The genome-wide determinants of human and chimpanzee microsatellite evolution

Kelkar YD, Tyekucheva S, Chiaromonte F and Makova KD.

Genome Research, 2008 18:30-38

Triinu Kõressaar

Seminar in Bioinformatics

TARTU 2008

What are microsatellites?

-known also as simple/short tandem repeats

-recurring tandemly

- short (motif size 1-6bp)

-undergo rapid length changes - ins/del of repeat units

-mutation rates - 10^-4-10^-2 per locus per generation in humans

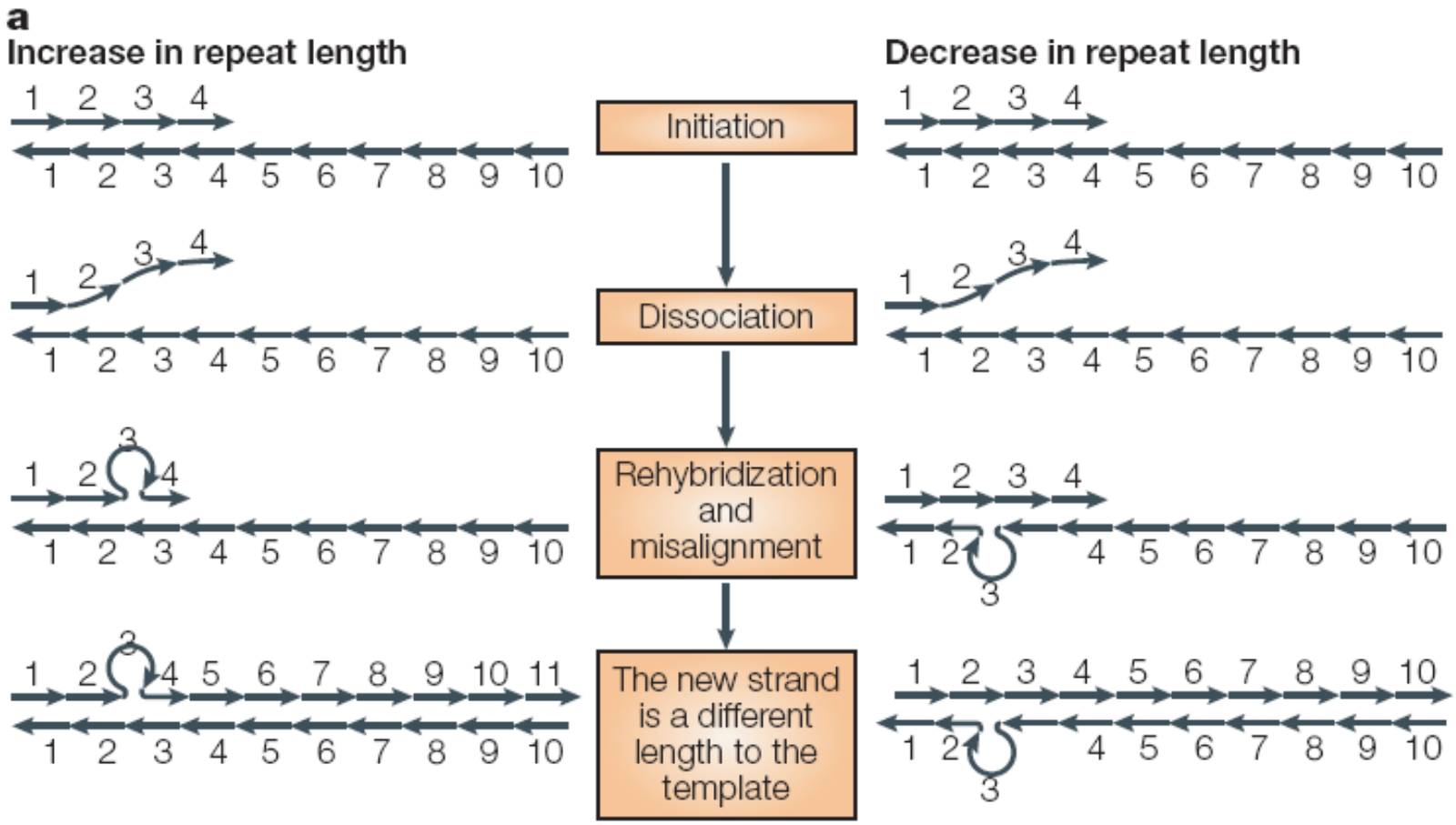
Why are microsatellites important?

The instability of microsatellites is the cause of:

- genesis of cancer
- neurological disorders

Markers for forensic and conservation genetics, genome mapping, population genetic studies etc

Mutation mechanisms: strand slippage



Replication slippage. After the replication of a repeat tract has been initiated, the two strands might dissociate. If the nascent strand then realigns out of register, continued replication will lead to a different length from the template strand. If misalignment introduced a loop on the nascent strand, the end result would be an increase in repeat length. A loop that is formed in the template strand leads to a decrease in repeat length

Mutation rate of microsatellites depends on (1/2):

- the number of repeated units
- length of microsatellite
- length of repeat unit
- the composition on repeated motif
- regionally varying genomic features
  (e.g. local substitution/recombination rates,
  localization respect to *Alu* repeats, GC content)

Mutation rate of microsatellites depends on (2/2):

- transcription
- replication
- localization of microsatellite in sex or autosomes

(Y > autosomes > X)

These features have not been considered **together** or on a **genome-wide scale**

Further...

1. Identification of orthologous microsatellites
2. Effect of:
    - -repeat number, motif size, motif length
    - - motif composition
    - - transcription
    - - location in isochores
    - - location in interspersed repeats
    - - chromosome type

on mutuability

3. Genomic features and microsatellite mutuability

Identification of orthologous microsatellites (1/2)

Motif size 1-4 bp

Uninterrupted microsatellites

For microsatellite identification Sputnik was used:

- minimum score of 4
- mismatch penalty -1000

For orthologous microsatellites BLASTZ alignments were used.

Identification of orthologous microsatellites (2/2)

Filtration:

no orthologous microsatellites in one of the species

low quality sequences in chimpanzee

different repeated motifs at orthologous locations in human and chimpanzee

neighboring microsatellites within 10 bp

<9 repeats for mononucleotide repeats

<4 repeats for di-tri-tetranucleotide repeats

2,107,841 orthologous microsatellite pairs (744,769, 952,382, 97,098, 76,074

mono-, di-, tri- and tetranucleotide intergenic and intronic autosomal microsatellites were used)

Calculating mutability (for groups containing at least 30 microsatellites)

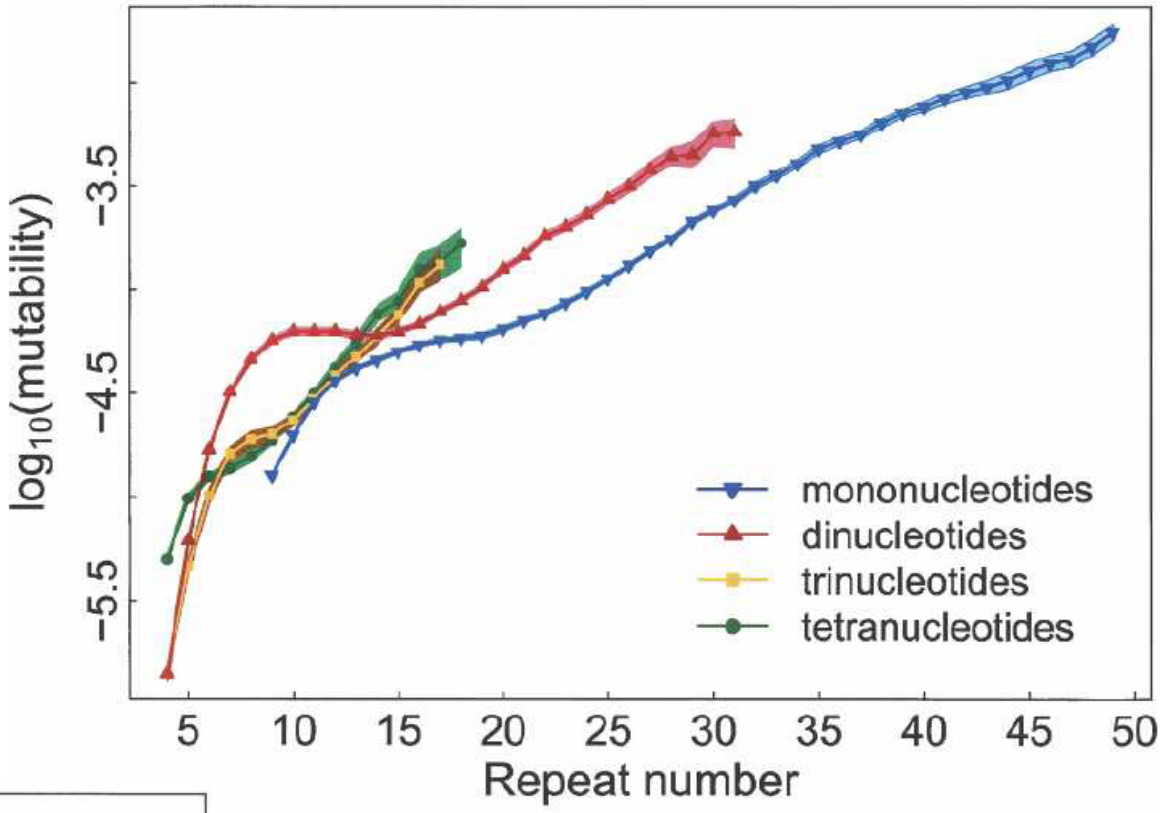
$$Mutability = \frac{\sum_{i=1}^{n_h} h_i (H_i - C_i)^2 + \sum_{i=1}^{n_c} c_i (C_i - H_i)^2}{\sum_{i=1}^{n_h} h_i + \sum_{i=1}^{n_c} c_i}$$

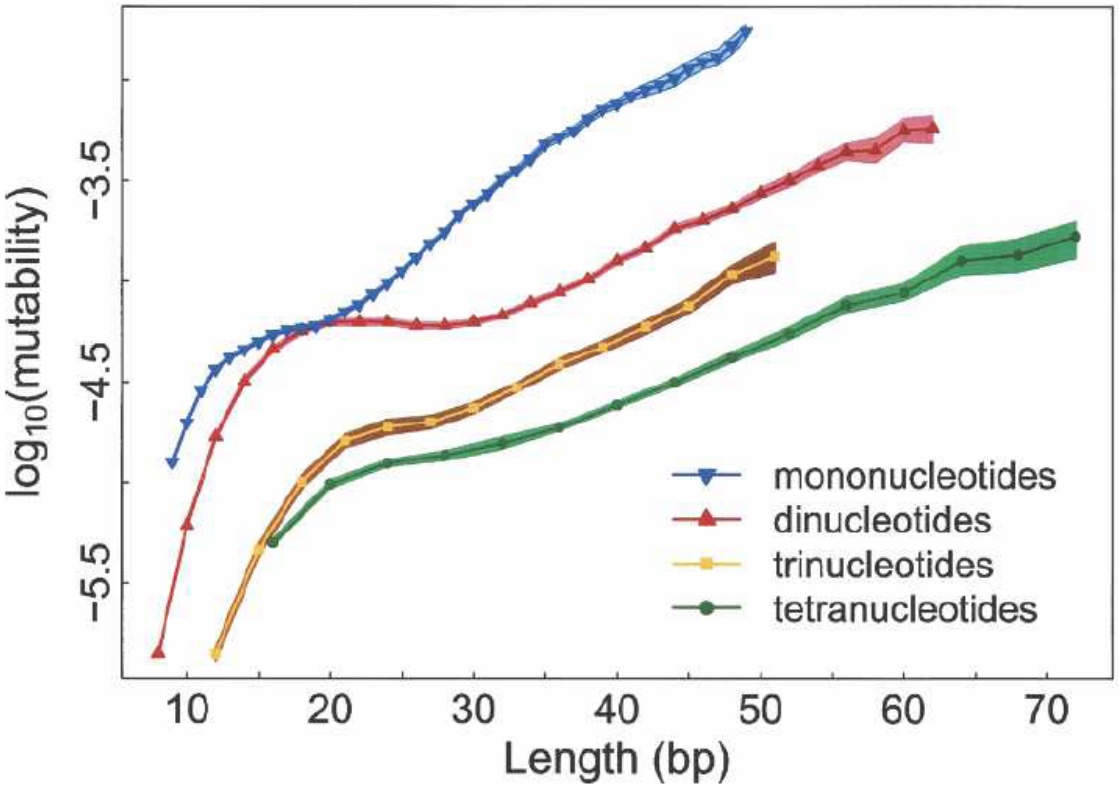
$H_{i(c)}$ human (chimpanzee) repeat number for each of $n_{h(c)}$ orthologous microsatellite pairs sorted according repeat number in human (chimpanzee)

h_i, c_i - correction parameters (=2 if correction for reverse mutation is required when human (chimpanzee) genome is considered ancestral,

e.g. human is ancestral, $C_i > H_i$ and $(H_i - (C_i - H_i)) <$ threshold)

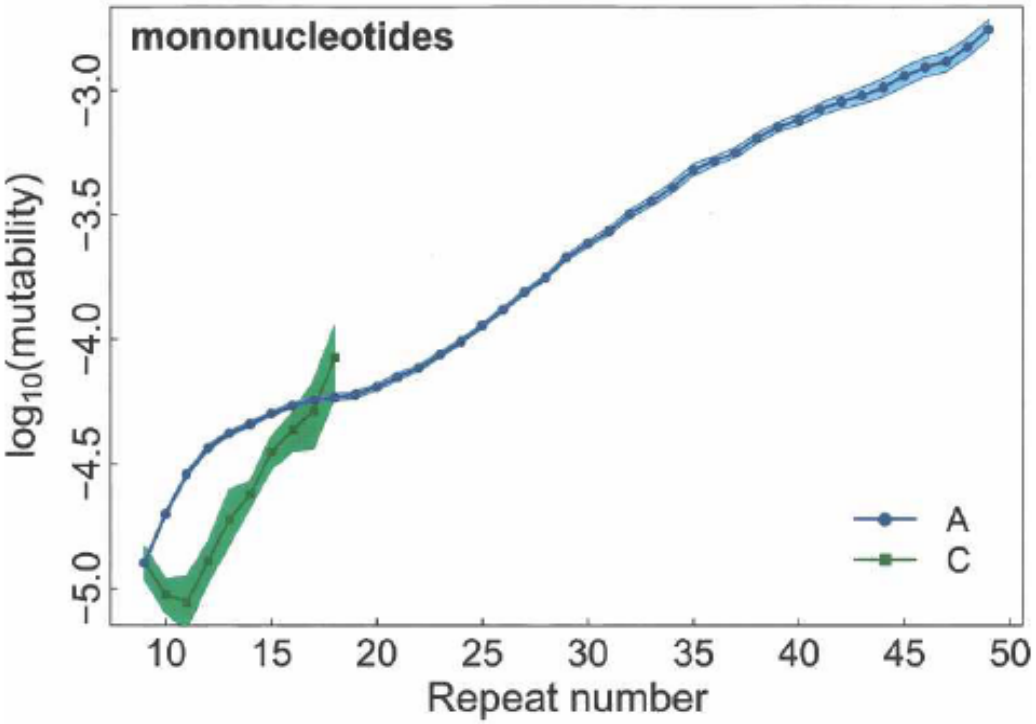
Effects of repeat number and motif size on mutability



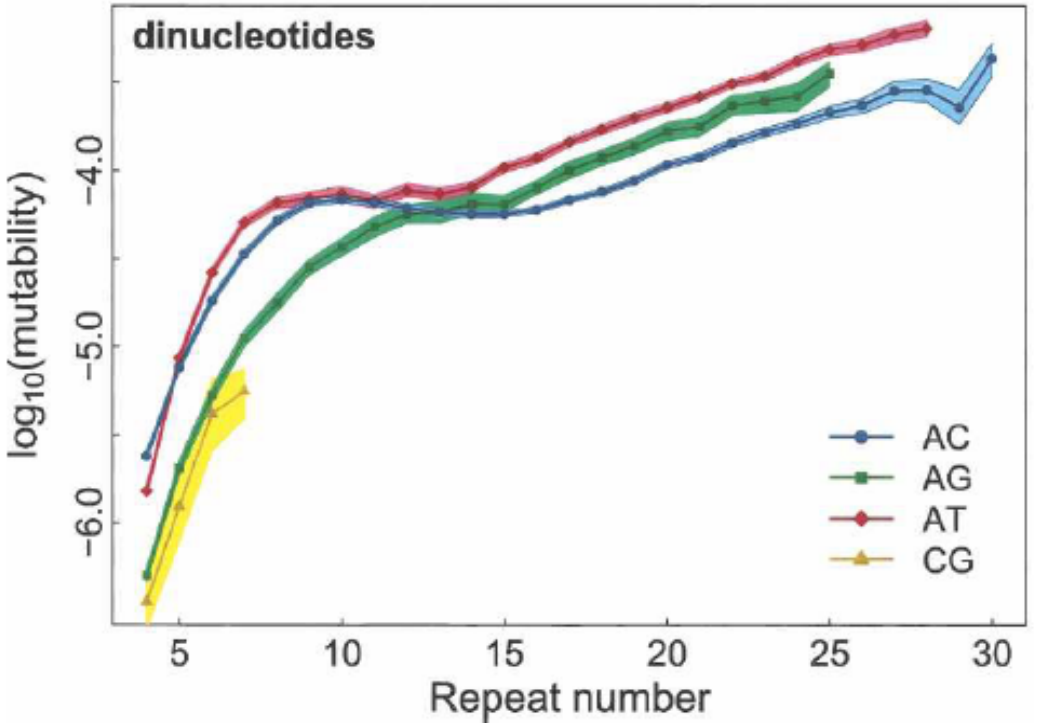


Effects of motif size and length on mutability

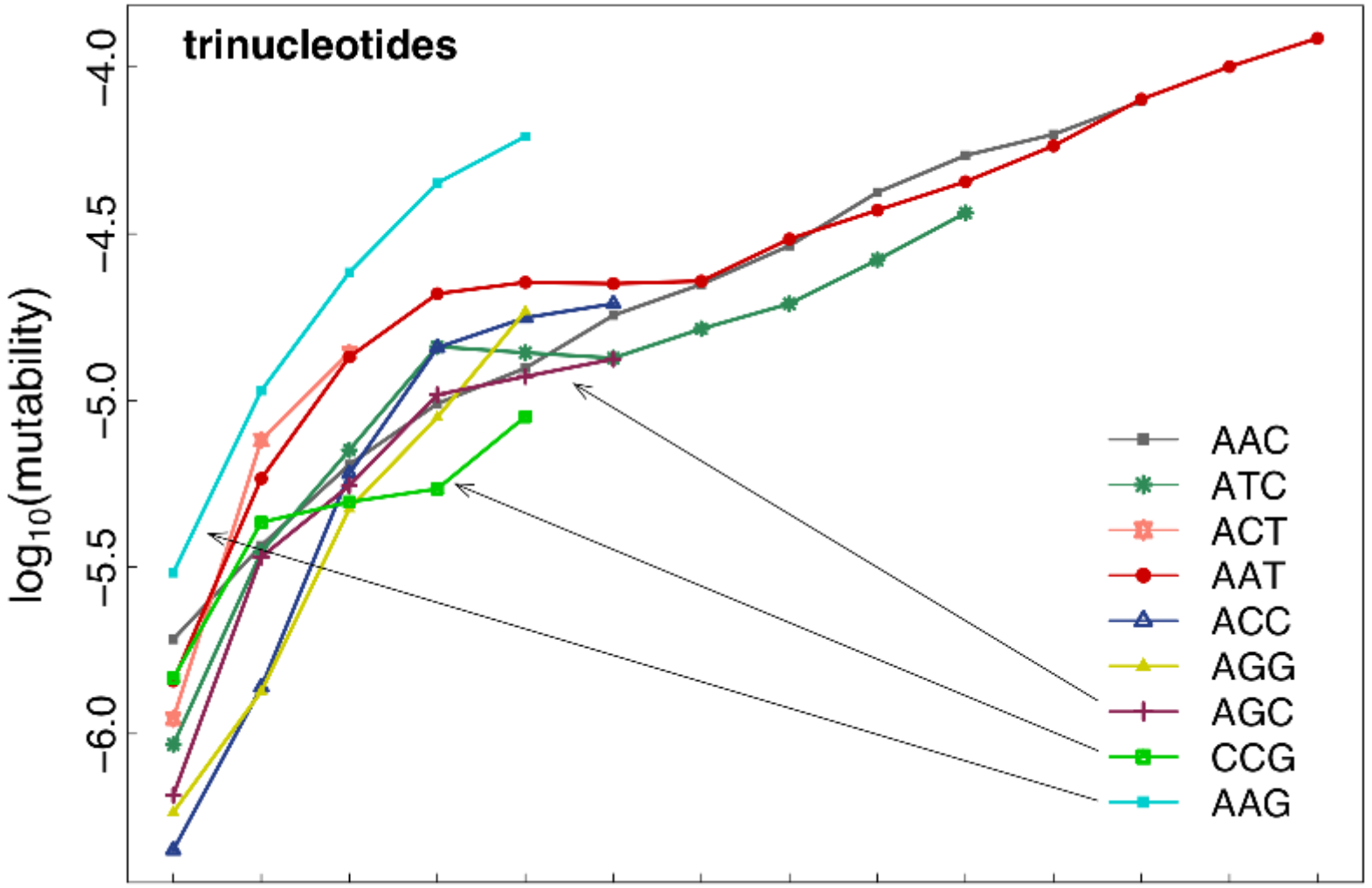
* The bands around the curves indicae the 2.5th and 97.5th percentiles of empirical distributions obtained through a resampling procedure



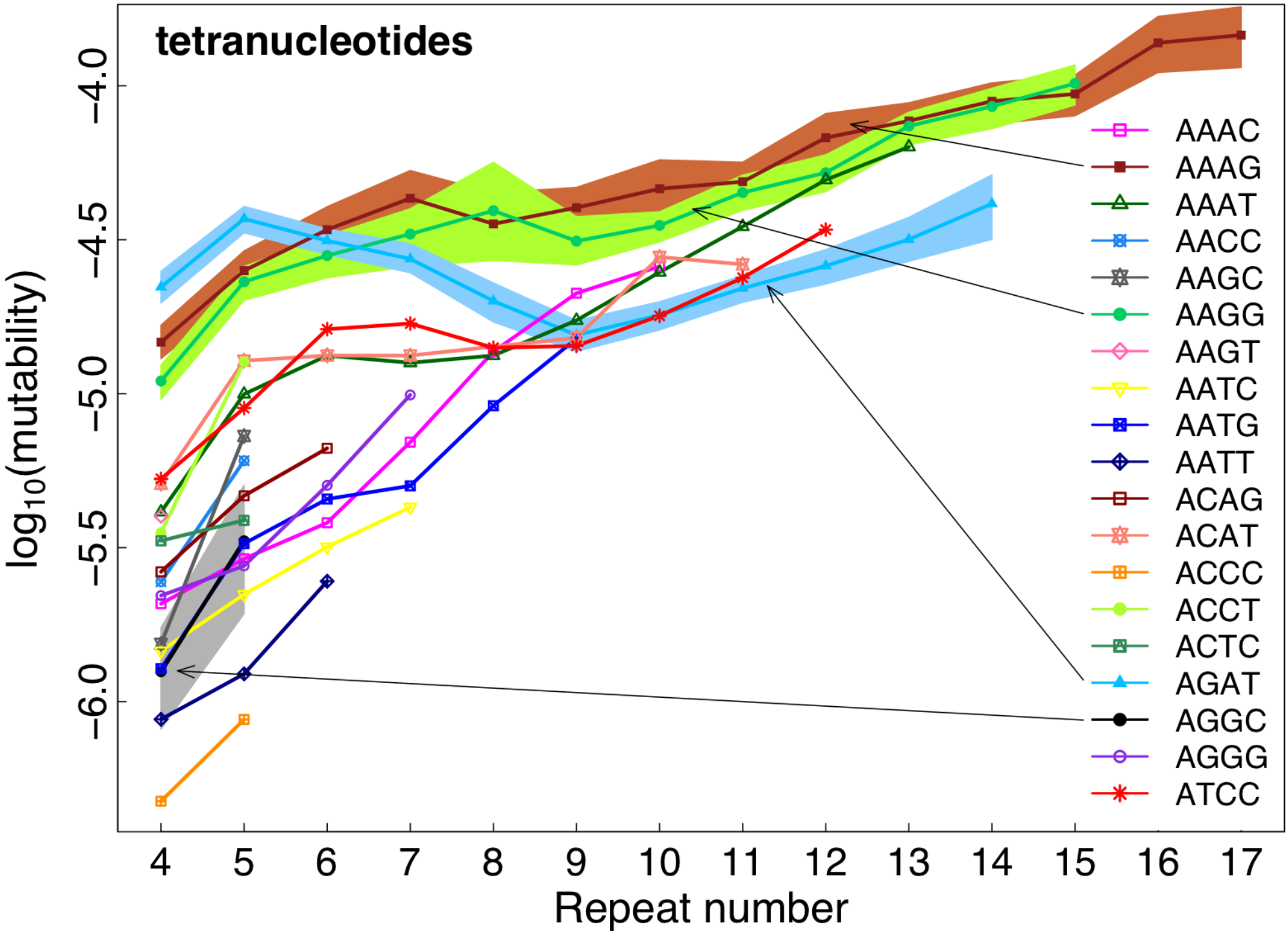
Effect of motif composition on mutability



Effect of motif composition on mutability



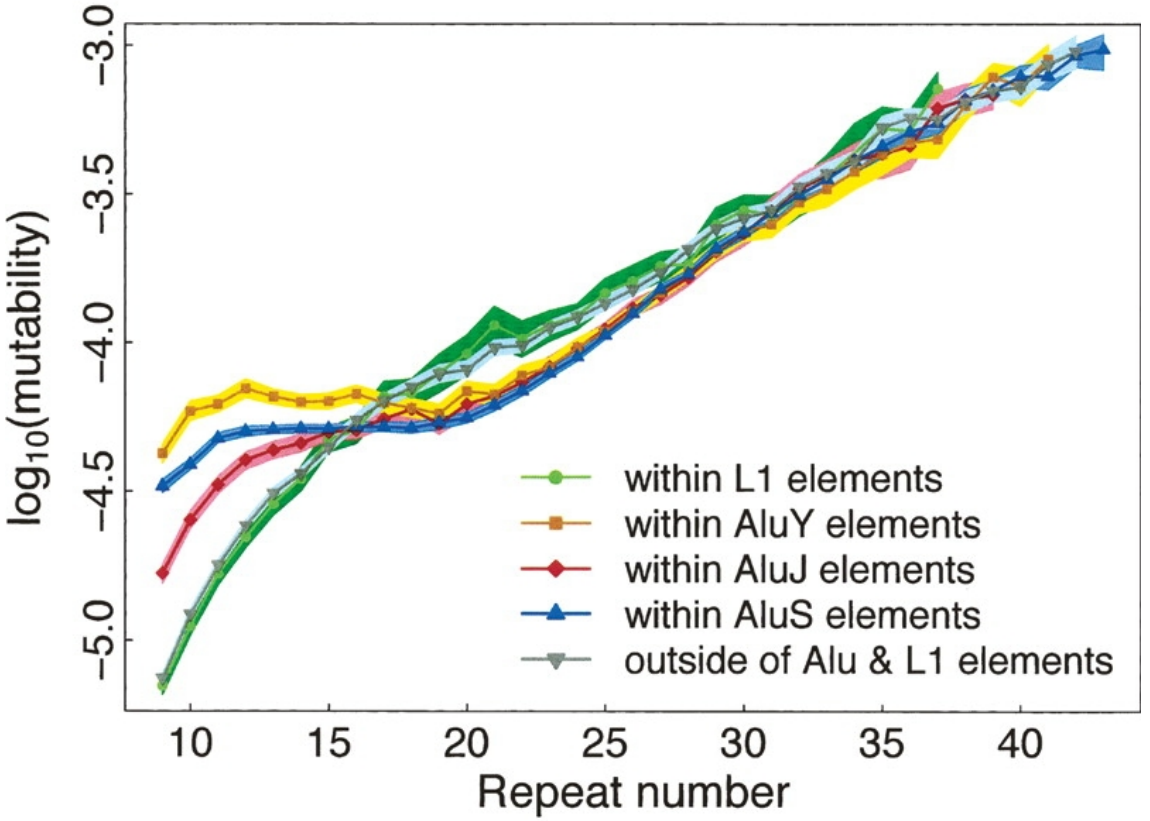
Effect of motif composition on mutability



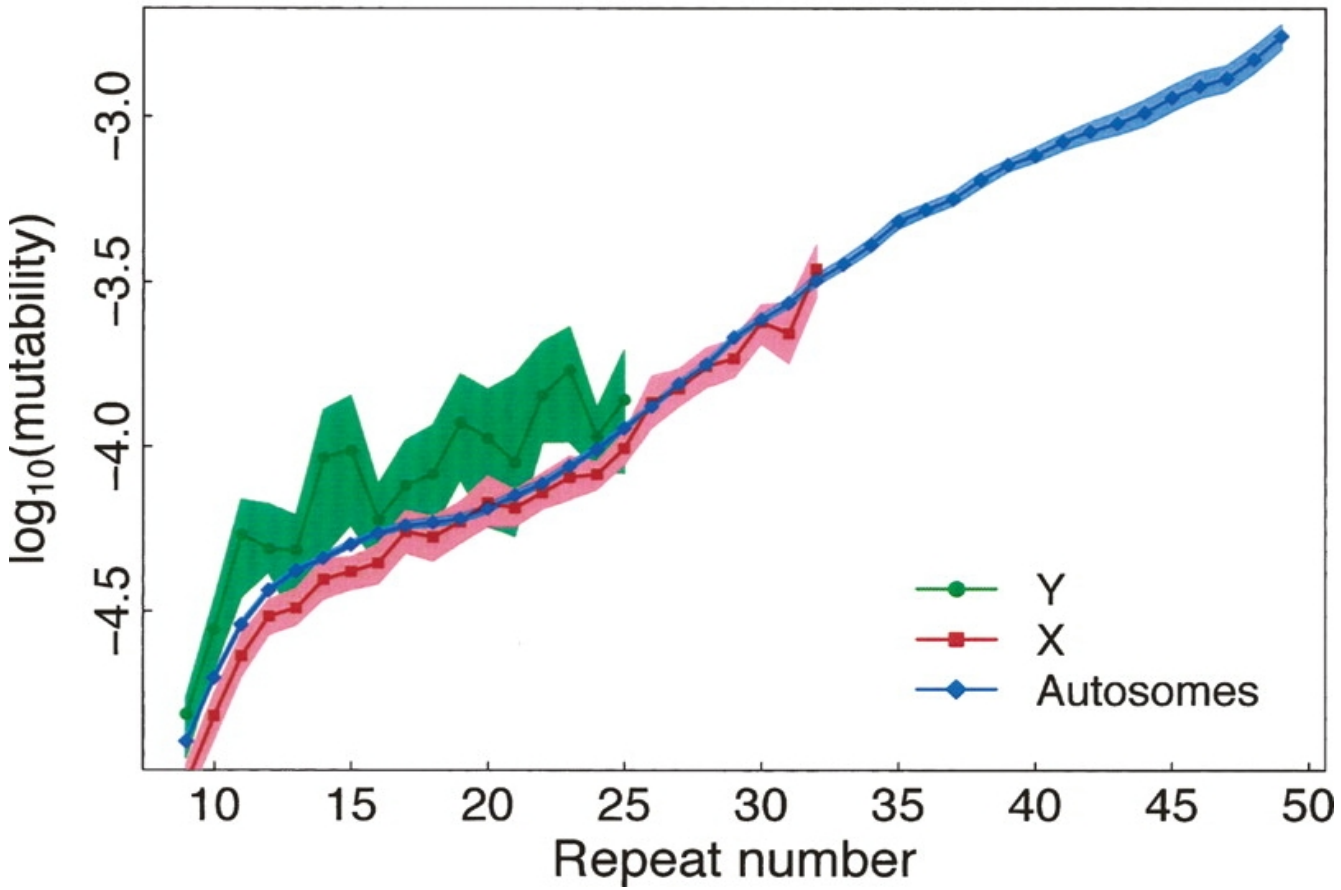
Effects of transcriptions, location in different isochores, and interspersed repeats on mutuality

Mutability does not differ significantly between:

- untranscribed and transcribed (intronic) microsatellites
- different isochores



Effects of chromosome type on mutability: mononucleotides



The male-to-female mutation rate ratio is equal 2.37, 2.03 and 2.31 for the Y/A, X/A and Y/X comparisions, respectively.

Regression analyses with R: best subset model-building technique

Quantitative predictors - repeat number, motif size, repeat length

Categorical predictor - chromosome type

Feature	Calculation
GC content	fraction of bases per window
Exon content	fraction of bases per window
Alu content	fraction of bases per window
L1 content	fraction of bases per window
Distance from telomere	distance of central base of window from nearest telomere
Human-macaque divergence	estimated using REV in ancestral repeats
Computational recombination rate	from Myers et al. (2005)
X-chromosome / autosome indicator variable	“0” for autosomes, “1” for X chromosome

Calculated in 5Mb and 1Mb windows based on human annotations

For each predictor

Relative **C**ontribution to **V**ariability **E**xplained

was calculated

$$RCVE = \frac{R^2_{full} - R^2_{reduced}}{R^2_{full}}$$

Table 1. Multiple regression models for log mutability (per locus per generation) of genomewide and local microsatellite groups.

Regressions	Genome-wide regressions					Regressions for local groups					
Microsatellites (window)/Feature	All	Mono-nucleotides	Di-nucleotides	Tri-nucleotides	Tetra-nucleotides	All (5 Mb)	Mono-nucleotides (5 Mb)[a]	Di-nucleotides (5 Mb)	Tri-nucleotides (5 Mb)	Tetra-nucleotides (5 Mb)	Mono-nucleotides (1 Mb)
Repeat number	0.03 (-12)	0.73 (-15)	0.77[b] (-15)	0.97[b] (-8)	0.35[b] (-5)	0.13[b] (-15)	0.94[b] (-15)	0.62[b] (-15)	0.97 (-15)	1 (-16)	0.43[b] (-15)
Length	0.07 (-15)	-	-	-	-	0.03 (-15)	-	-	-	-	-
Motif size	0.04 (-13)	-	-	-	-	<0.01 (-15)	-	-	-	-	-
Chromosome type[c]	<0.01 (-4)	0.02 (-9)	n. s.[d]	n. s.	n. s.	n. s.	<0.01 (-4)	n. s.	<0.01 (-2)	n. s.	0.02 (-15)
Motif composition[c]	-	0.01 (-8)	0.08 (-5)	0.12 (-10)	0.22 (-12)	-	-	0.26 (-15)	-	-	-
R^2 (predictors above)	**0.908**	**0.973**	**0.877**	**0.844**	**0.794**	**0.893**	**0.706**	**0.688**	**0.452**	**0.296**	**0.399**
GC content	-	-	-	-	-	<0.01 (-7)	n. s.	<0.01 (-13)	n. s.	n. s.	n. s.
Substitution rate	-	-	-	-	-	<0.01 (-5)	<0.01 (-6)	n. s.	n. s.	n. s.	n. s.
Distance from telomere	-	-	-	-	-	n. s.	<0.01 (-2)	<0.01 (-2)	n. s.	n. s.	n. s.
Recombination rate	-	-	-	-	-	<0.01 (-10)	<0.01 (-4)	0.01 (-2)	0.02 (-4)	n. s.	0.02 (-15)
Alu content	-	-	-	-	-	n. s.	n. s.	n. s.	n. s.	n. s.	0.09 (-15)
L1 content	-	-	-	-	-	<0.01 (-4)	< 0.01 (-15)	n. s.	n. s.	n. s.	n. s.
R^2	-	-	-	-	-	**0.894**	**0.715**	**0.693**	**0.462**	**0.296**	**0.450**

For each predictor, the relative contribution to the variability explained (RCVE; see Methods) is indicated and the significance ($\log_{10}$ *P*-value with Bonferroni correction for multiple tests applied) is given in parentheses. For significant quantitative predictors, red indicates a positive effect on mutability; blue, a negative effect. For each model, the multiple R_2 is indicated (adjusted R^2's were almost identical to multiple R^2).

[a]Only (A)n microsatellites were used.

[b]Repeat number was used in conjuction with its square root, the lower between the two $\log_{10}$*P*-values is provided.

[c]Categorical variable, the lower $\log_{10}$ *P*-value is provided.

[d]n.s. = not significant

Conclusions

Repeat number, motif size and repeat length determine most of the interlocus variation in microsatellite mutability (>90%)

Replication slippage is the predominant mechanism of mutagenesis

The effect of local genomic features on microsatellite mutability have to be re-evaluated at smaller scales

The regression models can be used to answer the questions:

- which disease-causing microsatellites are likely to have high rates of de novo mutations?

- which microsatellites are the most suitable for forensic applications?

- which are the polymorphic microsatellites suitable for population and conservation genetic studies?