BMC Genomics

RESEARCH ARTICLE Open Access

A comprehensive microsatellite landscape of human Y-DNA at kilobase resolution

Douyue Li[1†], Saichao Pan[1†], Hongxi Zhang[1†], Yongzhuo Fu[1], Zhuli Peng[1], Liang Zhang[1], Shan Peng[1], Fei Xu[2], Hanrou Huang[1], Ruixue Shi[1], Heping Zheng[1], Yousong Peng[1] and Zhongyang Tan[1*]

Abstract

Background: Though interest in human simple sequence repeats (SSRs) is increasing, little is known about the exact distributional features of numerous SSRs in human Y-DNA at chromosomal level. Herein, totally 540 maps were established, which could clearly display SSR landscape in every bin of 1 k base pairs (Kbp) along the sequenced part of human reference Y-DNA (NC_000024.10), by our developed differential method for improving the existing method to reveal SSR distributional characteristics in large genomic sequences.

Results: The maps show that SSRs accumulate significantly with forming density peaks in at least 2040 bins of 1 Kbp, which involve different coding, noncoding and intergenic regions of the Y-DNA, and 10 especially high density peaks were reported to associate with biological significances, suggesting that the other hundreds of especially high density peaks might also be biologically significant and worth further analyzing. In contrast, the maps also show that SSRs are extremely sparse in at least 207 bins of 1 Kbp, including many noncoding and intergenic regions of the Y-DNA, which is inconsistent with the widely accepted view that SSRs are mostly rich in these regions, and these sparse distributions are possibly due to powerfully regional selection. Additionally, many regions harbor SSR clusters with same or similar motif in the Y-DNA.

Conclusions: These 540 maps may provide the important information of clearly position-related SSR distributional features along the human reference Y-DNA for better understanding the genome structures of the Y-DNA. This study may contribute to further exploring the biological significance and distribution law of the huge numbers of SSRs in human Y-DNA.

Keywords: Simple sequence repeat, Human Y-DNA, SSR landscape, 1 Kbp differential unit, SSR density peak, Extremely low SSR density region

Background

Simple sequence repeats (SSRs/microsatellites) are ubiquitous in eukaryotic, prokaryotic, and also viral genomes with repeat-units of 1–6 bp/nt [1–5]. SSRs have been reported to nonrandomly occur in genomes and associate with different biological significances, which have been gradually recognized as important elements [2, 6, 7]. They have been discovered in both coding and noncoding regions with important roles in modifying morphological features [8], regulating gene expression [9], protecting sequence structures [7], acting as essential boundaries [10], modulating RNA structure and function [11], creating available variants to survive in the host [12] and contributing to genomic evolution [13]. Lots of medical studies have revealed abnormal SSRs in different genomic positions related with more than 40 genetic diseases like fragile X syndrome, Huntington's disease, Friedreich's ataxia and spinocerebellar ataxias type 8, or in

* Correspondence: zhongyangtan@yeah.net
†Douyue Li, Saichao Pan and Hongxi Zhang are co-first authors.
[1]Bioinformatics Center, College of Biology, Hunan University, Changsha 410082, China
Full list of author information is available at the end of the article

many cancers like colorectal cancer, endometrial cancer, gastrointestinal cancer and breast cancer [14–16].

SSRs have been reported to constitute ~ 12% of Japanese pufferfish genome, 15% of rabbit genome, 10% of primate genome and so on [17]; and it has been estimated that SSRs represent over 1 million sites covering 3% of human genome [1, 4]. Though numerous studies and interests were paid to the SSRs in human genome in past decades [1, 4, 15, 17], the clarified SSRs still involve only very few human genomic positions [2]. Rough SSR distributional features have been investigated in human genome [18, 19], and the Genome Browser also provides chance for surveying part of the elementary position of relatively longer SSRs in full human genome [20].

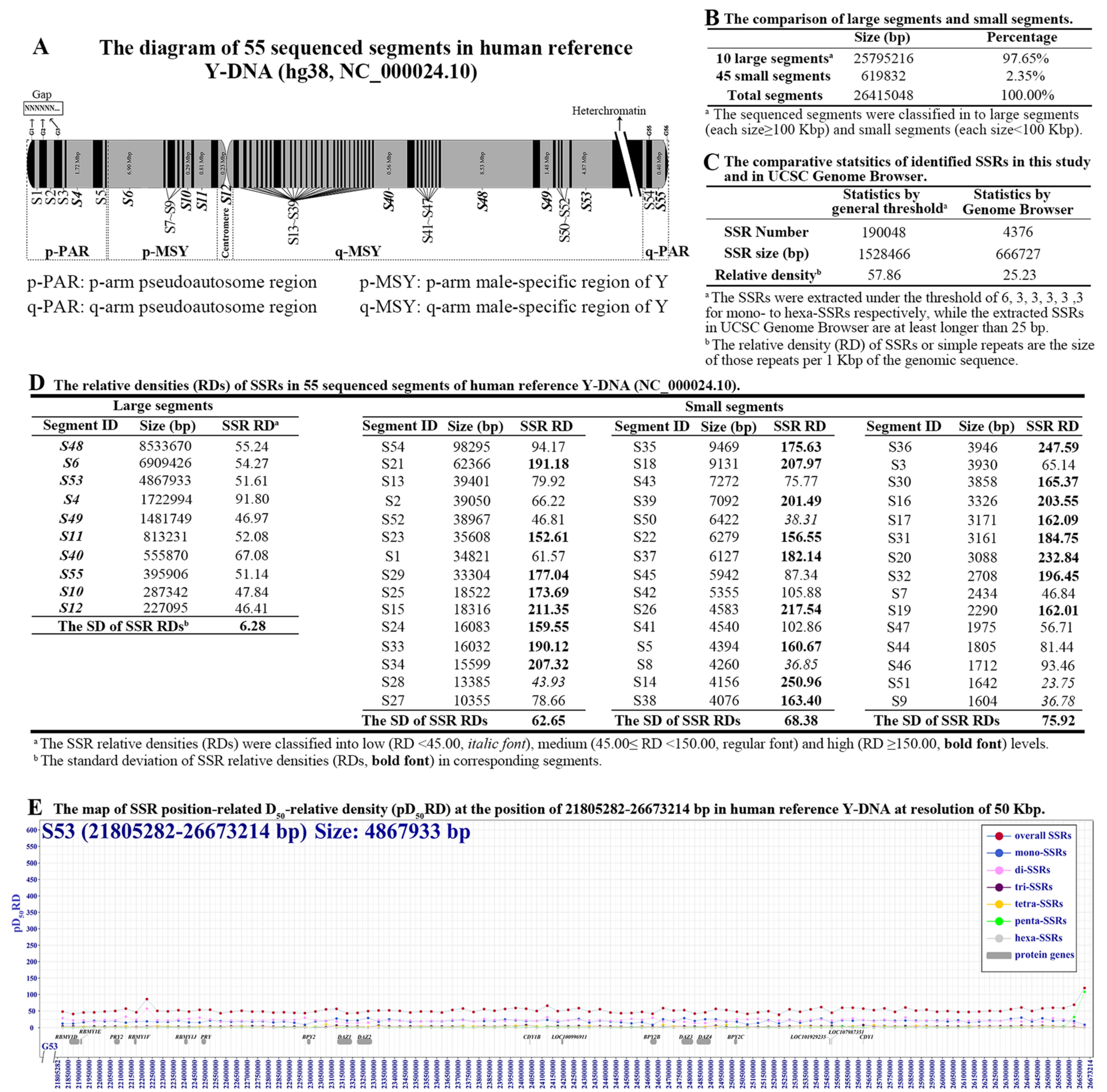
Human chromosome Y is unique with sex-determining genomic compositions and unusual evolutionary history [21, 22]. Human X and Y chromosomes originated from ordinary autosomes beginning at millions of years ago, and the Y chromosome specifically evolved with frequent gene decay and a lack of recombination, making it strikingly different from the X chromosome in size, structure and gene content [23, 24]. Owing to the gene decay and lacking recombination, Y chromosome formed a male-specific region (MSY) that comprise 95% of its length, and this region is flanked on both sides by pseudoautomosomal region (PAR), which can process mitotic recombination with chromosome X. Though human Y chromosome harbors a few genes, it is rich in repetitive and ampliconic elements, including SSRs [23–25]. The mutation rates of many SSRs are significantly high in this chromosome; Willems et al. predicted that the load of de novo SSR mutations is at least 75 mutations per generation in human Y chromosome [26] and Ballantyne et al. estimated that the mutation rates are from 3.78×10^{-4} to 7.44×10^{-2} per Y SSR marker they selected per generation [27]. And researches always prefer to work on these highly polymorphic SSRs in human Y-DNA, like DYS19 or called DYS394, whose sequence is $(TAGA)_3(TAGG)(TAGA)_{7-15}$ in Yp11.1 and mutation rate is 2.5×10^{-4}, as they can be widely applied in forensic investigation, paternity test, population study and evolutionary research [26, 28–33]. The investigations of SSR distributional features are limited to these several highly polymorphic sites, so it is necessary to reveal the exact distributional features of many thousands of SSRs in human Y-DNA at chromosomal level.

Here, we developed a differential calculating method for further exploring the exact distributional features of the SSRs with human reference Y-DNA (NC_000024.10). Hundreds of maps were established to clearly show SSR landscape in every 1 kilobase (Kbp) genomic region of human reference Y-DNA. These SSR landscapes revealed significant regional variation of SSR distributional features in this Y-DNA at the differential resolution of 1 Kbp. This study may provide an important guide for further exploring biological significance and distributional laws of numerous SSRs in human Y-DNA.

Results

The exact distributional features of the SSRs were investigated in the reference sequence of human Y-DNA (NC_000024.10), and this well reviewed Y-DNA is still incompletely sequenced with 55 sequenced segments and 56 gaps (Fig. 1a and Table S1). The sizes of 55 sequenced segments are in range of 1604 ~ 8,533,670 bp, which can be grouped into large (≥100 Kbp) and small (< 100 Kbp) size segments; the large sequenced segments are totally 25,805,216 bp representing about 97.65% of the sequenced part of the reference Y-DNA, and the other 45 small segments represent only 2.35% of the sequenced part (Fig. 1b and Table S1). The total number and size of SSRs are 190,048 and 1,528,466 in this reference Y-DNA under the threshold of 6, 3, 3, 3, 3, 3 (Fig. 1c); and the threshold was widely applied to analyze SSRs in many reported studies [34, 35], which could extract much more SSRs than those in The UCSC (University of California, Santa Cruz) Genome Browser, where the total number of SSRs is only 4376 due to excluding the SSRs shorter than 25 bp by the default settings [20, 36].

The SSR distributions were widely studied with the statistics of relative density (RD) [18, 34, 37]. The average relative density of the SSRs is 57.86 in the total 55 sequenced segments, but the relative density is very different in every sequenced segment. The relative densities of SSRs vary a little from 46.41 to 91.80 with a standard deviation (SD) equaling to 6.28 in the 10 large segments; and that vary a lot from 23.75 to 250.96 with standard deviations more than 62.65 in the 45 small segments, the standard deviation of SSR relative densities was showed to increase obviously as the sizes of investigated segments decreasing (Fig. 1d). These data suggested that the result of SSR relative density is seriously influenced by the statistical segment size; it may not correctly reveal the true features of SSR distributions in these large segments, and the big size may have masked the true distribution features of SSRs in those large segments. As the small segments were showed a great SSR relative density variation and separately located in different parts of the human Y-DNA, the SSR relative density variation may be significantly related to the genome position. These analyses indicate that such relative density method is possibly very limited for analysis of SSRs in big sequence like human genome, and it is necessary to develop new approaches for investigating the exact distribution feature of SSRs in large genomic sequence.



B The comparison of large segments and small segments.

	Size (bp)	Percentage
10 large segments[a]	25795216	97.65%
45 small segments	619832	2.35%
Total segments	26415048	100.00%

[a] The sequenced segments were classified in to large segments (each size≥100 Kbp) and small segments (each size<100 Kbp).

C The comparative statsitics of identified SSRs in this study and in UCSC Genome Browser.

	Statistics by general threshold[a]	Statistics by Genome Browser
SSR Number	190048	4376
SSR size (bp)	1528466	666727
Relative density[b]	57.86	25.23

[a] The SSRs were extracted under the threshold of 6, 3, 3, 3, 3 ,3 for mono- to hexa-SSRs respectively, while the extracted SSRs in UCSC Genome Browser are at least longer than 25 bp.

[b] The relative density (RD) of SSRs or simple repeats are the size of those repeats per 1 Kbp of the genomic sequence.

D The relative densities (RDs) of SSRs in 55 sequenced segments of human reference Y-DNA (NC_000024.10).

Large segments

Segment ID	Size (bp)	SSR RD[a]
S48	8533670	55.24
S6	6909426	54.27
S53	4867933	51.61
S4	1722994	91.80
S49	1481749	46.97
S11	813231	52.08
S40	555870	67.08
S55	395906	51.14
S10	287342	47.84
S12	227095	46.41
The SD of SSR RDs[b]		**6.28**

Small segments

Segment ID	Size (bp)	SSR RD	Segment ID	Size (bp)	SSR RD	Segment ID	Size (bp)	SSR RD
S54	98295	94.17	S35	9469	**175.63**	S36	3946	**247.59**
S21	62366	**191.18**	S18	9131	**207.97**	S3	3930	65.14
S13	39401	79.92	S43	7272	75.77	S30	3858	**165.37**
S2	39050	66.22	S39	7092	**201.49**	S16	3326	**203.55**
S52	38967	46.81	S50	6422	*38.31*	S17	3171	**162.09**
S23	35608	**152.61**	S22	6279	**156.55**	S31	3161	**184.75**
S1	34821	61.57	S37	6127	**182.14**	S20	3088	**232.84**
S29	33304	**177.04**	S45	5942	87.34	S32	2708	**196.45**
S25	18522	**173.69**	S42	5355	105.88	S7	2434	46.84
S15	18316	**211.35**	S26	4583	**217.54**	S19	2290	**162.01**
S24	16083	**159.55**	S41	4540	102.86	S47	1975	56.71
S33	16032	**190.12**	S5	4394	**160.67**	S44	1805	81.44
S34	15599	**207.32**	S8	4260	*36.85*	S46	1712	93.46
S28	13385	*43.93*	S14	4156	**250.96**	S51	1642	*23.75*
S27	10355	78.66	S38	4076	**163.40**	S9	1604	*36.78*
The SD of SSR RDs		**62.65**	**The SD of SSR RDs**		**68.38**	**The SD of SSR RDs**		**75.92**

[a] The SSR relative densities (RDs) were classified into low (RD <45.00, *italic font*), medium (45.00≤ RD <150.00, regular font) and high (RD ≥150.00, **bold font**) levels.

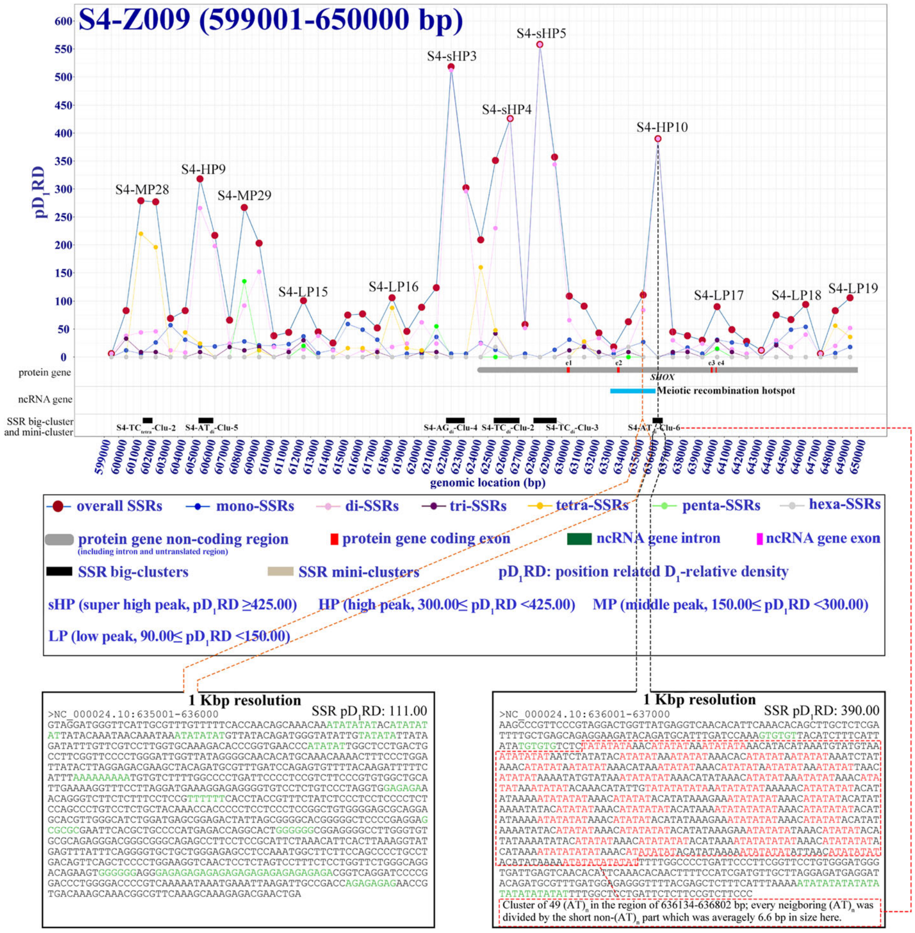
[b] The standard deviation of SSR relative densities (RDs, **bold font**) in corresponding segments.

Fig. 1 The densities of identified SSRs in the sequenced segments of reported human reference Y-DNA (NC_000024.10). **a** The diagram of 55 sequenced segments in human reference Y-DNA (NC_000024.10). **b** The comparison of sequenced segments and small segments. **c** The comparative statistics of identified SSRs in the study and in UCSC Genome Browser. **d** The relative densities (RDs) of SSRs in 55 sequenced segments in human reference Y-DNA (NC_000024.10). **e** The map of SSR position-related D_{50}-relative density at the position of 21,805,282–26,673,214 bp in human reference Y-DNA at resolution of 50 Kbp

To explore the exact features of SSR distributions in the large segment sequences, we developed a Differential Calculator of Microsatellites Version 2 (DCM V2) method, which can calculate SSR densities by dividing the large segments into many differential units, and the alteration of differential unit size may give different resolutions to reveal the feature of SSR distribution; herein, the differential unit size (D_n) was used as the resolutions of 100, 50, 10, 5, 2 and 1 Kbp in 10 large segments. So a SSR position-related D_n-relative density (pD_nRD) concept was introduced in this method. The differential resolutions more than 50 Kbp revealed that the SSR pD_nRD only vary a little around the average relative density value in the sequenced regions of the Y-DNA (Fig. 1e and Fig. S2). As the differential resolution size decreasing, the pD_nRD variation level usually increases in the large segments (Fig. S3), and the 1 Kbp resolution can reveal a clearest pD_1RD variation feature in these large segments of the reference Y-DNA (Fig. 2; Fig. S1.1-S1.540).

The SSRs landscape at 1 Kbp resolution

We obtained 540 maps of SSR position-related relative densities in the reference sequence of human Y-DNA by investigation at 1 Kbp differential resolution, and each map usually contains 51 bins of 1 Kbp with overlapping 1 bin to bilateral maps (Figs. S1.1-S1.540). These maps show an exact landscape of SSR distribution with significant variation of position D_1-relative SSR densities at different genomic positions as described in Figs. 2 and 3. The SSRs were observed to accumulate in 2040 differential bins of 1 Kbp genomic region forming mountain peak like SSR density peaks with pD_1RD much higher than the average relative density in sequenced part of human reference Y-DNA; the SSR density peaks can be divided into 4 levels including 36 super high density peaks (sHP, $pD_1RD \geq 425.00$), 76 high density peaks (HP, $300.00 \leq pD_1RD < 425.00$), 528 middle density peaks (MP, $150.00 \leq pD_1RD < 300.00$) and 1400 low density peaks (LP, $90.00 \leq pD_1RD < 150.00$) (Fig. 4 and Figs. S1.1–1.540). On the contrary, SSRs appear with extremely low densities in some genomic regions, and these regions can be grouped into 3 kinds including 2 big SSR extremely low density regions (bELR, RD < 25.00, size ≥100 Kbp), 137 small SSR extremely low density regions (sELR, RD < 25.00, 3 Kbp ≤ size < 100 Kbp) and 69 SSR desert regions (ZD, $pD_1RD = 0$, size ≤2 Kbp) (Fig. 4 and Figs. S1.1-S1.540). Therefore, the 51 bins usually have different pD_1RD making each map mixed with different SSR density peaks and extremely low density region, and the 540 maps can be typically classified into 6 types: 74 HML type maps with mix of high, middle and low density peaks, 202 ML type maps with mix of middle and low peaks, 212 L type maps with only low peaks, 16 Penta type maps with domination of pentanucleotide SSRs, 31 AV type maps with all pD_1RD close to the genomic average relative density, and 5 EL type maps with all pD_1RD very lower than average (Fig. 3 and Figs. S1.1-S1.540).

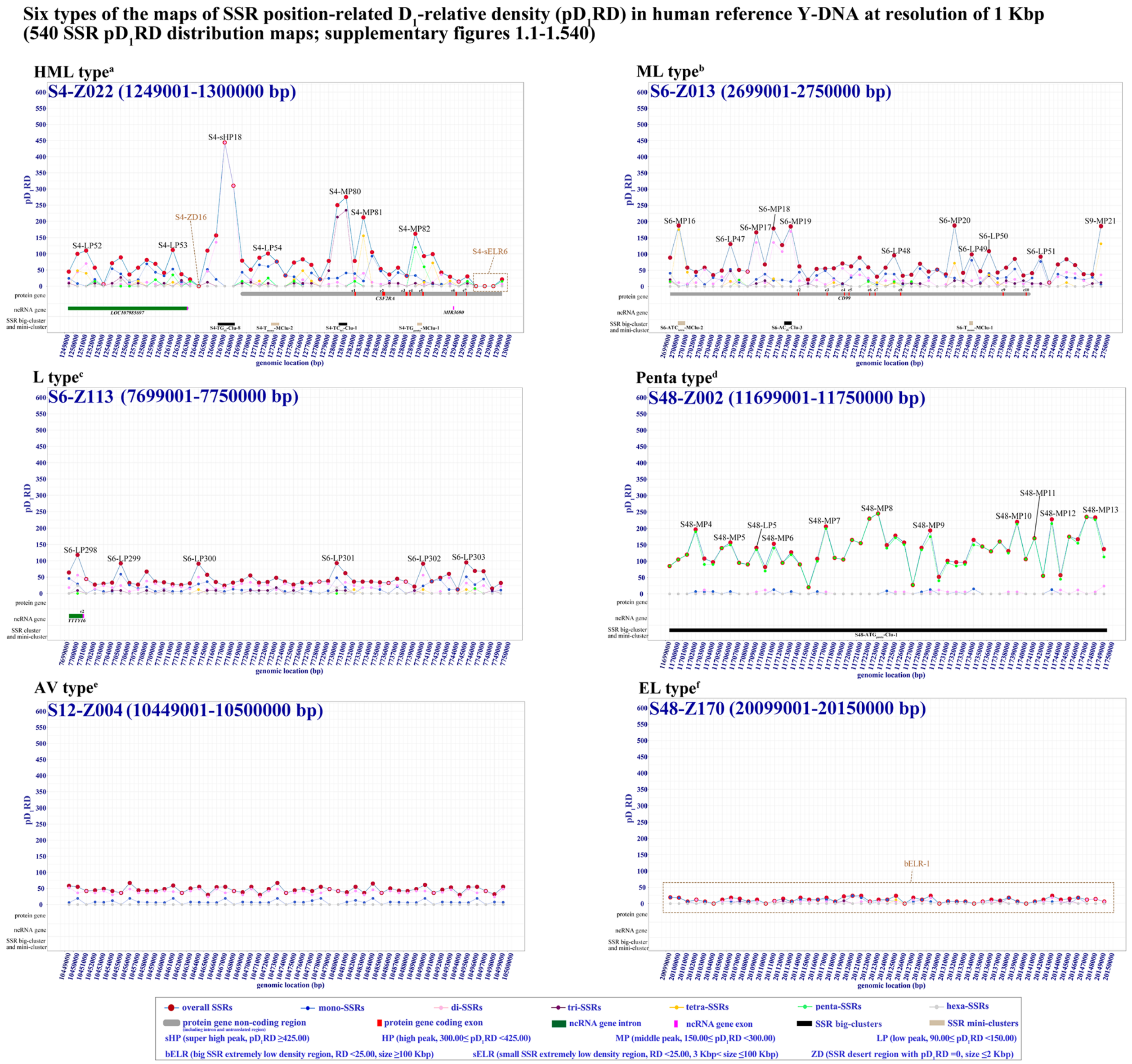


```
>NC_000024.10:635001-636000
GTAGGATGGGTTCATTGCGTTTGTTTTTCACCAACAGCAAACAAATATATATACATATAT
ATTATACAAATAACAAATAAATATATATGTTATACAGATGGGTATATTGTATATATTATA
GATATTTGTTCGTCCTTGGTGCAAAGACACCCGGTGAACCCATATATTGGCTCCTGACTG
CCTTCGGTTCCCCTGGGATTGGTTATAGGGGCAACACATGCAAACAAAACTTTCCCTGGA
TTATACTTAGGAGACGAAGCTACAGATGCGTTTGATCCAGAGTGTTTTACAAGATTTTTC
ATTTAAAAAAAAATGTGTCTTTTGGCCCCTGATTCCCCTCCGTCTTCCCGTGTGGCTGCA
TTGAAAAGGTTTCCTTAGGATGAAAGGAGAGGGGTGTCCTCTGTCCCTAGGTGGAGAGAA
ACAGGGTCTTCTCTTTCCTCCGTTTTTTCACCTACCGTTTCTATCTCCCTCCTCCCCTCT
CCAGCCCTGTCCTCTGCTACAAACCACCCCCTCCTCCCTCCGGCTGTGGGGAGCGCAGGA
GCACGTTGGGCATCTGGATGAGCGGAGACTATTAGCGGGGCACGGGGGCTCCCCGAGGAG
CGCGCGAATTCACGCTGCCCCATGAGACCAGGCACTGGGGGGCGGAGGGGCCTTGGGTGT
GCGCAGAGGGACGGGCGGGCAGAGCCTTCCTCCGCATTCTAAACATTCACTTAAAGGTAT
GAGTTTATTTCAGGGGTGCTGCTGGGAGAGCCTCCAAATGGCTTCTTCCAGCCCCTGCCT
GACAGTTCAGCTCCCCTGGAAGGTCAACTCCTCTAGTCCTTTCTCCTGGTTCTGGGCAGG
ACAGAAGTGGGGGGAGGGAGAGAGAGAGAGAGAGAGAGAGAGAGAGACGGTCAGGATCCCCG
GACCCTGGGGACCCCGTCAAAAATAAATGAAATTAAGATTGCCGACCAGAGAGAGAACCG
TGACAAAGCAAACGGCGTTCAAAGCAAAGAGACGAACTGA
```

```
>NC_000024.10:636001-637000
AAGCCCGTTCCCGTAGGACTGGTTATGAGGTCAACACATTCAAACACAGCTTGCTCTCGA
TTTTGCTGAGCAGAGGAAGATACAGATGCATTTGATCCAAAGTGTGTTACATCTTTCATT
ATATGTGTGTCTCTATATATATAAACATATATAAATATATATAAACATACATAAATGTATGTAA
ATATATATAATCTATATACATATATAAATATATAAACACATATATAATATATAAATCTAT
AAACATATATAATATATAAACATAAATATATAAACATATATAATATATAAATATATTAAC
ATATATAAAATATGTATAAATATATATAAACATATAAACATATATAAATATATAAACATA
TATAAATATATACAAACATATTGTATATATATAAATATATATAAAAACATATATATACAT
ATAAAAATATATATAAACATATATACATATAAAGAAATATATATAAACATATATACATAT
AAAATATACATATATAAACATATATATACATAAAATATATATATAAACATATATATACAT
ATAAAAATATATATAAACATATATACATATAAAGAAATATATATAAACATATATACATAT
AAAATATACATATATAAACATATATATACATATAAAGAAATATATATAAACATATATACA
TATAAAATATACATATATAAACATATATATACATAAAATATATATATAAACATATATATA
CATAAAATATATATATAAACATATATATACATATAAAAATATATATATTAACATATATAT
ACATATAAAAATATATATATATTTTTGGCCCCTGATTCCCTTCGGTTCCTGTGGGATGGG
TGATTGAGTCAACACATTCAAACACAACTTTTCCATCGATGTTGCTTAGGAGATGAGGAT
ACAGATGCGTTTGATGGAGAGGGTTTTACGAGCTCTTTCATTTAAAAATATATATATATA
TATATATATATTTTGGCTCCTGATTCTCTTCCGTCTTCCC
```

Cluster of 49 $(AT)_n$ in the region of 636134-636802 bp; every neigboring $(AT)_n$ was divided by the short non-$(AT)_n$ part which was averagely 6.6 bp in size here.

Fig. 2 A typical map of SSR position-related D_1-relative density (pD_1RD) in human reference Y-DNA (NC_000024.10) at resolution of 1 Kbp

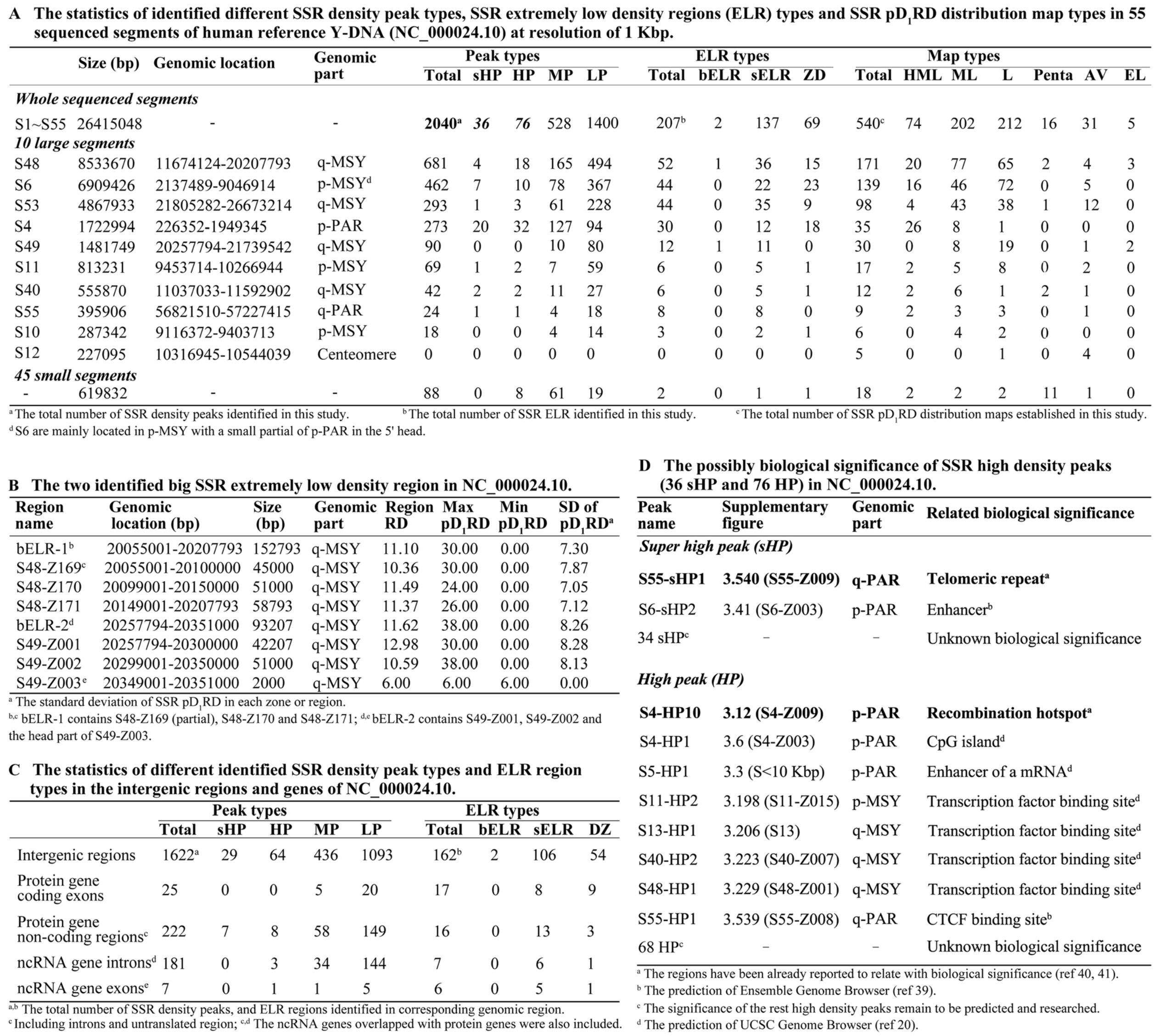


[a] The **HML** type map contains pD$_1$RD characteristics with a mix of high peak (H), middle peak (M) and low peak (L).
[b] The **ML** type map contains pD$_1$RD characteristics with a mix of middle peak (M) and low peak (L).
[c] The **L** type map contains pD$_1$RD characteristics with only low peak (L).
[d] The **Penta** type map is dominated with pentanucleotide SSRs.
[e] The **AV** type map shows SSR pD$_1$RD are mostly close to the average (AV) relative density in the whole Y-DNA.
[f] The **EL** type map contains characterstics with extremely low (EL) pD$_1$RD.

Fig. 3 The six types of SSR pD$_1$RD distribution maps in human reference Y-DNA (NC_000024.10) at resolution of 1 Kbp

Clusters of microsatellites

It was also found that there are large numbers of SSRs with same or similar motif which neighborly locate together without other SSR motif in many regions of this human reference Y-DNA (Table S2). Some of these regions even harbor hundreds of such kind of same or similar SSR motifs, for example, there are 430 $(CT/TC)_6$ without other SSR motif at the region of 95,647–133,828 bp of the Y-DNA (Fig. 5A.1); and some harbor dozens of or more than 3 same or similar SSR motifs, like 15 $(AT/TA)_n$ at the region of 7,426,653–7,426,857 bp (Fig. 5A.2) and 5 $(AAAG/AAGA/AGAA/GAAA)_n$ at the region of 56,858,319–56,858,540 bp of the Y-DNA (Fig. 5A.3). The regions of these specific SSR distributions can be defined as SSR clusters in this study; there are totally 8109 identified SSR clusters in sequenced part of the Y-DNA, which can be grouped into 3 levels including 203 big clusters (Clu, clustered same (similar) SSR number ≥ 26), 355 mini-clusters (MClu, 9 ≤ clustered same (similar) SSR number < 26)



A The statistics of identified different SSR density peak types, SSR extremely low density regions (ELR) types and SSR pD_1RD distribution map types in 55 sequenced segments of human reference Y-DNA (NC_000024.10) at resolution of 1 Kbp.

	Size (bp)	Genomic location	Genomic part	Peak types					ELR types				Map types						
				Total	sHP	HP	MP	LP	Total	bELR	sELR	ZD	Total	HML	ML	L	Penta	AV	EL
Whole sequenced segments																			
S1~S55	26415048	-	-	**2040**[a]	***36***	***76***	528	1400	207[b]	2	137	69	540[c]	74	202	212	16	31	5
10 large segments																			
S48	8533670	11674124-20207793	q-MSY	681	4	18	165	494	52	1	36	15	171	20	77	65	2	4	3
S6	6909426	2137489-9046914	p-MSY[d]	462	7	10	78	367	44	0	22	23	139	16	46	72	0	5	0
S53	4867933	21805282-26673214	q-MSY	293	1	3	61	228	44	0	35	9	98	4	43	38	1	12	0
S4	1722994	226352-1949345	p-PAR	273	20	32	127	94	30	0	12	18	35	26	8	1	0	0	0
S49	1481749	20257794-21739542	q-MSY	90	0	0	10	80	12	1	11	0	30	0	8	19	0	1	2
S11	813231	9453714-10266944	p-MSY	69	1	2	7	59	6	0	5	1	17	2	5	8	0	2	0
S40	555870	11037033-11592902	q-MSY	42	2	2	11	27	6	0	5	1	12	2	6	1	2	1	0
S55	395906	56821510-57227415	q-PAR	24	1	1	4	18	8	0	8	0	9	2	3	3	0	1	0
S10	287342	9116372-9403713	p-MSY	18	0	0	4	14	3	0	2	1	6	0	4	2	0	0	0
S12	227095	10316945-10544039	Centeomere	0	0	0	0	0	0	0	0	0	5	0	0	1	0	4	0
45 small segments																			
-	619832	-	-	88	0	8	61	19	2	0	1	1	18	2	2	2	11	1	0

[a] The total number of SSR density peaks identified in this study. [b] The total number of SSR ELR identified in this study. [c] The total number of SSR pD_1RD distribution maps established in this study.
[d] S6 are mainly located in p-MSY with a small partial of p-PAR in the 5' head.

B The two identified big SSR extremely low density region in NC_000024.10.

Region name	Genomic location (bp)	Size (bp)	Genomic part	Region RD	Max pD_1RD	Min pD_1RD	SD of pD_1RD[a]
bELR-1[b]	20055001-20207793	152793	q-MSY	11.10	30.00	0.00	7.30
S48-Z169[c]	20055001-20100000	45000	q-MSY	10.36	30.00	0.00	7.87
S48-Z170	20099001-20150000	51000	q-MSY	11.49	24.00	0.00	7.05
S48-Z171	20149001-20207793	58793	q-MSY	11.37	26.00	0.00	7.12
bELR-2[d]	20257794-20351000	93207	q-MSY	11.62	38.00	0.00	8.26
S49-Z001	20257794-20300000	42207	q-MSY	12.98	30.00	0.00	8.28
S49-Z002	20299001-20350000	51000	q-MSY	10.59	38.00	0.00	8.13
S49-Z003[e]	20349001-20351000	2000	q-MSY	6.00	6.00	6.00	0.00

[a] The standard deviation of SSR pD_1RD in each zone or region.
[b,c] bELR-1 contains S48-Z169 (partial), S48-Z170 and S48-Z171; [d,e] bELR-2 contains S49-Z001, S49-Z002 and the head part of S49-Z003.

C The statistics of different identified SSR density peak types and ELR region types in the intergenic regions and genes of NC_000024.10.

	Peak types					ELR types			
	Total	sHP	HP	MP	LP	Total	bELR	sELR	DZ
Intergenic regions	1622[a]	29	64	436	1093	162[b]	2	106	54
Protein gene coding exons	25	0	0	5	20	17	0	8	9
Protein gene non-coding regions[c]	222	7	8	58	149	16	0	13	3
ncRNA gene introns[d]	181	0	3	34	144	7	0	6	1
ncRNA gene exons[e]	7	0	1	1	5	6	0	5	1

[a,b] The total number of SSR density peaks, and ELR regions identified in corresponding genomic region.
[c] Including introns and untranslated region; [c,d] The ncRNA genes overlapped with protein genes were also included.

D The possibly biological significance of SSR high density peaks (36 sHP and 76 HP) in NC_000024.10.

Peak name	Supplementary figure	Genomic part	Related biological significance
Super high peak (sHP)			
S55-sHP1	**3.540 (S55-Z009)**	**q-PAR**	**Telomeric repeat**[a]
S6-sHP2	3.41 (S6-Z003)	p-PAR	Enhancer[b]
34 sHP[c]	–	–	Unknown biological significance
High peak (HP)			
S4-HP10	**3.12 (S4-Z009)**	**p-PAR**	**Recombination hotspot**[a]
S4-HP1	3.6 (S4-Z003)	p-PAR	CpG island[d]
S5-HP1	3.3 (S<10 Kbp)	p-PAR	Enhancer of a mRNA[d]
S11-HP2	3.198 (S11-Z015)	p-MSY	Transcription factor binding site[d]
S13-HP1	3.206 (S13)	q-MSY	Transcription factor binding site[d]
S40-HP2	3.223 (S40-Z007)	q-MSY	Transcription factor binding site[d]
S48-HP1	3.229 (S48-Z001)	q-MSY	Transcription factor binding site[d]
S55-HP1	3.539 (S55-Z008)	q-PAR	CTCF binding site[b]
68 HP[c]	–	–	Unknown biological significance

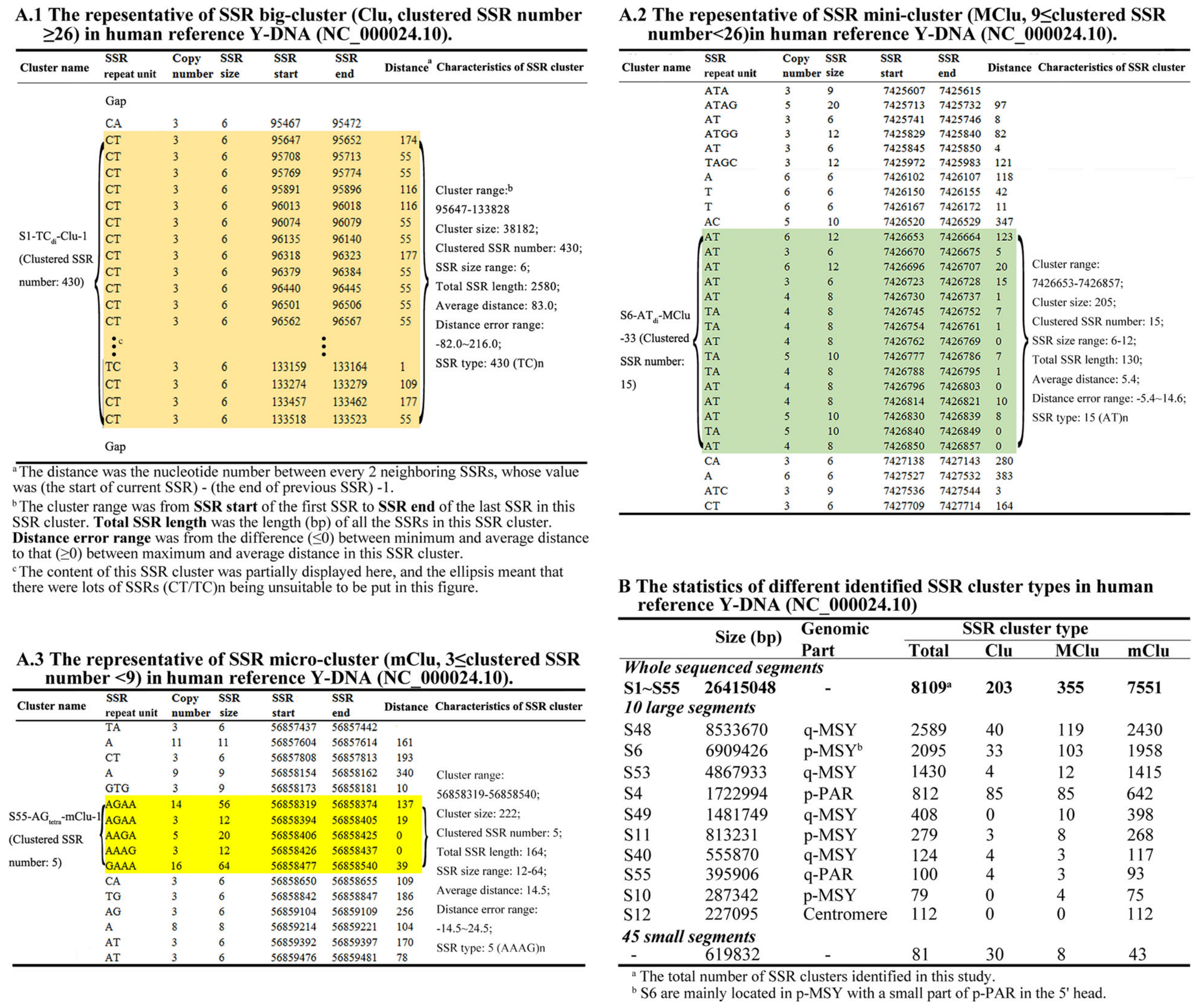
[a] The regions have been already reported to relate with biological significance (ref 40, 41).
[b] The prediction of Ensemble Genome Browser (ref 39).
[c] The significance of the rest high density peaks remain to be predicted and researched.
[d] The prediction of UCSC Genome Browser (ref 20).

Fig. 4 The statistics of different feature types of SSR pD_1RD distributions in human reference Y-DNA (NC_000024.10) at resolution of 1 Kbp**. a** The statistics of identified different SSR density peak types, SSR extremely low density regions (ELR) types and SSR pD_1RD distribution map types. **b** The two identified big SSR extremely low density region. **c** The statistics of different identified SSR density peak types and ELR region types in the intergenic regions and genes. **d** The possibly biological significance of SSR high density peaks (36 sHP and 76 HP), and details were listed in Table S4

and 7551 micro-clusters (mClu, 3 ≤ clustered same SSR number < 9) (Fig. 5b and Table S3).

Discussion

Our comprehensive survey of microsatellite distributions at 1 Kbp differential resolution to gain an exact landscape in the human reference Y-DNA (NC_000024.10), and 540 SSR landscape maps were obtained; these maps show that SSRs are accumulated significantly in some small regions and also seriously sparse in some regions; and many same or similar motif SSRs were observed to locate neighborly forming SSR clusters. Large numbers of SSRs in human Y-DNA have been previously understudied because the related studies usually focus on some significant Y SSR markers, or only analyzed the average distributions in the coding, noncoding and intergenic regions [7, 18, 19, 25, 28, 29, 38]. And UCSC Genome Browser might be not specific to highlight microsatellite distributional variation in every genomic position [20] (Fig. S4). The 540 SSR landscape maps in this study can provide a comprehensive view of clear SSR distributional features in every 1 Kbp genomic region along the human reference Y-DNA, and these maps can detailedly highlight the significant variations of



A.1 The repesentative of SSR big-cluster (Clu, clustered SSR number ≥26) in human reference Y-DNA (NC_000024.10).

Cluster name	SSR repeat unit	Copy number	SSR size	SSR start	SSR end	Distance[a]	Characteristics of SSR cluster
	Gap						
	CA	3	6	95467	95472		

S1-TC$_{di}$-Clu-1 (Clustered SSR number: 430)	CT	3	6	95647	95652	174	Cluster range:[b] 95647-133828
							Cluster size: 38182;
							Clustered SSR number: 430;
							SSR size range: 6;
							Total SSR length: 2580;
							Average distance: 83.0;
							Distance error range: -82.0~216.0;
							SSR type: 430 (TC)n

	CT	3	6	95708	95713	55	
	CT	3	6	95769	95774	55	
	CT	3	6	95891	95896	116	
	CT	3	6	96013	96018	116	
	CT	3	6	96074	96079	55	
	CT	3	6	96135	96140	55	
	CT	3	6	96318	96323	177	
	CT	3	6	96379	96384	55	
	CT	3	6	96440	96445	55	
	CT	3	6	96501	96506	55	
	CT	3	6	96562	96567	55	
	⋮[c]			⋮			
	TC	3	6	133159	133164	1	
	CT	3	6	133274	133279	109	
	CT	3	6	133457	133462	177	
	CT	3	6	133518	133523	55	
	Gap						

[a] The distance was the nucleotide number between every 2 neighboring SSRs, whose value was (the start of current SSR) - (the end of previous SSR) -1.

[b] The cluster range was from **SSR start** of the first SSR to **SSR end** of the last SSR in this SSR cluster. **Total SSR length** was the length (bp) of all the SSRs in this SSR cluster. **Distance error range** was from the difference (≤0) between minimum and average distance to that (≥0) between maximum and average distance in this SSR cluster.

[c] The content of this SSR cluster was partially displayed here, and the ellipsis meant that there were lots of SSRs (CT/TC)n being unsuitable to be put in this figure.

A.2 The repesentative of SSR mini-cluster (MClu, 9≤clustered SSR number<26)in human reference Y-DNA (NC_000024.10).

Cluster name	SSR repeat unit	Copy number	SSR size	SSR start	SSR end	Distance	Characteristics of SSR cluster
	ATA	3	9	7425607	7425615		
	ATAG	5	20	7425713	7425732	97	
	AT	3	6	7425741	7425746	8	
	ATGG	3	12	7425829	7425840	82	
	AT	3	6	7425845	7425850	4	
	TAGC	3	12	7425972	7425983	121	
	A	6	6	7426102	7426107	118	
	T	6	6	7426150	7426155	42	
	T	6	6	7426167	7426172	11	
	AC	5	10	7426520	7426529	347	

S6-AT$_{di}$-MClu -33 (Clustered SSR number: 15)	AT	6	12	7426653	7426664	123	Cluster range: 7426653-7426857;
							Cluster size: 205;
							Clustered SSR number: 15;
							SSR size range: 6-12;
							Total SSR length: 130;
							Average distance: 5.4;
							Distance error range: -5.4~14.6;
							SSR type: 15 (AT)n

	AT	3	6	7426670	7426675	5	
	AT	6	12	7426696	7426707	20	
	AT	3	6	7426723	7426728	15	
	AT	4	8	7426730	7426737	1	
	TA	4	8	7426745	7426752	7	
	TA	4	8	7426754	7426761	1	
	AT	4	8	7426762	7426769	0	
	TA	5	10	7426777	7426786	7	
	TA	4	8	7426788	7426795	1	
	AT	4	8	7426796	7426803	0	
	AT	4	8	7426814	7426821	10	
	AT	5	10	7426830	7426839	8	
	TA	5	10	7426840	7426849	0	
	AT	4	8	7426850	7426857	0	
	CA	3	6	7427138	7427143	280	
	A	6	6	7427527	7427532	383	
	ATC	3	9	7427536	7427544	3	
	CT	3	6	7427709	7427714	164	

A.3 The repesentative of SSR micro-cluster (mClu, 3≤clustered SSR number <9) in human reference Y-DNA (NC_000024.10).

Cluster name	SSR repeat unit	Copy number	SSR size	SSR start	SSR end	Distance	Characteristics of SSR cluster
	TA	3	6	56857437	56857442		
	A	11	11	56857604	56857614	161	
	CT	3	6	56857808	56857813	193	
	A	9	9	56858154	56858162	340	
	GTG	3	9	56858173	56858181	10	

S55-AG$_{tetra}$-mClu-1 (Clustered SSR number: 5)	AGAA	14	56	56858319	56858374	137	Cluster range: 56858319-56858540;
							Cluster size: 222;
							Clustered SSR number: 5;
							Total SSR length: 164;
							SSR size range: 12-64;
							Average distance: 14.5;
							Distance error range: -14.5~24.5;
							SSR type: 5 (AAAG)n

	AGAA	3	12	56858394	56858405	19	
	AAGA	5	20	56858406	56858425	0	
	AAAG	3	12	56858426	56858437	0	
	GAAA	16	64	56858477	56858540	39	
	CA	3	6	56858650	56858655	109	
	TG	3	6	56858842	56858847	186	
	AG	3	6	56859104	56859109	256	
	A	8	8	56859214	56859221	104	
	AT	3	6	56859392	56859397	170	
	AT	3	6	56859476	56859481	78	

B The statistics of different identified SSR cluster types in human reference Y-DNA (NC_000024.10)

	Size (bp)	Genomic Part	SSR cluster type			
			Total	Clu	MClu	mClu
Whole sequenced segments						
S1~S55	**26415048**	-	**8109**[a]	**203**	**355**	**7551**
10 large segments						
S48	8533670	q-MSY	2589	40	119	2430
S6	6909426	p-MSY[b]	2095	33	103	1958
S53	4867933	q-MSY	1430	4	12	1415
S4	1722994	p-PAR	812	85	85	642
S49	1481749	q-MSY	408	0	10	398
S11	813231	p-MSY	279	3	8	268
S40	555870	q-MSY	124	4	3	117
S55	395906	q-PAR	100	4	3	93
S10	287342	p-MSY	79	0	4	75
S12	227095	Centromere	112	0	0	112
45 small segments						
-	619832	-	81	30	8	43

[a] The total number of SSR clusters identified in this study.

[b] S6 are mainly located in p-MSY with a small part of p-PAR in the 5' head.

Fig. 5 The clusters of many SSRs with same or similar motif in human reference Y-DNA (NC_000024.10). **(A.1-A.3)** The typical 3 levels of SSR clusters including SSR big clusters, mini-clusters and micro-clusters. (**B**) The statistics of different identified SSR cluster types in human reference Y-DNA (NC_000024.10)

position-related microsatellite distributions in this Y-DNA. And our studies may be helpful to reveal the microsatellite distributional laws and to further explore the biological significance of SSRs in the human reference Y-DNA.

Our observation of significant SSR accumulations to form density peaks indicates an obviously statistic bias of SSR distributions in the human reference Y-DNA, and such accumulations were also observed in other human and mammal Y-DNA (Figs. S5 and S6), suggesting that these SSR accumulations with forming high density peaks were possibly selected for being related to some biological significances. There are 112 identified high density peaks including 36 super high peaks and 76 high peaks in this study, implicating that the highly significant SSR accumulating regions totally represent 0.4% (112 Kbp / 26,415 Kbp) of the whole sequenced regions of the human reference Y-DNA, which are worth focusing on. And 10 of these 112 peaks have been already reported to possibly be related with known biological significance (Fig. 4d) [10, 19, 20, 39–41], for example, S4-HP10 is in a reported recombination hotspot in the p-arm pseudoautosomal region (p-PAR) of the Y-DNA, which might contribute to the mitotic recombination (Fig. 2 and 4d, Table S4) [40]; S55-sHP1 is in the telomeric region at the q-arm of the Y-DNA, which might be the boundary between telomere and euchromatic region [7, 19] (Fig. 4d). Though the biological significances of the other 102 peaks are not reported as many SSRs being lacking of understanding in human Y-DNA originally, these peaks may also play some important biological roles potentially, which probably deserve to be further explored [6, 10, 42–45]. In addition, those middle and low peaks possibly be also helpful to some biological process. Therefore, our fine maps of SSR landscape may provide the guide for mining the biological significance of such significant SSR accumulations with high densities in the human reference Y-DNA.

SSRs are widely considered to be more in noncoding and intergenic regions than coding regions [1–3, 6]. However, it is inconsistent with the wide consideration that the SSRs were detected to occur with extremely low densities in 2 big intergenic regions and many small noncoding and intergenic regions of the human reference Y-DNA, moreover, the SSRs were even not discovered in dozens of noncoding and intergenic regions of this Y-DNA, which have never been reported before to our knowledge (Figs. 3, 4B and C, Table S5). These events indicate the SSRs occurring in these regions might be not well tolerated and be exposed to strong negative selection as the commonly accepted view illustrating that many SSRs have evolved under serious selective constraints in genomes [1, 13, 46]. Our SSR landscape maps showed significantly regional variation of SSR densities in human reference Y-DNA, and it may be worthwhile to deeply explore the relationship between the selection and the significantly regional variation of SSR densities. Thus, our studies could significantly augment the knowledge of SSR landscapes in every different coding, noncoding and intergenic region of human Y-DNA, and also may contribute to the study of SSR evolution in human Y-DNA.

Besides clearly showing SSR landscapes, this study also revealed that many SSRs with same or similar motif are located neighborly to form over 8000 different sizes of SSR clusters in the sequenced regions of human reference Y-DNA. Many SSR clusters were detected to distribute numbers of SSRs with same motif and small neighboring SSR distances, contributing to forming the SSR density peaks, and some SSR clusters even distribute hundreds of identical SSRs with regular neighboring SSR distances (Tables S2 and S3). SSRs were reported to likely have evolved from an initial expansion of a short existing sequence motif and to be the quickly expandable compositions in genomes [5, 13, 17, 46]; so, the SSR clusters can be also assumed to have evolved from the expansions of an existing SSR motif, and it may be valuable for further studying their evolutionary process in human Y-DNA.

Conclusions

SSRs are commonly thought to be not just the simple sequences randomly distributed in genomes. The distributional laws of SSRs in human Y-DNA still remain to be further studied, but we hope that this study with 540 exact SSR landscape maps in human reference Y-DNA will help to elucidate the genetic and evolutionary mechanism involved, and our DCM 2.0 method can contribute to clarifying the SSR landscapes in other genomic sequences.

Methods

Genomic sequence selection

We selected the reported reference genomic sequence of human chromosome Y (NC_000024.10), which is the published human Y-DNA with the highest sequencing completion (yet unfinished) and most convincing accuracy. The non-sequenced compositions (gaps) separated this reference Y-DNA into 55 different sequenced segments, which can be grouped into large (≥100 Kbp) and small (< 100 Kbp) size segments (Table S1).

SSR identification

IMEx, a program with friendly interface, was utilized to identify perfect SSRs in this analysis [47]. The lowest copy number of repeat units, which is the threshold to identify the SSRs, was 6, 3, 3, 3, 3, 3 for extracting mono-, di-, tri-, tetra-, penta- and hexanucleotide SSRs respectively according to empirical criterion and previous SSR studies [34, 35]. Some other programs like Tandem Repeat Finder and RepeatMasker, which usually exclude many short SSRs (with default parameters, SSRs shorter than 25 bp are filtered out), were mainly applied as the reference tools here [36, 48].

Method for fine maps of SSR landscapes in large genomic sequence

The former statistics of SSR relative density (RD) was usually calculated by the total SSR size dividing the total size of the sample genomic sequence [35, 49, 50], which could be described as the following formula:

$$\mathrm{D_n} = \mathrm{n_1} = \mathrm{n_2} = \mathrm{n_3} = \ldots = \mathrm{n_i} = \ldots = \mathrm{n_{la}} \qquad (1)$$

In this formula (1), M is the total size of SSRs (microsatellites) in the sample genomic sequence; N is the size of the sample genomic sequence. This method can only illustrate the global average value of SSR distributions in the sample genomic sequences, and the size of sample genomic sequence is usually very large, so this average value is actually not able to represent exact features of the SSR distributions in corresponding sequence.

We developed the Differential Calculator of Microsatellites Version 2.0 (DCM v2.0) method in the basis of DCM v1.0 [51] to calculate SSR relative density in every region along the large genomic sequence, which is aimed to reveal fine SSR landscapes in the genomic sequences. Firstly, DCM v2.0 can partition the large genomic sequence (N) into numerous differential units, which can be described as:

$$N = \sum n_i \tag{2}$$

$$D_n = n_1 = n_2 = n_3 = \ldots = n_i = \ldots = n_{la} \tag{3}$$

In these 2 formulas, n_i is the size of the i-th differential bin (i = 1, 2, 3, … i, i + 1, … la; la = the last) in large genomic sequence; la is equal to rounding (N/D_n) to up integer; D_n represents the resolution of differential unit size (Kbp), e.g., D_{50} means the differential resolution of 50 Kbp.

The differential resolution size (D_n) critically affects the exactness of revealing SSR landscapes as well as the pixels affect the image quality, and la is negative proportional to D_n. The large genomic sequence is like a large differential resolution size with la =1, which is not proper mentioned above, so it is necessary to process an investigation to adjust into the best size for revealing fine SSR landscapes.

Then DCM 2.0 can calculate the SSR relative density in each differential unit respectively, equaling to partitioning the total SSR size (M) into local SSR sizes in numerous differential units of the large genomic sequence, expressed as the following formula:

$$M = \sum m_i \tag{4}$$

In this formula (4), m_i is the size of SSRs in the i-th differential bin; it is also critically affected by the differential resolution size (D_n).

Each differential unit represents a small region of the large genomic sequence, so a concept of SSR position-related D_n-relative density (pD_nRD) was introduced in this method, which reflects SSR distributional features are critically influenced by both position (the i-th differential bin) and differential resolution size (D_n), and it can be expressed as:

$$pD_nRD_i = \frac{m_i}{D_n} \times 100 \tag{5}$$

pD_nRD_i is the pD_nRD in the i-th differential bin at the large genomic sequence. And the standard deviation (SD) of pD_nRD was used to test the exactness of SSR distributional features at the corresponding differential resolution size.

Visualization for exact SSR landscapes

Ggplot2, a R package, was utilized to visualize the SSR pD_nRD distributions in the reference human Y-DNA into the Figures. Among the maps of SSR pD_nRD distribution at 1 Kbp resolution (pD_1RD), a normal size map represents a zone including 51 differential units of 1 Kbp with overlapping 1 unit to bilateral maps, and each zone was labeled into a zone serial number, e.g., S4-Z003 represents the third zone in S4. Owing to the gaps in this Y-DNA, some zones are in unnormal size, including those of the starts and ends of large sequenced segments, and those of short sequenced segments (labeled into only segment name, e.g., S13); the much short sequenced segments (size < 10 Kbp) were all integrated into the same map. The reported protein coding genes and ncRNA genes were also marked in these maps according to the annotation of NC_000024.10.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07389-5.

Additional file 1: Table S1. The sequenced segments and SSR statistics of human reference Y-DNA (NC_000024.10).

Additional file 2: Table S2. The SSR clusters in the original statistics of SSR extraction in human reference Y-DNA (NC_000024.10)(S1: 10001–44,821 bp).

Additional file 3: Table S3. The features of identified SSR mini-clusters in 55 sequenced segments of human reference Y-DNA (NC_000024.10) (mini-cluster (MClu), 9 ≤ clustered same or similar SSR number < 26).

Additional file 4: Table S4. The features of identified super high density peaks in 55 segments of human reference Y-DNA (NC_000024.10) at 1 Kbp resolution (super high peak (sHP), pD1RD ≥ 425.00).

Additional file 5: Table S5. The statistics of identified different position related D_1-relative density (pD_1RD) map types in 55 sequenced segments of human reference Y-DNA (NC_000024.10).

Additional file 6: Figure S1. 540 SSR position related D_1-relative density maps in human Y-DNA (NC_000024.10) at 1 kilobase resolution.

Additional file 7: Figure S2. The SSR position related D_{50}-relative density ($pD_{50}RD$) map in 10 large segments of human reference Y-DNA (NC_000024.10).

Additional file 8: Figure S3. The comparison of SSR position related D_n-relative density maps at differential resolutions of 100, 50, 10, 5, 2, 1 Kbp in human reference Y-DNA (NC_000024.10).

Additional file 9: Figure S4. The comparison of showing SSR distributional features in SSR pD_1RD map and UCSC Genome Browser at the position of 1200001-1300000 bp in human reference Y-DNA (NC_000024.10).

Additional file 10: Figure S5. The example of comparing high SSR accumulations between human reference Y-DNA and other human Y-DNA at the same locations.

Additional file 11: Figure S6. The similar SSR accumulations of high densities located at the flanking region (about 15000 bp away) of SRY in human chimpanzee, rhesus monkey, mouse and rat reference Y.

Abbreviations

SSR: Simple sequence repeat; MSY: Male-specific region in Y; PAR: Pesudoautomosomal region; Kbp: kilobase; UCSC: University of California, Santa Cruz; RD: Relative density; SD: Standard deviation; D_n: Differential unit size; pD_nRD: Position-related D_n-relative density; DCM V2: Differential Calculator of Microsatellites Version 2; sHP: Super high peak; HP: High peak; MP: Middle peak; LP: Low peak; bELR: Big extremely low (density) region; sELR: Small extremely low (density) region; ZD: Zero (density) desert; Clu: Big-cluster; MClu: Mini-cluster; mClu: Micro-cluster

Acknowledgments

The authors thank Qingyi Wang, Yi He, Shiyan Huang, Jinyi Liu, Kang Sun, Jie Wang and Xiangyu Yan for their help at drawing maps.

Authors' contributions

ZT designed and directed this study. DL and SP1 performed the analysis of SSR densities in the human reference Y-DNA. HZ1 performed the comparative analysis of SSR densities in the reference Y-DNA of chimpanzee, rhesus, mouse and rat. ZP and LZ performed the analysis of SSR clusters in the human reference Y-DNA. YF and SP1 developed DCM method and built the SSR landscape maps. FX and YP directed on mathematic calculation of the SSR densities. SP2, HH and RS marked the annotations of the genes, regulator regions and specific genomic structures on the SSR landscape maps of the human reference Y-DNA. HZ2 gave advice for the revised manuscript and language. ZT and DL prepared the manuscript. The authors read and approved the final manuscript.

Funding

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

Availability of data and materials

The accession number of the reference human Y-DNA used in this analysis is NC_000024.10. The annotation and information of protein-coding and ncRNA genes were collected from the Genbank file of NC_000024.10 and the corresponding name is *Homo sapiens Updated Annotation Release 109.20201120* (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/109.20201120/). 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/) and *Ensemble Genome Browser* (www.ensemble.org).

The accession number of samples (alignment file) from 1000 human genome project (https://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/data/) are NA07051 (https://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/data/ERR3239281/) and NA18874 (https://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/data/ERR3243161/).

The accession number of the Y-DNA of chimpanzee, rhesus monkey, mouse and rat used in this analysis is NC_006492.4 (annotation: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Pan_troglodytes/105/), NC_027914.1 (annotation: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Macaca_mulatta/103/), NC_000087.8 (annotation: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Mus_musculus/109/) and NC_024475.1 (annotation: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Rattus_norvegicus/106/).

Tables S1-S5 are available in https://github.com/DooYal/human-Y-supplementary-material/tree/master/Supplementary%20tables

Figs. S1.1-S1.540 are available in https://dooyal.github.io/human_y_ssr_maps/

DCM 2.0 is available in https://github.com/DooYal/DCM

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Bioinformatics Center, College of Biology, Hunan University, Changsha 410082, China. [2]Department of Mathematics, Wilfrid Laurier University, Waterloo, Ontario N2L 3C5, Canada.

Received: 13 July 2020 Accepted: 13 January 2021

Published online: 22 January 2021

References

1. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018;19(5):286–98.
2. Gymrek M, Willems T, Reich D, Erlich Y. Interpreting short tandem repeat variations in humans using mutational constraint. Nat Genet. 2017;49(10):1495–501.
3. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016;48(1):22–9.
4. Mirkin SM. Expandable DNA repeats and human disease. Nature. 2007;447(7147):932–40.
5. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004;5(6):435–45.
6. Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. Unstable tandem repeats in promoters confer transcriptional evolvability. Science. 2009;324(5931):1213–6.
7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
8. Fondon JW, Garner HR. Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A. 2004;101(52):18058–63.
9. Yap K, Mukhina S, Zhang G, Tan JSC, Ong HS, Makeyev EV. A short tandem repeat-enriched RNA assembles a nuclear compartment to control alternative splicing and promote cell survival. Mol Cell. 2018;72(3):525–40.
10. Kumar RP, Krishnan J, Pratap Singh N, Singh L, Mishra RK. GATA simple sequence repeats function as enhancer blocker boundaries. Nat Commun. 2013;4:1844.
11. Santos-Pereira JM, Aguilera A. R loops: new modulators of genome dynamics and function. Nat Rev Genet. 2015;16(10):583–97.
12. Kita E, Katsui N, Emoto M, Sawaki M, Oku D, Nishikawa F, et al. Virulence of transparent and opaque colony types of Neisseria gonorrhoeae for the genital tract of mice. J Med Microbiol. 1991;34(6):355–62.
13. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017;14(6):590–2.
14. Mandal R, Samstein RM, Lee KW, Havel JJ, Wang H, Krishna C, et al. CANCER genetic diversity of tumors with mismatch repair deficiency influences anti-PD-1 immunotherapy response. Science. 2019;364(6439):485–91.
15. Chan EM, Shibue T, McFarland JM, Gaeta B, Ghandi M, Dumont N, et al. WRN helicase is a synthetic lethal target in microsatellite unstable cancers. Nature. 2019;568(7753):551–6.
16. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371–85 e318.
17. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol. 2004;21(6):991–1007.
18. Kofler R, Schlotterer C, Luschutzky E, Lelley T. Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. BMC Genomics. 2008;9.
19. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931–45.
20. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006.
21. Bachtrog D. The temporal dynamics of processes underlying Y chromosome degeneration. Genetics. 2008;179(3):1513–25.
22. Bachtrog D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet. 2013;14(2):113–24.
23. Hughes JF, Page DC. The biology and evolution of mammalian Y chromosomes. Annu Rev Genet. 2015;49:507–27.
24. Otto SP, Pannell JR, Peichel CL, Ashman TL, Charlesworth D, Chippindale AK, et al. About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. Trends Genet. 2011;27(9):358–67.
25. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003;423:825.
26. Willems T, Gymrek M, Poznik GD, Tyler-Smith C, Erlich Y. Y GPC: population-scale sequencing data enable precise estimates of Y-STR mutation rates. Am J Hum Genet. 2016;98(5):919–33.
27. Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, et al. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet. 2010;87(3):341–53.
28. Kayser M, Kittler R, Erler A, Hedman M, Lee AC, Mohyuddin A, et al. A comprehensive survey of human Y-chromosomal microsatellites. Am J Hum Genet. 2004;74(6):1183–97.
29. Kayser M. Forensic use of Y-chromosome DNA: a general overview. Hum Genet. 2017;136(5):621–35.
30. Claerhout S, Van der Haegen M, Vangeel L, Larmuseau MHD, Decorte R. A game of hide and seq: identification of parallel Y-STR evolution in deep-rooting pedigrees. Eur J Hum Genet. 2019;27(4):637–46.

31. Karmin M, Saag L, Vicente M, Sayres MAW, Jarve M, Talas UG, et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. 2015;25(4):459–66.
32. Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, et al. A direct characterization of human mutation based on microsatellites. Nat Genet. 2012;44(10):1161–5.
33. Summary List of Y Chromosome STR Loci and Available Fact Sheets [https://strbase.nist.gov//ystr_fact.htm].
34. Zhao XY, Tian YL, Yang RH, Feng HP, Ouyang QJ, Tian Y, et al. Coevolution between simple sequence repeats (SSRs) and virus genome size. BMC Genomics. 2012;13:435.
35. Rajendrakumar P, Biswal AK, Balachandran SM, Srinivasarao K, Sundaram RM. Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. Bioinformatics. 2007;23(1):1–4.
36. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.
37. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet. 2014;46(7):707–13.
38. Borstnik B, Pumpernik D. Tandem repeats in protein coding regions of primate genes. Genome Res. 2002;12(6):909–15.
39. Zerbino DR, Johnson N, Juetteman T, Sheppard D, Wilder SP, Lavidas I, et al. Ensembl regulation resources. Database (Oxford). 2016;2016:bav119.
40. May CA, Shone AC, Kalaydjieva L, Sajantila A, Jeffreys AJ. Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. Nat Genet. 2002;31(3):272–5.
41. Allshire RC, Dempster M, Hastie ND. Human telomeres contain at least three types of G-rich repeat distributed non-randomly. Nucleic Acids Res. 1989;17(12):4611–27.
42. Usdin K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res. 2008;18(7):1011–9.
43. Grunewald TGP, Bernard V, Gilardi-Hebenstreit P, Raynal V, Surdez D, Aynaud MM, et al. Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite. Nat Genet. 2015;47(9):1073–8.
44. Sun JH, Zhou LD, Emerson DJ, Phyo SA, Titus KR, Gong WF, et al. Disease-associated short tandem repeats co-localize with chromatin domain boundaries. Cell. 2018;175(1):224–38.
45. Sinai MIT, Salamon A, Stanleigh N, Goldberg T, Weiss A, Wang YH, et al. AT-dinucleotide rich sequences drive fragile site formation. Nucleic Acids Res. 2019;47(18):9685–95.
46. Duitama J, Zablotskaya A, Gemayel R, Jansen A, Belet S, Vermeesch JR, et al. Large-scale analysis of tandem repeat variability in the human genome. Nucleic Acids Res. 2014;42(9):5728–241.
47. Mudunuri SB, Nagarajaram HA. IMEx: imperfect microsatellite extractor. Bioinformatics. 2007;23(10):1181–7.
48. Tempel S. Using and understanding RepeatMasker. Methods Mol Biol. 2012;859:29–51.
49. Luo MC, Gu YQ, Puiu D, Wang H, Twardziok SO, Deal KR, et al. Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. Nature. 2017;551(7681):498–502.
50. Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, et al. The draft genome of tropical fruit durian (Durio zibethinus). Nat Genet. 2017;49(11):1633–41.
51. Li D, Jiao W, Zhou S, Fu Y, Peng S, Peng Y, et al. Comparative analysis on precise distribution-patterns of microsatellites in HIV-1 with differential statistical method. Gene Reports. 2018;12:141–8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.