A matter of life or death: How microsatellites emerge in and vanish from the human genome

Yogeshwar D. Kelkar, Kristin A. Eckert, Francesca Chiaromonte and Kateryna D. Makova

Genome Research, published online October 12, 2011

Triinu Kõressaar

Seminar in Bioinformatics

TARTU 2011

Microsatellites..

~3% of the human genome

High mutation rates

Popultaion genetics, forensic and association studies

Some of them associated with deseases

The evolution of microsatellites

(a) *Birth* - a locus acquires the number of repeats (a threshold) required for high rates of strand slippage

(b) *Adulthood* – characterized by rapid repeat number alterations due to slippage

(c) *Death* - a locus degrades to a repeat number below threshold, ceasing to sustain high slippage rates

**threshold* - the minimal number of repeats required to constitute a microsatellite

Identification of orthologous microsatellites

Sputnik is used to extract orthologous microsatellites from MULTIZ alignments of human, chimpanzee, orangutan, macaque and marmoset

Only uninterrupted microsatellites (pure microsatellite sequences were extended into flanking sequences to include interruptions whenever possible—the maximum allowed interruption length was equal to the length of the microsatellite motif, and the extension of the microsatellite was required to contain at least one complete repetition of the motif)

Only one repeated motif microsatellites (compound microsatellites were removed from analyzes - adjacent microsatellites separated by at most one non-microsatellite nucleotide joined together)

MULTIZ (Blanchette et al. 2004) – multiple alignment program, can be used even with sequences that are fragmented or have rearrangements such as inversions and duplications.

Sputnik (http://espressosoftware.com/sputnik/) - program that searches DNA sequence files in Fasta format for microsatellite repeats

Identification of orthologous microsatellites

Removed loci that, in any species
(a) had other microsatellite(s) in their 25 bp up- and downstream neighborhood (the central as well as neighborhood loci were removed, to minimize influences among neighboring loci);

(b) possessed nucleotides with phred score <20 within microsatellites or within flanks (10 bp up- and downstream);

(c) had 20 bp up-/downstream low-complexity flanks (i.e., flanks completely lacking one of the four nucleotides or harboring a six-repeat-unit long mononucleotide or four-repeat-unit long dinucleotide repeat);

(d) had flanks' sequence identity <85% in relation to any other species analyzed (to ensure orthology of microsatellites as well as to remove improperly aligned orthologous loci)

Table S1. The numbers of microsatellites remaining after each filtering step implemented.

Motif size	Threshold	Orthologous loci identified	No. of loci remaining after each filtering step		
			Interrupted, compound, & loci closer than n bp	Loci with low sequence quality (PHRED <20)	Loci with low complexity or flanking sequence similarity
1	5	9654415	3520671	439424	222386
	6	6725552	1837882	1349607	343147
	7	3445015	250922	225829	157905
	8	2370977	250866	225779	44451
	9*	1503526	1433786	1306667	127490
	9	1503526	1407654	1296474	21027
	10	92764	45835	39876	33226
2	3	675360	5366270	260297	123069
	4	588474	1234764	1311287	87135
	5*	478955	402763	387654	29290
	5	478955	394756	370989	4497
	6	200741	23456	21110	3283
	7	213983	119826	43348	2627
	8	87265	19746	16191	1133

Identification of microsatellite births/deaths

Human, chimpanzee, orangutan

Orthologous loci, where
at least one of the species possessed a microsatellite
at least one of the other two did not

Ancestral state inferred according to microsatellite presence/absence in the outgroup species

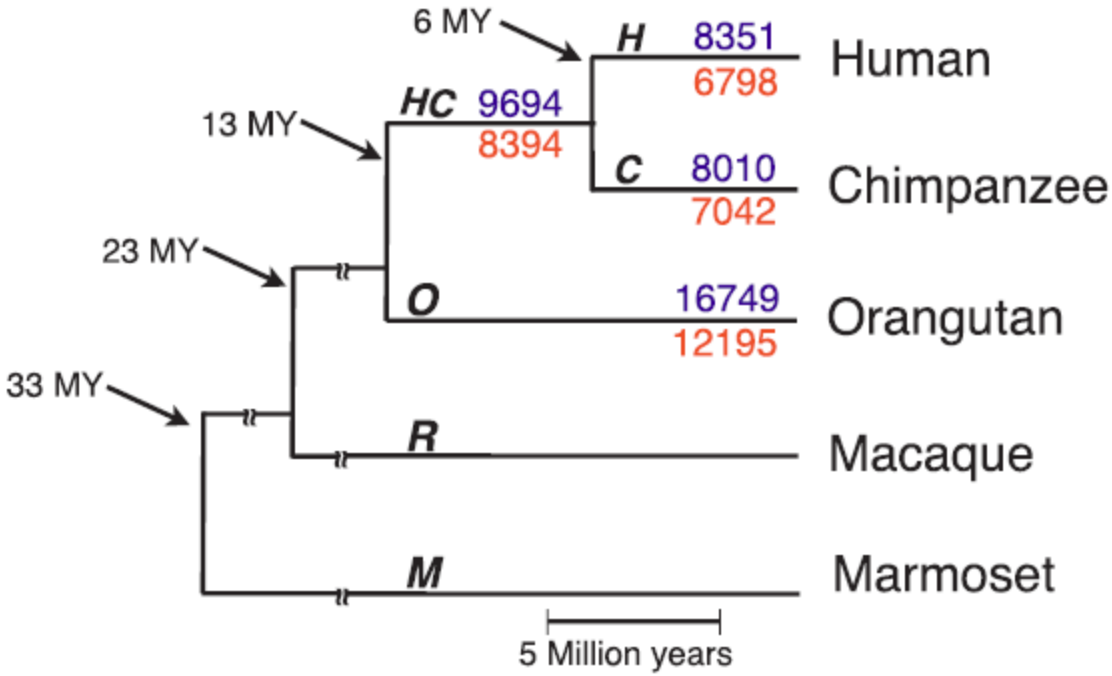


Figure 1. Number of microsatellite births (above) and deaths (below) along the *H*, *C*, *HC*, and *O* branches of the primate tree (thresholds [9,5,4,3]).

Birth-/death-causing mutations

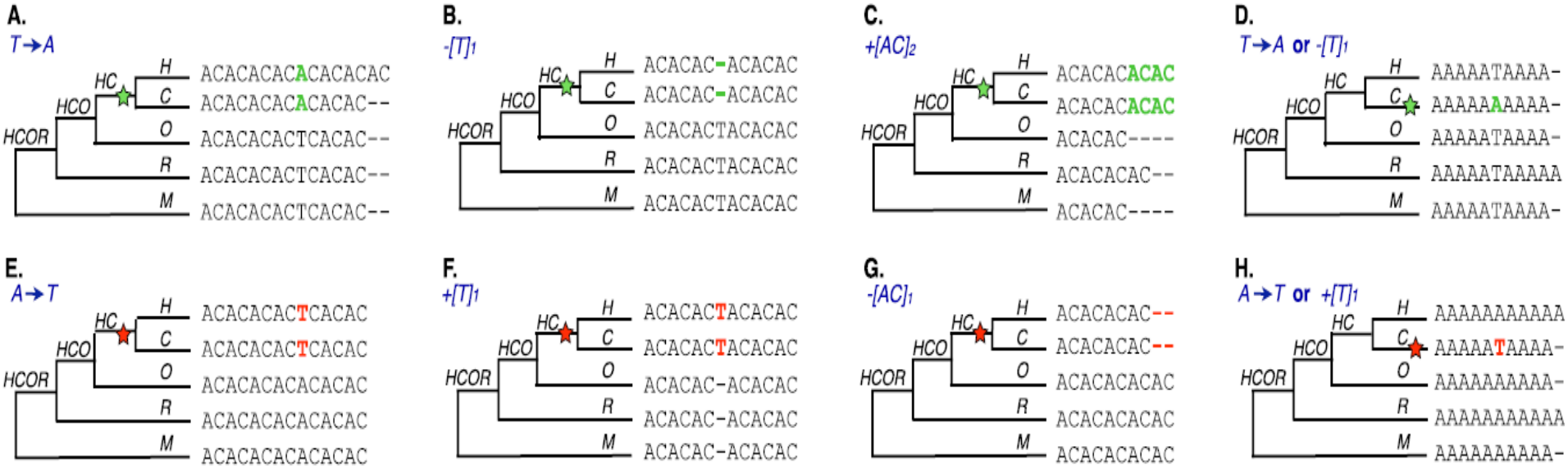


Figure 2. Inference of causal mutation mechanisms for microsatellite birth and death (with thresholds of 5 and 10 repeats for di- and mononucleotide microsatellites, respectively). The lineage experiencing a birth (death) is marked with a green (red) star. (*A*) Birth by substitution (see Mechanisms of birth/death in Supplementary Information). (*B*) Birth by non-motif deletion. (*C*) Birth by motif-insertion. (*D*) Births resulting from either substitutions or non-motif deletions cannot be distinguished for mononucleotides microsatellites. (*E*) Death by substitution. (*F*) Death by non-motif insertion. (*G*) Death by motif-deletion. (*H*) Deaths resulting from either substitutions or non-motif insertions cannot be distinguished for mononucleotide microsatellites.

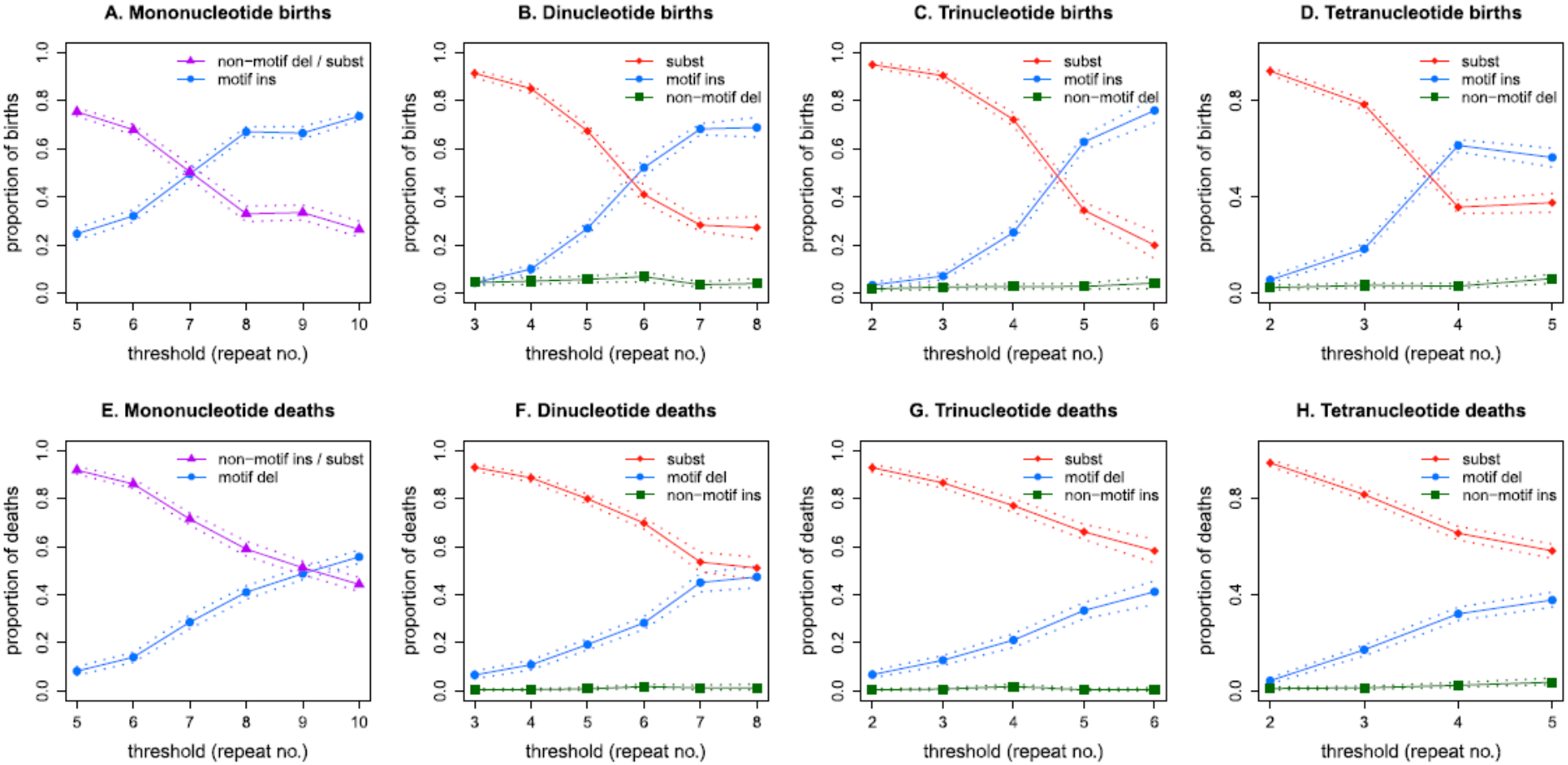
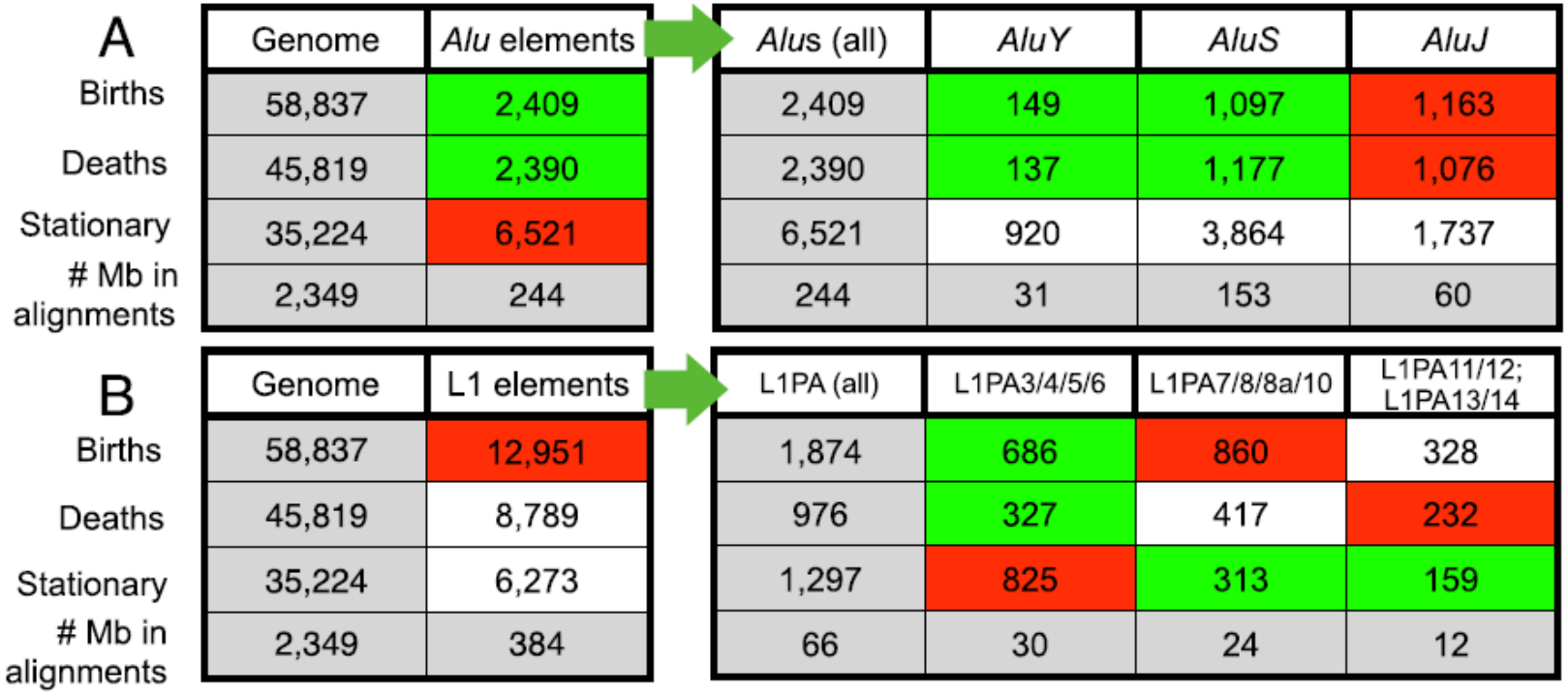


Figure 3. Proportion of various causal mutations as a function of the microsatellite threshold for (*A*) mono-, (*B*) di-, (*C*) tri-, and (*D*) tetranucleotide births; and (*E*) mono-, (*F*) di-, (*G*) tri-, and (*H*) tetranucleotide deaths. Dashed lines indicate 95% bootstrap confidence intervals that were computed for each threshold by re-sampling the genome-wide set of microsatellite loci with replacement.

Births/deaths in interspersed repeats



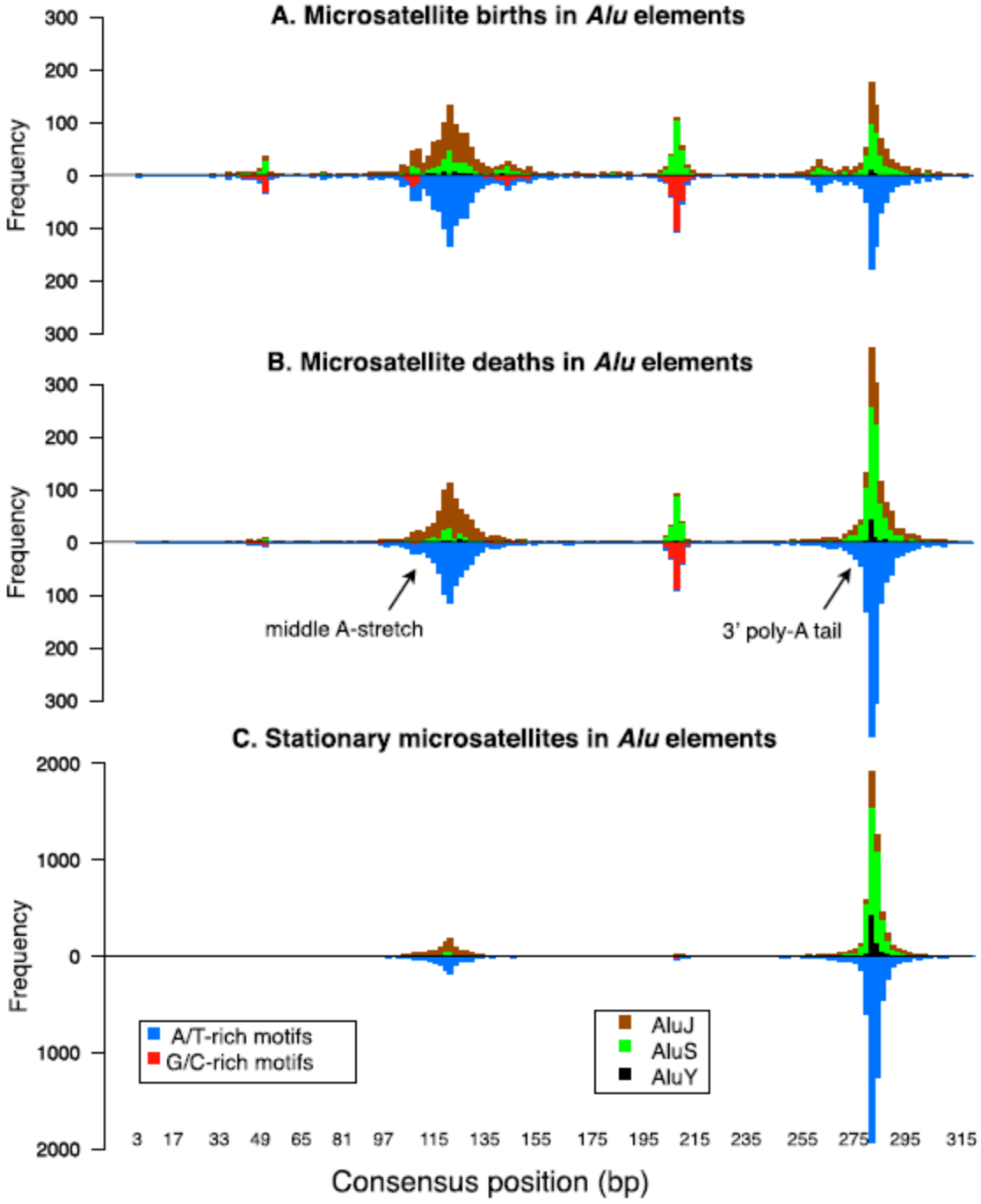
A

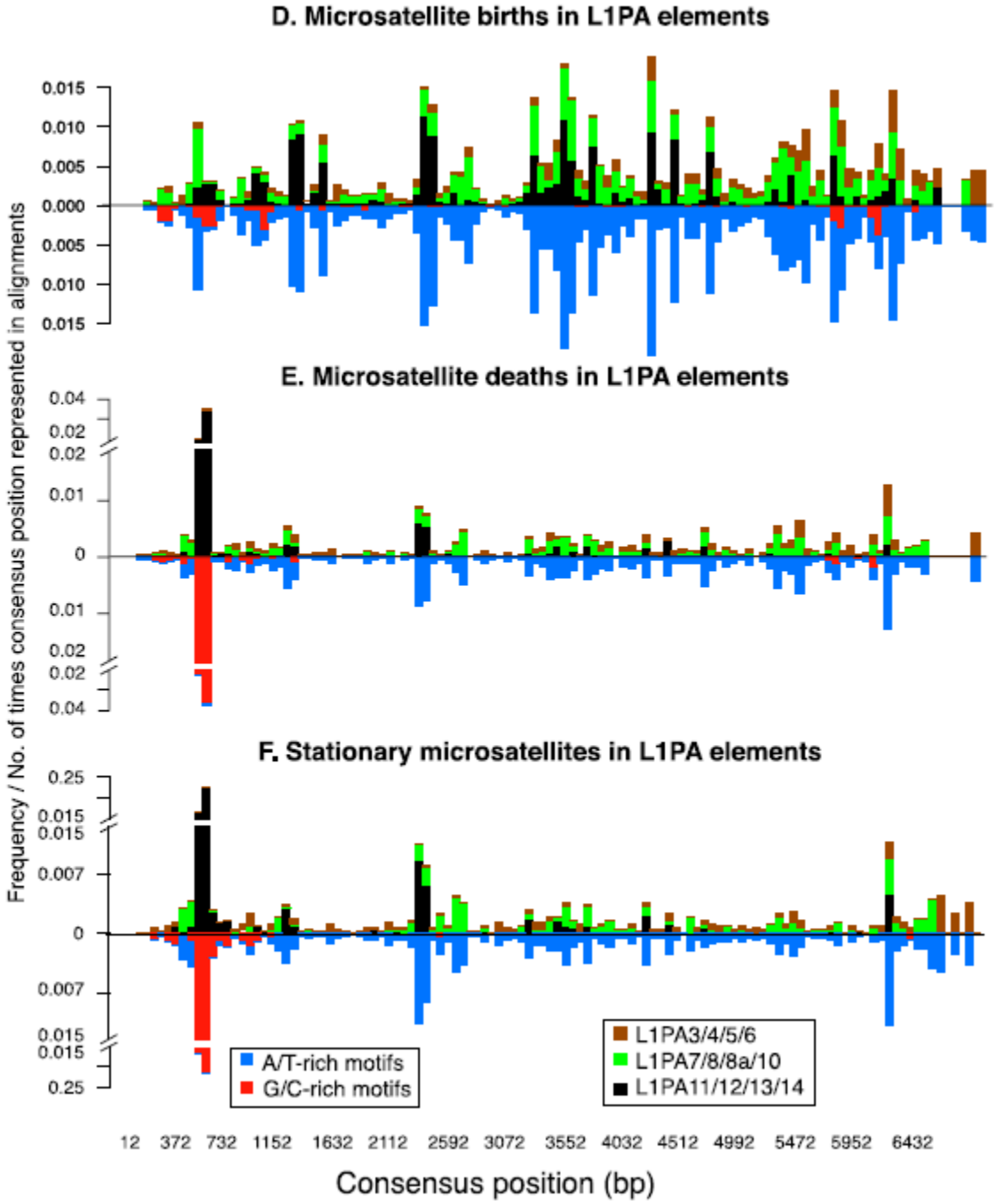
	Genome	*Alu* elements	*Alus* (all)	*AluY*	*AluS*	*AluJ*
Births	58,837	2,409	2,409	149	1,097	1,163
Deaths	45,819	2,390	2,390	137	1,177	1,076
Stationary	35,224	6,521	6,521	920	3,864	1,737
# Mb in alignments	2,349	244	244	31	153	60

B

	Genome	L1 elements	L1PA (all)	L1PA3/4/5/6	L1PA7/8/8a/10	L1PA11/12; L1PA13/14
Births	58,837	12,951	1,874	686	860	328
Deaths	45,819	8,789	976	327	417	232
Stationary	35,224	6,273	1,297	825	313	159
# Mb in alignments	2,349	384	66	30	24	12

Figure 4. Number of microsatellite births, deaths, and stationary loci mapping to (*A*) all *Alus* and different *Alu* subfamilies, and (*B*) all L1s and different L1PA subfamilies (thresholds [9,5,4,3]). Gray cells were used to derive expected counts in χ^2 tests for over- or under-representation of birth/death/stationary loci in all *Alus* and L1s (*left* panels), and in different *Alu* and L1 subfamilies (*right* panels). Loci corresponding to green and red colored cells have, respectively, significant under- and over-representation (*P*-values provided in Supplemental Fig. S10).


A. Microsatellite births in Alu elements
B. Microsatellite deaths in Alu elements
C. Stationary microsatellites in Alu elements
Frequency
300
200
100
0
100
200
300
2000
1000
middle A-stretch
3' poly-A tail
A/T-rich motifs
G/C-rich motifs
AluJ
AluS
AluY
3 17 33 49 65 81 97 115 135 155 175 195 215 235 255 275 295 315
Consensus position (bp)


D. Microsatellite births in L1PA elements
E. Microsatellite deaths in L1PA elements
F. Stationary microsatellites in L1PA elements
Frequency / No. of times consensus position represented in alignments
0.015
0.010
0.005
0.000
0.005
0.010
0.015
0.04
0.02
0.02
0.01
0
0.01
0.02
0.02
0.04
0.25
0.015
0.015
0.007
0
0.007
0.015
0.015
0.25
A/T-rich motifs
G/C-rich motifs
L1PA3/4/5/6
L1PA7/8/8a/10
L1PA11/12/13/14
12
372
732
1152
1632
2112
2592
3072
3552
4032
4512
4992
5472
5952
6432
Consensus position (bp)

Influence of regional genomic features on microsatellite births/deaths

Figure S9.

Feature	Window size					
	0.1 Kb	1 Kb	10 Kb	50 Kb	100 Kb	100 Kb*
GC content	0.31 (-15)	0.31 (-15)	0.37 (-15)	0.27 (-15)	0.19 (-15)	0.12 (-15)
L1 content	0.00	0.10 (-15)	0.08 (-15)	0.00	0.00	0.00
Alu content	0.19 (-15)	0.37 (-13)	0.46 (-15)	0.63 (-15)	0.72 (-15)	0.83 (-15)
Substitution rate	0.05 (-15)	0.06 (-15)	0.09 (-15)	0.06 (-15)	0.08 (-15)	0.04 (-15)
Protomicrosatellite	0.44 (-15)	0.15 (-2)	0.00	0.02 (-6)	0.01 (-2)	0.00
% deviance reduced	**0.07**	**0.02**	**0.01**	**0.03**	**0.05**	**0.13**

Conclusion

Microsatellites arise primarily by substitutions of interrupting nucleotides, and secondarily by slippage induced expansions

Mainly substitutions lead to microsatellites death