Ahmed et al. Hereditas (2017) 154:12
DOI 10.1186/s41065-017-0034-4

Hereditas

RESEARCH Open Access

A comparative genomics approach revealed evolutionary dynamics of microsatellite imperfection and conservation in genus *Gossypium*

Muhammad Mahmood Ahmed, Chao Shen, Anam Qadir Khan, Muhammad Atif Wahid, Muhammad Shaban and Zhongxu Lin*

Abstract

Background: Ongoing molecular processes in a cell could target microsatellites, a kind of repetitive DNA, owing to length variations and motif imperfection. Mutational mechanisms underlying such kind of genetic variations have been extensively investigated in diverse organisms. However, obscure impact of ploidization, an evolutionary process of genome content duplication prevails mostly in plants, on non-coding DNA is poorly understood.

Results: Genome sequences of diversely originated plant species were examined for genome-wide motif imperfection pattern, and various analytical tools were employed to canvass characteristic relationships among repeat density, imperfection and length of microsatellites. Moreover, comparative genomics approach aided in exploration of microsatellites conservation footprints in *G* evolution. Based on our results, motif imperfection in repeat length was found intricately related to genomic abundance of imperfect microsatellites among 13 genomes. Microsatellite decay estimation depicted slower decay of long motif repeats which led to predominant abundance of 5-nt repeat motif in *G* species. Short motif repeats exhibited rapid decay through the evolution of *G* lineage ensuing drastic decrease of 2-nt repeats, of which, "AT" motif type dilapidated in cultivated tetraploids of cotton.

Conclusion: The outcome could be a directive to explore comparative evolutionary footprints of simple non-coding genetic elements i.e., repeat elements, through the evolution of genus-specific characteristics in cotton genomes.

Keywords: *G* , Microsatellites, Motif imperfection, Comparative genomics, Evolution

Background

Microsatellites are DNA structural elements in which a short sequence pattern (motif) is repeated by various numbers and ubiquitously presented in genomes of eukaryotes and prokaryotes. Various mechanisms like replication slippage, unequal crossover, realignment after disassociation of replicating strands, mispairing and base substitution are causative to variations which ultimately lead to extensive polymorphism in microsatellites [1, 2]. Replication slippage events are majorly responsible for extensive repeat length variations, and mispairing of replication strands is key determinant for base substitution in DNA sequence [3]. Moreover, stable wobble mispairing and inefficiency of mismatch repair system could incorporate point mutations anywhere in genome sequence [4]. These processes could impinge on microsatellites and generate repeats with occasional mismatch, motif imperfection, in their repeat units.

During the replication process, insertion or deletion in repeat motif units due to reduced strand specificity or slippage errors could generate length variations. Various models comprehend and illustrate the mechanism of

* Correspondence: linzhongxu@mail.hzau.edu.cn
National Key Laboratory of Crop Genetic Improvement, College of Plant Science & Technology, Huazhong Agricultural University, Wuhan, Hubei 430070, China

replication slippage event [5]. A single unit repeat is inserted or deleted at a time according to stepwise mutation model [6], while repeat unit variations were found associated to a distinct probability distribution under specific assumptions [7]. Similarly, another model proposed generation of microsatellites as 3′ extension of retrotransposons in a way similar to polyadenylation of mRNA [8]; while, idea of probable generation and insertion of microsatellites in 5′ region and inside the mobile elements further extended the model [9]. On contrary, the mechanisms underlying mispairing and interruptions leading to single nucleotide changes are poorly understood [10].

Non-coding DNA elements, including microsatellites, also experience various evolutionary forces and show high mutation rates ($10^{-2} - 10^{-6}$) in response to selection forces as these usually go unnoticed [11]. Thus, higher mutability and hyper variability of microsatellites could be key determinants of a dynamic equilibrium state in which rapid loss or conservation of non-coding DNA elements exist in species among or within a clade [12]. Various studies reported microsatellite conservation over longer time [13], and recently a study reported conservation up to 450 million years ago (MYA) in vertebrates [14]. Similarly, evolutionary forces might regulate the mechanism which drives fate determination processes of non-coding or coding DNA elements. Thus, conservation and/or deterioration of DNA structural elements over a geographic time period determine impact of species-evolving processes like speciation, diversification, domestication and even duplication through polidization in plant genomes.

Non-coding elements constitute a major proportion of DNA in all forms of life and play a crucial role in modulating coding regions of DNA. Occurrence and applicability of microsatellites in coding sequence (CDS) have elucidated their functional importance [15]. Since microsatellites are implied to regulate biological functions [16], mutations in coding microsatellites could be informative and even point mutation might induce deleterious effects. However, our brief understandings about mechanisms underlying motif imperfection, either through base substitutions or point mutation, cause an eminent impediment in this regard.

In this report, genome-wide motif imperfection pattern were discerned in 13 plant genomes of diverse origins. The relationship between repeat length and degree of motif imperfection with its impact on the genomic abundance of imperfect microsatellites was determined, particularly for *G i* species. Moreover, evolutionary patterns of microsatellite conservation and/or loss among *G i* species were also established to ascertain structural consequences of whole genome duplication and allopolypolidization events through evolution of cultivated cotton tetraploids.

Methods

Genome assemblies of 13 plant species

The genome sequences of thirteen plant species were investigated comprising ten dicots and three monocots. The dicot species belonged to family Brassicaceae (*A abid i ha a* and *B a ica a a*), Fabaceae (*G ci e a* and *Pha e a i*), Salicaceae (*P ich ca a*) and majorly to Malvaceae (*G i a b e*, *G. ai dii*, *G. hi*, *G. ba bade ce* and *The b a caca*). The three monocots belonged to family Poaceae (*B ach - di di ach*, *O a a i a* and *Zea a*). In addition, coding sequences of annotated genes were also searched for imperfect repeats among four *G i* species. The genome assemblies and CDS sequences of *G i* species were retrieved from Cottongen [17], while genome sequences of other nine plant species were obtained from NCBI Genome Portal [18]. The name of each species was abbreviated to four letters where first capital letter denoted to *ge* and trailing three letters as *ecie* name.

Imperfect microsatellites identification

The SciRoKo (v3.4) program utility [19] was used with default imperfect search parameters to identify imperfect microsatellites of varying motif length from 1 to 6 nucleotide (nt) that were designated as 1-nt, 2-nt, 3-nt, 4-nt, 5-nt and 6-nt among 13 plant genomes. Imperfect search criteria regarding repeat length and mismatch penalty were modified, as previously described [20]. Maximum number of successive mismatch was set to '3', and minimum repeat length was set to 15 nt. Compound microsatellites were excluded where maximum distance for association between two repeats was less than 100 nt. For motif standardization, each microsatellite motif underwent a two-step complete standardization procedure. In the first step, repeat motifs like "AG" and "GA" were all categorized as "AG" and termed as partially standardized motifs. Subsequently, in the second step, reverse complement sequence of motifs was determined and all microsatellites were assigned to completely standardized motifs. Eventually, motifs like "AG", "GA", "TC" and "CT" were all designated to a single group "AG"" for further analysis. Such complete motif standardization resulted in 2, 4, 10, 33, 102 and 350 group categories for 1-nt, 2-nt, 3-nt, 4-nt, 5-nt and 6-nt repeat motifs, respectively.

Motif imperfection and repeat length analyses

The genomic abundance of imperfect repeats was ascertained with varying mismatch counts and repeat lengths. Initially, short length repeats were selected and motif imperfection was employed to discriminate between perfect (with no mismatch) and imperfect microsatellites (with mismatch >0) among all genomes. It was noticed that loci with repeat length < 20 nt did not carry

mismatch, thus were excluded from further analysis. While, repeats <25 nt had only '1' mismatch per locus and there were up to '2' mismatch in repeat loci <30 nt. The two datasets were mined separately, one having genomic abundance of microsatellites with '0' or '1' mismatch (no or low degree imperfection). While, other comprised of exactly '1' or '2' mismatch (low or higher degree imperfection) for repeats of length 20 to 24 nt (dataset I) and 25 to 29 nt (dataset II), respectively.

Thereafter, three different analytical strategies were employed. Firstly, principal coordinates method [21] was used to conduct generalized discriminant analysis on both datasets separately. The analysis generated canonical axes scores which were used to determine correlation among the variables, and statistical significance was estimated by permutation (= 9999) method implemented in 'CAP' program [22]. Moreover, the R package [23] was also used to test the significance of canonical correlations by add-on package 'CCP' [24]. Secondly, longer microsatellites of length 35, 40, 45, 50, 55, 60, 65, 70, 75 and 80 nt were targeted with different levels of imperfection (1, 2, 3 and 4 mismatch per loci). The relationships between motif imperfection and varied length repeats were estimated using permutational analysis of variance *PERMANOVA* [25] by calling "adonis" function under package vegan [26], and statistical significance of *F- a i ic* was measured (0.05) after permutations (= 999) in R environment. Thirdly, the genomic abundance of imperfect microsatellites was related to mismatch counts and repeat length through - *a i ic* and significant difference were determined (0.05). A linear regression model was fitted to mismatch count and repeat length in each *G i* species.

Transposable elements (TEs) distribution of perfect vs imperfect microsatellites

Perfect (no mismatch) and imperfect repeats (mismatch ≥1) were compared for relatedness to nearby intact TEs. Each chromosome was considered as independent segment, and TEs abundance was determined among the four cotton genomes in vicinity of microsatellites (500 nt flanked region on both sides) for both sets of repeats. Then, a linear regression model was fitted to the repeat density and TEs abundance in each *G i* species.

Motif imperfection in coding region of *Gossypium* species

Annotated coding sequences were searched for imperfect repeats under default criteria. All publically available annotated single nucleotide polymorphism (SNP) markers sequences [17] for cotton genomes were downloaded and employed for functional characterization of imperfect microsatellites in coding region.

Microsatellites conservation among *Gossypium* species

For conservation analysis, microsatellite libraries were developed separately for each species containing repeats flanked with 350 nt on both sides. Because of the drawbacks pertained to sequencing homo-polymers, 1-nt repeats were excluded. Moreover, all compound and overlapping sequences were filtered out and Basic Local Alignment Search Tool (BLAST) was used to validate uniqueness of each locus by NCBI BLAST 2.2.31 [27] through self-BLASTing each library and discarding false positives. A custom perl script was used to hard mask repeat sequences for each locus, and genome to sub-genome (diploid vs sub-diploid in tetraploids) BLAST searches was conducted among libraries. The output was filtered for algorithms like alignment ≥50% of query cover, identity ≥70% and E_value 10^{-10}. Since a whole genome duplication (WGD) event was reported through cotton tetraploids' evolution [28], duplication of each progenitor loci was allowed up to two physical positions for pair wise comparisons (A_2 vs A_T & D_5 vs D_T) among homologous chromosomes. The results were validated in reciprocal BLAST searches. Thereafter, conserved microsatellites proportion was employed to develop an intuitive diagram for genome-wide microsatellite conservation by motif size using circos tool [29].

Estimate of microsatellite decay during paleopolyploidization

The WGD event involved through paleopolyploidization of cultivated tetraploid species was estimated with varied divergence time [30]. A representative median divergence time as ~6 MYA was used. Assuming steady loss of microsatellite loci over the period of divergence time, an exponential fitted decay function was employed to estimate "comparative exponential decay" of microsatellite in *G. hi* and *G. ba bade ce*. On contrary, the "relative decay" estimates of varying motif repeats in tetraploids were determined by comparing proportion of conserved repeats. Thereafter, "relative decay" and "comparative exponential decay" were fitted to an exponential decay function. All non-parametric test statistics were measured and tested for significance (0.05) in R environment.

Microsatellite relative abundance

The repeat density patterns in cotton diploids exploited evolutionary footprints regarding impact of paleopolyploidization event on it in three different ways. Firstly, proportional abundance of microsatellites in diploid and tetraploid *G i* species was estimated by comparing their relative abundance in *T. caca* , a close relative belong to Malvaceae. Secondly, the relative abundance of microsatellites according to motif kind and size in four cotton genomes was determined individually. And

lastly, comparative abundance of standardized 2-nt motif was analyzed in four cotton genomes. All non-parametric test statistics were measured and tested for significance (0.05) in R environment.

Results

Frequency and distribution of microsatellites among monocots and dicots

Microsatellite distribution of varying motif length (1–6 nt) was examined in ten dicots and three monocots. Generally, 2-nt and 3-nt repeats were found in higher percentage in dicots and monocots (Table 1), respectively. While, Atha exhibited almost equal proportion of three motifs (1–3 nt). A plentiful abundance of 3-nt microsatellite was observed in Osat (33.67%) and the trend persisted among Bdis and Zmay. Among dicots, the proportion of 2-nt repeats prevailed and Brap featured the highest proportion of it (39.67%). However, frequency of 5-nt microsatellites predominated and featured specific to four *G i* genomes. Reckoning to the four cotton genomes, a drastic reduction in percentage of 2-nt repeats was observed in tetraploids (Ghir and Gbar) up to 11.95%. While, density of microsatellites with short motif repeats decreased, those with longer motifs (4–6 nt) exhibited incessant abundance in tetraploids compared to their progenitors (Table 1).

Our data depicted differential variations in repeat density (per Mb) and frequencies of imperfect microsa-

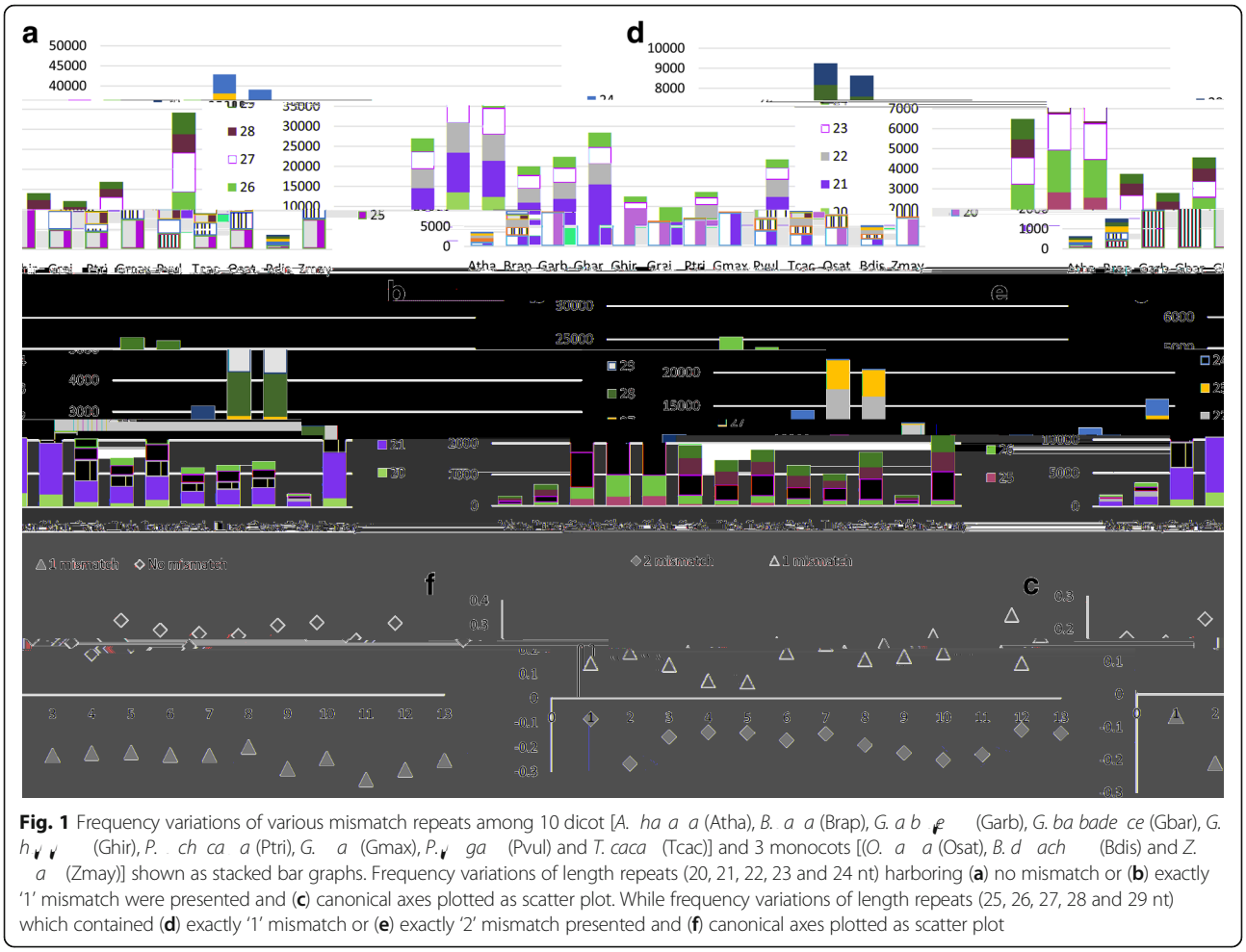

Fig. 1 Frequency variations of various mismatch repeats among 10 dicot [*A. ha a a* (Atha), *B. a a* (Brap), *G. a b e* (Garb), *G. ba bade ce* (Gbar), *G. h* (Ghir), *P. ch ca a* (Ptri), *G. a* (Gmax), *P. ga* (Pvul) and *T. caca* (Tcac)] and 3 monocots [(*O. a a* (Osat), *B. d ach* (Bdis) and *Z. a* (Zmay)] shown as stacked bar graphs. Frequency variations of length repeats (20, 21, 22, 23 and 24 nt) harboring (**a**) no mismatch or (**b**) exactly '1' mismatch were presented and (**c**) canonical axes plotted as scatter plot. While frequency variations of length repeats (25, 26, 27, 28 and 29 nt) which contained (**d**) exactly '1' mismatch or (**e**) exactly '2' mismatch presented and (**f**) canonical axes plotted as scatter plot

(F = 8.050, = 0.0029) and motif imperfection (F = 1.495, = 0.0055). These findings speculated existence of an evolutionary mechanism which determined length and motif heterogeneity of microsatellites and the pattern prevailed among the studied set of genomes.

Pair-wise *Pea* correlation estimates were calculated by controlling motif imperfection (Additional file 4: Table S4a-b) for different repeat lengths (35, 40, 45, 50, 55, 60, 65, 70 and 75 nt). The results established strong relationships among different length microsatellites carrying varying mismatch (Table 2). Genomic abundance of '1' mismatch repeats was found more related to '2' mismatch repeats, and the relationship maintained with '3' mismatch repeats. The scenario became more stringent for pair-wise comparisons between '2' mismatch and '3' mismatch repeats. Thus, the genomic abundance of loci with lower degree of imperfection (1 mismatch) was related to varying repeat lengths. Moreover, repeats with low imperfection (1 mismatch) seemed to modulate genomic abundance of repeats with moderate imperfection (2 mismatch). Similarly, repeats with higher imperfection

Table 2 Pair-wise correlation (*Pea*) estimates for imperfect microsatellites

Repeat length (nt)	Mismatch 1 vs mismatch 2	Mismatch 1 vs mismatch 3	Mismatch 2 vs mismatch 3
30 nt	0.983	0.921	0.952
35 nt	0.919	0.919	0.971
40 nt	0.911	0.807	0.894
45 nt	0.493	0.640	0.763
50 nt	0.767	0.545	0.821
55 nt	0.695	0.699	0.822
60 nt	0.558	0.605	0.643
65 nt	0.868	0.742	0.774
70 nt	0.776	0.629	0.423
75 nt	0.896	0.415	0.492
80 nt	0.742	0.852	0.867

The repeat length presented different lengths while frequencies of repeats harbored 1 mismatch (low imperfection), 2 mismatch (moderate imperfection) and 3 mismatch (high imperfection) were compared among 13 plant species. All estimates were significant while underlined were non-significant at 0.05

(3 mismatch) were more inflected by repeats which entertained '2' mismatch. These findings suggested a significant role of motif imperfection in determining length of repeats.

Motif imperfection in *Gossypium* genomes

To draw deep insights into evolutionary footprints, subsets of mismatched repeats in four *G i* species were considered for following analyses. Linear regression between repeat length and motif imperfection was determined and a prominent relationship was observed in each case (Fig. 2), but with varied degree of variation accounted in the model (Additional file 5: Table S5).

The genomic abundance of intact TEs was determined in 500 nt flanking region (both sides) of perfect and mismatched microsatellites (Additional file 6: Table S6). The two sets of repeats were compared for relatedness between repeat frequency and intact TEs, and significant relationships were observed (Additional file 7: Figure S1a-d). Generally, fewer TEs were observed in Grai and D_T sub-genomes; while, higher abundance of TEs discerned for Garb and A_T sub-genomes of tetraploids. Moreover, imperfect repeats were more colligated to genomic abundance of TEs than perfect ones except in Garb; where perfect repeats were more related than mismatched ones. This relatedness was further affirmed by *Pea* correlations and significance determined at 0.05 (Table 3). Correlation values of imperfect repeats were also expectedly higher than perfect ones, but not for Garb in which perfect repeats were most likely imbedded in TEs.

Differential abundance of imperfect repeats along with intact TEs distribution was determined among individual chromosomes. Interestingly, the frequency of imperfect microsatellites and intact TEs for all species was found at par (Additional file 6: Table S6). The higher abundance of intact TEs was observed in Garb than Grai, while vice versa was true for imperfect repeats. Chromosomes A05, A08, A10 and A11 exhibited higher density of both repeat elements, while a mixed pattern was observed for D03, D07 and D12 among tetraploids. Structural anomalies of chromosomes in two genetically distant progenitors, evolutionary processes, biased selection forces, mutations, deletions and translocations of larger DNA segments could be causative to such abrupt differentiations among tetraploids and diploids. Thus, comparative distributive pattern and its relatedness suggested more likely presence of intact TEs wherever a mismatch found in a microsatellite.

Motif interruptions in coding sequence

Coding sequences of *G i* species were also searched for imperfection and results were summarized in Additional file 8: Table S7. Generally, tri- and hexa- motif repeats prevailed, while only tri- motif repeats accounted for >70% of microsatellites in CDS regions (Table 4). It was expected due to triplet and degenerate nature of codons. The motif imperfection in Grai repeats was elevated and slight increment in average repeat length was noticed. The tandem repeat density was reduced than genomic ones, while the higher proportion of genes in Ghir harbored repeats in exons (Additional file 8: Table S7).

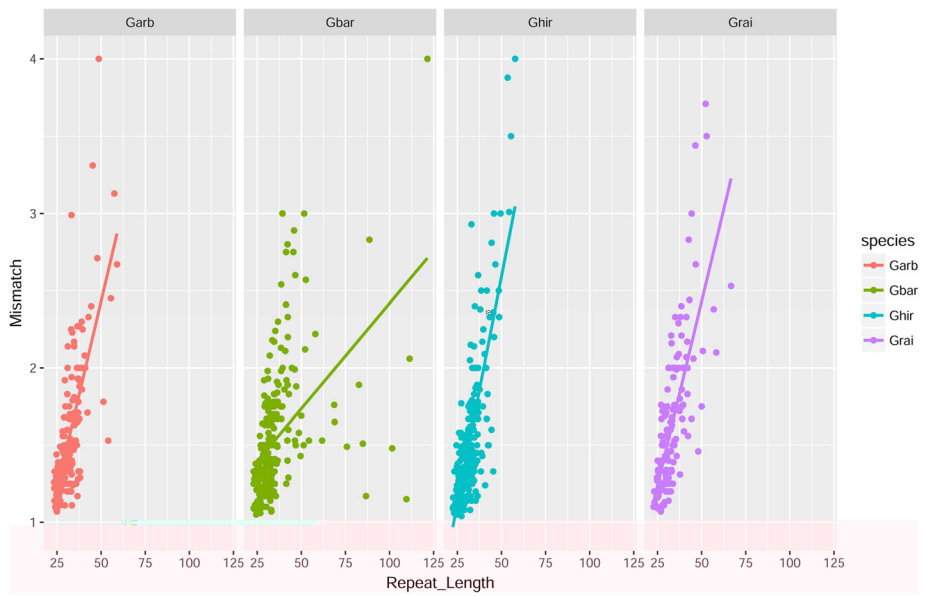

Fig. 2 The mismatch (count) and repeat length relationship split among *G* species. Average mismatch of each standardized motif was compared to their average repeat length (nt) in Garb (*G. a b e*), Grai (*G. a d*), Ghir (*G. h*) and Gbar (*G. ba bade ce*)

Table 3 A comparative correlation estimates between perfect (no mismatch) and imperfect (mismatch ≥1) repeats along with their intact TEs in cotton genomes

Genome/ Sub-genome	Perfect repeats			Imperfect repeats		
	Pea	correlation	value	*Pea*	correlation	value
Garb	0.944		1.13E-06	0.743		3.54E-03
Grai	0.631		2.01E-02	0.835		5.23E-04
$GhirA_T$	0.788		1.37E-03	0.911		1.47E-05
$GhirD_T$	0.615		2.50E-02	0.939		1.39E-06
$GbarA_T$	0.869		1.11E-04	0.879		7.49E-05
$GbarD_T$	0.648		1.64E-02	0.846		2.58E-04

Overall, the motif imperfection mechanisms targeted tri- and hexa- repeats in genomic and coding sequences respectively (Additional file 9: Figure S2).

Long motif repeats well conserved in *Gossypium* species

Our data showed predominant abundance of large microsatellites in *G i* species compared to other dicots and monocots. Microsatellite conservation pattern elucidated the possible influence of paleopolyploidization event and traced out the evolutionary footprints in *G i* species. The microsatellites proportion of each A_T and D_T sub-genome in two cultivated species was ascertained which retained from the progenitors (Fig. 3). The shorter repeats were found less conserved, while repeats with long motifs remained intact as in progenitor species (Additional file 10: Table S8). Among short repeats, the 2-nt microsatellites were found more deteriorated in tetraploids where Gbar retained the lowest proportion from two diploids. On the contrary, 3–6 nt repeats were more conserved, while 5-nt repeats were highly conserved (56.86–78.26%) in both tetraploids. Overall, Ghir retained higher proportion repeats from Garb (66.58%); while, slightly lower proportion retained (61.39%) from another progenitor. Since microsatellites were more deteriorated through series of evolutionary events in Gbar, thus fewer loci retained intact (37.83–43.99%).

Estimation of microsatellite decay

The conserved microsatellite proportion and divergence time were fitted to an exponential decay function and microsatellites decay rate of varied length motifs was calculated among *G i* species. A static decay was observed in Ghir for loci from both progenitors except 2-nt repeats; of which D_T originated repeats were lost significantly faster (*F ied a a i ic* = 48.44, = 0.00001) (Additional file 11: Figure S3a). Whereas in Gbar, 2-nt repeats of D_T origin were highly distorted and lost due to significantly faster decay (*F ied a a i - ic* = 55.083, = 0.00001) (Additional file 11: Figure S3b). In comparison, Ghir retained higher proportion of microsatellites from diploid species than Gbar. Similarly, large motif repeats (3-6 nt) were more evolutionary stable and conserved compared to 2-nt repeats in both A-genome and D-genome of *G i* lineage. Overall, repeats loci in Ghir depicted slower decay of repeats than Gbar (Fig. 4a), while a significant faster decay (*F ied a a i ic* = 18.762, = 0.0021) of Grai repeats was observed in tetraploids compared to Garb microsatellites. Considering both "relative decay" and "comparative exponential decay", all repeat motif exhibited faster decay in Gbar than Ghir (Fig. 4b).

Biased distribution of dinucleotide repeats in *Gossypium* species

A WGD event followed by allopolypolidization could be causative to observed evolutionary pattern of microsatellite conservation in *G i* tetraploids. The relative abundance estimates of four *G i* species' repeats relative to *T. caca* (as of their closest phylogenetic ancestor) further substantiated the microsatellites decay pattern in *G i* lineage (Additional file 12: Figure S4a-d). The short (1–3 nt) repeats were found less abundant in Garb (*K a Wa a i ic* = 24.15, = 0.0002) and Grai (*K a Wa a i ic* = 21.68, = 0.0006), while only 1-nt and 2-nt were drastically lost in Ghir (*K a Wa a i ic* = 17.23, = 0.0041) and Gbar (*K a Wa a i ic* = 17.84, = 0.0031). While comparing relative abundance among cotton genomes, the short repeats (1-nt and 2-nt) prevailed less (*K a Wa a i ic* = 15.70, = 0.0077) than longer (4-6 nt) ones (Fig. 5a). As decay rates were not determined for 1-nt repeats, thereby 2-nt repeat motif was pointed as scarcely abundant. Moreover, all 2-nt standardized motifs were lost in comparative way, while "AT" motif repeats

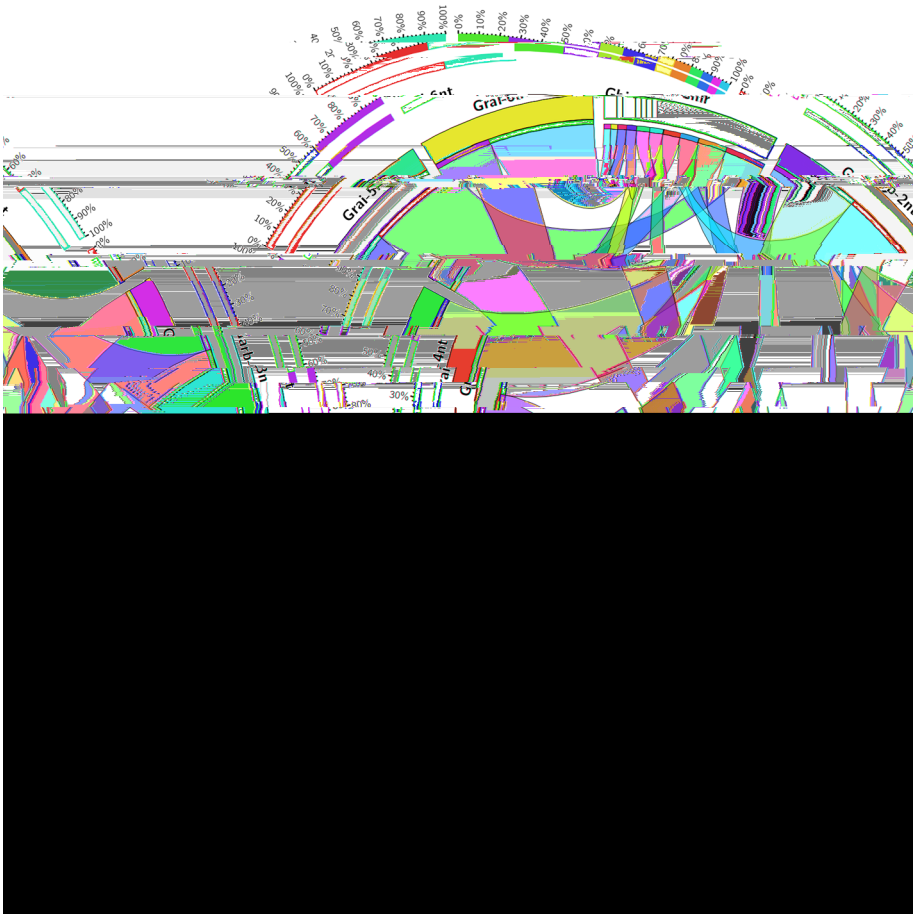

Fig. 3 Intuitive diagram shows conserved microsatellite proportion based on motif length. Motif repeats like 2-nt (dinucleotide), 3-nt (trinucleotide), 4-nt (tetra- nucleotide), 5-nt (penta- nucleotide) and 6-nt (hexa- nucleotide) were compared for conservation among Garb (*G. a b e*), Grai (*G. a d*), Ghir (*G. h*) and Gbar (*G. ba bade ce*)

were more rapidly lost in Ghir and Gbar (Fig. 5b). The ploidization could be a major factor for drastic loss of "AT" standardized motif as despite of ploidy increase, tetraploids conserved lower proportion of "AT" motif repeats than diploids.

Discussion

The study provided a comprehensive account of microsatellite imperfection in 13 plant genomes and explored motif imperfection and repeat length relationships in *G i* lineage. The microsatellite conservation pattern and evolutionary footprints of non-coding repetitive DNA elements were also discerned in closely related species of Malvaceae family.

Genomic abundance of imperfect microsatellites is well studied in simple organisms like viruses [31], mushroom [32], complex animal [33], insect [34], human [35] and even in green plants [36]. However, a clear understanding of imperfection mechanism, its role in genesis, preserving sequence variations in non-coding repeat elements and their potential impact on gene regulation are not well understood. As focus on evolution of microsatellites got reincarnated with the advent of high throughput sequencing technology [37, 38], this study emphasized on role of motif imperfection in microsatellites stability. Plant species were emphasized as microsatellites are being extensively employed in revealing diversity, heterozygosity and exploiting phenotypic variations in plants [39, 40]. Moreover, microsatellites have also been employed to unravel polyploidy in plants [41]. Evolutionary dynamics of microsatellites were revealed in genus *G i* , a best model to investigate ploidy increase impact on microsatellite imperfection and conservation.

Various tools are available for imperfect repeats search, but SciRoKo [19] was preferred since it provides a utility to efficiently digest large genome datasets into 50 Mb chunks. Moreover, assessing efficiency of various tools was beyond the scope of this study. Our results witnessed predomination of 3-nt and 2-nt repeats in

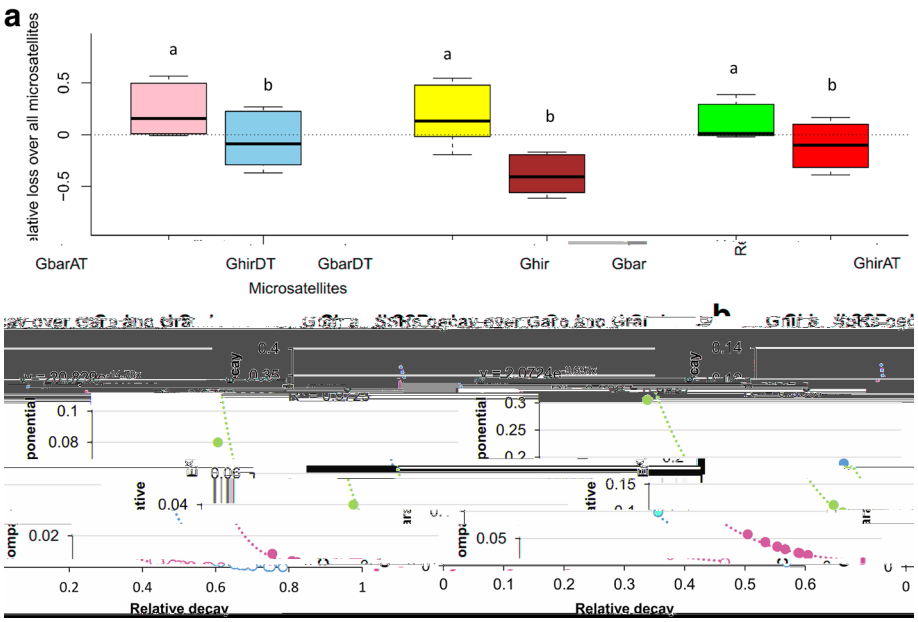

Fig. 4 Comprison of microsatellite decay between two cotton tetraploids compared to repeats originated in diploids (Garb and Grai). **a** Relative loss of all SSRs among A_T and D_T subgenomes of tetraploids (Ghir and Gbar). The *b ac ba h b* represens median and *d ed h a e* represent = 0, while significance was denoted as *e ca e e e*. **b** Comparing relationship of microsatellite relative decay and comparative exponential decay between Ghir and Gbar

monocots and dicots respectively. Many researchers have reported 3-nt repeats' abundance as characteristics in monocots [42, 43]. Similarly, predominate abundance of 2-nt repeats has been reported in dicots [44, 45], but paramount abundance of other motifs is not out of the scope. The resultant fluctuations might be owed to variations in search parameters. Appropriate parameters were employed for imperfect microsatellites search in all studied plant genomes (Table 1 & Additional file 1: Table S1). Likewise, the distribution and abundance of imperfect repeats varied among the studied species, and it also varied among homologous chromosomes of species (Additional file 2: Table S2). Previous studies supported varied distribution of imperfect repeats among different species of genus *D hi* [46]. The trend was not affected by varying quality of 13 genome assemblies as a consistent pattern of motif imperfection was observed.

The genomic abundance of imperfect microsatellites was intricately related to degree of motif imperfection and repeat length variations. Previously, repeat unit variations were found linked to repeat length and mutations [47, 48], while an equilibrium state was proposed to regulate abundance of mismatched repeats [1]. A varying stability of mismatched repeats due to sequence interruptions has been experimentally tested [49]. While, a study determined relationship between mutation and repeat units and explained why longer repeats likely to undergo contractions and short repeats experience expansion in case of slippage mutations [50]. The longer repeats (>40 nt) with no mismatch or lower imperfection (mismatch 1 and/or 2) were more abundant than those with higher degree of imperfection (Additional file 4: Table S4a-b). Moreover, a significant relationship between mismatch count and repeat length existed among *G i* species (Additional file 5: Table S5). Generally, the longer repeats do not remain stable and immediately split to give shorter repeats as these undergo contraction. However, our results predicted longer repeats were more likely to harbor mismatch, but these repeats were less prone to contraction as motif interruptions could impart stabilizing effect to such repeats [51]. Furthermore, a low to moderate positive correlation between microsatellites and TEs was recurrently reported [52, 53] but not between imperfect microsatellites and TEs. Due to sequence interruptions, the longer repeats were found more related to TEs abundance except for Garb, where perfect repeats were more likely to be present in the vicinity of TEs (Table 3). Thus, it can be speculated that microsatellite genesis, prolonged stability, length variations and motif imperfection of mismatched repeats could be better comprehended when the replication slippage errors, polyadenylation of 5′ and 3′ regions of TEs are considered simultaneously.

Recent studies reported that 5-nt repeats were predominated in Garb and Grai [54, 55]. However, another study reported 6-nt repeats prevailed in Ghir [56]; while our results affirmed that 5-nt repeats were consistently frequent

in four sequenced G i genomes. Although the dif-

Additional files

Additional file 1: Table S1. Characteristics of imperfect microsatellites investigated among 13 plant genomes. (XLS 25 kb)

Additional file 2: Table S2. Frequency of imperfect microsatellites distributed chromosomes and imperfection (%) in 13 plant genomes. (XLS 39 kb)

Additional file 3: Table S3. Average length comparison between perfect and imperfect repeats. (XLS 33 kb)

Additional file 4: Table S4. (a) Frequency of variable length repeats which harbored mismatch (1, 2, 3 and 4) in repeat motifs across 13 plant genomes; **(b)** Probability values for pair-wise length comparisons among imperfect repeats of 13 plant genomes. (XLS 33 kb)

Additional file 5: Table S5. Regression analysis of motif Imperfection (mismatch count) in relation to repeat length among four *G* species. (XLS 24 kb)

Additional file 6: Table S6. Chromosomal distribution of perfect and imperfect microsatellites in relation with intact (TEs), embedded in 500 nt flanked region on both sides, among four *G* species. (XLS 28 kb)

Additional file 7: Figure S1. Comparison between perfect (no mismatch) and imperfect repeats (mismatch ≥1) for correlation of microsatellites with intact TEs frequency in (a) *G. a b e* (Garb), (b) *G. a d* (Grai), (c) *G. h* (Ghir) and (d) *G. ba bade ce* (Gbar). (DOC 913 kb)

Additional file 8: Table S7. Motif distribution, density and imperfection of microsatellite repeats in coding sequences of four cotton genomes. (XLS 25 kb)

Additional file 9: Figure S2. Comparing motif imperfection pattern between genomic and coding microsatellites of varying motif sizes (2-6 nt) in *G. a b e* (Garb), *G. a d* (Grai), *G. h* (Ghir) and *G. ba bade ce* (Gbar). (DOC 51 kb)

Additional file 10: Table S8. Genome to sub-genome comparison of microsatellite conservation analysis. (XLS 28 kb)

Additional file 11: Figure S3. Relative loss of SSRs by motif length in (a) *G. h* (Ghir) and (b) *G. ba bade ce* (Gbar). The loss of 2-6 nt SSRs compared to the loss of all SSRs (y = 0, denoted by dotted line). Microsatellites of sub-genome A_T are shown in gray filling and D_T sub-genome shown in white. (DOC 78 kb)

Additional file 12: Figure S4. Relative abundance of microsatellite for *G* genomes, (a) *G. a b e* (Garb), (b) *G. a d* (Grai), (c) *G. h* (Ghir), and (d) *G. ba bade ce* (Gbar), compared to distribution of *T. caca* SSRs density by motif length (y = 0, denoted by dotted line). (DOC 153 kb)

Abbreviations

BLAST: Basic local alignment search tool; CDS: Coding sequences; DNA: Deoxyribonucleic acid; Gb: giga (10^9) bases; Mb: mega (10^6) bases; mRNA: messenger RNA; MYA: Million years ago; nt: nucleotide; SNP: Single nucleotide polymorphisms; SSR: Simple sequence repeats; TEs: Transposable elements; WGD: Whole genome duplication

Acknowledgements

We acknowledge the National Natural Science Foundation of China (Program #31371674) for providing financial support for this work.

Funding

This work was financially supported by the National Natural Science Foundation of China (Program #31371674).

Availability of data and materials

We have provided detailed information in material and methods section of our manuscript.

Authors' contributions

ZXL and MMA both conceived the idea and designed the study. MMA, SC and AQK performed data retrieval, handling and compilation. MMA, MAW and MS performed all analytical and statistical analyses. MMA and AQK designed and edited figures while MMA wrote manuscript and ZXL reviewed it. All authors contributed in editing and improving the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Publisher's Note

Received: 14 October 2016 Accepted: 27 April 2017
Published online: 18 May 2017

References

1. Kruglyak S, Durrett RT, Schug MD, Aquadro CF. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci U S A. 1998;95:10774–8.
2. Hancock JM. Microsatellites and other simple sequences: genomic context and mutational mechanisms. In: Goldstein DB, Schlötterer C, editors. Microsatellites: evolution and applications. Oxford: Oxford University Press; 1999. p. 1–9.
3. Echols H, Goodman MF. Fidelity mechanisms in DNA replication. Annu Rev Biochem. 1991;60:477–511.
4. Pray LA. DNA replication and causes of mutation. Nat Educ. 2008;1:214.
5. Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids Res. 1992;20:211–5.
6. Kimura M, Ohta T. Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc Natl Acad Sci U S A. 1978;75:2868–72.
7. Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB. Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci U S A. 1994;91:3166–70.
8. Nadir E, Margalit H, Gallily T, Ben-Sasson SA. Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. Proc Natl Acad Sci U S A. 1996;93:6470–5.
9. Tay WT, Behere GT, Batterham P, Heckel DG. Generation of microsatellite repeat families by RTE retrotransposons in lepidopteran genomes. BMC Evol Biol. 2010;10:144.
10. Wang W, Bittles AH. Imperfect units of an extended microsatellite structure involving single nucleotide changes. Electrophoresis. 2001;22:1095–7.
11. Schlotterer C. Evolutionary dynamics of microsatellite DNA. Chromosoma. 2000;109:365–71.
12. Blanquer-Maumont A, Crouau-Roy B. Polymorphism, monomorphism, and sequences in conserved microsatellites in primate species. J Mol Evol. 1995; 41:492–7.
13. Ezenwa VO, Peters JM, Zhu Y, Arévalo E, Hastings MD, Seppä P, et al. Ancient conservation of Trinucleotide microsatellite loci in Polistine wasps. Mol Phylogenet Evol. 1998;10:168–77.
14. Adams RH, Blackmon H, Reyes-Velasco J, Schield DR, Card DC, Andrew AL, et al. Microsatellite landscape evolutionary dynamics across 450 million years of vertebrate genome evolution. Genome. 2016;59:295–310.
15. Ranade SS, Lin YC, Van de Peer Y, Garcia-Gil MR. Comparative in silico analysis of SSRs in coding regions of high confidence predicted genes in Norway spruce (*P cea ab e*) and loblolly pine (*P aeda*). BMC Genet. 2015;16:149.
16. Wang Q, Zhang X, Wang X, Zeng B, Jia X, Hou R, et al. Polymorphism of CAG repeats in androgen receptor of carnivores. Mol Biol Rep. 2012;39: 2297–303.
17. CottonGen. https://www.cottongen.org/. 15 June 2016.
18. NCBI Genome Portal. https://www.ncbi.nlm.nih.gov/genome/?term. 15 June 2016.
19. Kofler R, Schlotterer C, Lelley T. SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics. 2007;23:1683–5.

20. Behura SK, Severson DW. Motif mismatches in microsatellites: insights from genome-wide investigation among 20 insect species. DNA Res. 2015;22:29–38.
21. Anderson MJ, Robinson J. Generalized discriminant analysis based on distances. Aust NZ J Stat. 2003;45:301–18.
22. Canonical Analysis of Principal Coordinates (CAP). http://www.esapubs.org/archive/ecol/E084/011/suppl-1.htm. 10 May 2016.
23. The R project for statistical computing. https://www.r-project.org/. 10 May 2016.
24. CCP: Significance tests for Canonical Correlation Analysis (CCA). https://cran.r-project.org/web/packages/CCP/index.html. 25 May 2016.
25. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Aust Ecol. 2001;26:32–46.
26. vegan: Community Ecology package. https://cran.r-project.org/web/packages/vegan/index.html. 25 May 2016.
27. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
28. Li FG, Fan GY, Lu CR, Xiao GH, Zou CS, Kohel RJ, et al. Genome sequence of cultivated upland cotton (*G h* TM-1) provides insights into genome evolution. Nat Biotechnol. 2015;33:524–U242.
29. Circos: Circular Visualization. http://circos.ca/. 20 July 2016.
30. Wendel JF, Grover CE. Taxonomy and evolution of the cotton genus. In: Fang D, Percy R, editors. Cotton, Agronomy. Madison: Monograph 24, ASA-CSSA-SSSA; 2015.
31. Alam CM, Singh AK, Sharfuddin C, Ali S. Genome-wide scan for analysis of simple and imperfect microsatellites in diverse carlaviruses. Infect Genet Evol. 2014;21:287–94.
32. Keirle MR, Avis PG, Feldheim KA, Hemmes DE, Mueller GM. Investigating the allelic evolution of an imperfect microsatellite locus in the Hawaiian mushroom Rhodocollybia laulaha. J Hered. 2011;102:727–34.
33. Gaspari Z, Ortutay C, Toth G. Divergent microsatellite evolution in the human and chimpanzee lineages. FEBS Lett. 2007;581:2523–6.
34. Stolle E, Kidner JH, Moritz RF. Patterns of evolutionary conservation of microsatellites (SSRs) suggest a faster rate of genome evolution in hymenoptera than in Diptera. Genome Biol Evol. 2013;5:151–62.
35. Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, et al. A direct characterization of human mutation based on microsatellites. Nat Genet. 2012;44:1161–5.
36. Kapil A, Rai PK, Shanker A. ChloroSSRdb: a repository of perfect and imperfect chloroplastic simple sequence repeats (cpSSRs) of green plants. Database. 2014;2014.
37. Yun YE, Yu JN, Nam GH, Ryu SA, Kim S, Oh K, et al. Next-generation sequencing identification and characterization of microsatellite markers in aconitum austrokoreense Koidz., an endemic and endangered medicinal plant of Korea. Genet Mol Res. 2015;14:4812–7.
38. Kang TH, Han SH, Park SJ. Development of seven microsatellite markers using next generation sequencing for the conservation on the Korean population of Dorcus Hopei (E. Saunders, 1854) (Coleoptera, Lucanidae). Int J Mol Sci. 2015;16:21330–41.
39. Han B, Wang C, Tang Z, Ren Y, Li Y, Zhang D, et al. Genome-wide analysis of microsatellite markers based on sequenced database in Chinese spring wheat (*T ae* L.). PLoS One. 2015;10:e0141540.
40. Ahmed M, Guo H, Huang C, Zhang X, Lin Z. Selection of core SSR markers for fingerprinting upland cotton cultivars and hybrids. Aust J Crop Sci. 2013;7:1912–20.
41. Li X, Jin X, Wang H, Zhang X, Lin Z. Structure, evolution, and comparative genomics of tetraploid cotton based on a high-density genetic linkage map. DNA Res. 2016;23:283–93.
42. Kalyana Babu B, Pandey D, Agrawal PK, Sood S, Kumar A. In-silico mining, type and frequency analysis of genic microsatellites of finger millet (*E c aca a* (L.) Gaertn.): a comparative genomic analysis of NBS-LRR regions of finger millet with rice. Mol Biol Rep. 2014;41:3081–90.
43. Shi J, Huang S, Fu D, Yu J, Wang X, Hua W, et al. Evolutionary dynamics of microsatellite distribution in plants: insight from the comparison of sequenced brassica, Arabidopsis and other angiosperm species. PLoS One. 2013;8:e59988.
44. Liu SR, Li WY, Long D, Hu CG, Zhang JZ. Development and characterization of genomic and expressed SSRs in citrus by genome-wide analysis. PLoS One. 2013;8:e75149.
45. Biswas MK, Xu Q, Mayer C, Deng X. Genome wide characterization of short tandem repeat markers in sweet orange (*C e*). PLoS One. 2014;9:e104182.
46. Ross CL, Dyer KA, Erez T, Miller SJ, Jaenike J, Markow TA. Rapid divergence of microsatellite abundance among species of drosophila. Mol Biol Evol. 2003;20:1143–57.
47. Xu X, Peng M, Fang Z. The direction of microsatellite mutations is dependent upon allele length. Nat Genet. 2000;24:396–9.
48. Falush D, Iwasa Y. Size-dependent mutability and microsatellite constraints. Mol Biol Evol. 1999;16:960–6.
49. Bacon AL, Farrington SM, Dunlop MG. Sequence interruptions confer differential stability at microsatellite alleles in mismatch repair-deficient cells. Hum Mol Genet. 2000;9:2707–13.
50. Lai Y, Sun F. The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol. 2003;20:2123–31.
51. Rolfsmeier ML, Lahue RS. Stabilizing effects of interruptions on Trinucleotide repeat expansions in *Saccha ce ce e ae*. Mol Cell Biol. 2000;20:173–80.
52. Ramsay L, Macaulay M, Cardle L, Morgante M, degli Ivanissevich S, Maestri E, et al. Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. Plant J. 1999;17:415–25.
53. Grandi FC, An W. Non-LTR retrotransposons and microsatellites: partners in genomic variation. Mob Genet Elements. 2013;3:e25674.
54. Lu C, Zou C, Zhang Y, Yu D, Cheng H, Jiang P, et al. Development of chromosome-specific markers with high polymorphism for allotetraploid cotton based on genome-wide characterization of simple sequence repeats in diploid cottons (*G a b e* L. and *G a d* Ulbrich). BMC Genomics. 2015;16:55.
55. Zou C, Lu C, Zhang Y, Song G. Distribution and characterization of simple sequence repeats in *G a d* genome. Bioinformation. 2012;8:801–6.
56. Wang Q, Fang L, Chen J, Hu Y, Si Z, Wang S, et al. Genome-wide mining, characterization, and development of microsatellite markers in *G* species. Sci Rep. 2015;5:10638.
57. Cronn RC, Small RL, Haselkorn T, Wendel JF. Rapid diversification of the cotton genus (*G* : Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. Am J Bot. 2002;89:707–25.
58. Flagel LE, Wendel JF, Udall JA. Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. BMC Genomics. 2012;13:1–13.