Advance Publication

The effect of perfection status on mutation rates of microsatellites in primates

Ming Yin NGAI[1,2], Naruya SAITOU[2,1]*

[1]*Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan*
[2]*Division of Population Genetics, National Institute of Genetics, Mishima 411-8540, Japan*

Received 29 June 2015; accepted 24 January 2016

Abstract Microsatellites are highly mutable elements in eukaryotic genomes. Although they are widely used by researchers as genetic markers and tools for evolutionary studies, the full mutation mechanisms and factors affecting the mutation rate of microsatellites are not well understood. Microsatellite loci shared by human, chimpanzee, gorilla, orangutan and rhesus macaque genomes were sorted into four different perfection status groups named Perfect, Imperfect, Perfect-Compound, and Imperfect-Compound. We found that compound microsatellite loci generally had no significant effect on mutation rates, while imperfect microsatellite loci significantly lowered mutation rates compared to perfect ones ($P = 1 \times 10^{-4}$). The significant difference resulted from a small amount of interruption (1–2 bp) within microsatellites, especially when the loci were shorter than 14–15 repeats. Furthermore, real perfect loci were used to compare with split up 'perfect loci' actually obtained from the imperfect group. We found that the mutation rates were significantly different from each other for small numbers of repeats, especially in 7–9 repeats ($P < 0.05$). This suggests that an imperfect locus should not be considered as two or more separated perfect loci. This also raises the question that the algorithms currently used to find microsatellite loci based on mismatch penalties may result not only in heterogeneous microsatellite data sets with heterogeneous numbers of imperfect loci but also in data sets that are biased in mutation rate because of these imperfect loci.

Key words: microsatellites, mutation rate, primates

Introduction

Microsatellites, also called short tandem repeats (STRs), are highly mutable repetitive short sequences 1–6 bp in size, and are abundant in the genome of eukaryotic organisms (Ellegren, 2004). Microsatellite instability is directly related to human cancers (Chung et al., 2010; Lacroix-Triki et al., 2011). Microsatellites are also widely used as genetic markers in forensic science (Song et al., 2010), and are often used in population genetic studies (e.g. Li et al., 2006; Sun et al., 2009). Many studies of microsatellites at the molecular level have already been published, focusing especially on their mutation mechanism (e.g. Kofler et al., 2008; Boyer et al., 2008; Amos, 2010). Slippage is commonly accepted to be the major mutation mechanism of microsatellites (Ellegren, 2004; Bhargava and Fuentes, 2010), when their repeat number exceeds the mutation threshold, which is nine repeats (rp) for mononucleotide microsatellites, and four repeats for other motif sizes (Lai and Sun, 2003; Kelkar et al., 2010). Factors affecting microsatellite slippage include repeat number, motif size, motif structure, chromosome type, and genomic location (Bhargava and Fuentes, 2010). Among these factors, the repeat number is the strongest factor positively correlated to mutation rates of microsatellites (Kelkar et al., 2008).

Perfection status of microsatellites

In addition to the factors mentioned above, the perfection of microsatellites also affects their mutation rates. The perfection status of microsatellites was first described by Oliveira et al. (2006). A perfect microsatellite refers to a repetitive sequence purely composed of one type of motif. An imperfect microsatellite refers to a locus having at least one base pair that does not match the repetitive sequence. Interruption refers to a short sequence within the repetitive sequences, while a composite (also called compound) microsatellite refers to two distinctive, consecutive repetitive sequences that are linked. However, this categorization is somehow idealistic: many microsatellites do not fall into any of the four categories in real cases. Algorithms used to search for microsatellites, such as Sputnik, RepeatMasker, and Tandem Repeat Finder (TRF), use different parameters, with major differences in the way they deal with mismatches (interruptions), resulting in non-uniform data sets (Leclercq et al., 2007). Interruptions inside microsatellites are long

* Correspondence to: Naruya Saitou, Division of Population Genetics, National Institute of Genetics, 1111 Yata, Mishima 411-8540, Japan.
E-mail: saitounr@nig.ac.jp

Published online 31 May 2016
in J-STAGE (www.jstage.jst.go.jp) DOI: 10.1537/ase.160124

known to have a stabilizing effect causing the mutation rate to be greatly lowered. However, direct measurement of mutation rates and a comparison between imperfect microsatellites and perfect microsatellites were only recently analyzed by Boyer et al. (2008). They proposed that interruptions could break an imperfect microsatellite into two short perfect microsatellites by an observed mutational behavioral change.

Different algorithms generate heterogeneous data sets, and researchers make use of these data sets for their own studies. This makes it impossible to compare results, especially for studies that depend on loci content and number (Almeida and Penha-Goncalves, 2004; Galindo et al., 2009). Although a mutational behavioral change of interrupted microsatellites was observed, its interrelationship with the degree of mutations and repeat number, etc., at the genomic level still remains to be explored.

The main aim of this study is to distinguish microsatellites according to their perfection characteristics. Microsatellites located on the non-coding regions of human chromosome 21 were compared with those in other primate species, including chimpanzee, gorilla, orangutan, and rhesus macaque, and were categorized into four perfection status groups, Perfect, Imperfect, Perfect-Compound, and Imperfect-Compound, following the new definitions proposed in this study.

Mutation rates were estimated in order to demonstrate if there was any significant difference among different perfection status groups. Further investigation was done to investigate the existence of interactions within the loci of imperfect microsatellites, which in turn provides evidence for the importance of microsatellite perfection status categorization.

Materials and Methods

Microsatellites and genomic data

Takezaki and Nei (2009) used human and chimpanzee chromosome 21 genomic sequence data, and microsatellite coordinates of intronic and intergenic regions data were kindly provided by Professor Takezaki Naoko of Kagawa University. Coordinates correspond to those of UCSC Genome Browser (www.genome.ucsc.edu). The human and chimpanzee genome builds used were hg18 and panTro2, respectively. Perfect microsatellite loci of either human or chimpanzee with at least six repeats were already selected in the file.

Chromosomes orthologous to the human chromosome 21 were examined using Ensembl Genome Browser. Therefore, DNA sequences of chromosome 21 for human (GRCh37.57), chimpanzee (CHIMP2.1.57), gorilla (gorGor3.57), orangutan (PPYG2.57), and chromosome 3 for rhesus macaque (MMUL_1.57) were downloaded in FASTA format from the Ensembl Genome Browser (http://uswest.ensembl.org/info/data/ftp/index.html). Chromosome data of human chromosome 21 (version hg18) were also downloaded from the UCSC database.

Orthologous loci shared among primates

Human–chimpanzee orthologous microsatellite loci together with 100 bp flanking regions were extracted from human chromosome 21 data available at the UCSC database. Extracted loci from the raw file were sorted by their motif size into four groups: dinucleotide (2905 loci), trinucleotide (188 loci), tetranucleotide (284 loci), and pentanucleotide (26 loci). There was no hexanucleotide in the data file. Only dinucleotide loci were used in this study because they were the most abundant.

The above-mentioned extracted regions were used as queries to search for orthologous regions in the other three primate species using BLASTN version 2.2.22 downloaded from the NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.22/) (Altschul et al., 1997). The strand used is decided according to the synteny results of Ensembl. The filtering query sequence option was turned off to prevent loci being filtered out. To account for the large number of gaps in microsatellites, the open gap penalty (parameter G) and the gap extension cost (parameter E) were each set to the lowest value allowed (G = 1, E = 1), which corresponds to the default mismatch penalty (parameter q) and match reward (parameter r) values. An E-value of E–10 was used (following Zhang et al., 2006), and the best hits in each primate species were extracted based on the blast scores.

Full multiple alignments were conducted using ClustalW2 version 2.0.12 (http://www.clustal.org/download/) (Larkin et al., 2007). In addition to usual indel mutations, gap opening is expected for microsatellites because of their variable sizes. Therefore the gap open penalty is increased to 4 while the gap extension penalty was decreased to 4, compared to default values. This allows the aligner to open a gap for at least two nucleotides, which is the minimum motif size of a dinucleotide microsatellite. Since a microsatellite can be long and simple, especially when it is in the SINE region, the delay divergence sequence switch was set to 95% identity so as to obtain high score sequences by giving priority to align to the human sequence first.

Only non-overlapping regions are used in this study to prevent double-counting and rate estimation for loci that are very close to each other, which might be considered as one imperfect locus. The latter issue will be explained in later sections.

Microsatellite perfectness and repeat counting

Microsatellites were categorized into four groups in this study: Perfect, Imperfect, Perfect-Compound, and Imperfect-Compound. A perfect microsatellite is defined as a locus with a perfect repetitive run with its own motif type, abbreviated as the locus motif (LM) (Figure 1a). An imperfect microsatellite is defined as a locus with a repetitive run that contains interruptions. Each interruption is 1–9 bp long (Figure 1b). When the interruption is more than 10 bp, it is considered as two perfect loci. Besides perfect and imperfect, a locus could also be either compound or non-compound. A compound microsatellite is defined as a locus which contains a repetitive sequence composed of a non-locus motif, abbreviated as nLM, where the repeat number passes the threshold value, and is within 10 bp flanking region of the locus (Figure 1c, d).

To obtain the repeat number in imperfect microsatellites, each locus run is first located by using the coordinates provided in the raw data, which is a perfect run of six repeats for

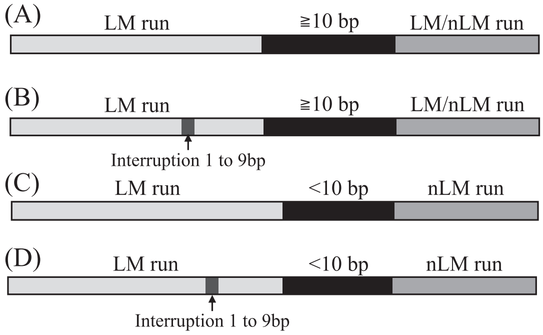

Figure 1. Illustrations of the four perfection status categorization groups. LM and nLM refers to locus motif and non-locus motif, respectively. 'LM run' region was used for mutation rate estimation. 'LM/nLM run' represents neighboring repetitive sequences. Black bars represent unique sequences. Each locus should be either compound or non-compound, either perfect or imperfect: (A) Perfect (perfect, non-compound) locus; (B) Imperfect (imperfect, non-compound) locus; (C) Perfect-Compound (perfect, compound) locus; and (D) Imperfect-Compound (imperfect, compound) locus.

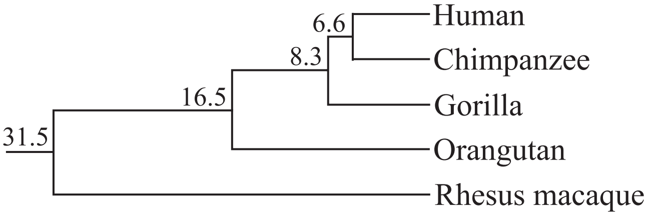

Figure 2. Divergence times (in million years ago) of primate species used for mutation rate estimation. Divergence times are from Perelman et al. (2011).

human and chimpanzee according to the threshold value applied by Takezaki and Nei (2009), and of two repeats (which infer no threshold value) for the other three primate species. If a sequence flanking the stated locus contains a LM run with at least two repeats in the other three species, and provided the interruption is less than 10 bp, the repeat number will be increased according to the size of extra LMs found, and newly considered as an imperfect locus. This process is repeated until either a non-LM run which passes the threshold value, or a unique sequence (≥10 bp) is found. Considering that there will be a chance that the interruption is due to a mutation occurring on a repeat, one repeat was compensated for the locus repeat number while the interruption size is ≤1 LM size.

Categorization was done on each locus for all five primate species. When locus perfection status among species was different, that locus was discarded in this study, and loci were binned according to the average repeat number of species considered (i.e. human and chimpanzee in stepwise mutation model, and all five species in total divergence time).

Slippage rate calculation

Slippage rates of microsatellite loci were estimated in this study by two methods: the stepwise mutation model (SMM) (Ohta and Kimura, 1973), and the total divergence time (TDT) method. Both methods are based on the counted repeat number (see the previous section). As there will be a chance that there is no orthologous microsatellite locus (or locus with less than two repeats) for a target primate species, that species will not be included in the TDT method for that locus. This step is introduced to eliminate seriously misaligned loci, and more importantly the sudden appearance or disappearance of microsatellite loci as they are sometimes sites for transposition (Kelkar et al., 2008), a case which is unrelated to slippage rate. Human and chimpanzee microsatellites were used in SMM as all the loci are orthologous between these two species.

Slippage rate is estimated by using the SMM, assuming the populations used are at mutation drift equilibrium. The $(\delta\mu)^2$ between the two populations defined by Goldstein et al. (1995) is:

$$(\delta\mu)^2 = (\mu_A - \mu_B)^2, \tag{1}$$

where μ_A and μ_B are the mean allele length in species (or population) A and B, respectively. By assuming mutation is non-directional and considering microsatellites evolve in a single-step manner, Goldstein et al. (1995) showed that:

$$(\delta\mu)^2 = 2vt, \tag{2}$$

where v and t are mutation rate (slippage rate in this study) and divergence time, respectively.

Slippage mutation rates can be estimated by counting the total slippage events divided by the total divergence time. The divergence time of primates from human is based on Perelman et al. (2011), as shown in Figure 2. The minimum number of slippage changes along the primate lineage evolution is counted, and then divided by the total divergence time. For instance, the repeat number counted in a locus is 6, 5, 6, 5, 5 repeats in human, chimpanzee, gorilla, orangutan, and rhesus macaque, respectively; the minimum change will be 2, i.e. an expansion of 1 in the common ancestor of human, chimpanzee, and gorilla, and a contraction of 1 in the chimpanzee lineage. An alternative explanation is two independent expansions—one in gorilla and one in human lineages. This new method is called the 'total divergence time' (TDT) method in this study.

Results

Basic statistics

We extracted 2905 regions containing dinucleotide microsatellite loci in human and chimpanzee from the raw data file. Among these, 2691 had flanking regions shared among five primate species and 2173 of them are non-overlapping. Eleven of the non-overlapping loci were discarded because of bad alignment, and 1766 of them contain target microsatellites in all five primate species. The number of loci and their relative proportions are presented in Figure 3. For imperfect and imperfect-compound loci, the minimum repeat number is 4, because of the minimum 2 rp as search starting point plus 2 rp extra LM run with interruption in between. For instance, $(AC)_2i(AC)_2$, where i denotes interruption.

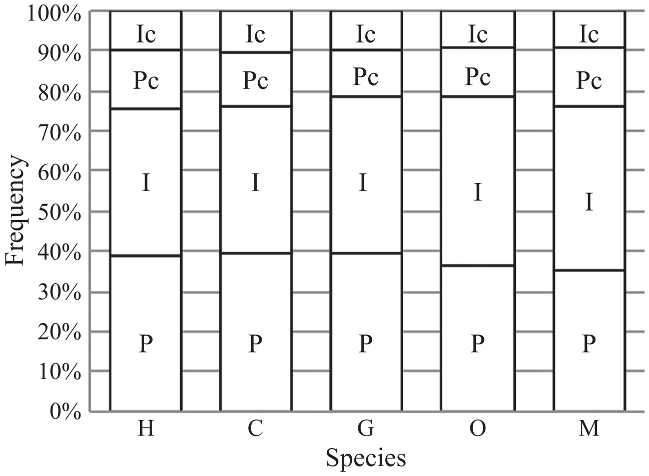

Figure 3. General statistics of dinucleotide loci extracted from genome data of five primate species. Ic, Pc, I, and P designate Imperfect-Compound, Perfect-Compound, Imperfect, and Perfect, respectively.

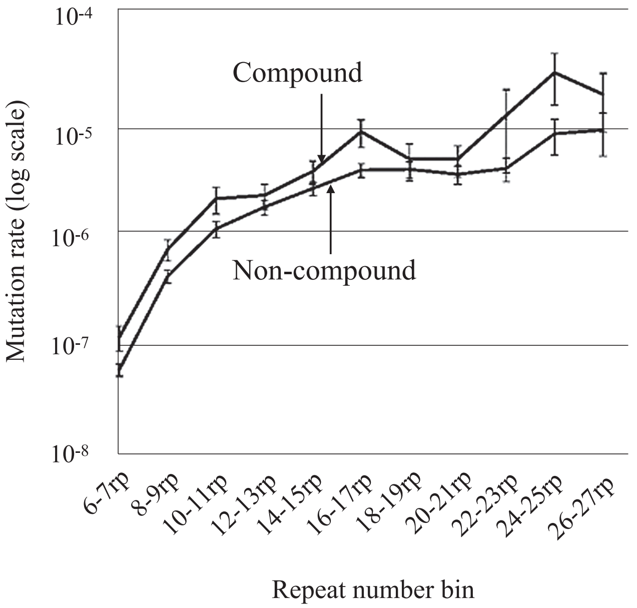

Figure 4. Comparison between compound and non-compound microsatellite loci. Rates are estimated by using SMM for human and chimpanzee data. Compound loci used in this graph are with non-LM run immediately adjacent to LM run.

Compound vs. non-compound microsatellites

The mutation rate profile of compound (immediately next to the LM run) and non-compound (without non-LM run within the 10 bp flanking region) dinucleotide microsatellites estimated by SMM, sorted by mean repeat number, are shown in Figure 4. Compound microsatellites have elevated mutation rates, and some of the data points (6–7 rp, 8–9 rp, 10–11 rp, 16–17 rp, 24–25 rp) even show statistically significant differences. This implies that certain interactions might exist between the LM run and the non-LM run, lowering the stability of the microsatellite. However, the general trend by analysis of covariance (ANCOVA) did not show significant results where closely located loci are expected to interact with each other and affect microsatellite mutation. This phenomenon might be explained by a relatively short non-LM run (from four repeats) which is considered as a compound microsatellite data set in this study. These non-LM runs may not be able to bring about significant rate elevation when the LM runs becomes relatively long. This explanation is supported by the observation that the difference is statistically significant mainly at shorter repeats (6–11 rp) but not at longer repeats, i.e. the relative effect of the non-LM run decreases when the LM run becomes relatively longer.

Perfect vs. imperfect microsatellites

A general picture showing how the mutation rates of imperfect microsatellites are different from those of perfect ones is shown in Figure 5. Microsatellite perfection status for each primate species was determined (see Materials and methods), and only non-compound microsatellites are considered to illustrate the difference between perfect and imperfect loci. The microsatellites were grouped according to the number of species that were of the same status. A gradual decrease in mutation rate estimated by the TDT method could be observed when the number of species that were imperfect increased. ANCOVA analysis on the difference between all perfect and imperfect species shows that the mutation rate is highly significantly different ($P = 1 \times 10^{-4}$). This agrees with previous studies proposing that interruptions have a stabilizing effect on microsatellites (Kunkel, 1985; Boyer et al., 2008).

Further analysis was done to investigate this phenomenon. To increase sample size for better resolution in repeat bins, only human and chimpanzee are considered. On the other hand, only non-compound microsatellites are used for perfect–imperfect separation to eliminate possible effects where an non-LM run may act on its LM run. A graph similar to Figure 5 is made using human and chimpanzee estimated by SMM, and is shown in Figure 6. 'Mixed' refers to non-compound microsatellites before separation according to the classification method as shown in Figure 1. The pattern of perfect repeats is highly significantly different from that of imperfect repeats ($P = 3 \times 10^{-4}$) by ANCOVA. This suggests that the effect of imperfection is much greater than that of the adjacently located non-LM run.

Our results show that the categorization algorithm of microsatellites in this study is able to categorize microsatellites that were claimed to be uninterrupted into significantly different groups, and demonstrate significant differences between perfect and imperfect groups.

Effect of interruption on mutation rate

Our results regarding the effect of imperfection agree with those of Boyer et al. (2008), whereby a small degree of interruption highly stabilizes microsatellites with a significantly lowered mutation rate. However, it is uncertain whether there is any interaction between two sides of a locus runinng across these interruptions. To investigate this, imperfect microsatellites are sorted by their proportion of interruption, at different bin lengths. The proportion of interruption for each locus is calculated, which is the ratio of the number of interrupting bases found within the LM run divided by the total

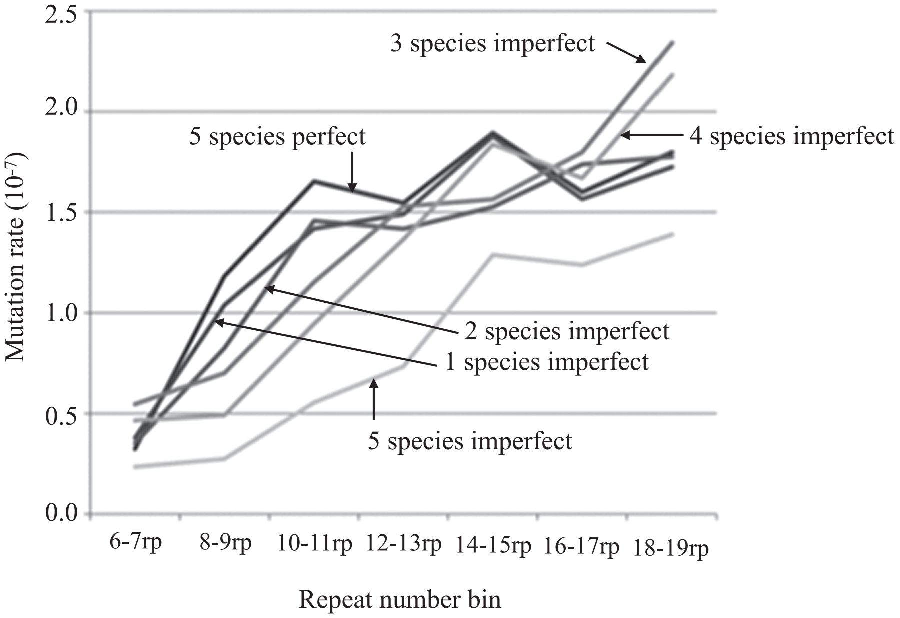

Figure 5. Comparison between perfect and imperfect microsatellites by using the TDT method. ANCOVA analysis on the difference between all species perfect and all species imperfect shows that the mutation rate is highly significantly different ($P = 1 \times 10^{-4}$).

number of bases of that locus. These proportions are then averaged for all loci in human and chimpanzee. Grubbs' (1969) test was used to examine the presence of outliers at each bin length, and significant outliers are marked with an asterisk in Table 1. The results show that the first data points (0% interruption, i.e. perfect) of 8–9 rp, 10–11 rp, and 12–13 rp are significant outliers ($P < 0.04$) (see Table 1), while there were no significant outliers for 16–17 rp, 18–19 rp, and 20–21 rp. There is no estimation for 14–15 rp because of this group was too small for Grubbs' test.

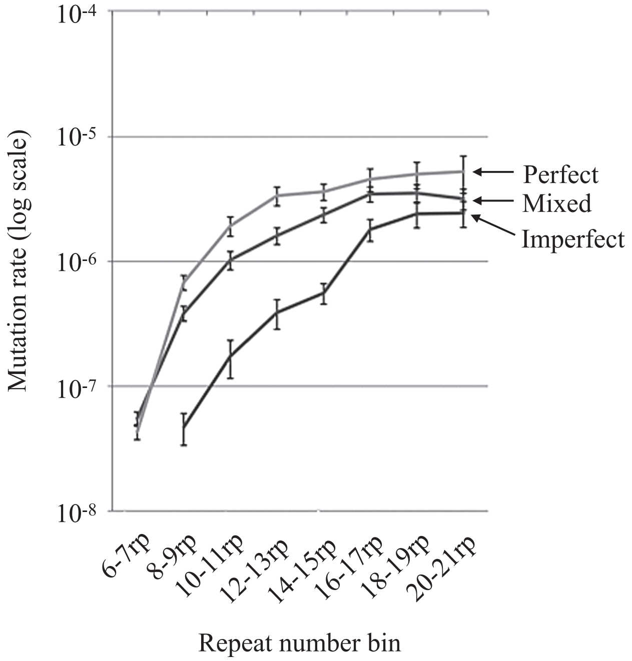

Figure 6. Comparison between perfect and imperfect microsatellites by using SMM for human and chimpanzee data. 'Mixed' refers to the result when perfect and imperfect loci are not separated (i.e. non-compound in Figure 4). Perfect loci are significantly different from imperfect loci ($P = 3 \times 10^{-4}$).

A comparison using perfect loci (0 bp interruption) with imperfect loci interrupted by one motif, i.e. 2 bp interruptions, is shown in Table 2. The ratio of the rate difference between perfect and imperfect loci is strongly negatively correlated with repeat number increase (Figure 7), as a small interruption could effectively lower the mutation rate of a microsatellite (Boyer et al., 2008). On the other hand, Student's *t*-test clearly shows that there is a sudden increase in *P*-value (loss of interruption effect) between the 14–15 rp and 16–17 rp bins (Figure 8), supported by a significant slope difference by regression analysis ($P = 0.04$) when comparing bin 14–15 rp against bin 16–17 rp (data not shown).

Figure 9 shows the comparison between perfect microsatellites vs. imperfect microsatellites with only one interruption site. Here the same pattern as previous analysis is observed where the difference between the two groups faded when the length increased for repeat numbers >15. All the above analyses show that interruptions are only one of the factors affecting mutation rates of microsatellites. This implies that an interaction exists between the two sides of imperfect loci. Careful attention should therefore be paid when we encounter imperfect microsatellites when searching for microsatellites.

Discussion

Microsatellites were categorized into four distinct groups and their mutation rates at different repeats were estimated, using intergenic and intronic dinucleotide microsatellites from primate species; the differences between perfect and

Table 1. *P*-values of Grubbs' test showing significant outliers for each graph in Figure 7

% Interruption	0%	1–5%	6–10%	11–15%	16–20%	21–25%	26–30%
8–9 rp	0.033*	0.335	0.308	0.484	0.338	0.364	0.345
10–11 rp	0.035*	0.293	0.340	0.417	0.299	0.455	0.308
12–13 rp	0.034*	0.331	0.288	0.415	0.466	0.387	0.296
14–15 rp				N/A			
16–17 rp	0.060	0.460	0.415	0.321	0.241	0.243	0.149
18–19 rp	0.091	0.491	0.304	0.234	0.283	0.253	0.106
20–21 rp	0.066	0.292	0.360	0.333	0.454	0.205	0.133

Asterisk indicates significant values ($P < 0.05$).

It shows that the 0% interruption (perfect microsatellites) in 8–9 rp, 10–11 rp, and 12–13 rp are significant outliers, indicating that the first short interruptions stabilized the loci. 14–15 rp was not analyzed because of insufficient data points (>6 points required for Grubbs' test).

Table 2. *P*-values of Student's *t*-test comparing loci having no interruption against up to 2 bp interruption and corresponding fold difference on mutation rate

	P-value[1]	Fold[2]
8–9 rp	4.53E–10*	28.68
10–11 rp	1.70E–06*	25.44
12–13 rp	1.61E–06*	18.90
14–15 rp	2.68E–06*	6.03
16–17 rp	1.29E–01	1.74
18–19 rp	1.24E–01	1.97
20–21 rp	8.24E–02	4.15

[1] Asterisk indicates significant values ($P < 0.05$).

[2] Fold is calculated from the 0 bp group divided by the 2 bp interruption group.

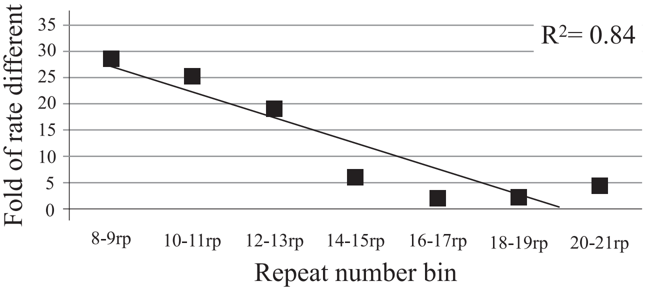

Figure 7. Correlation between ratio of rate difference and repeat numbers. Ratio of mutation rate differences between 0 rp (perfect) and 2 bp interrupted loci decrease when repeat number increases ($R^2 = 0.84$), indicating a negative correlation of stabilization power of interruptions against loci size.

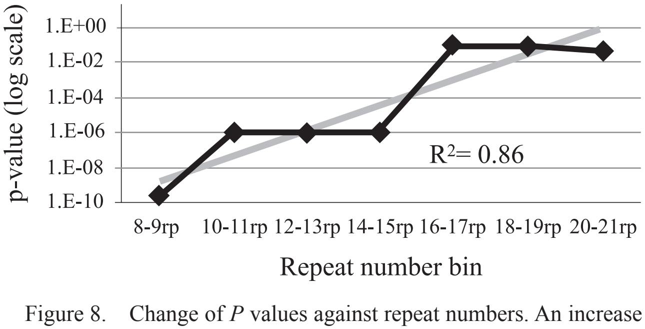

Figure 8. Change of *P* values against repeat numbers. An increase in log *P*-values refers to a decrease in stabilization power, and the distinct increase of P-value at 16–17 rp suggests some biological differences in addition to physical loci length difference.

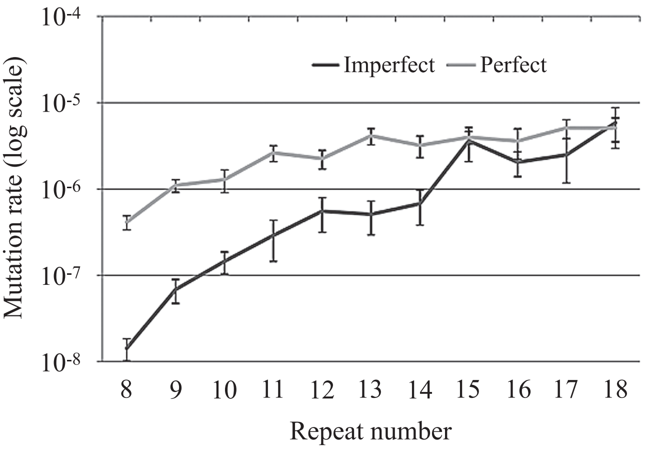

Figure 9. Comparison between perfect and imperfect microsatellites with one interruption site. Considering imperfect loci with only one interruption site, it is observed that two groups of microsatellites become not significantly different from each other; this is explained by the countering length factor on the microsatellites, and also shows the existence of an interaction between the two sides of imperfect loci.

imperfect microsatellites were found to be highly significant. This suggests there is a demand for microsatellite selection algorithms that are more advanced in handling interruptions, or pay attention to loci having closely located repetitive sequences, especially with the same motif.

The result of analyzing the effect of interruption on mutation rate suggested that the stabilizing effect of interruptions is particularly strong in shorter microsatellites (rp < 15), best illustrated by bin 8–9 rp: the first few base pairs of interruption led to the main decrease in mutation rates of these microsatellites. This stabilizing effect seems to fade out as the repeat number (opposite acting factor) increases. However, it is observed that the fade out was not gradual but occurred suddenly, between 14–15 rp and 16–17 rp. This phenomenon could be explained by a function of both physical and biological properties. Physically, it is found that the slippage process in one imperfect microsatellite is like two divided perfect loci when the interruption is introduced in the middle of a long microsatellite in vivo in a human mismatch repair defective cell line (Boyer et al., 2008). Although the position of interruption was not considered in the present study, assuming a random occurrence of point mutations on a locus, the overall effect of interruption should be like those that occurred in the middle of the loci, such that imperfect loci at group 16–17 rp here could be considered as two separate perfect 8 rp loci. Biologically, it was proposed in previous studies that exonucleolytic proofreading efficiency and mismatch repair efficiency sharply decrease when microsatellite loci length is longer than ~15–20 bp (8–10 rp in a dinucleotide motif) (Kroutil et al., 1996; Sia et al., 1997). Therefore imperfect loci at 14–15 rp seem to be largely stabilized compared to 16–17 rp.

Furthermore, the lost of significance from 15 rp shown in Figure 9 implies that interaction exists between microsatellites across interruptions. This in turn suggests that a locus should only be considered as a perfect when flanked by unique sequences. This is particularly important in short loci where the interruption effect is relatively strong. However, the exact value, or definition, of 'unique sequence' requires further investigation.

Inclusion of imperfect loci surely affects the locus number in the data set claimed to be perfect in this study, as in other related studies. Microsatellite searching programs, such as TRF, Sputnik, Mrep, etc., have different algorithms for selecting microsatellites, but the output microsatellite profiles differ considerably among algorithms (Leclercq et al., 2007). Parameters such as mismatch penalty are adjustable by the researcher. Some settings allow a small proportion of interruption but some will not. This causes a non-unified microsatellite data set. Almeida and Penha-Goncalves (2004) observed that there are large proportions of high identity long loci (in terms of percentage perfection) in vertebrate species. They suggested that the 'hump' from 10–22 rp in their data set could be due to an advanced mismatch repair system in vertebrates and backward mutation from long loci. However, microsatellite profiles given by Kelkar et al. (2008) showed that the same area (10–22 rp) was a plateau rather than a hump. They only used strictly perfect loci with 10 bp unique flanking sequences. Certainly, the mismatch repair system and backward mutation suggested by Almeida and Penha-Goncalves (2004) are reasonable explanations, but the results in this study implied that the data set structure might be one of the additional reasons. It is observed that the hump rises at ~11 rp and starts to drop from 16 rp. We suspect that the increase in the number of microsatellites from 10–15 rp forming the hump is a result of imperfect loci whose conservation may be more related to the interruption effect, and the number of loci drops after 16 rp where the interruption effect decreased.

Categorization is more appropriate when up to four levels of grouping were introduced—in order, these groupings are by motif size, perfection status, size, and finally by proportion of interruptions. However, sample sizes were sometimes insufficient in this study. Using whole genome data could hopefully solve this problem. Although the results in this study suggested that closely located repetitive sequences with same motif structure influence each other significantly, the required length of unique sequence which could safely separate them remains uncertain and could be the next step to be explored.

Acknowledgments

We would like to give sincere thanks to Professor Naoko Takezaki who generously provided us the raw data file for this study. We would also express our appreciation to Dr. Kirill Kryukov for assistance on computational analysis, and Mrs. Masako Mizuguchi and Mrs. Ai Iida for assisting preparation of figures. We also thank Dr. Timmothy A. Jinam, Dr. Hideaki Kanzawa-Kiriyama, Mr. Morteza Mahmoudisaber, Dr. Nilmini Hettiarachchi, Dr. Isaac Babarinde, and Ms Shayire Shokat, for their comments to the manuscript. We also appreciate comments of two reviewers.

References

Almeida P. and Penha-Goncalves C. (2004) Long perfect dinucleotide repeats are typical of vertebrates, show motif preferences and size convergence. Molecular Biology and Evolution, 21: 1226–1233.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25: 3389–3402.

Amos W. (2010). Mutation biases and mutation rate variation around very short human microsatellites revealed by human–chimpanzee–orangutan genomic sequence alignments. Journal of Molecular Evolution, 71: 192–201.

Bhargava A. and Fuentes F.F. (2010) Mutational dynamics of microsatellites. Molecular Biotechnology, 44: 250–266.

Boyer J.C., Hawk J.D., Stefanovic L., and Farber R.A. (2008) Sequence-dependent effect of interruptions on microsatellite mutation rate in mismatch repair-deficient human cells. Mutation Research, 640: 89–96.

Chung H., Lopez C.G., Holmstrom J., Young D.J., Lai J.F., Ream-Robinson D., and Carethers J.M. (2010) Both microsatellite length and sequence context determine frameshift mutation rates in defective DNA repair. Human Molecular Genetics, 19: 2638–2647.

Ellegren H. (2004) Microsatellites: simple sequence with complex evolution. Genetics, 5: 435–445.

Galindo C.L., Mclver L.J., McCormick J.F., Skinner M.A., Xie Y., Gelhausen R.A., Ng K., Kumar N.M., and Garner H.R. (2009) Global microsatellite content distinguishes humans, primates, animals, and plants. Molecular Biology and Evolution, 26: 2809–2819.

Goldstein D.B., Linares A.R., Cavalli-Sforza L.L., and Feldman M.W. (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. Proceedings of the National Academy of Sciences of the United States of America, 92: 6723–6727.

Grubbs F.E. (1969) Procedures for detecting outlying observations in samples. Technometrics, 11: 1–21.

Kelkar Y.D., Tyekucheva S., Chiaromonte F., and Makova K.D. (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Research, 18: 30–38.

Kelkar Y.D., Strubczewski N., Hile S.E., Chiaromonte F., Eckert K.A., and Makova K.D. (2010) What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. Genome Biology and Evolution, 2: 620–635.

Kofler R., Schlötterer C., Luschützky E., and Tamas L. (2008) Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. BMC Genomics, 9: 612.

Kroutil L.C., Register K., Bebenek K., and Kunkel T.A. (1996) Exonucleolytic proofreading during replication of repetitive

DNA. Biochemistry, 35: 1046–1053.

Kunkel T.A. (1985) The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Journal of Biological Chemistry, 260: 5787–5796.

Lacroix-Triki M., Lambros M.B., Geyer F.C., Suarez P.H., Reis-Filho J.S., and Weigelt B. (2011) Absence of microsatellite instability in mucinous carcinomas of the breast. International Journal of Clinical Experimental Pathology, 4: 22–31.

Lai Y. and Sun F. (2003) The relationship between microsatellite slippage mutation rate and the number of repeat units. Molecular Biology and Evolution, 20: 2123–2131.

Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., and Higgins D.G. (2007) Clustal W and Clustal X version 2.0. Bioinformatics, 23: 2947–2948.

Leclercq S., Rivals E., and Jarne P. (2007) Detecting microsatellites within genomes: significant variation among algorithms. BMC Bioinformatics, 8: 125.

Li S.L., Yamamoto T., Yoshimoto T., Uchihi R., Mizutani M., Kurimoto Y., Tokunaga K., Jin F., Katsumata Y., and Saitou N. (2006) Phylogenetic relationship of the populations within and around Japan using 105 short tandem repeat polymorphic loci. Human Genetics, 118: 695–707.

Ohta T. and Kimura M. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genetical Research, 22: 201–204.

Oliveira E.J., Pádua J.G., Zucchi M.I., Vencovsky R., and Vieira M.L. (2006) Origin, evolution and genome distribution of microsatellites. Genetics and Molecular Biology, 29: 294–307.

Perelman P., Johnson W.E., Roos C., Seuánez H.N., Horvath J.E., Moreira M.A., Kessing B., Pontius J., Roelke M., Rumpler Y., Schneider M.P.C., Silva A., O'Brien S.J., and Pecon-Slattery J. (2011) A molecular phylogeny of living primates. PLoS Genetics, 7: e1001342.

Sia E.A., Kokoska R.J., Dominska M., Greenwell P., and Petes T.D. (1997) Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Molecular and Cellular Biology, 17: 2851–2858.

Song X-B., Zhou Y., Ying B-W., Wang L-L., Li-YS., Liu J-F., Bai X-G., Zhang L., Lu X-J., Wang J., and Ye Y-X. (2010). Short-tandem repeat analysis in seven Chinese regional populations. Genetics and Molecular Biology, 33: 605–609.

Sun J.X., Mullikin J.C., Patterson N., and Reich D.E. (2009) Microsatellites are molecular clocks that support accurate inferences about history. Molecular Biology and Evolution, 26: 1017–1027.

Takezaki N. and Nei M. (2009). Genomic drift and evolution of microsatellite DNAs in human populations. Molecular Biology and Evolution, 26: 1835–1840.

Zhang L., Zuo K., Zhang F., Cao Y., Wang J., Zhang Y., Sun X., and Tang K. (2006) Conservation of noncoding microsatellites in plants: implication for gene regulation. BMC Genomics, 7: 323.