Comparative genomics of Tandem Repeat variation in apes

Carolina L. Adam[1*], Joana Rocha[2], Peter Sudmant[2,3*+], Rori Rohlfs[1,4*+]

[1]Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403, USA
[2]Department of Integrative Biology, University of California - Berkeley, Berkeley, CA 94720, USA
[3]Center for Genomics and Systems Biology, New York University, New York, NY 10003, USA
[4]School of Computer and Data Sciences, University of Oregon, Eugene, OR 97403, USA

* Corresponding authors e-mails: carolinaladam@gmail.com, psudmant@berkeley.edu, rori@uoregon.edu
+ These authors contributed equally

Abstract

Tandem repeats (TRs) are highly mutable DNA elements that comprise nearly 8% of the human genome and influence gene regulation, protein coding, and disease. Despite their functional importance, our understanding of TR evolution remains limited, as their repetitive nature has hindered accurate sequencing, annotation, and cross-species comparison. As a result, we lack a population-aware evolutionary framework to quantify TR conservation, divergence, and mutational dynamics across species. Here, using telomere-to-telomere (T2T) reference genomes for seven ape species and population-scale long reads for humans and chimpanzees, we generated a comprehensive comparative catalog of STRs and VNTRs. We identified over 3 million TR loci per ape genome and nearly 2 million homologous loci between humans and chimpanzees. TR diversity and conservation are strongly structured by genomic context, with coding and untranslated regions exhibiting reduced polymorphism and divergence, while intronic and intergenic regions show elevated variability. Heterozygosity varies systematically across species, motif lengths, and functional categories, and mutation rates show strong concordance between indirect and pedigree-based estimates. Using a divergence-diversity ratio framework, we identified TRs under extreme evolutionary regimes that are enriched in genes involved in nervous system development, synaptic function, and cell signaling. Together, these results establish a population and species-resolved framework for studying TR evolution and interpreting TR variation in functional contexts.

Main

Tandem repeats (TRs) are ubiquitous across metazoans and account for nearly 8% of the human genome (Nurk et al., 2022). Their repetitive structure makes them prone to replication slippage, leading to mutation rates that are orders of magnitude higher than single-nucleotide variants, or SNVs (Fan & Chu, 2007). Tandem repeat variants (TRVs) are more likely than point mutations to impact functional traits, and human medical genetics has repeatedly demonstrated their impact on development and disease, from neurodegeneration to cancer (Erwin et al., 2023; Schloissnig et al., 2024; Song et al., 2018). These properties make TRs strong candidates as evolutionary fuel for rapid adaptation (Gymrek, 2017; Horton et al., 2023; Y. Huang et al., 2025).

Hypotheses of the adaptive role of TRs have circulated for decades, and evidence has long pointed to their evolutionary significance (Chavali et al., 2017; Fondon & Garner, 2004; Vinces et al., 2009; Zhou et al., 2014). Yet, technical limitations in sequencing and genotyping long repetitive DNA have hindered systematic evaluation across species. This landscape has now shifted. Advances in long-read sequencing and telomere-to-telomere (T2T) assemblies resolved previously inaccessible repeat regions (Jarvis et al., 2022; Nurk et al., 2022) and are fundamentally reshaping our understanding of genome architecture and structural variation (Rocha et al., 2024). As a result, evolutionary TR studies continue to emerge. Human-specific TRs and TR expansions have been linked to differential gene expression and enhancer activity during primate evolution, particularly in the brain, highlighting their potential contribution to lineage-specific phenotypes (Kim et al., 2019; Liu & Tian, 2025; Sulovari et al., 2019). Still, we lack a framework that leverages long-read TR genotypes both between and within species, an approach that has proven transformative in SNP-based evolutionary studies (Leffler et al., 2013; Prado-Martinez et al., 2013). Thus, understanding how these repetitive regions evolve provides insight into the molecular processes underlying the emergence of lineage-specific traits.

Here, we leveraged T2T reference assemblies and long-read, assembly-level sequencing data, including newly sequenced chimpanzee genomes (Rocha et al. *in prep*), to chart the genomic landscape of TRs across seven ape genomes. We identified hundreds of thousands of homologous TR loci between humans and non-human apes, revealing broad patterns of TR retention and constraint across the evolutionary timescale that mirror the ape phylogeny. To capture the intraspecific diversity and better understand the evolutionary dynamics at play, we next examined assembly-level population data from 46 humans and 23 chimpanzees.

Species-specific TR catalogs, homology to humans, and genomic distribution

TR motif length distributions and catalog sizes were largely similar across all seven species (Figure 1a and b). Among short motif sizes (< 7 bp), mononucleotide repeats were the most abundant, while trinucleotides were the least abundant, consistent with previous studies (Sharma & Sowpati, 2025; Srivastava et al., 2019; Verbiest et al., 2023). Notably, hexanucleotide repeats were abundant in Sharma & Sowpati (2025) and Srivastava et al. (2019) but were strongly depleted in Verbiest et al. (2023), whereas we observed a mild reduction in hexanucleotides (Figure 1c). These discrepancies in motif length distribution between studies likely stem from differences in TR identification and filtering parameters. For instance, other studies have focused on perfect repeats, with 100% sequence constancy (Srivastava et al., 2019; Verbiest et al., 2023), or nearly perfect repeats, with ≥90% constancy (Sharma & Sowpati, 2025). In contrast, our analysis allowed for lower constancy (>60%). This approach generates a more inclusive TR catalog and may capture more variable repeats that stricter thresholds would have excluded. Motifs larger than 20 bp were considerably less common and have not been systematically explored in prior studies. Nonetheless, we observe distinct peaks in large motif sizes. When queried in the Dfam database, these motifs matched known repetitive elements, including Alu elements and VNTRs embedded within composite retrotransposons (Figure 1a). The GC content of TR motifs also varied with motif lengths, but showed broadly consistent patterns across species (Extended Figure 2). Mononucleotide repeats were almost exclusively A/T, resulting in GC content below 1%. In contrast, motifs 2-20 bp in length averaged ~30% GC content, increasing to ~50% in motifs > 20 bp.

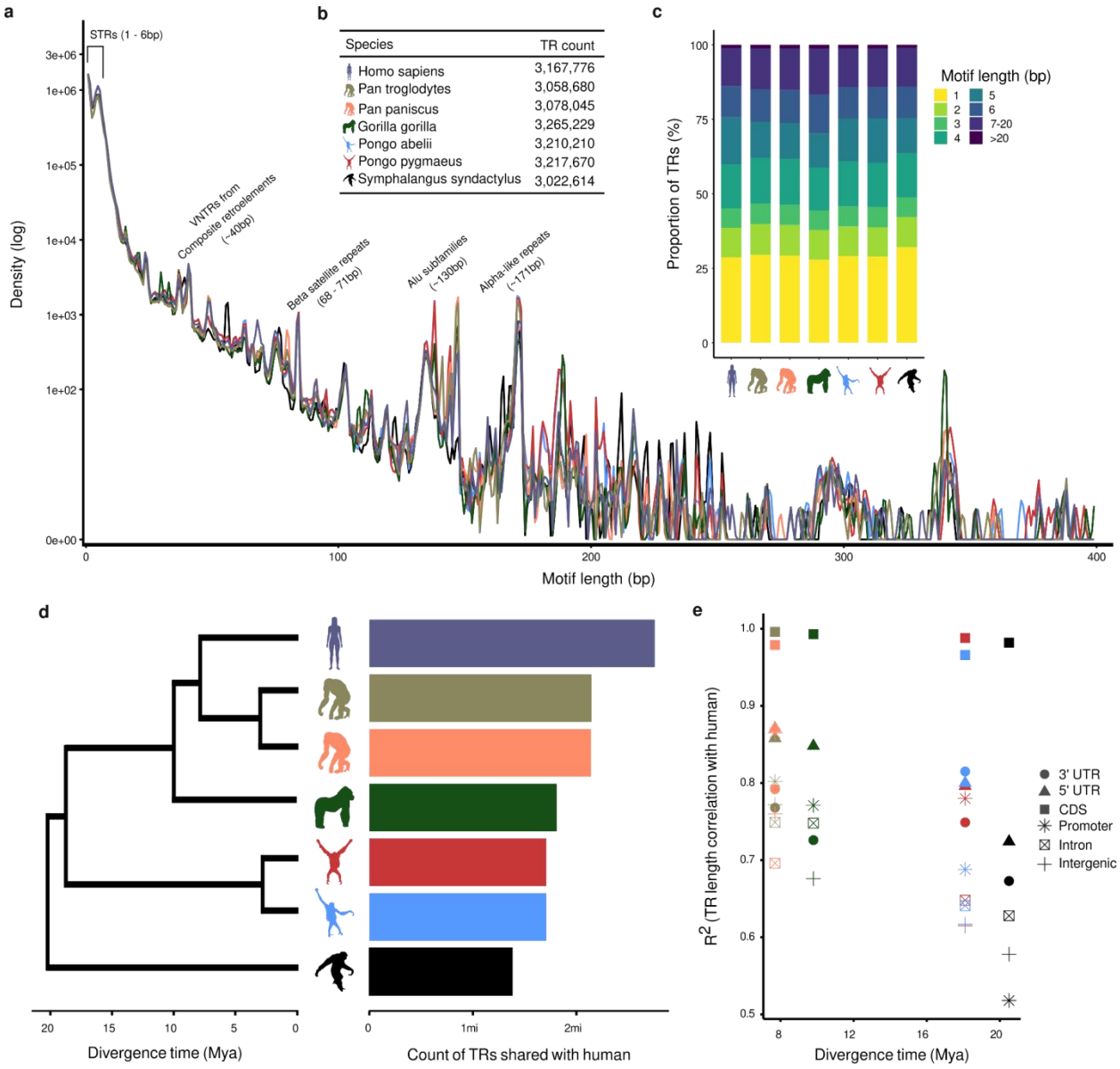


Figure 1. **Tandem repeats across T2T ape genomes.** a) Density plot showing the distribution of TRs by motif length across ape genomes. Labeled peaks indicate TRs with motif lengths >20 bp that show strong matches to entries in the Dfam database. b) Table with TR count per species. c) Stacked barplot showing the proportional distribution of TR motif lengths across seven telomere-to-telomere (T2T) ape genomes. d) Phylogenetic tree of the seven ape genomes with a barplot showing each species' number of homologous TRs with humans. The top bar represents the total number of TRs identified in the human genome. e) Scatterplot showing the relationship between divergence time (in millions of years) and the coefficient of determination (R^2) for reference TR length comparisons between humans and each ape species.

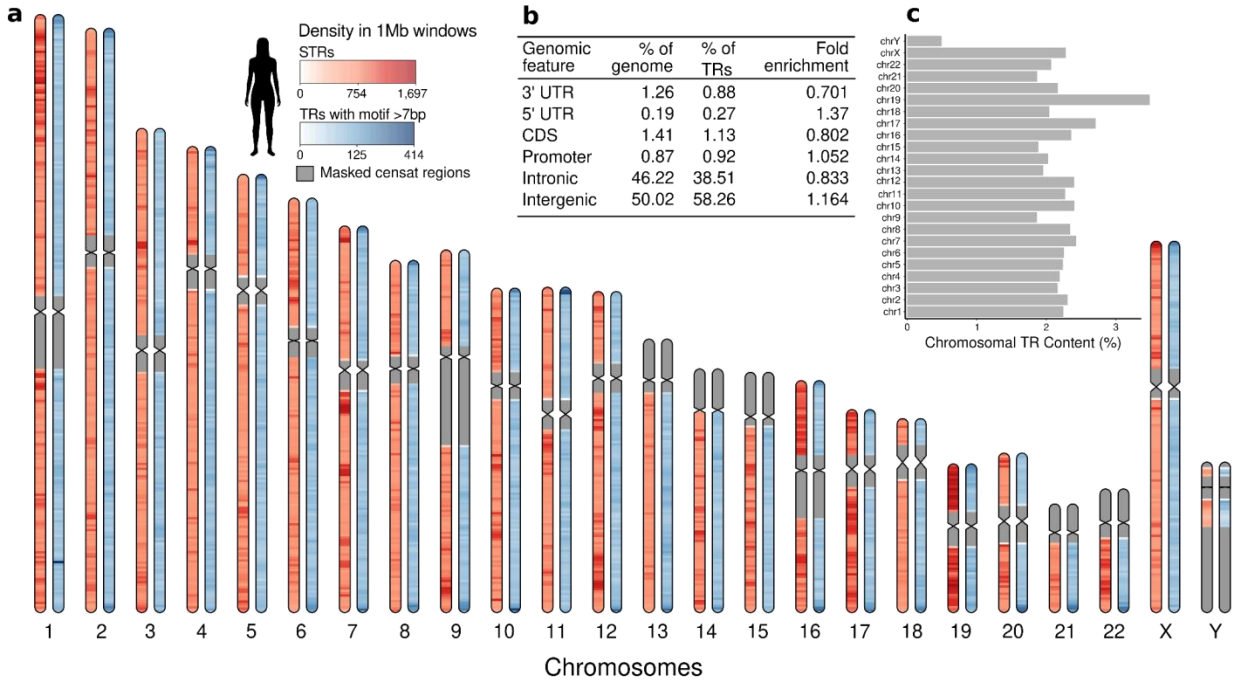
The abundance of homologous TRs between the human and each of the non-human T2T ape genomes reflects the phylogenetic distances, with the *Pan* genus—our closest living relatives, sharing a common ancestor with humans approximately 7.7 Mya (Shao et al., 2023)—retaining the largest overlap (Figure 1d). Comparison of TR length between species, after removing

centromere regions, reveals high conservation among TRs located within coding sequence (CDS) and 5' untranslated regions (UTRs) (Figures 1e and Extended Figure 3). Levels of conservation also followed the ape phylogeny, with correlations between homologous TR lengths decreasing as the time since the most recent common ancestor (TMRCM) increased. This trend is especially pronounced for TRs in promoters, introns, and intergenic regions, indicating reduced constraint compared to TRs in coding sequences and UTRs. Indeed, other studies have found that most protein-coding TRs are deeply conserved across mammals, with some extending across vertebrates and even to the base of eukaryotes (Schaper et al., 2014). This indicates that maintenance of TR length in coding regions may be crucial for functional protein structure (Madsen et al., 2008; Usdin, 2008) and in 5' UTRs to maintain translational regulation (Churbanov, 2005) and RNA structural elements (Byeon et al., 2021a). Variants in UTRs are known to disrupt transcription initiation and, in some cases, contribute to disease (Wieder et al., 2025).

Next, we estimated TR density across chromosomes in 1Mb windows for all species (Figure 2 and Extended Figure 4). Short tandem repeats (STRs, motifs of 1-6 bp) were generally uniformly distributed along the chromosome lengths. Chromosome 19 had the highest proportion of TRs relative to its length, followed by chromosome 17 (Figure 2a and c), a pattern consistent with previous reports (Grimwood et al., 2004; Subramanian et al., 2003). Although chromosome Y is known to harbor extensive repetitive regions, the majority consists of satellite repeats (Rhie et al., 2023), which were excluded during the catalog filtering. Thus, chromosome Y shows the lowest density of non-cenSat TRs. Repeats with motifs ≤7 bp showed elevated density in human telomeres (Figure 2a), consistent with reports of VNTR enrichment in the telomeres, particularly those with motifs ≤15 bp (Linthorst et al., 2020). In non-human primates, distinctive repeat structures have been described at chromosome ends, which are absent in human chromosomes: siamang gibbons harbor telomeric alpha satellites (Koga et al., 2012), while gorillas and certain chromosomes of chimpanzees and bonobos contain StSat (Ahmad et al., 2020). Because these repeat classes were filtered out from the catalogs, they are absent from the ideogram density plots for these species (Extended Figure 4a and f).

Considering their representation in different genomic elements, TRs were depleted in introns, 3' UTR, and coding regions, while being enriched in intergenic and 5' UTR regions (Figure 2b). When stratified by motif length, distinct distributions emerged across genomic

features (Extended Figure 5, Supplemental Table S1). Mono and dinucleotide repeats were strongly depleted in coding regions (0.01 and 0.05-fold, respectively), while repeats with motif lengths in multiples of three were overrepresented: for instance, tri and hexanucleotide repeats showed 5.2-fold and 2.7-fold enrichment, respectively. This pattern reflects the coding-frame constraint, where expansions or contractions in multiples of three preserve the reading frame and are therefore tolerated (Srivastava et al., 2019; Subramanian et al., 2003). A similar trend was observed in the 5' UTRs, which also showed depletion in monomers and dimers (0.1 and 0.3-fold), and enrichment of trimers (4.3-fold enrichment), a trend also observed in a large STR population panel (Huang et al., 2025).



Genomic feature	% of genome	% of TRs	Fold enrichment
3' UTR	1.26	0.88	0.701
5' UTR	0.19	0.27	1.37
CDS	1.41	1.13	0.802
Promoter	0.87	0.92	1.052
Intronic	46.22	38.51	0.833
Intergenic	50.02	58.26	1.164

Figure 2. **Genomic TR distribution in CHM13.** a) Ideogram of the human CHM13 genome showing the density of short tandem repeats (STRs) and TRs with motif length >7 bp across non-overlapping 1 Mb windows. Repeat density is plotted along each chromosome. b) Table showing the distribution of TRs across major annotated genomic elements and fold-enrichment indicating the magnitude of deviation of TR distribution from genome expectation for each feature. FDR-corrected p-values < 1e-16 for all comparisons. c) Barplot showing the proportion of each chromosome contained in TRs based on the CHM13 genome.

TR genotypes and length variation

After QC and filtering (see Methods), we identified 1,905,903 homologous TRs between

humans and chimpanzees. Chimpanzees showed higher TR heterozygosity than humans across all genomic features (Figure 3a), consistent with long-standing evidence that humans harbor lower overall genetic diversity (Li & Sadler, 1991; Prado-Martinez et al., 2013). In both species, heterozygosity was reduced in coding regions and 5' UTRs, likely due to stronger functional constraints. This pattern aligns with previous observations that coding TRs are both rarer and less polymorphic than those in noncoding regions (Huang et al., 2016; Press et al., 2018; Rockman & Wray, 2002; Willems et al., 2014).

Heterozygosity was also variable across motif lengths, highest for mono- and dinucleotides, then steadily decreased up to motifs of 20 bp, before increasing again for larger motifs (Figure 3b). A similar decrease in diversity with increasing motif size was reported by Jam et al., n.d., thought their analysis, limited to motifs ≤6 bp, missed the rise in heterozygosity at larger motifs. This overall trend holds across genomic features, except for coding regions, which have low heterozygosity also for mono and dinucleotide repeats (Extended Figure 6). This deviation from the overall trend showcase that TRs in coding regions have depleted variability even when harboring motif lengths known to have elevated mutation rates (Fan & Chu, 2007).

Next, we estimated per-generation mutation rates using Goldstein's genetic distance (du^2) framework under the stepwise mutation model (Goldstein et al., 1995) and compared these indirect estimates with direct *de novo* mutation rates from human trio-based data (Porubsky et al., 2025). The approaches were highly concordant (Figure 3c), showing higher mutation rates at mono- and dinucleotide TRs, with a subsequent increase that plateaued near 20 bp. The subsequent rise in diversity among longer motifs suggests a shift in the dominant mutational process, potentially from replication slippage in short motifs to recombination or gene conversion-associated mechanisms for longer arrays (Lai, 2003). These mutation rate patterns closely mirror our heterozygosity estimates across motif lengths (Figure 3a-c), underscoring how mutational dynamics shape both standing variation and *de novo* changes across motif lengths and genomic features.

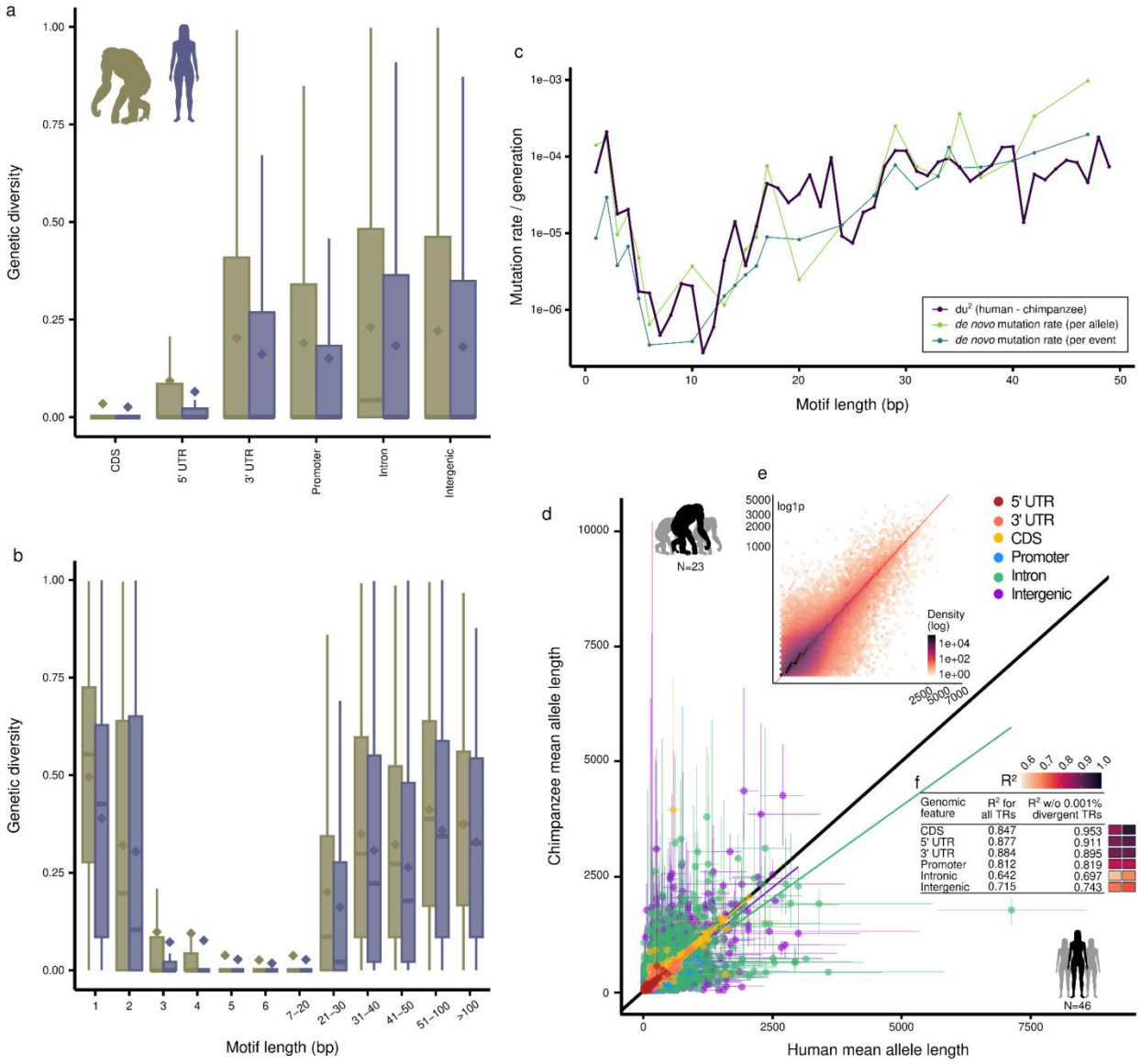


Figure 3. **Comparative analyses of homologous TRs between humans and chimapzness.** Expected heterozygosity a) across different genomic features and b) across motif lengths. c) Estimates of mutation rate across motif lengths. d) Scatterplot of mean TR allele lengths between humans (x-axis) and chimpanzees (y-axis) for TRs shared between species, grouped by genomic feature. Each point represents a single TR locus, with color indicating the genomic annotation in the CHM13 genome. Error bars represent the 90th percentile of the length distribution in each species. e) Heatmap of the same TR loci showing the joint distribution of human and chimpanzee mean allele lengths. Color intensity reflects the density of TRs within each bin (50 bins), highlighting regions of agreement and divergence in TR length between species. e) Correlation between human and chimpanzee TR length grouped by genomic feature.

Of the 1,905,903 TRs shared between the two species, 1,033,666 (54.2%) were invariant (iTRs), i.e. monomorphic across samples, and the remaining 872,237 (45.7%) were variable (TRVs). Across genomic features, the largest discrepancy between iTR and TRV proportions

was observed in coding and 5' UTRs (Extended Figure 7a). 89.7% of coding TRs were invariant, reflecting known functional constraint. Strikingly, 74.2% of 5' UTR TRs were also invariant, compared to roughly 55% and 50% of 3' UTR and intronic TRs, respectively. This unexpected degree of constraint in 5' UTRs is consistent with emerging evidence for extreme evolutionary conservation in subsets of vertebrate 5' UTRs, which is linked to RNA-mediated translational control (Byeon et al., 2021). iTRs also tend to have larger motif sizes than TRVs (Extended Figure 7b). The majority of homologous TRs were short—62.1% below the mean allele length in both species (~22 bp)—and showed high length correlation between species (Figure 3d-e), which was stronger in coding and UTR regions (Figure 3a and d), confirming reference length results (Extended Figure 1e). This strong cross-species concordance in TR lengths mirrors patterns from assembly-level comparisons of single human and chimpanzee genomes that reported only a 0.02 bp difference in orthologous STRs between the two species (Kronenberg et al., 2018). Within-species, TRs with species-specific expansions also exhibited elevated variation in the species where the expansion occurred (Extended Figure 8), reflecting an increased mutational potential of longer arrays. This is consistent with recombination-mediated mutation events capable of generating extensive repeat gains that exceed the scale expected from replication slippage alone (Li et al., 2002; Richard & Pâques, 2000). Thus, expansions contribute both to between-species divergence and within-species polymorphism.

Divergence-diversity ratio and functional context of extreme TRs

We calculated the divergence-diversity ratio (*D*) for all TRVs in humans and chimpanzees to identify loci with extreme patterns of divergence between species relative to within-species variation. *D* values were highly correlated between species (Figure 4a), indicating that the majority of TRVs are subject to similar mutational dynamics and selective pressures between humans and chimpanzees. To explore the selective regimes leading to these patterns, we focused on the 3 types of genic TRVs: those with strong signatures of divergent selection in both species (high *D*), those with strong signatures of divergence in a single species (high *D*_human or high *D*_chimp) and those with strong signatures of balancing selection in both species (low *D*) (Supplemental Table S2). Across categories, genic TRVs with signatures of divergent selection were longer and with larger motifs (Wilcoxon rank-sum test, FDR-adjusted $p < 0.0001$; Extended Figure 9a-d). While genic TRVs with signatures of balancing selection were also

longer than background TRVs (Wilcoxon rank-sum test, FDR-adjusted $p < 0.001$), no significant difference was observed for motif length (Wilcoxon rank-sum test, FDR-adjusted $p = 0.523$). Highly divergent TRVs also display higher GC content when compared to a set of background TRVs with the same TR length and motif length distribution (Wilcoxon rank-sum test, FDR-adjusted $p < 0.0001$) and a high proportion is located in coding regions (Extended Figure 9e).

To investigate the biological processes associated with TRVs with extreme divergence and diversity, we performed Gene Ontology (GO) enrichment analysis on the genes intersecting the top 1,000 genic TRVs from each category. Despite their contrasting evolutionary patterns, genes containing TRVs with either extreme divergence or diversity showed enrichment for biological process GO terms involved in similar pathways. While TR-containing genes are modestly enriched for broad developmental processes, primarily involving cellular component organization and anatomical structure morphogenesis, relative to the set of human genes (~1.2-fold; Extended Figure 10), genes with TRVs in the top ratio categories show >2-fold enrichment relative to all TR genes, converging on more specific pathways underlying nervous system development, synaptic organization and cell signaling (Figure 4b and c).

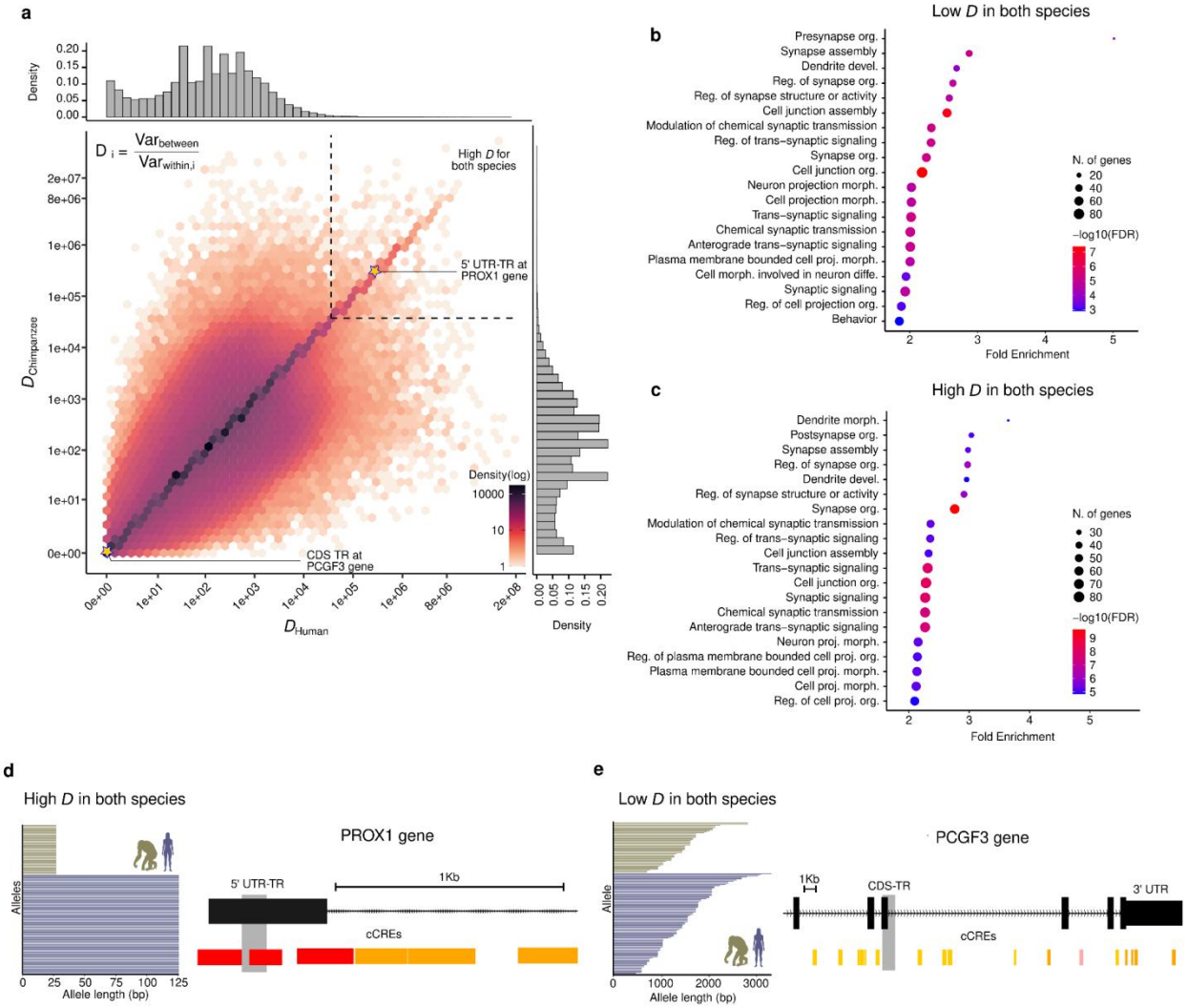


Figure 4. **Divergence–diversity landscapes of tandem repeats in humans and chimpanzees.** a) Heatmap showing the joint distribution of TR Divergence-Diversity Ratios (*D*) in humans and chimpanzees. Marginal histograms display the distribution of ratios for each species. Black dashed line shows the boundaries of high *D* in both species. Stars indicate example TRs belonging to extreme ratio categories in both species. b-c) Gene Ontology enrichment analysis for Biological Processes terms associated with genes intersecting the top 1,000 genic TRs with b) low *D* in both species and c) high *D* in both species. The set of all TR containing genes was used as background. d-e) Genomic location and allele length distribution for two example TRs classified as d) low *D* in both species and e) high *D* in both species. Grey vertical bars represent TR location. Candidate cis-regulatory elements (cCREs) indicate promoter-like signature (red), proximal (orange) and distal (yellow) enhancer-like signature, and DNase-H3K4me3 elements (pink).

The ten genic TRVs with the strongest signature for divergence (high *D*) are located in introns of functionally diverse genes, spanning roles in neural and sensory systems, epithelial integrity, and signal transduction (Supplemental Table S3). Among non-intronic TRVs within

the top 1,000 candidates, 13 of the 32 genes harboring TRVs in their CDS regions belonged to zinc-finger (ZNF) genes, a large and functionally diverse group involved in essential molecular processes, including transcriptional regulation, DNA repair, and cell migration (Cassandri et al., 2017). Several ZNFs are known to undergo rapid lineage-specific divergence and positive selection on DNA-binding domains between humans and chimpanzees (Jovanovic et al., 2021; Nowick et al., 2011), consistent with the high interspecies TRV divergence we observe. A compelling candidate is a TRV in the 5' UTR of PROX1, which exhibits pronounced between-species differences in mean allele length and is simultaneously highly homogeneous within each species (Figure 4e). PROX1 encodes a homeobox transcription factor essential for embryonic tissue development in mammals, including the formation of the central nervous system, eyes and the lymphatic system (Elsir et al., 2012).

Focusing on the 10 most extreme TRV candidates with an excess of diversity compared to divergence (low *D*), we recovered candidates located in genes related to immune function (Supplemental Table S3). Among them are PRKCE, which contributes to innate immune activation, with knockout mice showing impaired inflammatory responses and increased susceptibility to infections (Altman & Kong, 2016), XRCC4, a core DNA-repair factor required for B-cells to switch antibody classes (Soulas-Sprauel et al., 2007), and CALCR, a calcitonin receptor involved in bone metabolism with emerging roles in modulating inflammatory response and immune-associated signaling (Maleitzke et al., 2022; Wang et al., 2024). Considering non-intronic candidates in the top 1,000 low *D* category, we examined the TRV in the CDS region of PCGF3, which acts primarily as a transcriptional activator required for mesodermal differentiation (Zhao et al., 2017), also contributing to antiviral immunity by promoting interferon-responsive gene transcription (Da et al., 2024). This TRV shows high within-species diversity, with nearly every allele unique (Figure 4d). This extreme example of TRV diversity is consistent with the elevated SNP diversity observed in innate immune genes under balancing selection (Bitarello et al., 2018; Ferrer-Admetlla et al., 2008), suggesting that balancing selection related to immune response may act on this repeat.

We also consider TRVs with a high *D* in only one species, that is, where divergence is high compared to a low diversity in one species, but not compared to a higher diversity in the other species. This pattern may be consistent with either active directional selection or elevated mutation rate in the species with higher TRV diversity. Such genic TRVs with high *D*_chimp,

indicating high divergence between species with higher diversity in humans, were largely enriched for pathways associated with cell morphogenesis and nervous system development, particularly neurogenesis (Extended Figure 11a). The 10 highest-ranked genic TRVs are in introns of genes involved in cell signaling and intracellular trafficking, with many showing activity specifically in the brain (Supplemental Table S4). Among these genes are ADARB2, exclusively expressed in brain cells (Rodriguez De Los Santos et al., 2024), and DLGAP2, highly expressed in the striatum and associated with multiple brain disorders (Rasmussen et al., 2017). Considering non-intronic TRVs, one notable example is the VNTR located in the CDS region of MUC7, whose repeat copy-number polymorphism exhibits lineage-specific selective pressures tied to mucosal innate immunity (Xu et al., 2016). This study detected signatures of positive selection acting on MUC7 in the primate lineage compared to other mammals, based on analyses of non-repetitive coding region of the gene. In our dataset, this VNTR shows human-specific expansion and reduced within-species variation in chimpanzees, suggestion continued lineage-specific constraint or adaptation in this locus.

In contrast, high *D*_human TRVs, which are diverged between species with high chimpanzee diversity, showed significant GO enrichment for a single Biological Process pathway related to cell-cell adhesion. We further explored Cellular Component enrichment for these candidates, revealing a strong localization bias in neuronal synaptic compartments, particularly membrane-associated protein complexes within dendritic regions of neurons (Extended Figure 11b). The 10 highest-ranked genic TRVs are also located in introns, with gene function ranging RNA processing, cell signaling and cytoskeletal organization, with some showing brain-specific activity (Supplemental Table S4). Coding TRVs in this category include two associated with ZDHHC7, which regulates brain development and plasticity (Kerkenberg et al., 2021), and RTTN, required for embryonic axial rotation and successful organogenesis (Faisst et al., 2002).

Differential gene expression and TR divergence

Using RNA-seq expression data from Brawand et al., (2011), we analyzed 7,050 orthologous genes expressed in humans and chimpanzees across six tissues (brain, cerebellum, heart, kidney, liver and testis). While genes harboring TRVs in their promoters did not have generally greater expression divergence (Wilcoxon rank-sum test, FDR-adjusted $p > 0.05$ across

all tissues), we did find a significant but weak correlation between gene expression divergence and average divergence of TRV allele lengths (Pearson's r = 0.02, FDR-adjusted p = 2.2e-16; Extended Figure 12). This suggests that is not merely the presence of TRs that influences gene expression and divergence, but the variation in repeat length. Experimental studies in yeast indicate that TRs can modulate transcription in a non-linear, length-dependent manner, with optimal repeat lengths maximizing expression (Vinces et al., 2009). In humans, this trend is supported by fine-mapped eSTRs (FM-eSTRs), STRs statistically prioritized as causal for gene expression (Fotsing et al., 2019). FM-eSTRS show no consistent tendency for TR length to increase or decrease gene expression. Thus, divergence in repeat length is expected to translate into heterogeneous effects across loci, yielding subtle genome-wise associations.

Standard *limma* analysis (see Methods) identified 4,125 differentially expressed (DE) genes between humans and chimpanzees, 2,046 of which contained at least one TRV, but TRVs were not overrepresented among DE genes (Odds ratio = 0.84, p = 0.987). We next explored whether DE genes are enriched for TRVs previously associated with gene expression (eTRs), including FM-eSTRs (Fotsing et al., 2019) and expression-associated VNTRs (eVNTRs) (Eslami Rasekh et al., 2021). Consistently, DE genes were significantly more likely to harbor eTRs than non-DE genes (Odds ratio = 1.45, p = 0.012). Among these, 303 DE genes contained TRVs in the top D ratio categories, but no enrichment was observed (Odds ratio = 0.99, p = 0.518). Thus, differential expression between humans and chimpanzees is not broadly associated with TR variation. Instead, DE genes are preferentially enriched for TRs with prior evidence of regulatory function, highlighting a specific, rather than global, role of expression-associated repeats.

Discussion

Here, we present a comprehensive survey of TR variation across apes and highlight how mutational processes and selective pressures jointly shape repeat diversity and divergence. By integrating population-level long-read data for humans and chimpanzees, we achieved high-confidence genotyping of long TRs, extending the accessible spectrum of TR variation beyond what has been captured by short-read datasets. This design enabled not only characterization of genomic patterns of TR variation but also identification of locus-specific deviations in diversity-divergence ratios, providing insights into the selective forces shaping TR landscape in humans

and chimpanzees.

TR diversity and divergence bear strong signatures of stabilizing selection, particularly for functional variation. TRs in regions with strong functional constraints exhibit depleted polymorphism within species and reduced interspecies divergence. Constraint is also evident in 5' UTRs, consistent with vertebrate conservation in these regulatory regions (Byeon et al., 2021), suggesting that purifying selection extends beyond coding sequences. By contrast, intergenic, intronic, and to some extent promoter regions, harbor more dynamic TR landscapes, generating novel alleles that may contribute to species-specific traits (Press et al., 2018). This continuum, from constraint to flexible, mirrors patterns observed for other classes of genomic variants, including SNPs and indels (Yu et al., 2015). Mutational dynamics also vary across TRs, and mirrors patterns of standing genetic variation. Heterozygosity and mutation rates vary systematically with motif length, highest for mono- and dinucleotide repeats, then declining for intermediate length motifs before rising again in longer arrays. This pattern suggest that motif-dependent mutational mechanisms, potentially from strong replication slippage in short motifs to recombination or gene conversion-associated events in longer arrays, shape both standing variation and interspecies divergence across the genome.

TRVs at extremes of the diversity-divergence ratio, despite differing in motif properties and genomic context, converged functionally. In coding regions, highly divergent TRs were consistently observed in zinc finger genes, a family with documented lineage-specific positive selection, whereas highly diverse TRVs were enriched in immune-related genes, reflecting the well-established role of balancing selection in immunity (Minias & Vinkler, 2022). Nonetheless, both categories were enriched in genes involved in nervous system development, cell signalling, and synaptic processes, aligning with prior links between TR variation and human neurodevelopment and neurological disorders (Cui et al., 2025; Gymrek et al., 2016; Xiao et al., 2022). Further, genes containing expression-associated TRs (Eslami Rasekh et al., 2021; Fotsing et al., 2019) were more likely to be differentially expressed between humans and chimpanzees. Although overall TRV presence or extreme repeat divergence alone did not predict differential expression, divergence in TR length showed a weak but significant association with expression divergence, supporting the modulatory "tuning knob" model (King et al., 1997) in which TR variation influences regulatory evolution, likely on a context dependent manner (Fotsing et al., 2019).

A key limitation of our analysis is the lack of population-scale TR genotypes for non-human apes beyond chimpanzees. While T2T assemblies enable high-confidence, comprehensive catalog construction and homology inference across species, estimates of within-species diversity, mutation rate, and divergence-diversity ratios remain restricted to humans and chimpanzees. Future long-read population datasets from other apes will be essential to determine wheter the patterns observed here generalize across the ape phylogeny.

Altogether, our findings show that TR evolution reflects a balance between mutational dynamics and selective constraint that is strongly modulated by genomic and functional context. TRs in constrained regions, such as coding sequences or regulatory elements of essential genes, are stabilized by purifying selection, whereas TRs in regions with more permissive constraints, such as in immune-related loci, exhibit elevated diversity. Over evolutionary timescales, this interplay between stability and flexibility produces a heterogeneous landscape of conserved and divergent TRs, preserving genomic integrity while simultaneously generating substrate for adaptive change. By integrating population-scale long-read data with T2T reference assemblies, our study underscores the role of TRs as pervasive and dynamic components of genome evolution, shaping both functional constraint and evolutionary innovation across apes.

METHODS

Creating TR reference catalogs

The reference dataset comprises T2T genomes from seven ape species (Fig. 1b), obtained from the Telomere-to-Telomere Consortium (Nurk et al., 2022; Yoo et al., 2024). These genomes were generated using long-read, high-coverage PacBio HiFi sequencing (>50x) and Oxford Nanopore ultra-long reads (100 kb+ and >30x). We used the TRACK pipeline (Adam et al., 2024) to generate TR catalogs for each species from the T2T reference genomes. Specifically, TRs were identified using the Tandem Repeat Finder (Benson, 1999) and filtered based on total repeat length (<10 Kbp), copy number (>2.5), and constancy score, i.e., the percentage of matches between adjacent copies (>60%). Overlapping repeats were resolved by retaining the shortest motif, and motifs were normalized to their minimal repeat unit. Finally, we queried our catalog against the Dfam database to identify instances where TRs intersect known repetitive element families (Hubley et al., 2016).

After this initial characterization of the catalogs, we applied additional filtering to

exclude centromeric regions and regions containing alpha satellite DNA (cenSat) and subterminal satellites (StSat). Annotations were obtained from the CHM13 and T2T-Primate Consortium to identify and remove complex high-order repeat (HOR)-rich regions before homology assessment and length comparisons. We then characterize each TR according to its location within annotated genomic features in the CHM13 reference genome, i.e 3' UTRs, 5'UTRs, CDS, promoters, and introns. This step was performed using only the APPRIS principal isoform of each transcript (Rodriguez et al., 2013). A detailed description of the catalog creation is available in the Supplemental Material.

Identifying homologous TRs

The TRACK pipeline (Adam et al., 2024) was also used to identify homologous TRs between species. Briefly, TRACK utilizes chain files, which contain the longest and highest-scoring synthetic regions in a pairwise whole-genome alignment, to convert the TR genomic locations from one species' assembly to another using the UCSC Liftover tool (Hinrichs, 2006). To avoid biases in the homology detection process, TRACK lifts the TR catalogs bidirectionally, with the two genomes of interest serving as both target and query. The lists are compared to the target species' original TR catalog to confirm that the target species has a TR locus in the homologous regions. In this step, we retained loci with at least 10% overlap between putative homologous TR and the reference TR. Their sequence identity was then examined by pairwise alignment of the motifs from every shared TR, retaining those with a minimum alignment score of 95% or higher. A detailed description of the homology assessment is available in the Supplemental Material.

Determining TR genotypes

We genotyped the homologous TRs using long-read PacBio HiFi sequence data from 46 humans in the Human Pangenome Research Consortium (HPRC) (Liao et al., 2023) and 23 newly sequenced chimpanzee genomes (Rocha et al. *in prep*) using the Tandem Repeat Genotyping Tool implemented in the TRACK pipeline (Adam et al., 2024). The merged multisample VCF was filtered for missing data (--max-missing 1), minimum allele spanning depth (>3), and allele constancy score (>0.6). Loci with lengths < 11 bp were removed from further analyses.

Genetic diversity per locus, also referred to as expected heterozygosity (H_s), was computed as

$$H_s = 1 - \sum_{i=1}^{k} p_i^2$$

where k is the number of alleles at a locus, and p_i is the frequency of the i^{th} allele across all samples.

Per-locus mutation rates were estimated using the Goldstein genetic distance (du^2) framework under the stepwise mutation model. This distance quantifies allele size divergence between species while correcting for within-species variance, defined as

$$du^2 = ASD - (V_h + V_c),$$

where ASD represents the average square difference in allele copy number between species, and V_h and V_c denote the within-species variances for each species. The ASD term was computed as

$$ASD = \sum_i \sum_j (a_i - a_j)^2 p_i q_j$$

where a_i and a_j are allele copy numbers in humans and chimpanzees, and p_i and q_j are their corresponding allele frequencies. Within-species variances (V_h, V_c) were estimated analogously using allele frequencies within each species.

Under the stepwise mutation model, the expected value of du^2 increases linearly with the mutation rate (μ) and the divergence time (t). We therefore estimated per-generation mutation rates as

$$\mu = \frac{du^2}{t_{total}},$$

where t_{total} is the sum of the human and chimpanzee branch lengths. We assumed branch lengths of 6.2 million years for each lineage and generation times of 28 years for humans and 25 years for chimpanzees, yielding a total of approximately $t_{total} = 4.6 \times 10^5$ generations.

TR allele length variability

To investigate the relationship between TR lengths in humans and chimpanzees across different genomic contexts, we performed linear regression analyses stratified by the annotated

genomic features. For each category, we modeled chimpanzee TR length as a function of human TR length. The coefficient of determination (R^2) was computed to assess the strength of the linear relationship. Then, TRs were initially classified based on their degree of conservation across and within species. Loci with no allele length variation across all individuals were defined as *invariant* (iTR), while the remaining were classified as *variable* (TRV). TRVs were then used to compute Divergence-diversity ratios (*D*).

Divergence-diversity ratio (*D*)

To assess the relationship between a TR's divergence and diversity, we define the per locus Divergence-diversity ratio (*D*) as the ratio of between-species to within-species variance in allele length, computed separately for each species:

$$D_i = \frac{Var_{between}}{Var_{within,i}}$$

where between-species variance for each TR was defined as the weighted squared difference between each species-specific mean length and the overall mean length across both species:

$$Var_{between} = \sum_{i=1}^{2} n_i(\underline{x}_i - \underline{x})^2 , \qquad \underline{x} = \frac{\Sigma_{i=1}^{2} n_i \underline{x}_i}{\Sigma_{i=1}^{2} n_i}$$

where n_i is the number of alleles in species *i*, and $\underline{x}$ is the overall mean length across both species.

To identify TRs with extreme D_i values, we ranked all variable TRs independently for humans and chimpanzees. TRs with high D_i in both species were defined as the top 1,000 loci in each species, while TRs with low D_i in both species were defined as the bottom 1,000 loci after excluding repeats with within-species variance below 1. We implemented this threshold as a conservative noise floor in order to eliminate small TRs with low variance that produce small ratios that are likely not biologically relevant (Extended Figure 1). Species-specific extremes were defined as TRs exhibiting the greatest deviation from the diagonal in the human-chimpanzee *D* comparison. The top 1,000 TRs with the largest deviation were selected for each species, representing loci with pronounced divergence relative to diversity in a single lineage. This classification yielded four categories of TRs for downstream analyses: high *D* in both species, low *D* in both species, high *D* in humans, and high *D* in chimpanzees. Within each category, TRs were characterized with respect to genomic features, GC content, heterozygosity,

and motif length to identify features associated with conserved versus highly variable repeats.

Gene ontology enrichment analysis and gene expression data

We performed Gene Ontology (GO) enrichment analysis to test whether genes containing TRs with extreme ratio (*D*) signatures were associated with specific Biological Processes terms. We selected the genes intersecting the top 1,000 TRs within each ratio category as the gene sets to be tested for enrichment using the full set of genes intersecting TRs as the background. Analyses were conducted in ShinyGO v.0.82 (Ge et al., 2020) with a minimum pathway size of 15 and maximum pathway size of 1,000 genes. Significance was assessed using the false discovery rate (FDR), and only pathways with FDR-corrected $p < 0.01$ were considered enriched.

Gene expression data was obtained from Brawand et al., (2011), which provides normalized RPKM (reads per kilobase of exon per million mapped reads) values for six tissues (brain, cerebellum, heart, kidney, liver, and testis) from humans and chimpanzees. To reduce noise from lowly expressed genes, we applied a threshold of RPKM > 1 in at least one individual for each tissue and species. Differential expression (DE) between species was assessed per tissue *limma* (Ritchie et al., 2015) on log-transformed RPKM values, with empirical Bayes moderation of gene-wise standard errors. Genes with FDR-adjusted p-values < 0.05 were considered DE. TR length divergence was quantified as the log2 fold change of mean allele length per locus and compared to gene expression divergence using Pearson's correlation. To focus on genes containing TRs known to be associated with expression (eTRs), we retained those associated with fine-mapped expression STR (Fotsing et al., 2019) and expression-associated VNTRs (Eslami Rasekh et al., 2021).

Data availability

Code used for this project is available at https://github.com/caroladam/tr_analysis/tree/main

Supplemental Tables

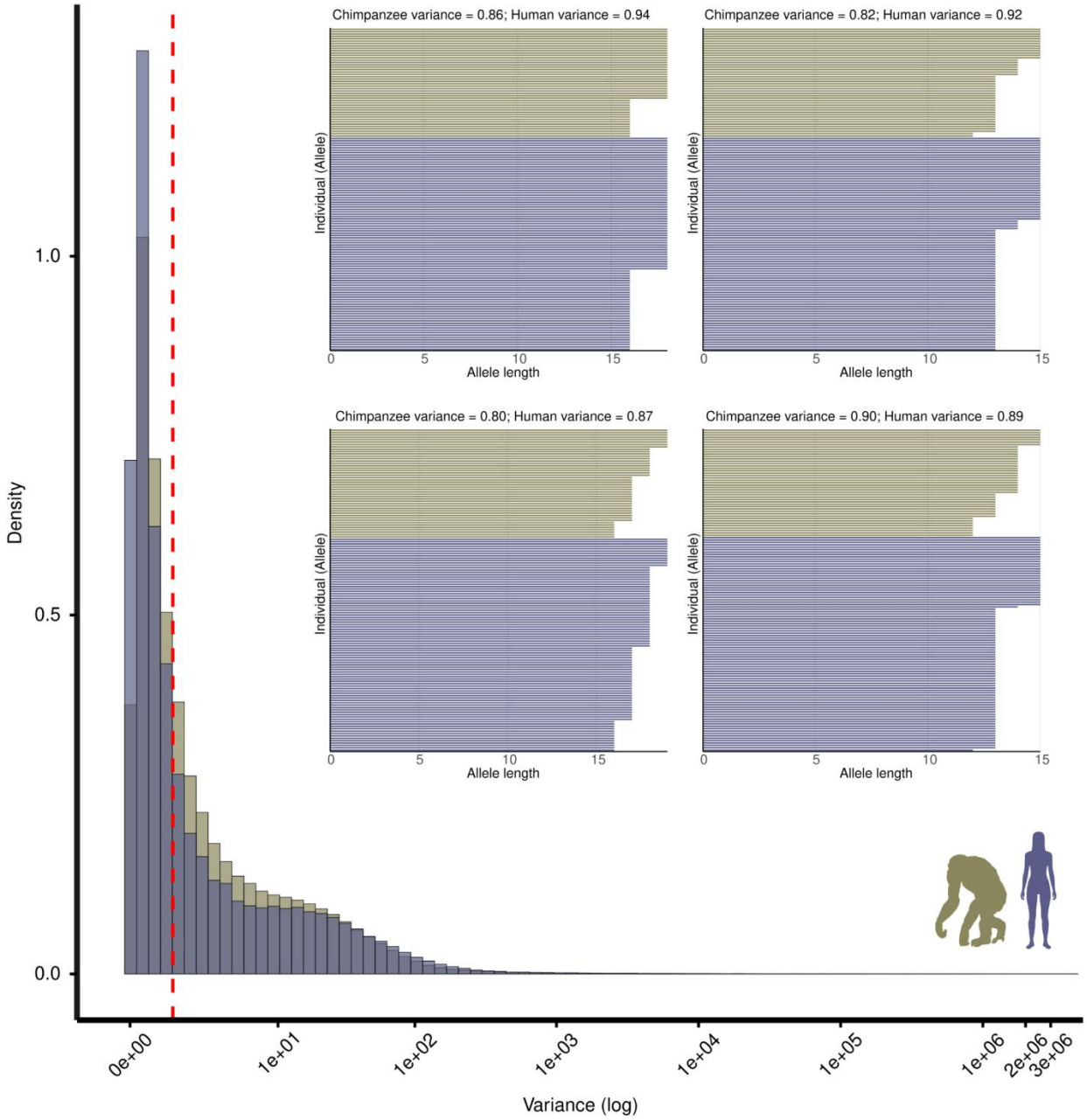
Supplemental Table S1 - Results of χ^2 enrichment tests for motif length distributions across genomic features. Shown are observed and expected TR counts, standardized residuals, p-values (raw and FDR-corrected), fold enrichment, and enrichment status relative to expectation (enriched or depleted).

Supplemental Table S2 - Description of genic Tandem Repeat Variants (TRVs) with the top 1,000 highest-ranked ratio categories.

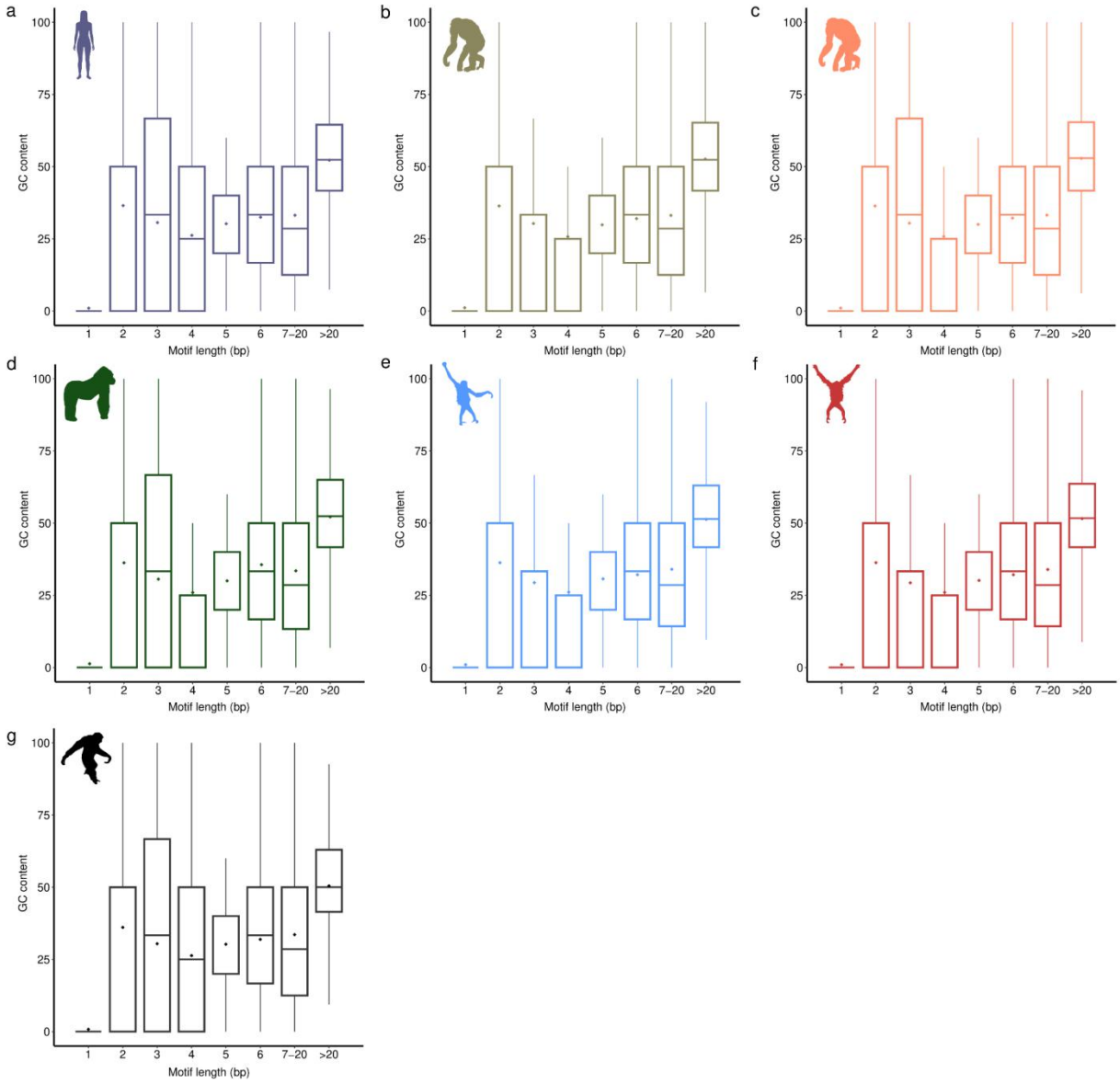
Supplemental Table S3. Description of genes harboring the top 10 highest-ranked TRVs in the high and low Divergence-Diversity Ratio categories.

Supplemental Table S4 - Description of genes harboring the top 10 highest-ranked TRVs in the high Divergence-Diversity Ratio categories in a single species.

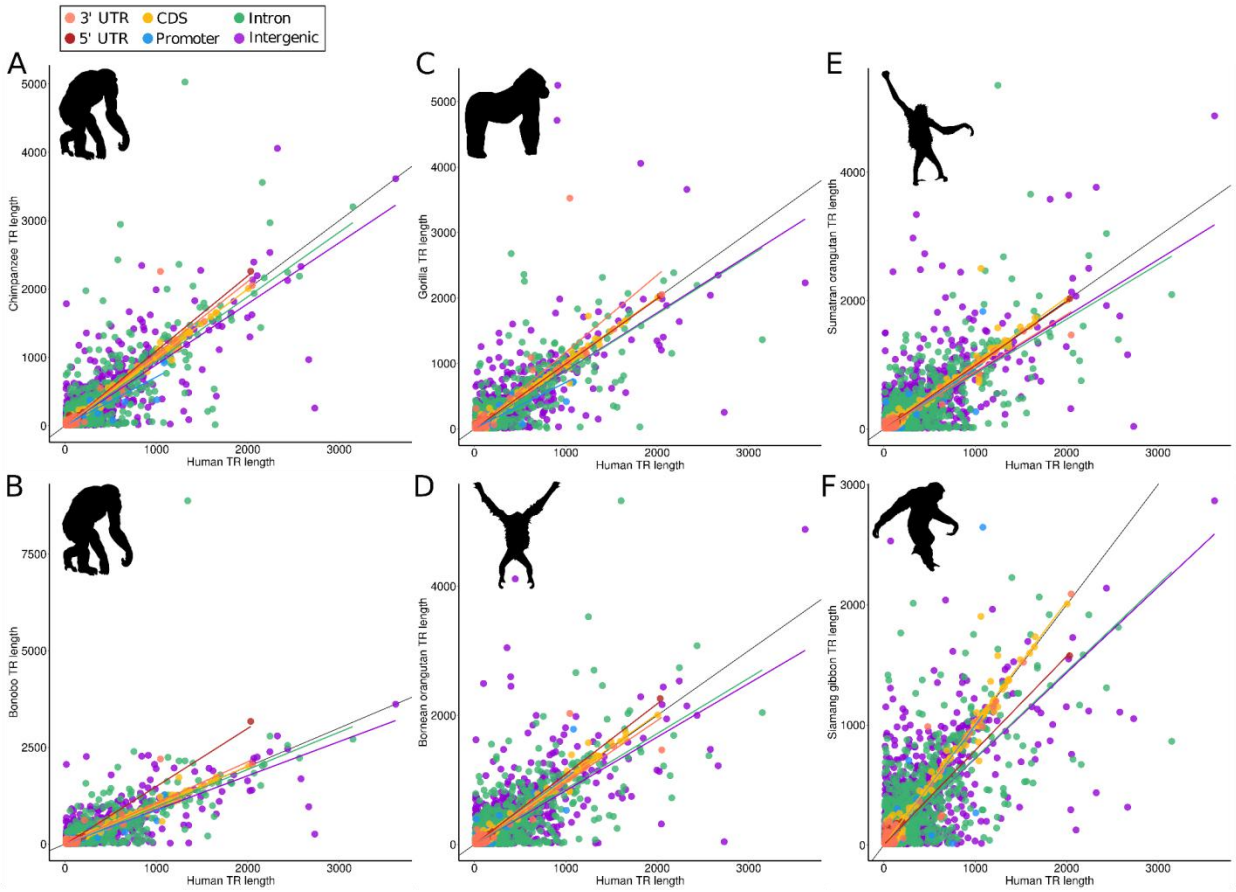
Extended Figures



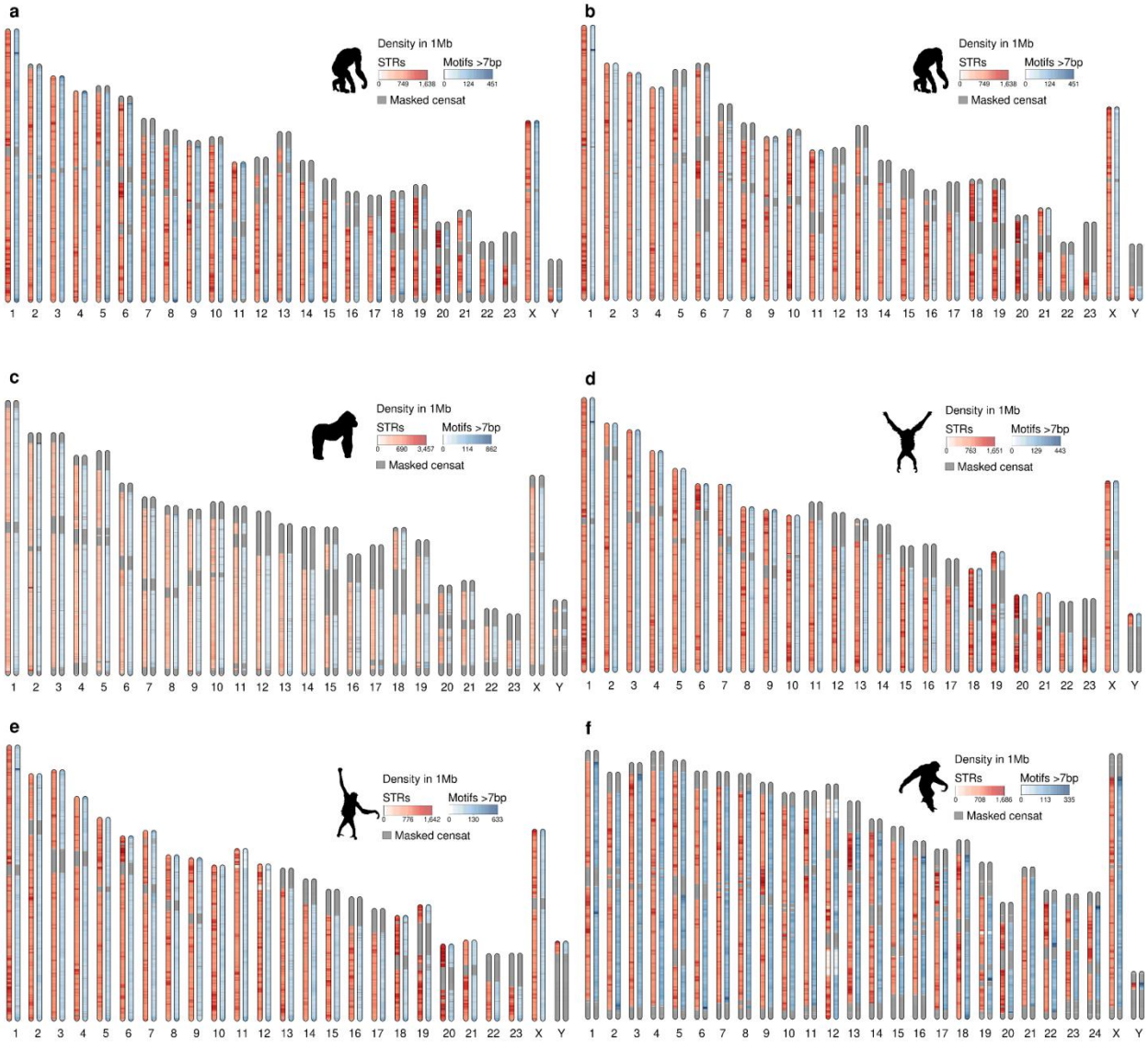
Extended Figure 1 - Histogram showing the distribution of non-zero within-species TR variance across humans and chimpanzees. The vertical line indicates the implemented threshold (minimum variance = 1), used to exclude loci with low variance that produce small ratios that are likely not biologically relevant. Inset barplots show four examples of the allele-length distributions of TRVs with within-species variance near but below 1.



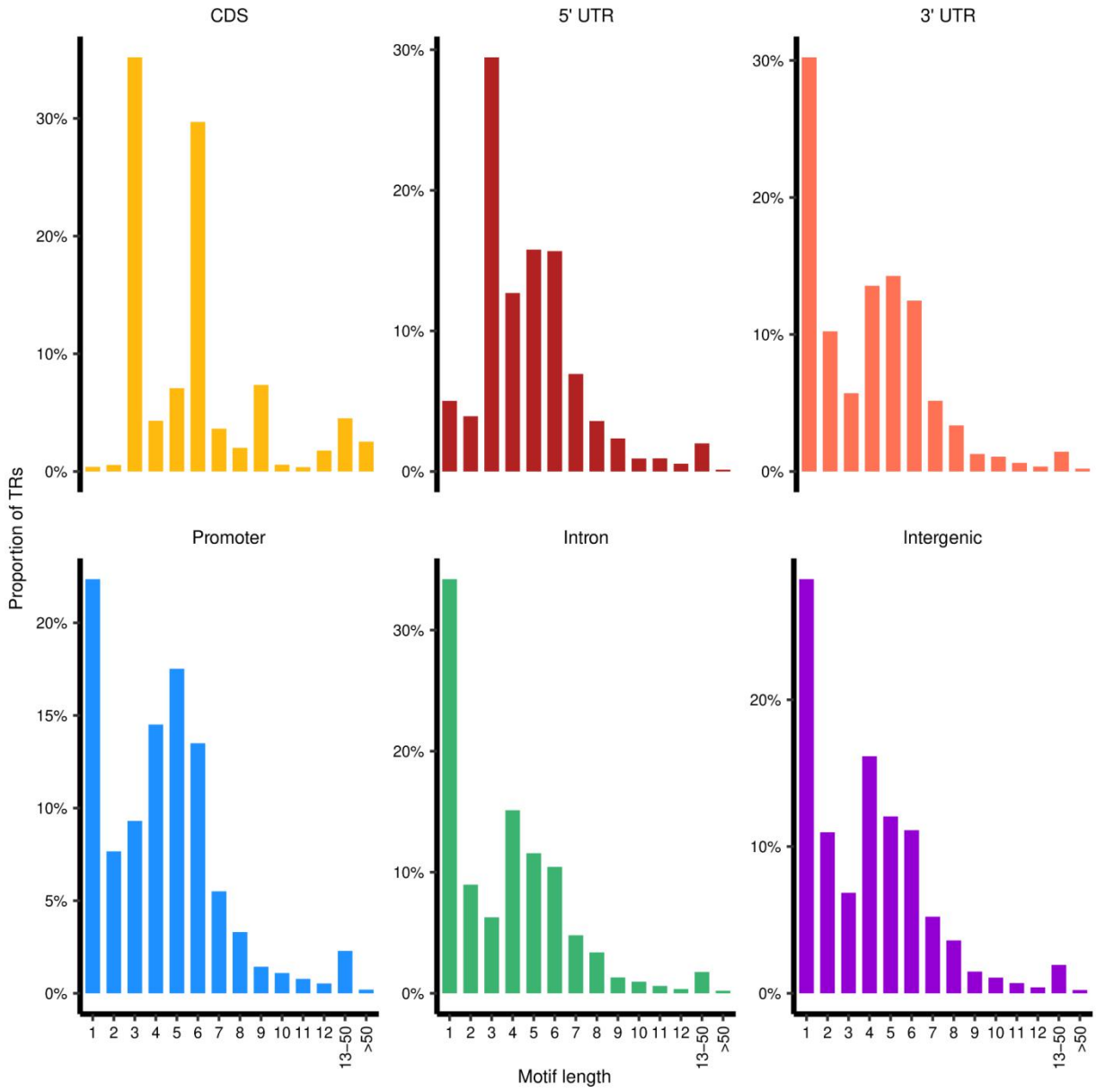
Extended Figure 2 - Boxplot showing the GC content of TR motifs across motif lengths for a) *Homo sapiens*, b) *Pan troglodytes*, c) *Pan paniscus*, d) *Gorilla gorilla*, e) *Pongo pygmaeus*, f) *Pongo abelii*, and g) *Symphalangus syndactylus.*



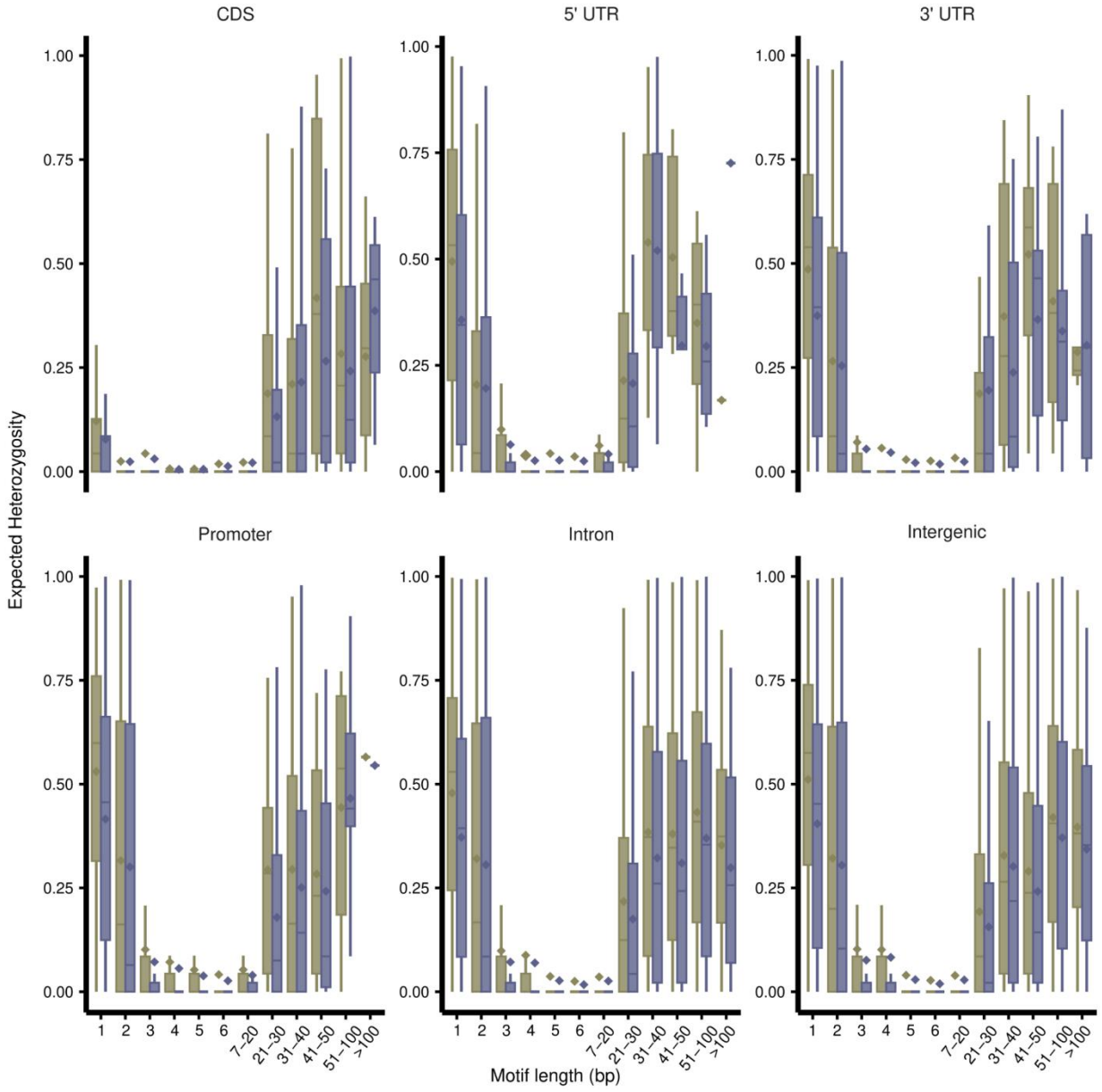
Extended Figure 3 -Scatterplot of reference tandem repeat (TR) length between humans (x-axis) and remaining T2T ape species (y-axis). Each point represents a single TR locus, with color indicating the genomic annotation in the CHM13 genome. a) *Pan troglodytes*, b) *Pan paniscus*, c) *Gorilla gorilla*, d) *Pongo pygmaeus*, e) *Pongo abelii*, and f) *Symphalangus syndactylus.*



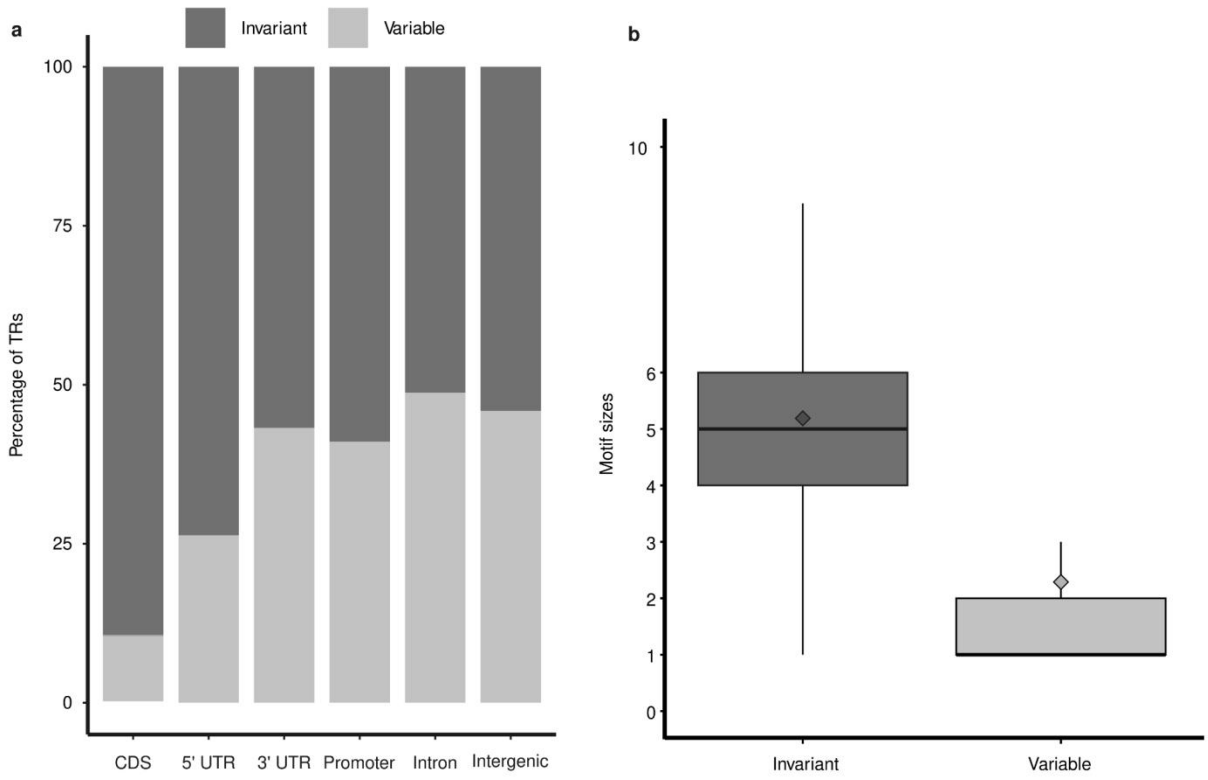
Extended Figure 4 - Ideogram of the non-human T2T genomes showing the density of short tandem repeats (STRs) in red and TRs with motif length >7 bp in blue across non-overlapping 1 Mb windows. Repeat density is plotted along each chromosome. TR densities are shown for a) *Pan troglodytes*, b) *Pan paniscus*, c) *Gorilla gorilla*, d) *Pongo pygmaeus*, e) *Pongo abelii*, and f) *Symphalangus syndactylus.*



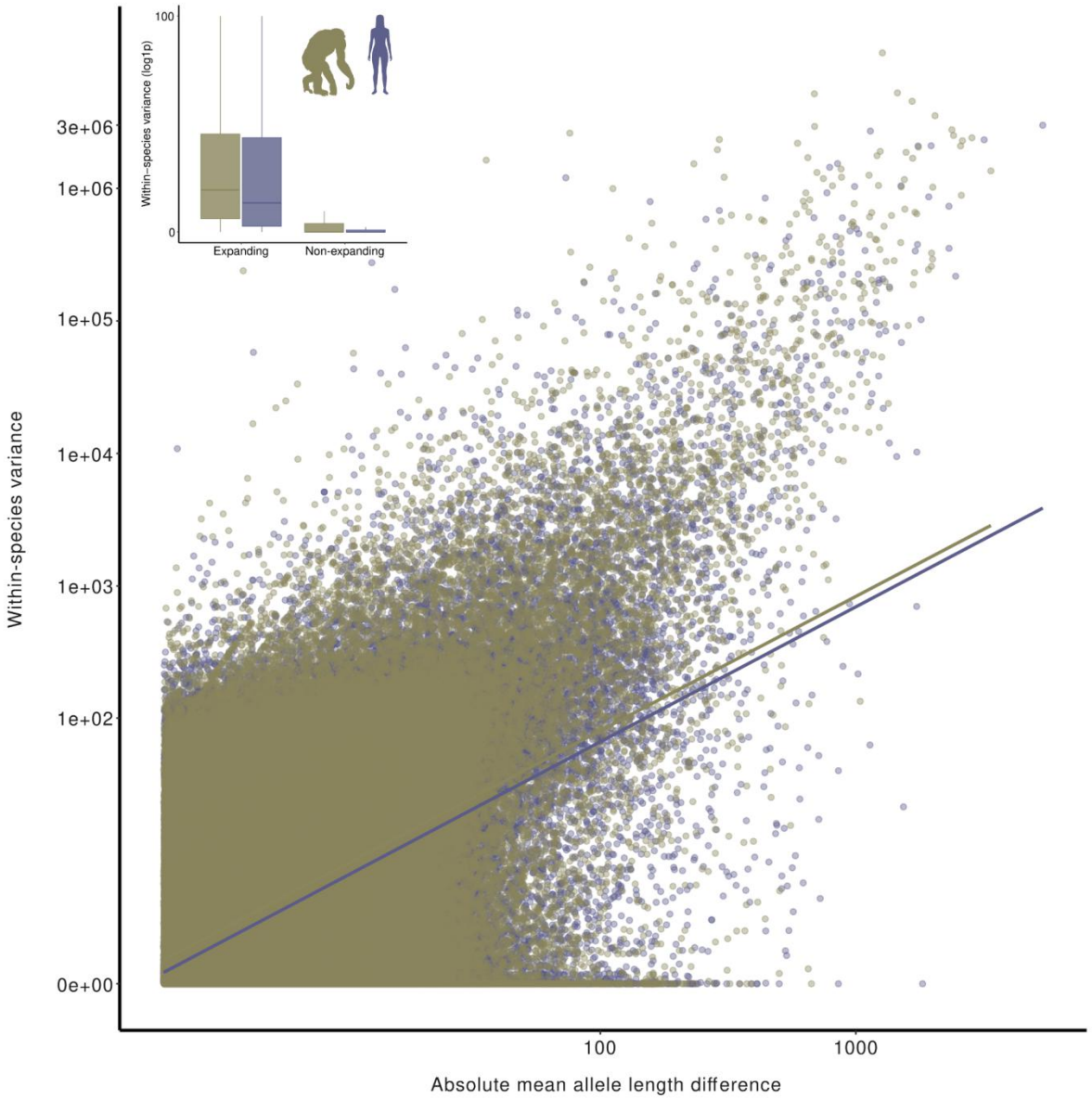
Extended Figure 5 - Barplots showing the proportional distribution of TR motif lengths across genomic features in the CHM13 reference genome.



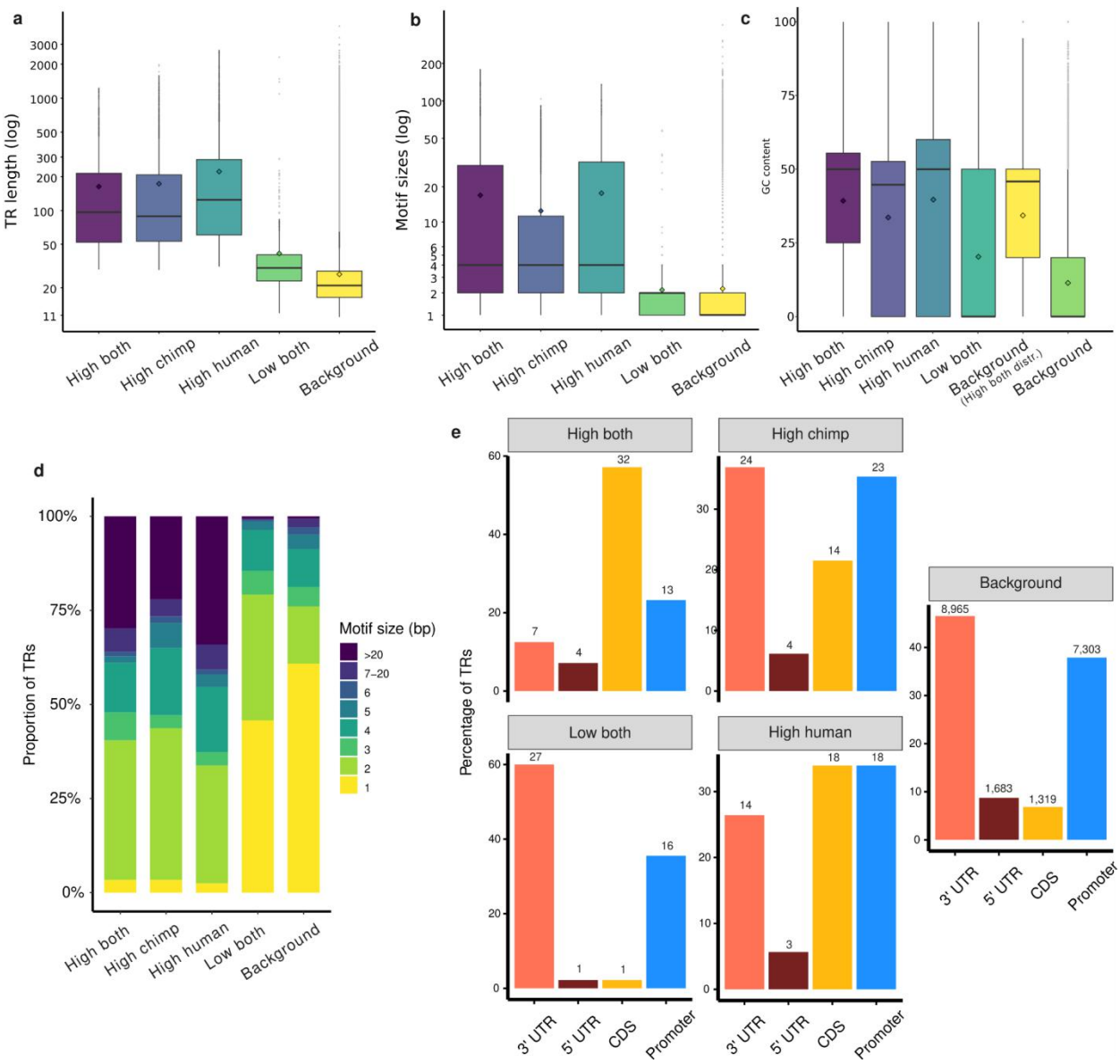
Extended Figure 6 - Expected heterozygosity of TRs shared between humans (purple) and chimpanzees (green) across genomic features, stratified by motif lengths.



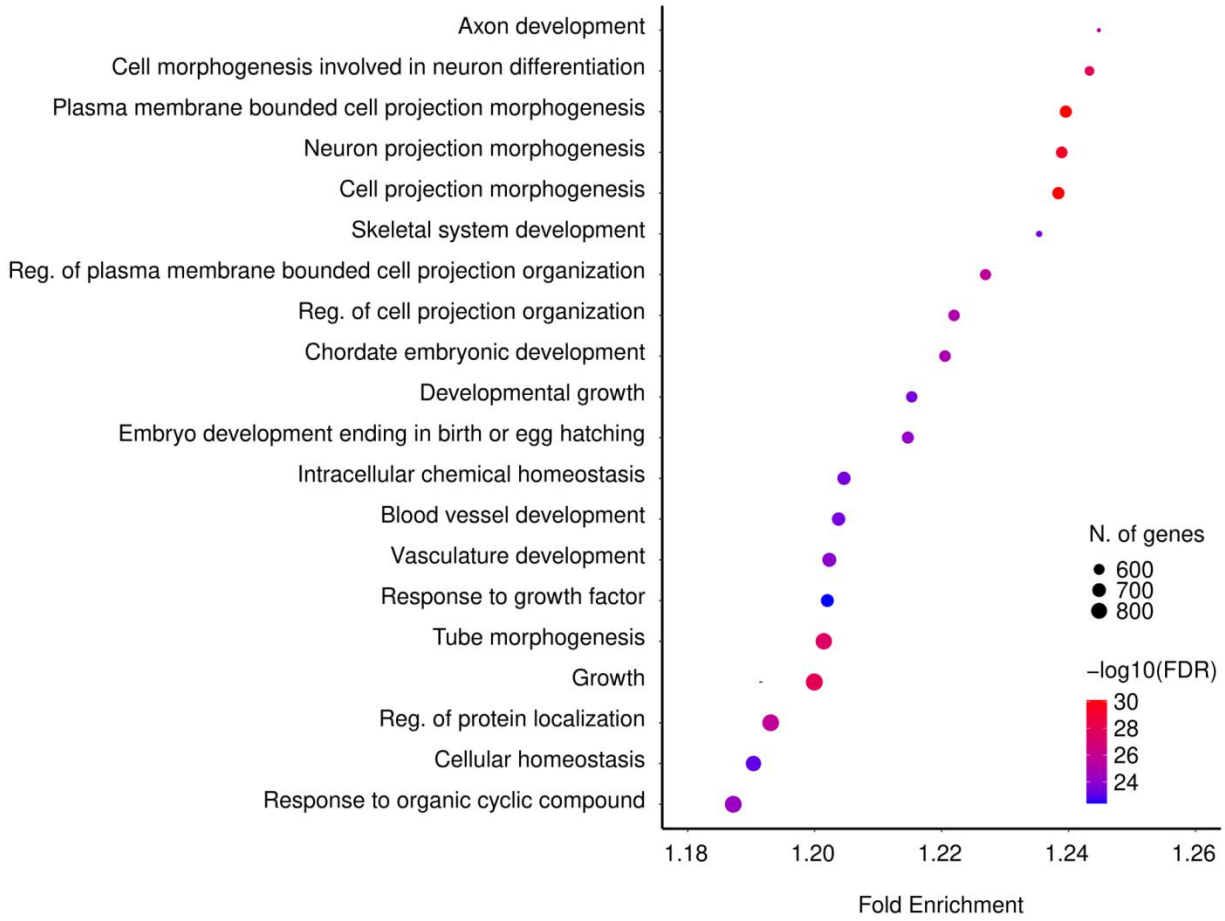
Extended Figure 7 - a) Barplots showing the percentage of invariant and variable TRs within each genomic feature. b) Boxplot showing the distribution of motif lengths across invariant and variable TRs. Outliers are omitted for clarity.



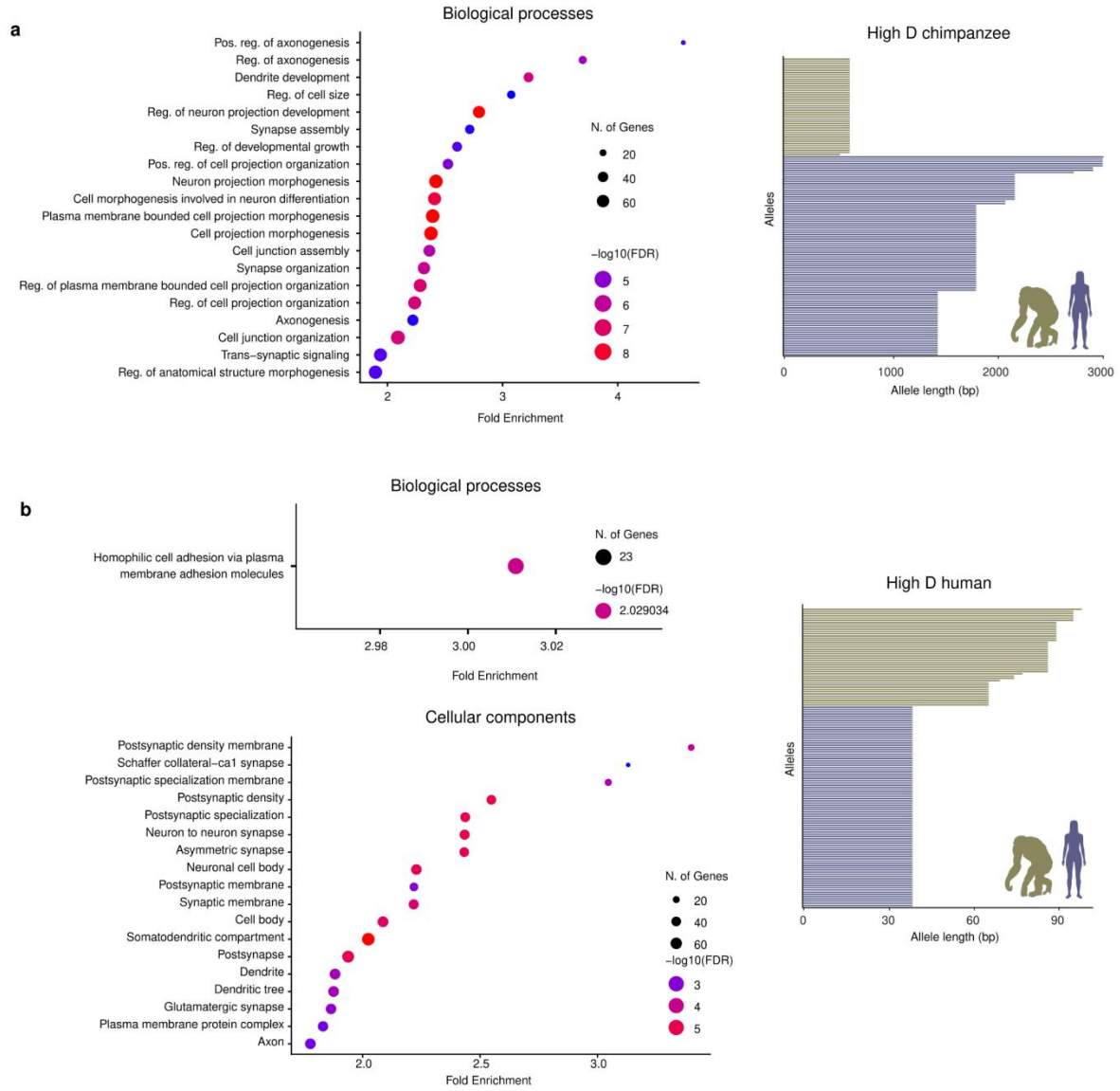
Extended Figure 8 - Scatterplot showing within-species variance versus absolute difference in mean allele length between humans and chimpanzees. The inset boxplot shows within-species variance, grouped by expansion status, for each species. Outliers are omitted for clarity.



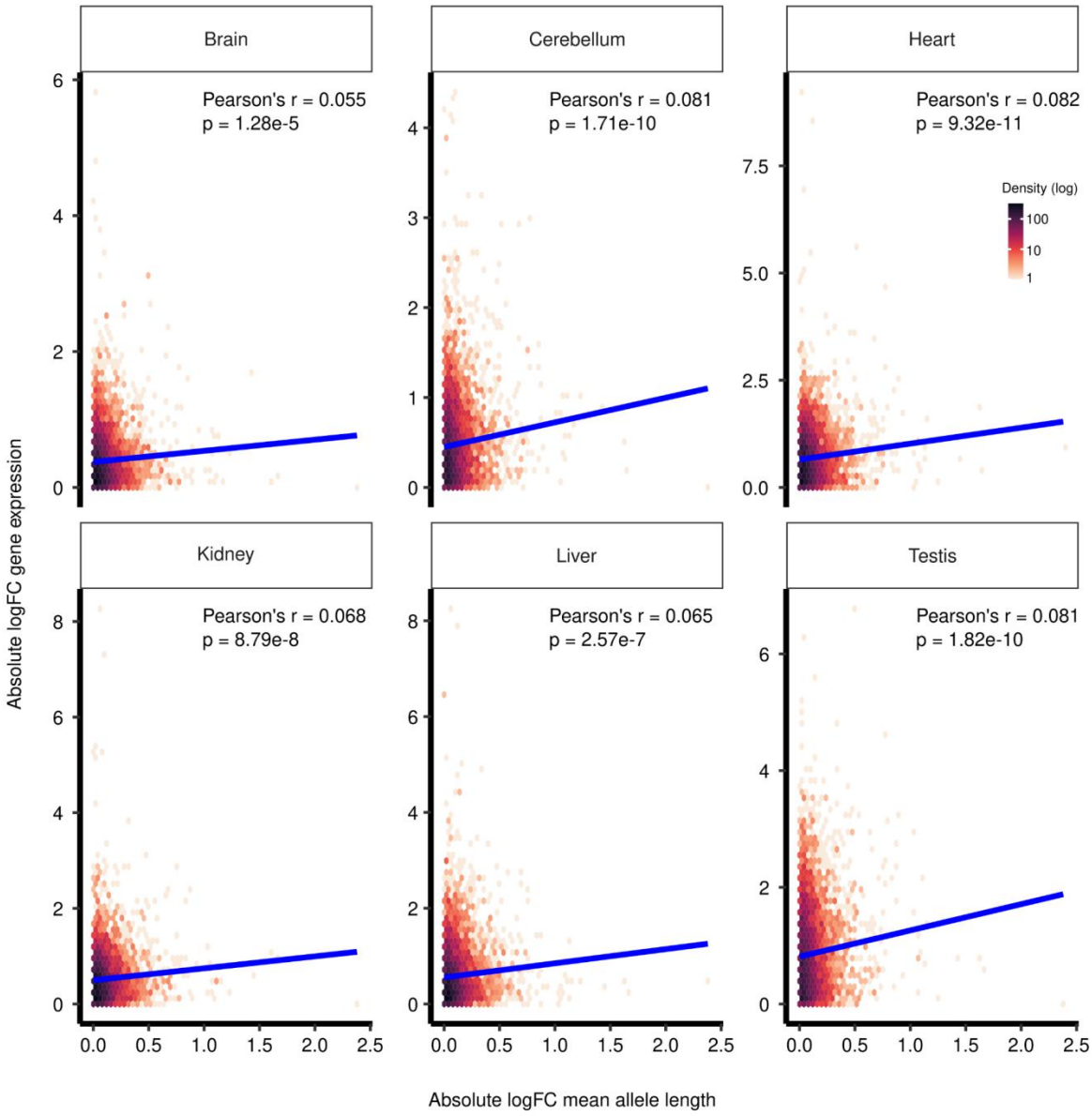
Extended Figure 9 - Boxplots showing the distribution of a) TR lengths, b) motif sizes, and c) GC content across *D* categories. d) Stacked barplot showing the proportion of TRs with different motif sizes across *D* categories, and e) Barplots showing the proportion of genic, non-intronic TRs within each *D* category made up by different genomic features.



Extended Figure 10 - Gene Ontology enrichment analysis for Biological Processes terms associated with genes containing Tandem Repeats (TRs) using the whole set of human genes as background.



Extended Figure 11 - Gene Ontology enrichment analysis for Biological Processes terms associated with genes intersecting the top 1,000 genic TRs with a) high chimpanzee *D* and b) high human *D*, which also includes enrichment for Cellular Components. Includes allele length distribution for one representative TR in each category.



Extended Figure 12 - Hexbin plot showing the correlation between absolute log fold change in mean TR allele length across genes (x-axis) and absolute log fold change in gene expression (y-axis) across multiple tissues.

References

Adam, C. L., Rocha, J., Sudmant, P., & Rohlfs, R. (2024). TRACKing tandem repeats: A customizable pipeline for identification and cross-species comparison. *Bioinformatics Advances*, *5*(1), vbaf066. https://doi.org/10.1093/bioadv/vbaf066

Ahmad, S. F., Singchat, W., Jehangir, M., Suntronpong, A., Panthum, T., Malaivijitnond, S., & Srikulnath, K. (2020). Dark Matter of Primate Genomes: Satellite DNA Repeats and Their Evolutionary Dynamics. *Cells*, *9*(12), 2714. https://doi.org/10.3390/cells9122714

Altman, A., & Kong, K.-F. (2016). Protein Kinase C Enzymes in the Hematopoietic and Immune Systems. *Annual Review of Immunology*, *34*(1), 511–538. https://doi.org/10.1146/annurev-immunol-041015-055347

Bitarello, B. D., De Filippo, C., Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D., & Andrés, A. M. (2018). Signatures of Long-Term Balancing Selection in Human Genomes. *Genome Biology and Evolution*, *10*(3), 939–955. https://doi.org/10.1093/gbe/evy054

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., & Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, *478*(7369), 343–348. https://doi.org/10.1038/nature10532

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, 27(2), 573-580. https://doi.org/10.1093/nar/27.2.573

Byeon, G. W., Cenik, E. S., Jiang, L., Tang, H., Das, R., & Barna, M. (2021). Functional and structural basis of extreme conservation in vertebrate 5′ untranslated regions. *Nature Genetics*, *53*(5), 729–741. https://doi.org/10.1038/s41588-021-00830-1

Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G., & Raschellà, G. (2017). Zinc-finger proteins in health and disease. *Cell Death Discovery*, *3*(1), 17071. https://doi.org/10.1038/cddiscovery.2017.71

Chavali, S., Chavali, P. L., Chalancon, G., De Groot, N. S., Gemayel, R., Latysheva, N. S., Ing-Simmons, E., Verstrepen, K. J., Balaji, S., & Babu, M. M. (2017). Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. *Nature Structural & Molecular Biology*, *24*(9), 765–777. https://doi.org/10.1038/nsmb.3441

Churbanov, A. (2005). Evolutionary conservation suggests a regulatory function of AUG triplets in 5'-UTRs of eukaryotic genes. *Nucleic Acids Research*, *33*(17), 5512–5520. https://doi.org/10.1093/nar/gki847

Cui, Y., Arnold, F. J., Li, J. S., Wu, J., Wang, D., Philippe, J., Colwin, M. R., Michels, S., Chen, C., Sallam, T., Thompson, L. M., La Spada, A. R., & Li, W. (2025). Multi-omic quantitative trait loci link tandem repeat size variation to gene regulation in human brain. *Nature Genetics*, *57*(2), 369–378. https://doi.org/10.1038/s41588-024-02057-2

Da, G., Wang, J., Shang, J., Xun, C., Yu, Y., Wang, Y., Tie, N., & Li, H. (2024). Nuclear PCGF3 inhibits the antiviral immune response by suppressing the interferon-stimulated gene. *Cell Death Discovery*, *10*(1), 429. https://doi.org/10.1038/s41420-024-02194-x

637 Elsir, T., Smits, A., Lindström, M. S., & Nistér, M. (2012). Transcription factor PROX1: Its role in
638 development and cancer. *Cancer and Metastasis Reviews*, *31*(3–4), 793–805.
639 https://doi.org/10.1007/s10555-012-9390-8

640 Erwin, G. S., Gürsoy, G., Al-Abri, R., Suriyaprakash, A., Dolzhenko, E., Zhu, K., Hoerner, C. R., White,
641 S. M., Ramirez, L., Vadlakonda, A., Vadlakonda, A., Von Kraut, K., Park, J., Brannon, C. M.,
642 Sumano, D. A., Kirtikar, R. A., Erwin, A. A., Metzner, T. J., Yuen, R. K. C., … Snyder, M. P.
643 (2023). Recurrent repeat expansions in human cancer genomes. *Nature*, *613*(7942), 96–102.
644 https://doi.org/10.1038/s41586-022-05515-1

645 Eslami Rasekh, M., Hernández, Y., Drinan, S. D., Fuxman Bass, J. I., & Benson, G. (2021). Genome-
646 wide characterization of human minisatellite VNTRs: Population-specific alleles and gene
647 expression differences. *Nucleic Acids Research*, *49*(8), 4308–4324.
648 https://doi.org/10.1093/nar/gkab224

649 Faisst, A. M., Alvarez-Bolado, G., Treichel, D., & Gruss, P. (2002). Rotatin is a novel gene required for
650 axial rotation and left–right specification in mouse embryos. *Mechanisms of Development*, *113*(1),
651 15–28. https://doi.org/10.1016/S0925-4773(02)00003-5

652 Fan, H., & Chu, J.-Y. (2007). A Brief Review of Short Tandem Repeat Mutation. *Genomics, Proteomics*
653 *& Bioinformatics*, *5*(1), 7–14. https://doi.org/10.1016/S1672-0229(07)60009-6

654 Ferrer-Admetlla, A., Bosch, E., Sikora, M., Marquès-Bonet, T., Ramírez-Soriano, A., Muntasell, A.,
655 Navarro, A., Lazarus, R., Calafell, F., Bertranpetit, J., & Casals, F. (2008). Balancing Selection Is
656 the Main Force Shaping the Evolution of Innate Immunity Genes. *The Journal of Immunology*,
657 *181*(2), 1315–1322. https://doi.org/10.4049/jimmunol.181.2.1315

658 Fondon, J. W., & Garner, H. R. (2004). Molecular origins of rapid and continuous morphological
659 evolution. *Proceedings of the National Academy of Sciences*, *101*(52), 18058–18063.
660 https://doi.org/10.1073/pnas.0408118101

661 Fotsing, S. F., Margoliash, J., Wang, C., Saini, S., Yanicky, R., Shleizer-Burko, S., Goren, A., & Gymrek,
662 M. (2019a). The impact of short tandem repeat variation on gene expression. *Nature Genetics*,
663 *51*(11), 1652–1659. https://doi.org/10.1038/s41588-019-0521-9

664 Fotsing, S. F., Margoliash, J., Wang, C., Saini, S., Yanicky, R., Shleizer-Burko, S., Goren, A., & Gymrek,
665 M. (2019b). The impact of short tandem repeat variation on gene expression. *Nature Genetics*,
666 *51*(11), 1652–1659. https://doi.org/10.1038/s41588-019-0521-9

667 Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: A graphical gene-set enrichment tool for animals and
668 plants. *Bioinformatics*, *36*(8), 2628–2629. https://doi.org/10.1093/bioinformatics/btz931

669 Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L., & Feldman, M. W. (1995). An evaluation of
670 genetic distances for use with microsatellite loci. *Genetics*, *139*(1), 463–471.
671 https://doi.org/10.1093/genetics/139.1.463

672 Grimwood, J., Gordon, L. A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., Hellsten, U., Goodstein, D.,
673 Couronne, O., Tran-Gyamfi, M., Aerts, A., Altherr, M., Ashworth, L., Bajorek, E., Black, S.,
674 Branscomb, E., Caenepeel, S., Carrano, A., Caoile, C., … Lucas, S. M. (2004). The DNA sequence
675 and biology of human chromosome 19. *Nature*, *428*(6982), 529–535.
676 https://doi.org/10.1038/nature02399

677 Gymrek, M. (2017). A genomic view of short tandem repeats. *Current Opinion in Genetics &*
678 *Development*, *44*, 9–16. https://doi.org/10.1016/j.gde.2017.01.012

679 Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M. J., Price, A. L.,
680 Pritchard, J. K., Sharp, A. J., & Erlich, Y. (2016). Abundant contribution of short tandem repeats to
681 gene expression variation in humans. *Nature Genetics*, *48*(1), 22–29. https://doi.org/10.1038/ng.3461

682 Hinrichs, A. S. (2006). The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Research*,
683 *34*(90001), D590–D598. https://doi.org/10.1093/nar/gkj144

684 Horton, C. A., Alexandari, A. M., Hayes, M. G. B., Marklund, E., Schaepe, J. M., Aditham, A. K., Shah,
685 N., Suzuki, P. H., Shrikumar, A., Afek, A., Greenleaf, W. J., Gordân, R., Zeitlinger, J., Kundaje, A.,
686 & Fordyce, P. M. (2023). Short tandem repeats bind transcription factors to tune eukaryotic gene
687 expression. *Science*, *381*(6664), eadd1250. https://doi.org/10.1126/science.add1250

688 Huang, J., Li, W., Jian, Z., Yue, B., & Yan, Y. (2016). Genome-wide distribution and organization of
689 microsatellites in six species of birds. *Biochemical Systematics and Ecology*, *67*, 95–102.
690 https://doi.org/10.1016/j.bse.2016.05.023

691 Huang, Y., Wang, M., Wang, Z., Liu, X., Feng, Y., Zhong, J., Huang, H., Geng, J., Tang, T., Liu, C., Lu,
692 Y., Cheng, J., Bu, F., He, G., & Yuan, H. (2025). Short tandem repeats in populations of the
693 Qinghai-Tibet Plateau and adjacent regions provide insights into high-altitude adaptation. *S c i e n c*
694 *e A d v a n c e s*.

695 Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J.
696 (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, *44*(D1), D81–D89.
697 https://doi.org/10.1093/nar/gkv1272

698 Jam, H. Z., Li, Y., DeVito, R., Mousavi, N., Ma, N., Lujumba, I., Adam, Y., Maksimov, M., Huang, B.,
699 Dolzhenko, E., Qiu, Y., Kakembo, E., Joseph, H., Onyido, B., Adeyemi, J., Park, J., Javadzadeh, S.,
700 Jjingo, D., Adebiyi, E., & Gymrek, M. (n.d.). *A deep population reference panel of tandem repeat*
701 *variation*.

702 Jarvis, E. D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., Tracey, A., Thibaud-Nissen, F.,
703 Vollger, M. R., Porubsky, D., Cheng, H., Asri, M., Logsdon, G. A., Carnevali, P., Chaisson, M. J. P.,
704 Chin, C.-S., Cody, S., Collins, J., Ebert, P., … Human Pangenome Reference Consortium. (2022).
705 Semi-automated assembly of high-quality diploid human reference genomes. *Nature*, *611*(7936),
706 519–531. https://doi.org/10.1038/s41586-022-05325-5

707 Jovanovic, V. M., Sarfert, M., Reyna-Blanco, C. S., Indrischek, H., Valdivia, D. I., Shelest, E., & Nowick,
708 K. (2021). Positive Selection in Gene Regulatory Factors Suggests Adaptive Pleiotropic Changes
709 During Human Evolution. *Frontiers in Genetics*, *12*, 662239.
710 https://doi.org/10.3389/fgene.2021.662239

711 Kerkenberg, N., Wachsmuth, L., Zhang, M., Schettler, C., Ponimaskin, E., Faber, C., Baune, B. T., Zhang,
712 W., & Hohoff, C. (2021). Brain microstructural changes in mice persist in adulthood and are
713 modulated by the palmitoyl acyltransferase ZDHHC7. *European Journal of Neuroscience*, *54*(6),
714 5951–5967. https://doi.org/10.1111/ejn.15415

715 Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M.,
716 & Pääbo, S. (2005). Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans

and Chimpanzees. *Science*, *309*(5742), 1850–1854. https://doi.org/10.1126/science.1108296

Kim, K., Bang, S., Yoo, D., Kim, H., & Suzuki, S. (2019). De novo emergence and potential function of human-specific tandem repeats in brain-related loci. *Human Genetics*, *138*(6), 661–672. https://doi.org/10.1007/s00439-019-02017-5

King, D. G., Soller, M., & Kashi, Y. (1997). Evolutionary tuning knobs. *Endeavour*, *21*(1), 36–40. https://doi.org/10.1016/S0160-9327(97)01005-3

Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., Underwood, J. G., Nelson, B. J., Chaisson, M. J. P., Dougherty, M. L., Munson, K. M., Hastie, A. R., Diekhans, M., Hormozdiari, F., Lorusso, N., Hoekzema, K., Qiu, R., Clark, K., Raja, A., … Eichler, E. E. (2018). High-resolution comparative analysis of great ape genomes. *Science*, *360*(6393), eaar6343. https://doi.org/10.1126/science.aar6343

Lai, Y. (2003). The Relationship Between Microsatellite Slippage Mutation Rate and the Number of Repeat Units. *Molecular Biology and Evolution*, *20*(12), 2123–2131. https://doi.org/10.1093/molbev/msg228

Leffler, E. M., Gao, Z., Pfeifer, S., Ségurel, L., Auton, A., Venn, O., Bowden, R., Bontrop, R., Wall, J. D., Sella, G., Donnelly, P., McVean, G., & Przeworski, M. (2013). Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science (New York, N.Y.)*, *339*(6127), 1578–1582. https://doi.org/10.1126/science.1234070

Li, W. H., & Sadler, L. A. (1991). Low nucleotide diversity in man. *Genetics*, *129*(2), 513–523. https://doi.org/10.1093/genetics/129.2.513

Li, Y., Korol, A. B., Fahima, T., Beiles, A., & Nevo, E. (2002). Microsatellites: Genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, *11*(12), 2453–2465. https://doi.org/10.1046/j.1365-294X.2002.01643.x

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., … Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312–324. https://doi.org/10.1038/s41586-023-05896-x

Linthorst, J., Meert, W., Hestand, M. S., Korlach, J., Vermeesch, J. R., Reinders, M. J. T., & Holstege, H. (2020). Extreme enrichment of VNTR-associated polymorphicity in human subtelomeres: Genes with most VNTRs are predominantly expressed in the brain. *Translational Psychiatry*, *10*(1), 369. https://doi.org/10.1038/s41398-020-01060-5

Liu, Q., & Tian, W. (2025). Association of human-specific expanded short tandem repeats with neuron-specific regulatory features. *S c i e n c e A d v a n c e s.*

Madsen, B. E., Villesen, P., & Wiuf, C. (2008). Short Tandem Repeats in Human Exons: A Target for Disease Mutations. *BMC Genomics*, *9*(1), 410. https://doi.org/10.1186/1471-2164-9-410

Maleitzke, T., Hildebrandt, A., Dietrich, T., Appelt, J., Jahn, D., Otto, E., Zocholl, D., Baranowsky, A., Duda, G. N., Tsitsilonis, S., & Keller, J. (2022). The calcitonin receptor protects against bone loss and excessive inflammation in collagen antibody-induced arthritis. *iScience*, *25*(1), 103689. https://doi.org/10.1016/j.isci.2021.103689

Minias, P., & Vinkler, M. (2022). Selection Balancing at Innate Immune Genes: Adaptive Polymorphism Maintenance in Toll-Like Receptors. *Molecular Biology and Evolution*, *39*(5), msac102. https://doi.org/10.1093/molbev/msac102

Nowick, K., Fields, C., Gernat, T., Caetano-Anolles, D., Kholina, N., & Stubbs, L. (2011). Gain, Loss and Divergence in Primate Zinc-Finger Genes: A Rich Resource for Evolution of Gene Regulatory Differences between Species. *PLoS ONE*, *6*(6), e21553. https://doi.org/10.1371/journal.pone.0021553

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). *The complete sequence of a human genome*.

Porubsky, D., Dashnow, H., Sasani, T. A., Logsdon, G. A., Hallast, P., Noyes, M. D., Kronenberg, Z. N., Mokveld, T., Koundinya, N., Nolan, C., Steely, C. J., Guarracino, A., Dolzhenko, E., Harvey, W. T., Rowell, W. J., Grigorev, K., Nicholas, T. J., Goldberg, M. E., Oshima, K. K., … Eichler, E. E. (2025). Human de novo mutation rates from a four-generation pedigree reference. *Nature*. https://doi.org/10.1038/s41586-025-08922-2

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., … Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471–475. https://doi.org/10.1038/nature12228

Press, M. O., McCoy, R. C., Hall, A. N., Akey, J. M., & Queitsch, C. (2018). Massive variation of short tandem repeats with functional consequences across strains of *Arabidopsis thaliana*. *Genome Research*, *28*(8), 1169–1178. https://doi.org/10.1101/gr.231753.117

Rasmussen, A. H., Rasmussen, H. B., & Silahtaroglu, A. (2017). The DLGAP family: Neuronal expression, function and role in brain disorders. *Molecular Brain*, *10*(1), 43. https://doi.org/10.1186/s13041-017-0324-9

Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., Hook, P. W., Koren, S., Rautiainen, M., Alexandrov, I. A., Allen, J., Asri, M., Bzikadze, A. V., Chen, N.-C., Chin, C.-S., Diekhans, M., Flicek, P., Formenti, G., Fungtammasan, A., … Phillippy, A. M. (2023). The complete sequence of a human Y chromosome. *Nature*, *621*(7978), 344–354. https://doi.org/10.1038/s41586-023-06457-y

Richard, G., & Pâques, F. (2000). Mini- and microsatellite expansions: The recombination connection. *EMBO Reports*, *1*(2), 122–126. https://doi.org/10.1093/embo-reports/kvd031

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47. https://doi.org/10.1093/nar/gkv007

Rocha, J., Lou, R. N., & Sudmant, P. H. (2024). Structural variation in humans and our primate kin in the era of telomere-to-telomere genomes and pangenomics. *Current Opinion in Genetics & Development*, *87*, 102233. https://doi.org/10.1016/j.gde.2024.102233

Rocha, J. et al. (in prep). [Study on structural variation and haplotype diversity in humans, bonobos, and chimpanzees through pangenomic approaches].

Rockman, M. V., & Wray, G. A. (2002). Abundant Raw Material for Cis-Regulatory Evolution in Humans. *Molecular Biology and Evolution*, *19*(11), 1991–2004. https://doi.org/10.1093/oxfordjournals.molbev.a004023

Rodriguez De Los Santos, M., Kopell, B. H., Buxbaum Grice, A., Ganesh, G., Yang, A., Amini, P., Liharska, L. E., Vornholt, E., Fullard, J. F., Dong, P., Park, E., Zipkowitz, S., Kaji, D. A., Thompson, R. C., Liu, D., Park, Y. J., Cheng, E., Ziafat, K., Moya, E., … Breen, M. S. (2024). Divergent landscapes of A-to-I editing in postmortem and living human brain. *Nature Communications*, *15*(1), 5366. https://doi.org/10.1038/s41467-024-49268-z

Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Valencia, A., & Tress, M. L. (2013). APPRIS: Annotation of principal and alternative splice isoforms. *Nucleic Acids Research*, *41*(D1), D110–D117. https://doi.org/10.1093/nar/gks1058

Schaper, E., Gascuel, O., & Anisimova, M. (2014). Deep Conservation of Human Protein Tandem Repeats within the Eukaryotes. *Molecular Biology and Evolution*, *31*(5), 1132–1148. https://doi.org/10.1093/molbev/msu062

Schloissnig, S., Pani, S., Rodriguez-Martin, B., Ebler, J., Hain, C., Tsapalou, V., Söylev, A., Hüther, P., Ashraf, H., Prodanov, T., Asparuhova, M., Hunt, S., Rausch, T., Marschall, T., & Korbel, J. O. (2024). *Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2024.04.18.590093

Sharma, A., & Sowpati, D. T. (2025). Analysis of tandem repeats in seven telomere-to-telomere primate genomes. *Journal of Genetics*, *104*(2), 14. https://doi.org/10.1007/s12041-025-01503-2

Song, J. H. T., Lowe, C. B., & Kingsley, D. M. (2018). Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia. *American Journal of Human Genetics*, *103*(3), 421–430. https://doi.org/10.1016/j.ajhg.2018.07.011

Soulas-Sprauel, P., Le Guyader, G., Rivera-Munoz, P., Abramowski, V., Olivier-Martin, C., Goujet-Zalc, C., Charneau, P., & De Villartay, J.-P. (2007). Role for DNA repair factor XRCC4 in immunoglobulin class switch recombination. *The Journal of Experimental Medicine*, *204*(7), 1717–1727. https://doi.org/10.1084/jem.20070255

Srivastava, S., Avvaru, A. K., Sowpati, D. T., & Mishra, R. K. (2019). Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics*, *20*(1), 153. https://doi.org/10.1186/s12864-019-5516-5

Subramanian, S., Mishra, R. K., & Singh, L. (2003). Genome-wide analysis of microsatellite repeats in humans: Their abundance and density in specific genomic regions. *Genome Biology*, *4*(2), R13. https://doi.org/10.1186/gb-2003-4-2-r13

Sulovari, A., Li, R., Audano, P. A., Porubsky, D., Vollger, M. R., Logsdon, G. A., Human Genome Structural Variation Consortium, Warren, W. C., Pollen, A. A., Chaisson, M. J. P., Eichler, E. E., Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., … Lee, C. (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proceedings of the National Academy of Sciences*, *116*(46),

836 23243–23253. https://doi.org/10.1073/pnas.1912175116

837 Usdin, K. (2008). The biological effects of simple tandem repeats: Lessons from the repeat expansion
838 diseases: Table 1. *Genome Research*, *18*(7), 1011–1019. https://doi.org/10.1101/gr.070409.107

839 Verbiest, M., Maksimov, M., Jin, Y., Anisimova, M., Gymrek, M., & Bilgin Sonay, T. (2023). Mutation
840 and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity
841 across species. *Journal of Evolutionary Biology*, *36*(2), 321–336. https://doi.org/10.1111/jeb.14106

842 Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., & Verstrepen, K. J. (2009). Unstable Tandem
843 Repeats in Promoters Confer Transcriptional Evolvability. *Science*, *324*(5931), 1213–1216.
844 https://doi.org/10.1126/science.1170097

845 Wang, S., Wang, W., & Zeng, J. (2024). Role of CALCR expression in liver cancer: Implications for the
846 immunotherapy response. *Molecular Medicine Reports*, *31*(2), 41.
847 https://doi.org/10.3892/mmr.2024.13406

848 Wieder, N., D'Souza, E. N., Dawes, R., Chan, A., Martin-Geary, A., & Whiffin, N. (2025). The role of
849 untranslated region variants in Mendelian disease: A review. *European Journal of Human Genetics*,
850 *33*(9), 1096–1105. https://doi.org/10.1038/s41431-025-01905-x

851 Willems, T., Gymrek, M., Highnam, G., The 1000 Genomes Project Consortium, Mittelman, D., & Erlich,
852 Y. (2014). The landscape of human STR variation. *Genome Research*, *24*(11), 1894–1904.
853 https://doi.org/10.1101/gr.177774.114

854 Xiao, X., Zhang, C.-Y., Zhang, Z., Hu, Z., Li, M., & Li, T. (2022). Revisiting tandem repeats in
855 psychiatric disorders from perspectives of genetics, physiology, and brain evolution. *Molecular*
856 *Psychiatry*, *27*(1), 466–475. https://doi.org/10.1038/s41380-021-01329-1

857 Xu, D., Pavlidis, P., Thamadilok, S., Redwood, E., Fox, S., Blekhman, R., Ruhl, S., & Gokcumen, O.
858 (2016). Recent evolution of the salivary mucin MUC7. *Scientific Reports*, *6*(1), 31791.
859 https://doi.org/10.1038/srep31791

860 Yoo, D., Rhie, A., Hebbar, P., Antonacci, F., Logsdon, G. A., Solar, S. J., Antipov, D., Pickett, B. D.,
861 Safonova, Y., Montinaro, F., Luo, Y., Malukiewicz, J., Storer, J. M., Lin, J., Sequeira, A. N.,
862 Mangan, R. J., Hickey, G., Anez, G. M., Balachandran, P., … Eichler, E. E. (2024). *Complete*
863 *sequencing of ape genomes*. https://doi.org/10.1101/2024.07.31.605654

864 Yu, F., Lu, J., Liu, X., Gazave, E., Chang, D., Raj, S., Hunter-Zinck, H., Blekhman, R., Arbiza, L., Van
865 Hout, C., Morrison, A., Johnson, A. D., Bis, J., Cupples, L. A., Psaty, B. M., Muzny, D., Yu, J.,
866 Gibbs, R. A., Keinan, A., … Boerwinkle, E. (2015). Population Genomic Analysis of 962 Whole
867 Genome Sequences of Humans Reveals Natural Selection in Non-Coding Regions. *PLOS ONE*,
868 *10*(3), e0121644. https://doi.org/10.1371/journal.pone.0121644

869 Zhao, W., Huang, Y., Zhang, J., Liu, M., Ji, H., Wang, C., Cao, N., Li, C., Xia, Y., Jiang, Q., & Qin, J.
870 (2017). Polycomb group RING finger proteins 3/5 activate transcription via an interaction with the
871 pluripotency factor Tex10 in embryonic stem cells. *Journal of Biological Chemistry*, *292*(52),
872 21527–21537. https://doi.org/10.1074/jbc.M117.804054

873 Zhou, K., Aertsen, A., & Michiels, C. W. (2014). The role of variable DNA tandem repeats in bacterial
874 adaptation. *FEMS Microbiology Reviews*, *38*(1), 119–141. https://doi.org/10.1111/1574-6976.12036