UNIVERSIDAD AUTONOMA DE MADRID

FACULTAD DE MEDICINA

TRABAJO FIN DE MÁSTER

REPETICIONES Y FUSIONES TELOMERICAS EN MAMIFEROS

Máster Universitario en Bioinformática y Biología Computacional

Autor: Rodriguez del Valle, Jonatan

Directores: Gómez Rodríguez, Manuel José
Unidad de Bioinformática, CNIC, ISCIII
Flores Hernández, Ignacio
Centro de Biología Molecular Severo Ochoa, CSIC-UAM

Tutor: Fátima Sánchez Cabo
Departamento de Bioquímica Universidad Autónoma de Madrid

CURSO: 2023-24
FECHA: Mayo, 2024

Index

REPETICIONES Y FUSIONES TELOMERICAS EN MAMIFEROS

Abstract

Telomeres, consisting of tandem repeats of the motif TTAGGG, safeguard DNA integrity by protecting chromosomal ends. Loss of telomeric sequences during cellular replication leads to senescence or apoptotic death in somatic cells. To counteract this, some stem and cancer cells employ telomerase reverse transcriptase (TERT) for elongation, while others activate the alternative lengthening of telomeres (ALT) pathway. Studies in mammals have linked ALT to telomere fragility, resulting in telomere fusions (TFs) and variability (TVRs). ALT is often associated with PML-bodies and break-induced replication (BIR) mechanisms, which has been observed to be activated in certain somatic cells. In this study, we introduce TeloFusVarfinder, a novel bioinformatic tool designed to analyse telomeric fusions, variability, and telomere length from raw next-generation sequencing (NGS) data. Leveraging data from the Zoonomia project encompassing 200 mammalian species, we compare telomeric features with the presence or absence of known proteins obtained from TOGA data. Initial investigations indicate enhanced variability and and increased telomere content in Chiroptera and suggest negative correlation between the number of telomeric fusions and the presence of orthologs of certain genes involved in telomere mantainance.

Key words: Telomere, telomere fusions, telomere variability, telomerase, PML-bodies, ALT, Zoonomia, telomere length, chromosome number,

Introduction

DNA AND DNA REPLICATION

DNA, or deoxyribonucleic acid, is a molecule that carries the genetic instructions used in the growth, development, functioning, and reproduction of all known living organisms and many viruses. DNA consists of long polymers of nucleotides of four types, represented by the letters A, C, G and T.

Nucleotide sequences in DNA determine the specific traits of an organism, such as its physical characteristics, susceptibility to certain diseases, metabolism, and so on. DNA is organized into structures called genes, which are segments of DNA that encode specific functional products such as non-coding RNA and messenger RNA, the latter of which can be translated by ribosomes into proteins.

Eukaryotic cells may divide, in order to reproduce themselves, once the DNA is replicated during the S (synthesis) phase of the cell cycle, which consists of mitosis followed by cytokinesis. During mitosis, the replicated chromosomes condense and align to the center of the cell. Then, chromosomes are pulled apart by spindle fibres and distributed into two daughter cells, ensuring that each new cell receives the same amount of DNA.

TELOMERES AND TELOMERE SHORTENING

Telomeres are structures located at the ends of chromosomes. They consist of tandem repeats of DNA sequences, called telomeric motifs, and associated proteins. Telomere sequences vary in length and composition among different organisms: while in mammals they consist of TTAGGG tandem repeats, in plants the repetition is TTTAGGG, and in yeast they are made of an irregular repeated sequence $T(G)_{1-3}/(C)_{1-2}A$ *(Srinivas et al., 2020)*. Telomeres play a crucial role in maintaining the stability and integrity of chromosomes during cell division, by protecting the ends of chromosomes from degradation and fusion with neighbouring chromosomes.

During each cycle of DNA replication, a small portion of the telomere is lost due to the inability of polymerase II to fully replicate the ends of the chromosomes, in a phenomenon known as the "end-replication problem". Telomere shortening has been associated to cell aging and may explain why cells can divide only a limited number of times. Eventually, when telomeres become critically short, cells may enter a state of senescence or undergo apoptosis, preventing the propagation of damaged or abnormal cells. *(Bize et al., 2009)*

Research indicates that individuals within a species population who exhibit shorter telomeres or experience faster telomere shortening typically face decreased survival rate. This accelerated telomere attrition is frequently increased by oxidative damage, aligning with cellular deterioration and aging. *(Bize et al., 2009)* In addition, telomere ends in humans most commonly consist in the sequence GGTTA, GGGTT or GTTAG in the 5' to 3' end, and CAATC and AATCC for the 3' to 5' end. *(Sfeir et al., 2005)(figure 1)*

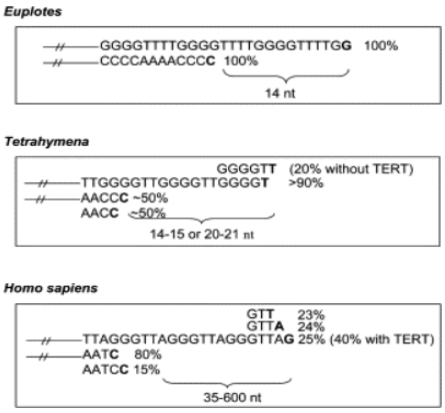


Figure 1 Comparison of the Telomeric-End Structures (Sfeir et al., 2005)

TELOMERE LENGTHENING

Telomere shortening can be reversed in some cells with the participation of telomerase (TERT), an enzyme that adds repetitive DNA sequences to the ends of chromosomes by means of its RNA-dependent DNA polymerase activity, using a non-coding RNA molecule, coded by gene TERC, which acts as a template to elongate the telomere *(Srinivas et al., 2020)*. While telomerase is usually highly activated in stem cells and certain types of cancer cells, in most somatic cells (differentiated cells),

telomerase activity is low, and telomeres gradually shorten with each cell division. In mammals, TRF1 and TRF2 (TTAGGG repeat binding factors) inhibit the telomere lengthening (*Collins, 2000) (Srinivas et al., 2020)*. A summary of genes involved in the regulation of telomerase activity or telomere lengthening is presented in figure 2.

PROTEIN	TELOMERE FUNCTION
TERT	Telomerase is a ribonucleoprotein enzyme essential for the replication of chromosome termini in most eukaryotes. Active in progenitor and cancer cells. Inactive, or very low activity, in normal somatic cells.
DAXX	Acts as a targeting component of the chromatin remodeling complex ATRX:DAXX which has ATP-dependent DNA translocase activity and catalyzes the replication-independent deposition of histone H3.3 in pericentric DNA repeats outside S-phase and telomeres, and the in vitro remodeling of H3.3-containing nucleosomes.
ATRX	Catalytic component of the chromatin remodeling complex ATRX:DAXX which has ATP-dependent DNA translocase activity and catalyzes the replication-independent deposition of histone H3.3 in pericentric DNA repeats outside S-phase and telomeres, and the in vitro remodeling of H3.3-containing nucleosomes.
TERF1	Binds the telomeric double-stranded 5'-TTAGGG-3' repeat and negatively regulates telomere length. Involved in the regulation of the mitotic spindle. Component of the shelterin complex (telosome) that is involved in the regulation of telomere length and protection. Shelterin associates with arrays of double-stranded 5'-TTAGGG-3' repeats added by telomerase and protects chromosome ends
TERF2	Binds the telomeric double-stranded 5'-TTAGGG-3' repeat and plays a central role in telomere maintenance and protection against end-to-end fusion of chromosomes. In addition to its telomeric DNA-binding role, required to recruit a number of factors and enzymes
RTEL1	ATP-dependent DNA helicase implicated in telomere-length regulation, DNA repair and the maintenance of genomic stability. Acts as an anti-recombinase to counteract toxic recombination and limit crossover during meiosis.
PIF1	Involved in the maintenance of telomeric DNA. Inhibits telomere elongation, de novo telomere formation and telomere addition to DSBs via catalytic inhibition of telomerase.

Figure 2 A schema of telomere associated proteins and their functions. (Uniprot.)

More than 80% of cancer cell types have the capability to reactivate telomerase to increase the telomere size and avoid the senescence or apoptotic phase. However, there are alternative ways to increase the length of telomeres that have been observed in around 15% of cancer types. These mechanisms are called Alternative Lengthening of Telomeres (ALT), and cells or tumour expressing them are referred to as ALT-positive (ALT+) *(Shen et al., 2021)*

ALT is generally mediated by break-induced replication (BIR), which is a fundamental mechanism that cells use to repair DNA breaks and maintain genomic integrity. It's especially important for dealing with breaks encountered during DNA replication and contributes to telomere maintenance. The mechanism copies the DNA from a template which could be a sister chromatid, an extrachromosomal fragment, or the same chromosome. *(Malkova & Ira, 2013)* (figure 3).

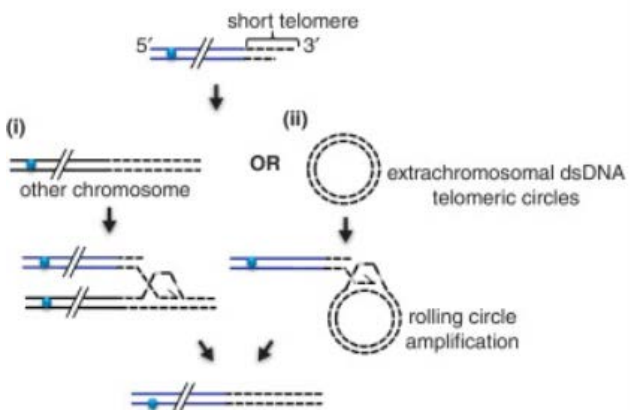


Figure 3: Break-Induced Replication (BIR) Pathway (Malkova & Ira, 2013) (I) represents the mechanism where an external chromosome is used as template to increase the length the telomere (II) Represents the mechanism where an extrachromosomal fragment is used as template.

Cells employing ALT pathways exhibit unique traits, such as extrachromosomal telomere repeats (ECTRs), random DNA damage, replication stress, and an unusual chromatin state specialised to ALT activity. These characteristics, coupled with increased double-strand breaks (DSBs) and BIR make ALT+ telomeres fragile. In recent studies some proteins have been found to be involved in the ALT-BIR mechanism, such as proliferating cell nuclear antigen (PCNA), and other proteins associated to promyelocytic leukaemia nuclear bodies (PML-NBs) *(Shen et al., 2021)*.

PML-NBs are nuclear protein complexes that interact with a large variety of proteins and that have been associated with different functions such as the regulation of cellular chromatin, antiviral activity, and telomere elongation. (Corpet et al., 2020).

Several studies suggest that PML-NBs play a role in modulating telomeric chromatin structure in non-neoplastic cells. In somatic cells, PML-NBs may participate in telomere surveillance. PML-NBs act as a scaffold, like a vacuole with no membrane, to concentrate specific DNA repair and recombination factors, along with chromatin modifiers, around telomeric chromatin. Experiments demonstrate that artificially created ALT-associated PML-NBs (APBs) (figure 4) promote the clustering of telomeres. APBs are closely associated with changes in heterochromatin dynamics at telomeres, alterations in histone chaperones such as ASF1, or modifications in histone enzymes like SETDB1, which impacts heterochromatin organization at telomeres (Corpet et al., 2020). Other proteins that are involved in PML-NBs-telomere interaction include the RPA complex, whose function is not well understood, although it has been suggested that it has an important role in the telomere maintenance in ALT cancer cells *(Loe et al., 2020)*

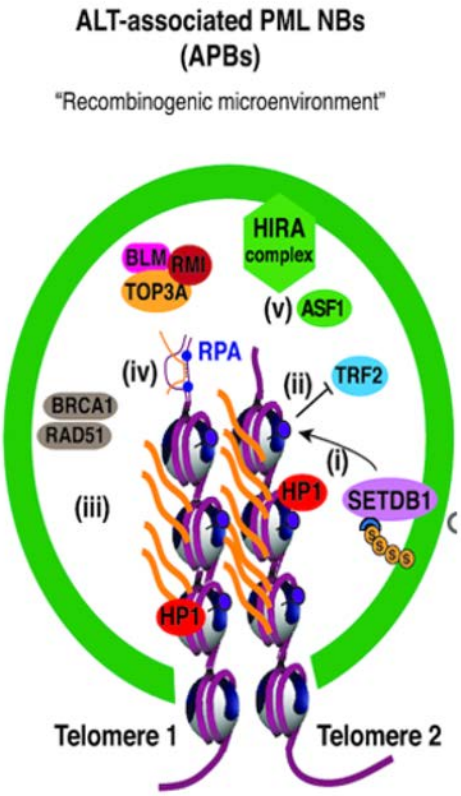


Figure 4: Interpretation of ALT associated PML-NBs (APBs). RPA complex appears to have an important role in using another chromosome as template. (The content of this slide may be subject to copyright, (Loe et al., 2020))

ALT telomeres contain degenerate telomeric repeats known as telomere variant repeats (TVRs), which are spread widely along telomeric DNA because of recombination-mediated templating. Accordingly, ALT+ cancer cells have a higher rate of TVRs than normal cells. On a different side, it has been observed that mutations on the genes encoding the ATRX/DAXX complex and the TERT promoter overall increased TTAGGG variation. *(Lee et al., 2018)*.

TELOMERIC FUSIONS

Telomeric fusions (TFs) may occur as consequence of a cellular response to protect the end of chromosomes that have suffered of extreme shortening of telomeres. Such end-to-end chromosomal fusions result in a specific sequence pattern that has been described as "inward", in which 5'-TTAGGG-3' tandem repeats (or circular permutations) are linked to 5'-CCCTAA-3' motifs (Figure 5).

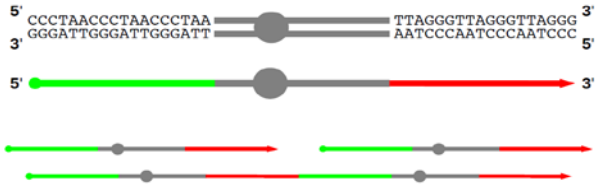


Figure 5: Inward fusion representation, where two telomeres fuse.

Inward telomeric fusions arise in normal cells under certain conditions of replicative stress, DNA damage or maturation, for example in cardiomyocyte maturation. *(Aix et al., 2023),* and can be seen "Chromatin bridges" under microscopy, under certain circumstances *(Hoang & O'Sullivan, 2020). (figure 6)* They could also be related to genome evolution processes that result in the reduction of chromosome numbers.

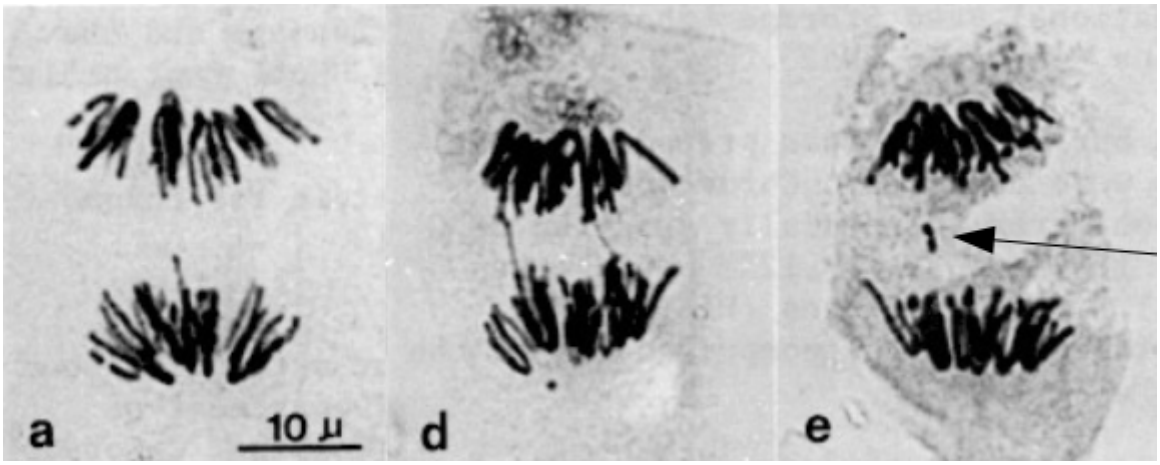


Figure 6: Telomere fusions.

In addition to "inward" telomeric fusions, a pattern referred to as "outward" has also been described, in which 5'-CCCTAA-3' tandem repeats (or circular permutations) are linked to 5'-TTAGGG-3' motifs (figure 7). The "outward" fusion pattern cannot occur as consequence of a simple end-to-end chromosomal fusion because it implies more severe reorganization, as it could happen in Breakage/Fusion/Bridge cycles (BFB, figure 8), or possibly in BIR, as mentioned before.

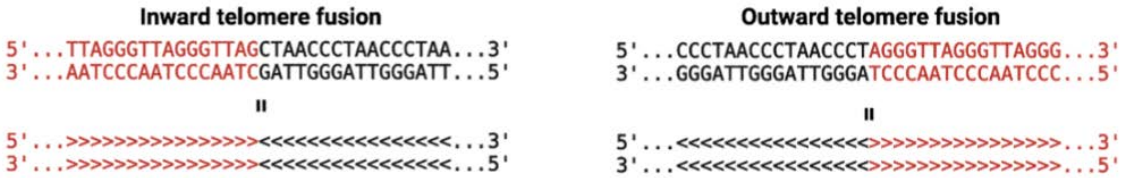


Figure 7 Types of telomere fusions (TFs). (A) "Inward TFs" are characterized by 5'-TTAGGG-3' repeats followed by 5'-CCCTAA-3' repeats, representing the expected genomic footprint of end-to-end TFs. (B) "Outward TFs" are characterized by 5'-CCCTAA-3' repeats followed by 5'-TTAGGG-3' repeats and are directly associated with the ALT (Alternative Lengthening of Telomeres) mechanism of telomere (Muyas et al., 2024)

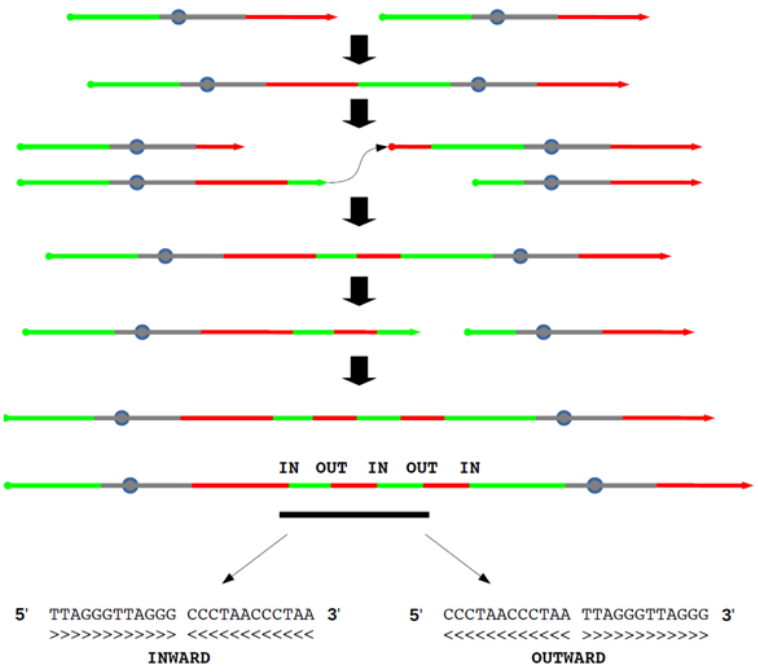


Figure 8: Scheme of Breakage/Fusion/Bridge (BFB) cycles *leading* to the production of inward and *outward fusions*.

Furthermore, a recent study has found that the occurrence of "outward" fusions, is especially high in ALT+ cancer cells, what has been used to suggest a role for ALT mechanisms and PML-bodies in their generation *(Muyas et al., 2024)*.

Although APBs are more commonly observed in cancerous cells, particularly those employing the alternative lengthening of telomeres (ALT) mechanism for telomere maintenance, they are not restricted to cancer cells. Emerging evidence suggests that ALT and APBs may have physiological functions beyond cancer development, such as contributing to telomere length regulation in normal somatic cells and participating in DNA repair processes. APBs have been observed in some mice cell types such as s*plenocytes*. *(Neumann et al., 2013)*

In that sense, the results of a recent study focused on telomere properties and longevity in bats suggest that high level of expression of ALT-related proteins, such as BL1, PARP1, RAD50, RPA1 and WRN, may be associated to increased telomere length and longevity *(Lagunas-Rangel, 2020)*.

Studies such the one described on bats made us to question whether differences in telomere metabolism exist across different branches of the mammalian tree of life, which could arise not only by differences in the regulation of genes involved in telomere maintenance, but also by their lack of conservation.

Given that genome-wide expression data across mammals could be scarce, we decided to use genomic sequences, both assembled and in the form of raw data, to define presence/absence profiles for genes related to telomere metabolism, including ALT-related genes, and to quantify the occurrence

of inward and outward fusions, as well as the frequency of telomeric motif variants, across mammals. Such genomic information was retrieved from the Zoonomia Project.

The Zoonomia Project focuses on studying the genomics of both shared and unique traits across some mammals. They have assembled genomes for around 130 species, most of which had not been previously characterized. Additionally, they have conducted whole-genome alignments for 240 species from diverse phylogenetic backgrounds. The project emphasizes prioritizing phylogenetic diversity and aims to make data accessible quickly and without restrictions. *(Zoonomia Consortium, 2020)* They have aligned the genomes to reference genomes to categorise and analyse thousands of genes via TOGA (Tool to infer Orthologs from Genome Alignments). *(Kirilenko et al., 2023)*

TOGA is a bioinformatic tool developed to integrate ortholog inference and gene annotation. It operates by taking gene annotations of a reference species, such as human, mouse, or chicken, and a whole-genome alignment with a query genome. TOGA identifies orthologous gene loci in the query genome, annotates and classifies orthologous genes, detects gene losses and duplications, and generates protein and codon alignments using machine learning techniques. TOGA extends its analysis to non-exon regions, including introns and intergenic regions. *(Kirilenko et al., 2023)*

The current project's objective was to conduct an in-depth bioinformatic analysis of TFs and TVRs with the raw data used in the *Zoonomia* project and compare the information with the presence and absence of genes involved in telomere metabolism across all the mammals analysed with TOGA.

Methods

Zoonomia information extraction (Extraction_zoonomia_information.py)

This Python script was crafted to perform text mining, enabling the extraction of valuable information from a range of websites:

Parsing of HTML content retrieved from the Zoonomia Project website was handled by the *parse_html* function. It involved utilizing the *BeautifulSoup* library to extract specific information such as species names, taxonomic order, family, and NCBI assembly accession numbers from the HTML content. The extracted data was then organized into a data frame for further processing.

Downloading and extracting ZIP files from specified URLs was facilitated by the *download_and_extract_zip* function. This process involved utilizing the requests library to download the ZIP file and the *zipfile* library to extract its contents into a temporary folder. This function was utilized for retrieving genomic information from JSONL.

The processing of genomic data stored in JSONL format was managed by the *extract_value* function. This function read the JSONL file and extracted relevant information such as accession numbers, assembly statistics, organism details, assembly methods, and SRA/SRR numbers associated with genomic sequences. The criteria for selecting SRR numbers associated with genomic sequences were

applied through multiple steps. Initially, experiment packages were extracted from the HTML content, and key parameters such as the size of the data and the number of runs associated with each experiment were identified. Distinctions were made between experiments with only one run and those with multiple runs. For experiments with a single run, priority was given to those with a size greater than 20 GB. Metadata such as instrument model, SRR value, total bases, total spots, average read length, and library layout were then retrieved for the selected experiment. In cases where multiple runs were present, the experiment with the largest individual run size was selected, with priority again given to those exceeding 20 GB. Subsequently, the same metadata was extracted for the chosen run.

The extraction of longevity information from web pages at the Animal Diversity Web (ADW) website was carried out by the *extractionLongevity* function. This function retrieved the webpage content using the requests library and utilized *BeautifulSoup* to parse the HTML content using a species name as input.

The coordination of the functions' execution was handled by the main function. Firstly, HTML content containing mammalian species information was read from the Zoonomia Project website and parsed using the *parse_html* function. Next, genomic data files were downloaded from external sources specified by the NCBI assembly accession numbers obtained earlier. The extract_values function was then utilized to extract genomic data details from these files. Additionally, the *extractionLongevity* function was employed to retrieve longevity information for each species from the ADW website. Finally, the collected data was organized into a data frame and saved to an Excel file for further analysis, named zoonomia_sp_info.xlsx.(table1)

TeloFusVarfinder

SrrDownloader

A class script was created to download and filter FASTQ files. SRR accession numbers were obtained from the table generated previously by Extraction_zoonomia_information.py. Then, SRR files were downloaded, extracted, and processed with three components of the sra- and fastx-toolkits.

- Prefetch: This component was responsible for retrieving the SRA data files from the NCBI servers and storing them locally in SRA format.
- Fastq-dump: Once the SRA files were obtained, fastq-dump was used to convert them into the FASTQ format, which is a standard file format for storing DNA sequences along with their quality scores.
- Fastq_quality_filter: This component was applied to FASTQ files to filter reads. It removed reads that did not contain at least 50% of bases with a quality of at least 30, helping to ensure the reliability of the data for downstream analysis.

This script had outputted fastq files depending on the type of library used, PAIRED or SINGLE which could be {specie}_filtered_1.fastq and {specie}_filtered_2.fastq, or {specie}_filtered.fastq, respectively.

TeloFusProcessor

The class process of telomere fusion detection in TeloFusVarfinder involves several key functions and methodologies:

Circular_rotations Function: Even though the principal canonical motifs are normally represented as TTAGGG or CCCTAA, fusion events could imply the connection of any of the corresponding circular permutations, as shortening of the telomere could have occurred up to any base. This function, therefore, generated a list of all possible circular permutations of three tandem repeats of the forward motif (5'-TTAGGG-3'), or three tandem repeats of the reverse motif (5'-CCCTAA-3'). The resulting list of patterns was used to search for reads that could contain sequences corresponding to telomeric fusions, as described below.

- *Process_pattern* Function: The *PairwiseAligner* class from the *Bio.Align* module was used by this function to perform local alignments between reads and the list of patterns generated by the previously described function. Results were stored as two BED files, corresponding to matches with forward and reverse patterns. The aligner was configured with parameters to allow for the identification of potential fusions with high sensitivity, while providing slight flexibility in matching (1 mismatch). To speed up the process, a preliminary search with a simple pattern of type TTAGGG or CCCTAA was performed.

- *Merge_coord* Function: Telomere fusions were detected by this function. The two BED files generated by the previous function, describing matches to forward and reverse patterns, were processed to identify potential fusion events by analysing the distance between matches. Intervals corresponding to the same read ID from each of the BED files were merged to provide a comprehensive view of potential fusions. The maximum allowed distance was 20 base pairs. The output of the process consisted in CSV tables with putative fusion details (table 2).

- *Fusion_coord* Function: This function used the coordinates identified by the *merge_coord* function to extract, from the original FASTQ files, the reads containing potential fusions. The output consisted in FASTA files with potential fusions. (figure S1).

- *candidate_fus* Function: All the processes described above were tied together by this main function to facilitate the search for telomere fusions. The overall workflow was managed by it, ensuring that each function was called in the correct sequence to efficiently identify potential fusions within the genome. Additionally, at the end of the process, BED files were deleted to free space.

TelomericVariantSearcher

The class process of telomere variability analysis in TeloFusVarfinder involved two key functions: are_similar and search_telomeric_variants.

- *are_similar* function was used to compare sequences to determine their similarity, accepting two sequences as input along with a maximum allowed difference *(max_error*). The number of differing characters between the sequences was calculated by the function and a boolean indicating whether the difference was within the specified error threshold was returned. A crucial role was played by this function in identifying telomeric variants that closely matched a reference sequence.
- *search_telomeric*_variants function was responsible for searching telomeric variants within the genome data (FASTQ files). It scanned through sequence data, using the *are_similar* to identify sequences that closely resembled known telomeric motifs. The inputs for this process included a threshold parameter (min_frequency) to filter the data based on minimum frequency, the fastq_file to be analysed, and a distinguishing number for the output file name. The function outputted a CSV file containing several key metrics: ID (the ID of the read or scaffold analysed), TelomereMotif (the original motif analysed, e.g., TTAGGG), Variant (the variant analysed from the original motif with permitted max_errors), Frequency (the number of times the motif appeared in the read without overlapping), *PorcentageTelomereSequence* (the percentage of the motif/variant appearing in the read, e.g., 30% of TTAGGG), and Length (the total length of the read). (table 3)

  The parameters were meticulously elected via experimentation. Min_frequency was selected as 1 to ensure any variant frequencies were not left behind. *Max_error* was decided as 1 to see only the frequency of the variants with only one mismatch, and the *original_motif* were selected as TTAGGG and CCCTAA, because these motifs tend to be the main motif across mammals. The function also included code to classify any given read as “telomeric” if the frequency of the original forward or reverse motif (TTAGGG or CCCATT) was the least 40%.

TeloFusVarfinder main script.

The functionalities of three separate scripts, *SRR_download_process.py, Telo_fusion_process.py and Telo_motif_process.py* (*SrrDownloader, TeloFusProcessor*, and *TelomericVariantSearcher, respectively*), were combined in the script telomere_analysis.py to automate the processes of downloading and filtering SRR data from repositories, detecting telomere fusions, and identifying telomeric motifs and variants. It was operated within a multiprocessing environment, allowing for the simultaneous execution of multiple tasks, thereby enhancing efficiency. It was designed to analyse and process SRR data from FASTQ files, which could be either single-end or paired-end.

In order to work in a cluster environment and further parallelize this code, using the strategy of job arrays, a Python function named *process_row* was defined. It was intended to process a specific row of

data from an Excel file. The function took one argument, *row_number*, which indicated the row number to be analysed. The Excel file named 'zoonomia_sp_info_c.xlsx' was read using pandas and stored in a *data frame*. It was then checked if the provided *row_number* was within the valid range of row indices in the *data frame*. If the *row_number* was invalid (i.e., less than 1 or greater than the number of rows in the *data frame)*, an error message was printed, and the function returned without further processing. If the *row_number* was valid, the row of data corresponding to the specified *row_number* was retrieved from the *data frame*. This row of data was then passed to another function named *process_species*, which encompassed the entire telomere analysis, parallelizing the three classes mentioned above.

A block of code was also included in the script to check if it was being executed as the main program. If so, the *row_number* was retrieved from an environment variable named 'SGE_TASK_ID' (which came from the SGE job scheduling system), converted to an integer, and passed to the *process_row* function for further processing. If the *row_number* was not within the valid range, an error message was printed.

Pipeline

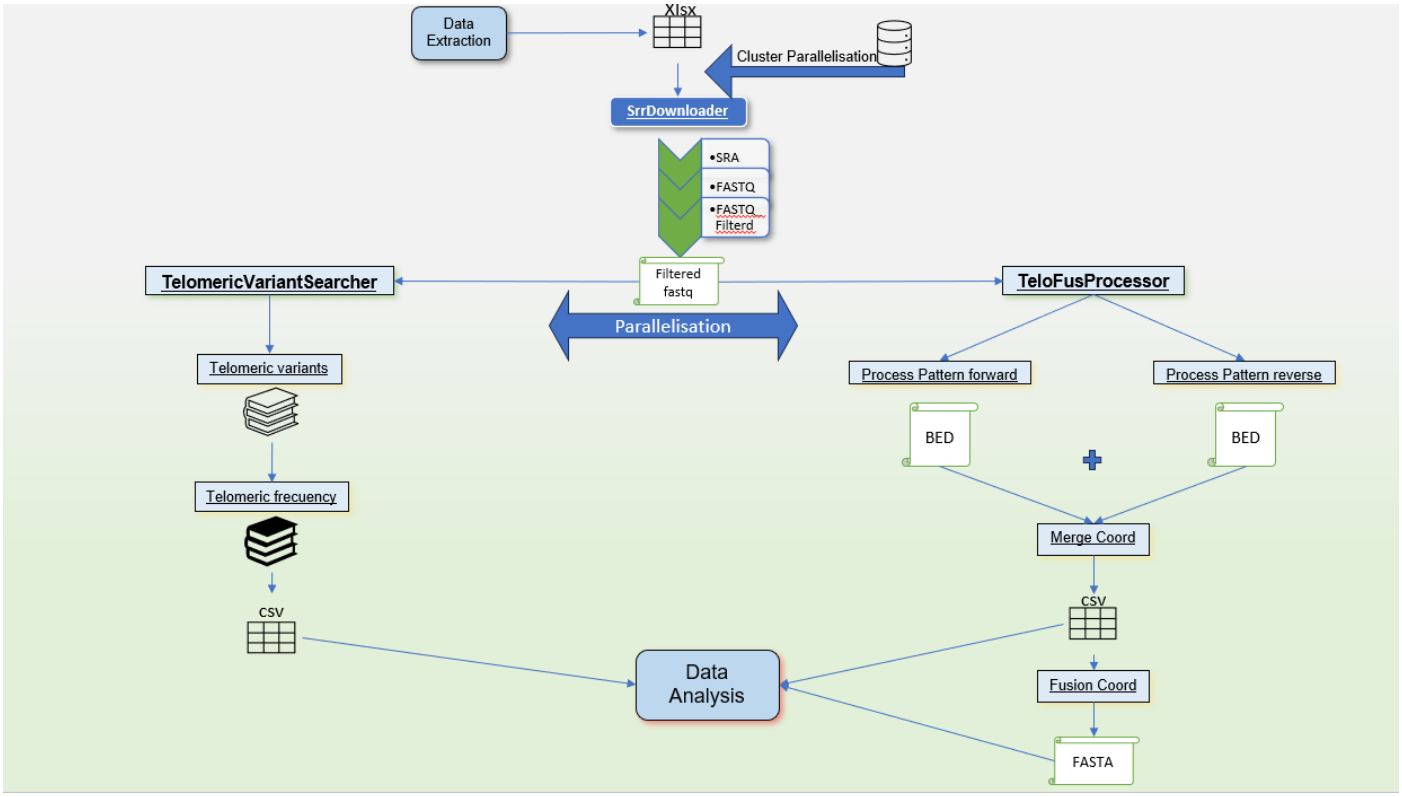


Figure 9: A representative figure of how the pipeline looks like.

Downstream data processing

Generation of primary data with TeloFusVarfinder took around a month and half and focused on 179 mammals from the Zoonomia project. Three main output files were generated for each mammalian species:

- Combined_fusions_{specie}_{number}.csv (table 2),
- Telo_fus_{specie}_{number}.fasta (figure S1)
- Telo_mv_{specie}_{number}.csv (table 3).

Output files were then processed to integrate information across all analysed species and to ensure better comprehension of the data by means of new scripts that were created in Python 3.12. In addition, complementary information regarding the biology of each species was retrieved from other resources.

Fusion extraction. (Fusion_analysis.py)

This script was used to consolidate all fusion events observed across the analysed species, resulting in the generation of two distinct tables: one containing raw fusion data labelled as "Fusion_results_raw.xlsx" and the other presenting data normalized by genome coverage, identified as "Fusion_results_coverage.xlsx". (table 6, table 7)

Data was categorized into different types of fusions based on their positional relationships and distances. These categories included inward fusions (where telomeric sequences were joined tail to tail e.g: TTAGGGxxCCCTAA), outward fusions (where sequences were joined in an opposite orientation, e.g: CCCTAAxxTTAGGG).

Within each category, a score of 35 (1 mismatch in the entire fusion) was required to further filter the candidate fusions. For instance, fusions with a distance of zero were categorized as {orientation}_distance_0 and those with distances greater than zero were categorized as {orientation}_other. Additionally, total frecuencies for different types of fusions were calculated. (eq.1)

Motif and variability extraction (Motif_variant_analysis.py)

The overall frequency of the primary motif (TTAGGG or CCCTAA) along with its variants was computed from the data stored in the table Telo_mv_{specie}_{number}.csv. (table3) The script began by defining a function called process_csv_file. Each CSV file was read, and its rows were iterated through, organizing the data into dictionaries based on the motif type ('TTAGGG' or 'CCCTAA'). For each motif type, the motif frequency, and its variants along with their frequencies were recorded. Once the telomere motif data was processed and organized, individual tables for each specie and number of FASTQ (R1, R2, or R) were generated, containing motif frequencies and variant frequencies. Each column in this merged table corresponded to the file name, while the respective frequency values constituted the values within these columns. (table13)

Moving forward, further analysis was performed on the merged Excel file to gain deeper insights into telomere variability. Sequences were compared to a reference sequence ('TTAGGG') to determine their

similarity and were categorized based on this criterion. The data was then separated into true and false rows. The false rows, which represented the CCCTAA motif and its variants, were then transformed into their complementary base sequence to be added into a new column, transforming the true data frame, and organizing the columns into table_{specie}_{number}_CCCTAA and table_{specie}_{number}_TTAGGG. Sums of columns for each species with the same name were calculated, removing the suffix _CCCTAA and _TTAGGG. This information was used to create a data frame that included total frequencies which were used to calculate percentages of variability and TTAGGG motifs, and additional metadata such as total spots run, total bases run, and genome size for each species then, percentage of total telomere size and the total telomere size was calculated from the total telomere reads previously calculated. (eq2 and eq3) Lastly, Telomere_variants_catalogue.xlsx was constructed. (table5)

TOGA information extraction (Extraction_gene_information.py)

TOGA (Transcript and Gene Ortholog Assignment) information was retrieved using BeautifulSoup and a URL in this script, with a specific focus on transcript and gene data derived from the assembled data obtained in TOGA using the zoonomia_sp_info.xlsx table. The script facilitated the download and organization of this data into {species}_orthologsClassification.tsv files, representing the orthology classification provided by TOGA. These classifications included distinct classes such as one2one (where a gene from the reference genome corresponded to one region of the query genome), one2many (where a gene from the reference genome corresponded to multiple regions of the query genome), many2many (where multiple genes from the reference genome corresponded to multiple regions of the query genome), and one2zero (where no orthologous genes were found in the assembled genome).

Furthermore, the script handled the extraction and structuring of TOGA data into {species}_loss_summ_data.tsv files, which encapsulated gene, transcript, and projection data categorized into various classes. These classes encompassed different scenarios:

PG - no orthologous chains identified, TOGA projected transcripts via paralogous chains and cannot make any conclusion.

PM - partial & missing. Most of the projection lies outside scaffold borders.

M - missing, assembly gaps mask >50% of the prediction CDS.

L - clearly lost.

UL - "uncertain loss", there are inactivating mutations but not enough evidence for "clearly lost" class. In other words: neither lost nor intact.

PI - partially intact: some fraction of CDS is missing, but most likely this is intact.

I - clearly intact.

Figure 10: Toga gene classification ((Kirilenko et al., 2023))

Both tables for each specie were added to 2 folders depending on whether the genome of reference was Orthologs_human_hg38 or Orthologs_mouse_mm10.

Gene annotation of proteins. (Specific_proteins_class.py and all_protein_class.py)

The gene information obtained from TOGA ({species}_orthologsClassification.tsv and {species}_loss_summ_data.tsv) was compiled by this script, and a table containing information about specific proteins previously selected was created. The species name was extracted from the tables loss_summ_data.tsv to obtain information. Subsequently, the gene ID was selected from loss_summ_data.tsv, and its respective classification was saved. Then, the orthologsClassification.tsv was utilized to locate the protein name of the corresponding gene ID. It was added as a column name with the structure of {protein}_g_{gene}_{reference}, and the classification name was stored as the value of the column, while the species name was added as the index row value.

Two scripts were used to extract specific proteins or all proteins. The selected proteins, 'ATRX', 'DAXX', 'TERT', 'RTEL1', 'PIF1', "TERF1", "TERF2", "TINF2", "TPP1", 'POT1', "POT1A", 'POT1B', "DCR1B", "ACD", "WRN", "BLM", "TP53", "SIRT1", "RAD52", "FEN1", "RPA1", 'RPAIN', 'RPA2', 'ASF1A', 'SETDB1', 'RMI', 'BLM', 'TOP3A', 'PML', 'ZBTB40', and 'RAD51', were added to the data frame then a table was created zoonomia_proteins_class.csv, For all proteins the other table was name Zoonomia_all_proteins_class_human.csv (table8).

Chromosome number extraction

Chromosome numbers for a certain collection of species was manually annotated using information from the literature *(Rumpler et al., 2011), (Sotero-Caio et al., 2017), (Lee Kolnicki, n.d.), ('List of Organisms by Chromosome Count', 2024), (Kurihara et al., 2017).*

Results

Selected mammal species

As mentioned above, the Zoonomia project covers at present 250 mammal species. The associated database (https://zoonomiaproject.org) provides access to raw sequence data files (via SRA), genomic assemblies (via NCBI genome), gene annotation and orthologs (via TOGA), and additional metadata mostly related with the sequencing and assembly projects (such as type of sequencing technology, read length, or number of contigs in the current assembly version).

Out of the 250 species described in Zoonomia, 179 were selected for the current analysis on the basis of the following criteria:

- Illumina sequencing technology
- All raw sequence files were selected in order to get the best criteria number with higher information, mentioned in the methods.
- Library layout paired or single end.

Computing strategy

The scripts described in methods, written in Python as part of the current project, were executed at CNIC's SGE HPCC cluster. The main pipeline used as main input a file that listed the collection of accession numbers for genome assemblies and raw sequence data. Given the volume of data and the existing limitation in storage capacity, data was downloaded, uncompressed, processed and removed from disk as a job array, allowing a maximum concurrency of 6. Once scripts and pipelines had been written and optimized, the whole process took 7 weeks to run. (figure 5)

Telomeric fusion detection and phylogenetic distribution

After extracting all the data explained above in the methodology section, we re-assembled it into a unique table (table9A, table9B, table9C) that combined information for the variables described in the following summary, which were used in the whole analysis (notebook):

Variables	Description
Inward_distance_0	Forward + reverse fusions with distance within the fusion equal to 0 e.g. TTAGGGCCCTAA
Outward_distance_0	Reverse + forward fusions with distance within the fusion equal to 0 e.g. CCCTAATTAGGG
Inward_other	Forward + reverse fusions with distance within the fusion higher than 0 and lower than 20 (e.g. TTAGGGxxxCCCTAA, where x represents any nucleotide)
Outward_other	Reverse + forward fusions with distance within the fusion higher than 0 and lower than 20 (e.g. CCCTAAxxxTTAGGG. where x represents any nucleotide)
Total_fusions	Sum of all fusions
Total_fusions_0	Sum of all fusions with distance 0
Total_inward	Sum of all inward fusions
Total_outward	Sum of all Outward fusions
Genome_coverage	The genome coverage previously calculated per run
Longevity	Lifespan of the species

Chrom_num	The number of chromosomes of each species
{protein}_g_{gene_ID}	Name of the protein selected
Total_telomere_size	Total number of reads classifed as telomeric multiplied by the average number of base pairs in the reads
Percentage_telomere_size	Total telomere size divided by the total number of base pairs of the genome
Porcentage_variability	Total frequency of all non-canonical telomeric motifs, divided by the total frequency of motifs, all multiplied by 100
Porcentage_TTAGGG	Total frequency of all TTAGGG motifs divided by the total frequency of motifs, all multiplied by 100

Results generated by the script TeloFusProcessor described the number of inward and outward fusions, detected in raw sequence data, with intervening sequences of either length equal to zero, or higher than zero.

Figure 11 describes the number of telomeric fusions of any type, normalized by genome coverage * 100, for the 20 organisms with more fusions. Several representatives of the order Chiroptera and Rodentia are among the species with more fusions.

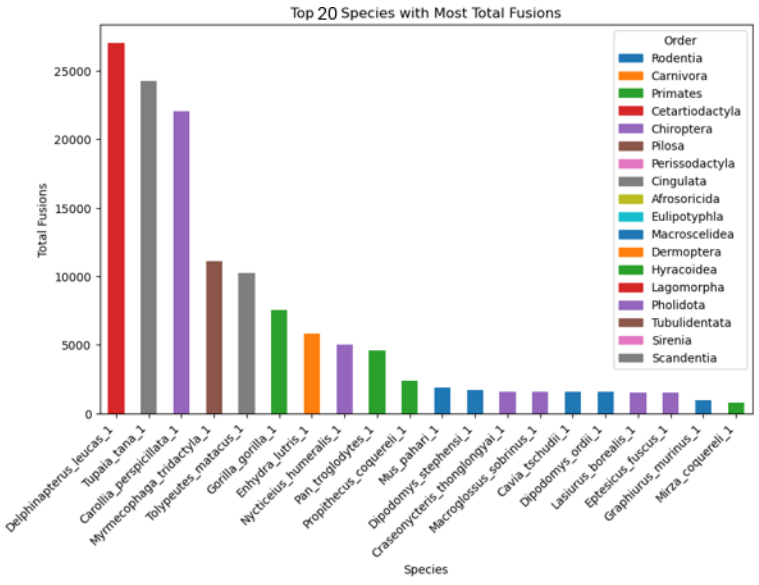


Figure 11 Top 20 organisms with more fusions.

We then analysed the data to determine whether the distributions of telomeric fusion frequencies were different between orders (Figure 12). For most mammal species the number of fusions was zero or near zero. However, fusion frequency distributions had higher median values for the order Chiroptera, when focusing on inward or outward fusions without intervening sequences (distance = 0), because the occurrence of specific mammal species with higher fusion numbers.

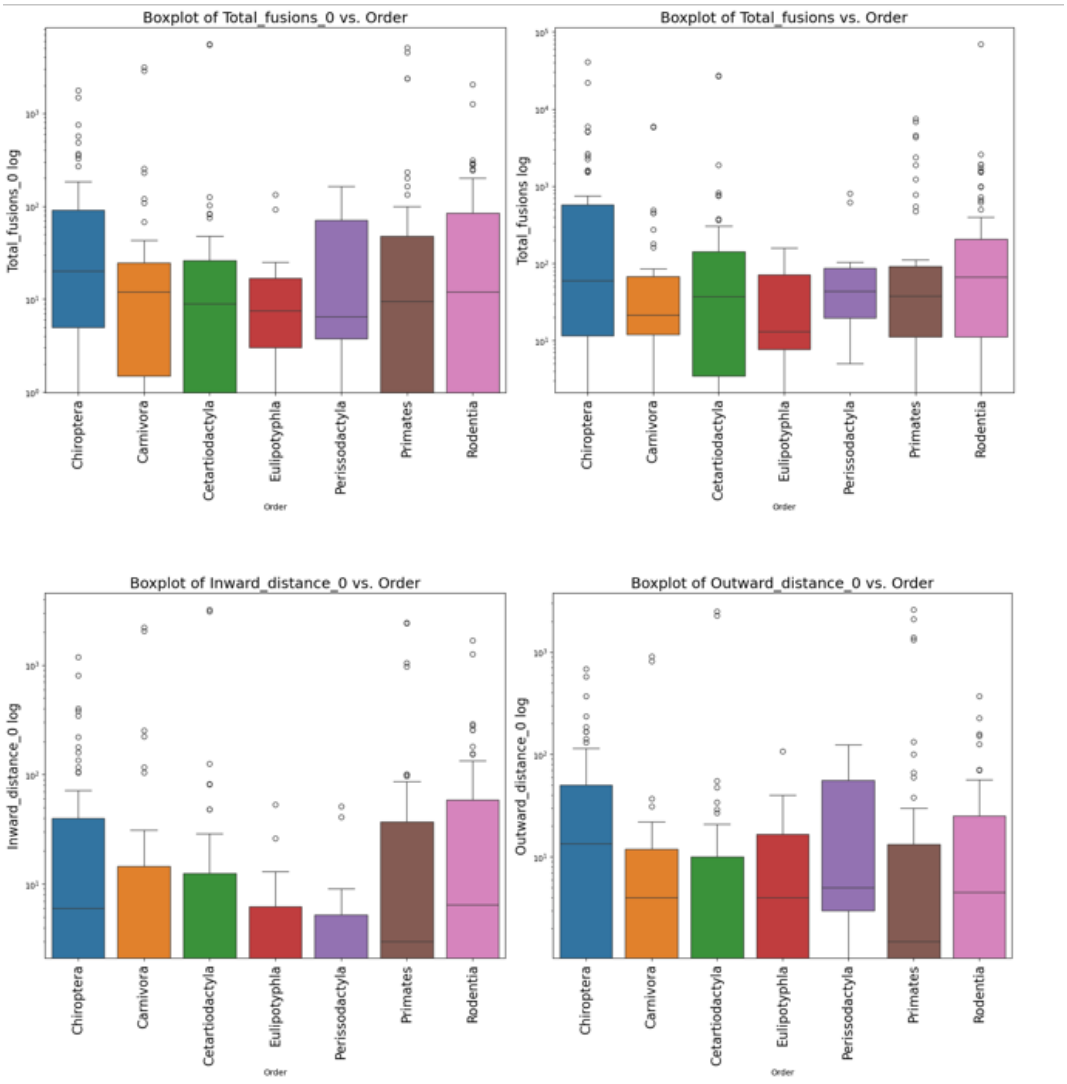


Figure 12: Occurence of telomeric fusions, by order: Total fusion_0 (upper_left), Total_fusion (upper-right), inward_distance_0 (down-left) and outward_distance_0 (down-right)

Characterization of fusions breakpoints

Having processed raw sequence data to find cases of putative telomeric fusions, we characterized the sequence at the fusion breakpoints. To do it, we compared each detected fusion with a catalogue of all possible fusions, considering that the end of any naturally occurring stretch of telomeric repeats would not necessarily finish with a canonical motif (TTAGGG or CCCTAA). Given that for each canonical motif there are six circular permutations, the total number of possible breakpoints both for inward and outward fusions is 36. (Table 3) We observed that the most frequent fusion breakpoints corresponded only to 7 or 8 of all 36 possible breakpoints, for inward and outward telomeric fusions, respectively (figure 13).

For inward telomeric fusions:

1. Around 25% of inward fusions have a breakpoint of AGGGTT/AACCCT or GGGTTA/ACCCTA
2. Around 23% of inward fusions have a breakpoint of TAGGGT/TAACCC or AGGGTT/ACCCTA
3. Around 20% of inward fusions have a breakpoint of GGGTTA/CCCTAA or TAGGGT/AACCCT

For outward telomeric fusions:

1. Around 20% of outward fusions have a breakpoint of ACCCTA/GGGTTA or AACCCT/AGGGTT
2. Around 10% of outward fusions have a breakpoint of TAACCC/TAGGGT
3. Around 11% of outward fusions have a breakpoint of TAACCC/TTAGGG or AACCCT/TAGGGT
4. Around 12% of outward fusions have a breakpoint of CCCTAA/GGGTTA or ACCCTA/AGGGTT

It is relevant to note that the most frequent breakpoint sequences detected for inward fusions match with the expected breakpoint sequences as derived from the observed sequence at telomere ends. (Sfeir et al., 2005).

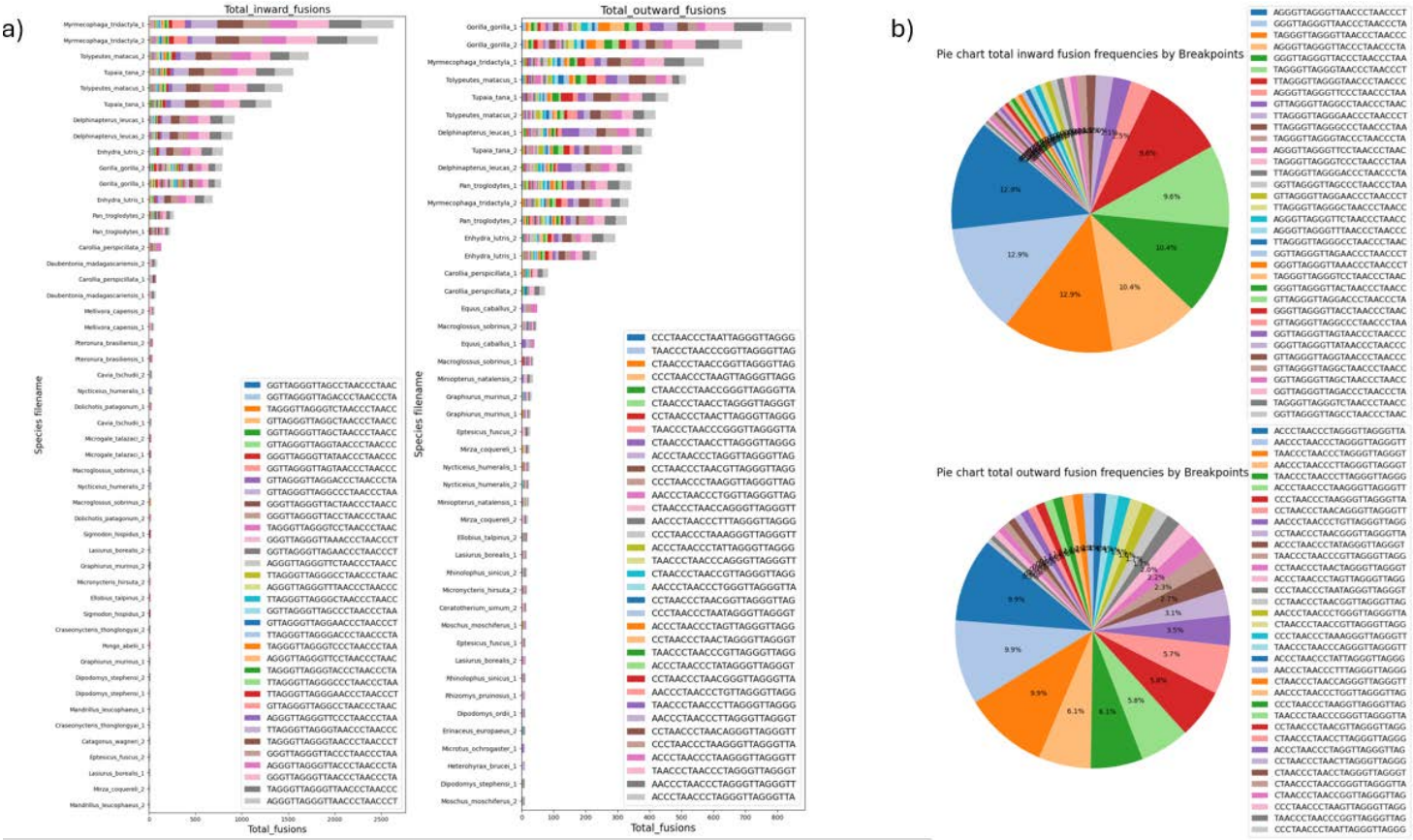


Figure 13: Frequency of detected breakpoints for inward and outward fusions, a) by species, disclosing the data for fusions detected in R1 (1) and R2 (2) reads for paired-end data, or in single read (0); b) globally, for inward and outward fusions. (figure S11)

PCA was then applied to the fusion breakpoint data, which consists of 36 variables representing different types of fusion breakpoints. By plotting the species points in the reduced-dimensional space defined by principal components 1 and 2, we could visually assess how the different mammal species cluster according to their breakpoint frequency profile. Additionally, by labelling the species points according to their taxonomic order (e.g., Chiroptera or Rodentia), we could observe that Rodentia and Chiroptera have different fusion breakpoint characteristics. (figure 14)

Figure 14 Dot plot with the PC1 and PCA2 components of the PCA analysis of telomere breakpoints.

After plotting the PCA cumulative explained variance we observed that the first two dimensions explained most of the variance (90%). (figure 15)

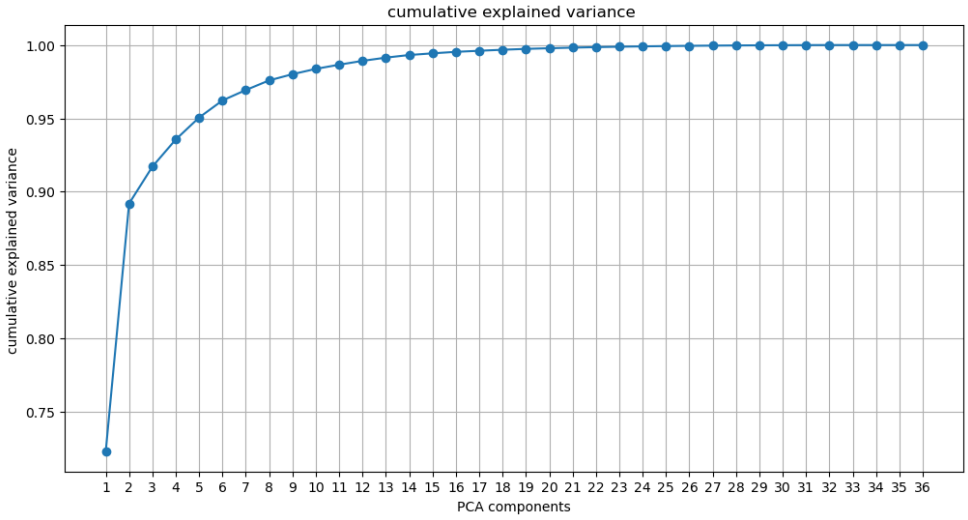


Figure 15: Cumulative explained variance from PCA.

Telomeric motif variability and telomere length

As part of the process of detecting putative telomeric fusions in raw sequencing data, reads were classified as "telomeric" if more than 40% of their length corresponded to the canonical motifs TTAGGG or CCCTAA. The fraction of reads classified as "telomeric" was considered a proxy of telomeric content for any individual genome. In addition, telomeric reads were further analysed to determine the frequency in which variations of the canonical motifs were found, to calculate the percentage they represented. Then, we explored how telomeric content values and telomeric motif variant percentages were distributed across mammal orders.

Telomeric content values had maximal values for some Chiroptera species (Figure 16 left). Similarly, telomeric motif variant percentages had higher values in Chiroptera (Figure 16 right).

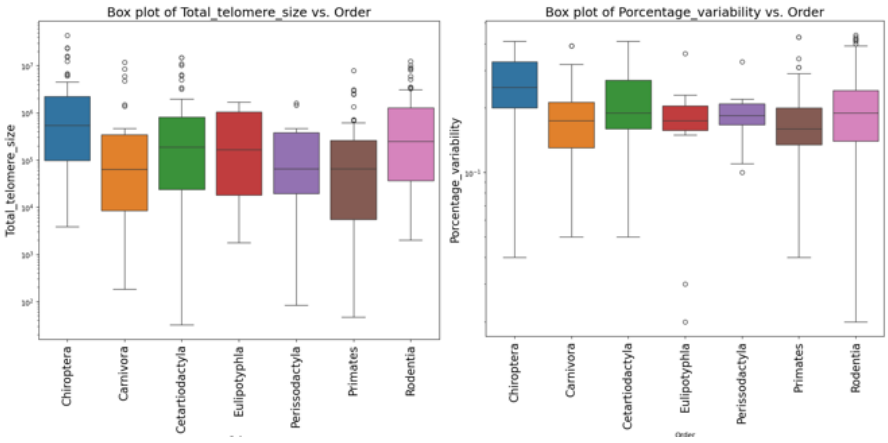


Figure 16: Box plot representing the distribution of telomere content (left), and percentage of telomeric motif variants (right), by order.

Interaction of telomeric data with the species-specific collection of proteins involved in telomere maintenance.

As mentioned in the introduction, a recent study with bats had suggested that high level of expression of ALT-related proteins, such as BL1, PARP1, RAD50, RPA1 and WRN, may be associated to increased telomere length and longevity *(Lagunas-Rangel, 2020)*.

We then hypothesized that, given the absence of expression data for most of the mammals, possible differences in the complement of proteins involved in telomere maintenance across mammal species could correlate with differences in the properties of telomeric sequences, for example, the number of detectable telomeric fusions, or the percentage of telomeric motif variations.

To perform such analysis, we took advantage of TOGA results. TOGA is a component of the Zoonomia project focused on gene annotation and ortholog detection, both of which are based on the construction of genome alignments against reference genomes such as those from *Homo sapiens*, *Mus musculus* and *Gallus gallus*. The output of this approach consists in species-specific annotations describing whether orthologous genes and potential transcripts are likely to exist by comparison with one of the reference genomes. The approach followed by TOGA is based on the calculation of alignment percentages for each gene or transcript. Then, categories are defined on the basis of conservation percentages but, also, in the presence of mutations that are predicted to inactivate any given gene. TOGA annotations consider six possible situations for any given gene already explained on the methods section: I, PI, UL, L, M and PM.

Given the complexity of the TOGA classification, we decided to simplify the classes to only two categories that would be represented by a binary or Boolean variable, where I and PI would be simplified to true values, and UL, L, M, and PM would be simplified to false values, representing the presence or absence of the protein (table9.B). We based our decision on the methodology employed by TOGA in classifying orthologs as one2one, one2zero, and so on, for aligning the data. While TOGA utilized UL as a reference sequence for alignment, upon observing several alignments, we observed discrepancies and subsequently classified UL as absent. This reclassification aimed to streamline complexity and enhance clarity within the dataset.

Presence/absence profiles were generated for the 179 mammal species analysed as part of this project, and a collection of 19,000 proteins. However, most of the downstream analyses were focused on 26 proteins that were selected from the literature because they are involved in telomere maintenance or they have been associated to APBs (fig2),(fig3) and (table8). We also defined two presence/absence profiles for each mammal species: one based on comparisons against Mus musculus and another one based on Homo sapiens.

Having defined gene conservation as a simple binary pattern (presence/absence), we decided to study possible interactions with numerical variables describing the frequency of telomeric fusions, the percentage of telomeric variants or the fraction of telomeric content.

To do it we decided to apply mostly two strategies:

- Determine the possible correlation between numerical variables (frequency of telomeric fusions, the percentage of telomeric variants or the fraction of telomeric content) and the binary gene presence/absence pattern, defined for each protein, used as a dummy variable. Correlation would be represented by the Pearson correlation coefficient.
- Determine whether differences in the value distribution for numerical variables (frequency of telomeric fusions, the percentage of telomeric variants or the fraction of telomeric content) exist between two categorical groups following the binary gene presence/absence pattern, defined for each protein. Differences would be assessed using the Kruskal-Wallis test.

Both strategies would be applied using binary gene presence/absence patterns derived from comparisons against Homo sapiens and Mus musculus.

As we explained above 26 proteins were selected from known reference information and then analysed with the data obtained.

Interaction of telomer*ic fusion frequency with* protein profiles

Results generated with presence/absence profiles based on a human reference:

When using presence/absence patterns based on a human reference the protein with a higher negative and overall correlation with telomeric fusion frequency was RPAIN. For total fusions with null intervening distance (*Total fusions 0)*, it had -0.38 correlation coefficient (r) and a p-value of $9.55*10^{-4}$. For total fusions, independently of the length of the intervening distance (*total fusions)*, the protein RPAIN had r=-0.43 and a p-value of $7.22*10^{-4}$.

It is noteworthy to stress a correlation with ASF1A with r=-0.28 and p-value of $1.27*10^{-2}$ in *"Total fusions 0"*. However, the correlation for "*total fusions"* was -0.09 and the p-value of $6.7 *10^{-3}$. Other protein that could be highlighted is PML, which had a negative correlation coefficient of r=-0.20 for "*total fusions";* however, the p-value, calculated with the Kruskal-Wallis test was higher than 0.05. (table11)(table12).

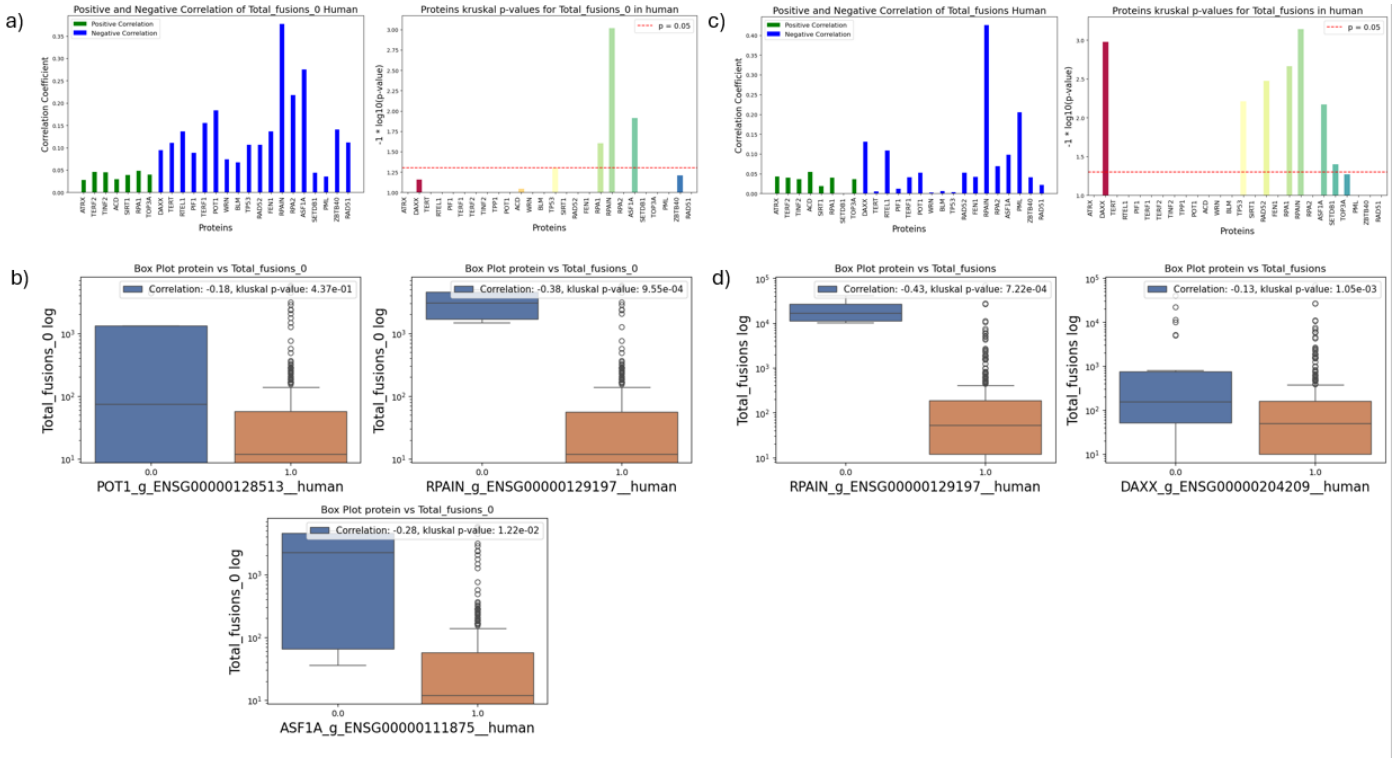


Figure 17: A representation of the correlation between total fusions 0 and total fusions vs the proteins selected in human reference. a) Data histograms of the correlation between total fusions 0 between the human proteins left side and the p-value of the proteins calculated with Kluskal Wallis. b) three representative box plots of POT1, ASF1A and PRAIN whit the corresponding correlation coefficient and its p-values. c) Data histograms of the correlation between total fusions between the human proteins left side and the p-value of the proteins calculated with Kluskal Wallis. d) two representative box plots of ASF1A and PRAIN whit the corresponding correlation coefficient and its p-values. figS7.

As we can observe in the box plots RPAIN has the highest p-value, highest correlation and appears with a significant difference between the absence and presence of the protein and the fusions. We could observe as well that POT1 appears to have a good correlation, however, its p-value does not support a significant difference.

To place the results obtained for the collection of 26 selected proteins in the context of a bigger collection of proteins, we calculated the distribution of correlation values, and its corresponding gaussian curve, for the whole collection of 19,000 proteins assembled on the basis of TOGA

information. Correlation values calculated for RPAIN were among the most extreme, in comparison with the values calculated for any protein. (figure 18)

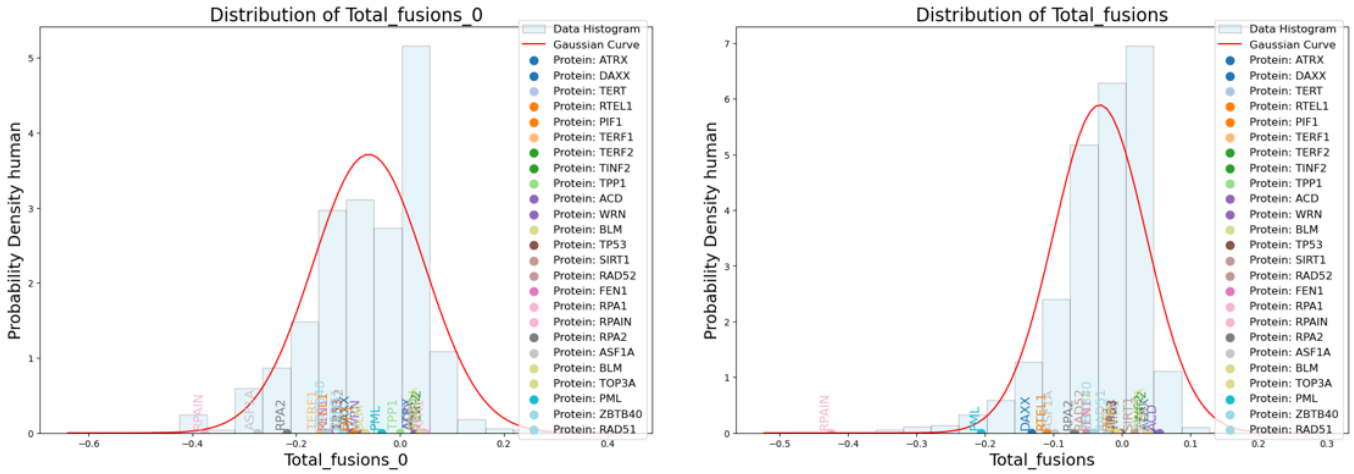


Figure 18: Two correlation matrix distributions for total fusions 0 and total fusions vs all 19000 human proteins and the visualization of the main specific proteins already selected.(figS6)

In our additional analysis using Random Forest, we observed that RPAIN exhibited the highest importance score, surpassing 0.30, with a particularly notable value of 0.45 for total fusions. Conversely, ASF1A demonstrated significant importance exclusively in the context of total fusions, with a distance score of 0 (figure 19).

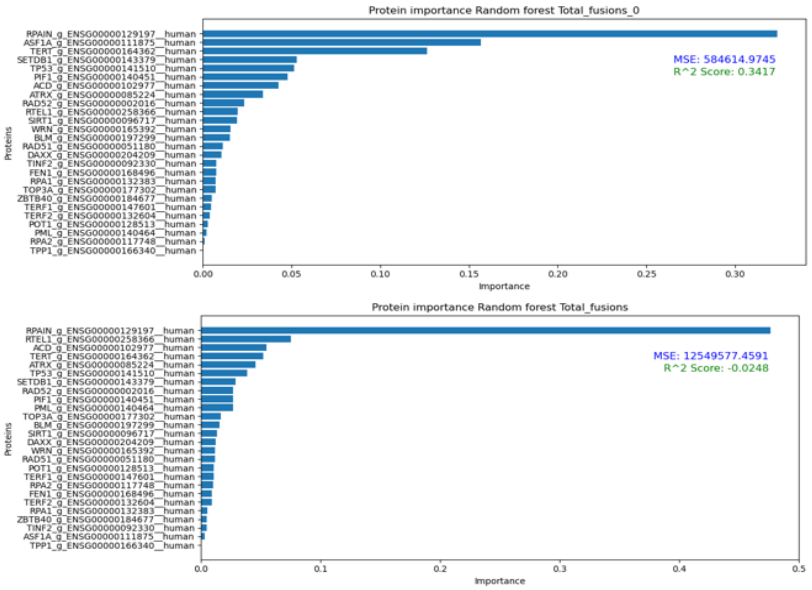


Figure 19: Random Forest protein importance against "total fusions" and total fusions 0", using human as reference to define orthologous genes.

Results generated with presence/absence profiles based on a mouse reference:

When using presence/absence patterns based on a mouse reference the protein with the highest negative correlation with the number of total fusions was RTEL1, however, it had a weak p-value. For total fusions the most representative is ASF1A as it has a balance between correlation coefficient and a p-value lower than 0.05.

For "total_fusions_0", RETL1 had r = -0.31 and a p-value of 0.24, however ASF1A had a negative correlation of r = -0.27 r and a p-value of 0.01. For "total fusions" DAXX and RAD52 were the only with a weak negative correlation with r = -0.13 and r = -0.11, respectively and a p-value of $5.7*10^{-4}$ and $2.7*10^{-3}$. We could also highlight the weak positive correlation of ATRX r = 0.14 and a p-value of $2.13*10^{-5}$. (table11)(table12).

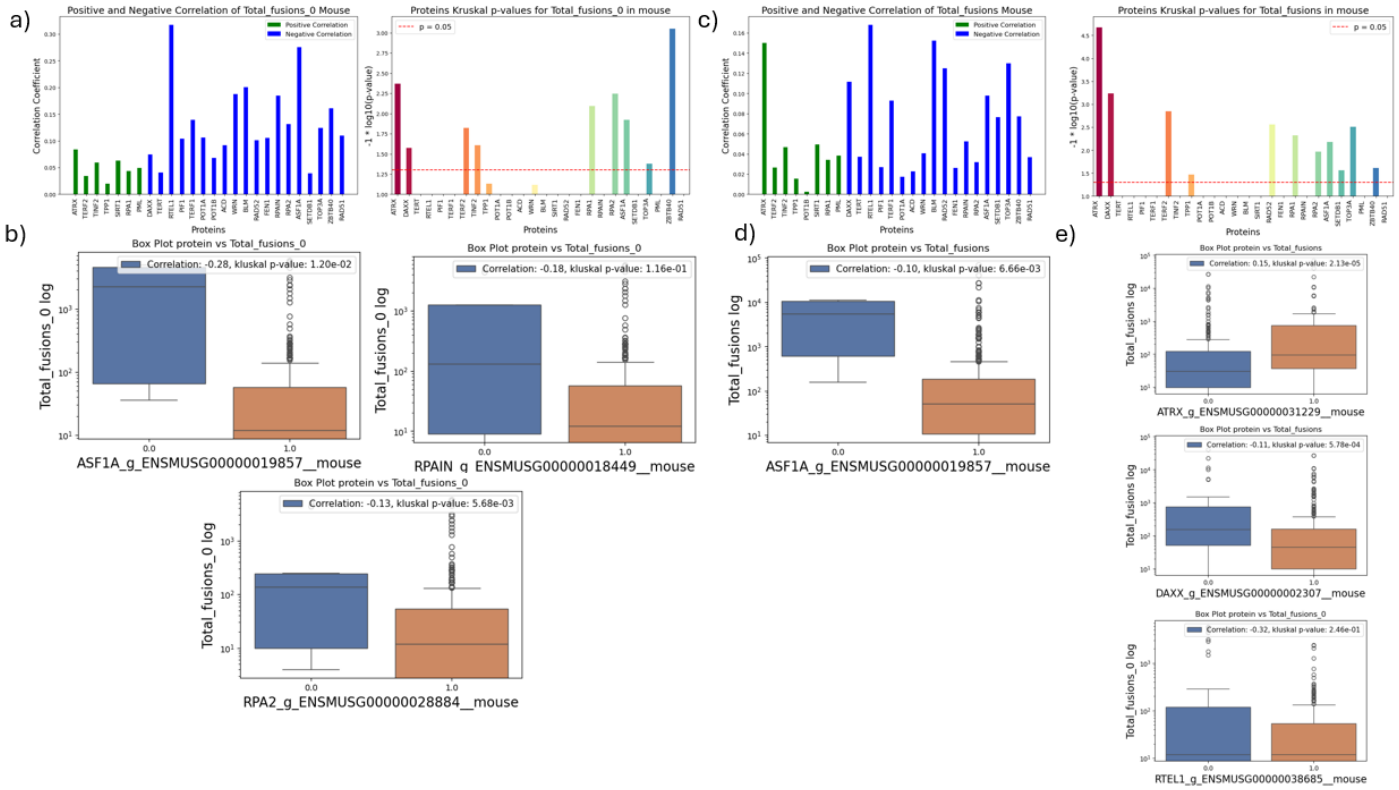


Figure 20: A representation of the correlation between total fusions 0 and total fusions vs the proteins selected in mouse reference. a) Data histograms with the correlation between total fusions 0 between the human proteins left side and the p-value of the proteins calculated with Kluskal Wallis, right side. b) Tree representative box plots of RPAIN, ASF1A and PRA2 whit the corresponding correlation coefficient and its p-values. c) Data histogram with the correlation between total fusions between the human proteins left side and the p-value of the proteins calculated with Kluskal Wallis. d) One representative box plot of ASF1A whit the corresponding correlation coefficient and its p-values. e) Box plots of the correlation between ATRX, DAXX and RTEL1 with total fusions. FigS7

As evidenced by the analysis, while RETL, DAXX, and RAD52 exhibit a negative correlation, and ATRX displays a positive correlation, with statistically significant p-values, only ASF1A demonstrates a notably strong correlation, as observed from the box plots.

The distribution of the correlation matrix, along with its corresponding Gaussian curve, was analysed. Notably, for mouse RTEL1, a considerably distance was observed compared to the other specific proteins under consideration.

Further examination reveals that ASF1A tends to cluster closer to the far negative end in both scenarios, albeit with a relatively weak probability. Conversely, RTEL1 is noticeably distant from the remaining proteins, indicating a stronger correlation. (figure 20)

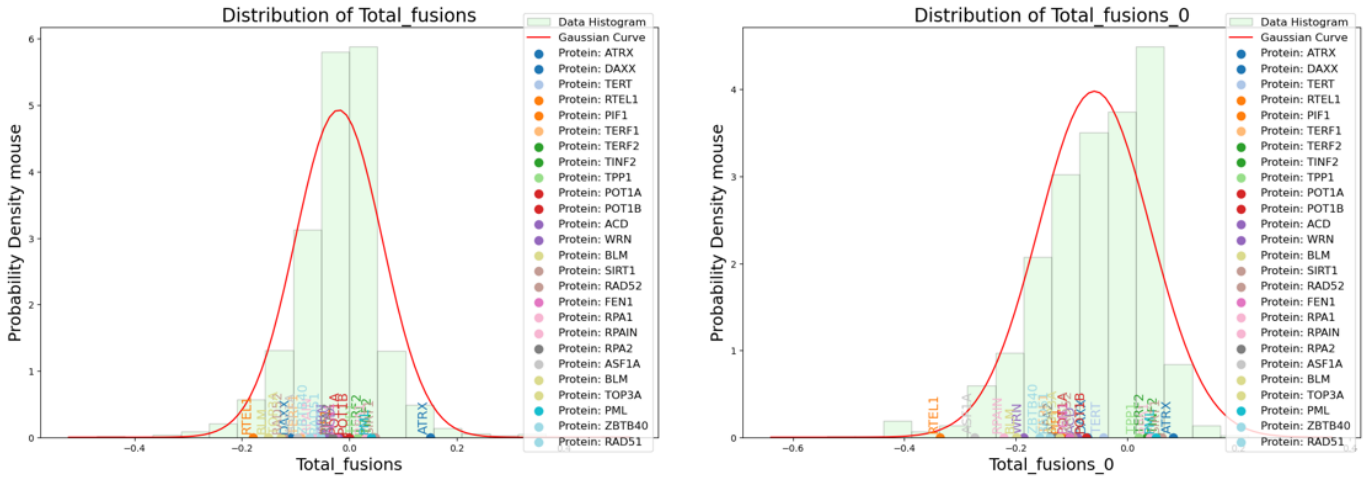


Figure 21 Two correlation matrix distributions of total fusions 0 and total fusions vs all 20000 mouse proteins and the visualization of the main specific proteins already selected. (figS6)

In the analysis of mouse proteins, the random forest model yielded suboptimal scores overall. However, it's worth noting that ASF1A exhibited a relatively importance score of around 0.10, particularly in the context of Total_fusions_0, therefore we can conclude that there is not a good correlation.

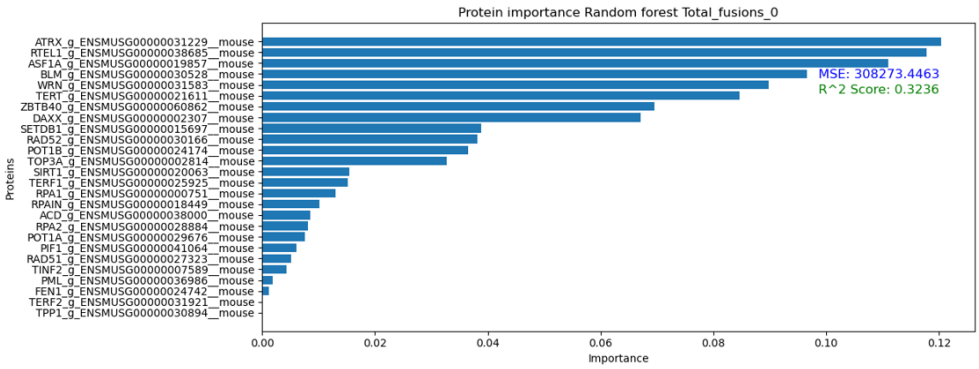


Figure 22 Random Forest proteins importance against "total fusions 0" in mouse as the reference.

Interaction of telomeric variability and telomere content with protein profiles:

Percentage of telomeric motif variants

The correlation between the profile for specific selected proteins and the percentage of variability has weak positive and negative correlations.

POT1B (from mouse) and POT1(from human) have a weak positive correlation of r = 0.17 and representative P-value of 9.5 *10^{-3} and r = 0.19 and a P-value of 5*10^{-4}, respectively. DAXX from both have weak negative correlation and a representative low p-value. RAD52 in both, as well we have a negative correlation and a low P-value. We perceive that the correlations are not higher than 0.20 in both negative and positive. (table11)(table12).

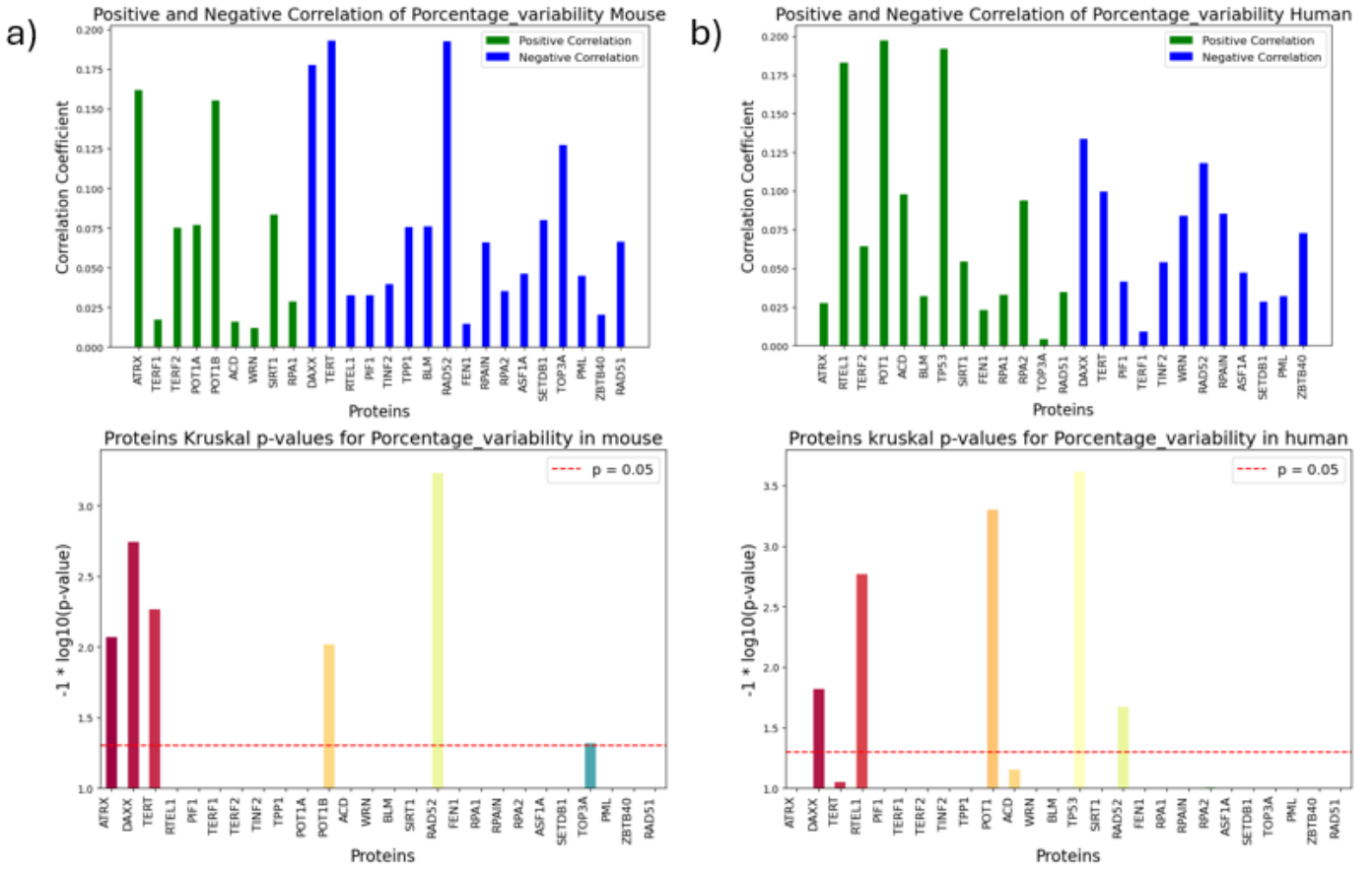


Figure 23: Data histograms represent the correlation coefficient (upper side) and the Kluskal Wallis P-value (downside) of proteins vs percentage of variability in mouse(a) and human(b).

Total telomere size

Surprisingly RPAIN has strong negative correlation in human as reference. The correlation coefficient is r = -0.38 and a p value of $1.2*10^{-3}$, however, in mouse reference RPAIN have a weak negative correlation and a p value higher than 0.05. In mouse aligned proteins the correlation coefficient of DAXX is r = -0.24 and its P-value higher than $2.1 * 10^{-5}$. In both, human and mouse reference we see a negative correlation with ASF1A with a representative low p-value in both cases. (table11)(table12).

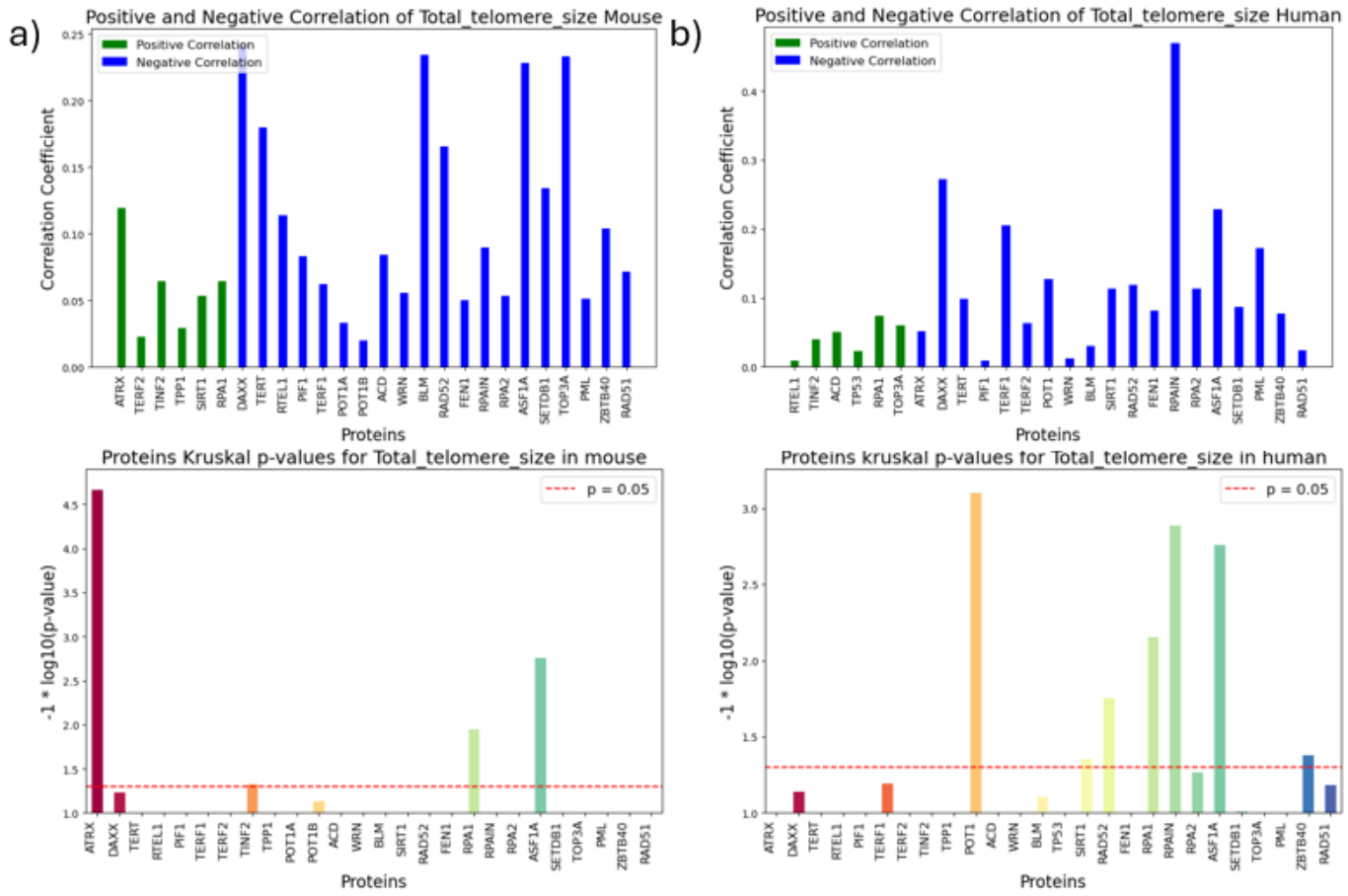


Figure 24: Data histogram which represents the correlation coefficient (upper side) and the Kluskal Wallis P-value (downside) of proteins vs telomere size in mouse(a) and humans (b).

Interactions between fusion frequency and telomeric variability with other variables

Fusions vs chromosomal number, longevity, and telomere variability:

We observed a positive correlation with all types of fusions. An interesting thing is that we cannot observe a relationship between fusions a variability and fusions with chromosome number. On the other hand, we see positive correlation with the telomere length.

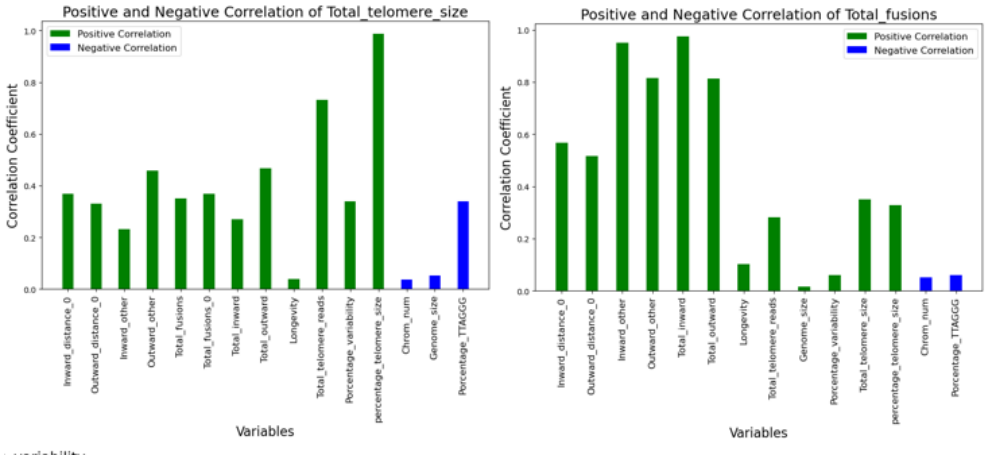


Figure 25: Data histogram of correlation between fusions and non-categorical variables.

Analysis of percentage of variability, chromosomal number, and total telomere length:

Generally, percentage of variability has low correlation coefficients with all non-categorical variables however, it seems to have a stronger positive correlation with the telomere size, which means as the telomere increase more provability to have higher variability.

Simultaneously total telomere size has a positive correlation with the fusions and as said before with variability.

For the chromosome number, the only weak correlation to be highlighted is longevity, which has a positive correlation not higher than r = 0.20

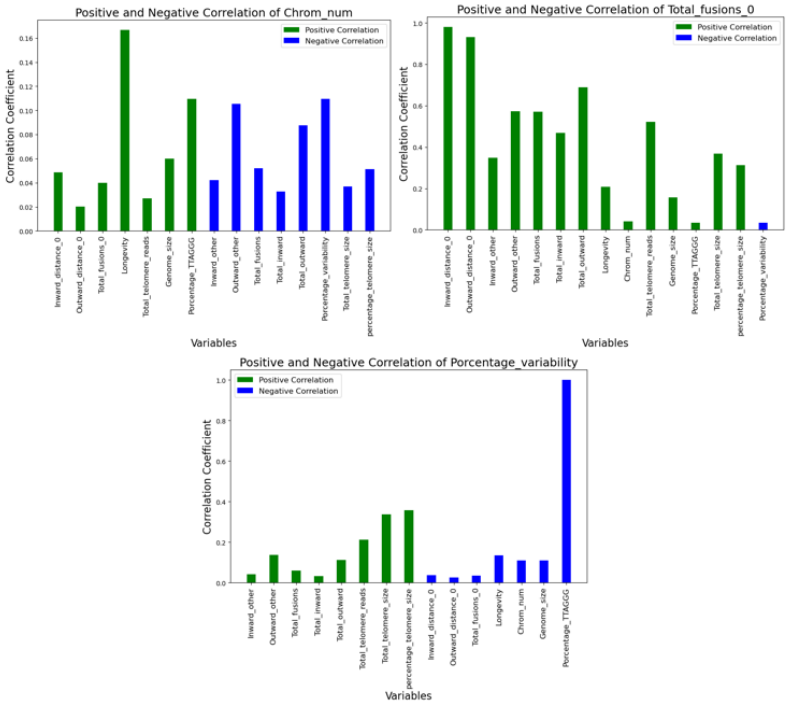


Figure 26: Data histogram of percentage of variability (upper side), total telomere size (down-left side) and chromosome number (down-right side)

Percentage of variability vs fusions by order:

We observed that rodents do not get the highest number of fusions as well as primates and they have a negative correlation with variability. Secondly *Cetartiodactyla* has a positive correlation with variability which means that to higher variability, higher is the probability to have fusions. Carnivora only has a positive correlation with outward distance 0 fusions. (figure 27)

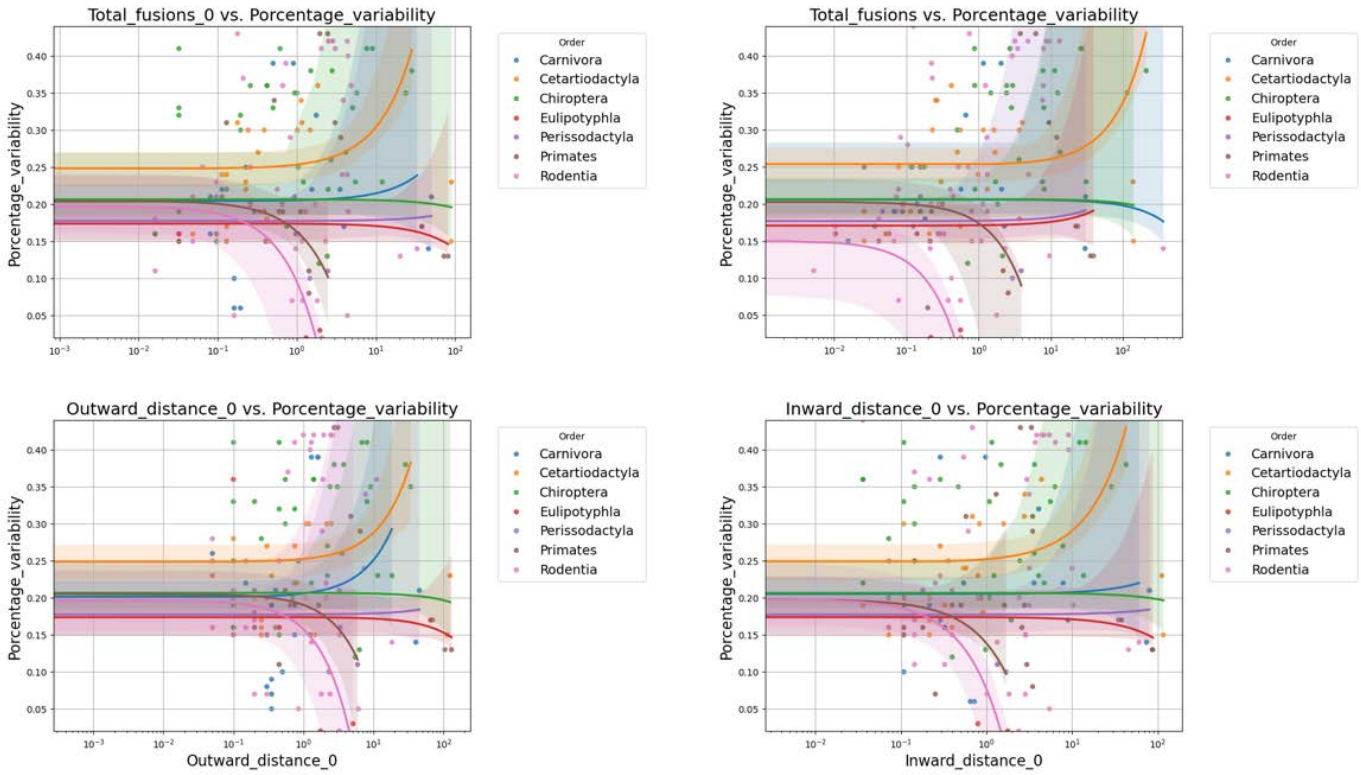


Figure 27: Dot plot of correlation coefficient of variability vs fusions by order.

Discussion

- We have used TeloFusVarfinder to analyse genomic sequencing datasets to quantify telomeric fusion occurrence across 179 mammal species. Our breakpoint analysis indicates that inward fusions are conformed to the typical configuration of telomere ends, as described in previous literature (Sfeir et al., 2005). This observation supports that the observed fusion evidence corresponds to real fusions, and not to sequencing artifacts. Observed breakpoint frequencies, however, differ from the data obtained in ALT+ cancer cells (Muyas et al., 2024).

- In our investigation of telomere reads and variability across mammalian species, we have observed slightly differences in classification by order, with Chiroptera exhibiting a particularly higher concentration of telomere variability and telomere content. This increased telomere content in *Chiroptera* compared to other mammals has been documented in the literature *(Lagunas-Rangel, 2020)*. Given that telomeres are susceptible to strand-specific sequencing

errors, giving reads with strict quality scores we minimized the errors, aligning with findings from recent studies (Lee et al., 2018). Notably, these studies suggest that telomeric variability and fragility are generally elevated in ALT+ cancer cells, providing context for the observed higher variability and telomere concentration in Chiroptera.

- Analysing the data and reading through the literature, the generation of fusions are not well understood however, some theories about how they are produced come in mind, the first theory is explained in the introduction section (figure 8) in which the outward fusion occurs by fusing fragmented telomeres originated by ALT activity (*Muyas et al., 2024*). I suggest a complementary pathway for Outward fusions, researching into alternative lengthening of telomeres (ALT) mechanisms, such as break-induced replication (BIR) unravelled by *(Malkova & Ira, 2013)* (figure 3), suggests that telomeric elongation can occur by diverse methods, they demonstrated that in order to lengthen the telomere, it needs a template, which could be another telomere or an extrachromosomal dsDNA, by copying an extrachromosomal dsDNA, with a telomeric fusion already there, the mechanism could elongates using that in any direction and therefore it could create the Outward fusion.

- Extensive exploration of the literature and previous studies have discovered numerous ALT-PML associated proteins. Motivated by this insight, we conducted a surface-level analysis of these proteins alongside with other telomere maintenance proteins provided by TOGA data, examining their presence or absence of the gene. Our analysis revealed a noteworthy finding regarding RPAIN in comparison to the human reference. This protein triggers the introduction of RPA1, RP2, and RP3 to the PML body, these proteins comprise a protein complex typically localized within the PML-body in ALT, believed to serve as a principal elongation factor *(Loe et al., 2020)*. As well, we can observe a distinction in the correlation between fusions and ASFI1A with a considerably low p-value in both human and mouse references mouse references which indicates a possible correlation with the fusions.

- The project possesses significant challenges. Telomeres present intricate analysis barriers, primarily due to limitations in Next-Generation Sequencing (NGS) techniques, such as the way Illumina divided the DNA by sonication or chemicals, the tag or barcode used, and so on. The samples catalogued in Zoonomia lack of information such as age and tissue type, crucial for analysing telomere size and variability. Despite the pre-selection of Illumina NGS methods, methodological discrepancies may persist. Moreover, the protein analysis conducted via TOGA may suffer from incomplete genomes as some proteins could have been mis-recorded as missing or loss, as well as the limiting insights into protein expression. Furthermore, the binary variables used in this analysis, namely absence and presence, introduce noise and diminish data richness.

- To enhance the analysis, we could refine the categorization of proteins by employing one-hot encoding in Random Forest. Aligning the proteins would help highlight the significance of any missing segments in relation to their functionality, prompting a reassessment of their categorization. Additionally, determining which Illumina methodology contributes to noise generation would enable us to mitigate its impact. We can reuse the program to identify endogenous fusions and remove them from the telomere data. These steps represent potential avenues for improving the analysis.

Conclusions

- I have developed a suite of Python scripts under a general application called TeloFusVarfinder that can be used to identify potential telomeric fusions from FASTQ files.
- TeloFusVarfinder can also be used to analyse telomeric content and telomeric motif variability from raw genome data.
- The application TeloFusVarfinder has a high efficiency due to parallelization.
- I have applied TeloFusVarfinder to analyse data from 179 mammal genomes, using data provided from the Zoonomia project.
- The genome of some Chiroptera seem to be characterized by higher telomere content and higher telomeric fusion frequencies.
- The presence of RPAIN and ASF1A orthologs has maximal negative correlations with the frequency of telomeric fusions and with telomere content across mammals.
- These analysis can guide future research as our work suggest that Chiroptera and Rodentia could have different ways to elongates the telomere, as their fusions seem to differ from the rest of the mammals (figure 14), they have a slightly higher amount of telomere fusions, variability and telomere size, all of those characteristics are associated with ALT. (figure 12, figure 16).
- We would recommend to:
    - Compare the gene expression of APBs associated proteins, such as RPA complex and ASF1A.
    - Telomeric enrichment to detect TFs and TRVs in Chiroptera and Rodentia and compare it with other mammals.
    - Investigate the ALT-PML body telomeric presence via fluorescent technics against other mammals to see if those mammals use them in higher amount than the others.

Data and materials availability.

Here we provide the TeloFusVarfinder material, where you can run the scripts available in

- https://github.com/JonathanRodriguez92/TeloFusVarfinder

All data are available in the Data analysis material and Data extraction, available for download in:

- https://github.com/JonathanRodriguez92/TeloFusVarfinder/tree/e335c6c0fbd2ca46a1dbfcda0bbacfd4bd731d2a/Zoonomia_project/Data_extraction
- https://github.com/JonathanRodriguez92/TeloFusVarfinder/tree/254ad3b610cdfba39d4535871e42e21d8682a7e7/Zoonomia_project/Data_analysis

Notebooks with all the analysis done:

- https://github.com/JonathanRodriguez92/TeloFusVarfinder/blob/2ddf95ca722f3a3b01191cb433a6c997507002e0/Zoonomia_project/Data_analysis/Analysis_notebook.ipynb

References

1. Aix, E., Gallinat, A., Yago-Díez, C., Lucas, J., Gómez, M. J., Benguría, A., Freitag, P., Cortez-Toledo, E., Fernández De Manuel, L., García-Cuasimodo, L., Sánchez-Iranzo, H., Montoya, M. C., Dopazo, A., Sánchez-Cabo, F., Mercader, N., López, J. E., Fleischmann, B. K., Hesse, M., & Flores, I. (2023). Telomeres fuse during cardiomyocyte maturation. Circulation, 147(21), 1634–1636. https://doi.org/10.1161/CIRCULATIONAHA.122.062229
2. Biopython · biopython. (n.d.). Retrieved 15 May 2024, from https://biopython.org/
3. Bize, P., Criscuolo, F., Metcalfe, N. B., Nasir, L., & Monaghan, P. (2009). Telomere dynamics rather than age predict life expectancy in the wild. Proceedings of the Royal Society B: Biological Sciences, 276(1662), 1679–1683. https://doi.org/10.1098/rspb.2008.1817
4. Chung, I., Osterwald, S., Deeg, K. I., & Rippe, K. (2012). PML body meets telomere: The beginning of an ALTernate ending? Nucleus, 3(3), 263–275. https://doi.org/10.4161/nucl.20326
5. Collins, K. (2000). Mammalian telomeres and telomerase. Current Opinion in Cell Biology, 12(3), 378–383. https://doi.org/10.1016/S0955-0674(00)00103-4
6. Corpet, A., Kleijwegt, C., Roubille, S., Juillard, F., Jacquet, K., Texier, P., & Lomonte, P. (2020). PML nuclear bodies and chromatin dynamics: Catch me if you can! Nucleic Acids Research, 48(21), 11890–11912. https://doi.org/10.1093/nar/gkaa828
7. Hoang, S. M., & O'Sullivan, R. J. (2020). Alternative lengthening of telomeres: Building bridges to connect chromosome ends. Trends in Cancer, 6(3), 247–260. https://doi.org/10.1016/j.trecan.2019.12.009
8. Homo_sapiens—Ensembl genome browser 112. (n.d.). Retrieved 15 May 2024, from https://useast.ensembl.org/Homo_sapiens/Gene

9. Illumina | Sequencing and array solutions to fuel genomic discoveries. (n.d.). Retrieved 15 May 2024, from https://www.illumina.com/
10. Kirilenko, B. M., Munegowda, C., Osipova, E., Jebb, D., Sharma, V., Blumer, M., Morales, A. E., Ahmed, A.-W., Kontopoulos, D.-G., Hilgers, L., Lindblad-Toh, K., Karlsson, E. K., Zoonomia Consortium‡, Hiller, M., Andrews, G., Armstrong, J. C., Bianchi, M., Birren, B. W., Bredemeyer, K. R., … Zhang, X. (2023). Integrating gene annotation with orthology inference at scale. Science, 380(6643), eabn3107. https://doi.org/10.1126/science.abn3107
11. Kurihara, N., Tajima, Y., Yamada, T. K., Matsuda, A., & Matsuishi, T. (2017). Description of the karyotypes of Stejneger's beaked whale (Mesoplodon stejnegeri) and Hubbs' beaked whale (M. carlhubbsi). Genetics and Molecular Biology, 40(4), 803–807. https://doi.org/10.1590/1678-4685-gmb-2016-0284
12. Lagunas-Rangel, F. A. (2020). Why do bats live so long?—Possible molecular mechanisms. Biogerontology, 21(1), 1–11. https://doi.org/10.1007/s10522-019-09840-3
13. Lee Kolnicki, R. (n.d.). Mammalian species origin and geographical dispersal patterns correlate with changes in chromosome structure, exemplified in lemurs (Madagascar) and bats(Worldwide). https://doi.org/10.7275/0V47-PE32
14. Lee, M., Teber, E. T., Holmes, O., Nones, K., Patch, A.-M., Dagg, R. A., Lau, L. M. S., Lee, J. H., Napier, C. E., Arthur, J. W., Grimmond, S. M., Hayward, N. K., Johansson, P. A., Mann, G. J., Scolyer, R. A., Wilmott, J. S., Reddel, R. R., Pearson, J. V., Waddell, N., & Pickett, H. A. (2018). Telomere sequence content can be used to determine ALT activity in tumours. Nucleic Acids Research, 46(10), 4903–4918. https://doi.org/10.1093/nar/gky297
15. List of organisms by chromosome count. (2024). In Wikipedia. https://en.wikipedia.org/w/index.php?title=List_of_organisms_by_chromosome_count&oldid=1222163272
16. Loe, T. K., Li, J. S. Z., Zhang, Y., Azeroglu, B., Boddy, M. N., & Denchi, E. L. (2020). Telomere length heterogeneity in ALT cells is maintained by PML-dependent localization of the BTR complex to telomeres. Genes & Development, 34(9–10), 650–662. https://doi.org/10.1101/gad.333963.119
17. Malkova, A., & Ira, G. (2013). Break-induced replication: Functions and molecular mechanism. Current Opinion in Genetics & Development, 23(3), 271–279. https://doi.org/10.1016/j.gde.2013.05.007
18. Matplotlib—Visualization with python. (n.d.). Retrieved 15 May 2024, from https://matplotlib.org/
19. Muyas, F., Rodriguez, M. J. G., Cascão, R., Afonso, A., Sauer, C. M., Faria, C. C., Cortés-Ciriano, I., & Flores, I. (2024). The ALT pathway generates telomere fusions that can be detected in the blood of cancer patients. Nature Communications, 15(1), 82. https://doi.org/10.1038/s41467-023-44287-8
20. National center for biotechnology information. (n.d.). Retrieved 15 May 2024, from https://www.ncbi.nlm.nih.gov/

21. Neumann, A. A., Watson, C. M., Noble, J. R., Pickett, H. A., Tam, P. P. L., & Reddel, R. R. (2013). Alternative lengthening of telomeres in normal mammalian somatic cells. Genes & Development, 27(1), 18–23. https://doi.org/10.1101/gad.205062.112
22. Reactome | pathway browser. (n.d.). Retrieved 15 May 2024, from https://reactome.org/PathwayBrowser/#/R-HSA-157579&SEL=R-HSA-180786&PATH=R-HSA-1640170,R-HSA-73886&FLG=P46100&DTAB=DT
23. Rumpler, Y., Hauwy, M., Fausser, J.-L., Roos, C., Zaramody, A., Andriaholinirina, N., & Zinner, D. (2011). Comparing chromosomal and mitochondrial phylogenies of the Indriidae (Primates, lemuriformes). Chromosome Research, 19(2), 209–224. https://doi.org/10.1007/s10577-011-9188-5
24. Sfeir, A. J., Chai, W., Shay, J. W., & Wright, W. E. (2005). Telomere-end processing. Molecular Cell, 18(1), 131–138. https://doi.org/10.1016/j.molcel.2005.02.035
25. Shen, M., Young, A., & Autexier, C. (2021). PCNA, a focus on replication stress and the alternative lengthening of telomeres pathway. DNA Repair, 100, 103055. https://doi.org/10.1016/j.dnarep.2021.103055
26. Sotero-Caio, C., Baker, R., & Volleth, M. (2017). Chromosomal evolution in chiroptera. Genes, 8(10), 272. https://doi.org/10.3390/genes8100272
27. Srinivas, N., Rachakonda, S., & Kumar, R. (2020). Telomeres and telomere length: A general overview. Cancers, 12(3), 558. https://doi.org/10.3390/cancers12030558
28. Stack overflow—Where developers learn, share, & build careers. (n.d.). Stack Overflow. Retrieved 15 May 2024, from https://stackoverflow.com/
29. The python standard library. (n.d.). Python Documentation. Retrieved 15 May 2024, from https://docs.python.org/3/library/index.html
30. Uniprot. (n.d.). Retrieved 14 May 2024, from https://www.uniprot.org/uniprotkb/
31. Wang, X., & Baumann, P. (2008). Chromosome fusions following telomere loss are mediated by single-strand annealing. Molecular Cell, 31(4), 463–473. https://doi.org/10.1016/j.molcel.2008.05.028
32. Zoonomia. (n.d.). Retrieved 15 May 2024, from https://zoonomiaproject.org/
33. Zoonomia Consortium. (2020). A comparative genomics multitool for scientific discovery and conservation. Nature, 587(7833), 240–245. https://doi.org/10.1038/s41586-020-2876-6
34. Nucleic Acids Res, Volume 48, Issue 21, 2 December 2020, Pages 11890–11912, https://doi.org/10.1093/nar/gkaa828.

Appendix

zoonomia_sp_info.xlsx

Species	Order	Family	NCBI_AssemblyAccNo	Common_name	Accession_number	Genome_coverage
Balaenoptera acutorostrata	Cetartiodactyla	Balaenopteridae	GCF_000493695.1		GCF_000493695	92.0
Balaenoptera bonaerensis	Cetartiodactyla	Balaenopteridae	GCA_000978805.1	Antarctic minke whale	GCA_000978805	60.0
Beatragus hunteri	Cetartiodactyla	Bovidae	GCA_004027495.1		GCA_004027495	23.1
Bison bison	Cetartiodactyla	Bovidae	GCF_000754665.1		GCF_000754665	60.0
Bos indicus	Cetartiodactyla	Bovidae	GCA_000247795.2	zebu cattle	GCA_000247795	52.0
Bos mutus	Cetartiodactyla	Bovidae	GCF_000298355.1	wild yak	GCF_000298355	130.0
Bos taurus	Cetartiodactyla	Bovidae	GCF_000003205.7	cattle	GCF_000003205	19.0
Bradypus variegatus	Pilosa	Bradypodidae	GCA_004027775.1	Brown-throated sloth	GCA_004027775	57.9
Bubalus bubalis	Cetartiodactyla	Bovidae	GCF_000471725.1	water buffalo	GCF_000471725	70.0
Callicebus donacophilus	Primates	Pitheciidae	GCA_004027715.1	Bolivian titi	GCA_004027715	23.7
Callithrix jacchus	Primates	Cebidae	GCA_002754865.1	white-tufted-ear marmoset	GCA_002754865	60.0
Camelus bactrianus	Cetartiodactyla	Camelidae	GCF_000767855.1	Bactrian camel	GCF_000767855	79.2
Camelus dromedarius	Cetartiodactyla	Camelidae	GCF_000767585.1	Arabian camel	GCF_000767585	65.0
Camelus ferus	Cetartiodactyla	Camelidae	GCF_000311805.1	Wild Bactrian camel	GCF_000311805	30.0
Canis lupus familiaris	Carnivora	Canidae	GCF_000002285.3	dog	GCF_000002285	
Capra aegagrus	Cetartiodactyla	Bovidae	GCA_000978405.1	wild goat	GCA_000978405	83.5
Capra hircus	Cetartiodactyla	Bovidae	GCF_001704415.1	goat	GCF_001704415	50.0
Capromys pilorides	Rodentia	Capromyidae	GCA_004027915.1	Desmarest's hutia	GCA_004027915	26.1
Carollia perspicillata	Chiroptera	Phyllostomidae	GCA_004027735.1	Seba's short-tailed bat	GCA_004027735	27.7
Castor canadensis	Rodentia	Castoridae	GCA_004027675.1	American beaver	GCA_004027675	30.4
Catagonus wagneri	Cetartiodactyla	Tayassuidae	GCA_004024745.1	Chacoan peccary	GCA_004024745	87.1
Cavia aperea	Rodentia	Caviidae	GCA_000688575.1	Brazilian guinea pig	GCA_000688575	333.0
Cavia porcellus	Rodentia	Caviidae	GCF_000151735.1	domestic guinea pig	GCF_000151735	6.8
Cavia tschudii	Rodentia	Caviidae	GCA_004027695.1	Montane guinea pig	GCA_004027695	34.8
Cebus albifrons	Primates	Cebidae	GCA_004027755.1	white-fronted capuchin	GCA_004027755	28.8
Cebus capucinus	Primates	Cebidae	GCF_001604975.1	Panamanian white-faced capuchin	GCF_001604975	81.0
Ceratotherium simum	Perissodactyla	Rhinocerotidae	GCF_000283155.1	southern white rhinoceros	GCF_000283155	91.0

Table 1. all mammals information of Zoonomia obtained from https://zoonomiaproject.org/the-mammal-tree-list-view/, and www.ncbi.nlm.nih.gov . See Extraction_zoonomia_information.py. see main scripts

Combined_distances_{specie}_1.csv (example)

ID	Distance	FusionStart	FusionEnd	FusionType	TotalScore
SRR893003.9925087	5	8	49	Outward	360
SRR893003.15745389	19	5	59	Inward	350
SRR893003.40504459	0	11	46	Inward	350
SRR893003.43266811	0	1	36	Inward	350
SRR893003.54579934	0	3	38	Outward	350
SRR893003.73489562	0	20	55	Outward	350
SRR893003.77193141	0	4	39	Inward	350
SRR893003.80935996	15	14	65	Outward	360
SRR893003.84132736	1	10	47	Outward	350
SRR893003.105918223	20	9	66	Outward	350
SRR893003.109556003	15	36	87	Outward	350
SRR893003.120503923	0	0	35	Outward	350
SRR893003.127079658	11	0	47	Outward	350
SRR893003.130492534	0	13	49	Outward	360
SRR893003.132359828	0	24	60	Outward	360
SRR893003.151118460	15	33	84	Outward	360
SRR893003.159798097	12	0	48	Inward	350
SRR893003.164121902	0	5	41	Inward	360
SRR893003.168549440	0	6	42	Outward	350
SRR893003.189940690	8	15	59	Outward	360
SRR893003.202776462	15	35	86	Outward	360
SRR893003.212138703	20	4	60	Inward	350
SRR893003.212408085	0	7	43	Outward	360
SRR893003.220851990	0	0	35	Inward	350
SRR893003.222352722	0	44	80	Inward	350
SRR893003.232665601	16	1	53	Inward	350
SRR893003.241252973	9	0	45	Inward	350
SRR893003.248952060	14	10	60	Outward	360
SRR893003.262004641	0	12	48	Outward	360
SRR893003.263709189	2	15	52	Outward	350
SRR893003.283060740	0	11	47	Inward	360
SRR893003.288307748	11	5	52	Inward	350
SRR893003.290575415	2	15	50	Outward	350
SRR893003.306137651	2	11	49	Outward	360
SRR893003.317422252	15	14	65	Outward	360
SRR893003.343157829	5	8	49	Outward	360
SRR893003.343277031	12	11	59	Inward	350
SRR893003.346509218	2	37	76	Outward	350
SRR893003.347310259	15	27	78	Outward	360
SRR893003.348640213	13	46	95	Outward	350

Table 2 Results obtained from TeloFusProcessor.py. The analysis is a compilation of the fusions presented in the FastQ dataset. Distance represents the distance between the two parts of the fusion. And the score represents how accurate is the fusion.

Telo_fus_{specie}_1.fasta (example)

```
>SRR7611052.27595685 Outward 15-51 TotalLength=250 bp TotalScore=36.0 fusion AACCCTAACCCTAACCCTGGTTAGGGTTAGGGTTAG distance = 0
CCTAACCCTAACCCTAACCCTAACCCTAACCCTGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGATAGGGAGAGCGTGAGGGA
TCGGGAGCGAGTTAGAGATCTCGTATGCGTGCGGCTGCGTGGTAGAGTAAGCGGTCGTCT
GAGGGATAGC
>SRR7611052.37621043 Outward 150-205 TotalLength=250 bp TotalScore=35.0 fusion CTAACCCTAACCCTAACCTTAACCCTAGCCCTAGATCTAGGGTTAAGGTTAGGGT distance = 19
CCTAAACCTAACCCTAACTCTAACACTAGCCCTAGCCCTAGCTCTAGGCCTAAACTAGCC
CTAAACCTATCCCTATCCCTATCCTAGCCCTAGCCCTAGCCCTAACCCTAACCCTAACCC
TAACCCCTAACCCTAAATCTAAAACTAAGCCTAACCCTAACCCTAACCTTAACCCTAGCC
CTAGATCTAGGGTTAAGGTTAGGGTTAGGGTTAGGCTTAGTTTTAGATTTAGGGTTAGAT
CGGAAGAGCA
>SRR7611052.45339309 Inward 149-201 TotalLength=250 bp TotalScore=35.0 fusion TTAGGGTTAGGGTTAGGGCTAACCCTAACCTAAACCTAACACTAACCCTAAC distance = 16
ACTCAGGGTTAGAGTTAAGGTTAGGCTTAGTGTTAGGGTTAGGGTTGTTAGGGTTAGAGT
TAGGGTTAGGGTTAGGTTTAGGGATAGGGTTAGGGTTATGGTTCAGGTTAGGGTTAGTGT
TAGGTTTAGGTTAGGGTTAGGACTCAGTATTAGGGTTAGGGTTAGGGCTAACCCTAACCT
AAACCTAACACTAACCCTAACCTGAACCATAACCCTAACCCTAGATCGGAAGAGCACACG
TCTGAACTCC
>SRR7611052.55111343 Inward 94-130 TotalLength=250 bp TotalScore=36.0 fusion TTAGGGTTAGGGTTAGGGTAACCCTAACCCTAACCC distance = 0
AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTT
AGGGGTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTCACCATACACCAA
ACCCTAACCC
>SRR7611052.56535162 Inward 45-81 TotalLength=250 bp TotalScore=35.0 fusion TAGGGTTAGGGTTAGGGTTAGCCCTAACCCTAACCC distance = 0
ATCTTGCCTCCTTAGCCTAGGACATTTTATCTCTTGCCTTCTGCTTAGGGTTAGGGTTAG
GGTTAGCCCTAACCCTAACCCTGATCTGGACACTTAGGCAGGAAGCTGGAGCCACCCCAG
AGATCGGTCCCCTGCCCCCACTTCCGTCTCTCAGAAAACTGTACTGAGACACAGACAGCT
GCACCCTGATGCCTATTTCTTTAAAGCTGTGTTACAGGTGTTTTCTGCTCTCTAATTTAA
TTTACATTTC
>SRR7611052.71274822 Outward 191-245 TotalLength=250 bp TotalScore=35.0 fusion CTAACCCTAACCCTAACCATAACCCTACCCCTATATGGATAGGGTTAGGGTTAG distance = 18
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACTATAACCCTAACCCTAACCCTAACTCTAACCCTAACCC
TAACCCTCCCCCTAACCCTAACCCTAACCATAACCCTACCCCTATATGGATAGGGTTAGG
GTTAGGGTTA
```

Figure S 1: Results obtained from TeloFusProcessor a fasta file with all the fusions present in the fastQ file.

Fusion_frequency_{direction}.xlsx (example)

Column1	GTTAGGGTTAGGCTAACCCTAAC	GTTAGGGTTAGGACCCTAACCCT	GTTAGGGTTAGGCCTAACCCTAA	GTTAGGGTTAGGCCCTAACCCTA	GTTAGGGTTAGGTAACCCTAACC
Acomys_cahirinus_1	0	0	0	0	0
Acomys_cahirinus_2	0	0	0	0	0
Ailurus_fulgens_1	0	0	0	0	0
Allactaga_bullata_1	0	0	0	1	0
Ammotragus_lervia_1	0	0	0	0	0
Ammotragus_lervia_2	0	0	0	0	0
Antrozous_pallidus_1	0	0	0	0	0
Antrozous_pallidus_2	0	0	0	0	0
Artibeus_jamaicensis_1	0	0	0	0	0
Artibeus_jamaicensis_2	0	0	0	0	0
Ateles_geoffroyi_1	0	0	3	0	0
Ateles_geoffroyi_2	0	0	2	0	0
Balaenoptera_acutorostrata_2	0	0	0	0	0
Beatragus_hunteri_1	0	0	0	0	0
Bradypus_variegatus_1	0	0	0	0	0
Bradypus_variegatus_2	0	0	0	0	0
Callithrix_jacchus_1	0	0	0	0	0

Table 3: A table where we can see the types of fusion that exists. This is important to analyse the types of breakpoints where the fusion occurred. This table was generated via the script Fasta_fusion_analysis.py. The values are the frequency, the first column the name of the file, and the rest of the columns the types of fusions. At the end appears the Order of each species. please see data_extraction

Telo_mv_{specie}_1.csv (example)

	A	B	C	D	E	F	G
1	Column	ID	TelomereMoti	Varian	Frequenc	PorcentageTelomereSequenc	Lengt
2	0	SRR7611052.2012	TTAGGG		40	960	250
3	1			TAAGGG	1	24	250
4	2	SRR7611052.3796	CCCTAA		37	888	250
5	3			CCCTCA	2	48	250
6	4			CCCAAA	1	24	250
7	5	SRR7611052.5752	CCCTAA		17	408	250
8	8			ACCTAA	5	120	250
9	12			CACTAA	4	96	250
10	7			TCCTAA	2	48	250
11	6			CCCTGA	1	24	250
12	9			CTCTAA	1	24	250
13	10			CCCTTA	1	24	250
14	11			GCCTAA	1	24	250
15	13			CCCTCA	1	24	250
16	14			CCCTAC	1	24	250
17	15			CCATAA	1	24	250
18	16	SRR7611052.8820	TTAGGG		18	432	250
19	22			TTAGGA	4	96	250
20	17			ATAGGG	3	72	250
21	19			TTAGGC	3	72	250
22	21			TTATGG	2	48	250
23	25			TTAGTG	2	48	250
24	18			TAAGGG	1	24	250
25	20			TTAGCG	1	24	250
26	23			TTGGGG	1	24	250
27	24			TCAGGG	1	24	250

Table 4 This table represents the frequency in each read of the motif and the variant with only one mismatch. This table was obtained from TelomericVariantSearcher. please see data_extraction

Telomere_variants_catalogue.xlsx (example)

Variant	Acomys_cahirinus_	Acomys_cahirinus_	Ailurus_fulgens_	Ailurus_fulgens_
TTAGGG	2953906	287426	78918	56692
TTAGGA	64222	20930	68	40
TAAGGG	27094	5018	64	56
ATAGGG	47102	4978	170	208
TCAGGG	39842	10610	90	78
CTAGGG	26950	1940	506	280
TGAGGG	77340	6812	180	162
TTAAGG	24662	8360	122	52
TTACGG	11906	4190	100	50
TTAGAG	23482	6350	162	100
TTAGCG	12556	3812	80	60
GTAGGG	97594	10642	2404	2206
TTAGGC	37058	12670	294	234
TTAGGT	90860	28806	204	178
TTAGTG	66352	19476	114	102
TTATGG	43286	11690	148	74
TTCGGG	16658	2614	182	218
TTGGGG	88862	12816	358	432
TTTGGG	66756	12040	172	262
Total_telomere_reads	52459	7099	2900	2156
Total_spots_run	143851292	143851292	149123819	149123819
Total_bases_run	71925646000	71925646000	30123011438	30123011438
Genome_size	2306070819	2306070819	2342939051	2342939051
Porcentage_TTAGGG	0,77	0,61	0,94	0,92
Porcentage_variability	0,23	0,39	0,06	0,08

Table 5. A better look of the variability where all CCCTAA and variants Ire added to the TTAGGG and variants values. In addition, there are extra calculations done, Porcentage_TTAGGG and Porcentage_variability. The table was obtained from Motif_variant_analysis.py.

Fusion_results_raw.xlsx (example)

Specie_name	Inward (Distance C	Outward (Distance C	Inward (Other	Outward (Other	Total_fusion	Total_fusions_	Total_inwar	Total_outwar	Genome_coverag
Acomys_cahirinus_1	2	1	3	2	8	3	5	3	15,59
Acomys_cahirinus_2	3	6	4	2	15	9	7	8	15,59
Ailurus_fulgens_1	0	0	0	0	0	0	0	0	6,43
Ailurus_fulgens_2	0	0	3	0	3	0	3	0	6,43
Allactaga_bullata_1	2	0	1	1	4	2	3	1	12,58
Allactaga_bullata_2	0	0	1	0	1	0	1	0	12,58
Alouatta_palliata_mexicana_1	0	0	0	6	6	0	0	6	14,93
Alouatta_palliata_mexicana_2	0	0	0	7	7	0	0	7	14,93
Ammotragus_lervia_1	0	0	0	0	0	0	0	0	14,41
Ammotragus_lervia_2	0	0	0	0	0	0	0	0	14,41
Anoura_caudifer_1	0	5	1	2	8	5	1	7	24,97
Anoura_caudifer_2	1	5	2	4	12	6	3	9	24,97
Antilocapra_americana_1	0	3	0	0	3	3	0	3	20,51
Antilocapra_americana_2	0	2	0	0	2	2	0	2	20,51
Antrozous_pallidus_1	1	6	62	26	95	7	63	32	18,12
Antrozous_pallidus_2	1	6	75	31	113	7	76	37	18,12
Aotus_nancymaae_1	0	0	1	0	1	0	1	0	9,11
Aotus_nancymaae_2	0	0	1	0	1	0	1	0	9,11
Aplodontia_rufa_1	0	0	1	1	2	0	1	1	13,19
Aplodontia_rufa_2	0	1	2	2	5	1	2	3	13,19
Artibeus_jamaicensis_1	1	1	21	7	30	2	22	8	13,52
Artibeus_jamaicensis_2	5	2	56	15	78	7	61	17	13,52
Ateles_geoffroyi_1	11	0	10	0	21	11	21	0	22,01
Ateles_geoffroyi_2	9	1	11	2	23	10	20	3	22,01

Table 6. A table with all the fusions and the genome coverage calculated previously in each sample. Table obtained from Fusion_analysis.py.

Fusion_results_coverage.xlsx (example)

Specie_name	Inward (Distance C	Outward (Distance C	Inward (Other	Outward (Other	Total_fusion	Total_fusions_	Total_inwar	Total_outwar	Genome_coverag
Acomys_cahirinus_1	12	6	19	12	49	18	31	18	15,59
Acomys_cahirinus_2	19	38	25	12	94	57	44	50	15,59
Ailurus_fulgens_1	0	0	0	0	0	0	0	0	6,43
Ailurus_fulgens_2	0	0	46	0	46	0	46	0	6,43
Allactaga_bullata_1	15	0	7	7	29	15	22	7	12,58
Allactaga_bullata_2	0	0	7	0	7	0	7	0	12,58
Alouatta_palliata_mexicana_1	0	0	0	40	40	0	0	40	14,93
Alouatta_palliata_mexicana_2	0	0	0	46	46	0	0	46	14,93
Ammotragus_lervia_1	0	0	0	0	0	0	0	0	14,41
Ammotragus_lervia_2	0	0	0	0	0	0	0	0	14,41
Anoura_caudifer_1	0	20	4	8	32	20	4	28	24,97
Anoura_caudifer_2	4	20	8	16	48	24	12	36	24,97
Antilocapra_americana_1	0	14	0	0	14	14	0	14	20,51
Antilocapra_americana_2	0	9	0	0	9	9	0	9	20,51
Antrozous_pallidus_1	5	33	342	143	523	38	347	176	18,12
Antrozous_pallidus_2	5	33	413	171	622	38	418	204	18,12
Aotus_nancymaae_1	0	0	10	0	10	0	10	0	9,11
Aotus_nancymaae_2	0	0	10	0	10	0	10	0	9,11
Aplodontia_rufa_1	0	0	7	7	14	0	7	7	13,19
Aplodontia_rufa_2	0	7	15	15	37	7	15	22	13,19
Artibeus_jamaicensis_1	7	7	155	51	220	14	162	58	13,52
Artibeus_jamaicensis_2	36	14	414	110	574	50	450	124	13,52
Ateles_geoffroyi_1	49	0	45	0	94	49	94	0	22,01
Ateles_geoffroyi_2	40	4	49	9	102	44	89	13	22,01
Balaenoptera_acutorostrata_1	0	27	41	27	95	27	41	54	14,49
Balaenoptera_acutorostrata_2	48	55	48	124	275	103	96	179	14,49

Table 7. A table with all fusions normalized with by the genome coverage multiply per 100 and the genome coverage calculated previously. Table obtained from Fusion_analysis.py. please see data_extraction

Genome Coverage

$$Genomecoverage = \frac{Totalreadsperrun * bpread}{genome}$$

*Equation 1. Genome coverage was calculated by multiplying total reads per run (SRR***) per the average bp in a read divided by the total bp genome size.*

{specie}_loss_summ_data.tsv (example)

Archivo Editar Ver

GENE	ENSG00000272525	I
GENE	ENSG00000117748	I
GENE	ENSG00000174600	I
GENE	ENSG00000186230	I
GENE	ENSG00000169714	I
GENE	ENSG00000166736	I
GENE	ENSG00000118058	PI
GENE	ENSG00000189030	L
GENE	ENSG00000197142	PI
GENE	ENSG00000178773	UL
GENE	ENSG00000138028	PM
GENE	ENSG00000136488	I
GENE	ENSG00000095485	I
GENE	ENSG00000137868	PI
GENE	ENSG00000070495	I
GENE	ENSG00000132341	I
GENE	ENSG00000102910	PI
GENE	ENSG00000055163	PI
GENE	ENSG00000160050	I
GENE	ENSG00000186562	L
GENE	ENSG00000179456	I

Figure S 2 Data collected from TOGA Ibsite, Zoonomia. https://genome.senckenberg.de/download/TOGA/human_hg38_reference/ first column correspond to Gene, transcription or projection, second column correspond to gene, transcription or projection ID and third correspond to the status of the ID in the genome obtained from the alignment in the Zoonomia project where the status could be I(intact), PI(partially intact), PG(paralogous projection), U (uncertain loss), L (loss), PM(partially missing) and M (missing). This table was obtained via the script Extraction_gene_to_gene_annotation.py.

{specie}_OrthologsClassification.tsv (example)

Archivo Editar Ver

t_gene	t_transcript	q_gene	q_transcript	orthology_class
ENSG00000129484	ENST00000429687.PARP2	reg_12621	ENST00000429687.PARP2.3769	one2one
ENSG00000129484	ENST00000527915.PARP2	reg_12621	ENST00000527915.PARP2.3769	one2one
ENSG00000129484	ENST00000250416.PARP2	reg_12621	ENST00000250416.PARP2.3769	one2one
ENSG00000089693	ENST00000539187.MLF2	reg_16214	ENST00000539187.MLF2.16730	one2one
ENSG00000121653	ENST00000395629.MAPK8IP1	reg_5548	ENST00000395629.MAPK8IP1.1571	one2one
ENSG00000121653	ENST00000241014.MAPK8IP1	reg_5548	ENST00000241014.MAPK8IP1.1571	one2one
ENSG00000140553	ENST00000394275.UNC45A	reg_6096	ENST00000394275.UNC45A.10224	one2one
ENSG00000140553	ENST00000418476.UNC45A	reg_6096	ENST00000418476.UNC45A.10224	one2one
ENSG00000140553	ENST00000639885.UNC45A	reg_6096	ENST00000639885.UNC45A.10224	one2one
ENSG00000108602	ENST00000494157.ALDH3A1	reg_4067	ENST00000494157.ALDH3A1.40566	one2one
ENSG00000108602	ENST00000395555.ALDH3A1	reg_4067	ENST00000395555.ALDH3A1.40566	one2one
ENSG00000108602	ENST00000225740.ALDH3A1	reg_4067	ENST00000225740.ALDH3A1.40566	one2one
ENSG00000173575	ENST00000626874.CHD2	reg_15886	ENST00000626874.CHD2.-1	one2one
ENSG00000173575	ENST00000394196.CHD2	reg_15886	ENST00000394196.CHD2.-1	one2one
ENSG00000155016	ENST00000332884.CYP2U1	reg_14780	ENST00000332884.CYP2U1.27468	one2one
ENSG00000196341	ENST00000641897.OR8D1	reg_759	ENST00000641897.OR8D1.103824	many2many
ENSG00000196341	ENST00000641897.OR8D1	reg_1819	ENST00000641897.OR8D1.114493	many2many
ENSG00000196341	ENST00000641897.OR8D1	reg_2737	ENST00000641897.OR8D1.36008	many2many
ENSG00000196341	ENST00000641897.OR8D1	reg_2379	ENST00000641897.OR8D1.104389	many2many
ENSG00000196341	ENST00000641897.OR8D1	reg_2816	ENST00000641897.OR8D1.105097	many2many
ENSG00000196341	ENST00000641897.OR8D1	reg_10212	ENST00000641897.OR8D1.105857	many2many

Figure S 3 Data collected from TOGA Ibsite, Zoonomia. https://genome.senckenberg.de/download/TOGA/human_hg38_reference/ this is organised by t_gene (reference gene ID) t_transcript (reference transcript isoforms ID) q_gene (query gene ID, region), q_transcript(query transcript ID) orthology_class which can be one2one (1 gene to 1 region) one2many(1 gene to many regions) many2many(many genes to many regions) one2zero (1 gene to zero regions). This table was obtained via the script Extraction_gene_annotation.py.

Zoonomia_all_proteins_class_human.csv (example).

	A	B	C	D	E	F
1	Species	ATXN3_g_ENSG00000123594_huma	ZNF10_g_ENSG00000197020_huma	ZNF10_g_ENSG00000103994_huma	ZNF10_g_ENSG00000196247_huma	ZNF10_g_ENSG00000181896_huma
2	Helogale parvula	L	I	I	L	M
3	Anoura caudifer	L	I	I	L	I
4	Hyaena hyaena	L	I	I	I	I
5	Artibeus jamaicensis	PG	I	PI	L	I
6	Nycticeius humeralis	PG	I	PI	PM	I
7	Carollia perspicillata	PG	I	I	PM	I
8	Macrotus californicus	PG	L	I	PM	I
9	Craseonycteris thonglongyai	PG	I	I	UL	I
10	Rhinolophus ferrumequinum	L	I	I	I	I
11	Hipposideros galeritus	PG	I	I	UL	I
12	Mellivora capensis	PG	I	I	UL	I
13	Lasiurus borealis	PG	I	I	PM	I
14	Mirounga angustirostris	L	I	I	I	UL
15	Macroglossus sobrinus	L	I	I	UL	I
16	Mungos mungo	L	I	I	L	I
17	Megaderma lyra	L	I	I	I	I
18	Panthera onca	PG	I	I	I	I
19	Micronycteris hirsuta	PG	I	I	PM	I
20	Chrysochloris asiatica	L	UL	I	L	I
21	Miniopterus schreibersii	PG	I	I	L	I
22	Pteronura brasiliensis	PG	I	I	I	I
23	Mormoops blainvillei	PG	I	I	L	I
24	Puma concolor	L	I	I	UL	I
25	Myotis myotis	M	I	I	L	I
26	Spilogale gracilis	PG	I	I	UL	I
27	Noctilio leporinus	PG	I	I	L	I

Table 8. An example of the file where all the data from {specie}_OrthologsClassification.tsv and {specie}_loss_summ_data.tsv was stored and organized in a table. Columns represent proteins named by its ID {protein}_g_{geneID}_human or mouse depending on the reference organism. There are two versions of this table, one done with specific proteins premediated selected and this one which has around 19000 proteins if the reference is human and 21000 proteins if the reference is muss musculus. This table was obtained via the script All_proteins_class.py or Specifi_proteins_class.py.

zoonomia_analysis_data_boolean_coverage_human.xlsx (part1)

Species	Inward (Distance C	Outward (Distance C	Inward (Othe	Outward (Othe	Total_fusior	Total_fusions_	Total_inwar	Total_outwar	Genome_coverag	Order	Instrument_moc
Acomys cahirinus	12	6	19	12	49	18	31	18	15,59	Rodentia	Illumina HiSeq 2500
Acomys cahirinus	19	38	25	12	94	57	44	50	15,59	Rodentia	Illumina HiSeq 2500
Ailurus fulgens	0	0	0	0	0	0	0	0	6,43	Carnivora	Illumina HiSeq 2000
Ailurus fulgens	0	0	46	0	46	0	46	0	6,43	Carnivora	Illumina HiSeq 2000
Allactaga bullata	15	0	7	7	29	15	22	7	12,58	Rodentia	Illumina HiSeq 2500
Allactaga bullata	0	0	7	0	7	0	7	0	12,58	Rodentia	Illumina HiSeq 2500
Alouatta palliata mexicana	0	0	0	40	40	0	0	40	14,93	Primates	Illumina HiSeq 2500
Alouatta palliata mexicana	0	0	0	46	46	0	0	46	14,93	Primates	Illumina HiSeq 2500
Ammotragus lervia	0	0	0	0	0	0	0	0	14,41	Cetartiodactyla	Illumina HiSeq 4000
Ammotragus lervia	0	0	0	0	0	0	0	0	14,41	Cetartiodactyla	Illumina HiSeq 4000
Anoura caudifer	0	20	4	8	32	20	4	28	24,97	Chiroptera	Illumina HiSeq 2500
Anoura caudifer	4	20	8	16	48	24	12	36	24,97	Chiroptera	Illumina HiSeq 2500
Antilocapra americana	0	14	0	0	14	14	0	14	20,51	Cetartiodactyla	Illumina HiSeq 2500
Antilocapra americana	0	9	0	0	9	9	0	9	20,51	Cetartiodactyla	Illumina HiSeq 2500
Antrozous pallidus	5	33	342	143	523	38	347	176	18,12	Chiroptera	Illumina HiSeq 2500
Antrozous pallidus	5	33	413	171	622	38	418	204	18,12	Chiroptera	Illumina HiSeq 2500
Aotus nancymaae	0	0	10	0	10	0	10	0	9,11	Primates	Illumina HiSeq 2000
Aotus nancymaae	0	0	10	0	10	0	10	0	9,11	Primates	Illumina HiSeq 2000
Aplodontia rufa	0	0	7	7	14	0	7	7	13,19	Rodentia	Illumina HiSeq 2500
Aplodontia rufa	0	7	15	15	37	7	15	22	13,19	Rodentia	Illumina HiSeq 2500
Artibeus jamaicensis	7	7	155	51	220	14	162	58	13,52	Chiroptera	Illumina HiSeq 2500
Artibeus jamaicensis	36	14	414	110	574	50	450	124	13,52	Chiroptera	Illumina HiSeq 2500
Ateles geoffroyi	49	0	45	0	94	49	94	0	22,01	Primates	Illumina HiSeq 2500
Ateles geoffroyi	40	4	49	9	102	44	89	13	22,01	Primates	Illumina HiSeq 2500
Balaenoptera acutorostrata	0	27	41	27	95	27	41	54	14,49	Cetartiodactyla	Illumina HiSeq 2000

Table 9.A. This table represents the unification of Fusion_results_coverage.xlsx, Telomere_variants_catalogue.xlsx, zoonomia_sp_info.xlsx, SupplementaryData.xls and Zoonomia_all_proteins_class_human.csv created to analyse the data. This part represents fusion results normalize by genome coverage and part of Instrument model and Order obtain from zoonomia_sp_info.xlsx. This table was obtained via the script FusionVariantProteinInfo.py. please see data_extraction

zoonomia_analysis_data_boolean_coverage_human.xlsx (part2)

Instrument_model	Average_bp_rea	Longevit	Chrom_nur	ATRX_g_ENSG00000085224__huma	DAXX_g_ENSG00000204209__huma	TERT_g_ENSG00000164362__huma	RTEL1_g_ENSG00000258366__huma
Illumina HiSeq 2500	250	4	38	TRUE	TRUE	TRUE	FALSE
Illumina HiSeq 2500	250	4	38	TRUE	TRUE	TRUE	FALSE
Illumina HiSeq 2000	101	14	36	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 2000	101	14	36	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 2500	251			TRUE	TRUE	FALSE	TRUE
Illumina HiSeq 2500	251			TRUE	TRUE	FALSE	TRUE
Illumina HiSeq 2500	250	7	54,53				
Illumina HiSeq 2500	250	7	54,53				
Illumina HiSeq 4000	100,45	20,9	58	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 4000	100,45	20,9	58	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 2500	250	10	30	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 2500	250	10	30	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 2500	250	16	58	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 2500	250	16	58	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 2500	250	9	46	TRUE	TRUE	FALSE	TRUE
Illumina HiSeq 2500	250	9	46	TRUE	TRUE	FALSE	TRUE
Illumina HiSeq 2000	101		54	TRUE	TRUE	TRUE	FALSE
Illumina HiSeq 2000	101		54	TRUE	TRUE	TRUE	FALSE
Illumina HiSeq 2500	250	6	46	FALSE	TRUE	TRUE	TRUE
Illumina HiSeq 2500	250	6	46	FALSE	TRUE	TRUE	TRUE
Illumina HiSeq 2500	251	7	30	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 2500	251	7	30	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 2500	250	47,1	34	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 2500	250	47,1	34	TRUE	TRUE	TRUE	TRUE
Illumina HiSeq 2000	100	45	42,44	FALSE	TRUE	FALSE	TRUE
Illumina HiSeq 2000	100	45	42,44	FALSE	TRUE	FALSE	TRUE

Table 9.B. This part represents the unification of some information from zoonomia_sp_info.xlsx and Zoonomia_all_proteins_class_human.csv where the categories where changed. I and PI Ire changed to True, this means presence of the gene and PG, U, M, PL and L Ire changed to False which means the absence of the gene. This table was obtained via the script FusionVariantProteinInfo.py.

zoonomia_analysis_data_boolean_coverage_human.xlsx (part3)

RAD51_g_ENSG00000051180__huma	CMR	ICM	Total_telomere_read	Genome_siz	Porcentage_TTAGG	Porcentage_variabilit	Total_telomere_siz	percentage_telomere_siz
TRUE			52459	2306070819	0,77	0,23	841228,3515	0,000364789
TRUE			7099	2306070819	0,61	0,39	113838,9994	4,93649E-05
TRUE	0,070707071	0,0733606	2900	2342939051	0,94	0,06	45552,09953	1,94423E-05
TRUE	0,070707071	0,0733606	2156	2342939051	0,92	0,08	33865,62986	1,44543E-05
TRUE			12924	3093575781	0,85	0,15	257863,593	8,33545E-05
TRUE			266	3093575781	0,79	0,21	5307,313196	1,71559E-06
			2111	3033617595	0,83	0,17	35348,29203	1,16522E-05
			163	3033617595	0,8	0,2	2729,403885	8,99719E-07
TRUE	0,040322581	0,085277433	29876	2650836958	0,95	0,05	208261,2214	7,85643E-05
TRUE	0,040322581	0,085277433	28440	2650836958	0,92	0,08	198251,0756	7,47881E-05
TRUE			13537	2206589520	0,78	0,22	135532,6392	6,14218E-05
TRUE			2290	2206589520	0,68	0,32	22927,51302	1,03905E-05
TRUE	0,03125	0,03792695	20645	2899768642	0,84	0,16	251645,5388	8,67812E-05
TRUE	0,03125	0,03792695	1034	2899768642	0,8	0,2	12603,608	4,34642E-06
TRUE			88154	2597469403	0,64	0,36	1216252,759	0,000468245
TRUE			87661	2597469403	0,64	0,36	1209450,883	0,000465627
TRUE			55115	2861668348	0,94	0,06	611044,4566	0,000213527
TRUE			61570	2861668348	0,96	0,04	682609,2206	0,000238535
TRUE			15780	3005535537	0,84	0,16	299090,2199	9,95131E-05
TRUE			814	3005535537	0,79	0,21	15428,35481	5,13331E-06
TRUE			360610	2424784351	0,59	0,41	6694756,657	0,00276097
TRUE			152580	2424784351	0,59	0,41	2832661,243	0,001168212
TRUE	0,142857143	0,190286811	4640	2897033911	0,81	0,19	52703,31667	1,81922E-05
TRUE	0,142857143	0,190286811	240	2897033911	0,66	0,34	2726,033621	9,40974E-07
TRUE			2162831	2431671281	0,7	0,3	14926369,91	0,006138317
TRUE			2148898	2431671281	0,7	0,3	14830213,94	0,006098774

Table 9.C This part represents the unification of SupplementaryData.xls, this data was rejected for the analysis as the data has been not enough for the analysis, and Telomere_variants_catalogue.xlsx where 2 more variables Ire added to the table Total_telomere_size and Percentage_telomere_size. This table was obtained via the script FusionVariantProteinInfo.py. please see data_extraction

Total telomere size

$$Totaltelomere\frac{TotalTelomereReads * BpReads}{GenomeCoverage}$$

Equation 2. Total telomere size was calculated by taking the to total telomere reads obtained in Telomere_variants_catalogue.xlsx multiplied by average of bp per read obtained in zoonomia_sp_info.xlsx and divided by the genome coverage obtained in Fusion_results_coverage.xlsx. The calculation of the equation was done by the script FusionVariantProteinInfo_feature_coverage.py.

Percentage telomere size

$$PercentagetelomereTotaltelomereGenome100$$

Equation 3. Percentage telomere size was calculated by taking Total telomere size previously analysed and divided by the total genome size represented in the table zoonomia_sp_info.xlsx. The calculation of the equation was done by the script FusionVariantProteinInfo_feature_coverage.py.

SupplementaryData.xls (example).

Species	Species sexual maturity (days)	Source of sexual maturity	CMR	ICM	Adult life expectancy
Echinops_telfairi		ZIMS	0,0571	0,0247	2720,27
Addax_nasomaculatus		ZIMS	0,0357	0,0616	3127,89
Aepyceros_melampus	374	ZIMS	0,0573	0,1412	2465,89
Alces_alces	374	ZIMS	0,0000	0,0000	2103,78
Ammotragus_lervia		ZIMS	0,0403	0,0853	3157,15
Antidorcas_marsupialis		ZIMS	0,0060	0,0054	1728,06
Antilocapra_americana		ZIMS	0,0313	0,0379	2057,88
Antilope_cervicapra		ZIMS	0,0000	0,0000	2307,02
Axis_axis		ZIMS	0,0811	0,0709	3332,79
Axis_porcinus		ZIMS	0,0345	0,0267	3381,52
Bison_bison		ZIMS	0,0548	0,0525	4678,87
Bison_bonasus		ZIMS	0,0408	0,0446	4130,22
Boselaphus_tragocamelus		ZIMS	0,0500	0,0495	3532,20
Budorcas_taxicolor		ZIMS	0,0357	0,0063	4042,38
Capra_caucasica		ZIMS	0,0000	0,0000	2178,04
Capra_falconeri		ZIMS	0,0137	0,1157	2226,85
Capra_ibex		ZIMS	0,0000	0,0000	2727,22
Capra_nubiana		ZIMS	0,0417	0,0291	2301,50
Cervus_canadensis		https://www.iucnredlist.org/species/55997823/55997871	0,0000	0,0000	4139,74
Cervus_elaphus	374	ZIMS	0,0000	0,0000	3589,92
Cervus_nippon		ZIMS	0,0164	0,0124	2727,23
Connochaetes_gnou		ZIMS	0,0000	0,0000	3279,07
Connochaetes_taurinus	374	ZIMS	0,0093	0,0079	3117,78
Dama_dama		ZIMS	0,0207	0,0131	2732,30
Dama_mesopotamica		https://genomics.senescence.info/species/entry.php?species=Dama_dama	0,1000	0,1516	3096,06

Table 10. This table represents where CMR and ICM was extracted. This data was not used in the final analysis.

Kruskal_{reference}_p_values.xlsx (example)

Column1	Total_fusions_	Total_fusion	Inward_distance_
RPAIN_g_ENSG00000129197__human	0,000954524	0,000722451	0,00062189
DAXX_g_ENSG00000204209__human	0,069667571	0,001051907	0,134813428
RPA1_g_ENSG00000132383__human	0,024881868	0,002162604	0,247996571
RAD52_g_ENSG00000002016__human	0,148761528	0,003352532	0,079930998
TP53_g_ENSG00000141510__human	0,049854514	0,006120418	0,219293375
ASF1A_g_ENSG00000111875__human	0,012244717	0,006762091	0,003963877
SETDB1_g_ENSG00000143379__human	0,164623908	0,039870599	0,003216202
TOP3A_g_ENSG00000177302__human	0,242404507	0,053719273	0,076256235
TERF2_g_ENSG00000132604__human	0,739522761	0,129609048	0,589238997
FEN1_g_ENSG00000168496__human	0,13789661	0,144497983	0,198918265
PML_g_ENSG00000140464__human	0,701682886	0,152754688	0,89091915
RPA2_g_ENSG00000117748__human	0,183745274	0,159870483	0,095612976
ZBTB40_g_ENSG00000184677__human	0,061513199	0,178691955	0,450730548
TINF2_g_ENSG00000092330__human	0,17999347	0,202658163	0,024345039
RTEL1_g_ENSG00000258366__human	0,507019445	0,242885091	0,890214741
BLM_g_ENSG00000197299__human	0,627535056	0,303673342	0,510290738
RAD51_g_ENSG00000051180__human	0,702032971	0,338071568	0,344158708

Table 11: Kruskal Wallis with the p-values of Total fusions, Total fusions 0, inward, outward, percentage of variability and so on.

Correlation_matrix_fusions_{reference}.xlsx

Column1	Inward_distance_	Outward_distance_	Inward_othe	Outward_othe	Total_fusion	Total_fusions_	Total_inwar	Total_outwar	Longevity	Chrom_nur
RPAIN_g_ENSG00000129197__human	-0,378583533	-0,333220207	-0,280788871	-0,601054481	-0,425296491	-0,37655846	-0,317813502	-0,590829332	-0,046419708	0,037178366
PML_g_ENSG00000140464__human	-0,030524503	-0,040229921	-0,147184755	-0,350212307	-0,205481881	-0,035326132	-0,142656312	-0,313183476	0,081268508	0,024833361
DAXX_g_ENSG00000204209__human	-0,098021797	-0,07817887	-0,084563616	-0,205197916	-0,131361986	-0,094488881	-0,093434278	-0,19452471	-0,162028229	-0,046362377
RTEL1_g_ENSG00000258366__human	-0,175979796	-0,050743746	-0,112769131	-0,016997823	-0,108949291	-0,136342577	-0,131050526	-0,025049478	0,035762289	-0,009972076
ASF1A_g_ENSG00000111875__human	-0,284123819	-0,230113627	-0,044569008	-0,075082921	-0,09769188	-0,27518604	-0,082379973	-0,111853432	-0,300843548	-0,098956833
RPA2_g_ENSG00000117748__human	-0,226394175	-0,180215542	-0,030459662	-0,042838671	-0,069480644	-0,21810485	-0,060892891	-0,073693808	-0,102489156	-0,088448601
Porcentage_TTAGGG	0,035869318	0,025162291	-0,042052384	-0,138683093	-0,06176884	0,033296228	-0,034402504	-0,115724789	0,132707494	0,109462127
POT1_g_ENSG00000128513__human	-0,1898894	-0,151731048	-0,020290849	-0,028880916	-0,05262618	-0,183150035	-0,046134063	-0,055785676	-0,102489156	-0,013185093
RAD52_g_ENSG00000002016__human	-0,095452736	-0,115173241	-0,020166447	-0,075786804	-0,052300311	-0,106518659	-0,032555914	-0,089264952	-0,214882824	0,057089648
Chrom_num	0,048706953	0,020325653	-0,042267626	-0,105464899	-0,052002381	0,040069791	-0,032782651	-0,087741548	0,166521785	1
FEN1_g_ENSG00000168496__human	-0,141211193	-0,113903391	-0,017995072	-0,024141975	-0,042017746	-0,136596512	-0,037038068	-0,044021872	0,051205879	0,037532186
TERF1_g_ENSG00000147601__human	-0,161118625	-0,12760543	-0,014318352	-0,020576595	-0,041437168	-0,154978165	-0,036420818	-0,043682007	-0,295903953	-0,003488755

Table 12: A correlation matrix table with all the values for human and mouse. please see data_extraction

Merged_table_all.xlsx (example)

Variant	table_Acomys_cahirinus_1	table_Acomys_cahirinus_2	table_Ailurus_fulgens_1	table_Ailurus_fulgens_2
ACCTAA	26414	7804	70	44
ATAGGG	19879	1529	16	15
CACTAA	20516	5246	23	15
CCATAA	14406	3303	31	14
CCCAAA	30130	4902	44	64
CCCCAA	36519	4199	100	94
CCCGAA	4453	943	24	50
CCCTAA	748974	86963	25635	18038
CCCTAC	16755	1873	1121	1044
CCCTAG	1552	486	217	117
CCCTAT	3672	960	69	89
CCCTCA	14549	1474	56	32
CCCTGA	7356	2391	24	12
CCCTTA	3316	983	20	14
CCGTAA	2626	922	17	6
CCTTAA	5995	1910	16	9
CGCTAA	2525	857	13	4
CTAGGG	11923	484	36	23
CTCTAA	4986	1415	49	19
GCCTAA	7912	2946	91	72
GTAGGG	32042	3448	81	59
TAAGGG	10231	1526	12	14
TCAGGG	12565	2914	21	27
TCCTAA	13955	4811	13	8
TGAGGG	24121	1932	34	49
TTAAGG	6336	2270	45	17

Table 13. The table represents from each sample (R1 and R2 fastQ files) The variant and the principal motif TTAGGG and CCCTAA frequency. The table was obtained from Motif_variant_analysis.py. please see data_extraction

Miniconda_env_lib.sh

```
$ miniconda_env_lib.sh
#!/bin/bash



mkdir -p /home/jrodriguezdv/miniconda3
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh -O /home/jrodriguezdv/miniconda3/miniconda.sh
bash /home/jrodriguezdv/miniconda3/miniconda.sh -b -u -p /home/jrodriguezdv/miniconda3
rm -rf /home/jrodriguezdv/miniconda3/miniconda.sh
miniconda3/bin/conda init bash
# Creación del environment
conda create -n telomere_env python=3.12
conda activate telomere_env
conda install -c bioconda sra-tools
conda install biopython
conda install jsonlines
conda install openpyxl
conda install pandas
conda install requests
conda install seaborn
conda install gprofiler-official
conda install xlrd
conda install sklearn
conda install natsort
```

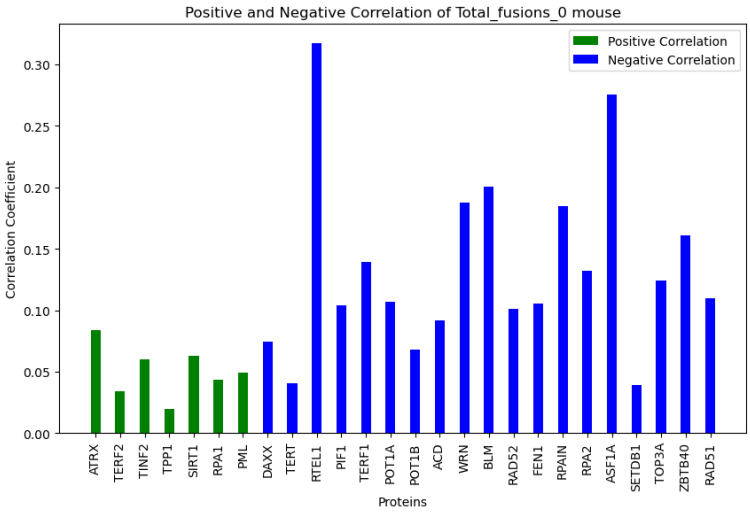
Figure S 4 This represents the environment, the libraries installed in the environment and the python version used.

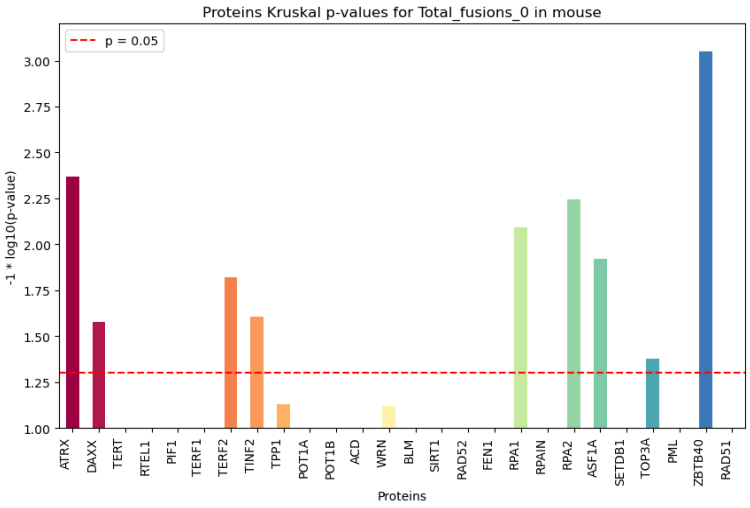
Telomere_process_raw_qsub.sh

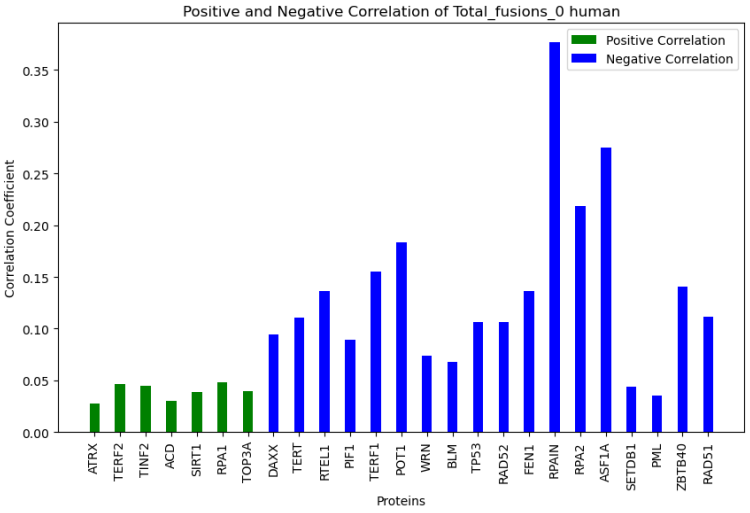
```
$ Telomere_process_raw_qsub.sh
#!/bin/bash
# Nombre del trabajo
#$ -t 1-250
#$ -tc 6
#$ -N RAW_2_TELO_zoo
# Salida estándar y de error
#$ -e error_telo_zoo.$TASK_ID.log
#$ -o output_telo_zoo.$TASK_ID.log
# Cola de trabajos
#$ -P AG
# cores necesarios.
#$ -pe pthreads 6
# memoria RAM requerida
#$ -l h_vmem=6G
#$ -A bioinformatica

export PYTHONPATH=
unset PKG_CONFIG_PATH
unset LD_LIBRARY_PATH
unset R_HOME
unset JAR_HOME
unset PERL5LIB
unset SAMTOOLS
.  /home/jrodriguezdv/miniconda3/bin/activate
conda activate telomere_env
python Telomere_analysis.py

```

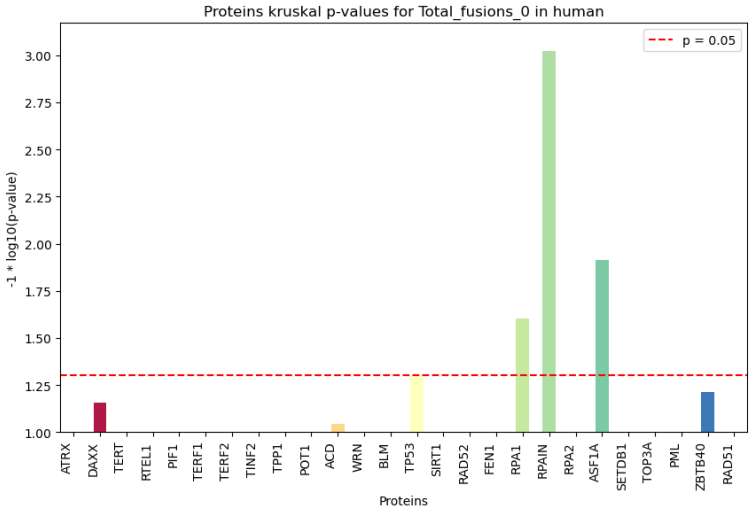
Figure S 5 This represents the script used in the laboratory which leads to the cluster used in the project

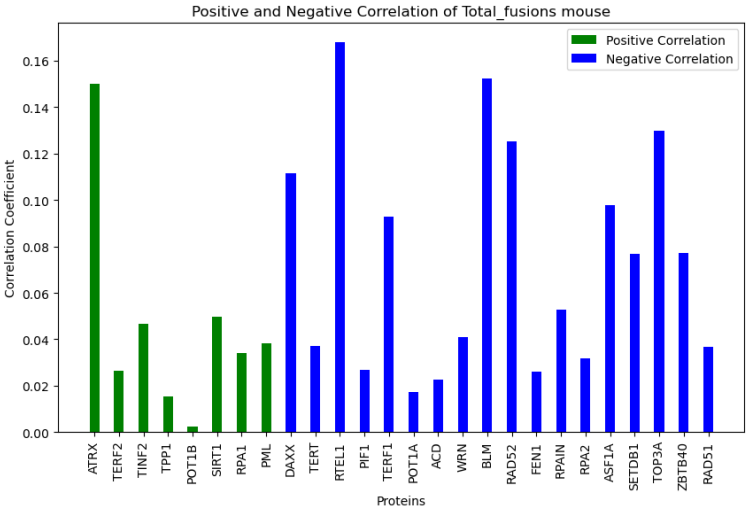
Correlation matrix and kluskal p-values representation.

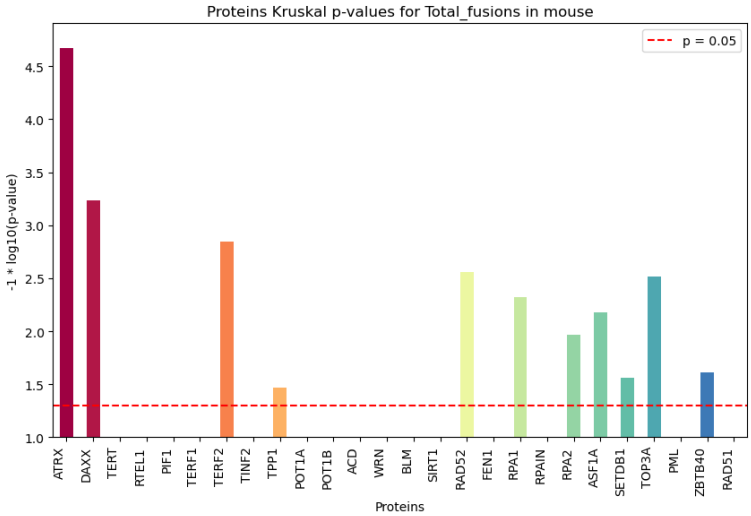

Positive and Negative Correlation of Total_fusions_0 mouse
Positive Correlation
Negative Correlation
Correlation Coefficient
Proteins
ATRX TERF2 TINF2 TPP1 SIRT1 RPA1 PML DAXX TERT RTEL1 PIF1 TERF1 POT1A POT1B ACD WRN BLM RAD52 FEN1 RPAIN RPA2 ASF1A SETDB1 TOP3A ZBTB40 RAD51

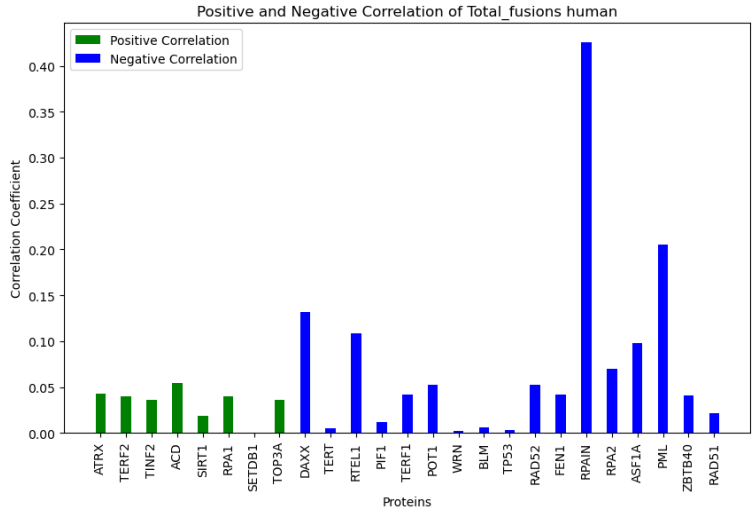

Proteins Kruskal p-values for Total_fusions_0 in mouse
p = 0.05
-1 * log10(p-value)
Proteins
ATRX DAXX TERT RTEL1 PIF1 TERF1 TERF2 TINF2 TPP1 POT1A POT1B ACD WRN BLM SIRT1 RAD52 FEN1 RPA1 RPAIN RPA2 ASF1A SETDB1 TOP3A PML ZBTB40 RAD51


Positive and Negative Correlation of Total_fusions_0 human
Positive Correlation
Negative Correlation
Correlation Coefficient
Proteins
ATRX TERF2 TINF2 ACD SIRT1 RPA1 TOP3A DAXX TERT RTEL1 PIF1 TERF1 POT1 WRN BLM TP53 RAD52 FEN1 RPAIN RPA2 ASF1A SETDB1 PML ZBTB40 RAD51


Proteins kruskal p-values for Total_fusions_0 in human
p = 0.05
-1 * log10(p-value)
Proteins
ATRX DAXX TERT RTEL1 PIF1 TERF1 TERF2 TINF2 TPP1 POT1 ACD WRN BLM TP53 SIRT1 RAD52 FEN1 RPA1 RPAIN RPA2 ASF1A SETDB1 TOP3A PML ZBTB40 RAD51


Positive and Negative Correlation of Total_fusions mouse
Positive Correlation
Negative Correlation
Correlation Coefficient
Proteins
ATRX TERF2 TINF2 TPP1 POT1B SIRT1 RPA1 PML DAXX TERT RTEL1 PIF1 TERF1 POT1A ACD WRN BLM RAD52 FEN1 RPAIN RPA2 ASF1A SETDB1 TOP3A ZBTB40 RAD51


Proteins Kruskal p-values for Total_fusions in mouse
p = 0.05
-1 * log10(p-value)
Proteins
ATRX DAXX TERT RTEL1 PIF1 TERF1 TERF2 TINF2 TPP1 POT1A POT1B ACD WRN BLM SIRT1 RAD52 FEN1 RPA1 RPAIN RPA2 ASF1A SETDB1 TOP3A PML ZBTB40 RAD51



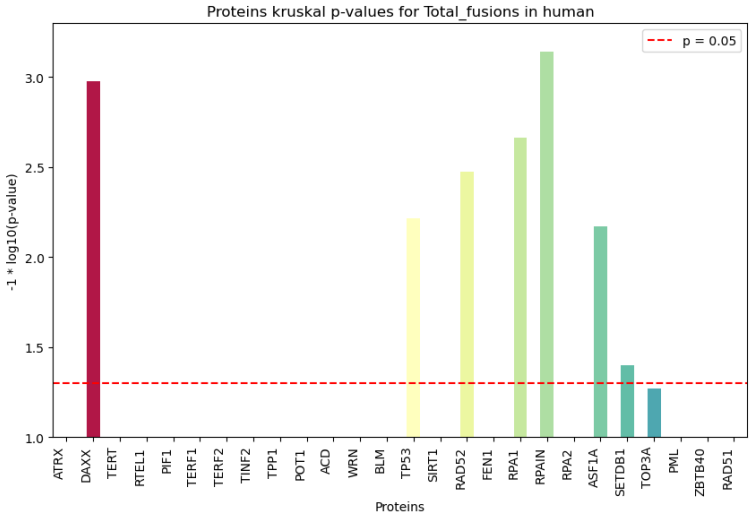
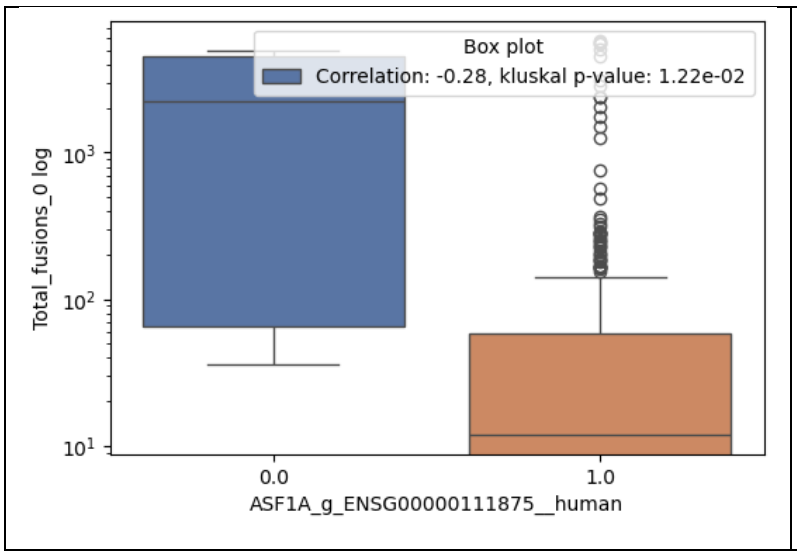
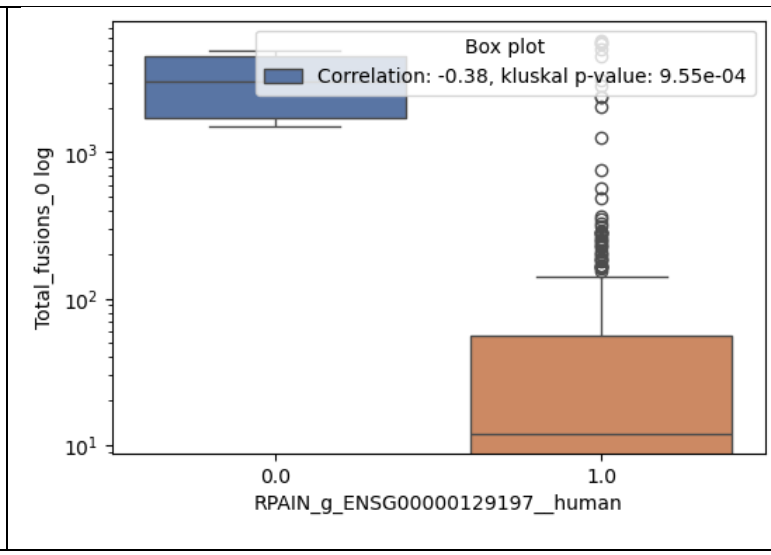
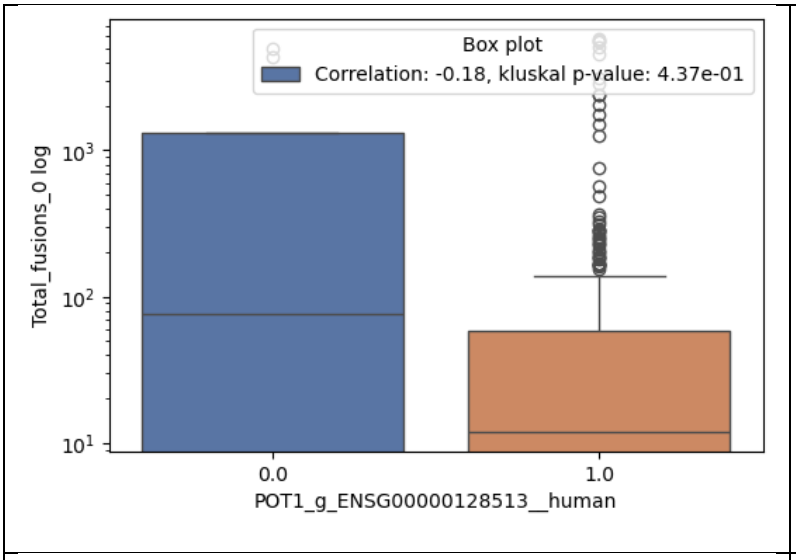


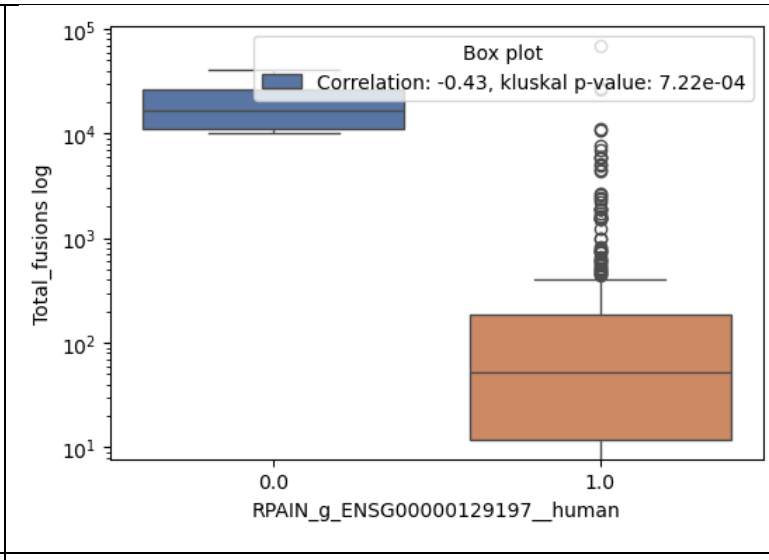
Figure S 6 Total_fusions_0 and Total_fusion representations.

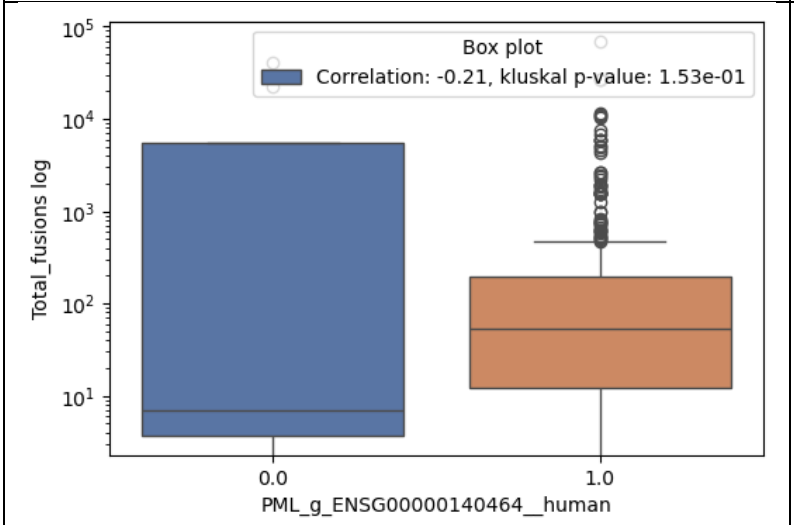
Correlation matrix and kluskal p-values representation.



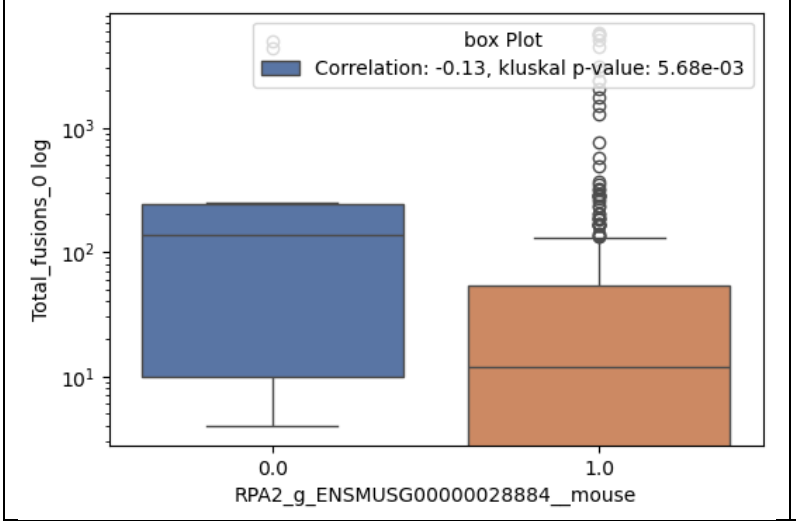


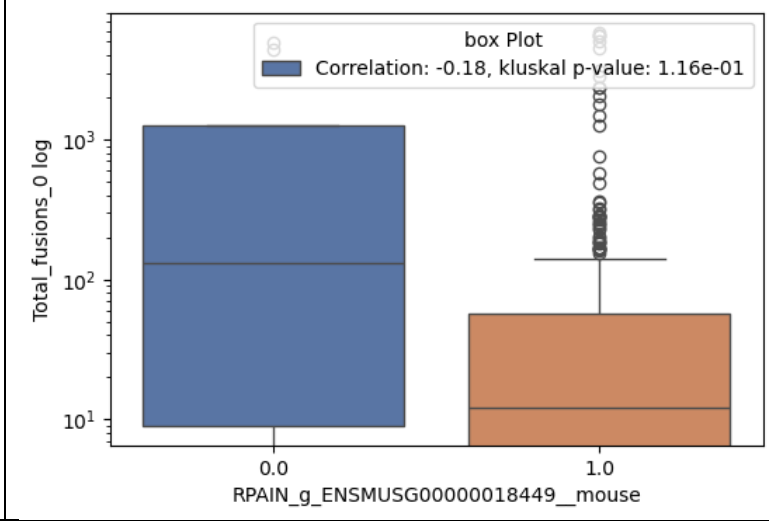

Box plot
Correlation: -0.18, kluskal p-value: 4.37e-01
Total_fusions_0 log
POT1_g_ENSG00000128513__human

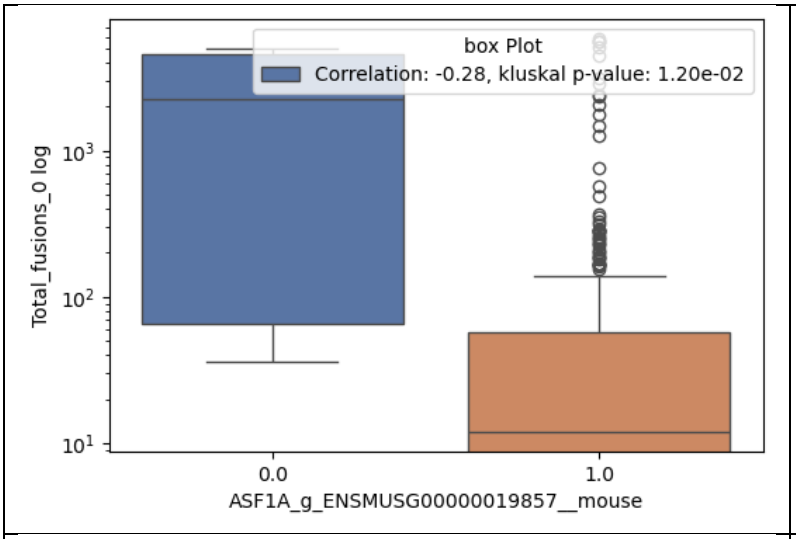

Box plot
Correlation: -0.43, kluskal p-value: 7.22e-04
Total_fusions log
RPAIN_g_ENSG00000129197__human

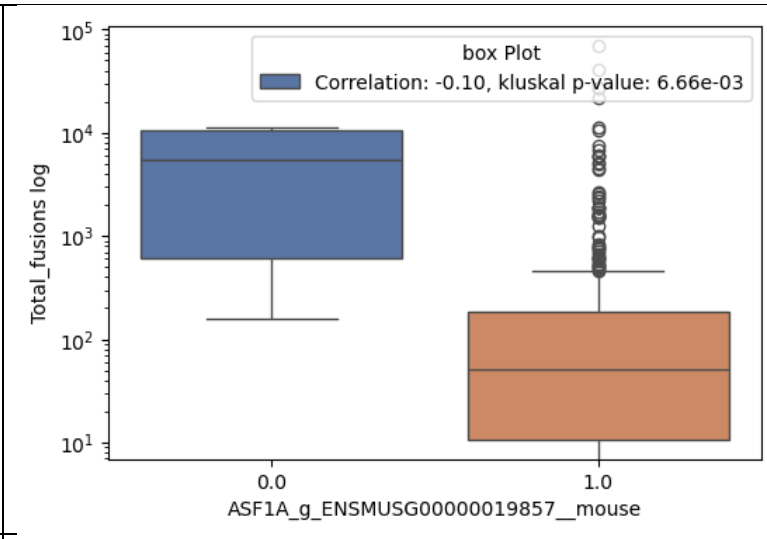

Box plot
Correlation: -0.21, kluskal p-value: 1.53e-01
Total_fusions log
PML_g_ENSG00000140464__human

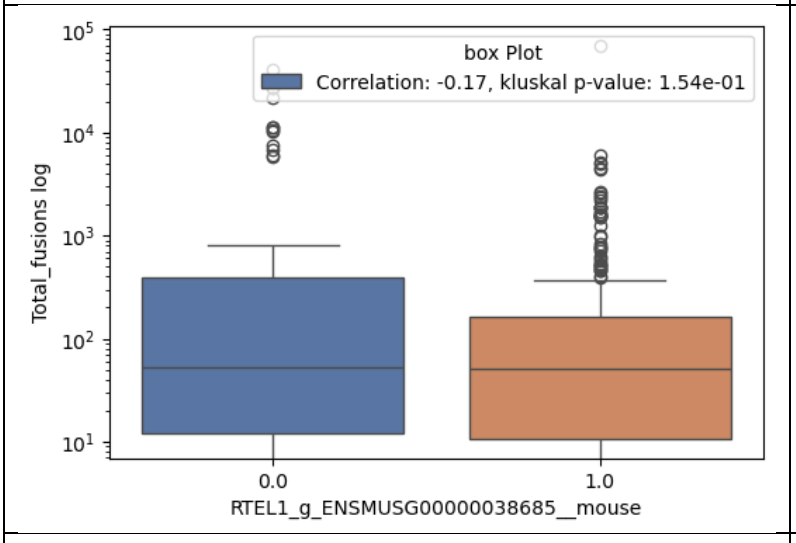
MOUSE

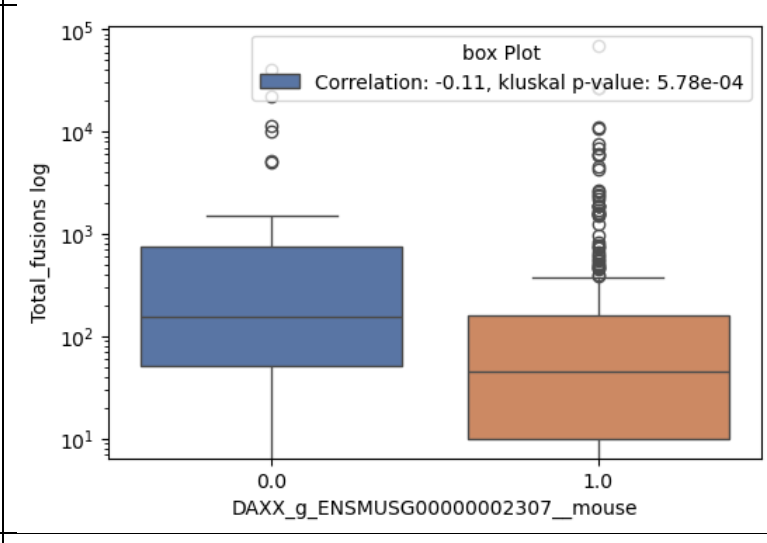

box Plot
Correlation: -0.13, kluskal p-value: 5.68e-03
Total_fusions_0 log
RPA2_g_ENSMUSG00000028884__mouse


box Plot
Correlation: -0.18, kluskal p-value: 1.16e-01
Total_fusions_0 log
RPAIN_g_ENSMUSG00000018449__mouse









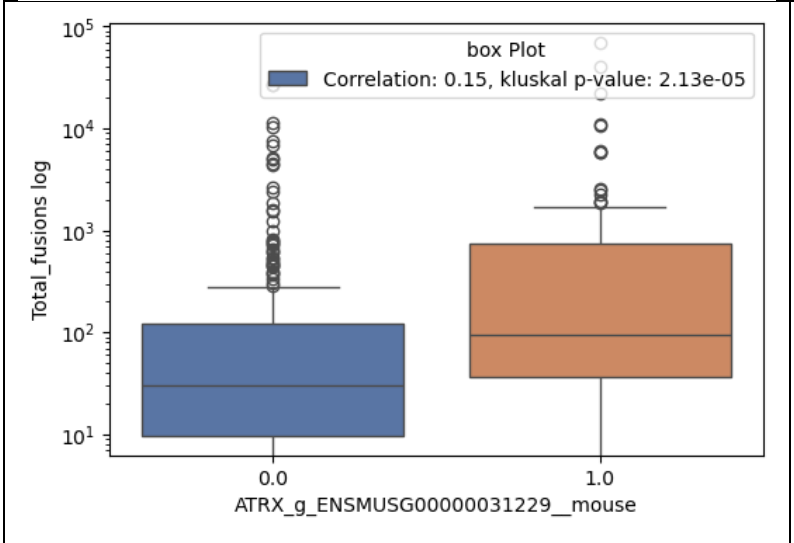


Figure S 7 Total_fusions_0 and Total_fusion representations.

Correlation matrix values, and distribution representation of all proteins.

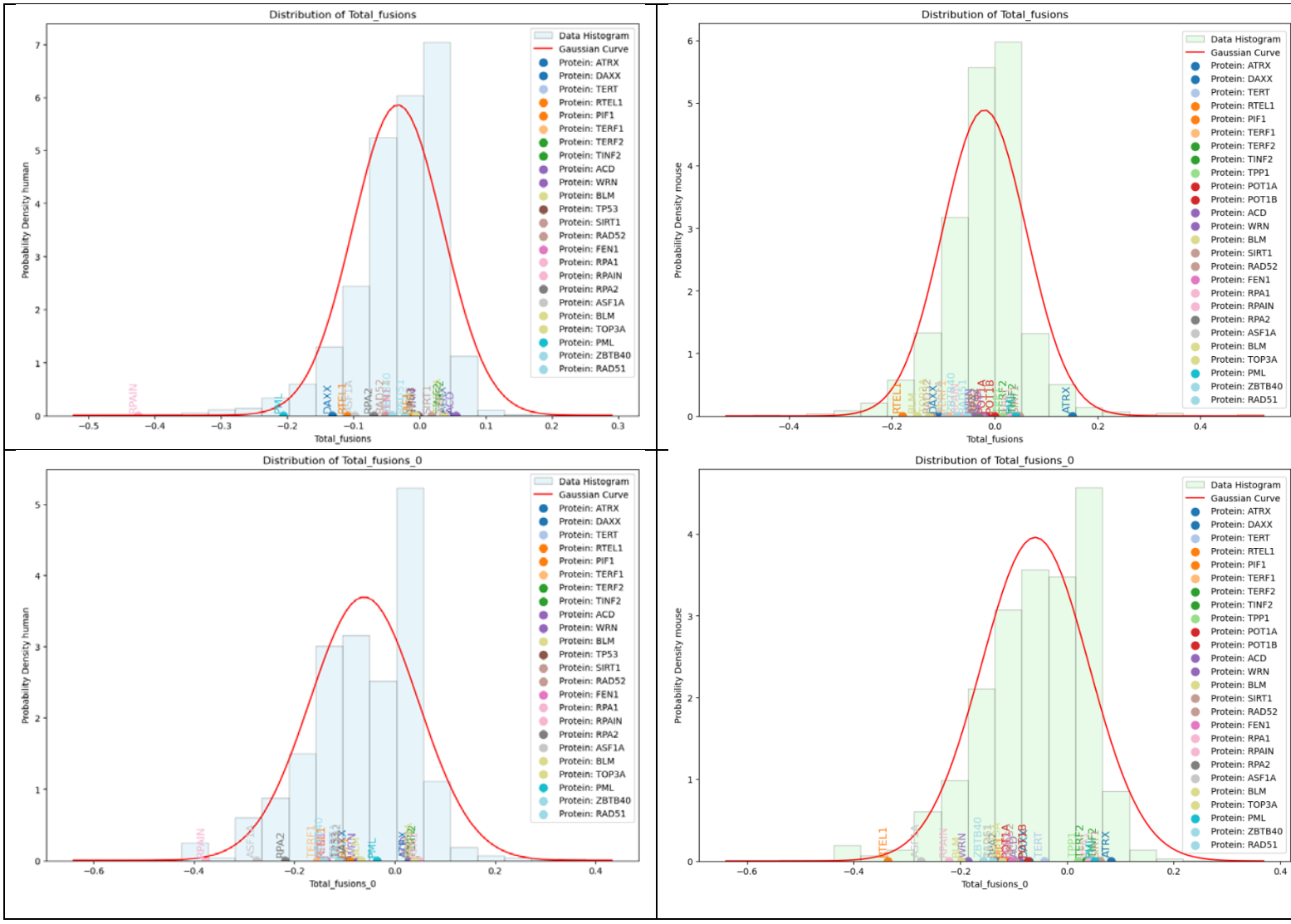
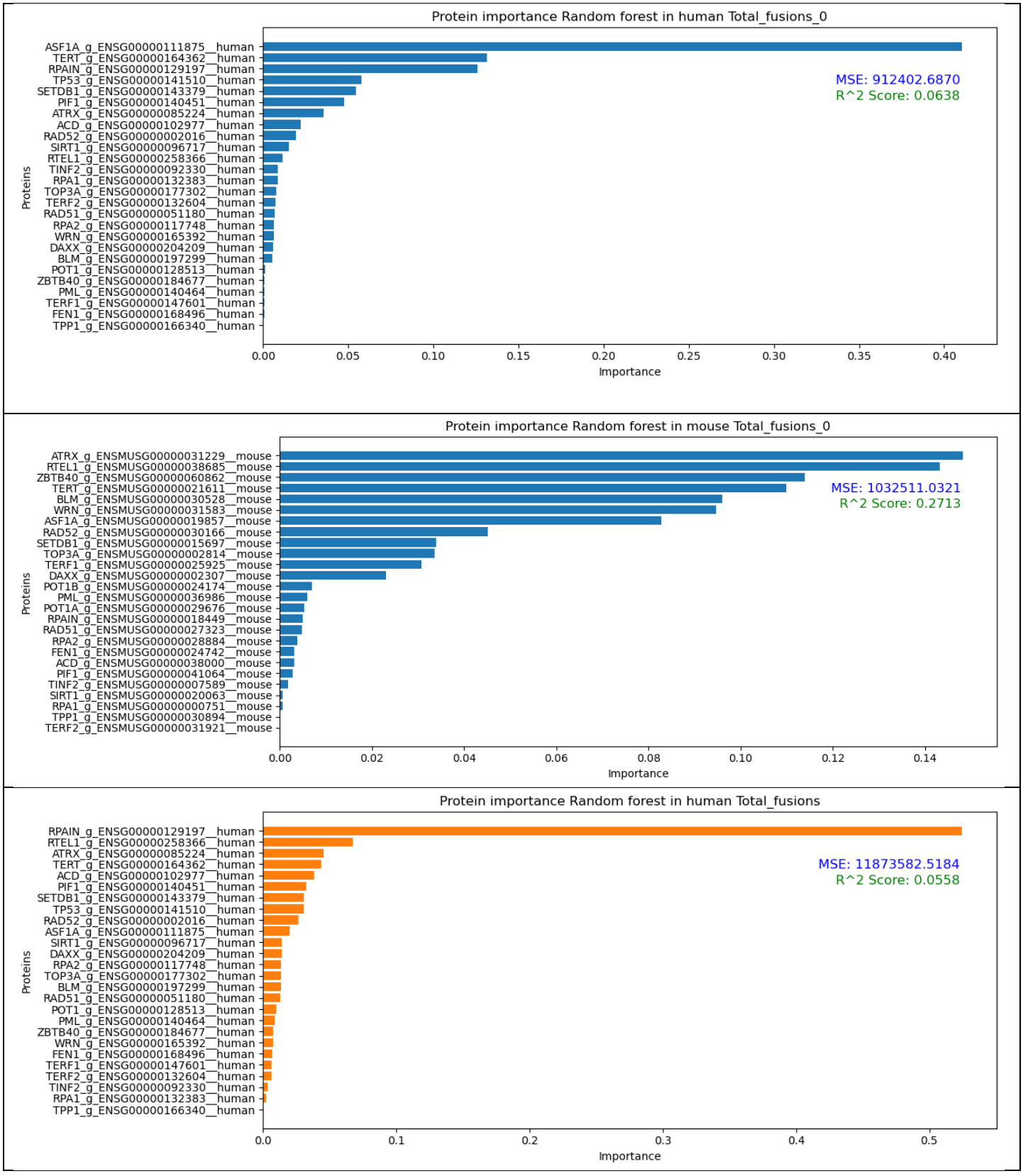
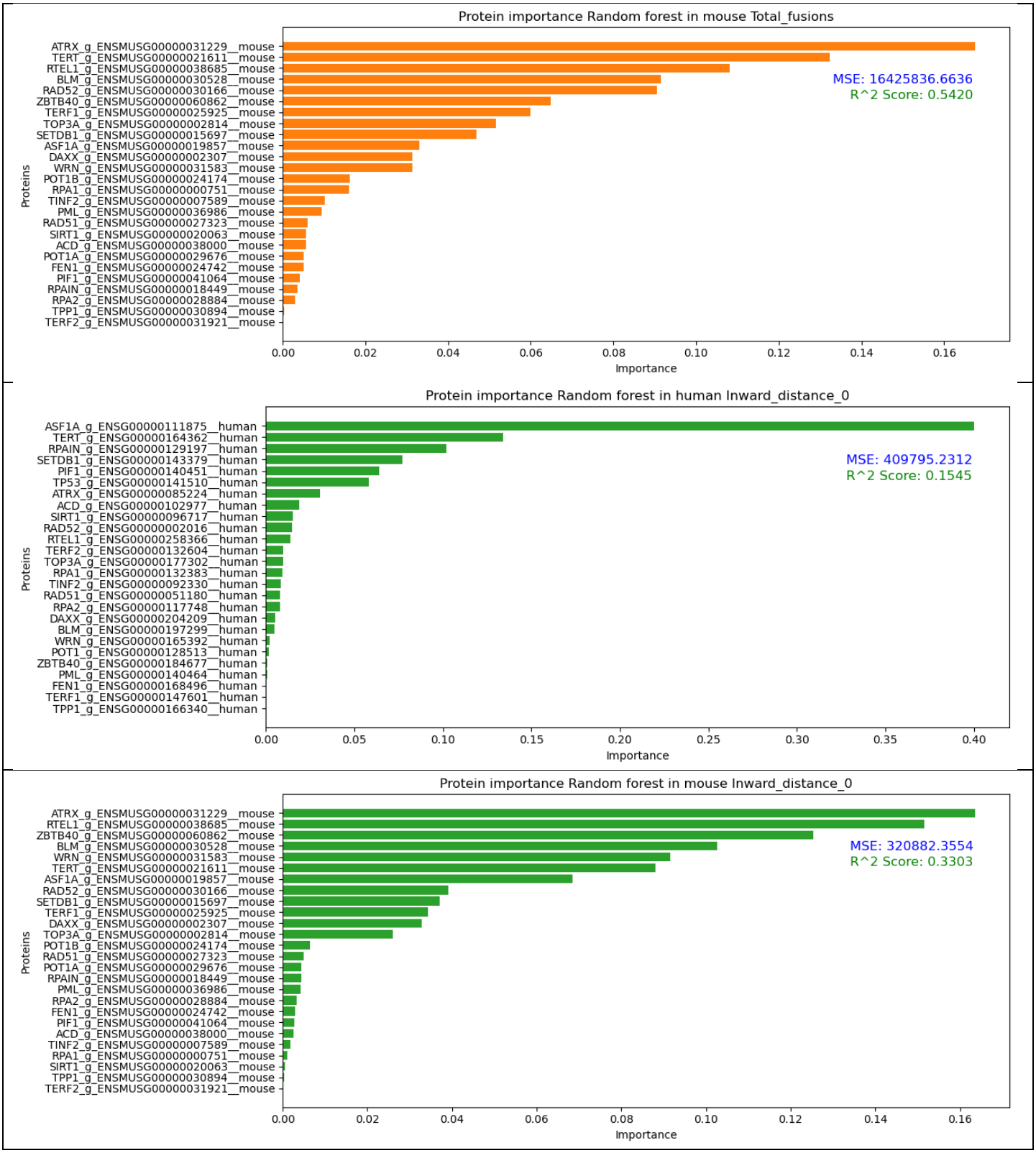
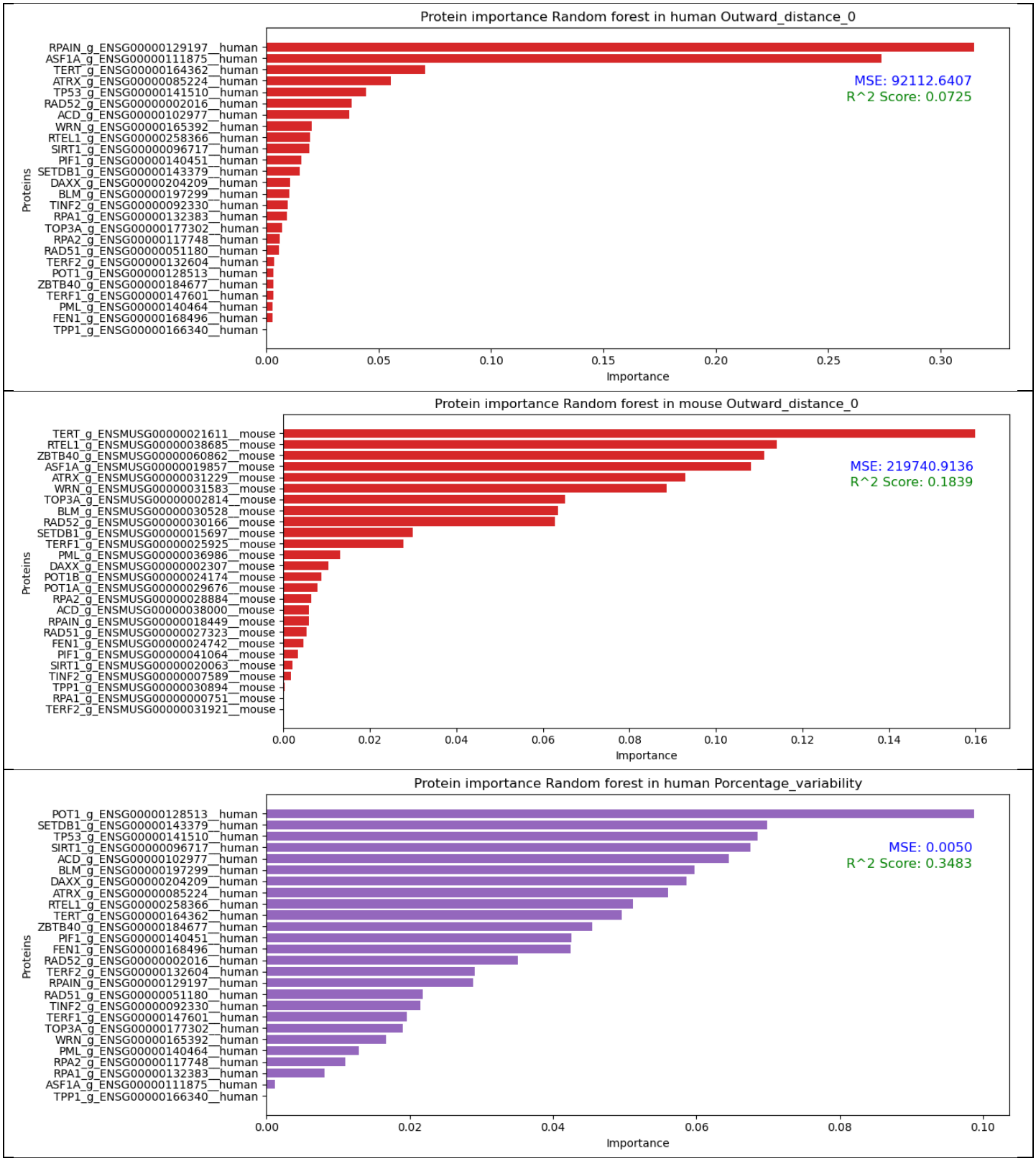


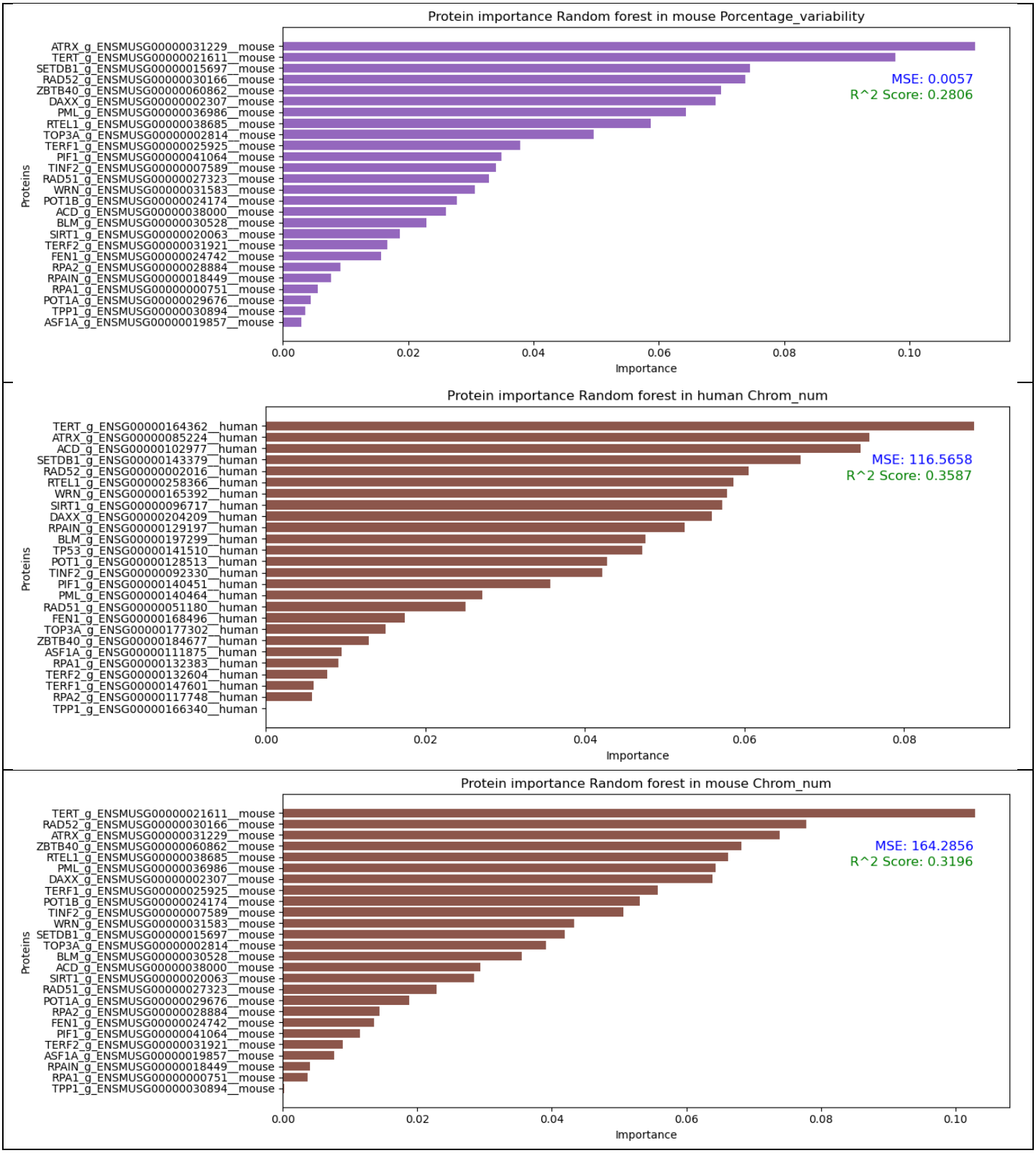
Figure S 8 Total_fusions_0 and Total fusions representations against all proteins highlighting the specific proteins.

Random Forest Proteins


Protein importance Random forest in human Total_fusions_0
ASF1A_g_ENSG00000111875__human
TERT_g_ENSG00000164362__human
RPAIN_g_ENSG00000129197__human
TP53_g_ENSG00000141510__human
SETDB1_g_ENSG00000143379__human
PIF1_g_ENSG00000140451__human
ATRX_g_ENSG00000085224__human
ACD_g_ENSG00000102977__human
RAD52_g_ENSG00000002016__human
SIRT1_g_ENSG00000096717__human
RTEL1_g_ENSG00000258366__human
TINF2_g_ENSG00000092330__human
RPA1_g_ENSG00000132383__human
TOP3A_g_ENSG00000177302__human
TERF2_g_ENSG00000132604__human
RAD51_g_ENSG00000051180__human
RPA2_g_ENSG00000117748__human
WRN_g_ENSG00000165392__human
DAXX_g_ENSG00000204209__human
BLM_g_ENSG00000197299__human
POT1_g_ENSG00000128513__human
ZBTB40_g_ENSG00000184677__human
PML_g_ENSG00000140464__human
TERF1_g_ENSG00000147601__human
FEN1_g_ENSG00000168496__human
TPP1_g_ENSG00000166340__human
Proteins
MSE: 912402.6870
R^2 Score: 0.0638
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Importance
Protein importance Random forest in mouse Total_fusions_0
ATRX_g_ENSMUSG00000031229__mouse
RTEL1_g_ENSMUSG00000038685__mouse
ZBTB40_g_ENSMUSG00000060862__mouse
TERT_g_ENSMUSG00000021611__mouse
BLM_g_ENSMUSG00000030528__mouse
WRN_g_ENSMUSG00000031583__mouse
ASF1A_g_ENSMUSG00000019857__mouse
RAD52_g_ENSMUSG00000030166__mouse
SETDB1_g_ENSMUSG00000015697__mouse
TOP3A_g_ENSMUSG00000002814__mouse
TERF1_g_ENSMUSG00000025925__mouse
DAXX_g_ENSMUSG00000002307__mouse
POT1B_g_ENSMUSG00000024174__mouse
PML_g_ENSMUSG00000036986__mouse
POT1A_g_ENSMUSG00000029676__mouse
RPAIN_g_ENSMUSG00000018449__mouse
RAD51_g_ENSMUSG00000027323__mouse
RPA2_g_ENSMUSG00000028884__mouse
FEN1_g_ENSMUSG00000024742__mouse
ACD_g_ENSMUSG00000038000__mouse
PIF1_g_ENSMUSG00000041064__mouse
TINF2_g_ENSMUSG00000007589__mouse
SIRT1_g_ENSMUSG00000020063__mouse
RPA1_g_ENSMUSG00000000751__mouse
TPP1_g_ENSMUSG00000030894__mouse
TERF2_g_ENSMUSG00000031921__mouse
Proteins
MSE: 1032511.0321
R^2 Score: 0.2713
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Importance
Protein importance Random forest in human Total_fusions
RPAIN_g_ENSG00000129197__human
RTEL1_g_ENSG00000258366__human
ATRX_g_ENSG00000085224__human
TERT_g_ENSG00000164362__human
ACD_g_ENSG00000102977__human
PIF1_g_ENSG00000140451__human
SETDB1_g_ENSG00000143379__human
TP53_g_ENSG00000141510__human
RAD52_g_ENSG00000002016__human
ASF1A_g_ENSG00000111875__human
SIRT1_g_ENSG00000096717__human
DAXX_g_ENSG00000204209__human
RPA2_g_ENSG00000117748__human
TOP3A_g_ENSG00000177302__human
BLM_g_ENSG00000197299__human
RAD51_g_ENSG00000051180__human
POT1_g_ENSG00000128513__human
PML_g_ENSG00000140464__human
ZBTB40_g_ENSG00000184677__human
WRN_g_ENSG00000165392__human
FEN1_g_ENSG00000168496__human
TERF1_g_ENSG00000147601__human
TERF2_g_ENSG00000132604__human
TINF2_g_ENSG00000092330__human
RPA1_g_ENSG00000132383__human
TPP1_g_ENSG00000166340__human
Proteins
MSE: 11873582.5184
R^2 Score: 0.0558
0.0
0.1
0.2
0.3
0.4
0.5
Importance


Protein importance Random forest in mouse Total_fusions
ATRX_g_ENSMUSG00000031229__mouse
TERT_g_ENSMUSG00000021611__mouse
RTEL1_g_ENSMUSG00000038685__mouse
BLM_g_ENSMUSG00000030528__mouse
RAD52_g_ENSMUSG00000030166__mouse
ZBTB40_g_ENSMUSG00000060862__mouse
TERF1_g_ENSMUSG00000025925__mouse
TOP3A_g_ENSMUSG00000002814__mouse
SETDB1_g_ENSMUSG00000015697__mouse
ASF1A_g_ENSMUSG00000019857__mouse
DAXX_g_ENSMUSG00000002307__mouse
WRN_g_ENSMUSG00000031583__mouse
POT1B_g_ENSMUSG00000024174__mouse
RPA1_g_ENSMUSG00000000751__mouse
TINF2_g_ENSMUSG00000007589__mouse
PML_g_ENSMUSG00000036986__mouse
RAD51_g_ENSMUSG00000027323__mouse
SIRT1_g_ENSMUSG00000020063__mouse
ACD_g_ENSMUSG00000038000__mouse
POT1A_g_ENSMUSG00000029676__mouse
FEN1_g_ENSMUSG00000024742__mouse
PIF1_g_ENSMUSG00000041064__mouse
RPAIN_g_ENSMUSG00000018449__mouse
RPA2_g_ENSMUSG00000028884__mouse
TPP1_g_ENSMUSG00000030894__mouse
TERF2_g_ENSMUSG00000031921__mouse
Proteins
MSE: 16425836.6636
R^2 Score: 0.5420
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16
Importance
Protein importance Random forest in human Inward_distance_0
ASF1A_g_ENSG00000111875__human
TERT_g_ENSG00000164362__human
RPAIN_g_ENSG00000129197__human
SETDB1_g_ENSG00000143379__human
PIF1_g_ENSG00000140451__human
TP53_g_ENSG00000141510__human
ATRX_g_ENSG00000085224__human
ACD_g_ENSG00000102977__human
SIRT1_g_ENSG00000096717__human
RAD52_g_ENSG00000002016__human
RTEL1_g_ENSG00000258366__human
TERF2_g_ENSG00000132604__human
TOP3A_g_ENSG00000177302__human
RPA1_g_ENSG00000132383__human
TINF2_g_ENSG00000092330__human
RAD51_g_ENSG00000051180__human
RPA2_g_ENSG00000117748__human
DAXX_g_ENSG00000204209__human
BLM_g_ENSG00000197299__human
WRN_g_ENSG00000165392__human
POT1_g_ENSG00000128513__human
ZBTB40_g_ENSG00000184677__human
PML_g_ENSG00000140464__human
FEN1_g_ENSG00000168496__human
TERF1_g_ENSG00000147601__human
TPP1_g_ENSG00000166340__human
Proteins
MSE: 409795.2312
R^2 Score: 0.1545
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
Importance
Protein importance Random forest in mouse Inward_distance_0
ATRX_g_ENSMUSG00000031229__mouse
RTEL1_g_ENSMUSG00000038685__mouse
ZBTB40_g_ENSMUSG00000060862__mouse
BLM_g_ENSMUSG00000030528__mouse
WRN_g_ENSMUSG00000031583__mouse
TERT_g_ENSMUSG00000021611__mouse
ASF1A_g_ENSMUSG00000019857__mouse
RAD52_g_ENSMUSG00000030166__mouse
SETDB1_g_ENSMUSG00000015697__mouse
TERF1_g_ENSMUSG00000025925__mouse
DAXX_g_ENSMUSG00000002307__mouse
TOP3A_g_ENSMUSG00000002814__mouse
POT1B_g_ENSMUSG00000024174__mouse
RAD51_g_ENSMUSG00000027323__mouse
POT1A_g_ENSMUSG00000029676__mouse
RPAIN_g_ENSMUSG00000018449__mouse
PML_g_ENSMUSG00000036986__mouse
RPA2_g_ENSMUSG00000028884__mouse
FEN1_g_ENSMUSG00000024742__mouse
PIF1_g_ENSMUSG00000041064__mouse
ACD_g_ENSMUSG00000038000__mouse
TINF2_g_ENSMUSG00000007589__mouse
RPA1_g_ENSMUSG00000000751__mouse
SIRT1_g_ENSMUSG00000020063__mouse
TPP1_g_ENSMUSG00000030894__mouse
TERF2_g_ENSMUSG00000031921__mouse
Proteins
MSE: 320882.3554
R^2 Score: 0.3303
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16
Importance


Protein importance Random forest in human Outward_distance_0
RPAIN_g_ENSG00000129197__human
ASF1A_g_ENSG00000111875__human
TERT_g_ENSG00000164362__human
ATRX_g_ENSG00000085224__human
TP53_g_ENSG00000141510__human
RAD52_g_ENSG00000002016__human
ACD_g_ENSG00000102977__human
WRN_g_ENSG00000165392__human
RTEL1_g_ENSG00000258366__human
SIRT1_g_ENSG00000096717__human
PIF1_g_ENSG00000140451__human
SETDB1_g_ENSG00000143379__human
DAXX_g_ENSG00000204209__human
BLM_g_ENSG00000197299__human
TINF2_g_ENSG00000092330__human
RPA1_g_ENSG00000132383__human
TOP3A_g_ENSG00000177302__human
RPA2_g_ENSG00000117748__human
RAD51_g_ENSG00000051180__human
TERF2_g_ENSG00000132604__human
POT1_g_ENSG00000128513__human
ZBTB40_g_ENSG00000184677__human
TERF1_g_ENSG00000147601__human
PML_g_ENSG00000140464__human
FEN1_g_ENSG00000168496__human
TPP1_g_ENSG00000166340__human
Proteins
MSE: 92112.6407
R^2 Score: 0.0725
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Importance
Protein importance Random forest in mouse Outward_distance_0
TERT_g_ENSMUSG00000021611__mouse
RTEL1_g_ENSMUSG00000038685__mouse
ZBTB40_g_ENSMUSG00000060862__mouse
ASF1A_g_ENSMUSG00000019857__mouse
ATRX_g_ENSMUSG00000031229__mouse
WRN_g_ENSMUSG00000031583__mouse
TOP3A_g_ENSMUSG00000002814__mouse
BLM_g_ENSMUSG00000030528__mouse
RAD52_g_ENSMUSG00000030166__mouse
SETDB1_g_ENSMUSG00000015697__mouse
TERF1_g_ENSMUSG00000025925__mouse
PML_g_ENSMUSG00000036986__mouse
DAXX_g_ENSMUSG00000002307__mouse
POT1B_g_ENSMUSG00000024174__mouse
POT1A_g_ENSMUSG00000029676__mouse
RPA2_g_ENSMUSG00000028884__mouse
ACD_g_ENSMUSG00000038000__mouse
RPAIN_g_ENSMUSG00000018449__mouse
RAD51_g_ENSMUSG00000027323__mouse
FEN1_g_ENSMUSG00000024742__mouse
PIF1_g_ENSMUSG00000041064__mouse
SIRT1_g_ENSMUSG00000020063__mouse
TINF2_g_ENSMUSG00000007589__mouse
TPP1_g_ENSMUSG00000030894__mouse
RPA1_g_ENSMUSG00000000751__mouse
TERF2_g_ENSMUSG00000031921__mouse
Proteins
MSE: 219740.9136
R^2 Score: 0.1839
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Importance
Protein importance Random forest in human Porcentage_variability
POT1_g_ENSG00000128513__human
SETDB1_g_ENSG00000143379__human
TP53_g_ENSG00000141510__human
SIRT1_g_ENSG00000096717__human
ACD_g_ENSG00000102977__human
BLM_g_ENSG00000197299__human
DAXX_g_ENSG00000204209__human
ATRX_g_ENSG00000085224__human
RTEL1_g_ENSG00000258366__human
TERT_g_ENSG00000164362__human
ZBTB40_g_ENSG00000184677__human
PIF1_g_ENSG00000140451__human
FEN1_g_ENSG00000168496__human
RAD52_g_ENSG00000002016__human
TERF2_g_ENSG00000132604__human
RPAIN_g_ENSG00000129197__human
RAD51_g_ENSG00000051180__human
TINF2_g_ENSG00000092330__human
TERF1_g_ENSG00000147601__human
TOP3A_g_ENSG00000177302__human
WRN_g_ENSG00000165392__human
PML_g_ENSG00000140464__human
RPA2_g_ENSG00000117748__human
RPA1_g_ENSG00000132383__human
ASF1A_g_ENSG00000111875__human
TPP1_g_ENSG00000166340__human
Proteins
MSE: 0.0050
R^2 Score: 0.3483
0.00
0.02
0.04
0.06
0.08
0.10
Importance


Protein importance Random forest in mouse Porcentage_variability
ATRX_g_ENSMUSG00000031229__mouse
TERT_g_ENSMUSG00000021611__mouse
SETDB1_g_ENSMUSG00000015697__mouse
RAD52_g_ENSMUSG00000030166__mouse
ZBTB40_g_ENSMUSG00000060862__mouse
DAXX_g_ENSMUSG00000002307__mouse
PML_g_ENSMUSG00000036986__mouse
RTEL1_g_ENSMUSG00000038685__mouse
TOP3A_g_ENSMUSG00000002814__mouse
TERF1_g_ENSMUSG00000025925__mouse
PIF1_g_ENSMUSG00000041064__mouse
TINF2_g_ENSMUSG00000007589__mouse
RAD51_g_ENSMUSG00000027323__mouse
WRN_g_ENSMUSG00000031583__mouse
POT1B_g_ENSMUSG00000024174__mouse
ACD_g_ENSMUSG00000038000__mouse
BLM_g_ENSMUSG00000030528__mouse
SIRT1_g_ENSMUSG00000020063__mouse
TERF2_g_ENSMUSG00000031921__mouse
FEN1_g_ENSMUSG00000024742__mouse
RPA2_g_ENSMUSG00000028884__mouse
RPAIN_g_ENSMUSG00000018449__mouse
RPA1_g_ENSMUSG00000000751__mouse
POT1A_g_ENSMUSG00000029676__mouse
TPP1_g_ENSMUSG00000030894__mouse
ASF1A_g_ENSMUSG00000019857__mouse
Proteins
0.00
0.02
0.04
0.06
0.08
0.10
Importance
MSE: 0.0057
R^2 Score: 0.2806
Protein importance Random forest in human Chrom_num
TERT_g_ENSG00000164362__human
ATRX_g_ENSG00000085224__human
ACD_g_ENSG00000102977__human
SETDB1_g_ENSG00000143379__human
RAD52_g_ENSG00000002016__human
RTEL1_g_ENSG00000258366__human
WRN_g_ENSG00000165392__human
SIRT1_g_ENSG00000096717__human
DAXX_g_ENSG00000204209__human
RPAIN_g_ENSG00000129197__human
BLM_g_ENSG00000197299__human
TP53_g_ENSG00000141510__human
POT1_g_ENSG00000128513__human
TINF2_g_ENSG00000092330__human
PIF1_g_ENSG00000140451__human
PML_g_ENSG00000140464__human
RAD51_g_ENSG00000051180__human
FEN1_g_ENSG00000168496__human
TOP3A_g_ENSG00000177302__human
ZBTB40_g_ENSG00000184677__human
ASF1A_g_ENSG00000111875__human
RPA1_g_ENSG00000132383__human
TERF2_g_ENSG00000132604__human
TERF1_g_ENSG00000147601__human
RPA2_g_ENSG00000117748__human
TPP1_g_ENSG00000166340__human
Proteins
0.00
0.02
0.04
0.06
0.08
Importance
MSE: 116.5658
R^2 Score: 0.3587
Protein importance Random forest in mouse Chrom_num
TERT_g_ENSMUSG00000021611__mouse
RAD52_g_ENSMUSG00000030166__mouse
ATRX_g_ENSMUSG00000031229__mouse
ZBTB40_g_ENSMUSG00000060862__mouse
RTEL1_g_ENSMUSG00000038685__mouse
PML_g_ENSMUSG00000036986__mouse
DAXX_g_ENSMUSG00000002307__mouse
TERF1_g_ENSMUSG00000025925__mouse
POT1B_g_ENSMUSG00000024174__mouse
TINF2_g_ENSMUSG00000007589__mouse
WRN_g_ENSMUSG00000031583__mouse
SETDB1_g_ENSMUSG00000015697__mouse
TOP3A_g_ENSMUSG00000002814__mouse
BLM_g_ENSMUSG00000030528__mouse
ACD_g_ENSMUSG00000038000__mouse
SIRT1_g_ENSMUSG00000020063__mouse
RAD51_g_ENSMUSG00000027323__mouse
POT1A_g_ENSMUSG00000029676__mouse
RPA2_g_ENSMUSG00000028884__mouse
FEN1_g_ENSMUSG00000024742__mouse
PIF1_g_ENSMUSG00000041064__mouse
TERF2_g_ENSMUSG00000031921__mouse
ASF1A_g_ENSMUSG00000019857__mouse
RPAIN_g_ENSMUSG00000018449__mouse
RPA1_g_ENSMUSG00000000751__mouse
TPP1_g_ENSMUSG00000030894__mouse
Proteins
0.00
0.02
0.04
0.06
0.08
0.10
Importance
MSE: 164.2856
R^2 Score: 0.3196

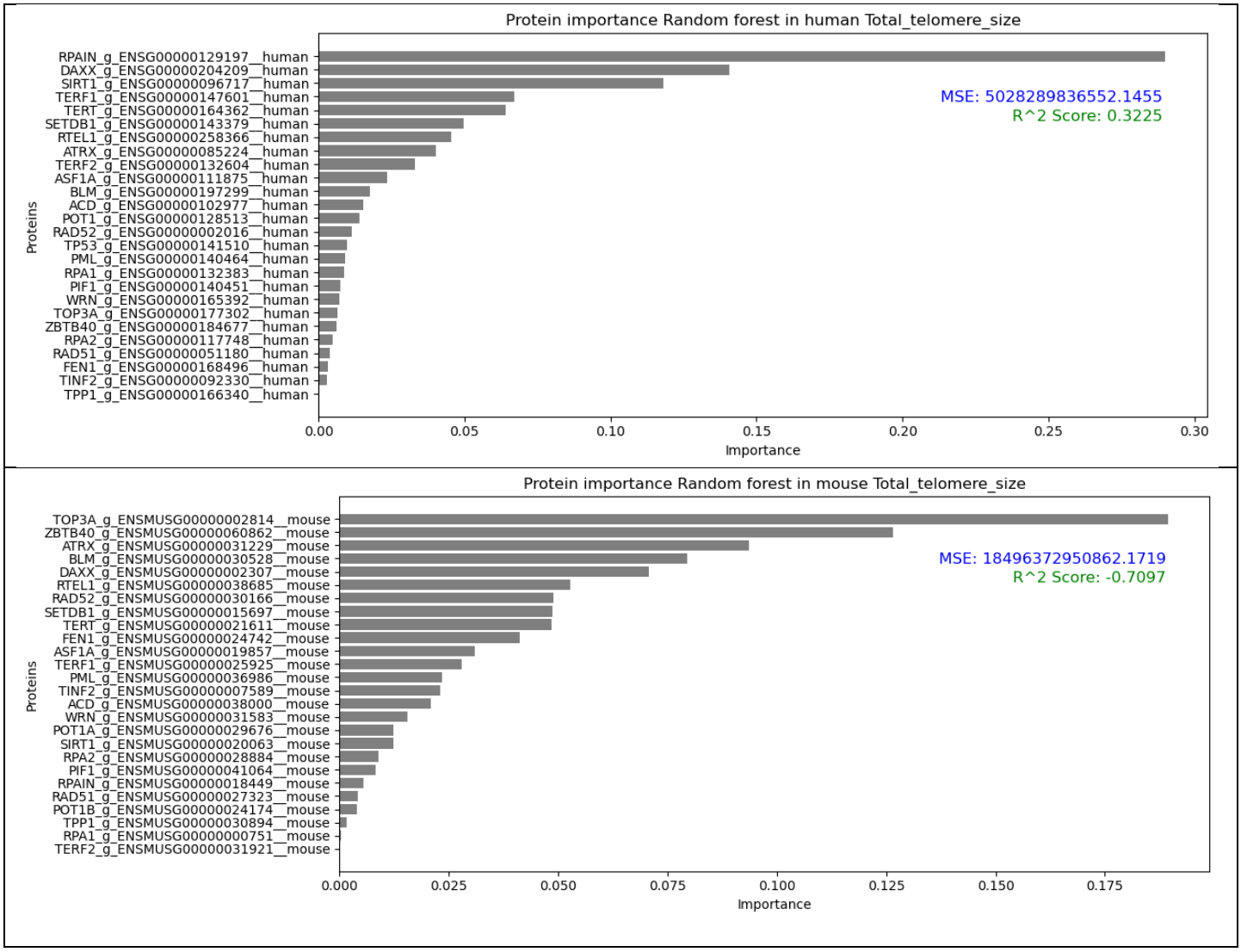
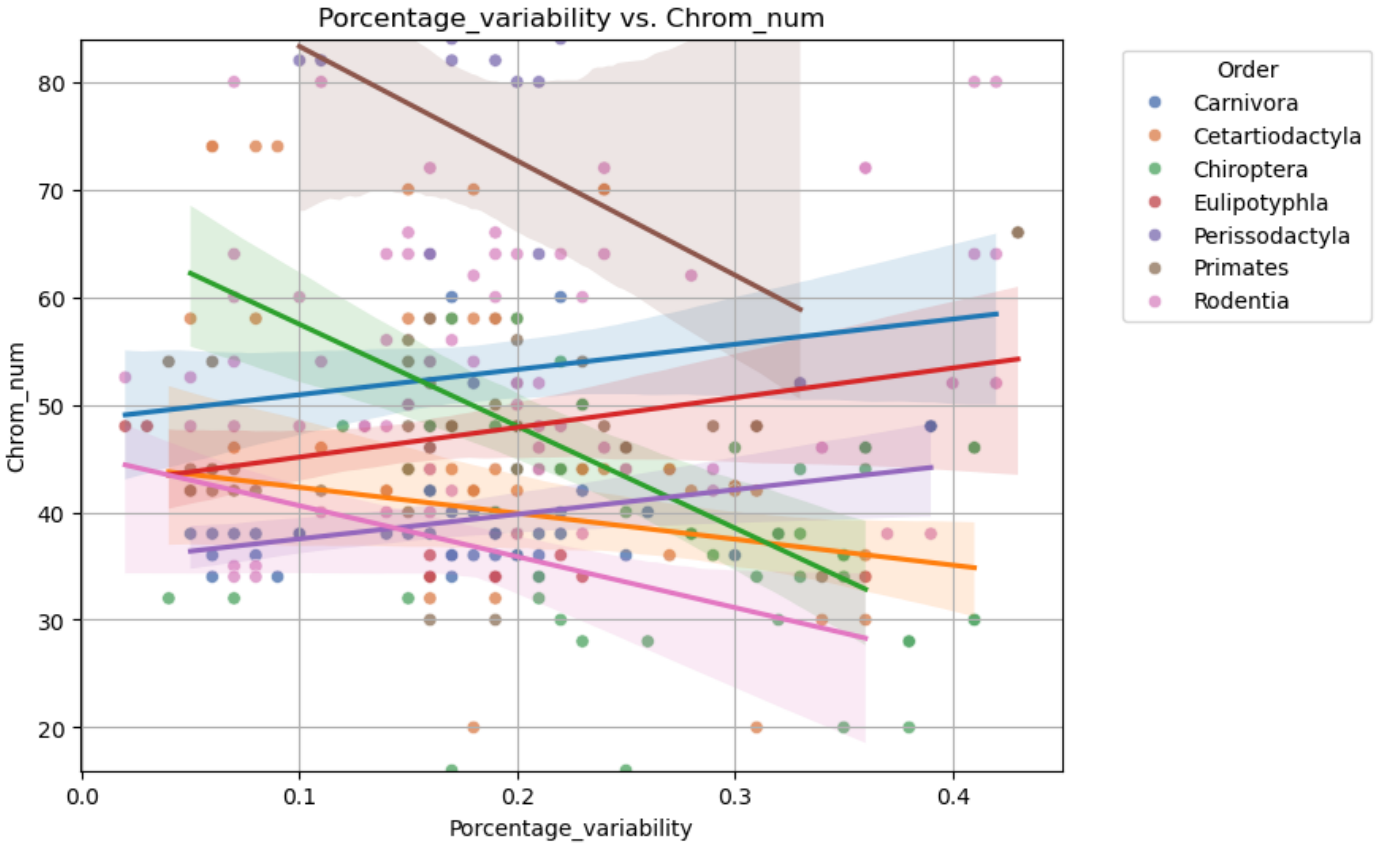


Figure S 9 Random forest done with Estimators number 10000, depth size 10 and random state 42


Porcentage_variability vs. Chrom_num
Chrom_num
Porcentage_variability
0.0
0.1
0.2
0.3
0.4
20
30
40
50
60
70
80
Order
Carnivora
Cetartiodactyla
Chiroptera
Eulipotyphla
Perissodactyla
Primates
Rodentia

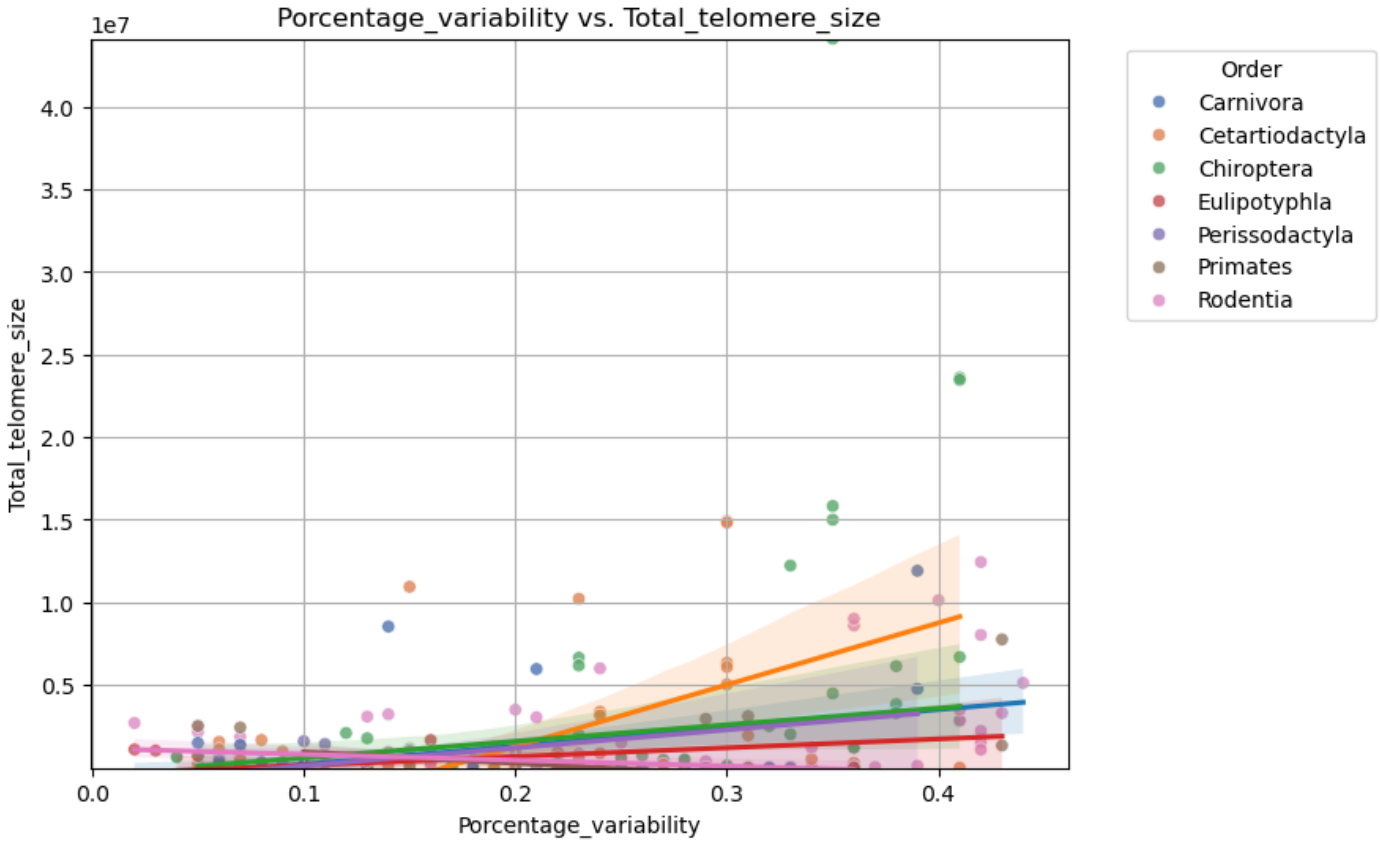
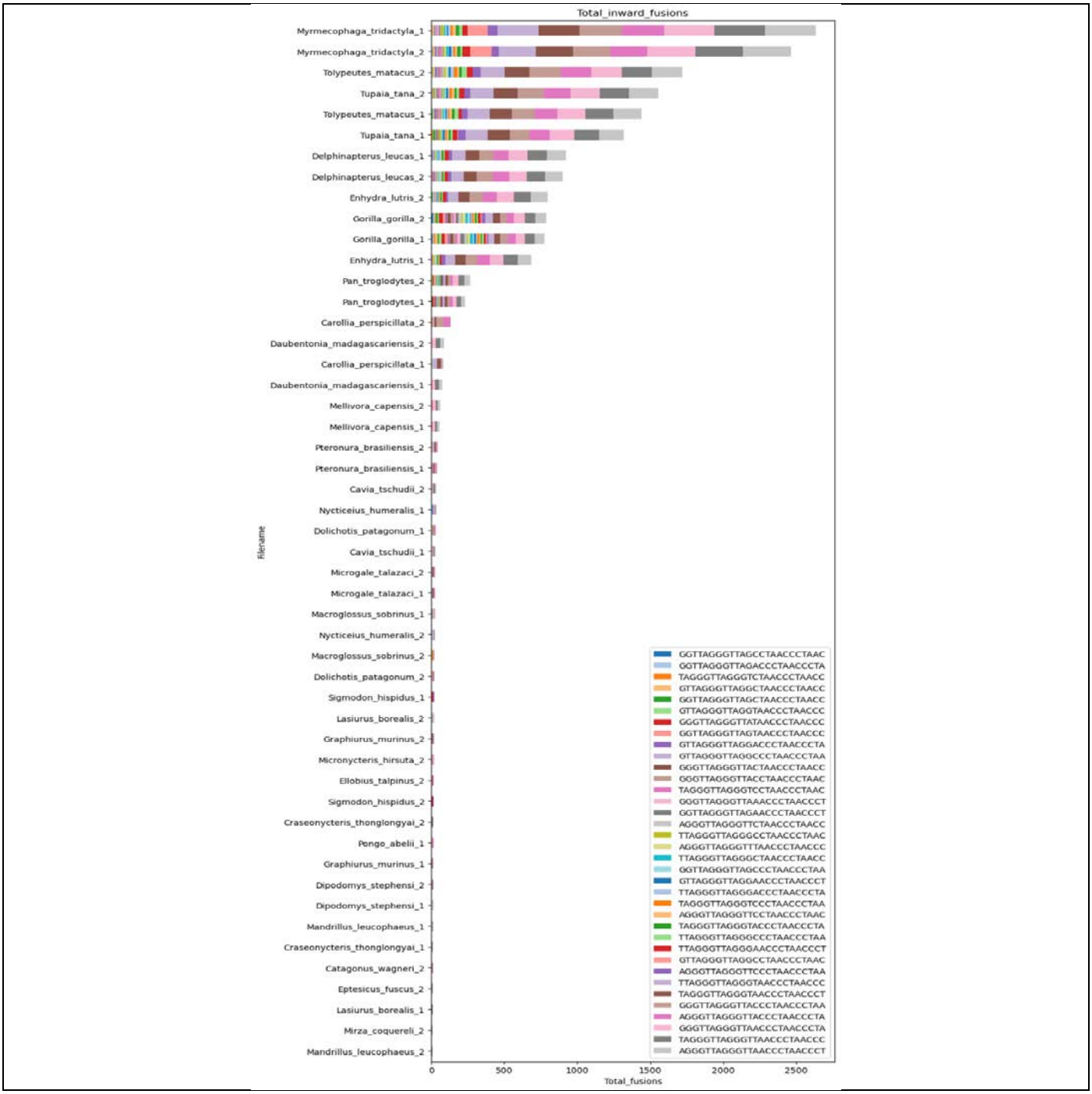
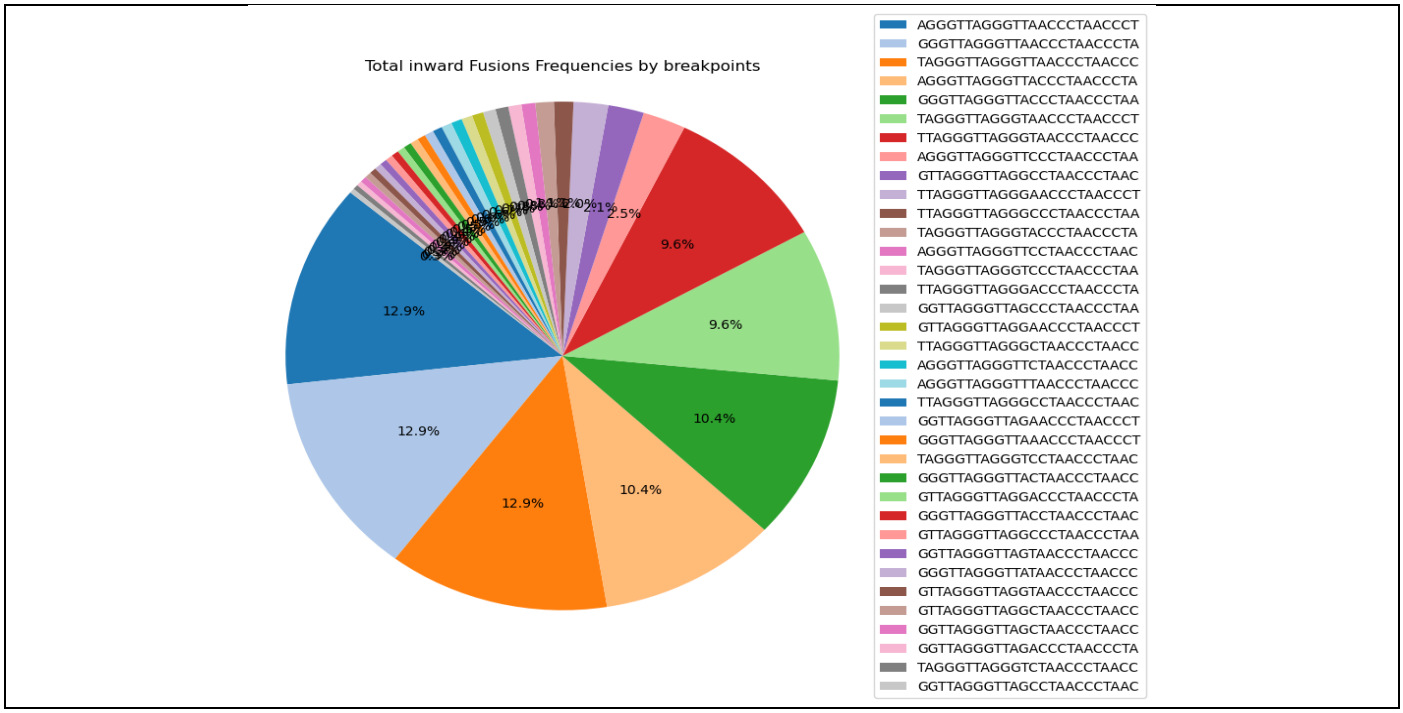
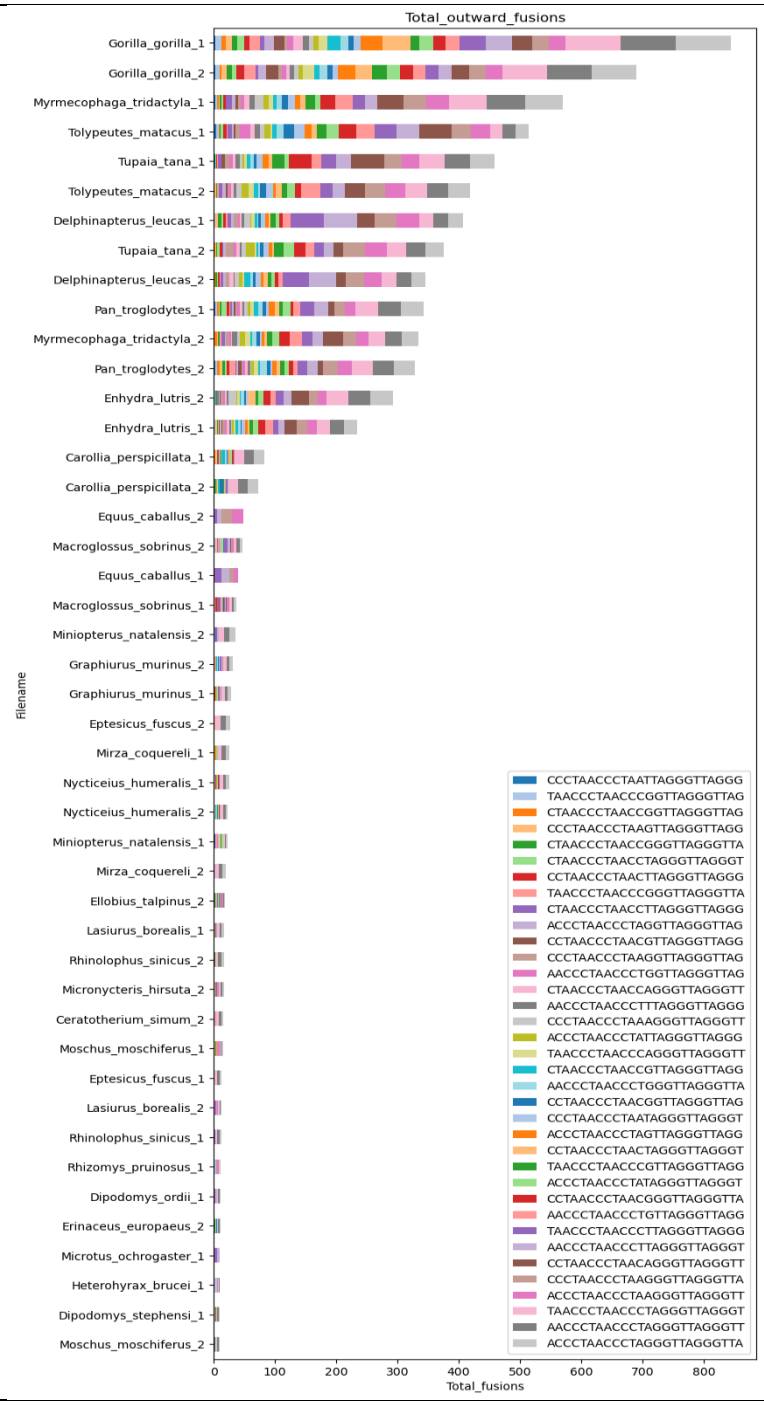


Figure S 10 Non-categorical variables analysed by order please see Analysis_notebook.ipynb

Breakpoints tables


Total_inward_fusions
Myrmecophaga_tridactyla_1
Myrmecophaga_tridactyla_2
Tolypeutes_matacus_2
Tupaia_tana_2
Tolypeutes_matacus_1
Tupaia_tana_1
Delphinapterus_leucas_1
Delphinapterus_leucas_2
Enhydra_lutris_2
Gorilla_gorilla_2
Gorilla_gorilla_1
Enhydra_lutris_1
Pan_troglodytes_2
Pan_troglodytes_1
Carollia_perspicillata_2
Daubentonia_madagascariensis_2
Carollia_perspicillata_1
Daubentonia_madagascariensis_1
Mellivora_capensis_2
Mellivora_capensis_1
Pteronura_brasiliensis_2
Pteronura_brasiliensis_1
Cavia_tschudii_2
Nycticeius_humeralis_1
Dolichotis_patagonum_1
Cavia_tschudii_1
Microgale_talazaci_2
Microgale_talazaci_1
Macroglossus_sobrinus_1
Nycticeius_humeralis_2
Macroglossus_sobrinus_2
Dolichotis_patagonum_2
Sigmodon_hispidus_1
Lasiurus_borealis_2
Graphiurus_murinus_2
Micronycteris_hirsuta_2
Ellobius_talpinus_2
Sigmodon_hispidus_2
Craseonycteris_thonglongyai_2
Pongo_abelii_1
Graphiurus_murinus_1
Dipodomys_stephensi_2
Dipodomys_stephensi_1
Mandrillus_leucophaeus_1
Craseonycteris_thonglongyai_1
Catagonus_wagneri_2
Eptesicus_fuscus_2
Lasiurus_borealis_1
Mirza_coquereli_2
Mandrillus_leucophaeus_2
Filename
0
500
1000
1500
2000
2500
Total_fusions
GGTTAGGGTTAGCCTAACCCTAAC
GGTTAGGGTTAGACCCTAACCCTA
TAGGGTTAGGGTCTAACCCTAACC
GTTAGGGTTAGGCTAACCCTAACC
GGTTAGGGTTAGCTAACCCTAACC
GTTAGGGTTAGGTAACCCTAACCC
GGGTTAGGGTTATAACCCTAACCC
GGTTAGGGTTAGTAACCCTAACCC
GTTAGGGTTAGGACCCTAACCCTA
GTTAGGGTTAGGCCCTAACCCTAA
GGGTTAGGGTTACTAACCCTAACC
GGGTTAGGGTTACCTAACCCTAAC
TAGGGTTAGGGTCCTAACCCTAAC
GGGTTAGGGTTAAACCCTAACCCT
GGTTAGGGTTAGAACCCTAACCCT
AGGGTTAGGGTTCTAACCCTAACC
TTAGGGTTAGGGCCTAACCCTAAC
AGGGTTAGGGTTTAACCCTAACCC
TTAGGGTTAGGGCTAACCCTAACC
GGTTAGGGTTAGCCCTAACCCTAA
GTTAGGGTTAGGAACCCTAACCCT
TTAGGGTTAGGGACCCTAACCCTA
TAGGGTTAGGGTCCCTAACCCTAA
AGGGTTAGGGTTCCTAACCCTAAC
TAGGGTTAGGGTACCCTAACCCTA
TTAGGGTTAGGGCCCTAACCCTAA
TTAGGGTTAGGGAACCCTAACCCT
GTTAGGGTTAGGCCTAACCCTAAC
AGGGTTAGGGTTCCCTAACCCTAA
TTAGGGTTAGGGTAACCCTAACCC
TAGGGTTAGGGTAACCCTAACCCT
GGGTTAGGGTTACCCTAACCCTAA
AGGGTTAGGGTTACCCTAACCCTA
GGGTTAGGGTTAACCCTAACCCTA
TAGGGTTAGGGTTAACCCTAACCC
AGGGTTAGGGTTAACCCTAACCCT


Total inward Fusions Frequencies by breakpoints
12.9%
12.9%
12.9%
10.4%
10.4%
9.6%
9.6%
2.5%
AGGGTTAGGGTTAACCCTAACCCT
GGGTTAGGGTTAACCCTAACCCTA
TAGGGTTAGGGTTAACCCTAACCC
AGGGTTAGGGTTACCCTAACCCTA
GGGTTAGGGTTACCCTAACCCTAA
TAGGGTTAGGGTAACCCTAACCCT
TTAGGGTTAGGGTAACCCTAACCC
AGGGTTAGGGTTCCCTAACCCTAA
GTTAGGGTTAGGCCTAACCCTAAC
TTAGGGTTAGGGAACCCTAACCCT
TTAGGGTTAGGGCCCTAACCCTAA
TAGGGTTAGGGTACCCTAACCCTA
AGGGTTAGGGTTCCTAACCCTAAC
TAGGGTTAGGGTCCCTAACCCTAA
TTAGGGTTAGGGACCCTAACCCTA
GGTTAGGGTTAGCCCTAACCCTAA
GTTAGGGTTAGGAACCCTAACCCT
TTAGGGTTAGGGCTAACCCTAACC
AGGGTTAGGGTTCTAACCCTAACC
AGGGTTAGGGTTTAACCCTAACCC
TTAGGGTTAGGGCCTAACCCTAAC
GGTTAGGGTTAGAACCCTAACCCT
GGGTTAGGGTTAAACCCTAACCCT
TAGGGTTAGGGTCCTAACCCTAAC
GGGTTAGGGTTACTAACCCTAACC
GTTAGGGTTAGGACCCTAACCCTA
GGGTTAGGGTTACCTAACCCTAAC
GTTAGGGTTAGGCCCTAACCCTAA
GGTTAGGGTTAGTAACCCTAACCC
GGGTTAGGGTTATAACCCTAACCC
GTTAGGGTTAGGTAACCCTAACCC
GTTAGGGTTAGGCTAACCCTAACC
GGTTAGGGTTAGCTAACCCTAACC
GGTTAGGGTTAGACCCTAACCCTA
TAGGGTTAGGGTCTAACCCTAACC
GGTTAGGGTTAGCCTAACCCTAAC


Total_outward_fusions
Gorilla_gorilla_1
Gorilla_gorilla_2
Myrmecophaga_tridactyla_1
Tolypeutes_matacus_1
Tupaia_tana_1
Tolypeutes_matacus_2
Delphinapterus_leucas_1
Tupaia_tana_2
Delphinapterus_leucas_2
Pan_troglodytes_1
Myrmecophaga_tridactyla_2
Pan_troglodytes_2
Enhydra_lutris_2
Enhydra_lutris_1
Carollia_perspicillata_1
Carollia_perspicillata_2
Equus_caballus_2
Macroglossus_sobrinus_2
Equus_caballus_1
Macroglossus_sobrinus_1
Miniopterus_natalensis_2
Graphiurus_murinus_2
Graphiurus_murinus_1
Eptesicus_fuscus_2
Mirza_coquereli_1
Nycticeius_humeralis_1
Nycticeius_humeralis_2
Miniopterus_natalensis_1
Mirza_coquereli_2
Ellobius_talpinus_2
Lasiurus_borealis_1
Rhinolophus_sinicus_2
Micronycteris_hirsuta_2
Ceratotherium_simum_2
Moschus_moschiferus_1
Eptesicus_fuscus_1
Lasiurus_borealis_2
Rhinolophus_sinicus_1
Rhizomys_pruinosus_1
Dipodomys_ordii_1
Erinaceus_europaeus_2
Microtus_ochrogaster_1
Heterohyrax_brucei_1
Dipodomys_stephensi_1
Moschus_moschiferus_2
Filename
0
100
200
300
400
500
600
700
800
Total_fusions
CCCTAACCCTAATTAGGGTTAGGG
TAACCCTAACCCGGTTAGGGTTAG
CTAACCCTAACCGGTTAGGGTTAG
CCCTAACCCTAAGTTAGGGTTAGG
CTAACCCTAACCGGGTTAGGGTTA
CTAACCCTAACCTAGGGTTAGGGT
CCTAACCCTAACTTAGGGTTAGGG
TAACCCTAACCCGGGTTAGGGTTA
CTAACCCTAACCTTAGGGTTAGGG
ACCCTAACCCTAGGTTAGGGTTAG
CCTAACCCTAACGTTAGGGTTAGG
CCCTAACCCTAAGGTTAGGGTTAG
AACCCTAACCCTGGTTAGGGTTAG
CTAACCCTAACCAGGGTTAGGGTT
AACCCTAACCCTTTAGGGTTAGGG
CCCTAACCCTAAAGGGTTAGGGTT
ACCCTAACCCTATTAGGGTTAGGG
TAACCCTAACCCAGGGTTAGGGTT
CTAACCCTAACCGTTAGGGTTAGG
AACCCTAACCCTGGGTTAGGGTTA
CCCTAACCCTAACGGTTAGGGTTAG
CCCTAACCCTAATAGGGTTAGGGT
ACCCTAACCCTAGTTAGGGTTAGG
CCTAACCCTAACTAGGGTTAGGGT
TAACCCTAACCCGTTAGGGTTAGG
ACCCTAACCCTATAGGGTTAGGGT
CCTAACCCTAACGGGTTAGGGTTA
AACCCTAACCCTGTTAGGGTTAGG
TAACCCTAACCCTTAGGGTTAGGG
AACCCTAACCCTTAGGGTTAGGGT
CCTAACCCTAACAGGGTTAGGGTT
CCCTAACCCTAAGGGTTAGGGTTA
ACCCTAACCCTAAGGGTTAGGGTT
TAACCCTAACCCTAGGGTTAGGGT
AACCCTAACCCTAGGGTTAGGGTT
ACCCTAACCCTAGGGTTAGGGTTA

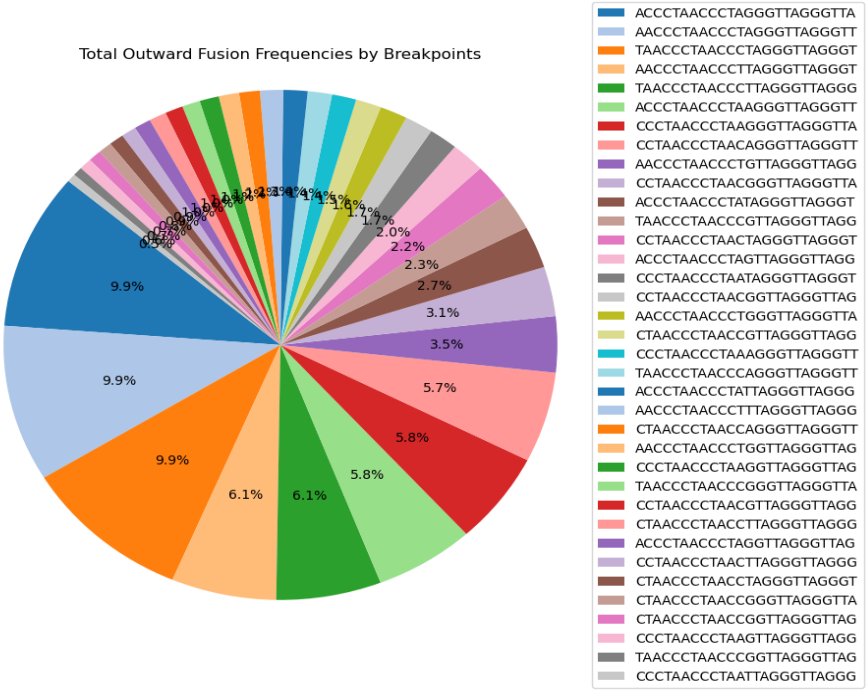


Figure S 11 breakpoints plots