Manuscript

Maps of high density microsatellites accumulation in human genome

Yun Xia[1,#], Hanrou Huang[1,#],Yulin Liang[1,#], Douyue Li[1,#], Tingyi Chen[1], Ruixue Shi[1], Saichao Pan[1], Fei Xu[2], Shaoliang Peng[1,*] and Zhongyang Tan[1,*]

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

[2] Department of Mathematics, Wilfrid Laurier University, Waterloo, Ontario N2L 3C5, Canada

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

[*] Corresponding author

E-mail: zhongyangtan@yeah.net or slpeng@hnu.edu.cn

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

Postal Code: 410082

Abstract

We systematically built more than 100,000 microsatellites landscape-maps with identifying 3,433 and 5,306 high density microsatellites accumulation (HDMA) peaks in the human reference genome (GRCh38, p13 version) and the first complete human genome (T2T-CHM13, 1.1 version) at 1 kilobase resolution. Intuitive HDMA peak-locus maps were constructed for every chromosome of the 2 genomes, revealing that these HDMA peaks are distributed regularly along every chromosome. Data mining and comparison of HDMA peaks illustrated that most of the HDMA peaks may play biological roles with highly variable genomic sequences. Therefore, comparing the variation of HDMA peaks will provide an unprecedented approach to study human genetic variation.

Keywords: Human genome; Microsatellite; Microsatellites accumulation; Peak-locus map; Human genetic variation

1. Introduction

Our understanding of the function of human genomes, complex-trait genetics and the genetics of human diseases, usually has been limited by technologies of analyzing single nucleotide polymorphisms (SNPs), particularly in genome-wide association studies (GWAS) [1-3]. Indeed, over half of the human genome consist of repetitive elements that make up the repeatome, microsatellites DNA, also known as short tandem repeats (STRs) or simple sequence repeats (SSRs), commonly occur with repeat-units of 1-6 bp/nt in eukaryotes, prokaryotes, and also viruses [4-7]. Short tandem repeats (STRs) have been estimated to include over 1 million discrete STR loci in the human genome [8,9]; they are increasingly realized to have biological roles [10-12], and are most intensively and comprehensively studied in linkage with diseases, where many medical studies have found that the STRs with abnormal repeat-unit numbers, also called microsatellite instability (MSI), are linked to more than 40 genetic diseases like fragile X syndrome, Huntington's disease and autism, or in many cancer genomes like colorectal and breast cancer [13-16]. Many non-coding tandem repeats are located in gene regulatory regions and are thought to fine-tune gene transcription through several mechanisms [5,12,17,18]. Though there are increasing numbers of microsatellites studies in human genomes in the past decades, these studies still have only considered relatively longer and small part of disease-associated STRs in a few genomic locations, or just processed primarily global statistics, or provided chances for surveying parts of the elementary positions of the relatively longer STRs in the full human genome [19-22].

Human genome contains 22 autosomes and 2 sex chromosomes, and the reported human reference genome is the most researched human genome, whose extensively researched version is GRCh38 released from the Genome Reference Consortium in 2013 (patch GRCh38.p13 in 2019), with 3,099,734,149 bp in total sequence length and 2,948,611,470 bp in total ungapped length, including overall 473 scaffolds and 999 contigs [23,24]. Recently, the Telomere-to-Telomere Consortium has firstly finished the sequencing of a complete human (female) genome without any gap called T2T-CHM13 (CHM13), with 3,054,815,472 bp in total sequence length, which contains all centromeric satellite arrays and the short arms of all 5 acrocentric chromosomes (chr13, 14, 15, 21 and 22), and this complete genome can provide a more comprehensive perspective to analyze microsatellites in human genome [25-28]. Previously, we investigated STRs landscape maps in the full human reference chromosome Y (GRCh38: NC_000024.10) at 1 kilo base pairs (Kbp) resolution by Differential Calculator of Microsatellite (DCM) method [29], revealing an exact distributional feature of STRs in every 1-Kbp locational bins of the chromosome Y [30]. In this study, we investigated STRs landscape maps in all chromosomes of human genome of GRC38 (p13 version) and CHM13 (1.1 version), and found that STRs tend to form HDMA peaks along each chromosome of GRCh38 and CHM13 human genomes.

2. Materials and methods

2.1. Genomic sequence selection

We collected the reported genomic sequences of the human reference genome (GRCh38, version p13) [23,24] and the human reference genome (T2T-CHM13, version 1.1) [25] from Genbank, which are the most authoritative reported sequences of the human genome. The related information of all investigated chromosome sequences was listed in Table S1.

2.2. STRs identification

IMEx 2.1 (Imperfect Microsatellite Extractor) was applied to identify STRs in the investigated genomes, which can find out relatively shorter STRs [31]. The threshold, which represents the lowest copy number of STR repeat units, was set at 6, 3, 3, 3, 3, 3 for extracting perfect mono-, di-, tri-, tetra-, penta- and hexanucleotide STRs respectively according to many previous STRs studies [30,32,33]. RepeatMasker and Tandem Repeat Finder were used as the reference tools for STRs identification [34,35].

2.3. STRs landscape maps establishment

The STRs landscape maps in the sample chromosomes were established by our developed method - Differential Calculator of Microsatellites (DCM) [29,30]; it could display the STRs density value in each genomic locational bin along the sample chromosome; this STRs density value was defined as STRs position-related differential relative density (pD_nRD), and pD_1RD represents the STRs density value at 1 Kbp differential resolution (pD_1RD = (STRs size in each 1-Kbp bin /1-Kbp bin size) × 1000). The STRs landscape maps in the 2 genomes were based on the corresponding statistics of the STRs pD_1RD and visualized by ggplot2 R package [36]. The approach used for building the chromosomal STR maps using DCM can be found at: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07389-5. A STRs landscape map represents a zone that typically including 50 1-Kbp bins, while the zone next to the gap contains <50 1-Kbp bins, in addition, the ungapped regions <50 Kbp were excluded. Each map was labeled by a zone serial number at the top.

2.4. HDMA peak locus maps establishment

The locus maps of HDMA peaks of the GRCh38 and the T2T-CHM13 were established strictly according to the actual position of the chromosome. It can illuminate the distributional features of HDMA peaks along sample chromosomes. The lengths of all chromosomes in the figure have uniform length units. The Peak of high-density STRs accumulation with a density value at 300~425 being defined as HP, and peak with a density value over 425 being defined as sHP. Different motif-type peaks were lighted in different colors.

2.5. Investigation on biological significance of HDMA peak region and variation of the HDMA peaks

The related biological significance of the HDMA peak region in GRCh38 human genome were searched in UCSC Genome Browser and its source database [37-40]. The packages nucmer, delta-filter, show-coords and mummerplot in MUMmer3.23 were used to compare each chromosome of GRCh38 and T2T-CHM13 genome [41]. Conservation and variation of HDMA peaks between GRCh38 and T2T-CHM13 genome were compared by ClustalX2.1 [42].

3. Result

3.1. Microsatellites tend to accumulate into high density peaks

The microsatellites were investigated in the sequenced regions of the human reference (GRCh38) genome (Table S1) under the widely accepted STRs threshold of 6, 3, 3, 3, 3, 3 for mono- to hexanucleotide repeat motifs. More than 5% of the GRCh38 human genome consists of STRs instead of about 3% [43] (Table S2). Furthermore, the microsatellites landscapes at 1 Kbp resolution, as we displayed in the reference human Y-DNA previously [30], were comprehensively surveyed in the reference sequences of the GRCh38 human genome. Then, a total 58,769 SSRs landscape maps were obtained in the GRCh38 human genome (Fig. S1.01-S1.24). These STRs landscape maps clearly and exactly revealed the STRs distributional characteristics in the GRCh38 human genome.

The STRs were observed to accumulate abundantly in the full sequences of the GRCh38 human genome, forming different density levels of STRs accumulation peaks in those landscape maps. There are numerous low-density STRs accumulation peaks, large numbers of middle-density STRs accumulation peaks, and thousands of high-density STRs accumulation peaks (HP) also including super high-density STRs accumulation peaks (sHP), which were clarified in the GRCh38 human genomes (Fig. S1.01-S1.24). Especially, those high-density microsatellites accumulation (HDMA) peaks (HP & sHP) showed the relative density values about six times or more of the average microsatellites relative density of the full GRCh38 genome (Fig. 1, Table S2); these HDMA peaks usually cover a very short genomic region with only one or dozens of 1-Kbp bins, a few HDMA peaks even cover hundreds of 1-Kbp bins or more. For example, there are 199 HDMA peaks in the sequenced GRCh38 chromosome 1, and 292 peaks in the sequenced GRCh38 chromosome 2, ..., and 100 peaks in the sequenced GRCh38 chromosome Y; totally 3433 HDMA peaks (2335 HP and 1098 sHP) were observed on the chromosomes of GRCh38 landscape maps (Fig. 2A, Table S3, Table S4). Similarly, the microsatellites landscapes were comprehensively surveyed in the recently published sequences of a complete human genome (T2T-CHM13). As the new T2T-CHM13 human genome finished with gapless assembles for all 22 autosomes and chromosome X, which includes many repetitive and polymorphic regions, the number of HDMA peaks was identified to be obviously increased in every chromosome of T2T-CHM13 with different degree; especially in the five acrocentric chromosomes (Chr13, 14, 15, 21 and 22), the HDMA peak numbers increased dramatically from 100, 83, 66, 83 and 87 to 473, 226, 350, 341 and 300 respectively; then total 5306 HDMA peaks (3,235 HP and 2,071 sHP) were identified in the chromosomes of T2T-CHM13 (Fig. 2A, Table S3, Table S5). In particular, among all these peaks of the CHM13 genome, the largest one is C13-09-sHP26$_{m27529}$ on chromosome 9 covering a genomic region of 27.5 Mbp.

3.2. Typical motif types for HDMA peaks

Analysis of the HDMA motif type showed that the HDMA peaks tend to be consisted of the same or several

STRs motifs with significant statistical bias. For instance, the peak G38-02-sHP49 at chromosome 2 of GRCh38 is accumulated by the motifs including AT_n, A_n, AC_n, AG_n, TG_n, TC_n with percentage of 87.7%, 3.0%, 3.0%, 1.7%, 1.3%, 1.3% respectively, can be classified into the $[AT]^h$ motif type of STRs high-density peaks (Fig. 2B). Then, the data showed that the $[AT]^h$ and $[AT]^m$ dinucleotide motif type peaks account for a very high percentage in almost all chromosomes of both GRC38 and CHM13 (Fig. 2C, Table S6), and the other dinucleotide motif type peaks also occur frequently in most of the investigated chromosomes, except only one $[GC]^{h+m}$ dinucleotide motif type peak was found in chromosome 5 of CHM13 genome. The $[TC_{mix}]$ motif type peaks especially consist of a group of different length motifs with the composition of only T and C nucleotides like $TTCC_n$, $TCCC_n$, C_n, TC_n and T_n, which occur with high frequencies in many chromosomes, and similar states were observed in $[AG_{mix}]$ motif type peaks (Fig. 2C). There are dozens of motif type peaks with low frequencies in some investigated chromosomes as described in Fig. 2D. The (CCA_n, TGG_n, $ATCC_n$, $ATGG_n$) motif-type peaks are the most common tri- and tetranucleotide motif type peaks in most of chromosomes; the tri- and tetranucleotide (ATG_n, CAT_n, $ATAG_n$, $ATCT_n$) motif type peaks also appear in many chromosomes. And the ($AATGG_n$, $ATTCC_n$, $AACCCT_n$, $TTAGGG_n$) peaks are the most common penta- and hexanucleotide motif-type peaks in most of chromosomes no matter in GRCh38 or CHM13 genomes (Fig 2D). Finally, dozens of other motif type peaks were observed to occasionally appear in one or several chromosomes with very low frequencies (Fig. 2E).

3.3. Maps construction for the HDMA peaks

We constructed 24 peak-locus maps to explore the features of high SSRs density peaks accumulation position in the 24 chromosomes of the GRCh38 genome (Fig. 3A-3C), and also constructed 23 peak-locus maps for 23 chromosomes of the CHM13 genome (Fig. 4). Analysis of the peak accumulation position on all the maps revealed that the di-nucleotide motif type HDMA peaks are widely distributed from one end to the other of all the chromosomes of both the two investigated genomes, and the $[AT]^h$ and $[AT]^m$ motif type HDMA peaks are most abundant distributed peaks; their loci are distributed along all the full chromosomes. $[TC_{mix}]$ and $[AG_{mix}]$ motif type HDMA peaks sparsely inset among the di-nucleotide motif type HDMA peaks, which is observed on all investigated chromosomes, except that many $[TC_{mix}]$ motif type HDMA peaks cluster at the upstream centromeric regions of the five acrocentric chromosomes (Chr13, Chr14, Chr15, Chr21, Chr22) of the CHM13 genome. The tri-nucleotide (CCA_n, TGG_n) motif type peaks and the tetra-nucleotide ($ATGG_n$, $ATCC_n$) motif type peaks are scattered among the di-nucleotide motif type HDMA peaks which are observed in almost all chromosomes, except that several motif type peaks ($ATGG_n$, $ATCC_n$, CCA_n, TGG_n) clustered in the centromeric region of Chr11 of CHM13. The penta-nucleotide ($AATGG_n$, $ATTCC_n$) motif type peaks were found clustered at centromeric region of most chromosomes but not found at centromeric region of Chr6, Chr8, Chr9, Chr11, Chr12, Chr18, Chr19, ChrX and ChrY. The hexa-nucleotide ($AACCCT_n$) and ($TTAGGG_n$) motif type HDMA peaks occur at telomeric genomic region of the p-arm and the q-arm in almost all chromosomes of T2T-CHM13 respectively, with only one exception that no ($AACCCT_n$) motif type HDMA peaks observed at telomeric genomic region of the p-arm in the Chr13 of CHM13 replaced by $[AATGG]^m$ motif type peaks clusters.

3.4. Biological significance of HDMA peak region

To explore the possible related biological significance of HDMA peak region in the human genome, we mined the reported data including reported papers, Genome Reference Consortium and UCSC Genome Browser [20,37,40]. The mined data clearly showed that over 60% HDMA peak region (Fig. 3D) were situated in the genomic region of transcription factor binding sites, DNase hypersensitive site, miRNA binding and recombination of the GRCh38 genome (Fig. 3, Table S7). These genomic regions are mostly transcription factor binding sites and DNase hypersensitive site, and only a few of them are miRNA binding site and recombination region; notably, over 80% of HDMA peak region contain the transcription factor binding sites in the Chr16, Chr17, Chr19 and Chr20 respectively. These data suggest that these HDMA peak region of human genome most possibly involve transcription factor binding, miRNA binding, DNAase hypersensitive sites, recombination site etc.

3.5. Variation of HDMA peaks between GRCh38 and CHM13

As above mentioned, total 3433 and 5306 HDMA peaks were found in the GRCh38 and CHM13 human genome respectively. The genomic region of HDMA peaks exhibits a higher sequence variation than up- and down-stream region between the GRCh38 and CHM13 genome (Fig. 6, Table S8). And the variation of HDMA peaks can be divided into three levels, the first is relatively conserved HDMA peaks, where 1656 HDMA peaks of GRCh38 are found relatively high similarity with 1667 high peaks of CHM13 in aligned genomic position, but the genomic region of HDMA peaks is still lower in similarity than up- and down-stream region with sequences difference from dozens of base pairs to hundreds of base pairs (Fig. 6A); the second is dramatically expanding or contracting HDMA peaks, where 895 HDMA peaks of GRCh38 are dramatically expanding or contracting in repeat units with 910 HDMA peaks of CHM13 in aligned genomic position; the third is that the HDMA peaks of GRCh38 and CHM13, where 784 HDMA peaks of GRCh38 and 1,491 HDMA peaks of CHM13 were found only in one of the genome, while middle and lower density accumulation peaks were found in aligned genomic position of the other, suggesting that dramatic expansion or contraction of repeat units occurred in those HDMA peaks position. In the second and third conditions, the sequence variation of HDMA peaks of HDMA is mostly thousands of base pairs in aligned genomic positions of GRCh38 and CHM13. In addition, 1,238 HDMA peaks of CHM13 were aligned to the gap region of GRCh38. Therefore, over half of the HDMA peaks are variant greatly in repeat units and genomic sequences in the corresponding genomic position of GRCh38 and CHM13.

4. Discussion

This work provides the first comprehensive genome-wide view of high-density microsatellites accumulation (HDMA) in the human genome. The first construction of the locus maps of HDMA peaks clearly revealed that these HDMA peaks are distributed regularly along every human chromosome in both the reference (GRCh38) and first complete (CHM13) human genome, suggesting that the high-density microsatellites accumulation is probably the necessary fundamental elements in human genome. We formerly presented a folded slippage model for 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 repeats, which are possibly selected fixation for different biological significance in the long-history evolution [44].

Among all motif types of HDMA peaks, more than half of HDMA peaks are AT_n motif type in the GRC38 and CHM13 human genome, suggesting that AT_n motif is the most specific STRs motif among all STRs motifs and tend to accumulate in human genome. The folded slippage replication possibly explain why short and A/T repeat-motifs occur much more frequently than long and C/G repeat-motifs [44], however, there is no A_n and T_n motif accumulation peak but many AT_n motif accumulation peaks (Fig 2C-2E, Table S6), further indicating that AT_n motifs might have been specifically selected as very essential elements for genome structures and functions to fix and accumulate in these chromosomes. And the comparison of HDMA peaks between the GRCh38 and CHM13 human genomes clarified that the sequence of HDMA peaks are higher variable with exhibiting repeat motifs expanding or contacting dramatically, this is consistent with many former observation that STRs are liable to mutation through repeat units expanding or contacting [45-47], indicating that the variation of HDMA peaks probably affect regulatory and then related to variation of gene expression profiles, and further may act to human phenotypes and human variation. The comparison also indicated that variation of HDMA peaks may be a main type of human genetic variation with large mount sequence difference, so it is far more variable than SNPs and other genomic regions, taking into account the HDMA peaks high possibly related to biological significance of regulation. Therefore, it is remarkable to compare variation of HDMA peaks in human genome, which is most possibly another very important approach to study the relationship between human genomic variation and human phenotypes even diseases, except the method of GWAS.

In the past decades, most works in STRs biology focused on the repeat-associated genetics and pathogenesis of a range of human disorders, and dozens of repeat motifs were revealed relevant to different genetic diseases and cancers [48-50]. However, few of those disease-associated repeat motifs were found in the clarified HDMA peaks, while the abundant peaks and frequently occurred motif type were almost not related to reported human disorder repeats (Fig. 2), so these data indicate that high density repeats accumulation is not certainly related to human diseases. And some of these HDMA peaks, like $AATGG_n$, $ATTCC_n$ motif type peaks cluster at the centromeric region in most of the investigated chromosomes, $AACCCT_n$ and $TTAGGG_n$ motif type peaks especially occur at telomeric region, also possibly involve in much biological significance of human genome structure, organization and so on, as $TTAGGG_n$ in telomeric region was reported as the boundary between human telomere and euchromatic region, and promoter-STRs studies suggested that STRs might play a role in local chromatin structure in many eukaryotic genomes [9,51,52].

The observation here clarified that the variation of HDMA peaks result in hundreds or thousands even more genomic sequences difference and may be associated with biological significance of regulation, genome structure, diseases and so on. So, the comparison of HDMA peak variation will provide an unprecedented opportunity to study human genetic variation and human variation except analysis of SNPs by GWAS, and our strategy to construct HDMA peaks locus maps here will also provide blueprints for genome-wide comparison of variation of HDMA peaks and new insight to deeply explore biological significance of STRs accumulation.

Funding

This work was jointly supported by funding from the “National Key Research and Development Program of China” (grant No. 2016YFC1200200 and 2016YFD0500300, Recipient: Zhongyang Tan).

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

Shaoliang Peng and Zhongyang Tan designed and directed this study. Yun Xia, Hanrou Huang, Douyue Li and Yulin Liang performed the data analysis. Fei Xu, Tingyi Chen and Saichao Pan developed the statistical method of the STRs density. Douyue Li, Ruixue Shi and Saichao Pan drew the STRs landscape maps in GRCh38 and CHM13 human genomes. Yun Xia, Hanrou Huang and Tingyi Chen drew the locus maps of HDMA peaks in GRCh38 and CHM13 human genomes. Zhongyang Tan, Fei Xu, Yun Xia, Hanrou Huang, Douyue Li and Yuling Liang prepared the manuscript.

Data Availability

The annotations of the regulatory regions and specific genomic structures were collected from Genome Browser at University of California, Santa Cruz (http://genome.ucsc.edu/).

Differential Calculator of Microsatellites method is access in https://github.com/DooYal/DCM.

All of the Supplementary Figures (Figure S1-S4) and Tables (Table S1-S8) are access in https://www.jianguoyun.com/p/DZicv3IQtK6xCBiW5YoE.

Acknowledgments

This work was jointly supported by funding from the "National Key Research and Development Program of China".

References

[1] P.M. Visscher, N.R. Wray, Q. Zhang, P. Sklar, M.I. McCarthy, M.A. Brown, J. Yang, 10 Years of GWAS Discovery: Biology, Function, and Translation, The American Journal Of Human Genetics 101 (2017) 5-22. 10.1016/j.ajhg.2017.06.005.

[2] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, D. Meyre, Benefits and limitations of genome-wide association studies, Nat. Rev. Genet. 20 (2019) 467-484. 10.1038/s41576-019-0127-1.

[3] M.D. Gallagher, A.S. Chen-Plotkin, The Post-GWAS Era: From Association to Function, The American Journal of Human Genetics 102 (2018) 717-730. https://doi.org/10.1016/j.ajhg.2018.04.002.

[4] M.O. Press, R.C. McCoy, A.N. Hall, J.M. Akey, C. Queitsch, Massive variation of short tandem repeats with functional consequences across strains of Arabidopsis thaliana, Genome Res. 28 (2018) 1169-1178. 10.1101/gr.231753.117.

[5] A.J. Hannan, Tandem repeats mediating genetic plasticity in health and disease, Nat. Rev. Genet. 19 (2018) 286-298. 10.1038/nrg.2017.115.

[6] M. Gymrek, T. Willems, D. Reich, Y. Erlich, Interpreting short tandem repeat variations in humans using mutational constraint, Nat. Genet. 49 (2017) 1495-1501. 10.1038/ng.3952.

[7] H. Ellegren, Microsatellites: Simple sequences with complex evolution, Nat. Rev. Genet. 5 (2004) 435-445. 10.1038/nrg1348.

[8] Y.C. Li, A.B. Korol, T. Fahima, E. Nevo, Microsatellites within genes: Structure, function, and evolution, Mol. Biol. Evol. 21 (2004) 991-1007. 10.1093/molbev/msh073.

[9] F.S. Collins, E.S. Lander, J. Rogers, R.H. Waterston, International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, Nature 431 (2004) 931-945. 10.1038/nature03001.

[10] M. Gymrek, T. Willems, A. Guilmatre, H. Zeng, B. Markus, S. Georgiev, M.J. Daly, A.L. Price, J.K. Pritchard, A.J. Sharp, Y. Erlich, Abundant contribution of short tandem repeats to gene expression variation in humans, Nat. Genet. 48 (2016) 22-29. 10.1038/ng.3461.

[11] S.F. Fotsing, J. Margoliash, C. Wang, S. Saini, R. Yanicky, S. Shleizer-Burko, A. Goren, M. Gymrek, The impact of short tandem repeat variation on gene expression, Nat. Genet. 51 (2019) 1652-+. 10.1038/s41588-019-0521-9.

[12] K. Yap, S. Mukhina, G. Zhang, J.S.C. Tan, H.S. Ong, E.V. Makeyev, A short tandem repeat-enriched RNA assembles a nuclear compartment to control alternative splicing and promote cell survival, Mol. Cell 72 (2018) 525-540. 10.1016/j.molcel.2018.08.041.

[13] I. Mitra, B. Huang, N. Mousavi, N. Ma, M. Lamkin, R. Yanicky, S. Shleizer-Burko, K.E. Lohmueller, M. Gymrek, Patterns of de novo tandem repeat mutations and their role in autism, Nature 589 (2021) 246-250. 10.1038/s41586-020-03078-7.

[14] C.P. Schaaf, C. Betancur, R.K.C. Yuen, J.R. Parr, D.H. Skuse, L. Gallagher, R.A. Bernier, J.A. Buchanan, J.D. Buxbaum, C.A. Chen, K.A. Dies, M. Elsabbagh, H.V. Firth, T. Frazier, N. Hoang, J. Howe, C.R. Marshall, J.L. Michaud, O. Rennie, P. Szatmari, W.K. Chung, P.F. Bolton, E.H. Cook, S.W. Scherer, J.A.S. Vorstman, A framework for an evidence-based gene list relevant to autism spectrum disorder, Nat. Rev. Genet. 21 (2020) 367-376. 10.1038/s41576-020-0231-2.

[15] A.J. Hannan, Repeat DNA expands our understanding of autism spectrum disorder, Nature 589 (2021) 200-202. 10.1038/d41586-020-03658-7.

[16] N. van Wietmarschen, S. Sridharan, W.J. Nathan, A. Tubbs, E.M. Chan, E. Callen, W. Wu, F. Belinky, V. Tripathi, N. Wong, K. Foster, J. Noorbakhsh, K. Garimella, A. Cruz-Migoni, J.A. Sommers, Y. Huang, A.A. Borah, J.T. Smith, J. Kalfon, N. Kesten, K. Fugger, R.L. Walker, E. Dolzhenko, M.A. Eberle, B.E. Hayward, K. Usdin, C.H. Freudenreich, R.M. Brosh, Jr., S.C. West, P.J. McHugh, P.S. Meltzer, A.J. Bass, A. Nussenzweig, Repeat expansions confer WRN dependence in microsatellite-unstable cancers, Nature 586 (2020) 292-298. 10.1038/s41586-020-2769-8.

[17] R.P. Kumar, J. Krishnan, N. Pratap Singh, L. Singh, R.K. Mishra, GATA simple sequence repeats function as enhancer blocker boundaries, Nat. Commun. 4 (2013) 1844. 10.1038/ncomms2872.

[18] P. Martin, K. Makepeace, S.A. Hill, D.W. Hood, E.R. Moxon, Microsatellite instability regulates transcription factor binding and gene expression, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 3800-3804. 10.1073/pnas.0406805102.

[19] D.R. Zerbino, N. Johnson, T. Juetteman, D. Sheppard, S.P. Wilder, I. Lavidas, M. Nuhn, E. Perry, Q. Raffaillac-Desfosses, D. Sobral, D. Keefe, S. Graf, I. Ahmed, R. Kinsella, B. Pritchard, S. Brent, R. Amode, A. Parker, S. Trevanion, E. Birney, I. Dunham, P. Flicek, Ensembl regulation resources, Database (Oxford) 2016 (2016) bav119. 10.1093/database/bav119.

[20] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (2002) 996-1006. 10.1101/gr.229102.

[21] A. Sulovari, R. Li, P.A. Audano, D. Porubsky, M.R. Vollger, G.A. Logsdon, C. Human Genome Structural Variation, W.C. Warren, A.A. Pollen, M.J.P. Chaisson, E.E. Eichler, Human-specific tandem repeat expansion and differential gene expression during primate evolution, Proc. Natl. Acad. Sci. U. S. A. 116 (2019) 23243-23253. 10.1073/pnas.1912175116.

[22] J.H. Sun, L.D. Zhou, D.J. Emerson, S.A. Phyo, K.R. Titus, W.F. Gong, T.G. Gilgenast, J.A. Beagan, B.L. Davidson, F. Tassone, J.E. Phillips-Cremins, Disease-associated short tandem repeats co-localize with chromatin domain boundaries, Cell 175 (2018) 224-238. 10.1016/j.cell.2018.08.005.

[23] The Genbank Staff, Homo sapiens GRCh38.p13 genome summary. https://www.ncbi.nlm.nih.gov/genome/?term=human, 2019.

[24] The Genbank Staff, Homo sapiens GRCh38.p13 assembly report. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39/#/st, 2019.

[25] S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A.V. Bzikadze, A. Mikheenko, M.R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S.J. Hoyt, M. Diekhans, G.A. Logsdon, M. Alonge, S.E. Antonarakis, M. Borchers, G.G. Bouffard, S.Y. Brooks, G.V. Caldas, N.C. Chen, H. Cheng, C.S. Chin, W. Chow, L.G. de Lima, P.C. Dishuck, R. Durbin, T. Dvorkina, I.T. Fiddes, G. Formenti, R.S. Fulton, A. Fungtammasan, E. Garrison, P.G.S. Grady, T.A. Graves-Lindsay, I.M. Hall, N.F. Hansen, G.A. Hartley, M. Haukness, K. Howe, M.W. Hunkapiller, C. Jain, M. Jain, E.D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V.V. Maduro, T. Marschall, A.M. McCartney, J. McDaniel, D.E. Miller, J.C. Mullikin, E.W. Myers, N.D. Olson, B. Paten, P. Peluso, P.A. Pevzner, D. Porubsky, T. Potapova, E.I. Rogaev, J.A. Rosenfeld, S.L. Salzberg, V.A. Schneider, F.J. Sedlazeck, K. Shafin, C.J. Shew, A. Shumate, Y. Sims, A.F.A. Smit, D.C. Soto, I. Sovic, J.M. Storer, A. Streets, B.A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B.P. Walenz, A. Wenger, J.M.D. Wood, C. Xiao, S.M. Yan, A.C. Young, S. Zarate, U. Surti, R.C. McCoy, M.Y. Dennis, I.A. Alexandrov, J.L. Gerton, R.J. O'Neill, W. Timp, J.M. Zook, M.C. Schatz, E.E. Eichler, K.H. Miga, A.M. Phillippy, The complete sequence of a human genome, Science 376 (2022) 44-53. 10.1126/science.abj6987.

[26] S. Aganezov, S.M. Yan, D.C. Soto, M. Kirsche, S. Zarate, P. Avdeyev, D.J. Taylor, K. Shafin, A. Shumate, C. Xiao, J. Wagner, J. McDaniel, N.D. Olson, M.E.G. Sauria, M.R. Vollger, A. Rhie, M. Meredith, S. Martin, J. Lee, S. Koren, J.A. Rosenfeld, B. Paten, R. Layer, C.-S. Chin, F.J. Sedlazeck, N.F. Hansen, D.E. Miller, A.M. Phillippy, K.H. Miga, R.C. McCoy, M.Y. Dennis, J.M. Zook, M.C. Schatz, A complete reference genome improves analysis of human genetic variation, Science 376 (2022). 10.1126/science.abl3533.

[27] A. Gershman, M.E.G. Sauria, X. Guitart, M.R. Vollger, P.W. Hook, S.J. Hoyt, M. Jain, A. Shumate, R. Razaghi, S. Koren, N. Altemose, G.V. Caldas, G.A. Logsdon, A. Rhie, E.E. Eichler, M.C. Schatz, R.J. O'Neill, A.M. Phillippy, K.H. Miga, W. Timp, Epigenetic patterns in a complete human genome, Science 376 (2022) eabj5089. 10.1126/science.abj5089.

[28] M.R. Vollger, X. Guitart, P.C. Dishuck, L. Mercuri, W.T. Harvey, A. Gershman, M. Diekhans, A. Sulovari, K.M. Munson, A.P. Lewis, K. Hoekzema, D. Porubsky, R. Li, S. Nurk, S. Koren, K.H. Miga, A.M. Phillippy, W. Timp, M. Ventura, E.E. Eichler, Segmental duplications and their variation in a complete human genome, Science 376 (2022). 10.1126/science.abj6965.

[29] D.Y. Li, W.L. Jiao, S.Q. Zhou, Y.Z. Fu, S.L. Peng, Y.S. Peng, X.X. He, L. Zhou, Z.L. Peng, J.Y. Zhou, H.X. Zhang, L. Zhang, J.L. Zhou, M.F. Li, Z.Y. Tan, Comparative analysis on precise distribution-patterns of microsatellites in HIV-1 with differential statistical method, Gene Reports 12 (2018) 141-148.

[30] D. Li, S. Pan, H. Zhang, Y. Fu, Z. Peng, L. Zhang, S. Peng, F. Xu, H. Huang, R. Shi, H. Zheng, Y. Peng, Z. Tan, A comprehensive microsatellite landscape of human Y-DNA at kilobase resolution, BMC Genomics 22 (2021) 76. 10.1186/s12864-021-07389-5.

[31] S.B. Mudunuri, H.A. Nagarajaram, IMEx: Imperfect Microsatellite Extractor, Bioinformatics 23 (2007) 1181-1187. 10.1093/bioinformatics/btm097.

[32] X.Y. Zhao, Y.L. Tian, R.H. Yang, H.P. Feng, Q.J. Ouyang, Y. Tian, Z.Y. Tan, M.F. Li, Y.L. Niu, J.H. Jiang, G.L. Shen, R.Q. Yu, Coevolution between simple sequence repeats (SSRs) and virus genome size, BMC Genomics 13 (2012) 435. Artn 435
10.1186/1471-2164-13-435.

[33] P. Rajendrakumar, A.K. Biswal, S.M. Balachandran, K. Srinivasarao, R.M. Sundaram, Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions, Bioinformatics 23 (2007) 1-4. 10.1093/bioinformatics/btl547.

[34] S. Tempel, Using and understanding RepeatMasker, Methods Mol. Biol. 859 (2012) 29-51. 10.1007/978-1-61779-603-6_2.

[35] G. Benson, Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Res. 27 (1999) 573-580. 10.1093/nar/27.2.573.

[36] The R Core Team, R: A language and environment for statistical computing. http://www.R-project.org/, 2016.

[37] D. Karolchik, A.S. Hinrichs, T.S. Furey, K.M. Roskin, C.W. Sugnet, D. Haussler, W.J. Kent, The UCSC Table Browser data retrieval tool, Nucleic Acids Res. 32 (2004) D493-D496. 10.1093/nar/gkh103.

[38] C.A. Sloan, E.T. Chan, J.M. Davidson, V.S. Malladi, J.S. Strattan, B.C. Hitz, I. Gabdank, A.K. Narayanan, M. Ho, B.T. Lee, L.D. Rowe, T.R. Dreszer, G. Roe, N.R. Podduturi, F. Tanaka, E.L. Hong, J.M. Cherry, ENCODE data at the ENCODE portal, Nucleic Acids Res. 44 (2016) D726-D732. 10.1093/nar/gkv1160.

[39] O.L. Griffith, S.B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M.C. Sleumer, M. Bilenky, M. Haeussler, M. Griffith, S.M. Gallo, B. Giardine, B. Hooghe, P. Van Loo, E. Blanco, A. Ticoll, S. Lithwick, E. Portales-Casamar, I.J. Donaldson, G. Robertson, C. Wadelius, P. De Bleser, D. Vlieghe, M.S. Halfon, W. Wasserman, R. Hardison, C.M. Bergman, S.J.M. Jones, O.R.A. Consort, ORegAnno: an open-access community-driven resource for regulatory annotation, Nucleic Acids Res. 36 (2008) D107-D113. 10.1093/nar/gkm967.

[40] The Genbank Staff, NCBI Homo sapiens updated annotation release 109.20210226. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/109.20210226/, 2019.

[41] A.L. Delcher, S. Kasif, R.D. Fleischmann, J. Peterson, O. White, S.L. Salzberg, Alignment of whole genomes, Nucleic Acids Res 27 (1999) 2369-2376. 10.1093/nar/27.11.2369.

[42] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Clustal W and Clustal X version 2.0, Bioinformatics 23 (2007) 2947-2948. 10.1093/bioinformatics/btm404.

[43] S. Subramanian, R.K. Mishra, L. Singh, Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions, Genome Biol. 4 (2003) R13. 10.1186/gb-2003-4-2-r13.

[44] H.X. Zhang, D.Y. Li, X.Y. Zhao, S.C. Pan, X.L. Wu, S. Peng, H.R. Huang, R.X. Shi, Z.Y. Tan, Relatively semi-conservative replication and a folded slippage model for short tandem repeats, BMC Genomics 21 (2020) 563. 10.1186/s12864-020-06949-5.

[45] T. Willems, D. Zielinski, J. Yuan, A. Gordon, M. Gymrek, Y. Erlich, Genome-wide profiling of heritable and de novo STR variations, Nat. Methods 14 (2017) 590-592. 10.1038/nmeth.4267.

[46] A. Fungtammasan, M. Tomaszkiewicz, R. Campos-Sanchez, K.A. Eckert, M. DeGiorgio, K.D. Makova, Reverse Transcription Errors and RNA-DNA Differences at Short Tandem Repeats, Mol. Biol. Evol. 33 (2016) 2744-2758.

10.1093/molbev/msw139.

[47] Y.D. Kelkar, K.A. Eckert, F. Chiaromonte, K.D. Makova, A matter of life or death: how microsatellites emerge in and vanish from the human genome, Genome Res. 21 (2011) 2038-2048. 10.1101/gr.122937.111.

[48] G.E.B. Wright, J.A. Collins, C. Kay, C. McDonald, E. Dolzhenko, Q.W. Xia, K. Becanovic, B.I. Drogemoller, A. Semaka, C.M. Nguyen, B. Trost, F. Richards, E.K. Bijlsma, F. Squitieri, C.J.D. Ross, S.W. Scherer, M.A. Eberle, R.K.C. Yuen, M.R. Hayden, Length of uninterrupted CAG, independent of polyglutamine size, results in increased somatic instability, hastening onset of Huntington disease, Am. J. Hum. Genet. 104 (2019) 1116-1126. 10.1016/j.ajhg.2019.04.007.

[49] H. Jacobi, S.T. du Montcel, S. Romanzetti, F. Harmuth, C. Mariotti, L. Nanetti, M. Rakowicz, G. Makowicz, A. Durr, M.L. Monin, A. Filla, A. Roca, L. Schols, H. Hengel, J. Infante, J.S. Kang, D. Timmann, C. Casali, M. Masciullo, L. Baliko, B. Melegh, W. Nachbauer, K. Burk-Gergs, J.B. Schulz, O. Riess, K. Reetz, T. Klockgether, Conversion of individuals at risk for spinocerebellar ataxia types 1, 2, 3, and 6 to manifest ataxia (RISCA): a longitudinal cohort study, The Lancet. Neurology 19 (2020) 738-747. 10.1016/S1474-4422(20)30235-0.

[50] T.M. Kim, P.W. Laird, P.J. Park, The landscape of microsatellite instability in colorectal and endometrial cancer genomes, Cell 155 (2013) 858-868. 10.1016/j.cell.2013.10.015.

[51] A. Ambrosini, S. Paul, S.F. Hu, H. Riethman, Human subtelomeric duplicon structure and organization, Genome Biol. 8 (2007). 10.1186/Gb-2007-8-7-R151.

[52] M.D. Vinces, M. Legendre, M. Caldara, M. Hagihara, K.J. Verstrepen, Unstable tandem repeats in promoters confer transcriptional evolvability, Science 324 (2009) 1213-1216. 10.1126/science.1170097.

Figure legends

Figure 1. The representative high density microsatellites accumulation (HDMA) peaks in STRs landscape maps of human genome. It listed the representative HDMA peaks in the STRs landscape maps of GRCh38 (G38) and CHM13 (C13) human genomes and the corresponding sequence. Totally >100,000 STRs landscape maps in the 2 genomes were built in Figs. S1.01-S1.24.

Figure 2. The different motif type HDMA peaks in every G38 and C13 chromosome. A. The total quantitative features of STRs and HDMA peaks in the G38 and C13 human genomes. The form of “data a / data b” meant “G38 data / C13 data”. Chr size, Seq size meant chromosome size (Mbp) and sequenced size of each chromosome (Mbp); LS map num meant STRs-landscape map number, W STRs RD represented STRs relative density (bp/Mbp) under the widely accepted standard, HDMA num was HDMA peak number, and HDMA dis was average distance (Mbp) between the neighboring peaks; the details were in Tables S1-S5. **B**. The composition features representative HDMA peaks. E_position was the end position of each peak. In integrated motif types, $[AT]^h$ meant ≥70% of motifs in a peak are AT_n, and $[AT]^m$ meant AT_n in a peak is in the highest percentage but <70%, $[AG_{mix}]$ meant many motifs in a peak are specifically composed by A and G. **C.** It listed high percentage motif types and the percentage >10% were in bold fonts. **D.** It listed middle-level percentage motif types, and the disappearance of penta- and hexanucleotide HDMA peaks in C13 chromosomes were in underline fonts. **E***. It listed low-frequency motif types in only the chromosomes containing such motifs, e.g. G38-05 represented the chromosome 5 of G38 genomes. **ii, iii, iv** meant very low frequency motif types. The details of motif type statistics were in Table S6.

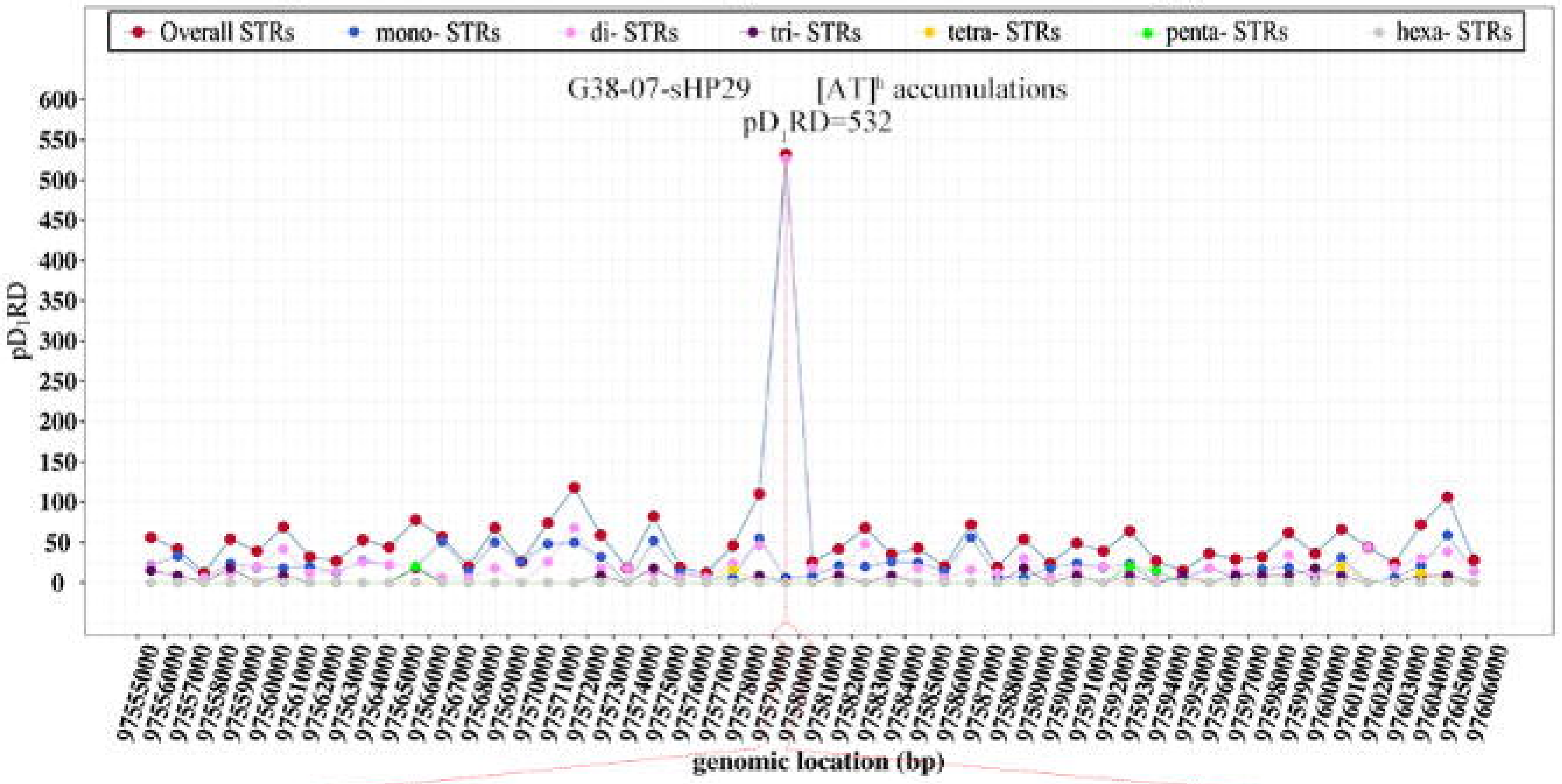
Figure 3. The HDMA peak-locus maps in GRCh38 (G38) human genome with possible biological significance. A. HDMA peak-locus maps in chromosomes 1-7 of G38 genome. **B.** HDMA peak-locus maps in chromosomes 8-13 and X of G38 genome. **C.** HDMA peak-locus maps in chromosomes 14-22 and Y of G38 genome. In these maps in **A, B, C**, the information of each investigated chromosome was listed in Table S1A; the information of non-sequenced gaps was listed in Table S1B, P.S. G38-01-G1~3 represented there are very close gaps (G38-01-G1, G38-01-G2, G38-01-G3), while the sequenced regions among them are too narrow to show in the maps; the features of each HDMA peak, including biological significance, motif type,

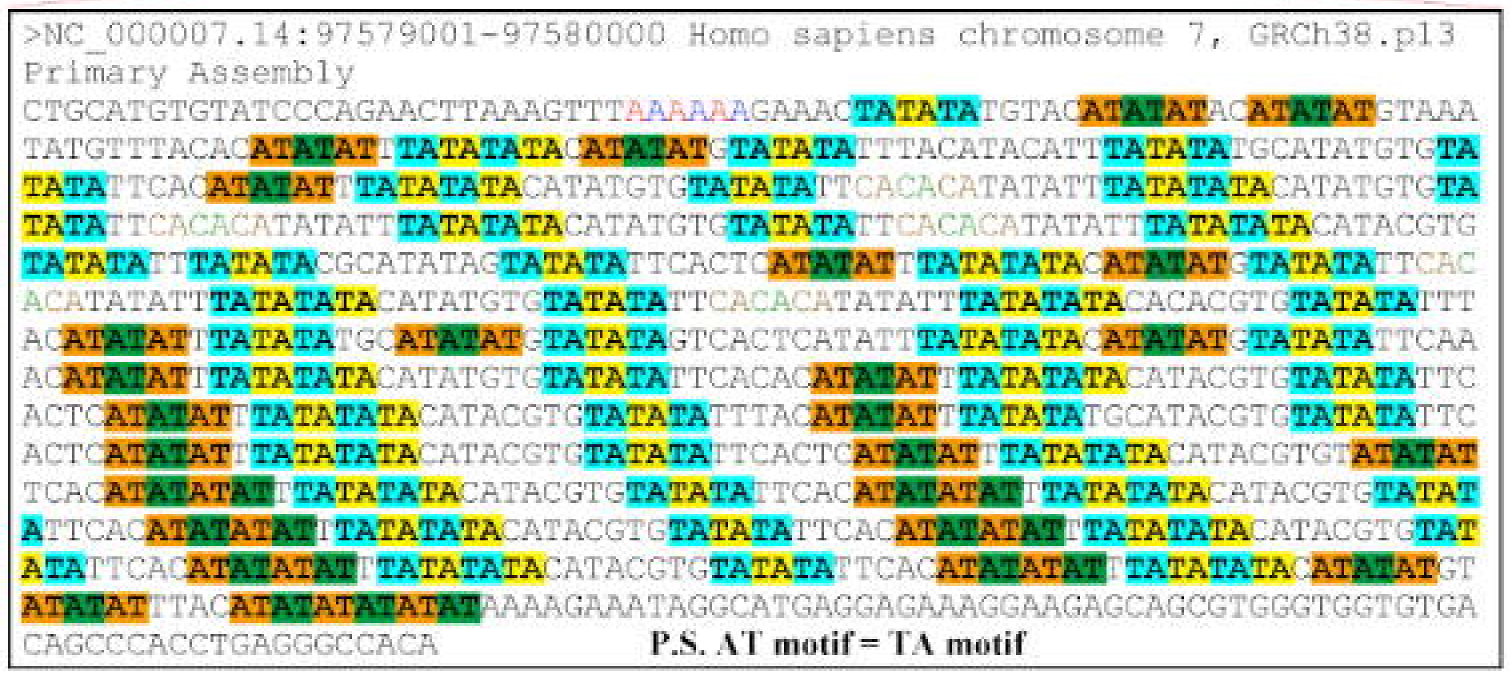
peak name (sHP and HP), were listed at the left side of each chromosome, whose details were in Tables S4, S6 and S7; the separated HDMA peak-locus maps in each investigated chromosome were in Figs. S2.01-S2.24. **D**. The statistics of HDMA peak-locus maps in G38 human genome with possible biological significance.

Figure 4. The peak-locus maps of HDMA peaks in CHM13 (C13) human genome A. HDMA peak-locus maps in chromosomes 1-8 of C13 genome. **B.** HDMA peak-locus maps in chromosome 9-16 of C13 genome. **C.** HDMA peak-locus maps in chromosomes 17-22 and X of C13 genome. The details of these HDMA peaks were listed in Table S5.

Figure 5. Frequency distribution of distance values between the neighboring peaks in these 2 genomes.

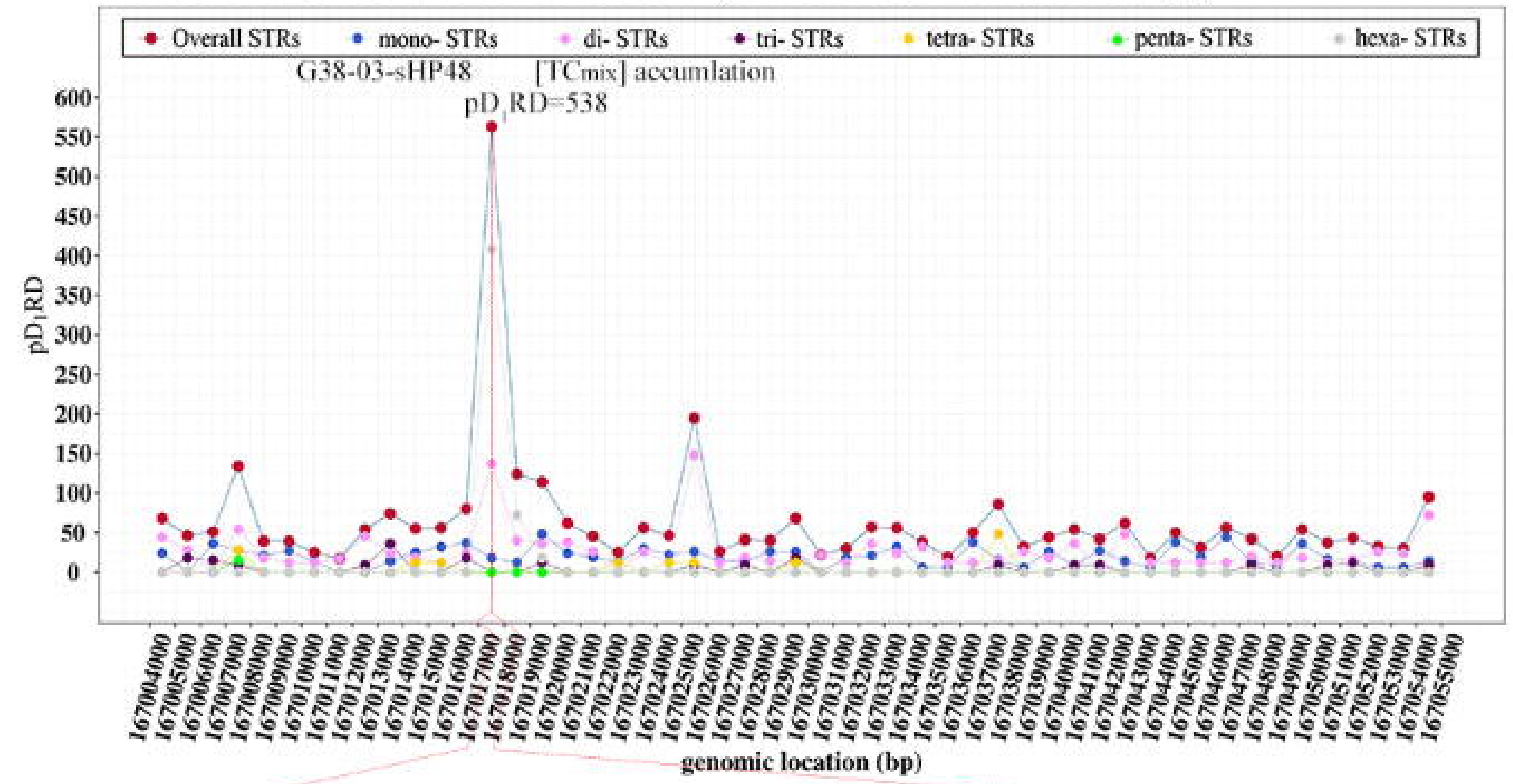
Figure 6. The comparison of HDMA peaks in C13 and G38 human genomes. A. The representative STRs landscape map of HDMA peaks in the two samples whose aligned sequence showed relatively conserved (identity ≥70%). **B**. The representative map of the peaks whose aligned sequence showed specific repeat expanded/contracted (identity<70%), and in the following sequence list, STRs were in colorful background. **C**. The peak number with different aligned features in C13 genome and those in G38 genome. In this table, HDMA num represented the number of HDMA peaks; rel_cons represented relatively conserved, exp/cont represented expanded/contracted, and there are some segmental duplication in C13 genome causing multiple peaks in C13 to align to the same peaks in G38; gap_seq represented C13 peaks aligned with non-sequenced gaps of G38 genome; HP>MP represented "number of G38 HDMA peak exactly aligned with the C13 lower peaks" vs. "number of C13 HDMA peak exactly aligned to the G38 lower peaks". And the comparisons were exactly listed in Table S8 and Figs. S3.01-S3.23. **D**. The non-conserved HDMA peaks possibly contribute to human variation.

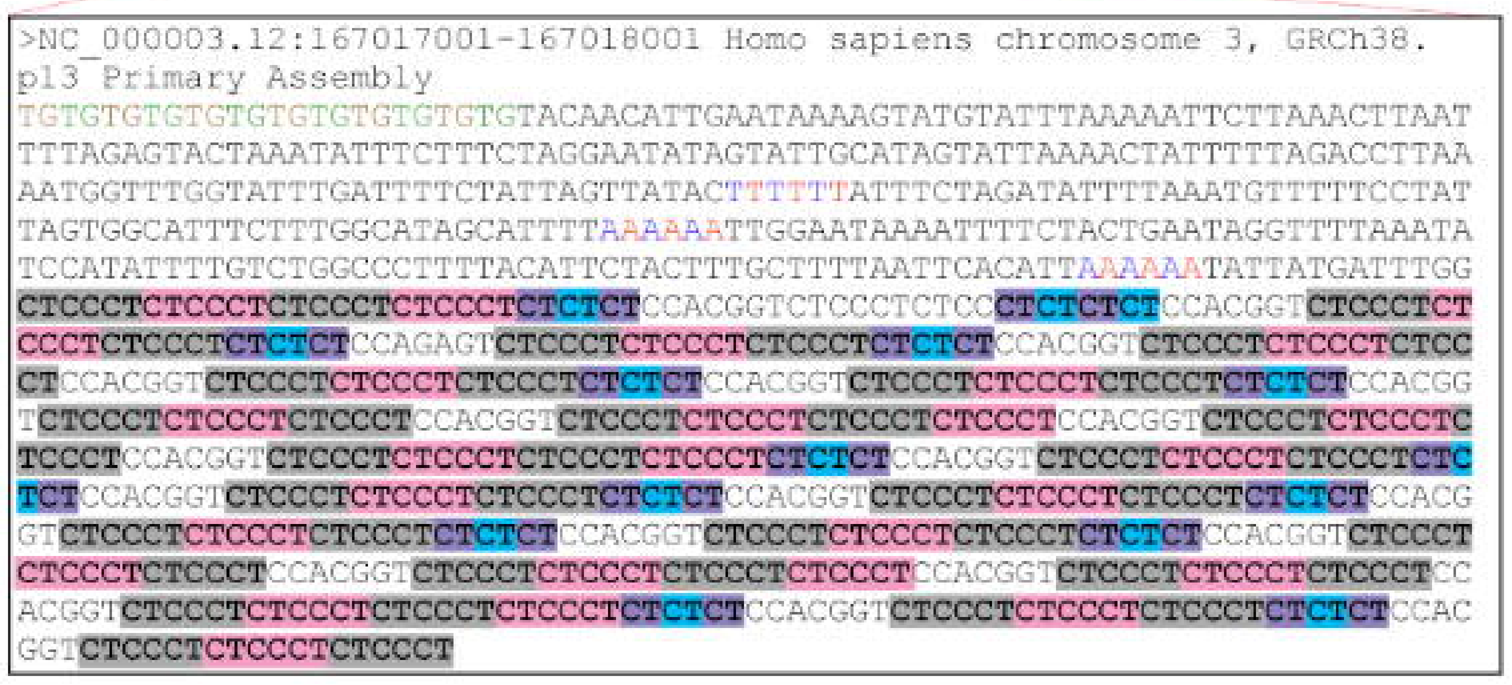

GRCh38_chr07: G38-07-Z1948 (97555001-97606000 bp)



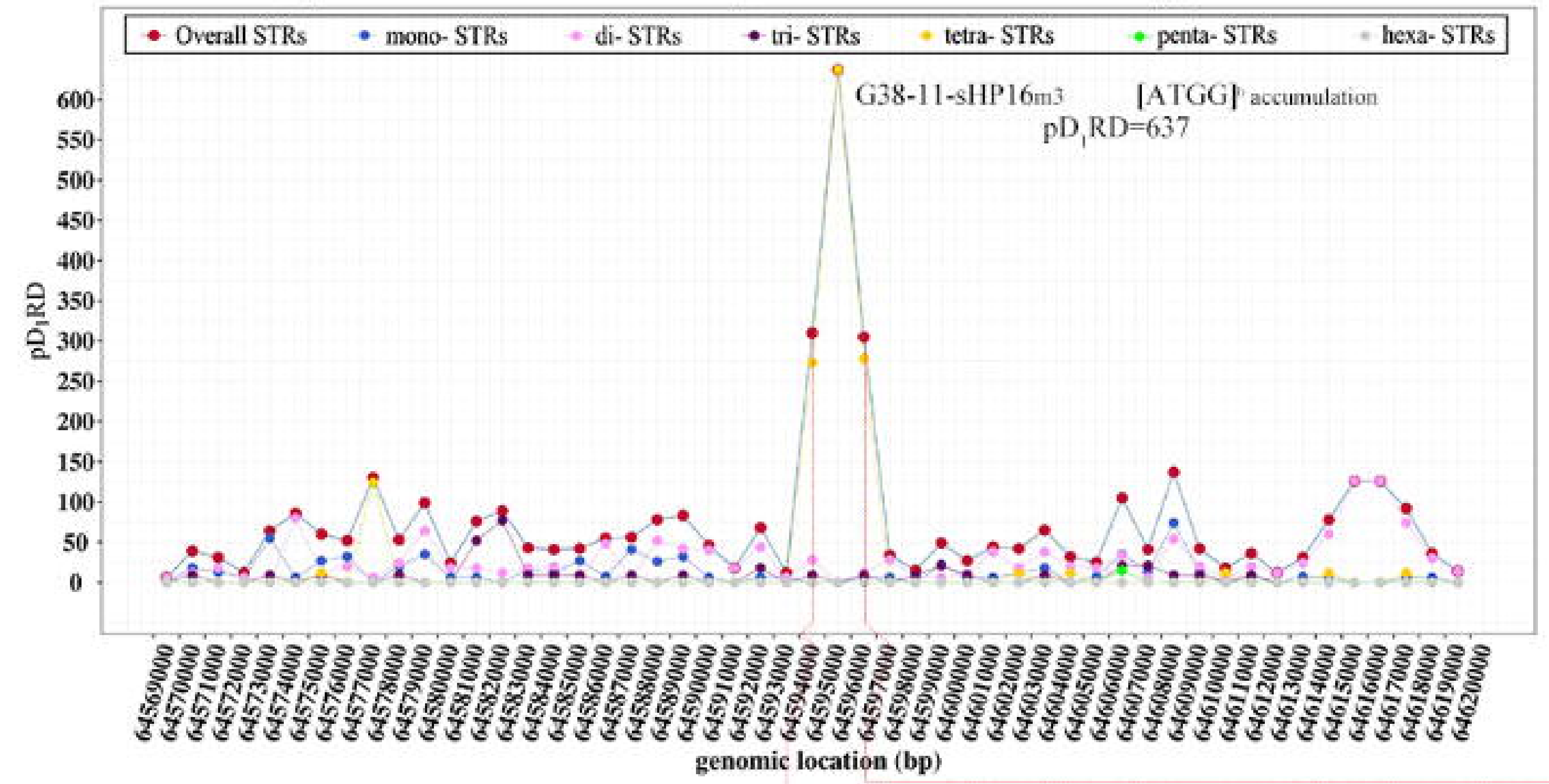
```
>NC_000007.14:97579001-97580000 Homo sapiens chromosome 7, GRCh38.p13
Primary Assembly
CTGCATGTGTATCCCAGAACTTAAAGTTTAAAAAAGAAACTATATATGTACATATATACATATATGTAAA
TATGTTTACACATATATTTATATATACATATATGTATATATTTACATACATTTATATATGCATATGTGTA
TATATTCACATATATTTATATATACATATGTGTATATATTCACACATATATTTATATATACATATGTGTA
TATATTCACACATATATTTATATATACATATGTGTATATATTCACACATATATTTATATATACATACGTG
TATATATTTATATACGCATATAGTATATATTCACTCATATATTTATATATACATATATGTATATATTCAC
ACATATATTTATATATACATATGTGTATATATTCACACATATATTTATATATACACACGTGTATATATTT
ACATATATTTATATATGCATATATGTATATAGTCACTCATATTTATATATACATATATGTATATATTCAA
ACATATATTTATATATACATATGTGTATATATTCACACATATATTTATATATACATACGTGTATATATTC
ACTCATATATTTATATATACATACGTGTATATATTTACATATATTTATATATGCATACGTGTATATATTC
ACTCATATATTTATATATACATACGTGTATATATTCACTCATATATTTATATATACATACGTGTATATAT
TCACATATATATTTATATATACATACGTGTATATATTCACATATATATTTATATATACATACGTGTATAT
ATTCACATATATATTTATATATACATACGTGTATATATTCACATATATATTTATATATACATACGTGTAT
ATATTCACATATATATTTATATATACATACGTGTATATATTCACATATATATTTATATATACATATATGT
ATATATTTACATATATATATATATAAAAGAAATAGGCATGAGGAGAAAGGAAGAGCAGCGTGGGTGGTGA
CAGCCCACCTGAGGGCCACA
```

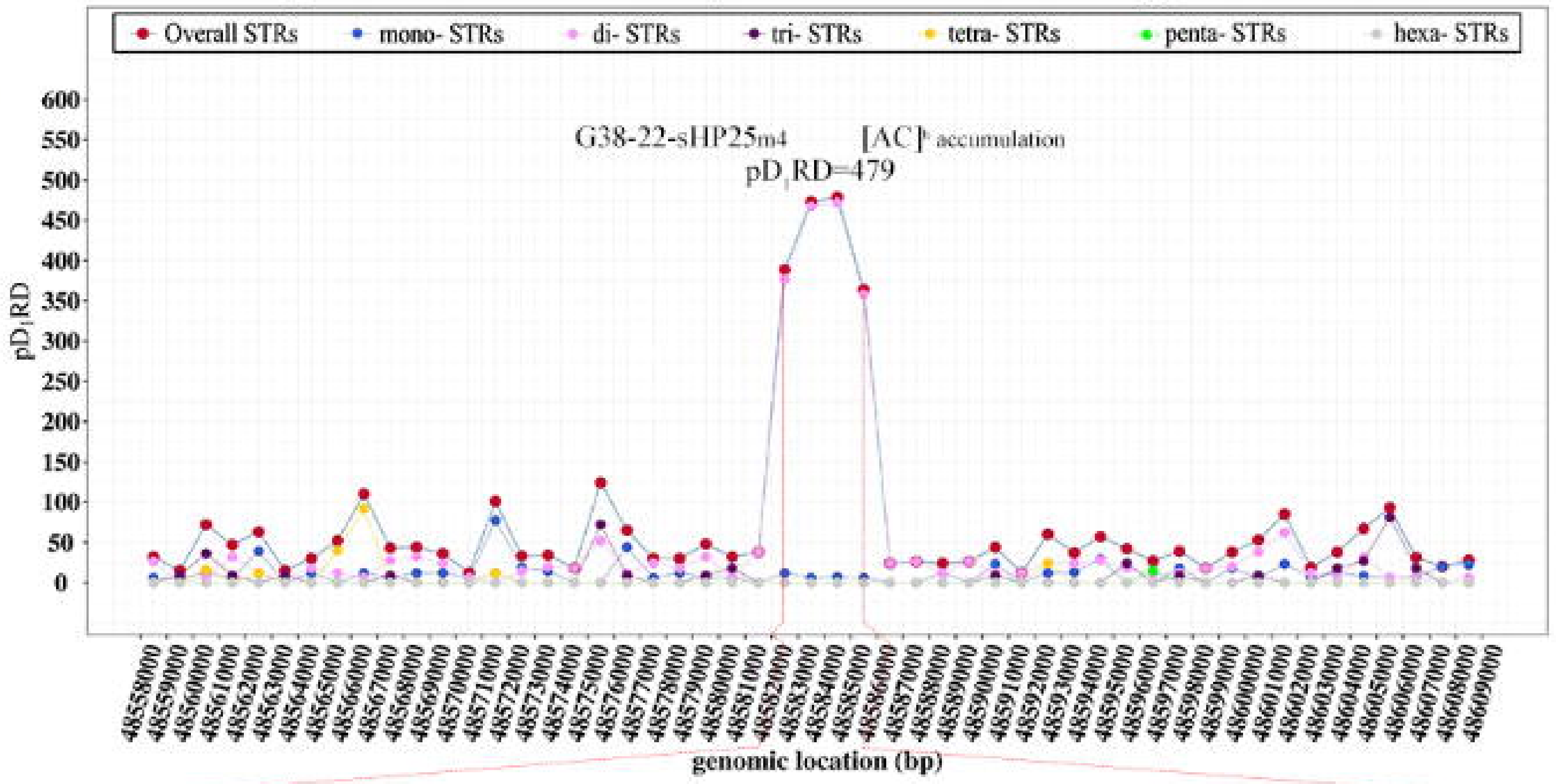
P.S. AT motif = TA motif

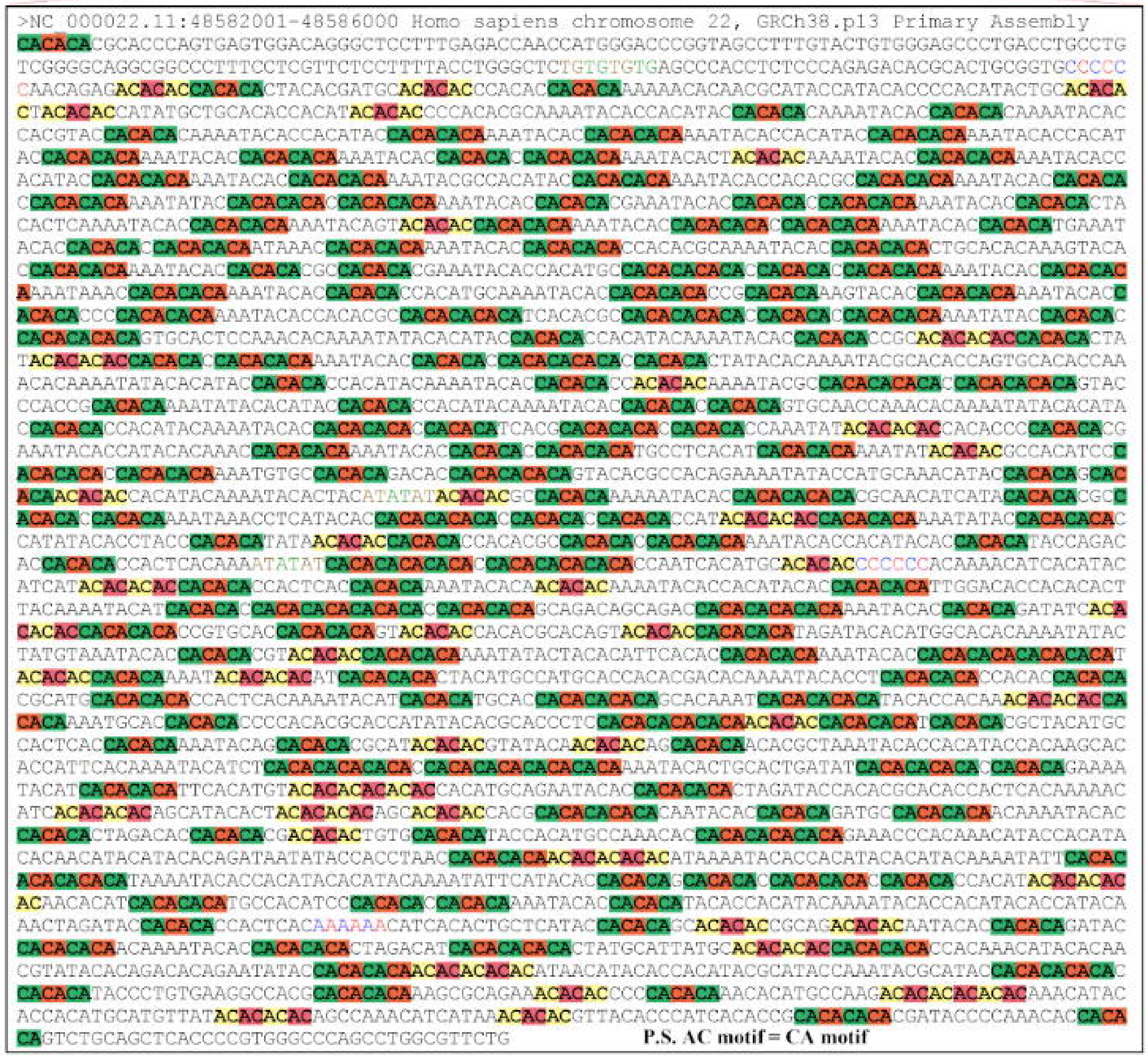

GRCh38_chr03: G38-03-Z3337 (167004001-167055000 bp)



```
>NC_000003.12:167017001-167018001 Homo sapiens chromosome 3, GRCh38.
p13 Primary Assembly
TGTGTGTGTGTGTGTGTGTGTGTGTACAACATTGAATAAAAGTATGTATTTAAAAATTCTTAAACTTAAT
TTTAGAGTACTAAATATTTCTTTCTAGGAATATAGTATTGCATAGTATTAAAACTATTTTTAGACCTTAA
AATGGTTTGGTATTTGATTTTCTATTAGTTATACTTTTTTATTTCTAGATATTTTAAATGTTTTTCCTAT
TAGTGGCATTTCTTTGGCATAGCATTTTAAAAAATTGGAATAAAATTTTCTACTGAATAGGTTTTAAATA
TCCATATTTTGTCTGGCCCTTTTACATTCTACTTTGCTTTTAATTCACATTAAAAAATATTATGATTTGG
CTCCCTCTCCCTCTCCCTCTCCCTCTCTCTCCACGGTCTCCCTCTCCCTCTCTCTCCACGGTCTCCCTCT
CCCTCTCCCTCTCTCTCCAGAGTCTCCCTCTCCCTCTCCCTCTCTCTCCACGGTCTCCCTCTCCCTCTCC
CTCCACGGTCTCCCTCTCCCTCTCCCTCTCTCTCCACGGTCTCCCTCTCCCTCTCCCTCTCTCTCCACGG
TCTCCCTCTCCCTCTCCCTCCACGGTCTCCCTCTCCCTCTCCCTCTCCCTCCACGGTCTCCCTCTCCCTC
TCCCTCCACGGTCTCCCTCTCCCTCTCCCTCTCCCTCTCTCTCCACGGTCTCCCTCTCCCTCTCCCTCTC
TCTCCACGGTCTCCCTCTCCCTCTCCCTCTCTCTCCACGGTCTCCCTCTCCCTCTCCCTCTCTCTCCACG
GTCTCCCTCTCCCTCTCCCTCTCTCTCCACGGTCTCCCTCTCCCTCTCCCTCTCTCTCCACGGTCTCCCT
CTCCCTCTCCCTCCACGGTCTCCCTCTCCCTCTCCCTCTCCCTCCACGGTCTCCCTCTCCCTCTCCCTCC
ACGGTCTCCCTCTCCCTCTCCCTCTCCCTCTCTCTCCACGGTCTCCCTCTCCCTCTCCCTCTCTCTCCAC
GGTCTCCCTCTCCCTCTCCCT
```

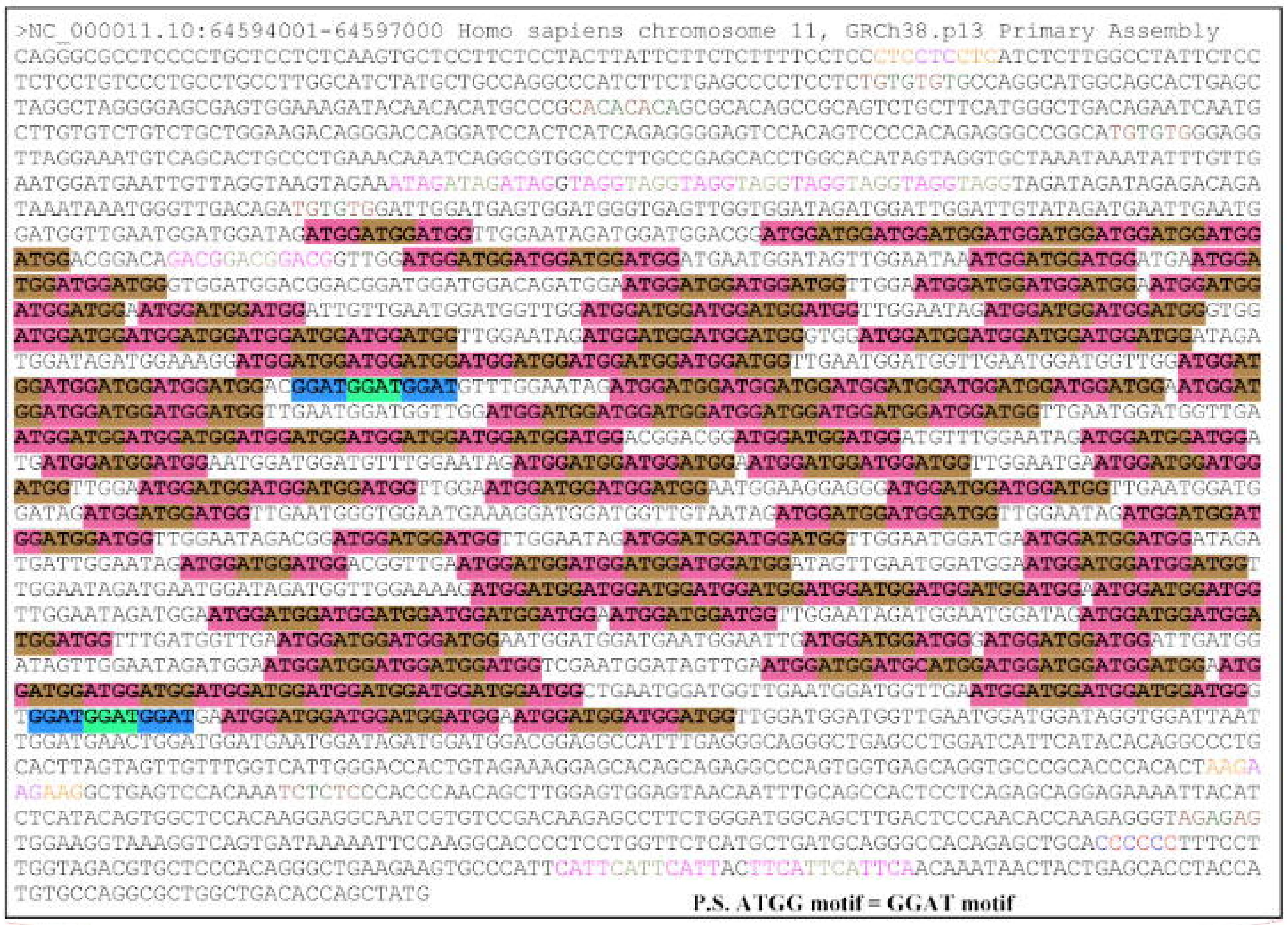

GRCh38_chr11: G38-11-Z1284 (64569001-64620000 bp)


GRCh38_chr22: G38-22-Z739 (48558001-48609000 bp)



```
>NC_000022.11:48582001-48586000 Homo sapiens chromosome 22, GRCh38.p13 Primary Assembly
CACACACGCACCCAGTGAGTGGACAGGGCTCCTTTGAGACCAACCATGGACCCGGTAGCCTTTGTACTGTGGGAGCCCTGACCTGCCTG
TCGGGGCAGGCGGCCCTTTCCTCGTTCTCCTTTTACCTGGGCTCTGTGTGTGAGCCCACCTCTCCCAGAGACACGGCACTGCGGTGCCCCC
CAACAGAGACACACCACACACTACACGATGCACACACCCACACCACACAAAAAACACAACGCATACCATACACCCCACATACTGCACACA
CTACACACCATATGCTGCACACCACATACACACCCCACACGCAAAATACACCACATACCACACACAAAATACACCACACACAAAATACAC
CACGTACCACACACAAAATACACCACATACCACACACAAAATACACCACACACAAAATACACCACATACCACACACAAAATACACCACAT
ACCACACACAAAATACACCACACACAAAATACACCACACACCACACACAAAATACACTACACACAAAATACACCACACACAAAATACACC
ACATACCACACACAAAATACACCACACACAAAATACGCCACATACCACACACAAAATACACCACACGCCACACACAAAATACACCACACA
CCACACACAAAATATACCACACACACCACACACAAAATACACCACACACGAAATACACCACACACCACACACAAAATACACCACACACTA
CACTCAAAATACACCACACACAAAATACAGTACACACCACACACAAAATACACCACACACACCACACACAAAATACACCACACATGAAAT
ACACCACACACCACACACAATAAACCACACACAAAATACACCACACACACCACACGCAAAATACACCACACACACTGCACACAAAGTACA
CCACACACAAAATACACCACACACGCCACACACGAAATACACCACATGCCACACACACACCACACACCACACACAAAATACACCACACAC
AAAATAAACCACACACAAAATACACCACACACCACATGCAAAATACACCACACACACCGCACACAAAGTACACCACACACAAAATACACC
ACACACCCCACACACAAAATACACCACACGCCACACACACATCACACGCCACACACACACCACACACCACACACAAAATATACCACACAC
CACACACACAGTGCACTCCAAACACAAAATATACACATACCACACACCACATACAAAATACACCACACACCGCACACACACCACACACTA
TACACACACCACACACCACACACAAAATACACCACACACCACACACACACCACACACTATACACAAAATACGCACACCAGTGCACACCAA
ACACAAAATATACACATACCACACACCACATACAAAATACACCACACACCACACACAAAATACGCCACACACACACCACACACACAGTAC
CCACCGCACACAAAATATACACATACCACACACCACATACAAAATACACCACACACCACACAGTGCAACCAAACACAAAATATACACATA
CCACACACCACATACAAAATACACCACACACACCACACATCACGCACACACACCACACACCAAATATACACACACCACACCCCACACACG
AAATACACCATACACAAACCACACACAAAATACACCACACACCACACACATGCCTCACATCACACACAAAATATACACACGCCACATCCC
ACACACACCACACACAAAATGTGCCACACAGACACCACACACACAGTACACGCCACAGAAAATATACCATGCAAACATACCACACAGCAC
ACAACACACCACATACAAAATACACTACATATATACACACGCCACACAAAAAATACACCACACACACACGCAACATCATACACACACGCC
ACACACCACACAAAATAAACCTCATACACCACACACACACCACACACCACACACCATACACACACCACACACAAAATATACCACACACAC
CATATACACCTACCCACACATATAACACACCACACACCACACGCCACACACCACACACAAAATACACCACATACACCACACATACCAGAC
ACCACACACCACTCACAAAATATATCACACACACACACCACACACACACACCAATCACATGCACACACCCCCCACAAAACATCACATAC
ATCATACACACACCACACACCACTCACCACACAAAATACACAACACACAAAATACACCACATACACCACACACATTGGACACCACACACT
TACAAAATACATCACACACCACACACACACACACCACACACAGCAGACAGCAGACCACACACACACAAAATACACCACACAGATATCACA
CACACCACACACACCGTGCACCACACACAGTACACACCACACGCACAGTACACACCACACACATAGATACACATGGCACACAAAATATAC
TATGTAAATACACCACACACGTACACACCACACACAAAATATACTACACATTCACACCACACACAAAATACACCACACACACACACACAT
ACACACCACACAAAATACACACACATCACACACACTACATGCCATGCACCACACGACACAAAATACACCTCACACACACCACACCACACA
CGCATGCACACACACCACTCACAAAATACATCACACATGCACCACACACACAGCACAAATCACACACACATACACCACAAACACACACCA
CACAAAATGCACCACACACCCACACGCACCATATACACGCACCCTCCACACACACACAACACACCACACACATCACACACGCTACATGC
CACTCACCACACAAAATACAGCACACACGCATACACACGTATACAACACACAGCACACAACACGCTAAATACACCACATACCACAAGCAC
ACCATTCACAAAATACATCTCACACACACACACCACACACACACACACAAAATACACTGCACTGATATCACACACACACCACACAGAAAA
TACATCACACACATTCACATGTACACACACACACCACATGCAGAATACACCACACACACTAGATACCACACGCACACCACTCACAAAAAC
ATCACACACACAGCATACACTACACACACAGCACACACCACGCACACACACACAATACACCACACAGATGCCACACACAACAAAATACAC
CACACACTAGACACCACACACGACACACTGTGCACACATACCACATGCCAAACACCACACACACACAGAAACCCACAAACATACCACATA
CACAACATACATACACAGATAATATACCACCTAACCACACACAACACACACACATAAAATACACCACATACACATACAAAATATTCACAC
ACACACACATAAAATACACCACATACACATACAAAATATTCATACACCACACAGCACACACCACACACACCACACACCACATACACACAC
ACAACACATCACACACATGCCACATCCCACACACCACACAAAATACACCACACATACACCACATACAAAATACACCACATACACCATACA
AACTAGATACCACACACCACTCACAAAAAACATCACACTGCTCATACCACACAGCACACACCGCAGACACACAATACACCACACAGATAC
CACACACAACAAAATACACCACACACACTAGACATCACACACACACTATGCATTATGCACACACACCACACACACCACAAACATACACAA
CGTATACACAGACACAGAATATACCACACACAACACACACACATAACATACACCACATACGCATACCAAATACGCATACCACACACACAC
CACACATACCCTGTGAAGGCCACGCACACACAAAGCGCAGAAACACACCCCCACACAAACACATGCCAAGACACACACACACAAACATAC
ACCACATGCATGTTATACACACACAGCCAAACATCATAAACACACGTTACACCCATCACACCGCACACACACGATACCCAAACACCACA
CAGTCTGCAGCTCACCCCGTGGGCCCAGCCTGGCGTTCTG
```

P.S. AC motif = CA motif



```
>NC_000011.10:64594001-64597000 Homo sapiens chromosome 11, GRCh38.p13 Primary Assembly
CAGGGCGCCTCCCCTGCTCCTCTCAAGTGCTCCTTCTCCTACTTATTCTTCTCTTTTCCTCCCTCCTCCTCATCTCTTGGCCTATTCTCC
TCTCCTGTCCCTGCCTGCCTTGGCATCTATGCTGCCAGGCCCATCTTCTGAGCCCCTCCTCTGTGTGTGCCAGGCATGGCAGCACTGAGC
TAGGCTAGGGGAGCGAGTGGAAAGATACAACACATGCCCGCACACACAGCGCACAGCCGCAGTCTGCTTCATGGGCTGACAGAATCAATG
CTTGTGTCTGTCTGCTGGAAGACAGGGACCAGGATCCACTCATCAGAGGGGAGTCCACAGTCCCCACAGAGGGCCGGCATGTGTGGGAGG
TTAGGAAATGTCAGCACTGCCCTGAAACAAATCAGGCGTGGCCCTTGCCGAGCACCTGGCACATAGTAGGTGCTAAATAAATATTTGTTG
AATGGATGAATTGTTAGGTAAGTAGAAATAGATAGATAGGTAGGTAGGTAGGTAGGTAGGTAGGTAGGTAGGTAGATAGATAGAGACAGA
TAAATAAATGGGTTGACAGATGTGTGGATTGGATGAGTGGATGGGTGAGTTGGTGGATAGATGGATTGGATTGTATAGATGAATTGAATG
GATGGTTGAATGGATGGATAGATGGATGGATGGTTGGAATAGATGGATGGACGGATGGATGGATGGATGGATGGATGGATGGATGGATGG
ATGGACGGACAGACGGACGGACGGTTGGATGGATGGATGGATGGATGGATGAATGGATAGTTGGAATAAATGGATGGATGGATGAATGGA
TGGATGGATGGGTGGATGGACGGACGGATGGATGGACAGATGGAATGGATGGATGGATGGTTGGAATGGATGGATGGATGGAATGGATGG
ATGGATGGAATGGATGGATGGATTGTTGAATGGATGGTTGGATGGATGGATGGATGGATGGTTGGAATAGATGGATGGATGGATGGGTGG
ATGGATGGATGGATGGATGGATGGATGGATGGTTGGAATAGATGGATGGATGGATGGGTGGATGGATGGATGGATGGATGGATGGATAGA
TGGATAGATGGAAAGGATGGATGGATGGATGGATGGATGGATGGATGGATGGATGGTTGAATGGATGGTTGAATGGATGGTTGGATGGAT
GGATGGATGGATGGATGGACGGATGGATGGATGTTTGGAATAGATGGATGGATGGATGGATGGATGGATGGATGGATGGATGGAATGGAT
GGATGGATGGATGGATGGTTGAATGGATGGTTGGATGGATGGATGGATGGATGGATGGATGGATGGATGGATGGTTGAATGGATGGTTGA
ATGGATGGATGGATGGATGGATGGATGGATGGATGGATGGATGGACGGACGGATGGATGGATGGATGTTTGGAATAGATGGATGGATGGA
TGATGGATGGATGGAATGGATGGATGTTTGGAATAGATGGATGGATGGATGGAATGGATGGATGGATGGTTGGAATGAATGGATGGATGG
ATGGTTGGAATGGATGGATGGATGGATGGTTGGAATGGATGGATGGATGGAATGGAAGGAGGATGGATGGATGGATGGTTGAATGGATG
GATAGATGGATGGATGGTTGAATGGGTGGAATGAAAGGATGGATGGTTGTAATAGATGGATGGATGGATGGTTGGAATAGATGGATGGAT
GGATGGATGGTTGGAATAGACGGATGGATGGATGGTTGGAATAGATGGATGGATGGATGGTTGGAATGGATGAATGGATGGATGGATAGA
TGATTGGAATAGATGGATGGATGGACGGTTGAATGGATGGATGGATGGATGGATGGATAGTTGAATGGATGGAATGGATGGATGGATGGT
TGGAATAGATGAATGGATAGATGGTTGGAAAAGATGGATGGATGGATGGATGGATGGATGGATGGATGGATGGATGGAATGGATGGATGG
TTGGAATAGATGGAATGGATGGATGGATGGATGGATGGATGGAATGGATGGATGGTTGGAATAGATGGAATGGATAGATGGATGGATGGA
TGGATGGTTTGATGGTTGAATGGATGGATGGATGGAATGGATGGATGAATGGAATTGATGGATGGATGGGATGGATGGATGGATTGATGG
ATAGTTGGAATAGATGGAATGGATGGATGGATGGATGGTCGAATGGATAGTTGAATGGATGGATGCATGGATGGATGGATGGATGGAATG
GATGGATGGATGGATGGATGGATGGATGGATGGATGGATGGCTGAATGGATGGTTGAATGGATGGTTGAATGGATGGATGGATGGATGGG
TGGATGGATGGATGAATGGATGGATGGATGGATGGAATGGATGGATGGATGGTTGGATGGATGGTTGAATGGATGGATAGGTGGATTAAT
TGGATGAACTGGATGGATGAATGGATAGATGGATGGACGGAGGCCATTTGAGGGCAGGGCTGAGCCTGGATCATTCATACACAGGCCCTG
CACTTAGTAGTTGTTTGGTCATTGGGACCACTGTAGAAAGGAGCACAGCAGAGGCCCAGTGGTGAGCAGGTGCCCGCACCCACACTAAGA
AGAAGGCTGAGTCCACAAATCTCTCCCACCCAACAGCTTGGAGTGGAGTAACAATTTGCAGCCACTCCTCAGAGCAGGAGAAAATTACAT
CTCATACAGTGGCTCCACAAGGAGGCAATCGTGTCCGACAAGAGCCTTCTGGGATGGCAGCTTGACTCCCAACACCAAGAGGGTAGAGAG
TGGAAGGTAAAGGTCAGTGATAAAAATTCCAAGGCACCCCTCCTGGTTCTCATGCTGATGCAGGGCCACAGAGCTGCACCCCCCTTTCCT
TGGTAGACGTGCTCCCACAGGGCTGAAGAAGTGCCCATTCATTCATTCATTACTTCATTCATTCAACAAATAACTACTGAGCACCTACCA
TGTGCCAGGCGCTGGCTGACACCAGCTATG
```

P.S. ATGG motif = GGAT motif

A

Chr No.	01	02	03	04	05	06	07	08
Chr size	249.7/248.4	242.5/242.7	198.5/201.1	190.4/193.6	181.6/182	170.8/172.1	159.3/160.6	145.1/146.3
Seq size	230.5/248.4	240.5/242.7	198.1/201.1	189.8/193.6	181.3/182	170.1/172.1	159.0/160.6	144.8/146.3
LS map num	4602/4965	4814/4855	3961/4021	3799/3873	3624/3639	3404/3447	3182/3212	2898/2924
W STRs RD	51.4/55.3	52.7/51.6	51.0/53.3	52.0/52.7	51.3/52.3	51.4/51.6	52.6/53.7	50.8/50.9
HDMA num	199/285	292/348	180/228	227/233	196/234	158/199	196/235	156/187
HDMA dis	1.2/0.9	0.8/0.7	1.1/0.9	0.8/0.8	0.9/0.8	1.0/0.9	0.8/0.7	0.9/0.8

Chr No.	09	10	11	12	13	14	15	16
Chr size	138.7/150.6	133.8/134.8	135.2/135.1	133.3/133.3	114.4/113.6	108.1/101.2	102.4/99.8	92.2/96.3
Seq size	121.8/150.6	133.3/134.8	134.5/135.1	133.1/133.3	98.0/113.6	90.6/101.2	84.6/99.8	81.8/96.3
LS map num	2440/3013	2665/2694	2692/2701	2662/2666	1960/2270	1810/2024	1696/1995	1638/1926
W STRs RD	50.9/126.5	51.8/53.3	49.8/50.1	52.3/52.5	51.6/66.6	51.0/55.7	50.8/60.4	53.1/56.0
HDMA num	121/184	160/195	142/191	154/189	100/473	83/226	66/350	105/129
HDMA dis	1.0/0.8	0.8/0.7	0.9/0.7	0.9/0.7	1.0/0.2	1.0/0.4	1.3/0.3	0.8/0.7

Chr No.	17	18	19	20	21	22	X	Y
Chr size	83.8/84.3	80.4/80.5	58.6/61.7	64.4/66.2	46.7/45.1	51.9/51.3	156/154.3	57.3/-
Seq size	82.9/84.3	80.1/80.5	58.4/61.7	63.9/66.2	40.1/45.1	39.2/51.3	154.9/154.3	26.4/-
LS map num	1658/1684	1600/1610	1170/1234	1270/1324	814/900	783/1027	3102/3086	525/-
W STRs RD	52.4/54.0	49.7/50.0	57.0/56.8	50.4/54.7	51.9/69.7	51.6/64.9	51.7/52.5	57.9/-
HDMA num	89/112	91/125	105/127	74/108	83/341	87/300	271/307	100/-
HDMA dis	0.9/0.8	0.9/0.6	0.6/0.5	0.9/0.6	0.5/0.1	0.5/0.2	0.6/0.5	0.3/-

B

Peak name	E_position	Original motif type	Percentage (%)	Integrated motif type	pD$_1$RD
G38-02-sHP49	107707000	AT/A/AC/AG/TG/TC	89.7/3.0/3.0/1.7/1.3/1.3	[AT]^{h}	465
G38-02-sHP80	179398000	AT/AC/ACGT/TG	46.6/43.2/9.0/1.1	[AT]^{m}	532
G38-02-HP135	191700000	AC/AT/A/TG/ATAC/AG	51.8/28.8/9.5/3.9/3.9/2.0	[AC]^{m}	305
G38-02-HP147	213408000	AG/TG/C	94.9/3.0/2.1	[AG]^{h}	331
G38-02-HP182	241862000	AC/C	98.2/1.8	[AC]^{h}	340
G38-02-HP177	239468000	TG/G/A/AT	92.1/3.8/2.2/1.9	[TG]^{h}	317
G38-02-HP156	224889000	TG/AT/CAT	60.5/37.2/2.3	[TG]^{m}	387
C13-01-HP10	3190000	TC/TCTCTG/AG	92.8/5.4/1.8	[TC]^{h}	332
C13-01-HP11	3238000	ATGG/TGGG/G/TG	62.8/26.3/7.8/3.0	[ATGG]^{m}	395
C13-01-HP38	18746000	ATGG/TTGG/TC/AC/TG/AG	91.0/3.0/1.5/1.5/1.5/1.5	[ATGG]^{h}	400
C13-22-HP186	47169000	ATCC/ACCC/TCCG/AC	64.3/26.9/7.0/1.8	[ATCC]^{m}	342
C13-22-sHP97	48172000	ATCC/AT/AC/AG	72.2/13.9/9.7/4.2	[ATCC]^{h}	432
C13-03-HP14	12447000	AAGG/AAG/AG/AGGG/A	72.2/20.3/3.0/3.0/1.5	[AG$_{mix}$]	399
C13-03-HP121	168581000	TTCC/TCCC/C/TC/T	49.6/34.9/10.4/3.4/1.7	[TC$_{mix}$]	355
G38-12-sHP16$_{m4}$	39712000	AT/TG/AC/T/ATGT/ATAC/CCT/C	83.0/6.7/3.0/3.0/1.9/1.3/0.7/0.4	[AT]^{h}	458

C

Chr No.	01	02	03	04	05	06	07	08	09	10	11	12
[AT]$^{h+m}$	**51.3/46.3**	**51.7/48.3**	**63.9**/4.0	**64.8/63.1**	**63.8/61.1**	**66.5/59.8**	**52.6/48.9**	**48.1/47.1**	**45.5/39.7**	**44.4/44.1**	**54.2/46.1**	**55.3/56.1**
[AC]$^{h+m}$	**10.6/11.2**	**14.4/11.8**	3.9/4.0	9.3/9.4	9.2/9.0	7.0/8.0	**12.8/11.5**	**11.5/11.2**	**12.4/11.4**	**11.9/12.8**	7.7/7.3	9.7/7.9
[TC]$^{h+m}$	2.0/2.1	0.3/0.9	2.2/1.3	0.4/0.4	0.5/2.1	1.9/0.5	1.0/1.7	1.9/0.5	0/0	3.8/2.6	2.8/3.1	0/0.5
[TG]$^{h+m}$	**10.6**/9.8	**12.7/13.8**	5.6/6.1	6.6/6.4	8.2/9.0	6.3/8.0	**10.7**/9.4	**14.7/11.2**	**25.6/22.8**	**13.1/10.3**	**11.3/10.5**	8.4/9.0
[AG]$^{h+m}$	1.0/1.1	2.7/2.0	0/0.4	0.9/0.4	0.5/0.4	0/0	0.5/0	1.3/1.6	0.8/3.3	2.5/1.5	2.1/1.6	2.6/2.1
[TC$_{mix}$]	8.5/7.4	6.5/7.2	7.8/7.9	3.1/4.3	5.1/5.6	7.9/8.5	5.6/8.9	4.5/4.3	3.3/4.4	9.4/8.2	2.1/4.2	7.1/6.4
[AG$_{mix}$]	4.0/9.8	4.1/6.9	**14.4**/8.8	9.3/8.6	8.2/6.4	6.3/8.0	6.7/6.8	**12.2/13.9**	5.0/6.0	5.6/8.2	7.8/**11.0**	4.6/6.9

Chr No.	13	14	15	16	17	18	19	20	21	22	X	Y
[AT]$^{h+m}$	**64.0/19.2**	**49.4/29.2**	**54.5/36.6**	**39.1/37.2**	**43.8/40.2**	**56.0/49.6**	**35.2/33.9**	**36.5/29.6**	**22.9/14.7**	**35.6/40.7**	**70.9/68.1**	**35.0/-**
[AC]$^{h+m}$	**11.0**/3.4	**12.0**/7.1	**13.6**/2.6	**15.2/12.4**	**12.4**/9.8	**11.0**/8.8	**13.3/10.2**	6.8/**12.0**	**24.1**/7.3	**17.2**/4.3	3.3/3.9	5.0/-
[TC]$^{h+m}$	2.0/**20.0**	1.2/7.5	1.5/8.6	2.9/3.1	2.3/1.8	0/0.8	5.7/4.7	2.7/2.8	4.8/**10.9**	2.3/5.0	2.2/2.0	7.0/-
[TG]$^{h+m}$	**12.0**/3.4	9.6/5.8	**10.6**/3.4	**14.3/11.6**	**18.0/15.2**	7.7/9.6	**14.3/10.2**	**17.6/13.0**	**16.9**/3.8	**16.1**/5.7	3.0/2.6	3.0/-
[AG]$^{h+m}$	0/0.2	2.4/2.2	0/0.6	0/1.6	0/0.9	1.1/0.8	3.8/4.7	4.1/5.6	0/0.3	5.8/2.7	1.9/3.6	4.0/-
[TC$_{mix}$]	5.0/**43.6**	6.0/**15.5**	9.1/**36.6**	**11.4/10.1**	4.5/5.4	5.5/8.8	6.7/9.5	6.8/**13.0**	**13.3/44.6**	2.3/**13.3**	6.3/6.8	**11.0/-**
[AG$_{mix}$]	4.0/1.7	7.2/4.4	4.6/2.0	2.9/7.8	6.7/5.4	5.5/4.8	6.7/7.9	6.8/6.5	3.6/1.5	2.3/1.7	6.6/7.5	**17.0/-**

D

Chr No.	01	02	03	04	05	06	07	08	09	10	11	12
[CCA]$^{h+m}$	1.0/1.8	0.3/1.4	0.6/0.4	1.8/1.3	0.5/0.9	0.6/1.0	0/0.4	1.9/2.1	1.7/2.2	0.6/1.5	0.7/0.5	0/0
[TGG]$^{h+m}$	3.0/1.8	0.7/0.9	0/0.9	0.4/0.9	0/1.7	0.6/1.0	1.5/1.7	0.6/2.1	0/1.1	0/0.5	1.4/3.7	0.7/1.1
[ATG]$^{h+m}$	0.5/0.7	0.3/0.3	0/0	0.4/0	0/0	0.6/0.5	0/0	0/0.5	0/0	0/1.0	0.7/0.5	1.3/0
[CAT]$^{h+m}$	0.5/0.4	1.0/0.3	0/0	0.4/0.4	0/0	0/0	0/0	0/1.1	0/0	0/0	0.7/1.0	1.3/0.5
[ATCC]$^{h+m}$	1.5/1.1	0.7/0.3	0.6/0.4	0.4/0.4	0.5/0.4	0.6/0	3.1/2.6	0.6/1.1	1.7/2.7	1.3/1.0	4.2/3.7	0.7/2.1
[ATGG]$^{h+m}$	1.5/1.4	0.3/0.9	0/0.4	0.9/0.9	1.5/0.9	0.6/0.5	1.5/1.3	0.6/0.5	1.7/2.2	0/0.5	2.1/3.7	0.7/0
[ATAG]$^{h+m}$	1/0.4	0/0	0/0	0/0.4	0.5/0.4	0/0	1.5/1.3	0/0.5	0/0	0.6/0.5	0/0.5	0/0
[ATCT]$^{h+m}$	0/0	0/0	0/0	0/0	0/0.4	0.6/0.5	1.0/0.9	0/0	0/0	0.6/0.5	0.7/0.5	0/0
[AATGG]$^{h+m}$	1.0/2.5	0.3/0.9	0/2.2	0.4/**0**	0/**0**	0/**0**	0/0.9	0/**0**	0/2.2	3.8/3.6	0/**0**	0.7/0.5
[ATTCC]$^{h+m}$	0/0.7	0.7/0.3	0.6/0.4	0.4/0.9	0/0.4	0/**0**	0/1.7	0/**0**	0/**0**	0.6/0.5	0/**0**	0/**0**
[AACCCT]$^{h+m}$	1.0/0.4	0/0.3	0.6/0.4	0/0.4	0.5/0.4	0/0.5	0/0.4	0/0.5	0.8/0.5	0.6/0.5	0/0.5	0.7/0.5
[TTAGGG]$^{h+m}$	0.5/0.4	0/0.3	0/0.4	0.4/0.4	0/0.4	0/0.5	0/0.4	0/0.5	0/0.5	0/0.5	0/0.5	0.7/0.5

Chr No.	13	14	15	16	17	18	19	20	21	22	X	Y
[CCA]$^{h+m}$	0/0.6	2.4/0.9	3/0.9	1.0/0	1.1/0.9	2.2/2.4	1.0/1.6	1.4/2.8	3.6/1.2	1.2/0.7	0/0	0/-
[TGG]$^{h+m}$	0/0	1.2/0.9	1.5/0.6	0/0.8	4.5/4.5	0/3.2	0/0.8	0/0	1.2/0.6	2.3/1.3	0/0	0/-
[ATG]$^{h+m}$	0/0	0/0	0/0	1.0/1.6	0/0	0/0	0/0.8	1.4/0.9	0/0	0/0	0/0	0/-
[CAT]$^{h+m}$	0/0	0/0	0/0	0/0	1.1/0.9	0/0	0/0	0/0	0/0	0/0.3	0.4/0.7	1.0/-
[ATCC]$^{h+m}$	1.0/0.2	2.4/0.9	0/0	2.9/3.1	1.1/0.9	3.3/3.2	6.7/6.3	0/1.9	1.2/0.9	3.5/2.0	0.7/0.7	2.0/-
[ATGG]$^{h+m}$	1.0/0.2	3.6/1.3	0/0.3	4.8/3.9	2.2/1.8	1.1/0.8	2.9/3.2	4.1/2.8	1.2/0.3	4.6/2	1.1/0.7	3.0/-
[ATAG]$^{h+m}$	0/0	0/0	0/0	0/0	0/0	1.1/0.8	0/0	0/0	0/0.3	1.2/0.3	0.4/0.7	1.0/-
[ATCT]$^{h+m}$	0/0.2	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0.6	0/0	1.1/0	2.0/-
[AATGG]$^{h+m}$	0/2.5	0/12.4	0/2.3	1.9/0.8	1.1/6.3	0/**0**	0/**0**	8.1/4.6	1.2/7	1.2/13.3	0/**0**	1.0/-
[ATTCC]$^{h+m}$	0/3.6	0/10.2	0/4	1/0.8	0/0.9	1.1/**0**	0/**0**	0/**0**	3.7/5.0	1.2/5.0	0/**0**	3.0/-
[AACCCT]$^{h+m}$	0/**0**	0/0.4	0/0.3	0/0.8	0/0.9	1.1/3.2	0/0.8	0/0.9	0/0.3	0/0.3	0/0.3	0/-
[TTAGGG]$^{h+m}$	0/0.2	0/0.4	0/0.3	0/0.8	0/0.9	0/0.8	0/0.8	1.4/1.9	1.2/0.3	1.2/0.3	0.4/0.3	1.0/-

E*

i

[AAAAT]$^{h+m}$	0.5 (G38-05)	0.6 (G38-08)	0.5 (C13-08)	0.5 (C13-10)	0.4 (G38-X)	0.3 (C13-X)	1.0 (G38-Y)
[G]$^{h+m}$	0.4 (C13-07)	0.7 (G38-12)	1.1 (C13-12)	1.0 (G38-16)	0.8 (C13-16)	0.3 (C13-22)	
[ATGT]$^{h+m}$	0.4 (C13-07)	1.4 (G38-20)	1.0 (C13-20)	0.4 (G38-X)	0.3 (C13-X)	1.0 (G38-Y)	
[ACTC]$^{h+m}$	0.5 (G38-07)	0.4 (C13-07)	0.6 (G38-08)	0.5 (C13-08)	0.8 (G38-09)	0.5 (C13-09)	
[AGTG]^{h}	0.3 (G38-02)	0.3 (C13-02)	0.7 (G38-12)	0.5 (C13-12)	2.0 (G38-19)	0.8 (C13-19)	
[AGCC]$^{h+m}$	0.3 (G38-02)	0.6 (C13-02)	1.5 (G38-15)	0.3 (C13-15)	0.8 (C13-16)	1.2 (G38-22)	
[C]$^{h+m}$	0.6 (G38-10)	0.5 (C13-11)	0.2 (C13-13)	0.4 (G38-X)	1.0 (G38-Y)		
[AGG]$^{h+m}$	0.6 (G38-06)	0.5 (C13-06)	0.8 (C13-19)	1.1 (G38-17)	0.3 (C13-22)		
[CTA]$^{h+m}$	0.3 (C13-02)	0.5 (G38-05)	0.4 (C13-05)	1.0 (G38-16)	0.8 (C13-16)		
[AATG]$^{h+m}$	0.4 (C13-04)	0.7 (G38-12)	0.5 (C13-12)	0.4 (G38-X)	1.0 (G38-Y)		
[AAAG]^{m}	0.5 (C13-06)	2.2 (G38-18)	0.8 (C13-18)	0.3 (C13-X)			
[ACCC]^{m}	0.5 (C13-11)	0.8 (C13-16)	1.0 (G38-19)	0.8 (C13-19)			
[TGGC]$^{h+m}$	0.3 (G38-02)	0.6 (C13-02)	0.3 (C13-15)	0.8 (C13-16)			
[ATT]$^{h+m}$	0.6 (G38-08)	0.7 (G38-12)	0.3 (C13-X)				
[CCG]$^{h+m}$	0.3 (C13-12)	0.5 (C13-12)	0.9 (C13-17)				
[ATCCT]$^{h+m}$	0.7 (G38-02)	0.3 (C13-02)	0.3 (C13-15)				
[AGTGG]$^{h+m}$	0.7 (G38-12)	0.5 (C13-12)	1.0 (C13-20)				
[AGCCC]^{h}	2.4 (G38-14)	0.9 (C13-14)	0.3 (C13-15)				
[ACTCC]^{m}	0.4 (C13-04)	0.4 (C13-13)	0.3 (C13-21)				

iii

[CG]^{m}	0.5 (C13-12)
[AAG]^{m}	0.5 (C13-06)
[AGT]^{m}	0.4 (C13-03)
[AGGT]^{h}	0.3 (C13-21)
[AGGG]^{m}	0.9 (C13-17)
[TTTC]^{m}	0.3 (C13-X)
[TCTG]^{h}	0.8 (C13-18)
[AATAT]^{h}	0.2 (C13-13)
[ATAGT]^{h}	0.9 (C13-17)
[ATATT]^{m}	0.3 (C13-X)
[ACACC]^{h}	0.9 (C13-17)
[ATAGGT]^{h}	0.2 (C13-13)
[ATCTCC]^{h}	0.4 (C13-04)
[ATGTGG]^{h}	0.3 (C13-02)
[AGGGTG]^{h}	0.5 (C13-10)

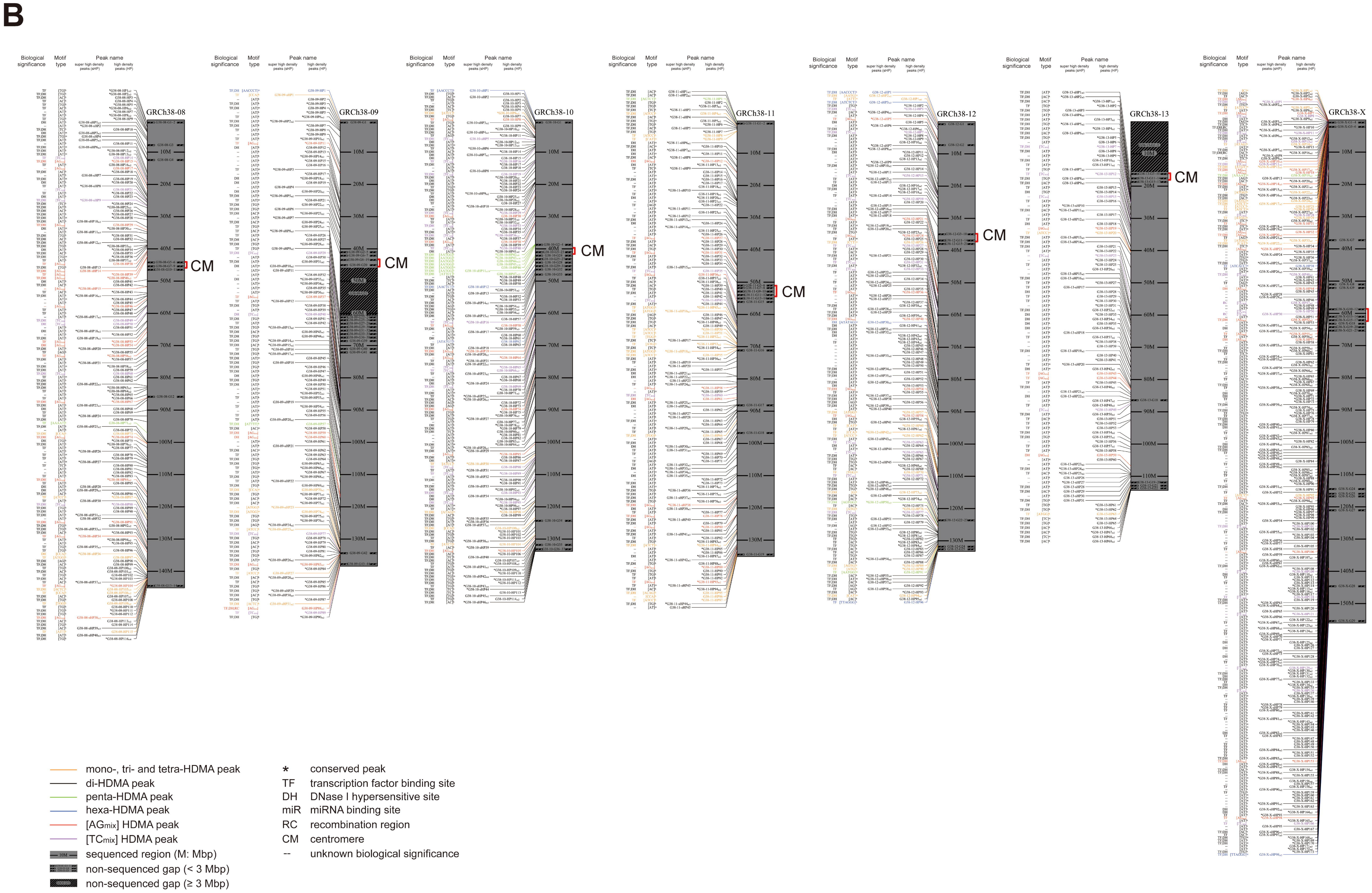
ii

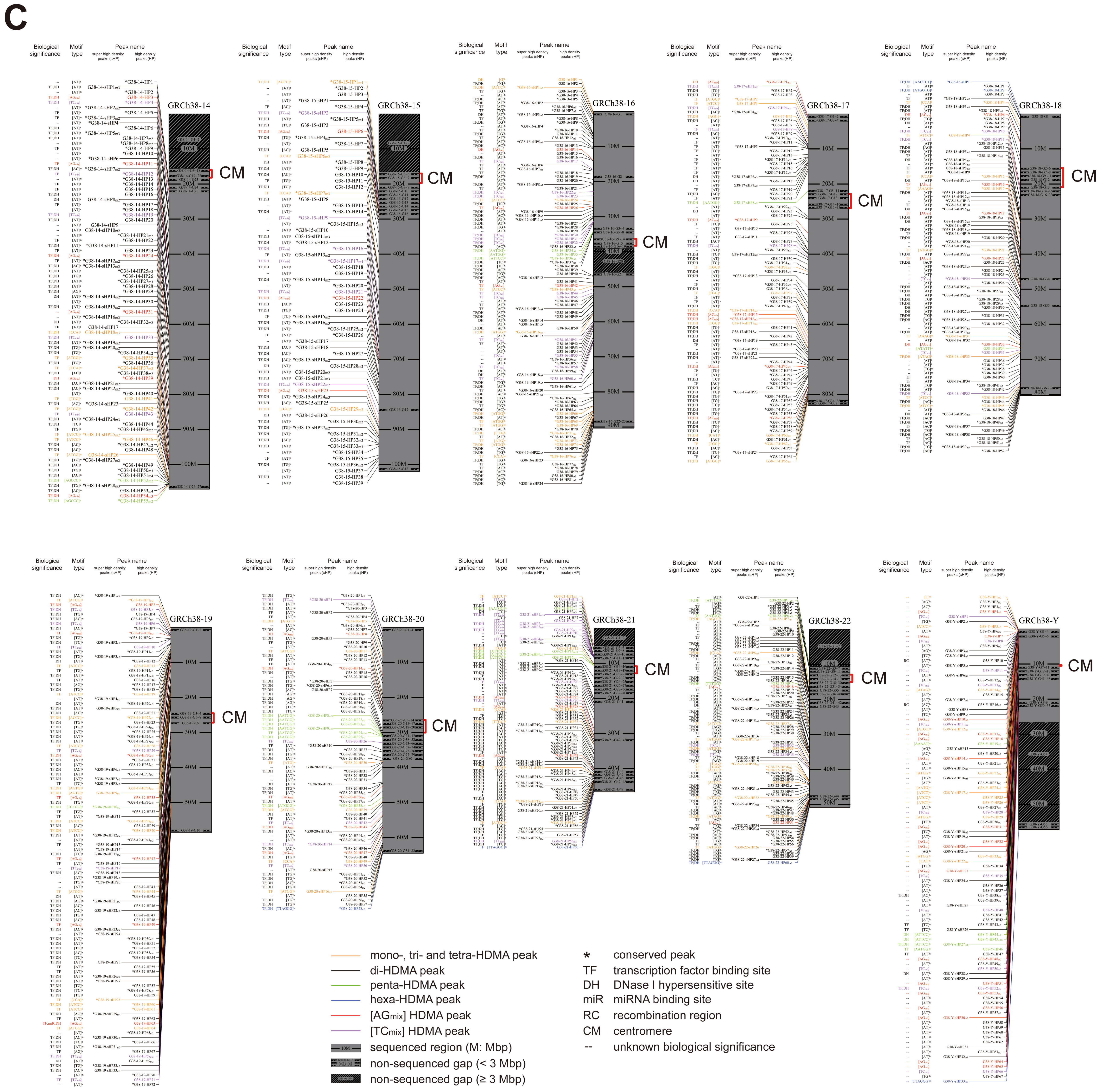
[T]^{h}	0.3 (G38-02)	0.5 (C13-06)
[TTG]$^{h+m}$	0.5 (C13-12)	0.2 (C13-13)
[CAG]$^{h+m}$	0.4 (C13-03)	0.5 (C13-06)
[CGG]$^{h+m}$	0.4 (C13-01)	0.3 (G38-02)
[ACCT]^{m}	0.5 (G38-01)	0.4 (C13-01)
[TTCC]^{m}	0.4 (C13-03)	0.5 (C13-10)
[ATTTT]^{h}	0.4 (C13-03)	0.8 (G38-09)
[ATGGG]$^{h+m}$	0.8 (C13-19)	1.4 (G38-20)
[TTTGG]^{h}	1.2 (G38-22)	0.3 (C13-22)
[TCTGG]^{h}	1.0 (G38-19)	0.8 (C13-19)
[TGCCC]$^{h+m}$	0.3 (G38-02)	0.3 (C13-02)
[ATATAG]^{m}	0.7 (G38-12)	0.5 (C13-12)
[ATCACC]$^{h+m}$	0.5 (C13-09)	0.4 (G38-X)
[ATCTCT]^{h}	0.7 (G38-12)	0.5 (C13-12)
[ATGGTG]^{h}	1.1 (G38-18)	0.8 (C13-18)

iv

[CTT]^{m}	0.7 (G38-12)
[ACAG]^{h}	0.7 (G38-11)
[TTGG]^{h}	1.2 (G38-21)
[AGTCT]^{h}	0.7 (G38-11)
[AACTCT]^{h}	0.6 (G38-10)
[ATATGT]^{h}	0.5 (G38-07)
[TCTGTG]^{m}	0.3 (G38-02)

A
Biological significance
Motif type
Peak name
super high density peaks (sHP)
high density peaks (HP)
GRCh38-01
GRCh38-02
GRCh38-03
GRCh38-04
GRCh38-05
GRCh38-06
GRCh38-07
CM
mono-, tri- and tetra-HDMA peak
di-HDMA peak
penta-HDMA peak
hexa-HDMA peak
[AGmix] HDMA peak
[TCmix] HDMA peak
sequenced region (M: Mbp)
non-sequenced gap (< 3 Mbp)
non-sequenced gap (≥ 3 Mbp)
* conserved peak
TF transcription factor binding site
DH DNase I hypersensitive site
miR miRNA binding site
RC recombination region
CM centromere
-- unknown biological significance


B
Biological significance
Motif type
Peak name
super high density peaks (sHP)
high density peaks (HP)
GRCh38-08
GRCh38-09
GRCh38-10
GRCh38-11
GRCh38-12
GRCh38-13
GRCh38-X
CM
10M
20M
30M
40M
50M
60M
70M
80M
90M
100M
110M
120M
130M
140M
150M
mono-, tri- and tetra-HDMA peak
di-HDMA peak
penta-HDMA peak
hexa-HDMA peak
[AGmix] HDMA peak
[TCmix] HDMA peak
sequenced region (M: Mbp)
non-sequenced gap (< 3 Mbp)
non-sequenced gap (≥ 3 Mbp)
* conserved peak
TF transcription factor binding site
DH DNase I hypersensitive site
miR miRNA binding site
RC recombination region
CM centromere
-- unknown biological significance

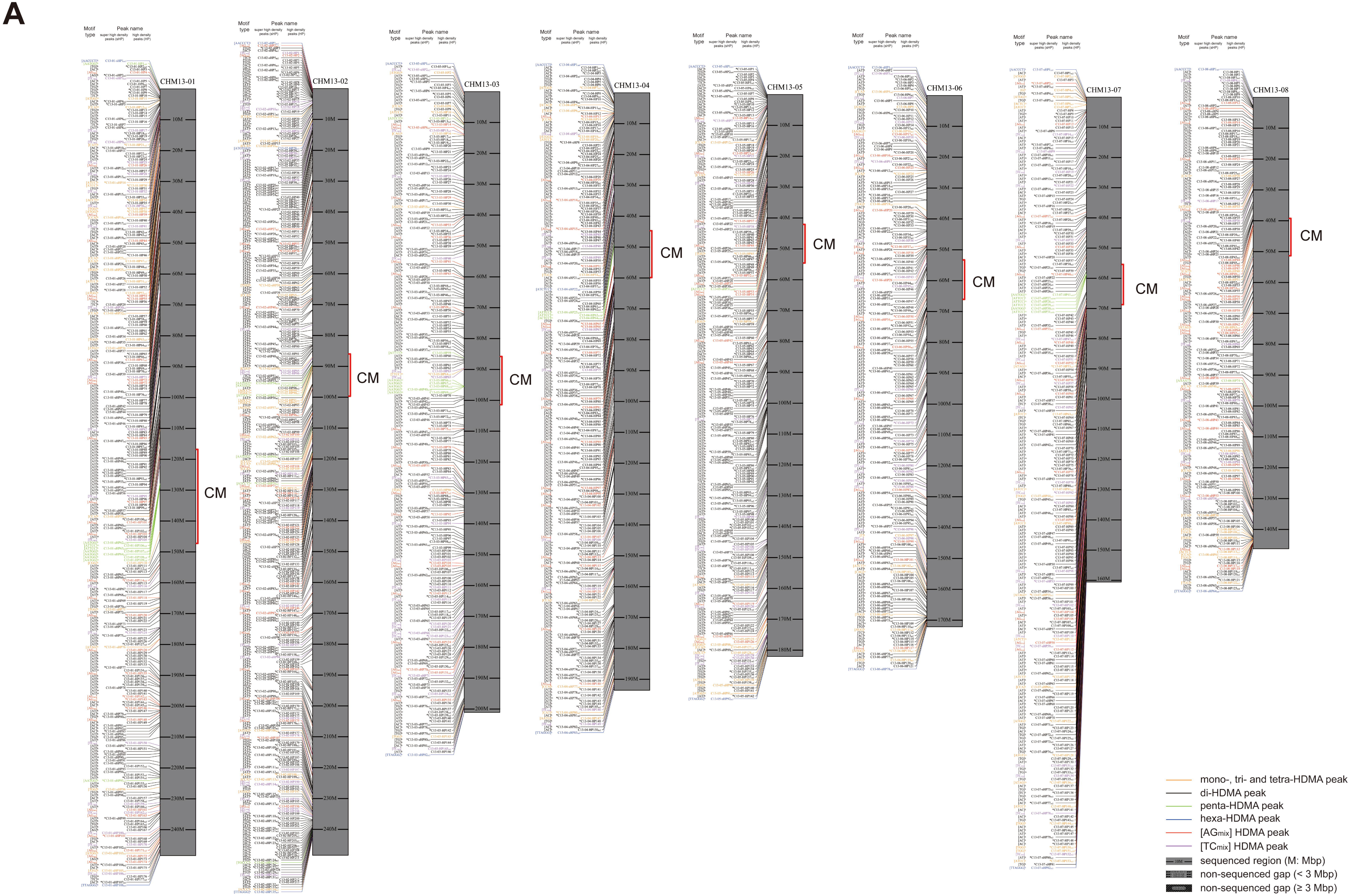

C
Biological significance
Motif type
Peak name
super high density peaks (sHP)
high density peaks (HP)
GRCh38-14
GRCh38-15
GRCh38-16
GRCh38-17
GRCh38-18
GRCh38-19
GRCh38-20
GRCh38-21
GRCh38-22
GRCh38-Y
CM
mono-, tri- and tetra-HDMA peak
di-HDMA peak
penta-HDMA peak
hexa-HDMA peak
[AGmix] HDMA peak
[TCmix] HDMA peak
10M
sequenced region (M: Mbp)
non-sequenced gap (< 3 Mbp)
non-sequenced gap (≥ 3 Mbp)
* conserved peak
TF transcription factor binding site
DH DNase I hypersensitive site
miR miRNA binding site
RC recombination region
CM centromere
-- unknown biological significance

D

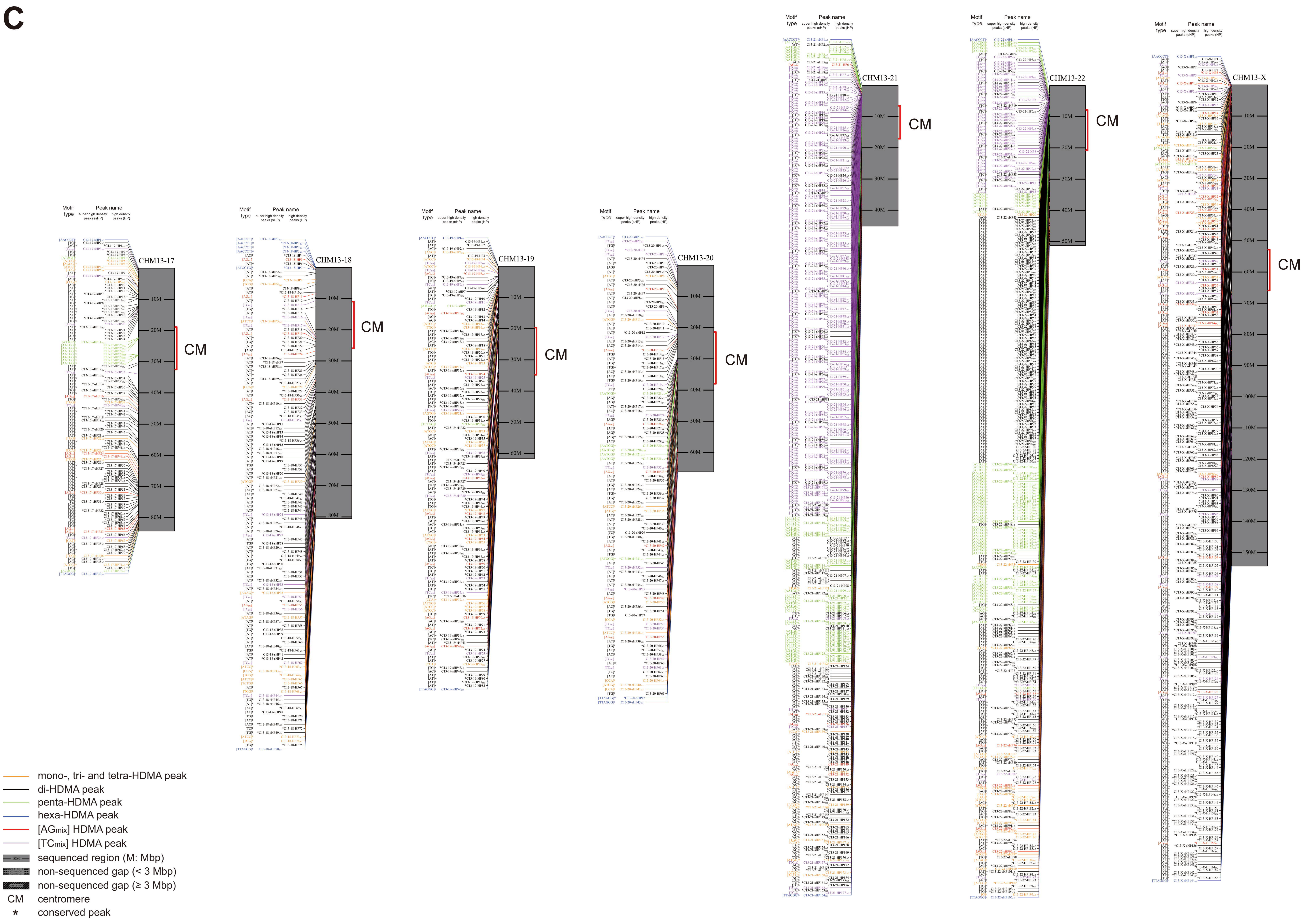
HDMA peaks possibly involve biological significance in the human reference (GRCh38) genome.

Chr No.	Peak num	Biological significance [a]			
		TF [b]	DH [c]	miR [d]	RC [e]
chr1	199	149	118	0	0
chr2	292	211	176	0	0
chr3	180	105	85	0	0
chr4	227	115	110	0	0
chr5	196	128	96	0	0
chr6	158	101	75	0	0
chr7	196	153	117	1	0
chr8	156	118	87	0	0
chr9	121	88	70	0	1
chr10	160	123	103	0	0
chr11	142	92	75	0	0
chr12	154	107	76	0	0
chr13	100	65	47	0	0
chr14	83	63	46	0	0
chr15	66	49	44	0	0
chr16	105	85	58	0	0
chr17	89	72	52	0	0
chr18	91	56	45	0	0
chr19	105	90	68	1	0
chr20	74	61	48	0	0
chr21	83	66	50	0	0
chr22	87	66	46	0	0
chrX	271	131	88	0	4
chrY	100	7	6	0	2

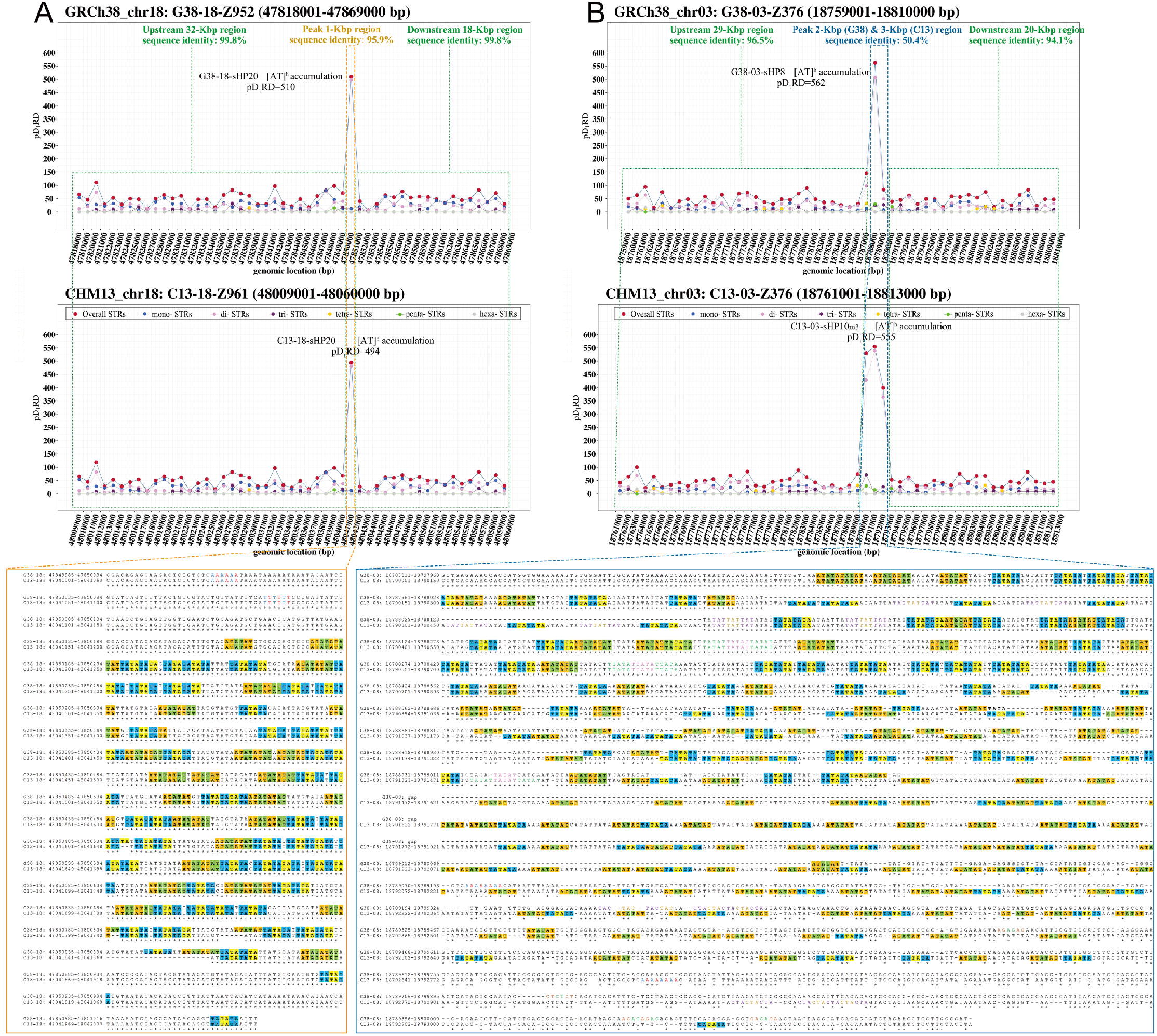
a The number of peaks possibly involving biological significance. b Transcription factor binding sites. c DNase I hypersensitive sites. d miRNA binding sites. e Recombination region.


A
Motif type
Peak name
super high density peaks (sHP)
high density peaks (HP)
CHM13-01
CHM13-02
CHM13-03
CHM13-04
CHM13-05
CHM13-06
CHM13-07
CHM13-08
10M
20M
30M
40M
50M
60M
70M
80M
90M
100M
110M
120M
130M
140M
150M
160M
170M
180M
190M
200M
210M
220M
230M
240M
CM
mono-, tri- and tetra-HDMA peak
di-HDMA peak
penta-HDMA peak
hexa-HDMA peak
[AGmix] HDMA peak
[TCmix] HDMA peak
sequenced region (M: Mbp)
non-sequenced gap (< 3 Mbp)
non-sequenced gap (≥ 3 Mbp)
CM centromere
* conserved peak

B
Motif type
Peak name
super high density peaks (sHP)
high density peaks (HP)
CHM13-09
CHM13-10
CHM13-11
CHM13-12
CHM13-16
CHM13-13
CHM13-14
CHM13-15
CM
10M
20M
30M
40M
50M
60M
70M
80M
90M
100M
110M
120M
130M
140M
150M
mono-, tri- and tetra-HDMA peak
di-HDMA peak
penta-HDMA peak
hexa-HDMA peak
[AGmix] HDMA peak
[TCmix] HDMA peak
sequenced region (M: Mbp)
non-sequenced gap (< 3 Mbp)
non-sequenced gap (≥ 3 Mbp)
CM centromere
* conserved peak

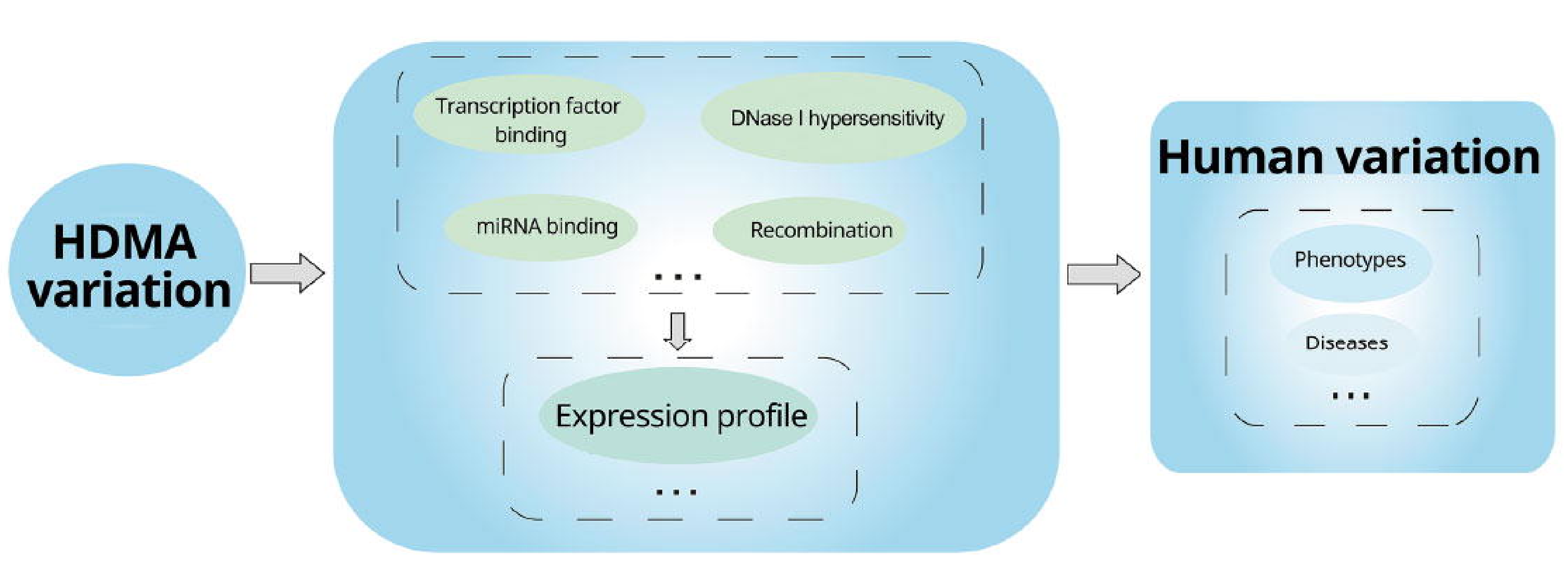

C
Motif type
Peak name
super high density peaks (sHP)
high density peaks (HP)
CHM13-17
CHM13-18
CHM13-19
CHM13-20
CHM13-21
CHM13-22
CHM13-X
CM
10M
20M
30M
40M
50M
60M
70M
80M
90M
100M
110M
120M
130M
140M
150M
mono-, tri- and tetra-HDMA peak
di-HDMA peak
penta-HDMA peak
hexa-HDMA peak
[AGmix] HDMA peak
[TCmix] HDMA peak
sequenced region (M: Mbp)
non-sequenced gap (< 3 Mbp)
non-sequenced gap (≥ 3 Mbp)
CM centromere
* conserved peak

Sample
G38
C13
distance (Mbp)
5.0
2.5
0
chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
chr16
chr17
chr18
chr19
chr20
chrX
*
4.0
2.0
chr13
chr14
chr15
chr21
chr22
chrY
full genome



C

Chr No.	01	02	03	04	05	06	07	08	09	10	11	12
HDMA num	199/285	292/348	180/228	227/233	196/234	158/199	196/235	156/187	121/184	160/195	142/191	154/189
rel_cons	95/98	145/145	102/103	131/131	78/80	48/48	87/87	63/63	66/66	75/75	86/86	77/77
exp/cont	55/58	86/90	37/37	43/43	77/80	72/72	57/58	46/46	31/31	45/45	25/25	39/39
gap_seq	0/8	0/1	0/6	0/2	0/2	0/2	0/6	0/2	0/5	0/2	0/2	0/2
HP>MP	49/121	61/112	41/82	53/57	41/72	38/77	52/84	47/76	24/82	40/73	31/78	38/71
Chr No.	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**X**	**Total**
HDMA num	100/473	83/226	66/350	105/129	89/112	91/125	105/127	74/108	83/341	87/300	271/307	3435/5306
rel_cons	34/34	55/55	38/38	60/60	47/47	62/62	55/55	40/40	42/44	44/44	126/129	1656/1667
exp/cont	46/46	12/12	18/18	11/11	20/20	16/16	21/21	15/15	11/11	25/29	87/87	895/910
gap_seq	0/352	0/129	0/271	0/1	0/9	0/3	0/1	0/7	0/228	0/196	0/1	0/1238
HP>MP	20/41	16/30	10/23	34/57	22/36	13/44	29/50	19/46	30/58	18/31	58/90	784/1491


D
HDMA variation
Transcription factor binding
DNase I hypersensitivity
miRNA binding
Recombination
...
Expression profile
...
Human variation
Phenotypes
Diseases
...