Analysis of simple sequence repeats in *Period* genes in mammals

Gunja Snehi*

Department of Zoology, JNVU, Jodhpur, Rajasthan, India.

ABSTRACT

Circadian rhythm is driven by the circadian clock. One of the core clock genes identified is *Period* (*Per*) in *Drosophila* and *Period1* (*Per1*), *Period2* (*Per2*), and *Period3* (*Per3*) in mammals. The present study examines the presence of perfect simple sequence repeats (SSRs) in human *Period* genes and their orthologues in chimpanzee, gorilla, orangutan, gibbon, and mouse. Comparisons of all *Per* genes show a higher density of SSRs in *Per1* of all organisms followed by *Per3* and *Per2* with exceptions. Possibly due to higher chances of mutations, CG-rich repeats are not abundant in *Per* genes. This analysis shows gene and species-specific distribution of repeats along with region-specific distribution of repeat types and repeat motifs in exons and introns of the *Per* genes. There is diverse distribution of CG-rich repeat types and motifs in the three *Per* and also in exons and introns of the three genes in different organisms. Long perfect SSRs are generally not common. Since perfect SSRs are prone to mutations, there is a need to study SSRs, especially found in exons of human *Per* genes which could serve as markers for investigators to study the role of such SSRs in *Per*-related disorders.

1. INTRODUCTION

Circadian rhythm is the rhythmicity of 24 h, which is adjusted by environmental cues such as light. This rhythmicity affects behavioral and physiological activities of organisms and is found in almost all organisms, from cyanobacteria to mammals [1,2]. There are three important features of circadian rhythm; endogenous free running period, entrainment, and temperature compensation [2-4]. Circadian clock that drives circadian rhythm [3] is present in all the cells and tissues of the body [5]. Circadian clocks or rhythms influence behavior, physiology, sleep-wake cycle, metabolism, hormonal secretion, body temperature, etc. [6].

Disruption of circadian genes and rhythm may lead to disorders [7] that may result in lung tumorigenesis [8], sleep disorders, cancer, obesity, cardiovascular disease, immune system dysfunction, neurobehavioral disorders, drug and alcohol addiction, psychiatric conditions, etc. [9].

The mammalian circadian entrainment is a complex process. Expression of core clock genes and interaction of the genes and proteins with other components of the clock within the primary circadian clock, suprachiasmatic nucleus (SCN) is essential for SCN cells to act as pacemaker [10]. *Per* gene, one of the core clock genes [11], is named so because it affects the period of the circadian rhythm [11].

The orthologues of *Drosophila Per* [12] genes in mammals are *Per1*, *Per2,* and *Per3* [13]. A putative gene sequence (gb: AC027390.3)

showing high similarity with the three *Per* genes is named *Per4* though it is possibly a pseudogene [14]. Other clock genes are *Cry* (Cryptochrome), *Bmal1* (Brain and Muscle Arnt-Like1), *Tim* (Timeless gene), *Clk* (*Clock*) [11], etc. *Clk* gene is one among many factors that affect the expression and regulation of *Per* genes [10].

1.1. *Period* gene(s) (*Per*)

Per gene is one of the important components that affects circadian clock [15]. Besides their role in circadian rhythm, *Per* genes may regulate cell proliferation, cell division (cell cycle), apoptosis, DNA damage (double stranded break) repair pathway, etc., and have role in cancer (breast, bone marrow, myeloid leukemia, lung, endometrial, and pancreatic) [16].

Three paralogues of *Per* in vertebrates have possibly evolved from single ancestral gene due to genome duplication events [17]. Probably two genome duplications resulted in the *Per* genes paralogues. Alternatively, the four *Per* paralogues could have been a result of independent tandem duplication of the region containing ancestral *Per* [18].

Schematic representation of human karyotype with marked positions of human *Per* genes (*Per1-Per4*) as per Ensembl 2019 [19] is given in Figure 1. The detail about the four *Per* genes is given below:

1.1.1. Per1 gene
The first discovered *Per* gene was called *RIGUI* (named later as *Per1*) which is present on chromosome 17p12-13.1 in humans [20]. *Per1* has essential role in maintaining circadian rhythmicity [21]. It shows strong expression in mammalian SCN [12]. *Per1* is also expressed in

---

Corresponding Author:
Gunja Snehi,
Department of Zoology, JNVU, Jodhpur, Rajasthan, India.
E-mail: gunjasnehi01@gmail.com
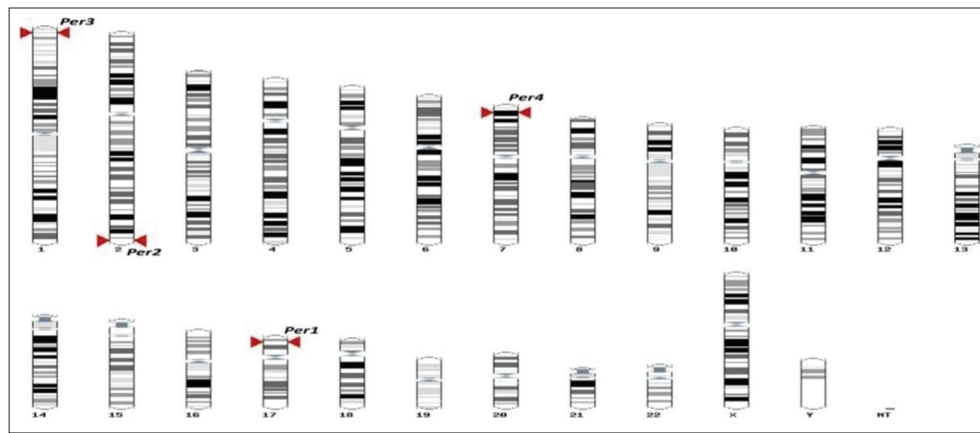
Figure 1: Schematic representation of human karyotype with marked positions of *Per* genes (*Per1 - Per4*) on respective chromosomes as per Ensembl 2019 [19]. Red arrows indicate positions of *Per* genes on respective chromosomes.

other parts of the brain [16,22] such as pineal and pituitary [16] and peripheral tissue [22]. Rhythmic expression of *Per1* is induced by light [16] which is differential [22].

Per1 is responsible for morning oscillation [23] as its rhythmic expression is reported in morning [24] possibly early morning along with CRY1 [23] which is high during day [22] or mid-day [25]. *Per1* or "morning phase clock gene" [26] is one of the circadian pacemakers expressed 4 h earlier than *Per2* [15,26]. Light-dark cycle induces PER1 that entrains the circadian clock by regulating PER2 protein. Thus, induction of *Per1* is essential for the expression of *Per2* [15].

1.1.2. Per2 gene
Human *Per2* is present on chromosome number 2 [14], to be precise on 2q37.3 [15]. *Per2* is one of the circadian pacemaker [15] that plays an important role in circadian time keeping in humans [27]. *Per2* is a positive regulator of circadian gene expression [11,16,21] but possibly delays the clock [11,23]. *Per2* is expressed in SCN [10,11,15,16] and in other regions of brain [16], eyes [22], and other peripheral tissue [11,16,22]. However, expression in SCN is higher [22,26].

Like other period genes, *Per2* is activated by light [28] though not directly [11] and entrained by dark-light cycle [11,22]. Light induces differential regulation of expression of *Per2* [22] during late day time [25] or in the evening [24] but expression is high during day [22]. Thus, *Per2* is also known as "afternoon phase clock gene" [26]. However, there is also report of highest expression during early evening. Thus, it is the evening oscillator along with CRY2 though *Per2* inducibility is limited during early night [23].

1.1.3. Per3 gene
Human *Per3* gene is present on chromosome number 1 [14,28]. *Per3* may not be essential as compared to *Per1* and *Per2* [11] for driving the central circadian clock [17] or core clock loop [21] but maintains period length of the circadian rhythm [29]. *Per3* expression in SCN [30] is rhythmic [17] which is highest in mammalian SCN besides its expression in other brain regions [28,30] and peripheral tissue [17,28,30] besides retina [30].

Unlike *Per1* and *Per2*, *Per3* is not induced by light [28,30] as the levels of *Per3* mRNA do not fluctuate in response to light pulse during night [30].

1.1.4. Per4 gene
Human *Per4*, presumed lost [17,18], is present on chromosome 7 [14]. However, contradictory position is reported on chromosome 3 near telomere [18]. Although *Per4* has sequence similarity with other three *Per* genes, it is a pseudogene [14,31] possibly due to silencing [18].

1.2. Simple Sequence Repeats (SSRs) or Microsatellites

SSRs or microsatellites are tandem repeats of nucleotides in DNA [32-35] and are also called short tandem repeats (STRs) [32]. SSRs have 2-6 bp per repeat unit and the array size should be 8 bps to less than one kilo base pairs (kbp) [33]. SSRs are found in most genomes [34] and are distributed in exon, intron and intergenic regions in eukaryotic genomes including primates [36]. These are abundant in non-coding regions [36,37] and in intergenic regions [36] compared with coding regions [33,38].

SSRs may be divided into different groups on different bases, but on the basis of the sequence of bases in repeat motifs, SSRs may be categorized in four types – perfect (when repeats are same such as $(GT)_9$), imperfect (when the non-repeated nucleotides are also present between the repeated ones such as $(GT)_6AT(GT)_8$), interrupted (when there is a small sequence within the repeats such as $(GT)_5CGTA(GT)_4$), and composite or compound (when two different repeats are present together such as $(CT)_5(AG)_7$) [39,40].

SSRs are vulnerable to mutations [34,41], as these repeats are hypervariable [42]. Mutations in SSRs may lead to disorders [43] such as cancer [44], colorectal cancer [41], and neurological disorder like Huntington's disease [32] and can affect protein interaction/binding or its flexibility [34]. SSRs with smaller units like dinucleotide repeats have slower evolution (mutation) rate compared with SSRs with larger units like tetranucleotide repeats possibly due to failure of mismatch repair for larger regions [33].

SSRs may have functional roles such as involvement in gene regulation [32,34,37,45], chromatin organization [34,45], alternate splicing, mRNA transport [45], affect evolutionary rates, adaptation, behavior [34,37], alter interaction between cells, circadian rhythm, and morphology [32], and phenotype [34]. SSRs translated as amino acid repeats, especially homopolymer repeats, generally have disordered regions and are associated with proteins involved in transcription and translation functions concerning DNA-protein or protein-protein interactions [43]. Further, SSRs may provide adaptability to environmental stress [42].

1.3. Aim of the Study and its Future Perspective

Repeats including tandem repeats have been found in circadian rhythm genes including *Per* genes [34]. Studies report presence of repeats in *Per* genes [17,46]. However, there has been no study on SSRs and in particular perfect SSRs in human *Per* genes and comparison of SSRs with orthologues. This study aims to fill this gap in analysis of repeats (if present) in human *Per* genes and compare the mono- to hexa-nucleotide perfect repeats with orthologue sequences in chimpanzee, gorilla, orangutan, gibbon, and mouse. Study of perfect SSRs is important because there are higher chances of mutations in such repeats [34] compared with imperfect SSRs [47,48]. The study may help in identification of sequences that may be useful for monitoring mutational regions that may be important for study of disorders in humans due to mutations in *Per* genes.

2. MATERIALS AND METHODS

2.1. Obtaining Gene IDs

Human (*Homo sapiens*) *Period* (*Per1*, *Per2,* and *Per3*) Ensembl IDs were obtained from the NCBI database (https://www.ncbi.nlm.nih.gov/gene/) [49]. The human Ensembl IDs obtained from NCBI were submitted to Ensembl genome database version 98 (https://asia.ensembl.org/index.html) through the BioMart interface [50] to obtain Ensembl gene IDs of human period genes and their orthologues in chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), gibbon (*Nomascus leucogenys*), orangutan (*Pongo abelii*), and mouse (*Mus musculus*).

2.2. Obtaining gene Information and Sequences

Unspliced period gene (gene containing both exons and introns) sequences of human and its above mentioned orthologs were retrieved using the Ensembl gene IDs. Other parameters such as gene start, gene end, transcript start, transcript end, strand, transcript count, gene CG% from features and 5'-UTR start, 5'-UTR end, 3'-UTR start, 3'-UTR end, CDS length, exon region start, exon region end, and exon rank associated with period genes of human and other orthologs were also obtained from Ensembl genome database version 98 (https://asia.ensembl.org/index.html) through the BioMart interface [50].

2.3. SSR Search

The software, SciRoKo [51], was used to search perfect SSRs in unspliced period gene sequences of human and respective orthologue sequences in chimpanzee, gorilla, gibbon, orangutan, and mouse. The default parameters were selected except for choosing minimum repeat as 3 instead of default value 4. The SciRoKo result was analyzed on the basis of motifs, SSR lengths, CG richness, and position of repeats (whether present in exon or in intron) in the gene sequences.

2.4. Calculating Repeat Densities and Position of Repeats in Exon/Intron

Repeat length less than 8nt were not considered for final analysis from the repeat search output obtained from SciRoko3.4 [51].

Repeat density per kbp in each gene was determined by dividing total number of repeats with gene length for each period gene. Further, it was determined that the repeats are present in the exonic or intronic region by analyzing the SSR start and SSR end and comparing these with the gene sequences obtained from Ensembl 98.

2.5. SSR Length, Motif Densities, and CG%

SSR lengths in period genes were measured based on number of nucleotides between SSR start and SSR end positions in each gene as per SciRoKo output. For SSR CG percentage, total numbers of occurrences of C/G in each SSR unit were considered. For example, CG richness in C_n/G_n repeats is 100% and in A_n/T_n is 0%. For dinucleotide repeats such as $(AG)_n/(TC)_n$, the CG richness is 50% (1/2*100) and for trinucleotide repeats, $(ACT)_n$, CG% is 33.33% (1/3*100), and so on. SSR motif densities of period genes were also calculated by dividing length of each motif repeat with the respective gene length. Heat maps of motif densities were produced for representing the densities by using Matrix2png [52].

3. RESULTS AND DISCUSSION

3.1. Density of SSRs in *Per* Genes

Among all the organisms, density of SSRs is highest in orangutan *Per1* gene followed by mouse *Per3*. SSR density is highest in *Per1*, followed by *Per3* with exceptions [Figure 2]. This is despite the fact that *Per3* gene is considered highly polymorphic compared with the *Per1* and *Per2* genes [17]. This indicates that possibly perfect SSRs are unlikely to be source of the polymorphism in *Per3* gene.

Among the SSR types, mono-, di-, and tri-nucleotide repeats are present in all genes of all organisms but tetra-, penta-, and hexa-nucleotide SSRs are not present in *Per* genes of all organisms. Further, comparison of SSR types shows highest density of trinucleotide SSRs in *Per1* of all organisms, dinucleotide SSRs in *Per2* of all organisms except gibbon, and mouse and mononucleotide SSRs in all *Per3* except mouse *Per3*. It is also found that mono-, di-, and tri-nucleotide SSRs are abundant in *Per3*, *Per2*, and *Per1* genes with some exceptions. Tetra-, penta-, and hexa-nucleotide SSRs are present in very less amount [Figure 3].

Further, there is gene-specific distribution of SSRs as *Per1* genes of all organisms have highest density of SSR per kbp compared with *Per2* and *Per3* genes. This is in corroboration with studies that show strain specific [46] or taxon/species-specific distribution in eukaryotic genomes including primates [36] and gene-specific distribution of SSRs [36]. Further, present study agrees with finding of all the repeat types that have different distribution in different species [53]. *Per1* gene is found to have the highest SSR density in all organisms except mouse. Higher SSR density may be due to some essential roles of SSRs in *Per1* gene since *Per1* has essential role in circadian rhythm [21].

The presence of SSR motifs and their densities show diverse patterns. Mononucleotide SSRs $(A/T)_n$ are present in all genes of all organisms. $(C/G)_n$ SSRs are present in very low amount in some organisms. Dinucleotide SSRs $(CG)_n$ are present only in *Per3* genes of human and mouse but $(GC)_n$ are absent in *Per3* genes of all organisms. Similarly, trinucleotide motif SSRs distribution is also very diverse. Few motifs show gene and organism-specific distributions such as $(AAC)_n$ is only found in mouse *Per2*, $(AAT)_n$ in orangutan *Per3*, $(ACA)_n$ in mouse *Per3*, and $(ACT)_n$ in gibbon *Per2* [Figure 4].

Mono- to tri-nucleotide repeats are present in *Per1* genes of all organisms. $(A/T)_n$ mononucleotide motifs are present in *Per1* of all organisms, $(C)_n$ is present only in mouse, and motif $(G)_n$ is present only in gorilla and gibbon genes. This shows that C/G rich repeats are avoided due to their higher chances of elongation or slippage [54]. SSRs with larger motifs are not present in *Per1* genes of all organisms except TCTG motif in mouse and GTTTTT motif in human, orangutan, and
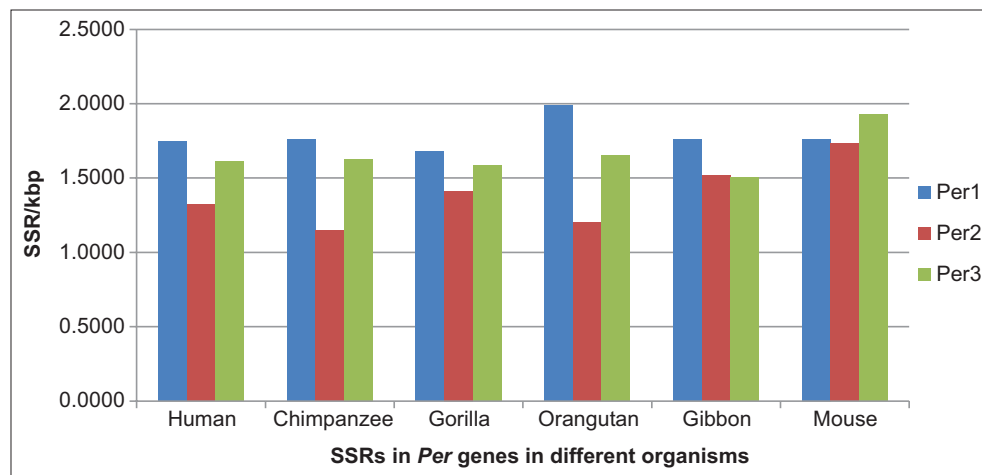
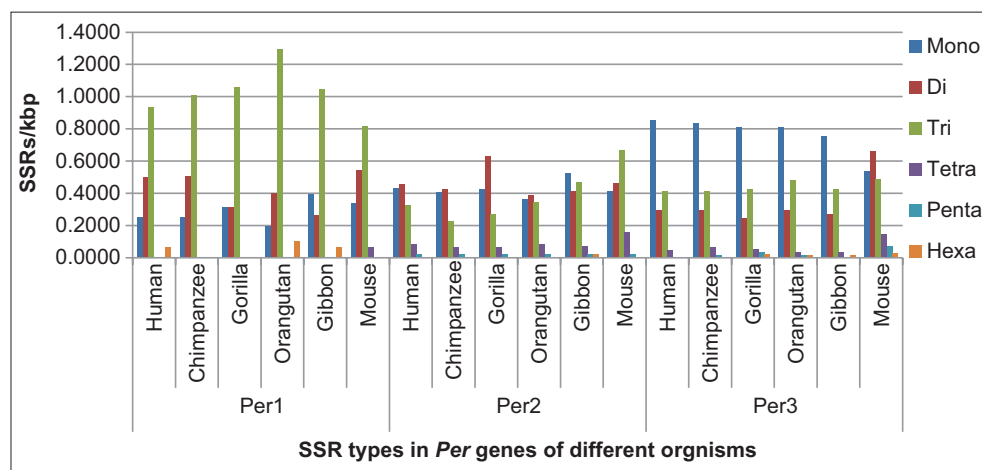**Figure 2:** Density of simple sequence repeats in *Per* genes.

Figure 3: Density of simple sequence repeat types in *Per* genes.

gibbon genes. Since larger motifs have higher chances of mutations compared with smaller motifs [38,48], possibly perfect tandem repeats of such motifs are avoided in *Per1* genes.

Per1 genes of all organisms do not have tetra-, penta-, and hexa-nucleotide repeat motifs except TCTG tetranucleotide repeat in mouse and GTTTTT hexa-nucleotide motif in human, orangutan, and gibbon genes. Similarly, numbers of tetra-, penta-, and hexa-nucleotide repeats are few in *Per2* and *Per3* genes of some organisms [Figure 5].

3.2. SSR Densities in Exons and Introns of *Per* Genes

The analysis shows higher density of SSRs in exons of *Per1* of all organisms except orangutan and mouse. *Per2* genes of all organisms except orangutan have high SSR densities in introns. Except human, gibbon, and mouse *Per3*, SSR densities are higher in introns of *Per3* genes of all organisms [Figure 6].

Per1 genes have mononucleotide repeats in exons of chimpanzee and mouse genes. Further, density of mononucleotide is lower in introns of chimpanzee gene but is higher in exons in mouse gene. Mononucleotide repeats are not present in gorilla and gibbon *Per3* exons. When present, mononucleotide repeat density is higher in introns of *Per2* and *Per3* genes in all organisms. Dinucleotide repeats show variations in densities in exons and introns of the *Per* genes. When present, dinucleotide repeats have higher density in introns of

Per1 genes of all organisms, but in *Per2,* it is higher in exons except mouse gene and in *Per3,* it is higher in exons of orangutan, gibbon, and mouse. Trinucleotide repeats, if present, have higher density in exons of *Per1* and *Per3* genes of all organisms. However, introns of *Per2* genes have higher density of trinucleotide repeats in introns of all organisms. Tetranucleotide repeats, if present in *Per* gene, are present in introns but present in exons of only *Per3* genes of human and mouse and is higher in exons compared with introns. Penta- and hexa-nucleotide repeats are not present in exons of the *Per* genes of any organism in the present study [Figure 7].

Possibly all the repeats, especially in exons or introns, found in *Per1* are required for the gene function. This is because SSR mutation may lead to loss of functional gene or protein [34]. If such mutations happen in *Per1* SSRs, loss of *Per1* function may lead to higher levels of PER2 by having effect on its post- transcriptional regulation by PER1 and affect period length [55] or loss of circadian rhythm [21].

Most SSRs were found to be present in introns as previously reported [35]. This result suggests that any mutation in such repeats does not affect gene expression. The repeats which have some functional roles and are triplets or multiple of triplets are present in exons. The SSRs, other than tri- and hexa-nucleotide SSRs, cause frameshift mutations, so these are present in less frequency in coding regions [56].
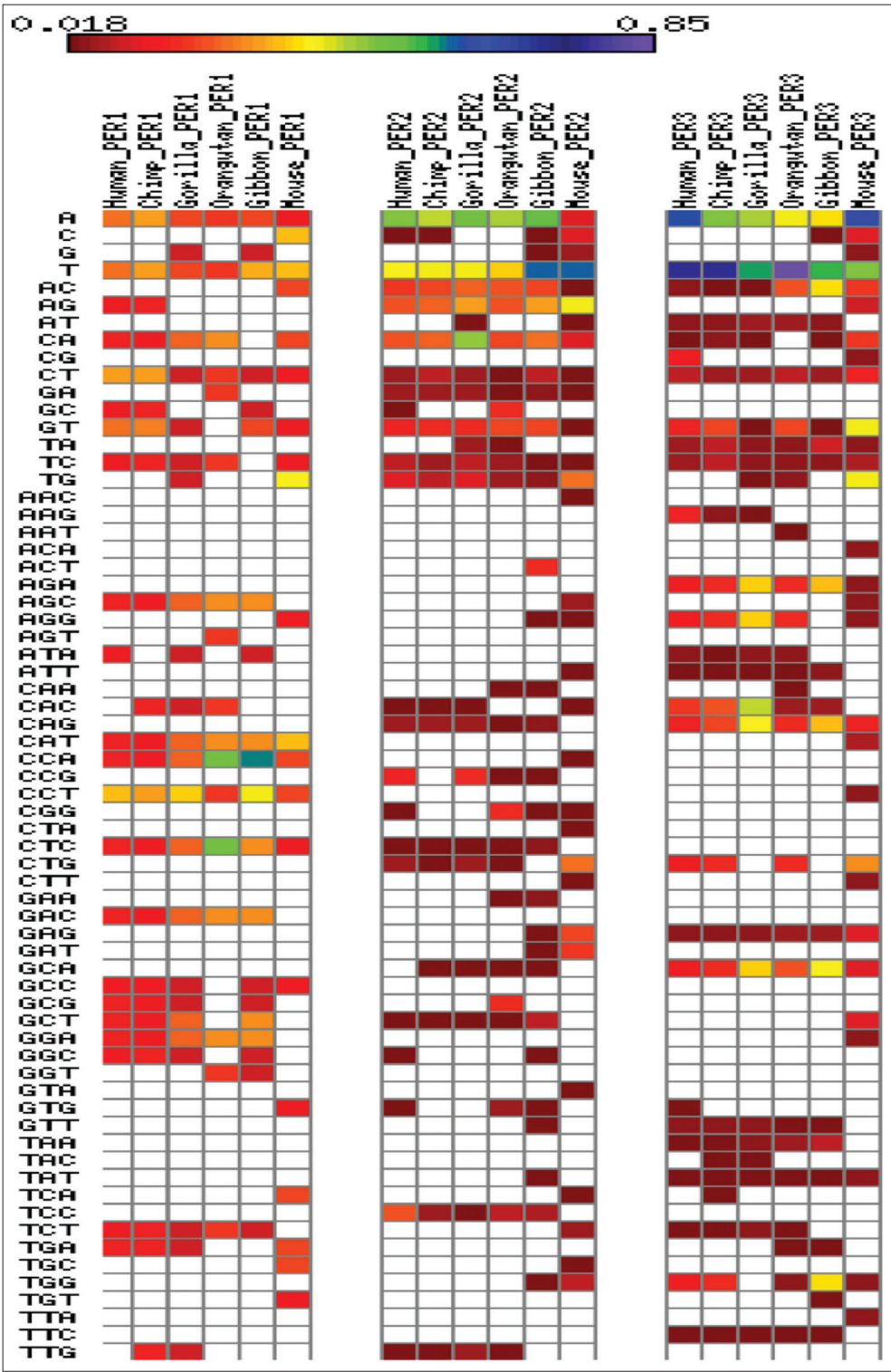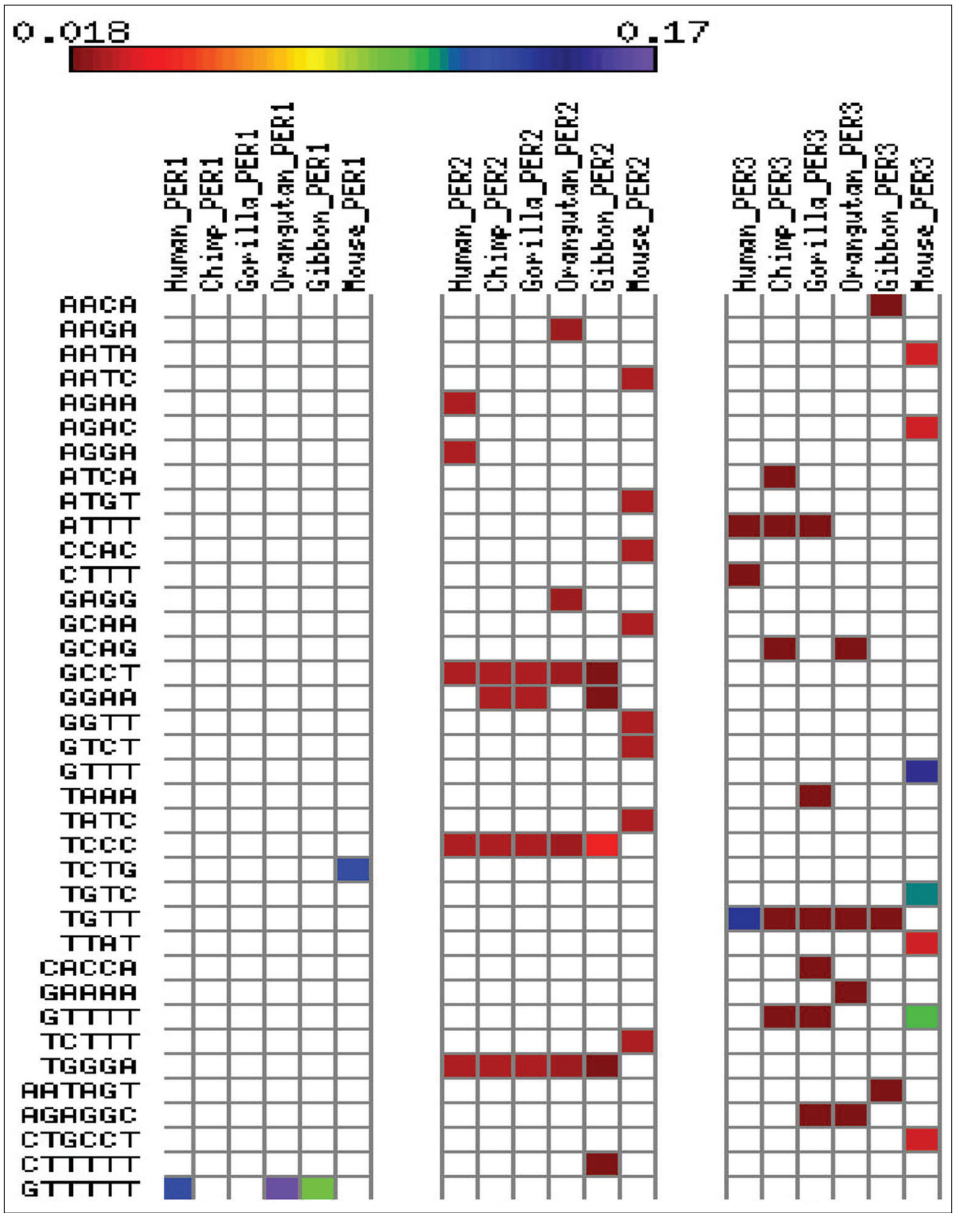
Figure 4: Density of mono-, di-, and tri-nucleotide motifs in *Per* genes of human, chimpanzee, gorilla, orangutan, gibbon, and mouse. Blank cells indicate absence of motifs. Legend color intensity indicates density per kbp of motifs.
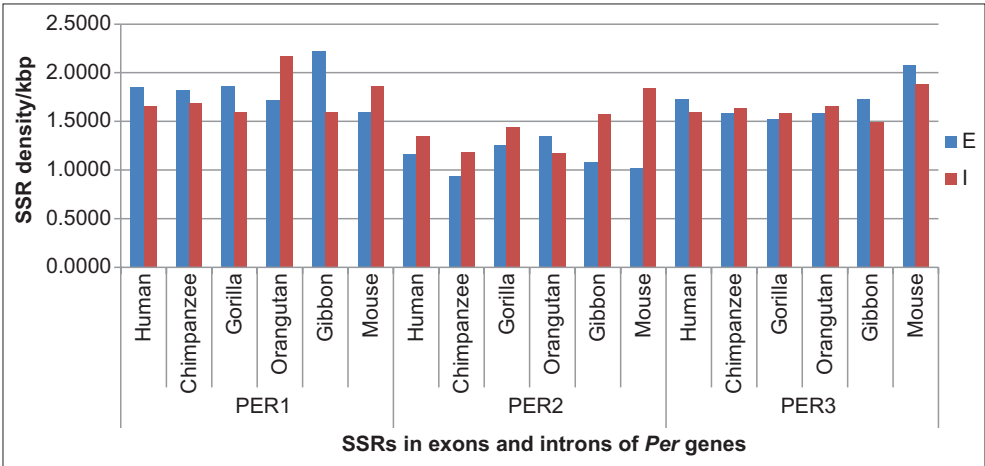
3.3. SSR Motifs in Exons and Introns of *Per* Genes

Some repeats show different distribution patterns in exons and introns of all *Per* genes of all organisms. For example, mononucleotide repeat A is present in exons of only chimpanzee *Per1*, but is present in exons of *Per2* genes of all organisms except mouse and present in exons of *Per3* of only human and mouse. This motif is present in introns of all *Per* genes of all organisms. Dinucleotide repeat motif CG is present only in human *Per3* exon and intron of mouse *Per3* gene. Trinucleotide motif TCC is present only in human *Per2* exon [Figure 8]. Tetranucleotide motifs are present in introns of *Per1* and *Per2* genes of some organisms. *Per3* gene exons have only tetranucleotide motif TGTT and present only in human gene

Figure 5: Density of tetra-, penta-, and hexa-nucleotide motifs in *Per* genes of human, chimpanzee, gorilla, orangutan, gibbon, and mouse. Blank cells indicate absence of motifs. Legend color intensity indicates density per kbp of motifs.

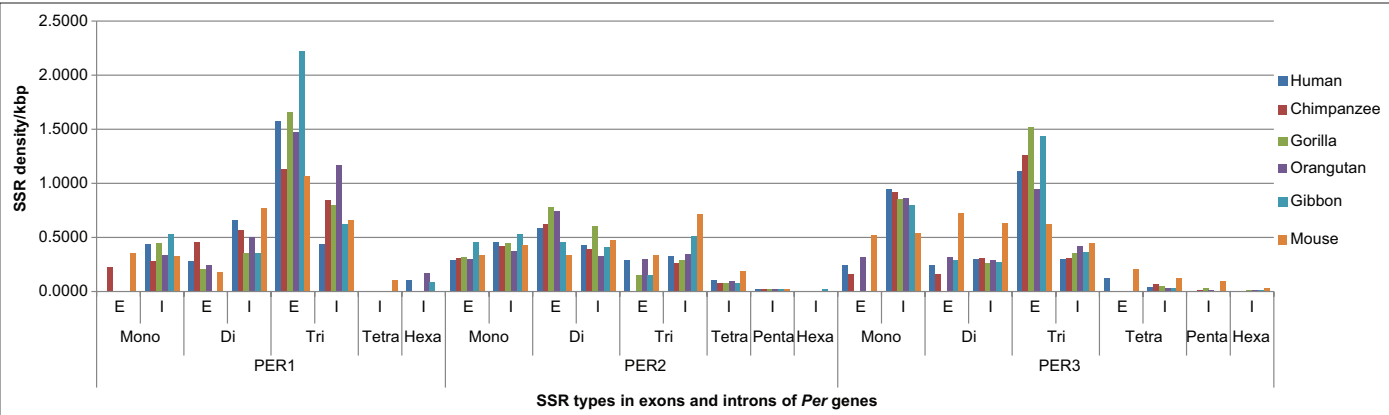**Figure 6:** Density of simple sequence repeats in exons and introns of *Per* genes.

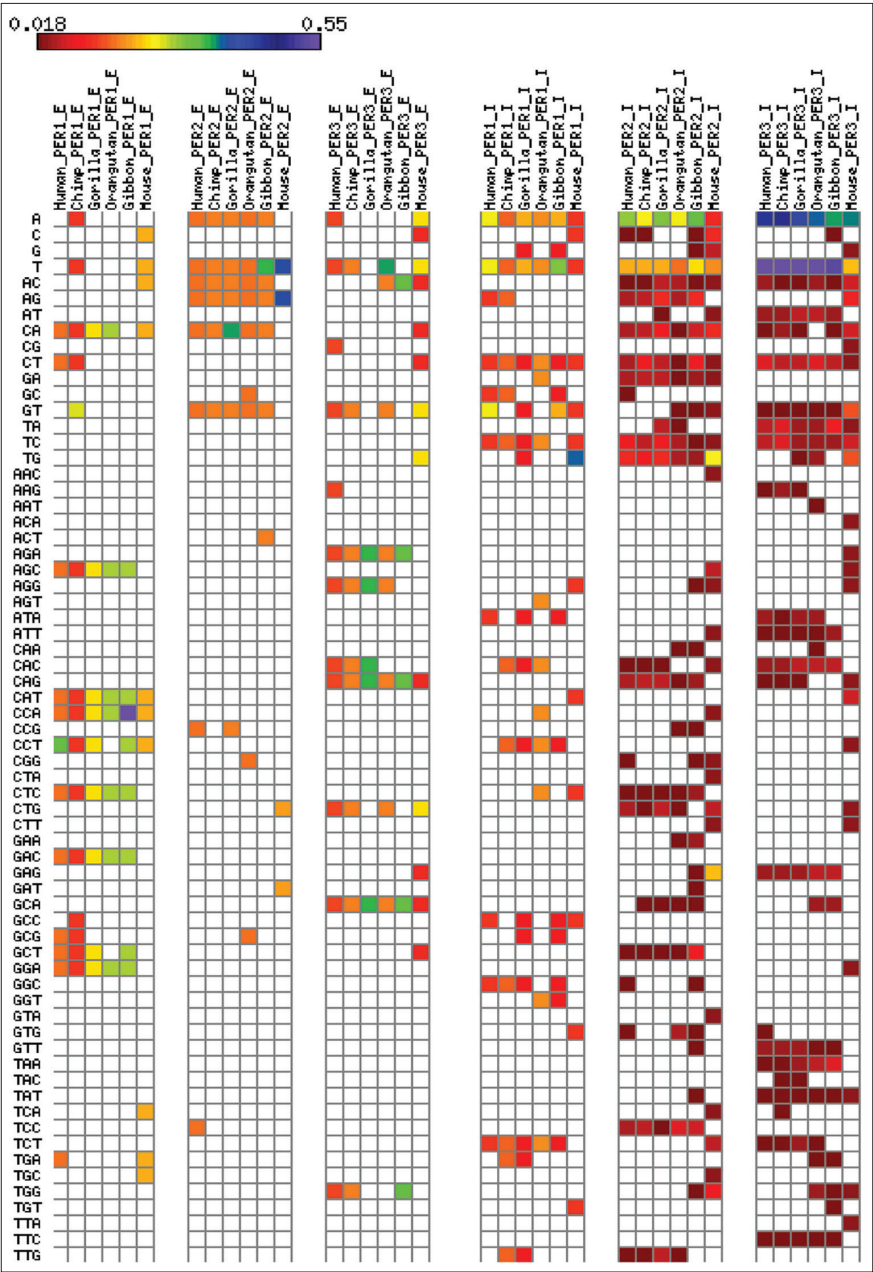**Figure 7:** Density of simple sequence repeat types in exons and introns of *Per* genes.

Figure 8: Density of mono-, di-, and tri-nucleotide motifs in exons (e) and introns (i) of *Per* genes of human and its orthologues. Blank cells indicate absence of motifs. Legend color intensity indicates density per kbp of motifs.

whereas GTTT and TGTC motifs are present only in mouse *Per3* exons. Penta- and hexa-nucleotide motifs are not present in exons of any *Per* gene of any organism. Similarly, introns of *Per1* genes of all organisms do not have tetra- and hexa-nucleotide motifs except TCTG in mouse gene and hexa-nucleotide motif GTTTTT in human, orangutan, and gibbon genes [Figure 9].

Unlike many studies reviewed in Ellegren 2004, that report abundance of repeats in introns [35], the present study shows abundance of repeats in exons of *Per1* (except orangutan) besides higher density in orangutan *Per2* exon and human and mouse *Per3* exons.

3.4. SSR CG Percentage in *Per* Genes

Per3 genes have high AT rich repeats in all organisms followed by *Per2* genes [Figure 10]. Motifs with 50% and 66.67% CG richness are present in all *Per* genes of all organisms. However, *Per1* genes of all organisms do not have repeats with 75% CG richness, present in *Per2* of all organisms but present only in chimpanzee and orangutan *Per3* genes. SSR motifs with 100% CG richness are present in *Per1* of all organisms except orangutan, *Per2* of all organisms, and *Per3* genes of human, gibbon, and mouse [Figure 11].
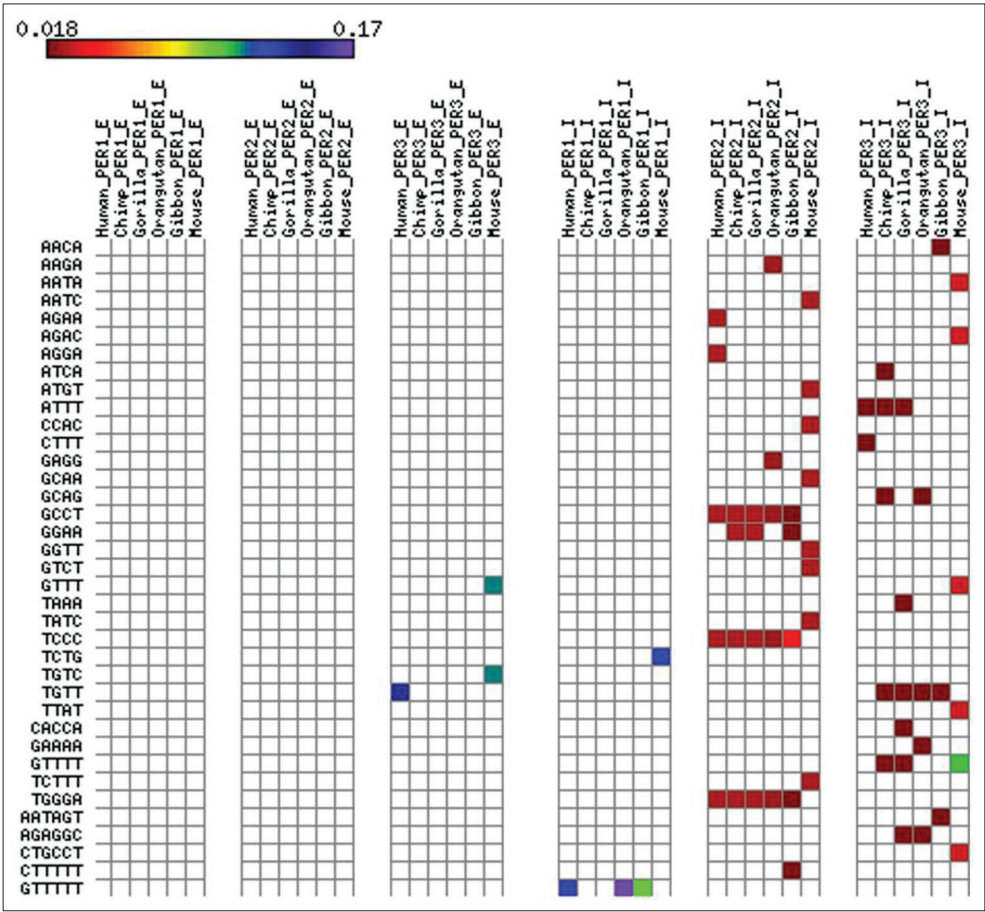

Figure 9: Density of tetra-, penta-, and hexa-nucleotide motifs in exons (e) and introns (i) of *Per* genes of human and its orthologues. Blank cells indicate absence of motifs. Legend color intensity indicates density per kbp of motifs.
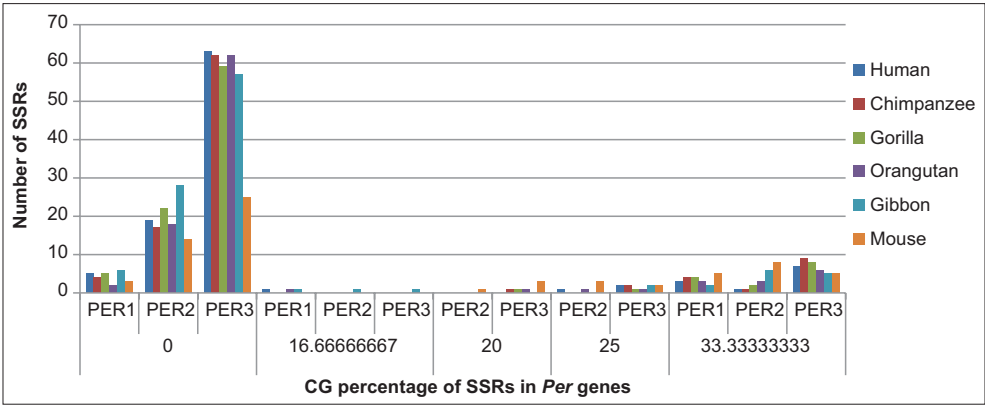

Figure 10: CG percentage of simple sequence repeats in *Per* genes.

AT rich mononucleotide repeats are present in all three *Per* genes of all organisms. *Per3* genes have highest number of AT rich repeats followed by *Per2* and *Per1*. However, numbers of CG-rich mononucleotide repeats are very low in the *Per* genes. AT rich dinucleotides are not present in *Per1* genes of any organism but 50% CG-rich dinucleotide repeats are present in all three *Per* genes of all organisms. However, 100% CG-rich dinucleotide repeats are not abundant [Figure 12]. AT rich trinucleotides are not abundant but are present in *Per3* genes of all

organisms. >30% and less than 70% CG-rich trinucleotide repeats are present in all three *Per* genes of all organisms. However, trinucleotides with 100% CG richness are not present in *Per3* and are not abundant in *Per1* and *Per2* genes. Tetranucleotide repeat with 100% CG-rich is not present in any *Per* gene of any organism and 50% CG-rich tetranucleotide repeats are present in *Per1* and *Per3* genes of mouse only [Figure 13]. Penta- and hexa-nucleotide repeats are generally not CG-rich as none have >66.67% CG richness in any *Per* gene [Figure 14].
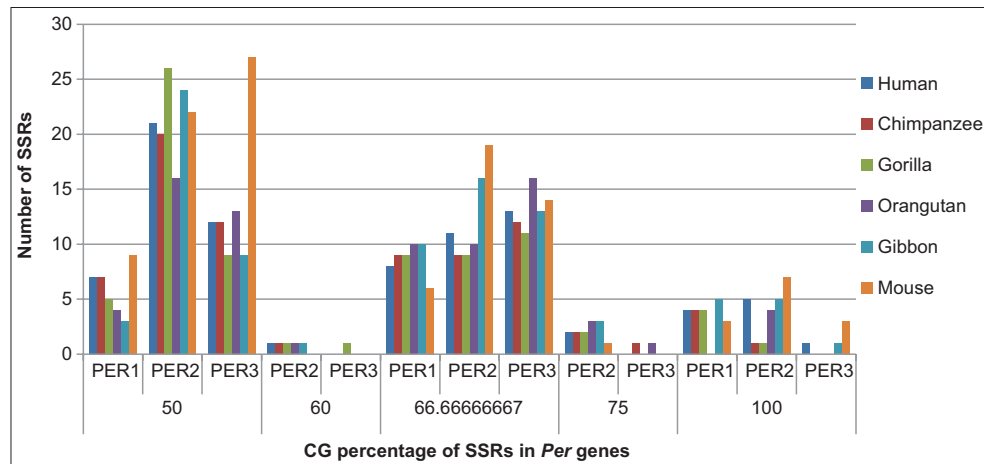

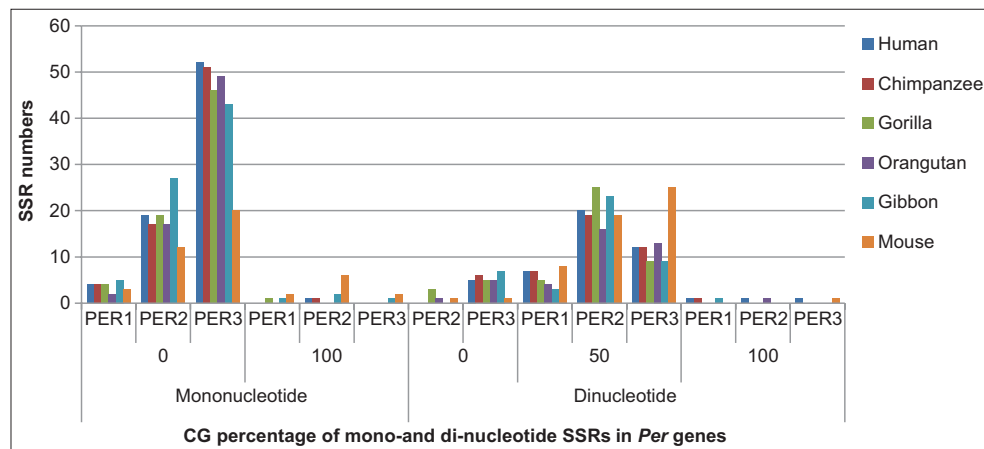**Figure 11:** CG percentage of simple sequence repeats in *Per* genes.

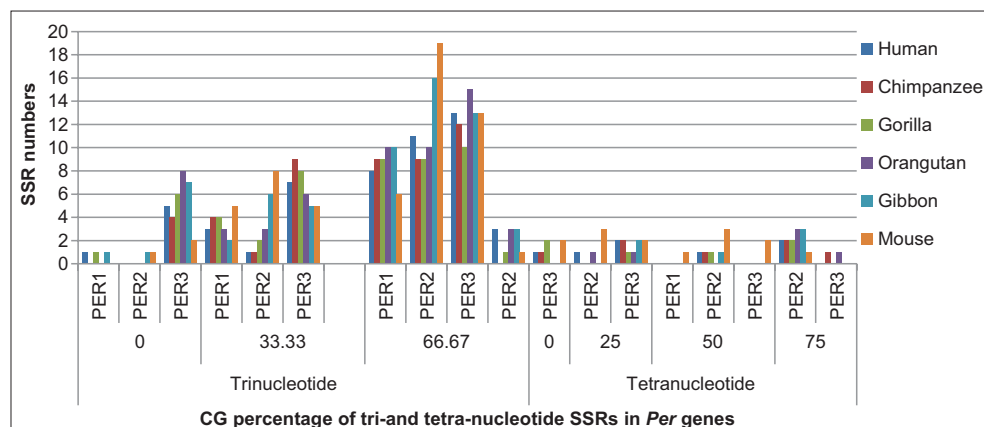**Figure 12:** CG percentage of mono- and di-nucleotide simple sequence repeats in *Per* genes.

Figure 13: CG percentage of tri- and tetra-nucleotide simple sequence repeats in *Per* genes.

Most SSRs were found to have low CG richness. Low percentage of CG is to avoid mutation due to strand slippage [38,57].

3.5. SSR CG Richness in Exons and Introns of *Per* Genes

Number of AT rich repeats is higher in introns compared with exons of all three genes in all organisms. However, repeats with

higher CG richness are also abundant in introns of *Per* genes [Figures 15 and 16].

All the repeat types mono- to hexa-nucleotides with different CG percentage are also present in introns and exons of all three genes of all organisms. However, generally, the number of repeats is higher in introns of all genes [Figures 17 and 18]. Tetranucleotide motifs with
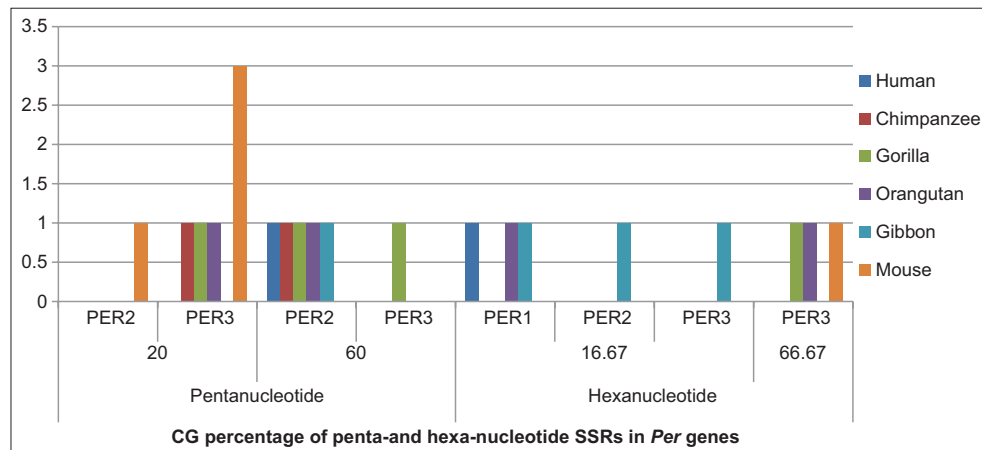

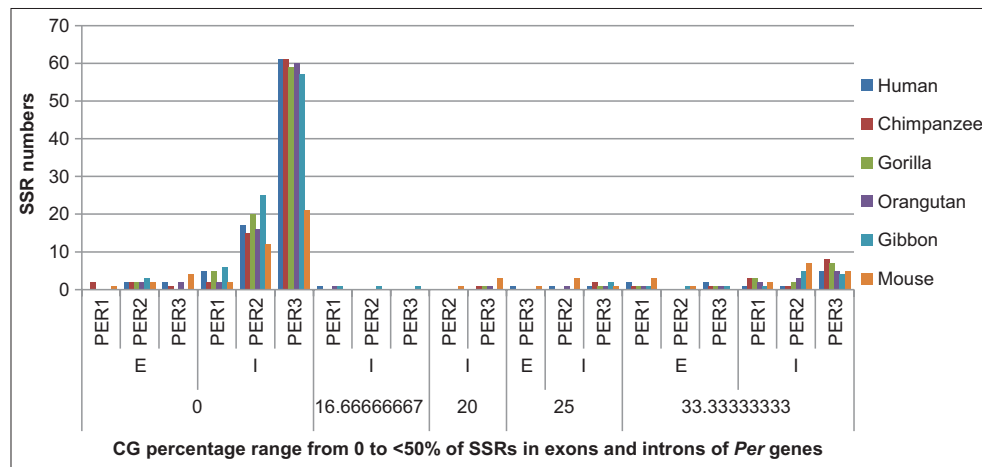**Figure 14:** CG percentage of penta- and hexa-nucleotide simple sequence repeats in *Per* genes.

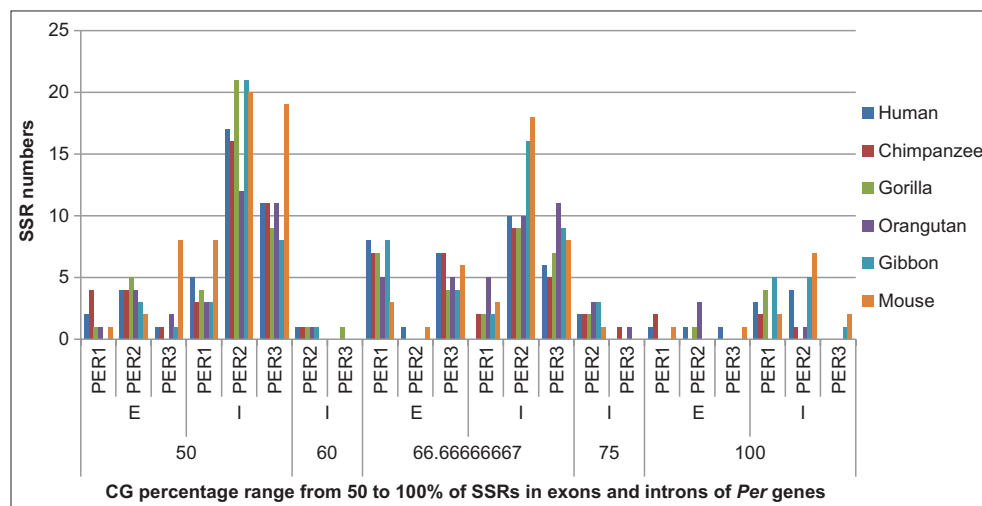**Figure 15:** CG percentage range from 0 to <50% of simple sequence repeats in exons and introns of *Per* genes.

Figure 16: CG percentage range from 50 to 100% of simple sequence repeats in exons and introns of *Per* genes.
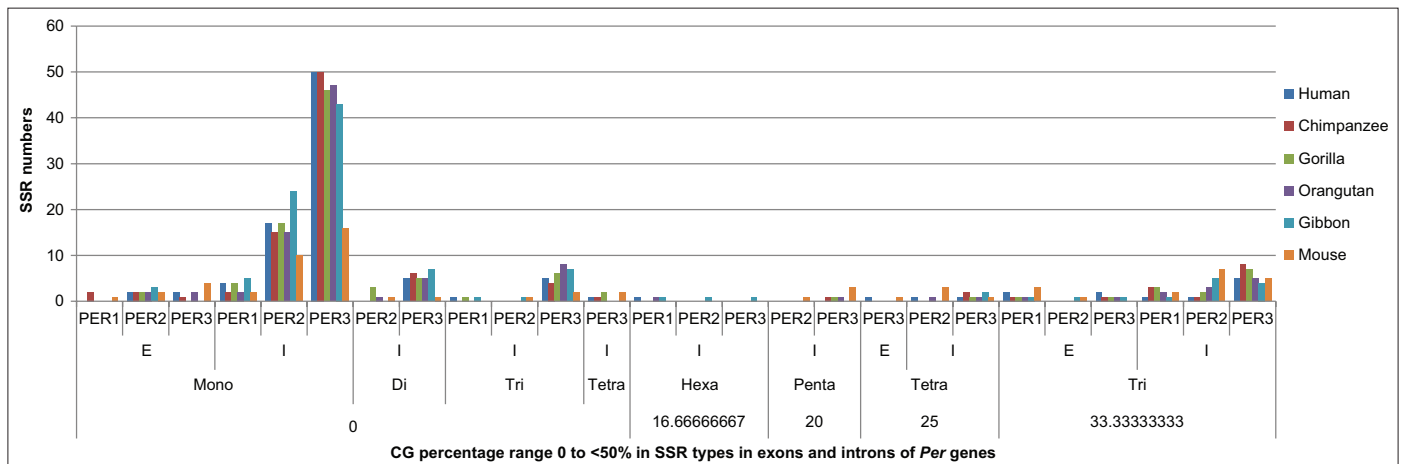
Figure 17: CG percentage range 0 to <50% in simple sequence repeat types in exons and introns of *Per* genes.
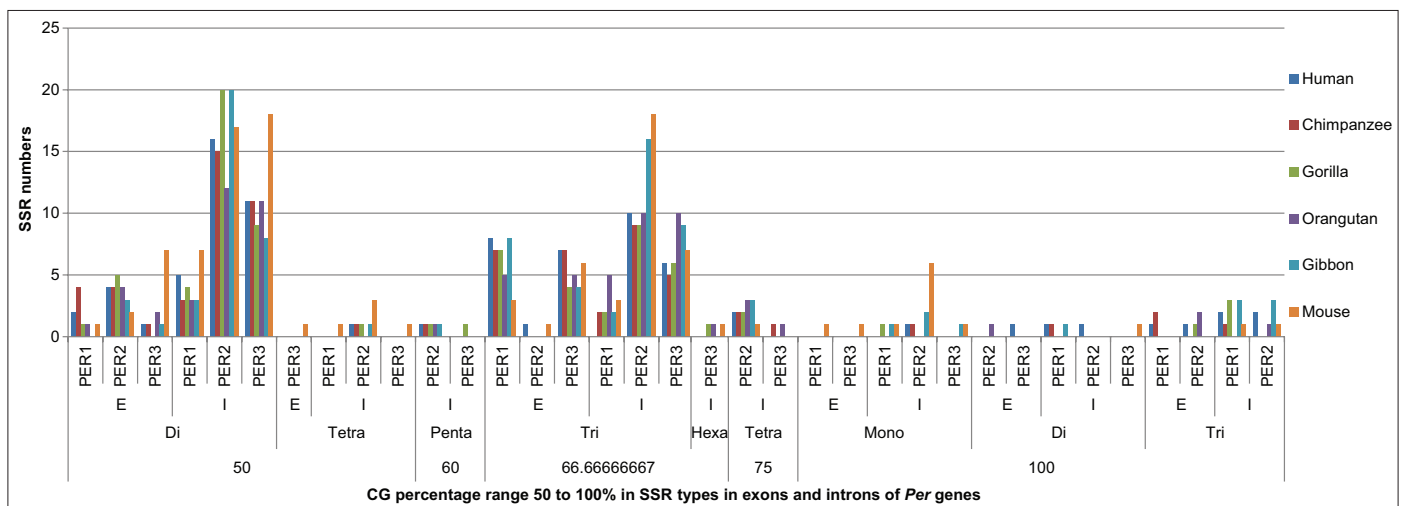

Figure 18: CG percentage range 50 to 100% in simple sequence repeat types in exons and introns of *Per* genes.

50% CG percentage are not present in exons of any *Per* gene except mouse *Per3*. SSR motifs with 60% and 75% CG percentage are not present in exons of any gene in any organism. However, repeat motifs with 66.67% CG percentage are present in exons of *Per1* and *Per3* genes of all organisms but in *Per2* exons of human and mouse only. The distribution of SSR types with 100% CG richness is different in exons of *Per* genes [Figures 17 and 18].

Most of the repeats are 50% CG-rich and are present in introns. This finding is in corroboration of other studies which suggest that CG content is low in genomes, especially in codons [58,59].

3.6. SSR Length in *Per* Genes

Longest repeat (79 nt long) is present in *Per3* gene of chimpanzee. Longest repeat in *Per2* genes is 44nt SSR and *Per1* genes have longest repeats of 28nt. Numbers of repeats with length ranging from 8nt to 28nt are generally higher in *Per3* followed by *Per2* and *Per1* genes in all organisms with exceptions [Figure 19].

Since longer repeats cause instability, short repeat motifs are more abundant [60] and longer repeats are avoided due to their higher chances of mutations [38,48]. In agreement with this fact, this study also finds that most SSRs are short to evade higher chances of slippage.

3.7. Length of SSR Types in *Per* Genes

Mononucleotide repeat lengths range from 8nt to 79nt. Dinucleotide repeat lengths range from 8 to 52nt. However, longest dinucleotide repeat is 9nt in *Per1* gene except human gene where it is 13nt long. Longest dinucleotide repeats in *Per2* and *Per3* genes are 41nt and 52nt, respectively, but only in mouse. Longest trinucleotide repeat is 18nt in *Per1* (only mouse), 21nt in *Per2* (only orangutan). and 22nt in *Per3* (only gorilla). *Per1* genes of all organisms do not have tetranucleotide repeats except mouse where it is 12nt long. Longest tetranucleotide repeat in *Per2* and *Per3* are 44nt and 25nt long, respectively, but present only in mouse genes. Longest pentanucleotide repeat lengths are 19nt and 20nt that are present only in *Per2* and *Per3* genes, respectively. Maximum length of hexa-nucleotide repeats is 24nt, 18nt, and 25nt in *Per1*, *Per2,* and *Per3* genes, respectively [Figure 20].

The dinucleotide repeats were most abundant as many studies suggested previously that the most occurring SSRs are dinucleotides in most species [35,61].

3.8. SSR Length in Exons and Introns of *Per* Genes

Longest SSR in exons is 28nt that is present only in human *Per2*. Long SSRs are not very common in introns of *Per1* where maximum SSR length is 28nt present in mouse gene. Maximum SSR length in *Per2*
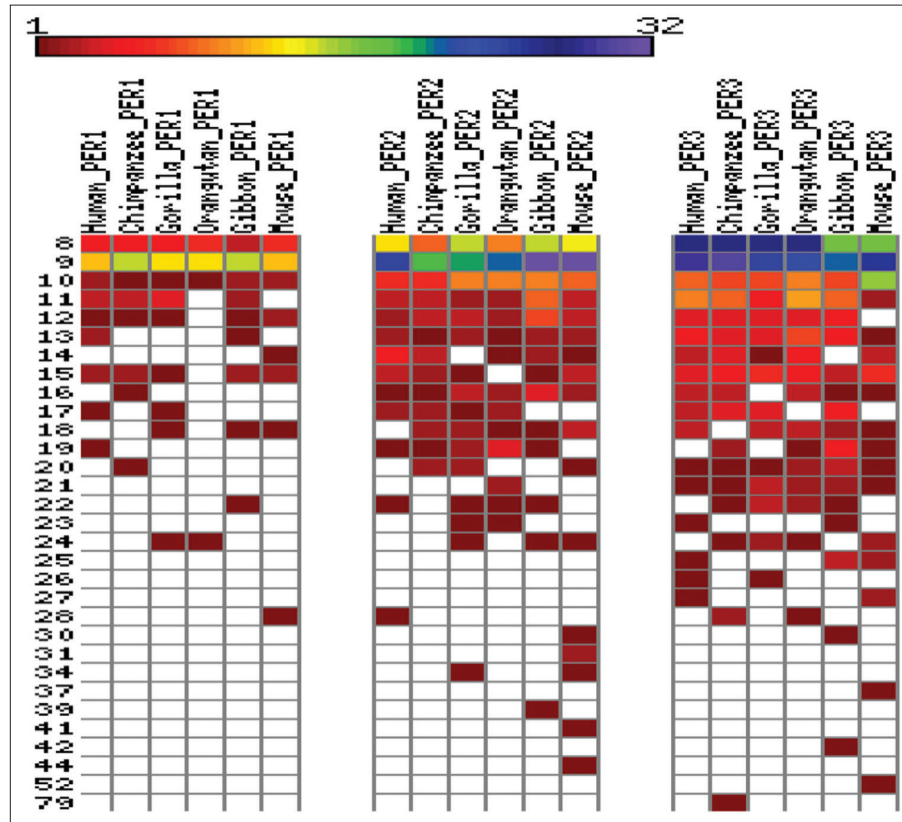
Figure 19: Simple sequence repeat length (nucleotides) in *Per* genes of human and its orthologues. Blank cells indicate absence of motifs. Legend color intensity indicates density per kbp of motifs.
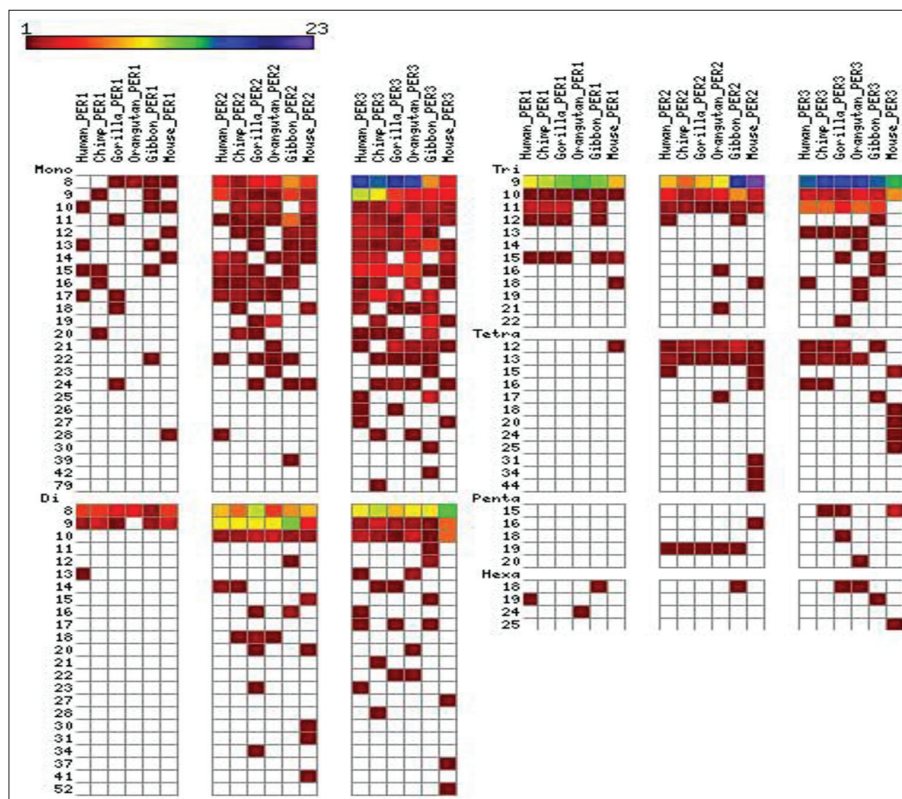
Figure 20: Mono- to hexa-nucleotide repeat lengths (nucleotides) in *Per* genes of human and its orthologues. Blank cells indicate absence of motifs. Legend color intensity indicates density per kbp of motifs.

intron is 44nt (mouse) and in *Per3* intron, it is 79nt (chimpanzee) [Figure 21]. Different SSR types are abundant in introns of *Per* genes with exceptions [Figures 22 and 23].

4. CONCLUSION

Comparison of all *Per* genes shows higher density of SSRs in *Per1* genes of all organisms with exceptions followed by *Per3* and *Per2* genes. This indicates that possibly perfect SSRs are unlikely to be the source of the polymorphism in *Per3* gene.

Among repeat types, trinucleotides, dinucleotides, and mononucleotides are the most abundant repeat types in *Per1*, *Per2*, and *Per3* with exceptions. Other repeat types are the least common. Many mono- to hexa-nucleotide repeat motifs show gene- and species-specific distribution. There is preference of some repeat types over others for each *Per* gene that could be due to their functional roles.

Exons and introns of the three *Per* genes also show diverse patterns of SSR distribution and density. The repeat types and motifs also have diverse distribution. This shows not only gene- and species-specific distribution of repeats but also region-specific distribution of repeat types and repeat motifs in exons and introns of the *Per* genes.

Distribution of CG-rich repeats is also gene- and species-specific. Similarly, there is diverse distribution of CG-rich repeat types and motifs in the three *Per* genes and their exons and introns in different organisms in the present study. Besides this, exons of *Per* genes generally do not have abundance of CG-rich repeats which indicates possible selection against CG-rich repeats in exons of the *Per* genes with exceptions.

Long perfect SSRs are generally not common in the three *Per* genes. Among repeat types, mono- and di- nucleotide repeats are longer in comparison with other repeats. Unlike introns, exons of all three *Per* genes generally do not have long repeats.
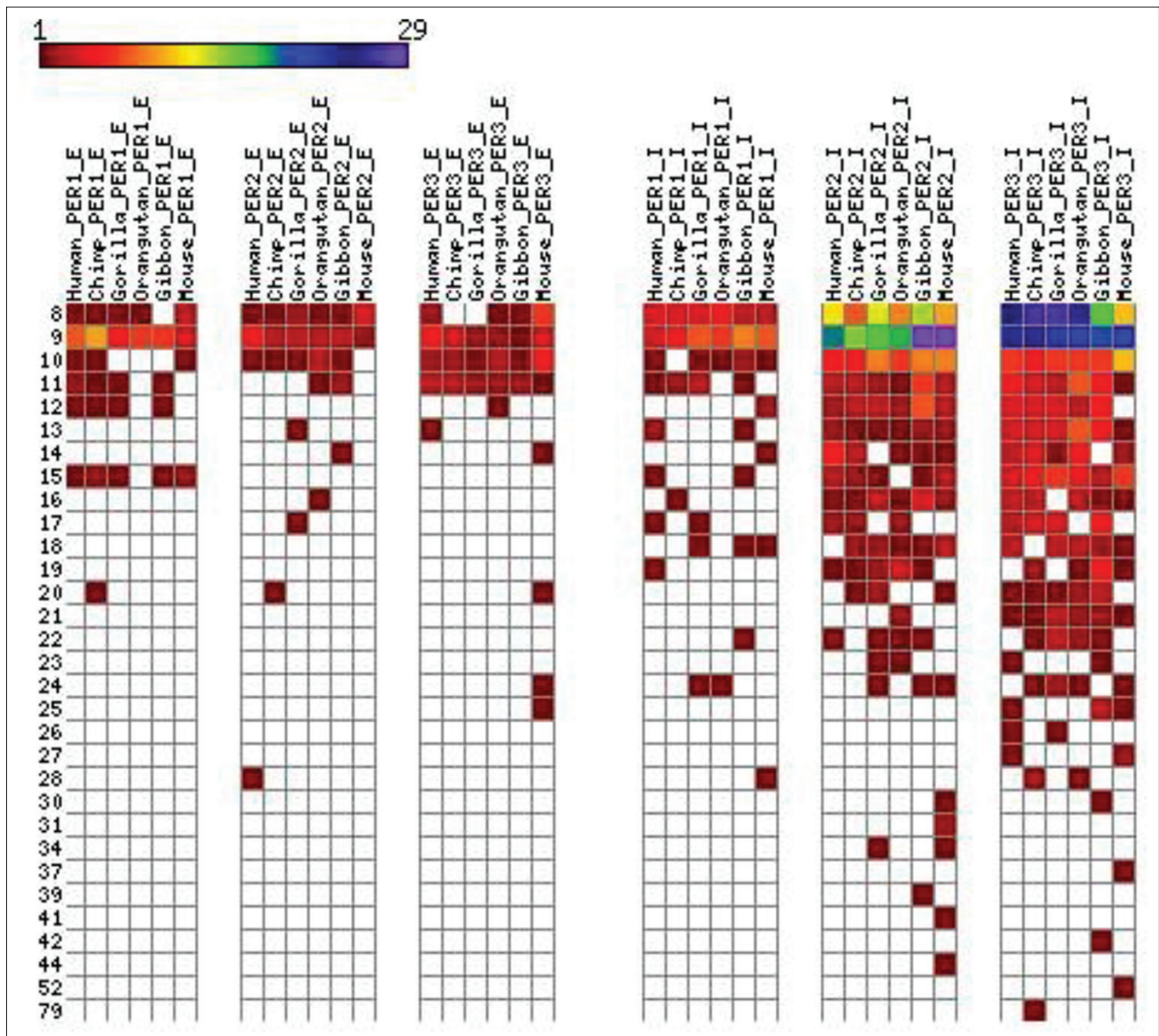
Figure 21: Lengths (nucleotides) of simple sequence repeats in exons (E) and introns (I) in *Per* genes of human and its orthologues. Blank cells indicate absence of motifs. Legend color intensity indicates density per kbp of motifs.
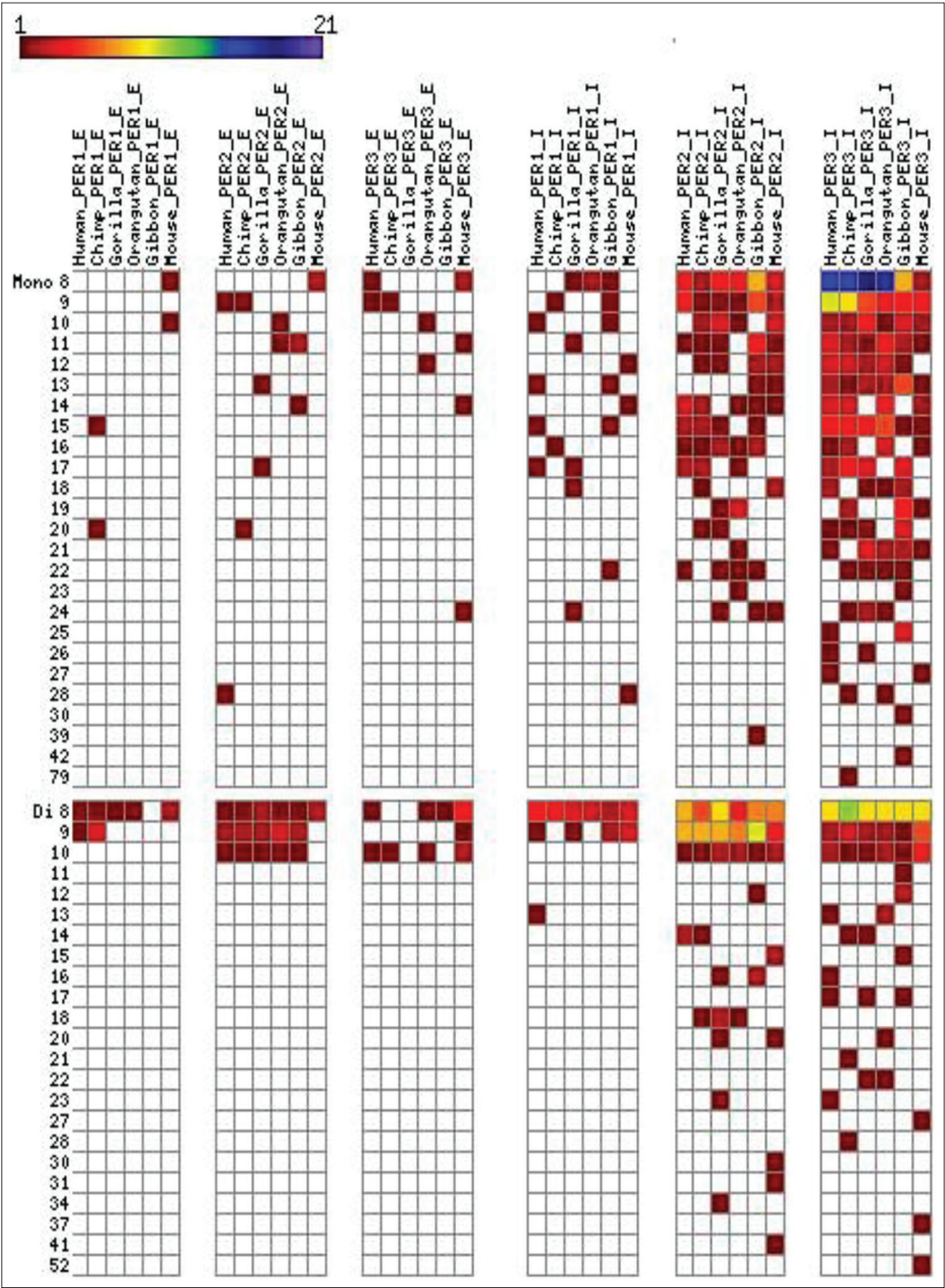
Figure 22: Lengths (nucleotides) of mono- and di- nucleotide repeats in exons (E) and introns (I) in *Per* genes of human and its orthologues. Blank cells indicate absence of motifs. Legend color intensity indicates density per kbp of motifs.

The SSRs found in the present study do show gene and species-specific variations but are present in all three *Per* genes of all organisms. Further, study of SSRs especially found in exons of human *Per* genes with reference to chronotype preference, physiological disorders and sleep disorders need to be taken forward since perfect SSRs are predisposed to mutations [34] compared with imperfect SSRs. The SSRs found in the human *Per* genes could serve as markers for investigators to study role of such SSRs in *Per* related disorders, if any.
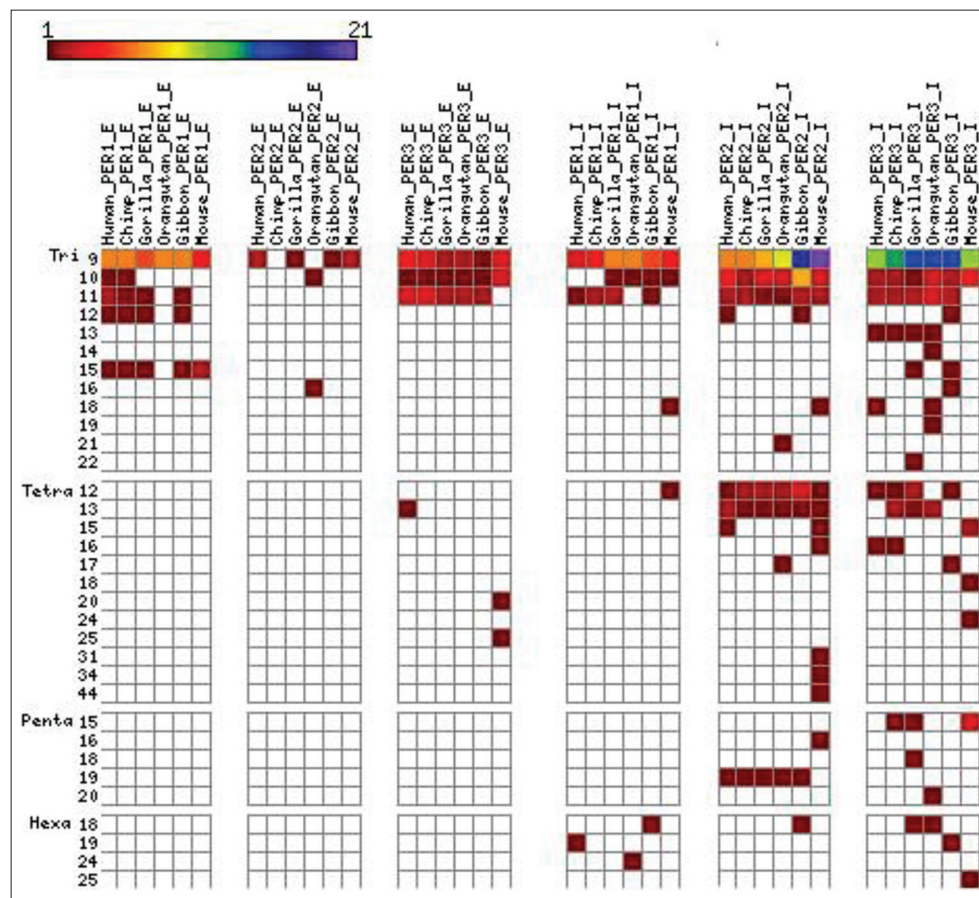
5. ACKNOWLEDGMENT

6. AUTHOR CONTRIBUTIONS

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in

Figure 23: Lengths (nucleotides) of tri- to hexa-nucleotide repeats in exons (E) and introns (I) in *Per* genes of human and its orthologues. Blank cells indicate absence of motifs. Legend color intensity indicates density per kbp of motifs.

drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; and agree to be accountable for all aspects of the work. All the authors are eligible to be an author as per the international committee of medical journal editors (ICMJE) requirements/guidelines.

7. CONFLICTS OF INTEREST

The authors report no financial or any other conflicts of interest in this work.

8. ETHICAL APPROVALS

Not applicable.

9. PUBLISHER'S NOTE

This journal remains neutral with regard to jurisdictional claims in published institutional affiliation.

REFERENCES

1. Kumar V, Sharma A. Common features of circadian timekeeping in diverse organisms. Curr Opin Physiol 2018;5:58-67.
2. Tataroglu O, Emery P. Studying circadian rhythms in *Drosophila melanogaster*. Methods 2014;68:140-50.
3. Bell-Pedersen D, Cassone VM, Earnest DJ, Golden SS, Hardin PE, Thomas TL, *et al*. Circadian rhythms from multiple oscillators: Lessons from diverse organisms. Nat Rev Genet 2005;6:544-56.
4. Lowrey P, Takahashi J. Genetics of circadian rhythms in mammalian model organisms. Adv Genet 2011;74:175-230.
5. Mohawk J, Green C, Takahashi J. Central and peripheral circadian clocks in mammals. Annu Rev Neurosci 2012;35:445-62.
6. Ko C, Takahashi J. Molecular components of the mammalian circadian clock. Hum Mol Genet 2006;15:R271-7.
7. Maury E. Off the clock: From circadian disruption to metabolic disease. Int J Mol Sci 2019;20:1597.
8. Papagiannakopoulos T, Bauer M, Davidson S, Heimann M, Subbaraj L, Bhutkar A, *et al*. Circadian rhythm disruption promotes lung tumorigenesis. Cell Metab 2016;24:324-31.
9. Bailey M, Silver R. Sex differences in circadian timing systems: Implications for disease. Front Neuroendocrinol 2014;35:111-39.
10. Allen GC, Farnell Y, Bell-Pedersen D, Cassone VM, Earnest DJ. Effects of altered clock gene expression on the pacemaker properties of SCN2.2 cells and oscillatory properties of NIH/3T3 cells. Neuroscience 2004;127:989-99.
11. Albrecht U. Invited review: Regulation of mammalian circadian clock genes. J Appl Physiol (1985) 2002;92:1348-55.
12. Tei H, Okamura H, Shigeyoshi Y, Fukuhara C, Ozawa R, Hirose M, *et al*. Circadian oscillation of a mammalian homologue of the *Drosophila* period gene. Nature 1997;389:512-6.
13. Blau J. PERspective on PER phosphorylation. Genes Dev 2008;22:1737-40.
14. Clayton J, Kyriacou C, Reppert S. Keeping time with the human genome. Nature 2001;409:829-31.
15. Albrecht U, Bordon A, Schmutz I, Ripperger J. The multiple facets of per2. Cold Spring Harb Symp Quant Biol 2007;72:95-104.
16. Chen-Goodspeed M, Lee C. Tumor suppression and circadian

function. J Biol Rhythms 2007;22:291-8.

17. Archer S, Schmidt C, Vandewalle G, Dijk D. Phenotyping of PER3 variants reveals widespread effects on circadian preference, sleep regulation, and health. Sleep Med Rev 2018;40:109-26.

18. von Schantz M, Jenkins A, Archer S. Evolutionary history of the vertebrate period genes. J Mol Evol 2006;62:701-7.

19. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, *et al*. Ensembl 2019. Nucleic Acids Res 2019;47:D745-51.

20. Sun Z, Albrecht U, Zhuchenko O, Bailey J, Eichele G, Lee C. RIGUI, a putative mammalian ortholog of the *Drosophila* period gene. Cell 1997;90:1003-11.

21. Bae K, Jin X, Maywood E, Hastings M, Reppert S, Weaver D. Differential functions of mPer1, mPer2, and mPer3 in the SCN circadian clock. Neuron 2001;30:525-36.

22. Shearman L, Zylka M, Weaver D, Kolakowski L, Reppert S. Two period homologs: Circadian expression and photic regulation in the suprachiasmatic nuclei. Neuron 1997;19:1261-9.

23. Oster H, Maronde E, Albrecht U. The circadian clock as a molecular calendar. Chronobiol Int 2002;19:507-16.

24. Spoelstra K, Daan S. Effects of constant light on circadian rhythmicity in mice lacking functional cry genes: Dissimilar from per mutants. J Comp Physiol A Neuroethol Sens Neural Behav Physiol 2008;194:235-42.

25. Yamaguchi Y, Okada K, Mizuno T, Ota T, Yamada H, Doi M, *et al*. Real-time recording of circadian per1 and per2 expression in the suprachiasmatic nucleus of freely moving rats. J Biol Rhythms 2016;31:108-11.

26. Takumi T, Matsubara C, Shigeyoshi Y, Taguchi K, Yagita K, Maebayashi Y, *et al*. A new mammalian period gene predominantly expressed in the suprachiasmatic nucleus. Genes Cells 1998;3:167-76.

27. Chang A, Duffy J, Buxton O, Lane JM, Aeschbach D, Anderson C, *et al*. Chronotype genetic variant in PER2 is associated with intrinsic circadian period in humans. Sci Rep 2019;9:5350.

28. Takumi T, Taguchi K, Miyake S, Sakakida Y, Takashima N, Matsubara C, *et al*. A light-independent oscillatory gene mPer3 in mouse SCN and OVLT. EMBO J 1998;17:4753-9.

29. Matsumura R, Akashi M. Role of the clock gene Period3 in the human cell-autonomous circadian clock. Genes Cells 2019;24:162-71.

30. Zylka M, Shearman L, Weaver D, Reppert S. Three period homologs in mammals: Differential light responses in the suprachiasmatic circadian clock and oscillating transcripts outside of brain. Neuron 1998;20:1103-10.

31. Gotter A, Reppert S. Analysis of human Per4. Brain research. Mol Brain Res 2001;92:19-26.

32. Rodriguez CM, Todd PK. New pathologic mechanisms in nucleotide repeat expansion disorders. Neurobiol Dis 2019;130:104515.

33. Chambers G, MacAvoy E. Microsatellites: Consensus and controversy. Comp Biochem Physiol B Biochem Mol Biol 2000;126:455-76.

34. Kashi Y, King D. Simple sequence repeats as advantageous mutators in evolution. Trends Genet 2006;22:253-9.

35. Ellegren H. Microsatellites: Simple sequences with complex evolution. Nat Rev Genet 2004;5:435-45.

36. Srivastava S, Avvaru A, Sowpati D, Mishra R. Patterns of microsatellite distribution across eukaryotic genomes. BMC Genomics 2019;20:153.

37. Abdurakhmonov I. Genomics era for plants and crop species-advances made and needed tasks ahead. In: Plant Genomics. London: InTech; 2016.

38. Ananda G, Walsh E, Jacob KD, Krasilnikova M, Eckert KA, Chiaromonte F, *et al*. Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. Genome Biol Evol 2013;5:606-20.

39. Senan S, Kizhakayil D, Sasikumar B. Methods for development of microsatellite markers: An overview. Not Sci Biol 2014;6:1-13.

40. Vieira M, Santini L, Diniz A, Munhoz C. Microsatellite markers:

What they mean and why they are so useful. Genet Mol Biol 2016;39:312-28.

41. Wheeler J, Bodmer W, Mortensen N. DNA mismatch repair genes and colorectal cancer. Gut 2000;47:148-53.

42. Trifonov E. Tuning function of tandemly repeating sequences: A molecular device for fast adaptation. In: Evolutionary Theory and Processes. Dordrecht, Netherlands: Springer; 2004. p. 115-38.

43. Bacolla A, Wells R. Non-B DNA conformations as determinants of mutagenesis and human disease. Mol Carcinog 2009;48:273-85.

44. Pin-On P, Aporntewan C, Siriluksana J, Bhummaphan N, Chanvorachote P, Mutirangura A. Targeting high transcriptional control activity of long mononucleotide A-T repeats in cancer by Argonaute 1. Gene 2019;699:54-61.

45. Vinces M, Legendre M, Caldara M, Hagihara M, Verstrepen K. Unstable tandem repeats in promoters confer transcriptional evolvability. Science 2009;324:1213-6.

46. Peixoto A. Evolutionary behavioral genetics in *Drosophila*. Adv Genet 2002;47:117-50.

47. Zavodna M, Bagshaw A, Brauning R, Gemmell N. The effects of transcription and recombination on mutational dynamics of short tandem repeats. Nucleic Acids Res 2018;46:1321-30.

48. Legendre M, Pochet N, Pak T, Verstrepen K. Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res 2007;17:1787-96.

49. O'Leary N, Wright M, Brister J, Ciufo S, Haddad D, McVeigh R, *et al*. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44:D733-45.

50. Kinsella R, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, *et al*. Ensembl BioMarts: A hub for data retrieval across taxonomic space. Database (Oxford) 2011;2011:bar030.

51. Kofler R, Schlötterer C, Lelley T. SciRoKo: A new tool for whole genome microsatellite search and investigation. Bioinformatics 2007;23:1683-5.

52. Pavlidis P, Noble W. Matrix2png: A utility for visualizing matrix data. Bioinformatics 2003;19:295-6.

53. Li B, Xia Q, Lu C, Zhou Z, Xiang Z. Analysis on frequency and density of microsatellites in coding sequences of several eukaryotic genomes. Genomics Proteomics Bioinformatics 2004;2:24-31.

54. Tian X, Strassmann J, Queller D. Genome nucleotide composition shapes variation in simple sequence repeats. Mol Biol Evol 2011;28:899-909.

55. Zheng B, Albrecht U, Kaasik K, Sage M, Lu W, Vaishnav S, *et al*. Nonredundant roles of the mPer1 and mPer2 genes in the mammalian circadian clock. Cell 2001;105:683-94.

56. Metzgar D, Bytof J, Wills C. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 2000;10:72-80.

57. Gu T, Tan S, Gou X, Araki H, Tian D. Avoidance of long mononucleotide repeats in codon pair usage. Genetics 2010;186:1077-84.

58. Trivedi S. Microsatellites (SSRs): Puzzles within puzzle. Indian J Biotechnol 2004;3:331-47.

59. Li Y, Korol A, Fahima T, Nevo E. Microsatellites within genes: Structure, function, and evolution. Mol Biol Evol 2004;21:991-1007.

60. 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;512:331-6.

61. Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 2000;156:847-54.