Global abundance of short tandem repeats is non-random in rodents and primates

Masoud Arabfard¹, Mahmood Salesi¹, Yazdan Hassani Nourian¹, Iman Arabipour², AliMohammad Ali Maddi³, Kaveh Kavousi³ and Mina Ohadi⁴*

**Abstract**

**Background** While of predominant abundance across vertebrate genomes and significant biological implications, the relevance of short tandem repeats (STRs) (also known as microsatellites) to speciation remains largely elusive and attributed to random coincidence for the most part. Here we collected data on the whole-genome abundance of mono-, di-, and trinucleotide STRs in nine species, encompassing rodents and primates, including rat, mouse, olive baboon, gelada, macaque, gorilla, chimpanzee, bonobo, and human. The collected data were used to analyze hierarchical clustering of the STR abundances in the selected species.

**Results** We found massive differential STR abundances between the rodent and primate orders. In addition, while numerous STRs had random abundance across the nine selected species, the global abundance conformed to three consistent <clusters>, as follows: <rat, mouse>, <gelada, macaque, olive baboon>, and <gorilla, chimpanzee, bonobo, human>, which coincided with the phylogenetic distances of the selected species (p < 4E-05). Exceptionally, in the trinucleotide STR compartment, human was significantly distant from all other species.

**Conclusion** Based on hierarchical clustering, we propose that the global abundance of STRs is non-random in rodents and primates, and probably had a determining impact on the speciation of the two orders. We also propose the STRs and STR lengths, which predominantly conformed to the phylogeny of the selected species, exemplified by (t)10, (ct)6, and (taa4). Phylogenetic and experimental platforms are warranted to further examine the observed patterns and the biological mechanisms associated with those STRs.

**Keywords** Global, Short tandem repeat, Abundance, Non-random, Rodent, Primate, Hierarchical clustering

---

**Introduction**

Speciation is the evolutionary process by which populations evolve to become distinct species. Several models and theories have been proposed for this highly complicated process, including gene regulatory networks, community ecology, and mating preferences (for a review see [1]). Natural selection may be considered a major outcome associated with, and linking the above propositions. With an exceptionally high degree of polymorphism and plasticity, short tandem repeats (STRs) (also known as microsatellites/simple sequence repeats) may be a spectacular source of variation required for speciation and...
evolution [2–6]. The impact of STRs on speciation is supported by their various functional implications in gene expression, alternative splicing, and translation [4, 7–13].

STRs are a source of rapid and continuous morphological evolution [14], for example, in the evolution of facial length in mammals [15]. These highly evolving genetic elements may also be ideal responsive elements to fluctuating selective pressures. A role in evolutionary selection and adaptation is consistent with deep evolutionary conservation of some STRs, as “tuning knobs”, including several in genes with neurological and neurodevelopmental function[16].

While a limited number of studies indicate that purifying selection and drift can shape the structure of STRs at the inter- and intra-species levels [17–22], the global abundance of STRs at the crossroads of speciation remains largely unknown.

Mononucleotide and dinucleotide STRs are the most common categories of STRs in the vertebrate genomes [23, 24]. In addition to their association with frameshifts in coding sequences and pathological [25] and possibly evolutionary consequences, recent evidence indicates surprising functions for the mononucleotide STRs, such as their proposed role in translation initiation site selection [12, 26]. Several groups have found evidence on the involvement of a number of dinucleotide STRs in gene regulation, speciation, and evolution [4, 23, 27–30].

Trinucleotide STRs are frequently linked to human neurological disorders, most of which are specific to this species [31, 32].

Here, we analyzed the global hierarchical clustering of all types of mono-, di-, and trinucleotide STRs in nine mammalian species, encompassing primates and rodents. Those species belong to the superordinal group of Euar- chontoglires [33], and form three distinct and unambiguous phylogenetic <clusters>. The aim of this analysis was to examine whether the global abundance of STRs in the selected species conforms to the phylogenetic <clusters> of the selected species, or not.

Materials and methods
Species and whole-genome sequences
The UCSC genome browser (https://hgdownload.soe.ucsc.edu) was used to download and analyze the latest genome assemblies of nine species as follows (genome sizes are indicated below each species): rat (Rattus norvegicus): 2,647,915,728, mouse (Mus musculus): 2,728,222,451, gelada (Ther- opithecus gelada): 2,889,630,685, olive baboon (Papio anubis): 2,869,821,163, macaque (Macaca mulatta): 2,946,843,737, gorilla (Gorilla gorilla gorilla): 3,063,362,754, chimpanzee (Pan troglodytes): 3,050,398,082, bonobo (Pan paniscus): 3,203,531,224, and human (Homo sapiens): 3,099,706,404. Those species encompassed rodents: rat and mouse, Old World monkeys: gelada, olive baboon, macaque, and great apes: gorilla, bonobo, chimpanzee, human.

Extraction of STRs from genomic sequences
The whole-genome abundance of mononucleotide STRs of ≥10-repeats, dinucleotide STRs of ≥6-repeats, and trinucleotide STRs of ≥4-repeats were studied in the nine selected species. To that end, we designed a software package in Java (https://github.com/arabfard/Java_STR_Finder). All possibilities of mononucleotide motifs, consisting of A, C, T, and G, all possibilities of dinucleotide motifs, consisting of AC, AG, AT, CA, CG, CT, GA, GC, GT, TA, TC, and TG, and all possibilities of trinucleotide motifs, consisting of AAC, AAG, ACA, ACC, ACT, ACG, ATA, ATC, ATT, ATG, AGA, AGC, AGT, AGG, CAA, CAC, CAT, CAG, CCA, CCT, CCG, CTA, CTC, CTG, CGA, CGC, CGT, CGG, TAA, TAC, TAT, TAG, TCA, TCC, TCT, TCG, TTA, TTC, TTG, TGA, TGC, TGT, TGG, GAA, GAC, GAT, GAG, GCA, GCC, GCT, GCG, GGA, GTA, GTC, GTG, GTT, GGA, GGC, and GGT were analyzed.

The written program calculated based on perfect (pure) STRs. The algorithm started from an initial point, which was the first nucleotide of each genome, and iteratively repeated a series of steps during walking on the genome, nucleotide by nucleotide. In the first step, it investigated a window frame of 2*N, where 2 was the definition of tandem repeats i.e., two identical continuous sequences, and N was the length of the STR core. If the first half of the sequence inside the window was not equal to the second half, the algorithm moved one nucleotide forward. If equal, the algorithm checked the nucleotides, and this process continued until all identical continuous nucleotides, which were the same as the core were found. The final selected sequence- M*N- was introduced as a new STR, which had a core with a length of N and M repeats. All steps were repeated to find new STRs from the end of the previous STR. We repeated the algorithm for different values of N (N was between 1 and 3 in each genome to detected mono, di, and trinucleotide STRs).

Whole-genome STR data aggregation, abundance, and hierarchical cluster analysis across species
Whole-genome chromosome-by-chromosome data were aggregated and analyzed in the nine species. STR abundances across the selected species were obtained and depicted by boxplot diagrams and hierarchical clustering, using boxplot and hclust packages [34] in R, respectively. Boxplots illustrate abundance differences among segments across the selected species, and hierarchical clustering plots demonstrate the level of similarity and differences across the obtained abundances. The input data to these packages were numerical arrays. Each array
consisted of a number of columns, each column corresponding to the STR abundance in different chromosomes. It should be noted that the focus of our analysis was to evaluate the global abundance of STRs across those species, regardless of the homologous regions.

Statistical analysis
The STR abundances across the nine selected species were compared by repeated measurements analysis, using one and two-way ANOVA tests. These analyses were confirmed by nonparametric tests.

Results
Global abundance of mono, di, and trinucleotide STRs coincides with the phylogenetic distance of the nine selected species
Whole-genome data was collected on the abundance of mononucleotide STRs across the nine species (Table 1). We found massive expansion of the mononucleotide STR compartment in all primate species versus rat and mouse. Hierarchical clustering yielded three<clusters> as follows: <rat, mouse>, <gelada, olive baboon, macaque>, and <gorilla, chimpanzee, bonobo, human>, which coincided with the phylogenetic distance of the nine selected species (P=6.3E-09) (Fig. 1) namely <rodents>, <Old World monkeys>, and <great apes>.

The whole-genome STR abundances from aggregated chromosome-by-chromosome analysis in the di nucleotide category (Table 2) was decremented in primates versus rodents. Similar to the mononucleotide STR compartment, the dinucleotide STR compartment conformed to the genetic distance among the three<clusters> of species (P=7.1E-08) (Fig. 2).

There was global shrinkage of the trinucleotide STR compartment in primates versus rodents (P=3.8E-05) (Table 3; Fig. 3). Remarkably, human stood out among all other species in the trinucleotide STR compartment.

Differential abundance patterns of various STRs and STR lengths across rodents and primates
Numerous STRs and STR lengths across the mono, di, and trinucleotide STR categories coincided to the phylogenetic distances of the nine selected species, for example, in the instance of T/A mononucleotides of 10, 11, and 12 repeats, which were the most abundant STRs across all nine species (Fig. 4). In another example, (ct)6 and (taa)4 conformed to the phylogeny of the studied species in the di and trinucleotide STR categories, respectively.

On the other hand, numerous STRs did not follow perfect phylogenetic patterns, such as (C)10, (at)8, and (ttg)4 (Fig. 5). Hierarchical clusters of all studied STRs across

Table 1 Mononucleotide STR abundance across the nine selected species

| Chromosome/Species | Rat | Mouse | Gelada | Baboon | Macaque | Gorilla | Chimpanzee | Bonobo | Human |
|--------------------|-----|-------|--------|--------|---------|---------|------------|--------|-------|
| 1                  | 53,318 | 47,294 | 90,549 | 87,241 | 83,595 | 77,718 | 79,390 | 79,173 | 82,820 |
| 2(A)               | 46,221 | 45,636 | 71,588 | 67,963 | 64,609 | 35,908 | 35,897 | 34,400 | 78,550 |
| 2(B)               | 0     | 0     | 0     | 0     | 0     | 40,245 | 39,968 | 39,837 | 0     |
| 3                  | 36,364 | 38,493 | 70,736 | 68,688 | 65,836 | 62,398 | 62,713 | 64,472 | 64,027 |
| 4                  | 34,818 | 39,019 | 62,831 | 60,726 | 57,817 | 54,996 | 54,855 | 53,287 | 56,495 |
| 5                  | 36,532 | 38,805 | 66,164 | 64,101 | 61,533 | 60,436 | 49,944 | 54,142 | 56,538 |
| 6                  | 28,617 | 35,751 | 63,104 | 61,642 | 59,150 | 53,872 | 53,769 | 53,420 | 55,185 |
| 7(A)               | 29,411 | 33,649 | 25,699 | 65,267 | 63,438 | 50,898 | 53,882 | 50,792 | 56,257 |
| 7(B)               | 0     | 0     | 42,663 | 0     | 0     | 0     | 0     | 0     | 0     |
| 8                  | 27,353 | 31,938 | 50,576 | 48,446 | 46,757 | 43,593 | 44,212 | 43,618 | 45,220 |
| 9                  | 23,532 | 31,142 | 50,050 | 47,879 | 46,910 | 36,797 | 38,035 | 37,493 | 41,744 |
| 10                 | 31,065 | 34,138 | 41,475 | 39,012 | 37,477 | 44,166 | 44,562 | 44,416 | 46,075 |
| 11                 | 17,071 | 33,869 | 54,287 | 54,284 | 51,654 | 37,218 | 41,059 | 40,757 | 42,217 |
| 12                 | 15,101 | 29,325 | 42,675 | 35,365 | 42,793 | 46,865 | 47,576 | 47,481 | 48,483 |
| 13                 | 21,673 | 29,496 | 40,602 | 39,101 | 38,022 | 27,902 | 28,481 | 28,479 | 29,430 |
| 14                 | 21,835 | 28,833 | 45,820 | 44,693 | 42,677 | 30,311 | 30,659 | 30,595 | 31,460 |
| 15                 | 20,351 | 25,753 | 43,334 | 41,671 | 40,009 | 28,611 | 29,752 | 29,049 | 31,402 |
| 16                 | 19,958 | 24,139 | 41,211 | 39,781 | 37,693 | 29,268 | 31,121 | 28,460 | 34,364 |
| 17                 | 18,458 | 24,234 | 32,308 | 31,285 | 30,378 | 29,884 | 36,791 | 37,010 | 38,947 |
| 18                 | 16,651 | 22,580 | 25,310 | 24,850 | 23,551 | 22,556 | 22,428 | 22,236 | 23,130 |
| 19                 | 14,266 | 16,221 | 35,819 | 32,702 | 30,470 | 23,832 | 31,405 | 30,614 | 32,423 |
| 20                 | 14,475 | 0     | 34,962 | 32,965 | 32,095 | 20,654 | 22,106 | 31,034 | 21,961 |
| 21                 | 0     | 0     | 0     | 0     | 0     | 10,462 | 10,633 | 10,467 | 12,050 |
| 22                 | 0     | 0     | 0     | 0     | 0     | 13,778 | 14,816 | 13,904 | 16,014 |
| X                  | 25,983 | 40,547 | 52,836 | 49,013 | 47,590 | 43,138 | 43,302 | 41,656 | 46,178 |
| Sum                | 549,053 | 650,864 | 1,084,599 | 1,036,675 | 1,004,054 | 925,406 | 946,356 | 946,792 | 990,970 |

On the other hand, numerous STRs did not follow perfect phylogenetic patterns, such as (C)10, (at)8, and (ttg)4 (Fig. 5). Hierarchical clusters of all studied STRs across
the three categories are available at: https://figshare.com/articles/figure/STR_Clustering/17054972.

Discussion

While the mechanisms underlying speciation are extremely complicated and largely based on theories and
models, the impact of genetics seems to be significant in respect of adaptation, gene flow, and natural selection. In fact, natural selection may be a central converging point of the evolutionary propositions for speciation. However, the various mechanisms involved in speciation have different impact on natural selection, and it is the net effect

Table 3  Trinucleotide STR abundance across the nine selected species

| Chromosome/Species | Rat   | Mouse | Gelada | Baboon | Macaque | Gorilla | Chimpanzee | Bonobo | Human |
|--------------------|-------|-------|--------|--------|---------|---------|------------|--------|-------|
| 1                  | 25,234| 18,913| 16,307 | 15,350 | 15,341  | 14,540  | 15,219     | 15,054 | 14,882|
| 2(A)               | 22,996| 17,856| 13,005 | 12,341 | 11,998  | 6800    | 6842       | 6537   | 14,521|
| 2(B)               | 0     | 0     | 0      | 0      | 0       | 7545    | 7545       | 0      | 0     |
| 3                  | 16,869| 15,022| 12,749 | 11,154 | 10,960  | 11,116  | 10,380     | 10,504 | 10,445|
| 4                  | 17,088| 15,204| 11,921 | 11,116 | 11,128  | 11,228  | 10,264     | 10,504 | 9,832 |
| 5                  | 16,339| 15,469| 13,001 | 12,514 | 12,312  | 10,581  | 9,665      | 10,640 | 10,649|
| 6                  | 13,495| 14,332| 12,150 | 11,743 | 10,264  | 10,504  | 10,504     | 10,445 | 29,430|
| 7(A)               | 14,317| 13,760| 3937   | 10,991 | 10,871  | 9,342   | 10,117     | 9,744  | 9,995 |
| 7(B)               | 0     | 0     | 0      | 0      | 0       | 0       | 0          | 0      | 0     |
| 8                  | 12,701| 13,518| 10,032 | 9,524  | 9,682   | 8,752   | 9,096      | 8,645  | 8,990 |
| 9                  | 11,646| 12,378| 9,295  | 8,755  | 8,659   | 6,898   | 7,328      | 7,157  | 7,580 |
| 10                 | 12,552| 13,968| 7,297  | 6,728  | 6,786   | 8,096   | 8,530      | 8,245  | 8,925 |
| 11                 | 7,987 | 13,232| 9,615  | 9,578  | 9,403   | 7,801   | 8,668      | 8,458  | 8,352 |
| 12                 | 6,060 | 11,817| 7,742  | 8,297  | 8,029   | 8,905   | 9,218      | 9,051  | 9,127 |
| 13                 | 10,852| 11,634| 7,266  | 6,823  | 6,860   | 5,273   | 5,479      | 5,452  | 5,391 |
| 14                 | 10,325| 11,865| 8,869  | 8,583  | 8,253   | 5,473   | 5,771      | 5,785  | 5,706 |
| 15                 | 10,075| 10,693| 7,727  | 7,339  | 7,152   | 4,869   | 5,168      | 5,082  | 5,297 |
| 16                 | 8,476 | 9,527 | 6,228  | 5,837  | 5,801   | 5,738   | 6,007      | 5,623  | 6,402 |
| 17                 | 9,502 | 10,045| 5,908  | 5,737  | 5,684   | 5,666   | 5,859      | 5,914  | 6,091 |
| 18                 | 8,124 | 9,154 | 4,738  | 4,645  | 4,603   | 4,722   | 4,625      | 4,584  | 4,566 |
| 19                 | 6,984 | 6,190 | 5,432  | 4,643  | 4,664   | 3,807   | 5,438      | 5,230  | 5,101 |
| 20                 | 6,445 | 0     | 6,655  | 6,016  | 5,945   | 4,072   | 4,472      | 4,155  | 4,130 |
| 21                 | 0     | 0     | 0      | 0      | 0       | 2,051   | 2,092      | 2,028  | 2,304 |
| 22                 | 0     | 0     | 0      | 0      | 0       | 2,721   | 2,825      | 2,601  | 2,915 |
| X                  | 10,411| 13,783| 11,449 | 10,609 | 10,666  | 9,547   | 9,838      | 9,140  | 10,062|
| Sum                | 258,478| 258,360| 198,875| 189,725| 186,787 | 176,152 | 183,317    | 179,714| 202,461|

Fig. 2  Whole-genome dinucleotide STR abundance in the nine selected species. Global decremented patterns were observed in all primate species versus mouse and rat (left graph). The global pattern conformed to the three clusters across the nine species and their phylogenetic distance (right graph)
which may ultimately result in the emergence of a new species.

As one of the most abundant genetic elements in various animal genomes, it is largely unknown whether at the crossroads of speciation, STRs evolved as a result of purifying selection, genetic drift, and/or in a directional manner.

Here, we selected multiple species across rodents and primates, and investigated the clustering patterns of all possible types and lengths of mononucleotides, dinucleotide, and trinucleotide STRs on the whole-genome scale in those species. Hierarchical clustering yielded clusters that predominantly conformed to the phylogenetic distances of the selected species. Hierarchical clustering is an unsupervised clustering method that is used to group data. This algorithm is unsupervised because it uses random, unlabeled datasets. As the number of clusters increases, the accuracy of the hierarchical clustering algorithm improves.

Our findings may be of significance in a number of aspects. Firstly, there were significant differential abundances separating rodents from primates, for example, massive decremented abundance of dinucleotide and trinucleotide STRs in primates versus the rodent species, and massive incremented abundance of mononucleotide STRs in primates versus rodents. Secondly, the three major clusters obtained from global hierarchical cluster analysis matched the phylogeny of the three clusters of species, i.e., rodents, Old World monkeys, and great apes. It is possible that there are mathematical channels/thresholds required for the abundance of STRs in various orders. This is in line with the hypothesis that STRs function as scaffolds for biological computers[35]. In addition, our data indicate that various STRs and STR lengths behave differently with respect to their colossal abundance. Not all the studied STRs conformed to the phylogenetic distances of the nine selected species. We hypothesize that those which did, had a link with the speciation of those species, whereas those which did not, apparently followed random patterns for the most part. The potential effect of STRs in non-genic regions is largely unknown. However, when located at genic regions, various STRs and repeat lengths can potentially recruit transcription factors (TFs), which differ in qualitative and quantitative terms (http://algen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3) [36]. Those various TF sets may differentially regulate expression of the relevant genes during the process of evolution. For example, T-blocks of 10, 12, and 14-repeats recruit various combinations of FOXD3, HNF-3, and Hb (Fig. 6). Interestingly, (T)10 and (T)12 were among the mononucleotide STRs, which conformed to the phylogenetic distance of the nine species (Fig. 4), and (t)14 did not (https://figshare.com/articles/figure/STR_Clustering/17054972). The concept of various TF sets stands for other STRs as well. For example, (ct)6 conforms to the phylogenetic clusters, and recruits a number of TFs, whereas (ct)7, which does not conform to those clusters, recruits quantitatively different set of those TFs (Fig. 7).

Mononucleotide STRs impact various processes, such as gene expression, translation alterations, and frame-shifts of various proteins, which may have evolutionary and pathological consequences[12, 25]. They can overlap
with G4 structures, many of which associate with evolutionary consequences[37].

In a number of instances, dinucleotide STRs located in the protein-coding gene core promoters have been subject to contraction in the process of human and non-human primate evolution[38]. A number of those STRs are identical in formula in primates versus non-primates, and the genes linked to those STRs are involved in characteristics that have diverged primates from other mammals, such as craniofacial development, neurogenesis, and spine morphogenesis. Structural variants are enriched near genes that diverged in expression across great apes[39], and genes with STRs in their regulatory regions are more divergent in expression than genes with fixed or no STRs[40]. STR variants are likely to have epistatic interactions, which can have significant consequences in complex traits, in human as well as model organisms[6, 41].

Trinucleotide STRs are predominantly focused on in human because of their link with several neurological disorders[42–45]. We found an exceptional global hierarchical distance between human and all other species in that compartment. In view of the fact that most of the phenotypes attributed to trinucleotide STRs are human-specific in nature, it is conceivable that their evolution is also significantly distant from all other species studied.

The observed abundances were independent of the genome sizes of the selected species. For example in the instances of di- and trinucleotide STRs, we observed higher abundances in rodents versus primates despite the smaller genome sizes of the former. These findings are in line with the previous reports of lack of relationship between genome size and abundance of STRs[46, 47].

It should be noted that this is a pilot study based on hierarchical clustering, and future studies are warranted to further examine our hypothesis, using phylogenetic platforms and additional orders and species. Functional studies are also warranted to examine the biological impact of the relevant STRs.
Conclusion

We propose that the global abundance of STRs is non-random across rodents and primates. We also propose the STRs and STR lengths, which predominantly conformed to the phylogenetic distances of those species, such as (t)10, (ct)6, and (taa4). Additional species encompassing other orders and phylogenetic platforms are warranted to further examine this proposition.

Limitations

This research was a pilot study based on hierarchical clustering of the collected data in a number of mammalian species. Phylogenetic platforms and additional orders of species are warranted to further examine our hypothesis.

Fig. 5 Example of STRs and STR lengths, abundance of which appeared to be predominantly random across the nine selected species. Three STRs are depicted as examples for each of mono, di, and trinucleotide categories. Data from all studied STRs are available at: https://figshare.com/articles/figure/STR_Clustering/17054972
Abbreviations

STR  Short tandem repeat
TF    Transcription factor

Fig. 6 Potential recruitment of qualitatively and quantitatively different TFs to various lengths of (T)-repeats. (T)10 (A) and (T)12 (B) conformed to the phylogenetic clusters, whereas (T)14 (C) did not. Differential recruitment of TFs may differentially regulate the relevant genes in evolutionary processes.

Fig. 7 Potential differential TF recruitments to various lengths of (ct)6 A) and (ct)7 B). Those two lengths result in alternative quantitative binding of three TFs. (ct)6 conformed and (ct)7 did not conform to the phylogenetic clusters.

Abbreviations

STR  Short tandem repeat
TF    Transcription factor

Acknowledgements

Not applicable.
Authors’ contributions
MA performed and coordinated the bioinformatics analyses. MS performed the biostatistics analysis. YHN, IA, and AMAM contributed to data collection. IK contributed to coordination. MO conceived and supervised the project, and wrote the manuscript with input from all authors.

Funding
Not applicable.

Data Availability
Raw data are available at: https://figshare.com/articles/dataset/Trends/15073329 and https://figshare.com/articles/figure/STR_Clustering/17054972.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Authors have no conflict of interest to declare.

Received: 29 June 2022 / Accepted: 18 October 2022
Published online: 03 November 2022

References
1. Gavrilets S. Models of speciation: where are we now? J Hered. 2014;105(1):743–55.
2. Mohammadparast S, Bayat H, Biglarian A, Ohadi M. Exceptional expansion and conservation of a CT-repeat complex in the core promoter of PAXBP1 in primates. Am J Primatol. 2014;76(8):747–56.
3. Bushehri A, Barez MRM, Mansouri SK, Biglarian A, Ohadi M. Genome-wide identification of human- and primate-specific core promoter short tandem repeat repeats. Gene. 2016;587(1):83–90.
4. Nikkhah M, Rezazadeh M, Khorshid HRK, Biglarian A, Ohadi M. An exception- long CA-repeat in the core promoter of SCGB2B2 links with the evolution of apes and Old World monkeys. Gene. 2016;576(1):109–14.
5. Reinar WB, Lalun VO, Reitan T, Jakobsen KS, Butenko MA. Length variation in short tandem repeats affects gene expression in natural populations of Arabidopsis thaliana. Plant Cell. 2021;33(7):2221–34.
6. Press MO, Carlson KD, Queitsch C. The overdue promise of short tandem repeat variation for heritability. Trends Genet. 2014;30(11):504–12.
7. Jakubosky D, D’Antonio M, Bonder MJ, Small C, Donovan MKR, Greenwald WWY, Matsu H, D’Antonio-Chronowski AS, Stegle O, Smith EN. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. Nat Commun. 2020;11(1):1–15.
8. Valipour E, Kowsari A, Bayat H, Banan M, Kazeminasab S, Mohammadparast S, Ohadi M. Polymorphic core promoter GA-repeats alter gene expression of the early embryonic developmental gene. Genet. 2013;53(2):175–9.
9. Ranathunge C, Wheeler GL, Chimahusky ME, Perkins AD, Pramod S, Welch MF. Transcribed microsatellite allele lengths are often correlated with gene expression in natural sunflower populations. Molecular Ecology. 2020.
10. Press MO, Hall AN, Morton EA, Queitsch C. Substitutions are boring: some arguments for parallel mutations and high mutation rates. Trends Genet. 2019;35(4):253–64.
11. Fotsing SF, Margolish J, Wang C, Sani S, Yaniyck R, Shleizer-Burko S, Goren A, Gymeek M. The impact of short tandem repeat variation on gene expression. Nat Genet. 2019;51(1):1652–9.
12. Arabfard M, Kavousi K, Delbari A, Ohadi M. Link between short tandem repeats and translation initiation site selection. Hum Genomics. 2018;12(1):47.
13. Yaq K, Mukhina S, Zhang G, Tan JSC, Ong HS, Makeyev EV. A short tandem repeat-enriched RNA assembles a nuclear compartment to control alternative splicing and promote cell survival. Mol Cell. 2018;72(2):525–40.
14. Fondon JN, Gerner HR. Molecular origins of rapid and continuous morphological evolution. Proceedings of the National Academy of Sciences. 2004;101(52):18058–18063.
15. Wren JQ, Forgacs E, Fondon JN, Pettisemido A, Cheng SY, Gallardo T, Williams RS, Shoget RV, Minna JD, Garner HR. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. Am J Hum Genet. 2000;67(2):345–56.
16. King DG. Evolution of simple sequence repeats as mutable sites. Tandem Repeat Polymorphisms 2012:10–25.
17. Srivastava S, Anvaru AK, Sowpati DT, Mishra RK. Patterns of microsatellite distribution in Eukaryotic genomes: BMC Genomics. 2019;20(1):153.
18. Pavlova A, Gan HM, Lee YP, Austin CM, Gilligan DM, Lintemans M, Sunnucks P. Purifying selection and genetic drift shaped Pleistocene evolution of the mitochondrial genome in an endangered Australian freshwater fish. Heredity. 2017;118(5):466–76.
19. Jorde PE, Søvik G, Westgaard JL, Albretsen J, André C, Hvingel C, Johansen T, Sandvik AD, Kingsley M, Jørstad KE. Genetically distinct populations of northern shrimp, Pandalus borealis, in the North Atlantic: adaptation to different temperatures as an isolation factor. Mol Ecol. 2015;24(8):1742–57.
20. Legrand D, Cheneil T, Campagne C, Lachaise D, Carou ML. Inter-island divergence within Drosophila mauritiana, a species of the D. simulans complex. Past history and/or speciation in progress? Mol Ecol. 2011;20(3):2787–804.
21. Sun G, McGarvey ST, Bayoumi R, Mulligan CJ, Barrantes R, Raskin S, Zhong Y, Ayvei J, Chakraborty R, Dekk R. Global genetic variation at nine short tandem repeat loci and implications on forensic genetics. Eur J Hum Genet. 2003;11(1):39–49.
22. Abe H, Gemmell NJ. Evolutionary footprints of short tandem repeats in avian promoters. Sci Rep. 2016;6(1):1–11.
23. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W. Initial sequencing and analysis of the human genome. 2001.
24. Fan H, Chu JY. A brief review of short tandem repeat mutation. Genom Proteom Bioinform. 2007;5(1):7–14.
25. Mo HY, Lee JH, Kim MS, Yoo NJ, Lee SH. Frameshift Mutations and Loss of Expression of CLCN4 Gene are Frequently in Colorocutaneous Cancers With Microsatellite Instability. App Immunohistochem Mol Morphology. 2020;28(7):489–99.
26. Maddi AMA, Kavousi K, Arabfard M, Ohadi M. Tandem repeats ubiquitously flank and contribute to translation initiation sites. BMC Genomic Data. 2022;23(1):59.
27. Corney BPA, Widnell CL, Rees DJ, Davies JS, Crunelli V, Carter DA. Regulatory architecture of the neuronal Ca2cng2/Tarp2 gene promoter: multiple repressive domains, a polymorphic regulatory short tandem repeat, and bidirectional organization with co-regulated IncRNAs. J Mol Neurosci. 2019;67(2):282–94.
28. Emamalizadeh B, Movafagh A, Darvish H, Kazeminasab S, Andarva M, Namdar-Aligoordazi P, Ohadi M. The human RT2 core promoter short tandem repeat predominant allele is species-specific in length: a selective advantage for human evolution? Mol Genet Genomics. 2017;392(1):611–7.
29. Hassi RL, Johnson RC, Paviera BA. The effects of microsatellite selection on linked sequence diversity. Genome Biol Evol. 2014;6(7):1843–6.
30. Yin J-J, Adams AA, Kim JH, Holland SM. Evolution of an intronic microsatellite polymorphism in Toll-like receptor 2 among primates. Immunogenetics. 2006;58(9):740–5.
31. Arnaiz DJ, Vandeweyer G, Elrick E, Sánchez-Juan A, French CE, Raymond L, Kooy RF. Abundancy of polymorphic CAG repeats in the human genome suggest a broad involvement in neurological disease. Sci Rep. 2021;11(1):1–11.
32. Tang H, Kirkness EF, Lippert C, Biggs WH, Fabani M, Guzman E, Ramakrishnan S, Lavenkio V, Kakarodav B, Hou C. Profiling of short tandem repeat variation in human diseases 12,632 human whole genomes. Am J Hum Genet. 2017;100(1):700–15.
33. Kumar V, Hallstrom BM, Janke A. Coalescent-based genome analyses resolve the early branches of the eukaryotigales. PLoS ONE. 2013;8(4):e60019.
34. Murtugh F, Legendre P. Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion? J Classif. 2014;31(3):274–95.
35. Herbert A. Simple Paragraphs as Building Blocks for Genetic Computers. Trends in Genetics. 2020.
36. Faré D, Roset R, Huerta M, Adusua JE, Rosello L, Albà MM, Messeguer X. Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALIGNEN. Nucleic Acids Res. 2003;31(13):3651–3.
37. Sawaya S, Baghwar A, Buschhazo E, Kumar P, Chowdhury S, Black MA, Gemmell N. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS ONE. 2013;8(2):e54710.
38. Ohadi M, Valipour E, Ghadimi-Haddadan S, Namdar-Aligoordazi P, Bagheri A, Kowsari A, Rezaadeh M, Darvish H, Kazeminasab S. Core promoter short
tandem repeats as evolutionary switch codes for primate speciation. Am J Primatol. 2015;77(1):34–43.

Kronenberg ZN, Fiddes T, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJ, Dougherty ML. High-resolution comparative analysis of great ape genomes. Science 2018, 360(6393).

Sonay TB, Carvalho T, Robinson MD, Greminger MP, Krüzen M, Comas D, Highnam G, Mittelman D, Sharp A, Marques-Bonet T. Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. Genome Res. 2015;25(11):1591–9.

Bagshaw ATM, Horwood LJ, Fergusson DM, Gemmell NJ, Kennedy MA. Microsatellite polymorphisms associated with human behavioural and psychological phenotypes including a gene-environment interaction. BMC Med Genet. 2017;18(1):1–12.

Sundblom J, Niemelä V, Ghazarian M, Strand A-S, Bergdahl IA, Jansson J-H, Söderberg S, Stattin E-L. High frequency of intermediary alleles in the HTT gene in Northern Sweden-The Swedish Huntingtin Alleles and Phenotype (SHAPE) study. Sci Rep. 2020;10(1):1–7.

Baker EK, Arpino M, Kraan C, Bui M, Rogers C, Field M, Bretherton L, Ling L, Ure A, Cohen J. FMR1 mRNA from full mutation alleles is associated with ABC-C FX scores in males with fragile X syndrome. Sci Rep. 2020;10(1):1–8.

Zhou X, Wang C, Ding D, Chen Z, Peng Y, Peng H, Hou X, Wang P, Ye W, Li T. Analysis of (CAG)n expansion in ATXN1, ATXN2 and ATXN3 in Chinese patients with multiple system atrophy. Sci Rep. 2018;8(1):1–5.

Zhang Q, Yang M, Sørensen KK, Madsen CS, Boesen JT, An Y, Peng SH, Wei Y, Wang Q, Jensen KJ. A brain-targeting lipidated peptide for neutralizing RNA-mediated toxicity in Polyglutamine Diseases. Sci Rep. 2017;7(1):1–13.

Neff BD, Gross MR. Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. Evolution. 2001;55(9):1717–33.

Park JY, An Y-R, An C-M, Kang J-H, Kim EM, Kim H, Cho S, Kim J. Evolutionary constraints over microsatellite abundance in larger mammals as a potential mechanism against carcinogenic burden. Sci Rep. 2016;6(1):1–5.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.