



Evolution of Transcript Abundance is Influenced by Indels in Protein Low Complexity Regions

Zachery W. Dickson¹ · G. Brian Golding¹

Received: 5 October 2023 / Accepted: 24 January 2024 / Published online: 14 March 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Protein Protein low complexity regions (LCRs) are compositionally biased amino acid sequences, many of which have significant evolutionary impacts on the proteins which contain them. They are mutationally unstable experiencing higher rates of indels and substitutions than higher complexity regions. LCRs also impact the expression of their proteins, likely through multiple effects along the path from gene transcription, through translation, and eventual protein degradation. It has been observed that proteins which contain LCRs are associated with elevated transcript abundance (TAb), despite having lower protein abundance. We have gathered and integrated human data to investigate the co-evolution of TAb and LCRs through ancestral reconstructions and model inference using an approximate Bayesian calculation based method. We observe that on short evolutionary timescales TAb evolution is significantly impacted by changes in LCR length, with insertions driving TAb down. But in contrast, the observed data is best explained by indel rates in LCRs which are unaffected by shifts in TAb. Our work demonstrates a coupling between LCR and TAb evolution, and the utility of incorporating multiple responses into evolutionary analyses.

Keywords Low-complexity region · Transcript abundance · Protein abundance · Approximate Bayesian calculation · Co-evolution

Introduction

In any given mammalian proteome, approximately a fifth of protein sequences contain at least one region where the amino acids are highly repetitive or compositionally biased (Karlin et al. 2002). These LCRs were once considered ‘junk’ protein sequences, at most spacers between more traditional protein domains (Golding 1999).

Proteins which contain any LCRs (LCR⁺ proteins) have since been shown to have important roles enabled by their LCRs which confer a host of properties depending on their specific amino acid composition. As compared to LCR free (LCR⁻) proteins, Many LCR⁺ are intrinsically disordered

at physiological conditions (Romero et al. 2001), but changes in those conditions can lead to conformational or phase shifts (Martin and Mittag 2018). Another common property is promiscuous binding. As the regions are unstructured and have little variation in amino acids they cannot discriminate binding targets unless there is some impact from the context of the protein itself (Mier et al. 2017). Non-specific binding to RNA and protein allows LCR⁺ proteins to function as hubs for protein interaction networks (Dosztányi et al. 2005), and in generalized complexes for transcription (eg. Nab3; Loya et al. 2017) and splicing (eg. hhrNPG; Zhou et al. 2019). The latter of which occurs in the spliceosome which is a membraneless organelle; a liquid droplet of RNA and protein. The combination of non-specific binding and inducible phase change also make LCRs critical in another type of membraneless organelle: stress granules (Fomicheva and Ross 2021). LCRs also make appearances in structural proteins like keratin (Parry and North 1998) and collagen (Per-sikov et al. 2000).

The same properties that make LCRs useful, can be harmful if the balance of properties shifts. Expanded LCRs are hallmarks of several neurodegenerative diseases such as

Communicated by Minh Bui.

✉ Zachery W. Dickson
dicksoz@mcmaster.ca

G. Brian Golding
golding@mcmaster.ca

¹ Department of Biology, McMaster University, Hamilton, ON, Canada

Huntington (Cummings and Zoghbi 2000). The number of associated diseases is likely related to the fact that LCRs are mutationally unstable. They evolve rapidly through replication slippage (Huntley and Golding 2006b), unequal crossing over (DePristo et al. 2006), and point mutations (Lenz et al. 2014). The tension between the multiple important roles LCR⁺ proteins have and the mutational risk of utilizing them create evolutionary pressures on the regulation of these proteins. LCR⁺ proteins tend to have lower levels of protein abundance (PAB) (Chavali et al. 2017; Dickson and Golding 2022) as compared to highly conserved, important proteins (Pál et al. 2001). Despite the lower PAB, it has been shown in mammals that LCR encoding transcripts have higher abundance than those which do not (Dickson and Golding 2022).

The disconnect between protein and TAB for LCR⁺ proteins may be potentially explained by any or all of the regulatory steps between gene transcription and eventual protein degradation. At the transcription level, Horton et al. (2023) recently showed that transcription factors interact directly with short tandem repeats (a type of DNA LCR) flanking canonical binding motifs, ultimately affecting the expression of the gene. Non-specific binding to other regulatory proteins may be a mechanism by which protein LCRs alter the abundance of their host transcripts and proteins.

Regardless of the particular changes, protein regulation and LCR sequences must co-evolve to maintain physiologically useful protein levels. Of particular interest is the question of temporal order, does the appearance or expansion of LCRs create selective pressures on the regulation of the proteins which contain them? Or is it that LCRs are only tolerated in proteins which have appropriate regulatory frameworks in place? To answer this we must understand the co-evolution of both LCRs sequences and regulation of PAB.

Most of what is known about LCR evolution has been through study of DNA microsatellites, short tandem repeats in intergenic regions. Most models of their evolution are length dependent stepwise models with slippage being more likely with longer repeats (Kruglyak et al. 1998; Dieringer and Schlotterer 2003; Sainudiin et al. 2004). Point mutations are also included in these models as a mechanism which breakup long repeats. These models indicate the balance of insertions, deletions and point mutations as the explanation for the observed distribution of microsatellite lengths. While the mechanisms of evolution may be similar for protein LCRs the selective pressures for coding regions are very different. Due to significant selection against frameshift mutations, only tri- and sometimes hexanucleotide repeats are tolerated. While the underlying evolutionary change is at the DNA level, models which only allow full codon indels functionally operate as amino acid indels.

As LCRs are ultimately features of primary sequences, they can only evolve through direct changes to the DNA

including both indels and point mutations. This contrasts with the cornucopia of ways to evolutionarily vary PAB, most of which stem from the many steps from gene transcription to protein degradation. Mutations to the gene itself can alter the rate of translation as well as protein stability. Changes altering TAB will also affect PAB, and even here there are multiple indirect mechanisms for evolution. Considering changes which only affect gene transcription, TAB can be altered by mutations in the sequences of transcription factor binding sites and proximal sequences (Odom et al. 2007; Bradley et al. 2010; He et al. 2011). TAB can also evolve through the loss and formation of binding sites (Ni et al. 2012), this is especially true for longer binding motifs which often evolve from transposable element and repeat expansion (Bourque et al. 2008). The level of sequence and binding conservation varies across the tree of life and differs between tissue specific and constitutive transcription factors (Villar et al. 2014). He et al. (2011) showed that binding can be combinatorial, allowing compensatory changes across multiple transcription factors and binding sites. Evolution of TAB is the net effect of a large number of possible effectors.

The combination of many effects will tend towards a normal distribution through the central limit theorem. Therefore the evolution of gene expression is often modelled as a stochastic process with Gaussian increments. Examples include Brownian motion (Bedford and Hartl 2009) or Ornstein-Uhlenbeck (OU) process (Rohlf et al. 2014). The former of which considers the expression to take a random walk over evolutionary time, radiating away from an ancestral state, while the latter introduces a selective optimum which exerts pressure on the abundance value. Either of these can be incorporated into a Bayesian framework to estimate the parameters of an evolutionary model. However to incorporate interactions with LCR length the likelihood calculations become analytically intractable and computationally prohibitive. As an alternative, ABC can be performed, where simulations are performed and compared to the data in order to estimate the likelihood. Beaumont et al. (2002) describe methods which compare summary statistics for observed and simulated data. Pritchard et al. (1999) used simulations and rejection sampling to investigate microsatellites on the human Y chromosome. Marjoram et al. (2003) illustrated a method to use Markov Chain Monte Carlo (MCMC) without likelihood functions. ABC methods can also infer parameters in a multivariate space. Habu and Kutsukake (2019) used ABC to jointly model both group size and sociality in naked mole rats, demonstrating an example of multivariate analysis over evolutionary time.

In addition to interactions between TAB and LCRs, the evolutionary age of proteins may be a lurking variable which could alternately explain the observed positive correlation between the two. Persi et al. (2023) demonstrated that the

relative contribution of LCRs and gene duplications to the evolution of protein families trades off as the families age and become established. Newer protein families evolve primarily through LCR evolution, while gene duplications are the primary mechanism for older families. Likewise there are differences in gene expression between young and old genes, as shown by Werner et al. (2018).

In this work we attempt to determine the evolutionary relationships between LCRs and TAB on the short timescale of human evolution. While of general interest, PAb data on a proteome level for individuals are not often available for outside of humans. In order to consistently compare LCRs between individuals on short evolutionary time scales we use the concept of conserved minimum entropy region (CMER), which is described in detail in the methods section. In short CMER is a region in a protein which has been identified as minimal entropy in any version of the protein across individuals. We characterize changes in CMER length and TAB across individuals, and use ABC to estimate the degree of interaction and temporal order of these evolutionary events.

Materials and Methods

Overview

To investigate the temporal order of changes in LCR and TAB it would be ideal to have a set of individuals where the evolutionary history of the individuals, as well as the sequences and abundances of their proteins is known. From that point it is possible to investigate evolutionary models. However such an ideal situation is not generally possible without intentional artificial evolution experiments. What follows is a general overview of our approach to reconstruct evolutionary histories from the observed data. The details of how this set of evolutionary histories was compiled and modelled are discussed in the following sections.

For mammals, proteome scale data is sparse as are complete evolutionary histories. We have used the available data for humans to build a common set of proteins which have quantified TABs and a consistent method for identifying LCRs. We then employ parsimony and Brownian motion models to reconstruct the evolutionary history for LCRs and TAB respectively. We investigate models of co-evolution using an ABC approach, where simulations are used to estimate the probability of observing the data given a particular model of evolution.

Genomic and Transcriptomic Data

Human data were acquired from the International Genome Sample Resource (IGSR), specifically the “1000 Genomes

30x on GRCh38” (Byrska-Bishop et al. 2022) and “Human Genome Structural Variation Consortium, Phase 2” (Ebert et al. 2021) datasets. Only individuals which had both high coverage genome assemblies and transcriptomic data were selected. A set of 28 human individuals and their accession ids can be found in Supp. Table 1. In addition to the genome assemblies and raw RNA-Seq reads, single nucleotide polymorphism (SNP) calls were also acquired.

In order to ensure consistent annotation of genes and transcripts across assemblies, annotations were transferred from the reference genome GRCh38 (GCF_000001405.39; Schneider et al. 2016) to each assembly using the annotation mapping program `Liftoff` (Shumate and Salzberg 2021) with the `polish` option. The reference annotation was filtered to only include entries for coding sequences to decrease runtime.

Construction of Phylogenetic Tree

The phylogenetic tree was constructed on the basis of SNPs on chromosome 19 of the human genome. As our focus is on co-evolutionary forces common to LCR⁺ proteins, a single tree was used to model the evolution for all proteins. A single chromosome was selected to reduce the time required for tree construction. Our analysis focuses on protein coding sequences, therefore chromosome 19 was selected as it is the most gene dense human chromosome (Grimwood et al. 2004), and most likely to represent the evolutionary relationships between human individuals. For the purposes of an outgroup, the chimpanzee reference genome (GCF_002880755.1; Sequencing and Consortium 2005) was used. Human SNP calls were acquired from the IGSR. Human SNP were generated by mapping fragments of the chimpanzee reference chromosome 19 to the homologous human chromosome 19. The 150 bp fragments were generated by sliding a 500 bp window in 200 bp overlapping increments and taking the first and last 150 bp in the window. These fragments were mapped to the human reference using `BWA` (Li and Durbin 2009). The mappings were sorted and indexed with `Samtools` (Li et al. 2009). `BCFtools` (Li 2011) was used to call chimpanzee SNPs as well as indexing all SNP calls, filtering human calls to the relevant samples, and merging the calls for both species. As human SNP calls were made with a larger set of individuals than the subset used in this study, some sites were invariant in the subset even when including the outgroup. These were discarded with a Perl (Wall et al. 2000) script utilizing the BioPerl package (Stajich et al. 2002). The final set of SNPs was converted to fasta format using `VCF-kit` (Cook and Andersen 2017). The tree was constructed using `IQ-TREE 2` (Minh et al. 2020) with a general time reversible model

including an ascertainment bias correction as only SNP data was used.

Processing of Transcriptomic Data

Adapter sequences present in the raw RNA-Seq data were identified with *FastQC* (Andrews 2015). Adapter removal as well as trimming, was performed with *fastp* (Chen et al. 2018). Reads from each sample were mapped to the genome from which they originated using the splice aware mapper *STAR* (Dobin et al. 2013). The splice junctions used by *STAR* were generated from the *Liftoff* generated genome annotations. Quantification of read counts was performed using *stringtie* (Pertea et al. 2015), which is also capable of assembling and quantifying transcripts outside of the reference annotation. The abundance value used in later analysis is the normalized read count per transcript rather than a sample specific value such as transcripts per million. Normalization for library size is performed using the median of geometric means method as described for *DESeq2* (Love et al. 2014). After normalization, the ‘primary’ transcript was identified as that which had the highest geometric mean abundance across individuals. Only the primary transcript was used in later analysis. Reconstruction of abundance values at the common ancestor of the human individuals was performed in R (R Core Team 2022) using the *Rphylopars* package (Goolsby 2017) supported with the *phytools* package (Revell 2012).

Conserved Minimum Entropy Regions

Protein encoding sequences were extracted from each assembly based on the transferred annotation. The following set of quality control steps were performed on the transferred annotations. Coding sequences which had an inconsistent number of exons across individuals were discarded. Selenocysteine residues were recoded as cysteine residues. Coding sequences where any individual appeared to have a nonsense mutation were discarded. For individuals where there was an apparent frame-shift mutation as indicated by an in-frame stop, and a consistent gap in alignments across individuals, the frameshift was repaired by deleting an apparent insertion or inserting the consensus residue for apparent deletions. If in-frame stops were still present after a single repair, the protein was abandoned. The attempts at repair were made rather than discarding due to the high frequency of apparent frame-shift mutations. This was interpreted as issues from genome assembly or annotation transfer rather than true biological variation leading to hundreds of faulty proteins in any given individual. After repair, the coding sequences were translated. We also used half-alignment ratios as an additional filter to remove proteins which were incorrectly annotated as the same isoform. Each alignment

was divided into two sequences, and for each half the harmonic mean of plurality residue proportion was calculated across sites, as well the proportion of gaps. The ratio of the value calculated for the first and second half should be near one for proper alignments of the same isoform across individuals. If a different isoform is incorrectly included, then the two halves will appear markedly different. We excluded the top 5% of proteins based on their euclidean distances from both ratios being 1. All of this was done using custom Perl scripts utilizing the *BioPerl* package, and performing alignments with *MAFFT* (Katoh and Standley 2013).

In this work, we use the low-entropy definition of LCRs, and perform identification with *Seg* (Wootton and Federhen 1993), using a window length of 15 amino acids, a lower entropy bound ($K1$) of 1.9 bits and an upper entropy bound ($K2$) of 2.2 bits. The default *Seg* parameters are designed to liberally identify LCRs for masking purposes. The values we used have been empirically found to be more useful if the LCRs themselves are of interest (Huntley and Golding 2000, 2002; Haerty and Golding 2010; Huntley and Golding 2006a). We also previously showed that the connection between TAB and LCRs was robust to reasonable choices of entropy thresholds (Dickson and Golding 2022). In this work we used a modified version of *Seg* which properly accounts for alphabet size as described in Enright et al. (2023). However binning proteins into LCR^+ and LCR^- is insufficient for temporal analyses as this categorization cannot distinguish between evolutionary events which nudge a sequence across the threshold and events which radically change the entropy of the sequence. We introduce the concept of CMER to deal with this. CMER can have variable lengths and entropy in different individuals, or not be present, but always refers to a homologous stretch of the protein. Additionally all proteins have a CMER regardless of their LCR status, and will generally have lower entropy for LCR^+ proteins.

To identify CMERs., the minimum entropy window is found for each individual’s version of a protein, *Seg* is then run on the protein with the same window length, $K1$ equal to the entropy of the minimum entropy window, and $K2$ 0.3 bits higher than $K1$. Each LCR identified in the protein is a minimum entropy region, and its location in the individual’s protein version are noted. All versions of the protein are aligned using *MAFFT*. The coordinates of individual minimum entropy regions are converted to alignment coordinates and all overlapping intervals are combined together into CMER. For individuals, the length and entropy of the CMER are calculated from the gap free sequence. An example can be found in Table 1. The length and entropy can also be calculated for the consensus sequence of the CMER. This is useful to compare different CMERs., for example when there are multiple CMERs. in a protein we analyze the one with minimum entropy, then maximum length, then earliest position in the protein sequence.

Table 1 Example of CMERs in NM_001466.4, Frizzled Class Receptor 2. This protein has two minimum entropy regions, the latter of which is perfectly conserved across all 28 individuals. In HG00732 two additional leucine residues have been inserted

Individual	Region 1 (6–22)			Region 2 (176–192)		
	Alignment	Entropy	Length	Alignment	Entropy	Length
Consensus	...ALPRLLLP--LLLLPAA...	1.673	15	...PGAGGTPGGPGGGGAP...	1.609	17
HG00096	...ALPRLLLP--LLLLPAA...	1.673	15	...PGAGGTPGGPGGGGAP...	1.609	17
HG00732	...ALPRLLLPLLLLLLPAA...	1.574	17	...PGAGGTPGGPGGGGAP...	1.609	17

Ancestral reconstruction of indel events in CMERs. was performed using a parsimony based method described by Fitch (1971) and implemented in Perl. The evolutionary states are the length of the CMER, and the probability of observing a change of a given length in a given time (branch length) is assumed to follow a Poisson distribution:

$$P(\Delta L = \ell) = \frac{\omega t^\ell e^{-\omega t}}{\ell!}, \quad (1)$$

where ω is an estimate of the indel rate. The estimate is calculated from the observed deviation in CMER length across individuals and the distances between individuals. Specifically:

$$\omega = \frac{2}{\overline{\text{MAD}} \cdot \bar{D}}, \quad (2)$$

where $\overline{\text{MAD}}$ is the median of mean absolute deviation (MAD) in CMER length across all proteins, and \bar{D} is the mean pairwise distance between tips of the tree. The MAD value for protein i is calculated as:

$$\text{MAD}_i = \frac{\sum_{u=1}^n |x_{i,u} - \bar{x}_i|}{n}, \quad (3)$$

where n is the number of tips in the tree, and $x_{i,u}$ is the CMER length of protein i for individual u . The mean pairwise distance between tips is calculated as:

$$\bar{D} = \frac{2}{n(n+1)} \sum_{u=1}^{n-1} \sum_{v=u+1}^n D_{u,v}, \quad (4)$$

where n is the number of tips in the tree, and $D_{u,v}$ is the distance between tips u and v .

Evolutionary Model

The evolution of CMERs. and TAB were modeled as stepwise and OU processes respectively. Each process also included a term which depended on the value of the other variable to model co-evolution. For TAB, fold changes in the CMER length relative to the length at the root of the tree alter the selective optimum of the OU process. Similarly fold changes in the TAB relative to the root of the tree alter the

rate of indels. Point mutations were also accounted for in the evolution of CMER length; when a mutation occurs the CMER is broken into two parts, and the longer part is then considered to be the CMER. Equations 5 and 10 describe the co-evolution of CMER and TAB along any given branch of the tree.

The length of CMER is the result of 3 processes: insertions, deletions, and point mutations. The length of a CMER at a node which is t time units diverged from a parent node at time T is:

$$L_{T+t} = M \cdot \max [L_T + N_\lambda - N_\kappa, 0], \quad (5)$$

where M is the proportional length of the longest fragment of the CMER after point mutations, L_T is the length at the parent node, and N_x is the number of insertions or deletions. All indels are Poisson distributed:

$$N_x = \text{Pois}(xL_T\Upsilon t), \quad (6)$$

where x is the length, time, and abundance dependent insertion (λ) or deletion (κ) rate. Υ is the effect of abundance on indels:

$$\Upsilon = (A_T/A_0)^v, \quad (7)$$

where A_T is the abundance at the parent node, A_0 is the abundance at the root node and v is the strength of indel dependence on abundance. A positive v indicates that as abundance rises, so too do indel rates. This could be equivalently considered as relaxed selection on indels. The opposite is indicated with a negative v .

If point mutations break the CMER at uniformly distributed points, it has been shown that length of the longest of N segments is distributed as the ratio of the maximum of N exponentially distributed random variables divided by their sum (Holst 1980).

$$M = \frac{\max_{i=1}^R (X_i)}{\sum_{i=1}^R X_i}, \quad (8)$$

where each X_i is an exponentially distributed random variable with a mean of one. The number of these variables (R) depends on number of mutations expected to occur in the CMER in time t , and is Poisson distributed:

$$R = \text{Pois}(\mu L_T t), \quad (9)$$

where μ is the length and time dependent substitution rate.

The TAB at a particular node is the result of two processes, drift which depends only on time, and selection which pushes the mean towards a selective optimum which depends on the CMER length. The TAB at a node which is t time units diverged from some parent node which is T time units diverged from the root is:

$$A_{T+t} = \bar{A}_{T+t} \cdot e^{\text{Norm}(0, \sigma^2)}, \quad (10)$$

where σ is the strength of drift, and \bar{A}_{T+t} is the modal TAB value at the node. The mode is the selection weighted average of the parental node's value A_T and the selective optimum adjusted by a length dependent factor (Δ_τ).

$$\bar{A}_{T+t} = A_T e^{-\delta t} + A_0 \Delta_\tau (1 - e^{-\delta t}), \quad (11)$$

where δ is the strength of selection, and A_0 is the TAB value at the root of the tree which is assumed to be the selective optimum. A_0 was reconstructed using Rphylopars which uses a Brownian motion model, setting the selective optimum centrally, after accounting for phylogeny, within the range of observed TABs. This models the TAB and CMER length at the root as being in equilibrium and only needing to change in response to mutations in one or the other. This choice of selective optimum is best given the short evolutionary timescale and the high variation in TAB variation observed at the tips of the tree. The selective optimum at any given node is inflated (or shrunk) based on the CMER length dependent factor:

$$\Delta_\tau = (L_T/L_0)^\tau, \quad (12)$$

where τ is the strength of length's effect on optimum abundance. As the length of the parent's CMER (L_T) grows relative to the length at the root of the tree (L_0), a positive τ would indicate an increasing demand for higher TAB, and a negative τ would indicate that longer CMERs select for lower TAB.

Our model includes multiplicative drift: the \bar{A}_{T+t} is multiplied by a log Normal deviate with a scale proportional to the strength of drift and time. Many biological processes are inherently multiplicative rather than additive and we found that a multiplicative drift lead to more consistent results.

ABC

As the parameters of interest (τ : effect of length on optimum abundance and v : effect of abundance on indel rate) could only be meaningfully assessed for proteins which had variation in length, the model was fitted using ABC using the subset of the protein data where variability in CMER length

was observed. We evaluated four versions of the model. The full Stepwise OU model's priors for τ and v were *Normal*(0, 2) and *Uniform*(-1, 1) respectively. The full priors for all models can be found in Supp. Table 2. Three special cases of the full model were investigated: τ was fixed at zero (-tau), v was fixed at zero (-upsilon), and both were fixed at zero (-tau-epsilon).

The fixing of the τ and v parameters makes assumptions about the co-evolution of CMER length and TAB. The τ parameter describes the impact of changes in CMER length on the selective optimum for TAB. Fixing τ at zero in the -tau model assumes that there is no impact: TAB is independent of CMER length. Similarly the v parameter describes the effect of changes in TAB on the indel rates in CMER. Therefore fixing v at zero in the -upsilon model assumes indel rates are unaffected by changes in TAB. Setting both to zero in the -tau-epsilon model assumes there is no co-evolution and both evolve independently of the other. By fixing different sets of parameters, we can compare the family of Stepwise OU models to evaluate which best describes biological reality.

As the specific parameter values are unknown, general priors were selected which gave appropriate bounds on the domain. For example, log-normal priors for λ and κ ensured both remained strictly positive. Additional restrictions were placed on the domains of the δ and μ parameters. The former was given a finite upper bound at a value which would result in the A_T having a negligible contribution to the value of \bar{A}_{T+t} in Eq. 11. Specifically the term would be less than one for even the highest abundance transcript along the longest branch of the tree. Any higher values are functionally equivalent to infinite selection strength and are unnecessary to explore. The lower bound of μ was set such that the probability of even one mutation in the combined length of all CMERs along the longest root to tip path in the tree was less than 10^{-9} . Any lower than this is functionally equivalent to a mutation rate of zero in Eq. 9, and was unnecessary to explore. The v parameter was bounded between negative and positive one, not because the value was known to lie in this interval but for numerical reasons. As the variation in TAB values was observed to be relatively high the result of Eq. 7 can be extreme for absolute v values above one. In many cases the tolerances of tools for numerically evaluating the results are exceeded. Bounding absolute v below one made computation possible while still allowing investigation of the qualitative outcomes of positive, negative, and zero values for v .

Our ABC implementation uses simulations to estimate the likelihood of the data given a set of model parameters. For each simulation, the root of the tree is initialized with the ancestrally reconstructed values and then the length and abundance values at each node of the tree are sampled according to Eq. 5, 10. For each human individual

i and protein j in the k^{th} simulation the absolute relative error between observed (O) and simulated (S) values is calculated as

$$E_{i,j,k,x} = \frac{|O_{i,j,k,x} - S_{i,j,k,x}|}{O_{i,j,k,x}}, \quad (13)$$

where x indicates that the same calculation is done for length and abundance. The simulated value is not considered to match the observed value if the absolute relative error is greater than some threshold (ϵ). For relative errors less than ϵ , a partial match between 1 (zero error) and 0 (ϵ or more error) is counted:

$$C_{i,j,k,x} = \begin{cases} 1 - \frac{E_{i,j,k,x}}{\epsilon}, & E_{i,j,k,x} < \epsilon \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

For observed values of zero, only an exact match or mismatch is possible. Partial matches were used rather than exact matches as the latter would happen rarely in a computationally feasible number of simulations, leading to severe underestimation of the likelihood. Bounding partial matches between zero and ϵ ensures that only positive matches are counted and precluding the possibility of negative estimated probabilities. In this work we used an ϵ value of 10% which gave a balance between underestimating from exact matches and accuracy of the simulated results. Pseudocounts were included to prevent counts of zero by increasing the observed match count by one, and the number of opportunities for matches by 2. The proportion of matches across all simulations is then the estimated likelihood for that value:

$$\hat{\mathcal{L}}_{i,j,x} = \frac{1 + \sum_{k=1}^s C_{i,j,k,x}}{r + 2}, \quad (15)$$

where s is the number of simulations. The product of all estimated likelihoods for length and abundance across n individuals, and proteins is the overall estimated likelihood:

$$\hat{\mathcal{L}} = \prod_{i=1}^n \prod_{j=1}^m \hat{\mathcal{L}}_{i,j,L} \cdot \hat{\mathcal{L}}_{i,j,A}, \quad (16)$$

where n is the number of tips in the tree, and m is the number of proteins.

Due to the simulated nature of this likelihood estimation there is variance around the estimate. Given the same set of parameters, multiple evaluations will give a range of likelihood values. We observed the likelihood estimate to be approximately log-normally distributed. This variation combined with asymmetrically preferring higher likelihoods can lead to chains becoming ‘stuck’. That is, a situation occurs where proposals are rarely accepted,

even for very similar model parameters. To illustrate this, consider a chain which always proposes the same model. The true likelihood for each proposal is always the same, and therefore, the proposal should always be accepted. However, using a simulation-based estimate of the likelihood will generate different estimates for each proposal. If by chance the likelihood of the proposal is evaluated to be much higher than the true likelihood. Then, all subsequent proposals are more likely to have lower likelihoods and be less likely to be accepted, despite the model parameters being identical. Proposals will only be accepted for rare events, even more unlikely high-likelihood estimates, or a low probability acceptance. In chains where new model parameters are proposed, this acts to slow mixing as the chain remains ‘stuck’ on proposals with overestimated likelihoods.

Ideally, the variance in the likelihood estimate due to simulation would be handled by evaluating each set of parameters multiple times to get a more accurate estimate. However, this becomes computationally prohibitive. Instead, we have added an adaptive parameter which controls how generously proposals are interpreted. For a single likelihood estimation we cannot know its deviation from the true likelihood. If we interpret proposals generously we assume that the proposal’s estimate was below its true likelihood, and the current model’s estimate is above its true likelihood. Assuming the variance in likelihood estimates is constant for both the proposal and the current model, we can adjust for the difference by multiplying the proposal’s likelihood by a value proportional to the variance. On a logarithmic scale, the generosity (G) is calculated as:

$$G = \text{Norm}_{1-\alpha}^* \cdot \sqrt{2} \sigma_{\ln \hat{\mathcal{L}}}, \quad (17)$$

where $\sigma_{\ln \hat{\mathcal{L}}}$ is the standard deviation of the estimated log likelihood, Norm_x^* is a standard normal quantile, and α is the complement of the assumed deviate from the mean of the log likelihood distribution. It falls in the interval (0, 1] where a value of one indicates that the estimate is assumed to be at the mean and no adjustment is necessary. Conversely as α approaches zero the adjustment grows without bound. By observing the current rate of proposal acceptances this value can be updated as necessary, decreasing α if the chain is ‘stuck’ and increasing α if it is exploring excessively. This adjustment is performed automatically and we consider α to be an adaptive parameter, which we describe more below. The value of $\sigma_{\ln \hat{\mathcal{L}}}$ is periodically estimated by evaluating the current parameter set multiple times.

We made use of heated chains to increase the rate at which the parameter space was explored. Each heated chain’s probability of accepting proposals is elevated based on an adaptive temperature increment, where the increment is varied to achieve a specified target swap rate. This adaptation

is described more below. Periodically, the chains are synced and an attempt to make a swap between two chains is made which depends on relative temperature and estimated likelihood of each chain's current parameter set. The effect of heating on proposal acceptance, and chain swapping is as described by Shi and Rabosky (2015).

For any particular parameter, a new value is proposed with a normal deviate from the current parameter's value. The proposal density is truncated to match the domain of the prior for the parameter. The scale of the proposal density is adaptive over the MCMC run: increased or decreased to keep proposal acceptance rates at a specified target value. Adaptive parameters are described more below. On any given iteration of the MCMC, some combination of parameters is allowed to vary. This is performed systematically by initially enumerating all combinations and then shuffling the combinations to break up runs where one parameter is altered or fixed many times consecutively. This shuffled order is then cycled through on each iteration.

Our MCMC is controlled by several adaptive parameters: generosity, temperature increment, and parameter specific proposal densities. In some cases, adaptation can violate the assumptions required for convergence of the MCMC (Andrieu and Thoms 2008). In our implementation no parameter depends directly on the current state, only indirectly through properties of the likelihood landscape of the recently visited states. Each parameter tracks some property of the system (X_i) over some time horizon, and compares it to a target value for that property (X_0). Temperature increment has a target chain swap rate of 50% over the last 50 attempted swaps. Generosity and proposal scales both target a proposal acceptance rate of 23.4% which has been shown to lead to optimal mixing in many cases (Schmon and Gagnon 2022). Generosity operates on a 51 iteration time scale, and each proposal scale parameter has a time horizon of 23 iterations where that particular parameter is not fixed. For each parameter the X_i value and the property's value on the previous time horizon X_{i-1} are tracked. After each time horizon both X_i and X_{i-1} are compared to X_0 and the adaptive parameter's value is updated if the trend between X_i and X_{i-1} is away from X_0 . A running average is not used to minimize the dependence of the current state on the history. An update consists of setting the current parameter value to a uniform random value between either double the current value or half the current value as necessary. The temperature increment increases if the swap rate is too low, while proposal scales are decreases if acceptance rates are too low. Within any given set of adaptive parameters, the prior corrected transition probabilities between model and proposal remain symmetrical which supports convergence. The adaptive parameters are disconnected from the particular model parameters, instead more related to the topology of the local likelihood landscape.

After estimating the likelihood of the proposal, the probability of accepting the proposal is calculated according to:

$$P = \min \left[1, \left(\frac{e^G \hat{\mathcal{L}}_p \pi(p)}{\hat{\mathcal{L}}_m \pi(m)} \right)^T \right], \quad (18)$$

where $\pi(x)$ is the prior density of the parameter set for the current model (m) or the proposed model (p), G is the generosity, and T is the temperature of the chain. The acceptance probability depends on the likelihood ratio; the Hastings ratio, which accounts for the asymmetric proposal densities (Hastings 1970); the chain temperature; and the simulation variance.

Iteration of MCMC chains was stopped based on multivariate effective sample size (mESS) as defined by Vats et al. (2017). Specifically iteration terminated after mESS crossed a threshold of 1000 effective samples or the expected Monte Carlo error fell below 15%

Model Analysis

After ABC evaluation, the maximum likelihood estimate for modal parameters is the parameter values at the multivariate mode of the posterior density. To estimate this value, the smoothed multivariate density was estimated for each sample using a multivariate normal kernel. The sample with the highest density was used as an initial estimate. This estimate was then iteratively improved by estimating the gradient in the smoothed density at the current point, then using golden section search (Kiefer 1953) to find the maximum density along the line in the gradient direction. This is repeated until there is no increase in density, or the gradient magnitude is sufficiently small.

The final likelihood of each model was evaluated as the geometric mean of 10 evaluation runs, each with 10000 simulations. Model selection was performed using Akaike information criteria (AIC; Akaike 1998). A $q\%$ credibility region was determined by standardizing all parameter estimates to bring them to the same scale, ordering the smoothed multivariate densities of each sample by the euclidean distance from the multivariate mode, and finding the distance at which the cumulative density is $q\%$ of the total smoothed density. Transforming a hypersphere with this radius results in the ellipsoid credibility region. The corresponding credibility interval for each parameter is then the range of values observed within the region.

Results

After implementing consistent annotation with `Liftoff`, sequence repair, and quality control filters, we identified 7331 primary transcripts and their associated proteins which were present in all 28 individuals. Of these, variation in CMER length was only observed in 57 proteins (See Supp. Table 3). As such the most parsimonious estimate of the number of indel events for all other proteins is zero. In the small subset of proteins with CMER length variation we inferred a maximum parsimony set of 132 events. The number of insertions and deletions were 84 and 48, respectively, which is significantly unbalanced by Chi-squared test ($p < 0.01$). The branches along which these indels were inferred can be seen in Fig. 1A.

The inference of indel events also provides a reconstruction of the CMER length for the lowest common ancestor (LCA) of the individuals. The proteins can be subdivided into LCR^+ and LCR^- by comparing the entropy of their most extreme CMER to (K1) of 1.9 at the LCA. We observe that 1856 (25.3%) of proteins contained LCR at the LCA. In contrast, 48 (84.2%) of the proteins where we observed CMER length variation were LCR^+ .

There were 2 proteins where indel events would cause entropy to cross K1 and change the LCR status of the protein. The transcripts encoding these proteins are NM_015440, which encodes a protein with a C-terminal poly-glycine tract, and NM_145269, which encodes a protein with a glutamate-rich N-terminal region. In both cases the deletion of a single residue increases the entropy in the region to just above K1, causing a ‘loss’ of the ancestral LCR. These two proteins are specifically marked points in Fig. 1.

In Fig. 1C the CMER lengths at the LCA are broken down by LCR status and CMER length variability. The same breakdown for the ancestrally reconstructed TAB is in Fig. 1B. For length, we observed that static CMERs in LCR^- proteins have a median length of 17 (95% CI [17, 17]) amino acids, longer than the 14 (95% CI [14, 15]) of those with LCRs. Proteins with variability in CMER length also had a longer median length of 17 (95% CI [16, 18]). The median TABs (in thousands of normalized reads) for LCR^+ positive proteins with static and variable CMER lengths were, respectively, 7.56 (95% CI [7.14, 8.03]) and 10.8 (95% CI [9.43, 14.1]). Both medians for LCR^+ proteins (static or variable) are higher than the median for static LCR^- proteins: 5.84 (95% CI [5.60, 6.10])

We used ABC to fit four evolutionary models describing the co-evolution of TAB and CMER length. A summary of all estimated parameter values and likelihoods can be found in Table 2. The model which consistently had the highest likelihood was -upsilon with log likelihoods

ranging from -5374 ± 2.137 to -5386 ± 3.043 . This model fixes the value of v (the degree to which shifts in TAB impact indel rates) at zero, and assumes CMER indels are TAB-independent. As v is fixed, it has fewer free parameters. Additionally, this model was more consistent than the full model which was the model with the next highest likelihood. The best replicate of the full model had a log likelihood of -5387 ± 1.924 . With an already higher likelihood, the -upsilon model also had the replicate with the lowest AIC ($1.076 \times 10^4 \pm 4.275$). This may indicate that indels have a bigger impact on TAB evolution than the reverse.

As additional evidence for indels having a bigger effect: both models which set τ at zero had significantly lower likelihoods. As τ describes the degree to which indels impact TAB evolution, fixing τ at zero assumes TAB is independent of CMER length. The highest likelihood between these two zero- τ models was replicate 3 of -tau which had a log likelihood of -5429 ± 3.398 : 25 natural orders of magnitude less likely than the worst estimate for a model which included non-zero τ . This may indicate that TAB evolution is impacted by the evolution of CMER length.

In all models where τ (the impact of indels on TAB evolution) was estimated, the 95% credibility interval included zero, however the multivariate modal value was consistently below zero. In contrast v values largely filled the range of possible values defined by its prior. Substitution rates, as measured by μ , were lower, and in no accepted sample did the mutation rate rise above 10^{-2} amino acid substitutions per site per unit time. The modal estimates range between 10^{-12} to 10^{-4} . The strength of selection appears to be able to take on any value so long as it is sufficiently high (above 10 per unit time). In contrast the strength of drift was consistent: the drift factor between two nodes of a tree follows a log-normal distribution with a scale factor between 9.0 and 9.7 per unit time.

Estimated insertion and deletion rates were consistently estimated as approximately equal, or at least insignificantly different. While their credibility intervals span 4 orders of magnitude, the modal values were consistently estimated between 0.001 and 0.01 amino acid indels per site per unit time. We ran two additional models equivalent to the full model and -upsilon, but explicitly setting insertion and deletion rates to be equal. Visualizations for these two models can be found in Supp. Figures 37 to 42. The AICs values and their standard deviations for the equal indel models which otherwise match the full and -upsilon models were $1.081 \times 10^4 \pm 5.002$ and $1.079 \times 10^4 \pm 5.411$, respectively. The former is at the upper range of values seen for the full model, while the later is outside the range (Table 2). The non-indel parameter estimates were not qualitatively different. Models which allow for even small imbalances between insertions and deletion rates better explain the data.

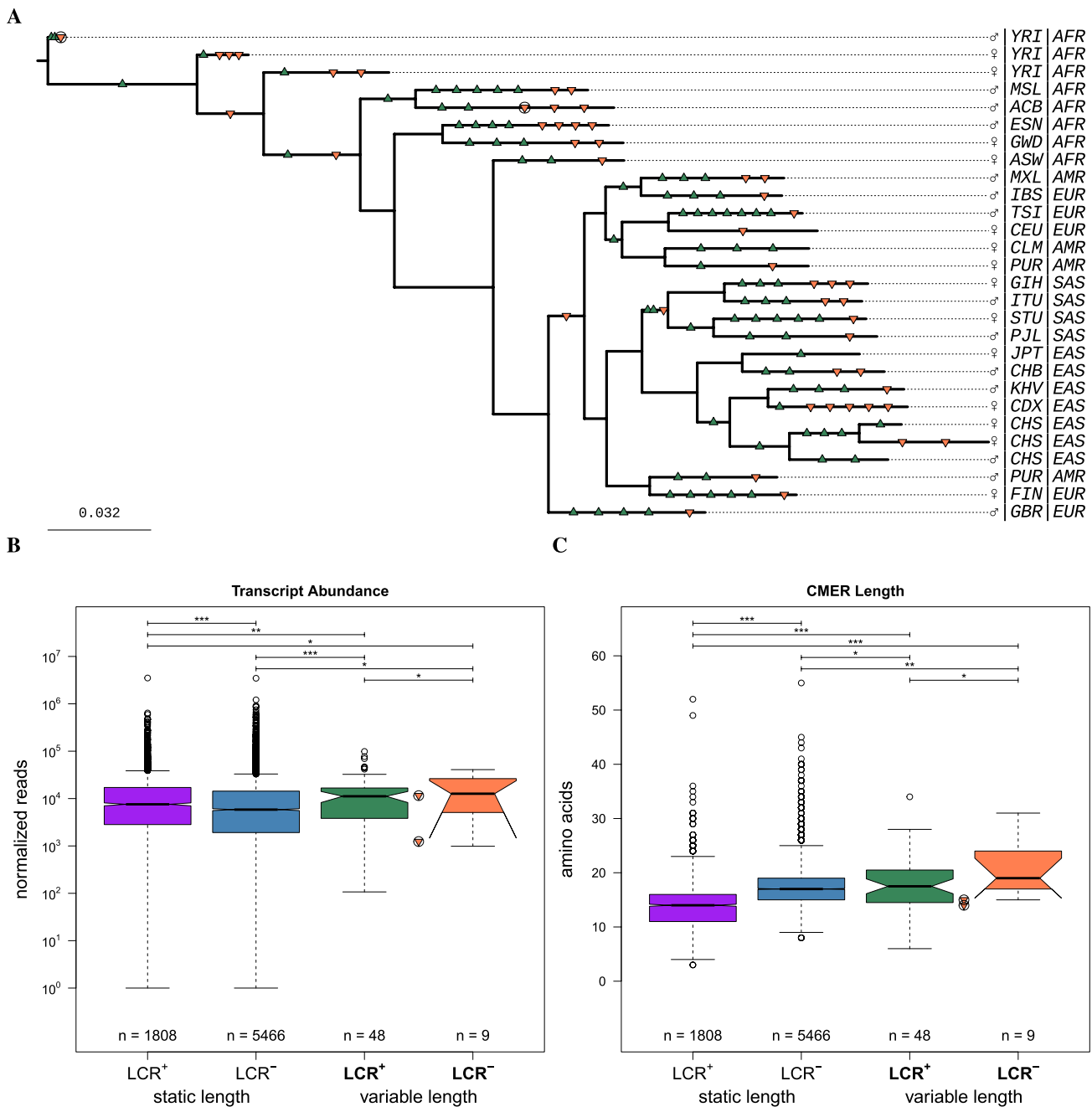


Fig. 1 Properties of the proteins used in ABC modeling. **A** The SNP tree for chromosome 19 of 28 human individuals. All 132 indel events in CMERs are shown on the branch along which they are inferred to have occurred. Green, Upward-pointing triangles indicate the 84 insertions, while orange, downward-pointing triangles indicate the 48 deletions. Sex as well as population and superpopulation codes are shown for each individual. Circles indicate indels which changed LCR status. The chimpanzee outgroup is not shown. **B** TAB data and

C length data reconstructed for the LCA of 28 human individuals broken down by LCR status and whether any variation in CMER length was observed. Notches indicate approximate 95% confidence intervals on the median, which may be wider than the interquartile distance. Circled points indicate events which would cause a change in LCR status. Significance assessed by Wilcox-test. One, two, and three asterisks represent significance at the 95, 99, and 99.1% significance levels, respectively

The posterior distribution for δ -upsilon replicate 1 can be found in Fig. 2. For the δ and μ parameters it can be seen that they can take on any value allowed by their priors except at the lower and upper extremes, respectively. That is, selection

can be any value so long as it is sufficiently high, and substitution rates can be any value so long as they are sufficiently low. Indel parameters κ and λ are constrained to the lower quadrant: both must be low. Both σ and τ appear to take on

Table 2 Summary of ABC modeling runs

Model	Replicate	Parameter modes and 95% credibility intervals										-ln L		AIC	
		δ	κ	λ	$\ln \mu$	σ	τ	ν	Mean	St.Dev	Mean	St.Dev			
Full	1	2912 (160.2 3300)	0.002906 (4.863E-6 0.02345)	0.005895 (4.033E-6 0.02101)	-13.82 (160.2 3300)	9.646 (7.511 11.15)	-1.104 (-7.661 6.412)	-0.1412 (-0.9999 0.9982)	5389	1.924	1.079E4	3.848			
	2	698.7 (10.39 3300)	0.00324 (2.557E-6 0.02428)	0.007198 (3.333E-6 0.0304)	-28.08 (10.39 3300)	8.989 (7.420 11.12)	-1.384 (-6.009 3.302)	0.392 (-0.9979 0.9996)	5404	2.349	1.082E4	4.698			
	3	792.5 (18.68 3299)	0.0005936 (1.081E-7 0.01941)	0.005722 (1.911E-6 0.02513)	-19.64 (18.68 3299)	9.356 (7.823 10.71)	-1.169 (-10.56 7.545)	0.1348 (-0.9995 0.9996)	5387	2.467	1.079E4	4.934			
-Tau	1	1342 (15.04 3299)	0.008613 (6.735E-7 0.02487)	0.006222 (6.332E-6 0.01854)	-28.38 (15.04 3299)	9.182 (7.964 10.79)	0 (0 0)	0.2141 (-0.9996 0.9998)	5457	4.549	1.093E4	9.098			
	2	2769 (22.21 3300)	0.002485 (7.450E-8 0.02048)	0.004264 (4.570E-6 0.02477)	-12.61 (22.21 3300)	9.686 (8.002 11.34)	0 (0 0)	0.270 (-0.9997 0.9997)	5431	3.992	1.087E4	7.983			
	3	2094 (18.04 3300)	0.005444 (2.351E-6 0.02175)	0.0009211 (6.756E-6 0.01474)	-23.51 (18.04 3300)	9.743 (8.029 11.35)	0 (0 0)	-0.2948 (-0.9992 0.9994)	5429	3.398	1.087E4	6.796			
-Upsilon	1	1013 (17.15 3300)	0.002174 (2.994E-7 0.02031)	0.00392 (1.945E-6 0.02409)	-12.96 (17.15 3300)	9.298 (7.756 10.95)	-1.787 (-7.58 5.146)	0 (0 0)	5374	2.137	1.076E4	4.275			
	2	1106 (13.77 3299)	0.005218 (1.331E-5 0.0232)	0.00561 (1.128E-7 0.02244)	-25.35 (13.77 3299)	9.192 (7.755 10.66)	-1.692 (-7.486 4.964)	0 (0 0)	5384	3.483	1.078E4	6.967			
	3	1801 (7.548 3300)	0.007334 (3.330E-6 0.02608)	0.006005 (1.368E-6 0.0255)	-27.49 (7.548 3300)	9.410 (7.609 11.15)	-1.51 (-8.00 5.442)	0 (0 0)	5386	3.043	1.078E4	6.086			
-Tau-epsilon	1	1373 (15.19 3298)	0.003638 (3.055E-6 0.01659)	0.007043 (1.167E-6 0.0209)	-22.38 (15.19 3298)	9.241 (7.991 10.62)	0 (0 0)	0 (0 0)	5457	3.753	1.092E4	7.506			
	2	1663 (8.020 3299)	0.003034 (1.569E-6 0.01805)	0.003262 (6.223E-7 0.01885)	-19.0 (8.020 3299)	9.258 (7.862 10.73)	0 (0 0)	0 (0 0)	5453	2.935	1.092E4	5.871			
	3	2748 (12.93 3300)	0.001863 (2.397E-7 0.02092)	0.007496 (5.490E-7 0.02696)	-10.34 (12.93 3300)	9.234 (7.831 11.03)	0 (0 0)	0 (0 0)	5455	2.997	1.092E4	5.994			

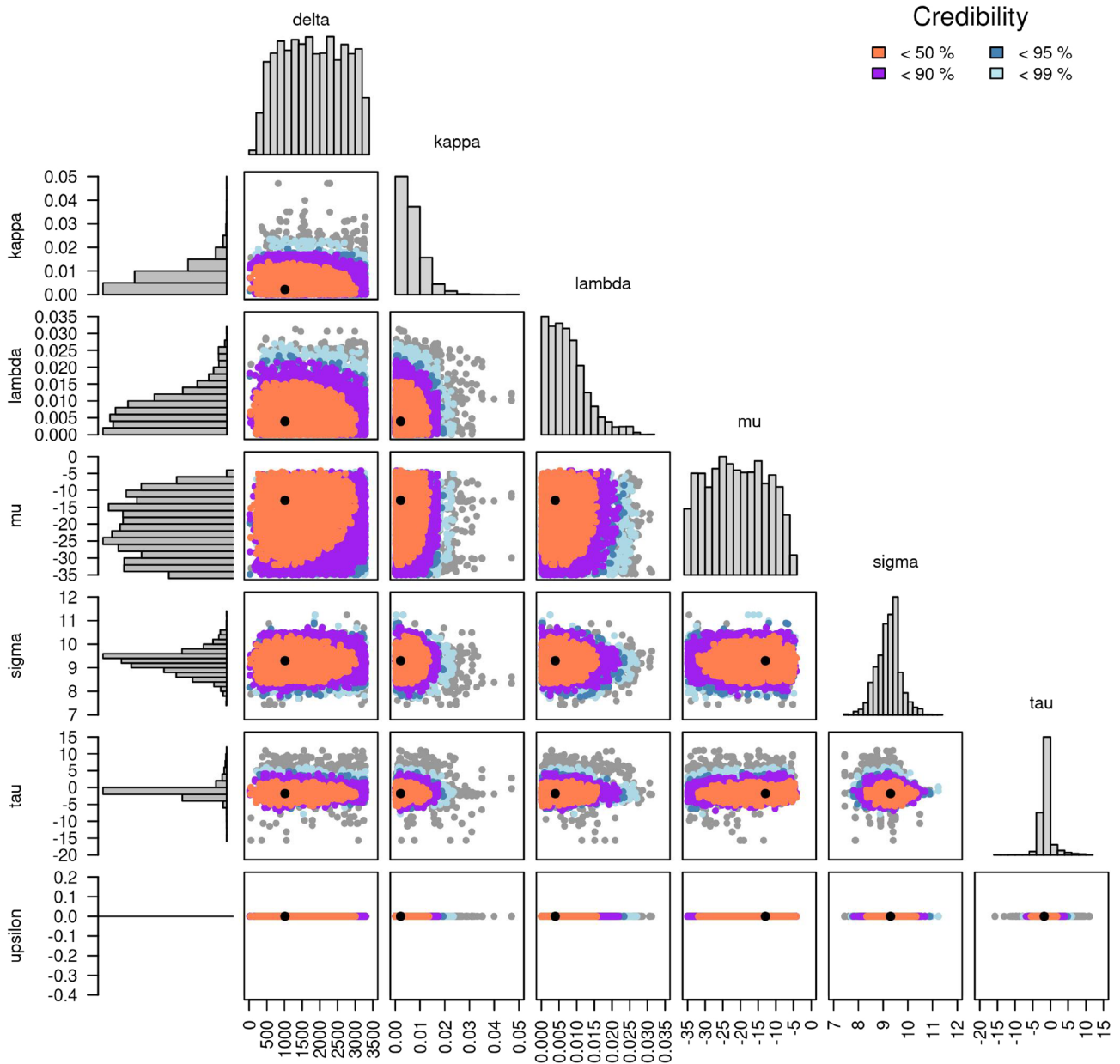


Fig. 2 A visual summary of the posterior distribution estimate by ABC of the StepwiseOU- ϵ evolutionary model which assumes indel rates are independent of TAB (epsilon is zero). The central black point indicates the multivariate mode of parameter estimates, with colours indicating the credibility interval within which each posterior sample fell. Each scatter plot is a projection of the posterior dis-

tribution down to two dimensions. The parameters are the selection strength (delta); deletion (kappa), insertion (lambda), and substitution (mu) rates; the scale of drift (sigma); the strength of CMER length on TAB (tau), and the impact of TAB on indel rates (epsilon). The parameter mu is shown in natural orders of magnitude

defined values with a specific, strong drift strength, and a specific, negative, and small magnitude dependence of TAB on CMER length. For most parameters, the simulations are robust to their values: the range of parameter values which produce acceptable simulation results is quite wide. However for τ there is a much narrower window where the model is able to predict observations. The predictive power of the model depends strongly on the value of τ .

The trends described are specific to the -epsilon model. The MCMC traces and a closer look at the univariate posteriors for each parameter can be found in Supp.Figures 20 and 21., respectively. For all other models the general trends for posterior distributions are qualitatively similar. These trends can be inspected in the visualizations for every modelling run which can be found in the supplement. As epsilon is fixed in the -epsilon model, we must look to other models

to examine its posterior distribution. In all the models which include ϵ (full and $-\tau$), the posterior distribution appears to match its prior. That is, it is uniformly distributed between negative and positive one. The particular value of ϵ does not appear to influence the ability of the model to describe the data.

As a test of the ability of the ABC method to fit the model, we can examine how closely a particular set of parameters predicts the observed data. In Supp. Figure 1 we show a test of this for the $-\epsilon$ model. Simulated results were compared to observations, where each simulation's model parameters were randomly drawn from either the prior or posterior distribution. Parameter sets drawn from the credibility region of the posterior simulate results which deviate less than parameter sets randomly drawn from the prior.

Discussion

We have identified human proteins which have had insertions or deletions amino acids in their evolutionary history. Constructing evolutionary models for TAB and CMER length for these proteins has demonstrated that co-evolution of the two is required by the data. The model which assumed the evolution of both were mutually independent ($-\tau$ - ϵ) was least likely to produce the data. More specifically TAB is more likely to be impacted by indels than the rate or tolerance of indels is to be impacted TAB. Any model which included a dependence of TAB on CMER length had a higher likelihood, than any which did not. Additionally, between the full model and $-\epsilon$ model, the later achieved a higher likelihood with fewer free parameters. Fixing ϵ at zero assumes that CMER length evolves independently of TAB, and therefore the apparent better fit of the $-\epsilon$ model over all others would indicate that the effect of TAB on LCR evolution is weaker than the reverse. However there are important factors in the data used which may limit the scope of conclusions that can be drawn.

We will first discuss the number, type, and distribution of inferred amino acid indel events. We observed significantly more insertions than deletions which is consistent with a finding by Gonzalez et al. (2019) showing higher tolerance of indels in β -lactamases. Indels in terminal regions also tend to have higher tolerance (Lin et al. 2017), and we did observe bias towards the termini. Taking the central position of CMER as a proportion of the protein length and fitting the two shape parameters of a beta distribution can give a general description of where the CMERs are located. In the 57 proteins where variation in CMER length were observed, a $\text{Beta}(0.56 \pm 0.094, 0.59 \pm 0.102)$ best describes the CMER position distribution. Values below one indicate bias towards the C- or N-terminal regions, respectively.

The probability of this fit for a beta distribution if it were truly $\text{Beta}(1, 1)$ distributed (uniformly), is less than 10^{-7} . Across all proteins the distribution of CMER positions is generally non-uniform, with a bias towards the C-terminus with a $\text{Beta}(0.80 \pm 0.012, 0.94 \pm 0.015)$ distribution ($P(\text{uniform}) < 10^{-67}$).

We also observe an uneven distribution of inferred indels across time. In Fig. 1A, events are biased towards terminal branches of the tree. Accounting for branch lengths, the indel rate is significantly higher on terminal branches than internal branches (Wilcox test: $p = 0.0001$). This is not likely to be the biological truth as there is no mechanism to explain increasing indel rates at tree tips. It is more likely that there are hidden insertions and deletions masking each other. If we assume an opposing insertion and deletion pair along each branch, the internal rates exceed that on terminal branches and the difference becomes insignificant (Wilcox test: $p = 0.16$). While we do not explicitly correct underestimated internal events, our evolutionary models allow simultaneous insertions and deletions, and the likelihood estimation depends only on the known tip data. However the reconstructed root state used as the start point for simulations may have been more similar to the observed tips than the true root state.

More important to our conclusions than the events themselves is the CMER length and TAB of the proteins in which we observed the events. In general these proteins would be classified as LCR^+ , with longer CMERs than the LCR^+ proteins where we did not observe variation in CMER length (Fig. 1C). The fact that we only observed changes in longer low-complexity regions on this short evolutionary timescale is consistent with indel rates being proportional to the length of repeats, which has been well established (Kruglyak et al. 1998; Dieringer and Schlotterer 2003; Sainudiin et al. 2004). Of note is that among proteins with static CMER lengths, LCR^- proteins tended to have longer CMERs than LCR^+ proteins. CMER length alone is not indicative of more extreme LCRs. A minimum entropy region is a minimum for that protein, therefore both a protein with uniform, high complexity and a protein with a long homo-repeat could have long minimum entropy regions. The difference would be in the entropy of those regions, with the former being high, and the latter low. The proteins which had variable CMER length had long and low-entropy CMERs.

Turning to TAB, we observe that the proteins which we were able to include in our evolutionary models had higher TAB than those where CMER lengths remained static. While this is consistent with our previous work showing that LCR^+ proteins are encoded by higher abundance transcripts (Dickson and Golding 2022), it also indicates that whatever mechanism causes the elevated transcript abundance has probably already had its effect for these proteins. As a result, our modelling does not capture the full evolutionary interplay

between TAb and LCR evolution. Our conclusions are limited to describing how these properties co-evolve after the regulatory or evolutionary machinery accommodating LCRs is in place.

This potentially explains why the effect of increasing CMER length appears to apply negative evolutionary pressure to TAb, despite the net observed effect of LCR presence being an elevation of TAb. In either case where regulatory changes allowed LCR to be tolerated, or the appearance of LCR induced compensatory increases in TAb, further increases in the LCR length may tip the fitness balance in the other direction. The benefits of maintaining the protein concentration may be outweighed by the increased deleterious effects of the longer LCRs.

Persi et al. (2023) showed that the evolutionary pressures and mechanisms differ depending on the age of the protein family. We made an effort to get relative ages for the proteins in our dataset. For each protein we constructed a consensus sequence based on the MAFFT alignment from all 28 individuals. Then we used BLAST (McGinnis and Madden 2004) to search for homologues to the consensus sequence. Ignoring synthetic, or other artificial constructs we identified the LCA across all proteins matching to the consensus sequence. This was done using a Perl script which made use of TaxonKit (Shen and Ren 2021). This assignment of LCA is taken as an approximate age for the protein ranging from human specific to shared by all eukaryotes. The median LCA for proteins with static length was at the superclass level (Sarcopterygii) while the median for proteins with variable length CMER was the class level (Mammalia). However by chi squared test the distribution across all 16 taxonomic ranks considered was not significantly different ($p = 0.17$). In general the proteins included in the modeling are ancient relative to the timescale analyzed. This is further evidence that the evolutionary impacts of LCR appearance have already been felt, and the evolution we modelled is nested within that effect.

Nevertheless, we observed a negative relationship between LCR length and TAb on short evolutionary timescales after the establishment of LCRs, despite an overall positive relationship between LCR presence and TAb. This offers hints as to temporal order of LCR establishment and TAb elevation. It suggests that regulatory frameworks may be in place prior to establishment of LCR, however further work is needed to determine this. Deeper time data-sets are needed to identifying the establishment of LCRs in protein families. This is a critical step to answering the temporal question of TAb and LCR evolution. Challenges to overcome include the fact that LCRs evolve rapidly which makes identifying evolutionary events increasingly difficult with deeper time. Also there is limited availability of high quality genomes and transcriptomes to properly bracket the required timescales.

While we cannot currently elucidate the original temporal order of TAb and LCR, our results indicate that TAb evolution is coupled to changes in LCRs. After establishment of an LCR further increases to the LCR may increase selective pressure against further elevating transcript abundance. Our work demonstrates the usefulness and importance of incorporating multiple evolutionary outcomes into models to fully understand the contributions of all factors.

Acknowledgements This work was funded by the Natural Sciences and Engineering Research Council of Canada (grants RGPIN-202-05733 to GBG and PGSD3-547476-2020 to ZWD).

Code Availability Custom Perl and R scripts used for quality control of input data and reconstruction of ancestral TAb and LCR states can be found on Github at: www.github.com/zacherydickson/AncRecon-LCR-TAb. The program written to perform ABC inference of co-evolutionary models can be found on Github at: www.github.com/zacherydickson/ABC-LCR-TAb

Declarations

Conflict of interest The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

- Akaike H (1998) Selected Papers of Hirotugu Akaike. Chapter Information Theory and an Extension of the Maximum Likelihood Principle. Springer, New York, pp 199–213. https://doi.org/10.1007/978-1-4612-1694-0_15
- Andrews S (2015) Fastqc. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Andrieu C, Thoms J (2008) A tutorial on adaptive MCMC. *Stat Comput* 18:343–373
- Beaumont M, Zhang W, Balding D (2002) Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035
- Bedford T, Hartl D (2009) Optimization of gene expression by natural selection. *Proc Natl Acad Sci USA* 106:1133–1138
- Bourque G, Leong B, Vega V, Chen X, Lee Y, Srinivasan K, Chew J, Ruan Y, Wei C, Ng H et al (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* 18:1752–1762
- Bradley R, Li X, Trapnell C, Davidson S, Pachter L, Chu H, Tonkin L, Biggin M, Eisen M (2010) Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol* 8:e1000343
- Byrska-Bishop M, Evani U, Zhao X, Basile A, Abel H, Regier A, Corvelo A, Clarke W, Musunuri R, Nagulapalli K et al (2022) High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell* 185:3426–3440. e19
- Chavali S, Chavali PL, Chalancon G, deGroot NS, Gemayel R, Latysheva NS, Ing-Simmons E, Verstrepen KJ, Balaji S, Babu MM (2017) Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. *Nat Struct Mol Biol* 24:765–777
- Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890

- Cook D, Andersen E (2017) VCF-kit: assorted utilities for the variant call format. *Bioinformatics* 33:1581–1582
- Cummings CJ, Zoghbi HY (2000) Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum Mol Genet* 9:909–16
- DePristo MA, Zilversmit MM, Hartl DL (2006) On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene* 378:19–30
- Dickson Z, Golding G (2022) Low complexity regions in mammalian proteins are associated with low protein abundance and high transcript abundance. *Mol Biol Evol* 39:mca087
- Dieringer D, Schlotterer C (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Res* 13:2242–2251
- Dobin A, Davis C, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras T (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21
- Dosztányi Z, Csizmók V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 347:827–839
- Ebert P, Audano P, Zhu Q, Rodríguez-Martin B, Porubsky D, Bonder M, Sulovari A, Ebler J, Zhou W, SerraMari R et al (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372:abf7177
- Enright J, Dickson Z, Golding G (2023) Low complexity regions in proteins and DNA are poorly correlated. *Mol Biol Evol* 40:msad084
- Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Biol* 20:406–416
- Fomicheva A, Ross E (2021) From prions to stress granules: defining the compositional features of prion-like domains that promote different types of assemblies. *Int J Mol Sci* 22:1251
- Golding GB (1999) Simple sequence is abundant in eukaryotic proteins. *Protein Sci* 8:1358–61
- Gonzalez CE, Roberts P, Ostermeier M (2019) Fitness effects of single amino acid insertions and deletions in tem-I beta-lactamase. *J Mol Biol* 431:2320–2330
- Goolsby E (2017) Rapid maximum likelihood ancestral state reconstruction of continuous characters: a rerooting-free algorithm. *Ecol Evol* 7:2791–2797
- Grimwood J, Gordon L, Olsen A, Terry A, Schmutz J, Lamerdin J, Hellsten U, Goodstein D, Couronne O, Tran-Gyamfi M et al (2004) The DNA sequence and biology of human chromosome 19. *Nature* 428:529–535
- Haba Y, Kutsukake N (2019) A multivariate phylogenetic comparative method incorporating a flexible function between discrete and continuous traits. *Evol Ecol* 33:751–768
- Haerty W, Golding G (2010) Low-complexity sequences and single amino acid repeats: not just “junk” peptide sequences. *Genome* 53:753–762
- Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109
- He Q, Bardet A, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J (2011) High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat Genet* 43:414–420
- Holst L (1980) On the lengths of the pieces of a stick broken at random. *J Appl Probab* 17:623–634
- Horton C, Alexandari A, Hayes M, Marklund E, Schaepe J, Aditham A, Shah N, Suzuki P, Shrikumar A, Afek A et al (2023) Short tandem repeats bind transcription factors to tune eukaryotic gene expression. *Science* 381:eadd1250
- Huntley M, Golding G (2000) Evolution of simple sequence in proteins. *J Mol Evol* 51:131–140
- Huntley M, Golding G (2002) Simple sequences are rare in the protein data bank. *Proteins* 48:134–140
- Huntley M, Golding G (2006) Selection and slippage creating serine homopolymers. *Mol Biol Evol* 23:2017–2025
- Huntley MA, Golding GB (2006) Selection and slippage creating serine homopolymers. *Mol Biol Evol* 23:2017–2025
- Karlin S, Brocchieri L, Bergman A, Mrázek J, Gentles AJ (2002) Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci* 99:333–338
- Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30:772–80
- Kiefer J (1953) Sequential minimax search for a maximum. *Proc Am Math Soc* 4:502–506
- Kruglyak S, Durrett R, Schug M, Aquadro C (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci USA* 95:10774–10778
- Lenz C, Haerty W, Golding GB (2014) Increased substitution rates surrounding low-complexity regions within primate proteins. *Genome Biol Evol* 6:655–65
- Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754–1760
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079
- Lin M, Whitmire S, Chen J, Farrel A, Shi X, Jt Guo (2017) Effects of short indels on protein structure and function in human genomes. *Sci Rep* 7:9313
- Love M, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550
- Loya T, O’Rourke T, Reines D (2017) The hnRNP-like Nab3 termination factor can employ heterologous prion-like domains in place of its own essential low complexity domain. *PLoS ONE* 12:e0186187
- Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci U S A* 100:15324–15328
- Martin E, Mittag T (2018) Relationship of sequence and phase separation in protein low-complexity regions. *Biochemistry* 57:2478–2487
- McGinnis S, Madden T (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 32:W20–5
- Mier P, Alanis-Lobato G, Andrade-Navarro MA (2017) Context characterization of amino acid homorepeats using evolution, position, and order. *Proteins* 85:709–719
- Minh B, Schmidt H, Chernomor O, Schrempf D, Woodhams M, von Haeseler A, Lanfear R (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the Genomic Era. *Mol Biol Evol* 37:1530–1534
- Ni X, Zhang Y, Negre N, Chen S, Long M, White K (2012) Adaptive evolution and the birth of CTCF binding sites in the *Drosophila* genome. *PLoS Biol* 10:e1001420
- Odom D, Dowell R, Jacobsen E, Gordon W, Danford T, MacIsaac K, Rolfe P, Conboy C, Gifford D, Fraenkel E (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* 39:730–732
- Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931
- Parry D, North A (1998) Hard alpha-keratin intermediate filament chains: substructure of the N- and C-terminal domains and the predicted structure and function of the C-terminal domains of type I and type II chains. *J Struct Biol* 122:67–75

- Persi E, Wolf Y, Karamycheva S, Makarova K, Koonin E (2023) Compensatory relationship between low-complexity regions and gene paralogy in the evolution of prokaryotes. *Proc Natl Acad Sci USA* 120:e2300154120
- Persikov A, Ramshaw J, Kirkpatrick A, Brodsky B (2000) Amino acid propensities for the collagen triple-helix. *Biochemistry* 39:14960–14967
- Perteu M, Perteu G, Antonescu C, Chang T, Mendell J, Salzberg S (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 33:290–295
- Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* 16:1791–1798
- R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
- Revell LJ (2012) Phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 3:217–223
- Rohlf R, Harrigan P, Nielsen R (2014) Modeling gene expression evolution with an extended Ornstein–Uhlenbeck process accounting for within-species variation. *Mol Biol Evol* 31:201–211
- Romero P, Obradovic Z, Li X, Garner E, Brown C, Dunker A (2001) Sequence complexity of disordered protein. *Proteins* 42:38–48
- Sainudiin R, Durrett R, Aquadro C, Nielsen R (2004) Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics* 168:383–395
- Schmon S, Gagnon P (2022) Optimal scaling of random walk Metropolis algorithms using Bayesian large-sample asymptotics. *Stat Comput* 32:28
- Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, et al. 2016. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *bioRxiv* <https://www.biorxiv.org/content/early/2016/08/30/072116>
- Sequencing C, Consortium A (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87
- Shen W, Ren H (2021) Taxonkit: a practical and efficient ncbi taxonomy toolkit. *J Genet Genomics* 48:844–850
- Shi J, Rabosky D (2015) Speciation dynamics during the global radiation of extant bats. *Evolution* 69:1528–1545
- Shumate A, Salzberg S (2021) Liftoff: accurate mapping of gene annotations. *Bioinformatics* 37:1639–1643
- Stajich J, Block D, Boulez K, Brenner S, Chervitz S, Dagdigian C, Fuellen G, Gilbert J, Korf I, Lapp H et al (2002) The bioperl toolkit: Perl modules for the life sciences. *Genome Res* 12:1611–1618
- Vats D, Flegal JM, Jones GL. (2017). Multivariate output analysis for Markov chain Monte Carlo. [arXiv:1512.07713](https://arxiv.org/abs/1512.07713)
- Villar D, Flicek P, Odom D (2014) Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat Rev Genet* 15:221–233
- Wall L, Christiansen T, Orwant J. 2000. Programming perl. " O'Reilly Media, Inc."
- Werner M, Sieriebriennikov B, Prabh N, Loschko T, Lanz C, Sommer R (2018) Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Res* 28:1675–1687
- Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Computers Chem* 17:149–163
- Zhou K, Shi H, Lyu R, Wylder A, Matuszek Z, Pan J, He C, Parisien M, Pan T (2019) Regulation of co-transcriptional pre-mRNA splicing by m(6)A through the low-complexity protein hnRNPG. *Mol Cell* 76:70–81.e9

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.