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.)

ABSTRACT

Accurate knowledge on the core components of mutation rates is of vital importance to understand genome dynamics. By performing a single-genome and model-free analysis of 39894 retrotransposon remnants, we reveal core, sequence-dependent, nucleotide substitution rates at each of the 3.2 billion positions of the human genome. Benefiting from the data made available in such detail, we show that a simulated genome generated by equilibrating a random DNA sequence solely using our rate constants, exhibits nucleotide organisation observed in the actual human genome, with or without repeat elements. This directly demonstrates the key role of the core nucleotide substitution rates in shaping the oligomeric composition of the human genome. We next generate the basal mutability profile of the human genome and show the depletion of the moieties with low basal mutability in the database of cancer mutations.

INTRODUCTION

The stability, dynamics and organisation of genomes are key factors that influence the molecular evolution of life[1]. Single-nucleotide substitutions occur an order of magnitude more frequently than common insertions/deletions[2,3], and are major contributors to the sampling pathways by which a genome changes over the passage of time. A thorough understanding of the descriptors that govern single-nucleotide substitutions (mutations hereafter) is thus essential to comprehend genome dynamics and its connection to the underlying first principle processes.

For a given genomic position and *i->j* nucleotide substitution, the mutation rate, as expressed by the rate constant $r_{i,j}$, can be presented as a single-base average value $r_{i,j}^{sb}$ and fluctuations contributed by

short-range context ($\delta r_{i,j}^{sr}$), CpG-associated ($\delta r_{i,j}^{CpG}$), long-range ($\delta r_{i,j}^{lr}$), gene/functional ($\delta r_{i,j}^{gene}$), and specific ($\delta r_{i,j}^{spec}$) effects:

$$r_{i,j} = \boldsymbol{r_{i,j}^{sb}} + \boldsymbol{\delta r_{i,j}^{sr}} + \delta r_{i,j}^{CpG} + \delta r_{i,j}^{lr} + \delta r_{i,j}^{gene} + \delta r_{i,j}^{spec} \quad (1)$$

The $r_{i,j}^{sb}$ term can be estimated through genomic averages for the individual *i->j* mutation rates, and is reported for the genomes of human[4,5] and other species[6-8]. By investigating the aggregation patterns in substitution frequencies, it was shown that the $r_{i,j}$ variation is subjected to two distinct short-range (<10 nt) and long-range (>1000 nt) effects[9,10]. In the equation above, the short-range effect is captured through the $\delta r_{i,j}^{sr}$ term and describes the totality of the intrinsic properties and sequence-dependent interactions of DNA with overall mutagenic and reparation processes in a given organism[11]. The better-studied mutation patterns at a CpG context[12-14] are separated in the $\delta r_{i,j}^{CpG}$ term, since besides having a specific short-range dyad context, CpG mutations also depend on a number of regional factors that alter the epigenetic targeting of the CpG sites[10,15-18]. Many relatively recent studies have shed light on the $\delta r_{i,j}^{lr}$ variation caused by the regional effects that depend on a long-range sequence context through secondary mechanisms, such as recombination and GC-biased gene conversion[1,4,19,20], transcription-coupled biased genome repair[21] and instability[22], chromatin organisation[23], replication-associated mutational bias[24] and inhomogenous repair[25], differential DNA mismatch repair[26], non-allelic gene conversion[27], and male mutation bias[28]. The term $\delta r_{i,j}^{gene}$ captures the change in mutation rates in genes and other functional elements under strong selection bias and reflects observations such as the increased neutral substitution rates in exons[29,30] and the possible reduction of the mutation rates in X-chromosome[31]. $\delta r_{i,j}^{spec}$ holds the highly specific increase or decrease in mutation rates governed by targeted mechanisms[32].

Herein, we obtain the core components (**Eq. 2**) of the spontaneous single-nucleotide substitution rates via the direct analysis of 39894 L1 mobile DNA remnants[33] in the same, human, genome (a single-genome approach).

$$r_{i,j}^{core} = r_{i,j}^{sb} + \delta r_{i,j}^{sr} \quad (2)$$

Our **tr**ansposon **e**xposed **k** (Trek) method provides the $r_{i,j}^{core}$ rates at single-nucleotide resolution in L1, where we demonstrate sufficient sequence variability to cover a wide-range of sequence contexts. We use this coverage to determine the core rate constants for all possible nucleotide substitutions (3 per position) at every single 3.2 billion position in the human genome. The Trek method reveals the $r_{i,j}^{core}$ variation in a model-free manner and at a level beyond accounting for only the two immediate neighbouring nucleotides[34]. Furthermore, we make our dataset, holding the time-dependent rate constants for individual substitutions that account for up to 7-mer sequence context effect, publically

available. Importantly, we demonstrate that $r_{i,j}^{core}$ values alone can generate a sequence, from first principles, starting from a random DNA sequence, whose key features reflect the oligomeric organisation of the actual human genome. Next, we calculate the core mutability profile of the human genome evaluating the basal predisposition to single-nucleotide substitutions and outline the decreased frequency of the stable sequence motifs among the sites linked to somatic cancer mutations.

RESULTS AND DISCUSSION

Revealing the core single-nucleotide substitution rates

The repetitive occurrence of mobile DNA elements in different regions within the same genome[33] provides the opportunity to obtain the core $r_{i,j}^{core}$ mutation rate constants that account for the $\delta r_{i,j}^{sr}$ immediate effects of neighbouring nucleotides. After the initial inactivation at different time epochs[35-38], individual remnants of many transposon subfamilies within a genome have been subjected to largely the same overall mutagenic and proofreading conditions as the rest of the genome[39], hence can also serve as markers of $r_{i,j}^{core}$ mutation rates applicable to genomic sites that share the immediate sequence-context. For the purpose of this study, we have used the hominoid lineage of the L1 LINE retrotransposons, spanning 3.1 to 20.4 myr (million years) of age[36]. The constituent subfamilies of the lineage are L1PA5, L1PA4, L1PA3, L1PA2 and the most recent L1Hs. Their respective age and the number of insertions in the human genome are presented in **Supplementary Table S1**. The choice was made through the following reasoning. The L1 elements have a long (~6k nt) sequence without extended repeats like in the LTR elements[33]. This enables their robust mapping on a chosen template and provides essential local sequence variability around different nucleotide positions within L1 elements. There are distinct L1 subfamilies that were active at different time epochs, with detailed molecular clock analyses available[35-38] to reveal and, importantly, validate the age of each subfamily. They are well-represented and, unlike other classes of transposable elements, are uniformly scattered across mostly the intergenic regions of the human genome[33,40,41]. Unlike SINEs and LTRs, LINE sites show very low level of RNA polymerase enrichment, as a marker of transcriptional association, in normal tissues[42]. The selected most-recent subfamilies are sufficiently young[36] a) to enable an unambiguous identification of the genomic coordinates of the borders for the remnants; b) to assume that each position in those elements would be unlikely to mutate multiple times over the studied period of their existence as remnants in the human genome (see **Methods**); c) to attribute a time-invariance to the rates during the analysed period of mutation accumulation[4,43,44]. The young L1 subfamilies have

most of their remnants coming from the genomic regions with G+C content close to the genomic average value[40] (see **Supplementary Fig. S1**). Finally, many matching positions in our studied five L1 representatives share the same consensus bases, hence, such positions are not polymorphic due to adaptive pressure and can serve as internal references for inferring the $r_{i,j}^{core}$ rates.

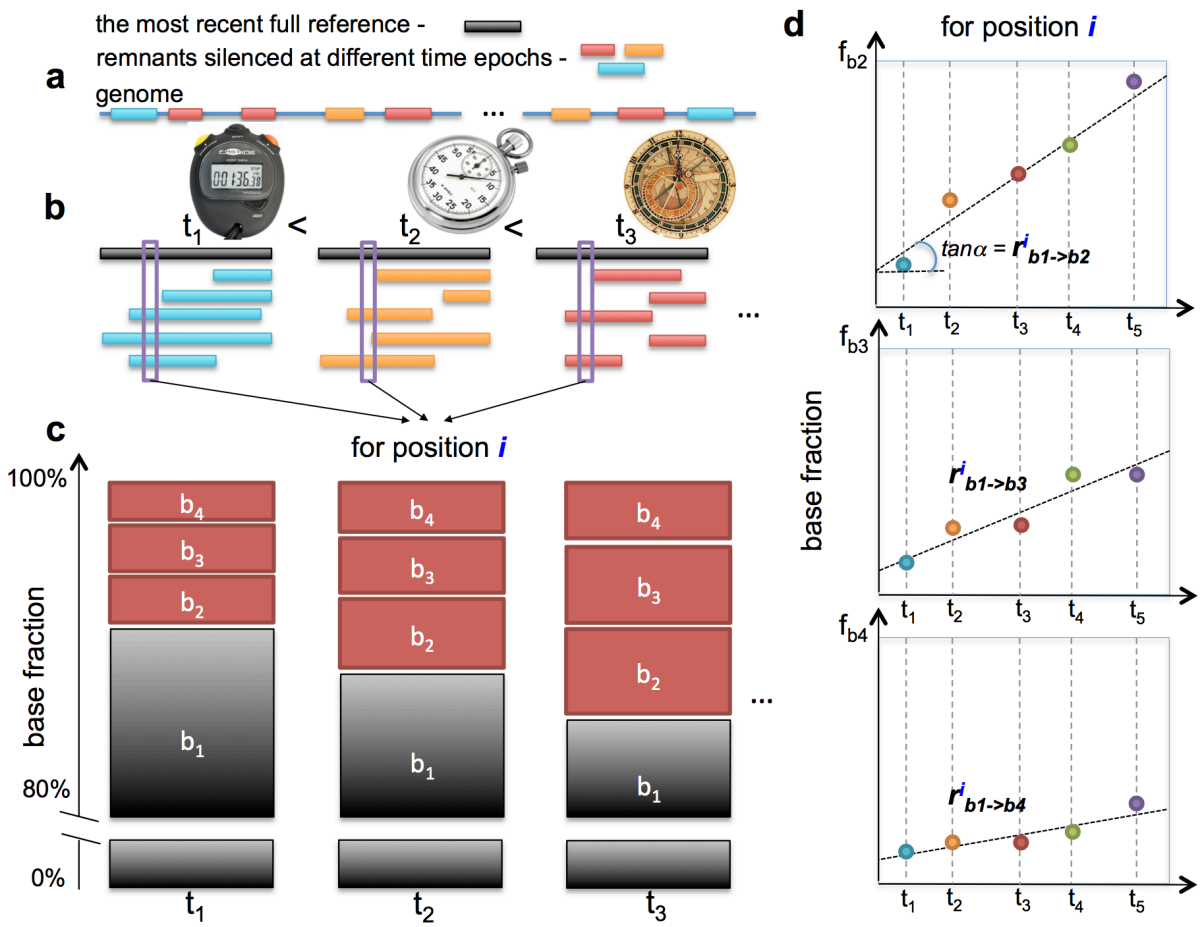

Figure 1. Trek methodology of determining the core single-nucleotide substitution rate constants. (**a**) The Trek approach is applicable to a genome containing multiple remnants of retrotransposon subfamilies silenced at different time epochs. We can consider those subfamilies as mutation counters that had different resetting ages (**b**). The full consensus sequence of the most recent subfamily is taken as a reference (**a**). The remnants are then grouped by their age and fully mapped onto the reference sequence (**b**). For each position *i* in the reference sequence, the fractions of the four bases in all the time groups are calculated (**c**). The comparison of these fractions coming from individual base types across different time periods enables a linear model fitting, through which we can reveal the rates of the mutations for the substitutions into the b_2, b_3 and b_4 bases from the consensus (b_1) state of the given position (**d**). The steps **c** and **d** are repeated for all the positions in the reference sequence, producing single-nucleotide resolution core mutation rate constants with a sequence-context dependency as sampled in the reference sequence of the mobile element. To assure the high quality and neutrality of the retrieved mutation rate constants, we accounted for the sites in the reference sequence that had at least 700 mapped occurrences in each time group (**b**), with the same wild-type variant being always the prevalent one (more than 80%) in each subfamily (**c**) and producing a Pearson's correlation coefficient of at least 0.7 in the time-evolution plots (**d**).

The Trek methodology of obtaining $r_{i,j}^{core}$ rates, along with the considerations for filtering out the possible selection and non-neutral mutation sites, is presented in **Fig. 1** with further details in **Methods** and **Supplementary Fig. S2**. The acquired data on the full set of position-specific mutation rates are presented in **Fig. 2**.

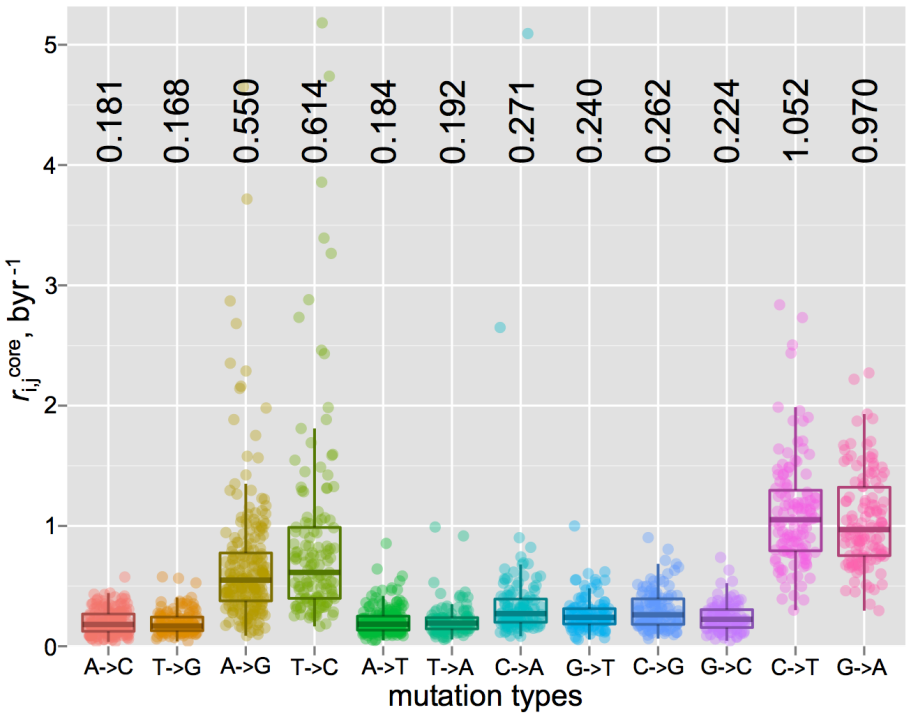

Figure 2. Transposon exposed (Trek) $r_{i,j}^{core}$ substitution rate constants of the human genome. The boxplots are shown for each *i->j* substitution type inferred from the hominoid L1 elements spread across the human genome. Each point comes from a specific position in the L1 element, reflecting the mutation rate constant averaged across multiple occurrences of that specific position with the same sequence-context in multiple regions of the genome. The complementary *i->j* pairs are plotted in adjacency. The median values of the overall mutation rates in byr^{-1} (billion years) unit, averaged across the varying sequence-context within the L1 elements, are shown on the top.

A total of 661 positions, at the 3' side of the L1 elements used, passed our robustness checks (see **Methods**) and were thus employed to infer the corresponding $r_{i,j}^{core}$ values from the analysis of all the young L1 remnants in the human genome. We recorded the data in the Trek database that contains a set of well-defined $r_{i,j}^{core}$ constants (see below for the extent of sequence context coverage in the Trek database) capturing the influence of the unique arrangement of neighbouring nucleotides at those positions. Owing to the nature of the selected L1 elements, as discussed above, and the Trek procedure design (**Fig. 1**), we expect the absence of the $\delta r_{i,j}^{gene}$ contribution, the elimination of $\delta r_{i,j}^{lr}$ at the averaging stage (**Fig. 1b**) and the removal of the $\delta r_{i,j}^{CpG}$ and $\delta r_{i,j}^{spec}$ effects through our robustness checks embedded within the Trek procedure (see **Fig. 1c,d** and **Methods**). Therefore, our method

provides the $r_{i,j}^{core} = r_{i,j}^{sb} + \delta r_{i,j}^{sr}$ core variation (**Fig. 2**) of the mutation rates at around the $r_{i,j}^{sb}$ genomic average values for each *i->j* base substitution. If the above is correct and our method indeed results in $r_{i,j}^{core}$ values, further averaging of the core $r_{i,j}^{sb} + \delta r_{i,j}^{sr}$ rates (median values shown in **Fig. 2**) should give us the single-base $r_{i,j}^{sb}$ genomic average substitution rates, cancelling out the remaining $\delta r_{i,j}^{sr}$ contribution. In fact, the comparisons of our Trek-derived $r_{i,j}^{sb}$ with two published datasets that report on the genomic average $r_{i,j}^{sb}$ rates[4,5] show an excellent correlation (**Supplementary Fig. S3**, Pearson's R>0.99) confirming the absence of any bias and unusual mutation rates in the time-accumulated substitutions at the L1 sites that pass the Trek procedure. The genome simulation, described later in this work, provides an additional validation for our rate constants. The $r_{i,j}^{core}$ values (**Eq. 1** and **2**) for all possible *i->j* substitutions inferred for each of the eligible individual L1 positions are thus assumed to be common for any other sites in the genome that share the short-range sequence context.

The influence range of neighbour nucleotides

To apply the $r_{i,j}^{core}$ constants to the human genome, we first established the optimal length of a DNA sequence (k-mer, where k is the length of the sequence) capturing most of the influences that modulate mutation rates of the base at the centre. For this, we evaluated the power of the knowledge of the neighbouring arrangement of nucleotides in predicting the $r_{i,j}^{core}$ constants for each of the twelve *i->j* substitution types, where *i* and *j* are the four DNA bases. We built test predictors for individual substitution types via a tree-based machine learning technique, while using varying lengths of sequences centred at the positions where the rate constants were to be predicted (see **Methods**). The aim of the machine learning procedure was to establish the optimal sequence length to minimise the error in the predicted rate constants (**Supplementary Fig. S4** and **S5**). In agreement with prior evidence[9,10,45,46], but now obtained for each individual *i->j* substitution type from Trek data, the optimal window was found to be 5-7-nt (both 5- and 7-nt resulting in comparable results for many substitution types) and was subsequently used as guidance for the direct mapping of the Trek rate constants from the L1 sequence onto any given human nuclear DNA sequence for the $r_{i,j}^{core}$ assignment.

Mapping the Trek $r_{i,j}^{core}$ data on any DNA sequence

The upper 7-nt size window for determining the single-nucleotide substitution rate constants at the central base accounts for 3 upstream and 3 downstream bases relative to each nucleotide position. Our substitution positions that pass the Trek criteria capture 636 unique 7-mers out of the possible 16384 (4^7). Therefore, for many loci in the human genome we need to use a smaller window (< 7-mer) as a

match criterion to assign to one of the Trek rate constant sets. By trimming the size of the k-mer to 5, hence accounting for 2 upstream and 2 downstream bases, we cover 404 unique sequences out of possible 1024 (4^5). Further reduction of the size to 3, allows having data for 56 unique triads out of 64 (leaving out only the CpG containing triads, see below). For the single-base case (1-mers), where we average out all short-range neighbour effects and longer-range sequence variability, we obtain data for all the 4 bases and 4×3 possible substitutions as shown in **Fig. 2** (the median values on the top of the figure). The coverage of the longer k-mers is however increased nearly twice when we account for the strand-symmetry, as described in **Methods**. Please note, that for each unique k-mer we obtained 3 $r_{i,j}^{core}$ constants via the described analysis of a large pool of L1 remnants from different genomic loci.

With the above considerations, we created a program (Trek mapper) to produce the full range of $r_{i,j}^{core}$ core mutation rate constants for any sequence, accounting for the context information for up to the 7-mer window and pulling the matching core data from the Trek database. Should a representative match be absent with the full 7-nt long sequence, the window around the given position in a query sequence is shortened into the longest variant possible (out of the 5-nt, 3-nt or 1-nt lengths) with a full match in the Trek database (see **Methods** and **Supplementary Fig. S6**). In this way, for all the possible 16384 7-mers, our Trek database reports 49152 rate constants (3×16384), of which 3168 (6.4%) account for the 7-mer context, 23232 (47.3%) account for the nested 5-mer context, 17120 (34.8%) for 3-mer and only 5632 (11.5%, CpG containing sequences) constants do not account for any context effect on the central base, since we eliminate those by design, due to the $\delta r_{i,j}^{CpG}$ contributions. We thus produce an unprecedented dataset that reports, and makes publically available (**Supplementary Data 2** and **Data 3**), the direct, $r_{i,j}^{core}$ rates for all individual *i->j* substitutions accounting for the context effects beyond the 64 triads[34]. If we consider only the unique values in the Trek database, we report 2078 unique rate constants (taking into account different extent of averaging, where multiple entries are present for the different context ranges), of which 1208 (58.1%), 782 (37.6%), 85 (4.1%) and 3 (0.1%) entries account for 7-, 5-, 3- and 1-mer contexts respectively. The 1-mer averaged data were used for the k-mers in that contain either C or G bases of a CpG dyad at the centre, to assign the overall mutation rate constants by the Trek mapper. This was done since none of the CpG sites in the L1 elements passed our robustness checks, due to the targeted epigenetic control ($\delta r_{i,j}^{CpG}$) of the single-nucleotide substitutions there[12-14], which were also non-uniform with time (active targeting, $\delta r_{i,j}^{spec}$, while in the viable epoch for each L1 subfamily) and were present to silence the active retrotransposons.

Our current data are for the human nuclear genome. However, the general approach for obtaining $r_{i,j}^{core}$ constants is applicable to any organism where the genome contains a set of well-characterised and related young mobile elements silenced at different time epochs and without notable context bias.

The origin of the oligomeric landscape in the human genome

The full set of sequence-dependent human $r_{i,j}^{core}$ mutation rates (all 3 constants per position) enabled us to perform a sophisticated *in silico* evolution of a random DNA sequence, guided solely by our $r_{i,j}^{core}$ values. We started from a random sequence of 5 million (mln) nt with a G+C content of 60% (substantially greater than the 40.45% G+C content for the human genome). We performed random mutations weighted by Trek-inferred probabilities (see **Methods**, **Supplementary Fig. S7** and **Supplementary Video 1**), where, after each cycle, the mutation rate constants were updated for the sequence positions that were either mutated or fell within the influence zone of the performed mutations. The simulation was continued until the overall G+C content of the simulated sequence became constant (see **Fig. 3a-c**).

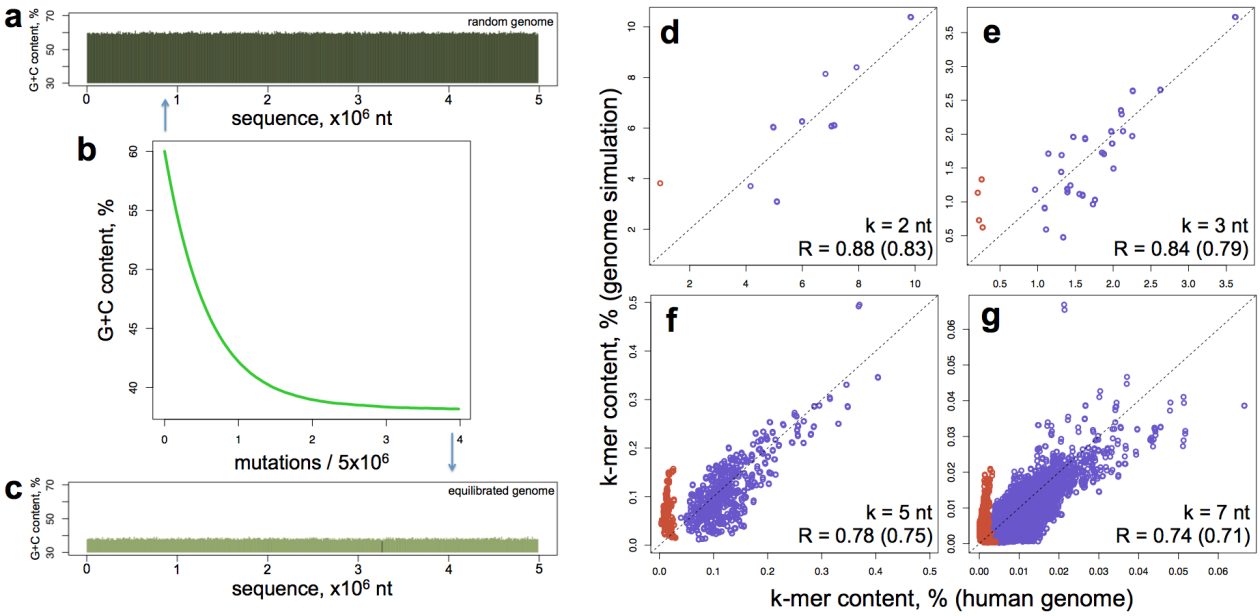

Figure 3. Comparison of the *in silico* evolved and actual human genomes. (**a**) The 5-mln-nt starting sequence is randomly generated with 60% G+C content. The sequence is then neutrally evolved using $r_{i,j}^{core}$ values only, until the base-compositional equilibrium is established (**a**-**c**). This was reached after about 20 mln mutations (or an average of 4 mutations per site (**b**). The equilibration converges faster when we start from a sequence with lower G+C content. The plots **d**-**g** show the correlation of the k-mer contents in the equilibrated genome 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 plots. Two correlation coefficients are shown with the exclusion and the inclusion (the value in the bracket) of CpG containing oligomers (red points in the plots). The dashed lines depict the diagonals for the ideal match of the k-mer contents.

The simulation converged to generate a sequence with the A, T, G and C compositions of 30.91, 30.90, 19.06 and 19.13% respectively. Note, that these values are in good agreement with the A, T, G and C compositions of the repeat-masked human genome of 29.75, 29.79, 20.24 and 20.22% respectively (see **Methods**). Furthermore, the simulated genome captures the contents of different individual oligomers (k-mers) in the human genome. The data for all the possible 16 dyads, 64 triads, 1024 pentads and 16384 heptads are presented in **Fig. 3d-g** and show a significant (see the correlation coefficients) correlation between the compositional landscapes of the Trek-simulated genome and the actual human genome. Regardless of the starting composition of the initial DNA sequences, our simulations always equilibrated to a state with similar oligomer (up to 7-mer) content. The k-mer contents shown in **Fig. 3d-g** for the actual human genome were calculated from the repeat-masked version of the RefSeq human genome, where all the identified repeat elements, including the L1, were disregarded. This assured the removal of a potential bias due to the presence of L1 elements in the human genome. As $r_{i,j}^{core}$ constants are free of the $\delta r_{i,j}^{CpG}$ contribution (see above), the simulated genome produced higher alterations in representing the k-mer contents that have CpGs (red points in **Fig. 3d-g**), which directly demonstrates the contribution of $\delta r_{i,j}^{CpG}$ to the background compositional landscape of the human genome.

The correlations in **Fig. 3** are from simulations where the rate constants were symmetrised according to the inherit strand-symmetry in double helical DNA (see **Methods**). The results without such equalisation are still significant, though producing slightly worse correlation coefficients (**Supplementary Fig. S8**).

To confirm that the observed correlations for different k-mer contents (**Fig. 3d-g**) arise due to our sequence-context-dependent core mutation rates, rather than as a side effect, by a pure chance, in a sequence where the simulation makes only the single-base composition converge to that of the real human genome (such as in sequence generated using an ideal 4×4 single-nucleotide substitution rate matrix), we calculated the expected distribution of different k-mers in a fully random genome but with the exact human A, T, G and C base compositions. In the complete absence of any sequence-context effects, the probability of the occurrence (fraction) of any k-mer in a sufficiently long sequence is equal to the product of the occurrence probabilities of their constituent bases. For instance, the probability of observing the AGT triad is the $p_{AGT} = p_A p_G p_T$ product, where the individual p_i probabilities are the base contents expressed in fractions. The comparison of the k-mer fractions obtained in this way with the human genome data (**Supplementary Fig. S9**) shows a substantially reduced correlation (for the genomic 7-mer content, Pearson's R=0.59 compared to 0.74 using Trek rates), supporting the

attribution of the important role to the $r_{i,j}^{core}$ values in shaping the compositional landscape of our genome.

The basal mutability profile of the human genome

Our Trek mapper provides the full set of $r_{i,j}^{core}$ constants for each position in the whole human genome. Such data enables us to calculate the context-dependent overall mutability by taking the sum of the individual rate constants for the 3 possible mutations at each base position, thus producing the core $r_{i,N}^{core}$ mutability constant for the substitution of a given base *i* by any other base *N*. **Supplementary Fig. S10** shows a comparison of the basal mutability profiles calculated for the individual chromosomes (red) with the whole genome profile (green), where most of the chromosomes exhibit the same overall distribution as the whole genome. Further grouping and analysis[47] of the unique sequences found in regions of different basal mutability for the whole human genome reveals motifs that tune the stabilities of the bases at the centre (see **Fig. 4** and the caption for further elaborations). Note that the observed sequence-determined mutational biases can potentially contribute to the initial nucleation of a more extensive base-content pattern formation in chromosomes[48,49].

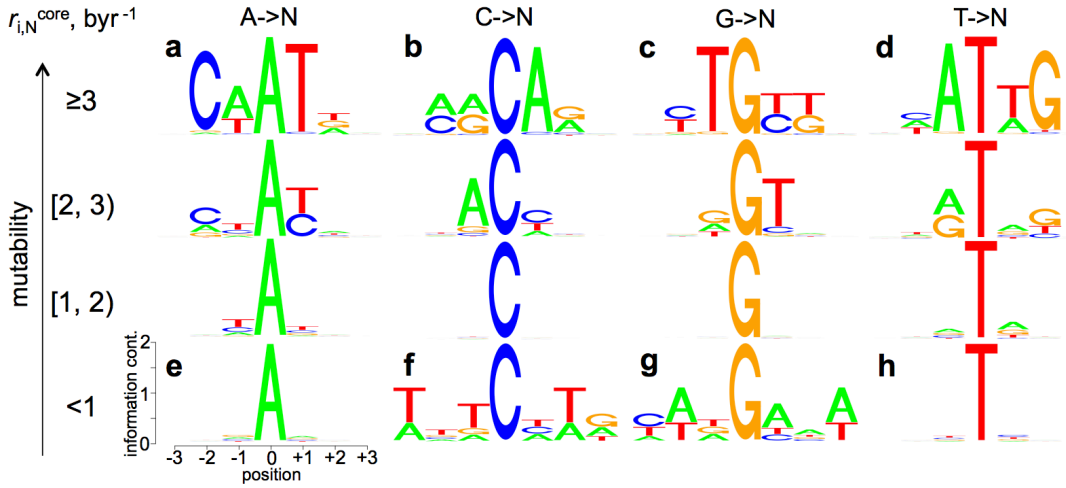

Figure 4. Sequence-context dependence of the $r_{i,N}^{core}$ basal mutability constants. Sequence logos[47] are shown for all the unique 7-mer sequences grouped by the central base type (columns) and the category of the mutability range the sequences fall in (rows, basal mutability range is shown in byr^{-1} rate constants). The y-axes in the individual sequence logos show the information content in bits. The x-axes outline the neighbouring base positions relative to the central base. For each sequence, the mutability constant of the central base (*i*) depicts the sum of the core rate constants for the mutations to the three other (non-*i*) bases, $r_{i,N}^{core} = r_{i,b2}^{core} + r_{i,b3}^{core} + r_{i,b4}^{core}$. As can be seen from the plots, the bases A and T are highly mutable when the neighbouring positions are enriched in the same, A and T, bases (compare the logos **a** and **d** with **e** and **h**). The adjacent enrichment in A increases the mutability of C (**b**), and decreases the mutability of G (**g**) bases. Conversely, the adjacent enrichment in T increases the mutability of G (**c**) and decreases that of the C (**f**) bases. Note, that our data are for $r_{i,N}^{core}$ and do not include the methylation-driven increased mutation rates in CpG dyads[12-14].

Basal mutability profile in cancer-linked sites

A recent study[50] suggested that the multi-etiologic nature of cancer is primarily linked to random chance. To this end, genomic sites with higher intrinsic mutability (lower stability) might exhibit a higher prevalence of cancer-related genome alterations, as compared to sites of lower intrinsic mutability (higher stability). Although the sequence context of cancer mutations and their dependence on cancer types is out of the scope of the present work and is covered in detail elsewhere[51-54], here we examined the simple relationship between our calculated basal mutability values and observed cancer-associated somatic mutations accessed via the annotated COSMIC cancer database[55] (see **Methods**).

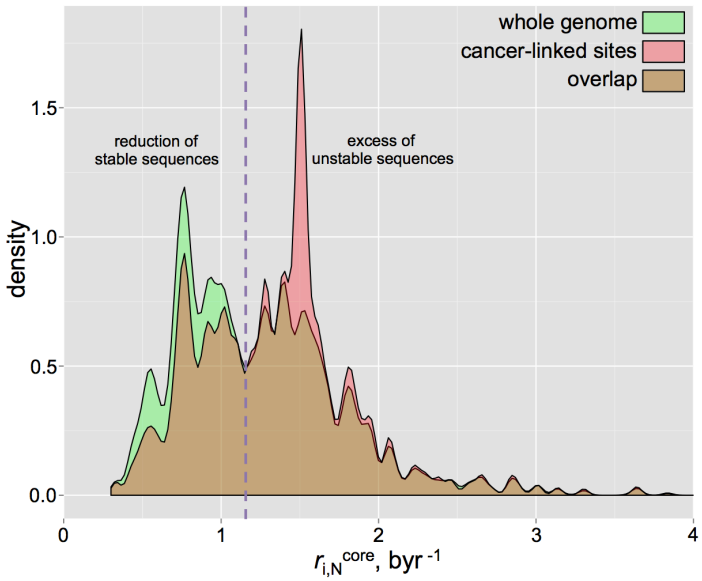

Figure 5. Basal mutability profiles of the whole human genome and cancer-linked mutation sites. The density (kernel density estimate) distribution of the mutation rates in the whole human genome (green), compared to the sites of the mutations associated with cancer (red) are shown. The overlaps of both distributions are in brown. The x-axis shows the mutability constant for the mutation to any other base ($r_{i,N}^{core} = r_{i,b2}^{core} + r_{i,b3}^{core} + r_{i,b4}^{core}$). The comparison clearly shows a reduction of the stable sites and excess of the unstable sites in cancer-linked loci.

Since the Trek data are for the core spontaneous substitutions, we restricted the analysis to the non-coding and non-polymorphic (not identified as SNP) single-nucleotide substitutions (6 mln mutations) in cancer. By mapping these sites to the human genome and retrieving the sequence-context information (7-nt long sequences centred at the mutation points), we processed the data with Trek mapper and obtained the $r_{i,N}^{core}$ profile for the non-coding sites detected in human cancer. The outcome in **Fig. 5**, overlapped with the whole-genome $r_{i,N}^{core}$ profile, shows that stable sites in the human genome, assigned by the Trek mapper to have mutability constants below 1.13 byr^{-1}, are significantly less represented in cancer.

Like many other disease-causing mutation sites[56], most of the sites that are highly enriched in cancer (**Fig. 6a**) are CpGs[57], which, even without accounting for the methylation driven increase[12-14] of the mutation rates, show high basal mutability values[58]. However, **Fig. 6b,c** demonstrates the discussed trend in the 7-mer cancer enrichment ratio (**Methods**) vs. basal mutability dependence even when all the CpG sites are removed from the analysis. Overall, our results show that the intrinsic basal mutability of different sites in DNA may contribute to their absence/presence in pathological genotypes. In particular, we observe that 7-mers with low mutability of the central base are relatively depleted in cancer-linked mutation data. They present a cancer enrichment ratio that is smaller than 1, whereas for the unstable 7-mers, the enrichment ratio, on average, tends to 1 (**Fig. 6c**), meaning that the enrichment in cancer is comparable to the one in the whole genome.

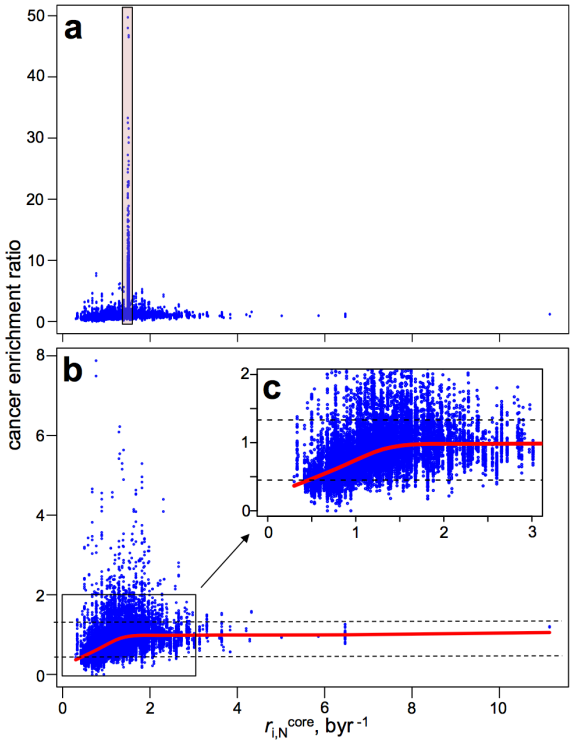

Figure 6. Enrichment of 7-mers with varying basal mutability in the cancer-linked mutation sites. The 4^7 points in the plots correspond to unique 7-mer sequences. The cancer enrichment score for each such sequence was calculated by dividing the occurrence fraction of the sequence in only the cancer-linked sites to the fraction in the whole repeat-masked human genome. All the 7-mers that had either C or G of a CpG dyad at the centre show a remarkable cancer enrichment ratio (up to 49, see the points in the red box in **a**). Since for the CpGs, Trek data report on only the average C and G mutation rate constants, the mutability values for those points can be underestimated in **a**. However, even the average mutabilities for C and G bases are higher than the discussed 1.13 byr^{-1} threshold. Accounting for the epigenetic methylation-driven increase in mutation rates for the CpG containing sequences would only increase their mutability values. The plots **b** and **c** represent the data that exclude the sequences with CpGs at the centre. The mean cancer enrichment ratio for such subset was 0.89, with standard deviation of 0.44. The data points within the 0.89±0.44 range of cancer enrichment ratio are contained in between the dashed lines in **b** and **c**. The red lines in **b** and **c** represent the Lowess[65] fit, showing the decrease of the cancer enrichment ratio with the decrease in mutability.

CONCLUSIONS

We have employed a single-genome and model-free approach (Trek) that reveals the core ($r_{i,j}^{core} = r_{i,j}^{sb} + \delta r_{i,j}^{sr}$) components of the spontaneous single-nucleotide substitution rates and basal mutability constants ($r_{i,N}^{core}$) for the human nuclear genome. Although the mobile DNA elements have been used before[4,39,43,59] for estimating averaged substitution rates, the increased quality of the human reference sequence and the detailed subfamily divergence studies for the L1 elements[35-38] done during the past decade enabled the construction of a specific direct method for the single-genome retrieval of the core $r_{i,j}^{core}$ rate constants at a single-nucleotide resolution, while also accounting for the comprehensive short-range context effects beyond the previous Bayesian estimates for the +1/-1 base effects[34]. The retrieval of our $r_{i,j}^{core}$ data in a single-genome manner adds additional value, since it ensures the absence of potential bias due to a) differences in the molecular machinery that influence the mutation rates while comparing the genomes of different species, and b) the short generation span where only the most probable mutations would become visible while comparing genomic data in families (parents/offspring) to detect *de novo* substitutions. As the major outcome of this study, we used our context-dependent rate constants to provide the first direct demonstration of the equilibration of a random DNA sequence into the one with overall (genomic base content) and short-range (different k-mer contents) characteristics that closely mirror that of the actual human genome. The L1-derived rates recapitulated fundamental properties of the repeat-masked and therefore L1-free human genome (**Supplementary Fig. S11**). This simulation thus provides significant evidence in support of the role of core neutral mutations in shaping the compositional dynamics of complex genomes[60] and additionally validates the reliability of the obtained Trek-derived $r_{i,j}^{core}$ constants. Importantly, our study demonstrates that the non-specific core mutation rates are capable of producing apparent selection or depletion patterns in the human genome. To this end, our *in silico* equilibrated sequences, obtained solely based on the full set of $r_{i,j}^{core}$ constants, can now serve as true background standards for the comparisons to reveal real selection[61] for or against different sequence motifs. The extended set of core mutation rate constants we report can potentially help advance our understanding in genome dynamics, with possible implications for the role of random mutations in the emergence of pathological genotypes and the evolution of proteomes.

METHODS

The human reference genome sequence was taken from the Ensembl database (www.ensembl.org), and was of the version hg19/GRCh37. The positions and span of the retrotransposons were taken from the output of the RepeatMasker[62] processing, accessed through the UCSC genome database (www.genome.ucsc.edu). The repeat annotations were those corresponding to the version of the used human RefSeq genome (hg19/GRCh37). The R programming language (www.r-project.org) was used for all the consecutive analyses. Most of the computations were performed on the available Linux workstation and computing cluster facilities hosted at the Department of Chemistry, University of Cambridge, and the Cancer Research UK Cambridge Institute.

Revealing the core mutation rate constants. All the remnant sequences of the mentioned subfamilies (**Supplementary Table S1**) were first aligned onto the 6064 nt reference sequence. As the reference, we took the consensus sequence of the human L1Hs retrotransposon (**Fig. 1a,b** and **Supplementary Fig. S2**). The alignment was done in a pairwise manner with high end-gap penalties ("overlap" mode) that, while allowing insertions and deletions, did not severely break the queried sequences for false mappings with a better global alignment score. R with the Biostrings library for alignment was used. After the alignment, all the relevant mutation fractions were collected for each position in the five L1 subfamilies reporting on a specific time epoch (**Fig. 1b,c**). For example, if the position i in the reference sequence was G (b_1), the mutation rate constants were calculated for the G->A (b_1->b_2) transition and G->C (b_1->b_3), G->T (b_1->b_4) transversions. First, the base fractions were calculated for five time-reporting L1 subfamilies; i.e. to get the fraction of mutations accumulated in ~20.4 myr (age of L1PA5[36]), all bases in L1PA5 remnants that were precisely mapped on the ith position of the reference sequence were counted and the fractions of G, A, C and T bases retrieved (**Fig. 1c**). Here, we applied one of the robustness checks and made sure that the fractions were estimated if at least 700 mapped bases were present for the ith position in each time-reporting subfamily (**Fig. 1b**). We also aimed to calculate such substitution rates for only the positions where the mutations are random and not specifically selected for or against. In other words, the position should not be a polymorphic or a subfamily speciation-defining nucleotide. We filtered out such cases by ensuring that any eligible i^{th} position had the same nucleotide of the reference sequence as its most prevalent variant with a minimum of 80% occurrence in all subfamilies (**Fig. 1c**). The average crude single-nucleotide substitution rate is noted to be 12.85×10^{-9} mutations per site per generation[5]. Assuming an average

generation length of 20 years[2], the mutation rate constant in a time domain can be crudely approximated as 0.64 byr^{-1}. In the course of 20.4 myr (the age of L1PA5), this should result in only a 1.31% mutated base fraction at a give site, caused by the average spontaneous substitution rates. Therefore, by assuming a threshold of 80%, we allow up to 15 times variation of the rates from the average estimate, which is a safe range[5] for the direct estimation of the single-nucleotide substitution rates and their core variation. Having the mutation fraction data from five different ages and for three (b_1->b_2, b_1->b_3, b_1->b_4) possible substitutions at the position *i* allowed to fit a linear model via the least squares methodology for the fraction-versus-time dependence for each mutation separately (**Fig. 1d**). If the data, hence the fitted line, were of high quality, the slope was expected to represent the $r_{i,j}^{core}$ mutation rate constant. We applied the third robustness filtering at this stage, by making sure that the rates were calculated for only the cases where the time correlation of the mutation fractions in **Fig. 1d** had greater than 0.7 Pearson's correlation coefficient. This ensured that the retrieved fractions of the mutations comprised of only the time-accumulated substitutions, rather than of targeted substitutions during the active life-span of the L1 elements, before their silencing. Please note, however, that the correlation coefficients in most of such time correlations that passed the whole Trek procedure were substantially higher (the observed Pearson's correlation coefficients were centred at 0.92 with 0.07 standard deviation). The procedure was done for all the 6064 positions in the L1 reference sequence, except the positions 5856-5895 and 6018-6064, close to the 3'-end (**Supplementary Fig. S2**) that engulf the low-complexity G-rich and A-rich sequences correspondingly, prone to alignment errors. One of the reasons for the usage of only the young L1 subfamilies (spanning 20.4 myr age) was to minimise the potential error in rate constant determination in the Trek procedure caused by repeated substitutions hitting the same position during the considered period of the mutation accumulation. The effect is indeed negligible for 20.4 myr span, as we can estimate using the above mentioned 0.64 byr^{-1} value[5] for the average *i*->*j* substitution rate constant, *r*. Since the rate constant is sufficiently small to induce only a small δf_j change in substituted base fraction during the δt = 0.0204 byr (20.4 myr) time period (see above), we can equate the δf_j change in the fraction of the base *j* (at the given position that had the original base *i* identity in a large population of homologous sequences) to the $p \approx \delta f_j \approx r\delta t$ substitution probability within δt period. We can thus make a crude estimation for the probability of the second substitution to another, $k \neq j$, base happening at the same position to be $(r\delta t)^2$, which is the product of individual substitution probabilities assuming that the rate constant does not change from our average estimate *r* across those two substitution types. We can permit this for the sake of the back-of-the-envelope estimation of the order of the effect expected from the repeated mutations hitting the

same site within 20.4 myr period. To this end, the δf_j^{app} apparent change in *i->j* substitution fraction that we would observe by neglecting the additional *j->k* substitution, would underestimate the more realistic δf_j and be equal to $\delta f_j^{app} = r\delta t - (r\delta t)^2$, as we would not count the *j* bases that emerged but became additionally substituted by *k*. Hence, the corresponding apparent rate constant that neglects second substitution would also underestimate the actual value *r*, and can be expressed as $r^{app} = \delta f_j^{app}/\delta t = (r\delta t - (r\delta t)^2)/\ \delta t = r(1-r\delta t)$. This means that the underestimation of the actual rate constant would be by $[r-r(1-r\delta t)]\times 100/r = 100\times r\delta t$ %. Putting 0.64 byr^{-1} for *r* and the 0.0204 byr for δt, we can expect only 1.3% contribution from the repeated second substitution at the same position within 20.4 myr. Since some of the other non-*j* bases could become *j*s and balance the underestimation of the δf_j fraction, the error could be even smaller. This shows that repeated substitutions can be safely ignored in 20.4 myr time-scale. Furthermore, the validity of our $r_{i,j}^{core}$ rate constants was further checked through the two independent analyses reflected in **Supplementary Fig. S3** and **Fig. S4**.

Finding the influence range of neighbour nucleotides. We have used generalised boosted models[63] (GBM) to elucidate the effective range for the core sequence-context effects. This was achieved by developing test models to evaluate the predictive strength of only the neighbouring bases in defining the core mutation rate of the central base. The GBM was used as implemented in the gbm library for R. For each *i->j* mutation type, all the found Trek data were taken without the possible outliers, which were filtered by allowing only the usage of the values that were within the 1.65 (a value that keeps ~90% data if normally distributed) times standard deviation range of the constants in a given mutation category. The sequences were then processed to produce pos/b_i uncoupled features that were associated with the relative adjacent positions (*pos*, - for upstream and + for downstream positions) and their possible four b_i base types. Those features took values 0 or 1, depending on whether the base at an associated relative position was of the b_i (1) or any other base type (0). For instance, if we wanted to develop a model accounting for only a single upstream (*pos* = -1) and a single downstream (*pos* = +1) nucleotides, hence predicting the mutation rates for different 3-mers, where the central base is the one that mutates, then we produced 8 pos/b_i features for the GBM fitting. There, 4 binary features (-1/A, -1/C, -1/G and -1/T) described whether the upstream -1 position is of base type A, C, G or T, and 4 binary features described the same for the downstream +1 position. We built the models using 3-, 5-, 7-, 9-, and 11-mers, thus accounting for 1, 2, 3, 4 and 5 upstream and the same number of downstream neighbour bases. The absence of the coupling in the binary features, unlike in the case where, for instance, one employs only two binary features per 4 states, enabled us to also investigate the predictive

significance of each nucleobase identity at a given neighbouring position, which was useful in deciding against the construction of more complex machine learning models (see below) using additional features with higher level of abstraction for the sequence information (overall base content, sequence-derivative properties). The GBM models were then fitted by systematically trying different permutations of the tuning values[63] for the number of trees (50-7500), interaction depth (1-10), shrinkage (0.001, 0.01 and 0.1), the number of minimum observations per node (1-28) and the bag fraction (0.25-0.65). The optimal combinations of the tuning parameters were found per mutation type and sequence length, via a 16-fold cross validation repeated 7 times. The found best parameters are accessible in the **Supplementary Data**, and the predictive performances of the best models, from the repeated cross validation studies, are presented in **Supplementary Fig. S5**. To make a predictor of $r_{i,j}^{core}$ based on a sequence, we found it much better to use direct values coming from the proposed Trek methodology, rather than the GBM models, as the Trek values were already well averaged across multiple occurrences of the same sequence in different loci of the human genome (**Fig. 1-3** and **Supplementary Fig. S3**). Furthermore, the overall poor performance of the GBM models implies that the influence of the immediate context is highly non-additive and non-Bayesian, which is expected taking into account the nature of the core context-dependent mutation rates. The latter rates reflect the intrinsic short-range sequence properties, interactions and recognition with the overall mutagenic and repair machinery present in a given organism. There, the whole sequence at a certain small scale[9] is what defines the interaction[11], and it is hard to represent such effects through even smaller-scale constituents. The direct model-free approach used in our Trek mapper methodology (see below) thus seems essential in mapping the $r_{i,j}^{core}$ rate constants throughout the human genome. To this end, the GBM models here had a sole purpose of identifying the optimal range of influence for accounting the neighbouring nucleotides. The optimal range was found to be captured, on average, by a 5-7-nt long window (**Supplementary Fig. S4** and **S5**) which is in an excellent agreement with the prior <10 nt estimate[9-11,45,46]. We used the maximum 7-nt length to stratify the Trek data for the further model-free mapping on any provided sequence, including the whole human genome.

Mapping the Trek mutation data on any sequence. We developed a Trek mapper program. For each *i* position in a query sequence (**Supplementary Fig. S6**), the program looks at the bases *i*-3 to *i*+3. If the exact 7-mer, with the associated rate constant values, is not available in the Trek database, the program reduces the size of the sequence to 5, by considering *i*-2 to *i*+2 positions, or, if necessary, to 3- or 1-mers, until an exact match is found in the database. This would essentially mean that some

reported mutation rates would come from the actual triad data. About half will come from pentads and some from heptads, accounting for more precise sequence-context information. A few will originate from the fully averaged single-base (1-mer) Trek mutation rates. Single-base values are also returned for the terminal positions in a query sequence. For each unique sequence in the discussed 7-, 5-, 3- and 1-mers, if the k-mer appears more than once in the reference L1 sequence, of course with different neighbours at the positions out of the k-mer range, we average the Trek values by taking the median. For instance, the $r_{G->A}$ mutation rate constant in the 3-mer AGT represents the rate averaged across all the appearances of AGT in the L1 reference sequence, which would normally be with varying other neighbour bases, out of the 3-mer range. The $r_{G->A}$ in AGT will therefore represent the average rate constant across all the representatives of the significant range, the NNAGTNN 7-mers, present in L1, where N can be any of the four bases. In the same way, the mutation rate constants for the single bases (1-mers) can be considered as fully averaged across all the possible neighbour effects in NNNGNNN sequences. Our algorithm therefore makes the most of transposon exposed mutation rate data of the human genome, returning the best possible values inferable from our Trek database and, where uncertainty is present, returning the best averaged values for a shorter context range. Furthermore, we have enabled an option to use symmetrised Trek parameterisation, assuming an overall strand-invariance of the mutation rates. In the latter case, the complementary mutation rate constants of the central bases in two reverse complementary k-mers were equalised. For example, the G->A mutation rate constant (r_1) in the 3-mer AGC was set equal to the complementary C->T rate constant (r_2) in the reverse complementary GCT. The data equalisation was done in the following way: if both r_1 and r_2 were of the same quality accounting for the whole sequence-context information in both 3-mer variants, then both values were set to $(r_1+ r_2)/2$; however, if one of the rate constants was determined with a better quality, since the full 3-mer data for the other case was missing and the 1-mer average was used as a replacement, then the rate constant of the better quality variant was assigned to both r_1 and r_2. Accounting for the strand symmetry improves the results of the validation studies, further refining the mutation rate constant values and increasing the coverage of longer k-mers in the Trek database.

The described Trek mapper program, along with the associated data can be accessed through the http://trek.atgcdynamics.org web page. Future improvements in data and the program, through extending the types of mobile DNA in the Trek procedure, will be reflected on the same web site. The Trek mapper server application was written in R, using the Shiny library and server application (http://shiny.rstudio.com).

Equilibration of a random DNA sequence with the Trek rates. A 5 million (mln) nt sized random genome was created with the initial A, T, G and C base contents set to 20, 20, 30 and 30% correspondingly, hence with 60% genomic G+C content. The length was selected to cope with the finite computational and time resources, though operating on lengthier sequences will not change the outcome of the calculations, since the captured sequence-context effects are within 7-nt window. We first calculated the probabilities of all the possible mutations in this random sequence, which basically meant the assignment of 3 mutation rate constants per position in the genome, describing the conversion into the three bases other than the base already present in the respective position. This was done using the Trek mapper described above. Next, at each step, we sampled 5000 mutations weighted by the calculated 3×5 mln rate constants. We then identified those 5000 positions and the corresponding mutation types that were sampled to happen (**Supplementary Fig. S7**), performed those mutations, and, updated the probability values via the Trek mapper. Repeated multiple times, the process evolved the sequence (**Supplementary Video 1**) ruled by the core spontaneous rate constants that are sensitive to the changes in the sequence composition at the immediate vicinity in the genome.

For the comparison of the simulated sequence at equilibrium with the real human genome (RefSeq), we calculated the fractions of different oligomers (k-mers) in both sequences (**Supplementary Data**). The k-mer contents of the human genome were calculated by sliding a window of size k (from 1 to 7) and counting the occurrence of each 4^k unique sequence. We used a direct calculation of the lexicological index[64] of a string to increase the computational efficiency of the k-mer counting. Although, data from the masked human genome were used in the k-mer analyses to rule out any bias from the presence of the same L1 elements in the object of application of Trek data, the comparison of the masked and unmasked genomes showed only negligible differences in both single base and short k-mer contents. If, however, we consider only the L1 elements, the k-mer content was quite different from the rest of the genome.

Basal mutability constants at cancer-linked sites. We took all the non-coding somatic single-nucleotide substitution data associated with cancer from the COSMIC database[55] (www.sanger.ac.uk/cosmic, NCV dataset accessed in February, 2015). Since our Trek rate constants are for the spontaneous mutations, we only considered the sites that were also not declared as known SNPs (the status was present in the NCV dataset). This was to ensure that we excluded sites where an active polymorphism is potentially encouraged by a natural selection. About 3.7% data from the remaining set of cancer-linked mutations were duplicates, with no differences found in genomic

location and mutation types. We removed those, keeping only the single first-encountered copies of such entries. The resulting data contained 5984711 mutation entries.

The cancer enrichment score (**Fig. 6**) for a given k-mer sequence was calculated by taking the ratio of the occurrence fractions, calculated in (numerator) only the cancer-linked sites (where the linked base is the central one in the k-mer) and (denominator) in the whole repeat-masked human genome (**Supplementary Data**).

Data access. The Trek $r_{i,j}^{core}$ database and the mapper are housed at http://trek.atgcdynamics.org. Additional figures and tables, referenced in the text, can be found in the **Supplementary Information** (**Table S1** and **Figures S1-S11**). We deposit the video showing the *in silico* dynamics of a random sequence upon equilibration (**Supplementary Video 1**), the raw data on the mutation rate constants in the reference L1 sequence (**Supplementary Data 1**), the resulting Trek database processed with (**Supplementary Data 2**) and without (**Supplementary Data 3**) the strand-symmetry considerations, the k-mer content for the masked and unmasked human genomes (**Supplementary Data 4**), the full set of 7-mer sequences with the respective cancer enrichment scores and mutability values (**Supplementary Data 5**), and the GBM parameters that were minimising the error of the tree-based test models (**Supplementary Data 6**). All the associated method and analyses source codes are available from the authors upon request.

REFERENCES

1. Lynch, M. *The origins of genome architecture*. (Sinauer Associates Inc., 2007).
2. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304 (2000).
3. Chen, J.-Q. *et al*. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol.* **26,** 1523–1531 (2009).
4. 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).
5. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 961–968 (2010).
6. Denver, D. R., Morris, K., Lynch, M., Vassilieva, L. & Thomas, K. High direct estimate of the mutation rate in the mitochondrial genome of caenorhabditis elegans. *Science* **289,** 2342–2344 (2000).
7. Lynch, M. *et al*. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 9272–9277 (2008).
8. Zhu, Y. O., Siegal, M. L., Hall, D. W. & Petrov, D. A. Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E2310–8 (2014).

9. Silva, J. C. & Kondrashov, A. S. Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet*. **18,** 544–547 (2002).
10. Ellegren, H., Smith, N. G. & Webster, M. T. Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Devel*. **13,** 562–568 (2003).
11. Zavolan, M. & Kepler, T. B. Statistical inference of sequence-dependent mutation rates. *Curr. Opin. Genet. Devel*. **11,** 612–615 (2001).
12. Sved, J. & Bird, A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. USA* **87,** 4692–4696 (1990).
13. Jiang, C. & Zhao, Z. Directionality of point mutation and 5-methylcytosine deamination rates in the chimpanzee genome. *BMC Genomics* **7,** 316 (2006).
14. Supek, F., Lehner, B., Hajkova, P. & Warnecke, T. Hydroxymethylated cytosines are associated with elevated C to G transversion rates. *PLoS Genet*. **10,** e1004585 (2014).
15. Majewski, J. & Ott, J. Distribution and characterization of regulatory elements in the human genome. *Genome Res*. **12,** 1827–1836 (2002).
16. Hellmann, I. *et al*. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res*. **13,** 831–837 (2003).
17. Fryxell, K. J. & Moon, W.-J. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol*. **22,** 650–658 (2005).
18. Mugal, C. F. & Ellegren, H. Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. *Genome Biol* **12,** R58 (2011).
19. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet*. **18,** 337–340 (2002).
20. Duret, L. & Arndt, P. F. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. **4,** (2008).
21. Hanawalt, P. C. & Spivak, G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat. Rev. Mol. Cell Biol*. **9,** 958–970 (2008).
22. Gaillard, H., Herrera-Moyano, E. & Aguilera, A. Transcription-associated genome instability. *Chem. Rev*. **113,** 8638–8661 (2013).
23. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504–507 (2012).
24. Agier, N. & Fischer, G. The mutational profile of the yeast genome is shaped by replication. *Mol. Biol. Evol*. **29,** 905–913 (2012).
25. Reijns, M. A. M. *et al*. Lagging-strand replication shapes the mutational landscape of the genome. *Nature* **518,** 502–506 (2015).
26. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521,** 81–84 (2015).
27. Ellison, C. E. & Bachtrog, D. Non-allelic gene conversion enables rapid evolutionary change at multiple regulatory sites encoded by transposable elements. *Elife* **4,** e05899 (2015).
28. Ellegren, H. Characteristics, causes and evolutionary consequences of male-biased mutation. *Proc. Biol. Sci*. **274,** 1–10 (2007).
29. Subramanian, S. & Kumar, S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res*. **13,** 838–844 (2003).
30. Chamary, J. V., Parmley, J. L. & Hurst, L. D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7,** 98–108 (2006).
31. McVean, G. T. & Hurst, L. D. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* **386,** 388–392 (1997).
32. Martincorena, I. & Luscombe, N. M. Non-random mutation: the evolution of targeted hypermutation and hypomutation. *Bioessays* **35,** 123–130 (2012).

33. Kazazian, H. H., Jr. *Mobile DNA. Finding treasure in junk*. (Pearson Education, 2011).
34. Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. USA* **101,** 13994–14001 (2004).
35. Boissinot, S., Chevret, P. & Furano, A. V. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.* **17,** 915–928 (2000).
36. Khan, H. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* **16,** 78–87 (2006).
37. Lee, J. *et al*. Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene* **390,** 18–27 (2007).
38. Giordano, J. *et al*. Evolutionary history of mammalian transposons determined by genome-wide defragmentation. *PLoS Comput. Biol.* **3,** e137 (2007).
39. Lander, E. S. *et al*. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
40. Medstrand, P., van de Lagemaat, L. N. & Mager, D. L. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res.* **12,** 1483–1495 (2002).
41. Rawal, K. & Ramaswamy, R. Genome-wide analysis of mobile genetic element insertion sites. *Nucl. Acids Res.* **39,** 6864–6878 (2011).
42. Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M. & Neretti, N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15,** 583 (2014).
43. Arndt, P. F., Petrov, D. A. & Hwa, T. Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. *Mol. Biol. Evol.* **20,** 1887–1896 (2003).
44. *The phylogenetic handbook: a practical approach to the phylogenetic analysis and hypothesis testing*. (Cambridge University Press, 2012).
45. Zhao, Z. & Boerwinkle, E. Neighboring-nucleotide effects on single nucleotide polymorphisms: A study of 2.6 million polymorphisms across the human genome. *Genome Res.* **12,** 1679–1686 (2002).
46. Nevarez, P. A., DeBoever, C. M., Freeland, B. J., Quitt, M. A. & Bush, E. C. Context dependent substitution biases vary within the human genome. *BMC Bioinformatics* **11,** 462 (2010).
47. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18,** 6097–6100 (1990).
48. Eyre-Walker, A. & Hurst, L. D. The evolution of isochores. *Nat Rev Genet* **2,** 549–555 (2001).
49. Costantini, M., Clay, O., Auletta, F. & Bernardi, G. An isochore map of human chromosomes. *Genome Res.* **16,** 536–541 (2006).
50. Tomasetti, C. & Vogelstein, B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347,** 78–81 (2015).
51. Hodgkinson, A., Chen, Y. & Eyre-Walker, A. The large-scale distribution of somatic mutations in cancer genomes. *Hum. Mutat.* **33,** 136–143 (2012).
52. Alexandrov, L. B. *et al*. Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
53. Kandoth, C. *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).
54. Jia, P., Pao, W. & Zhao, Z. Patterns and processes of somatic mutations in nine major cancers. *BMC Medical Genomics* **7,** 11 (2014).
55. Forbes, S. A. *et al*. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucl. Acids Res.* **39,** D945–50 (2011).

56. Cooper, D. N. & Krawczak, M. The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. *Hum Genet* **85,** 55–74 (1990).
57. Greenman, C. *et al*. Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
58. Rubin, A. F. & Green, P. Mutation patterns in cancer genomes. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 21766–21770 (2009).
59. Walser, J. C., Ponger, L. & Furano, A. V. CpG dinucleotides and the mutation rate of non-CpG DNA. *Genome Res*. **18,** 1403–1414 (2008).
60. Kimura, M. *The neutral theory of molecular evolution*. (Cambridge University Press, 1983).
61. Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting natural selection in genomic data. *Annu. Rev. Genet*. **47,** 97–120 (2013).
62. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0,* 2013-2015, at http://www.repeatmasker.org.
63. Kuhn, M. & Johnson, K. *Applied predictive modeling*. (Springer, 2013).
64. Compeau, P. & Pevzner, P. *Bioinformatics algorithms: an active learning approach*. (Active Learning Publishers, 2014).
65. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc*. **74,** 829–836 (1979).

ACKNOWLEDGEMENTS

This research was supported by the Herchel Smith Fund. SB is a Wellcome Trust Senior Investigator. We thank Dr. Chris Lowe for proofreading the manuscript.

AUTHOR CONTRIBUTIONS

A.B.S and S.B. designed the study, performed the research, interpreted the results and wrote the paper. The authors declare no competing financial interests.