Manuscript

Microsatellite Density Landscapes Illustrate Short Tandem Repeats Aggregation in The Complete Reference Human Genome

Yun Xia[1†], Douyue Li[1†], Tingyi Chen[1†], Saichao Pan[1†], Hanrou Huang[1†], Wenxiang Zhang[1†], Yulin Liang[1†], Yongzhuo Fu[1], Zhuli Peng[1], Hongxi Zhang[1], Liang Zhang[1], Shan Peng[1], Ruixue Shi[1], Xingxin He[1], Siqian Zhou[1], Weili Jiao[1], Xiangyan Zhao[1], Xiaolong Wu[1], Lan Zhou[1], Jingyu Zhou[1], Qingjian Ouyang[1], You Tian[1], Xiaoping Jiang[1], Yi Zhou[1], Shiying Tang[1], Junxiong Shen[1] , Kazusato Ohshima[2] and Zhongyang Tan[1*]

[1] Bioinformatic Center, College of Biology, Hunan University, Changsha 410082, China

[2] Faculty of Agriculture, Saga University, 840-8502 Saga, Japan

[†] Co-First Author, contributed equally to this work

[*] Corresponding author

E-mail: zhongyangtan@yeah.net

Address: Lushan Road (S), Yuelu District, Changsha, Hunan Province, China

Postal Code: 410082

Abstract

Background: Microsatellites are increasingly realized to have biological significance in human genome and health in past decades, the assembled complete reference sequence of human genome T2T-CHM13 brought great help for a comprehensive study of short tandem repeats in the human genome.

Results: Microsatellites density landscapes of all 24 chromosomes were built here for the first complete reference sequence of human genome T2T-CHM13. These landscapes showed that short tandem repeats

(STRs) are prone to aggregate characteristically to form a large number of STRs density peaks. We classified 8,823 High Microsatellites Density Peaks (HMDPs), 35,257 Middle Microsatellites Density Peaks (MMDPs) and 199, 649 Low Microsatellites Density Peaks (LMDPs) on the 24 chromosomes; and also classified the motif types of every microsatellites density peak. These STRs density aggregation peaks are mainly composing of a single motif, and AT is the most dominant motif, followed by AATGG and CCATT motifs. And 514 genomic regions were characterized by microsatellite density feature in the full T2T-CHM13 genome.

Conclusions: These landscape maps exhibited that microsatellites aggregate in many genomic positions to form a large number of microsatellite density peaks with composing of mainly single motif type in the complete reference genome, indicating that the local microsatellites density varies enormously along the every chromosome of T2T-CHM13.

Keywords: Human genome; Microsatellite density; STRs aggregation; Landscape.

1. Background

Tandem repeat biology is causing a revolution in genetics in the past two decades as many studies of understanding the evolution and biological functions of tandem repeats in the genomes of thousand species and also human diseases [1-4]. Indeed, more than half of the human genome is constituted of repetitive sequences making them the most challenging genomic regions to study, and repetitive sequences are generally classified into interspersed repeats and tandem repeats [5]. The Short Tandem Repeats (STRs), also called microsatellites, happen ubiquitously in human genome scattered in coding and non-coding regions, which commonly occur with repeat units of 1-6 base pairs and own the highest mutational rate in genome [1, 3, 6]. Microsatellites are increasingly realized to have biological significance in human genome and health, and are reported to involve in more than 30 disorders and several cancers [7-10], regulate gene expression in healthy genomes [11, 12], and also be related to genetic plasticity and missing heritability

[13-20]. Slipped-strand mispairing was suggested as a major mechanism for tandem repeat occurrence [21, 22], and we formerly presented a folded slippage model for short tandem repeats occurring mechanism, predicting that micro-disturbing in the process of replication may provide a lot of chances for the template chain folded to produce short tandem repeats, which are possibly selected and fixated for different biological significance in the long-history evolution [23].

Although short tandem repeats have been studied in human genome for several decades, many sequenced columns of human genome so far contain a large number of unsequenced gaps, and these unsequenced gaps are often composed of short tandem repeats [3, 5]. To date, studies about short tandem repeats on complete, gap-free human genomes have been very lacking. The recent assembly of T2T-CHM13 reference removes the gap filled regions of autosomes and Chromosome X [5, 24-28], and over 50% gaps in Chromosome Y [29], therefore it represents a truly complete human genome, this complete reference sequence of human genome T2T-CHM13 will certainly bring great helps for comprehensive study short tandem repeats in human genome. And so far, there is still no unified threshold for microsatellites studies, the repeats in animal genomes are generally considered to be longer than 12 bp [30, 31], and 3% of human genomes was reported consist of microsatellites under that standard [32]. And microsatellites of human sequences comprised of GA/TC/GC/AT bases are investigated using seq-requester microsatellite [24]. However, we applied a threshold of 6, 3, 3, 3, 3, 3 iterations for mono- to hexa-nucleotide repeat motifs for analyzing microsatellites in the complete reference human genome T2T-CHM13, the threshold was widely used to study microsatellites in the genome sequences of viruses, mitochondrial and chloroplast [2, 33-35]. And we have first applied the threshold to investigate microsatellites in the Y-DNA of the human reference genome (GRCh38, NC_000024.10) at 1 Kbp resolution by Differential Calculator of Microsatellite version 2.0 (DCM 2.0) method, revealing an exact distributional feature of STRs in every local bins of 1 Kbp sequence of the chromosome Y [36]. Herein, we built 24 Microsatellite landscape maps in all 24 chromosomes of the first complete human genome T2T-CHM13 (CHM13), including the positions and motif types of all classified High Microsatellite Density Peaks (HMDPs), Middle Microsatellite Density Peaks (MMDPs), and Low Microsatellite Density Peaks (LMDPs).

2. Materials and methods

2.1. Genome sequences

The complete 24 chromosome sequences of the human reference genome T2T-CHM13v2.0 were collected

from GenBank, and the accession No. of the complete 24 chromosome sequences were listed in Table S2. All chromosome sequences of human reference genome GRCh38.p14 were also obtained from GenBank (Table S2).

2.2. STRs identification

The Imperfect Microsatellite Extractor (IMEx 2.1) [37] was applied to identify STRs in all the sequences of T2T-CHM13v2.0 and GRCh38.p14 genomes. The threshold for extracting perfect tandem repeats was set at iterations of 6, 3, 3, 3, 3, 3 for mono-, di-, tri-, tetra-, penta- and hexa- repeat motif respectively.

2.3. Calculate local microsatellite density at 1 kb resolution

The local microsatellite density of 24 complete chromosome sequences of T2T-CHM13 were counted by the new promoting program of Differential Calculator of Microsatellites version 3.0 (https://github.com/zhongyangtan/DCM.git), which calculate the local STRs density by divided the every chromosome sequence into a large number of differential-units (bins) with size of 1kb, so microsatellite position-related Differential-unit$_{1kb}$ (D_1) Relative Density (pD_1RD) was calculated for every 1 kb sequence of the 24 chromosome sequences of T2T-CHM13, and also for GRCh38.p14, which was proven to be a better resolution to calculate local STRs density of human genomic sequence [36]. The calculation formula for pD_1RD is:

$$pD_1RD_i = \frac{m_i}{1000bp} \times 1000$$

In this formula, pD_1RD_i is the pD_1RD in the i-th 1 kb bins of the genome, m_i is the size (bp) of microsatellite in the i-th 1 kb bins.

2.4. Identifying Microsatellite Density Peaks

After the microsatellite density in every 1 kb bins, i.e, pD_1RD, of the genome was calculated. The adjacent bins with similar density ranges were merged, and each merged density bin is referred to as a peak. When the pD_1RD values of each bin in adjacent bins is greater than or equal to 90 and less than 150 ($150 > pD_1RD \geq 90$), those adjacent bins was identify as a Low Microsatellite Density Peak (LMDP); the pD_1RD values of each bin in adjacent bins is greater than or equal to 150 and less than 300 ($300 > pD_1RD \geq 150$),

those adjacent bins was identified as a Middle Microsatellite Density Peak (MMDP); and the pD_1RD values of each bin in adjacent bins is greater than or equal to 300 ($pD_1RD \geq 300$), those adjacent bins was identify as a High Microsatellite Density Peak (HMDP).

2.5. Building the STRs density landscape maps

A series of pD_1RD values were obtained by T2T-CHM13 chromosome sequences by the DCM v3.0. Maps displayed the series of pD_1RD values from telomere of p-arm to telomere of q-arm of every chromosome by using the UCSC Genome Browser's bedGraph format. Then, 24 landscape maps of STRs density for T2T-CHM13 were built in the viewer of UCSC Genome Browser window, in which a full chromosomal view of local STRs density is shown for every bin of 1 kb sequence of the chromosome from telomere of p-arm to telomere of q-arm. 24 landscape maps of STRs density for GRCh38.p14 were also built.

2.6. Sorting Microsatellite Density Peaks

High Microsatellite Density Peaks (HMDPs) ($pD_1RD \geq 300$), Middle Microsatellite Density Peaks (MMDPs) ($300 > pD_1RD \geq 150$) and Low Microsatellite Density Peaks (LMDPs) ($150 > pD_1RD \geq 90$) were sorted by the new developed program Microsatellite Density Peaks Sorter version 1.0 (MDPS.v1.0, https://github.com/zhongyangtan/MDPS.git). And the MDPS.v1.0 also further sorted HMDPs, MMDPs and LMDPs into different Main Motif Type (MMT) and sub-Main Motif Type (sub-MMT). (Table S5, S6).

2.7. Mapping Microsatellite Density Peaks

The sorted data of exact position, corresponding motif type and peak name of HMDPs, MMDPs and LMDPs in every chromosome were transferred by the UCSC Genome Browser's bed format and displayed under the landscapes with three tracks of HMDPs, MMDPs and LMDPs in the viewer of UCSC Genome Browser window.

2.8. Comparison of HMDPs between T2T-CHM13.v2.0 and GRCh38.p14

The similar HMDPs between T2T-CHM13.v2.0 and GRCh38.p14, were firstly determined allelic position

by MUMmer v3.23 [38], then comparing motif type by new developed program Genome HMDPs Comparator version 1.0 (GHC.v1.0, https://github.com/zhongyangtan/GHC.git), finally, aligned by ClustalX v2.1 [39].

2.9. The division of Genomic Regions

The full genome of T2T-CHM13 was divided into different Genome Regions according to local microsatellites density features (Table S9). The Genomic Regions with the extremum pD_1RD value ≥ 300 between bins were classified as High Variable Microsatellite Density Region (HVMD-R), those with (300 > the extremum pD_1RD value ≥ 150) between bins were classified as Middle Variable Microsatellite Density Region (MVMD-R), and those with (150 > the extremum pD_1RD value ≥ 90) between bins were classified as Low Variable Microsatellite Density Region (LVMD-R) subclass. The Genomic Regions with the extremum pD_1RD value < 90 between bins were classified as relatively Even microsatellite density (E-) Region class. Genomic Regions clustering with single dominant motif type of peaks were classified as Peak Cluster (PC-) Region class and Telomere Repeat (T-) Region class (located in the telomere regions).

3. Result

3.1. Microsatellite Density landscapes

The recently released complete assembles of T2T-CHM13 including all 22 autosomes, chromosome X and Y, comprise of 3,117,275,501 bp of DNA; and we obtained 22,198,470 microsatellites with total size of 183,133,984 bp and microsatellites relative density (RD) is 58.75 in the full genome (Table 1-a, Table S2 & S3). The widely used method of relative density for analyzing microsatellites was proved limited for analyzing STRs in big sequence like human genome; therefore, to explore the exact distributing feature, the microsatellites landscapes at 1 Kbp resolution, were formerly comprehensively surveyed in the Y-DNA of reference human GRCh38 by the differential calculator of microsatellites [36]. Herein, STRs of the complete

reference human genome T2T-CHM13 including all chromosomes were investigated by the differential calculate method, in which the values of position related relative density at 1 kilo-base resolution (pD_1RD) were calculated in every bin unit of 1kb DNA sequence, and the pD_1RD value was suggested to be possibly a better way to estimate the local microsatellite density variation in human genome [36]. The bedgraph format tool of UCSC Genome Browser was used to map the exact STRs distribution features with the pD_1RD data of the full T2T-CHM13 genome, a series of panoramic landscapes of microsatellites density were obtained to precisely exhibit the local microsatellites density in every bin of 1 kb sequence for the 22 autosome, chromosome X and Y of T2T-CHM13 in viewers of Genome browser window (Fig. 1, Fig. 2, Fig. S1, Fig. S2, Table S1). The 24 landscapes revealed that the relative density in every 1 kb sequence (the pD_1RD value) varies enormously in different site along the 24 chromosomes even from 0 to 970, and the STRs is prone to accumulate to a large numbers of different extent microsatellite density peaks; these density peaks are found genome-wide, and were classified into 8,823 high microsatellites density peaks (HMDPs) ($pD_1RD \geqq 300$), 35,257 middle microsatellites density peaks (MMDPs) ($300 > pD_1RD \geqq 150$) and 199, 649 low microsatellites density peaks (LMDPs) ($150 > pD_1RD \geqq 90$) (Table 1-b, Table S4.01-S4.06); and the bed format tool of UCSC Genome Browser was used to display the exact position and motif types of these HMDPs, MMDPs and LMDPs (Fig. 2, Fig. S2). Similarly, the microsatellites landscapes were comprehensively surveyed in the reference human genome GRCh38 as comparison (Table 1-b, Table S14 & S15, Fig. S12).

3.2. High Microsatellite Density Peaks (HMDPs)

The local relative density values ($pD_1RD \geqq 300$) of those HMDPs are approximately six times or more of the average relative density (RD=58.75) of the full T2T-CHM13 genome, the significant statistical bias of microsatellites relative density in local genome region hints their importance to human genome (Fig. 3-A, Fig. S2.01). It was found that all these HMDPs occur genome widely in all the 24 chromosomes of the

T2T-CHM13, especially pervade in the two arms of all chromosomes at different intervals, except absent at most centromeric region and some pericentromeiric region. The largest quantity of HMDPs was found in ChrY with 2,697 HMDPs, secondly in Chr13, the minimum number of HMDPs in Chr19 with 127 HMDPs identified, and only 305 HMDPs were found on chr1 with longest sequence; therefore, the distribution of HMDPs is not related to chromosome size directly. Furthermore, 89.26% of the HMDPs are identified in intergenic, 10.74% in intron region, but no HMDPs in exon region (Fig. 3-B-a, Table S10.01).

3.3. Main Motif Types (MMTs) of HMDPs

Analysis of the composition of the motifs in every HMDP demonstrated that same type of motif is prone to accumulate in the same HMDP, and therefore, we classified these HMDPs by the main motif composition in each HMDP; for example, when the AT motif account for the main motif composition in a HMDP, it was named $(AT)_n$ Main Motif Type (MMT) HMDP (Table S4.). And the MMT HMDPs can also be further classified into three sub-types: the high-percentage Main Motif Type (hMMT) HMDPs (main motif ≧ 66.7%), the middle-percentage Main Motif Type (mMMT) HMDPs (66.7% > main motif ≧ 50%), and the low-percentage Main Motif Type (lMMT) HMDPs (50% > main motif ≧ 33.4%) (Table 1-c, Table S4-S8). All 8,823 HMDPs be classified into Main Motif Type (MMT) class and Mix Motif Type class. Though the microsatellites generally comprising tandem repeats with motifs of 1-6 bp makes that there are possible total 964 motif types (Table S16), we actually identified only 87 MMTs from 8819 HMDPs belong to MMT class, and these 87 MMTs were further classified into 8 subclasses: mono-, di-, tri-, tetra-, penta-, hexa-, AG- and TC-MMT subclass. The di-MMT subclass includes 6944 HMDPs is the most abundant HMDPs subclass, secondly is the penta-MMT subclass. Notably, 86.1% of the total 8823 HMDPs are the hMMT HMDPs (Fig. 3-B-b,Table S10.01), suggesting that high percentage of same repeat motif is prone to accumulated into same high microsatellites density peak (HMDP) in the complete reference human genome T2T-CHM13. Moreover, many MMT HMDPs usually appears in pairs with motif reverse complementary in the complete

genome like the example of 386 $(AATGG)_n$ MMT HMDPs reversely complementary pairing with 311 $(CCATT)_n$ MMT HMDPs (Table 1-c).

3.4. $[AT]_n^h$ motifs type are the most abundant HMDPs

Among the all hMMT HMDPs, 4828 HMDPs are $[AT]_n^h$ motif type, representing 63.5% of the di-MMT subclass HMDPs and also 54.7% of all 8823 HMDPs, therefore, it is the most abundant hMMT HMDPs (Table 1-c, Fig. 3-C, Table S10.02). These $[AT]_n^h$ HMDPs were classified into two distributional patterns by average distance (AD): sparse-distributional (AD: 1,434,206 bp) and dense-distributional (AD: 4,476 bp) patterns, there are 1886 $[AT]_n^h$ HMDPs scatter sparsely and 2942 $[AT]_n^h$ HMDPs array densely in the full T2T-CHM13 genome (Fig. 3-D-a,b,c, Table S10.03, Fig. S3.01). As a whole, the distributional features can be summed up to 3 $[AT]_n^h$ HMDPs chromosomal distribution models, one model is that only $[AT]_n^h$ HMDPs sparse-distribution appear in the two arms of the 19 chromosomes, another is that $[AT]_n^h$ HMDPs dense-distribution occur in the short p-arm and sparse-distribution in the long q-arm as found in the five acrocentric chromosomes, the other is the $[AT]_n^h$ HMDPs distribution in the Chr Y which 35 $[AT]_n^h$ HMDPs with sparse-distribution are found in Chr Y euchromatin region and 2575 $[AT]_n^h$ HMDPs with dense-distributions in Chr Y heterochromatin region. However, the $[AT]_n^h$ HMDPs is abundantly representing 54.7% of all HMDPs, but 21.3% of all MMDPs and only 2.7% of all LMDPs (Fig. 3-D-d), suggesting that AT motif tandem repeats is easy to accumulate at high density in the complete reference genome. And $(AT)_n$ repeats aggregation are reported to related to chromosomal structure and rearrangements [40-42].

3.5. The $[AATGG]_n^h$ and $[CCATT]_n^h$ penta-hMMT HMDPs account for about half sequence size of all hMMT HMDPs

There are 773 penta-motif HMDPs identified in the full T2T-CHM13 genome, and ranks secondly in all 8 subclasses of the MMT HMDP class; however, the reverse complementary $[AATGG]_n^h$ and $[CCATT]_n^h$

hMMT HMDPs are found 331 and 252 in numbers accounting for 75.4% of all penta-hMMT HMDPs (Table 1-C, Table S4 & S5). Though the large majority of the two types are small in size that is lower than 5 bins (5 kbp); the HMDPs with size that is larger than 5 bin, are mainly $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs (Fig. 3-C-b & c, Table S10.02); therefore, total size of $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs are 3789 and 3362 bins respectively, and combination of two hMMT HMDPs accounting for about 43.55% of the size of all hMMT HMDPs, implied their importance to human genome structure. Analysis of occurred locations of the $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs reveals 4 distributional models about the two hMMT HMDPs in the T2T-CHM13 genome: (a) no $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs distribution occur in Chr 6, 8, 11, 12, 18, 19 and X, (b) both the two hMMT HMDPs appear in short arm of the five acrocentric chromosomes, (c) pericentric distribution of the two hMMT HMDPs in Chr 1, 2, 3, 4, 5, 7, 9, 10, 17 and Y, (d) the $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs are found on the main arms of the chromosomes (on arm distribution) that are majorly in Chr 1, 2, 16, 20 and 22) (Fig. 3-E-a, Table S10.04, Fig. S3.02). Remarkably, Chr 9 is clustered a large segment of $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs at the pericentromeric region of q-arm (Fig. 3-E-b), this region well corresponds to the heterochromatic regions consisting of Classical Human Satellite III [43, 44].

3.6. Comparison of HMDPs between T2T-CHM13 and GRCh38

The microsatellite density landscapes of 24 chromosomes of the reference human genome GRCh38.p14 were also built here, As the reference genome GRCh38 contained gaps and collapsed tandem repeats (Table 1-a, Table S14), the total sequence assembly of GRCh38.p14 is 2,937,639,396 bp, with unsequenced gap size of 150,630,436 bp. And only 3,480 HMDPs, 26,310 MMDPs and 178,848 LMDPs were identified in the microsatellite density landscapes of GRCh38.p14 genome (Table 1-b, Table S14); so HMDPs in GRCh38 genome are shown to be greatly different from that of T2T-CHM13, but MMDPs and LMDPs in GRCh38 genome are relatively close to those in T2T-CHM13 (Table 1-b). A detailed comparison of HMDPs in microsatellite density landscapes between full T2T-CHM13 and GRCh38 full genome, revealed that 1233 and 1290 are high and low-middle similar corresponding to those HMDPs of GRCh38 respectively (Fig. 3-F-b,c,d, Table S15, Fig. S12), above 2,000 HMDPs correspond to middle/low microsatellite density peaks (M/LMDPs) and no peaks region of GRCh38, more than half of HMDPs of T2T-CHM13 are corresponding to un-sequenced gaps region of GRCh38 (Fig. 3-F-d, Table S15); suggesting that HMDPs alleles variety possibly contribute a lot to the individual diversity of human genome in spite of that the no-sequenced gap region may influence the comparing result.

3.7. Genomic Regions divided by local microsatellites density features

Because the landscapes graphed by the differential calculating microsatellites density method, the STRs density distribution feature can be visualized in every 1 kb sequence local genomic region; we discovered from the landscapes that the full genome of T2T-CHM13 may be divided into 514 different microsatellites density related Genomic Regions (Table 1-d, Fig. 4-A, Fig. S4, Table S9). These Genomic Regions comprise of 4 Genomic Region class: Variable microsatellite density (V-) Region class, Even microsatellite density (E-) Region class, Peaks Cluster (PC-) Region class and Telomere repeat (T-) Region class. The V- Region class was further divided into High Variable Microsatellite Density Region (HVMD-R), Middle Variable Microsatellite Density Region (MVMD-R) and Low Variable Microsatellite Density Region (LVMD-R) subclass, the 154 Regions of this class principally correspond to gene coding regions and also some satellite distribution region (Table 1-d). The E- Region class includes relatively Even and Average Microsatellite Density Region (EAMD-R) subclass and relatively Even and extreme Low Microsatellite Density Region (EeLMD-R) subclass, 67 Regions of this class mainly match α satellites and other satellites. The PC- Region

class are subdivided into di-, tri-, tetra-, and penta- subclasses, Regions of this class showed that special motif type of Microsatellite Density Peaks (MDPs) usually gather in proximity to form peaks cluster. And the hexa-nucleotide motif microsatellite cluster Region class was discovered only locating in the telomeric region of T2T-CHM13, the 24 $(TTAGGG)_n$-Rs were found in the telomere of q-arm of all chromosomes, as well as the reverse complementary 23 $(CCCTAA)_n$-Rs were found in the telomere of p-arm of all chromosomes except the $H(AATGG)_n$-R in the p-arm of Chr13 (For simplicity, the Region name abbreviation are only shown here and below, Region name nomenclature please refer to Table S9).

3.8. The broad High Variable Microsatellite Density Region

The microsatellite density landscape maps display that the density of microsatellites on the main arms of most chromosomes has a huge rhythmic change, showing that HMDPs, MMDPs and LMDPs appear alternately, such regions are summarized as broad High Variable Microsatellite Density Regions (b-HVMD-Rs) with every genomic region size more than 10,000,000 bp (Fig. 4-A, Table S9). Total 42 b-HVMD-Rs were classified in the HVMD-R subclass, all chromosomes contain two b-HVMD-Rs in main central part of two arms except that the five acrocentric chromosomes include only one b-HVMD-R in each of their q-arm and ChrY include one in p-arm (Fig. 4-B-a, Table S9). Comparing all b-HVMD-Rs with CAT/Liftoff Genes, RefSeq mRNAs and CenSat Annotation tracks, demonstrated that the b-HVMD-Rs well corresponding to broad gene coding regions in every chromosome (Table 1-d, Table S9). The sum size of the 42 b-HVMD-Rs is 2,850,947,000 bp, representing 91.46% of the full sequence size of T2T-CHM13 genome; but only 4118 of the 8823 HMDPs located in the intergenic and intron region of all b-HVMD-Rs (Fig. 4-B-c, Table S10 & S11).

3.9. Microsatellite Density Peaks Clustering (PC-) Regions

In the 246 Genomic Regions of PC- Region class (Table 1-d, Fig. 4-C, Fig. S6-S9); we clarified 70 Regions in the di-subclass, these Regions were classified into 4 Region types as $HML(CT^{\#})_n$-cR, $HML(AT)_n$-cR,

M$(AT)_n$-cR and ML$(AT)_n$-cR; the 5 HML$(CT^{\#})_n$-cR are found in the near central part of short arm of every the five acrocentric chromosomes and correspond to the five rRNA genes coding regions; the 53 HML$(AT)_n$-cR occur in the five short arm of the five acrocentric chromosomes and the heterochromatin region of ChrY; and the 3 M$(AT)_n$-cR and 9 ML$(AT)_n$-cR appear mainly in the five short arm of the five acrocentric chromosomes and centromeric regions of Chr 3 and 4. Secondly, 13 Regions in the tri-subclass were clarified into 4 Region types as ML$(ATC)_n$-cR, L$(ATC)_n$-cR, ML$(GAT)_n$-cR and L$(GAT)_n$-cR, which are mainly presenting at the pericentromeric regions of chr1, 2, 7, 10, 16, and 17. Then, only 5 Regions were identified in the tetra-subclass as 4 Region types as ML$(AAGG)_n$-cR, L$(AAGG)_n$-cR, HML$(AGCC)_n$-cR and HML$(GGCT)_n$-cR, occurring at Chr2, 3 and 22. At last, we sorted out 158 Regions in the penta-subclass as 10 Region types; therein 8 Region types included 156 Regions containing the reverse complementary motif of $(AATGG)_n$ or $(CCATT)_n$, were found at almost all chromosomes except chr6, 8, 11, 12, 16, 19, and X, and this is consistent with the observation of aforesaid no $[AATGG]_n^h$ and $[CCATT]_n^h$ motif type HMDPs distribution; the other two Region types include only two Regions also containing another two reverse complementary motif of $(ACTCC)_n$ and $(GGAGT)_n$, are only found at the short arm of the acrocentric chromosome 14 and 15. The comparing with CenSat Annotation tracks in UCSC genome browser reveals that di-, tri-, tetra-, and penta- subclasses mainly correspond to diverse classical human satellite sequences region (Table 1-d, Table S9).

3.10. HML$(AT)_n$-cR and ML$(AATGG)_n$-cR alternatively connecting to form the Y heterochromatin

Previously we have constructed the landscapes of the incomplete Y-DNA sequence of human genome GRCh38, which includes many unsequenced gaps especially a big gap of more than 30,000,000 bp in the heterochromatin [36]. Herein, we made the landscape of Y-DNA again but for the first complete sequence of human genome T2T-CHM13 [29], illustrating the local microsatellite relative density feature of the Y heterochromatin (Fig. 4-D), in which Genomic Region HML$(AT)_n$-cR clustering of $(AT)_n$ HMDPs, MMDPs

and LMDPs alternatively link to Region ML(AATGG)$_n$-cR clustering (AATGG)$_n$ MMDPs and LMDPs; therefore, the Y heterochromatin contains 43 HML(AT)$_n$-cRs and 36 ML(AATGG)$_n$-cRs alternatively connected at intervals (Fig. 4-D), and only 2 ML(CCATT)$_n$-cRs and 5 L(AATGG)$_n$-cRs replace ML(AATGG)$_n$-cRs to separate HML(AT)$_n$-cRs. The above-mentioned densely distributing 2575 $[AT]_n^h$ HMDPs scatter among the 43 HML(AT)$_n$-cRs and jaggedly separated by (AATGG)$_n$ MMDPs and LMDPs. Comparing the landscape track with CenSat Annotation track in UCSC genome browser showed that HML(AT)$_n$-cRs and ML(AATGG)$_n$-cRs well correspond to alternating pattern of classical human satellite 1B (hsat1B) and classical human satellite 3 (hsat3) in the heterochromatin Yq12 region [29] (Fig. 4-D, Fig. S10).

3.11. Similar local microsatellites density features in short arms of the five acrocentric chromosomes

The landscapes showed that local microsatellites density features in the five short arm of the five acrocentric chromosomes is more complicate than other autosomes. A palisade arranging HML(CT$^{\#}$)$_n$ cluster Region situates in central or near central part of the short arm of these chromosomes, a set of HML(CCATT)$_n$-cR, H(AATGG)$_n$-cR, M(AT)$_n$-cR, HML(AATGG)$_n$-cR array in left side, 2HML(CCATT)$_n$-cR, one HML(AT)$_n$-cR and several other Regions lie orderly at the right side of the central HML(CT$^{\#}$)$_n$-cR; though these Regions are different in length, they arrange in similar order (Fig. 4-E, Fig. S4 & S5). Comparing with CAT/Liftoff Genes, RefSeq mRNAs and CenSat Annotation tracks, revealed that the HML(CT$^{\#}$)$_n$-cR well correspond to the rRNA gene cluster Regions. Moreover, there are also other PC- Regions arraying in either side of the HML(CT$^{\#}$)$_n$-cR. In a ward, the local microsatellites density features arraying in the landscapes are very similar in the five short arms of the acrocentric chromosomes.

3.12. Even and Low microsatellites density mainly distribute in the centromere

Although the centromeres are known as comprised of large arrays of tandem repeated alpha satellite [27] and most Genomic Regions of the T2T-CHM13 genome exhibit large variations local STRs density (ie. the

pD_1RD value) in the microsatellite density landscape maps, the centromeric regions showed that the feature of local relative microsatellite density is relatively even and low (Fig. 4-F, Fig. S10)；the average pD_1RD of microsatellite in all centromeric regions is 42 and far lower than average relative microsatellite density with value of 58.75 in the full genome, the lowest average pD_1RD is only 12.3 in centromere of Chr2, the average pD_1RD value of centromeres in 11 chromosomes (chr18, 11, 2, 15, 17, X, 8, 22, 20, 9, 14) are lower than 30, the average pD_1RD value of centromeres in 6 chromosomes (chr19, 1, 6, 12, 13, 21) are between 30 to 42 (the average pD_1RD of all centromeres), and 7 centromeres in chromosomes (chr 4, 3, 7, 5, 10, 16, Y) with the average pD_1RD value higher than 42. Two $M(AT)_n$-cRs are embedded in the landscapes of centromere in Chr 3 and 4, and cause the average pD_1RD of centromeres rise to 138.4 and 60.9 respectively.

3.13. $(CT^{\#})_n$ HMDPs groups separate rRNA genes in the five acrocentric short arms

As aforesaid, the five short arm of the acrocentric chromosomes contain five $HML(CT^{\#})_n$-cRs in their central part, series HMDPs group with $(CT)_n$ as main motif were found clustering palisadingly and followed by several MMDPs and LMDPs, forming a tandem Duplication Segment Unit (DSU), these DSUs connects tandemly composing these $HML(CT^{\#})_n$-cRs. The size of these DSUs are between 43000-47000 bp considered as rDNA array [24], and there are 76 DSUs found in the short arm of Chr13, 16 DSUs in Chr14, 50 DSUs in Chr15, 56 DSUs in Chr21, 21 DSUs in Chr22, and total 219 DSUs. Further comparison show that every $HML(CT^{\#})_n$-cR comprise of the rRNA genes Groups region and the $[CT^{\#}]_n$ HMDPs Groups region, remarkably, the $[CT^{\#}]_n$ HMDPs Groups region well one-by-one separate every the rRNA genes Group in every DSU of the five $HML(CT)_n$-cR Regions (Fig. 4-G-a, Fig. S5). Alignment results showed that the identities among the 219 DSUs, 219 rRNA genes Groups and 219 HMDPs Groups are higher than 0.82, 0.99 and 0.68 respectively (Fig. 4-G-b & c, Fig. S5), suggesting the $[CT^{\#}]_n$ HMDPs Groups region may functionally separate the rRNA genes region but be more variable than rRNA genes region.

4. Discussion

This work built the first comprehensive genome wide microsatellite density landscape maps of the first complete human reference genome T2T-CHM13 [24, 29]. These landscape maps exhibited that microsatellites aggregate in many genomic positions to form a large number of microsatellite density peaks, and these peaks array along the every human chromosomes arranged like notes of a beautiful piece of music; therefore, the 24 microsatellite density landscapes together look like to form a symphony of human life (Fig. 2). Though transcriptional and epigenetic state of human repeat elements were comprehensive analyzed [5], we mainly focus on the short tandem repeats of human genome, and our works revealed that the microsatellite density landscapes are compatible with human chromosome structure at a high extent, suggesting these landscapes possibly contribute to the further exploration the relationship between STRs aggregation and human genome structure, variation, evolution and so on.

The high microsatellites density peak (HMDP) aggregated predominantly by a single motif, is the most statistically significant short tandem repeats aggregation phenomenon. We identified 8823 HMDPs in T2T-CHM13, much more than the number of 3480 HMDPs in GRCh38; the HMDP number was substantially increased in T2T-CHM13, mainly corresponding to unassembled regions in GRCh38, especially the heterochromatin of Chr Y and the five short arms of the acrocentric Chromosomes. Almost all the HMDPs are single motif dominated, particularly, 86.1% of the 8823 HMDPs are high percentage dominated by a single motif. And there should be total 964 Main Motif Type of HMDPs ranging from mono- to hexa- main motif types, actually, we only identified 87 single motif dominated Main Motif Types (MMTs); these statistical significances are strongly implied the importance of aggregation of short tandem repeats on gene expression, regulation, epigenetics, genetic architecture and evolution of human genome. What is more, the $(AT)_n$ single motif dominated Main Motif Types (MMTs) account for more than half of the total number of HMDPs, and the reverse complementary motif $(AATGG)_n$ and $(CCATT)_n$ dominated two Main Motif

Types (MMTs) account for approximately half of the total bins of HMDPs; suggesting that the three MMT HMDPs are worthy to be further explored for their roles in gene function and chromosome structure of human genome.

In addition, we classified 514 characterized by microsatellite density in the complete reference genome, the Variable microsatellite density (V-) region class almost overlap the gene coding region, the Even microsatellite density (E-) region class mainly cover centromere, pericentromere and p-arm of acrocentric chromosomes, especially, the m-EAMD-R and m-EeLMD-R well correspond to centromeres (Table S9.01); the 246 PC-regions generally locate in pericentromere, p-arm of acrocentric chromosomes and heterochromatin of Chr Y; and the reverse complementary repeat motif $(CCCTAA)_n$ and $(TTAGGG)_n$ telomere repeat regions almost dominate in p-arms and q-arms respectively. Meanwhile, similar genomic regions structural characters were observed in the five acrocentric short arms by the view of microsatellite density features, and alternative microsatellite density regions structure observed in heterochromatin Chr Y [29]. all these indicate that local regional microsatellite density variation may be related to human genome structure. Furthermore, most (microsatellite density) peaks cluster (PC-) regions correspond to different Classical human satellites (hsat); but most the even microsatellite density regions especially in centromere, where microsatellite density are limited to very low and even, correspond to different α satellite higher-order repeats (αSat hor) (Table 1-d); suggesting that although microsatellites (STRs) are highly related to satellites (long tandem repeats), the mechanism of their occurrence are likely completely different.

The 219 near identical Duplication Segment Units (DSUs) of approximately 45-kbp that contain segments of encoding 45S rRNAs, are embedded in the short arms of the 5 acrocentric chromosomes [5]; we identified that the HMDPs also well separate every 45S rRNA gene of about 13300 bp. All the HMDPs were classified here in the intergenic and intron regions, implying that the HMDPs possibly own the basic role of separating genes and exons in the genome and may be very worthwhile to further explore. The first complete

reference T2T-CHM13 provided us the opportunity to study short tandem repeat sequences in human genome comprehensively, our results of microsatellite density landscape maps revealed that short tandem repeats are tend to aggregate characteristically throughout the full genome of T2T-CHM13, it may be very helpful to deepen the exploring of the mysterious roles of tandem repeats to human genome's structure, evolution, regulation, variation and also human diseases.

5. Conclusions:

These landscape maps exhibited that microsatellites aggregate in many genomic positions to form a large number of microsatellite density peaks with composing of mainly single motif type in the complete reference genome, indicating that the local microsatellites density varies enormously along the every chromosome of T2T-CHM13.

Declarations

Abbreviations

STRs: Short Tandem Repeats;

HMDPs: High Microsatellites Density Peaks;

MMDPs: Middle Microsatellites Density Peaks;

LMDPs: Low Microsatellites Density Peaks;

MDPs: Microsatellite Density Peaks;

DCM 2.0: Differential Calculator of Microsatellites version 2.0;

IMEx 2.1: Imperfect Microsatellite Extractor version 2.1;

pD_1RD: position-related D_1-Relative Density;

MDPS.v1.0: Microsatellite Density Peaks Sorter version 1.0;

MMT: Main Motif Type;

sub-MMT: sub-Main Motif Type;

GHC.v1.0: Genome HMDPs Comparator version 1.0;

RD: Relative Density;

hMMT: high-percentage Main Motif Type;

mMMT: middle-percentage Main Motif Type;

lMMT: low-percentage Main Motif Type;

AD: Average Distance;

V-region class: Variable microsatellite density region class;

E-region class: Even microsatellite density region class;

PC-region class: Peaks Cluster region class;

T-region class: Telomere repeat region class;

HVMD-R: High Variable Microsatellite Density Region;

MVMD-R: Middle Variable Microsatellite Density Region;

LVMD-R: Low Variable Microsatellite Density Region;

EAMD-R: Even and Average Microsatellite Density Region;

EeLMD-R: Even and extreme Low Microsatellite Density Region;

DSUs: Duplication Segment Units;

b-HVMD-Rs: broad High Variable Microsatellite Density Regions;

hsat: human satellites;

αSat hor: α satellite higher-order repeats.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The Web version of microsatellite density landscape maps of 24 chromosomes of T2T-CHM13v2.0 can be accessed in (http://genome.ucsc.edu/s/zhongyangtan/CHM13v2.0).

The Web version of microsatellite density landscape maps of 24 chromosomes of GRCh38.p14 can be accessed in (http://genome.ucsc.edu/s/zhongyangtan/GRCh38.p14).

All of the Supplementary Figures and Tables can be accessed in https://drive.google.com/file/d/1DiXYuUnSINpdkfCbZJAyMkPBKIFfwFEg/view?usp=sharing

Competing interests

The authors declare no competing interests.

Funding

National Key Research and Development Program of China (grant No. 19-163-12-ZT-005-004-04, Recipient: Zhongyang Tan).

Authors' contributions

Z.T. designed and directed this study. Z.T., K.O., Y.X., D.L., T.C., Sc.P., H.H., W.Z. and Y.L. prepared the manuscript. T.C., D.L., Sc.P., H.H., Y.F., W.Z. and Z.T. developed the statistical method and programs. Y.X., D.L., T.C., Sc.P., H.H., W.Z., Y.L., Z.P., H.Z., Li.Z., S.P., R.S., X.H., S.Z., W.J., X.Z., X.W., La.Z., J.Z., Q.O., Y.T., X.J., Y.Z., S.T., J.S. and Z.T. performed the data analysis, K.O. and Z.T. edited this manuscript. All authors read and approved the final version of the manuscript.

Acknowledgments

We thank Yongliang Wu (The Hong Kong Polytechnic University), Jingchen Xie, Junyu Zeng and Wenjing Zhou (Hunan University), Yasha He (South Central University), Puxi Tan (Xi'an Jiaotong University) for helping in data analysis.

References

1. Ellegren H. Microsatellites: simple sequences with complex evolution. Nature reviews Genetics. 2004; 5(6):435-445.
2. Zhao X, Tian Y, Yang R, Feng H, Ouyang Q, Tian Y, et al. Coevolution between simple sequence repeats (SSRs) and virus genome size. BMC genomics. 2012; 13:435.
3. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nature reviews Genetics. 2018; 19(5):286-298.
4. Hannan AJ. Repeat DNA expands our understanding of autism spectrum disorder. Nature. 2021; 589(7841):200-202.
5. Hoyt SJ, Storer JM. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science. 2022; 376(6588):eabk3112.
6. Hartl DL. Molecular melodies in high and low C. Nature reviews Genetics. 2000; 1(2):145-149.
7. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell. 2013; 155(4):858-868.
8. Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and characterization of microsatellite instability across 18 cancer types. Nature medicine. 2016; 22(11):1342-1350.
9. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature. 2019; 575(7781):210-216.
10. van Wietmarschen N, Sridharan S, Nathan WJ, Tubbs A, Chan EM, Callen E, et al. Repeat expansions confer WRN dependence in microsatellite-unstable cancers. Nature. 2020; 586(7828):292-298.
11. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nature genetics. 2016; 48(1):22-29.
12. Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y, et al. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Res. 2016; 44(8):3750-3762.
13. Verstrepen KJ, Jansen A, Lewitter F, Fink GR. Intragenic tandem repeats generate functional variability. Nature genetics. 2005; 37(9):986-990.
14. Fondon JW, 3rd, Hammock EA, Hannan AJ, King DG. Simple sequence repeats: genetic modulators of brain function and behavior. Trends in neurosciences. 2008; 31(7):328-334.
15. Hannan AJ. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'.

Trends in genetics : TIG. 2010; 26(2):59-65.
16. Nasrallah MP, Cho G, Simonet JC, Putt ME, Kitamura K, Golden JA. Differential effects of a polyalanine tract expansion in Arx on neural development and gene expression. Human molecular genetics. 2012; 21(5):1090-1098.
17. Willems T, Gymrek M, Highnam G, Genomes Project C, Mittelman D, Erlich Y. The landscape of human STR variation. Genome research. 2014; 24(11):1894-1904.
18. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nature methods. 2017; 14(6):590-592.
19. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016; 538(7624):201-206.
20. Gymrek M, Willems T, Reich D, Erlich Y. Interpreting short tandem repeat variations in humans using mutational constraint. Nature genetics. 2017; 49(10):1495-1501.
21. Levinson G, Gutman GA. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. Nucleic Acids Res. 1987; 15(13):5323-5338.
22. Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids Res. 1992; 20(2):211-215.
23. Zhang H, Li D, Zhao X, Pan S, Wu X, Peng S, et al. Relatively semi-conservative replication and a folded slippage model for short tandem repeats. BMC genomics. 2020; 21(1):563.
24. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022; 376(6588):44-53.
25. Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, et al. Epigenetic patterns in a complete human genome. Science. 2022; 376(6588):eabj5089.
26. Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental duplications and their variation in a complete human genome. Science. 2022; 376(6588):eabj6965.
27. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, et al. Complete genomic and epigenetic maps of human centromeres. Science. 2022; 376(6588):eabl4178.
28. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, et al. A complete reference genome improves analysis of human genetic variation. Science. 2022; 376(6588):eabl3533.
29. Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, et al. The complete sequence of a human Y chromosome. Nature. 2023; 621:344-354
30. Lei Y, Zhou Y, Price M, Song Z. Genome-wide characterization of microsatellite DNA in fishes: survey and analysis of their abundance and frequency in genome-specific regions. BMC genomics. 2021; 22(1):421.
31. Qi WH, Yan CC, Li WJ, Jiang XM, Li GZ, Zhang XY, et al. Distinct patterns of simple sequence repeats and GC distribution in intragenic and intergenic regions of primate genomes. Aging. 2016; 8(11):2635-2654.
32. Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. 2003; 4(2):R13.
33. de Freitas KEJ, Busanello C, Viana VE, Pegoraro C, de Carvalho Victoria F, da Maia LC, et al. An empirical analysis of mtSSRs: could microsatellite distribution patterns explain the evolution of mitogenomes in plants? Funct Integr Genomics. 2022; 22(1):35-53.
34. Chen M, Tan Z, Zeng G, Peng J. Comprehensive analysis of simple sequence repeats in pre-miRNAs. Molecular biology and evolution. 2010; 27(10):2227-2232.
35. Sahu BP, Majee P, Singh RR, Sahoo N, Nayak D. Genome-wide identification and characterization of microsatellite markers within the Avipoxviruses. 3 Biotech. 2022; 12(5):113.
36. Li D, Pan S, Zhang H, Fu Y, Peng Z, Zhang L, et al. A comprehensive microsatellite landscape of human Y-DNA at kilobase resolution. BMC genomics. 2021; 22(1):76.
37. Mudunuri SB, Nagarajaram HA. IMEx: Imperfect Microsatellite Extractor. Bioinformatics. 2007; 23(10):1181-1187.
38. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5(2):R12.

39. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23(21):2947-2948.
40. Helmrich A, Stout-Weider K, Hermann K, Schrock E, Heiden T. Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes. Genome research. 2006; 16(10):1222-1230.
41. Irony-Tur Sinai M, Salamon A, Stanleigh N, Goldberg T, Weiss A, Wang YH, et al. AT-dinucleotide rich sequences drive fragile site formation. Nucleic acids research. 2019; 47(18):9685-9695.
42. Inagaki H, Ohye T, Kogo H, Yamada K, Kowa H, Shaikh TH, et al. Palindromic AT-rich repeat in the NF1 gene is hypervariable in humans and evolutionarily conserved in primates. Human mutation. 2005; 26(4):332-342.
43. Ramesh KH, Verma RS. Breakpoints in alpha, beta, and satellite III DNA sequences of chromosome 9 result in a variety of pericentric inversions. Journal of medical genetics. 1996; 33(5):395-398.
44. Starke H, Seidel J, Henn W, Reichardt S, Volleth M, Stumm M, et al. Homologous sequences at human chromosome 9 bands p12 and q13-21.1 are involved in different patterns of pericentric rearrangements. European journal of human genetics : EJHG. 2002; 10(12):790-800.

Figure legends

Figure 1. The landscape of STR Density, HMDP (High Microsatellite Density Peak), MMDP (Middle Microsatellite Density Peak) and LMDP (Low Microsatellite Density Peak) in T2T-CHM13 chromosome 18 with smallest HMDPs. A. Landscape of STR Density is displayed on the first track, and the microsatellite density values are shown on the ordinate. **B.** HMDPs are displayed in pack mode on the second track, including its Motif Type, location (shown in colored vertical line) and name. The HMDP name consists of "Genome name", "chr No.", "HMDP No.", "HMDP size" and "sub-Main Motif Type (sub-MMT)". **C-D.** MMDPs and LMDPs are displayed in dense mode on the third and last track. Chromosome 18 with the lowest number of HMDP, is shown here, and other chromosomes of the T2T-CHM13 are shown in Figure S2.

Table 1. Statistic of microsatellite density landscape information on the human genome T2T-CHM13. a. The assembly information and microsatellite statistics of T2T-CHM13v2.0 and GRCh38.p14. **b.** The list of microsatellites density peaks in T2T-CHM13v2.0 and GRCh38.p14. **c.** The motif types of HMDP in T2T-CHM13v2.0. **d.** Genomic regions characterized by microsatellite density in T2T-CHM13v2.0.

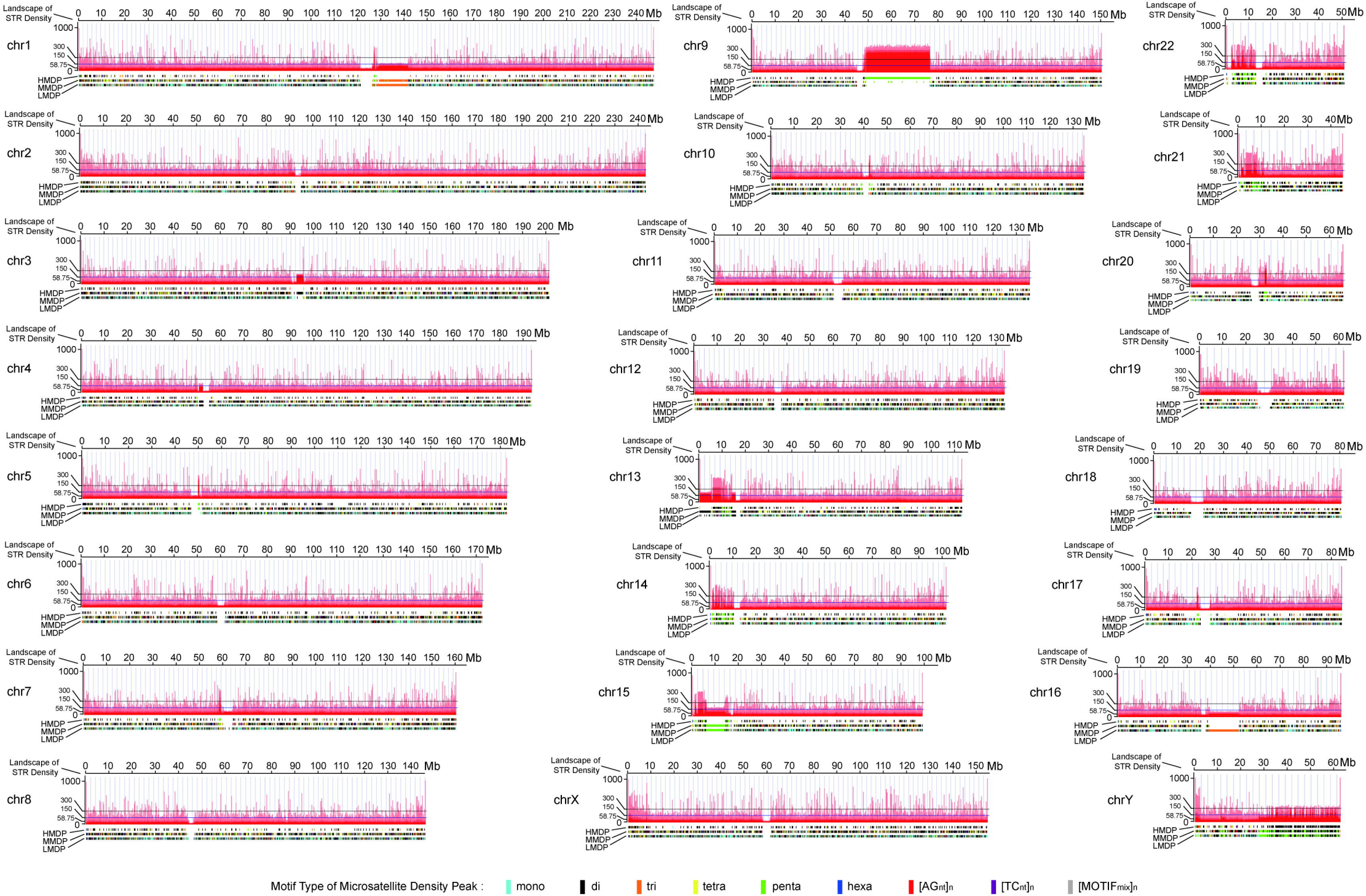
Figure 2. Landscape map of STR Density of 24 chromosomes of T2T-CHM13. Landscape of each chromosome, first track displayed landscape of STR Density, second track displayed HMDP, third track displayed MMDP, last track displayed LMDP.

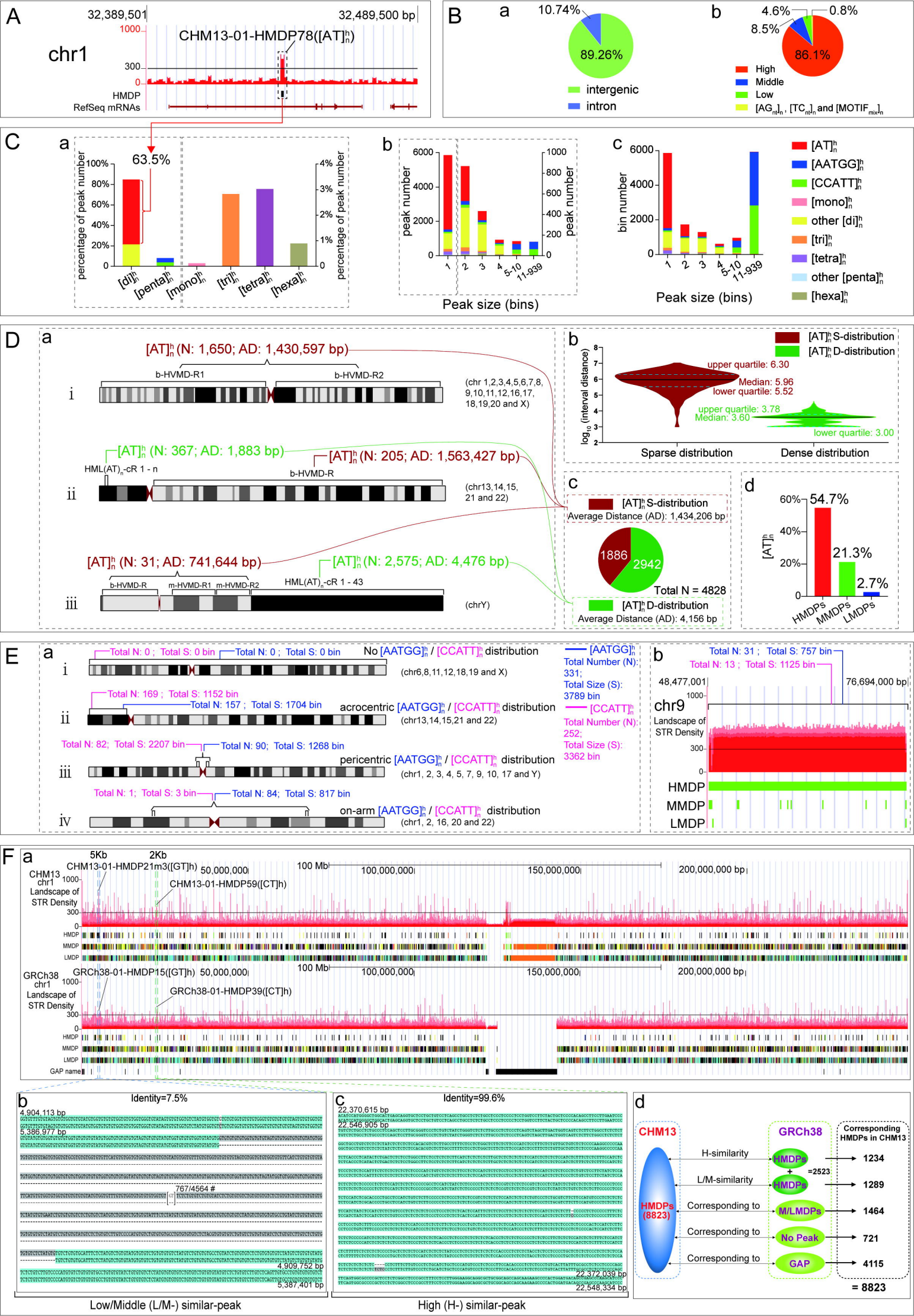
Figure 3. Summary of HMDPs in T2T-CHM13. A. An $[AT]_n^h$ HMDP in the T2T-CHM13. **B.** (a) Ratio of HMDP corresponding to intron and intergenic regions. (b) Ratio of hMMT, mMMT and lMMT HMDPs and other HMDPs. **C.** Statistic of HMDP number and bin, (a) The percentage of peak number of mono-, di-, tri-, tetra-, penta- and hexa- hMMT HMDPs in all hMMT HMDPs; (b, c) The number and total bin number of hMMT HMDP in each Peak sequence size (bins) group. **D.** $[AT]_n^h$ HMDP distribution, (a) The 3 $[AT]_n^h$ HMDP distribution models; (b) The distribution map of interval distance of $[AT]_n^h$ Sparse distribution (S-distribution) and Dense distribution (D-distribution) HMDPs, the ordinate indicated the logarithm base 10 of the interval distance (bp), and the values of the interval distance are listed in Table S10.03; (c) The number of $[AT]_n^h$ Sparse distribution and Dense distribution HMDPs; (d) The percentage of $[AT]_n^h$ HMDPs, MMDPs and LMDPs to all HMDPs, MMDPs and LMDPs. **E.** $[AATGG]_n^{\mathbf{h}}$ and $[CCATT]_n^{\mathbf{h}}$ HMDP distribution, (a) The 4 distributional models of $[CCATT]_n^h$ and $[AATGG]_n^h$ HMDPs; (b) A big cluster of $[AATGG]_n^{\mathbf{h}}$ and $[CCATT]_n^h$ motif type HMDPs at the p-arm of Chr9 . **F.** Comparison of landscapes T2T-CHM13.v2.0 and GRCh38.p14, (a) Comparison in Chr 1;

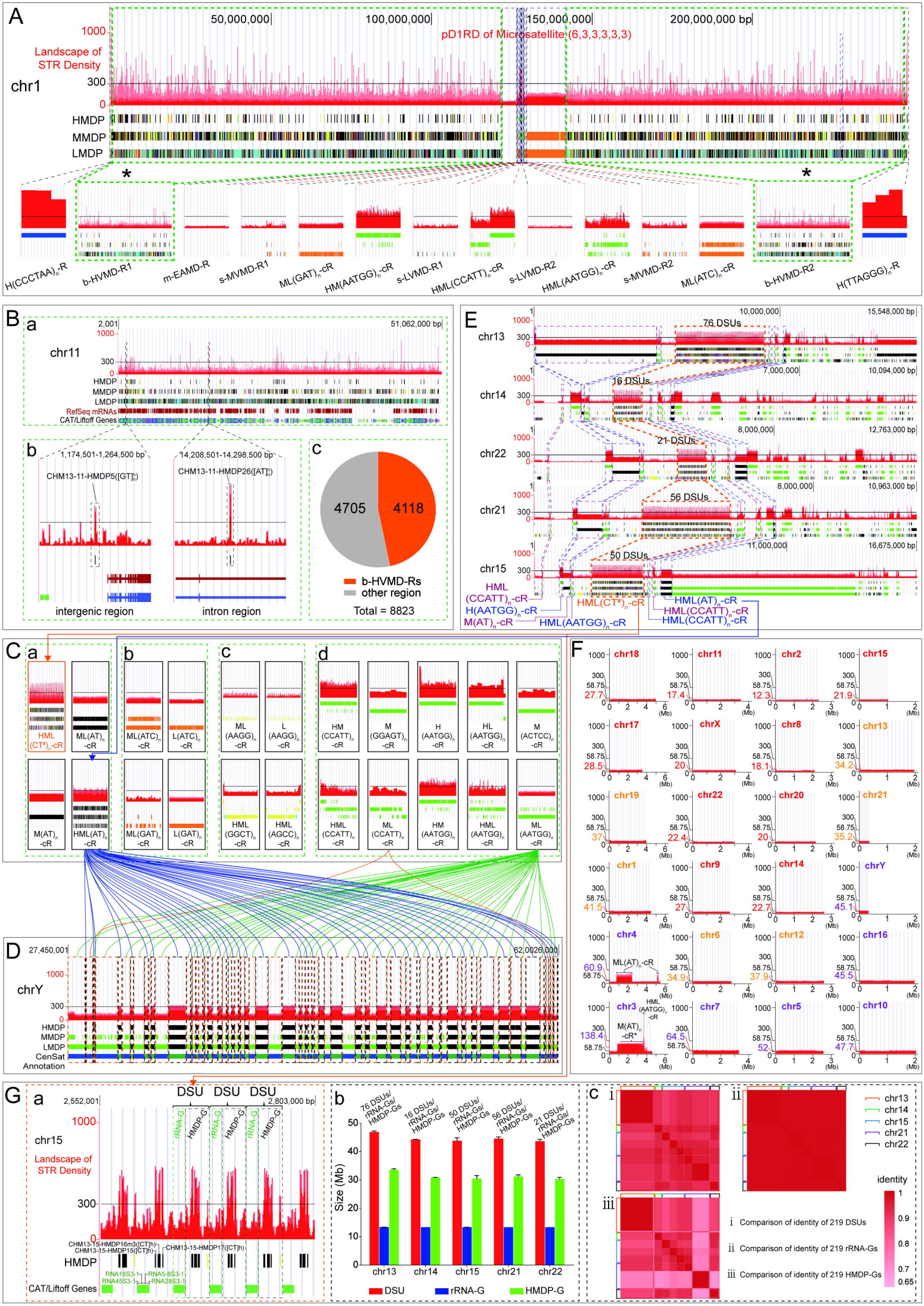
(b) identity of a pair low similar HMDPs; (c) identity of a pair high similar HMDPs; (d) corresponding HMDPs of the T2T-CHM13 corresponding to MDPs of GRCh38.p14.

Figure 4. Summary of Genomic Regions characterized by microsatellite density in T2T-CHM13. A. Chromosome 1 of the T2T-CHM13 characterized into 14 Genomic Regions. **B.** broad High Variable Microsatellite Density Region (b-HVMD-R), (a) example of a b-HVMD-R (chromosome 11, including RefSeq mRNAs and CAT / Liftoff Genes Track). (b) HMDPs corresponding to gene intron and intergenic region; (c) the number of HMDPs in b-HVMD-Rs and other regions. **C.** Feature maps of Peaks Cluster (PC-) Regions (a: di-, b: tri-, c: tetra- and d: penta-subclass). **D.** HML$(AT)_n$-cRs and ML$(AATGG)_n$-cRs alternatively connecting in the Chr Y heterochromatin. **E.** Similar local microsatellites density features in the short arms of acrocentric chromosomes (HML$(CT^{\#})_n$-cR in center). **F.** Microsatellite density feature of centromeres. The ordinate shown average pD_1RD value, the large font number : the average value of pD_1RD of each centromere (colors showing different value ranges), 58.75: the average pD_1RD of full genome. **G.** The $(CT\#)_n$ HMDPs groups separate rRNA genes in the five HML$(CT^{\#})_n$-cR, (a) Partial enlarged view of HML$(CT^{\#})_n$-cR in Chr15; (b) Sequence size of Duplication Segment Units (DSUs), rRNA-Groups (rRNA-Gs) and HMDP-Groups (HMDP-Gs) in HML$(CT^{\#})_n$-cRs; (c) Comparison of identity of DSUs, rRNA-Gs and HMDP-Gs.

T2T-CHM13_chr18
A
Landscape of STR Density
Scale
1000
300
0
20 Mb
hs1
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
45,000,000
50,000,000
55,000,000
60,000,000
65,000,000
70,000,000
75,000,000
80,000,000
B
CHM13-18-HMDP1m2([CCCTAA]h)
CHM13-18-HMDP2([CCCTAA]h)
CHM13-18-HMDP3([CCCTAA]h)
CHM13-18-HMDP4([CCCTAA]h)
CHM13-18-HMDP5([AC]h)
CHM13-18-HMDP6([AGAGGG]h)
CHM13-18-HMDP7([AT]m)
CHM13-18-HMDP8([GATGGT]h)
CHM13-18-HMDP9([AT]h)
CHM13-18-HMDP10([AT]h)
CHM13-18-HMDP11([ACC]h)
CHM13-18-HMDP12m2([GGT]h)
CHM13-18-HMDP13([AT]h)
CHM13-18-HMDP14([AT]h)
CHM13-18-HMDP15([AG]m)
CHM13-18-HMDP16([AT]h)
CHM13-18-HMDP17([CCTT]h)
CHM13-18-HMDP18([GT]h)
CHM13-18-HMDP19([AT]h)
CHM13-18-HMDP20([TCnt])
CHM13-18-HMDP21m2([ATCC]h)
CHM13-18-HMDP22([CCT]h)
CHM13-18-HMDP23([AT]h)
CHM13-18-HMDP24([AAAG]m)
CHM13-18-HMDP25([AT]m)
CHM13-18-HMDP26([GT]h)
CHM13-18-HMDP27([AT]h)
CHM13-18-HMDP28([AG]h)
CHM13-18-HMDP29([AGG]h)
CHM13-18-HMDP30m3([AT]h)
CHM13-18-HMDP31([AT]h)
CHM13-18-HMDP32m3([AT]h)
CHM13-18-HMDP33([AT]m)
CHM13-18-HMDP34([AT]m)
CHM13-18-HMDP35([AT]h)
CHM13-18-HMDP36([AT]h)
CHM13-18-HMDP37([ACC]m)
CHM13-18-HMDP38([AT]h)
CHM13-18-HMDP39([AT]h)
CHM13-18-HMDP40([AAAG]l)
CHM13-18-HMDP41m2([AT]l)
CHM13-18-HMDP42([AT]h)
CHM13-18-HMDP43([AC]h)
CHM13-18-HMDP44m2([AC]h)
CHM13-18-HMDP45m2([C]l)
CHM13-18-HMDP46([AT]h)
CHM13-18-HMDP47([AC]m)
CHM13-18-HMDP48([AT]h)
CHM13-18-HMDP49([AT]h)
CHM13-18-HMDP50([AT]h)
CHM13-18-HMDP51([AT]h)
CHM13-18-HMDP52m3([AT]h)
CHM13-18-HMDP53([AT]h)
CHM13-18-HMDP54([AT]h)
CHM13-18-HMDP55([AT]h)
CHM13-18-HMDP56([GT]h)
CHM13-18-HMDP57([AT]h)
CHM13-18-HMDP58([AT]h)
CHM13-18-HMDP59m2([AT]h)
CHM13-18-HMDP60([GGAT]h)
CHM13-18-HMDP61([AT]h)
CHM13-18-HMDP62m2([AC]h)
CHM13-18-HMDP63([AC]l)
CHM13-18-HMDP64([AT]h)
CHM13-18-HMDP65([AT]h)
CHM13-18-HMDP66([AT]h)
CHM13-18-HMDP67([AT]h)
CHM13-18-HMDP68([CTT]h)
CHM13-18-HMDP69([AT]h)
CHM13-18-HMDP70([AT]h)
CHM13-18-HMDP71([AT]h)
CHM13-18-HMDP72([AT]h)
CHM13-18-HMDP73([CCT]h)
CHM13-18-HMDP74([AT]h)
CHM13-18-HMDP75([GT]h)
CHM13-18-HMDP76([AT]h)
CHM13-18-HMDP77([AT]h)
CHM13-18-HMDP78m2([AT]h)
CHM13-18-HMDP79([GT]h)
CHM13-18-HMDP80([GT]h)
CHM13-18-HMDP81m2([AC]h)
CHM13-18-HMDP82([AT]h)
CHM13-18-HMDP83([AT]h)
CHM13-18-HMDP84m2([AT]h)
CHM13-18-HMDP85([CCTTT]h)
CHM13-18-HMDP86m2([AT]h)
CHM13-18-HMDP87([AAAG]l)
CHM13-18-HMDP88([CTTT]l)
CHM13-18-HMDP89([AT]h)
CHM13-18-HMDP90([AAAG]l)
CHM13-18-HMDP91([CTTT]m)
CHM13-18-HMDP92m2([AT]h)
CHM13-18-HMDP93([AGAT]h)
CHM13-18-HMDP94m2([AT]h)
CHM13-18-HMDP95([AT]h)
CHM13-18-HMDP96([GT]h)
CHM13-18-HMDP97([AT]h)
CHM13-18-HMDP98m2([AT]h)
CHM13-18-HMDP99([AT]h)
CHM13-18-HMDP100m2([AC]h)
CHM13-18-HMDP101([GT]h)
CHM13-18-HMDP102([AT]h)
CHM13-18-HMDP103([AT]h)
CHM13-18-HMDP104([CCCT]h)
CHM13-18-HMDP105([ATCC]h)
CHM13-18-HMDP106m2([ACC]h)
CHM13-18-HMDP107([GGT]h)
CHM13-18-HMDP108([ATCC]m)
CHM13-18-HMDP109([CTGT]h)
CHM13-18-HMDP110([AT]h)
CHM13-18-HMDP111([GGT]h)
CHM13-18-HMDP112m2([CCT]m)
CHM13-18-HMDP113m2([GT]h)
CHM13-18-HMDP114([AT]h)
CHM13-18-HMDP115([AC]h)
CHM13-18-HMDP116([AC]h)
CHM13-18-HMDP117([GT]h)
CHM13-18-HMDP118([AC]h)
CHM13-18-HMDP119m2([AC]h)
CHM13-18-HMDP120([CT]h)
CHM13-18-HMDP121m2([GT]h)
CHM13-18-HMDP122([ATCC]h)
CHM13-18-HMDP123m2([GGT]h)
CHM13-18-HMDP124([GT]h)
CHM13-18-HMDP125m4([TTAGGG]h)
C
MMDP
D
LMDP
Motif Type of Microsatellite Density Peak :
mono
di
tri
tetra
penta
hexa
[AGnt]n
[TCnt]n
[MOTIFmix]n


Landscape of STR Density
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chrX
chr22
chr21
chr20
chr19
chr18
chr17
chr16
chrY
HMDP
MMDP
LMDP
Mb
Motif Type of Microsatellite Density Peak :
mono
di
tri
tetra
penta
hexa
[AGnt]n
[TCnt]n
[MOTIFmix]n


A
32,389,501
32,489,500 bp
chr1
CHM13-01-HMDP78([AT]ⁿ)
HMDP
RefSeq mRNAs
B
a
10.74%
89.26%
intergenic
intron
b
4.6%
0.8%
8.5%
86.1%
High
Middle
Low
[AGnt]n, [TCnt]n and [MOTIFmix]n
C
a
63.5%
percentage of peak number
[di]ⁿ
[penta]ⁿ
[mono]ⁿ
[tri]ⁿ
[tetra]ⁿ
[hexa]ⁿ
b
peak number
Peak size (bins)
c
bin number
Peak size (bins)
[AT]ⁿ
[AATGG]ⁿ
[CCATT]ⁿ
[mono]ⁿ
other [di]ⁿ
[tri]ⁿ
[tetra]ⁿ
other [penta]ⁿ
[hexa]ⁿ
D
a
[AT]ⁿ (N: 1,650; AD: 1,430,597 bp)
b-HVMD-R1
b-HVMD-R2
(chr 1,2,3,4,5,6,7,8, 9,10,11,12,16,17, 18,19,20 and X)
[AT]ⁿ (N: 367; AD: 1,883 bp)
[AT]ⁿ (N: 205; AD: 1,563,427 bp)
HML(AT)n-cR 1 - n
b-HVMD-R
(chr13,14,15, 21 and 22)
[AT]ⁿ (N: 31; AD: 741,644 bp)
[AT]ⁿ (N: 2,575; AD: 4,476 bp)
b-HVMD-R
m-HVMD-R1
m-HVMD-R2
HML(AT)n-cR 1 - 43
(chrY)
b
[AT]ⁿ S-distribution
[AT]ⁿ D-distribution
log10 (interval distance)
upper quartile: 6.30
Median: 5.96
lower quartile: 5.52
upper quartile: 3.78
Median: 3.60
lower quartile: 3.00
Sparse distribution
Dense distribution
c
[AT]ⁿ S-distribution
Average Distance (AD): 1,434,206 bp
1886
2942
Total N = 4828
[AT]ⁿ D-distribution
Average Distance (AD): 4,156 bp
d
[AT]ⁿ
54.7%
21.3%
2.7%
HMDPs
MMDPs
LMDPs
E
a
Total N: 0 ; Total S: 0 bin
Total N: 0 ; Total S: 0 bin
No [AATGG]ⁿ / [CCATT]ⁿ distribution
(chr6,8,11,12,18,19 and X)
Total N: 169 ; Total S: 1152 bin
Total N: 157 ; Total S: 1704 bin
acrocentric [AATGG]ⁿ / [CCATT]ⁿ distribution
(chr13,14,15,21 and 22)
Total N: 82; Total S: 2207 bin
Total N: 90; Total S: 1268 bin
pericentric [AATGG]ⁿ / [CCATT]ⁿ distribution
(chr1, 2, 3, 4, 5, 7, 9, 10, 17 and Y)
Total N: 1; Total S: 3 bin
Total N: 84; Total S: 817 bin
on-arm [AATGG]ⁿ / [CCATT]ⁿ distribution
(chr1, 2, 16, 20 and 22)
[AATGG]ⁿ
Total Number (N): 331;
Total Size (S): 3789 bin
[CCATT]ⁿ
Total Number (N): 252;
Total Size (S): 3362 bin
b
Total N: 31 ; Total S: 757 bin
Total N: 13 ; Total S: 1125 bin
48,477,001
76,694,000 bp
chr9
Landscape of STR Density
HMDP
MMDP
LMDP
F
a
5Kb
2Kb
CHM13-01-HMDP21m3([GT]h)
CHM13-01-HMDP59([CT]h)
CHM13 chr1
Landscape of STR Density
50,000,000
100 Mb
100,000,000
150,000,000
200,000,000 bp
HMDP
MMDP
LMDP
GRCh38-01-HMDP15([GT]h)
GRCh38-01-HMDP39([CT]h)
GRCh38 chr1
Landscape of STR Density
GAP name
b
Identity=7.5%
4,904,113 bp
5,386,977 bp
767/4564 #
4,909,752 bp
5,387,401 bp
Low/Middle (L/M-) similar-peak
c
Identity=99.6%
22,370,615 bp
22,546,905 bp
22,372,039 bp
22,548,334 bp
High (H-) similar-peak
d
CHM13
HMDPs (8823)
GRCh38
Corresponding HMDPs in CHM13
H-similarity
HMDPs
=2523
1234
L/M-similarity
HMDPs
1289
Corresponding to
M/LMDPs
1464
Corresponding to
No Peak
721
Corresponding to
GAP
4115
= 8823


A
chr1
Landscape of STR Density
pD1RD of Microsatellite (6,3,3,3,3,3)
50,000,000
100,000,000
150,000,000
200,000,000 bp
HMDP
MMDP
LMDP
H(CCCTAA)n-R
b-HVMD-R1
m-EAMD-R
s-MVMD-R1
ML(GAT)n-cR
HM(AATGG)n-cR
s-LVMD-R1
HML(CCATT)n-cR
s-LVMD-R2
HML(AATGG)n-cR
s-MVMD-R2
ML(ATC)n-cR
b-HVMD-R2
H(TTAGGG)n-R
B
a
chr11
2,001
51,062,000 bp
RefSeq mRNAs
CAT/Liftoff Genes
b
1,174,501-1,264,500 bp
14,208,501-14,298,500 bp
CHM13-11-HMDP5([GT]n)
CHM13-11-HMDP26([AT]n)
intergenic region
intron region
c
4705
4118
b-HVMD-Rs
other region
Total = 8823
E
chr13
76 DSUs
10,000,000
15,548,000 bp
chr14
16 DSUs
7,000,000
10,094,000 bp
chr22
21 DSUs
8,000,000
12,763,000 bp
chr21
56 DSUs
8,000,000
10,963,000 bp
chr15
50 DSUs
11,000,000
16,675,000 bp
HML (CCATT)n-cR
H(AATGG)n-cR
M(AT)n-cR
HML(AATGG)n-cR
HML(CT#)n-cR
HML(AT)n-cR
HML(CCATT)n-cR
HML(CCATT)n-cR
C
a
HML (CT#)n-cR
ML(AT)n -cR
M(AT)n -cR
HML(AT)n -cR
b
ML(ATC)n -cR
L(ATC)n -cR
ML(GAT)n -cR
L(GAT)n -cR
c
ML (AAGG)n -cR
L (AAGG)n -cR
HML (GGCT)n -cR
HML (AGCC)n -cR
d
HM (CCATT)n -cR
M (GGAGT)n -cR
H (AATGG)n -cR
HL (AATGG)n -cR
M (ACTCC)n -cR
HML (CCATT)n -cR
ML (CCATT)n -cR
HM (AATGG)n -cR
HML (AATGG)n -cR
ML (AATGG)n -cR
D
27,450,001
62,0026,000
chrY
HMDP
MMDP
LMDP
CenSat
Annotation
F
chr18 27.7
chr11 17.4
chr2 12.3
chr15 21.9
chr17 28.5
chrX 20
chr8 18.1
chr13 34.2
chr19 37
chr22 22.4
chr20 20
chr21 35.2
chr1 41.5
chr9 27
chr14 22.7
chrY 45.1
chr4 60.9
ML(AT)n-cR
chr6 34.9
chr12 37.9
chr16 45.5
chr3 138.4
M(AT)n -cR*
HML (AATGG)n -cR
chr7 64.5
chr5 52
chr10 47.7
(Mb)
G
a
2,552,001
2,803,000 bp
DSU
DSU
DSU
chr15
Landscape of STR Density
rRNA-G
HMDP-G
CHM13-15-HMDP16m3([CT]n)
CHM13-15-HMDP15([CT]n)
CHM13-15-HMDP17([CT]n)
HMDP
RNA18S3-1
RNA5-8S3-1
RNA45S3-1
rRNA28S3-1
CAT/Liftoff Genes
b
76 DSUs/ rRNA-Gs/ HMDP-Gs
16 DSUs/ rRNA-Gs/ HMDP-Gs
60 DSUs/ rRNA-Gs/ HMDP-Gs
56 DSUs/ rRNA-Gs/ HMDP-Gs
21 DSUs/ rRNA-Gs/ HMDP-Gs
Size (Mb)
chr13
chr14
chr15
chr21
chr22
DSU
rRNA-G
HMDP-G
c
i
ii
iii
chr13
chr14
chr15
chr21
chr22
identity
i Comparison of identity of 219 DSUs
ii Comparison of identity of 219 rRNA-Gs
iii Comparison of identity of 219 HMDP-Gs