Diversity of ribosomes at the level of rRNA variation associated with human health and disease

Daphna Rothschild[1]*, Teodorus Theo Susanto[1]*, Xin Sui[3,4]*, Jeffrey P. Spence[1], Ramya Rangan[5], Naomi R. Genuth[1,2], Nasa Sinnott-Armstrong[1], Xiao Wang[3,4], Jonathan K. Pritchard[1,2]§, Maria Barna[1]§

[1] Department of Genetics, Stanford University, Stanford, CA, 94305, USA
[2] Department of Biology, Stanford University, Stanford, CA, 94305, USA
[3] Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA.
[4] Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA.
[5] Biophysics Program, Stanford University, Stanford, CA, 94305, USA
* Co-first author
§ Co-senior author. Email: pritch@stanford.edu (JKP), mbarna@stanford.edu (MB).

Summary

Ribosomal DNA and RNA (rDNA and rRNA) sequences are usually discarded from sequencing analyses. But with hundreds of copies of rDNA genes it is unknown whether they possess sequence variations that form different types of ribosomes that affect human physiology and disease. Here, we developed an algorithm for variant-calling between paralog genes (termed RGA) and compared rDNA variations found in short- and long-read sequencing data from the 1,000 Genomes Project (1KGP) and Genome In A Bottle (GIAB). We additionally developed a novel protocol for long-read sequencing full-length rRNA (RIBO-RT) from actively translating ribosomes. Our analyses identified hundreds of rDNA variants, most of which, surprisingly, are short insertion-deletions (indels) and dozens of highly abundant rRNA variants that are incorporated into translationally active ribosomes. To visualize variant ribosomes at the single cell level, we developed an *in-situ* rRNA sequencing method (SWITCH-seq) which revealed that variants are co-expressed within individual cells. Strikingly, by analyzing rDNA, we found that variants assemble into distinct ribosome subtypes. We discovered that these subtypes acquire different rRNA structures by successfully employing dimethyl sulfate (DMS) probing of full length rRNA. With this atlas we investigated rRNA variation changes across human tissues and cancer types. This revealed tissue-specific rRNA subtype expression in

endoderm/ectoderm-derived tissues. In cancer, low abundant rRNA variants can become highly expressed, which suggests the presence of cancer-specific ribosomes. Together, this study identifies and comprehensively characterizes the diversity of ribosomes at the level of rRNA variants which is dominated by indel variants, their chromosomal location and unique structure as well as the association of ribosome variation with tissue-specific biology and cancer.

Introduction

The ribosome is a complex, ancient machine responsible for all protein synthesis, with a core ribosomal RNA (rRNA) structure that is conserved across all kingdoms of life. The primary transcript of the rRNA genes is the large 45S pre-rRNA which contains the 18S, 5.8S, and 28S rRNAs as well as transcribed spacer regions. In humans, rRNA genes are present in hundreds of ribosomal DNA (rDNA) copies in tandem repeats that are spread across multiple chromosomal loci [1]. These high rDNA copy numbers are thought to be necessary to produce millions of ribosomes in each cell. Nevertheless, these hundreds of rDNA copies allow for sequence variation between copies, as was first noted in mice and humans almost 50 years ago [2].

It remains an outstanding challenge to understand whether ribosomes are different at the level of rRNA and how many ribosome subtypes may exist. For the last several decades we have therefore had a limited knowledge of fundamental differences in the translational machinery and limited insight beyond the textbook view of ribosome composition. A very significant study showed that rRNA sequence variations between different rDNA loci lead to novel rRNA functions, in *E.coli*, where sequence variations between the bacterial rDNA operons support growth under stress conditions [3]. In particular, they found that naturally occurring rRNA sequence variation can modulate bacterial ribosome function, central aspects of gene expression regulation, and cellular physiology. This inspiring work raises the possibility that this is general and could be the case in humans. In humans, by examining short-read sequencing data of the 1,000 Genomes Project (1KGP) [4], two previous studies discovered hundreds of positions in rDNA bearing sequence variants [5,6]. A major challenge faced by these former studies arises from limitations in short-read sequencing where there is the depletion of GC-rich sequences. Specifically certain regions within rDNA possess >80% GC content on average [6–8]. This results in two types of problems in variant calling: (1) false negatives, caused by the inability to identify variants in regions with poor sequence coverage, and (2) false positives,

caused by PCR errors in low-coverage regions. Moreover, these GC rich regions are highly repetitive which make short-read variant discovery tools inaccurate [9]. As a result, the latter two previous studies [5,6] reported contradictory results likely because of these major caveats. Nevertheless, these studies did raise interest in investigating rRNA sequence variation, inspiring future studies. Interestingly, the copy number of the rDNA loci was also shown to be important for gene expression and cellular homeostasis [10–13]. Moreover, rDNA copy number was associated with age and in cancer [14–16]. In cancer, both rDNA copy number gain and loss have been reported in multiple studies [17–21]. This calls for further investigation into whether rDNA copy number changes are coupled with sequence variations and if such changes affect human health.

Recently, long-read sequencing enabled the successful complete assembly of a human genome by the Telomere-2-Telomere (T2T) consortium, including positioning of rDNA copies in the five acrocentric chromosomes [22]. Moreover in the mouse and Arabidopsis plant genomes, rDNA variants were grouped into haplotypes and a few rRNA variants were found expressed in tissues using short-reads [5,23,24]. Yet in order to find low-frequency variations between full length rDNA paralog copies, new computational method development is necessary. Long-reads offer the possibility for distinguishing between paralog genes. However, existing common methods for long-read variant calling, such as DeepVariant [25] and Clair [26], are primarily designed for detecting variants in single-copy regions. For paralog genes where low-frequency variants exist between copies accurate variant calling is lacking [9]. Moreover, it remains an open question if rRNA haplotypes are expressed from the human genome, what their abundances are, and if such variability is linked to human physiology. There too, long-read sequencing and analysis of full-length rRNA is necessary This highlights the need for new approaches to comprehensively characterize human ribosome diversity.

To address this need in the field, we devised an efficient novel computational algorithm to detect all variations between paralogs (termed RGA). Applied to the long-read 1,000 Genomes Project dataset, we discovered hundreds of rDNA sequence variations enriched with previously undiscovered insertion-deletions. We further developed a novel methodology to perform long-read sequencing on rRNA in actively translating ribosomes to identify variants (RIBO-RT). Using this method, we discovered that ribosomes have different subtypes with rRNA variants that are genomically encoded by rDNA clustered on distinct chromosomes. Additionally, using an *in-situ* rRNA sequencing platform that we developed (SWITCH-seq), we discovered that variants belonging to different rRNA subtypes

are co-expressed in single cells. We then used structure probing coupled with long-read sequencing to find that 28S subtypes have different rRNA structures. Lastly, we found that these subtypes are differentially expressed in human tissues, and low abundance variants are elevated in certain cancers. Together, these results suggest that ribosomes with unique sequence variation may be used to modulate different cellular programs underlying human physiology and disease.

Results

Indels are the main variants of the human rDNA loci

How rDNA variation shapes the presence of unique ribosomes in the cell remains an important open question. Previous studies that analyzed the 1KGP dataset for discovery of rDNA variants reported discordant results. Parks et al. [5] reported hundreds of variants in both the 18S and the 28S, yet 75% of variants were not made publicly available, making a comparison to this dataset problematic. Nonetheless, Fan et al. [6] reported notable differences from Parks et al. by suggesting that the 18S has low variation, and also reporting many fewer variant positions in the 28S. Moreover, Parks et al. reported only 2.7% of variants being indels while Fan et al. reported 19.2% indels. Here, considering the limitations of short-reads in rDNA variant discovery and their inability to distinguish between rDNA paralogues we decided to re-evaluate the variants in the human rDNA genes.

Until recently, the 1KGP dataset included only short-read genome sequencing [27]. Yet as of 2022, the 1KGP includes PacBio's HiFi long-read sequencing for 30 individuals from diverse ancestral origins, which could serve as a better method for accurately calling variants. Here, we compared the rDNA variants captured by short- and long-reads from the same individuals and addressed the discrepancies between previous studies (**Figure 1A,B, Table S1**).

When analyzing the short-read data, we followed the pre-processing steps as performed in previous studies of marking duplicate reads which are suspected as PCR artifacts. This step discarded 97% of reads. However, it is unknown if duplicate reads are PCR biases given the high rDNA paralog copy numbers which highlights the limitation of short-read sequencing for rDNA variant discovery. Next, in order to call variants including rare variants which are not expected to follow germline variant frequencies in high paralogues rDNA copy numbers, we tested two common somatic variant calling

methods for short-reads: LoFreq* [28] which was used by Parks et al. and Mutect2 [29]. Specifically, Mutect2 was chosen instead of the germline variant caller used by Fan et al., because germline variant callers will not detect rare variants found between paralogues. Using the LoFreq* method which is known to be sensitive [28], we found 1,582 positions with variants compared to 861 positions with variants with Mutect2 (**Table S2-3**). Notably, both methods detected 23% indels, which is on par with the indel percentage reported by Fan et al.. Given the difference in the proportion of indel frequencies between the previous two studies, we compared variant quality scores of SNVs and indels (**Table S3** - Mutect2 FDR-corrected log 10 likelihood ratio score of variant existence). Here we found that indel variants were enriched with high confidence P-values (**Figure S1**, P-value < 10^{-15} comparing SNVs and indels likelihood ratio scores using Kolmogorov-Smirnov test for goodness of fit). Notably, tandem repeats and GC-rich sequences in the human genome were shown to be prone to chromosomal breakage and were found enriched in indels [30]. Therefore, the 23% indel frequency derived solely from short-read data may underrepresent the true indel frequency in these samples. To test this, we next examined the HiFi long-read data from the same samples.

LoFreq* and Mutect2 did not work on the long-read sequencing data. In order to identify all positions with sequence variants, we developed a new computational method for acurate variant calling between paralogs which we term Reference Gap Alignment (RGA, see **Methods**). We align all reads against a common reference, and report all variants at a given position with respect to this reference (**Figure 1C, S2, Methods**). Our method reports at every position with respect to the 18S/28S reference if that position has a variant and calls its identity. The only parameters in our method are the standard pairwise alignment parameters: a mismatch penalty, a gap opening penalty, and a gap extension penalty (see **Methods** for more details). When benchmarking the global sequence alignment parameters, these resulted in similar indel proportions (**Table S4**). In agreement with previous studies [5,6], we found that rDNA is highly variable, yet using our method, we discovered that the vast majority of variants are short indels and not SNVs in all 30 samples. Specifically, when examining each reference position individually we found that on average 95% of variants are GC-rich indels (**Table S5**).

Since this indel proportion found with long-reads is markedly different from the results obtained with short-reads (**Figure 1B)**, we cross-validated our reported variants in three ways. As first indel validation, we decided to compare our results from our primary long-read sequencing technology,

namely HiFi, with an alternative, Oxford Nanopore (ONT). Importantly, ONT is much more error prone compared to HiFi, with an estimated 13% error rate in ONT compared to 0.1% error rate in HiFi [31–33]. Since 1KGP long-read sequencing was only performed on HiFi, we tested this using the Genome In A Bottle (GIAB) dataset, which consists of two trio families, where both HiFi and ONT were performed on the same samples. Here, when examining the HiFi dataset of GIAB, in agreement with the 1KGP HiFi results, we discovered that 96% of variants are indels (**Table S6**, **Methods**). When using the ONT dataset as a validation dataset, 81% of variants found in HiFi were replicated in the ONT dataset (**Table S7, Methods**). Notably, the variants that were not identified by ONT consisted of insertions and SNVs but not deletions (**Table S8**). Additionally, after retaining variants found at frequencies above the ONT error rate, 87% of found variants were indels (**Table S7** filtering variants with allele frequency smaller than 0.13)**.** As another method to cross-validate our RGA variant caller, we examined variant frequencies in the family members of the GIAB trio dataset. We compared variant frequencies for variants that appeared only in the child, in a single parent, or in both parents. Here we expected that high abundant variants would be inherited and indeed all variants with frequency greater than 2% were found in at least one of the parents. This held true for both for SNVs and indels (**Figure S3**).

Finally, we tested if short-reads contain reads with indel variants that are not identified by short-read variant callers. To do this, we used the indels identified with long-reads by the RGA method as a reference of existing sequence variations and mapped the short-reads from the 1KGP to this reference (**Methods**). This directly tests if the variants found in the long-reads are also detected in the short-reads. Notably, mapping variants is different from de-novo variant discovery since here we count short-reads which perfectly match the indels as opposed to variant calling tools which are reference free and need to appear in high enough frequency to be discovered. Surprisingly, despite estimated low coverage of the rDNA loci we could detect 928 nucleotide variants which included 896 indels and 32 SNVs (**Methods, Table S9**). Previous variant callers which have analyzed the 1KGP were not able to identify most indels, likely because they are too low in abundance (LoFreq*, Mutect2 and DNAseq pipeline Sentieon, Release 201911). Our results show that short-reads can be mapped to indels given a reference of variants identified with long-read sequencing.

Our results highlight that previous studies which used short-read sequencing missed the majority of the variants found in the rDNA loci. Most of our identified rDNA variants were found at GC-rich

regions which are depleted from short-read sequencing. We conclude that indels are the main variants in rDNA both between and within individuals. These findings also highlight the need to curate a reference of rDNA variants and the importance of our accurate variant calling method together with long-read sequencing in confidently assigning rDNA variants.

An atlas of 18S and 28S human rDNA variants validated in rRNA of translating ribosomes and single cell microscopy

It is unknown if the rDNA copies with sequence variants found in the human genome are transcribed and are found in functionally translating ribosomes. With no human reference of different rRNA subtypes, studies performing RNA sequencing completely ignore rRNA variants which limits our understanding of the contribution of rRNA to human physiology and disease.

Sequencing of rRNA has been historically technically challenging [6,34]. Here, in addition to our computational RGA variant discovery method, by optimizing long-read sequencing, we have successfully developed an experimental method for full length rRNA sequencing of 18S and 28S from translating ribosomes which we named RIBO-RT. In order to extract rRNA from translationally active ribosomes, we first employed sucrose gradient fractionation wherein ribosomes can be separated into free ribosomal subunits and translationally active ribosomes, which contain one or more ribosomes bound to the same mRNA. We extracted RNA from translating ribosomes-containing fractions (**Figure S4**)**,** performed reverse transcription in denaturing conditions, and then sequenced complete 18S and 28S rRNA by HiFi long-read sequencing (**Figure 1D**, **Methods**). We selected a human embryonic stem cell line (H7-hESC) and using long-read sequencing, we sequenced the 18S and 28S from both its rDNA and rRNA (**Figure 1D**). Here, we obtained 58,495 sequences of the 18S and 14,430 sequences of the 28S rRNAs from translating ribosomes (**Methods**). With this approach and our variant discovery method, we were able to coherently characterize the 18S and 28S H7-hESC rRNA variants. Most importantly, since rRNA is known to be heavily modified [35], our strategy of matching rRNA to genomic rDNA from the same cell enables us to distinguish modifications or sequencing errors from true sequence variants belonging to different rDNA alleles.

In agreement with our 1KGP and GIAB rDNA results, we found that the H7-hESC rDNA is highly variable and is enriched with indels. Moreover, 96% and 84% of the hESC variants are also found in

the 1KGP genomes projects and GIAB datasets respectively. This is generally concordant with expected rates of replication based on these small sample sizes of the 1KGP and GIAB datasets. Moreover, we find high agreement in the frequency of variants between the hESC to other datasets (**Figure S5**). Additionally, rDNA variants are transcribed into functionally translating ribosomes as they are present in polysome fractions (**Figure S6**). Specifically, we found 270 positions with variants in the 18S and 858 positions with variants in the 28S, corresponding to about one variant every six rRNA positions (**Figure S6-7**). When comparing monosome to polysome fractions we observed high correlation in variant frequencies between fractions (**Figure S8, Table S10**). Additionally, we long read sequenced the 18S and 28S from translating ribosomes from an additional commonly used cell-line, K562 using the same extraction protocol as described in the methods (**Figure S9**). We found that 95% of the H7-hESC variants are also found in the tested human cell line (**Figure S10, Table S15**). Most variants (59%) are found in expansion segment (ES) regions (**Figure S11**, ES/non-ES regions are annotated). These regions vary in sequence both within and among different species, nearly doubling the eukaryotic rRNA sequence relative to that of prokaryotes [36]. ESs have recently been shown to bind ribosome associated proteins and transcripts, yet their functions remain poorly understood [36–40].

Most importantly, with accurate variant calling and full coverage of the underlying rDNA and transcribed rRNA, we can measure the frequency of each variant between the rDNA copies and rRNA expression levels. We distinguish possible modifications or sequencing errors from certain sequence variants belonging to different rDNA alleles by calling variants with similar frequencies measured in rDNA and rRNA as high-confidence alternate allele variants while those significantly deviating between their rDNA and rRNA frequencies as low-confidence alternate alleles (**Figure 1E, Table S11,** Fisher's Exact Test measuring association between rDNA and rRNA alternate allele frequencies - high-confidence alternate alleles are marked in red, low-confidence alternate alleles are marked in orange). Surprisingly, we did not identify any abundant high-confidence variants within the 18S. This suggests that rRNA variation is not tolerated within the small ribosome subunit. While most variants have low abundance, in the 28S we found 23 variants at 17 positions with frequencies above 10% in both rRNA and rDNA (**Figure 1E,F, Table S12, Figure S7** annotated positions, **S8C** minor allele frequency with dashed-like marking at 10% frequency). Notably, for the 28S high abundance variants, there is very good agreement between variant frequencies in rRNA and rDNA (**Figure 1E**, R=0.93 Pearson correlation, correlating all variants colored in red and orange).

Next, we focused on the 28S high abundance variants. For most positions, the RNA45S5 reference allele is the major allele found in the sequenced H7-hESC line (**Figure 1F**, reference allele in gray, **Extended Data 1-2 nucleotide atlas**). Yet some alternate alleles were more abundant than the 28S RNA45S5 reference alleles (**Figure 1F**, gray for the reference allele and red for variants). Notably, high abundant variants are only located in 4 ES regions (es7l, es15l, es27l and es39l) and one non-ES region, helix 28S:h11 (**Figure 1F**). Moreover, in the es7l, es15l and es27l regions we observed that variants can be grouped and characterized by indels of GGC in tandem-repeats. GGX tandem-repeats were recently suggested capable of forming G-quadruplex structures [41] while other works suggested that such repeats can form other higher order structures [42,43]. While the function of these ESs is largely unknown, a growing body of research supports various roles in translation regulation. For example, es27l has been shown to be important for control of translation fidelity and binding of ribosome-associated proteins for several processes, such as initiator methionine cleavage from the nascent polypeptide [44] or acetylation of nascent polypeptides [45]. Moreover, es39l interacts with the Signal Recognition Particle (SRP), which identifies the signal sequence on nascent polypeptides emerging from the translating ribosome [46]. Interestingly, the most abundant variant in the non-ES helices, a G to A substitution at position 60 in 28S:h11, is considered unique to humans. The alternative allele, A, is the reference allele for other mammals including chimpanzees [47,48].

For these aforementioned highly abundant variants, a strong correlation between the frequency of a variant's occurrence among rDNA copies and its expression levels in rRNA indicates both the authenticity of these sequence variants and their likely co-expression within individual cells. To explore this hypothesis, we developed a template-switching-based *in-situ* sequencing method (SWITCH-seq) to visualize variant ribosomes in individual HeLa cells (**Figure 2A**). This approach involved designing a reverse transcription primer to target constant non-variable regions downstream of the selected rRNA variant regions. Specifically, we selected regions for which we could design a primer for each of the ES regions with abundant variants (**Methods, Table S13**). The process of SWITCH-seq begins with performing reverse transcription on fixed HeLa cells, wherein a known sequence of choice is attached to the 3′ end of cDNA (template switching) (**Figure 2A**, **Methods**). This step incorporates the variant of interest into the cDNA, which is subsequently amplified into *in-situ* cDNA amplicons through enzymatic circularization and rolling circle amplification. These amplicons encapsulating the rRNA variants are then probed into a hydrogel network for cyclic imaging using a confocal microscope (**Methods**). We conducted multiple rounds of imaging which

capture two bases before the variants, and the consecutive base, where we expect to find the variants (**Figure 2B, S12**). As predicted, we successfully observed both the reference and alternate alleles (**Figure 2B,C, S12**). Furthermore, the frequencies of the reference and variant alleles corresponded with their frequencies in the H7-hESC samples (**Table S14**). We conclude that rRNA variants observed at high frequency in the H7-hESC rRNA and rDNA and in the rDNA across the 1KGP and GIAB samples form ribosomes that are co-expressed in individual cells that can be visualized at single cell resolution.

As an important resource for studying human rRNA variations, we create the first comprehensive atlas of all H7-ESC rRNA 18S and 28S rRNA variants at different resolutions from nucleotide variants to gene 28S haplotype-groups that we later describe as separate subtypes (**Figure S11, Table S17-18** for region annotation and atlas building is described in the **Methods**).

28S variants assemble to distinct ribosome subtypes

Since we successfully obtained full length 18S and 28S rRNA with variations, we address the outstanding question of whether variations lead to the formation of different ribosome subtypes. Here we focused on the 28S since 18S variants appeared at low frequency. For the 28S, we found high agreement between rRNA and rDNA variant frequency so we first asked which rDNA variants are co-located on the same 28S rDNA copy. To do this, we calculated the correlation coefficient (Pearson's r^2) between positions across all 28S H7-hESC rDNA sequences. This is analogous to measuring the linkage disequilibrium (LD) coefficient in population genetics, though across paralogous copies within a single genome rather than across individuals in a population. Notably, we found low global LD structure between highly abundant rDNA variants (**Figure 3A** showing LD for rDNA positions with found rRNA frequency > 10%), supporting recent findings indicating high rates of nonallelic gene conversion across the acrocentric chromosomes [49]. The highest LD ($r^2 > 0.2$) was found between the es27l to all other regions. Comparing different regions, we found LD between four regions: 28S:h11, es15l, es27l and es39l, where in each region we identified a position with higher linkage to the other three regions (**Figure 3A**, four positions are annotated, position 60 being 28S:h11, with higher linkage to other positions). By considering the variants at this subset of positions, we found a total of 21 different haplotypes both in rDNA and rRNA (**Figure S13**). For testing if haplotypes can be considered as different 28S subtype variants, we further analyzed two

independent long-read DNA datasets: (1) the fully assembled genome from the T2T with 219 rDNA copies with their chromosome location [22], and (2) the GIAB HiFi dataset [50]. In agreement with the H7-hESC results, 3 out of the 4 positions with higher linkage to other positions in the H7-hESC had higher LD in the GIAB dataset (**Figure 3B, S14**). Since these three variants are linked to variants at other positions, we define the haplotypes formed by positions 60, 3513 and 4913, belonging to regions 28S:h11, es27l and es39l respectively, as different 28S haplotypes (**Figure 3C**, **Extended Data 7 28S atlas**).

Previously it was shown that the rDNA array is composed of highly homogenized tandem clusters [51]. We therefore next asked whether different 28S haplotypes are spatially separated in the genome as different subtypes. For the H7-hESC we have 386 complete 28S rDNA sequences and in the GIAB dataset we randomly subsampled each GIAB sample to 386 complete 28S rDNA sequences. For these datasets, we do not know rDNA-chromosome positioning. Notably, by comparison of 28S rDNA sequence similarities, we detected distinct 28S sequence groups in both hESC and GIAB (**Figure 3D**, Principal Coordinate Analysis, PCoA, of Bray-Curtis dissimilarities between 28S sequences [52], **Methods**). Here, the different clusters in PCoA space match different 28S haplotypes. Specifically, we observed that 28S sequences of a given haplotype are more similar to one another in their entire sequence compared to 28S rRNAs of other haplotypes (**Figure 3D**). When plotting individual haplotypes, there is less observed structure in individual haplotypes as compared to the combined data (**Figure S15**). In the T2T assembly, rDNA copies have chromosome coordinates, which enables us to measure 28S subtype presence at the five acrocentric chromosomes. Remarkably, we discovered that 28S haplotypes are largely chromosome specific (**Figure 3E**). When analyzing the 1KGP, we find that all of the haplotypes found in high frequency in the H7-hESC, GIAB and the T2T CHM13 genome, are also present in the 1KGP genomes. However, when examining haplotype frequency changes, we find high standard deviation between their frequencies across individuals (**Table S18**). This variability limits our understanding of the 28S haplotypes, which may be caused by high rates of nonallelic gene conversion across rDNA copies. Taken together, our result support that 28S haplotypes are genomically separated and belong to different subtypes.

Ribosomes of different 28S subtypes have different structures.

We next asked if different 28S subtypes have different ribosome structures. Notably, the abundant variants in the hESC were found in four different ES regions which were never previously resolved by

CryoEM. Here, we treated the hESC sample with DMS, which covalently modifies the RNA at regions where the rRNA is accessible to allow for structure probing of the RNA (**Methods**). Using our RIBO-RT method for sequencing full length 28S with our RGA variant calling on DMS treated hESC cells, we obtained an accessibility map of the 28S. Importantly, we are able to predict the structure of the full length ES regions which was not previously possible.

We compared the two most abundant 28S subtypes and their linked variants and found they have different structures (**Figure 4A-C**, with 22% and 30% frequency of subtypes 1 and 2 respectively, **Methods**). While our method with DMS results in a full length accessibility map of the 28S, secondary structure prediction becomes less accurate for long RNA sequences. Given that ESs have tentacle-like extensions that protrude from the ribosome, we assumed that the core non-ES rRNA is not affected by changes in the ES regions which allowed us to focus on the structures of individual ESs. Most interestingly, we discovered that the ESs es7l, es15l and es27l have major DMS accessibility and structure differences when comparing the two subtypes observed at the GGC sites in es7l, es15l and es27l (**Figure 4A**, ES region box annotations, **Figure S16-20** for es7l, es15l and es27l). When focusing on es27l, the second longest ES, we noticed that the largest accessibility difference between the subtypes was at the site where es27l subtypes differ, at the GGC indel. Specifically, the subtype with one fewer tandem-repeat GGC insertion before the AG at position 3513 of the 28S showed greater DMS accessibility at position 3513 and its vicinity (**Figure 4D,E**). This GGC expands a six tandem-repeat GGCs, i.e. $(GGC)_6$, which changes the region's structure. Moreover, for the es27l region we found local structure changes near the sequence variants which opens the possibility that there are proteins or transcripts that may interact with the subtype with the GGCAG variant but not with the AG variant (**Figure 4D,E** region marked in red). Taken together, our DMS results provide evidence of structural differences for different ribosome subtypes.

Quantifying the relative abundance of rRNA variants in expression data.

Previously, 10 rRNA variants were annotated and showed changed expression between mouse tissues[5]. Here, we found that one of these rRNA variants replicates in our atlas. This promped us to check if rRNA variations accross human tissues by analyzing the publicly available Genotype-Tissue Expression (GTEx) short-read RNA-seq dataset to test if rRNA variant frequencies are associated with human tissue biology (**Methods** for altas usage instructions). Previous studies comparing mRNA across tissues in the GTEx dataset found tissue-specific, including brain-specific, gene expression

[53,54]. Here we analyzed 2,618 samples from 332 individuals and 44 tissues from GTEx and asked if rRNA subtypes differ in their expression in these tissues (**Figure 5A**). We hypothesized that the rRNA subtypes that we identified as highly expressed in the hESC may be important for tissue development. Strikingly, the most abundant subtypes in GTEx significantly differed in expression in many tissues (**Figure 5B-D, upper panels, Figure S21-25**, and **Table S19-20**: P-value < 0.05 FDR corrected Mann-Whitney U rank sum test). Notably, when comparing subtypes expression levels across tissues, we observed significant differences between tissues derived from the ectoderm and endoderm germ layers (**Figure 5B-D, lower panels, Table S21** FDR corrected ranksums test comparing subtype relative abundances of ectoderm-derived tissues in blue and endoderm-derived tissues in red). Most of the ectoderm-derived tissues belong to brain tissues, and most endoderm derived tissues are digestive-system tissues (**Figure 5A** endoderm and ectoderm derived tissues are labeled). Taken together, our results support major changes in the expression of rRNA subtypes across tissues.

Lastly, we asked if changes in the expression of rRNA variants are associated with cancer. For this, we used 10,030 samples of short-read RNA-seq with clinical phenotypes from The Cancer Genome Atlas (TCGA) [55]. When comparing cancer types, we found distinct expression patterns of rRNA regional variants across cancers (**Figure S26-31, Table S22** for region annotations). To test if rRNA variants are cancer-specific, we compared cancer biopsies to control biopsies from the same tissues. Surprisingly, we identified specific rRNA regional variants with significantly different expression levels in control and in cancer biopsies for 11 cancer types (**Figure 6**, **Table S22** for alternative allele regional variant abundances**, Table S23**, P-value < 0.05 after FDR correction, bootstrapping 10,000 times Mann-Whitney U rank sum test with sub sampling controls to match the number of cancer samples). These include rRNA variants that while they appeared in low abundance in both the H7-hESC and control biopsies, are found elevated in cancer biopsies. Thus even low abundance variants hold immense importance as disease biomarkers.

We conclude that our atlas enables direct measuring of rRNA variants changes in expression data. Moreover, we showed that atlas variants are present in translating ribosomes and that they are differentially expressed across tissues and cancer types.

Discussion

Here, by developing a pipeline for long-read sequencing and analysis of rDNA and rRNA from actively translating ribosomes, we measured for the first time variant frequencies in rDNA and rRNA and used *in-situ* sequencing microscopy to validate co-variant expression in individual cells. With this atlas we have enabled greater understanding of the often neglected yet ubiquitous rRNA sequencing data and built an atlas of functional human 18S and 28S rRNA variants at different resolutions, from nucleotide position variants, to 28S gene-level subtypes as a useful resource for studying rRNA variations, and composition across biological conditions (**Extended data 1-5**).

In our study, we have discovered chromosome-associated rDNA subtypes, revealing that different ribosome subtypes based on rRNA sequence variation exist. It may be possible that spatial separation of rDNA subtypes enable regulation of their expression at the chromosome level through allelic inactivation of rDNA loci or inactivation of nucleolar organizer regions (NORs) in the distal junction [56–58]. This may enable global remodeling of rDNA transcription and promote specific ribosome subtypes to be expressed within individual cells. Additionally, using DMS structure probing of full length 28S, we discovered that different rRNA subtypes have different structures at ES regions including different DMS accessibility profiles. Since these ES regions are solvent-exposed and highly flexible, these ES variations may fine-tune regulation of mRNA translation based on differential association with ribosome-associated proteins, mRNA transcripts, or other factors. Moreover, by analyzing the GTEx dataset, we observed differential expression of rRNA subtypes between tissues belonging to ectoderm and endoderm lineages. This pattern might hint at specialized functions of different ribosome subtypes. Long-lived cells associated with the nervous system might require ribosome subtypes that emphasize translation fidelity over speed, as compared to rapidly dividing cells in the digestive tract that require constant replacement given harsh local environments. Indeed, our lab and others have previously shown that es27l plays a role in translation fidelity through association with ribosome-associated proteins [44,59,60]. Such interactions that trade speed over fidelity might be fine-tuned by the expression of different rRNA subtypes.

Finally, by analyzing the TCGA dataset we discovered that some low abundant rRNA variants in control biopsies were elevated in cancer biopsies. Yet the mechanism of elevated expression of such variations remains unknown. One possible mechanism may be enhanced transcription of specific

rDNA copies bearing coding sequence variants and interestingly it was shown that Lung Adenocarcinoma samples were enriched with somatic and germline mutations at rDNA promoter regions [61]. Alternatively, de novo somatic mutations may increase certain rDNA variant frequencies. Future work is needed to understand whether they promote oncogenic ribosome activity and how they are regulated. Therefore, our results provide another layer of ribosome specificity wherein cancer cells might deploy a particular rRNA variant that is more compatible with their cellular fitness. Importantly, we found that specific rRNA variants may be used as biomarkers for disease. Notably, 5-fluorouracil, a common chemotherapy drug, was recently shown to incorporate into rRNA and promote drug resistance by changing mRNA translation [62]. It may be that drugs directly target specific rRNA variants and further examination would be needed to test whether they should be used for cancer specific therapies. Together, our results reveal the presence of structurally different ribosomes at the level of rRNA and provide the first atlas to distinguish different types of ribosomes and to link them to different cellular programs, including those underlying human health and disease.

Acknowledgments

We thank Rhiju Das for interpreting DMS results. We thank Xiangling Meng, Craig Kerr, Ali Wilkening and Michael Montgomery for help with early experiments that were not later followed in this study. We thank the Barna and Pritchard group members for discussions.

MB is supported by the New York Stem Cell Foundation and the National Institutes of Health grant R01HD086634. JKP is supported by RO1 HG008140. DR is supported by ALTF 1042-2019 EMBO and LT000218/2020-L HFSP postdoctoral scholarships. TTS is supported by a National Science Scholarship (PhD) from the Agency for Science, Technology and Research. MB is a NYSCF Robertson Investigator. XW is supported by Edward Scolnick Professorship, Ono Pharma Breakthrough Science Initiative Award, Merkin Institute Fellowship, and NIH DP2 New Innovator Award 1DP2GM146245-01.

Authors contributions

DR conceived the project, designed experimental and computational analyzes, conducted all computational analyses, interpreted the results, and wrote the manuscript. TTS designed and conducted all polysome and DMS sequencing experiments, interpreted the results, and wrote the manuscript. XS developed, optimized, and performed SWITCH-seq experiments. XW supervised the in-situ sequencing experiment. JPS helped in computational analyses. NG helped in experimental data collection. NSA helped interpret GTEx data. RR helped in DMS analyses. MB and JKP

conceived and directed the project and analyses, designed the analyses, interpreted the results and wrote the manuscript.

Declaration of interests

XW is a scientific cofounder of Stellaromics. X.W. and X.S. are inventors on pending patent applications related to SWITCH-seq

Data and materials availability

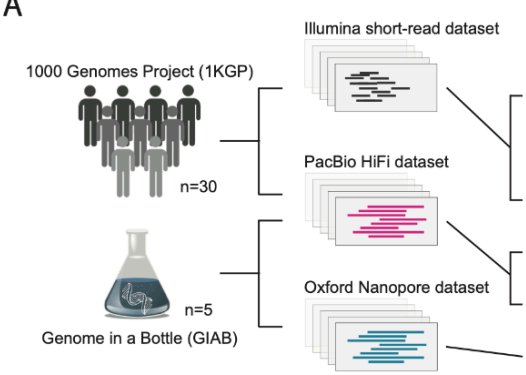
The atlas is available as Extended Data to this publication.
H7-hESC raw rDNA and rRNA sequencing data is available under BioProject ID PRJNA926787.

Limitations of study

In this paper we created an atlas of human rRNA sequence variations in translating ribosomes which we correlate with both development as well as to cancer. In this study we do not demonstrate that expression differences of rRNA variants have functional implications on human development and disease. In the TCGA dataset control samples do not belong to the same matched cancer biopsy and some cancer types have low control sample sizes. In our haplotype analysis we found high standard deviation between haplotype frequencies across individuals which limits our understanding of their functional significance. The expressed rRNA variants belong to the H7-hESC and K562 cell lines. It is likely that there are rRNA variants that are expressed in other human cells or samples that are not found in these cell lines. The RGA method is not limited to variant discovery between paralog genes, it can be applied for variant discovery between any related sequences for example in detecting variants between amplicon sequences.

Figures

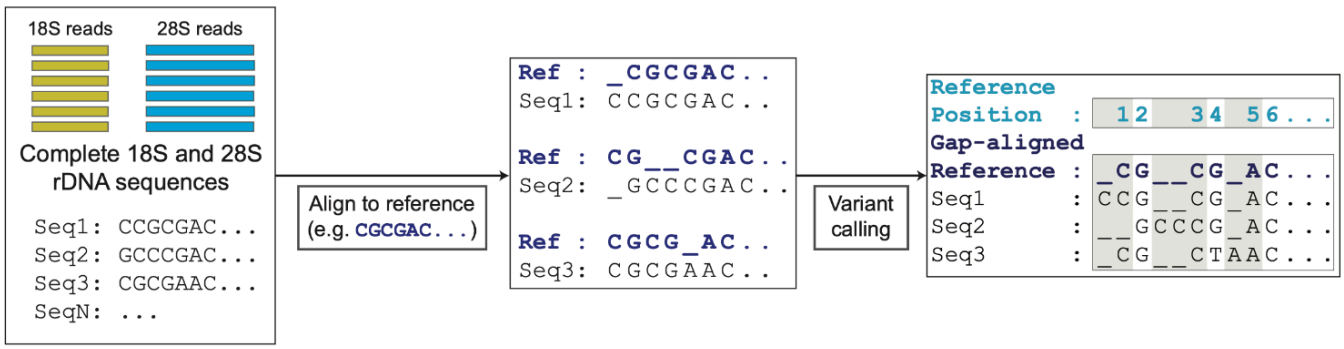
A



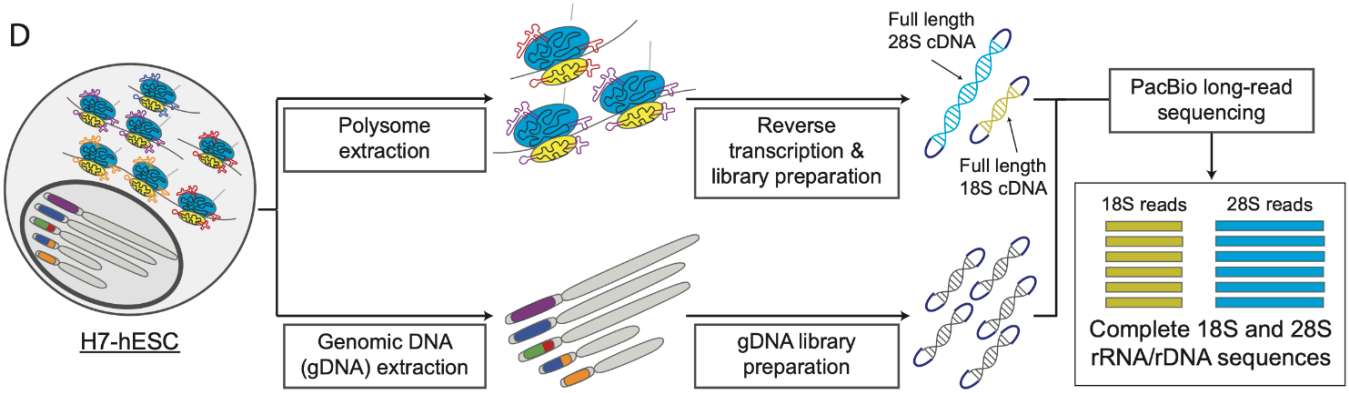
B

Study	Dataset	Technology	SNVs	Indels
Parks et al. (Sci Adv, 2018)	1KGP	Illumina short-read	97.3%	2.7%
Fan et al. (RNA, 2022)	1KGP	Illumina short-read	80.8%	19.2%
Our study	1KGP	Illumina short-read	77%	23%
	1KGP	HiFi long-read	5%	95%
	GIAB	HiFi long-read	4%	96%
	GIAB	ONT long-read	13%	87%

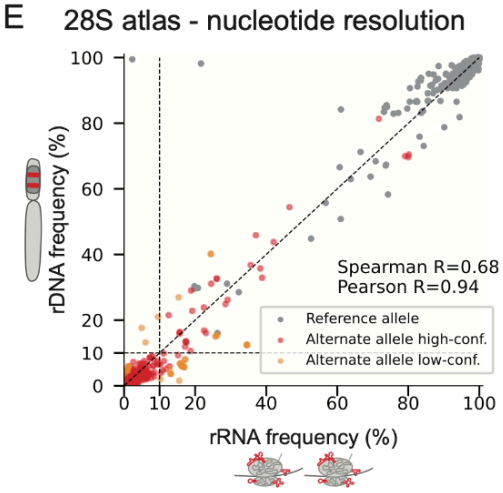
C



D



E



F

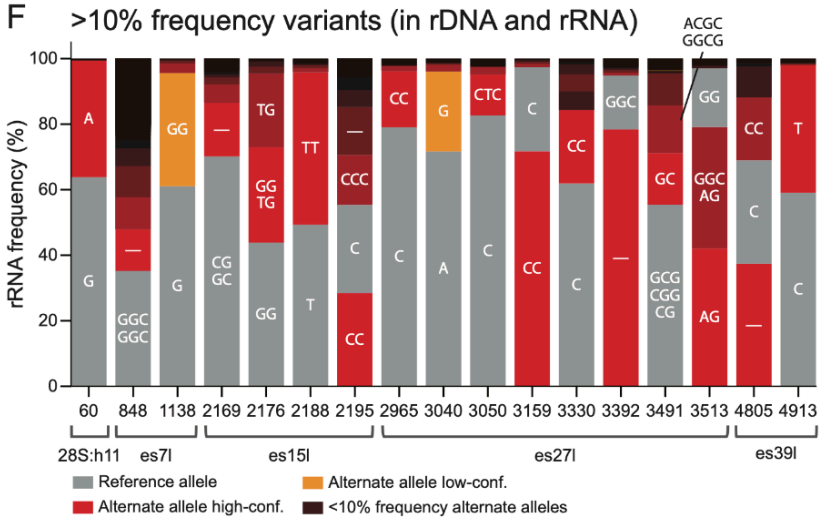


Figure 1. 1,000 Genome Project (1KGP) and H7-human embryonic stem cells (hESCs) rRNA and rDNA variant extraction pipeline with high correlation between 28S rDNA and rRNA high abundant variant frequencies.

A. Graphical illustration of the dataset analyzed consisting of 30 individuals from the 1000 genomes project (1KGP) with both short and long-read sequencing.
B. Comparison of Single Nucleotide Variants (SNVs) and Insertion-deletions (Indels) across studies.
C. Graphical illustration of the Reference Gap Alignment (RGA) alignment method used for variant discovery in 18S and 28S sequences
D. 18S and 28S rDNA / rRNA sequence extraction pipeline from H7-hESC
E. Scatter plot of 28S rRNA frequency (x-axis) and rDNA frequency (y-axis) for reference and alternate alleles. Alternate alleles are marked in red if their frequencies in rDNA and rRNA agree or in orange if they differ significantly. Reference alleles are marked in gray. A dashed black line indicates rRNA frequency equal to 10%. Spearman and Pearson correlations for rRNA frequency and rDNA frequency between alternate alleles alone are presented (calculated on variants, red dots alone).
F. Stacked bar plots of allele frequencies at positions with variants with frequency >10% in both rRNA and rDNA. The nucleotide sequence matching the alleles are indicated inside the bar plots for variants with >10% allele frequency ('-' indicates deletion). The reference allele is indicated in gray and alternate alleles are indicated in color.

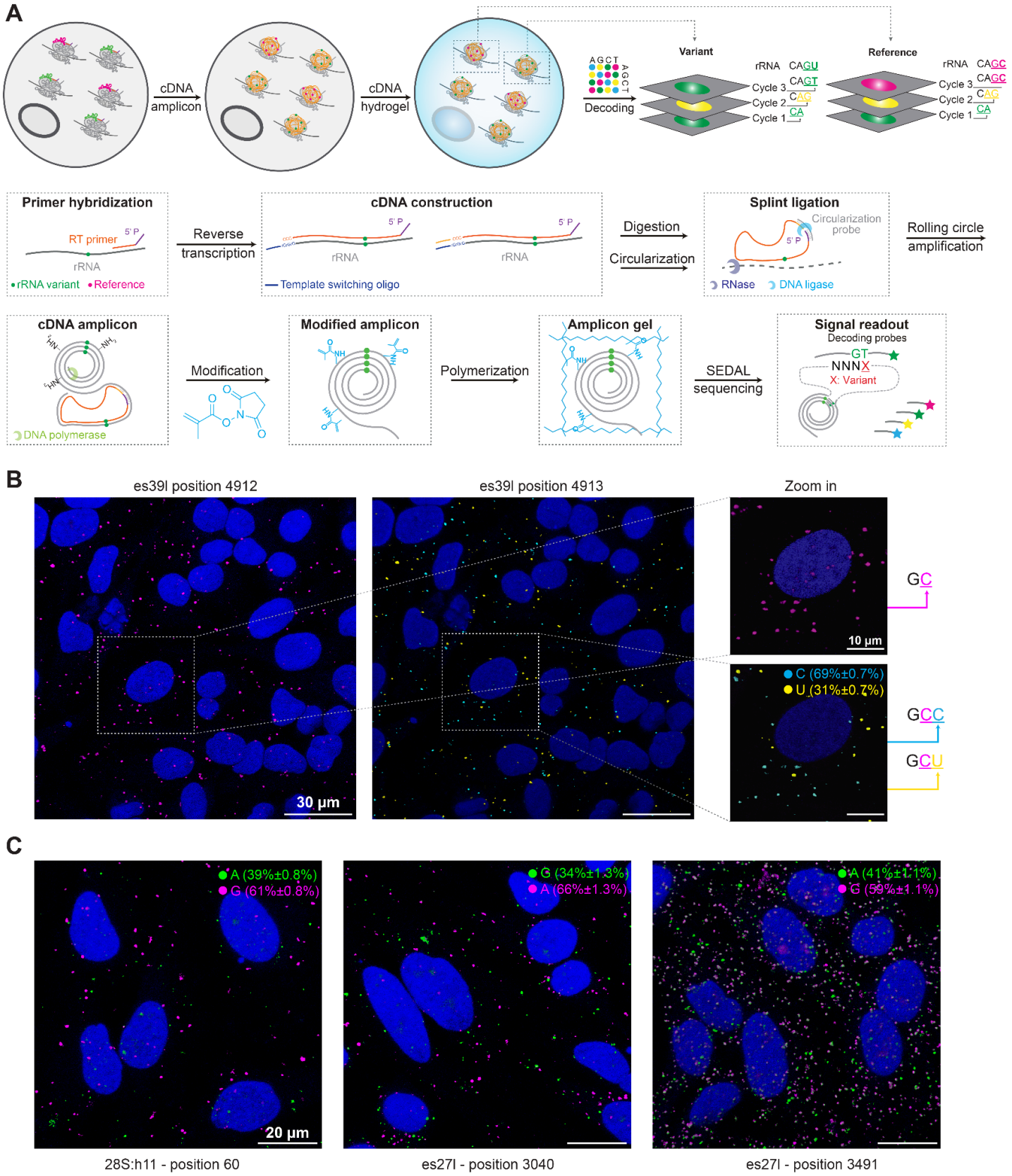
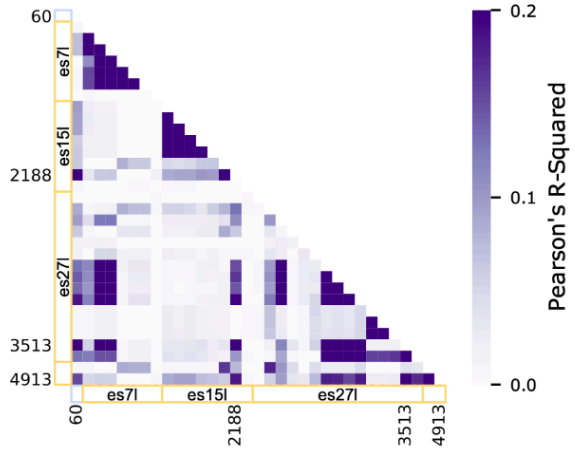

A
cDNA amplicon
cDNA hydrogel
Decoding
Variant
Reference
rRNA CAGU
Cycle 3 CAGT
Cycle 2 CAG
Cycle 1 CA
rRNA CAGC
Cycle 3 CAGC
Cycle 2 CAG
Cycle 1 CA
Primer hybridization
RT primer
5' P
rRNA
rRNA variant
Reference
Reverse transcription
cDNA construction
Template switching oligo
Digestion
Circularization
Splint ligation
Circularization probe
RNase
DNA ligase
Rolling circle amplification
cDNA amplicon
DNA polymerase
Modification
Modified amplicon
Polymerization
Amplicon gel
SEDAL sequencing
Signal readout
Decoding probes
GT
NNNX
X: Variant
B
es39l position 4912
es39l position 4913
Zoom in
30 µm
10 µm
C (69%±0.7%)
U (31%±0.7%)
GC
GCC
GCU
C
A (39%±0.8%)
G (61%±0.8%)
20 µm
G (34%±1.3%)
A (66%±1.3%)
A (41%±1.1%)
G (59%±1.1%)
28S:h11 - position 60
es27l - position 3040
es27l - position 3491

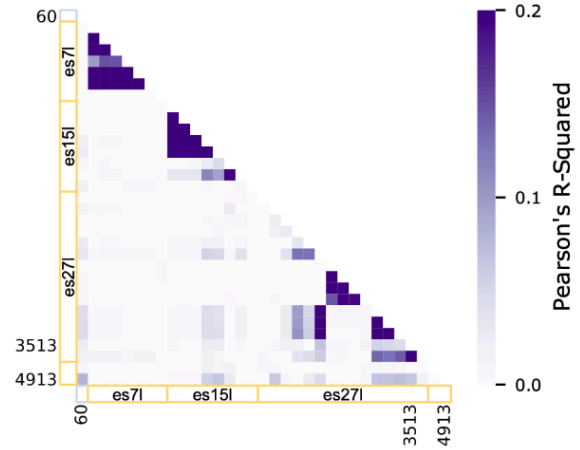
Figure 2. rRNA variants are co-expressed in individual cells as visualized by *in-situ* sequencing

A. Graphical illustration of SWITCH-seq pipeline.
B. Two rounds of representative fluorescent *in situ* sequencing images of HeLa cells (DAPI staining in blue) are presented for the es39l-probed region. We identified a non-variable base C (magenta) at position 4912. At position 4913, two alternative sequences were revealed: the known reference sequence C (cyan) and the alternative variant U (yellow). Data shown as mean percentage ± standard deviation. n = 4 images.
C. Representative fluorescent images of HeLa cells (DAPI staining in blue) showcase 3 highly abundant rRNA variants. The positions of the variants are indicated at the bottom of the images, while the reference (magenta) and alternative alleles (green) are indicated at the top, along with their respective rRNA frequencies. Data shown as mean percentage ± standard deviation. n = 4 images per allele.

A



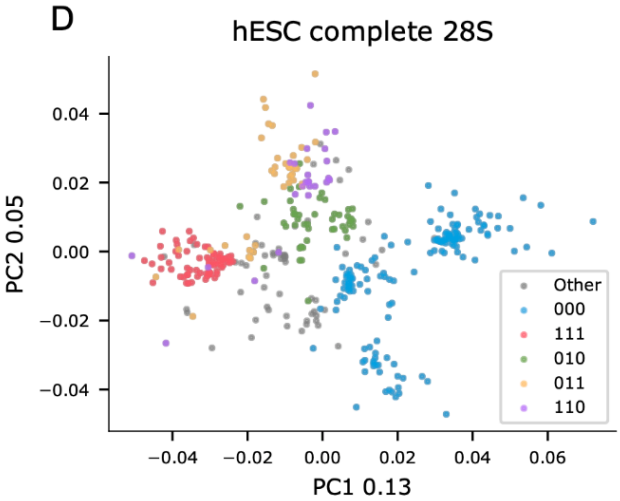
B

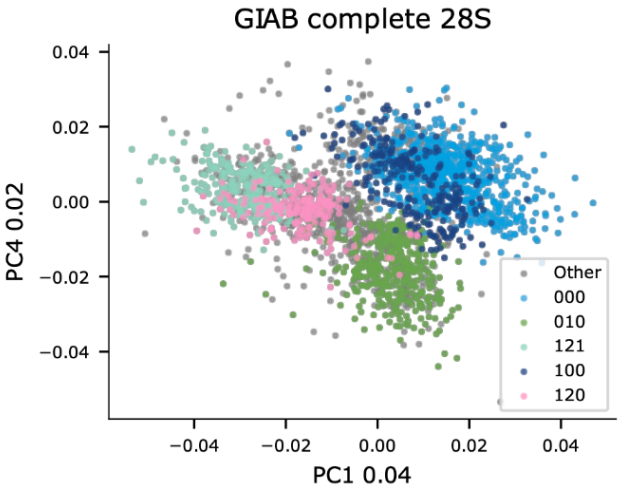


C

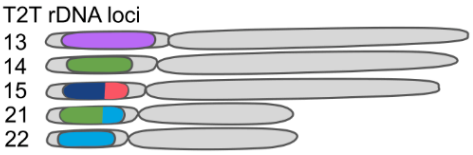
28S position	Digit	Nucleotide
60	0	G
	1	A
3513	0	AG
	1	GGCAG
	2	GG
4913	0	C
	1	T

D





E



Chromosome	Haplotype	T2T copy number
Chr13	110	73
	000	1
	100	1
	120	1
Chr14	010	14
Chr15	100	33
	111	16
	011	1
Chr21	010	39
	000	16
	100	1
Chr22	000	21

Figure 3. 28S subtypes found by haplotype analysis

A. Correlation coefficient (Pearson's r^2) heatmap between positions across H7-hESC 28S rDNA with variant frequency >10%. X-axis and Y-axis are annotated by regions. Helix regions are annotated by light blue and ES regions are annotated by yellow. Individual positions with higher r^2 between regions are also indicated.
B. Same as (a) for the Genome In A Bottle (GIAB) dataset.
C. Haplotype digit code to variant sequence conversion at the 3 positions with higher r^2 in (a) and (b)
D. Bray-Curtis Principal Coordinate Analysis (PCoA) of 386 H7-hESC 28S rDNA sequences (upper panel) and 386 28S rDNA sequences from each GIAB sample (lower panel). Each dot is a complete 28S rDNA sequence with similarity between sequences measured on 6-mers. The colors correspond to coloring an rDNA sequence by its 3 position haplotype described in (c). Numbers in the X and Y labels represent the PCoA explained variance.
E. Telomere-to-telomere (T2T) haplotype distribution across the 5 acrocentric chromosomes. The rDNA acrocentric arms are presented in a schematic cartoon with proportions of rDNA haplotypes in different colors as found in the matching table below. Haplotypes match the 3 position haplotypes in (c). We indicate the rDNA copy number of each haplotype in every chromosome.

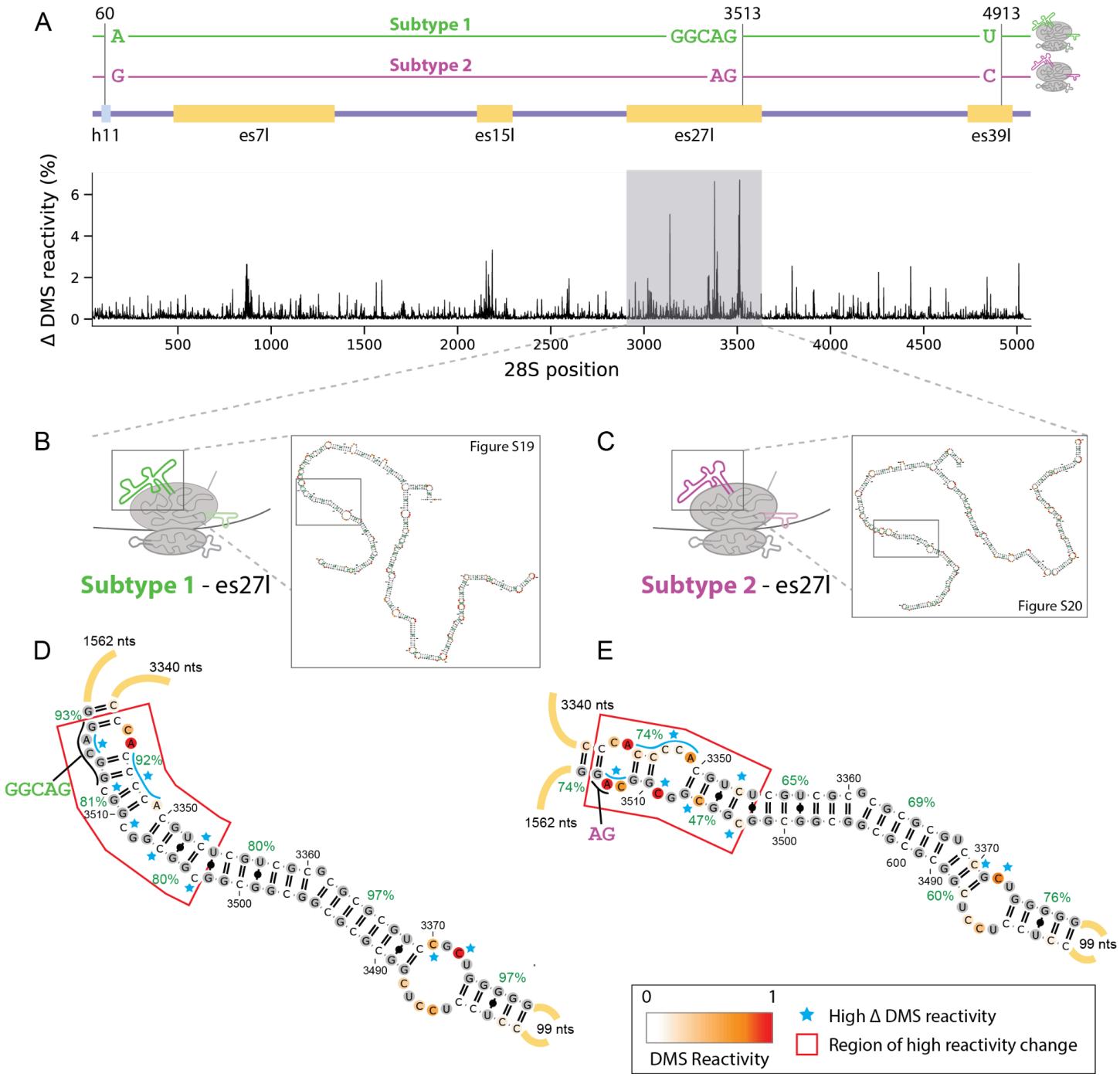


Figure 4. In-cell dimethyl sulfate (DMS) with long-read sequencing shows that 28S subtypes have different RNA 2D structure

A. Changes in DMS accessibility between two most abundant 28S subtypes across the complete 28S molecule. Subtypes are defined by the sequence variants observed at position 60 (28S:h11), 3513 (es27l), and 4913 (es39l), according to the numbering in NR_146117.1. Above

is an illustration of the two subtypes, together with the annotations for the aforementioned regions and other regions with large differences in accessibility. X-axis is the nucleotide position along the 28S. Y-axis is the absolute percentage of DMS accessibility differences at a given position for a window size of 10 nucleotides.

B. Illustration of es27l predicted secondary structure for subtype 1 (A, GGCAG, T). Detailed RNA 2D structure of the whole subtype 1 es27l is shown in Figure S19.
C. Illustration of es27l predicted secondary structure for subtype 2 (G, AG, C). Detailed RNA 2D structure of the whole subtype 2 es27l is shown in Figure S20.
D. Zoomed-in predicted RNA secondary structure of subtype 1 - es27l between position 3310 to 3552(+3). RNA secondary structures are colored by DMS reactivity and helix confidence estimates are depicted as green percentages. Regions with major differences are annotated by the red box. Nucleotides with differing accessibility between the two subtypes are highlighted by blue stars.
E. Zoomed-in predicted RNA secondary structure of subtype 2 - es27l between position 3310 to 3552. Detailed description of annotation is the same as (d).

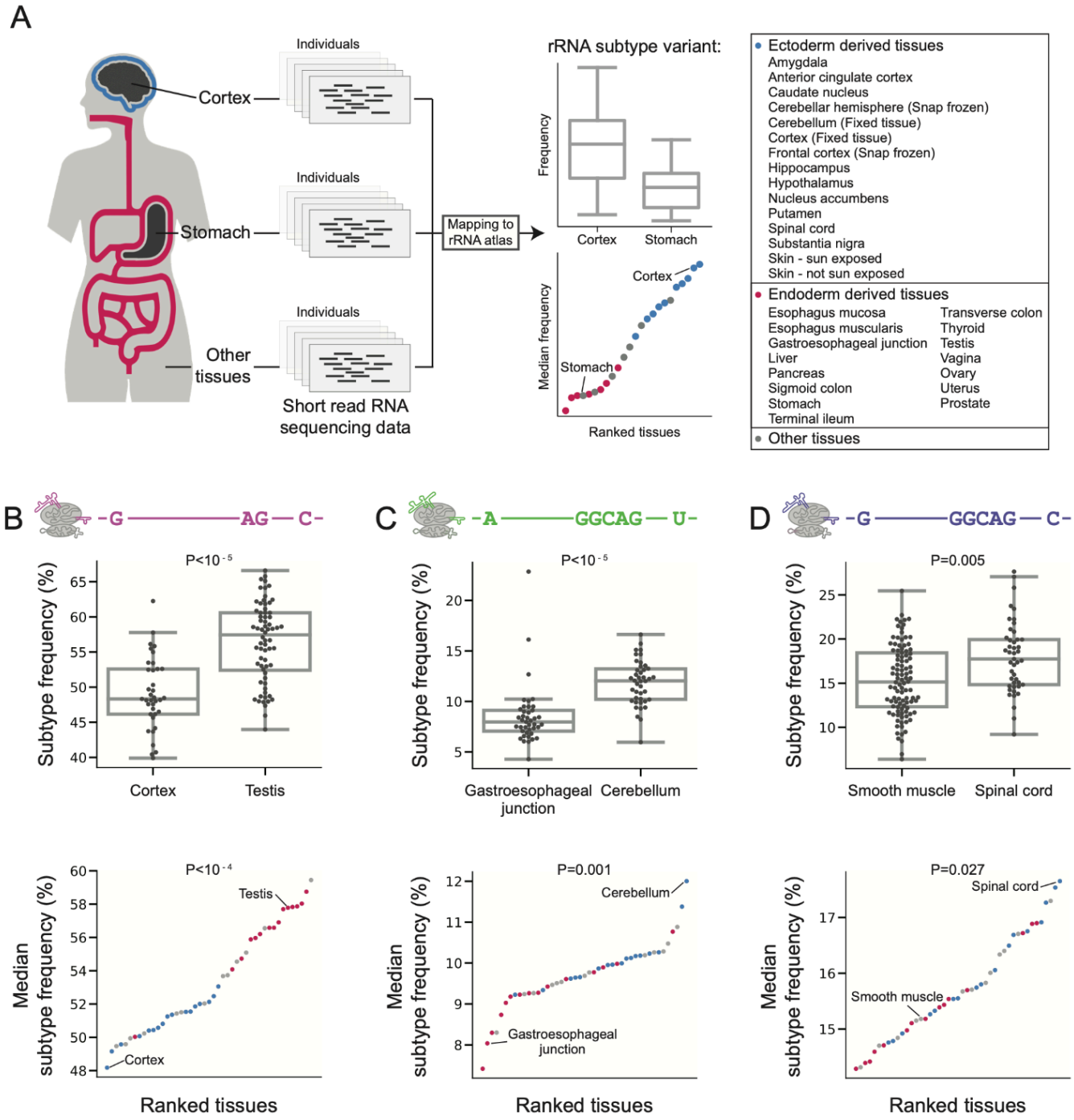

A
Individuals
Cortex
Individuals
Stomach
Individuals
Other tissues
Short read RNA sequencing data
Mapping to rRNA atlas
rRNA subtype variant:
Frequency
Cortex
Stomach
Median frequency
Cortex
Stomach
Ranked tissues
Ectoderm derived tissues
Amygdala
Anterior cingulate cortex
Caudate nucleus
Cerebellar hemisphere (Snap frozen)
Cerebellum (Fixed tissue)
Cortex (Fixed tissue)
Frontal cortex (Snap frozen)
Hippocampus
Hypothalamus
Nucleus accumbens
Putamen
Spinal cord
Substantia nigra
Skin - sun exposed
Skin - not sun exposed
Endoderm derived tissues
Esophagus mucosa
Esophagus muscularis
Gastroesophageal junction
Liver
Pancreas
Sigmoid colon
Stomach
Terminal ileum
Transverse colon
Thyroid
Testis
Vagina
Ovary
Uterus
Prostate
Other tissues
B
-G——AG—C-
P<10-5
Subtype frequency (%)
Cortex
Testis
P<10-4
Median subtype frequency (%)
Testis
Cortex
Ranked tissues
C
-A——GGCAG—U-
P<10-5
Subtype frequency (%)
Gastroesophageal junction
Cerebellum
P=0.001
Median subtype frequency (%)
Cerebellum
Gastroesophageal junction
Ranked tissues
D
-G——GGCAG—C-
P=0.005
Subtype frequency (%)
Smooth muscle
Spinal cord
P=0.027
Median subtype frequency (%)
Spinal cord
Smooth muscle
Ranked tissues

Figure 5. rRNA subtype expression is tissue specific and differs between tissues derived from the ectoderm and endoderm lineages

A. Schematic representation of the GTEx analysis focusing on Cortex versus Stomach comparison. We map rRNA reads from different samples to rRNA subtypes. Per rRNA subtype, we illustrate variant expression comparison between tissues: Upper panel with a box plot comparing rRNA subtype expression in Cortex and Stomach samples. Bottom panel shows the median rRNA subtype expression across all tissues. Cortex and Stomach are annotated, and all ectoderm and endoderm tissues are highlighted in blue/red marks.
B. Upper panel: Box plot comparing the expression levels of the rRNA subtype with the haplotype G,AG,C (positions 60,3513,4913) in Cortex and Testis samples.
   Bottom panel: Scatter plot showing the median frequency of the rRNA subtype from the upper panel across all tissues. Tissues derived from ectoderm are marked in blue, tissues derived from endoderm are marked in red, other tissues in gray. The Cortex and Testis that were shown in the top panel are annotated with a line.
C. Same as (b) for the rRNA subtype with the haplotype A,GGCAG,T highlighting Gastroesophageal Junction and Cerebellum samples.
D. Same as (b) for the rRNA subtype with the haplotype G,GGCAG,C highlighting Smooth muscle and Spinal cord samples.

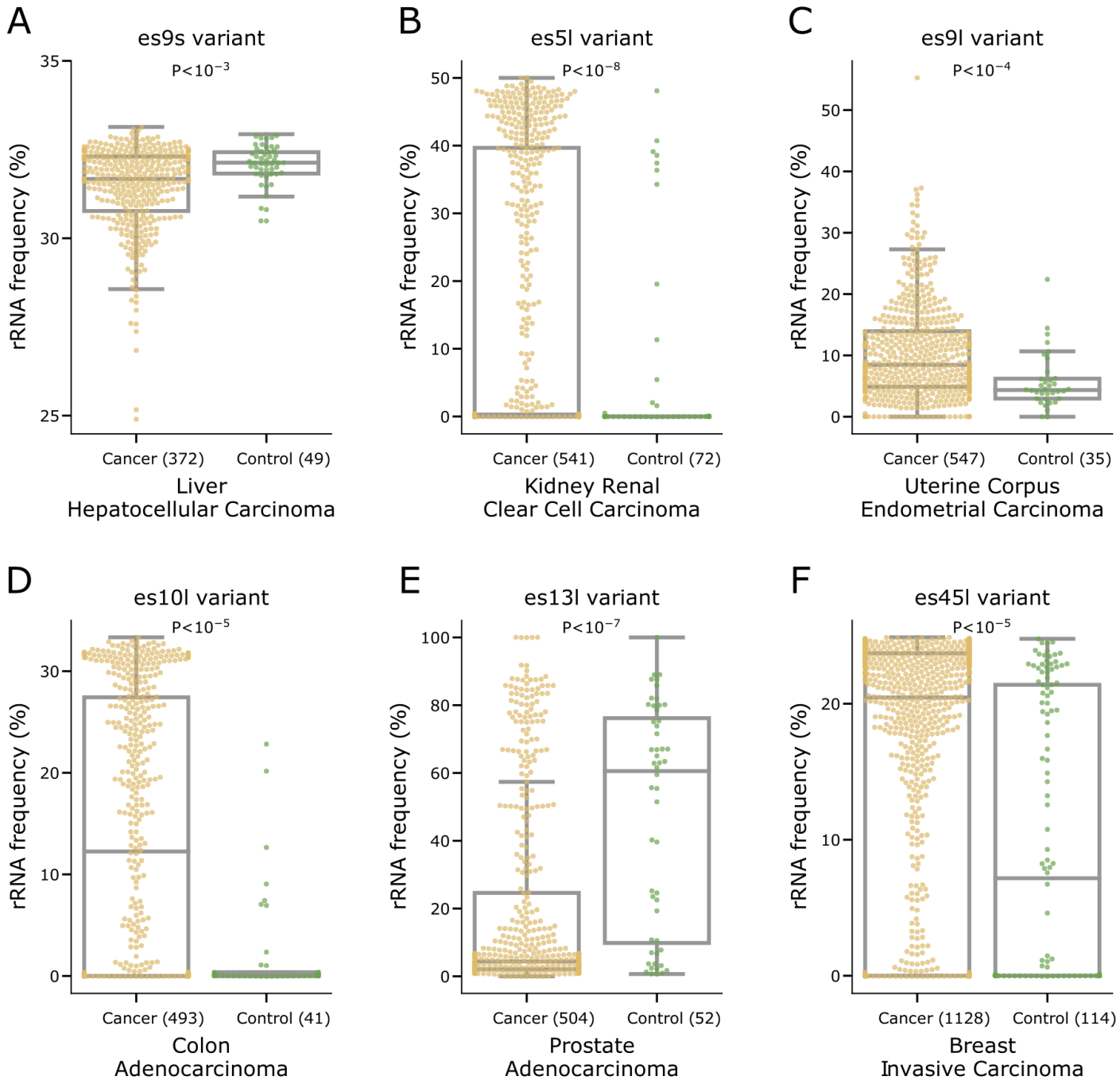


Figure 6. Cancer-specific rRNA variant expression

A. Box plot showing the distribution of rRNA frequencies of the top expressed, alternate allele regional variant, of es9s, ES:es9s:6_d14_r115, across TCGA cancer and control samples for Liver Hepatocellular Carcinoma. The box plot is overlaid with a categorical scatter plot which

saturates at values with over 20 samples and not all points are displayed. The sample sizes of cancer and controls is indicated in brackets and the P-value is indicated at the top (**Table S23**).

B. Same as (a) for region es5l, ES:es5l:12_d1_r1, in Kidney Renal Clear Cell Carcinoma.
C. Same as (a) for region es9l, ES:es9l:21_d2_r8, in Uterine Corpus Endometrial Carcinoma.
D. Same as (a) for region es10l, ES:es9l:21_d2_r8, in Colon Carcinoma.
E. Same as (a) for region es13l, ES:es13l:2_d3_r141, in Prostate Adenocarcinoma.
F. Same as (a) for region es45l, ES:es45l:7_d2_r45, in Breast Invasive Carcinoma.

References

1. Henderson, A.S., Warburton, D., and Atwood, K.C. (1972). Location of ribosomal DNA in the human chromosome complement. Proc Natl Acad Sci USA *69*, 3394–3398. 10.1073/pnas.69.11.3394.
2. Arnheim, N., and Southern, E.M. (1977). Heterogeneity of the ribosomal genes in mice and men. Cell *11*, 363–370. 10.1016/0092-8674(77)90053-8.
3. Kurylo, C.M., Parks, M.M., Juette, M.F., Zinshteyn, B., Altman, R.B., Thibado, J.K., Vincent, C.T., and Blanchard, S.C. (2018). Endogenous rRNA Sequence Variation Can Regulate Stress Response Gene Expression and Phenotype. Cell Rep. *25*, 236-248.e6. 10.1016/j.celrep.2018.08.093.
4. 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073. 10.1038/nature09534.
5. Parks, M.M., Kurylo, C.M., Dass, R.A., Bojmar, L., Lyden, D., Vincent, C.T., and Blanchard, S.C. (2018). Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. Sci. Adv. *4*, eaao0665. 10.1126/sciadv.aao0665.
6. Fan, W., Eklund, E., Sherman, R.M., Liu, H., Pitts, S., Ford, B., Rajeshkumar, N.V., and Laiho, M. (2022). Widespread genetic heterogeneity of human ribosomal RNA genes. RNA *28*, 478–492. 10.1261/rna.078925.121.
7. Clark, C.G., Tague, B.W., Ware, V.C., and Gerbi, S.A. (1984). Xenopus laevis 28S ribosomal RNA: a secondary structure model and its evolutionary and functional implications. Nucleic Acids Res. *12*, 6197–6220. 10.1093/nar/12.15.6197.
8. Wakeman, J.A., and Maden, B.E. (1989). 28 S ribosomal RNA in vertebrates. Locations of large-scale features revealed by electron microscopy in relation to other features of the sequences. Biochem. J. *258*, 49–56. 10.1042/bj2580049.
9. Barbitoff, Y.A., Abasov, R., Tvorogova, V.E., Glotov, A.S., and Predeus, A.V. (2022). Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. BMC Genomics *23*, 155. 10.1186/s12864-022-08365-3.
10. Gibbons, J.G., Branco, A.T., Godinho, S.A., Yu, S., and Lemos, B. (2015). Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. Proc Natl Acad Sci USA *112*, 2485–2490. 10.1073/pnas.1416878112.
11. Paredes, S., Branco, A.T., Hartl, D.L., Maggert, K.A., and Lemos, B. (2011). Ribosomal DNA deletions modulate genome-wide gene expression: “rDNA-sensitive” genes and natural variation. PLoS Genet. *7*, e1001376. 10.1371/journal.pgen.1001376.
12. Gibbons, J.G., Branco, A.T., Yu, S., and Lemos, B. (2014). Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans. Nat. Commun. *5*, 4850. 10.1038/ncomms5850.

13. Nelson, J.O., Watase, G.J., Warsinger-Pepe, N., and Yamashita, Y.M. (2019). Mechanisms of rDNA Copy Number Maintenance. Trends Genet. *35*, 734–742. 10.1016/j.tig.2019.07.006.

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

15. Malinovskaya, E.M., Ershova, E.S., Golimbet, V.E., Porokhovnik, L.N., Lyapunova, N.A., Kutsev, S.I., Veiko, N.N., and Kostyuk, S.V. (2018). Copy number of human ribosomal genes with aging: unchanged mean, but narrowed range and decreased variance in elderly group. Front. Genet. *9*, 306. 10.3389/fgene.2018.00306.

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

17. Chen, C., Feng, L., Chen, J., Shen, J., and Lin, L. (2023). Ribosomal DNA copy number alteration in blood sample from gastric cancer patients. Mol. Biol. Rep. *50*, 7155–7160. 10.1007/s11033-023-08630-y.

18. Stults, D.M., Killen, M.W., Williamson, E.P., Hourigan, J.S., Vargas, H.D., Arnold, S.M., Moscow, J.A., and Pierce, A.J. (2009). Human rRNA gene clusters are recombinational hotspots in cancer. Cancer Res. *69*, 9096–9104. 10.1158/0008-5472.CAN-09-2680.

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

20. Udugama, M., Sanij, E., Voon, H.P.J., Son, J., Hii, L., Henson, J.D., Chan, F.L., Chang, F.T.M., Liu, Y., Pearson, R.B., et al. (2018). Ribosomal DNA copy loss and repeat instability in ATRX-mutated cancers. Proc Natl Acad Sci USA *115*, 4737–4742. 10.1073/pnas.1720391115.

21. Feng, L., Du, J., Yao, C., Jiang, Z., Li, T., Zhang, Q., Guo, X., Yu, M., Xia, H., Shi, L., et al. (2020). Ribosomal DNA copy number is associated with P53 status and levels of heavy metals in gastrectomy specimens from gastric cancer patients. Environ. Int. *138*, 105593. 10.1016/j.envint.2020.105593.

22. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53. 10.1126/science.abj6987.

23. Rodriguez-Algarra, F., Seaborne, R.A.E., Danson, A.F., Yildizoglu, S., Yoshikawa, H., Law, P.P., Ahmad, Z., Maudsley, V.A., Brew, A., Holmes, N., et al. (2022). Genetic variation at mouse and human ribosomal DNA influences associated epigenetic states. Genome Biol. *23*, 54. 10.1186/s13059-022-02617-x.

24. Sims, J., Sestini, G., Elgert, C., von Haeseler, A., and Schlögelhofer, P. (2021). Sequencing of the Arabidopsis NOR2 reveals its distinct organization and tissue-specific rRNA ribosomal variants. Nat. Commun. *12*, 387. 10.1038/s41467-020-20728-6.

25. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. *36*, 983–987. 10.1038/nbt.4235.

26. Luo, R., Wong, C.-L., Wong, Y.-S., Tang, C.-I., Liu, C.-M., Leung, C.-M., and Lam, T.-W. (2020). Exploring the limit of using a deep neural network on pileup data for germline variant calling. Nat. Mach. Intell. *2*, 220–227. 10.1038/s42256-020-0167-4.

27. Fairley, S., Lowy-Gallego, E., Perry, E., and Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. *48*, D941–D947. 10.1093/nar/gkz836.

28. Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., and Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. *40*, 11189–11201. 10.1093/nar/gks918.

29. Benjamin, D.I., Sato, T., Cibulskis, K., Getz, G., Stewart, C., and Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. BioRxiv. 10.1101/861054.

30. Bose, P., Hermetz, K.E., Conneely, K.N., and Rudd, M.K. (2014). Tandem repeats and G-rich sequences are enriched at human CNV breakpoints. PLoS ONE *9*, e101607. 10.1371/journal.pone.0101607.

31. Dohm, J.C., Peters, P., Stralis-Pavese, N., and Himmelbauer, H. (2020). Benchmarking of long-read correction methods. NAR Genom. Bioinform. *2*, lqaa037. 10.1093/nargab/lqaa037.

32. Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. *37*, 1155–1162. 10.1038/s41587-019-0217-9.

33. Lal, A., Brown, M., Mohan, R., Daw, J., Drake, J., and Israeli, J. (2021). Improving long-read consensus sequencing accuracy with deep learning. BioRxiv. 10.1101/2021.06.28.450238.

34. Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. *40*, e72. 10.1093/nar/gks001.

35. Taoka, M., Nobe, Y., Yamaki, Y., Sato, K., Ishikawa, H., Izumikawa, K., Yamauchi, Y., Hirota, K., Nakayama, H., Takahashi, N., et al. (2018). Landscape of the complete RNA chemical modifications in the human 80S ribosome. Nucleic Acids Res. *46*, 9289–9298. 10.1093/nar/gky811.

36. Gerbi, S.A. (1996). Expansion segments: regions of variable size that interrupt the universal core secondary structure of ribosomal RNA. Ribosomal RNA—Structure, evolution, processing, and function in protein synthesis, 71–87.

37. Ramesh, M., and Woolford, J.L. (2016). Eukaryote-specific rRNA expansion segments function in ribosome biogenesis. RNA *22*, 1153–1162. 10.1261/rna.056705.116.

38. Morgan, D.G., Ménétret, J.F., Radermacher, M., Neuhof, A., Akey, I.V., Rapoport, T.A., and Akey, C.W. (2000). A comparison of the yeast and rabbit 80 S ribosome reveals the topology of the nascent chain exit tunnel, inter-subunit bridges and mammalian rRNA expansion segments. J. Mol. Biol. *301*, 301–321. 10.1006/jmbi.2000.3947.

39. van Nues, R.W., Venema, J., Planta, R.J., and Raué, H.A. (1997). Variable region V1 of Saccharomyces cerevisiae 18S rRNA participates in biogenesis and function of the small ribosomal subunit. Chromosoma *105*, 523–531. 10.1007/BF02510489.

40. Houge, G., Robaye, B., Eikhom, T.S., Golstein, J., Mellgren, G., Gjertsen, B.T., Lanotte, M., and Døskeland, S.O. (1995). Fine mapping of 28S rRNA sites specifically cleaved in cells undergoing apoptosis. Mol. Cell. Biol. *15*, 2051–2062. 10.1128/MCB.15.4.2051.

41. Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K., and Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. Nucleic Acids Res. *34*, 5402–5415. 10.1093/nar/gkl655.

42. Bing, T., Zheng, W., Zhang, X., Shen, L., Liu, X., Wang, F., Cui, J., Cao, Z., and Shangguan, D. (2017). Triplex-quadruplex structural scaffold: a new binding structure of aptamer. Sci. Rep. *7*, 15467. 10.1038/s41598-017-15797-5.

43. Mestre-Fos, S., Penev, P.I., Suttapitugsakul, S., Hu, M., Ito, C., Petrov, A.S., Wartell, R.M., Wu, R., and Williams, L.D. (2019). G-Quadruplexes in Human Ribosomal RNA. J. Mol. Biol. *431*, 1940–1955. 10.1016/j.jmb.2019.03.010.

44. Fujii, K., Susanto, T.T., Saurabh, S., and Barna, M. (2018). Decoding the function of expansion segments in ribosomes. Mol. Cell *72*, 1013-1020.e6. 10.1016/j.molcel.2018.11.023.

45. Knorr, A.G., Schmidt, C., Tesina, P., Berninghausen, O., Becker, T., Beatrix, B., and Beckmann, R. (2019). Ribosome-NatA architecture reveals that rRNA expansion segments coordinate N-terminal acetylation. Nat. Struct. Mol. Biol. *26*, 35–39. 10.1038/s41594-018-0165-y.

46. Halic, M., Becker, T., Pool, M.R., Spahn, C.M.T., Grassucci, R.A., Frank, J., and Beckmann, R. (2004). Structure of the signal recognition particle interacting with the elongation-arrested ribosome. Nature *427*, 808–814. 10.1038/nature02342.

47. Qu, L.H., Nicoloso, M., and Bachellerie, J.P. (1988). Phylogenetic calibration of the 5’ terminal domain of large rRNA achieved by determining twenty eucaryotic sequences. J. Mol. Evol. *28*, 113–124. 10.1007/BF02143502.

48. Qu, L.H., Nicoloso, M., and Bachellerie, J.P. (1991). A sequence dimorphism in a conserved domain of human 28S rRNA. Uneven distribution of variant genes among individuals. Differential expression in HeLa cells. Nucleic Acids Res. *19*, 1015–1019. 10.1093/nar/19.5.1015.

49. Guarracino, A., Buonaiuto, S., de Lima, L.G., Potapova, T., Rhie, A., Koren, S., Rubinstein, B., Fischer, C., Human Pangenome Reference Consortium, Gerton, J.L., et al. (2023). Recombination between heterologous human acrocentric chromosomes. Nature *617*, 335–343. 10.1038/s41586-023-05976-y.

50. Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason,

C.E., Alexander, N., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data *3*, 160025. 10.1038/sdata.2016.25.

51. Hori, Y., Shimamoto, A., and Kobayashi, T. (2021). The human ribosomal DNA array is composed of highly homogenized tandem clusters. Genome Res. *31*, 1971–1982. 10.1101/gr.275838.121.

52. Bray, J.R., and Curtis, J.T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol. Monogr. *27*, 325–349. 10.2307/1942268.

53. Taskesen, E., and Reinders, M.J.T. (2016). 2D Representation of Transcriptomes by t-SNE Exposes Relatedness between Human Tissues. PLoS ONE *11*, e0149853. 10.1371/journal.pone.0149853.

54. Donovan, M.K.R., D'Antonio-Chronowska, A., D'Antonio, M., and Frazer, K.A. (2020). Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. Nat. Commun. *11*, 955. 10.1038/s41467-020-14561-0.

55. Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell *173*, 400-416.e11. 10.1016/j.cell.2018.02.052.

56. Schlesinger, S., Selig, S., Bergman, Y., and Cedar, H. (2009). Allelic inactivation of rDNA loci. Genes Dev. *23*, 2437–2447. 10.1101/gad.544509.

57. van Sluis, M., van Vuuren, C., Mangan, H., and McStay, B. (2020). NORs on human acrocentric chromosome p-arms are active by default and can associate with nucleoli independently of rDNA. Proc Natl Acad Sci USA *117*, 10368–10377. 10.1073/pnas.2001812117.

58. Grob, A., Colleran, C., and McStay, B. (2014). Construction of synthetic nucleoli in human cells reveals how a major functional nuclear domain is formed and propagated through cell division. Genes Dev. *28*, 220–230. 10.1101/gad.234591.113.

59. Wild, K., Aleksić, M., Lapouge, K., Juaire, K.D., Flemming, D., Pfeffer, S., and Sinning, I. (2020). MetAP-like Ebp1 occupies the human ribosomal tunnel exit and recruits flexible rRNA expansion segments. Nat. Commun. *11*, 776. 10.1038/s41467-020-14603-7.

60. Shankar, V., Rauscher, R., Reuther, J., Gharib, W.H., Koch, M., and Polacek, N. (2020). rRNA expansion segment 27Lb modulates the factor recruitment capacity of the yeast ribosome and shapes the proteome. Nucleic Acids Res. *48*, 3244–3256. 10.1093/nar/gkaa003.

61. Ohashi, R., Umezu, H., Sato, A., Abé, T., Kondo, S., Daigo, K., Sato, S., Hara, N., Miyashita, A., Ikeuchi, T., et al. (2020). Frequent germline and somatic single nucleotide variants in the promoter region of the ribosomal RNA gene in japanese lung adenocarcinoma patients. Cells *9*. 10.3390/cells9112409.

62. Therizols, G., Bash-Imam, Z., Panthu, B., Machon, C., Vincent, A., Ripoll, J., Nait-Slimane, S., Chalabi-Dchar, M., Gaucherot, A., Garcia, M., et al. (2022). Alteration of ribosome function upon

5-fluorouracil treatment favors cancer cell drug-tolerance. Nat. Commun. *13*, 173. 10.1038/s41467-021-27847-8.