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

Ryan J. McGinty[1,*], Daniel J. Balick[1,*], Sergei M. Mirkin[2,†] & Shamil R. Sunyaev[1,†]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2] Department of Biology, Tufts University, Medford, MA, USA
[†] Corresponding authors: Sergei Mirkin (Sergei.Mirkin@tufts.edu), Shamil Sunyaev (ssunyaev@hms.harvard.edu).
[*] These authors contributed equally to this work.

Abstract

Using the Telomere-to-Telomere reference, we assembled the distribution of simple repeat lengths present in the human genome. Analyzing over two hundred mammalian genomes, we found remarkable consistency in the shape of the distribution across evolutionary epochs. All observed genomes harbor an excess of long repeats, which are prone to developing into repeat expansion disorders. We measured mutation rates for repeat length instability, quantitatively modeled the per-generation action of mutations, and observed the corresponding long-term behavior shaping the repeat length distribution. We found that short repetitive sequences appear to be a straightforward consequence of random substitution. Evolving largely independently, longer repeats (10+ nucleotides) emerge and persist in a rapidly mutating dynamic balance between expansion, contraction and interruption. These mutational processes, collectively, are sufficient to explain the abundance of long repeats, without invoking natural selection. Our analysis constrains properties of molecular mechanisms responsible for maintaining genome fidelity that underlie repeat instability.

Introduction

Over 2.5% of human genomic DNA consists of simple DNA repeats[1], Also known as short tandem repeats (STRs) or microsatellites, simple repeats consist of direct tandem repetitions of short sequence motifs, e.g., mononucleotides, dinucleotides, trinucleotides and so forth. In a randomized DNA sequence, the probability of encountering a simple repeat decreases as an exponential function of its length. Yet this relationship fails to predict the enormous overrepresentation of long simple repeats in genomic sequences, including in humans[2–4]. The origin of this overrepresentation remains to be elucidated.

This overrepresentation is even more striking in light of the existence of repeat expansion disorders (REDs), a growing list of severe human diseases caused by disruption of gene function owing to long simple repeats[5,6]. Decades of study have demonstrated that individual repeat tracts vary in length between and even within individuals[6], owing to frequent expansion and contraction mutations; the rate of these mutations increases with the length of a repeat, a phenomenon known as repeat length instability[7]. Instability is commonly ascribed to DNA strand slippage during replication and/or DNA repair, although a variety of other molecular mechanisms can also contribute[7]. Instability rates differ between various repeat motifs, being particularly pronounced for motifs that form non-B DNA secondary structures[8]. Importantly, when repeat length exceeds a threshold of approximately 75-90 bp, carriers frequently transmit a substantially longer repeat to the next generation. Known as 'genetic anticipation', this effect continues to compound in subsequent generations, which leads to more severe presentation and/or earlier age of onset[6]. Recently-developed techniques such as ExpansionHunter[9] and long-read sequencing have accelerated the discovery of pathogenic repeats, especially those lurking in introns and other non-coding regions[6]. Repeat expansions occur in various cancers[10-12] and serve as hotspots for genomic rearrangements[13].

While numerous studies focus on the instability of disease-length repeats, comparatively less is known about shorter repeats, including the so-called 'long-normal' alleles that sit immediately below the disease-length threshold. Carriers of long-normal repeat alleles are healthy, but risk transmitting a disease-length allele due to the higher rate of repeat expansion; additionally, some 'long-normal' alleles contain protective interruptions that may be lost[6]. 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 individuals[14]. Additionally, instability of A_8 and C_8 repeats in the coding sequences of mismatch repair genes MSH3 and MSH6, respectively, promotes tumor adaptability via frequent frameshifts and subsequent reversions[15]. The latter examples suggest that relatively short repeats, which comprise a much larger portion of the genome, have at least some medical 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 decades[2] and has led to speculation that natural selection preserves longer repeat lengths despite the risk of disease[16]. The best-supported examples of functionality are specific to telomeric and centromeric repeats[8,16,17], though some recent studies have suggested that simple repeats play a role in gene regulation[16]. 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 are solely a consequence of mutational processes. Several studies, mostly 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[4,18-31]. Indeed, all such studies of simple repeats have been limited by long-standing technical challenges to sequencing repetitive regions[32-34]. 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 GRCH38[1]. 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 explanation for the abundance and stability of long repeats does not require invoking selection; rather, extreme mutation rates cause long repeats to emerge as independently evolving elements. We discuss how inherent constraints of DNA replication and repair machineries could lead to persistent repeat length polymorphism.

Results

Features of the repeat length distribution and evolutionary stability

We first assembled a genome-wide distribution of repeat lengths for each simple tandem repeat motif using T2T-CHM13. We naively compared these to distributions from a randomly-shuffled genome sequence, which confirmed an excess of mid- to long-length repeats in the human genome, including those of near-disease length (**Fig. 1a**). The repeat distribution for a given motif generally consists of at least two qualitatively distinct regimes; short-length repeats (under ~10 nt) appear to be roughly randomly distributed, while mid- to long-length repeats are overrepresented (**Fig. 1a**). We found this to be the case regardless of motif length or sequence, with the long-length tail displaying some variation between motifs (**Fig. 1a, Fig. S1a**).

Despite technological differences in sequencing and assembly between T2T-CHM13 and GRCH38, the genome-wide repeat length distributions are nearly identical after normalization (**Fig. S1b**). We interpret the normalized distribution of repeat lengths, henceforth *P(L)*, as an estimate of the probability of randomly sampling a repeat of length *L* from the distribution of all repeats. Additionally, the estimated *P(L)* from T2T-CHM13 completely overlaps estimates from two individual genomes assembled from moderate coverage short-read sequences (**Fig. S1b**). Thus, short read sequencing, which inherently limits the length of identifiable repeats, is nonetheless sufficient to reconstruct the well-populated length classes of *P(L)*; long-read sequencing is only required to identify extreme length repeats that occur in low numbers.

We estimated distributions from genome sequences of over 200 mammals from the Zoonomia project, which employed a mixture of short- and long-read sequencing[35]. Each estimated *P(L)* was truncated at the first length with occupancy below 30 in the non-normalized distribution to account for differences in assembly length that affect sampling noise. The normalized distributions appear to be highly conserved in hominids, remain so across primate evolution (with the exception of the loris outgroup), and show surprisingly little change through mammalian evolution despite dramatic changes in genome size (**Fig. 1b**, **Fig. S2**). This suggests that both the repeat length distributions and, as a corollary, maintenance of the underlying mechanisms, were largely stable throughout at least 70 million years of primate evolution. The highly conserved evolution of *P(L)* serves as an interesting case in which the empirical observations directly suggest the emergence of a steady state equilibrium.

Given that the distribution is maintained on evolutionary time scales, a prominent extended tail of longer repeats appears to be a generic and conserved feature of *P(L)*. This may be surprising because repeats on the extreme end of this length range have the propensity to develop into repeat-expansion disorders in subsequent generations[5]. One proposed explanation is that longer repeats confer a selective advantage due to some repeat length-specific biological function[16]. As an alternative, we propose that

long repeats emerge and are maintained by the complex interplay between distinct mutational forces. Though these hypotheses are not mutually exclusive, we sought to understand the extent to which mutagenesis, alone, is capable of capturing the shape of the distribution, without introducing natural selection.

To avoid ambiguity related to repeat interruptions, all distributions were assembled by counting contiguous repetitions of each motif. This simultaneously defines length transitions due to each mutation type (either nucleotide substitutions or insertions/deletions), which can have one of four effects on a given repeat: lengthening, shortening, splitting one repeat into two (which we term 'fission'), or joining two repeats into one (which we term 'fusion'). Deletions and insertions of one or more entire repeat motif units are referred to as contractions and expansions, respectively, in contrast to partial deletions and non-motif insertions. The treatment of interrupted repeats as multiple distinct repeats is supported by previous studies, which have found that length-dependence of mutation rates associated with instability depends on the longest contiguous tract within an interrupted repetitive sequence[36-42].

Temporal evolution of the repeat length distribution

To understand the evolution of the repeat length distribution towards a steady state, we built a computational model that incorporates length-changing effects of substitutions, expansions, contractions, and non-motif insertions to follow the equilibration of $P(L)$. To reduce the computational time required to evolve a whole genome sequence and simultaneously track contiguous repeat lengths, we directly manipulated the repeat length distribution, treating the aggregate effects of mutation as deterministic. This deterministic assumption approximates the expectation of $P(L)$ under the repeated application of mutations over many generations, ignoring stochasticity in the mutational process and due to factors like genetic drift. The elementary step of the process is deceptively complex, owing to a combination of non-linear transitions in length (i.e., fusion depends on sampling the length distribution twice in each generation) and changes in the total number of repeats (e.g., fission splits one repeat into two). The computational model uses explicit length-dependent rates for each mutational process as inputs, which, ideally, would be fully defined by de novo rate estimates from parent-child trios. However, existing estimates of this relationship[43-49] lack granular information about each distinct mutagenic process, are available only for particular motifs, and/or are insufficient to estimate rates spanning the full range of observed repeat lengths. Instead, we used de novo data to motivate a parameterization (with minimal degrees of freedom) that characterizes a rapid increase in mutation rates with length, the hallmark of repeat instability. We search for parameters which show a close resemblance between computationally-propagated and empirical distributions.

To arrive at a set of realistic parameters for repeat mutagenesis, we first estimated the rates of each distinct mutagenic process using de novo substitutions and indels from available short-read trio sequencing datasets (n=9,387 trios) (see Methods). The resulting estimates of expansion, contraction and non-motif insertion rates show a rapid increase for lengths between 5-10 nt (**Fig. 2a**, **Fig. S3**) but unexpectedly decrease after ~10nt, inconsistent with our conceptual understanding of repeat length instability. To resolve this, we looked to a recent study from the deCODE group that analyzed instability at mid- to long-length repeats in short-read trio data (n=6,084) using a population structure-aware repeat-length caller named popSTR[49]. We reanalyzed this data to estimate the length-dependency of insertion and deletion rates (here, we specify 'insertion' because popSTR cannot distinguish between expansions and non-motif insertions). Consistent with our understanding of repeat length instability, these rates monotonically increase with length until the estimate becomes noisy (roughly L >20-30 nt, depending on the motif) (**Fig. 2a**, **Fig. S3**). Based on direct contradiction between these two independent estimates, we believe that our de novo estimates display a technical artifact that results from the reduced ability to accurately discriminate length changes within repetitive

sequences longer than 9 nt. However, publicly available popSTR calls omit measurements for $L \leq 10$ nt, extend only to $L<30$ nt, and, again, cannot differentiate expansions from non-motif insertions. Due to this conflation and unknown differences in the error profiles between estimates, we avoided directly merging the rate estimates for subsequent analyses. We further stratified each estimate by the number of units inserted or deleted per mutation (e.g., $L>L+1$, $L>L+2$, $L>L+3$, etc.) and observed that +1/-1 unit changes comprise the vast majority of indels, regardless of repeat length, in both de novo and popSTR data (**Fig. S4**). Because rate estimates for A_n/T_n mononucleotide repeats are the most robust, we focused on this motif in depth. De novo estimates for repeats up to $L=8$ nt (i.e., where they are reliable) were directly incorporated into our computational model. To describe expansion and contraction rates for longer repeats, we introduced three free parameters (**Fig. 2a**): m describes the rate increase from $L=8$ to 9 (multiplying the rate at $L=8$ to generate the $L=9$ point), a second parameter, τ_ϵ, determines the exponent of a power-law(i.e., the exponent τ in the dependence L^τ) representing an instability-driven increase in the per-nucleotide expansion rate beginning at $L=9$, and the third parameter, τ_κ, describes an analogous exponent for the contraction rate. Given that our empirical estimates lose accuracy within a rapidly-changing length range, m can capture a possible continued extreme increase in the rates, oversimplified as a discrete jump at L=9, prior to a transition to power-law like behavior. To limit the number of degrees of freedom, we assumed that the length dependence of non-motif insertions is dictated by τ_ϵ, the expansion rate exponent, due to their parallel increase in de novo rates (**Fig. 2a**) and because they likely arise from the same biological mechanism (e.g., synthesis of the inserted nucleotides by an error-prone polymerase). Power-law length dependencies were chosen as a minimal parameter family of curves that includes the possibility of a constant per-nucleotide rate (i.e., $\tau=0$, analogous to the constant per-nucleotide substitution rates) and linearity (i.e., $\tau=1$), a natural conceptual model for length dependence associated with repeat instability. This parametrization describes the large-length behavior on a scale consistent with the popSTR estimates, which superficially resembles this class of curves (**Fig. 2a**), but is not intended to represent a specific biological model; the true rate curves are likely more complex due to multiple contributing mechanisms.

Consistency between computational model results and empirical data

We analyzed the temporal evolution of the repeat length distribution over a wide range of parameter values to determine if any of the late-time distributions produced by the computational model are consistent with the empirical distribution. After 10^9 generations (sufficient time for substitution-based effects to equilibrate), we measured the distance between the computational and empirical *P(L)* using a goodness-of-fit metric explicitly sensitive to the distribution tail (**Methods**, **Equation 1**). **Figure 2b** displays metric values for a grid of parameters spanning $m=1$–32, $\tau_\epsilon=0$–5, and $\tau_\kappa=0$–5. The parameter combination with lowest metric value ($m=2$, $\tau_\epsilon=1.7$, $\tau_\kappa=2.0$) results in a late-time *P(L)* that closely resembles the observed distribution (**Fig. 2b**). A range of parameter combinations statistically consistent with the minimum appears along lines just above the diagonal defined by $\tau_\epsilon=\tau_\kappa$ for each multiplier (τ_κ-$\tau_\epsilon \approx 0.3$-$0.4$ with a weak m dependence; see Methods for description of statistical procedure). Inconsistent parameters broadly separate into two qualitative categories: τ_κ- $\tau_\epsilon \gtrsim 0.5$ yields distributions that underestimate the long repeat tail, while τ_κ- $\tau_\epsilon \lesssim 0.2$ yields distributions vastly overestimating the long tail, most of which are subject to explosive growth in all length bins (**Figs. S5**, **S6a**). Variation within these two classes is largely characterized by the value of τ_ϵ-τ_κ (**Fig. S5**). Notably, a number of parameter combinations consistent with the best fitting parameters resemble the popSTR-based rate estimates (m~8-16, τ_ϵ and τ_κ~1.5-3.5) (**Fig. 2d**, **S6b**). These popSTR-like parameters become consistent with the best-fitting combination within ~1-5x10^7 generations (**Fig. S6c**), suggesting equilibration can occur within primate-divergence timescales, even from a highly diverged state. Collectively, these results show the plausibility that the observed repeat

length distribution, including the excess of mid-to-long-length repeats, may simply be a result of the interplay between distinct mutational processes, rather than a consequence of selection.

Maintenance of the repeat length distribution in steady state

To understand the complex interplay between mutational processes that shapes and stabilizes the distribution of repeat lengths, we constructed an analytic model of the dynamics. This analytic approximation captures the behavior of *P(L)* after the mutational process reaches steady state (see **Methods** and **Supplementary Note SN1**). A number of previous studies have constructed mathematical models of repeat instability to study repeat length evolution[19,22,25-29,31], including a notable study by Lai and Sun[30] that incorporates many of the elements detailed herein. However, the combination of empirical rate estimates, a robust genome assembly, and our phylogenetic observations motivated the construction of a model from first principles that is directly informed by this collection of observations. In particular, our analytic construction was influenced by the observation of pervasive expansion-contraction bias at most repeat lengths (**Figs. 2a**, **S3**), the incorporation of non-motif insertions, and the primarily single unit changes in length that result from repeat instability in each generation (**Fig S4**).

We first constructed a discrete equation for the change in the number of repeats at a given length in a single generation due to the deterministic action of mutations (i.e., in the absence of selection and stochasticity in the mutational process, consistent with our computational model). We then imposed a steady state condition by requiring that the sum of all changes in and out of each length class vanishes at each time step after equilibration. Despite the simplifying assumption of steady state, the full dynamical equation cannot be solved generically. However, our estimates of de novo mutation rates suggested a dichotomy exists in the primary driver of changes in length between short and longer repeats (i.e., primarily substitutions for L<8 A-mononucleotide repeats vs. expansions and contractions for L>10). Accordingly, short and long repeat dynamics can be treated as separable (i.e., under the approximation of a separation of repeat length scales), leading to simpler approximations of both length regimes. Transitions between the short and long repeat regimes, while present, remain negligible in all realistic scenarios (see **SN1**).

For short repeats, we treated indel mutations as negligible and showed that a geometric distribution (see **Methods**, **Equation 2**) exactly solves the steady state equation under two-way substitutions alone (see **SN1**). For longer repeats, we constructed a partial differential equation (PDE) that approximates the discrete equation and studied its properties in steady state. To facilitate future use cases (e.g., updated empirical estimates or the study of other organisms), the relevant equations are provided in **Supplementary Note SN1** for generic length-dependent rates of expansion, contraction, and non-motif insertion. For direct comparison, we imposed the same parameterization of instability rates used in our computational model. To find the steady-state distributions, we simplified the PDE by treating fusion as negligible to long repeat dynamics (**Methods Equations 3** and **6**, **SN1**) and produced numerical solutions to the time-independent ordinary differential equation (ODE). For parameters that approach steady state, the late time behavior of the computational modeled distribution is accurately described for short and long repeats, respectively, by the geometric distribution and numerical solutions to the ODE without fusion. We further approximated the ODE (**Methods Equations 4** and **5**, **SN1**) to isolate the dominant effects within specific parameter regimes (**Fig. 3a-b**, **SN1**). Using our computational results, we decomposed the per-generation fluxes in and out of each length class into relative contributions from each mutational type to identify dominant mutational processes maintaining steady state (**Fig. 3c, SN1**); the accuracy of each approximation to the PDE was confirmed by analyzing the net magnitudes of fission and fusion within each length class and regime (**Supplementary Note Figs. SN1, SN4-7**).

We used this model to study the shape and stability of the empirical distribution and distinctions between repeats in different length regimes under mutational forces alone. Expansions and contractions remain non-negligible for any long repeat across the space of parameters that lead to stable late-time distributions, highlighting the importance of repeat length instability to the maintenance of long repeats. For extreme parameters that stabilize (i.e., $\tau_\kappa \gg \tau_\epsilon$), the dynamics of all long repeats are dominated by expansion and contraction, alone, leading to a distribution that truncates more rapidly than under substitutions alone (i.e., a *depletion* of long repeats relative to a geometric distribution). In contrast, for parameters consistent with the empirical A-mononucleotide distribution, an intermediate length regime emerges, characterized by the relevance of repeat fission. An accurate description of the shape of the distribution requires fission to account for the loss of repeats from the extreme tail (i.e., the longest populated length bins) and gain of intermediate length repeats. The relative contributions of fission due to substitutions and non-motif insertions are parameter-dependent; within the range of popSTR-consistent parameters, substitution is the primary driver of fission up to lengths of roughly 20 nt, while longer repeats are primarily interrupted by non-motif insertions (see **SN1**). Fission-based losses in the extreme tail are insufficient to fully counteract length increases due to expansion, independent of the mutational mechanism and parameter values. Instead, contraction is primarily responsible for truncating the distribution at finite repeat length but can be bolstered by both substitution- and non-motif insertion-based fission. The dynamics of the long repeat regime decouples from that of short repeats such that rapidly mutating long repeats effectively become independently evolving genomic elements, categorically distinct from random sequences of the same length. The abundance of long repeats in the genome may therefore be a consequence of their largely unencumbered evolution caused by rapid changes in length.

Application to longer motif lengths

Increasing motif length results in successively fewer observed repeats, resulting in statistical noise that precluded direct use of our computational inference procedure. However, extending our analytic understanding beyond mononucleotides is straightforward. For all longer motifs, we observed that expansions and contractions result in predominantly single unit changes in length (**Fig. S4**), suggesting the mechanics of repeat length instability are analogous to mononucleotides. Our estimates show that a sharp increase in instability rates with length is a consistent feature of all motifs (**Fig. S3**), which again suggests separability of length scales into substitution-dominated and instability-dominated dynamical regimes. However, the effective rates of substitutions differ by motif length: a single substitution is sufficient to disrupt any motif, while increasing repeat length for extended motifs can require multiple substitutions (up to the unit length), depending on the sequence context of flanking regions We estimated substitution rates for each motif to account for context dependence and incorporated this into a computational model of substitutions alone (**Methods**). This confirmed that the reduction in length-increasing substitutions results in a more rapid decay of the geometric distribution for short repeats, relative to mononucleotides (**Fig. S7**). Assuming the same instability rates, this would result in a more immediate transition to the instability-dominated regime (counting in number of repeated units L); this is consistent with empirical results for longer unit lengths (**Fig. 1a**, **S1a**). Interestingly, by plotting the distributions (**Figs. 1a**, **S1a**) and rate curves (**Fig. S3**) in total nucleotide length (L x unit length), we found that the transition between the substitution- and instability-dominated regimes occurs in roughly the same range (8-12 nt) for unit lengths of 1-4 nt.

Discussion

Motivated to understand the origin, prevalence and maintenance of simple tandem repeats in the genome, we constructed a model of repeat evolution under mutagenesis alone that bridges short- and long-timescale observations of repeat length instability. We demonstrate that mutation alone is capable of explaining the shape of the genome-wide distribution; furthermore, the observable transition to repeat length instability at some intermediate length, along with an apparent initial expansion bias, is largely responsible for the abundance of long repeats in the genome. This observation does not preclude selection at specific loci, whether beneficial or disease-associated, provided these comprise a small portion of repeats in the genome.

Length-dependent expansion-contraction bias is evident in our de novo estimates; incorporating associated degrees of freedom into the mutational process is sufficient to truncate the distribution at finite lengths due to substantial contraction-bias. This implicitly prevents the growth of repeats to disease-relevant lengths, suggesting natural selection as a disease-prevention mechanism may not be essential. Previous literature suggested that interruptions prevent disease by providing a stopping force that counteracts indefinite expansion[36,42,48,50-52], thereby truncating the distribution; based on our analyses, the rate of interruptions is far too low and must be augmented by contraction. However, truncation of the distribution occurs at lengths inaccessible in our rate estimates, requiring parameterization. If selection, rather than contraction bias, is responsible for terminating the distribution below disease length, it would have to be enormously efficient to counteract instability-driven expansion rates and act globally across all sufficiently long repeats. Further disentangling the roles of mutation and selection will require a concrete model of both processes simultaneously and an accurate measurement of de novo rates across all lengths present in the genomic distribution (which remains a technical challenge).

Analysis of the steady state dynamics led to the qualitative insight that the dynamics of short and long repeats decouple due to the rapid onset, and subsequent dominance, of repeat length instability. Short repeat dynamics are dictated by substitutions alone such that repeats within this regime are roughly indistinguishable from random strings of nucleotides of the same length. The long length tail of the distribution is produced and maintained in a distinct dynamic balance between expansion, contraction, and fission. Long repeats are primarily subject to distinct mutational forces, exhibiting rapid changes in length and a higher rate of repeat fission, which increases their total number; mid-length repeats primarily experience fission due to substitution, while the longest repeats are primarily interrupted by non-motif insertions. In this sense, long repeats emerge as independently evolving genomic elements (with parallels to the concept of selfish genetic elements[53-55]), in stark contrast to short repetitive sequences. Repeat length instability dominates above roughly 10 nt, leading to a natural boundary for the shortest 'unstable' repeat. However, the underlying mutational mechanisms may be better informed by measuring the lowest length where expansion or contraction rates start exceeding the background indel rate (as low as two units for many motifs; **Figs. 2, S3**). The corresponding scientific goals associated with these distinctions (i.e., identifying mechanistic onset vs. differentiating independently evolving elements) are indeed central to a debate concerning the definition of unstable repeats[56].

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

It remains unclear which biological mechanisms control key properties of repeat length instability. The proposed mechanism(s) should be able to explain length dependencies of instability

rates (**Fig. 2a, S3**) that show: a) rapid onset from ~6-10nt, surpassing the rate of substitutions, b) greater-than-linear increase in the expansion/contraction rate per target above 10nt, c) generically asymmetric rates of expansion and contraction with initial expansion-bias and terminal contraction-bias, and d) primarily single-unit changes in length (**Fig. S4**). Surprisingly, these observations appear to be largely independent of both motif sequence and unit length (**Fig. S3**), suggesting a common biological origin.

Two widely studied mechanisms, replication slippage and mismatch repair (MMR), likely explains part of the story[7,8,57-61]. Slippage, when newly synthesized DNA partially unwinds and realigns out of register, should strongly depend on the unit length; however, we see only minor variation associated with unit length (**Fig. S3**). While slippage during DNA replication produces loop-outs on both strands symmetrically[60], subsequent small loop-processing by MMR preferentially results in contractions[62] due to bias towards the nascent strand[63]. Slipped-strand structures may be a motif-independent source of loop-outs subject to the same MMR-processing; in contrast, other secondary structures are motif-specific and therefore cannot be the primary source of repeat instability but can potentially explain differences between motifs[64] (**Fig. S1a**). Importantly, the observation of mostly single-unit expansions/contractions rules out mechanisms involving larger structures (e.g., long hairpins that cannot be processed by MutSβ[65-70]), as these would be expected to generate multi-unit indels.

Single-unit expansions have also been observed in a different context: Okazaki fragment processing by flap-endonuclease FEN1[71]. Incomplete flap removal of the displaced lagging-strand primer may lead to expansion bias[72] and introduce an associated length scale. A secondary mechanism takes over when flaps exceed 30 nt[73]; speculatively, long repeats could give rise to long flaps. This illustrates how different mechanistic explanations may apply to repeats of distinct lengths, generating emergent properties like length-dependent expansion-contraction bias. Likewise, another nuclease, FAN1, is a genetic modifier of several repeat expansion disorders, may be involved in processing slip-outs resulting from replication and transcription, and has differential activity depending on flap length[74].

In sum, our data implies that a simple mutational process accounts for the abundance of mid-length tandem DNA repeats regardless of their sequence composition, which in turn, predisposes the genome for further instability. Subsequent large-scale expansions leading to genetic diseases are attributed to a subset of these repeats that can form stable secondary structures (a.k.a. structure-prone DNA repeats), and they are governed by separate mechanisms discussed elsewhere[7]. Incorporation of a broad range of recent data revealed new insights about long-held tenets of repeat length instability, yet many open questions remain. Inclusion of more comprehensive estimates, experimentation and new analytic approaches may help elucidate the complex biology of instability and evolutionary stability of repetitive sequences in our genomes.

Methods

Genome sources

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

Generation of empirical repeat length distributions

Repeat length distributions were generated by counting consecutive complete motifs (ie. perfect motifs, no interruptions and no partial motifs). For each motif, repeats were located by converting the

sequence of each chromosome into a binary array of matching and non-matching positions via 'regex' parsing, which then allowed for the generation of a histogram according to the length of consecutive motifs (or consecutive non-motifs). For motifs of unit length n>1 nt, the genome was split into n distinct frames, and each frame was used to generate a histogram for each motif. All histograms for equivalent motifs (eg. A=A/T, AC=AC/CA/TG/GT) were then summed. Finally, each histogram was then normalized by the sum of the counts in all length bins to generate a probability distribution *P(L)*.

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

For the various mammalian genomes, the same counting procedure was applied. Assemblies generated from short-read sequencing frequently contain many short contigs which typically originate from poorly-sequenced regions containing transposable elements; any contig of length <10 kb was discarded. Taxonomic data was retrieved from https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/. The median distribution of a given taxonomic group was assembled by gathering the *P(L)* distributions for every member of the group (i.e., for primates this includes humans, and for mammals this includes primates), and taking the value of the median species for each length bin.

Bioinformatic estimation of substitution and indel rates

De novo mutation datasets, representing a total of 10,912 parent-child trios, were acquired as VCF files (or equivalent) from various published sources[75-82]. We assumed that all individuals have the same underlying mutation rates. De novo data was mapped to GRCh38 either in the original study, or subsequently, using pyliftover (pypi.org/project/pyliftover/). Because of the limitations of the VCF file format (e.g., no read depth information on unmutated positions), 100kb regions lacking any substitutions across the combined dataset were assumed to suffer from mappability issues and masked when estimating rates. The average substitution rate was estimated to be $1.2\text{x}10^{-8}$, calculated as: number of substitutions / approximate number of sequenceable nucleotides in the diploid genome / number of offspring genomes in the dataset. We classified substitutions according to six categories based on trinucleotide context and the motif in question, as follows: for the example of A_n motifs, using B to represent non-A nucleotides, we determined rates (in parentheses) of ABB>AAB and BBA>BAA ($4.77\text{x}10^{-9}$) representing repeat-lengthening events, AAB>ABB and BAA>BBA ($8.08\text{x}10^{-9}$) representing repeat-shortening events, ABA>AAA ($2.87\text{x}10^{-9}$) representing fusion events, AAA>ABA ($4.54\text{x}10^{-9}$) representing fissions, BBB>BAB ($3.97\text{x}10^{-9}$) representing the rate of A_1 creation, and BAB>BBB ($6.45\text{x}10^{-9}$) representing the loss of A_1. Rates of substitutions of B which do not create an A were not estimated.

We calculated indel rates as a function of repeat length. Using positional information, upstream and downstream sequences for each event were pulled from the reference genome, under the assumption that the sequence of the parental genome is identical to the reference genome. For every focal motif, we used the reference sequence to determine repeat length. Indel rates per repeat length per motif were estimated by dividing by the number of repeats of that length in the masked GRCh38 genome (see above). Each indel was classified as an expansion, contraction or non-motif insertion, additionally measuring how many motifs were added/removed in the event. We limited mutations in our computational model to +1/-1 unit changes in length at appropriate rates. We also measured the rate of indels for all B positions (with respect to each motif; A_n rates in parentheses), separately estimating the rates of BB>BBB ($1.44\text{x}10^{-10}$), BBB>BB ($4.56\text{x}10^{-12}$) and ABA>AA ($2.89\text{x}10^{-10}$) events. Because

B strings were not modeled as having length-dependent instability, we measured the average rate, i.e., the rate per unit.

The popSTR repeat instability dataset, representing 6,084 parent-child trios, was acquired from the supplement of [49]. Files 'bpinvolved_extended' and 'mutRateDataAll.gz' were downloaded from https://github.com/DecodeGenetics/mDNM_analysisAndData. Due to our focus on uninterrupted repeats, we measured the longest contiguous repeat within the given coordinates for each event. We limited the dataset to loci where the popSTR-reported reference repeat length agreed without our own measurement in GRCh38. The popSTR dataset contains a mix of phased and unphased data; where the parental length for a given mutation was not assigned by phasing, we assumed that it originated from the parental copy which minimizes the difference in length between the proband repeat and any of the parental repeats. Skipping this phasing step and assuming that all events originated from the reference length allele (but retaining the size and direction of the event), as we do for the de novo dataset, results in only minor differences in counts per length bin. The 'mutRateDataAll.gz' file contains information on the number of trios where all three samples passed sequencing quality filters at a given locus, and the length of the repeat at each locus in GRCh38 but lacks information on the parental genotypes for each of these loci. For the denominator of the popSTR mutation rates, we thus generated a distribution of counts using the reference length for each locus multiplied by twice the number of passing trios (for two parental alleles).

For de novo substitution and indel rates, as well as popSTR rates, we calculated confidence intervals based on 200 Poisson samples of the mutation counts, removing the top 5 and bottom 5 values per length bin (see **Fig. 2a**, **Fig. S3**). We determined lines of best-fit (corresponding to empirical estimates of parameters τ_ϵ and τ_κ; see below) using the SciPy linregress package (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.linregress.html) to perform linear regression in log-log space for popSTR estimates corresponding to $A_{12\text{-}19}$ (i.e., the portion of the rate estimates resembling a power law and not subject to large uncertainty). The 99.9% confidence intervals for the slopes (i.e., power law exponents) were used to determine a range of intercepts at L=9 and the corresponding multiplier *m* (see below).

Computational modeling of repeat length dynamics

We used a custom-written script in Python 3 that models repeat dynamics by directly manipulating the distribution of repeat lengths. We simultaneously tracked and manipulated the length distribution of B strings. As detailed above, we assumed a binary genome consisting only of A and B sites, where A is a repeat unit and B represents any non-A unit; as a result, B strings do not *a priori* represent repetitive sequences. Mutations are applied in aggregate such that, in each generation, repeats transition between integer length bins according to rules associated with each mutational process, while the B distribution is updated accordingly (e.g., a substitution that lengthens a repeat simultaneously shortens a B string). Mutation rates were restricted to be sufficiently low to model only a single mutation event per repeat per generation. The non-normalized distribution was evolved and subsequently normalized to create a probability distribution for comparison to empirical data. This approach is far more computationally efficient than simulating an entire genomic sequence, subsequently applying mutations and generating a distribution; computational time in our script scales with the number of length bins rather than with the length of the genome. Tracking only the distribution discards information about the location of particular mutations, instead generating an expected number of mutations for each category per length bin per generation. Except where specified, we used a deterministic approximation to assess the behavior of the expectation value of each bin as the distribution evolves toward steady state via repeated application of the mutation kernel. To understand the impact of stochastic fluctuations on the steady state distribution, we additionally implemented a model that represents fluctuations by Poisson sampling the expected change to each length bin per generation. We model stochastic fluctuations around the applicable rates by sampling mutational

counts, but without constraining individual transitions (i.e., a net number of mutations may leave a given class, but the number introduced elsewhere, as a result, is appropriately distributed only on average due to an independent sampling procedure). All subsequent analyses were performed using the deterministic results, as modeling independent fluctuations in each bin showed no qualitative differences (**Fig. S8a-b**).

Mutations affect the distribution via the following well-defined rules for substitutions and indels. These rules assume that each mutation adds, subtracts or substitutes a complete repeat unit. Using the example of a repeat of L=5, a lengthening substitution subtracts one count from the L=5 bin and adds one to the L=6 bin. A shortening substitution subtracts one from the L=5 bin and adds one to the L=4 bin. A substitution causing repeat fission subtracts one from the L=5 bin and adds two new repeats, either one L=1 and one L=3, or two L=2 (when evolving the distribution as a whole, both occur simultaneously with appropriate relative rates). The reverse process of fission is fusion, in which an L=5 repeat can be generated by fusing an L=1 with an L=3, or by fusing two L=2 repeats, while the mutated B unit is replaced with an A unit and added to the repeat length. Lengthening and shortening substitutions act locally (i.e., counts leave the L bin and move to the adjacent L+1 and L-1 bins, respectively). Substitution of an L=1 in the A distribution also corresponds to fusion of B strings; the reverse, i.e., substitution of a length one B string, generates fusion in the A distribution. Fission and fusion substitutions inherently act non-locally: fission results in the loss of one count in the L bin and gain of two counts that are evenly distributed across all bins of length <=L-2; fusion evenly subtracts two counts from bins <=L-2 to add a count to L. The net effect of substitutions conserves the total length of the genome, i.e., the sum of the length of all A repeats plus the sum of the length of B strings remains constant under substitutions alone.

The rates of lengthening, shortening, fission and fusion substitutions per generation are separately estimated using the three-unit context: BBA>BAA (or ABB>AAB) for lengthening substitutions, AAB>ABB (or BAA>BBA) for shortening substitutions, AAA>ABA for fissions, and ABA>AAA for fusions. All substitution rates were assumed to be independent of repeat length, based on our previous observations showing little to no rate increase with increasing repeat length[83]. The target size for lengthening substitutions is two per repeat (i.e., the two sites adjacent to each repeat boundary). Likewise, the target size for shortening substitutions is also two per repeat, representing the two boundary units of the repeat (assuming L>1). The target size for fission substitutions is L-2 per repeat, representing all non-boundary units within the repeat. The target size for all fusion events is proportional to the L=1 count of the B distribution. Equations governing these processes are described in detail in **Supplementary Note SN1**.

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

To model evolution of the distribution, we apply this mutational process for a large number of iterations i, using the geometric distribution under substitution alone as the initial condition. Low mutation rates associated with low lengths, which are substitution dominated, require excessive

computational time to equilibrate (on the order of the inverse of the substitution rate). To reduce computational time, we introduced a time-rescaling factor r that multiplies all mutation rates by the same constant 10^r such that each iteration represents 10^r generations of evolution for a total of $T=10^{ir}$ generations (assuming a constant r at all time points). To ensure that we remain in the linear mutation regime (i.e., avoid introducing double, triple, etc. mutational events per repeat per generation), r was limited to values where the sum of all mutation rates per repeat per generation remains much less than one (with a maximum of 0.1) in all populated length bins. Imposing this bound inherently limits the space of explorable parameters for a given r; parameters with higher mutation rates must be run with lower r, which takes more computational time. Full exploration of the grid of allowed parameters would require infeasible computational times for low values of r (e.g., r=1), while large values of r (e.g., r=4) exit the linear mutation regime in at least one length class for a greater number of parameter combinations.

To compensate for this, we introduced a progressive time rescaling scheme within each run. Initially, $i=10^4$ iterations are performed at r=5 (for a total of $T=10^9$ generations) and parameters that do not exceed the total mutation rate bound are collected. For parameters that exceed the mutation rate bound at a given r, we set any individual mutation rate at a given length that exceeds 0.1 to 0.1; this results in deviation from a power-law length dependence due to saturation to a constant rate above this length. Using the final distribution from the first stage as the initial distribution, we subsequently repeat this procedure, running $i=10^4$ iterations with r=4, followed by $i=10^4$ iterations with r=3, $i=10^4$ iterations with r=2, $i=10^5$ iterations with r=1, and $i=10^6$ iterations with r=0. This procedure approximates the equilibration for a total of $T\geq10^9$ generations of evolution. While larger values of r are poor approximations to a linear mutation regime (i.e., they aggregate mutations over too many generations, temporarily exceeding the linear mutation bound), the output of each step provides an initial distribution for the subsequent step that is progressively closer to equilibrium. As a result, the distributions reach steady state (when applicable) in far fewer total iterations. For the subset of parameters that could feasibly be run for $T=10^9$ generations without progressive time rescaling, we confirmed that the grid of metric values is nearly identical (**Fig. S8c**) to those produced using the progressive time rescaling scheme (excluding parameter combinations that do not equilibrate, which are expected to differ). This allowed us to effectively model $T\geq10^9$ total generations of evolution to ensure that substitution-based processes with per-repeat rates on the order of 10^{-8} equilibrate to very near steady state.

We imposed a reflective boundary condition at very large lengths; this is necessary due to computation constraints, as the expansion, contraction, and non-motif insertion rates are length-dependent power laws that rapidly increase to unrealistic values at arbitrarily large lengths. To avoid indefinitely increasing rates, the reflective boundary maps all transitions to lengths above $L_{boundary}$ to the bin at $L_{boundary}$. This introduces an artifact in the distribution associated with the change in transition rules representing mutations, which leads to an observable aggregation at the boundary for any computational run resulting in a distribution with finite counts at or beyond $L_{boundary}$. This is primarily to limit computational time spent on unrealistic parameter combinations; this boundary must therefore be imposed at a length substantially longer than the longest repeats considered in the empirical distribution to avoid artifactual boundary effects for plausible parameter combinations. The resulting aggregation at $L_{boundary}$ also facilitates the identification of unstable parameter combinations that show a marked increase in counts at this boundary. When using progressive time-scaling, a count in excess of 1,000 in the L_{boundary} length class triggers the transition to the next speedup stage. L_{boundary} is specified as a command line option (default =100 for both A and B distributions).

The script relies on the following as inputs: an initial distribution for A and B (i.e., motif and non-motif) repeat lengths, per-target substitution rates in three-unit context, and per-target 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_{boundary}). For normalized length distributions that reach steady state, the initial distribution can be chosen arbitrarily; here, we used a distribution generated from the shuffled human genome sequence (T2T-CHM13), which is geometric in both A and B lengths.

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) around the expected number of mutational counts in each iteration.

After each run, we output a file containing repeat length counts reported after every $\Delta T=10^6$ generations 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.

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, 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.

Inference of optimal repeat instability rate curves

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 test the hypothesis that mutational forces alone can generate and stabilize the observed repeat distribution, we first incorporated the full extent of reliable empirical estimates of the substitution and indel rates described above. As discussed above, substitution rates were assumed to be length independent. We focused on A_n repeats, as both the distributions and rate curves were supported by the most empirical data. For expansion, contraction, and non-motif insertion rates for A_n repeats, reliable estimates from trio data extended only to repeat lengths of L=8, beyond which we parameterized the length dependence. Beginning at L=9, each instability rate was parameterized by a multiplier m describing a rapid local increase at L=9 relative to L=8, akin to an initial value of the asymptotic dependence at long lengths; the asymptotic rate curve was parameterized by a single parameter τ defining the exponent of a power law (i.e., rate(L) $\propto m\ \mathrm{L}^{\tau}$) extending from L=9 to a maximum computationally modeled length $L_{boundary}$. We note that the choice of a power-law dependence was arbitrary and instead could be replaced with any family of monotonic single-parameter curves (e.g., exponential growth). Expansion and contraction rates were defined with the same multiplier m, limiting the number of independent parameters, but with separate asymptotic exponents, τ_ϵ and τ_κ, respectively. To further limit the number of inferred parameters, non-motif insertions were fixed to the same asymptotic exponent as expansions. This is consistent with the empirical observation that the rate curves are roughly parallel and the assumption that the same biological mechanisms are responsible for all insertion rates (i.e., expansion and non-motif insertions are both single unit insertions into a repeat). In total, each run required specifying three degrees of freedom (m, τ_ϵ and τ_κ) that collectively parameterize the instability rates at lengths beyond our direct estimates.

Using this parameterization, we ran the computational model over a discrete grid of parameter combinations: multipliers m={1, 2, 4, 8, 16, 32} and exponents in the range $\tau_\epsilon,\tau_\kappa$=[0, 5] spaced in increments of $\Delta\tau$=0.1. These curves are restricted to monotonically increasing rates, consistent with our biological understanding of repeat instability, but includes a lower limit of τ=0 describing no length

dependence above L=8 (i.e., a local jump in rates at L=9, with no further rate increase). Mutation rates are definitionally bounded at a maximum value of 1 (i.e., every target mutates in each generation). Excessive monotonic rate increase (i.e., τ>>1) would result in unrealistically saturated mutation rates at measurable length scales, which is inconsistent with our estimates of indel rates; this corresponds roughly to a maximum power law of τ=5 such that the grid fully explores the space of allowed instability rates.

To assess how closely each computationally modeled distribution $P(L)$ (i.e., $P(L, t| m, \tau_\epsilon, \tau_\kappa)$ for each parameter combination $\{m, \tau_\epsilon, \tau_\kappa\}$ measured at the final time $t=T$) resembled the empirical distribution $P_{\text{emp}}(L)$, we defined the following metric. We compute the metric using the nonnormalized distributions (i.e., per genome counts at each length), denoted below as $\tilde{P}(\text{L})$ and $\tilde{P}_{\text{emp}}(\text{L})$ for the computationally modeled and empirical counts, respectively.

$$metric = \sum_{L=1}^{L=L_{boundary}} \left(log[Norm[\tilde{P}(L)+1]] - log\left[Norm\left[\tilde{P}_{emp}(L)+1\right]\right]\right)^2$$

(Equation 1)

Here, *log* is the natural logarithm and *Norm* indicates that $\tilde{P}(L)$ and $\tilde{P}_{emp}(L)$ are normalized only after adding a pseudocount of one at each length. The pseudocount of one was added to each length class to appropriately handle bins with zero counts. The distribution was then normalized to approximate a probability distribution used to evaluate the metric. This is a least-squares distance metric in logarithmic space (i.e., taking the log of the distribution at each length L), which was chosen to upweight larger length classes that are more informative about the shape of the tail of the distribution relevant to repeat instability. In contrast, a likelihood-based measure, for example, upweights bins with the largest number of counts (i.e., the first few bins L=1, L=2, etc.), which are entirely driven by substitution-based dynamics and therefore uninformative about repeat instability rates. All information about the late-time length distribution of B strings was discarded. This metric value was computed for each parameter combination and minimized to identify the best fit parameter combination.

We then assessed the effect of statistical errors associated with our empirical estimates on our ability to identify parameter combinations consistent with the best fit metric value. The following procedure was used to generate errors associated with the repeat length distribution and subsequent metric value for the best fit parameters. First, for each motif, we estimated statistical errors associated with the empirically estimated mutation rates, by randomly sampling from a Poisson distribution with mean equal to the observed mutation counts; samples were obtained for each relevant three-unit context for substitutions and for each length bin for indels. For A_n motifs, this generated a sampled set of substitution rates and sampled expansion, contraction, and non-motif insertion rate curves for lengths L=1–8. We extended these rate curves using the parameter combination associated with the minimum metric value amongst those tested (i.e., the best fit parameter combination). Using these mutation rates, we repeated our computational procedure to produce a late time repeat length distribution reflecting these sampled rates. This procedure was repeated 200 times, each with a distinct set of sampled rate estimates. From the resulting distributions, we calculated two-sided 95% confidence intervals (CIs) separately around each length bin of the distribution by throwing out the top and bottom five values. Bootstrap errors associated with estimating the number of repeats in the genome (see Methods describing generation of empirical repeat length distributions) are negligible in comparison (**Fig. 2c**) and were not considered for subsequent analyses. To determine which parameter combinations were consistent with the minimum metric value, we then used each set of Poisson sampled rate curves and substitution rates to compute a metric value from the resulting distribution; a one-sided 95% CI was estimated by removing the top 10 metric values. Here, a one-sided CI was used, as all other parameter

combinations have metric values above the minimum. Finally, we found the subset of parameter combinations consistent with the best fit parameters by identifying metric values within the 95% CI of the best fit value.

Analytic modeling of repeat length dynamics

To better understand the underlying dynamics that generate with the genome-wide repeat length distribution, we attempted to analytically model the effect of each mutational type on the number of repeats at a given length L from first principles. We were interested in describing the steady state distributions that emerge for a subset of parameter combinations, as seen in the results of our computational model. Our goal was to capture the balance between relevant mutative forces, which can vary by repeat length, by writing an appropriate approximation to the steady state equation; the solutions to these equations describe the shape of the normalized repeat length distribution, $P(L)$, restricted to the regime of validity of each approximation. Within this section, we have used the notation P_L to represent the distribution $P(L)$ more compactly when detailing the relevant equations. Each parameter combination defines a functional form for the per target (i.e., per unit) expansion, contraction, and non-motif insertion rates at lengths $L \geq 9$: $\epsilon(L \gg 1) = \epsilon_0 L^{\tau_\epsilon}$, $\kappa(L \gg 1) = \kappa_0 L^{\tau_\kappa}$, and $\iota(L \gg 1) = \iota_0 L^{\tau_\iota}$, respectively, where the constants ϵ_0, κ_0 , and ι_0 are set by the empirical value of these rates at L=8 and the multiplier m (noting that we set $\tau_\iota \equiv \tau_e$ to limit the number of free parameters; see inference Methods). Again, these length-dependent rates, in either discrete or continuous form, are denoted with a subscript L (e.g., $\epsilon_L \equiv \epsilon(L)$) in this section for brevity. For substitutions, we refer herein to rates $\mu \equiv \mu_{A\to B}$ and $\nu \equiv \mu_{B\to A}$ for lengthening and shortening mutations, respectively, but later specify separate mutation rates based on three-unit context (e.g., $\mu_{ABB\to AAB}$) when comparing directly to computational model results. While the mutation rates may be well defined by these rates, the combined effect of substitutions and indels on the repeat length distribution requires a description of a number of complicated behaviors, including both local and non-local transitions between lengths across the distribution, non-conservation of the number of repeats due to fission and fusion, and non-linear dependence on the state of the distribution due to fusion (i.e., the generic dynamics are non-Markovian). As a result, our aim was not to describe an exact solution, but instead an expression for the effective dynamics that dominate the maintenance of the distribution in steady state, specifically in the asymptotic regimes associated with the shortest and longest length repeats. Note that this analytic description was motivated by and is strictly applicable to mononucleotide repeat dynamics, where the species of repeat length-changing mutations are fewer, but the conceptual findings may be generalizable to longer motif repeats (**Fig. S3**).

Short repeat regime

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

$$P_{L \ll 10} \propto \left(\frac{\mu}{\mu+\nu}\right)^L \qquad \text{(Equation 2)}$$

Here, we have omitted a normalization constant that determines the relative weight of this geometric distribution to the weight of the long repeat tail. For comparison to the computational model (or the empirical distribution), we fixed the normalization constant using the mass of the $L = 1$ bin. The approximation that the effects of expansion, contraction, and non-motif insertion are negligible breaks down at a length determined by the estimated relative rates in **Fig. 2a**; the regime of validity for this approximation extends roughly to lengths of order $L = 10$.

Long repeat regime

The dynamics of long repeats, i.e., for asymptotically large repeat lengths $L \gg 1$, the analysis is complicated by the numerous length-dependent (and parameter-dependent) forces that can potentially contribute to stabilizing the distribution. While expansion and contraction describe inherently local transitions from L to $L + 1$ and from L to $L - 1$, respectively, the effects of non-motif insertions and substitutions on extended repeats are not strictly local. To model this regime, we first wrote a finite difference equation that describes the change in the distribution in a single time step Δt: $\Delta P_L \equiv P_L(t + \Delta t) - P_L(t)$, where $\mathrm{P_L(t)} = P(L, t; \mu, \nu, \epsilon_L, \kappa_L, \iota_L)$ is implicitly dependent on the length scaling of each rate (see **SN1**). From this discrete equation, we derived a partial differential equation (PDE) in the large-length continuum limit $\Delta L = 1 \ll L$ that approximates the dynamics in the large length regime (derivation provided in **SN1**). This PDE includes explicit terms depicting the combined local effects of repeat instability due expansion and contraction, each occurring at distinct length-dependent rates, and the separate effects of repeat fission and fusion, each introducing an integral that captures the aggregate effects of non-local transitions in length. Expansion and contraction collectively generate both symmetric (i.e., bidirectional) and asymmetric local length transitions, which correspond to a diffusion term represented by a second derivative and directional flux term expressed as a first derivative, respectively, each appropriately accounting for length dependent rates.

While local effects from substitutions and non-motif insertions exist (specifically, transitions $L \to L + 1$ or $L \to L - 1$), as well, they are negligible in comparison to expansion and contraction due to their low relative rates at long lengths and finite target size of two per repeat. Fission due to substitutions and non-motif insertions were both accounted for as separate non-local contributions to the change in P_L. Importantly, the probability of fission due to substitution is proportional to the target size $(L - 2) \approx L$; for insertions, the rate itself harbors an additional length dependence such that the per-repeat rate of fission scales as $L^{1+\tau_\epsilon}$. As a result, the relative importance of fission compared to local contributions is highly dependent on the parameters $\tau_\epsilon$and τ_κ; similarly, the relative importance of substitution- and insertion-based fission are parameter dependent due to distinct dependencies on length. Thus, a unified description across parameter space requires the inclusion of fission in full form and captures all four mutational effects. While we were able to explicitly describe the integral effects of length changes due repeat fusion in the continuum (see **SN1**), the inherent non-locality is additionally complicated by the nonlinearity introduced by pairing two repeats randomly sampled from the distribution. To make further progress, we proceeded under the assumption that fusion remains subdominant at large lengths, which we confirmed via our computational model to be generally true for parameters consistent with the empirical distribution. Stochastic fluctuations in the mutation rates were omitted, resulting in a deterministic approximation for the expected repeat length distribution.

Next, we imposed the assumption of steady state (i.e., $dP/dt = 0$), reducing the PDE to an ordinary differential equation in length to solve for the shape of the distribution in equilibrium. Despite excluding complications from fusion, the remaining approximation to the steady state equation is,

strictly speaking, a second order integro-differential equation, for which no explicit closed-form solutions exist. The following equation approximates the steady state dynamics in the absence of fusion (i.e., when fusion is subdominant). Here, ∂_x represents a derivative with respect to x (noting that partial derivatives with respect to L become total derivatives in steady state) and P_L is the steady state value of the continuous repeat length distribution at large length $L \gg 1$ up to an overall normalization constant (along with an arbitrary constant set to zero). Again, all continuous functions describing mutation rates (e.g., ϵ_L, κ_L) are expressed here as per-target rates.

$$\frac{dP_L}{dt} = 0 \approx \frac{1}{2}\partial_L^2[(\epsilon_L + \kappa_L)L\,P_L] - \partial_L[(\epsilon_L - \kappa_L)L\,P_L] - (\nu + \iota_L)L\,P_L + 2\int_L^\infty d\lambda\,(\nu + \iota_\lambda)\,P_L \qquad \text{(Equation 3)}$$

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

Contraction-biased rates stabilize the distribution

Importantly, we found that steady state could only be reached for the subset of parameter combinations with $\tau_\kappa > \tau_\epsilon$, corresponding to cases for which local transitions are asymptotically contraction-biased: $\lim_{L\to\infty}(\kappa_L - \epsilon_L) > 0$ (note that the edge case where $\tau_\kappa = \tau_\epsilon$ is asymptotically expansion-biased based on observations at $L = 8$ and implications of our parameterization). We therefore denote this as the contraction-biased regime, which is characterized by defining the variable $\Delta\tau \equiv \tau_\kappa - \tau_\epsilon$. When $\Delta\tau > 0$, the distribution is stabilized at some arbitrarily large length $L = L_{max}$ by sufficiently large contraction rates in excess of all processes that increase repeat length; a truncation of the distribution (i.e., when less than one repeat is expected in a genome of given size) occurs due to the more rapid increase of contraction rates than expansion rates that leads to contraction-biased dynamics at some point $L < L_{max}$. The necessity of asymptotic contraction-bias contrasts the notion that length-dependent interruptions (due to substitutions and non-motif insertions) counteract expansion at sufficiently long lengths, stabilizing the distribution[36,42,48,50-52]; based on our estimated mutation rates, this effect does not lead to a steady state in the absence of contractions, as the per-repeat rate of expansions far exceeds that of repeat fission (i.e., interruptions) at long lengths. As discussed below, the length at which the contraction 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 occur at non-integer values) is highly informative about the dynamics in each regime, as well as the behavior when all effects captured in **Equation 3** are simultaneously relevant; L^* is exponentially dependent on $\Delta\tau$ and more weakly controlled by the multiplier m, notably occurring at the same length across lines of constant $\Delta\tau$ in the parameter space (for a given m). For $m \gg 1$, the dynamics are nearly identical for parameter combinations with the same $\Delta\tau$, effectively collapsing the $\{\tau_\epsilon, \tau_\kappa\}$ plane to a single dimension. The functional dependence of L^* on the parameters and further discussion is provided in **Supplementary Note SN1**.

Effective equations approximating steady state dynamics

Given the complexity of **Equation 3** introduced by the nonlocal effects of fission, we first searched for subsets of the contraction-biased parameter space that could be well approximated under a further reduction of the dynamics. Such simplifications are, in principle, possible because the length scaling of each term in **Equation 3** is distinct; specifically, parameter combinations exist where the nonlocal behavior (i.e., the integral representing fission in) becomes subdominant and can be neglected in our analysis. Neglecting the integral results in a second order ODE approximation to the steady state equation. We identified two distinct dynamical regimes within the $\Delta\tau > 0$ region, which are each well-approximated by a subset of contributions that dominate the dynamics in their respective regimes of validity.

Balance between local dynamics in the highly contraction-biased regime

For parameter combinations with very large positive values of $\Delta\tau$ (i.e., for $\tau_\kappa \gg \tau_\epsilon$), the dynamics are entirely dominated by the diffusion and local directional flux terms appearing in **Equation 3**, as the contraction rate quickly outcompetes both the rate of fission in and fission out. This results in an effective steady state equation dominated only by local transitions.

$$\frac{1}{2}\partial_L^2[(\epsilon_L + \kappa_L)L\,P_L] - \partial_L[(\epsilon_L - \kappa_L)L\,P_L] \approx 0 \qquad \text{(Equation 4)}$$

In this case, the contraction rate exceeds the expansion rate almost immediately above the short length regime (i.e., L^* is of order 10) such that the dynamics are effectively uniform across the long length regime. The long length tail of the distribution decays in a super-exponential fashion such that the truncation occurs at low values of $L_{\max} \sim 20$, which dramatically limits the lengths of repeats that occur in a genome of realistic size. In this regime, a further simplification leads to an approximate closed-form analytic solution for the rough asymptotic shape of the distribution, however this approximation is only valid near the truncation point and rapidly loses accuracy. A more general solution was obtained by numerically solving the effective steady state equation (**Equation 4**) for comparison to computational model results. To obtain numerical values, two additional constraints must be applied, as with any second order ODE, which conceptually correspond to an overall normalization constant (in this case, fixing the relative weights of the short length and long length distributions) and a linear coefficient that defines the relative weights of two real solutions, if both exist. These constraints can be imposed by fixing the value of the distribution at two specific lengths, L_1 and L_2, (i.e., fixing $P_{L_1} = P_{L_1}^{comp}$ and $P_{L_2} = P_{L_2}^{comp}$, where P_L^{comp} is the value of the computationally modeled distribution at length L), with both lengths chosen to lie long length regime $L > 10$ where the continuum approximation remains valid. For consistency, we chose to constrain the numerical solutions at the two lengths of theoretical interest in stable distributions: $L_1 = L^*$ (rounded to the nearest integer) and $L_2 = L_{\max}$, both of which definitionally remain in the long length regime at a location with finite occupancy in a realistic genome and are well defined for all values of $\Delta\tau > 0$. All numerical solutions were obtained using the *NDSolve* function in Mathematica 14.0[84]. Comparisons between computational model results and numerical solutions to **Equation 4** showed that this approximate steady state equation remains highly accurate across the $\Delta\tau \gg 1$ regime (see **SN1**).

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

We found that at less extreme values of $\Delta\tau$, roughly on the order of $\Delta\tau \sim 1$, the integral contributions to **Equation 3** remained subdominant, but the effects of fission could not be omitted completely. In this regime, fission out non-negligibly impacts the dynamics, leading to an effective steady state equation that only omits incoming contributions from fission.

$$\frac{1}{2}\partial_L^2[(\epsilon_L + \kappa_L)L\,P_L] - \partial_L[(\epsilon_L - \kappa_L)L\,P_L] - (\nu + \iota_L)L\,P_L \approx 0 \qquad \text{(Equation 5)}$$

In this regime, contraction is aided by the length-reducing effects of fission out. However, the relevance of this contribution is limited roughly to lengths below L^* ; above L^*, the distribution remains well-described by **Equation 4** (see **SN1**). This indicates that contraction is largely responsible for truncating the distribution, even when fission is involved in shaping the distribution. This defines a range of intermediate lengths below L^* with distinguishable dynamics from asymptotic lengths, but this range is limited by the relatively small values of L^* on the order of $L^* \sim 15 - 20$. The approximation in **Equation 5** is again a second order ODE but is complicated by the introduction of an additional length scaling associated with substitution-based fission. However, even when substitution rates are negligible (e.g., for $m \gg 1$), no exact solution could be found due to the generic power laws associated with our parameterization. For comparison to the computational model, numerical solutions were obtained by again constraining the solution at lengths $L_1 = L^*$ and $L_2 = L_{\max}$. We found that the effective steady state equation (**Equation 5**) is a highly accurate approximation to the dynamics in this regime of moderate values of $\Delta\tau$. Additionally, this approximation remains accurate at large values of $\Delta\tau$ (i.e., **Equation 5** is applicable to the full subspace $\Delta\tau \gtrsim 1$), as the approximation in **Equation 4** is nested in **Equation 5**; the latter includes the additional effect of fission out, which becomes negligible for $\Delta\tau \gg 1$.

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

For values $\Delta\tau < 1$, the nonlocal effects described by the integral term in **Equation 3** become relevant to the maintenance of steady state. To further analyze this regime, we first eliminated the integral dependence by applying an overall length derivative to all terms on the right-hand side of **Equation 3** such that the equation becomes the following.

$$\partial_L\left[\frac{dP_L}{dt}\right] = 0 \approx \frac{1}{2}\partial_L^3[(\epsilon_L + \kappa_L)L\,P_L] - \partial_L^2[(\epsilon_L - \kappa_L)L\,P_L] - \partial_L[(\nu + \iota_L)LP_L] - 2(\nu + \iota_L)P_L \qquad \text{(Equation 6)}$$

This third order ODE now represents a constraint on the net flux, which must equal a time-independent constant. This can be seen by swapping the order of the derivatives on the left-hand side of **Equation 6**: $\partial_L[dP_L/dt] = d[\partial_L P_L]/dt = 0$. Taking this overall length derivative maps the nonlocal contributions from the fission of all repeats longer than L to an effectively local boundary effect on the net flux $\partial_L P_L$ through length L. However, this is not equivalent to steady state until applying an additional constraint that this net flux vanishes (i.e., the special case where the constant is zero, $\partial_L P_L = 0$). Obtaining numerical solutions to this third order ODE requires three constraints, which now includes the constraint that the net flux vanishes. For comparison to the computational model, this was imposed by again specifying $L_1 = L^*$ and $L_2 = L_{\max}$ along with the additional constraint $P_{L_3} = P_{L_3}^{comp}$ at length $L_3 = L_2 - 1$, chosen for convenience. We found good agreement between the resulting numerical solutions and our computational model results. Additionally, solutions to this equation accurately describe the parameter regimes that are well approximated by **Equations 4** and **5**, as the latter represent nested dynamics characterized by **Equation 6** that discard negligible contributions. Thus, **Equation 6** has a regime of validity that extends across the entire set of parameter combinations that result in stable distributions $\Delta\tau > 0$. As a corollary, the accuracy of this approximation to the full steady state dynamics across the space of computational model results indicates that the effects of repeat fusion remain negligible throughout. However, this statement is only applicable to the long

repeat dynamics for L>10; the effects of repeat fusion are everywhere relevant for short repeats, which, in part, shape the geometric distribution at steady state.

Details on the derivation, relevant approximations, dynamical regimes, and comparison between numerical and computational model results are provided in depth in **Supplementary Note SN1**.

Data Availability

The datasets analyzed during the current study are freely available from the NCBI (https://www.ncbi.nlm.nih.gov/datasets/genome/), the UCSC Genome Browser (https://genome.ucsc.edu), and other studies as cited. Instructions for accessing specific datasets are further detailed in the code repository (see Code Availability). Repeat length distributions for mammalian genomes analyzed in this study are available in Supplementary Data File SF1. Repeat length instability rates calculated in this study are available in Supplementary Data File SF2.

Code availability

The code to perform the analysis in the current study is available in a Github repository (https://github.com/ryanmcggg/repeat_distributions). For software/packages (with version numbers), please visit the Github repository.

Author contributions

RM conceived the study and performed bioinformatic analyses. RM and DB jointly designed and implemented the computational model and prepared the manuscript. DB and SS constructed the analytic model and mathematical analyses. SM and SS supervised the project, discussed results and prepared the manuscript.

Acknowledgements

Work in the SS lab is supported by grants from NIGMS R35GM127131, NIMH R01MH101244 and NHGRI U01HG012009. Work in the SM lab is supported by grants from NIGMS R35GM130322 and NSF-BSF 2153071. We thank Alexey Kondrashov and Alisa Lyskova for helpful discussions at the early stages of the project.

Competing Interests

The authors declare no competing interests.

References

1. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

2. Han, J., Hsu, C., Zhu, Z., Longshore, J. W. & Finley, W. H. Over-representation of the disease associated (CAG) and (CGG) repeats in the human genome. *Nucleic Acids Res.* **22**, 1735–1740 (1994).

3. Cox, R. & Mirkin, S. M. Characteristic enrichment of DNA repeats in different genomes. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 5237–5242 (1997).

4. Katti, M. V., Ranjekar, P. K. & Gupta, V. S. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.* **18**, 1161–1167 (2001).

5. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447**, 932–940 (2007).

6. Depienne, C. & Mandel, J.-L. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* **108**, 764–785 (2021).

7. Khristich, A. N. & Mirkin, S. M. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.* **295**, 4134–4170 (2020).

8. Wang, G. & Vasquez, K. M. Dynamic alternative DNA structures in biology and disease. *Nat. Rev. Genet.* **24**, 211–234 (2023).

9. Dolzhenko, E. et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).

10. van Wietmarschen, N. et al. Repeat expansions confer WRN dependence in microsatellite-unstable cancers. *Nature* **586**, 292–298 (2020).

11. Matos-Rodrigues, G. et al. S1-END-seq reveals DNA secondary structures in human cells. *Mol. Cell* **82**, 3538–3552.e5 (2022).

12. Erwin, G. S. et al. Recurrent repeat expansions in human cancer genomes. *Nature* **613**, 96–102 (2023).

13. Bacolla, A., Tainer, J. A., Vasquez, K. M. & Cooper, D. N. Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Res.* **44**, 5673–5688 (2016).

14. Grasberger, H. et al. STR mutations on chromosome 15q cause thyrotropin resistance by activating a primate-specific enhancer of MIR7-2/MIR1179. *Nat. Genet.* **56**, 877–888 (2024).

15. Kayhanian, H. et al. Homopolymer switches mediate adaptive mutability in mismatch repair-deficient colorectal cancer. *Nat. Genet.* **56**, 1420–1433 (2024).

16. Wright, S. E. & Todd, P. K. Native functions of short tandem repeats. *Elife* **12**, (2023).

17. Aksenova, A. Y. & Mirkin, S. M. At the Beginning of the End and in the Middle of the Beginning: Structure and Maintenance of Telomeric DNA Repeats and Interstitial Telomeric

Sequences. *Genes* **10**, (2019).

18. Cox, R. & Mirkin, S. M. Characteristic enrichment of DNA repeats in different genomes. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 5237–5242 (1997).

19. Bell, G. I. & Jurka, J. The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. *J. Mol. Evol.* **44**, 414–421 (1997).

20. Dechering, K. J., Cuelenaere, K., Konings, R. N. & Leunissen, J. A. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res.* **26**, 4056–4062 (1998).

21. Field, D. & Wills, C. Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutation pressures and a variety of selective forces. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 1647–1652 (1998).

22. Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 10774–10778 (1998).

23. Dokholyan, N. V., Buldyrev, S. V., Havlin, S. & Stanley, H. E. Distributions of dimeric tandem repeats in non-coding and coding DNA sequences. *J. Theor. Biol.* **202**, 273–282 (2000).

24. Ellegren, H. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet.* **16**, 551–558 (2000).

25. Kruglyak, S., Durrett, R., Schug, M. D. & Aquadro, C. F. Distribution and abundance of microsatellites in the yeast genome can Be explained by a balance between slippage events and point mutations. *Mol. Biol. Evol.* **17**, 1210–1219 (2000).

26. Xu, X., Peng, M. & Fang, Z. The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* **24**, 396–399 (2000).

27. Calabrese, P. P., Durrett, R. T. & Aquadro, C. F. Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. *Genetics* **159**, 839–852 (2001).

28. Sibly, R. M., Whittaker, J. C. & Talbot, M. A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. *Mol. Biol. Evol.* **18**, 413–417 (2001).

29. Dieringer, D. & Schlötterer, C. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Res.* **13**, 2242–2251 (2003).

30. Lai, Y. & Sun, F. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol. Biol. Evol.* **20**, 2123–2131 (2003).

31. Whittaker, J. C. et al. Likelihood-based estimation of microsatellite mutation rates. *Genetics* **164**, 781–787 (2003).

32. Eichler, E. E., Clark, R. A. & She, X. An assessment of the sequence gaps: unfinished business in

a finished human genome. *Nat. Rev. Genet.* **5**, 345–354 (2004).

33. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2011).

34. Kristmundsdóttir, S., Sigurpálsdóttir, B. D., Kehr, B. & Halldórsson, B. V. popSTR: population-scale detection of STR variants. *Bioinformatics* **33**, 4041–4048 (2017).

35. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).

36. Kunst, C. B. & Warren, S. T. Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. *Cell* **77**, 853–861 (1994).

37. Eichler, E. E. et al. Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nat. Genet.* **8**, 88–94 (1994).

38. Zoghbi, H. Y. & Orr, H. T. Spinocerebellar ataxia type 1. *Semin. Cell Biol.* **6**, 29–35 (1995).

39. Latham, G. J., Coppinger, J., Hadd, A. G. & Nolin, S. L. The role of AGG interruptions in fragile X repeat expansions: a twenty-year perspective. *Front. Genet*. **5**, 244 (2014).

40. Santoro, M., Masciullo, M., Silvestri, G., Novelli, G. & Botta, A. Myotonic dystrophy type 1: role of CCG, CTC and CGG interruptions within DMPK alleles in the pathogenesis and molecular diagnosis. *Clin. Genet.* **92**, 355–364 (2017).

41. Wright, G. E. B. et al. Interrupting sequence variants and age of onset in Huntington's disease: clinical implications and emerging therapies. *Lancet Neurol.* **19**, 930–939 (2020).

42. Völker, J. & Breslauer, K. J. How sequence alterations enhance the stability and delay expansion of DNA triplet repeat domains. *QRB Discov*. **4**, e8 (2023).

43. Kelkar, Y. D., Tyekucheva, S., Chiaromonte, F. & Makova, K. D. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.* **18**, 30–38 (2008).

44. Pumpernik, D., Oblak, B. & Borstnik, B. Replication slippage versus point mutation rates in short tandem repeats of the human genome. *Mol. Genet. Genomics* **279**, 53–61 (2008).

45. Molla, M., Delcher, A., Sunyaev, S., Cantor, C. & Kasif, S. Triplet repeat length bias and variation in the human transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 17095–17100 (2009).

46. Payseur, B. A., Jing, P. & Haasl, R. J. A genomic portrait of human microsatellite variation. *Mol. Biol. Evol.* **28**, 303–312 (2011).

47. Ananda, G. et al. Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. *Genome Biol. Evol.* **5**, 606–620 (2013).

48. Ananda, G. et al. Microsatellite interruptions stabilize primate genomes and exist as population-specific single nucleotide polymorphisms within individual human genomes. *PLoS Genet.* **10**, e1004498 (2014).

49. Kristmundsdottir, S. et al. Sequence variants affecting the genome-wide rate of germline microsatellite mutations. *Nat. Commun.* **14**, 3855 (2023).

50. Rolfsmeier, M. L. & Lahue, R. S. Stabilizing effects of interruptions on trinucleotide repeat expansions in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **20**, 173–180 (2000).

51. Kelkar, Y. D., Eckert, K. A., Chiaromonte, F. & Makova, K. D. A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Res.* **21**, 2038–2048 (2011).

52. Rajan-Babu, I.-S., Dolzhenko, E., Eberle, M. A. & Friedman, J. M. Sequence composition changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical implications. *Nat. Rev. Genet.* **25**, 476–499 (2024).

53. Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603 (1980).

54. Orgel, L. E. & Crick, F. H. Selfish DNA: the ultimate parasite. *Nature* **284**, 604–607 (1980).

55. Werren, J. H., Nur, U. & Wu, C. I. Selfish genetic elements. *Trends Ecol. Evol.* **3**, 297–302 (1988).

56. Kelkar, Y. D. et al. What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biol.* Evol. **2**, 620–635 (2010).

57. Sinden, R. R. & Wells, R. D. DNA structure, mutations, and human genetic disease. *Curr. Opin. Biotechnol.* **3**, 612–622 (1992).

58. Strand, M., Prolla, T. A., Liskay, R. M. & Petes, T. D. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365**, 274–276 (1993).

59. McMurray, C. T. Mechanisms of DNA expansion. *Chromosoma* **104**, 2–13 (1995).

60. Pearson, C. E. & Sinden, R. R. Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. *Curr. Opin. Struct. Biol.* **8**, 321–330 (1998).

61. Iyer, R. R. & Pluciennik, A. DNA Mismatch Repair and its Role in Huntington's Disease. *J. Huntingtons Dis.* **10**, 75–94 (2021).

62. Baptiste, B. A. et al. Mature microsatellites: mechanisms underlying dinucleotide microsatellite mutational biases in human cells. *G3* **3**, 451–463 (2013).

63. Putnam, C. D. Strand discrimination in DNA mismatch repair. *DNA Repair* **105**, 103161 (2021).

64. Murat, P., Guilbaud, G. & Sale, J. E. DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. *Genome Biol.* **21**, 209 (2020).

65. Habraken, Y., Sung, P., Prakash, L. & Prakash, S. Binding of insertion/deletion DNA mismatches by the heterodimer of yeast mismatch repair proteins MSH2 and MSH3. *Curr. Biol.* **6**, 1185–1187 (1996).

66. Sia, E. A., Kokoska, R. J., Dominska, M., Greenwell, P. & Petes, T. D. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol. Cell. Biol.* **17**, 2851–2858 (1997).

67. Gendrel, C. G. & Dutreix, M. (CA/TG) microsatellite sequences escape the inhibition of recombination by mismatch repair in Saccharomyces cerevisiae. *Genetics* **159**, 1539–1545 (2001).

68. Jensen, L. E., Jauert, P. A. & Kirkpatrick, D. T. The large loop repair and mismatch repair pathways of Saccharomyces cerevisiae act on distinct substrates during meiosis. *Genetics* **170**, 1033–1043 (2005).

69. Surtees, J. A. & Alani, E. Mismatch repair factor MSH2-MSH3 binds and alters the conformation of branched DNA structures predicted to form during genetic recombination. *J. Mol. Biol.* **360**, 523–536 (2006).

70. Panigrahi, G. B., Slean, M. M., Simard, J. P., Gileadi, O. & Pearson, C. E. Isolated short CTG/CAG DNA slip-outs are repaired efficiently by hMutSbeta, but clustered slip-outs are poorly repaired. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12593–12598 (2010).

71. Omer, S., Lavi, B., Mieczkowski, P. A., Covo, S. & Hazkani-Covo, E. Whole Genome Sequence Analysis of Mutations Accumulated in Yeast Strains with Defects in the Processing of Okazaki Fragments Indicates Template-Switching Events. *G3* **7**, 3775–3787 (2017).

72. Tsutakawa, S. E. et al. Phosphate steering by Flap Endonuclease 1 promotes 5'-flap specificity and incision to prevent genome instability. *Nat. Commun.* **8**, 15855 (2017).

73. Bae, S. H., Bae, K. H., Kim, J. A. & Seo, Y. S. RPA governs endonuclease switching during processing of Okazaki fragments in eukaryotes. *Nature* **412**, 456–461 (2001).

74. Deshmukh, A. L. et al. FAN1, a DNA Repair Nuclease, as a Modifier of Repeat Expansion Disorders. *J. Huntingtons Dis.* **10**, 95–122 (2021).

75. Goldmann, J. M. et al. Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48**, 935–939 (2016).

76. Yuen, R. K. C. et al. Genome-wide characteristics of mutations in autism. *NPJ Genome Med.* **1**, 160271–1602710 (2016).

77. Jónsson, H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).

78. An, J.-Y. et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, (2018).

79. Halldorsson, B. V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, (2019).

80. Sasani, T. A. et al. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife* **8**, (2019).

81. Jonsson, H. et al. Differences between germline genomes of monozygotic twins. *Nat. Genet.* **53**,

27–34 (2021).

82. Goes, F. S. et al. De novo variation in bipolar disorder. *Mol. Psychiatry* **26**, 4127–4136 (2021).

83. McGinty, R. J. & Sunyaev, S. R. Revisiting mutagenesis at non-B DNA motifs in the human genome. *Nat. Struct. Mol. Biol.* **30**, 417–424 (2023).

84. Wolfram Research, Inc., *Mathematica, Version 14.0*, Champaign, IL (2024).

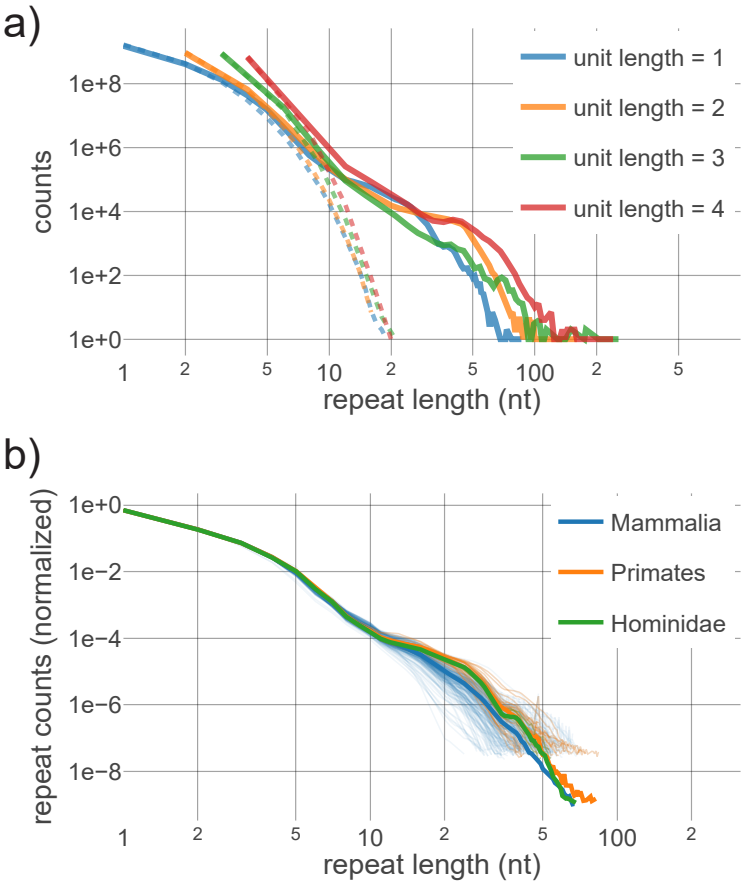


Fig. 1. *Repeat length distributions by motif length and across phylogenies.* **a)** Counts of repeats in human T2T genome pooled by motif unit length (e.g., unit 1 pools distributions for A/T and C/G). Dashed lines represent exponentially-distributed counts in a randomly-shuffled human genome sequence. **b)** Normalized distributions of A_n repeats in mammals, primates and hominids. Solid line indicates median values per length bin. Thin transparent lines show individual species within the phylogeny. To appropriately compare assemblies with different genome lengths (after normalization), individual distributions are cut off at the shortest bin containing 30 counts; median calculated without a cutoff. Phylogenies are inclusive (e.g., primates are included as a subset of mammals). Similarity within phylogenies suggests long-term stability of the distribution.

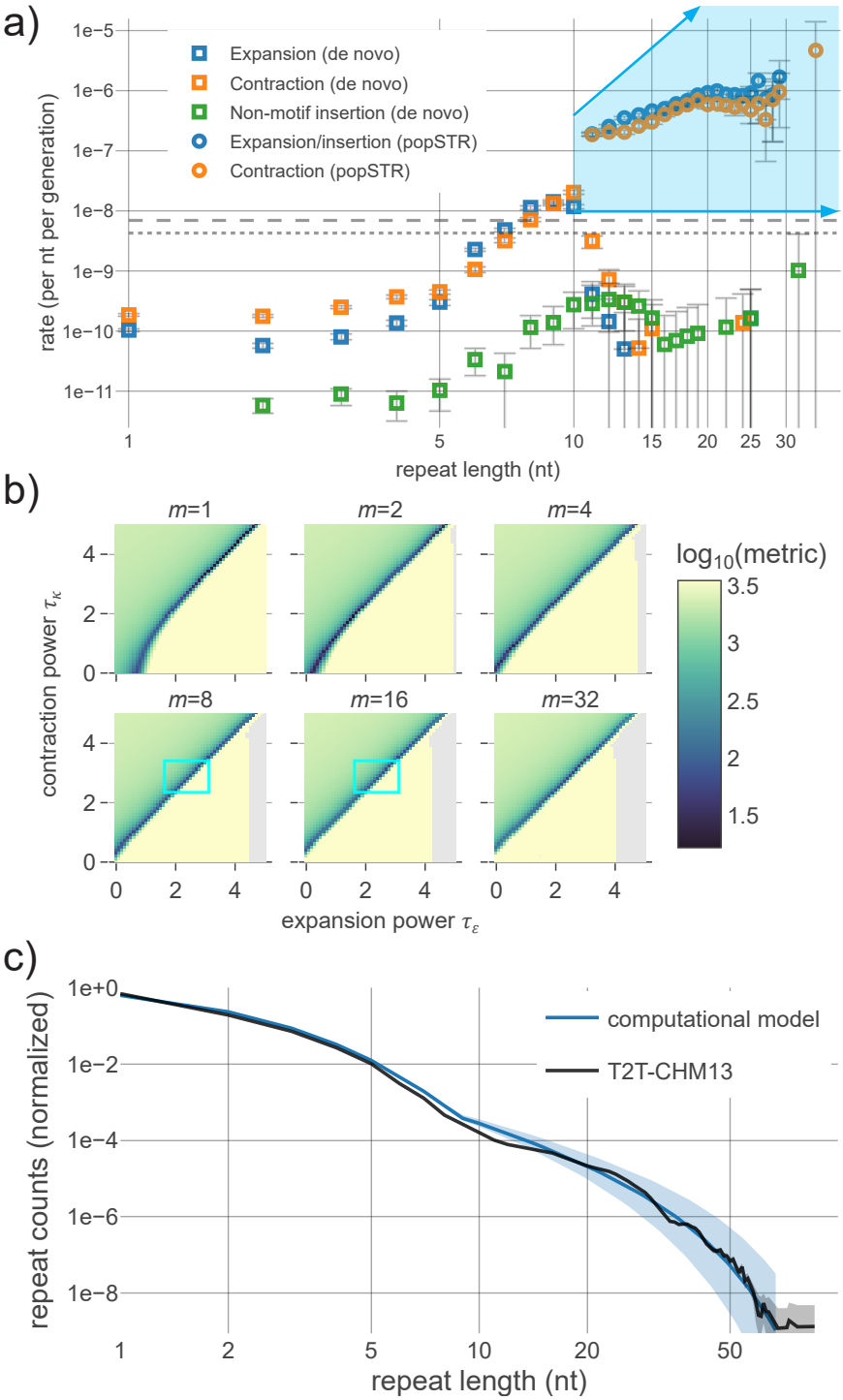
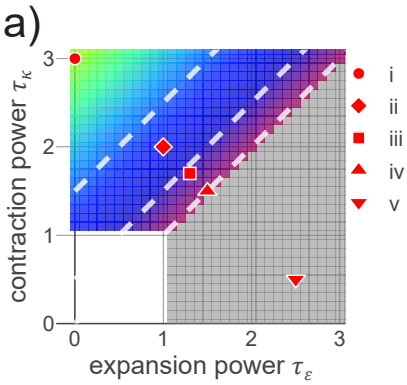
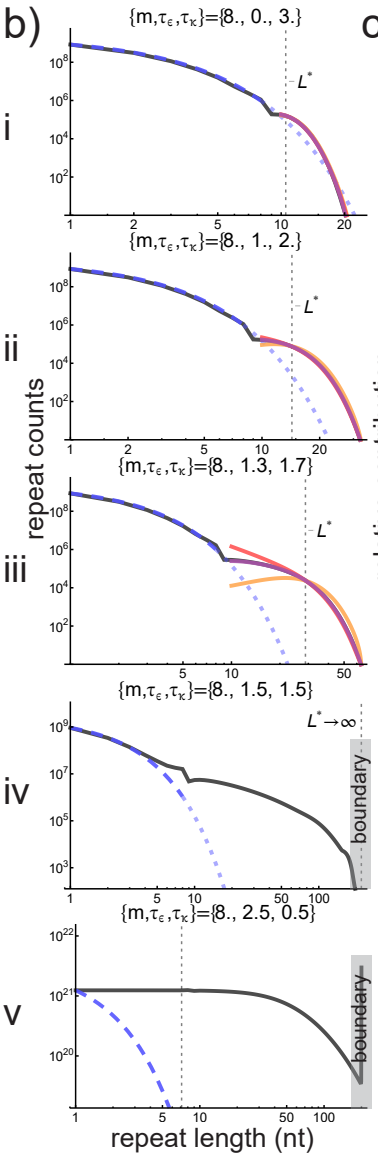


Fig. 2. *Estimated instability rates, parameterization, and consistency with maintenance of the repeat length distribution.* **a)** Separate rate estimates from de novo and popSTR datasets for expansion, contraction and non-motif insertions for A_n repeats. Statistical error bars show 95% confidence intervals. Gray dashed and dotted lines show average substitution rates for A>B and B>A (B=C,G or T), respectively. Light blue box illustrates the approximate parameter space explored by computational model; arrows illustrate power law functions (rate $\propto L^{\tau}$). **b)** Metric values comparing computationally-propagated and empirical distributions. Computational model runs across a range of parameters: multiplier $m=[2^0,2^5]$, expansion and contraction power law exponents $\tau_\varepsilon,\tau_\kappa=[0,5]$. Plane of $\tau_\varepsilon,\tau_\kappa$ is shown for slices of constant m. Color specifies $\log_{10}$ of metric values. Gray masks parameter combinations exceeding the linear mutation bound. All runs were initialized using the equilibrium distribution under substitutions alone and were propagated until reaching an approximate steady-state, if applicable. Blue boxes show 99.9% confidence intervals estimated from popSTR data via linear regression. **c)** Comparison of normalized empirical distribution and computational model of best-fit parameters (m=2, τ_ε=1.5, τ_κ=1.8). Blue transparency shows 95% confidence intervals generated by statistical errors around estimated de novo mutation rates. Gray shading indicates 95% bootstrap confidence intervals on empirical distribution.





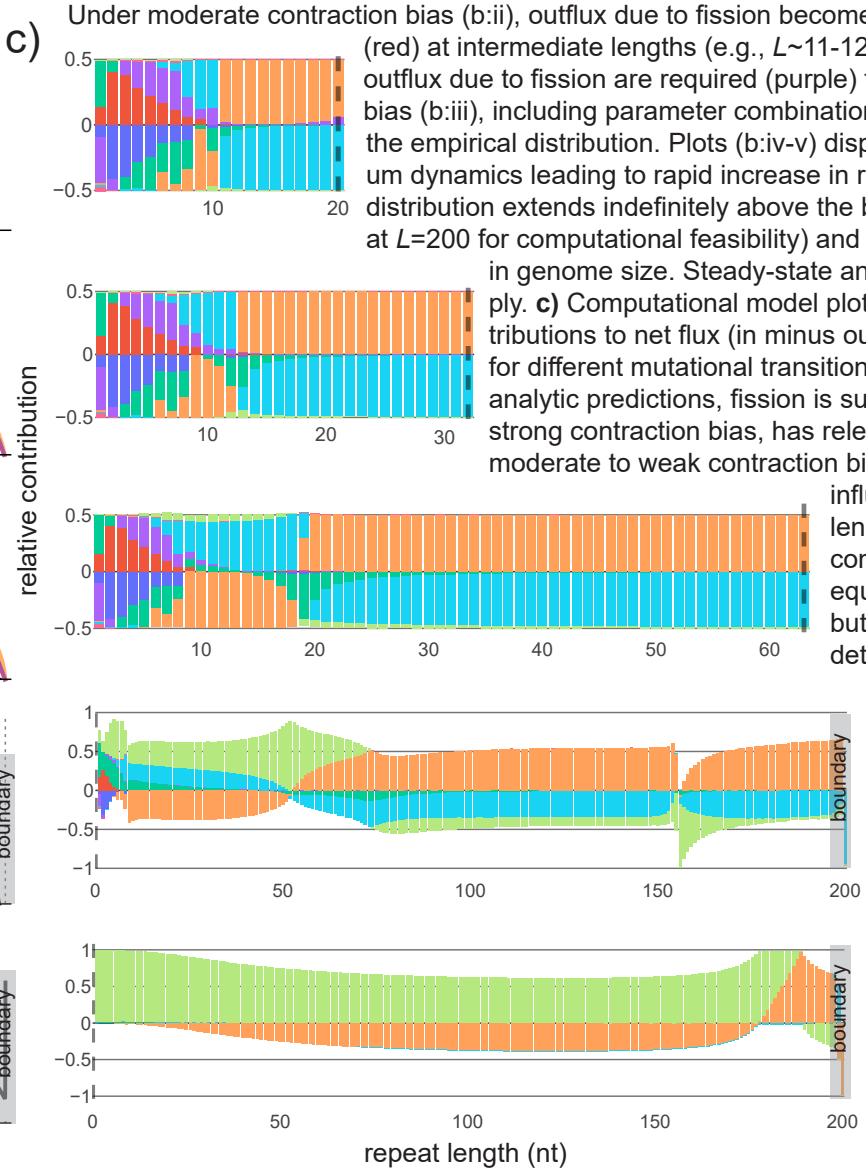


Fig. 3. *Dynamical regimes distinguishable by dominant mutational effects.* **a)** Five parameter combinations with $\tau_\varepsilon+\tau_\kappa=3$ in distinct dynamical regimes, corresponding to plots in (b) and (c). Dotted lines divide the parameter space into approximate dynamical regimes. **b)** Comparisons between computational model results, analytic approximations and numerical solutions of approximate steady state equations (see **Methods**, **Supplementary Note SN1**). 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 neglibile rates). Local transitions (L to L+/-1) were modeled as a combination of symmetric (diffusive) and asymmetric (directional) components. Under strong contraction-bias (b:i), all three approximations remain valid indicating that the dynamics are well approximated by neglecting fission entirely (orange). Under moderate contraction bias (b:ii), outflux due to fission becomes non-negligible (red) at intermediate lengths (e.g., L~11-12). Both influx and outflux due to fission are required (purple) for low contraction bias (b:iii), including parameter combinations consistent with the empirical distribution. Plots (b:iv-v) display non-equilibrium dynamics leading to rapid increase in repeat counts (true distribution extends indefinitely above the boundary imposed at L=200 for computational feasibility) and explosive growth in genome size. Steady-state analytics do not apply. **c)** Computational model plots of relative contributions to net flux (in minus out) per length bin for different mutational transitions. Consistent with analytic predictions, fission is subdominant under strong contraction bias, has relevant outflux under moderate to weak contraction bias, and relavant influx at intermediate lengths under weak contraction bias. In equilibrium, the distribution is stabilized in detailed balance (net influx = outflux). Nonequilibrium dynamics (c:iv-v) are generated by fluxes that do not sum to zero, leading to indefinite genome growth; distribution extends to length boundary of computational model.

- computational model
- short repeat geometric model
- 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
- substitution +1
- substitution fission
- substitution fusion
- expansion +1
- contraction -1
- deletion fusion
- non-motif insertion fission
- count <1

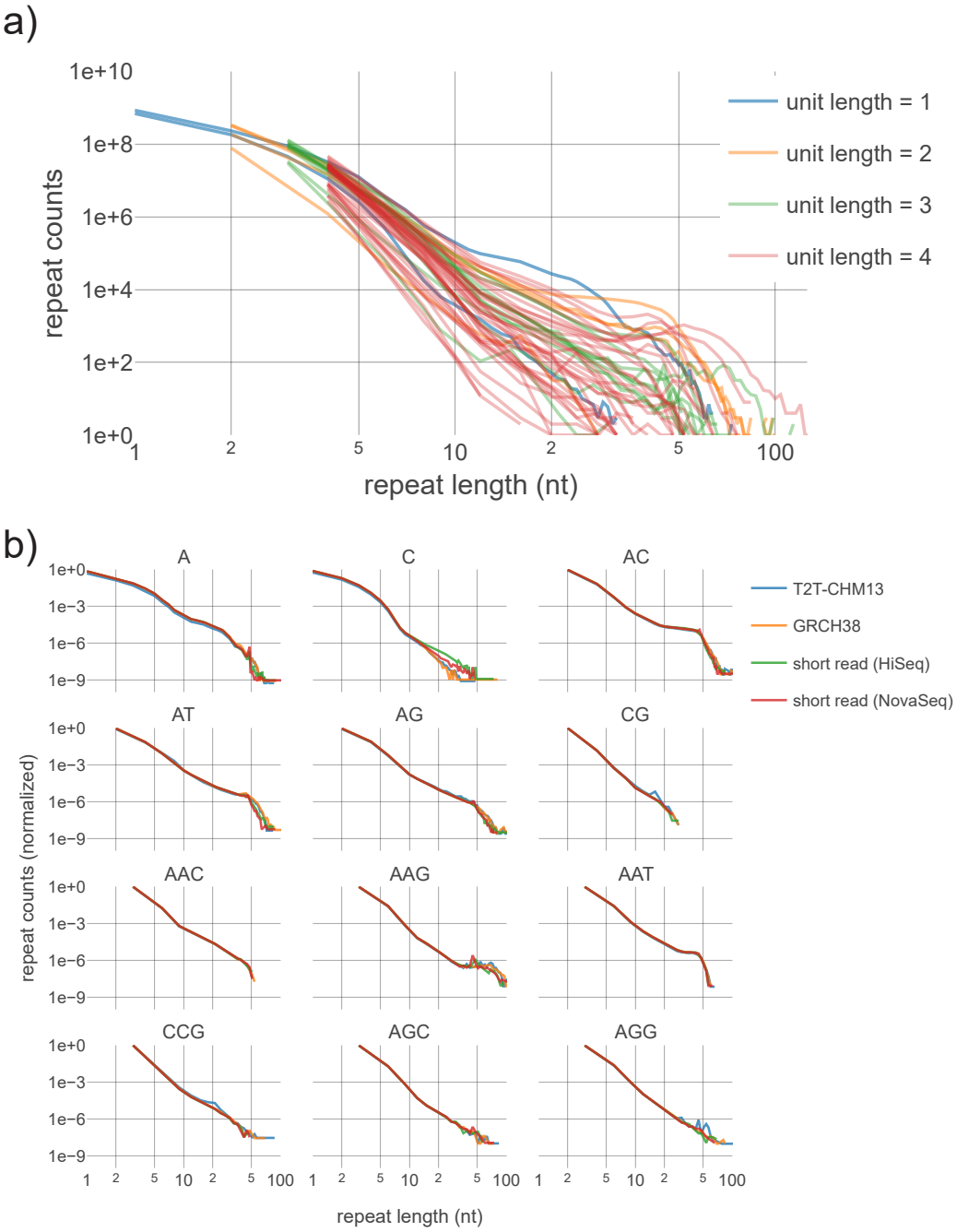


Fig. S1. *Repeat length distributions per motif and per sequencing technology.* **a)** Non-normalized distribution of repeat lengths for all distinct motifs for unit lengths 1-4. Motifs of the same unit length are shown in the same color. Distributions vary by motif but share similar qualitative features. **b)** Normalized distribution of repeat lengths in four distinct human genome assemblies with different sequencing technologies. T2T-CHM13 was assembled using multiple long read technologies, while other assembles employed shorter read technologies. Read length does not appear to affect the accurate counting of repeats of sub-disease lengths.

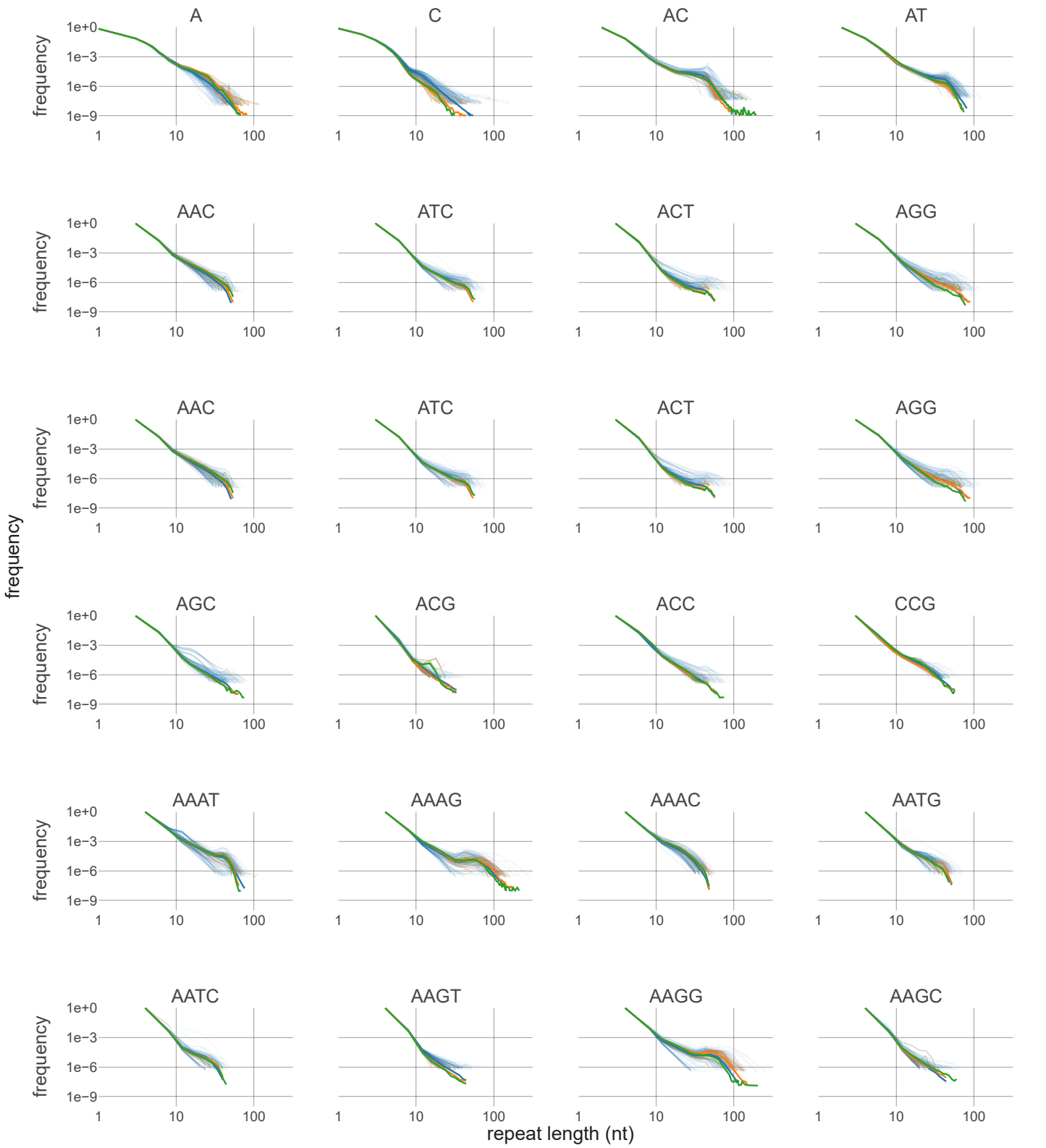
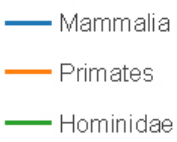


Fig. S2. *Repeat length distributions per motif across phylogenies.* Normalized distributions of repeat lengths for various motifs in mammals, primates and hominids. Solid line indicates median values per length bin for each phylogeny. Thin transparent lines show individual species within the phylogeny. Individual distributions are cut off at the shortest bin containing 30 counts. Phylogenies are inclusive (e.g., primates are included as a subset of mammals). Overlapping medians between primates and hominids suggests long-term stability of the distribution, while individual mammalian species display variability.



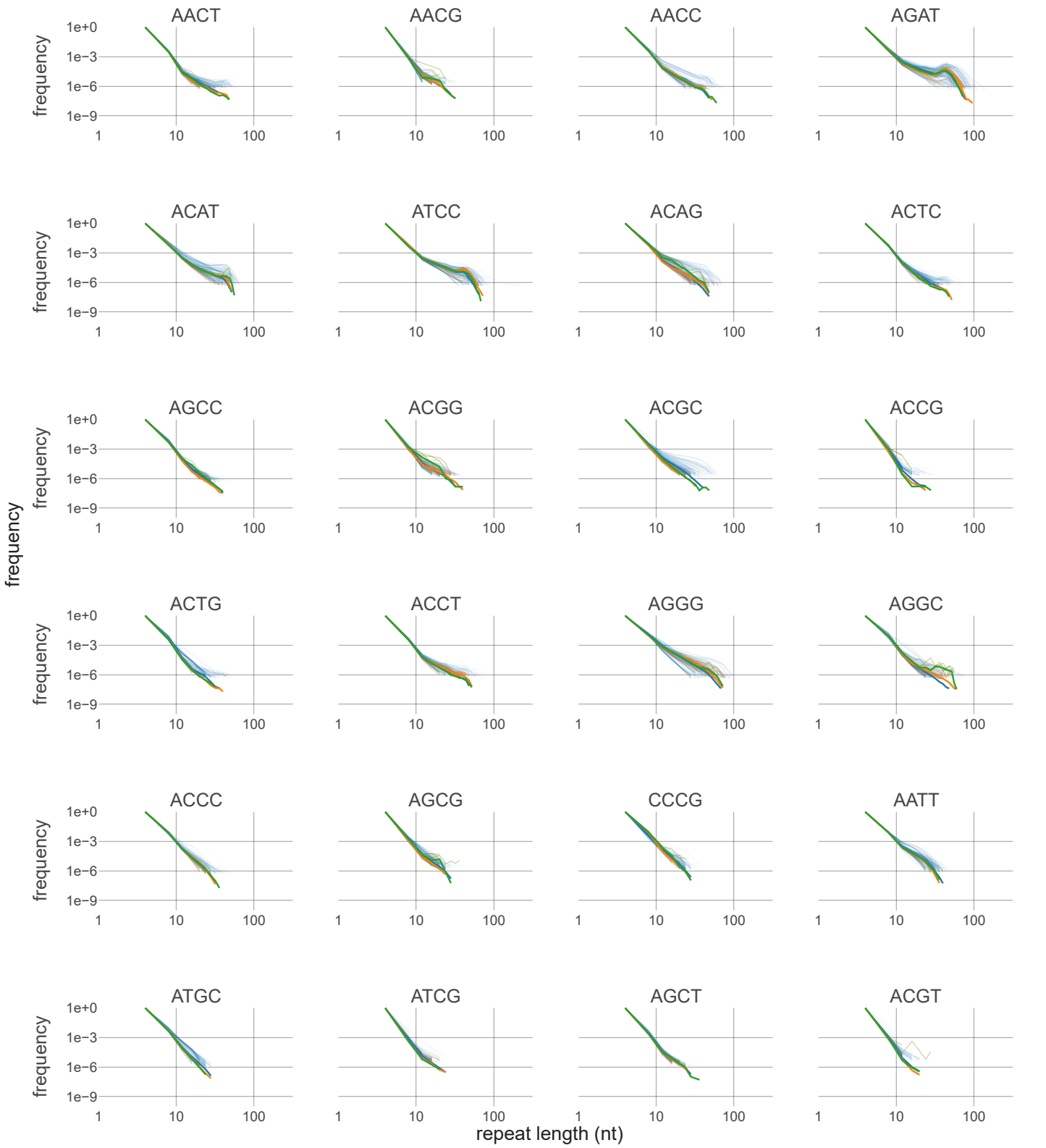


Fig. S2 (continued). *Repeat length distributions per motif across phylogenies.*

Mammalia
Primates
Hominidae

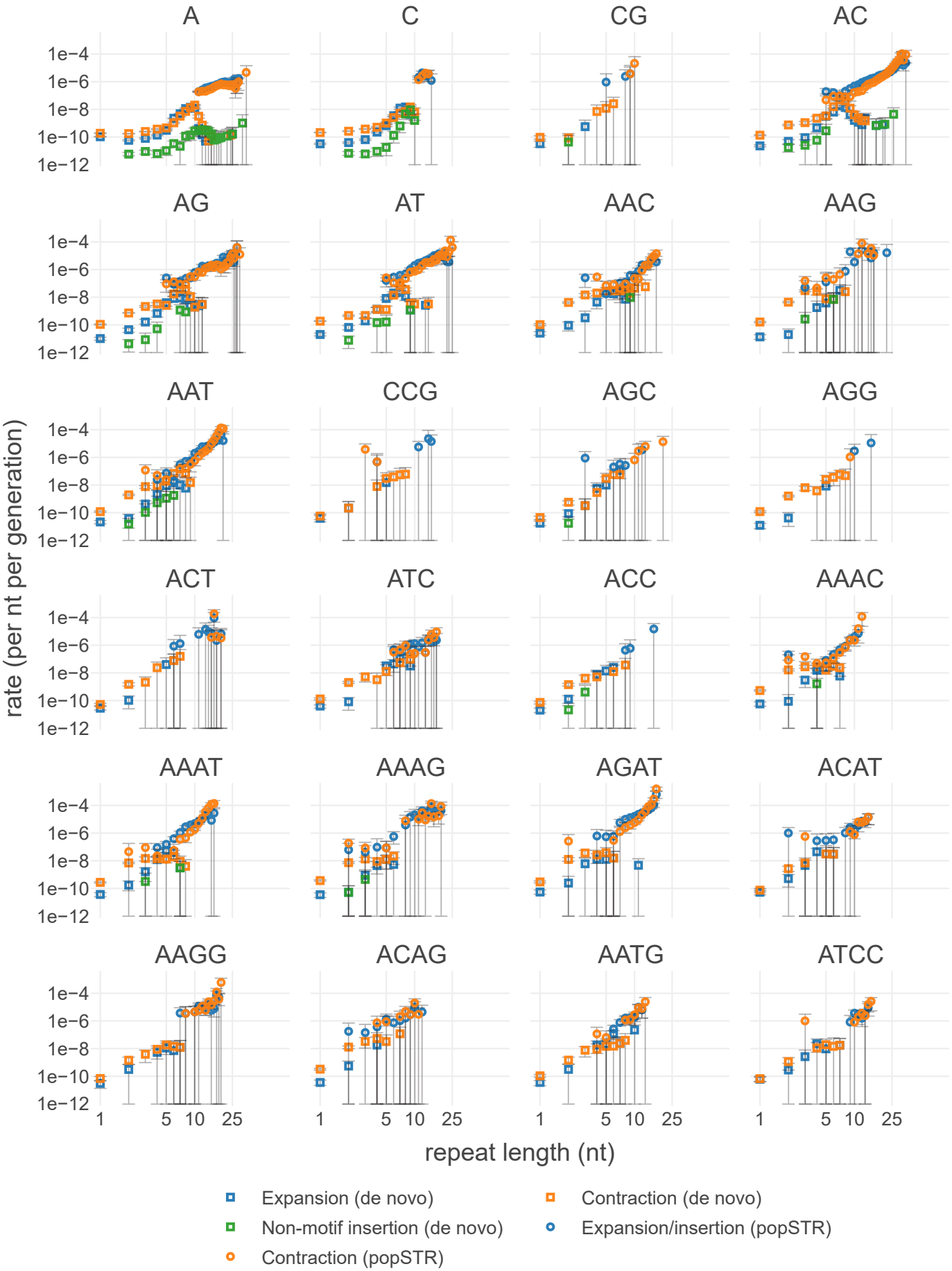


Fig. S3. *Instability rate estimates*. Separate rate estimates from de novo and popSTR datasets for expansion, contraction and non-motif insertions for various repeat motifs. Statistical error bars show 95% confidence intervals.

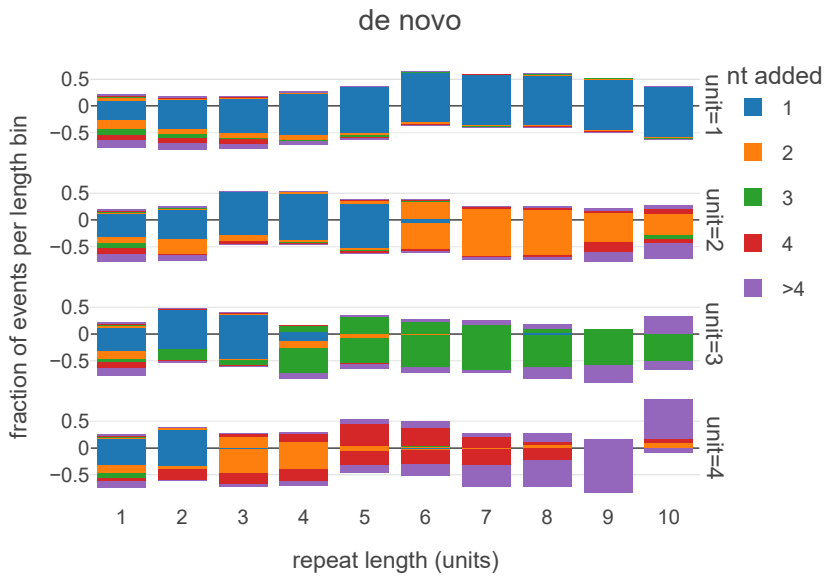
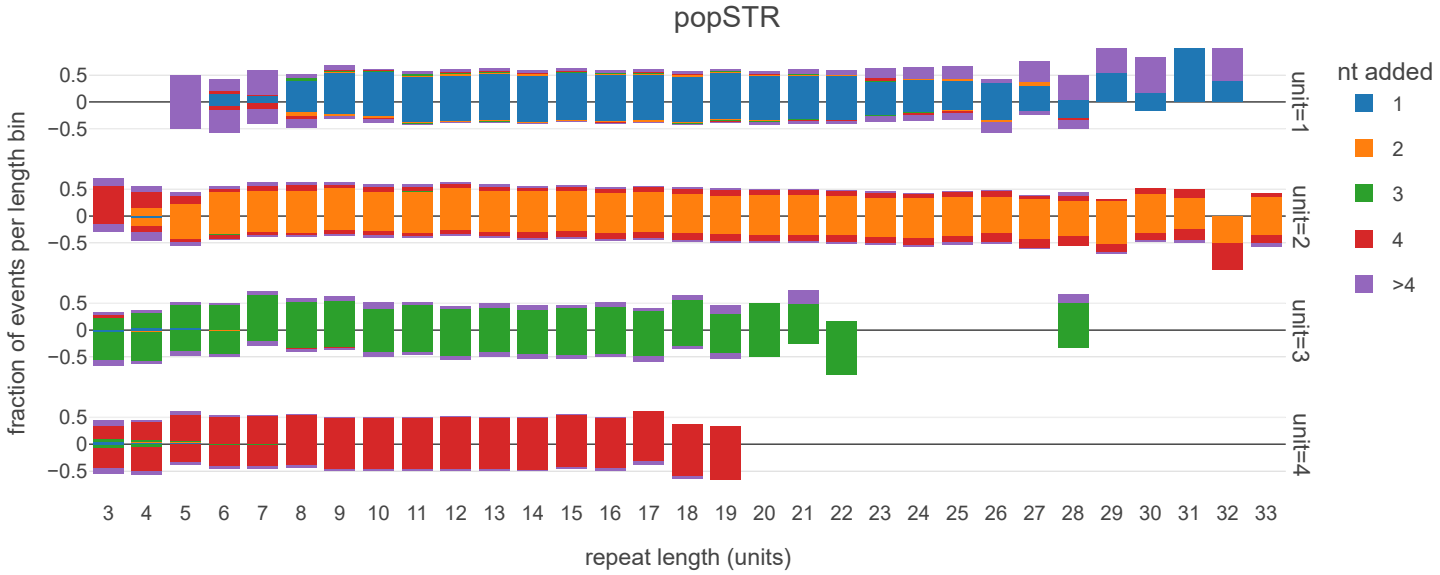


Fig. S4. *Nucleotides gained or lost per indel event.* Indels measured from two different data sources: de novo indels from trio sequencing (**top**) and popSTR estimates **(bottom**). Row labels indicate the motif unit length. Color indicates how many nucleotides were gained or lost per indel event. X-axis displays the length of the repeat in number of units. Y-axis displays the fraction of events per length bin. Y-axis sign indicates insertion (positive) or deletion (negative); overall shift up or down indicates bias. Plots show that most events involve the gain or loss of a single complete unit, independent of total repeat length; both partial-unit and multiple-unit changes are less frequent. A length threshold is visible, above which repeat instability exceeds the background indel process. Note that accurate calling of indels in the de novo dataset is reduced for STRs above 10 nt; likewise, the popSTR database is not sufficiently populated below 10 nt and above 30 nt.



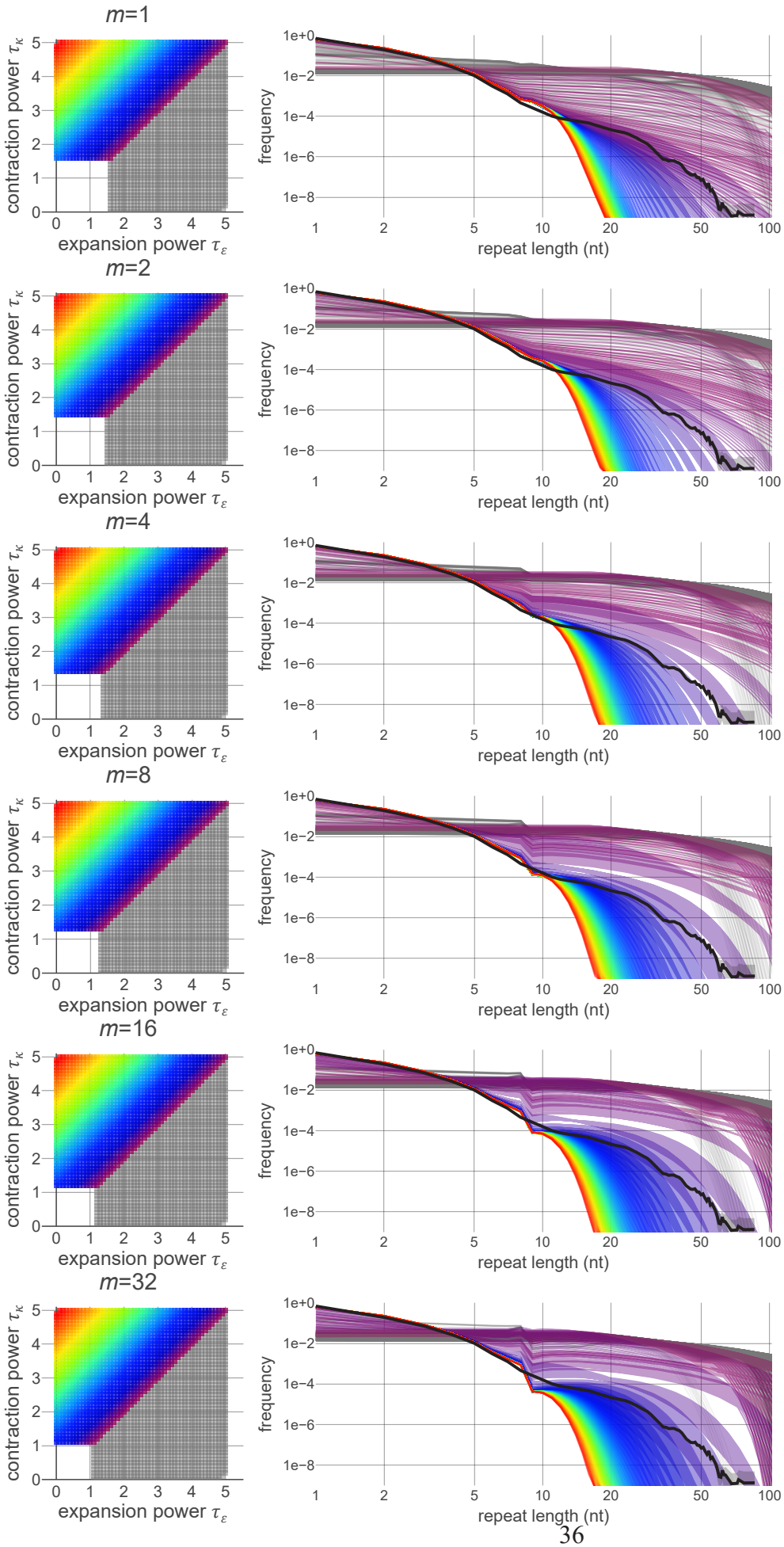


Fig. S5. *Final time point distributions from computational model across parameter space.* Lines of constant $\Delta\tau=\tau_\kappa-\tau_\varepsilon$ are shown in the same color. Red corresponds to large values of $\Delta\tau$, purple corresponds to low values of $\Delta\tau$, and gray represents negative values of $\Delta\tau$. **(Left)** Grid of $(\tau_\varepsilon,\tau_\kappa)$ parameter values for various multipliers *m*. For clarity, lower left region is not shown due to low rates insufficient to equilibrate in the alotted time. **(Right)** Plots of A_n repeat length distributions at the final time point, one for each point in the grid. Black line depicts the empirical distribution. Larger $\Delta\tau$ results in more rapid truncation of the distribution at lower lengths; smaller values of $\Delta\tau$ result in a more extended tail of long repeats.

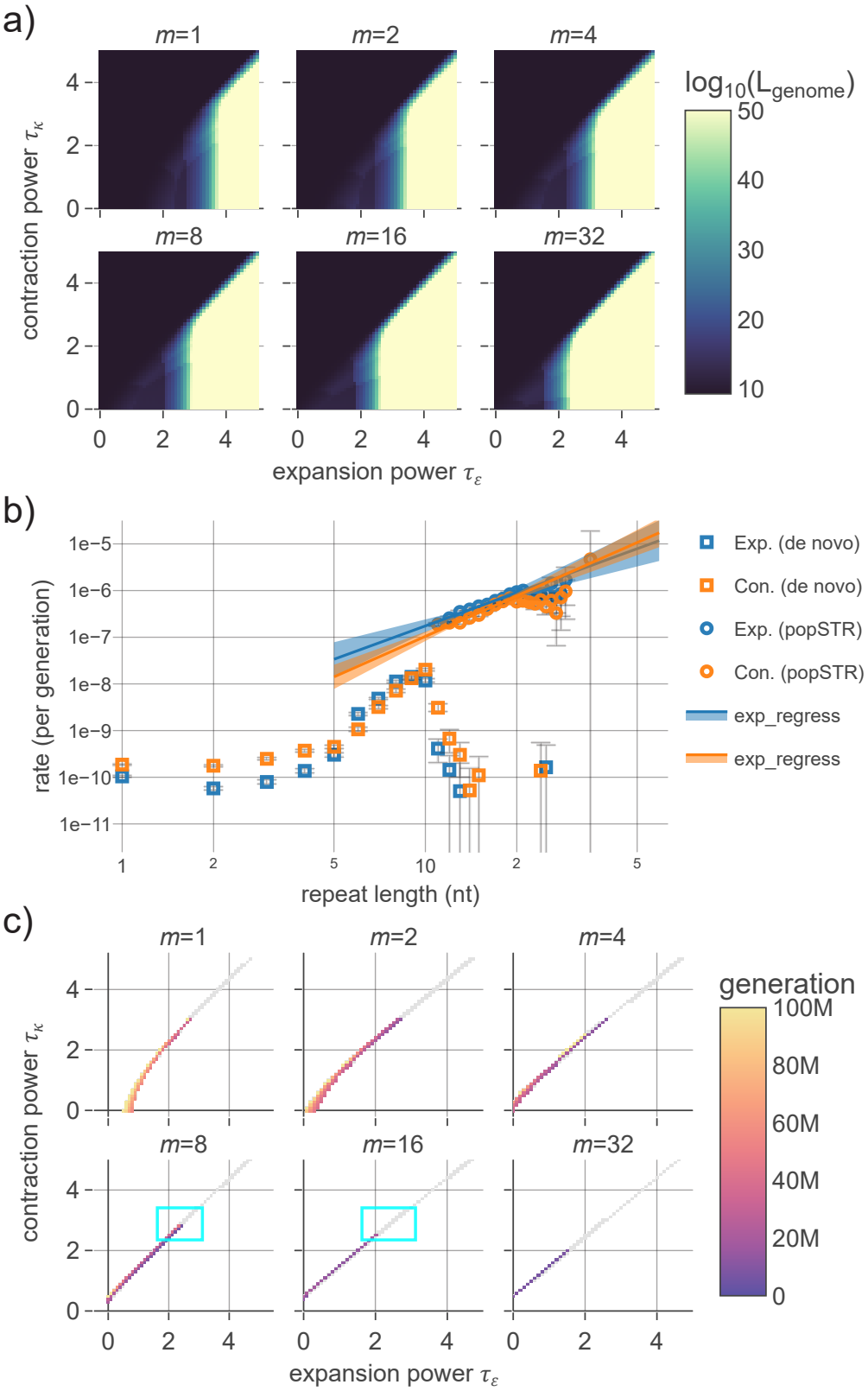


Fig. S6. *Constraints on realistic parameter values.* **a)** Plot of total number of repeat bases in genome at final time point across parameter space. Color specifies $\log_{10}$ of genome-wide count of A bases. Color range is truncated at 10^{50}. Computational model results in explosive genome growth for expansion-biased parameter combinations. **b)** Rate estimates from de novo and popSTR datasets for expansion, contraction and non-motif insertions for A_n repeats. Lines represent linear regression in log-log space of the popSTR expansion and contraction rate estimates for L=12-19. Statistical error bars show 99.9% confidence intervals, which suggest empirical bounds on parameters m, τ_ε and τ_κ. **c)** Plot of timepoint when each parameter combination becomes consistent with the best fit parameters. Only parameter values consistent at the final time point are shown. Here, a constant time rescaling was used, rather than progressive time rescaling. Blue boxes show 95% confidence intervals resulting from (b). The small subset of parameter combinations within these boxes is simultaneously consistent with de novo and popSTR rates estimates of repeat instability, the empirical distribution of repeat lengths in humans, and the convergence to steady state within realistic timescales. Grey points are consistent with the best-fit parameter combination but required progressive time-rescaling due to computational limitations, precluding generation time measurement.

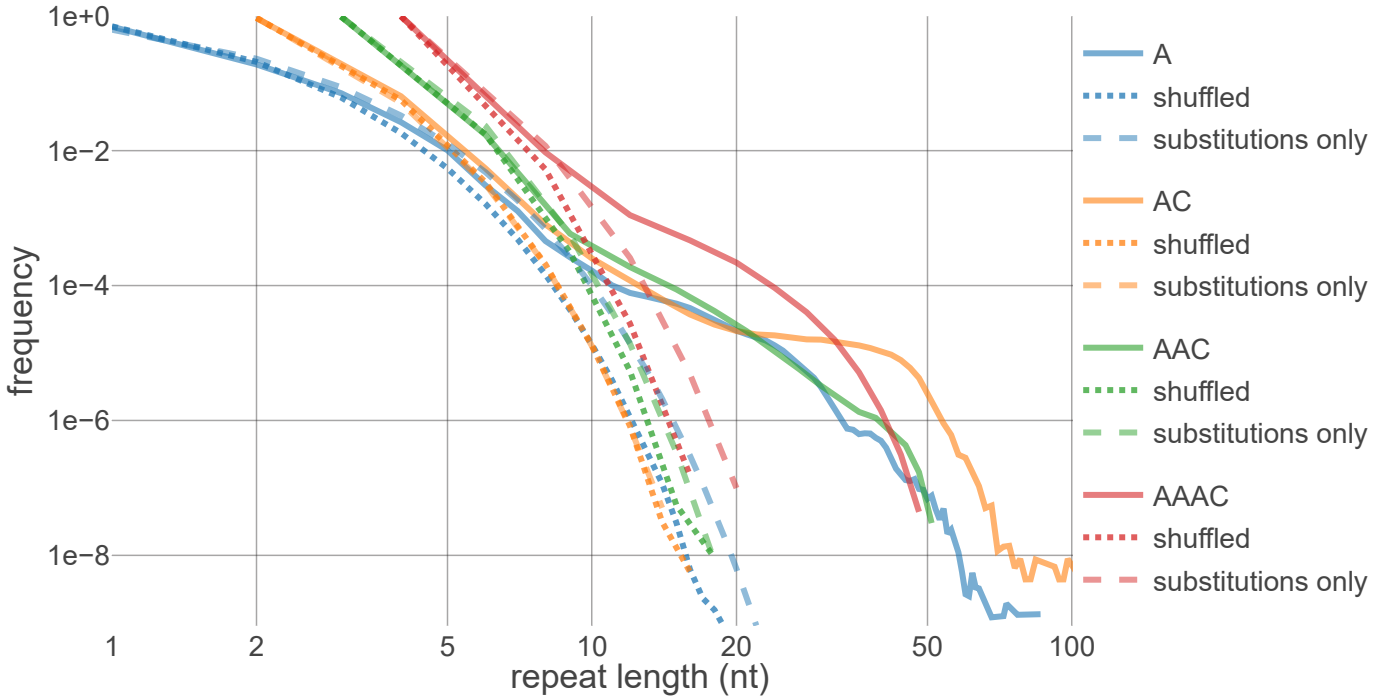


Fig. S7. *Assessing the onset of repeat instability for longer motifs.* Solid lines display normalized repeat length distributions in CHM13-T2T for several example motifs, highlighting the difference between mono-, di-, tri- and tetranucleotide motifs. Dotted lines represent counts in a randomly-shuffled human genome sequence. Dashed lines represent a computational model of the substitution process alone (omitting all indels). This results in a geometric distribution, which describes the low-length (i.e., <10 nt) portion of each empirical distribution. Empirical deviation from this distribution results from the relevance of repeat instability at longer lengths, roughly above 10 nt, independent of motif length.

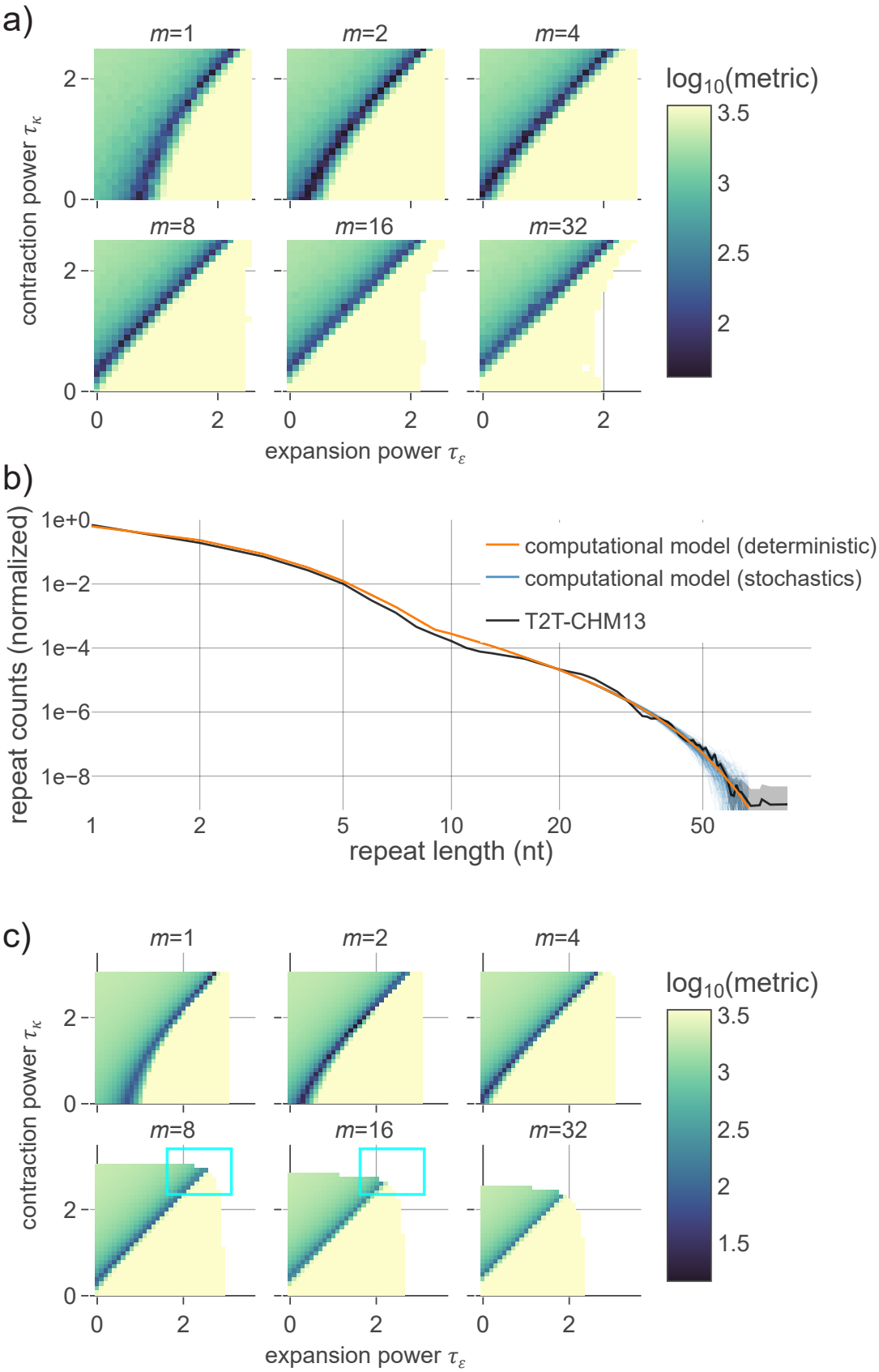


Fig. S8. *Tests of simplifying assumptions used in the computational model.* **a)** Metric plots for a computational model that incorporates stochastic fluctations. In each generation, the number of mutational transitions between length classes was Poisson sampled. Blank spaces indicate parameter combinations resulting in numerical errors. Results are nearly identical to the deterministic model results in Fig. 2b, indicating stochastic effects are largely unnecessary to identify parameters consistent with the empirical distribution. **b)** Comparison of normalized empirical distribution and computationally modeled distribution, with and without stochastics, for the best-fit parameters (m=2, τ_ε=1.5, τ_κ=1.8). Blue lines represent 200 individual runs with stochastic fluctuations, as described in (a). The distribution is not substantially altered. **c)** Metric values comparing empirical and computationally modeled distributions run for 10^9 generations without progressive time rescaling. Blank spaces indicate parameter combinations that were not run due to computational limitations. Results are nearly identical to the progressive time rescaling shown in Fig. 2b, indicating the validity of the progressive rescaling procedure.

Supplementary Note SN1

Contents

1 Analytic modeling of the repeat length distribution in steady state *42*
1.1 Summary of parameterized mutation rates at long lengths . . . *44*

2 Finite difference equation for repeat length changes *45*
2.1 Changes in repeat length due to expansion, contraction, and insertion . . . *45*
2.2 Changes in repeat length due to substitutions . . . *46*
2.3 Finite difference equation . . . *47*

3 Steady-state dynamics in the short repeat length regime *47*
3.1 Geometric solution to the substitution-only difference equation . . . *48*
3.2 Interactions between short and long repeats are largely restricted to boundary effects . . . *49*

4 Steady-state dynamics for asymptotically long repeat lengths *50*
4.1 Local contributions to the change in P_L . . . *50*
4.1.1 Competition between first-order local effects . . . *52*
4.1.2 Second-order corrections to the local behavior and diffusive dynamics . . . *53*
4.2 Repeat fission as a nonlocal contribution to changes in length . . . *54*
4.2.1 Substitution-based fission . . . *54*
4.2.2 Insertion-based fission . . . *55*
4.3 Repeat fusion under random sampling of the length distribution . . . *55*
4.4 Steady-state condition for long repeat dynamics . . . *56*
4.5 Decomposition of parameter space into dynamical regimes . . . *58*
4.5.1 Asymptotic length dependence of local transition rates and $\Delta\tau$ . . . *58*
4.5.2 Relative strengths of substitution- and insertion-driven fission and L_{fis} . . . *59*
4.5.3 Distinguishable dynamical regimes . . . *60*
4.6 Unstable dynamics in the asymptotically expansion-biased regime $\Delta\tau \leq 0$ . . . *61*
4.7 Stable dynamics in the asymptotically contraction-biased regime $\Delta\tau > 0$ . . . *62*
4.7.1 Strong asymptotic contraction bias . . . *62*
4.7.2 Strictly local approximation for strong asymptotic contraction bias $\Delta\tau \gg 1$ . . . *63*
4.8 Intermediate asymptotic contraction bias . . . *63*
4.9 Weak asymptotic contraction bias . . . *64*
4.10 Obtaining numerical solutions to the steady state dynamics for $\Delta\tau < 0$ . . . *65*

5 Comparison between numerical solutions and computationally modeled distributions *66*
5.0.1 Comparisons of approximations to the dynamics and steady-state distributions . . . *66*

6 Empirical constraints on parametric model *68*
6.1 Constraints on $\Delta\tau$ and L^* . . . *68*
6.2 Constraints on L_{fis} . . . *69*

Supplementary Note Figures

Figure SN1 *Comparison between computational model results and numerical solutions to steady state equations for $m = 8$ for parameter values with of constant $\tau_\epsilon + \tau_\kappa = 3$.* 70
Figure SN2 *Computational model results for the net flux per mutation type.* 71
Figure SN3 *Computational model results for directional flux per mutation type* 72
Figure SN4 *Computational model results for collective non-normalized fluxes showing the relevance of local transitions, fission, and fusion.* 73
Figure SN5 *Comparison between computational model results and numerical solutions to steady state equations for $m = 2$.* 74
Figure SN6 *Comparison between computational model results and numerical solutions to steady state equations for $m = 16$.* 75
Figure SN7 *Comparison between computational model results and numerical solutions to steady state equations for $m = 32$.* 76
Figure SN8 *Accuracy of analytic approximation for distribution falloff in the $\Delta\tau \gg 1$ regime.* . . . 77

1 Analytic modeling of the repeat length distribution in steady state

To better understand our observations, we sought to describe the equilibrium that emerges once the repeat length distribution (herein referred to as P_L in discrete form and $\rho(L)$ under a continuum approximation) has reached steady state at asymptotically late times (i.e., $P(L;t) \to P_{ss}(L)$ as $t \to \infty$, where the subscript ss denotes steady state). Importantly, this does not occur for all parameter combinations, as a subset remain inherently unstable. For the parameter values that stabilize, the shape of the distribution is described by a dynamic balance that emerges between mutational processes, with distinct mutational effects dominating the dynamics in different length regimes. For simplicity, the analysis provided here is restricted to the case of mononucleotide repeats consisting of various numbers of A bases. Alternative bases that terminate each sequence, labeled B, represent any T, C, or G base (i.e., any non-motif base for A repeats) at either end of an A repeat and the explicit distribution of lengths of B strings are ignored for the present purposes.

As described in the manuscript, the mutational processes that are included in our model are the following, explicitly enumerated here with corresponding variables used to represent and parameterize the rates of each mutation type for clarity.

1. *Lengthening substitutions* μ: Point mutations $B \to A$ that increase the length of a repeat while preserving the total length of the genome. Such mutations only increase repeat length when they occur on bases adjacent to an existing repeat. Note that two additional transitions result from such mutations: fusion of nearby repeats (merging of two shorter repeats to form a longer repeat) and generation of new 'repeats' of length L=1 (included in our description for completeness, despite the lack of repetition). The distinction between these processes can be categorized by the number of B bases adjacent to the mutated site, one, zero, and two, for lengthening, fusion, and generation, respectively. The per-target rate mu is assumed to be the same constant for any repeat length L such that the per-repeat rate of these substitutions is necessarily linear in length, scaling as $\mu \times L$ for all repeat lengths.

2. *Shortening substitutions* ν: Point mutations $A \to B$ that decrease repeat length while preserving genome length. Such mutations also result in repeat fission (interruption of a repeat that forms two smaller repeats) and repeat destruction (mutation of L=1 repeats, removing them from the distribution). Again, these transitions can be categorized by adjacency of the mutated base to one, zero, and two B bases for repeat shortening, fission, and destruction, respectively. The per-repeat rate of shortening substitutions is again linear in repeat length: $\nu \times L$.

3. *Contractions* κ_L: Deletions of a single A base that decrease repeat length and genome length by one unit (i.e., one nucleotide for mononucleotide repeats). Extended deletions of two or more bases are assumed to be subdominant for simplicity (consistent with estimates provided in **Supplementary Figure S3**). Repeat contractions occur at a length-dependent rate constrained by empirical estimates and further parameterized by two variables; all rates for lengths $L \leq 8$ are fixed by trio-based rate estimates (see **Methods** and **Figure 2a**), the rate at $L = 9$ is determined by a constant multiple m relative to the value at $L = 8$ (i.e., $\kappa_{L=9} = m \times \kappa_{L=8}$), and rates for $L > 9$ follow a power law dependence of the form $\kappa_{L>9} = C_\kappa L^{\tau_\kappa}$. Here, the length-independent constant C_κ sets the initial value of the power law that parameterizes the contraction rate for all $L \geq 9$ (for intuition, if the power law $C_\kappa L^{\tau_\kappa}$ was artificially extended below $L = 9$, C_κ would be the rate associated with $L = 1$: $C_\kappa L^{\tau_\epsilon}|_{L\to 1} = C_\kappa$); the subscript c refers to the crossover in behavior between substitution-dominated mutation rates at lengths $L < 9$ and the onset of repeat instability-dominated rates for $L > 9$. Under our parameterization, solving for $\kappa_9 = C_\kappa L^{\tau_\kappa}|_{L=9} = m\ \kappa_8$ yields the definition $C_\kappa = \kappa_{L=8} \times m/9^{\tau_\kappa}$, which is dependent on both τ_κ and m, but constant for a fixed parameter combination $\{m, \tau_\epsilon, \tau_\kappa\}$. The per-repeat target size for contractions in a repeat of length L is simply L, as the deletion of any base in an A repeat has an equivalent effect. Thus, the per-repeat rates are given by multiplying κ_L by an additional factor of L (i.e., the per-repeat contraction rate becomes $\kappa_L L = C_\kappa L^{1+\tau_\kappa}$ for long repeats $L \geq 9$). For all length-dependent rates (e.g., κ_L), we use the subscript L to denote the length dependence for both a discrete list of rates (e.g, κ_L

defined in the discrete range of positive definite integers $L \in \mathbb{Z}^+$) and in the continuum where these rates are interpreted as continuous functions (e.g., $\kappa_L \to \kappa(L)$ becomes a continuous function describing the contraction rate in the continuous space $L \in \mathbb{R}^+$, but we preserve the notation $\kappa_L \equiv \kappa(L)$ in this case for consistency).

4 *Expansions* ϵ_L: Insertions of a single A base that increase repeat length and genome length by one unit. Again, extended insertions are assumed to occur at subdominant rates (see **Supplementary Figure S3**). Expansions are defined to be dependent on length in the same way as contractions, defining low lengths empirically and including the same value for the parameter m, but with a distinct power law for rates at large lengths: $\epsilon_{L>9} = C_\epsilon L^{\tau_\epsilon}$. We have again expressed this dependence using the length-independent constant $C_\epsilon = \epsilon_8 \times m/9^{\tau_\epsilon}$ for notational convenience. We envision insertion as the replacement of a single A base with a pair of bases AA such that the per-target lengths are again multiplied by the target size L (e.g., asymptotically, $\epsilon_L L = C_\epsilon L^{1+\tau_\epsilon}$).

5 *Non-motif insertions* ι_L: Insertions of a single B base that interrupts a repeat, resulting in repeat fission. For conciseness, non-motif insertions will simply be referred to as *insertions* herein (with the implicit distinction between insertions leading to expansions and non-motif insertions). Extended insertions of more than a single base are again ignored (see **Supplementary Figure S3**). The per-target rate of insertions is again length dependent in the same way, with $\iota_{L\leq 8}$ derived from empirical data, with the parameterization-dependent $\iota_9 = m\iota_8$ defined using the same multiple m, and rates for lengths $L > 9$ defined by the power law $\iota_L = C_\iota L^{\tau_\iota}$. In this case, the per-repeat rate is dependent on a target size of $L - 1$ (B must be inserted between two A bases in the pre-mutated repeat and there are $L - 1$ pairs of adjacent motifs providing targets for $AA \to ABA$). Herein, we will assume that the rate ι_L was estimated by collecting the average per-repeat rate of insertions and dividing by the length of the repeat L (i.e., as an effective per-target rate), as in **Methods**; however, if the per-target rate ι_L is estimated directly (i.e., by separately counting each $AA \to ABA$ event), one may resort to the approximation $L - 1 \approx L$ in the asymptotic large L regime, which is the only relevant length range for insertion-based processes due to their highly subdominant rates (relative to, e.g., the expansion or contraction rates in the same regime). We impose one additional assumption here to limit the number of free parameters in our model: with empirical motivation (see manuscript text), we impose the constraint $\tau_\iota = \tau_\epsilon$ under the assumption that they are both consequences of the same biological mechanisms such that the rates scale in parallel as L^{τ_ϵ}, albeit with non-motif insertions occurring at significantly lower rates. Thus, the per-repeat rate of insertions in our model is given by $\iota_L L = C_\iota L^{1+\tau_\iota} = C_\iota L^{1+\tau_\epsilon}$ (i.e., multiplying ι_L by L, rather than $L - 1$, to recover the estimated per-repeat rate). As before, this is expressed in terms of the length-independent constant $C_\iota = \iota_{L=8} \times m/9^{\tau_\epsilon}$, where we have substituted in $\tau_\iota = \tau_\epsilon$.

All additional mutational processes and their impact on the repeat length distribution are assumed to be subdominant or ignored for analytic simplicity. For example, deletions of B bases are ignored throughout our analysis despite potential relevance to fusion rates, as the direct estimates of the deletion rate for length one B strings showed an orders of magnitude suppression relative to the substitution rate μ (a directly competing process resulting in repeat fusion; see **Methods**). We make no attempt to model the correlated dynamics of the B string length distribution such that insertions of B bases that lengthen B strings are also ignored, along with the insertion or deletion of extended sequences (i.e., $L \to L \pm k$, where $k > 1$; see above). With the exception of extended insertions and deletions, these effects were included in the results of our computational model at appropriate length-independent rates (e.g., the rate of $B \to BB$ was assumed to be the same as the rate of $BB \to BBB$, ignoring potential repeat instability due to the indistinguishability between $B \in \{C, T, G\}$), but do not change the qualitative behavior of the repeat length distribution. Further, we assume the constant m is positive definite and the exponents τ_κ and τ_ϵ are positive semidefinite, which together imply that $C_\kappa, C_\epsilon, C_\iota > 0$.

Having defined each mutational process, we note that the location of the same type of mutation along the repeat sequence can result in dramatically different effects, as alluded to above. The catalogue of changes to repeat length can be separated into local transitions (i.e., increases or decreases in length by

one motif unit: $L \to L \pm 1$) and nonlocal transitions (i.e., $L \pm k$ where $k > 1$). While expansions are strictly local processes describing transitions in repeat length ($L \to L+1$) with probabilities associated with $P_L \to P_{L+1}$, μ substitutions are only local when they occur at sites adjacent to only one repeat boundary (e.g., $BBAB \to BAAB$, rather than $ABAB \to AAAB$). Importantly, μ substitutions adjacent to two distinct repeats result in an inherently nonlocal process in which two repeats (e.g., length L_1 and L_2) are fused into a single longer repeat of length L_1+L_2+1. This transition is both nonlocal in length space and non-conservative with respect to the mass of the probability distribution P(L). Transitions generating new $L = 1$ 'repeats' via substitution similarly do not conserve mass, effectively corresponding to a source at the $L = 1$ boundary. The latter is counteracted by a sink at the same boundary due to ν mutations (with a distinct rate) that remove $L = 1$ repeats. Analogously, repeat fusion is counteracted by repeat fission due to ν substitutions that occur in the non-boundary bases (i.e., in the 'body') of the repeat and are again nonlocal in length space and non-conservative in distribution mass; notably, this process has a target linear in the length of the original repeat (strictly, $L - 2$, ignoring the boundaries), in contrast to the target for fusions, which is dependent on the number of B strings of length one). Local decreases in length can occur due to contraction (at a per-repeat rate *at least* linear in the length) or via ν substitutions (at a rate 2ν corresponding to the finite target size of two bases for repeats $L \geq 2$). Additional fission events can occur due to insertions, but, as noted above, additional fusion events due to deletions of B bases will be ignored throughout.

The nonlocal, non-conservative, and non-linear (e.g., power law dependencies and fusion rates dependent on two distinct length classes) nature of this ensemble of dynamics makes the difficulty of simultaneously modeling these effects immediately clear. Instead, motivated by the fundamental questions surrounding the maintenance and prevalence of long (and potentially disease-causing) repeats in the genome, we proceed with an analysis of the asymptotic dynamics. This approach provides a more straightforward understanding of the dominant processes appropriate in long repeat regime (i.e., $L \gg 1$), and how this contrasts from the forces shaping the short repeat end of the distribution (roughly defined here as $L < 10$). Due to its simplicity and an immediately recognizable shape, we first focused on the low L end of the distribution.

1.1 Summary of parameterized mutation rates at long lengths

For clarity, we summarize the mutation rate parameterization for each mutation type. Repeat lengths below $L = 9$ units are taken directly from empirical point estimates at each length (see **Figure 2a**) calculated for expansion, contraction, and insertions. Substitution rates μ and ν are taken from motif-dependent empirical estimates and assumed to be length independent. For repeat length $L \geq 9$, the per-target rates for expansion (ϵ_L), contraction (κ_L), and insertion (ι_L) are parameterized as follows.

$$\begin{aligned} \epsilon_L &= C_\epsilon L^{\tau_\epsilon} \equiv \epsilon_8 m \left(\frac{L}{9}\right)^{\tau_\epsilon} \\ \kappa_L &= C_\kappa L^{\tau_\kappa} \equiv \kappa_8 m \left(\frac{L}{9}\right)^{\tau_\kappa} \\ \iota_L &= C_\iota L^{\tau_\epsilon} \equiv \iota_8 m \left(\frac{L}{9}\right)^{\tau_\epsilon} \end{aligned} \tag{SN1}$$

The parameter-dependent constants C_ϵ, C_κ, and C_ι are computed from ϵ_8, κ_8, and ι_8, respectively, the empirical estimates at $L = 8$. The dependence on the number nine is a result of the largest length with reliable empirical estimates; for the purposes of the following analysis it is only important that this number is of order 10. Per-repeat rates are computed by multiplying each per-target rate by repeat length L (e.g., $\epsilon_L L$ for expansions, etc.). Noting that we have made the simplifying assumption that insertions and expansions obey the same power law (i.e., $\tau_\iota = \tau_\epsilon$), this provides a three-parameter description of the unobserved length dependencies of the mutation rates in terms of $\{m, \tau_\epsilon, \tau_\kappa\}$. This parameterization is used in our computational model to propagate the mutational process over time and substituted into more general analytic expressions below for direct comparison.

2 Finite difference equation for repeat length changes

We first modeled the dynamics of repeat length changes by writing a discretized finite difference equation that describes the combined effects of all five mutational processes on the distribution $\mathrm{P}_L(t)$ as it evolves over time.

$$\mathrm{P}_L(t+1) \approx \mathrm{P}_L(t) + \Delta_\epsilon \mathrm{P}_L(t) + \Delta_\kappa \mathrm{P}_L(t) + \Delta_\iota \mathrm{P}_L(t) + \Delta_\mu \mathrm{P}_L(t) + \Delta_\nu \mathrm{P}_L(t) \tag{SN2}$$

Here, we have assumed that all mutation rates are sufficiently small that the mutational processes can be linearized (i.e., the rate of multiple mutations on the same repeat in a single generation is negligibly small). This equation describes the discrete distribution at time $t+1$ as it evolves from the distribution at the previous time step due to changes induced by expansion, contraction, insertion, and substitutions (from both μ and ν mutations), respectively. This can be rearranged to express the change to the distribution for each length bin L in a single generation.

$$\Delta_t \mathrm{P}_L \approx (\Delta_\epsilon + \Delta_\kappa + \Delta_\iota + \Delta_\mu + \Delta_\nu)\mathrm{P}_L \tag{SN3}$$

Each term in this expression is defined by the single-generation effect of a given mutational type on the number of length L repeats P_L.

2.1 Changes in repeat length due to expansion, contraction, and insertion

Expansions introduce single-unit changes in length ($L \to L+1$), independent of their location within the repeat (target size L, as defined above). The change in the number of repeats in the Lth class can be written as a combination of an influx due to insertion from below ($L-1 \to L$) and an outflux due to expansion to the class above ($L \to L+1$), as follows.

$$\Delta_\epsilon \mathrm{P}_L = \epsilon_{L-1}(L-1)\mathrm{P}_{L-1} - \epsilon_L L \mathrm{P}_L \tag{SN4}$$

As defined above, ϵ_L represents the length-dependent expansion rate, which dramatically and monotonically increases above lengths of order ten. Contractions have the opposite effects: an influx from above ($L+1 \to L$) and an outflux to below ($L \to L-1$), with a length-dependent mutation rate κ_L and target size L.

$$\Delta_\kappa \mathrm{P}_L = \kappa_{L+1}(L+1)\mathrm{P}_{L+1} - \kappa_L L \mathrm{P}_L \tag{SN5}$$

Importantly, we have omitted the effect of contraction-based repeat fusion, under the assumption that this process is subdominant to substitution-based fusion at all repeat lengths (see description of substitution-based fusion, below). (Non-motif) insertions are described by more complicated transitions, due to the position-dependent effects on length. Additionally, insertions are non-conservative, replacing one repeat with two repeats of shorter length ($L \to k, L-k$, where $k \geq 1$). The outflux due to insertions occurs at a length-dependent rate ι_L with target size L, which is a sum over the L possible locations for the insertion that result in distinguishable transitions (up to the symmetry $k \leftrightarrow L-k$; e.g., transitions $5 \to 4, 2$ are equivalent to $5 \to 2, 4$, despite distinguishable mutational targets). Note that, for the purposes of writing the finite difference equation, one need not keep track of the eventual state(s) resulting from an outflux. A focal class L can gain counts due to insertions in any repeat longer than L: insertions that occur L units away from either repeat boundary result in an increase in P_L, such that each length class $l > L$ has 2 potential targets for transitions $l \to L, l-L$ (in the case where $l = 2L$, there is a single target, but two length L repeats are added to P_L). Together, the effects of insertion on a focal L class can be written as follows.

$$\Delta_\iota \mathrm{P}_L = -\iota_L L \mathrm{P}_L + 2 \sum_{l=L+1}^{\infty} \iota_l P_l \tag{SN6}$$

Here, we assume that P_L decays sufficiently rapidly as L increases such that $\sum_l \iota_l P_l$ remains finite and P_L is normalizable. This sum characterizes repeat fission due to insertions alone, which are inherently nonlocal and

non-conservative transitions. However, note that contributions from transitions $L+1 \to L$ are local (though non-conservative) effects included in the same contribution. In this sense, this sum can be considered a combination of local transitions $L+1 \to L$ (contributing $2\iota_{L+1}P_{L+1}$) and nonlocal transitions $l \geq L+2 \to L$ (contributing $2\sum_{l=L+2}^{\infty} \iota_l P_l$); this decomposition is relevant only when insertion rates are appreciable at long repeat length, which we approximate in the continuum limit (detailed below). Similarly, the target size L in the outflux in Equation SN6 is a sum of targets for local and nonlocal transitions (with targets 2 and $L-2$, respectively).

2.2 Changes in repeat length due to substitutions

Substitutions behave distinctly from the above transitions, though some aspects of the transitions they induce are analogous. For example, the outflux due to shortening substitutions is analogous to insertions with a target size L, but length-independent rate ν; this combines substitutions at the boundary that result in local transitions $L \to L-1$ (with target 2) and substitutions in the repeat body (target $L-2$) that result in repeat fission. The influx due to ν substitutions is also analogous to insertions, represented by a sum over contributions from fissions of repeats of length $l \geq L+2 \to L$ and local transitions $L+1 \to L$.

$$\Delta_\nu \mathrm{P}_L = -\nu L \mathrm{P}_L + 2\nu P_{L+1} + 2\nu \sum_{l=L+2}^{\infty} P_l = -\nu L \mathrm{P}_L + 2\nu \sum_{l=L+1}^{\infty} P_l \tag{SN7}$$

Again, we assume $\sum P_l$ rapidly converges to a finite value such that P_L is normalizable, though this condition is weaker than for insertions due to the (naively) monotonic increase in the per-target rate ι_L.

Length-increasing substitutions with length-independent rate μ per target result in an outflux restricted to sites adjacent to the repeat boundary (i.e., μ mutations $B \to A$ at either end of $BA...AB$). This includes local transitions when the mutated B is initially adjacent to less than two A bases (e.g., $BBA \to BAA$ for transitions $L \to L+1$ or $BBB \to BAB$ for transitions $L = 0 \to 1$) and nonlocal transitions when the mutated base results in repeat fusion ($ABA \to AAA$). Here, the total target size is 2 (i.e., not separated into local vs. nonlocal transitions) and the net outflux is independent of the resulting repeat length(s).

The influx due to μ substitutions has distinct contributions from local transitions ($L-1 \to L$) and nonlocal fusion of shorter repeats ($L_1, L_2 \to L$, where $L_1 + L_2 = L-1$). For the purposes of describing the discrete dynamics, we can introduce a time-dependent constant $p_F(t)$ that represents the steady-state probability that mutation of a B base terminating a repeat results in repeat fusion, rather than a local transition; $p_F(t)$ (henceforth, the time-dependence of p_F will be left implicit for brevity) is dependent on the genomic distribution of B string lengths at time t (i.e., the fraction of B strings of length one) or, alternatively, the probability that the three-unit context of the mutated B results in a substitution $ABA \to AAA$ (i.e., the probability that the mutated B is adjacent to two A repeats). The rate of substitution-based repeat fusions is therefore proportional to μp_F and the rate of local length-increasing substitutions is proportional to $\mu(1-p_F)$. The quantity p_F can be measured empirically at a given time (at any time in steady state) from the B string length distribution, computed from other genome-wide quantities like the length of the genome and total number of repeats, or computed theoretically in simple cases (see discussion below), but is ultimately unimportant to our analysis and results. The combined effects of μ mutations on repeats of length L can be characterized as follows.

$$\Delta_\mu \mathrm{P}_L = -2\mu \mathrm{P}_L + 2\mu(1-p_F)\mathrm{P}_{L-1} + \mu p_F \sum_{l=1}^{L-2} P_l P_{L-1-l} \tag{SN8}$$

The sum in the rightmost term is the discrete form of a convolution of the distribution P_L with itself, which has an intuitive interpretation. For integers distributed according to the discrete probability distribution P_L (assuming P_L is properly normalized), the probability of randomly sampling two integers $L_1, L_2 (< L-1)$ that sum to a fixed value $L-1$ is given by the convolution $\sum_{l=1}^{L-2} P_l P_{L-1-l}$; a substitution adds one unit to form a repeat of length L. Here, we have assumed that distinct classes of P_L remain uncorrelated in steady state for simplicity. This may be an oversimplification of the true steady state because fusion events may preferentially reverse repeat fission (i.e., fusion can re-form the original length of an interrupted repeat generated by fission;

the length-dependent per-target rate of insertion-based interruptions can result in correlation between length classes, further complicating the dynamics).

We note that there can be a similar contribution that originates from deletions of a B adjacent to two repeats (i.e., deletion-based repeat fusion). However, based on the estimated mutation rates (see Methods), background deletions are orders of magnitude less frequent than μ substitutions such that deletion-induced fusion transitions occur at negligible rates.

2.3 Finite difference equation

Summing the above contributions to $\Delta \mathrm{P}_L$ yields the full finite difference equation describing the change in each length class P_L subject to two-way substitutions, expansions, contractions, and insertions.

$$\begin{aligned}\Delta_t \mathrm{P}_L = &-(\nu L + 2\mu)\,\mathrm{P}_L + 2\nu \sum_{l=L+1}^{\infty} P_l + 2\mu(1-p_F)\mathrm{P}_{L-1} + \mu p_F \sum_{l=1}^{L-2} P_l P_{L-1-l} \\ &-(\epsilon_L + \kappa_L + \iota_L)\,L\mathrm{P}_L + \epsilon_{L-1}(L-1)\mathrm{P}_{L-1} + \kappa_{L+1}(L+1)\mathrm{P}_{L+1} + 2\sum_{l=L+1}^{\infty} \iota_l P_l \end{aligned} \tag{SN9}$$

For clarity, the first line of terms summarizes changes due to substitution and the second summarizes insertion- and deletion-based changes due to repeat instability. Inclusion of deletion-based fusion and/or correlations between length classes, both of which are treated as negligible, would add an additional convolution and/or a dependence on the covariance between length classes, respectively.

Guided by striking differences in the length-dependencies of substitutions and repeat instability rates (**Figure 2a**), further analyze the steady state dynamics by separating into distinct length regimes in which a subset of mutational processes dictate the vast majority of changes in length. Under this separation of length scales, the dynamics of short repeats (roughly, $L \leq 8$ for mononucleotide-A repeats) and long repeats (roughly $L > 10$) are largely dominated by distinct mutational forces: the distribution of short repeats is maintained in a dominant balance between the opposing effects of μ and ν substitutions (including both fission and fusion), while length changes to long repeats are dominated by a (parameter-dependent) balance between expansions, contractions, and fission (potentially including fission resulting from ν substitutions). The assumption of dominant balance allows for a dramatic simplification of the full set of contributions to Equation SN9; due to their low rates, all neglected terms provide minor corrections to the resulting approximation to the steady state distribution, which we demonstrate post-hoc.

Equation SN9 represents the deterministic change of the repeat length distribution in a reference sequence (i.e., a single individual) due to mutations alone. Here, we have assumed that natural selection is absent such that the accumulation of mutations in the reference results from mutations aggregating along the lineage ancestral to this individual. This process occurs over sufficiently long times that a steady state is eventually reached. Once in steady state, the time-averaged distribution of repeat lengths (i.e., averaging over stochastic behavior) can be equated to the steady state distribution obtained from Equation SN9.

3 Steady-state dynamics in the short repeat length regime

In contrast to previous studies, the empirical distribution we constructed includes very short sequences and, notably, single base sequences for comparison. The relative rates of expansions, contractions, and insertions to both types of substitutions makes it clear that repeat instability is are largely irrelevant to the maintenance of such short sequences (see **Figure 2a**). From a biological standpoint, this implies that repeat instability is technically irrelevant (i.e., occurs at highly suppressed rates) until repeats exceed roughly $L = 8$ (this differs slightly between motifs of distinct lengths and is more accurately described as roughly 8-10 nucleotides, rather than repeat units, for motifs of length $l_m \leq 4$ nucleotides; see **Supplementary Figure S2**,); this corresponds to the length range where expansion and contraction rates are comparable to substitution rates such that repeat instability becomes relevant to the dynamics.

The shape the distribution of short repeats is well-approximated as a balance between substitutions alone. The distribution is generated by the random process of sequence evolution under two-way substitution with distinct rates μ and ν. Given indefinite time, this process equilibrates to a steady state genome for which the probability of randomly sampling an A base is given by the fraction $p_A = \mu/(\mu+\nu)$ and the complementary probability of sampling a B base is given by $p_B = 1 - p_A = \nu/(\mu+\nu)$. Here, the relevant substitution rates μ and ν are the single-unit context rates (i.e., averaged over longer contexts) $\mu_{B\to A}$ and $\mu_{B\to A}$, respectively, as this distribution is not characterized by distinctions between local transitions, fission, and fusion. Under this model, a repeat is simply a contiguous sequence of A bases L bases long adjacent to a B base on either side. Conditioning on an initial B base, a length L string of A bases occurs at a frequency approximately given by the probability of randomly sampling an A base L successive times, followed by a terminating B base. The frequency of a length L repeat is therefore given by geometrically distributed distribution proportional to ${p_A}^L p_B$.

$$\mathrm{P}_L = \mathcal{N}\left(\frac{\mu}{\mu+\nu}\right)^L \left(\frac{\nu}{\mu+\nu}\right) \propto \left(\frac{\mu}{\mu+\nu}\right)^L \tag{SN10}$$

Here, $\mathcal{N}$ is a normalization constant defined as $\mathcal{N} = 1/\sum_{L=1}^{\infty} \mathrm{P}_L$, where $\sum_L \mathrm{P}_L$ is the total mass of the distribution. As the steady state distribution is no longer geometrically distributed when repeat instability becomes relevant roughly above $L = 8$, the value of $\mathcal{N}$ cannot be determined by the short repeat dynamics alone and must be evaluated only after identifying the steady state distribution over all length classes. Given that the constant $\mathcal{N}$ is unknown, the constant probability associated with sampling repeat-terminating B bases can be absorbed into the definition of $\mathcal{N}$.

3.1 Geometric solution to the substitution-only difference equation

To demonstrate the utility of the finite difference equation in a simpler setting, one can show that the geometric distribution, when normalized, provides the solution to the short length approximation to Equation SN9 in steady state. Under substitutions alone, the short length regime is well-approximated by the following difference equation.

$$\Delta \mathrm{P}_L \approx -\nu L \mathrm{P}_L - 2\mu \mathrm{P}_L + 2\nu \sum_{k=L+1}^{\infty} P_k + 2\mu(1-p_F)P_{L-1} + \mu p_F \sum_{k=1}^{L-2} P_k P_{L-1-k} = 0 \tag{SN11}$$

The finite difference equation becomes a steady state condition by imposing time independence (i.e., $\Delta \mathrm{P}_L = 0$). For consistency with Equations SN7 and SN8, the local influx from ν substitutions is included in the sum representing fission and the parameter $p_F = (1 - q_F)$ will be defined below to represent the appropriate rate of fusion in the present context. The normalized geometric distribution is given by the following expression in terms of p_A and $p_B = 1 - p_A$, defined above.

$$\mathrm{P}_L = p_B {p_A}^{L-1} = \frac{\nu \mu^{L-1}}{(\mu+\nu)^L} \tag{SN12}$$

Though fusion and fission are inherently non-conservative transitions that change the number of repeats in the distribution, we make the approximation that individual fission and fusion events do not significantly alter the normalization constant in steady state. The term representing fission is now the incomplete sum over a geometric series, which can be explicitly evaluated as follows.

$$2\nu \sum_{k=L+1}^{\infty} P_k = 2\nu p_B \sum_{k=L+1}^{\infty} {p_A}^{L-1} = 2\nu p_B \frac{{p_A}^L}{p_B} = 2\mu \mathrm{P}_L \tag{SN13}$$

Thus, the influx due to both repeat fission and local transitions, both ν substitution-based effects, exactly cancel the outlux due to μ mutations.

To evaluate the convolution term, note that the probability distribution for the length of B strings under two-way substitution $\tilde{P}_\lambda$ is geometric by the same arguments as for the A distribution, but with reversed

probabilities $p_A \leftrightarrow p_B$. Normalizing string lengths $\lambda = [1, \infty)$ (i.e., conditional on a length $\lambda \geq 0$ B string), the normalized probability distribution is as follows.

$$\tilde{P}_\lambda = p_A {p_B}^{\lambda-1} \tag{SN14}$$

The probability of the B string that terminates any given A string being length $\lambda = 1$ is simply $\tilde{P}_1 = p_A$. This is also the steady state probability of fusion p_F, for which a μ substitution in a length one B string results in the transition $L_1, L_2, \rightarrow L_1 + L_2 + 1$. The complement is the probability of locally increasing the length by a single unit (as opposed to fusion) due to an adjacent μ substitution, $q_F = 1 - \tilde{P}_1 = 1 - p_A = p_B$. After cancelling the ν substitution influx with the μ substitution outflux, we can rewrite the remaining terms in Equation SN11 by substituting in $p_F = p_A$, $q_F = p_B$ and evaluating the partial sum in the fusion term.

$$\begin{aligned}
-\nu L \mathrm{P}_L + 2\mu q_F P_{L-1} + \mu p_F \sum_{k=1}^{L-2} P_k P_{L-1-k} &= -\nu L \mathrm{P}_L + 2\mu p_B P_{L-1} + \mu p_A {p_B}^2 \sum_{k=1}^{L-2} {p_A}^{k-1} {p_A}^{L-k-2} \\
&= -\nu L \mathrm{P}_L + 2\mu \frac{p_B}{p_A} \mathrm{P}_L + \mu p_A {p_B}^2 \sum_{k=1}^{L-2} {p_A}^{L-2} \\
&= -\nu L \mathrm{P}_L + 2\nu \mathrm{P}_L + \nu (L-2) p_B p_A {p_A}^{L-2} \\
&= -\nu (L-2) \mathrm{P}_L + \mu (L-2) \mathrm{P}_L = 0
\end{aligned} \tag{SN15}$$

We find that a geometric distribution of both A and B repeats satisfies the steady state equation under two-way substitution, which justifies our aforementioned approximation for the substitution-dominated short length regime of the distribution.

3.2 Interactions between short and long repeats are largely restricted to boundary effects

For the present purposes, we are only concerned with the geometric falloff in Equation SN10 that defines the shape of the distribution in this length regime, which is given by the proportionality on the right hand side of the above expression. Importantly, this length dependence remains largely independent of the dynamics of the long length regime, which only affects the normalization constant. This separation of the dynamics for short and long length repeats is a reasonable approximation because the rate of transitions between these regimes is low and primarily limited to the intermediate lengths at $L \sim 8-10$ (i.e., the boundary between the short and long repeat regimes). As discussed in the context of the long length dynamics described below, we allow for an influx from the long length regime because it is inherently negligible due to the relative mass of short length repeats vastly outweighing the mass in the long repeat tail of the distribution. In this sense, the short repeat regime can be treated as a probability sink for the long repeat distribution for length transitions that exit the long length regime. Similarly, the mass in the long length regime is sourced by short repeats that approach the length regime boundary and are quickly subject to expansion-biased instability; this can be considered a local source at the boundary between the regimes without altering the dynamics elsewhere. If the steady state is in detailed balance (i.e., the flux in and out of each length class vanishes independently), fluxes at the length regime boundary must cancel such that both ends of the distribution maintain their relative weights.

While it is immediately clear that the distribution of A repeats deviates from a simple geometric decay due to the combined action of expansion, contraction, and insertion in the long length regime, the B string length distribution is unaltered by expansion and contraction, which solely manipulate A repeat length. While likely unrealistic, we proceeded under the assumption that B strings do not constitute repeats and are therefore not directly subject to repeat instability. Therefore, any deviation from a simple geometric distribution can only be due to the effects of insertion, which generates new length one B strings during each transition. Although the additional source of B strings will contribute to the total number of B strings in the eventual steady state balance between the A and B distributions, localization of this influx to the lowest length class implies that the normalized distribution of B lengths remains approximately geometric.

4 Steady-state dynamics for asymptotically long repeat lengths

The dynamics the steady state distribution at large lengths is qualitatively distinct in shape from the low length regime, with a much heavier tail than the geometric distribution that decays more slowly with length. Although one must speculate about the behavior in a length regime without direct estimates from de novo data, the assumption that repeat instability is associated with a (naively monotonic) length-dependent increase in the per-target expansion and contraction rates implies that at some length these mutational processes dominate over length-independent per-target substitution rates; this relative increase in rates can already be seen for lengths of order 10 (see **Figure 2a**). Consequently, these rates necessarily continue to increase in excess of the substitution rate, eventually dominating the mutational dynamics. However, the point at which substitution-based fission, which scales linearly with repeat length, can be ignored is entirely parameter dependent. As a result, we can infer that the forces relevant to the longest length repeats present in the distribution are some subset of expansion, contraction, substitution-based fission, and insertion-based fission (or the combination of all four). To analyze the dynamics at long repeat length and characterize the relative importance of each mutational process, we took the large length continuum limit of the full finite difference equation under the approximation that, for long repeats, $L \gg 1$ such that the distance between length bins can be treated as infinitesimal. While the continuum limit may be appropriate above lengths of order 10, this description remains an approximation to the discrete dynamics detailed in Equation SN9. In the continuum, the distinction between local and nonlocal terms becomes important, as local differences are approximated by continuous derivatives, while the nonlocal sums in Equation SN9 that capture transitions from across the rest of the distribution are approximated as integral dependencies.

4.1 Local contributions to the change in P_L

We first focus on local changes to the length distribution due to each of the forces. Starting with expansion in Equation SN4, the discrete change to a focal length class P_L is the difference between an influx due to $L-1 \to L$ length-increasing transitions and an outflux due to $L \to L+1$ length-increasing transitions. All effects of expansion are entirely local, which allows for a local approximation to $\Delta_\epsilon \mathrm{P}_L$ using a Taylor expansion in the continuum. To elucidate the continuum limit for large repeat length, we treat expansions in detail, below, but note that the continuum approximation for all local contributions can be obtained analogously. We first define the distance between discrete length classes $\Delta L = 1$, which suggests that we must change units before formally taking the continuum limit and assessing the accuracy of the resulting approximation.

$$\begin{aligned}
\Delta_\epsilon \mathrm{P}_L &= (L-1)\epsilon_{L-1}\mathrm{P}_{L-1} - L\epsilon_L \mathrm{P}_L \\
&= F_\epsilon(L-\Delta L) - F_\epsilon(L) \\
&= F_\epsilon(x-\Delta x) - F_\epsilon(x)
\end{aligned} \tag{SN16}$$

Here, ϵ_L is the length-dependence of the expansion rate (e.g., $\epsilon_L = C_\epsilon L^{\tau_\epsilon}$ under our parameterization for $L \geq 9$). The function $f_\epsilon(L) = \epsilon_L L P_L$ was defined solely for notational convenience. In the third line, repeat length L was rescaled under a change of variables defined by $x = L/c$ (with $\Delta x = \Delta L/c = 1/c$); c is an arbitrary dimensionful constant with the same units as L (i.e., measured in a number of motif units) such that x is dimensionless. If x is kept fixed while L grows large, Δx gets increasingly smaller. Thus, choosing an appropriate c allows the backward finite difference $\Delta F_\epsilon(x) = F_\epsilon(x-\Delta x) - F_\epsilon(x)$ to be approximated by a Taylor expansion in small Δx truncated at a some finite order (for all practical purposes, this is equivalent to a Taylor expansion around "small" ΔL after setting $\Delta L = 1$ when $L \gg 1$, but side steps the concern that $\Delta L = 1$ motif unit is definitionally not infinitesimal). Taking the continuum limit $L \to \infty$ and $\Delta x \to 0$, F_ϵ becomes the continuous function f_ϵ and the discrete derivative approaches the continuous derivative $\Delta F_\epsilon \to \partial_x f_\epsilon$. For finite L (and thus finite c), the first-order continuum approximation is accurate

to order Δx^2.

$$\Delta_\epsilon \mathrm{P}_L = \frac{\partial x}{\partial L}\frac{\partial f_\epsilon(x)}{\partial x} + \mathcal{O}(\Delta x^2) \approx -\partial_L\left[\epsilon_L L \rho(L)\right] \tag{SN17}$$

This is expressed in terms of a general, length-dependent expansion rate ϵ_L (at sufficiently long lengths) to allow for generic parameterizations, including Equation SN1. To arrive at this approximation, the variable change was inverted (i.e., imposing $L = cx$, where $\partial L/\partial x = 1/c$; the chain rule is explicit in the middle expression for clarity) after taking the continuum approximation. Here, the notation ∂_x^n represents the nth derivative with respect to x and will be used henceforth for brevity. In the length continuum, we assume that the discrete distribution P_L can be well-approximated by the continuous, differentiable function $\rho(L)$ (defined in the rescaled length units $\rho(x) \equiv \lim_{\Delta x \to 0} P(x)$) such that the derivative $\partial_L \rho(L)$ represents the flux through length L. The continuum approximation is only applicable for $L \gg 1$ and all expressions dependent on $\rho(L)$ should be considered implicitly conditioned on $L \gtrsim 10$ such that the long-length scaling behavior of the mutation rates can be used directly (i.e., empirically-derived rates for $L < 9$ can be ignored such that length scaling for each rate is dictated by the parameterization, e.g., Equation SN1). Importantly, at finite L, corrections to the lowest-order approximation to the finite difference $\Delta F(L)$ for a discrete function $F(L)$ cannot be treated as negligible when $\partial_L f(L)$ vanishes (see discussion of second-order corrections, below). Equation SN17 describes a strictly local transition that drives a length-dependent flux through the focal length L, which represents the collective impact of expansion rates on the distribution; if desired, the chain rule can be applied to show a separate flux and length-dependent loss of mass due to expansion (i.e., $\Delta_\epsilon \mathrm{P}_L \approx -\rho(L)\partial_L(\epsilon_L L) - \epsilon_L L \partial_L \rho(L)$). Noting that the derivative is negative throughout the roughly monotonic empirical distribution, the flux due to expansion may describe an overall increase or decrease in length, depending on the relative magnitude of the non-derivative contribution.

The effects of contraction $\Delta_\kappa \mathrm{P}_L$ can be summarized analogously in terms of contributions from and to the adjacent longer and shorter length bins, respectively, which becomes a local length-dependent flux in the opposite direction in the continuum limit. This sign difference emerges because the continuum approximation is to a forward finite difference for contractions and to a backward finite difference for expansions (i.e., the difference between continuum approximations for $L+1 \to L$ and $L-1 \to L$ transitions).

$$\Delta_\kappa \mathrm{P}_L = (L+1)\kappa_{L+1}\mathrm{P}_{L+1} - L\kappa_L \mathrm{P}_L \approx \partial_L\left(\kappa_L L \mathrm{P}_L\right) \tag{SN18}$$

Like expansions, contractions amount only to local changes to the distribution, but the length dependence of the per-repeat rate remains important to our understanding of the dynamics.

In contrast, substitutions induce both local and nonlocal transitions. We first isolate the local transitions from in Equation SN7 and SN8, treating the nonlocal contributions as fissions and fusions (see below). In discrete form, the local influx and outflux towards increasing length due to μ substitutions and towards decreasing length due to ν substitutions can be represented as follows.

$$\left(\Delta_\mu + \Delta_\nu\right)_{local} \mathrm{P}_L = 2\mu(1-p_F)\left(\mathrm{P}_{L-1} - \mathrm{P}_L\right) + 2\nu\left(\mathrm{P}_{L+1} - \mathrm{P}_L\right) \tag{SN19}$$

Unlike expansion and contraction, the target size for local changes in length due to substitution is a length-independent factor of 2 associated with mutations at either boundary (e.g., $AAA \to AAB$ and $\to BAA$ for ν substitutions; μ substitutions reverse these transitions), which have a distinct effect from mutations at non-boundary loci. The target size for μ substitutions is reduced further because a fraction p_F transitions on the boundary result in fusion, rather than local increases in length (i.e., leaving a target of $2(1-p_F) = 2q_F$ per repeat). The first-order continuum approximation to Equation SN19 is the following.

$$\left(\Delta_\mu + \Delta_\nu\right)_{local} \mathrm{P}_L \approx -\partial_L\left(2\left(\mu(1-p_F) - \nu\right)\rho(L)\right) \tag{SN20}$$

Insertions include a similar local term, but this comes with an additional complication. If an insertion occurs adjacent to either boundary (with target size 2: $AA...AA \to ABA...AA$ and $\to AA...ABA$), the resulting repeat fission desribes the replacement of one length L repeat with one length $L-1$ repeat and

one length 1 'repeat' ($L \to L-1, 1$). This is the only example of repeat fission that also contains a local transition, but will not be explicitly accounted for in our treatment of insertion-based fission below. The local component of this contribution can be written as follows.

$$(\Delta_\iota)_{local} \mathrm{P}_L = 2\iota_{L+1}\mathrm{P}_{L+1} - 2\iota_L \mathrm{P}_L \approx \partial_L \left(2\iota_L \rho(L)\right) \tag{SN21}$$

Due to the limited target of two possible insertions resulting in this local transition, the appropriate length dependence inside the derivative is twice the per-target rate ι_L, rather than a dependence on the per-repeat insertion rate. The correlated nonlocal contribution to the $L = 1$ bin can be ignored because the relatively low rate of insertions results in an exceedingly small influx into $L = 1$ relative to the geometrically distributed mass maintained by substitutions.

4.1.1 Competition between first-order local effects

The length-independent (and modest) target size, matching power law exponent, and globally suppressed rate of insertions relative to expansions together ensure that local insertions remain subdominant to expansions (i.e., $(\Delta_\iota)_{local} \mathrm{P}_L \ll \Delta_\epsilon \mathrm{P}_L$ for any L); as a result, local insertions may be neglected entirely. By the same argument, and given that expansion and contraction rates far exceed the substitution rates at long lengths, local changes in length due to substitutions are likely negligible in the $L \gg 1$ regime of interest. Additionally, we will at this point treat time as continuous (i.e., assuming infinitesimal generation time after appropriately rescaling the units of all rates) such that the per-generation change in occupancy of P_L can be approximated by the time-derivative $\partial_t \rho(L)$ (note that we revert back to the notation $\Delta \mathrm{P}_L$ when referencing the discrete equations). Under these approximations, the first-order approximation to the local dynamics can be summarized by the following.

$$\begin{aligned}(\partial_t \rho(L))_{local} &\approx -\partial_L \left[\left((\epsilon_L - \kappa_L)L - 2\iota_L + 2(\mu(1-p_F) - \nu)\right)\rho(L)\right] \\ &\approx -\partial_L \left[\left((\epsilon_L - \kappa_L)L\right)\rho(L)\right]\end{aligned} \tag{SN22}$$

The quantity $\epsilon_L - \kappa_L$ represents the bias between expansion and contraction at a given length, indicating that the lowest-order approximation to the local dynamics describes a flux through each length class due to the length-dependent expansion-contraction bias.

The sign difference between the expansion and contraction terms suggests that, if their magnitudes are nearly identical at a given length (i.e., $(\Delta_\epsilon + \Delta_\kappa)P_l \approx 0$ such that length changes are symmetric at length l), otherwise subdominant contributions from insertion and/or substitution could become relevant. Under the somewhat simplistic parameterization in Equation SN1, this can occur only at a specific (but parameter-dependent) length L^*.

$$L^* \equiv \left(\frac{C_\epsilon}{C_\kappa}\right)^{\frac{1}{\Delta\tau}} = 9\left(\frac{\epsilon_8}{\kappa_8}\right)^{\frac{1}{\Delta\tau}} \tag{SN23}$$

We have expressed this in terms of the collapsed parameter $\Delta\tau$, the difference between the contraction and expansion exponents.

$$\Delta\tau \equiv \tau_\kappa - \tau_\epsilon \tag{SN24}$$

L^* is independent of the multiplier m and dependent only on $\Delta\tau$, rather than τ_ϵ and τ_κ individually. This length is notable because subdominant local contributions (i.e., from insertions and substitutions) could theoretically become relevant to the maintenance of an eventual steady state at specific lengths. However, second-order corrections (due to expansion and contraction) to the finite difference equation dramatically exceed the local effects of insertions and substitutions due to their scaling with target size.

4.1.2 Second-order corrections to the local behavior and diffusive dynamics

When the first-order approximation to the local dynamics vanishes (due to nearly symmetric expansion an contraction rates across some range of lengths), subsequent corrections to the continuum approximation at finite L can dominate the leading-order local behavior. The space of allowed parameter values allows L^* to sit within the populated range of long lengths such that a description of the dynamics across the full set of $\{m, \tau_\epsilon, \tau_\kappa\}$ parameter values requires further approximation of local length changes. For parameters consistent with the empirical distribution, L^* typically sits within the well-occupied range of the long-length tail due to modest values of $\Delta\tau$; for comparatively small $\Delta\tau$, the length scalings for expansion and contraction are similar such that the neighborhood of L^* (i.e., the range of lengths with small expansion-contraction bias) extends to a wider range of lengths. In this sense, both the location and neighborhood of L^* are dictated by $\Delta\tau$ and together indicate expansion-contraction bias alone is insufficient to describe local length changes observed in humans.

With this in mind, we obtained the next-order approximation (i.e., the first strictly non-vanishing contribution) to Equation SN9 by Taylor expanding the expressions for expansion, contraction, insertion, and substitution to second order in the continuum limit. For expansion, we approximate Equation SN17 to order Δx^2 in the dimensionless variable x (i.e., the appropriately rescaled L) to produce the following approximation.

$$\begin{aligned}\Delta_\epsilon \mathrm{P}_L &= \left[f_\epsilon(x) - \left[\partial_x f_\epsilon(x)\right] \Delta x + \frac{1}{2}\left[\partial_x^2 f_\epsilon(x)\right] \Delta x^2 + \mathcal{O}(\Delta x^3) \right] - f_\epsilon(x) \\ &\approx -\partial_L \left[\epsilon_L \rho(L)\right] + \frac{1}{2}\partial_L^2 \left[\epsilon_L \rho(L)\right] \end{aligned} \tag{SN25}$$

All other mutational effects can be expanded to the same order analogously,. Importantly, the second-order contribution is strictly positive in each case such that their sum is strictly non-vanishing (in contrast to the sign difference that allows the first-order term to vanish at L^*). This results in the following approximation to the continuous-time local dynamics in the large length regime.

$$\begin{aligned}(\partial_t \rho(L))_{local} \approx &-\partial_L \left[\left((\epsilon_L - \kappa_L) L - 2\iota_L + 2\left(\mu(1-p_F) - \nu\right) \right) \rho(L) \right] \\ &+ \frac{1}{2}\partial_L^2 \left[\left((\epsilon_L + \kappa_L) L + 2\iota_L + 2(\mu(1-p_F) + \nu) \right) \rho(L) \right] \end{aligned} \tag{SN26}$$

The above (second-order) approximation in the continuum limit ($L \gg 1$) expresses the sum of local length changes for arbitrary length-dependent rates of expansion, contraction, and insertion, either estimated directly, or parameterized; higher order approximations would only become necessary in the neighborhood of a length at which the first and second derivative terms both vanish (for example, in the unrealistic scenario where inflection points in the length dependencies of $L\epsilon_L\rho(L)$, $L\kappa_L\rho(L)$, $\iota_L\rho(L)$, and $\rho(L)$ occur at the same length). Introducing the definitions of ϵ_L, κ_L, and ι_L under the power law parameterization allows for comparison between these length dependencies (along with the relative magnitudes of their local contributions).

$$\begin{aligned}(\partial_t \rho(L))_{local} \approx &-\partial_L \left[\left(C_\epsilon L^{1+\tau_\epsilon} - C_\kappa L^{1+\tau_\kappa} - 2C_\iota L^{\tau_\epsilon} + 2\left(\mu(1-p_F) - \nu\right) \right) \rho(L) \right] \\ &+ \frac{1}{2}\partial_L^2 \left[\left(C_\epsilon L^{1+\tau_\epsilon} + C_\kappa L^{1+\tau_\kappa} + 2C_\iota L^{\tau_\epsilon} + 2(\mu(1-p_F) + \nu) \right) \rho(L) \right] \end{aligned} \tag{SN27}$$

The second derivative terms in this approximation can be collectively viewed as describing diffusion-like changes to repeat length. Unlike the bias-driven fluxes represented by the first derivative term, the second-order term generates symmetric, bidirectional transitions (i.e., equiprobable local increases and decreases in length).

Interpreting the local dynamics, repeat instability drives a monotonic increase in the rate of expansion that results in a flux towards higher lengths represented by the first order term in Equation SN25; similarly, the second order term represents an increasing rate of diffusion with increasing length. Unlike the competing directional effects at first order, the strictly positive second-order contributions are net additive (i.e., both

expansion and contraction diffuse bidirectionally). Expansion-driven diffusion compounds with contraction-driven diffusion such that the length of longer repeats is increasingly unstable, including in the neighborhood of L^*. While the mutational forces do not counteract in this context, their relative rates can be compared to determine the dominant mutational processes underlying length diffusion. As with the first derivative, the minimal target sizes for local substitutions are dramatically outcompeted by expansion and contraction; length-independent substitution rates generate a uniform rate of diffusion with minimal magnitude. Despite the length-dependent per-target rate of insertion, the minimal target size and a lower initial rate (i.e.,. $C_\iota \ll C_\epsilon$ such that $\iota_L \ll \epsilon_L$, independent of parameter values) results in entirely negligible contributions to length diffusion relative to expansion. Given the negligible effects of both substitution and insertion, independent of parameter values, the local dynamics are well approximated by the following expression only dependent on expansion and contraction.

$$(\partial_t \rho(L))_{local} \approx -\partial_L \left[(\epsilon_L - \kappa_L)\, L\rho(L)\right] + \frac{1}{2}\partial_L^2 \left[(\epsilon_L + \kappa_L)\, L\rho(L)\right] \tag{SN28}$$

This general expression can again be evaluated under any parameterization of interest, including that of Equation SN1, provided our assumptions about the subdominant length-dependent rates of insertions remain appropriate. Local changes in length are thus a competition between bias-driven directional flux and diffusive changes in length. Clearly the latter dominates in the neighborhood of L^*, but the relative importance of these effects is also parameter dependent, comparing the net expansion-contraction bias (i.e., the asymmetric component of repeat instability) to the average magnitude of expansion and contraction (the symmetric component).

4.2 Repeat fission as a nonlocal contribution to changes in length

In contrast to the local effects described above, repeat fission and fusion are inherently nonlocal processes. Fission is a consequence of interruptions due to insertions or ν substitutions that generate both a flux out of the focal L class (by producing two shorter repeats) and a corresponding flux into the same class from fissions of longer repeats; the total number of perfect repeats is altered by replacing one longer repeat with two shorter repeats. Conceptual distinctions between substitution and insertion in this context are minor in the continuum: insertions allow transitions to and from adjacent length classes, while substitutions require transitions to and from at least two length classes away. Additionally, insertion within a repeat necessarily results in fission and does not conserve total genomic length (though, this does not *a priori* alter the mass of the A repeat distribution), unlike substitutions. The most pertinent difference owes to the length-dependent per-target rate of insertions, which can dramatically increase the number of such events as repeat length increases. In both cases, the nonlocal nature of fission-related transitions makes modeling the dynamics inherently more complex, as the influx from all bins of higher length amounts to a sum of distinct contributions to the change in P_L at each length (see Equations SN6 and SN7).

4.2.1 Substitution-based fission

Fission occurs only as a consequence of ν substitutions, with no dependence on the substitution rate μ, which only increase repeat length. After removing the target for local transitions via substitution at either repeat boundary (see Equation SN19), the remaining target for nonlocal transitions is $L-2$, corresponding to the body of the repeat. While the local transitions generate the derivatives shown in Equation SN26, the $L \gg 1$ continuum limit of the remaining terms in Equation SN7 results in an integral over the distribution of repeats longer than the focal class P_L.

$$\begin{aligned}(\Delta_\nu \mathrm{P}_L)_{fiss} &= -\nu(L-2)\mathrm{P}_L + \sum_{l=L+2}^{\infty} 2\nu P_l \\ &\approx -\nu(L-2\Delta L)\rho(L) + 2\nu \int_{L+2\Delta L}^{\infty} \rho(\lambda)d\lambda \\ &\approx -\nu L\rho(L) + 2\nu \int_{L}^{\infty} \rho(\lambda)d\lambda \end{aligned} \tag{SN29}$$

Here, the lower limit of the sum differs from Equation SN7 after removing local transitions $L+1 \to L$. The continuum approximation applies only in the asymptotically long length regime where $L \gg \Delta L = 1$, justifying the approximations $L + 2\Delta L \approx L$ (in the integration limit) and $L - 2 \approx L$ (in the target size dependence of the outflux term) in the final expression, above. The linear length scaling of the outflux characterizes the fact that repeat length changes due to any substitution in the repeat body. In contrast, there are exactly two targets for fission transitions from any given longer length class (i.e., specific substitutions L units away from either boundary of a length λ repeat result in a shorter repeat of length L; in the special case where $\lambda = 2L + 1$, only a substitution in the middle base is relevant, but yields two length L repeats). All longer repeats therefore contribute identically to the influx, independent of length. The net effect of fission depends on the focal length and many repeats of longer length sit in the tail of the distribution (i.e., how rapidly it decays and truncates).

4.2.2 Insertion-based fission

Fission due to insertions can be described analogously by taking the continuum limit of the nonlocal component of Equation SN6 after removing the local transitions included in Equations SN21 and SN26. Note that, in this case, the length dependence of the per-target rate ι_L appears under the sum.

$$\begin{aligned}(\Delta_\iota \mathrm{P}_L)_{fiss} &= -(L-2)\iota_L \mathrm{P}_L + \sum_{l=L+2}^{\infty} 2\iota_l P_l \\ &\approx -(L-2\Delta L)\iota_L \rho(L) + \int_{L+2\Delta L}^{\infty} 2\iota_\lambda \rho(\lambda) d\lambda \\ &\approx -L\iota_L \rho(L) + \int_{L}^{\infty} 2\iota_\lambda \rho(\lambda) d\lambda\end{aligned} \tag{SN30}$$

As with substitutions, the target size $L - 2$ and limit of summation $l = L + 2$ explicitly refer to nonlocal transition away from the repeat boundary, both of which we again subsequently approximate as L in the continuum under the asymptotic assumption $L \gg \Delta L = 1$. The resulting contributions resemble Equation SN29, but the length-dependent function ι_L alters the length scaling of both terms (e.g., using the parameterization in Equation SN1, the outflux scales as $L^{1+\tau_\epsilon}$). This results in a weighted integral over the tail of the distribution, which increases the rate of fission into the focal class, despite a finite target size of two per repeat. This is important to our asymptotic analysis, as the net direction of fission, i.e., the relative weight of the two terms in Equation SN30, now depends explicitly on the exponent τ_ϵ.

4.3 Repeat fusion under random sampling of the length distribution

Repeat fusion, the process by which either μ substitutions or, at a lower rate, deletions of length one B strings (henceforth *B-dels* for brevity) result in the merging of two shorter repeats into a longer repeat, substantially complicate our model of repeat length dynamics. For simplicity, we only describe μ substitution-based fusion (as shown in Equation SN8), but note that differences between substitution- and B-del-generated fusion are entirely analogous to those between substitution- and insertion-generated fission. The only notable exception is that we found no evidence that the rate of deletions between repeats (namely, deletions of single-nucleotide or single-unit interruptions) harbor any dependence on the length of the repeat. Additionally, empirical estimates of *de novo* rates from trio data indicate that the per-target rate of B-dels (roughly 2×10^{-10} per generation) is suppressed relative to the per-target rate for μ substitutions (roughly 4×10^{-9} per generation; see **Methods**) by more than an order of magnitude, suggesting B-dels provide an entirely negligible correction to substitution-based fusion rates. Henceforth, all discussion of fusion is focused on substitution-generated events.

Fusion can only occur due to mutations at B sites immediately adjacent to two A sites and thus occur at a rate proportional to the fraction B strings of single-unit length p_F (see Equation SN8 and subsequent discussion below Equation SN14). In the absence of insertions, this probability is given by $p_F = \mu/(\mu + \nu)$

(expansions and contractions of A repeats do not alter this rate). Insertion could in principle alter this rate, but, under the assumption that non-motif insertions are restricted to single-unit lengths, the inclusion of insertions results in a (slightly) greater influx into the $L = 1$ class of B strings. This amounts only to an additional source at the low length boundary; while this explicitly changes the total number of B strings, the normalized probability distribution remains geometric (i.e., Equation SN14) with a slightly modified rate constant. We confirmed via our computational model that this geometric distribution is nearly identical to that under two-way substitution alone, due to the overwhelming mass in the length one class and a negligible influx due to the low rate of insertions from longer repeats. However, we note that the computational model is an abstraction that assumes B-strings are not susceptible to repeat instability. In practice, the value of p_F could be better estimated from the empirical distribution of B strings, but this was unnecessary in the current setting, as any such estimate does not impact our analysis or results.

We proceed by taking the continuum limit of Equation SN8 after removing local contributions (i.e., those proportional to $(1 - p_F)$).

$$\begin{aligned}(\Delta_\mu)_{fuse}\mathrm{P}_L &= -2\mu p_F \mathrm{P}_L + \mu p_F \sum_{k=1}^{L-2} P_k P_{L-1-k} \\ &\approx -2\mu p_F\ \rho(L) + \mu p_F \int_{\Delta L}^{L-2\Delta L} \rho(\lambda)\rho(L - \Delta L - \lambda) d\lambda \\ &\approx -2\mu p_F \rho(L) + \mu p_F \int_0^L \rho(\lambda)\rho(L-\lambda) d\lambda \end{aligned} \tag{SN31}$$

As written, the sum contains an implicit factor of two associated with swapping the subscripts $k \leftrightarrow L - 1 - k$ (i.e., double counting when summing up to $k = L - 2$), which propagates to the integral (i.e., $\lambda \leftrightarrow L - \lambda$ when integrating to L). This contribution to changes in repeat length is explicitly nonlocal due to the quadratic, integral dependence on the distribution, which describes randomly sampling shorter repeats of the appropriate length. In the second line, we have again taken the large L asymptotic limit $L \gg \Delta L = 1$ to suppress subdominant terms. In this form, it becomes clear that the integral is simply a convolution of the distribution $\rho(L)$ with itself over a finite length window; this can be readily interpreted as the distribution of the sum of two random lengths (notably, $< L$) both drawn from the same probability distribution $\rho(L)$ (albeit not necessarily normalized appropriately due to non-conservative transitions). Noting that the p_F is strictly less than one and that empirical estimates show $\mu < \nu$, the rate of outflux due to fusion is strictly less than the outflux of fission due to substitution alone; this contribution becomes negligible in the long length regime due to the finite target size for fusions (i.e., $2\mu p_F \ll \nu L$). However, the integral terms in each cannot be analogously compared due to their non-overlapping limits and distinct functional form. Crucially, through exploration of the parameter space via our computational model, we observed that the relative contribution of fusion to the dynamics in the long length regime is generically subdominant to the rate of other transitions for all parameter combinations. This suggests that fusion may be an infrequent event leading to subdominant corrections to the dominant dynamics.

4.4 Steady-state condition for long repeat dynamics

Collecting the continuum approximations to each term of Equation SN9, the continuous-time dynamics of long repeats in the asymptotic large L regime can be described by the following partial differential equation (PDE).

$$\begin{aligned}\partial_t \rho(L) \approx &\left\{ \frac{1}{2}\partial_L^2 \left[\left((\epsilon_L + \kappa_L)L + 2\iota_L + 2(\mu(1-p_F) + \nu) \right) \rho(L) \right] - \partial_L \left[\left((\epsilon_L - \kappa_L)L - 2\iota_L + 2(\mu(1-p_F) - \nu) \right) \rho(L) \right] \right\}_{local} \\ &+ \left\{ -(\nu + \iota_L) L\rho(L) + \int_L^\infty 2(\nu + \iota_\lambda)\rho(\lambda) d\lambda \right\}_{fiss} + \left\{ -2\mu p_F \rho(L) + \mu p_F \int_0^L \rho(\lambda)\rho(L-\lambda) d\lambda \right\}_{fuse} \end{aligned} \tag{SN32}$$

The above expression is again left in terms of general length-dependent rates of expansion, contraction, and insertion and appropriately labeled to differentiate local, fission-, and fusion-based contributions. For completeness, we have reintroduced local contributions from insertion and substitution (these subdominant effects will be dropped again shortly). Setting the time derivative to zero, we find an integral ordinary differential equation (ODE) for the steady state distribution.

$$\partial_t \rho_{ss}(L) = 0 \tag{SN33}$$

Henceforth, the subscript ss (indicating steady state) will be dropped and assumed throughout. We now approximate this condition to focus on the dominant terms driving the distribution at large lengths. Again, substitutions and insertions are subdominant in the local terms due to a length-independent target size. Additionally, $\nu L \gg 2\mu p_F$ such that the fusion outflux remains subdominant at long lengths. These approximations yield a slightly simpler steady state condition in the asymptotic $L \gg 1$ regime.

$$\begin{aligned}\partial_t \rho(L) \approx \frac{1}{2}\partial_L^2 \left[(\epsilon_L + \kappa_L) L\rho(L)\right] - \partial_L \left[(\epsilon_L - \kappa_L) L\rho(L)\right] \\ - (\nu + \iota_L) L\rho(L) + \int_L^\infty 2(\nu + \iota_\lambda)\rho(\lambda)d\lambda + \mu p_F \int_0^L \rho(\lambda)\rho(L-\lambda)d\lambda \approx 0\end{aligned} \tag{SN34}$$

Expressing this in terms of our power law parameterization, we find the following.

$$\begin{aligned}\frac{1}{2}\partial_L^2 \left[(C_\epsilon L^{\tau_\epsilon} + C_\kappa L^{\tau_\kappa}) L\rho(L)\right] - \partial_L \left[(C_\epsilon L^{\tau_\epsilon} - C_\kappa L^{\tau_\kappa}) L\rho(L)\right] - (\nu + C_\iota L^{\tau_\epsilon}) L\rho(L) + \int_L^\infty 2(\nu + C_\iota \lambda^{\tau_\epsilon})\rho(\lambda)d\lambda \\ + \mu p_F \int_0^L \rho(\lambda)\rho(L-\lambda)d\lambda \approx 0\end{aligned} \tag{SN35}$$

Importantly, no generic closed-form solution to this ODE can be found. First, the complications introduced by fusion are significant, as integration must be performed over the short repeat length regime, limiting our ability to decouple the long length asymptotic dynamics. Second, even when omitting fusion entirely, the remaining terms describe a second-order integral ODE, which can be recast as a third order ODE to remove explicit nonlocal transitions; unfortunately, few third order ODEs are exactly soluble.

The functional form of the fusion term fundamentally limited our ability to further disentangle the dynamics. Motivated by our computational results, we proceeded under the *ansatz* that, for long lengths $L \gg 1$, the fusion term is everywhere negligible relative to fission and local transitions, the latter providing the largest contributions to the asymptotic dynamics due to length scaling. The validity of this assumption, which we confirmed by exploring the allowed parameter space with our computational model, is likely a consequence of both the low rate associated with $\mu p_F \ll \nu L$ and the rapid geometric decay in the short length regime followed by the further (though sub-geometric) monotonic decay at longer lengths. In contrast, Equation SN15 demonstrates the importance of fusion to the short length dynamics, which exactly balance length decreases due to ν substitutions. Exploring a wide range of parameters, we found non-negligible contributions from fusion only in the substitution-dominated short length regime, consistent with our analytic results, and in cases where the distribution is far from steady state. A more principled argument for the relative suppression of fusion at long lengths likely exists, but is unnecessary for the present purposes. For further analysis, we proceed under the following approximation of Equation SN35 for $L \gg 1$, which retains nonlocal transitions only in the form of repeat fission.

$$\frac{1}{2}\partial_L^2 \left[(\epsilon_L + \kappa_L) L\rho(L)\right] - \partial_L \left[(\epsilon_L - \kappa_L) L\rho(L)\right] - (\nu + \iota_L) L\rho(L) + \int_L^\infty 2(\nu + \iota_\lambda)\rho(\lambda)d\lambda \approx 0 \tag{SN36}$$

Expressed in terms of our parameterization, steady state is maintained under the following approximation.

$$\begin{aligned}\frac{1}{2}\partial_L^2 \left[(C_\epsilon L^{\tau_\epsilon} + C_\kappa L^{\tau_\kappa}) L\rho(L)\right] - \partial_L \left[(C_\epsilon L^{\tau_\epsilon} - C_\kappa L^{\tau_\kappa}) L\rho(L)\right] \\ - (\nu + C_\iota L^{\tau_\epsilon}) L\rho(L) + \int_L^\infty 2(\nu + C_\iota \lambda^{\tau_\epsilon})\rho(\lambda)d\lambda \approx 0\end{aligned} \tag{SN37}$$

The above expression describes four distinct effects that collectively lead to steady state, parameters permitting: bidirectional diffusion due to net repeat instability from the combined effects of expansion and contraction, expansion-contraction bias generating a directional flux, net (substitution- and/or insertion-based) outflux due to repeat fission, and net nonlocal (substitution- and/or insertion-based) influx due to fission of any longer repeats. The outflux due to fission is, strictly speaking, a local term in the equation (despite representing non-local transitions, appropriately accounted for by the nonlocal influx) that corresponds to the rate of interruptions of repeats in the focal length class.

4.5 Decomposition of parameter space into dynamical regimes

To better understand the dynamics, we characterized the behavior in qualitatively distinct parameter regimes as primarily controlled by a sum of two, three, or four of the terms in the more complete steady state approximation shown in Equation SN37. These regimes can be identified by studying the length scaling associated with each term, which indicates the primary difference between local and fission-based contributions to changes in length. In contrast to the generality of Equation SN34 (and, assuming subdominant fusion, Equation SN36), the following decomposition is dependent on the details of both our empirically estimated mutation rates and the parameterization in Equation SN1; differing estimates or parameterizations may result in differences in the quantities of importance to the dynamics, but are unlikely to fundamentally alter the subsequent qualitative conclusions about the steady state distribution.

4.5.1 Asymptotic length dependence of local transition rates and $\Delta\tau$

First focusing on local length changes, the quantity $\Delta\tau$ (see Equation SN24) provides an indicator variable for the sign of the bias term at asymptotically large lengths. This can be seen by manipulating the length dependence inside the first derivative term representing the directional (i.e., signed) per repeat rates.

$$(\kappa_L L - \epsilon_L L) = (C_\kappa L^{\tau_\kappa} - C_\epsilon L^{\tau_\epsilon})L = (C_\kappa L^{\Delta\tau} - C_\epsilon)L^{1+\tau_\epsilon} \tag{SN38}$$

The sign of this term determines the asymptotic dominance of either expansion or contraction (i.e., the bias at long lengths) and is dependent only on $\Delta\tau$ (i.e., the constants C_ϵ and C_κ do not scale with length). For clarity, the $\Delta\tau \equiv \tau_\kappa - \tau_\epsilon$ was defined to correspond to the sign of $\kappa_L - \epsilon_L$ such that positive $\Delta\tau$ indicates asymptotic contraction bias.

$$\lim_{L\to\infty} \text{sign}[\kappa_L - \epsilon_L] = \text{sign}[\Delta\tau] \tag{SN39}$$

As described in Equation SN23, $\Delta\tau$ (provided $\Delta\tau \neq 0$) specifies a length L^* at which the sign of $\kappa_L - \epsilon_L$ can reverse if the directional flux changes above $L = 9$ (relative to the initially expansion-biased rates at low lengths, $\epsilon_{L=9} > \kappa_{L=9}$). The (overly simplistic) power law parameterization limits the behavior to, at most, one such sign change in the long length regime: asymptotic contraction-bias ($\Delta\tau \geq 0$) displays one sign reversal, while asymptotic expansion bias ($\Delta\tau \leq 0$) depicts a positive definite (i.e.,, non-vanishing) directional flux throughout the long length regime (due to the empirically estimated expansion bias at $L = 8$). The asymptotic magnitude of the directional flux is determined by the scaling of the faster of the two rates.

$$\lim_{L\to\infty} (\kappa_L - \epsilon_L)\, L \propto \text{sign}[\Delta\tau] m L^{1+\max[\tau_\epsilon,\tau_\kappa]} \tag{SN40}$$

Here, we have included the linear dependence on m common to the constants C_ϵ and C_κ, but focus on the length dependence. This asymptotic dependence defines another useful variable that characterizes the asymptotic magnitude of the local flux $\tau_{\max}$, the maximum between τ_ϵ and τ_κ, which describes the more rapidly growing rate.

$$\tau_{\max} \equiv \max[\tau_\epsilon, \tau_\kappa] \tag{SN41}$$

As can be seen in Equation SN28, $\tau_{\max}$ also dictates the asymptotic dependence of the (strictly positive) diffusion coefficient in the second derivative term.

$$\lim_{L\to\infty}(\epsilon_L + \kappa_L)L \propto mL^{1+\max[\tau_\epsilon,\tau_\kappa]} \tag{SN42}$$

When $\|\Delta\tau\| \gg 1$, either due to large expansion or contraction rates, the asymptotic rates of diffusion and bias increase in magnitude in the same way (provided $L > L^*$ for $\Delta\tau > 0$).

While the quantities $\Delta\tau$ and $\tau_{\max}$ together characterize the asymptotic behavior of both local contributions to the dynamics, their values are related. The sign of $\Delta\tau$ dictates the value of $\tau_{\max}$: for $\Delta\tau < 0$, $\tau_{\max} = \tau_\epsilon$, for $\Delta\tau > 0$, $\tau_{\max} = \tau_\kappa$, and for $\Delta\tau = 0$, $\tau_{\max} = \tau_\epsilon = \tau_\kappa$. Additionally, because the values of τ_ϵ and τ_κ are positive semidefinite, allowed values of $\Delta\tau$ and $\tau_{\max}$ are correlated in the biologically-plausible positive quadrant (i.e., in the space $(\Delta\tau, \tau_{\max}) > (0,0)$). For example, the line defining $\Delta\tau = 1$ requires that $\tau_\kappa \geq 1$, which imposes the same constraint on $\tau_{\max} \geq 1$ (for positive $\Delta\tau$, $\tau_{\max} = \tau_\kappa$). The asymptotic magnitudes of the length-dependent rates of diffusion and directional flux are inherently bounded by the intersections with the axis at $\tau_\epsilon = 0$ or $\tau_\kappa = 0$ for $\Delta\tau > 0$ or $\Delta\tau < 0$, respectively, in the positive quadrant of interest.

$$\tau_{\max} \geq |\Delta\tau| \tag{SN43}$$

In this way, dynamical regimes associated with large $\tau_{\max}$ can be identified and bounded by $\Delta\tau$ values, provided the behavior at the intersection (either $\tau_\epsilon = 0$ or $\tau_\kappa = 0$) corresponds to similar dynamics at larger values of $\tau_{\max}$ (as can be seen in our computational model). Accordingly, subsequent decomposition of the parameter space is phrased in terms of diagonals associated only with constant values of $\Delta\tau$ (i.e., $\tau_\kappa = \tau_\epsilon + const.$), which adequately characterizes asymptotic properties of all local changes in repeat length. As the linear dependence on m is common to all local terms, it is unimportant for distinguishing between these terms; however, the dependence of fission on the substitution rate ν indicates that this parameter may inform comparisons between the rates of local and nonlocal changes in length.

4.5.2 Relative strengths of substitution- and insertion-driven fission and L_{fis}

In contrast to the local dynamics, repeat fission is asymptotically dominated by insertion, which outcompetes the substitution rate at asymptotically long lengths. A characteristic length L_{fis} emerges, above which the fission rate is dominated by insertion-based interruptions. This length can be found by comparing the per-target rates of substitution ν and insertion $C_\iota L^{\tau_\epsilon}$.

$$L_{\text{fis}} \equiv \left(\frac{\nu}{C_\iota}\right)^{1/\tau_\epsilon} = 9\left(\frac{\nu}{m \times \iota_8}\right)^{1/\tau_\epsilon} \tag{SN44}$$

Here, we used the definition of C_ι to show the explicit dependence on m. Our empirical estimate of the ratio ν/ι_8 is of order 50 for mononucleotide A-repeats (see **Figure 2a**); thus, for values of $m \lesssim 50$ (spanning the computationally explored parameter space), the ratio being exponentiated is greater than one. Consequently, as τ_ϵ increases, L_{fis} decreases, eventually exiting the long length regime (i.e., $L_{\text{fis}} < 9$ for $\tau_\epsilon \gg 1$) such that all long length fission events are insertion-dominated. Substitutions dominate more classes in the long length regime when both τ_ϵ and m are small (i.e., for $\tau_\epsilon < 1$ and $m \ll 50$), while substitutions are subdominant at more long lengths when τ_ϵ and m are large. Our computational inferences allow for a wide range of L_{fis} values, from substitutions remaining negligible at all long lengths to substitutions generating effectively all fissions of long repeats. Although intermediate values $m = 4 - 16$ are most consistent with empirical data, our inference showed a wide range of statistically consistent τ_ϵ values that include both $L_{\text{fis}} \sim 10$ (e.g., when $m \sim 16$ and $\tau_\epsilon \gtrsim 2.3$, insertions dominate for $L > L_{\text{fis}} \lesssim 15$) and L_{fis} values beyond the populated length classes (e.g., when $m = 4$ and $\tau_\epsilon < 1$, $L_{\text{fis}} > 100$ and substitutions dominate). This variation in qualitative behavior suggests that our inference has little power to decompose fission into distinct mutational types using available empirical data.

The quantities τ_ϵ and m together characterize the asymptotic length dependence of repeat fission by determining the value of L_{fis}, the length at which a switch occurs from substitution- to insertion-dominated

fission (at lengths $L < L_{\text{fis}}$ and $L > L_{\text{fis}}$, respectively). In the parameter regime where fission is entirely insertion-dominated at long length (i.e., $L_{\text{fis}} \sim 10$ when m, τ_ϵ, or both parameters are sufficiently large), all relevant terms in Equation SN37 are linear in m such that the dynamics are independent of this parameter; in this case, the relative importance of local transitions and fission can be characterized by the values of τ_ϵ and $\Delta\tau$ (or, equivalently, τ_ϵ and τ_κ). In this extreme, in addition to the decoupling between short and long repeat dynamics, long repeats are predominantly subject to insertion- and deletion-based mutations (i.e., repeat instability-generated expansions, contractions, and insertions) and evolve *mechanistically* independently (i.e., change length predominantly due to distinct mutational processes) from short repeats; importantly, this claim implicitly requires that long repeat dynamics are independent of μ (i.e., fusion remains negligible). In the opposite extreme in which insertions remain irrelevant at all populated lengths, the dynamics are independent of τ_ϵ and can be represented in the space of $(\Delta\tau, m)$. Our inference from empirical observations suggests that we cannot simultaneously determine the values of $\{m, \tau_\epsilon, \Delta\tau\}$ (equivalently, $\{m, \tau_\epsilon, \tau_\kappa\}$) and must rely on all three parameters to represent the space of empirically consistent dynamical models.

4.5.3 Distinguishable dynamical regimes

First, the sign of $\Delta\tau$ decomposes the space into two primary regions and the boundary between them, only one of which evolves towards a recognizable steady state equilibrium on reasonable timescales. The repeat length distribution only stabilizes for the subset of the parameter space with $\Delta\tau \leq 0$, which we refer to as the (asymptotically) *contraction-biased* regime. The intuitive explanation for this is that positive directional flux ($\Delta\tau < 0$) leads to indefinite repeat expansion in excess of the net shortening effects of repeat fission (the boundary at $\Delta\tau = 0$ is discussed below). In contrast, asymptotic contraction bias truncates the distribution at finite lengths, leading to a stable distribution, as was observed empirically. The collection of stable parameter combinations can be decomposed into parameter ranges exhibiting distinct dynamics that depend on the extent to which contraction dominates over expansion (i.e., the value of $\Delta\tau$). Within each subregime of $\Delta\tau > 0$, the full set of changes in repeat length in Equation SN32 that lead to a stable distribution can be reduced to a subset of effects that approximate the dominant contributions that shape the distribution; parameter combinations with $\Delta\tau \gg 1$ are controlled largely by a local balance between diffusion and a largely contraction-biased directional flux, both of which dominate over fission; intermediate values $1 \gtrsim \Delta\tau > 0.6$ show a balance between local transitions and length-decreasing outflux due to fission (from substitutions, insertions, or both); weak contraction bias (very roughly $\Delta\tau \lesssim 0.6$) requires a full accounting of local effects and both the influx and outflux due to fission (i.e., the full set of effects included in Equation SN37). At all points in the stable regime, the net increase in repeat length due to expansion (with minor local contributions from μ substitutions) is counteracted by the combined effects of contraction, which dominates over expansion above some length L^*, and insertions and ν substitutions in the repeat body that lead to fission. Based on our computational model, no stable distribution was maintained due to appreciable fusion, consistent with our approximation in Equation SN37. The boundary of this regime occurs near $\Delta\tau \approx 0$; however, very low positive values of $\Delta\tau$ (e.g., roughly $0.3 > \Delta\tau > 0$ when $m = 8$) remain expansion dominated across a large range of populated lengths because the intersection of the rates at $L^* \to \infty$ as $\Delta\tau \to 0$ (see Equation SN23). This results in unrealistic models inconsistent with the empirical distribution, as the asymptotic contraction bias is only apparent for repeats much longer than those observed in appreciable numbers in the human genome.

The remaining set of parameter combinations fall into two categories, both of which are implausible with respect to the collection of empirical observations presented in this manuscript: those with very slowly evolving distributions that do not equilibrate on evolutionarily-relevant timescales and those with unstable dynamics that lead to a rapid aggregation of repeats and consequent explosive growth in the length of the genome. Exceedingly slowly evolving dynamics occur for small values of m as both exponents approach zero (i.e., at values $\tau_\epsilon, \tau_\kappa < 1$ such that τ_ϵ and $\|\Delta\tau\|$ both remain small). These parameter combinations depict repeat instability rates inconsistent with empirical rate estimates (see **Figure 2a**) and, given indefinite time to evolve (in excess of the divergence timescale for primates), are unlikely to result in a stable distribution consistent with those observed across the primate phylogeny. We refer to this region of parameter space as the *slowly evolving* regime, as the repeat instability rates depicted remain scantly above the estimated

substitution rates, even for the longest well-populated lengths in the human genome. Parameter combinations that are dynamically unstable (denoted as the *unstable* regime) universally correspond to expansion rate-dominated dynamics ($\Delta\tau < 0$, excluding slowly evolving parameters with low τ_ϵ). This regime comprises the majority of the parameter space explored in our computational model, which can be dynamically disallowed under the sole observation of steady state, as they represent non-equilibrium dynamical regimes subject to strong nonlinear effects that result in increasingly rapid changes in the repeat length distribution. Under our parameterization, parameters with $\Delta\tau = 0$ (away from the slowly evolving regime) are similarly unstable due to an expansion-bias inherited from the empirically observed bias at $L = 8$ (i.e., $\kappa_9 - \epsilon_9 = m(\kappa_8 - \epsilon_8) > 0$) and subsequent parallel growth of expansion and contraction.

Last, we note that the decomposition of the dynamics is dramatically simplified for large multipliers $m \gtrsim 8$ due to the location of L_{fis}, which sits below the long length regime for sufficiently large m (see Equation SN44). We focus our discussion of the analytics on this case, where the dynamics can be characterized largely by $\Delta\tau$ alone and substitutions are subdominant to insertions for long repeats. However, the same dynamics apply to smaller m, values with the additional complication that the balance between substitutions and insertions breaks the dynamical similarity along lines of constant $\Delta\tau$.

4.6 Unstable dynamics in the asymptotically expansion-biased regime $\Delta\tau \leq 0$

We briefly discuss the dynamics of the unstable regime, as it informs our considerations for more realistic parameter combinations. There are two distinguishable cases where steady state cannot be assumed, contradicting the observation of long term maintenance of the distribution across the primate phylogeny. First, the expansion rate, which is initially dominant at $L = 9$ (i.e., $\epsilon_9 > \kappa_9$), may have a length dependence that rapidly outcompetes that of contraction, resulting in increasingly larger expansion-bias with increasing length. This corresponds to $\Delta\tau < 0$ with large magnitudes $\|\Delta\tau\| \gg 1$. In the second case, which occurs for relatively small values of $\|\Delta\tau\|$ approaching $\Delta\tau = 0$ (i.e., when $\tau_\epsilon = \tau_\kappa$) the length dependence of expansion increases with length comparably to, or slightly in excess of, the length-dependence of the contraction rate. In this case, the directional flux is minimized such that the relative importance of repeat fission becomes inflated. In the former case with highly dominant expansion, the bias generates a large directional flux that rapidly increases repeat lengths. In this regime, the flux is well approximated by the rate $C_\epsilon L^{1+\tau_\epsilon}$ (i.e., $C_\kappa L^{1+\tau_\kappa}$ detracts negligibly at all lengths $L > 10$) and the rate of length increase rapidly accelerates with increasing length. This nonlinearity results in an indefinitely extending tail, some of which feeds back into lower length classes due to fission, as large τ_ϵ also generates large rates of insertion-based repeat fission $C_\iota L^{1+\tau_\epsilon}$; however, the coefficient C_ι is roughly two orders of magnitude smaller than C_ϵ such that fission alone is unable to counteract expansion alone at any length. Repeat fission additionally increases the mass of the distribution, as it is a nonconservative transition: one repeat is replaced with two shorter length repeats. These shorter repeats are then subject to the large directional push due to expansion, which further increases the weight in the distribution tail. This feedback loop generates a rapid, indefinitely growing genome, which reshapes the distribution; rapid fission of any given repeat in the long tail equiprobably adds mass to the distribution in all shorter length classes, an integrated effect that is increased with increasing mass above a given length. This eventually leads to an extreme influx into the shortest length classes, inherently coupling the dynamics in the short and long length regimes (i.e., violating our assumption of separability and distorting the substitution-based geometric distribution). In addition to the unstable dynamics, the indefinite extension of the distribution tail quickly results in highly relevant expansion probabilities approaching one, which simultaneously makes computational modeling impractical across this regime and will invariably result in characteristic changes to the shape of the distribution as the power law parameterization of the rates must give way to saturation at this probabilistic bound.

In the second case, for $\Delta\tau$ near zero (again assuming sufficiently large τ_ϵ to avoid the slowly evolving regime), the directional flux is again expansion dominated and non-vanishing. Here, the relative importance of fission is inflated, which limits the rate at which the tail extends. However, the difference between the expansion and contraction rates, even at $L = 8$, is substantially in excess of the insertion rate. This leads to

the inability of repeat fission to independently counteract the directional flux at all lengths.

$$(C_\epsilon - C_\kappa)L^{1+\tau_\epsilon} \gg C_\iota L^{1+\tau_\epsilon} \tag{SN45}$$

In this sense, contraction must be sufficiently large to mitigate expansion in order for the distribution to truncate at finite length. This only occurs when $\Delta\tau > 0$ such that the contraction rate approaches, and eventually exceeds, the expansion rate (though this may occur at an extremely large length for the smallest values of $\Delta\tau > 0$). This defines a bound for the contraction-biased regime that leads to steady state. A small non-negligible contribution from fission is relevant above this bound, but is insufficient to control the directional push from expansion for very low $\|\Delta\tau\|$. We note that this effort may be aided by substitution-based fission, but the associated length scaling (i.e., the fission rate νL) only becomes relevant for small multipliers m when $\tau_\epsilon \lesssim 1$ (i.e., when $\nu/C_\iota \sim \mathcal{O}(1)$ and $\nu/C_\iota L^{\tau_\epsilon} \sim \mathcal{O}(1)$ across relevant lengths); parameters in this range correspond to sufficiently small rates that evolution proceeds exceedingly slowly, as discussed above.

4.7 Stable dynamics in the asymptotically contraction-biased regime $\Delta\tau > 0$

In contrast to the unstable regime, parameter combinations with asymptotic contraction bias $\Delta\tau > 0$ result in a distribution that approaches a stable steady state. Perhaps more intuitively, this regime can be equivalently characterized by L^*, the length at which expansion and contraction rates cancel such that the directional flux vanishes. For all values $\Delta\tau > 0$, evaluation of Equation SN23 shows that $L^* \geq 9$ (and $L^* \geq 10$ for realistic values of $\Delta\tau$) such that this intersection sits within, or near the boundary of, the long length regime. As the combination of our empirical estimates at $L = 8$ and our parameterization (namely, $m > 0$) dictate that the expansion rate exceeds the contraction rate at length $L = 9$, an intersection between the rates must occur prior to the asymptotic dominance of contraction. As a result, repeat lengths below L^* are expansion-biased and those above L^* are contraction-biased, with a vanishing directional flux at L^*. For large values of $\Delta\tau$, the value of L^* approaches a constant close to 10. For very small values, the location can sit at very large lengths, approaching $L^* \to \infty$ as $\Delta\tau \to 0$. At many values of $\Delta\tau > 0$, this occurs at lengths where the distribution is well-populated, assuming a total target size comparable to the length of the human genome.

The relative rates of expansion and contraction may remain on the same order of magnitude for smaller $\Delta\tau$, as the approach to this intersection is slow from either side; as a result, the first derivative term remains relatively small over a range of lengths that span part or all of the distribution tail, defining an extended length range associated with low directional flux. In this case, the diffusion term, which arises as a subdominant correction to the discrete local behavior (see Section 4.1.2) plays an important role in the extended neighborhood of L^*. This transition between expansion- and contraction-biased length ranges controls and complicates the dynamics. For the smallest values of $\Delta\tau \ll 1$, particularly for large values of $\tau_{\max}$, the intersection at L^* occurs at extreme lengths above those populated in the empirical distribution. This results in the tail extending dramatically due to a wide range of expansion-biased lengths until contraction bias becomes appreciable. Additionally, this generates a dramatically larger genome (with an excess of very long repeats) inconsistent with the observed range of mammalian genome sizes. This is somewhat similar to the unstable dynamics described above, but is eventually counteracted by sufficiently large contraction rates that truncate the distribution and stabilize the (presumably unrealistic) shape of the distribution.

As discussed above, the dynamics of long repeats across the space of parameters with $\Delta\tau > 0$ can be approximated by Equation SN37 under the assumption that repeat fusion is sufficiently infrequent. However, the dynamics reduce further in parts of this regime where fission (influx and, to a lesser extent, outflux) can be treated as negligible. Such approximations eliminate the need to explicitly treat the nonlocal effects of fission influx, which reduces the second-order integro-differential equation to a second-order ODE.

4.7.1 Strong asymptotic contraction bias

We first attempted to describe the dynamics in the regime with sufficiently large $\Delta\tau$ such that the majority of long repeats have contraction-biased rates (i.e., $L^* \sim 10$ occurs immediately above the short repeat regime).

In this case, the diffusion term remains relevant because L^* lies in the long length regime but the directional flux becomes increasingly relevant with increasing length, an effect magnified by necessarily larger $\Delta\tau$ (and consequently larger $\tau_{\max}$). For most contraction-biased parameter combinations, the dynamics of Equation SN37 are well approximated by the following.

$$\frac{1}{2}\partial_L^2\left[(\epsilon_L+\kappa_L)L\rho(L)\right]-\partial_L\left[(\epsilon_L-\kappa_L)L\rho(L)\right]-(\nu+\iota_L)\,L\rho(L)\approx 0 \tag{SN46}$$

Under the power law parameterization, this becomes the following.

$$\frac{1}{2}\partial_L^2\left[(C_\epsilon L^{\tau_\epsilon}+C_\kappa L^{\tau_\kappa})\,L\rho(L)\right]-\partial_L\left[(C_\epsilon L^{\tau_\epsilon}-C_\kappa L^{\tau_\kappa})\,L\rho(L)\right]-(\nu+C_\iota L^{\tau_\epsilon})\,L\rho(L)\approx 0 \tag{SN47}$$

Here, influx due to fission is treated as subdominant, as it is outcompeted by all remaining rates that scale asymptotically with length with an exponent greater than one (i.e., as $L^{1+\tau_\kappa}$ or $L^{1+\tau_\epsilon}$).

4.7.2 Strictly local approximation for strong asymptotic contraction bias $\Delta\tau \gg 1$

In the regime of largest $\Delta\tau$, corresponding to the largest asymptotic scaling $\kappa_L L \pm \epsilon_L L \approx C_\kappa L^{1+\tau_\kappa}$ for the rates of directional flux and diffusion, the outflux due to fission becomes negligible, as well (i.e., it is outcompeted because $\tau_\kappa \gg \tau_\epsilon$). This defines a dynamical sub-regime within the space of contraction-biased parameter values, with dynamics well approximated by the following balance.

$$\frac{1}{2}\partial_L^2\left[(C_\epsilon L^{\tau_\epsilon}+C_\kappa L^{\tau_\kappa})\,L\rho(L)\right]-\partial_L\left[(C_\epsilon L^{\tau_\epsilon}-C_\kappa L^{\tau_\kappa})\,L\rho(L)\right]\approx 0 \tag{SN48}$$

In this regime, sufficiently rapid decay in $\rho(L)$ results in a net negative directional flux (the correct sign associated with net contraction bias) that counteracts the strictly positive diffusion term. For example, when $m = 8$, this expression provides a good approximation to the dynamics primarily for $\Delta\tau > 2$ (equivalently, $\tau_{\max} = \tau_\kappa > 2$) and breaks down as fission becomes increasingly relevant for weaker contraction bias. In this regime, L^* occurs at or adjacent to $L = 10$ such that nearly all long lengths are contraction-biased. For example, the parameter combination $\{m, \tau_\epsilon, \tau_\kappa\} = \{8, 0, 2\}$ results in $L^* \approx 11$ and a contraction rate of roughly double the expansion rate at $L = 16$, while for $\{m, \tau_\epsilon, \tau_\kappa\} = \{8, 0, 4\}$, $L^* \approx 10$ and the contraction rate is roughly tenfold the expansion rate at $L = 16$. In the more extreme case of $\Delta\tau = 4$, we can better understand the truncation of the distribution by further approximating Equation SN48 under the very rough assumption that the asymptotic dependence is immediately relevant for lengths $L > L^*$.

$$\frac{1}{2}\partial_L^2\left[L^{1+\tau_\kappa}\rho(L)\right]\approx -\partial_L\left[L^{1+\tau_\kappa}\rho(L)\right] \tag{SN49}$$

Noting that the contraction rate constant C_κ cancels such that this equation is only dependent on the exponent τ_κ, an approximation for the asymptotic shape of the steady state distribution can be obtained in closed form (valid only for lengths $L \gg L^*$).

$$\lim_{\Delta\tau\gg 1}\rho(L)\approx c_1 L^{-(1+\tau_\kappa)}+c_2 L^{-(1+\tau_\kappa)}e^{-2L} \tag{SN50}$$

To evaluate the accuracy of this rough approximation, the arbitrary constants c_1 and c_2 can be found using values from our computational model at two lengths L_1 and L_2 (where $L_1, L_2 \gg L^*$) to constrain the distribution at $\rho(L_1)$ and $\rho(L_2)$. Comparing this expression to numerical solutions, we found reasonable agreement for the most extreme values of $\Delta\tau > 0$ (e.g., $\Delta\tau = 2-4$ for $m = 8$) at the longest lengths $L > L^*$ in the distribution and an expected departure as $L \to L^*$.

4.8 Intermediate asymptotic contraction bias

Intermediate values of $\Delta\tau > 0$ (e.g., roughly $\Delta\tau \sim 0.8-1.4$ for $m = 8$) require a description of the flux out due to fission, as shown in Equation SN47. At the larger end of this range (e.g., $\Delta\tau \sim 1$), the numerical solution to

Equation SN48, which omits all effects of fission, approximates the asymptotic shape of the distribution for lengths $L > L^*$. This indicates that the impact of outflux due to fission is primarily localized to intermediate lengths $L^* > L > 10$ and therefore most relevant when the rate of expansion exceeds contraction. At the same time, this suggests that the truncation of the distribution is driven by the contraction-biased directional flux, rather than fission alone or requiring the combined effects of fission and contraction.

4.9 Weak asymptotic contraction bias

For smaller values of $\Delta\tau$ that approach $\tau_\epsilon = \tau_\kappa$ (e.g., roughly $\Delta\tau \sim 0.3-0.6$ for $m = 8$), the dynamics revert to Equation SN37. The resulting distributions stabilize, in part, due to a nonlocal influx from fission of longer repeats. For these parameter combinations, L^* sits in the middle of the long length tail; for lengths $L > L^*$, the distribution is well-approximated by solutions to Equation SN47. This is consistent with a nonlocal net flow from lengths $L > L^*$ to lengths $L^* > L > 10$ and indicates that the dynamically relevant effects of fission influx are localized to the latter. The net effects of fission alone (i.e., fission influx minus outflux) result in a net loss of long repeats, with little gain from any longer length repeats that exist, and a compensatory net gain of intermediate length repeats within the distribution tail, in excess of the number lost to the short length regime. With decreasing values of $\Delta\tau \to 0$, L^* approaches very large values, extending the range of lengths receiving this influx. All effects represented in SN37 are thus required to adequately approximate the dynamics that lead to a steady state distribution when $\Delta\tau \ll 1$ (and a potentially broader range of fractional values $\Delta\tau < 1$, depending on m).

The nonlocal integral dependence in Equation SN37 that describes the influx due to fission complicates the steady state condition in this regime. To find solutions, we re-expressed the second-order integro-differential equation as a third order differential equation by applying an overall length derivative to each term.

$$\frac{1}{2}\partial_L^3\left[\left(C_\epsilon L^{\tau_\epsilon} + C_\kappa L^{\tau_\kappa}\right) L\rho(L)\right] - \partial_L^2\left[\left(C_\epsilon L^{\tau_\epsilon} - C_\kappa L^{\tau_\kappa}\right) L\rho(L)\right] - \partial_L\left(\nu + C_\iota L^{\tau_\epsilon}\right) L\rho(L) + \partial_L\left[\int_L^\infty 2\left(\nu + C_\iota \lambda^{\tau_\epsilon}\right)\rho(\lambda)d\lambda\right] \approx 0 \tag{SN51}$$

Taking this derivative allows us to apply the fundamental rule of calculus to replace the derivative the integral with the integrand evaluated at the bounds of integration. Under the assumption that the distribution decays sufficiently rapidly such that $L^{\tau_\epsilon}\rho(L) \to 0$ as $L \to \infty$, the integral term becomes the following.

$$\partial_L\left[\int_L^\infty 2\left(\nu + C_\iota \lambda^{\tau_\epsilon}\right)\rho(\lambda)d\lambda\right] = -2\left(\nu + C_\iota L^{\tau_\epsilon}\right)\rho(L) \tag{SN52}$$

This allows us to re-expresses the second-order integro-differential equation as the following third order ODE.

$$\frac{1}{2}\partial_L^3\left[\left(C_\epsilon L^{\tau_\epsilon} + C_\kappa L^{\tau_\kappa}\right) L\rho(L)\right] - \partial_L^2\left[C_\epsilon L^{\tau_\epsilon} - \left(C_\kappa L^{\tau_\kappa}\right) L\rho(L)\right] - \partial_L\left(\nu + C_\iota L^{\tau_\epsilon}\right) L\rho(L) - 2\left(\nu + C_\iota L^{\tau_\epsilon}\right)\rho(L) \approx 0 \tag{SN53}$$

After applying a length derivative to the steady state condition $\partial_L(\partial_t\rho(L)) = 0$, this now corresponds to a constraint on the flux $\partial_L\rho(L)$. This can be seen by swapping the order of the length and time derivatives (i.e., $\partial_t(\partial_L\rho(L)) = 0$), dictating that the fluxes through each length must sum to a time-independent constant $\phi_L i$. In the special case where $\phi_L = 0$, this corresponds to a steady state condition that maintains the shape of the distribution in equilibrium. We confirmed via our computational model that, once steady state was reached, the net flux through each individual bin independently vanished (see **Figures SN2-SN4**), indicating that the equilibrated state is maintained in a detailed balance. Insofar as our approximations remain valid, Equation SN53 provides a local expression for the steady-state flux through the repeat length distribution in the large length regime; this includes the nonlocal contributions of repeat fission to the flux, represented by boundary effects at length L (more accurately, at length $L + 2 \approx L$). While this equation cannot be solved analytically, numerical solutions can be readily obtained.

Assuming the effects of fusion remain subdominant in the long length regime, Equation SN53 captures the full set of dynamics associated with changes in repeat length. Solutions to this equation, obtained after applying the additional constraint that the fluxes vanish, are applicable across the full range of parameters that evolve towards steady state distributions $\Delta\tau > 0$. In contrast, Equations SN47 and SN48 are approximations to these dynamics appropriate in a subset of parameter space (very roughly, when $\Delta\tau > 1$ and when $\Delta\tau \gg 1$, respectively). However, in addition to aiding in our intuitive understanding of the dynamics, the absence of the third derivative in the latter equations makes numerical solutions more reliable, as they are less susceptible to instabilities in numerical techniques; this facilitates a slightly more reliable comparison between the numerical solutions and results of our computational model.

4.10 Obtaining numerical solutions to the steady state dynamics for $\Delta\tau < 0$

To compare our analytic understanding of the dynamics to the results of our computational model for generic parameters, we resorted to solving Equations SN47, SN48, and SN53 numerically. All numerical solutions were obtained using the NDSolve function in Mathematica 14.0 [1]. Solutions to second-order differential equations require the specification of two additional constraints that together fix the normalization constant and the linear coefficient specifying the relative weight of the two real solutions to the equation, if both exist. The third order equation for constant flux requires a third constraint that ensures that the flux vanishes. For the second-order equations, we chose to constrain the values of $\rho(L)$ at two lengths, L_1 and L_2, using the results of our computational model (i.e., $\rho(L_1) = \rho_{sim}(L_1)$ and $\rho(L_2) = \rho_{sim}(L_2)$, where $\rho_{sim}(L)$ is the value of the computationally propagated (i.e., 'simulated') distribution at length L once it has reached steady state). The choice of L_1 and L_2 is somewhat arbitrary, provided they are both in the long length regime where the continuum approximation is valid and that any additional constraints within the regime of validity for each approximation are respected (e.g., $\Delta\tau > 0$ and sufficiently far from $\Delta\tau = 0$, $\tau_\epsilon, \tau_\kappa \geq 0$, etc.). For the majority of comparisons, we chose two lengths that are well defined for any parameter combination: $L_1 = L^*$ (rounded to the nearest integer value) and $L_2 = L_{\max}$, where $L_{\max}$ is the length bin for which the occupancy of the non-normalized distribution first drops below a single count (i.e., $L_{\max} = \min[L \text{ for } \rho(L) < 1]$) and represents the truncation point of the distribution. $L_{\max}$ is uniquely defined because all computationally modeled distributions decay monotonically and stochastic effects were omitted to obtain the expected distribution under a mean field approximation. All comparisons were made using the non-normalized distributions to identify the truncation point of the distribution and obtain $L_{\max}$. To provide the third constraint needed for Equation SN53, we chose to use $L_3 = L_2 - 1$ for convenience; we note that inappropriate choice of L_3 outside of the regime of validity of the approximation can result in numerical instability in solutions to the third order ODE (e.g., when constraining the solution using a length class $\rho_{sim}(L_3)$ that has not yet equilibrated). We chose not to use the intuitive lower bound of the long length regime at $L = 10$ to better identify any potential effects associated with the breakdown of the continuum approximation. For any comparisons with $L^* < 10$ (which occur only for unstable parameter combinations with $\Delta\tau < 0$) we chose a lower bound at $L_1 = 10$ to avoid values in the short length regime; however, unstable dynamics were compared to numerical solutions primarily to confirm significant departure from any steady state characterized by Equation SN53 and to identify any common features that emerged. Finally, for all comparisons in the long length regime, the value of ν was replaced with $\nu_{fission}$, the appropriate rate estimated from the three-unit context $AAA \rightarrow ABA$ in which substitutions $A \rightarrow B$ result in repeat fission. Mutation rates μ and ν that appear in Equation SN10 were replaced with distinctly estimated rates relevant for the three-unit contexts associated with local transitions due to substitutions defined by μ_{local} (i.e., the summed rates of $ABB \rightarrow AAB$ and $BBA \rightarrow BAA$ substitutions) and ν_{local} (the summed rates of $AAB \rightarrow ABB$ and $BAA \rightarrow BBA$ substitutions), respectively.

5 Comparison between numerical solutions and computationally modeled distributions

As described above, numerical solutions to Equations SN47, SN48, and SN53 were compared across the space of parameter combinations that led to steady state distributions. Because our computational model obtained results over a large, but finite number of iterations, parameters corresponding to insufficiently high mutation rates failed to equilibrate in the allotted time (i.e, in a number of iterations corresponding to at least 10^9 generations of evolution, after accounting for a factor that progressively rescales time to increase computational speed). These slowly evolving computational results were localized to the lowest values of $\{\tau_\epsilon, \tau_\kappa\}$ for each multiplier m (i.e., points closest to the origin of the $\{\tau_\epsilon, \tau_\kappa\}$ plane) and spanned a larger range of parameter values for smaller m. For $m = 8$, this roughly corresponds to parameter values $\tau_\epsilon, \tau_\kappa \lesssim 1$; within this region, clines of roughly equivalent metric values (calculated by comparing computationally modeled distributions at the final time point to the empirical distribution) begin to deviate from lines of constant $\Delta\tau$. This corresponds to the point at which substitution becomes non-negligible and substitution-based fission occurs at a rate comparable to or greater than insertion-based fission. Given indefinite time to evolve, such parameter combinations would no doubt equilibrate, provided the combined action of contraction, substitution, and insertion is sufficient to truncate the distribution at finite length. As no equilibrium was reached, these points were excluded from our comparison to numerically produced steady-state distributions.

For all points outside of this slowly evolving region, equilibrated steady state distributions are quantitatively similar along lines of constant $\Delta\tau$. The following plots show comparisons at points along an anti-diagonal line perpendicular to $\Delta\tau = 0$ defined by $\tau_\epsilon + \tau_\kappa = 3$. Parameter combinations on this line are representative of the full set of computationally modeled $\Delta\tau$ outside of the slowly evolving region and could be obtained for most multipliers. Along the line of $\tau_\epsilon + \tau_\kappa = 3$, we selected examples that span the qualitatively distinct behaviors across the space of $\Delta\tau > 0$ for values $\Delta\tau = \{0, 0.2, 0.4, 0.6, 0.8, 1, 1.4, 2, 3\}$. The first two points were included for completeness, as they show examples of computationally modeled distributions that are not expected to be well described by numerical solutions. For $\Delta\tau = 0.2$ the location of $L^* \approx 92$, which is close to the maximum length included in our computational model at $L_{boundary} = 200$; a reflective boundary condition was imposed at $L = L_{boundary}$ to simultaneously prevent excessively mutation rates that preclude further computational iteration and to identify parameter combinations that result in excessively large genome size when truncation occurs at unrealistically large lengths $L_{max} > 200$. The point at $\Delta\tau = 0$ was included as an example of unstable dynamics; the computationally modeled distribution equilibrates to the boundary condition at $L_{boundary}$ (note the balanced fluxes in the largest lengths for computational results with $\Delta\tau = 0$ in Figures SN2I and SN3I), resulting in an artifactual shape (i.e., the non-monotonicity in computationally modeled curves shown in Figures SN1I, SN4I-SN6I). Interestingly, the asymptotic tail of the distribution is still somewhat well described by numerical solutions to Equation SN53, provided the constraint $L_2 = L_{boundary}$ is applied such that the boundary condition effectively acts as an artifactual source at $L_{boundary}$ (i.e., there is agreement between numerical and computationally modeled distributions for $\Delta\tau = 0$ in Figures SN1I, SN4I-SN6I, but only at long lengths that have equilibrated to the boundary). At longer timescales, the non-equilibrium behavior may overwhelm this artifactual effect.

5.0.1 Comparisons of approximations to the dynamics and steady-state distributions

Figure SN1 compares computationally modeled distributions for these example parameter combinations to the geometric distribution that describes the short length regime (Equation SN10 using values the average single-unit context rates $\mu = \mu_{B\to A}$ and $\nu = \mu_{A\to B}$) and to the three nested approximations of the long length tail of the distribution obtained by numerically solving Equation SN47, SN48, and SN53. These plots contrast numerical solutions produced under the full set of dynamics (excluding fusion), in the absence of influx due to fission, and in the absence of fission entirely. By observing the lengths at which each successive approximation breaks down, we localized the incoming and outgoing flux due to fission in length space.

For the same parameter combinations, Figure SN2 shows the total influx and outflux for each length associated with the individual effects of expansion, contraction, substitution-based fission, and insertion-

based fission (as well as fusion and and local transitions due to substitutions). The total flux was separately normalized for each bin (for visualization purposes, as the true magnitudes differ dramatically); bins with equal incoming and outgoing net flux have reached equilibrium (i.e., are maintained in a detailed balance). Slight deviation from this equilibrium occurred in every run, indicating a true steady state distribution was not yet reached (as expected in finite time). However, for all results of interest, the magnitude of deviation is very small. Some parameters with modest values of $\Delta\tau$ showed slight deviation from equilibrium at L=1 characteristic of disequilibrium between the A and B distributions (i.e., a source of new A counts at the $L = 1$ boundary) but showed detailed balance across all remaining classes. This is in contrast to unstable parameter combinations, which harbor large deviations from equilibrium in many or most length classes; at the same time, non-equilibrium simulations rapidly populated the largest length classes, subsequently equilibrating to the artificially imposed boundary condition at the large L boundary of the computationally modeled grid.

Figure SN3 provides an alternative characterization of the flux for each bin: the incoming and outgoing fluxes for each mutational effect were plotted separately, rather than computing their net (substitutions were separated into local and nonlocal contributions to demonstrate subdominance of the former relative to expansion and contraction). When separating into directional fluxes, the dominance of expansion and contraction over all other fluxes in both directions is made clear; the much higher rates of repeat instability dominate local fluxes throughout the long length regime. This also leads to large-scale diffusion, which captures the significant bidirectional flux that occurs for both expansions and contractions.

Figure SN4 shows a direct comparison between the computationally modeled sum of all fluxes (including fusion) at each length for comparison to three approximations of the full dynamics: local dynamics alone, local dynamics and fission outflux (treating influx as negligible), and local dynamics along with a full model of fission (all transitions other than fusion). The accuracy of each approximation can be seen at each length bin via the overlap with the net flux iunder the full model. In contrast to Figure SN1, which compares numerical solutions that approximate the steady-state distributions, these plots directly compare components of the finite difference equation (and, in a continuum approximation, the differential equations in steady state) for each model specified by Equations SN37, SN47, and SN48. In particular, the nonlocal interactions in Equation SN37 are accounted for directly, without requiring the intermediate step that leads to Equation SN53. This provides a complementary set of comparisons that lead to the same qualitative observations about the regime of validity and accuracy of each approximation across the parameter space, while retaining length-dependent information about the role of each effect across the long repeat regime. Additionally, this provides further justification for the assumption that repeat fusion remains negligible for long repeat dynamics, despite it's qualitative importance to short repeat dynamics.

Figures SN5, SN6, and SN7 show comparisons between numerical and computational results for the same values of τ_ϵ, τ_κ, and $\Delta\tau$ shown in Figure SN1, but with multipliers of $m = \{2, 16, 32\}$, respectively. For intuition about the effect of the dynamics on the genome size, all comparisons below are shown for non-normalized distributions. The total genome-wide target for A bases, $\bar{L}_A$, corresponds to the weighted mean of the non-normalized distribution: $\bar{L}_\mathrm{A} = \sum_{L=1}^{\infty} L\ P(L) \approx \int_0^\infty dL\ L\ \rho(L)$. The total genome size is the sum of $\bar{L}_A$ and the corresponding target for B bases $\bar{L}_B$. Comparing the same values of $\Delta\tau$ across multipliers m, we found qualitative consistency of the decomposition into three dynamical regimes but with boundaries that quantitatively depend on m. The approximations represented by numerical solutions to Equation SN47 (along with the rough analytic solution) and Equation SN48 break down at a larger values of $\Delta\tau$ for smaller m due to the added contribution of substitutions to the fission rates. For example, in Figure SN5 for $m = 2$, the purely local approximation in Equation SN47 breaks down at or above $\Delta\tau = 2$, rather than at roughly $\Delta\tau \sim 1.5$ for $m = 8$ shown in Figure SN1. Here, the relative rate of substitution is closer to the insertion rate and the total rate of fission is closer to the expansion and contraction rates. The relative increase in the strength of fission implies that fission plays a more substantial role in the maintenance of the steady state distribution. In contrast to the more aggressive approximations, numerical solutions to Equation SN53 (i.e., the continuum model with all effects except fusion) remain accurate across the $\tau_\epsilon + \tau_\kappa = 3$ line, even at low values of m.

Figure SN8 shows comparisons between the computationally modeled distribution, numerical solutions,

and the closed form approximation in Equation SN50 for parameter combination within and outside of the regime of validity of the latter (i.e., only appropriate when $\Delta\tau \gg 1$). Parameter combinations are shown for $\Delta\tau = \{1, 3, 5\}$ and $m = \{2, 8, 32\}$. This rough approximation captures the asymptotic falloff of the distribution when $\Delta\tau$ is sufficiently large (roughly $\Delta\tau \gtrsim 3$), failing to characterize the shape at lengths closest to the lower boundary of the long length regime around $L = 10$. Equation SN50 shows that the dependence on the constants C_ϵ and C_κ in Equation SN47 approximately cancels, yielding an m-independent expression (i.e., the long length tail of the distribution becomes insensitive to the multiplier). Equation SN10 is also independent of m below $L = 9$ because substitution dominates the dynamics. The shape of the distribution is therefore largely independent of m when $\Delta\tau \gg 1$, except at the transition between the short and long length asymptotic behavior, which is localized to a tight range of lengths; this insensitivity to m can be seen for larger $\Delta\tau$ values in Figure SN8. As m qualitatively captures the relative strength of the substitution and repeat instability rates, the dynamics of long length repeats decouples nearly perfectly from the short length regime when $\Delta\tau \gg 1$, effectively only interacting via a local flux across the length regime boundary at $L = 10$.

6 Empirical constraints on parametric model

Using our computational model, we compared distributions at the final time point (i.e., the steady state distribution for stable parameters) to the empirical distribution of A mononucleotide repeats using the metric defined in Equation 1 in **Methods**. Using the bootstrap procedure defined in **Methods**, we found the subset of parameters statistically consistent with the parameter combination that produced the minimum metric value (i.e., the *best fit* parameters). The best fit parameter values at the metric minimum are $\{m, \tau_\epsilon, \tau_\kappa\} = \{2, 1.5, 1.8\}$. Under our parameterization, a flat direction emerged in metric space roughly along lines of constant $\Delta\tau$ (best fit value $\Delta\tau = 0.3$), consistent with our understanding of the parameter space, detailed above. **Figure 2b** shows the metric values across slices of constant m; we believe that this fully explores the range of parameters consistent with all mathematical and biological constraints (e.g., monotonic rate increases for expansion, contraction, and non-motif insertion rates due to repeat instability, maintenance of the linear mutation regime such that transition probabilities at all lengths do not approach one, etc.). Here, we summarize these results interpreted in the context of the analytic model for long repeat dynamics presented in this supplementary note.

6.1 Constraints on $\Delta\tau$ and L^*

The subset of parameter values consistent with the metric minimum are poorly localized in τ_ϵ and τ_κ, but highly localized along a cline in $\Delta\tau$ (henceforth, *minimum cline*). For large multipliers (roughly $m \geq 8$), the minimum cline follows the off-diagonal line of constant $\Delta\tau = 0.4$ but includes a subset of parameters with $\Delta\tau = 0.3$ and 0.5. For smaller multipliers, the slowly evolving regime (i.e., roughly $\tau_\epsilon, \tau_\kappa \lesssim 1$) includes stable values at $\Delta\tau = 0.2$ (and a varying number of slowly evolving parameters with $\Delta\tau < 0.2$ that have not equilibrated in the allotted time or will slowly diverge); however, the steady state distribution in this regime is highly dependent on substitution-based fission, with a corresponding equilibration timescale greater than the inverse of the substitution rate (approximately $10^9 - 10^{10}$ generations), which is likely inconsistent with the rate of evolution of the distribution. Additionally, the cline deviating from constant $\Delta\tau$ includes values that do not equilibrate within the computationally modeled timescale (in excess of 10^9 generations). This suggests that the realistic and sufficiently equilibrated subset of inferred values to spans $\Delta\tau = 0.3 - 0.5$. Using Equation SN23, we find a range of values $L^* \approx 23 - 42$ (inclusion of low m values with $\Delta\tau = 0.2$ gives an upper value of $L^* = 92$). L^* localizes the transition point between expansion-biased (below) and contraction-biased length ranges and represents a dynamical shift from a net length-increasing to net length-decreasing flux due to the combined effects of expansion and contraction. We find that this occurs at intermediate lengths within the long length regime, below the truncation point of the distribution, which we estimate at roughly $L_{\max} = 65$. Notably, within the discrete grid of parameter values explored, values of $\Delta\tau < 0.3$ correspond to $L^* > 90 > L_{\max}$, significantly beyond the empirically observed truncation point of the distribution (i.e., the

value of $L_{\max}$ for the genome-wide distribution of A repeats). Indeed, distributions within the latter regime extend to near or beyond the boundary of the computational grid at $L_{boundary} = 200$; in contrast, we observe few populated length classes above roughly $L_{\max} \sim 65$ in the human genome.

6.2 Constraints on L_{fis}

As our inference showed statistically consistent metric values along a cline of $\Delta\tau$, we were unable to isolate the relative contribution of substitutions and insertions to repeat fission using the inference dataset alone. In other words, L_{fis} (see Equation SN44) could not be estimated from the metric cline, as consistent values of τ_ϵ span the whole parameter space (i.e., corresponding estimates of L_{fis} range from order ten to infinity). However, we analyzed a distinct dataset using popSTR [2] by linearly regressing the expansion and contraction rate power laws to estimate a range of consistent values of $\{m, \tau_\epsilon, \tau_\kappa\}$ as a validation step (see **Methods**; **Figure S6b**). This produced a small subset of parameters along the minimum cline (parameters within boxes in **Figure 2b**) and limited allowed values of the multiplier to roughly $m \sim 8 - 16$ and $\tau_\epsilon \sim 1.6 - 3.1$ (overlapping parameters follows the cline of $\Delta\tau = 0.3$ such that $\tau_\kappa \sim 1.9 - 3.4$). The overlapping region of our inference and popSTR-estimated parameters allowed for estimation of a range of consistent values of $L_{\text{fis}} \approx 12 - 26$.

These estimated values are reasonably well localized, suggesting that fission of repeats shorter than L_{fis} (in a range of roughly $10 - 25$) are primarily a result of substitution-based interruptions. At long lengths $L > L_{\text{fis}}$, fissions largely result from insertions. The inclusion of values closer to $L_{\text{fis}} \sim 10$ suggest the possibility that nearly the entire long length regime evolves independently of substitutions; the upper estimate of $L_{\text{fis}} \sim 25$ still describes substitution-independent asymptotic dynamics for the majority of populated long length classes in the genome. Above L_{fis}, the dynamics are entirely independent of substitution, controlled only by mutational effects categorized as repeat instability. This suggests that, in addition to the separability of long repeat dynamics from the substitution-driven short repeat regime, most long repeats are also *mechanistically independent* from short repeats; the primary mutational mechanisms that alter repeat length may be categorically different from replication and repair pathways that generate substitutions in shorter sequences. In this sense, two distinct boundaries can be placed on the lengths of repetitive sequences, corresponding to up to three distinct regimes: repeats below roughly 10 nucleotides are primarily subject to random substitutions; repeats below L_{fis} (where $25 \gtrsim L_{\text{fis}} \gtrsim 10$), which may experience expansion and contraction, are subject to substitution-based interruptions; repeats above L_{fis} primarily exhibit repeat instability-based length changes, evolving both dynamically and mechanistically independently from shorter-length repeats subject to substitution-based effects.

Supplementary References

1. Wolfram Research, Inc. *Mathematica, Version 14.0,* Champaign, IL, (2024).
2. Kristmundsdottir, S. *et al.* Sequence variants affecting the genome-wide rate of germline microsatellite mutations. en. *Nat Commun* **14,** 3855 (2023).

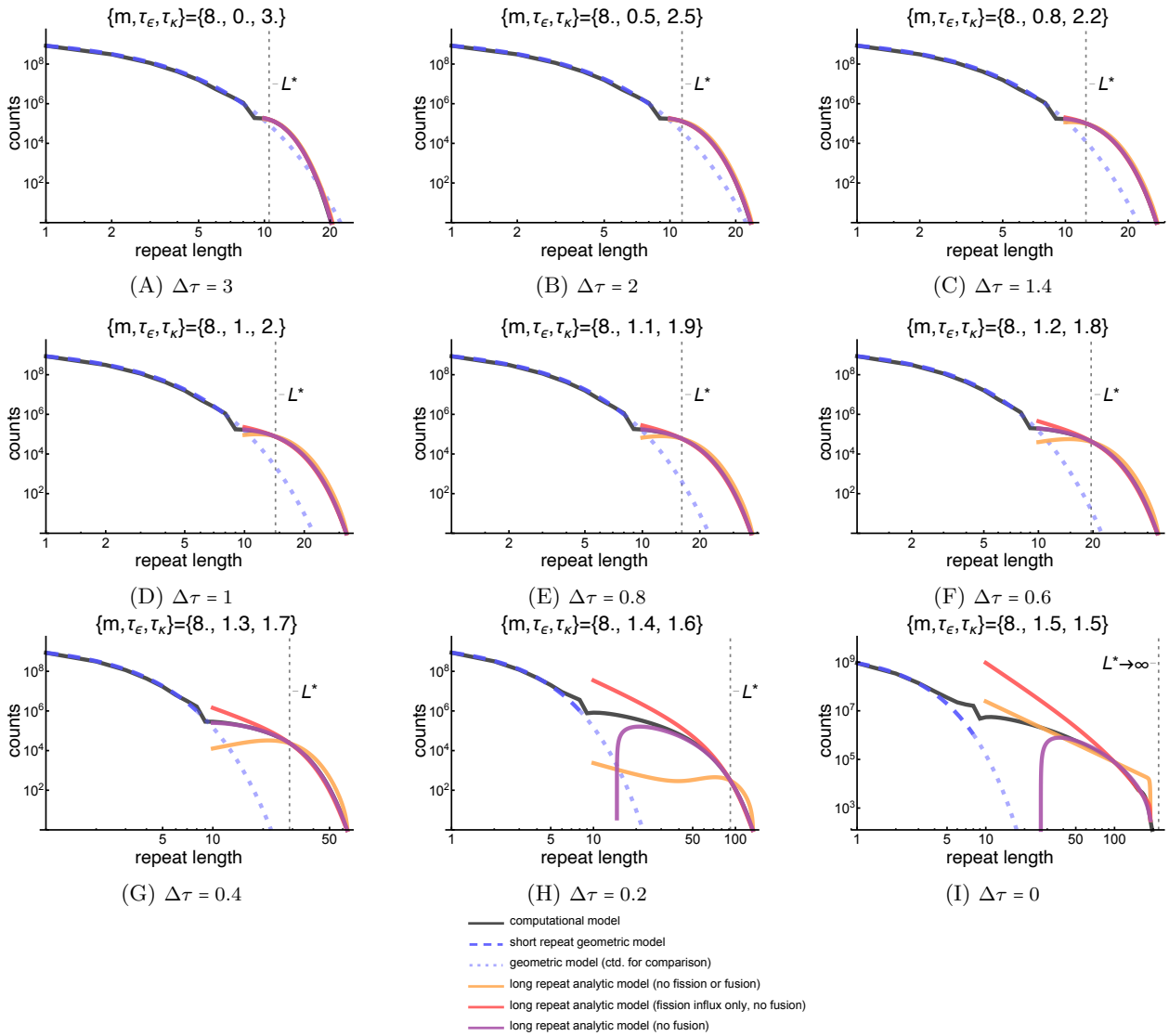
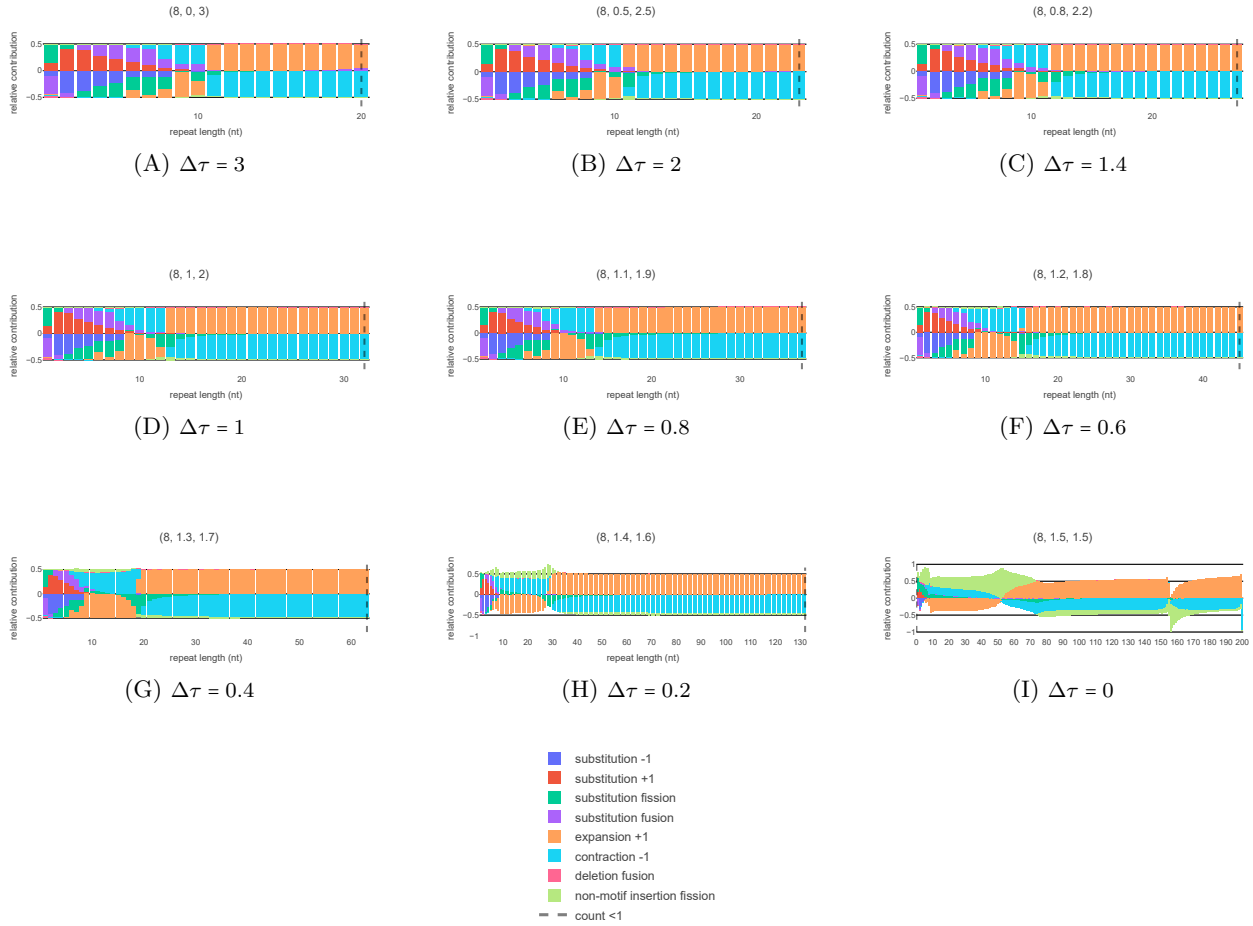


Figure SN1: *Comparison between computational model results and numerical solutions to steady state equations for* $m = 8$ *for parameter values with of constant* $\tau_\epsilon + \tau_\kappa = 3$. Each inset shows plots of the computationally modeled distribution at the final time point (black), geometric approximation for shorter repeats of length $L < 10$ (blue, continued as blue dashed line for comparison to distribution tail shape), numerical solutions to Equation SN48 with no fission (orange), numerical solutions to Equation SN47 with fission out but no fission in (red), and numerical solutions to Equation SN53 with fission (purple). (A-C) Comparisons for parameter combinations with $\Delta\tau \gtrsim 1.5$ (referred to as $\Delta\tau \gg 1$) show good agreement for all numerical solutions; fission is negligible. (D) Boundary between large $\Delta\tau$ and intermediate values where fission out first becomes relevant. (E) Fission out is relevant for $\Delta\tau \sim 1$, but only for repeats $L < L^*$ (i.e., deviation of orange line at $L \sim 10 - 13$). (F, G) Effects of fission in are relevant for $L^* > L \gtrsim 10$, while longer lengths are well approximated by considering only fission out and local dynamics; approximation with local dynamics alone remains reasonable for determining the location of L_{max}. For $m = 8$, $\Delta\tau = 0.4$ is the most consistent with the empirical distribution. (H) L^* lies close to the computational grid boundary at $L = 200$. This distribution would stabilize by truncating at a length above $L = 200$ if the grid was extended far beyond realistic lengths. (I) Unstable dynamical regime subject to nonlinear growth. The distribution shows clear interaction with the reflecting boundary.

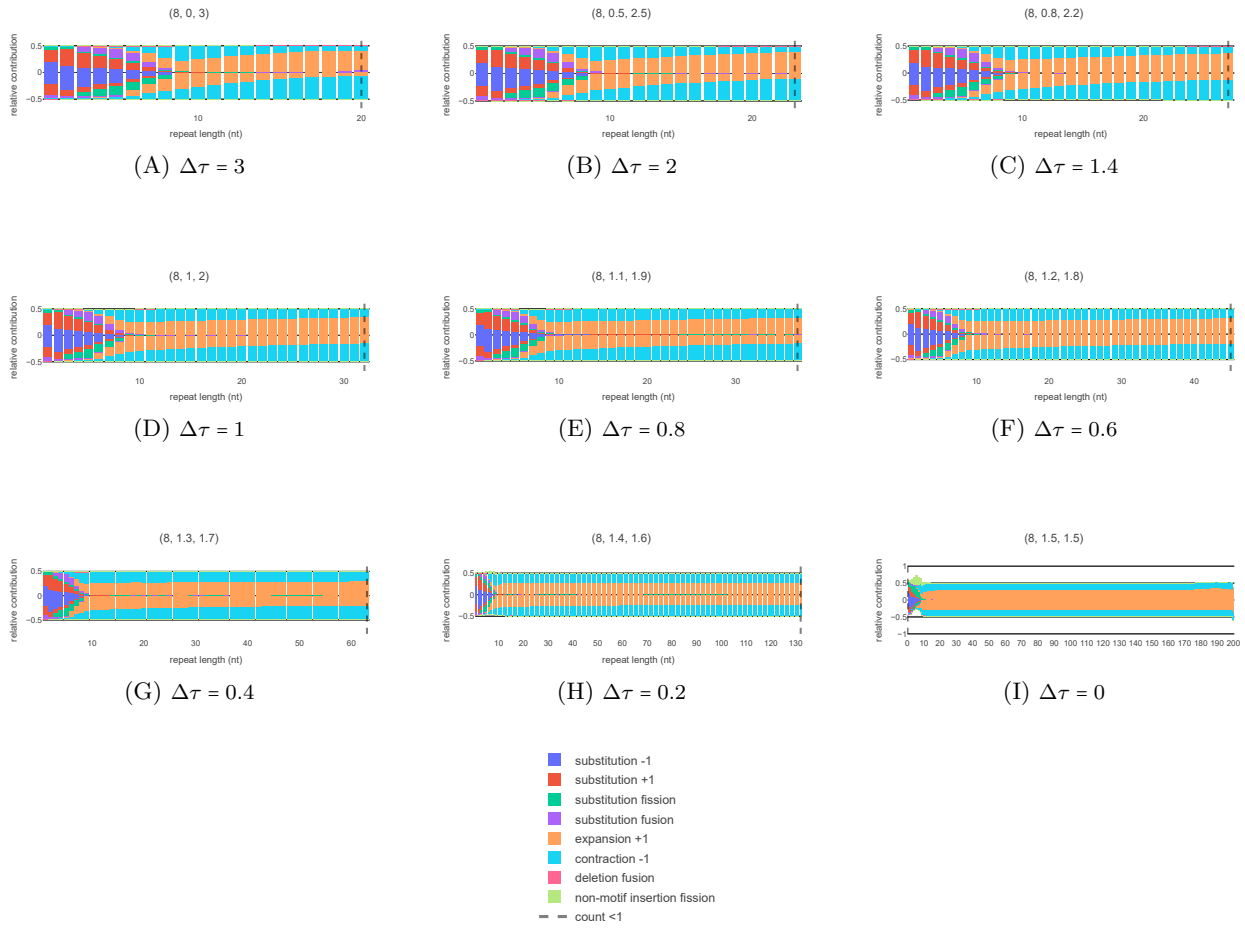


(A) $\Delta\tau = 3$ (B) $\Delta\tau = 2$ (C) $\Delta\tau = 1.4$

(D) $\Delta\tau = 1$ (E) $\Delta\tau = 0.8$ (F) $\Delta\tau = 0.6$

(G) $\Delta\tau = 0.4$ (H) $\Delta\tau = 0.2$ (I) $\Delta\tau = 0$

Figure SN2: *Computational model results for the net flux per mutation type.* Subplots (A)-(I) correspond to the same parameters shown in Figures SN1. After the final time point of each run, flux in and out of each bin was calculated and attributed to the following transitions: local length increase due to μ substitutions (darker blue); local length decrease due to ν substitutions (darker red); nonlocal fission generated by ν substitutions (darker green); nonlocal fusion due to μ substitutions (purple); local length increase due to expansion (orange); local length decrease due to contraction (lighter blue); nonlocal fusion resulting from deletion of a B base (lighter red); nonlocal fission due to non-motif insertions (lighter green). Dashed black line shows longest populated length L_{max} where $\rho(L > L_{max}) < 1$. Net flux per category was computed as flux in minus flux out (i.e., net change in the number of repeats per length class per mutational transition). After computing the net flux for each effect, the sum of magnitudes of all effects was separately normalized at each length (i.e., height of stacked bars sums to one). If a given transition results in a net influx (outflux), associated bar appears above (below) the axis. Bins showing identical heights above and below zero are maintained in detailed balance.

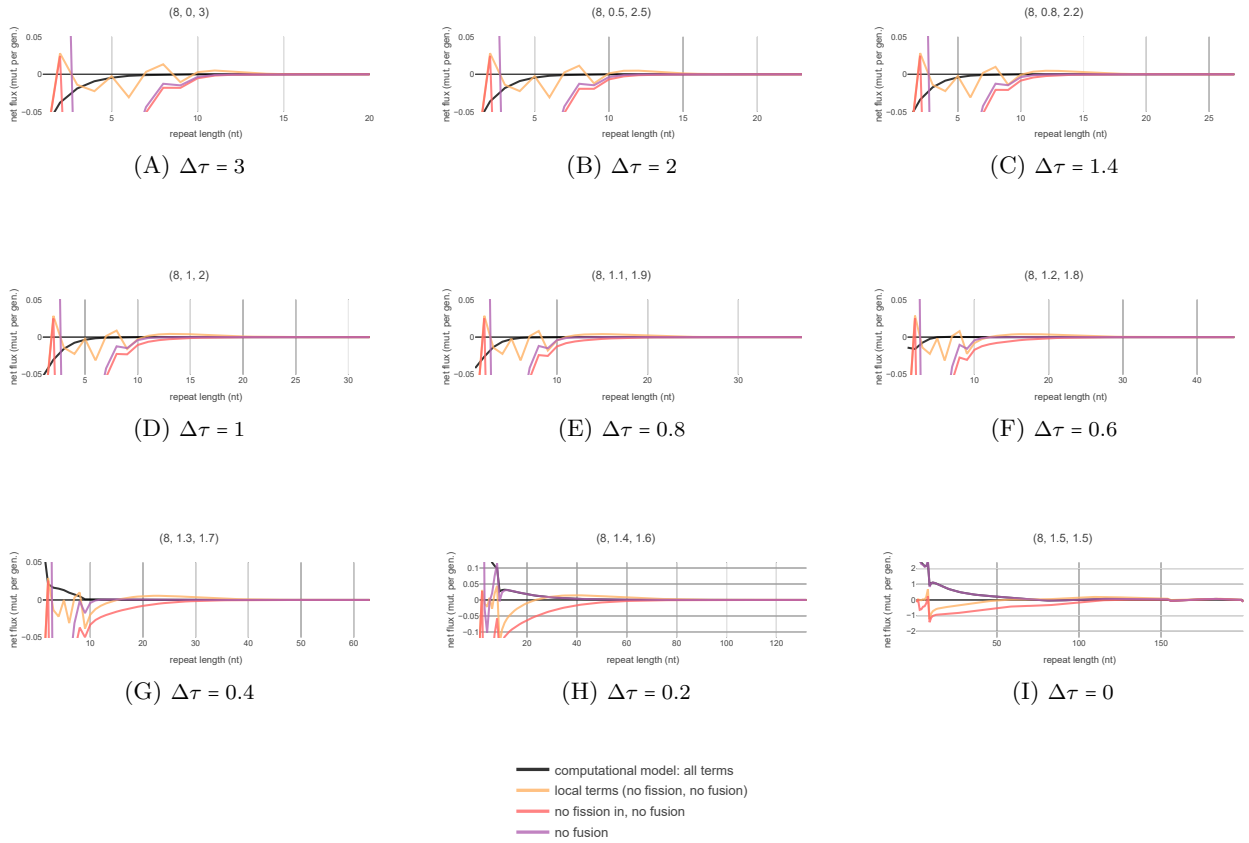


(A) $\Delta\tau = 3$ (B) $\Delta\tau = 2$ (C) $\Delta\tau = 1.4$

(D) $\Delta\tau = 1$ (E) $\Delta\tau = 0.8$ (F) $\Delta\tau = 0.6$

(G) $\Delta\tau = 0.4$ (H) $\Delta\tau = 0.2$ (I) $\Delta\tau = 0$

Figure SN3: *Computational model results for directional flux per mutation type* Subplots (A)-(I) correspond to the same parameters shown in Figures SN1. After the final time point of each run, flux in and out of each bin was calculated and attributed the same mutational processes as in Figure SN2 (shown in corresponding colors). For each category, flux in and flux out are plotted separately at each length (shown above and below zero, respectively). The sum of the magnitudes of all effects (influxes plus outfluxes) was separately normalized to one at each length. Bins showing identical heights above and below zero (i.e., influx equal to outflux) are maintained in detailed balance. In contrast to Figure SN2, each bar height represents the fraction of total number of transitions (in either direction) due to each signed mutational transition (e.g., fraction of number of transitions from expansion influx events, expansion outflux events, contraction influx events, etc.).



(A) $\Delta\tau = 3$ (B) $\Delta\tau = 2$ (C) $\Delta\tau = 1.4$

(D) $\Delta\tau = 1$ (E) $\Delta\tau = 0.8$ (F) $\Delta\tau = 0.6$

(G) $\Delta\tau = 0.4$ (H) $\Delta\tau = 0.2$ (I) $\Delta\tau = 0$

Figure SN4: *Computational model results for collective non-normalized fluxes showing the relevance of local transitions, fission, and fusion.* Subplots (A)-(I) correspond to the same parameters shown in Figures SN1. At the final iterated time point, fluxes were calculated for each length class and summed appropriately to test the accuracy of different analytic models of the long length regime specified in Equations SN37, SN47, and SN48. Each model specifies an approximate steady state equation assembled as a subset of the full collection of terms shown in Equation SN32. Equation SN32, which includes repeat fusion, was summarized by adding the fluxes due to all mutational effects separately at each length (shown in black); detailed balance occurs when all fluxes sum to zero at a given length. Each approximation is deemed appropriate at lengths where they overlap the black curve (restricted to $L > 10$). In contrast to Figure SN1, which tests the accuracy of solutions to the approximated steady state equations, this comparison tests the differential equation more directly by specifying the magnitude of individual terms in the expression (within a given parameter and length regime); in particular, this comparison captures nonlocal effects in Equation SN37 directly, without reference to Equation SN37. The model missing only fusion (Equation SN37; purple) deviates from the full model (Equation SN32; black) only for $L \lesssim 10$ indicating fusion is negligible in the long repeat regime. All three approximations overlap for large $\Delta\tau$, indicating the dominant behavior is local (described by Equation SN48; yellow); the model with fission treated strictly as an outflux (Equation SN47; red) remains a good approximation to the full effects of fission (purple) above roughly $\Delta\tau \sim 0.6$.

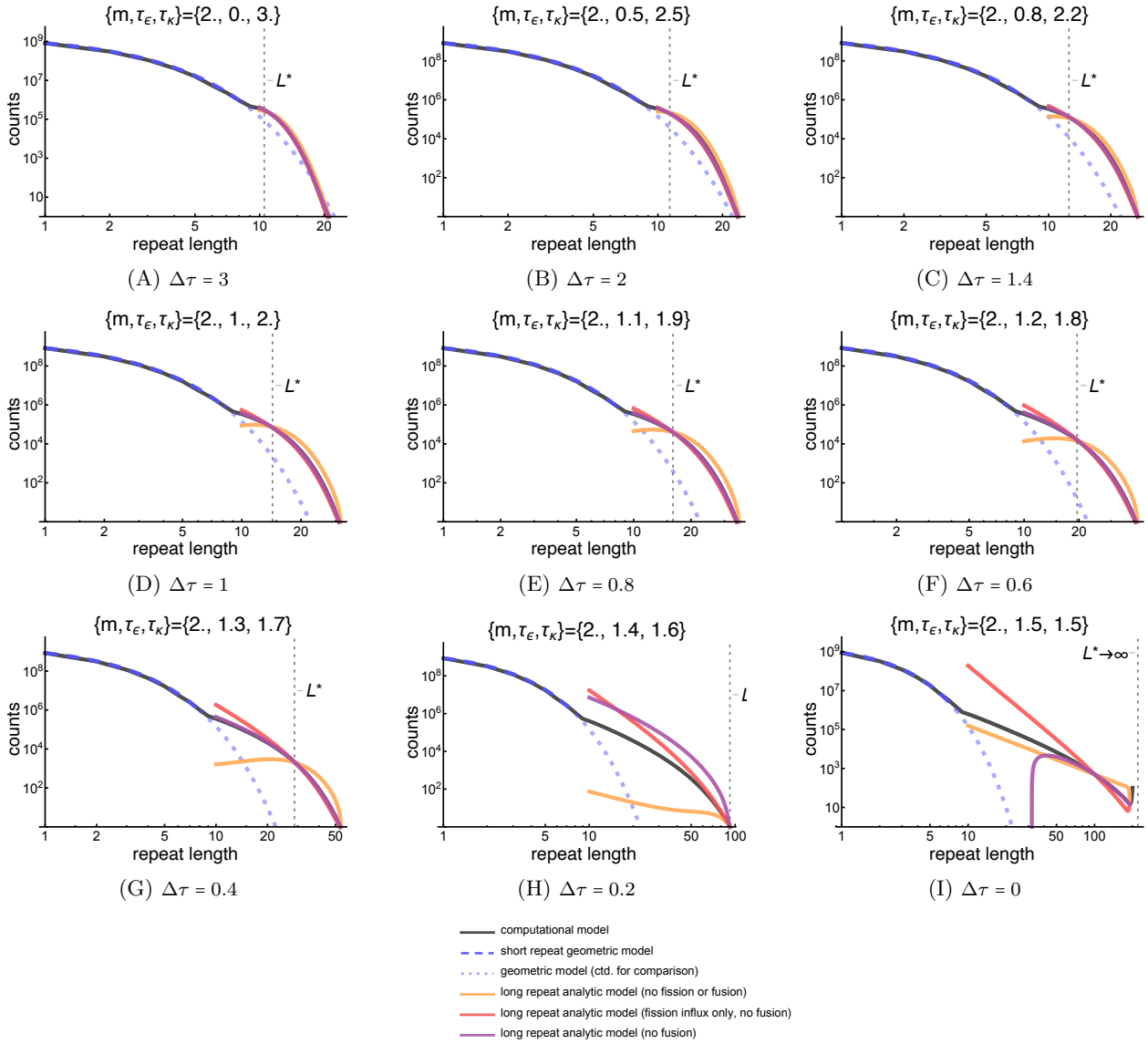


Figure SN5: *Comparison between computational model results and numerical solutions to steady state equations for* $m = 2$. Computationally modeled distributions are plotted for the same $\{\tau_\epsilon, \tau_\kappa\}$ (and therefore $\Delta\tau$) parameter combinations plotted in Figure SN1 (shown in the same location), but for $m = 2$. Each inset shows plots of the computationally modeled distribution at the final time point (black), geometric analytic approximation for shorter repeats of length $L < 10$ (blue, continued as blue dashed line for comparison to distribution tail shape), numerical solutions to Equation SN48 with no repeat fission (orange), numerical solutions to Equation SN47 with fission out but without fission in (red), and numerical solutions to Equation SN53 with fission out and fission in (purple).

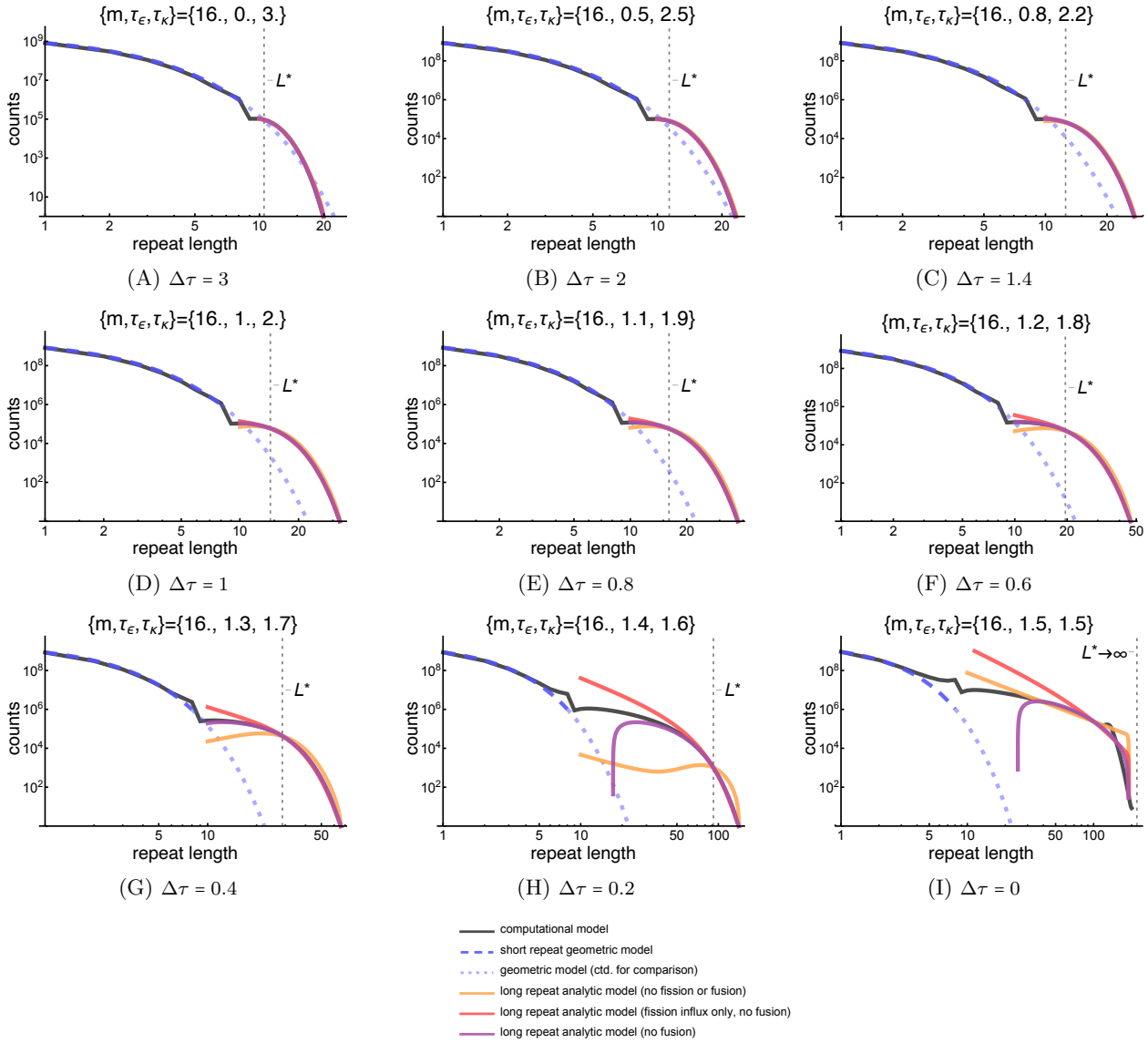


Figure SN6: *Comparison between computational model results and numerical solutions to steady state equations for* $m = 16$. Computationally modeled distributions are plotted for the same $\{\tau_\epsilon, \tau_\kappa\}$ (and therefore $\Delta\tau$) parameter combinations plotted in Figure SN1 (shown in the same location), but for $m = 16$. Each inset shows plots of the computationally modeled distribution at the final time point (black), geometric analytic approximation for shorter repeats of length $L < 10$ (blue, continued as blue dashed line for comparison to distribution tail shape), numerical solutions to Equation SN48 with no repeat fission (orange), numerical solutions to Equation SN47 with fission out but without fission in (red), and numerical solutions to Equation SN53 with fission out and fission in (purple).

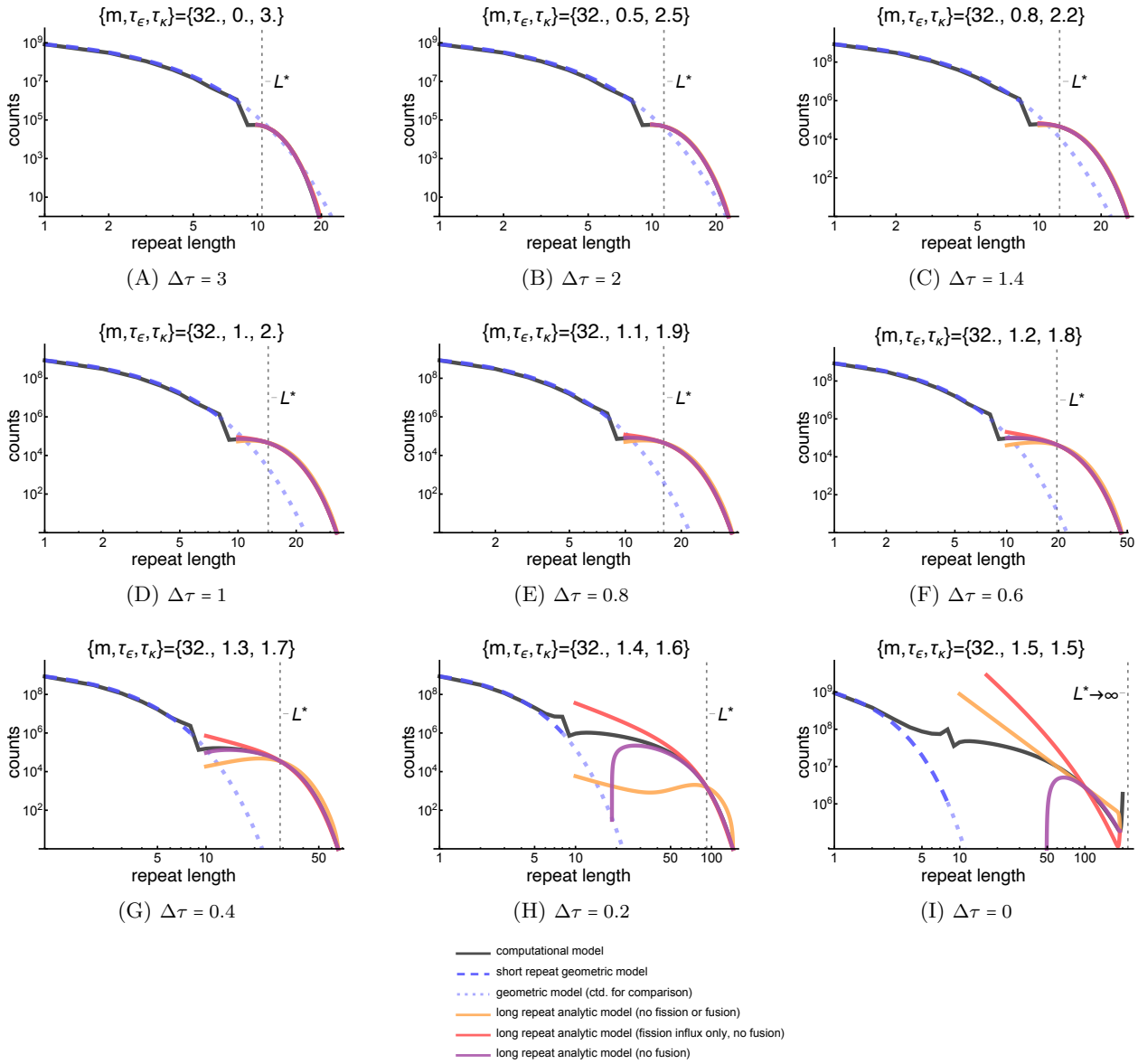


Figure SN7: *Comparison between computational model results and numerical solutions to steady state equations for* $m = 32$. Computationally modeled distributions are plotted for the same $\{\tau_\epsilon, \tau_\kappa\}$ (and therefore $\Delta\tau$) parameter combinations plotted in Figure SN1 (shown in the same location), but for $m = 32$. Each inset shows plots of the computationally modeled distribution at the final time point (black), geometric analytic approximation for shorter repeats of length $L < 10$ (blue, continued as blue dashed line for comparison to distribution tail shape), numerical solutions to Equation SN48 with no repeat fission (orange), numerical solutions to Equation SN47 with fission out but without fission in (red), and numerical solutions to Equation SN53 with fission out and fission in (purple).

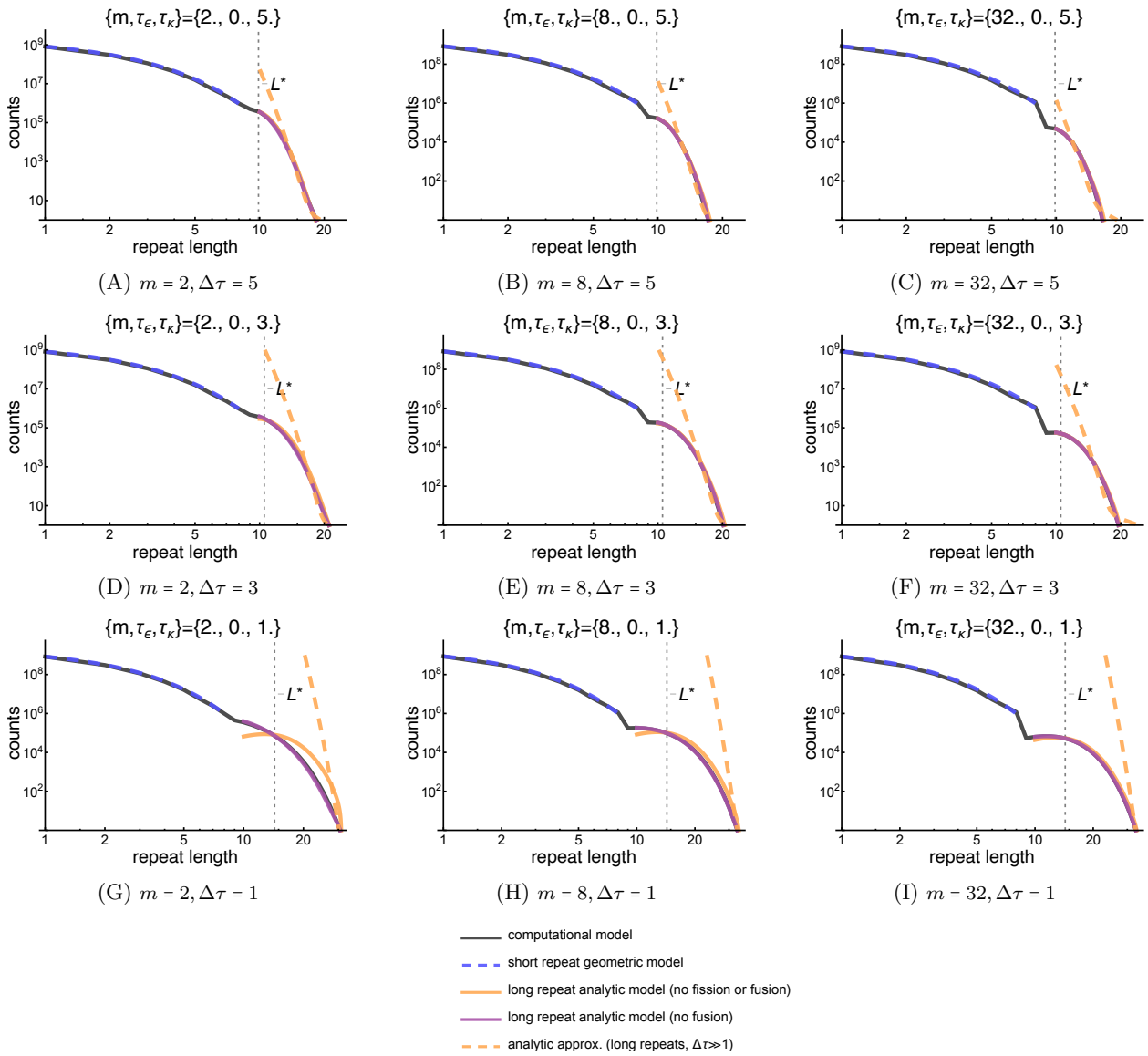


Figure SN8: *Accuracy of analytic approximation for distribution falloff in the $\Delta\tau \gg 1$ regime.* The rough approximation in Equation SN50 to the shape of the steady state distribution when $\Delta\tau \gg 1$ is shown for a wide range of parameter combinations with $\Delta\tau \geq 1$. Each row plots the same combination of $\{\tau_\epsilon, \tau_\kappa\}$ for multipliers $m = 2$ (left), $m = 8$ (middle), and $m = 32$. Each column plots the same value of m for $\Delta\tau = 5$ (top), $\Delta\tau = 3$ (middle), $\Delta\tau = 1$ (bottom). For $\Delta\tau = 5$, the analytic solution approximates the computationally modeled and numerically generated distributions, except at lengths adjacent to the short repeat regime (roughly $12 > L > 10$, noting axes are log spaced). The closed-form solution is independent of m; accuracy and the regime of validity show only a weak dependence on m for $\Delta\tau \gtrsim 3$.