rRNA paralogs with variations, rRNA-subtypes, affect diverse human phenotypes

Daphna Rothschild[1,3], Jonathan K. Pritchard[1,2,3], Maria Barna[1,3,4]

Correspondence: daphna@stanford.edu, pritch@stanford.edu, mbarna@stanford.edu

1 Department of Genetics, Stanford University, Stanford, CA 94305, USA

2 Department of Biology, Stanford University, Stanford, CA 94305, USA

3 Senior author

4 Lead contact

Abstract

Eukaryotic ribosomal RNA (rRNA) genes exhibit hyper-variability at non-conserved regions known as Expansion Segments (ESs). Due to the numerous rRNA copies in the genome, editing ESs is challenging, and their significance remains unclear. In this study, we analyze rRNA variant frequencies in the UK Biobank population, revealing that highly abundant ES variations are causally linked to human health and physiology. We developed a Ribosome Variation Analysis (RiboVAn) method, identifying both heritable germline variants and a larger proportion of low-heritability, likely somatic variants. The most heritable variants cluster within four ESs of the 28S rRNA, with specific variants in es15l associated with adiposity, es39l linked to body dimensions, and es27l associated with blood-related traits and diseases. Variant-chromosome specificity is observed where functional variants are linked to certain rDNA chromosomes. These findings causally link rRNA sequence variation to human traits and establish that ESs have distinct and important functions in human physiology.

Introduction

Ribosomal RNA (rRNA), presents significant challenges for genetic studies due to its highly repetitive nature, with hundreds of copies arranged in tandem arrays on the five acrocentric chromosomes [1]. Moreover, since rRNA is the most abundant RNA molecule in every cell, reaching 90% of total RNA [2], rRNAs are routinely depleted and discarded from standard transcriptome sequencing analyses. Population genetic studies have also largely ignored rDNA regions, as they are absent from SNP arrays given their repetitive nature. But rRNA are high copy number paralog genes which raises the possibility that sequence variations between copies may form new types of heterogeneous ribosomes that have specialized translation functions [3]. This idea was suggested already in the 1970s, where it was reported that there are variations between the rDNA repeats [4].

Recent studies including our previous work analyzed the population of the 1,000 Genome Project [5] and reported that humans carry hundreds of rDNA sequence variants [6–8]. Yet prior studies relied on short read sequencing which is depleted of highly GC-rich rRNA genes [9], and studies reported discordant findings [6,7]. In our previous study using newly available long-read sequencing in the 1,000 Genomes Project together with developing an accurate variant calling algorithm for paralog genes, Reference Gap Alignment (RGA), we curated the rRNA sequence variants found in the 1,000 genome project [8]. We found that the majority of rRNA variants are in the form of insertion deletion variants, including numerous new variants that were missed by short-reads. We identified tens of highly abundant variants and over a thousand low abundant ones. These variants were further grouped to (1) co-occurring variants, ones found in the same rRNA, termed rRNA-subtypes, which are highly abundant in the genome and (2) low abundant variants, found only in a few copies spread across the rDNA arrays.

These rRNA-subtypes may be new types of heterogenous ribosomes having distinct functional roles in normal physiology whereas low abundant variants could be envisioned as acquired mutations that may have the potential to be associated with specific pathological features. Moreover, we showed that the variants are found in translating ribosomes by successfully sequencing full length rRNA (Ribo-RT) from translationally active polysome fractions in both human embryonic stem cells (hESCs) and K562 cells [8]. We curated the variants into an atlas of rRNA sequence variations found in translating ribosomes. Using long-read DMS-seq in the hESC we showed that the rRNA-subtypes had different ribosome structures which could potentially lead to changes in ribosome function. To assess if rRNA-subtypes could have a role in physiology we tested if rRNA-subtypes expression levels are dynamic and by analyzing the GTEx dataset [10], we showed that rRNA-subtypes had differential expression across tissues belonging to different developmental germ layers [8]. In addition to normal physiology, we hypothesised that rRNA mutations found normally in low abundance may potentially affect disease and by analyzing the TCGA cancer dataset [11] we found specific low abundant variants that became elevated in cancer biopsies [8].

Interestingly, tens of highly abundant rRNA variants found in translating ribosomes were found in expansion segment (ES) regions of the 28S. ESs are enigmatic hyper-variable insertions within the rRNA, located on the outer surface of the ribosome, that vary in length and between eukaryotic species [12]. In humans, expansion segments (ESs) have nearly doubled the rRNA sequence length compared to prokaryotic rRNAs, with several ESs extending to hundreds of nucleotides. Since ESs are not evolutionary conserved, this has led to two notions: either they lack critical functions because they are outside of the conserved ribosome core or alternatively that they have more regulatory functions on the ribosomes that may even be species specific [12]. Yet it remains difficult to study ES functions. rRNA can not be genomically engineered given hundreds of copy numbers in the genome. Moreover, studying ESs is challenging because they are flexible and their structures remain unresolved by cryo-EM. Still, our previous work using DMS-seq showed that sequence variants at ESs change their structures around sites of variations. While very little is known about the function of ESs, multiple recent

studies suggest they are important[13,14,15,16,17]. One of the longest ESs, named es27l which is about 700 bases long, was shown to bind MetAP and to be important for translation fidelity[13,14,15]. Moreover, it was shown that mRNAs bind ESs[16], possibly regulating their translation, including Hoxa9 mRNA which was shown to bind specifically es9s[17]. Given that highly abundant variants were found in ES regions it is possible that these variants alter the function of ES regions and control ribosome function.

While overall changes in rRNA, largely at the level of copy number variation, were associated with human health and disease [18–21], it remains an outstanding question whether ESs have functional roles and whether common variants found within them causally affect human phenotypes. To test this, an ideal experiment would have been to genome edit the rRNA loci and test the outcome of adding or removing specific variations. While editing the rRNA is not possible given their high copy numbers, the UK Biobank (UKBB), with its extensive whole-genome sequencing (WGS) data [22], provides an unprecedented opportunity to dissect germline from somatic mutations and to establish a causal role of rRNA variations on human traits. Indeed, recent work leveraging WGS has shown that germline variations in rDNA are associated with complex human traits[19,23,24]. Here, we address whether distinct rRNA subtypes confer specialized functions that contribute to ribosome heterogeneity in normal human physiology.

Using the WGS from the UKBB and our curated atlas rRNA variants we show that rRNA-subtypes have different causal effects on human traits mediated by functional changes in ES regions. These findings not only demonstrate the critical importance in ES function for human physiology but also define how variations in the most highly paralogous genes in the human genome are associated with human traits.

Results

A Pipeline for Associating rRNA Variants with Human Traits

The human rDNA loci are organized in hundreds of tandem repeats on the short arms of five acrocentric chromosomes (**Figure 1A**). In our previous study, we generated a comprehensive atlas of rRNA sequence variants that are confirmed to be present in translating ribosomes [8]. Here, we investigated the role of rRNA variants in human health and by focusing our analysis on variants from our atlas, we ensure that we are studying variations found in translating ribosomes (**Figure 1A**). Here, we move beyond the associations found in our previous work to ask if these variants are causally linked to human traits. Specifically, since the direction of information flows from DNA to phenotype, establishing that germline mutations in rRNA associate with a phenotype implies their causal effect on the tested phenotype [25,26] (**Figure 1B**).

Since rRNA genes are highly repetitive, short read sequencing cannot identify the chromosomal location of variants. Instead, we use short-reads to call variant

frequencies within the rRNA. For calculating variant frequencies in short-read data, it is necessary for short-reads to align to longer sequences than the short-reads themselves. Since, UKBB short reads are 150 bases long and we expanded the published rRNA atlas [8] with 150 bases of non-variant reference sequences on both 5' and 3' ends in order to map reads to the atlas (**Data S1**, **Methods**). Yet in order to find single nucleotide variants after mapping to variant combinations there needs to be additional conversion from variant combinations to single nucleotide variants. Here we developed the Ribosome Variant Analysis (RiboVAn) pipeline to calculate atlas single nucleotide variant frequencies from short-read data.

RiboVAn reports the fraction of variant frequencies observed per individual for every atlas variant (**Figure 1C**). We created lookup tables, one for each ES/non-ES region, which contain for every combination of variants the positions and single nucleotide variants that are found in combination. With these tables, after mapping reads to the ES resolution atlas which contain combinations of variants, we sum nucleotide variants using the mapped reads and the conversion tables to report nucleotide relative frequencies. We released the RiboVAn tool as part of this study (**Methods** and **Data Availability**).

With this method we calculated the nucleotide variant frequencies for 928 variants, 318 variants from the 18S and 610 variants from the 28S, in 487,822 individuals in the UKBB cohort. Importantly, given GC-biases found in short-read data, as we later describe here, we analyze sequencing variant read depth statistics which we use to study the noise of variant calling. The resulting variant frequency table from RiboVAn is then used for further heritability analysis and association with phenotypic traits.

To control for technical biases, particularly the under-sequencing of GC-rich regions by Illumina platforms, we compared variant frequencies from UKBB short-reads to those from long-read sequencing of the hESCs from our atlas. We observed that highly abundant variants found in hESCs were all highly abundant variants in the UKBB and in agreeing frequencies, but we observed much lower agreement for the hESC low abundant variants (**Figure S1A**). When analyzing the UKBB rRNA variant sequencing depth, we found that discordant variants, those found in low frequency in the hESC and high frequency in the UKBB had low sequencing read count and likely resulted from high noise (**Figure S1B**).

Since hESCs had similar variant frequencies to the average sample in the UKBB, we next asked if variant frequencies vary within the population. Highly-abundant variants showed higher variation in the population and the top 10 most variable ones all belonged to the 28S large subunit rRNA gene (**Figure 1D**). These variants also had low noise levels when examining their sequencing raw read count. Surprisingly, we observed a wide distribution of variant frequencies in the UKBB population in these top variants. For example, the variant G:A at position 60 of the 28S varied dramatically in frequency between individuals from 14% to 80% of the copies carrying the G to A variation (**Figure 1D**, cyan histogram showing the distribution of 60:G:A variant). This

makes it particularly interesting to ask how such variability in the population affects human phenotypes. Moreover, in agreement to our previous study, most of these variants were located in the 28S rRNA ESs (**Figure 1D**, ES/non-ES are annotated on the cartoon ribosome). Indeed, these included high-frequency variants identified in our previous work that form distinct structural subtypes within the ES regions es15l, es27l and es39l. We conclude that highly abundant variants mostly within ES regions display high variations between the UKBB cohort.

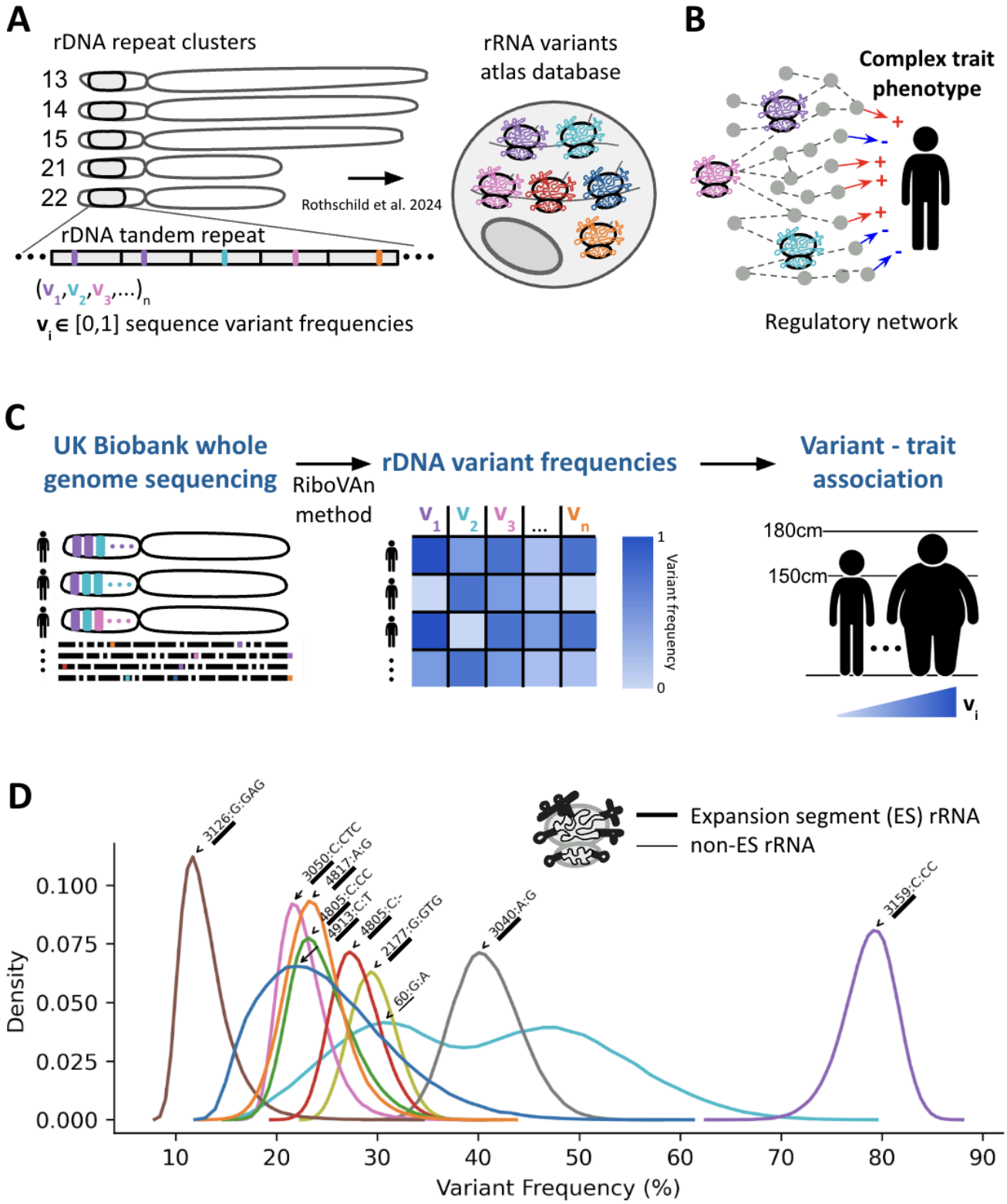

Figure 1: An rRNA sequence variants analysis pipeline
(A) Structure of the rDNA acrocentric chromosomes (13, 14, 15, 21, and 22) depicting the repeat clusters. The rRNA variants atlas database is shown, illustrating the sequence variant frequencies identified in translating ribosomes, as previously reported by Rothschild et al. (2024) [8].(B) Conceptual framework illustrating the hypothesis that rRNA variants affect complex trait phenotypes through regulatory networks. The figure highlights the potential connections between rRNA variants and various human health traits.

(C) Workflow for the analysis conducted in this study. Whole genome sequencing data from the UK Biobank is mapped to the rRNA variant atlas using the RiboVAn method for variant calling. The resulting rDNA variant frequencies are then associated with human traits.
(D) Distribution of variant frequencies for the top ten rRNA variants identified in the study. The density plot illustrates the frequency of variants, with annotations indicating the location of rRNA variants on the ribosome, emphasizing that most variants are located within expansion segments (ES).

rRNA Variants Include Both Heritable and Somatic Variations

Next we asked if rRNA variants are heritable, germline variations. To distinguish between germline and somatic variants, we assessed heritability using 177 monozygotic twins within the UKBB. We found that a subset of rRNA variants is highly heritable (**Table S1**). This included 53 variants with high correlation between twins (Spearman R>0.8), confirming their germline origin (**Figure 2A**, an example of one highly heritable rRNA variant, **Figure B2**, dashed line indicating Spearman correlation exceeding 0.8 between twins). This finding is in agreement with Rodriguez-Algarra et al. reporting germline rRNA variations [24]. However, the overall distribution of twin-variant-correlations (which we use to estimate heritability) is skewed towards lower values, suggesting that a majority of rRNA variants are likely somatic (**Figure 2B, Figure S2A** an example of one lowly heritable rRNA variant).

We asked if low-heritability was driven by noise. Indeed, variants belonging to regions with lower sequencing depth showed low-heritability, but this did not account for the entire signal (**Figure S2**). Specifically, we compared heritability to the variant read-depth in the twins. Out of 928 rRNA variants tested, 354 variants had low read-depth (<5000 reads in average mapped to the rRNA variant, **Figure S2** showing variant read depth). Using all variants, we found moderate correlation between heritability and sequencing-depth (Pearson R=0.25 between the mean variant read depth and twin-variant-correlation). After removing low read-depth variants, the remaining 570 variants showed relatively low correlation between heritability and sequencing depth (Pearson R=0.12 between the mean variant read depth and twin-variant-correlation). This supports that overall low heritability remains after controlling for GC biases, which likely reflects somatic mutations. We conclude that both heritable and somatic rRNA variants exist.

The most highly heritable variants are predominantly located in the 28S rRNA, specifically within four out of twenty nine mammalian ESs: es7l, es15l, es27l and es39l (**Figure 2C,** see ribosome structure video). These variants include mostly highly-abundant variants that show large abundance-variability within the cohort (**Figure 1D**). For this study, we focused only on these 53 most highly heritable variants (Spearman R>0.8), as their association with human traits provides a direct causal link between rRNA sequence variations and human physiology, thereby establishing a clear biological function for ESs.

In agreement with our previous study, these variants form haplotypes, combination of co-inherited variants, within the ES region but have much lower linkage to other regions

outside individual ES regions (**Figure 2C**, ES region annotations marked in different colors and with a ribosome cartoon). This suggests that ESs are largely independent from one another and could potentially have different functions. Additionally, we observed some weak linkage in between regions es27l, es39l and position 60 of the 28S and almost no linkage to es7l and es15l (**Figure S3**). In accordance with our previous study, while es7l is the longest ES, its variants are lowly linked to other ESs. Together, this supports that ES regions could have independent functions with potential cooperations between some ESs.

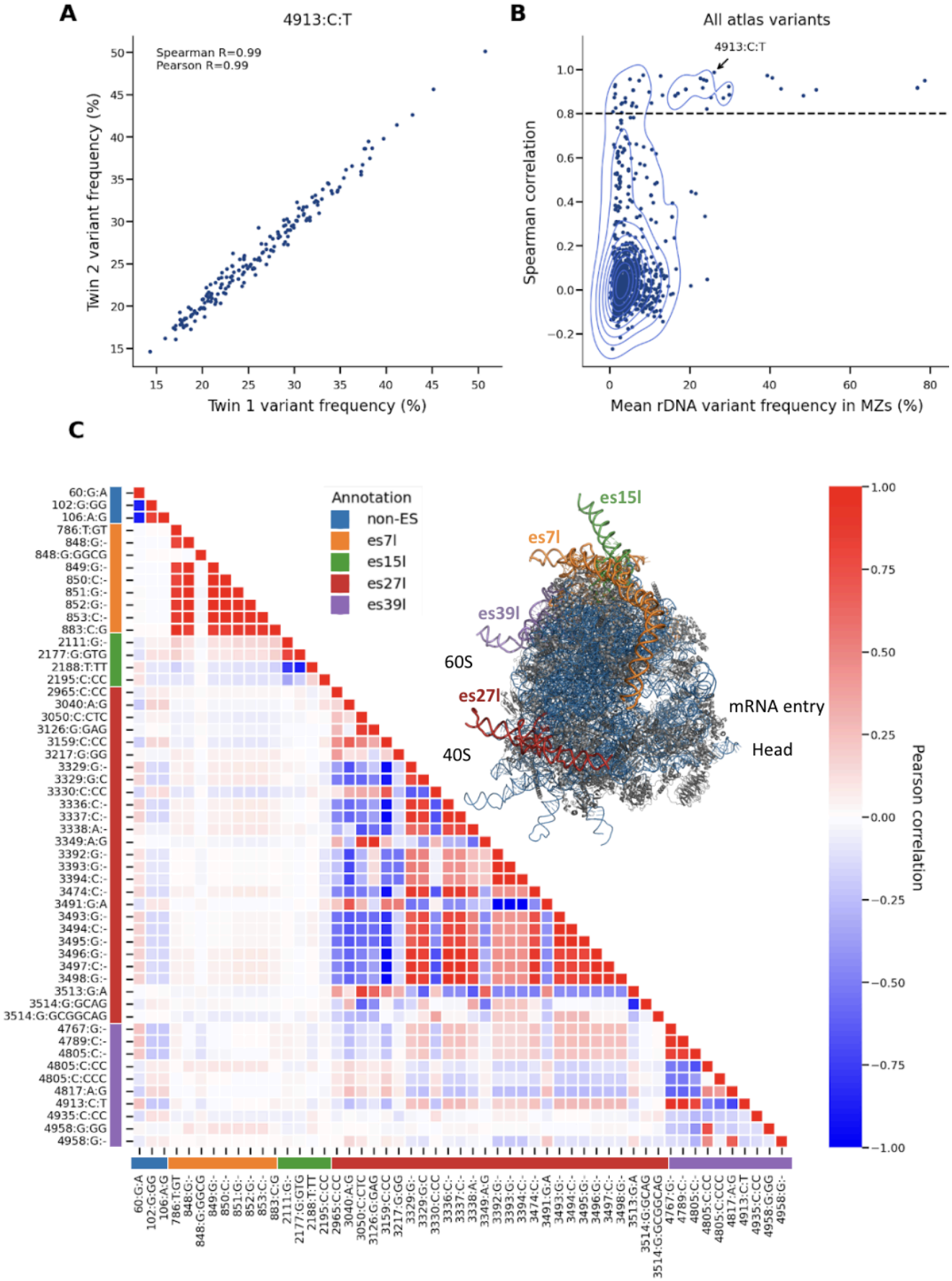
A
4913:C:T
Spearman R=0.99
Pearson R=0.99
Twin 2 variant frequency (%)
Twin 1 variant frequency (%)
B
All atlas variants
4913:C:T
Spearman correlation
Mean rDNA variant frequency in MZs (%)
C
Annotation
non-ES
es7l
es15l
es27l
es39l
60S
40S
mRNA entry
Head
Pearson correlation

Figure 2: rRNA variants include germline and somatic sequence variations
(A) Scatter plot showing the correlation of variant frequencies between monozygotic twins for the variant 4913:C:T of the 28S gene. The Spearman correlation coefficient (R = 0.99) and Pearson correlation coefficient are indicated, demonstrating high heritability for this specific variant.
(B) Contour plot illustrating the Spearman correlation of all atlas variants against their mean rDNA variant frequency in monozygotic twins. The dashed line indicates the threshold for significant heritability, with the variant 4913:C:T from panel (A) highlighted.
(C) Heatmap displaying high heritable variants (from panel B) correlation to one another. The color gradient represents the Pearson correlation values. A structure of a ribosome (PDB 4v6x [27]) with ES/non-ES region indicating the specific expansion segments (es7l, es15l, es27l, es39l) and non-ES variants in different colors. These colored annotations or ribosome regions are also annotated in the rows and columns of the heatmap.

Heritable Expansion Segment Variants Associate with Distinct Human Traits

To test for a causal role for rRNA variants on human traits, we performed an association study between the heritable variants and a wide range of human traits, including anthropometric traits, blood counts and diseases in the UKBB (**Table S2, S3, Methods**). Strikingly we observed many significant rRNA-variant-trait associations.

Blood calcium levels for example, showed no significant associations to any of the heritable rRNA variants (**Figure 3A**). In contrast, we found significant associations between ES variants and key physiological traits (**Figure 3B, 3C,** regression analysis P-value<0.05 after Bonferroni correction for the number of independent traits times the number of variants see **Methods**). Surprisingly, different variants belonging to different ESs were associated with the traits. Variants in es27l are associated significantly with IGF-1 levels, a crucial growth hormone involved in development and adult tissue maintenance [28]. In contrast, es15l and es39l variants are associated with standing height.

Importantly, Genome Wide Association Studies (GWAS) that aim to identify variants in the genome that affect human traits, mostly identify variants in non-coding parts of the genome [29]. These are usually interpreted as variants that are found in genomic proximity to the genes that have a causal effect on the focal trait of interest. However, unless these variants are located in known enhancer or promoter regions, they remain hard to interpret. In contrast, the tested atlas variants are coding variants, in the sense that they are part of a gene body as opposed to a genomic regulatory region. This means that variations in rRNA affect human traits. Furthermore, given that the linkage between ESs is relatively low (**Figure 2C**) this shows that different ESs have divergent effects on traits.

When taking a broader view of trait associations to rRNA variants, we observed clear clustering by trait category (**Figure 3D**, traits are hierarchically clustered by and annotated in yellow and teal). Variants in es15l were most strongly linked to traits related to adiposity, while variants in es39l showed the strongest association with body dimension traits, including standing height. Interestingly we also observed different

blood related traits that were associated with es27l, including platelet distribution width, and two disease associations; history of circulatory diseases to es27l and gastritis to es39l. This grouping highlights that es15l, es27l, and es39l perform unique functions. Interestingly, while two ESs, es15l and es39l are associated with some overlapping traits, es27l had strong association to other traits and es7l did not associate significantly to any trait. The overlap in es15l and es39l traits stem from es15l associating with both body dimension and adiposity related traits. Overall this suggests that es15l may have a broader ribosomal function in human health compared to more nuanced but important ribosome functions of es27l and es39l which affect distinct aspects of human health and disease.

It is important to note that the UKBB is a general population cohort, not a disease-focused one; therefore, the disease associations we identified likely represent an underestimation of the impact of rRNA variants on human disease. We conclude that rRNA variations as a whole display striking association with many human traits.

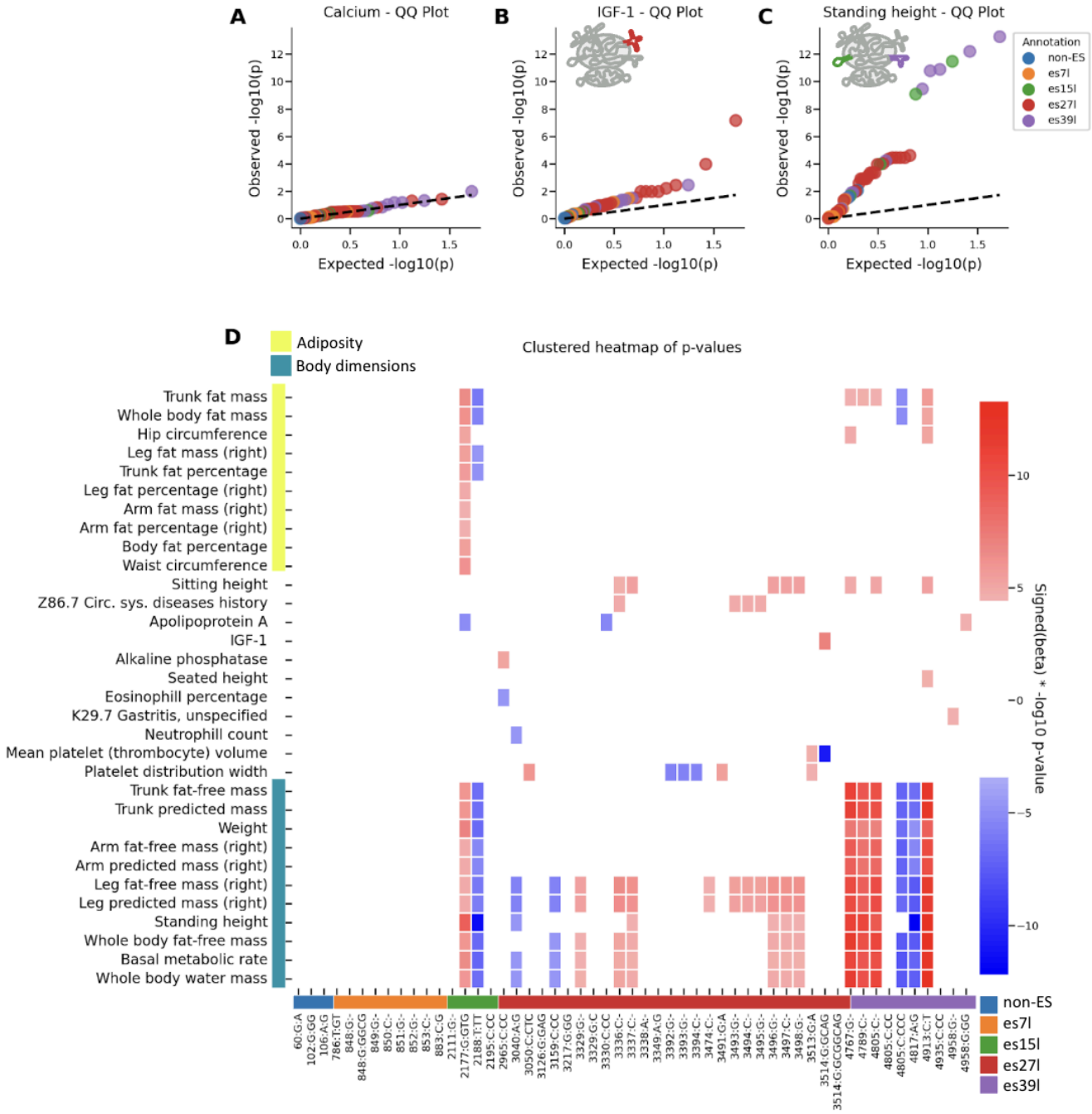

Figure 3: Expansion segments are associated with human traits

(A) Quantile-quantile (QQ) plot for the association of rRNA variants with blood calcium levels. The plot indicates no significant association, as evidenced by the distribution of observed versus expected p-values which follows the dashed line of random chance p-values.

(B) QQ plot showing the association of rRNA variants with IGF1 levels. The plot highlights significant associations, with expansion segments annotated in different colors corresponding to their respective ES. es27l is annotated in red on the ribosome cartoon.

(C) Same as (B) for standing height. es15l and es39l regions are annotated in green and purple on the ribosome cartoon.

(D) Clustered heatmap of p-values for the association of rRNA variants with various traits related to adiposity and body dimensions. Rows indicate traits and columns belong to variants. The rows are clustered and are annotated in yellow or teal if they belong to adiposity or body dimension related traits. The columns are annotated in color by their ES location. The heatmap visually represents the strength of

associations, with color gradients indicating significance levels for each trait. The colors red and blue indicate the sign of the linear regression beta where red indicates positive association with the trait and blue for negative association.

es15I and es39I Variants Exhibit Distinct Linkage Patterns to Acrocentric Chromosomes

In our previous study we found that variants were clustered in arrays and some were chromosome specific [8]. Here given that the UKBB has short-read data, we cannot map variants to their genomic location. Instead, we ran a GWAS analysis to associate rRNA variant frequencies to the rest of the genome. This allowed us to investigate the genomic origins of these heritable variants by analyzing their linkage to the five acrocentric chromosomes (13, 14, 15, 21, and 22).

As expected, all variants associated with the acrocentric chromosomes (**Figure 4**). When focusing on variants belonging to different ESs, we observed different linkage patterns. We focused on the top trait associated variants that belonged to different ESs, 2177:G:GTG from es15I, 3514:G:GCAG from es27I, and 4913:C:T from es39I. Interestingly, variants from es15I and es39I had a stronger association to one chromosome: The es15I variant showed the strongest linkage to chromosome 21 (**Figure 4A**) and the es39I variant was most associated with chromosome 14 (**Figure 4C**). In contrast, the es27I variant was strongly linked to multiple chromosomes 14, 15, 21, and 22 (**Figure 4B**). Given that es15I and es39I variants showed higher chromosome specificity, this could potentially point to regulation of adiposity and body dimension traits by regulating chromosome specific rRNA arrays.

Variant-chromosome specificity is interesting given the high propensity of translocations in acrocentric chromosomes that would homogenize the rRNA array across chromosomes. Specifically, Robertsonian translocations are the most common translocations in humans, happening in about 1 in every 1,000 newborns [30]. These translocations occur between the short arms of two acrocentric chromosomes. Here our data suggests that variant-chromosome specificity still occurs in the background of translocation events. However, since the UKBB has short-read data, we could not test if chromosome variants are clustered in rRNA arrays. Together, our data supports chromosome specificity of some variants while other variants are more evenly distributed across multiple chromosomes, consistent with the high tendency of translocation events in these chromosomes.

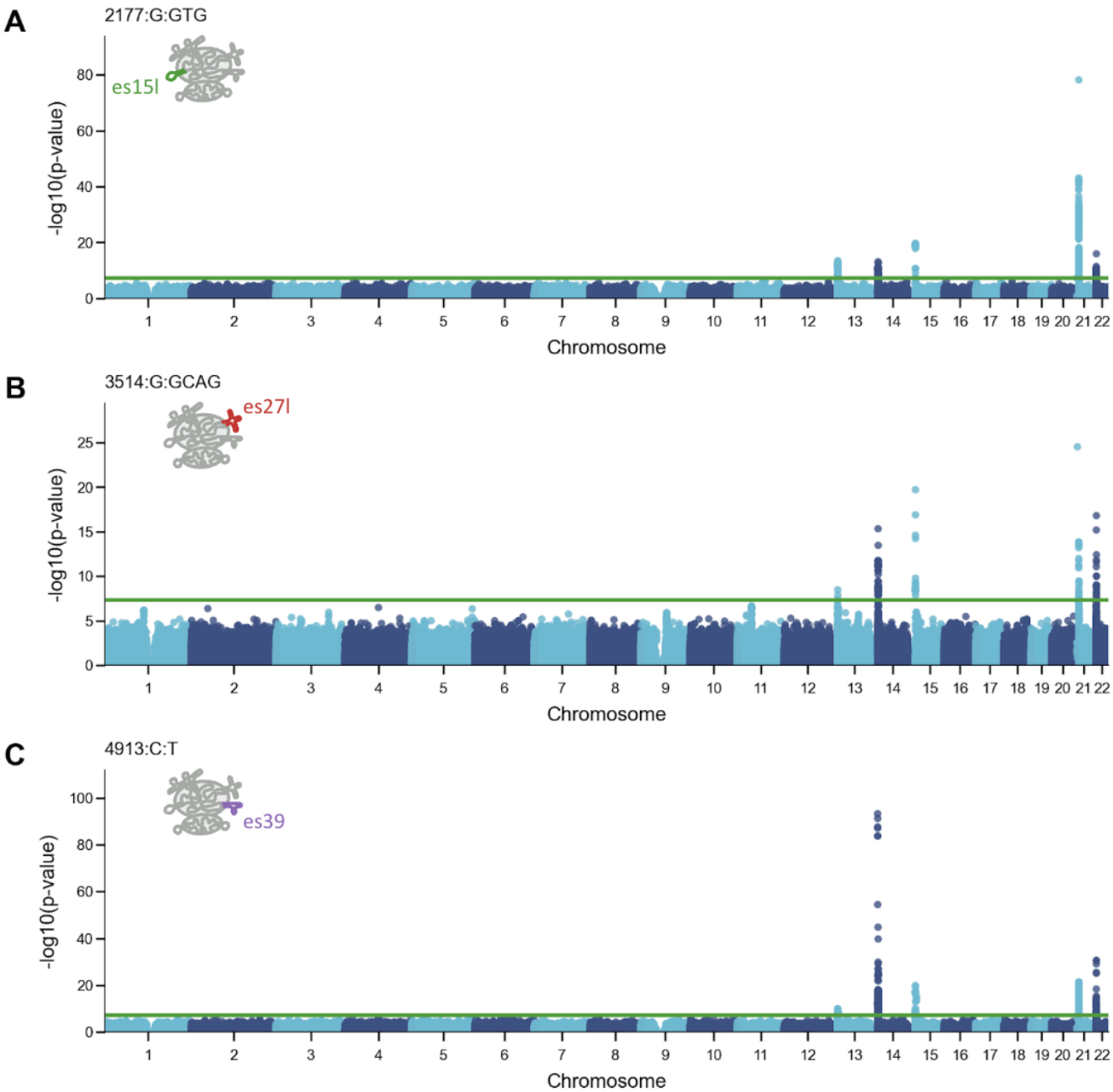

Figure 4: Varying genomic linkage between rRNA variants and acrocentric chromosomes
(A) Manhattan plot displaying the association of the top variant located in es15l (2177:G:GTG) with specific chromosomes. The y-axis represents -log10(p-value) for the association. es15l is annotated in green on the ribosome cartoon.
(B) Similar to (A) for the top variant located in es27l (3514:G:GCAG). es27l is annotated in red on the ribosome cartoon.
(C) Similar to (A) for the top variant located in es39l (4913:C:T). es39l is annotated in purple on the ribosome cartoon.

Discussion

This study establishes that rRNA variations located in ESs have clear phenotypic implications. While the precise mechanisms remain to be elucidated, our findings demonstrate that es15l influences both adiposity and body dimensions, es39l primarily affects body dimensions traits, and es27l impacts blood phenotypes and disease

susceptibility. We observed that several blood related traits such as platelet size are affected by variations in es27l. Interestingly there is accumulating literature that supports a critical role for ribosome function in hematopoiesis and that mutations in ribosome components in ribosomopathies lead to defects in hematopoiesis, including bone marrow failure [23,31–36]. These include Shwachman-Diamond syndrome affecting neutrophils, caused by a mutated ribosome biogenesis factor [35], Diamond Blackfan Anemia caused by mutations in ~20 ribosomal proteins that causes selective defects in erythroid progenitors [37], specific thrombocytopenia caused by mutation in SLFN14 that causes defects in rRNA regulation which affects platelets [32]. rRNA levels are also associated with hematopoiesis where hematopoietic transcription factors were shown to regulate rRNA transcription [33] and at the genome level copy number variations in rRNA affect hematological profiles [23]. Together this supports that ribosomes have broad roles in hematopoiesis, and future studies are needed to shed light on how variations in es27l affect hematopoiesis.

Additionally, we found variants in es15l and es39l that affect body dimension related traits, adiposity as well as linking these variants to specific chromosomes. This raises multiple interesting directions. In particular, in different ribosomopathies, patients have dwarfism [36,38,39] and it would be interesting to study if these diseases carry rRNA mutations that control body size. Moreover some variants in es15l and es39l are linked to specific chromosomes. It may be that there are specific transcription factors that bind specific rRNA arrays or that there is chromosome level regulation of rRNA transcription that is needed for metabolism or growth. There is a growing appreciation for tight metabolic control that is governed at the translation level [40]. ESs may expand our knowledge of how the ribosome itself can be linked to metabolic traits. The outstanding question is how different ESs function. The observed associations between the variations in the ES to different phenotype categories, i.e. body dimensions, adiposity and blood traits, raises the hypothesis that ESs show specificity as well as cooperativity between ESs. It may be that different ESs facilitate interactions with specific proteins, mRNAs, or even subcellular organelles that control translation. Moreover, ESs have also been shown to directly interact with each other, for example by direct tertiary RNA interactions [41] potentially working in concert for regulating important aspects of translation control linked to human physiology.

This study focused on heritable germline variants to establish a causal link to human traits. However somatic rRNA variations may also be critically important for human health. Building on previous work that associated rRNA variants and copy numbers with cancer [8,20,21,42–45], we hypothesize that rRNA variation contributes to oncogenesis through two potential mechanisms: germline variants acting as predisposition factors and the accumulation of somatic variants during an individual's lifetime. Given the low number of cancer cases in the UKBB cohort, we could not further investigate this here. Moreover, the UK Biobank is a population-based cohort, not one enriched for specific diseases. Consequently, our study was underpowered to detect associations between rRNA variants and disease states. Future investigations in dedicated disease-focused

cohorts will be necessary for uncovering the role of rRNA variation in conditions beyond what we could observe here.

Our finding that a majority of rRNA variants are likely somatic in origin opens the question of what generates these variations. There may be specific hotspots for mutagenesis within the rDNA repeats. As suggested in our previous work, this could include sequences prone to replication slippage, such as homopolymer tracts, which could explain the high prevalence of specific indels. Furthermore, the highly accessible chromatin state of active rDNA loci might render them more susceptible to DNA breaks and subsequent erroneous repair, contributing to the accumulation of somatic variants over time.

Together, using human genetics our findings highlight the relevance of ESs on the ribosome, supporting that they have key functions that affect human traits. This opens new directions for their study by highlighting which physiological processes, from control of body height to metabolism as well as others that may be guided by these unique rRNA regions associated with an expansion in ribosome size across evolution.

Data and code availability

In this publication we analyzed the whole genome sequencing and phenotypes of the UKBB cohort and the whole genome sequencing of the H7-hESC cell line. Both UKBB and H7-hESC are published datasets. The H7-hESC rDNA is available under BioProject ID PRJNA926787 under accession numbers SRR23196516 (H7-hESC whole genome sequencing). We generated an extended ES atlas with extended 150bp reference sequences as Extended Data to this publication (Data S1). This is the same ES atlas as we previously published [8] with 150bp flanking reference sequences to allow longer 150bp reads found in the UKBB dataset to map to the atlas. The RiboVAn pipeline is available at https://github.com/daphnar/rRNA/RiboVan

Methods

Datasets used in this study

The datasets analyzed in this study are the WGS and the genotype data available for 487,822 samples in the UKBB, and rDNA single nucleotide variant frequencies from the rRNA atlas, found in table Data S1 [8]. From the UKBB, monozygotic twins were calculated based on kinship greater than 0.4. For all regression analyses to control for population structure we used 402,610 unrelated White British individuals as our cohort, filtering based on sample QC characteristics as previously described in Sinnott-Armstrong et al., 2021[46].

Before extracting rRNA variant frequencies using the RiboVAn pipeline we first converted raw cram input files to fastq reads mapping to rRNA. For this we used the

following samtools [47] command as follows using bash command:

```
>> samtools view -C -o "${sample}.ribo" --reference "$reference" -M -L "$ribo_bed" -@ 20 "${cram_file}"
>> samtools fastq -c 9 -0 /dev/null --reference "$reference" "${sample}.ribo" -o "${sample}.ribo.fq.gz"
```

Here {sample} indicates the sample ID, {cram_file} is the input cram file and {reference} is the hg38 human genome reference fasta.

GWAS analyses were performed on genotyped data using the autosomal bi-allelic SNPs as described in Mostafavi et al., 2020 [48].

The analyses were run on the UKBB Research Analysis Platform.

Mapping UKBB WGS short-reads to the rRNA atlas

The UKBB WGS short-reads are 150 bases long and in order to map reads to the atlas we For step (2) in the RiboVAn we used an ES atlas with extended 150bp reference sequences that we generated as part of this paper, **Data S1**. To generate it we concatenated on the 5' and 3' end the rRNA reference sequence flanking the ES atlas [8].

Ribosome Analysis Variation (RiboVAn) pipeline

For every sample, for calculating rRNA variant frequencies we perform the following steps:

1. For every ES/non-ES region as defined in Table S17 in our previous study [8], we find all nucleotide variations associated with the ES/non-ES variant. We store this information as a lookup table that maps an ES variant to a list of all nucleotide variants found in this ES variant.
2. Using Bowtie2 [49] mapper we map short reads to the ES resolution atlas with flanking extended 150bp reference sequences (**Data S1**).
3. We count reads with perfect alignment to the ES resolution atlas variants.
4. Then we use the lookup table from step (1) and convert the ES resolution variant counts count to the nucleotide resolution variant counts.
5. We divide the nucleotide variant counts by the total number of reads mapped to the ES/non-ES region to get the relative abundances of nucleotide variants.

The RiboVAn pipeline is available at https://github.com/daphnar/rRNA/RiboVan

Regression Analysis

We associated rRNA variants to traits. For quantitative traits, we ran an ordinary least square linear regression model and for diseases we ran a logistic regression model (using python's statsmodels.api package). In the linear regressions, the number of individuals was dependent on the number of individuals having the quantitative trait.

The UKBB includes hundreds of ICD10 codes coding for diseases, but the UKBB is a general population cohort and is not a targeted disease cohort. In order to minimize tests, we focused on ICD10 codes with at least 30,000 cases. Moreover, to correct for multiple hypotheses we applied Bonferroni correction using the number of rRNA variants tested.

For logistic regressions, rRNA variants and covariates were normalized to one (by subtracting the mean and dividing by the standard deviation). The statsmodels.api.Logit models for some variant-traits pairs failed to converge in which case we did not consider the result.

In the regression models we control for age, sex, age times sex, age squared, rDNA copy number, population structure and batches. Specifically, we estimate rDNA copy number by the fraction of reads mapped to the rDNA out of total reads as described previously [23]. On top of using the subset of the UKBB of unrelated White British individuals, we add the 20 first genetic principal components to the model to account for population structure. The categorical batches of “Genotype measurement batch000”, “Release tranche050”, and “Shipment batch number053” were added to the model using one hot encoding.

GWAS analysis

For associating variant frequencies with the rest of the genome we used plink2 [50] with the following bash command for every chromosome-variant pair. The variant_id file stores variant frequencies for a selected rRNA variant and the rest of the chromosome files contain genotyped data.

```
phenofile="${variant_id}.txt"
pgenfile="chr${chromosome}.pgen"
psamfile="chr${chromosome}.psam"
pvarfile="chr${chromosome}.pvar"
famfile="chr${chromosome}.fam"
bimfile="chr${chromosome}.bim"
covarfile="ukbb_20pc.txt"
plink2 --pgen "$pgenfile" --bim "$bimfile" --fam "$famfile" --pheno "$phenofile" --covar "$covarfile" --glm hide-covar --covar-variance-standardize --out "$outfile" --threads 16 --vif 999
```

Acknowledgments

We thank Yuxiang Chen from the Barna lab for creating the video and caption of the ribosome structure. We thank the Barna and Pritchard group members for discussions. M.B. is supported by the National Institutes of Health grant R01HD086634. J.K.P. and D.R are supported by RO1 HG014005 and HG008140.

Author contributions

D.R. conceived and directed the project, designed and conducted all computational analyses, interpreted the results, and wrote the manuscript. M.B. and J.K.P. conceived and directed the project and analyses, designed the analyses, interpreted the results, and wrote the manuscript.

Declaration of interests

All authors declared no competing interests.

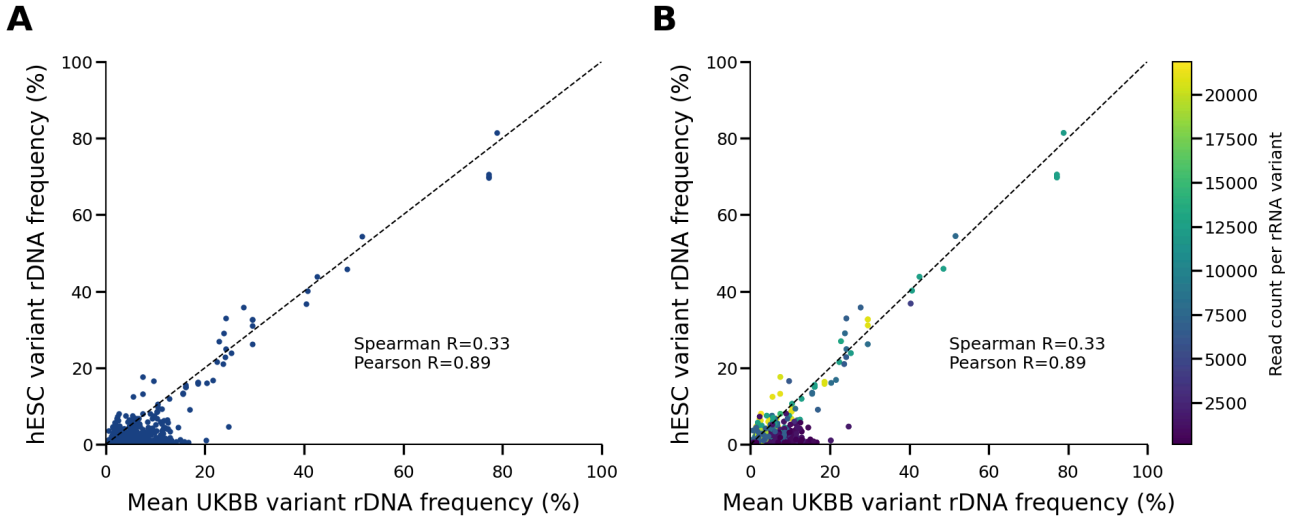

Figure S1: rRNA variants frequencies in hESC cell line and average UKBB profile agree for high coverage sequencing variants
(A) Scatter plot comparing the genomic rRNA variant profile of two datasets: mean UKBB and the hESC cell-line. Spearman and Pearson correlations are presented.
(B) Similar to (A) but variant frequencies are colored by the number of reads that were mapped to the rRNA variant. A color bar indicating read count is presented.

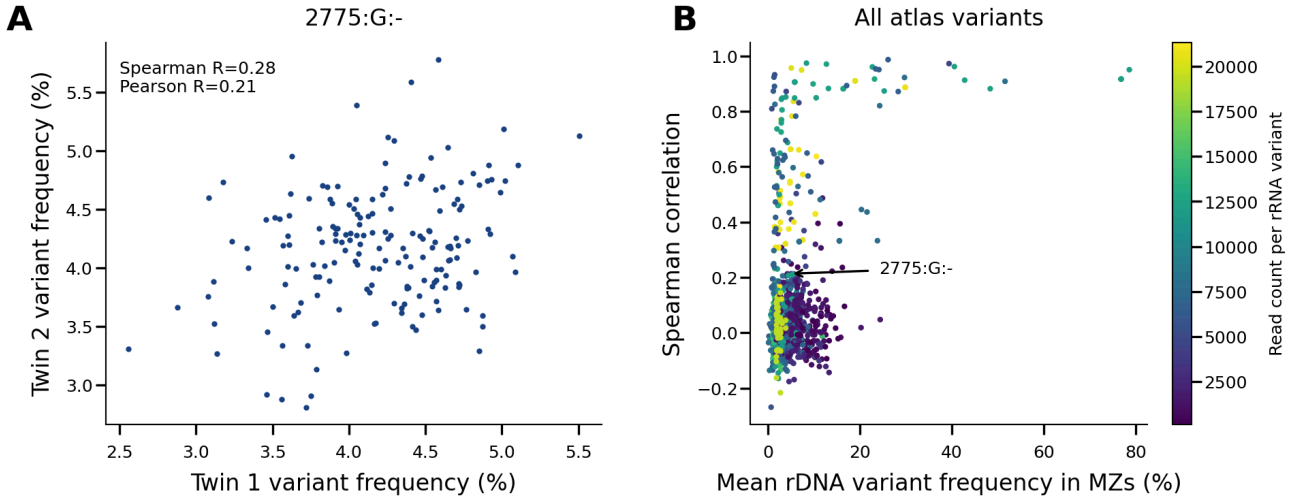

Figure S2: rRNA variants contain somatic sequence variations

(A) Scatter plot showing the correlation of variant frequencies between monozygotic twins for the variant 2775:C:T of the 28S gene. The Spearman correlation coefficient (R = 0.28) and Pearson correlation coefficient are indicated, demonstrating low heritability for this specific variant.

(B) Scatter plot showing the Spearman correlations of all atlas variants against their mean rDNA variant frequency in monozygotic twins. The Spearman correlation for variant frequencies are colored by the number of reads that were mapped to the rRNA variant. A color bar indicating the mean read count per sample is presented. The variant 2775:C:T from panel (A) highlighted.

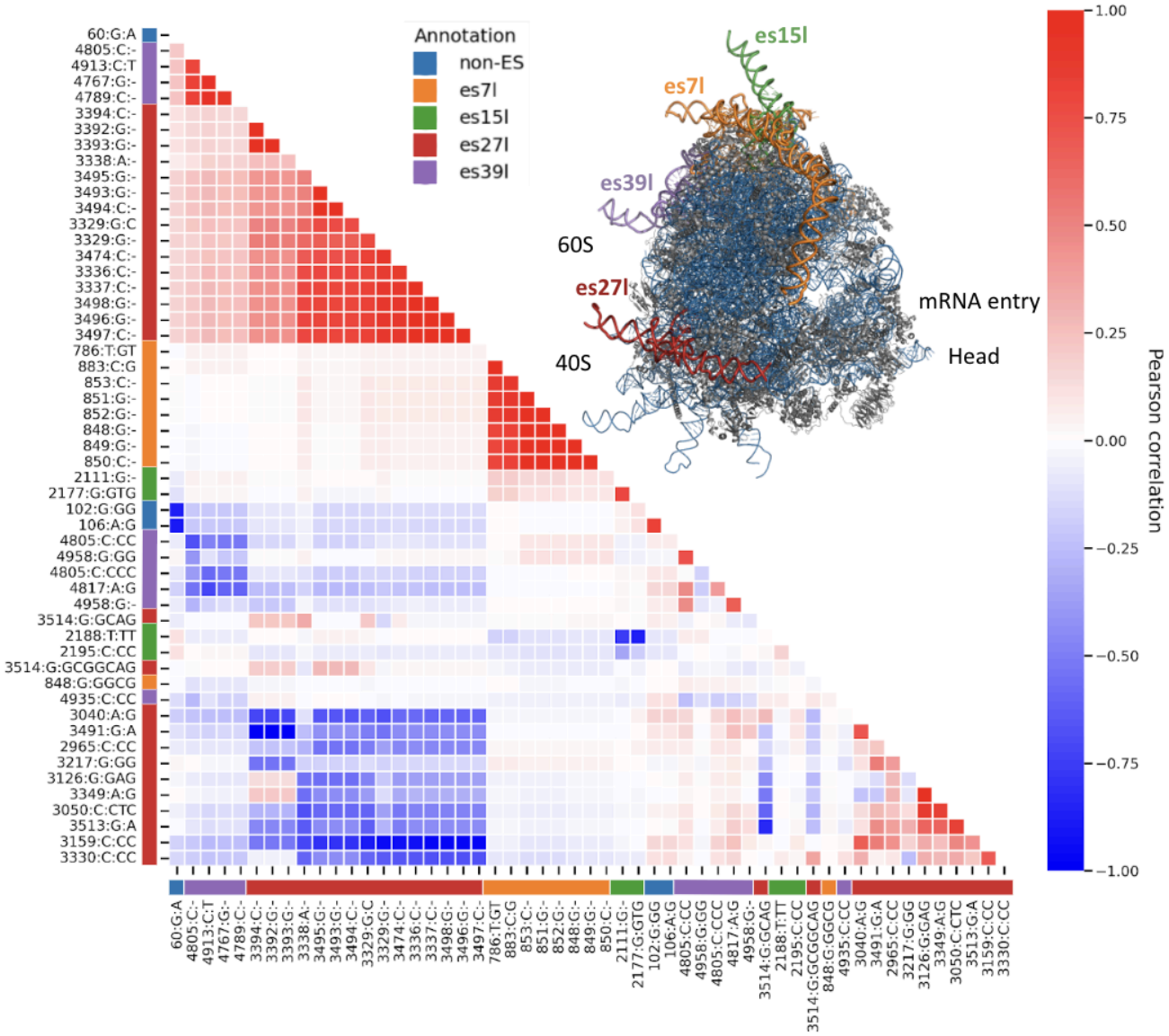

Figure S3: rRNA variants shows some haplotype linkage between rRNA regions

Heatmap displaying high heritable variants correlation to one another after hierarchical clustering. The color gradient represents the Pearson correlation values. A structure of a ribosome (PDB 4v6x [27]) with ES/non-ES region indicating the specific expansion segments (es7l, es15l, es27l, es39l) and non-ES variants in different colors. These colored annotations or ribosome regions are also annotated in the rows and columns of the heatmap.

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. Munafó, D.B., Langhorst, B.W., Chater, C.L., Sumner, C.J., Rodríguez, D.N., Russello, S., Gardner, A.F., Slatko, B.E., Stewart, F.J., Sinicropi, D., et al. (2016). Selective Depletion of Abundant RNAs to Enable Transcriptome Analysis of Low-Input and Highly Degraded Human RNA. Curr. Protoc. Mol. Biol. *113*, 7.22.1-7.22.9. 10.1002/0471142727.mb0722s113.
3. Genuth, N.R., and Barna, M. (2018). The discovery of ribosome heterogeneity and its implications for gene regulation and organismal life. Mol. Cell *71*, 364–374. 10.1016/j.molcel.2018.07.018.
4. 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.
5. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. 10.1038/nature15393.
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. 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.
8. Rothschild, D., Susanto, T.T., Sui, X., Spence, J.P., Rangan, R., Genuth, N.R., Sinnott-Armstrong, N., Wang, X., Pritchard, J.K., and Barna, M. (2024). Diversity of ribosomes at the level of rRNA variation associated with human health and disease. Cell Genomics *4*, 100629. 10.1016/j.xgen.2024.100629.
9. Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H., and Turner, D.J. (2008). A large genome center's improvements to the Illumina sequencing system. Nat. Methods *5*, 1005–1010. 10.1038/nmeth.1270.
10. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585. 10.1038/ng.2653.
11. 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.
12. 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.
13. 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.
14. 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.

15. Nagai, R., Milam, O.L., Niwa, T., Howell, W.J., Best, J.A., Yoshida, H., Freeburg, C.D., Koomen, J.M., and Fujii, K. (2025). Ribosomal expansion segment contributes to translation fidelity via N-terminal processing of ribosomal proteins. Nucleic Acids Res. *53*. 10.1093/nar/gkaf448.
16. Leppek, K., Byeon, G.W., Fujii, K., and Barna, M. (2021). VELCRO-IP RNA-seq reveals ribosome expansion segment function in translation genome-wide. Cell Rep. *34*, 108629. 10.1016/j.celrep.2020.108629.
17. Leppek, K., Fujii, K., Quade, N., Susanto, T.T., Boehringer, D., Lenarčič, T., Xue, S., Genuth, N.R., Ban, N., and Barna, M. (2020). Gene- and Species-Specific Hox mRNA Translation by Ribosome Expansion Segments. Mol. Cell *80*, 980-995.e13. 10.1016/j.molcel.2020.10.023.
18. Venturi, G., Zacchini, F., Ruzzi, F., Corradini, A.G., Serra, M., Penzo, M., Treré, D., Lollini, P.-L., and Montanaro, L. (2025). 5.8S rRNA forms and ribosome heterogeneity in breast cancer. Biochimie. 10.1016/j.biochi.2025.05.010.
19. Law, P.P., Mikheeva, L.A., Rodriguez-Algarra, F., Asenius, F., Gregori, M., Seaborne, R.A.E., Yildizoglu, S., Miller, J.R.C., Tummala, H., Mesnage, R., et al. (2024). Ribosomal DNA copy number is associated with body mass in humans and other mammals. Nat. Commun. *15*, 5006. 10.1038/s41467-024-49397-5.
20. 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.
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. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. 10.1038/s41586-018-0579-z.
23. Rodriguez-Algarra, F., Evans, D.M., and Rakyan, V.K. (2024). Ribosomal DNA copy number variation associates with hematological profiles and renal function in the UK Biobank. Cell Genomics *4*, 100562. 10.1016/j.xgen.2024.100562.
24. Rodriguez-Algarra, F., Cooper, M., Mardakheh, F.K., Evans, D.M., and Rakyan, V.K. (2025). Germline sequence variation within the ribosomal DNA is associated with human complex traits. BioRxiv. 10.1101/2025.02.06.635840.
25. Liu, X., Li, Y.I., and Pritchard, J.K. (2019). Trans effects on gene expression can drive omnigenic inheritance. Cell *177*, 1022-1034.e6. 10.1016/j.cell.2019.04.014.
26. Smith, G.D., and Ebrahim, S. (2003). "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. *32*, 1–22. 10.1093/ije/dyg070.
27. Anger, A.M., Armache, J.-P., Berninghausen, O., Habeck, M., Subklewe, M., Wilson, D.N., and Beckmann, R. (2013). Structures of the human and Drosophila 80S ribosome. Nature *497*, 80–85. 10.1038/nature12104.
28. Laron, Z. (2001). Insulin-like growth factor 1 (IGF-1): a growth hormone. MP, Mol Pathol *54*, 311–316. 10.1136/mp.54.5.311.

29. Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. *51*, 1339–1348. 10.1038/s41588-019-0481-0.

30. Lubs, H.A., and Ruddle, F.H. (1970). Chromosomal abnormalities in the human population: estimation of rates based on New Haven newborn study. Science *169*, 495–497. 10.1126/science.169.3944.495.

31. Barlow, J.L., Drynan, L.F., Trim, N.L., Erber, W.N., Warren, A.J., and McKenzie, A.N.J. (2010). New insights into 5q- syndrome as a ribosomopathy. Cell Cycle *9*, 4286–4293. 10.4161/cc.9.21.13742.

32. Ver Donck, F., Ramaekers, K., Thys, C., Van Laer, C., Peerlinck, K., Van Geet, C., Eto, K., Labarque, V., and Freson, K. (2023). Ribosome dysfunction underlies SLFN14-related thrombocytopenia. Blood *141*, 2261–2274. 10.1182/blood.2022017712.

33. Paralkar, V.R. (2024). Transcription factor regulation of ribosomal RNA in hematopoiesis. Curr. Opin. Hematol. *31*, 199–206. 10.1097/MOH.0000000000000816.

34. Da Costa, L.M., Marie, I., and Leblanc, T.M. (2021). Diamond-Blackfan anemia. Hematology Am Soc Hematol Educ Program *2021*, 353–360. 10.1182/hematology.2021000314.

35. Ganapathi, K.A., Austin, K.M., Lee, C.-S., Dias, A., Malsch, M.M., Reed, R., and Shimamura, A. (2007). The human Shwachman-Diamond syndrome protein, SBDS, associates with ribosomal RNA. Blood *110*, 1458–1465. 10.1182/blood-2007-02-075184.

36. Narla, A., and Ebert, B.L. (2010). Ribosomopathies: human disorders of ribosome dysfunction. Blood *115*, 3196–3205. 10.1182/blood-2009-10-178129.

37. Iskander, D., Wang, G., Heuston, E.F., Christodoulidou, C., Psaila, B., Ponnusamy, K., Ren, H., Mokhtari, Z., Robinson, M., Chaidos, A., et al. (2021). Single-cell profiling of human bone marrow progenitors reveals mechanisms of failing erythropoiesis in Diamond-Blackfan anemia. Sci. Transl. Med. *13*, eabf0113. 10.1126/scitranslmed.abf0113.

38. Mckusick, V.A., Eldridge, R., Hostetler, J.A., Ruangwit, U., and Egeland, J.A. (1965). Dwarfism in the amish. ii. cartilage-hair hypoplasia. Bull. Johns Hopkins Hosp. *116*, 285–326.

39. Trainor, P.A., and Merrill, A.E. (2014). Ribosome biogenesis in skeletal development and the pathogenesis of skeletal disorders. Biochim. Biophys. Acta *1842*, 769–778. 10.1016/j.bbadis.2013.11.010.

40. Biffo, S., Ruggero, D., and Santoro, M.M. (2024). The crosstalk between metabolism and translation. Cell Metab. *36*, 1945–1962. 10.1016/j.cmet.2024.07.022.

41. Alkemar, G., and Nygård, O. (2004). Secondary structure of two regions in expansion segments ES3 and ES6 with the potential of forming a tertiary interaction in eukaryotic 40S ribosomal subunits. RNA *10*, 403–411. 10.1261/rna.5135204.

42. 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.

43. 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.

44. 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.

45. 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.

46. Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G.R., Wainberg, M., Ollila, H.M., Kiiskinen, T., et al. (2021). Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat. Genet. *53*, 185–194. 10.1038/s41588-020-00757-z.

47. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.

48. Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J.K., and Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. eLife *9*. 10.7554/eLife.48376.

49. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. 10.1038/nmeth.1923.

50. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, s13742–015–0047–8. 10.1186/s13742-015-0047-8.