The long-range interaction map of ribosomal DNA arrays

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation
Yu, Shoukai, and Bernardo Lemos. 2018. “The Long-Range Interaction Map of Ribosomal DNA Arrays.” Edited by Giovanni Bosco. PLOS Genetics 14 (3) (March 23): e1007258. doi:10.1371/journal.pgen.1007258.

Published Version
doi:10.1371/journal.pgen.1007258

Accessed
July 25, 2018 8:23:56 AM EDT

Citable Link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:35979557

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP

(Article begins on next page)
The long-range interaction map of ribosomal DNA arrays

Shoukai Yu and Bernardo Lemos*

Program in Molecular and Integrative Physiological Sciences &
Department of Environmental Health,
Harvard T. H. Chan School of Public Health,
Boston, MA, USA

Address for correspondence:

Bernardo Lemos
Program in Molecular and Integrative Physiological Sciences &
Department of Environmental Health
665 Huntington Avenue, Bldg 2, Rm 219
Harvard T. H. Chan School of Public Health
Boston, MA, 02115
Email: blemos@hsph.harvard.edu

Keywords: long-range interaction, ribosomal DNA

Running title: Long-range rDNA interaction with Hi-C
Abstract
The repeated rDNA array gives rise to the nucleolus, an organelle that is central to cellular processes as varied as stress response, cell cycle regulation, RNA modification, cell metabolism, genome stability, and production of more than 70% of all cellular RNAs (the ribosomal RNAs). The rRNAs are produced from two sets of loci: the 5S rDNA array resides exclusively on human chromosome 1 while the 45S rDNA arrays reside on the short arm of five human acrocentric chromosomes. These critical genome elements have remained unassembled and have been excluded from all of the many Hi-C analyses to date. Here we built the first high resolution map of 5S and 45S rDNA array contacts with the rest of the genome combining over 15 billion Hi-C reads from several experiments. The data enabled sufficiently high coverage to map rDNA-genome interactions with <1MB resolution and identify rDNA-gene contacts. The map showed that the 5S and 45S arrays display preferential contact at common sites along the genome but are not themselves sufficiently close to yield 5S-45S Hi-C reads. Ribosomal DNA contacts are enriched in segments of closed, repressed, and late replicating chromatin, as well as CTCF binding sites. Finally, we identified categories whose dispersed genes coalesced in proximity to the rDNA arrays or instead avoided proximity with the rDNA arrays. The observations further our understanding of the spatial localization of rDNA arrays and their contribution to the architecture of the cell nucleus.
Author Summary

The repeated ribosomal DNA (rDNA) array gives rise to the nucleolus, an organelle that is involved in key cellular processes such as stress response, cell cycle regulation, RNA modification, and production of more than 70% of all cellular RNAs (the ribosomal RNAs). This critical genome element has remained unassembled and has been excluded from all of the many Hi-C analyses to date. Here we built the first map of 5S and 45S rDNA contacts with the rest of the genome. The map yielded a number of novel results and challenge the expectation that 5S and 45S arrays are close together in the nucleus. The rDNA arrays share common sites of contact across the genome, are biased towards segments of closed, repressed, and late replicating chromatin, and display greater proximity or avoidance to functionally coherent gene sets. The results further our understanding of the rDNA arrays and their localization in the nuclear environment.
**Introduction**

Ribosomal RNAs (rRNAs) are essential components of the cell, and are encoded in the 5S and 45S ribosomal DNA (rDNA) arrays of higher eukaryotes [1-4]. The 5S rDNA array resides on chromosome 1 and encodes the 5S rRNA, whereas the 45S rDNA array resides on five human acrocentric chromosomes and encodes the 18S, 5.8S, and 28S rRNA components of the ribosome [5-7]. The nucleolus, the first recognized nuclear organelle, is the site of 45S rRNA transcription [1, 2, 4, 8]. The lack of homology between the 5S rDNA and the subunits of the 45S rDNAs arrays reflect deep evolutionary separation. For instance, RNA polymerase I is exclusively dedicated to the transcription of the 45S rRNA, while RNA polymerase III transcribes the 5S rRNAs and tRNAs. The distinct RNA polymerase machineries required for transcription of 5S and 45S subunits are a conserved feature of yeasts, plants, fruit flies, and humans. Furthermore, distance to the nucleolus is thought to be relevant for global gene expression. For instance, proximity to the nucleolus can in some cases promote inactivation of certain pol II transcribed genes [9], although the observation has not been systematically tested across the genome. Finally, localization of the 5S array has been documented at the periphery of the nucleolus [9, 10], but also away from the organelle [11], with a substantial fraction of cells showing 5S arrays that are localized elsewhere in the nucleus [10]. Uncovering physical contacts between the rDNA arrays and the rest of the genome can expand our understanding of nuclear architecture, nucleolar structure and function, and the mechanism of concerted copy number variation between 5S and 45S rDNA arrays. However, studies of nuclear architecture have largely excluded analyses of spatial interactions with the 5S and 45S rDNA arrays.

Ligation-capture Hi-C sequencing technology [12-14] enabled a revolution in our understanding of nuclear organization with the identification of hundreds of topologically associated domains (TADs). Human TADs span an average 900 KB each and display remarkable conservation with TADs identified in mice. TADs display, moreover, remarkable structural stability through development and when cells are perturbed in gene knockdown experiments [15, 16]. On the other hand, deep sequencing of nucleoli led to the documentation of nucleoli associated DNA (naDNA) and the identification of nucleolus associated domains (NADs) [17-19]. While NADs
display size variation spanning multiple orders of magnitude, they are generally large. NADs covering less than 0.1 MB are relatively rare with most NADs around 1 MB or larger. The domains encompass about 5% of the human genome, are represented in all chromosomes, and are now recognized to be stably associated with nucleoli. Analysis of rDNA interactions with Hi-C might provide a complementary approach to localize the rDNA in the nuclear space possibly informing nucleolar interactions with the genome at a different scale than those afforded by analysis of naDNA.

Here we addressed the landscape of long-range rDNA interactions with 16,482,743 reads identified from a total of >15 billion (15,165,355,427) Hi-C reads in 5 cell types and 2 cell lines. The data enables a map of long-range rDNA interactions at 1MB resolution, and the identification of hundreds of segments displaying statistically significant differential contact density across cells. Our results revealed thousands of unique cell-specific contacts between the 5S and 45S rDNA arrays and their spatial localization relative to genomic regions across all chromosomes. The data identified functionally coherent GO categories whose dispersed genes either coalesce in proximity to the rDNA arrays or avoid proximity with the rDNA arrays.

Results

Ribosomal DNA containing reads in Hi-C

We investigated human Hi-C data for two cell lines and five cell types; the two cell lines represent the most replicated human Hi-C datasets to date, yet yielded a small number of rDNA informative reads. For instance, we mined 5,356,990,189 reads in LCL after quality control, to identify 13,528,436 reads with at least one end mapped to the 45S rDNA and 105,147 reads with at least one end mapped to the 5S rDNA (Supplementary Tables 1 and 2). Similarly, for K562 cells, we mined 903,837,936 reads after quality control to identify 1,698,063 reads with at least one end mapped to the 45S rDNA and 47,691 reads with at least one end mapped to the 5S rDNA. This represents a 0.25% and 0.19% recovery rate of 45S rDNA reads in shotgun Hi-C in the LCL and K562 sets, respectively. These numbers are substantially larger than the meager 0.002% and 0.005% recovery rate for 5S rDNA reads in LCL and K562, respectively. Similar recovery rates were obtained with the other five cell types studied (Table 1). Overall, we
uncovered 16,322,538 reads with at least one end mapped to the 45S rDNA reference and 160,205 reads with at least one end mapped to the 5S rDNA array (Table 1). Our mining effort illustrates the challenge in recovering rDNA information in shotgun Hi-C experiments. Nevertheless, the data revealed that rDNA contacts are dispersed across the entire genome, with segments differing in the density of interactions. The maps also revealed that naDNA and rDNA-contacts are not overlapping domains and must reflect different attributes of the nucleolus/rDNA (Figures S1).

Ribosomal DNA contact maps at 1MB resolution
Here we partitioned human autosomes (Chr 1 to 22) into 2897 segments of 1MB, 2465 and 2658 of which had no evidence of containing a 5S or 45S pseudogene, respectively. The remaining segments contained an rDNA pseudogene, were disproportionately found adjacent to centromeric and telomeric regions, and were excluded from all further analyzes. Unsurprisingly, all 1MB segments across all chromosomes displayed evidence of rDNA contact (Figure S1 and S2). Figure 1 illustrates the distribution of rDNA contact density for 1MB segments before normalization by library size. The difference in the average contact density reflects variation in the amount of Hi-C data for each cell type. The data shows a >10 fold variation in the logarithm of the contact density across segments, with all 1MB segments containing appreciable density of contacts in LCL and K562. However, ESC and the other cell types displayed a truncated distribution with many segments that contained too few rDNA contacts due to lower amounts of Hi-C reads for those cells (Figure 1A and 1C). Therefore, the following analyses focused mostly in the data from LCL and K562 cell lines but the 5 cell types are also used for comparisons.

Differential ribosomal DNA contacts density at 1MB resolution
Here we addressed variation in rDNA contact density across cell lines. We found 808 segments of 1MB with significantly different density of interactions (DI) of 45S rDNA contacts between LCL and K562 (P-value < 0.0001, Supplementary Table 3). We observed that 350 DI segments displayed increased density in LCL and 458 segments displayed increased density in K562. Among those 808 DI segments, 302 of them displayed a greater than 2-fold difference in contact density between LCL vs K562.
Similarly, nearly half of the 224 segments of 1MB in chromosome 1 showed evidence of DI density between LCL and K562 (chromosome 1: 106 segments significantly different, and 118 non-significant bins), with 97 segments displaying greater contact density in LCL and 9 segments containing greater contact density in K562. Chromosome 1 had the largest number of statistically significant DI, followed by Chr 13 (89), Chr 9 (63), and Chr 6 (61). Among the five cell types (ESC related), there are 193 segments of 1MB with DI density of rDNA contacts (P-value < 0.0001; Supplementary Table 4). For chromosome 1, for instance, there are only nine DI segments of 1MB across the five cell types (P-value < 0.0001). For the five cell types (Supplementary Table 4), Chr 22 has the largest number of segments with significant differences (22 segments); followed by Chr 2 (18 segments) and Chr 15 (14 segments). We visualized variation in 45S rDNA contact density for each segment and cell type using heat maps with the number of reads in each segment normalized by the sequencing effort in each cell line or cell type (Figure 2). Finally, we detected a meager 15 segments with evidence of differential density between LCL and K562 for the 5S rDNA (P-value < 0.0001); the small number of differential DI might partially reflect the lowered statistical power with few 5S rDNA reads, especially in the K561 set. Similarly, there is not enough Hi-C data to provide sensible statistical analysis of DI for 5S rDNA in the five ESC related cell types.

**Identification of rDNA-gene contacts**

Across 10,009 protein-coding genes, we identified 9,595 and 9,864 genes without evidence of pseudogenes for 5S and 45S, respectively. The remaining genes were excluded from all further analyzes. The data shows a continuous distribution of rDNA-gene contact density for the 45S and 5S rDNA (Figure 1). As expected, the rDNA contact density is correlated with gene length. We have thus calculated the contacted density per gene per nucleotide (“Contacts per gene, CPG”). This removed the correlations between gene length and 45S rDNA contacts and revealed that CPG is strongly correlated between LCL and K562 (rho = 0.65; P < 0.001) for the 45S rDNA arrays. This correlation is stronger than those between LCL and ESC (rho = 0.27; P < 0.001) and between K562 and ESC (rho = 0.34; P < 0.001). The lower correlations with the cell types might partially reflect the lower resolution of their 45S rDNA-gene contacts with a substantial fraction of genes showing less than 10 reads with rDNA contacts. Indeed,
although the overall amount of Hi-C data is large, the resolution to ascertain 45S rDNA-gene contacts is only sufficient for LCL and K562, the two biological sources with the largest amount of Hi-C reads to date. Although also detectable for 45S rDNA, this issue of lower rDNA-gene resolution was particularly evident for the 5S rDNA. Out of 9595 genes analyzed for 45S rDNA arrays, there were 67 and 612 genes with zero 5S contacts in LCL and K562, respectively. For ESC cell types, however, there were 1745 genes with zero contacts with 45S rDNA arrays. Out of 9864 genes analyzed for 5S rDNA arrays, there were 5916 and 7494 genes with zero 5S contacts in LCL and K562, respectively. For the ESC cell types, we observed that greater than 99% of the genes had zero 5S contacts. As in the case of the 45S the density of contacts is most strongly correlated with gene length (rho > 0.3, P < 0.001), but calculating the contact density per gene per nucleotide (“Contacts per gene, CPG”) removes the positive association. Among genes with at least one read showing 5S-rDNA contact in both LCL and K562 we find that CPG is strongly correlated between LCL and K562 (rho = 0.64, P < 0.001); the correlation remained significant but was substantially smaller when all genes are considered (rho = 0.64; P < 0.001). Finally, we find evidence for a strong positive association between the density of 45S and 5S rDNA contacts in LCL when genes with at least one rDNA contact read with both arrays are included (rho = 0.51, P < 2.2e10-16). The positive association is also evident, although substantially weakened, when all genes are considered (rho = 0.10, P < 2.2e10-16).

Differential rDNA-gene contact density
Here we tested for variation in rDNA-gene contact density across cell lines and cell types. For the 45S array, we found 731 genes with fold change in interaction density >2 for LCL vs K562 comparison (Figure S3 and Figure S4), with 40 genes displaying significantly different DI after multiple corrections with P < 0.0001 (Figure S4). For the analyses of five ESC related cell types of 45S array, there are 435 genes with significantly differential density of rDNA contacts at a P-value < 0.0001. As to rDNA-gene contact analyses of 5S array, we observed 954 genes with DI fold change >2 for LCL vs. K562 comparison. However, none of these genes reached the significance P-value < 0.0001, possibly due to the limited number of contacts for each gene especially
Per nucleotide rDNA contact rates

Here we estimated contact densities per base pair in three ways. First, the average contact per base pair across the whole genome was calculated by dividing the total numbers of mapped rDNA-genome reads by the genome length (3 billion base pairs). The average contact rate is $4.8 \times 10^{-5}$ and $3.7 \times 10^{-3}$ for the 5S and 45S rDNA, respectively (Supplementary Table 5). Hence, for the 45S rDNA each base pair in the genome is expected to have 0.37% mapped reads. Second, the average contact per base pair was estimated after filtering out bins with pseudo genes. Here we divided the total number of rDNA-genome reads within all 1MB segments without a pseudogene by the total sequence length in those segments, yielding an estimated average contact rate of $2.0 \times 10^{-5}$ and $1.7 \times 10^{-3}$ for the 5S and 45S rDNA, respectively. These numbers are comparable with those estimates using all rDNA reads and the whole genome. Third, we estimated the average contact rate per base pair in protein-coding genes by dividing the total number of rDNA-gene reads by the total length of nucleotides within genes, after excluding genes with evidence of containing rDNA pseudogenes. This yielded an average contact rate for genic segments of $2.2 \times 10^{-4}$ and 0.016 for the 5S and 45S rDNA, respectively (Supplementary Table 5). Collectively, these estimates of contact rate are useful in evaluating regions with putative enrichment or deficit in rDNA contacts.

rDNA contacts preferentially occur in close, repressed, late replicating domains

We examined the relationship between various functional elements or genomic attributes and the density of rDNA contacts. First, the data showed a significant association between the number of rDNA contacts and the A/B compartments. Specifically the B compartment of closed chromatin displays an enrichment in rDNA contacts, whereas the A compartment of open chromatin displays a deficit of rDNA contacts ($P < 0.01$, Chi-square test; Figure 3). In addition, we examined 15 functional annotations; significant enrichment were observed in segments of repressive chromatin, as well as in segments annotated as repetitive. Interestingly, insulator regions also display a highly significant enrichment in rDNA contacts ($P < 0.01$, Chi-
square test; Figure 3). Finally, we examined segments of CTCF binding; CTCF is a conserved 11-zinc finger DNA binding protein that regulates chromosome architecture [20]. Using the CTCF database we estimated that CTCF binding segments constitute <7.5% of the human genome. On the other hand, we observed that 37% and 29% of all rDNA-genome reads overlapped a CTCF binding segment in LCL and K562, respectively. These figures are in good agreement with the 35% of all rDNA-genome reads that overlapped a CTCF binding segment in the ESC cell set. These represent a >4 fold enrichment that indicate a significantly higher percentage of 45S rDNA contacts with CTCF binding sites (P < 0.05, one proportion test). On the other hand, the CTCF gene was itself depleted in rDNA contacts. CTCF is located on Chr16 and displayed a meager 118 contacts with the 45S rDNA in LCLs, which is significantly lower (P-value < 0.001, one proportion test) than the expected 1198 contacts calculated based on the genome wide average contacts per base pair (1.56%) and the length of the CTCF gene. Hence, the CTCF gene appears to be in repulsion to the rDNA arrays.

**Genes associated with rDNA CN variation are not enriched in rDNA contacts.**

We examined a small set of genes previously reported to regulate rDNA function or structure to be examined in greater detail [21-23]. To address whether these genes displayed disproportionally high or low contacts we divided the coverage total number by the gene length of each gene to obtain the average depth per base pair for the focal gene. All these genes inspected due to their co-variation with rDNA CN in natural populations had fewer hits than expected (P < 0.0001 for all of them, one proportion test). Next, we examined the top genes that are positively and negatively associated with rDNA CN in LCL [21]. These genes are neither enriched nor depleted rDNA contacts (Figure 4A), with a distribution of contacts that is undistinguishable from all other genes. Finally, ribosomal genes and mitochondrial genes were also associated with changes in the rDNA array CN [21]. Here we observed that ribosomal and mitochondrial genes show a distribution of contacts that is significantly shifted towards a greater than average number of contacts with the rDNA array.

**Ribosomal and mitochondrial genes are enriched in rDNA contacts**

Next, we addressed if the observations we made above regarding the higher density of rDNA contacts in ribosomal and mitochondrial genes would emerge as significant gene
ontology enrichments when genes with a high number of contacts are selected. To address the issue we examined the top 5% genes with higher number of 5S and 45S contacts in the LCL dataset after correction for gene length. For 5S rDNA-gene contacts the cell components category of mitochondrion (GO:0005739) emerged on the top of the list, with 56 candidates (out of 494 genes) localized to the mitochondrion. While functionally intriguing and also emerging in the K562 cell dataset, the class is not significantly enriched after correction for multiple comparisons (Supplementary Table 6). Nevertheless, the same class emerged among the top 5% in the 45S rDNA, with 62 candidates in the mitochondrion (GO:0005739; P < 0.05, after multiple correction). The class includes interesting candidates such as seryl-tRNA synthetase 2 (mitochondrial SARS2; ENSG00000104835), tRNA 5-methylaminomethyl-2-thiouridylate methyltransferase (TRMU; ENSG00000100416) and tRNA methyltransferase 1 (TRMT1; ENSG00000104907), Era like 12S mitochondrial rRNA chaperone 1 (ERAL1; ENSG00000132591). In addition, 10 other functionally coherent cell components emerged for 45S rDNA-gene contacts (Table 2 and Figure 4; P < 0.01, for all classes in LCL). Four of those are highly significant GO terms containing the protein-components of the ribosome (GO:0005840~ribosome, GO:0022625~cytosolic large ribosomal subunit, GO:0015935~small ribosomal subunit, and GO:0022627~cytosolic small ribosomal subunit). This suggests that highly transcribed genes encoding protein constituents of the ribosome are co-localized in proximity to the rDNA arrays. In addition, one GO term related to nucleolar function (GO:0005730~nucleolus) also emerged as significantly enriched with 39 genes in the top 5% of genes with higher numbers of 45S rDNA-gene contacts in LCL. These include intriguing candidates such as NOP2 nucleolar protein (NOP2; ENSG00000111641), FSHD region gene 1 (FRG1; ENSG00000109536), Sirtuin 6 (SIRT6; ENSG00000077463), and MDM2 (ENSG00000135679).

**Developmental genes are depleted in rDNA contacts**

Among the bottom 5% in 45S, we find seven HOX genes dispersed across several chromosomes (HOXA1, HOXA6, HOXA7, and HOXA11 on Chr 7, HOXB5 on Chr 17, HOXC11 on Chr 12, and HOXD13 on Chr 2), three of which showed zero 45S rDNA contacts [HOXA7(Chr 7), HOXC11(Chr 12), and HOXD13(Chr 2)]. This indicates that
developmentally regulated Hox genes are rarely localized in proximity to the rDNA arrays. Furthermore, we also found several other developmental genes in the set of 67 genes with zero contacts with 45S rDNA genes, further suggesting that developmental genes show “repulsion” from the rDNA genes. Interesting candidates include NK2 homeobox 3 (NKX2-3) on Chr 10, BMP3 on Chr 4, BMP5 and BMP6 on Chr 6, as well as NOTCH1 on Chr 9. Interestingly, the histone cluster 1 H1 family member d (HIST1H1D) on Chr 6 also emerged without a single 45S rDNA contact. Finally, we confirmed the lack of Hi-C contacts between the 5S and 45S arrays. The segments proximal to the 5S array also displayed depletion in 45S rDNA contacts. The gene RHOU, for instance, is located adjacent to the 5S array and emerged in the bottom 3% of the distribution of 45S rDNA contact density.

Discussion
Multicopy ribosomal DNA arrays are essential components of the genome. Yet ribosomal DNA arrays are also among the most variable segments of the genome. The arrays have lagged behind with limited assemblies and little understanding of their nuclear localization. Here we report a detailed contact map of spatial interactions between the rDNA arrays and the rest of the genome. Although there are huge amounts of Hi-C data, analyses of rDNA contact density for specific regions/genes remained a challenge because rDNA reads constitute a very small fraction of the Hi-C reads. Thus, we combined hundreds of Hi-C experiments to identify the subset of reads containing information on rDNA contacts. The compilation enabled us to conduct statistical tests of the differential density of rDNA interactions among two cell lines and five cell types and highlight both consistencies and variation of rDNA contact maps across cells. The effort was computational intensive because the number of rDNA reads in shotgun Hi-C is small. This is particularly evident in the case of the 5S rDNA array: the contact data remained sparse even for LCL, the cell line that has by far the largest amounts of data collected from hundreds of Hi-C experiments. Nevertheless, we identified consistency of rDNA-gene contacts across different cells (LCL and K562; especially for 45S), which point to replicable spatial interactions. The lower consistency of rDNA contacts between ESC cell types and the other two cell types (LCL and K562) could be due to the parity of rDNA Hi-C data in ESC cells. Heatmaps enabled visualization of rDNA contacts along the
human genome with statistical analyses pinpointing significant differences in the density of contacts. While the approach can be applied to other multicopy genes as well as single copy genes or regions, the resolution of shotgun Hi-C is not sufficiently high unless hundreds of datasets are available. This was the case for both the 5S rDNA and 45S rDNA arrays, which required hundreds of datasets to ascertain contacts with genic and non-genic segments of the genome.

Variation in rDNA contact density across genes reflects variation in proximity to the rDNA arrays. The data displayed over 100 fold variation in contact density across genes and revealed several intriguing patterns. First, as a class, the rDNA-proximity of genes previously identified as associated with rDNA CN variation across genotypes in human populations is undistinguishable from the background of genes. This indicates that genes globally regulated by rDNA CN are not spatially adjacent with rDNA genes and are not enriched in direct rDNA contacts. This is not an unexpected observation, because the association of gene expression variation with rDNA CN includes hundreds of genes. While genes associated with rDNA CN are presumably modulated by both direct and indirect effects emerging from the rDNA, the specific mechanism through which rDNA CN modulates gene expression is uncertain and the ratio of direct to indirect effects is unknown. Second, we observed that genes encoding protein components of the ribosome or localizing to the mitochondria or the nucleolus display a disproportionately large number of contacts with the 45S rDNA. Concordantly, genes localized to the mitochondria also emerged as enriched in 5S rDNA contacts. The data suggests that genes localized to the mitochondria might be jointly regulated through nuclear structures that are influenced by the ribosomal DNA. Noteworthy, connections between the rDNA array and mitochondrial gene expression and function have been uncovered before. In Drosophila, Paredes et al (2011) observed that engineered deletions in the rDNA array preferentially impacted the expression of genes whose protein products localized to the mitochondrion [24]. In humans, Gibbons et al (2014) observed that rDNA CN variation is associated with the expression of genes whose protein product localize to the mitochondrion as well as genes encoding protein components of the mitochondrial ribosome and mitochondrial DNA copy number [21]. Furthermore, in addition to its well-documented role as a structural component of the ribosome, the 5S
rRNA is also specifically imported into the mitochondria [25, 26]. Third, we observed that genes localized to the nucleolus and encoding protein components of the ribosome were significantly enriched for 45S rDNA contacts. The finding point to the specificity of rDNA-genome interactions and suggests that ribosomal gene regulation might be directly influenced by the rDNA array. This pattern of rDNA-gene contacts might partially explain the observation that genes whose expression was correlated with rDNA CN included components of the ribosome. Indeed, sequence specific inter-chromosomal interactions between the yeast rDNA array and an intergenic segment adjacent to the largest RNA pol I subunit has recently been demonstrated [27]. All in all, our study identified functionally coherent genes and GO terms that are depleted and enriched in the direct rDNA-genome contacts. Ribosomal DNA contacted regions for all chromosomes along the human genome suggest a structural component underlying the global regulatory consequence of rDNA CN variation [21].

Concerted copy number variation (cCNV) refers to the correlation in copy number of 5S and 45S rDNA [28]. This co-variation in copy number across genotypes with variable rDNA array size is observed in human lymphoblastoid cells (LCLs) and occurs despite 5S and 45S rDNA residence on different chromosomes and lack of sequence homology between 5S and 45S rDNA subunits. Therefore physical genome linkage between loci cannot explain the co-variation. On the other hand, spatial co-localization of the arrays as well as cellular processes of recombination such as those of micro-homology mediated end joining could conceivably contribute to the emergence of cCNV. Our results, however, confirmed a lack of direct 5S-45S contacts in Hi-C, an observation that is in agreement with a previous study [29]. This included a lack of 45S rDNA contacts with genes that are adjacent to the 5S rDNA array. The gene RHOU, for instance, is located next to the 5S and emerged in the bottom 3% of the distribution of 45S rDNA contact density. This indicates that the 5S and 45S rDNA are not in close enough proximity enough of the time to emerge with Hi-C contacts. The findings support the hypothesis that physical interactions occurring between 5S and 45S rDNA arrays are much more restricted than previously anticipated. On the other hand, the denser map presented here indicate that the 5S and 45S arrays share an overlapping contact map with many regions of the genome displaying a high density of contacts with both rDNA
arrays. For instance, the density of 5S-gene and 45S-gene contacts across genes is strongly correlated. Whether or not this shared contact maps is relevant for cCNV remains to be determined, but the evidence suggests that the two arrays are not completely independent. Coordination between them is likely to be relevant, with costs and benefits to 5S array proximity with the 45S arrays [29]. Finally, we note that as much as 29% of the 45S rDNA reads have both ends mapped in the 45S rDNA. These partially reflect linear proximity along the 45S rDNA unit but could also emerge from looping substructure with contacts between distant units, which has been suggested to facilitates ultra structural organization of the array and coordinate transcription among the rDNA repeat units [6, 7, 29-34]. All in all, the association between contact density for the 5S and 45S arrays suggest that cCNV might be facilitated by structural proximity. Similarly, rDNA mediated structural changes in the nucleus might partially explain the regulatory consequences of extensive naturally occurring variation in rDNA CN [21].

From an evolutionary perspective, the co-existence of two clusters of rDNA loci (5S and 45S) might incur costs and benefits compared to rDNA residency on a single location. In some plants and yeasts, the 5S and 45S/35S rDNA subunits are spatially adjacent in the genome [7, 35-39], whereas in Drosophila and mammals, the 5S and 45S arrays reside on different chromosomes. However, the correlated contact maps of rDNA arrays suggest however that 5S and 45S anchoring at common sites might narrow their spatial distances, and could explain why the 5S and 45S arrays can display apparent proximity to one another in a fraction of the cells as observed in cytological preparations. However, the lack of direct 5S-45S contacts might suggest a model of competitive exclusion for similar anchoring sites, and predicts that a segment is in close proximity to either the 5S or 45S rDNA at each time. In cases of cytological proximity between the 5S and 45S arrays, large protein complexes might be present and prevent the emergence of 5S-45S inter-chromosomal contacts in the scale captured by Hi-C technology. Furthermore, the enrichment of rDNA contacts with ribosomal protein coding genes is surprising but might help explain the association between rDNA CN and the expression of these genes [21]. It suggests a structural component to the regulatory effects of the rDNA and exerting potentially direct rDNA effects in the expression of these classes of genes. The data suggests that models that exclusively consider proximity to the rDNA
arrays/nucleolus as a repressive modifier of gene expression are overtly simplistic. Rather, the proximal association of genes with the rDNA arrays appears functionally motivated, as in the case of ribosomal genes whose proximity to the rDNA arrays might help facilitate coordinated Pol I, Pol II and Pol III responses. They might have partially evolved to mitigate the fitness costs of dosage imbalances among highly expressed RNA and protein components of the ribosome.

Methods

The 5S and 45S arrays

The human 5S rDNA along with flanking regions (chr1: 228,765,135-228,767,255) and the human 45S rDNA (GenBank reference number U13369.1, with modifications) were obtained as recently described [21, 28, 40, 41]. The 45S reference comprises the 18S, 5.8S and 28S rRNA encoding segments, external transcribed sequences (ETS) and internal transcribed segments 1 and 2 (ITS1 and ITS2), as well as a ~32 Kb non-coding intergenic spacer (IGS). Both 5S and 45S segments contain repetitive elements, such as Alu and Line1; all analysis carried out in this study used 5S and 45S sequences masked for these repeats.

Hi-C data sets

Raw Hi-C reads for LCLs and erythroleukemia K562 (K562) cells were downloaded from the Gene Expression Omnibus (GEO) repository with accession number GSE63525 [42]. For LCL and K562, Hi-C experiments were done with three restriction enzymes (MboI, HindIII, and Ncol; 6,017,877,658 reads in total for LCL) and one restriction enzyme (MboI; 1,366,228,845 reads in total for K562), respectively. All experiments with more than 1 technical replicate were included. In addition, raw Hi-C reads for 5 cell types were obtained from GEO data with SRA Study number SRP033089 [43]. The 5 cell types comprised the H1 embryonic stem (ES) cells and 4 differentiated cell-types derived from H1 [Mesendoderm (ME) cells, Mesenchymal stem (MS) cells, Neuronal Progenitor (NP) cells, trophoblast-like (TB) cells] [43, 44]. The number of reads studied and recovery rates for 5S and 45S informative reads was summarized in Table 1.

Data preparation and mapping
All data were downloaded in SRA format and converted into FASTQ files by the NCBI SRA Toolkit’s command (fastq-dump). FASTQ files were quality and adapter trimmed with Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The trimming criteria required minimal quality score (> 20) and length (>50 bp). Next, we identified Hi-C reads that mapped to the 5S rDNA array or the 45S rDNA array. In this step, both forward and reverse reads were mapped independently to the 5S rDNA and 45S rDNA using Bowtie2 [45]. We used unpaired mapping with ‘very-sensitive’ mode (combinations of parameters: -D 20 -R 3 -N 0 -L 20 -i S, 1, 0.50). The mapping results were sorted and converted into binary format using SAMtools [46] and bed format using BEDTools [47]. We then extracted reads that mapped to the rDNA array and mapped the opposite end to repeat libraries. Reads for which one end mapped to repeats library were excluded. Finally, in order to identify potential confounders due to rDNA pseudogenes, both rDNA references were blasted against the human genome separately. Putative pseudogenes were identified as significant hits (E-value <1 × 10⁻⁴) using BLASTN [48, 49]. A segment of 1 MB was excluded from the analysis if an rDNA blast hit is identified within it. Similarly, a gene was removed from analysis if an rDNA blast hit is identified within its boundaries.

_Detecting genome-wide physical Interaction with rDNA loci at 1MB resolution._
To identify spatial variation in genomic contact density along the chromosomes we segmented the human genome GRCh37/hg19 assembly into 3,173 bins of 1MB using BEDTools [47]. Bins with rDNA pseudogenes were excluded. Contact densities were summarized for each bin for each of 5 cell types and 2 cell lines. We calculated the number of Contacts Per Million reads (CPM) to normalize the data and control for different number of reads in each of the seven conditions. This placed all the data in a comparable scale, to enable visualization of contact density along the human genome using heat maps in the 'gplots' R package [50].

_Analyses of rDNA-gene contacts_
The term of “rDNA-gene contact” refers to reads with one end mapped to rDNA arrays and the other end mapped between the first and the last exon of an annotated gene in the human genome. We extracted coordinates of these reads using BEDTools [47] and
the Gene Transfer Format (GTF) file: Homo_sapiens.GRCh37.75.gtf from the Ensembl database. GC content and length were also computed for each gene. To normalize contact densities in genes of different length, we computed the number of contacts per gene length (Contacts reads per Gene, CPG). The web based tool DAVID v6.8 [51] was used to investigate gene ontology enrichments for the top 5% of genes with greater CPGs for 5S rDNA or 45S rDNA genes. This corresponds to 494 out of 9864 genes for 5S-gene contacts, and 480 out of 9595 genes for 45S-gene contacts. The “one proportion” test [52] was also applied to address whether the number of mapped reads per base pair within a gene is significantly different from the genome wide average.

**Statistical analysis for differential contact density among cell types**

We modeled differential contact density per 1MB and per gene using the edgeR package and statistical approaches adapted from RNA-seq analysis [53, 54]. Raw counts for physical contacts with rDNA loci within each bin along the human genome are modeled using generalized linear models (likelihood ratio tests) implemented in the edgeR package [53, 54]. These approaches were recently been used to detect differential interaction density (DIs) in Hi-C data [19, 55, 56]. The models identified statistically significant differences among cell lines/types in rDNA contacts density per MB and within genes. The Benjamini-Hochberg method was used for multiple testing correction [57], and statistical significance was denoted by P < 0.05. We applied the method separately to ascertain significant differences between LCL and K562 cell lines, and to ascertain significant between five ES derived cell types. Spearman rank correlation was used to measure the degree of association among variables.

**Functional annotation of rDNA contacts**

We cross-referenced the rDNA contact map with several sources of functional annotation. First, Hi-C studies proposed the partition of the genome into A and B compartments that are widely interpreted as open and closed chromatin, respectively [58]. A/B coordinates were downloaded for LCL cells and 12 cancer types [58]. Second, coordinates of 15 functional regions identified in hESC using ChromHMM [59] were downloaded. Third, information on replication timing along the genome was downloaded from the Replication Domain Database (www.replicationdomain.org).
Finally, CTCF binding coordinates were obtained from the CTCFBSDB database [60]. For all annotations we extracted the coordinates for all the segments and addressed the density of rDNA contacts. BEDTools was used to assess the number of mapped reads that overlapped with each annotated segment for each dataset. The percentage of mapped reads was calculated by dividing the number of reads mapped to the segment by the total number of mapped reads. The genome wide average read per base pair was used to compute the expected number of reads in the functional segment. Statistical significance was assessed with Chi-square tests. In addition, we applied the “one proportion” statistical test [52] to address whether the numbers of mapped reads per base pair within a functional segment (e.g., CTCF binding) is significantly different from the genome-wide average per nucleotide contact rate.

Acknowledgements

We thank members of the Lemos lab for discussions and numerous investigators who contributed Hi-C datasets used here. The computations in this paper were run on the Odyssey cluster supported by the Research Computing Group in the FAS Division of Science at Harvard University.
References

1. Warner JR. 1999 The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* **24**(11): p. 437-440.
2. Grummt I. 2003 Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. *Genes Dev* **17**(14): p. 1691-1702.
3. Pederson T. 2011 The nucleolus. *Cold Spring Harb Perspect Biol* **3**(3).
4. Woolford JL, Jr., Baserga SJ. 2013 Ribosome biogenesis in the yeast *Saccharomyces cerevisiae*. *Genetics* **195**(3): p. 643-681.
5. Henderson A, Warburton D, Atwood K. 1972 Location of ribosomal DNA in the human chromosome complement. *Proceedings of the National Academy of Sciences* **69**(11): p. 3394-3398.
6. Henderson A, Warburton D, Atwood K. 1973 Ribosomal DNA connectives between human acrocentric chromosomes.
7. Wicke S, Costa A, Muñoz J, Quandt D. 2011 Restless 5S: The re-arrangement (s) and evolution of the nuclear ribosomal DNA in land plants. *Molecular phylogenetics and evolution* **61**(2): p. 321-332.
8. Moss T, Langlois F, Gagnon-Kugler T, Stefanovsky V. 2007 A housekeeper with power of attorney: the rRNA genes in ribosome biogenesis. *Cell Mol Life Sci* **64**(1): p. 29-49.
9. Fedoriw AM, Starmer J, Yee D, Magnuson T. 2012 Nucleolar association and transcriptional inhibition through 5S rDNA in mammals. *PLoS Genet* **8**(1): p. e1002468.
10. Smith CL, et al. 2014 A separable domain of the p150 subunit of human chromatin assembly factor-1 promotes protein and chromosome associations with nucleoli. *Mol Biol Cell* **25**(18): p. 2866-2881.
11. Kaplan FS, et al. 1993 The topographic organization of repetitive DNA in the human nucleolus. *Genomics* **15**(1): p. 123-132.
12. Dekker J, Rippe K, Dekker M, Kleckner N. 2002 Capturing chromosome conformation. *science* **295**(5558): p. 1306-1311.
13. van Berkum NL, et al. 2010 Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp* (39).
14. Lieberman-Aiden E, et al. 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950): p. 289-293.
15. Nora EP, et al. 2012 Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**(7398): p. 381-385.
16. Dixon JR, et al. 2012 Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**(7398): p. 376-380.
17. Nemeth A, et al. 2010 Initial genomics of the human nucleolus. *PLoS Genet* **6**(3): p. e1000889.
18. Nemeth A, Langst G. 2011 Genome organization in and around the nucleolus. *Trends Genet* **27**(4): p. 149-156.
19. Matheson TD, Kaufman PD. 2016 Grabbing the genome by the NADs. *Chromosoma* **125**(3): p. 361-371.
20. van de Nobelen S, et al. 2010 CTCF regulates the local epigenetic state of ribosomal DNA repeats. *Epigenetics & chromatin* **3**(1): p. 19.
21. Gibbons JG, Branco AT, Yu S, Lemos B. 2014 Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans. *Nature communications* **5**.

22. Bell SP, Learned RM, Jantzen H-M, Tjian R. 1988 Functional cooperativity between transcription factors UBF1 and SL1 mediates human ribosomal RNA synthesis. *Science* **241**(4870): p. 1192-1197.

23. Li P, et al. 2013 ABH2 couples regulation of ribosomal DNA transcription with DNA alkylation repair. *Cell reports* **4**(4): p. 817-829.

24. Paredes S, Branco AT, Hartl DL, Maggert KA, Lemos B. 2011 Ribosomal DNA deletions modulate genome-wide gene expression: "rDNA-sensitive" genes and natural variation. *PLoS Genet* **7**(4): p. e1001376.

25. Magalhaes PJ, Andreu AL, Schon EA. 1998 Evidence for the presence of 5S rRNA in mammalian mitochondria. *Mol Biol Cell* **9**(9): p. 2375-2382.

26. Smirnov A, Entelis N, Martin RP, Tarassov I. 2011 Biological significance of 5S rRNA import into human mitochondria: role of ribosomal protein MRP-L18. *Genes Dev* **25**(12): p. 1289-1305.

27. Cahyani I, Cridge AG, Engelke DR, Ganley AR, O’Sullivan JM. 2015 A sequence-specific interaction between the Saccharomyces cerevisiae rRNA gene repeats and a locus encoding an RNA polymerase I subunit affects ribosomal DNA stability. *Mol Cell Biol* **35**(3): p. 544-554.

28. Gibbons JG, Branco AT, Godinho SA, Yu S, Lemos B. 2015 Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *Proceedings of the National Academy of Sciences*: p. 201416878.

29. Yu S, 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 biology and evolution* **8**(11): p. 3545-3558.

30. Raška I, Shaw PJ, Cmarko D. 2006 New insights into nucleolar architecture and activity. *International review of cytology* **255**: p. 177-235.

31. Bouteille M, Kalifat SR, Delarue J. 1967 Ultrastructural variations of nuclear bodies in human diseases. *J Ultrastruct Res* **19**(5): p. 474-486.

32. Goessens G. 1984 Nucleolar structure. *Int Rev Cytol* **87**: p. 107-158.

33. Fischer D, Weisenberger D, Scheer U. 1991 Assigning functions to nucleolar structures. *Chromosoma* **101**(3): p. 133-140.

34. Scheer U, Thiry M, Goessens G. 1993 Structure, function and assembly of the nucleolus. *Trends Cell Biol* **3**(7): p. 236-241.

35. Petes TD. 1979 Yeast ribosomal DNA genes are located on chromosome XII. *Proceedings of the National Academy of Sciences* **76**(1): p. 410-414.

36. Sone T, et al. 1999 Bryophyte 5S was inserted into 45S rDNA repeat units after the divergence from higher land plants. *Plant Molecular Biology* **41**: p. 679-685.

37. Ganley AR, Kobayashi T. 2007 Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research* **17**(2): p. 184-191.

38. Liu Y, Forrest LL, Bainard JD, Budke JM, Goffinet B. 2013 Organellar genome, nuclear ribosomal DNA repeat unit, and microsatellites isolated from a small-scale of 454 GS FLX sequencing on two mosses. *Mol Phylogenet Evol* **66**(3): p. 1089-1094.

39. Garcia S, Galvez F, Gras A, Kovarik A, Garnatje T. 2014 Plant rDNA database: update and new features. *Database (Oxford)* **2014**.
40. Gonzalez IL, Sylvester JE. 1995 Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. Genomics 27(2): p. 320-328.
41. Zentner GE, Saiakhova A, Manaenkov P, Adams MD, Scacheri PC. 2011 Integrative genomic analysis of human ribosomal DNA. Nucleic acids research : p. gkq1326.
42. Rao SS, et al. 2014 A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7): p. 1665-1680.
43. Dixon JR, et al. 2015 Chromatin architecture reorganization during stem cell differentiation. Nature 518(7539): p. 331-336.
44. Xie W, et al. 2013 Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell 153(5): p. 1134-1148.
45. Li H, et al. 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25(16): p. 2078-2079.
46. Quinlan AR, Hall IM. 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6): p. 841-842.
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool. Journal of molecular biology 215(3): p. 403-410.
48. Altschul SF, et al. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25(17): p. 3389-3402.
49. Warnes GR, et al. 2009 gplots: Various R programming tools for plotting data. R package version 2(4): p. 1.
50. Huang DW, Sherman BT, Lempicki RA. 2009 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4(1): p. 44-57.
51. Altman DG, Practical statistics for medical research1990: CRC press.
52. McCarthy DG, Chen Y, Smyth GK. 2012 Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic acids research 40(10): p. 4288-4297.
53. Robinson MD, McCarthy DJ, Smyth GK. 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1): p. 139-140.
54. Paulsen J, et al. 2014 HiBrowse: multi-purpose statistical analysis of genomewide chromatin 3D organization. Bioinformatics 30(11): p. 1620-1622.
55. Lun AT, Smyth GK. 2015 diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC bioinformatics 16(1): p. 258.
56. Benjamini Y, Hochberg Y. 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological): p. 289-300.
57. Fortin JP, Hansen KD. 2015 Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. Genome Biol 16: p. 180.
58. Ernst J, Kellis M. 2012 ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 9(3): p. 215-216.
59. Ziebarth JD, Bhattacharya A, Cui Y. 2012 CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. Nucleic acids research 41(D1): p. D188-D194.

Figure Legends
**Figure 1.** Extensive variation in ribosomal DNA (rDNA) contacts across protein-coding genes and 1MB genomic segments. (A) Distribution of the number of 45S contacts (log scale) in 1MB genomic segments for LCL (green), K562 (purple) and ESC (red). (B) Distribution of the number of 45S rDNA-gene contacts per base pair (log scale) in LCL and K562. (C) Distribution of the number of 5S contacts (log scale) in 1MB segments for LCL and K562. (D) Distribution of the number of 5S rDNA-gene contacts per base pair (log scale) in LCL and K562. Only segments and genes without rDNA pseudogenes are included.

**Figure 2.** Variation in 45S rDNA contact density for 1MB segments across seven datasets (five cell types and two cell lines). Heatmaps display data for the whole genome (Left panel), a region of Chromosome 9 (Right upper panel), and a region of chromosome 16 (Right lower panel). Statistically significant differences in rDNA contact density were assessed with edgeR and are described in the text.

**Figure 3.** Closed, repressed, late replicating chromatin segments are preferentially associated with the rDNA array. Shown is the number of observed (black) and expected (white) rDNA contacts with each functional annotation. Expected numbers are calculated with the genomewide per nucleotide contact rate. Shown is data for (A) A/B compartments in LCLs as computed in [58 106]; (B) three selected annotations from 15-label genomic segmentation of hESC using ChromHMM [59]; (C) the 5% most slow replicating segments of the genome as recorded in the Replication Domain Database.

**Figure 4.** Distribution of 45S-gene contacts across genes belonging to selected functional categories. (A) Genes associated with rDNA copy number (gRCN, green color); (B) Genes localized to the nucleolus (green color); (C) Genes encoding protein components of the mitochondrial ribosome (mRPGs, green color); (D) Genes encoding protein components of the cytosolic ribosome (cRPGs, green color). The red colored density display the variation in rDNA-gene contacts across all genes. Data for LCLs. Statistical significance assessed with Kolmogorov-Smirnov tests.
**Tables and Table Legends**

**Table 1.** Summary table for all of the analyzed Hi-C reads from two cell lines and five cell types. Shown are the total number of reads analyzed per dataset and the number of reads remaining after the quality control (QC) steps. Shown are the counts of reads mapped to rDNA arrays and the percentage of reads that mapped to the rDNA arrays after QC. Data is shown for both 5S and 45S rDNA arrays.

| Datasets | Number of reads | Number of reads after trimming and QC | % kept after QC | 45s reads | % of 45s reads | 5s reads | % of 5s reads |
|----------|----------------|--------------------------------------|----------------|-----------|---------------|----------|--------------|
| ESC      | 1,965,460,504  | 293,548,127                          | 14.94          | 300,903   | 0.10          | 1,627    | 0.0006       |
| Mesod    | 1,453,457,218  | 282,703,604                          | 19.45          | 160,118   | 0.06          | 1,217    | 0.0004       |
| Mesec    | 1,946,805,860  | 375,059,953                          | 19.27          | 167,365   | 0.04          | 765      | 0.0002       |
| Neuro    | 1,478,607,926  | 284,198,790                          | 19.22          | 287,756   | 0.10          | 1,449    | 0.0005       |
| Ectod    | 936,917,416    | 181,739,870                          | 19.40          | 179,897   | 0.10          | 2,309    | 0.0013       |
| K562     | 1,366,228,845  | 903,837,936                          | 66.16          | 1,698,063 | 0.19          | 47,691   | 0.0053       |
| LCL      | 6,017,877,658  | 5,356,990,189                        | 89.02          | 13,528,436| 0.25          | 105,147  | 0.0020       |
| SUM      | 15,165,355,427 | 7,678,078,469                        | 51.63          | 16,322,538| 0.21          | 160,205  | 0.0020       |
Table 2. Gene ontology enrichments for genes in the top 5% higher number of 45S rDNA contacts in LCLs. Shown are the number of genes represented in each GO category. Adjusted P-values were obtained after Benjamin-Hochberg correction.

| GO:0005840~ribosome | 18 | 0.000001 | 0.000 |
|----------------------|----|-----------|-------|
| GO:0022625~cytosolic large ribosomal subunit | 10 | 0.000050 | 0.010 |
| GO:0005739~mitochondrion | 62 | 0.000100 | 0.019 |
| GO:0005829~cytosol | 135 | 0.000300 | 0.028 |
| GO:0022627~cytosolic small ribosomal subunit | 7 | 0.000600 | 0.046 |
| GO:0005737~cytoplasm | 182 | 0.000700 | 0.047 |
| GO:0005681~spliceosomal complex | 11 | 0.002000 | 0.106 |
| GO:0015935~small ribosomal subunit | 5 | 0.003000 | 0.149 |
| GO:0005730~nucleolus | 39 | 0.006000 | 0.251 |

Supplementary Figures and Tables
**S1 Figure.** Comparison of results for positions of 45S rDNA contacts with Hi-C and nucleolus-associated chromatin domains (NADs). (A) part of chromosome 2q (chr2:90,052,599-169,971,888). (B) part of chromosome 15 q arm (chr15:19,990,398-63,840,300). Red boxes represent NADs [17], blue bars represent satellite repeats, orange box represent centromeres and blue horizontal bars represent a part of chromosome 2 or 15. Sites of 45S Hi-C contacts recovered with LCL and K562 are shown.

**S2 Figure.** (A) Plots of expected vs. observed number of hits per chromosome for LCL cell types. (B) Plots of expected vs. observed number of hits per chromosome for K562 cell types. (C) Plots of expected vs. observed number of hits per chromosome for ESC related cell types.

**S3 Figure.** Manhattan plots with P-values for differences in 45S rDNA-gene contacts in contrasts between two cell lines (LCL and K562) and among five cell types. The X-axis shows chromosomal position. The Y-axis shows −log10 (P). The blue line is the significance level (3 × 10⁻³).

**S4 Figure.** Volcano plot showing differences in 45s rDNA-gene contacts for LCL vs K562. Green dots represent larger than two fold change (two vertical lines). Red dots represent larger than two fold change and P-value smaller than 0.0001 (horizontal line).

**S1 Table.** Summary table of Hi-C reads mapping to the 5S rDNA repeat unit (2121bp) in each dataset. Shown are the number of reads for which both ends map to the 5S repeat unit and the number of reads for which one end maps to the 5S rDNA repeat unit and the other maps to the rest of the genome [whole genome (WG)]. Both the 5S rDNA unit and the whole genome were masked for repetitive elements.

**S2 Table.** Summary table of Hi-C reads mapping to the 45S rDNA repeat unit (45,337 bp) in each dataset. Shown are the number of reads for which both ends map to the 45S rDNA repeat unit and the number of reads for which one end maps to the 45S rDNA
repeat unit and the other maps to the rest of the genome [whole genome (WG)]. Both the 45S rDNA unit and the whole genome were masked for repetitive elements.

**S3 Table.** Number of 1MB segments displaying differential density of interactions with the 45S rDNA array in LCL and K562. Statistically significant segments ascertained with the EdgeR package. The number of segments with significant differences is shown for each chromosome.

**S4 Table.** Number of 1MB segments displaying differential density of interactions with the 45S rDNA array among five ESC-related cell types. Statistically significant segments ascertained with the EdgeR package. The number of segments with significant differences is shown for each chromosome.

**S5 Table.** Average number of contacts per nucleotide (CPN) obtained with three subsets of contacts. CPN obtained from (i) all rDNA-genome contacts (without controlling for pseudogenes), (ii) rDNA-genome contacts retrieved from 1MB bins without rDNA pseudogenes, and (iii) rDNA-gene contacts retrieved from genes without rDNA pseudogenes in them. Genome-wide contacts were divided by the genome size or the total size of the bins without pseudogenes. rDNA-gene contacts were divided by the total length of the genome with genic sequences (protein-coding genes only). Shown are estimates for both 5S and 45S rDNA arrays.

**S6 Table.** Summary table of statistically significant gene ontology enrichments for the top 5% of genes contacted with the 5S rDNA in the K562 and LCL cell types.