Bioinformatic Applications in Protein Low Complexity Regions and Targeted Metagenomics

BIOINFORMATIC APPLICATIONS IN PROTEIN LOW COMPLEXITY REGIONS AND TARGETED METAGENOMICS

By Zachery William DICKSON, Bachelor of Technology

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

McMaster University
Doctor of Philosophy (2023)
Hamilton, Ontario (Department of Biology)

TITLE: Bioinformatic Applications in Protein Low Complexity Regions and Targeted Metagenomics
AUTHOR: Zachery William DICKSON Bachelor of Technology (McMaster University)
SUPERVISOR: Dr. Brian GOLDING
NUMBER OF PAGES: xvi, 220

Lay Abstract

This thesis describes research in two fields: repetitive protein sequences and methods for sequencing the portions of a sample in which one is most interested. In the first part I describe the general properties of repetitive proteins, establish a connection between the presence of repeats in a protein and the amount of that protein which a cell maintains, and show that these two quantities evolve together. This informs our understanding of evolution and regulation with implications for repeat related diseases and further evolutionary research. In the second part I describe a method for selecting short nucleotide sequences which can be used to capture specifically the DNA of organisms of interest, as well as applications of this and other methods. These contributions are widely applicable as targeted sequencing is useful in fields as far apart as clinical sepsis diagnosis and determining the colour of ancient animals.

Abstract

Part I: Low complexity regions (LCRs) are common motifs in eukaryotic proteins, despite the fact that they are also mutationally unstable. For LCRs to be widely used and tolerated there must be regulatory mechanisms which compensate for their presence. I have endeavored to characterize the relationships and co-evolution of LCRs with the abundance of the proteins that host them as well as the transcripts which encode them. As the abundance of a gene product is ultimately responsible for its associated phenotype, any relationships have implications for the many neurodegenerative diseases associated with LCR expansion. I found that there are indeed relationships. LCRs are more associated with low abundance proteins, but the opposite is true at the RNA level: LCRs encoding transcripts have higher abundance. Investigating the co-evolution of LCRs and transcript abundance revealed that on short evolutionary timescales indels in LCRs influence the selective pressures on TAb. Viewing LCRs through the previously unexplored lens of abundance has generated new results. Results which, together with explorations of information flow and low-complexity in untranslated regions, expand our knowledge of the functional impacts of LCRs evolution.

Part II: A commonly encountered problem in DNA sequencing is a situation where the DNA of interest makes up a small proportion of the DNA in a sample. This challenge can be compounded when the DNA of interest may come from many different organisms. Targeted metagenomics is a set of techniques which aim to bias sequencing results towards the DNA of interest. Many of these techniques rely on carefully designed probes which are specific to targets of interest. I have developed a bioinformatic tool, `HUBDesign`, to design oligonucleotide probes to capture identifying sequences from a given set of targets of interest. Using `HUBDesign`, and other methods, I have contributed to projects ranging in context from clinical to ancient DNA.

Acknowledgements

I would first like to thank my supervisor, Dr. Brian GOLDING, for giving me the opportunity to join his lab. Our first conversation was *after* I had applied to graduate school and indicated I wanted him as my supervisor! Thank you, Brian, for giving me a chance, and being a fantastic mentor. You've fostered an environment that encourages curiosity and allows for exploration, while always being able to gently prevent me from floating off into the clouds.

I would also like to thank the members of my supervisory committee, Ben Evans and Elizabeth Weretilnyk. The former for all of his insightful questions: He may forget, but Chapter 5 stems wholly from a question at my first committee meeting. Elizabeth, I thank you for always drawing connections to other fields and making sure I explained myself clearly.

A thank you for Hendrik Poinar, who has been a deep wellspring of fruitful collaborations. It has been a pleasure to aid in so many research projects. I would also like to thank all of his students who often had to do the work of validating my computational results.

It has been a joy to be a part of the community of graduate students in the department of biology. Without this community I would never have left the quiet computer lab. I would like to specifically thank the rotating cast of friends at board game and trivia nights, and especially TEAM2023 for forcing me to wake up and write!

Thank you, Bia, I hope your investment pays off!

Contents

List of Figures

List of Tables

Acronyms

ABC approximate Bayesian calculation

AIC Akaike information criteria

CMER conserved minimum entropy region

`fLPS` fast low probability sequences

`GBA` graph based algorithm

IGSR International Genome Sample Resource

LC low-complexity

LCA lowest common ancestor

LCR low complexity region

MAD mean absolute deviation

MCMC Markov Chain Monte Carlo

mESS multivariate effective sample size

nTE normalized translation efficiency

OU Ornstein–Uhlenbeck

PAb protein abundance

PCR polymerase chain reaction

SNP single nucleotide polymorphism

TAb transcript abundance

TWnTE time weighted normalized translation efficiency

ULCM untranslated low-complexity motifs

UTR untranslated region

Declaration of Authorship

I, Zachery William DICKSON, declare that this thesis titled, "Bioinformatic Applications in Protein Low Complexity Regions and Targeted Metagenomics" and the work presented in it are my own. I confirm that:

- Chapter 1 is my own work with editorial feedback from Dr. Brian GOLDING,
- Chapter 2 is my own writing with conceptual and editorial feedback from Dr. Brian GOLDING,
- Chapter 3 is a published work to which I contributed significantly, as detailed in the chapter's preface,
- Chapter 4 is my own published work with contribution detailed in the chapter's preface,
- Chapter 5 is my own work aided by undergraduates as detailed in the chapter's preface,
- Chapter 6 is my own in-review work with contributions detailed in the chapter's preface,
- Chapter 7 is my own work with editorial feedback from Dr. Brian GOLDING,
- Chapter 8 is my own published work with contributions detailed in the chapter's preface, and
- Chapter 9 is a summary of work I have performed on seven additional projects with contributions detailed for each project.

Preamble

As my graduate career has been rooted in pure bioinformatics, I was not limited to data that I can generate. Rather I was able to explore a few areas of biological sciences. Unfortunately, given enough time and effort on diverging branches of research, one might look back and realize that the projects have little overlap, much like speciation on the tree of life. Two highly diverged topics form the two parts of my thesis. Both have been rewarding and pleasurable challenges, as evidenced by my inability to let either of them go. Part I details my exploration of LCRs and the impact they have on the expression, abundance, and evolution of the proteins which host them. In Part II, I describe the work I have done in the development of tools for targeted metagenomic sequencing. The second part is intended to be a matter of record for the work I have done over the course of my graduate career.

Early on, while still shaping my research on LCRs, I was asked to look into a program called BOND and its applications. Years of development later, a heavily modified version became the core of a family of probe design pipelines. The applications of which range widely from detecting parasites in the sediment from an ancient village to rapidly identifying potential pathogens in sepsis patients.

Throughout the work developing tools for metagenomic sequencing, I continued my branch of research on LCRs in proteins. The research program was done with the support of and in collaboration with a series of undergraduate volunteers and thesis students. The main trunk of that research had multiple branches from the flow of biological information to the temporal order of events in the evolution of gene expression and sequences.

These two research branches now appear to be completely distinct, tied together by their membership in the vast class of bioinformatic research. However, at the core both share a particular kind of problem that fascinates me. How to decompose multiple overlaid signals, cut through the noise, and reveal the nuggets of truth. The evolutionary landscape of LCRs is highly multidimensional, nucleotide and amino acid composition, secondary structure, binding partners, selection, and more all exert pressures. Teasing these apart to find the relationships between sequence and expression evolution has been a fulfilling challenge. Likewise, targeted metagenomics is centered on amplifying signals of interest to ease the challenge of disentangling noise from signal.

Either topic alone may be enough to constitute a thesis on its own, but together they better demonstrate how I have contributed and hope to continue contributing to the field of biology through bioinformatics.

Part I

Protein Low Complexity Regions

Chapter 1

Introduction

I was first introduced to low complexity regions (LCRs) while employed in the biotechnology sector. I was looking to return to academia and was researching potential bioinformatics research programs. I found a paper by Lenz et al. (2014) exploring increased mutation rates both in and around LCRs. I found the idea of such simple sequence features being associated with long term evolutionary dynamics fascinating as it played into two of my favourite ways of understanding biology at the molecular level. Both a beautiful dance of mechanical automation, and a concerted mixture of influences both large and small: noisy but with signals still discernible. My interest in the topic may have been helped along by discovering I realized my newly hired co-worker was the first author of the paper I had just read!

The instability of LCRs extends beyond the point mutations that Lenz et al. (2014) described. LCRs are also liable to rapid expansion and contraction via replication slippage (Huntley and Golding 2006) and unequal crossing over (DePristo et al. 2006). These indels can be pathological, with LCRs being associated with several neurodegenerative diseases (Cummings and Zoghbi 2000; Day and Ranum 2005; Verstrepen et al. 2005; Musova et al. 2009). Structurally, LCRs are also often intrinsically disordered (Romero et al. 2001; Dosztányi et al. 2006): having no defined globular structure under physiological conditions. The lack of structure lends itself to another property of promiscuity. Many LCR containing proteins are not limited to specific partners (Dosztányi et al. 2006; Ekman et al. 2006; Coletta et al. 2010; Fomicheva and Ross 2021). Despite the discordance these sequences seem to bring to the molecular ballet, they seem to be tolerated by selection. Indeed, LCRs are extremely common in eukaryotic proteins (Golding 1999; Huntley and Golding 2000; Karlin et al. 2002). The mutational instabilities and flexible binding give them uses on both evolutionary and physiological timescales. They play roles in the hubs of interaction networks (Dosztányi et al. 2006); participate in transient organelles (Kedersha and Anderson 2002; Kato and McKnight 2017; Fomicheva and Ross 2021); and can also serve as the raw materials on which evolution can act (Radó-Trilla et al. 2015; Persi et al. 2023).

Evolutionary biology is an undirected performance. Each player on the molecular stage acts according to its sequence defined form, and a sequence ensemble which forms a sufficiently cohesive whole is allowed to keep performing. We observe LCRs to be chaotic in their form and function, but with important roles to play. Furthering our understanding of the balancing act between LCR costs and benefits has been the focus of my research. This has been primarily

through the lens of abundance. The cellular machinery that is most heavily invested in tends to be the most important, sensitive to disruption, and therefore most highly conserved (Drummond et al. 2005). The abundance of proteins is highly regulated with a multitude of overlapping effects. It offers a medium through which to explore the impacts of LCR beyond changes in DNA sequence alone. How does the presence of LCR affect the abundance of proteins? Are there biases in which proteins are allowed to have LCR? What mechanisms exist to regulate LCR and abundance co-evolution?

In this part of my thesis, I describe my work to explore and understand the roles played by LCRs on the stage of biology. The range of properties that LCR can possess is wide presenting a challenge to concretely identifying them. In Chapter 2 I explore several approaches for identifying LCRs in the process demonstrating that no one method is completely satisfactory. For the questions I ultimately wished to explore, a definition with a low false-positive rate is preferable. Every 'normal' protein falsely included dilutes any effects which may be seen. With this in mind the definition chosen was based on information content. This low-entropy definition proves useful in Chapter 3 where I contributed to a tracing of the flow of information from DNA to protein via RNA. We found that this flow is not as simple as one might naïvely think, pointing to evolutionary forces beyond neutral selection influencing LCRs. As these motifs are involved in many processes it may be unsurprising that their evolution is intertwined with other properties. In Chapter 4, I examined the relationship with both transcript abundance (TAb) and protein abundance (PAb) as properties which may be connected to LCRs and found that these were indeed connected but having contrasting impacts at the RNA and protein levels. This prompted exploration of untranslated regions (UTRs) in Chapter 5 and evolutionary history of these properties in Chapter 6. UTRs are hotbeds of signaling motifs where low-complexity (LC) may be playing a previously unrecognized role. Understanding the temporal order of TAb shifts and LCRs appearance is important for informing the mechanisms that drive the intertwining of these two properties. Ultimately my research is an exploration of the highly complex interactions LCRs allow for in molecular interactions and evolution.

Chapter 2

Methods for Identifying Low Complexity Regions

2.1 Introduction

Once thought of as junk or spacer regions between more 'important' globular regions in proteins, low complexity regions (LCRs) are loosely defined as compositionally biased sequences. As there was little interest in these regions the original goal in identifying LCRs was to mask them so that they would not interfere with attempts to align homologous proteins (Wootton 1994b). Since then, protein LCRs have been revealed to have important functions and effects at both the nucleotide and protein level in neurodegenerative disease (Cummings and Zoghbi 2000), protein-protein interaction (Dosztányi et al. 2006), and formation of membraneless organelles (Dignon et al. 2018).

As such, identifying LCRs has become much more important to a variety of fields in biology. As the roles that LCRs have been revealed to play have expanded, so too has the variety of properties that could be used to identify them. A more fulsome review of these properties will follow in Section 2.2, but they can be broadly divided into sequence and structural properties. Mier et al. (2020) published a review of several statistical methods for identifying LCRs, as well as a tool for categorizing proteins on the basis of amino acid bias and repetitiveness. This review of methods includes some statistical methods they include, but also looks at others and attempts to evaluate them with a common dataset.

2.2 Tools for LCR detection

2.2.1 Sequence methods

One of the earliest tools specifically intended for the identification of LCRs in protein sequences was SEG developed by Wootton and Federhen (1993). The algorithm relies on the concept of sequence entropy and probability to flag and refine LCR hits. The program has three parameters used in the initial flagging of low entropy sequences. These are the window length, the lower entropy bound (K_1), and the upper entropy bound (K_2). Sliding the window along a sequence, the Shannon entropy is calculated according to Equation (3.1); If this is lower than

K_1, a putative LCR is called as the sequence in the window. This sequence is extended on both sides if the window in that direction has an entropy higher than K_2. After extension, the LCR is refined by selecting the lowest probability subsequence of the putative sequence. Here the probability is calculated as the probability of observing the particular composition of residues, given an unbiased selection of any particular residue.

The major advantage of `Seg` is speed. For example, NCBI's `BLAST` (McGinnis and Madden 2004) uses `Seg` as a preprocessing step to mask LCRs (Fassler and Cooper 2011) in proteins. It is also able to rank the compositional bias of sequences; homopolymers will have zero entropy with the value rising with less bias. The method is also applicable to DNA, RNA, and protein alphabets, however the widely used implementation hard-codes the use of a protein alphabet and tolerates the DNA alphabet as the latter is a subset of the former. While this limitation has no impact on protein calculations, it can affect studies at other levels. Modification made to the original `Seg` algorithm to overcome this limitation are described in Chapter 3.

One of the limitations of `Seg` is that it completely ignores the order of residues in determining if they are low complexity. For example, QPPQQP would have the same entropy as QQQPPP and QPQPQP, despite the latter two having recognizable repeat patterns. Li and Kahveci (2006) developed graph based algorithm (`GBA`) in an attempt to account for this by modifying the score for a particular residue depending on those which precede it. `GBA` also attempts to incorporate amino acid properties with the assumption that a string of similar amino acids may also be a low complexity region. This is achieved with the use of a scoring matrix like BLOSUM62 to determine similarity between amino acids.

While `Seg` uses low probability to refine the LCR hits it discovers with entropy, it is possible to entirely focus on the probability of subsequences. LCRs are inherently low probability sequences as their compositional bias is distinct from 'normal' amino acid usage. Fast low probability sequences (`FLPS`) is a program developed by Harrison (2017) to identify compositional bias, with the claim of higher speeds than `Seg`. The program works by calculating the probability of subsequences up to some maximum size, using an amino acid usage table. It includes a pre-calculated table constructed based on a curated set of protein domains but can also use unbiased or user defined usage tables. A probability threshold is used to define compositional bias.

2.2.2 Structural methods

Aside from amino acid properties and compositional bias, the structural properties of proteins can be used to identify LCRs. Often, LCRs do not form fixed 3D structures at physiological conditions; they are intrinsically disordered (Romero et al. 2001). As structural prediction for proteins is a popular topic in computational biology there are a plethora of tools which could be used in this area, however the majority of them are gargantuan models, generally focused on globular protein prediction, making them slow, unwieldy, and ill-suited to identification of LCRs.

Two tools developed specifically for intrinsic disorder are `IUPred` (Dosztányi et al. 2005) and `SPINE-D` (Zhang et al. 2012). The former uses pairwise amino acid energy content to

differentiate between structured and unstructured regions, while the latter is a pipeline using position specific scoring matrices and treating each amino acid as having two half shell neighbourhoods (for example a surface exposed half and a protein embedded half) to predict the final 3D structure. For both of these programs the output is a value for each residue: the probability that this particular residue is in an intrinsically disordered region. To collapse to a defined LCR, a probability threshold and a minimum length can be used.

2.3 Materials and Methods

Given that LCRs have several intersecting, but not fully overlapping properties, any method which seeks to identify LCRs based on any one property will be imperfect. Additionally comparing between methods can be difficult simply because defining a test dataset with perfectly known LCR^+ and LCR^- proteins is impossible. However, for the purposes of reviewing the methods presented above, an approximation of such a dataset was constructed.

The Uniprot database (The UniProt Consortium 2017) contains SwissProt, a manually annotated database of proteins. The set of human SwissProt proteins which had an annotation for compositional bias was selected as a set of LCR^+ proteins. The set of LCR^- proteins was defined as human SwissProt proteins, which lacked an annotation for compositional bias and had 3D structure resolved by x-ray crystallography. The intrinsic disorder of LCRs is most likely to affect the success of x-ray crystallography. Uniprot uses `MobiDB-lite` (Necci et al. 2017) to generate compositional bias annotations, which is a tool for identifying intrinsically disordered regions. The greatest bias in this dataset should be towards methods based on intrinsic disorder.

These requirements resulted in a set of 31,020 LCR^+ proteins and 17,605 LCR^- proteins; or 64% LCR^+. However, human proteomes are nearer to 20% LCR^+ (Karlin et al. 2002). To have a dataset which better represented biological reality, a random sample of 1000 proteins was selected, 200 from the LCR^+ set and 800 for the LCR^- set. Each method was run on this combined set and evaluated for accuracy and runtime, the parameters for each method are described below.

Each identification method was compared using the programs default parameters, as well as the set of parameters which give the lowest error rates. These parameters can be found in Table 2.1, and a clarification of some parameter names follows. Forget rate is the reduction in influence a preceding residue has on the current residue as it is more distant. If the value were 0.5, the preceding residue would have half influence, and the next before it one quarter. AA usage describes how the table of amino acid usages was constructed. It can be either precalculated from a curated set of protein domains or sampled from the set of proteins being evaluated. Search mode describes the strategy of `IUPred` and `SPINE-D`. The former doesn't have a default strategy, instead it can search for globular protein domains, where LCRs are the remainder. Alternatively, it can search for long or short stretches of intrinsic disorder. `SPINE-D` can also be optimized for long stretches of disorder. As an additional note, the optimized parameters for `Seg` have been previously described (Huntley and Golding 2000, 2002, 2006; Haerty and Golding 2010; Lenz et al. 2014).

TABLE 2.1: Parameters used for comparing LCR identification methods

Tool		Parameters		
		Window Size	K_1	K_2
Seg	Default Case	12 residues	2.2 bits	2.5 bits
	Best Case	15 residues	1.9 bits	2.2 bits
		Window Size	Forget Rate	Max Indels
GBA	Default Case	15 residues	0.95	3
	Best Case	3 residues	0.85	0
		Maximum Length	Min Probability	AA Usage
fLPS	Default Case	500 residues	1×10^{-3}	precalculated
	Best Case	500 residues	1×10^{-9}	sampled
		Search Mode	Min Probability	Min Length
IUPred	Default Case	globular	50%	3 residues
	Best Case	long	75%	15 residues
		Search Mode	Min Probability	Min Length
SPINE-D	Default Case	default	50%	3 residues
	Best Case	long	75%	15 residues

As the compositional bias annotation in UniProt are assigned to particular amino acid intervals in proteins, it is possible to assess both the ability of tools to separate LCR^+ and LCR^- proteins and the ability to assign specific amino acids to low and high complexity regions. The assessment of both is divided into calculating the proportion of false positives and negatives. These can be summed to an overall error rate.

2.4 Results

As the number of proteins one wishes to analyze increases, the speed of a computational tool becomes increasingly important. When analyzing whole proteomes, the differences of seconds or minutes can become days. The runtime of the tools is recorded in Table 2.2. The fastest tool was Seg at less than a second regardless of the parameters used as it performs simple calculations. On the other hand, GBA was the slowest, taking up to 40 minutes to process 1000 proteins. When preprocessing time is taken into account, SPINE-D, which performs the most intense structural predictions, was the slowest program. It required 4 days to construct the data necessary for structural predictions.

The ability of each tool to correctly categorize proteins and amino acids into LCR^+ and LCR^- can be found in Figure 2.1. At the protein level, the default parameters of most of the tools were not ideal for accurately determining LCR status. The best was IUPred with an overall error rate of $32\pm3\%$ while the worst was GBA at $80\pm3\%$. In every default case the error rate was driven almost entirely by incorrectly identifying LCRs in LCR^- proteins (false positives), while

TABLE 2.2: Runtime of LCR Identification Methods

Tool	Time		
	Preprocessing	Default Case	Best Case
`Seg`	NA	0.23 s	0.23 s
`GBA`	NA	39 min	4 min
`IUPred`	NA	10.3 s	10.3 s
`SPINE-D`	4 days	2.6s	2.6s
`fLPS`	1.0 s	4.8 s	0.25 s

correctly categorizing most LCR$^+$ proteins (true positives). This remains true when we look at the best case parameters, except for `Seg` and `fLPS` where the errors are more balanced or lean towards false negatives. These two programs have similar performance as well with `Seg` at an overall error rate of $14 \pm 3\%$ and `fLPS` at $13 \pm 3\%$. At $47 \pm 4\%$, `GBA` with optimized parameters still had the worst performance even being outperformed by the defaults for `IUPred`. `IUPred` was also the most consistent with its default and best cases performing similarly. `FLPS` was the most sensitive to parameters with a 60% shift in error rate depending on parameters.

At the amino acid level, the overall results are similar with several particularities which differ (Figure 2.1b). Overall, the error rates are lower, as the predicted LCRs tend to be small compared to the length of the proteins which contain them. The worst performing tool was `fLPS` with default parameters which incorrectly placed $62.6 \pm 0.5\%$ of amino acids. It was again most sensitive to parameters, flipping to the best performing with optimized parameters at $10.3 \pm 0.5\%$. With the exception of `fLPS` with default parameters, failure to place amino acids inside LCRs was a much more common error. The least parameter sensitive method in this case was `Seg`.

2.5 Discussion

I have described a small sample of the tools available for detecting LCRs based on sequence and structural properties, as well as compare their performance on a common dataset. This dataset has some deficiencies for the task which stem from the lack of a consensus for how to identify LCRs. Without this consensus it is not possible to construct a dataset where the LCR status is perfectly known. Instead, the dataset will always be biased towards the tools used to construct it. In this case `MobiDB-lite` was used to generate the annotations used to classify proteins as LCR$^+$. This tool is based on the identification of intrinsic disorder and would be expected to bias the results towards other tools with this basis. This is especially true for `IUPred` as the papers describing the two tools (Dosztányi et al. 2006; Necci et al. 2017) actually share an author. This doesn't appear to be the case as the false negative rates are similar with optimized parameters between `IUPred`, `SPINE-D`, `Seg`, and `fLPS`. The selection of LCR$^-$ proteins had an extra, structure-based filter in that those proteins must also not have an X-ray crystallography structure. Huntley and Golding (2002) showed that the sequence property of

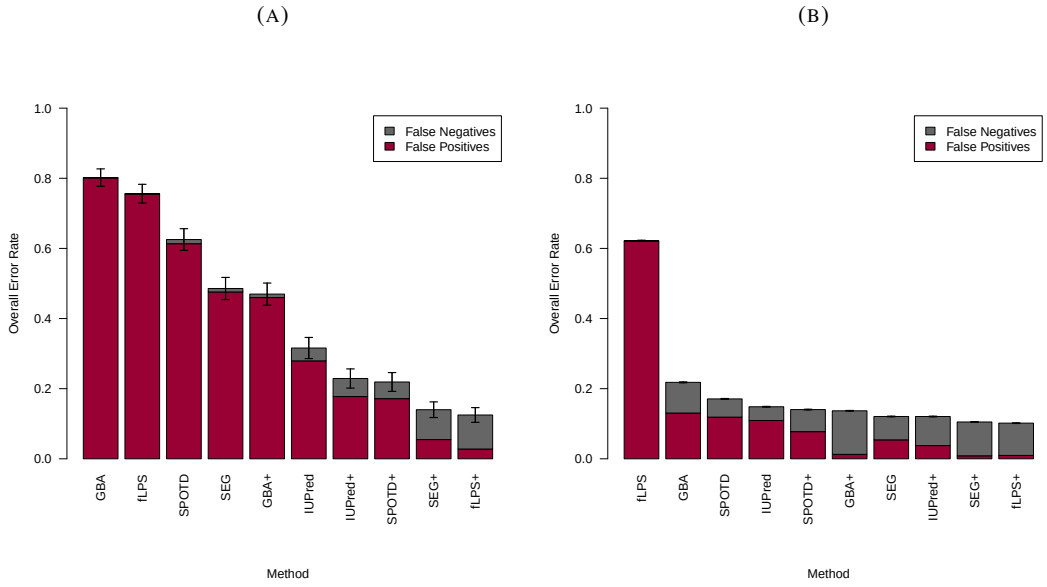

FIGURE 2.1: Probability and entropy based methods have lower error rates. A plus sign in the tool label indicates the use of best case parameters. Maroon, or the lower section of the bars indicates the proportion of proteins from the LCR$^-$ set incorrectly called as LCR$^+$, and grey (upper section) indicates the proportion of proteins from the LCR$^+$ set incorrectly called as LCR$^-$. a The error rates in categorization of 1000 proteins as LCR$^+$ and LCR$^-$. b The error rates in categorization of 460,902 amino acid residues from the 1000 proteins as being inside or outside of an LCR.

low entropy was associated with under-representation in a structural database. While imperfect, the dataset presented represents a reasonable common comparison for the methods.

In general, Figure 2.1 shows that when using optimized parameters `Seg` and `fLPS` have the best performance, and yet `GBA` another entropy based method was consistently the worst performer. The optimized parameters for all programs, except for `Seg`, were determined solely with the test data and may be over-fitted.

By examining differences between the protein and amino acid level performance we can gain insight into the types of mistakes each program makes. For example, `GBA` has very high false positive rates at the protein level, however at the amino acid level the overall error rate is lower and less biased towards false positives. As `GBA` predicts short LCRs in most proteins, it incorrectly assigns proteins to the LCR^+ category, but their relatively small size has little impact at the amino acid level. As `GBA` incorporates information on amino acid properties, it may be identifying regions where amino acids with similar properties are concentrated. The evolutionary forces bringing similar amino acids together are more likely to be due to selection for function and may be dissimilar to the forces creating repetitive amino acid tracts. The overall high false positive rates at the protein level imply an overzealousness in calling LCRs. When this is paired with higher false negative rates at the amino acid level, it is implied that even when a tool successfully calls a protein as LCR^+ it may be placing the LCR at the incorrect position within the protein.

Accepting that no tool is ever perfect, especially when what constitutes an LCR is up for debate, the important criteria are the ultimate goals of the study. If the goal is to characterize every LCR in a proteome, using a tool with a low false negative rate would be desirable. When attempting to correlate some other property with the presence of an LCR it is desirable to have the errors that do occur to be balanced between false positives and negatives. With a balance the effect of proteins near the boundary being incorrectly assigned is also balanced. This will increase the noise around any observed effect, but not the accuracy. Additionally, one must account for whether their property of interest is most affected at the structural or sequence level. If the former, a structural tool is most appropriate. For an analysis attempting to correlate the expression level of both mRNA and protein sequences with the presence of LCRs at the protein level, a balanced, low error rate program such as `Seg` is most appropriate.

Like Mier et al. (2020), our review has shown that no single tool is best at identifying LCRs, even in the case where the construction of the dataset may favour certain tools. There is significant room for improvement in this space, potentially in the form of tools which account for the evolutionary forces which spawn and maintain LCRs. With the current state of LCR identification tools it is imperative that researchers carefully consider which LCR definition and types of errors would be most appropriate for the questions they seek to answer.

Chapter 3

Correlating DNA and Protein sequence entropies

3.1 Preface

This chapter was published in Molecular Biology and Evolution in April, 2023 (https://doi.org/10.1093/molbev/msad084), with Johanna M Enright, Zachery W Dickson, and G Brian Golding as co-authors. JME wrote python scripts which parsed annotated genomes, called a modified version of `Seg` to identify low complexity regions (LCRs), calculated correlation coefficients, and simulated proteomes. JME also wrote the first rough draft of the manuscript, and passed it to me it to me for first revisions. JME and GBG contributed to later revisions. ZWD modified `Seg` to accept any alphabet and properly account for ambiguous residues in nucleotide sequences, devised approaches for simulations and interpretation of results, wrote first drafts of portions of the discussion, generated figures, as well as handled revisions and the publication process. GBG devised the initial project, edited the manuscript, and provided guidance to keep the project in a reasonable scope.

In this manuscript we set out to determine how well the entropy of DNA and protein sequences are correlated in LCRs. As information flows from DNA to protein the entropy of the two should be reasonably well correlated. However, we found that compared to simulations with some reasonable assumptions on LCR evolution the correlation was generally lower than expected, especially in LCRs with cryptic repeats. The impact of the work is in showing that evolutionary forces beyond neutral slippage, and retention of the amino acid sequence shape LCRs.

3.2 Introduction

Low complexity regions (LCRs) are segments of a protein or DNA sequence which are biased in composition (Wootton and Federhen 1993). LCRs can present as periodic repeats, ambiguous cryptic repeats, or can contain no apparent pattern at all, but simply deviate from a randomized composition (Tautz et al. 1986; Wootton 1994a). LCRs contain low information and have a low entropy (Wootton and Federhen 1993). Entropy, as measured by Shannon's Entropy equation (Shannon 1948), is a measure of compositional complexity which uses the

proportion of residue(s) in a subsequence to measure the compositional state of that subsequence (Wootton 1994b). A lower variety of residues would result in a lower entropy for that subsequence. Thus, minimal entropy would contain a subsequence consisting of only a single residue, whereas maximal entropy would contain all possible residues in an alphabet in equal proportions. In proteins, LCRs are typically composed of hydrophilic and small amino acid residues (Faux et al. 2005).

Interest in protein LCRs has grown in recent decades as involvement of LCRs in protein function and disease has been further illuminated. Due to a lack of motif conservation and a tendency to form non-globular protein domains, LCRs were once considered to be merely tolerated within their protein, offering no functional, biologic contribution (Huntley and Golding 2000, 2002). It is now believed that LCRs may offer a range of functions to various proteins, many which are linked to this non-globular, intrinsically disordered nature. Intrinsically disordered regions can allow for longer, more accessible protein domains, protein flexibility, and plasticity in molecular binding partners (Dosztányi et al. 2006; Ekman et al. 2006). As such, LCRs are often found in proteins involved in signaling pathways and can act as scaffolds in the formation of large protein complexes (Coletta et al. 2010; Dyson and Wright 2005). They are also enriched in transcription factors (Millard et al. 2020), developmental proteins (Huntley and Clark 2007) and can offer accessible regions for post-translational modifications (Jeronimo et al. 2016; Monahan et al. 2017).

As protein LCRs are ultimately the result of changes to the underlying DNA, their evolution is likely similar to that of inter-genic, noncoding DNA microsatellites (DePristo et al. 2006). Microsatellites are believed to evolve rapidly by expansion or contraction via two main mechanisms: the first and predominant mechanism being polymerase slippage, in which the DNA template and coding strand shift relative to one another and re-anneal with another repeat unit causing either insertion or deletion of a repeat unit (Levinson and Gutman 1987; Viguera et al. 2001). The second mechanism is unequal recombination which occurs via the misalignment of homologous repeat sequences during meiosis and results in the gain of repeats in the sequence of one chromosome and loss of repeats in other (Richard and Paques 2000). Other factors, including mismatch repair mechanisms (Levinson and Gutman 1987), repeat unit length (Schug et al. 1998), ability of the DNA to form structures (Dere et al. 2004; Moore et al. 1999; Murat et al. 2020), and repeat unit composition (Gragg et al. 2002) play a role in the rate of microsatellite slippage. Microsatellites above a certain threshold repeat length will undergo slippage and expand or contract, with longer microsatellites being more unstable and more likely to undergo slippage (Lai and Sun 2003). In coding regions, slippage of repeats whose units are multiples of three are more likely to be permitted compared to other repeat unit lengths because an insertion or deletion will not cause a frameshift mutation in the downstream coding sequence (Metzgar et al. 2000). Thus, they will be less likely to result in a deleterious mutation that will be selected against (Metzgar et al. 2000).

Codon homogeneity is also an important factor in LCR evolution. Because LCRs are believed to arise primarily via polymerase slippage, microsatellites of homogeneous codon runs are thought to be more unstable, evolve faster, and to be less conserved than sequences encoded by a heterogeneous mixture of synonymous codons (Albà et al. 1999). Over time, accumulation of

synonymous mutations in LCR coding regions can help to conserve the LCR by breaking up repeats and reducing the chance of slippage (Albà et al. 1999).

With important physiological roles and high mutation rates, it follows that LCRs have the potential for pathogenesis. In humans, one of the most notable examples is Huntington's disease (Everett and Wood 2004). Slippage of long tracts of CAG trinucleotide repeats result in an expanded polyglutamine tract. The mutant Huntingtin protein develops a toxic gain of function effect within the cell (Everett and Wood 2004). Other neurodegenerative diseases resulting from trinucleotide repeat expansion include spinocerebellar ataxia and muscular dystrophy, encoding polyglutamine and polyalanine tracts, respectively (Brown and Brown 2004; Everett and Wood 2004). As well, LCRs have been shown to contribute to antigenic variation and immune system evasion of human pathogens (Kebede et al. 2019; Velasco et al. 2013; Verstrepen et al. 2005).

Various studies have suggested that low entropy in nucleotide content correlates with LCRs in proteins. Li et al. (2015) showed how GC content constrains the types of amino acids which can be encoded, resulting in a bias towards amino acids encoded by codons with a high GC proportion and a bias against those with a lower GC proportion. The malaria parasite, *Plasmodium falciparum*, contains a high genomic AT content, which is strongly associated with the presence of protein LCRs, leading to preference for certain codons and amino acid types over others (DePristo et al. 2006). Xue and Forsdyke (2003) suggest that LCRs at the protein level are a result of AG content bias at the nucleotide level, and thus pressures at the DNA level can explain the presence of LCRs at the protein level. This was further supported by analyzing the nucleotide composition at first, second, and third codon bases, where AG content was higher in the first two, suggesting the importance of AG content to encode particular amino acids (Xue and Forsdyke 2003).

The structure of codons can significantly impact the entropy of the sequence they make up. Codons can be classified by the number of unique nucleotides in the codon, a property that we will refer to as nucleodicity. The DNA level entropy varies significantly among these codons: mononucleic codons such as AAA have an entropy of zero, dinucleic codons such as AGA have an entropy of 0.918, and trinucleic codons such as AGC have an entropy of 1.58. However, this property can only affect entropy at the DNA level as the information is lost when the codon is translated to a single amino acid. LCRs with codons of different nucleodicities may evolve differently as the repeats have variable abilities to form secondary structure during transcription and translation (Barik 2017). This property is thought to be an influencing factor in the likelihood of polymerase slippage (Murat et al. 2020).

Polymerase slippage, such as that seen in microsatellite expansion, is suggestive of a neutral model of evolution whereby the unstable LCR is merely tolerated within the protein so long as it does not impart deleterious effects (Radó-Trilla and Albà 2012). The LCR can then be preferentially retained if it confers a selective advantage. This is in contrast to a strict selective model of LCR evolution which maintains that LCRs within a protein are a result of selective pressures constraining the types and ordering of amino acids so as to create an amino acid motif which confers a particular function (Haerty and Golding 2010).

There have been multiple studies supporting both neutral evolution as well as selective evolution, with LCRs being created due to forces acting on the protein/amino acid level. Evidence for selective neutrality includes a large variation in LCR tract size both intra and interspecifically (Haerty and Golding 2010). Such high length polymorphic variability is also associated with a homogeneous codon tract and high slippage rates (Mularoni et al. 2007). Whereas conservation of LCR motifs and selective evolution may entail a heterogeneous codon tract of synonymous codons. That is, assuming the codons were a result of pressure from the protein level and not due to the degeneration of trinucleotide repeats (Albà et al. 1999; Huntley and Golding 2006). Neutral proteins could contain a high variation in repeat tract size as a result of unstable replicative slippage and also could undergo non-synonymous mutations which would be permitted due to the lack of purifying selection (Mularoni et al. 2007). However, increased repeat length has been observed to correspond with low non-synonymous mutation rate, suggesting the conservation of long LCRs (Mularoni et al. 2007). Studies showing synonymous mutations closer to LCRs have indicated that these regions may be evolutionarily conserved and hold functional significance (Lenz et al. 2014).

In this study, we have identified LCRs in proteins and assessed the correlation between their entropy and their corresponding DNA sequence entropy. We also identified LCRs in DNA sequences and comparing their entropy to that of their corresponding amino acid sequence. If the origin and evolution of LCRs were primarily a result of mutation acting at the DNA level via polymerase slippage and LCR expansion being allowed due to low selective constraints, we would expect to see a high correlation between protein entropy and its corresponding coding sequence entropy in LCRs. As DNA entropy decreased, codon types would be constrained thereby constraining and lowering the protein entropy as well as increasing chances of polymerase slippage for further LCR generation. If selection was the predominant mechanism by which LCRs were formed, we would expect to see a lower correlation between DNA and protein sequence entropy of corresponding sequences. This is because selection, unlike slippage, would not necessarily favour a homogeneous run of codons, but could instead allow a more random collection of synonymous codons for a particular amino acid residue. Ultimately, this would allow for a wider range of possible DNA entropies given a particular protein LCR.

3.3 Materials and Methods

All custom scripts and commands used in this analysis can be found on `GitHub` at `https://github.com/JohannaEnright/LCREntropyProject/`.

3.3.1 Sequence Data

Two correlation studies of sequence entropy were conducted. The first identified LCRs in proteins from the entire proteome of five model organisms *Saccharomyces cerevisiae*, *Homo sapiens*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. For each protein we identified LCRs and compared the entropy in the corresponding coding DNA sequence. The second study did the inverse; identifying LCRs in the coding DNA sequences and compared their entropy to the entropy of the amino sequence that they encode.

The genomes of the five organisms were downloaded from NCBI. Access dates and accession numbers are listed in Table A1.1.

Annotated sequences representing a haploid assembly for each organism were downloaded in genbank and fasta format. A custom python script was written to identify LCRs within coding sequences and locate their corresponding amino acid sequences and vice versa (Van Rossum and Drake 2009). LCRs were identified using the `Seg` algorithm (Wootton and Federhen 1993). Adjustments were made to `Seg` to account for alphabet size depending on the sequence type. Ambiguous characters were accounted for when identifying regions of low complexity by adding a fractional count to each residue represented by the ambiguous character. When searching for LCRs within proteins, `Seg` parameters were set to a window length (W) of 15, a trigger complexity (K1) of 1.9, and an extension complexity (K2) of 2.2 (see Table A1.2 for alternate parameters examined). To identify LCRs in coding sequences, parameters were set to 45 for W (three times the length used for amino acids) 1.3 for K1, and 1.5 for K2 (see Table A1.3 for alternate parameters examined).

A non-redundant set of coding sequences was selected by retaining only the longest isoform for each gene. This was done to reduce redundancy introduced by splice variants, duplicate genes in the pseudo-autosomal regions of the X and Y chromosomes, and duplicate genes present in alternate assemblies. If isoforms were the same length, the one which mapped to a chromosome was chosen over that from an alternate assembly. As well, an X chromosome isoform was chosen over a Y chromosome isoform. In addition, any coding sequences which encoded only a portion of a final protein product, such as immunoglobulin gene segments, or sequences which were not exactly three times the length of the amino acid sequence, were excluded as a direct mapping between amino acid sequence and coding sequence could not be made. For later simulations, the codon frequency, protein length, and proportion of proteins containing LCRs were calculated using this set.

For LCR analysis, only the longest LCR from each isoform was taken. We have observed in human data that less than 10% of LCR containing proteins have multiple LCRs, and the composition of LCRs tends to be similar within the same protein. Taking the longest allows for a simpler analysis with one signal per protein. If multiple LCRs from a single sequence were the same length, the one with the lowest entropy was chosen. Once all LCRs were obtained, the entropy of the corresponding amino acid or DNA sequence was calculated using the Shannon's Entropy equation (Shannon 1948):

$$H = \sum_{i=1}^{n} p_i \log_2 p_i \tag{3.1}$$

where p_i refers to the proportion of each unique letters in a sequence and n refers to the total number of unique letters (Equation (3.1)).

While calculating entropy, ambiguous characters were handled in the same manner as above. DNA LCRs were trimmed at the ends to ensure a direct correspondence between a codon and its amino acid. These end adjustments were taken into account before determining the longest LCR from a coding sequence.

Scatter plots were created for each organism and set of parameters. A linear regression and correlation coefficient calculation was performed for each plot in R (R Core Team 2022).

For the purpose of later simulations, the codons in the LCRs were classified by nucleodicity, the number of unique nucleotides in a codon. The nucleodicity of the LCR as whole was set to match the most frequent nucleodicity class amongst its codons. For example, an LCR made up of 6 AAA, 2 ATC, and 1 GCA would be assigned a nucleodicity of 1. In the case of a tie for the most frequent nucleodicity, an LCR would be partially assigned all tied nucleodicities. For example, an LCR with equal counts of mono-, di-, and tri- nucleic codons would be counted as one third for each class. Using these potentially partial counts the number of observed LCRs of each nucleodicity was counted. The number of expected LCRs for each nucleodicity class was calculated based on the codon usage for each organism by multiplying the total frequency for each class by the total number of observed LCRs. As an example, with completely unbiased codon usage, 6.6% ($\frac{4}{61}$) of LCRs would be expected to be mononucleic, 57% ($\frac{35}{61}$) dinucleic, and 36% ($\frac{22}{61}$) trinucleic. The actual proportions vary between species based their codon usage. The significance of differences between observed and expected numbers of LCRs in each class was evaluated using a chi squared test. A preference coefficient was calculated for each nucleodicity class to represent the observed preference for a codon class relative to the expected value. The coefficients are normalized relative to the most preferred codon class and are calculated as

$$P_{i,j} = \frac{O_{i,j}/E_{i,j}}{\max_k(O_{i,k}/E_{i,k})} \tag{3.2}$$

where P is the preference coefficient, O is the observed number of LCRs, E is the expected number of LCRs, organisms are indexed by i organism, and the number of unique codons in a codon class is indexed by j and k.

3.3.2 LCR Simulations

It was critical to have null expectations to which to compare the biologically observed values for entropy and correlation. To that end, several simulations were implemented with the python language (Van Rossum and Drake 2009). Simulations were performed separately for each organism studied and for several models of evolution. In each simulation, a set of 100000 equal length coding sequences were generated according to the relevant evolutionary model as well as the organism's codon usage. Each coding sequence had as many codons as the organism's average protein length (n) and a stop codon, for a total of $n+1$ codons. The evolutionary models considered are intended to simulate varying levels of replication slippage, codon nucleodicity class, and substitution. Overall, five models were used: Null, Slip, Slip + CC, Slip + Syn, and Slip + CC + Syn.

The first model, Null, is the simplest and is intended as a naïve model. Each coding sequence was constructed by randomly sampling from the 61 amino acid encoding codons. Each codon had an equal probability of being sampled. This model does not generate significant numbers of LCRs.

In the Slip model, codons are randomly selected according to each specific organism's genomic codon bias. However, once the same codon has been sampled at least twice the weighting for that codon is increased. As a result, runs of identical codons which encode LCRs are more likely. This elevated weighting is maintained until a different codon is selected at which point it returns to using the original genomic codon bias. The amount by which the probability is increased, the 'slope', is dynamically set such that the overall proportion of LCR containing proteins in the generated proteome matches that observed in the organism. The increase in weight is applied each time the codon is consecutively sampled. As a result, the probability of slippage increases linearly with the length of the LCR. Slippage on the basis of codons was used instead of nucleotide based slippage as there is strong selection against frameshift mutations in coding sequences (Metzgar et al. 2000).

There may be biological preferences for or against runs of identical codons in each nucleodicity class. Hence, the Slip+CC model generates sequences according to the Slip model and has a later additional step which attempts to mimic the species specific use of nucleodicity class. After a protein was constructed, according to the Slip model, `Seg` is used to identify any LCRs. The LCR is classified by its nucleodicity and is retained with a probability equal to the organism specific preference coefficient (See Equation (3.2)). If a protein is not retained, a new protein is generated. This process continues until a proteome of 100000 proteins with the same proportion of LCRs as observed biologically is constructed. In addition, this will generate a proteome which has LCRs in each nucleodicity class with the same proportions as is biologically observed.

As a final step, subsequent synonymous substitutions maintaining the amino acid sequence were simulated for the Slip model (Slip+Syn), and the Slip+CC model (Slip+CC+Syn). This was implemented by randomly selecting a codon within the previously simulated sequence. Any codon in the artificial protein could be selected, regardless of inclusion in an LCR. Then the first or third position nucleotide was randomly selected and randomly changed to any of the three other nucleotides with equal weight. This change was only accepted if the resulting codon was synonymous with the original. This process was repeated until 1000 accepted synonymous mutations were made. Each attempt was completely independent of any previous iterations, therefore potential mutation sites were sampled with replacement. Differences in probability between transitions and transversions were not explicitly accounted for, however the nature of the genetic code forces more synonymous transitions than synonymous transversions (Koonin and Novozhilov 2009) since transitions are more likely to be synonymous than transversions. The process of adding 1000 synonymous mutations was repeated for each of the proteins in the simulated proteome.

For all simulations, the same python script and `Seg` parameters described in the previous section were used to identify protein LCRs, calculate their entropy, and calculate the entropy of their corresponding coding sequences. The same was done for LCRs within coding sequences. See Tables A1.4 to A1.8 for alternate parameters examined in the Null, Slip, and Slip+Syn simulations. Each pair of entropy values were plotted and a linear regression and correlation coefficient were calculated (R Core Team 2022).

3.3.3 Confidence Intervals for Correlation Coefficient

To determine if the entropy correlations were significantly different, 95% confidence intervals ($\alpha = 0.05$) for the correlation coefficient were calculated. A Fisher transformation (Equation (3.3)) was first performed on the r values to improve normality with increasing sample size (David Shen 2006). The lower and upper confidence limits were then calculated (Equation (3.4)) and these limits were transformed back (Equation (3.5)) (David Shen 2006).

Calculations were performed using the following equations:

$$f_r = 0.5 \ln\left(\frac{1+r}{1-r}\right) \tag{3.3}$$

$$\begin{aligned} \zeta_l &= f_r - z_{(1-\alpha/2)}\sqrt{\frac{1}{n-3}} \\ \zeta_u &= f_r + z_{(1-\alpha/2)}\sqrt{\frac{1}{n-3}} \end{aligned} \tag{3.4}$$

$$\begin{aligned} r_l &= \tanh(\zeta_l) \\ r_u &= \tanh(\zeta_u) \end{aligned} \tag{3.5}$$

3.3.4 Identifying LCRs in Codons

LCRs at the codon level were identified and compared against the entropy of their encoded protein sequences. `Seg` was modified to be able to use an alphabet with 61 letters. `Seg` parameters used to identify codon LCRs were 15 for W, 2.5 for K1, and 2.9 for K2. All additional steps were performed as in the previous sections.

3.3.5 Identifying Periodic Repeats in LCRs

Mono-, di-, and tri- periodic repeats were identified at the protein level from the previously determined protein and DNA LCRs using a custom python script. Minimal repeat lengths for mono-, di-, and tri- repeats were 6, 5, and 4, respectively. LCRs which contained one or more of the three repeat types were classified as periodic LCRs, whereas LCRs which did not contain any of the three repeat types were classified as cryptic LCRs. Minimal repeat length parameters were varied to ensure consistent trends in sequence correlation for periodic repeats. Results for periodic LCRs and cryptic LCRs of alternate repeat lengths can be found in Tables A1.9 and A1.10.

3.4 Results

3.4.1 Entropy of LCRs in Protein and DNA Correlate Poorly with Corresponding Sequence Entropies

Protein and DNA sequence entropy comparisons were performed on the genome and proteome of five model organisms *Saccharomyces cerevisiae*, *Homo sapiens*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. In general, we observed a low correlation between corresponding sequence entropies when LCRs were identified in both coding regions and proteins. This lack of correlation was observed for all organisms, all of which had correlation coefficients at or below $r = 0.579$ for both LCR types. The low correlation suggests that DNA LCRs can encode a variety of amino acid sequence compositional complexities and that protein LCRs can be encoded by a mixture of nucleotide complexities and by a heterogeneous mixture of synonymous codons. To avoid being redundant, only the results from *S. cerevisiae* and *H. sapiens* will be described in detail. Corresponding results and figures for *A. thaliana*, *C. elegans*, and *D. melanogaster* can be found in Figures A1.1 to A1.3.

Despite the consistently low correlation between sequence entropies, there were significant differences in correlation coefficient values between LCR sequence types in some, but not all organisms. For example, in *S. cerevisiae*, sequence entropy comparisons in protein LCRs yield a correlation coefficient of $r = 0.488$ (95% CI: $0.435 - 0.538$; Figure 3.1a). The correlation between DNA LCRs and their corresponding protein sequences in *S. cerevisiae* was lower than for protein LCRs, although not significantly ($r = 0.428$, 95% CI: $0.281 - 0.555$; Figure 3.2a). In *H. sapiens*, the correlation coefficient for sequences entropies between protein LCRs and DNA was $r = 0.374$ (95% CI: $0.347 - 0.400$; Figure 3.1b). However, the correlation coefficient between DNA LCRs and their corresponding amino acid sequences was significantly higher at $r = 0.579$ (95% CI: $0.541 - 0.614$; Figure 3.2b). A summary of results for all five organisms can be found in Table A1.11.

Next, we examined the general trends in the entropy distributions. The majority of protein and DNA LCRs have entropies near or within the low and high cut `Seg` parameter values and quickly taper off as they near more extreme entropies. Protein LCRs are encoded predominantly by higher entropy coding sequences with very few being encoded by low entropy coding sequences (Figure 3.1). Comparatively, DNA LCRs typically encode relatively midrange entropy protein sequences and are more evenly distributed within the possible protein sequence entropy range (Figure 3.2). This indicates that low entropy DNA sequences encode comparatively lower entropy protein sequences whereas, low entropy protein sequences can still be encoded by relatively high entropy DNA sequences. At the extremes of the distribution, a vertical line at a protein entropy of 0 was observed in protein and DNA LCRs of both species (Figures 3.1 and 3.2). This line corresponds to homopeptide repeats of a single amino acid residue which was evidently encoded by codons with various nucleotide compositions as well as a potential mix of heterogeneous, synonymous codons, hence the wide range of corresponding DNA entropies.

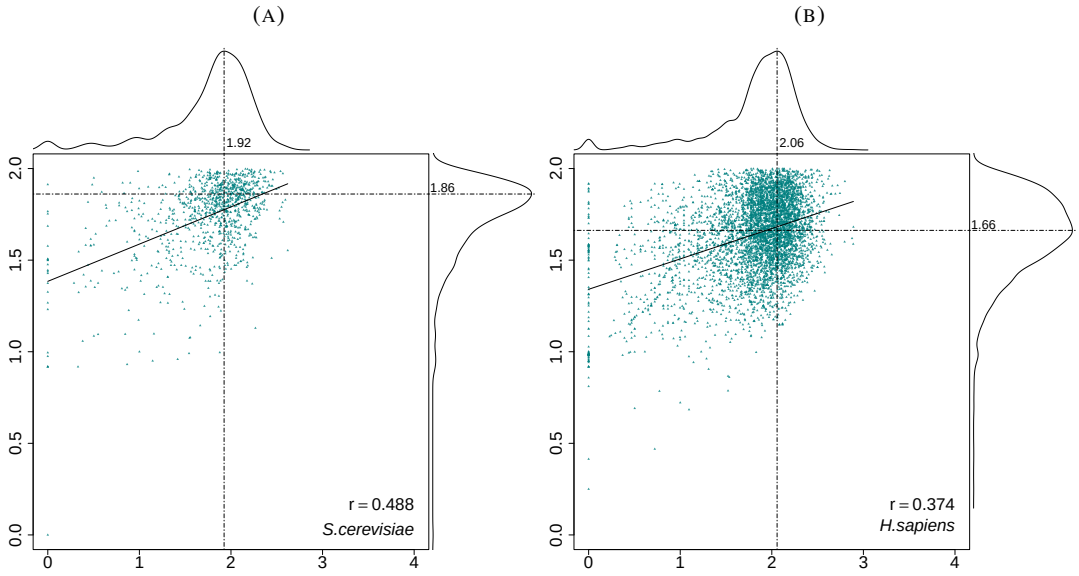

FIGURE 3.1: Entropy comparisons of protein LCRs and corresponding sequences in the *S. cerevisiae* and *H. sapiens* genome. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. a 1034 LCRs were identified from 6016 protein sequences in *S. cerevisiae* and their entropies were plotted against the entropies of the corresponding coding sequences ($r = 0.488$). b 5005 LCRs were identified from 133 689 protein sequences in *H. sapiens* and their entropies were plotted against the entropies of the corresponding coding sequences ($r = 0.374$).

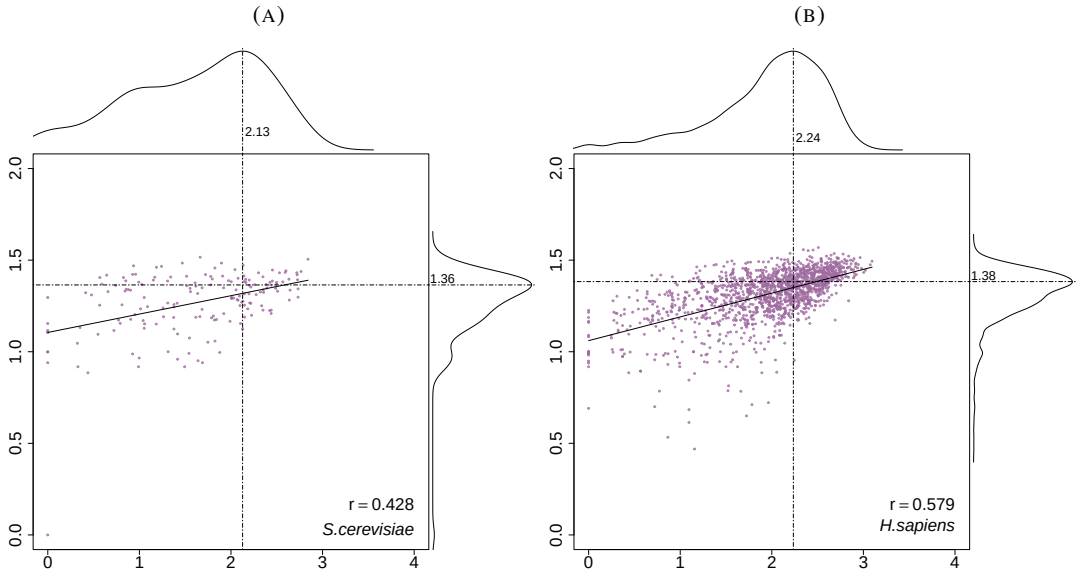

FIGURE 3.2: Entropy comparisons of LCRs and corresponding sequences in the *S. cerevisiae* and *H. sapiens* genome. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. a 171 LCRs were identified from 6016 coding sequences and their entropies were plotted against the entropies of the corresponding protein sequences ($r = 0.428$). b 1571 DNA LCRs were identified from 133 689 coding sequences and their entropies were plotted against the entropies of the corresponding protein sequences ($r = 0.579$).

Protein LCRs

> NP_001369.1 NC_000002 REEKKRKEEERKKKE	$H = 1.52$	> NP_001369.1 NC_000002 AGAGAGGAAAAGAAGAGAAAAGAA GAAGAAAGGAAAAAAAAAGAA	$H = 0.867$
> NP_056007.1 NC_000020 DDDDDDDD	$H = 0$	> NP_056007.1 NC_000020 GACGACGACGATGATGATGATGAC	$H = 1.92$

DNA LCRs

> XP_024307650.1 NC_000020 GGTTTTTTGTTTTTTGTTTTTTTG GTTTTTGTTTGTTTGTTTTTT	$H = 0.722$	> XP_024307650.1 NC_000020 GFLFFVFLVFVCLFF	$H = 1.96$
> XP_011508832.1 NC_000002 CCACCGCCGCCGCCGCCCCCTCC TCCACCTCCTCCTCCCCCACCGC CCCCTCCGCCTCCTCCTCTC	$H = 1.28$	> XP_011508832.1 NC_000002 PPPPPPPPPPPPPPPPPPPPL	$H = 0.267$

FIGURE 3.3: Example entropy comparisons from the opposing extremes among *H. sapiens* sequences, obtained from LCRs in protein sequences (Protein LCRs) and LCRs in DNA sequences (DNA LCRs). ***Protein LCRs)*** On the top, a relatively higher entropy protein LCR is encoded by a comparatively low entropy DNA sequence. On the bottom, an extremely low entropy protein LCR is encoded by a high entropy DNA sequence. ***DNA LCRs)*** On the top, a low entropy DNA LCR codes for a relatively high entropy protein sequence. On the bottom, a relatively higher entropy DNA LCR codes for a relatively low entropy protein sequence.

Other examples of extreme deviations include having high protein entropy but low DNA entropy or vice versa. Examples of both are shown in Figure 3.3 and provide insight into the low sequence correlations observed. Protein LCRs with high protein entropy and unexpectedly low DNA entropy consist of amino acid residues whose codons share the same nucleotides and contain two or fewer different nucleotides (Figure 3.3). For example, the protein LCR with relatively higher entropy composed of R, E, and K are encoded by the codons AGA, AGG (R), GAG, GAA (E), and AAG, AAA (K), all of which share the nucleotides, A and/or G. In contrast, low entropy protein regions with high entropy DNA sequences can be the result of a DNA sequence composed of distinct, but synonymous codons which are composed of three different nucleotides. In this case, the aspartic acid homopolymer encoded GAC and GAT. The DNA LCRs with low entropies and unexpectedly high protein entropies again tend to be encoded by codons which all share the same nucleotides in different rearrangements but encode different amino acids. Additionally, few distinct, synonymous codons are used for each amino acid. In this case, the sequences are composed of: G (GGT), F (TTT), L (TTG), V (GTT), and C (TGT). In DNA LCRs with relatively high DNA entropy and comparatively low protein entropy, different residues are encoded by synonymous codons which often do not share the same nucleotides. Hence, the degree of codon homogeneity, the codon nucleodicity, as well as the potential for shared nucleotides between codons, all affect the degree of correlation between entropies of protein and DNA sequences in LCRs.

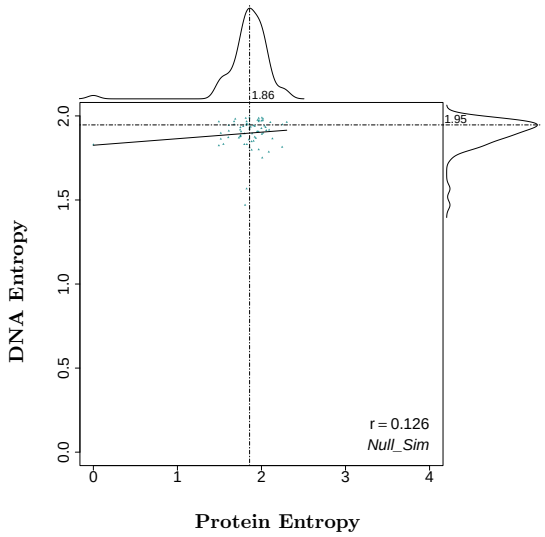

FIGURE 3.4: Entropy of LCRs from the null simulated proteomes. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. 70 LCRs were identified from a sample of 100 000 protein sequences.

3.4.2 Using Slippage and Substitution Models to Compare and Explain Observed Entropy Correlations in Biological Sequences

To further examine the significance of the sequence entropy correlations, proteomes were simulated according to five different slippage and substitution models. The first proteome, generated according to the Null model, contained only 70 proteins with LCRs. The correlation between protein and coding sequence entropy for these LCRs was low at $r = 0.126$ (95% CI: $-0.139 - 0.374$; Figure 3.4). The correlation coefficient for the random simulation was lower than that in both *H. sapiens* and *S. cerevisiae*, although this difference was only significant in *S. cerevisiae*. There was only one DNA LCR which was identified from the 100 000 Null model DNA sequences. Due to the lack of data points, a linear regression could not be performed. The small number of LCRs identified from the Null model sequences in both proteins and DNA, as well as the low sequence entropy correlation in protein LCRs suggests, not surprisingly, that LCRs are not the sole result of randomness in nature and are the result of some biological driving force. This is consistent with the literature which shows that LCRs are caused by replication error events like polymerase slippage which are exacerbated by an increase in repeat length (Viguera et al. 2001; Lai and Sun 2003; Levinson and Gutman 1987).

The second proteome, generated according to the Slip model, consisted of sequences with a propensity to form LCRs in a repeat length dependent manner in an attempt to mimic LCR formation by DNA polymerase slippage. In general, the Slip model resulted in higher LCR sequence entropy correlations than in the biological LCRs. When looking at protein LCRs and their corresponding DNA sequences in the *S. cerevisiae* specific simulation, the correlation coefficient was significantly higher than for the *S. cerevisiae* biological sequences at $r = 0.566$

(95% CI: 0.555 – 0.577; Figure 3.7a). DNA LCRs and their corresponding protein sequences in *S. cerevisiae* had a correlation coefficient of $r = 0.591$ (95% CI: 0.573 – 0.608; Figure 3.7b) which was also significantly higher than the biological *S. cerevisiae* DNA LCRs. In the Slip simulation specific to *H. sapiens*, when looking at protein LCRs and their corresponding DNA sequences, the correlation coefficient was significantly higher than for the biological protein LCRs in *H. sapiens* at $r = 0.658$ (95% CI: 0.637 – 0.678; Figure 3.7c). The correlation coefficient for the DNA LCRs and their corresponding protein sequences in the *H. sapiens* Slip model was slightly lower than the *H. sapiens* biological sequences although not significantly at $r = 0.573$ (95% CI: 0.503 – 0.636; Figure 3.7d). Overall, the higher correlations in the Slip model suggest that if LCRs were formed strictly in a neutral manner by DNA polymerase slippage, we would expect to see higher correlations in the biological sequences than what were actually observed.

To simulate conservation of the amino acid sequences, 1000 synonymous mutations were added to the coding sequences from the Slip model, generating a third proteome, the Slip+Syn model. In general, implementing synonymous mutations into the coding sequences decreased the correlation compared to the Slip model. Correlations when going from the Slip model to the Slip+Syn model in the *S. cerevisiae* specific simulation were significantly lower for protein LCRs and their corresponding sequences ($r = 0.459$; 95% CI: 0.446 - 0.472; Figure 3.8a) as well as the reverse ($r = 0.517$; 95% CI: 0.489 - 0.544; Figure 3.8b). In the *H. sapiens* specific Slip+Syn simulation, protein LCRs and their corresponding sequences had a significantly lower correlation compared to the Slip model ($r = 0.585$; 95% CI: 0.561 - 0.608; Figure 3.8c). However, DNA LCRs and their corresponding sequences were slightly higher although this was not significant ($r = 0.588$; 95% CI: 0.492 - 0.670; Figure 3.8d). This helped confirm that we could expect LCR sequence entropy correlation to be lower if a homogeneous run of codons was broken up by synonymous mutations and the LCR had thus been conserved or selected for. When comparing Slip+Syn simulations to the biological sequences, there were varying results depending on the organism and LCR sequence type. For protein LCRs and corresponding sequences in *S. cerevisiae*, the correlation was insignificantly lower. For DNA LCRs and their encoded protein sequences, the correlation was insignificantly higher. In *H. sapiens*, the correlation for protein LCRs and corresponding sequences had a significantly higher correlation. Also, DNA LCRs and corresponding sequences had an insignificantly higher correlation. Thus, the biological sequences have LCR sequence entropy correlation more similar to the LCRs generated from slippage followed by synonymous mutations although this model may still be limited in its ability to describe and predict LCR evolution.

The Slip and Slip+Syn simulations did have a greater tendency to generate mono-codon runs of LCRs as made evident by the large fraction of both DNA and protein LCRs with a protein entropy of 0. Since this trend was not observed in the biological sequences, this suggested that perhaps biology has a preference against runs of identical codons in LCRs. Thus, a preference for codon nucleodicity class was investigated in each organism. In the coding sequences for protein LCRs a strong nucleodicity class bias was observed for *S. cerevisiae* ($\chi^2 = 9.749 \times 10^{-30}$) and *H. sapiens* ($\chi^2 = 1.138 \times 10^{-280}$). Sequences in both organisms showed a greater number of codons with a nucleodicity of two and fewer codons with a nucleodicity of one and three. The fourth proteome, Slip+CC was generated taking this codon

nucleodicity bias into account. Correlation coefficients from this model specific to *S. cerevisiae* were $r = 0.699$ (95% CI: 0.690 - 0.707; Figure 3.9a) for Protein LCRs and corresponding coding sequences. This was significantly higher than the corresponding correlation coefficient for the *S. cerevisiae* Slip simulation as well as the *S. cerevisiae* biological sequences. A correlation coefficient of $r = 0.553$ (95% CI: 0.534 - 0.572; Figure 3.9b) was observed for DNA LCRs and their encoded protein sequences which was significantly lower than in the corresponding Slip simulation and insignificantly higher than for the biological *S. cerevisiae* DNA LCRs. For the *H. sapiens* specific model, the correlation coefficient for protein LCRs and their corresponding sequences was $r = 0.808$ (95% CI: 0.795 - 0.820; Figure 3.9c). Again, this was significantly higher than in the corresponding Slip simulation as well as for the *H. sapiens* biological sequences. The correlation coefficient for DNA LCRs and their corresponding sequences was $r = 0.546$ (95% CI: 0.479 - 0.607; Figure 3.9d) which was insignificantly lower than in the Slip simulation as well as for the biological *H. sapiens* DNA LCRs. Thus, the Slip+CC simulation did not model the LCR sequences as anticipated and result in higher correlations for protein LCRs and their corresponding sequences but have less of a discernible effect on the correlation of DNA LCRs and their corresponding sequences.

Lastly, 1000 synonymous mutations were added into the coding sequences from the Slip+CC simulations to produce a fifth proteome, Slip+CC+Syn. Similarly to the Slip and Slip+Syn simulations, the correlations from Slip+CC+Syn compared to Slip+CC were lower for both the *S. cerevisiae* and *H. sapiens* in both LCR sequence types. This was significant for all values except DNA LCRs and corresponding sequences in *H. sapiens*. In the *S. cerevisiae* specific version of this simulation, the correlation for protein LCRs and their corresponding sequences was close compared to the biological *S. cerevisiae* protein LCRs at $r = 0.490$ (95% CI: 0.477 - 0.503; Figure 3.10a). The correlation for DNA LCRs and their corresponding sequences was $r = 0.430$ (95% CI: 0.397 - 0.462; Figure 3.10b) which was also close to the biological *S. cerevisiae* DNA LCRs. For the *H. sapiens* specific simulation, the correlation between protein LCRs and their corresponding sequences was significantly higher than in the biological sequences at $r = 0.620$ (95% CI: 0.598 - 0.641; Figure 3.10c). The correlation between DNA LCRs and their corresponding sequences was insignificantly lower at $r = 0.536$ (95% CI: 0.446 - 0.615; Figure 3.10d). Thus, while this simulation seemed a good model for *S. cerevisiae*, this was not the case for *H. sapiens* suggesting length dependent slippage and incorporation of codon nucleodicity preferences followed by synonymous mutations is not sufficient to explain the correlations observed and therefore the mode of LCR evolution. Figure 3.5 summarizes the entropy correlations from the five simulations and compares them to the entropy correlations from the biological sequences. A summary table of the main results can be found in Table A1.11.

3.4.3 Comparing Correlations Between LCRs categorized as Periodic or Cryptic Repeats in Biological Sequences

The original analysis of LCRs and corresponding sequences of the five model organisms looked at sequence entropy correlations of LCRs as a whole. However, some studies have suggested that different types of LCRs, particularly protein LCRs with tandem periodic amino acid repeats may evolve differently than cryptic repeat LCRs with periodic repeat LCRs being

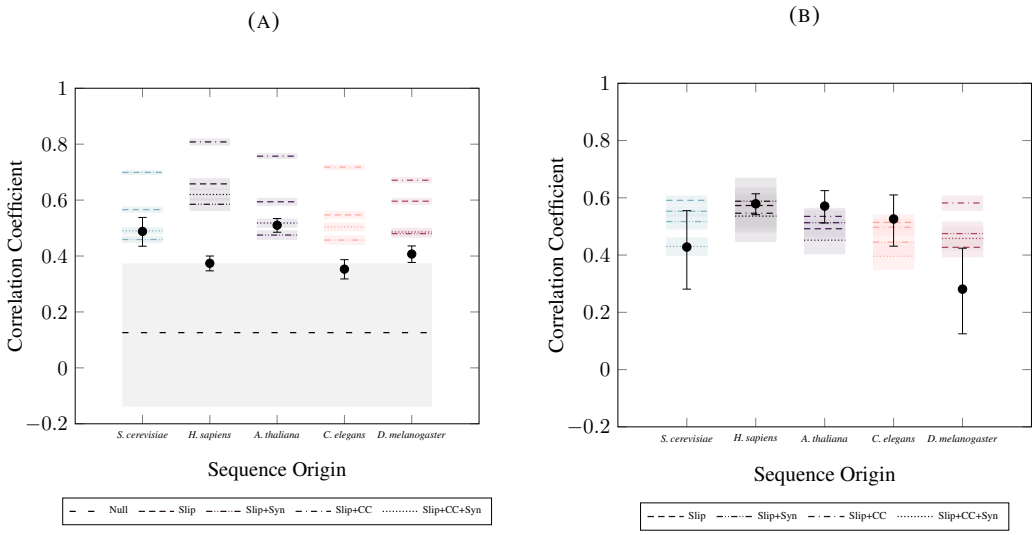

FIGURE 3.5: A summary of the LCR entropy correlation coefficients and 95% confidence intervals for each model organism and species specific simulated proteome. Exact correlation coefficient values can be found in Table A1.11. a Correlation coefficients are from protein LCRs and corresponding coding sequence linear regressions. The non-species specific Null proteome is also included (the shaded area which spans all 5 species). b Correlation coefficients are from DNA LCRs and corresponding protein sequence linear regressions. A linear regression could not be performed for DNA LCRs from the Null simulation due to lack of data points.

more likely to have evolved through DNA polymerase slippage (Battistuzzi et al. 2016). To test this theory, LCRs from the biological sequences were divided into two categories, those with periodic amino acid repeats and those without periodic amino acid repeats (cryptic repeats). Sequence entropy correlations between these two LCR classes were only investigated for protein LCRs as this seemed more biologically relevant for repeats at the amino acid level and because tri- or hexa- repeats at the DNA level would lead to repeats at the corresponding amino acid level and thus would likely not be informative of LCR evolution. For all organisms, LCR sequence entropy correlations were always significantly higher in periodic repeat LCRs compared to cryptic repeat LCRs Figure 3.6. Correlations for LCRs with periodic repeats were also either higher or significantly higher than the correlation for both LCR types combined, and correlations in cryptic repeats were significantly lower than the correlation for both LCR types combined. For *S. cerevisiae*, the correlation between protein LCRs and corresponding sequences with periodic repeats and cryptic repeats was $r = 0.573$ (95% CI: $0.481 - 0.652$) and $r = 0.322$ (95% CI: 0.248 - 0.392), respectively. For *H. sapiens*, the correlations for periodic repeats versus cryptic repeats were $r = 0.488$ (95% CI: $0.402 - 0.492$) and $r = 0.242$ (95% CI: 0.207 - 0.277), respectfully (Figure 3.6). Overall, these results suggest that LCRs containing periodic amino acid repeats are more likely to evolve via DNA polymerase slippage whereas cryptic repeat LCRs are more likely to be selected for. Thus, of the five slippage and substitution models, the Slip simulation should be the most accurate model for the evolution of the periodic repeat LCRs. Overall, the Slip simulation did bear the closest resemblance in correlation to the biological sequences with the correlation being similar between Slip and periodic LCRs in *S. cerevisiae* but significantly higher compared to periodic LCRs in *H. sapiens* (Figure 3.6). On the contrary, the cryptic LCRs did not bear resemblance to any of the slippage or substitution models and were significantly below the correlations for all species specific models in all organisms (Figure 3.6).

3.5 Discussion

At the outset, the fact that information flows from coding sequence to protein sequence might lead one to have a naïve expectation that the entropies of LCRs at each level would be highly correlated. However, for each of the genomes from the model organisms examined the correlations observed were low to moderate.

To ensure that the choice of parameters chosen for the measurement of LCRs via `Seg` was not the cause of this unusual effect, LCRs were identified using varying sets of parameters. To identify protein LCRs, parameters were chosen based on previous studies which found these parameters to work well for identifying highly repetitive LCRs while avoiding sequences that had higher complexity (Huntley and Golding 2002; Haerty and Golding 2010; Battistuzzi et al. 2016). These were then increased and decreased to explore the effect of the parameters as shown in Table A1.2. It was difficult to determine the biologically equivalent parameters for finding DNA LCRs as there are no known studies which have used `Seg` for DNA LCRs previously. We therefore chose to use the parameters suggested in the `Seg` manual and varied the parameters around these values as shown in Table A1.3. As can be seen in these tables, adjusting window length, low cut, and high cut parameters overall had little impact on the

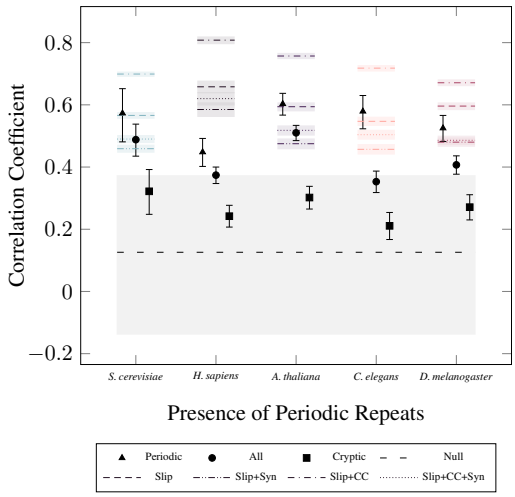

FIGURE 3.6: A summary of the correlation coefficients with 95% confidence intervals for protein LCRs with periodic amino acid repeats, all LCR types combined, and cryptic repeats for the five model organisms. The correlations for the four species specific slippage and substitution models and the Null model with 95% confidence intervals are also given for comparison. Exact correlation coefficient values can be found in Table A1.11

degree of correlation for the biological sequences (Tables A1.2 and A1.3). The value of the high cut parameter ($K2$) had the greatest impact on correlation. In general, and particularly in the DNA based results, increasing the values of the three parameters (window length, low and high cutoffs) resulted in higher correlation coefficients but at an extreme cost of many fewer regions considered to be low complexity. This could be because low entropies at long window lengths are less common, and higher low cut and high cut parameters results in less extreme LCRs being considered.

The analysis of the simulated proteomes, Null, Slip, and Slip+Syn, was also performed with the `Seg` parameter sets described in Tables A1.4 to A1.8. Regardless of the set of parameters used, organism examined, and sequence type in which the LCRs were identified in, the results were qualitatively the same: The correlations observed in biological data were lower than the correlations produced in any simulation. In most cases, the biological correlations were significantly lower, however the exceptions were concentrated at the high complexity extreme of `Seg` parameters tested. For example, in the instances where the biological correlation was significantly higher than the correlations in Slip and Slip+Syn for LCRs identified in proteins, all occurred at the settings with highest K2 values. K2 values much higher than K1 cause `Seg` to degenerate to finding the least probable subsequence in a protein regardless of entropy values. We observed that the excess correlation in the simulations was greatest when the definition of an LCR was strictest. This indicates that the evolutionary mechanisms embodied in the simulations: replication slippage as well as nucleodicity and synonymous substitutions

insufficiently explain the observed distribution of protein LCRs, especially for highly repetitive LCRs.

Even when the correlations seen in the simulations were not significantly different from the biological sequences, the distribution of sequence entropies was very different. All of the simulations which included slippage produced lower entropy LCRs both at the DNA and protein level (Figures 3.7 and 3.9). Slippage produced far more mono-amino acid repeats than observed biologically, indicating that slippage alone doesn't explain the abundance of less compositionally biased LCRs. Non-synonymous mutations which break up the amino acid tract would shift the distribution away from homopolymers towards the more commonly observed entropies (Figures 3.8 and 3.10). The distribution of DNA entropies in simulations which did not include synonymous mutations was also biased towards lower entropies but was also multimodal with peaks near the entropies corresponding to the nucleodicity of the plurality codon in the sequence (Figures 3.7 and 3.9). The addition of synonymous mutations brings the DNA entropy distribution more in line with what is seen biologically: unimodal with a peak at higher entropy (Figures 3.8 and 3.10). There may be a sample size effect, as we see the most dissimilarity between simulation and biology in *H. sapiens* which had the most LCR containing proteins at 5005 while *S. cerevisiae* had only 1034.

Comparing LCRs identified in proteins or in DNA coding sequences, the correlations for LCRs found in coding sequences were usually significantly higher in biological sequences. The exceptions are *S. cerevisiae* and *D. melanogaster* which both had the fewest LCRs identified in coding sequences at 176 and 187, respectively. The patterns are also unclear for the simulated proteomes: the Slip model often had higher correlation at the protein level, but the Slip+Syn model often showed the reverse. Without considering the bias in codon usage each organism has, correlations would be expected to be higher when identifying LCRs in coding sequences as some information is lost during the translation process. That is, the total entropy of the DNA sequence can never be lower than the protein sequence it encodes. The maximum entropy for a nucleotide is 2 bits, while the maximum entropy for an amino acid is roughly 4.32 bits. However, each amino acid is encoded by three nucleotides, for a maximum entropy of 6 bits. Thus, if the nucleotide variation is substantially constrained, as seen in DNA LCRs, the amino acids which can be encoded are limited to a select few. On the contrary, if the amino acid variation is limited, as seen with protein LCRs, there is still a possibility to have up to all four nucleotides comprising its coding sequence. Essentially, limiting DNA information content will limit protein information content, but the same is not necessarily true in the reverse direction. Hence, LCR sequences taken in one direction might be less correlated than LCR sequences taken in the other direction. Biological biases in codon usage, as well as the codons and amino acids tolerated in LCRs may modify this effect, and lead to the inconsistent pattern we observe.

It is interesting that DNA entropy for both protein and DNA LCRs rarely goes below one bit as there is evidence suggesting a bias toward the use of two nucleotides to drive particular codon usage which is thought to be associated with the presence of LCRs (Albà and Guigó 2004; DePristo et al. 2006; Knight et al. 2001; Li et al. 2015; Xue and Forsdyke 2003). In coding regions, there are rarely subsequences of DNA containing two or fewer nucleotides for 45 or more consecutive nucleotides. The percentage of protein LCRs with a corresponding DNA

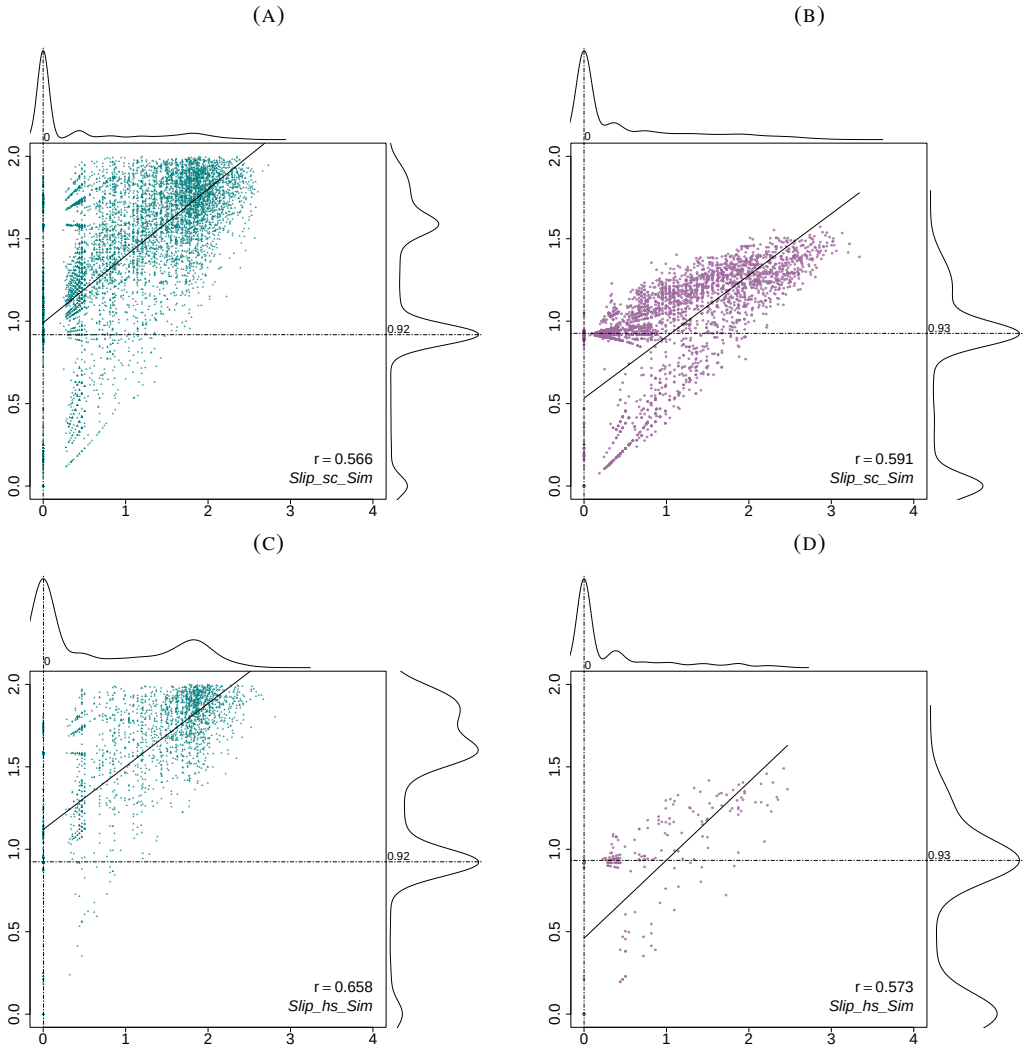

FIGURE 3.7: Entropy of LCRs from the Slip simulated proteomes specific to *S. cerevisiae* and *H. sapiens*. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. LCRs were identified from a sample of 100 000 sequences. a 17 239 LCRs were identified from protein sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.566$). b 6615 LCRs were identified from coding sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.591$). c 3765 LCRs were identified from protein sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.658$). d 488 LCRs were identified from coding sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.573$).

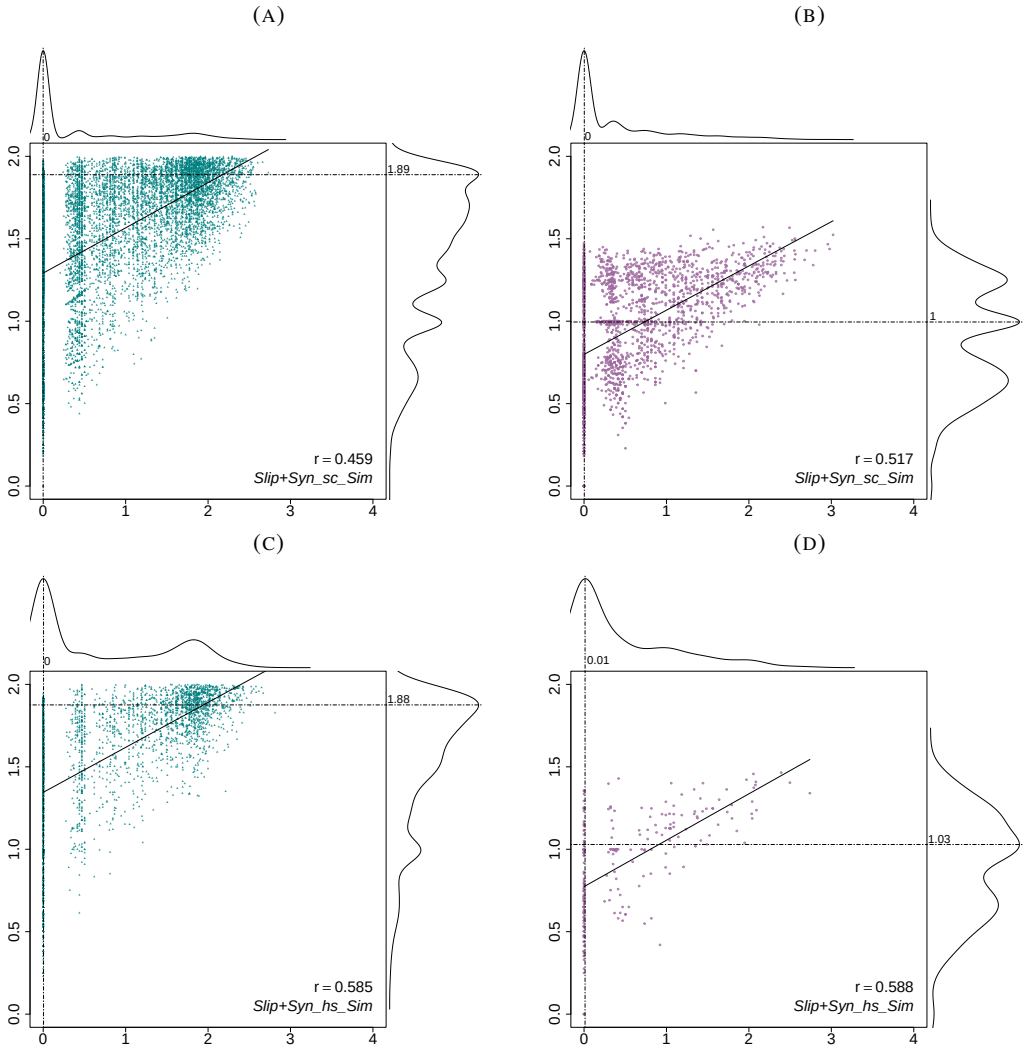

FIGURE 3.8: Entropy of LCRs from the Slip+Syn simulated proteomes specific to *S. cerevisiae* and *H. sapiens*. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. LCRs were identified from a sample of 100 000 sequences. a 17 239 LCRs were identified from protein sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.459$). b 3315 LCRs were identified from coding sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.517$). c 3765 LCRs were identified from protein sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.585$). d 260 LCRs were identified from coding sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.588$).

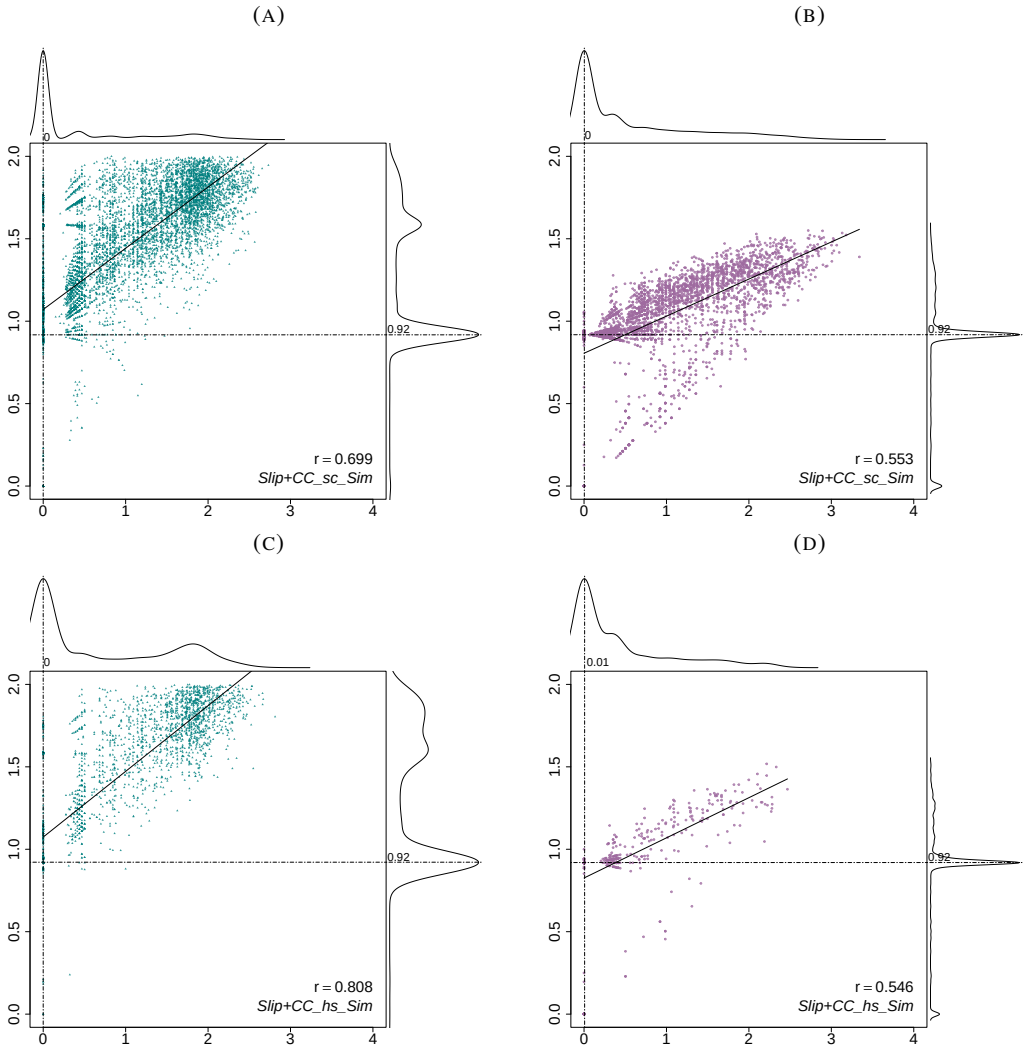

FIGURE 3.9: Entropy of LCRs from the Slip+CC simulated proteomes specific to *S. cerevisiae* and *H. sapiens*. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. LCRs were identified from a sample of 100 000 sequences. a 17 255 LCRs were identified from protein sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.699$). b 6505 LCRs were identified from coding sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.553$). c 3767 LCRs were identified from protein sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.808$). d 579 LCRs were identified from coding sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.546$).

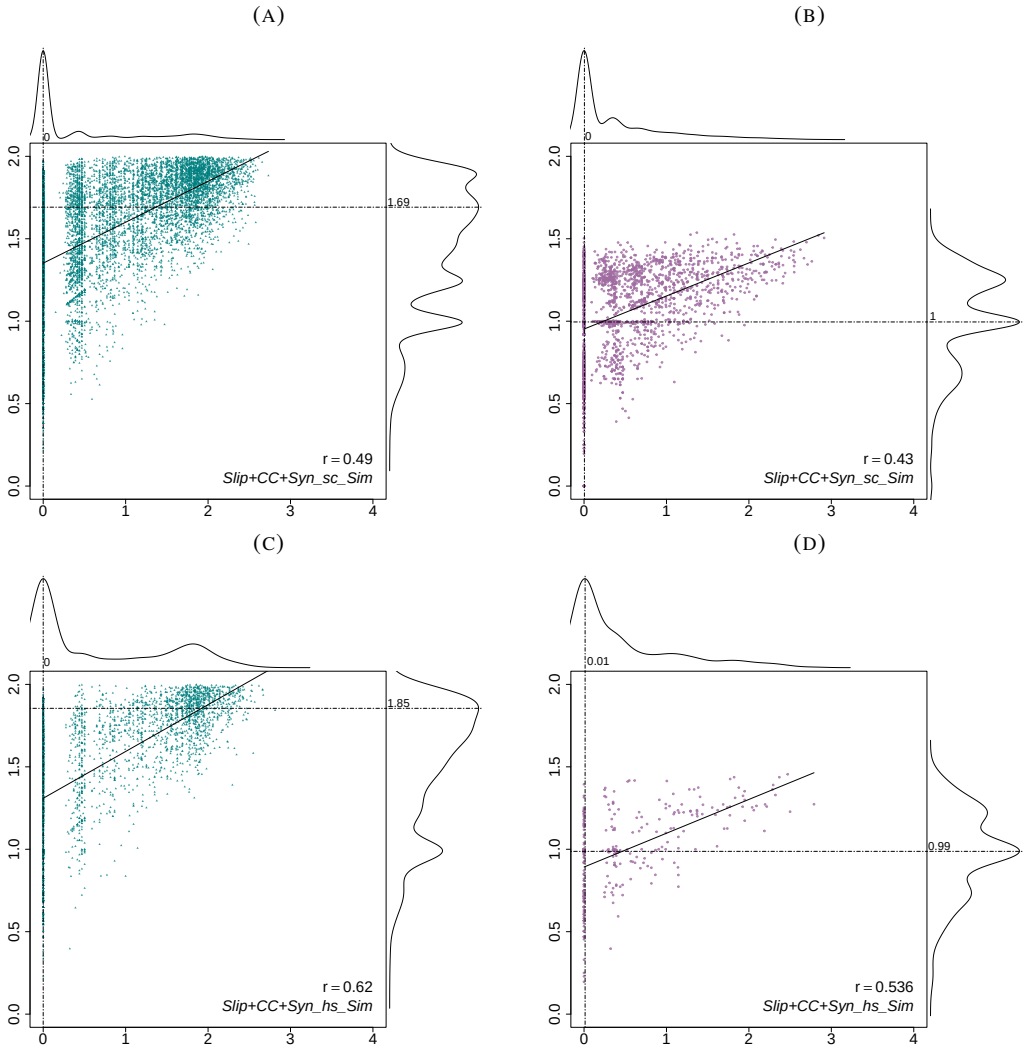

FIGURE 3.10: Entropy of LCRs from the Slip+CC+Syn simulated proteomes specific to *S. cerevisiae* and *H. sapiens*. LCRs were identified from a sample of 100 000 sequences. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. a 17255 LCRs were identified from protein sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.490$). b 3004 LCRs were identified from coding sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.430$). c 3767 LCRs were identified from protein sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.620$). d 340 LCRs were identified from coding sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.536$).

sequence entropy under 1 bit was 1.16% and 1.04% for *S. cerevisiae* and *H. sapiens*, respectively. None of the simulated proteomes had proportions as low, with the exception of the Null simulated proteome where no LCRs had DNA entropy less than 1. The Slip and Slip+CC models both produced proteomes where 31.6 (in *H. sapiens* Slip) to 53.2% (in *C. elegans* Slip+CC) of LCRs had DNA entropies less than 1.0, while the addition of synonymous mutations brought this proportion down to 14.6 (in *H. sapiens* Slip+Syn) to 21.2% (in *S. cerevisiae* Slip+Syn). Because the simulated values are much higher than observed in nature, it might indicate that biological sequences tend toward coding sequences with a higher variation of nucleotides than would be expected if both polymerase slippage resulting in codon repeats or a more heterogeneous mixture of codons was present.

It was possible that examining entropy at the codon level may have provided insights into LCR evolution. Each nucleotide triplet could be considered its own distinct character, especially considering that strong selection against frameshift mutations results in only replication slippage of whole codons being tolerated in coding regions (Metzgar et al. 2000). Because of this, we also calculated entropy at the codon level for each LCR. However, when comparing protein sequence entropy to codon entropy, there are mathematical constraints which force LCRs in the scatter plot to lie within a very constrained minimum and maximum threshold value, forcing a more linear relationship. (Figure A1.5). The maximal codon entropy for a homopolymeric sequence would occur if the amino acid had a 6 codon degeneracy at $\log_2 6$ which is 2.58 bits (Figure A1.5. While the minimal codon entropy is zero for the repetitive usage of a single codon. Due to these tight mathematical constraints, entropy at the codon level was not further considered.

The mathematical constraints on possible entropies are the result of codon nucleodicity. Nucleodicity bias would therefore play a role in the correlation between DNA and protein level entropies. In all organisms examined there was a significant bias against the formation or maintenance of mononucleic codon repeats. This definitely impacts the correlation between protein and DNA sequence entropies, as exemplified in Figure 3.3. However, including this bias in the simulations increases the correlation between DNA and protein sequence entropies, while leaving the distribution of protein entropies largely unaffected. Therefore, the nucleodicity bias not only doesn't explain the low biological correlation but seems to drive the correlation up. This indicates that other mechanisms must be acting as well.

One possible explanation for these low correlation results is that these LCRs are formed under a high selective pressure rather than through just polymerase slippage with low selective constraints. This result is surprising because there is a great deal of evidence to suggest that LCRs are the product of polymerase slippage at microsatellites, resulting in either the expansion or contraction of a repeat sequence (Hannan 2018; Levinson and Gutman 1987; Tompa 2003; Viguera et al. 2001; Wierdl et al. 1997). Also, evidence suggests that the polymorphic nature of LCRs is a result of this instability in combination with low selective pressure acting on either the protein as a whole, or this specific region of the protein (Fan and Chu 2007; Mularoni et al. 2007; Behura and Severson 2012). *In vitro* studies have also shown polymerase slippage can explain the observed microsatellite distributions within a genome (Madsen et al. 1993). Sequences consisting of pure codon repeats are more likely to undergo

slippage than codon tracts with synonymous mutations. Such mutations break up trinucleotide repeats and have been shown to help stabilize and conserve LCRs (Albà et al. 1999). Still other studies suggest that LCR expansion is in an equilibrium between insertions which decrease tract stability and point mutations which increase tract stability (Brandström and Ellegren 2008; Kruglyak et al. 1998). The length of repeat has also been shown to have a positive correlation with the chance of slippage (Lai and Sun 2003). Together this led us to hypothesize that as protein LCRs came nearer to a perfect repeat, and as the length of an uninterrupted tandem amino acid repeat became longer, that the chance of it being a result of slippage and having a corresponding DNA sequence consisting of pure codons also would increase.

Of course, if LCRs were a product of high selective pressures forcing an irregularly biased amino acid composition, the length and biochemical properties of the amino acids at the protein level would be the major important factor in determining the repetitiveness and ordering of an LCR. This suggests that the codon choice at the DNA level would be unimportant and could consist of any codons so long as they encoded for the correct amino acid. In this case, variety in codon usage would increase, likely resulting in a greater variety of nucleotides used. Thus, a high DNA entropy could encode for a wide range of protein LCR entropies, ultimately resulting in a lower correlation between sequence entropies. The correlations from the Slip+CC proteome was significantly higher than those observed for the biological sequence comparisons in both *S. cerevisiae* and *H. sapiens* (Figure 3.5). If slippage were the predominant LCR driving force with low selective constraints, it would be expected that the correlation coefficients for these organisms would be closer to those observed from the simulations. Instead, sequence entropy correlations were significantly different from that observed in the Slip and Slip+Syn proteomes. As well, the decrease in correlation coefficient between Slip and Slip+CC suggests that if LCRs were created predominantly by polymerase slippage and contained more pure codon repeats, the DNA and protein sequence entropies would be more highly correlated than if they were a product of selection and contained a greater mixture of synonymous codons (Figure 3.5).

Overall, the selective retention or loss or an LCR may be dependent on the location of the LCR within the protein (Coletta et al. 2010; Huntley and Clark 2007), the LCR type (Kobe and Kajava 2001; Radó-Trilla et al. 2015), the protein function (Coletta et al. 2010; Ekman et al. 2006), and the organism itself (Karlin et al. 2002). Battistuzzi et al. (2016) and Zilversmit et al. (2010) suggest that LCR type and periodicity may constitute factors which affect the mode of LCR evolution. They propose that slippage may be a more prominent mechanism as an LCR gets closer to a perfect repeat, and selection may be more important for LCRs with a higher complexity and lower periodicity. This would make sense, as slippage is thought to occur at higher rates at longer continuous repeat sequences (Leclercq et al. 2010; Lai and Sun 2003). When looking only at highly repetitive LCRs, our analyses agree. The results of the slippage based simulations do closely resemble the biological sequences for LCRs which are mainly periodic repeats (Figure 3.6). The differences seen when looking at all LCRs are driven by the non-repetitive LCRs. These compositionally biased domains with a lack of periodicity (cryptic repeats) would be less likely to undergo slippage and their presence would be more reasonably attributed to selection. But this discounts any suggestion that the cryptic repeats were at one point a periodic repeat which degenerated over time (Radó-Trilla and Albà 2012). Similarly,

Zilversmit et al. (2010) showed that in *P. falciparum* compositionally biased, aperiodic LCRs are less variable and evolve slower, whereas regions with long asparagine tracks are more variable and thought to evolve via replication slippage. We investigated the effect of periodicity on LCR entropy correlation using a range of minimum repeat lengths and found that when only sequences containing periodic repeats were compared, sequence entropy correlations were significantly higher in *H. sapiens* and higher, although not significantly, in *S. cerevisiae* compared to correlations of all LCR types combined (Figure 3.6). Correlations for periodic repeat LCRs were always significantly higher when comparing correlations for only cryptic repeat LCRs (Figure 3.6). The overall higher LCR sequence entropy correlation for periodic amino acid repeats is consistent with the findings of Battistuzzi et al. (2016) and Zilversmit et al. (2010).

The data presented here demonstrates that there is an unusually low correlation between the entropies of LCRs within proteins and their corresponding DNA coding sequences. This is largely driven by LCRs with cryptic rather than periodic repeats. Although LCRs are thought to be primarily created via polymerase slippage, the simulations conducted suggest that the correlations would be higher if this were the sole mechanism. Continuing evolution of cryptic LCRs via synonymous substitutions cannot reduce the size of the correlation and still maintain the size and entropy of the observed LCRs. Instead, the data suggests that these protein LCRs are maintained by genome wide, pervasive selection which acts to reduce the correlation by favouring synonymous substitutions that lower the correlation by lowering the repetitiveness of the LCRs at the DNA level and hence increasing the stability of the LCR. This may be partially facilitated through a bias against mononucleic codon repeats as we observe significantly fewer of these than codon usage frequencies would suggest. This would make the LCRs less prone to potentially deleterious slippage mutations. In the future, it is necessary to know if the correlations in sequence entropies change with the age of the proteins. If this hypothesis is true, we would expect a higher correlation in more recently evolved proteins (Toll-Riera et al. 2012) compared to older, more highly conserved proteins. Future investigations should also determine how preferred amino acid residue, protein function, protein age, and role of the LCR within the protein influences the relation between protein and DNA sequence entropies in LCRs.

3.6 Acknowledgements

We thank Natural Sciences and Engineering Research Counsel for funding provided for this project (grants RGPIN-202-05733 to GBG, PGSD3-547476-2020 to ZWD, USRA-526761 to JME).

3.7 Competing Interests

The authors declare there are no competing interests.

3.8 Funding Statement

This research was supported by Natural Sciences and Engineering Research Counsel (grants RGPIN-202-05733 to GBG, PGSD3-547476-2020 to ZWD, USRA-526761 to JME).

3.9 Data Availability

The most up-to-date data, code, and supplemental information for this research is publicly available on GitHub at https://github.com/JohannaEnright/LCREntropyProject/

Chapter 4

Characterizing Low Complexity Region associated Transcript and Protein Abundance

4.1 Preface

This chapter was published in Molecular Biology and Evolution in May, 2022 (https://doi.org/10.1093/molbev/msac087), with Zachery W Dickson and G Brian Golding as co-authors. ZWD devised the project idea, performed all analyses, and prepared the manuscript and figures. GBG provided guidance and edited the manuscript.

In this work our goal was to test a hypothesis that highly mutable low complexity regions (LCRs) are unlikely to be tolerated in important, high abundance proteins. We did show this but observed that this observation was not consistent at the transcript level. In general, transcripts which encode LCRs are more abundant than those which do not, an observation which was consistent across tested mammalian species. The implication is that more transcripts are required to maintain the same level of protein production if an LCR is present. Our results should also give pause to researchers who wish to draw physiological conclusions based on protein level effects based purely on transcript level data.

4.2 Introduction

Low complexity regions (LCRs) are some of the most common motifs in eukaryotic proteins (Golding 1999; Huntley and Golding 2000). These regions are highly repetitive and are enriched for one or a few amino acid residues. These regions are often intrinsically disordered, lacking fixed structures under normal physiological conditions (Romero et al. 2001). Perhaps as a result, these regions were thought of as a protein analog for 'junk-DNA', or as spacers between other protein regions (Golding 1999). However, more research has shown that these regions can perform various specific roles. They have been associated with phenotypic variation (Fondon and Garner 2004), implicated in neurodegenerative diseases (Cummings and Zoghbi 2000), suggested as hub proteins for interaction networks (Dosztányi et al. 2006), and shown to be essential to the normal functioning of some proteins (Loya et al. 2012).

LCRs are broadly defined by compositional bias (Mier et al. 2020), and there exist multiple methods for detecting and classifying LCRs. The basis for these methods ranges from sequence entropy (Wootton and Federhen 1993) and probability (Harrison 2017) to prediction of intrinsic disorder (Dosztányi et al. 2005). LCRs can be classified in several different manners including by primary amino acid, by location in the protein, by length, and by function (if known). A recent study found that for some proteins containing an essential LCR, the region could be replaced with some LCRs from other proteins without loss of function (Loya et al. 2017). The inter-operability of LCRs could thus be used as another classifier.

LCRs can expand and contract rapidly via slippage of DNA polymerase during replication (Huntley and Golding 2006) and can arise from unequal crossover events (DePristo et al. 2006). They may also evolve as the result of selection. Whether the LCR is retained in the protein once it has arisen is affected by several factors. Recent work has suggested that LCRs are preferentially retained in proteins which are already tightly regulated, possibly as the existing regulation ameliorates any deleterious effects from the LCR's presence (Chavali et al. 2017).

LCRs are also thought to arise in regions under relaxed selection, however previous work, examining serine homopolymers, found evidence of selection based on codon usage (Huntley and Golding 2006). Lenz et al. (2014) found that substitution rates increase in primate proteins in those regions flanking repetitive sequences like LCRs and microsatellites, and that these regions were under higher purifying selection. All of these results suggest that the presence of LCRs has evolutionary consequences for their host proteins.

It is well known that expression levels are positively correlated with selection pressure, with those genes which are most highly and broadly expressed being under strong selection (Pál et al. 2001). The abundance of these proteins makes their fitness sensitive to perturbations in their function as defined by their structure (whether globular or intrinsically disordered). The appearance, expansion, and deletion of an LCRs all have the capacity to dramatically alter the ability of a protein to perform its function. The majority of such mutations are deleterious and would be subject to purifying selection, and only tolerated where the effect is smaller, such as low abundance proteins under more relaxed selection. The intolerance for LCRs in high abundance proteins would result in a negative association between protein abundance and the presence of LCRs. It would then be expected that LCR positive (LCR^+) proteins would have lower expression than LCR^- proteins.

Previously, this relationship has only been incidentally examined. Some specific LCR^+ proteins have been studied for their influence on human health or their structural properties (Cornman and Willis 2009; Shin et al. 2016). A more general study of *Saccharomyces cerevisiae* proteins which contained homo-repeats found that these proteins are in lower abundance than other proteins (Chavali et al. 2017). This study examined only this one type of LCR which may have different properties from other LCR types.

Characterizing the relationship between LCRs, gene expression, and protein abundance (PAb) is an opportunity to shed light on the complex relationship between the latter two. There are multiple levels of regulation applying to protein expression at every step from transcription,

through translation, and protein stability. Not all of these processes are well understood and thus attempts to predict PAb from mRNA levels have been met with mixed success (Nie et al. 2006, 2007). This is a concern as gene expression research is increasingly being used to develop therapies, despite weak connections to the more physiologically relevant PAb. Nie et al. (2006) have used sequence characteristics to address a portion of the variation observed. Among the characteristics examined were amino acid and codon usage, but LCRs were not considered.

To our knowledge the research here is the first to comprehensively examine LCR^+ protein expression across mammals. We characterize the relationships between LCRs of different types and their expression in various tissues. We examine the apparent differences between transcript abundance (TAb) and PAb through the lens of LCRs and show that PAb is negatively associated with the presence of LCRs, but TAb is, unexpectedly, positively associated with the presence of LCRs.

4.3 Results

Human TAb and PAb from the GTEx project and the PaxDb were collected for 17975 proteins. Of these, 4246 (23.6%) were identified as containing an LCR as determined by the presence of a 15 amino acid window with Shannon entropy less than 1.9 bits. One million random permutations of the labeled LCR status were performed to generate distributions of expected quartile shifts. The observed median PAb for LCR^+ proteins was lower than that found in 99.2% of the permutations. On the other hand, the observed median TAb was higher than all medians found in the permutations (Figure 4.1). In both cases, the shift is significantly different from zero by the Mann-Whitney U test ($p_{PAb} < 10^{-16}$, $p_{TAb} = 1.06 \times 10^{-11}$).

Figure 4.2 shows the significance of the observed shift for the baseline, entropy based permutation and several potential biasing factors. Permutation testing using intrinsic disorder instead of LCR status yielded qualitatively similar results with two exceptions. The bottom quartile of abundance for LCR^+ proteins shows significantly greater abundance than that for LCR^- proteins, and there was no significant difference observed for the top quartile of TAb. Ischemia-time-adjusted TAb values give near identical results to the unadjusted TAb values. Qualitatively similar results are also observed when accounting for isoform redundancy and when restricting to only heteropolymer LCRs. The shifts observed for homopolymer LCRs are much less significant but are qualitatively similar.

The median tau index of LCR^+ proteins (0.74) is lower than that for LCR^- proteins (0.76) ($p < 10^{-5}$) indicating that LCR^+ proteins are more common among broadly expressed proteins. As a result, the proportion of LCR^+ proteins is higher than in the aggregate. These proportions vary from 24.1% in the testis to 25.1% in the brain. Regardless of these differences, the aggregate permutation results are qualitatively consistent with the results across tissues (Figure 4.3). Liver tissue is an exception. However, it had the smallest number of expressed proteins (12914), the second lowest proportion of LCRs among those proteins (24.6%), and the highest standard deviation in $\log_2$ scaled TAb (3.15).

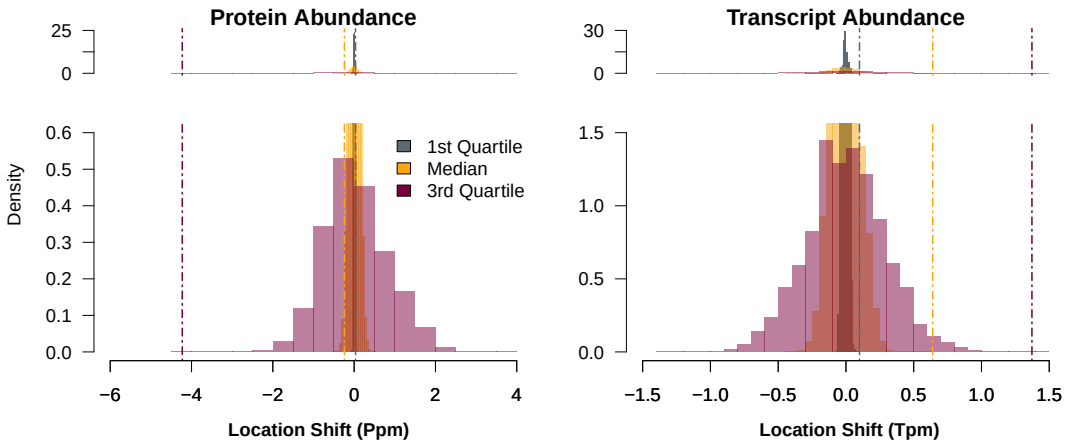

FIGURE 4.1: Under permutation all abundance quartiles are significantly different based on LCR status. The distribution of shifts in abundance quartiles after one million permutations is shown; the lower plots are insets of the upper plots. The observed shift for each quartile (dashed lines) can be compared to the matching distribution of location shifts under permutation. LCR^+ proteins have lower abundance for all quartiles ($p \leq 0.023$) but higher for TAb at all three quartiles ($p \leq 2 \times 10^{-6}$).

These results suggest that the presence of LCRs is associated with an increase in the level of TAb which might be required to maintain a particular PAb level, as compared to LCR^- proteins. While there are many processes in the pathway from gene transcription to protein degradation, we focused on the rates of protein degradation and translation. Coefficients of degradation (k_{deg}) for 3222 human proteins were aggregated, of which 965 (30.0%) were LCR^+. Schwanhäusser's data for 2180 mouse proteins included 450 (20.6%) LCR^+ proteins. In both datasets, LCR^+ proteins degrade 20-30% more rapidly (Table 4.1).

The perturbability and resupply of local codon supply were estimated using Schwanhäusser's mouse data. The estimated parameters indicated low perturbability but slow resupply meaning translation is only likely to be affected for longer, more repetitive transcripts. A single translation step consumes 5.78% of the tRNA isoacceptor supply of the least supplied codon, while 0.064% of the deficit between local and global supply is ameliorated. These parameters result in a correlation between measured translation rates and calculated time weighted normalized translation efficiency (TWnTE) values of 0.53 (95% CI [0.51,1.0]).

Using these translation parameters and selective wobble constraints optimized for each dataset (Table A2.2), TWnTE values were calculated using GTEX and Schwanhäusser data for 18054 (30.9% LCR^+) human and 3407 (34.0% LCR^+) mouse proteins. For both species, transcripts encoding LCR^+ proteins are translated 35-55% less efficiently (Table 4.1).

Logistic regression was used to estimate the relationship between abundance and LCR status while accounting for protein degradation, translation efficiency, and the increased odds of

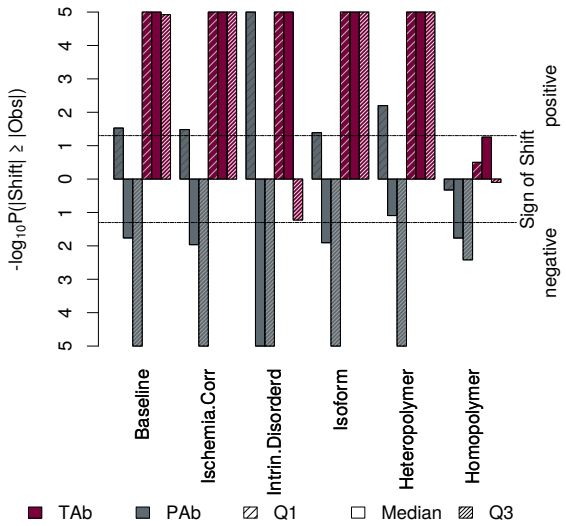

FIGURE 4.2: Observed shifts in LCR status remain after controlling for several technical explanations and known biological conditions. Bars represent empirical P values, calculated as the proportion of 100,000 permutations of GTEx and PaxDB human data with quartile shifts at least as large as the observed shift. Bars above the horizontal axis have observed shifts where LCR^+ proteins are greater than the null expectation, while bars below the horizontal axis represent observed shifts where LCR^+ proteins are below the null expectation. Dotted horizontal lines indicate a significance threshold of 0.05. All results are qualitatively similar to the baseline analysis.

TABLE 4.1: Summary of Protein Degradation and Translation

Process	Species	LCR Status	N	Median	95% CI	
					lower	upper
Degradation	Human	-	965	1.82×10^{-2}	1.75×10^{-2}	1.88×10^{-2}
k.deg (1/hr)		+	2257	2.42×10^{-2}	2.28×10^{-2}	2.57×10^{-2}
	Mouse	-	450	1.38×10^{-2}	1.30×10^{-2}	1.45×10^{-2}
		+	1730	1.57×10^{-2}	1.46×10^{-2}	1.79×10^{-2}
Translation	Human	-	4259	2.90×10^{-2}	2.84×10^{-2}	2.95×10^{-2}
TWnTE		+	13795	1.41×10^{-2}	1.37×10^{-2}	1.48×10^{-2}
	Mouse	-	865	2.28×10^{-4}	2.20×10^{-4}	2.37×10^{-4}
		+	2542	1.47×10^{-4}	1.38×10^{-4}	1.54×10^{-4}

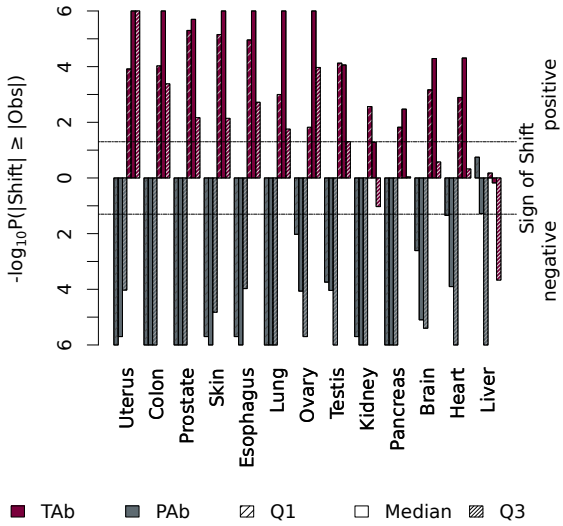

FIGURE 4.3: Differences in abundance between LCR$^+$ and LCR$^-$ proteins are consistent across tissues in humans. Bars represent empirical P values, calculated as the proportion of 1×10^6 permutations of GTEx and PaxDB human data with quartile shifts at least as large as the observed shift. Bars above the horizontal axis have observed shifts where LCR$^+$ proteins are greater than the null expectation, while bars below the horizontal axis represent observed shifts where LCR$^+$ proteins are below the null expectation. Dotted horizontal lines indicate a significance threshold of 0.05. TAb shifts are consistently and significantly positive while the PAb shifts are consistently, significantly negative.

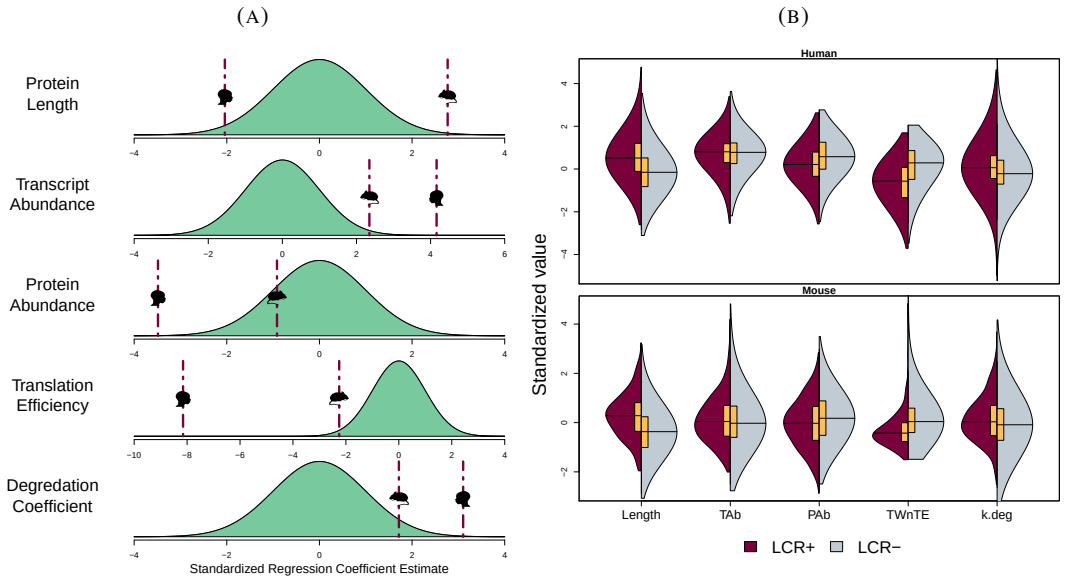

FIGURE 4.4: Logistic regression shows PAb and TAb are significantly associated with the probability of a protein containing an LCR. Regression is based on GTEx and PaxDB human abundance data as well as Schwanhäusser mouse data controlling for protein degradation and translation efficiency. Regressors are standardized so that the effect magnitudes may be compared (a) Estimated regression coefficients (maroon lines) are compared to a standard normal distribution (Teal). PAb and LCRs are negatively correlated, while the opposite is true for TAb. (b) Split violin plots showing the distributions of the regressors' values across LCR^+ (maroon) and LCR^- (grey) proteins. Yellow bars indicate the median, and interquartile for the distribution in which the bar is embedded.

finding LCRs in longer proteins. There were complete data for 3107 (29.1% LCR^+) human proteins and 2155 (20.7% LCR^+) mouse proteins. Figure 4.4 shows the standardized regression coefficients. Despite including two regulatory steps and the length of the proteins, PAb is still negatively associated with the presence of LCRs. From the coefficients in Table A2.3, we estimate that a human protein which has double the abundance of an otherwise similar protein would have 5.5 (95% CI [2.5,8.5])% lower odds of having an LCR. Conversely TAb is positively associated with the presence of LCRs. A doubling in TAb is associated with an 8.9 (95% CI [4.6,13])% increase in the odds of encoding an LCR. The PAb results are not significant for mice but tend towards a negative relationship with a 2.6 (95% CI [-3.0,7.9])% odds reduction for doubling PAb. The relationship with TAb is qualitatively the same between human and mouse proteins. For mice, the odds of encoding an LCR are 13 (95% CI [2.1,25])% higher for each doubling in TAb. There was no qualitative difference in the relationships to either PAb or TAb with changes to the translation parameters used when calculating TWnTE, or even whether TWnTE was included in the regression (Figure A2.1).

While PAb and degradation data are not as readily available across mammalian species, RNA-Seq data are plentiful. Therefore, transcriptomic data were processed together and

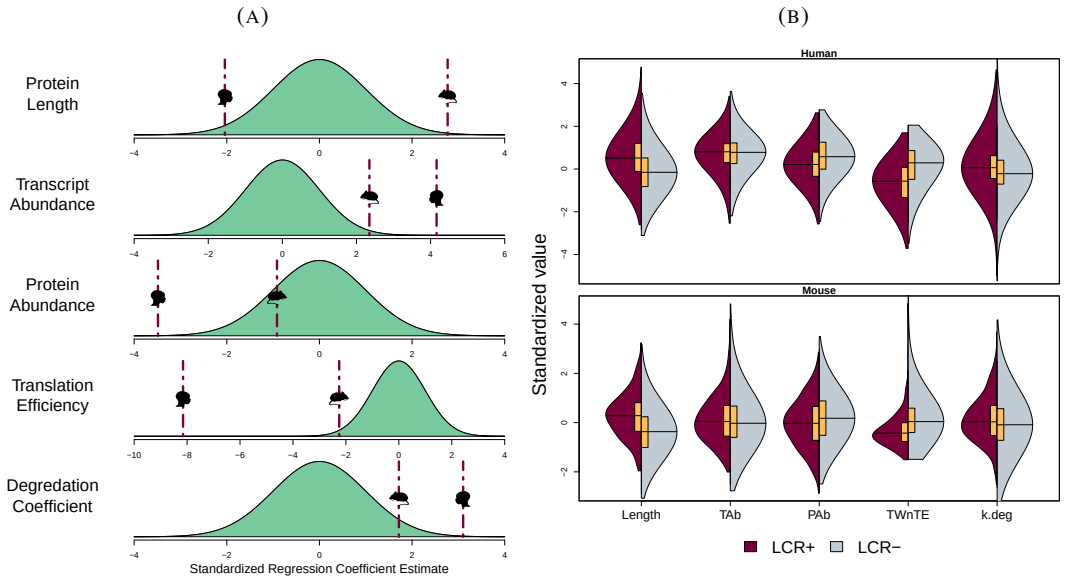

FIGURE 4.5: TAb is associated with an increased probability of an LCR being present, based on consistently processed RNA-Seq data from 9 mammalian species. Regressors are standardized so that the effect magnitudes may be compared. (a) Estimated regression coefficients (contrastingly coloured lines) are compared to a standard normal distribution (Teal). In all cases TAb is positively associated with the presence of LCRs. (b) Split violin plots showing the distributions of the regressors' values across LCR^+ (maroon) and LCR^- (grey) proteins. Yellow bars indicate the median, and interquartile for the distribution in which the bar is embedded.

aggregated for 9 mammalian species. As raw RNA-Seq results were not available for the GTEx or Schwanhäusser datasets, the human and mouse data are not the same as in the previous analyses. The results of a logistic regression of LCR status against TAb, protein length, and TWnTE can be seen in Figure 4.5. Each regression used data from at least 30k proteins, with LCR^+ rates between 23.0% (Human) and 27.6% (Horse). See Table A2.4 for details. The positive relationship between TAb and the presence of LCRs was qualitatively consistent across mammals. All estimates for the increase in odds of encoding an LCR were between 1.5% and 4.4% for each doubling in TAb.

4.4 Discussion

As would be expected if there were selective pressures against evolutionarily unstable regions in highly abundant proteins, we have found that PAb is negatively associated with LCRs. However, the opposite is true at the level of TAb where LCR encoding transcripts have higher abundance than expected. The observed associations are consistent across mammalian taxa. This is true even when accounting for two of the processes along the pathway from gene expression to protein degradation. This indicates that the associations between LCRs and

abundance cannot be explained solely by reduced translation efficiency of repetitive sequences or elevated degradation rates of LCR^+ proteins.

We investigated several technical explanations of the observed effect, the first of which was whether the effect was an artifact our choice of LCR threshold. We reanalyzed our data with minimum entropy thresholds ranging from 0 to 2.2 bits and observed that elevated TAb for LCR encoding transcripts is only observed for thresholds between 1 and 2. Increasing the entropy threshold dilutes the effect of LCRs via the inclusion of more false-positives in the LCR^+ category. This reduces the observed differences between the two groups. Conversely, the loss of signal with lower thresholds is due to a loss of statistical power. With the proportion of LCR^+ proteins dropping from 24% at a threshold of 1.9 down to 4% at a threshold of one, the power to detect an effect as large as we have observed drops below 0.5. Our chosen threshold strikes the balance between sample size, while still limiting the analysis to proteins with minimum entropies which are correlated with biological effects. This is further supported by the qualitatively similar results when looking at intrinsically disordered protein regions which often overlap LCRs (Figure 4.2).

We also investigated the possibility that bias in the mapping of short reads to highly repetitive sequences would explain the elevated TAb we observed. In that case, it would be expected that within a transcript, the LCR encoding region would have a higher depth of coverage than LCR^- regions. As we had access to the raw reads for the mammalian RNA-Seq experiments, we were able to evaluate this and found that there was no significant difference in depth of coverage for LCR^- encoding regions.

The data from the GTEx project is generated from human donors and time does pass between the death of the donor and stabilization of tissues for RNA-Seq. This ischemia time may have biased TAb towards more stable transcripts as unstable transcripts would be degraded during the ischemic window. If the observed shift in TAb were the result of this bias, it would suggest that LCR encoding transcripts are more stable. However, this is not the case. Almost identical effects are observed when using unadjusted or ischemia-time-adjusted TAb values (Figure 4.2). This indicates that the observed effect is not the result of a bias towards more stable transcripts.

The GTEx/PAb analysis had a small potential for redundancy as a result of protein isoforms. Of the 18,016 genes for which we obtained complete TAb, PAb, and LCR data, 49 had data from multiple isoforms. The abundance of each isoform was unique, and we observed near identical results to the baseline observations (Figure 4.2).

Chavali et al. (2017) have previously shown that yeast proteins which contain amino acid homopolymers have lower abundance than homopolymer free proteins. We repeated our permutation analysis twice: comparing only homopolymer containing proteins to LCR^- proteins and comparing only heteropolymer LCRs to LCR^- proteins (Figure 4.2). We found a much weaker signal for homopolymer containing proteins, likely due to a lack of statistical power as only 536 of the 4259 LCR^+ proteins had homopolymers. Homopolymer LCRs were qualitatively similar to the baseline results, but do not completely drive the effects we have observed, as they are consistent for heteropolymer LCRs as well.

Data availability presented a limitation to our ability to interrogate the biological mechanisms driving the elevated TAb of LCR encoding transcripts and the prevalence of elevated TAb and reduced PAb across mammals. For the latter, proteome wide PAb data are not widely available across mammals. However, the consistent observation of a positive association at the transcript level across mammals may indicate that the same relationship observed for humans and mice holds across mammals for PAb. Regardless, our work shows that there can be a disconnect between transcript and protein levels. This highlights the importance of carefully investigating RNA-Seq based conclusions to ensure that the physiologically relevant proteins are likewise up- or down-regulated.

The lack of translation rate data across mammals also limited the confidence in the accuracy of our calculations of TWnTE. However, we demonstrate that the particulars of the calculation did not significantly impact the conclusions about the relationships between TAb, PAb, and the presence of LCRs (Figure A2.1). Regardless, we believe that TWnTE is a useful method for calculating translation efficiency as it goes beyond merely considering sequence composition. In contrast to the standard normalized translation efficiency (nTE), TWnTE allows us to account for the inherently ordered nature of the codons in a transcript. This method of calculation clearly shows a difference in translation efficiency between transcripts for LCR^{+} and LCR^{-} proteins as seen in Figures 4.4 and 4.5.

The inclusion of TWnTE does make the interpretation of coefficients for the logistic regression more difficult. TWnTE, as calculated with a low coefficient of resupply, is highly correlated with protein length; longer proteins have lower TWnTE. This correlation and the differences in TWnTE between the GTEx human and Schwanhäusser mouse data cause the apparent difference in effect for protein length in Figure 4.4. However, the main goal of this analysis was to assess the relationships with TAb and PAb which are not strongly correlated with any of the other parameters. As a result, our conclusion that the presence of an LCRs is positively associated with TAb is unaffected.

Aside from data limitations, our analyses are also limited to an aggregate view across the wide variety of LCR compositions and properties. LCR composition is associated with variation in both transcript and protein abundance as shown by Cascarina and Ross (2018). They observed that PAb, nTE, and protein half-life can have qualitatively different relationships depending on the primary amino acid in an LCR (Cascarina and Ross 2018). Figure A2.2 shows the results of logistic regression for human GTEx data when the most prevalent amino acid in the minimum entropy regions of a protein is included as an interaction term with PAb and TAb. For statistically significant coefficient estimates, the observation made in aggregate holds true. PAb is negatively associated with LCRs, and TAb is positively associated with LCRs. Glycine, the least conformationally restricted amino acid, is the sole exception. The positive association of PAb with LCRs in proteins where glycine is the primary amino acid in low entropy regions may be driven by the high frequency of glycine in abundant structural proteins such as keratin (Parry and North 1998) and collagen (Persikov et al. 2000) which have repeating structures.

The structural function of these LCRs are undoubtedly a subset of the many important functional roles LCRs fulfil. These roles require a particular level of abundance to be maintained, leading to selective pressures on mechanisms which regulate abundance. Our

proposed explanation of the disconnect between TAb and PAb for LCR^+ proteins is that elevated TAb is an adaptive response to the appearance of LCRs in protein sequences. While the processes we investigated did not explain the disconnect, it is likely that through the combined effect on multiple regulatory processes LCRs lead to a reduction in steady-state protein levels. As these proteins still carry out important functions, there is a selective pressure to counter the LCR-associated reduction. Either increased transcription or stabilization of LCR encoding transcripts may be the specific adaptive response leading to elevated TAb. Follow-up studies will examine the ancestral states of the proteins and their abundance to examine this hypothesis. As well as to determine the specific biological mechanism leading to decreased PAb and yet increased TAb of LCR^+ proteins.

4.5 Materials and Methods

PAb data for human proteins were downloaded from PaxDb v4.1 (Wang et al. 2012). Data from brain, colon, esophagus, heart, kidney, liver, lung, ovary, pancreas, prostate, skin, testis, and uterus tissues were integrated with each protein being assigned abundance equal to the median abundance across tissues in which the protein was expressed. TAb data were downloaded from the GTEx project v8 (GTEx Consortium 2013). Data for the 13 tissues listed above were integrated in the same way to give a median across tissues where the transcript is expressed. The exclusion of tissues with zero measured expression maximizes the number of proteins which can be used in the analysis as half of transcripts have zero abundance in the majority of the selected tissues in the GTEx data. While there is variance in the abundance of a protein and its transcript across tissues, the sequences comprising LCRs remain constant across tissues.

Breadth of expression and initial expression were calculated from the raw GTEx TAb data. The former was quantified using the tau index (Yanai et al. 2005). This index ranges from 0 to 1, where 0 indicates a gene expressed in all tissues equally and a 1 indicates a tissue which is expressed only in one tissue. The TAb at time of death for GTEx data was estimated by fitting exponential curves to TAb as a function of ischemia time across samples for each transcript in each tissue.

Transcript and protein sequences were downloaded from the Ensembl database, release 99 (Howe et al. 2021). As identifiers used across studies differed, all sequence identifiers were mapped to UniProt protein identifiers using the UniProt Retrieve/ID mapping service (The UniProt Consortium 2017). The thirteen mitochondrial encoded proteins with both TAb and PAb data were excluded as mitochondrial genes are under fundamentally different constraints from the majority of nuclear genes.

LCRs in protein sequences were identified using the `Seg` algorithm (Wootton and Federhen 1993) using a window of 15 amino acids, a lower complexity bound of 1.9, and a higher complexity bound of 2.5 as these parameters were shown to detect longer, more repetitive regions in previous research (Golding 1999). This value also represents the lower inflection point in the distribution of minimum entropies across human proteins. The overlapping property of intrinsic disorder was also calculated. Proteins with intrinsically disordered regions

were identified using IUPred (Dosztányi et al. 2005) in glob mode. A protein was considered to have an intrinsically disordered region if it contained a non-globular region.

As LCRs range from homopolymer tracts to compositional bias we subdivided observed LCRs into homo- and heteropolymer LCRs. An LCR was considered a homopolymer if there was a contiguous tract of a single amino acid which made up at least half the length of the LCR.

Mouse TAb and PAb, as well as protein degradation rates, and translation rates were extracted from data generated by Schwanhäusser et al. (2011). Human Protein degradation rates were integrated from multiple sources (Doherty et al. 2009; Cambridge et al. 2011; Zhang et al. 2016; Zecha et al. 2018) by first converting all reported values to the coefficient of degradation. The geometric mean value of the coefficient across studies for each protein was used.

Translation efficiency was calculated based on the transcript sequences, in a method derived from the nTE scale (Pechmann and Frydman 2012). On this scale, the translation efficiency of a transcript is the geometric mean of the translation efficiencies for each codon within the transcript. The value for each codon is calculated as the ratio of the supply of tRNA isoacceptors for a codon to the global usage of that codon. This translation efficiency scale does not account for the ordering of codons within a transcript, which can have a profound effect through local tRNA depletion. For this work the codon usage values are calculated using Equation (4.1) as described by Pechmann and Frydman (2012), however the standard calculation of codon supply is treated as initial conditions for the translation of a transcript. For each subsequent codon in a transcript, the local supply of tRNA isoacceptors is updated according to Equation (4.2): accounting for perturbation of the local supply as well as resupply from the cellular environment. The perturbability is the proportion of the local supply used, scaled to the least supplied codon. Resupply is the portion of the local deficit which is ameliorated at each time step. Calculating TWnTE allows for codons which appear at the end of a repeat to have lower translation efficiency than those which appear alone or at the start of a repeat.

$$\mathrm{S}_{i,0} = \sum_{j=1}^{n_i} (1 - s_{i,j}) N_{i,j} / \max \mathrm{S}_0 \tag{4.1}$$

where:
- n_i = The number of tRNA isoacceptors for codon i
- $s_{i,j}$ = Wobble constraint between codon i and the jth tRNA
- $N_{i,j}$ = The copy number of codon i's jth tRNA

$$\mathrm{S}_{i,t} = \beta \mathrm{S}_{i,0} + (1 - \beta)(1 - \alpha_i)\mathrm{S}_{i,t-1} \tag{4.2}$$

where:
- $\mathrm{S}_{i,0}$ = Normalized initial codon supply
- $\mathrm{S}_{i,t}$ = Local codon supply for the codon i at time t
- α_i = Normalized perturbability for codon i
- β = Coefficient of resupply of tRNA isoacceptors

The selective constraints on wobble base pairing are a measure of how tolerant the ribosome is of different types of mismatches between codon-anticodon pairs. Most mismatches are not tolerated, but values were allowed to vary between 0 (tolerant) and 1 (intolerant) for A-A, U-G, G-U, and A-C mismatches. Wobble constraints were set for each organism by optimizing the correlation between codon supply and codon demand with R (R Core Team 2013) using the neldermead package (Bihorel and Baudin 2018), with initial conditions from estimates generated for yeast (dosReis et al. 2004). Codon supply was determined from genomic tRNA counts which can vary widely even in mammals. For example, *Bos taurus* (GCF_002263795.1) and *Rattus norvegicus* (GCF_000001895.5) respectively have 1637 and 377 annotated tRNA genes differentially distributed across potential anti-codons. Codon demand was determined from TAb weighted codon counts in the transcriptome. A consistent set of transcripts for codon usage calculations was constructed from across human transcripts which had data for both TAb and PAb available. For all other mammals the orthologous transcripts were determined based on mammalian orthogroups from PaxDb (Wang et al. 2012).

The perturbability and resupply parameters were selected by optimizing the correlation between calculated TWnTE values and measured translation rates across all proteins in the Schwänhausser dataset (Schwanhäusser et al. 2011). The optimization was performed with R (R Core Team 2013) using the neldermead package (Bihorel and Baudin 2018) with initial estimates of 0.5 for both parameters. The perturbability parameter is normalized to an organism's codon usage and tRNA availability, and the resupply parameter is based on basic diffusion. As only the Schwänhausser-based mouse dataset had translation rates, the TWnTE calculations for all other mammals used the same parameter values, under the necessary assumption that translation dynamics are consistent across the mammals tested.

Primate RNA-Seq data were acquired from the Non-Human Primate Reference Transcriptome Resource (Peng et al. 2015). Additional reads were downloaded via the Sequence Read Archive (Leinonen et al. 2011) for seven other transcriptomic studies (Brawand et al. 2011; Merkin et al. 2012; Fushan et al. 2015; Tang et al. 2017; Carelli et al. 2018; Valberg et al. 2018; Chen et al. 2019). The dataset assembled represents nine mammalian species: humans, chimpanzees, macaques, mice, rats, dogs, horses, cows, and pigs with data from six tissues: brain, heart, kidney, liver, lung, and muscle tissues. All RNA-Seq data were processed through the same pipeline to maximize consistency between datasets. Adapter removal, quality control, and read merging was performed using fastp (Chen et al. 2018) with quality windows of 4bp, minimum quality thresholds of 20, a minimum read length of 30bp, and merging any paired reads which overlapped by at least 20bp with 80% similarity. TAb quantification was performed with Salmon (Patro et al. 2017) using reference transcriptomes acquired from RefSeq (O'Leary et al. 2016), and the validate mappings flag. As orphan reads were generated during quality control, quantification was performed separately for orphaned and paired reads for each sample before pooling the library-size-adjusted results together.

The number of genes or proteins for which data were acquired can be found in Table A2.1. Total and LCR^{+} counts are broken down by data set, data type, species, and tissue.

Permutation testing was performed on the GTEx and PaxDB data by randomly shuffling the LCR status of proteins which had both TAb and PAb data. For each permutation, the 1st

quartile, median, and 3rd quartile of abundance was calculated for both the LCR^+ and LCR^- groups. The difference between the values was recorded to establish null distributions for each quantile where LCR status is unrelated to abundance. The results of the permutation test of median abundance were verified with a Mann-Whitney U test. The difference between each quantile was used rather than simply the difference between the median as it provides a better view of how LCRs are correlated with abundance across the wide distribution of abundances observed.

Permutation testing was also done as described above for several alternative conditions. To compare entropy-based and structure-based LCR identification, intrinsic disorder status was used as the permuted factor. To assess the effect of differential transcript stability, raw TAb was substituted with ischemia-time-adjusted TAb. To examine the effect of different classes of LCRs, separate permutation tests were performed which compared whether homo- or heteropolymer LCR^+ proteins to LCR^- proteins. To assess if redundancy from protein isoforms was affected results, permutations were done using a subset of the data such that only one transcript-protein pair was used for each gene.

Logistic regression was used to assess the probability of a protein containing an LCR given the TAb and PAb, while accounting for differences in protein degradation rates, translation efficiency, and protein length. All regressors were log transformed to meet the assumptions of linearity, then all regressors were standardized to allow comparisons of their effects on LCR probability. The fold change in odds for a unit change in each regressor can be obtained by natural exponentiation of the estimated regression coefficients. When performing logistic regression on the mammalian RNA-Seq data, only TAb and protein length were included as regressors. Regressions for all organisms were performed independently.

To evaluate the robustness of the analysis to the assumption that mouse translation parameters are applicable to other mammals, the logistic regression above was repeated for humans using standard nTE calculations, excluding translation efficiency from the model completely, and 25 pairs of parameter values evenly spread across the valid parameter space. The change to the estimated PAb and TAb coefficients was evaluated for qualitative changes to the conclusions.

4.6 Acknowledgments

This work was supported by the Natural Sciences and Engineering Council of Canada (PGSD3-547476-2020 and RGPIN-2020-0573).

Chapter 5

Low Complexity Regions as signals in Untranslated Regions

5.1 Preface

This chapter was aided by undergraduate volunteers, under my supervision. Daniel Ruiz performed the initial analysis showing that low complexity regions in untranslated regions (UTRs) have significant impacts on both the protein abundance and transcript abundance of downstream products. Megan Bilodeau performed a battery of analyses to examine which, if any, known signaling motifs in UTRs have overlap in presence and effect with untranslated low-complexity motifs. She also performed analysis to identify novel motifs. In both cases, I devised the research questions and methodology as well as interpretation of the results.

5.2 Introduction

Repetitive sequences are common not just in protein sequences but also non-coding regions throughout eukaryotic genomes. Repeats can be structurally important as in telomeres and centromeres (Price 1992) and they can also be sources for genetic variation (Tautz et al. 1986). These repetitive regions are ubiquitous, for example SINEs and LINEs (short and long interspersed nuclear elements), which make up 17 (Lander et al. 2001) and 13% (Zhang et al. 2021) of the human genome, respectively. Smaller scale repeats also exist: microsatellites are tandemly repeating elements in intergenic regions (Ellegren 2004). The evolution of these has best informed how we understand the evolution of repeats in coding regions as well.

The distinction between coding and non-coding repeats is important due to the significant physiological impacts for protein sequences. Protein low complexity regions (LCRs) are associated with abundance shifts at both the transcript and protein level (Dickson and Golding 2022). Some of the effects may be at the RNA level through changes in translation rate (Liu et al. 2016; Li et al. 2017). However, repeats in transcribed sequences can also have an effect during the transient RNA phase (Choi et al. 2009; Wang et al. 2017; DelCampo et al. 2015).

Transcribed but untranslated gene sequences can affect abundance through signaling transcription factors, transcript mobility and stability, and ribosomal occupancy (Kedersha and

Anderson 2002; Faux et al. 2005; Everett and Wood 2004; Horton et al. 2023). For example, myotonic dystrophy type 2 is caused by a tetra-nucleotide repeat in an intron of the CNBP gene (Ranum and Day 2002; Dere et al. 2004; Day and Ranum 2005; Musova et al. 2009). Even this transient inclusion of a repetitive element in the mRNA dis-regulates the abundance of the transcript and its protein. Beyond the extreme cases of pathological states, signaling motifs in the five and three prime untranslated regions (UTRs) of transcripts are well known and catalogued (Dalphin et al. 1996; Jacobs et al. 2000, 2002).

Some of the UTR signaling motifs are repetitive in nature, such as C-Rich stability elements (Kong and Liebhaber 2007). However, the overall effect of low-complexity (LC) motifs in UTR is not well understood. If the impacts of LC on abundance are purely the result of peri- and post-translational pressures, it would be expected that LC motifs in UTRs would not be strongly associated with changes in abundance. However, if LC motifs in UTR do have significant effects it may be due to overlap with known or potentially novel signaling motifs.

In this work we investigate the relationship between transcript abundance (TAb), protein abundance (PAb), and the presence of untranslated low-complexity motifs (ULCM). And further investigate the degree of overlap between known signaling motifs in UTRs.

5.3 Methods

PAb data for human proteins were downloaded from PaxDb v4.1 (Wang et al. 2012). Data from brain, colon, esophagus, heart, kidney, liver, lung, ovary, pancreas, prostate, skin, testis, and uterus tissues were integrated with each protein being assigned abundance equal to the median abundance across tissues in which the protein was expressed. Tissues where no expression was detected, or enumerated were excluded from this calculation. TAb data were downloaded from the GTEx project v8 (GTEx Consortium 2013). Data for the 13 tissues listed above were integrated in the same way to give a median across tissues where the transcript is expressed.

Gene sequences were taken from the annotated human genome (GRCh38 Schneider et al. 2016). LCRs were identified in coding sequences using `Seg` (Wootton and Federhen 1993), with a window of 15 amino acids, a lower entropy bound (K_1) of 1.9 bits, and an upper entropy bound (K_2) of 2.2bits. A K_1 of 1.9 has been used in several previous publications (Huntley and Golding 2000, 2002; Lenz et al. 2014; Dickson and Golding 2022) and found empirically to give good results. It is also the lower inflection point in the distribution of minimum entropies across human proteins. A modified version of `Seg` which properly accounts for the DNA alphabet (Enright et al. 2023) was used to identify ULCMs and LCRs. Window length of 45bp for DNA and 15 amino acids for protein were used. The K_1 values in each of the $5'$ and $3'$ UTRs, the coding sequence, and the protein sequence were determined by taking the lower inflection point in the distribution of minimum entropies across human transcripts. The mode of this distribution was taken to be the value of K_2. Each transcript region has its own distribution and K_1 and K_2 values.

The relationship of PAb and TAb with the presence of LCRs and ULCM was investigated using multivariate linear models implemented in R (R Core Team 2022). Later analyses which

included known signaling motifs and subdivisions of ULCM were also evaluated in this manner.

A set of known motifs in 5′ and 3′ UTRs was acquired from the TransTerm database (Jacobs et al. 2002). `PatScan` (Overbeek et al. 2010) was used to search for the known motifs in 5′ ULCMs. Motifs from both UTRs were acquired and searched for in both the 5′ and 3′ to investigate whether motifs canonically known to have effects when in one region have similar effects in the other.

A python script (Van Rossum and Drake 2009) was written to *de novo* identify ULCM motifs associated with changes in TAb. The general approach is seed based where the highest scoring motif is found between the first two sequences, then the best match to that motif is found in all subsequent sequences. Averaging the motifs resulting from repeatedly, randomly selecting the pair of sequences for the seed each time converges to the identified motif.

In general, the input sequences to such a program will have some *a priori* knowledge that a motif will be present, for example a collection of regions upstream of a transcription start site in genes known to be regulated by the same transcription factor. In our case, the only knowledge we have is associations with TAb. It is not reasonable to expect that all UTR with a ULCM will have the same TAb associated motif. To address this, we take random bootstraps of the ULCM sequences, where the weighting for each sequence is proportional to the TAb. In this way we can bias the input sequences to those more likely to have a common motif which can be identified. In our implementation, sequences were randomly selected from the input with replacement with the probability of a particular ULCM being selected weighted based on TAb. The ULCM with the lowest TAb is given a weight of 1 and all others are given a logarithmically scaling weight of at least 1. All weights (W) were truncated to the nearest integer. The weight was calculated as

$$W_i = \left\lfloor \log_2 \frac{x_i}{\min x} \right\rfloor + 1, \tag{5.1}$$

where x_i is the TAb for the i^{th} ULCM. As the entire bootstrap was randomly selected, the first two sequences, which are used for the seed, were also random.

The score used when comparing motifs was based on the information content of the resulting nucleotide profile. The profile being the frequency of each nucleotide at each position in the motif. The score for profile P was calculated as:

$$P = \sum_{i=1}^{k} 2 - H(f_i), \tag{5.2}$$

where k is the length of the motif, f_i is the frequency of each nucleotide at the i^{th} position of the profile, and $H(x)$ is the Shannon Entropy of a frequency table, calculated according to Equation (3.1). The total information content score was used as high conservation is correlated with high information content while still accounting for potential degeneracy. Additionally, the scale of information content does not depend on the number of input sequences making scores comparable between differently sized sets.

To find the seed for a motif, every pairwise comparison of k length subsequence from the first two ULCMs was made, and the seed was the pair with the highest scoring profile. Every k length subsequence of the third ULCM was then compared to the seed to find the profile of three sequences with the highest score. This was repeated until all ULCM in the bootstrap have been added to the profile. At this point the frequency of each nucleotide at each position in the profile was weighted by the score of the profile and added to an overall profile. In this way, high scoring profiles contribute more to the overall profile than low scoring profiles. After the desired number of 'permutations' have been completed the final identified motif was the consensus of the overall profile. This entire procedure was repeated for values of k between 10 and 30 bp, each with 400 permutations.

5.4 Results

To determine entropy thresholds for `Seg` to use in the identification of ULCM we examined the distribution of minimum entropies in 5′ UTRs of 17,919 mRNA sequences. The distributions for each transcript region, and the translated protein sequence can be seen in Figure 5.1. The K_1 for 5′ and 3′ UTRs was found to be 1.38 bits, while the higher complexity coding and protein sequences had K_1 values of 1.52, and 1.97 bits respectively.

We compiled data from 15,184 proteins for which non-zero transcript and protein abundance data were available and identified ULCMs in their 5′ and 3′ UTRs, as well as LCRs in the translated coding sequences. We observed 2255 (14.8%) of 5′ UTRs contained a ULCM, while 3874 (25.5%) of 3′ UTRs had a ULCM and 4733 (31.2%) of proteins had an LCR. The correlation in LC was low between the transcript regions. The highest observed correlation was between the presence of a ULCM in the 5′ UTR and the presence of an LCR in the protein sequence ($r = 0.10$). We also observed low correlation between the lengths of the transcript regions, the highest correlation being between the lengths of the UTRs at $r = 0.08$.

We performed multivariate linear regression with TAb and PAb as regressands and the Boolean status of the presence of a ULCM or an LCR in the UTRs or protein sequence as regressors. The lengths of each region were included as nuisance variables. The full results of this regression can be found in Table 5.1. The presence of a ULCM in the 5′ UTR was significantly associated with any change in abundance: a positive association with TAb. Protein LCRs were positively associated with TAb and negatively associated with PAb.

All further investigations focused on the 5′ UTR and attempted to identify known or novel motifs associated with the observed impacts on TAb. PatScan was used to find known translation motifs from the TransTerm database in 5′ UTRs of 16,049 transcripts. Considering the UTRs in which known motifs are found and the UTRs in which ULCM were found, the known motifs with the most similar pattern of presence and absence were m^1A methylation sites, pseudoknot like structures, and stem loop structures. The former is post-transcriptional modification where an adenine ribonucleotide has a methyl group appended to the pyrimidine ring. This interferes with base pairing with functional consequences during translation (Shima and Igarashi 2020). For the top three motifs, Jaccard similarity indices with ULCMs were 0.23, 0.17, and 0.14, respectively. All known motifs detected as well as their correlation and Jaccard

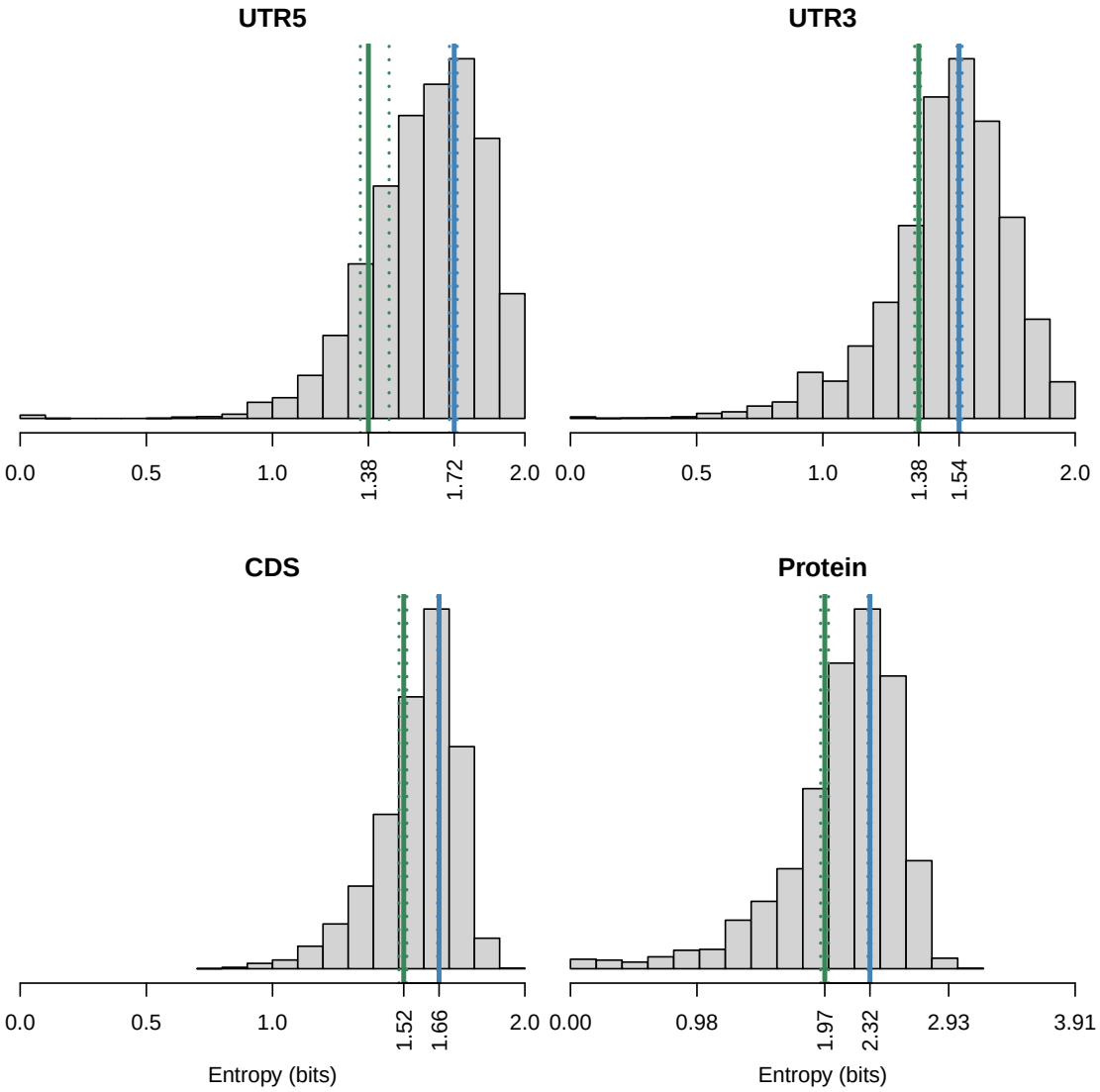

FIGURE 5.1: The distribution of minimum entropies in 45 bp (or 15 amino acid) windows across 17,919 mRNA sequences separated by transcript region. The lower inflection point and maximum of the distribution are marked, as well as 95% confidence intervals based on 1000 bootstraps of the mRNA sequences. These values were used as lower and upper entropy thresholds for ULCM identification with `Seg`.

TABLE 5.1: Multivariate Linear Regression of Abundance against LC presence

Region	Variable	TAb		PAb		ANOVA $\log_{10}$P
		Coef	(std.err)	Coef	(std.err)	
UTR5	ULCM$^+$	0.559	(0.065)	2.28	(0.056)	-16
	Length	-8.09×10^{-4}	(9.2×10^{-5})	-8.48×10^{-4}	1.2×10^{-4}	-21
Protein	LCR$^+$	0.302	(0.050)	-0.213	(0.067)	-22
	Length	-5.81×10^{-4}	(3.7×10^{-5})	-1.08×10^{-3}	5.0×10^{-5}	-114
UTR3	ULCM$^+$	-0.114	(0.058)	-0.104	(0.077)	-4.4
	Length	-2.19×10^{-5}	(1.4×10^{-5})	-1.64×10^{-5}	1.8×10^{-5}	-0.59

similarity with ULCMs can be found in Table 5.2. Of the 7561 UTRs where m^1A methylation sites were found, in 1060 of them the motif was entirely included within a ULCM.

We repeated the linear regression above but altered the 5′ UTR regressor to include information on the presence of m^1A methylation sites, the motif with the highest Jaccard similarity to ULCM. Instead of a Boolean variable it was considered as a factor with three levels: no ULCM, a ULCM without an m^1A methylation site, and a ULCM containing an m^1A methylation site. The estimated regression coefficient for TAb values for the latter (0.894 ± 0.090) was 1.6 times higher than the original estimate for simple ULCM presence (0.559 ± 0.065 Table 5.1). The coefficient for ULCMs without m^1A methylation sites was 2.1 times smaller at 0.264 ± 0.085. The coefficients for other transcript regions were not significantly different between the two regressions.

We used a method which bootstraps the ULCM sequences, weighted by their transcripts' abundance to *de novo* identify any motifs in 5′ UTR associated with elevated TAb. For all motif lengths tested we found that poly-cytosine motifs were most conserved and most TAb associated. (Figure 5.2). The sequences identified also indicate that guanine residues are also often present in the motif.

5.5 Discussion

We have examined human abundance data for both proteins and transcripts and investigated their relationship with the presence of ULCM. Accounting for LC in UTRs we were able to recapitulate previous results demonstrating differential impacts at the transcript and protein levels associated with the presence of LCRs in protein sequences. We have additionally shown that the presence of 5′ ULCM is positively associated with TAb, even after removing the influence of LC and length in other transcript regions. Further investigating 5′ ULCM, we found the known motif whose pattern of presence and absence was most similar to and correlated with ULCM presence was m^1A methylation sites. These sites account for a large fraction of the relationship between 5′ ULCM and TAb. We attempted to *de novo* identify any novel TAb associated motifs and identified GC rich sequences with a cytosine bias as the best candidate.

TABLE 5.2: Known translation motifs found in human 5′ UTR, sorted by Jaccard Index

Motif Name	Proteins with Motif Count	(%)	Overlap with ULCM Correlation	Jaccard Index
m1A methylation site	7561	(47.11)	0.26	0.23
Pseudoknot like structure	9944	(61.96)	0.12	0.17
Stem loop structure	15854	(98.78)	-0.01	0.14
CAG Element	3569	(22.24)	0.04	0.11
Musashi binding element (MBE)	5310	(33.09)	-0.01	0.11
Plant Polyadenylation Element	7782	(48.49)	-0.03	0.11
CUG Element	2849	(17.75)	0.03	0.1
Mammalian Polyadenylation Element	816	(5.08)	0.02	0.05
K-Box (KB)	778	(4.85)	0	0.04
Readthrough Element BYDV	303	(1.89)	0.09	0.04
Translational control sequence (TCS)	576	(3.59)	0.03	0.04
In vitro selected consensus stability element	686	(4.27)	-0.01	0.03
15-LOX-DICE Element	195	(1.22)	0.05	0.02
GU-Rich Element (GRE)	83	(0.52)	0.07	0.02
Yeast Polyadenylation element	259	(1.61)	0.01	0.02
AU-Rich Stability Element (ARE)	65	(0.41)	0.03	0.01
Brd-Box (Brd)	137	(0.85)	0	0.01
C-Rich Stability Element	95	(0.59)	0.05	0.01
Cytoplasmic polyadenylation element consensus (CPE)	84	(0.52)	0.01	0.01
GU-rich destabilization element UTRSite	51	(0.32)	0.04	0.01
Pumilio binding element (PBE)	88	(0.55)	0.01	0.01
tRNA like structure	114	(0.71)	0.02	0.01
Actin Localising Element	3	(0.02)	0.02	0
ADH_DRE Stability Element	95	(0.59)	-0.01	0
ARE database (ARED) Cluster III	4	(0.02)	0.02	0
ARE database (ARED) Cluster V	25	(0.16)	0.03	0
Yeast Puf4 consensus element	10	(0.06)	0	0
Cytoplasmic polyadenylation element (CPE) UTRSite	2	(0.01)	0	0
Cytoplasmic polyadenylation element non-consensus	36	(0.22)	0	0
Dinucleotide Repeat	5	(0.03)	0.04	0
Elastin G3A 3'UTR stability motif (G3A)	7	(0.04)	0.02	0
FMRP Translational regulator G-quartet	8	(0.05)	0.01	0
GLUT1 3 Prime Stability Element	2	(0.01)	0	0
Yeast Puf5 consensus element	20	(0.12)	0.02	0
HAC1 3' Binding Element (3'BE) yeast	3	(0.02)	0.01	0
Histone 3 prime stemloop UTRSite	1	(0.01)	0	0
Iron Responsive Element (IRE)	6	(0.04)	0	0
Yeast Puf3 consensus element	12	(0.07)	0	0
Mitochondrial Ribosomal Prot S12 Transl Cont Element	1	(0.01)	0	0
Musashi binding element (MBE) UTRSite	9	(0.06)	0.03	0
MYC mRNA localisation element	1	(0.01)	0	0
Nanos translation control element (NANOS_TCE) UTRSIte	62	(0.39)	0.01	0
Readthrough Element -from TMV	26	(0.16)	0	0
SAUR Plant Stability Element	1	(0.01)	0	0
Yeast Down Stream Element	5	(0.03)	-0.01	0

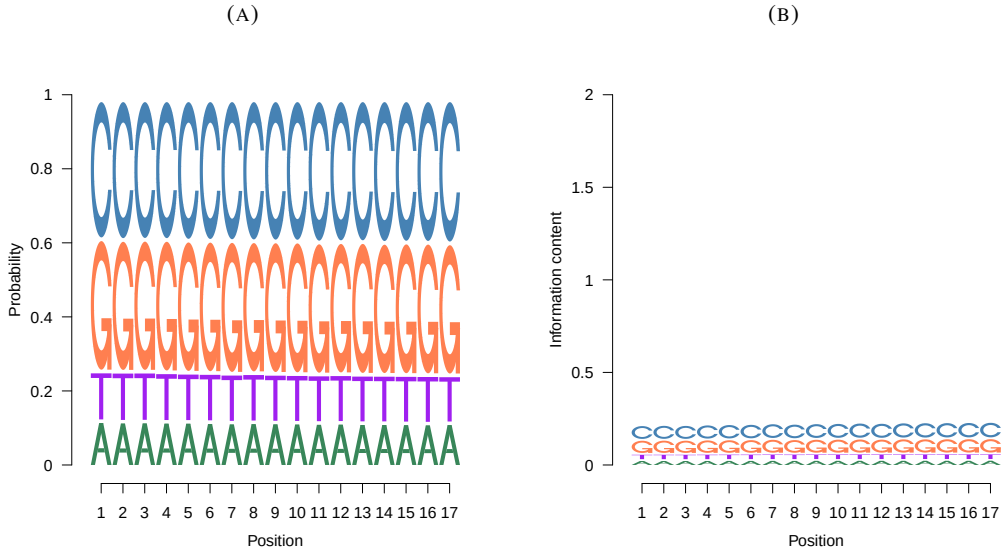

FIGURE 5.2: Sequence logos for best 17bp motif *de novo* identified in 5′ ULCMs. The probability of each nucleotide is displayed in a, and the information content in b. The LC Sequences most associated with TAb elevation in 5′ UTRs are C/G rich sequences.

It must be noted that the results of the analyses performed are influenced by the choice of entropy thresholds. We examined this impact by varying K_1 for each transcript region independently and observing how this would affect the linear regression. Specifically, we examined whether different choices of K_1 for any transcript region might alter the estimated change in TAb and PAb due to the presence of LC in each transcript region. The results of this test are shown in Figure 5.3. For all cases where the K_1 value was varied in one transcript region, the coefficient estimates for the other two regions never qualitatively changed: the coefficient values were always had the same sign regardless of K_1 value used. However, signs do change for the transcript region being varied. In all three transcript regions, if the K_1 value for that region was less than one bit, the relationship with either abundance would be negative, however these tend to be extreme sequences with very few transcripts having entropies in this range (Figure 5.1). For all three transcript regions the coefficient estimates are qualitatively stable for values near the inflection-point-based choice of K_1. Conclusions again change for high K_1 values. However, these are unreasonable thresholds as they are higher than the maximum-based choice of K_2; a definition of low entropy sequences which include those which minimum entropies higher than the modal minimum entropy value is a poor definition of low. While the conclusions are not completely robust to choice of K_1, our analysis is based on choices which adequately bracket LC sequences. Our conclusions are robust to choices of K_1 in the neighbourhood of our chosen K_1 values.

The known motifs which we observed to have the highest Jaccard similarity in presence or absence to 5′ ULCM were ubiquitous structural motifs and m^1A methylation sites. The former include sequences associated with pseudoknot-like and stem loop structures and were near

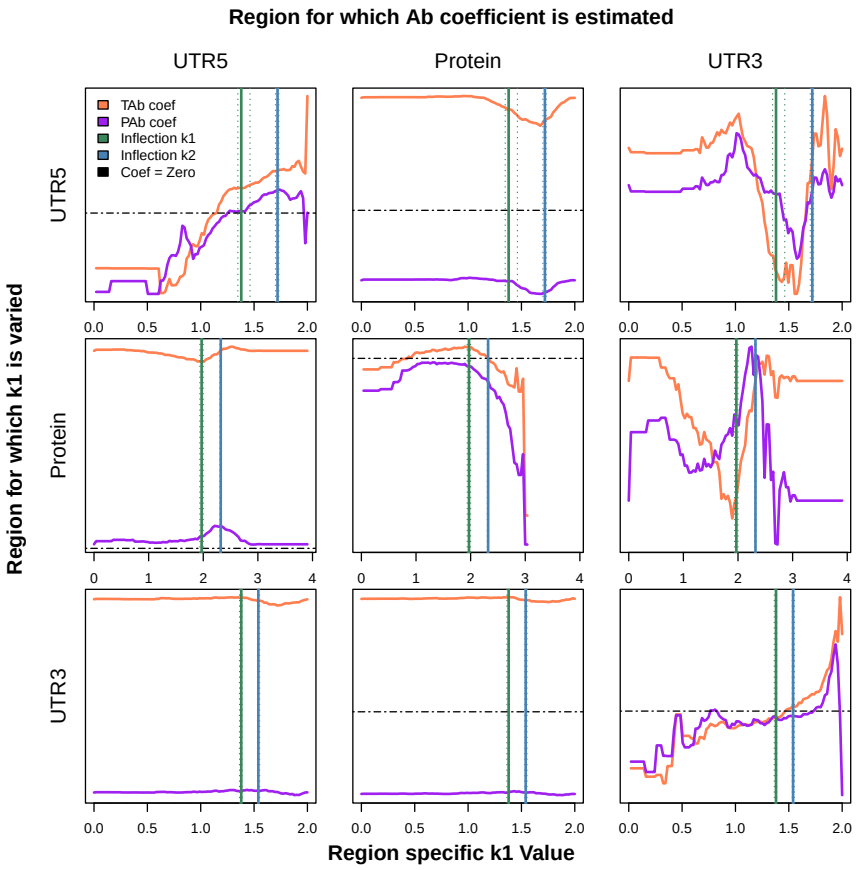

FIGURE 5.3: Qualitative conclusions about the relationship between ULCM and abundance are robust to choice of lower entropy bound (K_1). Each plot shows how the K_1 used to identify LC in the y-axis transcript region affects the estimated change in both TAb and PAb based on the LC status of the x-axis transcript region. The main diagonal shows how changing K_1 for a region affects its own abundance coefficients, while the off diagonals show the varied K_1 affects estimates in other regions. The orange and purple traces show the TAb and PAb coefficients respectively. The entropy thresholds identified based on critical points in the distribution are indicated by the green and blue vertical lines. The horizontal dotted line indicates a coefficient estimate of zero, if there is no such line then the coefficient estimates are always qualitatively the same regardless of K_1 value. The abundance traces crossing the zero-line indicate a qualitative change in conclusions. No trace ever crosses the zero line in the off diagonals. On the main diagonal the inflection point based K_1 value used in the analysis is in a region where small changes in the K_1 value do not qualitatively alter the conclusions.

ubiquitous: respectively, 62 and 99%, of 5′ UTRs having at least one of these structure-associated motifs. While mRNA structure is established as influencing TAb and PAb (DelCampo et al. 2015; Wang et al. 2017), the ubiquity of these motifs makes it unlikely that they explain the observed association between TAb and ULCM. In contrast, m^1A methylation sites are somewhat more rare (47% of UTRs) and have higher correlation and similarity to the pattern of ULCMs. These motifs are well studied for their critical role in tRNA structure and stabilization (Basavappa and Sigler 1991; Liu et al. 2016), but less well characterized in mRNA. There is evidence that m^1A methylation in coding sequences hampers translation, however in the 5′ UTR this modification is associated with greater translation efficiency (Li et al. 2017). We observed that m^1A methylation sites embedded in ULCMs account for a large portion of the positive association TAb between 5′ ULCMs. This may indicate that m^1A methylation also plays a role in stabilizing mRNA.

The m^1A methylation site motif from TransTerm (Dalphin et al. 1996) consists of a 17bp sequence with a central adenine residue (the methylated base) preceded by a GC dinucleotide, the remainder of the motif is either guanine or cytosine residues. While our attempts to *de novo* identify TAb associated motifs did not recapitulate this sequence, we consistently found GC rich sequences with a cytosine bias to be the TAb associated motif in 5′ ULCM. This was regardless of the motif length investigated. Poly-cytosine motifs may be targets for poly-C binding proteins which are known regulators of gene expression (Choi et al. 2009). These bonding proteins effects can be tissue specific, for example they are well known regulators in red blood cell development (Zhao et al. 2022). While C-rich elements are known to promote stability, they are generally observed in 3′ UTRs. Our observations show an expanded range of action for poly-C binding proteins is possible.

Our work shows that ULCMs in 5′ UTRs are associated with elevated transcript abundance. This appears to be largely, but not entirely, explained by GC-rich, and especially C-rich LC regions associated with post-transcriptional modification and protein binding regions which increase transcript stability and translation efficiency. Next steps could include experimental verification of m^1A methylation. It is unclear if these potential sites are truly modified *in vivo*. If so the impact of the surrounding context of the larger ULCM on the effectiveness and frequency of the modification could be investigated. It is also unclear whether the potential regulatory mechanism has any tissue specificity. Our work advances our understanding of the breadth of potential regulation of TAb and opens avenues towards further investigation of this complex and critical biological process.

Chapter 6

Modeling the Co-evolution of Low Complexity Regions and Transcript Abundance

6.1 Preface

At the time of writing, this chapter has been submitted to the Journal of Molecular Evolution and is under review. The submitted manuscript is entitled "Evolution of Transcript Abundance is Influenced by Indels in Encoded Protein Low Complexity Regions". The Authors of the paper are Zachery W Dickson, and G Brian Golding. GBG contributed to the conceptualization of the project, especially the suggestion of an ABC approach, and edited the manuscript. ZWD executed the project including data collection and processing, coding, and writing to the manuscript.

The goal of this work was to identify the temporal order of the evolution of low complexity regions (LCRs) and changes in transcript abundance (TAb). Based on the results described in Chapter 4 we hypothesized the observed elevation in TAb was an adaptive response to the appearance of LCRs in the protein sequence. Using data from human data to examine a short evolutionary time scale in combination with ancestral reconstruction and an ABC based modelling approach, we showed that co-evolution between LCRs and TAb best explains the data. The short timescale examined does not completely describe the evolutionary history of these data but demonstrates that examining multiple outcomes gives a more complete image of that history.

6.2 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 low complexity regions (LCRs) were once considered ‘junk’ protein sequences, at most spacers between more traditional protein domains (Golding 1999).

The proteins which contain them have since been shown to have important roles enabled by their LCRs which confer a host a properties depending on their specific amino acid composition. Many LCRs are intrinsically disordered at physiological conditions (Romero et al. 2001), but changes in those conditions can lead to conformal 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 (e.g. Nab3; Loya et al. 2017) and splicing (e.g. 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. 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 (Persikov 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 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 2006), 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 transcript abundance (TAb) for LCR^+ proteins may be 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 evolutionary 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 such as Brownian motion (Bedford and Hartl 2009) or Ornstein–Uhlenbeck (OU) process (Rohlfs 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, an approximate Bayesian calculation (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. Using these methods one can also infer parameters in a multivariate space. Haba and Kutsukake (2019) used an 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 interest, PAb data on a proteome level for individuals is not generally widely available. We characterize changes in LCR sequences and TAb across individuals and use an ABC to estimate the degree of interaction and temporal order of these evolutionary events.

6.3 Materials and Methods

6.3.1 Overview

In order 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.

6.3.2 Genomic and Transcriptomic Data

Human data was 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 Table A3.1. 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.

6.3.3 Construction of phylogenetic tree

The phylogenetic tree was constructed on the basis of SNPs on chromosome 19 of the human genome. 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). For the purposes of outgrouping, the chimpanzee reference genome (GCF_002880755.1; Sequencing and Consortium 2005) was used. Human SNP calls were acquired from the IGSR. Chimpanzee SNPs were generated by mapping fragments of the chimpanzee reference chromosome 19 to the homologous human chromosome 19. The 150bp fragments were generated by sliding a 500bp window in 200bp overlapping increments and taking the first and last 150bp 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.

6.3.4 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).

6.3.5 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 non-sense 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 as 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 (K_1) of 1.9 bits and an upper entropy bound (K_2) of 2.2 bits. We use 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 a conserved minimum entropy region (CMER) to deal with this. A 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, a K_1 equal to the entropy of the minimum entropy window, and a K_2 0.3 bits higher than K_1. 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 a CMER. For individuals, the length and entropy of the CMER are calculated from the gap free sequence. An example can be found in Table 6.1. 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.

Ancestral reconstruction of indel events in CMERs was performed using a parsimony based

TABLE 6.1: 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 (176 - 192)		
	Alignment	Entropy	Length	Alignment	Entropy	Length
Consensus	...ALPRLLLP--LLLLPAA...	1.574	15	...PGAGGTPGGPGGGGAP...	1.609	17
HG00096	...ALPRLLLP--LLLLPAA...	1.673	15	...PGAGGTPGGPGGGGAP...	1.609	17
HG00732	...ALPRLLLPLLLLLLPAA...	1.574	17	...PGAGGTPGGPGGGGAP...	1.609	17

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:

$$P(\Delta L = \ell) = \frac{\omega t^{\ell} e^{-\omega t}}{\ell!}, \tag{6.1}$$

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:

$$\omega = \frac{2}{\widetilde{\text{MAD}} \cdot \bar{D}}, \tag{6.2}$$

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

$$\text{MAD}_i = \frac{\sum_{u=1}^{m} |x_{i,u} - \bar{x}_i|}{m}, \tag{6.3}$$

where m is the number of individuals, and $x_{i,u}$ is the CMER length of protein i for individual u. The mean pairwise distance between leaves is calculated as:

$$\bar{D} = \frac{2}{n(n+1)} \sum_{u=1}^{n-1} \sum_{v=u+1}^{n} D_{u,v}, \tag{6.4}$$

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

6.3.6 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 (6.5) and (6.10) describe the co-evolution of CMER and TAb along any given branch of the tree.

The length of a 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:

$$L_{T+t} = M \cdot \max\left[L_T + N_\lambda - N_\kappa, 0\right], \tag{6.5}$$

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:

$$N_x = Pois(xL_T\Upsilon t), \tag{6.6}$$

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

$$\Upsilon = (A_T/A_0)^\upsilon, \tag{6.7}$$

where A_T is the abundance at the parent node, A_0 is the abundance at the root node and υ is the strength of indel dependence on abundance. A positive υ 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 υ.

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).

$$M = \frac{\max_{i=1}^{R}(X_i)}{\Sigma_{i=1}^{R} X_i}, \tag{6.8}$$

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:

$$R = Pois(\mu L_T t), \tag{6.9}$$

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:

$$A_{T+t} = \bar{A}_{T+t} \cdot e^{Norm(0,\sigma t)}, \tag{6.10}$$

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).

$$\bar{A}_{T+t} = A_T e^{-\delta t} + A_0 \Delta_\tau (1 - e^{\delta t}), \tag{6.11}$$

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 reasonable given the short evolutionary timescale and the amount of 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:

$$\Delta_\tau = (L_T/L_0)^\tau, \tag{6.12}$$

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 led to more consistent results.

6.3.7 Approximate Bayesian calculation

As the parameters of interest (τ: effect of length on optimum abundance and υ: effect of abundance on indel rate) could only be meaningfully assessed for proteins which had variation in length, the model was fitted using an 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 υ were $Normal(0,2)$ and $Uniform(-1,1)$ respectively. The full priors for all models can be found in Table A3.2. Three special cases of the full model were investigated: τ was fixed at zero (-tau), υ was fixed at zero (-upsilon), and both were fixed at zero (-tau-upsilon).

The fixing of the τ and υ 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 υ parameter describes the effect of changes in TAb on the indel rates in CMER. Therefore, fixing υ at zero in the -upsilon model assumes indel rates are unaffected by changes in TAb. Setting both to zero in the -tau-upsilon 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 Equation (6.11). 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 Equation (6.9) and was unnecessary to explore. The υ 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 TAb values was observed to be relatively high the result of Equation (6.7) can be extreme for absolute υ values above one. In many cases the tolerances of tools for numerically evaluating the results are exceeded. Bounding absolute υ below one made computation possible while still allowing investigation of the qualitative outcomes of positive, negative, and zero values for υ.

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 Equations (6.5) and (6.10). 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

$$E_{i,j,k,x} = \frac{|O_{i,j,k,x} - S_{i,j,k,x}|}{O_{i,j,k,x}}, \tag{6.13}$$

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:

$$C_{i,j,k,x} = \begin{cases} 1 - \frac{E_{i,j,k,x}}{\varepsilon}, & E_{i,j,k,x} < \varepsilon \\ 0, & \text{otherwise} \end{cases} \tag{6.14}$$

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:

$$\hat{\mathscr{L}}_{i,j,x} = \frac{1 + \sum_{k=1}^{s} C_{i,j,k,x}}{r+2}, \tag{6.15}$$

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

$$\hat{\mathscr{L}} = \prod_{i=1}^{n} \prod_{j=1}^{m} \hat{\mathscr{L}}_{i,j,L} \cdot \hat{\mathscr{L}}_{i,j,A}. \tag{6.16}$$

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', not accepting any proposals. If a model was evaluated on the higher end of the distribution for a given set of parameters, nearby proposals will appear to have lower relative likelihoods. This could be corrected for by evaluating each set of parameters multiple times, but this becomes computationally prohibitive. Instead, a parameter which adjusts how generously proposals are interpreted was included. For a single likelihood estimation, we cannot know where it came from in the distribution. If we interpret proposals generously, we assume that the proposals evaluation was in the lower end of its distribution, and the current model's evaluation was in the higher end of its distribution. Assuming the variance is constant between the two distributions 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:

$$G = Norm^{*}_{1-\alpha} \cdot \sqrt{2}\sigma_{\ln \hat{\mathscr{L}}}, \tag{6.17}$$

where $\sigma_{\ln \hat{\mathscr{L}}}$ 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, we can update this value as necessary, decreasing α if the chain is 'stuck' and increasing α if it is exploring excessively. We periodically estimated the chain's simulation variance 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 some target swap rate. 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 Markov Chain Monte Carlo (MCMC) run: increased or decreased in order to keep proposal acceptance rates at some target value. Both proposal scaling and proposal generosity are adapted in an attempt to achieve a

target acceptance rate of 23.4% which has been shown to lead to optimal mixing in many cases (Schmon and Gagnon 2022). 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.

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

$$P = \min\left[1, \left(\frac{e^{G}\hat{\mathscr{L}}_{p}\pi(p)}{\hat{\mathscr{L}}_{m}\pi(m)}\right)^{T}\right], \tag{6.18}$$

where $\pi(x)$ is the prior density of the parameter set for the current model (m) or the proposed model (p), 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%

6.3.8 Model Analysis

After an 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.

6.3.9 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:
`github.com/zacherydickson/AncRecon-LCR-TAb`

The program written to perform ABC inference of co-evolutionary models can be found on GitHub at: `github.com/zacherydickson/ABC-LCR-TAb`

6.4 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. 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 Figure 6.1a.

The inference of indel events also provides a reconstruction of the CMER length for the LCA of the individuals. The proteins can be subdivided into LCR$^+$ and LCR$^-$ by comparing the entropy of their most extreme CMER to a K_1 of 1.9 at the LCA. We observe that 1856 (25.3%) of proteins contained an 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 K_1 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 K_1, causing a 'loss' of the ancestral LCR. These two proteins are specifically marked points in Figure 6.1.

In Figure 6.1c 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 Figure 6.1b. 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 6.2. 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 υ (the degree to which shifts in TAb impact indel rates) at zero, and assumes CMER indels are TAb-independent. As υ is fixed it has fewer free parameters and was more consistent than the full model which was the model with the next highest likelihood. The best full replicate had a log likelihood of -5387 ± 1.924. With an already higher likelihood, the -upsilon model also

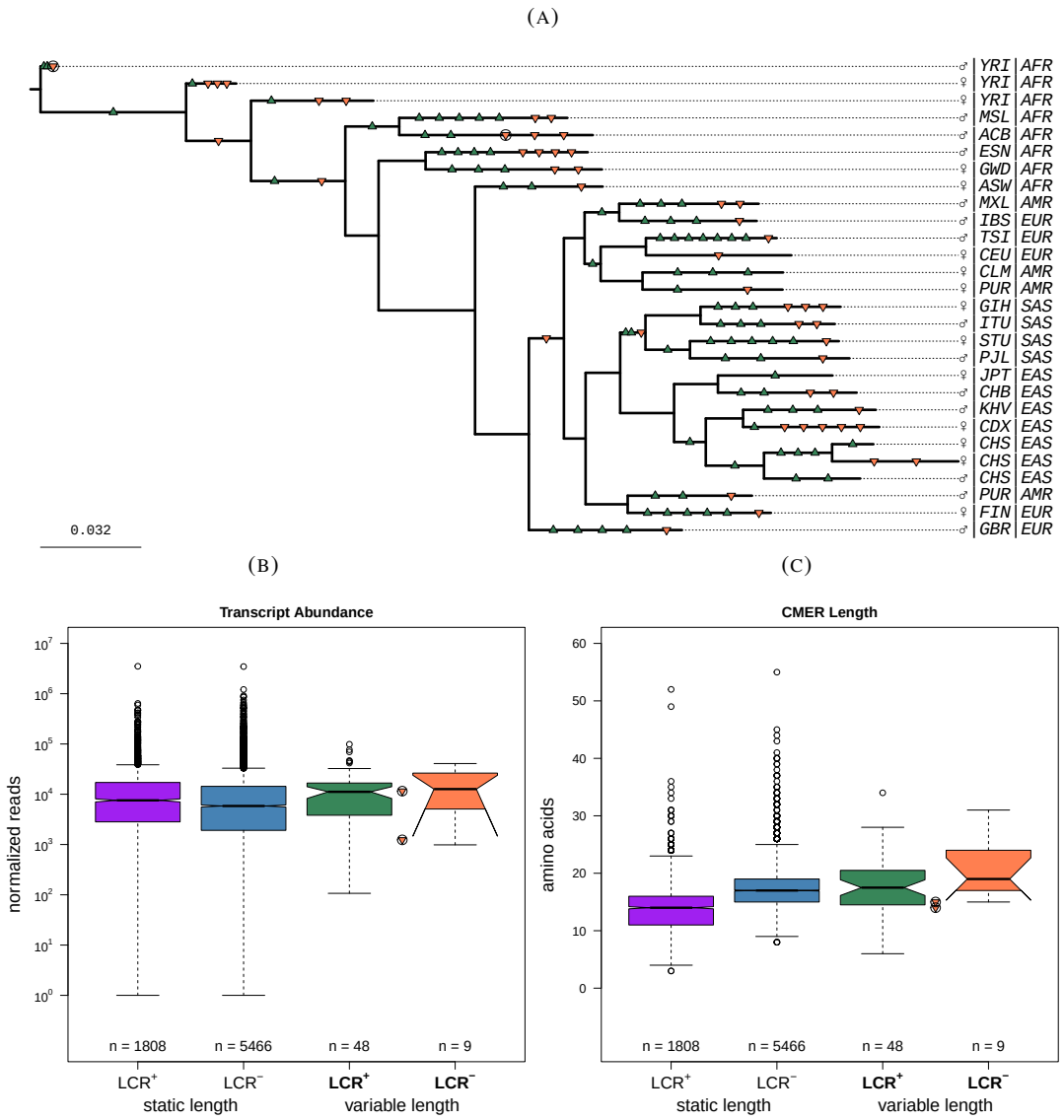

FIGURE 6.1: Properties of the proteins used in ABC modeling. a The SNP tree for chromosome 19 of 28 human individuals. All 132 indel events in CMERs are shown on the branch along which they are inferred to have occurred. Green, Upward-pointing triangles indicate the 84 insertions, while orange, downward-pointing triangles indicate the 48 deletions. Sex as well as population and super-population 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.

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 υ 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 Figures A3.36 to A3.41. 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 latter is outside the range (Table 6.2). 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.

The posterior distribution for -upsilon replicate 1 can be found in Figure 6.2. 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 strength 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 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 depends strongly on the value of τ.

TABLE 6.2: Summary of ABC modeling runs

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

The MCMC traces and a closer look at the univariate posteriors for each parameter can be found in Figures A3.19 and A3.20, respectively. Similar visualizations for every modeling run can be found in Appendix C.

6.5 Discussion

Constructing evolutionary models for TAb and CMER length based on the human proteins where we were able to definitively identify amino acid indel events 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-upsilon) 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 -upsilon model, the later achieved a higher likelihood with fewer free parameters. Fixing upsilon at zero assumes that CMER length evolves independently of TAb, and therefore the apparent better fit of the -upsilon 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 may be drawn.

First 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 a 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 $\text{Beta}(0.56 \pm 0.094, 0.59 \pm 0.102)$ distribution best describes the CMER position distribution. Values below one indicate bias towards the C- or N-terminal regions, respectively. Given that the unform distribution is a special case of the beta distribution with both parameters equal to one, the probability of the observed Beta fit given trully uniform data 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 Figure 6.1a, 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.

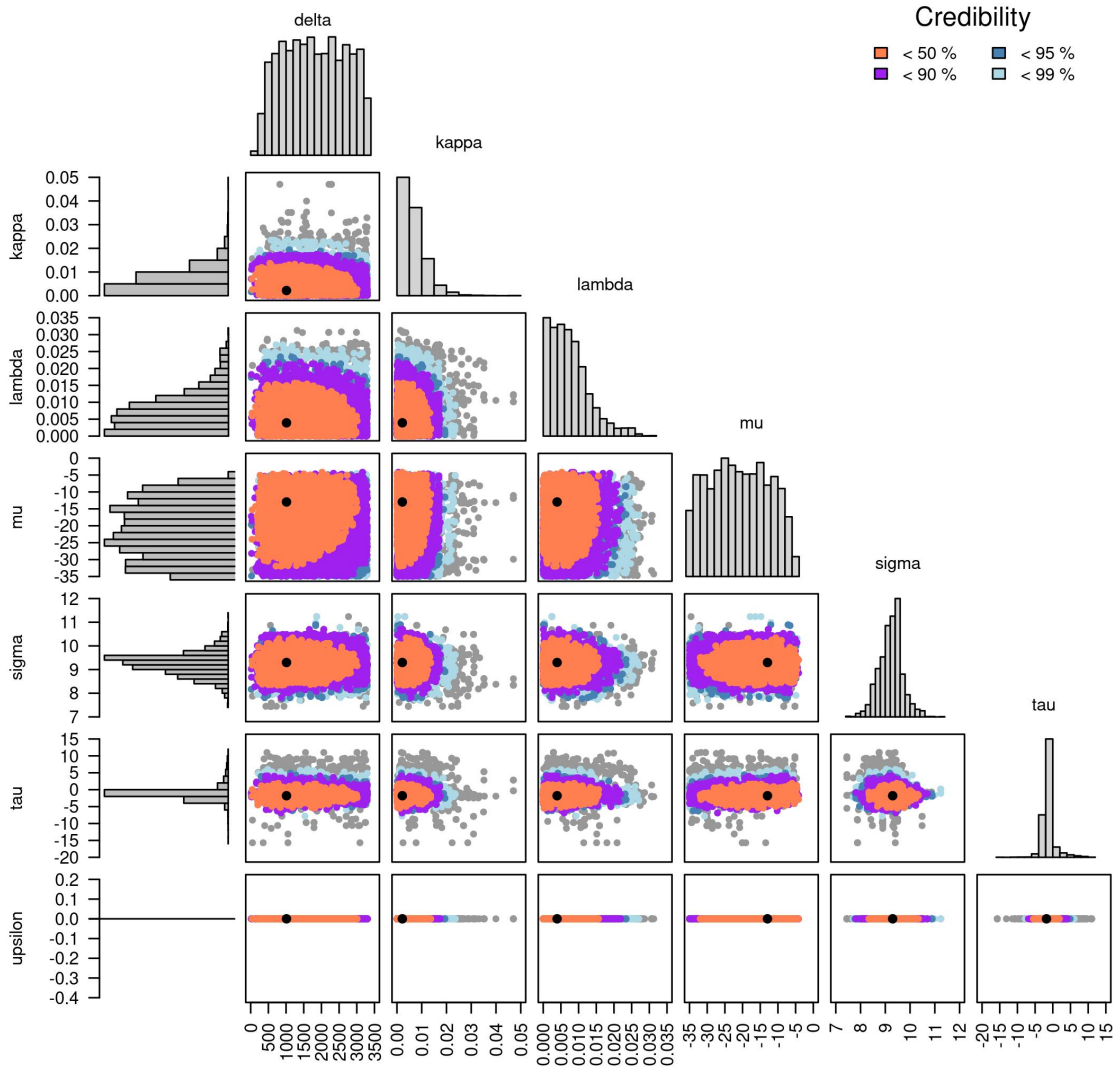

FIGURE 6.2: A visual summary of the posterior distribution estimate by ABC of the StepwiseOU-upsilon evolutionary model which assumes indel rates are independent of TAb . The central black point indicates the multivariate mode of parameter estimates, with colours indicating the credibility interval within which each posterior sample fell. Each scatter plot is a projection of the posterior distribution 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 (upsilon). The parameter mu is shown in natural orders of magnitude. Note that the parameter upsilon is fixed at zero in this model, which sets the indel rates as independent of TAb.

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 (Figure 6.1c). 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 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 an LCR to be tolerated, or the appearance of an 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 an 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.

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. Deeper time datasets are needed to identify 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 it is most likely 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.

Chapter 7

Discussion

The name of low complexity regions (LCRs) is potentially a misnomer. While defined by relatively simple primary sequences, most of their biological impacts are highly complicated. The answers to most questions around LCRs are rarely simple. Answering these questions requires careful analysis to process the complex mixture of signals and reveal the signatures of LCRs.

How do we identify LCRs? In Chapter 2, I demonstrated that selecting an all-encompassing definition of LCRs is difficult, with no one method fully sufficient. The choice of a low-entropy, information content based definition of LCRs naturally lends itself to a question on the flow of information necessary to the functioning of known biological systems. How do LCRs affect the flow of information from DNA to RNA to protein? Johanna Enright and I showed, in Chapter 3, that despite the constraints on information present in LCRs, it is a less than ideal predictor of the information present in the DNA. This implies that forces beyond neutral evolution are influencing LCRs. As it was known that these sequences are mutationally unstable we explored these forces by asking about the proteins which tolerate them. Are LCRs constrained to low abundance, side roles in the interplay of cellular machinery? My work in Chapter 4 demonstrated the importance of widening the field of view: despite low protein abundance (PAb) of LCR containing proteins, the messenger RNA encoding those proteins have higher transcript abundance (TAb). The modes of abundance regulation are widely varied with mechanisms at every step between gene transcription and eventual protein degradation. As the LCRs I examined were in coding sequences, a natural process to examine was translation. Is the influence of low-complexity (LC) restricted to coding regions of genes? In Chapter 5 my undergraduate volunteers and I demonstrated that compositional bias in $5'$ untranslated region (UTR) are associated with higher TAb, and that untranslated low-complexity motifs (ULCM) offer sites for regulatory control. Observing LC associated changes in PAb led to asking about evolutionary effects as the amount of protein affects how well its function is fulfilled. How do LCRs affect the evolution of expression and regulation? In Chapter 6, I showed that sequence evolution through indels and TAb are coupled together. Specifically, the former influences the latter more than the reverse; at least in proteins which have already tolerated the evolution of an LCR.

I have used the lens of abundance to expand the view of the importance of LCRs to the evolution of biological systems. While previously explored only incidentally, this view of the

evolutionary landscape makes clearer the complexity of the systems which drive biology. The wide variety in LCRs goes in hand with the variety of their impacts. The varied constraints along the path from gene to protein require both stringent control, and flexibility. High mutability serves as an engine of evolution not only for the development and refinement of new protein domains, but also in the regulation of abundance.

As most lines of inquiry do, we are ultimately left with more questions than when we began. The largest of these is that the mechanism which causes the positive association of TAb with LCRs is unknown. There are solid pointers for the negative association with PAb: lower translation efficiency and more rapid degradation of LCR^+ proteins as well as other similar effects for other processes. However, it is still unclear whether TAb elevation is an adaptive response to counter the protein level effects of LCR or LCRs are only tolerated in proteins with the appropriate regulatory framework. The process of answering this question will open opportunities to examine both the past and present.

Experimental validation of the effects we have observed would likely hint at a mechanistic solution. Specifically, targeted mutants could be generated where the size or presence of an LCR is varied, and the resulting impacts on abundance are observed. Adaptive TAb would imply that smaller or removed LCRs, when not deleterious, would lead to increased PAb without significantly altering TAb. Whereas little to no effect on abundance would imply an entrenched, robust regulatory framework. However, even this is limited by our observation that recent evolution is happening in the context of ancient changes.

Answering the temporal question has several challenges and overcoming them would advance our understanding significantly. We examined a short evolutionary timescale so that LCR would be 'easily' recognizable between individuals. However, it is clear that the appearance of LCR, and the TAb framework around them, was already established for these proteins, and therefore a deeper time approach is required. The question of LCR identification rears its head here as the imperfect solutions we use, combined with rapid LCR evolution, make identifying conserved, novel, and unique LCRs increasingly more difficult as one looks deeper in time. There is an opportunity for more robust methods of LCR identification which take into account evolution. These could not only accurately identify extant LCRs, but also distinguish between medium complexity regions and 'fossil' LCRs which have been fixed and degraded by substitutions. A reliable method for identifying proteins which are LCR^+ in some species but not others would allow us to finally examine how the presence or absence of an LCR is associated with abundance on an evolutionary timescale.

Beyond the direct effects of LCR in coding sequences there are many avenues of exploration for LC in signaling and indirect effects. The context of any signaling motif is known to modulate its efficacy (Reiter et al. 2023), and Horton et al. (2023) showed that direct interaction between short tandem repeats (a type of LCR) can be the mechanism for this. A similar mechanism may be at work in UTRs, where ULCM are modulating the effectiveness of other signals in the sequence. If this were the case, and it could be properly understood it offers a mechanism by which we might alter untranslated sequences to have more desirable properties in experimental setups or therapeutics like mRNA vaccines. In general, a better understanding

of the regulatory mechanisms which interact with LCRs to affect abundance may lead to treatments or prevention for LCR associated neurodegenerative diseases.

I believe my work has added a new lens through which we can examine LCRs and other sequence motifs, and established a foundation from which several new avenues of research can be built. In addition to abundance there are likely other important properties that LCRs are associated with. My work may serve as a template for integrating information together into an evolutionary context.

Part II

Targeted Metagenomics

Chapter 8

HUBDesign: Probe Design for Broad yet Targeted DNA Capture

8.1 Preface

This chapter was published in Cell Reports Methods in October of 2021 (https://doi.org/10.1016/j.crmeth.2021.100069, with Zachery W Dickson, Dirk Hackenberger, Melanie Kuch, Art Marzok, Arinjay Banerjee, Laura Rossi, Jennifer Ann Klowak, Alison Fox-Robichaud, Karen Mossmann, Matthew S Miller, Michael G Surette, G Brian Golding, and Hendrik Poinar as co-authors. See Section 8.7 for specifics on author contributions.

Additionally, an erratum was published in July of 2022 (https://doi.org/10.1016/j.crmeth.2022.100246). The main text was unaffected, but corrected versions of Figures 8.4 and 8.5 were made available. The corrected versions are the ones included in this chapter.

A wide variety of metagenomic research efforts are hampered by the same challenge: low concentrations of targets of interest combined with overwhelming amounts of background signal. While PCR or naïve DNA capture can be used when there are a small number of organisms of interest, design challenges become untenable for large numbers of targets. We present HUBDesign, a bioinformatic pipeline which designs probes for targeted DNA capture which leverages sequence homology to identify probes sets which maximize the breadth of coverage for targets while maintaining specificity. We validated HUBDesign by generating probe sets targeting the breadth of coronavirus diversity, as well as a suite of bacterial pathogens often underlying sepsis. In separate experiments demonstrating significant, simultaneous enrichment, we captured SARS-CoV-2 and HCoV-NL63 in a human RNA background and seven bacterial strains in human blood. HUBDesign (https://github.com/zacherydickson/HUBDesign) has broad applicability wherever there are multiple organisms of interest.

8.2 Introduction

Several critical monitoring, clinical, and research efforts are hampered by the same challenge: low concentrations of targets of interest combined with overwhelming amounts of background signal. Whether it be monitoring the reservoirs, disease ecology, and transmission of zoonotic infections, such as COVID-19 (Rodriguez-Morales et al. 2020; Boni et al. 2020); or attempting to determine which of a huge array of potential pathogens is present in a patient displaying sepsis, the combination of low signal in a high background presents a significant challenge. Attempts to overcome this have been stymied by the difficulty associated with detecting or culturing these microbes (Wade 2002; Papafragkou et al. 2014).

The advent of next generation sequencing and the ever declining cost of sequencing has made feasible a wide variety of research including transcriptomics, ancient genomics, and microbial metagenomics. It is now viable to use RNA or DNA sequencing to rapidly identify organisms and characterize the diversity of nucleic acids in heterogeneous samples (Wang et al. 2019). However, there remain limits to sequencing depth and cost. In some cases, rare and interesting microbes may remain undetected.

Rare taxa can be clouded by high backgrounds from host or environmental sources which often make up 99% of sequencing depth. In addition to the obscuring effects from sample backgrounds, differentiating true signals from contaminants becomes increasingly difficult as the organisms of interest often make up a small fraction of the sample. A naïve approach of simply sequencing deeper is an unbiased, yet costly way to overcome these issues. Pathogens in clinical or wildlife settings can easily make up less than 1 millionth of a sample, especially in early stages of infection where detection would be most useful for patients (Opota et al. 2015). Even with inexpensive sequencing costs, it becomes extremely wasteful and inefficient to spend sometimes critical time and resources to acquire and analyze these data when the majority is ultimately uninformative.

One way to alleviate the issue of cost is to bias detection towards targets of interest. Polymerase chain reaction (PCR) is one such technique used in many rapid detection systems (Tatti et al. 2011; Benirschke et al. 2019), including those used to detect individual sepsis pathogens and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of COVID-19 (Coronaviridae Study Group of the International Committee on Taxonomy of Viruses 2020). The technique relies on primers that bind to nucleic acid sequences specific to an organism or group of organisms. While capable of sensitive, rapid detection and quantification of a particular target, PCR is limited when multiple loci are targeted by primers. Identifying 'barcoding' regions has been used to amplify related organisms (Stahlberg et al. 2017; Adamowicz 2015), and multiplexed PCR can allow for the amplification of multiple disparate targets (Hayden et al. 2008). The former is only possible for closely related groups, and the latter can be prone to bias and interference between the various primers in use (Elnifro et al. 2000). Additionally, PCR is susceptible to failure in rapidly evolving organisms like viruses where mutations occurring at priming sites can prevent amplification, as seen in SARS-CoV-2 (Rahman et al. 2020).

Another important technique in this area is microarrays. Oligonucleotide probes are designed which specifically hybridize to sequences of interest. These probes are then immobilized such that each probe sequence is in a known position, and the entire array is exposed to a DNA sample. Target sequences will be retained, while the remainder are washed away. Fluorescently labeling the libraries allows captured targets to be visualized to determine the presence of key taxa within a sample (Brown and Botstein 1999). Such microbial detection arrays have demonstrated effectiveness (Gardner et al. 2010). However, they are limited to detection and identification of only known sequences, and there are challenges in efficiently designing probes to capture the targets of interest. A complementary solution is targeted enrichment. Oligonucleotide probes are designed to hybridize to target nucleic acids, however 'capture' is performed in solution and preferentially retains them over non-target sequences (Mertes et al. 2011). This leads to an enrichment of the target relative to the background and less effort and resources expended on sequencing and identifying uninformative molecules. A major advantage over PCR is the ability to design probes capturing multiple loci simultaneously, like those designed to capture ~2000 antimicrobial resistance genes (Guitor et al. 2019). Where identification and detection are important, capturing multiple independent loci in a genome provides more confidence of an organism's presence. Having multiple loci also assists in tracking variation between strains as they emerge and evolve. The simplest probe design for a single organism is to select probes with a window which slides along the genome. Typically, each subsequent window overlaps the previous one. The resulting overlapping probes tile the target and are likely more effective than non-overlapping probes (Bertone et al. 2006). This approach can be extended to multiple organisms; however, the number of probes increases rapidly as more genomes are targeted, and there is this method makes no effort to ensure the probes are specific to the organisms of interest. Each additional genome adds its length in probes increasing the chances for probe sequences to match multiple genomes. These matches are most often due to sequence homology between related organisms. As hybridization between probe and target is not perfectly specific (Mason et al. 2011), imperfect matches increase the chance of cross-reactivity and make most of the probes generated in this manner redundant.

Fortunately, sequence homology and variable hybridization are beneficial for the design of more efficient probes which are capable of specifically and simultaneously capturing targets from known and novel members of a group of organisms. While a probe will preferentially hybridize to its exact complement in a competitive environment, hybridization to sequences up to 20% divergent is possible (Mason et al. 2011; Delsuc et al. 2016). This has been used to design probes based on sequences from extant organisms facilitating the capture and enrichment of DNA from distantly related extinct taxa (Wagner et al. 2014; Enk et al. 2016; Delsuc et al. 2016). Increased success was obtained by designing probes based on ancestral reconstructions (Delsuc et al. 2016). Ancestral reconstruction infers past character states from the diversity of modern states (Joy et al. 2016), and in the context of probe design may be seen as constructing a sequence representing the diversity of a set of input sequences. The representation may also capture diversity that is not represented by existing nodes on a tree. More generally the concept of representative sequences will be used to design probes capable of capturing a broad set of sequences.

Genomes, especially those of bacteria and viruses, are mosaic in nature with different genes

TABLE 8.1: Statistics for various probe sets produced

Dataset	Method	Number of Probes	Nucleotide Coverage (%)	Depth of Coverage	Efficiency	Runtime (hr)	Peak Memory (GB)
Coronavirus	HUBDesign	13500	25.0%	4.72x	1.87	0.87	0.5
	CATCH strict	3846	20.0%	1.01x	1.13	1.7	6
	CATCH permissive	1474	22.6%	1.29x	4.23	0.73	4
	Naïve	21267	20%	5x	1	0.02	0.02
Sepsis	HUBDesign	26,870	2.09%	3.64x	29.3	6.1	7
	Naïve	2 million	2%	5x	1	3.5	7

and genomic regions offering unique evolutionary histories (Pedulla et al. 2003; Martin 1999). As a result, hierarchical trees constructed based on sequence similarity for entire organisms may differ from those constructed for individual genes (Goodman et al. 1979). Given a gene tree, a representative sequence can be constructed for each node in the tree by collapsing the sequences of all tips of the tree descended from that node. We have developed a pipeline that designs probes based on representative sequences at multiple hierarchical levels (e.g. family, genus, species). The resulting probes specifically target and enrich nested clades allowing for enrichment and identification of sequences from known and novel organisms.

Here we present and describe HUBDesign, a bioinformatic pipeline which leverages sequence homology and flexible DNA hybridization to design probes which can efficiently target sequences from a broad selection of organisms while maintaining specificity. To demonstrate the capabilities and effectiveness of HUBDesign we have designed and tested two probe sets. A coronavirus probe set capable of simultaneously detecting all sequenced coronaviruses, and a set of probes targeting bacterial pathogens associated with sepsis.

8.3 Results

8.3.1 Probe Design

Multiple methods of designing probes were performed, and information on each is detailed in Table 8.1. Given differences in breadth and depth of coverage a comparable metric of efficiency was calculated as the average number of distinct genomes any given probe maps to. For the relatively small coronavirus dataset, the runtime (52min) and effectiveness (1.87) of HUBDesign falls within the performance range CATCH given reasonable hybridization parameters, (44-102 min, and 1.13 - 4.23). However, HUBDesign is more memory efficient which allows it to scale to the much larger sepsis dataset, for which CATCH failed with all tested parameter sets. Both methods produce more compact and efficient probe sets than a naïve strategy.

The HUBDesign probe set for coronaviruses was tiled such that each taxon was targeted by approximately 400 probes. As seen in Figure 8.1 all genomes, where possible, are targeted by a minimum of 200 probes. The four lowest probe counts are for two gammacoronaviruses (Turkey Coronavirus: txid11152; 23 probes, and Infectious Bronchitis virus: txid11120; 8) and two alphacoronaviruses (BtRf-AlphaCoV/YN2012: txid1503293, and Rhinolophus bat coronavirus HKU2: txid693998) with zero probes.

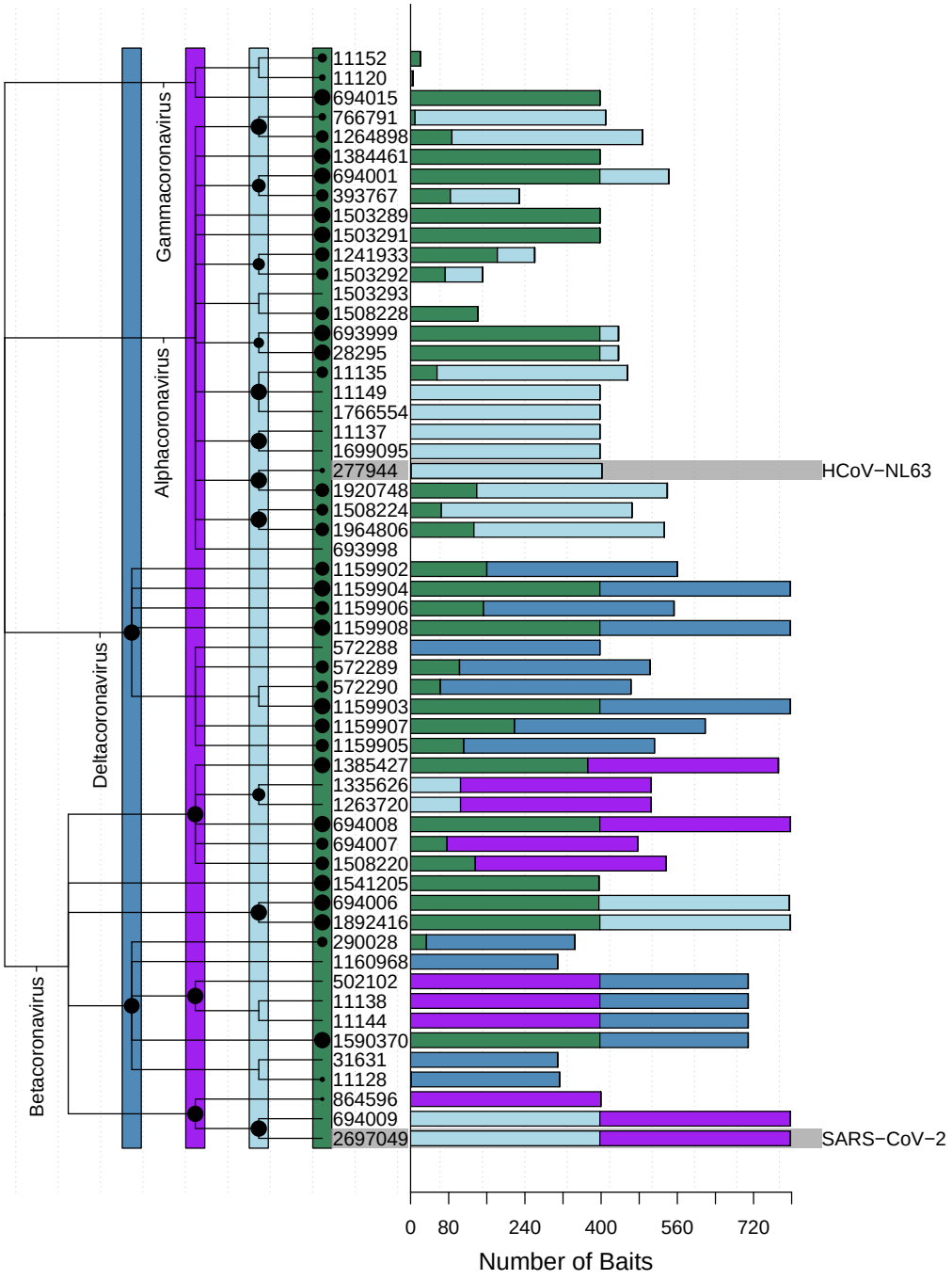

FIGURE 8.1: Tip labels indicate taxon ids for the viruses, and the size of each node and the width of the matching bar give the number of probes targeting that genome. The hierarchical level of the probes is colour coded according to the node height in the dendrogram. The two viruses used are highlighted with grey bars.

The majority (62.5%) of the probes have targets which are specific to one virus. Of the probes targeting multiple viruses, most (78.1%) target two or three. The remaining three sets of probes target loci specific to merbecoviruses and embecoviruses (both are *Betacoronavirus* subgenera) and loci common to the *Deltacoronavirus* genus. Both SARS-CoV-2 and Human coronavirus NL63 (HCoV-NL63) have probes at two levels in the hierarchy. For SARS-CoV-2 there are nearly 400 probes which target sequences common to SARS-CoV-2 and severe acute respiratory syndrome coronavirus 1 and an additional 400 probes which target *Sarbecovirus* sequences in general. While there are no probes which target SARS-CoV-2 specific loci, the virus is easily differentiated by its sequence at those bait positions. HCoV-NL63 has 400 probes targeting *Setracovirus* sequences and an additional 4 probes which specifically target HCoV-NL63 loci.

The HUBDesign probe set for sepsis pathogens contained 26,870 probes targeting bacterial pathogens covering 2.09% of all nucleotides in the input dataset at an average depth of coverage of 3.64x. A naïve tiling achieving 2% coverage at 5x would require over 2 million probes. All 1926 bacterial strains are targeted by probes which are at least at the genus level, and 53.3% of strains are targeted at the species level. The only genus which did not have any probes was *Clostridium*, but all strains in the genus were targeted at the species level. These species *C. botulinum*, *C. perfringens*, and *C. tetani*, also had the lowest probe counts at 12, 44, and 71, respectively. The next lowest were *Rickettsia prowazekii* and *Borrellia burgdorferi* at 53 and 90 probes, respectively. All other species have at least 100 probes, with an overall median of 478 probes per species. The seven spiked strains are targeted by at least 110 probes (*S. sanguinis*) and up to 564 probes (*Burkholderia multivorans*). *S. sanguinis* is the only spiked strain targeted at only the genus level, as it was not included in the dataset used to design the probes. Details on the number of probes per genus and species are in Table A4.6.

8.3.2 Coronavirus Probe Validation

Figure A4.1 shows how the amplicon levels compare to the rest of the coronavirus genomes. While there is a peak in the coverage in this region, it is within the variability of nearby genomic regions. The estimated copy masses of HCoV-NL63 and SARS-CoV-2 are 29.5 attograms/copy and 10.8 attograms/copy, respectively. Note that the copy mass of HCoV-NL63 was nearly 3x higher and therefore the nominal ratios based on copy number will not be represented in the sequencing results. For example, the EH sample was prepared with an amount of viral extract expected to result in 20000 copies of each virus. However, based on the shotgun baseline, the actual amounts of viral RNA are 215 and 589 femtograms of SARS-CoV-2 and HCoV-NL63, respectively. This gives a ratio closer to 1:3 rather than the original PCR estimated ratio of 1:1.

The proportion of reads assigned to SARS-CoV-2, HCoV-NL63, human, or otherwise can be seen for each sample in Figure 8.2. The proportion of viral reads is significantly and highly enriched relative to a shotgun sample with the same amount of viral RNA. The combined number of reads assigned to either of the two spiked viruses was considered to be the number of on-target reads. This and the number of off-target reads were used to perform logistic regression. We observed fold enrichment in on target reads of 97.6x (95%CI: 96.4-98.9x). The

summary of the logistic regression can be found in Table A4.8. To examine the performance of individual probes, the two genomes were divided into alternating regions with and without probes. For example, the SARS-CoV-2 genome was broken into 20 regions, the first of which covers the first 4,950 bp in the genome and was targeted by 166 probes. It is followed by a 956 bp region not targeted by any probes, which is in turn followed by a 773 bp region with one probe at its center. Region boundaries were defined as 350 bp upstream and downstream of overlapping probes. This allows the analysis to account for the extended field of influence of probes resulting from capturing fragments with significant overhang. The following values were calculated for each region: the GC content, the number of probes, the average level of divergence from the genome sequence across the probes, and the fold enrichment. Differences in library size were accounted for by adjusting the enriched library's read counts by the relative size of the paired shotgun library. The fold enrichment for each region was calculated by dividing the observed read count in the enriched sample by the adjusted read count in the paired shotgun sample.

Due to the low sample concentration of viral RNA, there were several regions with no read coverage in the shotgun samples. The shotgun baseline was used to adjust shotgun read counts to reduce zeros. For each region, the proportion of reads from the shotgun baseline in that region was calculated. The adjusted read count for each region was the proportion in the shotgun baseline multiplied by an estimate of total reads across the genome taken from a weighted average across all regions.

Based on linear regression, increases in GC content from the genomic mean are negatively associated with fold enrichment. An increase in GC content of 11.3% is associated with a halving of fold enrichment, however this effect is not significant in genomic regions targeted by probes. This amelioration of negative relationships when probes are present holds across all predictors. Another linear regression was also done which only used regions which had probes. Surprisingly, GC content, probe divergence, and shotgun baseline levels had little if any significant effect. However, probe density was significantly and positively associated with fold enrichment. Every additional 77probes/kb resulted in a doubling of fold enrichment. Both regressions show a positive relationship with viral load over the range of viral loads tested. The relationship was weaker for HCoV-NL63 which had more RNA per nominal copy, consistent with expected diminishing returns in fold enrichment at high viral loads. A complete summary of these regressions can be found in Table A4.9. GC content and baseline shotgun levels only have a significant effect in probe-free regions. It is likely that these factors increase the number of sequenced reads overall rather than affecting the enrichment specifically.

Figure 8.3 shows the fold enrichment, probe coverage, and GC content at each position in both genomes. There is no discernible relationship between GC content and fold enrichment, especially given the correlation between probe coverage and GC content. HUBDesign's selection of probes is based on finding unique sequences, and biased nucleotide content makes unique sequences less likely. Coronaviruses have an average of 30% GC content overall, but there are genomic regions with GC content at parity with AT content. As unique sequences are more likely to be found in these regions, it explains the correlation between increasing GC content and higher numbers of probes.

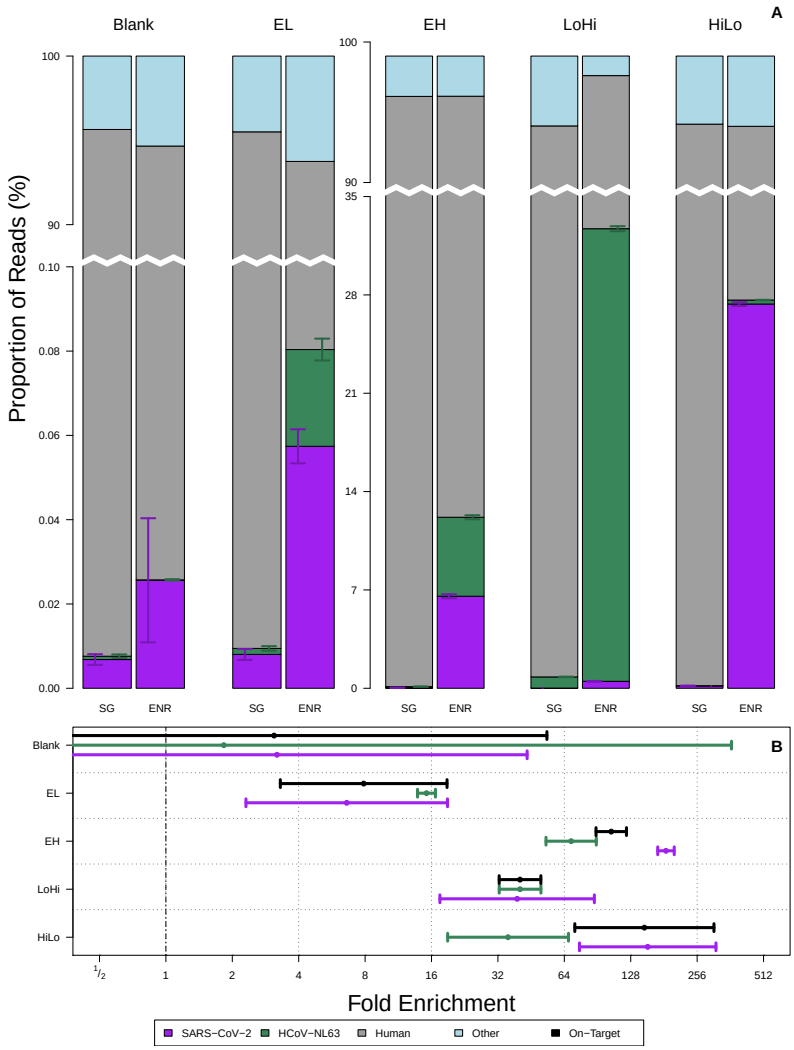

FIGURE 8.2: (A) The proportion of reads assigned to the two spiked viruses and the human background. Reads assigned to any other taxa are grouped together. The right column in each pair represents the sample enriched with the probes. Error bars are the 95% confidence interval on the proportion. The Blank and **E**qual **L**ow samples are on a different scale to show that enrichment is observable even at the lowest tested viral concentrations. (B) The Fold Enrichment of on-target viral RNA in each sample. While the proportion of SARS-CoV-2 in the blank is significantly greater in the enriched sample, the calculated fold enrichment is insignificant. Also see Table A4.8.

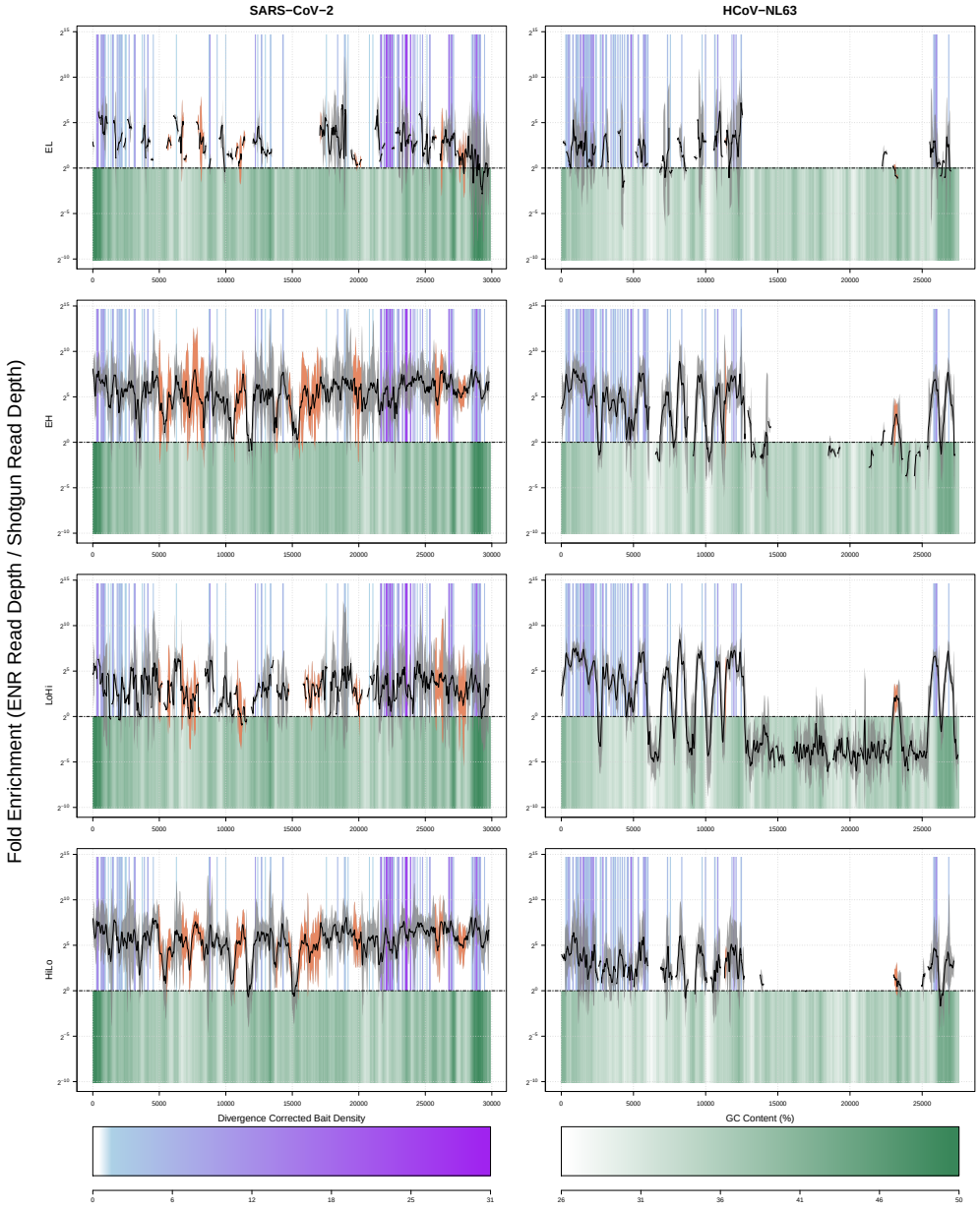

FIGURE 8.3: Each row is a single set of samples in the order, EL, EH, LoHi, and HiLo. The grey area indicates 3 standard deviations around the mean. The number of probes covering a particular position is indicated by the blue-purple intensity This number of probes is adjusted by the divergence, such that more divergent probes contribute to the density less. GC content is indicated by the green intensity. Breaks in the line indicate that there were no reads covering that position in enrichment samples. Regions with apparent enrichment without being targeted by a probe are highlighted with orange confidence intervals; See also Table A4.7. Large Gaps, or low fold enrichment regions for HCoV-NL63 are strongly correlated with regions with no probe coverage. Probes cover a larger proportion of the SARS-CoV-2 Genome, and there are fewer gaps or low-fold enrichment regions of the genome. Also see Table A4.9 and Figure A4.2.

8.3.3 Sepsis Probe Assessment

The proportion of reads assigned to each of the spiked strains, with all *Streptococcus spp.* grouped together, as well the proportion of human or other organisms can be seen in Figure 8.4. Enrichment of the spiked taxa, but not the human background, is observed in the blood blanks. We detected genomic sequences from every spiked strain in the blank shotgun samples, and these contaminant sequences were captured by the probes intended to do so. Enrichment of sequences targeted by probes, but unintentionally present in the samples is also apparent when examining the 'Other' category. The majority of these reads are assigned to probes targeting *Shigella* (69%) and *Escherichia* (30%) sequences. Adjusting for library size, there were 466x more reads for these two genera in the water blank samples than the blood blank samples. Fold enrichments estimated with logistic regression were 11.8x (95% CI: 8.87-15.7x) in the Low sample, 64.3x (95% CI: 40.1-103x) in the Medium sample, and 18.6x (95% CI: 12.4-27.9x) in the High sample.

To assess the difference in performance between probes targeting at the genus and species levels, all reads were remapped competitively to the genomes of the spiked bacterial strains with BLASTn being used to disambiguate reads which mapped to multiple positions within a genome or reads which mapped to multiple genomes. The genomes were broken up into regions targeted by probes at each taxonomic level, and one large region composed of all untargeted genomic regions. Within each region the log ratio of enriched reads to shotgun reads was calculated. The difference in library depth was accounted for by adjusting read counts in the larger library down by the ratio in size between the two libraries. Linear regression was used to account for properties of the baits and assess the difference between species level and genus level probes. This difference was only significant for *Staphylococcus aureus*, which also had the greatest disparity between the number of regions targeted at the genus and species level (Table A4.10). In all cases the variation due to properties of the probes, especially probe density and probe divergence, was larger than the variation due to taxonomic level. The performance in the probe regions across the spiked strains can be seen in Figure 8.5.

8.4 Discussion

The HUBDesign pipeline was able to rapidly design a compact and efficient probe-set covering almost every one of the targeted coronaviruses. Overall design time was less than a day, the majority of which was spent exhaustively filtering candidate probes against the human genome. The collapsing of the genomes into representative sequences required less than an hour, and the identification of candidates was completed in under a minute and required less than one GB of memory. However, processing 56 viral genomes is a minor task compared to the capabilities of the pipeline. Memory requirements for candidate identification scale linearly with genome size and number of organisms, but time requirements grow much more quickly. The pipeline has also been tested on three other input sets. Table A4.5 details the performance of SA_BOND on each set of organisms. These datasets come from different stages in the development of the HUBDesign pipeline, but the SA_BOND step has remained constant, and is also the most memory intensive step. It should be noted that the amount of diversity within a set of organisms

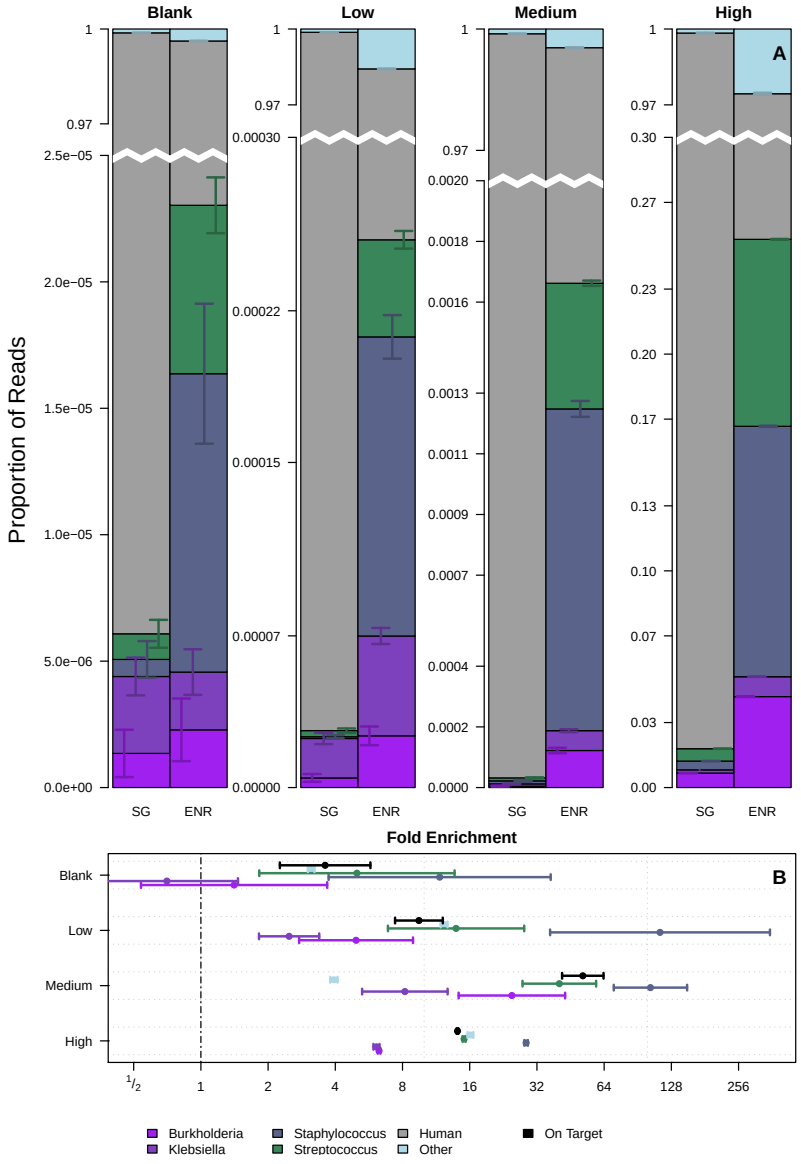

FIGURE 8.4: (A) The proportion of reads assigned to four different spiked genera and the human background. Reads assigned to other taxa are grouped together. The right column in each pair represents the sample enriched with the probes. Error bars are the 95% confidence interval of the proportion. Each sample is on a different scale to show the enrichment of the targeted genera. (B) The fold enrichment of non-human sequences in each sample. The error bars indicate 95% confidence intervals of difference between log read counts in the enriched and matching shotgun libraries. The only insignificant enrichment among the spiked genera is for *Burkholderia* and *Klebsiella* in the Blanks. The blanks in both panels refer to libraries prepared from blood only without any spiked bacteria.

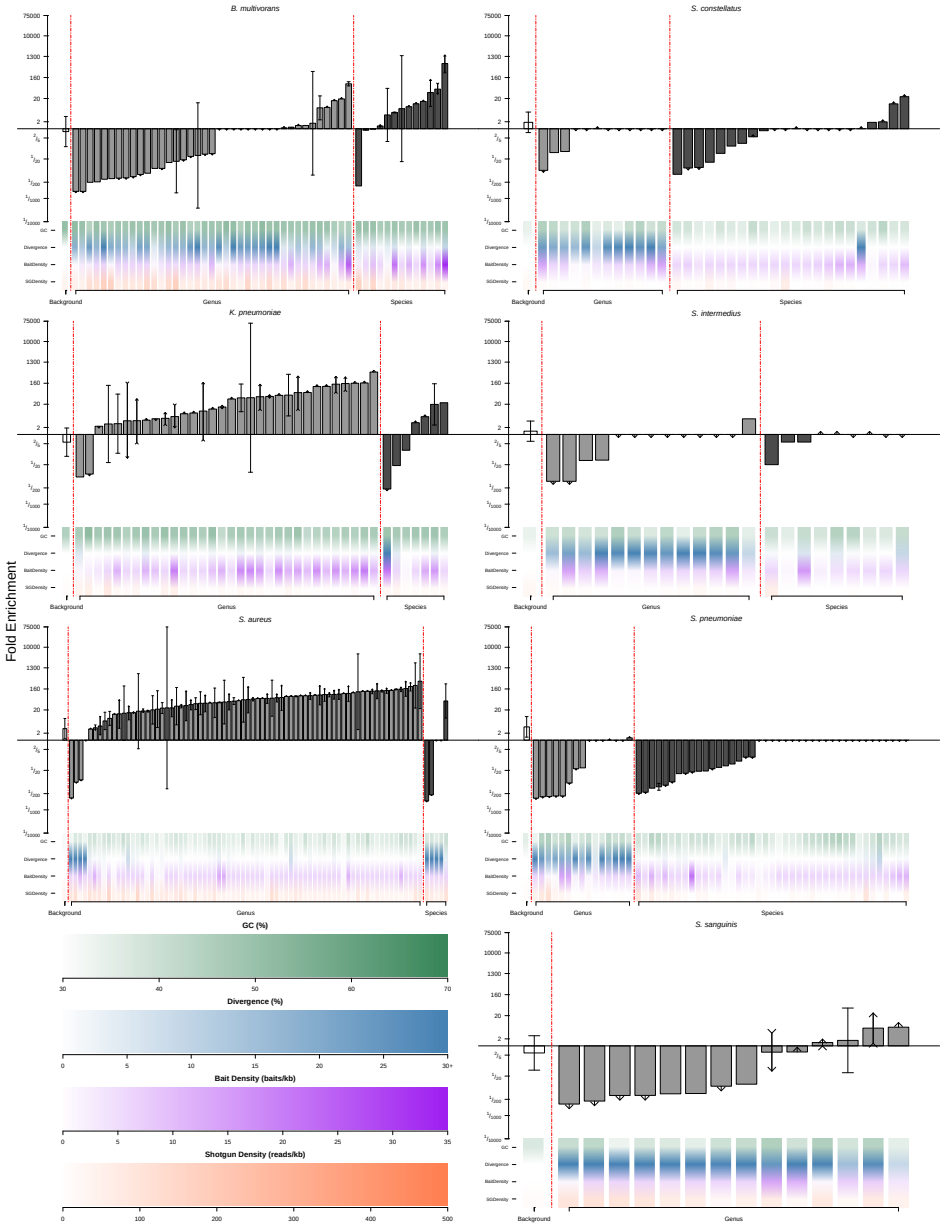

FIGURE 8.5: Bar heights indicate the fold enrichment observed for each genomic region. Bars are in ascending order of fold enrichment. Error bars represent the 95% confidence level of the difference between the log read count between enriched and shotgun samples. The arrows on the error bars are an attempt to incorporate incomplete information. Upward pointing arrows indicate that in some replicates reads were present in the enriched sample and not the shotgun sample, and therefore the finite values observed in the other replicates indicate the minimum fold enrichment. Downward facing arrows conversely indicate that the finite observed values are the maximum fold enrichment. Arrows in both directions indicated both scenarios were observed, and there is very likely no enrichment or depletion. Colour tracks below each bar plot indicate the bait or genomic properties likely to affect fold enrichment, note that bait divergence ranges from 0% to 30% or more. While there are differences between the performance of genus and species level baits, the direction of this is inconsistent as the properties of the baits, especially bait divergence and bait density have a larger effect than bait level. Also see Table A4.10 and Figure A4.3.

is very important. For example, a dataset composed of 1473 common gut bacteria was tested. While there are fewer genomes than the sepsis dataset, those genomes are spread over 161 genera compared to 35 for sepsis. As a result, there is less overall sharing of sequences, allowing for less information to be collapsed into representative sequences, and the corresponding runtime of SA_BOND is longer.

There were two viruses for which no candidate probes were found at any taxonomic level: txid1503293 and txid693998. In the tree used during assignment of gene clusters, these two viruses are labelled as alphacoronaviruses, without any indication of being more closely related. (Figure 8.1). However, all but one of their genes consistently formed clusters which were distinct from all other viruses. Given the tree, the apparent lowest common ancestor (LCA) node for these two is the *Alphacoronavirus* node. As the two viruses represent only 8.7% of the strains in the genus, their clusters did not meet the penetrance threshold, thus their sequences were not included in the pseudo-genome. Without any sequence to select from, no probes were found for these viruses. We have recognized this and the current version of HUBDesign constructs a new tree to be used during cluster assignment, which is based on the observed clusters, and optionally guided by a user provided tree to resolve ambiguities.

Eleven mutations across SARS-CoV-2 isolates have been identified which can be used to classify the virus into 5 clades (Guan et al. 2020). Our probes, which target SARS-CoV-2, cover these mutations well. Four of the loci are in positions directly covered by the probes, five more have a probe within 100 bp, and the remaining two are 267 and 546 bp from the nearest probe. The validation results demonstrate that enrichment of genomic regions adjacent to the probes occurs at least as far as 350 bp if not further (Figure 8.3). For these loci, all positions demonstrated an average of at least 3-fold enrichment across all samples. With the two loci furthest from a probe averaging 17-fold and 19-fold enrichment, respectively.

During analysis, we removed reads which mapped to amplicons generated during qPCR of the viruses. This was performed to reduce potential contamination from any amplicons that had escaped into the environment, which can occur relatively easily throughout the procedures. (Rys and Persing 1993). Analysis of the reads filtered out suggested that our approach was quite conservative, as out of the blank samples (a total of 6 million reads), only one read mapped to an amplicon. The number of reads mapping to the amplicons correlates well (cor = 0.997) with the nominal copy number of SARS-CoV-2 in the sample with one read mapping to an amplicon for every 10.45 nominal copies of the virus present ($p < 0.0001$). Despite conservatively removing these reads which were mostly true viral reads, high levels of enrichment were observed.

In preparing each sample's read data, only one read was carried forward for each unique sequence observed. All other copies of that sequence are string duplicates which could be true biological reads. When multiple copies of the genome are randomly fragmented, identical sequences can be produced. However, given the experimental set up, it is much more likely that these reads are the result of PCR duplication. To evaluate the effect of duplication, all analysis were performed again without the deduplication step. The mean±SEM duplication rates in human and viral reads was observed at $8.2 \pm 0.7\%$ and $15.8 \pm 5.5\%$, respectively in shotgun samples. In enriched samples duplication rates were $20.3 \pm 3.5\%$ for human reads and

$54.2 \pm 5.6\%$ for viral reads. Fold enrichment of viral reads was approximately doubled when reanalyzing without duplication, consistent with the approximately doubled rates of duplication in viral reads. The duplication rates for both human and viral reads are elevated in the enriched samples, consistent with going through additional rounds of PCR. As the viral genomes are much shorter than the human genome, it is much more likely for identical fragments to arise by chance. This may explain the elevated duplication rates relative to humans. Deduplication would then be reducing the true read count on viral reads, but failure to deduplicate artificially inflates read counts in the enrichment relative to the shotgun. Despite our conservative approach by removing duplicates, we still observed significant enrichment of both SARS-CoV-2 and HCoV-NL63.

While the specifics of individual probe performance vary depending on the resolution and method of analysis, the clear enrichment of sequences from both viruses is robust and apparent with every analysis we performed. The target sequences of probes at all relevant hierarchical levels were significantly enriched. Fold enrichment levels are highest in the EH pool and lowest, but still significant, in the EL pool, the sample with the lowest viral input. The relationship does not appear to be linear. Fold enrichment is about double for SARS-CoV-2 than in HCoV-NL63, however there were also nearly twice as many probes targeting the former (796 probes) than the latter (402 probes). While the number of probes targeting a single locus was not observed to have a significant effect, the global number of probes targeting an organism does have an effect. This emphasizes the importance of balancing probe numbers across organisms. The fact that there were nearly twice as many probes targeting loci in SARS-CoV-2, and that these loci are more evenly spread across the genome contributes to the apparent difference in enrichment profile observed in Figure 8.3. Enrichment across the SARS-CoV-2 genome is relatively even, potentially because of the closely spaced probes.

There is evidence of enrichment in regions that are not targeted by probes for SARS-CoV-2 or HCoV-NL63. A notable example is the peak visible near position 23k of HCoV-NL63 in Figure 8.3. Three potential explanations for this are an extended field of influence for nearby probes, off-target capture between viruses, and within genome off-target capture. We identified 14 regions that were significantly enriched which were also at least 350 bp away from the nearest probe. Two were in HCoV-NL63, and the remaining in SARS-CoV-2. These regions are highlighted in Figure 8.3 and are detailed in Table A4.7.

We identified 45,791 unique molecules in these regions across all enriched samples, making up 7.1% of all viral molecules. The large majority (92.4%) of these molecules best match SARS-CoV-2, with nearly half of those mapping in the vicinity of on-target probes, consistent with the range of enrichment around a probe being larger than 350 bp. This can be explained by overhanging fragments. For example, if a 75 bp probe perfectly matches the end of a 300 bp fragment, this leaves 225 bp of the fragment to potentially capture overlapping fragments from the opposite strand. This allows the probe to enrich beyond its immediate target. If this were occurring, we would expect to see bias in the strandedness of our captured viral sequence data. The probes are designed to capture negative strand cDNA synthesized by reverse transcribing the positive strand RNA of the virus. Most of the captured fragments should then be from the negative strand, but fragments pulled down indirectly should be positive stranded. As can be

seen in Figure A4.2, we do indeed see that most reads are negative stranded, but in regions flanking enriched areas the strand bias flips. These indirectly enriched areas in some cases explain regions where enrichment is occurring without targeted probes.

All off-target-enriched reads were remapped to probe regions excluding probes which capture HCoV-NL63 and SARS-CoV-2. Only 763 (1.7% of all reads in noted regions) of these molecules successfully mapped, but they always did so to a probe in the correct genera (HCoV-NL63 reads mapping to other alphacoronavirus probes, and SARS-CoV-2 to other betacoronavirus probes). When the remappings are broken down by their region in the genome, only 3 of the 14 regions appear to be enriched by off-target probes. The region from positions 23k to 23.3k in the HCoV-NL63 genome appears to have been captured by a group of 27 probes targeting the same region of the Camel alphacoronavirus (txid1699095), which is about 25% divergent from HCoV-NL63. Two regions of the SARS-CoV-2 genome covering positions 15.4k-17.2k appear to have been captured by three sets of probes all targeting similar genomic regions. The first set of 118 probes targets Betacoronavirus HKU24 (txid1590370), the second set of 36 probes targets Rat coronavirus Parker (txid502102), and the third set is a single probe targeting Rousettus bat coronavirus HKU9 (txid694006). Due to the overlapping nature of probes, it is difficult to say which specific probe is responsible, except in the case of the txid694006 probe which had 100 reads attributable to it. In general, the fold enrichment in these off-target regions is lower than in on-target regions, and the off-target effects would be expected to diminish if the true target of the probes were present. When not present these probes can still be an advantage when the goal is hunting for novel organisms or identifying and monitoring members of a community.

The third possibility for apparent enrichment at a distance from a probe is off-target capture by probes within the same genome. The reads which map to the genome in probe-free enrichment regions were mapped directly to probes, and the positions targeted by those probes was compared to the original genomic position of the read. Of the 15,669 (34.2%) reads which mapped to a probe almost all (97.9%) mapped to probe directly adjacent to a probe-free enrichment region. This again indicates an extended field of influence of the probes. However, there were 330 reads which mapped to a probe targeting a position at least 1000 bp away from read's genomic position, with the furthest being nearly 27000 bp away. These 330 reads fall exclusively into 13 of the 14 probe-free enrichment regions identified. The only apparent probe-free enrichment not at least partly explained by within genome off-target effects is the region from 23k to 23.3k in HCoV-NL63.

A final concern for off-target enrichment is inadvertent capture of the background. Of the nearly 29 million reads which mapped to the human genome only 57 (0.0002%) also mapped to a probe or the 350 bp region immediately up- or downstream of the probe. Broken down by sample, these reads are found more in the HiLo and LoHi samples (46 reads) than in the lower viral load samples (11 reads), and none were found in the negative samples. This is the opposite of what would be expected if the probes were enriching the human background. The amount of human input is at least 10,000 times higher than viral input in the samples with the highest concentration (and over 10 million times higher in the lowest). Therefore, the number of off-target reads would be expected to be either constant or decrease with higher viral load as

competition between the probes true target and a partial human match would favour capture of the virus. As there is an increase with viral load, it is more likely that our human filtration step over-zealously filtered out reads from the spiked viruses. Human background enrichment does not appear to be occurring for the sepsis probes either, as we observed 0.00055% of human reads also mapping to a probe. Most (67%) are in the blood blank samples, followed by the blood-free positive controls (22%). While the same counter pattern as seen for the coronavirus probes was not observed, given the overwhelming amount of human DNA present the low number of reads at worst indicates extremely inefficient off-target enrichment. We also observed enrichment in untargeted genomic regions of the spiked bacterial species. These samples were prepared with a double stranded library preparation protocol and thus the same strand bias as seen for the coronaviruses, would not be expected. Instead, we calculated the minimum distance to the nearest on-target probe for each read, and calculated the fold enrichment of reads at each distance. While there was clear enrichment far from probe regions there was almost none observed near to, but outside of targeted regions (Figure A4.3). This indicates that daisy-chain-enrichment was not a significant factor for these enrichments. This may be due to differences in library prep and the strandedness of the input nucleic acids. Another important factor is that the coronavirus genomes are orders of magnitude smaller, and the pool of fragments from which to sample hybridizations during enrichment is also less diverse. This makes it far more likely for a complementary fragment to be pulled down during enrichment.

The majority of observed off-target enrichment is the result of probes meant for other taxa targeting the spiked strains. There were 215 probes (0.8% of probes) which mapped to the spiked genomes but had different nominal targets. Of these, 178 nominally targeted another species in the correct genus. There were 31 probes which targeted at various levels within the non-*Klebsiella* members of the *Enterobacteriaceae* family. Most notable were probes targeting *Enterobacter aerogenes*, however since these probes were designed, this bacterium has since been classified as a member of the *Klebsiella* genus. The remaining 6 probes are 5 *S. intermedius* probes mis-targeting *Klebsiella pneumoniae*, and one *Klebsiella* probe miss-targeting *S. sanguinis*. The *Enterobacteriaceae* probes highlight a flaw in the design for the sepsis probe set: the partitioning of clusters primarily based on the nominal taxonomy, rather than observed sequence similarity. The sequences targeted by these probes are shared at the family level, but the design considered the genera independently and selected candidate probes which were not truly specific to their nominal targets. These flaws present in the older version of HUBDesign have already been corrected in newer versions of HUBDesign, as we continue to develop and improve it. The same cluster partitioning issue is likely the reason for some highly divergent probes being included in the probe set, especially for the *Streptococcus spp.*. Despite these points to improve, the probes were able to enrich *S. sanguinis*, a strain 'unknown' to HUBDesign, with some genomic regions enriched over 100x (Figure 8.5).

While not among the bacteria intentionally spiked into the samples, we also observed significant enrichment (2-16x) of reads mapping to *Shigella* and *Escherichia* probes. The latter is a common reagent contaminant (Salter et al. 2014), and these two bacteria are closely related. As a result of the cluster partitioning described above, the more numerous *Shigella* probes Table A4.6 are also likely capturing contaminating *Escherichia* sequences. Both are

also common human pathogens included in the design dataset for the sepsis probes. To avoid the capture of reagent contaminants, backlist databases of common contaminant sequences could be provided to HUBDesign during the filtration phase.

Overall, the fold enrichment observed in total reads for target organisms ranged between 10 and 100x for both probe sets. With individual regions of the viral genomes having mean fold enrichment up to 1000x for the coronavirus probes and up to 20000x for some *S. aureus* probe regions. This is comparable to the reported performance of CATCH (Metsky et al. 2019). Using a set of ~350 thousand probes targeting 356 viral species on 30 patient samples with known viral infections. The median fold enrichment at genomic positions ranged from 1x to 53x. They observed fold changes for the number of reads for a virus within a sample as high as 1000x. The enrichment achieved by HUBDesign's probes can be translated in savings to sequencing costs. To attain the same depth of coverage we observed in our enrichments with a shotgun library, one would need to sequence 10 to 100x more deeply, with a commensurate increase in sequencing costs!

8.5 Limitations of the Study

The validation experiments used artificial samples generated by pooling the desired background with genomic extracts from the targets of interest. Nucleic acids in patient samples and environmental extracts may have damage or modifications which reduce their availability for capture, reducing efficacy without altering specificity. As the performance of probes is dependent on the sequence properties of the targets, the level of enrichment will vary for each produced probeset.

HUBDesign relies on annotated genomes to efficiently cluster sequences. Probes cannot be designed for targets with unknown, or unannotatable genomes. HUBDesign may be able to capture these sequences by targeting closely related taxa. Development on the pipeline is ongoing to improve efficiency and reduce barriers to use.

8.6 Acknowledgements

- This work was partially supported by CIHR COVID-19 Rapid Response funding to MSM.
- MSM was supported in part by a CIHR New Investigator Award and an Early Career Researcher Award from the Government of Ontario.
- AB, HP, GBG, and ZWD were funded through Natural Sciences and Engineering Research Council of Canada.
- This work was partially funded by a CIHR COVID grant to Principal applicant KM and co-applicant AB.
- HP was generously funded by the Boris Family Fund, the Micheal G. DeGroote Institute for Infectious Disease Research, and the CIFAR Humans and the Microbiome Program

8.7 Author Contributions

Conceptualization, A.F.-R., G.B.G., H.P., M.G.S, and Z.D.; Methodology, H.P., D.H., M.K., H.P., and Z.W.D.; Software, Z.W.D.; Formal Analysis, Z.W.D.; Investigation, D.H. and M.K.; Resources A.B, A.M, L.R., K.M, M.S.M, and M.G.S; Data Curation, J.A.K. and A.F.-R.; Writing - Original Draft, Z.W.D., D.H, A.B, A.M; Writing - Review and Editing, All authors; Visualization, Z.W.D.; Supervision, A.F.-R., G.B.G, H.P., K.M., M.S.M, and M.G.S; Project Administration, H.P.; Funding Acquisition, G.B.G, H.P., M.S.M, and K.M.

8.8 Declaration of Interests

The authors declare no competing interests.

8.9 STAR Methods text

8.9.1 Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Zachery W Dickson (dicksoz@mcmaster.ca).

Materials Availability

- This study did not generate new unique reagents.
- Generated Probe Sequences and associated metadata can be found at https://github.com/zacherydickson/HUBDesign/probes.

Data and Code Availability

- All sequencing data generated in the course of this work is available on the Sequence Read Archive under the BioProject accession PRJNA674643.
- The source code for HUBDesign is available under the terms of the GPL-3.0 license at https://github.com/zacherydickson/HUBDesign. DOI: 10.5281/zenodo.5156877 .
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request

8.9.2 Experimental Model and Subject Details

Cell lines

VeroE6 cells were cultured in Dulbecco's modified Eagle medium containing 10% fetal bovine serum, 0.02M L-glutamine, 1,000 Units/mL Penicillin, and 1,000 μg/mL Streptomycin, as previously described (Banerjee et al. 2020).

Viruses

A clinical isolate of SARS-CoV-2 (SARS-CoV-2/SB3-TYAGNC) was propagated in Vero E6 cells and virus stocks were quantified and sequenced as previously mentioned (Banerjee et al. 2020). Virus stocks were maintained at -80°C. Work with SARS-CoV-2 was performed in a containment level 3 laboratory and all protocols were approved by the McMaster Presidential Biosafety Advisory Committee.

HCoV-NL63 (NR-470; BEI) was propagated on VeroE6 cells and viral titers were quantified using the 50% Tissue Culture Infectious Dose (TCID50) method. TCID50 values were determined using the Reed-Muench method (Reed and Muench 1938). Viral stocks were stored at -80°C.

Bacteria

Seven bacterial strains were used in this work *B. multivorans* (ATCC17616), *K. pneumoniae* (N25C9), *S. aureus* (IIDRC0017), *S. constellatus* (C1050), *S. intermedius* (B196), *S. pneumoniae* (R6), and *S. sanguinis* (GC83). All strains with the exception of *B. multivorans* were provided by Michael G Surette.

Frozen bacterial strains were independently cultured for 48 hours on agar.

8.9.3 Method Details

HUBDesign Pipeline

The Hierarchical Unique Bait Design (HUBDesign) pipeline aims to identify oligonucleotides (probes) which will specifically hybridize with nucleic acids from any member of a clade, and to do this for as many clades as possible within a given set of organisms. The probes designed are intended to enrich loci which are common to members of a clade, while being unique to that clade. This does not necessarily allow for whole genome enrichment, however by having probes at multiple hierarchical levels conserved genomic regions can be targeted by higher level probes while more variable regions are targeted by probes specific to a species or genome. The design of these hierarchical and unique probes is achieved through three design phases: clustering, identifying, and filtering.

In the clustering phase, sequences from the input organisms are grouped together and collapsed into representative sequences which are then used in the subsequent phase (Figure 8.6.2). To avoid computationally expensive all-vs-all comparisons, HUBDesign requires annotated genomes which allows rapid identification of gene families for clustering. The grouping of similar sequences serves as the source of hierarchical information and reduces computational effort. This reduction makes it possible to rapidly design probes for inputs ranging from dozens of viruses to thousands of bacteria. This implementation of the HUBDesign pipeline performs clustering on gene sequences as annotated using Prokka (Seemann 2014). For each annotated gene family, sequences are aligned using MAFFT (Katoh and Standley 2013), and then a neighbour-joining (Saitou and Nei 1987) tree is generated based on the uncorrected edit distance. The uncorrected distance is used rather than the evolutionary distance as the actual

difference between sequences is more important in a probe design context. Clusters are generated from the tree by selecting sub-trees which have a maximum root-to-tip divergence less than the maximum amount of divergence which still allows hybridization between a probe and target. A representative consensus sequence is generated for each gene cluster, and these sequences are assigned to the LCA of each organism represented (Figure 8.6.3). The representative sequence may not be appropriate to use if the represented organisms do not make up a significant portion of all descendants of the LCA. Laterally transferred elements are cases where a shared sequence is not useful in identifying a group of organisms. If an element is horizontally transferred to a distantly related organism, the representative sequence would be assigned to a node further from the tips of the tree. To prevent the use of these non-identifying sequences, a penetrance threshold is set. The penetrance for a representative sequence is calculated after its cluster has been assigned to the LCA of the genomes represented. The proportion of the LCA node's descendants which actually possess a member of the cluster is the penetrance of the representative sequence. All representative sequences assigned to a given node which pass the penetrance threshold are concatenated into a pseudo-genome for that node. The individual representative sequences are buffered to prevent selecting probes which straddle non-adjacent sequences. (Figure 8.6.4)

In the identification phase of the pipeline, pseudo-genomes are provided to a modified version of the program Basic OligoNucleotide Design (BOND) (Ilie et al. 2013) (Figure 8.6.5). BOND was originally designed for the rapid identification of a single unique oligonucleotide for each gene on a chromosome. Unique is defined as sharing no more than 15 consecutive identities and no more than 75% overall identity with any other oligo in the input. The entire program was modified to handle larger inputs allowing for the identification of multiple unique oligos for each genome in a set of genomes. The modified version (SA_BOND) is strand aware and tolerant of sequences which are repeated within a single genome. In terms of memory, this is the most computationally expensive phase in the pipeline. This is also the step where the specificity of the probes is improved far beyond a naïve tiling strategy. All probes are unique to the taxa they were designed for, allowing a probe targeting a taxon higher in the tree to capture all that node's descendants without also capturing unrelated organisms. In the filtration phase, oligonucleotides are removed which hybridize to off-target or known background sequences like the human genome or transcriptome (Figure 8.6.6). This implementation of the pipeline utilizes BLAST (Altschul et al. 1990) to identify and exclude candidate oligos with significant hits against background. The thresholds for this can be set based on how conservative one wishes to be. Remaining candidate oligos which are overlapping are collapsed into contiguous regions, and then low-complexity intervals are excluded using sdust (Morgulis et al. 2006). The last step of the filtering phase selects the final set of oligonucleotides from the candidates in a manner which attempts to reduce bias between organisms. The goal is to have the number of probes targeting each organism to be as close as possible across the organisms (Figure 8.6.7). Probe count balancing is achieved by varying tiling density such that oligos targeting over-represented organisms are tiled less densely than oligos targeting under-represented organisms. It has been shown that higher tiling density can improve the capture efficiency of probes (Bertone et al. 2006). Varying tiling density in this way is a trade-off between the number of unique targets and the efficiency with which those targets are captured. The

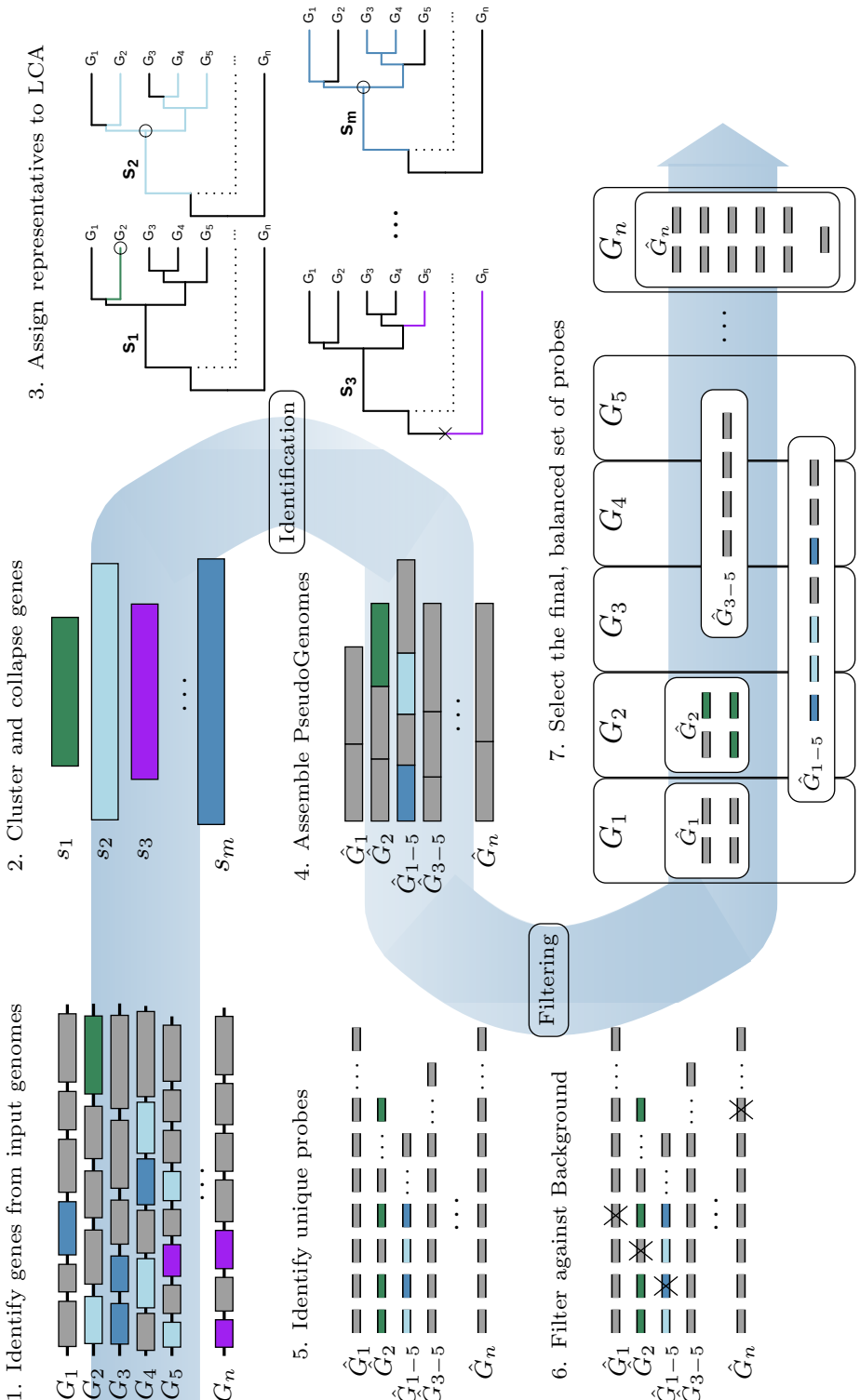

FIGURE 8.6: HUBDesign takes annotated genomes as input. In step 2, genes with sequences which are at most 15% divergent (Noted by colour) are collapsed into representative sequences. In step 3, Representative sequences are assigned to the LCA of organisms which possess the sequences represented. s_1 represents only one organism and is assigned to the leaf node. The LCA of organisms represented by s_1 and s_m is the same node and both represent the majority of the 5 descendants of this node; they are considered valid representatives. The LCA for s_3 is the root of the tree, however it is only representing a small fraction of the organisms in the tree. It is excluded in subsequent steps. In step 4, all representatives assigned to a node are concatenated into a pseudo-genome for that organism, note that s_2 and s_m were both assigned to the same node, and therefore are in the same pseudo-genome. In step 5, SA_BOND is run to identify all probes unique to each pseudo-genome. In step 6, probes which would capture off-target sequences are removed. In The final step, a set of probes is selected from the candidate probes which balances the number of probes per input genome.

hierarchical nature of the probes constrains the ability to balance across organisms as probes can target multiple organisms which have different levels of coverage. HUBDesign takes an iterative approach to this problem. Tiling density is performed on probe regions which are composed of candidate probes with contiguous start positions, and which all target the same taxon. The entire probe region is said to target that taxon, or all the descendent genomes if the taxon is at an internal node in the tree. On each iteration of the procedure a target number of probes per genome is set based on evenly dividing probes across organisms. Tiling strategies are determined for each probe region based on length, leaf taxa targeted, and a minimum tiling density specified by the user. If it would not be possible to bring the number of probes for an organism down to the target level even if all probe regions targeting that organism were tiled at minimum density, then all of those probe regions are assigned to be minimally tiled. If instead it is not possible to bring the number of probes for an organism up to the target even by tiling at maximum density (probes spaced apart by only 1 bp), then all probe regions targeting that organism are assigned to be maximally tiled. If a probe region targets both an under-targeted and an over-targeted organism, it will still be maximally tiled to ensure probes are available for the under-targeted organism. On subsequent iterations the target number of probes per organism is updated to reflect that the extremes are accounted for and should be excluded from consideration. Organisms with below target numbers of probes leave more probes available for the other organisms, and those with above target leave fewer. The new target is set by dividing the available probes evenly across the remaining organisms. Tiling classes are then reassigned based on this new target and iteration continues until tiling classes cease to change. Tiling density of the probe regions not assigned to the extremes are set in the order of constraint: Probe regions which target organisms which are targeted by the fewest probe regions are processed first. Within each organism, probe regions that target the most organisms are processed first. The tiling density is set as the weighted average of the minimum and maximum tiling density, weighted towards the max when there are few potential probes for the organism, or if the current number of probes is far from the target. Processing probe regions in this order allows coarse adjustments to the balance from the most constrained regions, and fine tuning from the least constrained.

Coronavirus Probe Design

In this section we describe the implementation of HUBDesign used to design probes for 56 coronaviruses taken from RefSeq (O'Leary et al. 2016). The set of viruses covers the *Alpha-*, *Beta-*, *Gamma-*, and *Deltacoronavirus* genera and includes the four major seasonally circulating human coronaviruses, as well as SARS-CoV-2 and viruses responsible for earlier novel coronavirus outbreaks. (Table A4.1).

In the clustering phase, distances were calculated using the distmat tool from EMBOSS (Rice et al. 2000), and neighbour joining trees constructed using the neighbour tool from the phylip package (Felsenstein 1989). Clusters were generated from sub-trees with a maximum root-to-tip divergence of 15%. This threshold was selected as probe sequences which diverge from their targets by less than this are most likely to successfully hybridize (Mason et al. 2011; Delsuc et al. 2016). Each cluster was assigned to the LCA in a dendrogram based on the lineage recorded in NCBI's taxonomy database for each genome (Schoch et al. 2020). A

penetrance threshold of 50% was used, therefore only representative sequences which were based on at least half of the descendants of the LCA were included in the pseudo-genome for the LCA. We observed that 90% of all clusters had at least 50% coverage, and this value ensured that pseudo-genomes representing all input taxa were constructed.

Candidate probes were identified using SA_BOND to search for all unique oligonucleotides of length 75 across all pseudo-genomes. BLASTn was used to find and exclude any candidate probe which matched the human genome (GRCh38 (Schneider et al. 2016)). Matches were considered significant if they had at least 75% identity, were at least 30 bp long, and had an e-value less than 0.01. Low-complexity regions were excluded using sdust (Morgulis et al. 2006) with the default parameters: a 64 bp window, and a score threshold of 20. The latter approximately corresponds to a sequence where 80% of the nucleotide triples in the window are the same. The final probe set was selected with a target minimum tiling density of 5x and a maximum of 13500 probes. This is the smallest number of probes for which the most balanced distribution of probes across targets still allows for tiling densities to vary between 5x and the maximum (1bp spacing). Any fewer and the most balanced configuration has all probe regions tiled at either of these extremes. Probes were balanced by treating all hierarchical levels independently, which maximized the number of probes for taxa represented at only one hierarchical level.

Sepsis Probe Design

The sepsis probe set was produced with an earlier version of HUBDesign. The input database contained 1926 bacterial genomes across 81 species and 35 genera (Table A4.2). All genomes were acquired from the PATRIC database (Davis et al. 2020).

Gene clusters were generated and assigned based on nominal taxonomy. Each genus was independently and recursively processed. All genes of the same name with a minimum penetrance of 95% in the particular genus were aligned using MAFFT (Katoh and Standley 2013). Then the conservation was calculated, and a consensus sequence was generated. A minimum of 85% conservation was required. Clusters meeting the penetrance and conservation thresholds were added to the genus's pseudo-genome, while those which failed to meet the criteria were broken up into clusters based on species. Each of these was realigned and tested against the thresholds once more. Only clusters which passed at the genus or species level were included in any pseudo-genome.

Up to 100 candidate 100bp probe regions per pseudo-genome were identified using SA_BOND. All 75 bp sub-sequences of these probe regions were considered candidate probes, and a maximum number of tiled probes were produced.

BLASTn was used to find and exclude any candidate probes which matched the human genome (GRCh38 (Schneider et al. 2016)). Matches were considered significant if they had at least 75% identity and were at least 20 bp long. All remaining contiguous probe regions after filtering were tiled with probes which were spaced apart by 5 bp. As 100 bp regions were identified this resulted in most loci being targeted by 5 probes.

Comparative Probe Design

As a baseline comparison, naïve tiling was performed on the genomes of the 56 reference coronavirus genomes, and on the 1926 sepsis pathogens. Probes were identified by selecting each 75bp subsequence of each genome, spaced apart by 15bp, and retaining only unique sequences.

A recently described computational method for probe design is the CATCH python package (Metsky et al. 2019). It can also be configured and applied to the task of finding probes, meant for identification. CATCH was run on the 56 reference coronavirus genomes. We selected two sets of hybridization parameters to account for the fact that CATCH and HUDesign use opposing approaches to probe selection. HUBDesign identifies candidate probes via elimination. A probe is only considered if it is unlikely to hybridize to another target in the dataset. As a result, liberal hybridization parameters make HUBDesign more strict and the resulting probes more specific. The opposite is true for CATCH. It identifies targets to which each probe will likely hybridize and selects an optimal set of probes with desired coverage. The two sets of hybridization parameters provided to CATCH were a strict set which did not allow mismatches, and a more permissive set allowing up to 18 mismatches but requiring an island of exact matches 15bp long. The permissive parameters are similar to those used by BOND (Ilie et al. 2013) to eliminate non-specific probes. Probes which were 75bp long and spaced out by 5bp were designed using the identify flag, targeting 20% coverage of the target genomes. The human genome (GRCh38 (Schneider et al. 2016)) was provided as a blacklist sequence.

Viral RNA Extraction

For SARS-CoV-2 infections, 2×10^5 Calu-3 cells were seeded in each well of a 6-well plate. A clinical isolate of SARS-CoV-2 (SARS-CoV-2/SB3-TYAGNC (Banerjee et al. 2020)) was used to infect Calu-3 cells at a multiplicity of infection of 0.01. Cells were harvested 48 hours post infection and total RNA was extracted from infected Calu-3 cells using the QIAamp viral RNA Mini kit (Qiagen) according to the protocol outlined by Banerjee *et al.* (Banerjee et al. 2020).

Viral RNA from HCoV-NL63-infected VeroE6 cells was extracted using the Qiagen RNeasy kit with minor modifications. Briefly, 100 μL of supernatant was mixed with an equal volume of RLT lysis buffer and 25 μL of 20 mg/mL proteinase K (Invitrogen). Samples were vortexed and incubated at 56°C for 15 minutes. Following, 200 μL of 70% ethanol was added to the solution and RNA was eluted as outlined in the RNeasy manufacturer's protocol.

Human total RNA was extracted from peripheral blood mononuclear cells using the RNeasy extraction kit (Qiagen) according to manufacturer's protocols.

Coronavirus Sample Preparation

The copy number of both viral extracts was determined using the Luna Universal Probe One-Step RT-qPCR Kit (NEB) and the primers in Table A4.3, while human RNA was quantified using Qubit RNA HS Assay Kit (Thermo Fisher). All three sets of RNA were separately treated with DNase I (NEB), and the human ribosomal RNA depletion kit (NEB)

according to manufacturer's protocols. Following this, samples were thermally fragmented to roughly similar fragment size distributions.

Four mock samples and a negative control were prepared by combining RNA extracts from the two viruses and total human RNA background. The samples were prepared at a low level of two hundred RNA copies and a higher level of twenty thousand RNA copies in a 1:1 ratio of both viruses (EL and EH). Two additional samples were prepared with two thousand RNA copies of one virus and two hundred thousand RNA copies of the other. In the LoHi sample SARS-CoV-2 was the low-level virus, and HCoV-NL63 was in the HiLo sample. All samples including the negative control contained one hundred nanograms total human RNA.

While qPCR copy numbers were used to prepare the mock pools, these copy numbers do not represent full genome copies present at that level, only that the PCR amplicon is present at that estimated copy number. Variation occurs as sub-genomic RNA molecules are created during the coronavirus life cycle (Kim et al. 2020; Fehr and Perlman 2015). Different regions of the genome are therefore likely to be better represented than others, with the expectation that the $3'$ regions of the genome will have the highest copy numbers (Figure A4.1).

To generate a confident baseline to account for this, an additional sample containing a mixture of the viruses was shotgun sequenced. The sample nominally contained 2.68 million copies of the HCoV-NL63 genome and 5.1 million copies of the SARS-CoV-2 genome.

All pooled samples were prepared in triplicate. First strand synthesis was performed using Superscript III Reverse Transcriptase (Thermo Fisher) according to manufacturer's protocols with 250ng random hexamers. This reaction was purified using AMPure XP beads (Beckman Coulter) at a 1.8X ratio to sample. Single-stranded libraries were then prepared using the SRSLY Nanoplus kit from Claret Bioscience. Attachment of indexing adapters was performed according to the protocol outlined by Kircher *et al.* (Kircher et al. 2012), after which the indexed libraries were purified using MinElute spin columns (Qiagen). Each replicate was split in two where one half was enriched prior to sequencing and the other shotgun sequenced.

Probes were synthesized through Ann Arbor Biosciences myBaits custom DNA-Seq program. Enriched samples were processed according to the Ann Arbor Biosciences myBaits targeted enrichment protocol version 4.01 (Ann Arbor Biosciences 2018) with a 72-hour capture step. After enrichment, both shotgun and enriched libraries were quantified alongside the Illumina PhiX standard, before being pooled to equimolar quantities and undergoing a gel-based size selection for fragments between 150-500 bp. Pools were then sequenced on an Illumina Hiseq 2x90 flow cell.

Bacterial Sample Preparation

Three mock pools were prepared in triplicate which contained human blood spiked with the seven bacterial strains listed above. With the exception of *S. sanguinis*, the genomes of the spiked bacteria were in the dataset used to design the probes. The three pools were at Low, Medium, and High concentrations (10^1, 10^3, and 10^6 CFU/mL) of each bacteria. One additional pool at each concentration was also prepared where all the bacteria were included

but the blood was omitted. Three negative controls were prepared with water and five negative controls were prepared with blood only.

Cultures of each of the strains to be spiked were suspended and diluted to the above concentrations in 0.85% saline. At each of the Low, Medium, and High concentrations the matching CFU counts for each strain were pooled together and pelleted. Fresh human blood was drawn from a healthy donor into tubes containing EDTA as an anticoagulant. Bacterial pellets for each pool were resuspended in the blood, or sterile saline for the no blood control.

DNA was extracted using the High Pure Viral Nucleic Acid Extraction large volume kit from Roche, according to their version 7 protocol for a 1mL sample. Pools were sonicated to ~200bp using a Covaris S220 focused ultrasonicator. Pools were divided such that each pool would have one shotgun library and two enriched libraries. Samples for both shotgun and enrichment were processed for library preparation using the NEBNext Ultra II DNA Library Prep Kit for Illumina (CN E7645) according to manufacturer's specifications. Probes were synthesized through Ann Arbor Biosciences myBaits custom DNA-Seq program. Samples to be enriched were processed using the Arbor Biosciences myBaits v5.0 kit with the high sensitivity protocol, according to manufacturer's specifications with a 63°C hybridization temperature.

After enrichment, all samples including those prepared for shotgun sequencing were quantified using KAPA SYBR FAST Bio-Rad iCycler Master Mix (Sigma Aldrich, CN KK4608), run alongside PhiX Control Standards (Illumina, CN FC-110-3001). Based on these concentrations, samples were pooled to equimolar amounts, then concentrated using a Minelute PCR purification column (Qiagen, CN 28006). This concentrated pool was then size-selected using NuSieve GTG Agarose (Lonza, CN 50081) in a 3% 1X TAE gel. Only molecules with a total length between 200bp and 500bp were excised from the gel. The final pool was purified from the gel using a Minelute Gel Extraction column (Qiagen, CN 28604), and eluted in 20μL. This final pool was sequenced on an Illumina HiSeq 2x90 flow cell.

Sample Analysis

After demultiplexing, samples were trimmed and merged using FastP (Chen et al. 2018). Any orphaned reads were treated as single ended reads going forward. Reads were string deduplicated using prinseq (Schmieder and Edwards 2011) as identical reads are much more likely to be PCR duplicates generated during sample preparation than identical templates generated during fragmentation. String deduplication was selected over mapping-based deduplication as a balance between removing PCR duplicates and retaining true biological duplicates from high-copy numbers. Reads for were then filtered against the GRCh38 (Schneider et al. 2016) version of the human transcriptome (coronaviruses) or genome (sepsis) using BWA (Li and Durbin 2009). Reads from coronavirus libraries were also mapped against amplicon sequences in Table A4.4, to assess possible aerosolized contamination from previous PCR reactions performed in our and neighbouring labs.

Reads were mapped to contiguous probe regions, flanked by up to 350 bp of upstream and downstream sequence. Reads mapping at this step were assigned to a probe and that probe's associated taxa. Non-mapped reads were competitively mapped to the genomes of the spiked

organisms: SARS-CoV-2 and HCoV-NL63 coronavirus or *B. multivorans*, *K. pneumoniae*, *S. aureus*, *S. constellatus*, *S. intermedius*, *S. pneumoniae*, and *S. sanguinis*. If any read overlapped a region targeted by a probe, the read was assigned to that probe. Unassigned reads were mapped to the set of other genomes used to design the respective probe sets. SAMtools (Li et al. 2009) was used at each filtering step, and to calculate depth of coverage.

Conversion from nominal copy number to mass of viral RNA was calculated for both viruses using the high copy number shotgun baseline sample. For both viruses, the genome was broken into non-overlapping regions of the same length as the PCR amplicon generated during copy number quantification. The ratio of the depth of coverage in each region to the reference region then was used to calculate the copy number of each region across the genome. Using a conversion of 320 g/mol/nt for ssRNA, and the length of the reference region for in each virus, the mass in each genomic region was calculated and integrated across the genome to estimate the total viral RNA given the nominal copy number.

To assess potential off-target enrichment of the human background, regions of interest in the human reference were identified as any region with read counts above the 95th percentile assuming read counts at each position are Poisson distributed with a mean equal to the average read depth across the reference. Reads from enriched libraries which mapped to these regions were additionally mapped to the probes and target genomes as above to determine by which, if any, probes these reads were captured.

8.9.4 Quantification and Statistical Analysis

Logistic regression was used to assess the effect of enrichment. The probability of any given read in a library mapping to a target genome was used as the regressand with enrichment input concentrations as regressors. Fold enrichment for this analysis was the fold change in the odds of observing an on target read, as determined by the value of coefficient estimate. For the CoV dataset, the mass of input RNA for each virus and the interaction between were used as continuous predictors, while whether the samples were enriched was used as a categorical predictor. For the Sepsis dataset, the number of on target reads was corrected based on the proportion of on target reads observed in the blank samples. This correction was done on a by sample basis by subtracting a number of reads equal to the number of reads observed in the blanks adjusted by the ratio in library size between each sample and the blanks. With this adjusted read count, the proportion of on target reads was used as the regressand, with spike level, enrichment status, and their interaction as categorical variables. Significance of enrichment was determined using the Wald test on the regression coefficient estimate for the enrichment parameter.

Linear regression on the with the number of doublings in fold enrichment as the regressand and various parameters of the baits as regressors was used to assess the performance of baits targeting different genomic regions. For both the CoV and Sepsis datasets continuous predictors were transformed as necessary to meet the assumptions of linearity between the predictors and the response.

Two linear regressions were performed for the CoV dataset. The first considered all genomic regions, while the latter only considered genomic regions targeted by baits. When considering all regions separate coefficients were estimated for targeted and untargeted regions for the following predictors: the deviation in GC content from the mean GC content across the genome; the proportion of reads observed in a region in the shotgun baseline sample; and the number of doublings in mass of RNA for both viruses. These predictors were untransformed. All regions were assumed to be in the SARS-CoV-2 genome, with a coefficient estimating the effect of actually being from HCoV-NL63. When considering only genomic regions, the density of probes (number of probes per kilobase), and the mean divergence of probes from their target were also included as predictors.

Linear regression performed for sepsis dataset with GC content, probe divergence, probe density, the baseline levels in the shotgun as continuous regressors. Whether the sample was in a blood or water background was used as a categorical predictor. Only data from the High concentration samples and positive control were used as most probe regions had no data in the shotgun samples at lower concentrations.

Chapter 9

Applications of Targeted Metagenomics

9.1 Preface

What follows is a brief description of several other projects to which I have contributed. These projects have in general been targeted at ancient DNA, and in all cases the goal was a set of probes to enrich the particular targets of interest, ranging from entire parasite species to particular genes involved in horse coat colour.

At the time of writing these projects are still in progress with collaborators. Where possible preliminary results have been included.

9.2 Intestinal Parasites

9.2.1 Contributions

This work was done in collaboration with Marissa Ledger, who was at the time affiliated with the University of Cambridge, and Tyler Murchie from the McMaster ADNA Lab. Marissa provided a set of parasite taxa of interest, with Tyler and Marissa performing the wet lab work.

9.2.2 Abstract

A late bronze age settlement known as Must Farm was previously described by (Ledger et al. 2019). At the site, parasite eggs were microscopically identified from a wide variety of parasite species. While this expanded our knowledge of human associated parasites in the bronze age, targeted DNA capture for parasite species could further expand this picture. To this end I used `SA_BOND` (Dickson et al. 2021) in an iterative manner to design hierarchical probes to capture the desired diversity of intestinal parasites. Probes were designed primarily for mitochondrial sequences, as well as rRNA and a selection of nuclear genes. Of the 130 parasite species the final set of 30,240 probes could target 98 (75%) of the species. Preliminary sequencing results largely recapitulated the original findings.

9.3 Human Gut Microbiome

9.3.1 Contributions

This work was done in collaboration with Henrik Poinar with funding from the Canadian Institute for Advanced Research. Dr. Poinar provided the list of taxa of interest across human proteomes. I was responsible for assembling a set of probes to capture as many as possible of the taxa of interest.

9.3.2 Abstract

In almost every case where one might wish to examine the human gut microbiome, from ancient DNA to clinical settings, the human background will be a huge component of the DNA sequenced. Selectively depleting the entire human genome can be quite challenging and is insufficient in the cases where more than a human background is expected, such as sediments from historical latrines where environmental background is also of concern. Instead, this is an ideal case for `HUBDesign` (Dickson et al. 2021). There is a known diversity of targets of interest (human gut microbiota) but in any given sample it is unknowable which if any are present.

I applied `HUBDesign` as previously described (Dickson et al. 2021), with modifications to the initial phase of clustering genes into representative sequences. In many applications the annotated gene names can be used as a shortcut to avoid computationally expensive all-vs-all comparisons of gene sequences. However, as the diversity of taxa included in the design increases the failures of this strategy become more apparent. Some genes which are very similar in sequence have completely distinct gene names, and name collisions can also become a problem with inconsistent annotation. To resolve this, `VSEARCH` (Rognes et al. 2016) was used to rapidly pre-cluster the gene sequences. The clustering and creation of representative sequences then proceeds as previously described using the pre-clustering IDs in place of gene names.

Using this pre-processing in combination with `HUBDesign` I was able to construct a set of 12,142 75 bp probes which could be used to efficiently identify 1792 different strains of human gut microbiota spread across 435 species, 51 genera, and 24 families.

9.4 Butyrate Metabolism Genes

9.4.1 Contributions

This work was also performed in collaboration with Dr. Hendrik Poinar, as well as Melanie Kuch and George S. Long. The former provided the sequences corresponding genes in the butyrate metabolic pathway. Melanie performed the wet lab work, and George performed the downstream analysis. I was responsible for designing a set of probes to efficiently capture the diversity of sequences involved in butyrate metabolism.

9.4.2 Abstract

Butyrate is a metabolic product often used as an indicator of the interactions between host and microbiota health (Zhang et al. 2021). It is the main product of gut fermentation, and its levels are influenced by diet and the composition of the gut microbiome. In this project the goal was to sequence genes involved in butyrate metabolism across humans and primates through both fecal and coprolite samples as well as explore the evolutionary history of gut microbiomes and host behaviour. `HUBDesign` was used to design the probes, however rather than its usual goal of probes which can identify taxa, the goal was identification of butyrate genes. Metabolic processes can be divided across bacterial species, and the genes involved can be horizontally transferred, therefore a functional perspective was more useful. The final probe set of 38,420 probes was designed to capture 31,807 alleles across 29 different genes involved in butyrate metabolism.

9.5 CAPDesign: Multiplex primer set design

9.5.1 Contributions

This work is part of a larger collaboration with Safeguard Biosystems attempting to create a reusable solid state array which can capture DNA sequences which are identifying for pathogens as well as their AMR genes. While the probes for the array can be designed using `HUBDesign`, a pre-amplification of all targets for the probes is required to achieve the desired sensitivity. Thus, the motivation for designing highly multiplexed PCR primer sets.

I am responsible for the conceptualization of the primer set design, as well as the code to create it. Rabia Raees has performed the work of testing the designed primer set, including running PAGE gels to visualize the amplification products of various primer pools. While testing results are so far are promising, the work is not yet complete.

9.5.2 Abstract

PCR can also be used to bias sequencing efforts towards targets of interest. For single, or small numbers of targets this is relatively straightforward, however the challenge of primer design scales exponentially with the number of targets. This is because each additional primer can potentially interact with every other primer with a major consequence being primer dimers which preferentially amplify over the target of interest. Recent work by (Xie et al. 2022) describes `SADDLE`: a system for designing highly multiplexed primer sets with a low likelihood of dimer formation based on simulated annealing. I have extended the concept with the concept of primer interchangeability: where two candidate primers can be swapped without a change to the set of targets amplified. I have implemented this with a set-cover algorithm to design minimal sets of multiplex primers with reduced probability of dimer formation. As an application, a set of 140 primers was designed to amplify 314 target sequences from 10 strains across 4 genera. Based on the preliminary validation work, these primers effectively amplify the targets of interest, without amplification of a human background, and without significant dimer formation.

9.5.3 Figures

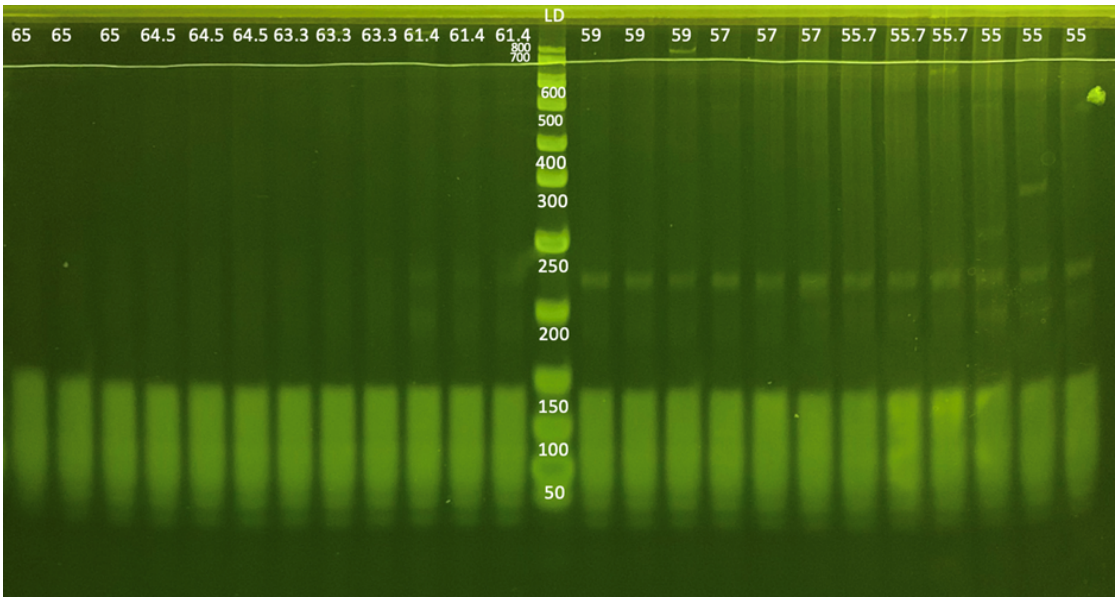

FIGURE 9.1: A primer set designed with CAPDesign has a low propensity to form primer dimers. 10% Urea PAGE of template free amplification reactions with varying annealing temperatures. Primer dimers begin to form as the annealing temperature drops below 61.4°C, however compared to the high concentration of primers present (140μM across 140 primers), the rate of dimer formation is low.

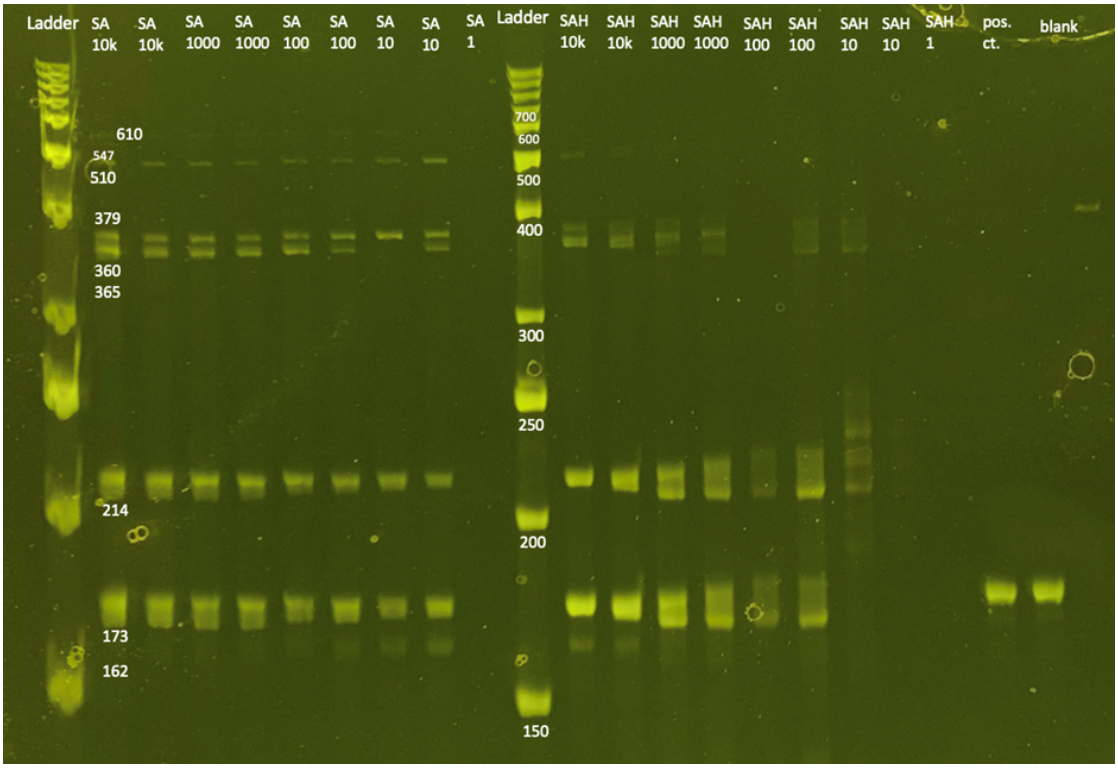

FIGURE 9.2: 10% Urea PAGE of amplifications with genome copy numbers from 1 to 10×10^4 of *S. aureus* without (SA) and with 11.1ng of human DNA spiked-in (SAH). On-target amplification is observed down to 10 copies of the bacterial genome. The addition of the human background reduces the sensitivity for larger amplicons but does not introduce significant off-target products.

9.6 Horse Coat Colour

9.6.1 Contributions

This project is in collaboration with Rachel Miller from the McMaster Ancient DNA center. Rachel collected the current literature on the genetics of horse coat colour and provided the known variants of interest for probe design. She is also performed the wet-lab work. I was responsible for generating a probe set which could capture the diversity in horse coat colour.

9.6.2 Abstract

The outward appearance of an organism is a critical part of their adaptation to their environment: Camouflage, conspecific communication, and thermotolerance. The genetics of coat variation are unusually well understood in horses due to sophisticated artificial breeding programs, but little is known of the genetics or appearance of ancient horses. The genes involved in coat colour are often pleiotropic, with potentially deleterious effects in some cases. Elucidating the diversity in coat colour in North American horses, prior to the influence of

humans will give insight into the selective pressures balanced across the pleiotropic effects of coat colour variants.

Capturing nuclear DNA in an ancient context is challenging, and we are employing a targeted DNA capture approach. Probes were designed to capture 66 variants in 19 genes associated with coat colour. Probes were additionally designed to capture the exons in these genes. The final set of 4782 probes target all but 3 of the variants, all of which are single nucleotide polymorphisms (SNPs) in the *KIT* genes.

9.7 Ancient Antibodies

9.7.1 Contributions

This in progress work focuses on enriching the genomic DNA of immune cells which have had their populations expanded in response to exposure to a pathogen. This can provide insights into past infections of extant individuals but can also be used to define the pathogen landscape in the past. RNA viruses are of special interest here as their genomes are too labile to be preserved and detected for all but the most recent past.

Sergey Yegorov collected the germline sequences for immune gene segments. I conceptualized the probe design framework, and wrote the code for generating, filtering, and selecting the probes. The work of validating the probe set designed was performed by Tess Wilson. Peter Zeng and I performed analysis generated sequencing data.

9.7.2 Abstract

The immune system is one of the most complicated systems in human biology as it must face constant challenge from a huge diversity of potential pathogens. The genomes of immune cells are modified via recombination to generate the diversity of immune receptors to combat pathogens. The entire population of immune cells defines an individuals' immune repertoire. As this varies over time a snapshot of repertoire provides information of the diversity of pathogens and the frequency of exposures an individual has faced. This can provide a window into studying pathogens which are difficult to identify directly such as RNA viruses in an ancient context. To that end I have designed a set of probes to capture the recombined germline sequences of B- and T-cells, using a graph based method to identify minimal sets of probes covering all unique junctions between recombined immune segments.

The preliminary set of 9435 probes was tested on fractionated blood samples of individuals both pre- and post- vaccination for influenza. We were able to show an average fold enrichment of immune gene segments of 1000x. However, there are indications that random insertion of nucleotides during the recombination process hampers the effectiveness of the probes. This is especially true for immune loci where two phases of recombination occur. We are currently working to explore this further.

9.7.3 Figures

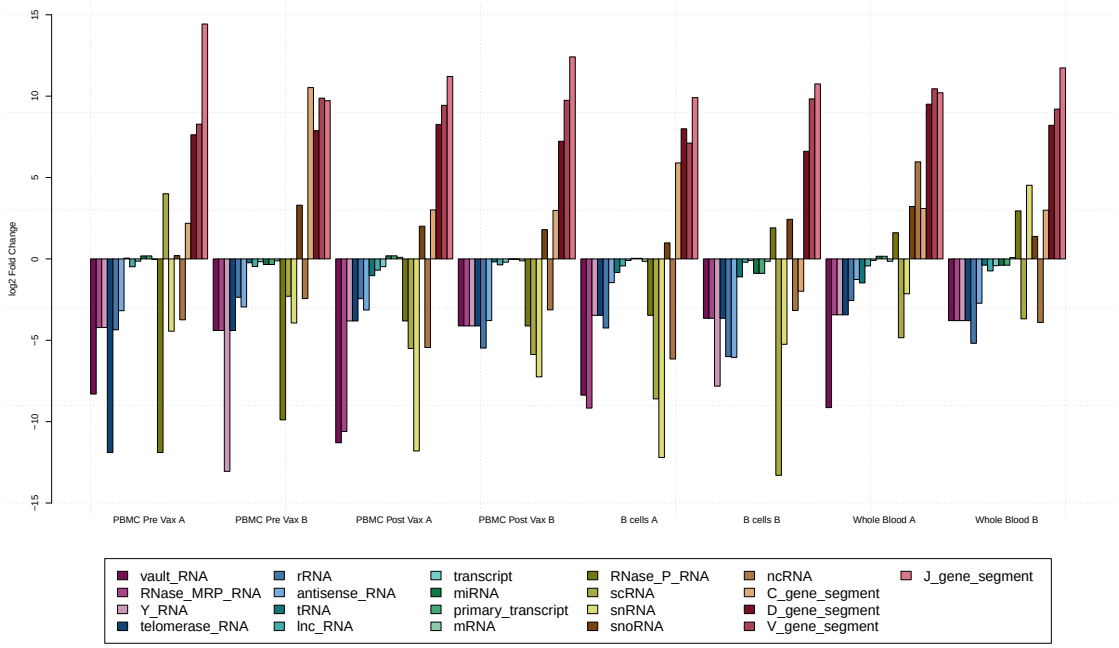

FIGURE 9.3: Immune loci are preferentially enriched using probes designed for the recombined junctions between immune segments. A barplot of fold-enrichment in proportion of reads assigned to particular transcript bio-types comparing proportions in enriched samples to a shotgun background. C, V, D, and J segments are recombined segments of immune loci. The proportion of reads mapping to these genomic regions are between 32 and 32000 times higher in enriched samples as compared to a shotgun sequenced baseline.

9.8 Discussion

Each of these projects has the core goal of investigating some targets of interest in samples with potentially complex backgrounds. In many cases we can leverage the fact that biological sequences are shared between related individuals. By targeting the shared sequences at different taxonomic levels, we can efficiently and specifically capture a wide array of targets. In cases where it is more important that the sequences of interest are captured preferentially over the background than it is that all the targets are immediately differentiable, we can even make use of sequences that are shared by chance or through horizontal gene transfer. With clever consideration of what the candidate sequences are, and how these candidates relate to each other it is possible to efficiently design probes and primers applicable to a wide variety of contexts.

Afterword

This thesis has been the result of exploration on many branches of bioinformatics, with two main trunks. The LCR trunk and its branches stem from an interest in understanding the fundamental relationships between sequence characteristics and their consequences. The metagenomics trunk and its many, continually sprouting, limbs are much more applied with a focus on how to use our understanding and the body of knowledge of sequences to more efficiently probe the world around us. Leaping between the branches of the tree has been a challenge and I trust this thesis to demonstrate the heights I, my supervisors, my collaborators, and my volunteers have climbed together.

Appendix

Appendix A

Chapter 3 Supplement

TABLE A1.1: NCBI genome access date (mm/dd/yyyy) and assembly version numbers.

	S. cerevisiae	*H. sapiens*	*A. thaliana*	*C. elegans*	*D. melanogaster*
Access Date	04/30/2022	05/30/2022	06/09/2022	06/09/2022	06/10/2022
Assembly Version	GCF_000146045.2	GCF_000001405.40	GCF_000001735.4	GCF_000002985.6	GCF_000001215.4

TABLE A1.2: Alternate Seg parameters used to identify LCRs from protein sequences in the five model organisms. Correlation coefficients and number of LCR data points are given for each parameter set.

parameters			*S. cerevisiae*				*H. sapiens*				*A. thaliana*				*C. elegans*				*D. melanogaster*			
W	**K1**	**K2**	**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**	
		1.7	0.420	1182	0.367	0.471	0.341	5812	0.315	0.366	0.454	5601	0.431	0.477	0.291	3588	0.257	0.324	0.350	4222	0.320	0.379
	1.7	**2**	0.474	1182	0.423	0.522	0.361	5812	0.336	0.386	0.475	5601	0.452	0.497	0.325	3588	0.292	0.357	0.382	4222	0.353	0.410
		2.7	0.591	1182	0.548	0.630	0.478	5812	0.456	0.500	0.596	5601	0.577	0.615	0.477	3588	0.448	0.505	0.535	4222	0.511	0.559
		1.9	0.404	2078	0.363	0.443	0.347	9944	0.328	0.366	0.452	10314	0.435	0.469	0.288	6456	0.263	0.313	0.358	6200	0.334	0.382
12	**1.9**	**2.2**	0.443	2078	0.403	0.480	0.340	9944	0.320	0.359	0.468	10314	0.451	0.485	0.341	6456	0.317	0.365	0.394	6200	0.370	0.417
		2.9	0.545	2078	0.510	0.578	0.481	9944	0.464	0.498	0.595	10314	0.581	0.609	0.496	6456	0.475	0.516	0.560	6200	0.541	0.579
		2.1	0.416	3359	0.384	0.446	0.296	14282	0.279	0.313	0.427	16318	0.413	0.441	0.296	10530	0.276	0.315	0.369	8669	0.349	0.389
	2.1	**2.4**	0.443	3359	0.412	0.473	0.308	14282	0.291	0.324	0.457	16318	0.443	0.470	0.326	10530	0.307	0.345	0.406	8669	0.386	0.425
		3.1	0.550	3359	0.523	0.575	0.556	14282	0.543	0.569	0.650	16318	0.640	0.660	0.580	10530	0.566	0.594	0.631	8669	0.617	0.645
		1.7	0.425	654	0.352	0.492	0.347	3167	0.312	0.381	0.483	2712	0.450	0.515	0.276	1976	0.230	0.321	0.336	2880	0.299	0.372
	1.7	**2**	0.466	654	0.396	0.530	0.361	3167	0.327	0.394	0.510	2712	0.478	0.540	0.303	1976	0.258	0.347	0.394	2880	0.359	0.428
		2.7	0.581	654	0.521	0.634	0.480	3167	0.450	0.509	0.613	2712	0.586	0.639	0.447	1976	0.407	0.485	0.537	2880	0.507	0.565
		1.9	0.458	1034	0.402	0.510	0.357	5005	0.330	0.384	0.482	4470	0.457	0.507	0.314	3074	0.278	0.349	0.371	3859	0.340	0.401
15	**1.9**	**2.2**	0.488	1034	0.435	0.538	0.374	5005	0.347	0.400	0.510	4470	0.485	0.534	0.353	3074	0.318	0.387	0.407	3859	0.377	0.436
		2.9	0.582	1034	0.535	0.625	0.502	5005	0.479	0.525	0.614	4470	0.593	0.634	0.470	3074	0.439	0.500	0.562	3859	0.537	0.586
		2.1	0.463	1661	0.420	0.504	0.345	8326	0.324	0.366	0.465	7889	0.445	0.484	0.315	5149	0.287	0.342	0.389	5349	0.363	0.414
	2.1	**2.4**	0.505	1661	0.464	0.544	0.356	8326	0.335	0.377	0.502	7889	0.483	0.520	0.360	5149	0.333	0.386	0.401	5349	0.376	0.426
		3.1	0.587	1661	0.551	0.621	0.504	8326	0.486	0.522	0.616	7889	0.601	0.631	0.511	5149	0.488	0.533	0.575	5349	0.555	0.595
		1.7	0.412	320	0.306	0.509	0.349	1428	0.297	0.399	0.420	1010	0.362	0.475	0.191	892	0.120	0.261	0.378	1710	0.332	0.422
	1.7	**2**	0.445	320	0.341	0.538	0.390	1428	0.340	0.438	0.473	1010	0.418	0.525	0.228	892	0.157	0.296	0.432	1710	0.388	0.474
		2.7	0.604	320	0.520	0.676	0.527	1428	0.484	0.568	0.561	1010	0.512	0.606	0.369	892	0.304	0.431	0.550	1710	0.512	0.586
		1.9	0.454	437	0.367	0.534	0.349	2163	0.307	0.390	0.478	1601	0.435	0.519	0.248	1388	0.192	0.302	0.357	2261	0.316	0.396
20	**1.9**	**2.2**	0.524	437	0.443	0.595	0.387	2163	0.346	0.426	0.509	1601	0.467	0.548	0.259	1388	0.203	0.313	0.385	2261	0.345	0.423
		2.9	0.629	437	0.561	0.688	0.520	2163	0.485	0.553	0.597	1601	0.561	0.631	0.417	1388	0.367	0.464	0.542	2261	0.509	0.574
		2.1	0.515	699	0.451	0.573	0.363	3564	0.331	0.394	0.499	2752	0.467	0.530	0.285	2242	0.242	0.327	0.365	3052	0.330	0.399
	2.1	**2.4**	0.538	699	0.477	0.594	0.391	3564	0.360	0.422	0.525	2752	0.494	0.555	0.324	2242	0.282	0.365	0.395	3052	0.361	0.428
		3.1	0.607	699	0.552	0.657	0.536	3564	0.509	0.562	0.624	2752	0.598	0.649	0.468	2242	0.431	0.503	0.585	3052	0.558	0.610

TABLE A1.3: Alternate Seg parameters used to identify LCRs from DNA sequences in the five model organisms. Correlation coefficients and number of LCR data points are given for each parameter set.

parameters			*S. cerevisiae*				*H. sapiens*				*A. thaliana*				*C. elegans*				*D. melanogaster*			
W	**K1**	**K2**	**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**	
		1	0.279	565	0.192	0.362	0.207	3663	0.172	0.241	0.220	3013	0.182	0.258	0.263	1205	0.203	0.321	0.184	620	0.098	0.267
	1	**1.2**	0.269	565	0.182	0.352	0.257	3663	0.223	0.290	0.238	3013	0.200	0.275	0.331	1205	0.274	0.386	0.113	620	0.026	0.199
		1.3	0.387	565	0.306	0.462	0.405	3663	0.374	0.435	0.414	3013	0.380	0.446	0.423	1205	0.370	0.473	0.230	620	0.145	0.311
		1.3	0.509	3175	0.480	0.537	0.478	12142	0.463	0.493	0.461	12706	0.446	0.476	0.469	8163	0.450	0.488	0.307	3844	0.275	0.339
21	**1.3**	**1.5**	0.572	3175	0.545	0.598	0.557	12142	0.543	0.571	0.541	12706	0.527	0.555	0.521	8163	0.503	0.538	0.371	3844	0.340	0.401
		1.6	0.626	3175	0.602	0.649	0.622	12142	0.610	0.634	0.621	12706	0.609	0.633	0.588	8163	0.572	0.604	0.430	3844	0.401	0.458
		1.7	0.626	5982	0.609	0.643	0.606	19868	0.596	0.616	0.606	27407	0.598	0.614	0.578	19847	0.568	0.588	0.555	13806	0.542	0.568
	1.7	**1.9**	0.631	5982	0.614	0.648	0.602	19868	0.592	0.612	0.634	27407	0.626	0.642	0.614	19847	0.604	0.624	0.591	13806	0.579	0.603
		2	0.773	5982	0.761	0.784	0.740	19868	0.733	0.747	0.814	27407	0.809	0.818	0.773	19847	0.767	0.779	0.807	13806	0.800	0.813
		1	0.101	23	-0.369	0.530	-0.190	144	-0.360	-0.008	0.354	53	0.061	0.591	0.336	28	-0.087	0.656	0.669	12	0.081	0.912
	1	**1.2**	0.094	23	-0.375	0.525	0.111	144	-0.072	0.287	0.439	53	0.161	0.653	0.683	28	0.378	0.854	0.526	12	-0.143	0.865
		1.3	0.294	23	-0.183	0.659	0.305	144	0.130	0.461	0.460	53	0.186	0.668	0.796	28	0.572	0.909	0.512	12	-0.161	0.860
		1.3	0.379	171	0.226	0.513	0.433	1571	0.387	0.477	0.459	676	0.390	0.523	0.526	314	0.431	0.610	0.232	182	0.073	0.380
45	**1.3**	**1.5**	0.428	171	0.281	0.555	0.579	1571	0.541	0.615	0.571	676	0.512	0.625	0.526	314	0.431	0.610	0.281	182	0.125	0.424
		1.7	0.528	171	0.396	0.639	0.649	1571	0.616	0.680	0.622	676	0.568	0.671	0.590	314	0.503	0.665	0.365	182	0.216	0.497
		1.7	0.623	4362	0.602	0.643	0.606	16102	0.595	0.617	0.617	18815	0.607	0.627	0.541	12953	0.527	0.554	0.443	6721	0.421	0.464
	1.7	**1.9**	0.740	4362	0.725	0.755	0.736	16102	0.728	0.744	0.748	18815	0.741	0.755	0.700	12953	0.690	0.710	0.612	6721	0.595	0.628
		2	0.850	4362	0.841	0.859	0.877	16102	0.873	0.881	0.902	18815	0.899	0.905	0.858	12953	0.853	0.863	0.859	6721	0.852	0.866
		1	0.214	10	-0.543	0.779	-0.156	46	-0.455	0.174	0.788	15	0.410	0.935	0.732	8	-0.044	0.957	0.620	6	-0.490	0.963
	1	**1.2**	0.204	10	-0.550	0.775	0.298	46	-0.026	0.565	0.594	15	0.053	0.865	0.596	8	-0.282	0.931	0.625	6	-0.484	0.964
		1.3	0.381	10	-0.401	0.842	0.242	46	-0.086	0.523	0.707	15	0.245	0.907	0.297	8	-0.586	0.857	0.746	6	-0.289	0.977
		1.3	0.364	55	0.078	0.594	0.352	607	0.272	0.427	0.531	195	0.409	0.635	0.490	108	0.312	0.635	0.363	38	0.011	0.635
60	**1.3**	**1.5**	0.434	55	0.160	0.646	0.573	607	0.510	0.630	0.584	195	0.471	0.678	0.607	108	0.455	0.725	0.227	38	-0.137	0.537
		1.7	0.570	55	0.332	0.740	0.673	607	0.621	0.719	0.655	195	0.556	0.736	0.529	108	0.359	0.665	0.316	38	-0.042	0.602
		1.7	0.634	2711	0.608	0.658	0.631	11069	0.618	0.643	0.628	10319	0.615	0.641	0.508	6909	0.488	0.527	0.420	3493	0.389	0.450
	1.7	**1.9**	0.759	2711	0.741	0.776	0.759	11069	0.750	0.768	0.761	10319	0.752	0.770	0.687	6909	0.673	0.701	0.601	3493	0.577	0.624
		2	0.852	2711	0.840	0.863	0.882	11069	0.877	0.887	0.901	10319	0.897	0.905	0.851	6909	0.844	0.858	0.853	3493	0.843	0.863

TABLE A1.4: Alternate Seg parameters used to identify LCRs from protein (a) and DNA (b) sequences in the Null simulated proteomes. Correlation coefficients, number of LCR data points, and 95% confidence intervals are given for each parameter set.

(A)

parameters			Null			
W	**K1**	**K2**	**r**	**n**	**95% CI**	
		1.7	0.162	203	0.009	0.308
	1.7	**2**	0.222	203	0.071	0.363
		2.7	0.485	203	0.358	0.594
		1.9	0.319	1587	0.269	0.367
12	**1.9**	**2.2**	0.351	1587	0.302	0.398
		2.9	0.533	1587	0.493	0.571
		2.1	0.318	7303	0.295	0.341
	2.1	**2.4**	0.369	7303	0.347	0.391
		3.1	0.597	7303	0.580	0.613
		1.7	0.652	18	0.211	0.872
	1.7	**2**	0.658	18	0.221	0.875
		2.7	0.188	18	-0.357	0.638
		1.9	0.217	70	-0.046	0.452
15	**1.9**	**2.2**	0.126	70	-0.139	0.374
		2.9	0.323	70	0.068	0.538
		2.1	0.211	638	0.127	0.292
	2.1	**2.4**	0.255	638	0.172	0.334
		3.1	0.447	638	0.375	0.514
		1.7	NA	0	NA	NA
	1.7	**2**	NA	0	NA	NA
		2.7	NA	0	NA	NA
		1.9	-1.000	2	NaN	NaN
20	**1.9**	**2.2**	-1.000	2	NaN	NaN
		2.9	1.000	2	NaN	NaN
		2.1	0.047	17	-0.491	0.559
	2.1	**2.4**	0.048	17	-0.490	0.559
		3.1	0.294	17	-0.274	0.710

(B)

parameters			Null			
W	**K1**	**K2**	**r**	**n**	**95% CI**	
		1	0.200	64	-0.077	0.448
	1	**1.2**	0.308	64	0.039	0.536
		1.3	0.473	64	0.230	0.660
		1.3	0.606	2211	0.576	0.635
21	**1.3**	**1.5**	0.675	2211	0.649	0.700
		1.6	0.704	2211	0.680	0.727
		1.7	0.649	92462	0.645	0.653
	1.7	**1.9**	0.728	92462	0.725	0.731
		2	0.868	92462	0.866	0.870
		1	NA	0	NA	NA
	1	**1.2**	NA	0	NA	NA
		1.3	NA	0	NA	NA
		1.3	NA	1	NA	NA
45	**1.3**	**1.5**	NA	1	NA	NA
		1.6	NA	1	NA	NA
		1.7	0.645	2853	0.620	0.668
	1.7	**1.9**	0.733	2853	0.713	0.751
		2	0.825	2853	0.811	0.838
		1	NA	0	NA	NA
	1	**1.2**	NA	0	NA	NA
		1.3	NA	0	NA	NA
		1.3	NA	0	NA	NA
60	**1.3**	**1.5**	NA	0	NA	NA
		1.6	NA	0	NA	NA
		1.7	0.660	159	0.550	0.748
	1.7	**1.9**	0.666	159	0.557	0.752
		2	0.735	159	0.644	0.806

TABLE A1.5: Alternate Seg parameters used to identify LCRs from species specific Slip simulated protein sequences. Correlation coefficients, number of LCR data points, and 95% confidence intervals are given for each parameter set.

parameters			*S. cerevisiae*				*H. sapiens*				*A. thaliana*				*C. elegans*				*D. melanogaster*			
W	**K1**	**K2**	**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**	
		1.7	0.543	24346	0.533	0.553	0.639	7042	0.623	0.654	0.578	14052	0.566	0.590	0.536	15094	0.523	0.549	0.582	18911	0.571	0.592
	1.7	**2**	0.574	24346	0.565	0.583	0.658	7042	0.643	0.673	0.605	14052	0.593	0.617	0.562	15094	0.550	0.574	0.610	18911	0.600	0.620
		2.7	0.629	24346	0.620	0.637	0.690	7042	0.676	0.703	0.650	14052	0.639	0.661	0.608	15094	0.597	0.619	0.652	18911	0.643	0.661
		1.9	0.661	37295	0.655	0.667	0.726	17604	0.718	0.734	0.691	24174	0.684	0.698	0.655	24552	0.647	0.663	0.688	30851	0.681	0.694
12	**1.9**	**2.2**	0.680	37295	0.674	0.686	0.738	17604	0.730	0.745	0.708	24174	0.701	0.715	0.673	24552	0.665	0.681	0.706	30851	0.700	0.712
		2.9	0.728	37295	0.723	0.733	0.759	17604	0.752	0.766	0.742	24174	0.736	0.748	0.712	24552	0.705	0.719	0.743	30851	0.737	0.749
		2.1	0.743	57393	0.739	0.747	0.749	41835	0.744	0.754	0.763	43014	0.759	0.767	0.743	40865	0.738	0.748	0.765	50823	0.761	0.769
	2.1	**2.4**	0.756	57393	0.752	0.760	0.756	41835	0.751	0.761	0.774	43014	0.770	0.778	0.755	40865	0.750	0.760	0.778	50823	0.774	0.782
		3.1	0.800	57393	0.797	0.803	0.769	41835	0.765	0.773	0.800	43014	0.796	0.804	0.787	40865	0.783	0.791	0.807	50823	0.804	0.810
		1.7	0.459	12507	0.443	0.474	0.535	2062	0.500	0.568	0.484	6281	0.463	0.505	0.450	7002	0.429	0.471	0.498	8847	0.480	0.515
	1.7	**2**	0.484	12507	0.469	0.499	0.552	2062	0.518	0.585	0.504	6281	0.483	0.524	0.472	7002	0.451	0.492	0.518	8847	0.501	0.535
		2.7	0.530	12507	0.516	0.544	0.600	2062	0.568	0.630	0.548	6281	0.528	0.567	0.514	7002	0.495	0.533	0.560	8847	0.544	0.576
		1.9	0.539	17239	0.527	0.551	0.638	3765	0.616	0.659	0.569	9274	0.553	0.584	0.520	10087	0.504	0.536	0.570	12577	0.557	0.583
15	**1.9**	**2.2**	0.566	17239	0.555	0.577	0.658	3765	0.637	0.678	0.594	9274	0.579	0.608	0.547	10087	0.532	0.562	0.596	12577	0.583	0.608
		2.9	0.618	17239	0.608	0.628	0.689	3765	0.670	0.707	0.638	9274	0.624	0.651	0.594	10087	0.580	0.608	0.637	12577	0.625	0.648
		2.1	0.654	24896	0.646	0.662	0.739	8510	0.728	0.750	0.688	14729	0.678	0.697	0.639	15109	0.628	0.649	0.678	19109	0.669	0.686
	2.1	**2.4**	0.674	24896	0.666	0.681	0.752	8510	0.742	0.762	0.704	14729	0.695	0.713	0.657	15109	0.647	0.667	0.696	19109	0.688	0.704
		3.1	0.720	24896	0.713	0.727	0.769	8510	0.759	0.779	0.737	14729	0.729	0.745	0.697	15109	0.688	0.706	0.733	19109	0.726	0.740
		1.7	0.343	4821	0.315	0.370	0.393	369	0.292	0.485	0.365	2100	0.323	0.406	0.342	2424	0.302	0.381	0.377	2988	0.342	0.411
	1.7	**2**	0.363	4821	0.335	0.390	0.417	369	0.318	0.507	0.379	2100	0.337	0.419	0.357	2424	0.318	0.395	0.397	2988	0.363	0.430
		2.7	0.397	4821	0.370	0.423	0.461	369	0.367	0.546	0.412	2100	0.372	0.451	0.383	2424	0.344	0.420	0.425	2988	0.392	0.457
		1.9	0.413	6253	0.390	0.436	0.474	600	0.402	0.540	0.429	2773	0.395	0.462	0.408	3212	0.375	0.440	0.437	4030	0.409	0.464
20	**1.9**	**2.2**	0.435	6253	0.412	0.457	0.488	600	0.417	0.553	0.449	2773	0.415	0.482	0.426	3212	0.394	0.457	0.454	4030	0.426	0.481
		2.9	0.472	6253	0.450	0.493	0.531	600	0.464	0.592	0.485	2773	0.453	0.516	0.449	3212	0.418	0.479	0.491	4030	0.464	0.517
		2.1	0.496	8816	0.478	0.513	0.592	1131	0.548	0.633	0.525	4191	0.500	0.549	0.488	4649	0.463	0.512	0.532	5869	0.511	0.552
	2.1	**2.4**	0.523	8816	0.506	0.540	0.609	1131	0.566	0.648	0.548	4191	0.524	0.571	0.504	4649	0.480	0.528	0.554	5869	0.534	0.573
		3.1	0.568	8816	0.552	0.584	0.643	1131	0.603	0.680	0.588	4191	0.565	0.610	0.541	4649	0.518	0.563	0.589	5869	0.570	0.607

TABLE A1.6: Alternate `Seg` parameters used to identify LCRs from species specific Slip simulated DNA sequences. Correlation coefficients, number of LCR data points, and 95% confidence intervals are given for each parameter set.

parameters			*S. cerevisiae*				*H. sapiens*				*A. thaliana*				*C. elegans*				*D. melanogaster*			
W	**K1**	**K2**	**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**	
		1	0.311	30077	0.300	0.322	0.392	8091	0.371	0.412	0.286	15254	0.270	0.302	0.286	17855	0.271	0.301	0.195	15938	0.178	0.212
	1	**1.2**	0.375	30077	0.364	0.386	0.428	8091	0.408	0.448	0.329	15254	0.313	0.345	0.340	17855	0.325	0.354	0.244	15938	0.228	0.260
		1.3	0.454	30077	0.444	0.464	0.488	8091	0.469	0.506	0.400	15254	0.385	0.415	0.414	17855	0.400	0.427	0.337	15938	0.322	0.352
		1.3	0.685	71346	0.681	0.689	0.709	35126	0.703	0.715	0.662	42112	0.656	0.668	0.671	47726	0.665	0.676	0.605	36893	0.598	0.612
21	**1.3**	**1.5**	0.748	71346	0.744	0.752	0.759	35126	0.754	0.764	0.721	42112	0.716	0.726	0.728	47726	0.723	0.733	0.684	36893	0.678	0.690
		1.6	0.781	71346	0.778	0.784	0.783	35126	0.778	0.787	0.750	42112	0.745	0.755	0.758	47726	0.754	0.762	0.720	36893	0.714	0.725
		1.7	0.748	100000	0.745	0.751	0.638	100000	0.634	0.642	0.719	100000	0.716	0.722	0.734	100000	0.731	0.737	0.732	100000	0.729	0.735
	1.7	**1.9**	0.743	100000	0.740	0.746	0.602	100000	0.598	0.606	0.709	100000	0.706	0.712	0.724	100000	0.721	0.727	0.731	100000	0.728	0.734
		2	0.289	100000	0.283	0.295	0.630	100000	0.626	0.634	0.519	100000	0.514	0.524	0.436	100000	0.430	0.442	0.595	100000	0.591	0.599
		1	0.193	3187	0.155	0.230	0.313	154	0.145	0.463	0.179	1092	0.114	0.242	0.205	1434	0.149	0.260	0.144	1352	0.085	0.202
	1	**1.2**	0.301	3187	0.265	0.336	0.351	154	0.187	0.496	0.271	1092	0.209	0.331	0.289	1434	0.235	0.341	0.237	1352	0.180	0.292
		1.3	0.342	3187	0.307	0.376	0.351	154	0.187	0.496	0.301	1092	0.240	0.360	0.312	1434	0.259	0.363	0.261	1352	0.205	0.316
		1.3	0.537	6615	0.518	0.556	0.533	488	0.458	0.600	0.453	2290	0.416	0.489	0.473	3034	0.442	0.503	0.395	2664	0.359	0.430
45	**1.3**	**1.5**	0.591	6615	0.573	0.608	0.573	488	0.503	0.636	0.492	2290	0.457	0.526	0.514	3034	0.484	0.543	0.427	2664	0.392	0.461
		1.6	0.610	6615	0.593	0.627	0.580	488	0.510	0.642	0.505	2290	0.470	0.538	0.529	3034	0.500	0.557	0.437	2664	0.402	0.471
		1.7	0.837	85352	0.835	0.839	0.868	37583	0.865	0.871	0.850	44526	0.847	0.853	0.855	52501	0.852	0.858	0.835	39088	0.832	0.838
	1.7	**1.9**	0.884	85352	0.882	0.886	0.889	37583	0.887	0.891	0.876	44526	0.874	0.878	0.883	52501	0.881	0.885	0.864	39088	0.861	0.867
		2	0.935	85352	0.934	0.936	0.945	37583	0.944	0.946	0.944	44526	0.943	0.945	0.946	52501	0.945	0.947	0.938	39088	0.937	0.939
		1	0.200	1419	0.144	0.255	0.352	35	-0.018	0.637	0.225	439	0.124	0.322	0.183	565	0.093	0.270	0.198	561	0.108	0.285
	1	**1.2**	0.288	1419	0.234	0.340	0.428	35	0.071	0.688	0.295	439	0.197	0.387	0.258	565	0.170	0.342	0.265	561	0.177	0.349
		1.3	0.315	1419	0.262	0.366	0.428	35	0.071	0.688	0.311	439	0.214	0.402	0.258	565	0.170	0.342	0.272	561	0.184	0.355
		1.3	0.456	2773	0.423	0.488	0.508	113	0.338	0.646	0.380	878	0.315	0.441	0.406	1135	0.350	0.459	0.331	1041	0.269	0.390
60	**1.3**	**1.5**	0.490	2773	0.458	0.521	0.531	113	0.366	0.664	0.391	878	0.327	0.452	0.430	1135	0.376	0.481	0.359	1041	0.299	0.417
		1.6	0.502	2773	0.470	0.532	0.531	113	0.366	0.664	0.402	878	0.338	0.462	0.444	1135	0.390	0.495	0.361	1041	0.301	0.418
		1.7	0.875	42378	0.872	0.877	0.887	5849	0.881	0.893	0.840	11511	0.834	0.846	0.856	16008	0.851	0.861	0.794	9491	0.786	0.802
	1.7	**1.9**	0.901	42378	0.899	0.903	0.899	5849	0.893	0.904	0.859	11511	0.854	0.864	0.875	16008	0.871	0.879	0.820	9491	0.813	0.827
		2	0.947	42378	0.946	0.948	0.935	5849	0.931	0.938	0.925	11511	0.922	0.928	0.939	16008	0.937	0.941	0.901	9491	0.897	0.905

TABLE A1.7: Alternate Seg parameters used to identify LCRs from species specific Slip+Syn simulated protein sequences. Correlation coefficients, number of LCR data points, and 95% confidence intervals are given for each parameter set.

parameters			*S. cerevisiae*				*H. sapiens*				*A. thaliana*				*C. elegans*				*D. melanogaster*			
W	**K1**	**K2**	**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**	
		1.7	0.431	24346	0.42	0.442	0.548	7042	0.53	0.566	0.452	14052	0.437	0.467	0.438	15094	0.424	0.452	0.462	18911	0.449	0.474
	1.7	**2**	0.465	24346	0.454	0.476	0.571	7042	0.553	0.588	0.483	14052	0.469	0.497	0.467	15094	0.453	0.481	0.492	18911	0.48	0.504
		2.7	0.531	24346	0.521	0.541	0.618	7042	0.602	0.634	0.54	14052	0.527	0.553	0.522	15094	0.509	0.535	0.544	18911	0.533	0.555
		1.9	0.555	37295	0.547	0.563	0.65	17604	0.64	0.659	0.573	24174	0.563	0.582	0.562	24552	0.552	0.571	0.579	30851	0.571	0.587
12	**1.9**	**2.2**	0.58	37295	0.572	0.587	0.667	17604	0.658	0.676	0.595	24174	0.586	0.604	0.584	24552	0.575	0.593	0.6	30851	0.592	0.608
		2.9	0.647	37295	0.64	0.654	0.706	17604	0.698	0.714	0.652	24174	0.644	0.66	0.639	24552	0.631	0.647	0.656	30851	0.649	0.663
		2.1	0.653	57393	0.648	0.658	0.683	41835	0.677	0.689	0.665	43014	0.659	0.671	0.661	40865	0.655	0.667	0.67	50823	0.665	0.675
	2.1	**2.4**	0.673	57393	0.668	0.678	0.698	41835	0.692	0.703	0.683	43014	0.677	0.689	0.68	40865	0.674	0.686	0.69	50823	0.685	0.695
		3.1	0.749	57393	0.745	0.753	0.742	41835	0.737	0.747	0.744	43014	0.739	0.749	0.742	40865	0.737	0.747	0.75	50823	0.746	0.754
		1.7	0.357	12507	0.34	0.374	0.462	2062	0.423	0.499	0.369	6281	0.345	0.393	0.358	7002	0.335	0.381	0.38	8847	0.36	0.4
	1.7	**2**	0.381	12507	0.364	0.398	0.48	2062	0.442	0.516	0.388	6281	0.364	0.411	0.381	7002	0.358	0.403	0.399	8847	0.379	0.418
		2.7	0.43	12507	0.414	0.446	0.536	2062	0.501	0.569	0.435	6281	0.412	0.457	0.422	7002	0.4	0.443	0.445	8847	0.426	0.463
		1.9	0.431	17239	0.417	0.444	0.557	3765	0.532	0.581	0.448	9274	0.43	0.466	0.427	10087	0.409	0.445	0.451	12577	0.435	0.466
15	**1.9**	**2.2**	0.459	17239	0.446	0.472	0.585	3765	0.561	0.608	0.475	9274	0.457	0.492	0.457	10087	0.44	0.474	0.48	12577	0.465	0.495
		2.9	0.519	17239	0.507	0.531	0.628	3765	0.606	0.649	0.528	9274	0.511	0.544	0.507	10087	0.491	0.523	0.529	12577	0.515	0.543
		2.1	0.548	24896	0.538	0.558	0.667	8510	0.654	0.68	0.568	14729	0.556	0.58	0.545	15109	0.532	0.557	0.567	19109	0.556	0.578
	2.1	**2.4**	0.572	24896	0.563	0.581	0.685	8510	0.672	0.697	0.591	14729	0.579	0.603	0.566	15109	0.554	0.578	0.588	19109	0.578	0.598
		3.1	0.635	24896	0.627	0.643	0.719	8510	0.707	0.73	0.641	14729	0.63	0.651	0.619	15109	0.608	0.63	0.64	19109	0.631	0.649
		1.7	0.26	4821	0.23	0.289	0.301	369	0.194	0.401	0.27	2100	0.225	0.314	0.259	2424	0.217	0.3	0.283	2988	0.246	0.319
	1.7	**2**	0.277	4821	0.248	0.306	0.326	369	0.22	0.424	0.284	2100	0.24	0.327	0.272	2424	0.23	0.313	0.303	2988	0.266	0.339
		2.7	0.31	4821	0.281	0.338	0.374	369	0.272	0.468	0.314	2100	0.27	0.356	0.298	2424	0.257	0.338	0.328	2988	0.292	0.363
		1.9	0.323	6253	0.298	0.348	0.393	600	0.315	0.466	0.327	2773	0.289	0.364	0.321	3212	0.286	0.355	0.335	4030	0.304	0.365
20	**1.9**	**2.2**	0.343	6253	0.318	0.367	0.408	600	0.331	0.48	0.347	2773	0.31	0.383	0.337	3212	0.302	0.371	0.355	4030	0.325	0.385
		2.9	0.38	6253	0.356	0.403	0.453	600	0.379	0.521	0.38	2773	0.344	0.415	0.362	3212	0.328	0.395	0.388	4030	0.358	0.417
		2.1	0.393	8816	0.373	0.412	0.523	1131	0.474	0.569	0.411	4191	0.383	0.439	0.397	4649	0.37	0.424	0.416	5869	0.392	0.439
	2.1	**2.4**	0.421	8816	0.402	0.44	0.542	1131	0.494	0.586	0.436	4191	0.408	0.463	0.413	4649	0.386	0.439	0.44	5869	0.417	0.463
		3.1	0.469	8816	0.451	0.487	0.583	1131	0.538	0.624	0.478	4191	0.452	0.504	0.454	4649	0.428	0.479	0.479	5869	0.457	0.501

TABLE A1.8: Alternate Seg parameters used to identify LCRs from species specific Slip+Syn simulated DNA sequences. Correlation coefficients, number of LCR data points, and 95% confidence intervals are given for each parameter set.

parameters			*S. cerevisiae*				*H. sapiens*				*A. thaliana*				*C. elegans*				*D. melanogaster*			
W	**K1**	**K2**	**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**	
		1	0.351	15258	0.335	0.366	0.371	4586	0.343	0.399	0.336	8132	0.314	0.357	0.317	9039	0.296	0.338	0.279	9185	0.258	0.300
	1	**1.2**	0.398	15258	0.383	0.413	0.399	4586	0.371	0.426	0.369	8132	0.348	0.390	0.350	9039	0.330	0.370	0.309	9185	0.288	0.329
		1.3	0.473	15258	0.459	0.487	0.487	4586	0.462	0.512	0.454	8132	0.435	0.473	0.435	9039	0.416	0.453	0.410	9185	0.391	0.429
		1.3	0.602	51110	0.596	0.608	0.651	29814	0.644	0.658	0.617	31451	0.609	0.625	0.602	34230	0.594	0.609	0.578	34035	0.570	0.586
21	**1.3**	**1.5**	0.664	51110	0.659	0.669	0.707	29814	0.700	0.713	0.672	31451	0.665	0.679	0.657	34230	0.650	0.664	0.638	34035	0.631	0.645
		1.6	0.704	51110	0.699	0.709	0.740	29814	0.734	0.745	0.707	31451	0.701	0.713	0.693	34230	0.687	0.699	0.677	34035	0.671	0.683
		1.7	0.713	100000	0.710	0.716	0.609	100000	0.604	0.613	0.675	100000	0.671	0.679	0.695	100000	0.691	0.699	0.677	100000	0.673	0.681
	1.7	**1.9**	0.722	100000	0.719	0.725	0.578	100000	0.573	0.583	0.675	100000	0.671	0.679	0.705	100000	0.702	0.708	0.676	100000	0.672	0.680
		2	0.596	100000	0.592	0.600	0.580	100000	0.575	0.584	0.669	100000	0.665	0.673	0.695	100000	0.691	0.699	0.619	100000	0.615	0.623
		1	0.245	1338	0.188	0.300	0.303	73	0.052	0.518	0.200	497	0.104	0.292	0.197	610	0.110	0.281	0.135	624	0.048	0.220
	1	**1.2**	0.340	1338	0.286	0.392	0.365	73	0.121	0.567	0.293	497	0.201	0.380	0.330	610	0.249	0.407	0.261	624	0.178	0.341
		1.3	0.391	1338	0.339	0.440	0.446	73	0.215	0.630	0.352	497	0.263	0.435	0.371	610	0.292	0.445	0.352	624	0.273	0.426
		1.3	0.419	3315	0.387	0.450	0.496	260	0.387	0.592	0.428	1255	0.376	0.477	0.368	1564	0.319	0.415	0.388	1553	0.340	0.434
45	**1.3**	**1.5**	0.517	3315	0.489	0.544	0.588	260	0.492	0.670	0.513	1255	0.466	0.557	0.445	1564	0.400	0.488	0.475	1553	0.431	0.517
		1.6	0.558	3315	0.531	0.584	0.605	260	0.511	0.684	0.549	1255	0.504	0.591	0.474	1564	0.430	0.516	0.512	1553	0.470	0.552
		1.7	0.812	58431	0.809	0.815	0.834	33600	0.830	0.837	0.824	31815	0.820	0.828	0.824	35489	0.820	0.828	0.800	36624	0.796	0.804
	1.7	**1.9**	0.857	58431	0.855	0.859	0.870	33600	0.867	0.873	0.856	31815	0.853	0.859	0.858	35489	0.855	0.861	0.841	36624	0.838	0.844
		2	0.948	58431	0.947	0.949	0.951	33600	0.950	0.952	0.933	31815	0.931	0.935	0.939	35489	0.938	0.940	0.935	36624	0.934	0.936
		1	0.176	602	0.088	0.261	0.125	17	-0.428	0.611	0.161	204	0.008	0.306	0.158	262	0.024	0.287	0.129	260	-0.007	0.260
	1	**1.2**	0.274	602	0.190	0.354	0.405	17	-0.153	0.767	0.214	204	0.063	0.355	0.317	262	0.190	0.433	0.280	260	0.150	0.400
		1.3	0.303	602	0.220	0.382	0.405	17	-0.153	0.767	0.211	204	0.060	0.353	0.347	262	0.222	0.460	0.312	260	0.184	0.429
		1.3	0.351	1355	0.298	0.402	0.491	57	0.236	0.683	0.305	480	0.212	0.393	0.305	615	0.223	0.383	0.328	600	0.246	0.405
60	**1.3**	**1.5**	0.455	1355	0.407	0.501	0.582	57	0.352	0.746	0.421	480	0.335	0.500	0.400	615	0.323	0.471	0.421	600	0.345	0.492
		1.6	0.471	1355	0.423	0.516	0.582	57	0.352	0.746	0.432	480	0.347	0.510	0.416	615	0.340	0.486	0.446	600	0.372	0.515
		1.7	0.833	18986	0.828	0.838	0.878	4791	0.871	0.885	0.823	6952	0.814	0.831	0.814	8280	0.806	0.822	0.793	8096	0.784	0.802
	1.7	**1.9**	0.863	18986	0.859	0.867	0.897	4791	0.891	0.903	0.848	6952	0.840	0.855	0.842	8280	0.835	0.849	0.823	8096	0.815	0.831
		2	0.944	18986	0.942	0.946	0.942	4791	0.939	0.946	0.902	6952	0.897	0.907	0.906	8280	0.902	0.910	0.900	8096	0.895	0.905

TABLE A1.9: Alternate minimum repeat lengths for protein LCRs containing periodic amino acid repeats. Correlation coefficients, number of LCR data points, and 95% confidence intervals are given for each parameter set.

Repeat Lengths			*S. cerevisiae*				*H. sapiens*				*A. thaliana*				*C. elegans*				*D. melanogaster*			
mono	di	tri	r	n	95% CI		r	n	95% CI		r	n	95% CI		r	n	95% CI		r	n	95% CI	
		3	0.524	485	0.448	0.592	0.422	2396	0.385	0.458	0.566	2457	0.535	0.595	0.476	1363	0.429	0.521	0.485	2171	0.448	0.520
	3	4	0.534	470	0.458	0.602	0.421	2364	0.383	0.457	0.569	2436	0.538	0.598	0.480	1283	0.432	0.526	0.486	2136	0.449	0.521
		5	0.534	466	0.458	0.603	0.422	2349	0.384	0.458	0.569	2435	0.538	0.598	0.481	1269	0.432	0.527	0.487	2116	0.450	0.522
		3	0.539	437	0.460	0.609	0.425	2214	0.386	0.462	0.570	2302	0.538	0.600	0.494	1267	0.446	0.539	0.492	2007	0.454	0.528
5	4	4	0.550	421	0.471	0.620	0.423	2180	0.384	0.461	0.573	2280	0.541	0.603	0.499	1186	0.450	0.545	0.493	1970	0.455	0.529
		5	0.551	417	0.472	0.621	0.424	2165	0.385	0.462	0.573	2279	0.541	0.603	0.501	1171	0.452	0.547	0.494	1949	0.456	0.531
		3	0.549	423	0.470	0.619	0.426	2140	0.387	0.464	0.570	2250	0.538	0.600	0.505	1230	0.457	0.550	0.499	1945	0.461	0.535
	5	4	0.561	407	0.482	0.631	0.424	2105	0.384	0.462	0.573	2228	0.541	0.603	0.511	1148	0.462	0.557	0.499	1908	0.460	0.536
		5	0.562	403	0.483	0.632	0.425	2090	0.385	0.463	0.573	2227	0.541	0.603	0.514	1133	0.465	0.560	0.501	1887	0.462	0.538
		3	0.532	379	0.446	0.608	0.445	1847	0.403	0.485	0.595	1842	0.561	0.627	0.520	991	0.467	0.569	0.505	1715	0.465	0.543
	3	4	0.545	364	0.459	0.621	0.443	1813	0.401	0.483	0.600	1817	0.566	0.632	0.530	906	0.476	0.580	0.506	1678	0.465	0.545
		5	0.546	360	0.460	0.622	0.444	1798	0.402	0.484	0.599	1815	0.565	0.631	0.531	891	0.476	0.582	0.508	1658	0.467	0.547
		3	0.544	327	0.453	0.624	0.451	1641	0.407	0.493	0.599	1659	0.563	0.632	0.547	877	0.493	0.597	0.517	1526	0.475	0.557
6	4	4	0.560	311	0.469	0.639	0.448	1604	0.403	0.491	0.604	1633	0.569	0.637	0.559	791	0.503	0.610	0.517	1487	0.474	0.557
		5	0.561	307	0.469	0.641	0.449	1589	0.404	0.492	0.603	1631	0.567	0.636	0.562	775	0.506	0.613	0.520	1466	0.477	0.560
		3	0.556	313	0.464	0.636	0.451	1562	0.406	0.494	0.598	1600	0.562	0.632	0.566	837	0.512	0.615	0.525	1453	0.482	0.565
	5	4	0.573	297	0.481	0.652	0.448	1524	0.402	0.492	0.603	1574	0.567	0.637	0.579	750	0.523	0.630	0.525	1413	0.482	0.566
		5	0.574	293	0.482	0.654	0.449	1509	0.403	0.493	0.603	1572	0.567	0.637	0.584	734	0.528	0.635	0.527	1392	0.483	0.568
		3	0.532	379	0.446	0.608	0.445	1847	0.403	0.485	0.595	1842	0.561	0.627	0.520	991	0.467	0.569	0.505	1715	0.465	0.543
	3	4	0.545	364	0.459	0.621	0.443	1813	0.401	0.483	0.600	1817	0.566	0.632	0.530	906	0.476	0.580	0.506	1678	0.465	0.545
		5	0.546	360	0.460	0.622	0.444	1798	0.402	0.484	0.599	1815	0.565	0.631	0.531	891	0.476	0.582	0.508	1658	0.467	0.547
		3	0.539	265	0.436	0.628	0.468	1209	0.417	0.516	0.626	1129	0.585	0.664	0.566	627	0.504	0.623	0.556	1186	0.511	0.598
7	4	4	0.557	248	0.453	0.646	0.464	1168	0.412	0.513	0.632	1100	0.591	0.670	0.584	540	0.518	0.643	0.560	1139	0.514	0.603
		5	0.558	244	0.454	0.647	0.465	1153	0.413	0.514	0.632	1098	0.591	0.670	0.588	523	0.522	0.647	0.563	1118	0.517	0.606
		3	0.550	250	0.446	0.640	0.467	1128	0.415	0.516	0.626	1068	0.584	0.665	0.592	581	0.530	0.648	0.566	1105	0.520	0.609
	5	4	0.568	233	0.463	0.658	0.463	1085	0.409	0.514	0.632	1039	0.589	0.671	0.612	493	0.547	0.670	0.570	1057	0.523	0.614
		5	0.570	229	0.464	0.660	0.464	1070	0.410	0.515	0.632	1037	0.589	0.671	0.618	476	0.552	0.676	0.573	1036	0.526	0.617
		3	0.532	379	0.446	0.608	0.445	1847	0.403	0.485	0.595	1842	0.561	0.627	0.520	991	0.467	0.569	0.505	1715	0.465	0.543
	3	4	0.545	364	0.459	0.621	0.443	1813	0.401	0.483	0.600	1817	0.566	0.632	0.530	906	0.476	0.580	0.506	1678	0.465	0.545
		5	0.546	360	0.460	0.622	0.444	1798	0.402	0.484	0.599	1815	0.565	0.631	0.531	891	0.476	0.582	0.508	1658	0.467	0.547
		3	0.519	214	0.401	0.620	0.452	884	0.392	0.509	0.646	771	0.598	0.690	0.566	474	0.494	0.631	0.562	918	0.511	0.609
10	4	4	0.539	197	0.419	0.641	0.445	841	0.382	0.503	0.654	742	0.606	0.698	0.589	385	0.511	0.657	0.568	868	0.516	0.616
		5	0.541	193	0.419	0.643	0.447	826	0.384	0.506	0.653	740	0.604	0.697	0.593	367	0.514	0.662	0.572	845	0.519	0.620
		3	0.504	161	0.363	0.622	0.486	594	0.414	0.552	0.646	454	0.582	0.702	0.543	315	0.450	0.624	0.549	637	0.486	0.607
	5	4	0.518	120	0.355	0.650	0.463	420	0.375	0.543	0.678	298	0.603	0.741	0.580	164	0.454	0.683	0.564	447	0.489	0.631
		5	0.521	116	0.356	0.655	0.464	405	0.374	0.545	0.676	296	0.601	0.739	0.594	146	0.463	0.700	0.572	423	0.496	0.639

TABLE A1.10: Alternate repeat lengths for cryptic repeat protein LCRs defined as LCRs which do not contain sequences with the minimal periodic amino acid repeat length. Correlation coefficients, number of LCR data points, and 95% confidence intervals are given for each parameter set.

Repeat Lengths			***S. cerevisiae***				***H. sapiens***				***A. thaliana***				***C. elegans***				***D. melanogaster***			
mono	**di**	**tri**	**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**		**r**	**n**	**95% CI**	
		3	0.325	549	0.239	0.406	0.233	2609	0.192	0.273	0.272	2013	0.226	0.316	0.210	1711	0.159	0.260	0.224	1688	0.173	0.274
	3	**4**	0.328	564	0.243	0.408	0.235	2641	0.194	0.275	0.266	2034	0.220	0.310	0.212	1791	0.162	0.261	0.234	1723	0.184	0.283
		5	0.332	568	0.248	0.411	0.233	2656	0.193	0.273	0.266	2035	0.220	0.310	0.213	1805	0.163	0.262	0.231	1743	0.181	0.280
		3	0.314	597	0.231	0.392	0.239	2791	0.200	0.278	0.273	2168	0.229	0.316	0.205	1807	0.155	0.254	0.232	1852	0.183	0.279
5	**4**	**4**	0.316	613	0.234	0.393	0.242	2825	0.203	0.280	0.267	2190	0.223	0.310	0.207	1888	0.158	0.255	0.241	1889	0.193	0.288
		5	0.320	617	0.239	0.397	0.240	2840	0.201	0.278	0.267	2191	0.223	0.310	0.207	1903	0.159	0.254	0.239	1910	0.191	0.286
		3	0.302	611	0.219	0.380	0.238	2865	0.199	0.276	0.285	2220	0.242	0.327	0.201	1844	0.152	0.249	0.224	1914	0.176	0.271
	5	**4**	0.304	627	0.223	0.381	0.241	2900	0.202	0.279	0.279	2242	0.236	0.321	0.203	1926	0.155	0.250	0.233	1951	0.186	0.279
		5	0.309	631	0.228	0.386	0.239	2915	0.200	0.277	0.279	2243	0.236	0.321	0.203	1941	0.155	0.250	0.231	1972	0.184	0.277
		3	0.339	655	0.261	0.412	0.233	3158	0.196	0.269	0.299	2628	0.260	0.337	0.219	2083	0.173	0.264	0.267	2144	0.223	0.310
	3	**4**	0.340	670	0.263	0.413	0.237	3192	0.200	0.273	0.294	2653	0.255	0.332	0.219	2168	0.174	0.263	0.274	2181	0.230	0.317
		5	0.344	674	0.268	0.416	0.235	3207	0.198	0.271	0.294	2655	0.255	0.332	0.221	2183	0.176	0.265	0.272	2201	0.228	0.315
		3	0.331	707	0.256	0.402	0.240	3364	0.204	0.275	0.298	2811	0.260	0.335	0.215	2197	0.170	0.259	0.270	2333	0.228	0.311
6	**4**	**4**	0.332	723	0.258	0.402	0.243	3401	0.207	0.278	0.293	2837	0.255	0.330	0.215	2283	0.171	0.258	0.276	2372	0.234	0.317
		5	0.335	727	0.261	0.405	0.242	3416	0.206	0.277	0.294	2839	0.256	0.331	0.215	2299	0.171	0.258	0.275	2393	0.233	0.316
		3	0.321	721	0.246	0.392	0.239	3443	0.204	0.274	0.307	2870	0.270	0.343	0.211	2237	0.166	0.255	0.264	2406	0.222	0.305
	5	**4**	0.322	737	0.248	0.392	0.242	3481	0.207	0.277	0.302	2896	0.265	0.338	0.211	2324	0.167	0.254	0.271	2446	0.230	0.311
		5	0.325	741	0.251	0.395	0.241	3496	0.206	0.275	0.303	2898	0.266	0.339	0.211	2340	0.167	0.254	0.269	2467	0.228	0.309
		3	0.339	655	0.261	0.412	0.233	3158	0.196	0.269	0.299	2628	0.260	0.337	0.219	2083	0.173	0.264	0.267	2144	0.223	0.310
	3	**4**	0.340	670	0.263	0.413	0.237	3192	0.200	0.273	0.294	2653	0.255	0.332	0.219	2168	0.174	0.263	0.274	2181	0.230	0.317
		5	0.344	674	0.268	0.416	0.235	3207	0.198	0.271	0.294	2655	0.255	0.332	0.221	2183	0.176	0.265	0.272	2201	0.228	0.315
		3	0.347	769	0.276	0.414	0.260	3796	0.227	0.293	0.336	3341	0.302	0.369	0.258	2447	0.216	0.299	0.283	2673	0.244	0.321
7	**4**	**4**	0.351	786	0.281	0.417	0.264	3837	0.231	0.297	0.332	3370	0.298	0.365	0.257	2534	0.216	0.297	0.287	2720	0.248	0.325
		5	0.353	790	0.283	0.419	0.262	3852	0.229	0.294	0.332	3372	0.298	0.365	0.257	2551	0.216	0.297	0.285	2741	0.246	0.323
		3	0.338	784	0.267	0.405	0.259	3877	0.226	0.291	0.342	3402	0.308	0.375	0.254	2493	0.213	0.294	0.279	2754	0.240	0.317
	5	**4**	0.342	801	0.272	0.408	0.262	3920	0.229	0.294	0.338	3431	0.305	0.371	0.253	2581	0.212	0.293	0.284	2802	0.246	0.322
		5	0.344	805	0.274	0.410	0.261	3935	0.228	0.293	0.339	3433	0.306	0.372	0.253	2598	0.212	0.293	0.282	2823	0.244	0.319
		3	0.339	655	0.261	0.412	0.233	3158	0.196	0.269	0.299	2628	0.260	0.337	0.219	2083	0.173	0.264	0.267	2144	0.223	0.310
	3	**4**	0.340	670	0.263	0.413	0.237	3192	0.200	0.273	0.294	2653	0.255	0.332	0.219	2168	0.174	0.263	0.274	2181	0.230	0.317
		5	0.344	674	0.268	0.416	0.235	3207	0.198	0.271	0.294	2655	0.255	0.332	0.221	2183	0.176	0.265	0.272	2201	0.228	0.315
		3	0.376	820	0.309	0.440	0.312	4121	0.281	0.342	0.368	3699	0.337	0.399	0.285	2600	0.245	0.324	0.328	2941	0.292	0.363
10	**4**	**4**	0.377	837	0.310	0.440	0.314	4164	0.283	0.344	0.364	3728	0.333	0.395	0.283	2689	0.244	0.321	0.330	2991	0.294	0.365
		5	0.379	841	0.313	0.442	0.313	4179	0.282	0.343	0.364	3730	0.333	0.395	0.283	2707	0.244	0.321	0.329	3014	0.293	0.364
		3	0.399	873	0.335	0.459	0.321	4411	0.291	0.350	0.437	4016	0.409	0.464	0.322	2759	0.284	0.359	0.366	3222	0.332	0.399
	5	**4**	0.427	914	0.366	0.484	0.342	4585	0.313	0.370	0.455	4172	0.428	0.481	0.333	2910	0.297	0.369	0.377	3412	0.344	0.409
		5	0.428	918	0.367	0.485	0.341	4600	0.312	0.369	0.455	4174	0.428	0.481	0.333	2928	0.297	0.368	0.376	3436	0.344	0.408

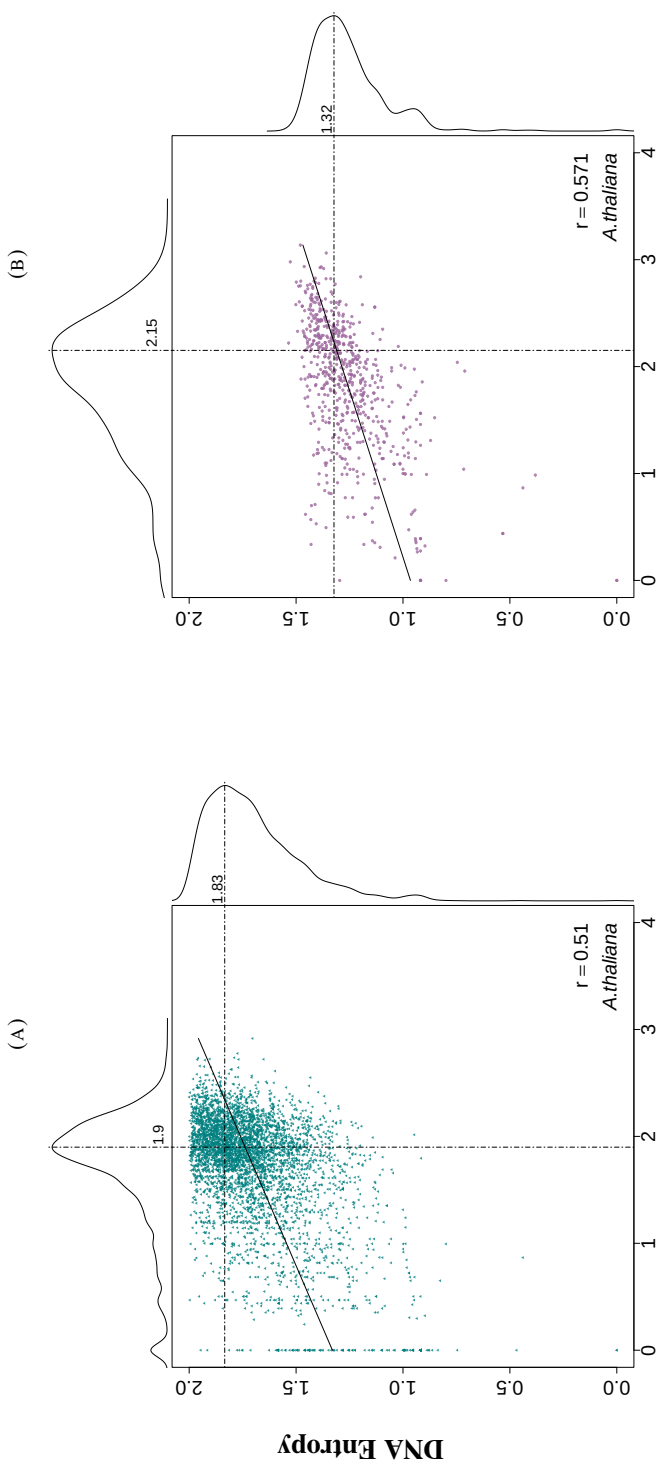

FIGURE A1.1: Entropy comparisons of LCRs and corresponding sequences in the *A. thaliana* genome. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. a 4470 LCRs were identified from 48 265 protein sequences and their entropies were plotted against entropies of the corresponding DNA sequences ($r = 0.510$). b 676 LCRs were identified from 48 265 coding sequences and their entropies were plotted against the entropies of the corresponding protein sequences ($r = 0.571$).

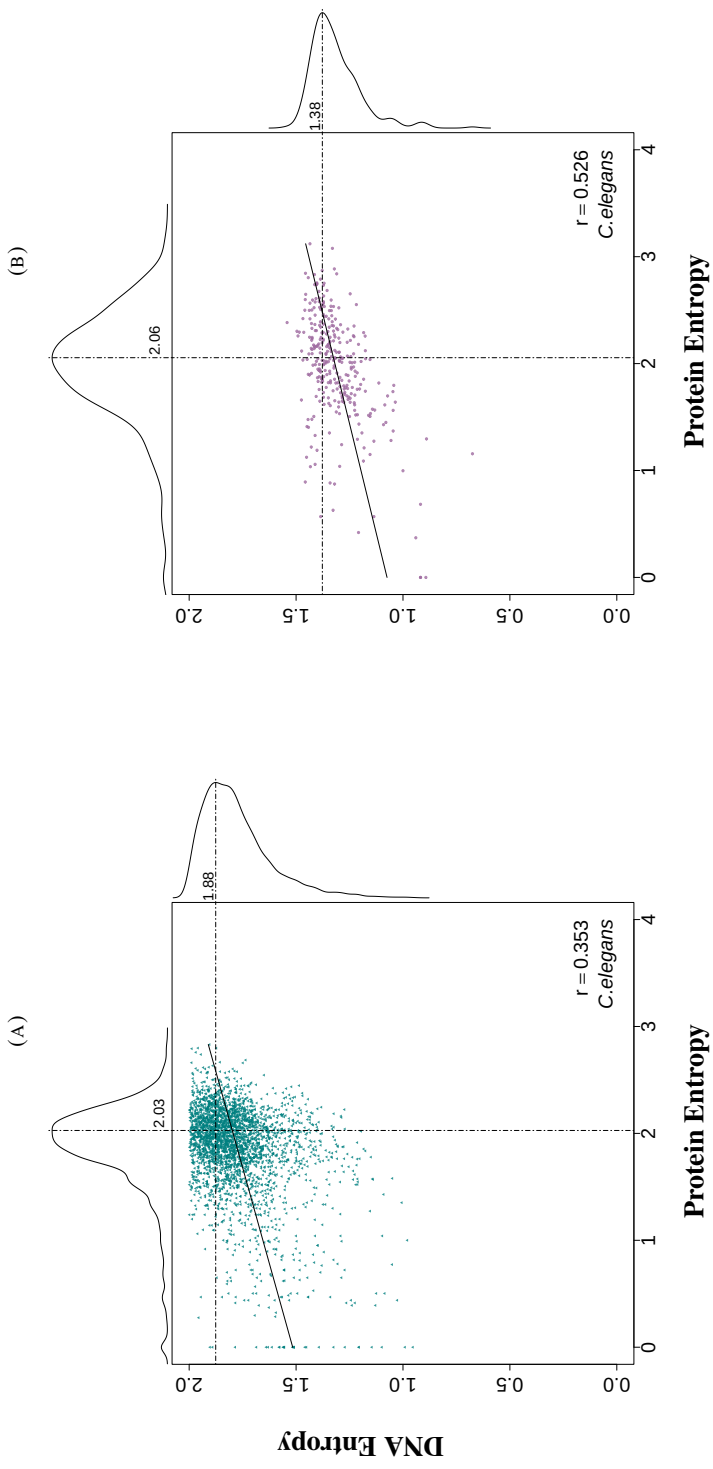

FIGURE A1.2: Entropy comparisons of LCRs and corresponding sequences in the *C. elegans* genome. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. a 3074 LCRs were identified from 30 502 protein sequences and their entropies were plotted against the entropies of the corresponding DNA sequences ($r = 0.353$). b 314 LCRs were identified from 30 502 coding sequences and their entropies were plotted against the entropies of the corresponding protein sequences ($r = 0.526$).

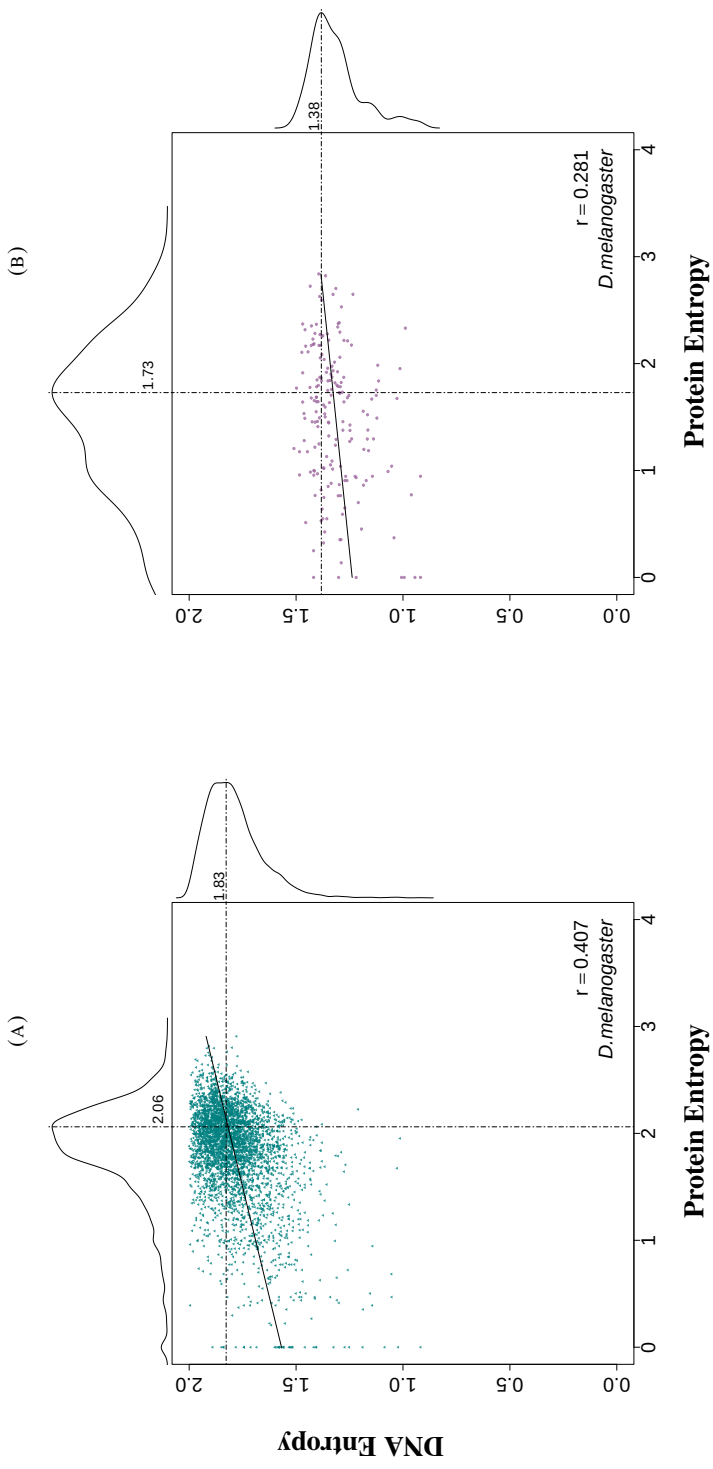

FIGURE A1.3: Entropy comparisons of LCRs and corresponding sequences in the *D. melanogaster* genome. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. a 3859 LCRs were identified from 30 726 protein sequences and their entropies were plotted against the entropies of the corresponding DNA sequences ($r = 0.407$). b 182 LCRs were identified from 30 726 coding sequences and their entropies were plotted against the entropies of the corresponding protein sequences ($r = 0.281$).

TABLE A1.11: Summary of correlation coefficients and 95% confidence intervals for the main parameters: $W = 15$, $K1 = 1.9$, $K2 = 2.2$ (protein LCRs), $W = 45$, $K1 = 1.3$, $K2 = 1.5$ (DNA LCRs), LCRs with amino acid repeats of minimal length 6, 5, and 4 for mono-, di- and tri- repeats, respectively (periodic LCRs). LCRs without repeats of minimal length 6, 5, and 4 for mono-, di-, and tri- repeats, respectively (cryptic LCRs). Results for the main parameters are given for the five model organisms (bio), the periodic repeat and cryptic LCRs, as well as the four species specific slippage and substitution models. Information for the Null simulation which is not species specific is given at the bottom.

		S. cerevisiae		***H. sapiens***		***A. thaliana***		***C. elegans***		***D. melanogaster***	
Bio	**Protein LCRs**	0.488	(0.435-0.538)	0.374	(0.347-0.400)	0.510	(0.485-0.534)	0.353	(0.318-0.387)	0.407	(0.377-0.436)
	DNA LCRs	0.428	(0.281-0.555)	0.579	(0.541-0.614)	0.571	(0.512-0.625)	0.526	(0.431-0.610)	0.281	(0.125-0.424)
Slip	**Protein LCRs**	0.566	(0.555-0.577)	0.658	(0.637-0.678)	0.594	(0.579-0.608)	0.547	(0.532-0.562)	0.596	(0.583-0.608)
	DNA LCRs	0.591	(0.573-0.608)	0.573	(0.503-0.636)	0.492	(0.457-0.526)	0.514	(0.484-0.543)	0.427	(0.392-0.461)
Slip+Syn	**Protein LCRs**	0.459	(0.446-0.472)	0.585	(0.561-0.608)	0.475	(0.457-0.492)	0.457	(0.440-0.474)	0.480	(0.465-0.495)
	DNA LCRs	0.517	(0.489-0.544)	0.588	(0.492-0.670)	0.513	(0.466-0.557)	0.445	(0.400-0.488)	0.475	(0.431-0.517)
Slip+CC	**Protein LCRs**	0.699	(0.690-0.707)	0.808	(0.795-0.820)	0.757	(0.747-0.767)	0.718	(0.707-0.728)	0.671	(0.660-0.682)
	DNA LCRs	0.553	(0.534-0.572)	0.546	(0.479-0.607)	0.535	(0.503-0.565)	0.497	(0.467-0.526)	0.582	(0.554-0.608)
Slip+CC+Syn	**Protein LCRs**	0.490	(0.477-0.503)	0.620	(0.598-0.641)	0.518	(0.501-0.534)	0.504	(0.488-0.520)	0.486	(0.471-0.501)
	DNA LCRs	0.430	(0.397-0.462)	0.536	(0.446-0.615)	0.452	(0.403-0.498)	0.396	(0.348-0.442)	0.458	(0.415-0.499)
Protein LCRs	**Periodic**	0.573	(0.481-0.652)	0.448	(0.402-0.492)	0.603	(0.567-0.637)	0.579	(0.523-0.630)	0.525	(0.482-0.566)
	Cryptic	0.322	(0.248-0.392)	0.242	(0.207-0.277)	0.302	(0.265-0.338)	0.211	(0.167-0.254)	0.271	(0.230-0.311)
Null	**Protein LCRs**	0.126	(-0.139-0.374)								
	DNA LCRs	NA	NA								

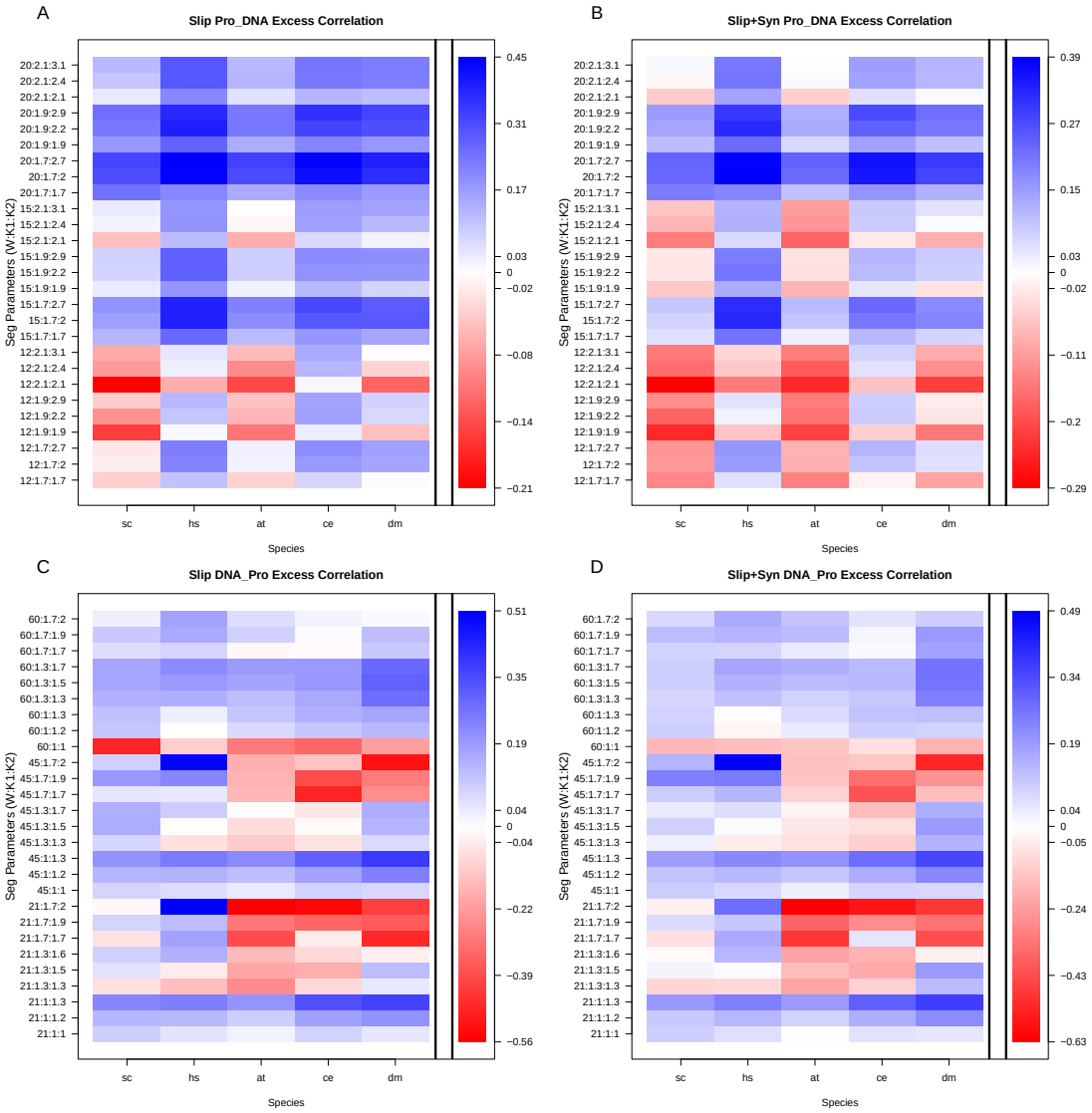

FIGURE A1.4: Heat maps showing the excess correlation of the slippage and substitution models compared to the biological correlations in *S. cerevisiae* (sc), *H. sapiens* (hs), *A. thaliana* (at), *C. elegans* (ce), and *D. melanogaster* (dm). Blue indicates a higher correlation in the simulated sequences relative to the biological sequences while red indicates a lower correlation in the simulated sequences relative to the biological sequences. **A)** Excess correlation in protein LCRs and corresponding coding sequences in the Slip simulated sequences. **B)** Excess correlation in DNA LCRs and corresponding protein sequences in the Slip simulated sequences. **C)** Excess correlation in protein LCRs and corresponding coding sequences in the Slip+Syn simulated sequences. **D)** Excess correlation in DNA LCRs and corresponding protein sequences in the Slip+Syn simulated sequences.

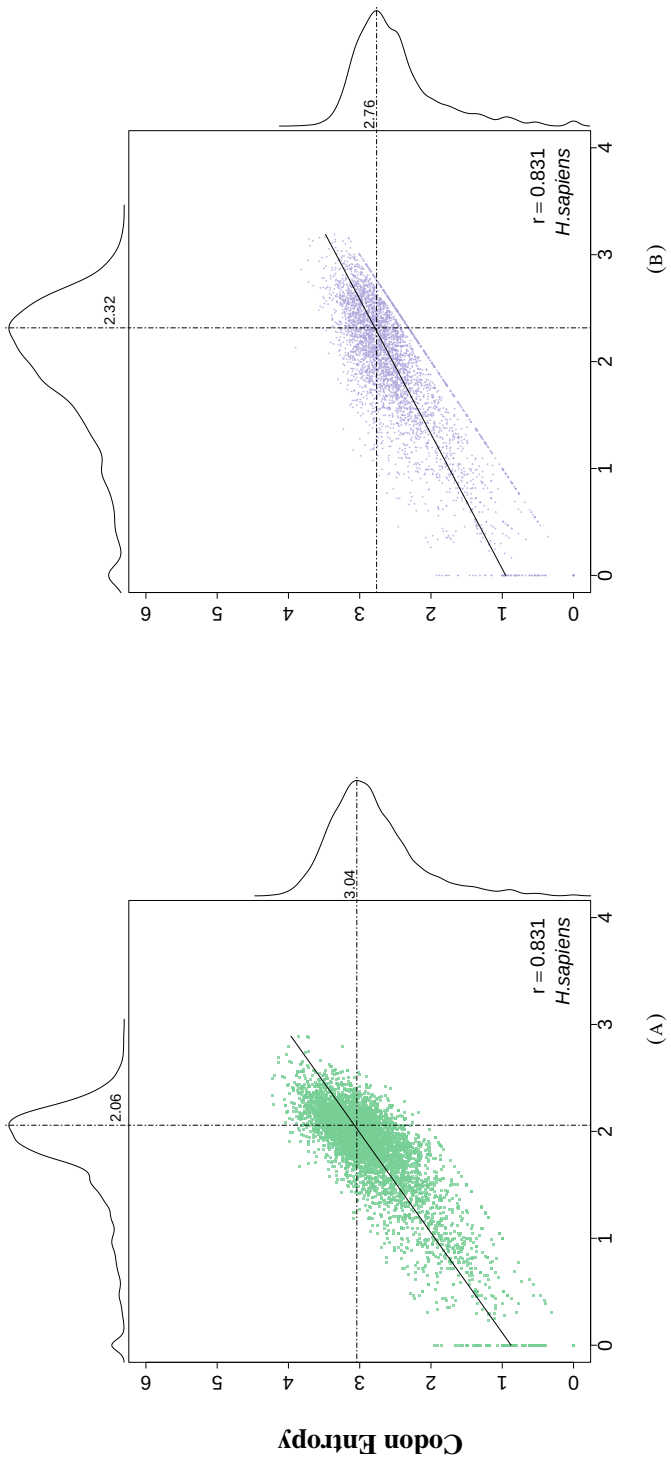

FIGURE A1.5: Entropy comparisons of LCRs and corresponding sequences in the *H. sapiens* genome. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. a 5005 LCRs were identified from 133 689 protein sequences and their entropies were plotted against the codon entropies from the corresponding DNA sequences ($r = 0.831$). b 4493 LCRs were identified from the codons in 133 689 coding sequences and their entropies were plotted against the entropies of the corresponding protein sequences ($r = 0.831$).

Appendix B

Chapter 4 Supplement

TABLE A2.1: Number of genes/proteins with data by species and tissue

Dataset	Type	Species	Tissue	Total	LCR$^+$
GTEx	TAb	Human	Aggregate	18067	4259 (23.6%)
			Brain	13903	3536 (25.4%)
			Colon	13981	3508 (25.1%)
			Esophagus	13694	3435 (25.1%)
			Heart	13223	3330 (25.2%)
			Kidney	13913	3437 (24.7%)
			Liver	12914	3180 (24.6%)
			Lung	14146	3524 (24.9%)
			Ovary	13582	3432 (25.3%)
			Pancreas	13291	3333 (25.1%)
			Prostate	14118	3532 (25.0%)
			Skin	13935	3491 (25.1%)
			Testis	15931	3839 (24.1%)
PaxDB	PAb	Human	Aggregate	20108	4246 (21.1%)
			Brain	8771	2437 (27.8%)
			Colon	7833	1944 (24.8%)
			Esophagus	6125	1434 (23.4%)
			Heart	10388	2675 (25.8%)
			Kidney	7427	1782 (24.0%)
			Liver	13341	3340 (25.0%)
			Ovary	11684	2986 (25.6%)
			Pancreas	9450	2437 (25.8%)
			Prostate	8992	2328 (25.9%)
			Skin	4267	950 (22.3%)
			Testis	12305	3158 (25.7%)
Schwanhäusser	TAb	Mouse	Fibroblast	3407	865 (25.4%)
	PAb			3407	865 (25.4%)
	k. Deg			4746	450 (9.5%)
	Trans. Eff.			3457	885 (25.6%)
Degredation	k. Deg	Human	Aggregate	8315	965 (11.6%)
RNA-Seq	TAb	Mouse	Brn/Hrt/Kdn/Lvr	70989	18094 (25.5%)
		Rat		55312	12915 (23.3%)
		Macaque		67926	18097 (26.6%)
		Chimpanzee		80690	20777 (25.7%)
		Human		108155	25233 (23.3%)
		Dog		58683	14908 (25.4%)
		Horse		60473	16543 (27.4%)
		Pig		63500	17571 (27.7%)
		Cow		63508	16756 (26.4%)

TABLE A2.2: Dataset specific estimates of wobble base-pairing selective constraints

Dataset	Species	AA	TG	GT	AC
GTEx	Human	1.000	0.3040	0.8406	1.370×10^{-2}
Schwanhäusser	Mouse	0.9945	1.190×10^{-2}	0.9678	1.470×10^{-2}
RNA-Seq	Mouse	1.000	0.5059	0.6665	1.049×10^{-3}
	Rat	1.000	0.2250	0.5793	9.999×10^{-7}
	Macaque	0.9991	0.6875	0.2960	2.600×10^{-4}
	Chimpanzee	0.9992	0.5311	0.2136	0.1583
	Human	0.9997	0.831	0.0629	0.1059
	Dog	1.000	0.6399	0.1149	4.700×10^{-4}
	Horse	0.9998	0.1196	0.1166	0.1417
	Pig	1.000	0.4881	0.3497	0.1288
	Cow	0.9999	3.528×10^{-5}	0.9999	0.9973

TABLE A2.3: Logistic regression using standardized GTEx and Schwänhausser data

Mammal	Proteins	LCR^+	Parameter	Transformation	$\bar{X}$	$SD_{\bar{X}}$	β	SE_β	Z value
Human	3107	903	(Intercept)	NA	NA	NA	-1.16	0.0628	-18.4
			Length	log2	8.78	1.060	-0.237	0.116	-2.040
			TAb	log2	0.638	2.87	0.244	0.0585	4.16
			PAb	log2	1.23	3.74	-0.212	0.0608	-3.49
			TWnTE	log2	-5.48	1.36	-0.910	0.112	-8.15
			k.deg	log	-3.83	0.852	0.147	0.0475	3.10
Mouse	2155	446	(Intercept)	NA	NA	NA	-1.38	0.0631	-21.8
			Length	log2	8.77	1.080	0.409	0.148	2.77
			TAb	log2	6.83	1.53	0.188	0.0801	2.35
			PAb	log2	3.81	3.24	-0.0843	0.0920	-0.916
			TWnTE	log2	-12.0	1.40	-0.449	0.200	-2.25
			k.deg	log	-4.20	1.030	0.115	0.0668	1.72

TABLE A2.4: Logistic regression using standardized mammalian RNA-Seq data

Mammal	Proteins	LCR$^+$	Parameter	Transformation	$\bar{X}$	$SD_{\bar{X}}$	β	SE_{β}	Z value
Chimp	73001	18614	(Intercept)	NA	NA	NA	-1.21	0.00948	-128
			Length	log2	9.020	1.13	-0.289	0.0291	-9.91
			TAb	log2	-0.564	2.44	0.0902	0.00914	9.87
			TWnTE	log2	-4.50	1.31	-1.080	0.0280	-38.7
Cow	58860	15584	(Intercept)	NA	NA	NA	-1.15	0.0105	-110
			Length	log2	8.98	1.090	0.561	0.0188	29.8
			TAb	log2	-0.0590	2.60	0.0901	0.0101	8.91
			TWnTE	log2	-14.9	1.14	-0.241	0.0211	-11.5
Dog	32599	8075	(Intercept)	NA	NA	NA	-1.27	0.0145	-87.1
			Length	log2	8.96	1.080	-0.381	0.0361	-10.6
			TAb	log2	0.457	2.74	0.0821	0.0141	5.81
			TWnTE	log2	-4.61	1.15	-1.080	0.0342	-31.7
Horse	52030	14389	(Intercept)	NA	NA	NA	-1.060	0.0107	-99.3
			Length	log2	9.040	1.11	-0.295	0.0289	-10.2
			TAb	log2	0.354	3.030	0.0477	0.0108	4.42
			TWnTE	log2	-4.55	1.29	-1.030	0.0282	-36.4
Human	88026	20201	(Intercept)	NA	NA	NA	-1.37	0.00908	-151
			Length	log2	8.94	1.090	-0.312	0.0291	-10.7
			TAb	log2	-0.252	2.64	0.115	0.00862	13.3
			TWnTE	log2	-5.89	1.35	-1.090	0.0282	-38.6
Macaque	63441	16578	(Intercept)	NA	NA	NA	-1.18	0.0101	-117
			Length	log2	8.99	1.12	-0.366	0.0312	-11.7
			TAb	log2	-0.0632	2.57	0.0830	0.00988	8.40
			TWnTE	log2	-4.58	1.39	-1.14	0.0302	-37.8
Mouse	54544	13644	(Intercept)	NA	NA	NA	-1.28	0.0118	-108
			Length	log2	8.96	1.10	0.330	0.0368	8.96
			TAb	log2	-0.419	3.060	0.0938	0.0107	8.75
			TWnTE	log2	-12.2	0.999	-0.569	0.0444	-12.8
Pig	55521	15144	(Intercept)	NA	NA	NA	-1.080	0.0104	-104
			Length	log2	9.030	1.070	-0.0985	0.0270	-3.65
			TAb	log2	-0.635	2.82	0.0418	0.0104	4.030
			TWnTE	log2	-4.22	1.20	-0.820	0.0260	-31.6
Rat	35562	8429	(Intercept)	NA	NA	NA	-1.37	0.0153	-90.1
			Length	log2	8.88	1.080	-0.280	0.0452	-6.20
			TAb	log2	-0.0723	3.060	0.0868	0.0136	6.40
			TWnTE	log2	-10.7	1.11	-1.15	0.0515	-22.3

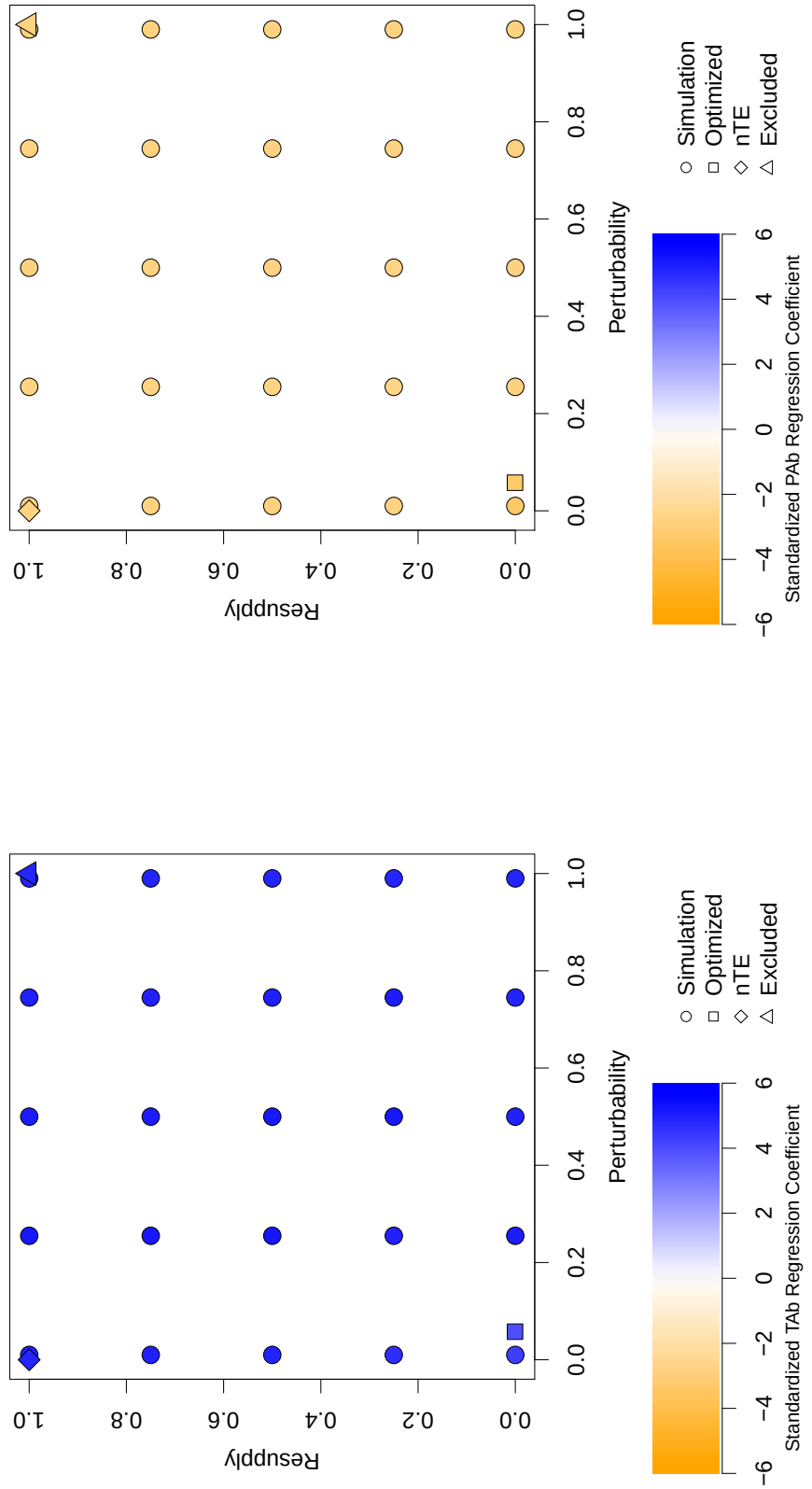

FIGURE A2.1: Choice of translation parameters has no effect on the strength of the relationship between the presence of an LCR and TAb (a) and PAb (b). Colour indicates the z value of the regression coefficient when the indicated pair of translation parameters are used. Using standard nTE calculations is equivalent to a resupply of 1 and perturbability of 0. Excluding translation efficiency from the regression is positioned at (1,1) despite not having a true position in the parameter space.

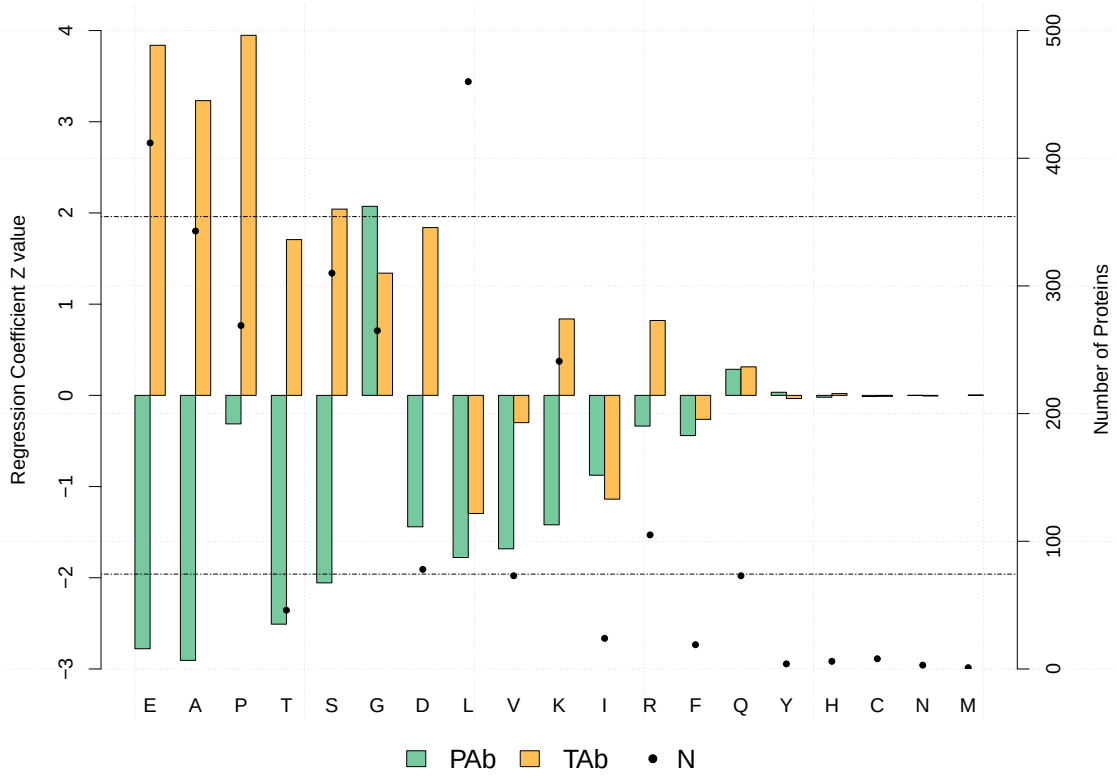

FIGURE A2.2: Aggregate LCR associations are also present for individial amino acids. Bars represent the z values of estimated logistic regression coefficients for the TAb and PAb with interaction terms for each amino acid. The amino acid for each protein is the amino acid which appears most often in the 15 AA windows of the protein sequence which have minimum entropy. The dotted line represents the z value corresponding to the 95% threshold. Individual points are the number of proteins with complete data which have the corresponding primary AA.

Appendix C

Chapter 6 Supplement

A1 Human Accessions

TABLE A3.1: IGSR Samples used to analyze temporal order of LCRs and TAb

Sample ID	Sex	Ancestry
HG00732	female	American
HG00864	female	East Asian
HG00512	male	East Asian
HG01596	male	East Asian
HG02492	male	South Asian
HG02587	female	African
HG03065	male	African
HG03371	male	African
HG03732	male	South Asian
NA18534	male	East Asian
NA18939	female	East Asian
NA19239	male	African
NA20509	male	European
HG00171	female	European
HG00096	male	European
HG00513	female	East Asian
HG00731	male	American
HG01114	female	American
HG00514	female	East Asian
HG02011	male	African
HG01505	male	European
NA12329	female	European
HG03683	female	South Asian
NA19238	female	African
NA19240	female	African
NA19983	female	African
NA19650	male	American
NA20847	female	South Asian

A2 Model Priors

TABLE A3.2: Priors for Stepwise OU TAb and LCR co-evolutionary models

Model	Parameter	Distribution Lower Bound	Upper Bound	
all Models	δ	Normal(0,10)	0	3300
	κ	LogNormal(-0.5,0.5)	0	∞
	λ	LogNormal(-0.5,0.5)	0	∞
	μ	LogNormal(-13.5,2.0)	e^{-35}	1
	σ	Normal(0,10)	0	∞
full	τ	Normal(0,2)	$-\infty$	∞
	υ	Uniform(-1,1)	-1	1
-tau	τ	Fixed(0)	0	0
	υ	Uniform(-1,1)	-1	1
-upsilon	τ	Normal(0,2)	$-\infty$	∞
	υ	Fixed(0)	0	0
-tau-upsilon	τ	Fixed(0)	0	0
	υ	Fixed(0)	0	0

A3 Modeling Summaries

A3.1 StepwiseOU-full-1

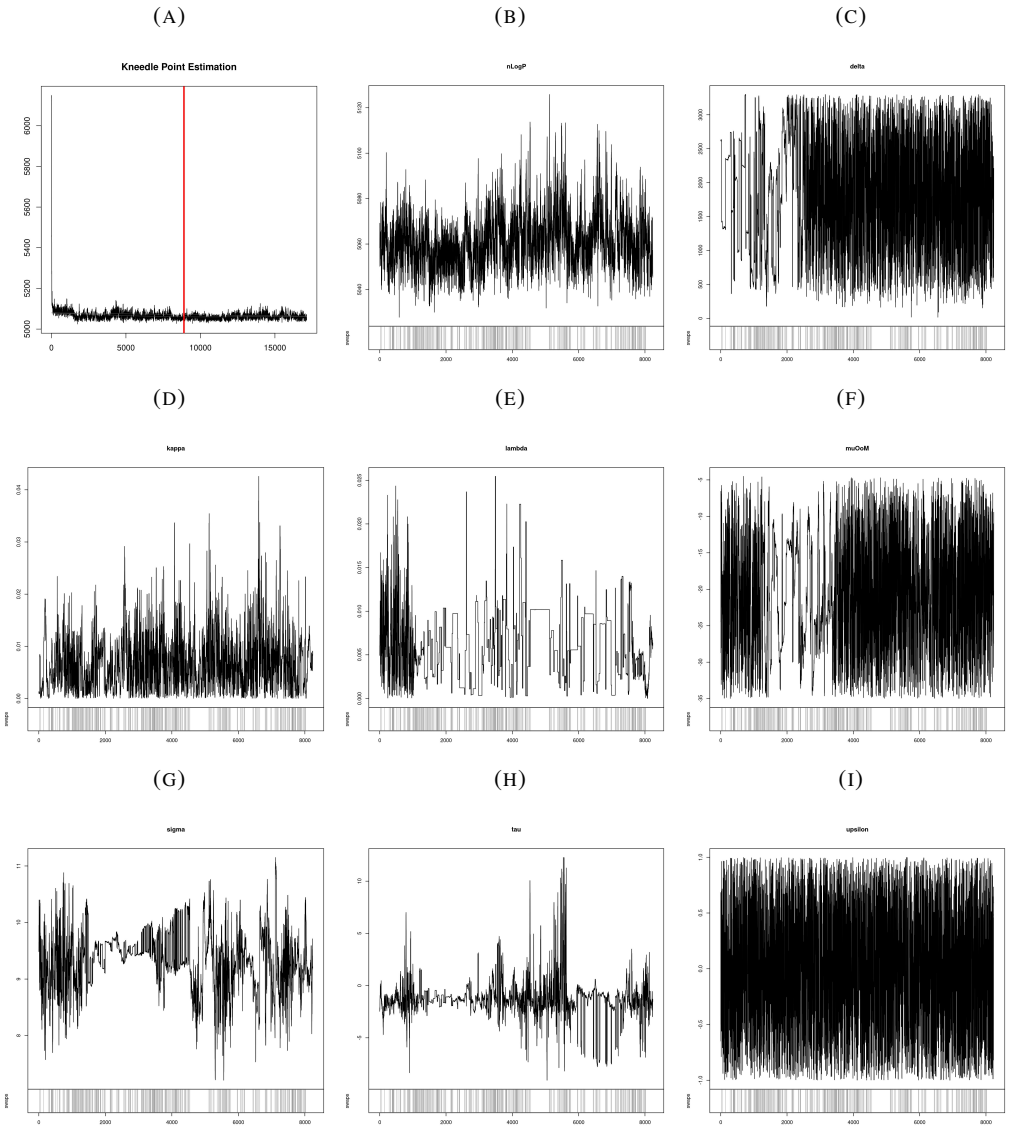

FIGURE A3.1: MCMC traces for full replicate 1. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

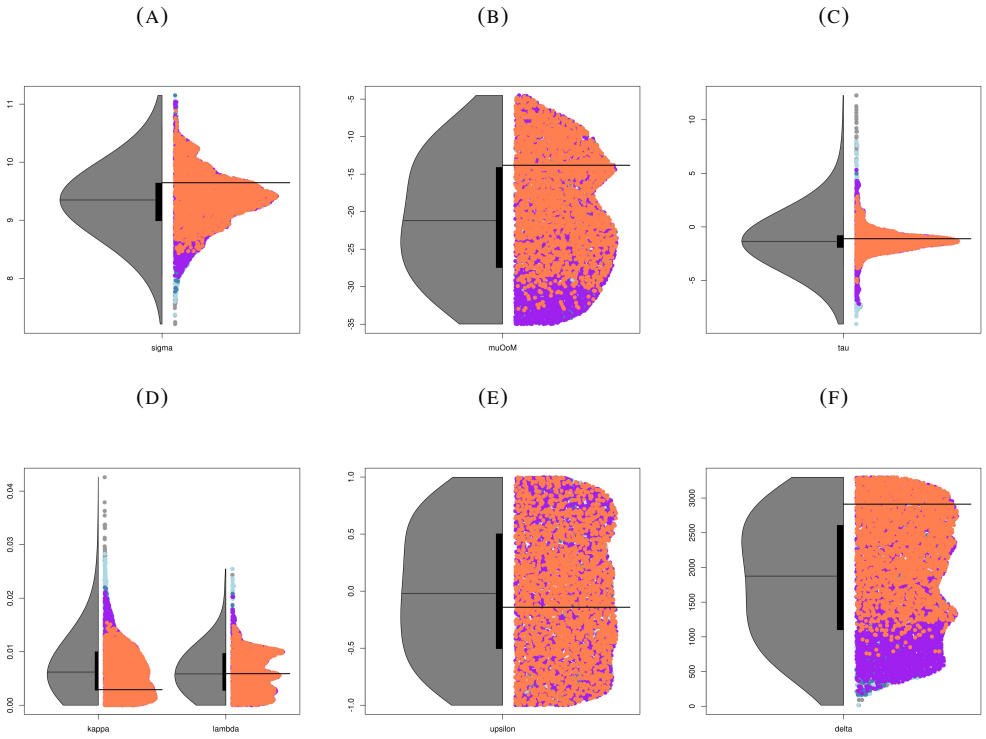

FIGURE A3.2: Density plots for full replicate 1. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

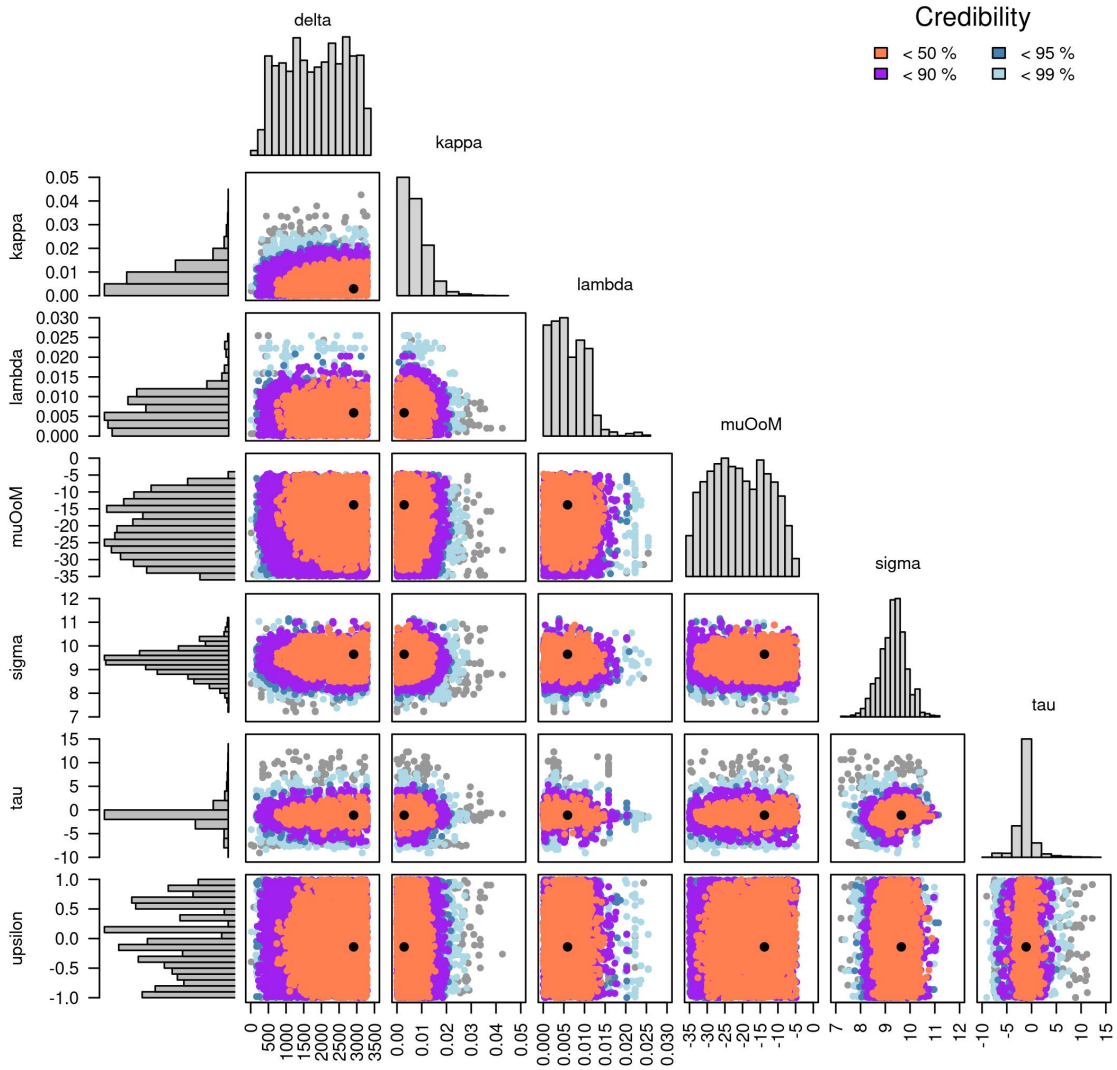

FIGURE A3.3: A visual summary of the posterior distribution estimate by ABC of the full Stepwise OU model (full replicate 1).

A3.2 StepwiseOU-full-2

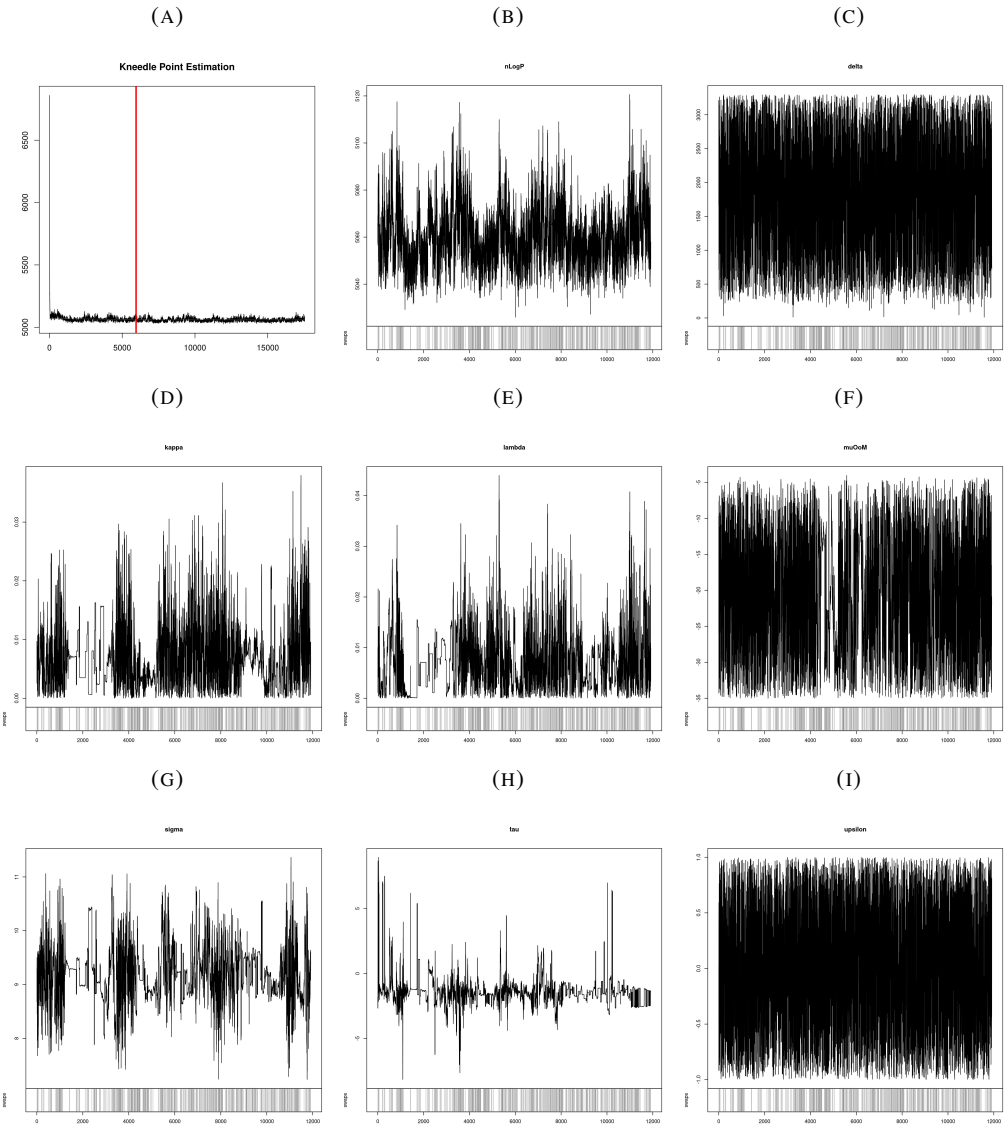

FIGURE A3.4: MCMC traces for full replicate 2. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

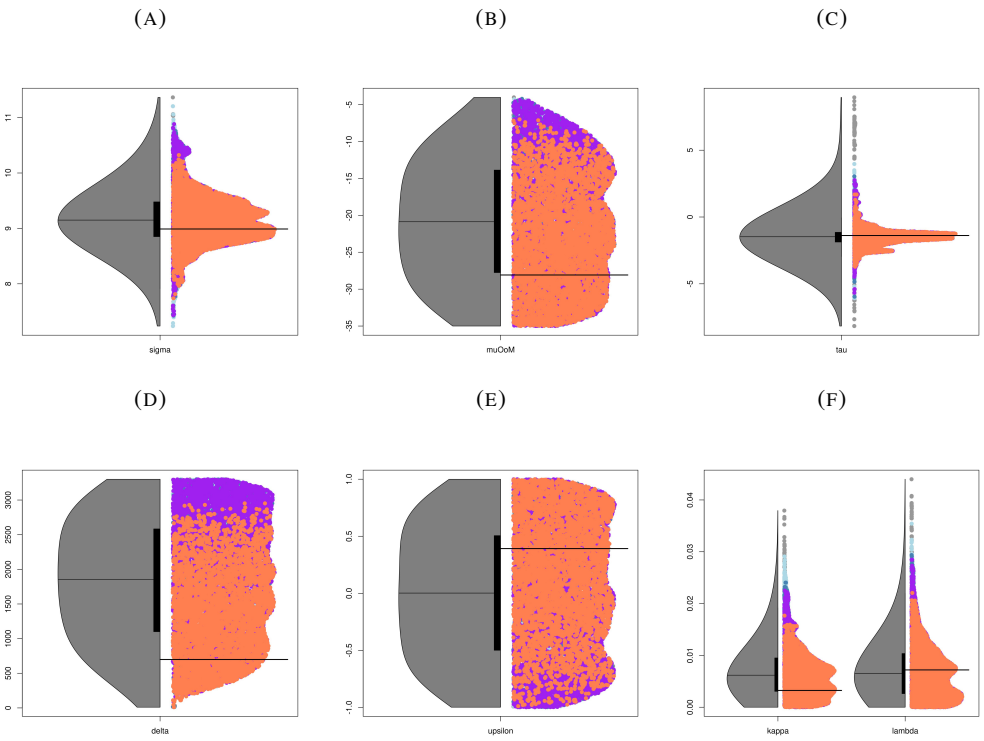

FIGURE A3.5: Density plots for full replicate 2. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

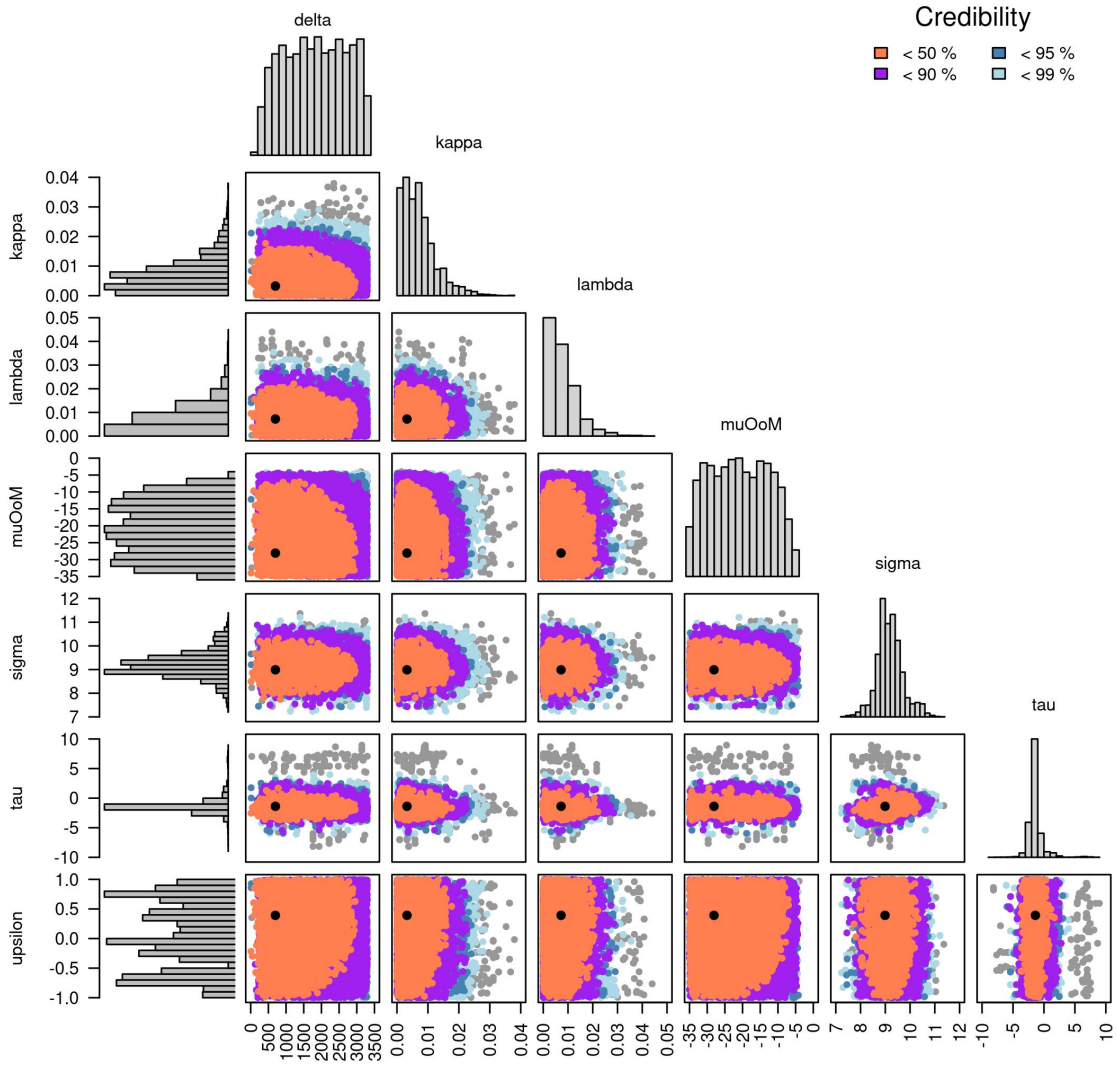

FIGURE A3.6: A visual summary of the posterior distribution estimate by ABC of the full Stepwise OU model (full replicate 2).

A3.3 StepwiseOU-full-3

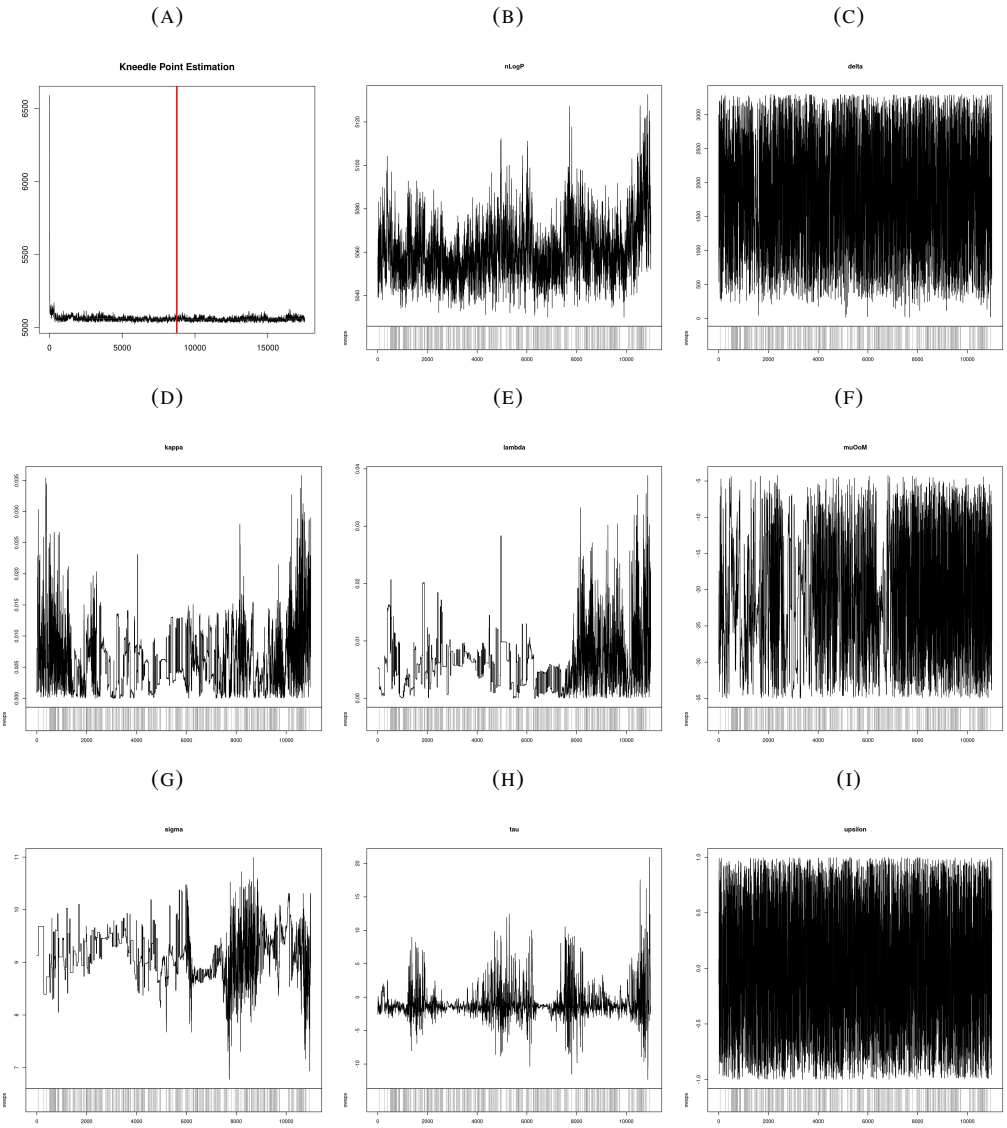

FIGURE A3.7: MCMC traces for full replicate 3. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

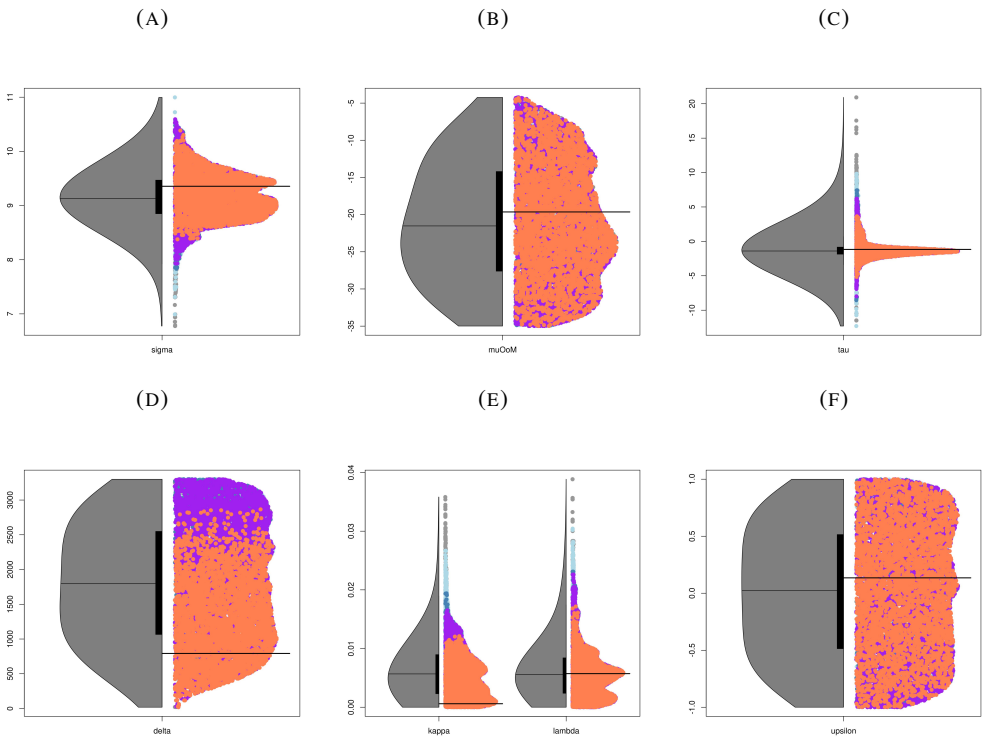

FIGURE A3.8: Density plots for full replicate 3. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

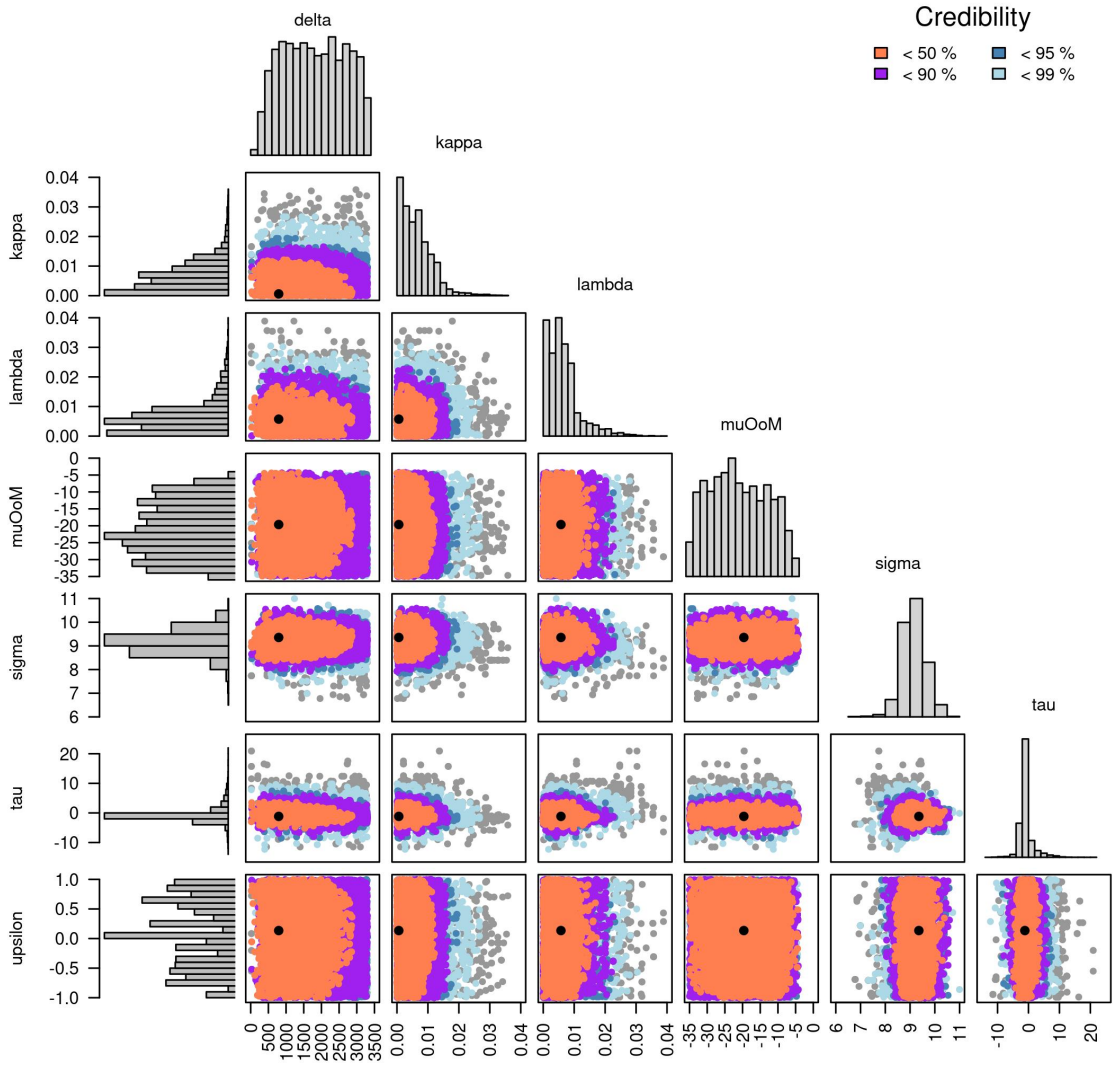

FIGURE A3.9: A visual summary of the posterior distribution estimate by ABC of the full Stepwise OU model (full replicate 3).

A3.4 StepwiseOU-tau-1

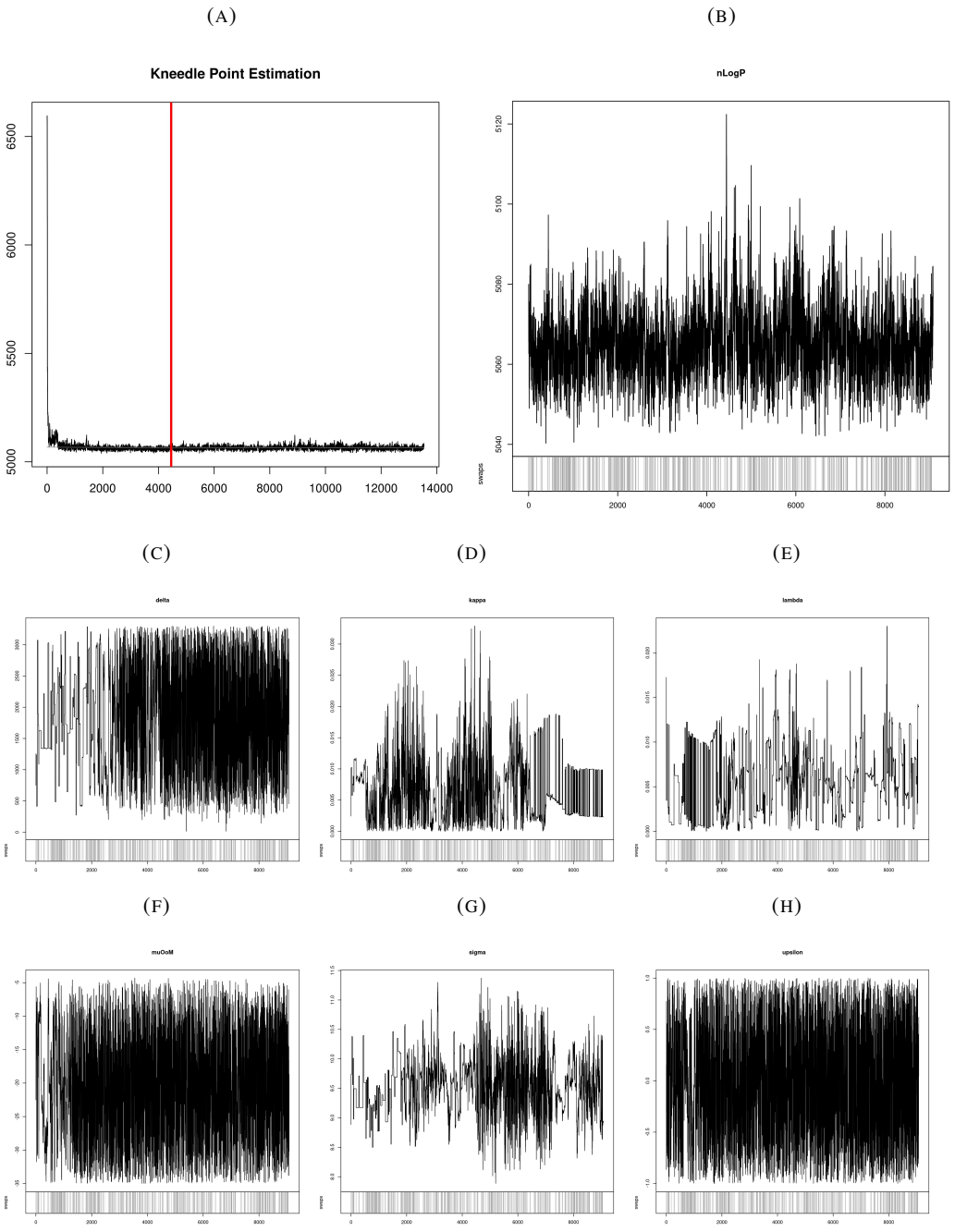

FIGURE A3.10: MCMC traces for -tau replicate 1. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

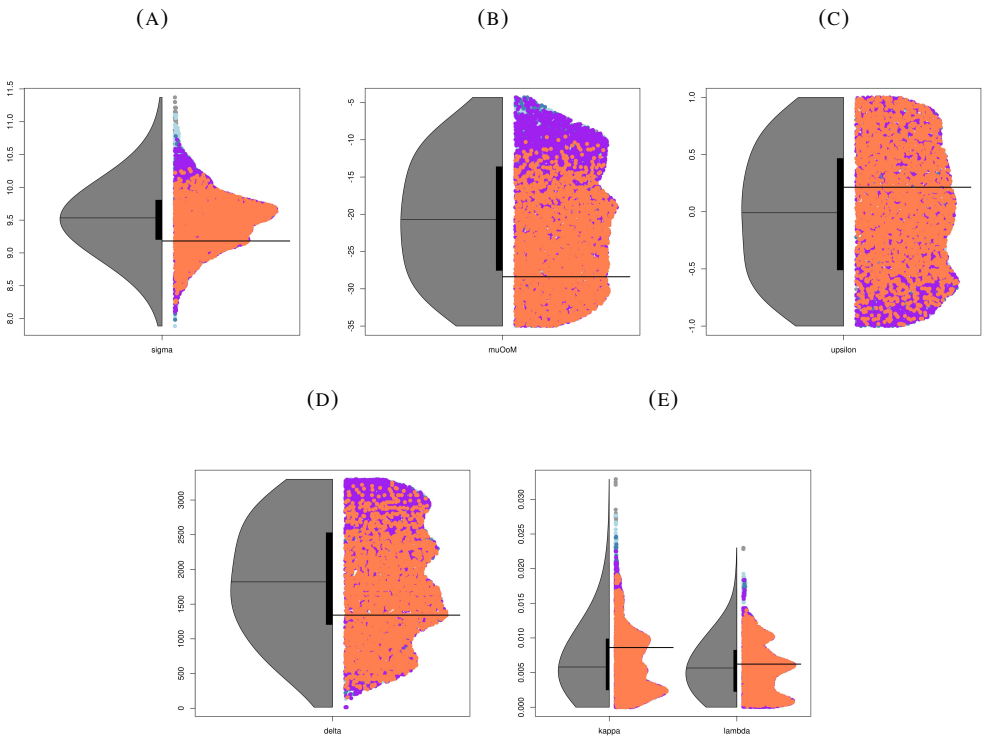

FIGURE A3.11: Density plots for -tau replicate 1. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

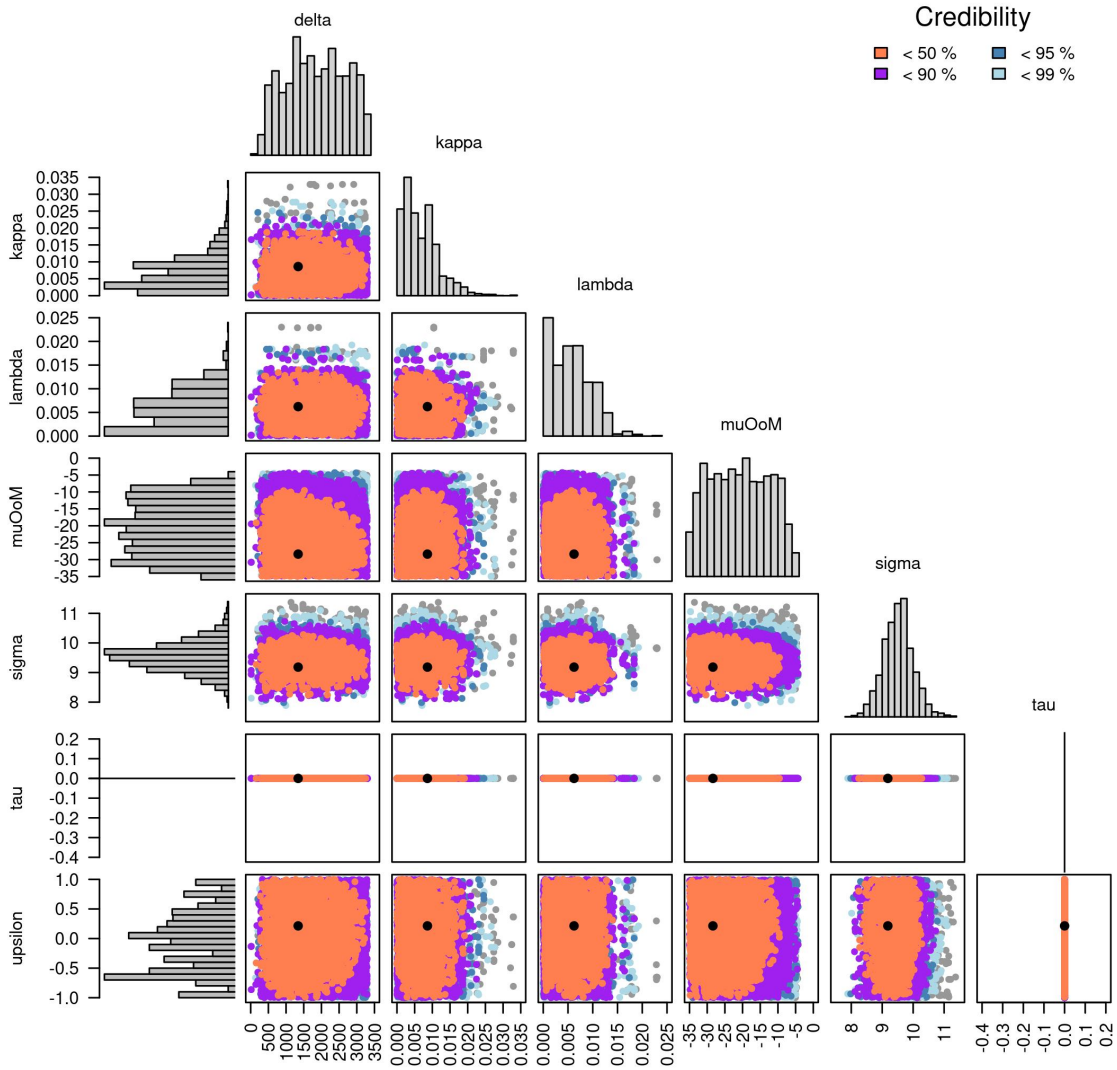

FIGURE A3.12: A visual summary of the posterior distribution estimate by ABC of an evolutionary model which assumes TAb is independent of CMER length (-tau replicate 1).

A3.5 StepwiseOU-tau-2

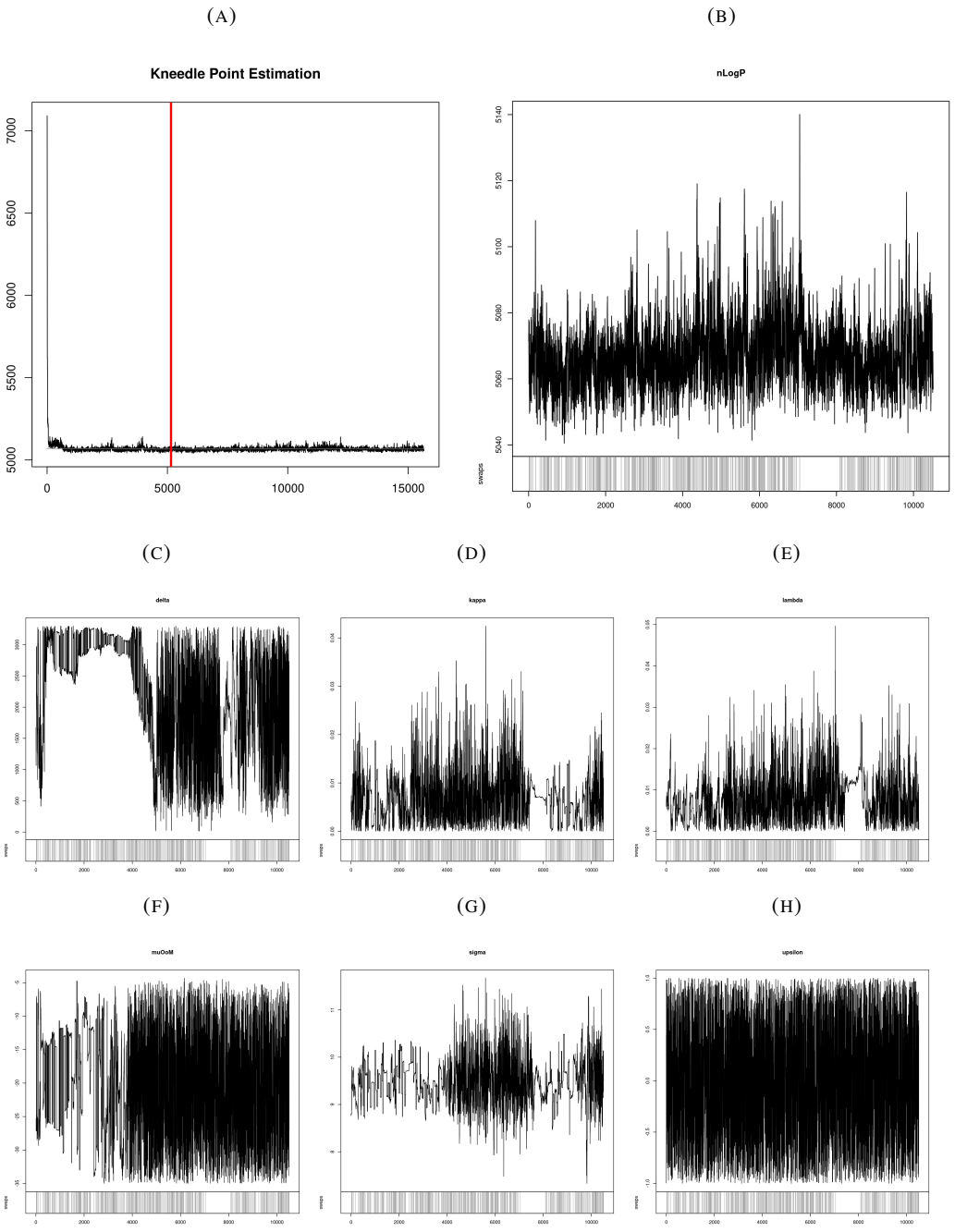

FIGURE A3.13: MCMC traces for -tau replicate 2. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

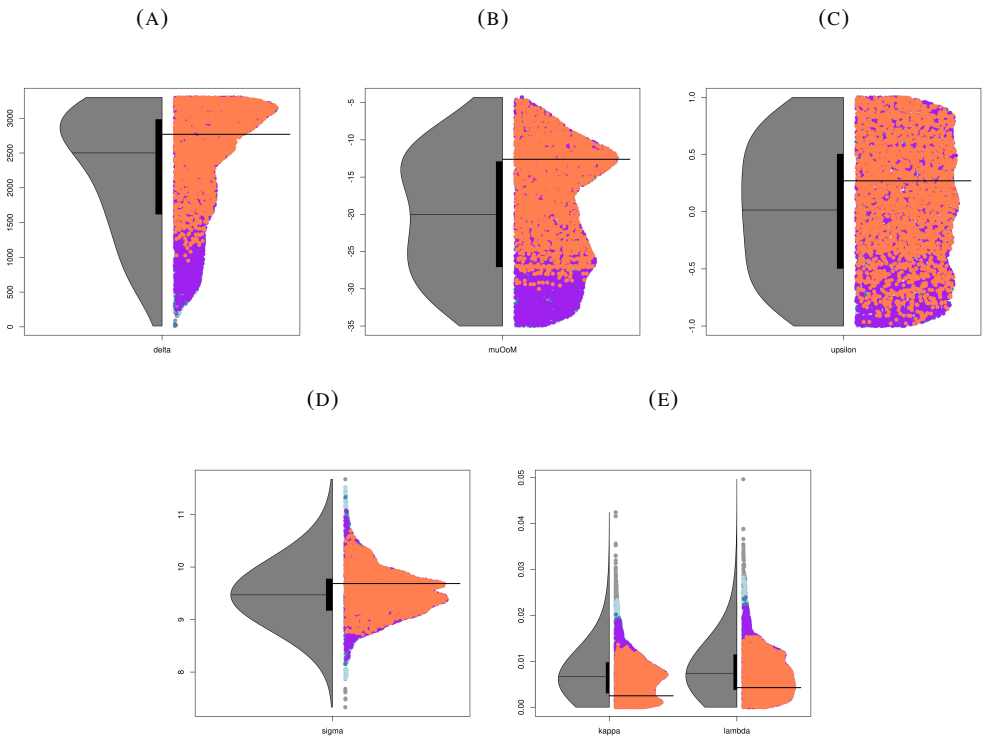

FIGURE A3.14: Density plots for -tau replicate 2. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

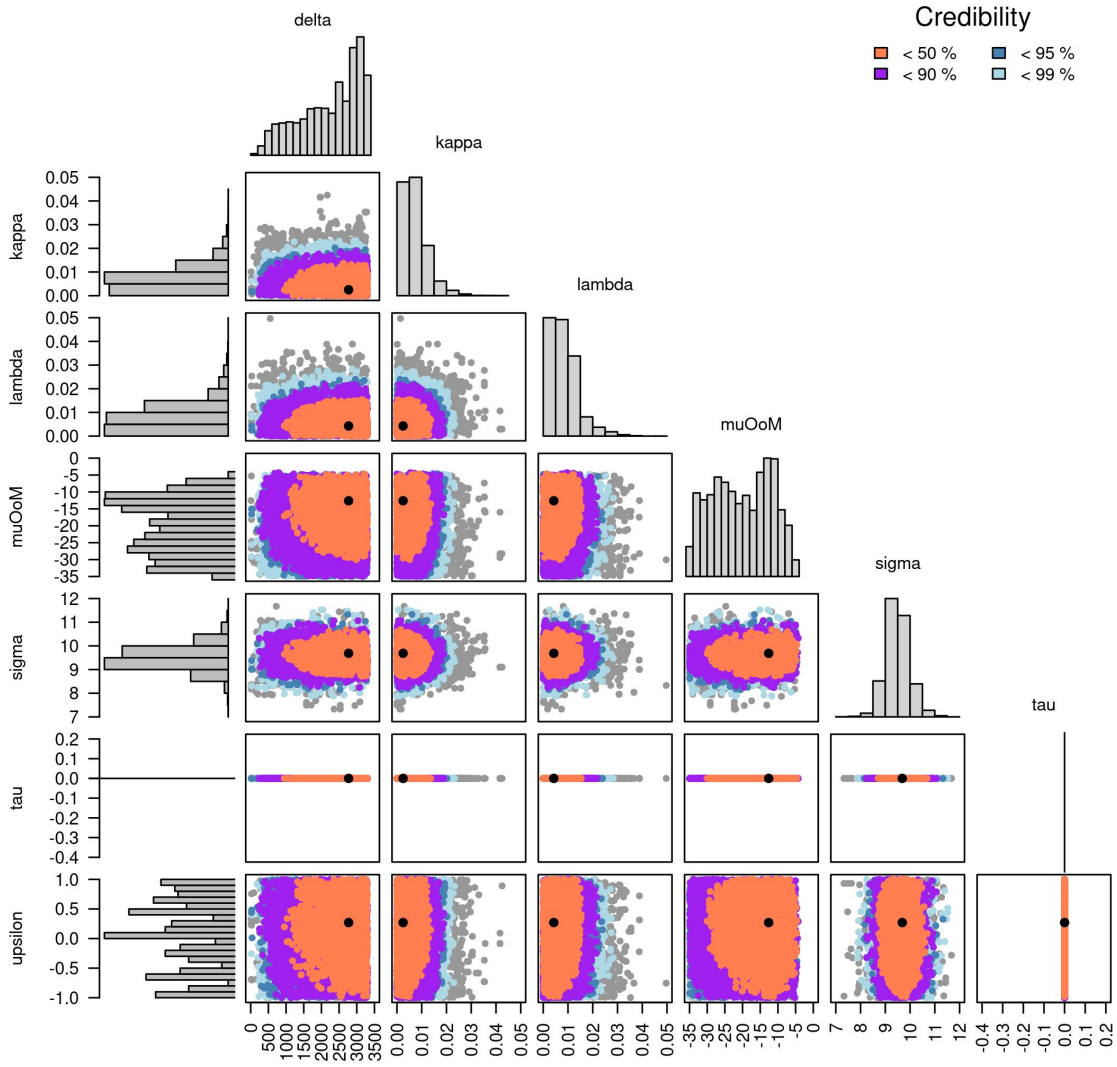

FIGURE A3.15: A visual summary of the posterior distribution estimate by ABC of an evolutionary model which assumes TAb is independent of CMER length (-tau replicate 2).

A3.6 StepwiseOU-tau-3

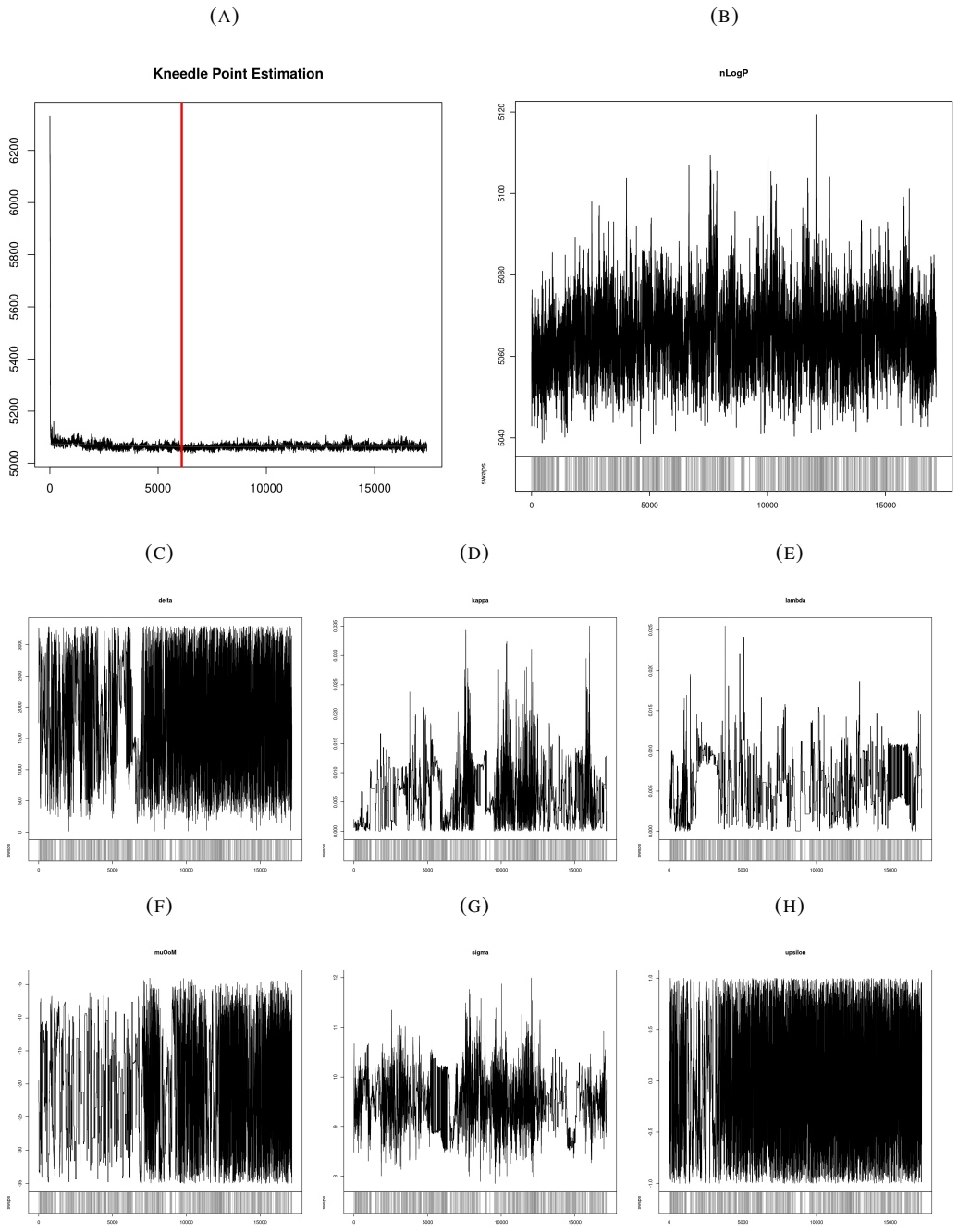

FIGURE A3.16: MCMC traces for -tau replicate 3. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

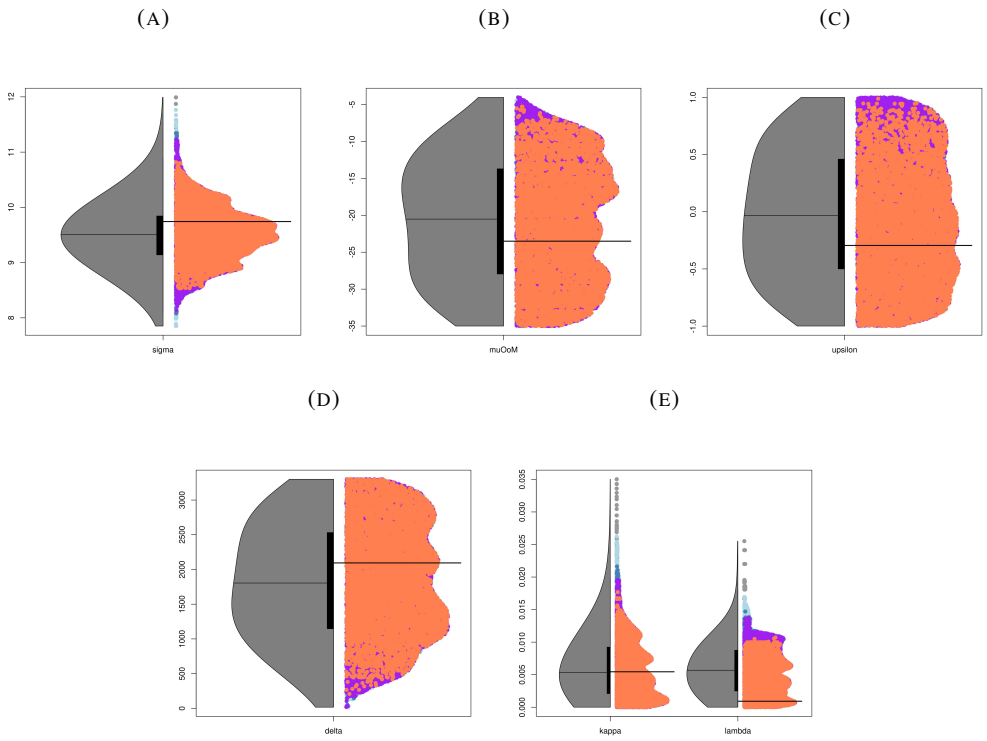

FIGURE A3.17: Density plots for -tau replicate 3. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

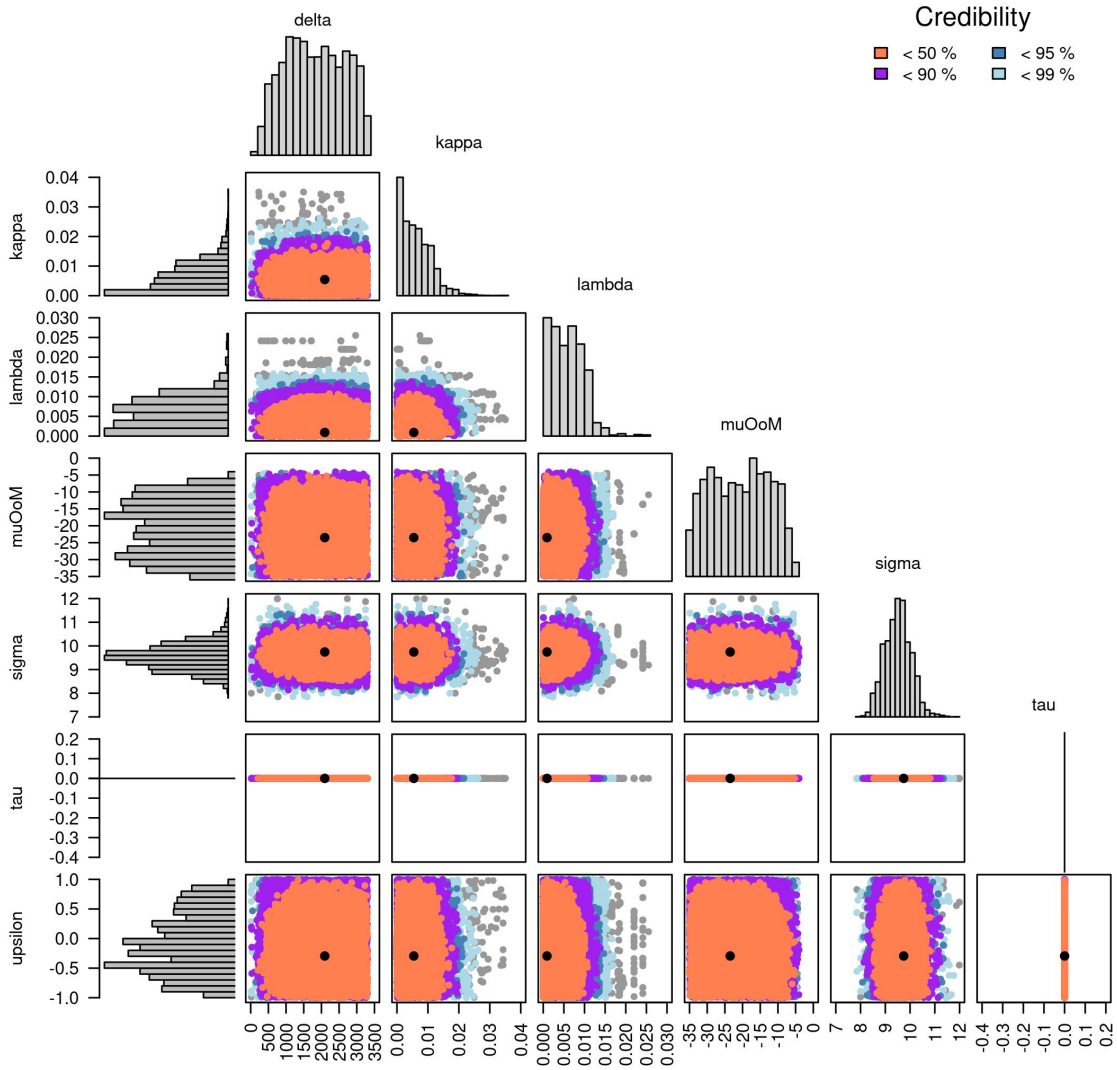

FIGURE A3.18: A visual summary of the posterior distribution estimate by ABC of an evolutionary model which assumes TAb is independent of CMER length (-tau replicate 3).

A3.7 StepwiseOU-upsilon-1

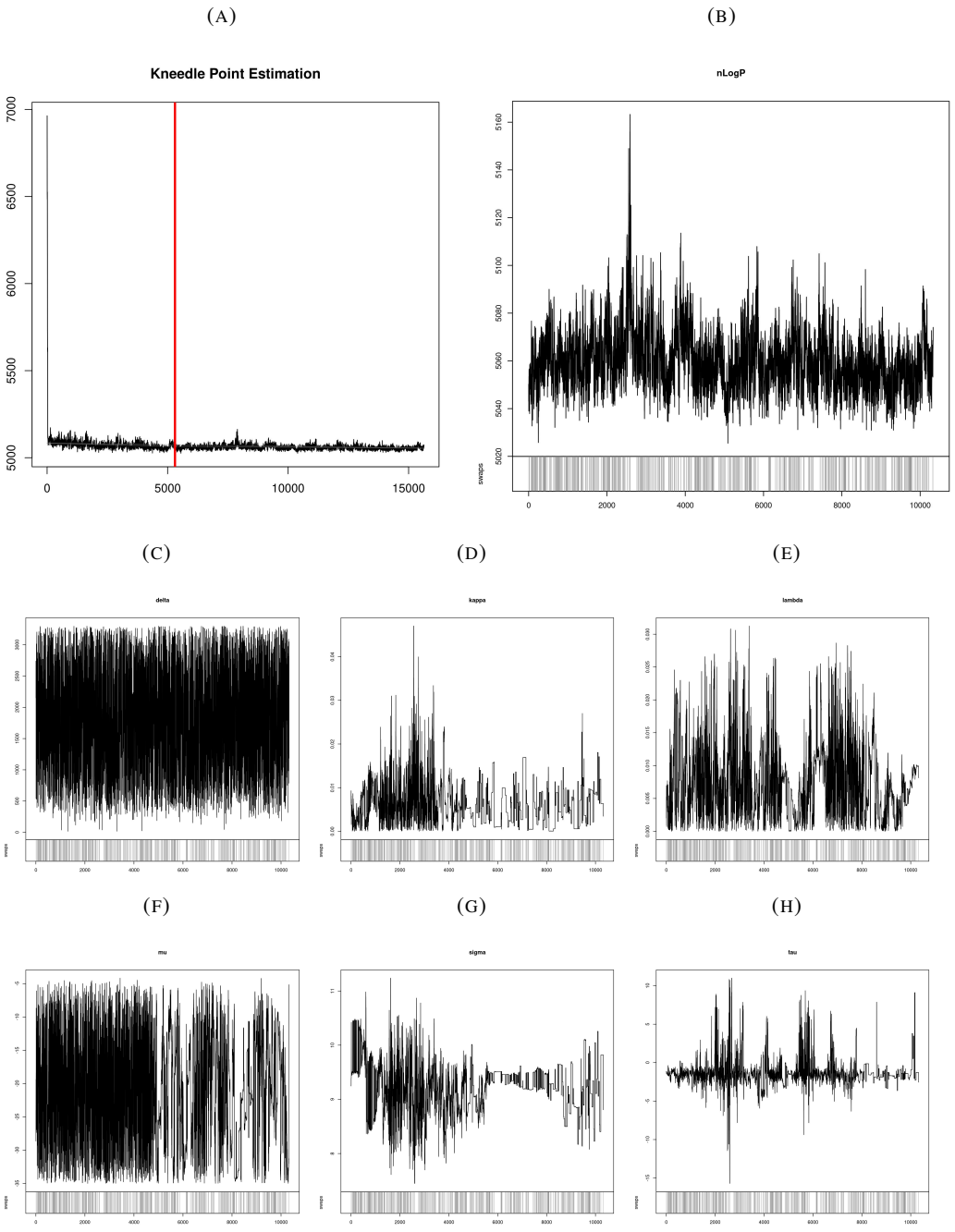

FIGURE A3.19: MCMC traces for -upsilon replicate 1. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

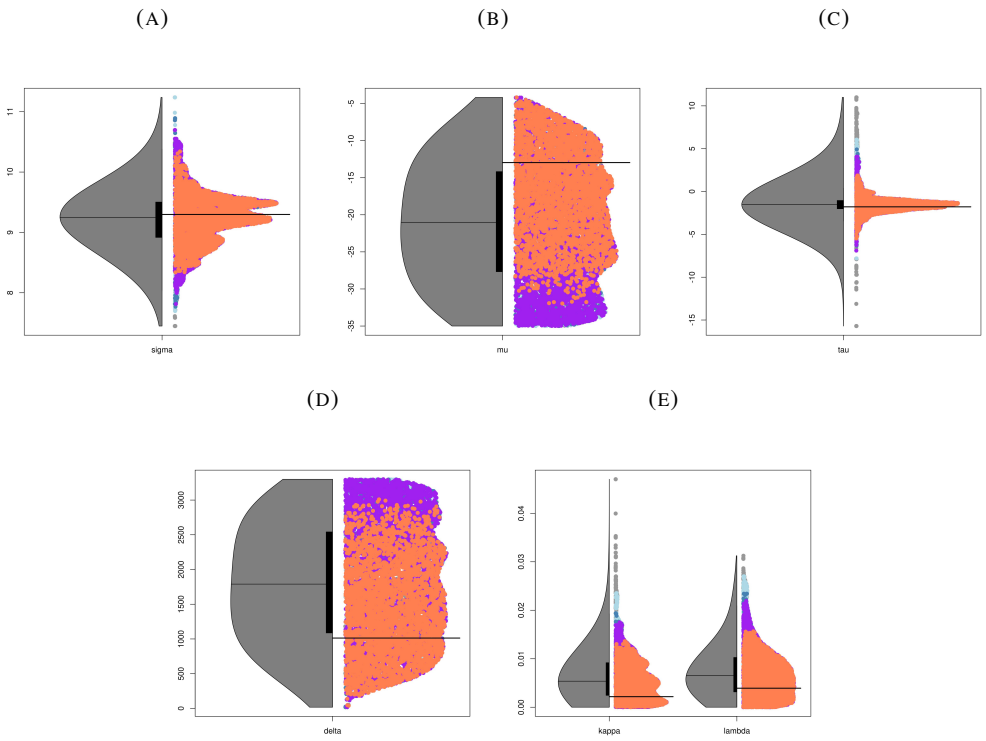

FIGURE A3.20: Density plots for -upsilon replicate 1. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

A3.8 StepwiseOU-upsilon-2

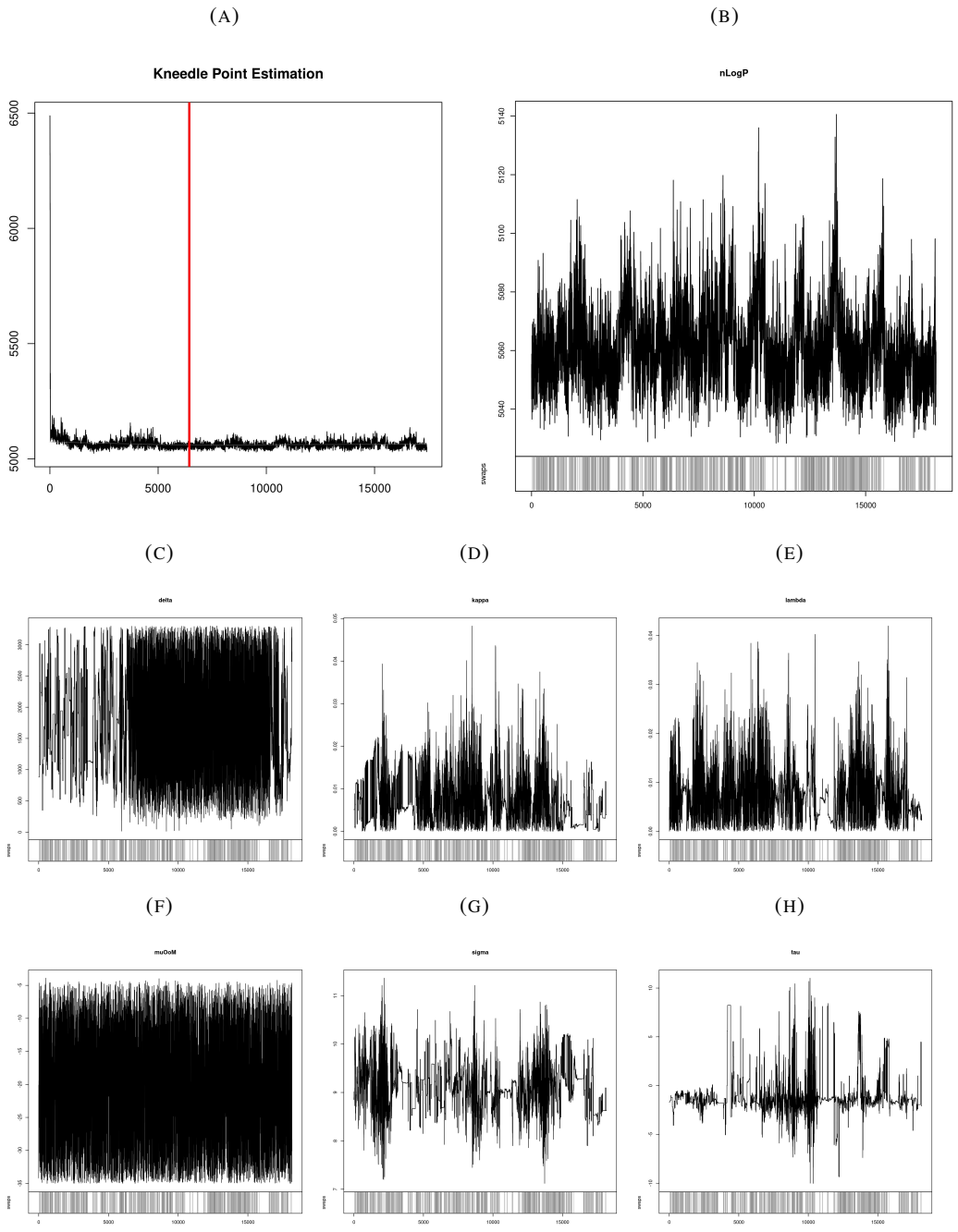

FIGURE A3.21: MCMC traces for -upsilon replicate 2. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

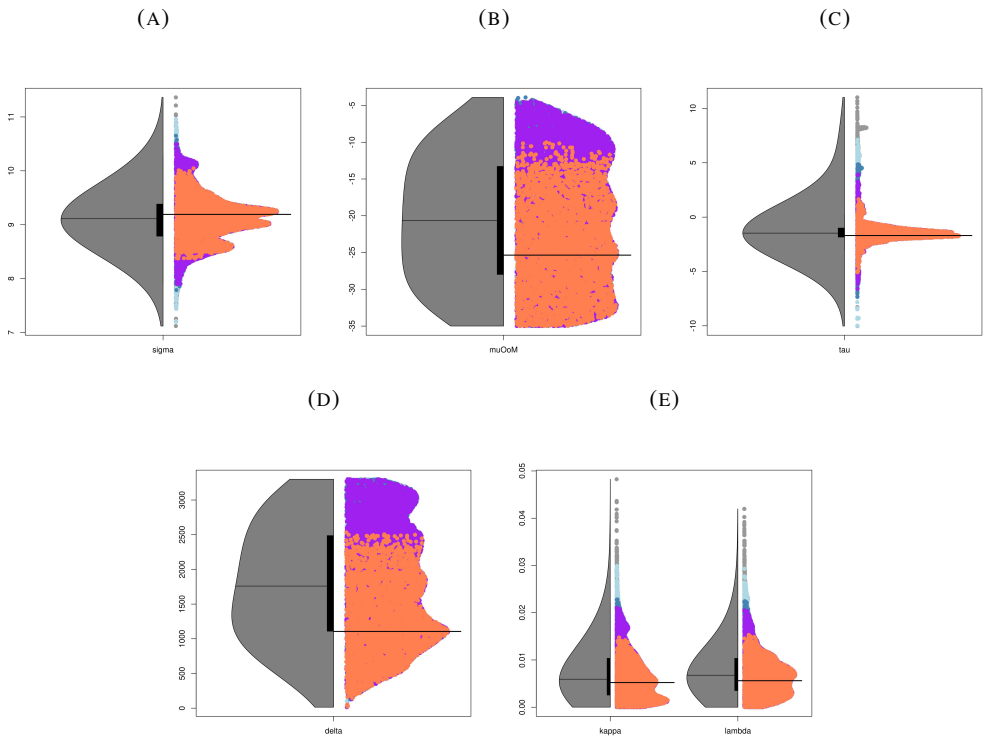

FIGURE A3.22: Density plots for -upsilon replicate 2. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

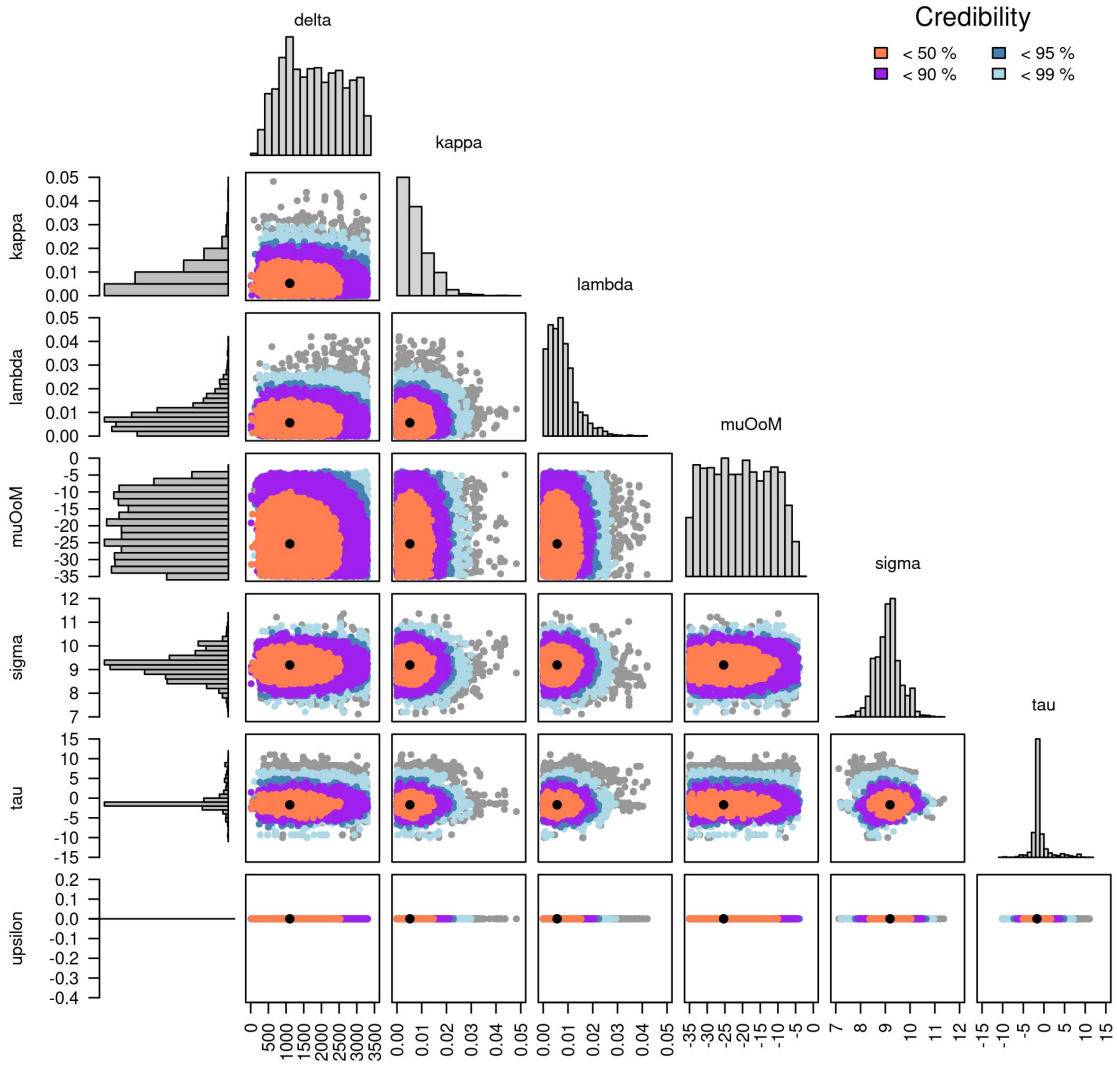

FIGURE A3.23: A visual summary of the posterior distribution estimate by ABC of an evolutionary model which assumes indel rates are independent of TAb (-upsilon replicate 2).

A3.9 StepwiseOU-upsilon-3

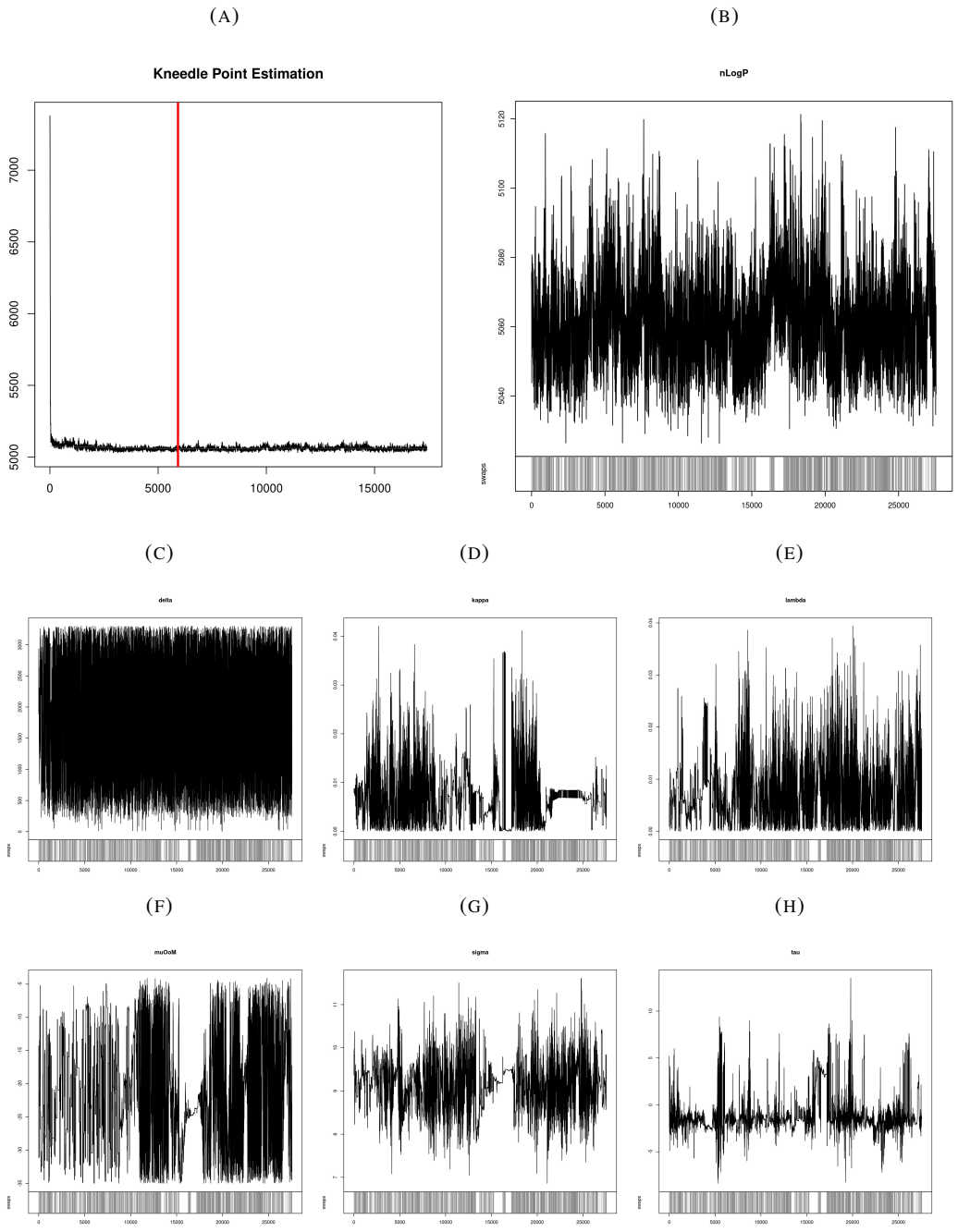

FIGURE A3.24: MCMC traces for -upsilon replicate 3. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

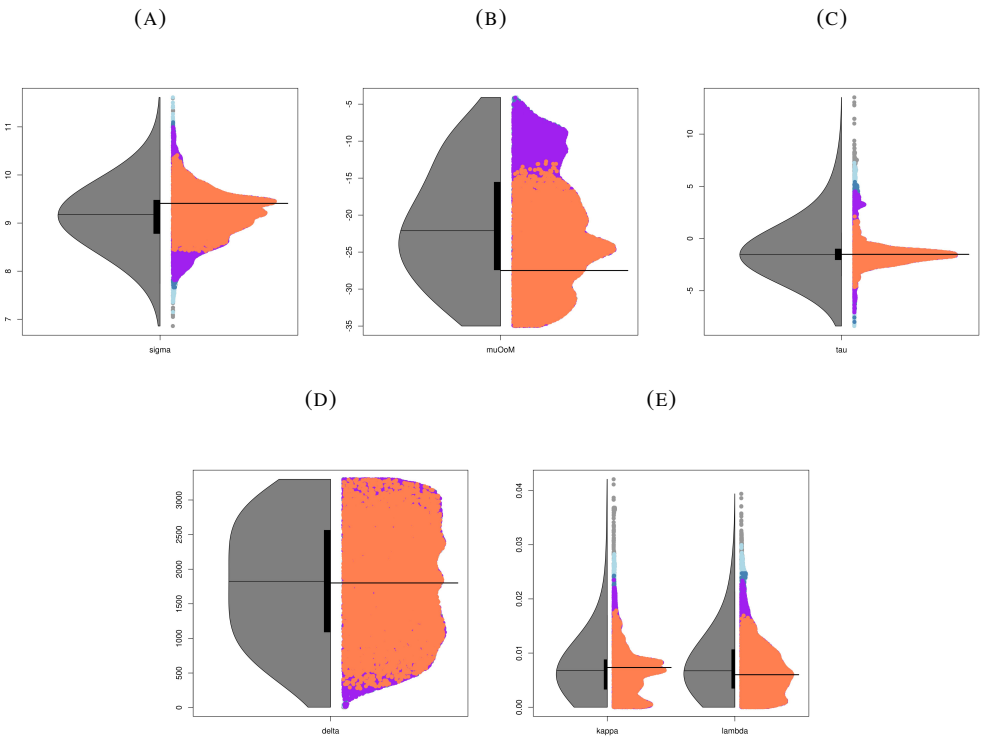

FIGURE A3.25: Density plots for -upsilon replicate 3. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

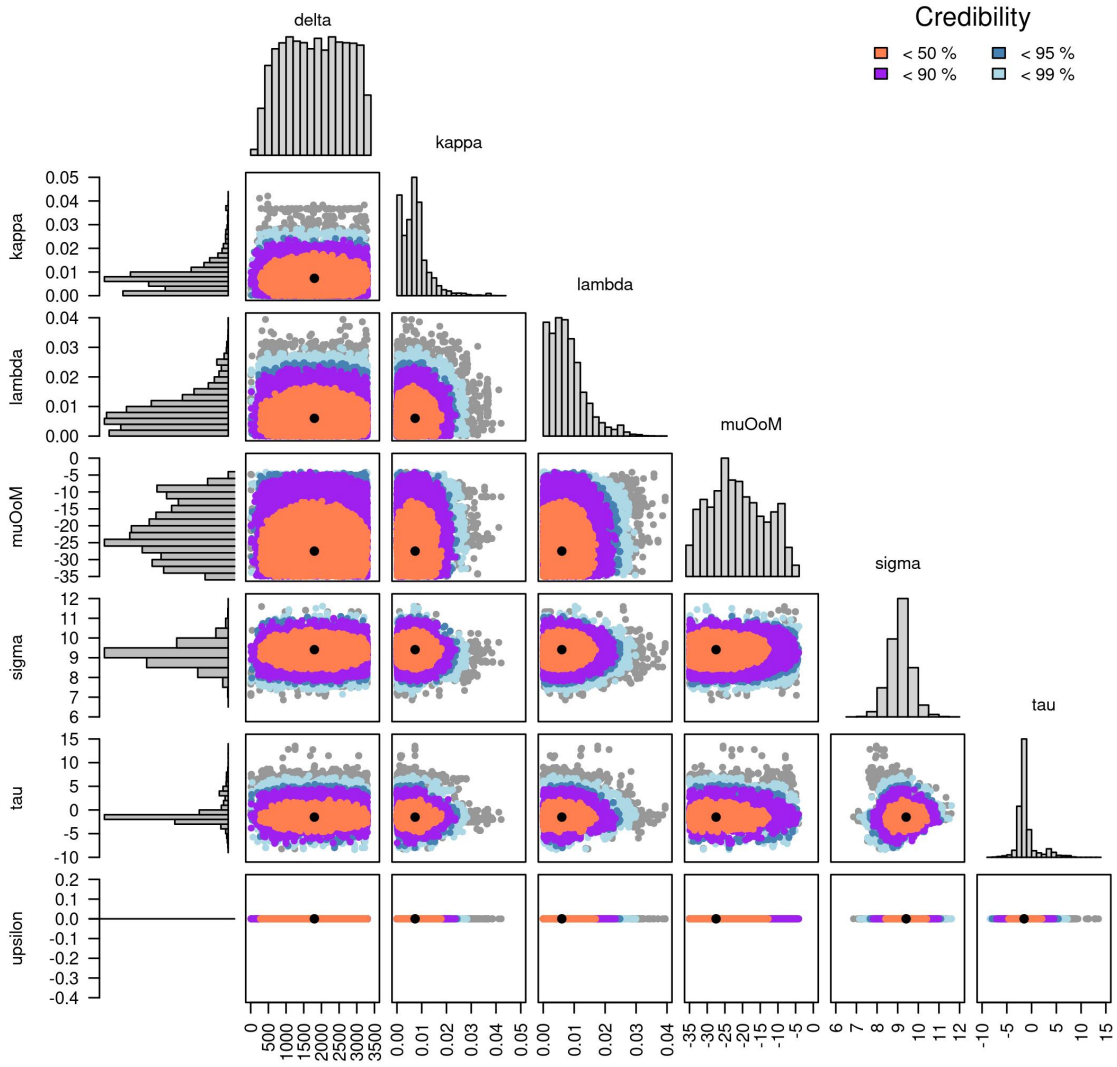

FIGURE A3.26: A visual summary of the posterior distribution estimate by ABC of an evolutionary model which assumes indel rates are independent of TAb (-upsilon replicate 3).

A3.10 StepwiseOU-tau-upsilon-1

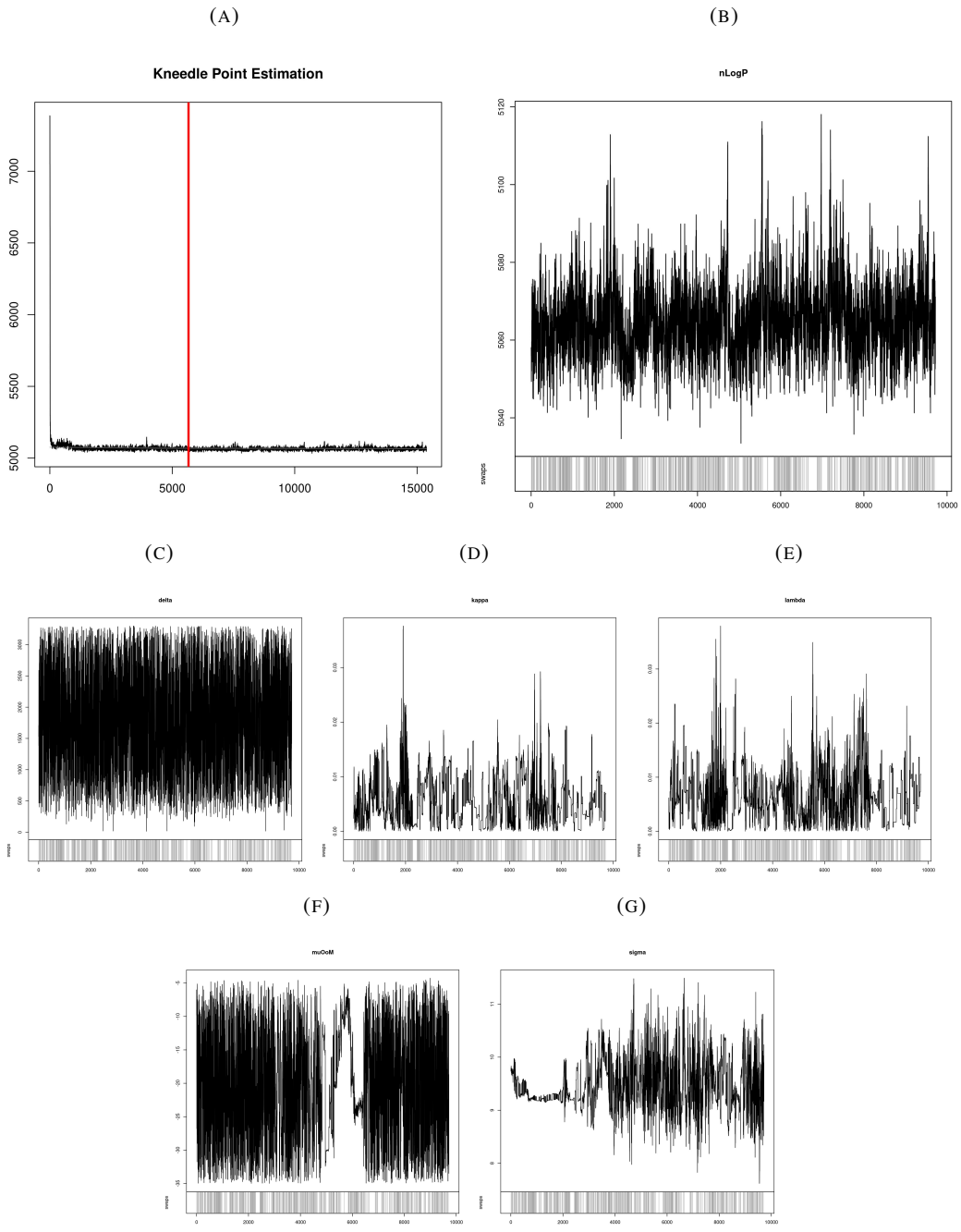

FIGURE A3.27: MCMC traces for -tau-upsilon replicate 1. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

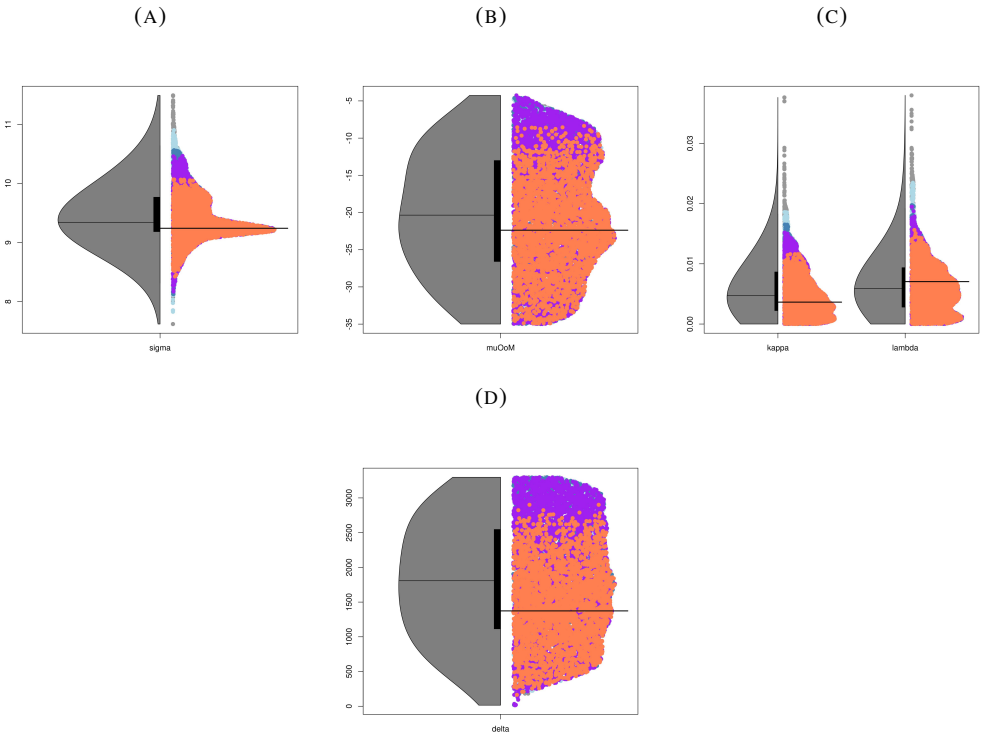

FIGURE A3.28: Density plots for -tau-upsilon replicate 1. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

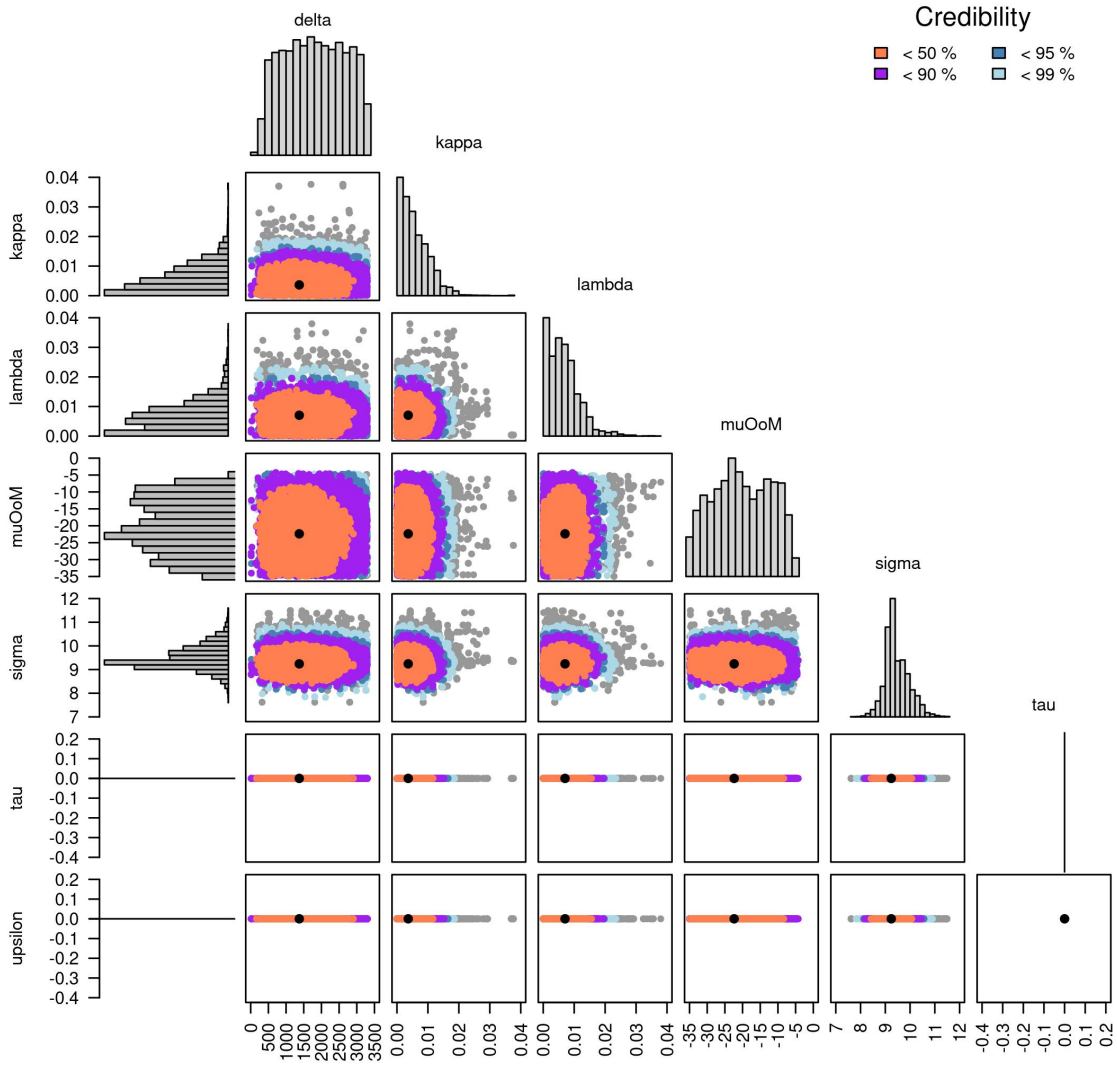

FIGURE A3.29: A visual summary of the posterior distribution estimate by ABC of an evolutionary model which assumes acstab and CMER length are independent of eachother (-tau-upsilon replicate 1).

A3.11 StepwiseOU-tau-upsilon-2

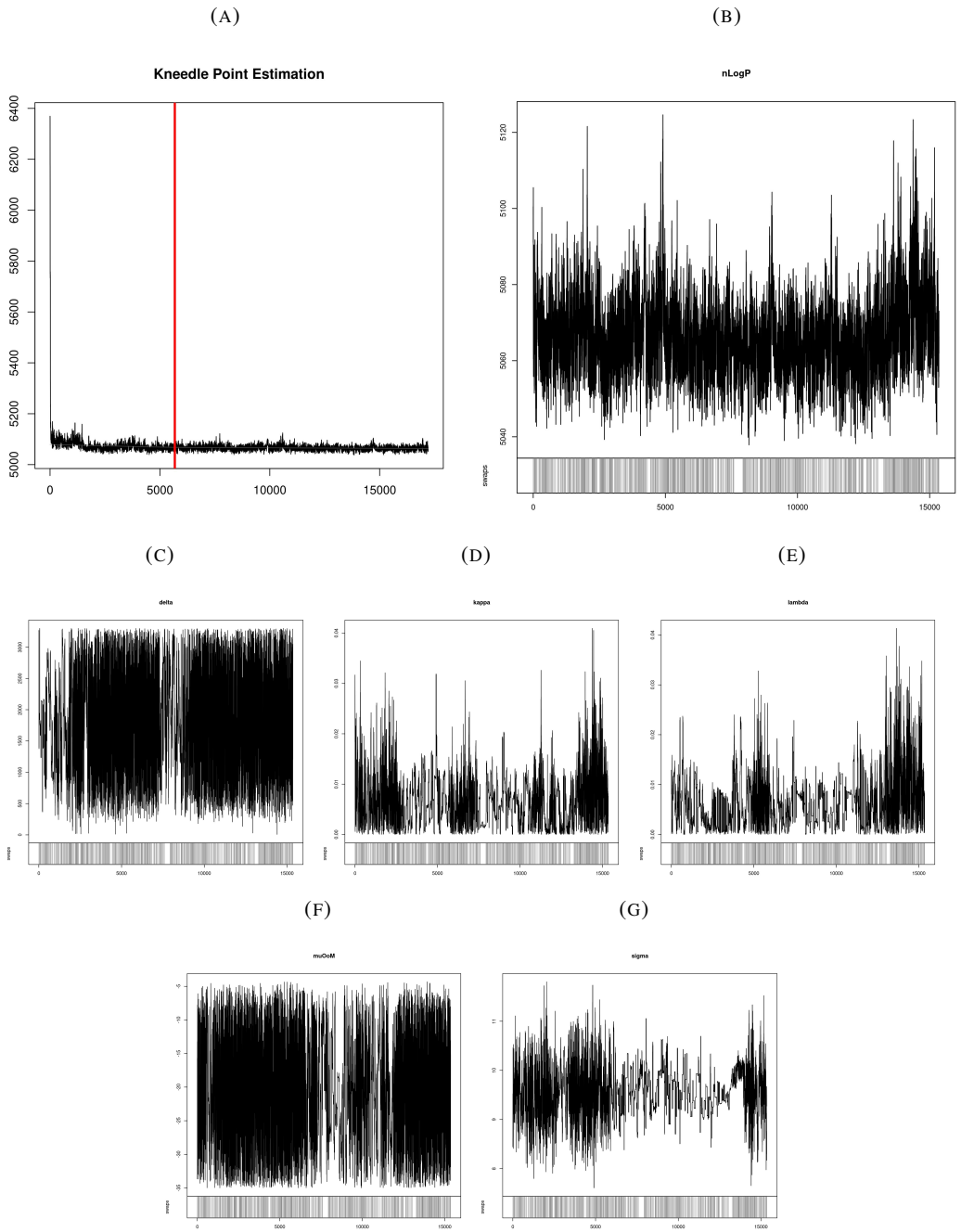

FIGURE A3.30: MCMC traces for -tau-upsilon replicate 2. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

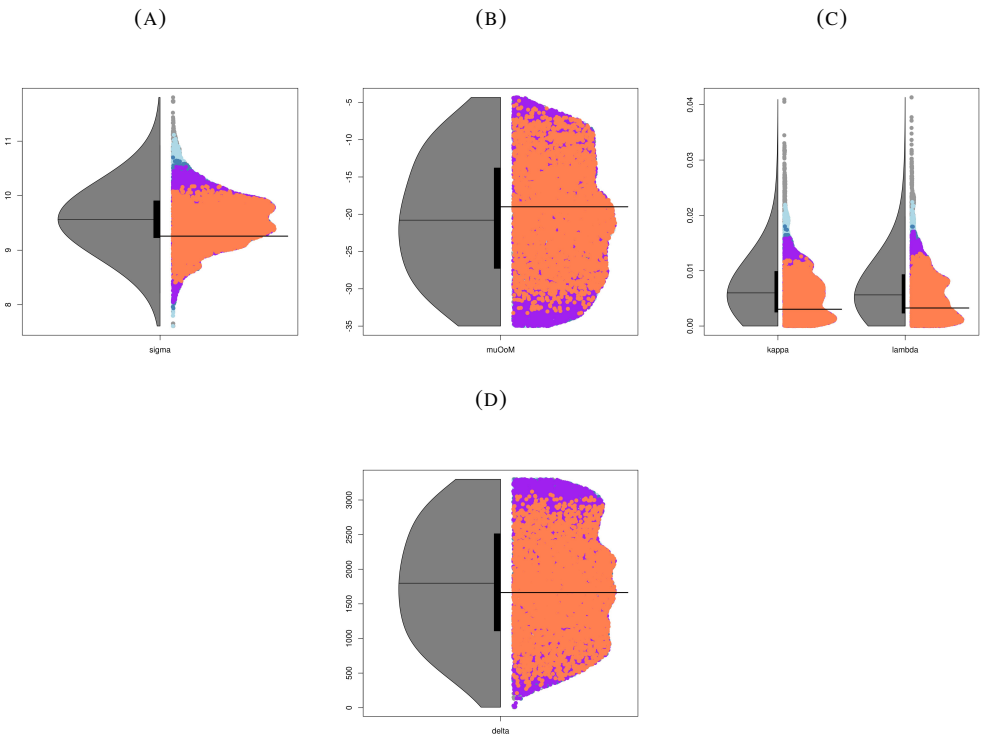

FIGURE A3.31: Density plots for -tau-upsilon replicate 2. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

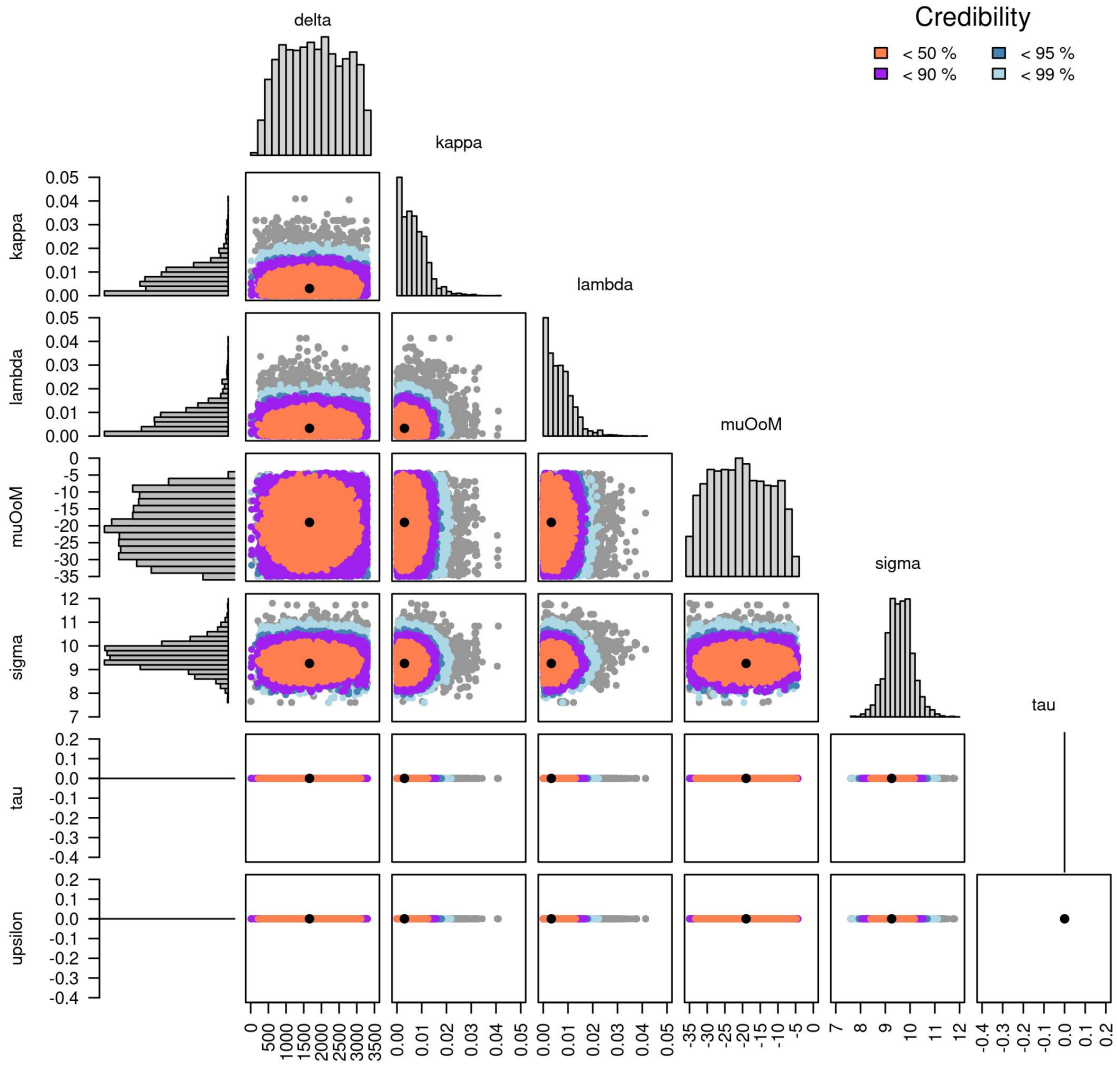

FIGURE A3.32: A visual summary of the posterior distribution estimate by ABC of an evolutionary model which assumes acstab and CMER length are independent of eachother (-tau-upsilon replicate 2).

A3.12 StepwiseOU-tau-upsilon-3

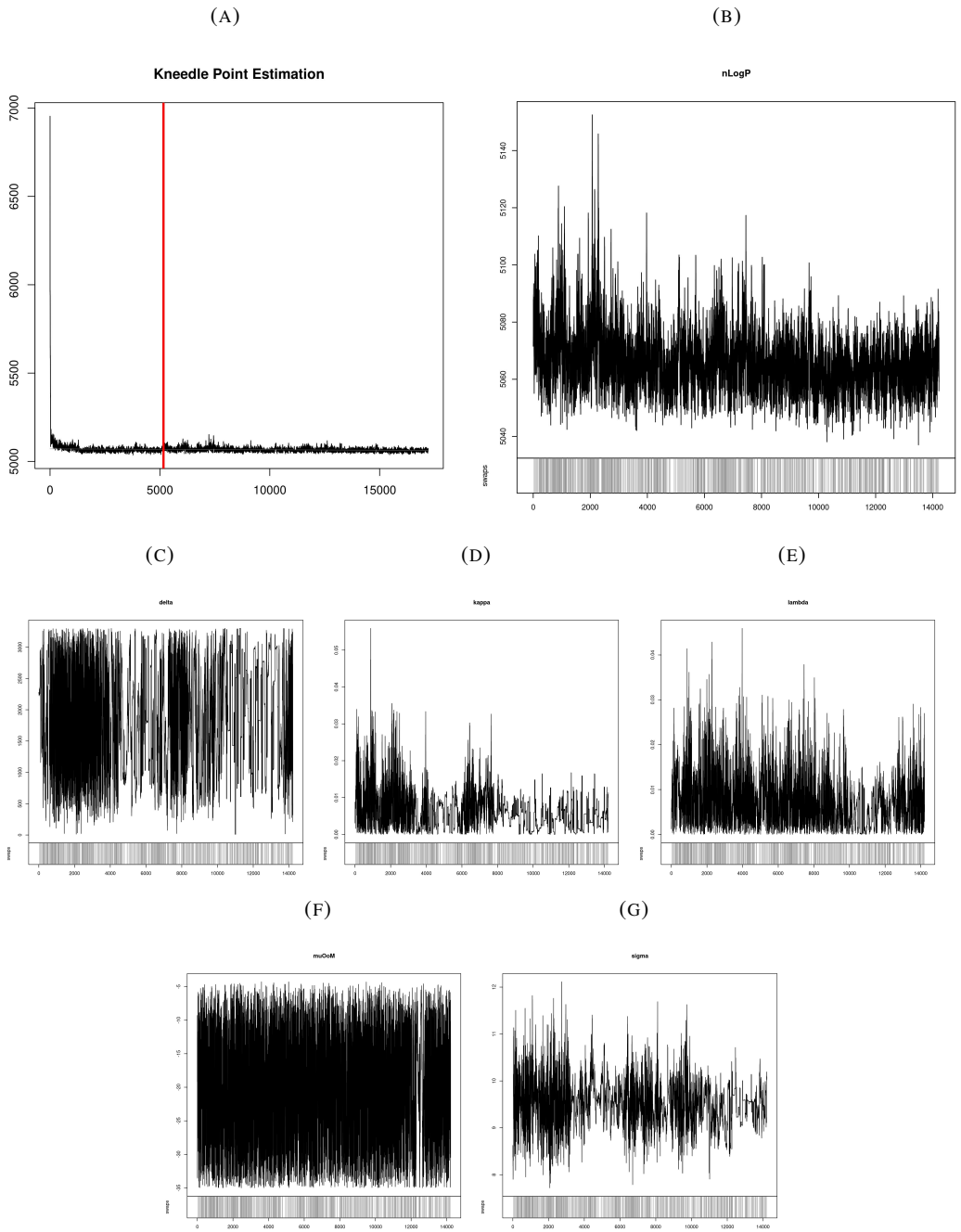

FIGURE A3.33: MCMC traces for -tau-upsilon replicate 3. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

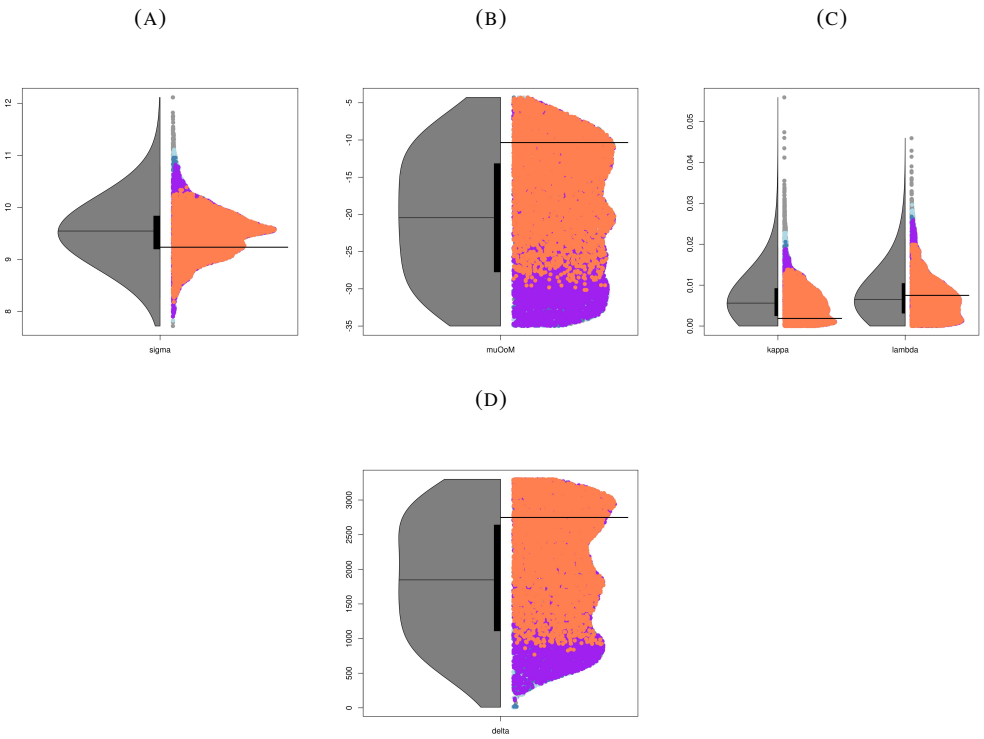

FIGURE A3.34: Density plots for -tau-upsilon replicate 3. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

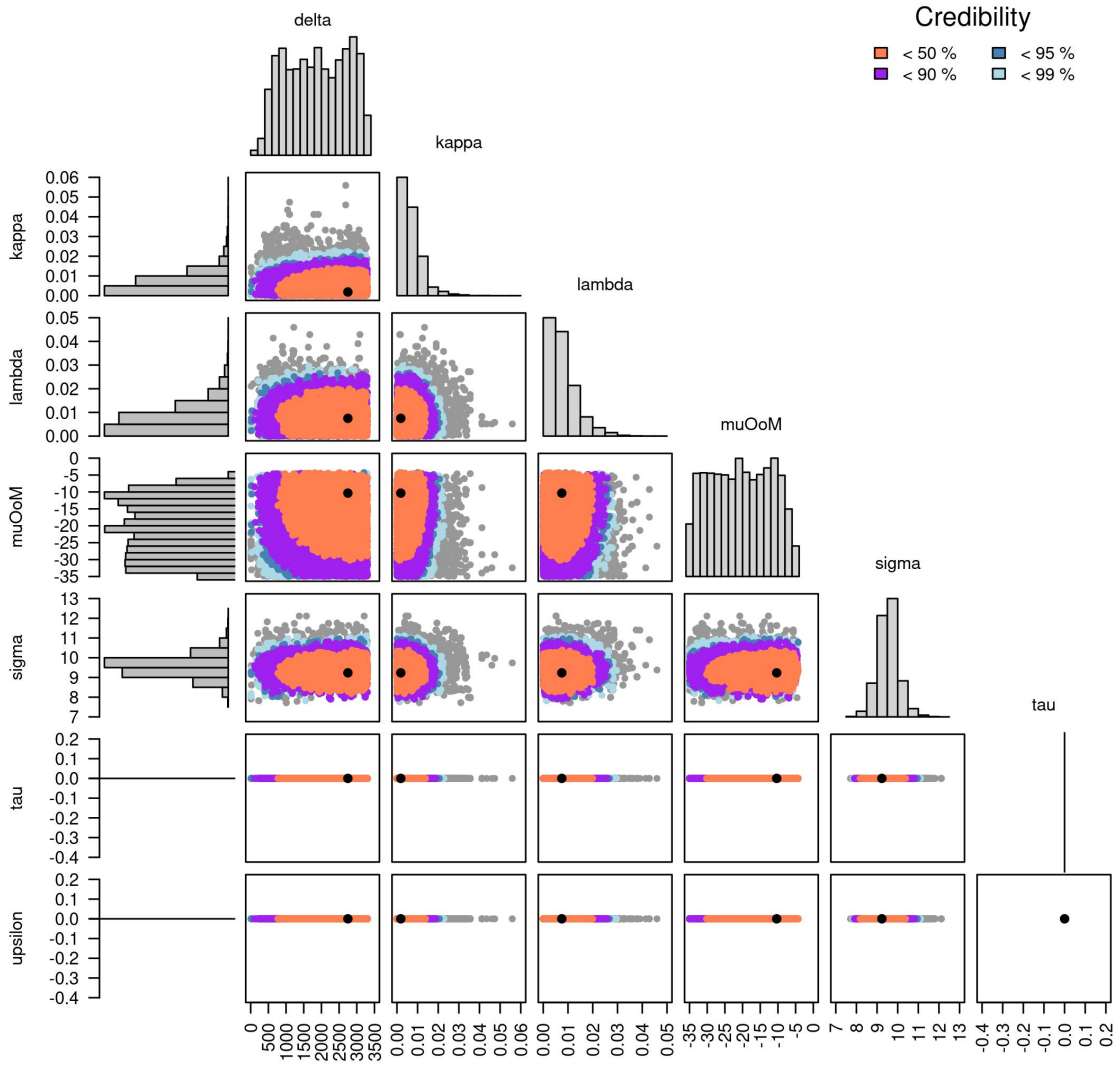

FIGURE A3.35: A visual summary of the posterior distribution estimate by ABC of an evolutionary model which assumes acstab and CMER length are independent of eachother (-tau-upsilon replicate 3).

A3.13 StepwiseOU-eqIndel-1

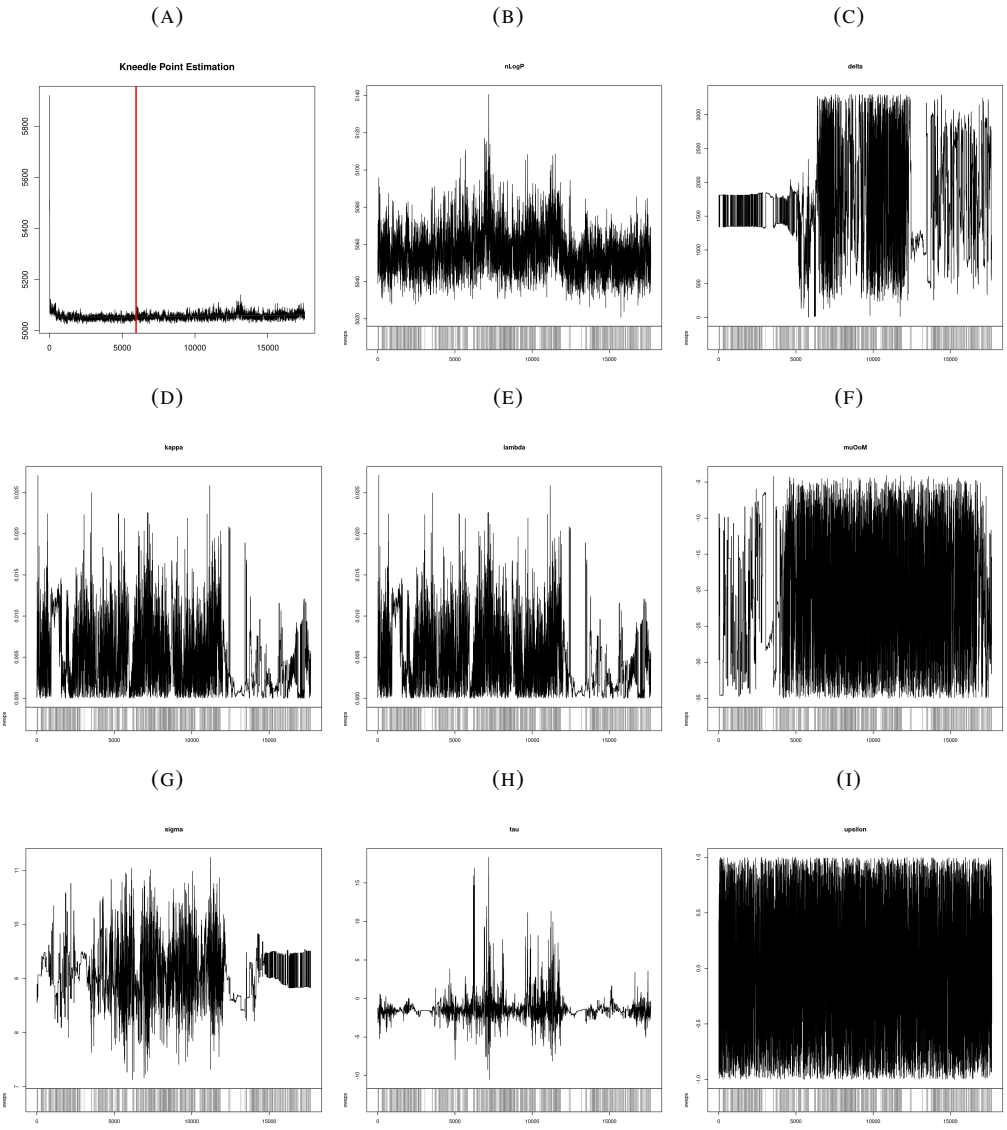

FIGURE A3.36: MCMC traces for -eqIndel1. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

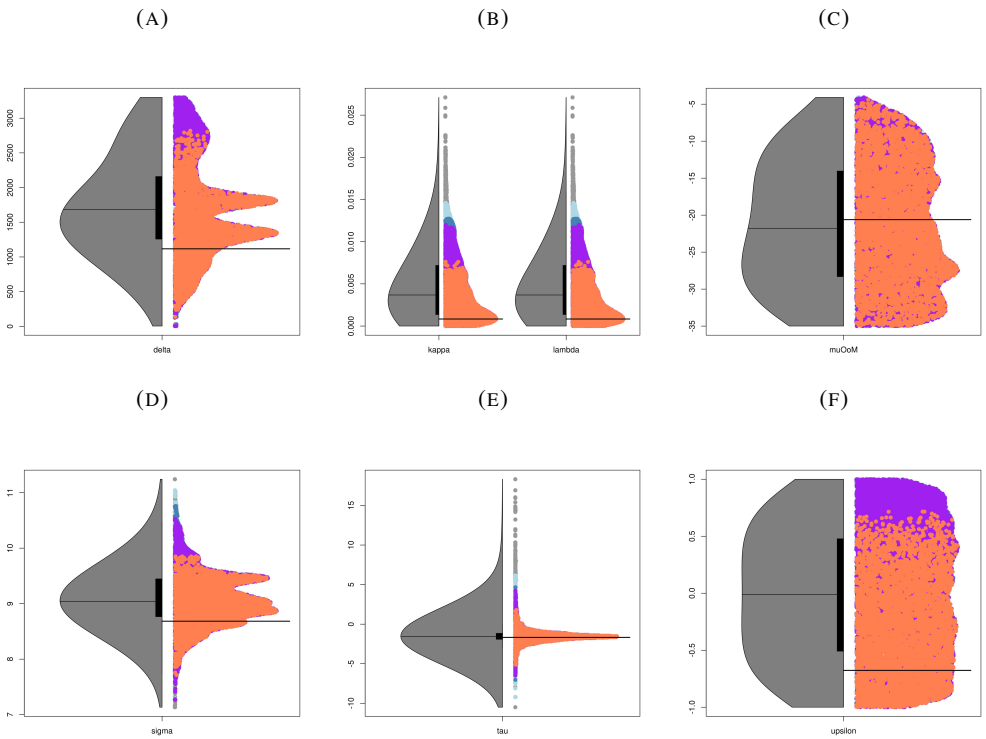

FIGURE A3.37: Density plots for -eqIndel1. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

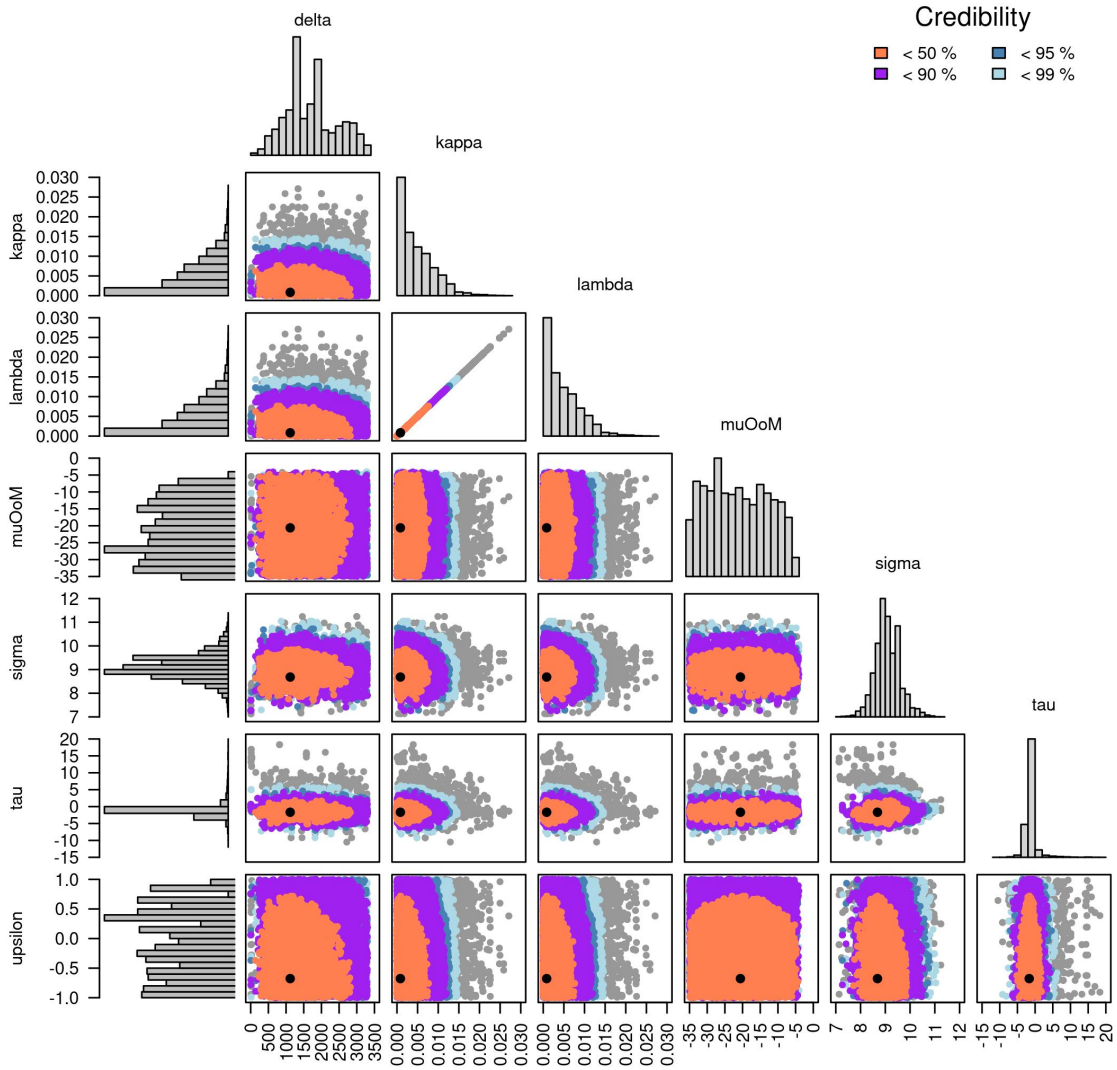

FIGURE A3.38: A visual summary of the posterior distribution estimate by ABC of an evolutionary model which assumes insertion and deletion rates are equal (-eqIndel1).

A3.14 StepwiseOU-upsilon-eqIndel-1

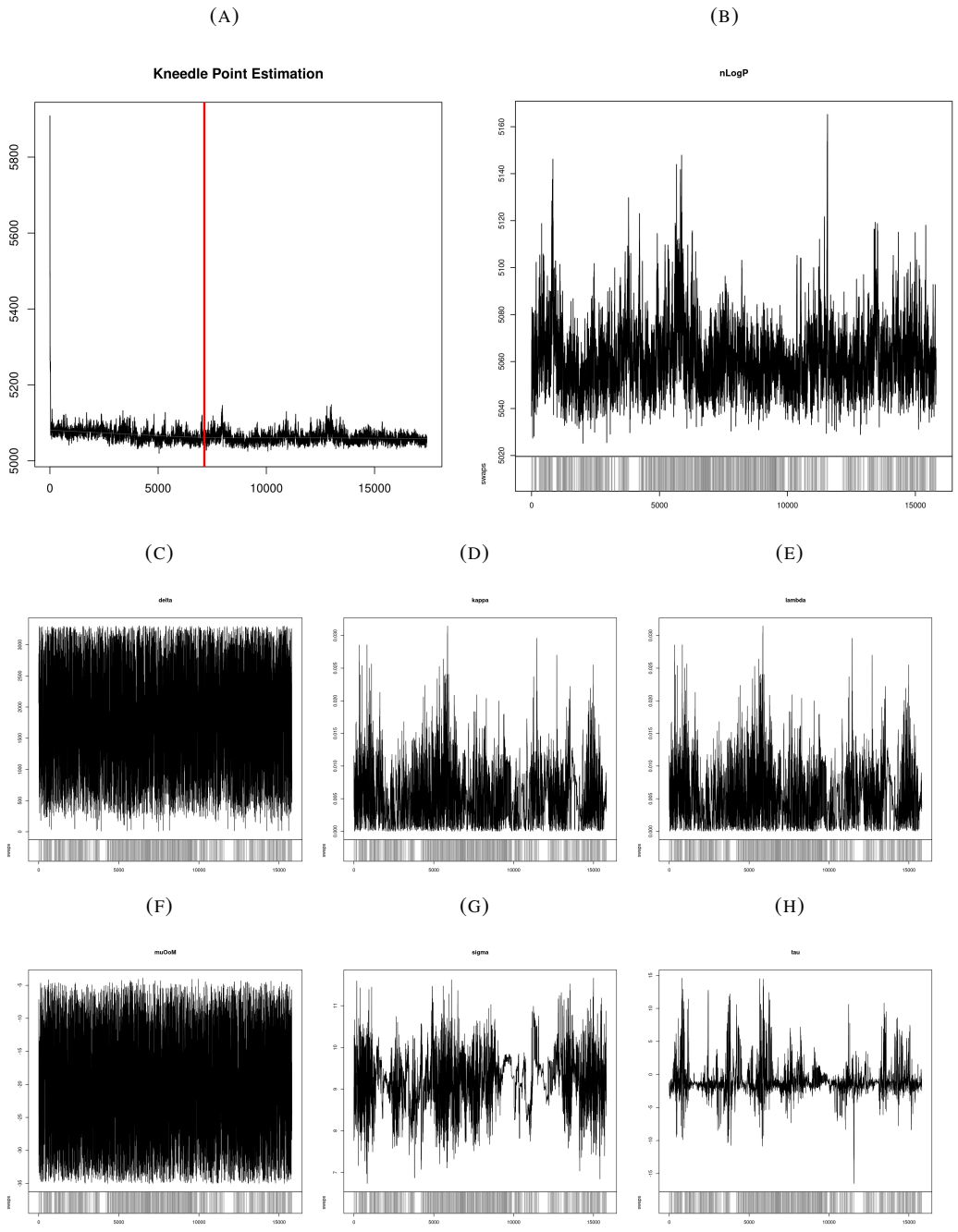

FIGURE A3.39: MCMC traces for -upsilon-eqIndel1. a and b show the entire trace for negative log likelihood and the trace after removing the burn-in. Remaining plots show the traces for the model parameters, after burning. Grey track along the bottom indicates where the chain swapped between the main chain and a heated chain.

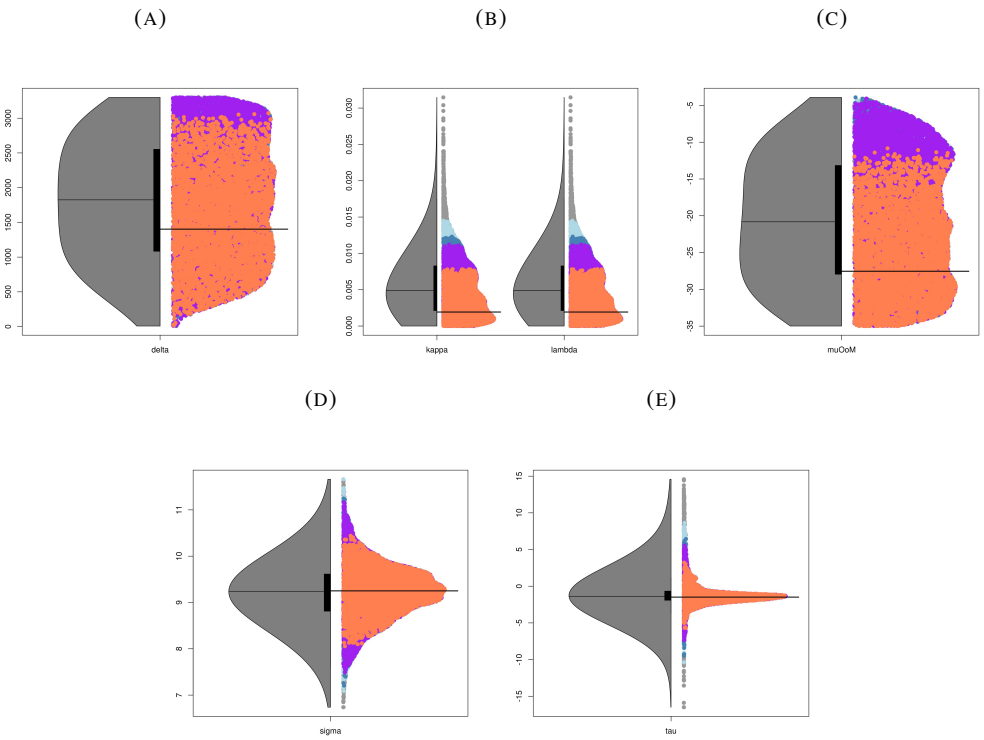

FIGURE A3.40: Density plots for -upsilon-eqIndel1. Each plot show the univariate posterior density for each modelling parameter. On the left is a purely univariate, high bandwidth density with the median marked. On the right is a projection of the multivariate density with the multivariate mode marked. Colours indicate the tightest credibility region each sample resides in. Respectively, 50, 90, 95, and 99% correspond to coral, purple, dark blue, and light blue.

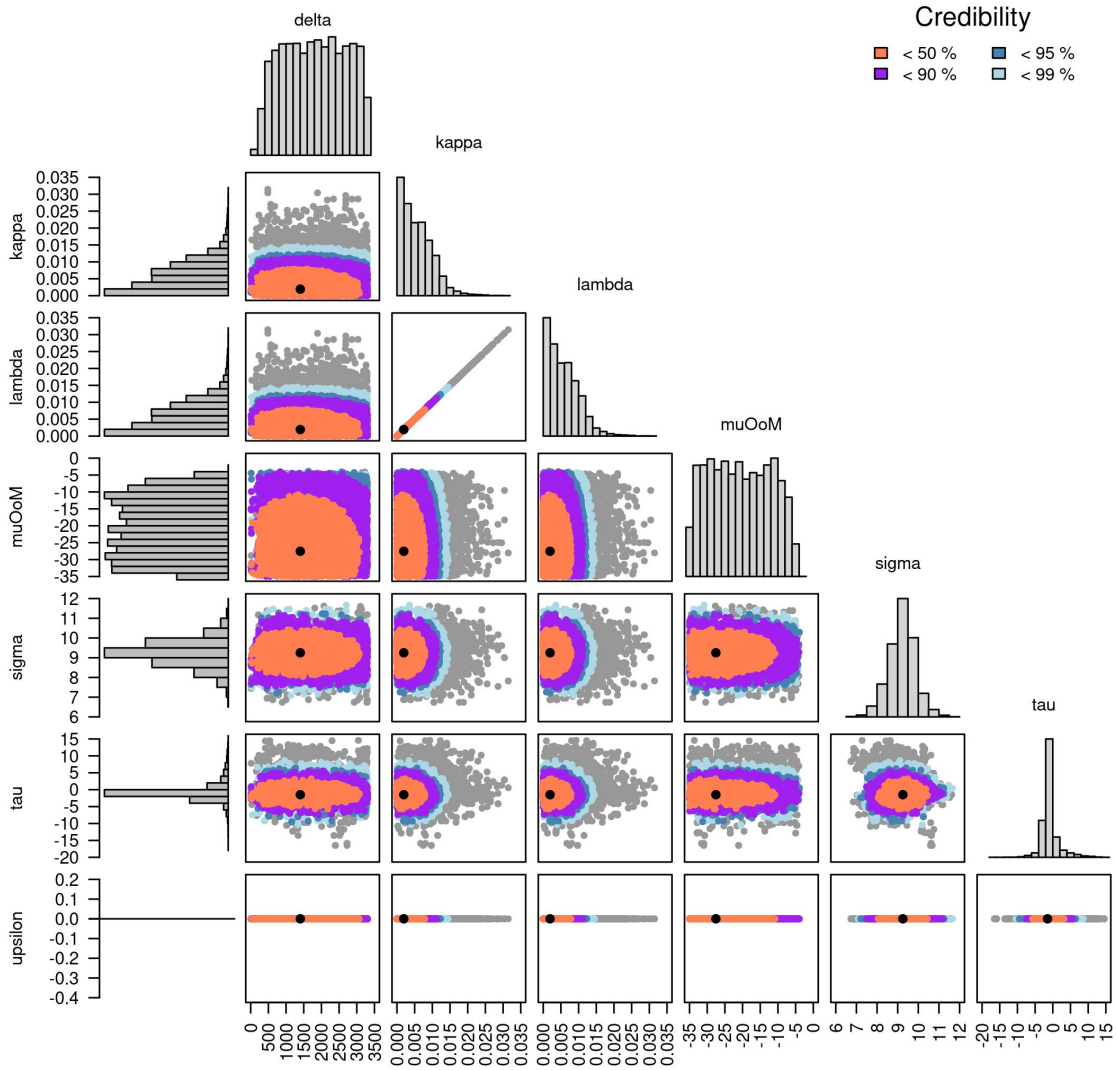

FIGURE A3.41: A visual summary of the posterior distribution estimate by ABC of an evolutionary model which assumes insertion and deletion rates are equal and assumes CMER length is eindependent of TAb (-upsilon-eqIndel1).

Appendix D

Chapter 8 Supplement

TABLE A4.2: Bacteria included in the design dataset for the Sepsis probe set

Taxa	PATRIC ID (or other ID)					
Achromobacter xylosoxidans	562971.11	85698.16	85698.28	85698.48	85698.49	85698.50
Acinetobacter baumannii	1096995.4	1096996.4	1096997.4	1413216.3	1455315.5	405416.6
	470.1288	470.1294	470.1295	470.1310	470.1311	470.1345
	470.1375	470.1405	470.1574	470.1575	470.1576	470.1579
	470.1763	470.1864	470.1865	470.1866	470.1867	470.1869
	470.2026	470.2122	470.2668	470.2669	470.2670	470.2671
	470.2672	470.2673	470.2674	470.2675	470.2908	470.2911
	470.2912	470.2913	470.2928	470.2929	470.2930	470.2931
	470.3044	470.3106	470.3347	470.3348	470.3349	470.3350
	470.3351	470.3352	470.3353	470.3354	470.3355	470.3356
	470.3357	470.3358	470.3362	470.3774	470.4258	470.4273
	470.4286	470.4287	470.4288	470.4289	470.773	470.775
	480119.5	557600.4				
Acinetobacter nosocomialis	106654.21	106654.48				
Bacillus anthracis	1392.86	1392.87	1392.92	1452727.3	592021.13	768494.3
Bordetella bronchiseptica	518.17					
Bordetella hinzii	103855.14	103855.15				
Bordetella holmesii	35814.12					
Bordetella parapertussis	257311.4	519.5				
Bordetella pertussis	257313.5	520.171	520.172	520.173	520.174	520.175
	520.176	520.177	520.178	520.179	520.180	520.181
	520.336	520.337	520.347	520.360	520.361	520.362
	520.363	520.364	520.365	520.366	520.367	520.368
	520.369	520.370	520.371	520.372	520.373	520.374
	520.375	520.376	520.377	520.378	520.379	520.380
	520.381	520.382	520.383	520.384	520.385	520.386
	520.387	520.388	520.389	520.390	520.391	520.392
	520.393	520.405	520.408	520.412	520.414	520.415
	520.416	520.417	520.420	520.421	520.422	520.423
	520.424	520.425	520.426	520.427	520.459	520.460
	520.461	520.462	520.505	520.506	520.507	520.508
	520.509	520.510	520.511	520.512	520.513	520.514
	520.515	520.516	520.517	520.518	520.519	520.520
	520.521	520.522	520.523	520.524	520.525	520.526
	520.527	520.528	520.529	520.530	520.531	520.532
	520.533	520.534	520.535	520.536	520.537	520.538
	520.539	520.540	520.541	520.542	520.543	520.544
	520.545	520.546	520.547	520.548	520.549	520.550
	520.551	520.552	520.553	520.554	520.555	520.556
	520.557	520.558	520.559	520.560	520.561	520.562
	520.563	520.564	520.565	520.566	520.567	520.568
	520.569	520.570	520.571	520.572	520.573	520.574
	520.575	520.576	520.577	520.578	520.579	520.580
	520.581	520.582	520.583	520.584	520.585	520.586
	520.587	520.588	520.589	520.590	520.591	520.592
	520.593	520.594	520.595	520.596	520.597	520.598
	520.599	520.600	520.601	520.602	520.603	520.604
	520.605	520.606	520.607	520.608	520.609	520.610
	520.611	520.612	520.613	520.614	520.615	520.616
	520.617	520.618	520.619	520.620	520.621	520.622
	520.623	520.624	520.625	520.626	520.627	520.628
	520.629	520.630	520.631	520.632	520.633	520.634
	520.635	520.636	520.637	520.638	520.639	520.640
	520.641	520.642	520.643	520.644	520.645	520.646
	520.647	520.648	520.649	520.650	520.651	520.652
	520.653	520.654	520.655	520.656	520.657	520.658
	520.659	520.660	520.661	520.662	520.663	520.664

Taxa	PATRIC ID (or other ID)					
	520.665	520.666	520.667	520.668	520.669	520.670
	520.671	520.672	520.673	520.674	520.675	520.676
	520.677	520.678	520.679	520.680	520.681	520.682
	520.683	520.684	520.685	520.686	520.687	520.688
	520.689	520.690	520.691	520.692	520.693	520.694
	520.695	520.696	520.697	520.698	520.699	520.700
	520.701	520.702	520.703	520.704	520.705	520.706
	520.707	520.708	520.709	520.710	520.711	520.712
	520.713	520.714	520.715	520.716	520.717	520.718
	520.719	520.720	520.721	520.722	520.729	520.730
	520.731	520.737	520.738	520.739	520.740	520.742
	520.743	520.744	520.745	520.746	520.747	520.748
	520.749	520.750	520.751	520.752	520.753	520.754
	520.755	520.756	520.757	520.758	520.761	520.769
	520.770	520.771	520.772	520.773	520.774	520.775
	520.776	520.777	520.778	520.779		
Borrelia burgdorferi	139.88	139.89	139.96			
Burkholderia ambifaria	398577.6					
Burkholderia cenocepacia	331271.8	95486.299	95486.300	95486.301	95486.302	95486.303
	95486.315	95486.74	95486.85	95486.86		
Burkholderia cepacia	292.26					
Burkholderia dolosa	1385930.3					
Burkholderia latens	488446.5					
Burkholderia mallei	13373.17	13373.25	13373.58	13373.65	243160.12	
Burkholderia multivorans	87883.77	985079.5	ATCC_17616			
Burkholderia oklahomensis	441162.11					
Burkholderia pseudomalli	1241583.3	1249468.3	1249469.3	1249470.6	1249475.3	1249658.4
	1249659.3	1306417.5	1306419.4	1306420.5	1306421.4	1335307.3
	1435985.3	1435994.3	1437000.3	272560.51	272560.6	28450.155
	28450.156	28450.157	28450.244	28450.245	28450.246	28450.247
	28450.248	28450.362	28450.365	28450.377	28450.378	28450.382
	28450.383	28450.384	28450.385	28450.644	28450.645	28450.646
	28450.647	28450.648	28450.651	28450.652	28450.653	28450.654
	28450.655	28450.656	28450.657	28450.659	28450.660	28450.662
	28450.663	28450.666	28450.667	28450.711	28450.713	28450.714
	28450.83	28450.90	360118.8	884204.3	884204.6	
Burkholderia stabilis	95485.5					
Burkholderia thailandensis	1241582.3	1249663.3	57975.9			
Burkholderia vietnamiensis	60552.42					
Campylobacter coli	1358410.3	195.282				
Campylobacter fetus	360106.6					
Campylobacter jejuni	1338035.3	1340842.3	1349827.3	1357994.3	1380767.3	1380768.3
	1383068.3	192222.6	197.11692	197.11720	197.11723	197.11724
	197.11725	197.11726	197.11830	197.5229	197.5260	197.5261
	197.5262	197.5263	197.5264	197.5265	197.5266	197.5267
	197.5268	197.5269	197.5270	197.5271	197.5272	197.5273
	197.5274	197.5275	197.5276	197.5277	197.5278	197.5279
	197.5280	197.5281	197.5282	197.5283	197.5284	197.5285
	197.5286	197.5287	197.5288	197.5289	197.5290	197.5291
	197.5292	197.5293	197.5294	197.5295	197.5296	197.5297
	197.5298	197.5299	197.5300	197.5301	197.5302	197.5303
	197.5304	197.5305	197.5306	197.5307	197.5308	197.5309
	197.5310	197.5311	197.5312	197.5313	197.5314	197.5323
	32022.145	32022.146	32022.147	32022.148	32022.158	32022.159
	32022.293	32022.294	407148.6	645464.3		
Chlamydia pneumoniae	115711.10	138677.3	182082.4	83558.13	83558.14	83558.15
	83558.16	83558.17	83558.18	83558.19		
Chlamydia trachomatis	1260222.3	1260223.3	813.101	813.102	813.103	813.104
	813.105	813.126	813.135	813.136	813.137	813.138
	813.139	813.140	813.141	813.142	813.143	813.144
	813.145	813.146	813.147	813.148	813.31	813.32
	813.33	813.34	813.35	813.36	813.37	813.38
	813.39	813.40	813.41	813.44	813.45	813.46
	813.88	813.89	813.90	813.91	813.92	813.93
	813.94	887712.6				
Clostridioides difficile	645462.3	645463.3	699034.5	699035.5	699036.5	699037.5
Clostridium botulinum	1408283.5	1491.354	1491.355	1491.356	1491.662	515621.3
	536232.3					
Clostridium perfringens	1502.206	1502.338				
Clostridium tetani	1231072.4	212717.8				
Enterobacter aerogenes	548.115	548.154	548.155	548.156	935296.3	
Enterobacter cloacae	1333849.3	1333850.3	1333851.3	1812934.3	1812935.7	550.1143
	550.1232	550.1235	550.153	550.154	550.155	550.252
	550.253	550.254	550.388	550.486	550.487	550.488
	550.622	550.642	718254.4			
Enterococcus faecalis	1287066.3	1351.496	1351.556	226185.9	936153.3	
Enterococcus faecium	1155766.14	1305849.3	1352.1358	1352.1406	1352.1408	1352.1610
	1352.1611	1352.1641	1352.1643	1352.1644	1352.1645	1352.1708
	1352.1709	1352.1711	1352.1712	1352.1713	1352.1714	1352.1716
	1352.1760	1352.1761	1352.2590	1352.2704	1352.2735	1352.2736
	1352.2737	1352.2738	1352.657	1352.674	1352.770	1352.804
Escherichia albertii	1440052.3					
Escherichia coli	1038927.31	1045010.15	1045010.16	1045010.21	1048254.4	1248902.3

Taxa	PATRIC ID (or other ID)					
	1248915.3	1355100.3	1355101.3	168807.6	199310.4	2048778.3
	316435.10	331112.6	362663.9	364106.8	409438.11	431946.3
	498388.3	511145.208	562.10006	562.10007	562.10008	562.10009
	562.10826	562.10827	562.10887	562.10890	562.11294	562.11317
	562.11318	562.11327	562.11469	562.11479	562.12823	562.12824
	562.12825	562.12826	562.12907	562.12941	562.13544	562.13545
	562.13546	562.13547	562.13550	562.13551	562.13552	562.13553
	562.13554	562.13557	562.13559	562.14013	562.14016	562.15191
	562.15192	562.15193	562.15194	562.15195	562.15196	562.15197
	562.15198	562.15199	562.15200	562.15201	562.15202	562.15203
	562.15204	562.15205	562.15206	562.15208	562.15209	562.15210
	562.16428	562.16465	562.16466	562.16467	562.16499	562.17359
	562.17710	562.17711	562.17734	562.17735	562.18998	562.19158
	562.19192	562.19193	562.19574	562.20516	562.20517	562.22306
	562.22307	562.22323	562.22326	562.22331	562.22333	562.22350
	562.5740	562.6958	562.7071	562.7228	562.7235	562.7257
	562.7300	562.7382	562.7564	562.7585	562.7586	562.7587
	562.7588	562.7622	562.7692	562.7736	562.7943	562.7955
	562.7956	562.8472	562.8489	562.8507	562.8517	562.8518
	562.9096	562.9097	562.9098	562.9099	562.9316	569579.3
	574521.7	591946.4	701177.3	741093.3	83334.221	941323.4
	BW25113					
Escherichia fergusonii	564.14					
Francisella tularensis	119856.8	1341656.4	1386968.3	177416.18	263.138	263.91
	264.23	418136.12				
Genus Species	PIDOID					
Haemophilus influenzae	1232659.5	1295140.4	262727.7	262728.6	281310.6	374930.9
	374931.10	727.1050	727.305	727.532	727.533	727.534
	727.972					
Haemophilus parainfluenzae	862965.3					
Helicobacter pylori	102611.3	1055527.3	1055531.3	1055532.3	1127122.3	1163739.3
	1163740.3	1163741.3	1163742.3	1163743.3	1311573.5	1321939.6
	1321940.5	1352356.3	1382920.3	1382921.3	1382922.3	1382923.3
	1382924.3	1382925.3	1382926.3	1382927.3	1391726.3	1407462.3
	1407463.3	1431450.3	210.1380	210.1381	210.1916	210.1917
	210.1918	210.1969	210.1970	210.2070	210.2071	210.2072
	210.2437	210.2438	210.2439	210.2440	210.2441	210.2442
	210.2712	210.2901	210.2902	210.2903	210.2904	210.2905
	210.2906	210.2907	210.2908	210.2909	210.2910	210.2911
	210.2912	210.2913	210.2914	210.2915	210.2916	210.2917
	210.2918	210.2919	210.2920	210.2921	210.2922	210.2923
	210.2924	210.2925	210.2926	210.2927	210.2928	210.2929
	290847.5	357544.13	570508.6	585538.3	765963.4	794851.3
	85962.8	85963.30	85963.7			
Klebsiella oxytoca	1191061.3	1333852.3	571.108	571.110	571.199	571.213
	571.33	571.62	571.63			
Klebsiella pneumoniae	1123862.3	1225181.3	1244085.3	1263871.10	1328325.3	1380908.3
	1392499.4	1420012.3	1420013.3	1463165.4	272620.9	573.12457
	573.12458	573.12459	573.12460	573.12792	573.12796	573.1352
	573.1358	573.1361	573.1369	573.1374	573.14839	573.1497
	573.14988	573.1499	573.15001	573.1500	573.15234	573.15309
	573.15310	573.15312	573.15317	573.15318	573.15319	573.15320
	573.15321	573.15329	573.15365	573.15379	573.1761	573.1921
	573.1922	573.1923	573.1924	573.1961	573.4021	573.4022
	573.4026	573.4029	573.4030	573.4031	573.4032	573.4038
	573.5649	573.5783	573.5786	573.6972	573.6973	573.6974
	573.6995	573.6996	573.6997	573.7004	573.7015	573.7020
	573.7226	573.7238	573.7239	573.7240	573.7241	573.7242
	573.7243	573.7244	573.7245	573.7246	573.7247	573.7248
	573.7249	573.7351	573.7352	573.7353	573.7354	573.7355
	573.7356	573.7357	573.7358	573.7359	573.7360	573.9629
	573.9630	573.9631	573.9632	573.9643	573.9645	573.9779
	72407.108	72407.109	72407.110	72407.122	72407.123	72407.396
	72407.626	72407.693	72407.733	72407.75	72407.77	72407.78
	72407.79	72407.81	N25C9			
Legionella pneumophila	1199191.3	400673.7	423212.4	446.556	446.588	446.637
	446.638	446.639	446.640	446.641	446.650	446.651
	91891.37					
Listeria monocytogenes	1126011.4	1196159.3	1196160.3	1196162.3	1196163.3	1196166.3
	1196167.3	1196169.3	1196171.3	1196172.3	1196174.3	1196176.3
	1196177.3	1196178.3	1196179.3	1196180.3	1196181.4	1196184.3
	1196186.3	1196189.3	1196190.3	1196191.3	1639.1083	1639.1084
	1639.1085	1639.1160	1639.1207	1639.1511	1639.2298	1639.2421
	1639.2613	1639.2624	1639.2641	1639.756	1639.758	1639.956
	1639.965	1639.988	1639.989	1639.990	1639.991	1639.992
	1639.993	265669.9				
Micrococcus luteus	1270.31	465515.4				
Moraxella catarrhalis	1236608.7	480.268	480.269	749219.3		
Mycobacterium abscessus	1132508.3	1185650.3	1185650.4	1299325.4	1303024.3	1962118.4
	319705.13	319705.14	319705.15	36809.213	36809.363	36809.364
	36809.365	36809.366	36809.367	36809.368	36809.369	36809.370
	36809.371	36809.372	36809.373	36809.374	36809.380	36809.381
	36809.382	36809.383	36809.384			

Taxa	PATRIC ID (or other ID)					
Mycobacterium bovis	33892.16	561275.19				
Mycobacterium tuberculosis	1249615.4	1262525.3	1262526.3	1262529.3	1266446.3	1295765.3
	1344043.3	1344044.3	1344045.3	1344046.3	1344047.3	1344048.3
	1346765.3	1346766.3	1346767.3	1346768.3	1346770.3	1346771.3
	1346772.3	1346773.3	1346774.3	1346775.3	1346776.3	1346777.3
	1346778.3	1346779.3	1346780.3	1346781.3	1346782.3	1346783.3
	1346784.3	1346785.3	1346786.3	1346787.3	1346788.3	1352570.3
	1352571.3	1352572.3	1352573.3	1352574.3	1352575.3	1352576.3
	1352577.3	1352578.3	1352579.3	1352580.3	1352581.3	1352582.3
	1352583.3	1352584.3	1352585.3	1352586.3	1352587.3	1352588.3
	1352589.3	1352590.3	1352591.3	1352592.3	1352593.3	1352594.3
	1352595.3	1352596.3	1352597.3	1352598.3	1352599.3	1352600.3
	1773.2409	1773.249	1773.31	1773.349	1773.522	1773.523
	1773.554	1773.5931	1773.5970	1773.5971	1773.6029	1773.6081
	1773.6103	1773.6106	1773.6113	1773.6127	1773.6130	1773.6141
	1773.6150	1773.6153	1773.6158	1773.6164	1773.6172	1773.6185
	1773.6249	1773.6273	1773.6322	1773.6471	1773.6476	1773.6486
	1773.6607	1773.6623	1773.6709	1773.6725	1773.6742	1773.6743
	1773.6744	1773.6745	1773.6746	1773.6747	1773.6748	1773.8213
	1773.8214	1773.8215	1773.8216	1773.8217	1773.8218	1773.836
	1773.8399	1773.8650	1773.8651	1773.8652	1773.8653	1773.8657
	1773.8658	1773.8659	1773.8660	1773.8661	1773.8662	1773.8663
	1773.8664	1773.8665	1773.8666	1773.8667	1773.8668	1773.8669
	1773.8670	1773.8671	1773.8672	1773.8673	1773.8674	1773.8675
	1773.8676	1773.8677	1773.8678	1773.8679	1773.8680	1773.8681
	1773.8682	1773.8683	1773.8684	1773.8685	1773.8686	1773.8687
	1773.8688	1773.8689	1773.8690	1773.8691	1773.8692	1773.8693
	1773.8694	1773.8695	1773.8696	1773.8697	1773.8698	1773.8699
	1773.8700	1773.8701	1773.8702	1773.8703	1773.8704	1773.8705
	1773.8706	1773.8707	1773.8708	1773.8709	1773.8710	1773.8711
	1773.8712	1773.8713	1773.8714	1773.8715	1773.8716	1773.8717
	1773.8718	1773.8719	1773.8720	1773.8721	1773.8722	1773.8723
	1773.8724	336982.7	419947.17	419947.8	443150.4	478434.4
	83331.22	83332.12	83332.293			
Mycoplasma pneumoniae	1112856.4	1238993.3	1263756.3	1263757.3	1263758.3	1263760.3
	1263761.3	1263762.3	1263763.3	1263835.4	1280940.3	1441379.3
	2104.102	2104.136	2104.137	2104.162	2104.163	2104.164
	2104.165	2104.166	2104.167	2104.168	2104.169	2104.189
	2104.190	272634.6	722438.5			
Neisseria gonorrhoeae	242231.10	485.508	485.509	485.510	485.879	485.881
	521006.8	940296.3				
Neisseria meningitidis	1095685.4	1386087.3	272831.7	487.1096	487.1102	487.1105
	487.1106	487.1114	487.1118	487.1119	487.1124	487.1125
	487.1129	487.1130	487.1134	487.1135	487.1136	487.1137
	487.1138	487.1140	487.1143	487.1144	487.1145	487.1146
	487.1147	487.1148	487.1149	487.1150	487.1151	487.1152
	487.1153	487.1154	487.1155	487.1156	487.1157	487.1158
	487.1159	487.1160	487.1161	487.1162	487.1163	487.1164
	487.1165	487.1166	487.1167	487.1168	487.1169	487.1170
	487.1171	487.1172	487.1173	487.1174	487.1175	487.1176
	487.1177	487.1203	487.1231	487.1300	487.1301	487.1302
	487.1303	487.1306	487.1548	487.1549	487.485	487.486
	487.487	487.488	487.517			
Proteus mirabilis	529507.6	584.106	584.140	584.217	584.218	584.82
	584.84	584.85				
Proteus vulgaris	585.12	585.15				
Pseudomonas aeruginosa	1089456.3	1093787.3	1280938.3	1340851.3	1352355.3	1356855.3
	1407059.3	1427342.3	208963.12	287.1309	287.1494	287.1773
	287.1997	287.2542	287.2548	287.2549	287.2550	287.2551
	287.2552	287.2553	287.2554	287.2555	287.2556	287.2557
	287.2558	287.2559	287.2560	287.2561	287.2562	287.2570
	287.2571	287.2578	287.2579	287.2580	287.2658	287.2665
	287.2692	287.2816	287.2899	287.2900	287.2901	287.2902
	287.2903	287.2904	287.2906	287.2907	287.2965	287.2966
	287.2967	287.2970	287.3104	287.3867	287.3868	287.3869
	287.3877	287.3878	287.4050	287.4051	287.4090	287.4091
	287.4092	287.4094	287.4464	287.4554	287.4566	381754.6
Rickettsia prowazekii	1290427.3	1290428.3	272947.5			
Salmonella enterica	1016998.12	1124936.4	1243585.3	1243619.3	1243621.3	1244111.3
	1244112.3	1244113.3	1244118.3	1244119.3	1244120.3	1267753.3
	1299044.4	1412451.3	1412452.3	1412592.3	1454583.3	1454592.3
	1454593.3	1454594.3	1454596.3	1454606.3	1454607.3	1454608.3
	1454610.3	1454611.3	1454612.3	1454627.3	1454640.3	1454641.3
	1454642.3	1454643.3	1454644.3	1454645.3	149539.316	220341.7
	28150.13	28901.1106	28901.2334	28901.2706	321314.9	54388.137
	54388.139	54388.5	54388.6	59201.158	59204.12	59204.13
	59204.14	600.8	600.9	611.112	611.119	611.120
	611.125	611.128	611.129	611.65	90105.125	90105.85
	90105.86	90370.2206	90370.929	90370.930	90371.1185	90371.1186
	90371.1187	90371.1940	90371.2143	90371.2552	90371.686	90371.883
	90371.903					
Serratia marcescens	615.102	615.105	615.109	615.142	615.364	615.516
	615.517	615.518	615.519	615.520	615.521	615.522

Taxa	PATRIC ID (or other ID)					
	615.523	615.524	615.525	615.526		
Shigella boydii	344609.11	621.85				
Shigella dysenteriae	754093.4					
Shigella flexneri	1435046.4	1617964.3	1935181.3	198214.7	374923.3	374923.6
	374923.7	42897.19				
Shigella sonnei	624.1184	624.1185	624.1272	624.1273	624.1274	624.1285
	624.635	624.637				
Staphylococcus aureus	1006543.3	1123523.3	1229492.3	1241616.6	1280.10152	1280.10341
	1280.10405	1280.10759	1280.11542	1280.11642	1280.11675	1280.11676
	1280.11677	1280.11678	1280.11679	1280.11680	1280.11681	1280.11682
	1280.11683	1280.11685	1280.11686	1280.11687	1280.11688	1280.11689
	1280.11690	1280.12234	1280.12239	1280.12240	1280.12241	1280.12242
	1280.12243	1280.12244	1280.12279	1280.12905	1280.12906	1280.12910
	1280.14567	1280.14568	1280.14569	1280.14571	1280.14573	1280.3349
	1280.3352	1280.3356	1280.3366	1280.3367	1280.3368	1280.3369
	1280.3371	1280.3378	1280.3522	1280.3537	1280.3566	1280.3568
	1280.3574	1280.3583	1280.3589	1280.4353	1280.4355	1280.4797
	1280.4800	1280.4803	1280.4805	1280.4808	1280.4809	1280.4810
	1280.4813	1280.4824	1280.4826	1280.4829	1280.4832	1280.4847
	1280.4848	1280.4849	1280.4850	1280.4851	1280.4852	1280.5046
	1280.5047	1280.5050	1280.5203	1280.5204	1280.5205	1280.5230
	1280.7175	1280.7176	1280.7177	1280.7178	1280.7179	1280.7180
	1280.7181	1280.7188	1280.7242	1280.7243	1280.7244	1280.7245
	1280.7246	1280.8872	1280.8873	1321369.3	1343064.4	1392476.3
	158879.11	196620.5	359786.13	359787.11	426430.24	46170.102
	46170.127	46170.148	46170.149	46170.154	46170.155	46170.181
	46170.182	46170.183	46170.244	46170.245	46170.246	46170.247
	46170.248	46170.249	46170.276	46170.277	46170.290	46170.303
	46170.304	46170.338	46170.86	IIDRC0017		
Staphylococcus epidermidis	1282.1164	1282.2148	1282.2244	1282.2268	1282.2269	176279.9
Staphylococcus lugdunensis	28035.20	28035.28	28035.29	28035.30	28035.8	28035.9
	698737.3					
Staphylococcus pseudointermedius	1266717.69	984892.3				
Staphylococcus saprophyticus	342451.11					
Stenotrophomonas maltophilia	40324.126	40324.127	40324.128	40324.225	40324.252	
Streptococcus agalactiae	1311.1246	1311.1314	1311.1315	1311.132	1311.133	1311.1349
	1311.1451	1311.1452	1311.1453	1311.1454	1311.1455	1311.1456
	1311.1457	1311.1458	1311.1459	1311.1504	1311.246	1311.337
	1311.349	1311.350	1311.625	1311.694	1311.695	1311.983
	1427374.3	211110.3				
Streptococcus anginosus	1328.20	1328.24				
Streptococcus constellatus	696216.3	862968.3	C1050			
Streptococcus intermedius	1338.26	1338.30	1338.31	B196		
Streptococcus pneumoniae	1313.13073	1313.13078	1313.13079	1313.13832	1313.13833	1313.13834
	1313.13835	1313.15185	1313.15186	1313.15187	170187.11	525381.4
	574093.3	R6				
Streptococcus pyogenes	1010840.4	1048264.3	1150773.3	1207470.4	1235829.3	1314.131
	1314.132	1314.134	1314.189	1314.191	1314.194	1314.195
	1314.199	1314.200	1314.201	1314.202	1314.212	1314.213
	1314.214	1314.239	1314.260	1314.261	1314.262	1314.263
	1314.267	1314.375	1314.470	1314.492	1314.496	1314.497
	1314.515	1314.534	1314.536	1437007.3	1437008.3	1440772.3
	160490.10	160491.19	193567.3	286636.3	471876.6	487215.4
Vibrio cholerae	1224154.5	1420885.6	666.1986	666.1992	666.3404	686.14
Vibrio parahaemolyticus	1338032.3	1338034.3	670.1110	670.1228	670.961	
Vibrio vulnificus	672.143	672.153	672.170	672.171		
Yersinia enterocolitica	1262462.5	1262464.3	1262466.3	1262467.6	630.129	630.16
	630.18	630.33	930944.6			
Yersinia pestis	214092.181	214092.21	360102.15	547048.4	632.127	632.129
	632.175					

TABLE A4.1: Coronaviruses to which the HUBDesign pipeline was applied

Virus	Genus	TaxID	Acc.No.
Feline infectious peritonitis virus	Alphacoronavirus	11135	NC_002306.3
Human coronavirus 229E	Alphacoronavirus	11137	NC_002645.1
Transmissible gastroenteritis virus	Alphacoronavirus	11149	NC_038861.1
Porcine epidemic diarrhea virus	Alphacoronavirus	28295	NC_003436.1
Human coronavirus NL63	Alphacoronavirus	277944	NC_005831.2
Bat coronavirus 1A	Alphacoronavirus	393767	NC_010437.1
Rhinolophus bat coronavirus HKU2	Alphacoronavirus	693998	NC_009988.1
Scotophilus bat coronavirus 512	Alphacoronavirus	693999	NC_009657.1
Miniopterus bat coronavirus HKU8	Alphacoronavirus	694001	NC_010438.1
Mink coronavirus strain WD1127	Alphacoronavirus	766791	NC_023760.1
Rousettus bat coronavirus HKU10	Alphacoronavirus	1241933	NC_018871.1
Ferret coronavirus	Alphacoronavirus	1264898	NC_030292.1
Bat coronavirus CDPHE15/USA/2006	Alphacoronavirus	1384461	NC_022103.1
BtMr-AlphaCoV/SAX2011	Alphacoronavirus	1503289	NC_028811.1
BtNv-AlphaCoV/SC2013	Alphacoronavirus	1503291	NC_028833.1
BtRf-AlphaCoV/HuB2013	Alphacoronavirus	1503292	NC_028814.1
BtRf-AlphaCoV/YN2012	Alphacoronavirus	1503293	NC_028824.1
Lucheng Rn rat coronavirus	Alphacoronavirus	1508224	NC_032730.1
Wencheng Sm shrew coronavirus	Alphacoronavirus	1508228	NC_035191.1
Camel alphacoronavirus	Alphacoronavirus	1699095	NC_028752.1
Swine enteric coronavirus	Alphacoronavirus	1766554	NC_028806.1
NL63-related bat coronavirus	Alphacoronavirus	1920748	NC_032107.1
Coronavirus AcCoV-JC34	Alphacoronavirus	1964806	NC_034972.1
Bovine coronavirus	Betacoronavirus	11128	NC_003045.1
Murine hepatitis virus	Betacoronavirus	11138	NC_001846.1
Murine hepatitis virus strain JHM	Betacoronavirus	11144	AC_000192.1
Human coronavirus OC43	Betacoronavirus	31631	NC_006213.1
Human coronavirus HKU1	Betacoronavirus	290028	NC_006577.2
Rat coronavirus Parker	Betacoronavirus	502102	NC_012936.1
Rousettus bat coronavirus HKU9	Betacoronavirus	694006	NC_009021.1
Tylonycteris bat coronavirus HKU4	Betacoronavirus	694007	NC_009019.1
Pipistrellus bat coronavirus HKU5	Betacoronavirus	694008	NC_009020.1
Severe acute respiratory syndrome-related coronavirus	Betacoronavirus	694009	NC_004718.3
Bat coronavirus BM48-31/BGR/2008	Betacoronavirus	864596	NC_014470.1
Rabbit coronavirus HKU14	Betacoronavirus	1160968	NC_017083.1
Betacoronavirus England 1	Betacoronavirus	1263720	NC_038294.1
Middle East respiratory syndrome-related coronavirus	Betacoronavirus	1335626	NC_019843.3
Betacoronavirus Erinaceus/VMC/DEU/2012	Betacoronavirus	1385427	NC_039207.1
Bat coronavirus	Betacoronavirus	1508220	NC_034440.1
Bat Hp-betacoronavirus/Zhejiang2013	Betacoronavirus	1541205	NC_025217.1
Betacoronavirus HKU24	Betacoronavirus	1590370	NC_026011.1
Rousettus bat coronavirus	Betacoronavirus	1892416	NC_030886.1
Severe acute respiratory syndrome coronavirus 2	Betacoronavirus	2697049	NC_045512.2
Infectious bronchitis virus	Gammacoronavirus	11120	NC_001451.1
Turkey coronavirus	Gammacoronavirus	11152	NC_010800.1
Beluga whale coronavirus SW1	Gammacoronavirus	694015	NC_010646.1
Bulbul coronavirus HKU11-934	Deltacoronavirus	572288	NC_011547.1
Munia coronavirus HKU13-3514	Deltacoronavirus	572289	NC_011550.1
Thrush coronavirus HKU12-600	Deltacoronavirus	572290	NC_011549.1
Common moorhen coronavirus HKU21	Deltacoronavirus	1159902	NC_016996.1
Magpie-robin coronavirus HKU18	Deltacoronavirus	1159903	NC_016993.1
Night heron coronavirus HKU19	Deltacoronavirus	1159904	NC_016994.1
Porcine coronavirus HKU15	Deltacoronavirus	1159905	NC_039208.1
Sparrow coronavirus HKU17	Deltacoronavirus	1159906	NC_016992.1
White-eye coronavirus HKU16	Deltacoronavirus	1159907	NC_016991.1
Wigeon coronavirus HKU20	Deltacoronavirus	1159908	NC_016995.1

TABLE A4.3: Primers used for Viral RNA Quantification

Virus	Target	Genome Position	Annealing Temp		Primer
SARS-CoV-2	E Gene	26,269-26,381	57°C	F	ACAGGTACGTTAATAGTTAATAGCGT
				R	ATATTGCAGCAGTACGCACACA
HCoV-NL63	RdRp Gene	13,547-13,743	60°C	F	CTTCTTCCCCAGCACTCGTT
				R	AGCATCACCATTCTGTGCGA

TABLE A4.4: Contaminant amplicon sequences filtered against (Primers in bold)

Amplicon	Sequence
HCoV-NL63	**CTTCTTCCCCAGCACTCGTT**GATCAACGCACTATTTGTTT TTCTGTTGCAGCATTGAGTACTGGTTTGACAAATCAAGT TGTTAAGCCAGGTCATTTTAATGAAGAGTTTTATAACTTT CTTCGTTTAAGAGGTTTCTTTGATGAAGGTTCTGAACTTA CATTAAAACATTTCTTCTT**CGCACAGAATGGTGATGCT**
SARS-CoV-2 N	**TTACAAACATTGGCCGCAA**ATTGCACAATTTGCCCCCAGC GCTTCAGCG**TTCTTCGGAATGTCGCGC**ATTGGCATGGAAG TCACACCTTCGGGAACG
SARS-CoV-2 RdRp	AGTGTGCTCAAGTATTGA**GTGAAATGGTCATGTGTGGCGG** TTCACTATATGTTAAACCAGGTGGAACCTCATCAGGAGAT GCCACAACTGCT**TATGCTAATAGTGTTTTTAACATTTG**TC AAGCTGT
SARS-CoV-2 E	CGGAAGAG**ACAGGTACGTTAATAGTTAATAGCGT**ACTTCT TTTTCTTGCTTTCGTGGTATTCTTGCTAGTTACACTAGCC ATCCTTACTGCGCTTCGATTG**TGTGCGTACTGCTGCAATA** **T**TGTTAACGT
SARS-CoV-2 ORF3a	GTGAAATCAAGGATGCTACTCCTTCAGAT**TTTGTTCGCGC** **TACTGCAAC**GATACCGATACAAGCCTCACTCCCTTT**CGGA** **TGGCTTATTGTTGGCG**TTGCACTTCTTGCTGTTTTTCAGA GCGCTTCCA

TABLE A4.5: Performance of SA_BOND on various input sets

Description	Genomes	Collapsed Input (Mbp)	Genomes Covered	Run Time (min)	Memory (GB)
Coronaviruses	56	0.8	96%	0.5	0.5
Respiratory Viruses	110	0.6	100%	1.5	0.5
Sepsis Bacteria	1926	108	100%	25	7.1
Gut Bacteria	1473	4168	89%	1165	272

TABLE A4.6: Probe counts by taxa in the Sepsis probe set

Genus	Species	Genus Level	Species Level	Total Probes
Achromobacter	xylosoxidans	283	0	283
Acinetobacter	baumannii	356	10	366
Acinetobacter	nosocomialis	356	378	734
Bacillus	anthracis	432	0	432
Bordetella	pertussis	291	0	291
Bordetella	bronchiseptica	291	254	545
Bordetella	hinzii	291	377	668
Bordetella	holmesii	291	420	711
Bordetella	parapertussis	291	454	745
Borrelia	burgdorferi	90	0	90
Burkholderia	dolosa	255	69	324
Burkholderia	cepacia	255	142	397
Burkholderia	pseudomalli	255	269	524
Burkholderia	stabilis	255	298	553
Burkholderia	multivorans	255	309	564
Burkholderia	cenocepacia	255	338	593
Burkholderia	ambifaria	255	354	609
Burkholderia	latens	255	373	628
Burkholderia	mallei	255	382	637
Burkholderia	thailandensis	255	397	652
Burkholderia	oklaholmensis	255	406	661
Burkholderia	vietnamiensis	255	419	674
Campylobacter	jejuni	149	0	149
Campylobacter	coli	149	24	173
Campylobacter	fetus	149	363	512
Chlamydia	trachomatis	414	155	569
Chlamydia	pneumoniae	414	396	810
Clostridioides	difficile	177	0	177
Clostridium	botulinum	0	12	12
Clostridium	perfringens	0	44	44
Clostridium	tetani	0	71	71
Enterobacter	aerogenes	338	338	676
Enterobacter	cloacae	338	383	721
Enterococcus	faecium	414	57	471
Enterococcus	faecalis	414	441	855
Escherichia	coli	167	0	167
Escherichia	albertii	167	245	412
Escherichia	fergusonii	167	435	602
Francisella	tularensis	155	0	155
Haemophilus	influenzae	383	19	402
Haemophilus	parainfluenzae	383	389	772
Helicobacter	pylori	389	0	389
Klebsiella	pneumoniae	394	84	478
Klebsiella	oxytoca	394	431	825
Legionella	pneumophila	449	0	449
Listeria	monocytogenes	408	0	408
Micrococcus	luteus	238	0	238
Moraxella	catarrhalis	439	0	439
Mycobacterium	tuberculosis	427	81	508
Mycobacterium	abscessus	427	336	763
Mycobacterium	bovis	427	358	785
Mycoplasma	pneumoniae	416	0	416
Neisseria	meningitidis	380	242	622
Neisseria	gonorrhoeae	380	277	657
Proteus	vulgaris	368	24	392
Proteus	mirabilis	368	42	410
Pseudomonas	aeruginosa	344	0	344
Rickettsia	prowazekii	53	0	53
Salmonella	enterica	377	0	377
Serratia	marcescens	362	0	362
Shigella	boydii	240	202	442
Shigella	sonnei	240	238	478
Shigella	flexneri	240	303	543
Shigella	dysenteriae	240	341	581
Staphylococcus	aureus	390	25	415
Staphylococcus	epidermidis	390	45	435
Staphylococcus	lugdunensis	390	134	524
Staphylococcus	saprophyticus	390	220	610
Staphylococcus	pseudointermedius	390	426	816
Stenotrophomonas	maltophilia	218	0	218
Streptococcus	intermedius	111	81	192
Streptococcus	anginosus	111	85	196
Streptococcus	agalactiae	111	176	287
Streptococcus	constellatus	111	197	308
Streptococcus	pneumoniae	111	367	478
Streptococcus	pyogenes	111	417	528
Vibrio	cholerae	432	338	770
Vibrio	parahaemolyticus	432	368	800
Vibrio	vulnificus	432	435	867
Yersinia	enterocolitica	413	361	774
Yersinia	pestis	413	463	876

TABLE A4.7: Regions of Viral Genomes with Apparent Bait Free Enrichment

Virus	Region		Number of Off Target Enriched Reads			
	Start	End	All	(+) Stranded	Intragenomic Bait	Extragenomic Bait
HCoV-NL63	11359	11445	6020	1602	7	1
	23051	23304	971	698	0	390
SARS-CoV-2	4953	5255	905	167	2	0
	5551	5910	538	318	5	0
	6686	7214	2415	1017	24	0
	7243	8356	7751	3847	31	0
	10641	11512	7410	5733	44	0
	13779	13891	145	115	1	2
	14707	14912	858	59	1	17
	15447	16107	2550	1336	8	609
	16158	17169	26677	1595	2	114
	19570	20200	1456	3593	13	41
	25730	26339	18539	10645	68	0
	27487	28060	14583	7590	124	0

TABLE A4.8: Logistic Regression of on-target coronavirus reads

Coefficient	Unit	Estimate	Std Error	95% CI	
Intercept	NA	-11.48	2.17×10^{-2}	-11.51	-11.44
Enriched	NA	4.58	6.59×10^{-3}	4.57	4.59
$\text{Mass}_{\text{SARS-CoV-2}}$	fg^{-1}	6.77×10^{-4}	2.57×10^{-6}	6.72×10^{-4}	6.81×10^{-4}
$\text{Mass}_{\text{HCoV-NL63}}$	fg^{-1}	2.62×10^{-4}	9.21×10^{-5}	2.60×10^{-4}	2.63×10^{-4}
Mass Interaction	fg^{-2}	3.61×10^{-5}	1.68×10^{-9}	3.57×10^{-5}	3.63×10^{-5}

TABLE A4.9: Linear Regression of log2 Fold Enrichment in Coronavirus genomic regions

Data	Coefficient	Unit	Estimate	Std Error	95% CI	
All Regions	Intercept	NA	15.1	0.445	14.4	15.8
	GC Devation	$\%^{-1}$	-8.85×10^{-2}	2.41×10^{-2}	-0.128×10^{-2}	-4.8×10^{-2}
	GC Devation + Baits	$\%^{-1}$	-1.02×10^{-3}	4.90×10^{-2}	-8.18×10^{-2}	7.98×10^{-2}
	Baseline	$\%^{-1}$	-0.190	1.62×10^{-2}	-0.217	-0.164
	Baseline + Baits	$\%^{-1}$	1.17×10^{-2}	1.57×10^{-2}	-1.42×10^{-2}	3.75×10^{-2}
	SARS-CoV-2 Load	$\log2(ng)^{-1}$	0.961	3.50×10^{-2}	0.903	1.02
	SARS-CoV-2 Load + Baits	$\log2(ng)^{-1}$	0.947	3.26×10^{-2}	0.893	1.00
	HCoV-NL63 Load	$\log2(ng)^{-1}$	-5.61×10^{-2}	6.61×10^{-2}	-0.165	5.29×10^{-2}
	HCoV-NL63 Load + Baits	$\log2(ng)^{-1}$	0.204	9.13×10^{-2}	0.107	0.300
	HCoV-NL63 Regions	NA	-10.78	0.828	-12.1	-9.42
	HCoV-NL63 Regions + Baits	NA	-8.02	0.772	-9.29	-6.75
Probe Regions	Intercept	NA	14.8	0.530	13.9	15.7
	Bait Density	$bait\cdot kb^{-1}$	1.30×10^{-2}	6.18×10^{-3}	2.74×10^{-3}	2.32×10^{-2}
	GC Deviation	$\%^{-1}$	-2.68×10^{-2}	4.81×10^{-2}	-0.106	5.29×10^{-2}
	Bait Divergence	$\%^{-1}$	-2.99×10^{-2}	2.46×10^{-2}	-7.05×10^{-2}	1.08×10^{-2}
	Baseline	$\%^{-1}$	-2.91×10^{-3}	1.53×10^{-2}	-2.82×10^{-2}	2.23×10^{-2}
	SARS-CoV-2 Viral Load	$\log2(ng)^{-1}$	0.916	3.10×10^{-2}	0.865	0.968
	HCoV-NL63 Viral Load	$\log2(ng)^{-1}$	0.204	4.70×10^{-2}	0.126	0.281
	HCoV-NL63 Regions	NA	-7.53	0.634	-6.48	-6.61

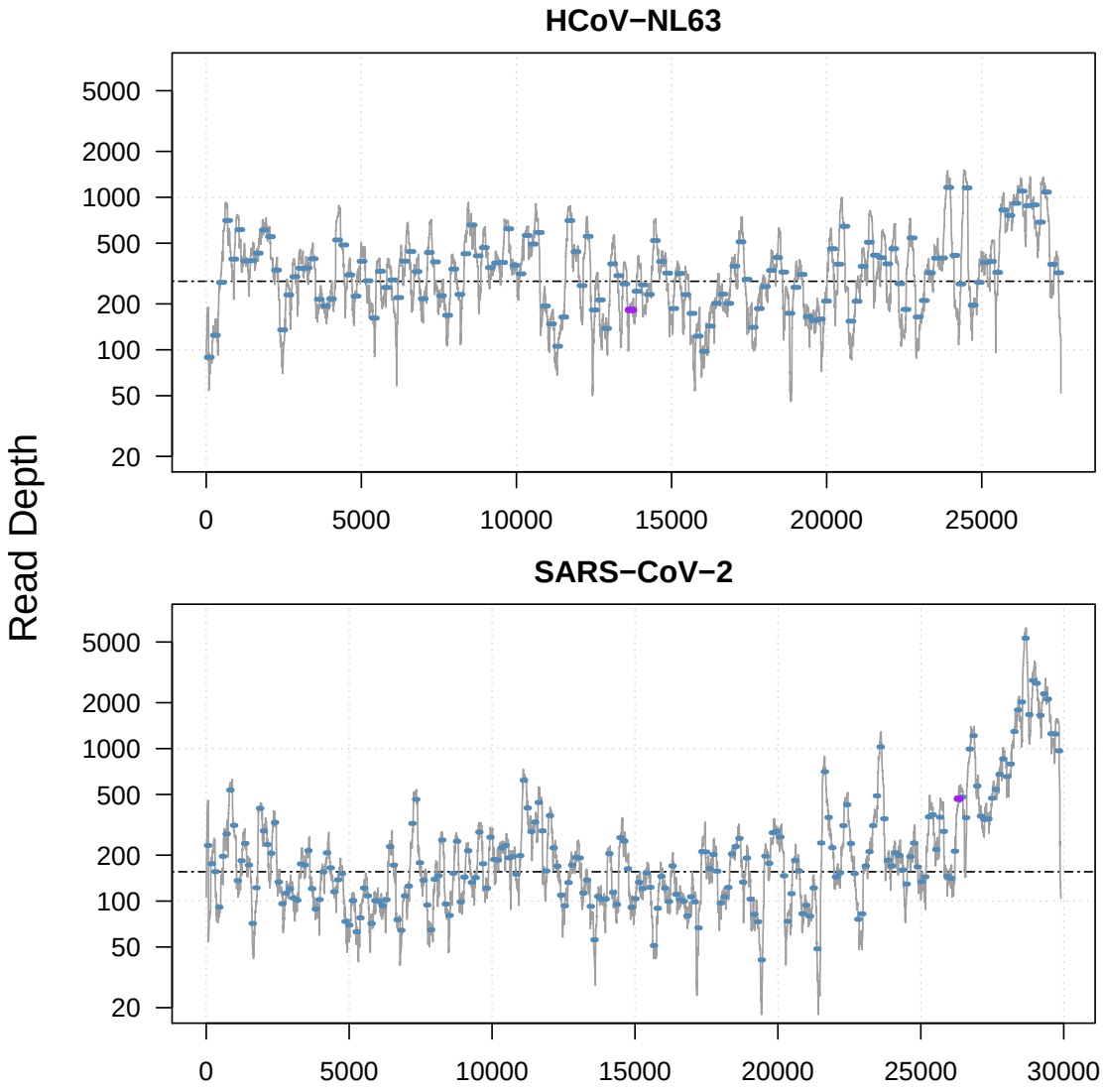

FIGURE A4.1: The grey line is the observed read depth at every position. The blue line segments are the mean depth across a region the same length as the amplicon generated during qPCR. The purple line segment is the position within the genome targeted by primers used to quantify the copy number of the viral samples. The dotted line is the average depth across all primer-length regions in the genome. In HCoV-NL63 the dotted line is higher than the purple reference region indicating that one copy quantified during qPCR results in a mass of viral RNA greater than one genome's worth. The opposite is true for SARS-CoV-2. The mass of viral RNA present when a single copy is quantified is the sum of the blue line segments relative to the reference region.

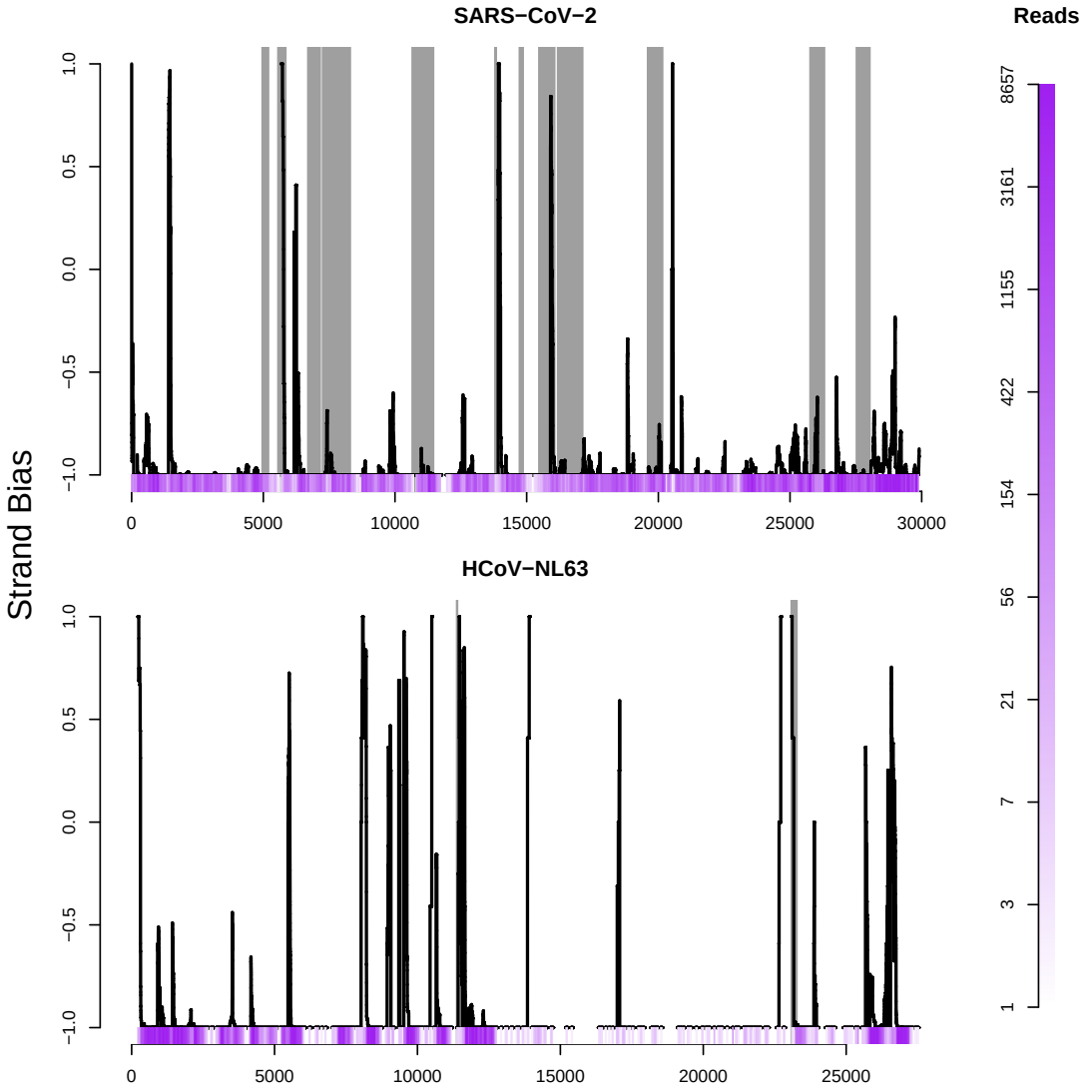

FIGURE A4.2: SARS-CoV-2 data from HiLo samples, and HCoV-NL63 data from LoHi Samples. The colour intensity indicates the number of reads mapping to a position. Regions of the genome which appear to be enriched without having nearby baits are indicated by grey shading. Most reads, based on cDNA, are negative stranded which indicates most RNA in the sample is positive stranded. Strand bias flips on the flanks of enriched regions indicating indirect enrichment.

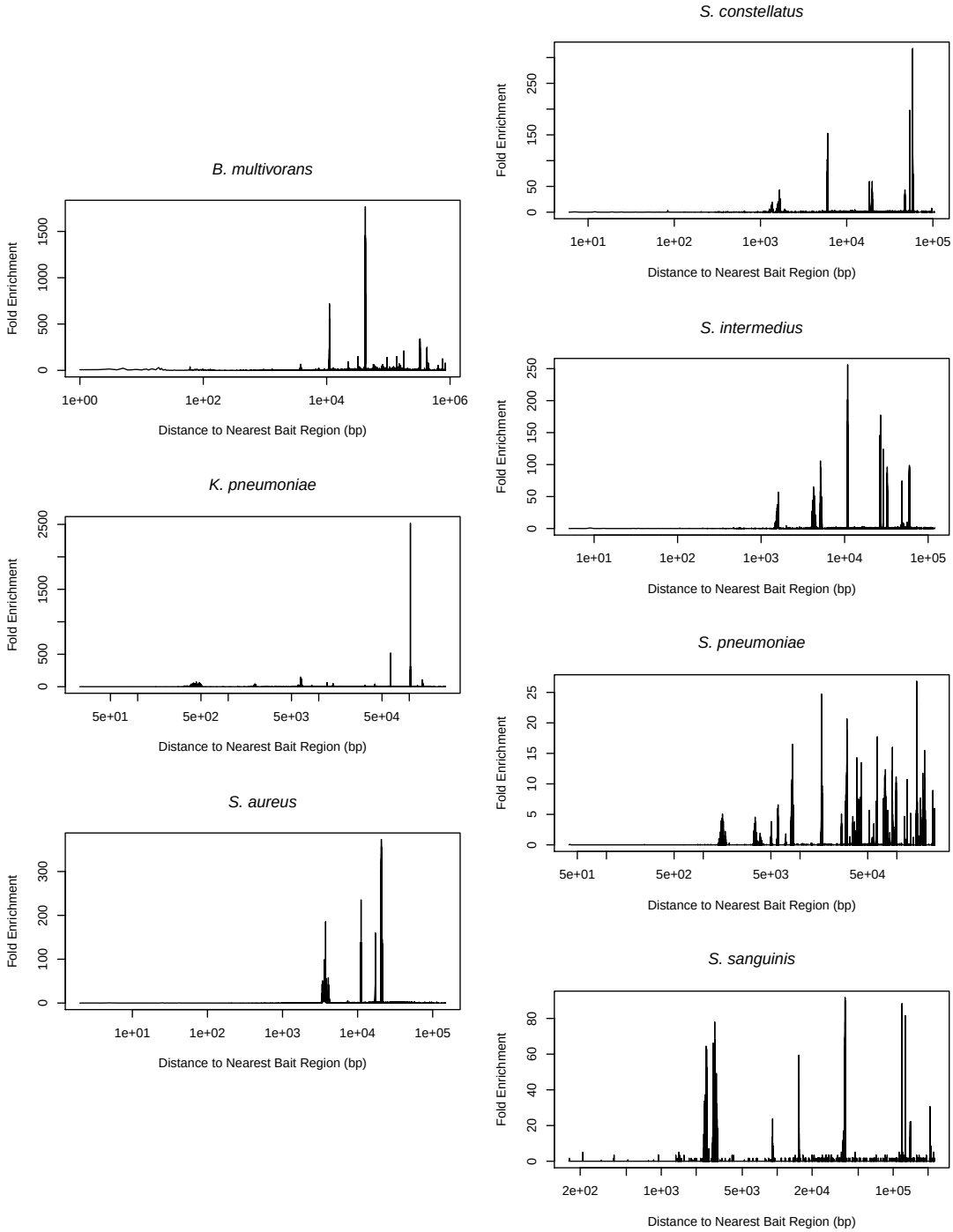

FIGURE A4.3: . There is very little enrichment at distances near to, but outside of bait regions, while there is clear enrichment at further positions.

TABLE A4.10: Estimated change in log fold enrichment from species level to genus level probes corrected for GC content, divergence, probe density, and baseline read counts

Strain	Estimate	StdErr	95% CI	
Burkholderia multivorans	0.355	1.02	-1.32	2.03
Klebsiella pneumoniae	0.951	0.890	-0.514	2.41
Staphylococcus aureus	5.10	1.14	3.23	6.97
Streptococcus constellatus	13.1	11.1	-5.12	31.2
Streptococcus intermedius	2.43	2.46	-1.63	6.48
Streptococcus pneumoniae	NA	NA	NA	NA
Streptococcus sanguinis	NA	NA	NA	NA

Bibliography

Adamowicz S. 2015. International Barcode of Life: Evolution of a global research community. Genome 58:151–162.

Akaike H. 1998. Selected Papers of Hirotugu Akaike. chapter Information Theory and an Extension of the Maximum Likelihood Principle. New York, NY: Springer New York. p. 199–213. doi:10.1007/978-1-4612-1694-0_15.

Albà M, Guigó R. 2004. Comparative analysis of amino acid repeats in rodents and humans. Genome Res 14:549–554.

Albà M, Santibáñez-Koref M, Hancock J. 1999. Conservation of polyglutamine tract size between mice and humans depends on codon interruption. Mol Biol Evol 16:1641–1644.

Altschul S, Gish W, Miller W, Myers E, Lipman D. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Andrews S. 2015. Fastqc. https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Ann Arbor Biosciences. 2018. myBaits - Hybridization Capture for targeted NGS, v4.01.

Banerjee A, Nasir J, Budylowski P, Yip L, Aftanas P, Christie N, Ghalami A, Baid K, Raphenya A, Hirota J, et al. 2020. Isolation, Sequence, Infectivity, and Replication Kinetics of Severe Acute Respiratory Syndrome Coronavirus 2. Emerg Infect Dis 26:2054–2063.

Barik S. 2017. Amino acid repeats avert mRNA folding through conservative substitutions and synonymous codons, regardless of codon bias. Heliyon 3:e00492.

Basavappa R, Sigler P. 1991. The 3 A crystal structure of yeast initiator tRNA: functional implications in initiator/elongator discrimination. EMBO J 10:3105–3111.

Battistuzzi F, Schneider K, Spencer M, Fisher D, Chaudhry S, Escalante A. 2016. Profiles of low complexity regions in *Apicomplexa*. BMC Evol Biol 16:47.

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 U S A 106:1133–1138.

Behura S, Severson D. 2012. Genome-wide comparative analysis of simple sequence coding repeats among 25 insect species. Gene 504:226–232.

Benirschke R, McElvania E, ThomsonRB J, Kaul K, Das S. 2019. Clinical Impact of Rapid Point-of-Care PCR Influenza Testing in an Urgent Care Setting: a Single-Center Study. J Clin Microbiol 57:01281–18.

Bertone P, Trifonov V, Rozowsky J, Schubert F, Emanuelsson O, Karro J, Kao M, Snyder M, Gerstein M. 2006. Design optimization methods for genomic DNA tiling arrays. Genome Res 16:271–281.

Bihorel S, Baudin M. 2018. neldermead: R Port of the 'Scilab' Neldermead Module. R package version 1.0-11.

Boni M, Lemey P, Jiang X, Lam T, Perry B, Castoe T, Rambaut A, Robertson D. 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol 5:1408–1417.

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.

Brandström M, Ellegren H. 2008. Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. Genome Res 18:881–887.

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343–348.

Brown L, Brown S. 2004. Alanine tracts: the expanding story of human illness and trinucleotide repeats. Trends Genet 20:51–58.

Brown P, Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. Nat Genet 21:33–37.

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.

Cambridge SB, Gnad F, Nguyen C, Bermejo JL, Krüger M, Mann M. 2011. Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. J Proteome Res 10:5275–84.

Carelli F, Liechti A, Halbert J, Warnefors M, Kaessmann H. 2018. Repurposing of promoters and enhancers during mammalian evolution. Nat Commun 9:4066.

Cascarina SM, Ross ED. 2018. Proteome-scale relationships between local amino acid composition and protein fates and functions. PLOS Computational Biology 14:1–33.

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 J, Swofford R, Johnson J, Cummings B, Rogel N, Lindblad-Toh K, Haerty W, Palma F, Regev A. 2019. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. Genome Res 29:53–63.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890.

Choi H, Hwang C, Song K, Law P, Wei L, Loh H. 2009. Poly(C)-binding proteins as transcriptional regulators of gene expression. Biochem Biophys Res Commun 380:431–436.

Coletta A, Pinney J, Solís D, Marsh J, Pettifer S, Attwood T. 2010. Low-complexity regions within protein sequences have position-dependent roles. BMC Syst Biol 4:43.

Cook D, Andersen E. 2017. VCF-kit: assorted utilities for the variant call format. Bioinformatics 33:1581–1582.

Cornman RS, Willis JH. 2009. Annotation and analysis of low-complexity protein families of *Anopheles gambiae* that are associated with cuticle. Insect Mol Biol 18:1365–2583.

Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 5:536–544.

Cummings CJ, Zoghbi HY. 2000. Fourteen and counting: unraveling trinucleotide repeat diseases. Hum Mol Genet 9:909–16.

Dalphin M, Brown C, Stockwell P, Tate W. 1996. TransTerm: a database of translational signals. Nucleic Acids Res 24:216–218.

David Shen ZL. 2006. Computation of Correlation Coefficient and Its Confidence Interval in SAS. In: R Mitchell, editor. SAS Users Group International Proceedings. volume 170 of *SUGI 31*. p. 170.

Davis J, Wattam A, Aziz R, Brettin T, Butler R, Butler R, Chlenski P, Conrad N, Dickerman A, Dietrich E, et al. 2020. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic Acids Res 48:D606–D612.

Day J, Ranum L. 2005. RNA pathogenesis of the myotonic dystrophies. Neuromuscul Disord 15:5–16.

DelCampo C, Bartholomaus A, Fedyunin I, Ignatova Z. 2015. Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. PLoS Genet 11:e1005613.

Delsuc F, Gibb G, Kuch M, Billet G, Hautier L, Southon J, Rouillard J, Fernicola J, Vizcaino S, MacPhee R, et al. 2016. The phylogenetic affinities of the extinct glyptodonts. Curr Biol 26:R155–6.

DePristo M, Zilversmit M, Hartl D. 2006. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. Gene 378:19–30.

Dere R, Napierala M, Ranum L, Wells R. 2004. Hairpin structure-forming propensity of the (CCTG.CAGG) tetranucleotide repeats contributes to the genetic instability associated with myotonic dystrophy type 2. J Biol Chem 279:41715–41726.

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:msac087.

Dickson Z, Hackenberger D, Kuch M, Marzok A, Banerjee A, Rossi L, Klowak J, Fox-Robichaud A, Mossmann K, Miller M, et al. 2021. Probe design for simultaneous, targeted capture of diverse metagenomic targets. Cell Rep Methods 1:100069.

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.

Dignon GL, Zheng W, Kim YC, Best RB, Mittal J. 2018. Sequence determinants of protein phase behavior from a coarse-grained model. PLOS Computational Biology 14:1–23.

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.

Doherty MK, Hammond DE, Clague MJ, Gaskell SJ, Beynon RJ. 2009. Turnover of the human proteome: Determination of protein intracellular stability by dynamic silac. Journal of Proteome Research 8:104–112. PMID: 18954100.

dosReis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32:5036–5044.

Dosztányi Z, Chen J, Dunker A, Simon I, Tompa P. 2006. Disorder and sequence repeats in hub proteins and their implications for network evolution. J Proteome Res 5:2985–2995.

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. Journal of molecular biology 347:827–839.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences 102:14338–14343.

Dyson H, Wright P. 2005. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6:197–208.

Ebert P, Audano P, Zhu Q, Rodriguez-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.

Ekman D, Light S, Björklund A, Elofsson A. 2006. What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae?* Genome Biol 7:R45.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445.

Elnifro E, Ashshi A, Cooper R, Klapper P. 2000. Multiplex PCR: optimization and application in diagnostic virology. Clin Microbiol Rev 13:559–570.

Enk J, Devault A, Widga C, Saunders J, Szpak P, Southon J, Rouillard JM, Shapiro B, Golding GB, Zazula G, et al. 2016. Mammuthus population dynamics in late pleistocene north america: Divergence, phylogeography, and introgression. Frontiers in Ecology and Evolution 4:42.

Enright J, Dickson Z, Golding G. 2023. Low Complexity Regions in Proteins and DNA are Poorly Correlated. Mol Biol Evol 40:msad084.

Everett C, Wood N. 2004. Trinucleotide repeats and neurodegenerative disease. Brain 127:2385–2405.

Fan H, Chu J. 2007. A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics 5:7–14.

Fassler J, Cooper P. 2011. BLAST® Help [Internet]. chapter BLAST Glossary. Bethesda, MD, US: National Center for Biotechnology Information. p. 53–60.

Faux N, Bottomley S, Lesk A, Irving J, Morrison J, delaBanda M, Whisstock J. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res 15:537–551.

Fehr A, Perlman S. 2015. Coronaviruses: an overview of their replication and pathogenesis. Methods Mol Biol 1282:1–23.

Felsenstein J. 1989. Phylip - phylogeny inference package (version 3.2). Cladistics 5:164–166.

Fitch WM. 1971. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Systematic Biology 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.

Fondon JW, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A 101:18058–18063.

Fushan A, Turanov A, Lee S, Kim E, Lobanov A, Yim S, Buffenstein R, Lee S, Chang K, Rhee H, et al. 2015. Gene expression defines natural changes in mammalian lifespan. Aging Cell 14:352–365.

Gardner S, Jaing C, McLoughlin K, Slezak T. 2010. A microbial detection array (MDA) for viral and bacterial detection. BMC Genomics 11:668.

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-1 beta-lactamase. Journal of Molecular Biology 431:2320–2330.

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Systematic Zoology 28:132–163.

Goolsby E. 2017. Rapid maximum likelihood ancestral state reconstruction of continuous characters: A rerooting-free algorithm. Ecol Evol 7:2791–2797.

Gragg H, Harfe B, Jinks-Robertson S. 2002. Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by mismatch repair in *Saccharomyces cerevisiae*. Mol Cell Biol 22:8756–8762.

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.

GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 45:580–585.

Guan Q, Sadykov M, Mfarrej S, Hala S, Naeem R, Nugmanova R, Al-Omari A, Salih S, AlMutair A, Carr M, et al. 2020. A genetic barcode of SARS-CoV-2 for monitoring global distribution of different clades during the COVID-19 pandemic. Int J Infect Dis 100:216–223.

Guitor A, Raphenya A, Klunk J, Kuch M, Alcock B, Surette M, McArthur A, Poinar H, Wright G. 2019. Capturing the Resistome: a Targeted Capture Method To Reveal Antibiotic Resistance Determinants in Metagenomes. Antimicrob Agents Chemother 64:01324–19.

Haba Y, Kutsukake N. 2019. A multivariate phylogenetic comparative method incorporating a flexible function between discrete and continuous traits. Evolutionary Ecology 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.

Hannan A. 2018. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet 19:286–298.

Harrison PM. 2017. fLPS: Fast discovery of compositional biases for the protein universe. BMC Bioinformatics 18:476.

Hastings WK. 1970. Monte carlo sampling methods using markov chains and their applications. Biometrika 57:97–109.

Hayden M, Nguyen T, Waterman A, Chalmers K. 2008. Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping. BMC Genomics 9:80.

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. Journal of Applied Probability 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.

Howe K, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode M, Armean I, Azov A, Bennett R, Bhai J, et al. 2021. Ensembl 2021. Nucleic Acids Res 49:D884–D891.

Huntley M, Clark A. 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. Mol Biol Evol 24:2598–2609.

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.

Ilie L, Mohamadi H, Golding G, Smyth W. 2013. BOND: Basic OligoNucleotide Design. BMC Bioinformatics 14:69.

Jacobs G, Rackham O, Stockwell P, Tate W, Brown C. 2002. Transterm: a database of mRNAs and translational control elements. Nucleic Acids Res 30:310–311.

Jacobs G, Stockwell P, Schrieber M, Tate W, Brown C. 2000. Transterm: a database of messenger RNA components and signals. Nucleic Acids Res 28:293–295.

Jeronimo C, Collin P, Robert F. 2016. The RNA Polymerase II CTD: The Increasing Complexity of a Low-Complexity Protein Domain. J Mol Biol 428:2607–2622.

Joy J, Liang R, McCloskey R, Nguyen T, Poon A. 2016. Ancestral Reconstruction. PLoS Comput Biol 12:e1004763.

Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles A. 2002. Amino acid runs in eukaryotic proteomes and disease associations. Proc Natl Acad Sci U S A 99:333–338.

Kato M, McKnight SL. 2017. Cross-beta polymerization and hydogel formation by low-complexity sequence proteins. Methods 126:3–11.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–80.

Kebede A, Tadesse F, Feleke A, Golassa L, Gadisa E. 2019. Effect of low complexity regions within the PvMSP3alpha block II on the tertiary structure of the protein and implications to immune escape mechanisms. BMC Struct Biol 19:6.

Kedersha N, Anderson P. 2002. Stress granules: sites of mrna triage that regulate mrna stability and translatability. Biochemical Society Transactions 30:963–969.

Kiefer J. 1953. Sequential minimax search for a maximum. Proceedings of the American Mathematical Society 4:502–506.

Kim D, Lee J, Yang J, Kim J, Kim V, Chang H. 2020. The Architecture of SARS-CoV-2 Transcriptome. Cell 181:914–921.e10.

Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 40:e3.

Knight R, Freeland S, Landweber L. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol 2:RESEARCH0010.

Kobe B, Kajava A. 2001. The leucine-rich repeat as a protein recognition motif. Curr Opin Struct Biol 11:725–732.

Kong J, Liebhaber S. 2007. A cell type-restricted mRNA surveillance pathway triggered by ribosome extension into the 3' untranslated region. Nat Struct Mol Biol 14:670–676.

Koonin E, Novozhilov A. 2009. Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61:99–111.

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 U S A 95:10774–10778.

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

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Leclercq S, Rivals E, Jarne P. 2010. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. Genome Biol Evol 2:325–335.

Ledger M, Grimshaw E, Fairey M, Whelton H, Bull I, Ballantyne R, Knight M, Mitchell P. 2019. Intestinal parasites at the Late Bronze Age settlement of Must Farm, in the fens of East Anglia, UK (9th century B.C.E.). Parasitology 146:1583–1594.

Leinonen R, Sugawara H, Shumway M. 2011. The sequence read archive. Nucleic Acids Res 39:D19–21.

Lenz C, Haerty W, Golding G. 2014. Increased substitution rates surrounding low-complexity regions within primate proteins. Genome Biol Evol 6:655–665.

Levinson G, Gutman G. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221.

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.

Li J, Zhou J, Wu Y, Yang S, Tian D. 2015. GC-Content of Synonymous Codons Profoundly Influences Amino Acid Usage. G3 (Bethesda) 5:2027–2036.

Li X, Kahveci T. 2006. A Novel algorithm for identifying low-complexity regions in a protein sequence. Bioinformatics 22:2980–2987.

Li X, Xiong X, Zhang M, Wang K, Chen Y, Zhou J, Mao Y, Lv J, Yi D, Chen X, et al. 2017. Base-Resolution Mapping Reveals Distinct m(1)A Methylome in Nuclear- and Mitochondrial-Encoded Transcripts. Mol Cell 68:993–1005.e9.

Lin M, Whitmire S, Chen J, Farrel A, Shi X, Guo Jt. 2017. Effects of short indels on protein structure and function in human genomes. Scientific Reports 7:9313.

Liu F, Clark W, Luo G, Wang X, Fu Y, Wei J, Wang X, Hao Z, Dai Q, Zheng G, et al. 2016. ALKBH1-Mediated tRNA Demethylation Regulates Translation. Cell 167:816–828.e16.

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.

Loya TJ, O'Rourke TW, Reines D. 2012. A genetic screen for terminator function in yeast identifies a role for a new functional domain in termination factor Nab3. Nucleic Acids Res 40:7476–7491.

Madsen C, Ghivizzani S, Hauswirth W. 1993. In vivo and in vitro evidence for slipped mispairing in mammalian mitochondria. Proc Natl Acad Sci U S A 90:7671–7675.

Martin E, Mittag T. 2018. Relationship of Sequence and Phase Separation in Protein Low-Complexity Regions. Biochemistry 57:2478–2487.

Martin W. 1999. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. Bioessays 21:99–104.

Mason V, Li G, Helgen K, Murphy W. 2011. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. Genome Res 21:1695–1704.

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.

Merkin J, Russell C, Chen P, Burge C. 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science 338:1593–1599.

Mertes F, Elsharawy A, Sauer S, vanHelvoort J, vanderZaag P, Franke A, Nilsson M, Lehrach H, Brookes A. 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics 10:374–386.

Metsky H, Siddle K, Gladden-Young A, Qu J, Yang D, Brehio P, Goldfarb A, Piantadosi A, Wohl S, Carter A, et al. 2019. Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. Nat Biotechnol 37:160–168.

Metzgar D, Bytof J, Wills C. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 10:72–80.

Mier P, Alanis-Lobato G, Andrade-Navarro MA. 2017. Context characterization of amino acid homorepeats using evolution, position, and order. Proteins 85:709–719.

Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltész B, Urbanek A, Gruca A, Plewczynski D, Grynberg M, Bernadó P, et al. 2020. Disentangling the complexity of low complexity proteins. Breifings in Bioinformatics 21:458–472.

Millard P, Bugge K, Marabini R, Boomsma W, Burow M, Kragelund B. 2020. IDDomainSpotter: Compositional bias reveals domains in long disordered protein regions-Insights from transcription factors. Protein Sci 29:169–183.

Minh B, Schmidt H, Chernomor O, Schrempf D, Woodhams M, vonHaeseler 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.

Monahan Z, Ryan V, Janke A, Burke K, Rhoads S, Zerze G, O'Meally R, Dignon G, Conicella A, Zheng W, et al. 2017. Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. EMBO J 36:2951–2967.

Moore H, Greenwell P, Liu C, Arnheim N, Petes T. 1999. Triplet repeats form secondary structures that escape DNA repair in yeast. Proc Natl Acad Sci U S A 96:1504–1509.

Morgulis A, Gertz E, Schaffer A, Agarwala R. 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol 13:1028–1040.

Mularoni L, Veitia R, Albà M. 2007. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. Genomics 89:316–325.

Murat P, Guilbaud G, Sale J. 2020. DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. Genome Biol 21:209.

Musova Z, Mazanec R, Krepelova A, Ehler E, Vales J, Jaklova R, Prochazka T, Koukal P, Marikova T, Kraus J, et al. 2009. Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene. Am J Med Genet A 149A:1365–1374.

Necci M, Piovesan D, Dosztanyi Z, Tosatto S. 2017. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinformatics 33:1402–1404.

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.

Nie L, Wu G, Culley D, Scholten J, Zhang W. 2007. Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. Crit Rev Biotechnol 27:63–75.

Nie L, Wu G, Zhang W. 2006. Correlation of mRNA Expression and Protein Abundance Affected by Multiple Sequence Features Related to Translational Efficiency in Desulfovibrio vulgaris: A Quantitative Analysis. Genetics 174:2229–2243.

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.

O'Leary N, Wright M, Brister J, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–45.

Opota O, Jaton K, Greub G. 2015. Microbial diagnosis of bloodstream infection: towards molecular diagnosis directly from blood. Clin Microbiol Infect 21:323–331.

Overbeek R, Joerg D, Price M. 2010. Patscan. https://blog.theseed.org/servers/2010/07/scan-for-matches.html.

Pál C, Papp B, Hurst LD. 2001. Highly Expressed Genes in Yeast Evolve Slowly. Genetics 158:927–931.

Papafragkou E, Hewitt J, Park G, Greening G, Vinje J. 2014. Challenges of culturing human norovirus in three-dimensional organoid intestinal cell culture models. PLoS One 8:e63485.

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.

Patro R, Duggal G, Love M, Irizarry R, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14:417–419.

Pechmann S, Frydman J. 2012. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nature Structural and Molecular Biology 20:237.

Pedulla M, Ford M, Houtz J, Karthikeyan T, Wadsworth C, Lewis J, Jacobs-Sera D, Falbo J, Gross J, Pannunzio N, et al. 2003. Origins of highly mosaic mycobacteriophage genomes. Cell 113:171–182.

Peng X, Thierry-Mieg J, Thierry-Mieg D, Nishida A, Pipes L, Bozinoski M, Thomas M, Kelly S, Weiss J, Raveendran M, et al. 2015. Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTR). Nucleic Acids Res 43:D737–42.

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 U S A 120:e2300154120.

Persikov A, Ramshaw J, Kirkpatrick A, Brodsky B. 2000. Amino acid propensities for the collagen triple-helix. Biochemistry 39:14960–14967.

Pertea M, Pertea 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.

Price C. 1992. Centromeres and telomeres. Curr Opin Cell Biol 4:379–384.

R Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

R Core Team. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Radó-Trilla N, Albà M. 2012. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evol Biol 12:155.

Radó-Trilla N, Arató K, Pegueroles C, Raya A, delaLuna S, Albà M. 2015. Key Role of Amino Acid Repeat Expansions in the Functional Diversification of Duplicated Transcription Factors. Mol Biol Evol 32:2263–2272.

Rahman MS, Islam MR, Ul Alam ASMR, Islam I, Hoque MN, Akter S, Rahaman MM, Sultana M, Hossain MA. 2020. Evolutionary dynamics of sars-cov-2 nucleocapsid protein (n protein) and its consequences. bioRxiv .

Ranum L, Day J. 2002. Dominantly inherited, non-coding microsatellite expansion disorders. Curr Opin Genet Dev 12:266–271.

Reed L, Muench H. 1938. A Simple Method of Estimating Fifty Per Cent Endpoints. American Journal of Epidemiology 27:493–497.

Reiter F, deAlmeida B, Stark A. 2023. Enhancers display constrained sequence flexibility and context-specific modulation of motif function. Genome Res 33:346–358.

Revell LJ. 2012. phytools: an r package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3:217–223.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277.

Richard G, Paques F. 2000. Mini- and microsatellite expansions: the recombination connection. EMBO Rep 1:122–126.

Rodriguez-Morales A, Bonilla-Aldana D, Balbin-Ramon G, Rabaan A, Sah R, Paniz-Mondolfi A, Pagliano P, Esposito S. 2020. History is repeating itself: Probable zoonotic spillover as the cause of the 2019 novel Coronavirus Epidemic. Infez Med 28:3–5.

Rognes T, Flouri T, Nichols B, Quince C, Mahe F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584.

Rohlfs 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.

Rys P, Persing D. 1993. Preventing false positives: quantitative evaluation of three protocols for inactivation of polymerase chain reaction amplification products. J Clin Microbiol 31:2356–2360.

Sainudiin R, Durrett R, Aquadro C, Nielsen R. 2004. Microsatellite mutation models: insights from a comparison of humans and chimpanzees. Genetics 168:383–395.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425.

Salter S, Cox M, Turek E, Calus S, Cookson W, Moffatt M, Turner P, Parkhill J, Loman N, Walker A. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12:87.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864.

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 .

Schoch C, Ciufo S, Domrachev M, Hotton C, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, et al. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020:baaa062.

Schug M, Hutter C, Wetterstrand K, Gaudette M, Mackay T, Aquadro C. 1998. The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. Mol Biol Evol 15:1751–1760.

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. Nature 473:337.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

Sequencing C, Consortium A. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.

Shannon C. 1948. A Mathematical Theory of Communication. The Bell System Technical Journal 27:379–423, 623–656.

Shen W, Ren H. 2021. Taxonkit: A practical and efficient ncbi taxonomy toolkit. Journal of Genetics and Genomics 48:844–850. Special issue on Microbiome.

Shi J, Rabosky D. 2015. Speciation dynamics during the global radiation of extant bats. Evolution 69:1528–1545.

Shima H, Igarashi K. 2020. N 1-methyladenosine (m1A) RNA modification: the key to ribosome control. J Biochem 167:535–539.

Shin J, Salameh JS, Richter JD. 2016. Impaired neurodevelopment by the low complexity domain of CPEB4 reveals a convergent pathway with neurodegeneration. Scientific Reports .

Shumate A, Salzberg S. 2021. Liftoff: accurate mapping of gene annotations. Bioinformatics 37:1639–1643.

Stahlberg A, Krzyzanowski P, Egyud M, Filges S, Stein L, Godfrey T. 2017. Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. Nat Protoc 12:664–682.

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.

Tang Q, Gu Y, Zhou X, Jin L, Guan J, Liu R, Li J, Long K, Tian S, Che T, et al. 2017. Comparative transcriptomics of 5 high-altitude vertebrates and their low-altitude relatives. Gigascience 6:1–9.

Tatti K, Sparks K, Boney K, Tondella M. 2011. Novel multitarget real-time PCR assay for rapid detection of Bordetella species in clinical specimens. J Clin Microbiol 49:4059–4066.

Tautz D, Trick M, Dover G. 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656.

The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Research 45:D158–D169.

Toll-Riera M, Radó-Trilla N, Martys F, Albà M. 2012. Role of low-complexity sequences in the formation of novel protein coding sequences. Mol Biol Evol 29:883–886.

Tompa P. 2003. Intrinsically unstructured proteins evolve by repeat expansion. Bioessays 25:847–855.

Valberg S, Perumbakkam S, McKenzie E, Finno C. 2018. Proteome and transcriptome profiling of equine myofibrillar myopathy identifies diminished peroxiredoxin 6 and altered cysteine metabolic pathways. Physiol Genomics 50:1036–1050.

Van Rossum G, Drake FL. 2009. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Vats D, Flegal JM, Jones GL. 2017. Multivariate output analysis for markov chain monte carlo. 1512.07713.

Velasco A, Becerra A, Hernández-Morales R, Delaye L, Jiménez-Corona M, Ponce-de Leon S, Lazcano A. 2013. Low complexity regions (LCRs) contribute to the hypervariability of the HIV-1 gp120 protein. J Theor Biol 338:80–86.

Verstrepen K, Jansen A, Lewitter F, Fink G. 2005. Intragenic tandem repeats generate functional variability. Nat Genet 37:986–990.

Viguera E, Canceill D, Ehrlich S. 2001. Replication slippage involves DNA polymerase pausing and dissociation. EMBO J 20:2587–2595.

Villar D, Flicek P, Odom D. 2014. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat Rev Genet 15:221–233.

Wade W. 2002. Unculturable bacteria–the uncharacterized organisms that cause oral infections. J R Soc Med 95:81–83.

Wagner D, Klunk J, Harbeck M, Devault A, Waglechner N, Sahl J, Enk J, Birdsell D, Kuch M, Lumibao C, et al. 2014. Yersinia pestis and the plague of Justinian 541-543 AD: a genomic analysis. Lancet Infect Dis 14:319–326.

Wall L, Christiansen T, Orwant J. 2000. Programming perl. " O'Reilly Media, Inc.".

Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf S, Hengartner M, von Mering C. 2012. PaxDb, a database of protein abundance averages across all three domains of life. Mol Cell Proteomics 11:492–500.

Wang P, Yan Z, Yang S, Wang S, Zheng X, Fan J, Zhang T. 2019. Environmental DNA: An Emerging Tool in Ecological Assessment. Bull Environ Contam Toxicol 103:651–656.

Wang X, Li P, Gutenkunst RN. 2017. Systematic effects of mrna secondary structure on gene expression and molecular function in budding yeast. bioRxiv .

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.

Wierdl M, Dominska M, Petes T. 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics 146:769–779.

Wootton J. 1994a. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 18:269–285.

Wootton J. 1994b. Sequences with 'unusual' amino acid compositions. Current Opinion in Structural Biology 4:413–421.

Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. Computers Chem 17:149–163.

Xie N, Wang M, Song P, Mao S, Wang Y, Yang Y, Luo J, Ren S, Zhang D. 2022. Designing highly multiplex PCR primer sets with Simulated Annealing Design using Dimer Likelihood Estimation (SADDLE). Nat Commun 13:1881.

Xue H, Forsdyke D. 2003. Low-complexity segments in *Plasmodium falciparum* proteins are primarily nucleic acid level adaptations. Mol Biochem Parasitol 128:21–32.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659.

Zecha J, Meng C, Zolg DP, Samaras P, Wilhelm M, Kuster B. 2018. Peptide level turnover measurements enable the study of proteoform dynamics. Molecular & Cellular Proteomics 17:974–992.

Zhang L, Liu C, Jiang Q, Yin Y. 2021. Butyrate in Energy Metabolism: There Is Still More to Learn. Trends Endocrinol Metab 32:159–169.

Zhang T, Faraggi E, Xue B, Dunker AK, Uversky VN, Zhou Y. 2012. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J Biomol Struct Dyn. 29:799–813.

Zhang T, Shen S, Qu J, Ghaemmaghami S. 2016. Global analysis of cellular protein flux quantifies the selectivity of basal autophagy. Cell reports 14:2426–39.

Zhao H, Wei Z, Shen G, Chen Y, Hao X, Li S, Wang R. 2022. Poly(rC)-binding proteins as pleiotropic regulators in hematopoiesis and hematological malignancy. Front Oncol 12:1045797.

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.

Zilversmit M, Volkman S, DePristo M, Wirth D, Awadalla P, Hartl D. 2010. Low-complexity regions in *Plasmodium falciparum*: missing links in the evolution of an extreme genome. Mol Biol Evol 27:2198–2209.