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

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. And 514 genomic regions were characterized

by microsatellite density feature in the full T2T-CHM13 genome. The landscape maps revealed that the local microsatellites density varies enormously and rhythmically along the every chromosome of T2T-CHM13.

Keywords: Human genome; Microsatellite density; STRs aggregation; landscape map.

1. Introduction

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 (Ellegren 2004; Zhao et al. 2012; Hannan 2018; Hannan 2021). 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 (Hoyt and Storer 2022). 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 (Hartl 2000; Ellegren 2004; Hannan 2018). 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 (Kim et al. 2013; Hause et al. 2016; Priestley et al. 2019; van Wietmarschen et al. 2020), regulate gene expression in healthy genomes (Gymrek et al. 2016; Quilez et al. 2016), and also be related to genetic plasticity and missing heritability (Verstrepen et al. 2005; Fondon et al. 2008; Hannan 2010; Nasrallah et al. 2012; Willems et al. 2014; Mallick et al. 2016; Gymrek et al. 2017; Willems et al. 2017). Slipped-strand mispairing was suggested as a major mechanism for tandem repeat occurrence (Levinson and Gutman 1987; Schlötterer and Tautz 1992), 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 (Zhang et al. 2020).

The recent assembly of T2T-CHM13 reference removes the gap filled regions of euchromosomes and Chromosome X (Aganezov et al. 2022; Altemose et al. 2022; Gershman et al. 2022; Hoyt and Storer 2022;

Nurk et al. 2022; Vollger et al. 2022), and over 50% gaps in Chromosome Y (Rhie et al. 2023), 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 (Qi et al. 2016; Lei et al. 2021), and 3% of human genomes was reported consist of microsatellites under that standard (Subramanian et al. 2003). And microsatellites of human sequences comprised of GA/TC/GC/AT bases are investigated using seq-requester microsatellite (Nurk et al. 2022). 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 (Chen et al. 2010; Zhao et al. 2012; de Freitas et al. 2022; Sahu et al. 2022). 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 (Li et al. 2021). 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 the 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) (Mudunuri and Nagarajaram 2007) was applied to identify STRs in all the sequences of T2T-CHM13v2.0 and GRCh38.p14 genomes. The threshold for extract

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 densety 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 large sequence segment into many differential units, improve the counting resolution by altering differential unit size (D_n); the D_1 have differential unit size (D_n) of 1kb, which was prove to be a better resolution to calculate local STRs density of human genomic sequence (Li et al. 2021). So STRs position-related D1-Relative Density (pD_1RD) was calculated for every 1 kb sequence of the 24 chromosome sequences of T2T-CHM13, and also for GRCh38.p14.

2.4. 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 p-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 p-arm. 24 landscape maps of STRs density for GRCh38.p14 were also built.

2.5. Sorting Microsatellite Density Peaks

Microsatellite Density Peaks classified into the three categories of High Microsatellite Density Peaks (HMDPs) ($pD_1RD \geqq 300$), Middle Microsatellite Density Peaks (MMDPs) ($300 > pD_1RD \geqq 150$) and Low Microsatellite Density Peaks (LMDPs) ($150 > pD_1RD \geqq 90$) 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). (Tables S5, S6).

2.6. 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.

2.7. Comparison of HMDPs betweem T2T-CHM13.v2.0 and GRC38.p14

The similar HMDPs between T2T-CHM13.v2.0 and GRC38.p14, were firstly determined allelic position by MUMmer v3.23 (Kurtz et al. 2004), 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 (Larkin et al. 2007).

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 (Li et al. 2021). Herein, STRs of the complete reference human genomeT2T-CHM13 including all chromosomes were investigated by the

129 differential calculate method, in which the values of position related relative density at 1 kilo-base resolution
130 (pD_1RD) were calculated in every bin unit of 1kb DNA sequence, and the pD_1RD value was suggested to be
131 possibly a better way to estimate the local microsatellite density variation in human genome (Li et al. 2021).
132 The bedgraph format tool of UCSC Genome Browser was used to map the exact STRs distribution features
133 with the pD_1RD data of the full T2T-CHM3 genome, a series of panoramic landscapes of microsatellites
134 density were obtained to precisely exhibit the local microsatellites density in every bin of 1 kb sequence for
135 the 22 autosome, chromosome X and Y of T2T-CHM13 in viewers of Genome browser window (Fig. 1, Fig.
136 S1, Table S1). The 24 landscapes revealed that the relative density in every 1 kb sequence (the pD_1RD value)
137 varies enormously in different site along the 24 chromosomes even from 0 to 970, and the STRs is prone to
138 accumulate to a large numbers of different extent micro-satellite density peaks; these density peaks are
139 found genome-wide, and were classified into 8,823 high microsatellites density peaks (HMDPs) ($pD_1RD \geqq$
140 300), 35,257 middle microsatellites density peaks (MMDPs) ($300 > pD_1RD \geqq 150$) and 199, 649 low
141 microsatellites density peaks (LMDPs) ($150 > pD_1RD \geqq 90$) (Table 1-b, Table S4.01-S4.06); and the bed
142 format tool of UCSC Genome Browser was used to display the exact position and motif types of these
143 HMDPs, MMDPs and LMDPs (Fig. 1, Fig. S2). Similarly, the microsatellites landscapes were
144 comprehensively surveyed in the reference human genome GRCh38 as comparison (Table 1-b, Table S14 &
145 S15, Fig. S12).

146 **3.2. High Microsatellite Density Peaks (HMDPs)**

147 The local relative density values ($pD_1RD \geqq 300$) of those HMDPs are approximately six times or more of the
148 average relative density (RD=58.75) of the full T2T-CHM13 genome, the significant statistical bias of
149 micro-satellites relative density in local genome region hints their importance to human genome (Fig. 2-A,
150 Fig. S2.01). It was found that all these HMDPs occur genome widely in all the 24 chromosomes of the
151 T2T-CHM13, especially pervade in the two arms of all chromosomes at different intervals, except absent at

152 most centromeric region and some pericentromeiric region. The largest quantity of HMDPs were found in
153 ChrY with 2,697 HMDPs, secondly in Chr13, the minimum number of HMDPs in Chr19 with 127 HMDPs
154 identified, and only 305 HMDPs were found on chr1 with longest sequence; therefore, the distribution of
155 HMDPs is not related to chromosome size directly. Furthermore, 89.26% of the HMDPs are identified in
156 intergenic, 10.74% in intron region, but no HMDPs in exon region (Fig. 2-B-a, Table S10.01).

157 **3.3. Main Motif Types (MMTs) of HMDPs**

158 Analysis of the composition of the motifs in every HMDPs demonstrated that same type of motif is prone to
159 accumulate in the same HMDP, and therefore, we classified these HMDPs by the main motif composition in
160 each HMDP; for example, when the AT motif account for the main motif composition in a HMDP, it was
161 named $(AT)_n$ Main Motif Type (MMT) HMDP (Table S4.). And the MMT HMDPs can also be further
162 classified into three sub-types: the high-percentage Main Motif Type (hMMT) HMDPs (main motif ≧
163 66.7%), the middle-percentage Main Motif Type (mMMT) HMDPs (66.7%>main motif ≧50%), and the
164 low-percentage Main Motif Type (lMMT) HMDPs (50%>main motif ≧33.4%) (Table 1-c, Table S4-S8).
165 All the 8,823 HMDPs be classified into Main Motif Type (MMT) class and Mix Motif Type class. Though
166 the microsatellites generally comprising tandem repeats with motifs of 1-6 bp makes that there are possible
167 total 964 motif types (Table S16), we actually identified only 87 MMTs from 8819 HMDPs belong to MMT
168 class, and the 87 MMTs was further classified into 8 subclass : mono-, di-, tri-, tetra-, penta-, hexa-, AG- and
169 TC-MMT subclass. The di-MMT subclass includes 6944 HMDPs is the most abundant HMDPs subclass,
170 secondly is the penta-MMT subclass. Notably, 86.1% of the total 8823 HMDPs are the hMMT HMDPs
171 (Fig.2-B-b,Table S10.01), suggesting that high percentage of same repeat motif is prone to accumulated into
172 same high microsatellites density peak (HDMP) in the complete reference human genome T2T-CHM13.
173 Moreover, many MMT HMDPs usually appears in pairs with motif reverse complementary in the complete
174 genome like the example of 386 (AATGG)n MMT HMDPs reversely complementary pairing with 311

175 (CCATT)n MMT HMDPs (Table 1-c).

176 **3.4. $[AT]_n^h$ motifs type are the most abundant HMDPs**

177 Among the all hMMT HMDPs, 4828 HMDPs are $[AT]_n^h$ motif type, representing 63.5% of the di-MMT
178 subclass HMDPs and also 54.7% of all 8823 HMDPs, therefore, it is the most hMMT HMDPs HDMP
179 (Table 1-c, Fig. 2-C, Table S10.02). These $[AT]_n^h$ HMDPs were classified into two distributional patterns by
180 average distance (AD): sparse-distributional (AD: 1,434,206bp) and dense-distributional (AD: 4,476bp)
181 patterns, there are 1886 $[AT]_n^h$ HMDPs scatter sparsely and 2942 $[AT]_n^h$ HMDPs array densely in the full
182 T2T-CHM13 genome (Fig. 2-D-a & b, Table S10.03, Fig. S3.01). As a whole, the distributional features can
183 be summed up to 3 $[AT]_n^h$ HMDPs chromosomal distribution models, one model is that only $[AT]_n^h$ HMDPs
184 sparse-distribution appear in the two arm of the 19 chromosomes, another is that $[AT]_n^h$ HMDPs
185 dense-distribution occur in the short p-arm and sparse-distribution in the long q-arm as found in the five
186 acrocentric chromosome, the other is the $[AT]_n^h$ HMDPs distribution in the Chr Y which 35 $[AT]_n^h$ HMDPs
187 with sparse-distributions are found in Chr Y euchromatin region and 2575 $[AT]_n^h$ HMDPs with
188 dense-distributions in Chr Y heterochromatin region. However, the $[AT]_n^h$ HMDPs is abundantly
189 representing 54.7% of all HMDPs, but 21.3% of all MMDPs and only 2.7% of all LMDPs (Fig. 2-D-c),
190 suggesting that AT motif tandem repeats is easy to accumulate at high density in the complete reference
191 genome.

192 **3.5. The $[AATGG]_n^h$ and $[CCATT]_n^h$ penta- hMMT HMDPS account for about half sequence size of**
193 **all hMMT HMDPs**

194 There are 773 penta- motif HMDPs identified in the full T2T-CHM13 genome, and ranks secondly in all 8
195 subclass of the MMT HMDP class; however, the reverse complementary $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT
196 HMDPs are found 331 and 252 in numbers accounting for 75.4% of all penta- MMT HMDPs (Table 1-C,
197 Table S4 & S5). Though the large majority of the two type of are small in size that is lower than 5 bins (5

198 kbp); the HMDPs with size that is larger than 5 bin, are mainly $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs
199 (Fig. 2-C-b & c, Table S10.02); therefore, total size of $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs are 3789
200 and 3362 bins respectively, and combination of two hMMT HMDPs accounting for about 43.55% the size of
201 all hMMT HMDPs, implied their importance to human genome structure. Analysis of occurred locations of
202 the $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs reveals 4 distributional models about the two hMMT
203 HMDPs in the T2T-CHM13 genome: (a) no $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs distribution occur
204 in Chr 6, 8, 11, 12, 18, 19,and X, (b) both the two hMMT HMDPs appear in short arm of the five acrocntric
205 chromosomes, (c) pericentric distribution of the two hMMT HMDPs in Chr 1, 2, 3, 4, 5, 7, 9, 10, 17 and Y,
206 (d) the $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs are found on the main arms of the chromosomes (on
207 arm distribution) that are majorly in Chr 1, 2, 16, 20 and 22) (Fig. 2-E-a, Table S10.04, Fig. S3.02).
208 Remarkably, Chr 9 is clustered a large segment of $[AATGG]_n^h$ and $[CCATT]_n^h$ hMMT HMDPs at the
209 pericentromeric region of p-arm (Fig. 2-E-b).

210 **3.6. Comparison of HMDPs between T2T-CHM13 and GRCh38**

211 The micro-satellite density landscapes of 24 chromosomes of the reference human genome GRCh38.p14
212 were also built here, As the reference genome GRCh38 contained gaps and collapsed tandem repeats (Table
213 1-a, Table S14), the total sequence assembly of GRCh38.p14 is 2,937,639,396 bp, with un-sequenced gap
214 size of 150,630,436 bp. And only 3,480 HMDPs, 26,310 MMDPs and 178,848 LMDPs were identified in
215 the microsatellite density landscapes of GRCh38.p14 genome (Table 1-b, Table S14); so HMDPs in
216 GRCh38 genome are shown to be greatly different from that of T2T-CHM13, but MMDPs and LMDPs in
217 GRCh38 genome are relatively close to those in T2T-CHM13 (Table 1-b). A detailed comparison of HMDPs
218 in micro-satellite density landscapes between full T2T-CHM13 and GRCh38 full genome, revealed that
219 1233 and 1290 are high and low-middle similar corresponding to those HMDPs of GRCh38 respectively
220 (Fig. 2-F-b,c,d, Table S15, Fig. S12), above 2,000 HMDPs correspond to middle/low microsatellite density

221 peaks (M/LMDPs) and no peaks region of GRCh38, more than half HMDPs of T2T-CHM13 are
222 corresponding to un-sequenced gaps region of GRCh38 (Fig. 2-F-d, Table S15); suggesting that HMDPs
223 contribute a lot to the individual diversity of human genome in spite of that the no-sequenced gap region
224 may influence the comparing result.

225 **3.7. Genomic regions divided by local microsatellites density features**

226 Because the landscapes graphed by the differential calculating microsatellites density method, the STRs
227 density distribution feature can be visualized in every 1kb sequence local genomic region; we discovered
228 from the landscapes that the full genome of T2T-CHM13 may be divided into 514 different microsatellites
229 density related genomic regions (Table 1-d, Fig. 3-A, Fig. S4, Table S9). These genomic regions comprise of
230 4 genomic region class: Variable microsatellite densty (V-) region class, Even microsatellite densty (E-)
231 region class, Peaks Cluster (PC-) region class and Telomere repeat (T-) region class. The V-region class was
232 further divided into High Variable Microsatellite Densty Region (HVMD-R), Middle Variable Microsatellite
233 Densty Region (MVMD-R) and Low Variable Microsatellite Densty Region (LVMD-R) subclass, the 154
234 regions of this class principally correspond to gene coding regions and also some satellite distribution region
235 (Table 1-d). The E-region class includes relatively Even and Average Microsatellite Density Region
236 (EAMD-R) subclass and relatively Even and extreme Low Microsatellite Density Region (EeLMD-R)
237 subclass, 67 regions of this class mainly match α satellites and other satellites. The PC-region class are
238 subdivided into di-, tri-, tetra-, and penta- sub-classes, regions of this class showed that special motif type of
239 Microsatellite Density Peaks (MDPs) usually gather in proximity to form peaks cluster. And the
240 hexa-nucleotide motif microsatellite cluster region class was discovered only locating in the telomeric
241 region of T2T-CHM13, the 24 $(TTAGGG)_n$-Rs were found in the telomere of q-arm of all chromosomes, as
242 well as the reverse complementary 23 $(CCCTAA)_n$-Rs were found in the telomere of p-arm of all
243 chromosomes except the $H(AATGG)_n$-R in the p-arm of Chr13 (For simplicity, the region name abbreviation

244 are only shown here and below, region name nomenclature please refer to Table S9).

245 **3.8. The broad High Variable Microsatellite Density Region**

246 The microsatellite density landscape maps display that the density of microsatellites on the main arms of
247 most chromosomes has a huge rhythmic change, showing that HMDPs, MMDPs and LMDPs appear
248 alternately, such regions are summarized as broad High Variable Microsatellite Density Regions
249 (b-HVMD-Rs) with every genomic region size more than 10,000,000 bp (Fig. 3-A,Table S9). Total 42
250 b-HVMD-Rs were classified in the HVMD-R sun-class, all chromosomes contain two b-HVMD-Rs in main
251 central part of two arms except that the five acrocentric chromosomes include only one b-HVMD-R in each
252 of their q-arm and ChrY include one in p-arm (Fig 3-B-a, Table S9). Comparing all b-HVMD-Rs with
253 CAT/Liftoff Genes, RefSeq mRNAs and CenSat Anotation tracks, demonstrated that the b-HVMD-Rs well
254 corresponding to broad gene coding regions in every chromosomes (Table 1-d, Table S9). The sum size of
255 the 42 b-HVMD-Rs is 2,850,947,000 bp, representing 91.46% of the full sequence size of T2T-CHM13
256 genome; but only 4118 of the 8823 HMDPs located in the intergenic and intron region of all b-HVMD-Rs
257 (Fig. 3-B-c, Table S10 & S11).

258 **3.9. Microsatellite density Peaks Clustering (PC-) regions**

259 In the 246 genomic regions of PC-region class (Table 1-d, Fig. 3-C, Fig. S6-S9); we clarified 70 regions in
260 the di-sub-region-class , these regions were classed into 4 region types as HML$(CT^{\#})_n$-cR, HML$(AT)_n$-cR,
261 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
262 the five acrocentric chromosomes and correspond to the five rRNA genes coding regions; the 53
263 HML$(AT)_n$-cR occur in the five short arm of the five acrocentric chromosomes and the hetro-chromatin
264 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
265 acrocentric chromosomes and centromeric regions of Chr 3 and 4. Secondly, 13 regions in the tri-sub-class
266 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

267 mainly presenting at the pericentrometic regions of chr1, 2, 7, 10, 16, and 17. Then, only 5 regions were
268 identified in the tetra-sub-class as 4 region types as ML$(AAGG)_n$-cR, L$(AAGG)_n$-cR, HML$(AGCC)_n$-cR and
269 HML$(GGCT)_n$-cR, occurring at Chr2, 3 and 22. At last, we sorted out 158 regions in the penta-sub-class as
270 10 region types; therein 8 region types include 156 regions containing the reverse complementary motif of
271 $(AATGG)_n$ or $(CCATT)_n$, were found at almost all chromosomes except chr6, 8, 11, 12, 16, 19, and X, and
272 this is consistent with the observation of aforesaid no $[AATGG]_n^h$ and $[CCATT]_n^h$ motif type HMDPs
273 distribution; the other two region types include only two regions also containing another two reverse
274 complementary motif of $(ACTCC)_n$ and $(GGAGT)_n$, are only found at the short arm of the acrocentric
275 chromosome 14 and 15. The comparing with CenSat Anotation tracks in UCSC genome browser reveals that
276 di-, tri-, tetra-, and penta- sub-region-classes mainly correspond to divers classical human satellite sequences
277 region (Table 1-d, Table S9).

278 **3.10. HML$(AT)_n$-cR and ML$(AATGG)_n$-cR alternatively connecting to form the Y heterochromatin**

279 Previously we have constructed the landscapes of the in complete Y-DNA sequence of human genome
280 GRCh38, which includes many unsequenced gaps especially a big gap of more than 30,000,000 bps in the
281 heterochromatin (Li et al. 2021). Herein, we made the landscape of Y-DNA again but for the first complete
282 sequence of human genome T2T-CHM13 (Rhie et al. 2023), illustrating the local microsatellite relative
283 density feature of the Y heterochromotin (Fig. 3-D), in which genomic region HML$(AT)_n$-cR clustering of
284 $(AT)_n$ HMDPs, MMDPs and LMDPs alternatively link to region ML$(AATGG)_n$-cR clustering $(AATGG)_n$
285 MMDPs and LMDPs; therefore, the Y heterochromotin contains 43 HML$(AT)_n$-cRs and 36
286 ML$(AATGG)_n$-cRs alternatively connected at intervals (Fig. 3-D), and only 2 ML$(CCATT)_n$-cRs and 5
287 L$(AATGG)_n$-cRs replace ML$(AATGG)_n$-cRs to separate HML$(AT)_n$-cRs. The above-mentioned densely
288 distributing 2575 $(AT)_n^h$ HMDPs scatter among the 43 HML$(AT)_n$-cRs and jaggedly separated by $(AATGG)_n$
289 MMDPs and LMDPs. Copmparing the landscape track with CenSat Anotation track in UCSC genome

290 browser showed that HML$(AT)_n$-cRs and ML$(AATGG)_n$-cRs well correspond to alternating pattern of
291 classical human satellite 1B (hsat1B) and classical human satellite 3 (hsat3) in the heterochromatin Yq12
292 region (Rhie et al. 2023) (Fig. 3-D, Fig. S10).

293 **3.11. Similar local microsatellites density features in short arms of the five acrocentric**
294 **chromosomes**

295 The landscapes showed that local microsatellites density features in the five short arm of the five acrocentric
296 chromosomes is more complicate than other autosomes. A palisade arranging HML$(CT^{\#})_n$ cluster region
297 situates in central or near central part of the short arm of these chromosomes, a set of HML$(CCATT)_n$-cR,
298 H$(AATGG)_n$-cR, M$(AT)_n$-cR, HML$(AATGG)_n$-cR array in left side, 2HML$(CCATT)_n$-cR, one
299 HML$(AT)_n$-cR and several other regions lie orderly at the right side of the central HML$(CT^{\#})_n$-cR; though
300 these regions are different in length, they arrange in similar order (Fig. 3-E, Fig. S4 & S5). Comparing with
301 CAT/Liftoff Genes, RefSeq mRNAs and CenSat Anotation tracks, revealed that the HML$(CT^{\#})_n$-cR well
302 correspond to the rRNA gene cluster regions. Moreover, there are also other PC- regions arraying in either
303 side of the HML$(CT^{\#})_n$-cR. In a ward, the local microsatellites density features arraying in the landscapes
304 are very similar in the five short arms of the acrocentric chomosomes.

305 **3.12. Even and Low microsatellites density mainly distribute in the centromere**

306 Although the centromeres are known as comprised of large arrays of tandam repeated alpha satellite
307 (Altemose et al. 2022) and most genomic regions of the T2T-CHM13 genome exhibit large variations local
308 STRs density (ie. the pD_1RD value) in the microsatellite density landscape maps, the centromeric regions
309 showed that the feature of local relative microsatellite density is relatively even and low (Fig. 3-F, Fig. S10);
310 the average pD_1RD of microsatellite in all centromeric regions is 42 and far lower than average relative
311 microsatellite density with value of 58.7 in the full genome, the lowest average pD_1RD is only 12.3 in
312 centromere of Chr2, the average pD_1RD value of centromeres in11 chromosomes (chr18, 11, 15,17, X, 8, 22,

313 20, 9, 14) are lower than 30, the average pD_1RD value of centromeres in 6 chromosomes (chr19, 1, 6, 12, 13,
314 21) are between 30 to 42 (the average pD_1RD of all centromeres), and 7 centromeres in chromosomes (chr 4,
315 3, 7, 5, 10, 16) with the average pD_1RD value higher than 42. Two $M(AT)_n$-cRs are embedded in the
316 landscapes of centromere in Chr 4 and 3, and cause the average pD_1RD of centromeres rise to 138.4 and
317 60.9 respectively.

318 **3.13. $(CT^{\#})_n$ HMDPs groups separate rRNA genes in the five acrocentric short arms**

319 As aforesaid, the five short arm of the acrocentric chromosomes contain five $HML(CT^{\#})_n$-cR regions in their
320 central part, series HMDPs group with $(CT)_n$ as main motif were found clustering palisadingly and followed
321 by several MMDPs and LMDPs, forming a tendamly duplicating segmental unit (DSU), these DSUs
322 connects tandemly composing these $HML(CT^{\#})_n$-cR regions. The size of these DSUs are between
323 43000-47000 bp considered as rDNA array (Nurk et al. 2022), and there are 76 DSUs found in the short arm
324 of Chr13, 16 DSUs in Chr14, 50 DSUs in Chr15, 56 DSUs in Chr21, 21 DSUs in Chr22, and total 219
325 DSUs. Further comparison show that every $HML(CT^{\#})_n$-cR comprise of a rRNA genes groups region and a
326 the $[CT^{\#}]_n$ HMDPs groups region, remarkably, the $[CT^{\#}]_n$ HMDPs groups region well one-by-one separate
327 every the rRNA genes group in every DSUs of the five $HML(CT)_n$-cR regions (Fig. 3-G-a, Fig. S5).
328 Alignment results showed that the identities among the 219 DSUs, 219rRNA genes groups and 219 HMDPs
329 groups are higher than 0.82, 0.99 and 0.68 respectively (Fig. 3-G-b & c, Fig. S5). suggesting the $[CT^{\#}]_n$
330 HMDPs groups region may functionally separate the rRNA genes region but be more variable than rRNA
331 genes region.

332 **4. Discussion**

333 This work built the first comprehensive genome wide microsatellite density landscape maps of the first
334 complete human reference genome T2T-CHM13 (Nurk et al. 2022; Rhie et al. 2023). These landscape maps
335 exhibited that microsatellites aggregate in many genomic positions to form a large number of microsatellite

336 density peaks, and these peaks array rhythmically along the every human chromosomes arranged like notes
337 of a beautiful piece of music; therefore, the 24 microsatellite density landscapes together look like to form a
338 symphony of human life (Fig. 1). Though transcriptional and epigenetic state of human repeat elements were
339 comprehensive analyzed (Hoyt and Storer 2022), we mainly focus on the short tandem repeats of human
340 genome, and our works revealed that the microsatellite density landscapes are compatible with human
341 chromosome structure at a high extent, possibly highlight the importance of microsatellites aggregation to
342 human genome structure, variation, evolution and so on.

343 The high microsatellites density peak (HMDP) aggregated predominantly by a single motif, is the most
344 statistically significant short tandem repeats aggregation phenomenon. We identified 8823 HMDPs in
345 T2T-CHM13, much more than the number of 3480 HMDPs in GRCh38; the HMDP number was
346 substantially increased in T2T-CHM13, mainly corresponding to unassembled regions in GRCh38,
347 especially the hetero-chromatin of Chr Y and the five short arms of the acrocentric Chromosomes. Almost
348 all the HMDPs are single motif dominated, particularly, 86.1% of the 8823 HMDPs are high percentage
349 dominated by a single motif. And there should be total 964 Main Motif Type of HMDPs ranging from mono-
350 to hexa- main motif types, actually, we only identified 87 single motif dominated Main Motif Types (MMTs);
351 these statistical significances are strongly implied the importance of aggregation of short tandem repeats on
352 gene expression, regulation, epigenetics, genetic architecture and evolution of human genome. What is more,
353 the $(AT)_n$ single motif dorminated Main Motif Types (MMTs) account for more than half of the total number
354 of HMDPs, and the reverse complementary motif $(AATGG)_n$ and $(CCATT)_n$ dorminated two Main Motif
355 Types (MMTs) account for approximately half of the total bins of HMDPs; suggesting that the three MMT
356 HMDPs are worthy to be further explored for their roles in gene function and chromosome structure of
357 human genome.

358 In addition, we classified 514 characterized by microsatellite density in the complete reference genome,

359 the Variable microsatellite density region (V-) class almost overlap the gene coding region, the Even
360 microsatellite density region (E-） class mainly cover centromere, pericentromere and P-arm of acrocentric
361 chromosomes, especially, the m-EAMD-R and m-EeLMD-R) well correspond to centromeres (Table S9.01);
362 the 246 PC- regions generally locate in pericentromere, p-arm of acrocentric chromosomes and
363 heterochromatin of Chr Y; and the reverse complementary repeat motif $(CCCTAA)_n$ and $(TTAGGG)_n$
364 telomere repeat regions almost dominate in p-arms and q-arms respectively. Meanwhile, similar genomic
365 regions structural characters were observed in the five acrocentric short arms by the view of microsatellite
366 density features, and alternative microsatellite density regions structure observed in heterochromatin Chr Y
367 (Rhie et al. 2023). all these indicate that local regional microsatellite density variation may be related to
368 human genome structure. Furthermore, most (microsatellite density) peaks cluster (PC-) regions correspond
369 to different Classical human satellites (hsat); but most the even microsatellite density regions especially in
370 centromere, where microsatellite density are limited to very low and even, correspond to different α satellite
371 higher-order repeats (αSat hor) (Table 1-d); suggesting that although microsatellites (STRs) are highly
372 related to satellites (long tandem repeats), the mechanism of their occurrence are likely completely different.
373 The 219 near identical duplicational segment units (DSUs) of approximately 45-kbp that contain
374 segments of encoding 45S rRNAs, are enbedded in the short arms of the 5 acrocentric chromosomes (Hoyt
375 and Storer 2022); we identified that the HMDPs also well separate every 45S rRNA gene of about 13300 bp.
376 All the HMDPs were classified here in the intergenic and intron regions, implying that the HMDPs possibly
377 own the basic role of separating genes and exons in the genome and may be very worthwhile to further
378 explore. The first complete reference T2T-CHM13 provided us the opportunity to study short tandem repeat
379 sequences in human genome comprehensively, our results of microsatellite density landscape maps revealed
380 that short tandem repeats are tend to aggregate characteristically throughout the full genome of T2T-CHM13,
381 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.

Competing Interests

The authors declare that they have no competing interests.

Data Availability

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

The 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/1VCetOopxhYf7L6ErvvOYfh-5IjpF5WvL/view?usp=sharing

Acknowledgments

This work was supported by funding from the "National Key Research and Development Program of China" (grant No. 19-163-12-ZT-005-004-04, Recipient: Zhongyang Tan). 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.

Author Contributions

Zhongyang Tan designed and directed this study. Zhongyang Tan, Kazusato Ohshima, Yun Xia, Douyue Li, Tingyi chen, Saichao Pan Hanrou Huang, Wenxing Zhang and Yuling Liang prepared the manuscript. Tingyi Chen, Douyue Li, Saichao Pan, Hanrou Huang, Yongzhuo Fu, Wenxing Zhang and Zhongyang Tan developed the statistical method and programs. Yun Xia, Douyue Li, Tingyi chen, Saichao Pan Hanrou Huang, Wenxing Zhang, Yuling Liang, Zhuli Peng, Hongxi, Zhang, Liang Zhang, San Peng Ruixue Shi, Xinxin He, Siqian Zhou, Weili Jiao, Xiangyan Zhao, Xiaolong Wu, lan Zhou, Jingyu Zhou, Qingjian

Ouyang, You Tian, Xiaoping Jiang, Yi Zhou, Shiying Tang, Junxiong, Shen and Zhongyang Tan performed the data analysis, Kazusato Ohshima and Zhongyang Tan edited this manuscript.

References

Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C et al. 2022. A complete reference genome improves analysis of human genetic variation. *Science* **376**: eabl3533. doi:10.1126/science.abl3533

Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ et al. 2022. Complete genomic and epigenetic maps of human centromeres. *Science* **376**: eabl4178. doi:10.1126/science.abl4178

Chen M, Tan Z, Zeng G, Peng J. 2010. Comprehensive analysis of simple sequence repeats in pre-miRNAs. *Molecular biology and evolution* **27**: 2227-2232. doi:10.1093/molbev/msq100

de Freitas KEJ, Busanello C, Viana VE, Pegoraro C, de Carvalho Victoria F, da Maia LC, Costa de Oliveira A. 2022. An empirical analysis of mtSSRs: could microsatellite distribution patterns explain the evolution of mitogenomes in plants? *Funct Integr Genomics* **22**: 35-53. doi:10.1007/s10142-021-00815-7

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nature reviews Genetics* **5**: 435-445. doi:10.1038/nrg1348

Fondon JW, 3rd, Hammock EA, Hannan AJ, King DG. 2008. Simple sequence repeats: genetic modulators of brain function and behavior. *Trends in neurosciences* **31**: 328-334. doi:10.1016/j.tins.2008.03.006

Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, Jain M, Shumate A, Razaghi R, Koren S et al. 2022. Epigenetic patterns in a complete human genome. *Science* **376**: eabj5089. doi:10.1126/science.abj5089

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ et al. 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature genetics* **48**: 22-29. doi:10.1038/ng.3461

Gymrek M, Willems T, Reich D, Erlich Y. 2017. Interpreting short tandem repeat variations in humans using mutational constraint. *Nature genetics* **49**: 1495-1501. doi:10.1038/ng.3952

Hannan AJ. 2010. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends in genetics : TIG* **26**: 59-65. doi:10.1016/j.tig.2009.11.008

Hannan AJ. 2018. Tandem repeats mediating genetic plasticity in health and disease. *Nature reviews Genetics* **19**: 286-298. doi:10.1038/nrg.2017.115

Hannan AJ. 2021. Repeat DNA expands our understanding of autism spectrum disorder. *Nature* **589**: 200-202. doi:10.1038/d41586-020-03658-7

Hartl DL. 2000. Molecular melodies in high and low C. *Nature reviews Genetics* **1**: 145-149. doi:10.1038/35038580

Hause RJ, Pritchard CC, Shendure J, Salipante SJ. 2016. Classification and characterization of microsatellite instability across 18 cancer types. *Nature medicine* **22**: 1342-1350. doi:10.1038/nm.4191

Hoyt SJ, Storer JM. 2022. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**: eabk3112. doi:10.1126/science.abk3112

Kim TM, Laird PW, Park PJ. 2013. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* **155**: 858-868. doi:10.1016/j.cell.2013.10.015

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for

447 comparing large genomes. *Genome Biol* **5**: R12. doi:10.1186/gb-2004-5-2-r12
448 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al.
449 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948. doi:10.1093/bioinformatics/btm404
450 Lei Y, Zhou Y, Price M, Song Z. 2021. Genome-wide characterization of microsatellite DNA in fishes: survey and analysis of their
451 abundance and frequency in genome-specific regions. *BMC genomics* **22**: 421. doi:10.1186/s12864-021-07752-6
452 Levinson G, Gutman GA. 1987. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage
453 M13 in Escherichia coli K-12. *Nucleic Acids Res* **15**: 5323-5338. doi:10.1093/nar/15.13.5323
454 Li D, Pan S, Zhang H, Fu Y, Peng Z, Zhang L, Peng S, Xu F, Huang H, Shi R et al. 2021. A comprehensive microsatellite
455 landscape of human Y-DNA at kilobase resolution. *BMC genomics* **22**: 76. doi:10.1186/s12864-021-07389-5
456 Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A et al. 2016. The
457 Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**: 201-206.
458 doi:10.1038/nature18964
459 Mudunuri SB, Nagarajaram HA. 2007. IMEx: Imperfect Microsatellite Extractor. *Bioinformatics* **23**: 1181-1187.
460 doi:10.1093/bioinformatics/btm097
461 Nasrallah MP, Cho G, Simonet JC, Putt ME, Kitamura K, Golden JA. 2012. Differential effects of a polyalanine tract expansion in
462 Arx on neural development and gene expression. *Human molecular genetics* **21**: 1090-1098. doi:10.1093/hmg/ddr538
463 Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A et al.
464 2022. The complete sequence of a human genome. *Science* **376**: 44-53. doi:10.1126/science.abj6987
465 Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, Duyvesteyn K, Haidari S, van Hoeck A, Onstenk W et al.
466 2019. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**: 210-216.
467 doi:10.1038/s41586-019-1689-y
468 Qi WH, Yan CC, Li WJ, Jiang XM, Li GZ, Zhang XY, Hu TZ, Li J, Yue BS. 2016. Distinct patterns of simple sequence repeats
469 and GC distribution in intragenic and intergenic regions of primate genomes. *Aging* **8**: 2635-2654.
470 doi:10.18632/aging.101025
471 Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y, Joshi RS, Mittelman D, Sharp AJ. 2016. Polymorphic tandem
472 repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res*
473 **44**: 3750-3762. doi:10.1093/nar/gkw219
474 Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, Hook PW, Koren S, Rautiainen M, Alexandrov IA et al. 2023. The
475 complete sequence of a human Y chromosome. *Nature* **621**: 344-354 doi:10.1038/s41586-023-06457-y
476 Sahu BP, Majee P, Singh RR, Sahoo N, Nayak D. 2022. Genome-wide identification and characterization of microsatellite
477 markers within the Avipoxviruses. *3 Biotech* **12**: 113. doi:10.1007/s13205-022-03169-4
478 Schlötterer C, Tautz D. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* **20**: 211-215.
479 doi:10.1093/nar/20.2.211
480 Subramanian S, Mishra RK, Singh L. 2003. Genome-wide analysis of microsatellite repeats in humans: their abundance and
481 density in specific genomic regions. *Genome Biol* **4**: R13. doi:10.1186/gb-2003-4-2-r13
482 van Wietmarschen N, Sridharan S, Nathan WJ, Tubbs A, Chan EM, Callen E, Wu W, Belinky F, Tripathi V, Wong N et al. 2020.
483 Repeat expansions confer WRN dependence in microsatellite-unstable cancers. *Nature* **586**: 292-298.
484 doi:10.1038/s41586-020-2769-8
485 Verstrepen KJ, Jansen A, Lewitter F, Fink GR. 2005. Intragenic tandem repeats generate functional variability. *Nature genetics* **37**:
486 986-990. doi:10.1038/ng1618
487 Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, Diekhans M, Sulovari A, Munson KM, Lewis AP et al.
488 2022. Segmental duplications and their variation in a complete human genome. *Science* **376**: eabj6965.
489 doi:10.1126/science.abj6965
490 Willems T, Gymrek M, Highnam G, Genomes Project C, Mittelman D, Erlich Y. 2014. The landscape of human STR variation.
491 *Genome research* **24**: 1894-1904. doi:10.1101/gr.177774.114
492 Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. 2017. Genome-wide profiling of heritable and de novo STR
493 variations. *Nature methods* **14**: 590-592. doi:10.1038/nmeth.4267

Zhang H, Li D, Zhao X, Pan S, Wu X, Peng S, Huang H, Shi R, Tan Z. 2020. Relatively semi-conservative replication and a folded slippage model for short tandem repeats. *BMC genomics* **21**: 563. doi:10.1186/s12864-020-06949-5

Zhao X, Tian Y, Yang R, Feng H, Ouyang Q, Tian Y, Tan Z, Li M, Niu Y, Jiang J et al. 2012. Coevolution between simple sequence repeats (SSRs) and virus genome size. *BMC genomics* **13**: 435. doi:10.1186/1471-2164-13-435

Figure legends

Figure 1. 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 (High Microsatellite Density Peak), third track displayed MMDP (Middle Microsatellite Density Peak), last track displayed LMDP (Low Microsatellite Density Peak).

Figure 2. Summary of High Microsatellite Density Peaks (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 number of $[AT]_n^h$ Sparse distribution (S-distribution) and Dense distribution (D-distribution) HMDPs; (c) 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 $[CCATT]_n^{\mathbf{h}}$ 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 3. Summary of genomic regions charcterized 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, inluding 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-cRin 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 HDMPs 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 segmental 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.

Table 1-a. The assembly information and microsatellite statistics of T2T-CHM13v2.0 and GRCh38.p14.

Genome	Assembled size (bp)	Sequenced size (bp)	Gap number	Gap size (bp)	Microsatellite number	Microsatellite size (bp)	Microsatellite Relative Density (bp/Kbp)
CHM13	**3,117,275,501**	**3,117,275,501**	**0**	**0**	**22,198,470**	**183,133,984**	**58.75**
GRCh38	3,088,269,832	2,937,639,396	783	150,630,436	19,672,293	151,222,113	51.48

Table 1-b. The list of microsatellites density peaks in T2T-CHM13v2.0 and GRCh38.p14.

Chr No.	HMDPs		MMDPs		LMDPs		Chr No.	HMDPs		MMDPs		LMDPs	
	CHM13	GRCh38	CHM13	GRCh38	CHM13	GRCh38		CHM13	GRCh38	CHM13	GRCh38	CHM13	GRCh38
01	**305**	199	**2,703**	1,968	**16,891**	14,086	13	**587**	100	**1,682**	825	**7,098**	5,976
02	**357**	294	**2,015**	1,972	**14,802**	14,395	14	**342**	83	**1,205**	728	**6,005**	5,405
03	**231**	180	**1,475**	1,404	**11,798**	11,624	15	**507**	67	**2,638**	637	**7,772**	4,968
04	**234**	230	**1,955**	1,579	**11,889**	11,421	16	**130**	105	**1,000**	989	**7,688**	5,509
05	**258**	198	**1,455**	1,372	**10,580**	10,572	17	**135**	89	**880**	786	**5,771**	5,681
06	**200**	158	**1,379**	1,365	**10,172**	10,128	18	**125**	91	**643**	645	**4,429**	4,420
07	**249**	197	**1,581**	1,506	**10,002**	9,978	19	**127**	105	**1,046**	956	**5,102**	4,869
08	**188**	156	**1,291**	1,313	**8,481**	8,352	20	**184**	75	**806**	650	**3,691**	3,753
09	**237**	121	**1,058**	997	**7,239**	7,199	21	**449**	87	**1,088**	505	**3,009**	2,392
10	**208**	167	**1,328**	1,203	**8,226**	8,143	22	**378**	99	**1,039**	462	**3,074**	2,403
11	**191**	143	**1,125**	1,060	**7,581**	7,706	X	**313**	276	**1,673**	1,648	**9,465**	9,470
12	**191**	155	**1,215**	1,203	**8,470**	8,433	Y	**2,697**	105	**3,977**	537	**10,414**	1,965
							Total	**8,823**	3,480	**36,257**	26,310	**199,649**	178,848

Table 1-c. The motif types of HMDP in T2T-CHM13v2.0.

Main Motif Type (MMT) class

di-MMT subclass

MMT [a]	sub-MMT [b]	
$(AT)_n$ 5063	$[AT]_n^h$	**4,828**
	$[AT]_n^m$	188
	$[AT]_n^l$	47
$(CT)_n$ 905	$[CT]_n^h$	748
	$[CT]_n^m$	95
	$[CT]_n^l$	62
$(AG)_n$ 101	$[AG]_n^h$	77
	$[AG]_n^m$	15
	$[AG]_n^l$	9
$(GT)_n$ 443	$[GT]_n^h$	414
	$[GT]_n^m$	15
	$[GT]_n^l$	14
$(AC)_n$ 431	$[AC]_n^h$	394
	$[AC]_n^m$	25
	$[AC]_n^l$	12
$(CG)_n$ 1	$[CG]_n^h$	1
	$[CG]_n^m$	0
	$[CG]_n^l$	0
6[d]		6,944[e]

tri-MMT subclass

MMT	num [c]
$(GGT)_n$	64
$(ACC)_n$	56
$(AGG)_n$	39
$(CCT)_n$	37
$(CTT)_n$	25
$(AAG)_n$	23
$(ATC)_n$	12
$(GAT)_n$	12
$(CCG)_n$	3
$(CGG)_n$	2
$(ACT)_n$	3
$(AGT)_n$	1
$(AGC)_n$	2
$(ATT)_n$	1
$(GTT)_n$	1
15	281

tetra-MMT subclass

MMT	num	MMT	num
$(CTTT)_n$	215	$(GGCT)_n$	6
$(AAAG)_n$	130	$(AGCC)_n$	3
$(ATCC)_n$	69	$(ACTC)_n$	3
$(GGAT)_n$	59	$(GAGT)_n$	3
$(AAGG)_n$	54	$(ACCC)_n$	2
$(CCTT)_n$	35	$(GGGT)_n$	1
$(AGGG)_n$	28	$(ACCT)_n$	1
$(CCCT)_n$	27	$(AGGT)_n$	1
$(AGAT)_n$	16	$(AATG)_n$	3
$(ATCT)_n$	10	$(ATGT)_n$	3
		$(CTGT)_n$	1
		21	670

penta-MMT subclass

MMT	sub-MMT	
$(AATGG)_n$ 386	$[AATGG]_n^h$	**331**
	$[AATGG]_n^m$	50
	$[AATGG]_n^l$	5
$(CCATT)_n$ 311	$[CCATT]_n^h$	**252**
	$[CCATT]_n^m$	56
	$[CCATT]_n^l$	3

MMT	num
$(ACTCC)_n$	10
$(GGAGT)_n$	8
$(AAAGG)_n$	8
$(CCTTT)_n$	5
$(AGGGG)_n$	8
$(CCCCT)_n$	4
$(AAGGG)_n$	5
$(CCCTT)_n$	3

MMT	num
$(AAAAT)_n$	4
$(ATTTT)_n$	2
$(ATATT)_n$	2
$(AATAT)_n$	1
$(CCTCT)_n$	2
$(AGAGG)_n$	1
$(AAAAG)_n$	1
$(CTTTT)_n$	1
$(AGCCC)_n$	2
$(ATCCT)_n$	2
$(GGGAT)_n$	2
$(ACACC)_n$	1
$(ATAGT)_n$	1
$(GCCCT)_n$	1
$(CTGGT)_n$	1
$(GGTTT)_n$	1
26	773

hexa-MMT subclass

MMT	num
$(CCCTAA)_n$	26
$(TTAGGG)_n$	25
$(CCCTCT)_n$	8
$(AGAGGG)_n$	4
$(ACCATC)_n$	1
$(GATGGT)_n$	1
$(AGATAT)_n$	1
$(GAGGGT)_n$	1
$(ATAGGT)_n$	1
$(GGAGGT)_n$	1
$(CCATCT)_n$	1
$(ATCTCT)_n$	1
$(GGATGT)_n$	1
$(CCCCCT)_n$	1
14	73

mono-MMT subclass

MMT	num
$(G)_n$	8
$(C)_n$	6
$(T)_n$	1
3	15

AG-MMT subclass

MMT	num
$(AG_{nt})_n$	33
1	33

TC-MMT subclass

MMT	num
$(TC_{nt})_n$	30
1	30

MMTs/HMDPs in MMT class: 87/8819 (mono-: 3/15, di-: 6/6944, tri-: 15/281, tetra-: 21/670, penta-: 26/773, hexa-: 14/73, AG-: 1/33, TC-: 1/30)

Mix Motif Type (Mix) class

Motif type	num
Mix	4

HMDPs in Mix class: 4

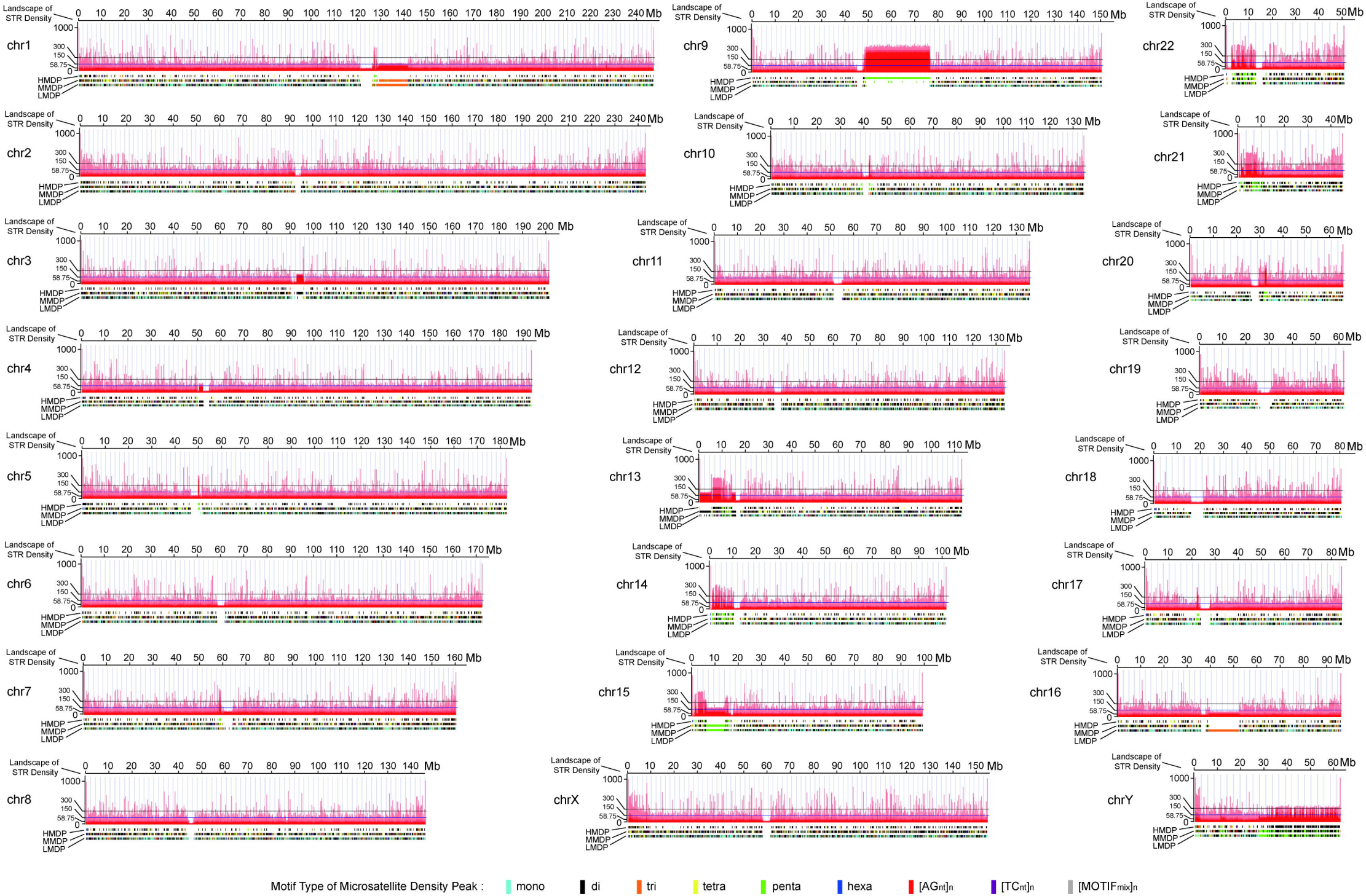
Total HMDPs: 8823 (HMDPs in MMT class: 8819 (h-: 7600, m-: 749, l-: 407, AG-: 33, TC-: 30); HMDPs in Mix class: 4)

a-c: Main Motif Type (MMT), sub-Main Motif Type (sub-MMT), and HMDP number (num); d: MMT number in each subclass, e: HMDP number in each subclass. The brace represents the sub-MMT corresponding to MMT, Reverse complementary MMT in dotted box.

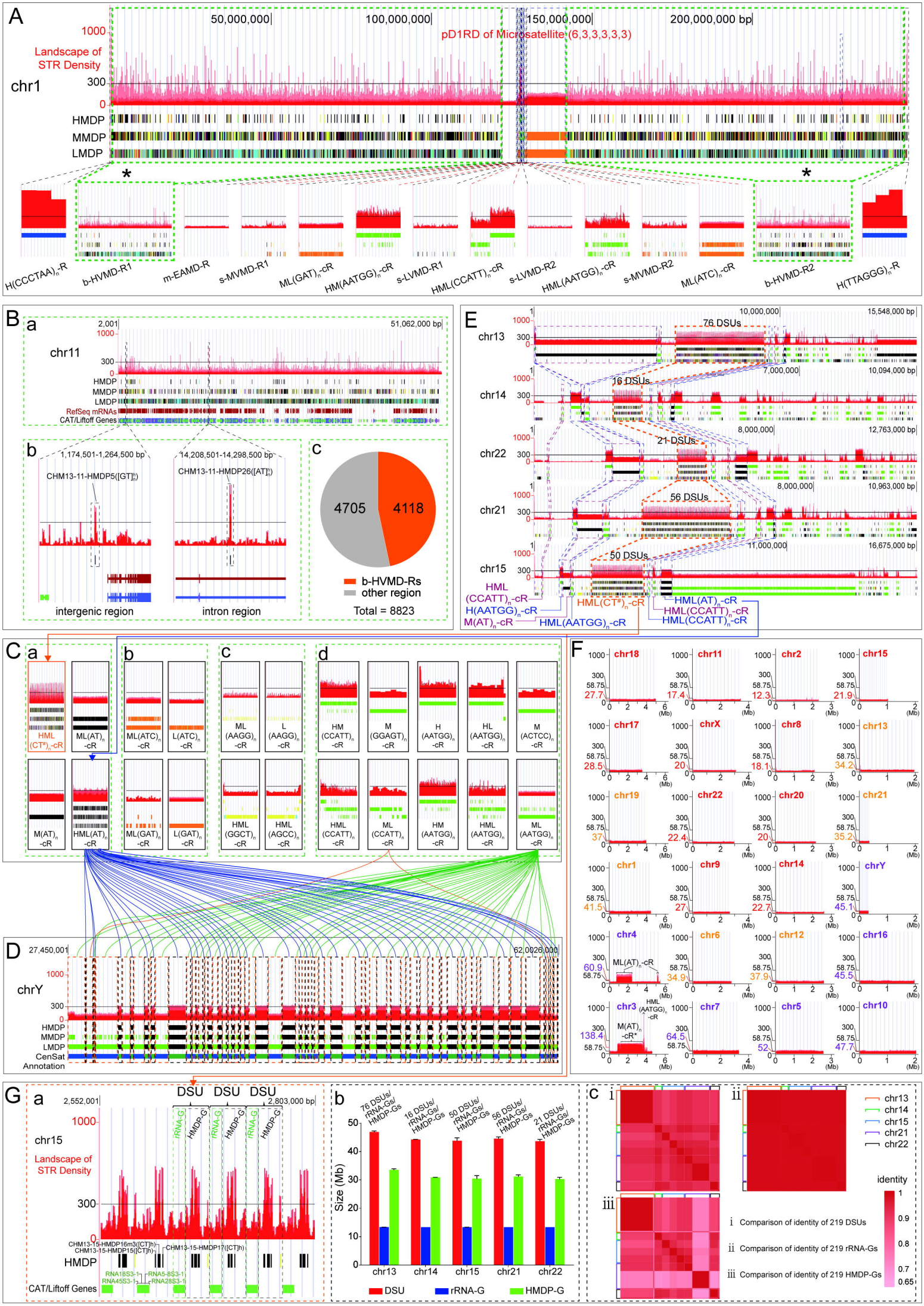
Table 1-d. Genomic regions characterized by microsatellite density in T2T-CHM13v2.0.

Region class	Region subclass	Region type	Region num	Corresponding annotation	Located chr
V-[*a] (154)	HVMD-R (78)	b-HVMD-R	42	Broad GCR; (ct, censat, etc)[*e]	all chr
		m-HVMD-R	7	Medium GCR; (censat, ct, etc)	13,14,16,20,21,Y
		s-HVMD-R	29	Short GCR; (ct, censat, etc)	2,9,12-15,21,22,Y
	MVMD-R	s-MVMD-R	37	Sparse GCR; (ct, censat, etc)	1-4,7,10,13-15,17,21,22,Y
	LVMD-R (39)	m-LVMD-R	1	bsat	22
		s-LVMD-R	38	Sparse GCR; (ct, censat, etc)	1,3,5-7,10,12-15,17,19,21,22
E-[*b] (67)	EAMD-R (31)	m-EAMD-R	9	αSat (hor, dhor, mon), bsat, ct	1,5-7,10,12,13,16,19
		s-EAMD-R	22	αSat (mon, hor, dhor), ct	3,6,7,14,15,19,21,22,Y
	EeLMD-R (36)	m-EeLMD-R	12	αSat (hor, dhor, mon), ct	2,4,8,9,11,14,15,17,18,20,22,X
		s-EeLMD-R	24	bsat, αSat (hor, mon), censat, ct	3-5,9,10,13-15,21,22,Y
PC-[*c] (246)	di- (70)	$HML(CT^{\#})_n$-cR	5	rRNA GCR; (ct, censat)	13,14,15,21,22
		$HML(AT)_n$-cR	53	hsat1B, hsat1A, hsat3, ct, bsat	13,14,15,21,22,Y
		$M(AT)_n$-cR	3	hsat1A	3,14,15
		$ML(AT)_n$-cR	9	hsat1A, bsat	3,4,13,14,15,21,22
	tri- (13)	$ML(ATC)_n$-cR	2	hsat2, ct	1,2
		$L(ATC)_n$-cR	5	hsat2	2,7,10
		$ML(GAT)_n$-cR	4	hsat2	1,2,17
		$L(GAT)_n$-cR	2	hsat2	7,16
	tetra- (5)	$ML(AAGG)_n$-cR	1	censat, ct	22
		$L(AAGG)_n$-cR	1	censat	3
		$HML(AGCC)_n$-cR	2	censat	2
		$HML(GGCT)_n$-cR	1	censat	2
	penta- (158)	$ML(AATGG)_n$-cR	44	hsat3	14,15,17,Y
		$HML(AATGG)_n$-cR	39	hsat3, αSat (mon), hsat2, ct	1-3,7,10,13-15,17,20-22,Y
		$HM(AATGG)_n$-cR	10	hsat3, ct	1,9,10,14,15,21,22
		$H(AATGG)_n$-cR	4	hsat3, censat; (p-TRR[*f])	13,14,21,22
		$HL(AATGG)_n$-cR	1	hsat3	7
		$HML(CCATT)_n$-cR	44	hsat3, αSat (mon, dhor), etc	1,2,4,7,9,10,13-15,21,22,Y
		$HM(CCATT)_n$-cR	10	hsat3	2,5,7,13,14,15,17,22,Y
		$ML(CCATT)_n$-cR	4	hsat3	15,22,Y
		$M(ACTCC)_n$-cR	1	hsat3	15
		$M(GGAGT)_n$-cR	1	hsat3	14
T-[*d] (47)	hexa- (47)	$H(CCCTAA)_n$-R	16	p-TRR; (censat)	1-3,6,7,9-12,16,18-21,X,Y
		$HM(CCCTAA)_n$-R	7	p-TRR; (censat)	4,5,8,14,15,17,22
		$H(TTAGGG)_n$-R	18	q-TRR[*g]; (censat)	1-7,11-14,17-19,21,22,X,Y
		$HM(TTAGGG)_n$-R	6	q-TRR	8,9,10,15,16,20
		Total regions	514		

*a: Variable microsatellite density (V-) region class; *b: Even microsatellite density (E-) region class; *c: Peaks Cluster (PC-) region class; *d: Telomere repeat (T-) region class. *e: Gene Coding Region (GCR), Outside the parentheses were dominant annotations, inside the parentheses were secondary annotations; *f: Telomere Repeat Region in chromosome p-arm (p-TRR); *g: Telomere Repeat Region in chromosome q-arm (q-TRR). Reverse complementary motif region type in dotted box.


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
1000
CHM13-01-HMDP78([AT]ⁿₙ)
chr1
300
0
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
100%
80%
60%
40%
20%
0
4%
3%
2%
1%
0
[di]ⁿₙ
[penta]ⁿₙ
[mono]ⁿₙ
[tri]ⁿₙ
[tetra]ⁿₙ
[hexa]ⁿₙ
b
peak number
6000
4000
2000
0
1000
800
600
400
200
0
1
2
3
4
5-10
11-939
Peak size (bins)
c
bin number
6000
4000
2000
0
1
2
3
4
5-10
11-939
Peak size (bins)
[AT]ⁿₙ
[AATGG]ⁿₙ
[CCATT]ⁿₙ
[mono]ⁿₙ
other [di]ⁿₙ
[tri]ⁿₙ
[tetra]ⁿₙ
other [penta]ⁿₙ
[hexa]ⁿₙ
D
a
[AT]ⁿₙ (N: 1,650; AD: 1,430,597 bp)
b-HVMD-R1
b-HVMD-R2
i
(chr 1,2,3,4,5,6,7,8, 9,10,11,12,16,17, 18,19,20 and X)
[AT]ⁿₙ (N: 367; AD: 1,883 bp)
[AT]ⁿₙ (N: 205; AD: 1,563,427 bp)
HML(AT)n-cR 1 - n
b-HVMD-R
ii
(chr13,14,15, 21 and 22)
[AT]ⁿₙ (N: 31; AD: 741,644 bp)
[AT]ⁿₙ (N: 2,575; AD: 4,476 bp)
b-HVMD-R
m-HVMD-R1
m-HVMD-R2
HML(AT)n-cR 1 - 43
iii
(chrY)
b
[AT]ⁿₙ S-distribution
Average Distance (AD): 1,434,206 bp
1886
2942
Total N = 4828
[AT]ⁿₙ D-distribution
Average Distance (AD): 4,156 bp
c
60%
40%
20%
0
54.7%
21.3%
2.7%
[AT]ⁿₙ
HMDPs
MMDPs
LMDPs
E
a
Total N: 0 ; Total S: 0 bin
Total N: 0 ; Total S: 0 bin
No [AATGG]ⁿₙ / [CCATT]ⁿₙ distribution
i
(chr6,8,11,12,18,19 and X)
Total N: 169 ; Total S: 1152 bin
Total N: 157 ; Total S: 1704 bin
acrocentric [AATGG]ⁿₙ / [CCATT]ⁿₙ distribution
ii
(chr13,14,15,21 and 22)
Total N: 82; Total S: 2207 bin
Total N: 90; Total S: 1268 bin
pericentric [AATGG]ⁿₙ / [CCATT]ⁿₙ distribution
iii
(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]ⁿₙ / [CCATT]ⁿₙ distribution
iv
(chr1, 2, 16, 20 and 22)
[AATGG]ⁿₙ
Total Number (N): 331;
Total Size (S): 3789 bin
[CCATT]ⁿₙ
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
1000
chr9
Landscape of STR Density
300
0
HMDP
MMDP
LMDP
F
a
5Kb
2Kb
CHM13-01-HMDP21m3([GT]h)
50,000,000
100 Mb
100,000,000
150,000,000
200,000,000 bp
CHM13 chr1
1000
Landscape of STR Density
CHM13-01-HMDP59([CT]h)
300
0
HMDP
MMDP
LMDP
GRCh38-01-HMDP15([GT]h)
50,000,000
100 Mb
100,000,000
150,000,000
200,000,000 bp
GRCh38 chr1
1000
Landscape of STR Density
GRCh38-01-HMDP39([CT]h)
300
0
HMDP
MMDP
LMDP
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
1234
L/M-similarity
HMDPs
=2523
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