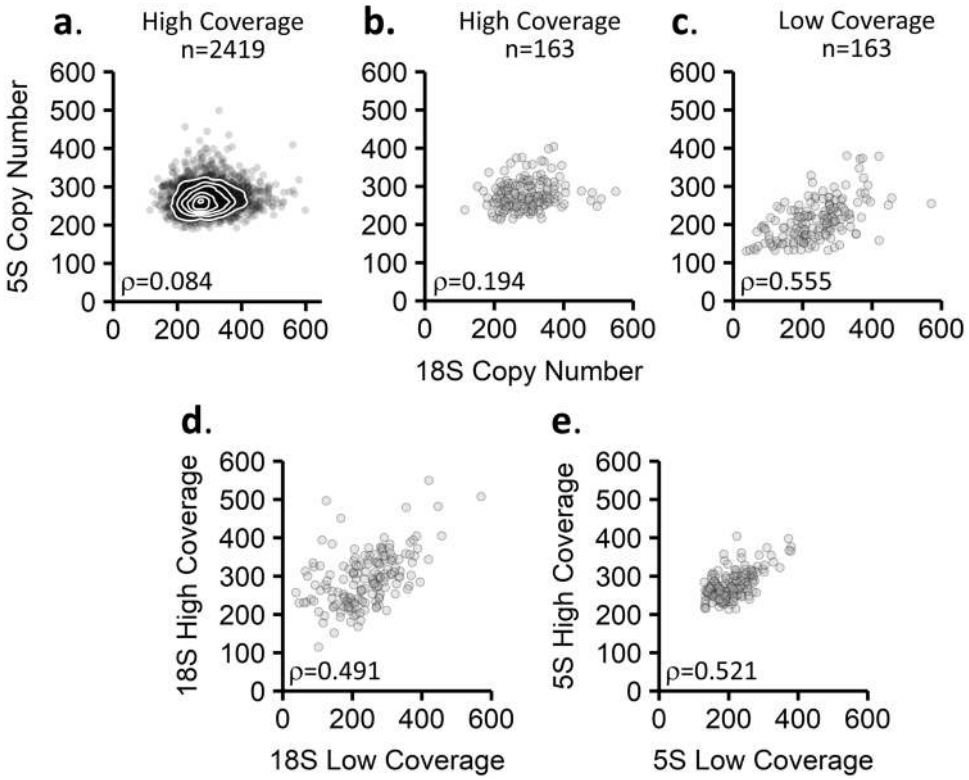

Figure 2. Correlations of the 5S and 45S rDNA copy numbers in 1000 Genomes Project data. (**a**) Correlation of the 18S to the 5S in the high-coverage 1000 Genomes Project data (n = 2419). (**b**) Correlation of the 18S to the 5S in the subset of 1000 Genomes Project data samples also analyzed in the 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.

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 2. Correlations between 45S and 5S rDNA copy numbers in the high-coverage 1000 Genomes Project dataset. The Spearman correlation coefficient and linear models that describe the relationships between the 45S and 5S rDNA copy numbers are shown.

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 (Fig. 1d,e, Table 1).

Concerted copy number variation in high-coverage 1000 Genomes Project 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 2500 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 ~ 30×. 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 (Fig. 2a, Table 2, and Supplemental Fig. 2). Despite the significance of the correlation, there is no predictive power to the

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

Table 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. The Spearman correlation coefficient and linear models that describe the relationships between the 45S and 5S rDNA copy numbers are shown.

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

Table 4. Correlations between 45S and 5S rDNA copy numbers in the low-coverage 1000 Genomes Project dataset. The Spearman correlation coefficient and linear models that describe the relationships between the 45S and 5S rDNA copy numbers are shown.

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).

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 (Fig. 2b, Table 3, and Supplemental Fig. 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[23].

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[41], while bowtie2 was used in the original study[23,42]. 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[23]. 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 4, Fig. 2c, Supplemental Fig. 2). These coefficients are similar to those previously published, which are 0.61, 0.80, and 0.73, respectively[23]. 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 (Fig. 2d,e, Supplemental Fig. 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[14,43]. Taken together, our results are consistent with previous reports that different library preparations of the same samples often yield different rDNA copy number estimates[30].

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[44], and library preparation methods and batch effects can influence rDNA copy number estimates[30]. 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.

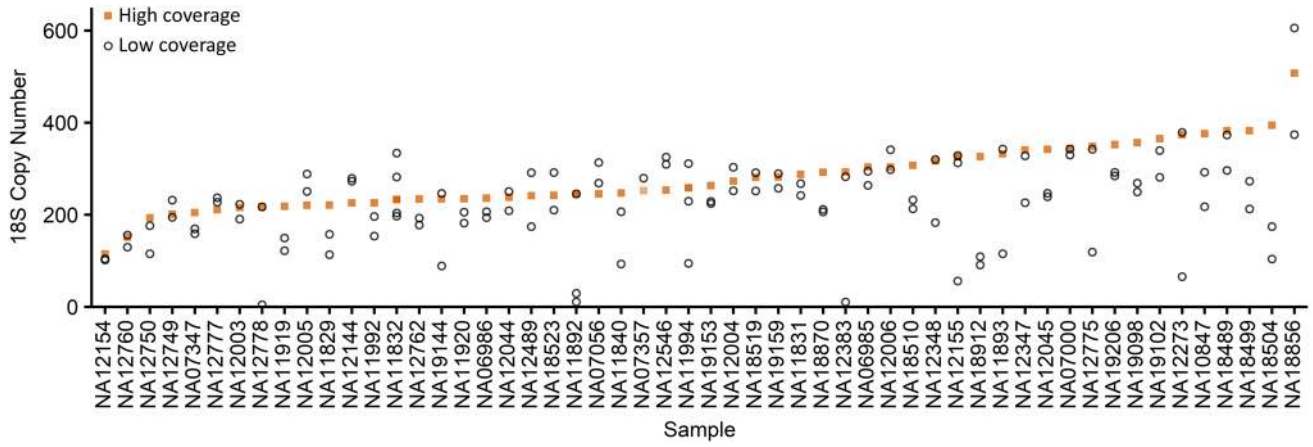

Figure 3. Comparison of 18S copy number estimates for different libraries made from the same cell lines. Of the low-coverage 1000 Genomes Project samples, 54 contained data from multiple sequencing libraries. Orange squares denote the high-coverage 1000 Genomes Project estimate. Open black circles indicate individual library estimates for each sample. Samples are ordered by high-coverage 18S estimate.

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 0.24× to 14.4×, with an average of 5.4×. 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 (Supplemental Fig. 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 (Fig. 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.

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 of 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 (Fig. 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.

For comparison, the individual libraries for the low-coverage 1000 Genomes Project samples have a mean coverage of 5.4× 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 4.5×. For the four samples analyzed, at ~ 4.5× coverage, the copy number estimates are at most ± 13 copies of the full dataset estimate. Even at 10 million reads (~ 0.45× 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.

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[1]. Because these components are part of the same array, they should correlate

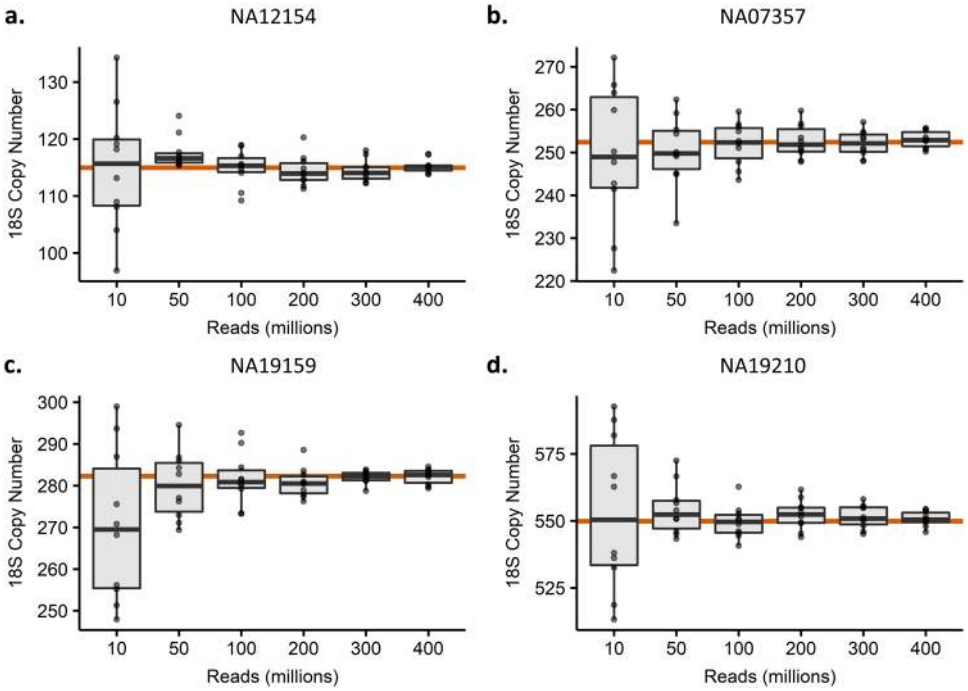

Figure 4. Read coverage downsampling of four high-coverage 1000 Genomes Project samples. (**a**) NA12154 has an 18S copy number of 115 in the full dataset. (**b**) NA07357 has an 18S copy number of 252 in the full dataset. (**c**) Has an 18S copy number of 282 in the full dataset. (**d**) NA19210 has 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.

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 (Fig. 5a, Table 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 (Fig. 5b,c, Tables 6 and 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 6 and 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 (Supplemental Fig. 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 Project 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 (Fig. 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) (Fig. 5e,f, Supplemental Fig. 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[30], we found reasonably high correlation between the 18S copy number estimates arising from the duplicated samples (Spearman correlation 0.81, Fig. 5g, Table 8). The 5S copy number estimates also showed high correlation between the twin and duplicated samples (Spearman correlation 0.902, Fig. 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 (Supplemental Fig. 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

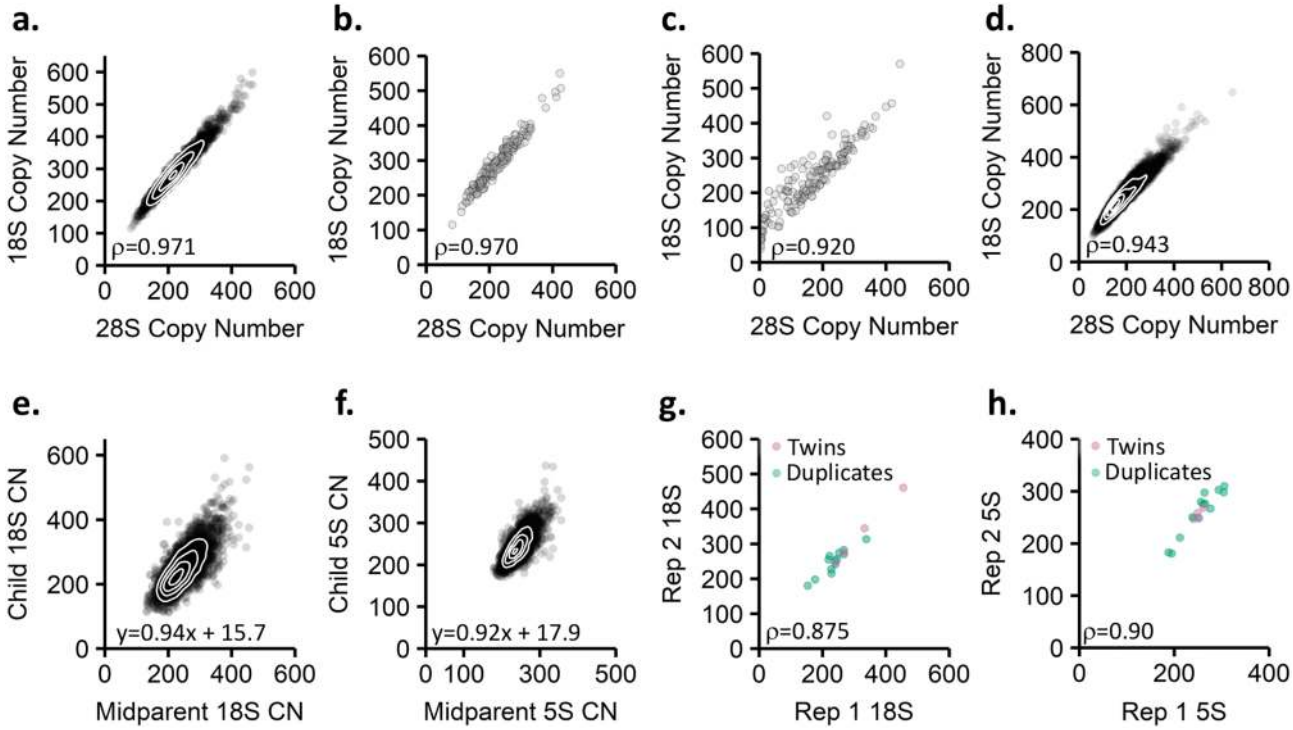

Figure 5. Data quality metrics for rDNA copy number estimates. (**a**–**d**) Correlations between the 18S and 28S regions of the 45S rDNA repeat unit for the (**a**) high-coverage 1000 Genomes Project data (n = 2,419), (**b**) subset of high-coverage 1000 Genomes Project data also analyzed in the low-coverage 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.

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 5. Correlations between the three rRNA genes encoded in the 45S repeat unit in the high-coverage 1000 Genomes Project dataset. The Spearman correlation coefficient and linear models that describe the relationships between the 18S, 5.8S, and 28S copy numbers are shown.

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 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. The Spearman correlation coefficient and linear models that describe the relationships between the 18S, 5.8S, and 28S copy numbers are shown.

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 7. Correlations between the three rRNA genes encoded in the 45S repeat unit in the low-coverage 1000 Genomes Project dataset. The Spearman correlation coefficient and linear models that describe the relationships between the 18S, 5.8S, and 28S copy numbers are shown.

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

Table 8. Correlations between the three rRNA genes encoded in the 45S repeat unit in the Simons Simplex Collection. The Spearman correlation coefficient and linear models that describe the relationships between the 18S, 5.8S, and 28S copy numbers are shown.

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.

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[30]. 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 dataset and the newer high-coverage 1000 Genomes 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 species such as *A. thaliana* and *Xenopus* species[27,45–48]. 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 45 and 5S arrays was previously observed in model organism studies. Co-variation between 45 and 5S rDNA copy numbers was weak in a large mutant collection of over 2000 strains and 40 natural isolates of *Caenorhabditis elegans*, which estimated rDNA copy number from high-coverage sequence data[49]. 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[50,51]. 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 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[14,43]

and multicellular organisms such as *C. elegans*[30,49], except for when rDNA copy number is reduced to a level that causes fitness defects such as in *Saccharomyces cerevisiae* or *Drosophila melanogaster*[52,53].

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 45 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 rely on rDNA copy number estimates from short read sequencing data[1,8,23]. However, even studies that develop their own rDNA copy number estimation methods still compare to short read sequencing data to comment on their accuracy[4,13]. 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. Even still, better sequencing and bioinformatics methods must be developed for the accurate assessment of repetitive DNA regions so that repeat copy numbers can be calculated unambiguously. 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.

Methods

Alignment and copy number estimation. *Sequence analysis.* Samtools version 1.9 and bwa version 0.7.15 were used for all analyses.

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[23]. These are positions 3841–3985 of U13369.1 for the 18S, and 8049–8198 of U13369.1 for the 28S.

Copy number estimates, ribosomal DNA read coverage and average chromosome 1 coverage for all samples in this study are provided as supplementary data files.

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

Data availability

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.

Received: 31 August 2020; Accepted: 15 December 2020
Published online: 11 January 2021

References

1. Gibbons, J. G., Branco, A. T., Yu, S. & Lemos, B. Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans. *Nat. Commun. Lond.* **5**, 4850 (2014).
2. Parks, M. M. *et al.* Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. *Sci. Adv.* **4**, eaao0665 (2018).

3. Press, M. O., Hall, A. N., Morton, E. A. & Queitsch, C. Substitutions are boring: Some arguments about parallel mutations and high mutation rates. *Trends Genet.* **35**, 253–264 (2019).
4. Xu, B. *et al.* Ribosomal DNA copy number loss and sequence variation in cancer. *PLOS Genet.* **13**, e1006771 (2017).
5. Wang, M. & Lemos, B. Ribosomal DNA copy number amplification and loss in human cancers is linked to tumor genetic context, nucleolus activity, and proliferation. *PLoS Genet.* **13**, e1006994 (2017).
6. Valori, V. *et al.* Human rDNA copy number is unstable in metastatic breast cancers. *Epigenetics* https://doi.org/10.1080/15592294.2019.1649930 (2019).
7. Stults, D. M. *et al.* Human rRNA gene clusters are recombinational hotspots in cancer. *Cancer Res.* **69**, 9096–9104 (2009).
8. Udugama, M. *et al.* Ribosomal DNA copy loss and repeat instability in ATRX-mutated cancers. *Proc. Natl. Acad. Sci.* https://doi.org/10.1073/pnas.1720391115 (2018).
9. Strehler, B. L., Chang, M.-P. & Johnson, L. K. 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 (1979).
10. Johnson, L. K., Johnson, R. W. & Strehler, B. L. Cardiac hypertrophy, aging and changes in cardiac ribosomal RNA gene dosage in man. *J. Mol. Cell. Cardiol.* **7**, 125–133 (1975).
11. Zafiropoulos, A., Tsentelierou, E., Linardakis, M., Kafatos, A. & Spandidos, D. A. Preferential loss of 5S and 28S rDNA genes in human adipose tissue during ageing. *Int. J. Biochem. Cell Biol.* **37**, 409–415 (2005).
12. Hallgren, J., Pietrzak, M., Rempala, G., Nelson, P. T. & Hetman, M. Neurodegeneration-associated instability of ribosomal DNA. *Biochim. Biophys. Acta.* **1842**, 860–868 (2014).
13. Watada, E. *et al.* Age-dependent ribosomal DNA variations in mice. *Mol. Cell. Biol.* https://doi.org/10.1128/MCB.00368-20 (2020).
14. Peterson, C. R. D., Cryar, J. R. & Gaubatz, J. W. 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 (1984).
15. Ono, T., Okada, S., Kawakami, T., Honjo, T. & Getz, M. J. Absence of gross change in primary DNA sequence during aging process of mice. *Mech. Ageing Dev.* **32**, 227–234 (1985).
16. Halle, J. P., Müller, S., Simm, A. & Adam, G. Copy number, epigenetic state and expression of the rRNA genes in young and senescent rat embryo fibroblasts. *Eur. J. Cell Biol.* **74**, 281–288 (1997).
17. Malinovskaya, E. M. *et al.* Copy number of human ribosomal genes with aging: Unchanged mean, but narrowed range and decreased variance in elderly group. *Front. Genet.* **9**, 2 (2018).
18. Romão-Corrêa, R. F., Maria, D. A., Ruiz, I. R. G., Neto, C. F. & Sanches, J. A. Ribosomal DNA exhibits few alterations in human skin cancers. *J. Dermatol. Sci.* **34**, 109–111 (2004).
19. Chestkov, I. V. *et al.* Abundance of ribosomal RNA gene copies in the genomes of schizophrenia patients. *Schizophr. Res.* https://doi.org/10.1016/j.schres.2018.01.001 (2018).
20. Veiko, N. N. *et al.* 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 (2003).
21. Pietrzak, M., Rempala, G., Nelson, P. T., Zheng, J.-J. & Hetman, M. Epigenetic silencing of nucleolar rRNA genes in Alzheimer's disease. *PLoS ONE* **6**, e22585 (2011).
22. Ershova, E. S. *et al.* Copy number variations of satellite III (1q12) and ribosomal repeats in health and schizophrenia. *Schizophr. Res.* https://doi.org/10.1016/j.schres.2020.07.022 (2020).
23. Gibbons, J. G., Branco, A. T., Godinho, S. A., Yu, S. & Lemos, B. Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *Proc. Natl. Acad. Sci.* **112**, 2485–2490 (2015).
24. French, S. L., Osheim, Y. N., Cioci, F., Nomura, M. & Beyer, A. L. 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 (2003).
25. Dammann, R., Lucchini, R., Koller, T. & Sogo, J. M. Chromatin structures and transcription of rDNA in yeast Saccharomyces cerevisiae. *Nucleic Acids Res.* **21**, 2331–2338 (1993).
26. Jackson, D. A., Iborra, F. J., Manders, E. M. M. & Cook, P. R. Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. *Mol. Biol. Cell* **9**, 1523–1536 (1998).
27. Douet, J. & Tourmente, S. Transcription of the 5S rRNA heterochromatic genes is epigenetically controlled in *Arabidopsis thaliana* and *Xenopus laevis*. *Heredity* **99**, 5–13 (2007).
28. Smith, J. S. & Boeke, J. D. An unusual form of transcriptional silencing in yeast ribosomal DNA. *Genes Dev.* **11**, 241–254 (1997).
29. Ford, E. *et al.* Mammalian Sir2 homolog SIRT7 is an activator of RNA polymerase I transcription. *Genes Dev.* **20**, 1075–1080 (2006).
30. Morton, E. A. *et al.* Challenges and approaches to genotyping repetitive DNA. *G3 Genes Genomes Genet.* https://doi.org/10.1534/g3.119.400771 (2019).
31. Stults, D. M., Killen, M. W., Pierce, H. H. & Pierce, A. J. Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res.* **18**, 13–18 (2008).
32. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
33. Sanders, S. J. *et al.* Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
34. Coe, B. P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
35. Porokhovnik, L. Individual copy number of ribosomal genes as a factor of mental retardation and autism risk and severity. *Cells* **8**, 1151 (2019).
36. Turner, T. N. *et al.* Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710-722.e12 (2017).
37. An, J.-Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, 2 (2018).
38. Brandler, W. M. *et al.* Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
39. Werling, D. M. *et al.* An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018).
40. Zhou, J. *et al.* Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.* **51**, 973–980 (2019).
41. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arxiv.org* (2013).
42. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
43. Killen, M. W., Stults, D. M., Adachi, N., Hanakahi, L. & Pierce, A. J. Loss of bloom syndrome protein destabilizes human gene cluster architecture. *Hum. Mol. Genet.* **18**, 3417–3428 (2009).
44. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
45. Douet, J., Tutois, S. & Tourmente, S. A pol V-mediated silencing, independent of RNA-directed DNA methylation, applies to 5S rDNA. *PLOS Genet.* **5**, e1000690 (2009).
46. Douet, J., Blanchard, B., Cuvillier, C. & Tourmente, S. Interplay of RNA pol IV and ROS1 during post-embryonic 5S rDNA chromatin remodeling. *Plant Cell Physiol.* **49**, 1783–1791 (2008).
47. Blevins, T., Pontes, O. & Pikaard, C. S. Heterochromatic siRNAs and DDM1 independently silence aberrant 5S rDNA transcripts in arabidopsis. *PLoS ONE* **4**, e5932 (2009).
48. Peterson, R. C., Doering, J. L. & Brown, D. D. Characterization of two xenopus somatic 5S DNAs and one minor oocyte-specific 5S DNA. *Cell* **20**, 131–141 (1980).

49. Thompson, O. *et al.* The million mutation project: A new approach to genetics in *Caenorhabditis elegans*. *Genome Res.* **23**, 1749–1762 (2013).
50. Simon, L. *et al.* Genetic and epigenetic variation in 5S ribosomal RNA genes reveals genome dynamics in *Arabidopsis thaliana*. *Nucleic Acids Res.* **46**, 3019–3033 (2018).
51. Sochorová, J., Garcia, S., Gálvez, F., Symonová, R. & Kovařík, A. Evolutionary trends in animal ribosomal DNA loci: Introduction to a new online database. *Chromosoma* **127**, 141–150 (2018).
52. Boncinelli, E., Graziani, F., Polito, L., Malva, C. & Ritossa, F. rDNA magnification at the bobbed locus of the Y chromosome in Drosophila melanogaster. *Cell Differ.* **1**, 133–142 (1972).
53. Kobayashi, T., Heck, D. J., Nomura, M. & Horiuchi, T. 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 (1998).

Acknowledgements
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.

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

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).

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-80049-y.

Correspondence and requests for materials should be addressed to T.N.T. or C.Q.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.