Cell Genomics Transparent Peer Review Record

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

Daphna Rothschild[1,6], Teodorus Theo Susanto[1,6], Xin Sui[3,4,6], Jeffrey P. Spence[1], Ramya Rangan[5], Naomi R. Genuth[1,2], Nasa Sinnott-Armstrong[1], Xiao Wang[3,4], Jonathan K. Pritchard[1,2,7], Maria Barna[1,7,8,*]

Summary

Initial submission:	Received : January 18, 2024
	Scientific editor: Laura Zahn
First round of review:	Number of reviewers: 3 Revision invited : 2/28/2024 Revision received : 5/7/2024
Second round of review:	Number of reviewers: 3 Accepted : 7/14/2024
Data freely available:	Yes
Code freely available:	Yes

Referees' reports, first round of review

Reviewer #1: The authors use long read sequencing data to define some novel 18S/28S variants rRNA variants. They use some novel approaches to show that the variants are expressed in cell lines, model that they may cause some structural changes and also explore if the variants are differentially expressed in different human tissues using GTEX data and in the context of cancer using TGCA data.

The finding that the majority of the variation that occurs is in the form of INDELs is of interest to those interested in the contribution of rRNA to variant ribosomes. However, the manuscript does not demonstrate any functional consequences of the variants described and represents a relatively incremental step in light of the many other previous studies.

The finding that rRNA variation exists and is also included in translating ribosomes is not particularly novel.

(see, Parks et al. (2018), Fan et al. (2022).)

Technical concerns.

1. It would be really useful to have a table for each data source showing the number of reads mapping to each variant at each position to improve the transparency of the data as this can have a large impact on the capacity of variant calling. Their initial criteria seems very lenient, including anything above the estimated background error rate for the technology.

2. The addition of the matched Ribo-RT derived long read rRNA sequencing of 18S and 28S to long read sequencing data derived from the DNA of the same cell line add confidence that the variants detected in this context (present in both datasets) are genomically encoded rather than induced by RT-based modification induced errors. This is further supported by the only high abundance 18S rRNA variant not correlating between datasets. What are the high confidence variants validated in this approach that are also present in the long read GIAB and 1000 genomes datasets?

3. Can the trio data in the GIAB be leveraged to further cross-validate their variant calling (using heritability?)

4. The evidence for assigning haplotypes based on linkage shows relatively weak linkage and the clusters are not clearly separated. I'm not overly convinced by this analysis. Helpful to understand the other sources of variability in the data.

5. I have reserves about the GTEX and TGCA data analyses above just the use of short read data to estimate variant frequency. Firstly, as inter-individual variation in variant frequency occurs, how has this been eliminated as a potential confounder? It is clear in the cancer analysis for example, that there are limited numbers of controls in comparison to cancer samples which could be statistically problematic. In an ideal analysis only matched tissues from the same donors would be included to eliminate unknown batch variation and equalise numbers. Also, the relevance of this finding should be interpreted in light of the many studies that show rDNA copy number instability in the context of cancer.

Reviewer #2:
The authors have significantly revised and expanded the manuscript, including new experimental results. The presentation is much improved, and the addition of long-read and in situ sequencing, in particular, are helpful for validating the key variants reported. My remaining concerns can largely be addressed by further dampening a few of the claims.

My biggest concern is the authors' assertion that the chromosomal location of these subtypes is known. There was a recent paper published "Recombination between heterologous human acrocentric chromosomes" (Guarracino et al. Nature 2023) that reports on recombinational exchange between different acrocentric chromosomes. While the authors of this work could not look specifically at the rDNA arrays, the finding raises the possibility that the human rDNA arrays are not chromosome specific. This hypothesis is also weakly supported by a statement in the CHM13 paper "A chromosome 15 rDNA morph shows the highest identity (98.9%) to the current KY962518.1 rDNA reference sequence, originally derived from a human chromosome 21 BAC clone" (Nurk et al. Science 2021), i.e. that rDNA units may not cluster by chromosome across different human genomes. More complete human genomes will be needed to confirm this, but given the emerging evidence, I would suggest the authors avoid language like that found in the abstract "this study identifies and comprehensively characterizes the diversity of ribosomes at the level of rRNA variants, __their chromosomal location__ and unique structure" since the only chromosome location data presented thus far is from the single CHM13 genome.

The addition of long-read data is great, but the fact so many more indel variants are found makes me a nervous, especially since this stands in stark contrast with the prior work. Indel errors in homopolymers and short tandem repeats remain the biggest error mode for PacBio and Nanopore sequencing. Even though their average error rates are low, they have a locally elevated error rate in these sequence contexts, especially in the presence of certain structures (e.g. "Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate" Guiblet et al. Genome Research 2018). I am glad the authors report how many of these variants are found in both technologies, but it would also be nice to know how many of the indels observed in the long-read data are found in _some_ short-read dataset, just for additional support of the claim that "Indels are the main variants of the human rDNA loci".

Lastly, I find it hard to follow the connection between the "haplotypes" presented in Fig 3 (110, 010, etc.) and the "subtypes" in Figs 4&5. Are the words haplotype and subtype meaning the same thing in these figures? Are the colors the same between the three figures (it doesn't appear so)? It would be nice to use consistent names/colors.

Reviewer #3: The authors have added new experiments and analyses to address most of my previous comments. While the authors argue that demonstrating the functional significance of rRNA variants is beyond the scope of this study, they state in the last sentence of the introduction that "these results suggest that ribosomes with unique sequence variation may be used to modulate different cellular programs underlying human physiology and disease". I still consider this statement an overreach and the authors should address the tentative nature of this conclusion in a "Limitations of study" section. Another point to add to such a section would be the use of RIBO-RT on rRNA from transcripts containing "one" or more ribosomes - meaning, the variant ribosomes may have originated primarily from the monosome population, which could be stalled or otherwise inactive ribosomes. A more balanced discussion of the caveats of the study will address my concerns.

Authors' response to the first round of review

Reviewer #1: The authors use long read sequencing data to define some novel 18S/28S variants rRNA variants. They use some novel approaches to show that the variants are expressed in cell lines, model that they may cause some structural changes and also explore if the variants are differentially expressed in different human tissues using GTEX data and in the context of cancer using TGCA data.

The finding that the majority of the variation that occurs is in the form of INDELs is of interest to those interested in the contribution of rRNA to variant ribosomes. However, the manuscript does not demonstrate any functional consequences of the variants described and represents a relatively incremental step in light of the many other previous studies.

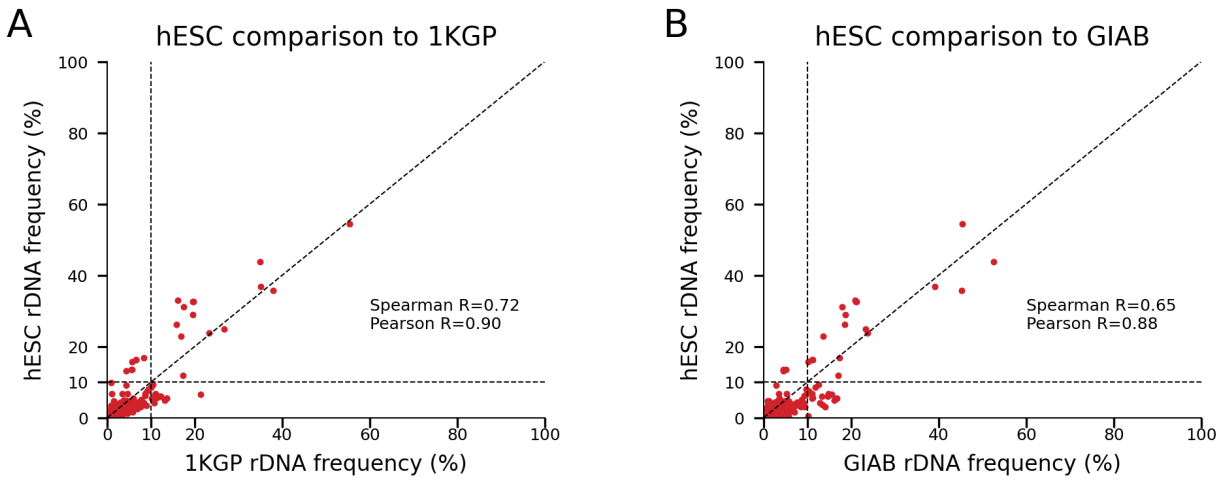
The finding that rRNA variation exists and is also included in translating ribosomes is not particularly novel. (see, Parks et al. (2018), Fan et al. (2022).)
Technical concerns.

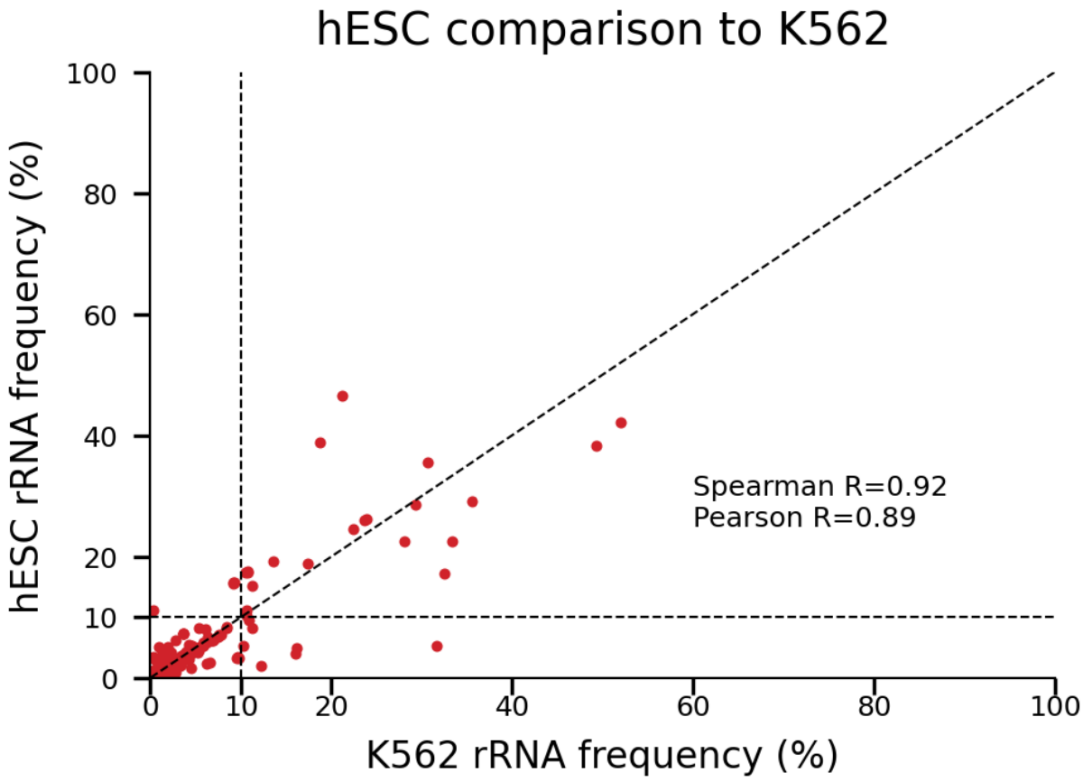
We thank the review for their endorsement and for acknowledging that the vast majority of rRNA variants were unknown until our work. We believe that the work is novel because it revealed that Parks et al. and Fan et al. had the wrong annotation for the majority of rRNA variants and this would have tremendously hindered further work on the function of ribosome rRNA heterogeneity. In the Parks et al. paper the majority of variants were also not reported. Here, as the reviewer wrote we advance the field by developing and employing multiple new methods including a state of the art variant calling algorithm in paralog genes as well as developing methods to sequence full length 18S and 28S, in-situ sequencing, and use DMS structure probing for structure prediction of ribosome bearing variants. This is the first demonstration that rRNA variants can change the structure of the ribosome, showing a function in variants in control of the expansion segments of the ribosome for which there is no CryoEM analysis. We also developed a completely novel in-cell sequencing technology to visualize rRNA variants. With these methods we were able to curate an atlas at different resolutions including single nucleotide resolution and structured regions of expansion segments. We make all of our data and tools publicly available including detailed usage instructions for replicating our results as well as provide instructions on how to use the atlas for variant calling in existing data which we demonstrate in the GTEx and TCGA data. Specifically, in short read datasets it is useful to map the short reads to long atlas regions for variant calling (instead of de-novo calling nucleotide variants). Here we describe how we use the atlas for mapping short reads against the atlas and explain how to estimate variant expression levels. We hope and anticipate that future work will make use of our atlas and methods for further investigating the rRNA variations in different biological contexts and is important for the field at large.

1. It would be really useful to have a table for each data source showing the number of reads mapping to each variant at each position to improve the transparency of the data as this can have a large impact on the capacity of variant calling. Their initial criteria seems very lenient, including anything above the estimated background error rate for the technology.
2.

Unlike previous studies we are able to quantify the expression of variants and put much of our focus on high abundant variants which we are confident are higher than 10% in their abundance (Figure 1F). Still, we find and report hundreds of rRNA variants with lower frequency that allow people to follow up on.
In this new version of the manuscript, we added the raw rDNA read count and rRNA read count to all called variants at every atlas resolution.
1) Specifically, for nucleotide resolution atlas, where variants are not grouped, we added a column "rdna_raw_read_count" and "rdna_raw_read_count".
2) Our atlas is available at different resolutions which is useful for variant calling (mapping short reads to longer sequences). There, the variant name ID indicates in the name the raw rDNA and rRNA read count. The naming convention is "atlas_resolution:regional_variant:ID_Raw-rDNA-count_Raw-rRNA-count".
So for example, at the atlas resolution of expansion segments, the first regional variant of region es2s is named: ES:es2s:0_d730_r43137. In this example, this nucleotide sequence containing sequence variations was observed 730 in rDNA and 43,137 in rRNA.
We now add this description to our supplemental information under "Atlas variant calling".
2. The addition of the matched Ribo-RT derived long read rRNA sequencing of 18S and 28S to long read sequencing data derived from the DNA of the same cell line add confidence that the variants detected in this context (present in both datasets) are genomically encoded rather than induced by RT-based modification induced errors. This is further supported by the only high abundance 18S rRNA variant not correlating between datasets. What are the high confidence variants validated in this approach that are also present in the long read GIAB and 1000 genomes datasets?
We thank the reviewer for this comment and in this revised version for assessing if the hESC atlas variants are present in the general population, we compare the hESC validated variants to the 1000 genomes project and GIAB datasets.
Specifically, we find that 96% and 84% of the hESC variants are also found in the 1000 genomes projects and GIAB datasets respectively. This is generally concordant with expected rates of replication based on these small sample sizes of the 1000 genomes projects and GIAB datasets.

Moreover, all high abundant variants in the hESC (those indicated in Figure 1F) are found in both 1000 genomes project and GIAB datasets.
Notably, by calculating the median variant frequency in the 1000 genomes projects and GIAB datasets, we find high agreement in the frequency of variants between the hESC to other datasets.



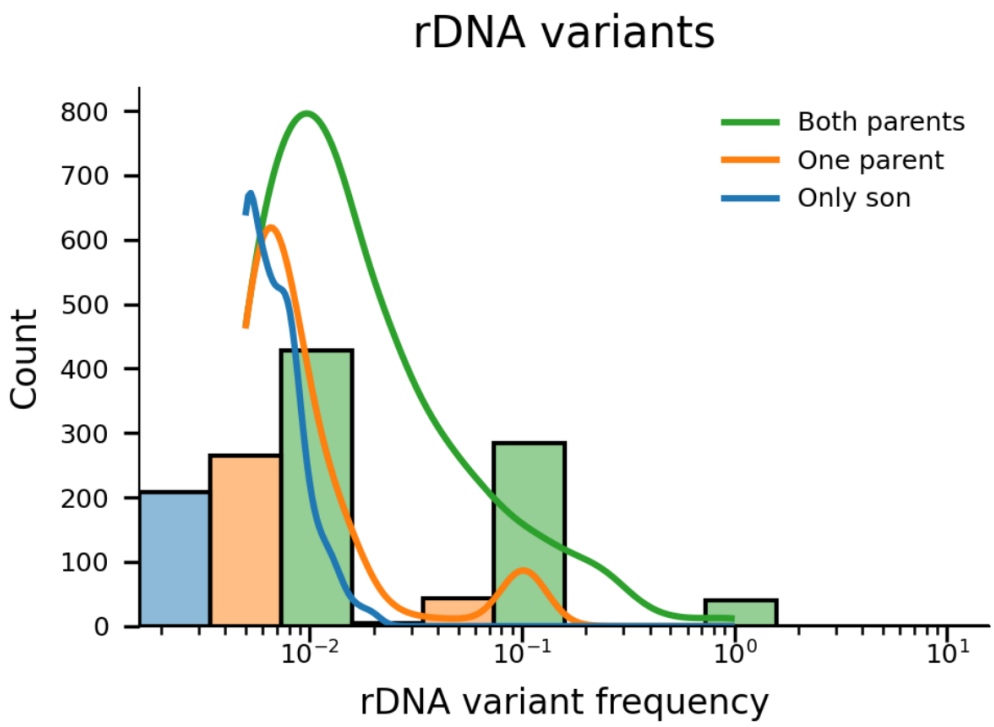


3. Can the trio data in the GIAB be leveraged to further cross-validate their variant calling (using heritability?)

Yes. We thank the reviewer for suggesting this analysis which we now add to the manuscript. Here, the GIAB dataset was used to check which variant frequencies exist for variants that appear only in the child, in a single parent, or both parents. For this we use the Ashkenazy family trio and not the Chinese Han family trio as HiFi long-read data is not available for the Chinese Han father.

In the plot below, rDNA variant frequencies of the GIAB Ashkenazy son are shown in a histogram and are color coded in blue if the are not found in neither parents (they are only in the son), orange if found in only one of the parents, and in green if found in both parents.

When comparing the variant frequencies in the child, we found that all variants with frequency >2% are in at least one of the parents both for single nucleotide variant polymorphisms (SNVs) as well as insertion deletion variants (indels).

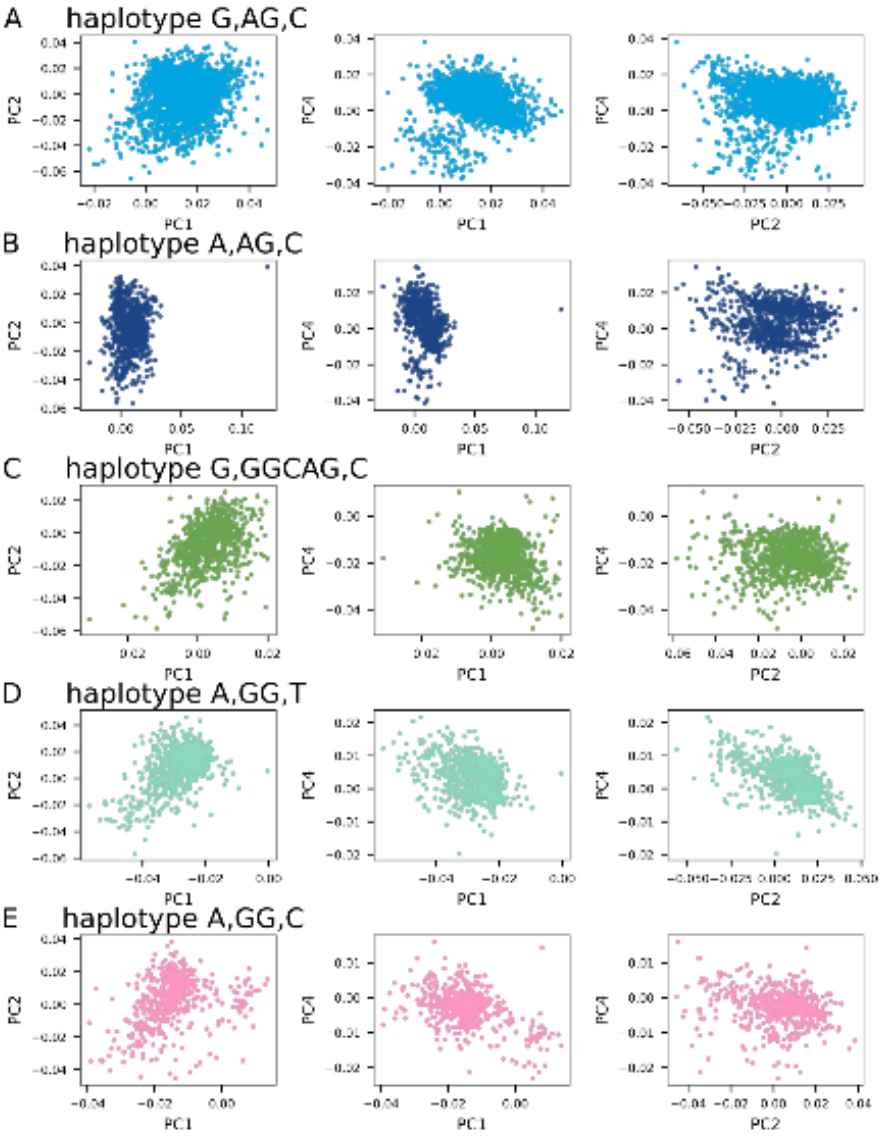


This analysis is now part of the main text with variant frequencies of the GIAB dataset being reported in **Table S6** and the plot is found in **Figure S3.**

4. The evidence for assigning haplotypes based on linkage shows relatively weak linkage and the clusters are not clearly separated. I'm not overly convinced by this analysis. Helpful to understand the other sources of variability in the data.

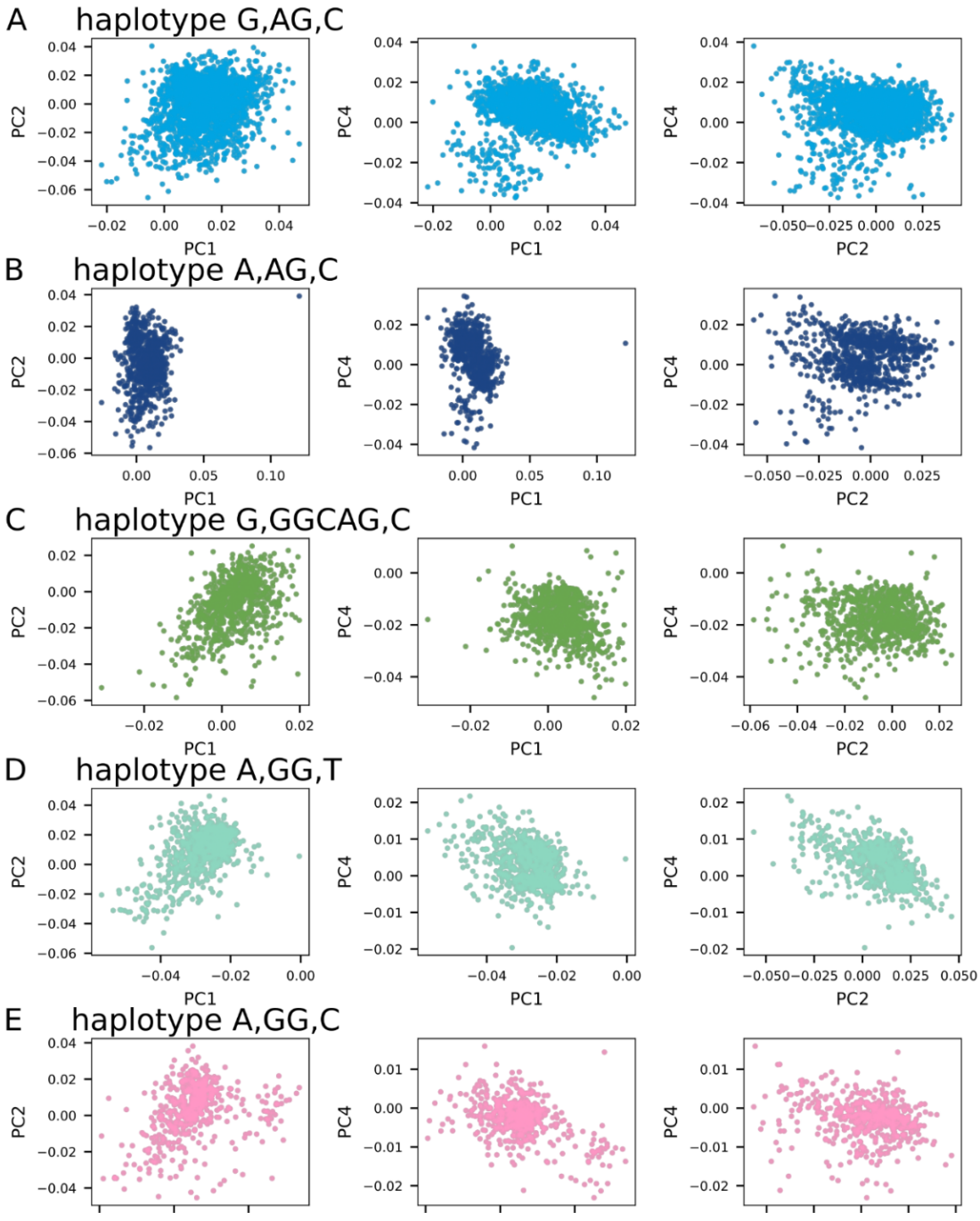
We thank the reviewer for this comment and in this revision we further analyze the haplotypes.

As the reviewer has pointed out correctly, the clusters observed in the GIAB dataset have some overlap as seen in the Bray-Curtis Principal Coordinate Analysis (PCoA). Yet, we believe that it interesting that haplotypes can be used for tagging the 28S which carry hundreds of sequence variants. In the main figure, the clusters defined by the haplotype identity seem to form groups in the PCoA plot (**Figure 3D** where the haplotype groups have different colors). To support this, in



this version we add a supplementary plot that shows the same top abundant haplotype groups when plotted separately. PC1 and PC4 were shown in the main plot with all data points, and here we plot the groups

individually as well as add the second PC (PC2) which was not presented in the main figure as it does not seem to separate the data. What can be seen in this figure is that there is less observed structure in individual haplotypes when plotted separately as compared to the combined data.



This analysis is now added to the main test and **Figure S15**.

Additionally, we agree with the reviewer that there is weak linkage between the 28S positions with sequence variations. This is likely a result of nonallelic gene conversion across rDNA copies which we now emphasize in the text and cite a recent study by Guarracino et al. Nature 2023 focusing on recombination between the 5 accrocentric chromosomes:

"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 perhaps indicating high rates of nonallelic gene conversion across the acrocentric chromosomes (Guarracino et al. 2023)". To further support this, in this version we add an analysis of the observed variability between haplotype frequencies across datasets.

Here, we quantify the haplotype frequencies and their standard deviation in the 1,000 genomes project. When examining the identity of the haplotypes, we find that all of the haplotypes found in high frequency in the hESC, T2T CHM13 genome and the GIAB, are also present in the 1,000 genomes. Moreover, high frequency variants in the hESC are consistently high frequency variants in all datasets analyzed including the 1,000 genomes project (this is a new analysis done as part of this revision and is described at the second reviewer comment - **Figure S5**, **Table S5-6)**. Yet when examining haplotype frequencies we find high standard deviation across individuals. These changes in allele frequencies can explain part of the low LD structure in addition to our hypothesis that there is high rates of nonallelic gene conversion across rDNA copies. In this revision we add this analysis as a new supplementary table (**Table S18**) and are cautious in our interpretation of the haplotypes. This is also now added as a caveat to our "Limitations of study" section. Specifically, we tone down both the abstract as well as in the main result section by reporting that while the 28S clusters by haplotypes and are found in the T2T CHM13 cell line on distinct chromosomes, the overall linkage is low and haplotype frequencies changes across individuals which limits our understanding of rRNA subtypes.

5. I have reserves about the GTEX and TGCA data analyses above just the use of short read data to estimate variant frequency. Firstly, as inter-individual variation in variant frequency occurs, how has this been eliminated as a potential confounder?

As the reviewer pointed out, there is inter-individual variation across individuals. In the GTEx dataset we account for this by considering only one sample per individual when comparing tissues.

In all GTEx analyses, in the main **Figure 5** and supplementary **Figures S21-S25**, we each time compared pairs of tissues. When examining a pair of tissues (tissue1 and tissue2), we divided individuals to those having only tissue1, only tissue2 or both tissues. The individuals that have both tissues were randomly split, half were assigned to tissue1 and the other half to tissue2. This way, by keeping only one sample per person, inter-individual variations do not confound tissue comparisons.

In this revision we expand the methods section. In the previous methods section we wrote a short description that we keep one sample per individual. In this revision we explain that this is done for controlling for inter-individual variations.

We new method section is as followed:

GTEx and TCGA sample handling:
GTEx: Most individuals have multiple organs sequenced. To control for inter-individual variations

when comparing tissues, we select one sample per individual in the GTEx dataset. For each compared tissue pair, individuals that have both tissues are randomly divided into two halves, from the first group we keep one tissue and from the second group we keep the other tissue. This way we only have one tissue per individual when comparing tissues. In all analyses we compared tissues with at least 10 samples.

In the TCGA cancer/control comparison, we selected cancers with at least 50 samples.

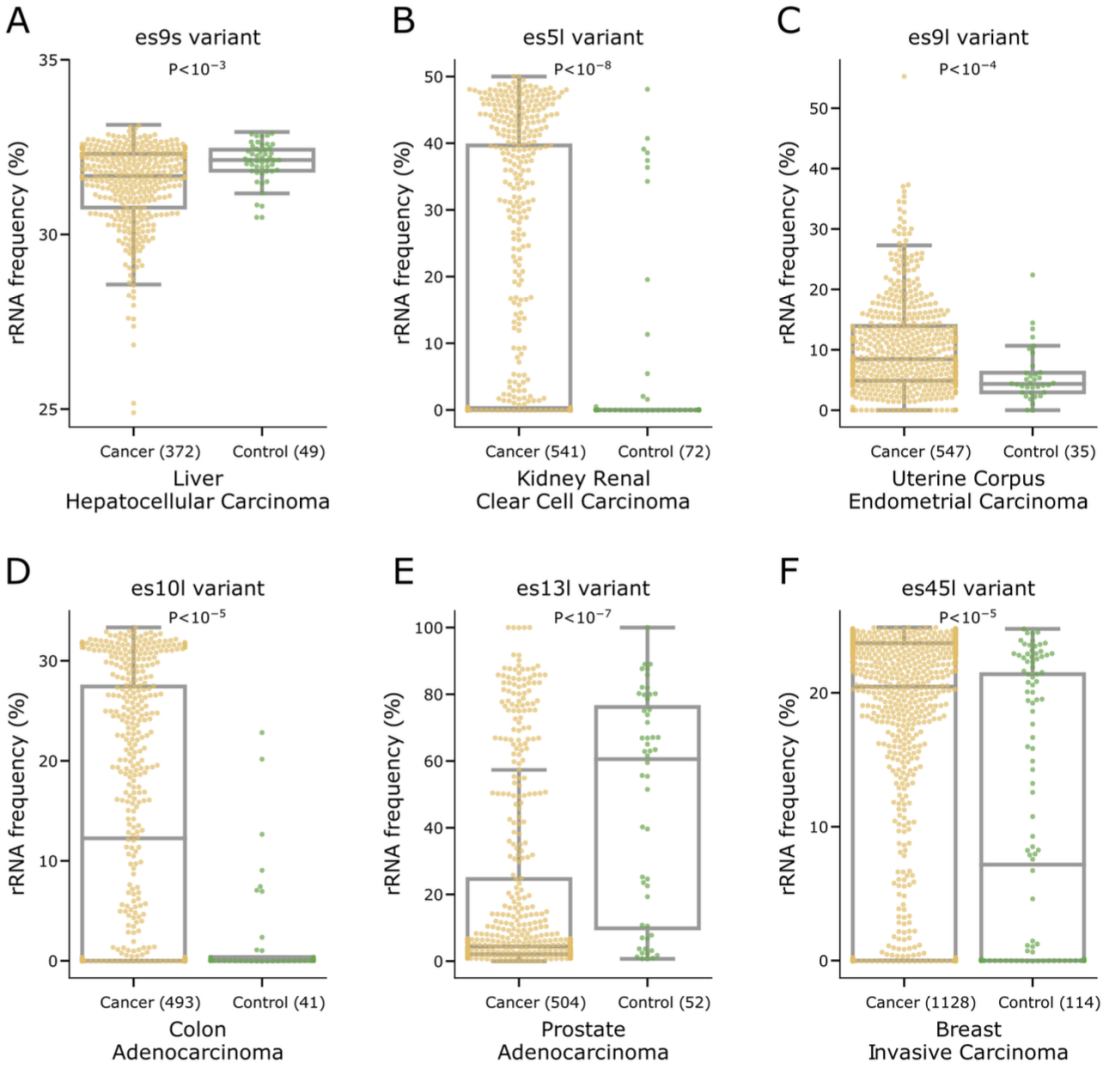
It is clear in the cancer analysis for example, that there are limited numbers of controls in comparison to cancer samples which could be statistically problematic. In an ideal analysis only matched tissues from the same donors would be included to eliminate unknown batch variation and equalise numbers. Also, the

relevance of this finding should be interpreted in light of the many studies that show rDNA copy number instability in the context of cancer.

Indeed in the TCGA dataset control biopsies do not have matched cancer biopsies and some cancer types have a small number of controls. We add this caveat to our "Limitations of study" section. Moreover, as the reviewer highlighted, we don't know the underlying mechanism of how rRNA variations are associated with cancer and indeed there are multiple studies that study rDNA copy number variations in cancer.

However, we reached statistical significance for 11 different cancer types. In the main figure, the cancer types have at least 35 control samples and in breast cancer in particular there were 114 control samples. Following this comment we realized that this was not clearly evident in our figure and in this revision we add the number of samples (in brackets) for cancer and control samples as well as explain in the figure legend that box plots are overlayed with categorical scatter plots yet they saturate at values with over 20 samples.

Not all samples are therefore shown in individual points. In panel B for example (in the figure here below), there are 72 control samples of Kidney Renal Clear Cell Carcinoma. Yet 60 control samples have 0 expression of the alternate allele tested and the scatter plot saturates at value 0 and only shows 20 samples at 0 expression level. Now, the sample size is indicated and the legend hopefully adds clarity. We note that the scatter plot is only for visualization purposes while the bar plot and the calculated P-value are based on all samples.



In the legend we now add "The box plot is overlaid with a categorical scatter plot which saturates at values with over 20 samples and not all points are displayed".

Additionally, we reference multiple studies that found rDNA changes in cancer as well as add a discussion about somatic variations that may lead to the rRNA variant changes that we observed.

In the introduction we add:

"Moreover, rDNA copy number was associated with age and in cancer [14–16] where in cancer, both rDNA copy number gain and loss of was reported in multiple studies [17–21]„

In the discussion we add:

"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 [59]. Alternatively, de novo somatic mutations may increase certain rDNA variant frequencies. ”

Reviewer #2:
The authors have significantly revised and expanded the manuscript, including new experimental results. The presentation is much improved, and the addition of long-read and in situ sequencing, in particular, are helpful for validating the key variants reported. My remaining concerns can largely be addressed by further dampening a few of the claims.

We thank the reviewer for their kind endorsement.

My biggest concern is the authors' assertion that the chromosomal location of these subtypes is known. There was a recent paper published "Recombination between heterologous human acrocentric chromosomes" (Guarracino et al. Nature 2023) that reports on recombinational exchange between different acrocentric chromosomes. While the authors of this work could not look specifically at the rDNA arrays, the finding raises the possibility that the human rDNA arrays are not chromosome specific. This hypothesis is also weakly supported by a statement in the CHM13 paper "A chromosome 15 rDNA morph shows the highest identity (98.9%) to the current KY962518.1 rDNA reference sequence, originally derived from a human chromosome 21 BAC clone" (Nurk et al. Science 2021), i.e. that rDNA units may not cluster by chromosome across different human genomes. More complete human genomes will be needed to confirm this, but given the emerging evidence, I would suggest the authors avoid language like that found in the abstract "this study identifies and comprehensively characterizes the diversity of ribosomes at the level of rRNA variants, __their chromosomal location__ and unique structure" since the only chromosome location data presented thus far is from the single CHM13 genome.

In this revised version we tone down the abstract and main text result. We removed the chromosome specificity from the abstract. In this version we cite the Guarracino et al. Nature 2023 paper as well. We furrther expand on this in the “limitations of study section”.

The addition of long-read data is great, but the fact so many more indel variants are found makes me a nervous, especially since this stands in stark contrast with the prior work. Indel errors in homopolymers and short tandem repeats remain the biggest error mode for PacBio and Nanopore sequencing. Even though their average error rates are low, they have a locally elevated error rate in these sequence contexts, especially in the presence of certain structures (e.g. "Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate" Guiblet et al. Genome Research 2018). I am glad the authors report how many of these variants are found in both technologies, but it would also be nice to know how many of the indels observed in the long-read data are found in _some_ short-read dataset, just for additional support of the claim that "Indels are the main variants of the human rDNA loci".

The reviewer has suggested to cross validate variants - testing if variants that were found in the long-read data are also detected in short-reads, especially given our main result that the vast majority of variants between the rDNA loci are indels. This test can be used to both validate the RGA method for variant discovery in paralog genes as well as used to compare short- and long-reads. We believe that this is a very important comparison that was missing from the

previous version and in this revision, we add the following analysis to the manuscript and thank the reviewer for proposing it.

In this revision we provide another method for cross validating our variant calling by testing if variants that were found in the long-read data are also detected in short-reads. Here we both validate our variant calling as well as compare results found by short- and long-reads. 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. This directly tests if the variants found in the long-reads are also detected in the short-reads (**Methods**).

Specifically, we were able to get about 0.2X coverage of the rDNA loci for the 1,000 genome project tested samples. We get this estimate using the following back of the envelope calculation: Each individual is expected to have ~9 million rDNA nucleotides (the 45S gene which has the 18S and 28S is 45,000 bases

long and it is estimated that individuals have about 200 rDNA copies). Short-read length is ~100 bases and from the 30 individuals tested we extracted 480,000 short reads that mapped to the rDNA.

While this is a low coverage of the rDNA we detected most of the long-read atlas variants.

The short-read that mapped to the rDNA contained 928 nucleotide variants which are 60% of all atlas variants. 32 were SNVs and 896 were indels (SNVs are color coded in light blue and indels are colored dark blue) in the figure below.

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.
These results are now added to the main text as well as in **Table S9** and in the methods under "Validation of atlas nucleotide variants using 1KGP short-read data".

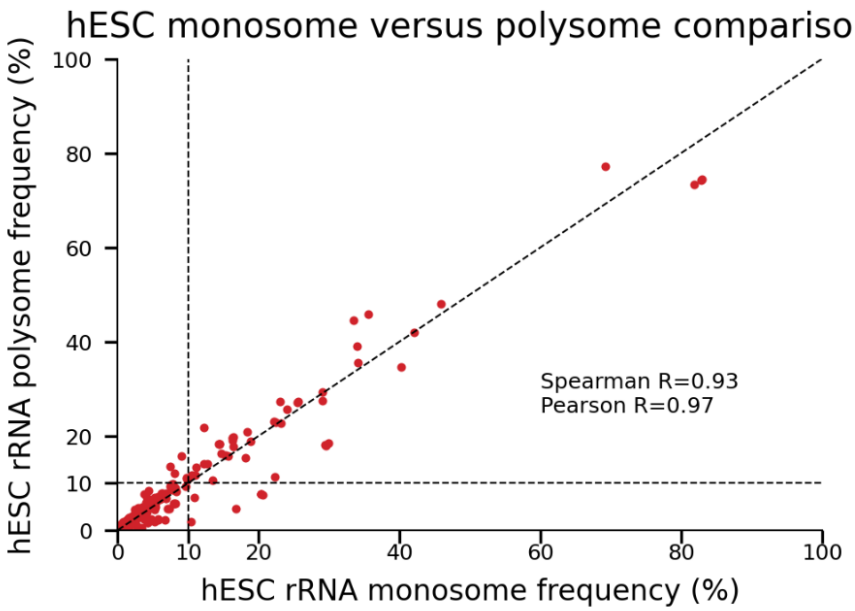
Lastly, I find it hard to follow the connection between the "haplotypes" presented in Fig 3 (110, 010, etc.) and the "subtypes" in Figs 4&5. Are the words haplotype and subtype meaning the same thing in these figures? Are the colors the same between the three figures (it doesn't appear so)? It would be nice to use consistent names/colors.

We thank the review for correcting us and believe that is this version we have corrected inconsistencies.

Reviewer #3: The authors have added new experiments and analyses to address most of my previous comments. While the authors argue that demonstrating the functional significance of rRNA variants is beyond the scope of this study, they state in the last sentence of the introduction that "these results suggest that ribosomes with unique sequence variation may be used to modulate different cellular programs underlying human physiology and disease". I still consider this statement an overreach and the authors should address the tentative nature of this conclusion in a "Limitations of study" section. Another point to add to such a section would be the use of RIBO-RT on rRNA from transcripts containing "one" or more ribosomes - meaning, the variant ribosomes may have originated primarily from the monosome population, which could be stalled or otherwise inactive ribosomes. A more balanced discussion of the caveats of the study will address my concerns.

We thank the reviewer for suggesting to compare monosome and polysomes. We believe that this is an important point, testing if rRNA variants may result in inactive ribosomes and we did not address this possible concern in previous versions. In this revision of the manuscript we add a new analysis of the variants present in the monosomes versus polysomes alone.

We recreated our atlas with only monosomes, only polysomes, and both (which is our previous atlas). When comparing the atlases created with and without the monosomes, we find that only low abundant variants are missing from the solely-polysomes atlas. Moreover, when comparing the abundances of variants found in monosomes vs polysomes we observed high correlation between variant frequencies. Together our results support that rRNA variants are found in all translating ribosome fractions and variants are not specific to monosome or polysomes.



We add this analysis to in the main test in the results section and in **Figure S8** and **Table S10**. Additionally, we add a "Limitations of study" section as followed:

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.

Referees' reports, second round of review

Reviewer #1: Overall, the authors have adjusted the manuscript to further improve the confidence in the presence of the high frequency 28S variants that show conservation across multiple data sources, as well as including more relevant data quality control information in the supplementary section.
Further requests:
*** For the GTEx data: I suspect the approach taken is due to a low number of donors that contribute more than one tissue resulting in an underpowered analysis for directly comparing Tissue 1 and Tissue 2 only from the same donor? This is the only away to truly eliminate confounding by any biased subsampling with regards to underlying rDNA genotype differences. It would be useful to include this analysis as supplementary, as even if underpowered, a directional trend that confirms the current analysis would add further confidence to the conclusions.**
*** For the TCGA data, there are extreme differences in sample sizes. Although, the Mann-Whitney U test applied by the authors can be used when comparing unequal sample sizes, the large difference in control and cancer sample sizes (from approx. 8 to 15 times in TGCA datasets) reduces the statistical power of the test. The best way to improve the test validity is to increase the number of control samples which is clearly problematic. I again think the manuscript could be further strengthened by including an alternative analysis in the supplementary. The alternative way here can be random sub-sampling: the cancer samples are randomly selected to match the control sample sizes, Mann-Whitney U test is performed, and p-value is recorded. Then, the procedure is repeated, for example, 10 000 times. The resulting distribution of p-values**

can shed the light on the validity of the variant frequency difference observed in control and cancer samples.

Reviewer #2: I am satisfied with the authors revisions.

Reviewer #3: All of my comments have been adequately addressed.

Authors' response to the second round of review

Reviewer #1: Overall, the authors have adjusted the manuscript to further improve the confidence in the presence of the high frequency 28S variants that show conservation across multiple data sources, as well as including more relevant data quality control information in the supplementary section.
Further requests:
* For the GTEx data: I suspect the approach taken is due to a low number of donors that contribute more than one tissue resulting in an underpowered analysis for directly comparing Tissue 1 and Tissue 2 only from the same donor? This is the only away to truly eliminate confounding by any biased subsampling with regards to underlying rDNA genotype differences. It would be useful to include this analysis as supplementary, as even if underpowered, a directional trend that confirms the current analysis would add further confidence to the conclusions.

In this revision Table S20 contains pairwise tissue comparisons of rRNA-subtypes from the GTEx dataset. There, we only include samples that have both tissues. When considering tissue pairs with at least 10 samples, we find 37 pairs of tissues that have significant difference in their rRNA-subtype frequencies (FDR corrected Mann-Whitney U test).

* For the TCGA data, there are extreme differences in sample sizes. Although, the Mann-Whitney U test applied by the authors can be used when comparing unequal sample sizes, the large difference in control and cancer sample sizes (from approx. 8 to 15 times in TGCA datasets) reduces the statistical power of the test. The best way to improve the test validity is to increase the number of control samples which is clearly problematic. I again think the manuscript could be further strengthened by including an alternative analysis in the supplementary. The alternative way here can be random sub-sampling: the cancer samples are randomly selected to match the control sample sizes, Mann-Whitney U test is performed, and p-value is recorded. Then, the procedure is repeated, for example, 10 000 times. The resulting distribution of p-values can shed the light on the validity of the variant frequency difference observed in control and cancer samples.

In this revision, Table S23 contains the Mann-Whitney U test performed on subsampled cancer samples to match control sample sizes and is bootrapped 10,000 times. For 11 cancer types we observed rRNA variants that significantly differ between cancer and control samples.
