ESCI upf.
School of International Studies

MEMÒRIA DEL TREBALL DE FI DE GRAU DEL GRAU (ESCI-UPF)

Developing a Variant Calling Pipeline for Ribosomal DNA

AUTOR/A: Alba Jiménez Lupiáñez	**NIA:** 106466
GRAU: Bachelor's Degree in Bioinformatics	

CURS ACADÈMIC: 2023-2024

DATA: 18/06/2024

TUTOR/S: José Miguel Ramírez and Marta Melé

FULL DE RESUM DEL TREBALL DE FI DE GRAU DEL BDBI (ESCI-UPF)

TÍTOL DEL PROJECTE: Developing a Variant Calling Pipeline for Ribosomal DNA

AUTOR/A: Alba Jiménez Lupiáñez | **NIA:** 106466

CURS ACADÈMIC: 2023-2024

DATA: 18/06/2024

TUTOR/S: José Miguel Ramírez and Marta Melé

PARAULES CLAU (mínim 3)

- Català: ADN ribosómic, detección de variantes, evaluació de referència.
- Castellà: ADN ribosómico, detección de variantes, evaluación de referencia.
- Anglès: Ribosomal DNA, Variant calling, benchmark analysis.

RESUM DEL PROJECTE (extensió màxima: 100 paraules per llengua)

- Català:

L'ADN ribosòmic (rADN) és vital degut al seu paper en la traducció de proteïnes, però l'alt contingut de GC i la repetició de les regions on estan codificats fan que sigui difícil seqüenciar-lo i analitzar-lo. En aquest estudi s'ha desenvolupat un preocés per analitzar la variabilitat de l'rADN, provat amb lectures simulades a causa de la disponibilitat limitada de dades d'rADN. Amb aquest procés s'ha aconseguit una alta puntuació F1 de 0.9358 validada amb lectures de seqüenciació del genoma complet de l'assemblatge T2T-CHM13.

- Castellà:

El ADN ribosómico (rADN) es vital debido a su papel en la traducción de proteínas; sin embargo, el alto contenido de GC y la repetitividad de las regiones donde están codificados hacen que sea difícil secuenciarlas y analizarlas. En este estudio se ha desarrollado una procedimiento para analizar la variabilidad del rADN, probada con lecturas simuladas debido a la limitada disponibilidad de datos de rADN. Con el procedimiento se ha logrado una alta puntuación F1 de 0.9358 validada con lecturas de secuenciación del genoma completo del ensamblaje T2T-CHM13.

- Anglès:

Ribosomal DNA (rDNA) is vital due to its role in protein translation, however, the high GC content and repetitiveness of the regions where they are encoded make it difficult to sequence and analyze. This study developed a pipeline to analyze rDNA variability, tested on simulated reads due to limited rDNA data. The pipeline achieved a high F1 score of 0.9358 and was validated with whole genome sequencing reads from the T2T-CHM13 assembly.

Developing a Variant Calling Pipeline for Ribosomal DNA

Alba Jiménez Lupiáñez

Scientific directors: Marta Mele[1], Jose Miguel Ramírez[1]

[1]Transcriptomics and Functional Genomics Lab, Barcelona Supercomputing Center, Plaça d'Eusebi Güell, 1-3, 08034 Barcelona

Abstract

The study of ribosomal DNA (rDNA), is crucial given to the ribosome's essential role in protein translation. However, the high GC content and repetitive nature of the regions where rDNA is encoded challenge their sequencing, assembling, and variant characterization. The release of the first complete Telomere-to-Telomere assembly (T2T-CHM13) marks significant milestones in resolving the gaps and inaccuracies of the human reference genome. Particularly, these assemblies provided the first comprehensive reference sequences for rDNA regions, which are crucial for understanding ribosome composition and its implications in health and disease.

This study develops and benchmarks a variant-calling pipeline to analyze rDNA variability. The little available rDNA data limits the amount of testing that can be done, therefore we tested the pipeline in two sets of simulated reads. The first was meant to identify any conflict raised in a perfect case scenario. The second one, containing realistic rDNA and sequencing variability was used to fine-tune the parameters to optimize the precision and recall. In the end, we obtained a maximum F1 score of 0.9358. We validated the procedure pipeline using whole genome sequencing (WGS) reads of the T2T-CHM13 assembly.

Introduction

Ribosomes are macromolecule complexes in charge of translating mRNAs into proteins. They are composed of two subunits, each constituted by ribosomal RNA (rRNA) and proteins. The small subunit (40S) contains the 18S rRNA and 33 proteins, and the large subunit (60S) contains the rRNAs 25S, 5.8S, and 5S along with 49 proteins [1].

Over 80% of all cellular RNA corresponds to rRNAs [2]. In humans, 18S, 5.8S, and 28S rRNAs are encoded in an operon, a group of genes that are transcribed together, called 45S (Figure 1A). The 45S operon is concealed in near-identical repeats separated by intergenic spacers (IGS) arranged in large tandem repeat arrays. These large tandem are embedded within the short arms of the acrocentric chromosomes (13, 14, 15, 21, and 22) [3]. The 5S rRNA gene is found separately, in repeats between IGSs, in chromosome 1. The total number of rDNA copies varies between individuals, ranging from tens to thousands of copies [4, 5]. Each copy of the 45S operon consists of an external transcribed spacer (5'ETS), the 18S rRNA gene, the internal transcribed spacer 1 (ITS1), the 5.8S rRNA gene, the internal transcribed spacer 2 (ITS2), the 28S rRNA gene and an external transcribed spacer (3' ETS). The copies are separated by intergenic spacers (IGS) about 30 kb long, in total conforming rDNA units about 45kb long. The operons are transcribed in bulk by an RNA polymerase I [6], each copy resulting in a 13.4 kb 45S long precursor rRNA (pre-rRNA) [2], and the 5S rDNA is transcribed separately by RNA polymerase III. After transcription, the 45S pre-rRNA is processed, losing the transcribed spacers, and being slit into pre-60S and pre-40S particles. In this step, the 5S transcript is incorporated into the pre-60S subunit. After maturation occurs in the cytoplasm, both mature rRNAs are assembled into several ribosomal proteins to form the two subunits constituting the ribosome [1].

Originally, ribosomes were thought to be homogenous (Moore et al. 1968), but studies consistently show variability within ribosome composition among tissues [7, 8]. Given the essential role of ribosomes in protein synthesis, there is a major interest in understanding rDNA coding variation and rRNA expression.

It is known that rDNA copy number varies between chromosomes within each individual, between individuals within the same population, and between ancestry backgrounds [9], nevertheless, rDNA copy number alterations have also been linked to some diseases. Within the tandem repeats, the IGS regions accumulate variation at a much higher rate than the transcribed regions [10], although, variants in the 18S and 28S rRNA sequences have also been identified. Despite most being in expansion segments, which may affect ribosomal structure and flexibility [11], a few of these variants have been linked to functionally conserved regions modulating RNA folding and RNA-RNA interactions [11]. Moreover, there is evidence of functionally distinct ribosomes [9, 12], which exhibit stratification across various tissues [13, 14]. This functionality divergence across tissues has only been linked to ribosomal proteins [8] but the potential effects of rRNA variation on ribosomal functionality have only been theorized.

The characterization of variants in rDNA has posed a significant challenge due to several impediments. On the one hand, several second-generation sequencing limitations challenge the assembly of the tandem arrays given the high GC content and the repetitiveness of these regions[15,

16, 17]. On the other hand, the characterization of these regions was challenged by the absence of a high-quality reference rDNA sequence.

Up until recently, the GRCh38 was the highest-quality, most complete genome assembly. However, despite continuous improvements since 2013 [18], 8% of the genome remained unresolved due to technological limitations struggling to correctly assemble repetitive regions. This corresponded to 151 million unknown base pairs, which included centromeric and subtelomeric regions, segmental duplications, and ribosomal DNA (rDNA) [19]. This left a gap filled in 2018 with the release of the KY962518.1 rDNA sequence [5]. The sequence corresponds to one rDNA copy from a single individual. This reference has allowed the study of rRNA variation across tissues, which has suggested that rDNA alleles can be expressed differently across tissues [17]. However, previous studies doing rDNA variant calling have never benchmarked their pipeline [17, 24].

Ultimately, the release of the Telomere-to-Telomere assemblies, T2T-CHM13 [19], and T2T-YAO [20] allowed, for the first time, to obtain complete human genomes. They managed to resolve the missing regions by combining high-quality PacBio HiFi with ultra-long reads from Oxford Nanopore Technology (ONT). Notably, the T2Ts assemblies offer resolved continuous rDNA regions to be studied.

Objectives

The recent publications of the rDNA sequences in the T2T-CHM13 individual allow the benchmarking of a pipeline to the variability within 45S rDNA regions. However, this is still challenging and due to the little WGS data currently available, limiting the amount of testing that can be done, no previous study on rDNA variant calling has ever benchmarked it [9]. To tackle this, we first simulated different sequences to obtain artificial rDNA sequencing reads samples to adjust and analyze the performance of a variant calling. We then performed a benchmark to optimize the different parameter configurations used in the variant calling. Finally, we validated the pipeline with the real data from the T2T-CHM13 assembly.

Methods

Simulating reads with no sequencing errors

We analyzed the performance of the variant calling pipeline for two different sets of simulated reads. The first was simulated reads of the 24 unique 45S rDNA copies of the T2T-CHM13 assembly (defined in the following method section ***Evaluation***) assigned with the highest Illumina base quality, thus being referred to as "perfect reads". The lack of sequencing errors was used to evaluate the performance of the pipeline in a perfect case scenario in which there is no prospect of sequencing errors being classified as variants. The set was generated using a Python script which, given a FASTA input file, generates FASTQ files of paired-end reads of specified length and insert size. We generated three sets of perfect reads with 151 paired-end length and coverages 30x, 100x, and 1000x.

Callable regions

Due to the high GC content and repetitiveness of the rDNA regions, there are areas whose variants may be harder to detect. On one hand, the repetitiveness may lead to multimapping reads [17], which could result in false positives. On the other hand, certain regions within the internal and external transcribed spacers and in the 28S possess GC content over 80% (Figure S1), which could lead to false negatives due to low coverage. For the characterization of these regions, we used RepeatMasker [30].

Simulating reads with sequencing errors

The second way of obtaining WGS data was the creation of realistic reads through a two-step process. First, generating samples with genomic variability and then, simulating second-generation reads of these samples to introduce sequencing errors to the data (Figure 1B). For the first step, we used a *simulator* previously developed by the group [https://github.com/Rbbt-Workflows/SyntheticRDNA], which takes a sequence and produces a selected number of samples with sequence variability. It begins by creating a variant catalog by introducing an input number of insertions, deletions, and single nucleotide polymorphisms (SNPs) to an input reference sequence. Then, an input-defined number of model sequences are developed by introducing a chosen number of polymorphisms from the variant catalog to a reference sequence. Sourcing variants from a common catalog ensures that the resulting model sequences have shared variability. Finally, to create each sample, a random number of the model sequences are selected, each with a varying copy number. For the second step, we generated artificial second-generation sequencing reads of the sequences for each sample using NEAT, therefore introducing sequencing variability.

We ran a simulator using the ChrR sequence as a reference, which consists of the KY962518.1 rDNA sequence with the final third of the IGS, containing enhancers and promoters, shifted to the front, before the 5' ETS [2]. The input variables were adjusted to match the T2T-CHM13 statistics as much as possible. We created 100 artificial samples with 50 sequences each. Then, we used NEAT [27] to generate 151 paired-end DNA sequencing reads and 36x depth for each of the sequences of each of the samples' FASTA files. We also created short reads for the sequences of "rDNA-like" regions in the GRCh38 genome. These regions constitute the masked regions of the GRCh38 as defined in George et al. and adding them allowed us to have a pool of rDNA reads with realistic genomic variability. In the end, we obtained a FASTQ of simulated rDNA sequencing reads that contained variability and sequencing errors, just like real WGS data would.

Preprocessing step of the variant calling pipeline

Once we obtained both the perfect and simulated reads, we processed them using a pipeline previously developed by the team to call variants with WGS and RNA-seq data (Figure 1C). Our input data at this point was treated as real WGS data, which contained reads of 45S rDNA regions and rDNA-like reads, including pseudogenes and incomplete fractions of rDNA arrays. Consequently, the first step is to isolate the rDNA reads.

To obtain the reads coming from rDNA sequences, we first needed to identify candidate rDNA-like reads and remove those with less than 20% identity with the reference rDNA sequence. For this purpose, we used a MUMmer [22] aligner called NUCMER (NUCleotide MUMmer). In the first step of the alignment, it finds the longest possible unique subsequences (MUMs) for both the reference and

the aligned input sequences. Pairwise comparisons are made between the two sets of MUMs and matches are sorted by similarity, allowing us to set a threshold on the identity level between the pairs. Setting a minimum number of matches for the pairs to be considered increases the specificity, as it limits the chances of false positive matches. For this step, we used the ChrR sequence as a reference with a 500bp padding and we set the threshold on 30 nucleotides using the parameter "maxmatch -l". This way, we considered only the reads that matched over 80% to rDNA regions. Once we identified the rDNA reads, we used Seqkit [23] to extract their sequences and discard the rDNA-like reads.

The second step of the pipeline consisted of removing reads that would map better to any other genomic region. To do this, we mapped the isolated rDNA reads to the reference, T2T with rDNA regions masked and ChrR at the end. We used Burrows-Wheeler Aligner-Maximal Exact Match (BWA-MEM) [24], a popular tool to align sequencing reads to a reference genome, which allowed us to add a read group tag for each of the sequences of the resulting BAM file. This was useful to later be able to discern sequences coming from different samples (in the case of the simulated reads) or unique sequences (in the case of the perfect sequences).

In the third step, we used samtools [25] to filter out reads that were not uniquely mapped, which were tagged as alternative alignments (with the tag "*XA*"). The discarded reads had multiple potential alignment positions and since the rDNA tandem arrays were masked, they were mapped to rDNA pseudogenes. If pseudogenes had been considered for the variant calling, they would have mapped to the rDNA region and their variability would lead to false positives.

Variant Calling Pipeline

The fourth step consisted of the variant calling of the rescued uniquely mapped rDNA reads. First, we used the HaplotypeCaller from Genome Analysis Toolkit (GATK) [26] to perform variant calling for each of the reads sets and generated a genotype variant call format (gVCF) file for each of the samples. We used the complete chrR as the reference sequence, based on the results of the perfect reads (section results), we decided to only consider the callable regions (section methods) for the simulated reads. The resulting files contained the variant calling evaluation of all positions, independent of whether a variant was found in them.

We tested two versions of the pipelines diverging on the fifth step. The first version first uses GenomicsDBImport which takes all samples' gVCF files and constructs a database. Then, performs join-genotyping using GenotypeGVCFs followed by hard filtering of the VCF file (see methods). The second version performs hard filtering of the gVCF file based on the analysis of the evaluation of the output files of the Haplotype caller (see methods). Either way, in the end, the resulting VCF file contains a unified analysis report on the positions where variants were detected for each of the samples.

By default, the variant caller assumes that the organism from which the samples come from is diploid, however, the *--sample-ploidy* parameter can be modified as if the organism had more than two sets of each chromosome. For the three sets of perfect reads of depth 30x, 100x, and 100x, we ran the variant calling with ploidy 5. Using a low ploidy may lead to variants with very low allele depth being undetected and resulting in false negatives. Intending to reduce the number of false

negatives, we tested the performance of the variant caller when considering ploidy 2, 5, 7, 10, 20, 30, 40, 50, and 100. Based on the results we focused on ploidy 2, 5, 7, 10, 50, and 100 for the second version of the pipeline. We would expect that the higher the ploidy value, the more low-frequency variants are considered, and therefore, the more the sensitivity would increase. Additionally, in all cases, we used the default parameters for the maximum number of alleles called per position which is 6, and the maximum number of genotypes considered which is 1024.

Filtering step of the Variant Calling Pipeline

For the first version of the pipeline, we analyzed the distribution of all of the parameters computed by GenotypeGVCFs in the join-genotyping step. Based on the results we gradually implemented filters to reduce the number of false positives while avoiding implementing redundant filters. Out of all parameters computed by default by GenotypeGVCFs, we ended up using the quality (*QUAL*), the Genotype Quality *GQ*, the Quality by Depth *QD*, the Symmetric Odds Ratio test value, and the Rank sum test value. The quality scores are encoded in the "*QUAL*" column, and each corresponds to the position-specific phred-scaled probability a polymorphism exists in the input sequencing data,

$$QUAL = \ -10 \ * \ log(1 - p).$$

This means that a variant with a lower score indicates a higher probability that the decision is incorrect. The Genotype Quality represents the phred-scaled confidence that the genotype assignment (GT) is correct. It is the difference between the likelihoods of the two most likely genotypes. The SOR parameter is computed by performing a Symmetric Odds Ratio test, which estimates the bias that may come from sequencing when one strand is favored over the other [27]. The ReadPosRankSum value comes from the Rank sum test, which tries to identify whether there is evidence of bias in the position of reference and alternative alleles within the reads that support them. It tries to account for second-generation sequencing's tendency to concentrate errors at the extremes of the reads. Based on the distribution of our variant positions for each of the samples. We decided to remove those variant positions QD under 1, SOR above 3, and those with quality scores under 30. For the third version of the filter, we added the removal of variants with ReadPosRankSum values above 6, based on the visualization of the distribution of the positions after applying the second filter.

For the second version of the pipeline, after evaluating the performance of HaplotypeCaller (see evaluation), we analyzed the distribution of each of the VCF parameters computed by the tool per variant position while clustering by evaluation category. Based on the results, we implemented 4 filters of increasingly higher quality score thresholds using VariantFiltration. The GenotypeGVCFs tool needs a gVCF file as input since it needs to reevaluate each of the positions of the reference. This version of the pipeline was implemented to preserve the quality filter we applied, so we performed our version of a join-genotyping step in the form of a filter removing all positions reported in fewer than a given number of samples.

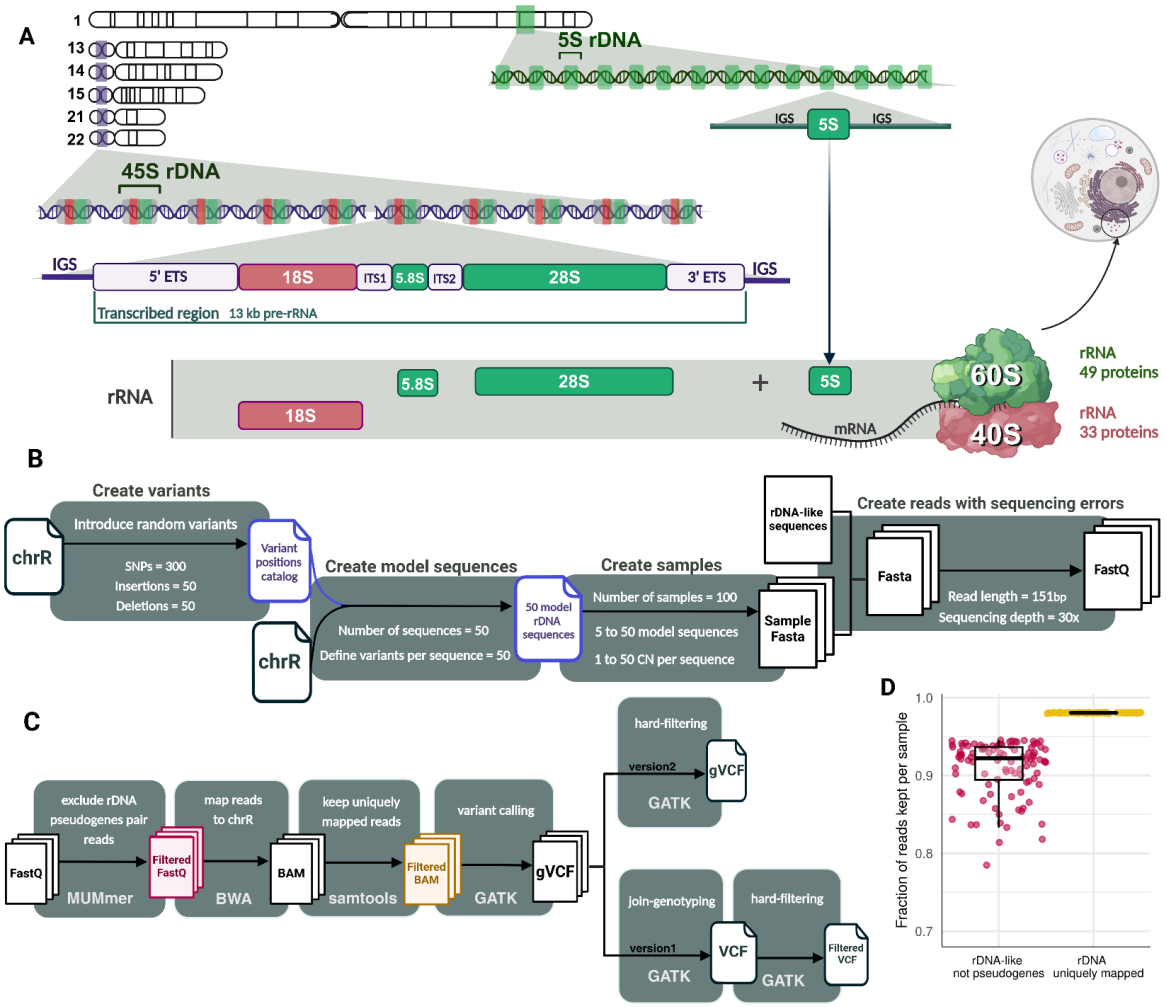

Figure 1. Diagrams of the project's pipelines.
A) Diagram of the 45S and 5S rDNA location, transcription, maturation process, and eventual ribosomal location. **B**) WGS data simulation diagram. We use a simulator to generate the Sample FASTA files, using the ChrR sequence as a reference. Then, we use NEAT to create artificial second-generation reads. **C**) Diagram of the pipeline, WGS data processing, and variant calling process including the two versions. The files free of pseudogenes are colored in red and uniquely mapped paired reads in yellow. **D**) Resulting fraction of rDNA simulated reads after removing pseudogenes in red, and a fraction of them that are uniquely mapped (right in yellow) for each of the 100 samples.

Validation process

Once we identified the adequate parameters to optimize the performance of the pipeline, we validated the process by using real sequencing reads. We downloaded the T2Ts genome assemblies along with their annotations and identified all 45S rDNA copies reported by He et. al. [20,19]. The final T2T-CHM13 assembly contains 219 complete rDNA copies of the 45S operon, totaling 9.9 Mbp of sequence [19]. T2T-YAO reports less than half with 98 rDNA copies for the Chinese individual (F1) [20]. First, we extracted from the annotation files, all regions containing attribute terms related to rRNA and excluding mitochondrial rRNA. The filtered gff3 files contained all the continuous 45S operons of all the acrocentric chromosomes along with pseudogenes, mitochondrial rRNA which avoided the filter, 5S rDNA genes, and incomplete 45S sequences, all of which were removed. The incomplete 45S rDNA sequences were removed and, in the end, we obtained all the coordinates of all continuous 45S operons which matched the number reported by the authors. Next, we used the

resulting gff3 file to extract the 45S rDNA sequences from each of the assemblies. To find them, we performed a set of MSA using Clustal-Omega [28] with iterations 1, 2, 5, 10, 25, and 50 for all the sequences found in the previous step for both parents' assemblies. Then, we selected one sequence out of every cluster sharing 100% identity to identify the unique sequences. Finally, we performed an additional heatmap containing all unique rDNA sequences we now had available; 22 from the paternal and 14 from the maternal T2T-YAO assembly, 24 from the T2T-CHM13 assembly, and the ChrR sequences.

We downloaded the four technical replicate samples available from the T2T-CHM13 and removed the adapters using Trimmomatric before running both versions of the pipeline with the best-performing parameters in each case (see results). To obtain the true variants for the evaluation of the validation process, we performed an alignment with the 24 unique sequences and created a table with all diverging sites.

Results

Variant calling in perfect reads shows great accuracy in non-repetitive regions

We first evaluated the performance of the pipeline with perfect reads to see if, even in the best-case scenario, there were regions in which variation was difficult to detect. We divided the rDNA array into repetitive and non-repetitive regions (see next section results) and, for both types of regions, we compared the GATK results with all the identified variants of the 24 unique rDNA sequences of the T2T-CHM13 genome (last section methods validation). As a result, we could expect all variants to be identified in this scenario, however, this was not the case. There are 41 variants not falling under repetitive regions, 38 were correctly identified for all three sets of kmers (Figure 2B). Only 3 positions were uncalled by GATK and thus, classified as false negatives (FN): positions 22309, 11727, and 18402 on ChrR. To understand why these variants were undetected, we used Integrative Genomics Viewer (IGV) [29] to visualize the perfect reads distribution when mapped to the 24 unique copies of the regions covering these positions, and took screenshots of the visualizations.

All three positions reside in high-GC regions containing homopolymers, and two of them present low coverage. The summary table **Table 1** shows how many of the 24 units did not have reads covering the positions, and in case they did, which allele they presented. The lack of coverage was either due to their mapping to other regions of the rDNA sequences or their removal during the Preprocessing step of the variant calling pipeline (see methods) which has the goal to remove pseudogenes and other rDNA-like fragments.

The 14bp deletion starting at position 11727 was only present in unit 3, which had no reads mapped covering the position, hence, it could not be detected (Figure S2A). The variant in position 18402 comes from unit 23 which had no coverage using 30x (Figure S2B), and 12 out of the 24 units presented half to no coverage in the surrounding bases. This variant was correctly identified when increasing the depth, seeing that in the set of depth 100X, all 22 reads mapped to unit 23 presented the allele. Finally, the variant falling on position 22309, is a multiallelic variant, and only one of the alternative alleles was identified, whereas the allele "C", which comes from unit 3, was undetected. In the case of the set of depth 30X, the position had half the expected coverage, but all 15 reads contained the allele "C". In the case of the set of depth 100X, there were 83 reads covering the position (Figure S2C). The majority, 70 reads, presented the allele "C", but the presence of 13 reads with the reference allele may have led GATK to evaluate it as a non-variant position.

Depth	Position	reference	alternative	alternative2	Evaluation	IGV ref	IGV alt	IGV alt2	no coverage
30X	11727	CGGTGGCGCCCCGCGTGGGGCC	-		FN	23	0		1
100X	11727	CGGTGGCGCCCCGCGTGGGGCC	-		FN	23	0		1
30X	18402	C	T		FN	22	0		2
100X	18402	C	T		TP	23	1		0
30X	22309	CCGCG	CCG	C	FN allele	12	3	1	8
100X	22309	CCGCG	CCG	C	FN allele	14	4	1	5

Table 1. Summary table of the IGV visualizations.
Summary of the IGV visualizations of the three positions evaluated as false negative, for the 30X and 100X perfect reads sets.

Overall, the results of the non-repetitive regions were favorable with no false positive positions found, a precision of 1, and a recall over 90% for the three sets of perfect reads (Figure 3D). However, the results in the repetitive regions were much worse, with the false variants being called and the majority of true variants being missed. This suggests that, even in the best-case scenario without sequencing errors, we cannot trust results in these regions. Therefore, we should focus our analysis on non-repetitive regions which, on average, correspond to 69.5% of the reference rDNA sequence (Figure 2A).

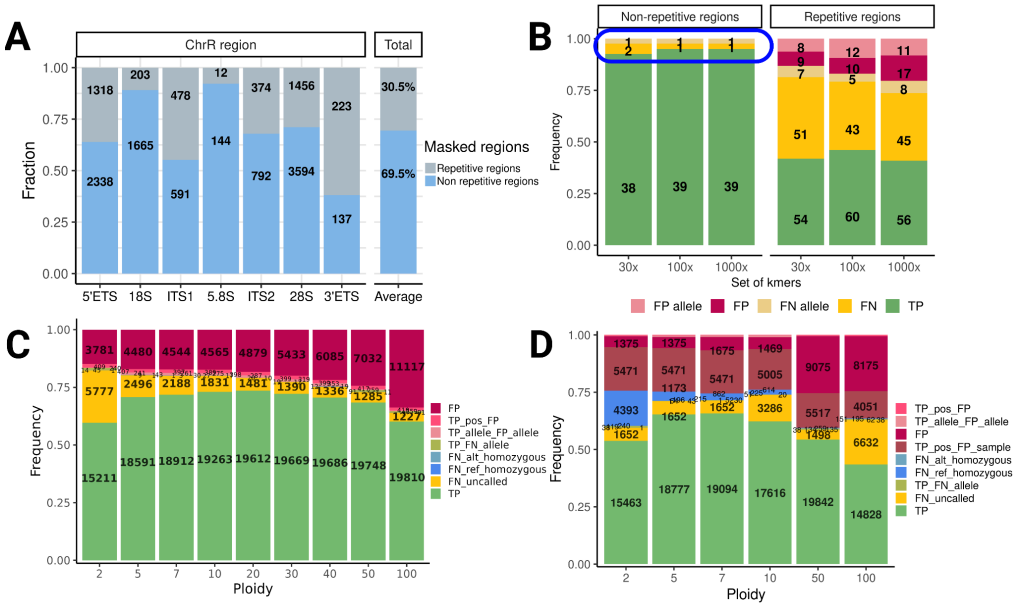

Figure 2. Results summary statistics barplots.
A) Fraction of repetitive regions blacklisted in the reference rDNA sequence per region. B) Results of the evaluation of the reads without sequencing errors. C) Results of the evaluation of the reads with sequencing errors using the second version of the pipeline D) Results of the evaluation of the reads with sequencing errors using the first version of the pipeline.

Retrieving candidate rDNA reads

The simulated second-generation reads in each of the samples contained varying copy numbers and coverage(Figure S3). In the first step of the pipeline 75 to 95% of them, corresponding to pseudogenes, were removed (Figure 1D). After the BWA, around 2% of the remaining reads were not uniquely mapped to chrR and were thus removed from the final pool of rDNA reads fed to GATk.

Simulating the reads with sequencing errors

We then simulated reads with sequencing errors, as they would be closer to real data. We used a simulator that initially generates samples containing sequences with shared and diverging variability, and then, generates artificial second-generation sequencing reads (see methods). In the first step, we generated a variant catalog with 300 SNPs, 50 insertions, and 50 deletions (Figure 1B), out of which only 293 SNPs and 93 insertions and deletions were included in the final 50 model sequences comprehended in the 100 artificial samples (Figure S4B). On average, each variant position was present in 78 samples, given that it was present in at least 1 sequence out of the 50 included in each sample (Figure S4A).

When generating the artificial sequencing reads (see methods) we wanted to account for the fact that, in a real set of WGS sequencing reads, we would find reads from pseudogenes, rDNA incomplete fragments, and other rDNA-like regions. Therefore, we decided to include all masked regions of the GRCh38 as defined in George et al in the simulation of the sort reads. These are rDNA-like regions masked to use the sequence as a rDNA sequence and avoid multimapping. In total, we included 473 sequences with an average length of 1860 base pairs, most around 300bp (Figure S5) which corresponded to around 13% of the total length of the sets of sequences, including the generation of sequencing reads. In the end, we obtained a FASTQ of simulated rDNA sequencing reads that contained variability and sequencing errors, just like real WGS data would. In the preprocessing step of the variant calling, all samples kept 75% to 95% of the reads after applying this filter (Figure 1D).

Evaluation result of the reads with sequencing errors

For all the calls of the simulated reads, we compared the final VCF files with the variants introduced during the simulation. The simulator records which of the variants of the catalog are introduced to each model sequence and which model sequence is included in each sample. Therefore, we could evaluate the performance by comparing the variant alleles expected in each sample with the ones found in the pipeline. We used an RScript that evaluates each variant found in the pipeline by comparing the genotypes estimated by GATK with the alleles that the sample was supposed to include.

As expected, the results of the simulated reads were less favorable than those of the reads with no sequencing errors or rDNA-like sequences introduced (Figure 2C). For the first version of the pipeline, after performing join-genotyping for ploidy values 2, 5, and 7, around 200 more variants were correctly identified (Figure 2D). compared to the previous step's results (Figure 2C); for ploidy 50 there were 100 more. However, for ploidy 10 and 1000 the number of correctly called positions decreased and, overall, the proportion of false positives was higher. Given that around 5000 were variants generated by the simulator, we can infer that GATK was trying to assign variants present in other samples to those that did not have them. To identify a way to remove them from the final VCF file we compared the distribution of all parameters in the file with those positions correctly identified. We observed that most positions with a genotype quality value (*GT*) falling under the 25 percentile for that sample resulted in a non-true positive evaluation. Thus, our first version of the filter removes those variants whose *GT* falls below the 25 percentile (Figure S6). This increased the

F1 score by 10% for all ploidy sets but 2 and 100 (Figure S7). For the second version of the filter, we considered the recommendations by GATk; quality (*QUAL*) above 30, Quality by Depth (*QD)* above 2, the Symmetric Odds Ratio test value below 3, and the Rank sum test value above 4 (see methods). After seeing the distribution of our data sets in each of the parameters (Figure S6), we decided that the second version of the filter would, in addition, remove variants presenting either *QUAL* lower than 30, *SOR* above 3, or *QD* under 2. This only improved the results of the sets of ploidy 50 and 100 (Figure S7). Hence, for the third version of the filter we added the removal of those variants with a Rank sum test value above 6 which resulted in a slight decrease in the recall compared to the previous version. In the end, the best best-performing filter was the second one (Figure S7), and the highest F1 score,

$$F1 = 2 \cdot precision \cdot recall \,/\, (precision + recall),$$

was achieved when using 7 as the value for the ploidy parameter, reaching 0.909. In the end, we the majority of the wrongly called positions for the selected set of parameters (Figure 3B), in the set of ploidy 7, using the second filter the removed percentage 87% .

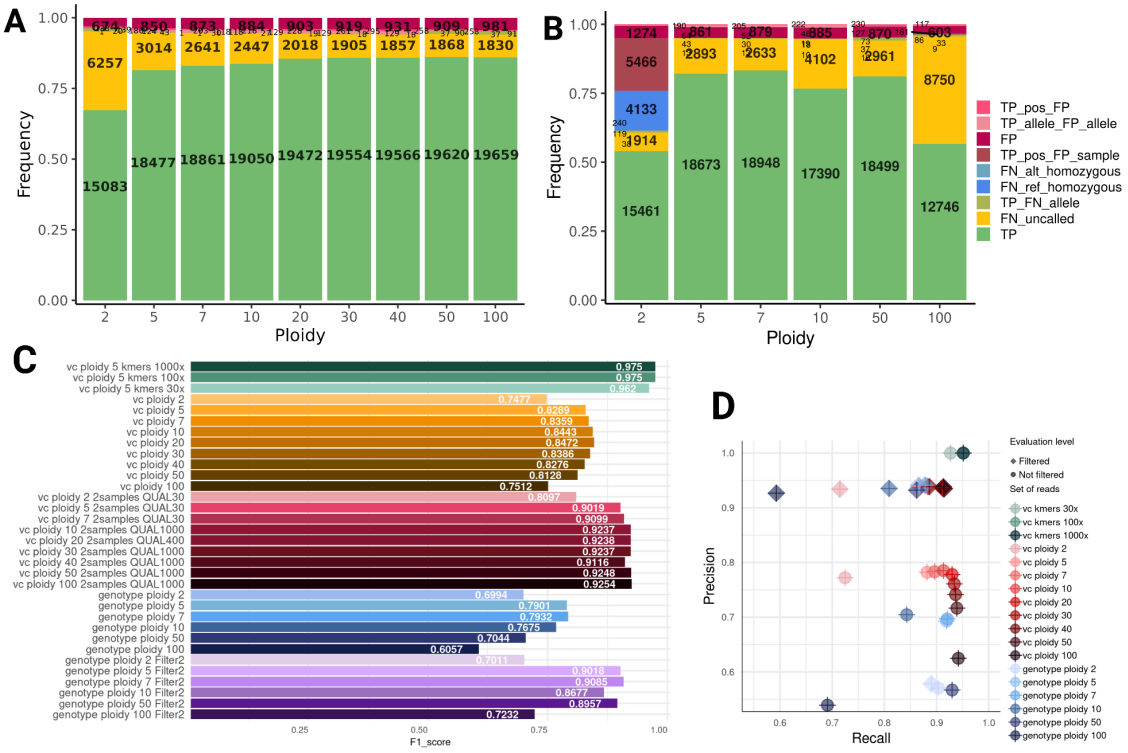

Figure 3. Results of the variant calling pipeline with the simulated reads

A) Results of the evaluation of the reads with sequencing errors using the second version of the pipeline when removing positions with quality below 1000 and appearing in less than two samples. B) Results of the evaluation of the reads with sequencing errors using the first version of the pipeline after applying the second version of the filter. C) F1 score best-performing sets in each of the evaluated conditions per ploidy. D) Precision by recall distribution of best-performing sets in each of the evaluated conditions per ploidy.

For the second version of the pipeline, out of all the default parameters computed for each of the positions in the output VCF files, only the quality showed a clear pattern that could be extrapolated into a filter (Figure S8), Some of the parameters showed a similar tendency since they were estimated using the quality value, therefore, applying filters based on them would be redundant and more costly. For the first version of the filter, we removed all positions with quality below 1 since none of them were true positives. The second version of the filter removed values with quality below

30 as recommended in the GATk documentation. Finally, the third and fourth filters were decided based on visualization of the quality distribution of the calls, using ploidy higher than 10. When using ploidy higher than 20, the majority of false positive values fell below 400 and, in the case of ploidy 50 and 1000, the threshold could be set at 1000. After evaluating the resulting filtered files, we observed that different filters worked best for different ploidies (Figure S9). The highest F1 score was achieved when using 1000 as the quality threshold and the ploidy value 50. For ploidy 50 and 100, using softer filters with a lower quality threshold, mildly increased the recall but also reduced the precision and in the end, resulted in a lower F1 score.

In our manual version of join-genotyping, initially, we thought of removing those positions called by only 1 sample to filter out those variants introduced in the artificial sequencing step, since we wouldn't expect a sequencing error to be introduced in the same position in two samples enough times to be considered a variant. However, after plotting the number of times each variant position was classified in each of the samples (Figure S10A), we decided to test the removal of those positions that were called by fewer than 2, 5, 10, and 15 samples. We considered 15 as the threshold, given that nine FP and no TP positions were called in less than 15 samples. We tested all combinations of the filters for ploidy 50 and 100 and the highest F1 score, 0.9258, was achieved using ploidy value 50 (Figure S10C), quality threshold 1000, and a minimum of 15 samples. However, we intend the pipeline to be used with samples from different donors, and setting the threshold to 15 samples could remove low-frequency alleles. Therefore, using 2 should be enough with real data to keep rare variants present in a few donors. The F1 score using 2 instead of 15 as the join-genotyping filter, remained about the same, with a value of 0.9236. In the end, after removing all variants with quality below 1000 and those appearing in less than two samples we still had around 900 false positive variants for all ploidy values which are not be removed (Figure 3A). However, out of the two the second version of the pipeline presented the the best results (Figure 3C).

Using our final set of parameters for version 2 of the pipeline, we found 18 missed and 11 false positive variants present in 50 to 80 samples (Figure S11C). All 11 false positives were within 1 to 3 positions after a missed variant (Figure S11D). This means that all of them are being called due to identified variability produced by the real variants and not the noise we introduced to the samples with the rDNA-like reads. We expected to have a higher proportion of missed alleles in regions with higher GC-content as we expect a lower coverage of those regions, however, this isn't the case with the simulated reads (Figure S11E). Additionally, GATK doesn't struggle to identify closely located variants (Figure S11A), but insertions and deletions are missed on a higher proportion than SNPs (Figure S11B).

Validation

We applied the two versions of the pipeline to T2T-CHM13 WGS data using the best-performing parameters in each case. For the first version of the pipeline, we didn't apply the filter removing the lower 25 percentile of the GQ distribution as it removed all variants (Figure S12), however, the remaining filters were enough to remove the identified false positives (Figure 4C). In the case of the second version of the pipeline, applying the filter removing all variants with quality value below 1000 removed all of the False positives (Figure 4D). In the end, both versions of the pipeline had precision 1, as no false variants were called (Figure 4A, 4B), and out of the two the second version had a higher recall (0.708) than the first one (0.595), finding 4 to 5 variants more per sample. Thus the F1 score of version 1 with real data is 0.746 and for version 2 is 0.829.

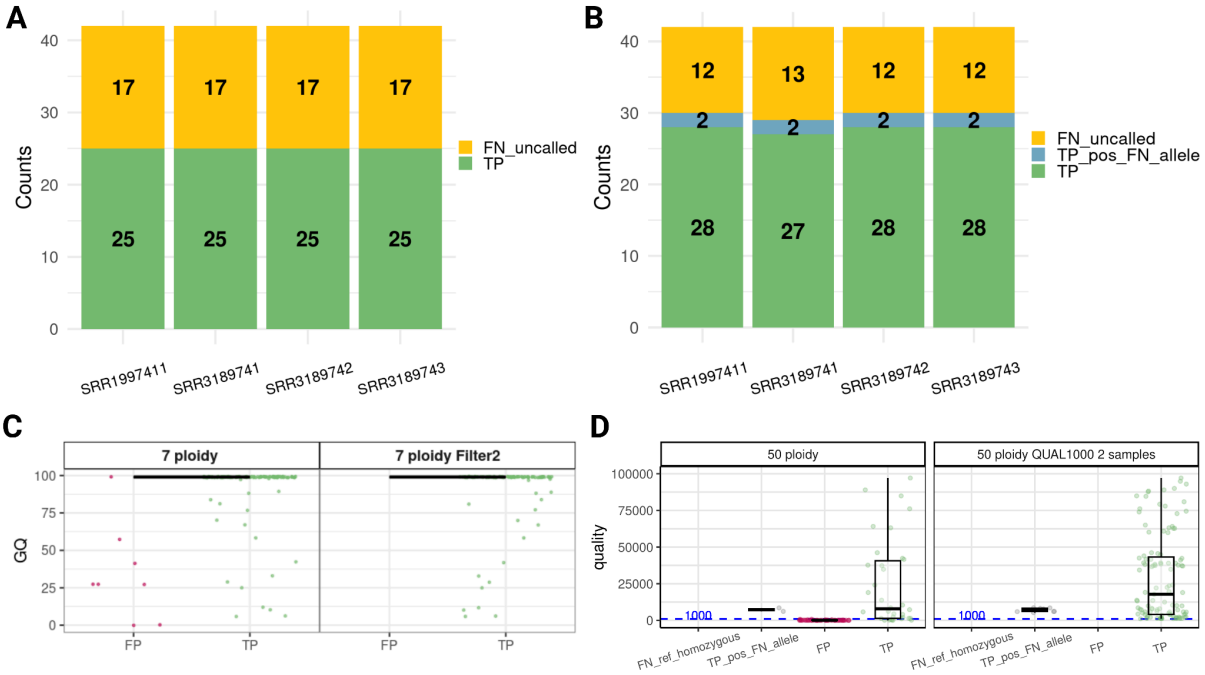

Figure 4. Results of the variant calling pipeline with real sequencing data

A) Results of the evaluation of the reads using the first version of the pipeline B) Results of the evaluation of the reads using the second version of the pipeline. C) Genotype quality distribution of all called positions before (left) and after (right) applying the filter using the first version of the pipeline. D) Quality distribution of all called positions before (left) and after (right) applying the filter suing the first version of the pipeline.

Discussion

We successfully developed a variant calling pipeline and identified the optimal performing values for the parameters explored in two versions of the pipeline. However, there is still room for improvement. On the one hand, the first version of the pipeline would be flawed for a highly diverse cohort. This is because GATK's GenotypeGVCFs tries to assign the same alleles to all samples (Figure 2B), which would lead to a false homogeneity of the results when trying to discover the variability of samples from diverging ancestry. Additionally, the validation process showed that out of the two versions, this wouldn't be the best choice, even when using technical replicates.

On the other hand, using the best-performing version out of the two, with the current set of parameters, all multiallelic variants are reported as biallelic, and the less frequent allele for each of the donors is always left unreported. This is an issue given that we know that in a single position of the real rDNA sequences, we can find more than two alleles (Figure 4B). We realized that the reason for the lack of detection of any third allele of multiallelic variants is the same as for the drop in precision in the join-genotyping step of version 1 of the pipeline (Figure 2B). Throughout our analysis we use the default values for both the *--max-alternate-alleles* and *--max-genotype-count* parameters which are 6 and 1024 respectively. However, the number of genotypes as defined by GATK given a ploidy and a number of alternate alleles corresponds to the next formula:

$$F(num\_alleles, ploidy) = (num\_alleles + ploidy - 1)! \,/\, ((num\_alleles - 1)! * ploidy!).$$

This means that for example, in ploidy 50 the maximum number of alleles that would be reported per position would be 2, as

$$F(3, 50) = 1326,$$

which already surpasses the default maximum of HaplotypeCaller, 1024. Therefore, to avoid calling only biallelic positions with ploidy 50, we should update the parameter *--max-genotype-count* for every ploidy to keep constant the maximum number of alleles to 6. This means that for ploidy 10, we should use 3003, for ploidy 50, 3478761, and ploidy 100, 96560646. If we modified these values we could only use the second version of the pipeline since GenotypeGVCFs only takes files computed using at most 1024 genotypes.

All the analyses we performed were position-specific. For example, positions evaluated as "TP_allele_FP_allele" are variants for which the correct alternative allele was found but an additional incorrect allele was assigned to position. When evaluating the precision we assigned these positions false positives since they were wrongly called variants. Moving forward these could be split and considered as true positive and false positive alleles. Additionally, during the manual join-genotyping step of version 2 of the pipeline, in which we remove all positions that are only present in one sample, performing this filter at the allele level could lead to the selection of the true positive alleles.

Conclusion and Future Directions

In this paper, we developed and tested a variant calling pipeline to detect rDNA variants with a final 0.938 precision and 0.913 recall (Figure 3D). We identified which regions were unevaluable and should be excluded from the analysis as their detection was challenging in an ideal case scenario. Then, we tested the performance of two versions of the pipeline using a simulated set of realistic rDNA reads presenting shared variability and genomic noise. In the end, we identified the optimal parameters for each version of the pipeline and validated them using WGS data from the T2T-CHM13 assembly.

So far, out of the two T2T assemblies, only CHM13 has WGS data publicly available. However, both the CHM13 and YAO rDNA unique copies were identified to perform the variant calling of the T2T-YAO WGS data once it is released (Figure S13).

Finally, we evaluated the performance of the pipeline using a single set of simulated reads. Right now our estimates are the average performance of the 100 samples generated. However, we would get a better estimate of the performance of the pipeline by computing the mean across different i sets of simulated reads.

Supplementary Data

GitHub: https://github.com/AlbaJimLup/FTP
Supplementary Figures and Tables: Supplementary materials

Acknowledgments

I would like to thank Marta Melé, who served as a PI, and my supervisor, José Miguel Ramírez for their assistance, guidance and support throughout the Final Degree Project.

References

1. Panse, Vikram Govind, and Arlen W. Johnson. 2010. “Maturation of Eukaryotic Ribosomes: Acquisition of Functionality.” Trends in Biochemical Sciences 35 (5): 260–66.

2. George, Subin S., Maxim Pimkin, and Vikram R. Paralkar. 2023. “Construction and Validation of Customized Genomes for Human and Mouse Ribosomal DNA Mapping.” The Journal of Biological Chemistry 299 (6): 104766.

3. Caburet, Sandrine, Chiara Conti, Catherine Schurra, Ronald Lebofsky, Stuart J. Edelstein, and Aaron Bensimon. 2005. “Human Ribosomal RNA Gene Arrays Display a Broad Range of Palindromic Structures.” Genome Research 15 (8): 1079–85.

4. Shiao, Yih-Horng. 2021. “Discoveries and Biological Implications of Mammalian 45S rDNA Variants and Non-Structural rRNAs.” http://dx.doi.org/.

5. Kim, Jung-Hyun, Alexander T. Dilthey, Ramaiah Nagaraja, Hee-Sheung Lee, Sergey Koren, Dawood Dudekula, William H. Wood Iii, et al. 2018. “Variation in Human Chromosome 21 Ribosomal RNA Genes Characterized by TAR Cloning and Long-Read Sequencing.” Nucleic Acids Research 46 (13): 6712–25.

6. Miller, O. L., Jr, and B. R. Beatty. 1969. “Visualization of Nucleolar Genes.” Science 164 (3882): 955–57.

7. Shi, Zhen, and Maria Barna. 2015. “Translating the Genome in Time and Space: Specialized Ribosomes, RNA Regulons, and RNA-Binding Proteins.” Annual Review of Cell and Developmental Biology 31 (Volume 31, 2015): 31–54.

8. Milenkovic, Ivan, Helaine Graziele Santos Vieira, Morghan C. Lucas, Jorge Ruiz-Orera, Giannino Patone, Scott Kesteven, Jianxin Wu, et al. 2023. “Dynamic Interplay between RPL3- and RPL3L-Containing Ribosomes Modulates Mitochondrial Activity in the Mammalian Heart.” Nucleic Acids Research 51 (11): 5301–24.

9. Parks, Matthew M., Chad M. Kurylo, Randall A. Dass, Linda Bojmar, David Lyden, C. Theresa Vincent, and Scott C. Blanchard. 2018. “Variant Ribosomal RNA Alleles Are Conserved and Exhibit Tissue-Specific Expression.” Science Advances 4 (2): eaao0665.

10. Gonzalez, I. L., and J. E. Sylvester. 1995. “Complete Sequence of the 43-Kb Human Ribosomal DNA Repeat: Analysis of the Intergenic Spacer.” Genomics 27 (2): 320–28.

11. Kim, Jung-Hyun, Vladimir N. Noskov, Aleksey Y. Ogurtsov, Ramaiah Nagaraja, Nikolai Petrov, Mikhail Liskovykh, Brian P. Walenz, et al. 2021. “The Genomic Structure of a Human Chromosome 22 Nucleolar Organizer Region Determined by TAR Cloning.” Scientific Reports 11 (1): 2997.

12. Rothschild, Daphna, Teodorus Theo Susanto, Jeffrey P. Spence, Naomi R. Genuth, Nasa Sinnott-Armstrong, Jonathan K. Pritchard, and Maria Barna. 2023. “A Comprehensive rRNA Variation Atlas in Health and Disease.” bioRxiv. https://doi.org/10.1101/2023.01.30.526360.

13. Barna, Maria. 2017. “Specialized Ribosomes: A New Frontier in Gene Regulation, Organismal Biology, & Evolution.” FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology 31 (S1). https://doi.org/10.1096/fasebj.31.1_supplement.387.1.

14. McCann, Kathleen L., and Susan J. Baserga. 2013. “Genetics. Mysterious Ribosomopathies.” Science 341 (6148): 849–50.

15. Fan, Wenjun, Eetu Eklund, Rachel M. Sherman, Hester Liu, Stephanie Pitts, Brittany Ford, N. V. Rajeshkumar, and Marikki Laiho. 2022. “Widespread Genetic Heterogeneity of Human Ribosomal RNA Genes.” RNA 28 (4): 478–92.

16. Rothschild, Daphna, Teodorus Theo Susanto, Xin Sui, Jeffrey P. Spence, Ramya Rangan, Naomi R. Genuth, Nasa Sinnott-Armstrong, Xiao Wang, Jonathan K. Pritchard, and Maria Barna. 2023. “Diversity of Ribosomes at the Level of rRNA Variation Associated with Human Health and Disease.” bioRxiv: The Preprint Server for Biology, December. https://doi.org/10.1101/2023.01.30.526360.

17. Morton, Elizabeth A., Ashley N. Hall, Elizabeth Kwan, Calvin Mok, Konstantin Queitsch, Vivek Nandakumar, John Stamatoyannopoulos, Bonita J. Brewer, Robert Waterston, and Christine Queitsch. 2020. “Challenges and Approaches to Genotyping Repetitive DNA.” G3 10 (1): 417–30.

18. Schneider, Valerie A., Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, et al. 2017. “Evaluation of GRCh38 and de Novo Haploid Genome Assemblies Demonstrates the Enduring Quality of the Reference Assembly.” Genome Research 27 (5): 849–64.

19. Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. “The Complete Sequence of a Human Genome.” Science 376 (6588): 44–53.

20. He, Yukun, Yanan Chu, Shuming Guo, Jiang Hu, Ran Li, Yali Zheng, Xinqian Ma, et al. 2023. “T2T-YAO: A Telomere-to-Telomere Assembled Diploid Reference Genome for Han Chinese.” Genomics, Proteomics & Bioinformatics, August. https://doi.org/10.1016/j.gpb.2023.08.001.

21. Stephens, Zachary D., Matthew E. Hudson, Liudmila S. Mainzer, Morgan Taschuk, Matthew R. Weber, and Ravishankar K. Iyer. 2016. “Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models.” PloS One 11 (11): e0167047.

22. The MUMmer 3 Manual.” n.d. Accessed March 27, 2024. https://mummer.sourceforge.net/manual/.

23. Shen, Wei, Shuai Le, Yan Li, and Fuquan Hu. 2016. “SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation.” PloS One 11 (10): e0163962.

24. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at http://arxiv.org/abs/1303.3997 (2013).

25. Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” Bioinformatics 25 (16): 2078–79.

26. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. 201178 Preprint at https://doi.org/10.1101/201178 (2018).

27. “Tool Documentation Index.” n.d. GATK. Accessed March 27, 2024. https://gatk.broadinstitute.org/hc/en-us/articles/360036856231--Tool-Documentation-Index.

28. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539 (2011).

29. “IGV: Integrative Genomics Viewer.” n.d. Accessed March 29, 2024. https://igv.org/.
30. RepeatMasker Web Server. https://www.repeatmasker.org/cgi-bin/WEBRepeatMasker.

All figures were created with BioRender.com