
Inherent instability of simple DNA repeats shapes an evolutionarily stable distribution of repeat lengths

Received: 22 January 2025

Accepted: 13 November 2025

Cite this article as: McGinty, R.J., Balick, D.J., Mirkin, S.M. *et al.* Inherent instability of simple DNA repeats shapes an evolutionarily stable distribution of repeat lengths. *Nat Commun* (2025). <https://doi.org/10.1038/s41467-025-66725-5>

Ryan J. McGinty, Daniel J. Balick, Sergei M. Mirkin & Shamil R. Sunyaev

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

1 Inherent instability of simple DNA repeats shapes an evolutionarily stable distribution of repeat
2 lengths

3

4 Ryan J. McGinty^{1,*}, Daniel J. Balick^{1,*}, Sergei M. Mirkin^{2,†} & Shamil R. Sunyaev^{1,†}

5

6

7 ¹ Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

8 ² Department of Biology, Tufts University, Medford, MA, USA

9 [†] Corresponding authors: Sergei Mirkin (Sergei.Mirkin@tufts.edu), Shamil Sunyaev
10 (ssunyaev@hms.harvard.edu).

11 * These authors contributed equally to this work.

12

13

14 Abstract

15 Using the Telomere-to-Telomere reference, we assemble the distribution of simple repeat lengths
16 present in the human genome. Analyzing over three hundred mammalian genomes, we find
17 remarkable consistency in the shape of the distribution across evolutionary epochs. All observed
18 genomes harbor an excess of long repeats, which are potentially prone to developing into repeat
19 expansion disorders. We measure mutation rates for repeat length instability, quantitatively
20 model the per-generation action of mutations, and observe the corresponding long-term behavior
21 shaping the repeat length distribution. We find that short repetitive sequences appear to be a
22 straightforward consequence of random substitution. Evolving largely independently, longer
23 repeats (above roughly 10 nt) emerge and persist in a rapidly mutating dynamic balance between
24 expansion, contraction and interruption. These mutational processes, collectively, are sufficient
25 to explain the abundance of long repeats, without invoking natural selection. Our analysis
26 constrains properties of molecular mechanisms responsible for maintaining genome fidelity that
27 underlie repeat instability.

28

29

30 **Introduction**

31 Over 2.5% of human genomic DNA consists of simple DNA repeats¹. Also known as short
32 tandem repeats (STRs) or microsatellites, simple repeats consist of direct tandem repetitions of
33 short sequence motifs, e.g., mononucleotides, dinucleotides, trinucleotides and so forth. In a
34 randomized DNA sequence, the probability of encountering a simple repeat is exponentially
35 decreased with increasing tract length. Yet this relationship fails to predict the enormous
36 overrepresentation of long simple repeats in most genomic sequences, including in humans^{2,3,4}.
37 The origin of this overrepresentation remains to be elucidated.

38 This overrepresentation is even more striking in light of the existence of repeat expansion
39 disorders, a growing list of severe human diseases caused by disruption of gene function due to
40 long STRs^{5,6}. Decades of study have demonstrated that repeat tract lengths vary between and
41 within individuals⁷, owing to frequent expansion and contraction mutations. The rate of these
42 mutations increases with the length of a repeat, a phenomenon known as repeat length
43 instability⁸. Length instability is commonly ascribed to DNA strand slippage during replication
44 and/or DNA repair, although a variety of other molecular mechanisms can also contribute⁸.
45 Instability rates differ between various repeat motifs, being particularly pronounced for motifs
46 that form non-B DNA secondary structures⁹. Importantly, when repeat length exceeds a threshold
47 of approximately 75–90 nt, carriers frequently transmit a substantially longer repeat to the next
48 generation. Known as ‘genetic anticipation’, this effect continues to compound in subsequent
49 generations, which leads to more severe presentation and/or earlier age of onset⁶. Recently
50 developed techniques such as ExpansionHunter¹⁰ and long-read sequencing have accelerated the
51 discovery of pathogenic repeats; in particular, the growing number of repeat expansion disorders
52 mapped to introns and other non-coding regions sheds light on repeat disease biology beyond
53 coding regions. Repeat expansions are also observed in various cancers^{11,12,13} and serve as
54 hotspots for genomic rearrangements¹⁴.

55 While numerous studies focus on the instability of disease-length repeats, comparatively less
56 is known about shorter repeats, including the so-called ‘long-normal’ alleles that sit immediately
57 below the disease-length threshold. Carriers of long-normal repeat alleles are healthy, but risk
58 transmitting a disease-length allele due to the higher rate of repeat expansion; additionally, some
59 long-normal alleles contain protective interruptions that, if lost, result in reversion to disease

length⁶. Complementing our understanding of long disease-causing repeats, a recent finding identified an autosomal dominant thyroid disorder linked to a (TTTG)₄ repeat, with a recurrent deletion to (TTTG)₃ in affected individuals¹⁵. Additionally, instability of A₈ and C₈ repeats in the coding sequences of mismatch repair genes *MSH3* and *MSH6*, respectively, promotes tumor adaptability via frequent frameshifts and subsequent reversions¹⁶. The latter examples suggest that relatively short repeats, which comprise a much larger portion of the genome, also have biomedical relevance.

In light of the rapidly growing list of repeat-associated diseases, it is surprising to find repeats harbored in abundance in the genome. Interest in this discrepancy goes back at least three decades² and has led to speculation that natural selection preserves longer repeat lengths despite the risk of disease¹⁷. The best-supported examples of functionality are specific to telomeric and centromeric repeats^{9,17,18}, though some recent studies have suggested that simple repeats play a role in gene regulation¹⁷. However, before assuming the overabundance of repeats is evidence of functionality, a more basic explanation should be considered: the excess of repeats in the genome is solely a consequence of mutational processes. Several studies, largely pre-dating the human genome era, considered this premise, but were limited by the availability of sufficiently long genome sequences, lacked robust direct measurements of repeat instability, and/or considered oversimplified mutational models^{4,19–34}. Indeed, all such studies of simple repeats have been limited by long-standing technical challenges to sequencing repetitive regions^{35,36,37,38}. Technological developments led to the release of the human Telomere-to-telomere genome (T2T-CHM13), which more than doubled the number of mapped simple repeats compared to the previous reference genome GRCH38¹. This warranted a fresh look at the distribution of repeat lengths and whether mutational processes, in the absence of selection, can explain their abundance.

In this study, we measured genome-wide distributions of repeat lengths across mammals, observing that the distribution, including the prevalence of long repeats, is remarkably stable over evolutionary timescales. We modeled the effects of repeat length instability on evolution of the distribution, finding that the observed repeat length distribution can emerge and be maintained solely due to the interplay between distinct mutational processes. After incorporating empirical estimates and inference of repeat length instability rates, the most parsimonious

90 explanation for the abundance and stability of long repeats does not require invoking selection;
91 rather, extreme mutation rates cause long repeats to emerge as independently evolving elements.
92 We discuss how inherent constraints of DNA replication and repair machineries could lead to
93 persistent repeat length polymorphism.

94

95 Results

96 *Features of the repeat length distribution and evolutionary stability*

97 Using T2T-CHM13, we first assembled a genome-wide Distribution of Repeat tract Lengths
98 (henceforth, DRL) for each simple tandem repeat motif, pooling over bioinformatically
99 indistinguishable permutations (see **Supplementary Fig. 1a, Methods**). Each distribution was
100 assembled by counting contiguous, uninterrupted repetitions of a specified motif, allowing for
101 straightforward bioinformatic assembly of the DRL (see **Methods**). Each DRL showed a marked
102 excess of repeats longer than ~10 nt., relative to a randomly shuffled control. This was apparent
103 for nearly all motifs (**Supplementary Fig. 1b**) but with motif-specific variation in the shape of
104 the extended tail of long tract lengths. **Fig. 1a** plots the DRLs after pooling motifs of the same
105 unit length (e.g., mononucleotide repeats, dinucleotide repeats, etc. up to hexamer repeats), each
106 with a clear tail of long repeats. We found that short read sequencing was sufficient to
107 reconstruct the well-populated length classes of nearly all DRLs, lacking estimates only for the
108 very longest repeats (**Supplementary Fig. 1c**). We were therefore able to estimate distributions
109 from genome sequences of over 300 mammals from the Zoonomia project³⁹ and compare them
110 to humans. Due to differences in total assembly length, direct comparison was performed on the
111 normalized DRL (see **Supplementary Fig. 2, Methods**). There was surprisingly little variation
112 in the shape of the DRLs between primates; DRL shapes were qualitatively similar but more
113 variable in mammalian DRLs, consistent with the longer divergence time. This comparison is
114 shown in **Fig. 1b** for mono-A/mono-T repeats, which are the most prevalent in the human
115 genome and are the primary focus of our subsequent analyses (normalized DRLs for additional
116 motifs shown in **Supplementary Fig. 3**). Consistency of the shape of the DRLs across the
117 primate lineage suggests that both the repeat tract length distributions and, as a corollary,
118 maintenance of the underlying mechanisms, were largely stable for at least 70 million years. This
119 highly conserved DRL evolution directly suggests the emergence of a steady state equilibrium.

120 The empirical DRLs extend to lengths that, at disease loci, would be subject to genetic
121 anticipation, risking progression to repeat expansion disorders in subsequent generations⁵;
122 despite the associated disease risk, this tail of long repeats appears to be a generic and
123 evolutionarily conserved feature of repeat length distributions. One proposed explanation is that
124 longer repeats confer a selective advantage due to some repeat length-specific biological
125 function¹⁷. As an alternative, we propose that long repeats emerge and are maintained by the
126 complex interplay between distinct mutational forces. Though these hypotheses are not mutually
127 exclusive, we sought to understand the extent to which mutagenesis, alone can maintain the
128 shape of the distribution, without introducing natural selection.

129

130 *Mutational transitions in repeat tract length*

131 As described above, literature suggests that repeat instability emerges as a very rapid increase in
132 the rate of length changes as tract length increases, with the longest repeats mutating nearly
133 every generation. In light of such high mutation rates, the observation that the DRL evolves in
134 steady state over long timescales is somewhat surprising, suggesting that the maintenance of the
135 distribution results from a dynamic balance between the ensemble of mutational processes that
136 alter repeat length.

137 To better understand the genome-wide distribution, we therefore require a comprehensive
138 understanding of all involved mutational processes (e.g., nucleotide substitutions, insertion,
139 deletion; see **Fig. 2** for schematic of mutational processes) and how they differ by repeat length.
140 Estimating repeat tract lengths from sequencing data is a notorious bioinformatic challenge,
141 particularly for homopolymer repeats^{35,36,37,38}. Published results only sparsely cover the full
142 range of lengths observed in the genome, largely focusing on disease relevant lengths and loci⁴⁰—
143 ⁴⁶. In contrast, there is little information about mutation rates at short tract lengths despite
144 comprising the vast majority of repeats in the genome.

145 In order to study mutations across a wide range of tract lengths, we first subdivided
146 insertions and deletions into repeat-relevant mutational categories; we refer to expansions and
147 contractions as mutations that alter repeat length by whole motif units and maintain one
148 contiguous repeat tract, in contrast to partial deletions and non-motif insertions. Rates of each

149 mutagenic process were estimated by pooling existing short-read trio sequencing datasets
150 ($n=9,387$ trios; henceforth, ‘pooled trio’ dataset). This data was sufficient to directly estimate
151 length-dependent rates for short repeats (up to roughly $L=6-8$ units, depending on motif, where L
152 is the number of repeated units in a tract) but observed that sequencing errors dramatically
153 reduced mutation counts for longer repeat tract lengths (see **Methods**). We complemented these
154 estimates by length-stratifying data from a recent study⁴⁶ that used a population structure-aware
155 caller (named ‘popSTR’) to study repeat mutations in the mid-to-long length range in short-read
156 trio data ($n=6,084$). Due to a variety of technical considerations (see **Methods**), estimates were
157 only reliable within a limited range of tract lengths for each motif, which differed by dataset.

158 It was previously observed that the majority of mutations within a repeat increase or decrease
159 length by one unit (i.e., $L \rightarrow L \pm 1$)^{47,48}. To expand on this, we length-stratified the mutation data
160 and found that single-unit length changes dominate above a clear length threshold, consistent
161 with the onset of repeat instability (**Fig. 3a**, **Supplementary Fig. 4**). Accordingly, we estimated
162 the rates of single-unit length changes, separately estimating the contraction, expansion, and non-
163 motif insertion rates from the pooled trio data; popSTR-based estimates combine expansion and
164 non-motif insertion rates due to technical limitations (see **Methods**). For mono-A repeats, all
165 instability rates increase rapidly between roughly 5-10 nt (**Fig. 3b**, see **Supplementary Fig. 5** for
166 all motifs), consistent with a threshold-like onset of repeat instability in this length range
167 (detailed below). The popSTR-based estimates suggest that repeat instability rates continue to
168 increase monotonically, at least until the length range where the data becomes noisy (**Fig. 3b**,
169 **Supplementary Fig. 5**).

170 The combination of both datasets recapitulates the hallmark of repeat instability^{5,6,8,9}: a rapid
171 increase in the rates of expansion and contraction as length increases. Beyond confirming this
172 property, available data was insufficient to robustly estimate the length dependence of each
173 mutational process across the tract length range observed in the genome. Such an estimate is a
174 necessary component of any quantitative understanding of the approach of the DRL towards the
175 steady state (as observed in primates). For further analyses, the length dependence of these
176 mutation rates was extended to longer tract lengths by parameterization in an inference
177 framework described below.

178

179 *Computational modeling of DRL dynamics*

180 We sought to assess whether our three observations of the empirical human DRL, the
 181 existence of a long-term steady state, and the estimated repeat instability could be simultaneously
 182 incorporated into a self-consistent model of repeat length evolution. To this end, we built a
 183 computational model that incorporates length-changing effects of substitutions, expansions,
 184 contractions, and non-motif insertions to follow the evolution of the DRL towards a steady state.
 185 We modeled the distribution of mono-A repeats, in part, because they are subject to the simplest
 186 ensemble of mutational changes in length (related to the lack of distinction between tract length
 187 and the number of repeated units). As a consequence of counting of contiguous repeats to
 188 assemble the DRL, mutational processes alter the length of a given repeat in one of four ways
 189 (**Fig. 2**): lengthening, shortening, joining two repeats into one (which we term ‘fusion’), or
 190 splitting one repeat into two (i.e., repeat interruption, which we term ‘fission’). This treatment of
 191 interruptions as effectively splitting a repeat is consistent with previous observations that
 192 interruptions result in locus-wide rates scaling with the longest contiguous subunit^{49–55} or,
 193 alternatively, a rate reduction that scales with distance from the repeat boundary⁷.

194 To reduce the computational time required to evolve a whole genome sequence and
 195 simultaneously count contiguous repeat tracts, we directly evolved the DRL by manipulating the
 196 occupancy of each length bin. In this formulation, length-altering mutations are reframed as
 197 transitions between length bins and the DRL evolves under repeated application of mutations
 198 over many generations. However, the elementary step of the process is deceptively complex, as
 199 repeat fusion precludes framing the mutational process as a standard transition rate matrix and
 200 both fission and fusion are non-conservative transitions. We treated the aggregate mutational
 201 effects as deterministic, ignoring stochasticity in the mutational process and due to factors like
 202 genetic drift, to approximate the expectation of the DRL at late times. We interpret this late-time
 203 expectation as an approximation to the steady-state distribution, if one exists, resulting from the
 204 modeled mutational processes.

205

206 *Bayesian inference and parametric model comparison*

207 We constructed a Bayesian inference procedure (**Fig. 4**) to constrain properties of repeat
 208 instability consistent with the steady-state evolution of the observed human DRL (see **Methods**).
 209 The computational model uses explicit length-dependent rates for each mutational process as
 210 inputs. We directly incorporated the subset of estimated rates from the pooled trio data shown in
 211 **Fig. 3b** (i.e., for $L=1-8$); the instability rate curves were then extended to longer lengths to
 212 model a rapid, monotonic increase with length. The inclusion of empirical estimates at low
 213 lengths, which include a rapid rate increase, limits the number of parameters required to describe
 214 complex length-dependent rates of repeat instability with both a rapid transition and a distinct
 215 asymptotic functional form for long repeats.

216 Resemblance to the monotonic increase seen in popSTR estimates at intermediate lengths
 217 (**Fig 3b**) motivated a class of parameterizations with power-law increase in the rates (see **Table**
 218 **1, Methods**). We first specified a model with minimal degrees of freedom that describes equal
 219 rates of expansion and contraction (i.e., rates for $L>8$ represent a simplistic model of replication
 220 slippage based on previous literature^{8,20,23,28}). This was treated as a null model for comparison to
 221 parameterizations with additional degrees of freedom that characterize expansion-contraction
 222 bias. We used the popSTR-based rate estimates to define plausible, empirically based Bayesian
 223 priors for each parameter space, representing varying degrees of confidence in this dataset
 224 (**Supplementary Fig. 6**), including a (naively uninformed) uniform prior. We then used the
 225 results of our computationally modeled DRLs in an Approximate Bayesian Computation
 226 framework (following the prescription in Wilkinson, 2013⁵⁶) to compute a posterior probability
 227 distribution for each parametric model. We used the range of primate DRLs to define a rejection
 228 probability by comparing them to the human DRL (using Kullback-Leibler divergence, which
 229 quantifies the difference between two distributions). The same quantity was computed for each
 230 computationally modeled DRL and used to approximate the posterior probability distribution
 231 over the parameter space (see **Methods** for details).

232 The results of our inference for each parameterization are summarized in Table 1. We
 233 assessed the relative statistical support for various model comparisons via the Bayes factor ratio
 234 (see **Methods**). The relative Bayes factors strongly suggested discarding the two-parameter null
 235 model in favor of further degrees of freedom that introduce asymmetry (i.e., bias) between the
 236 expansion and contraction rates. The power-law parameterization with the largest Bayes factor,

237 regardless of prior, was a three dimensional model of expansion and contraction with distinct
238 exponents and related multiplicative constants (see **Table 1, Methods**). This model is a modest
239 improvement over the four-parameter description (i.e., completely decoupled expansion and
240 contraction rates), which otherwise provides a dramatically better description than lower-
241 dimensional parameterizations.

242 To test how reliant these conclusions are on the power-law functional form, we defined an
243 alternate class of parameterizations with logarithm-based growth rates (i.e., with slower-growing
244 rates at large lengths to better approximate saturation). Model comparison within this class of
245 parameterizations provided qualitatively consistent results to those comparing power law
246 parameterizations (see **Table 1**). The inequivalence of priors across functional forms (along with
247 additional necessary approximations; see **Methods**) suggests caution should be taken in direct
248 comparisons between models with distinct functional forms. Due to the relative analytic
249 simplicity of the functional form, subsequent analyses were focused on power-law
250 parameterization results.

251

252 *Inference of instability rates from the steady-state repeat length distribution*

253 Amongst power-law models, we focused on the three-parameter multiplier-coupled model, as it
254 showed the strongest statistical support, regardless of choice of prior. Above $L=8$, this model is
255 parameterized by exponents, τ_ϵ and τ_κ , and a common multiplier m (representing a discrete
256 jump in rates immediately above the empirical estimates; explicit definition in **Methods**), which
257 together characterize the length dependence of expansion $\epsilon(L; m, \tau_\epsilon)$ and contraction $\kappa(L; m, \tau_\kappa)$
258 (see **Methods, Table 1** for full definitions). The common multiplier for expansion and
259 contraction, which limits the dimensionality, may be interpreted as representing the onset of
260 repeat instability due to some common biological mechanism. Our trio rate estimates rapidly rise
261 in the length range where they lose accuracy; m , which describes a potentially dramatic jump
262 immediately above this range, can provide an oversimplified characterization of a rapid
263 transition to power-law like behavior. To limit further degrees of freedom, we assumed that the
264 length dependence of non-motif insertions is dictated by τ_ϵ , the expansion rate exponent, due to
265 their parallel increase in de novo rates (**Fig. 3b**) and because they likely arise from the same
266 biological mechanism (e.g., synthesis of the inserted nucleotides by an error-prone polymerase).

267 The parameter space we explored includes the possibility of a constant per-nucleotide rate (i.e., τ
 268 =0, analogous to the constant per-nucleotide substitution rates), linearity (i.e., $\tau=1$), a natural
 269 conceptual model for length dependence associated with repeat instability, and more rapid
 270 growth on par with popSTR-based estimates (**Fig. 3b**). However, the parameterization itself is
 271 not intended to represent a specific biological model; the true rate curves are likely more
 272 complex due to multiple contributing mechanisms.

273 To interpret the resulting posteriors, we first approximated highest density regions
 274 (HDRs) comprising 68%, 95%, and 99.7% of posterior probability on the finite grid for the
 275 multiplier-coupled model (**Fig. 5a**, **Supplementary Fig. 7a**; alternate parameterizations shown
 276 in **Supplementary Figs. 8,9**). The posterior is largely localized along a ridge of constant values
 277 of $\Delta\tau \equiv \tau_\kappa - \tau_\epsilon$ (roughly, $\Delta\tau \approx 0.3 - 0.6$) and roughly between multipliers $m=1-4$, with larger
 278 average exponents for smaller multipliers (i.e., anticorrelation between m and τ values for the
 279 most probable parameters); the difference between expansion and contraction rate exponents
 280 appears to be more relevant than their specific values. Under the uniform prior, some parameter
 281 combinations within the 95% HDR deviate from this range of $\Delta\tau$ (extending to both lower values
 282 of $\Delta\tau$ and lower $\tau_\epsilon, \tau_\kappa$; **Fig. 5a**) but are excluded when applying the popSTR-based prior, which
 283 requires consistency with the larger estimated instability rates (**Fig. 5a**, **Supplementary Fig. 7a**).

284 We computed the posterior-weighted DRL (i.e., the expectation value of the DRL; see
 285 **Methods**) and the range of DRLs consistent with the 95% HDR (**Fig. 5b**). The posterior-
 286 weighted DRLs closely resemble the human genome-wide distribution, while the 95% HDR
 287 parameters roughly span the range of primate DRLs used in our inference procedure. This
 288 demonstrates that the coarse features of the empirical DRL can be recapitulated from mutational
 289 dynamics alone.

290 We then computed the posterior-weighted length-dependent rates of expansion and
 291 contraction for each prior and found rough consistency with popSTR-estimated rates (**Fig. 5c**).
 292 One salient feature emerged, regardless of prior: expansion bias at intermediate tract lengths
 293 transitions to contraction bias at longer lengths due to the faster increase in contraction rate with
 294 length (i.e., $\tau_\kappa > \tau_\epsilon$; see **Figs. 5a,c**). This likely explains the preference for the multiplier-
 295 coupled model, which necessarily inherits a modest initial expansion bias directly from empirical
 296 rate estimates. However, if the apparent expansion bias at $L=8$ is simply a consequence of

297 homopolymer sequencing errors, correcting the direction of this bias would lead to statistical
 298 rejection of the multiplier-coupled model in favor of the four-parameter model (which lacks the *a*
 299 *priori* imposition of initial expansion bias on the parametrized rates). Regardless, inference
 300 results under the four-parameter model recapitulate the importance of a transition from
 301 expansion to contraction bias (**Supplementary Fig. 10**).

302 To gain intuition for the preference of expansion-to-contraction biased parameters in the
 303 multiplier-coupled model, we contrasted the DRLs to the parameter combinations outside of the
 304 95% HDR. Excluding slowly evolving rates, the remaining parameter space broadly separates
 305 into three qualitative categories, largely characterized by $\Delta\tau$: $\Delta\tau \gtrsim 0.6$ yields DRLs that
 306 underestimate the long repeat tail (i.e., early truncation), while $0 < \Delta\tau \lesssim 0.2$ yields distributions
 307 that overestimate the long tail (**Supplementary Fig. 11**). The roughly half of parameter space
 308 with $\Delta\tau < 0$ showed a clear reason for near-zero posterior probabilities: regardless of multiplier,
 309 these parameter combinations do not converge to steady state at late times and are subject to
 310 explosive growth in all length bins (**Supplementary Figs. 11, 12a**).

311 In addition to $\Delta\tau$ values outside of the high posterior ridge, larger values of m generally
 312 remained beyond the 95% HDR. This was likely due, in part, to the discrete jump between
 313 estimated and parameterized rates at $L=9$ that results in a discontinuous DRL, which can
 314 artificially inflate the Kullback-Leibler divergence. To investigate this, we repeated our inference
 315 after smoothing the instability rate length dependences via interpolation around this transition
 316 (**Supplementary Fig. 8**). The results suggested that a (naively more realistic) interpolated length
 317 dependence results in less penalization for larger multipliers and posterior probabilities more
 318 robust to the choice of prior. Additionally, inference using interpolated rates under the restrictive
 319 informative prior results in posterior-weighted instability rate estimates that overlap the popSTR
 320 estimates (**Supplementary Fig. 8c**); this suggests a smoother length dependence may represent
 321 more realistic instability rates. Indeed, this mutational model incorporates all available data and
 322 describes a self-consistent picture of a steady-state DRL shaped only by mutational dynamics.

323 To better understand the approach to steady state for realistic parameters within the 95%
 324 HDR, we followed the temporal evolution of the DRL, starting from a highly diverged initial
 325 state (see **Methods**, **Supplementary Fig. 12b**). This analysis suggested a two-stage equilibration
 326 process with two distinct timescales. The bulk of the long repeat tail establishes exponentially

327 quickly, followed by a slower fine-scale equilibration of mutational processes at each length.
328 Finally, we tested for robustness to potential confounders (e.g., differing initial conditions, use of
329 a step-wise speed-up factor, lack of stochastic fluctuations, etc.) and found no major changes in
330 the qualitative results (see **Methods**, **Supplementary Fig. 13**). Collectively, these results show
331 that mutational dynamics, rather than natural selection, may be responsible for maintenance of an
332 excess of mid-to-long tract length repeats in the human genome.

333

334 *Maintenance of the repeat length distribution in steady state*

335 To understand the complex interplay between mutational processes that shapes and stabilizes the
336 distribution of repeat lengths, we constructed an analytic model of the dynamics. This analytic
337 approximation captures the behavior of the DRL after the mutational process reaches steady state
338 (see **Methods** and **Supplementary Note**), focusing primarily on the previously described
339 multiplier-coupled three-dimensional parameterization. A number of previous studies have
340 constructed mathematical models of repeat instability to study repeat length
341 evolution^{20,23,26,27,28,29,30,32,34,57}, including a notable study by Lai and Sun³¹ that incorporates many
342 of the elements detailed herein. However, the combination of empirical rate estimates, a robust
343 genome assembly, and our phylogenetic observations motivated the construction of a model from
344 first principles that is directly informed by this collection of observations. In addition to
345 differences in mathematical machinery, the analytic construction differs from previous efforts by
346 incorporating pervasive length-dependent expansion-contraction bias (**Fig. 3b**, **Supplementary**
347 **Fig. 5**) and explicit effects from non-motif insertions.

348 We first constructed a discrete equation for the change in the number of repeats at a given
349 length in a single generation due to the deterministic action of mutations (i.e., in the absence of
350 selection and stochasticity in the mutational process, consistent with our computational model).
351 We then imposed a steady state condition by requiring that the sum of all changes in and out of
352 each length class vanishes at each time step after equilibration. Despite the simplifying
353 assumption of steady state, the full dynamical equation cannot be solved generically. However,
354 our estimates of de novo mutation rates suggested a dichotomy exists in the primary driver of
355 changes in length between short and longer repeats (i.e., primarily substitutions for $L < 8$ A-
356 mononucleotide repeats vs. expansions and contractions for $L > 10$; see below for direct inference

357 of this length range). Accordingly, short and long repeat dynamics can be treated as separable
358 (i.e., under the approximation of a separation of repeat length scales), leading to simpler
359 approximations of both length regimes. Transitions between the short and long repeat regimes,
360 while present, remain negligible in all realistic scenarios (see **Supplementary Note**).

361 For short repeats, we treated indel mutations as negligible and showed that a geometric
362 distribution (see **Methods, Equation 8**) exactly solves the steady state equation under two-way
363 substitutions alone (see **Supplementary Note**). For longer repeats, we constructed a partial
364 differential equation (PDE) that approximates the discrete equation and studied its time-
365 independent properties in steady state; dynamical equations are derived in the **Supplementary**
366 **Note** in terms of generic parameterization of the length-dependent instability rates. Focusing on
367 the multiplier-coupled parameterization, we obtained numerical solutions to the steady-state
368 dynamical equations under various approximations and under the assumption that fusion-based
369 contributions are negligible to long repeats (**Methods, Equations 9—11; Supplementary Note**).
370 These solutions, along with the geometric distribution for short repeats, accurately describe the
371 late-time DRLs produced by our computational model across the range of parameters that
372 approach a steady state (**Fig. 6; Supplementary Figs. 14—17; Supplementary Note**). Using
373 these comparisons, we found that, within some parameter regimes, the dynamics simplify to a
374 less complex balance of mutational processes (**Methods, Equations 10—11; Supplementary**
375 **Note**) and assessed the appropriate regime of validity (**Figs. 6a,b; Supplementary Fig. 14—17**).
376 To more directly test the accuracy of the PDE, we used our computational results to decompose
377 the per-generation fluxes in and out of each length class into relative contributions from each
378 mutational type. This allowed for identification of the dominant mutational processes
379 maintaining steady state (**Fig. 6c; Supplementary Note**); the accuracy of each approximation
380 was confirmed by analyzing the net magnitudes of fission and fusion within each length class
381 and regime (**Supplementary Figs. 18—20**).

382 We used this model to study the shape and stability of the empirical DRL and distinctions
383 between repeats in different length regimes under mutational forces alone. Expansions and
384 contractions remain non-negligible for any long repeat across the space of parameters that lead to
385 stable late-time DRLs, highlighting the importance of repeat length instability to the maintenance
386 of long repeats. For extreme parameters that stabilize (i.e., $\tau_\kappa \gg \tau_\epsilon$), the dynamics of all long

387 repeats are dominated by expansion and contraction, alone, leading to a DRL that truncates more
388 rapidly than under substitutions alone (i.e., a *depletion* of long repeats relative to a geometric
389 distribution). In contrast, for realistic parameters (i.e., within the 95% HDR for A-
390 mononucleotide repeats), an intermediate length regime emerges, characterized by the relevance
391 of repeat fission. An accurate description of the shape of the DRL requires fission to account for
392 the loss of repeats from the extreme tail (i.e., the longest populated length bins) and gain of
393 intermediate length repeats. The relative contributions of fission due to substitutions and non-
394 motif insertions are parameter-dependent; within the rough neighborhood of the maximum
395 posterior parameters (informative prior), substitution is the primary driver of fission up to lengths
396 of ~20 nt, while longer repeats are primarily interrupted by non-motif insertions (see
397 **Supplementary Note**). Fission-based losses in the extreme tail are insufficient to fully
398 counteract length increases due to expansion, independent of the mutational mechanism and
399 parameter values. Instead, contraction is primarily responsible for truncating the DRL at finite
400 repeat length but can be bolstered by both substitution- and non-motif insertion-based fission.
401 The dynamics of the long repeat regime decouples from that of short repeats such that rapidly
402 mutating long repeats effectively become independently evolving genomic elements,
403 categorically distinct from random sequences of the same length. The abundance of long repeats
404 in the genome may therefore be a consequence of their largely unencumbered evolution caused
405 by rapid changes in length.

406

407 *Inferring the onset of instability from the shape of the DRL*

408 We sought to better characterize the length at which repeats become independently evolving
409 genomic elements. Our analyses thus far suggests that this occurs roughly at the length where
410 expansion and contraction rates exceed substitution rates. This length was explicitly fixed in our
411 inference via reliance on empirical rate estimates at $L=1-8$; however, this precluded exploration
412 of the onset length of repeat instability. To study the encoding of this information within the
413 shape of the DRL, we defined fully parameterized rate curves (omitting all empirical rate
414 estimates) that include the length at which instability rates exceed substitution as an explicit
415 parameter. Expansion and contraction rates are each parameterized by an independent power law
416 at all lengths (i.e., with no reference to empirical estimates) in terms of an exponent, τ , and λ ,

417 the length at which the rate intersects the relevant substitution rate (μ for expansion, ν for
 418 contraction; full parameterizations defined in **Table 1, Methods**). This four-parameter model
 419 depicts an oversimplified rate dependence but serves as a toy model to probe the instability onset
 420 length λ .

421 Applying the same Bayesian inference pipeline, we estimated the posterior probability
 422 using both a uniform prior (i.e., excluding all rate data) and a prior informed by the combined set
 423 of empirical rate estimates (i.e., from both the pooled trio and popSTR data; see **Methods**). We
 424 marginalized the posterior to specify values of the onset lengths for expansion (λ_ϵ) and
 425 contraction (λ_κ) and found highly restrictive marginal distributions with 95% HDRs isolated to
 426 $\lambda_\epsilon = 9, \lambda_\kappa = 12 - 13$ (informative prior: $\lambda_\epsilon, \lambda_\kappa = 9, 12$; **Supplementary Fig. S21, Table 1**).
 427 This recapitulates the range of lengths observed in direct empirical rate estimates, despite
 428 excluding all such data from the inference (i.e., isolating the influence of the shape of the DRL).
 429 The posterior-weighted DRLs reproduce the informative features of the human DRL (e.g.,
 430 deviation from the substitution-driven geometric distribution at roughly 10 nt; **Supplementary**
 431 **Fig. 21b**), despite the oversimplified model of instability. This suggests that the transition in
 432 shape of the DRL corresponds to the onset length of repeat instability, allowing for rough
 433 estimation of this key feature from visual inspection of the distribution.

434

435 *Application to repeats with longer unit length*

436 Given that the DRL is informative about the onset length of repeat instability, we next compared
 437 this quantity across motifs of differing unit lengths (e.g. dinucleotides, trinucleotides). Empirical
 438 rate estimates for all motifs showed qualitatively similar properties to mono-A repeats (i.e.,
 439 predominantly single-unit expansions and contractions with rates that scale rapidly with tract
 440 length; **Fig. 3a, Supplementary Figs.4,5**), suggesting an analogous dynamical competition
 441 between substitutions and repeat instability that shapes the steady state DRL. We compared two
 442 distinct measures of the onset of repeat instability, representing long- and short-timescale
 443 information: first, rough lengths at which empirical DRLs first deviate from geometric decay
 444 (i.e., the expected DRL under substitutions alone) and second, the approximate lengths at which
 445 per-repeat expansion and/or contraction rates first exceed either substitution rate (i.e. rates
 446 comparable to μ or ν that perturb the geometric dependence on L). Both measures showed
 447 reasonable agreement confined to a range of onset lengths between roughly 6 and 12 nt, despite

448 differing unit lengths (**Supplementary Fig. 22a**). This suggests a universal description of repeat
449 dynamics that shapes the extended tail of the DRL, despite apparent differences in the geometric
450 portion at short lengths. Rapid geometric falloff is an immediate consequence of increasing the
451 unit length (**Supplementary Fig. 22b**): while a single substitution is sufficient to shorten tract
452 lengths regardless of motif, lengthening of repeat tracts can require multiple substitutions (up to
453 the unit length of the motif). Given initially comparable expansion and contraction rates across
454 motifs, longer motifs show a more immediate transition to the repeat instability-dominated
455 dynamics when measured in the number of repeated units but largely agree when measuring the
456 onset length of repeat instability in nucleotides (**Fig. 1a**).

457

458 Discussion

459 Motivated to understand the origin, prevalence and maintenance of simple tandem repeats in the
460 genome, we constructed a model of repeat evolution under mutagenesis alone that bridges short-
461 and long-timescale observations of repeat length instability. We demonstrate that mutations alone
462 are sufficient to explain the shape of the genome-wide distribution of tract lengths. The
463 abundance of long repeats in the genome reflects the rapid onset of repeat instability with an
464 initial expansion bias, rather than natural selection. This observation does not preclude selection
465 at specific loci, whether beneficial or disease-associated, provided these comprise a small portion
466 of repeats in the genome.

467 Length-dependent expansion-contraction bias is evident in our de novo estimates;
468 incorporating this property into the mutational model is sufficient to truncate the distribution at
469 finite lengths due to substantial contraction-bias. The long length tail of the distribution is
470 produced and maintained in a dynamic balance between expansion, contraction and fission. This
471 implicitly prevents the growth of repeats to disease-relevant lengths, suggesting natural selection
472 as a disease-prevention mechanism may not be essential. If selection, rather than contraction
473 bias, is responsible for terminating the distribution below disease length, it would have to be
474 enormously efficient to counteract instability-driven expansion rates and act globally across all
475 sufficiently long repeats. If pervasive selection plays a role in shaping the distribution, this must
476 be inferred as a deviation from the DRL under mutation alone built on a more complete model of
477 repeat mutagenesis.

478 Our analysis of the genome-wide properties of repeats is complementary to studies of
479 individual loci harboring disease, which generally occurs at or above the longest lengths present
480 in the reference genome (i.e., stochastically driven length classes in the present context). Such
481 elongated repeats can form motif-specific secondary structures that can disrupt replication and
482 repair, causing instability with qualitatively distinct properties^{8,9}. Furthermore, even amongst
483 repeats of the same motif, locus-specific properties can introduce variability in the length-
484 dependent rates of expansion and contraction and directional differences in bias (e.g., for long
485 CAG repeat loci⁷). One well-studied example is the CAG repeat locus responsible for
486 Huntington's disease. A recent analysis showed that a secondary phase of expansion-biased,
487 accelerated instability rates best explains somatic repeat expansion and its association with
488 disease progression⁵⁷. This locus-specific inference does not conflict with our observation that
489 contraction bias terminates the bulk of the genome-wide distribution; indeed, this may indicate
490 that, at lengths well-above those studied in the present manuscript, additional directional flips in
491 bias may occur. This, along with potential inter-locus variability, may contribute to the modest
492 number of repeats at lengths above the truncation point.

493 Our analysis offers a potential explanation for the prevalence of repeats at lengths that risk
494 progression to disease. First, the dynamics of short and long repeats decouple due to the rapid
495 onset, and subsequent dominance, of repeat length instability. Short repeat dynamics are dictated
496 by substitutions alone such that repeats within this regime are roughly indistinguishable from
497 random strings of nucleotides of the same length. Longer repeats are primarily subject to distinct
498 mutational forces, exhibiting rapid expansions and contractions and a higher rate of repeat
499 fission, which increases the total number of repeat tracts. Amongst long repeats, those of mid-
500 length primarily experience substitution-based fission, while mutations in the longest repeats are
501 effectively substitution-independent (i.e., fission is driven by non-motif insertion). This is
502 inconsistent with previous literature that suggested substitutions prevent disease by providing a
503 stopping force that counteracts indefinite expansion^{45,49,55,58,59,60}; instead, our analyses suggest
504 this is primarily a consequence of contraction bias at long tract lengths (similar to previous
505 proposals based on very early data^{25,27,33,47}). Given the negligible role of substitution, there is
506 little overlap in the mutational forces—and, subsequently, the underlying mechanisms—between
507 the shortest and longest repetitive sequences included in our analyses. In this sense, long repeats
508 emerge as independently evolving genomic elements (with parallels to the concept of selfish

509 genetic elements^{61,62,63}). Monotonically increasing instability rates generates length-dependent
510 dynamics under which expansions lead to further instability, while decreasing length is
511 effectively stabilizing; the former results in frequent forays into long length bins that may be the
512 precursors to disease. The onset of this process leads to a natural definition for the shortest
513 ‘unstable’ repeat (roughly 6—12 nt, far below disease length). This dynamical definition is
514 distinct from measuring the lowest length where expansion or contraction rates start exceeding
515 the background indel rate (as low as two units for many motifs; **Fig. 3b, Supplementary Fig. 5**),
516 which may better inform the molecular underpinnings of repeat instability. This difference in
517 scientific goals underlies the debate in the literature concerning the definition of unstable
518 repeats⁶⁴.

519 Provided selection plays little role in directly modifying repeat length, the conservation of the
520 distribution in steady state implies that the underlying mutational mechanisms (i.e., DNA
521 replication and repair) are highly conserved. Generically, such mechanisms play a broad role in
522 maintaining sequence fidelity of the entire genome, primarily preventing single nucleotide
523 mutations; due to the substantially larger target size, it is unlikely that machinery responsible for
524 both single site mutations and instability-driven length changes are optimized to properties of the
525 latter. The abundance of long repeats may thus be an inescapable consequence of the pleiotropic
526 function of the machinery maintaining genome-wide sequence fidelity.

527 It remains unclear which biological mechanisms control the key properties of repeat
528 length instability described in our study. The proposed mechanism(s) should be able to explain
529 length dependencies of instability rates (**Fig. 3b, Supplementary Fig. 5**) that show: a) rapid
530 onset from ~6—12 nt, surpassing the rate of substitutions, b) greater-than-linear increase in the
531 expansion/contraction rate per target above ~10 nt, c) generically asymmetric rates of expansion
532 and contraction with initial expansion-bias, followed by terminal contraction-bias, and d) single-
533 unit expansions/contractions, regardless of tract length (**Fig. 3a, Supplementary Fig. 4**).
534 Surprisingly, these observations appear to be largely independent of both motif sequence and unit
535 length (**Supplementary Fig. 5**), suggesting a common biological origin.

536 Two widely studied mechanisms, replication slippage and mismatch repair (MMR), likely
537 explains part of the story^{8,9,65—69}. Slippage, when newly synthesized DNA partially unwinds and
538 realigns out of register, should strongly depend on the unit length; however, we see only minor

539 variation associated with unit length (**Supplementary Fig. 5**). While slippage during DNA
540 replication produces loop-outs on both strands symmetrically⁶⁸, subsequent small loop-
541 processing by MMR preferentially results in contractions⁷⁰ due to bias towards the nascent
542 strand⁷¹. Slipped-strand structures may be a motif-independent source of loop-outs subject to the
543 same MMR-processing; in contrast, other secondary structures are motif-specific and therefore
544 cannot be the primary source of repeat instability but can potentially explain differences between
545 motifs⁷² (**Supplementary Fig. 1b**). Importantly, the observation of mostly single-unit
546 expansions/contractions argues against mechanisms involving larger structures (e.g., long
547 hairpins that cannot be processed by MutS β ^{73–78}), as these would be expected to generate multi-
548 unit indels.

549 Single-unit expansions have also been observed in a different context: Okazaki fragment
550 maturation by flap-endonuclease *FEN1*⁷⁹. Imprecise removal of the flap formed by the displaced
551 5'-flank of an Okazaki fragment may lead to expansion bias⁸⁰ and introduce an associated length
552 scale. A secondary mechanism takes over when flaps exceed 30 nt⁸¹; speculatively, long repeats
553 could give rise to long flaps. Likewise, another flap-endonuclease, *FAN1*, which recently
554 emerged as a genetic modifier of several repeat expansion disorders, was implicated in
555 processing of various slip-outs and demonstrated differential activity depending on flap length⁸².
556 Altogether, this illustrates how different mechanistic explanations may apply to repeats of
557 distinct lengths, generating emergent properties like length-dependent expansion-contraction
558 bias.

559 In addition to advancing a mechanistic understanding, substantial effort continues to be
560 dedicated to both assembling datasets and developing estimation techniques specific to repeat
561 instability, due to the inherent difficulties associated with repetitive DNA. Given the difficulty of
562 this task, the present work demonstrates how direct rate estimates can be informed by orthogonal
563 data. The comparative robustness of estimates of the distribution of repeat lengths provides
564 constraints on properties of instability that can serve as a new means for evaluating the quality of
565 differing rate estimates. The DRL may also serve as a summary statistic informative about the
566 evolutionary history of mutation rates and mechanisms, including in species where no population
567 data exists. Indeed, our rapidly improving understanding of repetitive elements, which have

568 historically evaded sequencing efforts, unlocks a range of new questions about the composition
569 and evolution of the genome.

570

571 **Methods**

572 Genome sources

573 Genome fasta files for T2T-CHM13_2.0 were downloaded from UCSC:
574 <http://hgdownload.soe.ucsc.edu/downloads.html#human>. Alternate human assemblies and
575 mammalian genomes were downloaded from the NCBI genome database:
576 <https://www.ncbi.nlm.nih.gov/datasets/genome/>

577

578 Motif labeling

579 Throughout the present study, repeat motifs are given a standardized label according to
580 alphabetical order within the list of all cyclical permutations of a given motif (e.g., CAG, AGC,
581 GCA) and their reverse complements (e.g., CTG, GCT, TGC). Outside of coding regions,
582 cyclical permutations of a motif become mostly indistinguishable, both bioinformatically and
583 biologically (after exceeding some minimal length relevant to processes such as protein binding
584 site recognition). Likewise, if not considering specific hypotheses such as transcription direction,
585 reverse complementary motifs should be treated as equivalent because Watson and Crick strands
586 are assigned to each chromosome arbitrarily. The present study does not investigate any of these
587 specific biological hypotheses, and so we combine results for all equivalent motifs under a single
588 label to increase statistical power. In this arrangement, well-studied motifs may receive a label
589 that differs from that commonly used in the literature (e.g. Huntington's disease (CAG)_n repeats
590 and myotonic dystrophy (CTG)_n repeats are both labeled 'AGC'.)

591

592 Generation of empirical repeat tract length distributions

593 Repeat tract length distributions were generated by counting consecutive complete motifs (i.e.,
594 perfect motifs, no interruptions and no partial motifs). Each distribution was assembled by

595 counting contiguous, uninterrupted repetitions of a specified motif. Instead of introducing an
596 arbitrary tolerance for interruptions when counting repeats of a given length, this strict definition
597 allows for straightforward bioinformatic assembly of the DRL. The regex pattern
598 ‘([ATGC]{1,6}?)1+’ detects arbitrarily long tracts of repeated nucleotides, finding any motif
599 with unit length of 1-6 nt. Using a ‘regex’ implementation in Python 3 (pypi.org/project/regex/,
600 version 2024.11.6), all motifs can be detected simultaneously by using the ‘finditer’ command
601 with the ‘overlapped=True’ option. Because this pattern detects repetitions of motifs, separate
602 regex patterns were used to detect single instances of each motif (i.e. $L=1$), taking the form
603 ‘([ATGC]{n})\1{0}(?!\\1)’, where n is each motif unit length 1-6. The results of all regex searches
604 were combined to generate a histogram of counts for each motif (pooled under the appropriate
605 label) at all tract lengths present in the genome (i.e., the distribution of repeat tract lengths,
606 DRL). Histograms representing counts of non-motifs (i.e., the lengths of contiguous regions
607 where a particular motif is absent; required for computational modeling) were generated on a
608 per-motif basis, using the regex pattern ‘(?:(!motif)+)[ATGC]+’ and combining the results
609 for all cyclical and reverse-complementary permutations of the given motif.

610 Bootstrap confidence intervals were generated around the T2T-CHM13 repeat length
611 distribution. The genome was divided into 1Mb contiguous non-overlapping segments,
612 discarding any sub-1Mb chromosome ends. DRLs were measured for each segment. A
613 distribution for the full-length genome was then reconstituted by randomly sampling from these
614 segments, allowing replacements, and summing the distributions from each segment. This
615 process was repeated 1000 times, after which 95% confidence intervals were generated by
616 separately taking the minimum and maximum in each length bin by removing the top and bottom
617 25 counts.

618 For the various mammalian genomes, the same counting procedure was applied. Assemblies
619 generated from short-read sequencing frequently contain many short contigs which typically
620 originate from poorly sequenced regions containing transposable elements; any contig of length
621 <10 kb was discarded. Taxonomic data was retrieved from
622 <https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/>. The median distribution of a given taxonomic group
623 was assembled by gathering the normalized DRLs (see below) for every member of the group

624 (i.e., for primates this includes humans, and for mammals this includes primates) and taking the
625 value of the median species for each length bin.

626

627 Distribution normalization

628 After initially computing the DRL for each motif from the T2T genome, we sought to compare
629 the shape of each histogram of raw counts to those assembled from distinct human reference
630 assemblies (**Supplementary Fig. 1c**) and from references for various species (**Fig. 1b**,
631 **Supplementary Fig. 2**). To compare distributions estimated from assemblies with differing total
632 target size, it was necessary to normalize each distribution (i.e., divide by the total number of
633 counts, summing over length bins) to standardize the overall scale. We refer to the normalized
634 DRL as the probability distribution of repeat tract lengths, which we interpret as an estimate of
635 the probability of randomly sampling a repeat of length L from the set of all contiguous motifs
636 (including $L=1$) in the assembly; when specific length classes are omitted, this becomes a related
637 conditional probability distribution (i.e., $P(L|L > L_{\min})$). Shorter assemblies (particularly due to
638 lower quality and read depth at repetitive loci) have a reduced overall number of sequenced
639 repeats and a threshold for statistical (and potentially stochastic dynamical) noise at lower length.
640 To ensure we are comparing estimates robust to statistical noise, we truncate each DRL above
641 the lowest length bin containing less than 30 counts (roughly 1.5 in logarithmic space). This
642 results in otherwise comparable normalized DRLs (assuming the same motif) with distinct
643 truncation points based on non-normalized counts. Qualitative differences between the shape of
644 the resulting normalized DRLs in remaining comparable length bins are indicative of differences
645 in the evolutionary parameters (e.g., mutation rate, selection, etc.) or systematic error profiles (or
646 both) between compared assemblies.

647 In addition to normalization for empirical comparisons, the empirical DRL and
648 parameterized theoretical DRLs generated by our computational model were normalized by
649 summing only over length classes above a specified minimum length to produce a comparable
650 normalized DRL, conditional on $L \geq L_{\min}$. This improved the summary statistic used to
651 characterize differences between these distributions. Further details and justification for the
652 specific choice of L_{\min} are provided in *Bayesian inference procedure*, below. To make figures
653 easier to interpret, normalized DRLs from the computational model were subsequently rescaled

654 (where noted) to match the non-normalized counts for T2T-CHM13 by multiplying each
655 normalized DRL by the sum of counts for bins $L \geq L_{\min}$ in the T2T-CHM13 DRL.

656

657 Bioinformatic estimation of substitution and indel rates

658 De novo mutation datasets were acquired as VCF files (or equivalent) from various published
659 sources^{83–90}, representing a total of 10,912 parent-child trios with available SNV data and 9,387
660 trios with available indel data. This dataset was compiled in McGinty and Sunyaev (2023)⁹¹ and
661 comprised of all freely available trio samples at the time of analysis; samples from distinct VCFs
662 were pooled to increase statistical power. We assumed that all individuals have the same
663 underlying mutation rates. Variants were mapped to GRCh38 either in the original study, or
664 subsequently, using ‘pyliftover’ (pypi.org/project/pyliftover/, version 1.3.2). The average
665 substitution rate was estimated to be 1.2×10^{-8} , calculated as: number of substitutions /
666 approximate number of sequenceable nucleotides in the diploid genome (see below) / number of
667 offspring genomes in the dataset. We classified substitutions according to six categories based on
668 trinucleotide context and the motif in question, as follows: for the example of mono-A motifs,
669 using B to represent non-A nucleotides, we determined rates (in parentheses) of ABB>AAB and
670 BBA>BAA (4.58×10^{-9}) representing repeat-lengthening events, AAB>ABB and BAA>BBA
671 (7.74×10^{-9}) representing repeat-shortening events, ABA>AAA (2.74×10^{-9}) representing fusion
672 events, AAA>ABA (4.35×10^{-9}) representing fissions, BBB>BAB (3.80×10^{-9}) representing the
673 rate of A₁ creation, and BAB>BBB (6.17×10^{-9}) representing the loss of A₁. Rates of substitutions
674 of B which do not create an A were not estimated.

675 We calculated indel rates as a function of repeat tract length. Using positional
676 information, upstream and downstream sequences for each event were pulled from the reference
677 genome, under the assumption that the sequence of the parental genome is identical to the
678 reference genome. For every focal motif, we used the reference sequence to determine tract
679 length. Indel rates per tract length per motif were estimated by dividing by the number of repeats
680 of that length, obtained by generating DRLs in a GRCh38 genome masked for low quality
681 regions (see below). Each indel was classified as an expansion, contraction or non-motif
682 insertion, additionally measuring how many motif units were added/removed in the event. We
683 limited mutations in our computational model to +1/-1 unit changes in length at appropriate

684 rates. We also measured the rate of indels for all B positions (with respect to each motif; mono-A
685 rates in parentheses), separately estimating the rates of BB>BBB (1.38×10^{-10}), BBB>BB
686 (4.37×10^{-12}) and ABA>AA (2.76×10^{-10}) events. Because B strings were not modeled as having
687 length-dependent instability, we measured the average rate, i.e., the rate per unit.

688 Limitations of the VCF file format, namely the lack of any information at unmutated
689 positions, forces the treatment of the pooled-trio VCFs as a complete record of variants in all
690 individuals. At a coarse level, this problem was minimized by assuming that 100kb regions
691 lacking any substitutions across the combined dataset suffer from regional mappability issues.
692 These regions were masked in GRCh38 when estimating the denominator for rate calculations.
693 At the fine level, this issue persists: mutations may have been filtered (prior to populating the
694 VCF file) due to localized drops in sequencing quality, resulting in false negative calls and
695 undercounting in the pooled estimates. This results in an underestimate of mutation rates,
696 because counts from GRCh38 used in the denominator remain static. This may particularly affect
697 estimation of instability rates as repeat tract length increases, because long repeats are known to
698 interfere with several facets of the sequencing and bioinformatic processes^{35,36,37,38}. We believe
699 this systematic error mode, leading to progressively more severe underestimation of instability
700 rates with increasing tract length, is the underlying cause of non-monotonicity observed in these
701 rate estimates (**Supplementary Fig. 5**). Mononucleotide repeats may be especially susceptible to
702 systematic rate underestimation, as they are among the most difficult motifs to sequence³⁸.

703 The popSTR repeat instability dataset, representing 6,084 parent-child trios, was acquired
704 from the supplement of Kristmundsdottir et al. (2023)⁴⁶. This dataset was incorporated into our
705 inference due to the unique methodology, which provided high quality calls of mutations
706 extending beyond short tract length repeats that allowed us to produce length-stratified rate
707 estimates. Files ‘bpinvolved_extended’ and ‘mutRateDataAll.gz’ were downloaded from
708 https://github.com/DecodeGenetics/mDNM_analysisAndData. Due to our focus on uninterrupted
709 repeats, we measured the longest contiguous repeat tract within the provided coordinates for each
710 event. We limited the dataset to loci where the popSTR-reported reference tract length agreed
711 with our own measurement in GRCh38. The ‘bpinvolved_extended’ file contains a mix of
712 phased and unphased data; where the parental length for a given mutation was not assigned by
713 phasing, we assumed that it originated from the parental copy which minimizes the difference in

714 tract length between the proband repeat and any of the parental repeats. Skipping this phasing
715 step under the assumption that all events originated from the reference length allele (but retaining
716 the size and direction of the event), as we do for the pooled-trio dataset, results in relatively
717 minor differences in counts per length bin. The ‘mutRateDataAll.gz’ file contains information on
718 the number of trios where all three samples passed sequencing quality filters at a given locus,
719 and the length of the repeat tract at each locus in GRCh38, but lacks information on the parental
720 genotypes for each of these loci (i.e., the file does not report pass/fail counts stratified by
721 parental tract length). For the denominator of the popSTR mutation rates, we thus generated a
722 distribution of passing counts (using the reference length for each locus), multiplied by two
723 parental alleles. This assumption leads to some amount of misestimation of rates: loci containing
724 long repeats show higher tract length variance in the population (due to higher instability rates),
725 and thus individuals are more likely to differ from the reference genome. It is unclear whether a
726 related effect (owing to the absence of loci with reference tract lengths below 10 nt in the
727 popSTR dataset), or some other unknown error mode, is responsible for apparent overestimation
728 of popSTR-based rates at shorter tract lengths where direct comparisons to reliable pooled-trio
729 estimates are possible (**Supplementary Fig. 5**).

730 We note that the popSTR dataset differs from the pooled trio data in several aspects: the
731 popSTR caller provides no estimates for reference tract lengths below 10 nt; mutations are
732 classified by length change, failing to distinguish between expansions and non-motif insertions;
733 and, due to data access limitations, we were unable to assess the nature and magnitude of
734 potential systematic errors detailed above. These distinctions precluded direct merging of
735 instability rate estimates with the pooled trio data; popSTR-based estimates were instead
736 incorporated into our inference by informing the prior (see below). Insertion rates, which are the
737 sum of expansion and non-motif insertion rates, were used as a surrogate for expansion rates
738 under the assumption that non-motif insertion remains far more infrequent than expansion
739 (consistent with estimates from the pooled trio data at roughly 1% of total insertions).

740 For substitution and indel rate estimates based on either the pooled-trio or popSTR datasets,
741 we calculated 95% confidence intervals based on 200 Poisson samples of the mutation counts,
742 removing the top 5 and bottom 5 values per length bin (see **Fig. 3b**, **Supplementary Fig. 5**). We
743 note that error bars provided on each estimate represent only Poisson distributed statistical error

744 bars associated with point estimated counts in the numerator of each rate and are therefore
745 subject to the above and any additional systematic errors underlying variant calling.

746

747 Computational modeling of repeat length dynamics

748 We used a custom-written script in Python 3 that models repeat dynamics by directly
749 manipulating the distribution of repeat lengths. We simultaneously tracked and manipulated the
750 length distribution of B strings. As detailed above, we assumed a binary genome consisting only
751 of A and B sites, where A is a repeat unit and B represents any non-A unit; as a result, B strings
752 do not a priori represent repetitive sequences. Mutations are applied in aggregate such that, in
753 each generation, repeats transition between integer length bins according to rules associated with
754 each mutational process, while the B distribution is updated accordingly (e.g., a substitution that
755 lengthens a repeat simultaneously shortens a B string). Mutation rates were restricted to be
756 sufficiently low to model only a single mutation event per repeat per generation. The non-
757 normalized distribution was evolved and subsequently normalized to create a probability
758 distribution for comparison to empirical data. This approach is far more computationally efficient
759 than simulating an entire genomic sequence, subsequently applying mutations and generating a
760 distribution; computational time in our script scales with the number of length bins rather than
761 with the length of the genome. Tracking only the distribution discards information about the
762 location of particular mutations, instead generating an expected number of mutations for each
763 category per length bin per generation. Except where specified, we used a deterministic
764 approximation to assess the behavior of the expectation value of each bin as the distribution
765 evolves toward steady state via repeated application of the mutation kernel. To understand the
766 impact of stochastic fluctuations on the steady state distribution, we additionally implemented a
767 model that represents fluctuations by Poisson sampling the expected change to each length bin
768 per generation. We model stochastic fluctuations around the applicable rates by sampling
769 mutational counts, but without constraining individual transitions (i.e., a net number of mutations
770 may leave a given class, but the number introduced elsewhere, as a result, is appropriately
771 distributed only on average due to an independent sampling procedure). All subsequent analyses
772 were performed using the deterministic results, as modeling independent fluctuations in each bin
773 showed no qualitative differences (**Supplementary Fig. 13e**).

774 Mutations affect the distribution via the following well-defined rules for substitutions and
 775 indels (see **Fig. 2** for illustration). These rules assume that each mutation adds, subtracts or
 776 substitutes a single, complete repeat unit (i.e., the most prevalent class of length changes, seen in
 777 **Fig. 3a, Supplementary Fig. 4**). Using the example of a repeat of $L=6$, a lengthening
 778 substitution subtracts one count from the $L=6$ bin and adds one to the $L=7$ bin. A shortening
 779 substitution subtracts one from the $L=6$ bin and adds one to the $L=5$ bin. A substitution causing
 780 repeat fission subtracts one from the $L=6$ bin and adds two new repeats, either one $L=1$ and one
 781 $L=4$, or one $L=2$ and one $L=3$ (when evolving the distribution in aggregate, both occur
 782 simultaneously with appropriate relative rates). The reverse process of fission is fusion, in which
 783 an $L=6$ repeat can be generated by fusing an $L=1$ with an $L=4$, or by fusing one $L=2$ and one $L=3$
 784 repeats, while the mutated B unit is replaced with an A unit and added to the repeat length.
 785 Lengthening and shortening substitutions act locally (i.e., counts leave the L bin and move to the
 786 adjacent $L + 1$ and $L - 1$ bins, respectively). Substitution of an $L=1$ in the A distribution also
 787 corresponds to fusion of B strings; the reverse, i.e., substitution of a length one B string,
 788 generates fusion in the A distribution. Fission and fusion substitutions inherently act non-locally
 789 in length space: fission results in the loss of one count in the L bin and gain of two counts that
 790 are evenly distributed across all bins of length $\leq L - 2$; fusion evenly subtracts two counts from
 791 bins $\leq L - 2$ to add a count to L . The net effect of substitutions conserves the total length of the
 792 genome, i.e., the sum of the length of all A repeats plus the sum of the length of B strings
 793 remains constant under substitutions alone.

794 The rates of lengthening, shortening, fission and fusion substitutions per generation are
 795 separately estimated using the three-unit context: BBA>BAA (or ABB>AAB) for lengthening
 796 substitutions, AAB>ABB (or BAA>BBA) for shortening substitutions, AAA>ABA for fissions,
 797 and ABA>AAA for fusions. All substitution rates were assumed to be independent of repeat
 798 length, based on our previous observations showing little to no rate increase with increasing
 799 repeat length⁹¹. The target size for lengthening substitutions is two per repeat (i.e., the two sites
 800 adjacent to each repeat boundary). Likewise, the target size for shortening substitutions is also
 801 two per repeat, representing the two boundary units of the repeat (assuming $L > 1$). The target
 802 size for fission substitutions is $L - 2$ per repeat, representing all non-boundary units within the
 803 repeat. The target size for all fusion events is proportional to the $L=1$ count of the B distribution.
 804 Equations governing these processes are described in detail in the **Supplementary Note**.

805 Indel mutations operate under an analogous logic, but with a few important distinctions.
 806 Indels, by definition, do not conserve the length of the genome. Expansions and contractions act
 807 strictly locally, but the location of the event is indistinguishable within the repeat, affecting any
 808 of the units rather than just the boundaries; this results in a per-repeat target size L for these
 809 mutations, rather than 2. Non-motif insertions (i.e., AA>ABA) cause fission, resulting in the loss
 810 of one count in the L bin and gain of two counts that are evenly distributed across all bins of
 811 length $\leq L - 1$; deletion of a B string of $L=1$ (i.e., ABA>AA) causes fusion, which evenly
 812 subtracts two counts from bins of length $\leq L - 1$ and adds one count to bin L . Indel rates for
 813 expansions and contractions are incorporated in a length-dependent manner, described above, in
 814 contrast to substitution rates. We did not model length dependence for B indels, as most B strings
 815 represent a combination of nucleotides and not necessarily STRs with any biological relevance.
 816 This assumption should not impact the evolution of the A distribution after normalization, which
 817 is only coupled to the $L=1$ class of the B distribution; this length class is dominated by
 818 substitution rate dynamics and not subject to repeat instability.

819

820 *Time-rescaling using a constant speed-up factor*

821 Due to the large number of iterations required to reach a steady-state DRL, propagating the
 822 mutational process directly was computationally prohibitive. Instead, we approximated the DRL
 823 by first rescaling time by multiplying all mutation rates by the same constant 10^r such that each
 824 iteration represents 10^r generations of evolution for a total of $T = 10^{Ir}$ generations (assuming a
 825 constant r at all time points run for I iterations); r was limited to integer values for convenience.
 826 This defines a set of time-rescaled substitution, expansion, contraction, and non-motif insertion
 827 rates as a function of length.

828 Due to the rapid growth of instability rates with increasing length, these rates quickly
 829 saturate, reaching probabilities of one at some length L . To avoid multiple mutations per repeat
 830 per iteration, we defined a saturation length L_{\max} as the length at which the sum of all mutation
 831 rates first exceeds 0.1 (e.g., $L_{\max} \rightarrow \infty$ if the sum of rates remains below 0.1 at all lengths). L_{\max}
 832 thus demarcates the linear mutation regime, below which multiple mutation events remain rare.
 833 In addition to dependence on the instability rate parameters, L_{\max} is dependent on r : increasing
 834 the speed-up factor increases mutation rates by 10^r , which decreases L_{\max} . To ensure reasonable

835 computational time, L_{\max} was limited to a maximum of 200 (i.e., $\min\{L_{\max}, 200\}$), computed
 836 separately for A -repeats and B -string lengths), which extends well beyond the empirical DRLs.
 837 To prevent loss of mass associated with the finite length grid, we imposed a reflective boundary
 838 condition at $L_{\text{bound}} = \min\{L_{\max}, 200\}$, (i.e., all transitions from lengths $L < L_{\text{bound}}$ to
 839 $L \geq L_{\text{bound}}$ were assigned to L_{bound}). This results in artefactual behavior near the boundary but
 840 provides a reasonable approximation when the expected number of counts drops below one at
 841 lengths far below L_{bound} . Substantial counts at the boundary are indicative of unrealistic
 842 distributions (often associated with diverging total genome size), provided L_{bound} is sufficiently
 843 far from the maximum well-populated lengths in the comparable empirical distribution.

844 For a given parameterization, producing a grid of DRLs requires choosing a constant
 845 speed up r and the boundary length L_{bound} appropriate for each parameter combination (the
 846 required computational time is largely determined by the number of parameter combinations
 847 with the lowest value of r). This procedure can be used to produce a coarse grid of parameters
 848 (e.g., for comparison to alternative approximations) but proved computationally prohibitive for
 849 the dense grid needed for inference.

850

851 *Step-wise speed-up procedure*

852 To produce finer grids of DRLs (for several parameterizations), we implemented a
 853 procedure that reduces overall computational time, while producing approximately the same
 854 DRL as that under a constant speedup (described above). This procedure models the evolution of
 855 the DRL by performing several, discrete phases of evolution, each with successively smaller
 856 time-rescaling factors 10^r . Each stage is allowed 10^6 iterations of evolution under the specified r
 857 (and the associated reflective boundary at $L_{\text{bound}} = \min\{L_{\max}, 200\}$ for each parameter
 858 combination); $r = 3$ for the first stage, and is reduced to $r = 2, 1$, and 0 for subsequent stages.
 859 L_{bound} is altered at each stage to maintain the linear mutation regime (i.e., ensuring $L_{\max} \geq$
 860 L_{bound}). In total, parameter combinations can experience up to four stages (equivalent to 4×10^6
 861 iterations, or 1.111×10^9 generations).

862 For computational efficiency, we first separated parameter combinations that rapidly
 863 equilibrate in a single stage of 10^6 iterations under a sufficiently large rescaling factor $r \geq 3$
 864 (easily identified by $L_{\max} \geq L_{\text{bound}} = 200$; equivalent to propagation using a single, constant

865 speed-up factor). These were each run at the largest allowed integer r for the equivalent of 10^9
 866 generations and removed from the grid. All other parameter combinations were subjected to
 867 several stages with progressively decreased r ; after each stage, parameter combinations deemed
 868 equilibrated were removed from the grid (again identified by $L_{\max} \geq 200$, determined by the
 869 preceding r -rescaled instability rates). For parameter combinations with $L_{\max} < 200$ in the
 870 absence of any speed-up factor ($r = 0$), counts in all length bins between L_{\max} and 200 were set
 871 to zero prior to analysis of the DRL.

872 To ensure that the multi-step procedure provides a reasonable approximation to the DRL
 873 produced under a constant speed-up, we compared inference results over a coarse grid of
 874 parameter combinations and found negligible differences (**Supplementary Fig. 13**). Intuitively,
 875 this procedure takes advantage of the faster mutation rates at longer lengths, which equilibrate on
 876 much more quickly than shorter length bins.

877

878 *Initial conditions at $t = 0$*

879 The computational model was initialized with an initial distribution that is approximately
 880 geometric (created by propagating substitutions alone, setting all instability rates to zero) for the
 881 equivalent of 10^{10} generations (using the largest allowable rescaling, $r = 5$). Using this
 882 approximation to the substitution-only steady-state as a pre-simulation substantially reduces
 883 equilibration times because the lowest length bins (i.e., those dominated by substitutions) require
 884 the most time to equilibrate due to low mutation rates.

885 Although the eventual steady-state DRL should not depend on the initial state, the choice
 886 of initial distributions can dramatically affect equilibration times. We confirmed that the final
 887 timepoint DRLs are effectively independent of choice of initial condition by comparing the
 888 results of two distinct initial conditions with similar equilibration times (geometric vs. geometric
 889 plus uniform; **Supplementary Fig. 12**), finding only minor differences in the deterministic late-
 890 time distribution.

891

892 *Computational model inputs and outputs*

893 The script relies on the following as inputs: an initial distribution for A and B (i.e., motif
 894 and non-motif) repeat lengths, per-target substitution rates in three-unit context, and per-target
 895 mutation rate curves for expansions, contractions and non-motif insertions. Substitution rates and

length-dependent indel rate curves are imported from external files (see above for estimated substitution rates, below for generation of parameterized rate curves); these files, along with the initial repeat length distribution table, can be replaced with appropriate tables for other purposes, if desired. This table must specify rates for each mutational process at all lengths intended to be computationally modeled (i.e., from 1 to L_{bound}). For normalized length distributions that reach steady state, the initial distribution can be chosen arbitrarily, in principle, but any specific choice affects equilibration time; due to equilibration time differences, minor differences between the deterministic late time distributions arise from distinct initial conditions (see **Supplementary Fig. 13** for comparison between two initial distributions).

Stochastics can be introduced using a command line option to model fluctuations in the mutational process; the number of mutations in and out of each length bin are separately Poisson sampled (using `numpy.random.poisson`, version 2.2.1) around the expected number of mutational counts in each iteration.

After each run, we output a file containing repeat length counts reported at various time points to show the temporal evolution of the distribution. We subsequently normalized the resulting distributions by dividing each length bin by the total number of repeats in the distribution (see Methods on normalization).

The relative contribution of each mutational force was assessed by producing a single-generation plot of the transitions in and out of each length bin at the final time point (i.e., once steady state was reached, if applicable). To produce these plots (see **Fig. 6**), we applied the mutation kernel for a single generation and separately computed the number of fission, fusion and local changes for substitutions and indels. For each length bin, the magnitude of total flux in and out was normalized to one. Length bins that have equilibrated should contain equal fluxes in and out; steady state occurs only when all bins show equilibrated fluxes.

920

921 *Bayesian inference procedure*

Given our observations indicating a stable distribution of repeat lengths over phylogenetic time scales, we sought to identify mutation rates capable of explaining this observation. To study the extent to which mutational processes alone can recapitulate the repeat tract length distribution, we constructed a Bayesian inference framework to compare models (i.e., parameterizations) of the length-dependent rates of repeat instability. Each parameterization

927 describes the length-dependent rates of expansion and contraction as following a simple
 928 functional form; as discussed above, substitution rates are assumed to be length independent in
 929 all cases. Within the Bayesian framework, a prior probability distribution on the parameter space
 930 is specified (several priors were used for interpretation) and used to weight the likelihood to
 931 calculate a posterior probability that a given parameter combination accurately describes the
 932 length dependence of the repeat instability rates. In the present setting, the likelihood is
 933 constructed by comparing the empirical repeat tract length distribution to the late-time
 934 distribution generated by computationally modeling a given parameter combination. Due to the
 935 analytic intractability of this likelihood, we used Approximate Bayesian Computation (ABC)⁵⁶ to
 936 approximate the posterior probability distribution, which avoids specifying the likelihood
 937 explicitly. Additionally, length bins are presumably correlated due to the complex mutational
 938 transitions underlying the distribution, complicating naïve construction of the likelihood. In
 939 contrast, ABC-based inference circumvents this issue by approximating the posterior in terms of
 940 summary statistics that appropriately characterize the DRL (see discussion of summary statistics
 941 below). After specifying summary statistics, the late-time distribution for each parameter
 942 combination was summarized for comparison to the empirical distribution (e.g., mononucleotide
 943 A repeat tract lengths in T2T-CHM13). For each parameterization, we specified a discrete grid of
 944 parameters for comparison to the empirical distribution, the result of which was weighted by the
 945 prior probabilities for those parameters (equivalent to randomly sampling the prior as prescribed
 946 in ABC⁵⁶; see below for specific priors used) to compute the posterior.

947 We chose to use the Kullback-Leibler (KL) divergence, a well-established statistic for
 948 distribution comparison, to characterize the difference between the empirical (P_L^{emp}) and
 949 parameterized repeat tract length distributions. The KL divergence quantifies the extent to which
 950 each parameterized distribution diverges from the empirical distribution and was calculated for
 951 all parameter combinations (denoted θ below) on the discrete grid using the following definition.
 952

$$D_{\text{KL}} = \sum_{L=L_{\min}}^{L=L_{\max}} P_L^{\text{emp}} \log \left(\frac{P_L^{\text{emp}}}{P_L(\theta)} \right) \quad (1)$$

953 We note that comparing the empirical distribution to itself results in a divergence of zero such
 954 that the KL values are equivalent to the difference between the modeled and empirical
 955 distributions (i.e., $\Delta D_{\text{KL}} = D_{\text{KL}}(P_L(\theta), P_L^{\text{emp}}) - D_{\text{KL}}(P_L^{\text{emp}}, P_L^{\text{emp}}) = D_{\text{KL}}(P_L(\theta), P_L^{\text{emp}})$). To

958 define a cutoff for ABC rejection, we estimated the divergence between the human empirical
 959 distribution and the ensemble of primate genomes using the same statistic. Under the assumption
 960 that primates evolved towards the same steady state (i.e., the mutational parameters remain
 961 constant across the phylogeny), we proceeded under the assumption that differences between the
 962 repeat tract length distributions in distinct species are due to a combination of stochastics and
 963 bioinformatic errors due to the lower coverage and short read technologies used to assemble
 964 primate reference genomes. Due to the difference in assembly lengths, we added a pseudocount
 965 of one to all length classes in all species to avoid divergence of the statistic and confirmed that
 966 our results were qualitatively independent of the choice of pseudocount between 0.01 and 100.
 967 L_{\max} was set the longest modeled length bin, $L = 200$. We set the lower bound to $L_{\min} = 4$ to
 968 ensure that the ordering of D_{KL} statistics computed for all primates remains roughly consistent:
 969 setting $L_{\min} \geq 4$ resulted in the smallest values for the human HG38 reference (which was not
 970 included in subsequent analyses) and the largest values for the most divergent primates (i.e.,
 971 those on the loris branch). We used the range of 36 computed primate KL values to very roughly
 972 define a rejection threshold by throwing out the largest 2 values; we considered the remaining 34
 973 values (i.e., roughly, the closest 95% of ranked primates) and used this to approximate the
 974 variance of D_{KL} (i.e., $\sigma_{D_{KL}}^2 \approx (D_{KL}^{95th \text{ percentile}}/2)^2$) associated with stochasticity and sequencing
 975 errors.

976 We approximated the posterior probability (up to a normalization constant) following the
 977 prescription in Wilkinson⁵⁶ wherein ABC is applied with a soft rejection threshold by rejecting
 978 values of $D_{KL}(\theta)$ with the following probability based on the primate-estimated variance.
 979

$$980 \quad Pr(\theta|\text{data})Pr(\text{data}) = Pr(\text{data}|\theta)Pr(\theta) \sim Pr(\theta) e^{-\frac{D_{KL}^2(\theta)}{2\sigma_{D_{KL}}^2}} \quad (2)$$

981
 982 Here, $Pr(\theta|\text{data})$ is the posterior probability distribution over the grid of parameter
 983 combinations θ , $Pr(\theta)$ is the prior, and the Gaussian falloff is the rejection probability (up to
 984 normalization) for a given parameter combination. This soft-rejection procedure provides a
 985 slightly better approximation for the posterior distribution than rejection with probability one
 986 (e.g., reject all parameter combinations with $D_{KL}(\theta) > D_{KL}^{95th \text{ percentile}}$, roughly corresponding to
 987 the primate-estimated 95% confidence interval). Additionally, Wilkinson argues that the

988 Gaussian rejection probability quantifies model misspecification inherent in the procedure, as all
 989 parameterizations (i.e., models) employed herein are imperfect approximations of the true
 990 instability rate parameters.

991

992 *Model comparison using Bayes factors*

993 To assess the relative explanatory power of each parameterization M_θ modeling the
 994 length dependence of repeat instability, we computed a Bayes factor for each model $BF(M_\theta)$
 995 using the definition below.

996

$$997 \quad BF(M_\theta) = \int d\theta Pr(\theta | M_\theta) Pr(\text{data} | \theta, M_\theta) \sim \int d\theta Pr(\theta) e^{-\frac{D_{KL}^2(\theta)}{2\sigma^2_{KL}}} \quad (3)$$

998

999 Here, the right-hand side is our previously computed approximation to the posterior, integrated
 1000 over the parameter space for a given model. As we were only interested in the relative Bayes
 1001 factor between models, proportionality constants can be ignored, including the overall
 1002 normalization and an assumed uniform prior over model space. The Bayes factor for a model
 1003 naturally controls for the number of degrees of freedom in each parameterization because
 1004 integration over the weighted posterior is performed in parameter spaces with differing
 1005 dimensionalities. Once computed, models were compared by interpreting the Bayes factor ratio
 1006 (BFR) as indicative of the relative statistical support between two parameterizations of interest.

1007

$$1008 \quad BFR(M_1, M_2) = BF(M_1)/BF(M_2) \quad (4)$$

1009

1010 We then used the Jeffrey's scale to interpret the strength of statistical support for each model.

1011

1012 *Parameterizations of repeat instability rates*

1013 We tested several parameterizations to assess consistency with the empirical distribution
 1014 of repeat tract lengths. We focused on mononucleotide A repeats, as both the distributions and
 1015 rate estimates were supported by the most empirical data. For computational convenience, we
 1016 defined a sequence of nested parameterizations (see **Supplementary Fig. 23**) that could be
 1017 computed simultaneously across the grid of the parameter combinations under the model with
 1018 the largest number of degrees of freedom (DoF). To define the most general set of length-

1019 dependent instability rate models, we parameterized expansion and contraction rates, $\epsilon(L)$ and
 1020 $\kappa(L)$, respectively, as independent power law functions at all lengths $L > 8$.

1021

$$1022 \quad \epsilon(L > 8) = c_\epsilon \left(\frac{L}{9}\right)^{\tau_\epsilon}, \quad \kappa(L > 8) = c_\kappa \left(\frac{L}{9}\right)^{\tau_\kappa} \quad (5)$$

1023

1024 Rates for $L = 1—8$ were taken directly from empirical rate estimates; the expansion and
 1025 contraction rates at longer lengths were parameterized in terms of c_ϵ and c_κ , which denote their
 1026 respective values at $L = 9$ (i.e., the first parametrized length bin). Guided by our empirical
 1027 estimates, we assumed that the rate of non-motif insertion $\iota(L)$ is directly related to the rate of
 1028 expansion with the same length dependence at 1% of the rate (i.e., $\iota(L) = \epsilon(L)/100$ at all
 1029 lengths. This results in instability rates characterized by four independent parameters
 1030 ($c_\epsilon, \tau_\epsilon, c_\kappa, \tau_\kappa$). We then constructed a series of nested lower-dimensional models for comparison.
 1031 A natural way to reduce the dimensionality of the parameter space is to introduce symmetries
 1032 corresponding to $c_\epsilon = c_\kappa$ and/or $\tau_\epsilon = \tau_\kappa$. The simplest model assumes fully symmetric
 1033 expansion and contraction rates (i.e. both $c_\epsilon = c_\kappa \equiv c$ and $\tau_\epsilon = \tau_\kappa \equiv \tau$) with a two-dimensional
 1034 parameter space (c, τ). We treat this as a null model corresponding to a frequently
 1035 discussed^{8,20,23,28} biological interpretation of repeat slippage. The parameter space can be reduced
 1036 to three DoF by restricting to either $c_\epsilon = c_\kappa$ or $\tau_\epsilon = \tau_\kappa$, which need not have straightforward
 1037 biological interpretations. We constructed an additional 3 DoF model parameterized by
 1038 ($m, \tau_\epsilon, \tau_\kappa$) by treating the expansion and contraction rates at $L = 9$ as increased by a common
 1039 multiplier m relative to their values at $L = 8$ (i.e., $\epsilon(9) = m \epsilon(8), \kappa(9) = m \kappa(8)$). For
 1040 computational expediency, we embedded this model within the four-dimensional grid of
 1041 parameters by appropriately choosing intervals for c_ϵ and c_κ when defining the grid
 1042 discretization.

1043 We used two distinct, non-nested parameterizations for subsequent analyses. To test the
 1044 reliance of our inference on the functional form of the length dependence (i.e., power-law
 1045 parameterization), we defined an additional parameterization by replacing the length dependence
 1046 for $L > 8$ with logarithmic growth in the following form:

1047

$$1048 \quad \epsilon(L > 8) = c_\epsilon \left(\frac{\log(L-7)}{\log 2}\right)^{\tau_\epsilon}, \quad \kappa(L > 8) = c_\kappa \left(\frac{\log(L-7)}{\log 2}\right)^{\tau_\kappa} \quad (6)$$

1049

1050 Here, the dependence on $\log 2$ ensures that c_ϵ and c_κ parameterize the values at $L = 9$ of the
 1051 expansion and contraction rates, respectively. Under this parameterization, empirical estimates
 1052 were again used for all lengths $L \leq 8$. This functional form retains monotonicity while growing
 1053 more slowly at longer lengths to model a saturation-like effect.

1054 We analyzed a second version of the power-law parameterization that extends the
 1055 functional form to all lengths such that the rates are fully independent of empirical estimates. We
 1056 re-parameterized the functional dependence in terms of the parameters $(\lambda_\epsilon, \tau_\epsilon, \lambda_\kappa, \tau_\kappa)$, where λ_ϵ
 1057 and λ_κ correspond to the length at which each instability rate exceeds the relevant substitution
 1058 rate (i.e., $\epsilon(\lambda_\epsilon) = \mu$ and $\kappa(\lambda_\kappa) = \nu$).
 1059

$$\epsilon(L) = \mu \left(\frac{L}{\lambda_\epsilon} \right)^{\tau_\epsilon}, \kappa(L) = \nu \left(\frac{L}{\lambda_\kappa} \right)^{\tau_\kappa} \quad (7)$$

1060 Here, μ and ν are point estimates of the average lengthening and shortening substitution rates,
 1061 respectively, for a given motif; note that this parameterization is defined at all tract lengths
 1062 (including $L \leq 8$). This allowed us to directly infer the length scale of the instability-substitution
 1063 rate crossover and assess the extent to which our inferences from the above parameterizations
 1064 rely on direct use and accuracy of empirical rate estimates below $L = 9$. After confirming that
 1065 expansion-biased parameters do not approach steady state DRLs, the parameter space was
 1066 further limited to asymptotically contraction-biased parameter combinations (i.e., with $\tau_\kappa > \tau_\epsilon$)
 1067 to limit computational time.
 1068

1069 The functional form of each of the aforementioned parameterizations is specified in Table
 1070 1.
 1071

1072 *Construction of prior distributions*

1073 For each parameterization, we constructed an uninformative prior by treating each
 1074 parameter combination as equally probable with probability equal to $1/n$, where n is the number
 1075 of computationally modeled points on a discrete grid. We next generated informative priors
 1076 using approximations derived from our empirical estimates of the expansion and contraction
 1077 rates. For power-law parameterizations that include empirical estimates at low lengths (Models
 1078 1—5 in **Table 1**), we performed a linear fit to the popSTR rate estimates (at lengths $L=11$ —29)
 1079 in log-log space to estimate parameters of best-fitting power laws. Curve fitting was performed
 1080

1081 in Python 3 (`scipy.optimize.curve_fit()`, version 1.15.1) with the sigma option to specify an array
 1082 of approximately symmetric log error bars (i.e., approximating a rescaled Poisson as log-
 1083 normal). We note that Poisson regression does not appropriately model statistical noise due to
 1084 target size rescaling when estimating rates from mutational counts. We fit using rate estimates at
 1085 all available lengths. However, we artificially inflated the variance at lengths above and below
 1086 $L=13-21$ to model potential systematic errors that generate observed non-monotonicity, likely
 1087 due to miscalling at the shortest and longest lengths accessible to popSTR. The optimization
 1088 package produced a covariance matrix for the best-fit line expressed in terms of the slope and
 1089 intercept in log-log space. We used this covariance matrix to approximate lines representing the
 1090 95% confidence bounds around the best-fit line. Using these lines, we estimated the value of the
 1091 best-fit line and standard deviation at $L = 9$. Assuming no correlation between expansion rate
 1092 and contraction rate parameters, we used these values to approximate a block diagonal
 1093 covariance matrix for the parameters $(c_\epsilon, \tau_\epsilon, c_\kappa, \tau_\kappa)$ in the four-dimensional model. We then
 1094 inflated the variance by a constant (100-fold for ‘restrictive’ informative prior; 1000-fold for
 1095 ‘permissive’ prior) and used the rescaled covariance matrix to define a multivariate normal
 1096 distribution centered at the point-estimates for the best-fit parameters. The informative prior for
 1097 the four-parameter model was constructed by normalizing this multivariate normal over the
 1098 discrete grid. Analogous priors for nested models were defined by restricting to the appropriate
 1099 subset of parameter space, maintaining the relative weights specified by the normal distribution,
 1100 and normalizing by the number of discrete grid points in this subset.

1101 The approximate nature of our ABC-based inference procedure prohibited construction of
 1102 strictly uninformative priors (i.e., Jeffreys priors) for fair comparison between models with
 1103 differing parametric functional forms. We instead treated the uniform prior as naively
 1104 uninformative; however, despite the similarity of their parameters, we caution that uniform priors
 1105 are inequivalent for distinct functional forms. To facilitate very rough model comparison, we
 1106 constructed a restrictive prior for the logarithm-based model by again fitting the functional form
 1107 to popSTR rates using `scipy.optimize.curve_fit()`. This produced a point estimate of the best-fit
 1108 parameters that was used to specify the mean of a multivariate normal distribution. To attempt to
 1109 define a multivariate normal very roughly comparable to the priors for the power-law
 1110 parameterizations, we defined the normal distribution in terms of the covariance matrix
 1111 estimated under the four-parameter power-law model. This comparison relies on the fact that

1112 both parameterizations use the same parameters to represent nearly identical quantities (i.e.,
 1113 parameters c_ϵ , τ_ϵ , c_κ , and τ_κ define constants and exponents in the same way). The procedure
 1114 described above was then used to define restrictive and informative priors (with 100-fold and
 1115 1000-fold inflated variances, respectively) for two- and four-dimensional logarithm-based
 1116 parameterizations from four-dimensional multivariate normal distributions with appropriately
 1117 shifted means.

1118 To construct priors for pure power-law models, we again used a uniform prior over the
 1119 discrete grid to define a naively uninformative prior over the parameter space. Informative priors
 1120 were defined by again using `scipy.optimize.curve_fit` to estimate best-fit parameters and a
 1121 covariance matrix from empirical instability rate estimates. Unlike the previous models, rates at
 1122 all lengths (including $L \leq 8$) were fully parameterized by **Equation 7** (see also **Table 1**); data
 1123 from both the pooled trio and popSTR datasets was used to estimate the mean and covariance
 1124 matrix by fitting the functional form to expansion and contraction rate estimates for $L = 4\text{--}15$.
 1125 The covariance matrix was again inflated by a factor of 100 and used to define a four-
 1126 dimensional multivariate normal around the point-estimated mean values of $(\lambda_\epsilon, \tau_\epsilon, \lambda_\kappa, \tau_\kappa)$.
 1127

1128 *Calculation of expectation values from posterior probability distribution*

1129 In addition to identifying maximum posterior probability parameter combinations, we
 1130 used the posterior distribution to weight various quantities to calculate their expectation values.
 1131 Expectations of an arbitrary parameter-dependent function $f(\theta)$ were computed as: $E[f(\theta)] =$
 1132 $\int d\theta f(\theta) Pr(\theta|\text{data})$. Here, the ABC-approximated posterior was used for $Pr(\theta|\text{data})$ after
 1133 normalizing over the discrete computational grid of parameters θ such that $E[1] =$
 1134 $\int d\theta Pr(\theta|\text{data}) = 1$. We used this to compute: the length dependence of the posterior-weighted
 1135 repeat instability rates $E[\epsilon(L; \theta)]$ and $E[\kappa(L; \theta)]$ for comparison; and the posterior-weighted
 1136 distribution of repeat tract lengths $E[P(L; \theta)]$ (see **Fig. 5**).
 1137

1138 Analytic modeling of repeat length dynamics

1139 To better understand the underlying dynamics that generate with the genome-wide repeat length
 1140 distribution, we attempted to analytically model the effect of each mutational type on the number
 1141 of repeats at a given length L from first principles. We were interested in describing the steady

1142 state distributions that emerge for a subset of parameter combinations, as seen in the results of
1143 our computational model. Our goal was to capture the balance between relevant mutative forces,
1144 which can vary by repeat length, by writing an appropriate approximation to the steady state
1145 equation; the solutions to these equations describe the shape of the normalized repeat length
1146 distribution, $P(L)$, restricted to the regime of validity of each approximation. Within this section,
1147 we have used the notation P_L to represent the distribution $P(L)$ more compactly when detailing
1148 the relevant equations. Each parameter combination defines a functional form for the per target
1149 (i.e., per unit) expansion, contraction, and non-motif insertion rates at lengths $L \geq 9$:
1150 $\epsilon(L \gg 1) = \epsilon_0 L^{\tau_\epsilon}$, $\kappa(L \gg 1) = \kappa_0 L^{\tau_\kappa}$, and $\iota(L \gg 1) = \iota_0 L^{\tau_\iota}$, respectively, where the constants
1151 ϵ_0 , κ_0 , and ι_0 are set by the empirical value of these rates at $L=8$ and the multiplier m (noting
1152 that we set $\tau_\iota \equiv \tau_\epsilon$ to limit the number of free parameters; see inference Methods). Again, these
1153 length-dependent rates, in either discrete or continuous form, are denoted with a subscript L (e.g.,
1154 $\epsilon_L \equiv \epsilon(L)$) in this section for brevity. For substitutions, we refer herein to rates $\mu \equiv \mu_{A \rightarrow B}$ and
1155 $\nu \equiv \mu_{B \rightarrow A}$ for lengthening and shortening mutations, respectively, but later specify separate
1156 mutation rates based on three-unit context (e.g., $\mu_{ABB \rightarrow AAB}$) when comparing directly to
1157 computational model results. While the mutation rates may be well defined by these rates, the
1158 combined effect of substitutions and indels on the repeat length distribution requires a
1159 description of a number of complicated behaviors, including both local and non-local transitions
1160 between lengths across the distribution, non-conservation of the number of repeats due to fission
1161 and fusion, and non-linear dependence on the state of the distribution due to fusion (i.e., the
1162 generic dynamics are non-Markovian). As a result, our aim was not to describe an exact solution,
1163 but instead an expression for the effective dynamics that dominate the maintenance of the
1164 distribution in steady state, specifically in the asymptotic regimes associated with the shortest
1165 and longest length repeats. Note that this analytic description was motivated by and is strictly
1166 applicable to mononucleotide repeat dynamics, where the species of repeat length-changing
1167 mutations are fewer, but the conceptual findings may be generalizable to longer motif repeats
1168 (**Supplementary Fig. 3**).

1169

1170 *Short repeat regime*

1171 First, we focused on the regime of asymptotically short repeats, as their behavior is more
 1172 straightforward. By assessing the relative rates of substitution and indel processes in the
 1173 estimated per-target rates (**Fig. 2b**), one can immediately see that substitutions must dominate the
 1174 dynamics for the lowest length repeats. Short repeats can be characterized by a straightforward
 1175 balance between opposing types of substitutions, μ and ν , which is equivalent to sequence
 1176 evolution under a two-way point mutation process. At steady state, the resulting distribution is
 1177 equivalent to the probability of randomly assembling specific strings of length L when the whole
 1178 genome is randomly sampled between A and B bases with probability $p_A = \mu/(\mu + \nu)$ and $p_B =$
 1179 $\nu/(\mu + \nu)$, respectively. The frequency of a length L string of A's (i.e., an A repeat) is
 1180 geometrically distributed in proportion to p_A^L (i.e., sampling an A, L successive times).
 1181

$$1182 P_{L \ll 10} \propto \left(\frac{\mu}{\mu + \nu} \right)^L \quad (8)$$

1183
 1184 Here, we have omitted a normalization constant that determines the relative weight of this
 1185 geometric distribution to the weight of the long repeat tail. For comparison to the computational
 1186 model (or the empirical distribution), we fixed the normalization constant using the mass of the
 1187 $L = 1$ bin. The approximation that the effects of expansion, contraction, and non-motif insertion
 1188 are negligible breaks down at a length determined by the estimated relative rates in **Fig. 3b**; the
 1189 regime of validity for this approximation extends roughly to lengths of order $L = 10$.
 1190

1191 *Long repeat regime*

1192 The dynamics of long repeats, i.e., for asymptotically large repeat lengths $L \gg 1$, the
 1193 analysis is complicated by the numerous length-dependent (and parameter-dependent) forces that
 1194 can potentially contribute to stabilizing the distribution. While expansion and contraction
 1195 describe inherently local transitions from L to $L + 1$ and from L to $L - 1$, respectively, the effects
 1196 of non-motif insertions and substitutions on extended repeats are not strictly local. To model this
 1197 regime, we first wrote a finite difference equation that describes the change in the distribution in
 1198 a single time step Δt : $\Delta P_L \equiv P_L(t + \Delta t) - P_L(t)$, where $P_L(t) = P(L, t; \mu, \nu, \epsilon_L, \kappa_L, \iota_L)$ is

1199 implicitly dependent on the length scaling of each rate (see **Supplementary Note**). From this
 1200 discrete equation, we derived a partial differential equation (PDE) in the large-length continuum
 1201 limit $\Delta L = 1 \ll L$ that approximates the dynamics in the large length regime (derivation
 1202 provided in the **Supplementary Note**). This PDE includes explicit terms depicting the combined
 1203 local effects of repeat instability due expansion and contraction, each occurring at distinct length-
 1204 dependent rates, and the separate effects of repeat fission and fusion, each introducing an integral
 1205 that captures the aggregate effects of non-local transitions in length. Expansion and contraction
 1206 collectively generate both symmetric (i.e., bidirectional) and asymmetric local length transitions,
 1207 which correspond to a diffusion term represented by a second derivative and directional flux
 1208 term expressed as a first derivative, respectively, each appropriately accounting for length
 1209 dependent rates.

1210 While local effects from substitutions and non-motif insertions exist (specifically,
 1211 transitions $L \rightarrow L + 1$ or $L \rightarrow L - 1$), as well, they are negligible in comparison to expansion and
 1212 contraction due to their low relative rates at long lengths and finite target size of two per repeat.
 1213 Fission due to substitutions and non-motif insertions were both accounted for as separate non-
 1214 local contributions to the change in P_L . Importantly, the probability of fission due to substitution
 1215 is proportional to the target size $(L - 2) \approx L$; for insertions, the rate itself harbors an additional
 1216 length dependence such that the per-repeat rate of fission scales as $L^{1+\tau_\epsilon}$. As a result, the relative
 1217 importance of fission compared to local contributions is highly dependent on the parameters
 1218 τ_ϵ and τ_κ ; similarly, the relative importance of substitution- and insertion-based fission are
 1219 parameter dependent due to distinct dependencies on length. Thus, a unified description across
 1220 parameter space requires the inclusion of fission in full form and captures all four mutational
 1221 effects. While we were able to explicitly describe the integral effects of length changes due
 1222 repeat fusion in the continuum (see **Supplementary Note**), the inherent non-locality is
 1223 additionally complicated by the nonlinearity introduced by pairing two repeats randomly
 1224 sampled from the distribution. To make further progress, we proceeded under the assumption that
 1225 fusion remains subdominant at large lengths, which we confirmed via our computational model
 1226 to be generically true across parameter space. Stochastic fluctuations in the mutation rates were
 1227 omitted, resulting in a deterministic approximation for the expected repeat length distribution.

1228 Next, we imposed the assumption of steady state (i.e., $dP/dt = 0$), reducing the PDE to
 1229 an ordinary differential equation in length to solve for the shape of the distribution in
 1230 equilibrium. Despite excluding complications from fusion, the remaining approximation to the
 1231 steady state equation is, strictly speaking, a second order integro-differential equation, for which
 1232 no explicit closed-form solutions exist. The following equation approximates the steady state
 1233 dynamics in the absence of fusion (i.e., when fusion is subdominant). Here, ∂_x represents a
 1234 derivative with respect to x (noting that partial derivatives with respect to L become total
 1235 derivatives in steady state) and P_L is the steady state value of the continuous repeat length
 1236 distribution at large length $L \gg 1$ up to an overall normalization constant (along with an arbitrary
 1237 constant set to zero). Again, all continuous functions describing mutation rates (e.g., ϵ_L , κ_L) are
 1238 expressed here as per-target rates.

1239

$$1240 \quad \frac{dP_L}{dt} = 0 \approx \frac{1}{2} \partial_L^2 [(\epsilon_L + \kappa_L)L P_L] - \partial_L [(\epsilon_L - \kappa_L)L P_L] - (\nu + \iota_L)L P_L + 2 \int_L^\infty d\lambda (\nu + \iota_\lambda) P_\lambda \quad (9)$$

1241

1242

1243 In order from left to right, the terms appearing on the right hand side describes: length-dependent
 1244 diffusion (arising from local transitions due to expansion and contraction), a length-dependent
 1245 local directional flux (due to the bias between expansion and contraction), a net loss of due to
 1246 fissions that break up length L repeats (i.e., substitutions or insertions that interrupt the repeat
 1247 sequence; referred to herein as *fission out*), and a net gain due fissions of repeats longer than L
 1248 (referred to as *fission in*). Fission in represents the sole integral effect, which substantially
 1249 complicates our analysis; elimination of the integral dependence is discussed below and results
 1250 in a third order ordinary differential equation (ODE) that maps to this second order integro-
 1251 differential equation.

1252

1253 *Contraction-biased rates stabilize the distribution*

1254 Importantly, we found that steady state could only be reached for the subset of parameter
 1255 combinations with $\tau_\kappa > \tau_\epsilon$, corresponding to cases for which local transitions are asymptotically

1256 contraction-biased: $\lim_{L \rightarrow \infty} (\kappa_L - \epsilon_L) > 0$ (note that the edge case where $\tau_\kappa = \tau_\epsilon$ is asymptotically
 1257 expansion-biased based on observations at $L = 8$ and implications of our parameterization). We
 1258 therefore denote this as the contraction-biased regime, which is characterized by defining the
 1259 variable $\Delta\tau \equiv \tau_\kappa - \tau_\epsilon$. When $\Delta\tau > 0$, the distribution is stabilized at some arbitrarily large length
 1260 $L = L_{\text{trunc}}$ by sufficiently large contraction rates in excess of all processes that increase repeat
 1261 length; a truncation of the distribution (i.e., when less than one repeat is expected in a genome of
 1262 given size) occurs due to the more rapid increase of contraction rates than expansion rates that
 1263 leads to contraction-biased dynamics at some point $L < L_{\text{trunc}}$. The necessity of asymptotic
 1264 contraction-bias contrasts the notion that length-dependent interruptions (due to substitutions and
 1265 non-motif insertions) counteract expansion at sufficiently long lengths, stabilizing the
 1266 distribution^{45,49,55,58,59,60} based on our estimated mutation rates, this effect does not lead to a steady
 1267 state in the absence of contractions, as the per-repeat rate of expansions far exceeds that of repeat
 1268 fission (i.e., interruptions) at long lengths. As discussed below, the length at which the contraction
 1269 rate is equal to the expansion rate L^* (i.e., L^* is the unique length $L \geq 8$ where $\kappa_L = \epsilon_L$, which may
 1270 occur at non-integer values) is highly informative about the dynamics in each regime, as well as
 1271 the behavior when all effects captured in **Equation 9** are simultaneously relevant; L^* is
 1272 exponentially dependent on $\Delta\tau$ and more weakly controlled by the multiplier m , notably occurring
 1273 at the same length across lines of constant $\Delta\tau$ in the parameter space (for a given m). For $m > 2.5$,
 1274 the dynamics are nearly identical for parameter combinations with the same $\Delta\tau$, effectively
 1275 collapsing the $(\tau_\epsilon, \tau_\kappa)$ plane to a single dimension. The functional dependence of L^* on the
 1276 parameters and further discussion is provided in the **Supplementary Note**.

1277

1278 *Effective equations approximating steady state dynamics*

1279 Given the complexity of **Equation 9** introduced by the nonlocal effects of fission, we
 1280 first searched for subsets of the contraction-biased parameter space that could be well
 1281 approximated under a further reduction of the dynamics. Such simplifications are, in principle,
 1282 possible because the length scaling of each term in **Equation 9** is distinct; specifically, parameter
 1283 combinations exist where the nonlocal behavior (i.e., the integral representing fission in)
 1284 becomes subdominant and can be neglected in our analysis. Neglecting the integral results in a
 1285 second order ODE approximation to the steady state equation. We identified two distinct

1286 dynamical regimes within the $\Delta\tau > 0$ region, which are each well-approximated by a subset of
 1287 contributions that dominate the dynamics in their respective regimes of validity.

1288

1289 *Balance between local dynamics in the highly contraction-biased regime*

1290 For parameter combinations with very large positive values of $\Delta\tau$ (i.e., for $\tau_\kappa \gg \tau_\epsilon$), the
 1291 dynamics are entirely dominated by the diffusion and local directional flux terms appearing in
 1292 **Equation 9**, as the contraction rate quickly outcompetes both the rate of fission in and fission
 1293 out. This results in an effective steady state equation dominated only by local transitions.

1294

1295
$$\frac{1}{2} \partial_L^2 [(\epsilon_L + \kappa_L)L P_L] - \partial_L [(\epsilon_L - \kappa_L)L P_L] \approx 0 \quad (10)$$

1296

1297 In this case, the contraction rate exceeds the expansion rate almost immediately above the short
 1298 length regime (i.e., L^* is of order 10) such that the dynamics are effectively uniform across the
 1299 long length regime. The long length tail of the distribution decays in a super-exponential fashion
 1300 such that the truncation occurs at low values of $L_{\text{trunc}} \sim 20$, which dramatically limits the lengths
 1301 of repeats that occur in a genome of realistic size. In this regime, a further simplification leads to
 1302 an approximate closed-form analytic solution for the rough asymptotic shape of the distribution,
 1303 however this approximation is only valid near the truncation point and rapidly loses accuracy. A
 1304 more general solution was obtained by numerically solving the effective steady state equation
 1305 (**Equation 10**) for comparison to computational model results. To obtain numerical values, two
 1306 additional constraints must be applied, as with any second order ODE, which conceptually
 1307 correspond to an overall normalization constant (in this case, fixing the relative weights of the
 1308 short length and long length distributions) and a linear coefficient that defines the relative
 1309 weights of two real solutions, if both exist. These constraints can be imposed by fixing the value
 1310 of the distribution at two specific lengths, L_1 and L_2 , (i.e., fixing $P_{L_1} = P_{L_1}^{\text{comp}}$ and $P_{L_2} = P_{L_2}^{\text{comp}}$,
 1311 where P_L^{comp} is the value of the computationally modeled distribution at length L), with both
 1312 lengths chosen to lie long length regime $L > 10$ where the continuum approximation remains
 1313 valid. For consistency, we chose to constrain the numerical solutions at the two lengths of

1314 theoretical interest in stable distributions: $L_1 = L^*$ (rounded to the nearest integer) and $L_2 =$
 1315 L_{trunc} , both of which definitionally remain in the long length regime at a location with finite
 1316 occupancy in a realistic genome and are well defined for all values of $\Delta\tau > 0$. All numerical
 1317 solutions were obtained using the *NDSolve* function in Mathematica 14.0⁹². Comparisons
 1318 between computational model results and numerical solutions to **Equation 10** showed that this
 1319 approximate steady state equation remains highly accurate across the $\Delta\tau \gg 1$ regime (see
 1320 **Supplementary Note**).

1321

1322 *Relevant effects of fission out in the intermediate contraction-biased regime*

1323 We found that, at less extreme values of $\Delta\tau$, roughly on the order of $\Delta\tau \sim 1$ (e.g., roughly
 1324 $1.5 > \Delta\tau > 0.7$ for $m = 4$), the integral contributions to **Equation 9** remained subdominant, but
 1325 the effects of fission could not be omitted completely. In this regime, fission out non-negligibly
 1326 impacts the dynamics, leading to an effective steady state equation that only omits incoming
 1327 contributions from fission.

1328

1329
$$\frac{1}{2} \partial_L^2 [(\epsilon_L + \kappa_L)L P_L] - \partial_L [(\epsilon_L - \kappa_L)L P_L] - (\nu + \iota_L)L P_L \approx 0 \quad (11)$$

1330

1331 In this regime, contraction is aided by the length-reducing effects of fission out. However, the
 1332 relevance of this contribution is limited roughly to lengths below L^* ; above L^* , the distribution
 1333 remains well-described by **Equation 10** (see **Supplementary Note**). This indicates that
 1334 contraction is largely responsible for truncating the distribution, even when fission is involved in
 1335 shaping the distribution. This defines a range of intermediate lengths below L^* with
 1336 distinguishable dynamics from asymptotic lengths, but this range is limited by the relatively
 1337 small values of L^* on the order of $L^* \sim 15-20$. The approximation in **Equation 11** is again a
 1338 second order ODE but is complicated by the introduction of an additional length scaling
 1339 associated with substitution-based fission. However, even when substitution rates are negligible
 1340 (e.g., for $m \gg 1$), no exact solution could be found due to the generic power laws associated
 1341 with our parameterization. For comparison to the computational model, numerical solutions were
 1342 obtained by again constraining the solution at lengths $L_1 = L^*$ and $L_2 = L_{\text{trunc}}$. We found that the

1343 effective steady state equation (**Equation 11**) is a highly accurate approximation to the dynamics
 1344 in this regime of moderate values of $\Delta\tau$. Additionally, this approximation remains accurate at
 1345 large values of $\Delta\tau$ (i.e., **Equation 11** is applicable to the full subspace $\Delta\tau \gtrsim 1$), as the
 1346 approximation in **Equation 10** is nested in **Equation 11**; the latter includes the additional effect
 1347 of fission out, which becomes negligible for $\Delta\tau \gg 1$.

1348

1349 *Inclusion of the nonlocal dynamics in the weakly contraction-biased regime*

1350 For values $\Delta\tau \lesssim 0.5$ (roughly $\Delta\tau < 0.7$ for $m=4$), the nonlocal effects described by the
 1351 integral term in **Equation 9** become relevant to the maintenance of steady state. To further
 1352 analyze this regime, we first eliminated the integral dependence by applying an overall length
 1353 derivative to all terms on the right-hand side of **Equation 9** such that the equation becomes the
 1354 following.

1355

$$1356 \quad \partial_L \left[\frac{dP_L}{dt} \right] = 0 \approx \frac{1}{2} \partial_L^3 [(\epsilon_L + \kappa_L)L P_L] - \partial_L^2 [(\epsilon_L - \kappa_L)L P_L] - \partial_L [(\nu + \iota_L)L P_L] - 2(\nu + \iota_L)P_L$$

1357 (12)

1358 This third order ODE now represents a constraint on the net flux, which must equal a time-
 1359 independent constant. This can be seen by swapping the order of the derivatives on the left-hand
 1360 side of **Equation 12**: $\partial_L[dP_L/dt] = d[\partial_L P_L]/dt = 0$. Taking this overall length derivative maps
 1361 the nonlocal contributions from the fission of all repeats longer than L to an effectively local
 1362 boundary effect on the net flux $\partial_L P_L$ through length L . However, this is not equivalent to steady
 1363 state until applying an additional constraint that this net flux vanishes (i.e., the special case where
 1364 the constant is zero, $\partial_L P_L = 0$). Obtaining numerical solutions to this third order ODE requires
 1365 three constraints, which now includes the constraint that the net flux vanishes. For comparison to
 1366 the computational model, this was imposed by again specifying $L_1 = L^*$ and $L_2 = L_{\text{trunc}}$ along
 1367 with the additional constraint $P_{L_3} = P_{L_3}^{\text{comp}}$ at length $L_3 = L_2 - 1$, chosen for convenience. We
 1368 found good agreement between the resulting numerical solutions and our computational model
 1369 results. Additionally, solutions to this equation accurately describe the parameter regimes that are
 1370 well approximated by **Equations 10** and **11**, as the latter represent nested dynamics characterized

1371 by **Equation 12** that discard negligible contributions. Thus, **Equation 12** has a regime of validity
1372 that extends across the entire set of parameter combinations that result in stable distributions
1373 $\Delta\tau > 0$. As a corollary, the accuracy of this approximation to the full steady state dynamics
1374 across the space of computational model results indicates that the effects of repeat fusion remain
1375 negligible throughout. However, this statement is only applicable to the long repeat dynamics for
1376 $L > 10$; the effects of repeat fusion are everywhere relevant for short repeats, which, in part,
1377 shape the geometric distribution at steady state.

1378 Details on the derivation, relevant approximations, dynamical regimes, and comparison
1379 between numerical and computational model results are provided in depth in the **Supplementary**
1380 **Note**.

1381

1382 **Data Availability**

1383 The datasets analyzed during the current study are freely available from the NCBI
1384 (<https://www.ncbi.nlm.nih.gov/datasets/genome/>), the UCSC Genome Browser
1385 (<https://genome.ucsc.edu>), and other studies as cited⁸³⁻⁹⁰. Instructions for accessing specific
1386 datasets are further detailed in the code repository (see Code Availability). DRLs for mammalian
1387 genomes analyzed in this study are provided in Supplementary Data 1. Length-dependent
1388 instability rates calculated in this study are provided in Supplementary Data 2.

1389

1390 **Code availability**

1391 The code to perform the analysis in the current study is available in a Github repository
1392 (https://github.com/ryanmcgigg/repeat_distributions)⁹³. Software/packages (including version
1393 numbers) are further detailed therein.

1394

1395 **References**

- 1396 1. Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. *Science*.
1397 2022;376(6588):44-53. doi:10.1126/science.abj6987
1398

- 1399 2. Han J, Hsu C, Zhu Z, Longshore JW, Finley WH. Over-representation of the disease
1400 associated (CAG) and (CGG) repeats in the human genome. *Nucleic Acids Res.*
1401 1994;22(9):1735-1740. doi:10.1093/nar/22.9.1735
- 1402
- 1403 3. Cox R, Mirkin SM. Characteristic enrichment of DNA repeats in different genomes. *Proc Natl
1404 Acad Sci U S A.* 1997;94(10):5237-5242. doi:10.1073/pnas.94.10.5237
- 1405
- 1406 4. Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in
1407 eukaryotic genome sequences. *Mol Biol Evol.* 2001;18(7):1161-1167.
1408 doi:10.1093/oxfordjournals.molbev.a003903
- 1409
- 1410 5. Mirkin SM. Expandable DNA repeats and human disease. *Nature.* 2007;447(7147):932-940.
1411 doi:10.1038/nature05977
- 1412
- 1413 6. Depienne C, Mandel JL. 30 years of repeat expansion disorders: What have we learned and
1414 what are the remaining challenges?. *Am J Hum Genet.* 2021;108(5):764-785.
1415 doi:10.1016/j.ajhg.2021.03.011
- 1416
- 1417 7. Hujoel MLA, Handsaker RE, Kamitaki N, et al. Insights into the causes and consequences of
1418 DNA repeat expansions from 700,000 biobank participants. Preprint. *bioRxiv.*
1419 2024;2024.11.25.625248. Published 2024 Nov 26. doi:10.1101/2024.11.25.625248
- 1420
- 1421 8. Khristich AN, Mirkin SM. On the wrong DNA track: Molecular mechanisms of repeat-
1422 mediated genome instability. *J Biol Chem.* 2020;295(13):4134-4170.
1423 doi:10.1074/jbc.REV119.007678
- 1424
- 1425 9. Wang G, Vasquez KM. Dynamic alternative DNA structures in biology and disease. *Nat Rev
1426 Genet.* 2023;24(4):211-234. doi:10.1038/s41576-022-00539-9
- 1427
- 1428 10. Dolzhenko E, Deshpande V, Schlesinger F, et al. ExpansionHunter: a sequence-graph-based
1429 tool to analyze variation in short tandem repeat regions. *Bioinformatics.* 2019;35(22):4754-4756.
1430 doi:10.1093/bioinformatics/btz431
- 1431
- 1432 11. van Wietmarschen N, Sridharan S, Nathan WJ, et al. Repeat expansions confer WRN
1433 dependence in microsatellite-unstable cancers. *Nature.* 2020;586(7828):292-298.
1434 doi:10.1038/s41586-020-2769-8
- 1435
- 1436 12. Matos-Rodrigues G, van Wietmarschen N, Wu W, et al. S1-END-seq reveals DNA
1437 secondary structures in human cells. *Mol Cell.* 2022;82(19):3538-3552.e5.
1438 doi:10.1016/j.molcel.2022.08.007
- 1439
- 1440 13. Erwin GS, Gürsoy G, Al-Abri R, et al. Recurrent repeat expansions in human cancer
1441 genomes. *Nature.* 2023;613(7942):96-102. doi:10.1038/s41586-022-05515-1
- 1442

- 1443 14. Bacolla A, Tainer JA, Vasquez KM, Cooper DN. Translocation and deletion breakpoints in
1444 cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids*
1445 *Res.* 2016;44(12):5673-5688. doi:10.1093/nar/gkw261
- 1446
- 1447 15. Grasberger H, Dumitrescu AM, Liao XH, et al. STR mutations on chromosome 15q cause
1448 thyrotropin resistance by activating a primate-specific enhancer of MIR7-2/MIR1179. *Nat Genet.*
1449 2024;56(5):877-888. doi:10.1038/s41588-024-01717-7
- 1450
- 1451 16. Kayhanian H, Cross W, van der Horst SEM, et al. Homopolymer switches mediate adaptive
1452 mutability in mismatch repair-deficient colorectal cancer. *Nat Genet.* 2024;56(7):1420-1433.
1453 doi:10.1038/s41588-024-01777-9
- 1454
- 1455 17. Wright SE, Todd PK. Native functions of short tandem repeats. *eLife.* 2023;12:e84043.
1456 Published 2023 Mar 20. doi:10.7554/eLife.84043
- 1457
- 1458 18. Aksanova AY, Mirkin SM. At the Beginning of the End and in the Middle of the Beginning:
1459 Structure and Maintenance of Telomeric DNA Repeats and Interstitial Telomeric Sequences.
1460 *Genes (Basel).* 2019;10(2):118. Published 2019 Feb 5. doi:10.3390/genes10020118
- 1461
- 1462 19. Cox R, Mirkin SM. Characteristic enrichment of DNA repeats in different genomes. *Proc
1463 Natl Acad Sci U S A.* 1997;94(10):5237-5242. doi:10.1073/pnas.94.10.5237
- 1464
- 1465 20. Bell GI, Jurka J. The length distribution of perfect dimer repetitive DNA is consistent with its
1466 evolution by an unbiased single-step mutation process. *J Mol Evol.* 1997;44(4):414-421.
1467 doi:10.1007/pl00006161
- 1468
- 1469 21. Dechering KJ, Cuelenaere K, Konings RN, Leunissen JA. Distinct frequency-distributions of
1470 homopolymeric DNA tracts in different genomes. *Nucleic Acids Res.* 1998;26(17):4056-4062.
1471 doi:10.1093/nar/26.17.4056
- 1472
- 1473 22. Field D, Wills C. Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and
1474 the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from
1475 strong mutation pressures and a variety of selective forces. *Proc Natl Acad Sci U S A.*
1476 1998;95(4):1647-1652. doi:10.1073/pnas.95.4.1647
- 1477
- 1478 23. Kruglyak S, Durrett RT, Schug MD, Aquadro CF. Equilibrium distributions of microsatellite
1479 repeat length resulting from a balance between slippage events and point mutations. *Proc Natl
1480 Acad Sci U S A.* 1998;95(18):10774-10778. doi:10.1073/pnas.95.18.10774
- 1481
- 1482 24. Dokholyan NV, Buldyrev SV, Havlin S, Stanley HE. Distributions of dimeric tandem repeats
1483 in non-coding and coding DNA sequences. *J Theor Biol.* 2000;202(4):273-282.
1484 doi:10.1006/jtbi.1999.1052
- 1485
- 1486 25. Ellegren H. Microsatellite mutations in the germline: implications for evolutionary inference.
1487 *Trends Genet.* 2000;16(12):551-558. doi:10.1016/s0168-9525(00)02139-9
- 1488

- 1489 26. Kruglyak S, Durrett R, Schug MD, Aquadro CF. Distribution and abundance of
1490 microsatellites in the yeast genome can Be explained by a balance between slippage events and
1491 point mutations. *Mol Biol Evol.* 2000;17(8):1210-1219.
1492 doi:10.1093/oxfordjournals.molbev.a026404
1493
- 1494 27. Xu X, Peng M, Fang Z. The direction of microsatellite mutations is dependent upon allele
1495 length. *Nat Genet.* 2000;24(4):396-399. doi:10.1038/74238
1496
- 1497 28. Calabrese PP, Durrett RT, Aquadro CF. Dynamics of microsatellite divergence under
1498 stepwise mutation and proportional slippage/point mutation models. *Genetics.* 2001;159(2):839-
1499 852. doi:10.1093/genetics/159.2.839
1500
- 1501 29. Siby RM, Whittaker JC, Talbot M. A maximum-likelihood approach to fitting equilibrium
1502 models of microsatellite evolution. *Mol Biol Evol.* 2001;18(3):413-417.
1503 doi:10.1093/oxfordjournals.molbev.a003817
1504
- 1505 30. Dieringer D, Schlötterer C. Two distinct modes of microsatellite mutation processes:
1506 evidence from the complete genomic sequences of nine species. *Genome Res.* 2003;13(10):2242-
1507 2251. doi:10.1101/gr.1416703
1508
- 1509 31. Lai Y, Sun F. The relationship between microsatellite slippage mutation rate and the number
1510 of repeat units. *Mol Biol Evol.* 2003;20(12):2123-2131. doi:10.1093/molbev/msg228
1511
- 1512 32. Whittaker JC, Harbord RM, Boxall N, Mackay I, Dawson G, Siby RM. Likelihood-based
1513 estimation of microsatellite mutation rates. *Genetics.* 2003;164(2):781-787.
1514 doi:10.1093/genetics/164.2.781
1515
- 1516 33. Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet.*
1517 2004;5(6):435-445. doi:10.1038/nrg1348
1518
- 1519 34. Messer PW, Arndt PF, Lässig M. Solvable sequence evolution models and genomic
1520 correlations. *Phys Rev Lett.* 2005;94(13):138103. doi:10.1103/PhysRevLett.94.138103
1521
- 1522 35. Eichler EE, Clark RA, She X. An assessment of the sequence gaps: unfinished business in a
1523 finished human genome. *Nat Rev Genet.* 2004;5(5):345-354. doi:10.1038/nrg1322
1524
- 1525 36. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational
1526 challenges and solutions. *Nat Rev Genet.* 2011;13(1):36-46. Published 2011 Nov 29.
1527 doi:10.1038/nrg3117
1528
- 1529 37. Kristmundsdóttir S, Sigurpálsdóttir BD, Kehr B, Halldórsson BV. popSTR: population-scale
1530 detection of STR variants. *Bioinformatics.* 2017;33(24):4041-4048.
1531 doi:10.1093/bioinformatics/btw568
1532

- 1533 38. Jeanjean SI, Shen Y, Hardy LM, et al. A detailed analysis of second and third-generation
1534 sequencing approaches for accurate length determination of short tandem repeats and
1535 homopolymers. *Nucleic Acids Res.* 2025;53(5):gkaf131. doi:10.1093/nar/gkaf131
- 1536
- 1537 39. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and
1538 conservation. *Nature*. 2020;587(7833):240-245. doi:10.1038/s41586-020-2876-6
- 1539
- 1540 40. Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. The genome-wide determinants of
1541 human and chimpanzee microsatellite evolution. *Genome Res.* 2008;18(1):30-38.
1542 doi:10.1101/gr.7113408
- 1543
- 1544 41. Pumpernik D, Oblak B, Borstnik B. Replication slippage versus point mutation rates in short
1545 tandem repeats of the human genome. *Mol Genet Genomics*. 2008;279(1):53-61.
1546 doi:10.1007/s00438-007-0294-1
- 1547
- 1548 42. Molla M, Delcher A, Sunyaev S, Cantor C, Kasif S. Triplet repeat length bias and variation
1549 in the human transcriptome. *Proc Natl Acad Sci U S A*. 2009;106(40):17095-17100.
1550 doi:10.1073/pnas.0907112106
- 1551
- 1552 43. Payseur BA, Jing P, Haasl RJ. A genomic portrait of human microsatellite variation. *Mol
1553 Biol Evol*. 2011;28(1):303-312. doi:10.1093/molbev/msq198
- 1554
- 1555 44. Ananda G, Walsh E, Jacob KD, et al. Distinct mutational behaviors differentiate short
1556 tandem repeats from microsatellites in the human genome. *Genome Biol Evol*. 2013;5(3):606-
1557 620. doi:10.1093/gbe/evs116
- 1558
- 1559 45. Ananda G, Hile SE, Breski A, et al. Microsatellite interruptions stabilize primate genomes
1560 and exist as population-specific single nucleotide polymorphisms within individual human
1561 genomes. *PLoS Genet*. 2014;10(7):e1004498. Published 2014 Jul 17.
1562 doi:10.1371/journal.pgen.1004498
- 1563
- 1564 46. Kristmundsdottir S, Jonsson H, Hardarson MT, et al. Sequence variants affecting the
1565 genome-wide rate of germline microsatellite mutations. *Nat Commun*. 2023;14(1):3855.
1566 Published 2023 Jun 29. doi:10.1038/s41467-023-39547-6
- 1567
- 1568 47. Ellegren H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat
1569 Genet*. 2000;24(4):400-402. doi:10.1038/74249
- 1570
- 1571 48. Mitra I, Huang B, Mousavi N, et al. Patterns of de novo tandem repeat mutations and their
1572 role in autism. *Nature*. 2021;589(7841):246-250. doi:10.1038/s41586-020-03078-7
- 1573
- 1574 49. Kunst CB, Warren ST. Cryptic and polar variation of the fragile X repeat could result in
1575 predisposing normal alleles. *Cell*. 1994;77(6):853-861. doi:10.1016/0092-8674(94)90134-1
- 1576
- 1577 50. Eichler EE, Holden JJ, Popovich BW, et al. Length of uninterrupted CGG repeats determines
1578 instability in the FMR1 gene. *Nat Genet*. 1994;8(1):88-94. doi:10.1038/ng0994-88

- 1579
1580 51. Zoghbi HY, Orr HT. Spinocerebellar ataxia type 1. *Semin Cell Biol.* 1995;6(1):29-35.
1581 doi:10.1016/1043-4682(95)90012-8
1582
1583 52. Latham GJ, Coppinger J, Hadd AG, Nolin SL. The role of AGG interruptions in fragile X
1584 repeat expansions: a twenty-year perspective. *Front Genet.* 2014;5:244. Published 2014 Jul 29.
1585 doi:10.3389/fgene.2014.00244
1586
1587 53. Santoro M, Masciullo M, Silvestri G, Novelli G, Botta A. Myotonic dystrophy type 1: role of
1588 CCG, CTC and CGG interruptions within DMPK alleles in the pathogenesis and molecular
1589 diagnosis. *Clin Genet.* 2017;92(4):355-364. doi:10.1111/cge.12954
1590
1591 54. Wright GEB, Black HF, Collins JA, et al. Interrupting sequence variants and age of onset in
1592 Huntington's disease: clinical implications and emerging therapies. *Lancet Neurol.*
1593 2020;19(11):930-939. doi:10.1016/S1474-4422(20)30343-4
1594
1595 55. Völker J, Breslauer KJ. How sequence alterations enhance the stability and delay expansion
1596 of DNA triplet repeat domains. *QRB Discov.* 2023;4:e8. Published 2023 Nov 6.
1597 doi:10.1017/qrd.2023.6
1598
1599 56. Wilkinson RD. Approximate Bayesian computation (ABC) gives exact results under the
1600 assumption of model error. *Stat Appl Genet Mol Biol.* 2013;12(2):129-141. Published 2013 May
1601 6. doi:10.1515/sagmb-2013-0010
1602
1603 57. Handsaker RE, Kashin S, Reed NM, et al. Long somatic DNA-repeat expansion drives
1604 neurodegeneration in Huntington's disease. *Cell.* 2025;188(3):623-639.e19.
1605 doi:10.1016/j.cell.2024.11.038
1606
1607 58. Rolfsmeier ML, Lahue RS. Stabilizing effects of interruptions on trinucleotide repeat
1608 expansions in *Saccharomyces cerevisiae*. *Mol Cell Biol.* 2000;20(1):173-180.
1609 doi:10.1128/MCB.20.1.173-180.2000
1610
1611 59. Kelkar YD, Eckert KA, Chiaromonte F, Makova KD. A matter of life or death: how
1612 microsatellites emerge in and vanish from the human genome. *Genome Res.* 2011;21(12):2038-
1613 2048. doi:10.1101/gr.122937.111
1614
1615 60. Rajan-Babu IS, Dolzhenko E, Eberle MA, Friedman JM. Sequence composition changes in
1616 short tandem repeats: heterogeneity, detection, mechanisms and clinical implications. *Nat Rev
1617 Genet.* 2024;25(7):476-499. doi:10.1038/s41576-024-00696-z
1618
1619 61. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution.
1620 *Nature.* 1980;284(5757):601-603. doi:10.1038/284601a0
1621
1622 62. Orgel LE, Crick FH. Selfish DNA: the ultimate parasite. *Nature.* 1980;284(5757):604-607.
1623 doi:10.1038/284604a0
1624

- 1625 63. Werren JH, Nur U, Wu CI. Selfish genetic elements. *Trends Ecol Evol.* 1988;3(11):297-302.
1626 doi:10.1016/0169-5347(88)90105-X
1627
- 1628 64. Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, Makova KD. What is a
1629 microsatellite: a computational and experimental definition based upon repeat mutational
1630 behavior at A/T and GT/AC repeats. *Genome Biol Evol.* 2010;2:620-635.
1631 doi:10.1093/gbe/evq046
1632
- 1633 65. Sinden RR, Wells RD. DNA structure, mutations, and human genetic disease. *Curr Opin
1634 Biotechnol.* 1992;3(6):612-622. doi:10.1016/0958-1669(92)90005-4
1635
- 1636 66. Strand M, Prolla TA, Liskay RM, Petes TD. Destabilization of tracts of simple repetitive
1637 DNA in yeast by mutations affecting DNA mismatch repair. *Nature.* 1993;365(6443):274-276.
1638 doi:10.1038/365274a0
1639
- 1640 67. McMurray CT. Mechanisms of DNA expansion. *Chromosoma.* 1995;104(1):2-13.
1641 doi:10.1007/BF00352220
1642
- 1643 68. Pearson CE, Sinden RR. Trinucleotide repeat DNA structures: dynamic mutations from
1644 dynamic DNA. *Curr Opin Struct Biol.* 1998;8(3):321-330. doi:10.1016/s0959-440x(98)80065-1
1645
- 1646 69. Iyer RR, Pluciennik A. DNA Mismatch Repair and its Role in Huntington's Disease. *J
1647 Huntingtons Dis.* 2021;10(1):75-94. doi:10.3233/JHD-200438
1648
- 1649 70. Baptiste BA, Ananda G, Strubczewski N, et al. Mature microsatellites: mechanisms
1650 underlying dinucleotide microsatellite mutational biases in human cells. *G3 (Bethesda).*
1651 2013;3(3):451-463. doi:10.1534/g3.112.005173
1652
- 1653 71. Putnam CD. Strand discrimination in DNA mismatch repair. *DNA Repair (Amst).*
1654 2021;105:103161. doi:10.1016/j.dnarep.2021.103161
1655
- 1656 72. Murat P, Guilbaud G, Sale JE. DNA polymerase stalling at structured DNA constrains the
1657 expansion of short tandem repeats. *Genome Biol.* 2020;21(1):209. Published 2020 Aug 21.
1658 doi:10.1186/s13059-020-02124-x
1659
- 1660 73. Habraken Y, Sung P, Prakash L, Prakash S. Binding of insertion/deletion DNA mismatches
1661 by the heterodimer of yeast mismatch repair proteins MSH2 and MSH3. *Curr Biol.*
1662 1996;6(9):1185-1187. doi:10.1016/s0960-9822(02)70686-6
1663
- 1664 74. Sia EA, Kokoska RJ, Dominska M, Greenwell P, Petes TD. Microsatellite instability in
1665 yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol Cell Biol.*
1666 1997;17(5):2851-2858. doi:10.1128/MCB.17.5.2851
1667
- 1668 75. Gendrel CG, Dutreix M. (CA/TG) microsatellite sequences escape the inhibition of
1669 recombination by mismatch repair in *Saccharomyces cerevisiae*. *Genetics.* 2001;159(4):1539-
1670 1545. doi:10.1093/genetics/159.4.1539

- 1671
1672 76. Jensen LE, Jauert PA, Kirkpatrick DT. The large loop repair and mismatch repair pathways
1673 of *Saccharomyces cerevisiae* act on distinct substrates during meiosis. *Genetics*.
1674 2005;170(3):1033-1043. doi:10.1534/genetics.104.033670
1675
1676 77. Surtees JA, Alani E. Mismatch repair factor MSH2-MSH3 binds and alters the conformation
1677 of branched DNA structures predicted to form during genetic recombination. *J Mol Biol*.
1678 2006;360(3):523-536. doi:10.1016/j.jmb.2006.05.032
1679
1680 78. Panigrahi GB, Slean MM, Simard JP, Gileadi O, Pearson CE. Isolated short CTG/CAG DNA
1681 slip-outs are repaired efficiently by hMutSbeta, but clustered slip-outs are poorly repaired. *Proc
1682 Natl Acad Sci U S A*. 2010;107(28):12593-12598. doi:10.1073/pnas.0909087107
1683
1684 79. Omer S, Lavi B, Mieczkowski PA, Covo S, Hazkani-Covo E. Whole Genome Sequence
1685 Analysis of Mutations Accumulated in *rad27Δ* Yeast Strains with Defects in the Processing of
1686 Okazaki Fragments Indicates Template-Switching Events. *G3 (Bethesda)*. 2017;7(11):3775-
1687 3787. Published 2017 Nov 6. doi:10.1534/g3.117.300262
1688
1689 80. Tsutakawa SE, Thompson MJ, Arvai AS, et al. Phosphate steering by Flap Endonuclease 1
1690 promotes 5'-flap specificity and incision to prevent genome instability. *Nat Commun*.
1691 2017;8:15855. Published 2017 Jun 27. doi:10.1038/ncomms15855
1692
1693 81. Bae SH, Bae KH, Kim JA, Seo YS. RPA governs endonuclease switching during processing
1694 of Okazaki fragments in eukaryotes. *Nature*. 2001;412(6845):456-461. doi:10.1038/35086609
1695
1696 82. Deshmukh AL, Porro A, Mohiuddin M, et al. FAN1, a DNA Repair Nuclease, as a Modifier
1697 of Repeat Expansion Disorders. *J Huntingtons Dis*. 2021;10(1):95-122. doi:10.3233/JHD-
1698 200448
1699
1700 83. Goldmann JM, Wong WS, Pinelli M, et al. Parent-of-origin-specific signatures of de novo
1701 mutations. *Nat Genet*. 2016;48(8):935-939. doi:10.1038/ng.3597
1702
1703 84. Yuen RK, Merico D, Cao H, et al. Genome-wide characteristics of *de novo* mutations in
1704 autism. *NPJ Genom Med*. 2016;1:160271-1602710. doi:10.1038/npjgenmed.2016.27
1705
1706 85. Jónsson H, Sulem P, Kehr B, et al. Parental influence on human germline de novo mutations
1707 in 1,548 trios from Iceland. *Nature*. 2017;549(7673):519-522. doi:10.1038/nature24018
1708
1709 86. An JY, Lin K, Zhu L, et al. Genome-wide de novo risk score implicates promoter variation in
1710 autism spectrum disorder. *Science*. 2018;362(6420):eaat6576. doi:10.1126/science.aat6576
1711
1712 87. Halldorsson BV, Palsson G, Stefansson OA, et al. Characterizing mutagenic effects of
1713 recombination through a sequence-level genetic map. *Science*. 2019;363(6425):eaau1043.
1714 doi:10.1126/science.aau1043
1715

- 1716 88. Sasani TA, Pedersen BS, Gao Z, et al. Large, three-generation human families reveal post-
1717 zygotic mosaicism and variability in germline mutation accumulation. *Elife*. 2019;8:e46922.
1718 Published 2019 Sep 24. doi:10.7554/eLife.46922
1719
1720 89. Jonsson H, Magnusdottir E, Eggertsson HP, et al. Differences between germline genomes of
1721 monozygotic twins. *Nat Genet*. 2021;53(1):27-34. doi:10.1038/s41588-020-00755-1
1722
1723 90. Goes FS, Pirooznia M, Tehan M, et al. De novo variation in bipolar disorder. *Mol Psychiatry*.
1724 2021;26(8):4127-4136. doi:10.1038/s41380-019-0611-1
1725
1726 91. McGinty RJ, Sunyaev SR. Revisiting mutagenesis at non-B DNA motifs in the human
1727 genome. *Nat Struct Mol Biol*. 2023;30(4):417-424. doi:10.1038/s41594-023-00936-6
1728
1729 92. Wolfram Research, Inc., *Mathematica, Version 14.0*, Champaign, IL (2024).
1730
1731 93. McGinty RJ, Balick DJ, Mirkin SM, Sunyaev SR. Inherent instability of simple DNA repeats
1732 shapes an evolutionarily stable distribution of repeat lengths. *Github*.
doi:10.5281/zenodo.17517596 (2025).
1733

1734 **Acknowledgements**

1735 This study was supported by grants from NIGMS R35GM127131 (SS), NIMH R01MH101244
1736 (SS), NHGRI U01HG012009 (SS), NIGMS R35GM130322 (SM) and NSF-BSF 2153071 (SM).
1737 Portions of this research were conducted on the O2 High Performance Compute Cluster,
1738 supported by the Research Computing Group, at Harvard Medical School. We thank Alexey
1739 Kondrashov and Alisa Lyskova for helpful discussions at the early stages of the project.

1740 **Author contributions**

1741 RM conceived the study and performed bioinformatic analyses. RM and DB jointly designed and
1742 implemented the computational model and prepared the manuscript. DB and SS constructed the
1743 analytic model and mathematical analyses. RM, DB and SS devised the statistical inference
1744 procedure. SM and SS supervised the project, discussed results and prepared the manuscript.

1745 **Competing Interests**

1746 The authors declare no competing interests.

1747

Table 1: Repeat instability rate parameterizations and Bayesian inference results.

Model name	Parameters	Functional form	Prior	Bayes factor (ratio to null)	Max posterior [mean posterior]
Symmetric power law [†] (L<9 from empirical rates)	2; $\theta = (c, \tau)$	$\epsilon(L > 8) = \kappa(L > 8) = c \left(\frac{L}{9}\right)^{\tau}$	Uniform	5.1e-18 (1.0)	$\theta = (4.3\text{e-}8, 0.6)$ [(4.3e-8, 0.67)]
			Permissive	2.8e-21 (1.0)	$\theta = (4.3\text{e-}8, 1.2)$ [(4.1e-8, 1.3)]
			Restrictive	7.6e-31 (1.0)	$\theta = (2.7\text{e-}8, 3.0)$ [(3.4e-8, 2.8)]
Decoupled power laws (L<9 empirical)	4; $\theta = (c_{\epsilon}, c_{\kappa}, \tau_{\epsilon}, \tau_{\kappa})$	$\epsilon(L > 8) = c_{\epsilon} \left(\frac{L}{9}\right)^{\tau_{\epsilon}}$ $\kappa(L > 8) = c_{\kappa} \left(\frac{L}{9}\right)^{\tau_{\kappa}}$	Uniform	2.0e-5 (3.9e12)	$\theta = (2.7\text{e-}8, 1.7\text{e-}8, 1.6, 2.0)$ [(2.9e-8, 1.8e-8, 1.9, 2.3)]
			Permissive	1.9e-5 (6.9e15)	$\theta = (2.7\text{e-}8, 1.7\text{e-}8, 3.0, 3.5)$ [(3.9e-8, 2.5e-8, 2.9, 3.3)]
			Restrictive	2.2e-5 (2.9e25)	$\theta = (4.3\text{e-}8, 2.7\text{e-}8, 3.1, 3.6)$ [(4.7e-8, 2.9e-8, 3.1, 3.6)]
Power laws with independent constants (L<9 empirical)	3; $\theta = (c_{\epsilon}, c_{\kappa}, \tau)$	$\epsilon(L > 8) = c_{\epsilon} \left(\frac{L}{9}\right)^{\tau}$ $\kappa(L > 8) = c_{\kappa} \left(\frac{L}{9}\right)^{\tau}$	Uniform	1.3e-5 (2.6e12)	$\theta = (1.7\text{e-}8, 1.1\text{e-}8, 0.9)$ [(2.0e-8, 1.2e-8, 0.6)]
			Permissive	9.9e-14 (3.5e7)	$\theta = (1.7\text{e-}8, 1.1\text{e-}8, 1.0)$ [(1.7e-8, 1.1e-8, 1.0)]
			Restrictive	2.8e-31 (3.7e-1)	$\theta = (2.7\text{e-}8, 2.7\text{e-}8, 3.0)$ [(3.4e-8, 3.4e-8, 2.8)]
Power laws with independent exponents (L<9 empirical)	3; $\theta = (c, \tau_{\epsilon}, \tau_{\kappa})$	$\epsilon(L > 8) = c \left(\frac{L}{9}\right)^{\tau_{\epsilon}}$ $\kappa(L > 8) = c \left(\frac{L}{9}\right)^{\tau_{\kappa}}$	Uniform	9.7e-7 (1.9e11)	$\theta = (2.7\text{e-}8, 0.6, 0.1)$ [(2.7e-8, 2.3, 2.1)]
			Permissive	2.3e-6 (8.2e14)	$\theta = (2.7\text{e-}8, 3.4, 3.3)$ [(2.8e-8, 3.3, 3.2)]
			Restrictive	2.0e-6 (2.6e24)	$\theta = (2.7\text{e-}8, 3.6, 3.5)$ [(2.9e-8, 3.5, 3.4)]
Multiplier-coupled power laws (L<9 empirical)	3; $\theta = (m, \tau_{\epsilon}, \tau_{\kappa})$	$\epsilon(L > 8) = \epsilon(8) \times m \left(\frac{L}{9}\right)^{\tau_{\epsilon}}$ $\kappa(L > 8) = \kappa(8) \times m \left(\frac{L}{9}\right)^{\tau_{\kappa}}$ (empirical $\epsilon(8), \kappa(8)$)	Uniform	1.8e-4 (3.5e13)	$\theta = (2.5, 1.6, 2.0)$ [(2.7, 1.9, 2.3)]
			Permissive	1.4e-4 (4.8e16)	$\theta = (4.0, 2.8, 3.3)$ [(3.6, 2.9, 3.3)]
			Restrictive	1.3e-4 (1.7e26)	$\theta = (4.0, 3.1, 3.6)$ [(4.3, 3.1, 3.6)]
Symmetric logarithmic power ^{††} (L<9 from empirical rates)	2; $\theta = (c, \tau)$	$\epsilon(L > 8) = \kappa(L > 8) = c \left(\frac{\log(L - 7)}{\log 2}\right)^{\tau}$	Uniform	8.4e-18 (1.0)	$\theta = (2.7\text{e-}8, 0.8)$ [(3.2e-8, 0.7)]
			Permissive	3.4e-19 (1.0)	$\theta = (2.7\text{e-}8, 0.9)$ [(3.4e-8, 0.7)]
			Restrictive	8.3e-29 (1.0)	$\theta = (4.3\text{e-}8, 1.0)$ [(4.2e-8, 1.0)]
Decoupled logarithmic power (L<9 empirical)	4; $\theta = (c_{\epsilon}, c_{\kappa}, \tau_{\epsilon}, \tau_{\kappa})$	$\epsilon(L > 8) = c_{\epsilon} \left(\frac{\log(L - 7)}{\log 2}\right)^{\tau_{\epsilon}}$ $\kappa(L > 8) = c_{\kappa} \left(\frac{\log(L - 7)}{\log 2}\right)^{\tau_{\kappa}}$	Uniform	1.8e-4 (2.1e13)	$\theta = (2.7\text{e-}8, 1.1\text{e-}8, 2.1, 2.8)$ [(3.4e-8, 1.4e-8, 1.7, 2.4)]
			Permissive	5.1e-4 (1.5e15)	$\theta = (4.3\text{e-}8, 1.7\text{e-}8, 2.0, 2.7)$ [(3.8e-8, 1.4e-8, 2.1, 2.9)]
			Restrictive	1.8e-3 (2.2e25)	$\theta = (4.3\text{e-}8, 1.7\text{e-}8, 2.0, 2.7)$ [(3.8e-8, 1.3e-8, 2.2, 3.1)]
Multiplier-coupled logarithmic power (L<9 empirical)	3; $\theta = (m, \tau_{\epsilon}, \tau_{\kappa})$	$\epsilon(L > 8) = \epsilon(8) m \left(\frac{\log(L - 7)}{\log 2}\right)^{\tau_{\epsilon}}$ $\kappa(L > 8) = \kappa(8) m \left(\frac{\log(L - 7)}{\log 2}\right)^{\tau_{\kappa}}$ (empirical $\epsilon(8), \kappa(8)$)	Uniform	3.9e-4 (4.6e13)	$\theta = (2.5, 1.3, 0.9)$ [(2.8, 1.0, 1.2)]
			Permissive	2.7e-4 (7.7e14)	$\theta = (4.0, 1.4, 1.7)$ [(3.5, 1.5, 1.8)]
			Restrictive	1.7e-4 (2.0e24)	$\theta = (4.0, 1.8, 2.1)$ [(4.0, 1.9, 2.2)]
Pure power law (parameterized at all lengths)	4; $\theta = (\lambda_{\epsilon}, \lambda_{\kappa}, \tau_{\epsilon}, \tau_{\kappa})$	$\epsilon(L) = \mu \left(\frac{L}{\lambda_{\epsilon}}\right)^{\tau_{\epsilon}}$ $\kappa(L) = \nu \left(\frac{L}{\lambda_{\kappa}}\right)^{\tau_{\kappa}}$ (empirical sub. rates μ, ν)	Uniform	3.1e-17	$\theta = (9.13, 3.6, 4.0)$ [(9.1, 12.9, 3.7, 4.0)]
			Restrictive	8.7e-17	$\theta = (9.12, 3.8, 4.0)$ [(9.0, 12.5, 3.7, 4.0)]

1748

null model for [†]power-law and ^{††}logarithmic power models; **bold**: largest Bayes factor amongst power-law models

1749

1750

1751 **Figure Legends**

1752 **Fig. 1.** *Distributions of repeat tract lengths (DRLs) by motif length and across phylogenies.* **a**
1753 Counts of repeats in human T2T genome pooled by motif unit length (e.g., unit length 1 pools
1754 DRLs for A/T and C/G). Dashed lines represent counts in a randomly-shuffled human genome
1755 sequence. Canonical centromeric and telomeric motifs are excluded from unit lengths 5 and 6,
1756 respectively, due to qualitative differences in the DRLs. **b** Normalized DRLs of mononucleotide-
1757 A repeats in mammals (blue; n=315), primates (orange; n=37) and hominids (green; n=6). (See
1758 **Supplementary Fig. 3** for other motifs.) Counts necessarily normalized to account for different
1759 genome lengths (see **Methods**, **Supplementary Fig. 2**). Solid line indicates median values per
1760 length bin. Phylogenies are inclusive (e.g., primates are included as a subset of mammals). Thin
1761 lines show individual species. Similarity within phylogenies suggests long-term stability of the
1762 DRLs.

1763 **Fig. 2.** *Mutational transitions between repeat tract lengths.* **a** Mutation types, using the example
1764 of transitions to and from repeat tract length $L=6$. ‘A’ represents a given STR motif; ‘B’
1765 represents any other sequence with length equal to A. Arrows indicate mutations, either
1766 substitutions (left) or indels (right), that affect the length of A repeat tract(s). Mutations can
1767 lengthen/shorten (top) or interrupt/rejoin (bottom) repeat tracts. The latter we term repeat
1768 ‘fission’ and ‘fusion,’ respectively. **b** Depiction of the same mutational processes as length
1769 transitions in the DRL. Lengthening/shortening mutations increase or decrease length by one unit
1770 ('local' transitions) and maintain the same total count of repeats ('conservative'), while
1771 fission/fusion processes are non-local and non-conservative.

1772 **Fig. 3.** *Estimated instability rates stratified by repeat tract length.* **a** Dependence of indel size on
1773 repeat tract length (x-axis) for repeat unit lengths 1–4 (different colors). Y-axis measures fraction
1774 of all indels that result in single unit length changes (i.e. number of inserted/deleted bases is
1775 equal to repeat unit length). Point estimates shown for pooled trio (squares; n=9,387) and
1776 popSTR datasets (circles; n=6,084). Above a threshold of ~5 units, repeat instability primarily
1777 consists of +/- 1 unit changes. Tract lengths subject to severe technical artifacts were omitted for
1778 clarity. See **Supplementary Fig. 3** for additional detail. **b** Mononucleotide-A mutation rate
1779 estimates from pooled trio and popSTR datasets for expansions (blue), contractions (orange) and
1780 non-motif insertions (green; note that popSTR dataset combines expansions and non-motif

1781 insertions). Rates calculated only from +/-1 unit changes. Point estimates from pooled trio
 1782 (squares; n=9,387) and popSTR datasets (circles; n=6,084); statistical error bars represent 95%
 1783 confidence intervals assuming Poisson mutation counts. Gray dashed and dotted lines show point
 1784 estimates of substitution rates μ (A>B, where B=C,G or T) and ν (B>A), respectively. Tract
 1785 lengths subject to severe technical artifacts were omitted (see **Supplementary Fig. 4** for
 1786 complete estimates and additional motifs).

1787 **Fig. 4.** *Schematic representation of Bayesian inference procedure.* Inference of parameters
 1788 representing length-dependent instability rates via Approximate Bayesian Computation (ABC).
 1789 Empirical data sources informing the inference are shown in purple. First, expansion and
 1790 contraction rates are parameterized at lengths where estimates are unreliable or unavailable. Each
 1791 parameter combination specifies a complete set of mutation rates for repeats of all lengths. Given
 1792 an initial state, mutational transitions (see **Fig. 2**) are repeatedly applied to evolve the DRL for a
 1793 large number of generations. The late time DRL is treated as an approximation to the steady state
 1794 DRL (when applicable). For each parameter combination, the difference between the late-time
 1795 DRL and the human T2T DRL is summarized by the Kullback-Leibler (KL) divergence. The
 1796 approximate posterior is proportional to the product of the prior and the probability of acceptance
 1797 under an ABC rejection strategy⁵⁶. The prior is informed by popSTR-estimated mutation rates.
 1798 The acceptance probability is treated as Gaussian-distributed in the KL divergence, with mean
 1799 zero and variance defined by divergences for an ensemble of primates (see **Methods**). This
 1800 quantity is calculated for each parameter combination and subsequently normalized over
 1801 parameter space to approximate the posterior probability distribution.

1802 **Fig. 5.** *Inference results and self-consistency of a mutation-only model.* **a** Bayesian posterior
 1803 probabilities inferred for three-parameter multiplier-based power-law model of mono-A repeat
 1804 instability rates; uninformative prior (top row), popSTR-based informative prior (bottom row;
 1805 restrictive condition, see **Supplementary Fig. 5**). Each coordinate represents a distinct set of
 1806 length-dependent rates defined by parameters (m , τ_e , τ_k); τ_e (x-axes) and τ_k (y-axes) determine the
 1807 power laws for expansion and contraction, respectively, and m (columns) represents a
 1808 multiplicative jump at L=9 (parameterization in **Table 1**). Color indicates highest density range
 1809 (HDR) of the posterior for various total probabilities; black region sums to 0.01% of the
 1810 probability. Red arrows show maximum posterior for each prior. Informative prior results in a

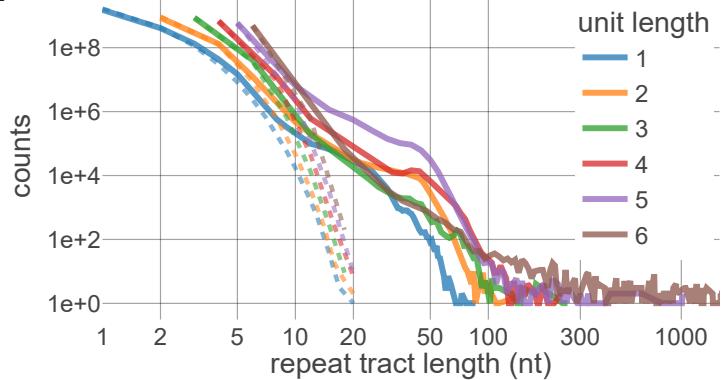
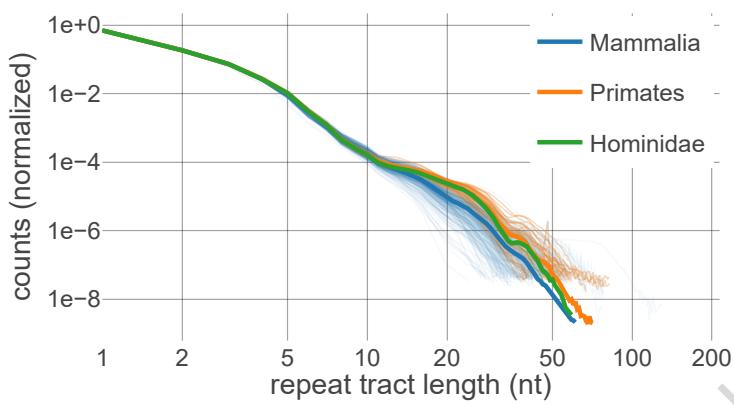
more rapid increase in instability rates with length. **Supplementary Figs. 6–9** show posteriors under various parameterizations. **b** Comparison of inference to empirical DRLs. DRLs are necessarily normalized for comparison (conditional on $L > 3$, see **Methods**); y-axis indicates normalized fractions (parentheses) and counts rescaled to match the number of repeats in the T2T genome (bold labels). Blue lines represent posterior-weighted DRLs (average of all DRLs weighted by the posterior probability for each parameter combination; see **Methods**) for informative (dashed) and uninformative (solid) priors; modeled DRLs are largely consistent with the empirical T2T DRL (black). Green region shows minimum and maximum counts at each length bin across all parameters within the 95% HDR (uninformative prior). Purple region shows min-max range generated from non-human primate genomes ($n=34$, after removing the two most-diverged DRLs and truncating each DRL where raw counts drop below 30; see **Methods**). Overlap between these regions indicates that the posterior under the uninformative prior largely reflects the ensemble of primates. **c** Posterior-weighted repeat instability rates. Tract length dependencies of expansion (blue) and contraction rates (orange) for uninformative (solid) and informative (dashed) priors. Empirical estimates from pooled trios (squares; directly incorporated in model) and popSTR data (circles; used to construct informative priors) are shown for comparison. Informative prior imposes consistency with popSTR-estimated rates, while the posterior-weighted DRL (panel b) remains consistent with the T2T genome.

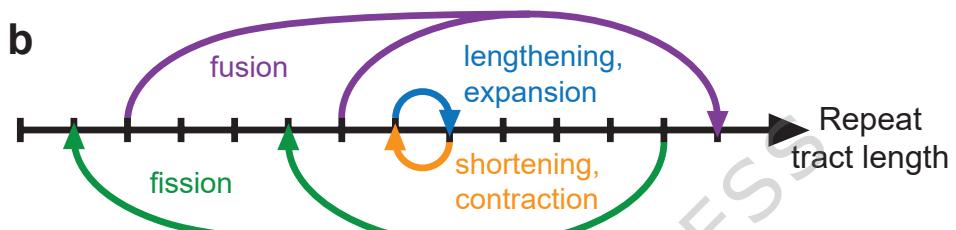
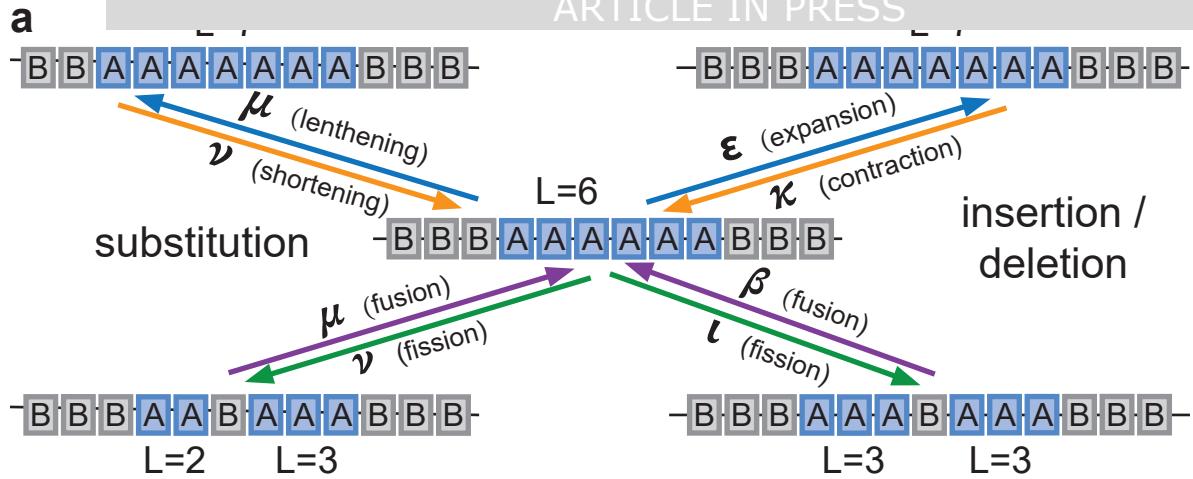
Fig. 6. Dynamical regimes distinguishable by dominant mutational effects. **a** Slice of parameter space for the three-parameter multiplier-based model (parameterization in **Table 1**). Five example parameter combinations with $m=4$ and $\tau_e+\tau_k=3.5$ are shown, corresponding to plots in (b) and (c). Colors roughly divide the parameter space into dynamical regimes. **b** Comparisons between computational model results and numerical solutions of approximate steady state equations (see **Methods**, **Supplementary Note**). Short length regime at equilibrium is geometrically distributed (blue dashed lines). For long repeats, numerical solutions are shown for three nested approximations to the steady state equation in the continuum limit ($L \gg 1$) in the absence of fusion (due to negligible rates); solutions to approximations in Equations 10–12 shown in orange, red and purple, respectively. Local transitions (L to $L+/-1$) were modeled as a combination of symmetric (diffusive) and asymmetric (directional bias) components. L^* represents length at which expansion and contraction rates are equal. For strong asymptotic contraction-bias (b:i), all three approximations remain valid indicating that the dynamics are well

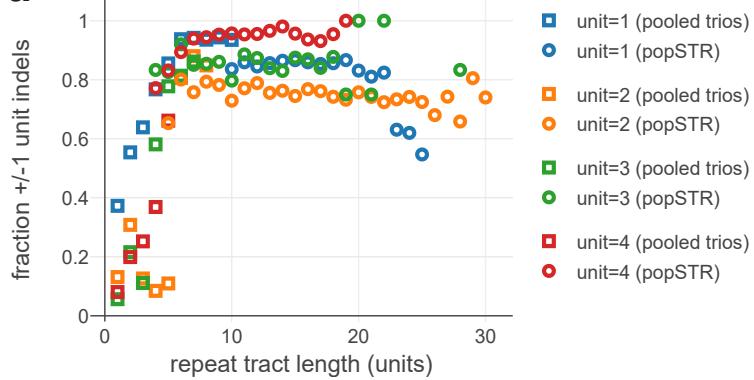
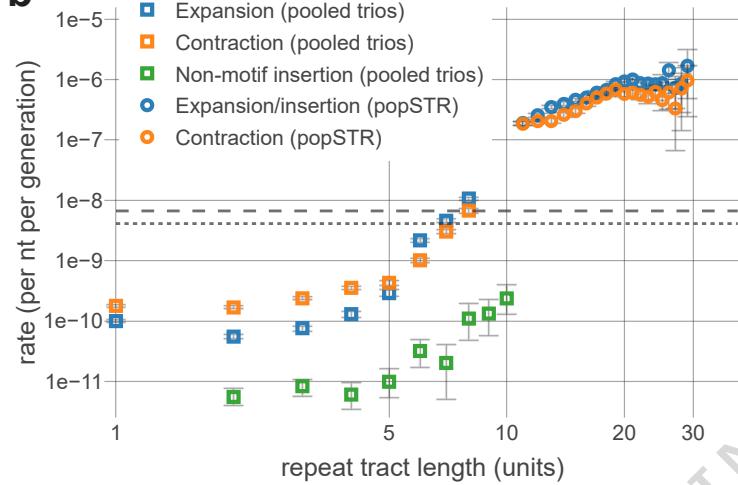
1842 approximated by neglecting fission entirely (orange curve aligns). For moderate contraction bias
 1843 (b:ii), outflux due to fission becomes non-negligible (orange begins to deviate). For realistic
 1844 parameter combinations with lower contraction bias (b:iii), outflux due to fission is required at
 1845 all lengths (orange fails); influx due to fission is required at intermediate lengths (red deviates);
 1846 fusion remains negligible (purple remains accurate). Plots (b:iv-v) display non-equilibrium
 1847 dynamics leading to rapid increase in repeat counts and explosive growth in genome size. Under
 1848 universal expansion bias (b:v), DRL extends indefinitely above the boundary imposed at $L=200$
 1849 (for computational feasibility). Steady-state analytics do not apply. **c** Relative contributions from
 1850 each mutational transition to net flux (in minus out) per length bin, produced by computational
 1851 model. Dashed line indicates DRL truncation (counts <1). Consistent with analytic predictions,
 1852 fission is subdominant under strong contraction bias, has relevant outflux under moderate to
 1853 weak contraction bias, and relevant influx at intermediate lengths under weak contraction bias.
 1854 Equilibrium distribution is stabilized in detailed balance (net influx = outflux). Influx > outflux
 1855 (c:iv-v) leads to nonequilibrium dynamics and indefinite genome growth.

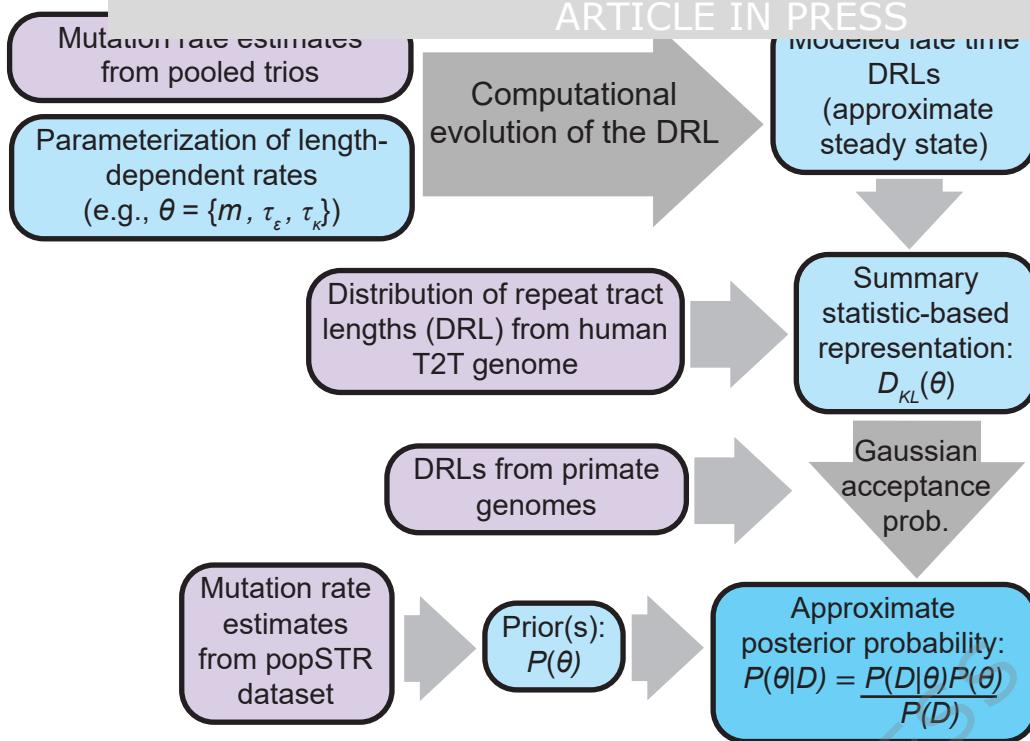
1856 **Table 1:** *Parametric models of instability rates and summary of Bayesian inference results.* For
 1857 each parameterization used in our analyses, this table specifies the model name (as referred to in
 1858 the text), the tract lengths described by the parameterization, the inference parameters, the
 1859 functional forms for length-dependent expansion and contraction rates, and a summary of
 1860 inference results. For each model, the following quantities are given for each prior: Bayes factor
 1861 (and Bayes factor ratio to null model within the same nesting, denoted by symbols), parameter
 1862 combination with maximum posterior probability, and mean posterior parameter combination.
 1863 Distinct nestings separated by bold borders. Primary model considered is shown in bold text.
 1864 Further detail on prior construction, calculation of Bayes factors (and model comparison), and
 1865 expectation used to compute mean posterior parameters provided in **Methods**.

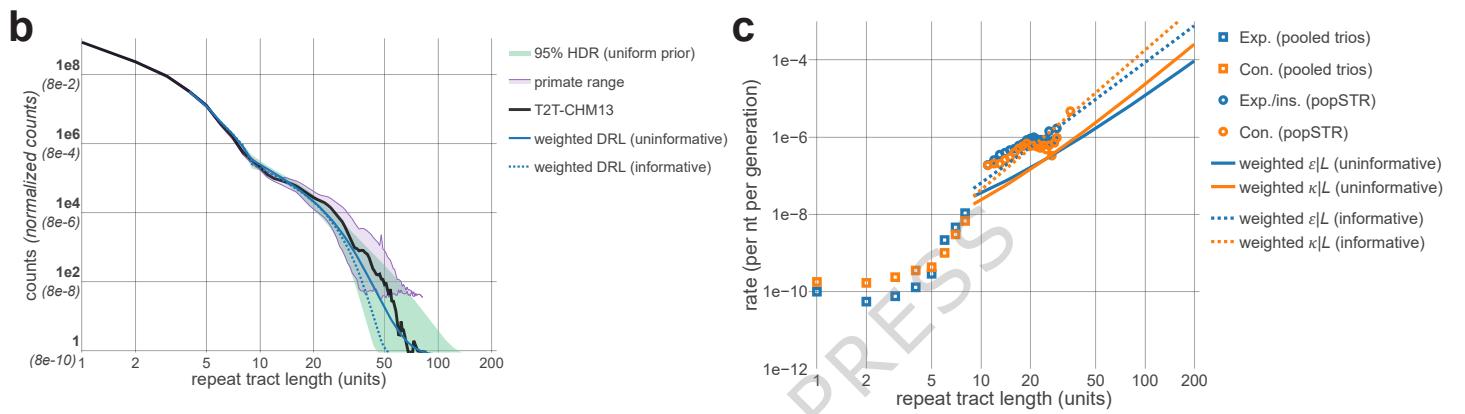
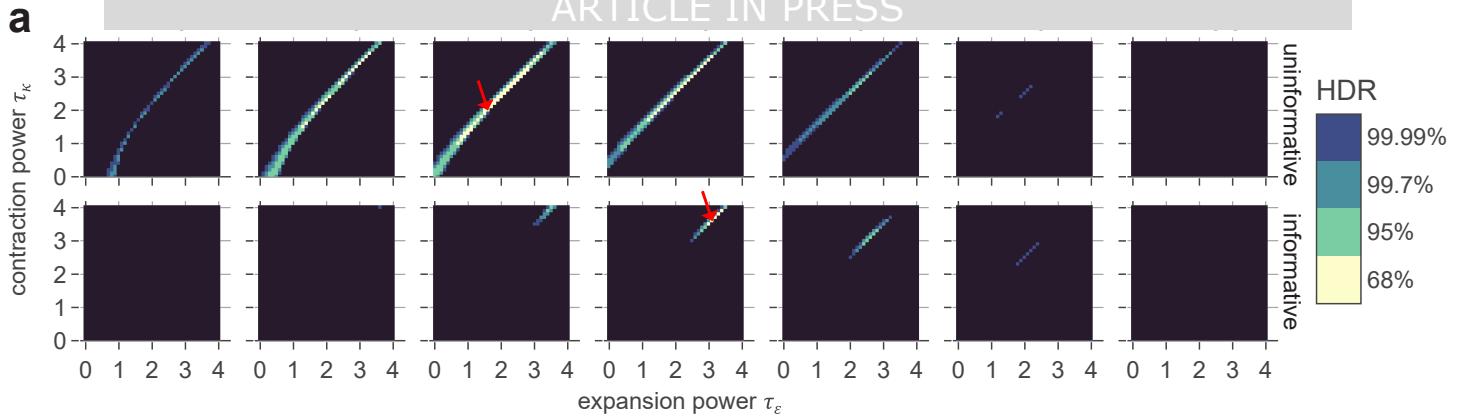
1866

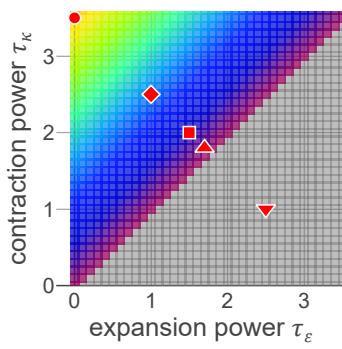
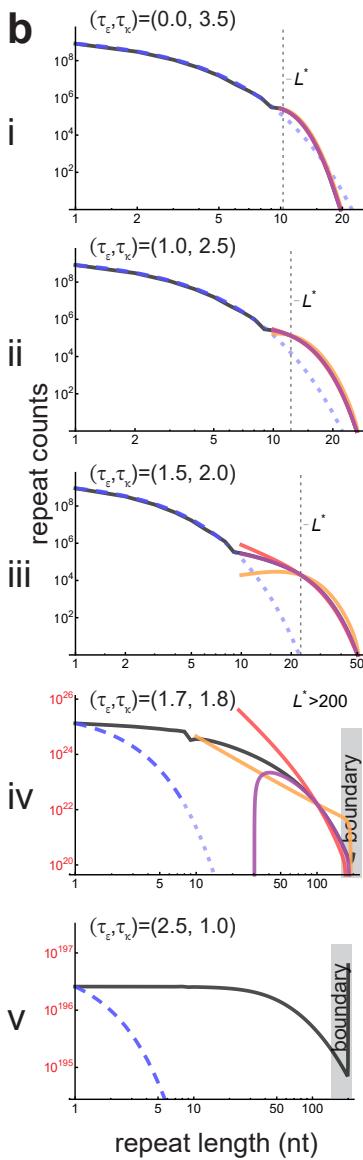
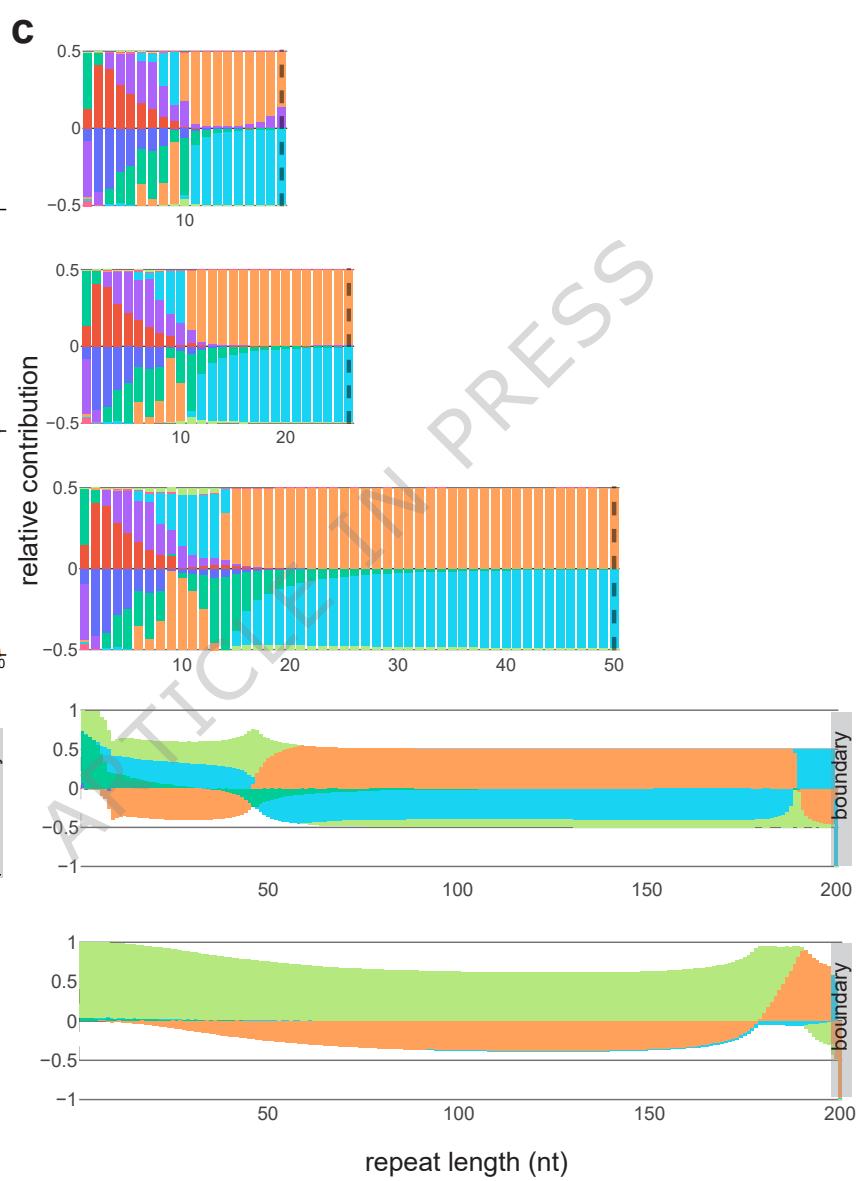
a**b**



a**b**





a**b****c**

— computational model
— dashed line short repeat geometric model

— dotted line geometric model (continued for comparison)

— long repeat analytic model (no fission or fusion)

— long repeat analytic model (fission influx only, no fusion)

— long repeat analytic model (no fusion)

■ substitution -1	■ expansion +1
■ substitution +1	■ contraction -1
■ substitution fission	■ deletion fusion
■ substitution fusion	■ non-motif insertion fission
— count <1	