Core variability in mutation rates shape the basal sequence characteristics of the human genome

Aleksandr B. Sahakyan[1,*] & Shankar Balasubramanian[1,2,3,*]

[1]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK.
[2]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK.
[3]School of Clinical Medicine, University of Cambridge, Cambridge CB2 0SP, UK.

*Correspondence to as952@cam.ac.uk (A.B.S.) and sb10031@cam.ac.uk (S.B.)

SUPPLEMENTARY INFORMATION

CONTENT

Table S1 --- ii
Figure S1 --- iii
Figure S2 --- iv
Figure S3 --- v
Figure S4 --- vi
Figure S5 --- vii
Figure S6 --- viii
Figure S7 --- ix
Figure S8 --- x
Figure S9 --- xi
Figure S10 --- xii
Figure S11 --- xiii
Description of the Supplementary Video and Data Files --- xiv
Supplementary References --- xv

L1 type	N^{hg}	age, myr
L1Hs	1528	3.1
L1PA2	4867	7.6
L1PA3	10565	12.5
L1PA4	11763	18.0
L1PA5	11171	20.4

Table S1. Number of genomic copies and estimated age of hominoid L1 subfamilies. The numbers of L1 mobile elements (N^{hg}) in the human genome were revealed through the RepeatMasker[1] processing of the genome. The divergence age estimation was obtained from the published molecular clock analysis[2].

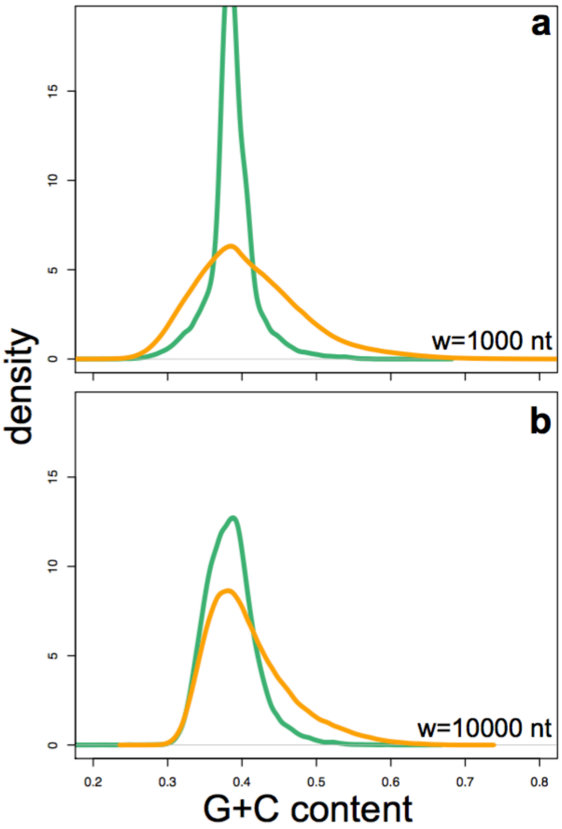


Figure S1. Long-range G+C context of the L1 insertion sites in the human genome. The distribution of the G+C contents for all the w-sized (1000-nt in **a** and 10000-nt in **b**) bins in the human genome (orange lines) is shown, as compared to the same distribution but using only the bins centred at the midpoints of all the remnants of young L1 elements (L1Hs, L1PA2, L1PA3, L1PA4, L1PA5).

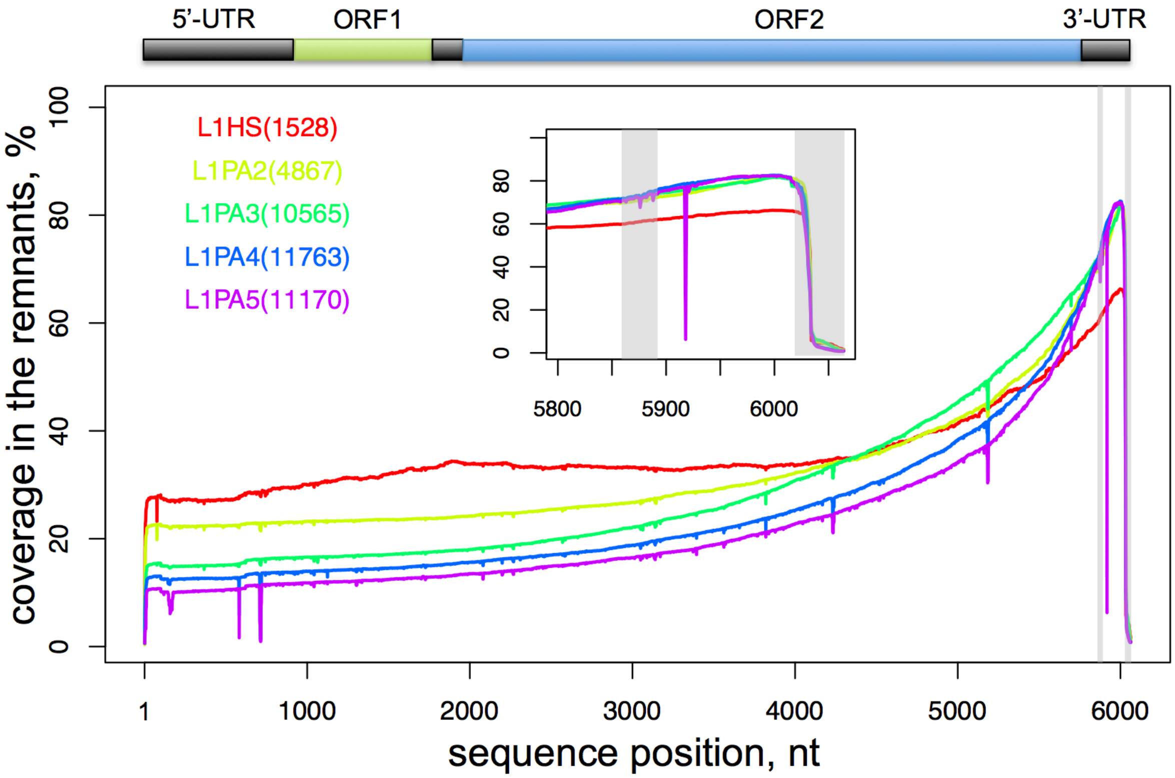
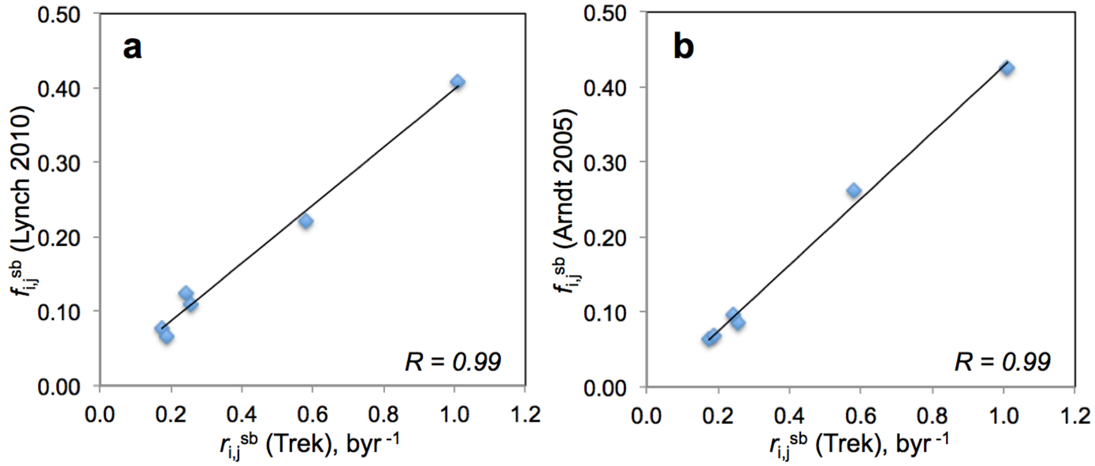


Figure S2. The position-wise coverage of the five subfamilies of L1 remnants in the human genome. All retrotransposon types, which are not too ancient for the time-accumulated mutations to severely decrease the information content, were pairwise aligned on the 6064-nt consensus sequence of the human-specific L1Hs subfamily (reference sequence). The graph shows the percentage of cases where, for each considered L1 subfamily, the remnant sequences were mapped onto the corresponding position (x-axis). The gene organisation in these L1 elements is displayed on top, highlighting the terminal untranslated regions, along with two open reading frames (ORF1 and ORF2) and the short inter-ORF region. The positions 5856-5895 and 6018-6064, close to the 3'-end (also zoomed in the sub-plot) that engulf the low-complexity G-rich and A-rich sequences are marked with grey bands and excluded from the mutation rate analyses. The colour coding of the examined hominoid L1 retrotransposons, along with the genomic copy numbers in the human genome, is shown on the plot. The abrupt drops in the coverage at different positions along the sequence are because of deletions. The figure highlights the incomplete 5'-reverse transcription, characteristic to LINE elements. The poly-A tail at the 3'-end is also largely incomplete in most L1 remnants. It is interesting to note that the further back in time we go in terms of the activity period of the L1 subfamily, the less preserved the 5'-side of the L1 elements become relative to the 3'-end. An overall decrease of the preservation, hence coverage, is expected for the more ancient subfamilies, but the observed decrease relative to the 3'-end can indicate that the more recent subfamilies evolved a more efficient retrotransposition and/or improved transposonic RNA stability that results in more complete insertions.



c

	Trek (this work)	*Lynch (2010)*	*Arndt (2005)*
Mutation type	$r_{i,j}^{sb}$, byr^{-1}	$f_{i,j}^{sb}$	$f_{i,j}^{sb}$
`A:T` to `C:G`	0.175	0.078	0.064
`A:T` to `G:C`	0.582	0.221	0.262
`A:T` to `T:A`	0.188	0.067	0.068
`G:C` to `T:A`	0.256	0.110	0.085
`G:C` to `C:G`	0.243	0.125	0.095
`G:C` to `A:T`	1.011	0.408	0.425

Figure S3. Demonstration of the unbiased averaging of the Trek mutation rates. The median values of the Trek-reported $r_{i,j}^{core} = r_{i,j}^{sb} + \delta r_{i,j}^{sr}$ mutation rates, which should be reasonable estimates of single-base genomic average $r_{i,j}^{sb}$ rates (in time domain), are compared (**a**, **b**) with two published datasets[3,4] for $r_{i,j}^{sb}$, expressed by normalised mutation fractions. Since both datasets report on strand-symmetry-accounted six unique mutation rates, we have performed the same strand-symmetry averaging of the median values shown in **Fig. 2** before the comparison. The numerical data are presented in **c**. The Pearson's correlation coefficients are shown on the plots **a** and **b**.

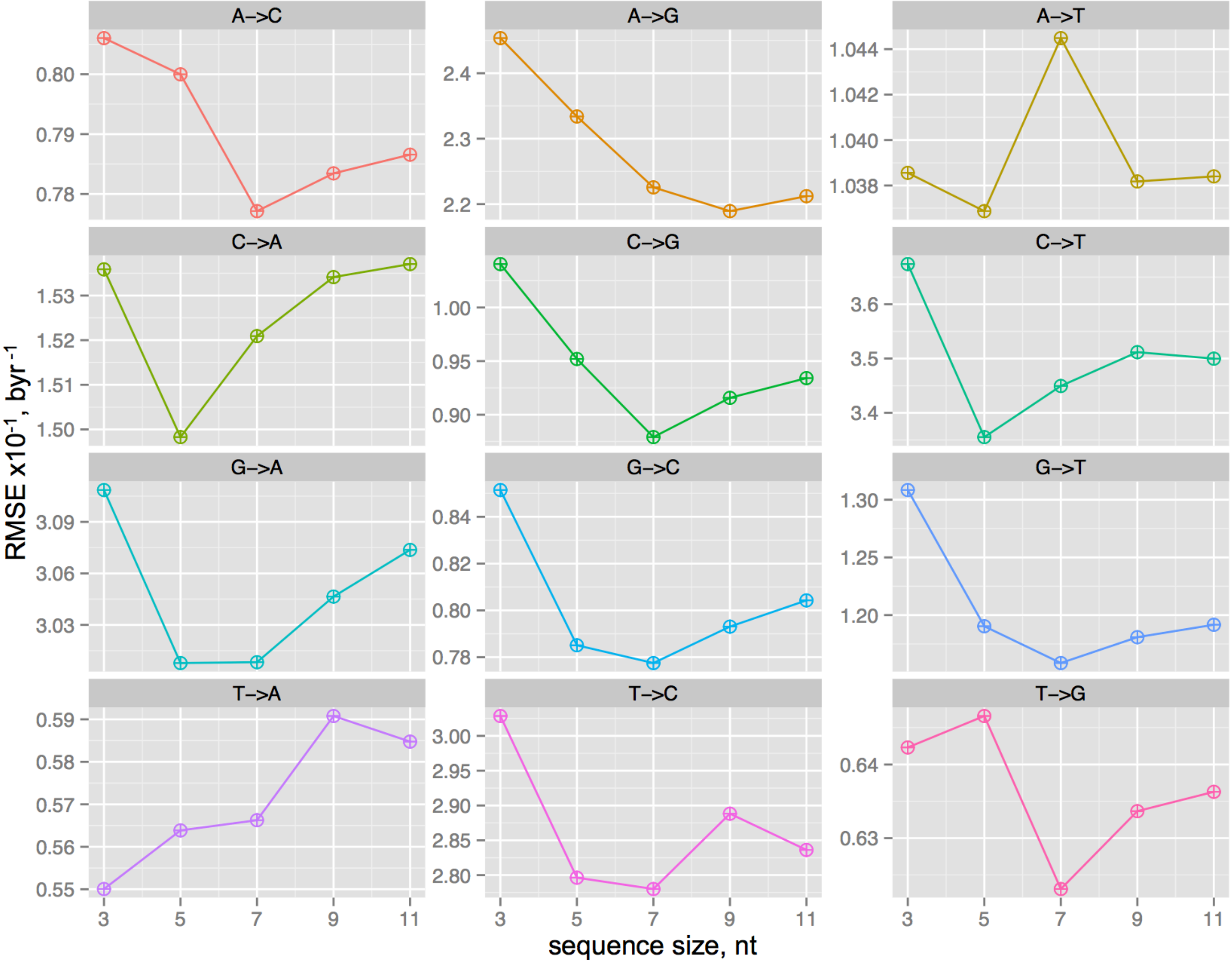


Figure S4. Selection of the optimal short-range sequence length directly influencing the mutation rates. The root-mean-squared errors of the rate constant prediction from only the neighbouring base information is presented as a function of the accounted sequence length, where the mutations occur at the central base. The examined lengths are thus odd numbers, to allow equal number of upstream and downstream bases around the mutation point. The test models for the mutation rates to assess the optimal sequence length were built using the machine learning, generalised boosted models. The 5- and 7-nt sequence length are found to be the best for general applicability for all the mutation types. Note that the machine learning predictions here are solely for finding out the optimal sequence length and, for the actual $r_{i,j}^{\text{core}}$ determination, a direct model-free mapping to the Trek mutation database is used.

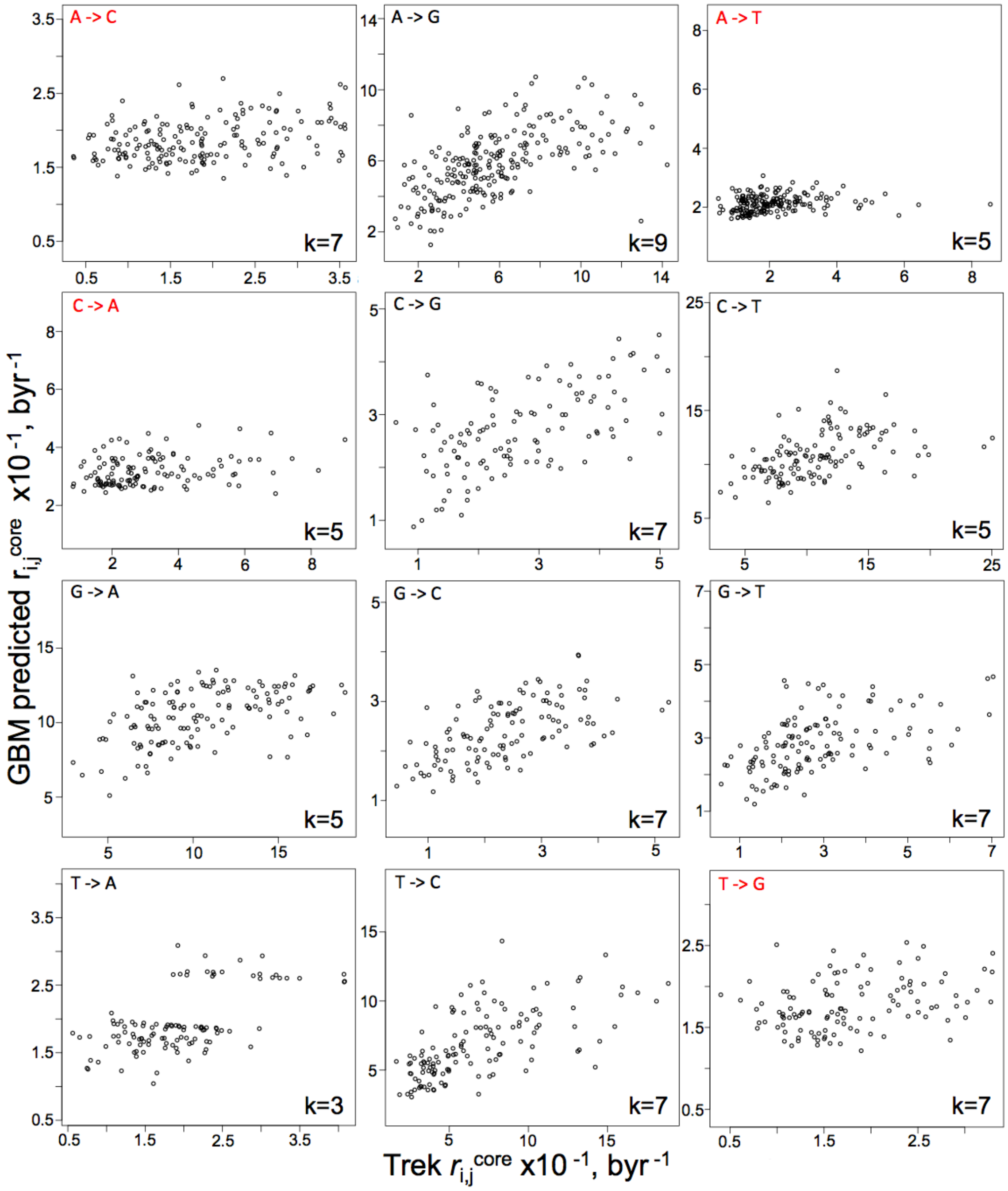


Figure S5. Performance of the optimal GBM models in our test constructs. Machine learning models were built to predict the $r_{i,j}^{core}$ constants using only the knowledge of the neighbouring bases. The resulting highest performing lengths of the sequences (k-mers) are shown at the bottom right corners of each plot. The best models identified for all *i->j* substitution types are presented here as an example of the predictability of the mutation rates from the neighbouring residues. We have used these analyses to infer the optimal sequence (k-mer) length, found to be around 7 nt. We then utilised the found length for all the mutation types as a factor to stratify the Trek $r_{i,j}^{core}$ mutation rate constants for the direct mapping to any given sequence.

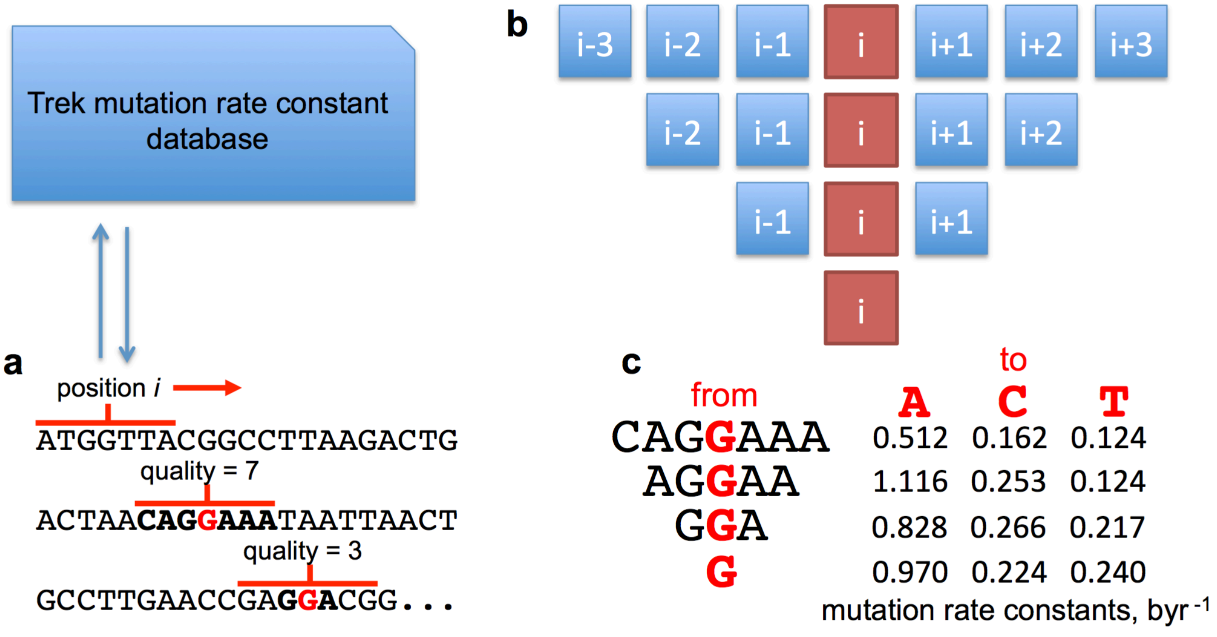


Figure S6. The procedure of mapping the $r_{i,j}^{core}$ constants onto any sequence. A query sequence (**a**) is analysed by examining each base via a 7-nt window centred at the position of the base. To reveal the rate constants for the three mutations of the central base into the three other bases, the Trek database is searched. The Trek data contain information on the full set of unique k-mers (k = 1, 3, 5 and 7) found in the reference L1 element, along with their respective three mutation rate constants. The Trek data are averaged where multiple values are found because of multiple k-mer copies in the same L1 reference sequence. In case a representative match is not found with the full 7-nt long sequence, the window around the given position in a query sequence is shortened (**b**) into the longest variant (5-nt, 3-nt or 1-nt) with a match detected in the Trek database. **c** demonstrates a case for the average mutation rate constants of the base G mutating to A, C or T, with different extent of context information in the Trek data.

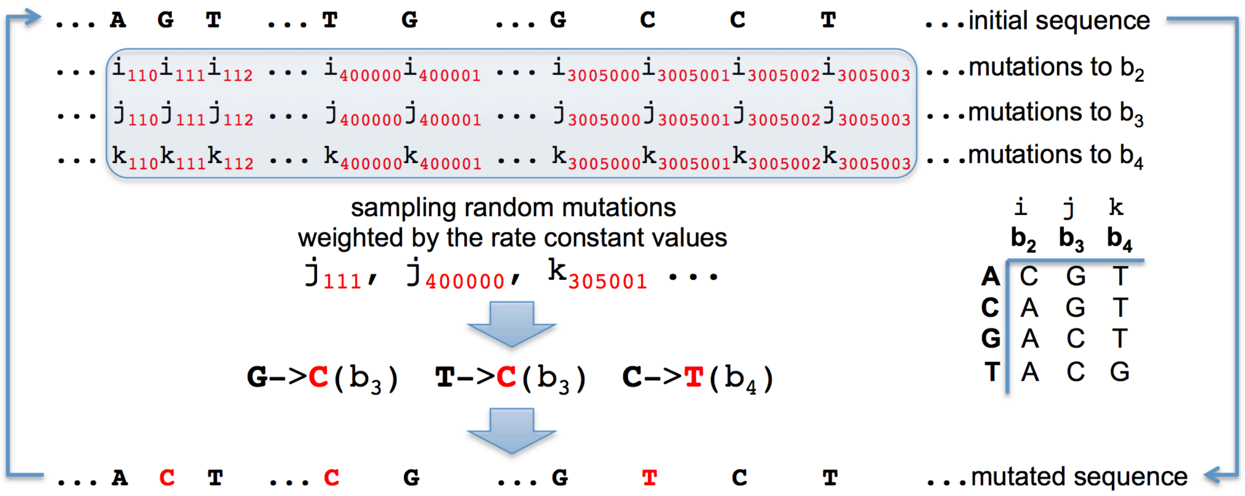

... A G T ... T G ... G C C T ...initial sequence
... i110 i111 i112 ... i400000 i400001 ... i3005000 i3005001 i3005002 i3005003 ...mutations to b2
... j110 j111 j112 ... j400000 j400001 ... j3005000 j3005001 j3005002 j3005003 ...mutations to b3
... k110 k111 k112 ... k400000 k400001 ... k3005000 k3005001 k3005002 k3005003 ...mutations to b4
sampling random mutations
weighted by the rate constant values
j111, j400000, k305001 ...
G->C(b3) T->C(b3) C->T(b4)
... A C T ... C G ... G T C T ...mutated sequence
i j k
b2 b3 b4
A C G T
C A G T
G A C T
T A C G

Figure S7. The *in silico* evolution of a genome with only $r_{i,j}^{core}$ mutation rate constants. A random sequence of 5-mln-nt is generated with 60% G+C content (30% G, 30% C, 20% A, 20% T). Next, for each position, *pos*, the i_{pos}, j_{pos} and k_{pos} rate constants for the mutations into the other three b_2, b_3 and b_4 bases are obtained from the Trek database, using the 7-nt sequence window centred at each *pos*. This generates 15-mln (3 times the sequence size) rate constant values that, for the first-order kinetic processes, are proportional to the mutation probabilities. We then sampled 5000 instances from the 15 mln generated set of *i*, *j* and *k*, with the sampling process weighted by the i_{pos}, j_{pos} and k_{pos} values. Those sampled instances contain information on both the mutation positions (*pos* subscript) and the mutation types (*i*, *j* and *k* describing the mutations to b_2, b_3 and b_4 bases respectively, the latter ones being bases other than the initial base, always ordered alphabetically). After performing the sampled mutations, we repeat the cycle and continue the process each time with mutation rates updated wherever the already performed mutations affect the values. We continue the simulation until the equilibration of the sequence composition.

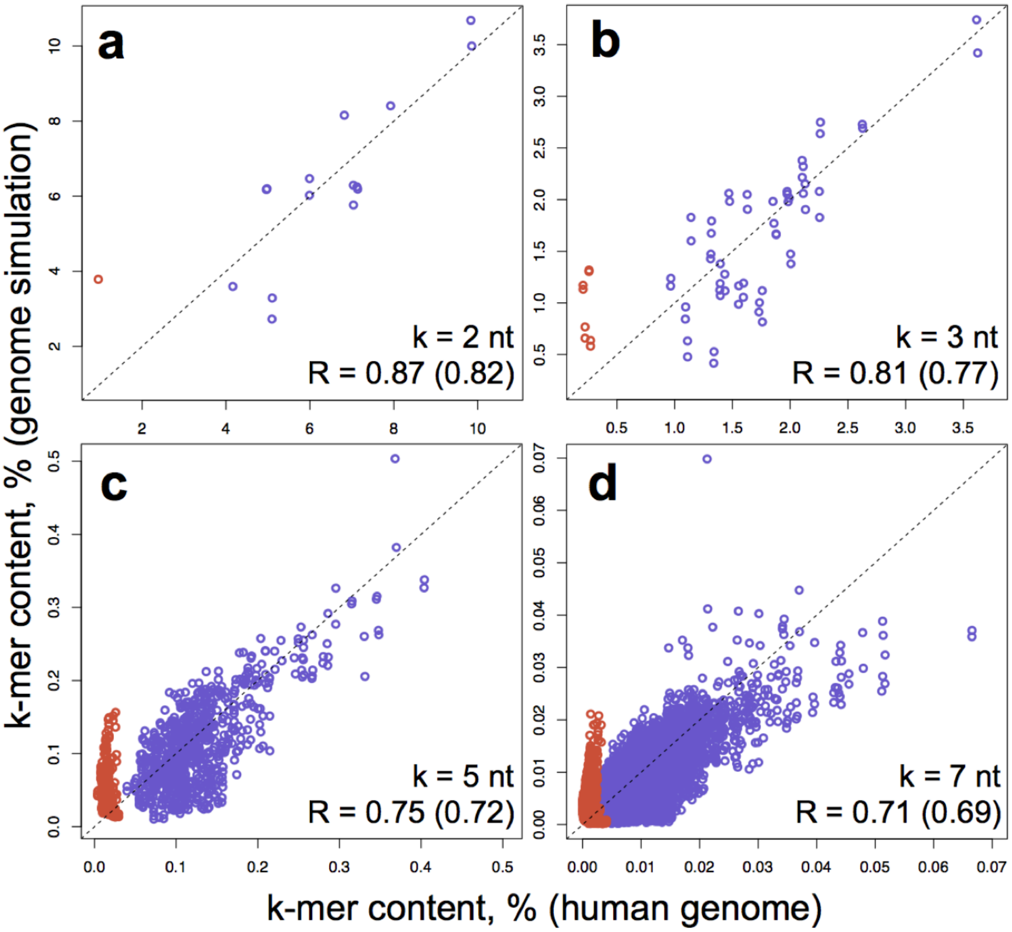


Figure S8. The comparison of the *in silico* (no strand-symmetries) evolved and real human genomes. The *in silico* genome is equilibrated here by using the raw Trek database but without accounting for the strand-symmetries of the $r_{i,j}^{core}$ mutation rates (contrasting with **Fig. 3d-g**). The plots **a**-**d** show the correlation of different k-mer contents in the equilibrated sequence with the corresponding content in the real human genome. The lengths of the k-mers along with the correlation coefficients are shown on the bottom right corners of the figure. Two correlation coefficients are shown with the exclusion and the inclusion (the value in the bracket) of the CpG containing oligomers (red points in the plots). The dashed lines depict the diagonals for the ideal match of the k-mer contents.

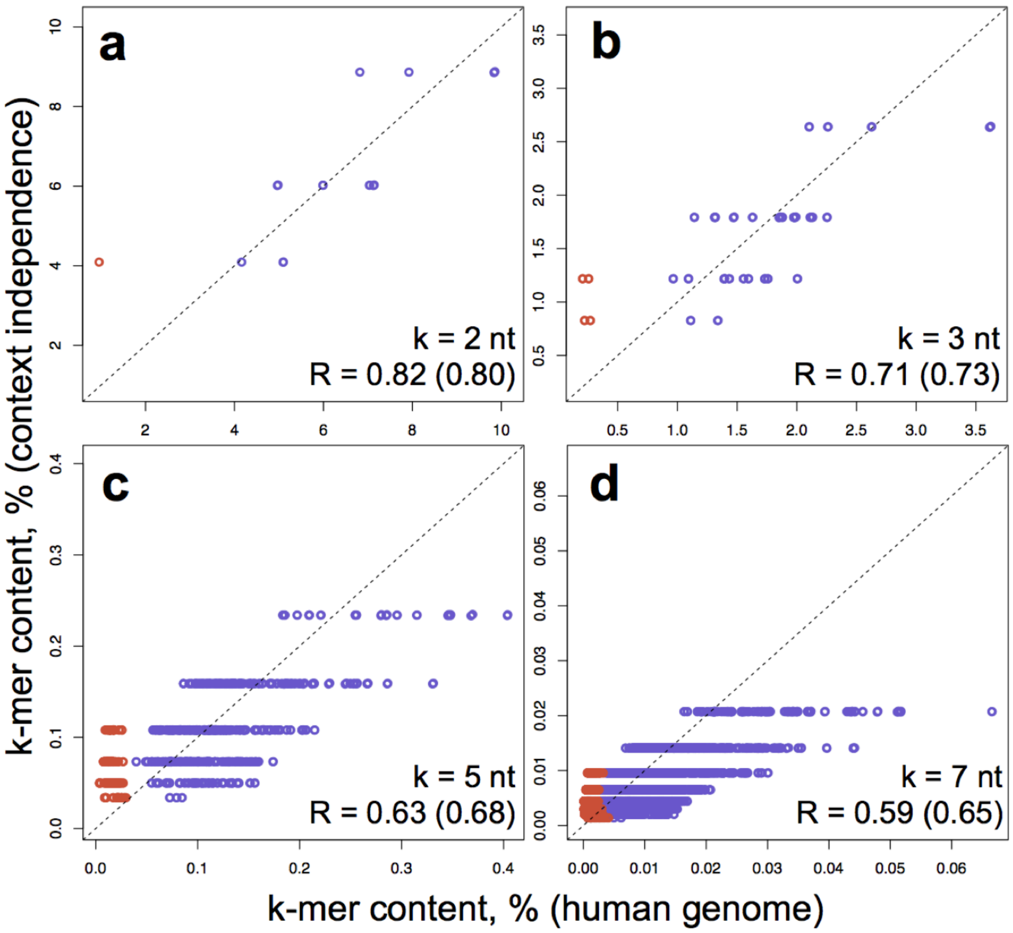


Figure S9. The oligomeric composition of a neighbour-invariant sequence. The oligomeric content of the human genome (x-axis) is compared to the content expected by a chance (y-axis) in a sequence that has the exact single-base composition as the human genome, but has mutation rates that are absolutely context independent. This corresponds to hypothetic genome simulation with perfectly corect, ideal $r_{i,j}^{sb}$ single-base rate constants, but without any $\delta r_{i,j}^{sr}$ sequence-context dependence present. The lengths of the k-mers, along with the Pearson's correlation coefficients without and with (the values in the brackets) the CpG containing oligomer data (red points) are shown on the bottom right corners of the plots. The correlation coefficients are notably smaller compared to the *in silico* sequence equilibrated based on the full set of context-dependent Trek $r_{i,j}^{core}$ constants. The dashed lines depict the diagonals for the ideal match of the k-mer contents.

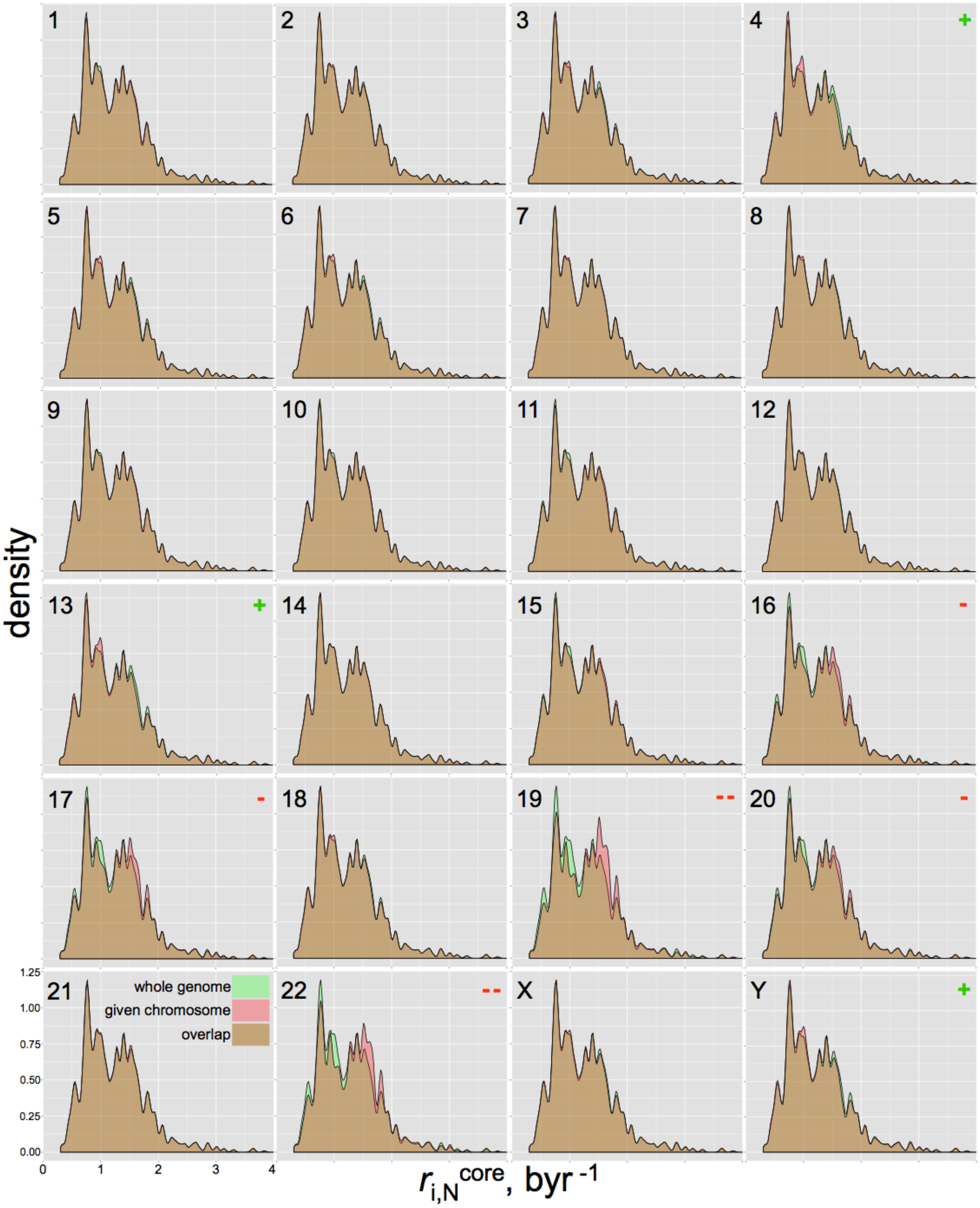


Figure S10. Basal mutability profiles of the human genome and its individual chromosomes. The plots show the density (kernel density estimate) distribution of the mutability rates in the whole human genome (green), as compared to the individual chromosomes (red). The chromosome types are shown at the top-left corners of the plots. The overlaps of both distributions are in brown. The x-axis shows the mutability rate constant for the mutation to any other base ($r_{i,N}^{core} = r_{i,b2}^{core} + r_{i,b3}^{core} + r_{i,b4}^{core}$), as inferred from mapping the positions with the context information (up to 7-mers) to the Trek database. Most of the chromosomes repeat the whole-genome mutation profile, with the significant exceptions noted for the chromosomes 19 and 22 that are relatively “destabilised”, in part due to their high G+C contents. The type (+ for stabilisation and - for destabilisation) of the difference between the profiles are shown at the top-right corners of the plots.

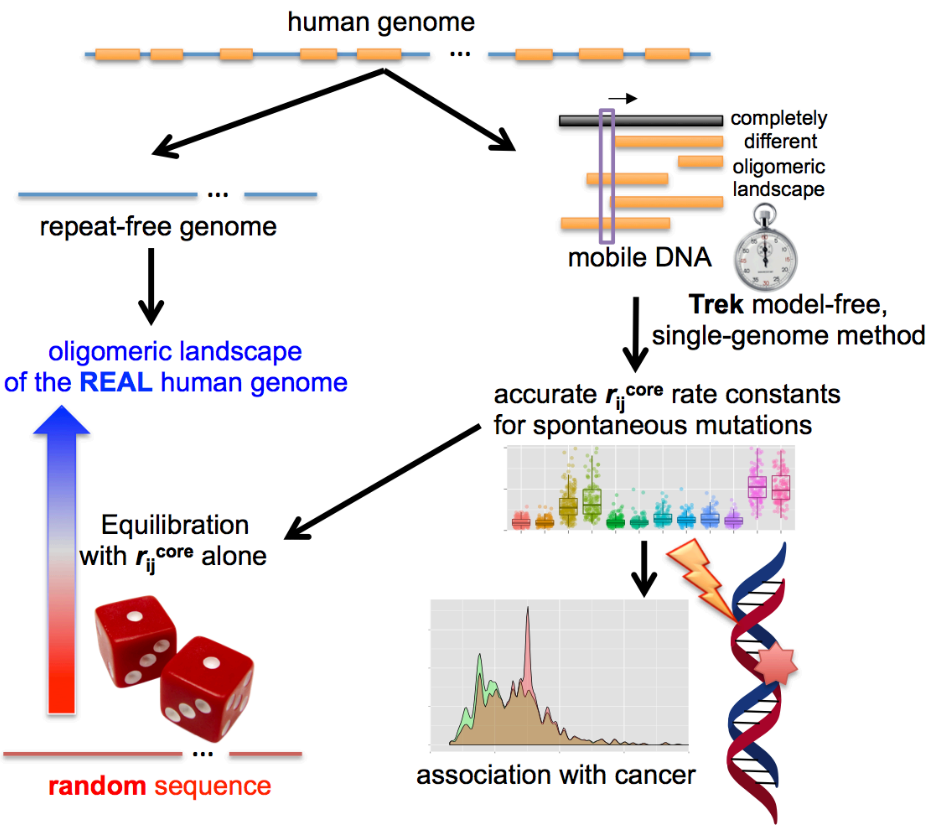

human genome
...
completely different oligomeric landscape
mobile DNA
repeat-free genome
...
Trek model-free, single-genome method
oligomeric landscape of the **REAL** human genome
accurate r_{ij}^{core} rate constants for spontaneous mutations
Equilibration with r_{ij}^{core} alone
association with cancer
...
random sequence

Figure S11. Schematic representation of the overall design and major outcomes of this work.

DESCRIPTION OF THE SUPPLEMENTARY VIDEO AND DATA FILES

Supplementary_Video_1.mp4

Simulated evolution of a random genome guided solely by the $r_{\mathrm{i,j}}^{\mathrm{core}}$ mutation rate constants.
The starting 5mln-nt-long DNA sequence has 60% G+C content. The simulation goes on till the equilibration of the overall G+C content.

Supplementary_Data_1.txt

Data on the Trek $r_{\mathrm{i,j}}^{\mathrm{core}}$ core mutation rate constants in the L1 reference sequence (L1Hs). The columns hold the substitution types, sequence positions, sequence contexts with 5 upstream and 5 downstream bases, $r_{\mathrm{i,j}}^{\mathrm{core}}$ constants (byr^{-1}) and the correlation coefficient of the linear fit behind the $r_{\mathrm{i,j}}^{\mathrm{core}}$ inference.

Supplementary_Data_2.txt

Trek database processed with strand-symmetry consideration. The file contains all 7-mers, outlining the central base (*i*) where the rate constants (byr^{-1}) are given for the core neutral mutations into the 3 non-*i* bases (shown in an alphabetical order). The final columns hold the quality scores for $r_{\mathrm{i,j}}^{\mathrm{core}}$ (actual sequence lengths matched with L1).

Supplementary_Data_3.txt

Trek database processed without strand-symmetry consideration. The file contains all 7-mers, outlining the central base (*i*) where the rate constants (byr^{-1}) are given for the core neutral mutations into the 3 non-*i* bases (shown in an alphabetical order). The final columns hold the quality scores for $r_{\mathrm{i,j}}^{\mathrm{core}}$ (actual sequence lengths matched with L1).

Supplementary_Data_4.txt

k-mer content for the repeat masked and unmasked versions of the human genome (RefSeq, hg19/GRCh37). The text file contains sections with headers showing the k-mer size and the genome masking status.

Supplementary_Data_5.txt

The full set of all 7-mer sequences with the respective cancer enrichment scores and mutability constants. These data are based on the comparison of the COSMIC dataset with the 7-mer distribution in the human RefSeq. Only the non-SNP and non-coding sites undergoing single-point mutations are taken from COSMIC (NCV set).

Supplementary_Data_6.txt

The GBM parameters that minimise the error of the test tree-based models, used to infer the optimal length window of neighbour effects. The columns represent the found best tuning parameters, along with the RMSE values for each *i->j* substitution type and window length.

SUPPLEMENTARY REFERENCES

1. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0,* 2013-2015, at http://www.repeatmasker.org.
2. Khan, H. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* **16,** 78–87 (2006).
3. Arndt, P. F., Hwa, T. & Petrov, D. A. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J. Mol. Evol.* **60,** 748–763 (2005).
4. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 961–968 (2010).