JOHANNES KEPLER
UNIVERSITY LINZ

Author
Mehdin Masinovic

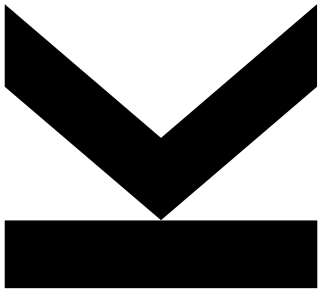
Submission
Institute of Biophysics

Thesis Supervisor
Assoc. Prof. Dr. Irene Tiemann-Boege

Assistant Thesis Supervisor
Dr. Philipp Hermann

November 2020

SHORT TANDEM REPEAT ANALYSIS IN RECOMBINATION HOTSPOTS ACROSS THE HUMAN GENOME



Bachelor's Thesis

to confer the academic degree of

Bachelor of Science

in the Bachelor's Program

Bioinformatics

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

Bibliographical Detail

Masinovic, M., 2020: Short Tandem Repeat Analysis in Recombination Hotspots across the Human Genome. Bachelor Thesis, in English. - 44 p., Institute for Biophysics, Johannes Kepler University, Linz, Austria

Annotation

Short tandem repeats are one of the most abundant tandem repeat types and are an important source of genetic variation. Comparative studies have analyzed their abundance in promoters, genes, and other relevant regions in the human genome, but short tandem repeats in recombination hotspots have yet to be fully characterized. Using the R package entitled *STRAH*, we analyzed the frequency of 310 distinct short tandem repeats in 37527 recombination hotspots across the human reference genome. We generated pattern-specific comparisons for all repeat types among recombination hotspots, regions directly surrounding them, and the remaining genomic region of the human genome. We detected that C/G-rich repeats tend to be enriched in recombination hotspots, observed that A/T-rich repeats are more enriched in regions unrelated to recombination, and found that repeats are present in very low numbers if they do not contain consecutively repeated DNA bases. Collectively, our results provide a standardized, genome-wide characterization of short tandem repeats in recombination hotspots and their surrounding regions, highlight pattern-specific differences that depend on repeat length and repeat type, and give insight into short tandem repeat enrichment in relation to recombination hotspots across the human genome.

Declaration

I hereby declare that I have worked on my Bachelor's thesis independently and used only the sources listed in the bibliography.

I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my Bachelor's thesis, which is kept in full form in the Faculty of Science archive and in electronic form in the publicly accessible part of the STAG database operated by University of South Bohemia in České Budějovice accessible through its web pages.

Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defense in accordance with the aforementioned Act No. 111/1998. I also agree to the comparison of the text of my thesis with the Theses.cz thesis database operated by the National Registry of University Theses and a plagiarism detection system.

....

Place, Date Mehdin MASINOVIC

Contents

List of Figures

List of Tables

1 Introduction

Short tandem repeats (STRs) are segments of DNA where a specific pattern of nucleotides occurs repeatedly (Lander et al., 2001). STRs are one of the most abundant tandem repeat types; they have a non-random distribution (Sawaya et al., 2013; Kozlowski et al., 2010) and constitute approximately 1% of the human genome (Gymrek et al., 2016). About 17% of all human genes contain STRs in their coding region (Gemayel et al., 2010), and 18.8% of genes contain at least one STR in their upstream regulatory region (Bolton et al., 2013).

Compared to single-nucleotide polymorphisms (SNPs) and non-repeating regions of DNA, short tandem repeats have a higher mutation rate (Brinkmann et al., 1998; Weber & Wong, 1993; Verstrepen et al., 2005), which is due to either the slippage of the DNA polymerase at the time of DNA replication (Weber & Wong, 1993; Wells et al., 2005) or the imprecise repair of double-strand breaks in the DNA (Wells et al., 2005). As a result, short tandem repeats are one of the most polymorphic classes of alleles and an important source of genetic variation (Gymrek et al., 2016; Willems et al., 2014; Gemayel et al., 2010; Bolton et al., 2013).

STRs can have various functional properties, such as the regulation of transcription-factor binding (Contente et al., 2002; Martin et al., 2005; Deaton & Bird, 2011), the regulation of gene expression (Gymrek et al., 2016; Contente et al., 2002; Martin et al., 2005; Kouzine & Levens, 2007), and the formation of DNA secondary structures (Hefferon et al., 2004; Qin & Hurley, 2008; Guedin et al., 2010). Moreover, there is an enrichment of short tandem repeats near transcriptional start sites (TSSs) such as the promoter region and the 5' untranslated region (UTR) of genes (Lawson & Zhang, 2008; Bolton et al., 2013; Sawaya et al., 2013; Vinces et al., 2009), which further highlights the role of short tandem repeats in the regulation of gene expression. Lastly, STRs have been associated with several diseases (Manolio et al., 2009) such as Huntington's disease or the fragile-X syndrome (Mirkin, 2007; Sathasivam et al., 2013).

Several studies have investigated the enrichment of short tandem repeats in relation to certain biological processes, such as gene expression (Sawaya et al., 2013) or disease (Mirkin, 2007; Sathasivam et al., 2013). In the following study, we investigate a possible association between short tandem repeat types and recombination hotspots in the human reference genome. Meiotic recombination is a biological process that leads to the reshuffling of genetic material (Jensen-Seaman et al., 2004). The rates of recombination differ in the genome, where regions with high recombination rates are defined as recombination hotspots (Thomsen et al., 2001; Paigen & Petkov, 2010).

Previous studies have suggested that short tandem repeats can be enriched, suppressed, or unaffected by recombination (Myers et al., 2005; Majewski & Ott, 2000). In yeast, the presence of certain repeats can affect recombination activity (Treco & Arnheim, 1986; Schultes & Szostak, 1991). Moreover, Bagshaw et al. (2008) have shown that in the S. cerevisiae genome, long mono-, di-, and trinucleotide microsatellites are more frequent in hotspots compared to non-hotspot regions. An analysis of microsatellites on human chromosome 22 has shown a significant correlation between long GT repeats and recombination hotspots (Majewski & Ott, 2000). As certain short tandem repeat types are known to have functional properties, it is relevant to investigate the relationship between short tandem repeats and recombination hotspots.

We extended the functionality of our R package entitled *STRAH* - a short tandem repeat finder - to search for any repeat type or DNA motif in a DNA sequence. In this thesis, we use *STRAH* to screen the human reference genome for all distinct combinations of mono-, di-, tri-, and tetranucleotide repeats, as well as certain DNA motifs. Using the double-strand break map (DSB-map) of Pratto et al. (2014), we analyze the presence and density of detected repeats in relation to recombination hotspots in the human genome.

2 Materials

The data analysis of this thesis is conducted on the assembly version GRCh38/hg38 of the human reference genome (Team, 2015). The reference sequence is produced by the Genome Reference Consortium (GRC) and provides highly accurate sequence information on the human genome (I. H. G. S. Consortium et al., 2004; G. R. Consortium et al., 2019). We use the *Bioconductor* package (Huber et al., 2015) in *R* (R Core Team, 2020) to retrieve genomic data in the form of *Biostrings* objects (Pagès et al., 2019) using the starting and ending positions of every chromosome of the human reference genome shown in Table 1.

Chromosome	Start	End	Chromosome	Start	End	Chromosome	Start	End
chr1	1	248956422	chr9	1	138394717	chr17	1	83257441
chr2	1	242193529	chr10	1	133797422	chr18	1	80373285
chr3	1	198295559	chr11	1	135086622	chr19	1	58617616
chr4	1	190214555	chr12	1	133275309	chr20	1	64444167
chr5	1	181538259	chr13	1	114364328	chr21	1	46709983
chr6	1	170805979	chr14	1	107043718	chr22	1	50818468
chr7	1	159345973	chr15	1	101991189	chrX	1	156040895
chr8	1	145138636	chr16	1	90338345	chrY	1	57227415

Table 1: A position matrix listing the starting and ending coordinates of every chromosome in the human reference genome. It includes the full genomic range of all chromosomes of the human reference genome (version GRCh38/hg38).

We relate the position of detected short tandem repeats to the position of recombination hotspots. The recombination hotspot coordinates are retrieved from the comprehensive maps of meiotic recombination initiation (double-stranded break maps) produced by (Pratto et al., 2014). In our data application, we define recombination hotspots as ± 500 base pairs upstream and downstream from the hotspot coordinates identified by (Pratto et al., 2014).

The data analysis was conducted in *R* (R Core Team, 2020). The short tandem repeats were detected using the *R* package entitled *STRAH* (Hermann et al., 2020). The resulting data was analyzed with *R* scripts, which are available at `https://github.com/MehdinMasinovic/Human-Genome-STR-Screening`. The data was summarized using base *R* (R Core Team, 2020), *plyr* (Wickham, 2011) and *dplyr* (Wickham et al., 2020). The figures were produced using *ggplot2* (Wickham, 2016) and *cowplot* (Wilke, 2019), and the tables were produced using *kableExtra* (Zhu, 2019).

Differences in repeats were tested using a Kruskal-Wallis test (Kruskal & Wallis, 1952) and a post-hoc analysis using a Dunn's test (Dunn, 1964) with Holm's error correction (Holm, 1979). The data was tested for non-normality using the Kolmogorov-Smirnov test (Massey Jr, 1951) and for the homogeneity of variances using the Fligner-Killeen test (Fligner & Killeen, 1976).

3 Methods

3.1 *STRAH*

3.1.1 Update

STRAH is an *R* package developed to detect short tandem repeats in a DNA sequence (Hermann et al., 2020). Originally, *STRAH* (Version: 1.0.1) was designed to detect mononucleotide repeats, which are defined as a repeated segment of the same nucleotide (e.g. "AAAAAAAA" is a mononucleotide repeat of length 8). With the new update (Version: 1.1.0) *STRAH*'s functionality was extended to enable the detection of any type of repeat (mono-, di-, tri-, and polynucleotide repeats) with any length. The extension was added within the *R* function *STR_detection()*, which detects user-defined repeats in any DNA sequence specified by the user. The *matchPattern()* function of the *Biostrings* package is used in *STR_detection()* to search for repeats with a repeat unit greater than one (Hermann et al., 2020; Pagès et al., 2019). *STRAH* also allows a user-defined number of mismatches in the repeat search.

3.1.2 Workflow

For any user-defined repeat type, *STRAH* detects all occurences of the repeat in a DNA sequence and stores the coordinates. As an additional feature, *STRAH* relates the position of a detected repeat in relation to the closest recombination hotspot. The genomic region around each hotspot is structured into several "zones". By default, *STRAH* partitions the region on each side of a hotspot into five zones of user-defined length. The zones surrounding the hotspot are then numbered from 1 to 5 based on their proximity to the hotspot, i.e. the closer the zone is to the hotspot, the lower the zone number. The remaining DNA sequence (excluding hotspots and zones) is considered to be the "outside zone". Finally, *STRAH* stores the coordinates of the detected repeat and the zone number in which the repeat is found. Figure 1 provides a visual representation of the partitioning of a genomic region around a hotspot into several zones.

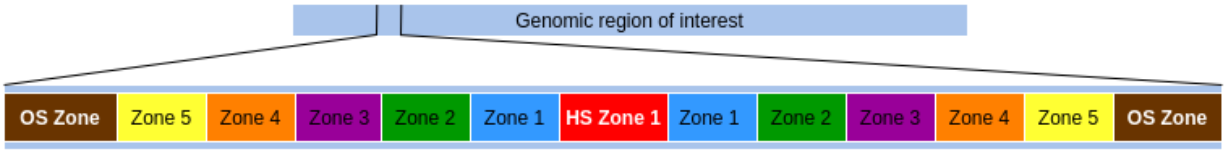


Figure 1: A recombination hotspot and the zones surrounding it. *STRAH* uses the recombination hotspot coordinates to structure the genomic region surrounding every hotspot into zones. By default, every hotspot is flanked by five zones on each end of the hotspot. The remaining DNA sequence (excluding hotspots and flanking zones) is treated as the "outside zone". Therefore, for any detected repeat within the genomic region of interest, the position is recorded together with the zone it is found in (Hotspot Zone (HS Zone), Zone 1-5, Outside Zone (OS Zone)).

3.1.3 Input

To detect a user-defined repeat in a DNA sequence, *STRAH* requires a DNA sequence and the coordinates of recombination hotspots as input. *STRAH* can either use a FASTA file or genomic information stored in the form of *Biostrings* objects as input. A *Biostrings* object is a memory-efficient string container used to store DNA sequences (Pagès et al., 2019). When using *Biostrings* a object, a position matrix must be provided to set the genomic region of interest for the analysis. Another required input argument of *STRAH* is a recombination hotspot matrix, which consists of the starting and ending positions of recombination hotspots in a DNA sequence. The information derived from the recombination-hotspot matrix is then used to set the zones flanking each recombination hotspot.

3.2 Validation of *STRAH*'s output

In order to validate the correctness of *STRAH*'s output, we developed a set of *R* functions to generate a DNA sequence and a corresponding recombination hotspot matrix. The simulated DNA sequence contains a user-defined set of repeats of different length. The repeat types occur at frequencies individually specified for each zone. As a result, the generated DNA sequence emulates a genomic region that contains hotspots with several repeat types. Finally, *STRAH* screens the emulated sequence for repeats. According to the parameters defined to simulate the DNA sequence, the correctness of *STRAH*'s output can be assessed. Figure 2 provides a visual representation of the simulation process.

The correctness of *STRAH*'s result is tested in regard to the correct detection and zone determination of the repeats. For this reason, we generate a DNA sequence with equal repeat density in each zone (Hotspot Zone (HS Zone), Zone 1-5, Outside Zone (OS Zone)). The simulated sequence is 200,000 base pairs long and it contains mononucleotide repeats of adenosine ranging from 6 to 10 base pairs. The generated sequence contains 10 hotspots in total, each being 2,000 base pairs long. Every hotspot is flanked by five zones on each end, where each zone in total is 2,000 base pairs long. Each zone is given an equal density of adenosine nucleotides, i.e. the total number of adenosine bases is the same for all zones. The total number of adenosine bases of each zone is equally distributed among repeat types, i.e. each repeat type has a share of 20%. Next, the sequence-generating function produces a pool of possible poly-A repeats and creates the simulated DNA segment for each zone where each repeat is separated by at least one of the remaining nucleotides (cytosine, guanine, thymine). The occurrence of the remaining nucleotides is simulated randomly.

By running *STRAH* on the simulated sequence with equal repeat density, we expect an approximately uniform distribution of poly-A density across all zones (Hotspot Zone (HS Zone), Zone 1-5, Outside Zone (OS Zone)), which is shown in Figure 3A. Figure 3B shows a second simulation where a sequence is generated under the same conditions except that Zone 1 and Zone 5 do not contain any repeats.

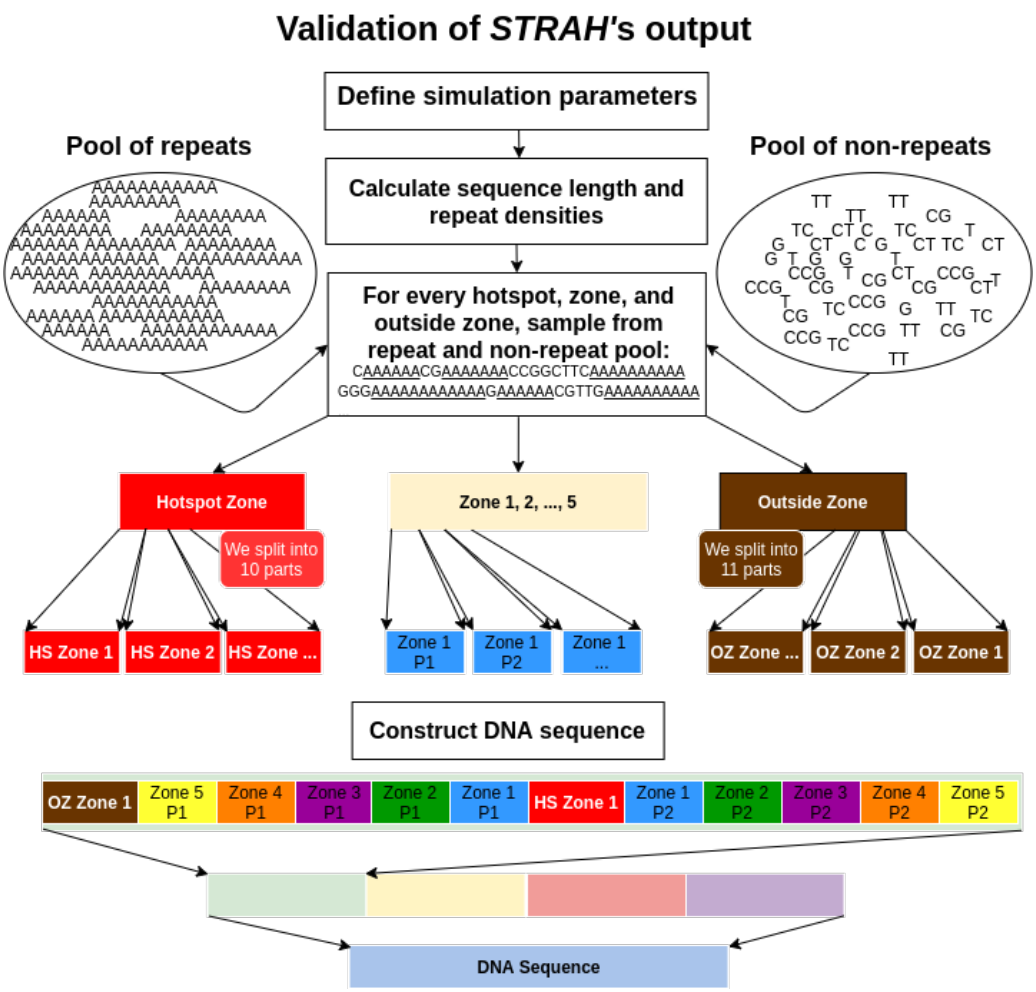


Figure 2: A description of the simulation process that tests the output of *STRAH* using a simulated DNA sequence. The sequence is generated by defining simulation parameters which are used to set the sequence length and repeat densities. The DNA sequence for each zone (Hotspot zone, Zone 1-5, Outside zone) is generated by sampling from a pool of repeats and non-repeats. The complete DNA sequence is constructed by combining the zone parts in the specified order (Outside Zone, Zone 5-1, Hotspot Zone, Zone 1-5, Outside Zone, Zone 5-1, ...).

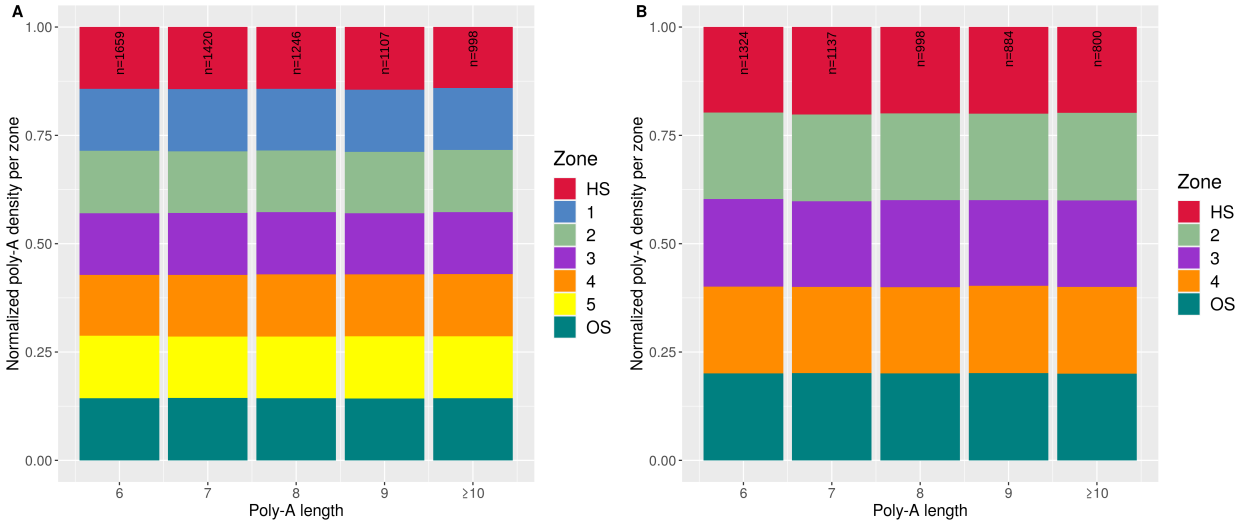


Figure 3: Simulation output. This figure shows the result of applying *STRAH* to a simulated sequence of 200,000 base pairs length. The simulated sequence contains hotspot zones, five zones flanking the hotspot, and an outside zone. All zones are given the same poly-A density, which results into a uniform density over all zones, as shown in panel A. Panel B shows a simulation where no repeats are contained in Zone 1 and 5.

3.3 Statistical analysis and data visualization

Short tandem repeats occur at different frequencies in the genome (Sawaya et al., 2013; Kozlowski et al., 2010). In our analysis, we investigate the frequency of a repeat type and compare it among i) recombination hotspots, ii) zones directly surrounding them, and iii) the remaining genomic region of the human genome (Outside Zone). As the compared regions are of different length, we compare the frequencies of repeats relative to the length of the zone they are found in, i.e. we assess their relative frequencies.

In the following formulations, we describe how the repeat densities of each repeat type are obtained. We refer to a repeat type as "poly-X", where X stands for any DNA motif that occurs repeatedly. For example, a poly-A repeat is a type of mononucleotide repeat that consists of adenosine bases only, i.e. "AAAAAA" is a poly-A repeat of 6 base pairs in length. As an additional instance, a poly-AT repeat is a dinucleotide repeat where the pattern "AT" occurs repeatedly. The same holds for the tri- and tetranucleotide repeats, and DNA motifs under study.

The poly-X density is defined as the number of poly-X repeats found in a zone divided by the corresponding zone length. The frequency of each poly-X tract with the length i is divided by the length of the zone j it is found in. The zone number j ranges from 1 to 7, where 1 is the hotspot zone, 2-6 are the surrounding 5 zones, and 7 is the outside zone. The poly-X density of a repeat type with length i detected in zone j is calculated as the following:

$$\text{Poly-}X\text{ Density}_{ij} = \frac{\text{Frequency of Poly-}X_{ij}}{\text{Zone Length}_j}; 1 \leq j \leq 7$$

The normalized poly-X density for each poly-X repeat length i is obtained by dividing the poly-X density by the sum of densities across all seven zones of the corresponding poly-X repeat length i.

$$\text{Normalized Density}_i = \frac{\text{Poly-}X\text{ Density}_{ij}}{\sum_{1 \leq j \leq 7} \text{Poly-}X\text{ Density}_{ij}}; 1 \leq j \leq 7$$

The density of X's of a repeat type found in a zone is obtained by multiplying the length i of a repeat by its frequency n, and dividing it by the length of the zone j.

$$\text{Density of } X_j = \frac{(\text{Length of Poly-}X_{ij} \times n_{ij})}{\text{Zone Length}_j}; 1 \leq j \leq 7$$

The poly-X density relative to the hotspot zone is obtained by dividing the poly-X density of a repeat with length i in zone j by the poly-X density of the hotspot zone j, i.e. $j = 1$.

$$\text{Relative Poly-}X\text{ Density} = \frac{\text{Poly-}X\text{ Density}_{ij}}{\text{Poly-}X\text{ Density}_{i,j=1}}; 1 \leq j \leq 7$$

4 Results

STRAH (Hermann et al., 2020) is applied to version GRCh38/hg38 of the human reference genome (G. R. Consortium et al., 2019). The detected repeats are then grouped by their repeat type, repeat length, and relative position to recombination hotspots. To relate the position of detected repeats to recombination hotspots, we use the recombination-hotspot coordinates from (Pratto et al., 2014). In our data application, we define recombination hotspots as ±500 base pairs both upstream and downstream from the hotspot coordinates of (Pratto et al., 2014), leading to an average hotspot length of ~2,000 base pairs. The flanking zones are chosen as 1,000 base-pair segments both upstream and downstream from the hotspot zone (2,000 base pairs in total per zone per hotspot), and each hotspot is flanked by five zones.

The short tandem repeat analysis comprises a whole genome screening of the human reference genome of several repeat types. We screen for the mononucleotide repeats of all four DNA bases (adenosine, cytosine, guanine, thymine) that are consecutively repeated at least six times. Additionally, we screen the human reference genome for di- and trinucleotide repeats that are consecutively repeated at least three and two times, respectively. Lastly, we detect the positions of all distinct combinations of tetranucleotide repeats and several DNA motifs that are repeated at least once. In our analysis, we only consider perfectly matching repeats, i.e. we do not allow mismatches. The detected repeats are then presented in the form of relative densities, where the frequency of each repeat type is considered relative to the length of the zone it is found in. Section 4.3 provides a more comprehensive overview of how the relative densities are obtained.

4.1 Mononucleotide repeats

We screened the human reference genome for the mononucleotide repeats of all DNA bases (adenosine, cytosine, guanine, and thymine) of a minimum length of 6 base pairs.

The absolute number of detected repeats of the complementary bases of the set (adenosine, thymine) and the set (cytosine, guanine) is comparably similar, as shown in Table 2. Moreover, poly-C and poly-G repeats exist in markedly lower absolute numbers (with a ratio of approximately 1 to 12) compared to poly-A and poly-T repeats.

	Absolute Repeat Frequencies per Zone						
STR	HS	1	2	3	4	5	OZ
A	818,975	650,012	623,484	596,481	578,349	556,129	23,585,191
T	822,549	646,394	621,369	601,449	578,485	558,141	23,872,526
C	76,057	57,492	53,572	49,762	46,898	46,124	1,591,162
G	75,148	56,988	52,555	49,749	47,594	46,369	1,589,151
Zone Length (in bp)	93,730,487	75,054,000	75,054,000	75,054,000	75,054,000	75,054,000	2,619,269,345

Table 2: A comparison of absolute repeat frequencies among DNA bases. The detected repeats are summarized per zone to compare the absolute frequencies among complementary bases. The length of every zone (Hotspot Zone (HS Zone), Zone 1-5, Outside Zone (OS Zone)) is listed in base pairs.

Figure 4 shows a cumulative relative frequency plot for all mononucleotide repeat types. In the poly-A (blue) and poly-T (violet) repeats, 95% of the detected repeats consist of repeats smaller than or equal to 15 base pairs. In the poly-C (green) and poly-G (yellow) repeats, 95% of the detected repeats comprise repeats smaller than or equal to 7 base pairs.

We investigated the repeat densities of all mononucleotide repeats and stratified them by their repeat length. The repeat densities of poly-A and poly-C repeats are presented in Figure 5 and 6, respectively. Figures 17 and 18 in the supplementary material provide the same representation for poly-G and poly-T repeats, respectively.

The repeat density between the recombination hotspots (HS Zone) and their surrounding zones (Zone 1-5, Outside Zone) is compared in panel A and C. Here, the repeat densities are stratified by repeat length, where the frequency of each repeat is divided by the zone length (panel C) and then divided by the sum of the densities per repeat length (panel A). The repeat densities of poly-A and poly-T repeats decrease with increasing distance from the hotspot, except for the outside zone (Figures 5C and 18C). Across the repeat length range [6,12], the detected repeats are more enriched in the outside zone (Figures 5D and 18D). In contrast, poly-C and poly-G densities show a decrease in repeat densities with increasing distance from the hotspot, including the outside zone (Figures 6C-D and 17C-D). The decreasing repeat density is further elucidated in panel B, which presents the variation in repeat density per zone. In panel D, the repeat densities per zone are visualized relative to the hotspot. Here, the repeat density of each repeat length in a zone is divided by the repeat density of the hotspot zone of the same repeat length.

We tested for differences in repeat densities with a Kruskal-Wallis test and identified statistically significant differences ($p < 0.001$) for all DNA bases. Multiple comparisons using Dunn's test and error control using Holm's method are listed in Table 3. We observed that the repeat density of the hotspot zone of poly-C and poly-G repeats is significantly different ($p < 0.006$) from zone 5 and the outside zone, which persists after controlling the error rate. Similarly, poly-A and poly-T repeats show significant differences ($p < 0.001$) between the hotspot zone and zones 3-5, but not the outside zone. Here, the outside zone shows a significantly larger enrichment ($p < 0.003$) in repeat density compared to zones 3-5.

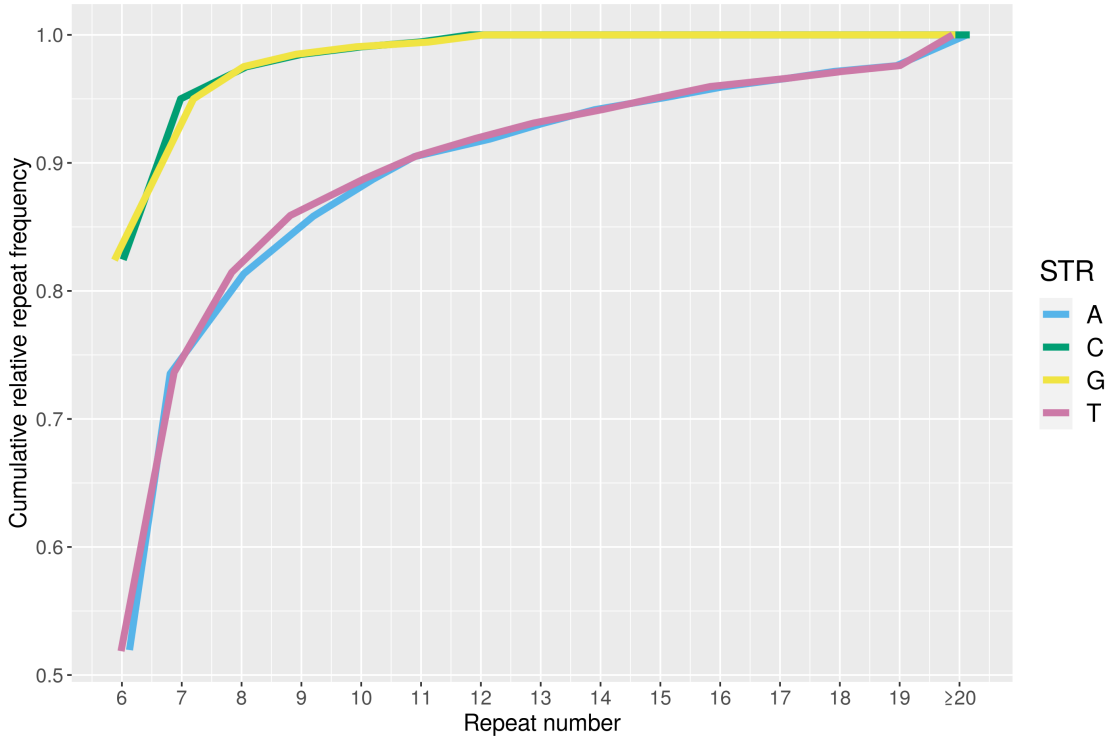


Figure 4: A cumulative relative frequency plot for all mononucleotide repeat types (A (blue), C (green), G (yellow), T (violet)) that shows the distribution of repeat frequency per repeat number.

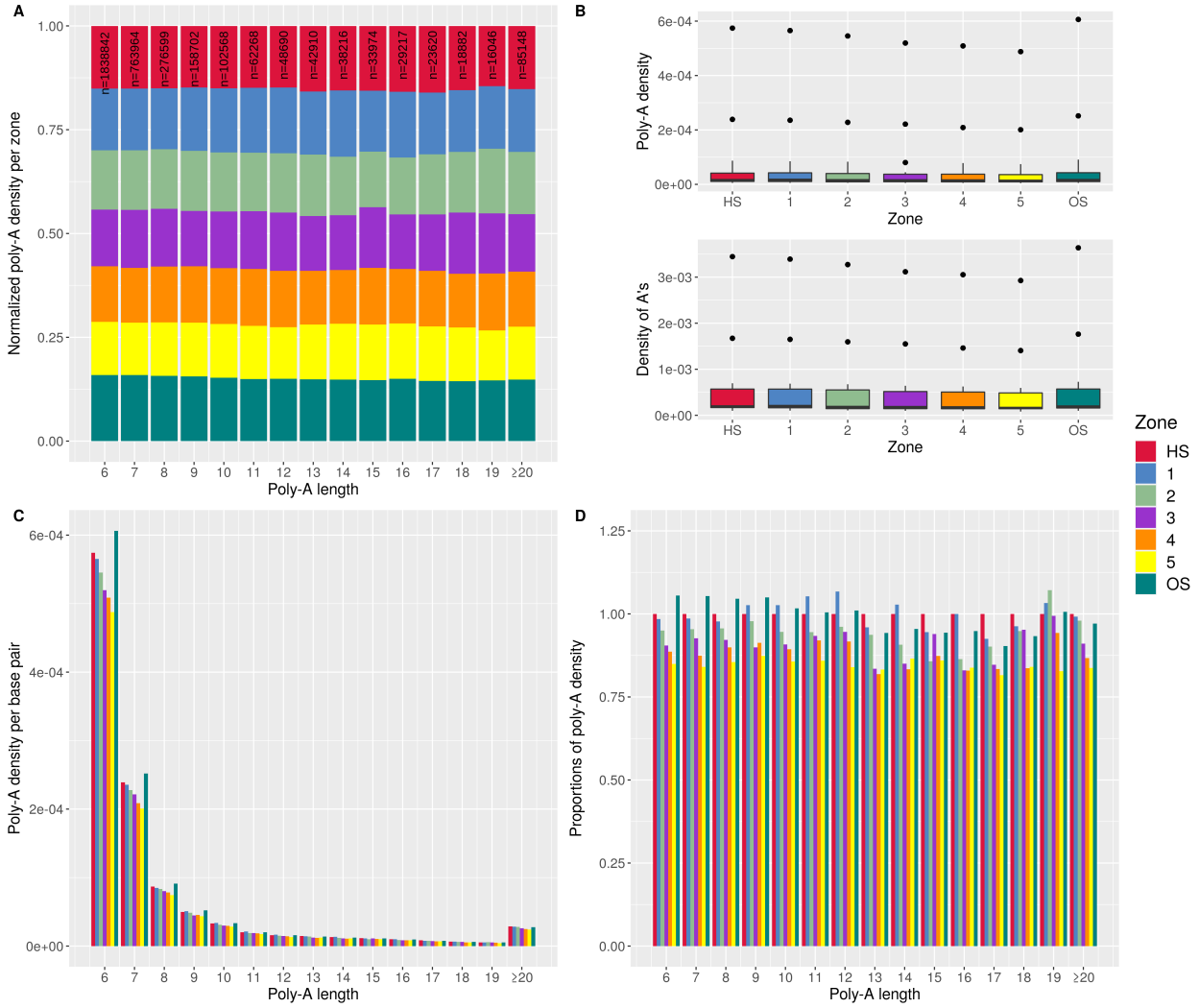


Figure 5: Densities of poly-A repeats (in version GRCh38/hg38 of the human reference genome). The poly-A density per zone of a certain repeat is obtained by dividing the frequency of the repeat by the length of the zone it is found in. Repeat types longer than 20 base pairs are grouped with repeats equal to 20 due to their low frequency. Figure 5A presents the normalized poly-A density, where the poly-A density per repeat per zone is divided by the total sum of poly-A densities across all seven zones (Hotspot Zone, Zones 1-5, Outside Zone) of the poly-A repeat. The total count of a certain repeat found across all seven zones is shown at the top of each bar of the plot. The top panel of Figure 5B describes the distribution of poly-A densities of all repeat lengths found in a given zone. The bottom panel of Figure 5B shows the density of A's of all repeat types found in a given zone, where the density of A's is obtained by multiplying the length of a repeat by its frequency and dividing it by the length of the zone it is in. Figure 5C presents the poly-A densities per base pair of all zones, stratified with respect to the repeat length. Figure 5D presents the poly-A densities of each zone (Hotspot Zone, Zones 1-5, Outside Zone) relative to the hotspot zone.

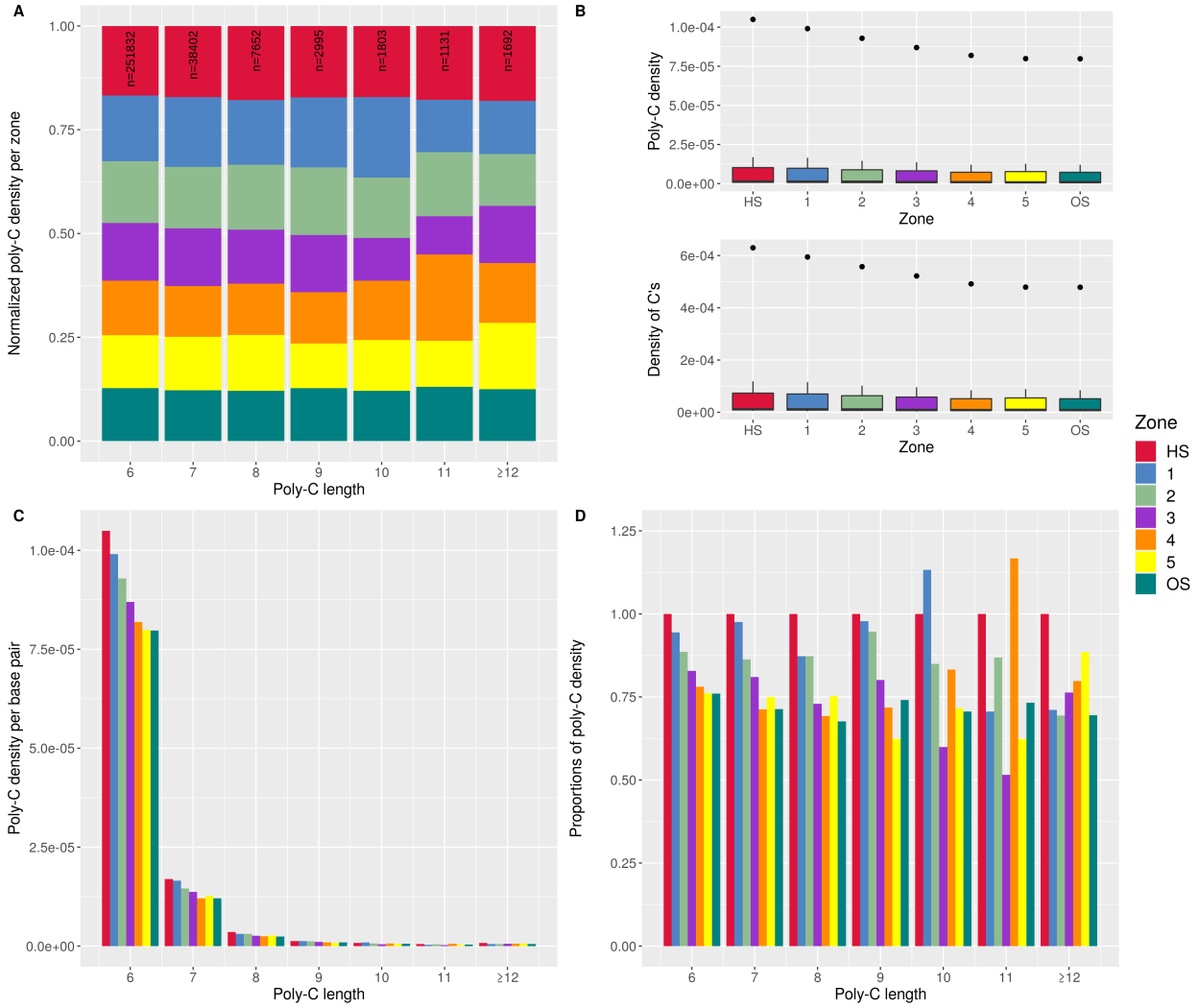


Figure 6: Densities of poly-C repeats (in version GRCh38/hg38 of the human reference genome). The poly-C density per zone of a certain repeat is obtained by dividing the frequency of the repeat by the length of the zone it is found in. Repeat types longer than 12 base pairs are grouped with repeats equal to 12 due to their low frequency. Figure 6A presents the normalized poly-C density, where the poly-C density per repeat per zone is divided by the total sum of poly-C densities across all seven zones (Hotspot Zone, Zones 1-5, Outside Zone) of the poly-C repeat. The total count of a certain repeat found across all seven zones is shown at the top of each bar of the plot. The top panel of Figure 6B describes the distribution of poly-C densities of all repeat lengths found in a given zone. The bottom panel of Figure 6B shows the density of C's of all repeat types found in a given zone, where the density of C's is obtained by multiplying the length of a repeat by its frequency and dividing it by the length of the zone it is in. Figure 6C presents the poly-C densities per base pair of all zones, stratified with respect to the repeat length. Figure 6D presents the poly-C densities of each zone (Hotspot Zone, Zones 1-5, Outside Zone) relative to the hotspot zone.

STR	Comparison	Z	P.adj
A	1 - 3	4.06	<0.001
A	1 - 4	5.42	<0.001
A	2 - 4	3.46	0.006
A	1 - 5	6.95	<0.001
A	2 - 5	4.99	<0.001
A	3 - 5	2.89	0.038
A	3 - HS	-4.58	<0.001
A	4 - HS	-5.94	<0.001
A	5 - HS	-7.47	<0.001
A	3 - OS	-3.76	0.002
A	4 - OS	-5.12	<0.001
A	5 - OS	-6.65	<0.001
C	3 - HS	-3.40	0.013
C	5 - HS	-3.65	0.005
C	HS - OS	4.17	<0.001
G	1 - 5	3.45	0.010
G	5 - HS	-3.63	0.005
G	1 - OS	3.96	0.002
G	HS - OS	4.13	<0.001
T	1 - 3	3.87	0.001
T	1 - 4	4.63	<0.001
T	1 - 5	6.19	<0.001
T	2 - 5	3.89	0.001
T	2 - HS	-3.72	0.002
T	3 - HS	-5.29	<0.001
T	4 - HS	-6.05	<0.001
T	5 - HS	-7.61	<0.001
T	3 - OS	-4.18	<0.001
T	4 - OS	-4.94	<0.001
T	5 - OS	-6.49	<0.001

Table 3: Significant differences in mononucleotide repeat densities using Dunn's multiple comparison with Holm's error control among the hotspot zone (HS), the surrounding five zones (1-5), and the outside zone (OS) in the human reference genome (version GRCh38/hg38) are shown. The repeat type, the zone comparison, the Z-value, and the adjusted p-value of the Dunn's test are listed.

4.2 Dinucleotide repeats

We screened the human reference genome for the dinucleotide repeat types AT, CT, AG, GT, AC, and CG, where the dinucleotide motif is consecutively repeated at least three times. We excluded the remaining ten combinations, since four of them (AA, CC, GG, TT) are already screened for in the analysis of mononucleotide repeats, and the remaining six repeat types (TA, TC, GA, TG, CA, and GC) are reversed patterns of the analyzed dinucleotide repeats.

Figure 7 provides a comparison of repeat densities of the analyzed dinucleotide repeat types. Among the analyzed motifs, the poly-AT repeats occur most frequently, the motifs CT, AG, GT, and AC appear in similar repeat densities, and CG-repeats exist in very low repeat densities. Figure 8 shows a cumulative relative frequency plot for all dinucleotide repeat types. In the poly-AT repeats (green), 95% of the detected repeats comprise repeats smaller than or equal to 20 base pairs. In the poly-CG (yellow), poly-AG (blue) and poly-CT (red) repeats, 95% of the detected repeats consist of repeats smaller than or equal to 16 base pairs. In the poly-GT (violet) and poly-AC (orange) repeats, 95% are contained in repeats smaller than or equal 14 base pairs.

In Figure 9, we further investigate the repeat densities of the poly-AT repeat equivalently to the mononucleotide repeats. Similarly to the poly-A and poly-T repeats, the poly-AT repeat density seems to decrease with increasing distance from the hotspot for the five zones directly surrounding the hotspot (Figure 9C). In the repeat length ranging from 6 to 16 base pairs, there is a noticeable difference in repeat densities between the outside zone and the hotspot zone (Figure 9D). We do not observe a specific pattern of enrichment in the variable repeat densities occurring in repeat lengths of 22-28 base pairs (Figures 9A and 9D).

The same representation of the dinucleotide repeat types CT, AG, GT, and AC is shown in supplementary Figures 19-22, respectively. Here, we observe a pattern distinct from the mononucleotide analysis and the poly-AT repeat. For shorter repeats (6-16 base pairs in length), the repeat density decreases with increasing distance to the hotspot zone, except for the outside zone (Figures 19-22C). In contrast to the poly-AT repeat, the repeat density of the outside zone is consistently lower than the hotspot zone, but is not the lowest among all zones (Figures 19-22 C-D).

The poly-CG repeat occurs with the smallest frequency compared to all other dinucleotide repeats (Figure 7). Similarly to poly-C and poly-G repeats, the repeat density of the hotspot zone is the highest compared to all other zones including the outside zone (Figures 10B-C). Compared to other dinucleotide repeats, the repeat densities stratified by repeat length (Figure 10C) show a stronger decrease in repeat frequency with increasing repeat length, which could be explained by the higher mutation rate in those repeats (Fryxell & Moon, 2004) leading to interrupted repeat segments that are not detected in our analysis.

We tested for differences in repeat densities with a Kruskal-Wallis test and identified statistically significant differences ($p < 0.004$) for all six repeat combinations. Multiple comparisons using Dunn's test and error control using Holm's method are listed in Table 4. Among these, the repeat densities of repeats containing AT, GT, CT, and CG show the hotspot zone to be significantly different ($p < 0.05$) from zones 3-5 after p-values are adjusted. Intriguingly, the outside zone of the poly-AT repeat is significantly different ($p < 0.001$) from zones 2-5. Similarly, the repeat densities of poly-GT and poly-AC repeats show significant differences ($p < 0.02$) between the outside zone and zone 5.

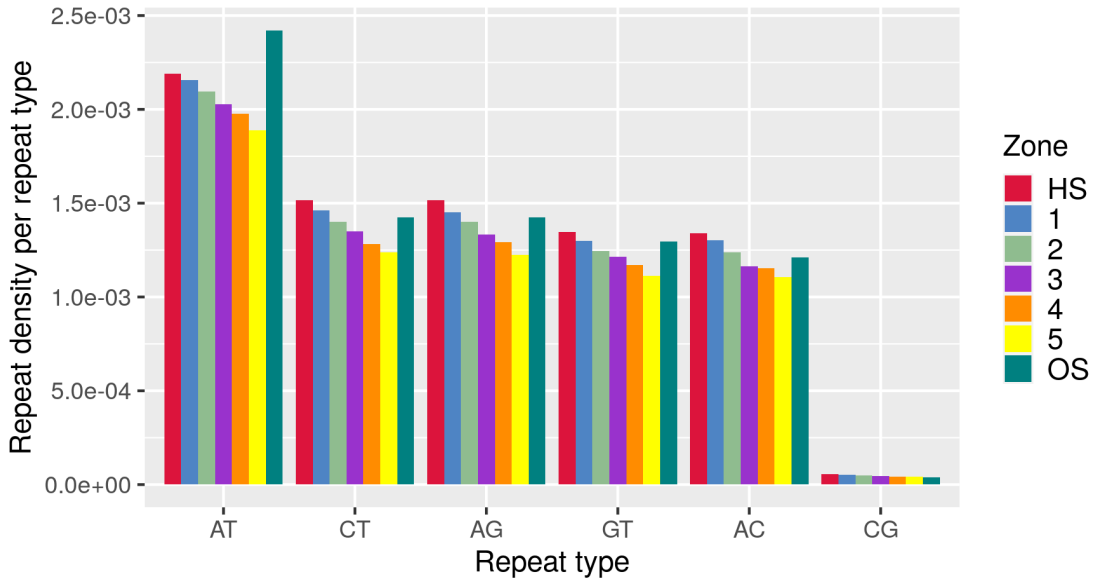


Figure 7: Repeat density per dinucleotide repeat type. In this figure, the repeat densities of six dinucleotide repeat types (AT, CT, AG, GT, AC, and CG) of the human reference genome (version GRCh38/hg38) are shown. The repeat density per zone of a repeat type is obtained by dividing the frequency of the repeat by the length of the zone it is found in.

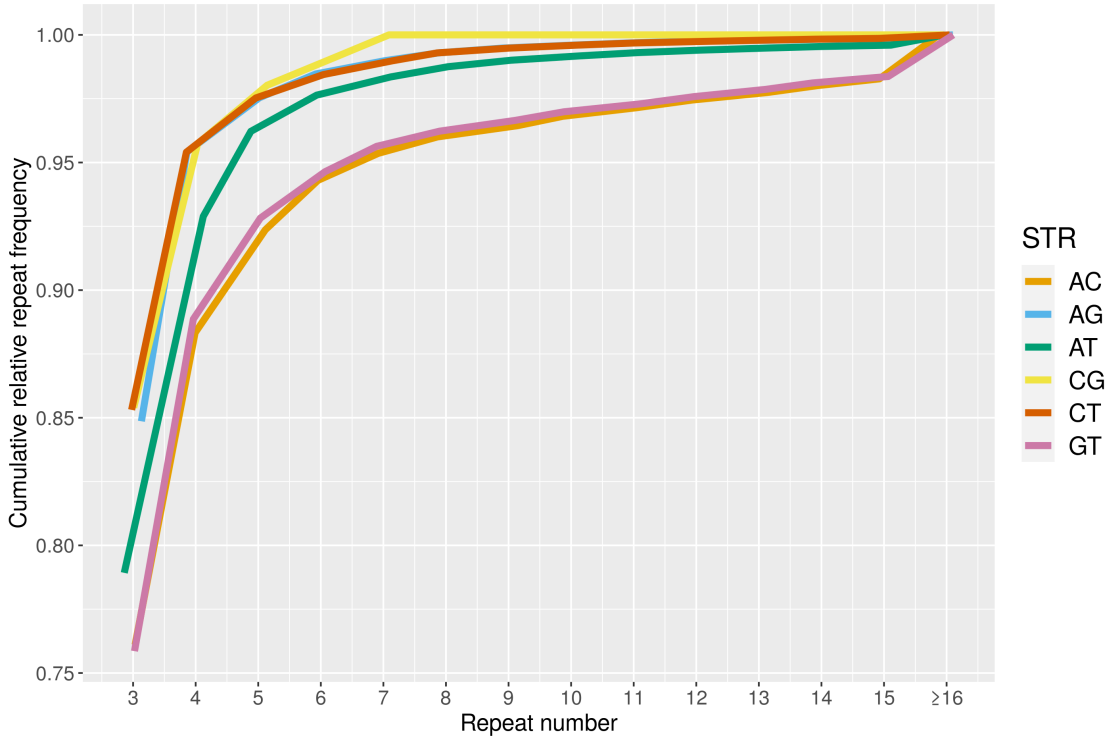


Figure 8: A cumulative relative frequency plot for all dinucleotide repeat types (AC (orange), AG (blue), AT (green), CG (yellow), CT (red), GT (violet)) that shows the distribution of repeat frequency per repeat number.

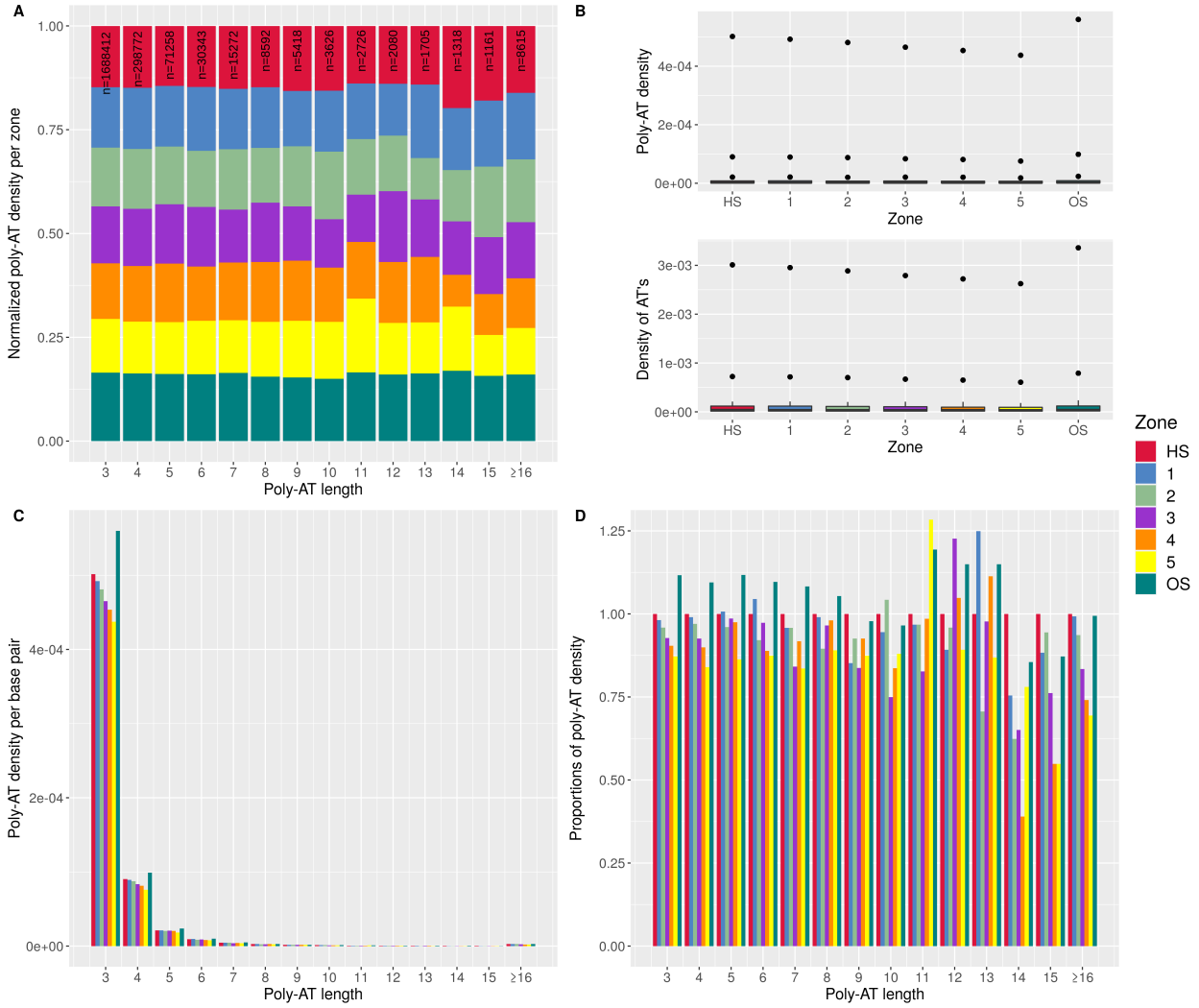


Figure 9: Densities of poly-AT repeats (in version GRCh38/hg38 of the human reference genome). The poly-AT density per zone of a certain repeat is obtained by dividing the frequency of the repeat by the length of the zone it is found in. Repeat types longer than 16 base pairs are grouped with repeats equal to 16 due to their low frequency. Figure 9A presents the normalized poly-AT density, where the poly-AT density per repeat per zone is divided by the total sum of poly-AT densities across all seven zones (Hotspot Zone, Zones 1-5, Outside Zone) of the poly-AT repeat. The total count of a certain repeat found across all seven zones is shown at the top of each bar of the plot. The top panel of Figure 9B describes the distribution of poly-AT densities of all repeat lengths found in a given zone. The bottom panel of Figure 9B shows the density of AT's of all repeat types found in a given zone, where the density of AT's is obtained by multiplying the length of a repeat by its frequency and dividing it by the length of the zone it is in. Figure 9C presents the poly-AT densities per base pair of all zones, stratified with respect to the repeat length. Figure 9D presents the poly-AT densities of each zone (Hotspot Zone, Zones 1-5, Outside Zone) relative to the hotspot zone.

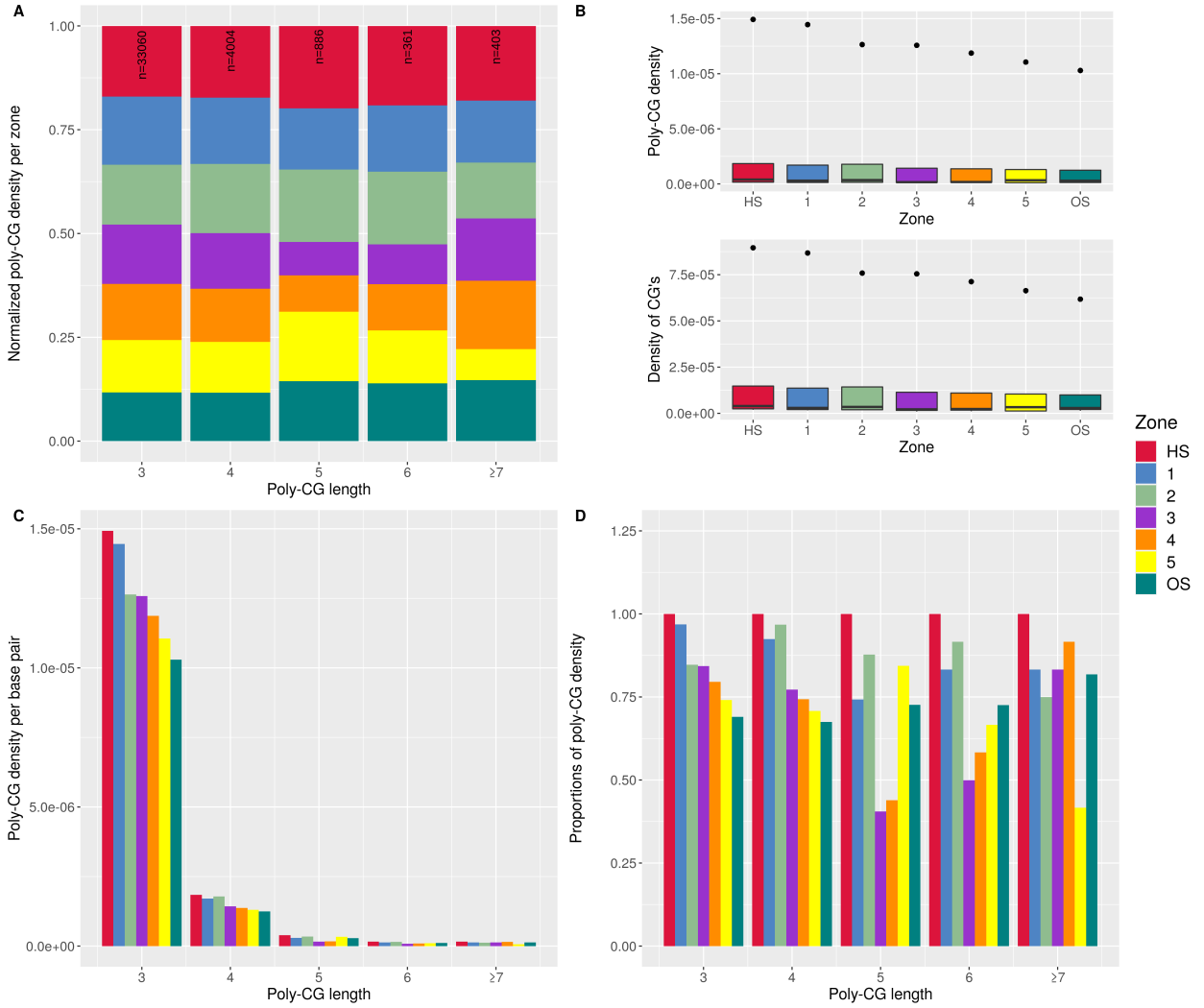


Figure 10: Densities of poly-CG repeats (in version GRCh38/hg38 of the human reference genome). The poly-CG density per zone of a certain repeat is obtained by dividing the frequency of the repeat by the length of the zone it is found in. Repeat types longer than 7 base pairs are grouped with repeats equal to 7 due to their low frequency. Figure 10A presents the normalized poly-CG density, where the poly-CG density per repeat per zone is divided by the total sum of poly-CG densities across all seven zones (Hotspot Zone, Zones 1-5, Outside Zone) of the poly-CG repeat. The total count of a certain repeat found across all seven zones is shown at the top of each bar of the plot. The top panel of Figure 10B describes the distribution of poly-CG densities of all repeat lengths found in a given zone. The bottom panel of Figure 10B shows the density of CG's of all repeat types found in a given zone, where the density of CG's is obtained by multiplying the length of a repeat by its frequency and dividing it by the length of the zone it is in. Figure 10C presents the poly-CG densities per base pair of all zones, stratified with respect to the repeat length. Figure 10D presents the poly-CG densities of each zone (Hotspot Zone, Zones 1-5, Outside Zone) relative to the hotspot zone.

STR	Comparison	Z	P.adj
AC	1 - 3	3.14	0.029
AC	1 - 5	3.82	0.003
AC	3 - HS	-3.72	0.004
AC	5 - HS	-4.40	<0.001
AC	5 - OS	-3.28	0.019
AG	1 - 5	3.05	0.042
AG	2 - HS	-3.44	0.012
AG	4 - HS	-3.13	0.033
AG	5 - HS	-4.54	<0.001
GT	1 - 5	3.62	0.005
GT	2 - 5	3.16	0.023
GT	3 - HS	-4.43	<0.001
GT	4 - HS	-3.45	0.009
GT	5 - HS	-5.17	<0.001
GT	3 - OS	-3.37	0.012
GT	5 - OS	-4.12	<0.001
CG	3 - HS	-3.26	0.024
CG	4 - HS	-3.18	0.028
CG	5 - HS	-3.24	0.024
CT	1 - 5	3.49	0.009
CT	2 - HS	-3.86	0.002
CT	3 - HS	-3.47	0.009
CT	4 - HS	-4.14	<0.001
CT	5 - HS	-5.01	<0.001
AT	1 - 5	2.95	0.044
AT	3 - HS	-2.98	0.043
AT	4 - HS	-3.11	0.030
AT	5 - HS	-3.78	0.003
AT	2 - OS	-3.57	0.006
AT	3 - OS	-4.46	<0.001
AT	4 - OS	-4.59	<0.001
AT	5 - OS	-5.25	<0.001

Table 4: Significant differences in dinucleotide repeat densities using Dunn's multiple comparison with Holm's error control among the hotspot zone (HS), the surrounding five zones (1-5), and the outside zone (OS) in the human reference genome (version GRCh38/hg38) are shown. The repeat type, the zone comparison, the Z-value, and the adjusted p-value of the Dunn's test are listed.

4.3 Trinucleotide repeats

We screened the human reference genome for the full set of trinucleotide repeat combinations, in which the trinucleotide motif is at least two times consecutively repeated. We excluded the repeat types AAA, CCC, GGG, and TTT, as they are included in the mononucleotide repeat analysis. The repeat densities stratified by repeat length are available at `https://github.com/MehdinMasinovic/Human-Genome-STR-Screening`.

The repeat densities for all screened repeat combinations are presented in Figure 11. The repeat densities stratified by repeat length are shown in Figures 12-13. We observe a higher repeat density for trinucleotide repeats rich in either adenosine or thymine, and the motifs ATT, AAT, AGA, TCT, TTC, GAA, TAT, ATA, TTA, and TAA represent the set with the highest repeat density. The A/T-rich repeats are followed by G/C-rich repeat types, namely CCT, AGG, CCA, TGG, GGA, TCC, CAG, CTG, GTG, CAC, CTC, and GAG. Furthermore, the set of trinucleotide repeats comprising solely cytosine and guanine - namely GCC, CGC, CCG, CGG, GCG, and GGC - are present in very low repeat densities, and the repeat types ACG, CGT, CGA, and TCG exist in the lowest density in the human genome.

In all repeat combinations, the repeat density of the hotspot zone is higher than the five zones directly flanking the hotspot. Intriguingly, if at least two positions of a motif are filled with either adenosine or thymine, the repeat density of the outside zone is consistently higher or equal to the hotspot zone; the largest difference being present in the motifs ATT, AAT, TAT, ATA, TTA, and TAA. In contrast, the repeat density of the outside zone is consistently smaller than the hotspot zone and Zone 1, and persistently smaller or equal than Zone 2 in repeat combinations where at least two positions comprise cytosine or guanine; the strongest difference being apparent in the motifs CCA, AGG, TCC, CAC, GAG, and CTG.

A Kruskal-Wallis test suggests statistically significant differences ($p < 0.05$) among zones in the densities of A/T-rich repeats such as ATT, AAT, TAT, and ATA. Similarly, G/C-rich repeats such as CCA, TCC, GAG, and CAC have significant differences ($p < 0.05$) in repeat densities among the hotspot zone, the five zones surrounding it, and the outside zone. The complete list of all trinucleotide repeat combinations tested with a Kruskal-Wallis test is shown in Table 5. Multiple comparisons using Dunn's test and error control using Holm's method are listed in Table 6.

Repeats containing the motif AAG, TTA, GAT, and TGA show the hotspot zone and Zone 1 to be significantly different ($p < 0.05$) from Zone 5. The repeat densities of the repeat-triplets (CAA, ACA, AAC) and (GTT, TGT, TTG) result in significant differences ($p < 0.05$) between the outside zone and Zone 4, and repeats containing (ATT, TAT, TTA) show significant differences ($p < 0.05$) between the outside zone and Zone 5.

We observed particularly interesting differences in repeats with the motifs (TTG, TGT, GTT), (AAC, CAA), TAG, and TCT, where only the outside zone resulted in significant differences ($p < 0.05$) compared to the surrounding zones, which suggests possible functional activities of these repeat types in regions unrelated to recombination. Moreover, the hotspot zone is significantly different from the outside zone in G/C-rich repeats such as CGC, CCG, and GAG, which might indicate possible functionalities of these repeat types in recombination.

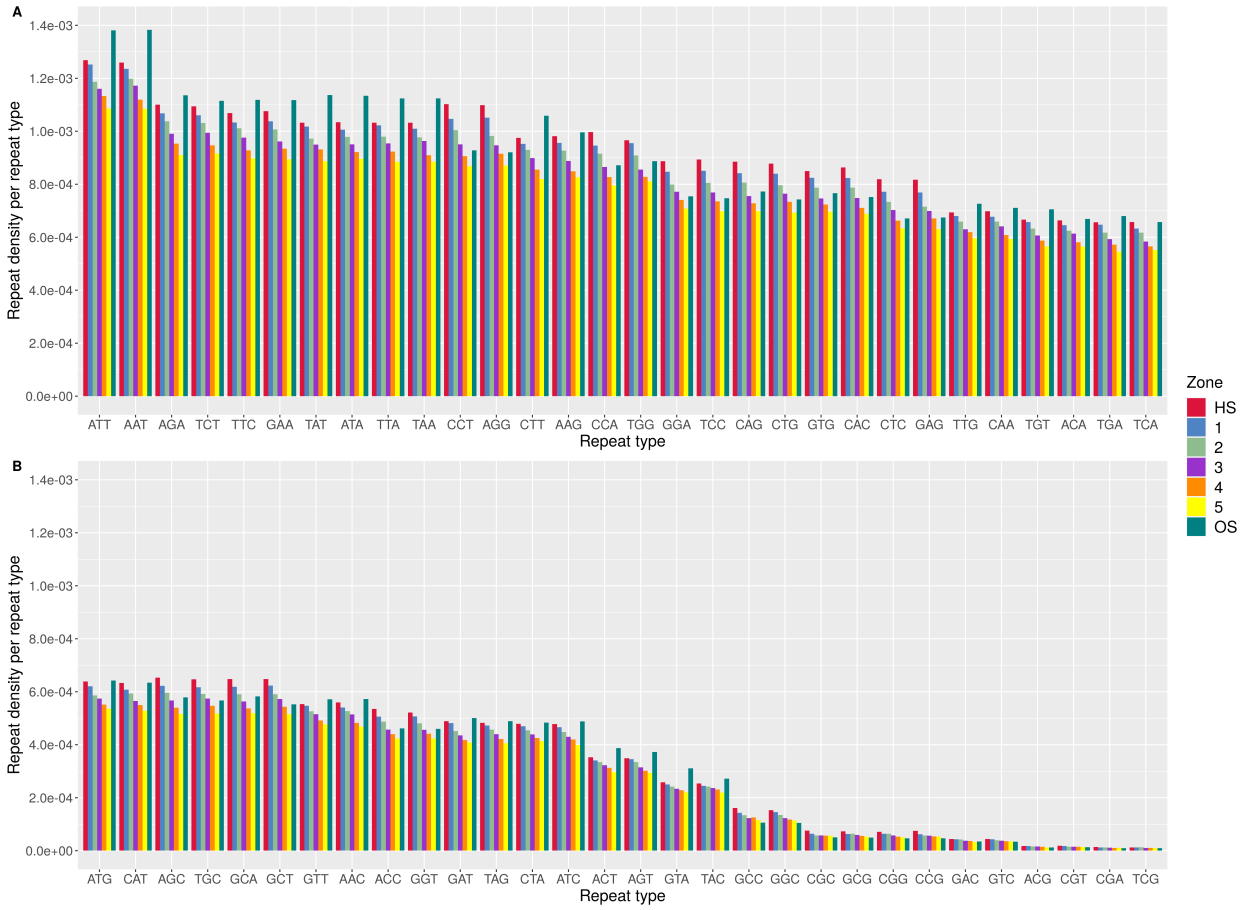


Figure 11: Repeat density per trinucleotide repeat type. The repeat densities of all distinct trinucleotide repeat types of the human reference genome (version GRCh38/hg38) are shown, excluding the repeat types AAA, CCC, GGG, and TTT. The repeat density per zone of a repeat is obtained by dividing the frequency of the repeat by the length of the zone.

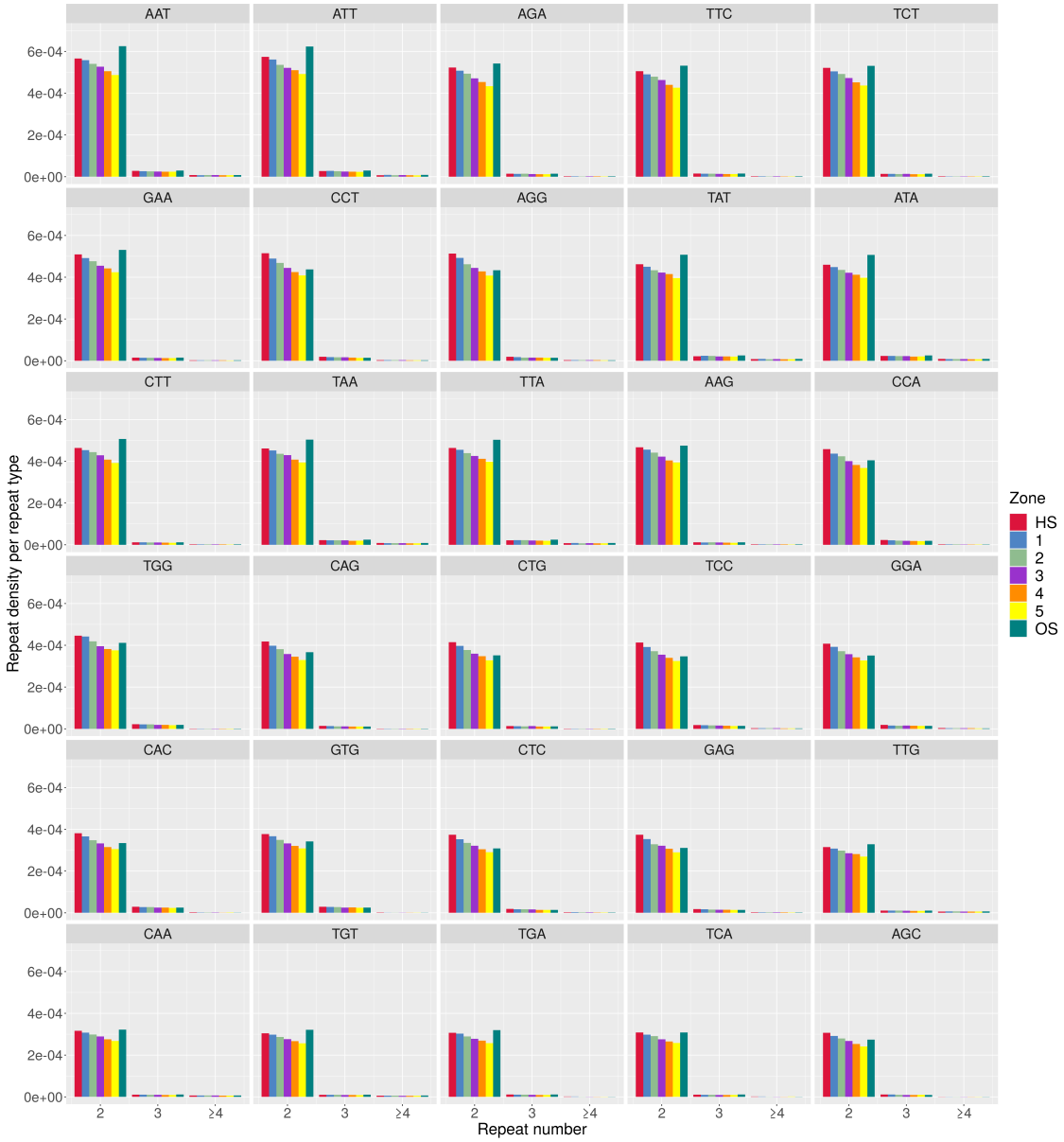


Figure 12: Repeat densities of trinucleotide repeats stratified by repeat length. The repeat densities of 30 distinct trinucleotide repeat types of the human reference genome (version GRCh38/hg38) are shown. The repeat density of a repeat is obtained by dividing the frequency of a repeat by the length of the zone. Repeat numbers longer than or equal to 4 are grouped with repeats equal to 4 repeat numbers due to their low frequency.

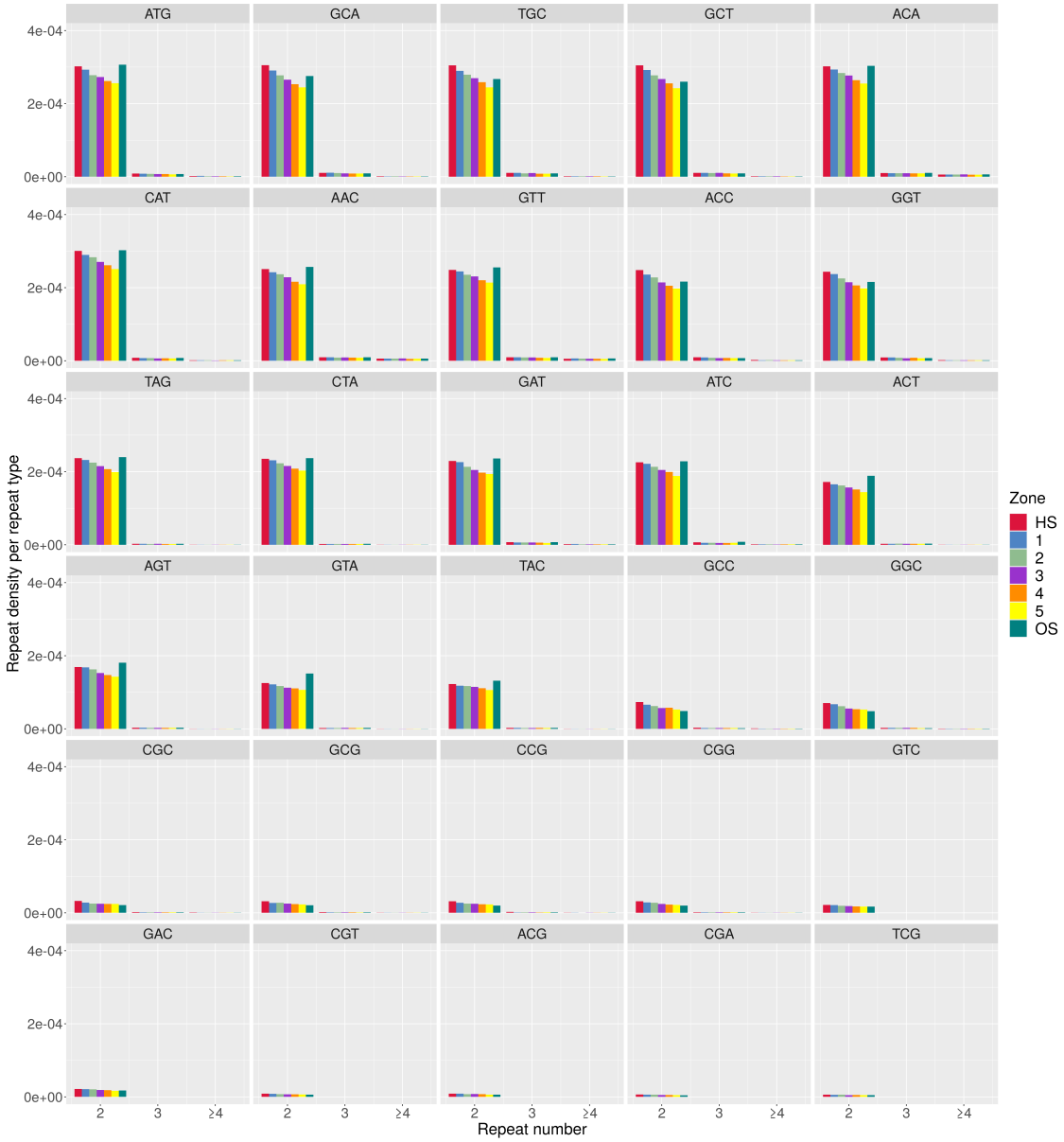


Figure 13: Repeat densities of trinucleotide repeats stratified by repeat length. The repeat densities of 30 distinct trinucleotide repeat types of the human reference genome (version GRCh38/hg38) are shown. The repeat density of a repeat is obtained by dividing the frequency of a repeat by the length of the zone. Repeat numbers longer than or equal to 4 are grouped with repeats equal to 4 repeat numbers due to their low frequency.

STR	Chi-Squared	P-Value	STR	Chi-Squared	P-Value	STR	Chi-Squared	P-Value
CAA	**15.03**	**0.020**	**GCC**	**13.33**	**0.038**	**TGG**	**14.54**	**0.024**
GAA	**21.54**	**0.001**	**TCC**	**16.97**	**0.009**	**ATG**	**29.11**	**<0.001**
TAA	11.68	0.070	AGC	8.79	0.186	**CTG**	**16.31**	**0.012**
ACA	**20.12**	**0.003**	**CGC**	**17.08**	**0.009**	**GTG**	**18.88**	**0.004**
CCA	7.10	0.311	GGC	10.70	0.098	**TTG**	**18.54**	**0.005**
GCA	**15.17**	**0.019**	TGC	10.72	0.097	**AAT**	**14.90**	**0.021**
TCA	7.88	0.247	ATC	9.32	0.156	CAT	11.21	0.082
AGA	**20.28**	**0.002**	**CTC**	**13.54**	**0.035**	**GAT**	**26.59**	**<0.001**
CGA	4.57	0.600	GTC	8.52	0.203	**TAT**	**25.72**	**<0.001**
GGA	11.52	0.073	**TTC**	**20.54**	**0.002**	ACT	2.84	0.829
TGA	**30.44**	**<0.001**	**AAG**	**23.30**	**<0.001**	CCT	12.42	0.053
ATA	10.57	0.103	CAG	10.92	0.091	**GCT**	**14.71**	**0.023**
CTA	3.19	0.784	**GAG**	**14.56**	**0.024**	**TCT**	**15.43**	**0.017**
GTA	**15.36**	**0.018**	**TAG**	**14.77**	**0.022**	AGT	12.40	0.054
TTA	**33.17**	**<0.001**	ACG	9.21	0.162	CGT	6.64	0.355
AAC	**17.44**	**0.008**	**CCG**	**24.00**	**<0.001**	GGT	10.06	0.122
CAC	**15.28**	**0.018**	GCG	5.10	0.531	**TGT**	**13.69**	**0.033**
GAC	10.74	0.097	TCG	4.47	0.614	**ATT**	**30.71**	**<0.001**
TAC	4.82	0.567	**AGG**	**14.18**	**0.028**	CTT	12.16	0.058
ACC	5.86	0.439	CGG	8.56	0.200	**GTT**	**13.08**	**0.042**

Table 5: The trinucleotide repeat type, chi-squared value, and the p-value of a Kruskal-Wallis test are listed, which examines possible differences in repeat density in the human reference genome (version GRCh38/hg38) among the hotspot zone, the surrounding five zones, and the outside zone. A repeat type is set in bold if the p-value is below the significance threshold ($p<0.05$).

STR	Comparison	Z	P.adj
TTA	1 - 2	3.25	0.020
TTA	1 - 5	4.02	0.001
TTA	3 - 5	3.42	0.011
TTA	5 - HS	-3.07	0.034
TTA	2 - OS	-3.47	0.010
TTA	5 - OS	-4.24	<0.001
ATT	1 - 2	3.30	0.017
ATT	1 - 5	3.57	0.007
ATT	3 - 5	3.22	0.022
ATT	2 - OS	-3.74	0.004
ATT	5 - OS	-4.01	0.001
TAT	1 - 2	3.39	0.012
TAT	1 - 5	3.47	0.010
TAT	2 - OS	-3.54	0.008
TAT	5 - OS	-3.62	0.006
AAG	1 - 5	3.39	0.014
AAG	3 - 5	3.09	0.038
AAG	5 - HS	-4.16	<0.001
GAA	1 - 5	3.27	0.021
GAA	5 - HS	-3.82	0.003
TTG	4 - OS	-3.83	0.003
ATG	1 - 4	3.29	0.019
ATG	1 - 5	3.98	0.001
ATG	4 - HS	-3.26	0.020
ATG	5 - HS	-3.94	0.002
TGA	1 - 4	3.37	0.014
TGA	1 - 5	4.12	<0.001
TGA	4 - HS	-3.40	0.013
TGA	5 - HS	-4.16	<0.001
CGC	2 - HS	-3.45	0.012
CGC	3 - HS	-3.22	0.024
CGC	HS - OS	3.32	0.018
CCG	2 - HS	-4.17	<0.001
CCG	3 - HS	-3.65	0.005
CCG	HS - OS	3.74	0.004
TTC	5 - HS	-3.49	0.009
TTC	5 - OS	-4.08	<0.001
GAG	3 - HS	-3.07	0.045
GAG	HS - OS	3.03	0.049
ACA	4 - HS	-3.10	0.039
ACA	4 - OS	-3.99	0.001
GTG	3 - HS	-3.10	0.038
GTG	5 - HS	-3.38	0.015
CTG	1 - 5	3.07	0.043
CTG	1 - OS	3.17	0.032
TAG	4 - OS	-3.16	0.033
CAA	4 - OS	-3.44	0.012
GCA	5 - HS	-3.10	0.040
AGA	5 - HS	-3.61	0.006
AAC	4 - OS	-3.76	0.004
CAC	5 - HS	-3.49	0.010
TCC	5 - HS	-3.19	0.030
TGG	5 - HS	-3.51	0.010
AAT	4 - HS	-3.34	0.017
GAT	1 - 5	3.95	0.002
GAT	5 - HS	-4.12	<0.001
TCT	5 - OS	-3.28	0.022
TGT	4 - OS	-3.42	0.013
GTT	4 - OS	-3.18	0.031

Table 6: Significant differences in trinucleotide repeat densities using Dunn's multiple comparison with Holm's error control among the hotspot zone (HS), the surrounding five zones (1-5), and the outside zone (OS) in the human reference genome are shown. The repeat type, the zone comparison, the Z-value, and the adjusted p-value of the Dunn's test are listed if the adjusted p-value is below the significance threshold ($p<0.05$).

4.4 Tetranucleotide repeats

Additionally to the whole genome analysis of mono-, di-, and trinucleotide repeats, we searched for all possible combinations of tetranucleotide repeats. The full set of distinct combinations of tetranucleotide repeats reaches the count of 240 repeats, excluding repeat types that have similarly been screened for in the mono- and dinucleotide repeat sets, such as AAAA, ACAC, CACA, etc. In contrast to repeat types with a repeat unit smaller than four, the tetranucleotide repeats have to occur only once to be considered in the analysis. Figures 14 and 15 represent the most and least abundant tetranucleotide repeat types, respectively. The remaining motifs are visualized in the supplementary Figure 23.

Equivalently to repeat types with a smaller repeat unit (mono-, di-, and trinucleotide repeats), the hotspot zone of the tetranucleotide repeats has the highest repeat density compared to the five zones directly surrounding the hotspot, and the repeat density is decreasing with increasing distance to the hotspot. In accordance to the patterns observed in trinucleotide motifs, the outside zone of A/T-rich repeat combinations is consistently higher or equal to the hotspot zone. This pattern is clearest in motifs with the highest repeat density, namely AAAT, TAAA, GAAA, and TTTG. Contrary to repeats with a smaller repeat unit, there are very few cases with a low repeat density in the outside zone, namely CCTG, CAGG, CTGG, and CCAG. Instead, there seems to be an uniform distribution across all seven zones for most motifs. The low repeat densities shown in Figure 15C suggest a trend of low repeat density for repeat motifs that do not comprise consecutively repeated DNA bases, namely TCGC, TCGA, ATCG, and CGAT among others.

To test for significant differences among zones, we grouped the tetranucleotide repeats into four categories: repeats that have three or four positions filled with either cytosine and guanine or adenosine and thymine. For example, repeats such as ATTT, AAAT, AATA, etc. are grouped into the category of having four positions filled with adenosine or thymine (later on referred to as "ATs4"), and repeats such as ATTC, ATAG, AAAC, etc. are grouped into the category of having three positions filled with adenosine or thymine and one position filled with cytosine or guanine ("ATs3"), and vice versa for repeats filling at least three ("CGs3") or four ("CGs4") positions with cytosine or guanine.

A Kruskal-Wallis test suggests statistically significant differences ($p < 0.05$) in the repeat densities of the categories ATs3, ATs4, CGs3, but not for the CGs4 category ($p = 0.051$), as shown in Table 7. To investigate differences among the hotspot zone, the five zones surrounding it, and the outside zone, we conducted multiple comparisons using Dunn's test and error control using Holm's method, which are listed in Table 8. Here, we observe significant differences between the hotspot zone and Zone 5 in the two categories ATs3 and CGs3. Furthermore, the outside zone of the ATs3 category tested significant against Zones 4 and 5. The significant differences persist after adjusting the p-values.

Figure 14: Repeat density per tetranucleotide repeat type. The repeat densities of 81 distinct tetranucleotide repeat types of the human reference genome (version GRCh38/hg38) are shown. The repeat density per zone of a given repeat is obtained by dividing the frequency of the repeat by the length of the zone.

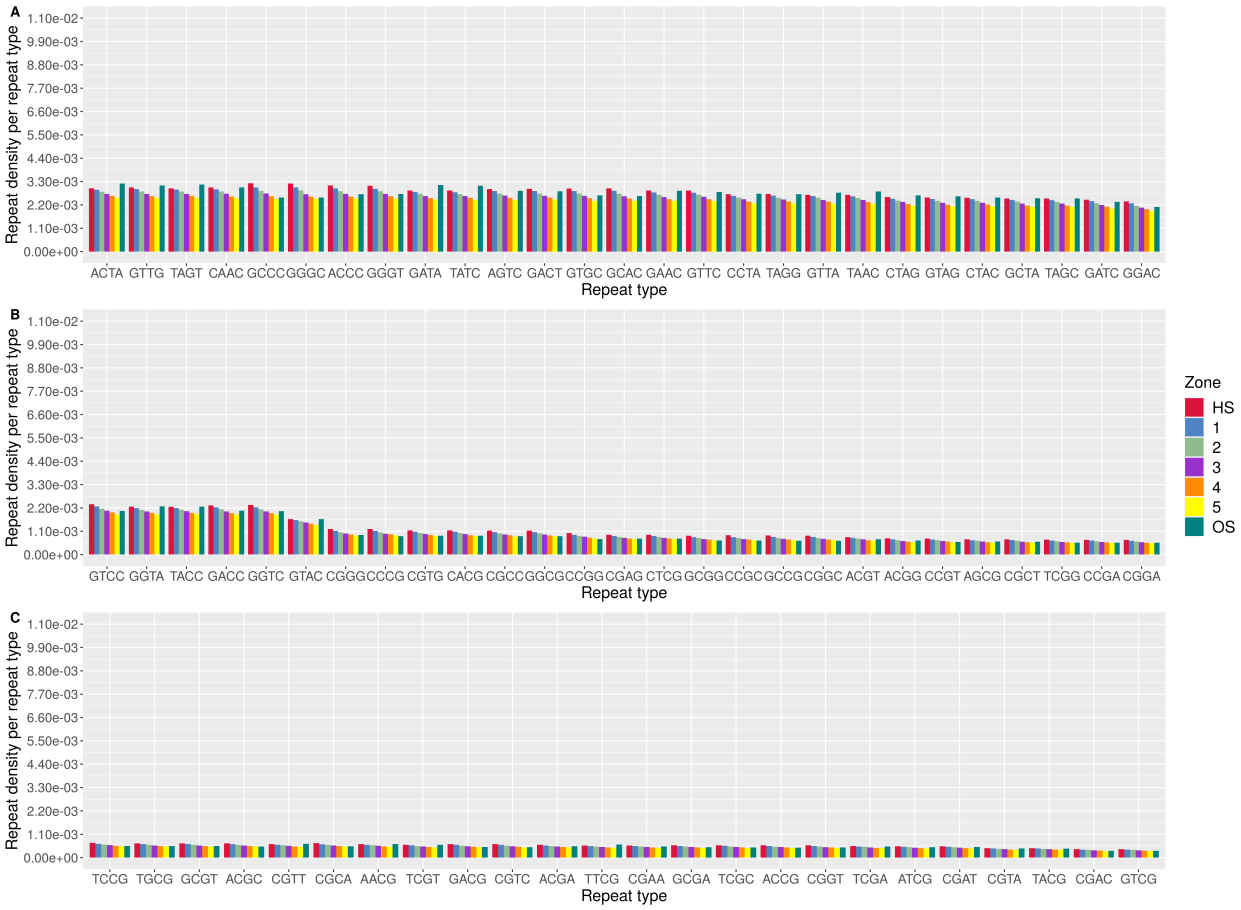


Figure 15: Repeat density per tetranucleotide repeat type. The repeat densities of 78 distinct tetranucleotide repeat types of the human reference genome (version GRCh38/hg38) are shown. The repeat density per zone of a given repeat is obtained by dividing the frequency of the repeat by the length of the zone.

STR	Chi-Squared	P-Value
ATs3	25.52	<0.001
ATs4	12.73	0.047
CGs3	14.02	0.029
CGs4	12.56	0.051

Table 7: The tetranucleotide repeat category, chi-squared value, and the p-value of a Kruskal-Wallis test are listed, which examines possible differences in repeat density in the human reference genome (version GRCh38/hg38) among the hotspot zone, the surrounding five zones, and the outside zone. ATs3 and ATs4 include all repeat types that have three or four positions occupied with adenosine or thymine, respectively. CGs3 and CGs4 include all repeat types that have three or four positions occupied with cytosine or guanine, respectively.

STR	Comp.	Z	P.adj
ATs3	5 - HS	-3.23	0.023
ATs3	4 - OS	-3.46	0.011
ATs3	5 - OS	-4.14	<0.001
CGs3	5 - HS	-3.10	0.041

Table 8: Significant differences in tetranucleotide repeat densities using Dunn's multiple comparison with Holm's error control among the hotspot zone (HS), the surrounding five zones (1-5), and the outside zone (OS) in the human reference genome (version GRCh38/hg38) are shown. The repeat type, the zone comparison, the Z-value, and the adjusted p-value of the Dunn's test are listed. ATs3 and ATs4 include all repeat types that have three or four positions occupied with adenosine or thymine, respectively. CGs3 and CGs4 include all repeat types that have three or four positions occupied with cytosine or guanine, respectively.

4.5 Motifs

Similarly to the screening of tetranucleotide repeats, we analyzed the presence of several DNA motifs, namely (AAAAN, CCCCN, GGGGN, TTTTN) and (AAANAAA, CCCNCCC, GGGNGGG, TTTNTTT), where N stands for all the DNA bases that would not result in a mononucleotide repeat (g.e. AAAAA, CCCCC, etc.) if it is added to the motif, i.e. AAAAG, AAAAC, AAAAT, etc. The relative densities of the analyzed motifs are listed in Figure 16.

In agreement to the repeat types already discussed in this thesis, the repeat densities of the A/T-rich motifs AAAAN and TTTTN are higher than the C/G-rich motifs CCCCN and GGGGN. Furthermore, the outside zone is consistently higher than any other zone in A/T-rich motifs, but lower than the hotspot zone in C/G-rich motifs. A Kruskal-Wallis test suggests no statistically significant differences among zones in the repeat densities of any repeat type, as shown in Table 9.

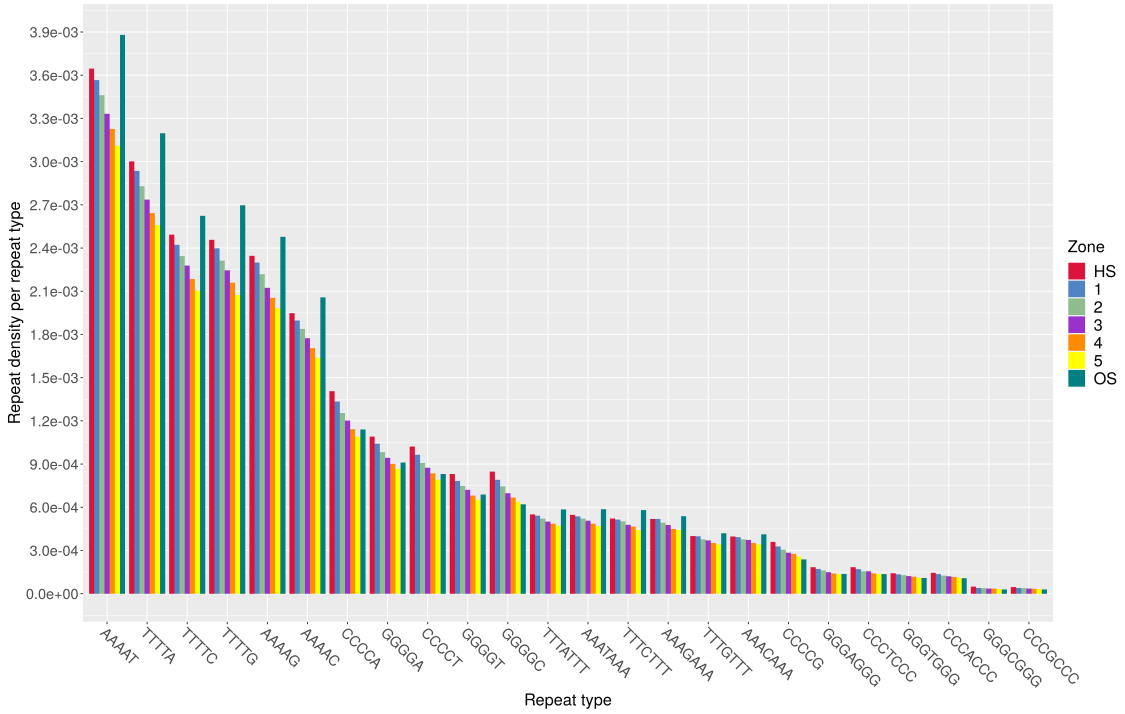


Figure 16: Repeat densities of several DNA motifs. The relative densities of 24 DNA motifs of the human reference genome (version GRCh38/hg38) are shown. The density per zone of a DNA motif is obtained by dividing the frequency of the motif by the length of the zone.

STR	Chi-Squared	P-Value	STR	Chi-Squared	P-Value
AAAAN	2.81	0.833	AAANAAA	5.35	0.500
CCCCN	2.15	0.906	CCCNCCC	3.03	0.805
GGGGN	5.75	0.452	GGGNGGG	3.03	0.805
TTTTN	7.65	0.265	TTTNTTT	5.63	0.466

Table 9: The DNA motif group, chi-squared value, and the p-value of a Kruskal-Wallis test are listed, which examines possible differences in repeat density in the human reference genome (version GRCh38/hg38) among the hotspot zone, the surrounding five zones, and the outside zone. N stands for all the DNA bases that would not result in a mononucleotide repeat (g.e. AAAAA, CCCCC, etc.) if it is added to the motif, i.e. AAAAG, AAAAC, AAAAT, etc.

5 Discussion

Several studies have investigated the frequency and distribution of short tandem repeats across the human genome (Lander et al., 2001), near to genes (Bolton et al., 2013) or promoters (Sawaya et al., 2013), and in other biologically important contexts like disease (reviewed in Polyzos & McMurray, 2017). The purpose of this study was to complement the literature by analyzing repeat patterns in relation to recombination hotspots.

We provided pattern-specific results for all distinct repeat combinations with a repeat unit smaller than or equal to four. Here, the repeats were screened for in the human reference genome to analyze whether there are observable differences in the occurrence of repeats i) within the hotspot zone and the surrounding five regions, ii) zones that are significantly enriched compared to the hotspot zone, or iii) whether certain repeat types are present in other biologically relevant regions (in the outside zone). Similarly, we analyzed the occurrence of all distinct tetranucleotide repeats and certain DNA motifs.

In agreement with the results of a previous study (Subramanian et al., 2003), repeat motifs that consist mainly of adenosine and thymine exist in greatly larger numbers compared to motifs comprising cytosine and guanine. The greatest contrast is present in poly-A/T repeats compared to poly-C/G, poly-AT compared to poly-CG, and poly-ATT/AAT compared to poly-CGG/GGC among others. Moreover, repeat motifs that do not comprise consecutively repeated DNA bases seem to exist in very low densities, such as those in TCG, CGA, CGAC, and TCGA among others.

In all analyzed repeat types, the repeat density generally decreases with increasing distance from the hotspot for zones 1-5. The repeat density further decreases in the outside zone for C/G-rich repeats but not for A/T-rich repeats. In our statistical testing where we assume independence between zones, we observe the hotspot zone and the outside zone to be significantly different from the outermost zones (zones 4 & 5) in C/G-rich repeats and A/T-rich repeats, respectively. Nevertheless, we hardly observe statistically significant differences between the hotspot zone and the outside zone. A significantly higher repeat density in the hotspot zone compared to the outside zone would indicate an enrichment and possibly a functional role of the repeat type in recombination hotspots. In contrast, a significantly higher outside zone compared to other zones would indicate that the detected repeat type could have functional properties in regions unrelated to recombination.

Recombination occurs only in a small proportion (~5%) of the human genome (Myers et al., 2005). In our analysis, the hotspot region under study comprises 3% of the human genome, in contrast to the surrounding five zones (12%) and the outside zone (~85%). Therefore, any significant differences in the outside zone should be interpreted with caution unless more analyses are performed in further research. As was pointed out by Heissl et al., 2019, it is not fully understood to what extent other biological mechanisms could drive the enrichment of short tandem repeats at a broad scale, which makes interpretation of the markedly larger outside zone difficult. To mitigate the effect, we considered the repeat densities relative to the length of the zone. Furthermore, we wish to note that we did not account for multiple testing across the number of tested hypothesis as the results of this study should only provide a first insight into differences regarding repeat structure and their location with respect to recombination hotspots.

In agreement with previous studies, our findings indicate C/G-rich repeats to be enriched in recombination hotspots (Pratto et al., 2014; Majewski & Ott, 2000), whereas A/T-rich repeats seem to exist in higher numbers (Sawaya et al., 2013) but possibly in regions unrelated to recombination (Myers et al., 2005). Repeat types having C, G, CGC, GAG, or CCG as the repeated segment of DNA result in significant differences between the hotspot zone and the outside zone after error control. Similarly to human promoters (Sawaya et al., 2013), we report a possibly strong association of C/G-rich repeats with recombination hotspots. Here, *STRAH* could be applied to the human genome repeatedly with varying zone length to focus on possible functional regions or chromosome-specific differences.

In the work of Heissl et al., 2019, *STRAH* was used to analyze poly-A and poly-T repeats in relation to recombination hotspots, and a significant enrichment of poly-A and poly-T repeats in recombination hotspots was reported. We detected a coding error in *STRAH* that the authors were not aware of, and we report the newly estimated repeat densities of poly-A repeats in the human reference genome (version GRCh37/hg19) in the supplementary section 8.4.

In our analysis the outside zone was not significantly larger than the hotspot zone for any repeat type, but several A/T-rich repeat types showed markedly noticable differences between the outside zone and the five zones surrounding the hotspot, such as A, T, AT, TTA, and TAT among others. *STRAH* is not restricted to the detection of STRs in recombination hotspots; depending on the position coordinates, it is possible to fully screen any number of biologically relevant regions in a genomic sequence, such as up- and downstream regions of promoters, genes, or SNP-positions. Therefore, *STRAH* could be used to investigate the relative densities of certain repeat types in other regions of the genome.

The tetranucleotide repeat analysis resulted in significant differences for the categories ATs3 and CGs3 - repeats having three positions filled with adenosine or thymine and one position filled with cytosine or guanine ("ATs3"), and vice versa ("CGs3") - between the hotspot zone and Zone 5, and in differences between the outside zone and Zones 4-5. We did not observe significant differences in any of the studied motifs with a repeat unit of longer than four. To further investigate these repeat types, the repeats could be grouped by other characteristics, such as the tendency to form secondary structures, or sharing the same base at specific positions of the motif.

6 Conclusion

In this thesis, we analyzed short tandem repeat patterns in the human reference genome using an update of the *STRAH* R package. We used its pattern-matching functionality to screen recombination hotspots and the regions surrounding them for all distinct combinations of mono-, di-, tri-, tetranucleotide repeats, and certain DNA motifs. We analyzed a total of 310 short tandem repeats and generated pattern-specific comparisons among recombination hotspots, the regions directly surrounding them, and the remaining genomic region of the human genome.

In total, we analyzed 37527 recombination hotspots across the human reference genome. The screened repeat types were then presented in relative densities and tested for differences using a Kruskal-Wallis test and a post-hoc analysis using Dunn's test. We detected a higher density of A/T-rich repeats compared to C/G-rich repeats, and assume that repeat motifs exist in very low numbers if they do not consist of consecutively repeated DNA bases. Furthermore, we observed that C/G-rich repeats tend to be enriched in regions considered to have high recombination, and that A/T-rich repeats are possibly enriched in other biologically important regions of the human genome.

Lastly, the analysis of all distinct short tandem repeats with a repeat unit smaller than or equal to four provides a useful comparison among repeat types for researchers interested in the presence of certain STRs in recombination hotspots throughout the human genome. Furthermore, directions were given on how the versatile functionalities of STRAH can be used in future research to search for other regions in the genome that might show an enrichment in short tandem repeats.

References

Bolton, K. A., Ross, J. P., Grice, D. M., Bowden, N. A., Holliday, E. G., Avery-Kiejda, K. A., & Scott, R. J. (2013). Starrrt: a table of short tandem repeats in regulatory regions of the human genome. *BMC genomics*, *14*(1), 795.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., & Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics*, *62*(6), 1408–1415.

Consortium, G. R., et al. (2019). Genome reference consortium human build 38 patch release 13 (grch38. p13). *NCBI https://www. ncbi. nlm. nih. gov/assembly/GCF_000001405*, *39*.

Consortium, I. H. G. S., et al. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931.

Contente, A., Dittmer, A., Koch, M. C., Roth, J., & Dobbelstein, M. (2002). A polymorphic microsatellite that mediates induction of pig3 by p53. *Nature genetics*, *30*(3), 315–320.

Deaton, A. M., & Bird, A. (2011). Cpg islands and the regulation of transcription. *Genes & development*, *25*(10), 1010–1022.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*(3), 241–252.

Fligner, M. A., & Killeen, T. J. (1976). Distribution-free two-sample tests for scale. *Journal of the American Statistical Association*, *71*(353), 210–213.

Fryxell, K. J., & Moon, W.-J. (2004, 11). CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content. *Molecular Biology and Evolution*, *22*(3), 650-658. Retrieved from `https://doi.org/10.1093/molbev/msi043` doi: 10.1093/molbev/msi043

Gemayel, R., Vinces, M. D., Legendre, M., & Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics*, *44*, 445–477.

Guedin, A., Gros, J., Alberti, P., & Mergny, J.-L. (2010). How long is too long? effects of loop size on g-quadruplex stability. *Nucleic acids research*, *38*(21), 7858–7868.

Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., ... others (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature genetics*, *48*(1), 22–29.

Hefferon, T. W., Groman, J. D., Yurk, C. E., & Cutting, G. R. (2004). A variable dinucleotide repeat in the cftr gene contributes to phenotype diversity by forming rna secondary structures that alter splicing. *Proceedings of the National Academy of Sciences*, *101*(10), 3504–3509.

Heissl, A., Betancourt, A. J., Hermann, P., Povysil, G., Arbeithuber, B., Futschik, A., ... Tiemann-Boege, I. (2019). The impact of poly-a microsatellite heterologies in meiotic recombination. *Life science alliance*, *2*(2).

Hermann, P., Heinzl, M., & Masinovic, M. (2020). Strah [Computer software manual]. Retrieved from `https://github.com/PhHermann/STRAH` (R package version 1.0.1)

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., ... Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. Retrieved from `http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html`

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y., Roskin, K. M., Chen, C.-F., ... Jacob, H. J. (2004). Comparative recombination rates in the rat, mouse, and human genomes. *Genome research*, *14*(4), 528–538.

Kouzine, F., & Levens, D. (2007). Supercoil-driven dna structures regulate genetic transactions. *Front Biosci*, *12*(8-12), 4409.

Kozlowski, P., de Mezer, M., & Krzyzosiak, W. J. (2010). Trinucleotide repeats in human genome and exome. *Nucleic acids research*, *38*(12), 4027–4039.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583–621.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... others (2001). Initial sequencing and analysis of the human genome.

Lawson, M. J., & Zhang, L. (2008). Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5'-utr region. *Gene*, *407*(1-2), 54–62.

Majewski, J., & Ott, J. (2000). Gt repeats are associated with recombination on human chromosome 22. *Genome Research*, *10*(8), 1108–1114.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... others (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.

Martin, P., Makepeace, K., Hill, S. A., Hood, D. W., & Moxon, E. R. (2005). Microsatellite instability regulates transcription factor binding and gene expression. *Proceedings of the National Academy of Sciences*, *102*(10), 3800–3804.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68–78.

Mirkin, S. M. (2007). Expandable dna repeats and human disease. *Nature*, *447*(7147), 932–940.

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, *310*(5746), 321–324.

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2019). Biostrings: Efficient manipulation of biological strings [Computer software manual]. (R package version 2.52.0)

Paigen, K., & Petkov, P. (2010). Mammalian recombination hot spots: properties, control and evolution. *Nature Reviews Genetics*, *11*(3), 221–233.

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., & Camerini-Otero, R. D. (2014). Recombination initiation maps of individual human genomes. *Science*, *346*(6211).

Qin, Y., & Hurley, L. H. (2008). Structures, folding patterns, and functions of intramolecular dna g-quadruplexes found in eukaryotic promoter regions. *Biochimie*, *90*(8), 1149–1171.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Sathasivam, K., Neueder, A., Gipson, T. A., Landles, C., Benjamin, A. C., Bondulich, M. K., ... others (2013). Aberrant splicing of htt generates the pathogenic exon 1 protein in huntington disease. *Proceedings of the National Academy of Sciences*, *110*(6), 2366–2370.

Sawaya, S., Bagshaw, A., Buschiazzo, E., Kumar, P., Chowdhury, S., Black, M. A., & Gemmell, N. (2013). Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PloS one*, *8*(2), e54710.

Schultes, N. P., & Szostak, J. W. (1991). A poly (da. dt) tract is a component of the recombination initiation site at the arg4 locus in saccharomyces cerevisiae. *Molecular and cellular biology*, *11*(1), 322–328.

Subramanian, S., Mishra, R. K., & Singh, L. (2003). Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome biology*, *4*(2), R13.

Team, T. B. D. (2015). Bsgenome.hsapiens.ucsc.hg38: Full genome sequences for homo sapiens (ucsc version hg38) [Computer software manual]. (R package version 1.4.1)

Thomsen, H., Reinsch, N., Xu, N., Bennewitz, J., Looft, C., Grupe, S., ... others (2001). A whole genome scan for differences in recombination rates among three bos taurus breeds. *Mammalian genome*, *12*(9), 724–728.

Treco, D., & Arnheim, N. (1986). The evolutionarily conserved repetitive sequence d (tg. ac) n promotes reciprocal exchange and generates unusual recombinant tetrads during yeast meiosis. *Molecular and Cellular Biology*, *6*(11), 3934–3947.

Verstrepen, K. J., Jansen, A., Lewitter, F., & Fink, G. R. (2005). Intragenic tandem repeats generate functional variability. *Nature genetics*, *37*(9), 986–990.

Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., & Verstrepen, K. J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. *Science*, *324*(5931), 1213–1216.

Weber, J. L., & Wong, C. (1993). Mutation of human short tandem repeats. *Human molecular genetics*, *2*(8), 1123–1128.

Wells, R. D., Dere, R., Hebert, M. L., Napierala, M., & Son, L. S. (2005). Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic acids research*, *33*(12), 3785–3798.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29. Retrieved from `http://www.jstatsoft.org/v40/i01/`

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from `https://ggplot2.tidyverse.org`

Wickham, H., François, R., Henry, L., & Müller, K. (2020). dplyr: A grammar of data manipulation [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=dplyr` (R package version 1.0.0)

Wilke, C. O. (2019). cowplot: Streamlined plot theme and plot annotations for 'ggplot2' [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=cowplot` (R package version 1.0.0)

Willems, T., Gymrek, M., Highnam, G., Mittelman, D., Erlich, Y., Consortium, . G. P., et al. (2014). The landscape of human str variation. *Genome research*, *24*(11), 1894–1904.

Zhu, H. (2019). kableextra: Construct complex table with 'kable' and pipe syntax [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=kableExtra` (R package version 1.1.0)

Supplementary Material

6.1 Mononucleotide repeats

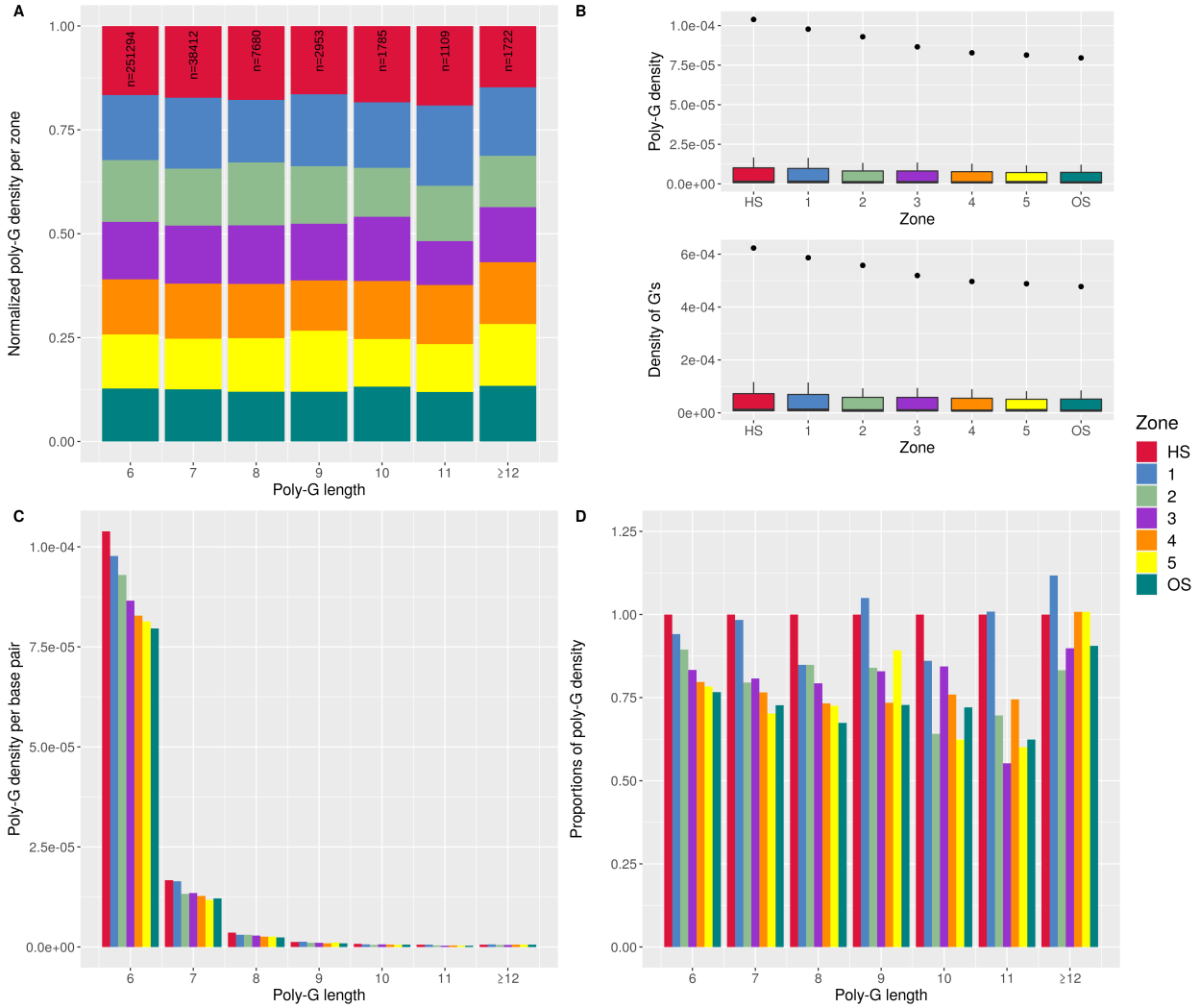


Figure 17: Densities of poly-G repeats (in version GRCh38/hg38 of the human reference genome). The poly-G density per zone of a certain repeat is obtained by dividing the frequency of the repeat by the length of the zone it is found in. Repeat types longer than 12 base pairs are grouped with repeats equal to 12 due to their low frequency. Figure 17A presents the normalized poly-G density, where the poly-G density per repeat per zone is divided by the total sum of poly-G densities across all seven zones (Hotspot Zone, Zones 1-5, Outside Zone) of the poly-G repeat. The total count of a certain repeat found across all seven zones is shown at the top of each bar of the plot. The top panel of Figure 17B describes the distribution of poly-G densities of all repeat lengths found in a given zone. The bottom panel of Figure 17B shows the density of G's of all repeat types found in a given zone, where the density of G's is obtained by multiplying the length of a repeat by its frequency and dividing it by the length of the zone it is in. Figure 17C presents the poly-G densities per base pair of all zones, stratified with respect to the repeat length. Figure 17D presents the poly-G densities of each zone (Hotspot Zone, Zones 1-5, Outside Zone) relative to the hotspot zone.

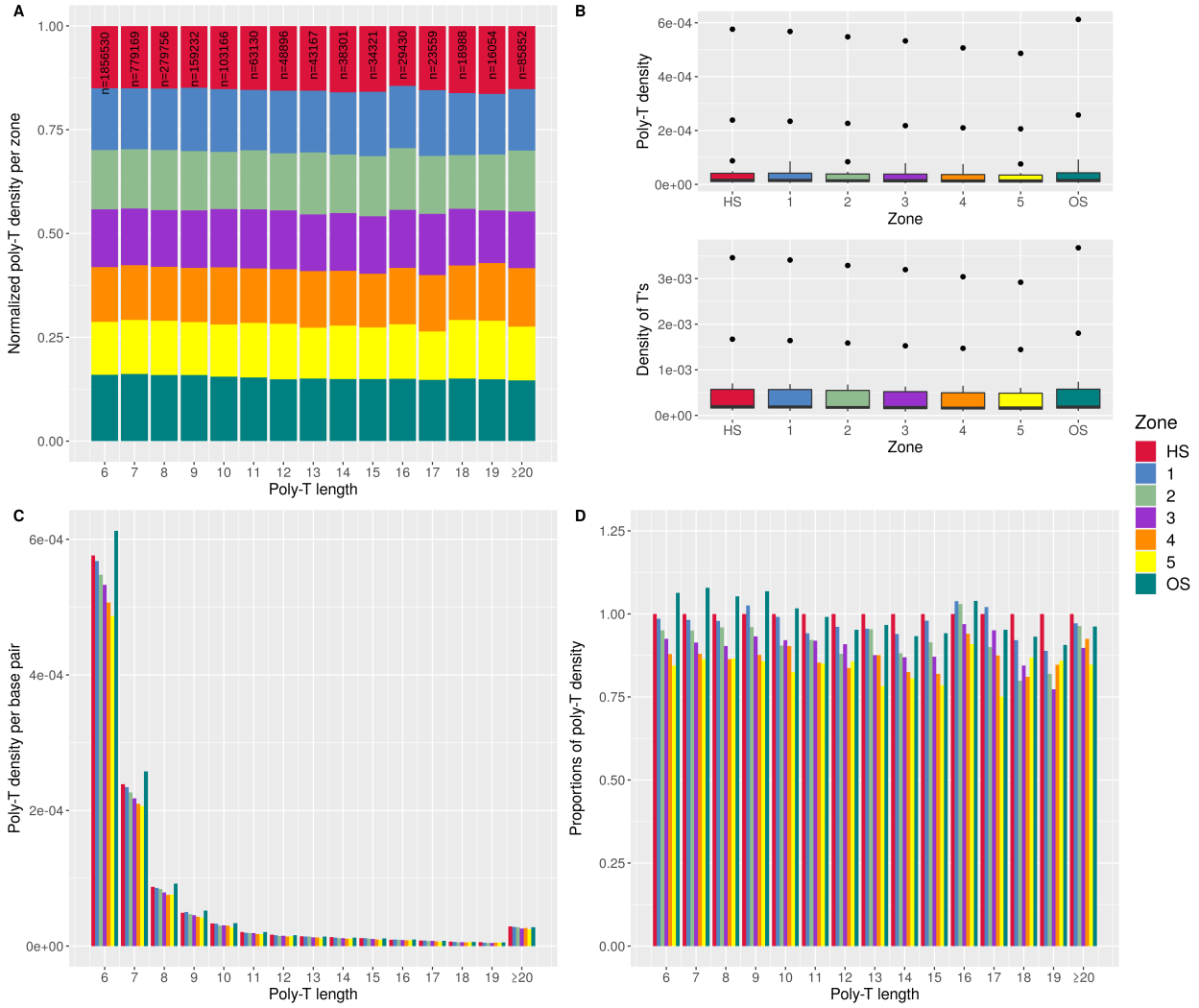


Figure 18: Densities of poly-T repeats (in version GRCh38/hg38 of the human reference genome). The poly-T density per zone of a certain repeat is obtained by dividing the frequency of the repeat by the length of the zone it is found in. Repeat types longer than 20 base pairs are grouped with repeats equal to 20 due to their low frequency. Figure 18A presents the normalized poly-T density, where the poly-T density per repeat per zone is divided by the total sum of poly-T densities across all seven zones (Hotspot Zone, Zones 1-5, Outside Zone) of the poly-T repeat. The total count of a certain repeat found across all seven zones is shown at the top of each bar of the plot. The top panel of Figure 18B describes the distribution of poly-T densities of all repeat lengths found in a given zone. The bottom panel of Figure 18B shows the density of T's of all repeat types found in a given zone, where the density of T's is obtained by multiplying the length of a repeat by its frequency and dividing it by the length of the zone it is in. Figure 18C presents the poly-T densities per base pair of all zones, stratified with respect to the repeat length. Figure 18D presents the poly-T densities of each zone (Hotspot Zone, Zones 1-5, Outside Zone) relative to the hotspot zone.

6.2 Dinucleotide repeats

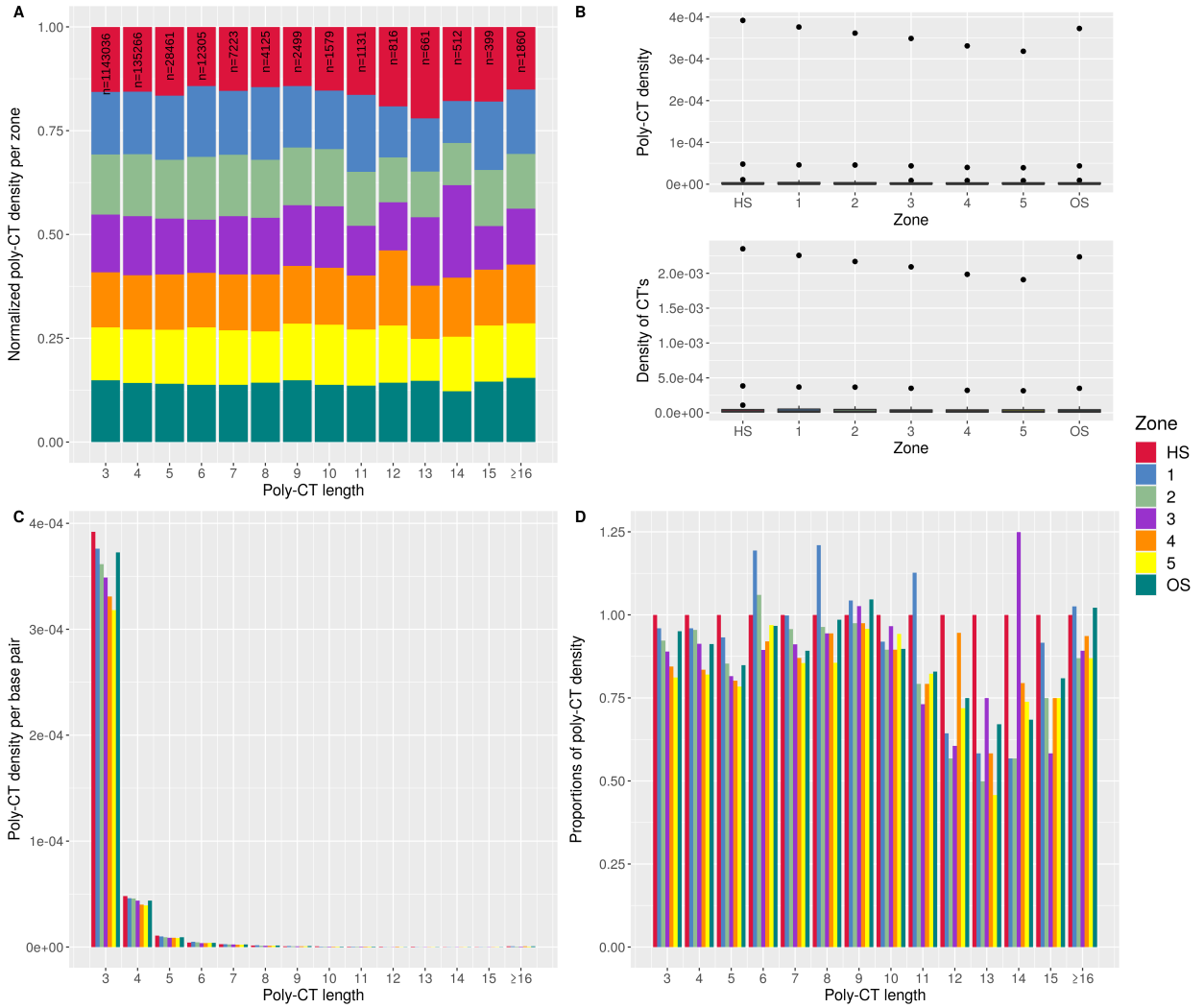


Figure 19: Densities of poly-CT repeats (in version GRCh38/hg38 of the human reference genome). The poly-CT density per zone of a certain repeat is obtained by dividing the frequency of the repeat by the length of the zone it is found in. Repeat types longer than 16 base pairs are grouped with repeats equal to 16 due to their low frequency. Figure 19A presents the normalized poly-CT density, where the poly-CT density per repeat per zone is divided by the total sum of poly-CT densities across all seven zones (Hotspot Zone, Zones 1-5, Outside Zone) of the poly-CT repeat. The total count of a certain repeat found across all seven zones is shown at the top of each bar of the plot. The top panel of Figure 19B describes the distribution of poly-CT densities of all repeat lengths found in a given zone. The bottom panel of Figure 19B shows the density of CT's of all repeat types found in a given zone, where the density of CT's is obtained by multiplying the length of a repeat by its frequency and dividing it by the length of the zone it is in. Figure 19C presents the poly-CT densities per base pair of all zones, stratified with respect to the repeat length. Figure 19D presents the poly-CT densities of each zone (Hotspot Zone, Zones 1-5, Outside Zone) relative to the hotspot zone.

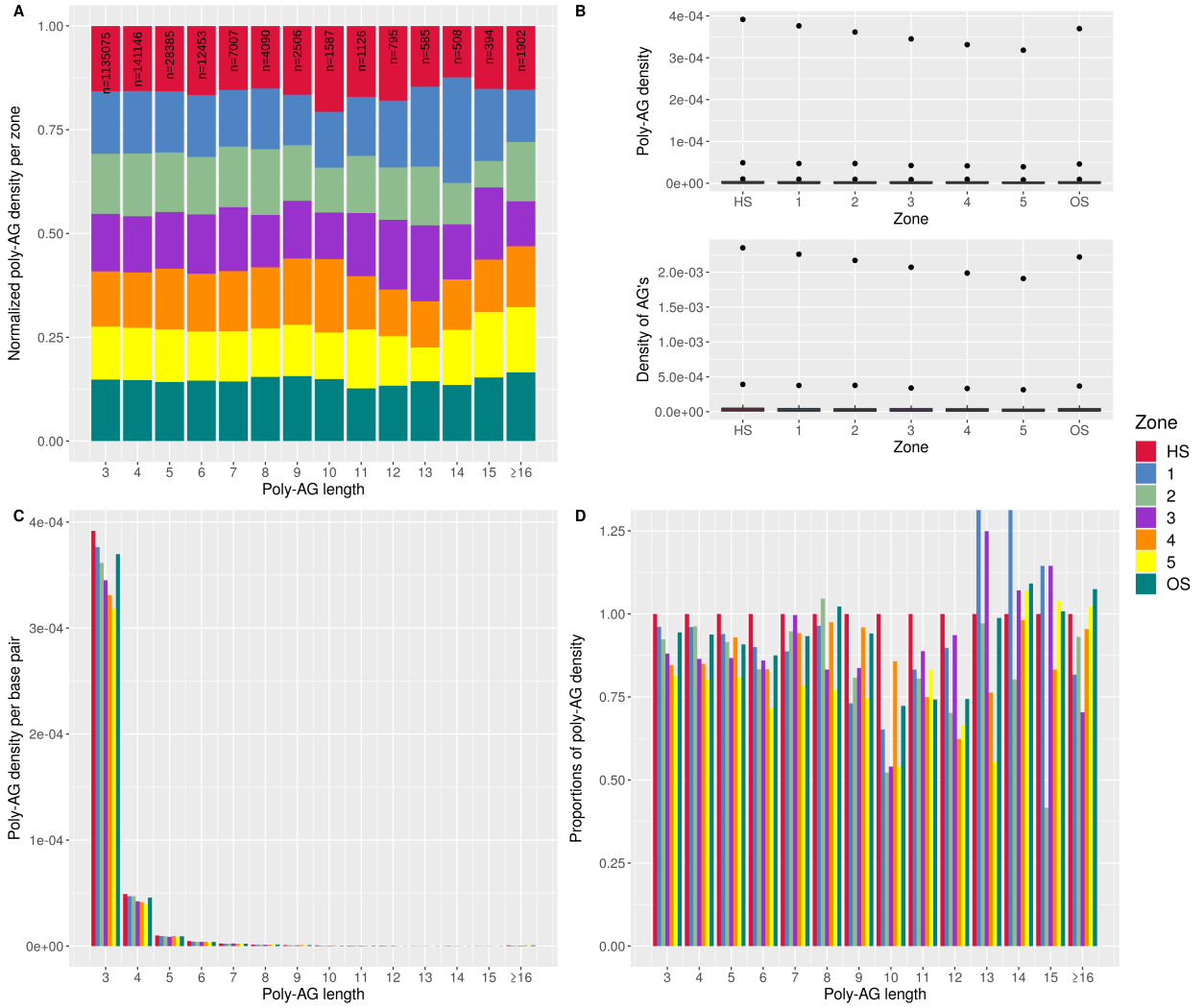


Figure 20: Densities of poly-AG repeats (in version GRCh38/hg38 of the human reference genome). The poly-AG density per zone of a certain repeat is obtained by dividing the frequency of the repeat by the length of the zone it is found in. Repeat types longer than 16 base pairs are grouped with repeats equal to 16 due to their low frequency. Figure 20A presents the normalized poly-AG density, where the poly-AG density per repeat per zone is divided by the total sum of poly-T densities across all seven zones (Hotspot Zone, Zones 1-5, Outside Zone) of the poly-AG repeat. The total count of a certain repeat found across all seven zones is shown at the top of each bar of the plot. The top panel of Figure 20B describes the distribution of poly-AG densities of all repeat lengths found in a given zone. The bottom panel of Figure 20B shows the density of AG's of all repeat types found in a given zone, where the density of AG's is obtained by multiplying the length of a repeat by its frequency and dividing it by the length of the zone it is in. Figure 20C presents the poly-AG densities per base pair of all zones, stratified with respect to the repeat length. Figure 20D presents the poly-AG densities of each zone (Hotspot Zone, Zones 1-5, Outside Zone) relative to the hotspot zone.

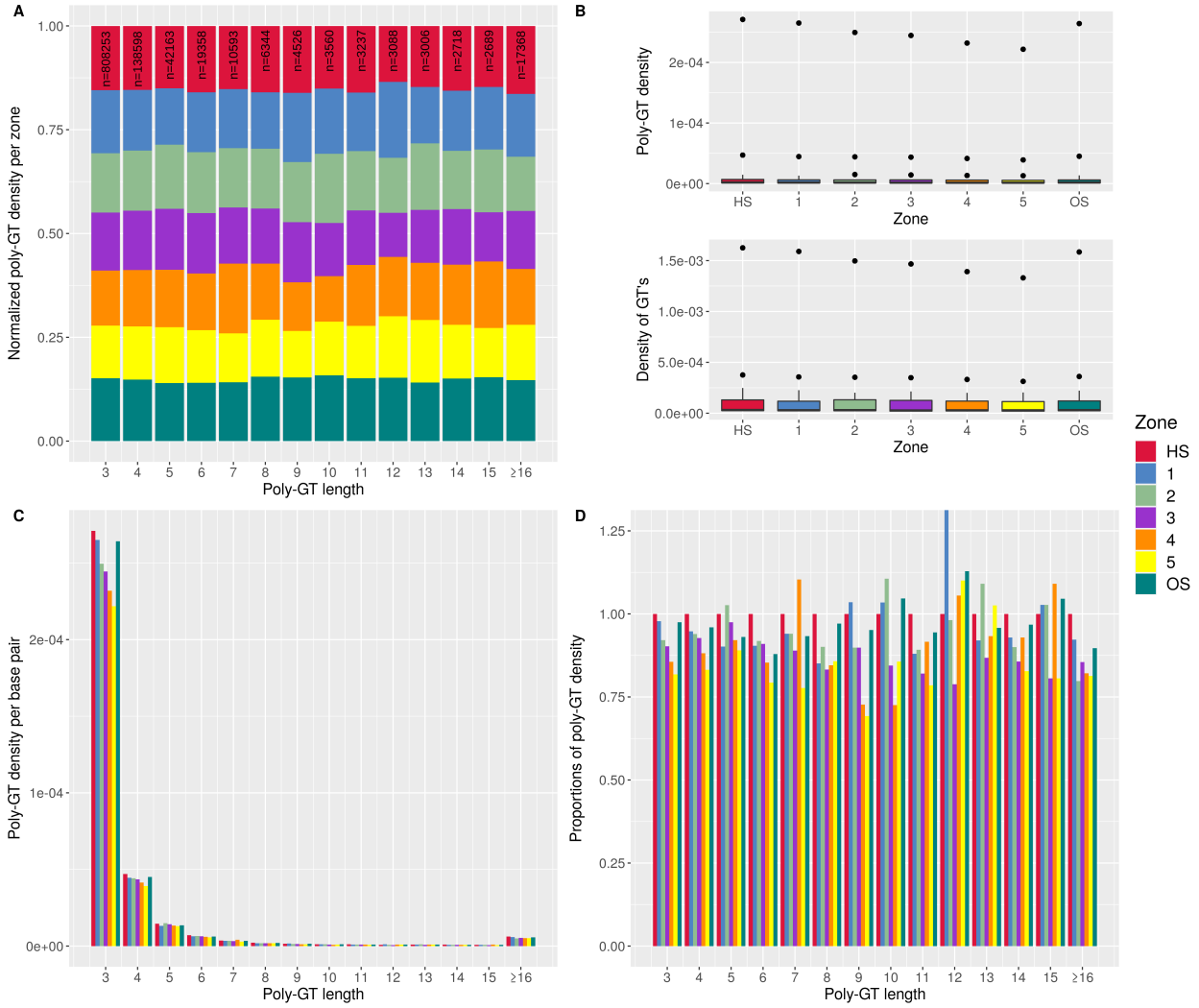


Figure 21: Densities of poly-GT repeats (in version GRCh38/hg38 of the human reference genome). The poly-GT density per zone of a certain repeat is obtained by dividing the frequency of the repeat by the length of the zone it is found in. Repeat types longer than 16 base pairs are grouped with repeats equal to 16 due to their low frequency. Figure 21A presents the normalized poly-GT density, where the poly-GT density per repeat per zone is divided by the total sum of poly-GT densities across all seven zones (Hotspot Zone, Zones 1-5, Outside Zone) of the poly-GT repeat. The total count of a certain repeat found across all seven zones is shown at the top of each bar of the plot. The top panel of Figure 21B describes the distribution of poly-GT densities of all repeat lengths found in a given zone. The bottom panel of Figure 21B shows the density of GT's of all repeat types found in a given zone, where the density of GT's is obtained by multiplying the length of a repeat by its frequency and dividing it by the length of the zone it is in. Figure 21C presents the poly-GT densities per base pair of all zones, stratified with respect to the repeat length. Figure 21D presents the poly-GT densities of each zone (Hotspot Zone, Zones 1-5, Outside Zone) relative to the hotspot zone.

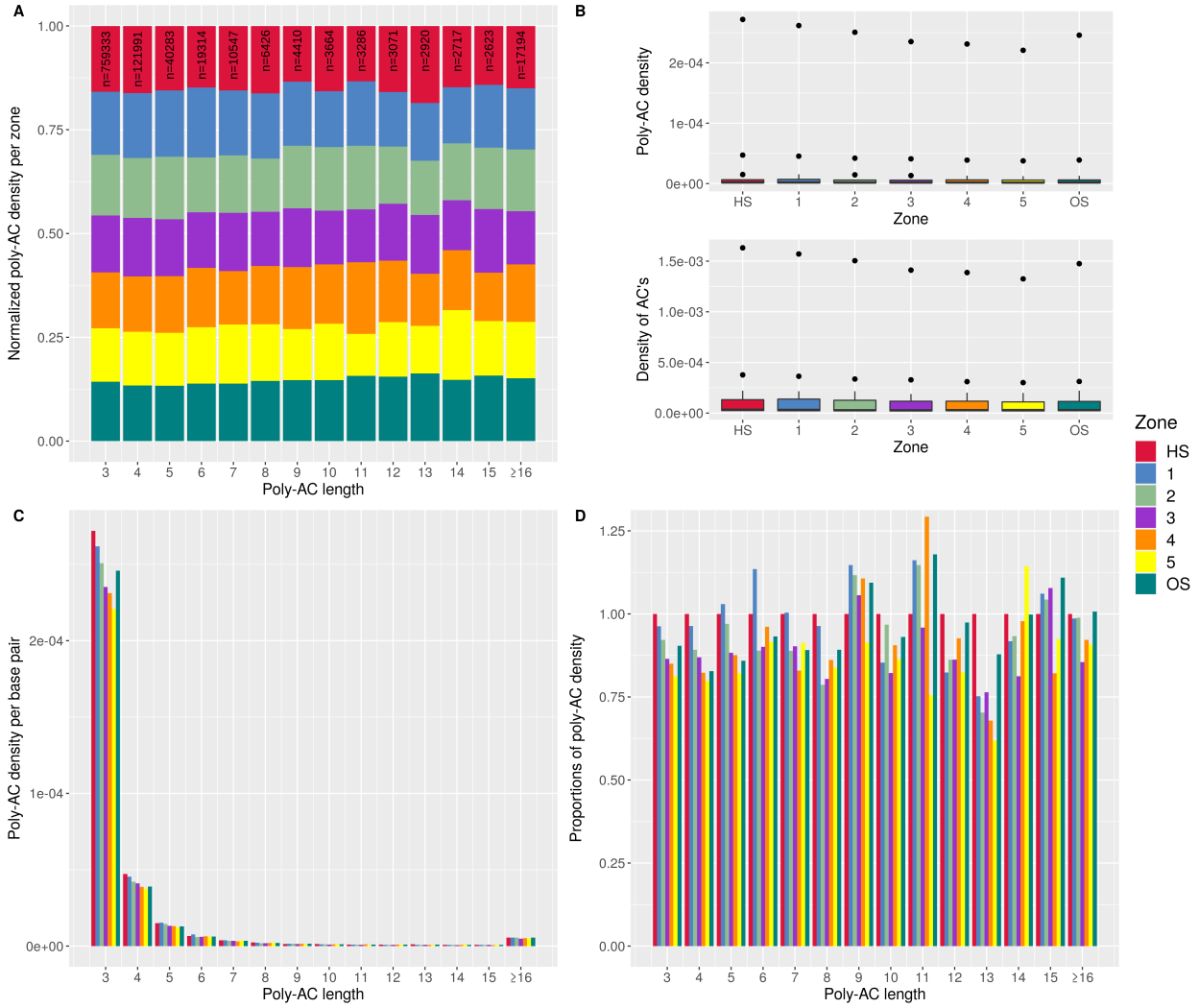


Figure 22: Densities of poly-AC repeats (in version GRCh38/hg38 of the human reference genome). The poly-AC density per zone of a certain repeat is obtained by dividing the frequency of the repeat by the length of the zone it is found in. Repeat types longer than 16 base pairs are grouped with repeats equal to 16 due to their low frequency. Figure 22A presents the normalized poly-AC density, where the poly-AC density per repeat per zone is divided by the total sum of poly-AC densities across all seven zones (Hotspot Zone, Zones 1-5, Outside Zone) of the poly-AC repeat. The total count of a certain repeat found across all seven zones is shown at the top of each bar of the plot. The top panel of Figure 22B describes the distribution of poly-AC densities of all repeat lengths found in a given zone. The bottom panel of Figure 22B shows the density of AC's of all repeat types found in a given zone, where the density of AC's is obtained by multiplying the length of a repeat by its frequency and dividing it by the length of the zone it is in. Figure 22C presents the poly-AC densities per base pair of all zones, stratified with respect to the repeat length. Figure 22D presents the poly-AC densities of each zone (Hotspot Zone, Zones 1-5, Outside Zone) relative to the hotspot zone.

6.3 Tetranucleotide repeats

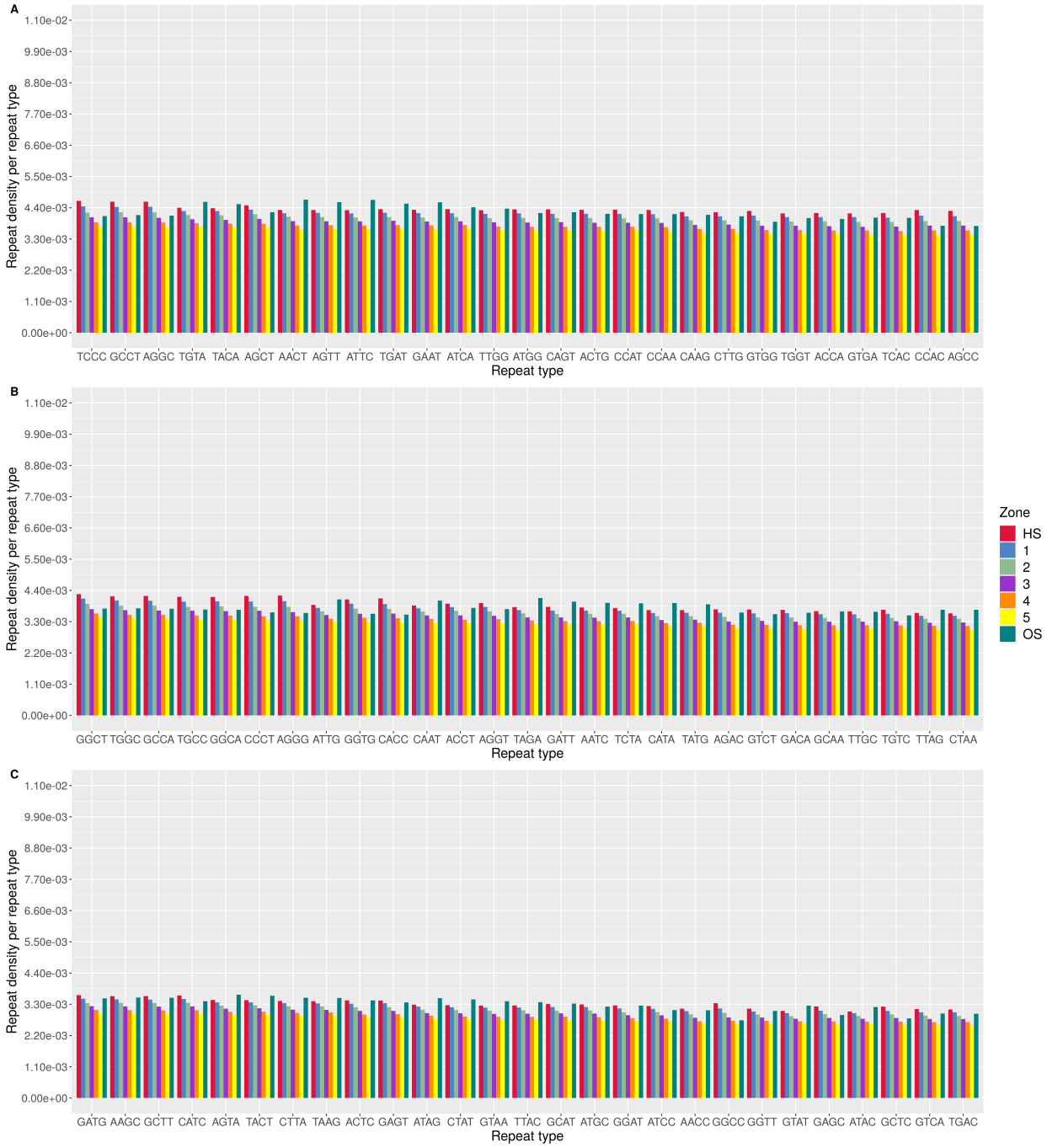


Figure 23: Repeat density per tetranucleotide repeat type. The repeat densities of 81 distinct tetranucleotide repeat types of the human reference genome are shown. The repeat density per zone of a given repeat is obtained by dividing the frequency of the repeat by the length of the zone.

6.4 Correction

The screening of repeats in artificially-generated DNA sequences indicated that the original version of *STRAH* (Version: 1.0.1) overestimated the frequency of repeats in hotspots due to a coding error, which resulted into all repeats found to the right side of a hotspot being assigned to the outside zone, as illustrated in Figure 24. *STRAH* correctly assigned all repeats that were found to the left of the hotspot, but failed to do so for any repeat to the right of the hotspot zone, such that all repeats on the right hand side were assigned to the outside zone.

Any data application, therefore, suggests a two-fold increase of repeats in hotspots, but is actually observed due to the fact that only half of the repeats are assigned to the surrounding zones. Heissl et al. (2019) used *STRAH* to analyze poly-A repeats using genome version GRCh37/hg19, and reported a significant approximate twofold increase in poly-A repeat densities in hotspots compared to flanking regions. Here, we list a pairwise comparison of all panels of Figure 4 in the work of Heissl et al. (2019) to present the newly calculated densities obtained from the corrected version of *STRAH* (Version: 1.1.0). The results are compared in Figure 25-29.

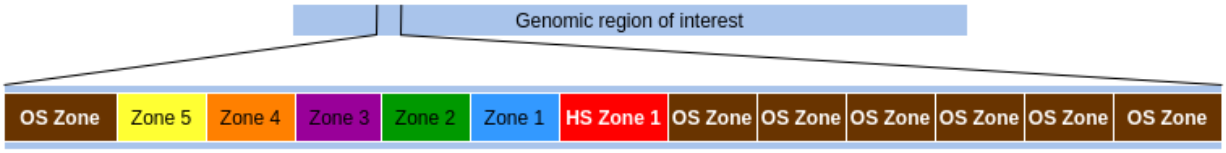


Figure 24: Explanation of the incorrect output of *STRAH*. The result can be explained by an incorrect zone-assignment, where *STRAH* fails to correctly assign all repeats to the right side of a hotspot zone and assigns the repeats to the outside zone.

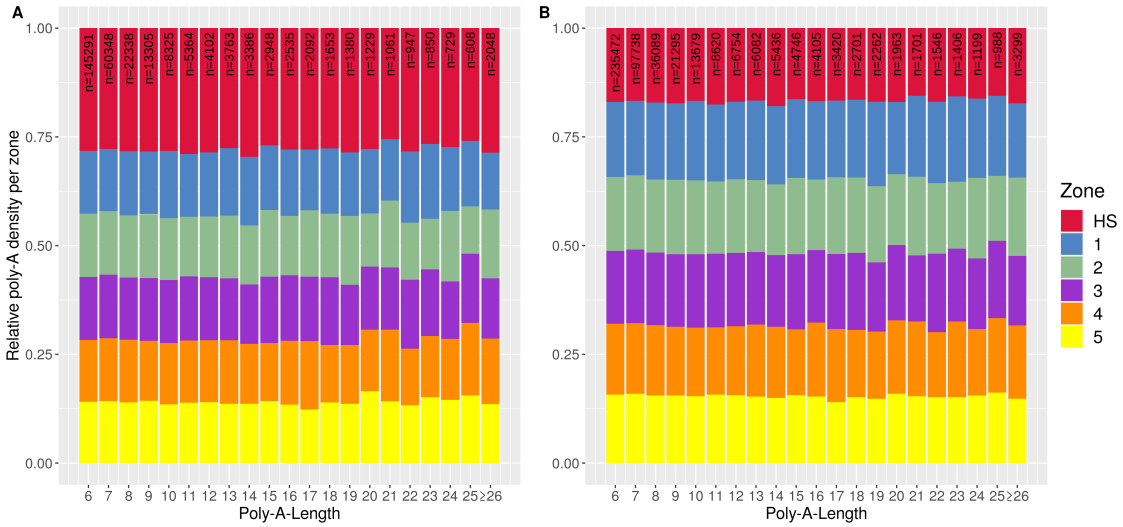


Figure 25: Comparison of poly-A density of each poly-A tract length (Figure 4A) using genome-version GRCh37/hg19. Panel A presents the results obtained by (Heissl et al., 2019), panel B shows the densities when the corrected version of *STRAH* is used.

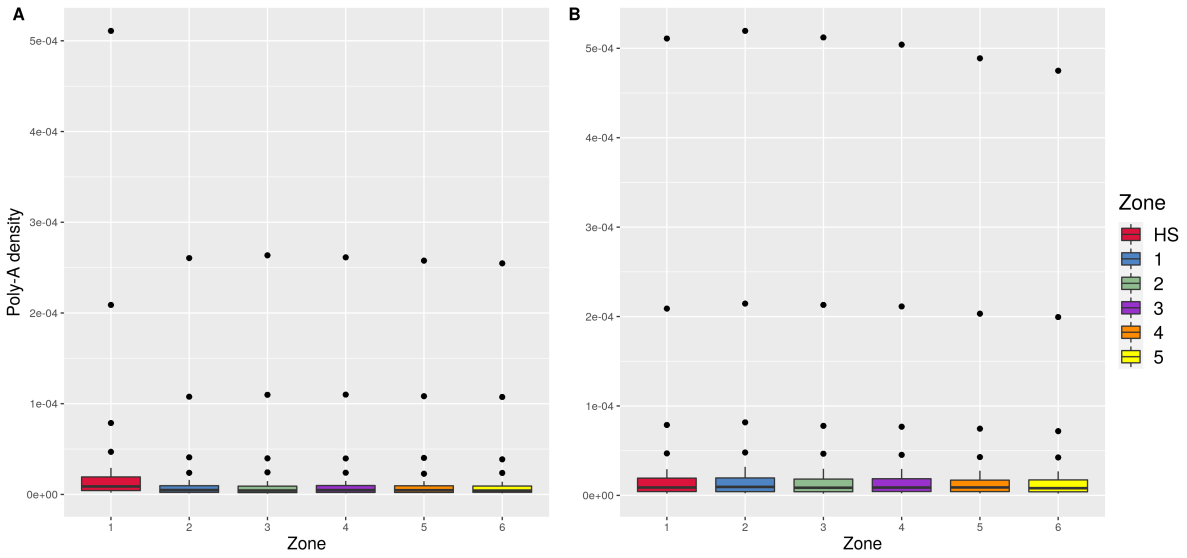


Figure 26: Comparison of poly-A density per zone (Figure 4B) using genome-version GRCh37/hg19. Panel A presents the results obtained by (Heissl et al., 2019), panel B shows the densities when the corrected version of *STRAH* is used.

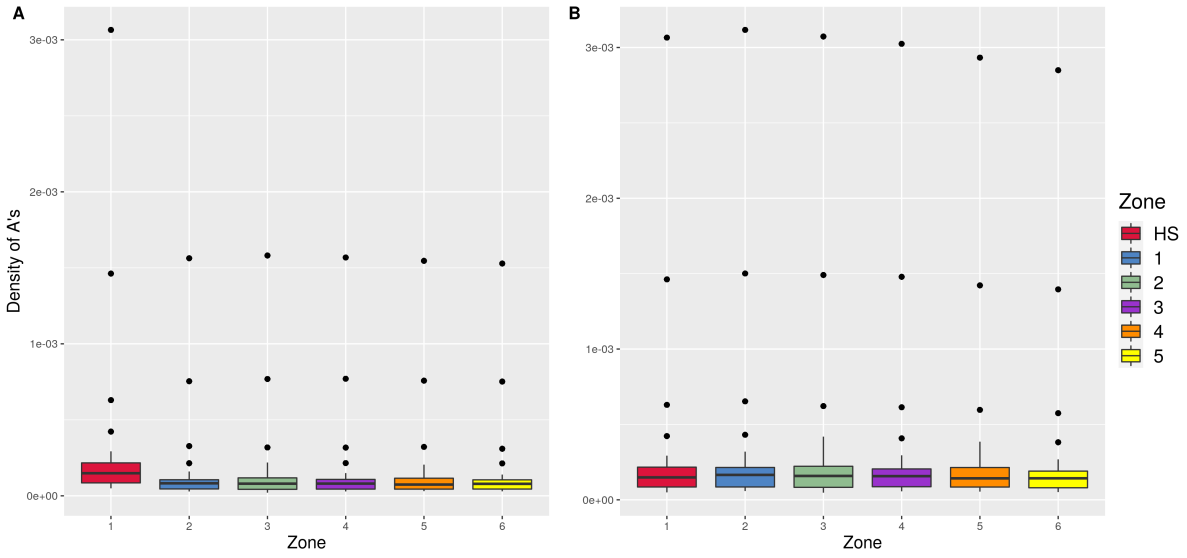


Figure 27: Comparison of the density of A's per zone (Figure 4B) using version GRCh37/hg19. Panel A presents the results obtained by (Heissl et al., 2019), panel B shows the densities when the corrected version of *STRAH* is used.

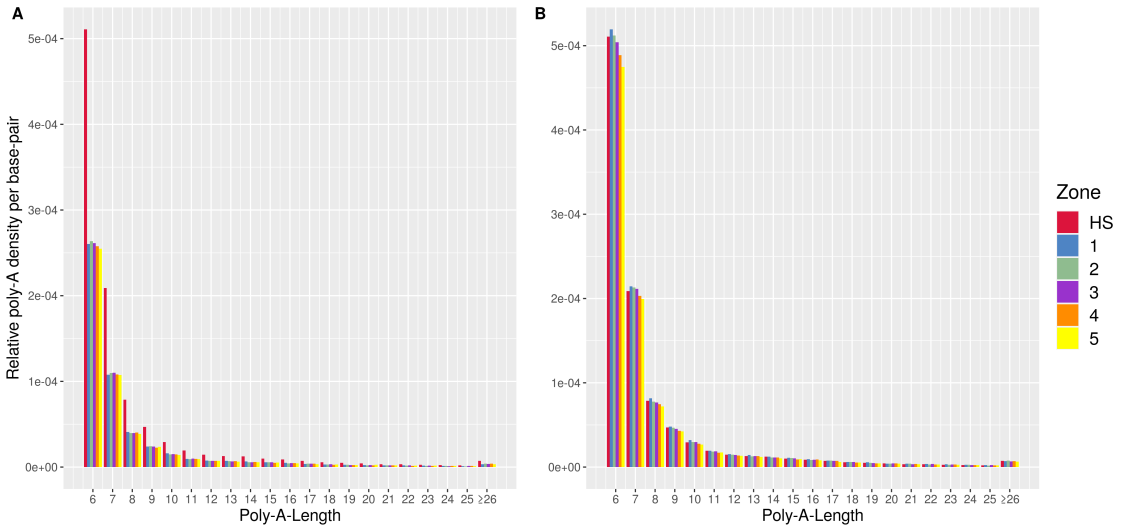


Figure 28: Comparison of poly-A density per base pair per zone (Figure 4C) using version GRCh37/hg19. Panel A presents the results obtained by (Heissl et al., 2019), panel B shows the densities when the corrected version of *STRAH* is used.

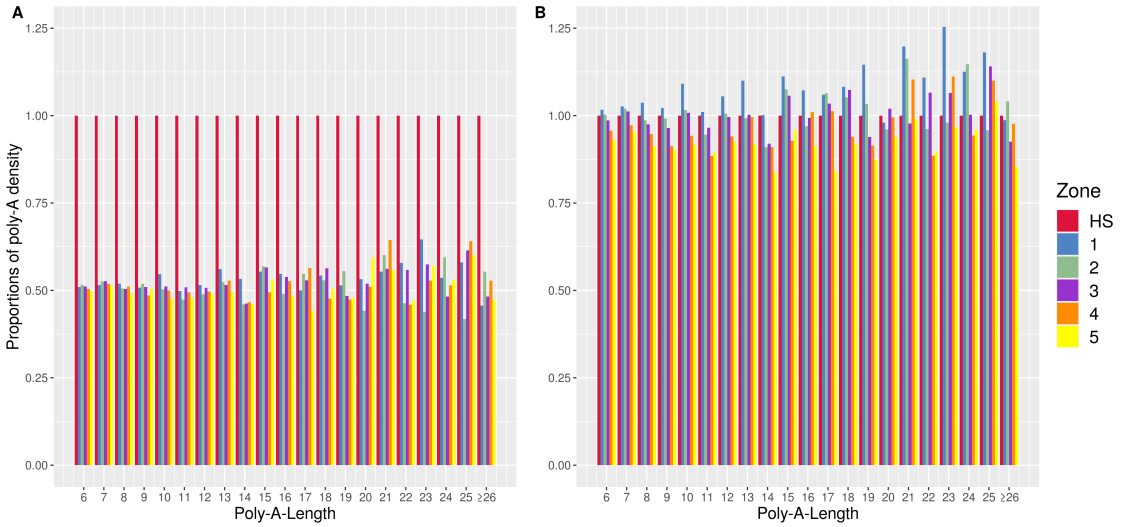


Figure 29: Comparison of poly-A densities per poly-A tract length (Figure 4D) using version GRCh37/hg19. Panel A presents the results obtained by (Heissl et al., 2019), panel B shows the densities when the corrected version of *STRAH* is used.