Lei et al. BMC Genomics (2021) 22:421
https://doi.org/10.1186/s12864-021-07752-6

BMC Genomics

RESEARCH Open Access

# Genome-wide characterization of microsatellite DNA in fishes: survey and analysis of their abundance and frequency in genome-specific regions

Yi Lei[1], Yu Zhou[1], Megan Price[1] and Zhaobin Song[1,2*]

## Abstract

**Background:** Microsatellite repeats are ubiquitous in organism genomes and play an important role in the chromatin organization, regulation of gene activity, recombination and DNA replication. Although microsatellite distribution patterns have been studied in most phylogenetic lineages, they are unclear in fish species.

**Results:** Here, we present the first systematic examination of microsatellite distribution in coding and non-coding regions of 14 fish genomes. Our study showed that the number and type of microsatellites displayed nonrandom distribution for both intragenic and intergenic regions, suggesting that they have potential roles in transcriptional or translational regulation and DNA replication slippage theories alone were insufficient to explain the distribution patterns. Our results showed that microsatellites are dominant in non-coding regions. The total number of microsatellites ranged from 78,378 to 1,012,084, and the relative density varied from 4925.76 bp/Mb to 25,401.97 bp/Mb. Overall, (A + T)-rich repeats were dominant. The dependence of repeat abundance on the length of the repeated unit (1–6 nt) showed a great similarity decrease, whereas more tri-nucleotide repeats were found in exonic regions than tetra-nucleotide repeats of most species. Moreover, the incidence of different repeated types appeared species- and genomic-specific. These results highlight potential mechanisms for maintaining microsatellite distribution, such as selective forces and mismatch repair systems.

**Conclusions:** Our data could be beneficial for the studies of genome evolution and microsatellite DNA evolutionary dynamics, and facilitate the exploration of microsatellites structural, function, composition mode and molecular markers development in these species.

**Keywords:** Microsatellites, Genomic regions, Distribution patterns, Fish species

* Correspondence: zbsong@scu.edu.cn
[1]Sichuan Key Laboratory of Conservation Biology on Endangered Wildlife, College of Life Sciences, Sichuan University, Chengdu 610065, People's Republic of China
[2]Key Laboratory of Bio-Resources and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, People's Republic of China

## Background

Microsatellites, also termed as simple sequence repeats (SSRs), are short tandemly repeated sequences with 1-to-6 base pair (bp) motifs [1, 2]. They are ubiquitous and highly abundant in eukaryote, prokaryote and virus genomes [3–5], making up around 3% of the human genome [6]. Microsatellite instability is an important and unique form of mutation that is responsible for, or strongly implicated in, over 40 human neurological, neurodegenerative and neuromuscular disorders [7] and associations have also been observed in other complex diseases [2, 3, 8, 9]. Undoubtedly, microsatellites have attracted considerable attention due to their roles in the organization of chromosome structure, DNA recombination and replication, and gene expression and cell cycle dynamics [10].

Microsatellite analysis is used for a wide range of biological questions. Unique polymorphism of normal and disease-causing repeats can be used for disease diagnosis and prognosis [11–13]. Microsatellite repeats are advantageous as genetic markers due to their high polymorphism, informativeness and co-dominance, and have been used to construct quantitative trait loci (QTL) maps, genetic linkage maps [14–18] and DNA fingerprinting [19]. These features also provide the foundation for their successful application in other fundamental and applied fields of biology, including population and conservation genetics, genetic dissection of complex traits and marker-assisted breeding programs [10, 20–22].

Microsatellite content generally correlates positively with genome size [23–25]. The distribution of microsatellites exhibit different properties in genomes with different functionality [26–31], contradicting earlier studies stating that they are randomly distributed and simply represent “junk” DNA sequences [32]. Microsatellites are ubiquitously distributed across the entire genome, including protein-coding and non-coding regions [6, 33–35]. Previous studies have indicated that microsatellite occurrence differs significantly in coding and non-coding regions [36], and some microsatellite types were preferred and often common in genome-specific regions [26, 29]. Excessive microsatellite repeats occur in non-coding regions of eukaryotic organisms [37], whereas they are relatively rare in coding regions, ranging between 7 and 10% of higher plants [38, 39] and between 9 and 15% of vertebrates [40–42]. Meanwhile, multiple studies have demonstrated that the hotspots of microsatellite distribution may be related with various phenotypic traits [43, 44]. In the genome of *Saccharomyces cerevisiae* about 17% of genes contain microsatellite repeats in open reading frames (ORFs) [45, 46] and the repeats are specifically enriched in regulatory genes that encode transcription factors, DNA-RNA binding proteins and chromatin modifiers [47]. Microsatellite repeats in *cis*-regulatory elements and promoters, which frequently occur (e.g., ~ 25% promoters in yeast contain tandem repeats), regulate the process of gene expression [48, 49]. The (TTAGGG) n tracts constitute a substantial portion of the telomeric regions and are recognized by telomerase, which can be related to stability of chromosomes and nucleolus organizing regions [10, 50, 51].

xMicrosatellites are inherently unstable with high mutation rates from about $10^{-6}$ to $10^{-2}$ per locus per generation, resulting from DNA replication slippage [52, 53]. Mutation rates vary among microsatellite types (perfect, compound or interrupted), base composition of the repeat [54], repeat types (di-, tri- and tetranucleotide) [55, 56] and lengths [21], and heterozygosity [57, 58], but also among chromosome position, cell division, the GC content in flanking DNA and taxonomic groups [59–62]. Microsatellite instability has a strong influence on genomic microsatellite abundance and various functions and is explained by two mutually exclusive mutational mechanisms: (i) DNA replication slippage theory suggests that during DNA replication, the nascent and template strand realign out of register, and if DNA synthesis continues unabated on this molecule the repeat number of the microsatellite is altered [21, 63, 64]. The stability of the slipped structure has been maintained by hairpin, triplex, cruciform or quadruplex arrangement of DNA strands [65–69]. (ii) Unequal recombination theory assumes that large scale contractions and expansions of the repeat array involved the processes of DNA unequal recombination, including unequal crossing over and gene conversion [70], via a number of transposable elements, the best known are *Alu* and other short interspersed elements [64, 71]. Non-reciprocal recombination, random genetic drift and selective forces could have a significant effect on the accumulation of tandem-repetitive sequences in genomes [63, 65, 70].

So far, systematic research regarding microsatellite variation and characterization have been conducted on phylogenetic lineages, including humans [41], primates [72–74], plants and fungi [36, 75–82], and viruses [83, 84]. Yet microsatellite distribution patterns in fishes, an important branch of biological evolution, remained unclear. Here, 14 fish genomes have been used to indicate the microsatellite distribution patterns. The main objectives of the present study were to examine the distribution patterns of microsatellite in different fish genomes. The specific aims were 1) to examine the abundance and frequency of microsatellites in several important fish genomes, and 2) to compare the compositional differences of microsatellites in different taxa and genome-specific regions. We anticipate our study will provide foundational knowledge of microsatellite dynamics in fish species, helping us to better understand microsatellite distribution, and provide strong support for further exploration of genome structure and microsatellite functions.

## Materials and methods

### Genomic sequences

Genome sequences from 14 fish species, including model fishes (*Danio rerio, Oryzias latipes, Astyanax mexicanus*), commercial species (*Cyprinus carpio, Oncorhynchus mykiss, Oncorhynchus kisutch, Oreochromis niloticus, Ictalurus punctatus, Esox Lucius, Cynoglossus semilaevis,*), ornamental fishes (*Poecilia reticulate, Takifugu rubripes, Nothobranchius furzeri*), and "living fossil" fish species (*Lepisosteus oculatus*), were used in this study. Most genome sequences were downloaded from the Ensembl Genome Browser (Ensembl, Available online: http://asia.ensembl.org/index.html). The sequences of *Cyprinus carpio*, *Nothobranchius furzeri*, *Oncorhynchus kisutch*, and *Oncorhynchus mykiss* were obtained from the National Centre for Biotechnology Information (NCBI. Available online: http://www.ncbi.nlm.nih.gov/). We also obtained genome annotations to identify microsatellite locations in the genomes. The genomic (chromosomal) sequences that had complete genome annotations were included in this study. We filtered the unknown bases (Ns) in genome sequences using the Perl script and obtained the valid length of sequences for further analysis. The details of the genome sequences are listed in Table 1.

### Microsatellite identification

Microsatellites were identified from genome sequences using the Krait v0.9.0 program, a robust and ultrafast tool with a user-friendly graphic interface for genome-wide investigation of microsatellites [85]. We employed the perfect search model of the program to investigate the all motifs according to minimum repeats or minimum length of microsatellite. In the present study, we defined the perfect microsatellites as being mononucleotide repeats ≥12-bp, dinucleotide repeats ≥14-bp, trinucleotide repeats ≥15-bp, tetranucleotide repeats ≥16-bp, pentanucleotide repeats ≥20-bp and hexanucleotidue repeats ≥24-bp, and the length of flanking sequence was constrained to 200 bp, as previously described [72, 74]. We mainly examined the distribution of perfect repeats ≥12-bp long. The rationale for choosing the small cutoff value was that the microsatellites are often disrupted by single base substitutions [6, 33]. The occurrence of repeats in exons, introns and intergenic regions have been identified from the annotations of the 14 fish genome sequences using Perl scripts. The SciRoKo software tool [86] and the NCBI Graphical Sequence Viewer program (https://www.ncbi.nlm.nih.gov/projects/sviewer/) were employed to increase the reliability of the results for examined microsatellite repeats.

**Table 1** Data source and genome sizes of the 14 fish species studied in the present study

| Order | Family | Species | Abbreviation | Source | The number of chromosomes | Total length of sequences (Mbp)[a] | Total valid length of sequences (Mbp)[a] | Unknown bases (Ns) in sequences (%) |
|---|---|---|---|---|---|---|---|---|
| Characiformes | Characidae | *Astyanax mexicanus* | Amex | Ensemble | 25 | 930.61 | 896.76 | 3.75 |
| Cypriniformes | Cyprinidae | *Cyprinus carpio* | Ccar | NCBI | 50 | 825.07 | 809.20 | 1.91 |
| Cypriniformes | Cyprinidae | *Danio rerio* | Drer | Ensemble | 25 | 1345.10 | 1340.61 | 0.33 |
| Pleuronectiformes | Cynoglossidae | *Cynoglossus semilaevis* | Csem | Ensemble | 22 | 445.14 | 424.00 | 5.28 |
| Esociformes | Esocidae | *Esox lucius* | Eluc | Ensemble | 24 | 796.47 | 787.67 | 1.12 |
| Siluriformes | Ictaluridae | *Ictalurus punctatus* | Ipun | Ensemble | 29 | 762.01 | 752.24 | 1.29 |
| Lepisosteiformes | Lepisosteidae | *Lepisosteus oculatus* | Locu | Ensemble | 29 | 891.14 | 834.92 | 9.31 |
| Cyprinodontiformes | Nothobranchiidae | *Nothobranchius furzeri* | Nfur | NCBI | 19 | 1078.72 | 715.84 | 32.96 |
| Beloniformes | Adrianichthyidae | *Oryzias latipes* | Olat | Ensemble | 24 | 723.44 | 582.14 | 19.51 |
| Salmoniformes | Salmonidae | *Oncorhynchus kisutch* | Okis | NCBI | 30 | 1686.58 | 1628.60 | 3.43 |
| Salmoniformes | Salmonidae | *Oncorhynchus mykiss* | Omyk | NCBI | 29 | 1949.96 | 1754.52 | 10.09 |
| Cichliformes | Cichlidae | *Oreochromis niloticus* | Onil | Ensemble | 23 | 868.59 | 864.36 | 0.49 |
| Cyprinodontiformes | Poeciliidae | *Poecilia reticulata* | Pret | Ensemble | 23 | 696.70 | 637.75 | 8.43 |
| Tetraodontiformes | Tetraodontidae | *Takifugu rubripes* | Trub | Ensemble | 22 | 281.57 | 268.85 | 4.64 |

[a](Mbp) megabase pair

Repeats with unit patterns being circular permutations and/or reverse complements of each other were grouped together as one type. The total number of the non-overlapping type was 501 for 1–6 nt long motifs, with 1-nt motif containing two 2 types: A and C (A = T and C = G), 2-nt motif containing 4 types: AT, AG, AC and GC (AT = TA, AG = GA = CT = TC, AC=CA = GT = TG, and GC=CG), and 3–6-nt motif containing 10, 33, 102 and 350 types [41, 87].

## Results

### Distribution patterns of microsatellite repeats in the fish genomes

We examined the number, relative frequency (microsatellite numbers per Mb of the sequence), relative density (total microsatellite length per Mb of the sequence), GC content and the coverage degree (percentage of total microsatellites length in sequence) of microsatellites with motif lengths of 1–6 nucleotides in the 14 fish genomes (Table 2). We assigned 4-letter name abbreviations to the 14 species and these have been henceforth used to simplify results and discussion (e.g. *Danio rerio* = Drer, *O. niloticus* = Onil; see Table 2). The total number of microsatellites, ranging from 78,378 (Locu) to 1,012,084 (Drer), differed between fish species and the coverage degree varied from 0.18% (Locu) to 5.29% (Trub) (Table 2). The lowest relative frequency and relative density of microsatellites were both found in Olat (249.99 loci/Mb and 4925.76 bp/Mb, respectively) (Table 2). The highest relative frequency and density of microsatellites was found in Csem (3445.94 loci/Mb) and Ipun (25,401.97 bp/Mb), respectively. The GC content ranged from 10.94% (Trub) to 48.20% (Okis) (Table 2).

The main distribution pattern of di- (mononucleotide SSRs) > mono- (dinucleotide SSRs) > tetra- (trinucleotide SSRs) > tri- (tetranucleotide SSRs) > penta- (pentanucleotide SSRs) > hexanucleotide (hexanucleotide SSRs) was shared by six fish genomes (i.e., Drer, Trub, Ipun, Amex, Eluc, and Omyk), while a mono- > di- > tetra- > tri- > penta- > hexanucleotide pattern was observed in Olat, Ccar and Pret (Table 2). The di- > mono- > tri- > tetra- > penta- > hexanucleotide pattern was shared by Onil and Csem, whereas Nfur exhibited a di- > tetra- > mono- > tri- > penta- > hexanucleotide pattern (Table 2). The 1-nt or 2-nt repeats had a higher percentage motif abundance in the fish genomes than any other motif length, while the 2-nt repeats represented more than 60% motif abundance in Omyk, Eluc and Okis (Table 3). There was an almost equal distribution of motif abundance percentages between the first three motifs (1–3 nt) in Locu, with 3-nt and 1-nt repeats being almost identical (34.92, 34.99%, respectively). The percentage of 4-nt repeats was remarkably uniform across all taxa except for Drer and Locu, which had marginally greater or lesser percentages of this motif length (Table 3). Microsatellites with longer motifs (5–6 nt) showed lower percentages compared to the short motif repeats (1–4 nt). The 6-nt repeats had the lowest percentages among these motif lengths, ranging from 0.21% (Drer) to 0.85% (Trub) (Table 3).

### Mononucleotide repeats

Motif abundance percentages of mononucleotide repeats within intergenic regions, introns and exons varied across species, with intergenic regions ranging from 0.19% (Okis) to 7.19% (Trub), introns ranging from 9.87% (Nfur) to 45.06% (Olat) and exons ranging from 0.37% (Okis) to 3.58% (Pret) (Table 3). Among the two types of mononucleotide repeats, poly(A/T) was generally far more abundant than poly(C/G) in these fish genomes, except that the reverse was found in the Trub, Omyk and Okis genome sequences (Supplement Table 1, Tables 4 and 5). Drer had the maximum repeat number of A (or T) (192,264) followed by Ccar, Ipun, Amex and Pret. Pret contained the maximum number (4549) of C (or G). Although poly(A/T) tracts were clearly more abundant than poly(C/G) in exons (Table 4), this difference was not consistently observed in introns (Table 5) and intergenic regions (Table 6).

### Dinucleotide repeats

Among genome-specific regions, there was a lower percentage of dimer repeats (AT, AC, AG, CG) in exons compared to non-coding regions, ranging from 0.37% (Ccar) to 1.77% (Onil). Within the non-coding regions, intronic regions have a higher proportion of dinucleotide repeats compared to the intergenic regions (Table 3). We found that (AC) n repeats were generally more numerous in specific genomic regions, except that Amex had greater numbers of (AG) n repeats in exonic regions and Okis had greater (AT) n repeats in intronic regions (Tables 4 and 5). The number of (AT) n repeats observed the greatest variation between genome-specific regions and species. For example, intronic or intergenic regions of Drer have similar numbers of (AT) n repeats to (AG) n repeats, whereas exon numbers of (AT) n repeats were considerably less than (AG) n repeats. Olat had more (AT) n repeats than (AG) n in exons, but the opposite was found in other genomes. Finally, (CG) n repeats were very infrequent or absent in these genomes.

### Trinucleotide repeats

Motif abundance percentages of trinucleotide repeats in the exons of six fish species were greater than in intergenic regions, these six species being Olat (1.90%), Csem (1.81%), Pret (1.63%), Onil (1.53%) and Eluc (0.78%) (Table 3). Meanwhile, motif abundance percentages of trinucleotide repeats in the exons of Locu, Csem and Olat were greater than other motif lengths (e.g. mono-,

**Table 2** Microsatellite distribution as frequency, density and GC content of different fish genomes

| Repeat type | Species | Abbreviation | Total numbers | Total length (bp) | Relative abundace (loci/Mb) | Relative Density (bp/Mb) | Coverage degree (%) | G + C | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Length (bp) | content (%) |
| Mono- (1-nt) | *D. rerio* | Drer | 207,951 | 3,278,888 | 155.12 | 2445.81 | 0.24 | 156,865 | 4.78 |
| | *O. niloticus* | Onil | 85,303 | 1,316,072 | 98.69 | 1522.59 | 0.15 | 139,335 | 10.59 |
| | *O. latipes* | Olat | 76,805 | 1,120,875 | 131.94 | 1925.44 | 0.19 | 198,155 | 17.68 |
| | *T. rubripes* | Ttub | 34,748 | 504,976 | 129.25 | 1878.28 | 0.19 | 264,888 | 52.46 |
| | *I. punctatus* | Ipun | 176,856 | 2,991,388 | 237.37 | 4013.86 | 0.40 | 343,854 | 11.49 |
| | *C. semilaevis* | Csem | 97,303 | 1,483,637 | 232.06 | 3539.85 | 0.35 | 231,161 | 15.58 |
| | *E. lucius* | Eluc | 45,139 | 667,053 | 62.88 | 928.80 | 0.08 | 326,439 | 48.94 |
| | *O. mykiss* | Omyk | 90,887 | 1,336,990 | 51.64 | 759.16 | 0.08 | 690,373 | 51.64 |
| | *N. furzeri* | Nfur | 50,594 | 805,921 | 71.06 | 1131.40 | 0.11 | 364,314 | 45.20 |
| | *L. oculatus* | Locu | 26,639 | 347,118 | 39.34 | 517.74 | 0.04 | 80,197 | 23.10 |
| | *A. mexicanus* | Amex | 170,345 | 3,398,972 | 189.44 | 3753.31 | 0.38 | 266,645 | 7.84 |
| | *C. carpio* | Ccar | 186,935 | 2,883,267 | 230.83 | 3558.04 | 0.37 | 84,049 | 2.91 |
| | *P. reticulata* | Pret | 119,330 | 1,775,344 | 188.83 | 2810.96 | 0.23 | 66,159 | 3.73 |
| | *O. kisutch* | Okis | 96,581 | 1,542,599 | 59.64 | 952.88 | 0.09 | 993,999 | 64.44 |
| Di- (2-nt) | *D. rerio* | Drer | 382,761 | 14,766,368 | 285.51 | 11,014.63 | 1.10 | 3,485,035 | 23.60 |
| | *O. niloticus* | Onil | 123,095 | 3,294,102 | 142.41 | 3811.02 | 0.38 | 1,337,732 | 40.61 |
| | *O. latipes* | Olat | 25,449 | 515,736 | 43.72 | 885.93 | 0.09 | 203,111 | 39.38 |
| | *T. rubripes* | Trub | 63,408 | 12,729,440 | 235.85 | 6432.72 | 4.73 | 837,010 | 6.58 |
| | *I. punctatus* | Ipun | 359,022 | 10,686,166 | 484.03 | 14,362.27 | 1.42 | 3,777,407 | 35.35 |
| | *C. semilaevis* | Csem | 127,607 | 2,668,806 | 2998.81 | 6238.79 | 0.63 | 1,071,149 | 40.14 |
| | *E. lucius* | Eluc | 150,812 | 3,215,024 | 212.80 | 4533.65 | 0.41 | 1,466,606 | 45.62 |
| | *O. mykiss* | Omyk | 390,181 | 11,517,856 | 225.44 | 7684.48 | 0.66 | 5,186,111 | 45.03 |
| | *N. furzeri* | Nfur | 186,289 | 6,543,766 | 261.89 | 9175.17 | 0.91 | 3,065,853 | 46.85 |
| | *L. oculatus* | Locu | 16,853 | 289,242 | 30.03 | 527.85 | 0.03 | 122,671 | 42.41 |
| | *A. mexicanus* | Amex | 446,448 | 12,194,636 | 498.90 | 13,549.68 | 1.36 | 3,706,651 | 30.40 |
| | *C. carpio* | Ccar | 181,614 | 5,801,364 | 224.78 | 7173.27 | 0.72 | 2,057,824 | 35.47 |
| | *P. reticulata* | Pret | 70,237 | 1,810,180 | 110.24 | 2843.81 | 0.28 | 858,108 | 47.40 |
| | *O. kisutch* | Okis | 502,660 | 12,625,074 | 310.50 | 7784.45 | 0.78 | 6,023,864 | 47.71 |
| Tri- (3-nt) | *D. rerio* | Drer | 140,175 | 3,374,835 | 104.56 | 2517.38 | 0.25 | 262,890 | 7.79 |
| | *O. niloticus* | Onil | 33,649 | 662,808 | 38.93 | 766.82 | 0.08 | 166,799 | 25.17 |
| | *O. latipes* | Olat | 15,732 | 295,236 | 27.02 | 507.16 | 0.05 | 104,498 | 35.39 |
| | *T. rubripes* | Ttub | 12,671 | 262,275 | 47.13 | 975.54 | 0.10 | 117,321 | 44.73 |
| | *I. punctatus* | Ipun | 91,842 | 2,466,039 | 123.29 | 3302.90 | 0.32 | 324,772 | 13.17 |
| | *C. semilaevis* | Csem | 48,581 | 1,052,241 | 114.46 | 2478.47 | 0.25 | 408,981 | 38.87 |
| | *E. lucius* | Eluc | 11,632 | 208,425 | 15.88 | 284.64 | 0.03 | 62,534 | 30.03 |
| | *O. mykiss* | Omyk | 37,290 | 704,511 | 21.39 | 404.12 | 0.05 | 210,431 | 29.87 |
| | *N. furzeri* | Nfur | 39,742 | 918,591 | 55.80 | 1288.97 | 0.13 | 221,669 | 24.13 |
| | *L. oculatus* | Locu | 26,590 | 644,715 | 35.80 | 822.92 | 0.08 | 93,309 | 14.47 |
| | *A. mexicanus* | Amex | 61,775 | 2,572,995 | 69.16 | 2833.21 | 0.29 | 333,193 | 12.95 |
| | *C. carpio* | Ccar | 53,774 | 1,273,065 | 66.86 | 1588.10 | 0.16 | 127,770 | 10.04 |
| | *P. reticulata* | Pret | 24,815 | 616,578 | 39.01 | 966.24 | 0.10 | 144,894 | 23.50 |
| | *O. kisutch* | Okis | 34,901 | 643,725 | 21.72 | 401.60 | 0.04 | 213,491 | 33.16 |
| Tetra- (4-nt) | *D. rerio* | Drer | 243,740 | 8,195,596 | 181.81 | 6113.32 | 0.61 | 2,326,652 | 28.39 |

**Table 2** Microsatellite distribution as frequency, density and GC content of different fish genomes *(Continued)*

| Repeat type | Species | Abbreviation | Total numbers | Total length (bp) | Relative abundace (loci/Mb) | Relative Density (bp/Mb) | Coverage degree (%) | G + C | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Length (bp) | content (%) |
| | *O. niloticus* | Onil | 32,971 | 731,160 | 38.14 | 845.90 | 0.08 | 210,529 | 28.79 |
| | *O. latipes* | Olat | 22,084 | 776,628 | 37.94 | 1334.10 | 0.13 | 250,470 | 32.25 |
| | *T. rubripes* | Trub | 16,187 | 541,848 | 60.21 | 2015.43 | 0.20 | 231,326 | 42.69 |
| | *I. punctatus* | Ipun | 108,518 | 2,422,148 | 145.59 | 3347.91 | 0.32 | 526,093 | 21.72 |
| | *C. semilaevis* | Csem | 34,100 | 768,504 | 79.96 | 1807.35 | 0.18 | 277,262 | 36.08 |
| | *E. lucius* | Eluc | 21,952 | 411,372 | 30.05 | 562.93 | 0.05 | 191,704 | 46.60 |
| | *O. mykiss* | Omyk | 70,423 | 1,850,584 | 40.46 | 1062.57 | 0.11 | 826,372 | 44.65 |
| | *N. furzeri* | Nfur | 55,909 | 1,531,664 | 78.34 | 2142.29 | 0.21 | 530,688 | 34.65 |
| | *L. oculatus* | Locu | 4439 | 92,396 | 7.76 | 153.69 | 0.01 | 35,673 | 38.61 |
| | *A. mexicanus* | Amex | 84,721 | 3,160,820 | 95.30 | 3525.04 | 0.35 | 884,624 | 27.99 |
| | *C. carpio* | Ccar | 62,915 | 1,749,772 | 77.78 | 2162.41 | 0.22 | 374,219 | 21.39 |
| | *P. reticulata* | Pret | 32,055 | 1,138,844 | 50.22 | 1777.93 | 0.18 | 421,839 | 37.04 |
| | *O. kisutch* | Okis | 90,546 | 2,527,184 | 55.65 | 1551.12 | 0.16 | 1,163,069 | 46.02 |
| Penta- (5-nt) | *D. rerio* | Drer | 35,377 | 1,419,895 | 26.39 | 1059.14 | 0.11 | 104,811 | 7.38 |
| | *O. niloticus* | Onil | 12,089 | 316,950 | 13.99 | 366.69 | 0.04 | 69,140 | 21.81 |
| | *O. latipes* | Olat | 4973 | 145,115 | 8.54 | 249.28 | 0.02 | 36,644 | 25.25 |
| | *T. rubripes* | Ttub | 3871 | 164,760 | 14.40 | 612.83 | 0.06 | 87,874 | 53.33 |
| | *I. punctatus* | Ipun | 9556 | 234,630 | 12.84 | 315.11 | 0.03 | 50,793 | 21.65 |
| | *C. semilaevis* | Csem | 5828 | 186,075 | 13.57 | 428.73 | 0.04 | 54,885 | 29.50 |
| | *E. lucius* | Eluc | 1140 | 26,200 | 1.55 | 35.77 | 0.00 | 11,874 | 45.32 |
| | *O. mykiss* | Omyk | 13,985 | 342,135 | 7.96 | 194.80 | 0.02 | 122,566 | 35.82 |
| | *N. furzeri* | Nfur | 10,079 | 275,340 | 14.04 | 382.28 | 0.04 | 76,366 | 27.74 |
| | *L. oculatus* | Locu | 2292 | 56,690 | 3.10 | 73.65 | 0.01 | 20,527 | 36.21 |
| | *A. mexicanus* | Amex | 16,919 | 1,094,190 | 18.90 | 1187.90 | 0.12 | 282,448 | 25.81 |
| | *C. carpio* | Ccar | 15,817 | 686,070 | 19.57 | 850.94 | 0.08 | 79,829 | 11.64 |
| | *P. reticulata* | Pret | 3815 | 125,295 | 5.97 | 195.30 | 0.02 | 32,651 | 26.06 |
| | *O. kisutch* | Okis | 17,343 | 428,935 | 10.64 | 263.32 | 0.03 | 166,246 | 38.76 |
| Hexa- (6-nt) | *D. rerio* | Drer | 2080 | 58,950 | 1.55 | 43.97 | 0.00 | 19,507 | 33.09 |
| | *O. niloticus* | Onil | 966 | 26,856 | 1.12 | 31.07 | 0.00 | 8832 | 32.89 |
| | *O. latipes* | Olat | 485 | 13,884 | 0.83 | 23.85 | 0.00 | 5484 | 39.50 |
| | *T. rubripes* | Trub | 1127 | 34,446 | 4.19 | 128.12 | 0.01 | 18,573 | 53.92 |
| | *I. punctatus* | Ipun | 1671 | 44,820 | 2.24 | 59.92 | 0.01 | 15,581 | 34.76 |
| | *C. semilaevis* | Csem | 2996 | 106,962 | 7.08 | 252.92 | 0.03 | 50,107 | 48.85 |
| | *E. lucius* | Eluc | 712 | 23,232 | 1.01 | 33.21 | 0.00 | 11,962 | 51.49 |
| | *O. mykiss* | Omyk | 4590 | 140,352 | 2.66 | 81.33 | 0.01 | 71,447 | 50.91 |
| | *N. furzeri* | Nfur | 1664 | 44,028 | 2.33 | 61.70 | 0.01 | 20,000 | 45.42 |
| | *L. oculatus* | Locu | 1565 | 53,070 | 1.84 | 62.81 | 0.01 | 23,037 | 43.41 |
| | *A. mexicanus* | Amex | 2179 | 193,794 | 2.43 | 209.55 | 0.02 | 60,384 | 31.16 |
| | *C. carpio* | Ccar | 1724 | 69,300 | 2.13 | 88.33 | 0.01 | 20,736 | 29.92 |
| | *P. reticulata* | Pret | 786 | 21,594 | 1.23 | 33.73 | 0.00 | 11,193 | 51.83 |
| | *O. kisutch* | Okis | 5178 | 169,488 | 3.20 | 104.70 | 0.01 | 85,527 | 50.46 |
| Total | *D. rerio* | Drer | 1,012,084 | 31,094,532 | 754.94 | 23,194.25 | 2.31 | 6,355,760 | 20.44 |
| | *O. niloticus* | Onil | 288,073 | 6,347,948 | 333.28 | 7344.09 | 0.73 | 1,932,367 | 30.44 |

**Table 2** Microsatellite distribution as frequency, density and GC content of different fish genomes *(Continued)*

| Repeat type | Species | Abbreviation | Total numbers | Total length (bp) | Relative abundace (loci/Mb) | Relative Density (bp/Mb) | Coverage degree (%) | G + C | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Length (bp) | content (%) |
| | *O. latipes* | Olat | 145,528 | 2,867,474 | 249.99 | 4925.76 | 0.48 | 798,362 | 27.84 |
| | *T. rubripes* | Trub | 132,012 | 14,237,745 | 491.03 | 12,042.92 | 5.29 | 1,556,992 | 10.94 |
| | *I. punctatus* | Ipun | 747,465 | 18,845,191 | 1005.36 | 25,401.97 | 2.50 | 5,038,500 | 26.74 |
| | *C. semilaevis* | Csem | 316,415 | 6,266,225 | 3445.94 | 14,746.11 | 1.48 | 2,093,545 | 33.41 |
| | *E. lucius* | Eluc | 231,387 | 4,551,306 | 324.17 | 6379.00 | 0.57 | 2,071,119 | 45.51 |
| | *O. mykiss* | Omyk | 607,356 | 15,892,428 | 349.55 | 10,186.46 | 0.93 | 7,107,300 | 44.72 |
| | *N. furzeri* | Nfur | 344,277 | 10,119,310 | 483.46 | 14,181.81 | 1.41 | 4,278,890 | 42.28 |
| | *L. oculatus* | Locu | 78,378 | 1,483,231 | 117.87 | 2158.66 | 0.18 | 375,414 | 25.31 |
| | *A. mexicanus* | Amex | 782,387 | 22,615,407 | 874.13 | 25,058.69 | 2.52 | 5,533,945 | 24.47 |
| | *C. carpio* | Ccar | 502,779 | 12,462,838 | 621.95 | 15,421.09 | 1.56 | 2,744,427 | 22.02 |
| | *P. reticulata* | Pret | 251,038 | 5,487,835 | 395.50 | 8627.97 | 0.81 | 1,534,844 | 27.97 |
| | *O. kisutch* | Okis | 747,209 | 17,937,005 | 461.35 | 11,058.07 | 1.11 | 8,646,196 | 48.20 |

di-). Among the different trinucleotide repeats, (AAT) n repeats were generally the most numerous repeats in intronic and intergenic regions of different taxa (Tables 5 and 6), except for Okis where (ACT) n repeats were the most numerous in intergenic regions (Table 6). There was no one trinucleotide repeat in exonic regions that was typically more numerous than another across the different fish species. For example, (AAT) n repeats were most numerous in Drer and Ipun, while (ATC) n repeats were greater in Ccar and Nfur and (AGG) n repeats were greatest in the 10 remaining species (Table 4). Repeats such as ACT, ACT, AGC, ACG and CCG were generally in low numbers in each specific genomic region. Furthermore, CCG repeats were absent in the intergenic regions of Eluc, Omyk, Ccar and Okis (Table 6).

### Tetranucleotide repeats

Tetranucleotide repeats were frequent in each genomic region and were generally dependent on the base composition of the repeat unit (Tables 7, 8, 9 and Supplement Table 1). Overall, repeats with > 50% of A + T (e.g. AAAT, ATAG and AATC repeats) were more abundant in studied fish genomes (Supplement Table 1). There were, however, a few notable exceptions. For example, (ACAG) n repeats were the most numerous in Eluc, Omyk and Okis (Supplement Table 1). We found that the (AAAB) n repeats (where B denotes any base other than A) were most numerous in exonic regions in five fish species (i.e. Olat, Drer, Onil, Ccar and Ipun), the (ACAG) n repeats were numerous in Eluc, Omyk and Okis, and (ATCC) n repeats were most common of the remaining four fish species (Table 7). Similar to exons, the most common tetranucleotide repeat in intergenic regions was (AAAB) n, except for (ATCC) n in Olat, (AATC) n in Eluc and (ATAG) n in Amex (Table 9). In introns, (AATB) n or (ACAG) n were the most common tetranucleotide repeats in studied fish (Table 8). We also found some repeats with > 50% of C + G (e.g. ACGC, AGGG and AGCG repeats) were in the top 50% of tetranucleotide repeats in specific genome regions (Tables 7, 8, 9).

### Pentanucleotide repeats

As expected, the occurrence pentanucleotide repeats was less than tetranucoeitde repeats in different genome regions. We found a general distribution pattern of pentanucleotide repeats for all species, where (A + T)-rich repeats were the most abundant. Yet, we still found notable exceptions where (C + G)-rich repeats were dominant in specific genomic locations, including AGAGG and ACTGG in introns or intergenic regions of Trub, Csem and Okis and ACTGC in exons of Eluc (Tables 7, 8, 9). Although AGAGG repeats in introns and exons were relatively abundant in Csem, it was also the only species that lacked this repeat in intergenic regions in this study (Supplement Table 1). We also found that the CpG-containing repeats were present in the top 50% of pentanucleotide repeats, including (ATACG) n or (CCCGG) n tracts in intronic regions of Eluc and Locu, (CCCGG) n, (AATCG) n or (ACCGG) n tracts in exonic regions of Trub, Amex and Pret, and (ATACG) n or (ACCGG) n tracts in intergenic regions of Eluc and Pret (Tables 7, 8, 9).

### Hexanucleotide repeats

Hexanucleotide repeats were the least numerous in specific genomic regions, except for the exons of Trub (Table 3). In exonic and intronic regions, a dominance of (C + G)-rich repeats was found in the majority of the genomes (Tables 7 and 8). The repeat motifs present in intergenic regions were highly variable and relatively

**Table 3** Microsatellite distribution as percent motif abundance (%) among 14 genomes

| Species | Abbreviation | Genomic region | Mono- (1-nt) | Di- (2-nt) | Tri- (3-nt) | Tetra- (4-nt) | Penta- (5-nt) | Hexa- (6-nt) |
|---|---|---|---|---|---|---|---|---|
| *D. rerio* | Drer | all | 20.54 | 37.82 | 13.85 | 24.08 | 3.50 | 0.21 |
| | | intergenic regions | 2.10 | 5.66 | 1.79 | 3.42 | 0.40 | 0.02 |
| | | introns | 18.06 | 31.55 | 11.72 | 20.54 | 3.10 | 0.18 |
| | | exons | 0.39 | 0.60 | 0.34 | 0.12 | 0.01 | 0.00 |
| *O. niloticus* | Onil | all | 29.61 | 42.73 | 11.68 | 11.45 | 4.20 | 0.34 |
| | | intergenic regions | 1.91 | 4.23 | 1.15 | 0.89 | 0.53 | 0.07 |
| | | introns | 25.15 | 36.73 | 9.01 | 10.02 | 3.55 | 0.24 |
| | | exons | 2.56 | 1.77 | 1.53 | 0.53 | 0.12 | 0.02 |
| *O. latipes* | Olat | all | 52.77 | 17.49 | 10.81 | 15.18 | 3.42 | 0.33 |
| | | intergenic regions | 6.26 | 2.19 | 1.37 | 2.21 | 0.55 | 0.05 |
| | | introns | 45.06 | 14.57 | 7.54 | 12.56 | 2.81 | 0.27 |
| | | exons | 1.45 | 0.73 | 1.90 | 0.40 | 0.06 | 0.02 |
| *T. rubripes* | Trub | all | 26.32 | 48.03 | 9.60 | 12.26 | 2.93 | 0.85 |
| | | intergenic regions | 7.19 | 13.91 | 2.79 | 4.15 | 1.11 | 0.28 |
| | | introns | 17.07 | 32.71 | 5.40 | 7.81 | 1.78 | 0.52 |
| | | exons | 2.06 | 1.42 | 1.42 | 0.30 | 0.04 | 0.05 |
| *N. furzeri* | Nfur | all | 14.71 | 54.15 | 11.55 | 16.2 | 2.91 | 0.48 |
| | | intergenic regions | 4.53 | 17.11 | 3.83 | 5.66 | 1.13 | 0.16 |
| | | introns | 9.87 | 37.05 | 7.25 | 10.79 | 1.85 | 0.34 |
| | | exons | 0.57 | 1.25 | 0.73 | 0.24 | 0.04 | 0.01 |
| *O.mykiss* | Omyk | all | 14.91 | 64.33 | 6.12 | 11.59 | 2.29 | 0.75 |
| | | intergenic regions | 0.26 | 1.27 | 0.24 | 0.3 | 0.11 | 0.04 |
| | | introns | 14.17 | 61.7 | 5.27 | 11.22 | 2.18 | 0.7 |
| | | exons | 0.57 | 1.27 | 0.63 | 0.24 | 0.02 | 0.02 |
| *A. mexicanus* | Amex | all | 21.75 | 57 | 7.29 | 10.88 | 2.17 | 0.28 |
| | | intergenic regions | 1.05 | 3.26 | 0.53 | 0.77 | 0.16 | 0.04 |
| | | introns | 20.03 | 52.99 | 6.84 | 9.89 | 1.97 | 0.23 |
| | | exons | 0.68 | 0.81 | 0.52 | 0.17 | 0.03 | 0.01 |
| *E. lucius* | Eluc | all | 19.51 | 65.18 | 5.03 | 9.49 | 0.49 | 0.31 |
| | | intergenic regions | 1.1 | 3.45 | 0.33 | 0.7 | 0.06 | 0.04 |
| | | introns | 17.19 | 60.03 | 3.92 | 8.54 | 0.42 | 0.26 |
| | | exons | 1.22 | 1.7 | 0.78 | 0.25 | 0.01 | 0.01 |
| *I. punctatus* | Ipun | all | 23.68 | 48 | 12.3 | 14.25 | 1.28 | 0.22 |
| | | intergenic regions | 1.17 | 3.05 | 0.82 | 0.89 | 0.09 | 0.02 |
| | | introns | 20.97 | 43.26 | 10.64 | 13.18 | 1.12 | 0.19 |
| | | exons | 1.52 | 1.71 | 0.82 | 0.45 | 0.07 | 0.01 |
| *C. semilaevis* | Csem | all | 31.05 | 39.78 | 15.25 | 11.11 | 1.85 | 0.96 |
| | | intergenic regions | 1.79 | 1.63 | 0.78 | 0.76 | 0.11 | 0.07 |
| | | introns | 27.47 | 37.29 | 12.76 | 9.72 | 1.66 | 0.8 |
| | | exons | 1.49 | 1.41 | 1.81 | 0.3 | 0.07 | 0.07 |
| *L. oculatus* | Locu | all | 34.99 | 21.5 | 34.92 | 5.66 | 2.93 | 2 |
| | | intergenic regions | 6.57 | 4.81 | 7.46 | 1.26 | 0.73 | 0.51 |
| | | introns | 24.71 | 15.27 | 23.44 | 4.06 | 2.05 | 1.36 |
| | | exons | 2.71 | 1.42 | 3.01 | 0.35 | 0.14 | 0.13 |

**Table 3** Microsatellite distribution as percent motif abundance (%) among 14 genomes *(Continued)*

| Species | Abbreviation | Genomic region | Mono- (1-nt) | Di- (2-nt) | Tri- (3-nt) | Tetra- (4-nt) | Penta- (5-nt) | Hexa- (6-nt) |
|---|---|---|---|---|---|---|---|---|
| *C. carpio* | Ccar | all | 37.18 | 36.12 | 10.7 | 12.51 | 3.15 | 0.34 |
| | | intergenic regions | 0.54 | 0.6 | 0.17 | 0.19 | 0.05 | 0.01 |
| | | introns | 36.2 | 35.16 | 10.2 | 12.22 | 3.08 | 0.33 |
| | | exons | 0.44 | 0.37 | 0.32 | 0.11 | 0.02 | 0.01 |
| *P. reticulata* | Pret | all | 46.3 | 29.54 | 9.69 | 12.69 | 1.48 | 0.31 |
| | | intergenic regions | 2.25 | 2.01 | 0.59 | 0.1 | 0.12 | 0.03 |
| | | introns | 40.47 | 26.13 | 7.47 | 11.34 | 1.27 | 0.24 |
| | | exons | 3.58 | 1.39 | 1.63 | 0.36 | 0.09 | 0.04 |
| *O.kisutch* | Okis | all | 12.85 | 67.38 | 4.61 | 12.15 | 2.31 | 0.7 |
| | | intergenic regions | 0.19 | 1.16 | 0.13 | 0.22 | 0.06 | 0.02 |
| | | introns | 12.29 | 65.73 | 4.17 | 11.84 | 2.23 | 0.68 |
| | | exons | 0.37 | 0.49 | 0.32 | 0.09 | 0.01 | 0.00 |

(A + T)-rich (Table 9). Except for in Olat, Onil, Ccar, Okis and Okis, the CpG-containing repeats were common in the top 50% of hexanucleotide repeats in intronic and exonic regions, and half of species had CpG-containing in the top 50% of hexanucleotide repeats in intergenic regions (Tables 7, 8, 9). A few telomere-like repeats were found in introns or intergenic regions, excluding Pret. However, the (AATCCC) n and (AACC CT) n tracts were observed in exonic regions of Trub and Omyk, respectively (Table 8).

### Iteration number and length distribution of microsatellites in fish genomes

Iteration number and length of microsatellites are both important factors determining microsatellite mutation rates, and it could be extremely important not only for genomic stability, but also with regard to the evolution of additional genomic features such as codon usage. To assess expandability of the repeats, iteration number of microsatellites was plotted against microsatellite length of various quantity intervals: <20, 20–50, 50–100, 100–

**Table 4** Total numbers of Mono-, Di-, and Trinucleotide repeats in exons among 14 fish genomes

| Repeated unit | *O. latipes* | *D. rerio* | *O. niloticus* | *T. rubripes* | *E. lucius* | *L. oculatus* | *O. mykiss* | *C. carpio* | *I. punctatus* | *C. semilaevis* | *A. mexicanus* | *N. furzeri* | *P. reticulata* | *O. kisutch* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1938 | 3830 | 6968 | 2079 | 2069 | 1652 | 2766 | 2178 | 10,981 | 4478 | 5149 | 1376 | 9181 | 1453 |
| C | 177 | 105 | 397 | 643 | 758 | 359 | 686 | 57 | 364 | 232 | 207 | 471 | 242 | 1341 |
| AC | 746 | 4274 | 3272 | 1429 | 2858 | 493 | 4435 | 1179 | 8327 | 2870 | 3180 | 3345 | 2238 | 2205 |
| AG | 97 | 1021 | 1121 | 272 | 679 | 459 | 2236 | 457 | 3402 | 1273 | 3234 | 412 | 877 | 991 |
| AT | 196 | 787 | 683 | 174 | 382 | 156 | 1020 | 203 | 1010 | 314 | 824 | 237 | 525 | 535 |
| CG | 24 | 10 | 9 | 1 | 5 | 4 | 5 | 3 | 68 | 3 | 14 | 39 | 12 | – |
| AAC | 145 | 332 | 435 | 86 | 172 | 182 | 365 | 171 | 850 | 437 | 150 | 136 | 252 | 206 |
| AAG | 235 | 346 | 542 | 154 | 151 | 108 | 434 | 153 | 680 | 587 | 385 | 178 | 641 | 274 |
| AAT | 170 | 775 | 665 | 118 | 179 | 156 | 350 | 309 | 2124 | 338 | 729 | 314 | 544 | 191 |
| ACC | 152 | 142 | 249 | 126 | 186 | 114 | 377 | 69 | 199 | 296 | 236 | 145 | 185 | 262 |
| ACG | 382 | 289 | 591 | 274 | 102 | 291 | 247 | 209 | 343 | 1159 | 528 | 417 | 704 | 184 |
| ACT | 14 | 111 | 133 | 30 | 43 | 20 | 62 | 46 | 123 | 148 | 84 | 265 | 66 | 51 |
| AGC | 322 | 172 | 239 | 203 | 102 | 255 | 202 | 113 | 199 | 519 | 614 | 55 | 474 | 124 |
| AGG | 1031 | 574 | 1141 | 591 | 576 | 777 | 1245 | 220 | 853 | 1731 | 856 | 46 | 923 | 782 |
| ATC | 199 | 660 | 353 | 121 | 235 | 93 | 495 | 321 | 726 | 397 | 401 | 752 | 306 | 304 |
| CCG | 120 | 32 | 51 | 165 | 48 | 247 | 27 | 19 | 64 | 129 | 84 | 14 | 194 | 11 |

**Table 5** Total numbers of Mono-, Di-, and Trinucleotide repeats in introns among 14 fish genomes

| Repeated unit | *O. latipes* | *D. rerio* | *O. niloticus* | *T. rubripes* | *E. lucius* | *L. oculatus* | *O. mykiss* | *C. carpio* | *I. punctatus* | *C. semilaevis* | *A. mexicanus* | *N. furzeri* | *P. reticulata* | *O. kisutch* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 54,469 | 168,546 | 65,095 | 10,417 | 20,201 | 13,989 | 41,783 | 176,323 | 142,517 | 74,854 | 143,668 | 17,966 | 102,376 | 34,546 |
| C | 11,110 | 8113 | 7349 | 12,124 | 19,575 | 3979 | 44,066 | 5666 | 14,260 | 12,080 | 13,067 | 13,704 | 4070 | 58,278 |
| AC | 14,896 | 151,933 | 72,301 | 36,853 | 107,530 | 6473 | 219,574 | 102,632 | 180,297 | 71,964 | 142,073 | 99,236 | 48,930 | 25,072 |
| AG | 1311 | 16,700 | 12,758 | 4675 | 19,458 | 3157 | 116,077 | 26,618 | 78,058 | 20,922 | 112,135 | 11,308 | 9216 | 195,437 |
| AT | 4937 | 129,210 | 20,624 | 1546 | 11,767 | 2307 | 38,935 | 47,442 | 64,634 | 25,056 | 160,352 | 8320 | 10,458 | 35,958 |
| CG | 55 | 255 | 135 | 100 | 145 | 33 | 162 | 69 | 392 | 63 | 58 | 322 | 134 | 113 |
| AAC | 896 | 9018 | 4035 | 543 | 796 | 1633 | 3550 | 5763 | 10,973 | 4836 | 1646 | 146 | 1911 | 2654 |
| AAG | 762 | 2806 | 2699 | 350 | 173 | 393 | 1874 | 2606 | 5733 | 4440 | 4239 | 3781 | 2253 | 1683 |
| AAT | 4610 | 88,156 | 11,327 | 1362 | 3558 | 11,224 | 10,690 | 37,268 | 45,617 | 10,492 | 31,950 | 8898 | 9074 | 8967 |
| ACC | 539 | 677 | 941 | 448 | 921 | 764 | 1297 | 293 | 1217 | 783 | 1134 | 964 | 394 | 1954 |
| ACG | 706 | 508 | 1153 | 692 | 96 | 264 | 681 | 589 | 428 | 5941 | 2032 | 1125 | 1183 | 940 |
| ACT | 115 | 3189 | 858 | 354 | 652 | 226 | 2025 | 1107 | 3542 | 1476 | 2492 | 4481 | 388 | 2846 |
| AGC | 517 | 296 | 511 | 749 | 69 | 223 | 715 | 288 | 242 | 2754 | 2028 | 1291 | 926 | 542 |
| AGG | 1671 | 1427 | 2156 | 1359 | 679 | 694 | 3286 | 630 | 1678 | 6136 | 2180 | 78 | 1747 | 4327 |
| ATC | 1005 | 8440 | 2186 | 929 | 2092 | 975 | 7870 | 2708 | 10,070 | 3209 | 5727 | 1278 | 1605 | 7564 |
| CCG | 151 | 64 | 77 | 337 | 38 | 113 | 17 | 46 | 61 | 303 | 121 | 35 | 162 | 17 |

**Table 6** Total numbers of Mono-, Di-, and Trinucleotide repeats in intergenic regions among 14 fish genomes

| Repeated unit | *O. latipes* | *D. rerio* | *O. niloticus* | *T. rubripes* | *E. lucius* | *L. oculatus* | *O. mykiss* | *C. carpio* | *I. punctatus* | *C. semilaevis* | *A. mexicanus* | *N. furzeri* | *P. reticulata* | *O. kisutch* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 7053 | 19,888 | 4667 | 4118 | 1523 | 3883 | 846 | 2603 | 7815 | 4738 | 7386 | 8467 | 5671 | 515 |
| C | 2058 | 1337 | 827 | 5367 | 1013 | 1170 | 740 | 108 | 919 | 921 | 867 | 7139 | 237 | 907 |
| AC | 2238 | 32,637 | 7840 | 15,342 | 5880 | 2113 | 3698 | 1873 | 13,324 | 2766 | 8393 | 48,648 | 3896 | 4585 |
| AG | 202 | 4756 | 1780 | 2175 | 1082 | 962 | 3056 | 445 | 6202 | 1023 | 7170 | 5467 | 705 | 3532 |
| AT | 741 | 19,891 | 2563 | 780 | 1020 | 683 | 979 | 693 | 3277 | 1346 | 9921 | 4626 | 684 | 606 |
| CG | 6 | 45 | 9 | 61 | 6 | 13 | 4 | – | 31 | 7 | 4 | 169 | 13 | 10 |
| AAC | 109 | 1260 | 438 | 255 | 49 | 537 | 190 | 79 | 866 | 281 | 92 | 797 | 133 | 64 |
| AAG | 117 | 476 | 296 | 167 | 19 | 146 | 70 | 51 | 500 | 376 | 291 | 2127 | 193 | 74 |
| AAT | 782 | 13,857 | 1526 | 673 | 369 | 4014 | 368 | 597 | 3336 | 621 | 2247 | 5103 | 699 | 170 |
| ACC | 74 | 95 | 65 | 250 | 58 | 230 | 30 | 5 | 87 | 30 | 104 | 566 | 25 | 28 |
| ACG | 160 | 103 | 131 | 421 | 6 | 68 | 5 | 16 | 44 | 373 | 108 | 576 | 90 | 68 |
| ACT | 22 | 214 | 52 | 290 | 48 | 86 | 197 | 26 | 298 | 84 | 269 | 837 | 25 | 234 |
| AGC | 167 | 57 | 74 | 335 | 2 | 42 | 12 | 5 | 16 | 173 | 144 | 39 | 65 | 8 |
| AGG | 349 | 364 | 345 | 653 | 70 | 224 | 187 | 11 | 183 | 340 | 175 | 734 | 143 | 197 |
| ATC | 180 | 214 | 367 | 467 | 143 | 326 | 422 | 56 | 782 | 162 | 715 | 2384 | 155 | 122 |
| CCG | 30 | 14 | 13 | 169 | – | 34 | – | – | 8 | 30 | 12 | 28 | 19 | – |

**Table 7** The most frequent Tetra-, Penta-, and Hexanucleotide repeats in introns[a]

| Length of repeated unit | Taxa | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *O. latipes* | *D. rerio* | *O. niloticus* | *T. rubripes* | *E. lucius* | *L. oculatus* | *O. mykiss* | *C. carpio* | *I. punctatus* | *C. semilaevis* | *A. mexicanus* | *N. furzeri* | *P. reticulata* | *O. kisutch* |
| 4-nt | AAAT(4298) | AAAT(46819) | AAAT(7252) | ATCC(2657) | ACAG(4505) | AAAC(394) | ACAG(12845) | AAAT(26188) | AAAT(36182) | ATCC(7288) | ATAG(25367) | ATCC(7081) | ATCC(7303) | ACAG(30164) |
| | ATCC(4197) | ATAG(27227) | AAAC(4710) | ACAG(1263) | AGGG(2259) | ATCC(338) | AAAT(9010) | ATAG(11853) | AAAC(11498) | ATAG(4977) | AAAT(13454) | AAAT(5962) | AAAT(6468) | AGGG(9858) |
| | ATAG(1888) | ATCC(23910) | ATCC(4352) | ATAG(812) | AATC(2201) | AAAT(282) | AGGG(6350) | | ATCC(10151) | AAAT(3814) | | ACGC(4103) | ATGG(1988) | AAAT(9303 |
| | | AATG(22059) | | AAAT(747) | ACGC(1493) | AATC(253) | ACGC(5380) | | | | | ATAG(3514) | | |
| | | | | | | ACAG(228) | | | | | | | | |
| 5-nt | AATTC(892) | AATA T(12313) | AATA G(2669) | AGAG G(1286) | ATACC(157) | AATTC(168) | AAATC(5447) | AAAA T(3745) | AAAA C(1282) | AGAGG(592) | AATGT(2997) | AATTC(1286) | AAAAT(580) | AAATC(7042) |
| | AAAT C(552) | AAAAT(4838) | AATTC(2375) | | AGAGG(59) | AAAAC(104) | AGAG G(1922) | AATAT(3347) | AATTC(1214) | AATCT(465) | AATAT(1814) | ATACC(775) | AAAAC(457) | AGAGG(4021) |
| | AATAG(494) | | AAAAT(868) | | ATACG(49) | AATAC(80) | | AAATT(1092) | AATAT(1079) | AAAAT(383) | AAAAT(1119) | AATAG(684) | AATAG(326) | |
| | AAAAT(273) | | | | AACGG(44) | AATAT(70) | | | AAAAT(965) | AATAT(346) | AAGAG(918) | AAAAT(327) | AATAT(244) | |
| | | | | | ACCCC(36) | AAATC(66) | | | | AAAAC(326) | AATAG(535) | | AAAAG(158) | |
| | | | | | AGGGG(35) | AAAAT(61) | | | | ACTGG(270) | ATATC(496) | | | |
| | | | | | AAAAC(34) | ATGAC(60) | | | | ATGTC(250) | | | | |
| | | | | | ATCGC(29) | AAATT(41) | | | | | | | | |
| | | | | | AACTG(28) | AAGCG(35) | | | | | | | | |
| | | | | | AAGGC(28) | ACTGG(34) | | | | | | | | |
| | | | | | | CCCGG(34) | | | | | | | | |
| 6-nt | AATC AG(75) | AAATGT(244) | AATC AG(107) | AACCCT(67) | ACAC GC(94) | AACT TG(150) | ACAG AG(730) | AAAA AT(282) | ATACTG(99) | ACCA GG(156) | AATCCC(216) | ACAC GC(236) | AACC AG(236) | ACAG AG(1106) |
| | ATAT AG(28) | ACAC GC(234) | ATATAC(87) | AGAGGG(42) | AACCCT(78) | AACA GG(51) | ACAC GG(468) | AAAT GT(187) | AAATGT(96) | ACAC AG(134) | AAACTG(93) | AATC AG(102) | ACACGC(53) | AGAGGG(472) |
| | AAAT TC(18) | AAATAC(211) | ATATAG(52) | AATCCC(37) | ACAC AG(45) | ACACTC(39) | AGAG GG(367) | AATC AG(174) | ATATAC(94) | AACC AG(101) | AAATGT(84) | AAGACG(71) | AAAAAT(26) | ACACGC(383) |
| | AACC CT(16) | ATATAC(113) | AAAAAC(43) | ACAGAG(32) | AATCCC(41) | ACACGC(34) | AACCCT(321) | AAAG TG(124) | ATAGTC(65) | ATATAC(94) | AAAAAT(61) | AAATAC(52) | | ATACAC(279) |
| | AATC CC(12) | AAAAAT(89) | AAAAAT(30) | ACCCCC(24) | ATATAC(40) | AATAAG(31) | ACAC AG(272) | AATCCC(92) | ATATAG(62) | AAGAGG(90) | AAATAT(53) | AAACAT(50) | | AAAGAG(250) |
| | AAAAAT(10) | | AATAAG(26) | AATCAG(21) | AGCG GG(39) | AATCTG(30) | | | AAATAT(60) | ACCTGG(77) | ATATAC(53) | ATACAG(34) | | ACACAG(223) |
| | ATAGAG(10) | | AAATTC(23) | ACACGC(21) | | AAATAT(29) | | | AACCCT(57) | ACAGAG(69) | AATCGT(52) | AATCCC(32) | | |
| | AAATAC(9) | | | AATCTG(20) | | ACACAG(28) | | | AAAAAT(52) | AAATAT(68) | AATGCT(51) | | | |
| | AAAAAG(8) | | | AATTCT(18) | | AATCCC(25) | | | AAAAAC(43) | AGAGGG(54) | ATATAG(45) | | | |
| | AAAATT(7) | | | ACTGGG(18) | | AATAAC(21) | | | ATACAC(33) | AACCCT(49) | AAAGTC(44) | | | |
| | | | | AAGAGG(17) | | AACG GG(20) | | | AATCCC(32) | ACTCCC(48) | ACACTC(40) | | | |
| | | | | ACACTC(17) | | AACCCT(19) | | | AATAGT(27) | ACCTCC(45) | AAAGAG(38) | | | |

**Table 7** The most frequent Tetra-, Penta-, and Hexanucleotide repeats in introns[a] *(Continued)*

| Length of repeated unit | Taxa | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *O. latipes* | *D. rerio* | *O. niloticus* | *T. rubripes* | *E. lucius* | *L. oculatus* | *O. mykiss* | *C. carpio* | *I. punctatus* | *C. semilaevis* | *A. mexicanus* | *N. furzeri* | *P. reticulata* | *O. kisutch* |
| | | | | ACACAG(16) | | | | | | AAAAAT(43) | ATGTCC(37) | | | |
| | | | | | | | | | | AATCCC(43) | AAATCC(36) | | | |
| | | | | | | | | | | ACTGGG(43) | AACCCT(34) | | | |
| | | | | | | | | | | ATCAGC(39) | | | | |
| | | | | | | | | | | AAGATG(38) | | | | |
| | | | | | | | | | | AGTCGG(37) | | | | |
| | | | | | | | | | | ATGACC(37) | | | | |

[a]Only the repeat motifs that together comprise 50% of all repeats of the particular unit length are shown here
[b]Hexanucleotide repeats for which the number of the tandem repeat is < 2 are not shown

**Table 8** The most frequent Tetra-, Penta-, and Hexanucleotide repeats in exons[a]

| Length of repeated unit | Taxa | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *O. latipes* | *D. rerio* | *O. niloticus* | *T. rubripes* | *E. lucius* | *L. oculatus* | *O. mykiss* | *C. carpio* | *I. punctatus* | *C. semilaevis* | *A. mexicanus* | *N. furzeri* | *P. reticulata* | *O. kisutch* |
| 4-nt | AAAT(123) | AAAT(336) | AAAC(474) | AAAC(57) | ACAG(99) | AAAC(44) | ACAG(286) | AAAT(181) | AAAT(1063) | AAAC(245) | AAAT(264) | AAAT(212) | AAAT(231) | ACAG(117) |
| | AAAC(91) | AAAC(229) | AAAT(339) | AAAT(57) | ACGC(68) | AGGG(32) | AAAT(119) | AAAC(121) | AAAC(928) | AAAT(124) | AAAC(232) | AAAC(109) | AAAC(166) | AAAT(61) |
| | ATCC(65) | AATG(109) | | ACGC(42) | AAAT(53) | AAAT(17) | ATAC(106) | | | AAAG(76) | AAAG(104) | ACGC(109) | AAAG(89) | ATAC(61) |
| | AAAG(39) | | | ACAG(39) | ATCG(41) | AGCG(16) | AGGG(97) | | | ACAG(66) | ACAG(95) | | | ATCG(53) |
| | | | | AAAG(32) | ATAC(39) | ACGG(14) | ATCG(79) | | | | | | | ACTG(41) |
| | | | | | | AAAG(13) | ACGC(76) | | | | | | | |
| 5-nt | AAAAT(18) | AATAT(19) | AAAAC(79) | AAAAT(6) | ACTGC(2) | AAAAC(13) | AAATC(26) | AAAAC(13) | AAAA C(145) | AAAAC(30) | AAAAT(35) | AAAAT(26) | AAAAC(53) | AAATC(17) |
| | AATAG(8) | AAAAC(17) | AAAAT(49) | AAAAC(5) | AGCCC(2) | ACCGG(11) | AGAGG(25) | AAAAT(11) | AAAAT(67) | AGAGG(28) | AAAAC(25) | AATAG(13) | AAAAT(37) | AAAAC(6) |
| | AATTC(7) | AAAAT(13) | AAAAG(43) | AGAGG(5) | AGGGG(2) | CCCGG(9) | AATAC(7) | AATAT(7) | AAAAG(44) | AAAAG(24) | AATGT(20) | AATTC(11) | ACCGG(29) | AGAGG(6) |
| | AAAAG(5) | AAATT(7) | | AAAAG(4) | ATCTC(2) | AGGCG(8) | AAAAC(5) | AACTG(4) | AATAT(20) | ACTGG(14) | AATAT(12) | AGAGG(8) | AAAAG(16) | AAAAT(4) |
| | AAACT(5) | AATCG(7) | | CCCGG(4) | CCCCG(2) | AGAGG(5) | ACCCC(5) | AATAG(4) | | AAAAT(13) | AAAAG(11) | AATGT(7) | | AACAG(4) |
| | | | | ACAGG(3) | | AAAAT(4) | | AGAGG(4) | | | AAGCT(11) | | | |
| | | | | AAATT(2) | | ACCCG(4) | | | | | AATCG(10) | | | |
| | | | | | | | | | | | AAGAG(9) | | | |
| 6-nt | ACTC GG(3) | ACAC GC(6) | AAAAAC(6) | ATACCG(8) | ACACCG(4) | AAGA GG(5) | AGAG GG(12) | AGCC GG(4) | AAAAAT(8) | ACCCAG(36) | ACGAGG(4) | ACACGC(5) | AACC AG(24) | AGAG GG(5) |
| | AACC AG(2) | ACACTC(3) | AAAAAT(3) | ACCA GG(3) | ACACGC(4) | ACTCCC(5) | ACAGAG(7) | AAGA GG(2) | AAAAAC(6) | AACCAG(12) | AAAAAT(3) | AATCAG(3) | ACCTGG(6) | AAAACC(3) |
| | AAGA TG(2) | AAGA TG(2) | AACCAG(3) | ACGA GG(3) | AACCAG(2) | AGCC GG(5) | AAGAGG(5) | ACAG GG(2) | AAAAAG(6) | ACCTGG(12) | AAGAGG(3) | AAAAAC(2) | AAGAGG(5) | AACCAC(2) |
| | ATGTCC(2) | ACCTCC(2) | AAAAAG(2) | ACGG GG(3) | ACCG AG(2) | ACGA GG(4) | ACACGC(5) | ACCCCG(2) | AAGAGG(4) | ATCAGC(12) | AATCGT(2) | AACCAG(2) | ACCTCC(4) | AACCCT(2) |
| | etc.[b] | AGCC CG(2) | AAGAGG(2) | ACTCGG(3) | ACCTCC(2) | AGAG GG(4) | ACTCCC(4) | ACGC AG(2) | ATATAC(4) | ACCTCC(9) | ACACTC(2) | AAGA GG(2) | ATCTGG(4) | AAGA GG(2) |
| | | etc.[b] | ACCCGG(2) | AGCC GG(3) | AGGC GG(2) | ACCCGG(3) | AGGGGG(3) | ATCCTC(2) | ACACCC(3) | AAGAGG(8) | ACCAGG(2) | AGCTCC(2) | ACGCCG(3) | AGCTCC(2) |
| | | | ACTCCG(2) | ATCCTC(3) | | ACCG GG(3) | AAAAAT(2) | | ACCCCC(3) | AGCAGG(8) | ACCTCC(2) | | ACTCCG(3) | |
| | | | AGCTCC(2) | AACACC(2) | | ACGG GG(3) | AAAGAG(2) | | ACGAGG(3) | ACAGAG(7) | ACTCTC(2) | | AGCGGC(3) | |
| | | | ATGCCC(2) | AAGG AG(2) | | AGTCGG(3) | AACACC(2) | | AGGGGG(3) | ACCACG(5) | AGTCCG(2) | | AATACT(2) | |
| | | | | AATCCC(2) | | ATGCCC(3) | AACCCT(2) | | ATACAC(3) | ACGAGG(5) | | | | |

**Table 8** The most frequent Tetra-, Penta-, and Hexanucleotide repeats in exons[a] *(Continued)*

| Length of repeated unit | Taxa | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *O. latipes* | *D. rerio* | *O. niloticus* | *T. rubripes* | *E. lucius* | *L. oculatus* | *O. mykiss* | *C. carpio* | *I. punctatus* | *C. semilaevis* | *A. mexicanus* | *N. furzeri* | *P. reticulata* | *O. kisutch* |
| | | | | AATCTG(2) | | AACAGC(2) | ACCCAG(2) | | ATCCTC(3) | ACTGGG(5) | | | | |
| | | | | ACAG AG(2) | | ACACTC(2) | ACCGCC(2) | | | ATCCTC(5) | | | | |
| | | | | | | ACCCCC(2) | | | | AAAAAC(5) | | | | |
| | | | | | | ACCGGC(2) | | | | | | | | |
| | | | | | | AGCCCC(2) | | | | | | | | |

[a]Only the repeat motifs that together comprise 50% of all repeats of the particular unit length are shown here
[b]Hexanucleotide repeats for which the number of the tandem repeat is < 2 are not shown

**Table 9** The most frequent Tetra-, Penta-, and Hexanucleotide repeats in intergenic regions[a]

| Length of repeated unit | Taxa | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *O. latipes* | *D. rerio* | *O. niloticus* | *T. rubripes* | *E. lucius* | *L. oculatus* | *O. mykiss* | *C. carpio* | *I. punctatus* | *C. semilaevis* | *A. mexicanus* | *N. furzeri* | *P. reticulata* | *O. kisutch* |
| 4-nt | ATCC(889) | AAAT(7753) | AAAT(505) | ATCC(1513) | AATC(327) | AAAC(153) | ACAG(453) | AAAT(418) | AAAT(2306) | ATAG(657) | ATAG(2538) | ATCC(4513) | ATCC(787) | ACAG(600) |
| | AAAT(597) | ATAG(4886) | ATCC(468) | ATAG(527) | ACAG(295) | AAAT(112) | AAAT(246) | ATAG(147) | AAAC(794) | ATCC(525) | AAAT(778) | AAAT(3225) | AAAT(477) | AGGG(237) |
| | ATAG(347) | AATG(4285) | AAAC(431) | ACAG(502) | ATCC(169) | ATCC(108) | AGGG(167) | | ATAG(790) | AAAT(199) | | ATAG(2429) | | |
| | | ATCC(4223) | | AAAT(380) | ATAC(111) | AATC(106) | ACTG(134) | | | | | | | |
| | | | | | | ACGC(51) | | | | | | | | |
| 5-nt | AATT C(203) | AATA T(1421) | AATTC(466) | AGAG G(865) | ATACC(27) | AATTC(83) | AAAT C(209) | AAAA T(61) | AATTC(81) | ACTGG(47) | AATGT(176) | AATC C(866) | AAAAT(43) | AGAG G(167) |
| | AAAT C(119) | AAAAT(601) | AATAG(384) | | ATACG(15) | AATAC(41) | AGAGG(25) | AATAT(37) | AAAAC(75) | AATAT(31) | AATAT(98) | ATAC C(540) | AATAG(30) | AAAT C(155) |
| | AATAG(86) | AAATT(405) | | | AACAC(8) | AAAAC(39) | | AATTC(19) | AATAT(71) | AAAAT(27) | ATATG(96) | AATA G(533) | AAAAC(22) | |
| | | | | | AAGGC(7) | AATAT(29) | | AATA G(16) | AAAAT(63) | AACTG(26) | AAAAT(74) | | AAAAG(21) | |
| | | | | | AATGC(7) | AAATC(25) | | | ATCAC(41) | ATGTC(25) | AATAG(73) | | AATAT(20) | |
| | | | | | AACGG(6) | AAAAT(24) | | | AAAAG(35) | AATCT(21) | AAGAG(62) | | ACCGG(15) | |
| | | | | | | AATAG(17) | | | | | AAACC(60) | | AGAGG(11) | |
| | | | | | | ACTGG(16) | | | | | | | | |
| | | | | | | ATGAC(16) | | | | | | | | |
| 6-nt | AATC AG(12) | AAAT GT(33) | AATC CC(113) | AACC CT(51) | AACC CT(16) | AACT TG(56) | AACC CT(74) | AAAG TG(6) | AACC CT(43) | ACCAGG(18) | AACC CT(136) | ACAC GC(95) | AACC AG(40) | ACAG AG(31) |
| | ACCTGG(4) | ACAC GC(33) | | AATC CC(43) | AATC CC(10) | AACA GG(33) | ACAG AG(37) | AAAA AT(5) | AATC CC(21) | AACCAG(15) | AATCCC(47) | AATC AG(74) | ACACGC(9) | AACC CT(18) |
| | AAACTG(3) | AAAT AC(27) | | AATCTG(16) | ACAC GC(7) | AAAT AT(17) | AATC CC(31) | AATC CC(5) | ATCCAC(8) | ACCTGG(14) | | AAGA CG(43) | | ACAC AG(18) |
| | AACCCT(3) | ATATAC(13) | | AATC AG(15) | ACAC AG(5) | ACAC TG(17) | | | AAATGT(6) | AAGTCG(12) | | AAAT AC(41) | | ACACGG(9) |
| | AAGA GG(3) | AATC CC(12) | | ACCC CC(15) | ACTCTC(5) | AATC TG(13) | | | ATATAC(6) | ATCAGC(12) | | ATAC AG(30) | | |
| | AATCCC(3) | AAAA AT(11) | | ACAG AG(14) | | ACAC GC(10) | | | | ACACAG(11) | | | | |
| | ATCTGG(3) | | | AGAG GG(14) | | AATAAG(9) | | | | ATGACC(11) | | | | |
| | AAATTG(2) | | | AAGA GG(10) | | ACACTC(9) | | | | AAAAAT(7) | | | | |
| | AACCAG(2) | | | AATT | | | | | | ACACTG(7) | | | | |

**Table 9** The most frequent Tetra-, Penta-, and Hexanucleotide repeats in intergenic regions[a] *(Continued)*

| Length of repeated unit | Taxa | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *O. latipes* | *D. rerio* | *O. niloticus* | *T. rubripes* | *E. lucius* | *L. oculatus* | *O. mykiss* | *C. carpio* | *I. punctatus* | *C. semilaevis* | *A. mexicanus* | *N. furzeri* | *P. reticulata* | *O. kisutch* |
| | | | | AG(10) | | | | | | | | | | |
| | | | | | | | | | | AAGATG(6) | | | | |
| | | | | | | | | | | ACTGGG(6) | | | | |

[a]Only the repeat motifs that together comprise 50% of all repeats of the particular unit length are shown here
[b]Hexanucleotide repeats for which the number of the tandem repeat is < 2 are not shown

200, 200–300, and >300 (Fig. 1). The details of all iteration numbers and densities of microsatellites in fish genomes are given in the Supplement Table 2. Usually, the frequency of microsatellites has a tendency to converge to a small iteration number. In other words, short microsatellites were observed more frequently in the fish genomes than long microsatellites. When the iteration number was less than 20, the repeat tracts varying motif lengths from mono- to hexa-nucleotide (1–6 nt) comprised more than 83.93, 67.22, 90.38, 88.93, 92.58 and 90.42%, respectively (Fig. 1 and Supplement Table 2). However, a few special microsatellites were found where the iteration number exceeded 300, for example 1-nt microsatellites in Csem, Eluc, Amex, Ccar and Okis, 2-nt microsatellites in Drer, Csem, Eluc, Nfur, Amex, Ccar and Okis, 3-nt microsatellites in Nfur, Amex, Ccar, 4-nt microsatellites in Drer, Csem, Nfur, Amex and Ccar, 5-nt microsatellites in Amex and Ccar, and 6-nt microsatellites in Csem, Amex and Ccar (Supplement Table 2).

## Discussion

In this study, we examined the microsatellites composed of motifs 1–6 bp long in the entire genomes of 14 fish species and analyzed their distribution and frequency in different genomic regions. Microsatellite occurrence significantly differed with the coverage degree varying from 0.18 to 5.29%. Comparison of microsatellite repeat occurrence in the genomes of humans (3%) [6], primates (0.83–0.88%) [72–74], birds (0.13–0.49%) [88], plants and fungi (0.04–0.15%) [75, 76, 80, 89, 90], with our data indicates that microsatellite occurrence differs between different species and this might be a general phenomenon across taxa [33]. In fact, differences might even occur between closely related species as humans and chimpanzees [91], and within the genus of *Drosophila* [92, 93].

Another clear trend to emerge from this analysis was that the observed dependence of microsatellite abundance on repeated unit length and iteration number was very much biased from the expected trend of gradual decrease, which was consistent with a previous study [36]. Our research also indicated that microsatellite density is not strictly positively correlated with genome size. Although it was well known that the microsatellite density generally correlates positively with genome size [26, 36, 94], our contradictory results have been found in other studies [72, 83, 88, 95]. Overall, the comparative analysis of microsatellites indicated that there was great variation of microsatellite content across the 14 fish species. This might be indicative that differential selective constraints may play an important role in microsatellite evolution and result in

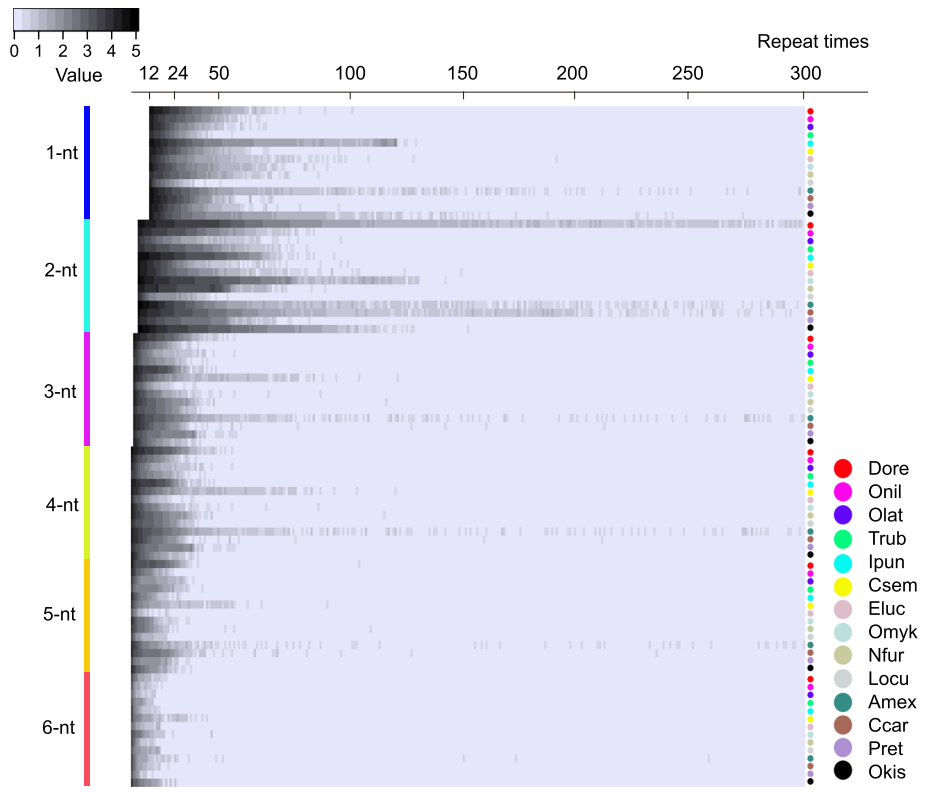


**Fig. 1** Heat map of the microsatellite distribution frequency of different motif length (1–6 nt) based on the iteration number among 14 fish genomes

the accumulated preference for different microsatellite types (Saeed2016&Ellegren2004& Schlötterer2000).

During genome evolution, microsatellite repeats mutation may provide a molecular mechanism for faster adaptation to environmental stress by increasing the quantities of DNA and providing the raw materials for adaptive evolution of organisms. Generally, microsatellite instability of dinucleotide repeats is higher than trinucleotide, tetranucleotide and pentanucleotide repeats [96]. In other words, the mutation rate of microsatellite dependence on repeated unit length is biased from the trend of gradual decrease. This could explain the high numbers of mono−/di-nucleotide motif microsatellites and the low numbers of penta−/hexa-nucleotide motif microsatellites in the genomes. We should note that the frequency of tetranucleotide repeats was more than trinucleotide repeats in most of the 14 genomes. However, there was a trend that trinucleotide repeats were more frequent than tetranucleotide repeats in exonic regions, and less than tetranucleotide repeats in intronic and intergenic regions of most genomes. We suggest that the lower number of trinucleotide repeats cannot only be explained by conservation since they attribute triplet codes to form parts of genes. However, there may be a mechanism (e.g. mismatch repair system) in the exonic regions to maintain the higher number of trinucleotide repeats.

As is evident from Tables 2, 3, 4, 5, 6, 7, 8 and 9, poly(A/T) tracts were more common than poly(C/G) tracts in these genomes. Poly(A/T) tracts were particularly common in exonic and intergenic regions, but this was opposite in intronic regions of some taxa (e.g., Trub, Omyk and Okis) and this has also been observed in the human genome [6]. The higher frequency of poly(A) tracts can be attributed to the re-integration of processed genes into the genome from mRNA with an attached poly(A) tail, while poly(C/G) are not part of this integrative mechanism. An alternative explanation is that a long A-rich tail is known to be necessary for the universal retrotransposon in eukaryotic genomes, such as *Alu*, LINE-1 and L1 retrotransposons [97–99]. Meanwhile, the formation of pseudogenes may attribute to this higher proportion of (A + T)-rich repeats [36, 100]. However, the mutation mechanism of microsatellite DNA provides a basis for this phenomenon. The variable frequencies of poly(A) and poly(C) could be due to the difference in stability between (GC) n and (AT) n repeats. (GC) n repeats are more stable than (AT) n repeats and hence it would be more difficult for the poly(C) sequences to slip during replication during the evolution of microsatellite DNA [6, 95, 101]. In the intronic regions, the higher than expected frequencies of poly(C/G) tracts in some species may be due to duplication events of key DNA sequences during evolution or the integrity of chromosomes may depend on a higher order DNA sequence organization that includes the presence of poly(C/G) tracts [102].

In the case of dimeric repeats, we found (AC) n tract was common and the (GC) n tract was rare. Assuming that, on the microsatellite DNA stability, (GC)-rich regions are relatively stable, there is less replication slippage generating the repeated motifs of microsatellites [103]. On a genomic scale, microsatellite sequences are presumably at equilibrium, where (AC) n or (AG) n repeats should be more abundant than (AT) n or n repeats. However, we found the opposite distribution of microsatellite motifs in the genome of Amex. We suggest that there is interspecific variation in the mechanisms of mutation or repair of specific motifs [63] or there might be variation in the selective constraints that are associated with different microsatellite motifs [33].

Compared to other microsatellite motifs, the trinucleotide repeat undergoes strict regulation under evolutionary stress. While the (AAT) n tracts were common in intronic and intergenic regions of the fish genomes, (AGG) n tracts were typically more numerous than other repeat types in exons. Therefore, different genome fractions may characterize different microsatellite abundances resulting from the functions of genome evolution and selective constraints [104]. Combined with the above, inconsistent distribution patterns where (ACT) n tracts were numerous in intergenic regions of Okis and (AAT) n tracts were common in exons of Drer and Ipun indicated that the distribution of microsatellites reflected the bias of the base composition in the genomes fractions. Other biases, such as the (CCG) n tracts in Trub and the (ACC) n tracts in Ccar, suggest that selective forces probably play various roles in specific genomes and differ from each other in a species-specific manner [36].

It should be noted that we found extremely rare (CCG) n and (ACG) n repeats in these genomes. A reasonable explanation for this rarity is the presence of the highly mutable CpG dinucleotide within the motif. Rarity of CpG is almost certainly a consequence of the methylation. In vertebrate genomes, a CpG-containing island occurs at about one-fifth of the expected frequency [105, 106] because between 60 and 90% of CpGs are methylated at the 5 position on the cytosine ring and there is a failure of the DNA repair mechanism to recognize deamination of 5-methylcytosine to produce thymine [107, 108]. However, experiments have shown that clusters of non-methylated CpG may attribute to the lack of CpG suppression in the HTF islands, where an approximate 1% DNA fraction accounted for the total genome from a variety of vertebrates [109, 110]. The HTF fraction is extremely rich in cleavable sites for mCpG-sensitive restriction enzymes and sequences chosen at random from the HTF fraction belong to

islands of DNA several hundred base pairs long that contain CpG at more than 10 times its density in bulk DNA. This would help to explain the phenomenon that (ACG) n or (CCG) n tracts were abundant in introns of all fishes, in contrast to the rarity or absence of this motif in intergenic regions. An alternative explanation is that a specific mechanism exists to maintain the observed level of CpG-containing repeats in introns. The role of cytosine methylation in histone deacetylation, chromatin remodeling, and gene silencing may account for this phenomenon [111].

In the tetranucleotide microsatellites, the (AAAB) n tracts (B denotes any base other than A) seem to be more common, followed by 25% G + C content, and then 75% G + C content and 100% G + C content. Previous studies have indicated that DNA sequence composition could have a profound influence on microsatellite incidence [26, 33]. Kristitin et al. (2002) suggested that the G + C content of microsatellites might have influenced the mutation rate because the tetranucleotide repeats with 25% G + C content were not statistically different from each other, but each was significantly different from the repeats with 50% G + C content [112]. Meanwhile, the attribution of selective forces and DNA mismatch repair system for the distribution patterns could not be ignored, because of several exceptions observed in our study, for example (ACAG) n tracts were abundant in Omky and Okis.

The longer microsatellites (5–6 nt) have an advantage of being more polymorphic than the shorter ones (1–4 nt), as mutation rates generally increase with an increase in the number of repeat units [33, 113]. The significant differences in the repeat types and motif length of microsatellites between studied fish species seems to be due to their genome-specific characteristics. In conclusion, though it remains unclear why certain repeat motifs are more common than others, or the reason they vary so much between different fish species, several observations presented here suggest that individual genomes and genome-specific regions may be characterized by unique microsatellite profiles. This was also supported by the reports of taxon-specific repeats or genome-specific region repeats [6, 36]. The study of microsatellites may help us understand numerous aspects of genome organization and functions.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07752-6.

**Additional file 1: Supplement Table 1.** Total numbers of different microsatellite repeats in 14 fish species.

**Additional file 2: Supplement Table 2.** Total numbers of Mono-, Di-, Tri- Tetra-, Penta-, and Hexanucleotide repeats in 14 fish species.

### Acknowledgements

We would like to thank Lianming Du for assistance with microsatellite extraction.

### Authors' contributions

Zhaobin Song proposed the research and directed the study. Yi Lei and Yu Zhou analyzed and wrote the paper. Megan Price wrote and revised the paper. All authors read and approved the final manuscript.

### Funding

This work was supported by the Yalong River Hydropower Development Company, Ltd. (No. YLDC-ZBA-2018116).

### Availability of data and materials

Genome data are available using the links provided by the Ensembl team (*Astyanax mexicanus*: http://ftp.ensembl.org/pub/release-103/fasta/astyanax_mexicanus/dna/; *Danio rerio*: http://ftp.ensembl.org/pub/release-103/fasta/danio_rerio/dna/; *Cynoglossus semilaevis*: http://ftp.ensembl.org/pub/release-103/fasta/cynoglossus_semilaevis/dna/; *Esox Lucius*: http://ftp.ensembl.org/pub/release-103/fasta/esox_lucius/dna/; *Ictalurus punctatus*: http://ftp.ensembl.org/pub/release-103/fasta/ictalurus_punctatus/dna/; *Lepisosteus oculatus*: http://ftp.ensembl.org/pub/release-103/fasta/lepisosteus_oculatus/dna/; *Oryzias latipes*: http://ftp.ensembl.org/pub/release-103/fasta/oryzias_latipes_hni/dna/; *Oreochromis niloticus*: http://ftp.ensembl.org/pub/release-103/fasta/oreochromis_niloticus/dna/; *Poecilia reticulate*: http://ftp.ensembl.org/pub/release-103/fasta/poecilia_reticulata/dna/; *Takifugu rubripes*: http://ftp.ensembl.org/pub/release-103/fasta/takifugu_rubripes/dna/), and NCBI team (*Cyprinus carpio*: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/951/615/GCF_000951615.1_common_carp_genome/; *Nothobranchius furzeri*: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/465/895/GCF_001465895.1_Nfu_20140520/; *Oncorhynchus kisutch*: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/021/735/GCF_002021735.2_Okis_V2/; *Oncorhynchus mykiss*: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/013/265/735/GCF_013265735.2_USDA_OmykA_1.1/).

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no conflict of interest.

Received: 8 March 2021 Accepted: 24 May 2021
Published online: 07 June 2021

### References

1. Dieringer D, Schlötterer C. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res. 2003;13(10):2242–51. https://doi.org/10.1101/gr.1416703.
2. Zavodna M, Bagshaw A, Brauning R, Gemmell NJ. The effects of transcription and recombination on mutational dynamics of short tandem repeats. Nucleic Acids Res. 2018;46(3):1321–30. https://doi.org/10.1093/nar/gkx1253.
3. Fungtammasan A, Ananda G, Hile SE, Su MS, Sun C, Harris R, et al. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. 2015;25(5):736–49. https://doi.org/10.1101/gr.185892.114.
4. Ahmed MM, Shen C, Khan AQ, Wahid MA, Shaban M, Lin Z. A comparative genomics approach revealed evolutionary dynamics of microsatellite imperfection and conservation in genus Gossypium. Hereditas. 2017;154(1):1–12.
5. Hatcher E, Wang C, Lefkowitz E. Genome variability and gene content in chordopoxviruses: dependence on microsatellites. Viruses. 2015;7(4):2126–46. https://doi.org/10.3390/v7042126.
6. Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. 2003;4(2):1–10.

7. Pearson CE, Edamura KN, Cleary JD. Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet. 2005;6(10):729–42. https://doi.org/10.1038/nrg1689.
8. Gelsomino F, Barbolini M, Spallanzani A, Pugliese G, Cascinu S. The evolving role of microsatellite instability in colorectal cancer: a review. Cancer Treat Rev. 2016;51:19–26. https://doi.org/10.1016/j.ctrv.2016.10.005.
9. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018;19(5):286–98. https://doi.org/10.1038/nrg.2017.115.
10. Chistiakov DA, Hellemans B, Volckaert FAM. Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics. Aquaculture. 2006;255(1):1–29. https://doi.org/10.1016/j.aquaculture.2005.11.031.
11. Brouwer JR, Willemsen R, Oostra BA. Microsatellite repeat instability and neurological disease. BioEssays. 2009;31(1):71–83. https://doi.org/10.1002/bies.080122.
12. Gao FB, Richter JD. Microsatellite expansion diseases: repeat toxicity found in translation. Neuron. 2017;93(2):249–51. https://doi.org/10.1016/j.neuron.2017.01.001.
13. Sinden RR. Origins of instability. Nature. 2001;411(6839):757–8. https://doi.org/10.1038/35081234.
14. Dib C, Fauré S, Fizames C, Samson D, Drouot N, Vignal A, et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature. 1996;380(6570):152–4. https://doi.org/10.1038/380152a0.
15. Dietrich WF, Miller JC, Steen RG, Merchant M, Damron D, Nahf R, et al. A genetic map of the mouse with 4,006 simple sequence length polymorphisms. Nat Genet. 1994;7(2):220–45. https://doi.org/10.1038/ng0694supp-220.
16. Kaye C, Milazzo J, Rozenfeld S, Lebrun MH, Tharreau D. The development of simple sequence repeat markers for *Magnaporthe grisea* and their integration into an established genetic linkage map. Fungal Genet Biol. 2003;40(3):207–14. https://doi.org/10.1016/j.fgb.2003.08.001.
17. Ren P, Peng W, You W, Huang Z, Guo Q, Chen N, et al. Genetic mapping and quantitative trait loci analysis of growth-related traits in the small abalone Haliotis diversicolor using restriction-site-associated DNA sequencing. Aquaculture. 2016;454:163–70. https://doi.org/10.1016/j.aquaculture.2015.12.026.
18. Campoy JA, Ruiz D, Egea J, Rees DJG, Celton JM, Martínez-Gómez P. Inheritance of flowering time in apricot (Prunus armeniaca L.) and analysis of linked quantitative trait loci (QTLs) using simple sequence repeat (SSR) markers. Plant Mol Biol Rep. 2011;29(2):404–10. https://doi.org/10.1007/s11105-010-0242-9.
19. Chambers GK, Curtis C, Millar CD, Huynen L, Lambert DM. DNA fingerprinting in zoology: past, present, future. Invest Genet. 2014;5(1):1–11.
20. Rafiei V, Banihashemi Z, Jiménez-Díaz RM, Navas-Cortés JA, Landa BB, Jiménez-Gasco MM, et al. Comparison of genotyping by sequencing and microsatellite markers for unravelling population structure in the clonal fungus *Verticillium dahliae*. Plant Pathol. 2018;67(1):76–86. https://doi.org/10.1111/ppa.12713.
21. Bhargava A, Fuentes FF. Mutational dynamics of microsatellites. Mol Biotechnol. 2010;44(3):250–66. https://doi.org/10.1007/s12033-009-9230-4.
22. Vieira MLC, Santini L, Diniz AL, Munhoz CDF. Microsatellite markers: what they mean and why they are so useful. Genet Mol Biol. 2016;39(3):312–28. https://doi.org/10.1590/1678-4685-GMB-2016-0027.
23. Garner TWJ. Genome size and microsatellites: the effect of nuclear size on amplification potential. Genome. 2002;45(1):212–5. https://doi.org/10.1139/g01-113.
24. Hancock J. Microsatellites and other simple sequences: genomic context and mutational mechanisms. New York: Oxford University Press; 1999.
25. Primmer CR, Raudsepp T, Chowdhary BP, Moller AP, Ellegren H. Low frequency of microsatellites in the avian genome. Genome Res. 1997;7(5):471–82. https://doi.org/10.1101/gr.7.5.471.
26. Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol. 2001;18(7):1161–7. https://doi.org/10.1093/oxfordjournals.molbev.a003903.
27. Karlin S, Brocchieri L, Bergman A, Mrázek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations. Proc Natl Acad Sci U S A. 2002;99(1):333–8. https://doi.org/10.1073/pnas.012608599.
28. Rockman MV, Wray GA. Abundant raw material for *cis*-regulatory evolution in humans. Mol Biol Evol. 2002;19(11):1991–2004. https://doi.org/10.1093/oxfordjournals.molbev.a004023.
29. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol. 2004;21(6):991–1007. https://doi.org/10.1093/molbev/msh073.
30. Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and characterization of microsatellite instability across 18 cancer types. Nat Med. 2016;22(11):1342–50. https://doi.org/10.1038/nm.4191.
31. Ranathunge C, Wheeler GL, Chimahusky ME, Kennedy MM, Morrison JI, Baldwin BS, et al. Transcriptome profiles of sunflower reveal the potential role of microsatellites in gene expression divergence. Mol Ecol. 2018;27(5):1188–99. https://doi.org/10.1111/mec.14522.
32. Orgel LE, Crick FHC. Selfish DNA: the ultimate parasite. Nature. 1980;284(5757):604–7. https://doi.org/10.1038/284604a0.
33. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004;5(6):435–45. https://doi.org/10.1038/nrg1348.
34. Rajendrakumar P, Biswal AK, Balachandran SM, Srinivasarao K, Sundaram RM. Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. Bioinformatics. 2007;23(1):1–4. https://doi.org/10.1093/bioinformatics/btl547.
35. Kim CK, Lee GS, Mo JS, Bae SH, Lee TH. Molecular marker database for efficient use in agricultural breeding programs. Bioinformation. 2015;11(9):444–6. https://doi.org/10.6026/97320630011444.
36. Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 2000;10(7):967–81. https://doi.org/10.1101/gr.10.7.967.
37. Metzgar D, Bytof J, Wills C. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. 2000;10(1):72–80.
38. Wang Z, Weber JL, Zhong G, Tanksley SD. Survey of plant short tandem DNA repeats. Theor Appl Genet. 1994;88(1):1–6. https://doi.org/10.1007/BF00222386.
39. Varshney RK, Thiel T, Stein N, Langridge P, Graner A. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett. 2002;7(2A):537–46.
40. Moran C. Microsatellite repeats in pig (*Sus domestica*) and chicken (*Gallus domesticus*) genomes. J Hered. 1993;84(4):274–80. https://doi.org/10.1093/oxfordjournals.jhered.a111339.
41. Jurka J, Pethiyagoda C. Simple repetitive DNA sequences from primates: compilation and analysis. J Mol Evol. 1995;40(2):120–6. https://doi.org/10.1007/BF00167107.
42. Lith HA, Zutphen LFM. Characterization of rabbit DNA micros extracted from the EMBL nucleotide sequence database. Anim Genet. 1996;27(6):387–95. https://doi.org/10.1111/j.1365-2052.1996.tb00505.x.
43. Hammock EAD, Young LJ. Microsatellite instability generates diversity in brain and sociobehavioral traits. Science. 2005;308(5728):1630–4. https://doi.org/10.1126/science.1111427.
44. Gylfe AE, Tuupanen S, Hänninen U, Kondelin J, Ristolainen H, Katainen R, et al. Abstract 5193: novel candidate oncogenes with mutation hot spots in microsatellite unstable colorectal cancer. Cancer Res. 2014;74(19):5193.
45. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet. 2010;44(1):445–77. https://doi.org/10.1146/annurev-genet-072610-155046.
46. Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, de la Banda MG, et al. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res. 2005;15(4):537–51. https://doi.org/10.1101/gr.3096505.
47. Mularoni L, Ledda A, Toll-Riera M, Albà MM. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. Genome Res. 2010;20(6):745–54. https://doi.org/10.1101/gr.101261.109.
48. Gemayel R, Cho J, Boeynaems S, Verstrepen KJ. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. Genes. 2012;3(3):461–80. https://doi.org/10.3390/genes3030461.
49. Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. Unstable tandem repeats in promoters confer transcriptional evolvability. Science. 2009;324(5931):1213–6. https://doi.org/10.1126/science.1170097.
50. Morin GB. The human telomere terminal transferase enzyme is a ribonucleoprotein that synthesizes TTAGGG repeats. Cell. 1989;59(3):521–9. https://doi.org/10.1016/0092-8674(89)90035-4.
51. Casas-Vila N, Scheibe M, Freiwald A, Kappei D, Butter F. Identification of TTAGGG-binding proteins in *Neurospora crassa*, a fungus with vertebrate-like telomere repeats. BMC Genomics. 2015;16(1):1–9.

52. Sand L, Szuhai K, Hogendoorn P. Sequencing overview of Ewing sarcoma: a journey across genomic, epigenomic and transcriptomic landscapes. Int J Mol Sci. 2015;16(7):16176–215. https://doi.org/10.3390/ijms160716176.
53. Lai Y, Sun F. The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol. 2003;20(12):2123–31. https://doi.org/10.1093/molbev/msg228.
54. Bachtrog D, Agis M, Imhof M, Schlötterer C. Microsatellite variability differs between dinucleotide repeat motifs—evidence from *Drosophila melanogaster*. Mol Biol Evol. 2000;17(9):1277–85. https://doi.org/10.1093/oxfordjournals.molbev.a026411.
55. Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci U S A. 1997;94(3):1041–6. https://doi.org/10.1073/pnas.94.3.1041.
56. Schug MD, Hutter CM, Wetterstrand KA, Gaudette MS, Mackay TF, Aquadro CF. The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. Mol Biol Evol. 1998;15(12):1751–60. https://doi.org/10.1093/oxfordjournals.molbev.a025901.
57. Amos W, Flint J, Xu X. Heterozygosity increases microsatellite mutation rate, linking it to demographic history. BMC Genet. 2008;9(1):1–10.
58. Amos W. Heterozygosity increases microsatellite mutation rate. Biol Lett. 2016;12(1):20150902.
59. Primmer CR, Ellegren H, Saino N, Møller AP. Directional evolution in germline microsatellite mutations. Nat Genet. 1996;13(4):391–3. https://doi.org/10.1038/ng0896-391.
60. Ellegren H. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet. 2000;24(4):400–2. https://doi.org/10.1038/74249.
61. Whittaker JC, Harbord RM, Boxall N, Mackay I, Dawson G, Sibly RM. Likelihood-based estimation of microsatellite mutation rates. Genetics. 2003;164(2):781–7.
62. Seyfert AL, Cristescu MEA, Frisse L, Schaack S, Thomas WK, Lynch M. The rate and spectrum of microsatellite mutation in *Caenorhabditis elegans* and *Daphnia pulex*. Genetics. 2008;178(4):2113–21. https://doi.org/10.1534/genetics.107.081927.
63. Schlötterer C. Evolutionary dynamics of microsatellite DNA. Chromosoma. 2000;109(6):365–71. https://doi.org/10.1007/s004120000089.
64. Noble L. Microsatellites — evolution and applications. Heredity. 1999;83(5):633–4. https://doi.org/10.1038/sj.hdy.6886482.
65. Madesis P, Ganopoulos I, Tsaftaris A. Microsatellites: evolution and contribution. In: Kantartzi SK, Totowa NJ, editors. Microsatellites: Methods and Protocols. New York: Humana Press; 2013. p. 1–13.
66. Saeed AF, Wang R, Wang S. Microsatellites in pursuit of microbial genome evolution. Front Microbiol. 2016;6:1462.
67. Weber JL, Wong C. Mutation of human short tandem repeats. Hum Mol Genet. 1993;2(8):1123–8. https://doi.org/10.1093/hmg/2.8.1123.
68. Pearson CE, Sinden RR. Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. Curr Opin Struct Biol. 1998;8(3):321–30. https://doi.org/10.1016/S0959-440X(98)80065-1.
69. Sinden RR. Biological implications of the DNA structures associated with disease-causing triplet repeats. Am J Hum Genet. 1999;64(2):346–53. https://doi.org/10.1086/302271.
70. Richard GF, Pâques F. Mini- and microsatellite expansions: the recombination connection. EMBO Rep. 2000;1(2):122–6. https://doi.org/10.1093/embo-reports/kvd031.
71. Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 1994;371(6494):215–20. https://doi.org/10.1038/371215a0.
72. Liu S, Hou W, Sun T, Xu Y, Li P, Yue B, et al. Genome-wide mining and comparative analysis of microsatellites in three macaque species. Mol Gen Genomics. 2017;292(3):537–50. https://doi.org/10.1007/s00438-017-1289-1.
73. Xu Y, Li W, Hu Z, Zeng T, Shen Y, Liu S, et al. Genome-wide mining of perfect microsatellites and tetranucleotide orthologous microsatellites estimates in six primate species. Gene. 2018;643:124–32. https://doi.org/10.1016/j.gene.2017.12.008.
74. Xu Y, Hu Z, Wang C, Zhang X, Li J, Yue B. Characterization of perfect microsatellite based on genome-wide and chromosome level in rhesus monkey (*Macaca mulatta*). Gene. 2016;592(2):269–75. https://doi.org/10.1016/j.gene.2016.07.016.
75. Karaoglu H, Lee CMY, Meyer W. Survey of simple sequence repeats in completed fungal genomes. Mol Biol Evol. 2005;22(3):639–49. https://doi.org/10.1093/molbev/msi057.
76. Li C-Y, Liu L, Yang J, Li J-B, Su Y, Zhang Y, et al. Genome-wide analysis of microsatellite sequence in seven filamentous fungi. Interdiscip Sci: Comput Life Sci. 2009;1(2):141–50. https://doi.org/10.1007/s12539-009-0014-5.
77. Lim S, Notley-McRobb L, Lim M, Carter DA. A comparison of the nature and abundance of microsatellites in 14 fungal genomes. Fungal Genet Biol. 2004;41(11):1025–36. https://doi.org/10.1016/j.fgb.2004.08.004.
78. Murat C, Riccioni C, Belfiori B, Cichocki N, Labbé J, Morin E, et al. Distribution and localization of microsatellites in the Perigord black truffle genome and identification of new molecular markers. Fungal Genet Biol. 2011;48(6):592–601. https://doi.org/10.1016/j.fgb.2010.10.007.
79. Ohm RA, de Jong JF, Lugones LG, Aerts A, Kothe E, Stajich JE, et al. vanKuyk PA, Horton JS, Grigoriev IV, Wösten HAB. Genome sequence of the model mushroom *Schizophyllum commune*. Nat Biotechnol. 2010;28(9):957–63. https://doi.org/10.1038/nbt.1643.
80. Qian J, Xu H, Song J, Xu J, Zhu Y, Chen S. Genome-wide analysis of simple sequence repeats in the model medicinal mushroom *Ganoderma lucidum*. Gene. 2013;512(2):331–6. https://doi.org/10.1016/j.gene.2012.09.127.
81. Zhao X, Tan Z, Feng H, Yang R, Li M, Jiang J, et al. Microsatellites in different potyvirus genomes: survey and analysis. Gene. 2011;488(1):52–6. https://doi.org/10.1016/j.gene.2011.08.016.
82. Mrázek J, Guo X, Shah A. Simple sequence repeats in prokaryotic genomes. Proc Natl Acad Sci U S A. 2007;104(20):8472–7. https://doi.org/10.1073/pnas.0702412104.
83. Burranboina K, Abraham S, Murugan K, Bayyappa M, Yogisharadhya R, Raghavendra G. Genome wide identification and analysis of microsatellite repeats in the largest DNA viruses (Poxviridae family): an insilico approach. Annu Res Rev Biol. 2018;22(1):1–11. https://doi.org/10.9734/ARRB/2018/38367.
84. Zhou L, Deng L, Fu Y, Wu X, Zhao X, Chen Y, et al. Comparative analysis of microsatellites and compound microsatellites in T4-like viruses. Gene. 2016;575(2):695–701. https://doi.org/10.1016/j.gene.2015.09.053.
85. Du L, Zhang C, Liu Q, Zhang X, Yue B. Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. Bioinformatics. 2018;34(4):681–3. https://doi.org/10.1093/bioinformatics/btx665.
86. Kofler R, Schlötterer C, Lelley T. SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics. 2007;23(13):1683–5. https://doi.org/10.1093/bioinformatics/btm157.
87. Luo W, Nie Z, Zhan F, Wei J, Wang W, Gao Z. Rapid development of microsatellite markers for the endangered fsh *Schizothorax biddulphi* (Günther) using next generation sequencing and cross-species amplification. Int J Mol Sci. 2012;13(11):14946–55. https://doi.org/10.3390/ijms131114946.
88. Huang J, Li W, Jian Z, Yue B, Yan Y. Genome-wide distribution and organization of microsatellites in six species of birds. Biochem Syst Ecol. 2016;67:95–102. https://doi.org/10.1016/j.bse.2016.05.023.
89. Cai G, Leadbetter CW, Muehlbauer MF, Molnar TJ, Hillman BI. Genome-wide microsatellite identification in the fungus *Anisogramma anomala* using Illumina sequencing and genome assembly. PLoS One. 2013;8(11):e82408. https://doi.org/10.1371/journal.pone.0082408.
90. Wang Y, Chen M, Wang H, Wang JF, Bao D. Microsatellites in the genome of the edible mushroom, *Volvariella volvacea*. Biomed Res Int. 2014;2014:1–10.
91. Webster MT, Smith NGC, Ellegren H. Microsatellite evolution inferred from human– chimpanzee genomic sequence alignments. Proc Natl Acad Sci U S A. 2002;99(13):8748–53. https://doi.org/10.1073/pnas.122067599.
92. Pascual M, Schug MD, Aquadro CF. High density of long dinucleotide microsatellites in *Drosophila subobscura*. Mol Biol Evol. 2000;17(8):1259–67. https://doi.org/10.1093/oxfordjournals.molbev.a026409.
93. Schlötterer C, Harr B. *Drosophila virilis* has long and highly polymorphic microsatellites. Mol Biol Evol. 2000;17(11):1641–6. https://doi.org/10.1093/oxfordjournals.molbev.a026263.
94. Hancock JM. Simple sequences in a 'minimal ' genome. Nat Genet. 1996;14(1):14–5. https://doi.org/10.1038/ng0996-14.
95. Qi WH, Jiang XM, Du LM, Xiao GS, Hu TZ, Yue BS, et al. Genome-wide survey and analysis of microsatellite sequences in bovid species. PLoS One. 2015;10(7):e0133667. https://doi.org/10.1371/journal.pone.0133667.
96. Perinchery G, Nojima D, Goharderakhshan R, Tanaka Y, Alonzo J, Dahiya R. Microsatellite instability of dinucleotide tandem repeat sequences is higher than trinucleotide, tetranucleotide and pentanucleotide repeat sequences in prostate cancer. Int J Oncol. 2000;16(6):1203–9.
97. Borodulina OR, Golubchikova JS, Ustyantsev IG, Kramerov DA. Polyadenylation of RNA transcribed from mammalian SINEs by RNA polymerase III: complex requirements for nucleotide sequences. Biochim

Biophys Acta. 2016;1859(2):355–65. https://doi.org/10.1016/j.bbagrm.2015.12.003.
98. Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet. 2009;10(1):19–31. https://doi.org/10.1038/nrg2487.
99. Richardson SR, Morell S, Faulkner GJ. L1 retrotransposons and somatic mosaicism in the brain. Annu Rev Genet. 2014;48(1):1–27. https://doi.org/10.1146/annurev-genet-120213-092412.
100. Prasad MD. Survey and analysis of microsatellites in the silkworm, *Bombyx mori*: frequency, distribution, mutations, marker potential and their conservation in heterologous species. Genetics. 2005;169(1):197–214. https://doi.org/10.1534/genetics.104.031005.
101. Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y. Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. Genome Res. 2000;10(1):62–71.
102. Murray V. The frequency of poly(G) tracts in the human genome and their use as a sensor of DNA damage. Comput Biol Chem. 2015;54:13–7. https://doi.org/10.1016/j.compbiolchem.2014.11.006.
103. Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids Res. 1992;20(2):211–5. https://doi.org/10.1093/nar/20.2.211.
104. Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet. 2002;30(2):194–200. https://doi.org/10.1038/ng822.
105. Russell GJ, Walker PMB, Elton RA, Subak-Sharpe JH. Doublet frequency analysis of fractionated vertebrate nuclear DNA. J Mol Biol. 1976;108(1):1–20. https://doi.org/10.1016/S0022-2836(76)80090-3.
106. Swartz MN, Trautner TA, Kornberg A. Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. J Biol Chem. 1962;237(6):1961–7. https://doi.org/10.1016/S0021-9258(19)73967-2.
107. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature. 1978;274(5673):775–80. https://doi.org/10.1038/274775a0.
108. Bird AP. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 1980;8(7):1499–504. https://doi.org/10.1093/nar/8.7.1499.
109. Cooper DN, Taggart MH, Bird AP. Unmethlated domains in vertebrate DNA. Nucleic Acids Res. 1983;11(3):647–58. https://doi.org/10.1093/nar/11.3.647.
110. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell. 1985;40(1):91–9. https://doi.org/10.1016/0092-8674(85)90312-5.
111. Razin A. CpG methylation, chromatin structure and gene silencing—a three-way connection. EMBO J. 1998;17(17):4905–80. https://doi.org/10.1093/emboj/17.17.4905.
112. Eckert KA, Yan G, Hile SE. Mutation rate and specificity analysis of tetranucleotide microsatellite DNA alleles in somatic human cells. Mol Carcinog. 2002;34(3):140–50. https://doi.org/10.1002/mc.10058.
113. Wierdl M, Dominska M, Petes TD. Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics. 1997;146(3):769–79. https://doi.org/10.1093/genetics/146.3.769.

## Publisher's Note