# THE ROLE OF COMPLEX EVOLUTIONARY DYNAMICS IN MOLECULAR SEQUENCE ANALYSIS

---

A Dissertation
Submitted to
the Temple University Graduate Board

---

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

By
Alexander George Lucaci
May 2023

Examining Committee Members:

Professor Sergei L Kosakovsky Pond, Advisory Chair, Department of Biology
Professor Sudhir Kumar, Department of Biology
Professor Jody Hey, Department of Biology
Professor Spencer V Muse, External Member, North Carolina State University

©

**COPYRIGHT**
2023

by

Alexander G Lucaci
All rights reserved.

## ABSTRACT

Codon substitution models can be used to quantify selective pressure on molecular sequences. The work contained within this dissertation represents my efforts to create, validate, and test codon models, and to apply them to biologically diverse data sets. Specifically, my objective is to determine the impact and consequences of instantaneous multi-nucleotide mutational events (MH) on the statistical inference of evolutionary rate parameters. In evolutionary terms, MH events represent rare phenomena which may have strong effects on protein-coding genes. Evidence for MH events directly impact the patterns and processes of protein-coding gene evolution across species and time. My central hypothesis is that accounting for multi-nucleotide mutational events alters the estimation of evolutionary rate parameters and represents an alternative path a gene may embark upon during their evolutionary history. This hypothesis is fundamentally based on a synthesis of my own work, and others within the field, on the relationship of the evolutionary properties of functional coding regions of the genome. My rationale is that completion of my dissertation will result in the generation of new statistical methodologies and readily available computational implementations designed to identify key targets of evolutionary mechanisms, genomic patterns, and biological processes important for the functional adaptation of the genome as it relates to MH.

My long-term goal is to develop novel strategies to improve the biological realism of statistical models of molecular sequence evolution. I accomplish this goal by addressing gaps in conventional protein-coding gene model assumptions and with the inclusion of

additional biological, statistical, physiological, and evolutionary information. In this effort, I have created a codon model to account for MH events, with details described in Chapter 2. We have also made the model easily available on Datamonkey.org and provided several useful visualizations to aide in the interpretations of results.

To better understand the evolution of protein-coding genes, I applied my methods and other existing methods in the field to explore potentially novel biological insights into a gene family, the heat shock proteins, which play an important role as a protein chaperone, and in another paper, Brain-Derived Neurotrophic Factor (BDNF), which plays an important role in brain development. I developed software pipelines, curated data, and provided visualizations for the interpretation of results. The publications associated with this work are highlighted in Chapters 3 and 4, respectively.

To enable the exploration of molecular virology and the analysis of viral pathogens I developed "Rapid Assessment of Selection in CLades (RASCL)": a novel application for the rapid assessment of molecular sequence evolution viral clades. I used RASCL to study the emergence and ongoing evolution of SARS-CoV-2 lineage and to identify key sites subject to adaptive evolution and the development of new viral lineages. Near-real-time pathogen molecular surveillance is an important part of understanding the spread of disease. Our development of scalable tools to analyze big datasets of viral pathogen sequences is a critical step forward for global public health.

My methods and results can be used to translate existing molecular sequence data into novel insights, and to improve the understanding of important evolutionary systems, and all together constitute an accessible platform for quantifying selective pressure on molecular sequences.

## DEDICATION

I dedicate this dissertation to my mother and to the memory of my father,

Aurelia and Viorel Lucaci, and to my brother Raul Lucaci

for their continued love and support.

My parents arrived in the United States as political refugees from communist Romania,

to give their children a brighter future.

I am their American Dream.

To all my friends in the Bioinformatics program, thank you for making graduate school

bearable, and for the lifelong memories we have created.

## ACKNOWLEDGEMENTS

I would like to take this moment to thank all the people that have helped me through my scientific career. I enjoyed the camaraderie of my colleagues in the Bioinformatics program, especially the members of my graduate cohort Alyssa Pivirotto, Jordan Zehr, Steven Weaver, and Francisco McGee. I have also been fortunate to have an excellent group of lab mates including Sadie Wisotsky, who provided me with critical guidance in the early part of my PhD, Hannah Verdonk, Avery Grace Ann Selberg, Hannah Kim, Jordan Zehr, Stephen Shank, Ryan Velazquez, and Steven Weaver, who made the long hours of lab work a little easier. My graduate school experience was marked by the SARS-CoV-2 global pandemic, to which I made important contributions that led to publications in high profile journals and large international collaborations. The scientific process is a global effort to create and spread knowledge, I have been truly blessed to be among creative, intelligent, and hard-working people.

I would like to thank the fellow founding members of the Bioinformatics Studio (Alyssa Pivirotto, Jordan Zehr, and Francisco McGee), a hackerspace for the data sciences, that was created at Temple University, including our academic advisors Rob Kulathinal and Vincenzo Carnevale. This endeavor has led to a strong spirit in exploring computational biology, and the building of community at the undergraduate and graduate, and faculty level, and development of talent in the data sciences, with collaborations across university departments, and organizations. In addition, I would like to thank the members of my oversight and dissertation committee, Dr. Jody Hey, Dr. Sudhir Kumar, and Dr.

Spencer Muse for their mentorship and guidance. I would especially like to thank Drs. Jody Hey and Rob Kulathinal for their courage and vision, and members of the Bioinformatics program steering committee for the effort and commitment to create our Bioinformatics PhD program at Temple University, of which I am a member of the first-ever Ph.D. cohort.

I am forever indebted to Dr. Sergei L Kosakovsky Pond, for taking me under his scientific wing and providing me mentorship, room to learn, grow, fail, and explore ideas around molecular evolution. He has shown me what it is to be a scientific leader, advisor, and the importance of encouraging curiosity, and the power of consistently showing up. He is truly a scientist's scientist, and I am happy to have been fortunate enough for him to take chance on me and to have worked with him in the search for the truth on important scientific problems.

Chapter 2, in full, is a reprint of the material as it appears in "Lucaci AG, Wisotsky SR, Shank SD, Weaver S, Kosakovsky Pond SL. Extra base hits: Widespread empirical support for instantaneous multiple-nucleotide changes. PloS One. 2021 Mar 12;16(3):e0248337. Doi: 10.1371/journal.pone.0248337. PMID: 33711070; PMCID: PMC7954308.". The dissertation author was the primary investigator and a co-first author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in "Benndorf R, Velazquez R, Zehr JD, Pond SLK, Martin JL, Lucaci AG. Human HspB1, HspB3, HspB5 and HspB8: Shaping these disease factors during vertebrate evolution. Cell Stress

Chaperones. 2022 Jul;27(4):309-323. Doi: 10.1007/s12192-022-01268-y. Epub 2022 Jun 9. PMID: 35678958; PMCID: PMC9346038.”. The dissertation author made significant contributions as a senior investigator and last author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in “Lucaci, A.G., Notaras, M.J., Kosakovsky Pond, S.L. et al. The evolution of BDNF is defined by strict purifying selection and prodomain spatial coevolution, but what does it mean for human brain disease?. Transl Psychiatry 12, 258 (2022). https://doi.org/10.1038/s41398-022-02021-w”. The dissertation author was the primary investigator and a co-first author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in “Lucaci AG, Zehr JD, Shank SD, Bouvier D, Ostrovsky A, et al. (2022) RASCL: Rapid Assessment of Selection in CLades through molecular sequence analysis. PLOS ONE 17(11): e0275623. https://doi.org/10.1371/journal.pone.0275623”. The dissertation author was the primary investigator and the first author of this paper.

## TABLE OF CONTENTS

ABSTRACT ..... iii

DEDICATION ..... vi

ACKNOWLEDGEMENTS ..... vii

TABLE OF CONTENTS ..... x

LIST OF TABLES ..... xvii

LIST OF FIGURES ..... xx

CHAPTER 1 INTRODUCTION ..... 1

The Genetic World ..... 1

A Framework for Natural Selection Detection. ..... 2

Improving the Biological Realism of Evolutionary Models through MH Accountability ..... 6

Shedding Light on the Patterns and Processes of Evolution ..... 9

CHAPTER 2. EXTRA BASE HITS: WIDESPREAD EMPIRICAL SUPPORT FOR INSTANTANEOUS MULTIPLE-NUCLEOTIDE CHANGES ..... 12

Abstract ..... 12

Introduction ..... 13

Methods ..... 17

Substitution Models ..... 17

Site-level Support for MH .....20

Hypothesis Testing.....20

How Do Our Models Relate to Previously Published Multiple-hit Models? 21

Empirical Data .....22

Simulated Data.....22

Implementation and Availability .....24

Results..... 25

Benchmark Alignments .....25

Large-scale Empirical Analyses .....28

Strong Evidence for Non-zero Multiple-hit Rates. .....31

Factors Associated with MH Detection. .....31

Strong MH Signals Come from a Small Fraction of Sites. .....32

Patterns of Substitution Associated with MH Rates. .....33

Interaction Between Rate Estimates. .....35

Impact on Branch Length Estimates. .....36

Simulations .....36

Discussion..... 39

CHAPTER 3. HUMAN HSPB1, HSPB3, HSPB5 AND HSPB8: SHAPING THESE DISEASE FACTORS DURING VERTEBRATE EVOLUTION. .....42

Abstract ..... 42

Introduction ..... 43

Methods ..... 48

Retrieval of Vertebrate sHSP Sequences from the Databases ..... 48

Inclusion Criteria ..... 49

Species Included in the sHSP Sequence Analysis ..... 50

Determination of Codon-specific Evolutionary Pressure Along the Entire Length of the Human sHSPs ..... 51

Sequence Partitions ..... 56

Determination of the Evolutionary Pressure on Full-length Sequences and Sequence Partitions ..... 56

Results ..... 58

Selective Pressure Profiles Along the Human sHSP Sequences ..... 58

Differences in the Selective Pressures between the sHSPs (full-length sequences) ..... 60

Differences in the Selective Pressures between the Regions within the sHSPs ..... 61

Differences in the Selective Pressures between the Homologous Regions of the Four Studied sHSPs ..... 65

Differences in the Selective Pressures between the Exons .....66

Selective Pressure on the Missense Mutation Sites .....67

Discussion ..... 70

CHAPTER 4. THE EVOLUTION OF BDNF IS DEFINED BY STRICT PURIFYING SELECTION AND PRODOMAIN SPATIAL COEVOLUTION, BUT WHAT DOES IT MEAN FOR HUMAN BRAIN DISEASE? .....79

Abstract ..... 79

Introduction ..... 80

A Primer of the Molecular Biology of BDNF and its Functional Topology .82

The Conservation of BDNF and Neurotrophins: a Signal that Evolution is Important .....83

Purpose of this Study .....85

Results ..... 85

Evolutionary History of Mammalian BDNF .....86

Predominant Purifying Selection in BDNF .....87

Specific Sites that are Evolving Non-neutrally .....90

Evidence of Coevolutionary Forces .....91

Discussion ..... 95

Predominant Purifying Selection Across BDNF in Mammalia .....96

Non-neutral Positive Diversifying Evolution Sites in the BDNF Gene .....97

Discovery of Proximal and Distal Coevolving Site-Pairs in the BDNF Gene .....98

Potential Structural Implications of Evolving Sites.....99

Limitations of our Computational Evolutionary Analysis.....100

Future Directions: Understanding the Remainder of the Neurotrophin Family .....101

Conclusion ..... 102

Methods..... 102

Data Retrieval .....102

Data Cleaning.....135

Analysis of Orthologous Collections (AOC): Alignment, Recombination detection, Tree Inference, and Selection Algorithms.....136

Recombination Detection.....139

Data Availability..... 140

CHAPTER 5. RASCL: RAPID ASSESSMENT OF SELECTION IN CLADES THROUGH MOLECULAR SEQUENCE ANALYSIS. .....141

Abstract..... 141

Introduction..... 142

Materials and Methods ..... 144

RASCL Application Overview ..... 144

Map and Compress ..... 145

Prepare for Selection Analysis. ..... 146

Selection Analyses ..... 146

Software Availability ..... 149

Visualization and Downstream Post-Hoc Analyses ..... 149

Results and Discussion ..... 154

An Overview of Molecular Surveillance of Important SARS-CoV-2 Viral Clades ..... 154

A Closer Examination of the SARS-CoV-2 BA.5 clade ..... 156

Assessing Evolutionary Pressures on the SARS-CoV-2 μ (B.1.621) Clade ..... 160

Evidence of Natural Selection History Operating on SARS-CoV-2 Genomes ..... 162

Potential Biological and Clinical Significance of Mutations ..... 163

Conclusions ..... 171

CHAPTER 6. SUMMARY ..... 175

A Review of the Dissertation Chapters. ..... 175

What have I Accomplished? ..... 176

Why is it Important? ..... 177

What does the Future Hold for This Work? ..... 178

BIBLIOGRAPHY ..... 181

APPENDICIES ..... 207

APPENDIX A: Supplemental Material for Chapter 2 ..... 207

APPENDIX B: Supplemental Material for Chapter 3 ..... 213

APPENDIX C: Supplemental Material for Chapter 4 ..... 305

APPENDIX D: Supplemental Material for Chapter 5 ..... 315

## LIST OF TABLES

Table Page

1. Key parameters of our models……..18

2. Expressions for different types of substitutions in the model rate matrix……...19

3. Analysis of 13 benchmark datasets for evidence of multi-nucleotide substitutions…24

4. Evidence for multiple hit rates in empirical datasets…….29

5. Power to detect MH rates……..……39

6. MEME analysis of the BDNF gene…….91

7. The BGM analysis of BDNF……...95

8. Tabulation of species included in our analysis, comprising NCBI ortholog gene IDs, symbols, mammalian species, common name, and RefSeq accessions…….103

9. Examples of various recent RASCL analyses on variants of interest and variants of concern for SARS-CoV-2……..155

10. BUSTED[S] selection results on the BA.5 SARS-CoV-2 clade across segments…158

11. Site profiles for selected sites in the BA.5 Spike gene……160

12. A table of high-priority sites in SARS-CoV-2 μ B.1.621 sequences…….……164

13. Estimated omega rate distributions for benchmark datasets……..206

14. Reported disease-associated missense mutations in human HspB1, HspB3, HspB5 and HspB8a…….…….214

15. List of sHSP mRNAs (GeneBank and ENSMBLE identifiers) as used for the FEL analysis, the species of origin and their taxonomic ranking……219

16. FEL analysis of the coding sequences of human HspB1, HspB3, HspB5 and HspB8……..264

17. Aggregate omega values as returned by the FitMG94 algorithm of full-length sequences and of selected sequence partitions of human HspB1, HspB3, HspB5, and HspB8 ……299

18. Examples for reported molecular and cellular implications of sHSP missense mutations with potential toxic gain-of-function mechanism.....300

19. The FEL analysis of the BDNF gene found 174 of 261 (66.7%) sites to be statistically significant (LRT p-value <= 0.1) for pervasive negative (purifying) selection……...304

20. List of sites found to be under diversifying positive selection by MEME (p≤0.05) along internal branches in BA.1, as well as biochemical properties that are important at this site (via the PRIME method)……..315

21. List of sites found to be under diversifying positive selection by MEME (p≤0.05) along all branches in BA.1, as well as biochemical properties that are important at this site (via the PRIME method)…….322

22. List of sites found to be selected differentially along internal branches between BA.1 and background sequences (FDR ≤ 0.2) using the Contrast-FEL method……...333

23. List of sites found to be evolving under directional selection in the entire tree, using a FUBAR-like implementation of the DEPS method……...…….335

24. Pairs of sites (BA.1) found to have epistatic (co-evolving) substitution patterns by BGM method……..337

25. List of sites found to be under pervasive negative selection by FEL (p≤0.05)……340

26. BUSTED[S] selection results on the BA.1 SARS-CoV-2 clade across segments…...342

27. List of sites found to be under diversifying positive selection by MEME (p≤0.05) along internal branches in BA.5, as well as biochemical properties that are important at this site (via the PRIME method)……..344

28. List of sites found to be under diversifying positive selection by MEME (p≤0.05) along all branches in BA.5, as well as biochemical properties that are important at this site (via the PRIME method)…….348

29. List of sites found to be selected differentially along internal branches between BA.5 and background sequences (FDR ≤ 0.2) using the Contrast-FEL method……...353

30. List of sites found to be evolving under directional selection in the entire tree, using a FUBAR-like implementation of the DEPS method……354

31. Pairs of sites (BA.5) found to have epistatic (co-evolving) substitution patterns by BGM method……..355

32. List of sites found to be under pervasive negative selection by FEL ($p \leq 0.05$) along internal branches in BA.5……...357

33. List of sites found to be under diversifying positive selection by MEME ($p \leq 0.05$) along internal branches in B.1.621, as well as biochemical properties that are important at this site (via the PRIME method)……358

34. List of sites found to be evolving under directional selection in the entire tree, using a FUBAR-like implementation of the DEPS method……362

## LIST OF FIGURES

Figure Page

1. Exploring molecular evolution…….7

2. Archetypal sites based on model preferences……..28

3. Intersections of likelihood ratio test significance……30

4. Multiple hit detection rate……32

5. Three-hit substitutions commonly occurring in empirical data……...34

6. Three-hit non-synonymous substitutions and two-hit substitutions occurring in empirical data……..35

7. Selective pressures along the sequences of human HspB1 (a), HspB3 (b), HspB5 (c) and HspB8 (d) as returned by the FEL algorithm……55

8. Selective pressures estimated for the full-length sHSP sequences……..58

9. Selective pressures estimated for the various sHSP domains and regions……..64

10. Selective pressures estimated for the exons of HspB1, HspB5, and HspB8……...67

11. Summary of the selective pressures detected at the disease associated missense mutation sites of the four studied sHSPs…….69

12. Mammalian BDNF exhibits overwhelmingly strict purifying selection, but also evidence of specific evolutionary pressures at particular sites……89

13. The BGM analysis of BDNF……...93

14. BDNF NT-4 heterodimer structural analysis to highlight coevolving and adaptive sites at the 3D protein level…….……..94

15. The RASCL application overview…….151

16. Example visualization using our interactive notebook…….……..153

17. Evolutionary trajectories of 40 high-priority selected sites……..……...166

18. Temporal trends of the substitution combinations at sites in Spike……..……...168

19. Temporal trends of the substitution combinations at sites in RdRp……...……..169

20. Spike protein crystal structure annotation with MEME sites……..……170

# CHAPTER 1

# INTRODUCTION

## The Genetic World

Genes and their fundamental building blocks, deoxyribonucleic acid (DNA), are the storage information of life, and are central to the passage of information from generation to generation and are of direct interest in the study of evolutionary genomics. Genes are fundamental in the study of how populations and species change over time to meet selective pressure and life challenges. DNA is a part of the central dogma of biology, and it is critical to study the patterns and processes which contribute to functionality and importantly allow for subsequent genetic diversity to push evolutionary innovation. Genetic variation and its consequences are a useful source of information for studying the differences in mutational events across species in modern comparative genomic analyses. The velocity of gene evolution can vary by orders of magnitude from rapidly evolving genes involved in host-pathogen interactions to the slowly evolving so-called "housekeeping" genes involved in ancient pathways and mechanisms including cellular metabolism and regulatory DNA interaction dynamics.

Codons are regions of the DNA where a short motif (three nucleotides in the standard "universal" code, although alternatives exist in synthetic biology) are translated to protein coding amino acids, are often a source of genetic mutations and frequently and generally take the form of a common single nucleotide polymorphism (SNP).

**A Framework for Natural Selection Detection.**

A first-order Markov chain is a mathematical system that undergoes transitions from one state to another according to certain probabilistic rules. In other words, the probability of transitioning to any particular state is dependent solely on the current state and time elapsed. Our current operational framework for examining evolutionary history consists of finite state and continuous-time Markov chains which are used to model the evolution of molecular sequences. In the context of molecular evolution, a first-order Markov chain can be used to model the probability of a sequence of nucleotides (or amino acids) changing over time, especially over higher branches in the evolutionary tree. This is often used in bioinformatics to study how substitutions arise and propagate across species in a comparative genomic study. In a first-order Markov chain model of molecular evolution, the current state of the system is represented by a sequence of nucleotides (or amino acids), and the possible future states are all the possible sequences that can be derived from the current sequence by a single mutation (typically), which I move to challenge across my body of work. The probability of transitioning from one state to another is determined by the rules of the model, which may be based on biological data such as the rates of different types of substitutions. The key features of a first-order Markov chain in molecular evolution are: the current state of the system is represented by a sequence of nucleotides, or codons (or amino acids). The possible future states are determined by the rules of the model, which may be based on biological data. The probability of transitioning from one state to another is determined by the current state and

the time elapsed, but not by the past states of the system. Overall, a first-order Markov chain in molecular evolution is a probabilistic model that can be used to study the evolution of nucleotide, codon (or amino acid) sequences over time.

For molecular sequence evolution, the sites are typically individual nucleotides (typically corresponding to the Watson-Crick nucleotides and base pairs), forming members of the codon table, or the standard set of amino acids. The way we build out evolutionary models typically consists of a large rate matrix defined by our sequence substitution model and describes the exchanges between evolutionary states. In molecular evolution, the Q matrix and P matrix are used to describe the rates of change between different nucleotide or amino acid states over time. The Q matrix, also known as the substitution matrix, specifies the rate at which one nucleotide or amino acid state is expected to change into another state. Where each $q_{ij}$ is the instantaneous rate of change from state *i* to state *j*, with each change assumed to occur at different rates. Most models define the Q matrix to only allow for a single change in state to occur, a limiting factor which I have endeavored to overcome. The General Time Reversible (GTR) (Tavaré 1986) is a nucleotide model with 4 possible states, corresponding to the Watson-Crick nucleotide. For the GTR model, it is assumed that the frequencies of each nucleotide $\pi$ are not equal, the parameters (estimated from the data or from prior knowledge) can reflect the composition of a given sequence, such as higher GC content. The P matrix, also known as the transition probability matrix or P(t), describes the probability of any given nucleotide or amino acid state changing into any other state after a specific discrete period. These

matrices are used to model the evolution of biological sequences over time and can be estimated from empirical data. The transition probability of a site *s* changing from state *i* to state *j* in time t is denoted $p_{ij}(t)$. The relationship between the transition probability matrix P(t) and the instantaneous rate matrix Q can be solved to give the relationship $P(t) = e^{Qt}$.

Finite state continuous-time Markov chains provide a simple and convenient way to model sequence evolution at the molecular level. Importantly, codon substitution models which build upon Markov chains can also incorporate other biologically relevant parameters. Codon models utilize the codon (a triplet of nucleotides, described above) as the unit of evolution instead of individual nucleotides or amino acids. A codon model's advantage lies in its ability to account for the redundancy, and translation of the genetic code (which determines the behavior of a mutation) and incorporate dependencies among sites within a codon. Codon models are used to account for the rates of synonymous (dS) and nonsynonymous (dN) changes in nucleotide sequences, while amino acid models can only describe nonsynonymous changes.

The likelihood method is a statistical technique that is used to estimate the parameters of a model. The general idea behind this method is to calculate the likelihood of the observed data, given a particular set of model parameters. This likelihood is then used to determine the most likely values for the model parameters.

The likelihood of the observed data, given a set of model parameters, can be expressed mathematically as follows: $L(\theta \mid D) = P(D \mid \theta)$ where $L(\theta \mid D)$ is the likelihood of the data D, given the model parameters $\theta$.

In the context of molecular evolution, the maximum likelihood (ML) approach is a method for estimating the evolutionary parameters that best explain a given set of data. This method is based on the principle of maximum likelihood, which involves finding the set of parameters that maximize the likelihood of the data given the model. The maximum likelihood approach involves specifying a probabilistic model of molecular evolution, which describes the processes that give rise to the data. This model is used to compute the likelihood of the data for different sets of parameters, and the set of parameters that maximizes the likelihood is chosen as the best estimate. Overall, the maximum likelihood approach is a powerful tool for studying molecular evolution, by providing estimates of the evolutionary parameters that best explain the data.

As it relates to Felsenstein's work, this likelihood function is often used to estimate the parameters of evolutionary models, such as the rates of genetic mutation and the patterns of inheritance among different species. Felsenstein developed several algorithms (Felsenstein 1981) and statistical methods that are based on the likelihood method, and these have been widely used in the field of evolutionary biology including a seminal paper that introduced an efficient algorithm for performing maximum likelihood calculations to estimate evolutionary trees from DNA sequence data (Felsenstein 1981). In summary, the likelihood method is a statistical technique that is used to estimate the parameters of a

model, and it is particularly useful in the context of evolutionary biology where the behavior of complex systems needs to be understood and predicted.

## Improving the Biological Realism of Evolutionary Models through MH Accountability

The standard approach for codon substitution models in widespread use is to assume that the probability of complex multi-nucleotide events occurring within a codon (MH) is negligible and therefore unnecessary to model in the generating process of the data. The central foundation to this stepwise mutation model is the construction of a sparsely coded rate (Q) matrix (introduced above). In contrast to the simple single mutation stepwise codon substitution model, our models consider broadly applicable MH mutational events and directly improves upon the reality of biological models. Our work fits the Q matrix with densely coded information, and thus does not assume that gene evolution occurs outside of the light of MH events thereby “illuminating the darkness” in molecular evolution.

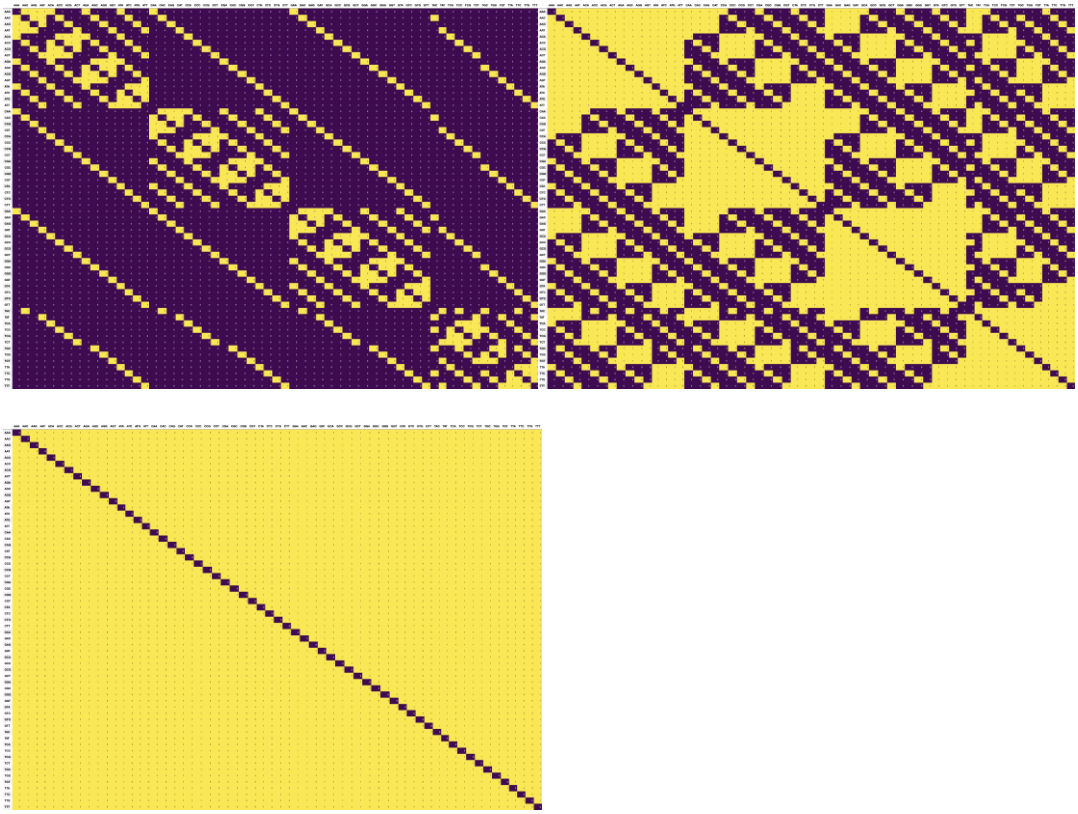


Figure 1.1 Exploring molecular evolution. These figures demonstrate how the Q matrix is encoded in a common codon substitution model for single-hit and multi-hit models. These figures demonstrate the concept of allowing for MH events and transition from a sparsely coded matrix to a densely coded one. Yellow designates a cell which receives an entry in the Q matrix, the color Purple that it does not, therefore the entry designates zero versus non-zero values. If you remove the diagonal (e.g., AAA → AAA) from calculation 14.37% of the SH matrix is being encoded, while 57.21% for SH/DH inclusive model and 100% for SH/DH/TH model. Rows designate the "starting" codon, Columns the "target" codon.

To accomplish this goal, I developed and validated a modified version of the Muse-Gaut 1994 (Muse and Gaut, 1994) (MG94) codon substitution model, which I use to investigate the nature and extent of MH events. I constructed individual-gene level results, providing a functional analysis of the underlying biological mechanisms and creating a profile from which to identify the best scenarios to apply MH-aware models.

Throughout our analysis, I make scientific predictions based on the properties of the models and the rates of evolution. In application, I find that the pattern of gene MH events preservation is consistent with occurrences on short branches in the phylogenetic tree, where the effect of saturation of mutations at neighboring nucleotides is minimized. This result is of scientific interest, since it changes the conclusions of earlier work in the MH literature and provides insights in this persistent biological phenomenon, and sheds light on alternative evolutionary paths genes may embark upon during their life history. Multinucleotide substitutions are changes to a sequence of nucleotides that affect multiple nucleotides at once. These changes can have a significant impact on the evolutionary relationships between different species, because they can cause large changes to the sequences of nucleotides that are used to estimate phylogenetic trees and in the inference of evolutionary rate parameters. In the context of short branches on a phylogenetic tree, multinucleotide substitutions can cause rapid changes in the sequences of nucleotides, leading to significant differences between the sequences of closely related species.

Overall, multinucleotide mutations on short branches of the phylogenetic tree can have a significant impact on the accuracy of phylogenetic reconstructions, and evolutionary rate inference, and should be considered when studying molecular evolution.

In this dissertation, I construct and apply bioinformatics principles and statistical models towards the understanding of the evolution of genes due to the special circumstances of multiple instantaneous mutations (MH) occurring within a codon. I analyze my models to validate them, observe their behavior, and to make measured scientific predictions from which I can derive novel results.

**Shedding Light on the Patterns and Processes of Evolution**

Empirical studies have suggested that the rate of MH events can depend on several biological mechanisms including the error-prone polymerase zeta. However, I find few instances in the literature that account for MH events and they are based on models that are parameterized differently from ours and do not have easily available software implementations. To address this gap, I introduce a series of codon models based on continuous-time Markov processes that account for multiple instantaneous mutations in a single codon. Our models account for the biological process of multiple mutation and its observed dynamics and effect on the rate of molecular evolution. Specifically, our proposed models' events in which multiple nucleotide mutations occur instantaneously within a codon of a protein coding sequence and have risen to fixation within a population. I accomplish this within the Hypothesis Testing using Phylogenies (HyPhy) software suite,

a platform for the analysis of molecular sequence data (Kosakovsky et al., 2005). Additionally, I use the Datamonkey (https://www.datamonkey.org), a web-based gateway to offer an easy-to-use interface with our model (Kosakovsky and Frost, 2005) along with providing graphical review of results via Hyphy-Vision (http://vision.hyphy.org/). A current implementation of our software is available at: http://datamonkey.org/multihit.

In Chapter 2, I will present the FitMultiModel (FMM) model, implemented in HyPhy, which is a codon substitution model which accounts for MH events. I will describe its implementation design, demonstrate its results in empirical data, and compare its performance on simulated data, all against standard models.

In Chapter 3, I will present a shift in the narrative of this dissertation from the creation (validation and testing) of codon models to their application in empirical datasets for biological investigation. I will show our results, data aggregation and curation, and selection analysis on a part of the Heat Shock Protein (HSP) family, across a broad taxonomic range in vertebrates.

In Chapter 4, I will present an evolutionary analysis of the Brain Derived Neurotrophic Factor (BDNF) gene, which plays a critical role in brain development. I will show our data aggregation method through publicly available databases, data curation, genetic recombination testing, and the novel results of our selection analysis using modern techniques.

In Chapter 5, I will present recent work in virology and molecular evolution which involves workflow and application development for understanding viral gene evolution.

I developed an application “Rapid Assessment of Selection in Clades” (RASCL). When applied, this work contributes to the SARS-CoV-2 effort by discovering areas of ongoing evolution in viral clades outside of signature defining mutations. Specifically, the near-real-time monitoring of the population-level selective processes that will underlie the emergence of future variants of concern in rapidly evolving pathogens with extensive genomic surveillance.

In summary, this dissertation encompasses both the creation and application of codon models for quantifying selective pressure on molecular sequences. At every step I emphasize usability, and interpretability of our results through useful visualizations, and contextualize our work in relation to what is known in the literature.

# CHAPTER 2.

# EXTRA BASE HITS: WIDESPREAD EMPIRICAL SUPPORT FOR INSTANTANEOUS MULTIPLE-NUCLEOTIDE CHANGES


## Abstract

Despite many attempts to introduce evolutionary models that permit substitutions to instantly alter more than one nucleotide in a codon, the prevailing wisdom remains that such changes are rare and generally negligible or are reflective of non-biological artifacts, such as alignment errors. Codon models continue to posit that only single nucleotide changes have non-zero rates. Here, we develop and test a simple hierarchy of codon-substitution models with non-zero evolutionary rates for only one-nucleotide (1H), one- and two-nucleotide (2H), or any (3H) codon substitutions. Using over 42,000 empirical alignments, we find widespread statistical support for multiple hits: 61% of alignments prefer models with 2H allowed, and 23%—with 3H allowed. Analyses of simulated data suggest that these results are not likely to be due to simple artifacts such as model misspecification or alignment errors. Further modeling reveals that synonymous codon island jumping among codons encoding serine, especially along short branches, contributes significantly to this 3H signal. While serine codons were prominently involved in multiple-hit substitutions, there were other common exchanges contributing to better model fit. It appears that a small subset of sites in most alignments have unusual evolutionary dynamics

not well explained by existing model formalisms, and that commonly estimated quantities, such as dN/dS ratios may be biased by model misspecification.

Our findings highlight the need for continued evaluation of assumptions underlying workhorse evolutionary models and subsequent evolutionary inference techniques. We provide a software implementation for evolutionary biologists to assess the potential impact of extra base hits in their data in the HyPhy package and in the Datamonkey.org server.

## Introduction

Most modern codon models in widespread use assume that any changes within a codon happen as a sequence of single instantaneous nucleotide changes, enforced by setting instantaneous rates between codons that differ in more than one nucleotide to zero. This choice was made independently for the mechanistic models of Muse and Gaut (Muse and Gaut, 1994) and Goldman and Yang (Goldman and Yang, 1994), and adopted by subsequent model developers and model users. For example, when Halpern and Bruno (Halpern and Bruno, 1998) introduced their mutation-selection models, they considered the general multi-hit (MH) case first, but then largely abandoned it, noting that the single hit reduction "..has very little effect on our results under the conditions we have investigated." This assumption is both computationally convenient and biologically sound in the majority of cases, since randomly occurring mutations "hitting" the same codon is a negligibly rare event. While these events are indeed rare, evidence for substitutions

occurring in tandem at adjacent nucleotide sites had been reported at about the same time the codon models were being introduced (Wolfe and Sharp, 1993).

Averof et al (Averof et al., 2000) reported significant rates of changes between TCN and AGY codon islands in perfectly conserved serine residues, and argued against going through intermediary non-synonymous changes due to their likely deleterious effects. Rogozin et al (Rogozin et al., 2016) took the opposite view, namely that strong purifying selection on single nucleotide changes is a more plausible explanation for such island hops in general.

Neither of those studies had considered an explicit evolutionary model, however. Serine is the only amino-acid with synonymous codon islands in the universal genetic code, but several other codes have other amino acids with this property: leucine in the *Chlorophycean* and *Scenedesmus obliquus* mitochondrial codes (TAG and CTH), and alanine in the *Pachysolen tannophilus* nuclear code (CTG and GCH).

Recent studies estimated that 2% of nucleotide substitutions are part of larger multiple nucleotide changes that occur simultaneously (Harris and Nielsen, 2014; Kaplanis *et al.*, 2019), due in part to an error-prone DNA polymerase zeta. Human germline tandem mutations may constitute up to 0.4% of all mutations (Chen et al, 2014), and individual cases of such mutations have significant phenotypic consequences, e.g., via their effects on protein folding (Okada et al 2017).

A number of codon model extensions have incorporated MH, invariably finding improvement in fit and (if the model allowed testing) statistically significant evidence of non-zero rates involving multiple nucleotide changes. Kosiol et al (Kosiol et al, 2007) developed a general MH empirical codon substitution model estimated jointly from a large collection of training alignments and noted that it was overwhelmingly preferred to standard SH models on a sample of biological data from the PANDIT database.

Several groups have independently created alternative codon model parameterizations to allow for MH, including Whelan and Goldman (Whelan and Goldman, 2004) ("… these events [MH] are far more prevalent than previously thought"), Zaheri et al (Zaheri et al 2014), and Dunn et al (Dunn et al, 2019) (the latter two studies show a dramatically better model fit to empirical alignments when allowing MH). Other studies have used evolutionary models with varying degrees of accommodation for multiple hits (Doron-Faigenboim and Pupko, 2007; Miyazawa et al, 2011; Zoller et al 2012; De Maio, 2013). Jones et al (Jones et al, 2019) implemented a complex model to detect adaptive evolution with a discrete-state phenotype, allowing for double and triple mutations to be absorbed into the parameter estimates. Despite multiple introductions to the field, these models have been unable to gain attraction in applied evolutionary analyses, and for some of these methods, software implementing them is no longer available.

Failure to include multiple hits in codon substitution models may mislead evolutionary hypothesis testing. Venkat et al (Venkat et al, 2018) found that the addition of a double-hit rate parameter improved model fit and impacted branch-specific inferences

of positive selection (MH along short branches can inflate false positives). Dunn et al (Dunn et al, 2019) used principled simulation studies to show that fitting 1H models to data generated with low rates of multiple hits can increase false positive rates and dilute power for identifying individual sites subject to positive selection.

In this study, we develop simple extensions to the Muse-Gaut (Muse and Gaut, 1994) codon model which add double, and triple instantaneous (2H, 3H, respectively) changes and compare them to simpler models in large collections of empirical data.

Our models are mechanistic and simpler than those proposed by Whelan and Goldman (Whelan and Goldman, 2004) and Dunn et al (Dunn et al, 2019). This relative simplicity allows our models to be implemented and fitted quickly, and offers straightforward interpretation, including the ability to identify individual sites that benefit from the addition of MH. The primary goals of our data analyses are to establish how often evidence for multiple hits can be detected in large-scale empirical databases (something that no other study looking at evolutionary models has done), identify the codons that are frequently involved in such events, and explore plausible biological explanations for why these rates are non-zero for a majority of alignments.

## Methods

### *Substitution Models*

The most general model considered here is the 3H+ substitution model and all others can be derived from it as special cases (see Table 2.1 for key parameter notation).

The model is a straightforward extension of the Muse-Gaut style of time-reversible, continuous Markov processes model (Muse and Gaut, 1994). In this study we compare five models:

- **1H**. is the standard Muse-Gaut style model which only permits single nucleotides to substitute instantaneously. Non-synonymous changes occur at rate $\omega$ (relative to the synonymous rate), and this rate varies from site to site according to a three-bin general discrete distribution (GDD).
- **2H**. is the 1H model extended to allow two nucleotides in a codon to substitute instantaneously with rate $\delta$ (relative to 1H synonymous rate).
- **3HSI**. is the 2H model extended to allow three nucleotides in a codon to substitute instantaneously if the change is synonymous (e.g., serine islands), with relative rate $\psi_s$.
- **3H**. is the 2H model extended to also permit any three-nucleotide substitutions, with relative rate $\psi$.
- **3H+**. is the 3HSI model extended to also permit any three-nucleotide substitutions, with relative rate $\psi$.

All codon substitutions in these models fall into one of six categories defined by (i) whether or not they are synonymous or non-synonymous, and (ii) by how many nucleotides are being replaced (1, 2, or 3). The instantaneous rate expression for substitutions between codons $i$ and $j$ ($i \neq j$) for these six classes, and how many of all possible codon substitutions are in each class,

Table 2.1. Key parameters of our models. Estimation strategies for each parameter for the five different models are shown. GDD = general discrete distribution; 1H, 2H, 3H = instantaneous changes involving one, two, or three nucleotides.

| **Parameter** | **Description** | **Model** | | | | |
|---|---|---|---|---|---|---|
| | | **1H** | **2H** | **3HSI** | **3H** | **3H+** |
| $\omega_i$ | Site-specific dN/dS ratio | Random effect 3-bin GDD distribution | | | | |
| $\delta$ | Global 2H/1H rate ratio | 0 | Estimated | Estimated | Estimated | Estimated |
| $\psi_s$ | Global 3H/1H rate ratio for synonymous codon islands | 0 | Estimated | Estimated | $= \psi$ | Estimated |
| $\psi$ | Global 3H/1H rate ratio | 0 | 0 | 0 | Estimated | Estimated |

Table 2.2. Expressions for different types of substitutions in the model rate matrix. Six cases for instantaneous rates ($q_{ij}$) of substituting codon *i* with codon *j* ($i \neq j$). The "Count" columns show the number of rate matrix entries in each class (excluding the diagonal) for two commonly used genetic codes.

| Type | Expression for $q_{ij}$ | Example | Count | |
|---|---|---|---|---|
| | | | Universal | mtDNA |
| 1H synonymous | $\theta_{ij}\pi_j$ | $ACA \rightarrow ACT: \ \theta_{CT}\pi_T^3$ | 134 | 128 |
| 1H non-synonymous | $\omega^k\theta_{ij}\,\pi_j$ | $AAA \rightarrow AGA: \omega^k\theta_{AG}\pi_G^2$ | 392 | 380 |
| 2H synonymous | $\delta\prod_{n=1}^{2}\theta_{ij}^n\pi_j^n$ | $CTC \rightarrow TTA: \delta\theta_{CT}\theta_{AC}\pi_T^1\pi_A^3$ | 28 | 16 |
| 2H non-synonymous | $\delta\omega^k\prod_{n=1}^{2}\theta_{ij}^n\pi_j^n$ | $AAA \rightarrow ACC: \delta\omega^k\theta_{AC}\theta_{AC}\pi_C^2\pi_C^3$ | 1540 | 1500 |
| 3H synonymous | $\psi_s\prod_{n=1}^{3}\theta_{ij}^n\pi_j^n$ | $AGC \rightarrow TCA: \psi_s\theta_{AT}\theta_{CG}\theta_{AC}\pi_T^1\pi_C^2\pi_A^3$ | 12 | 12 |
| 3H non-synonymous | $\psi\omega^k\prod_{n=1}^{3}\theta_{ij}^n\pi_j^n$ | $GTG \rightarrow TAC: \psi\omega^k\theta_{GT}\theta_{AT}\theta_{CG}\pi_T^1\pi_A^2\pi_C^3$ | 1554 | 1504 |

are shown in Table 2.2. In addition to key model parameters defined in Table 2.1, the model contains a number of other, standard parameters, which are not the main focus of inference and can be viewed, for the most part, as nuisance parameters. They include $\theta_{ij}$: nucleotide-level biases coming from the general time reversible model (5 parameters), and $\pi_j$ are codon-position specific nucleotide frequencies estimated from counts using the CF3x4 procedure (Kosakovsky Pond et al, 2010). $\omega^k$ are non-synonymous / synonymous rate ratios which vary from site to site using a random effect (D-bin general discrete distribution, D = 3 by default, 2D − 1 parameters). The key parameters are global relative rates of multiple hit substitutions: $\delta$ is the rate for 2H substitutions relative to the synonymous 1H rate (baseline), $\psi$—the relative rate for non-synonymous 3H substitutions, and $\psi_s$—the relative rate for synonymous 3H substitutions.

All parameters, except π, including branch lengths are fitted using directly optimized phylogenetic likelihood in HyPhy (Kosakovsky Pond et al, 2020). Initial estimates for branch lengths and θ are obtained using the standard nucleotide general time reversible model. Following this initialization, models are fitted in the order of increasing complexity (1H, then 2H, then 3HSI, then 3H+), using parameter estimates from each stage as initial points for the next stage.

### *Site-level Support for MH*

In order to identify which individual sites show preference for MH models, we use evidence ratios (ER), defined as the ratio of site likelihoods under two models being compared, e.g., ER(2H:1H) $\coloneqq \frac{L(s_i|\,2H)}{L(s_i\,|\,1H)}$ . We previously showed that ER are useful for identifying the sites driving support for one model over another (Murrell et al, 2015), and they incur trivial additional overhead to compute once model fits have been performed.

### *Hypothesis Testing*

Nested models are compared using likelihood ratio tests with χ2 asymptotic distribution used to assess significance. The degrees of freedom (d) parameter are as follows: d = 1 for 2H:1H, 3SHI:2H, and 3H+:3HSI comparisons; d = 2 for 3H+:2H comparison; d = 3 for 3H+:1H comparison.

***How Do Our Models Relate to Previously Published Multiple-hit Models?***

The BS+MNM model (Venkat et al, 2018) was designed for testing subsets of branches for episodic diversifying selection. It is very similar to our 2H model, except that $\theta_{ij}$ in their model follows the HKY85 parameterization, it is possible to allow $\kappa$ (transition/transversion ratio) to be different between 1H and 2H changes, and target codon frequencies are used in $q_{ij}$ (Goldman and Yang, 1994). The Empirical Codon Model or ECM (Kosiol et al, 2007) directly estimates numerical rates for all pairs of codon exchanges in the (Goldman and Yang, 1994) GY94 frequency framework from a large training dataset. The SDT model (Whelan and Goldman, 2004) uses a context-averaging approach to include the effect of substitutions that span codon boundaries and is difficult to directly relate to our models; the 3H model might be the closest to the SDT model. Regrettably, there does not seem to exist a working implementation of the SDT model (pers. comm from Simon Whelan), which makes direct comparison to our approaches impractical. The KCM model (Zaheri et al, 2014) only has a single rate for multiple hits (double or triple), and has position-specific nucleotide substitution rates ($\theta$ in our notation), so it would be most comparable to the 3H model with $\delta = \psi$.

The GPP model class (Dunn et al, 2019) can be parametrized to recapitulate our models because it can capture (in a log-linear parametric form) arbitrary rate matrices with suitable parametric complexity. Several of the models in the GPP class include multiple hits, but they are not directly comparable to ours, mostly because they also incorporate $\omega$

rates that depend on physicochemical properties of amino acids, and because the exact parametric form of the models are hard to glean from available description.

### *Empirical Data*

The Selectome (Moretti et al, 2014) data collection consists of 13,714 gene alignments from the Euteleostomi clade of Bony Vertebrates from Version 6 of the database (Moretti et al, 2014) and can be downloaded from data.hyphy.org/web/busteds/. The Shultz et al data collection (Shultz et al, 2019) contains 11,262 orthologous protein coding genes from 39 different species of birds and is freely available at https://datadryad.org/stash/dataset/doi:10.5061/dryad.kt24554. The Enard et al data collection (Enard et al, 2016) includes 9,861 orthologous coding sequence alignments of 24 mammalian species and is available at https://datadryad.org/stash/dataset/doi:10.5061/dryad.fs756. Our mtDNA dataset consists of both invertebrate and vertebrate Metazoan orders with mitochondrial gene alignments. This dataset was originally published in Mannino et al (Mannino et al, 2020), and can be found at https://github.com/srwis/variancebound.

### *Simulated Data*

We generated simulated alignments of two sequences in HyPhy using the 'SimulateMG94' package from https://github.com/veg/hyphy-analyses/. These alignments were simulated under the 1H (no site-to-site rate variation) with varying sequence and branch lengths as well as varied but constant ω across sites but no multiple hits. We created

1000 simulations scenarios to capture a range of important model parameters and drew 5 replicates per scenario. $\omega$ was drawn uniformly from U(0.01, 2.0), branch length was drawn Exp(U(0.01, 1.0)), and codon lengths as an integer from 100 to 5000 uniformly. Parameter values were sampled using the Latin Hypercube approach to improve parameter space coverage. Multiple sequence simulations were based on the fits to one of four benchmark datasets: Drosophila adh, Hepatitis D antigen, HIV *vif*, and the Vertebrate rhodopsin data. We took all model parameters estimated under the 3H+ model as the starting point and generated 500 replicates per dataset of which 35% were null (1H), 10% each from 2H, 3SHI or restricted 3H+ ($\psi_s = 0$), and 35% from 3H+. $\delta$, $\psi$ and $\psi_s$ parameters, when allowed to be non-zero by the model, were sampled from U(0, 1), U(0, 1), and U(0, 10), respectively.

Table 2.3. Analysis of 13 benchmark datasets for evidence of multi-nucleotide substitutions. N—number of sequences, S—number of codons, T—total tree length (expected subs/site) under the 1H model, two-hit (δ) rate estimate under the 3H model (2H model in parentheses), there-hit synonymous island date (ψs) estimate under the 3H model, three-hit rate (ψ) estimate under the 3H model. Likelihood ratio test p-values for all pairs of nested models e.g., 2H:1H—2H alternative, 1H null. Values <0.05 are bolded. # sites with ER > 5 lists the number of sites which show strong preferences for 2H or 3H model using evidence ratios of at least 5 (see text).

| Gene | N | S | T | $\delta$ | $\psi_s$ | $\psi$ | LRT p-value | | | | | # sites with $ER > 5$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | 2H:1H | 3H+:1H | 3H+:2H | 3H+:3HSI | 3HSI:2H | 2H:1H | 3H+:2H |
| *β*-globin | 17 | 144 | 2.5 | 0.7 (0.81) | >100 | 0 | **<0.001** | **<0.001** | **<0.001** | 1 | **<0.001** | 10 | 6 |
| Flavivirus NS5 | 18 | 342 | 6 | 0.49 (0.73) | 2.3 | 0.6 | **<0.001** | **<0.001** | 0.056 | 0.062 | 0.13 | 16 | 0 |
| Primate Lysozyme | 19 | 130 | 0.24 | 0 (0) | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| COXI | 21 | 510 | 5.3 | 0.4 (0.4) | 0 | 0 | **<0.001** | **0.0018** | 1 | 0.94 | 0.98 | 3 | 0 |
| Drosophila *adh* | 23 | 254 | 1.4 | 0.31 (0.4) | 0 | 0.42 | **<0.001** | **<0.001** | 0.19 | 0.067 | 0.99 | 4 | 0 |
| Encephalitis *env* | 23 | 500 | 0.84 | 0.076 (0.076) | 0 | 0 | 0.19 | 0.42 | 1 | 1 | 0.98 | 0 | 0 |
| Sperm lysin | 25 | 134 | 2.8 | 0.4 (0.46) | 2.3 | 0.3 | **<0.001** | **<0.001** | **0.04** | **0.015** | 0.49 | 21 | 1 |
| HIV-1 *vif* | 29 | 192 | 0.96 | 0.007 (0.044) | 0 | 0.17 | 0.058 | **0.0013** | **0.0077** | **0.0018** | 0.95 | 0 | 2 |
| Hepatitis D virus antigen | 33 | 196 | 1.9 | 0.34 (0.37) | 0 | 0.2 | **<0.001** | **<0.001** | 0.25 | 0.098 | 0.99 | 15 | 0 |
| Vertebrate Rhodopsin | 38 | 330 | 3.9 | 0.54 (0.72) | 9.2 | 0.9 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **0.0029** | 43 | 3 |
| Camelid VHH | 212 | 96 | 15 | 0.29 (0.32) | 0 | 0.13 | **<0.001** | **<0.001** | **0.011** | **0.0026** | 0.92 | 46 | 0 |
| Influenza A virus HA | 349 | 329 | 1.4 | 0.06 (0.06) | 0 | 0.0093 | **<0.001** | **<0.001** | 0.95 | 0.74 | 0.98 | 5 | 0 |
| HIV-1 *RT* | 476 | 335 | 6.6 | 0.086 (0.093) | 0 | 0.048 | **<0.001** | **<0.001** | 0.15 | 0.052 | 1 | 17 | 1 |

### *Implementation and Availability*

All analyses were performed in HyPhy version 2.5.1 or later (Kosakovsky Pond et al, 2020). The fmm (FitMultihitModel) module used to fit the standard 1H model along with 2H, 3H and 3HSI versions is available from: https://github.com/veg/hyphy-analyses/, and is a part of the standard library (invoked with hyphy fmm) in HyPhy version 2.5.7 or later and in our datamonkey.org server (Weaver et al, 2018).

The result of an fmm analysis is a JSON file which can be visualized using a web-application at http://vision.hyphy.org/multihit.

## Results

### *Benchmark Alignments*

We introduce the models using a collection of thirteen representative alignments Table 2.3 that we and others have recently used to benchmark selection analyses (Wisotsky et al, 2020). We also include a primate lysozyme alignment originally analyzed with early codon models by Yang (Yang, 1998). We consider five models (see Table 2.1 and the methods section for details), which form a nested hierarchy (with the exception of 3HSI and 3H which are not nested), each with one additional alignment-wide parameter.

1. Evidence for multiple hits is pervasive. In ten of thirteen datasets the analyses strongly reject the hypothesis that 2H have zero rates, with $p < 0.001$ (2H:1H comparison). For five of thirteen datasets, we can further reject the hypothesis that 3H have zero rates (3H+:2H comparison) at $p \leq 0.05$.
2. Varied patterns for rate preferences. Even in this small collection of datasets, the entire spectrum of options is present. For the Primate Lysozyme dataset there is no evidence for anything other than 1H changes, while for the Vertebrate Rhodopsin dataset each of the individual rates is significantly different from 0. HIV-1 *vif* dataset is the only dataset that does not support 2H rates but does support 3H rates. Five datasets share a pattern: reject 1H in favor of 2H, and 1H in favor of 3H+, but

none of the others, which can be interpreted as support for 2H rates, but none of the 3H rates.

3. Varied extent of site-level support for MH. Ratios between site-level likelihoods under individual models, denoted here as ER (evidence ratios), can indicate which model provides better fit to the data at a particular site. The number of sites with strong (ER > 5) preference for 2H vs 1H model was positive for all models rejecting 1H in favor of 2H with LRT, and ranged from 3 to 46, while a smaller number of sites (0 − 6) preferred 3H+ to 1H. Interestingly, for Camelid VHH, where the LRT rejects 1H in favor of 3H+, no individual sites had ER > 5, implying that the support for this model came from a number of individual weak site contributions.
4. Interaction between 1H, 2H and 3H rates. Assuming that the biological process of evolution does include MH events, not modeling them appropriately might have the effect of inflating other rate estimates.

In line with other studies (Dunn et al, 2019), the addition of 2H rates lowers the point estimates of ω rates for all datasets where 2H:1H comparison is significant at $p \leq 0.05$ (Appendix A, Table A.S1) sometimes dramatically (e.g., by a factor of 0.6× for the β−globin gene). This could be indicative of estimation bias due to model misspecification. Similarly, the δ rate under the 2H model is always higher than the rate estimate under the 3H+ model, implying that the 2H rate may be "absorbing" some of the 3H variation. We will later see the same pattern emerge in large-scale sequence screens.

To bolster our intuitive understanding of model preferences, we visualized inferred substitutions at four archetypal sites in the Vertebrate Rhodopsin alignment (Yokoyama et al, 2008), for which every single rate in the 3H+ model was significantly non-zero (Figure 2.1). We used joint maximum likelihood ancestral state reconstruction (Pupko et al, 2000) under the 3H+ model to estimate the number and kind of substitutions that occurred at each site (this number is a lower bound and is subject to estimation uncertainty; here we use it for illustration purposes).

1. **Single-hit site.** Site 37 is what one might call a traditional single-hit substitution site, where the 1H model is preferred to all other models based on ER values; all apparent substitutions involve changes at a single nucleotide, hence the standard 1H is perfectly adequate. Of 330 codons, 149 preferred the 1H model compared to the 2H model.
2. **Two-hit site.** Site 144 has a dramatic preference for the 2H model over the 1H model (ER > 300); of 6 total substitutions, 4 involved a change at 2 nucleotides (and none—at 3).
3. **Serine island site.** Site 281 prefers the 3HSI model over the 2H model (ER = 39), and has a complex substitution pattern: nine 1H, four 2H, and two 3H substitutions; both 3H substitutions at this site involve synonymous changes between serine codon islands (TCN and AGY). 148 other sites had a preference (ER > 1) for 3HSI over 2H.

4. **Three-hit site.** Site 236 prefers 3H to 3HSI (ER = 5.4) as the only apparent 3H substitution at that site does not involve serine.

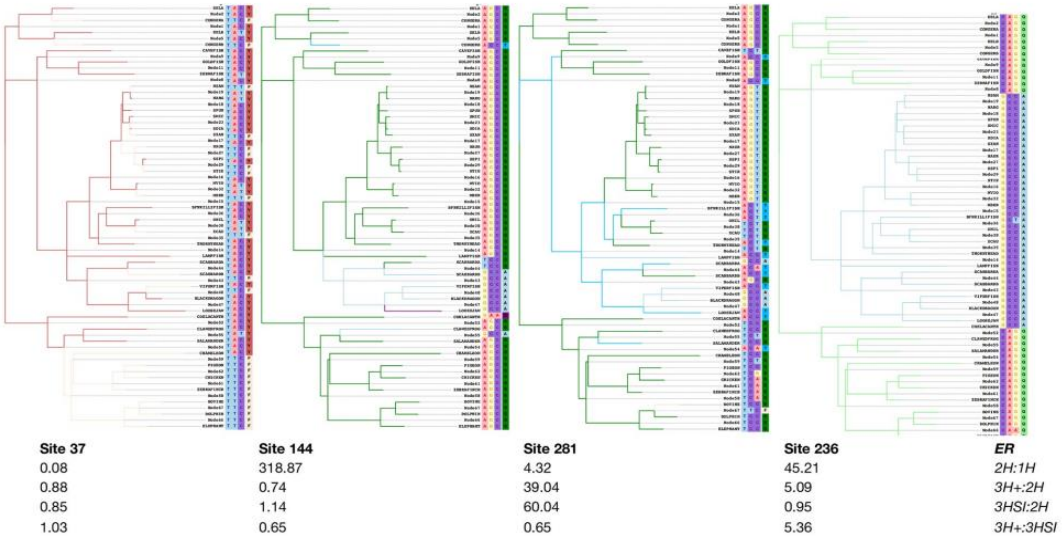



Figure 2.1. Archetypal sites based on model preferences. Four alignment sites from the Vertebrate Rhodopsin dataset (Yokoyama et al, 2008) chosen to illustrate substitution patterns which give rise to support for specific rate models. Branches are colored by the amino acid that is observed/estimated to exist at the end of the branch. Internal nodes are labeled with ancestral states inferred under the 3H+ model. Evidence ratios, which are the ratios of MLE site likelihoods under the respective models, for four pairwise model comparisons are listed below each site.

### *Large-scale Empirical Analyses*

We fitted the hierarchy of MH models to 42, 498 empirical datasets, assembled from three large-scale studies of natural selection of nuclear genes (Moretti et al, 2014;

Shultz et al, 2019; Enard et al, 2016), and a smaller collection vertebrate and invertebrate mitochondrial genes (Mannino et al, 2020), which represent a different evolutionary landscape (e.g., not affected by polymerase zeta).

Table 2.4. Evidence for multiple hit rates in empirical datasets. For each collection of alignments, the table shows the fraction with significant ($p < 0.01$, based on a 5-way conservative Bonferroni correction for FWER of 5%) LRT test results, and the average value of the likelihood ratio test statistic (for significant tests) in parentheses. Masked vs unmasked refers to the two versions of data in Moretti et al (2014): alignments in the masked version have some low-quality sites removed.

| **Dataset** | **2H:1H** | **3H+:2H** | **3H:1H** | **3H+:3HSI** | **3HSI:2H** |
|---|---|---|---|---|---|
| Invertebrate mtDNA | 92% (119.2) | 7.4% (17.12) | 92% (122.2) | 8.9% (19.97) | 2.3% (9.089) |
| Vertebrate mtDNA | 54% (33.30) | 3.0% (16.60) | 50% (36.92) | 3.2% (15.11) | 0.69% (7.986) |
| Shultz *et al* (2009) | 62% (32.39) | 20% (17.76) | 63% (39.87) | 21% (13.63) | 7.4% (12.84) |
| Moretti *et al* (2014) Unmasked | 76% (55.99) | 37% (21.82) | 77% (67.67) | 20% (13.56) | 29% (16.73) |
| Moretti *et al* (2014) Masked | 76% (67.35) | 32% (27.3) | 74% (83.23) | 17% (18.24) | 23% (20.92) |
| Enard *et al* (2006) | 28% (15.69) | 5.4% (14.18) | 28% (20.39) | 5.3% (10.49) | 3.4% (11.07) |
| Overall | 61.02% (53.99) | 23.07% (19.13) | 60.79% (61.71) | 15.66% (15.17) | 15.01% (13.11) |

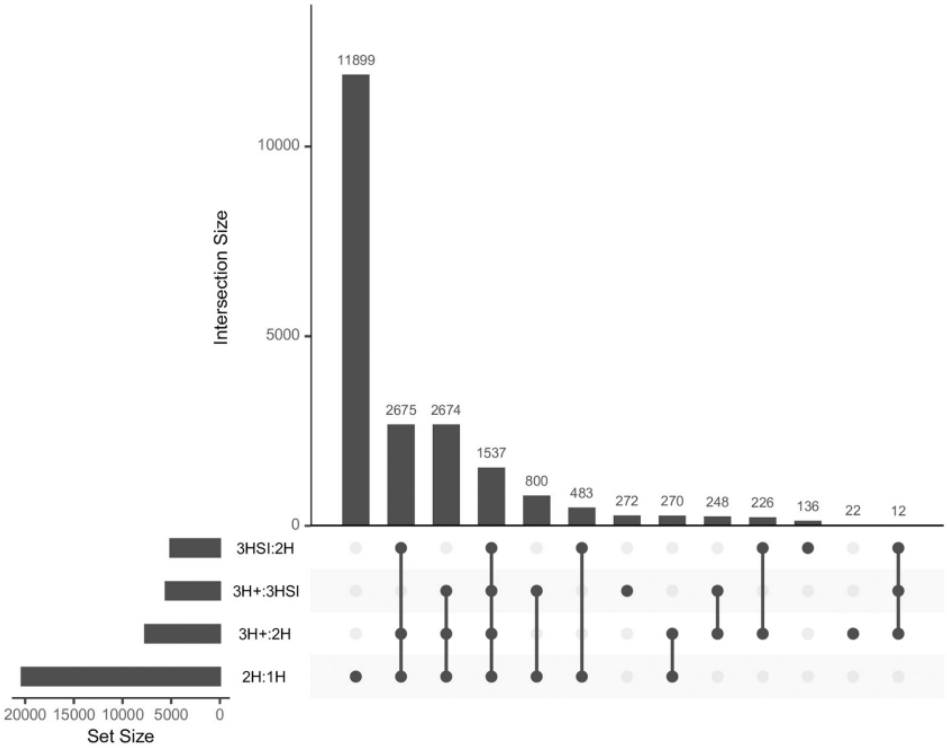



Figure 2.2. Intersections of likelihood ratio test significance. Overlaps of empirical alignments with $p \leq 0.01$ according to each of four LRTs performed for the combined empirical datasets. Groups of alignments for which a particular combination of tests was significant are shown in the main chart. The auxiliary chart in the lower left shows the number alignments belonging to a particular comparison category (row "sums"), with the significant tests indicated with filled dots. For example, there are 1537 alignments where all 4 tests are significant, and 136 alignments where the only significant test is 3SHI:2H.

***Strong Evidence for Non-zero Multiple-hit Rates.***

We found widespread statistical support for models that include non-zero rates involving multiple nucleotides. The 1H model was overwhelmingly rejected in favor of the 2H model (Table 2.4), and the improvement in fit was quite dramatic on average, for all but the Enard et al (Enard et al, 2016) collection. A substantial fraction of alignments preferred models that allowed non-zero three-hit rates over the two-hit model, and the 3H+ model which does not limit 3H instantaneous changes to only synonymous codons. Based on the results of the four likelihood ratio tests, each dataset could be assigned to a unique rate preference category Fig 2.2. For example, 11, 899 alignments preferred the 2H to 1H model, but none of the other comparisons were significant, i.e., there was no evidence for non-zero 3H instantaneous rates. 2, 675 alignments preferred 2H to 1H, and 3H+ to 2H, i.e., provided evidence for non-zero 3H instantaneous rates. 483 alignments preferred 2H to 1H and 3HSI to 2H, but not 3H+ to 3HSI, implying that all 3H changes were constrained to synonymous codon islands.

***Factors Associated with MH Detection.***

The rates at which 2H, 3H and 3HSI rates were detected with $p < 0.01$ as functions of simple statistics of the alignments, are shown in Fig 2.3. Larger (more sequences) and longer (more codons) alignments generally elicited higher detection rates for all types of multiple hits. Increasing overall divergence levels between sequences, measured by the total tree length, also corresponded to increasing detection rates, up to a saturation point.

The mean strength of selection, measured by the gene-average ω had little effect on detection rates, except for a dip in the tail. In a simple logistic regression using 2H:1H $p < 0.01$ as the outcome variable, sequence length, and number of sequences were positively associated with the detection rate ($p < 0.0001$), while tree length was confounded with the number of sequences and was not independently predictive, and ω was not significantly predictive.

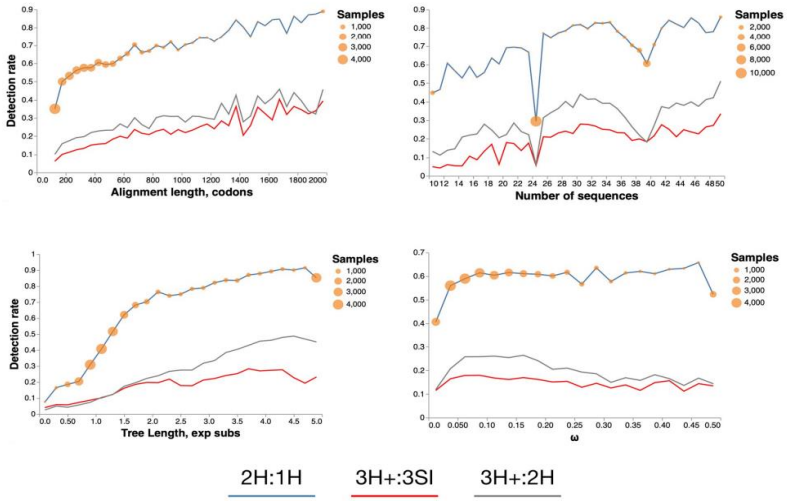



Figure 2.3. Multiple hit detection rate. The fraction of alignments where the corresponding test was significant at $p \leq 0.01$ as a function of one of four alignment properties. Orange circles depict the binning steps and the number of alignments falling into each bin. For tree lengths and ω values we used estimates under the 1H model.

***Strong MH Signals Come from a Small Fraction of Sites.***

For alignments where there was significant evidence for non-zero 2H and/or 3H rates ($p < 0.01$), a small fraction of sites strongly (ER > 5) supported the corresponding

MH model. For the 2H:1H comparison, a median of 0.67% (interquartile range, IQR [0.21% − 1.7%]), and for the 3H:2H comparison, a median of 0.52% (IQR [0.26% − 0.94%]) (Appendix A A.S2 Fig) sites in an alignment had high ER in support of the respective model.

***Patterns of Substitution Associated with MH Rates.***

Substitutions between serine islands (AGY and TCN) appear to be the most frequent inferred 3H change in biological alignments (see Fig 2.4). Six of the most common substitutions at sites with high ER in support of the 3H+ model involve island jumping. However, other amino-acid pairs are also involved in hundreds of apparent substitutions, e.g., ATG(M)↔GCA(A). Of the 7664 datasets that reject the 2H model in favor of the general 3H+ model, 2901(37.9%) fail to reject 3HSI in favor of 3H+, implying that they only require non-zero rates for synonymous island jumps. However, many of the same changes frequently appear at sites that do not strongly prefer 3H+ to 2H model, but strongly prefer 2H to 1H model (i.e., 2H sites). A key determinant of whether or not an AGY: TCN or other 3H change benefits from non-zero ψ rates is the length of the branch where the change is inferred to occur. Branches with 3H changes that supported 3H+ model were significantly shorter than those where 2H model was sufficient: median 0.09 substitutions/site, vs median 0.26 substitutions/site. Consequently, the need to explain 3H changes happening over short branches (shorter evolutionary time, slower overall rates) provides evidence in support of 3H+ models.`

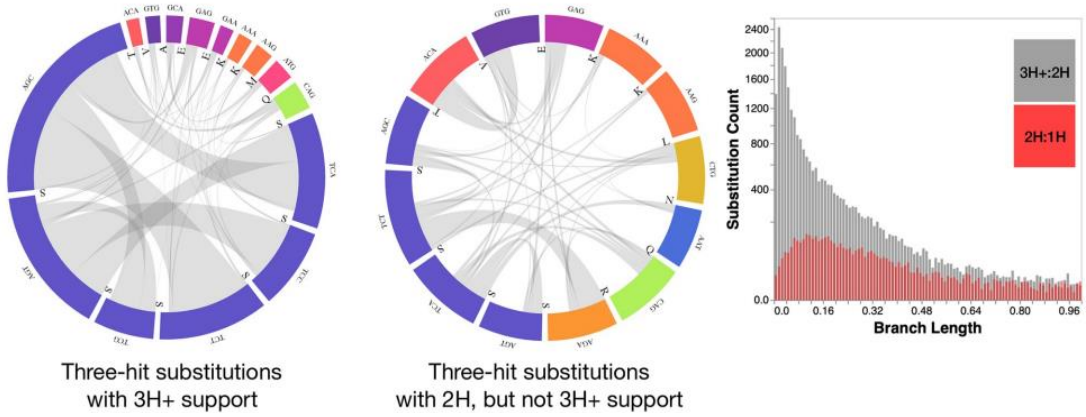



Figure 2.4. Three-hit substitutions commonly occurring in empirical data. A subset of common three-hit substitutions across all empirical datasets. Three-hit substitutions with 3H+ support are defined as those occurring at sites with ER(3H+:2H)>5. Three-hit substitutions with 2H but not 3H + support are defined as those occurring at sites with ER(3H+:2H)<1 and ER(2H: 1H)>5. Branch lengths along which the two types of substitutions are inferred to occur are shown in the histogram.

Among 3H non-synonymous substitutions (see Fig 2.5) codons encoding for serine are still prominently represented, but not as dominant, with numerous substitutions involving methionine and other amino-acids. Serine codons are similarly frequently involved in 2H substitutions, along both short and long branches (e.g., between codons such as AGC ↔ TCC and AGT ↔ TCT), but other pairs are exchanged at least 90 times, including ACA(T)↔ATG(M) and CAG(Q)↔TGG(W) (short branches) and ATT(I)↔TTA(L) (long branches).

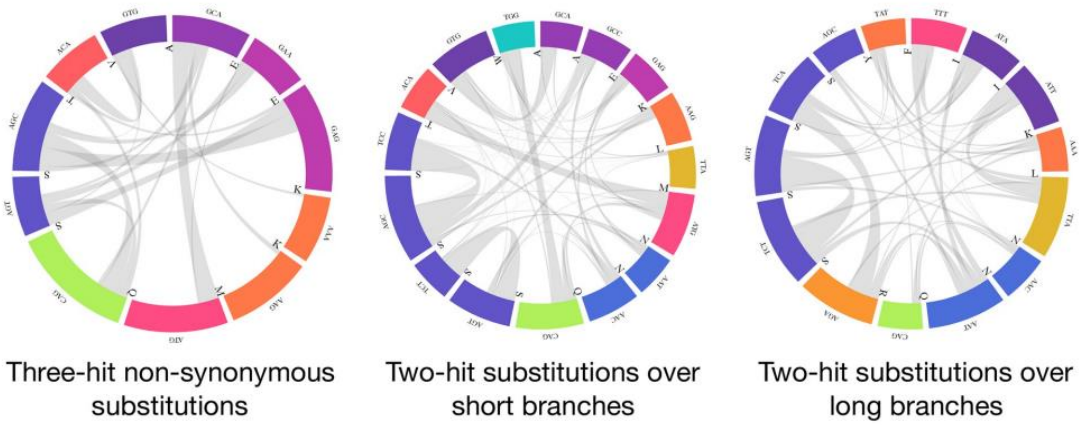



Figure 2.5. Three-hit non-synonymous substitutions and two-hit substitutions occurring in empirical data. A subset of common substitutions across all empirical datasets. Three-hit substitutions with non-synonymous support are defined as those occurring at sites with ER(3H+:3HSI) > 5. Two-hit substitutions over short branches are defined as those occurring at sites with ER(3H+:2H) < 1 and ER(2H: 1H) > 10 and branch length is ≤ 0.05 subs/site. Two-hit substitutions over short branches are defined as those occurring at sites with ER(3H+:2H) < 1 and ER(2H: 1H) > 10 and branch length is ≥ 0.25 subs/site

***Interaction Between Rate Estimates.***

As with the benchmark datasets, the inclusion of multiple hit rates in models has an effect on other substitution rates. The gene wide point-estimate of ω is systematically lowered by the inclusion of non-zero δ rates, even though there are rare instances when the ω estimates are increased (Appendix A, A.S1 Fig). A Theil Sen robust linear regression estimate yields ω(2H) ~ 0.965 × ω(1H), but for 1150(5.7%) of the datasets with where

2H:1H comparison was significant, the $\omega(2H) < 0.75 \times \omega(1H)$. Consequently, the estimation bias in important evolutionary rates due to model misspecification for some of the datasets could be significant. The inclusion of 3H components in the model lowers the 2H rate as well $\delta(3H+) \sim 0.77 \times \delta(2H)$.

***Impact on Branch Length Estimates.***

Branch lengths estimated with 2H and 3H models on the Selectome dataset were, on average, 0.93× the length of the standard (1H) estimate, while they were effectively identical between 2H and 3H estimates (Appendix A, A.S5 Fig). On data simulated without MH, all three models (1H, 2H, 3H) yielded branch length estimates that were nearly identical, and a slight underestimate of the true values due to a bias in estimating the length of very short branches. However, on data simulated with MH, 1H models overestimate branch lengths compared to the 2H/3H models. A plausible explanation is that 1H models "expand" branch lengths slightly to compensate for the multi-hit events, while 2H and 3H models do not introduce bias in branch length estimates when there is no underlying multi-hit component because they can account for this situation by setting some parameters to 0.

***Simulations***

False positive rates.

We evaluated operating characteristics of the likelihood ratio tests (LRT) for MH model testing on parametrically simulated data. In the simplest case of a single-branch

(two-sequence) null data generated under the 1H model, Type I error rates for 2H:1H and 3H:2H tests were on average below nominal. However, for very divergent sequences (e.g., >3 expected substitution per site), the test became somewhat anti-conservative, which is not surprising for such severely saturated data (Appendix A Figure S3). Individual branches that are this long are highly abnormal in real biological datasets. Expanded to multiple sequence alignments generated using parameter estimates from four biological datasets, simulations confirmed that all the tests employed appear to be somewhat conservative; this is by design because asymptotic distributions of LRT statistics when null hypotheses are on the boundaries of the parameter space are less conservative than the 1- or 2-degrees of freedom $\chi^2$ distributions we use here (Self et al, 1987).

Power

The tests are generally well-powered, especially if the effect sizes (magnitudes of MH rates) are sufficiently large (Table 2.5). The power to detect two-hit substitution (2H:1H) is especially high (>90%) across all simulations. The test which attempts to identify non-zero triple-hit synonymous island rates (3HSI:2H) is the least powerful, because its signal is derived from a tiny fraction of all substitutions (substitutions between synonymous islands), i.e., the effective sample size is smaller than for the other tests.

False positives due to alignment errors.

Whelan and Goldman (Whelan and Goldman, 2004) suggested that non-zero estimates of triple-hit rates could be at least partially attributed to alignment errors. It is

impossible, with a few rare exceptions, to declare that any particular alignment of biological sequences is correct. Hence, in order to estimate what, if any, effect potential multiple sequence alignment errors might have on our inference, we simulated null data (1H model) with varying indel rates with Indelible (Fletcher and Yang, 2009), inferred multiple sequence alignments MAFFT (Katoh et al, 2002) in a codon-aware fashion, inferred trees using neighbor-joining, and performed our hierarchical model fit. This procedure induces multiple levels of model misspecification, and errors: Indelible uses a different model (GY94 M3) to simulate sequences, there is alignment error, and there is phylogeny inference error. Sufficiently high indel rates coupled with other inference errors can indeed bias our tests to become anti-conservative, although these levels are higher than what we see (based on per-sequence "gap"/character) ratios for our biological alignments (Appendix A, A.S4 Fig). Empirical alignments have gap content that is consistent with alignments simulated with $0.01 - 0.015$ indel rates, for which test performance is nominal. The Selectome dataset (Moretti et al, 2014) can also be retrieved masked for alignments with unreliability at certain sites. We compared the masked and unmasked alignments and found similar results: 76% of alignments in both cases favor 2H over 1H; 77% of unmasked alignments and 74% of masked alignments choose 3H over 1H (Table 2.4). While there is certainly some false signal due to misalignment, it is unlikely to be the dominant factor here. Nonetheless, care must be taken not to over-interpret MH findings when the alignments are uncertain.

Table 2.5. Power to detect MH rates. The fractions of simulated datasets that had $p < 0.05$ for the corresponding test. N = number of simulations in each category, and the explicit definition effect size is shown in parentheses.

| Test | All | | Large effect | |
|---|---|---|---|---|
| | N | Power | N | Power |
| 2H:1H | 1956 | 94% | 967 | 99%($\delta > 0.5$) |
| 3H+:2H | 1940 | 64% | 1056 | 83%($\psi > 0.5$) |
| 3HSI:2H | 447 | 33% | 114 | 51%($\psi_s > 5.0$) |
| 3H+:3HSI | 1940 | 66% | 1056 | 86%($\psi > 0.5$) |

## Discussion

Nearly three in five empirical alignments considered here provide strong statistical support that at least some of the substitutions are not well modeled by standard codon models. More than one in five prefer to have direct three-hit substitutions accounted for explicitly in the models. Substitutions involving serine codons, which are unique among the amino acids in that they comprise two islands which are two or three nucleotide changes from each other, are prominent in driving statistical signal for these preferences, especially if they occur along short branches. Many other amino-acid pairs are also involved in such exchanges, indicating that not all the statistical signal is due to serine codons, although in a typical alignment only a small fraction of sites (about 1%) prefer multiple hit models strongly. Many previous studies have provided evidence that evolutionary models with multiple hits provide better fit to the data, but the scale of this phenomenon in the comparative evolutionary context has not been fully appreciated, although the interest in model development in this area is being rekindled. Our results also show that the inclusion

of multiple-hit model parameters changes ω estimates, and with them—potentially alter inferences of positive selection, which was demonstrated for branch-site tests, (Venkat et al, 2018), and for data simulated with multiple hits but analyzed with standard models (Dunn et al, 2019). Additionally, traditional models may slightly overestimate branch lengths for data where multiple-hit models provide better fits.

How much of this apparent support for multiple-hits comes from biological reality, and how much from statistical artifacts, or other unmodeled evolutionary processes—the so-called phenomenological load (Jones et al, 2018)? Our simulation studies provide compelling evidence that the tests we use here are statistically well-behaved and possess good power, i.e., our positive findings are unlikely to be primarily a result of statistical misclassification. Other confounders, especially alignment error, have the potential to mislead the tests, but only at levels that appear higher than what is likely present in most biological alignments. In addition, there are some datasets (e.g., HIV reverse transcriptase), where alignment is not in question due to low biological insertion/deletion rates or structural information, and these data still support non-zero multiple-hit rates as well.

There is an abundance of data and examples of doublet substitutions in literature, and mechanistic explanations for them, e.g., due to polymerase zeta (Harris and Nielsen, 2014) exist. There are several papers arguing that the number of apparent triple-hits occurring in sequences is greater than what we would expect solely from random mutation (Bazykin et al, 2004; Schrider et al, 2011; Smith and Hurst, 1999), however the mechanism (if it exists) by which they might occur is speculative. As one option, Sakofsky et al.

(Sakofsky et al, 2014) have suggested that DNA repair mechanisms could help explain multi-nucleotide mutations.

Our analyses indicate that much, but not all, of the support for non-zero triple-hit rates derived from serine codon island jumping, particularly in cases when these jumps must occur over a short branch in the tree. Comparative species data might lack the requisite resolution to discriminate between instant multiple base changes and a rapid succession of single nucleotide changes spurred on by selection; the literature is split on which mechanism is primal (Averof et al, 2000; Rogozin et al, 2016). Such a common phenomenon is worth further investigation, in our opinion.

Our evolutionary models are broadly comparable to several others that have been published in this domain, some of which have more parametric complexity (Dunn et al, 2019) or consider effects of substitutions spanning codon boundaries (Whelan and Goldman, 2004). Our novel contributions are direct tests for the importance of synonymous island jumping, and a simple evidence ratio approach to identify and categorize specific sites that benefit from non-zero multiple hit rates. These models are easy to fit computationally, with roughly the same cost as would be required for an ω−based positive selection analysis, and we provide an accessible implementation for researchers to use them. Further modeling extensions, e.g., the inclusion of synonymous rate variation, branch site effects, etc., can be easily incorporated.

# CHAPTER 3.

# HUMAN HSPB1, HSPB3, HSPB5 AND HSPB8: SHAPING THESE DISEASE FACTORS DURING VERTEBRATE EVOLUTION.


## Abstract

Small heat shock proteins (sHSPs) emerged early in evolution and occur in all domains of life and nearly in all species, including humans. Mutations in four sHSPs (HspB1, HspB3, HspB5, HspB8) are associated with neuromuscular disorders. The aim of this study is to investigate the evolutionary forces shaping these sHSPs during vertebrate evolution. We performed comparative evolutionary analyses on a set of orthologous sHSP sequences, based on the ratio of non-synonymous: synonymous substitution rates for each codon. We found that these sHSPs had been historically exposed to different degrees of purifying selection, decreasing in this order: HspB8 > HspB1, HspB5 > HspB3. Within each sHSP, regions with different degrees of purifying selection can be discerned, resulting in characteristic selective pressure profiles. The conserved α-crystallin domains were exposed to the most stringent purifying selection compared to the flanking regions, supporting a 'dimorphic pattern' of evolution. Thus, during vertebrate evolution the different sequence partitions were exposed to different and measurable degrees of selective pressures. Among the disease-associated mutations, most are missense mutations primarily in HspB1 and to a lesser extent in the other sHSPs. Our data provide an explanation for

this disparate incidence. Contrary to the expectation, most missense mutations cause dominant disease phenotypes.

Theoretical considerations support a connection between the historic exposure of these sHSP genes to a high degree of purifying selection and the unusual prevalence of genetic dominance of the associated disease phenotypes. Our study puts the genetics of inheritable sHSP-borne diseases into the context of vertebrate evolution.

## Introduction

The human genome encodes ten small heat shock proteins (sHSP) (Fontaine et al. 2003; Kappé et al. 2003), now systematically named HspB1 through HspB10, whether or not their expression is induced by stress factors (Kampinga et al. 2009). The defining feature of sHSPs is a conserved ~ 85 amino acid long sequence stretch called the α-crystallin domain (αCD) (de Jong et al. 1998; Kappé et al. 2010; Mymrikov et al. 2011). It is believed that the primary function of sHSPs is to act as holdase chaperones with a depot function for unfolded proteins, thereby shaping protein–protein interactions. This, in turn, controls many cellular processes and structures such as cell signaling, cell architecture, differentiation, redox homeostasis, apoptosis, and others (Carra et al. 2019).

sHSPs occur in all domains of life, including Archea, Bacteria, and Eucarya, suggesting that they emerged early in evolution. Nearly all organisms contain genes encoding sHSPs except a few pathogens that have lost these genes (Kriehuber et al. 2010).

The number of paralogous sHSP genes per genome can vary between 2 (average) in Archea and Bacteria, 3 (average) in fungi, 8 (average) in Metazoa, and up to 50 (maximum) in some plants (Kriehuber et al. 2010). Many vertebrates, including humans, contain ten paralogous sHSP genes, although some variation may occur, notably in the fish taxa. Among the vertebrate sHSPs, orthology to the human sHSPs can be easily established by phylogenetic analysis in most cases, whereas this is usually not possible across the other taxa.

Multiple sequence alignment (Fontaine et al. 2003; Franck et al. 2004; Kriehuber et al. 2010) reveals that the conserved αCD is flanked by regions of elevated sequence variability: the moderately conserved N-terminal region (NTR) and the highly variable C-terminal extension (CTE). Sometimes a short central region (CeR) with high sequence variability, positioned between the NTR and the αCD (possibly serving as a 'hinge' between both regions), is considered separately (Wang et al. 2000; Fontaine et al. 2003). This organization of the primary structure is summarized in Fig. S1. The αCD contains several β-strands forming two anti-parallel β-sheets (sandwich structure), whereas the flanking regions (NTR, CeR, CTE) are less ordered or even disordered (Carra et al. 2019; Webster et al. 2019; Boelens 2020). Whether or not such regions without a compact folding structure should be considered domains is debatable. Together, these domains and regions contribute to both the dynamic association of the monomers into high molecular mass complexes and to their holdase chaperone function.

The evolutionary history of sHSPs through all domains of life is characterized by certain unique features (Kriehuber et al. 2010):

(1) According to the PFAM classification (PF00011 at http://pfam.xfam.org/family/PF00011#tabview=tab0) the overwhelming majority of sHSPs from all domains of life contain just one structural domain, the highly conserved αCD, contrary to other studied protein superfamilies which typically diverged by recruiting additional domains;

(2) The evolution of sHSPs follows a 'dimorphic pattern': the αCD has a monophyletic origin reflecting the evolution of species, as opposed to the flanking regions which were historically remodeled several times in parallel but independent of the αCD. Thus, the αCD and the flanking regions have a fundamentally different evolutionary history. It is believed that this combination of the conserved αCD with the variable flanking regions (NTR, CeR, CTE) provides a unique and high degree of structural and functional speciation.

Four of these sHSP genes carry known mutations which have been associated with neuromuscular disease phenotypes in humans: HspB1 (Hsp27, Hsp25), HspB3, HspB5 (αB-crystallin), and HspB8 (Hsp22) (Datskevich et al. 2012; Benndorf et al. 2014; Vendredy et al. 2020), not counting polymorphisms or other sequence variants with unclear relation to disease. Mutant alleles of HspB1, HspB3 and HspB8 primarily associate with neuropathies with a spectrum of symptoms ranging from the clinical characteristics 'Distal

Hereditary Motor Neuropathy' and 'Charcot-Marie-Tooth Disease' to some forms of 'Amyotrophic Lateral Sclerosis,' although a small number of mutations in these sHSPs associate also with some forms of myopathies. Thus, myopathies are also a relevant and primary disease related to mutations in these sHSPs (Appendix B Table B.S1; Benndorf et al. 2014; Echaniz-Laguna et al. 2017; Adriaenssens et al. 2017; Katz et al. 2020; Chen et al. 2021).

However, the pathogenicity of the HspB3 mutations has been recently disputed (Adriaenssens et al. 2017; Vendredy et al. 2020). Quite differently, mutations in HspB5 cause exclusively various forms of myopathies, including cardiomyopathies, and cataracts in the lens of the eye, the latter partially in association with the myopathies. The overwhelming majority among these disease-associated mutations are missense mutations, as opposed to a small number of identified frame shift, nonsense and elongation mutations (Benndorf et al. 2014; Vendredy et al. 2020). Disorders caused by the mutant sHSP alleles with their high penetrance belong to the group of Mendelian diseases (Quintana-Murci and Barreiro 2010), in spite of the fact that in most patients the symptoms develop only later in life (late onset diseases). Most of these missense mutations are associated with a dominant disease phenotype, or this association can be assumed, e.g., when the mutations occur sporadically (Appendix B Table B.S1). Four of the known missense mutations in HspB1 and HspB5 are associated with a recessive disease phenotype.

One approach for understanding the evolutionary forces that have shaped proteins is to measure the ratio (ω) of the non-synonymous (β) and synonymous (α) substitution

rates ($\omega = \beta/\alpha$) at each codon position in a gene of interest (Kosakovsky Pond and Frost 2005). Non-synonymous substitutions change the amino acid being coded for at that site, which can directly affect protein structure and influence the fitness of an organism through negative (purifying) selective pressure. In contrast, synonymous substitutions do not change the amino acid being coded for, leaving the amino acid sequence unchanged. Synonymous substitutions are typically viewed as neutral and provide a baseline rate against which non-synonymous evolutionary rates can be calibrated. Therefore, the ratio $\omega$ of relative rates of non-synonymous and synonymous substitutions can provide information as to the type of selection that has acted upon a given set of protein-coding sequences, upon selected sequence segments (e.g., structural domains), or upon single codons (e.g., mutation sites). This ratio $\omega$ has become a standard measure of selective pressure in evolutionary biology (Frost et al. 2005; Arenas 2015). When there are more non-synonymous changes relative to synonymous changes, the $\omega$ ratio will be greater than 1, indicative of positive, diversifying selection. Conversely, if there are fewer non-synonymous changes relative to synonymous changes, the $\omega$ ratio will be less than 1, indicative of negative, purifying selection.

In this study, we have estimated the ratio $\omega$ for each codon along the entire length of the sequences of human HspB1, HspB3, HspB5, and HspB8, using the aligned vertebrate sequences of the orthologs. We provide evidence for the exposure of the NTR, CeR, αCD, and CTE in these sequences to different degrees of purifying selection during vertebrate evolution. This disparate pattern of selective pressure within vertebrates fits well to the

wider pattern of the dimorphic evolution that has shaped sHSPs across all domains of life, i.e., from the origin of the tree of life onward (Kriehuber et al. 2010). Thus, the same forces that shaped the sHSPs early in evolution are detectable also in the vertebrate evolution in recent 400–500 million years. Although similar, the ω profiles of the four studied human sHSPs exhibit differences indicating exposure of each sHSP to variable degrees of evolutionary pressure. For the disease-associated missense mutation sites, we find that all sites were historically exposed to detectable purifying selection, although to different degrees. Additionally, our data provide an explanation for the different incidence by which the affected sHSPs harbored disease-associated mutations.

## Methods

### *Retrieval of Vertebrate sHSP Sequences from the Databases*

This study was restricted to the vertebrate sHSP sequences, specifically to the taxon Gnathostomata. Gnathostomata emerged more than 400 million years ago in evolution (Kuraku et al. 2016) and comprises all vertebrate species except those from the taxon Agnatha. Gnathostomata orthologs of human HspB1, HspB3, HspB5, and HspB8 were retrieved by a BLAST-P search at the NCBI website (https://www.ncbi.nlm.nih.gov/) using the full-length human protein sequences as the query sequence. Additionally, several sequences were retrieved from the ENSEMBL website (https://useast.ensembl.org/index.html?redirect=no) adding in total to sequences from at least 130 species for each studied sHSP, representing the major taxons of Gnathostomata

(see below). Thereafter, the matching cDNA sequences were retrieved. Only full-length sequences of high quality were retained for analysis. When the translatable sHSP nucleotide sequences were embedded in larger DNA sequence blocks, the sHSP open reading frames were identified in silico by using the translation tool (https://web.expasy.org/translate/).

### *Inclusion Criteria*

Each retrieved candidate protein sequence was aligned to the human ortholog using the EMBOSS-Needle algorithm (https://www.ebi.ac.uk/Tools/psa/emboss_needle/) returning the degree of identity and similarity of each sequence to be tested with the human sequence. Sequences exhibiting $\geq 50\%$ identity and $\geq 60\%$ similarity were included in further analysis, with all other sequences being excluded. When aligned by the EMBOSS-Needle algorithm, a few candidate sequences exhibit a putative N-terminal extension, compared to the human orthologs and to the orthologs of nearly all other species. However, these putative N-terminal extensions were deduced in silico by algorithms based on DNA-sequencing data without any experimental support. Therefore, these putative N-terminal extensions can be assumed to result from an erroneous in silico-identification of the initiation codons and were consequently excluded from further processing. In these sequences, we used the canonical initiation codons based on the alignment with the respective human sHSPs.

Orthology of each candidate sequence with the human sHSPs was tested by constructing gene trees using alignment and phylogenetic inference (Kristensen et al. 2011) at the Clustal Omega web site (https://www.ebi.ac.uk/Tools/msa/clustalo/). Phylogenetic trees were constructed using all ten human sHSPs and the query sequence. Orthology was determined through analysis of patterns of sequence clustering within the reconstructed phylogenetic tree. Sequences were labeled as orthologs of human HspB1, HspB3, HspB5 or HspB8 if they grouped with these sHSPs, whereas sequences that did not group with these sHSPs were discarded.

Applying the criteria above resulted in the exclusion of several candidate sHSP-like sequences, notably from the fish taxa. The numbers of sHSP sequences that finally entered analyses for the detection of natural selection within the taxon Gnathostomata were 152 for HspB1, 130 for HspB3, 147 for HspB5, and 144 for HspB8 (Appendix B Table B.S2).

***Species Included in the sHSP Sequence Analysis***

The designation of species and their taxonomic ranking was according to the NCBI web site (https://www.ncbi.nlm.nih.gov/taxonomy/). This study includes sHSP sequences from Gnathostomata (Vertebrata: Gnathostomata) species representing the major vertebrate taxa: Mammalia (Gnathostomata: Teleostomi: Euteleostomi: Sarcopterygii: Dipnotetrapodomorpha: Tetrapoda: Amniota: Mammalia; includes primates, rodents, lagomorphs, ungulates, whales, bats, shrews, carnivorians, mammals of African origin,

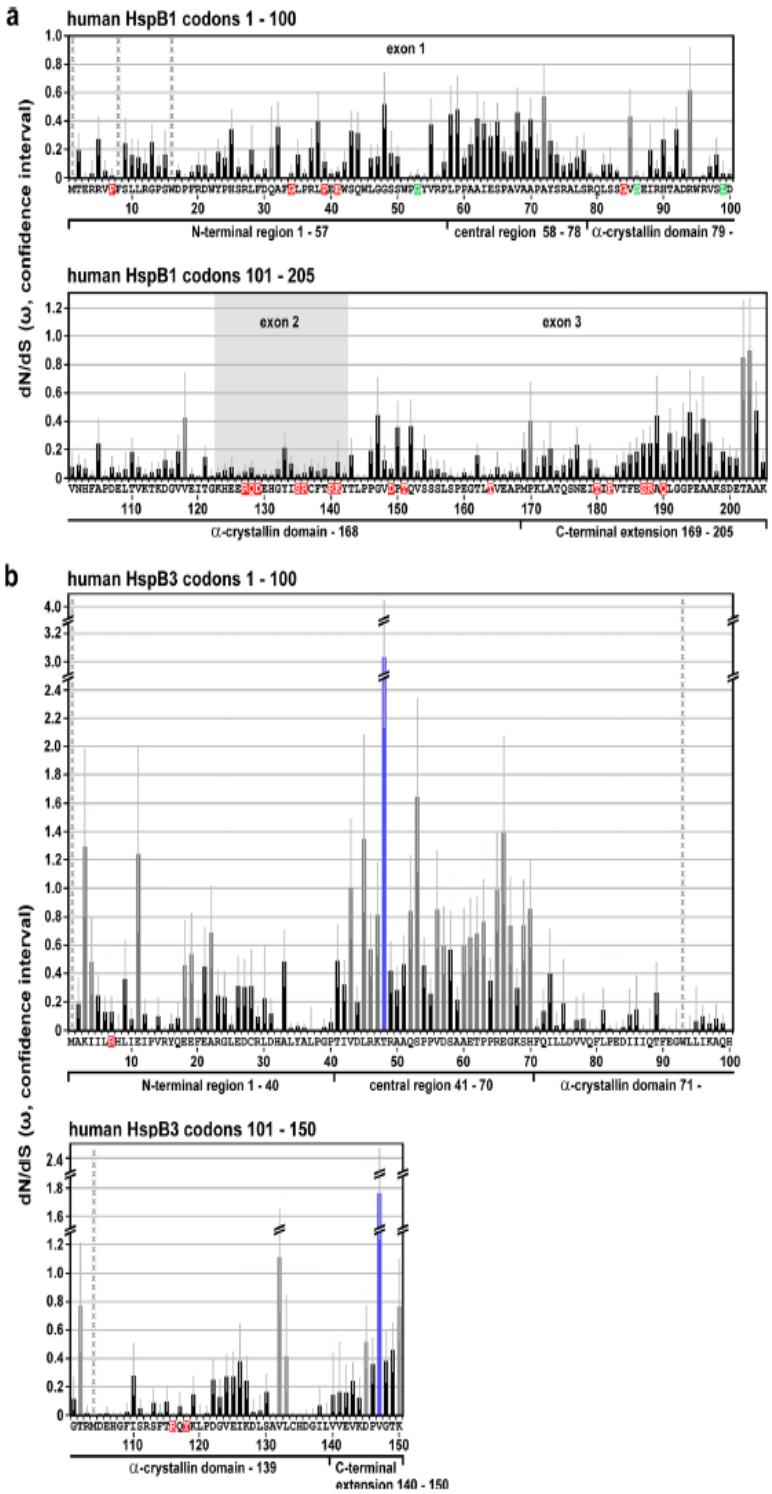
marsupials, armadillos, others), Sauria (Amniota: Sauropsida: Sauria; includes crocodiles, lizards, snakes, birds, turtles), Amphibia (Tetrapoda: Amphibia; includes frogs and toads), Coelacanthimorpha (Sarcopterygii: Coelacanthimorpha; includes just one species, the coelacanth), Neopterygii (Euteleostomi: Actinopterygii: Neopterygii, includes various taxa of ray-finned fish), and Chondrichthyes (Gnathostomata: Chondrichthyes or cartilaginous fish). The names of all species included in this study and their taxonomic ranking are given in Appendix B Table B.S2.

***Determination of Codon-specific Evolutionary Pressure Along the Entire Length of the Human sHSPs***

The strength and direction of pervasive natural selection (i.e., purifying versus positive or diversifying) was assessed at the level of individual codon sites in each of the four sHSPs using the Fixed Effects Likelihood (FEL) method implemented in the HyPhy package (Pond et al. 2005) and the Datamonkey (https://www.datamonkey.org) web application (Pond and Frost 2005). A dN/dS point estimate with confidence intervals was computed for each site using the FitMG94 model (Kosakovsky Pond et al. 2010). The confidence interval around the maximum likelihood point estimate of ω was derived by computing occurrences where the parameter estimate was not rejected in favor of the maximum likelihood estimate at a defined significance level (95% confidence of profile likelihood). A phylogenetic tree was constructed, using FastTree2.1 (Price et al. 2010) with the Generalized Time Reversible (GTR) model of nucleotide evolution. We estimated the ratio (ω) of non-synonymous (β) to synonymous (α) substitution rates for each codon along

the entire multiple sequence alignment by maximum likelihood. These ω-values are a commonly used measure of selection, expected to be i) significantly (using the likelihood ratio test, LRT) less than one when purifying selection is predominant, ii) not significantly different from one for neutral evolution, and iii) significantly greater than one if positive selection is predominant. We report ω-value estimates for all codons of the human sHSP sequences and omit all codons in vertebrate sequences that have no homologous sites in the human sequences (Appendix B, Table B.S3).

The estimated numeric ω-values also indicate the strength of the selective pressure for a given codon, e.g., with values of 0 or near 0 suggesting a greater degree of purifying selection. p values indicating statistical significance were derived using the asymptotic Chi-square test distribution for the LRT. Statistical significance (LRT p-value $\leq 0.1$) for detecting purifying selection was observed for all codons with $\omega = 0$ (no columns; cf. Figure 1) and typically for most codons with $0 < \omega < 0.5$ (indicated by black columns). Statistical significance (LRT p-value $\leq 0.1$) for detecting diversifying selection was observed for two codons in HspB3 with $\omega > 1.5$ (indicated by blue columns). For codons with $\omega \approx 1$, no selective force was detected (neutral evolution; indicated by gray columns). Appendix B Table B.S3 gives the numeric values of β, α, ω (dN/dS point estimates), and the dN/dS confidence intervals with the upper and lower bounds.


a
human HspB1 codons 1 - 100
exon 1
N-terminal region 1 - 57
central region 58 - 78
α-crystallin domain 79 -
human HspB1 codons 101 - 205
exon 2
exon 3
α-crystallin domain - 168
C-terminal extension 169 - 205
dN/dS (ω, confidence interval)
b
human HspB3 codons 1 - 100
N-terminal region 1 - 40
central region 41 - 70
α-crystallin domain 71 -
human HspB3 codons 101 - 150
α-crystallin domain - 139
C-terminal extension 140 - 150

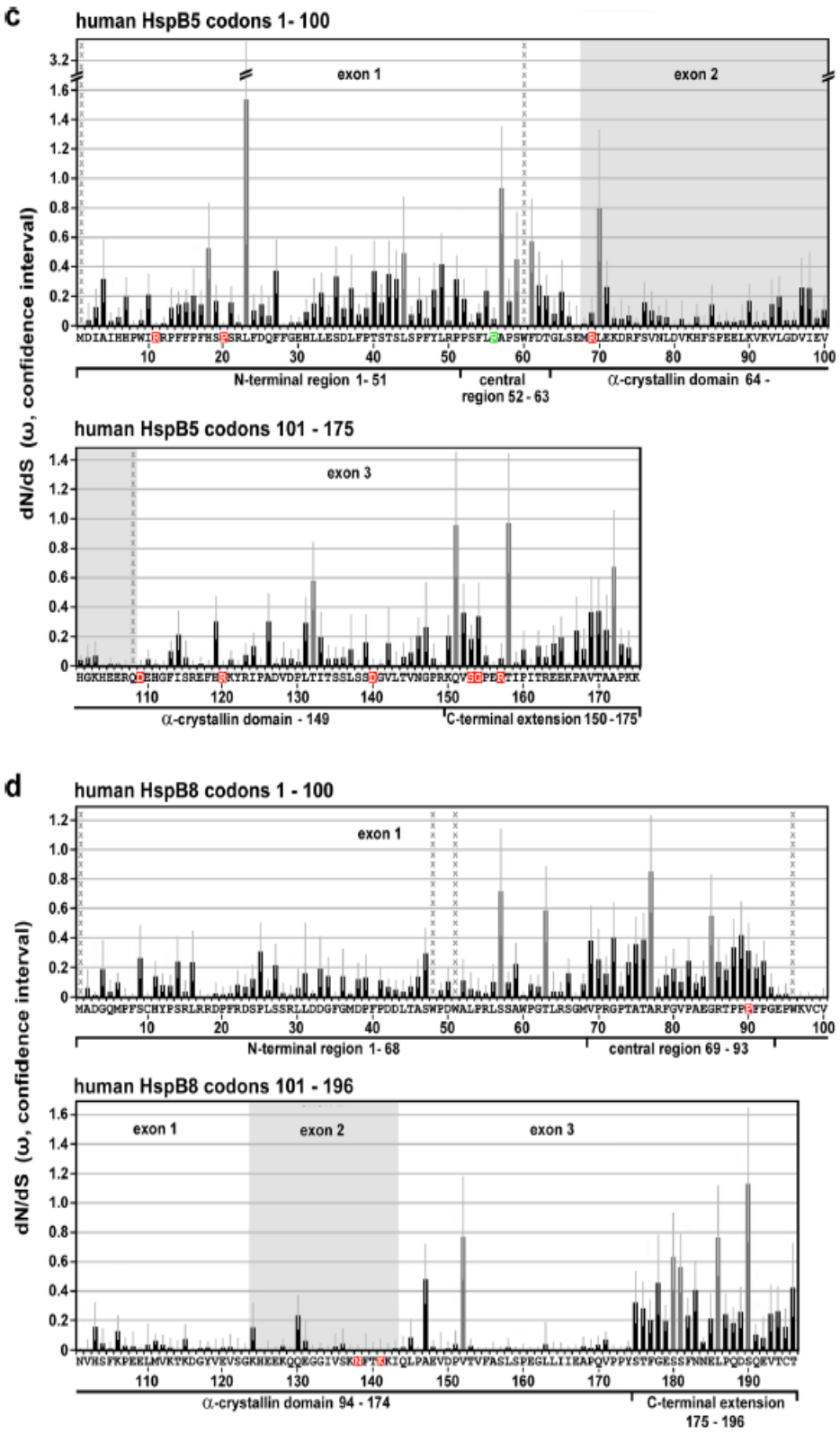

c
human HspB5 codons 1- 100
exon 1
exon 2
N-terminal region 1- 51
central region 52 - 63
α-crystallin domain 64 -
human HspB5 codons 101 - 175
exon 3
α-crystallin domain - 149
C-terminal extension 150 - 175
dN/dS (ω, confidence interval)
d
human HspB8 codons 1 - 100
exon 1
N-terminal region 1- 68
central region 69 - 93
human HspB8 codons 101 - 196
exon 1
exon 2
exon 3
α-crystallin domain 94 - 174
C-terminal extension 175 - 196
dN/dS (ω, confidence interval)

Figure 3.1. Selective pressures along the sequences of human HspB1 (a), HspB3 (b), HspB5 (c) and HspB8 (d) as returned by the FEL algorithm. The dN/dS point estimates were determined for each codon of the aligned Gnathostomata sHSP orthologs, omitting all codons without homologous amino acid residues in the human sequences. The plots show the principal ω-values with the respective confidence intervals (lower and upper bound; light gray error bars) along the primary sequences. No columns or black columns: $\omega < 1$, indicating that these positions were historically exposed to purifying selection, using $p \leq 0.1$ as statistical significance threshold; dark gray columns: $\omega \approx 1$ or near 1, with $p > 0.1$, indicating neutral evolution; blue columns: $\omega > 1$, with $p < 0.1$, indicating positive selection. Positions with undefined ω-values are indicated by columns of gray x. Amino acid residues affected by disease associated missense mutations are highlighted in color (red, green: dominant and recessive disease phenotypes, respectively). Sequence segments (partitions) with relatively low, intermediate or high ω-values in average, as they can be discerned, are indicated and were used to delineate the sequence partitions as used in this study (cf. Appendix B Table B.S4). These partitions correspond approximately, though not perfectly, to the NTR, CeR, αCD, and CTE. The exons are indicated where applicable because methionine is encoded solely by ATG, the ω-value for the strictly aligned (invariable) initiator methionine's is not defined ($\beta/\alpha$; $\beta = 0$, $\alpha = 0$). Depending on the alignments, the ω-values for several internal methionine and tryptophan (encoded solely by TGG) residues, or occasionally for other amino acid residues within sequences may also be undefined ($\beta/\alpha$; $\beta = 0$, $\alpha = 0$) or approaching infinity ($\beta/0$; $\beta/\alpha$ with $\alpha \rightarrow 0$). These codons (marked in Table B.S3 and by columns of gray x in Fig. 3.1) which were included in the

FEL analysis were excluded from the subsequent selection analyses using the FitMG94 method (see below), while other internal, variable methionine, and tryptophan residues were included in both analyses.

### *Sequence Partitions*

For comparison of the selective pressures acting historically on different parts of the sHSP sequences, the human sequences were separated into defined partitions. These partitions were primarily inferred from the profiles of the ω-values as returned by the FEL-method (cf. Figure 3.1), thereby separating regions with relatively high, medium, or low average ω-values. The obtained partitions correspond approximately, though not perfectly, to the domains and regions as identified earlier by multiple alignment of the sHSPs (cf. Appendix B Fig. B.S1): The NTR, CeR, αCD, and CeR can be discerned in most cases. Where this ω-plot-based discrimination was not possible (e.g., for some regions of HspB5), multiple alignment data were also used to demarcate the partitions as used in this study. Partitions corresponding to the exons of HspB1, HspB5, and HspB8 were separated according to the ENSEMBL web site (https://useast.ensembl.org/index.html?redirect=no).

### *Determination of the Evolutionary Pressure on Full-length Sequences and Sequence Partitions*

In addition to the evaluation of selection at the level of individual codon sites, selection at the level of the coding regions of the sHSPs as a whole (full-length sequences) or at the level of the partitions (corresponding approximately to the NTR, CeR, αCD, CTE)

was also evaluated. The goal of this partition-level selection analysis was to determine the selective pressure on defined regions relative to the other regions of that same sHSP. Additionally, the collected partition-level data permit the comparison of homologous partitions between the various sHSPs. Also, partitions corresponding to the exon structure of these sHSPs were compared to provide insight into the evolutionary history of the structure of these genes. Aggregate estimates of the omega values (ω) were calculated using the standard MG94xREV codon substitution model implemented in the FitMG94 algorithm (https://github.com/veg/hyphy-analyses/tree/master/FitMG94) (Kosakovsky Pond et al. 2010) and represent either the full-length sequence-wide inference or the partition-wide inference of the alignment-wide omega (ω) values (point estimates) with associated confidence intervals, an upper and lower bound. This approach provides a quantitative measure of the selective pressures historically applied to each partitioned region on average. The calculated alignment-wide aggregate ω values are listed in Appendix B Table B.S4 with upper and lower bounds and are marked by indices corresponding either to the full-length sequences or to the various partitions, e.g., $\omega_{FL}$, $\omega_{NTR}$, etc. To test for significantly different aggregate ω estimates, we performed a one degree of freedom likelihood ratio test where the null hypothesis enforced the same aggregate value on both genes or partitions, and the alternative hypothesis estimated them separately (FitMG94-Compare).

Statistical relationships with significantly different full-length or partitioned sequences, as identified by this analysis, are indicated in the Hasse diagrams in Figs. 2

through 4 by arrows, pointing toward the sequences with the greater numeric aggregate ω values, i.e., with less stringent purifying selection.

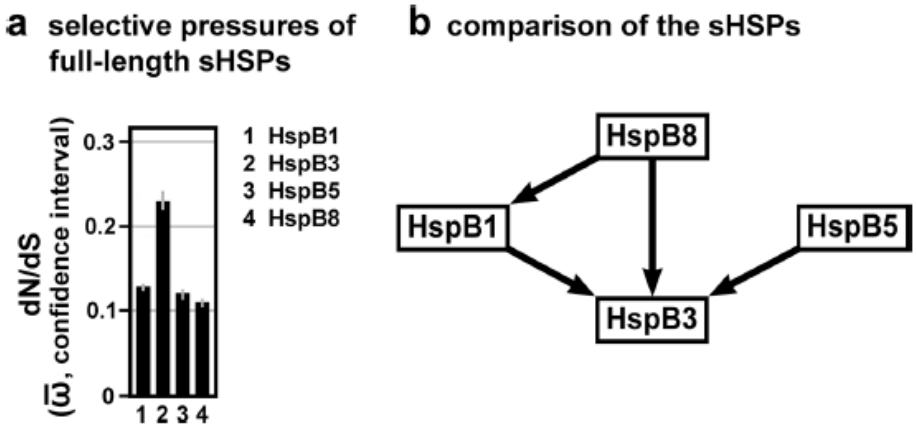



Figure 3.2. Selective pressures estimated for the full-length sHSP sequences. a, Plot of the dN/dS estimates (aggregate ω-values with confidence intervals) of the full-length sHSPs, as returned by the FitMG94 algorithm. The data were taken from Appendix B Table B.S4. b, Hasse diagram, as inferred by the FitMG94-Compare method, demonstrating the statistical relationship between the aggregate ω-values of the four human sHSPs. The arrows point to the sHSPs with the higher ω-values, i.e., they were historically exposed to a less stringent purifying selection.

## Results

### *Selective Pressure Profiles Along the Human sHSP Sequences*

The ω-values (dN/dS point estimates) for each codon of the Gnathostomata HspB1, HspB3, HspB5, and HspB8 alignments, as returned by the FEL model (Appendix B Table

B.S3), were plotted along the sequences of the human sHSPs, omitting all codons which have no homologous sites in the human sequences (Fig. 3.1). Overall, these results reveal that most of the codons were exposed to varying degrees of purifying selection over their evolutionary history (no columns, black columns). A minor fraction of codons in each sHSP do not reveal a detectable signal for purifying selective pressure, tentatively suggesting that they have evolved neutrally (gray columns). Obviously, there are differences between these sHSPs in terms of (1) the fraction of these neutrally evolving codons, (2) their distribution along the sHSPs sequences, and (3) the strength of the detected selective pressures (see below). In HspB1, the codons with neutral signature (8/205; 3.9%) are distributed relatively evenly along the sequence with just one cluster of two codons (T202, A203) found in the distal CTE (Fig. 3.1a). In HspB3, the fraction of neutrally evolving codons is highest among the studied sHSPs (28/150; 18.7%) with most of these codons clustered in the CeR (Fig. 3.1b). Interestingly, HspB3 is the only studied sHSP with codons that were historically exposed to a detectable positive diversifying selection (T48, V147), and not by chance these codons are found in regions with less stringent conservation or even neutral evolution: in the CeR and CTE. While in HspB5 the codons with neutral signature (11/175; 6.3%) are relatively evenly distributed over the entire sequence (Fig. 3.1c), in HspB8 these codons (9/196; 4.6%) are clustered, to a certain extent, in the CeR and CTE (Fig. 1d).

A closer inspection of FEL ω-plots (cf. Figure 3.1) reveals alternating patterns with regions of lower and higher ω-values, indicating variable degrees of higher and lower

purifying selection, respectively, acting over the evolutionary history of these sequence sections. Based on these profiles, the sequences of the human sHSPs were partitioned as specified in Fig. 1a through d (for HspB1, HspB3, HspB5, and HspB8, respectively) and in Appendix B Table B.S4, in order to separate these sequence segments. These partitions correspond approximately to the domains and regions in the sHSPs as identified earlier by sequence alignments: NTR, CeR, αCD, and CTE (cf. Appendix B Fig. B.S1). Such differentiation of the ω-profile into regions with lower and higher ω-values is most pronounced in HspB3 (Fig. 3.1b) and HspB8 (Fig. 3.1d) and to a lesser extend in HspB1 (Fig. 3.1a). In HspB5, it is least pronounced (Fig. 1c) thus complicating the allocation of the partitions. In regions with unclear differentiation (i.e., low vs. high ω-values), information from the sequence alignment (Fontaine et al. 2003; cf. Appendix B Fig. B.S1) was also used for the delimitation of the partitions. The partitions as defined in Fig. 3.1 and Appendix B Table B.S4 were used to quantitatively determine the selective pressures (measured as aggregate ω) to which the respective regions of these sHSPs were historically exposed to (see below).

***Differences in the Selective Pressures between the sHSPs (full-length sequences)***

Next, we attempted to quantitatively determine differences in the selective pressures acting over the evolutionary history of the Gnathostomata sHSPs, considering the full-length sequences. Based on the codon-specific ω-values (dN/dS point estimates) as shown in Appendix B Table B.S3, aggregate ω-values for each full-length sHSP ($\omega_{FL}$; cf. Appendix B Table B.S4) were inferred using the FitMG94 algorithm (Fig. 2a). Clearly,

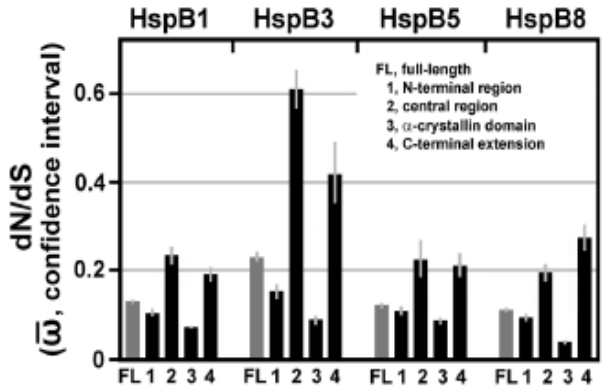
the obtained aggregate $\omega_{FL}$-values between these sHSPs differ: HspB8 and HspB3 were historically exposed to the most stringent (lowest $\omega_{FL}$) and most relaxed (highest $\omega_{FL}$) purifying selection, respectively, with HspB1 and HspB5 being placed in-between. Further statistical testing using the FitMG94-Compare algorithm confirms that several of these differences indeed exhibit a great statistical confidence, as represented by the Hasse diagrams (Fig. 2b). The arrows point to the sHSPs with higher aggregate $\omega_{FL}$-values: (1) HspB3 evolved under a less stringent purifying selection compared to the other studied sHSPs; and (2) HspB1 evolved under a less stringent level of purifying selection than HspB8. No differences were detected between HspB5 on the one hand and HspB8 and HspB1 on the other hand, applying the same statistical confidence criteria. Such differences in the overall selective pressures between some of the sHSPs, as demonstrated here, may be related to the frequency by which these sHSPs harbored disease-associated mutations (cf. Discussion).

***Differences in the Selective Pressures between the Regions within the sHSPs***

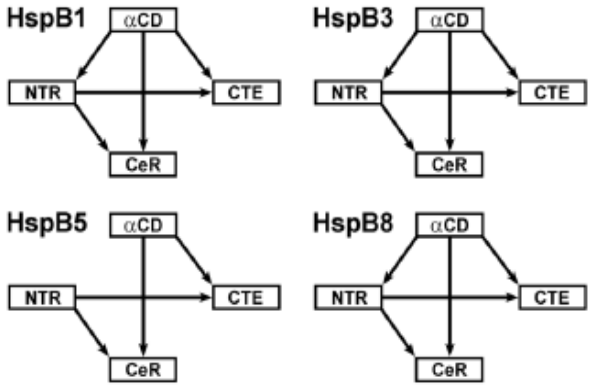
In the next step, we sought to verify that different parts of the four studied sHSP sequences were indeed historically exposed to different degrees of selective pressures, as is suggested by Fig. 3.1. Again, we used the codon-specific ω-values (dN/dS point estimates; Appendix B Table B.S3) and inferred, using the FitMG94 algorithm, the aggregate ω-values for each sequence partition as defined in Fig. 3.1. As expected, the aggregate ω-values for the various partitions exhibit differences within each sequence (Appendix B Table B.S4). Plotting these partition-specific aggregate ω-values results in

profiles which are, to a certain extent, characteristic for the studied sHSPs (Fig. 3.3a). For example, the highest and least degrees of purifying selection, among all partitions, are seen in the αCD of HspB8 and CeR of HspB3, respectively, not counting the exon-specific partitions (see below).

**a** sHSP region-specific selective pressures




**b** comparison of the regions within each sHSP




**c** comparison of the homologous regions of HspB1, HspB3, HspB5 and HspB8

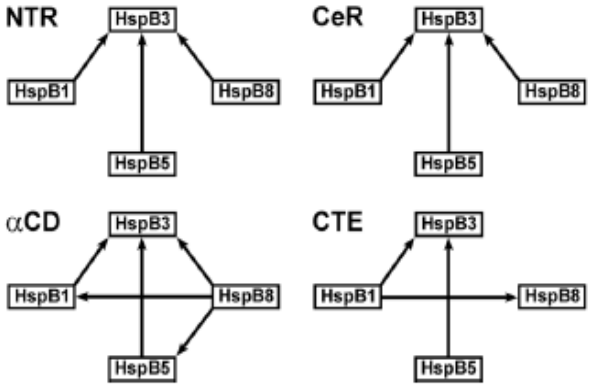


Figure 3.3. Selective pressures estimated for the various sHSP domains and regions. a, Plot of the dN/dS estimates (aggregate ω-values with confidence intervals) of the region-specific partitions of the four studied sHSPs, as returned by the FitMG94 algorithm. The data was taken from Appendix B Table B.S4. For comparison, the aggregate ω-values of the full-length sequences are also included (gray columns). b, c, Hasse diagrams, as determined by the FitMG94-Compare method, demonstrating the statistical relationship between the aggregate ω-values of the regions within each sHSP sequence (b) or between the homologous regions of the four sHSPs (c). The arrows are used as in Fig. 3.2.

Further testing using the FitMG94-Compare algorithm reveals the high statistical confidence for most of the differences between the partitions within each sHSP sequence (Fig. 3.3b). The resulting Hasse diagrams demonstrate that the αCDs of HspB1, HspB3 and HspB8 were exposed to a more stringent purifying selection than the other regions within the same sHSP sequences. For HspB5, however, no such statistical difference between the αCD and the NTR could be discerned, although the αCD was under a more stringent purifying selection than the CeR and CTE. In all four studied sHSPs, the NTR was under a more stringent purifying selection than the CeR and the CTE, whereas no differences were detected between the CeRs and the CTEs.

In summary, within each studied sHSP, sequence partitions can be identified which exhibit detectable and significant differences in the degrees of purifying selection they were historically exposed to.

***Differences in the Selective Pressures between the Homologous Regions of the Four Studied sHSPs***

We used the collected aggregate ω values (Appendix B Table B.S4) also for the comparison among the homologous regions, as defined in Fig. 3.1 and Appendix B Fig. B.S1, of the four studied sHSPs (Fig. 3.3a). Applying the FitMG94-Compare algorithm provided the Hasse diagrams shown in Fig. 3.3c, resulting in these relationships: The NTR, CeR, and αCD of HspB3 exhibit the least degrees of purifying selection when compared to the homologous partitions of the other sHSPs, while the CTE of HspB3 exhibits a lesser degree of purifying selection when compared to the homologous partitions of HspB1 and HspB5, but not of HspB8. The αCD of HspB8 was exposed to a stronger purifying selection than the αCDs of all other sHSPs. Quite differently, the CTE of HspB8 was exposed to a lesser degree of purifying selection than the CTE of HspB1. None of the partitions exhibits a significant difference between HspB1 and HspB5, a finding that is consistent with the fact that we could not detect a significant difference between the full-length sequences of both proteins (cf. Figure 3.2b). Similarly, most of the partitions (NTR, CeR, CTE) do not exhibit a difference between HspB5 and HspB8, except the αCD which clearly exhibits a difference. This difference, however, does not have a sufficient weight to cause a difference at the level of the full-length sequences (cf. Figure 3.2b). In summary, several significant differences in the degrees of purifying selection between the homologous domains and regions of the studied sHSPs were detected.

### *Differences in the Selective Pressures between the Exons*

HspB1, HspB5, and HspB8 are encoded by genes consisting of three exons (two introns), whereas HspB3 is encoded by a one-exon (intronless) gene (cf. Figure 3.1). In HspB1, HspB5 and HspB8, exon 2 encodes the core part of the αCD, the most conserved region in these sHSPs (Fontaine et al. 2003; Franck et al. 2004). Applying the FitMG94 algorithm, we used the codon-specific ω-values (dN/dS point estimates; Table B.S3) and inferred the aggregate ω-values for each exon-specific partition. As expected, plotting the exon-specific aggregate ω-values for HspB1, HspB5 and HspB8 confirms that exons 2 were historically exposed to the strongest purifying selection, compared to exons 1 and 3, within each sHSP, with HspB5 exhibiting the least pronounced effect (Fig. 3.4a; cf. Table B.S4). When the FitMG94-Compare algorithm was applied, this finding was confirmed with a high statistical confidence (Fig. 3.4b). This analysis reveals also that exons 1 of HspB1 and HspB8 were under a more stringent purifying selection than exons 3 of the same sHSPs. In HspB5, no such difference between exons 1 and 3 can be demonstrated. When the homologous exons among the three sHSPs were compared, the FitMG94-Compare algorithm did not detect such differences which would hold out against this level of statistical confidence, with one exception: Exon 2 of HspB5 was historically exposed to a less stringent selective pressure than exons 2 of HspB1 and HspB8 (Fig. 3.4c). This relation is noteworthy and may have consequences for the frequency by which exons 2 in these sHSPs harbor disease-associated mutations (cf. Discussion). In summary, the various exons of HspB1, HspB5, and HspB8 were historically exposed to differential degrees of

purifying selection. This relationship is most pronounced when exons 2 are compared with exons 1 and 3 within the same sHSP on the one hand and among one another on the other hand.

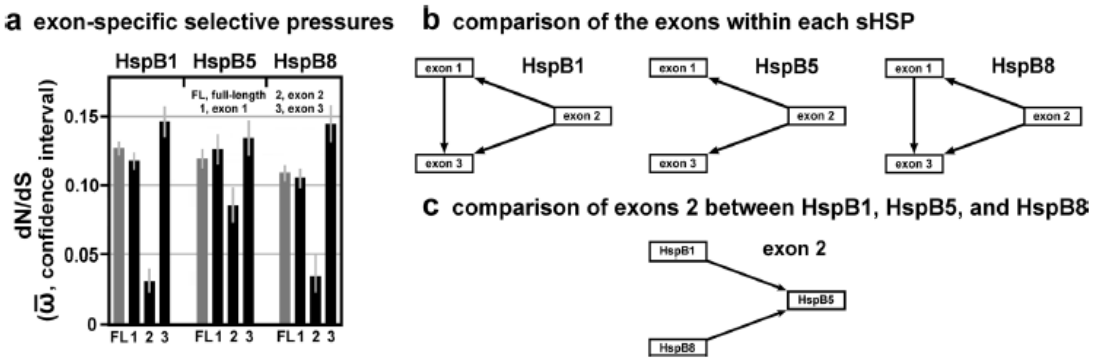



Figure 3.4. Selective pressures estimated for the exons of HspB1, HspB5, and HspB8. a, Plot of the dN/dS estimates (aggregate ω-values with confidence intervals) of the various exon-specific partitions, as returned by the FitMG94 algorithm. The data were taken from Table B.S4. For comparison, the aggregate ω-values of the full-length sequences are also included (gray columns). b, c, Hasse diagrams, as determined by the FitMG94-Compare method, demonstrating the statistical relationship between the aggregate ω-values of the exons within each sHSP sequence (b) or between exons 2 of the three sHSPs (c). The arrows are used as in Fig. 3.2

### *Selective Pressure on the Missense Mutation Sites*

Based on the used set of Gnathostomata orthologs of HspB1, HspB3, HspB5, and HspB8, stringent purifying selection was detected at all mutation sites harboring missense mutations (cf. Figure 3.1, Table B.S3; with the relevant data summarized in Fig. 3.5). Two

further observations are noteworthy: (1) All mutation sites in the αCD of HspB3 and HspB8, and three out of four mutations sites in the αCD of HspB5, exhibit only synonymous variation ($\omega = 0$), as opposed to most mutation sites in the αCD of HspB1 that exhibit both, synonymous and nonsynonymous variation; and (2) mutation sites positioned in the CTEs of HspB1 and HspB5 exhibit, by trend, weaker purifying selective pressure than mutation sites in the other partitions. The four known mutation sites that harbored missense mutations and that are associated with recessive disease phenotypes (G53, S86, L99 in HspB1; R56 in HspB5) all exhibit ω-values $< 0.05$, indicating highly stringent purifying selection.

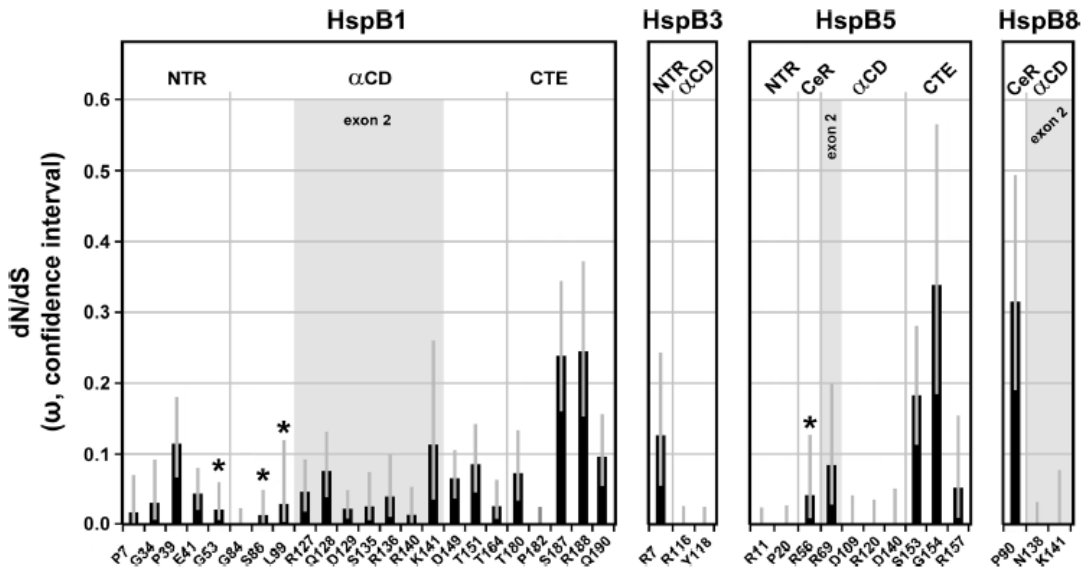



Figure 3.5. Summary of the selective pressures detected at the disease associated missense mutation sites of the four studied sHSPs. The dN/dS point estimates (ω-values with confidence intervals) for each mutation site were taken from Table B.S3. The positions of the mutation sites within the various sequence partitions (domains, regions, exons) of the sHSPs are indicated (cf. Table B.S1, Fig. 3.1). The asterisks mark mutation sites whose mutations are associated with a recessive disease phenotype. All other mutation sites are affected by mutations that are associated with a dominant disease phenotype, or this association can be assumed.

Interestingly, there is a disparate distribution of the known disease-associated missense mutations among the exons in the various sHSPs. Exons 2 in HspB1 and HspB8 harbored a substantial fraction of the known mutations in the respective sHSPs, whereas exon 2 in HspB5 harbored a minor fraction (just a single mutation), among all mutations known to

date (Fig. 3.5). The reason for this disparate distribution is not known, although it may be related to the relatively relaxed purifying selection seen in exon 2 of HspB5 (cf. Figure 3.4; see Discussion).

## Discussion

In this study, we explore the evolutionary history of four human sHSPs: HspB1, HspB3, HspB5, and HspB8, all known disease factors as mutant alleles have been found associated with various forms of neuropathies, myopathies, and with cataracts in the lens of the eye. We found that the Gnathostomata sHSPs have been generally under strong purifying selection that is typical of highly conserved and functionally constrained genes. We observed significantly different degrees of purifying selection between these sHSPs on the one hand and between the various sequence partitions on the other hand, with relatively higher and lower selective pressures within each sHSP that correspond approximately to the domains and regions NTR, CeR, αCD, and CTE. Our study demonstrates that the obtained profiles of the strength of the selective pressures are characteristic attributes for the studied sHSPs.

Our results provide a partial explanation for some of the unusual features of the sHSPs, including (1) the dimorphic pattern of their evolution, (2) the observed differences in the mutation rates between the sHSPs in humans, (3) and the prevalence of genetically dominant disease phenotypes resulting from gain-of-function mutations. These features are discussed in detail below.

(1) The dimorphic pattern of sHSP evolution. As determined earlier by phylogenetic analysis (Kriehuber et al. 2010), the evolutionary history of the sHSPs with the dimorphic pattern exhibits a striking feature: While the isolated central part of these proteins, the αCD, evolved along the species tree across all domains of life from the beginning, the flanking regions (NTR + CeR, CTE) have trees different from the αCD being consistent with extensive remodeling and complex evolutionary dynamics independently of the αCD. The same analysis revealed also that all Metazoa (including Gnathostomata) sHSPs are derived from a single ancestral sequence, with the gene number increasing by gene duplication mostly after the separation of the different Bilateria taxa during the Cambrian species explosion. Interestingly, certain characteristic N-terminal sequences found otherwise scattered all over the αCD-based phylogenetic tree and which served to establish the evolutionary dimorphic pattern, were not detected in the Metazoa branch of this phylogenetic tree (Kriehuber et al. 2010). In this context, our collected data provide an argument for the extension of this dimorphic evolutionary pattern also to the branch of the Gnathostomata sHSPs. Differences in the selective pressures between the αCD on the one hand and the flanking regions (NTR + CeR, CTE) on the other hand imply that this dimorphic evolutionary pattern continued also in the taxon Gnathostomata (cf. Figure 3.3b).

A closer inspection reveals also certain differences in the selective pressures between the different partitions of the flanking regions in all four studied sHSPs. For example, the purifying selective pressures in the NTR was historically stronger than in the

CeR and CTE of all four studied sHSPs (cf. Figure 3.3b). These differential selective pressures likely played a role in shaping the vertebrate sHSPs, even if the specifics are not understood at this time. The importance of these flanking regions for both the structure and function of the mammalian sHSPs has been demonstrated (Selig et al. 2020).

(2) Differences in the mutation rates between the sHSPs in humans. As shown in Table B.S1 and in the references therein, sHSPs harbor disease-associated mutations with different incidences. This is especially evident if HspB1, HspB3, and HspB8 are compared, all of them causing, when mutated, a similar spectrum of peripheral motor neuropathies. HspB1 harbored by far most of the mutations as compared to HspB3 and HspB8: In a systematic study of a cohort of 510 unrelated index patients with peripheral neuropathies, 32 patients were found to have mutations in sHSPs (32/510; 6.3%) (Echaniz-Laguna et al. 2017). Among them, 28 patients carried mutations in HspB1, four patients in HspB8, and none in HspB3. Thus, in the patient population of this study, HspB1 is approximately seven times more frequently affected than HspB8, while HspB3 apparently plays only a marginal role, if any. In another study of 163 patients belonging to 108 families with distal hereditary motor neuropathies, a genetic diagnosis was achieved in 37/108 families (34.2%) or in 78/163 of all patients (47.8%) (Frasquet et al. 2021). Among them, the most frequent cause of disease was mutations in HspB1, besides mutations in several other genes. In that study, no mutations were found in HspB8 or HspB3. Similarly, in other systematic studies with smaller patient cohorts, a number of mutations in HspB1 were found, but none in the other sHSPs (Capponi et al. 2011; Luigetti et al. 2016).

This disparate incidence may be directly related to the different degrees of purifying selection these sHSPs have been historically subjected to (measured as aggregate ω-values for the full-length sequences ($\omega_{FL}$). HspB3 and HspB8 were exposed to the most relaxed and most stringent purifying selection, respectively, with HspB1 being placed in-between (cf. Table B.S4, Fig. B.2). Because of the historically most stringent purifying selection, it is plausible that mutations in HspB8 would lead more frequently to impaired fitness, thus eliminating the disease alleles over time from the human population. Conversely, mutations in HspB3 may have fewer consequences for the fitness, thus allowing the accumulation of mutated alleles in the population. If this were the case, there should exist a relatively large number of HspB3 allelic variants in the healthy human population. HspB1, positioned in-between, exhibits a certain degree of 'plasticity', allowing it to harbor mutations at an increased rate compared to HspB8, although this comes with a cost for the health. However, since a significant proportion of mutant HspB1-associated disorders have a late onset in life (references in Table B.S1), these mutations are likely to come with relatively little fitness costs in the human population in evolutionary terms. In this scenario, most mutant HspB1 alleles should be under weak purifying selection in the human population, if any, thus permitting their accumulation.

These considerations cannot simply be extended to HspB5 as the disease phenotypes associated with its mutations are quite different. However, in terms of both the estimated overall selective pressure for full-length HspB5 (cf. Figure 3.2) and the number of the harbored missense mutations (cf. Figure 3.5), HspB5 shares some similarity with

HspB1. Yet, another feature is noteworthy: Exon 2 of HspB5 harbored only a minor fraction (just one) of the known mutations, quite different from the exons 2 of HspB1 and HspB8 that harbored substantial fractions of the known mutations (cf. Figure 3.5). This may be related to the fact that exon 2 of HspB5 is distinguished by the relatively relaxed purifying selection it was historically subjected to, compared to HspB1 and HspB8 (cf. Figure 3.4a, c). Following the above logic, exon 2 of HspB5 may have a relatively high tendency to accumulate a number of allelic variants in the healthy human population without major fitness costs in evolutionary terms.

(3) The prevailing of genetically dominant disease phenotypes resulting from gain-of-function mutations. The fact that most missense mutations in sHSPs are associated with dominant (or semi-dominant) disease phenotypes is contrary to the expectation, as mutagenesis studies in many organisms revealed that ~ 90% of the wild-type alleles are dominant over the mutant alleles, with dominant mutant phenotypes being the exception in the wider picture (Wilkie 1994). Thus, in addition to the peculiar dimorphic evolutionary history, sHSPs are also distinguished by their atypical genotype–phenotype relation.

Heritable diseases in general are believed to persist in the human population because of a balance between mutations, genetic drift, and natural selection, the latter eventually eliminating the mutation from the population by purifying selection (Blekhman et al. 2008). For alleles causing highly penetrant Mendelian disorders, purifying selection can be quite strong, unless the disease outbreak has a late onset as is the case with most mutant sHSP-associated disorders (see references in Table B.S1). Thus, in the human

population the majority of the mutant sHSP alleles can be expected to be under reduced purifying selection, as the late onset is unlikely to impose major fitness costs in evolutionary terms. This is, however, not in conflict with the fact that all sHSP mutation sites themselves were exposed to a stringent purifying selection, together with the entire genes, during the Gnathostomata evolution, as we show here. Indeed, it was observed that Mendelian disease genes in general are under a more widespread and stronger purifying selection if the disease-causing alleles are dominant, as compared both to recessive and more complex disease genes (Blekhman et al. 2008). While this interrelation is highly interesting, the mechanisms behind it still await their elucidation.

Disease-causing mutated alleles of the four studied sHSPs with their high phenotypic penetrance and segregation of the disease traits in a Mendelian fashion qualify them as Mendelian disease genes. All four sHSP genes harboring such mutations were historically subject to purifying selection. These facts and observations offer an explanation for both the high proportion of dominant mutations among neuromuscular disease-associated sHSP mutations, and also for the relatively high proportion of sHSP mutations among all mutations (6.3% and 10.4%; Echaniz-Laguna et al. 2017; Frasquet et al. 2021; respectively) in patients with inherited peripheral neuropathies.

Historically, the phenomenon of genetic dominance has been fiercely debated (summarized in Benndorf et al. 2014). One aspect is that the existence of multiple paralogous genes, such as the 10 sHSP genes in humans, seems to be related to their receptivity to harboring dominant-negative mutations. The evolutionary retention of such

paralogs in polyploid species, or species with polyploid ancestry that includes Homo sapiens, was explained by gene dosage effects and dominant-negative effects involving supramolecular structures (Comai 2005; Veitia 2010; Veitia and Birchler 2010).

Such oligomeric complexes are indeed a hallmark of all studied sHSPs (Mymrikov et al. 2011; Boelens 2020; Muranova et al. 2020). Mutations associated with dominant phenotypes fall into two categories: loss-of-function type (i.e., haploinsufficiency) and gain-of-function type (Wilkie 1994), although mixed types might also occur. For the dominant sHSP mutations, a loss-of-function situation can be tentatively dismissed. Arguments come from HspB1-, HspB5-, and HspB8-knock-out rodents that do not reproduce overt disease phenotypes as are seen for the dominant missense mutations (Brady et al. 2001; Huang et al. 2007; Bouhy et al. 2018), despite the fact that some impaired cell biological pathways, e.g., the autophagosome formation, could be identified (Haidar et al. 2019). Additional biochemical evidence supports this notion: Mutations in sHSPs do not necessarily result in the loss of key functions such as chaperoning and anti-apoptotic activities (Krishnan et al. 2008; Almeida-Souza et al. 2010), even if some modification, e.g., of the chaperoning specificity, may occur (Weeks et al. 2018). These findings tentatively suggest that loss-of-function mechanisms may not be the primary mechanism responsible for the manifestation of the disorders.

Regarding the dominant gain-of-function mutations, there are only a handful principal mechanisms that seem to accommodate most situations (Wilkie 1994): (1) 'classic' dominant negative effects (e.g., through protein–protein interactions); (2)

formation of toxic gene products (e.g., of amyloids); (3) impact on cytoskeletal systems resulting in altered cell architecture; and (4) increased enzymatic activities.

Existing experimental, genetic and clinical evidence suggests that most if not all studied dominant disease-associated sHSP mutant alleles fit this pattern. Table S5 summarizes some of the molecular and cellular consequences of the expression of dominant mutant sHSP alleles, all supporting gain-of-function mechanisms. The expression of sHSP mutant alleles results in abnormal protein–protein interactions and in abnormal quaternary structures (implying dominant negative effects, summarized, e.g., for a number of HspB1 mutants by Muranova et al. (2020)), in the deposition of protein aggregates (implying the formation of highly toxic amyloid precursors and altered cell architecture), and in increased downstream enzymatic activities. This situation should have direct consequences for any future therapeutic interventions which should aim to reduce these toxic effects of the mutated proteins, e.g., by reducing the expression of the mutant proteins, by 'detoxification' of toxic amyloid precursors of the aggregates, or by inhibition of the abnormally increased downstream enzymatic activities. The latter point was experimentally addressed by d'Ydewalle et al. (2011) who found that expression of mutant HspB1 decreased the abundance of acetylated α-tubulin in mice that was paralleled by axonal transport deficits. A pharmacological inhibition of the histone deacetylase 6 restored the α-tubulin acetylation together with the axonal transport, and thus, rescued the disease phenotype. Also, pharmacological targeting sHSPs and their abnormal interactions

with other proteins resulting from mutations are within reach of future approaches (Boelens 2020).

Taken together, our data shed light onto some unusual features of the evolution of four of the vertebrate sHSPs (HspB1, HspB3, HspB5, HspB8) and how their evolutionary history has shaped these proteins. We provide evidence that the vertebrate evolution of these sHSPs exhibits a dimorphic pattern, with different degrees of purifying selection acting historically upon the various domains and regions of these sHSPs. When compared to one another, the studied sHSPs have been exposed to different degrees of purifying selection. These features have implications for the role of these proteins in human health and disease and may explain both the different incidences by which the various sHSPs harbored disease-associated missense mutations in the human population, and also the unusual genotype–phenotype relation in patients with neuromuscular diseases resulting from these mutations. Experimental findings and theoretical considerations support the prevalence of toxic gain-of-function mechanisms for the manifestation of the mutant sHSP-associated diseases in most individuals. Thus, our study puts both the genetics and pathomechanisms of inheritable sHSP-borne diseases into the context of vertebrate evolution. In so doing, we provide support for Theodosius G. Dobzhansky's famous realization that 'nothing in biology makes sense except in the light of evolution' (Dobzhansky 1973). The extension of this paradigm to inheritable human diseases may help to advance the emerging field of 'Evolutionary Medicine'.

# CHAPTER 4.

# THE EVOLUTION OF BDNF IS DEFINED BY STRICT PURIFYING SELECTION AND PRODOMAIN SPATIAL COEVOLUTION, BUT WHAT DOES IT MEAN FOR HUMAN BRAIN DISEASE?


## Abstract

Brain-Derived Neurotrophic Factor (BDNF) is an essential mediator of brain assembly, development, and maturation. BDNF has been implicated in a variety of brain disorders such as neurodevelopmental disorders (e.g., autism spectrum disorder), neuropsychiatric disorders (e.g., anxiety, depression, PTSD, and schizophrenia), and various neurodegenerative disorders (e.g., Parkinson's, Alzheimer's, etc.). To better understand the role of BDNF in disease, we sought to define the evolution of BDNF within Mammalia. We conducted sequence alignment and phylogenetic reconstruction of BDNF across a diverse selection of >160 mammalian species spanning ~177 million years of evolution. The selective evolutionary change was examined via several independent computational models of codon evolution including FEL (pervasive diversifying selection), MEME (episodic selection), and BGM (structural coevolution of sites within a single molecule). We report strict purifying selection in the main functional domain of BDNF (NGF domain, essentially comprising the mature BDNF protein). Additionally, we discover six sites in our homologous alignment which are under episodic selection in early

regulatory regions (i.e., the prodomain) and 23 pairs of coevolving sites that are distributed across the entirety of BDNF.

Coevolving BDNF sites exhibited complex spatial relationships and geometric features including triangular relations, acyclic graph networks, double-linked sites, and triple-linked sites, although the most notable pattern to emerge was that changes in the mature region of BDNF tended to coevolve along with sites in the prodomain. Thus, we propose that the discovery of both local and distal sites of coevolution likely reflects 'evolutionary fine-tuning' of BDNF's underlying regulation and function in mammals. This tracks with the observation that BDNF's mature domain (which encodes mature BDNF protein) is largely conserved, while the prodomain (which is linked to regulation and its own unique functionality) exhibits more pervasive and diversifying evolutionary selection. That said, the fact that negative purifying selection also occurs in BDNF's prodomain also highlights that this region also contains critical sites of sensitivity which also partially explains its disease relevance (via Val66Met and other prodomain variants). Taken together, these computational evolutionary analyses provide important context as to the origins and sensitivity of genetic changes within BDNF that may help to deconvolute the role of BDNF polymorphisms in human brain disorders.

## Introduction

Brain-derived neurotrophic factor (BDNF) is one of the most ubiquitously studied molecules in modern neuroscience (Notaras and van den Buuse, 2015). BDNF is a

neurotrophin that binds with high affinity to its cognate tyrosine kinase receptor, TrkB (Nagappan and Lu, 2005), to elicit rapid induction of synaptic plasticity (Black, 1999; Yoshii and Constantine-Paton, 2010; Sakuragi et al, 2013) and neuronal spine remodeling (Horch et al, 1999; Giza et al, 2018).

Additionally, BDNF has been implicated in a variety of brain disorders (Notaras and van den Buuse, 2015), including depression (Martinowich et al, 2007; Angelucci et al, 2005; Kim et al, 2007), PTSD (Notaras and van den Buuse, 2020; Pivac et al, 2012; Pitts et al, 2019; Zhang et al, 2014), schizophrenia (Angelucci et al, 2005; Notaras et al, 2015; Gratacòs et al, 2007; Zakharyan et al, 2011), Parkinson's disease (Howells et al, 2000; Palasz et al, 2020), and autism spectrum disorders (Correia et al, 2010; Ricci et al, 2013; Tsai, 2005) amongst many more. BDNF has correspondingly been the primary target, or an ancillary factor, of many novel therapeutics including small molecule mimetics (Massa et al, 2010; Kingwell, 2010) and existing drugs (e.g., antidepressants (Chen et al, 2001; Björkholm and Monteggia, 2016)). Yet, nascent research has provided the humbling reminder that much remains to be discovered about BDNF. In recent years, new BDNF ligands have been discovered (Anastasia et al, 2013), new receptor interactions unveiled (Anastasia et al, 2013; Glerup et al, 2016), and mechanisms of behavioral function unlocked (Giza et al, 2018). This is a timely reminder that while BDNF has remained a seminal molecule of interest across the broader neuroscience literature, much remains to be discovered about its origins, evolution, function, and disease relevance.

### *A Primer of the Molecular Biology of BDNF and its Functional Topology*

BDNF is encoded by the BDNF gene (Notaras and van den Buuse, 2019), whose expression is regulated in humans by an antisense gene (BDNF-AS) that can form RNA-duplexes to attenuate translation (Pruunsild et al, 2007). Thus, the natural antisense for BDNF is capable of directly downregulating endogenous expression on demand (Lipovich et al, 2012). The BDNF gene in humans comprises 11 exons (Pruunsild et al, 2007) and can produce at least 17 detectable transcript isoforms (Notaras and van den Buuse, 2019). Different transcripts are induced in response to activity and/or cellular states, allowing the BDNF gene to adjust to environmental stimuli and potential selection pressures. However, all transcripts ultimately yield a singular preproBDNF protein that (prior to intracellular processing, cleavage, and transport) can be partitioned into three domains (Notaras and van den Buuse, 2020; Notaras and van den Buuse, 2019): a signal peptide, a prodomain, and the mature domain. The signal domain is only 18 amino acid residues long (with ambiguously defined functionality) with the majority of BDNFs functional outputs reflecting sequence specificity to the prodomain and mature domain. The BDNF prodomain encodes binding sites for intracellular transport of both BDNF mRNA (Chiaruttini et al, 2009) and BDNF protein (Chen et al, 2005), and contains numerous posttranslational modification sites (Notaras and van den Buuse, 2019). The BDNF prodomain is also the resident location of a widely studied Single Nucleotide Polymorphism (SNP) in neuroscience (Val66Met, or rs6265) (Notaras et al, 2015), and the Furin consensus sequence (Arg 125) for cleavage to its mature form (including by plasmin

(Gray and Ellis, 2008). The prodomain is composed of 110 amino acids within the N-terminus, and must be processed via proteases to generate mature BDNF (Sakuragi et al, 2013). The mature domain of BDNF is almost exclusively composed of the Nerve growth factor (NGF) domain and is responsible for the canonical trophic actions associated with BDNF (e.g., long-term potentiation, rapid-acting antidepressant effects, etc.). Following intracellular handling, processing, and transport, the preproBDNF isoform is cleaved to yield the mature BDNF peptide (which only contains the mature NGF domain).

For many years the prodomain was thought to be degraded following the facilitation of BDNF trafficking. However, recent work has shown that the cleaved prodomain can be secreted and bind as a ligand to novel receptors (e.g., SorCS2) (Anastasia et al, 2013). Thus, the BDNF prodomain can influence brain circuits as well as behavior (Giza et al, 2018). For a comprehensive, detailed, analysis of the various intricacies of the BDNF gene, protein, and its regulation, more information is provided in (Notaras and van den Buuse, 2019).

***The Conservation of BDNF and Neurotrophins: a Signal that Evolution is Important***

One of the interesting curiosities surrounding BDNF is its relationship to other neurotrophic (NT) growth factors, comprising NGF, NT-3, and NT-4. Specifically, all neurotrophins retain some intercalated functionality. Neurotrophins also share some commonalities in structure (pre-, pro-, and mature-domains) (Notaras and van den Buuse, 2019), post-translational modification potential (e.g., glycosylation (del Carmen Cardenas-

Aguayo, 2013), as well as catalytic processing, trafficking, and composition (Maisonpierre et al, 1991). Specifically, neurotrophins share approximately 50% sequence homology (Notaras and van den Buuse, 2019), and a comparison of domains and motifs reveals that each comprises a prototypic NGF domain as the principal component of the mature pro-growth peptide (see PFAM database (Finn et al, 2016)). While each neurotrophin elicits functionality via binding to cognate receptors, neurotrophins also exhibit cross-affinity amongst neurotrophin receptors (Rodriguez-Tebar et al, 1990) presumably due to their high rates of structural homology. Not surprisingly then, there is some redundancy in the trophic effects of neurotrophins, yet each still maintains nuanced functionality which remains specific to each factor during central nervous system development (Castellani and Bolz, 1999). Differences in the evolution and temporal dynamics of regulatory sequences, which target gene products to specific destinations within cell-compartments (e.g., dendrites) (Kuczewski et al, 2009) or to processing routes (e.g., the activity-dependent release pathway) which alter secretory dynamics and/or bioavailability (Chen et al, 2004), likely contribute to both similarities and differences between neurotrophins. However, almost nothing is known about how the BDNF prodomain has evolutionarily adapted to specifically regulate BDNF dynamics. While evolution has almost certainly shaped the sequences, structure, and function of BDNF, the modeling of such remains relatively unexplored but could provide important insight into the phylogenetic evolutionary history of BDNF, its selection pressure sensitivity across lineages, and quantitative metrics of evolutionary change across species.

### *Purpose of this Study*

Here, we use computational methods to explore the molecular evolution of BDNF. To reconstruct phylogenetic trees of BDNF, we utilized sequence alignments of over 160 mammalian species (all available mammalian sequences) to determine the genomic attributes of BDNF evolution that are specific to Mammalia. This analysis was specific by being constrained to sequences that have the most direct evolutionary relevance to humans. Notably, we sought to identify sites in BDNF that are subject to pervasive (i.e., consistently across the entire phylogeny) diversifying selection (FEL) or pervasive/episodic (i.e., only on a single lineage or subset of lineages, diversifying selection (MEME). Likewise, utilizing multiple models for the inference of selective pressure and the evaluation of evolutionary change, we identify novel sites within the BDNF prodomain and mature peptide coding regions that are susceptible to synonymous and nonsynonymous changes. Additionally, we investigate which sites in BDNF may be coevolving (BGM). Taken together, these computational evolutionary analyses provide an important context as to the origins and sensitivity of genetic changes within the BDNF gene, which may be important for providing insight into genetic risk factors linked to disease in humans.

## Results

We find that unique evolutionary pressures have shaped the BDNF gene across time. These forces have mostly operated through strict purifying selection. Of note, BDNF elicits tight regulation and specific functionality that can be separated from other

neurotrophins, yet these growth factors remain closely related in their structure and sequence, especially in the conserved NGF domain.

### *Evolutionary History of Mammalian BDNF*

Prior to conducting our primary evolutionary analysis, we ported our mammalian species into a platform (timetree.org, see refs. (Kumar et al, 2017, Hedges et al, 2006)) to examine the epoch events that may have influenced the analysis described here. This was an important pre-analysis step to frame the age of our genomes, and the broad-stroke evolutionary pressures that these species have been exposed to (which, in theory, could contribute to subsequent purifying selection and coevolution analyses). As expected, this revealed BDNF as an ancient gene that has been preserved throughout the mammalian lineage and has both survived and been shaped under all major evolutionary events of the past ~177 million years (data not shown). We identified several examples of species-level evolutionary epochs that cross-referenced with major earth events (e.g., bottleneck events) that have historically been believed to drive evolutionary adaptation. This included major geologic periods that are cross-referenced against earth impacts, oxygenation changes across time, atmospheric carbon dioxide concentrations, and solar luminosity. This indicates that even under extreme evolutionary pressures, the BDNF gene has exhibited (relatively speaking) very specific adaptation events (see results below) over millions of years within Mammalia. This tracks with the idea that “old genes” tend to be highly conserved, evolve more slowly, and therefore are more likely to exhibit both specific and selective changes as opposed to more dramatic permutations (e.g., gene duplications, etc.).

### *Predominant Purifying Selection in BDNF*

A common approach to gain an increased understanding of the evolutionary forces that have shaped proteins is to measure the omega ratio ω consisting of the nonsynonymous (β or dN) and synonymous (α or dS) substitution rates, with $\omega = \beta/\alpha$ for each site in a particular gene of interest (Kosakovsky Pond and Frost, 2005). We define two major changes for the amino acid being coded for at each site: synonymous changes, which keep the same amino acid coded for at a particular site, and nonsynonymous changes, which change the amino acid coded for at a particular site. Non-synonymous changes can have strong influences on the structural, functional, and fitness measures of an organism. This is in contrast to synonymous changes which leave the amino acid at a particular site unchanged but can confer weak fitness effects through the emergent properties of codon usage bias, mRNA structural stability, translation, and tRNA availability. However, synonymous changes are typically understood to represent neutral selection acting on coding sequences and provide a baseline rate against which nonsynonymous evolutionary rates can be compared. The omega ratio ω of relative rates of nonsynonymous and synonymous substitutions is a common measure in evolutionary biology of the selective pressure acting on protein-coding sequences. These estimates provide increased information availability as to the type of selection (positive, with $\omega > 1$ or negative, with omega $<1$, or neutral with omega $=1$) that has acted upon any given set of protein-coding sequences.

As FEL analysis is a sensitive measure of negative (purifying) selection, for this analysis we observe a predominant amount of purifying selection (over 66% of sites, 174 sites out of 261; Table C.S1) in our recombination-free alignment for BDNF. The dN/dS estimates for the entire alignment were plotted including 95% lower- and upper-bound estimates (see Fig. 4.1 or Table C.S1). Overwhelmingly, the mature NGF domain of the BDNF exhibited evidence of greater pervasive negative purifying selection relative to the prodomain region of BDNF. Thus, over the evolutionary history of Mammalia, negative selection has predominantly occurred in the regions of BDNF that encode the functional mature protein that binds TrkB to elicit neurotrophic effects. The mature domain of BDNF has exhibited remarkable conservation across innumerous epochs that have been defined by rapid evolutionary adaptation in other genes and taxa.

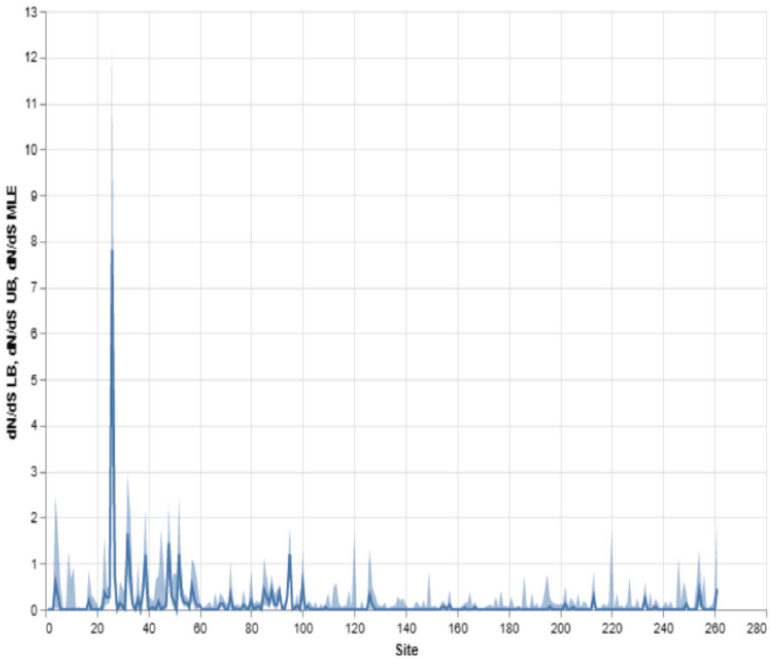



Figure 4.1. Mammalian BDNF exhibits overwhelmingly strict purifying selection, but also evidence of specific evolutionary pressures at particular sites. The FEL analysis of the BDNF gene found 174 of 261 (66.7%) sites to be statistically significant (LRT p value ≤ 0.1) for pervasive negative (purifying) selection. We plot the estimated values of omega (dN/dS) for each site in the alignment. Additionally, we plot 95% confidence intervals (CI) for each site. These results are also available in Table C.S1. We observe a high degree of strict purifying selection in the Human NGF region. The region for Human NGF corresponds to alignment sites 144–254 (NP_001700.2 and https://github.com/aglucaci/AnalysisOfOrthologousCollections/blob/main/tables/BDNF/BDNF_AlignmentMap.csv). This alignment of BDNF across all selected species (Mammalia, see Table 4.3) reveals a site-specific positive/adaptive diversifying selection and negative purifying selection. The thick line represents the point estimate (i.e., the

evolutionary pressure) and the shadings reflect 95% confidence intervals which relate to the upper and lower bound of the point estimates. As shown, the prodomain sites exhibit more pervasive/episodic and positive/diversifying evolutionary selection, consistent with the fact that more disease-associated SNPs occur in this topological region of the BDNF gene in humans (early prodomain mapping not further shown due to nuanced variation across mammalian species).

***Specific Sites that are Evolving Non-neutrally***

To examine specific sites for episodic adaptive evolutionary selection, we utilized an algorithm known as MEME which is fundamentally similar to our FEL analysis (described above) except that it applies a more sensitive method for the detection of both pervasive (persistent) and episodic selection (transient selection occurring only on one or a subset of branches in the phylogenetic tree) as compared to only pervasive selection which occurs across all branches of the phylogenetic tree. Essentially, only a subset of the lineages (i.e., species) are affected allowing for a more granular/sensitive method of detecting selection (whereas FEL is better geared towards broad changes). This analysis revealed that for all sites, only 2.3% (6 of 261; see Table 4.1) exhibit evidence for episodic diversifying selection (i.e., positive selection) in at least one branch within the phylogeny. Spatially, these mutations occur outside of the NGF functional region of BDNF. Further, this result is essentially relevant as the MEME analysis is a sensitive measure of episodic selection. The sites we observe as statistically significant were 26, 27, 30, 38, 249, and 254. For comparison, these specific sites were realigned to the respective human sites with indel

(insertion/deletion) events accounting for any respective discrepancy in specific site numbers. When mapping these sites to the human BDNF coordinate system, they correspond to sites 26, 27, 29, 36, 238, and 240, respectively.

Table 4.1. MEME analysis of the BDNF gene found 6 of 261 (2.3%) of sites to be statistically significant (LRT p-value ≤0.1).

| # | CodonSite | Human CodonSite | alpha | beta- | *p*- | beta+ | *p*+ | LRT | *p* value | # branches under selection | MEME LogL | FEL LogL | Omega |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 26 | 26 | 0.14 | 0.00 | 0.02 | 1.12 | 0.98 | 6.89 | 0.01 | 9 | −86.53 | −86.53 | 7.9 |
| 2 | 27 | 27 | 0.25 | 0.08 | 0.99 | 28.91 | 0.01 | 4.30 | 0.05 | 1 | −39.25 | −37.01 | 117.6 |
| 3 | 30 | 29 | 0.00 | 0.00 | 0.00 | 0.38 | 1.00 | 4.63 | 0.05 | 6 | −35.88 | −35.88 | inf |
| 4 | 38 | 36 | 0.72 | 0.00 | 0.93 | 8.07 | 0.07 | 3.80 | 0.07 | 5 | −69.22 | −66.28 | 11.3 |
| 5 | 249 | 238 | 0.92 | 0.00 | 0.99 | 203.11 | 0.01 | 8.25 | 0.01 | 1 | −50.74 | −44.13 | 221.6 |
| 6 | 254 | 240 | 0.23 | 0.00 | 0.99 | 1357.12 | 0.01 | 21.81 | 0.00 | 1 | −44.50 | −31.95 | 5850.4 |

### *Evidence of Coevolutionary Forces*

To examine the coevolution of sites, i.e., if one particular amino acid was evolving in-tandem with another, we subjected our protein-coding gene sequences to the BGM algorithm which leverages Bayesian graphical models (Poon et al, 2008). The BGM algorithm infers substitution history through the use of maximum-likelihood analyses for ancestral sequences and maps these to the phylogenetic tree, which allows for the detection of correlated patterns of substitution (Poon et al, 2008). For our BGM analysis, we find evidence for 23 pairs of coevolving sites. This suggests interaction dynamics in tertiary space of the 3D, folded, protein level (see relevant sites in Fig. 4.2) BDNF protein structure. Alternatively, this data may be evidence that coevolving sites may be related to other fitness consequences (e.g., compensation) for maladaptive changes in another part of the protein sequence that may have occurred. When we review these sites, we notice that several pairs (see Fig. 4.2) occur within alignment sites, which correspond to the Human

BDNF coordinate system (Table 4.1). These include pair-sites of (89, 184), (94, 155), (103, 233), and (135, 154). Of note, several other sites also display interesting geometric features including triangular relations [(81, 93), (93, 98), and (81, 98)], an acyclic graph network of site connections [(70, 74), (74, 94), (94, 155) and (25, 49), (49, 85), (49, 86)], more complex double-linked coevolutionary sites [(39, 103) and (103, 233)], and triple-linked coevolutionary sites [(30, 119), (33, 119), and (91, 119)]. Additionally, three-dimensional reconstruction—here focusing on a specific heterodimer configuration of BDNF and NT-4 as an example of a spatial protein–protein interaction—highlights that coevolving sites, as well as positively evolving sites, are likely to have been fine-tuned over time to help support BDNF's cognate functionality (see Fig. 4.3). Mapping our FEL purifying sites in a structural configuration was not shown due to the overwhelming nature of negative selection acting on BDNF within mammals.

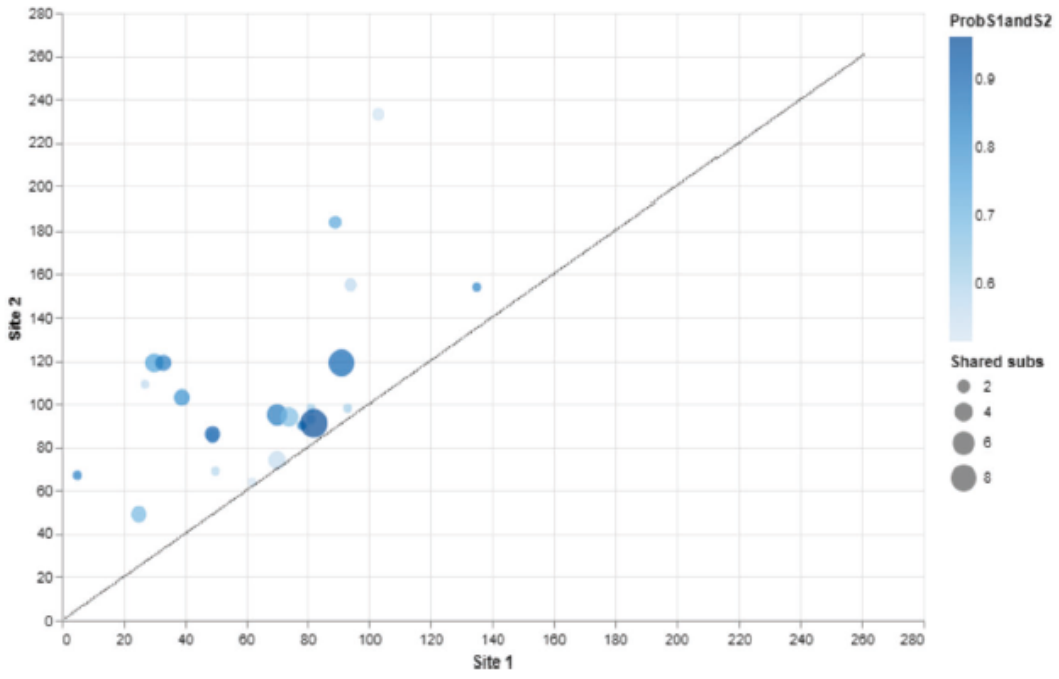



**Fig.**

Figure 4.2**.** The BGM analysis of BDNF found 23 pairs of coevolving sites out of 261 total sites to be statistically significant (with a posterior probability threshold of 0.5). Here, we plot only the statistically significant coevolving pairs. The number of shared substitutions between pairs of coevolving sites is visualized by the size of the circle, with larger circles indicating more shared substitutions. Poster probability of the interaction (coevolving pair) corresponds to the color blue, with dark blue indicating higher values. Individual BDNF sites are mapped on both the X and Y axis so that readers can view which sites are coevolving with another. Once more, note that the coevolution tends to be focally constrained to the broader BDNF prodomain region at a topological level, which is once more consistent with the idea that the NGF domain (site >144; see Fig. 3) is highly conserved and probably deleteriously impacted by variation. However, we did discover four sites of coevolution in the NGF domain (basically, the mature BDNF protein) that are

evolving with early prodomain sites. This highlights that both proximal and distal sites in BDNF can, and indeed are, evolving together over time.

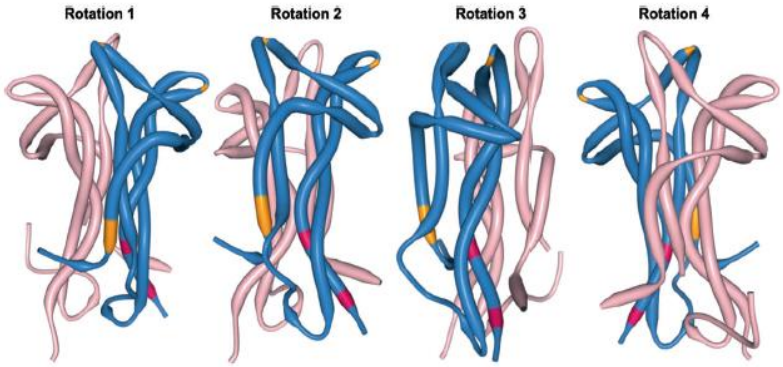



Figure 4.3. BDNF NT-4 heterodimer structural analysis to highlight coevolving and adaptive sites at the 3D protein level. Demonstrating the structural configuration of the BDNF (blue) and NT-4 (pink) heterodimer (see https://www.rcsb.org/structure/1B8M), with rotations (arbitrary degrees) shown to accentuate the view of coevolving sites (orange, see also Fig. 4.2 and Table 4.2) and positively evolving sites (red; see Table 4.1). The PDB structure is limited to the NGF domain which limits our ability to highlight sites of interest (SOI), therefore we have limited our annotation only to the modeled sites in the structure. The relative positioning of coevolving and positively evolving sites in this heterodimer visualization are in proximity to looping and other macro tertiary structures of protein. An interactive figure that is rotatable in 3D space, where occupations occur in three dimensions (i.e., teasing out relative proximity in 2D linear space), is available here: https://observablehq.com/@aglucaci/bdnf-structure.

Table 4.2. The BGM analysis of BDNF found 23 pairs of coevolving sites out of 261 total sites to be statistically significant (with a posterior probability threshold of 0.5).

| # | Site 1 | Site 2 | P [Site 1 –> Site 2] | P [Site 2 –> Site 1] | P [Site 1 < –> Site 2] | Site 1 subs | Site 2 subs | Shared subs |
|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 67 | 0.52 | 0.32 | 0.84 | 1 | 1 | 1 |
| 2 | 25 | 49 | 0.022 | 0.67 | 0.69 | 8 | 6 | 3 |
| 3 | 27 | 109 | 0.12 | 0.45 | 0.57 | 2 | 1 | 1 |
| 4 | 30 | 119 | 0.2 | 0.56 | 0.75 | 6 | 12 | 4 |
| 5 | 33 | 119 | 0.12 | 0.72 | 0.84 | 3 | 12 | 3 |
| 6 | 39 | 103 | 0.22 | 0.58 | 0.8 | 8 | 4 | 3 |
| 7 | 49 | 85 | 0.48 | 0.049 | 0.53 | 6 | 3 | 2 |
| 8 | 49 | 86 | 0.36 | 0.55 | 0.91 | 6 | 5 | 3 |
| 9 | 50 | 69 | 0.31 | 0.27 | 0.58 | 1 | 2 | 1 |
| 10 | 62 | 64 | 0.25 | 0.27 | 0.52 | 1 | 1 | 1 |
| 11 | 70 | 74 | 0.17 | 0.38 | 0.55 | 7 | 14 | 4 |
| 12 | 70 | 95 | 0.59 | 0.26 | 0.85 | 7 | 21 | 5 |
| 13 | 74 | 94 | 0.034 | 0.64 | 0.68 | 14 | 7 | 4 |
| 14 | 78 | 90 | 0.46 | 0.38 | 0.84 | 1 | 1 | 1 |
| 15 | 81 | 93 | 0.27 | 0.31 | 0.58 | 1 | 1 | 1 |
| 16 | 81 | 98 | 0.27 | 0.33 | 0.61 | 1 | 1 | 1 |
| 17 | 82 | 91 | 0.045 | 0.92 | 0.96 | 15 | 29 | 9 |
| 18 | 89 | 184 | 0.24 | 0.5 | 0.74 | 2 | 6 | 2 |
| 19 | 91 | 119 | 0.22 | 0.67 | 0.89 | 29 | 12 | 8 |
| 20 | 93 | 98 | 0.31 | 0.3 | 0.61 | 1 | 1 | 1 |
| 21 | 94 | 155 | 0.28 | 0.29 | 0.57 | 7 | 2 | 2 |
| 22 | 103 | 233 | 0.4 | 0.13 | 0.53 | 4 | 3 | 2 |
| 23 | 135 | 154 | 0.4 | 0.43 | 0.82 | 1 | 1 | 1 |

## Discussion

In this study, we explore the evolutionary history of the BDNF gene in Mammalia. The BDNF gene is implicated in a number of human diseases including a variety of brain disorders such as neurodevelopmental disorders (e.g., autism spectrum disorder), neuropsychiatric disorders (e.g., depression, PTSD, and schizophrenia), and some neurodegenerative disorders (Notaras and van den Buuse, 2015). By using orthologous

BDNF sequences within the Mammalia taxonomic group, our results indicate that unique site-specific changes within BDNF have evolved over time. We performed a number of comparative evolutionary analyses to tease out signals from our orthologous gene collection in BDNF. Of note, the BDNF gene elicits tight regulation and specific functionality that can be separated from other neurotrophins, yet these growth factors remain closely related in their structure and sequence and conservation of the NGF functional domain. In the NGF domain, we observe a high degree of conservation (via purifying selection) across species, owing to the functional importance of this region in protein–protein interactions. This work additionally provides broad comparative insights into the evolutionary history of the BDNF gene family. Our MEME method identified novel substitutions (see Table 4.1) in regions of BDNF that may provide significant areas of interest for designing molecular therapeutic approaches, and their potential broader significance are outlined in further detail below.

***Predominant Purifying Selection Across BDNF in Mammalia***

Over time, evolution drives the divergence of genetic sequences, but what can we learn from the direct comparison of the sequences of the BDNF gene in Mammalia? By comparing the BDNF products of orthologous sequences in different species, we observe the accumulation of mutations at different sites with varying degrees of insight into both BDNF functionality (see (Notaras and van den Buuse, 2019) for site annotation) and potential disease (Notaras and van den Buuse, 2015). These are summarized in full within Table C.S1 and Table 4.1. Coding sequences with highly constrained structures are

expected to fix nonsynonymous mutations at a slower rate due to the maladaptive nature of changes such as what we observe with FEL negatively selected sites across BDNF. Additionally, we observe a high degree of negative (purifying) selection across the main functional domain (NGF) of BDNF. While structures for the NGF domain in most species under analysis do not exist, based on our findings we expect a highly conserved tertiary structure. Based on the high degree of purifying selection observed across BDNF, we hypothesize that BDNF plays a critical role in the underlying network of genes governing homeostasis and normal organismal development. This may have happened because BDNF is particularly useful specifically for the phylogenetic branch in question (i.e., mammals). This interpretation is also consistent with the observation that BDNF is essential for normative development and is lethal in non-conditional full knock-out mammalian models.

***Non-neutral Positive Diversifying Evolution Sites in the BDNF Gene***

It has been described that BDNF plays a particular role as a foundational gene for brain development (Kim et al, 2007). Despite a significant level of purifying selection shaping the evolutionary history of BDNF (Fig. 4.1), we observe several novel statistically significant sites under positive episodic diversifying selection across the BDNF gene (see Table 4.1). Traditionally, the evolution of this variety consists of amino acid diversifying events that may promote phylogenetic adaptation and/or functionality. These results are entirely novel—they have not been previously reported (to the best of our knowledge) and MEME is an established and sensitive method for the analysis of episodic diversifying sites. Thus, the very specific and limited sites within BDNF to exhibit such patterns is a

highly promising result from which to further disentangle BDNF's complex functionality and disease linkage. We would encourage biologists to consider these sites as those that may contain important adaptive functions within the BDNF gene. However, where our results fall within the context of a core protein–protein interaction network of required genes for neural cellular diversity and development is yet to be determined. We do note that at least one identified site (238) overlaps with potential posttranslational modifications to the human BDNF peptide (specifically, a disulfide bridge; see UniProt and (Notaras and van den Buuse, 2019). This supports the idea that non-neutral positive diversifying sites within BDNF are not spurious and likely reflect specialized, regulatory, or functional capacities that may have yet to be annotated in full. Given that this manuscript is devoted to the analysis of BDNF's evolution in mammals, we highlight the potential importance of these sites but emphasize that their importance remains a hypothesis that should be tested in well-defined experiments under controlled laboratory conditions.

***Discovery of Proximal and Distal Coevolving Site-Pairs in the BDNF Gene***

Another novel, and potentially important, series of findings in this manuscript was the presence of numerous sites that exhibit coevolution. In fact, we observe a significant number of coevolving sites within the BDNF gene (see Fig. 4.2 and Table 4.2), and these too reflect an entirely novel aspect of BDNF biology that has not previously been reported. Evidence of coevolving sites are not limited to a particular domain (e.g., prodomain vs mature) nor specific motifs. Instead, coevolving pairs seem to be distributed across the entirety of BDNF with, perhaps unsurprisingly, an increased density of interactions early

in the prodomain region. However, we also note that there are coevolving sites in the mature NGF domain which are "linked" to early domain sites. Importantly, these relationships may confer strong epistatic interactions shaping the continued evolution of this critically important gene. The new evidence for coevolution may point to the importance of these sites in shaping the early regulatory or main functional (NGF) domain of BDNF. These residues may form important interactions for the functional integrity of BDNF and, importantly, the highly specific pairs which span the BDNF prodomain and its mature region point to a new mechanism by which the BDNF prodomain may have coregulated the mature domain (or vice versa). Alternatively, these coevolving pairs may be part of a network of residues occupying a shifted fitness landscape in order to accommodate new or species-specific functional requirements.

### *Potential Structural Implications of Evolving Sites*

In considering our observation of both diversifying selection and coevolving sites in the BDNF gene, we considered the potential implications this may have at a protein structural level in three-dimensional space (see Fig. 4.3). While protein structural impacts from evolution remain poorly understood and cannot be completely experimentally disentangled in a confirmatory sense, the implications fall upon our understanding of basic BDNF neurobiology. Here we note that our BGM and FEL analyses implicate the prodomain—the primary topological region of BDNF known for polymorphic variability (e.g., Val66Met and Gln75His) that is often linked to disease (Notaras and van den Buuse, 2015; Notaras and van den Buuse, 2020; Notaras and van den Buuse, 2019), and our 3D

modeling suggests that two of our coevolving sites appear to be associated with looping structures that could have important yet to be discovered functionality. In this regard, we predict that the evolutionary changes described here are likely to reflect some form of specialization and/or divergence in function and/or interaction partners at different points of BDNF's evolutionary history in mammals. Thus, further work may unveil yet more novel sites that could provide further insight into the origins of BDNF's diverse functionality and its role in disease.

***Limitations of our Computational Evolutionary Analysis***

This analysis focused on BDNF sequences contained in the taxonomic group Mammalia in lieu of examining a more inclusive dataset for BDNF containing sequences from all of Gnathostomata (jawed vertebrates) or extension into invertebrate clades which may contain BDNF or BDNF-like analog genes. Our results are applicable to mammals, which are our intentional taxonomic group of study, but we nonetheless emphasize that our results do not capture the entirety of BDNF's evolutionary history (e.g., there could be more to learn about BDNF from birds, lizards, fish, and higher-order taxonomic groups which we do not evaluate here). In addition, we do not explore the patterns or mutational processes occurring outside of coding-sequence evolution which include complex structure and dynamics of non-coding regions in the BDNF gene. Therefore, evolutionary temporality is important in the context and interpretation of our results because Mammalia represents only a portion of the long evolutionary history of BDNF. Although we failed to find evidence for recombination in our dataset, species where we may find evidence for

recombination may have been precluded from our analysis due to our decision to focus on mammalian BDNF evolution. Further, a limitation of the current analysis is owed to the presence of indel events, especially in the early region of the alignment but which also occur in other spatially distributed regions of the BDNF gene. These indel events are not currently modeled in existing codon substitution models but may represent an additional pathway of evolutionary change. Nonetheless, the prominence of indels in our observations indicates that several regions of BDNF may evolve significantly through indel events across species. Lastly, although there is a risk that the "gappy" nature of the early region of our multiple sequence alignment may be a computational artifact of the alignment procedure, based on all other outputs we believe that our results are reasonably interpreted and have subsequently tolerated these potential effects.

***Future Directions: Understanding the Remainder of the Neurotrophin Family***

We hypothesize that the similarities between neurotrophins reflects conserved evolutionary selection for motifs and domains which support common functionality in neurotrophic factors between sites and lineages. While we note significant isotropy in mature peptide sequences for these factors, anisotropic pressures likely influenced the prodomain sequences of neurotrophins leading to alterations in processing, trafficking, regulation, and secretion. As such, we also predict differences in the evolutionary fate of other neurotrophins which also exhibit compartmentalized functionality due to similar alterations within their prodomains (i.e., similar results may be reasonably anticipated in NGF, NT-3, and NT-4).

## Conclusion

To sum up, our research modeled the natural evolutionary history of changes in BDNF across >160 mammalian genomes. Conservatively, this analysis spans ~177 million years of evolution—and going deeper could yet reveal more information on the ontogenesis of BDNF and its topological structure (and, consequently, function). Notably, we observed strict purifying selection in the main functional domain of the BDNF gene in mammals and discovered 6 specific sites in our homologous alignment that are under episodic selection in the early regulatory region of BDNF (i.e., the prodomain) and in the terminal region of BDNF. We also make the case for spatial coevolution within this gene, with 23 pair-sites that have evolved together. In sum, these data go above and beyond the common trope that "BDNF is highly conserved" by defining exactly where and how the mammalian BDNF has evolved. Thus, we confirm the widespread belief that the BDNF prodomain is more prone to change than the mature BDNF protein, having important implications for how we think about and consider genetic variation in BDNF and its linkage to disease.

## Methods

### *Data Retrieval*

For this study, we queried the NCBI Ortholog database via https://www.ncbi.nlm.nih.gov/kis/ortholog/627/?scope=7776. For the purpose of this study, as we are interested in mammalian BDNF evolution, we limited our search to only include species from this taxonomic group (mammals, Mammalia). This returned 162 full

gene transcripts and protein sequences. We downloaded all available files: RefSeq protein sequences, RefSeq transcript sequences, Tabular data (CSV, metadata). In Table 4.3, we provide a table of the species included in this analysis, but we also make this accessible via GitHub. Furthermore, we also make these species NCBI accessions (see also Table 4.3) available for download on GitHub: https://github.com/aglucaci/AnalysisOfOrthologousCollections/blob/main/data/BDNF/BDNF_orthologs.csv

Table 4.3. Tabulation of species included in our analysis, comprising NCBI ortholog gene IDs, symbols, mammalian species, common name, and RefSeq accessions.

| **Gene ID** | **Gene symbol** | **Description** | **Scientific name** | **Common name** | **RefSeq Transcript accessions** | **RefSeq Protein accessions** |
|---|---|---|---|---|---|---|
| 627 | BDNF | brain derived neurotrophic factor | Homo sapiens | human | NM_001709.5 | NP_001700.2 |
| 12064 | Bdnf | brain derived neurotrophic factor | Mus musculus | house mouse | NM_001048142.1 | NP_001041607.1 |
| 24225 | Bdnf | brain-derived neurotrophic factor | Rattus norvegicus | Norway rat | NM_001270630.1 | NP_001257559.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 397495 | BDNF | brain derived neurotrophic factor | Sus scrofa | pig | XM_005654684.3 | XP_005654741.1 |
| 403461 | BDNF | brain derived neurotrophic factor | Canis lupus familiaris | dog | XM_038429434.1 | XP_038285362.1 |
| 493690 | BDNF | brain derived neurotrophic factor | Felis catus | domestic cat | NM_001009828.1 | NP_001009828.1 |
| 503511 | BDNF | brain derived neurotrophic factor | Pan troglodytes | chimpanzee | NM_001012441.1 | NP_001012443.1 |
| 554233 | BDNF | brain derived neurotrophic factor | Monodelphis domestica | gray short-tailed opossum | XM_007497196.2 | XP_007497258.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 617701 | BDNF | brain derived neurotrophic factor | Bos taurus | cattle | XM_005216334.4 | XP_005216391.1 |
| 701245 | BDNF | brain derived neurotrophic factor | Macaca mulatta | Rhesus monkey | XM_015114598.2 | XP_014970084.1 |
| 100009689 | BDNF | brain derived neurotrophic factor | Equus caballus | horse | NM_001081787.1 | NP_001075256.1 |
| 100081142 | BDNF | brain derived neurotrophic factor | Ornithorhynchus anatinus | platypus | XM_029059317.2 | XP_028915150.1 |
| 100356949 | BDNF | brain derived neurotrophic factor | Oryctolagus cuniculus | rabbit | XM_017345633.1 | XP_017201122.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 100409412 | BDNF | brain derived neurotrophic factor | Callithrix jacchus | white-tufted-ear marmoset | XM_009007854.3 | XP_009006102.1 |
| 100447350 | BDNF | brain derived neurotrophic factor | Pongo abelii | Sumatran orangutan | XM_002821931.2 | XP_002821977.1 |
| 100467162 | BDNF | brain derived neurotrophic factor | Ailuropoda melanoleuca | giant panda | XM_011226480.3 | XP_011224782.2 |
| 100594402 | BDNF | brain derived neurotrophic factor | Nomascus leucogenys | northern white-cheeked gibbon | XM_003254347.2 | XP_003254395.1 |
| 100667885 | BDNF | brain derived neurotrophic factor | Loxodonta africana | African savanna elephant | XM_023550772.1 | XP_023406540.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 100730257 | Bdnf | brain derived neurotrophic factor | Cavia porcellus | domestic guinea pig | XM_013147262.2 | XP_013002716.1 |
| 100768664 | Bdnf | brain derived neurotrophic factor | Cricetulus griseus | Chinese hamster | XM_007653166.4 | XP_007651356.1 |
| 100934810 | BDNF | brain derived neurotrophic factor | Sarcophilus harrisii | Tasmanian devil | XM_031944159.1 | XP_031800019.1 |
| 100958946 | BDNF | brain derived neurotrophic factor | Otolemur garnettii | small-eared galago | XM_012814047.1 | XP_012669501.1 |
| 100983866 | BDNF | brain derived neurotrophic factor | Pan paniscus | pygmy chimpanzee | XM_034932318.1 | XP_034788209.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 101007866 | BDNF | brain derived neurotrophic factor | Papio anubis | olive baboon | XM_017948537.3 | XP_017804026.1 |
| 101037414 | BDNF | brain derived neurotrophic factor | Saimiri boliviensis | Bolivian squirrel monkey | XM_003919940.3 | XP_003919989.2 |
| 101117275 | BDNF | brain derived neurotrophic factor | Ovis aries | sheep | XM_012096129.4 | XP_011951519.2 |
| 101134399 | BDNF | brain derived neurotrophic factor | Gorilla gorilla | western gorilla | XM_004050851.2 | XP_004050899.1 |
| 101281702 | BDNF | brain derived neurotrophic factor | Orcinus orca | killer whale | XM_004263941.3 | XP_004263989.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 101338565 | BDNF | brain derived neurotrophic factor | Tursiops truncatus | common bottlenose dolphin | XM_019947883.2 | XP_019803442.1 |
| 101350169 | LOC101350169 | brain derived neurotrophic factor | Trichechus manatus latirostris | Florida manatee | XM_004369733.1 | XP_004369790.1 |
| 101372507 | BDNF | brain derived neurotrophic factor | Odobenus rosmarus divergens | Pacific walrus | XM_004408653.1 | XP_004408710.1 |
| 101398458 | LOC101398458 | brain derived neurotrophic factor | Ceratotherium simum simum | southern white rhinoceros | XM_004418542.2 | XP_004418599.1 |
| 101428552 | BDNF | brain derived neurotrophic factor | Dasypus novemcinctus | nine-banded armadillo | XM_012523604.2 | XP_012379058.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 101528222 | BDNF | brain derived neurotrophic factor | Ochotona princeps | American pika | XM_004585440.1 | XP_004585497.1 |
| 101553155 | BDNF | brain derived neurotrophic factor | Sorex araneus | European shrew | XM_004607860.1 | XP_004607917.1 |
| 101566518 | Bdnf | brain derived neurotrophic factor | Octodon degus | degu | XM_004644113.2 | XP_004644170.1 |
| 101600646 | Bdnf | brain derived neurotrophic factor | Jaculus jaculus | lesser Egyptian jerboa | XM_004660846.2 | XP_004660903.1 |
| 101632951 | BDNF | brain derived neurotrophic factor | Condylura cristata | star-nosed mole | XM_004682964.2 | XP_004683021.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 101643198 | BDNF | brain derived neurotrophic factor | Echinops telfairi | small Madagascar hedgehog | XM_004708253.2 | XP_004708310.1 |
| 101687038 | BDNF | brain derived neurotrophic factor | Mustela putorius furo | domestic ferret | XM_004755891.2 | XP_004755948.1 |
| 101704465 | Bdnf | brain derived neurotrophic factor | Heterocephalus glaber | naked mole-rat | XM_004851581.3 | XP_004851638.1 |
| 101837384 | Bdnf | brain derived neurotrophic factor | Mesocricetus auratus | golden hamster | XM_005064810.4 | XP_005064867.1 |
| 101954451 | Bdnf | brain derived neurotrophic factor | Ictidomys tridecemlineatus | thirteen-lined ground squirrel | XM_040284582.1 | XP_040140516.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 101993181 | Bdnf | brain derived neurotrophic factor | Microtus ochrogaster | prairie vole | XM_005364052.2 | XP_005364109.1 |
| 102016703 | Bdnf | brain derived neurotrophic factor | Chinchilla lanigera | long-tailed chinchilla | XM_005401751.2 | XP_005401808.1 |
| 102126294 | BDNF | brain derived neurotrophic factor | Macaca fascicularis | crab-eating macaque | XM_005578346.2 | XP_005578403.1 |
| 102180782 | BDNF | brain derived neurotrophic factor | Capra hircus | goat | XM_005690025.2 | XP_005690082.2 |
| 102247090 | BDNF | brain derived neurotrophic factor | Myotis brandtii | Brandt's bat | XM_014545223.1 | XP_014400709.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 102279714 | BDNF | brain derived neurotrophic factor | Bos mutus | wild yak | XM_005890667.2 | XP_005890729.1 |
| 102395046 | BDNF | brain derived neurotrophic factor | Bubalus bubalis | water buffalo | XM_006068006.2 | XP_006068068.1 |
| 102440705 | BDNF | brain derived neurotrophic factor | Myotis lucifugus | little brown bat | XM_023756374.1 | XP_023612142.1 |
| 102475289 | BDNF | brain derived neurotrophic factor | Tupaia chinensis | Chinese tree shrew | XM_014591303.2 | XP_014446789.1 |
| 102511705 | BDNF | brain derived neurotrophic factor | Camelus ferus | Wild Bactrian camel | XM_006195083.3 | XP_006195145.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 102532564 | BDNF | brain derived neurotrophic factor | Vicugna pacos | alpaca | XM_031681074.1 | XP_031536934.1 |
| 102730927 | BDNF | brain derived neurotrophic factor | Leptonychotes weddellii | Weddell seal | XM_031025064.1 | XP_030880924.1 |
| 102754946 | BDNF | brain derived neurotrophic factor | Myotis davidii | | XM_006753948.2 | XP_006754011.1 |
| 102839210 | BDNF | brain derived neurotrophic factor | Chrysochloris asiatica | Cape golden mole | XM_006864851.1 | XP_006864913.1 |
| 102870215 | BDNF | brain derived neurotrophic factor | Elephantulus edwardii | Cape elephant shrew | XM_006883713.1 | XP_006883775.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 102896087 | BDNF | brain derived neurotrophic factor | Pteropus alecto | black flying fox | XM_006907988.2 | XP_006908050.1 |
| 102911929 | Bdnf | brain derived neurotrophic factor | Peromyscus maniculatus bairdii | prairie deer mouse | XM_006980818.2 | XP_006980880.1 |
| 102958723 | BDNF | brain derived neurotrophic factor | Panthera tigris altaica | Amur tiger | XM_015544869.1 | XP_015400355.1 |
| 102977934 | BDNF | brain derived neurotrophic factor | Physeter catodon | sperm whale | XM_028501134.1 | XP_028356935.1 |
| 103017938 | BDNF | brain derived neurotrophic factor | Balaenoptera acutorostrata scammoni | | XM_007180991.1 | XP_007181053.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 103085069 | BDNF | brain derived neurotrophic factor | Lipotes vexillifer | Yangtze River dolphin | XM_007461679.1 | XP_007461741.1 |
| 103126099 | BDNF | brain derived neurotrophic factor | Erinaceus europaeus | western European hedgehog | XM_016194241.1 | XP_016049727.1 |
| 103206496 | BDNF | brain derived neurotrophic factor | Orycteropus afer afer | | XM_007951952.1 | XP_007950143.1 |
| 103238537 | BDNF | brain derived neurotrophic factor | Chlorocebus sabaeus | green monkey | XM_008003703.2 | XP_008001894.1 |
| 103252117 | BDNF | brain derived neurotrophic factor | Carlito syrichta | Philippine tarsier | XM_008050727.1 | XP_008048918.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 103292699 | BDNF | brain derived neurotrophic factor | Eptesicus fuscus | big brown bat | XM_008148947.2 | XP_008147169.1 |
| 103552694 | BDNF | brain derived neurotrophic factor | Equus przewalskii | Przewalski's horse | XM_008522855.1 | XP_008521077.1 |
| 103599960 | BDNF | brain derived neurotrophic factor | Galeopterus variegatus | Sunda flying lemur | XM_008584165.1 | XP_008582387.1 |
| 103659367 | BDNF | brain derived neurotrophic factor | Ursus maritimus | polar bear | XM_008686880.2 | XP_008685102.2 |
| 103748489 | Bdnf | brain derived neurotrophic factor | Nannospalax galili | Upper Galilee mountains blind mole rat | XM_029555090.1 | XP_029410950.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 104671126 | BDNF | brain derived neurotrophic factor | Rhinopithecus roxellana | golden snub-nosed monkey | XM_010374529.2 | XP_010372831.1 |
| 104873642 | Bdnf | brain derived neurotrophic factor | Fukomys damarensis | Damara mole-rat | XM_033764060.1 | XP_033619951.1 |
| 104982427 | BDNF | brain derived neurotrophic factor | Bison bison bison | | XM_010831788.1 | XP_010830090.1 |
| 105076551 | BDNF | brain derived neurotrophic factor | Camelus bactrianus | Bactrian camel | XM_010964743.1 | XP_010963045.1 |
| 105106448 | BDNF | brain derived neurotrophic factor | Camelus dromedarius | Arabian camel | XM_031459858.1 | XP_031315718.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 105293721 | BDNF | brain derived neurotrophic factor | Pteropus vampyrus | large flying fox | XM_011362565.2 | XP_011360867.1 |
| 105463281 | BDNF | brain derived neurotrophic factor | Macaca nemestrina | pig-tailed macaque | XM_011710310.2 | XP_011708612.1 |
| 105507710 | BDNF | brain derived neurotrophic factor | Colobus angolensis palliatus | | XM_011936307.1 | XP_011791697.1 |
| 105528120 | BDNF | brain derived neurotrophic factor | Mandrillus leucophaeus | drill | XM_011964715.1 | XP_011820105.1 |
| 105572781 | BDNF | brain derived neurotrophic factor | Cercocebus atys | sooty mangabey | XM_012031697.1 | XP_011887087.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 105717219 | BDNF | brain derived neurotrophic factor | Aotus nancymaae | Ma's night monkey | XM_012452238.1 | XP_012307661.1 |
| 105814556 | BDNF | brain derived neurotrophic factor | Propithecus coquereli | Coquerel's sifaka | XM_012649556.1 | XP_012505010.1 |
| 105860148 | BDNF | brain derived neurotrophic factor | Microcebus murinus | gray mouse lemur | XM_020286534.1 | XP_020142123.1 |
| 105988089 | Bdnf | brain derived neurotrophic factor | Dipodomys ordii | Ord's kangaroo rat | XM_013019592.1 | XP_012875046.1 |
| 106824927 | BDNF | brain derived neurotrophic factor | Equus asinus | ass | XM_014831462.1 | XP_014686948.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 106970278 | BDNF | brain derived neurotrophic factor | Acinonyx jubatus | cheetah | XM_027072960.1 | XP_026928761.1 |
| 107134338 | Bdnf | brain derived neurotrophic factor | Marmota marmota marmota | Alpine marmot | XM_015476449.1 | XP_015331935.1 |
| 107512170 | BDNF | brain derived neurotrophic factor | Rousettus aegyptiacus | Egyptian rousette | XM_036221046.1 | XP_036076939.1 |
| 107531942 | BDNF | brain derived neurotrophic factor | Miniopterus natalensis | | XM_016206154.1 | XP_016061640.1 |
| 108311306 | BDNF | brain derived neurotrophic factor | Cebus imitator | Panamanian white-faced capuchin | XM_017538722.1 | XP_017394211.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 108401589 | BDNF | brain derived neurotrophic factor | Manis javanica | Malayan pangolin | XM_017667612.2 | XP_017523101.2 |
| 108519651 | BDNF | brain derived neurotrophic factor | Rhinopithecus bieti | black snub-nosed monkey | XM_017858805.1 | XP_017714294.1 |
| 109271282 | BDNF | brain derived neurotrophic factor | Panthera pardus | leopard | XM_019456577.1 | XP_019312122.1 |
| 109387036 | BDNF | brain derived neurotrophic factor | Hipposideros armiger | great roundleaf bat | XM_019650718.1 | XP_019506263.1 |
| 109569212 | BDNF | brain derived neurotrophic factor | Bos indicus | zebu cattle | XM_019974522.1 | XP_019830081.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 109695020 | Bdnf | brain derived neurotrophic factor | Castor canadensis | American beaver | XM_020177315.1 | XP_020032904.1 |
| 110143608 | BDNF | brain derived neurotrophic factor | Odocoileus virginianus texanus | | XM_020903256.1 | XP_020758915.1 |
| 110202027 | BDNF | brain derived neurotrophic factor | Phascolarctos cinereus | koala | XM_020977976.1 | XP_020833635.1 |
| 110285237 | Bdnf | brain derived neurotrophic factor | Mus caroli | Ryukyu mouse | XM_021151236.2 | XP_021006895.1 |
| 110318089 | Bdnf | brain derived neurotrophic factor | Mus pahari | shrew mouse | XM_021192829.2 | XP_021048488.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 110576669 | BDNF | brain derived neurotrophic factor | Neomonachus schauinslandi | Hawaiian monk seal | XM_021685880.1 | XP_021541555.1 |
| 111150176 | LOC111150176 | brain derived neurotrophic factor | Enhydra lutris kenyoni | | XM_022507574.1 | XP_022363282.1 |
| 111179820 | BDNF | brain derived neurotrophic factor | Delphinapterus leucas | beluga whale | XM_022583733.1 | XP_022439441.1 |
| 111554690 | BDNF | brain derived neurotrophic factor | Piliocolobus tephrosceles | Ugandan red Colobus | XM_023230347.2 | XP_023086115.1 |
| 112303343 | BDNF | brain derived neurotrophic factor | Desmodus rotundus | common vampire bat | XM_024558903.1 | XP_024414671.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 112412908 | BDNF | brain derived neurotrophic factor | Neophocaena asiaeorientalis asiaeorientalis | Yangtze finless porpoise | XM_024764694.1 | XP_024620462.1 |
| 112607193 | BDNF | brain derived neurotrophic factor | Theropithecus gelada | gelada | XM_025358305.1 | XP_025214090.1 |
| 112667789 | BDNF | brain derived neurotrophic factor | Canis lupus dingo | dingo | XM_025459816.2 | XP_025315601.1 |
| 112829628 | BDNF | brain derived neurotrophic factor | Callorhinus ursinus | northern fur seal | XM_025879326.1 | XP_025735111.1 |
| 112865997 | BDNF | brain derived neurotrophic factor | Puma concolor | puma | XM_025929035.1 | XP_025784820.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 112935019 | BDNF | brain derived neurotrophic factor | Vulpes vulpes | red fox | XM_026018600.1 | XP_025874385.1 |
| 113199355 | Bdnf | brain derived neurotrophic factor | Urocitellus parryii | Arctic ground squirrel | XM_026412272.1 | XP_026268057.1 |
| 113265088 | BDNF | brain derived neurotrophic factor | Ursus arctos horribilis | | XM_026512203.1 | XP_026367988.1 |
| 113624889 | BDNF | brain derived neurotrophic factor | Lagenorhynchus obliquidens | Pacific white-sided dolphin | XM_027115138.1 | XP_026970939.1 |
| 113905455 | BDNF | brain derived neurotrophic factor | Bos indicus x Bos taurus | hybrid cattle | XM_027562786.1 | XP_027418587.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 113913517 | BDNF | brain derived neurotrophic factor | Zalophus californianus | California sea lion | XM_027577807.1 | XP_027433608.1 |
| 114036167 | BDNF | brain derived neurotrophic factor | Vombatus ursinus | common wombat | XM_027852578.1 | XP_027708379.1 |
| 114093712 | Bdnf | brain derived neurotrophic factor | Marmota flaviventris | yellow-bellied marmot | XM_027936596.1 | XP_027792397.1 |
| 114220007 | BDNF | brain derived neurotrophic factor | Eumetopias jubatus | Steller sea lion | XM_028117708.1 | XP_027973509.1 |
| 114500881 | BDNF | brain derived neurotrophic factor | Phyllostomus discolor | pale spear-nosed bat | XM_028517638.2 | XP_028373439.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 114620789 | Bdnf | brain derived neurotrophic factor | Grammomys surdaster | | XM_028766921.1 | XP_028622754.1 |
| 114702098 | Bdnf | brain derived neurotrophic factor | Peromyscus leucopus | white-footed mouse | XM_037204438.1 | XP_037060333.1 |
| 114886864 | BDNF | brain derived neurotrophic factor | Monodon monoceros | narwhal | XM_029207934.1 | XP_029063767.1 |
| 115306639 | BDNF | brain derived neurotrophic factor | Suricata suricatta | meerkat | XM_029957152.1 | XP_029813012.1 |
| 115525821 | BDNF | brain derived neurotrophic factor | Lynx canadensis | Canada lynx | XM_030333208.1 | XP_030189068.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 115857612 | BDNF | brain derived neurotrophic factor | Globicephala melas | long-finned pilot whale | XM_030864833.1 | XP_030720693.1 |
| 116090620 | Bdnf | brain derived neurotrophic factor | Mastomys coucha | southern multimammate mouse | XM_031371313.1 | XP_031227173.1 |
| 116476405 | BDNF | brain derived neurotrophic factor | Hylobates moloch | silvery gibbon | XM_032166370.1 | XP_032022261.1 |
| 116555191 | BDNF | brain derived neurotrophic factor | Sapajus apella | tufted capuchin | XM_032283691.1 | XP_032139582.1 |
| 116599082 | BDNF | brain derived neurotrophic factor | Mustela erminea | ermine | XM_032358760.1 | XP_032214651.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 116638304 | BDNF | brain derived neurotrophic factor | Phoca vitulina | harbor seal | XM_032414183.1 | XP_032270074.1 |
| 116758559 | BDNF | brain derived neurotrophic factor | Phocoena sinus | vaquita | XM_032641481.1 | XP_032497372.1 |
| 116881443 | BDNF | brain derived neurotrophic factor | Lontra canadensis | Northern American river otter | XM_032881246.1 | XP_032737137.1 |
| 116901726 | Bdnf | brain derived neurotrophic factor | Rattus rattus | black rat | XM_032903861.1 | XP_032759752.1 |
| 117029758 | BDNF | brain derived neurotrophic factor | Rhinolophus ferrumequinum | greater horseshoe bat | XM_033118944.1 | XP_032974835.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 117080419 | BDNF | brain derived neurotrophic factor | Trachypithecus francoisi | Francois's langur | XM_033205475.1 | XP_033061366.1 |
| 117704234 | Bdnf | brain derived neurotrophic factor | Arvicanthis niloticus | African grass rat | XM_034496309.1 | XP_034352200.1 |
| 118015716 | BDNF | brain derived neurotrophic factor | Mirounga leonina | Southern elephant seal | XM_035012477.1 | XP_034868368.1 |
| 118546649 | BDNF | brain derived neurotrophic factor | Halichoerus grypus | gray seal | XM_036109493.1 | XP_035965386.1 |
| 118582264 | Bdnf | brain derived neurotrophic factor | Onychomys torridus | southern grasshopper mouse | XM_036184983.1 | XP_036040876.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 118628031 | BDNF | brain derived neurotrophic factor | Molossus molossus | Pallas's mastiff bat | XM_036258707.1 | XP_036114600.1 |
| 118664943 | BDNF | brain derived neurotrophic factor | Myotis myotis | | XM_036325692.1 | XP_036181585.1 |
| 118713266 | BDNF | brain derived neurotrophic factor | Pipistrellus kuhlii | Kuhl's pipistrelle | XM_036427676.1 | XP_036283569.1 |
| 118853508 | BDNF | brain derived neurotrophic factor | Trichosurus vulpecula | common brushtail | XM_036763631.1 | XP_036619526.1 |
| 118900201 | BDNF | brain derived neurotrophic factor | Balaenoptera musculus | Blue whale | XM_036862552.1 | XP_036718447.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 118909009 | BDNF | brain derived neurotrophic factor | Manis pentadactyla | Chinese pangolin | XM_036879079.1 | XP_036734974.1 |
| 118983123 | BDNF | brain derived neurotrophic factor | Sturnira hondurensis | | XM_037040421.1 | XP_036896316.1 |
| 119049902 | BDNF | brain derived neurotrophic factor | Artibeus jamaicensis | Jamaican fruit-eating bat | XM_037146035.1 | XP_037001930.1 |
| 119255407 | BDNF | brain derived neurotrophic factor | Talpa occidentalis | Iberian mole | XM_037522538.1 | XP_037378435.1 |
| 119536361 | BDNF | brain derived neurotrophic factor | Choloepus didactylus | southern two-toed sloth | XM_037839114.1 | XP_037695042.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 119815449 | Bdnf | brain derived neurotrophic factor | Arvicola amphibius | Eurasian water vole | XM_038331579.1 | XP_038187507.1 |
| 119943514 | BDNF | brain derived neurotrophic factor | Tachyglossus aculeatus | Australian echidna | XM_038764545.1 | XP_038620473.1 |
| 120220096 | BDNF | brain derived neurotrophic factor | Hyaena hyaena | striped hyena | XM_039217098.1 | XP_039073029.1 |
| 120605221 | BDNF | brain derived neurotrophic factor | Pteropus giganteus | Indian flying fox | XM_039866065.1 | XP_039721999.1 |
| 120876831 | BDNF | brain derived neurotrophic factor | Oryx dammah | scimitar-horned oryx | XM_040258919.1 | XP_040114853.1 |

Table 4.3 (Continued)

| Gene ID | Gene symbol | Description | Scientific name | Common name | RefSeq Transcript accessions | RefSeq Protein accessions |
|---|---|---|---|---|---|---|
| 121043625 | BDNF | brain derived neurotrophic factor | Puma yagouaroundi | jaguarundi | XM_040495572.1 | XP_040351506.1 |
| 121156520 | BDNF | brain derived neurotrophic factor | Ochotona curzoniae | black-lipped pika | XM_040980297.1 | XP_040836231.1 |
| 121465693 | Bdnf | brain derived neurotrophic factor | Microtus oregoni | creeping vole | XM_041679035.1 | XP_041534969.1 |
| 121476611 | BDNF | brain derived neurotrophic factor | Vulpes lagopus | Arctic fox | XM_041730708.1 | XP_041586642.1 |

### *Data Cleaning*

We used the protein sequence and full gene transcripts to derive coding sequences (CDS) (via a custom script, “scripts/codons.py”). However, this process was met with

errors in 20 “PREDICTED” protein sequences, which had invalid characters such as sequences, which have incorrect “X”, or unresolved amino acids and these sequences were subsequently exempt from the analysis. This process removes low-quality protein sequences from analysis which may inflate rates of nonsynonymous change.

***Analysis of Orthologous Collections (AOC): Alignment, Recombination detection, Tree Inference, and Selection Algorithms***

The analysis of orthologous collections (AOC) application is designed for comprehensive protein-coding molecular sequence analysis (https://github.com/aglucaci/AnalysisOfOrthologousCollections). It accomplishes this through a series of comparative evolutionary methods. AOC allows for the inclusion of recombination detection, a powerful force in shaping gene evolution and interpreting analytic results. As well, it allows for lineage assignment and annotation. This feature (lineage assignment) allows between-group comparisons of selective pressures. This application currently accepts two input files: a protein sequence unaligned fasta file, and a transcript sequence unaligned fasta file for the same gene. Typically, this can be retrieved from public databases such as NCBI Orthologs. Although other methods of data compilation are also acceptable. In addition, the application is easily modifiable to accept a single CDS input, if that data is available.

If protein and transcript files are provided, a custom script “scripts/codons.py” is executed and returns coding sequences where available. Note that this script currently is

set to use the standard genetic code, this will need to be modified for alternate codon tables. This script also removes "low-quality" sequences if no match is found, see the above Data cleaning section.

**Step 1. Alignment.** We used the HyPhy (Kosakovsky Pond et al, 2020) codon-aware multiple sequence alignment procedure available at (https://github.com/veg/hyphy-analyses/tree/master/codon-msa). This was performed with a Human BDNF coding sequence NM_001709.5 Homo sapiens brain-derived neurotrophic factor (BDNF), transcript variant 4, mRNA as a reference-based alignment. Our alignment procedure retained 126 unique in-frame sequences.

**Step 2. Recombination detection.** Performed manually via RDP v5 (Martin et al, 2015), see below, the "Recombination detection" section for additional details. A recombination-free file is placed in the following folder: results/BDNF/Recombinants. For the purpose of this study, we did not detect recombination in our dataset.

**Step 3. Tree inference and selection analyses.** For the recombination-free fasta file, we perform maximum-likelihood phylogenetic inference via IQ-TREE (Minh et al, 2020). Next, the recombination-free alignment and an unrooted phylogenetic tree is evaluated through a standard suite of molecular evolutionary methods. This set of selection analyses includes the following but for the sake of brevity, some of these results were not shown (essentially, most were not statistically significant or not meaningful as relevant to the evolutionary results presented here).

- FEL: locates codon sites with evidence of pervasive positive diversifying or negative selection (Kosakovsky Pond et al, 2005).
- BUSTED[S]: tests for gene-wide episodic selection (Wisotsky et al, 2020).
- MEME: locates codon sites with evidence of episodic positive diversifying selection

  (Murrell et al, 2012).
- aBSREL: tests if a positive selection has occurred on a proportion of branches (Smith et al, 2015).
- SLAC: performs substitution mapping (Kosakovsky Pond et al, 2005).
- BGM: identifies groups of sites that are apparently coevolving (Poon et al, 2008).
- RELAX: compare gene-wide selection pressure between the query clade and background sequences (Wertheim et al, 2015).
- CFEL: comparison site-by-site selection pressure between query and background sequences (Kosakovsky Pond et al, 2021).
- FMM: examines model fit by permitting multiple instantaneous substitutions (Lucaci et al, 2021).

**Step 4A. Lineage assignment and tree annotation.** For the unrooted phylogenetic tree, we perform lineage discovery, via NCBI and the python package ete3 toolkit. Assigning lineages to a K (by default, K = 20) number of taxonomic groups. Here, the aim is to have a broad representation of taxonomic groups, rather than the species being heavily

clustered into a single group. As a reasonable approximation, we aim for <40% of species to be assigned to any one particular taxonomic group.

**Step 4B. We perform tree labeling via the hyphy-analyses/Label-Trees method.** Available at https://github.com/veg/hyphy-analyses/tree/master/LabelTrees. Resulting in one annotated tree per lineage designation. For the purpose of this study, we will only consider the following five lineages for additional analyses (Artiodactyla, Carnivora, Chiroptera, Glires, Primates) as they are the most populated lineages.

**Step 5. Selection analyses on lineages**. Here, the recombination-free fasta file and the set of annotated phylogenetic trees (where labeling was performed in Step 4) is provided for analysis with the RELAX and Contrast-FEL methods.

***Recombination Detection***

Manually tested via RDP v5.5 with modified settings as follows:

We also included the following algorithms/analyses: RDP (Martin and Rybicki, 2000), GENECONV (Padidam et al, 1999), Chimaera (Posada and Crandall, 2001), MaxChi (Maynard Smith, 1992), BootScan (Martin et al, 2005) (Primary and Secondary Scan), SiScan (Gibbs et al, 2000) (Primary and Secondary Scan), 3Seq (Lam et al, 2018).

- Recombination events are “accepted” in cases where three or more methods are in agreement.
- We slightly modified default parameters, such that

- Require topological evidence.
- Polish breakpoints.
- Check alignment consistency.
- Sequences are linear.
- List events detected by >2 methods.
- We manually recheck all of the events via “Recheck all identified events with all methods”.
- We manually accept events detected by >2 methods.
- The resulting alignment was saved as a distributed alignment (with recombinant regions separated).

Recombination was not detected within our Human reference-based alignment. Therefore, we used the single recombination-free alignment for analyses.

**Data Availability**

The AOC application is freely available via a dedicated GitHub repository at: https://github.com/aglucaci/AnalysisOfOrthologousCollections. Raw data for this study is available on GitHub: https://github.com/aglucaci/AnalysisOfOrthologousCollections/tree/main/data/BDNF. Full results for this study include all HyPhy selection analyses JSON-formatted result files are available on GitHub: https://github.com/aglucaci/AnalysisOfOrthologousCollections/tree/main/results/BDNF.

# CHAPTER 5.

# RASCL: RAPID ASSESSMENT OF SELECTION IN CLADES THROUGH MOLECULAR SEQUENCE ANALYSIS.


## Abstract

An important unmet need revealed by the COVID-19 pandemic is the near-real-time identification of potentially fitness-altering mutations within rapidly growing SARS-CoV-2 lineages. Although powerful molecular sequence analysis methods are available to detect and characterize patterns of natural selection within modestly sized gene-sequence datasets, the computational complexity of these methods and their sensitivity to sequencing errors render them effectively inapplicable in large-scale genomic surveillance contexts. Motivated by the need to analyze new lineage evolution in near-real time using large numbers of genomes, we developed the Rapid Assessment of Selection within CLades (RASCL) pipeline. RASCL applies state of the art phylogenetic comparative methods to evaluate selective processes acting at individual codon sites and across whole genes. RASCL is scalable and produces automatically updated regular lineage-specific selection analysis reports: even for lineages that include tens or hundreds of thousands of sampled genome sequences. Key to this performance is (i) generation of automatically subsampled high quality datasets of gene/ORF sequences drawn from a selected "query" viral lineage; (ii) contextualization of these query sequences in codon alignments that include high-quality "background" sequences representative of global SARS-CoV-2 diversity; and (iii)

the extensive parallelization of a suite of computationally intensive selection analysis tests. Within hours of being deployed to analyze a novel rapidly growing lineage of interest, RASCL will begin yielding JavaScript Object Notation (JSON)-formatted reports that can be either imported into third-party analysis software or explored in standard web-browsers using the premade RASCL interactive data visualization dashboard. By enabling the rapid detection of genome sites evolving under different selective regimes, RASCL is well-suited for near-real-time monitoring of the population-level selective processes that will likely underlie the emergence of future variants of concern in measurably evolving pathogens with extensive genomic surveillance.

## Introduction

Rapid characterization and assessment of clade-specific mutations that are found in persistent or rapidly expanding SARS-CoV-2 lineages have become an important component of efforts to monitor and manage the COVID19 pandemic. Identifying the most relevant mutations, e.g., those likely to impact transmission or immune escape, is a priority when new lineages are discovered, as these are key to assessing a lineage's potential threat-level. If observed mutations have not been previously characterized, computational and laboratory-based analytical approaches to inferring whether the mutations provide transmission or immune escape advantages are generally too slow to inform early public health responses.

Epidemiologically relevant mutations are likely subject to natural selection because they provide a fitness advantage (Harvey et al, 2021). Such mutations can be identified by detecting the subtle patterns of nucleotide variation within gene-sequence datasets that are indicative of selective processes. There are a multitude of powerful computational techniques that, given sufficiently informative sequence data, can identify individual codons within genes that are evolving under a range of different selective regimes (Arenas, 2015). Hundreds of papers and preprints have used some of the methods implemented in HyPhy (Kosakovsky Pond et al, 2020) and Datamonkey (Weaver et al, 2018) for SARS-CoV-2 selection analyses e.g., (Benvenuto et al, 2020; Dearlove et al, 2020; Li et al, 2020; Viana et al, 2022), but on small, commonly hand-curated, datasets. This is because codon-based selection analyses do not scale well to more than a few hundred sequences when using the generic out-of-the-box versions of these analyses, and because noisy sequencing data (errors in assembled consensus genomes) can drive false positives.

We developed RASCL to standardize and accelerate comparative selection detection analyses of SARS-CoV-2 variants of interest (VOI) or variants of concern (VOC). The tool has been used to study selective forces which, at least in-part, drove the emergence of the Alpha, Beta, Gamma, and Omicron VOCs (Tegally et al, 2021; Faria et al, 2021; Martin et al, 2021; Martin et al, 2022). More broadly, a tool like RASCL enables near-real-time monitoring of emergent lineages, which in turn can be used both to detect potentially adaptive mutations before they rise to high frequencies, and to help establish relationships between individual mutations and key viral characteristics including

pathogenicity, transmissibility, immune evasiveness, and drug resistance (Hamed et al, 2021; Young et al, 2021; Luchsinger et al, 2021; Abdool et al, 2021; Maher et al, 2022). Through routine analysis of patterns of ongoing selection within individual major lineages, we can reveal the variants or circulating sub-lineages that carry potentially concerning fitness-enhancing mutations, and which would therefore most likely drive future viral transmission (Rambaut et al, 2020).

## Materials and Methods

### *RASCL Application Overview*

The "query" set of whole genome sequences is compared against a diverse set of "background" sequences, chosen to represent globally circulating SARS-CoV-2 sequences (throughout the pandemic), and the query data set is the set of sequences which are the target of selection analyses (e.g., BA.5 clade sequences). Our background dataset is available at https://github.com/veg/RASCL/tree/main/data/ReferenceSetViPR and was assembled from Virus Pathogen Database and Analysis Resource (ViPR, viprbrc.org) (Pickett et al, 2012), a curated database of publicly available viral pathogen sequences, assemblies, and genome annotations (Appendix D, D.S1 File). The inclusion of the background dataset also provides an “outgroup” clade, enabling the study of selection on branches basal to the clade of interest (COI), and to discover regions under selection pressure which are unique to the COI. The application uses several open-source tools, as well as selection analysis modules from the HyPhy software package and assembles the

results from the analysis into JSON files, which can then be visualized with our full-featured ObservableHQ (Perkel, 2021) notebook.

***Map and Compress***

Specifically, RASCL takes as input (i) a "query" dataset comprising a single FASTA file containing unaligned SARS-CoV-2 full or partial genomes belonging to a clade of interest (e.g. all sequences from the PANGO (Maher et al, 2022; Rambaut et al, 2020) lineage B.1.617.2) and (ii) a generic "background" dataset that might comprise, for example, a set of sequences that are representative of global SARS-CoV-2 genomic diversity, e.g. those assembled from ViPR. It is not necessary to remove sequences in the query dataset that are duplicated in the background dataset – the pipeline will do this automatically.

The choice of query and background datasets is analysis specific. For example, if another clade of interest is provided as a background, it is possible to directly identify the sites that are evolving differentially between the two clades. Other sensible choices of query sequences might be sequences from a specific country/region, or sequences sampled during a particular time-period. Note that the analysis does not require the two sets to be reciprocally monophyletic, but in many applications, this will be the case. Following the automated mapping of whole genome datasets into individual coding sequences (based on the NCBI reference annotation), the gene datasets (each containing a set of query and background sequences) are processed in parallel.

***Prepare for Selection Analysis.***

Using complete linkage distance clustering implemented in the TN93 package (tn93-cluster tool, https://github.com/veg/tn93), RASCL subsamples from available sequences while maintaining overall genomic diversity; the clustering threshold distance is chosen automatically to include no more than a user-specified number of genomes "D" (e.g., 300). In the Results section below, we discuss our recent analyses of several SARS-CoV-2 clades and while other available subsampling methods for SARS-CoV-2 genomes exist which rely on spatiotemporal distributions (Cheng et al, 2022), our method relies on increasing sequence diversity to enhance evolutionary signal for downstream selection analyses, while reducing computational complexity. Core method implementations in HyPhy can handle up to 25,000 subsampled sequences in a reasonable time, but the computational cost increases rapidly, and for faster turnaround 1000 sequences are the recommended setting. Following dataset compression, RASCL creates a combined (query and background) alignment with only the sequences that are divergent enough to be useful for subsequent selection analyses. Inference of a maximum likelihood phylogenetic tree with RAxML-NG, (Kozlov et al, 2019), or IQ-TREE (Nguyen et al, 2015), is performed on the merged dataset and the query and background branches of this tree are labeled as Query or Background; internal branches of the tree are labeled using maximum parsimony.

***Selection Analyses***

Selection analyses are performed with state-of-the-art molecular evolution (Spielman et al, 2019) models implemented in HyPhy. To partially mitigate the potentially

confounding influences of within-host evolution (Kosakovsky Pond and Frost, 2006; Pybus et al, 2007), where mutations occurring within an individual have not been filtered by selection at the broader population-level, and sequencing errors, selection analyses are performed only on the internal branches of phylogenetic trees, where at least one or more rounds of virus transmission are captured (Lorenzo-Redondo et al, 2016). The following individual selection tests are applied to each gene-level alignment of merged query and background sequences.

• Branch-site Unrestricted Statistical Test for Episodic Diversification with Synonymous rate variation (BUSTED[S]): this method tests for gene-wide selection which is either pervasive (occurring throughout the evolutionary tree) or episodic (occurring only on some lineages). (Wisotsky et al, 2020).

• Single-Likelihood Ancestor Counting (SLAC), uses a combination of maximum-likelihood and counting approaches to infer pervasive selection through nonsynonymous (dN) and synonymous (dS) substitution rates on a per-site basis for a given coding alignment and corresponding phylogeny. We use the results from SLAC to create substitution mapping of genomic sites and selection analysis results across methods (Kosakovsky Pond and Frost, 2005).

• Coevolution detection using Bayesian Graphical Models (BGM): this method identifies groups of sites that might be co-evolving using the joint distribution of substitutions (Poon et al, 2008).

• Fixed Effects Likelihood (FEL): this method locates codon sites within a gene with evidence of pervasive positive diversifying or negative selection by inferring nonsynonymous (dN) and synonymous (dS) substitution rates on a per-site basis for a given coding alignment and corresponding phylogeny (Kosakovsky Pond et al, 2005).

• Mixed Effects Model of Evolution (MEME): a more sensitive analysis as compared to FEL, this method locates codon sites with evidence of episodic positive diversifying selection, (Murrell et al, 2012).

• Relaxed Selection (RELAX): compares gene-wide selection pressure and looks for evidence that the strength of selection has been relaxed (or intensified) between the query clade and background sequences (Wertheim et al, 2015).

• Contrast-FEL: comparison of site-by-site selection pressure between query and background sequences to detect evidence indicative of different selective regimes (Kosakovsky Pond et al, 2021)

• A FUBAR (Murrell et al, 2013) Approach to Directional Selection (FADE): this method identifies amino-acid sites with evidence of directional selection (Kosakovsky Pond et al, 2008). FUBAR refers to our previously published method Fast, Unconstrained Bayesian AppRoximation for Inferring Selection.

• FitMultiModel (FMM): this method identifies genes with complex multiple instantaneous substitutions that occur within a codon, a rare but potent source of evolutionary signal. (Lucaci et al, 2021).

### *Software Availability*

The RASCL application, depicted at high level in Fig 5.1, is implemented:

- As a standalone pipeline (https://github.com/veg/RASCL) in Snakemake (Mölder et al, 2021).
- As a web application (https://galaxy.hyphy.org/u/hyphy/w/rapid-assessment-of- selection-on-clades-and-lineages), integrated as a workflow in the Galaxy (Afgan et al, 2018) framework, that is freely available for use on powerful public computing infrastructure (https://usegalaxy.org).

### *Visualization and Downstream Post-Hoc Analyses*

Results are combined using a Python script "generate-report.py" into two machine-readable JSON files ("summary.json" and "annotation.json") that represent detailed analysis results for gene segments and individual sites, respectively. JSON is an open standard text-based file format which is also human-readable. It is well-suited for representing structured data and is commonly used for transmitting data in web applications. These JSON files can either be used as input for other software or visualized within a standard web-browser via a feature-rich interactive RASCL dashboard hosted on ObservableHQ (Fig 5.2). For the web application implementation of RASCL, alignments,

trees and analysis results are stored and made web-accessible via the Galaxy platform. Results are visualized with interactive notebooks hosted on ObservableHQ (Fig 5.2) that include an alignment viewer, a visualization of individual codons/amino acid states at user-selected sites mapped onto the tips of phylogenetic trees, and detailed tabulated information on analysis results for individual genes and codon-sites.

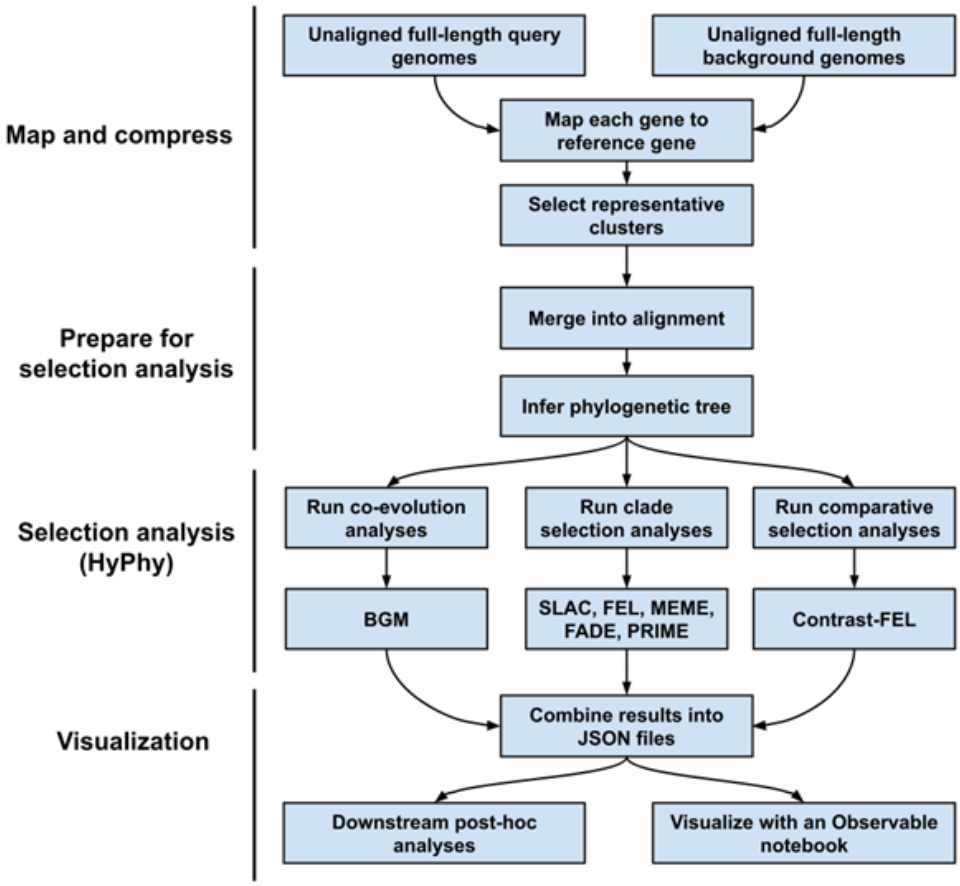



Figure 5.1. The RASCL application overview**.** We highlight the high-level architecture of the RASCL workflow. These include what we call multiple Phases, including a (1) Map and Compress step, where input query and background whole genome sequences are separated into individual genes from the viral genome by mapping to the reference gene. For each gene we then extract the representative gene diversity using genetic distance clustering. (2) is where we prepare our gene alignments for selection analysis. We accomplish this by merging alignments from the background and query datasets into a "combined" dataset. From this, we infer a phylogenetic tree and annotate it based on query and background sequences. (3) We perform selection analyses in HyPhy (described in

further detail in the Methods section). (4) We combine the results of selection analyses across the viral genome by mapping substitutions to each position in the viral genome to create a selection 'profile' for each statistically significant site into an interpretable JSON-formatted file. These combined results are then used for further post-hoc or downstream analysis or ingested by our interactive notebook.

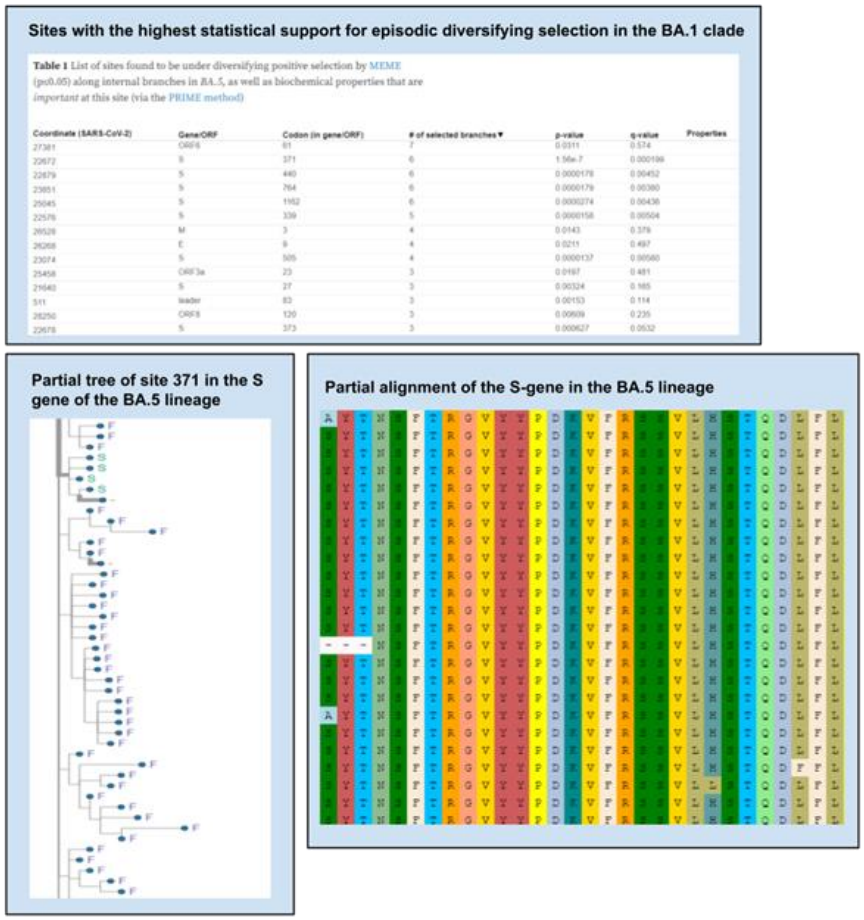



Figure 5.2. Example visualization using our interactive notebook. Here, we highlight some of the features of our interactive notebook which was created to facilitate result exploration. Key features include: (1) tables with statistically significant results for each selection analysis, (2) the ability to explore the full phylogenetic tree or a site-level tree to explore selection acting on individual sites and (3) we provide a multiple sequence alignment viewer for any of the genes in the results.

## Results and Discussion

RASCL uses molecular sequence data from genotypically distinct viral lineages to identify distinguishing features and evolution within lineages. A query set of sequences is compared against a globally diverse set of background sequences. The background data set typically contains globally circulating viral sequences, and the query data set is the set of sequences of particular interest to the user. Below, we describe our analyses of several variants of SARS-CoV-2 whole genome sequences, but RASCL is applicable to any measurably evolving pathogen with sufficient surveillance data.

### *An Overview of Molecular Surveillance of Important SARS-CoV-2 Viral Clades*

As a concrete example of the utility of RASCL consider our analysis of 112,017 BA.1 (WHO Omicron, all available BA.1 sequences as of January 2, 2022) sequences. RASCL selected a median (per gene) of 524 BA.1 sequences (a compression ratio of 99.53%) and a median of 145 background sequences (from a dataset of over 150,000 publicly available sequences from ViPR) to represent genomic diversity in SARS-CoV-2. Using the Spike gene as an exemplar, we compressed all available sequences down to 933 representative sequences, reflecting a compression ratio of 99.17%. This level of compression is representative of recent analyses of a few VOIs/VOCs (see Table 5.1) with RASCL.

Table 5.1. Examples of various recent RASCL analyses on variants of interest and variants of concern for SARS-CoV-2. Full results, and more recent updates, are available through our interactive notebooks (see Methods section for additional details). The lower compression ratios for Omicron reflects the higher genomic variability of this lineage, which gave rise to many sublineages soon after emergence, combined with a high-volume of viral genome sequencing.

| Lineage | Available | Median analyzed | Compression Ratio | Positively selected sites | Positively selected sites |
|---|---|---|---|---|---|
| | | | | $p \leq 0.05$ | $p \leq 0.0001$ |
| BA.2 (WHO Omicron) | 48,623 | 319 | 1:152 | 168 | 29 |
| BA.1 (WHO Omicron) | 112,017 | 524 | 1:208 | 200 | 42 |
| B.1.617.2 (WHO Delta) | 1,983 | 59 | 1:29 | 77 | 1 |
| B.1.621 (WHO Mu) | 3,288 | 101 | 1:32 | 67 | 3 |
| C.37 (WHO Lambda) | 2,127 | 80 | 1:26 | 53 | 1 |
| P.1 (WHO Gamma) | 2,070 | 47 | 1:43 | 49 | 4 |
| B.1.1.7 (WHO Alpha) | 8,586 | 169 | 1:50 | 44 | 1 |

Importantly, there is evidence of diversifying positive selection acting on the BA.1 sequences, on 42 (p 0.0001) individual sites (there are 4312 sites that are polymorphic in the amino-acid space among clade sequences; selection also includes the basal branch of the clade) along the internal branches of the clade (Appendix D, D.S1 Table). There is evidence of diversifying positive selection acting on the BA.1 sequences, on 359 individual sites along all branches of the clade (Appendix D, D.S2 Table), with 40 sites with an LRT p-value of 0.0001. When comparing the strength of selection on BA.1 to background sequences along internal tree branches, 21 individual sites along the internal branches of the clade (Appendix D, D.S3 Table) showed statistically significant differences. Over the entire tree, 31 sites demonstrate evidence of directional selection (Appendix D, D.S4 Table).

Along the internal branches of the BA.1 clade, 81 pairs of sites showed evidence of coevolution (Appendix D, D.S5 Table). Along the internal branches of the BA.1 clade, 47 sites showed evidence of negative selection (Appendix D, D.S6 Table). Along the internal branches of the BA.1 clade, 15 (out of 21 segments considered 71.4%) genes/ORFs showed evidence of episodic diversifying selection (BUSTED, q-value ≤ 0.1, Appendix D, D.S7 Table).

***A Closer Examination of the SARS-CoV-2 BA.5 clade***

We investigated the nature and extent of selective forces acting on the viral genes in BA.5 clade (all available BA.5 sequences as of August 9, 2022) sequences by performing a series of comparative phylogenetic analyses on a median of 258 BA.5 sequences and a median of 113 sequences from available sequences chosen to represent genomic diversity in SARS-CoV-2. We compiled our background dataset from the globally subsampled Nextstrain (Hadfield et al, 2018) (https://nextstrain.org/ncov/gisaid/global/all-time) build (last accessed June 12, 2022, genomes were sampled from the beginning of the SARS-CoV-2 pandemic) (Appendix D, D.S1 File). We observe that:

- There is evidence of diversifying positive selection acting on the BA.5 sequences, on 94 individual sites (there are 2737 sites that are polymorphic in the amino-acid space among clade sequences; selection also includes the basal branch of the clade) along the internal branches of the clade (Appendix D, D.S8 Table).

- There is evidence of diversifying positive selection acting on the BA.5 sequences, on 133 individual sites along all branches of the clade (Appendix D, D.S9 Table).
- When comparing the strength of selection on BA.5 to background sequences along internal tree branches, 12 individual sites along the internal branches of the clade (Appendix D, D.S10 Table) showed statistically significant differences.
- Over the entire tree 11 sites show evidence of directional selection (Appendix D, D.S11 Table).
- Along the internal branches of the BA.5 clade, 37 pairs of sites showed evidence of coevolution (Appendix D, D.S12 Table).
- Along the internal branches of the BA.5 clade, 14 sites showed evidence of negative selection (Appendix D, D.S13 Table).
- Along the internal branches of the BA.5 clade, 14 (out of 21 segments considered, 66.7%) genes/ORFs showed evidence of episodic diversifying selection (BUSTED[S], q-value $\leq 0.1$, Table 5.2).

Table 5.2. BUSTED[S] selection results on the BA.5 SARS-CoV-2 clade across segments. Segment corresponds to the gene or ORF does under analysis. omega1 refers to the first omega rate class, p1 refers to proportion of sites which fit this rate class. omega2 refers to the first omega rate class, p2 refers to proportion of sites which fit this rate class. omega3 refers to the first omega rate class which captures the episodic diversifying features, p3 refers to proportion of sites which fit this rate class. p-value, the p-value for the likelihood ratio test. q-value refers to the multiple-test corrected q-value (Benjamini-Hochberg).

| Segment | omega1 | p1 | omega2 | p2 | omega3 | p3 | p | q |
|---|---|---|---|---|---|---|---|---|
| 3C | 0.03 | 0.94 | 0.07 | 0.01 | 4.93 | 0.05 | 0.2449 | 0.3429 |
| E | 0 | 0.32 | 1 | 0 | 1.18 | 0.68 | 0.4462 | 0.5511 |
| M | 0 | 0.17 | 0 | 0.79 | 8.69 | 0.04 | 0.0006 | 0.0009 |
| N | 0 | 0.93 | 0.69 | 0.05 | 51.71 | 0.02 | 0 | 0 |
| ORF3a | 0 | 0.35 | 0 | 0.61 | 25.77 | 0.04 | 0 | 0 |
| ORF6 | 0 | 0.03 | 0 | 0.93 | 34.85 | 0.04 | 0 | 0 |
| ORF7a | 0.21 | 0.91 | 0.29 | 0.05 | 28.91 | 0.03 | 0 | 0 |
| ORF8 | 0.42 | 0.96 | 0.43 | 0.01 | 45.23 | 0.02 | 0 | 0 |
| RdRp | 0.11 | 0.98 | 0.47 | 0.01 | 76.27 | 0.01 | 0 | 0 |
| S | 0.75 | 0.98 | 0.87 | 0.02 | 64042.59 | 0 | 0 | 0 |
| endornase | 0 | 0.59 | 0 | 0.39 | 18.75 | 0.03 | 0.0005 | 0.0008 |
| helicase | 0 | 0.73 | 0 | 0.09 | 2.3 | 0.17 | 0.32 | 0.4201 |
| leader | 1 | 0 | 1 | 1 | 8003.48 | 0 | 0.0002 | 0.0004 |
| methyltransferase | 0 | 0.06 | 0 | 0.93 | 67.61 | 0.01 | 0 | 0 |
| nsp10 | 0.32 | 0.49 | 0.5 | 0.51 | 1 | 0 | 0.5 | 0.5833 |
| nsp2 | 0 | 0.89 | 0.74 | 0.09 | 72.38 | 0.01 | 0 | 0 |
| nsp3 | 0 | 0.76 | 0 | 0.23 | 47.93 | 0.01 | 0 | 0 |
| nsp6 | 0 | 0.07 | 0 | 0.46 | 1 | 0.47 | 0.5 | 0.5526 |
| nsp7 | 0.32 | 0.85 | 0.34 | 0.1 | 1 | 0.05 | 0.5 | 0.525 |
| nsp8 | 0 | 0.92 | 0 | 0.07 | 34.97 | 0.02 | 0 | 0 |
| nsp9 | 0.06 | 0 | 0.36 | 1 | 1.11 | 0 | 0.5 | 0.5 |

We consider our results from gene-wide estimates of adaptation where we observed that 3 structural and 11 non-structural proteins yield statistically significant results. Within the set of structural proteins, we find Spike (S), Membrane glycoprotein (M), and the nucleocapsid phosphoprotein (N), these genes have been implicated in complex biological functions, including as a highly conserved target, M, (Shen et al, 2021) and have been the

focus of studies on viral infection (Rathnasinghe et al, 2022), pathology (Zhao et al, 2022) and vaccination and therapeutic intervention (Costa et al, 2022). Interestingly, within the set of non-structural proteins we find ORF3a, ORF6, ORF7a, and ORF8, which have been implicated in novel biological mechanisms in the SARS-CoV-2 virus including the induction of autophagy and role as a viral ion channel, ORF3a, (Su et al, 2021; Kern et al, 2021) disruption of nucleocytoplasmic transport, ORF6, (Addetia et al, 2021) inhibition of host interferon response, ORF7a, ORF8 (Pawlica et al, 2021; Chen et al, 2022). We also find several members of the ORF1ab polyprotein including RNA-dependent RNA polymerase (RdRp), endoRNAse, leader, methyltransferase, nsp2, nsp3, nsp8. Several important sites in Spike from Table 5.1 are discussed, including: S/339, S/371, S/440, S/764, S/1162. We highlight these sites due to the level of statistical signal associated with them (we find 5 branches selected in the exploratory MEME analysis) and provide selection profiles for each of these sites below.

Table 5.3. Site profiles for selected sites in the BA.5 Spike gene. Genomic position (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome. Codon in gene: the location of the codon in the corresponding Spike gene. Number Of selected branches: the number of tree branches (internal branches BA.5 clade) that have evidence of diversifying positive selection at this site (empirical Bayes factor ≥100). FEL p-value: the p-value for the likelihood ratio test that non-synonymous rate / synonymous rate ≠ 1 at this site. MEME Internal p-value: the p-value for the likelihood ratio test that a non-zero fraction of internal branches have omega > 1 (i.e., episodic diversifying selection at this branch). CFEL p-value: the p-value for the likelihood ratio test that omega ratios between the internal branches of the two clades are different. p-values reported in this table are not corrected for multiple testing. Amino acid composition (including gaps) is reported for the BA.5 and background dataset at the corresponding site.

| # | Position | Codon | Branches | FEL | MEME | CFEL | Composition BA.5 | Composition Background |
|---|---|---|---|---|---|---|---|---|
| 1 | 22576 | 339 | 5 | 0.00000645 | 0.0000158 | 0.000363 | D327 G15 -8 N1 | G169 D14 -1 |
| 2 | 22672 | 371 | 6 | 1.01E-07 | 1.56E-07 | 0.0075 | F310 S28 -9 L4 | S168 L8 F6 -2 |
| 3 | 22879 | 440 | 6 | 0.00000818 | 0.0000178 | 0.0716 | K298 N44 -9 | N168 K15 -1 |
| 4 | 23797 | 764 | 6 | 0.00000729 | 0.0000179 | 0.000463 | K325 N17 -8 I1 | N167 K13 -4 |
| 5 | 25045 | 1162 | 6 | 0.0000201 | 0.0000274 | 0.0208 | P308 L41 -2 | P182 Q1 S1 |

### *Assessing Evolutionary Pressures on the SARS-CoV-2 μ (B.1.621) Clade*

We identify genomic sites in B.1.621 (μ) (Halfmann et al, 2022) clade sequences that may be subject to selective forces and could be prioritized for further studies but have not yet reached high frequencies. We present an analysis of the B.1.621 variant, performed on individual genes/protein products, using a median of 101 μ sequences subsampled from all available sequences in GISAID (Khare et al, 2021) to represent the genomic diversity

in this clade (all sequences as of September 7, 2021). A similarly subsampled global SARS-CoV-2 background dataset from publicly available sequences via the ViPR database is used as background and provided in our Github repository, (linked to in our Methods section). Interactive results for this analysis can be explored in our ObservableHQ notebook (https://observablehq.com/@aglucaci/rascl-mu) and our Virological (https://virological.org) post (https://virological.org/t/assessing-evolutionary- pressures-on-the-sars-cov-2-mu-b-1-621-clade/760). Our analysis identified 67 (Appendix D, D.S14 Table) individual codon sites (among 1643 sites that are polymorphic in the amino-acid space) that showed evidence of episodic diversifying selection along internal branches of this clade using the MEME method at $q \leq 0.20$ false discovery rate (FDR). A total of 5 sites (Appendix D, D.S15 Table) were found to be subject to directional selection using the FADE method.

We identify high-priority sites in SARS-CoV-2 μ (B.1.621) sequences (Table 5.4), with a "Rank" for each site based on a point system described below. We identify and rank sites (+1 for each category) according to the following protocol:

- Inferred to be under positive selective pressure.
- Are not clade-defining mutations.
- Contain mutations that are not predictable based on the evolution of Sarbecovirus sequences (Lytras et al, 2022).

- Contain mutations that occur in a large fraction of unique haplotypes. This was shown to be predictive of near-term growth in a separate analysis from our global SARS-CoV-2 analysis (Maher et al, 2022).

Additionally, where available, we provide interpretation of our identified sites in terms of known functional significance, temporal growth trends and location on the 3D structure of the protein. Briefly, each site is given a point for each of the follow requirements that are met: found to be positively selected, the site occurs outside of the clade defining site set, the site has any unexpected mutations as described in our Sarbecoviruses evolutionary analysis notebook (https://observablehq.com/@spond/sars-cov-2-pvo), the site has a mutation present above the minimum threshold in the SARS-CoV-2 Global Haplotype analysis (https://observablehq.com/@spond/sc2-haplotypes), for this we remove any mutation present in background clade, and remove gaps.

***Evidence of Natural Selection History Operating on SARS-CoV-2 Genomes***

For the set of high-priority SARS-CoV-2 genomic sites (taken from Table 5.4), sites inferred from B.1.621 sequences, we observe when and how selection (positively or negatively) operated on them, through a series of 3-month overlapping intervals going back to the beginning of the pandemic (Fig 5.3). The earliest intervals end in February 2020 and the latest - in September 2021. In selected sites we observe the temporal trends of high-priority sites in Spike (Fig 4) and RdRp (Fig 5.5) in B.1.621 sequences. We also describe

the spatial location of sites inferred to be under positive selective pressure in the Spike gene (Fig 5.6) from B.1.621 (μ) sequences on the structure of the protein (from Table 5.4).

***Potential Biological and Clinical Significance of Mutations***

The ongoing monitoring of emergent VOIs and VOCs can detect adaptive mutations before they rise to high frequency and help establish their relationship to key clinical parameters including pathogenicity and transmissibility. Additionally, continued evolution within a particular clade may form the foundation for a subclade with further functional sites of interest. Based on current information for the Spike gene from Stanford Coronavirus Antiviral Resistance Database (CoVDB, https://covdb.stanford.edu) (Tzou et al, 2022) we include several annotations with clinical relevance. From the SARS-CoV-2 B.1.621 (μ) Spike gene, we identify the following sites of interest from Table 5.4 due to their interaction with epitope binding in monoclonal antibodies (mAbs): 144, 145, 146, 147, 148, 417, and 501.

Table 5.4. A table of high-priority sites in SARS-CoV-2 μ B.1.621 sequences. Genomic position (in SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome. Gene/ORF: which gene or ORF does this site belong to. Codon in gene: the location of the codon in the corresponding gene/ORF. p-value: the p-value for the likelihood ratio test that a non-zero fraction of branches have omega > 1 (i.e., episodic diversifying selection at this branch). This is not corrected for multiple testing; the MEME test is generally conservative on real data. q-value: multiple-test corrected q-value (Benjamini-Hochberg). We assign a "Rank" to each site based on a point system described above. Properties: which, if any, of the five compositive biochemical properties (Atchley et al, 2005) are conserved or changed at this site.

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | p-value | q-value | Rank | Property |
|---|---|---|---|---|---|---|
| 16075 | RDRP/ORF1b | 879/870Y | 0.0452115 | 0.140312 | 4 | |
| 28873 | N | 201G | 0.0460285 | 0.138086 | 4 | |
| 28253 | ORF8 | 121VHF | 0.0300385 | 0.180231 | 4 | |
| 25336 | S | 1259HV | 0.0157386 | 0.134903 | 4 | |
| 25333 | S | 1258D | 0.00852771 | 0.0959367 | 4 | |
| 13516 | RDRP/ORF1b | 26/17I | 0.0336153 | 0.168077 | 4 | |
| 14122 | RDRP/ORF1b | 228/219D | 0.036867 | 0.170155 | 4 | |
| 14530 | RDRP/ORF1b | 364/355F | 0.0416466 | 0.144161 | 4 | |
| 14767 | RDRP/ORF1b | 443/434V | 0.0384624 | 0.161005 | 4 | |
| 14785 | RDRP/ORF1b | 449/440T | 0.0383252 | 0.168257 | 4 | charge |
| 17976 | Helicase/ORF1b | 580/1503F | 0.0113156 | 0.107201 | 3 | |
| 20550 | Endornase/ORF1b | 310/2361I | 0.00645087 | 0.0893198 | 3 | |
| 21234 | Methyltransferase/ORF1b | 192/2589Y | 0.049474 | 0.134929 | 3 | |
| 21640 | S | 27LT | 0.000824192 | 0.0247258 | 3 | |
| 21997 | S | 146YNTPS | 0.000598969 | 0.0215629 | 3 | |
| 22003 | S | 148S | 0.0221244 | 0.165933 | 3 | |
| 22000 | S | 147HQNP | 0.0110123 | 0.1166 | 3 | “Overall, secondary” |
| 19482 | Exonuclease/ORF1b | 481/2005V | 0.0474395 | 0.133424 | 3 | |
| 25707 | ORF3a | 106FPIL | 0.0410847 | 0.145005 | 3 | |
| 27210 | ORF6 | 4PHI | 7.99E-06 | 0.000479141 | 3 | |
| 29023 | N | 251S | 0.0425228 | 0.141743 | 3 | |
| 19548 | Exonuclease/ORF1b | 503/2027VS | 0.0470548 | 0.134442 | 3 | |
| 29443 | N | 391AN | 0.0325738 | 0.167522 | 3 | |
| 14470 | RDRP/ORF1b | 344/335I | 0.0339005 | 0.164921 | 3 | |
| 18327 | Exonuclease/ORF1b | 96/1620I | 0.0395957 | 0.15494 | 3 | |
| 2944 | NSP3/ORF1a | 633/894S | 0.0246692 | 0.158588 | 3 | |
| 17820 | Helicase/ORF1b | 528/1451S | 0.0392635 | 0.160624 | 3 | |
| 17025 | Helicase/ORF1b | 263/1186I | 0.0241484 | 0.167182 | 3 | |
| 15001 | RDRP/ORF1b | 521/512C | 0.0457987 | 0.139725 | 3 | |
| 9472 | NSP4/ORF1a | 307/3070T | 0.040304 | 0.145094 | 3 | |
| 9424 | NSP4/ORF1a | 291/3054T | 0.0402514 | 0.147862 | 3 | |
| 9139 | NSP4/ORF1a | 196/2959S | 0.0111345 | 0.111345 | 3 | |
| 8614 | NSP4/ORF1a | 21/2784VT | 0.0400482 | 0.153376 | 2 | |
| 6535 | NSP3/ORF1a | 1273/2091D | 0.0246588 | 0.164392 | 2 | |
| 8578 | NSP4/ORF1a | 9/2772H | 0.0198022 | 0.162018 | 2 | volume |
| 19479 | Exonuclease/ORF1b | 480/2004CV | 0.00358349 | 0.0586389 | 2 | “volume, charge” |
| 8659 | NSP4/ORF1a | 36/2799N | 0.0441831 | 0.142017 | 2 | charge |
| 18960 | Exonuclease/ORF1b | 307/1831V | 0.0462054 | 0.134145 | 2 | |
| 16260 | Helicase/ORF1b | 8/931Y | 0.0309811 | 0.17989 | 2 | |
| 2230 | NSP2/ORF1a | 476/656SA | 0.0316127 | 0.172433 | 2 | |

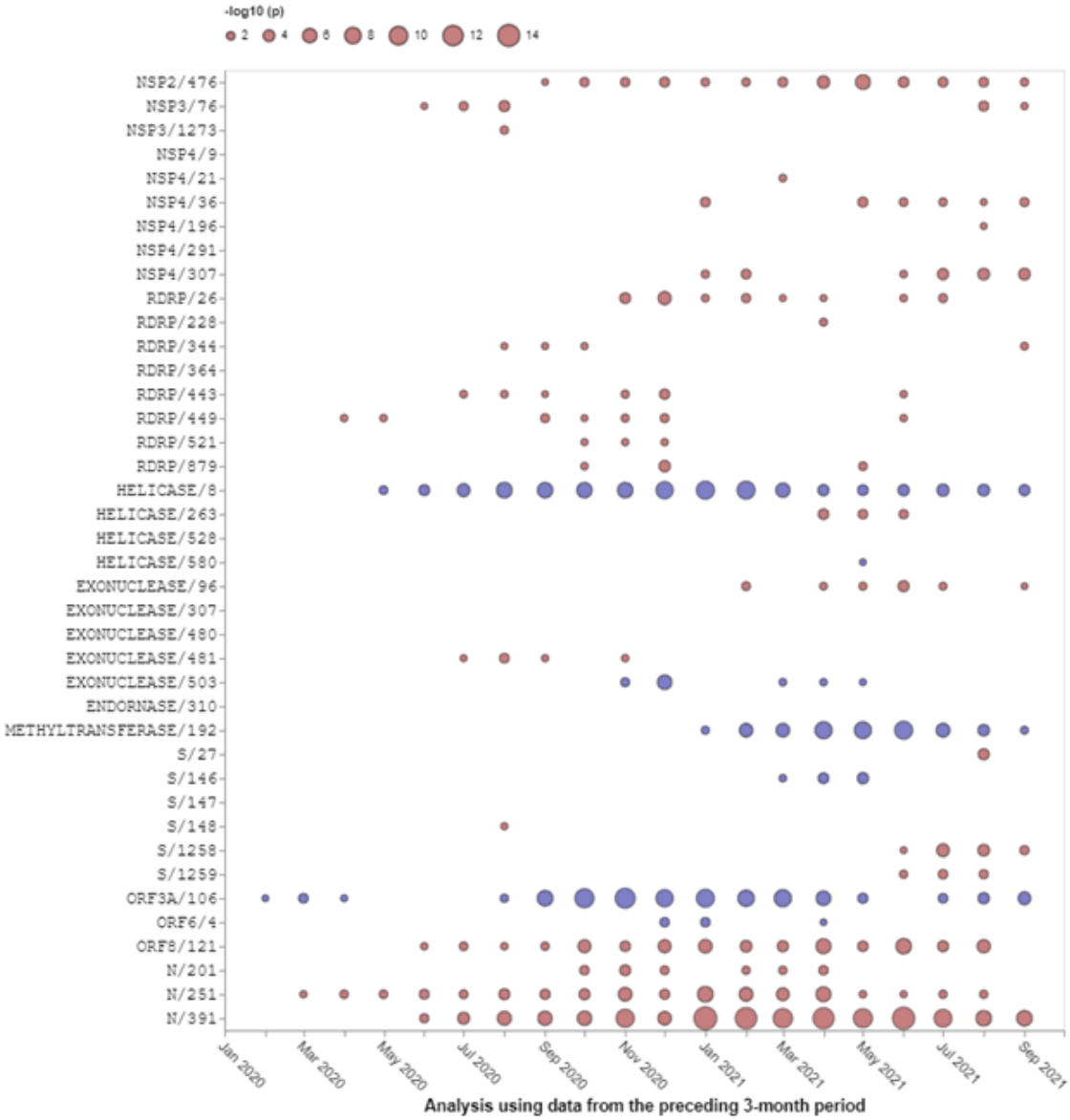



Figure 5.3. Evolutionary trajectories of 40 high priority selected sites (from Table 5.4). If a site was found to be positively (red) or negatively (blue) selected during a specific time, a bubble will be drawn at a corresponding point on the plot. The area of the bubble is scaled as -log10 p, where p is the p-value of the FEL likelihood ratio test. Larger bubbles correspond to smaller p-values; p-values are not directly comparable between different

time windows and different genes due to differences in sample sizes and other factors. The x-axis shows the endpoint of the time-window, e.g., March 30th, 2021, will correspond to the analysis performed with the data from January 1, 2021, to March 30, 2021. Figures like this can be generated with the "Evidence of natural selection history operating on SARS-CoV-2 genomes" ObservableHQ notebook (https://observablehq.com/@spond/sars-cov-2-selected-sites).

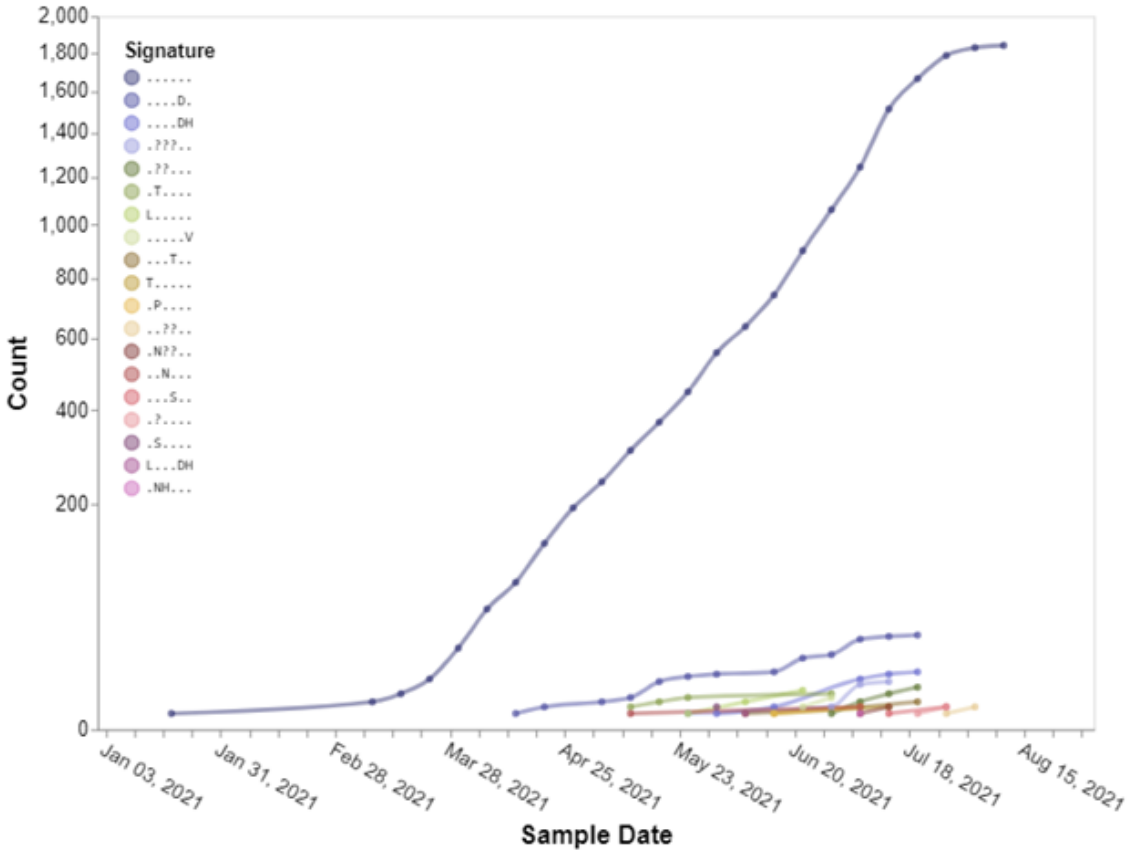



Figure 5.4. Temporal trends of the substitution combinations at selected sites represented in Table 5.4 in the Spike gene for B.1.621 (μ) sequences in 2021 (from left to right: S/27, S/146, S/147, S/1258, S/1259). The symbol "." denotes the reference residue at that site. Figures like this can be generated using Trends in mutational patterns across SARS-CoV-2 Spike enabled by data from https://observablehq.com/@spond/spike-trends. Additional search parameters include "B.1.621[pangolin] AND 20210101[after]". Notebook link: https://observablehq.com/@spond/spike-trends.

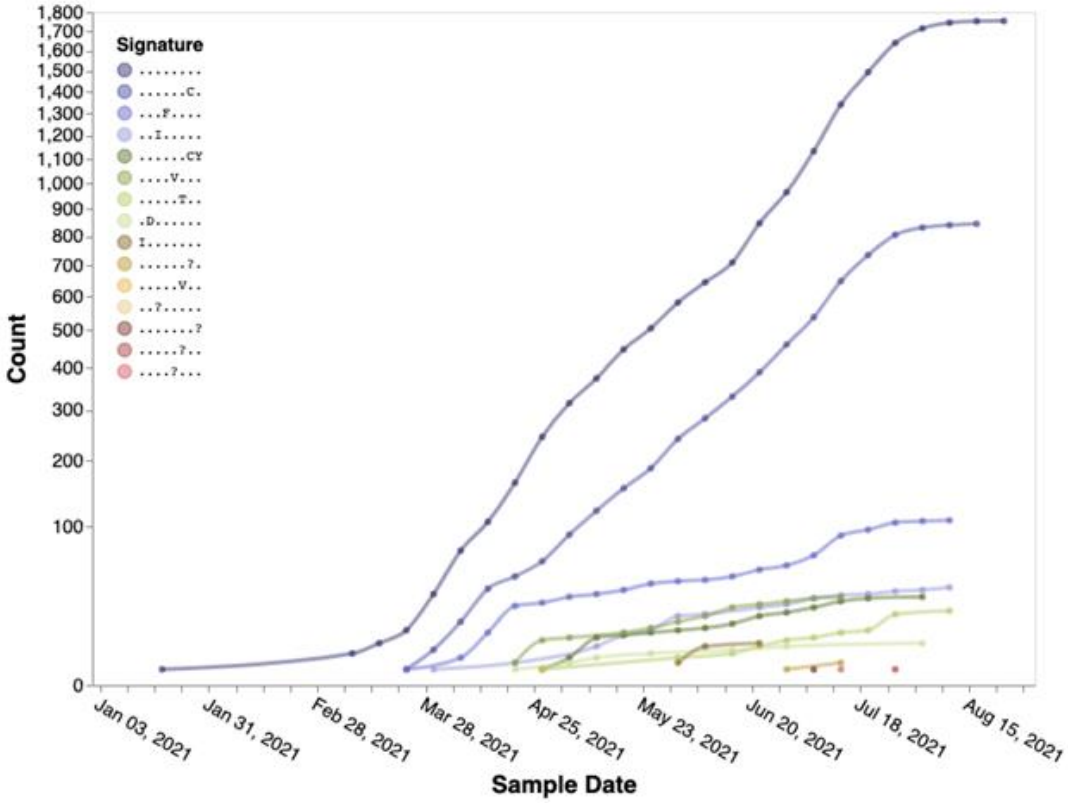



Figure 5.5. Temporal trends of the substitution combinations at all sites represented in Table 5.4 in the RDRP (RNA-dependent RNA polymerase) gene for B.1.621 (μ) sequences in 2021 (from left to right: RDRP/26, RDRP/228, RDRP/344, RDRP/364, RDRP/443, RDRP/449, RDRP/521, RDRP/879). The symbol "." denotes the reference residue at that site. Figures like this can be generated using Trends in mutational patterns across SARS-CoV-2 Spike enabled by data from https://observablehq.com/@spond/spike-trends. Additional search parameters include "B.1.621[pangolin] AND 20210101[after]". Notebook link https://observablehq.com/@spond/spike-trends

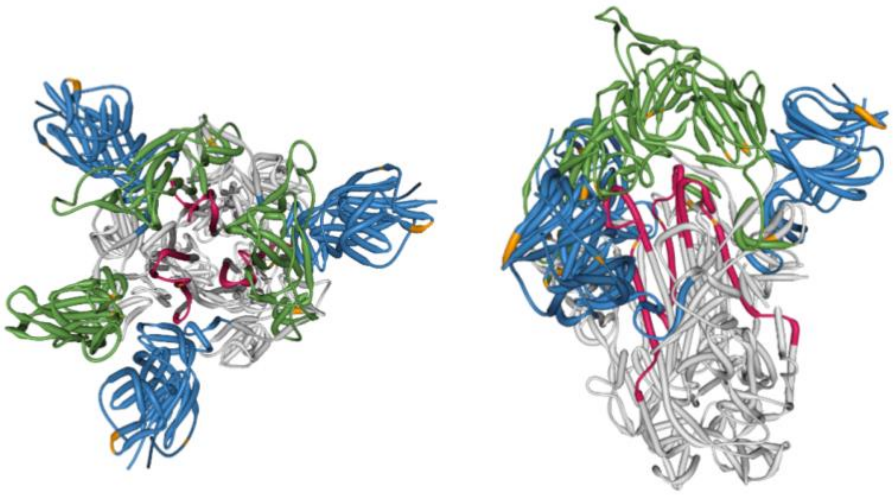


Figure 5.6. Spike protein crystal structure annotation 6CRZ (https://www.rcsb.org/structure/6CRZ) with MEME sites, a measure of episodic selection (These sites are listed in Table 5.4). The color legend for these figures is as follows: the N-Terminal domain (NTD) region is highlighted in Blue, the Receptor binding domain (RBD) region is highlighted in Green, The Heptad Repeat (HR) region is highlighted in Ruby, MEME (Positively selected) sites are highlighted in Orange. To interact with the figure above visit: https://observablehq.com/@aglucaci/categorical-ngl-rascl-mu.

interpret how these changes impact the fitness and host-interaction of the virus. Additionally, this information is crucial to provide public health officials with the most up-to-date information when making public health decisions. To gather this information, computational and laboratory-based analytical approaches have been used to test and

## Conclusions

A need that has reoccurred throughout the course of the COVID-19 pandemic is to rapidly identify molecular changes as they arise within the SARS-CoV-2 genome and to validate hypotheses about the observed genotype and phenotypic implications (McCallum et al, 2021).

These current approaches require both significant effort and time to complete, therefore the results may be gained too slowly to inform early public health responses. Computational methods that detect natural selection can be leveraged to identify sites of interest within viral clades. Due to the massive amount of sequence availability of SARS-CoV-2 genomes, many such methods are rendered uncapable of providing results in a timely manner, bottlenecked by the increased computational complexity associated with large-scale analysis. We address this limitation with RASCL, an agile method that can be used to rapidly characterize and assess natural selection at sites and across proteins within viral genomes. Now, SARS-CoV-2 VOI/VOCs can be screened for signals of selection in a standardized manner and at an accelerated rate, while providing easily interpretable, near-real-time results.

The novelty of RASCL lies in its design; it is highly modular and easily adaptable to rapidly analyze any molecular evolving pathogen. Regarding the modularity component, there are phases of the pipeline, described in Fig 5.1, each of which can be parallelized across either a high-performance computing environment or a personal computer. I take

the intermediate and terminal files created by our methods throughout the analysis and combine the pertinent output files together into the commonly used, standardized JSON format. To make interpretation and visualization of the results easy for the user, I created a customizable RASCL dashboard page using ObservableHQ which runs in the browser of any internet browser. The results page is dynamic and interactive, allowing the user to inspect the results for different signals of selection with ease. The modularity of RASCL makes it highly scalable, yielding near-real-time results at any stage of an outbreak. At the beginning stages of pathogen emergence when very few sequences exist, analyses run quickly, and as the outbreak persists and the number of sequences increases, subsampling can be increased, limiting the computational bottleneck.

RASCL is available in two forms, as a standalone pipeline that uses Snakemake, as well as a web application integrated as a workflow in the Galaxy framework. By implementing the method in these two ways, users at any level of bioinformatics expertise benefit, making RASCL highly accessible. RASCL has been designed such that with minimal modifications to the background and reference genomes, genes under analysis, and default thresholding settings any other evolving pathogens can be rapidly scrutinized to immediately inform public health measures. RASCL can be modified to screen for signals of selection in gene sequences from the current Monkeypox outbreak. Differences in virulence have been reported between Monkeypox isolates from two different geographic regions (Chen et al, 2005; Isidro et al, 2022), thus rapidly identifying the evolution within viral genes is highly relevant to public health efforts.

To date RASCL has been used to characterize the role of natural selection in the emergence of the Beta, Gamma, Omicron (Maher et al, 2022) and BA.4/BA.5 (Tegally et al, 2022) VOC lineages, as well as to identify patterns of convergent evolution in the Alpha, Beta and Gamma lineages (Martin et al, 2021). Whole genome sequences from any viral clade of interest (i.e., emerging pathogens), can be separated into a query and background sequence dataset representing the global diversity of viral sequences serves as the input to the tool, where the selective forces associated within and between the two sets of sequences are identified. RASCL has also been used to monitor the evolution of several lineages (see Table 5.1) and will be applied to future SARS-CoV-2 sublineages as they emerge. Therefore, whenever future genomic surveillance efforts reveal new potentially problematic SARS-CoV-2 lineages, we will use RASCL to analyze these too.

Among the limitations of the current study is that at this moment the RASCL application does not take recombination within genes into account; recombination between genes is handled by performing gene-by-gene analyses. While recombination plays a generally recognized role in the evolution of coronaviruses between species, only a limited amount of recombination is observed in the globally circulating viral population of SARS-CoV-2 (Turakhia et al, 2022). In addition, the types and modalities of selection analyses employed in the RASCL application (described in the Methods section) have robust statistical inference that are only biased when significant recombination changes the topology of the inferred phylogenetic relationships. Future versions of RASCL will include an optional configuration to detect genetic recombination using state-of-the-art methods

(Samson et al, 2022; Martin et al, 2020; Kosakovsky Pond et al; 2006) and will include an updated interactive notebook to visualize and interpret these kinds of complex evolutionary signals. By taking recombination into account, I look forward to increasing the role that the RASCL application can play in the global monitoring and surveillance of evolution in SARS-CoV-2 and other important pathogens.

Our focus in this study is on the rapid, near-real-time monitoring and analysis of emerging pathogens, where the RASCL software application provides interpretable results for molecular surveillance of continued natural evolution. While an area of active research and both proliferative and heated debate, the question of SARS-CoV-2 origins (Boni et al, 2020; Domingo et al, 2021) is out of scope for our study. Indeed, while our analysis of VOIs/VOCs in SARS-CoV-2, has uncovered important and emerging regions of interest in key viral proteins, RASCL has broader applicability to global health threats with known natural origins and existing animal reservoirs. However, further investigation in SARS-CoV-2 origins is critical for understanding the true biological and epidemiological context and complex evolutionary history (Makerenkov et al, 2021) of global pathogens.

# CHAPTER 6.

# SUMMARY

## A Review of the Dissertation Chapters.

Chapter 1 (Introduction) highlighted the current understanding of our scientific problem under investigation. It briefly described the biological process and the underpinnings of the genetic world. In this chapter, I will conclude my discussion of this work, and highlight important components for the reader to take away.

Chapter 2 forms the foundation of my work and encompasses the development and validation of statistical methods built to understand how natural selection shapes gene evolution and lays the groundwork for future expansions. The biological insight compiled from the building blocks of scientific researchers before me leads to the hypothesis generation process and software development of a codon substitution model that accounts for MH changes. Our recent publication highlights the models I developed in order to understand complex multi-nucleotide mutation changes within molecular sequences.

Chapter 3 and 4 demonstrated the utility of applying codon models to empirical data sets. Chapter 3 discovered novel insights into the relationship of purifying selection in shaping the relationship between members of the heat shock gene family and their evolutionary history in the context of their functional domains. Chapter 4 highlighted

efforts in understanding BDNF evolution in Mammalia and its potential relationship to disease. With a larger dataset than previous studies and using modern techniques I was able to uncover novel sites under episodic selection and show extensive coevolution among functional domains.

Chapter 5 highlights contributions towards rapid molecular sequence analysis of global viral pathogens during the COVID-19 pandemic. We are living through history, and this viral outbreak is a generation defining event on the scale of the early HIV pandemic in the 1980s. The acquisition, genomic sequencing, and dissemination of patient-derived sample information proved to be a critical component of the global response. As is the rapid assessment and qualification of key sites (responsible for biological response) inside and outside of viral lineage defining mutations (and their relationship to relevant clinical parameters). Our software application, RASCL, led to significant results in the literature describing the early and continued evolution of the Omicron SARS-CoV-2 variant.

**What have I Accomplished?**

Overall, the style and substance of this work contributes to the appreciation of nuanced biological mechanisms, rare mutations, and the expansion of thought surrounding the evolutionary process. In sum: my work builds upon seminal research in the areas of phylogenetics, DNA sequencing, evolutionary theory and establishes a modern approach

for inclusive gene history exploration. The development and validation of a MH-inclusive codon substitution model, a workhorse of molecular evolution and comparative genomics, allows for fundamental inquiries which probe the landscape of the mutational process involved in shaping protein coding gene evolution and DNA repair.

On the viral molecular surveillance front, I have established a robust application for the rapid assessment of global viral pathogens. While our contributions to date have largely focused on the COVID-19 pandemic (with our direct contributions to the literature mentioned above), I envision small modifications to the input parameters and sequence type to enable analysis of any viral strain of global public health interest, including respiratory transmitted viruses such as Influenza and viruses transmitted through additional vectors: Zika, Ebola. Together, I developed a strong bioinformatics pipeline capable of generating information-rich visualizations along with interactive and actionable results important for public health officials.

**Why is it Important?**

Comparative genomics analysis is an important component of the modern scientific enterprise. The subsequent results drawn from such analyses are “high-viz” and regularly part of studies published in high impact journals where they are part of the early data exploratory process and can influence the thoughts (in the field) and hypotheses of the

modern researcher. The results also lend them directly towards experimental study design and further analysis via *in-vivo* and *in-vitro* methods, costing money, time, experience to train experimental researchers, and opportunity cost of the experiment. By examining the fundamental assumptions of such models, I can open new doors in the biological sciences for the exploration of underappreciated mutational mechanisms and the biological diversity they enable.

By providing easy to use bioinformatic software, I can democratize the process of going from hypothesis to novel result, the so called "day science" or allow for rapid data exploration leading to novel scientific paths of inquiry, the so called "night science" (Yanai and Lercher et al, 2019). I also regard the interpretability of our models as a significant feature, by removing the "black box" algorithm of overcomplicated and non-generalizable models I offer an approachable suite of methods for scientific researchers who are often met with an uphill battle to fully contextualize their results with the evolutionary process.

**What does the Future Hold for This Work?**

I envision a broad future where researchers apply MH-enabled models to diverse datasets. This assumes continued adoption and implementation of the base MH model as a core part of modern codon substitution models. While I have begun some of this work, I see amenable situations for the addition of MH existing HyPhy methods. I view serine to

serine degeneracy and MH island exchanges as an important component of evolutionary history, as this exchange is less favored through a physicochemical unmatched intermediate. Particularly important is where this S-to-S exchange intersects with the codon usage literature and the unraveling of the historically ignored and biologically relevant contributions due to synonymous codons. Included in this are the important contributions of synonymous codons to protein science in the form of mRNA translation speed and backbone structural folding.

I also hold the view that the plethora of previously underappreciated DH exchange as contributing to the profile of patterns and processes present in molecular sequence evolution. I also view the future of the emerging field of synthetic biology for such diverse applications as food science and chemical manufacturing as a favorable area for MH studies. By understanding the available genetic landscape, in terms of diversity of molecular sequence and diversity of molecular function, novel application can be enabled to look beyond current limitations of thought in the mutational process to access targeted functions. Lastly, as knowledge around the plasticity and emergence of the codon table, both in its standard (universal) form and its additional modified forms, expands, I can view MH components as an evolutionary preserved feature of stable genetic codes and not as a vestigial organ. I can also review and speculate on the context and convergence of the codon table as triplets, with current work in the synthetic biology space where primordial

genetic codons codes may have been efficiently organized non-triplet-based systems, the so called Hachimoji codon table (Hoshika et al 2019). As the codon table itself represents a simple mapping between the less populated amino acid space, and the more populated codon space, I can speculate on synthetically engineered approaches taking advantage of the accessibility of high degeneracy (MH enabled) codon table codes to foster biological innovation.

## BIBLIOGRAPHY


Abdool Karim SS, de Oliveira T. New SARS-CoV-2 Variants - Clinical, Public Health, and Vaccine Implications. N Engl J Med. 2021 May 13;384(19):1866–8.

Ackerley S, James PA, Kalli A, French S, Davies KE, Talbot K. A mutation in the small heat-shock protein HSPB1 leading to distal hereditary motor neuronopathy disrupts neurofilament assembly and the axonal transport of specific cellular cargoes. Hum Mol Genet. 2006 Jan 15;15(2):347–54.

Addetia A, Lieberman NAP, Phung Q, Hsiang TY, Xie H, Roychoudhury P, et al. SARS-CoV-2 ORF6 Disrupts Bidirectional Nucleocytoplasmic Transport through Interactions with Rae1 and Nup98. mBio. 2021 Apr 13;12(2):e00065-21.

Adriaenssens E, Geuens T, Baets J, Echaniz-Laguna A, Timmerman V. Novel insights in the disease biology of mutant small heat shock proteins in neuromuscular diseases. Brain. 2017 Oct 1;140(10):2541–9.

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018 Jul 2;46(W1):W537–44.

Almeida-Souza L, Asselbergh B, d'Ydewalle C, Moonens K, Goethals S, de Winter V, et al. Small heat-shock protein HSPB1 mutants stabilize microtubules in Charcot-Marie-Tooth neuropathy. J Neurosci. 2011 Oct 26;31(43):15320–8.

Almeida-Souza L, Goethals S, de Winter V, Dierick I, Gallardo R, Van Durme J, et al. Increased monomerization of mutant HSPB1 leads to protein hyperactivity in Charcot-Marie-Tooth neuropathy. J Biol Chem. 2010 Apr 23;285(17):12778–86.

Anastasia A, Deinhardt K, Chao M, Will N, Irmady K, Lee F. Val66Met polymorphism of BDNF alters prodomain structure to induce neuronal growth cone retraction. Nat Comm [Internet]. 2013;4. Available from: https://doi.org/10.1038/ncomms3490

Angelucci F, Brene S, Mathe A. BDNF in schizophrenia, depression and corresponding animal models. Mol Psychiatry [Internet]. 2005;10. Available from: https://doi.org/10.1038/sj.mp.4001637

Arenas M. Trends in substitution models of molecular evolution. Front Genet. 2015;6:319.

Atchley WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. Proceedings of the National Academy of Sciences. 2005 May 3;102(18):6395–400.

Averof M, Rokas A, Wolfe KH, Sharp PM. Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science. 2000 Feb;287(5456):1283–6.

Bazykin GA, Kondrashov FA, Ogurtsov AY, Sunyaev S, Kondrashov AS. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. Nature. 2004 Jun;429(6991):558–62.

Benedetti S, Previtali SC, Coviello S, Scarlato M, Cerri F, Di Pierri E, et al. Analyzing histopathological features of rare charcot-marie-tooth neuropathies to unravel their pathogenesis. Arch Neurol. 2010 Dec;67(12):1498–505.

Benndorf R, Martin JL, Kosakovsky Pond SL, Wertheim JO. Neuropathy- and myopathy-associated mutations in human small heat shock proteins: Characteristics and evolutionary history of the mutation sites. Mutat Res Rev Mutat Res. 2014 Sep;761:15–30.

Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. The 2019-new coronavirus epidemic: Evidence for virus evolution. J Med Virol. 2020 Apr;92(4):455–9.

Björkholm C, Monteggia L. BDNF–a key transducer of antidepressant effects. Neuropharmacol [Internet]. 2016;102. Available from: https://doi.org/10.1016/j.neuropharm.2015.10.034

Black IB. Trophic regulation of synaptic plasticity. J Neurobiol [Internet]. 1999;41. Available from: https://doi.org/3.0.CO;2-U

Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, et al. Natural selection on genes that underlie human disease susceptibility. Curr Biol. 2008 Jun 24;18(12):883–9.

Boelens WC. Structural aspects of the human small heat shock proteins related to their functional activities. Cell Stress Chaperones. 2020 Jul;25(4):581–91.

Boni MF, Lemey P, Jiang X, Lam TTY, Perry BW, Castoe TA, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol. 2020 Nov;5(11):1408–17.

Bouhy D, Juneja M, Katona I, Holmgren A, Asselbergh B, De Winter V, et al. A knock-in/knock-out mouse model of HSPB8-associated distal hereditary motor neuropathy

and myopathy reveals toxic gain-of-function of mutant Hspb8. Acta Neuropathol. 2018 Jan;135(1):131–48.

Brady JP, Garland DL, Green DE, Tamm ER, Giblin FJ, Wawrousek EF. AlphaB-crystallin in lens development and muscle integrity: a gene knockout approach. Invest Ophthalmol Vis Sci. 2001 Nov;42(12):2924–34.

Brodehl A, Gaertner-Rommel A, Klauke B, Grewe SA, Schirmer I, Peterschröder A, et al. The novel αB-crystallin (CRYAB) mutation p.D109G causes restrictive cardiomyopathy. Hum Mutat. 2017 Aug;38(8):947–52.

Bugiardini E, Rossor AM, Lynch DS, Swash M, Pittman AM, Blake JC, et al. Homozygous mutation in HSPB1 causing distal vacuolar myopathy and motor neuropathy. Neurol Genet. 2017 Aug;3(4):e168.

Capponi S, Geroldi A, Fossa P, Grandis M, Ciotti P, Gulli R, et al. HSPB1 and HSPB8 in inherited neuropathies: study of an Italian cohort of dHMN and CMT2 patients. J Peripher Nerv Syst. 2011 Dec;16(4):287–94.

Capponi S, Geuens T, Geroldi A, Origone P, Verdiani S, Cichero E, et al. Molecular Chaperones in the Pathogenesis of Amyotrophic Lateral Sclerosis: The Role of HSPB1. Hum Mutat. 2016 Nov;37(11):1202–8.

Carmen Cardenas-Aguayo M, Kazim S, Grundke-Iqbal I, Iqbal K. Neurogenic and neurotrophic effects of BDNF peptides in mouse hippocampal primary neuronal cell cultures. PLoS ONE [Internet]. 2013;8. Available from: https://doi.org/10.1371/journal.pone.0053596

Carra S, Alberti S, Benesch JLP, Boelens W, Buchner J, Carver JA, et al. Small heat shock proteins: multifaceted proteins with important implications for life. Cell Stress Chaperones. 2019 Mar;24(2):295–308.

Carra S, Boncoraglio A, Kanon B, Brunsting JF, Minoia M, Rana A, et al. Identification of the Drosophila ortholog of HSPB8: implication of HSPB8 loss of function in protein folding diseases. J Biol Chem. 2010 Nov 26;285(48):37811–22.

Castellani V, Bolz J. Opposing roles for neurotrophin-3 in targeting and collateral formation of distinct sets of developing cortical neurons. Development [Internet]. 1999;126. Available from: https://doi.org/10.1242/dev.126.15.3335

Chen B, Dowlatshahi D, MacQueen G, Wang J, Young L. Increased hippocampal BDNF immunoreactivity in subjects treated with antidepressant medication. Biol

Psychiatry [Internet]. 2001;50. Available from: https://doi.org/10.1016/S0006-3223(01)01083-6

Chen JM, Cooper DN, Férec C. A new and more accurate estimate of the rate of concurrent tandem-base substitution mutations in the human germline: 0.4% of the single-nucleotide substitution mutation rate. Hum Mutat. 2014 Mar;35(3):392–4.

Chen J, Lu Z, Yang X, Zhou Y, Gao J, Zhang S, et al. Severe Acute Respiratory Syndrome Coronavirus 2 ORF8 Protein Inhibits Type I Interferon Production by Targeting HSP90B1 Signaling. Front Cell Infect Microbiol. 2022;12:899546.

Chen J, Liu X, Xu Y, Fan D. [Rare variants of HSPB1 are probably associated with amyotrophic lateral sclerosis]. Nan Fang Yi Ke Da Xue Xue Bao. 2021 Jan 30;41(1):75–8.

Chen N, Li G, Liszewski MK, Atkinson JP, Jahrling PB, Feng Z, et al. Virulence differences between monkeypox virus isolates from West Africa and the Congo basin. Virology. 2005 Sep 15;340(1):46–63.

Chen Z. Y, Leraci A, Teng H, Dall H, Meng C, Herrera D, et al. Sortilin controls intracellular sorting of brain-derived neurotrophic factor to the regulated secretory pathway. J Neurosci [Internet]. 2005;25. Available from: https://doi.org/10.1523/JNEUROSCI.1017-05.2005

Chen Z. Y, Patel P, Sant G, Meng C, Teng K, Hempstead B. Variant brain-derived neurotrophic factor (BDNF)(Met66) alters the intracellular trafficking and activity-dependent secretion of wild-type BDNF in neurosecretory cells and cortical neurons. J Neurosci [Internet]. 2004;24. Available from: https://doi.org/10.1523/JNEUROSCI.0348-04.2004

Cheng Y, Ji C, Han N, Li J, Xu L, Chen Z, et al. covSampler: A subsampling method with balanced genetic diversity for large-scale SARS-CoV-2 genome data sets. Virus Evolution. 2022 Jul 1;8(2):veac071.

Chiaruttini C, Vicario A, Li Z, Baj G, Braiuca P, Wu Y. Dendritic trafficking of BDNF mRNA is mediated by translin and blocked by the G196A (Val66Met) mutation. Proc Nat Acad Sci USA [Internet]. 2009;106. Available from: https://doi.org/10.1073/pnas.0902833106

Chung KW, Kim SB, Cho SY, Hwang SJ, Park SW, Kang SH, et al. Distal hereditary motor neuropathy in Korean patients with a small heat shock protein 27 mutation. Exp Mol Med. 2008 Jun 30;40(3):304–12.

Comai L. The advantages and disadvantages of being polyploid. Nat Rev Genet. 2005 Nov;6(11):836–46.

Correia C, Coutinho A, Sequeira A, Sousa I, Venda L, Almeida J. Increased BDNF levels and NTRK2 gene association suggest a disruption of BDNF/TrkB signaling in autism. Genes Brain Behav [Internet]. 2010;9. Available from: https://doi.org/10.1111/j.1601-183X.2010.00627.x

Costa CFS, Barbosa AJM, Dias AMGC, Roque ACA. Native, engineered and de novo designed ligands targeting the SARS-CoV-2 spike protein. Biotechnol Adv. 2022 Oct;59:107986.

d'Ydewalle C, Krishnan J, Chiheb DM, Van Damme P, Irobi J, Kozikowski AP, et al. HDAC6 inhibitors reverse axonal loss in a mouse model of mutant HSPB1-induced Charcot-Marie-Tooth disease. Nat Med. 2011 Jul 24;17(8):968–74.

Datskevich PN, Nefedova VV, Sudnitsyna MV, Gusev NB. Mutations of small heat shock proteins and human congenital diseases. Biochemistry (Mosc). 2012 Dec;77(13):1500–14.

de Jong WW, Caspers GJ, Leunissen JA. Genealogy of the alpha-crystallin--small heat-shock protein superfamily. Int J Biol Macromol. 1998 Jun;22(3–4):151–62.

De Maio N, Holmes I, Schlötterer C, Kosiol C. Estimating empirical codon hidden Markov models. Mol Biol Evol. 2013 Mar;30(3):725–36.

Dearlove B, Lewitus E, Bai H, Li Y, Reeves DB, Joyce MG, et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. Proc Natl Acad Sci U S A. 2020 Sep 22;117(38):23652–62.

den Engelsman J, Keijsers V, de Jong WW, Boelens WC. The small heat-shock protein alpha B-crystallin promotes FBX4-dependent ubiquitination. J Biol Chem. 2003 Feb 14;278(7):4699–704.

Dierick I, Baets J, Irobi J, Jacobs A, De Vriendt E, Deconinck T, et al. Relative contribution of mutations in genes for autosomal dominant distal hereditary motor neuropathies: a genotype-phenotype correlation study. Brain. 2008 May;131(Pt 5):1217–27.

Domingo JL. What we know and what we need to know about the origin of SARS-CoV-2. Environ Res. 2021 Sep;200:111785.

Doron-Faigenboim A, Pupko T. A combined empirical and mechanistic codon model. Mol Biol Evol. 2007 Feb;24(2):388–97.

Dunn KA, Kenney T, Gu H, Bielawski JP. Improved inference of site-specific positive selection under a generalized parametric codon model when there are multinucleotide mutations and multiple nonsynonymous rates. BMC Evolutionary Biology. 2019 Dec;19(1):22.

Echaniz-Laguna A, Geuens T, Petiot P, Péréon Y, Adriaenssens E, Haidar M, et al. Axonal Neuropathies due to Mutations in Small Heat Shock Proteins: Clinical, Genetic, and Functional Insights into Novel Mutations. Hum Mutat. 2017 May;38(5):556–68.

Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall. 2017 Jan;1(1):33–46.

Enard D, Cai L, Gwennap C, Petrov DA. Viruses are a dominant driver of protein adaptation in mammals. eLife. 2016 May;5.

Evgrafov OV, Mersiyanova I, Irobi J, Van Den Bosch L, Dierick I, Leung CL, et al. Mutant small heat-shock protein 27 causes axonal Charcot-Marie-Tooth disease and distal hereditary motor neuropathy. Nat Genet. 2004 Jun;36(6):602–6.

Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science. 2021 May 21;372(6544):815–21.

Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368–76.

Finn RD, Coggill P, Eberhardt R, Eddy S, Mistry J, Mitchell A. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res [Internet]. 2016;44. Available from: https://doi.org/10.1093/nar/gkv1344

Fletcher W, Yang Z. INDELible: A flexible simulator of biological sequence evolution. Molecular Biology and Evolution. 2009;26(8):1879–88.

Fontaine JM, Rest JS, Welsh MJ, Benndorf R. The sperm outer dense fiber protein is the 10th member of the superfamily of mammalian small stress proteins. Cell Stress Chaperones. 2003;8(1):62–9.

Fontaine JM, Sun X, Hoppe AD, Simon S, Vicart P, Welsh MJ, et al. Abnormal small heat shock protein interactions involving neuropathy-associated HSP22 (HSPB8) mutants. FASEB J. 2006 Oct;20(12):2168–70.

Fortunato F, Neri M, Geroldi A, Bellone E, De Grandis D, Ferlini A, et al. A CMT2 family carrying the P7R mutation in the N- terminal region of the HSPB1 gene. Clin Neurol Neurosurg. 2017 Dec;163:15–7.

Franck E, Madsen O, van Rheede T, Ricard G, Huynen MA, de Jong WW. Evolutionary diversity of vertebrate small heat shock proteins. J Mol Evol. 2004 Dec;59(6):792–805.

Frasquet M, Rojas-García R, Argente-Escrig H, Vázquez-Costa JF, Muelas N, Vílchez JJ, et al. Distal hereditary motor neuropathies: Mutation spectrum and genotype-phenotype correlation. Eur J Neurol. 2021 Apr;28(4):1334–43.

Frost SDW, Wrin T, Smith DM, Kosakovsky Pond SL, Liu Y, Paxinos E, et al. Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection. Proc Natl Acad Sci U S A. 2005 Dec 20;102(51):18514–9.

Gaeta M, Mileto A, Mazzeo A, Minutoli F, Di Leo R, Settineri N, et al. MRI findings, patterns of disease distribution, and muscle fat fraction calculation in five patients with Charcot-Marie-Tooth type 2 F disease. Skeletal Radiol. 2012 May;41(5):515–24.

Geuens T, De Winter V, Rajan N, Achsel T, Mateiu L, Almeida-Souza L, et al. Mutant HSPB1 causes loss of translational repression by binding to PCBP1, an RNA binding protein with a possible role in neurodegenerative disease. Acta Neuropathol Commun. 2017 Jan 11;5(1):5.

Ghaoui R, Palmio J, Brewer J, Lek M, Needham M, Evilä A, et al. Mutations in HSPB8 causing a new phenotype of distal myopathy and motor neuropathy. Neurology. 2016 Jan 26;86(4):391–8.

Gibbs MJ, Armstrong JS, Gibbs AJ. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics [Internet]. 2000;16. Available from: https://doi.org/10.1093/bioinformatics/16.7.573

Giza JI, Kim J, Meyer H, Anastasia A, Dincheva I, Zheng C. The BDNF Val66Met prodomain disassembles dendritic spines altering fear extinction circuitry and behavior. Neuron [Internet]. 2018;99. Available from: https://doi.org/10.1016/j.neuron.2018.05.024

Glerup S, Bolcho U, Molgaard S, Boggild S, Vaegter C, Smith A. SorCS2 is required for BDNF-dependent plasticity in the hippocampus. Mol Psychiatry [Internet]. 2016;21. Available from: https://doi.org/10.1038/mp.2016.108

Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular biology and evolution. 1994;11(5):725–36.

Gratacòs M, Gonzalez J, Mercader J, Cid R, Urretavizcaya, Estivill X. Brain-derived neurotrophic factor Val66Met and psychiatric disorders: meta-analysis of case-control studies confirm association to substance-related disorders, eating disorders, and schizophrenia. Biol Psychiatry [Internet]. 2007;61. Available from: https://doi.org/10.1016/j.biopsych.2006.08.025

Gray K, Ellis V. Activation of pro-BDNF by the pericellular serine protease plasmin. FEBS Lett [Internet]. 2008;582. Available from: https://doi.org/10.1016/j.febslet.2008.02.026

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018 Dec 1;34(23):4121–3.

Haidar M, Asselbergh B, Adriaenssens E, De Winter V, Timmermans JP, Auer-Grumbach M, et al. Neuropathy-causing mutations in HSPB1 impair autophagy by disturbing the formation of SQSTM1/p62 bodies. Autophagy. 2019 Jun;15(6):1051–68.

Halfmann PJ, Kuroda M, Armbrust T, Theiler J, Balaram A, Moreno GK, et al. Characterization of the SARS-CoV-2 B.1.621 (Mu) variant. Science Translational Medicine. 2022 May 17;14(657):eabm4908.

Halpern AL, Bruno WJ. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol. 1998 Jul;15(7):910–7.

Hamed SM, Elkhatib WF, Khairalla AS, Noreddin AM. Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology. Sci Rep. 2021 Apr 19;11(1):8435.

Harris K, Nielsen R. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Research. 2014 Sep;24(9):1445–54.

Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. 2021 Jul;19(7):409–24.

Hedges BS, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics, [Internet]. 2006;22. Available from: https://doi.org/10.1093/bioinformatics/btl505

Ho CC, Tai SM, Lee ECN, Mak TSH, Liu TKT, Tang VWL, et al. Rapid Identification of Pathogenic Variants in Two Cases of Charcot-Marie-Tooth Disease by Gene-Panel Sequencing. Int J Mol Sci. 2017 Apr 5;18(4):E770.

Holmgren A, Bouhy D, De Winter V, Asselbergh B, Timmermans JP, Irobi J, et al. Charcot-Marie-Tooth causing HSPB1 mutations increase Cdk5-mediated phosphorylation of neurofilaments. Acta Neuropathol. 2013 Jul;126(1):93–108.

Horch HW, Kruttgen A, Portbury S, Katz LC. Destabilization of cortical dendrites and spines by BDNF. Neuron [Internet]. 1999;23. Available from: https://doi.org/10.1016/S0896-6273(00)80785-0

Hoshika S, Leal NA, Kim MJ, Kim MS, Karalkar NB, Kim HJ, et al. Hachimoji DNA and RNA: A genetic system with eight building blocks. Science. 2019 Feb 22;363(6429):884–7.

Houlden H, Laura M, Wavrant-De Vrièze F, Blake J, Wood N, Reilly MM. Mutations in the HSP27 (HSPB1) gene cause dominant, recessive, and sporadic distal HMN/CMT type 2. Neurology. 2008 Nov 18;71(21):1660–8.

Howells D, Porritt M, Wong J, Batchelor P, Kalnins R, Hughes A, et al. Reduced BDNF mRNA expression in the Parkinson's disease substantia nigra. Exp Neurol [Internet]. 2000;166. Available from: https://doi.org/10.1006/exnr.2000.7483

Huang L, Min JN, Masters S, Mivechi NF, Moskophidis D. Insights into function and regulation of small heat shock protein 25 (HSPB1) in a mouse model with targeted gene disruption. Genesis. 2007 Aug;45(8):487–501.

Ikeda Y, Abe A, Ishida C, Takahashi K, Hayasaka K, Yamada M. A clinical phenotype of distal hereditary motor neuronopathy type II with a novel HSPB1 mutation. J Neurol Sci. 2009 Feb 15;277(1–2):9–12.

Inagaki N, Hayashi T, Arimura T, Koga Y, Takahashi M, Shibata H, et al. Alpha B-crystallin mutation in dilated cardiomyopathy. Biochem Biophys Res Commun. 2006 Apr 7;342(2):379–86.

Irobi J, Holmgren A, De Winter V, Asselbergh B, Gettemans J, Adriaensen D, et al. Mutant HSPB8 causes protein aggregates and a reduced mitochondrial membrane potential in dermal fibroblasts from distal hereditary motor neuropathy patients. Neuromuscul Disord. 2012 Aug;22(8):699–711.

Irobi J, Van Impe K, Seeman P, Jordanova A, Dierick I, Verpoorten N, et al. Hot-spot residue in small heat-shock protein 22 causes distal motor neuropathy. Nat Genet. 2004 Jun;36(6):597–601.

Isidro J, Borges V, Pinto M, Sobral D, Santos JD, Nunes A, et al. Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus. Nat Med. 2022 Aug;28(8):1569–72.

J C, X L, Y X, D F. [Rare variants of HSPB1 are probably associated with amyotrophic lateral sclerosis]. Nan fang yi ke da xue xue bao = Journal of Southern Medical University [Internet]. 2021 Jan 30 [cited 2022 Nov 3];41(1). Available from: https://pubmed.ncbi.nlm.nih.gov/33509756/

J I, K VI, P S, A J, I D, N V, et al. Hot-spot residue in small heat-shock protein 22 causes distal motor neuropathy. Nature genetics [Internet]. 2004 Jun [cited 2022 Nov 3];36(6). Available from: https://pubmed.ncbi.nlm.nih.gov/15122253/

James PA, Rankin J, Talbot K. Asymmetrical late onset motor neuropathy associated with a novel mutation in the small heat shock protein HSPB1 (HSP27). J Neurol Neurosurg Psychiatry. 2008 Apr;79(4):461–3.

Jl E, M DP, Ar P, Ka J, Gh K, Ra Q. The specificity of the interaction between αB-crystallin and desmin filaments and its impact on filament aggregation and cell viability. Philosophical transactions of the Royal Society of London Series B, Biological sciences [Internet]. 2013 Mar 25 [cited 2022 Nov 3];368(1617). Available from: https://pubmed.ncbi.nlm.nih.gov/23530264/

Jones CT, Youssef N, Susko E, Bielawski JP. Phenomenological Load on Model Parameters Can Lead to False Biological Conclusions. Mol Biol Evol. 2018 Jun;35(6):1473–88.

Jones CT, Youssef N, Susko E, Bielawski JP. A Phenotype–Genotype Codon Model for Detecting Adaptive Evolution. Systematic Biology. 2019 Dec;69(4):722–38.

K N, Jm P, Yj K, Br Y, Jh Y, H K, et al. A novel Lys141Thr mutation in small heat shock protein 22 (HSPB8) gene in Charcot-Marie-Tooth disease type 2L. Neuromuscular disorders : NMD [Internet]. 2013 Aug [cited 2022 Nov 3];23(8). Available from: https://pubmed.ncbi.nlm.nih.gov/23796487/

Kampinga HH, Hageman J, Vos MJ, Kubota H, Tanguay RM, Bruford EA, et al. Guidelines for the nomenclature of the human heat shock proteins. Cell Stress Chaperones. 2009 Jan;14(1):105–11.

Kaplanis J, Akawi N, Gallone G, McRae JF, Prigmore E, Wright CF, et al. Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations. Genome research. 2019 Jul;29(7):1047–56.

Kappé G, Boelens WC, de Jong WW. Why proteins without an alpha-crystallin domain should not be included in the human small heat shock protein family HSPB. Cell Stress Chaperones. 2010 Jul;15(4):457–61.

Kappé G, Franck E, Verschuure P, Boelens WC, Leunissen JAM, de Jong WW. The human genome encodes 10 alpha-crystallin-related small heat shock proteins: HspB1-10. Cell Stress Chaperones. 2003;8(1):53–61.

Katoh K, Misawa K, Kuma K ichi, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul;30(14):3059–66.

Katz M, Davis M, Garton FC, Henderson R, Bharti V, Wray N, et al. Mutations in heat shock protein beta-1 (HSPB1) are associated with a range of clinical phenotypes related to different patterns of motor neuron dysfunction: A case series. J Neurol Sci. 2020 Jun 15;413:116809.

Kern DM, Sorum B, Mali SS, Hoel CM, Sridharan S, Remis JP, et al. Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs. Nat Struct Mol Biol. 2021 Jul;28(7):573–82.

Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, et al. GISAID's Role in Pandemic Response. China CDC Wkly. 2021 Dec 3;3(49):1049–51.

Kijima K, Numakura C, Goto T, Takahashi T, Otagiri T, Umetsu K, et al. Small heat shock protein 27 mutation in a Japanese patient with distal hereditary motor neuropathy. J Hum Genet. 2005;50(9):473–6.

Kim YK, Lee H, Won S, Park E, Lee H, Lee B. Low plasma BDNF is associated with suicidal behavior in major depression. Prog Neuropsychopharmacol Biol Psychiatry [Internet]. 2007;31. Available from: https://doi.org/10.1016/j.pnpbp.2006.06.024

Kingwell K. BDNF copycats. Nat Rev Drug Disco [Internet]. 2010;9. Available from: https://doi.org/10.1038/nrd3190

Kolb SJ, Snyder PJ, Poi EJ, Renard EA, Bartlett A, Gu S, et al. Mutant small heat shock protein B3 causes motor neuropathy: utility of a candidate gene approach. Neurology. 2010 Feb 9;74(6):502–6.

Kosakovsky Pond SL, Wisotsky SR, Escalante A, Magalis BR, Weaver S. Contrast-FEL-A test for differences in selective pressures at individual sites among clades and sets of branches. Mol Biol Evol [Internet]. 2021;38. Available from: https://doi.org/10.1093/molbev/msaa263

Kosakovsky Pond S, Delport W, Muse SV, Scheffler K. Correcting the Bias of Empirical Frequency Parameter Estimators in Codon Models. Mailund T, editor. PLoS ONE. 2010 Jul;5(7):e11230.

Kosakovsky Pond SL, Frost SDW, Grossman Z, Gravenor MB, Richman DD, Brown AJL. Adaptation to Different Human Populations by HIV-1 Revealed by Codon-Based Analyses. PLoS Comput Biol. 2006 Jun;2(6):e62.

Kosakovsky Pond SL, Frost SDW, Muse SV. HyPhy: Hypothesis testing using phylogenies. Bioinformatics. 2005 Mar;21(5):676–9.

Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005 May;22(5):1208–22.

Kosakovsky Pond SL, Muse SV. Site-to-site variation of synonymous substitution rates. Molecular Biology and Evolution. 2005 Dec;22(12):2375–85.

Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, Frost SDW. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. Mol Biol Evol. 2008 Sep;25(9):1809–24.

Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Mol Biol Evol. 2020 Jan 1;37(1):295–9.

Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, Frost SDW. A Maximum Likelihood Method for Detecting Directional Evolution in Protein Sequences and Its Application to Influenza A Virus. Mol Biol Evol. 2008 Sep;25(9):1809–24.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm. Molecular Biology and Evolution. 2006 Oct 1;23(10):1891–901.

Kosakovsky Pond SL, Wisotsky SR, Escalante A, Magalis BR, Weaver S. Contrast-FEL-A Test for Differences in Selective Pressures at Individual Sites among Clades and Sets of Branches. Mol Biol Evol. 2021 Mar 9;38(3):1184–98.

Kosiol C, Holmes I, Goldman N. An empirical codon model for protein sequence evolution. Molecular Biology and Evolution. 2007 Mar;24(7):1464–79.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019 Nov 1;35(21):4453–5.

Kriehuber T, Rattei T, Weinmaier T, Bepperling A, Haslbeck M, Buchner J. Independent evolution of the core domain and its flanking sequences in small heat shock proteins. FASEB J. 2010 Oct;24(10):3633–42.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. Computational methods for Gene Orthology inference. Brief Bioinform. 2011 Sep;12(5):379–91.

Kuczewski N, Porcher C, Lessmann V, Medina I, Gaiarsa JL. Activity-dependent dendritic release of BDNF and biological consequences. Mol Neurobiol [Internet]. 2009;39. Available from: https://doi.org/10.1007/s12035-009-8050-7

Kumar S, Stecher G, Suleski M, Hedges B. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol [Internet]. 2017;34. Available from: https://doi.org/10.1093/molbev/msx116

Kuraku S, Feiner N, Keeley SD, Hara Y. Incorporating tree-thinking and evolutionary time scale into developmental biology. Dev Growth Differ. 2016 Jan;58(1):131–42.

Lam HM, Ratmann O, Boni MF. Improved algorithmic complexity for the 3SEQ recombination detection algorithm. Mol Biol Evolution [Internet]. 2018;35. Available from: https://doi.org/10.1093/molbev/msx263

Laššuthová P, Šafka Brožková D, Krůtová M, Neupauerová J, Haberlová J, Mazanec R, et al. Improving diagnosis of inherited peripheral neuropathies through gene panel analysis. Orphanet J Rare Dis. 2016 Aug 22;11(1):118.

Lewis-Smith DJ, Duff J, Pyle A, Griffin H, Polvikoski T, Birchall D, et al. Novel HSPB1 mutation causes both motor neuronopathy and distal myopathy. Neurol Genet. 2016 Dec;2(6):e110.

Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong XP, et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Sci Adv. 2020 Jul;6(27):eabb9153.

Lin KP, Soong BW, Yang CC, Huang LW, Chang MH, Lee IH, et al. The mutational spectrum in a cohort of Charcot-Marie-Tooth disease type 2 among the Han Chinese in Taiwan. PLoS One. 2011;6(12):e29393.

Lipovich L, Dachet F, Cai J, Bagla S, Balan K, Jia H. Activity-dependent human brain coding/noncoding gene regulatory networks. Genetics [Internet]. 2012;192. Available from: https://doi.org/10.1534/genetics.112.145128

Liu M, Ke T, Wang Z, Yang Q, Chang W, Jiang F, et al. Identification of a CRYAB mutation associated with autosomal dominant posterior polar cataract in a Chinese family. Invest Ophthalmol Vis Sci. 2006 Aug;47(8):3461–6.

Liu Y, Zhang X, Luo L, Wu M, Zeng R, Cheng G, et al. A novel alphaB-crystallin mutation associated with autosomal dominant congenital lamellar cataract. Invest Ophthalmol Vis Sci. 2006 Mar;47(3):1069–75.

Lorenzo-Redondo R, Fryer HR, Bedford T, Kim EY, Archer J, Pond SLK, et al. Persistent HIV-1 replication maintains the tissue reservoir during therapy. Nature. 2016 Feb 4;530(7588):51–6.

Lucaci AG, Wisotsky SR, Shank SD, Weaver S, Kosakovsky Pond SL. Extra base hits: widespread empirical support for instantaneous multiple-nucleotide changes. PLoS ONE [Internet]. 2021;16. Available from: https://doi.org/10.1371/journal.pone.0248337

Luchsinger LL, Hillyer CD. Vaccine efficacy probable against COVID-19 variants. Science. 2021 Mar 12;371(6534):1116.

Luigetti M, Fabrizi GM, Madia F, Ferrarini M, Conte A, Del Grande A, et al. A novel HSPB1 mutation in an Italian patient with CMT2/dHMN phenotype. J Neurol Sci. 2010 Nov 15;298(1–2):114–7.

Luigetti M, Fabrizi GM, Bisogni G, Romano A, Taioli F, Ferrarini M, et al. Charcot-Marie-Tooth type 2 and distal hereditary motor neuropathy: Clinical, neurophysiological and genetic findings from a single-centre experience. Clin Neurol Neurosurg. 2016 May;144:67–71.

Lytras S, Hughes J, Martin D, Swanepoel P, de Klerk A, Lourens R, et al. Exploring the Natural Origins of SARS-CoV-2 in the Light of Recombination. Genome Biology and Evolution. 2022 Feb 1;14(2):evac018.

MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol. 2021 Mar;19(3):e3001115.

Maeda K, Idehara R, Hashiguchi A, Takashima H. A family with distal hereditary motor neuropathy and a K141Q mutation of small heat shock protein HSPB1. Intern Med. 2014;53(15):1655–8.

Maher MC, Bartha I, Weaver S, di Iulio J, Ferri E, Soriaga L, et al. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Sci Transl Med. 2022 Feb 23;14(633):eabk3445.

Maisonpierre PC, Beau M, Espinosa R, Ip N, Belluscio L, Monte S. Human and rat brain-derived neurotrophic factor and neurotrophin-3: gene structures, distributions, and chromosomal localizations. Genomics [Internet]. 1991;10. Available from: https://doi.org/10.1016/0888-7543(91)90436-I

Makarenkov V, Mazoure B, Rabusseau G, Legendre P. Horizontal gene transfer and recombination analysis of SARS-CoV-2 genes helps discover its close relatives and shed light on its origin. BMC Ecol Evol. 2021 Jan 21;21(1):5.

Maloyan A, Sanbe A, Osinska H, Westfall M, Robinson D, Imahashi K ichi, et al. Mitochondrial dysfunction and apoptosis underlie the pathogenic process in alpha-B-crystallin desmin-related cardiomyopathy. Circulation. 2005 Nov 29;112(22):3451–61.

Mannino F, Wisotsky S, Kosakovsky Pond SL, Muse SV. Equiprobable discrete models of site-specific substitution rates underestimate the extent of rate variability. PLoS One. 2020;15(3):e0229493.

Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: detection and analysis of recombination patterns in virus genomes. Virus Evol [Internet]. 2015;1. Available from: https://doi.org/10.1093/ve/vev003

Martin DP, Posada D, Crandall KA, Williamson C. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses [Internet]. 2005;21. Available from: https://doi.org/10.1089/aid.2005.21.98

Martin D, Rybicki E. RDP: detection of recombination amongst aligned sequences. Bioinformatics [Internet]. 2000;16. Available from: https://doi.org/10.1093/bioinformatics/16.6.562

Martin DP, Varsani A, Roumagnac P, Botha G, Maslamoney S, Schwab T, et al. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. Virus Evol. 2020 Apr 12;7(1):veaa087.

Martin DP, Lytras S, Lucaci AG, Maier W, Grüning B, Shank SD, et al. Selection Analysis Identifies Clusters of Unusual Mutational Changes in Omicron Lineage BA.1 That Likely Impact Spike Function. Mol Biol Evol. 2022 Apr 11;39(4):msac061.

Martin DP, Weaver S, Tegally H, San JE, Shank SD, Wilkinson E, et al. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. Cell. 2021 Sep 30;184(20):5189-5200.e7.

Martinowich K, Manji H, Lu B. New insights into BDNF function in depression and anxiety. Nat Neurosci [Internet]. 2007;10. Available from: https://doi.org/10.1038/nn1971

Massa SM, Yang T, Xie Y, Shi J, Bilgen M, Joyce J. Small molecule BDNF mimetics activate TrkB signaling and prevent neuronal degeneration in rodents. J Clin Investig [Internet]. 2010;120. Available from: https://doi.org/10.1172/JCI41356

Maynard Smith J. Analyzing the mosaic structure of genes. J Mol Evol. 1992;34.

McCallum M, Bassi J, De Marco A, Chen A, Walls AC, Di Iulio J, et al. SARS-CoV-2 immune evasion by the B.1.427/B.1.429 variant of concern. Science. 2021 Aug 6;373(6555):648–54.

Minh BQ, Schmidt HA, Chernomor O, Schremph D, Haeseler A. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37.

Miyazawa S. Selective Constraints on Amino Acids Estimated by a Mechanistic Codon Substitution Model with Multiple Nucleotide Changes. Martin D, editor. PLoS ONE. 2011 Mar;6(3):e17244.

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10:33.

Morelli FF, Verbeek DS, Bertacchini J, Vinet J, Mediani L, Marmiroli S, et al. Aberrant Compartment Formation by HSPB2 Mislocalizes Lamin A and Compromises Nuclear Integrity and Function. Cell Rep. 2017 Aug 29;20(9):2100–15.

Moretti S, Laurenczy B, Gharib WH, Castella B, Kuzniar A, Schabauer H, et al. Selectome update: quality control and computational improvements to a database of positive selection. Nucleic Acids Research. 2014 Jan;42(Database issue):917–21.

Mp B, O Y, Q H, L D, Da H, Pl S, et al. Mutation R120G in alphaB-crystallin, which is linked to a desmin-related myopathy, results in an irregular structure and defective

chaperone-like function. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 1999 May 25 [cited 2022 Nov 3];96(11). Available from: https://pubmed.ncbi.nlm.nih.gov/10339554/

Muranova LK, Sudnitsyna MV, Strelkov SV, Gusev NB. Mutations in HspB1 and hereditary neuropathies. Cell Stress Chaperones. 2020 Jul;25(4):655–65.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet [Internet]. 2012;8. Available from: https://doi.org/10.1371/journal.pgen.1002764

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, et al. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol. 2013 May;30(5):1196–205.

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, et al. Gene-Wide Identification of Episodic Selection. Molecular Biology and Evolution. 2015 May;32(5):1365–71.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012;8(7):e1002764.

Muse SV, Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Molecular Biology and Evolution. 1994;11(5):715–24.

Mymrikov EV, Seit-Nebi AS, Gusev NB. Large potentials of small heat shock proteins. Physiol Rev. 2011 Oct;91(4):1123–59.

Nagappan G, Lu B. Activity-dependent modulation of the BDNF receptor TrkB: mechanisms and implications. Trends Neurosci [Internet]. 2005;28. Available from: https://doi.org/10.1016/j.tins.2005.07.003

Nam DE, Nam SH, Lee AJ, Hong YB, Choi BO, Chung KW. Small heat shock protein B3 (HSPB3) mutation in an axonal Charcot-Marie-Tooth disease family. J Peripher Nerv Syst. 2018 Mar;23(1):60–6.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan;32(1):268–74.

Notaras M, Buuse M. Brain-derived neurotrophic factor (BDNF): novel insights into regulation and genetic variation. Neuroscientist [Internet]. 2019;25. Available from: https://doi.org/10.1177/1073858418810142

Notaras M, Buuse M. Neurobiology of BDNF in fear memory, sensitivity to stress, and stress-related disorders. Mol Psychiatry [Internet]. 2020;25. Available from: https://doi.org/10.1038/s41380-019-0639-2

Notaras M, Hill R, Buuse M. A role for the BDNF gene Val66Met polymorphism in schizophrenia? A comprehensive review. Neurosci Biobehav Rev [Internet]. 2015;51. Available from: https://doi.org/10.1016/j.neubiorev.2014.12.016

Notaras M, Hill R, Buuse M. The BDNF gene Val66Met polymorphism as a modifier of psychiatric disorder susceptibility: progress and controversy. Mol Psychiatry [Internet]. 2015;20. Available from: https://doi.org/10.1038/mp.2015.27

Oberstadt M, Mitter D, Classen J, Baum P. Late onset dHMN II caused by c.404C>G mutation in HSPB1 gene. J Peripher Nerv Syst. 2016 Jun;21(2):111–3.

Okada M, Misumi Y, Ueda M, Yamashita T, Masuda T, Tasaki M, et al. A novel transthyretin variant V28S (p.V48S) with a double-nucleotide substitution in the same codon. Amyloid. 2017 Dec;24(4):231–2.

Ov E, I M, J I, L VDB, I D, Cl L, et al. Mutant small heat-shock protein 27 causes axonal Charcot-Marie-Tooth disease and distal hereditary motor neuropathy. Nature genetics [Internet]. 2004 Jun [cited 2022 Nov 3];36(6). Available from: https://pubmed.ncbi.nlm.nih.gov/15122254/

Padidam M, Sawyer S, Fauquet CM. Possible emergence of new geminiviruses by frequent recombination. Virology [Internet]. 1999;265. Available from: https://doi.org/10.1006/viro.1999.0056

Palasz E, Wysocka A, Gasiorowska A, Chalimoniuk M, Niewiadomski W, Niewiadomski G. BDNF as a promising therapeutic agent in Parkinson's disease. Int J Mol Sci [Internet]. 2020;21. Available from: https://doi.org/10.3390/ijms21031170

Pawlica P, Yario TA, White S, Wang J, Moss WN, Hui P, et al. SARS-CoV-2 expresses a microRNA-like small RNA able to selectively repress host genes. Proc Natl Acad Sci U S A. 2021 Dec 28;118(52):e2116668118.

Perkel JM. Reactive, reproducible, collaborative: computational notebooks evolve. Nature. 2021 May;593(7857):156–7.

Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res. 2012 Jan;40(Database issue):D593-598.

Pilotto A, Marziliano N, Pasotti M, Grasso M, Costante AM, Arbustini E. alphaB-crystallin mutation in dilated cardiomyopathies: low prevalence in a consecutive series of 200 unrelated probands. Biochem Biophys Res Commun. 2006 Aug 11;346(4):1115–7.

Pitts BL, Whealin J, Harpaz-Rotem I, Duman S, Krystal J, Southwick S. BDNF Val66Met polymorphism and posttraumatic stress symptoms in US military veterans: Protective effect of physical exercise. Psychoneuroendocrinol [Internet]. 2019;100. Available from: https://doi.org/10.1016/j.psyneuen.2018.10.011

Pivac N, Kozaric-Kovacic D, Grubisic-Ilic M, Nedic G, Rakos I, Nikolac M. The association between brain-derived neurotrophic factor Val66Met variants and psychotic symptoms in posttraumatic stress disorder. World J Biol Psychiatry [Internet]. 2012;13. Available from: https://doi.org/10.3109/15622975.2011.582883

Pond SLK, Frost SDW. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics. 2005 May 15;21(10):2531–3.

Poon AFY, Lewis FI, Frost SDW, Kosakovsky Pond SL. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. Bioinformatics [Internet]. 2008;24. Available from: https://doi.org/10.1093/bioinformatics/btn313

Posada D, Crandall KA. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci USA [Internet]. 2001;98. Available from: https://doi.org/10.1073/pnas.241370698

Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010 Mar 10;5(3):e9490.

Pruunsild P, Kazantseva A, Aid T, Palm K, Timmusk T. Dissecting the human BDNF locus: bidirectional transcription, complex splicing, and multiple promoters. Genomics [Internet]. 2007;90. Available from: https://doi.org/10.1016/j.ygeno.2007.05.004

Pybus OG, Rambaut A, Belshaw R, Freckleton RP, Drummond AJ, Holmes EC. Phylogenetic Evidence for Deleterious Mutation Load in RNA Viruses and Its Contribution to Viral Evolution. Molecular Biology and Evolution. 2007 Mar 1;24(3):845–52.

Quintana-Murci L, Barreiro LB. The role played by natural selection on Mendelian traits in humans. Ann N Y Acad Sci. 2010 Dec;1214:1–17.

Rajasekaran NS, Connell P, Christians ES, Yan LJ, Taylor RP, Orosz A, et al. Human alpha B-crystallin mutation causes oxido-reductive stress and protein aggregation cardiomyopathy in mice. Cell. 2007 Aug 10;130(3):427–39.

Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020 Nov;5(11):1403–7.

Rathnasinghe R, Jangra S, Ye C, Cupic A, Singh G, Martínez-Romero C, et al. Characterization of SARS-CoV-2 Spike mutations important for infection of mice and escape from human immune sera. Nat Commun. 2022 Jul 7;13(1):3921.

Reilich P, Schoser B, Schramm N, Krause S, Schessl J, Kress W, et al. The p.G154S mutation of the alpha-B crystallin gene (CRYAB) causes late-onset distal myopathy. Neuromuscul Disord. 2010 Apr;20(4):255–9.

Ricci S, Businaro R, Ippoliti F, Lo Vasco V, Massoni F, Onofri E. Altered cytokine and BDNF levels in autism spectrum disorder. Neurotox Res [Internet]. 2013;24. Available from: https://doi.org/10.1007/s12640-013-9393-4

Rodriguez-Tebar A, Dechant G, Barde Y. Binding of brain-derived neurotrophic factor to the nerve growth factor receptor. Neuron [Internet]. 1990;4. Available from: https://doi.org/10.1016/0896-6273(90)90107-Q

Rogozin IB, Belinky F, Pavlenko V, Shabalina SA, Kristensen DM, Koonin EV. Evolutionary switches between two serine codon sets are driven by selection. Proceedings of the National Academy of Sciences of the United States of America. 2016 Nov;113(46):13109–13.

Sacconi S, Féasson L, Antoine JC, Pécheux C, Bernard R, Cobo AM, et al. A novel CRYAB mutation resulting in multisystemic disease. Neuromuscul Disord. 2012 Jan;22(1):66–72.

Sadeh M, Rahat D, Meiner V, Fellig Y, Arad M, Schueler-Furman O, et al. Multi-system neurological disorder associated with a CRYAB variant. Neurogenetics. 2021 May;22(2):117–25.

Safieh LA, Khan AO, Alkuraya FS. Identification of a novel CRYAB mutation associated with autosomal recessive juvenile cataract in a Saudi family. Mol Vis. 2009 May 15;15:980–4.

Sakofsky CJ, Roberts SA, Malc E, Mieczkowski PA, Resnick MA, Gordenin DA, et al. Break-induced replication is a source of mutation clusters underlying kataegis. Cell Rep. 2014 Jun;7(5):1640–8.

Sakuragi S, Tominaga-Yoshino K, Ogura A. Involvement of TrkB-and p75 NTR-signaling pathways in two contrasting forms of long-lasting synaptic plasticity. Sci Rep [Internet]. 2013;3. Available from: https://doi.org/10.1038/srep03185

Samson S, Lord É, Makarenkov V. SimPlot++: a Python application for representing sequence similarity and detecting recombination. Bioinformatics. 2022 Jun 1;38(11):3118–20.

Sanbe A. Molecular mechanisms of α-crystallinopathy and its therapeutic strategy. Biol Pharm Bull. 2011;34(11):1653–8.

Scarlato M, Viganò F, Carrera P, Previtali SC, Bolino A. A novel heat shock protein 27 homozygous mutation: widening of the continuum between MND/dHMN/CMT2. J Peripher Nerv Syst. 2015 Dec;20(4):419–21.

Schrider DR, Hourmozdi JN, Hahn MW. Pervasive multinucleotide mutational events in eukaryotes. Current Biology. 2011;

Self SG, Liang KY. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. J Am Stat Assoc. 1987;82(398):605–310.

Selig EE, Zlatic CO, Cox D, Mok YF, Gooley PR, Ecroyd H, et al. N- and C-terminal regions of αB-crystallin and Hsp27 mediate inhibition of amyloid nucleation, fibril binding, and fibril disaggregation. J Biol Chem. 2020 Jul 17;295(29):9838–54.

Shemetov AA, Gusev NB. Biochemical characterization of small heat shock protein HspB8 (Hsp22)-Bag3 interaction. Arch Biochem Biophys. 2011 Sep 1;513(1):1–9.

Shen L, Bard JD, Triche TJ, Judkins AR, Biegel JA, Gai X. Emerging variants of concern in SARS-CoV-2 membrane protein: a highly conserved target with potential pathological and therapeutic implications. Emerg Microbes Infect. 2021 Dec;10(1):885–93.

Shultz AJ, Sackton TB. Immune genes are hotspots of shared positive selection across birds and mammals. eLife. 2019 Jan;8.

Simon S, Fontaine JM, Martin JL, Sun X, Hoppe AD, Welsh MJ, et al. Myopathy-associated alphaB-crystallin mutants: abnormal phosphorylation, intracellular location, and interactions with other small heat shock proteins. J Biol Chem. 2007 Nov 23;282(47):34276–87.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol [Internet]. 2015;32. Available from: https://doi.org/10.1093/molbev/msv022

Smith NG, Hurst LD. The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate? Genetics. 1999 Jun;152(2):661–73.

Spielman SJ, Weaver S, Shank SD, Magalis BR, Li M, Kosakovsky Pond SL. Evolution of Viral Genomes: Interplay Between Selection, Recombination, and Other Forces. Methods Mol Biol. 2019;1910:427–68.

Stancanelli C, Fabrizi GM, Ferrarini M, Cavallaro T, Taioli F, Di Leo R, et al. Charcot-Marie-Tooth 2F: phenotypic presentation of the Arg136Leu HSP27 mutation in a multigenerational family. Neurol Sci. 2015 Jun;36(6):1003–6.

Su WQ, Yu XJ, Zhou CM. SARS-CoV-2 ORF3a Induces Incomplete Autophagy via the Unfolded Protein Response. Viruses. 2021 Dec 9;13(12):2467.

Sun X, Fontaine JM, Hoppe AD, Carra S, DeGuzman C, Martin JL, et al. Abnormal interaction of motor neuropathy-associated mutant HspB8 (Hsp22) forms with the RNA helicase Ddx20 (gemin3). Cell Stress Chaperones. 2010 Sep;15(5):567–82.

Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993 May;10(3):512–26.

Tang B sha, Zhao G hua, Luo W, Xia K, Cai F, Pan Q, et al. Small heat-shock protein 22 mutated in autosomal dominant Charcot-Marie-Tooth disease type 2L. Hum Genet. 2005 Feb;116(3):222–4.

Tang B, Liu X, Zhao G, Luo W, Xia K, Pan Q, et al. Mutation analysis of the small heat shock protein 27 gene in chinese patients with Charcot-Marie-Tooth disease. Arch Neurol. 2005 Aug;62(8):1201–7.

Tavare S. Some probabilistic and statistical problems in the analysis of DNA sequences. Some mathematical questions in biology / DNA sequence analysis edited by

Robert M Miura [Internet]. 1986 [cited 2022 May 23]; Available from: https://scholar.google.com/scholar_lookup?title=Some+probabilistic+and+statistical+problems+in+the+analysis+of+DNA+sequences&author=Tavare%2C+S.&publication_year=1986

Tegally H, Moir M, Everatt J, Giovanetti M, Scheepers C, Wilkinson E, et al. Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. Nat Med. 2022 Jun 27;1–6.

Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature. 2021 Apr;592(7854):438–43.

Tsai S. J. Is autism caused by early hyperactivity of brain-derived neurotrophic factor? Med Hypotheses [Internet]. 2005;65. Available from: https://doi.org/10.1016/j.mehy.2005.01.034

Turakhia Y, Thornlow B, Hinrichs A, McBroome J, Ayala N, Ye C, et al. Pandemic-Scale Phylogenomics Reveals The SARS-CoV-2 Recombination Landscape. Nature. 2022 Aug 11;1–2.

Tzou PL, Tao K, Pond SLK, Shafer RW. Coronavirus Resistance Database (CoV-RDB): SARS-CoV-2 susceptibility to monoclonal antibodies, convalescent plasma, and plasma from vaccinated persons. PLOS ONE. 2022 Mar 9;17(3):e0261045.

Veitia RA. A generalized model of gene dosage and dominant negative effects in macromolecular complexes. FASEB J. 2010 Apr;24(4):994–1002.

Veitia RA, Birchler JA. Dominance and gene dosage balance in health and disease: why levels matter! J Pathol. 2010 Jan;220(2):174–85.

Vendredy L, Adriaenssens E, Timmerman V. Small heat shock proteins in neurodegenerative diseases. Cell Stress Chaperones. 2020 Jul;25(4):679–99.

Venkat A, Hahn MW, Thornton JW. Multinucleotide mutations cause false inferences of lineage-specific positive selection. Nature Ecology and Evolution. 2018 Aug;2(8):1280–8.

Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. Nature. 2022 Mar;603(7902):679–86.

Vicart P, Caron A, Guicheney P, Li Z, Prévost MC, Faure A, et al. A missense mutation in the alphaB-crystallin chaperone gene causes a desmin-related myopathy. Nat Genet. 1998 Sep;20(1):92–5.

Wang K, Gawinowicz MA, Spector A. The effect of stress on the pattern of phosphorylation of alphaA and alphaB crystallin in the rat lens. Exp Eye Res. 2000 Oct;71(4):385–93.

Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. Mol Biol Evol. 2018 Mar;35(3):773–7.

Webster JM, Darling AL, Uversky VN, Blair LJ. Small Heat Shock Proteins, Big Impact on Protein Aggregation in Neurodegenerative Disease. Front Pharmacol. 2019;10:1047.

Weeks SD, Muranova LK, Heirbaut M, Beelen S, Strelkov SV, Gusev NB. Characterization of human small heat shock protein HSPB1 α-crystallin domain localized mutants associated with hereditary motor neuron diseases. Sci Rep. 2018 Jan 12;8(1):688.

Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. RELAX: detecting relaxed selection in a phylogenetic framework. Mol Biol Evol [Internet]. 2015;32. Available from: https://doi.org/10.1093/molbev/msu400

Whelan S, Goldman N. Estimating the frequency of events that cause multiple-nucleotide changes. Genetics. 2004 Aug;167(4):2027–43.

Wilkie AO. The molecular basis of genetic dominance. J Med Genet. 1994 Feb;31(2):89–98.

Wisotsky SR, Kosakovsky Pond SL, Shank SD, Muse SV. Synonymous site-to-site substitution rate variation dramatically inflates false positive rates of selection analyses: ignore at your own peril. Mol Biol Evol [Internet]. 2020;37. Available from: https://doi.org/10.1093/molbev/msaa037

Wolfe KH, Sharp PM. Journal of Molecular Evolution Mammalian Gene Evolution: Nucleotide Sequence Divergence Between Mouse and Rat. Vol. 37, J Mol Evol. 1993 p. 441–56.

Yanai I, Lercher M. Night science. Genome Biology. 2019 Aug 28;20(1):179.

Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution. 1998;15(5):568–73.

Ylikallio E, Johari M, Konovalova S, Moilanen JS, Kiuru-Enari S, Auranen M, et al. Targeted next-generation sequencing reveals further genetic heterogeneity in axonal Charcot-Marie-Tooth neuropathy and a mutation in HSPB1. Eur J Hum Genet. 2014 Apr;22(4):522–7.

Yokoyama S, Tada T, Zhang H, Britt L. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proceedings of the National Academy of Sciences. 2008 Sep;105(36):13480–5.

Yoshii A, Constantine-Paton M. Postsynaptic BDNF-TrkB signaling in synapse maturation, plasticity, and disease. Dev Neurobiol. 2010;70.

Young BE, Wei WE, Fong SW, Mak TM, Anderson DE, Chan YH, et al. Association of SARS-CoV-2 clades with clinical, inflammatory and virologic outcomes: An observational study. EBioMedicine. 2021 Apr;66:103319.

Zaheri M, Dib L, Salamin N. A generalized mechanistic codon model. Molecular Biology and Evolution. 2014;

Zakharyan R, Boyajyan A, Arakelyan A, Gevorgyan A, Mrazek F, Petrek M. Functional variants of the genes involved in neurodevelopment and susceptibility to schizophrenia in an Armenian population. Hum Immunol [Internet]. 2011;72. Available from: https://doi.org/10.1016/j.humimm.2011.05.018

Zhai J, Lin H, Julien JP, Schlaepfer WW. Disruption of neurofilament network with aggregation of light neurofilament protein: a common pathway leading to motor neuron degeneration due to Charcot-Marie-Tooth disease-linked mutations in NFL and HSPB1. Hum Mol Genet. 2007 Dec 15;16(24):3103–16.

Zhang L, Benedek D, Fullerton C, Forsten R, Naifeh J, Li X. PTSD risk is associated with BDNF Val66Met and BDNF overexpression. Mol Psychiatry [Internet]. 2014;19. Available from: https://doi.org/10.1038/mp.2012.180

Zhang R xu, Yang X, Zi X hong, Li X bo, Xia K, Liu T, et al. [Cellular expression of (R127W)HSPB1 and its co-localization with neurofilament light chain]. Zhonghua Yi Xue Yi Chuan Xue Za Zhi. 2011 Oct;28(5):496–500.

Zhao LP, Roychoudhury P, Gilbert P, Schiffer J, Lybrand TP, Payne TH, et al. Mutations in viral nucleocapsid protein and endoRNase are discovered to associate with COVID19 hospitalization risk. Sci Rep. 2022 Jan 24;12(1):1206.

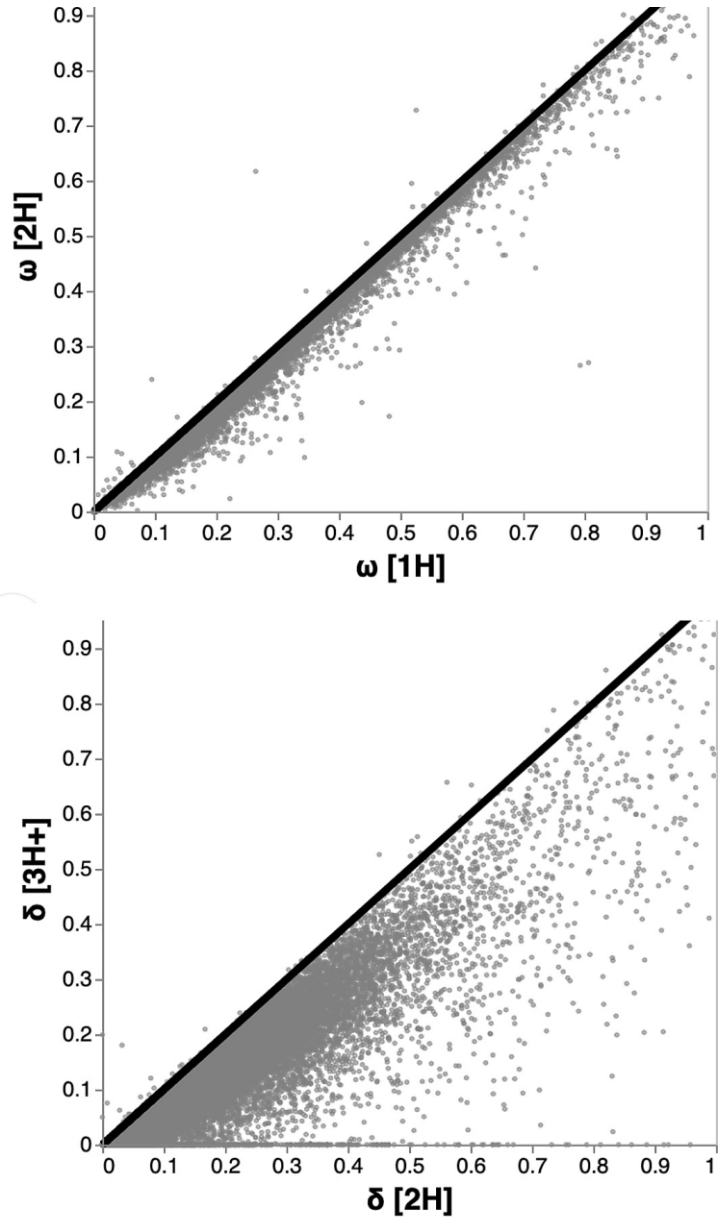
Zoller S, Schneider A. A new semiempirical codon substitution model based on principal component analysis of mammalian sequences. Mol Biol Evol. 2012 Jan;29(1):271–7.

# APPENDICIES

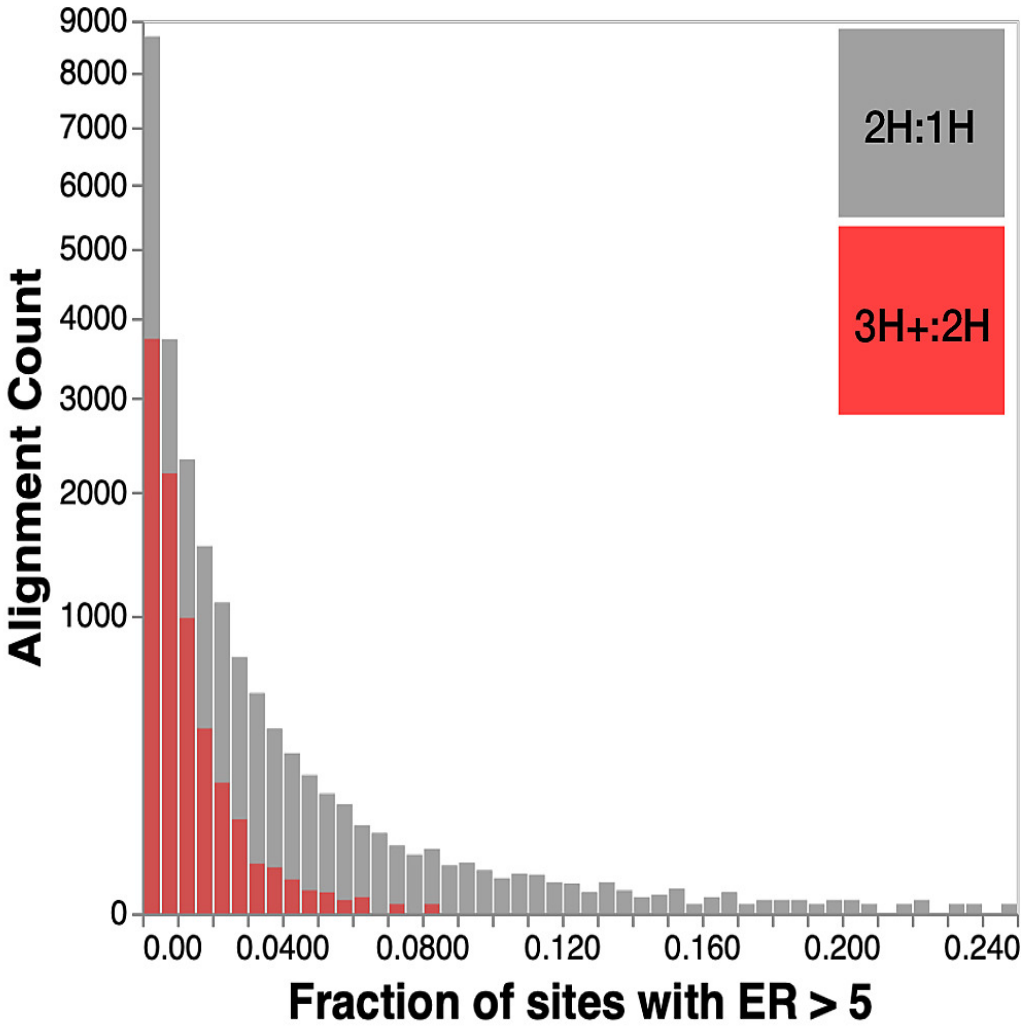
## APPENDIX A: Supplemental Material for Chapter 2

A.S1 Table. Estimated ω rate distributions for benchmark datasets for different models on the benchmark datasets. E[ω] : the mean ω value for the 1H model. $\frac{\mathrm{E}[\omega]:2\mathrm{H}}{\mathrm{E}[\omega]:1\mathrm{H}}$ the ratio of mean ω estimates from 2H and 1H models. $\frac{\delta:3\mathrm{H}+}{\delta:2\mathrm{H}}$ the ratio of estimates from 3H+ and 2H models. The datasets are sorted by increasing values of the $\frac{\mathrm{E}[\omega]:2\mathrm{H}}{\mathrm{E}[\omega]:1\mathrm{H}}$ column. Genes where there was significant evidence (LRT p < 0:05) of non-zero 2H rates are bolded, and those where there is evidence of non-zero 3H rates are underlined.

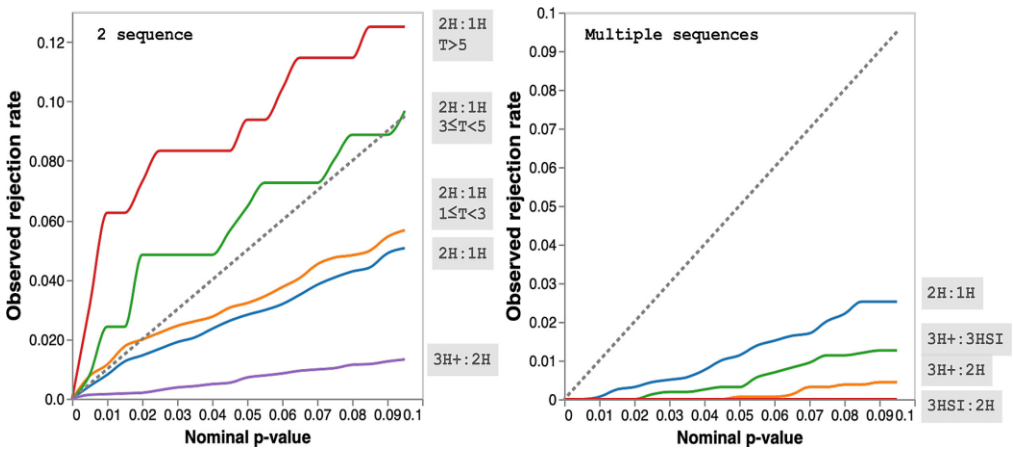
| Gene | 1H | | | | 2H | | | | 3H+ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $E[\omega]$ | $\omega_1(p_1)$ | $\omega_2(p_2)$ | $\omega_3(p_3)$ | $\frac{E[\omega]:2H}{E[\omega]:1H}$ | $\omega_1(p_1)$ | $\omega_2(p_2)$ | $\omega_3(p_3)$ | $\frac{\delta:3H+}{\delta:2H}$ | $\omega_1(p_1)$ | $\omega_2(p_2)$ | $\omega_3(p_3)$ |
| **$\beta$-globin** | 0.28 | 0.0073 (32.4%) | 0.28 (58.9%) | 1.3 (8.66%) | 0.6 | 0.00071 (28.4%) | 0.17 (61.5%) | 0.64 (10.1%) | 0.86 | 0.0035 (30.2%) | 0.19 (61.1%) | 0.7 (8.68%) |
| **Vertebrate Rhodopsin** | 0.12 | 0.0097 (58.8%) | 0.17 (30.3%) | 0.54 (10.9%) | 0.7 | 0.0085 (60.6%) | 0.12 (29.6%) | 0.4 (9.79%) | 0.75 | 0.0088 (60.6%) | 0.13 (29.8%) | 0.41 (9.63%) |
| Flavivirus NS5 | 0.047 | 0.0026 (63.9%) | 0.07 (26.5%) | 0.28 (9.57%) | 0.73 | 0.0031 (67.1%) | 0.062 (25.5%) | 0.22 (7.4%) | 0.68 | 0.0031 (66.8%) | 0.063 (25.7%) | 0.22 (7.55%) |
| **Drosophila *adh*** | 0.1 | 0 (50.1%) | 0.1 (33.8%) | 0.43 (16%) | 0.73 | 0 (38.1%) | 0.047 (42.5%) | 0.29 (19.4%) | 0.76 | 0 (37.2%) | 0.046 (43.2%) | 0.29 (19.6%) |
| **COXI** | 0.07 | 0.0027 (75.4%) | 0.033 (17.3%) | 0.14 (7.33%) | 0.79 | 0.0031 (79%) | 0.035 (14%) | 0.12 (6.99%) | 1 | 0.0031 (79.3%) | 0.036 (13.9%) | 0.12 (6.87%) |
| **Sperm lysin** | 1.1 | 0.11 (36.9%) | 1.1 (41.7%) | 2.9 (21.4%) | 0.83 | 0.096 (37.2%) | 1 (43.8%) | 2.5 (19%) | 0.87 | 0.1 (37.6%) | 1 (43.7%) | 2.4 (18.7%) |
| **Hepatitis D virus antigen** | 0.48 | 0.033 (46.2%) | 0.4 (32.4%) | 1.6 (21.4%) | 0.84 | 0.038 (48.7%) | 0.37 (30.2%) | 1.3 (21.2%) | 0.9 | 0.037 (48.3%) | 0.36 (29.9%) | 1.3 (21.7%) |
| **Camelid VHH** | 0.95 | 0.12 (34%) | 0.73 (40.8%) | 2.5 (25.1%) | 0.89 | 0.1 (34.7%) | 0.66 (40.4%) | 2.2 (24.9%) | 0.92 | 0.1 (34.4%) | 0.66 (40.2%) | 2.2 (25.5%) |
| **HIV-1 *RT*** | 0.19 | 0.016 (71.4%) | 0.35 (22.2%) | 1.6 (6.38%) | 0.93 | 0.015 (70.9%) | 0.31 (22.3%) | 1.5 (6.82%) | 0.93 | 0.015 (71.1%) | 0.31 (22.2%) | 1.5 (6.74%) |
| Encephalitis *env* | 0.054 | 0.024 (77.9%) | 0.028 (15.5%) | 0.48 (6.62%) | 0.96 | 0.023 (71.8%) | 0.026 (21.6%) | 0.46 (6.64%) | 1 | 0.023 (75.8%) | 0.026 (17.6%) | 0.46 (6.64%) |
| **Influenza A virus HA** | 0.47 | 0.095 (72%) | 0.93 (23%) | 3.7 (5.06%) | 0.98 | 0.097 (72.8%) | 0.95 (22.6%) | 3.7 (4.69%) | 0.99 | 0.098 (73.2%) | 0.98 (22.4%) | 3.8 (4.41%) |
| HIV-1 *vif* | 0.84 | 0.14 (61.5%) | 0.83 (20.6%) | 3.2 (17.9%) | 0.99 | 0.16 (66.2%) | 0.96 (16.4%) | 3.2 (17.4%) | 0.16 | 0.16 (65.6%) | 0.9 (16.2%) | 3.2 (18.2%) |
| Primate Lysozyme | 0.61 | 0.14 (0%) | 0.22 (82%) | 2.4 (18%) | 1 | 0.14 (0%) | 0.22 (82%) | 2.4 (18%) | 1 | 0.21 (16%) | 0.22 (66%) | 2.4 (18%) |



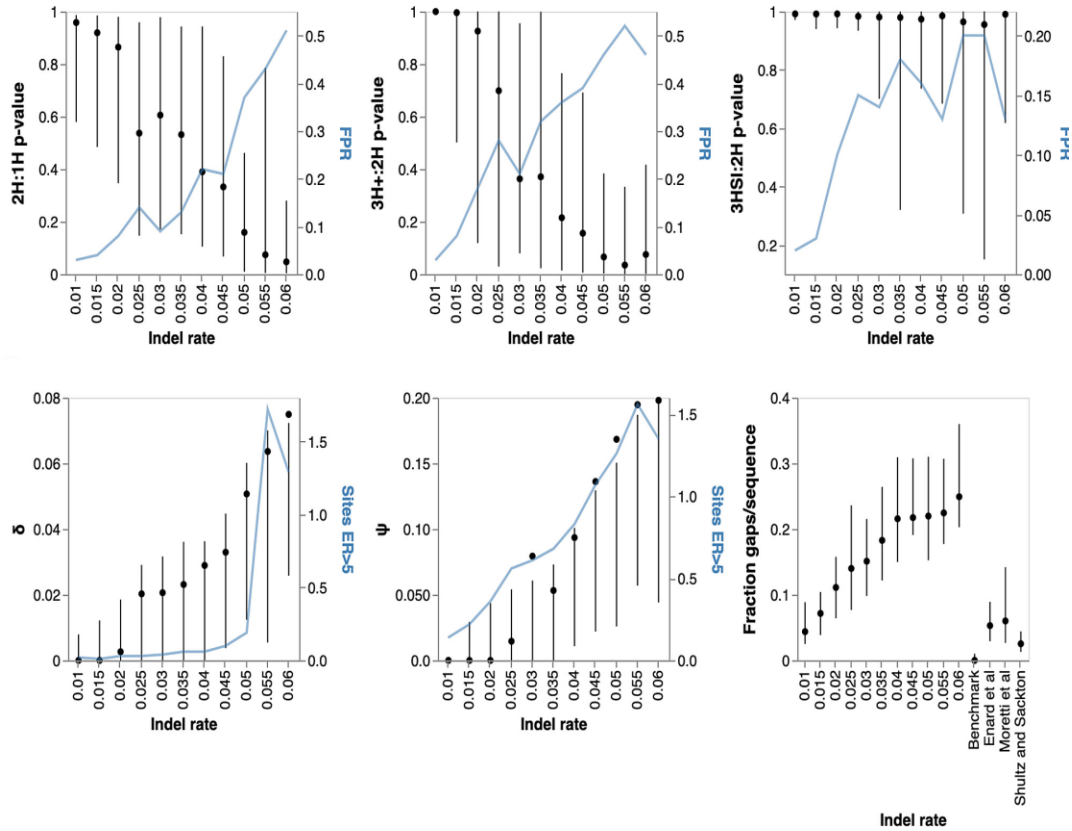

A.S1 Figure. The effect of model choice on rate estimates. Point estimates of global rate parameters under different models for each of the empirical datasets.



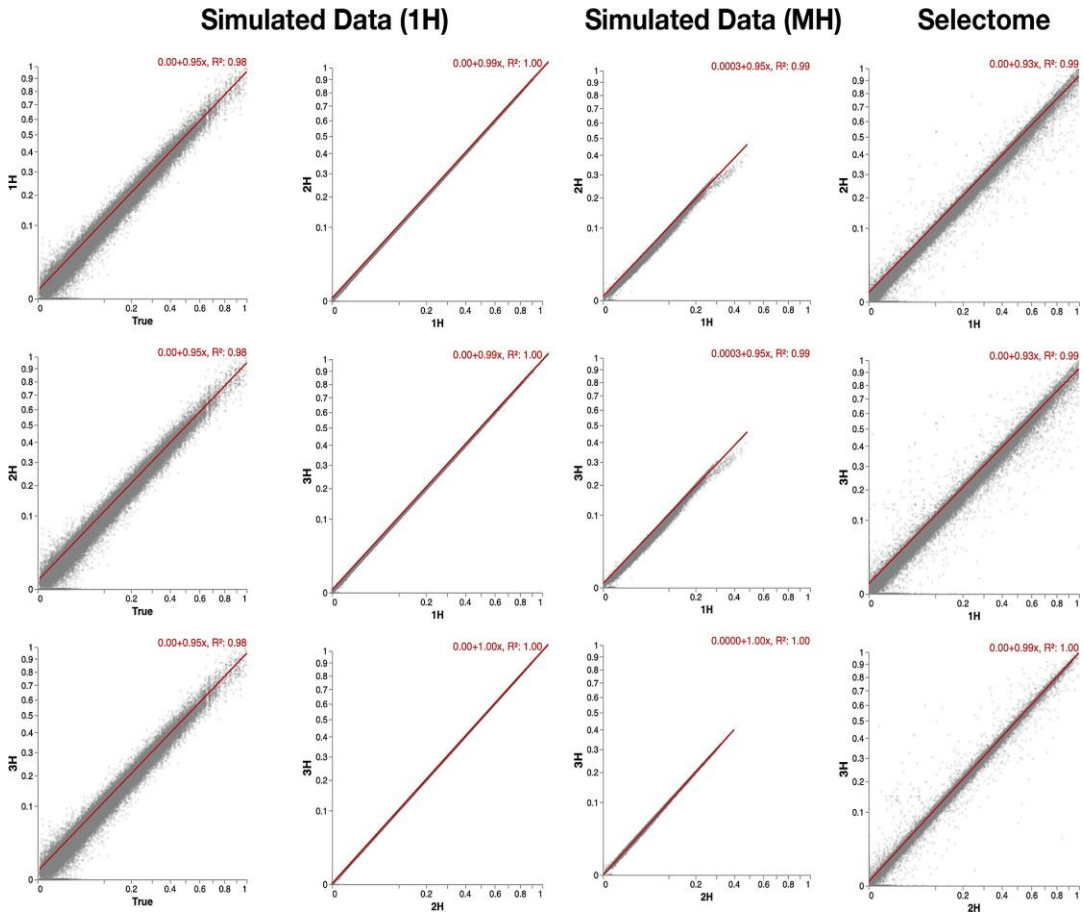

A.S2 Figure. The fraction of sites with strong MH model preference. Histograms are over alignments where there was significant ($p < 0.01$) support for the corresponding model: 20, 338 for 2H:1H (gray) and 7664 for 3H:2H (red).




A.S3 Figure. <u>False positive rates for LRTs on simulated data.</u> Results are shown for two sequence analyses (left) and multiple sequence analyses (right). For the two sequence simulations, we stratified the simulations by the length of the branch, *T*, (the range is labeled in the figure) measured in expected substitutions per site. The dotted line shows the nominal expectation (rejection rate = nominal p-value).

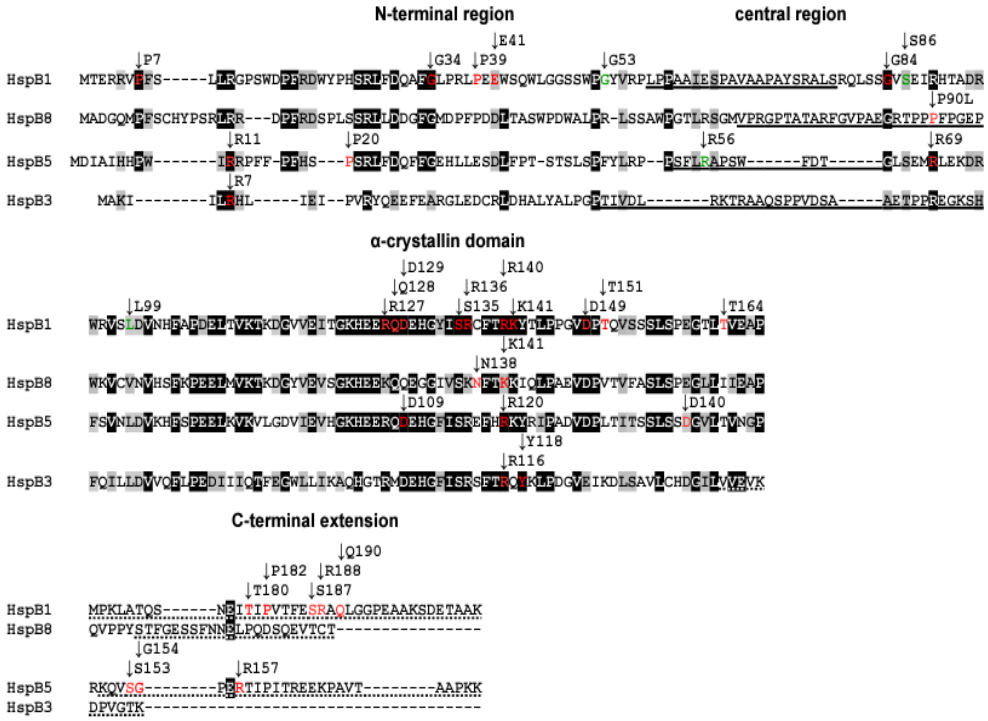



A.S4 Figure. <u>Indel rate verse TH rate.</u> Alignments with indel were simulated using INDELible across using the Drosophila *adh* tree and alignment length using GY94 M3 model with site-to-site $\omega$ variation. LRT p-values and rejection rates (FPR, at $p \leq 0.05$) are shown for different tests in the top row. The bottom row shows estimated $\delta$ and $\psi$ rates as a function of simulated indel rates, as well as the number of sites inferred to have high evidence ratios (ER) for 2H or 3H modes. The plot on the bottom right shows the average fraction of a sequence that in an alignment that is comprised of gaps is shown for simulated data, and empirical collections.




A.S5 Figure. Branch-length estimate behavior under different models. Simulated Data (1H): null simulations (1H model). Simulated Data (MH): power simulations. Selectome: empirical data. Red lines are drawn with least squares linear regression whose estimates slopes and intercepts as well as proportions of variance ($R^2$) explained are added to each plot.

## APPENDIX B: Supplemental Material for Chapter 3




B.S1 Figure. Domains and regions in the aligned human sequences of HspB1, HspB3, HspB5 and HspB8, and positions of the disease-associated missense mutations. The alignment was according to Fontaine et al. (2003) with minor modifications. Identical amino acid residues are highlighted in black if they occur in at least three sequences. Similar amino acid residues (E/D; A/G; H/F/W/Y; S/T; I/L/V; H/R/K) are highlighted in gray if they occur in at least three sequences, or in addition to two or three identical amino acid residues. Amino acid residues affected by missense mutations associated with various forms of neuropathy, myopathy, or with cataracts in the eye lens, are marked with color: red, mutation sites associated with a dominant disease phenotype (or this can be assumed); green, mutation sites associated with a recessive disease phenotype. Sequence sections

with relatively high average dN/dS point estimates (ω-values; cf., Fig. 3.1), compared to the α-crystallin domain and the N-terminal region, are underlined (solid and dotted underline) and correspond approximately to the central regions and the C-terminal extensions, respectively. This demarcation served to define the sequence partitions as used in this study.

B.S1 Table. Reported disease-associated missense mutations in human HspB1, HspB3, HspB5 and HspB8a.

[a] Abbreviations and symbols. Inheritance of phenotype: D, dominant; R, recessive; SD, semi-dominant; I, isolated (positive family history referred, parents not examined); S, sporadic (parents not affected); PM, parental mosaicism; Associated disease phenotypes: ALS, amyotrophic lateral sclerosis; dHMN, distal hereditary motor neuropathy; CMT, Charcot-Marie-Tooth disease; C, cataract; CC, congenital cataract; JC, juvenile cataract; MFM/AxM/DM, myofibrillar/axial/distal myopathy; DCM/RCM/HCM, dilated/restrictive/hypertrophic cardiomyopathy; FP, feet paresthesia; LLS, lower limb spasticity; N/D, no data available. Assumed, though not observed, inheritance is given in parenthesis.

[b] Individuals or families with observed or assumed dominant or semi-dominant disease phenotypes exhibited heterozygosity with both wild-type and mutant alleles being present. The four individuals or families with recessive disease phenotypes were homozygous with regard to the mutant alleles.

[c] Personal communication by Dr. J. Irobi, Antwerp, Belgium.
[d] The mutation p.R140G in HspB1 occurred in both heterozygous and homozygous individuals, suggesting tentatively a semi-dominant inheritance pattern.
[e] The pathogenicity of the HspB3 mutations has been disputed (Adriaenssens et al. 2017; Vendredy et al. 2020).

[f] References:

| sHSP | mutation (protein) | mutation (cDNA) | associated disease phenotype | | references[f] |
|---|---|---|---|---|---|
| | | | symptoms | inheritance[b] | |
| HspB1 | p.P7S | c.19C>T | dHMN | D | Echaniz-Laguna et al. 2017 |
| | p.P7R | c.20C>G | CMT | D | Fortunato et al. 2017 |
| | p.G34R | c.100G>A | dHMN | I (D) | Capponi et al. 2011 |
| | p.P39L | c.116C>T | dHMN, CMT | D, S | Houlden et al. 2008; Capponi et al. 2011; EchanizLaguna et al. 2017; Rossor et al. 2017 |
| | p.E41K | c.121G>A | dHMN | D | Capponi et al. 2011 |
| | p.G53D | c.158G>A | dHMN + cerebrellar ataxia | R | Echaniz-Laguna et al. 2017 |
| | p.G84R | c.250G>C c.250G>A | dHMN, CMT | D, S | Houlden et al. 2008; James et al. 2008, Rossor et al. 2017; Ho et al. 2017 |
| | p.S86L | c.257C>T | dHMN + ALS-like | R | Scarlato et al. 2015 |
| | p.L99M | c.295C>A | dHMN | R | Houlden et al. 2008; Rossor et al. 2017 |
| | p.R127W | c.379C>T | dHMN, CMT, FP, LLS, ALS | D | Evgrafov et al. 2004; Tang et al. 2005a; Dierick et al. 2008; Echaniz-Laguna et al. 2017; Benedetti et al. 2010; Chen et al. 2021 |

| sHSP | mutation (protein) | mutation (cDNA) | associated disease phenotype | | references[f] |
|---|---|---|---|---|---|
| | | | symptoms | inheritance[b] | |
| | p.R127L | c.380G>T | CMT | D | Ylikallio et al. 2014 |
| | p.Q128R | c.383A>G | dHMN, LLS | D | Echaniz-Laguna et al. 2017 |
| | p.D129E | c.387C>G | dHMN + myopathy | D | Lewis-Smith et al. 2016 |
| | p.S135F | c.404C>T | dHMN, CMT+FP | D, S | Evgrafov et al. 2004; Houlden et al. 2008; Chung et al. 2008; Echaniz-Laguna et al. 2017; Rossor et al. 2017 |
| | p.S135C | c.404C>G | dHMN, CMT | D | Benedetti et al. 2010; Oberstadt et al. 2016 |
| | p.S135Y | c.404C>A | CMT | D, S | Ylikallio et al. 2014; Rossor et al. 2017 |
| | p.R136W | c.406C>T | CMT | Dc | Evgrafov et al. 2004 |
| | p.R136L | c.407G>T | dHMN, CMT + deafness + pyramidal signs | D, I | Capponi et al. 2011; Gaeta et al. 2012; Stancanelli et al. 2015 |
| | p.R136H | c.407G>C | dHMN | (D) | Frasquet et al. 2021 |
| | p.R140G | c.418C>G | dHMN, DM | S, SD[d] | Houlden et al. 2008; Rossor et al. 2017; Bugiardini et al. 2017 |
| | p.K141Q | c.421A>C | dHMN | D | Ikeda et al. 2009; Maeda et al. 2014 |

| sHSP | mutation (protein) | mutation (cDNA) | associated disease phenotype | | references[f] |
|---|---|---|---|---|---|
| | | | symptoms | inheritance[b] | |
| | p.D149A | c.446A>C | ALS | (D) | Chen et al. 2021 |
| | p.T151I | c.452C>T | dHMN | D | Evgrafov et al. 2004; Dierick et al. 2008; EchanizLaguna et al. 2017 |
| | p.T151P | c.451A>C | ALS | (D) | Chen et al. 2021 |
| | p.T164A | c.490A>G | CMT | D | Lin et al. 2011 |
| | p.T180I | c.539C>T | dHMN, CMT | D, S | Luigetti et al. 2010; Capponi et al. 2011; EchanizLaguna et al. 2017 |
| | p.P182L | c.545C>T | dHMN | PM | Evgrafov et al. 2004; Dierick et al. 2008 |
| | p.P182S | c.544C>T | dHMN | S (D) | Kijima et al. 2005 |
| | p.P182A | N/D | dHMN | D | Rossor et al. 2017 |
| | p.S187L | c.560C>T | dHMN | S | Echaniz-Laguna et al. 2017 |
| | p.R188W | c.562C>T | CMT | I (D) | Capponi et al. 2011 |
| | p.Q190H | c.570C>G | ALS + cognitive impairment | S | Capponi et al. 2016 |
| HspB3[e] | p.R7S | c.21G>T | dHMN | D | Kolb et al. 2010; Laššuthová et al. 2016 |
| | p.R116P | c.347G>C | myopathy + axonal neuropathy | D | Morelli et al. 2017 |
| | p.Y118H | c.352C>T | CMT | D | Nam et al. 2018 |
| HspB5 | p.P20S | c.58C>T | CC | D | Liu et al. 2006a |

| sHSP | mutation (protein) | mutation (cDNA) | associated disease phenotype | | references[f] |
|---|---|---|---|---|---|
| | | | symptoms | inheritance[b] | |
| | p.R56W | c.166C>T | JC | R | Safieh et al. 2009 |
| | p.D109H | c.325G>C | MFM + DCM + C | D | Sacconi et al. 2012 |
| | p.D109G | c.326A>G | AxM + RCM | D | Brodehl et al. 2017 |
| | p.R120G | c.358A>G | MFM + HCM+C | D | Vicart et al. 1998 |
| | p.D140N | c.418G>A | CC | D | Liu et al. 2006b |
| | p.S153F | c.458>T | myopathy + HCM + C + cerebrellar ataxia + optic atrophy + cognitive impairment | (D) | Sadeh et al. 2021 |
| | p.G154S | c.460G>A | DM + DCM | (D) | Pilotto et al. 2006; Reilich et al. 2010 |
| | p.R157H | c.470G>A | DCM | (D) | Inagaki et al. 2006 |
| HspB8 | p.P90L | c.413A>C | dHMN | S | Echaniz-Laguna et al. 2017 |
| | p.N138T | c.413A>C | dHMN | D | Echaniz-Laguna et al. 2017 |
| | p.K141E | c.421A>G | dHMN, distal myopathy | D | Irobi et al. 2004; Dierick et al. 2008; Ghaoui et al. 2016 |
| | p.K141N | c.423G>C | dHMN, CMT | D | Irobi et al. 2004; ck et al. 2008; z-Laguna et al. 2017 |
| | | c.423G>T | CMT | D | Tang et al. 2005b |
| | p.K141T | c.422A>C | CMT | S (D) | Nakhro et al. 2013 |
| | p.K141M | c.422A>T | dHMN + FP | D | Echaniz-Laguna t al. 2017 |

B.S2 Table. List of sHSP mRNAs (GeneBank and ENSMBLE identifiers) as used for the FEL analysis, the species of origin and their taxonomic ranking. If the coding part of the sHSP mRNAs were embedded in larger sequence blocks, Expasy translation tool (https://web.expasy.org/translate/) was used to identify the coding parts of the sequences.

| **HspB1** |
|---|
| ***Gnathostomata: Teleostomi: Euteleostomi: Sarcopterygii: Dipnotetrapodomorpha: Tetrapoda:*** |
| ***Amniota: Mammalia: Theria: Eutheria: Boreoeutheria: Euarchontoglires:*** <br> ***Primates*** <br> 001. CR536489.1 (Homo sapiens) |
| 002. XM_002743949.4 (Callithrix jacchus) |
| 003. ENSCANT00000043608.1 (Colobus angolensis) |
| 004. XM_017533485.1 (Cebus capucinus) |
| 005. ENSCATT00000043996.1 (Cerocebus atys) |
| 006. XM_004045617.2 (Gorilla gorilla) |
| 007. NM_001283885.1 (Macaca fascicularis) |
| 008. NM_001260949.2 (Macaca mulatta) |
| 009. XM_012759201.2 (Microcebus murinus) |

| |
|---|
| 010. XM_012498702.1 (Nomascus leucogenys) |
| 011. XM_003794184.3 (Otolemur garnettii) |
| 012. ENSPPAT00000052328.1 (Pan paniscus) |
| 013. GABE01006326.1 (Pan troglodytes) |
| 014. XM_003895845.2 (Papio anubis) |
| 015. XM_024249722.1 (Pongo abelii) |
| 016. XM_012640279.1 (Propithecus coquereli) |
| 017. XM_010387433.1 (Rhinopithecus roxellana) |
| 018. AY518309.1 (Saguinus oedipus) |
| ***Euarchontoglires: Glires*** |
| 019. XM_020156395.1 (Castor canadensis) |
| 020. XM_003470110.4 (Cavia porcellus) |
| 021. XM_005415130.2 (Chinchilla lanigera) |
| 022. XM_010619058.2 (Fukomys damarensis) |

| |
|---|
| 023. XM_004840094.3 (Heterocephalus glaber) |
| 024. XM_004666274.1 (Jaculus jaculus) |
| 025. XM_015490689.1 (Marmota marmota) |
| 026. XM_021649341.1 (Meriones unguiculatus) |
| 027. XM_005080371.3 (Mesocricetus auratus) |
| 028. XM_005344605.3 (Microtus ochrogaster) |
| 029. NM_013560.2 (Mus musculus) |
| 030. XM_021186957.1 (Mus pahari) |
| 031. MGP_SPRETEiJ_T0074122.1 (Mus spretus) |
| 032. XM_008832266.2 (Nannospalax galili) |
| 033. XM_004587193.1 (Ochotona princeps) |
| 034. XM_004630450.2 (Octodon degus) |
| 035. XM_002711905.3 (Oryctolagus cuniculus) |
| 036. XM_006971277.2 (Peromyscus maniculatus) |

| |
|---|
| 037. NM_031970.4 (Rattus norvegicus) |

| |
|---|
| |
| ***Euarchontoglires: Scandentia*** |
| 038. XM_014590610.1 (Tupaia chinensis) |
| ***Euarchontoglires: Dermoptera: Cynocephalidae*** |
| 039. XM_008594581.1 (Galeopterus variegatus) |
| ***Boreoeutheria: Laurasiatheria*** |
| 040. NM_001304892.1 (Ailuropoda melanoleuca) |
| 041. XM_007186570.1 (Balaenoptera acutorostrata) |
| 042. KX784505.1 (Bos indicus) |
| 043. NM_001025569.1 (Bos taurus) |

| |
|---|
| 044. XM_010980735.1 (Camelus dromedarius) |

| |
|---|
| 045. NM_001003295.2 (Canis lupus) |

| |
|---|
| 046. XM_004442133.2 (Ceratotherium simum) |
| 047. XM_004688019.2 (Condylura cristata) |
| 048. XM_022588028.1 (Delphinapterus leucas) |

| |
|---|
| 049. XM_022500951.1 (Enhydra lutris) |
| 050. XM_008141457.1 (Eptesicus fuscus) |
| 051. XM_001504478.4 (Equus caballus) |
| 052. XM_007517945.2 (Erinaceus europaeus) |
| 053. XM_003998540.5 (Felis catus) |
| 054. XM_007468802.1 (Lipotes vexillifer) |
| 055. XM_017654032.1 (Manis javanica) |

| |
|---|
| 056. XM_004761627.2 (Mustela putorius) |
| 057. XM_005879150.2 (Myotis brandtii) |

| |
|---|
| 058. XM_014455390.2 (Myotis lucifugus) |
| 059. XM_021700598.1 (Neomonachus schauinslandi) |
| 060. XM_004399079.2 (Odobenus rosmarus) |
| 061. XM_004268817.2 (Orcinus orca) |
| 062. XM_012123070.2 (Ovis aries) |

| |
|---|
| 063. XM_019421634.1 (Panthera pardus) |
| 064. XM_006918567.3 (Pteropus alecto) |
| 065. XM_007103974.2 (Physeter catodon) |
| 066. XM_011371550.2 (Pteropus vampyrus) |
| 067. XM_019752763.1 (Rhinolophus sinicus) |
| 068. XM_016146541.1 (Rousettus aegyptiacus) |
| 069. XM_004620853.1 (Sorex araneus) |
| 070. NM_001007518.1 (Sus scrofa) |
| 071. XM_026516042.1 (Ursus arctos) |
| ***Eutheria: Afrotheria*** |
| 072. XM_004705622.1 (Echinops telfairi) |
| 073. XM_007940206.1 (Orycteropus afer) |
| 074. ENSPCAT00000013780.1 (Procavia capensis)<br><br>075. XM_004391282.2 (Trichechus manatus) |
| ***Theria: Metatheria*** |

| |
|---|
| 076. ENSMEUT00000014075.1 (Macropus eugenii) |
| 077. XM_001378935.4 (Monodelphis domestica) |
| 078. ENSSHAT00000014800.1 (Sarcophilus harrisii) |
| ***Eutheria: Xenarthra*** |
| 079. XM_004448274.3 (Dasypus novemcinctus) |
| ***Amniota: Sauropsida: Sauria*** |
| 080. XM_006260278.3 (Alligator mississippiensis) |
| 081. NM_001287309.1 (Alligator mississippiensis) |
| 082. XM_006033234.3 (Alligator sinensis) |
| 083. XM_027471928.1 (Anas platyrhynchos) |
| 084. XM_011581588.1 (Aquila chrysaetos) |
| 085. XM_005284435.3 (Chrysemys picta) |
| 086. XM_005501958.3 (Columba livia) |
| 087. XM_015881044.1 (Coturnix japonica) |
| 088. XM_019545392.1 (Crocodylus porosus) |

| |
|---|
| 089. XM_009560237.1 (Cuculus canorus) |
| 090. XM_026107252.1 (Dromaius novaehollandiae) |
| 091. XM_005439800.1 (Falco cherrug) |
| 092. XM_005240472.1 (Falco peregrinus) |
| 093. XM_005056512.2 (Ficedula albicollis) |
| 094. NM_205290.1 (Gallus gallus) |
| 095. XM_015408020.1 (Gekko japonicus) |
| 096. XM_010573540.1 (Haliaeetus leucocephalus) |

| |
|---|
| 097. XM_017828875.1 (Lepidothrix coronata) |

| |
|---|
| 098. XM_026050202.1 (Nothoprocta perdicaria) |
| 099. XM_015646306.1 (Parus major) |

| |
|---|
| 100. XM_025181738.1 (Pelodiscus sinensis) |
| 101. XM_015823845.1 (Protobothrops mucrosquamatus) |
| 102. XM_005525843.2 (Pseudopodoces humilis) |

| |
|---|
| 103. XM_009094688.2 (Serinus canaria) |
| 104. XM_002194667.3 (Taeniopygia guttata) |
| 105. XM_024207685.2 (Terrapene mexicana) |

| ***Tetrapoda: Amphibia: Batrachia: Anura*** |
|---|
| 106. XM_018571094.1 (Nanorana parkeri) |
| 107. NM_001093816.1 (Xenopus laevis) |
| 108. NM_001079349.1 (Xenopus tropicalis) |

| ***Sarcopterygii: Coelacanthimorpha: Coelacanthiformes: Coelacanthidae*** |
|---|
| 109. XM_005992232.2 (Latimeria chalumnae) |

| ***Teleostomi: Euteleostomi: Actinopterygii: Actinopteri: Neopterygii*** |
|---|
| 110. XM_022205999.1 (Acanthochromis polyacanthus) XM_023269342.1 (Amphiprion ocellaris) |
| 111. ENSACLT00000008013.1 (Astatotilapia calliptera) |
| 112. XM_007255827.3 (Astyanax mexicanus) |
| 113. XM_014019664.1 (Austrofundulus limnaeus) |
| 114. KT359728.1 (Clarias batrachus) |

| | |
|---|---|
| 115. | XM_026240617.1 (Carassius auratus) |
| 116. | ENSCSET00000017302.1 (Cynoglossus semilaevis) |
| 117. | XM_015383644.1 (Cyprinodon variegatus) |
| 118. | NM_001008615.2 (Danio rerio) |
| 119. | XM_027010259.1 (Electrophorus electricus) |
| 120. | ENSGMOT00000013588.1 (Gadus morhua) |
| 121. | ENSGAFT00000022102.1 (Gambusia affinis) |
| 122. | ENSGACT00000026785.1 (Gasterosteus aculeatus) |
| 123. | ENSHCOT00000014698.1 (Hippocampus comes) |
| 124. | XM_017460709.1 (Ictalurus punctatus) |
| 125. | XM_017412492.2 (Kryptolebias marmoratus) |
| 126. | ENSLBET00000008022.1. Labrus bergylta) |
| 127. | NM_001303311.1 (Larimichthys crocea) |
| 128. | XM_018692336.1 (Lates calcarifer) |

| |
|---|
| 129. XM_006640987.2 (Lepisosteus oculatus) |
| 130. ENSMAMT00000009140.1 (Mastacembelus armatus) |
| 131. XM_004557456.4 (Maylandia zebra) |
| 132. ENSMMOT00000007488.1 (Mola mola) |
| 133. ENSMALT00000001108.1 (Monopterus albus) |
| 134. XM_015961780.1 (Nothobranchius furzeri) |
| 135. NM_001279530.1 (Oreochromis niloticus) |
| 136. XM_026921457.1 (Pangasianodon hypophthalmus) |
| 137. XM_020085355.1 (Paralichthys olivaceus) |
| 138. ENSPKIT00000014737.1 (Paramormyrops kingsleyae) XM_014989045.1 (Poecilia mexicana) |
| 139. ENSPLAT00000000657.1 (Poecilia latipinna) |
| 140. ENSPFOT00000008950.1 (Poecilia latipinna) |
| 141. ENSPRET00000012065.1 (Poecilia reticulata) |

| | |
|---|---|
| 142. | ENSPNYT00000022694.1 (Pundamilia nyererei) |
| 143. | ENSPNAT00000009103.1 (Pygocentrus nattereri) |
| 144. | ENSSMAT00000000761.1 (Scophthalmus maximus) |
| 145. | ENSSLDT00000026576.1 (Seriola lalandi) |
| 146. | XM_027132842.1 (Tachysurus fulvidraco) |
| 147. | ENSTNIT00000011572.1 (Tetraodon. nigroviridis) |
| 148. | ENSXCOT00000007329.1 (Xiphophorus couchianus) |
| 149. | XM_005800786.3 (Xiphophorus maculatus) |

| |
|---|
| ***Gnathostomata: Chondrichthyes*** |
| 150. XM_007896558.1 (Callorhinchus milii) |

| |
|---|
| |
| **HspB3** |
| ***Gnathostomata: Teleostomi: Euteleostomi: Sarcopterygii: Dipnotetrapodomorpha: Tetrapoda:*** |
| ***Amniota: Mammalia: Theria: Eutheria: Boreoeutheria: Euarchontoglires: Primates***<br>001. NM_006308.2 (Homo sapiens) |
| 002. XM_008992067.3 (Callithrix jacchus) |

| |
|---|
| 003. XM_008057475.2 (Carlito syrichta) |
| 004. XM_017533213.1 (Cebus capucinus) |
| 005. XM_012086998.1 (Cercocebus atys) |
| 006. XM_007972294.1 (Chlorocebus sabaeus) |
| 007. XM_011939355.1 (Colobus angolensis) |
| 008. XM_004058861.2 (Gorilla gorilla) |
| 009. XM_005556891.2 (Macaca fascicularis) |
| 010. XM_001096576 (Macaca mulatta) |
| 011. XM_012756743 (Microcebus murinus) |
| 012. XM_003276527 (Nomascus leucogenys) |
| 013. XM_003782777 (Otolemur garnettii) |
| 014. XM_003827356.2 (Pan paniscus) |
| 015. XM_517764.6 (Pan troglodytes) |

| |
|---|
| 016. XM_003899666.3 (Papio anubis) |
| 017. XM_002815553.2 (Pongo abelii) |
| 018. XM_012655284.1 (Propithecus coquereli) |
| 019. XM_010353161.1 (Rhinopithecus roxellana) |
| 020. XM_003925862.1 (Saimiri boliviensis) |
| ***Euarchontoglires: Glires*** |
| 021. XM_020172000.1 (Castor canadensis) |
| 022. XM_003470185.4 (Cavia porcellus) |
| 023. XM_005392716.1 (Chinchilla lanigera) |
| 024. XM_027401441.1 (Cricetulus griseus) |
| 025. XM_013018573.1 (Dipodomys ordii) |
| 026. XM_010626859.2 (Fukomys damarensis) |
| 027. XM_004848525.2 (Heterocephalus glaber) |
| 028. XM_004664944.1 (Jaculus jaculus) |

| |
|---|
| 029. XM_015497380.1 (Marmota marmota marmota) |
| 030. XM_021638917.1 (Meriones unguiculatus) |
| 031. XM_005065524.3 (Mesocricetus auratus) |
| 032. XM_005356816.3 (Microtus ochrogaster) |
| 033. 041 NM_019960.2 (Mus musculus ) |
| 034. XM_021208803.1 (Mus pahari) |

| |
|---|
| 035. XM_008832949.1 (Nannospalax galili) |
| 036. XM_004583794.1 (Ochotona princeps) |

| |
|---|
| 037.XM_004623098.2 (Octodon degus) |
| 038. XM_002714030.3 (Oryctolagus cuniculus) |
| 039. XM_016002221.1 (Peromyscus maniculatus bairdii) |
| 040. NM_031750.1 (Rattus norvegicus) |

| |
|---|
| ***Euarchontoglires: Scandentia*** |
| 041. XM_006149012.1 (Tupaia chinensis) |
| ***Euarchontoglires: Dermoptera: Cynocephalidae*** |
| 042. XM_008586990.1 (Galeopterus variegatus) |
| ***Boreoeutheria: Laurasiatheria*** |
| 043. XM_002927469.3 (Ailuropoda melanoleuca) |
| 044. XM_007195583.1 (Balaenoptera acutorostrata) |
| 045. NM_001046570.2 (Bos taurus) |
| 046. XM_006185923.2 (Camelus ferus) |

| |
|---|
| 047. XM_010977776.1 (Camelus dromedarius) |
| 048. XM_005619327.3 (Canis lupus) |

| |
|---|
| 049. XM_014781626.1 (Ceratotherium simum) |
| 050. XM_012721798.1 (Condylura cristata) |
| 051. XM_022551605.1 (Delphinapterus leucas) |
| 052. XM_022493473.1 (Enhydra lutris) |

| |
|---|
| 053. XM_005604272.3 (Equus caballus) |
| 054. XM_007518898.2 (Erinaceus europaeus) |
| 055. XM_006928019.4 (Felis catus) |
| 056. XM_019641426.1 (Hipposideros armiger) |
| 057. XM_007469286.1 (Lipotes vexillifer) |
| 058. XM_017649540.1 (Manis javanica) |
| 059. XM_016199828.1 (Miniopterus natalensis) |
| 060. XM_004737891.2 (Mustela putorius) |

| |
|---|
| 061. XM_004413285.2 (Odobenus rosmarus) |
| 062. XM_004275133.1 (Orcinus orca) |

| |
|---|
| 063. XM_012135022.2 (Ovis aries musimon) |
| 064. XM_019464339.1 (Panthera pardus) |
| 065. XM_007076644.2 (Panthera tigris) |

| |
|---|
| 066. XM_006912453.1 (Pteropus aletco) |
| 067. XM_028492905.1 (Physeter catodon) |
| 068. XM_028492906.1 (Physeter catodon) |
| 069. XM_011380607.1 (Pteropus vampyrus) |
| 070. XM_016142353.1 (Rousettus aegyptiacus) |
| 071. XM_004608527.1 (Sorex araneus) |
| 072. XM_003133967.4 (Sus scrofa) |
| 073. XM_026510651.1 (Ursus arctos) |
| 074. XM_008692539.1 (Ursus maritimus) |
| 075. XM_015240724.1 (Vicugna pacos) |
| ***Eutheria: Afrotheria*** |
| 076. XM_004703719.1 (Echinops telfairi) |
| 077. XM_006889568.1 (Elephantulus edwardii) |
| 078. XM_003408031.2 (Loxodonta africana) |

| |
|---|
| 079. XM_007951536.1 (Orycteropus afer) |
| 080. ENSPCAT00000013456.1 (Procavia capensis) |
| 081. XM_004380462.1 (Trichechus manatus latirostis) Theria: Metatheria |
| 082. ENSMEUT00000015135.1 (Macropus eugenii) |
| 083. XM_001380851.2 (Monodelphis domestica) |
| 084. XM_003763125.1 (Sarcophilus harrisii) |
| ***Eutheria: Xenarthra*** |
| 085. XM_004483480.3 (Dasypus novemcinctus) |
| ***Amniota: Sauropsida: Sauria*** |
| 086. XM_006017287.2 (Alligator sinensis) |
| 087. XM_003216187.3 (Anolis carolinensis) |
| 088. XM_011596908.1 (Aquila chrysaetos canadensis) |
| 089. XM_014964827.1 (Calidirs pugnax) |
| 090. XM_008500142.1 (Calypte anna) |
| 091. XM_005282165.2 (Chrysemys picta bellii) |

| |
|---|
| 092. XM_021285444.1 (Columba livia) |
| 093. XM_017744316.1 (Corvus brachyrhynchos) |

| |
|---|
| 094. XM_015848720.1 (Coturnix japonica) |
| 095. XM_019532496 (Crocodylus porosus) |

| |
|---|
| 096. XM_009555858.1 (Cuculus canorus) |
| 097. XM_026120936 (Dromaius novaehollandiae) |
| 098. 131. XM_005443917.2 (Falco cherrug) |
| 099. XM_005241994 (Falco peregrinus) |

| |
|---|
| 100. XM_005060621 (Ficedula albicollis) |
| 101. XM_001231557 (Gallus gallus) |
| 102. XM_015419159 (Gekko japonicus) |
| 103. XM_017811136 (Lepidothrix coronata) |

| | |
|---|---|
| 104. | XM_021553896 (Lonchura striata domestica) |
| 105. | XM_009461137 (Nipponia nippon) |
| 106. | XM_015615583 (Parus major) |
| 107. | XM_006139541.3 (Pelodiscus sinensis) |
| 108. | XM_015822883 (Protobothrops mucrosquamatus) |
| 109. | XM_014259875 (Pseudopodoces humilis) XM_018920640 (Serinus canaria) |
| 110. | XM_002195233.1 (Taeniopygia guttata) |
| 111. | XP_024065889.2 (Terrapene carolina triunguis) |
| 112. | XM_026795640 (Zonotrichia albicollis) |
| ***Tetrapoda: Amphibia: Batrachia: Anura*** | |
| 113. | XM_018563319 (Nanorana parkeri) |
| 114. | XM_018226893 (Xenopus laevis) |
| 115. | XM_002941028.4 (Xenopus tropicalis) |

| | |
|---|---|
| ***Sarcopterygii: Coelacanthimorpha: Coelacanthiformes: Coelacanthidae*** | |
| 116. | XM_006006980.2 (Latimeria chalumnae) |

| ***Teleostomi: Euteleostomi: Actinopterygii: Actinopteri: Neopterygii*** |
|---|
| 117. XM_007250482.2 (Astyanax mexicanus) |
| 118. XM_019095081.1 (Cyprinus carpio) |
| 119. NM_001099452.1 (Danio rerio) |
| 120. XM_020052862.1 (Esox lucius) |
| 121. XM_015365087.1 (Lepisosteus oculatus) |
| 122. XM_023820693.1 (Paramormyrops kingsleyae) |
| 123. XM_017705104.1 (Pygocentrus nattereri) |
| 124. XM_016561090.1 (Sinocyclocheilus anshuiensis) |
| 125. XM_016290694.1 (Sinocyclocheilus grahami) |
| 126. XM_016561090.1 (Sinocyclocheilus rhinocerous) |

| ***Actinopteri: Chondrostei: Acipenseriformes: Acipenseroidei: Acipenseridae: Acipenserinae:*** |
|---|
| ***Acipenserini*** |
| 127. MH777919.1 (Acipenser oxyrinchus) |

| ***Gnathostomata: Chondrichthyes*** |
|---|
| 128. XM_007892194.1 (Callorhinchus milii) |

| |
|---|
| 129. XM_020511366.1 (Rhincodon typus) |

| |
|---|
| **HspB5** |
| ***Gnathostomata: Teleostomi: Euteleostomi: Sarcopterygii: Dipnotetrapodomorpha: Tetrapoda:*** |
| ***Amniota: Mammalia: Theria: Eutheria: Boreoeutheria: Euarchontoglires: Primates*** |
| 001. NM_001885.2 (Homo sapiens) |
| 002. XM_011926733.1 (Colobus angolensis) |
| 003. XM_002754383.3 (Callithrix jacchus) |
| 004. XM_008065587.2 (Carlito syrichta) |
| 005. XM_017517682.1 (Cebus capucinus) |
| 006. XM_012066401.1 (Cercocebus atys) |
| 007. XM_008020848.1 (Chlorocebus sabaeus) |
| 008. XM_004052118.2 (Gorilla gorilla) |
| 009. XM_015435350.1 (Macaca fascicularis) |
| 010. NM_001260901.1 (Macaca mulatta) |

| |
|---|
| 011. XM_012749592.2 (Microcebus murinus) |
| 012. XM_004090563.2 (Nomascus leucogenys) |
| 013. XM_003794732.3 (Otolemur garnettii) |
| 014. XM_003805460.4 (Pan paniscus) |
| 015. XM_016921980.2 (Pan troglodytes) |
| 016. XM_009247045.2 (Pongo abelii) |
| 017 XM_012659198.1 (Propithecus coquereli) |
| 018. XM_010382343.1 (Rhinopithecus roxellana) XM_010334574.1 (Saimiri boliviensis) |
| ***Euarchontoglires: Glires*** |
| 020. XM_020175744.1 (Castor canadensis) |
| 021. NM_001173076.1 (Cavia porcellus) |
| 022. XM_005378100.2 (Chinchilla lanigera) |
| 023. XM_016968238.1 (Cricetulus griseus) |
| 024. XM_010610684.2 (Fukomys damarensis) |

| |
|---|
| 025. NM_001279857.2 (Heterocephalus glaber) |
| 026. XM_004666484.2 (Jaculus jaculus) |
| 027. XM_015489896.1 (Marmota marmota marmota) |
| 028. XM_021658661.1 (Meriones unguiculatus) |
| 029. XM_013113005.2 (Mesocricetus auratus) |
| 030. XM_005347316.3 (Microtus ochrogaster) |
| 031. NM_001289785.1 (Mus musculus) |
| 032. XM_021206500.1 (Mus pahari) |
| 033. XM_008832795.1 (Nannospalax galili) |
| 034. XM_012926934.1 (Ochotona princeps) |
| 035. XM_004646354.3 (Octodon degus) |
| 036. NM_001082407.1 (Oryctolagus cuniculus) |
| 037. XM_015987087.1 (Peromyscus maniculatus) |
| 038. NM_012935.4 (Rattus norvegicus) |

| |
|---|
| ***Euarchontoglires: Scandentia*** |
| 039. XM_006150984.2 (Tupaia chinensis) |
| ***Euarchontoglires: Dermoptera: Cynocephalidae*** |
| 040. XM_008577609.1 (Galeopterus variegatus) |
| ***Boreoeutheria: Laurasiatheria*** |
| 041. NM_001304874.1 (Ailuropoda melanoleuca) |
| 042. XM_019975270.1 (Bos indicus) |
| 043. NM_174290.2 (Bos taurus) |

| |
|---|
| 044. XM_006190373.2 (Camelus ferus) |
| 045. KJ174932.1 (Camelus dromedarius) |

| |
|---|
| 046. XM_857165.5 (Canis lupus) |
| 047. XM_004427339.2 (Ceratotherium simum) |
| 048. XM_004689273.2 (Condylura cristata) |
| 049. XM_022561589.1 (Delphinapterus leucas) |

| |
|---|
| 050. XM_022512857.1 (Enhydra lutris) |
| 051. XM_008150161.1 (Eptesicus fuscus) |
| 052. XM_001501779.5 (Equus caballus) |
| 053. XM_007537847.2 (Erinaceus europaeus) |
| 054. XM_011286345.3 (Felis catus) |
| 055. XM_019653686.1 (Hipposideros armiger) |
| 056. XM_007464652.1 (Lipotes vexillifer) |
| 057. XM_017659392.1 (Manis javanica) |
| 058. XM_004749799.2 (Mustela putorius) |
| 059. XM_012567097.1 (Odobenus rosmarus) |
| 060. XM_004273414.2 (Orcinus orca) |
| 061. NM_001012457.2 (Ovis aries) |
| 062. XM_019469186.1 (Panthera pardus) |
| 063. XM_015542578.1 (Panthera tigris) |

| |
|---|
| 064. XM_007128267.2 (Physeter catodon) |
| 065. XM_006912821.1 (Pteropus alecto) |
| 066. XM_011365409.2 (Pteropus vampyrus) |
| 067. XM_016149055.1 (Rousettus aegyptiacus) |
| 068. XM_004604685.2 (Sorex araneus) |
| 069. XM_021062779.1 (Sus scrofa) |
| 070. XM_008690086.1 (Ursus maritimus) |
| 071. XM_006217871.2 (Vicugna pacos) |
| ***Eutheria: Afrotheria*** |
| 072. XM_006833817.1 (Chrysochloris asiatica) |
| 073. XM_004708973.2 (Echinops telfairi) |
| 074. XM_006890772.1 (Elephantulus edwardii) |
| 075. XM_003415624.3 (Loxodonta africana) |
| 076. XM_007936553.1 (Orycteropus afer) |

| |
|---|
| 077. ENSPCAT00000001647.1 (Procavia capensis) |
| 078. XM_004382499.3 (Trichechus manatus) |
| ***Theria: Metatheria*** |
| 079. XM_001371207.3 (Monodelphis domestica) |
| 080. XM_020989782.1 (Phascolarctos cinereus) |
| 081. XM_003764224.2 (Sarcophilus harrisii) |
| ***Eutheria: Xenarthra*** |
| 082. XM_004481144.3 (Dasypus novemcinctus) |
| ***Amniota: Sauropsida: Sauria*** |
| 083. XM_006271349.3 (Alligator mississippiensis) |
| 084. NM_001310366.1 (Anas platyrhynchos) |
| 085. XM_008121072.2 (Anolis carolinensis) |
| 086. XM_010169584.1 (Antrostomus carolinensis) |
| 087. XM_009870055.1 (Apaloderma vittatum) |
| 088. XM_011601882.1 (Aquila chrysaetos) |

| |
|---|
| 089. XM_026863824.1 (Athene cunicularia) |
| 090. XM_010298661.1 (Balearica regulorum) |
| 091. XM_010142606.1 (Buceros rhinoceros) |
| 092. XM_014949523.1 (Calidris pugnax) |
| 093. XM_008492420.1 (Calypte anna) |
| 094. XM_009696584.1 (Cariama cristata) |
| 095. XM_009997410.1 (Chaetura pelagica) |
| 096. XM_009882443.1 (Charadrius vociferus) |
| 097. XM_005291386.3 (Chrysemys picta) |
| 098. XM_021285292.1 (Columba livia) |
| 099. XM_008636923.2 (Corvus brachyrhynchos) |

| |
|---|
| 100. XM_015883834.1 (Coturnix japonica) |
| 101. XM_019547110.1 (Crocodylus porosus) |
| 102. XM_009569900.1 (Cuculus canorus) |

| | |
|---|---|
| 103. | XM_026105876.1 (Dromaius novaehollandiae) |
| 104. | XM_009638415.1 (Egretta garzetta) |
| 105. | XM_005440198.1 (Falco cherrug) |
| 106. | XM_005236602.1 (Falco peregrinus) |
| 107. | XM_005058800.1 (Ficedula albicollis) |
| 108. | NM_205176.1 (Gallus gallus) |
| 109. | XM_009810695.1 (Gavia stellata) |
| 110. | XM_015415588.1 (Gekko japonicus) |
| 111. | XM_010575003.1 (Haliaeetus leucocephalus) |
| 112. | XM_009917256.1 (Haliaeetus albicilla) |
| 113. | XM_017809764.1 (Lepidothrix coronata) |
| 114. | XM_021530106.1 (Lonchura striata) |
| 115. | XM_008935459.1 (Merops nubicus) |

| |
|---|
| 116. XM_009472271.1 (Nipponia nippon) XM_026041046.1 (Nothoprocta perdicaria) |
| 117. XM_009944637.1 (Opisthocomus hoazin) |
| 118. XM_015649982.1 (Parus major) |
| 119. XM_009485675.1 (Pelecanus crispus) |
| 120. XM_005531058.2 (Pseudopodoces humilis) 122. XM_009098028.2 (Serinus canaria) |
| 123. XM_002192884.3 (Taeniopygia guttata) |
| 124. XM_024207189.2 (Terrapene mexicana) |

| ***Tetrapoda: Amphibia: Batrachia: Anura*** |
|---|
| 125. XM_018564029.1 (Nanorana parkeri) |
| 126. XM_018225915.1 (Xenopus laevis) |
| 127. XM_002932918.4 (Xenopus tropicalis) |

| ***Sarcopterygii: Coelacanthimorpha: Coelacanthiformes: Coelacanthidae*** |
|---|
| 128. XM_005987256.2 (Latimeria chalumnae) |

| ***Teleostomi: Euteleostomi: Actinopterygii: Actinopteri: Neopterygii*** |
|---|

| | |
|---|---|
| 129. | ENSAPOT00000005575.1 (Acanthochromis polyacanthus) |
| 130. | ENSAMXT00000007540.2 (Astyanax mexicanus) |
| 131. | XM_026212446.1 (Carassius auratus) |
| 132. | 132. XM_012821983.1 (Clupea harengus) |
| 133. | XM_019095482.1 (Cyprinus carpio) |
| 134. | NM_131157.1 (Danio rerio) |
| 135. | XM_027031825.1 (Electrophorus electricus) |
| 136. | ENSIPUT00000034725.1 (Ictalurus punctatus) |
| 137. | XM_006642147.2 (Lepisosteus oculatus) |
| 138. | XM_026921851.1 (Pangasianodon hypophthalmus) |
| 139. | XM_023813558.1 (Paramormyrops kingsleyae) |
| 140. | XM_017699999.1 (Pygocentrus nattereri) |
| 141. | XM_018755058.1 (Scleropages formosus) |
| 142. | XM_016490540.1 (Sinocyclocheilus anshuiensis) |

| |
|---|
| 143. XM_016295944.1 (Sinocyclocheilus grahami) |
| 144. XM_016518971.1 (Sinocyclocheilus rhinocerous)<br><br>XM_027178619.1 (Tachysurus fulvidraco) |

| **Gnathostomata: Chondrichthyes** |
|---|
| 145. XM_007911619.1 (Callorhinchus milii) |
| 146. XM_020530352.1 (Rhincodon typus) |

| **HspB8** |
|---|
| ***Gnathostomata: Teleostomi: Euteleostomi: Sarcopterygii: Dipnotetrapodomorpha: Tetrapoda:*** |
| ***Amniota: Mammalia: Theria: Eutheria: Boreoeutheria: Euarchontoglires: Primates*** |
| 001. NM_014365.2 (Homo sapiens) |
| 002. ENSCJAT00000017713 (Callithrix jacchus) |
| 003. XM_008053630.1 (Carlito syrichta) |
| 004. XM_017529563.1 (Cebus capucinus) |
| 005. ENSCSAT00000000316 (Chlorocebus sabaeus) |
| 006. ENSGGOT00000013431 (Gorilla gorilla) |
| 007. XM_005572374.2 (Macaca fascicularis) |

| |
|---|
| 008. ENSMMUT00000029873 (Macaca mulatta) |
| 009. ENSMICT00000012392 (Microcebus murinus) |
| 010. ENSNLET00000000278 (Nomascus leucogenys) |
| 011. XM_003789928.1 (Otolemur garnettii) |
| 012. ENSPTRT00000010131 (Pan troglodytes) |
| 013. ENSPPYT00000005936 (Pongo abelii) |
| 014. XM_012650230.1 (Propithecus coquereli) |
| 015. XM_003932197.2 (Saimiri boliviensis) |
| ***Euarchontoglires: Glires*** |
| 016. XM_020163619.1 (Castor canadensis) |
| 017. XM_003478311.3 (Cavia porcellus) |
| 018. XM_005396529.2 (Chinchilla lanigera) |
| 019. XM_007630429.2 (Cricetulus griseus) |
| 020. XM_013014338.1 (Dipodomys ordii) |

| |
|---|
| 021. XM_010604080.2 (Fukomys damarensis) |
| 022. XM_004843691.2 (Heterocephalus glaber) |
| 023. XM_004664142.1 (Jaculus jaculus) |
| 024. XM_015505174.1 (Marmota marmota) |
| 025. XM_021631814.1 (Meriones unguiculatus) |
| 026. XM_005078985.3 (Mesocricetus auratus) |
| 027. XM_005344327.1 (Microtus ochrogaster) |
| 028. NM_030704.3 (Mus musculus) |
| 029. XM_021186952.1 (Mus pahari) |
| 030. XM_008824419.2 (Nannospalax galili) |
| 031. ENSOPRT00000016991 (Ochotona princeps) |
| 032. XM_004636239.2 (Octodon  degus) |
| 033. XM_006970817.2 (Peromyscus maniculatus) |
| 034. NM_053612.2 (Rattus norvegicus) |

| |
|---|
| ***Euarchontoglires: Scandentia*** |
| 035. XM_006151554.2 (Tupaia chinensis) |
| ***Euarchontoglires: Dermoptera: Cynocephalidae*** |
| 036. XM_008592760.1 (Galeopterus variegatus) |
| ***Boreoeutheria: Laurasiatheria*** |
| 037. XM_002919673.3 (Ailuropoda melanoleuca) |

| |
|---|
| XM_007189481.1 (Balaenoptera acutorostrata) |

| |
|---|
| 039. NM_001014955.1 (Bos taurus) |
| 040. XM_006192699.2 (Camelus ferus) |
| 041. NM_001003029.1 (Canis familiaris) |
| 042. XM_004429927.2 (Ceratotherium simum) |
| 043. XM_004690995.2 (Condylura cristata) |
| 044. XM_022597851.1 (Delphinapterus leucas) |
| 045. XM_022517444.1 (Enhydra lutris) |
| 046. XM_008142767.1 (Eptesicus fuscus) |

| |
|---|
| 047. XM_001490413.4 (Equus caballus) |
| 048. XM_007528642.2 (Erinaceus europaeus) |
| 049. XM_003994712.5 (Felis catus) |
| 050. XM_019649044.1 (Hipposideros armiger) |
| 051. XM_007464865.1 (Lipotes vexillifer) |
| 052. XM_017666754.1 (Manis javanica) |
| 053. XM_016201564.1 (Miniopterus natalensis) |
| 054. ENSMPUT00000002968 (Mustela putorius) |
| 055. XM_005863777.2 (Myotis brandtii) |
| 056. ENSMLUT00000016852 (Myotis lucifugus) |
| 057. XM_021703393.1 (Neomonachus schauinslandi) |
| 058. XM_004396682.2 (Odobenus rosmarus) |
| 059. XM_004276749.1 (Orcinus orca) |
| 060. XM_012110114.2 (Ovis aries) |

| |
|---|
| 061. XM_007079634.2 (Panthera tigris) |
| 062. XM_006908601.2 (Pteropus alecto) |
| 063. XM_007114647.2 (Physeter catodon) |
| 064. ENSPVAT00000017488 (Pteropus vampyrus) |
| 065. XM_016149531.1 (Rousettus aegyptiacus) |
| 066. XM_019753966.1 (Rhinolophus sinicus) |
| 067. XM_004611071.1 (Sorex araneus) |
| 068. XM_001929585.5 (Sus scrofa) |
| 069. XM_008707686.1 (Ursus maritimus) |
| 070. XM_006204799.2 (Vicugna pacos) |
| ***Eutheria: Afrotheria*** |
| 071. XM_006865455.1 (Chrysochloris asiatica) |
| 072. XM_006890415.1 (Elephantulus edwardii) |
| 073. XM_003419267.3 (Loxodonta africana) |

| |
|---|
| 074. XM_007938210.1 (Orycteropus afer) |
| 075. XM_004379029.1 (Trichechus manatus) |
| ***Theria: Metatheria*** |
| 076. ENSSHAT00000020774 (Sarcophilus harrisii) |
| ***Amniota: Sauropsida: Sauria*** |
| 077. XM_014593637.2 (Alligator mississippiensis) |
| 078. XM_006027370.2 (Alligator sinensis) |

| |
|---|
| 079. ENSACAT00000004096 (Anolis carolinensis) |
| 080. XM_011592666.1 (Aquila chrysaetos) |
| 081. XM_014965486.1 (Calidris pugnax) |
| 082. XM_008497491.1 (Calypte anna) |
| 083. XM_005303256.3 (Chrysemys picta) |
| 084. XM_017746547.1 (Corvus brachyrhynchos) |
| 085. XM_019539248 (Crocodylus porosus) |

| |
|---|
| 086. XM_009566591.1 (Cuculus canorus) |
| 087. XM_005445283.1 (Falco cherrug) |

| |
|---|
| 088. XM_005233215.1 (Falco peregrinus) |
| 089. XM_005054867.2 (Ficedula albicollis) |
| 090. ENSGALT00000011938 (Gallus gallus) |
| 091. XM_015425324.1 (Gekko japonicus) |
| 092. XM_010574425.1 (Haliaeetus leucocephalus) |
| 093. XM_017814476.1 (Lepidothrix coronata) |

| |
|---|
| XM_021549813.1 (Lonchura striata) |

| |
|---|
| 095. XM_009463728.1 (Nipponia nippon) |
| 096. XM_006134168.2 (Pelodiscus sinensis) |
| 097. XM_015817475.1 (Protobothrops mucrosquamatus) |

| |
|---|
| 098. XM_005524414.2 (Pseudopodoces humilis) |
| 099. XM_009092308.2 (Serinus canaria) |

| |
|---|
| 100. XM_002199452.1 (Taeniopygia guttata) |
| 101. XM_005493452.2 (Zonotrichia albicollis) |

| ***Tetrapoda: Amphibia: Batrachia: Anura*** |
|---|
| 102. XM_018570531.1 (Nanorana parkeri) |
| 103. NM_001086313.1 (Xenopus laevis) |
| 104. ETT00000054022 (Xenopus silurana) |

| ***Sarcopterygii: Coelacanthimorpha: Coelacanthiformes: Coelacanthidae*** |
|---|
| 105. ENSLACT00000015233 (Latimeria chalumnae) |

| ***Teleostomi: Euteleostomi: Actinopterygii: Actinopteri: Neopterygii*** |
|---|
| 106. XM_023277314.1 (Amphiprion ocellaris) |
| 107. XM_026344341.1 (Anabas testudineus) |
| 108. XM_014011157.1 (Austrofundulus limnaeus) |
| 109. XM_012820350.1 (Clupea harengus) |

| | |
|---|---|
| 110. | XM_015403675.1 (Cyprinodon variegatus) |
| 111. | XM_019104600.1 (Cyprinus carpio) |
| 112. | NM_001100957.2 (Danio rerio) |
| 113. | XM_013136582.2 (Esox lucius) |
| 114. | ENSGACT00000018865.1 (Gasterosteus aculeatus) |
| 115. | XM_017427754.2 (Kryptolebias marmoratus) |
| 116. | XM_010732969.2 (Larimichthys crocea) |
| 117. | XM_018703469.1 (Lates calcarifer) |
| 118. | ENSLOCT00000003798.1 (Lepisosteus oculatus) |
| 119. | XM_004566676.5 (Maylandia zebra) |
| 120. | XM_006803878.1 (Neolamprologus brichardi) |
| 121. | XM_015960520.1 (Nothobranchius furzeri) XM_021606608.1 (Oncorhynchus mykiss) |
| 122. | XM_021622152.1 (Oncorhynchus mykiss) |

| | |
|---|---|
| 123. | XM_020490486.1 (Oncorhynchus kisutch) |
| 124. | XM_020475741.1 (Oncorhynchus kisutch) |
| 125. | XM_024402897.1 (Oncorhynchus tshawytscha) |
| 126. | XM_024418238.1 (Oncorhynchus tshawytscha) |
| 127. | XM_003454160 (Oreochromis niloticus) |
| 128. | XM_004072074.4 (Oryzias latipes) |
| 129. | XM_020099324.1 (Paralichthys olivaceus) |
| 130. | ENSPFOT00000014634.1 (Poecilia formosa) |
| 131. | XM_014999665.1 (Poecilia mexicana) |
| 132. | XM_015059347 (Poecilia latipinna) |
| 133. | XM_014170575.1 (Salmo salar) |
| 134. | XM_014160033.1 (Salmo salar) |
| 135. | XM_024141472.1 (Salvelinus alpinus) |
| 136. | XM_024001519.2 (Salvelinus alpinus) |

| | |
|---|---|
| 137. | XM_022208447.1 (Scanthochromis polyacanthus) |
| 138. | XM_016444754.1 (Sinocyclocheilus anshuiensis) |
| 139. | XM_016528695.1 (Sinocyclocheilus rhinocerous) |
| 140. | XM_008279881.1 (Stegastes partitus) |
| 141. | XM_005803859.2 (Xiphophorus maculatus) |

| ***Gnathostomata: Chondrichthyes*** | |
|---|---|
| 142. | XM_007899064.1 (Callorhinchus milii) |
| 143. | XM_020525166.1 (Rhincodon typus) |

B.S3 Table**.** FEL analysis of the coding sequences of human HspB1, HspB3, HspB5 and HspB8. Abbreviations and symbols: LB, lower bound; MLE, maximum likelihood estimate; UB, upper bound; ND, not defined; PurS, purifying selection; NE, neutrally evolving (highlighted in gray); DivS, diversifying positive selection on (highlighted in blue); †, codon excluded from further processing**.**

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 1 | M | 0 | 0 | n/a | n/a | n/a | n/a | ND | † |
| 2 | T | 0.939 | 1.80E-01 | 8.21E-07 | 0.109 | 0.192 | 0.311 | PurS | |
| 3 | E | 0.54 | 0.00E+00 | 7.97E-07 | 0 | 0 | 0.032 | PurS | |
| 4 | R | 0.362 | 1.00E-02 | 1.19E-06 | 0.002 | 0.027 | 0.121 | PurS | |
| 5 | R | 0.721 | 0.194 | 0.000365 | 0.154 | 0.27 | 0.436 | PurS | |
| 6 | V | 0.624 | 2.90E-02 | 3.35E-09 | 0.011 | 0.046 | 0.12 | PurS | |
| 7 | P | 0.797 | 1.20E-02 | 5.54E-12 | 0.001 | 0.015 | 0.067 | PurS | |
| 8 | F | 0 | 0.042 | n/a | n/a | n/a | n/a | ND | † |
| 9 | S | 0.503 | 0.121 | 0.000938 | 0.124 | 0.241 | 0.419 | PurS | |
| 10 | L | 0.683 | 1.08E-01 | 9.56E-05 | 0.082 | 0.158 | 0.275 | PurS | |
| 11 | L | 0.731 | 1.03E-01 | 3.71E-06 | 0.064 | 0.141 | 0.265 | PurS | |
| 12 | R | 0.597 | 5.80E-02 | 7.15E-07 | 0.035 | 0.096 | 0.21 | PurS | |
| 13 | G | 0.873 | 0.217 | 0.001088 | 0.157 | 0.249 | 0.375 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 14 | P | 1.124 | 8.90E-02 | 2.41E-10 | 0.034 | 0.079 | 0.155 | PurS | |
| 15 | S | 0.293 | 0.048 | 0.030078 | 0.065 | 0.164 | 0.336 | PurS | |
| 16 | W | 0 | 0.011 | n/a | n/a | n/a | n/a | ND | † |
| 17 | D | 1.658 | 8.20E-02 | 8.48E-09 | 0.024 | 0.05 | 0.09 | PurS | |
| 18 | P | 0.893 | 0.00E+00 | 1.11E-16 | 0 | 0 | 0.026 | PurS | |
| 19 | F | 0.922 | 3.40E-02 | 6.20E-06 | 0.011 | 0.037 | 0.086 | PurS | |
| 20 | R | 0.821 | 7.00E-02 | 1.61E-09 | 0.033 | 0.085 | 0.174 | PurS | |
| 21 | D | 0.422 | 0.035 | 0.006346 | 0.026 | 0.083 | 0.197 | PurS | |
| 22 | W | 1.64 | 0.044 | 0.063044 | 0.008 | 0.027 | 0.063 | PurS | |
| 23 | Y | 1.175 | 0.21 | 0.002229 | 0.11 | 0.179 | 0.273 | PurS | |
| 24 | P | 1.429 | 1.97E-01 | 7.99E-08 | 0.08 | 0.138 | 0.22 | PurS | |
| 25 | H | 1.311 | 0.446 | 0.002575 | 0.232 | 0.34 | 0.482 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 26 | S | 0.983 | 7.10E-02 | 2.33E-05 | 0.034 | 0.072 | 0.131 | PurS | |
| 27 | R | 1.157 | 2.20E-02 | 3.23E-14 | 0.003 | 0.019 | 0.059 | PurS | |
| 28 | L | 0.431 | 0.084 | 0.001812 | 0.088 | 0.195 | 0.367 | PurS | |
| 29 | F | 1.224 | 2.50E-02 | 7.88E-13 | 0.005 | 0.02 | 0.053 | PurS | |
| 30 | D | 1.003 | 6.30E-02 | 5.35E-06 | 0.027 | 0.063 | 0.123 | PurS | |
| 31 | Q | 0.197 | 0.042 | 0.117112 | 0.066 | 0.212 | 0.497 | NE | |
| 32 | A | 0.837 | 0.298 | 0.002596 | 0.227 | 0.356 | 0.533 | PurS | |
| 33 | F | 1.631 | 0 | 0 | 0 | 0 | 0.01 | PurS | |
| 34 | G | 0.731 | 2.10E-02 | 4.22E-10 | 0.005 | 0.029 | 0.09 | PurS | |
| 35 | L | 0.742 | 0.121 | 0.001216 | 0.091 | 0.163 | 0.267 | PurS | |
| 36 | P | 1.89 | 0.049 | 0 | 0.008 | 0.026 | 0.061 | PurS | |
| 37 | R | 0.83 | 1.73E-01 | 3.48E-05 | 0.115 | 0.209 | 0.347 | PurS | |
| 38 | L | 0.616 | 0.244 | 0.015391 | 0.243 | 0.396 | 0.608 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 39 | P | 1.884 | 2.11E-01 | 6.12E-12 | 0.065 | 0.112 | 0.18 | PurS | |
| 40 | E | 1.443 | 3.70E-02 | 6.38E-10 | 0.008 | 0.026 | 0.061 | PurS | |
| 41 | E | 1.798 | 7.40E-02 | 7.62E-10 | 0.019 | 0.041 | 0.078 | PurS | |
| 42 | W | 1.094 | 0.119 | 0.000937 | 0.056 | 0.109 | 0.188 | PurS | |
| 43 | S | 1.418 | 0.463 | 0.001173 | 0.22 | 0.327 | 0.471 | PurS | |
| 44 | Q | 0.912 | 0.283 | 0.001585 | 0.199 | 0.311 | 0.463 | PurS | |
| 45 | W | 18.963 | 5.10E-02 | 9.46E-09 | 0.001 | 0.003 | 0.006 | PurS | |
| 46 | L | 0.981 | 1.32E-01 | 8.89E-07 | 0.069 | 0.134 | 0.233 | PurS | |
| 47 | G | 1.097 | 1.64E-01 | 4.01E-05 | 0.089 | 0.149 | 0.233 | PurS | |
| 48 | G | 0.681 | 0.352 | 0.088595 | 0.347 | 0.517 | 0.742 | PurS | |
| 49 | S | 0.781 | 0.132 | 0.009819 | 0.076 | 0.17 | 0.327 | PurS | |
| 50 | S | 1.565 | 2.34E-01 | 2.75E-06 | 0.093 | 0.15 | 0.227 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 51 | W | 24.794 | 0.022 | 0.007826 | 0 | 0.001 | 0.003 | PurS | |
| 52 | P | 1.47 | 0.012 | 0 | 0 | 0.008 | 0.037 | PurS | |
| 53 | G | 1.174 | 2.20E-02 | 3.44E-15 | 0.003 | 0.019 | 0.058 | PurS | |
| 54 | Y | 1.447 | 1.00E-02 | 5.26E-13 | 0 | 0.007 | 0.03 | PurS | |
| 55 | V | 0.692 | 0.259 | 0.023872 | 0.235 | 0.373 | 0.563 | PurS | |
| 56 | R | 1.859 | 0.014 | 0 | 0 | 0.008 | 0.032 | PurS | |
| 57 | P | 1.291 | 1.41E-01 | 3.75E-10 | 0.056 | 0.109 | 0.189 | PurS | |
| 58 | L | 0.884 | 0.394 | 0.019165 | 0.297 | 0.446 | 0.643 | PurS | |
| 59 | P | 0.818 | 0.393 | 0.038631 | 0.306 | 0.48 | 0.72 | PurS | |
| 60 | P | 3.144 | 4.40E-01 | 8.17E-11 | 0.091 | 0.14 | 0.207 | PurS | |
| 61 | A | 2.526 | 0.59 | 0.000109 | 0.152 | 0.234 | 0.35 | PurS | |
| 62 | A | 1.625 | 0.675 | 0.012651 | 0.287 | 0.416 | 0.586 | PurS | |
| 63 | I | 1.372 | 0.524 | 0.0221 | 0.262 | 0.382 | 0.54 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 64 | E | 2.359 | 0.68 | 0.004997 | 0.211 | 0.288 | 0.388 | PurS | |
| 65 | S | 2.277 | 0.898 | 0.00291 | 0.294 | 0.394 | 0.527 | PurS | |
| 66 | P | 1.381 | 2.50E-01 | 4.35E-07 | 0.108 | 0.181 | 0.284 | PurS | |
| 67 | A | 3.759 | 5.81E-01 | 4.96E-07 | 0.107 | 0.155 | 0.218 | PurS | |
| 68 | V | 1.431 | 0.654 | 0.022452 | 0.322 | 0.457 | 0.632 | PurS | |
| 69 | A | 1.612 | 4.05E-01 | 7.90E-06 | 0.165 | 0.251 | 0.367 | PurS | |
| 70 | A | 1.7 | 0.69 | 0.001617 | 0.287 | 0.406 | 0.562 | PurS | |
| | P | 1.423 | 2.90E-01 | 2.79E-06 | 0.127 | 0.204 | 0.31 | PurS | |
| 72 | A | 0.991 | 5.67E-01 | 0.139144 | 0.401 | 0.572 | 0.792 | NE | |
| 73 | Y | 0.896 | 2.29E-01 | 0.004463 | 0.162 | 0.256 | 0.384 | PurS | |
| 74 | S | 1.26 | 2.06E-01 | 5.98E-07 | 0.1 | 0.164 | 0.253 | PurS | |
| 75 | R | 1.448 | 0.12 | 1.29E-12 | 0.041 | 0.083 | 0.148 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 76 | A | 1.106 | 1.11E-01 | 5.52E-09 | 0.048 | 0.101 | 0.184 | PurS | |
| 77 | L | 0.529 | 7.50E-02 | 4.87E-05 | 0.056 | 0.142 | 0.29 | PurS | |
| 78 | S | 0.824 | 1.59E-01 | 0.000309 | 0.112 | 0.193 | 0.308 | PurS | |
| 79 | R | 1.214 | 0.035 | 5.88E-15 | 0.007 | 0.029 | 0.075 | PurS | |
| 80 | Q | 2.338 | 2.10E-02 | 3.94E-14 | 0.001 | 0.009 | 0.027 | PurS | |
| 81 | L | 0.958 | 8.50E-02 | 6.21E-08 | 0.04 | 0.089 | 0.167 | PurS | |
| 82 | S | 0.34 | 0.032 | 0.008182 | 0.029 | 0.093 | 0.216 | PurS | |
| 83 | S | 2.778 | 0.114 | 1.01E-13 | 0.022 | 0.041 | 0.068 | PurS | |
| 84 | G | 0.975 | 0.00E+00 | 2.89E-15 | 0 | 0 | 0.021 | PurS | |
| 85 | V | 0.575 | 0.248 | 0.118893 | 0.283 | 0.431 | 0.628 | NE | |
| 86 | S | 1.128 | 1.20E-02 | 0 | 0.001 | 0.011 | 0.047 | PurS | |
| 87 | E | 0.441 | 0 | 9.10E-06 | 0 | 0 | 0.04 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 88 | I | 0.305 | 0.059 | 0.02827 | 0.083 | 0.193 | 0.377 | PurS | |
| 89 | R | 1.063 | 0.063 | 1.88E-10 | 0.023 | 0.059 | 0.12 | PurS | |
| 90 | H | 0.678 | 1.81E-01 | 0.042527 | 0.155 | 0.267 | 0.425 | PurS | |
| 91 | T | 1.82 | 0.07 | 0 | 0.017 | 0.039 | 0.076 | PurS | |
| 92 | A | 1.149 | 0.39 | 0.000526 | 0.224 | 0.34 | 0.497 | PurS | |
| 93 | D | 0.896 | 0.054 | 1.86E-05 | 0.024 | 0.061 | 0.124 | PurS | |
| 94 | R | 0.463 | 2.85E-01 | 0.375464 | 0.394 | 0.615 | 0.918 | NE | |
| 95 | W | 5.15 | 2.20E-02 | 0.006792 | 0.001 | 0.004 | 0.013 | PurS | |
| 96 | R | 2.305 | 2.70E-02 | 0 | 0.003 | 0.012 | 0.03 | PurS | |
| 97 | V | 0.874 | 6.70E-02 | 1.52E-08 | 0.03 | 0.076 | 0.154 | PurS | |
| 98 | S | 0.773 | 1.25E-01 | 4.14E-05 | 0.086 | 0.162 | 0.276 | PurS | |
| 99 | L | 0.508 | 1.40E-02 | 2.08E-08 | 0.002 | 0.027 | 0.118 | PurS | |
| 100 | D | 1.174 | 2.70E-02 | 1.11E-15 | 0.006 | 0.023 | 0.06 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 101 | V | 0.548 | 4.40E-02 | 3.74E-06 | 0.025 | 0.081 | 0.189 | PurS | |
| 102 | N | 0.781 | 0.072 | 7.56E-05 | 0.044 | 0.093 | 0.168 | PurS | |
| 103 | H | 0.945 | 6.70E-02 | 9.68E-06 | 0.03 | 0.071 | 0.138 | PurS | |
| 104 | F | 1.261 | 0.027 | 1.39E-14 | 0.005 | 0.021 | 0.054 | PurS | |
| 105 | A | 0.554 | 1.32E-01 | 0.000359 | 0.119 | 0.239 | 0.423 | PurS | |
| 106 | P | 0.947 | 1.20E-02 | 6.66E-16 | 0.001 | 0.013 | 0.057 | PurS | |
| 107 | D | 0.713 | 5.60E-02 | 0.000161 | 0.031 | 0.079 | 0.161 | PurS | |
| 108 | E | 1.019 | 3.90E-02 | 1.26E-06 | 0.012 | 0.039 | 0.09 | PurS | |
| 109 | L | 0.461 | 2.80E-02 | 1.66E-06 | 0.01 | 0.061 | 0.189 | PurS | |
| 110 | T | 1.018 | 1.84E-01 | 6.79E-07 | 0.105 | 0.18 | 0.286 | PurS | |
| 111 | V | 1.137 | 9.00E-02 | 1.22E-10 | 0.036 | 0.079 | 0.148 | PurS | |
| 112 | K | 1.03 | 2.40E-02 | 2.07E-09 | 0.006 | 0.024 | 0.062 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 113 | T | 0.723 | 0.031 | 2.90E-09 | 0.011 | 0.042 | 0.111 | PurS | |
| 114 | K | 0.921 | 0.05 | 6.89E-07 | 0.022 | 0.054 | 0.111 | PurS | |
| 115 | D | 1.026 | 1.32E-01 | 9.10E-09 | 0.072 | 0.128 | 0.209 | PurS | |
| 116 | G | 1.295 | 8.50E-02 | 3.18E-14 | 0.03 | 0.065 | 0.123 | PurS | |
| 117 | V | 0.737 | 1.36E-01 | 0.0002 | 0.098 | 0.185 | 0.314 | PurS | |
| 118 | V | 0.286 | 1.21E-01 | 0.113726 | 0.211 | 0.423 | 0.744 | NE | |
| 119 | E | 0.983 | 2.70E-02 | 4.60E-11 | 0.007 | 0.028 | 0.072 | PurS | |
| 120 | I | 1.83 | 1.70E-02 | 4.44E-16 | 0.002 | 0.009 | 0.029 | PurS | |
| 121 | T | 1.262 | 1.83E-01 | 3.68E-11 | 0.086 | 0.145 | 0.228 | PurS | |
| 122 | G | 0.748 | 1.10E-02 | 1.37E-11 | 0.001 | 0.014 | 0.062 | PurS | |
| 123 | K | 1.278 | 4.80E-02 | 4.76E-13 | 0.015 | 0.037 | 0.076 | PurS | |
| 124 | H | 1.1 | 5.70E-02 | 2.11E-12 | 0.021 | 0.052 | 0.106 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 125 | E | 0.78 | 0.059 | 4.36E-08 | 0.03 | 0.076 | 0.155 | PurS | |
| 126 | E | 1.871 | 3.70E-02 | 1.83E-12 | 0.006 | 0.02 | 0.046 | PurS | |
| 127 | R | 1.226 | 0.054 | 1.11E-16 | 0.017 | 0.044 | 0.09 | PurS | |
| 128 | Q | 1.362 | 1.00E-01 | 1.77E-08 | 0.037 | 0.074 | 0.13 | PurS | |
| 129 | D | 1.856 | 0.037 | 0 | 0.006 | 0.02 | 0.047 | PurS | |
| 130 | E | 2.047 | 5.60E-02 | 3.33E-16 | 0.011 | 0.027 | 0.056 | PurS | |
| 131 | H | 1.77 | 3.60E-02 | 0 | 0.006 | 0.02 | 0.048 | PurS | |
| 132 | G | 0.897 | 5.50E-02 | 1.15E-11 | 0.022 | 0.061 | 0.133 | PurS | |
| 133 | Y | 0.906 | 1.91E-01 | 0.00028 | 0.13 | 0.21 | 0.319 | PurS | |
| 134 | I | 1.228 | 1.21E-01 | 6.86E-09 | 0.054 | 0.098 | 0.163 | PurS | |
| 135 | S | 0.961 | 2.20E-02 | 1.76E-13 | 0.004 | 0.023 | 0.072 | PurS | |
| 136 | R | 0.796 | 2.90E-02 | 1.39E-13 | 0.009 | 0.037 | 0.097 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 137 | C | 1.502 | 0.124 | 5.82E-08 | 0.045 | 0.083 | 0.137 | PurS | |
| 138 | F | 0.86 | 4.00E-02 | 3.16E-05 | 0.015 | 0.047 | 0.105 | PurS | |
| 139 | T | 0.658 | 4.10E-02 | 6.18E-09 | 0.019 | 0.062 | 0.146 | PurS | |
| 140 | R | 0.89 | 1.00E-02 | 2.22E-16 | 0.001 | 0.012 | 0.051 | PurS | |
| 141 | K | 0.284 | 3.20E-02 | 0.000141 | 0.034 | 0.111 | 0.259 | PurS | |
| 142 | Y | 0.884 | 2.90E-02 | 6.01E-07 | 0.008 | 0.033 | 0.087 | PurS | |
| 143 | T | 0.705 | 0.114 | 1.92E-06 | 0.083 | 0.162 | 0.28 | PurS | |
| 144 | L | 0.817 | 0.00E+00 | 2.89E-14 | 0 | 0 | 0.031 | PurS | |
| 145 | P | 0.799 | 0.00E+00 | 4.99E-13 | 0 | 0 | 0.032 | PurS | |
| 146 | P | 1.109 | 0.198 | 6.77E-09 | 0.1 | 0.178 | 0.292 | PurS | |
| 147 | G | 0.37 | 1.62E-01 | 0.026478 | 0.245 | 0.439 | 0.717 | PurS | |
| 148 | V | 0.804 | 0.103 | 1.06E-08 | 0.061 | 0.128 | 0.232 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 149 | D | 2.022 | 1.28E-01 | 1.71E-12 | 0.035 | 0.063 | 0.103 | PurS | |
| 150 | P | 0.789 | 2.87E-01 | 0.001729 | 0.228 | 0.363 | 0.548 | PurS | |
| 151 | T | 1.495 | 1.24E-01 | 2.20E-11 | 0.044 | 0.083 | 0.141 | PurS | |
| 152 | Q | 0.712 | 2.61E-01 | 0.00621 | 0.231 | 0.367 | 0.552 | PurS | |
| 153 | V | 0.915 | 4.10E-02 | 4.51E-10 | 0.014 | 0.045 | 0.105 | PurS | |
| 154 | S | 1.348 | 2.75E-01 | 2.78E-06 | 0.128 | 0.204 | 0.306 | PurS | |
| 155 | S | 0.588 | 3.40E-02 | 1.28E-07 | 0.014 | 0.057 | 0.149 | PurS | |
| 156 | S | 1.123 | 0.07 | 8.11E-12 | 0.025 | 0.062 | 0.128 | PurS | |
| 157 | L | 0.653 | 2.60E-02 | 9.64E-10 | 0.007 | 0.04 | 0.124 | PurS | |
| 158 | S | 0.679 | 1.10E-02 | 2.02E-12 | 0.001 | 0.016 | 0.072 | PurS | |
| 159 | P | 0.622 | 0 | 8.24E-14 | 0 | 0 | 0.037 | PurS | |
| 160 | E | 1.747 | 6.90E-02 | 1.81E-12 | 0.017 | 0.04 | 0.077 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 161 | G | 1.438 | 2.20E-02 | 0 | 0.003 | 0.016 | 0.048 | PurS | |
| 162 | T | 1.364 | 2.20E-01 | 2.71E-08 | 0.098 | 0.162 | 0.248 | PurS | |
| 163 | L | 0.644 | 1.30E-02 | 2.73E-10 | 0.001 | 0.02 | 0.089 | PurS | |
| 164 | T | 1.262 | 3.00E-02 | 0 | 0.006 | 0.024 | 0.061 | PurS | |
| 165 | V | 0.562 | 0.043 | 5.64E-07 | 0.024 | 0.077 | 0.181 | PurS | |
| 166 | E | 1.155 | 2.90E-02 | 2.72E-11 | 0.006 | 0.025 | 0.065 | PurS | |
| 167 | A | 1.087 | 4.80E-02 | 7.31E-14 | 0.014 | 0.044 | 0.104 | PurS | |
| 168 | P | 1.149 | 0.037 | 6.66E-16 | 0.008 | 0.032 | 0.084 | PurS | |
| 169 | M | 1.001 | 2.01E-01 | 2.49E-05 | 0.116 | 0.201 | 0.32 | PurS | |
| 170 | P | 0.375 | 1.50E-01 | 0.073742 | 0.213 | 0.4 | 0.678 | PurS | |
| 171 | K | 1.156 | 9.80E-02 | 1.60E-08 | 0.044 | 0.085 | 0.147 | PurS | |
| 172 | L | 0.925 | 1.44E-01 | 1.57E-08 | 0.08 | 0.155 | 0.27 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 173 | A | 0.419 | 8.70E-02 | 0.001972 | 0.088 | 0.207 | 0.403 | PurS | |
| 174 | T | 1.659 | 0.081 | 1.19E-14 | 0.023 | 0.049 | 0.089 | PurS | |
| 175 | Q | 0.966 | 8.70E-02 | 2.60E-05 | 0.041 | 0.09 | 0.168 | PurS | |
| 176 | S | 1.319 | 1.74E-01 | 1.07E-10 | 0.073 | 0.132 | 0.218 | PurS | |
| 177 | N | 1.065 | 0.243 | 1.35E-06 | 0.134 | 0.229 | 0.359 | PurS | |
| 178 | E | 1.657 | 2.00E-02 | 3.31E-14 | 0.002 | 0.012 | 0.037 | PurS | |
| 179 | I | 0.678 | 8.80E-02 | 0.00017 | 0.062 | 0.13 | 0.236 | PurS | |
| 180 | T | 1.235 | 8.70E-02 | 6.82E-12 | 0.032 | 0.071 | 0.132 | PurS | |
| 181 | I | 1.271 | 0.009 | 6.66E-15 | 0 | 0.007 | 0.032 | PurS | |
| 182 | P | 1.025 | 0.00E+00 | 0 | 0 | 0 | 0.023 | PurS | |
| 183 | V | 0.933 | 7.80E-02 | 2.60E-09 | 0.036 | 0.083 | 0.162 | PurS | |
| 184 | T | 0.937 | 0.099 | 1.50E-07 | 0.053 | 0.105 | 0.187 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 185 | F | 1.183 | 1.94E-01 | 1.15E-06 | 0.099 | 0.164 | 0.253 | PurS | |
| 186 | E | 1.088 | 0.2 | 0.001631 | 0.111 | 0.184 | 0.283 | PurS | |
| 187 | S | 1.823 | 0.433 | 1.76E-05 | 0.159 | 0.238 | 0.343 | PurS | |
| 188 | R | 1.116 | 0.271 | 1.36E-05 | 0.151 | 0.243 | 0.371 | PurS | |
| 189 | A | 0.536 | 0.234 | 0.040467 | 0.241 | 0.436 | 0.716 | PurS | |
| 190 | Q | 2.223 | 0.209 | 1.70E-07 | 0.052 | 0.094 | 0.155 | PurS | |
| 191 | L | 0.852 | 0.267 | 0.001371 | 0.189 | 0.313 | 0.49 | PurS | |
| 192 | G | 1.374 | 2.67E-01 | 0.000134 | 0.105 | 0.195 | 0.331 | PurS | |
| 193 | G | 0.654 | 1.87E-01 | 0.006046 | 0.139 | 0.286 | 0.526 | PurS | |
| 194 | P | 0.924 | 4.27E-01 | 0.034146 | 0.263 | 0.463 | 0.76 | PurS | |
| 195 | E | 0.596 | 1.86E-01 | 0.029479 | 0.164 | 0.313 | 0.543 | PurS | |
| 196 | A | 0.768 | 0.32 | 0.032965 | 0.225 | 0.417 | 0.714 | PurS | |

| Hsp B1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dNdS LB** | **dNdS MLE** | **dNdS UB** | **selection** | |
| 197 | A | 1.19 | 0.295 | 0.001497 | 0.146 | 0.248 | 0.395 | PurS | |
| 198 | K | 1.221 | 0.056 | 7.66E-08 | 0.016 | 0.046 | 0.099 | PurS | |
| 199 | S | 2.577 | 4.75E-01 | 3.75E-06 | 0.115 | 0.184 | 0.284 | PurS | |
| 200 | D | 1.674 | 0.245 | 1.83E-05 | 0.088 | 0.146 | 0.229 | PurS | |
| 201 | E | 2.754 | 0.383 | 2.75E-05 | 0.088 | 0.139 | 0.21 | PurS | |
| 202 | T | 0.638 | 0.538 | 0.62552 | 0.547 | 0.844 | 1.256 | NE | |
| 203 | A | 0.577 | 0.516 | 0.734621 | 0.611 | 0.894 | 1.265 | NE | |
| 204 | A | 0.693 | 0.326 | 0.097775 | 0.311 | 0.47 | 0.68 | PurS | |
| 205 | K | 1.081 | 0.117 | 2.95E-06 | 0.059 | 0.109 | 0.182 | PurS | |

| Hsp B3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE (omega)** | **dN/dS UB** | **selection** | |
| 1 | M | 0 | 0 | n/a | n/a | n/a | n/a | ND | † |
| 2 | A | 0.689 | 0.132 | 0.000604 | 0.082 | 0.191 | 0.374 | PurS | |

| Hsp B3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE (omega)** | **dN/dS UB** | **selection** | |
| 3 | K | 0.318 | 0.414 | 0.612466 | 0.802 | 1.299 | 1.981 | NE | |
| 4 | I | 0.555 | 0.268 | 0.10483 | 0.263 | 0.483 | 0.806 | NE | |
| 5 | I | 1.164 | 0.282 | 0.000161 | 0.138 | 0.242 | 0.393 | PurS | |
| 6 | L | 1.36 | 0.179 | 6.87E-07 | 0.062 | 0.132 | 0.241 | PurS | |
| 7 | R | 0.947 | 0.118 | 4.94E-06 | 0.053 | 0.124 | 0.243 | PurS | |
| 8 | H | 0.25 | 0 | 0.000789 | 0 | 0 | 0.113 | PurS | |
| 9 | L | 0.613 | 0.224 | 0.070153 | 0.184 | 0.365 | 0.639 | PurS | |
| 10 | I | 1.38 | 0.105 | 1.31E-07 | 0.03 | 0.076 | 0.157 | PurS | |
| 11 | E | 0.251 | 0.312 | 0.713835 | 0.71 | 1.246 | 2.017 | NE | |
| 12 | I | 0.942 | 0.108 | 4.79E-07 | 0.049 | 0.115 | 0.224 | PurS | |
| 13 | P | 1.105 | 0 | 4.20E-14 | 0 | 0 | 0.026 | PurS | |
| 14 | V | 0.872 | 0.089 | 3.60E-06 | 0.032 | 0.102 | 0.241 | PurS | |
| 15 | R | 1.353 | 0 | 2.33E-15 | 0 | 0 | 0.024 | PurS | |
| 16 | Y | 1.399 | 0.065 | 1.50E-07 | 0.011 | 0.046 | 0.121 | PurS | |
| 17 | Q | 1.101 | 0.1 | 1.18E-05 | 0.033 | 0.091 | 0.198 | PurS | |
| 18 | E | 0.541 | 0.251 | 0.160981 | 0.248 | 0.465 | 0.788 | NE | |
| 19 | E | 0.91 | 0.496 | 0.188972 | 0.342 | 0.545 | 0.824 | NE | |
| 20 | F | 0.778 | 0.073 | 2.40E-05 | 0.029 | 0.094 | 0.22 | PurS | |
| 21 | E | 0.701 | 0.318 | 0.073819 | 0.26 | 0.454 | 0.737 | PurS | |
| 22 | A | 0.837 | 0.581 | 0.303172 | 0.452 | 0.694 | 1.024 | NE | |
| 23 | R | 0.959 | 0.238 | 0.000258 | 0.127 | 0.249 | 0.434 | PurS | |
| 24 | G | 0.884 | 0.21 | 0.000965 | 0.122 | 0.237 | 0.413 | PurS | |
| 25 | L | 1.278 | 0.049 | 2.67E-10 | 0.006 | 0.039 | 0.121 | PurS | |
| 26 | E | 0.825 | 0.256 | 0.012887 | 0.165 | 0.31 | 0.527 | PurS | |
| 27 | D | 0.954 | 0.292 | 0.007884 | 0.174 | 0.306 | 0.496 | PurS | |
| 28 | C | 0.459 | 0.146 | 0.062671 | 0.153 | 0.319 | 0.579 | PurS | |
| 29 | R | 1.765 | 0.17 | 1.63E-09 | 0.049 | 0.096 | 0.176 | PurS | |

| Hsp B3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE (omega)** | **dN/dS UB** | **selection** | |
| 30 | L | 0.319 | 0.073 | 0.01824 | 0.057 | 0.229 | 0.601 | PurS | |
| 31 | D | 1.122 | 0.128 | 8.41E-06 | 0.049 | 0.114 | 0.223 | PurS | |
| 32 | H | 2.367 | 0 | 0 | 0 | 0 | 0.013 | PurS | |
| 33 | A | 1.044 | 0.51 | 0.039376 | 0.314 | 0.488 | 0.723 | PurS | |
| 34 | L | 1.954 | 0.027 | 5.77E-15 | 0.001 | 0.014 | 0.06 | PurS | |
| 35 | Y | 2.922 | 0.081 | 1.03E-13 | 0.009 | 0.028 | 0.065 | PurS | |
| 36 | A | 1.349 | 0.017 | 1.78E-14 | 0.001 | 0.012 | 0.054 | PurS | |
| 37 | L | 0.295 | 0 | 0.000136 | 0 | 0 | 0.167 | PurS | |
| 38 | P | 3.177 | 0 | 0 | 0 | 0 | 0.009 | PurS | |
| 39 | G | 0.957 | 0.018 | 1.22E-10 | 0.001 | 0.019 | 0.084 | PurS | |
| 40 | P | 0.571 | 0.031 | 1.04E-06 | 0.009 | 0.054 | 0.166 | PurS | |
| 41 | T | 0.742 | 0.362 | 0.050813 | 0.3 | 0.488 | 0.749 | PurS | |
| 42 | I | 1.721 | 0.571 | 0.000857 | 0.217 | 0.332 | 0.485 | PurS | |
| 43 | V | 0.698 | 0.702 | 0.990205 | 0.65 | 1.005 | 1.503 | NE | |
| 44 | D | 1.566 | 0.312 | 0.000159 | 0.114 | 0.2 | 0.323 | PurS | |
| 45 | L | 0.489 | 0.665 | 0.420895 | 0.838 | 1.361 | 2.097 | NE | |
| 46 | R | 1.04 | 0.599 | 0.145834 | 0.392 | 0.577 | 0.824 | NE | |
| 47 | K | 0.725 | 0.593 | 0.63272 | 0.534 | 0.818 | 1.193 | NE | |
| 48 | T | 0.352 | 1.068 | 0.009986 | 2.132 | 3.036 | 4.231 | DivS | |
| 49 | R | 1.248 | 0.53 | 0.026246 | 0.272 | 0.424 | 0.634 | PurS | |
| 50 | A | 2.692 | 0.774 | 0.001895 | 0.173 | 0.287 | 0.456 | PurS | |
| 51 | A | 2.003 | 0.944 | 0.016864 | 0.318 | 0.471 | 0.676 | PurS | |
| 52 | Q | 1.335 | 1.139 | 0.69843 | 0.578 | 0.853 | 1.238 | NE | |
| 53 | S | 0.771 | 1.276 | 0.184576 | 1.144 | 1.654 | 2.349 | NE | |
| 54 | P | 1.633 | 0.75 | 0.022294 | 0.307 | 0.46 | 0.666 | PurS | |
| 55 | P | 3.108 | 0.803 | 1.41E-05 | 0.174 | 0.258 | 0.375 | PurS | |

| Hsp B3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE (omega)** | **dN/dS UB** | **selection** | |
| 56 | V | 1.3 | 1.129 | 0.679405 | 0.58 | 0.869 | 1.271 | NE | |
| 57 | D | 1.218 | 0.741 | 0.288474 | 0.406 | 0.608 | 0.883 | NE | |
| 58 | S | 1.234 | 0.712 | 0.093031 | 0.376 | 0.577 | 0.855 | PurS | |
| 59 | A | 3.773 | 0.813 | 2.01E-06 | 0.145 | 0.215 | 0.314 | PurS | |
| 60 | A | 1.777 | 1.078 | 0.138705 | 0.412 | 0.607 | 0.87 | NE | |
| 61 | E | 1.64 | 1.085 | 0.315841 | 0.451 | 0.662 | 0.943 | NE | |
| 62 | T | 1.866 | 1.286 | 0.341367 | 0.491 | 0.689 | 0.952 | NE | |
| 63 | P | 1.084 | 0.841 | 0.459303 | 0.54 | 0.775 | 1.087 | NE | |
| 64 | P | 1.775 | 0.63 | 0.00139 | 0.237 | 0.355 | 0.513 | PurS | |
| 65 | R | 1.227 | 1.234 | 1 | 0.706 | 1.005 | 1.401 | NE | |
| 66 | E | 0.426 | 0.596 | 0.456166 | 0.901 | 1.398 | 2.08 | NE | |
| 67 | G | 0.697 | 0.517 | 0.471394 | 0.469 | 0.741 | 1.109 | NE | |
| 68 | K | 1.713 | 0.521 | 0.002998 | 0.197 | 0.304 | 0.45 | PurS | |
| 69 | S | 0.816 | 0.606 | 0.381937 | 0.499 | 0.743 | 1.073 | NE | |
| 70 | H | 0.965 | 0.831 | 0.659106 | 0.596 | 0.862 | 1.21 | NE | |
| 71 | F | 1.408 | 0.035 | 2.06E-10 | 0.004 | 0.025 | 0.076 | PurS | |
| 72 | Q | 0.897 | 0.126 | 0.000179 | 0.056 | 0.141 | 0.289 | PurS | |
| 73 | I | 0.513 | 0.208 | 0.066749 | 0.202 | 0.405 | 0.723 | PurS | |
| 74 | L | 0.792 | 0.025 | 8.80E-08 | 0.002 | 0.031 | 0.14 | PurS | |
| 75 | L | 0.406 | 0.078 | 0.004346 | 0.048 | 0.193 | 0.508 | PurS | |
| 76 | D | 4.885 | 0 | 0 | 0 | 0 | 0.007 | PurS | |
| 77 | V | 0.594 | 0.044 | 3.04E-05 | 0.012 | 0.074 | 0.23 | PurS | |
| 78 | V | 0.497 | 0.043 | 0.000104 | 0.014 | 0.086 | 0.267 | PurS | |
| 79 | Q | 0.536 | 0 | 8.50E-06 | 0 | 0 | 0.071 | PurS | |
| 80 | F | 1.353 | 0 | 1.10E-10 | 0 | 0 | 0.023 | PurS | |
| 81 | L | 0.816 | 0.116 | 6.09E-05 | 0.056 | 0.143 | 0.293 | PurS | |

| Hsp B3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE (omega)** | **dN/dS UB** | **selection** | |
| 82 | P | 1.849 | 0.016 | 0 | 0 | 0.008 | 0.037 | PurS | |
| 83 | E | 0.733 | 0 | 2.83E-08 | 0 | 0 | 0.048 | PurS | |
| 84 | D | 1.632 | 0.035 | 1.07E-11 | 0.004 | 0.021 | 0.066 | PurS | |
| 85 | I | 0.444 | 0.05 | 0.000527 | 0.028 | 0.113 | 0.297 | PurS | |
| 86 | I | 0.352 | 0.052 | 0.003904 | 0.037 | 0.148 | 0.386 | PurS | |
| 87 | I | 0.745 | 0 | 3.97E-09 | 0 | 0 | 0.043 | PurS | |
| 88 | Q | 0.243 | 0 | 0.001879 | 0 | 0 | 0.157 | PurS | |
| 89 | T | 0.607 | 0.161 | 0.003534 | 0.128 | 0.266 | 0.484 | PurS | |
| 90 | F | 4.093 | 0.034 | 0 | 0.001 | 0.008 | 0.025 | PurS | |
| 91 | E | 1.353 | 0 | 2.35E-11 | 0 | 0 | 0.026 | PurS | |
| 92 | G | 0.226 | 0 | 0.000254 | 0 | 0 | 0.153 | PurS | |
| 93 | W | 0 | 0 | n/a | n/a | n/a | n/a | ND | † |
| 94 | L | 0.42 | 0 | 2.02E-06 | 0 | 0 | 0.099 | PurS | |
| 95 | L | 0.352 | 0.024 | 0.000533 | 0.004 | 0.07 | 0.308 | PurS | |
| 96 | I | 1.466 | 0.141 | 1.27E-07 | 0.044 | 0.096 | 0.182 | PurS | |
| 97 | K | 1.07 | 0.051 | 3.69E-08 | 0.012 | 0.048 | 0.124 | PurS | |
| 98 | A | 1.044 | 0.088 | 1.89E-08 | 0.03 | 0.084 | 0.182 | PurS | |
| 99 | Q | 2.131 | 0.122 | 1.49E-09 | 0.023 | 0.057 | 0.117 | PurS | |
| 100 | H | 4.398 | 0 | 0 | 0 | 0 | 0.007 | PurS | |
| 101 | G | 0.674 | 0.079 | 2.24E-05 | 0.036 | 0.117 | 0.273 | PurS | |
| 102 | T | 0.409 | 0.315 | 0.559709 | 0.455 | 0.771 | 1.215 | NE | |
| 103 | R | 0.928 | 0.016 | 9.58E-11 | 0.001 | 0.017 | 0.075 | PurS | |
| 104 | M | 0 | 0 | n/a | n/a | n/a | n/a | ND | † |
| 105 | D | 5.743 | 0.017 | 0 | 0 | 0.003 | 0.013 | PurS | |
| 106 | E | 1.336 | 0.018 | 2.05E-10 | 0.001 | 0.014 | 0.062 | PurS | |
| 107 | H | 2.707 | 0.016 | 7.77E-15 | 0 | 0.006 | 0.026 | PurS | |
| 108 | G | 1.728 | 0 | 0 | 0 | 0 | 0.02 | PurS | |
| 109 | F | 1.278 | 0.035 | 1.66E-09 | 0.005 | 0.027 | 0.084 | PurS | |
| 110 | I | 0.648 | 0.185 | 0.009526 | 0.142 | 0.286 | 0.507 | PurS | |
| 111 | S | 1.504 | 0.075 | 9.31E-12 | 0.016 | 0.05 | 0.118 | PurS | |

| Hsp B3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE (omega)** | **dN/dS UB** | **selection** | |
| 112 | R | 1.353 | 0 | 3.76E-14 | 0 | 0 | 0.023 | PurS | |
| 113 | S | 0.732 | 0.065 | 5.64E-05 | 0.032 | 0.089 | 0.192 | PurS | |
| 114 | F | 0.92 | 0.016 | 2.85E-08 | 0.001 | 0.018 | 0.079 | PurS | |
| 115 | T | 0.783 | 0.077 | 8.74E-07 | 0.035 | 0.098 | 0.211 | PurS | |
| 116 | R | 1.353 | 0 | 3.84E-14 | 0 | 0 | 0.025 | PurS | |
| 117 | Q | 0.916 | 0.06 | 5.20E-06 | 0.016 | 0.065 | 0.169 | PurS | |
| 118 | Y | 1.713 | 0 | 9.32E-13 | 0 | 0 | 0.024 | PurS | |
| 119 | K | 0.957 | 0.141 | 6.87E-05 | 0.067 | 0.147 | 0.278 | PurS | |
| 120 | L | 0.606 | 0 | 9.31E-08 | 0 | 0 | 0.082 | PurS | |
| 121 | P | 0.95 | 0.016 | 2.83E-11 | 0.001 | 0.016 | 0.071 | PurS | |
| 122 | D | 1.434 | 0.361 | 0.000765 | 0.152 | 0.252 | 0.392 | PurS | |
| 123 | G | 1.103 | 0.144 | 1.72E-06 | 0.055 | 0.13 | 0.254 | PurS | |
| 124 | V | 1.7 | 0.463 | 0.00011 | 0.164 | 0.273 | 0.424 | PurS | |
| 125 | E | 1.123 | 0.307 | 0.002228 | 0.152 | 0.273 | 0.451 | PurS | |
| 126 | I | 0.592 | 0.224 | 0.023645 | 0.202 | 0.379 | 0.64 | PurS | |
| 127 | K | 0.799 | 0.195 | 0.002183 | 0.126 | 0.244 | 0.423 | PurS | |
| 128 | D | 0.652 | 0.017 | 4.07E-06 | 0.002 | 0.027 | 0.118 | PurS | |
| 129 | L | 0.8 | 0.026 | 3.17E-07 | 0.002 | 0.033 | 0.145 | PurS | |
| 130 | S | 1.058 | 0.178 | 1.14E-05 | 0.084 | 0.168 | 0.298 | PurS | |
| 131 | A | 1.623 | 0 | 0 | 0 | 0 | 0.02 | PurS | |
| 132 | V | 0.453 | 0.503 | 0.809258 | 0.716 | 1.11 | 1.653 | NE | |
| 133 | L | 0.272 | 0.113 | 0.141451 | 0.164 | 0.415 | 0.848 | NE | |
| 134 | C | 1.353 | 0 | 1.27E-12 | 0 | 0 | 0.022 | PurS | |
| 135 | H | 2.167 | 0 | 3.33E-16 | 0 | 0 | 0.014 | PurS | |
| 136 | D | 0.923 | 0 | 3.79E-10 | 0 | 0 | 0.038 | PurS | |
| 137 | G | 0.763 | 0 | 2.22E-10 | 0 | 0 | 0.046 | PurS | |
| 138 | I | 0.506 | 0.034 | 5.17E-05 | 0.011 | 0.067 | 0.208 | PurS | |
| 139 | L | 0.58 | 0 | 2.79E-07 | 0 | 0 | 0.088 | PurS | |
| 140 | V | 0.312 | 0.045 | 0.008808 | 0.024 | 0.143 | 0.445 | PurS | |
| 141 | V | 0.272 | 0.046 | 0.02439 | 0.028 | 0.169 | 0.523 | PurS | |

| Hsp B3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE (omega)** | **dN/dS UB** | **selection** | |
| 142 | E | 0.597 | 0.097 | 0.001908 | 0.058 | 0.163 | 0.355 | PurS | |
| 143 | V | 1.767 | 0.431 | 6.66E-05 | 0.151 | 0.244 | 0.374 | PurS | |
| 144 | K | 0.797 | 0.099 | 0.000805 | 0.045 | 0.124 | 0.265 | PurS | |
| 145 | D | 1.013 | 0.522 | 0.1279 | 0.328 | 0.515 | 0.773 | NE | |
| 146 | P | 1.496 | 0.538 | 0.00253 | 0.226 | 0.36 | 0.544 | PurS | |
| 147 | V | 0.982 | 1.727 | 0.090552 | 1.241 | 1.76 | 2.46 | DivS | |
| 148 | G | 1.176 | 0.452 | 0.00749 | 0.233 | 0.384 | 0.603 | PurS | |
| 149 | T | 1.457 | 0.67 | 0.018887 | 0.309 | 0.46 | 0.662 | PurS | |
| 150 | K | 1.064 | 0.816 | 0.535735 | 0.517 | 0.766 | 1.101 | NE | |

| Hsp B5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 1 | M | 0 | 0 | n/a | n/a | n/a | n/a | ND | † |
| 2 | D | 0.464 | 0.016 | 8.21E-05 | 0.002 | 0.034 | 0.152 | PurS | |
| 3 | I | 0.759 | 0.094 | 9.81E-05 | 0.049 | 0.123 | 0.252 | PurS | |
| 4 | A | 0.534 | 0.169 | 0.017232 | 0.151 | 0.316 | 0.579 | PurS | |
| 5 | I | 1.093 | 0.031 | 2.74E-09 | 0.005 | 0.028 | 0.087 | PurS | |
| 6 | H | 1.361 | 0.072 | 2.50E-06 | 0.016 | 0.053 | 0.124 | PurS | |
| 7 | H | 1.08 | 0.207 | 0.000421 | 0.102 | 0.191 | 0.326 | PurS | |
| 8 | P | 1.039 | 0 | 3.28E-12 | 0 | 0 | 0.033 | PurS | |
| 9 | W | 1.147 | 0.038 | 0.000182 | 0.006 | 0.033 | 0.104 | PurS | |
| 10 | I | 1.006 | 0.208 | 0.001876 | 0.111 | 0.207 | 0.351 | PurS | |
| 11 | R | 1.5 | 0 | 5.55E-16 | 0 | 0 | 0.023 | PurS | |
| 12 | R | 1.476 | 0.017 | 1.09E-13 | 0.001 | 0.012 | 0.051 | PurS | |
| 13 | P | 0.832 | 0.098 | 2.49E-05 | 0.042 | 0.118 | 0.255 | PurS | |
| 14 | F | 1.604 | 0.22 | 4.03E-06 | 0.075 | 0.137 | 0.229 | PurS | |
| 15 | F | 2.02 | 0.309 | 1.12E-05 | 0.089 | 0.153 | 0.244 | PurS | |
| 16 | P | 0.796 | 0.163 | 0.00019 | 0.093 | 0.204 | 0.383 | PurS | |
| 17 | F | 1.519 | 0.211 | 3.76E-05 | 0.072 | 0.139 | 0.241 | PurS | |

| Hsp B5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 18 | H | 0.633 | 0.33 | 0.217278 | 0.306 | 0.521 | 0.83 | NE | |
| 19 | S | 1.54 | 0.253 | 3.85E-06 | 0.09 | 0.165 | 0.275 | PurS | |
| 20 | P | 1.353 | 0 | 6.89E-13 | 0 | 0 | 0.026 | PurS | |
| 21 | S | 1.158 | 0.178 | 0.000151 | 0.079 | 0.154 | 0.266 | PurS | |
| 22 | R | 1.125 | 0.018 | 8.88E-12 | 0.001 | 0.016 | 0.072 | PurS | |
| 23 | L | 0.054 | 0.083 | 0.682006 | 0.546 | 1.535 | 3.328 | NE | |
| 24 | F | 0.751 | 0.073 | 9.02E-06 | 0.035 | 0.098 | 0.212 | PurS | |
| 25 | D | 0.838 | 0.117 | 0.001104 | 0.059 | 0.139 | 0.272 | PurS | |
| 26 | Q | 0.661 | 0.041 | 0.001081 | 0.01 | 0.062 | 0.192 | PurS | |
| 27 | F | 0.833 | 0.307 | 0.026922 | 0.217 | 0.368 | 0.583 | PurS | |
| 28 | F | 8.155 | 0.028 | 0 | 0.001 | 0.003 | 0.011 | PurS | |
| 29 | G | 1.163 | 0.018 | 8.09E-13 | 0.001 | 0.016 | 0.069 | PurS | |
| 30 | E | 2.028 | 0.033 | 1.65E-11 | 0.003 | 0.016 | 0.05 | PurS | |
| 31 | H | 1.219 | 0.104 | 1.78E-05 | 0.034 | 0.085 | 0.176 | PurS | |
| 32 | L | 0.712 | 0.105 | 9.99E-05 | 0.052 | 0.147 | 0.321 | PurS | |
| 33 | L | 1.826 | 0.397 | 4.55E-05 | 0.123 | 0.217 | 0.357 | PurS | |
| 34 | E | 0.957 | 0.051 | 5.30E-05 | 0.013 | 0.053 | 0.14 | PurS | |
| 35 | S | 0.864 | 0.288 | 0.003324 | 0.193 | 0.334 | 0.535 | PurS | |
| 36 | D | 1.387 | 0.159 | 1.54E-05 | 0.055 | 0.115 | 0.21 | PurS | |
| 37 | L | 0.684 | 0.173 | 0.002663 | 0.114 | 0.253 | 0.477 | PurS | |
| 38 | F | 1.473 | 0.104 | 4.64E-07 | 0.028 | 0.071 | 0.144 | PurS | |
| 39 | P | 1.471 | 0.17 | 1.08E-08 | 0.053 | 0.115 | 0.215 | PurS | |
| 40 | T | 1.062 | 0.389 | 0.004335 | 0.221 | 0.366 | 0.573 | PurS | |
| 41 | S | 1.052 | 0.159 | 5.07E-05 | 0.072 | 0.151 | 0.278 | PurS | |
| 42 | T | 0.846 | 0.291 | 0.010596 | 0.19 | 0.344 | 0.577 | PurS | |
| 43 | S | 1.077 | 0.337 | 0.010209 | 0.179 | 0.313 | 0.509 | PurS | |
| 44 | L | 0.766 | 0.372 | 0.13855 | 0.244 | 0.485 | 0.872 | NE | |
| 45 | S | 2.862 | 0.202 | 3.71E-08 | 0.034 | 0.071 | 0.129 | PurS | |
| 46 | P | 0.94 | 0.162 | 5.09E-05 | 0.078 | 0.172 | 0.326 | PurS | |
| 47 | F | 0.362 | 0.016 | 0.001554 | 0.003 | 0.043 | 0.197 | PurS | |
| 48 | Y | 0.8 | 0.193 | 0.005884 | 0.12 | 0.241 | 0.43 | PurS | |
| 49 | L | 1.049 | 0.432 | 0.018328 | 0.257 | 0.411 | 0.623 | PurS | |
| 50 | R | 0.632 | 0.018 | 9.14E-08 | 0.002 | 0.029 | 0.128 | PurS | |
| 51 | P | 1.407 | 0.439 | 0.000613 | 0.194 | 0.312 | 0.476 | PurS | |
| 52 | P | 1.11 | 0.196 | 7.13E-05 | 0.088 | 0.177 | 0.314 | PurS | |

| Hsp B5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 53 | S | 3.43 | 0.059 | 4.97E-10 | 0.001 | 0.017 | 0.077 | PurS | |
| 54 | F | 1.878 | 0.152 | 4.34E-07 | 0.04 | 0.081 | 0.144 | PurS | |
| 55 | L | 1.305 | 0.303 | 0.000125 | 0.128 | 0.232 | 0.383 | PurS | |
| 56 | R | 0.879 | 0.035 | 6.99E-08 | 0.007 | 0.04 | 0.125 | PurS | |
| 57 | A | 0.695 | 0.647 | 0.879723 | 0.619 | 0.931 | 1.351 | NE | |
| 58 | P | 0.84 | 0.136 | 6.54E-05 | 0.069 | 0.162 | 0.317 | PurS | |
| 59 | S | 0.358 | 0.159 | 0.213992 | 0.229 | 0.444 | 0.769 | NE | |
| 60 | W | 10000 | 0.035 | 0.589362 | 0 | 0 | 0 | ND | † |
| 61 | F | 0.83 | 0.469 | 0.116165 | 0.352 | 0.565 | 0.858 | NE | |
| 62 | D | 0.567 | 0.153 | 0.045666 | 0.129 | 0.27 | 0.496 | PurS | |
| 63 | T | 1.024 | 0.204 | 7.56E-05 | 0.106 | 0.199 | 0.339 | PurS | |
| 64 | G | 0.901 | 0.072 | 1.49E-07 | 0.025 | 0.08 | 0.186 | PurS | |
| 65 | L | 0.62 | 0.139 | 0.002768 | 0.095 | 0.223 | 0.439 | PurS | |
| 66 | S | 0.944 | 0.059 | 1.85E-08 | 0.016 | 0.063 | 0.164 | PurS | |
| 67 | E | 0.24 | 0 | 0.003876 | 0 | 0 | 0.132 | PurS | |
| 68 | M | 1.792 | 0.037 | 0.002361 | 0.005 | 0.021 | 0.054 | PurS | |
| 69 | R | 0.926 | 0.077 | 1.87E-07 | 0.026 | 0.084 | 0.197 | PurS | |
| 70 | L | 0.331 | 0.262 | 0.656399 | 0.429 | 0.791 | 1.335 | NE | |
| 71 | E | 0.882 | 0.23 | 0.01965 | 0.142 | 0.26 | 0.436 | PurS | |
| 72 | K | 1.009 | 0.047 | 4.37E-06 | 0.012 | 0.047 | 0.122 | PurS | |
| 73 | D | 0.414 | 0.016 | 0.00042 | 0.002 | 0.039 | 0.171 | PurS | |
| 74 | R | 1.72 | 0.117 | 2.29E-09 | 0.029 | 0.068 | 0.133 | PurS | |
| 75 | F | 1.161 | 0.014 | 3.21E-11 | 0.001 | 0.012 | 0.054 | PurS | |
| 76 | S | 1.408 | 0.245 | 2.38E-06 | 0.093 | 0.174 | 0.295 | PurS | |
| 77 | V | 0.702 | 0.068 | 4.94E-06 | 0.03 | 0.097 | 0.228 | PurS | |
| 78 | N | 1.454 | 0.093 | 1.22E-08 | 0.025 | 0.064 | 0.13 | PurS | |
| 79 | L | 0.786 | 0.044 | 4.72E-07 | 0.009 | 0.056 | 0.174 | PurS | |
| 80 | D | 0.861 | 0 | 2.21E-10 | 0 | 0 | 0.035 | PurS | |
| 81 | V | 0.492 | 0.019 | 2.16E-05 | 0.002 | 0.038 | 0.167 | PurS | |
| 82 | K | 1.616 | 0 | 1.09E-10 | 0 | 0 | 0.018 | PurS | |
| 83 | H | 0.902 | 0.049 | 5.25E-07 | 0.013 | 0.054 | 0.142 | PurS | |
| 84 | F | 0.422 | 0 | 1.54E-05 | 0 | 0 | 0.064 | PurS | |
| 85 | S | 0.879 | 0.123 | 2.06E-05 | 0.06 | 0.14 | 0.274 | PurS | |
| 86 | P | 1.068 | 0.018 | 6.09E-12 | 0.001 | 0.017 | 0.075 | PurS | |
| 87 | E | 1.18 | 0.033 | 2.11E-07 | 0.005 | 0.028 | 0.087 | PurS | |

| Hsp B5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 88 | E | 3.036 | 0.065 | 3.44E-15 | 0.007 | 0.022 | 0.051 | PurS | |
| 89 | L | 0.702 | 0.022 | 1.94E-07 | 0.002 | 0.032 | 0.141 | PurS | |
| 90 | K | 1.377 | 0.226 | 3.13E-05 | 0.087 | 0.164 | 0.282 | PurS | |
| 91 | V | 0.874 | 0.018 | 2.05E-08 | 0.001 | 0.021 | 0.092 | PurS | |
| 92 | K | 1.421 | 0.047 | 6.93E-08 | 0.008 | 0.033 | 0.085 | PurS | |
| 93 | V | 0.652 | 0.091 | 0.000128 | 0.05 | 0.14 | 0.305 | PurS | |
| 94 | L | 1.915 | 0.398 | 4.42E-06 | 0.123 | 0.208 | 0.331 | PurS | |
| 95 | G | 1.391 | 0.045 | 3.54E-12 | 0.006 | 0.033 | 0.094 | PurS | |
| 96 | D | 1.4 | 0.048 | 9.24E-10 | 0.009 | 0.035 | 0.09 | PurS | |
| 97 | V | 0.64 | 0.165 | 0.019232 | 0.129 | 0.258 | 0.456 | PurS | |
| 98 | I | 0.422 | 0.106 | 0.010093 | 0.107 | 0.252 | 0.492 | PurS | |
| 99 | E | 1.451 | 0.066 | 5.45E-09 | 0.014 | 0.046 | 0.107 | PurS | |
| 100 | V | 1.186 | 0.122 | 1.59E-07 | 0.044 | 0.103 | 0.203 | PurS | |
| 101 | H | 3.295 | 0.12 | 1.09E-12 | 0.016 | 0.036 | 0.071 | PurS | |
| 102 | G | 0.995 | 0.054 | 1.87E-08 | 0.013 | 0.055 | 0.142 | PurS | |
| 103 | K | 0.902 | 0.063 | 2.45E-06 | 0.022 | 0.07 | 0.164 | PurS | |
| 104 | H | 3.303 | 0 | 0 | 0 | 0 | 0.009 | PurS | |
| 105 | E | 1.334 | 0.016 | 9.98E-11 | 0.001 | 0.012 | 0.054 | PurS | |
| 106 | E | 5.768 | 0.049 | 0 | 0.002 | 0.009 | 0.022 | PurS | |
| 107 | R | 0.963 | 0 | 5.68E-12 | 0 | 0 | 0.035 | PurS | |
| 108 | Q | 0 | 0 | n/a | n/a | n/a | n/a | ND | † |
| 109 | D | 0.763 | 0 | 6.56E-10 | 0 | 0 | 0.04 | PurS | |
| 110 | E | 1.183 | 0.049 | 1.80E-07 | 0.01 | 0.042 | 0.109 | PurS | |
| 111 | H | 1.799 | 0.016 | 1.70E-14 | 0.001 | 0.009 | 0.04 | PurS | |
| 112 | G | 0.966 | 0 | 6.79E-12 | 0 | 0 | 0.034 | PurS | |
| 113 | F | 1.794 | 0.177 | 6.70E-07 | 0.051 | 0.099 | 0.171 | PurS | |
| 114 | I | 0.852 | 0.181 | 0.00061 | 0.105 | 0.212 | 0.376 | PurS | |
| 115 | S | 0.655 | 0.036 | 2.06E-06 | 0.009 | 0.055 | 0.171 | PurS | |
| 116 | R | 0.874 | 0 | 9.39E-12 | 0 | 0 | 0.036 | PurS | |
| 117 | E | 1.147 | 0.016 | 4.12E-07 | 0.001 | 0.014 | 0.064 | PurS | |
| 118 | F | 0.703 | 0 | 9.26E-08 | 0 | 0 | 0.039 | PurS | |
| 119 | H | 1.178 | 0.355 | 0.009477 | 0.179 | 0.301 | 0.474 | PurS | |
| 120 | R | 0.877 | 0 | 3.98E-11 | 0 | 0 | 0.033 | PurS | |
| 121 | K | 1.161 | 0.046 | 6.16E-09 | 0.01 | 0.04 | 0.104 | PurS | |
| 122 | Y | 0.919 | 0 | 5.29E-08 | 0 | 0 | 0.037 | PurS | |

| Hsp B5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 123 | R | 1.409 | 0.107 | 1.45E-08 | 0.03 | 0.076 | 0.155 | PurS | |
| 124 | I | 1.61 | 0.211 | 1.30E-06 | 0.069 | 0.131 | 0.223 | PurS | |
| 125 | P | 0.682 | 0 | 4.10E-10 | 0 | 0 | 0.051 | PurS | |
| 126 | A | 0.941 | 0.282 | 0.00135 | 0.166 | 0.3 | 0.494 | PurS | |
| 127 | D | 1.918 | 0.033 | 2.11E-15 | 0.003 | 0.017 | 0.054 | PurS | |
| 128 | V | 0.699 | 0.036 | 8.91E-07 | 0.009 | 0.052 | 0.163 | PurS | |
| 129 | D | 1.425 | 0.068 | 7.75E-09 | 0.015 | 0.047 | 0.111 | PurS | |
| 130 | P | 0.698 | 0.018 | 1.47E-08 | 0.001 | 0.026 | 0.116 | PurS | |
| 131 | L | 1.334 | 0.387 | 0.000695 | 0.168 | 0.29 | 0.466 | PurS | |
| 132 | T | 1.08 | 0.625 | 0.117146 | 0.379 | 0.579 | 0.844 | NE | |
| 133 | I | 0.631 | 0.122 | 0.001298 | 0.088 | 0.194 | 0.365 | PurS | |
| 134 | T | 0.732 | 0.033 | 1.27E-07 | 0.007 | 0.045 | 0.14 | PurS | |
| 135 | S | 1.153 | 0.056 | 4.87E-10 | 0.012 | 0.049 | 0.127 | PurS | |
| 136 | S | 1.293 | 0.071 | 5.19E-10 | 0.017 | 0.055 | 0.13 | PurS | |
| 137 | L | 0.405 | 0.046 | 0.000728 | 0.018 | 0.113 | 0.348 | PurS | |
| 138 | S | 0.828 | 0 | 8.00E-12 | 0 | 0 | 0.04 | PurS | |
| 139 | S | 0.591 | 0.095 | 0.000234 | 0.057 | 0.16 | 0.349 | PurS | |
| 140 | D | 0.623 | 0 | 1.55E-08 | 0 | 0 | 0.049 | PurS | |
| 141 | G | 1.884 | 0.017 | 7.77E-16 | 0.001 | 0.009 | 0.04 | PurS | |
| 142 | V | 0.356 | 0.055 | 0.004949 | 0.038 | 0.155 | 0.404 | PurS | |
| 143 | L | 0.435 | 0 | 2.06E-06 | 0 | 0 | 0.098 | PurS | |
| 144 | T | 0.83 | 0.05 | 4.49E-08 | 0.015 | 0.06 | 0.156 | PurS | |
| 145 | V | 1.003 | 0.089 | 1.37E-06 | 0.031 | 0.088 | 0.192 | PurS | |
| 146 | N | 1.368 | 0.279 | 3.17E-05 | 0.118 | 0.204 | 0.33 | PurS | |
| 147 | G | 0.355 | 0.093 | 0.013072 | 0.094 | 0.263 | 0.57 | PurS | |
| 148 | P | 0.749 | 0.037 | 1.09E-07 | 0.008 | 0.049 | 0.153 | PurS | |
| 149 | R | 1.105 | 0 | 1.97E-11 | 0 | 0 | 0.027 | PurS | |
| 150 | K | 1.052 | 0.216 | 0.000561 | 0.112 | 0.205 | 0.342 | PurS | |
| 151 | Q | 0.694 | 0.663 | 0.915204 | 0.601 | 0.956 | 1.458 | NE | |
| 152 | V | 1.121 | 0.404 | 0.011081 | 0.22 | 0.361 | 0.558 | PurS | |
| 153 | S | 2.044 | 0.371 | 2.24E-06 | 0.112 | 0.182 | 0.28 | PurS | |
| 154 | G | 0.794 | 0.267 | 0.011134 | 0.182 | 0.336 | 0.565 | PurS | |
| 155 | P | 1.663 | 0.114 | 1.03E-10 | 0.027 | 0.069 | 0.14 | PurS | |
| 156 | E | 1.055 | 0 | 8.02E-09 | 0 | 0 | 0.03 | PurS | |
| 157 | R | 0.736 | 0.036 | 7.39E-07 | 0.008 | 0.049 | 0.153 | PurS | |

| Hsp B5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 158 | T | 0.509 | 0.495 | 0.94903 | 0.63 | 0.972 | 1.448 | NE | |
| 159 | I | 0.717 | 0.015 | 3.19E-08 | 0.001 | 0.021 | 0.093 | PurS | |
| 160 | P | 0.84 | 0.093 | 8.07E-06 | 0.04 | 0.111 | 0.241 | PurS | |
| 161 | I | 0.55 | 0 | 2.31E-07 | 0 | 0 | 0.052 | PurS | |
| 162 | T | 1.143 | 0.152 | 1.30E-06 | 0.063 | 0.133 | 0.243 | PurS | |
| 163 | R | 1.244 | 0.074 | 7.83E-10 | 0.018 | 0.06 | 0.139 | PurS | |
| 164 | E | 1 | 0.151 | 0.000123 | 0.072 | 0.151 | 0.275 | PurS | |
| 165 | E | 0.972 | 0.188 | 0.003976 | 0.099 | 0.193 | 0.336 | PurS | |
| 166 | K | 3.021 | 0.016 | 2.42E-14 | 0 | 0.005 | 0.024 | PurS | |
| 167 | P | 0.594 | 0.143 | 0.001539 | 0.102 | 0.24 | 0.47 | PurS | |
| 168 | A | 0.912 | 0.105 | 5.77E-06 | 0.041 | 0.116 | 0.253 | PurS | |
| 169 | V | 0.835 | 0.305 | 0.021004 | 0.201 | 0.366 | 0.609 | PurS | |
| 170 | T | 1.031 | 0.384 | 0.006385 | 0.214 | 0.372 | 0.605 | PurS | |
| 171 | A | 0.626 | 0.153 | 0.002736 | 0.104 | 0.245 | 0.482 | PurS | |
| 172 | A | 0.627 | 0.423 | 0.324325 | 0.406 | 0.675 | 1.061 | NE | |
| 173 | P | 1.588 | 0.234 | 1.43E-05 | 0.076 | 0.147 | 0.256 | PurS | |
| 174 | K | 0.956 | 0.117 | 0.000342 | 0.053 | 0.123 | 0.239 | PurS | |
| 175 | K | 1.353 | 0 | 4.24E-10 | 0 | 0 | 0.022 | PurS | |

| Hsp B8 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 1 | M | 0 | 0 | n/a | n/a | n/a | n/a | ND | † |
| 2 | A | 0.585 | 0.034 | 3.58E-07 | 0.01 | 0.059 | 0.182 | PurS | |
| 3 | D | 3.454 | 0.043 | 0 | 0.003 | 0.012 | 0.032 | PurS | |
| 4 | G | 0.5 | 0.094 | 0.000811 | 0.075 | 0.188 | 0.383 | PurS | |
| 5 | Q | 0.831 | 0.031 | 6.27E-06 | 0.006 | 0.037 | 0.116 | PurS | |
| 6 | M | 2.524 | 0.246 | 0.001441 | 0.053 | 0.097 | 0.162 | PurS | |
| 7 | P | 0.858 | 0 | 1.92E-08 | 0 | 0 | 0.055 | PurS | |
| 8 | F | 1.607 | 0 | 1.08E-10 | 0 | 0 | 0.022 | PurS | |
| 9 | S | 0.867 | 0.229 | 0.00559 | 0.126 | 0.264 | 0.483 | PurS | |
| 10 | C | 1.353 | 0 | 1.04E-06 | 0 | 0 | 0.029 | PurS | |
| 11 | H | 1.135 | 0.157 | 1.13E-05 | 0.071 | 0.138 | 0.239 | PurS | |
| 12 | Y | 1.55 | 0.125 | 1.15E-06 | 0.037 | 0.081 | 0.152 | PurS | |

| Hsp B8 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 13 | P | 1.545 | 0.112 | 7.43E-12 | 0.031 | 0.072 | 0.141 | PurS | |
| 14 | S | 0.793 | 0.192 | 0.000842 | 0.128 | 0.242 | 0.41 | PurS | |
| 15 | R | 0.864 | 0.03 | 1.37E-09 | 0.006 | 0.035 | 0.109 | PurS | |
| 16 | L | 0.617 | 0.146 | 0.007011 | 0.107 | 0.237 | 0.445 | PurS | |
| 17 | R | 0.623 | 0 | 2.35E-09 | 0 | 0 | 0.045 | PurS | |
| 18 | R | 1.818 | 0 | 0 | 0 | 0 | 0.017 | PurS | |
| 19 | D | 0.639 | 0.014 | 1.03E-06 | 0.001 | 0.021 | 0.095 | PurS | |
| 20 | P | 1.081 | 0.016 | 8.60E-13 | 0.001 | 0.015 | 0.066 | PurS | |
| 21 | F | 0.988 | 0.024 | 5.43E-09 | 0.004 | 0.025 | 0.077 | PurS | |
| 22 | R | 1.042 | 0.085 | 3.29E-09 | 0.029 | 0.082 | 0.178 | PurS | |
| 23 | D | 1.06 | 0.074 | 1.09E-06 | 0.025 | 0.07 | 0.151 | PurS | |
| 24 | S | 1.468 | 0.18 | 1.72E-08 | 0.061 | 0.122 | 0.217 | PurS | |
| 25 | P | 0.795 | 0.244 | 0.00362 | 0.171 | 0.307 | 0.505 | PurS | |
| 26 | L | 1.158 | 0.053 | 1.21E-09 | 0.011 | 0.046 | 0.119 | PurS | |
| 27 | S | 1.156 | 0.25 | 3.55E-05 | 0.121 | 0.216 | 0.355 | PurS | |
| 28 | S | 1.096 | 0.031 | 1.32E-12 | 0.005 | 0.029 | 0.089 | PurS | |
| 29 | R | 0.983 | 0.016 | 9.20E-12 | 0.001 | 0.017 | 0.074 | PurS | |
| 30 | L | 0.516 | 0.031 | 4.05E-06 | 0.01 | 0.06 | 0.185 | PurS | |
| 31 | L | 0.193 | 0.031 | 0.02092 | 0.027 | 0.161 | 0.501 | PurS | |
| 32 | D | 3.78 | 0.087 | 7.77E-16 | 0.009 | 0.023 | 0.047 | PurS | |
| 33 | D | 0.367 | 0.07 | 0.00384 | 0.068 | 0.19 | 0.411 | PurS | |
| 34 | G | 0.875 | 0.124 | 9.70E-05 | 0.064 | 0.141 | 0.265 | PurS | |
| 35 | F | 0.675 | 0 | 2.49E-10 | 0 | 0 | 0.034 | PurS | |
| 36 | G | 0.492 | 0.068 | 0.0003 | 0.042 | 0.138 | 0.321 | PurS | |
| 37 | M | 4.038 | 0.077 | 1.41E-05 | 0.008 | 0.019 | 0.037 | PurS | |
| 38 | D | 1.203 | 0.147 | 2.98E-06 | 0.056 | 0.122 | 0.228 | PurS | |
| 39 | P | 0.647 | 0.085 | 1.90E-05 | 0.047 | 0.131 | 0.285 | PurS | |
| 40 | F | 1.353 | 0 | 2.29E-12 | 0 | 0 | 0.017 | PurS | |
| 41 | P | 1.618 | 0.177 | 4.31E-09 | 0.055 | 0.11 | 0.194 | PurS | |
| 42 | D | 2.351 | 0.161 | 4.80E-09 | 0.034 | 0.068 | 0.122 | PurS | |
| 43 | D | 1.356 | 0.07 | 2.09E-08 | 0.018 | 0.051 | 0.112 | PurS | |
| 44 | L | 1.019 | 0.037 | 5.11E-10 | 0.006 | 0.036 | 0.114 | PurS | |
| 45 | T | 1.732 | 0.125 | 3.78E-12 | 0.033 | 0.072 | 0.137 | PurS | |
| 46 | A | 0.838 | 0.115 | 4.53E-05 | 0.062 | 0.137 | 0.256 | PurS | |
| 47 | S | 0.927 | 0.276 | 0.003047 | 0.178 | 0.298 | 0.465 | PurS | |

| Hsp B8 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 48 | W | 0 | 0.016 | n/a | n/a | n/a | n/a | ND | † |
| 49 | P | 1.878 | 0.081 | 8.77E-15 | 0.015 | 0.043 | 0.094 | PurS | |
| 50 | D | 1.126 | 0.118 | 1.22E-05 | 0.048 | 0.105 | 0.197 | PurS | |
| 51 | W | 0 | 0 | n/a | n/a | n/a | n/a | ND | † |
| 52 | A | 0.631 | 0.069 | 4.39E-06 | 0.034 | 0.11 | 0.258 | PurS | |
| 53 | L | 0.686 | 0.037 | 1.15E-07 | 0.009 | 0.053 | 0.166 | PurS | |
| 54 | P | 0.798 | 0.033 | 7.68E-09 | 0.007 | 0.042 | 0.129 | PurS | |
| 55 | R | 1.274 | 0.033 | 6.65E-14 | 0.004 | 0.026 | 0.081 | PurS | |
| 56 | L | 0.794 | 0.08 | 9.56E-07 | 0.036 | 0.101 | 0.218 | PurS | |
| 57 | S | 0.369 | 0.263 | 0.46021 | 0.414 | 0.714 | 1.142 | NE | |
| 58 | S | 2.117 | 0.208 | 1.20E-11 | 0.053 | 0.098 | 0.169 | PurS | |
| 59 | A | 1.174 | 0.263 | 3.05E-05 | 0.128 | 0.224 | 0.363 | PurS | |
| 60 | W | 2.507 | 0.03 | 1.19E-07 | 0.002 | 0.012 | 0.038 | PurS | |
| 61 | P | 2.258 | 0.196 | 1.11E-11 | 0.046 | 0.087 | 0.148 | PurS | |
| 62 | G | 1.383 | 0.101 | 4.94E-10 | 0.029 | 0.073 | 0.149 | PurS | |
| 63 | T | 0.644 | 0.376 | 0.130099 | 0.368 | 0.585 | 0.884 | NE | |
| 64 | L | 0.521 | 0.02 | 1.42E-06 | 0.002 | 0.038 | 0.17 | PurS | |
| 65 | R | 1.151 | 0.044 | 1.39E-11 | 0.009 | 0.038 | 0.099 | PurS | |
| 66 | S | 1.786 | 0.291 | 1.09E-08 | 0.094 | 0.163 | 0.261 | PurS | |
| 67 | G | 1.013 | 0 | 3.20E-13 | 0 | 0 | 0.031 | PurS | |
| 68 | M | 1.442 | 0.127 | 5.24E-05 | 0.044 | 0.088 | 0.155 | PurS | |
| 69 | V | 0.76 | 0.292 | 0.015354 | 0.223 | 0.384 | 0.615 | PurS | |
| 70 | P | 0.874 | 0.224 | 0.001567 | 0.126 | 0.257 | 0.463 | PurS | |
| 71 | R | 1.331 | 0.211 | 2.29E-07 | 0.083 | 0.158 | 0.269 | PurS | |
| 72 | G | 0.808 | 0.326 | 0.0296 | 0.237 | 0.403 | 0.64 | PurS | |
| 73 | P | 0.899 | 0.065 | 4.64E-08 | 0.022 | 0.073 | 0.17 | PurS | |
| 74 | T | 1.759 | 0.415 | 4.64E-06 | 0.15 | 0.236 | 0.354 | PurS | |
| 75 | A | 1.233 | 0.442 | 0.002777 | 0.225 | 0.358 | 0.544 | PurS | |
| 76 | T | 1.222 | 0.475 | 0.004137 | 0.257 | 0.389 | 0.568 | PurS | |
| 77 | A | 0.646 | 0.55 | 0.680018 | 0.566 | 0.852 | 1.236 | NE | |
| 78 | R | 1.94 | 0.127 | 5.96E-11 | 0.03 | 0.065 | 0.121 | PurS | |
| 79 | F | 0.853 | 0.128 | 0.000203 | 0.072 | 0.15 | 0.273 | PurS | |
| 80 | G | 1.144 | 0.221 | 7.29E-05 | 0.104 | 0.193 | 0.323 | PurS | |
| 81 | V | 1.526 | 0.156 | 1.19E-08 | 0.049 | 0.102 | 0.186 | PurS | |
| 82 | P | 1.006 | 0.246 | 0.000162 | 0.136 | 0.244 | 0.401 | PurS | |

| Hsp B8 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 83 | A | 3.17 | 0.297 | 4.29E-12 | 0.052 | 0.094 | 0.155 | PurS | |
| 84 | E | 1.194 | 0.166 | 1.65E-05 | 0.069 | 0.139 | 0.246 | PurS | |
| 85 | G | 0.806 | 0.443 | 0.11839 | 0.347 | 0.55 | 0.829 | NE | |
| 86 | R | 0.985 | 0.236 | 0.000394 | 0.133 | 0.239 | 0.397 | PurS | |
| 87 | T | 1.609 | 0.297 | 0.001139 | 0.112 | 0.184 | 0.285 | PurS | |
| 88 | P | 1.054 | 0.356 | 0.001957 | 0.207 | 0.337 | 0.52 | PurS | |
| 89 | P | 0.963 | 0.404 | 0.009959 | 0.26 | 0.42 | 0.643 | PurS | |
| 90 | P | 1.088 | 0.342 | 0.002434 | 0.188 | 0.315 | 0.493 | PurS | |
| 91 | F | 1.887 | 0.37 | 7.38E-05 | 0.126 | 0.196 | 0.29 | PurS | |
| 92 | P | 1.428 | 0.349 | 2.31E-05 | 0.147 | 0.244 | 0.378 | PurS | |
| 93 | G | 1.023 | 0.068 | 3.65E-08 | 0.02 | 0.066 | 0.155 | PurS | |
| 94 | E | 0.962 | 0.015 | 1.82E-07 | 0.001 | 0.015 | 0.068 | PurS | |
| 95 | P | 0.805 | 0.016 | 3.12E-09 | 0.001 | 0.02 | 0.088 | PurS | |
| 96 | W | 0 | 0 | n/a | n/a | n/a | n/a | ND | † |
| 97 | K | 0.725 | 0 | 2.27E-10 | 0 | 0 | 0.033 | PurS | |
| 98 | V | 1.013 | 0 | 2.41E-12 | 0 | 0 | 0.031 | PurS | |
| 99 | C | 1.616 | 0 | 0 | 0 | 0 | 0.015 | PurS | |
| 100 | V | 0.526 | 0 | 5.79E-08 | 0 | 0 | 0.059 | PurS | |
| 101 | N | 2.633 | 0 | 0 | 0 | 0 | 0.009 | PurS | |
| 102 | V | 0.444 | 0 | 6.97E-07 | 0 | 0 | 0.07 | PurS | |
| 103 | H | 0.533 | 0.084 | 0.006884 | 0.062 | 0.157 | 0.32 | PurS | |
| 104 | S | 0.628 | 0.035 | 7.04E-05 | 0.014 | 0.055 | 0.144 | PurS | |
| 105 | F | 0.984 | 0.012 | 3.60E-08 | 0.001 | 0.012 | 0.054 | PurS | |
| 106 | K | 1.037 | 0.136 | 5.09E-06 | 0.065 | 0.131 | 0.232 | PurS | |
| 107 | P | 1.653 | 0.053 | 5.55E-16 | 0.008 | 0.032 | 0.082 | PurS | |
| 108 | E | 0.533 | 0.015 | 8.80E-05 | 0.002 | 0.028 | 0.123 | PurS | |
| 109 | E | 1.662 | 0 | 7.56E-13 | 0 | 0 | 0.017 | PurS | |
| 110 | L | 0.49 | 0.019 | 2.99E-06 | 0.002 | 0.039 | 0.176 | PurS | |
| 111 | M | 2.236 | 0.141 | 1.95E-08 | 0.033 | 0.063 | 0.109 | PurS | |
| 112 | V | 1.29 | 0.049 | 3.84E-11 | 0.009 | 0.038 | 0.099 | PurS | |
| 113 | K | 0.698 | 0.012 | 3.16E-08 | 0.001 | 0.018 | 0.078 | PurS | |
| 114 | T | 0.426 | 0 | 3.19E-07 | 0 | 0 | 0.063 | PurS | |
| 115 | K | 0.699 | 0.052 | 8.29E-06 | 0.023 | 0.074 | 0.173 | PurS | |
| 116 | D | 1.693 | 0 | 0 | 0 | 0 | 0.016 | PurS | |
| 117 | G | 1.123 | 0.017 | 3.30E-13 | 0.001 | 0.015 | 0.066 | PurS | |

| Hsp B8 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 118 | Y | 1.901 | 0.041 | 2.67E-13 | 0.005 | 0.021 | 0.056 | PurS | |
| 119 | V | 0.412 | 0 | 3.38E-07 | 0 | 0 | 0.075 | PurS | |
| 120 | E | 0.829 | 0.014 | 8.32E-08 | 0.001 | 0.017 | 0.077 | PurS | |
| 121 | V | 0.703 | 0.016 | 2.46E-08 | 0.001 | 0.023 | 0.101 | PurS | |
| 122 | S | 0.456 | 0 | 8.63E-08 | 0 | 0 | 0.077 | PurS | |
| 123 | G | 0.819 | 0 | 1.08E-11 | 0 | 0 | 0.039 | PurS | |
| 124 | K | 0.479 | 0.074 | 0.000548 | 0.061 | 0.155 | 0.314 | PurS | |
| 125 | H | 4.476 | 0 | 0 | 0 | 0 | 0.006 | PurS | |
| 126 | E | 0.425 | 0 | 6.44E-07 | 0 | 0 | 0.063 | PurS | |
| 127 | E | 1.353 | 0 | 3.17E-12 | 0 | 0 | 0.021 | PurS | |
| 128 | K | 1.366 | 0.04 | 9.04E-10 | 0.007 | 0.029 | 0.076 | PurS | |
| 129 | Q | 1.353 | 0 | 2.79E-11 | 0 | 0 | 0.022 | PurS | |
| 130 | Q | 1.15 | 0.27 | 0.000307 | 0.139 | 0.235 | 0.37 | PurS | |
| 131 | E | 0.439 | 0.028 | 6.64E-05 | 0.011 | 0.064 | 0.199 | PurS | |
| 132 | G | 0.993 | 0 | 2.71E-14 | 0 | 0 | 0.033 | PurS | |
| 133 | G | 1.518 | 0.017 | 1.57E-14 | 0.001 | 0.011 | 0.049 | PurS | |
| 134 | I | 1.782 | 0.012 | 0 | 0 | 0.007 | 0.029 | PurS | |
| 135 | V | 0.641 | 0.016 | 2.03E-08 | 0.001 | 0.025 | 0.111 | PurS | |
| 136 | S | 0.681 | 0.031 | 7.38E-08 | 0.008 | 0.045 | 0.141 | PurS | |
| 137 | K | 0.648 | 0 | 8.60E-08 | 0 | 0 | 0.037 | PurS | |
| 138 | N | 0.743 | 0 | 5.14E-09 | 0 | 0 | 0.03 | PurS | |
| 139 | F | 0.208 | 0 | 0.00276 | 0 | 0 | 0.111 | PurS | |
| 140 | T | 0.533 | 0 | 1.29E-09 | 0 | 0 | 0.051 | PurS | |
| 141 | K | 0.326 | 0 | 0.000201 | 0 | 0 | 0.075 | PurS | |
| 142 | K | 0.535 | 0 | 2.53E-08 | 0 | 0 | 0.044 | PurS | |
| 143 | I | 0.874 | 0.012 | 1.97E-10 | 0.001 | 0.014 | 0.059 | PurS | |
| 144 | Q | 0.775 | 0.016 | 6.99E-06 | 0.001 | 0.021 | 0.094 | PurS | |
| 145 | L | 0.676 | 0.06 | 1.02E-06 | 0.027 | 0.089 | 0.208 | PurS | |
| 146 | P | 1.077 | 0 | 5.00E-15 | 0 | 0 | 0.028 | PurS | |
| 147 | A | 1.046 | 0.505 | 0.042923 | 0.313 | 0.483 | 0.715 | PurS | |
| 148 | E | 1.798 | 0.043 | 3.33E-12 | 0.006 | 0.024 | 0.063 | PurS | |
| 149 | V | 0.258 | 0 | 8.77E-05 | 0 | 0 | 0.12 | PurS | |
| 150 | D | 1.052 | 0.014 | 2.93E-11 | 0.001 | 0.013 | 0.058 | PurS | |
| 151 | P | 0.769 | 0.032 | 3.42E-09 | 0.007 | 0.042 | 0.131 | PurS | |
| 152 | V | 0.519 | 0.398 | 0.516926 | 0.477 | 0.767 | 1.17 | NE | |

| Hsp B8 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 153 | T | 1.448 | 0.042 | 1.55E-15 | 0.007 | 0.029 | 0.076 | PurS | |
| 154 | V | 1.963 | 0 | 0 | 0 | 0 | 0.016 | PurS | |
| 155 | F | 1.353 | 0 | 8.88E-16 | 0 | 0 | 0.017 | PurS | |
| 156 | A | 0.626 | 0 | 3.03E-09 | 0 | 0 | 0.052 | PurS | |
| 157 | S | 1.353 | 0 | 0 | 0 | 0 | 0.023 | PurS | |
| 158 | L | 0.99 | 0.018 | 9.43E-12 | 0.001 | 0.018 | 0.079 | PurS | |
| 159 | S | 1.353 | 0 | 2.89E-15 | 0 | 0 | 0.023 | PurS | |
| 160 | P | 0.818 | 0 | 1.19E-12 | 0 | 0 | 0.037 | PurS | |
| 161 | E | 1.599 | 0 | 5.48E-14 | 0 | 0 | 0.017 | PurS | |
| 162 | G | 1.353 | 0 | 0 | 0 | 0 | 0.023 | PurS | |
| 163 | L | 0.426 | 0.018 | 7.58E-06 | 0.002 | 0.042 | 0.185 | PurS | |
| 164 | L | 0.696 | 0 | 5.03E-10 | 0 | 0 | 0.049 | PurS | |
| 165 | I | 0.505 | 0 | 1.99E-07 | 0 | 0 | 0.046 | PurS | |
| 166 | I | 0.852 | 0 | 4.75E-11 | 0 | 0 | 0.027 | PurS | |
| 167 | E | 0.707 | 0 | 1.59E-09 | 0 | 0 | 0.038 | PurS | |
| 168 | A | 1.191 | 0.034 | 5.13E-13 | 0.005 | 0.029 | 0.089 | PurS | |
| 169 | P | 1.03 | 0.017 | 5.84E-12 | 0.001 | 0.016 | 0.071 | PurS | |
| 170 | Q | 1.662 | 0.067 | 1.00E-07 | 0.013 | 0.041 | 0.095 | PurS | |
| 171 | V | 2.398 | 0.17 | 3.69E-14 | 0.037 | 0.071 | 0.121 | PurS | |
| 172 | P | 0.907 | 0 | 3.76E-13 | 0 | 0 | 0.034 | PurS | |
| 173 | P | 1.353 | 0 | 4.44E-16 | 0 | 0 | 0.023 | PurS | |
| 174 | Y | 1.631 | 0.031 | 4.73E-11 | 0.003 | 0.019 | 0.058 | PurS | |
| 175 | S | 0.756 | 0.245 | 0.007861 | 0.181 | 0.324 | 0.531 | PurS | |
| 176 | T | 0.937 | 0.266 | 0.000614 | 0.161 | 0.284 | 0.458 | PurS | |
| 177 | F | 0.913 | 0.187 | 0.000225 | 0.112 | 0.204 | 0.339 | PurS | |
| 178 | G | 0.412 | 0.187 | 0.084631 | 0.237 | 0.455 | 0.78 | PurS | |
| 179 | E | 1.778 | 0.357 | 0.000613 | 0.127 | 0.201 | 0.301 | PurS | |
| 180 | S | 0.608 | 0.383 | 0.400169 | 0.41 | 0.63 | 0.924 | NE | |
| 181 | S | 1.041 | 0.586 | 0.137147 | 0.393 | 0.562 | 0.785 | NE | |
| 182 | F | 1.565 | 0.368 | 0.00044 | 0.148 | 0.235 | 0.352 | PurS | |
| 183 | N | 1.019 | 0.415 | 0.026661 | 0.263 | 0.407 | 0.6 | PurS | |
| 184 | N | 7.278 | 0.417 | 3.22E-15 | 0.038 | 0.057 | 0.083 | PurS | |
| 185 | E | 0.71 | 0.154 | 0.006381 | 0.108 | 0.217 | 0.384 | PurS | |
| 186 | L | 0.645 | 0.492 | 0.484256 | 0.505 | 0.763 | 1.106 | NE | |
| 187 | P | 1.452 | 0.353 | 0.000217 | 0.146 | 0.243 | 0.379 | PurS | |

| Hsp B8 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **site** | **amino acid** | **alpha** | **beta** | **p-value** | **dN/dS LB** | **dN/dS MLE** | **dN/dS UB** | **selection** | |
| 188 | Q | 1.889 | 0.353 | 1.50E-05 | 0.115 | 0.187 | 0.287 | PurS | |
| 189 | D | 0.821 | 0.214 | 0.015173 | 0.146 | 0.26 | 0.424 | PurS | |
| 190 | S | 0.369 | 0.414 | 0.805903 | 0.739 | 1.121 | 1.632 | NE | |
| 191 | Q | 1.921 | 0.2 | 1.63E-07 | 0.055 | 0.104 | 0.177 | PurS | |
| 192 | E | 0.511 | 0.043 | 0.000113 | 0.021 | 0.084 | 0.219 | PurS | |
| 193 | V | 0.747 | 0.185 | 0.002043 | 0.123 | 0.247 | 0.44 | PurS | |
| 194 | T | 0.869 | 0.231 | 0.000493 | 0.151 | 0.266 | 0.429 | PurS | |
| 195 | C | 0.945 | 0.149 | 1.16E-05 | 0.078 | 0.157 | 0.278 | PurS | |
| 196 | T | 0.485 | 0.205 | 0.05592 | 0.224 | 0.423 | 0.719 | PurS | |

B.S4 Table. Aggregate omega values as returned by the FitMG94 algorithm of full-length sequences and of selected sequence partitions of human HspB1, HspB3, HspB5, and HspB8.

| sHSP | full-length sequence/ sequence partition | region | codons excluded from calculation | $\bar{\omega}$ (aggregate omega) | confidence interval (lower bound) | confidence interval (upper bound) |
|---|---|---|---|---|---|---|
| HspB1 | M1 - K205 | full-length | M1, F8, W16 | $\bar{\omega}_{FL}$ = 0.127 | 0.122 | 0.133 |
| | M1 - P57 | N-terminal region | M1, F8, W16 | $\bar{\omega}_{NTR}$ = 0.104 | 0.095 | 0.113 |
| | L58 - S78 | central region | | $\bar{\omega}_{CeR}$ = 0.232 | 0.212 | 0.254 |
| | R79 - P168 | α-crystallin domain | | $\bar{\omega}_{\alpha CD}$ = 0.069 | 0.063 | 0.075 |
| | M169 - K205 | C-terminal extension | | $\bar{\omega}_{CTE}$ = 0.190 | 0.173 | 0.207 |
| | M1 - G122 | exon 1 | M1, F8, W16 | $\bar{\omega}_{Ex1}$ = 0.118 | 0.112 | 0.125 |
| | K123 - Y143 | exon 2 | | $\bar{\omega}_{Ex2}$ = 0.032 | 0.024 | 0.041 |
| | T144 - K205 | exon 3 | | $\bar{\omega}_{Ex3}$ = 0.146 | 0.135 | 0.158 |
| HspB3 | M1 - K150 | full-length | M1, W93, M104 | $\bar{\omega}_{FL}$ = 0.231 | 0.220 | 0.242 |
| | M1 - P40 | N-terminal region | M1 | $\bar{\omega}_{NTR}$ = 0.150 | 0.134 | 0.168 |
| | T41 - H70 | central region | | $\bar{\omega}_{CeR}$ = 0.610 | 0.567 | 0.656 |
| | F71 - L139 | α-crystallin domain | W93, M104 | $\bar{\omega}_{\alpha CD}$ = 0.086 | 0.076 | 0.097 |
| | V140 - K150 | C-terminal extension | | $\bar{\omega}_{CTE}$ = 0.416 | 0.350 | 0.491 |
| HspB5 | M1 - K175 | full-length | M1, W60, Q108 | $\bar{\omega}_{FL}$ = 0.120 | 0.113 | 0.127 |
| | M1 - P51 | N-terminal region | M1 | $\bar{\omega}_{NTR}$ = 0.106 | 0.096 | 0.117 |
| | P52 - T63 | central region | W60 | $\bar{\omega}_{CeR}$ = 0.223 | 0.183 | 0.269 |
| | G64 - R149 | α-crystallin domain | Q108 | $\bar{\omega}_{\alpha CD}$ = 0.084 | 0.076 | 0.093 |
| | K150 - K175 | C-terminal extension | | $\bar{\omega}_{CTE}$ = 0.210 | 0.183 | 0.238 |
| | M1 - E67 | exon 1 | M1, W60 | $\bar{\omega}_{Ex1}$ = 0.126 | 0.116 | 0.137 |
| | M68 - R107 | exon 2 | | $\bar{\omega}_{Ex2}$ = 0.086 | 0.074 | 0.099 |
| | Q108 - K175 | exon 3 | Q108 | $\bar{\omega}_{Ex3}$ = 0.134 | 0.122 | 0.147 |
| HspB8 | M1 - T196 | full-length | M1, W48, W51, W96 | $\bar{\omega}_{FL}$ = 0.109 | 0.103 | 0.115 |
| | M1 - M68 | N-terminal region | M1, W48, W51 | $\bar{\omega}_{NTR}$ = 0.092 | 0.083 | 0.101 |
| | V69 - G93 | central region | | $\bar{\omega}_{CeR}$ = 0.194 | 0.174 | 0.215 |
| | E94 - Y174 | α-crystallin domain | W96 | $\bar{\omega}_{\alpha CD}$ = 0.037 | 0.032 | 0.043 |
| | S175 - T196 | C-terminal extension | | $\bar{\omega}_{CTE}$ = 0.272 | 0.244 | 0.303 |
| | M1 - G123 | exon 1 | M1, W48, W51, W96 | $\bar{\omega}_{Ex1}$ = 0.106 | 0.099 | 0.113 |
| | K124 - I143 | exon 2 | | $\bar{\omega}_{Ex2}$ = 0.035 | 0.024 | 0.050 |
| | Q144 - T196 | exon 3 | | $\bar{\omega}_{Ex3}$ = 0.145 | 0.132 | 0.158 |

B.S5 Table. Examples for reported molecular and cellular implications of sHSP missense mutations with potential toxic gain-of-function mechanism. All listed mutations are associated with dominant or semi-dominant neuromuscular disease phenotypes. The molecular and cellular consequences fall into four categories, all known dominant gain-of-function mechanisms (Wilkie et al. 1994a): (1) dominant-negative effects through altered protein-protein interactions (impaired quaternary structure), (2) accumulation of toxic products (e.g., formation of protein aggregates and their highly toxic amyloid precursors), (3) impaired cell architecture (e.g., through modified cytoskeletal components), and (4) increased enzymatic (downstream) activities.

| **sHSP** | **Mutation** | **Observed Molecular and Cellular Gain-of-Function Mechanisms** | **References** |
|---|---|---|---|
| HspB1 | p.P7S | hyperphosphorylation of neurofilaments | Echaniz-Laguna et al. 2017 |
| | p.R127W | increased chaperone activity; increased fraction of the protein residing in the monomeric state; enhanced binding to tubulin | Almeida-Souza et al. 2010; 2011 |
| | | abnormal neurofilament staining pattern and aggregation | Zhang et al. 2011 |
| | | increased propensity for large oligomeric complex formation at high concentrations; increased propensity to dissociate at low concentrations | Weeks et al. 2018 |
| | | impaired formation of SQSTM1/p62 bodies in autophagy | Haidar et al. 2019 |
| | p.Q128R | hyperphosphorylation of neurofilaments | Echaniz-Laguna et al. 2017 |
| | p.S135F | impaired neurofilament assembly; formation of protein aggregates | Evgrafov et al. 2004 |

| | | | |
|---|---|---|---|
| | | increased interaction with HspB8 | Fontaine et al. 2006 |
| | | disruption of neurofilament network and aggregation of neurofilament light chain, resulting in progressive degeneration of motor neurons | Zhai et al. 2007 |
| | | increased chaperone activity; increased fraction of the protein residing in the monomeric state; enhanced binding to tubulin; abnormal stabilization of microtubules | Almeida-Souza et al. 2010; 2011 |
| | | decreased level of acetylated □-tubulin possibly resulting from increased activity of histone deacetylase in sciatic nerve; impaired axonal transport | d'Ydewalle et al. 2011 |
| | | hyperphosphorylation of cyclin-dependent protein kinase Cdk5 resulting in hyperphosphorylation of neurofilaments, affecting their binding to the motor protein kinesin and thereby the anterograde transport | Holmgren et al. 2013 |
| | | increased propensity for large oligomeric complex formation at high concentrations; increased propensity to dissociate at low concentrations | Weeks et al. 2018 |
| | | impaired formation of SQSTM1/p62 bodies in autophagy | Haidar et al. 2019 |
| | p.R136W | increased chaperone activity; increased fraction of the protein residing in the monomeric state; enhanced binding to tubulin | Almeida-Souza et al. 2010; 2011 |
| | | increased propensity for large oligomeric complex formation at high concentrations, increased propensity to dissociate at low concentrations | Weeks et al. 2018 |
| | p.P182L | formation of protein aggregates (sequestration of neurofilament middle chain subunit and p150 dynactin); disruption of neurofilament assembly and axonal transport | Ackerley et al. 2006 |
| | | decreased level of acetylated □-tubulin possibly resulting from increased activity of histone | d'Ydewalle et al. 2011 |

| | | | |
|---|---|---|---|
| | | deacetylase in sciatic nerve; impaired axonal transport | |
| | | hyperphosphorylation of cyclin-dependent protein kinase Cdk5 resulting in hyperphosphorylation of neurofilaments, affecting their binding to the motor protein kinesin and thereby the anterograde transport | Holmgren et al. 2013 |
| | | increased interaction with the RNA binding protein PCBP1, resulting in reduced translational repressive activity | Geuens et al. 2017 |
| | | impaired formation of SQSTM1/p62 bodies in autophagy | Haidar et al. 2019 |
| | p.S187L | formation of protein aggregates | Echaniz-Laguna et al. 2017 |
| HspB8 | p.K141E | formation of protein aggregates | Irobi et al. 2004 |
| | | altered (largely increased) sHSP interactions; increased formation of protein aggregates | Fontaine et al. 2006 |
| | | increased interaction with the RNA-helicase Ddx20 (gemin 3) | Sun et al. 2010 |
| | | reduced interaction with BAG3 | Shemetov and Gusev 2011 |
| | | impaired clearing of protein aggregates | Carra et al. 2010 |
| | p.K141N | formation of protein aggregates | Irobi et al. 2004 |
| | | altered (largely increased) sHSP interactions; increased formation of protein aggregates | Fontaine et al. 2006 |
| | | increased interaction with the RNA-helicase Ddx20 (gemin 3) | Sun et al. 2010 |
| | | impaired clearing of protein aggregates | Carra et al. 2010 |
| | | formation of protein aggregates; reduced mitochondrial membrane potential | Irobi et al. 2012 |
| | | impact on the muscular cell architecture: Z-disk disorganization; granulofilamentous material accumulating along with mutant HspB8; desmin aggregation; impaired autophagy | Bouhy et al. 2018 |
| | p.K141M | increased binding to BAG3 | Echaniz-Laguna et al. 2017 |

| | | | |
|---|---|---|---|
| HspB5 | p.R120G | mutant HspB5-desmin aggregate formation | Vicart et al. 1998 |
| | | irregular quaternary structure of HspB5; defective chaperone function | Bova et al. 1999 |
| | | translocation of FBX4, a component of the ubiquitin-protein ligase system to the detergent-insoluble fraction and stimulation of ubiquitination | den Engelsman et al. 2003 |
| | | alterations in mitochondrial organization and architecture; impaired cytoskeletal network | Maloyan et al. 2005 |
| | | oxidative stress due to augmented expression or activity of glucose-6phosphate dehydrogenase, glutathione reductase, and glutathione peroxidase; protein aggregation | Rajasekaran et al. 2007 |
| | | abnormal sHSP interactions; increased tendency for aggregate formation; increased degree of mutant HspB5 phosphorylation | Simon et al. 2007 |
| | | perinuclear formation of protein aggregates | Sanbe 2011 |
| | | promotes desmin filament aggregation | Elliott et al. 2013 |

## APPENDIX C: Supplemental Material for Chapter 4

C.S1 Table. The FEL analysis of the BDNF gene found 174 of 261 (66.7%) sites to be statistically significant (LRT p-value <= 0.1) for pervasive negative (purifying) selection. Alpha (Synonymous substitution rate at a site). beta (Non-synonymous substitution rate at a site). alpha=beta (The rate estimate under the neutral model). LRT (Likelihood ration test statistic for beta = alpha, versus beta &neq; alpha). p-value (Likelihood ration test statistic for beta = alpha, versus beta &neq; alpha). Total branch length (The total length of branches contributing to inference at this site, and used to scale dN-dS)**.**

| Site | alpha | beta | alpha= beta | LRT | p-value | Total branch length | dN/dS LB | dN/dS MLE | dN/dS UB |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 4 | 0.1116 | 0.0635 | 0.0811 | 0.1573 | 0.6916 | 0.6594 | 0.0325 | 0.569 | 2.5205 |
| 5 | 0 | 0.0564 | 0.0468 | 0.3606 | 0.5482 | 0.3806 | 1462.797 | 10000 | 10000 |
| 6 | 0.2221 | 0 | 0.0805 | 4.043 | 0.0444 | 0.6545 | 0 | 0 | 0.5478 |
| 7 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 9 | 0.0928 | 0 | 0.0397 | 2.0198 | 0.1553 | 0.3227 | 0 | 0 | 1.2848 |
| 10 | 0.1594 | 0 | 0.0409 | 2.7359 | 0.0981 | 0.3326 | 0 | 0 | 0.658 |
| 11 | 0.1729 | 0 | 0.09 | 3.1185 | 0.0774 | 0.7317 | 0 | 0 | 0.9176 |
| 12 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 13 | 1.963 | 0 | 0.285 | 22.6368 | 0 | 2.3174 | 0 | 0 | 0.0549 |
| 14 | 0 | 0.2172 | 0.1438 | 3.0861 | 0.079 | 1.1694 | 6185.591 | 10000 | 10000 |
| 15 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

| Site | alpha | beta | alpha= beta | LRT | p-value | Total branch length | dN/dS LB | dN/dS MLE | dN/dS UB |
|---|---|---|---|---|---|---|---|---|---|
| 17 | 0.2673 | 0.053 | 0.0885 | 1.169 | 0.2796 | 0.7199 | 0.0113 | 0.1982 | 0.8804 |
| 18 | 0.3413 | 0 | 0.1102 | 6.7673 | 0.0093 | 0.8958 | 0 | 0 | 0.3054 |
| 19 | 0.4344 | 0 | 0.1182 | 7.7643 | 0.0053 | 0.9612 | 0 | 0 | 0.2399 |
| 20 | 0.7437 | 0 | 0.1993 | 12.7922 | 0.0003 | 1.6204 | 0 | 0 | 0.1425 |
| 21 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 22 | 0.7734 | 0 | 0.1935 | 13.6496 | 0.0002 | 1.5734 | 0 | 0 | 0.1283 |
| 23 | 0.1435 | 0.0527 | 0.0773 | 0.475 | 0.4907 | 0.6284 | 0.021 | 0.367 | 1.6325 |
| 24 | 1.7963 | 0.4672 | 0.8629 | 8.8413 | 0.0029 | 7.017 | 0.1183 | 0.2601 | 0.4893 |
| 25 | 1.4891 | 0.3979 | 0.5537 | 4.2623 | 0.039 | 4.5023 | 0.1213 | 0.2672 | 0.5042 |
| 26 | 0.1396 | 1.0896 | 0.7817 | 6.8872 | 0.0087 | 6.356 | 4.5038 | 7.8049 | 12.599 |
| 27 | 0.2849 | 0.2004 | 0.2436 | 0.1449 | 0.7035 | 1.9806 | 0.1198 | 0.7033 | 2.1145 |
| 28 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 29 | 0.4139 | 0.0566 | 0.1834 | 4.1399 | 0.0419 | 1.4914 | 0.0078 | 0.1367 | 0.6064 |
| 30 | 0 | 0.3797 | 0.2579 | 4.6268 | 0.0315 | 2.0973 | 7260.052 | 10000 | 10000 |
| 31 | 0.2774 | 0 | 0.0763 | 5.1487 | 0.0233 | 0.6206 | 0 | 0 | 0.3653 |
| 32 | 0.2572 | 0.4242 | 0.3976 | 0.2465 | 0.6196 | 3.2327 | 0.7845 | 1.6493 | 3.0082 |
| 33 | 0.303 | 0.2456 | 0.2654 | 0.0484 | 0.8259 | 2.1581 | 0.2009 | 0.8105 | 2.1234 |
| 34 | 1.0536 | 0.1089 | 0.3602 | 9.954 | 0.0016 | 2.9291 | 0.0172 | 0.1034 | 0.3222 |
| 35 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 36 | 0.2906 | 0.0686 | 0.1107 | 0.9484 | 0.3301 | 0.9 | 0.0134 | 0.236 | 1.0454 |
| 37 | 0.8023 | 0 | 0.2613 | 15.6626 | 0.0001 | 2.1248 | 0 | 0 | 0.1321 |
| 38 | 0.7113 | 0.3303 | 0.4528 | 1.6696 | 0.1963 | 3.6817 | 0.1837 | 0.4644 | 0.9547 |
| 39 | 0.4421 | 0.525 | 0.503 | 0.0485 | 0.8256 | 4.0903 | 0.542 | 1.1877 | 2.2205 |
| 40 | 0.4097 | 0 | 0.1578 | 7.5858 | 0.0059 | 1.2834 | 0 | 0 | 0.3013 |
| 41 | 0 | 0.0551 | 0.0393 | 0.6774 | 0.4105 | 0.3195 | 1462.759 | 10000 | 10000 |
| 42 | 1.2391 | 0.0506 | 0.2862 | 13.8667 | 0.0002 | 2.3274 | 0.0023 | 0.0409 | 0.1816 |
| 43 | 0.1559 | 0 | 0.0396 | 2.7457 | 0.0975 | 0.3218 | 0 | 0 | 0.6535 |
| 44 | 0.3404 | 0.0556 | 0.1494 | 2.9408 | 0.0864 | 1.2148 | 0.0093 | 0.1634 | 0.7237 |
| 45 | 0.0928 | 0 | 0.0443 | 1.4933 | 0.2217 | 0.3605 | 0 | 0 | 1.749 |
| 46 | 0 | 0.164 | 0.137 | 1.0523 | 0.305 | 1.114 | 5270.097 | 10000 | 10000 |
| 47 | 1.8271 | 0.1373 | 0.3286 | 11.8421 | 0.0006 | 2.6721 | 0.0187 | 0.0751 | 0.1965 |
| 48 | 0.7118 | 1.0257 | 0.9372 | 0.4209 | 0.5165 | 7.6211 | 0.8165 | 1.4411 | 2.3527 |

| Site | alpha | beta | alpha= beta | LRT | p-value | Total branch length | dN/dS LB | dN/dS MLE | dN/dS UB |
|---|---|---|---|---|---|---|---|---|---|
| 49 | 1.3089 | 0.3744 | 0.585 | 4.9417 | 0.0262 | 4.757 | 0.1222 | 0.286 | 0.5605 |
| 50 | 0.2882 | 0.0532 | 0.1161 | 2.0383 | 0.1534 | 0.944 | 0.0105 | 0.1844 | 0.8138 |
| 51 | 0.1419 | 0 | 0.0395 | 2.5243 | 0.1121 | 0.3214 | 0 | 0 | 0.747 |
| 52 | 0.2736 | 0.3297 | 0.3204 | 0.0312 | 0.8599 | 2.6057 | 0.4762 | 1.205 | 2.4754 |
| 53 | 1.5426 | 0.4637 | 0.8202 | 7.0701 | 0.0078 | 6.6692 | 0.1365 | 0.3006 | 0.5668 |
| 54 | 1.5617 | 0.2216 | 0.6252 | 13.3081 | 0.0003 | 5.0838 | 0.044 | 0.1419 | 0.3335 |
| 55 | 3.6302 | 0.5463 | 1.624 | 22.5001 | 0 | 13.2053 | 0.0643 | 0.1505 | 0.2957 |
| 56 | 1.5021 | 0.0545 | 0.4719 | 20.6584 | 0 | 3.8374 | 0.0021 | 0.0363 | 0.1612 |
| 57 | 0.5526 | 0.2782 | 0.3576 | 0.981 | 0.322 | 2.908 | 0.1797 | 0.5034 | 1.095 |
| 58 | 0.3979 | 0.0869 | 0.2313 | 2.3382 | 0.1262 | 1.8808 | 0.0124 | 0.2184 | 0.9655 |
| 59 | 11.119 | 0.8319 | 1.4313 | 1.5693 | 0.2103 | 11.6383 | 0.0038 | 0.0748 | 0.6522 |
| 60 | 4.372 | 0.3686 | 1.3037 | 33.3567 | 0 | 10.6014 | 0.0332 | 0.0843 | 0.1727 |
| 61 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 62 | 1.7721 | 2193.285 | 2856.939 | 0 | 1 | 23231.30 76 | 0.3292 | 1237.702 | 10000 |
| 63 | 0 | 3.644 | 2.0642 | 0.9793 | 0.3224 | 16.7849 | 1464.979 | 10000 | 10000 |
| 64 | 7.4868 | 0 | 1.0937 | 6.3427 | 0.0118 | 8.8931 | 0 | 0 | 0.1725 |
| 65 | 122.5088 | 0 | 1.174 | 7.2662 | 0.007 | 9.5464 | 0 | 0 | 0.0108 |
| 66 | 2.8095 | 0 | 0.7275 | 4.7774 | 0.0288 | 5.9161 | 0 | 0 | 0.3619 |
| 67 | 10.8121 | 0 | 1.6286 | 13.5042 | 0.0002 | 13.2432 | 0 | 0 | 0.0777 |
| 68 | 4.1586 | 0.4872 | 1.5564 | 8.0811 | 0.0045 | 12.6556 | 0.0193 | 0.1172 | 0.3731 |
| 69 | 3.2764 | 0.3962 | 1.3489 | 22.0059 | 0 | 10.9685 | 0.0431 | 0.1209 | 0.263 |
| 70 | 0.9289 | 0 | 0.3524 | 13.2397 | 0.0003 | 2.8657 | 0 | 0 | 0.1693 |
| 71 | 1.1053 | 0 | 0.3383 | 20.8817 | 0 | 2.7513 | 0 | 0 | 0.0943 |
| 72 | 0.2921 | 0.0985 | 0.1261 | 0.663 | 0.4155 | 1.0251 | 0.0558 | 0.3373 | 1.0461 |
| 73 | 1.1053 | 0 | 0.3395 | 20.5444 | 0 | 2.7603 | 0 | 0 | 0.0954 |
| 74 | 2.9165 | 0.0575 | 0.6252 | 32.3768 | 0 | 5.0841 | 0.0011 | 0.0197 | 0.0877 |
| 75 | 0.8228 | 0 | 0.2007 | 13.9151 | 0.0002 | 1.632 | 0 | 0 | 0.1241 |

| Site | alpha | beta | alpha= beta | LRT | p-value | Total branch length | dN/dS LB | dN/dS MLE | dN/dS UB |
|---|---|---|---|---|---|---|---|---|---|
| 76 | 5.5185 | 0 | 0.57 | 53.7705 | 0 | 4.6353 | 0 | 0 | 0.0174 |
| 77 | 0.628 | 0.0619 | 0.2215 | 5.6574 | 0.0174 | 1.801 | 0.0056 | 0.0985 | 0.4366 |
| 78 | 2.1641 | 0.0489 | 0.3787 | 23.0125 | 0 | 3.0792 | 0.0013 | 0.0226 | 0.1005 |
| 79 | 1.8388 | 0 | 0.4162 | 29.0685 | 0 | 3.384 | 0 | 0 | 0.0556 |
| 80 | 0.2839 | 0.0541 | 0.0909 | 1.2074 | 0.2718 | 0.7393 | 0.0109 | 0.1905 | 0.8475 |
| 81 | 1.3529 | 0 | 0.5463 | 20.2275 | 0 | 4.442 | 0 | 0 | 0.1126 |
| 82 | 1.8857 | 0.083 | 0.7606 | 19.4503 | 0 | 6.1851 | 0.0025 | 0.044 | 0.195 |
| 83 | 2.4273 | 0.1547 | 0.5523 | 20.675 | 0 | 4.491 | 0.0157 | 0.0637 | 0.1663 |
| 84 | 0.5934 | 0 | 0.0911 | 7.3673 | 0.0066 | 0.7411 | 0 | 0 | 0.1756 |
| 85 | 0.3406 | 0.1489 | 0.1729 | 0.4375 | 0.5083 | 1.4059 | 0.1081 | 0.4371 | 1.1399 |
| 86 | 0 | 0.3235 | 0.2567 | 2.3672 | 0.1239 | 2.0875 | 6809.739 | 10000 | 10000 |
| 87 | 1.2767 | 0.1562 | 0.4047 | 9.9233 | 0.0016 | 3.2907 | 0.0303 | 0.1223 | 0.3201 |
| 88 | 1.341 | 0.5707 | 0.8242 | 3.2658 | 0.0707 | 6.7019 | 0.2024 | 0.4256 | 0.781 |
| 89 | 1.2511 | 0.1289 | 0.5092 | 12.0135 | 0.0005 | 4.1408 | 0.017 | 0.1031 | 0.3194 |
| 90 | 0.6067 | 0.0552 | 0.202 | 6.0849 | 0.0136 | 1.6429 | 0.0052 | 0.091 | 0.403 |
| 91 | 4.6374 | 1.7289 | 2.1855 | 6.6899 | 0.0097 | 17.7717 | 0.2477 | 0.3728 | 0.5401 |
| 92 | 1.8962 | 0.0535 | 0.4607 | 23.2267 | 0 | 3.7463 | 0.0016 | 0.0282 | 0.1245 |
| 93 | 3.5662 | 0.0542 | 0.7407 | 40.9892 | 0 | 6.0226 | 0.0009 | 0.0152 | 0.0676 |
| 94 | 1.3252 | 0.3547 | 0.4855 | 3.6636 | 0.0556 | 3.9476 | 0.1139 | 0.2676 | 0.5226 |
| 95 | 1.3323 | 1.5991 | 1.5317 | 0.1473 | 0.7011 | 12.4555 | 0.7548 | 1.2003 | 1.8159 |
| 96 | 1.3529 | 0 | 0.1787 | 14.6186 | 0.0001 | 1.4535 | 0 | 0 | 0.0749 |
| 97 | 1.8434 | 0.0489 | 0.2953 | 17.5494 | 0 | 2.4011 | 0.0015 | 0.0265 | 0.1178 |
| 98 | 0.651 | 0.0536 | 0.2046 | 6.5994 | 0.0102 | 1.6636 | 0.0047 | 0.0824 | 0.3692 |
| 99 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 100 | 0.5931 | 0.3419 | 0.4207 | 0.4996 | 0.4797 | 3.4205 | 0.1807 | 0.5764 | 1.3251 |
| 101 | 5.6241 | 0.0678 | 0.7121 | 36.3681 | 0 | 5.7902 | 0.0007 | 0.012 | 0.0536 |
| 102 | 2.4986 | 0.1708 | 0.7664 | 25.4012 | 0 | 6.2322 | 0.0169 | 0.0684 | 0.1784 |

| Site | alpha | beta | alpha= beta | LRT | p-value | Total branch length | dN/dS LB | dN/dS MLE | dN/dS UB |
|---|---|---|---|---|---|---|---|---|---|
| 103 | 0.696 | 0 | 0.2116 | 11.7725 | 0.0006 | 1.7204 | 0 | 0 | 0.1687 |
| 104 | 2.3048 | 0 | 0.5043 | 34.9881 | 0 | 4.101 | 0 | 0 | 0.0444 |
| 105 | 0.6641 | 0 | 0.1786 | 10.1702 | 0.0014 | 1.452 | 0 | 0 | 0.1795 |
| 106 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 107 | 0.9618 | 0 | 0.3094 | 15.6026 | 0.0001 | 2.5158 | 0 | 0 | 0.1311 |
| 108 | 1.7507 | 0 | 0.2683 | 25.8194 | 0 | 2.1813 | 0 | 0 | 0.0495 |
| 109 | 1.2561 | 0.0472 | 0.2636 | 13.7621 | 0.0002 | 2.1434 | 0.0021 | 0.0376 | 0.1661 |
| 110 | 0.3333 | 0 | 0.09 | 5.1186 | 0.0237 | 0.7317 | 0 | 0 | 0.3586 |
| 111 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 112 | 0.2221 | 0 | 0.0736 | 4.3993 | 0.036 | 0.5984 | 0 | 0 | 0.4774 |
| 113 | 0.27 | 0 | 0.1012 | 3.9075 | 0.0481 | 0.8232 | 0 | 0 | 0.5785 |
| 114 | 0.6015 | 0 | 0.0925 | 7.5591 | 0.006 | 0.7518 | 0 | 0 | 0.1729 |
| 115 | 2.1843 | 0 | 0.595 | 37.4897 | 0 | 4.838 | 0 | 0 | 0.0485 |
| 116 | 1.7115 | 0 | 0.4751 | 29.9207 | 0 | 3.8634 | 0 | 0 | 0.0619 |
| 117 | 1.3529 | 0 | 0.4841 | 22.7688 | 0 | 3.9361 | 0 | 0 | 0.0965 |
| 118 | 0.2799 | 0 | 0.0871 | 4.6517 | 0.031 | 0.7081 | 0 | 0 | 0.4354 |
| 119 | 2.6166 | 0 | 0.5451 | 36.1605 | 0 | 4.4325 | 0 | 0 | 0.0423 |
| 120 | 0.0928 | 0 | 0.0445 | 1.5555 | 0.2123 | 0.3619 | 0 | 0 | 1.7079 |
| 121 | 3.0538 | 0 | 0.9175 | 44.7049 | 0 | 7.4605 | 0 | 0 | 0.0424 |
| 122 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 123 | 0.5018 | 0 | 0.121 | 8.4758 | 0.0036 | 0.984 | 0 | 0 | 0.2034 |
| 124 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 125 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 126 | 0.1674 | 0.0519 | 0.0785 | 0.6645 | 0.415 | 0.6381 | 0.0174 | 0.3101 | 1.3547 |
| 127 | 0.591 | 0.068 | 0.1651 | 3.2197 | 0.0728 | 1.3427 | 0.0066 | 0.1151 | 0.5099 |
| 128 | 0.4829 | 0 | 0.224 | 7.5898 | 0.0059 | 1.8217 | 0 | 0 | 0.3344 |
| 129 | 0.7281 | 0 | 0.1574 | 12.1458 | 0.0005 | 1.2801 | 0 | 0 | 0.1324 |
| 130 | 1.3529 | 0 | 0.4223 | 26.429 | 0 | 3.4343 | 0 | 0 | 0.077 |
| 131 | 1.3529 | 0 | 0.4072 | 24.6703 | 0 | 3.3113 | 0 | 0 | 0.077 |

| Site | alpha | beta | alpha= beta | LRT | p-value | Total branch length | dN/dS LB | dN/dS MLE | dN/dS UB |
|---|---|---|---|---|---|---|---|---|---|
| 132 | 0.5994 | 0 | 0.0843 | 7.7423 | 0.0054 | 0.6853 | 0 | 0 | 0.1579 |
| 133 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 134 | 2.4575 | 0 | 0.6712 | 37.7462 | 0 | 5.4579 | 0 | 0 | 0.0483 |
| 135 | 0 | 0.0435 | 0.0438 | 0.0008 | 0.9774 | 0.3558 | 1462.167 | 10000 | 10000 |
| 136 | 1.0696 | 0 | 0.2395 | 17.487 | 0 | 1.9473 | 0 | 0 | 0.0925 |
| 137 | 0.4335 | 0 | 0.1301 | 7.1875 | 0.0073 | 1.0578 | 0 | 0 | 0.275 |
| 138 | 0.6509 | 0 | 0.2392 | 11.8685 | 0.0006 | 1.9451 | 0 | 0 | 0.1867 |
| 139 | 0.4217 | 0 | 0.1161 | 7.6814 | 0.0056 | 0.9442 | 0 | 0 | 0.2443 |
| 140 | 0.9836 | 0 | 0.1277 | 11.8927 | 0.0006 | 1.0384 | 0 | 0 | 0.0967 |
| 141 | 3.5104 | 0 | 0.9082 | 50.5522 | 0 | 7.3849 | 0 | 0 | 0.0344 |
| 142 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 143 | 3.889 | 0 | 0.8487 | 57.0286 | 0 | 6.9012 | 0 | 0 | 0.0273 |
| 144 | 0.5803 | 0 | 0.1576 | 10.3294 | 0.0013 | 1.2818 | 0 | 0 | 0.1796 |
| 145 | 0.8551 | 0 | 0.2336 | 15.3582 | 0.0001 | 1.8993 | 0 | 0 | 0.1209 |
| 146 | 3.1941 | 0 | 0.8296 | 52.1403 | 0 | 6.7458 | 0 | 0 | 0.0335 |
| 147 | 0.4576 | 0 | 0.1172 | 8.0791 | 0.0045 | 0.9527 | 0 | 0 | 0.2215 |
| 148 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 149 | 0.1879 | 0 | 0.0883 | 2.9817 | 0.0842 | 0.7177 | 0 | 0 | 0.8603 |
| 150 | 2.4574 | 0 | 0.2729 | 29.1158 | 0 | 2.2192 | 0 | 0 | 0.0347 |
| 151 | 1.3529 | 0 | 0.3641 | 20.8617 | 0 | 2.961 | 0 | 0 | 0.0881 |
| 152 | 1.8431 | 0 | 0.3369 | 26.2702 | 0 | 2.7398 | 0 | 0 | 0.0543 |
| 153 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 154 | 1.1326 | 0.0449 | 0.2254 | 12.6183 | 0.0004 | 1.8326 | 0.0023 | 0.0397 | 0.1759 |
| 155 | 1.9612 | 0.1105 | 0.5215 | 19.7057 | 0 | 4.241 | 0.0093 | 0.0564 | 0.1754 |
| 156 | 9.5442 | 0 | 0.6185 | 72.668 | 0 | 5.0296 | 0 | 0 | 0.009 |
| 157 | 0.5748 | 0.0547 | 0.1375 | 3.7909 | 0.0515 | 1.1182 | 0.0054 | 0.0951 | 0.4215 |
| 158 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 159 | 2.8095 | 0 | 0.8095 | 44.3396 | 0 | 6.5824 | 0 | 0 | 0.0424 |

| Site | alpha | beta | alpha= beta | LRT | p-value | Total branch length | dN/dS LB | dN/dS MLE | dN/dS UB |
|---|---|---|---|---|---|---|---|---|---|
| 160 | 2.4575 | 0 | 0.6733 | 41.9754 | 0 | 5.4754 | 0 | 0 | 0.0429 |
| 161 | 3.1758 | 0 | 0.7134 | 48.2008 | 0 | 5.8007 | 0 | 0 | 0.0328 |
| 162 | 1.3529 | 0 | 0.3849 | 24.2341 | 0 | 3.1298 | 0 | 0 | 0.077 |
| 163 | 1.6026 | 0.0501 | 0.3633 | 19.2676 | 0 | 2.9542 | 0.0018 | 0.0313 | 0.1388 |
| 164 | 0.6094 | 0 | 0.1545 | 10.8904 | 0.001 | 1.2565 | 0 | 0 | 0.1629 |
| 165 | 0.2724 | 0 | 0.0441 | 3.6202 | 0.0571 | 0.3588 | 0 | 0 | 0.3729 |
| 166 | 1.3529 | 0 | 0.3828 | 23.3448 | 0 | 3.1125 | 0 | 0 | 0.0779 |
| 167 | 1.3492 | 0.0541 | 0.4779 | 20.1541 | 0 | 3.8862 | 0.0023 | 0.0401 | 0.1781 |
| 168 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 169 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 170 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 171 | 2.3512 | 0 | 0.6698 | 30.8621 | 0 | 5.4463 | 0 | 0 | 0.0603 |
| 172 | 1.3529 | 0 | 0.351 | 23.6433 | 0 | 2.8544 | 0 | 0 | 0.075 |
| 173 | 0.9159 | 0 | 0.2362 | 16.1752 | 0.0001 | 1.9207 | 0 | 0 | 0.111 |
| 174 | 2.5514 | 0 | 0.585 | 39.0254 | 0 | 4.7569 | 0 | 0 | 0.0413 |
| 175 | 0.4238 | 0 | 0.129 | 7.0778 | 0.0078 | 1.0492 | 0 | 0 | 0.2813 |
| 176 | 3 | 0 | 0.802 | 50.0929 | 0 | 6.5213 | 0 | 0 | 0.0351 |
| 177 | 0.2903 | 0 | 0.0867 | 4.8037 | 0.0284 | 0.705 | 0 | 0 | 0.4106 |
| 178 | 2.0769 | 0 | 0.7435 | 33.684 | 0 | 6.0457 | 0 | 0 | 0.0637 |
| 179 | 1.3529 | 0 | 0.2916 | 20.9924 | 0 | 2.3712 | 0 | 0 | 0.0756 |
| 180 | 2.8095 | 0 | 0.5579 | 41.9086 | 0 | 4.537 | 0 | 0 | 0.0356 |
| 181 | 0.4269 | 0 | 0.1294 | 7.1083 | 0.0077 | 1.0522 | 0 | 0 | 0.2792 |
| 182 | 2.0769 | 0 | 0.5715 | 36.1609 | 0 | 4.6469 | 0 | 0 | 0.051 |
| 183 | 2.4177 | 0 | 0.7387 | 39.0146 | 0 | 6.0071 | 0 | 0 | 0.0493 |
| 184 | 4.3195 | 0.0699 | 1.0443 | 43.8566 | 0 | 8.4918 | 0.0009 | 0.0162 | 0.0716 |
| 185 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 186 | 0.1373 | 0 | 0.0381 | 2.5665 | 0.1092 | 0.3095 | 0 | 0 | 0.7378 |

| Site | alpha | beta | alpha= beta | LRT | p-value | Total branch length | dN/dS LB | dN/dS MLE | dN/dS UB |
|---|---|---|---|---|---|---|---|---|---|
| 187 | 1.3529 | 0 | 0.3201 | 18.1218 | 0 | 2.6032 | 0 | 0 | 0.0888 |
| 188 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 189 | 0.272 | 0 | 0.0441 | 3.6125 | 0.0573 | 0.3584 | 0 | 0 | 0.3735 |
| 190 | 1.8571 | 0 | 0.3924 | 23.7914 | 0 | 3.1908 | 0 | 0 | 0.0653 |
| 191 | 0.9044 | 0 | 0.1654 | 9.9503 | 0.0016 | 1.345 | 0 | 0 | 0.1449 |
| 192 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 193 | 0.9156 | 0 | 0.166 | 10.0482 | 0.0015 | 1.3502 | 0 | 0 | 0.1431 |
| 194 | 0.2724 | 0 | 0.0451 | 3.5786 | 0.0585 | 0.3665 | 0 | 0 | 0.3827 |
| 195 | 0.1373 | 0 | 0.0392 | 2.51 | 0.1131 | 0.3184 | 0 | 0 | 0.7677 |
| 196 | 0.8673 | 0.0531 | 0.1785 | 6.8492 | 0.0089 | 1.4516 | 0.0035 | 0.0612 | 0.2703 |
| 197 | 0.6237 | 0 | 0.0904 | 7.774 | 0.0053 | 0.7354 | 0 | 0 | 0.1619 |
| 198 | 0.7734 | 0 | 0.1617 | 12.4902 | 0.0004 | 1.3145 | 0 | 0 | 0.126 |
| 199 | 0.9136 | 0 | 0.2416 | 15.6381 | 0.0001 | 1.9647 | 0 | 0 | 0.116 |
| 200 | 0 | 0.0439 | 0.0439 | 0.0088 | 0.9254 | 0.3566 | 1462.186 | 10000 | 10000 |
| 201 | 0.8954 | 0 | 0.2625 | 17.0957 | 0 | 2.1349 | 0 | 0 | 0.1132 |
| 202 | 0.5743 | 0.069 | 0.1664 | 3.1111 | 0.0778 | 1.3533 | 0.0068 | 0.1202 | 0.5326 |
| 203 | 3 | 0 | 0.8142 | 50.6222 | 0 | 6.6207 | 0 | 0 | 0.0351 |
| 204 | 0.4 | 0 | 0.0825 | 6.2389 | 0.0125 | 0.6712 | 0 | 0 | 0.2512 |
| 205 | 1.2072 | 0.0537 | 0.2628 | 11.61 | 0.0007 | 2.1373 | 0.0025 | 0.0445 | 0.1973 |
| 206 | 1.3529 | 0 | 0.3132 | 21.2564 | 0 | 2.547 | 0 | 0 | 0.075 |
| 207 | 0.2877 | 0 | 0.0444 | 3.7112 | 0.054 | 0.361 | 0 | 0 | 0.3519 |
| 208 | 2.6056 | 0 | 0.5748 | 39.9407 | 0 | 4.6744 | 0 | 0 | 0.0376 |
| 209 | 0.5731 | 0 | 0.1538 | 10.39 | 0.0013 | 1.2509 | 0 | 0 | 0.177 |
| 210 | 0.5502 | 0 | 0.0759 | 7.9203 | 0.0049 | 0.617 | 0 | 0 | 0.1531 |
| 211 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 212 | 1.6439 | 0 | 0.3321 | 24.7877 | 0 | 2.7006 | 0 | 0 | 0.0609 |
| 213 | 0.3655 | 0.1001 | 0.1571 | 1.5333 | 0.2156 | 1.2773 | 0.0454 | 0.274 | 0.8525 |
| 214 | 4.5596 | 0 | 0.6112 | 53.3306 | 0 | 4.9697 | 0 | 0 | 0.0213 |
| 215 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

| Site | alpha | beta | alpha= beta | LRT | p-value | Total branch length | dN/dS LB | dN/dS MLE | dN/dS UB |
|---|---|---|---|---|---|---|---|---|---|
| 216 | 1.8153 | 0 | 0.2159 | 20.6059 | 0 | 1.7556 | 0 | 0 | 0.0522 |
| 217 | 0.9243 | 0 | 0.2666 | 14.572 | 0.0001 | 2.1675 | 0 | 0 | 0.1312 |
| 218 | 2.0769 | 0 | 0.3184 | 21.3914 | 0 | 2.5891 | 0 | 0 | 0.059 |
| 219 | 0.6452 | 0 | 0.091 | 7.887 | 0.005 | 0.7397 | 0 | 0 | 0.1567 |
| 220 | 0.0928 | 0 | 0.0395 | 1.1939 | 0.2746 | 0.321 | 0 | 0 | 1.7623 |
| 221 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 222 | 0.2835 | 0 | 0.0789 | 5.0825 | 0.0242 | 0.6414 | 0 | 0 | 0.3719 |
| 223 | 2.3567 | 0 | 0.3682 | 24.3183 | 0 | 2.9942 | 0 | 0 | 0.0521 |
| 224 | 1.6066 | 0 | 0.4605 | 22.0347 | 0 | 3.7447 | 0 | 0 | 0.0858 |
| 225 | 11.2648 | 0 | 1.0274 | 68.4271 | 0 | 8.354 | 0 | 0 | 0.0116 |
| 226 | 0.7404 | 0 | 0.2181 | 12.062 | 0.0005 | 1.7737 | 0 | 0 | 0.161 |
| 227 | 0.1729 | 0 | 0.0784 | 3.7735 | 0.0521 | 0.6379 | 0 | 0 | 0.7082 |
| 228 | 1.7586 | 0 | 0.454 | 30.8744 | 0 | 3.692 | 0 | 0 | 0.0593 |
| 229 | 2.3195 | 0 | 0.6722 | 35.6279 | 0 | 5.4661 | 0 | 0 | 0.053 |
| 230 | 0.4314 | 0 | 0.1189 | 7.6758 | 0.0056 | 0.9671 | 0 | 0 | 0.2444 |
| 231 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 232 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 233 | 0.5711 | 0.1372 | 0.1973 | 2.031 | 0.1541 | 1.6047 | 0.0597 | 0.2403 | 0.6289 |
| 234 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 235 | 0.2636 | 0 | 0.044 | 3.5732 | 0.0587 | 0.3576 | 0 | 0 | 0.3853 |
| 236 | 2.6511 | 0 | 0.7753 | 46.0604 | 0 | 6.3044 | 0 | 0 | 0.039 |
| 237 | 1.0663 | 0.056 | 0.329 | 12.7008 | 0.0004 | 2.6757 | 0.003 | 0.0525 | 0.2331 |
| 238 | 0.6357 | 0 | 0.1581 | 11.0293 | 0.0009 | 1.2852 | 0 | 0 | 0.1594 |
| 239 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 240 | 2.0993 | 0 | 0.8145 | 34.8044 | 0 | 6.6232 | 0 | 0 | 0.0635 |
| 241 | 0.6014 | 0 | 0.0939 | 7.3236 | 0.0068 | 0.7636 | 0 | 0 | 0.1788 |
| 242 | 1.3529 | 0 | 0.197 | 20.0704 | 0 | 1.6019 | 0 | 0 | 0.0637 |
| 243 | 0.4994 | 0 | 0.116 | 8.6482 | 0.0033 | 0.9429 | 0 | 0 | 0.1943 |
| 244 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

| Site | alpha | beta | alpha= beta | LRT | p-value | Total branch length | dN/dS LB | dN/dS MLE | dN/dS UB |
|---|---|---|---|---|---|---|---|---|---|
| 245 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 246 | 0.0928 | 0 | 0.0366 | 2.1737 | 0.1404 | 0.2972 | 0 | 0 | 1.1363 |
| 247 | 0.7803 | 0 | 0.2206 | 12.6274 | 0.0004 | 1.7937 | 0 | 0 | 0.1503 |
| 248 | 0.1649 | 0 | 0.0395 | 2.8905 | 0.0891 | 0.3211 | 0 | 0 | 0.6024 |
| 249 | 1.0218 | 0.1248 | 0.4662 | 10.2373 | 0.0014 | 3.7913 | 0.0202 | 0.1222 | 0.3798 |
| 250 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 251 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 252 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 253 | 0.1668 | 0 | 0.0394 | 2.8836 | 0.0895 | 0.3208 | 0 | 0 | 0.5956 |
| 254 | 0.2793 | 0.1183 | 0.1771 | 0.8135 | 0.3671 | 1.4401 | 0.0704 | 0.4234 | 1.3076 |
| 255 | 1.2093 | 0.0823 | 0.5111 | 11.518 | 0.0007 | 4.156 | 0.0039 | 0.068 | 0.3022 |
| 256 | 0.1388 | 0 | 0.0395 | 2.5133 | 0.1129 | 0.3213 | 0 | 0 | 0.7658 |
| 257 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 258 | 1.963 | 0 | 0.4167 | 30.3846 | 0 | 3.3887 | 0 | 0 | 0.0513 |
| 259 | 0.9679 | 0 | 0.2005 | 15.4029 | 0.0001 | 1.6302 | 0 | 0 | 0.1003 |
| 260 | 0.2929 | 0 | 0.1083 | 5.9411 | 0.0148 | 0.8807 | 0 | 0 | 0.3766 |
| 261 | 0.1136 | 0.0508 | 0.0703 | 0.3166 | 0.5737 | 0.572 | 0.0255 | 0.4468 | 1.9804 |

## APPENDIX D: Supplemental Material for Chapter 5

https://doi.org/10.1371/journal.pone.0275623.s001 (ZIP)

D.S1 File. GISAID accession IDs for the analyses reported in Table 5.1. We also report the GISAID accession IDs for our Nextstrain background dataset.

D.S1 Table. List of sites found to be under diversifying positive selection by MEME ($p \leq 0.05$) along internal branches in BA.1, as well as biochemical properties that are important at this site (via the PRIME method). Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome. Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. % Of branches with omega>1: the fraction of tree branches (internal branches BA.1 clade) that have evidence of diversifying positive selection at this site (100%—pervasive selection, <100% –episodic selection). p-value: the p-value for the likelihood ratio test that a non-zero fraction of branches have omega > 1 (i.e., episodic diversifying selection at this branch). This is not corrected for multiple testing; the MEME test is generally conservative on real data. q-value: multiple-test corrected q-value (Benjamini-Hochberg). Properties: which, if any, of the five compositive biochemical properties are conserved or changed at this site.

| **Coordinate (SARS-CoV-2)** | **Gene/ORF** | **Codon (in gene/ORF)** | **# of selected branches** | **p-value** | **q-value** | **Properties** |
|---|---|---|---|---|---|---|
| 21574 | S | 5 | 4 | 0.015844 | 0.13269 | |
| 13492 | RdRp | 9 | 1 | 2.29E-05 | 0.000883 | |
| 13495 | RdRp | 10 | 1 | 0.001016 | 0.017964 | |
| 310 | leader | 16 | 3 | 0.029483 | 0.223404 | |
| 26627 | M | 36 | 1 | 0.000499 | 0.01115 | |
| 26630 | M | 37 | 1 | 0.013005 | 0.116586 | |
| 21700 | S | 47 | 1 | 0.012004 | 0.108375 | |
| 13627 | RdRp | 54 | 1 | 0.029008 | 0.222452 | |
| 27375 | ORF6 | 59 | 2 | 0.049607 | 0.31575 | |
| 25572 | ORF3a | 61 | 2 | 0.000278 | 0.007089 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| 27381 | ORF6 | 61 | 2 | 0.004124 | 0.050475 | |
| 21748 | S | 63 | 1 | 0.007205 | 0.078391 | |
| 21760 | S | 67 | 6 | 0.003726 | 0.047431 | |
| 475 | leader | 71 | 2 | 0.040071 | 0.268473 | |
| 26777 | M | 86 | 1 | 0.008946 | 0.088277 | |
| 520 | leader | 86 | 1 | 0.01483 | 0.128423 | |
| 21820 | S | 87 | 1 | 0.014878 | 0.127969 | |
| 21829 | S | 90 | 2 | 0.001736 | 0.026629 | |
| 21832 | S | 91 | 3 | 0.000126 | 0.003637 | |
| 21835 | S | 92 | 2 | 0.001631 | 0.025314 | |
| 21838 | S | 93 | 3 | 0.001477 | 0.0238 | |
| 21844 | S | 95 | 7 | 0.000741 | 0.01407 | |
| 21850 | S | 97 | 3 | 0.000342 | 0.008531 | |
| 21853 | S | 98 | 0 | 0.018952 | 0.156661 | |
| 21868 | S | 103 | 1 | 0.010576 | 0.098997 | |
| 21886 | S | 109 | 2 | 0.011796 | 0.107255 | |
| 21895 | S | 112 | 2 | 0.008227 | 0.084456 | |
| 21898 | S | 113 | 1 | 0.017605 | 0.146476 | |
| 21901 | S | 114 | 1 | 0.003994 | 0.049841 | |
| 21913 | S | 118 | 2 | 0.033717 | 0.241136 | |
| 21916 | S | 119 | 1 | 0.010857 | 0.100884 | |
| 21919 | S | 120 | 2 | 0.002841 | 0.039748 | |
| 28253 | S | 121 | 0 | 0.000624 | 0.013234 | |
| 21925 | S | 122 | 1 | 0.001542 | 0.02423 | |
| 21931 | S | 124 | 1 | 0.003304 | 0.044276 | |
| 21952 | S | 131 | 1 | 0.015139 | 0.129341 | |
| 26927 | M | 136 | 2 | 0.013118 | 0.115964 | |
| 21970 | S | 137 | 1 | 0.000344 | 0.00841 | |
| 21985 | S | 142 | 3 | 0.0015 | 0.023865 | |
| 21988 | S | 143 | 0 | 0.000489 | 0.011122 | |
| 21991 | S | 144 | 1 | 0.006734 | 0.075864 | |
| 21994 | S | 145 | 1 | 0.008682 | 0.08635 | |
| 709 | leader | 149 | 1 | 0.003024 | 0.041396 | |
| 22012 | S | 151 | 1 | 0.000226 | 0.006002 | |
| 25854 | ORF3a | 155 | 1 | 0.015593 | 0.131457 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| 22027 | S | 156 | 8 | 9.34E-10 | 1.70E-07 | |
| 22033 | S | 158 | 0 | 0.00354 | 0.046455 | |
| 22036 | S | 159 | 0 | 0.038196 | 0.260021 | |
| 22042 | S | 161 | 0 | 0.007428 | 0.078803 | |
| 22045 | S | 162 | 1 | 0.007384 | 0.079657 | |
| 13957 | RdRp | 164 | 1 | 0.030988 | 0.225419 | |
| 22054 | S | 165 | 1 | 0.000838 | 0.015465 | |
| 22060 | S | 167 | 2 | 0.000396 | 0.009161 | |
| 22063 | S | 168 | 1 | 0.002964 | 0.041017 | |
| 22075 | S | 172 | 2 | 0.003277 | 0.044382 | |
| 22078 | S | 173 | 1 | 3.27E-05 | 0.001189 | |
| 22081 | S | 174 | 1 | 0.008959 | 0.087733 | |
| 22084 | S | 175 | 2 | 2.24E-09 | 3.17E-07 | |
| 808 | nsp2 | 182 | 1 | 0.036615 | 0.253321 | |
| 22138 | S | 193 | 5 | 0.000115 | 0.003418 | |
| 27137 | M | 206 | 2 | 0.029604 | 0.222993 | |
| 27140 | M | 207 | 1 | 0.009762 | 0.092737 | |
| 27146 | M | 209 | 2 | 0.031871 | 0.230524 | |
| 22192 | S | 211 | 7 | 1.37E-10 | 3.49E-08 | |
| 22195 | S | 212 | 20 | 1.11E-15 | 7.07E-13 | |
| 22198 | S | 213 | 3 | 7.82E-05 | 0.00237 | |
| 22201 | S | 214 | 5 | 0.006607 | 0.075098 | |
| 14107 | RdRp | 214 | 2 | 0.030762 | 0.226358 | |
| 27176 | M | 219 | 0 | 0.003364 | 0.044602 | |
| 27185 | M | 222 | 0 | 0.036138 | 0.251388 | |
| 22237 | S | 226 | 1 | 0.000224 | 0.006065 | |
| 26076 | ORF3a | 229 | 0 | 0.042947 | 0.284747 | |
| 22267 | S | 236 | 1 | 1.35E-05 | 0.000555 | |
| 22270 | S | 237 | 1 | 0.003577 | 0.046467 | |
| 22279 | S | 240 | 1 | 5.4E-06 | 0.000312 | |
| 1084 | nsp2 | 274 | 1 | 0.026747 | 0.210175 | |
| 1105 | nsp2 | 281 | 1 | 0.009651 | 0.092378 | |
| 14407 | RdRp | 314 | 2 | 0.006851 | 0.0765 | |
| 22576 | S | 339 | 13 | 3.49E-07 | 3.41E-05 | |
| 22597 | S | 346 | 41 | 0 | 0 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| 22624 | S | 355 | 1 | 3.9E-06 | 0.000261 | |
| 22630 | S | 357 | 1 | 0.001302 | 0.022402 | |
| 22639 | S | 360 | 1 | 0.000844 | 0.015348 | |
| 22648 | S | 363 | 1 | 0.009609 | 0.092668 | |
| 22672 | S | 371 | 4 | 2.57E-05 | 0.000961 | |
| 22675 | S | 372 | 3 | 7.05E-07 | 5.98E-05 | |
| 22678 | S | 373 | 7 | 1.56E-05 | 0.00062 | |
| 22684 | S | 375 | 9 | 6.87E-06 | 0.000336 | |
| 22687 | S | 376 | 3 | 0.002359 | 0.033737 | |
| 22777 | S | 406 | 1 | 9.52E-06 | 0.000433 | |
| 22810 | S | 417 | 23 | 1.13E-09 | 1.80E-07 | |
| 22855 | S | 432 | 1 | 0.001837 | 0.027833 | |
| 22879 | S | 440 | 18 | 3.92E-10 | 8.32E-08 | |
| 22897 | S | 446 | 13 | 2.92E-07 | 3.09E-05 | |
| 22915 | S | 452 | 8 | 8.85E-09 | 1.13E-06 | |
| 22930 | S | 457 | 1 | 0.000695 | 0.013825 | |
| 14857 | RdRp | 464 | 1 | 0.045963 | 0.298523 | |
| 22990 | S | 477 | 16 | 1.12E-10 | 3.56E-08 | |
| 22993 | S | 478 | 14 | 3.34E-06 | 0.000236 | |
| 23008 | S | 483 | 1 | 0.000348 | 0.00837 | |
| 23011 | S | 484 | 12 | 4.6E-06 | 0.000293 | |
| 23020 | S | 487 | 1 | 0.000583 | 0.012804 | |
| 23038 | S | 493 | 8 | 5.51E-06 | 0.000292 | |
| 23047 | S | 496 | 6 | 0.001315 | 0.022019 | |
| 23053 | S | 498 | 7 | 7.75E-05 | 0.002406 | |
| 23059 | S | 500 | 1 | 0.001028 | 0.017926 | |
| 23062 | S | 501 | 5 | 0.001915 | 0.028674 | |
| 23071 | S | 504 | 3 | 3.00E-08 | 3.48E-06 | |
| 23074 | S | 505 | 10 | 4.82E-07 | 4.38E-05 | |
| 23098 | S | 513 | 1 | 0.001383 | 0.022573 | |
| 23101 | S | 514 | 1 | 5.82E-06 | 0.000297 | |
| 15196 | RdRp | 577 | 1 | 0.046789 | 0.300817 | |
| 2035 | nsp2 | 591 | 2 | 0.01171 | 0.108019 | |
| 23488 | S | 643 | 5 | 0.039473 | 0.267282 | |
| 23524 | S | 655 | 3 | 0.036984 | 0.254493 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| 15451 | RdRp | 662 | 2 | 0.005997 | 0.069397 | |
| 23596 | S | 679 | 9 | 6.13E-05 | 0.002108 | |
| 23602 | S | 681 | 11 | 1.02E-05 | 0.000447 | |
| 2305 | nsp2 | 681 | 2 | 0.015522 | 0.131726 | |
| 15520 | RdRp | 685 | 1 | 0.045779 | 0.298856 | |
| 23662 | S | 701 | 20 | 3.68E-11 | 1.56E-08 | |
| 2449 | nsp2 | 729 | 1 | 0.035035 | 0.247779 | |
| 23821 | S | 754 | 1 | 6.65E-05 | 0.00217 | |
| 23851 | S | 764 | 10 | 0.000273 | 0.007088 | |
| 15763 | RdRp | 766 | 1 | 0.011755 | 0.107652 | |
| 15826 | RdRp | 787 | 1 | 0.001303 | 0.022115 | |
| 23947 | S | 796 | 11 | 0.000148 | 0.004097 | |
| 2830 | nsp3 | 856 | 4 | 0.008044 | 0.083251 | |
| 24127 | S | 856 | 3 | 0.020423 | 0.16773 | |
| 24262 | S | 901 | 1 | 0.000685 | 0.014071 | |
| 2980 | nsp3 | 906 | 1 | 0.02996 | 0.223036 | |
| 2989 | nsp3 | 909 | 1 | 0.013554 | 0.118993 | |
| 2992 | nsp3 | 910 | 1 | 0.005688 | 0.067049 | |
| 3016 | nsp3 | 918 | 2 | 8.81E-07 | 7.01E-05 | |
| 24421 | S | 954 | 4 | 0.027933 | 0.216819 | |
| 24460 | S | 967 | 1 | 4.7E-06 | 0.000285 | |
| 24466 | S | 969 | 4 | 0.0098 | 0.092414 | |
| 24502 | S | 981 | 5 | 0.005357 | 0.063734 | |
| 24802 | S | 1081 | 3 | 0.008457 | 0.086123 | |
| 3544 | nsp3 | 1094 | 1 | 0.001966 | 0.029094 | |
| 24943 | S | 1128 | 2 | 0.02581 | 0.205355 | |
| 3652 | nsp3 | 1130 | 2 | 0.020453 | 0.166901 | |
| 3682 | nsp3 | 1140 | 1 | 0.00366 | 0.047065 | |
| 3829 | nsp3 | 1189 | 1 | 0.003999 | 0.04942 | |
| 25342 | S | 1261 | 1 | 0.005913 | 0.069055 | |
| 4087 | nsp3 | 1275 | 1 | 0.035538 | 0.249941 | |
| 17409 | helicase | 1314 | 2 | 0.021622 | 0.175321 | |
| 4258 | nsp3 | 1332 | 1 | 0.004306 | 0.05221 | |
| 17820 | helicase | 1451 | 2 | 0.040558 | 0.270316 | |
| 18021 | helicase | 1518 | 2 | 0.027976 | 0.215838 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| 18027 | helicase | 1520 | 2 | 6.93E-05 | 0.002206 | |
| 18030 | helicase | 1521 | 2 | 0.013034 | 0.116026 | |
| 18267 | exonuclease | 1600 | 2 | 0.000714 | 0.013992 | |
| 5260 | nsp3 | 1666 | 2 | 5.48E-06 | 0.000303 | |
| 5263 | nsp3 | 1667 | 0 | 0.007031 | 0.077159 | |
| 5266 | nsp3 | 1668 | 1 | 0.000609 | 0.013136 | |
| 5272 | nsp3 | 1670 | 0 | 0.032228 | 0.231784 | |
| 5278 | nsp3 | 1672 | 1 | 0.000388 | 0.009151 | |
| 5701 | nsp3 | 1813 | 1 | 0.025992 | 0.205512 | |
| 5818 | nsp3 | 1852 | 1 | 0.002397 | 0.033901 | |
| 5830 | nsp3 | 1856 | 1 | 0.046869 | 0.299818 | |
| 5923 | nsp3 | 1887 | 3 | 0.002338 | 0.033818 | |
| 19218 | exonuclease | 1917 | 2 | 0.023424 | 0.188723 | |
| 6055 | nsp3 | 1931 | 1 | 8.61E-06 | 0.000406 | |
| 19281 | exonuclease | 1938 | 2 | 0.005997 | 0.068778 | |
| 19332 | exonuclease | 1955 | 0 | 6.25E-05 | 0.002095 | |
| 19338 | exonuclease | 1957 | 0 | 0.00073 | 0.014089 | |
| 6367 | nsp3 | 2035 | 1 | 1.09E-05 | 0.000461 | |
| 6400 | nsp3 | 2046 | 3 | 0.008641 | 0.08661 | |
| 6445 | nsp3 | 2061 | 1 | 0.000692 | 0.01398 | |
| 6448 | nsp3 | 2062 | 2 | 0.034731 | 0.246998 | |
| 6460 | nsp3 | 2066 | 2 | 1.39E-06 | 0.000104 | |
| 6511 | nsp3 | 2083 | 1 | 0.003735 | 0.047077 | |
| 6514 | nsp3 | 2084 | 2 | 0.009006 | 0.08752 | |
| 6562 | nsp3 | 2100 | 1 | 0.00227 | 0.033219 | |
| 6577 | nsp3 | 2105 | 2 | 0.000882 | 0.015819 | |
| 19998 | endornase | 2177 | 2 | 0.044163 | 0.291291 | |
| 20016 | endornase | 2183 | 1 | 0.00688 | 0.076157 | |
| 6883 | nsp3 | 2207 | 1 | 0.001363 | 0.022532 | |
| 20283 | endornase | 2272 | 1 | 0.007402 | 0.079178 | |
| 7333 | nsp3 | 2357 | 1 | 5.11E-05 | 0.001808 | |
| 7924 | nsp3 | 2554 | 2 | 0.030867 | 0.225828 | |
| 21135 | | 2556 | 2 | 0.04427 | 0.29049 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| 21168 | | 2567 | 1 | 0.035717 | 0.249823 | |
| 21303 | | 2612 | 0 | 0.02788 | 0.217735 | |
| 21534 | | 2689 | 1 | 0.029845 | 0.223488 | |
| 21543 | | 2692 | 1 | 0.014442 | 0.125926 | |
| 8392 | nsp3 | 2710 | 3 | 0.025402 | 0.203376 | |
| 10447 | | 3395 | 3 | 0.007662 | 0.080613 | |
| 10624 | | 3454 | 1 | 0.005074 | 0.06093 | |
| 10630 | | 3456 | 1 | 0.008006 | 0.083539 | |
| 10633 | | 3457 | 3 | 0.000787 | 0.014735 | |
| 10891 | | 3543 | 1 | 0.029466 | 0.224614 | |
| 11080 | nsp6 | 3606 | 2 | 0.008496 | 0.085834 | |
| 11278 | nsp6 | 3672 | 1 | 0.037287 | 0.255193 | |
| 11281 | nsp6 | 3673 | 1 | 0.030085 | 0.222662 | |
| 11284 | nsp6 | 3674 | 0 | 0.000645 | 0.013465 | |
| 11287 | nsp6 | 3675 | 1 | 0.000132 | 0.003731 | |
| 11290 | nsp6 | 3676 | 1 | 0.046357 | 0.299558 | |
| 11635 | nsp6 | 3791 | 1 | 0.039802 | 0.268083 | |

D.S2 Table. List of sites found to be under diversifying positive selection by MEME ($p \leq 0.05$) along all branches in BA.1, as well as biochemical properties that are important at this site (via the PRIME method). Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome. Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. % Of branches with omega>1: the fraction of tree branches (internal branches BA.1 clade) that have evidence of diversifying positive selection at this site (100%—pervasive selection, <100% –episodic selection). p-value: the p-value for the likelihood ratio test that a non-zero fraction of branches have omega > 1 (i.e., episodic diversifying selection at this branch). This is not corrected for multiple testing; the MEME test is generally conservative on real data. q-value: multiple-test corrected q-value (Benjamini-Hochberg). Properties: which, if any, of the five compositive biochemical properties are conserved or changed at this site.

| **Coordinate (SARS-CoV-2)** | **Gene/ORF** | **Codon (in gene/ORF)** | **# of selected branches** | **p-value** | **p-value** | **q-value** | **Properties** |
|---|---|---|---|---|---|---|---|
| 25395 | ORF3a | 2 | 3 | 0.002832 | 0.666667 | 0.017844 | |
| 21568 | S | 3 | 3 | 0.001605 | 0.666667 | 0.01195 | |
| 21574 | S | 5 | 14 | 0.003256 | 0.015844 | 0.020121 | |
| 21580 | S | 7 | 1 | 0.016153 | 0.666667 | 0.071153 | |
| 21583 | S | 8 | 1 | 0.002091 | 0.666667 | 0.01416 | |
| 286 | ORF1a | 8 | 4 | 0.019012 | 0.068106 | 0.081216 | |
| 13492 | ORF1b | 9 | 3 | 0.000313 | 2.29E-05 | 0.003793 | |
| 25425 | ORF3a | 12 | 3 | 5.84E-05 | 0.248579 | 0.001162 | |
| 21595 | S | 12 | 2 | 0.000538 | 0.666667 | 0.005394 | |
| 21598 | S | 13 | 1 | 0.018268 | 0.666667 | 0.078563 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 13519 | ORF1b | 18 | 1 | 0.002334 | 0.666667 | 0.015478 | |
| 316 | ORF1a | 18 | 1 | 0.015556 | 0.197439 | 0.06924 | |
| 21616 | S | 19 | 7 | 0.036496 | 0.165796 | 0.139938 | |
| 21637 | S | 26 | 5 | 0.036118 | 0.666667 | 0.138909 | |
| 26627 | M | 36 | 3 | 0.000952 | 0.000499 | 0.007921 | |
| 25515 | ORF3a | 42 | 5 | 0.028568 | 0.067609 | 0.114005 | |
| 25527 | ORF3a | 46 | 2 | 0.007163 | 0.666667 | 0.037838 | |
| 21700 | S | 47 | 1 | 0.016444 | 0.012004 | 0.071934 | |
| 27564 | ORF7a | 58 | 2 | 0.007837 | 0.666667 | 0.040067 | |
| 25572 | ORF3a | 61 | 3 | 0.00142 | 0.000278 | 0.010887 | |
| 27381 | ORF6 | 61 | 1 | 0.043466 | 0.004124 | 0.158092 | |
| 21748 | S | 63 | 1 | 0.010338 | 0.007205 | 0.050231 | |
| 21760 | S | 67 | 13 | 0.005262 | 0.003726 | 0.029638 | |
| 27591 | ORF7a | 67 | 1 | 0.044244 | 0.666667 | 0.160007 | |
| 469 | ORF1a | 69 | 3 | 0.042171 | 0.666667 | 0.15471 | |
| 13675 | ORF1b | 70 | 1 | 2.30E-07 | 0.666667 | 1.95E-05 | |
| 21769 | S | 70 | 0 | 0.005099 | 0.13061 | 0.028975 | |
| 21784 | S | 75 | 2 | 0.004093 | 0.666667 | 0.024459 | |
| 21790 | S | 77 | 1 | 0.000418 | 0.666667 | 0.004505 | |
| 21793 | S | 78 | 3 | 0.002188 | 0.666667 | 0.014736 | |
| 21799 | S | 80 | 1 | 0.01216 | 0.666667 | 0.056703 | |
| 21805 | S | 82 | 1 | 0.000124 | 0.666667 | 0.002134 | |
| 508 | ORF1a | 82 | 2 | 0.015916 | 0.310557 | 0.070595 | |
| 26765 | M | 82 | 6 | 0.039033 | 0.136749 | 0.147011 | |
| 21814 | S | 85 | 4 | 0.00134 | 0.104074 | 0.010402 | |
| 520 | ORF1a | 86 | 4 | 0.011659 | 0.01483 | 0.055382 | |
| 21820 | S | 87 | 4 | 1.37E-05 | 0.014878 | 0.000397 | |
| 21829 | S | 90 | 3 | 0.001757 | 0.001736 | 0.012639 | |
| 21832 | S | 91 | 2 | 0.001312 | 0.000126 | 0.010244 | |
| 21835 | S | 92 | 2 | 0.001138 | 0.001631 | 0.009228 | |
| 21838 | S | 93 | 2 | 0.007333 | 0.001477 | 0.038415 | |
| 21844 | S | 95 | 14 | 2.20E-05 | 0.000741 | 0.000519 | |
| 27675 | ORF7a | 95 | 4 | 0.020635 | 0.364173 | 0.086695 | |
| 21850 | S | 97 | 3 | 1.07E-05 | 0.000342 | 0.000341 | |
| 21853 | S | 98 | 2 | 0.027611 | 0.018952 | 0.11123 | |
| 26816 | M | 99 | 2 | 0.024211 | 0.666667 | 0.100067 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 25695 | ORF3a | 102 | 3 | 0.037981 | 0.666667 | 0.144761 | |
| 21895 | S | 112 | 4 | 0.016344 | 0.008227 | 0.071746 | |
| 21898 | S | 113 | 2 | 0.009122 | 0.017605 | 0.045361 | |
| 21901 | S | 114 | 0 | 0.012637 | 0.003994 | 0.058499 | |
| 26864 | M | 115 | 1 | 0.000523 | 0.666667 | 0.005327 | |
| 21910 | S | 117 | 1 | 0.012547 | 0.666667 | 0.058292 | |
| 27741 | ORF7a | 117 | 2 | 0.034795 | 0.666667 | 0.134632 | |
| 27744 | ORF7a | 118 | 1 | 0.01092 | 0.666667 | 0.052657 | |
| 21916 | S | 119 | 1 | 0.049579 | 0.010857 | 0.175805 | |
| 21919 | S | 120 | 2 | 0.025315 | 0.002841 | 0.103956 | |
| 21925 | S | 122 | 1 | 0.003836 | 0.001542 | 0.023368 | |
| 21931 | S | 124 | 1 | 0.008839 | 0.003304 | 0.044126 | |
| 25767 | ORF3a | 126 | 1 | 0.017015 | 0.666667 | 0.073674 | |
| 21946 | S | 129 | 1 | 3.88E-06 | 0.666667 | 0.000176 | |
| 21952 | S | 131 | 1 | 0.013626 | 0.015139 | 0.061509 | |
| 21955 | S | 132 | 1 | 0.040262 | 0.051957 | 0.149426 | |
| 25797 | ORF3a | 136 | 2 | 0.000384 | 0.666667 | 0.004254 | |
| 26927 | M | 136 | 3 | 0.004945 | 0.013118 | 0.028229 | |
| 21970 | S | 137 | 1 | 0.00034 | 0.000344 | 0.003896 | |
| 21976 | S | 139 | 2 | 8.38E-06 | 0.666667 | 0.000296 | |
| 685 | ORF1a | 141 | 2 | 0.038371 | 0.666667 | 0.145376 | |
| 25815 | ORF3a | 142 | 5 | 0.011473 | 0.666667 | 0.054699 | |
| 21988 | S | 143 | 0 | 0.000324 | 0.000489 | 0.00385 | |
| 26951 | M | 144 | 2 | 0.000686 | 0.134694 | 0.006283 | |
| 21991 | S | 144 | 1 | 0.012882 | 0.006734 | 0.058989 | |
| 25821 | ORF3a | 144 | 5 | 0.024794 | 0.087388 | 0.102144 | |
| 21994 | S | 145 | 3 | 0.000311 | 0.008682 | 0.003805 | |
| 709 | ORF1a | 149 | 3 | 0.00515 | 0.003024 | 0.029136 | |
| 22012 | S | 151 | 3 | 0.00032 | 0.000226 | 0.00384 | |
| 25845 | ORF3a | 152 | 3 | 0.016917 | 0.666667 | 0.0735 | |
| 25854 | ORF3a | 155 | 6 | 0.000695 | 0.015593 | 0.006323 | |
| 22024 | S | 155 | 2 | 0.013114 | 0.666667 | 0.059621 | |
| 22030 | S | 157 | 2 | 1.33E-05 | 0.450037 | 0.000394 | |
| 22033 | S | 158 | 4 | 0.004857 | 0.00354 | 0.027976 | |
| 22042 | S | 161 | 0 | 0.034425 | 0.007428 | 0.133607 | |
| 22045 | S | 162 | 2 | 0.023031 | 0.007384 | 0.09581 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 22048 | S | 163 | 1 | 0.000271 | 0.666667 | 0.003675 | |
| 22051 | S | 164 | 1 | 0.040273 | 0.666667 | 0.149033 | |
| 22054 | S | 165 | 1 | 0.004457 | 0.000838 | 0.026025 | |
| 25884 | ORF3a | 165 | 4 | 0.047364 | 0.066381 | 0.168893 | |
| 22060 | S | 167 | 4 | 0.005534 | 0.000396 | 0.030761 | |
| 22063 | S | 168 | 5 | 0.027424 | 0.002964 | 0.111181 | |
| 22069 | S | 170 | 4 | 0.011698 | 0.398533 | 0.05536 | |
| 22072 | S | 171 | 1 | 0.00013 | 0.666667 | 0.002172 | |
| 22075 | S | 172 | 4 | 0.008288 | 0.003277 | 0.041702 | |
| 22078 | S | 173 | 4 | 1.53E-05 | 3.27E-05 | 0.000434 | |
| 27038 | M | 173 | 2 | 0.001919 | 0.231893 | 0.013349 | |
| 22081 | S | 174 | 2 | 0.000171 | 0.008959 | 0.002625 | |
| 22084 | S | 175 | 3 | 5.33E-09 | 2.24E-09 | 9.69E-07 | |
| 22087 | S | 176 | 2 | 0.003591 | 0.255798 | 0.022084 | |
| 25929 | ORF3a | 180 | 2 | 0.00023 | 0.666667 | 0.003286 | |
| 22138 | S | 193 | 8 | 0.000286 | 0.000115 | 0.003645 | |
| 22144 | S | 195 | 3 | 3.13E-05 | 0.666667 | 0.000686 | |
| 22150 | S | 197 | 2 | 0.001659 | 0.666667 | 0.012277 | |
| 26001 | ORF3a | 204 | 3 | 0.047038 | 0.666667 | 0.168201 | |
| 27137 | M | 206 | 3 | 0.008182 | 0.029604 | 0.041332 | |
| 22180 | S | 207 | 1 | 4.22E-05 | 0.666667 | 0.000854 | |
| 27140 | M | 207 | 3 | 0.00549 | 0.009762 | 0.03065 | |
| 22183 | S | 208 | 2 | 0.000677 | 0.666667 | 0.006335 | |
| 27146 | M | 209 | 3 | 0.020348 | 0.031871 | 0.086058 | |
| 22189 | S | 210 | 2 | 0.002697 | 0.666667 | 0.017165 | |
| 22192 | S | 211 | 2 | 9.62E-09 | 1.37E-10 | 1.36E-06 | |
| 22195 | S | 212 | 30 | 2.92E-12 | 1.11E-15 | 1.86E-09 | |
| 22198 | S | 213 | 5 | 0.000941 | 7.82E-05 | 0.007883 | |
| 22201 | S | 214 | 9 | 0.000157 | 0.006607 | 0.002556 | |
| 22204 | S | 215 | 1 | 0.005666 | 0.633016 | 0.031362 | |
| 22207 | S | 216 | 1 | 0.043485 | 0.666667 | 0.157709 | |
| 22213 | S | 218 | 2 | 0.010052 | 0.47262 | 0.049216 | |
| 27176 | M | 219 | 0 | 0.006461 | 0.003364 | 0.034706 | |
| 27182 | M | 221 | 3 | 4.33E-06 | 0.666667 | 0.000178 | |
| 22225 | S | 222 | 6 | 0.013803 | 0.666667 | 0.06209 | |
| 931 | ORF1a | 223 | 2 | 8.34E-05 | 0.666667 | 0.001584 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 937 | ORF1a | 225 | 2 | 0.000599 | 0.666667 | 0.00582 | |
| 22237 | S | 226 | 1 | 0.000292 | 0.000224 | 0.003682 | |
| 22243 | S | 228 | 2 | 0.002249 | 0.666667 | 0.015067 | |
| 22246 | S | 229 | 2 | 0.000785 | 0.150733 | 0.006894 | |
| 26076 | ORF3a | 229 | 7 | 0.03362 | 0.042947 | 0.13088 | |
| 22267 | S | 236 | 2 | 1.32E-05 | 1.35E-05 | 0.0004 | |
| 22270 | S | 237 | 2 | 0.002634 | 0.003577 | 0.016847 | |
| 22279 | S | 240 | 3 | 1.60E-05 | 5.40E-06 | 0.000423 | |
| 22315 | S | 252 | 1 | 0.000633 | 0.666667 | 0.006103 | |
| 1045 | ORF1a | 261 | 2 | 0.001011 | 0.666667 | 0.008305 | |
| 1084 | ORF1a | 274 | 1 | 0.035543 | 0.026747 | 0.137109 | |
| 1105 | ORF1a | 281 | 2 | 0.021264 | 0.009651 | 0.089043 | |
| 22417 | S | 286 | 1 | 1.92E-06 | 0.666667 | 0.000106 | |
| 22576 | S | 339 | 31 | 3.38E-08 | 3.49E-07 | 3.59E-06 | |
| 22597 | S | 346 | 69 | 2.35E-14 | 0 | 3.00E-11 | |
| 22624 | S | 355 | 2 | 8.90E-06 | 3.90E-06 | 0.000306 | |
| 22630 | S | 357 | 1 | 0.001203 | 0.001302 | 0.009632 | |
| 22639 | S | 360 | 1 | 0.001578 | 0.000844 | 0.011817 | |
| 22648 | S | 363 | 1 | 0.017815 | 0.009609 | 0.076877 | |
| 22672 | S | 371 | 10 | 7.89E-07 | 2.57E-05 | 5.58E-05 | |
| 22675 | S | 372 | 3 | 2.39E-07 | 7.05E-07 | 1.90E-05 | |
| 22678 | S | 373 | 15 | 0.000247 | 1.56E-05 | 0.003383 | |
| 22684 | S | 375 | 18 | 3.08E-06 | 6.87E-06 | 0.000151 | |
| 22687 | S | 376 | 3 | 0.012037 | 0.002359 | 0.056545 | |
| 22708 | S | 383 | 1 | 0.000106 | 0.666667 | 0.001903 | |
| 14620 | ORF1b | 385 | 2 | 0.001659 | 0.666667 | 0.01221 | |
| 14623 | ORF1b | 386 | 1 | 0.007524 | 0.666667 | 0.038933 | |
| 22777 | S | 406 | 2 | 3.00E-06 | 9.52E-06 | 0.000153 | |
| 22810 | S | 417 | 38 | 2.45E-08 | 1.13E-09 | 3.12E-06 | |
| 1516 | ORF1a | 418 | 1 | 0.001896 | 0.666667 | 0.013262 | |
| 22855 | S | 432 | 1 | 0.001243 | 0.001837 | 0.009886 | |
| 22873 | S | 438 | 2 | 1.64E-06 | 0.666667 | 9.51E-05 | |
| 22879 | S | 440 | 28 | 6.46E-08 | 3.92E-10 | 6.33E-06 | |
| 22894 | S | 445 | 1 | 0.000843 | 0.666667 | 0.007158 | |
| 22897 | S | 446 | 24 | 3.83E-07 | 2.92E-07 | 2.87E-05 | |
| 14806 | ORF1b | 447 | 1 | 0.03131 | 0.666667 | 0.123018 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 14809 | ORF1b | 448 | 1 | 0.000434 | 0.666667 | 0.004643 | |
| 22906 | S | 449 | 2 | 0.000303 | 0.666667 | 0.003782 | |
| 14812 | ORF1b | 449 | 1 | 0.010867 | 0.666667 | 0.052601 | |
| 14818 | ORF1b | 451 | 1 | 0.000782 | 0.666667 | 0.006909 | |
| 22915 | S | 452 | 18 | 4.43E-12 | 8.85E-09 | 1.88E-09 | |
| 22930 | S | 457 | 1 | 0.000773 | 0.000695 | 0.006879 | |
| 22981 | S | 474 | 2 | 0.004131 | 0.666667 | 0.024572 | |
| 22990 | S | 477 | 1 | 4.15E-09 | 1.12E-10 | 8.81E-07 | |
| 22993 | S | 478 | 21 | 0.000127 | 3.34E-06 | 0.002163 | |
| 22999 | S | 480 | 2 | 2.13E-05 | 0.08769 | 0.000512 | |
| 23002 | S | 481 | 2 | 0.031897 | 0.234057 | 0.124937 | |
| 23008 | S | 483 | 1 | 0.000353 | 0.000348 | 0.003977 | |
| 23011 | S | 484 | 31 | 9.45E-08 | 4.60E-06 | 8.60E-06 | |
| 23020 | S | 487 | 2 | 0.00185 | 0.000583 | 0.013013 | |
| 14932 | ORF1b | 489 | 1 | 0.010061 | 0.666667 | 0.049073 | |
| 14935 | ORF1b | 490 | 2 | 0.001277 | 0.293426 | 0.010034 | |
| 23038 | S | 493 | 20 | 8.09E-09 | 5.51E-06 | 1.29E-06 | |
| 14944 | ORF1b | 493 | 1 | 0.002463 | 0.666667 | 0.016076 | |
| 23041 | S | 494 | 1 | 0.000915 | 0.666667 | 0.007715 | |
| 23047 | S | 496 | 11 | 0.000836 | 0.001315 | 0.007145 | |
| 23053 | S | 498 | 11 | 0.00055 | 7.75E-05 | 0.005382 | |
| 23059 | S | 500 | 2 | 2.92E-06 | 0.001028 | 0.000155 | |
| 23062 | S | 501 | 15 | 0.000121 | 0.001915 | 0.002146 | |
| 23071 | S | 504 | 4 | 2.91E-08 | 3.00E-08 | 3.36E-06 | |
| 23074 | S | 505 | 29 | 2.12E-10 | 4.82E-07 | 6.73E-08 | |
| 14998 | ORF1b | 511 | 3 | 0.000353 | 0.666667 | 0.004007 | |
| 23098 | S | 513 | 1 | 0.001811 | 0.001383 | 0.012809 | |
| 23101 | S | 514 | 1 | 4.30E-06 | 5.82E-06 | 0.000182 | |
| 15016 | ORF1b | 517 | 3 | 0.005476 | 0.666667 | 0.03071 | |
| 23128 | S | 523 | 3 | 0.039849 | 0.07043 | 0.148763 | |
| 15037 | ORF1b | 524 | 2 | 0.001727 | 0.666667 | 0.012635 | |
| 15043 | ORF1b | 526 | 1 | 0.001509 | 0.666667 | 0.011432 | |
| 23143 | S | 528 | 1 | 9.22E-05 | 0.666667 | 0.001727 | |
| 15163 | ORF1b | 566 | 1 | 0.001464 | 0.666667 | 0.011162 | |
| 23320 | S | 587 | 2 | 0.004698 | 0.666667 | 0.027184 | |
| 23473 | S | 638 | 1 | 0.002536 | 0.666667 | 0.01639 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 23524 | S | 655 | 11 | 0.029543 | 0.036984 | 0.117524 | |
| 2233 | ORF1a | 657 | 1 | 0.000164 | 0.666667 | 0.002638 | |
| 15454 | ORF1b | 663 | 1 | 7.69E-06 | 0.666667 | 0.00028 | |
| 23554 | S | 665 | 1 | 2.78E-05 | 0.666667 | 0.000622 | |
| 2269 | ORF1a | 669 | 3 | 0.04493 | 0.666667 | 0.162028 | |
| 23596 | S | 679 | 18 | 0.000168 | 6.13E-05 | 0.00261 | |
| 2302 | ORF1a | 680 | 2 | 0.000534 | 0.666667 | 0.005396 | |
| 23602 | S | 681 | 21 | 3.92E-05 | 1.02E-05 | 0.000806 | |
| 23662 | S | 701 | 43 | 3.37E-10 | 3.68E-11 | 8.57E-08 | |
| 2428 | ORF1a | 722 | 1 | 0.008116 | 0.666667 | 0.04116 | |
| 23821 | S | 754 | 2 | 0.000276 | 6.65E-05 | 0.003623 | |
| 23851 | S | 764 | 18 | 0.002547 | 0.000273 | 0.016377 | |
| 23908 | S | 783 | 2 | 0.040132 | 0.666667 | 0.149381 | |
| 15826 | ORF1b | 787 | 2 | 0.001952 | 0.001303 | 0.01343 | |
| 23947 | S | 796 | 22 | 0.00051 | 0.000148 | 0.005283 | |
| 24004 | S | 815 | 2 | 0.029663 | 0.666667 | 0.117635 | |
| 24127 | S | 856 | 15 | 0.003217 | 0.020423 | 0.019979 | |
| 2905 | ORF1a | 881 | 3 | 0.007335 | 0.666667 | 0.038267 | |
| 24262 | S | 901 | 2 | 0.001752 | 0.000685 | 0.012676 | |
| 2989 | ORF1a | 909 | 1 | 0.016026 | 0.013554 | 0.070838 | |
| 3016 | ORF1a | 918 | 2 | 9.66E-07 | 8.81E-07 | 6.47E-05 | |
| 24409 | S | 950 | 7 | 0.019034 | 0.055632 | 0.081036 | |
| 24421 | S | 954 | 12 | 0.007635 | 0.027933 | 0.03935 | |
| 24460 | S | 967 | 1 | 5.55E-06 | 4.70E-06 | 0.000214 | |
| 24466 | S | 969 | 10 | 0.027784 | 0.0098 | 0.111573 | |
| 24475 | S | 972 | 2 | 0.00028 | 0.666667 | 0.003597 | |
| 24481 | S | 974 | 1 | 0.000174 | 0.666667 | 0.002613 | |
| 24490 | S | 977 | 1 | 0.002263 | 0.666667 | 0.015084 | |
| 24493 | S | 978 | 1 | 0.014581 | 0.666667 | 0.06513 | |
| 24496 | S | 979 | 1 | 0.007233 | 0.666667 | 0.038048 | |
| 24499 | S | 980 | 1 | 0.004071 | 0.666667 | 0.024443 | |
| 24502 | S | 981 | 17 | 0.000246 | 0.005357 | 0.003407 | |
| 3343 | ORF1a | 1027 | 1 | 0.006304 | 0.666667 | 0.034147 | |
| 24649 | S | 1030 | 1 | 0.001742 | 0.666667 | 0.012675 | |
| 3415 | ORF1a | 1051 | 2 | 0.042214 | 0.666667 | 0.15442 | |
| 3424 | ORF1a | 1054 | 3 | 0.045403 | 0.666667 | 0.16327 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 24781 | S | 1074 | 3 | 0.048972 | 0.666667 | 0.174137 | |
| 24802 | S | 1081 | 15 | 0.000338 | 0.008457 | 0.003916 | |
| 3544 | ORF1a | 1094 | 2 | 1.24E-05 | 0.001966 | 0.000384 | |
| 16782 | ORF1b | 1105 | 1 | 0.000391 | 0.666667 | 0.004286 | |
| 3634 | ORF1a | 1124 | 1 | 1.19E-06 | 0.666667 | 7.20E-05 | |
| 3640 | ORF1a | 1126 | 2 | 0.000165 | 0.666667 | 0.002621 | |
| 24943 | S | 1128 | 4 | 0.027391 | 0.02581 | 0.1114 | |
| 25048 | S | 1163 | 2 | 0.000829 | 0.666667 | 0.007182 | |
| 17031 | ORF1b | 1188 | 1 | 0.003712 | 0.666667 | 0.022716 | |
| 3829 | ORF1a | 1189 | 1 | 0.01383 | 0.003999 | 0.06199 | |
| 25162 | S | 1201 | 2 | 0.000835 | 0.666667 | 0.007185 | |
| 25183 | S | 1208 | 2 | 6.58E-06 | 0.666667 | 0.000246 | |
| 3943 | ORF1a | 1227 | 1 | 0.000401 | 0.666667 | 0.004359 | |
| 25285 | S | 1242 | 1 | 0.000274 | 0.666667 | 0.003674 | |
| 17205 | ORF1b | 1246 | 2 | 0.002842 | 0.666667 | 0.017819 | |
| 25342 | S | 1261 | 1 | 0.007424 | 0.005913 | 0.038573 | |
| 4258 | ORF1a | 1332 | 1 | 0.004885 | 0.004306 | 0.028014 | |
| 17478 | ORF1b | 1337 | 2 | 1.57E-05 | 0.666667 | 0.000434 | |
| 17574 | ORF1b | 1369 | 1 | 0.00067 | 0.666667 | 0.006318 | |
| 4420 | ORF1a | 1386 | 3 | 0.000824 | 0.666667 | 0.007185 | |
| 4510 | ORF1a | 1416 | 1 | 0.000361 | 0.666667 | 0.004027 | |
| 17781 | ORF1b | 1438 | 1 | 0.004214 | 0.666667 | 0.02495 | |
| 4630 | ORF1a | 1456 | 1 | 0.000307 | 0.666667 | 0.003799 | |
| 18027 | ORF1b | 1520 | 4 | 9.97E-05 | 6.93E-05 | 0.001814 | |
| 18030 | ORF1b | 1521 | 2 | 0.045872 | 0.013034 | 0.164494 | |
| 4897 | ORF1a | 1545 | 3 | 0.039089 | 0.050136 | 0.146786 | |
| 18258 | ORF1b | 1597 | 3 | 0.007813 | 0.666667 | 0.040103 | |
| 18261 | ORF1b | 1598 | 2 | 0.000702 | 0.666667 | 0.006336 | |
| 18267 | ORF1b | 1600 | 3 | 0.002368 | 0.000714 | 0.01562 | |
| 5119 | ORF1a | 1619 | 1 | 0.027496 | 0.666667 | 0.11112 | |
| 5122 | ORF1a | 1620 | 1 | 0.005938 | 0.666667 | 0.03258 | |
| 5269 | ORF1a | 1669 | 2 | 0.012808 | 0.583413 | 0.05886 | |
| 5275 | ORF1a | 1671 | 4 | 0.000186 | 0.423608 | 0.002749 | |
| 5278 | ORF1a | 1672 | 2 | 0.001006 | 0.000388 | 0.008313 | |
| 5281 | ORF1a | 1673 | 4 | 0.009245 | 0.215603 | 0.045795 | |
| 5284 | ORF1a | 1674 | 2 | 0.010039 | 0.177374 | 0.049342 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 18549 | ORF1b | 1694 | 1 | 0.000471 | 0.666667 | 0.004996 | |
| 18555 | ORF1b | 1696 | 2 | 0.001122 | 0.666667 | 0.009156 | |
| 5671 | ORF1a | 1803 | 8 | 0.026607 | 0.109787 | 0.108911 | |
| 5728 | ORF1a | 1822 | 7 | 0.018549 | 0.088453 | 0.079504 | |
| 18990 | ORF1b | 1841 | 1 | 0.000277 | 0.666667 | 0.003603 | |
| 19005 | ORF1b | 1846 | 2 | 0.001943 | 0.666667 | 0.013444 | |
| 19089 | ORF1b | 1874 | 1 | 0.000122 | 0.666667 | 0.002131 | |
| 19095 | ORF1b | 1876 | 2 | 0.002052 | 0.666667 | 0.013966 | |
| 5923 | ORF1a | 1887 | 10 | 0.004061 | 0.002338 | 0.024502 | |
| 19194 | ORF1b | 1909 | 1 | 0.000274 | 0.666667 | 0.003637 | |
| 19257 | ORF1b | 1930 | 1 | 0.000242 | 0.666667 | 0.00338 | |
| 6055 | ORF1a | 1931 | 1 | 8.97E-06 | 8.61E-06 | 0.000301 | |
| 19281 | ORF1b | 1938 | 2 | 0.006731 | 0.005997 | 0.03585 | |
| 6115 | ORF1a | 1951 | 1 | 3.52E-06 | 0.666667 | 0.000166 | |
| 19320 | ORF1b | 1951 | 2 | 1.76E-05 | 0.666667 | 0.000439 | |
| 19323 | ORF1b | 1952 | 2 | 0.020375 | 0.666667 | 0.085887 | |
| 19332 | ORF1b | 1955 | 4 | 0.000166 | 6.25E-05 | 0.002608 | |
| 19338 | ORF1b | 1957 | 2 | 0.00049 | 0.00073 | 0.005151 | |
| 6133 | ORF1a | 1957 | 1 | 0.027119 | 0.666667 | 0.110648 | |
| 6136 | ORF1a | 1958 | 1 | 0.000329 | 0.666667 | 0.003845 | |
| 6169 | ORF1a | 1969 | 1 | 0.000329 | 0.666667 | 0.003876 | |
| 19545 | ORF1b | 2026 | 2 | 0.011373 | 0.666667 | 0.054429 | |
| 19548 | ORF1b | 2027 | 5 | 2.09E-05 | 0.208443 | 0.000511 | |
| 6349 | ORF1a | 2029 | 2 | 6.11E-05 | 0.666667 | 0.001196 | |
| 19557 | ORF1b | 2030 | 1 | 0.004496 | 0.666667 | 0.026132 | |
| 6367 | ORF1a | 2035 | 1 | 3.62E-05 | 1.09E-05 | 0.000756 | |
| 6379 | ORF1a | 2039 | 2 | 0.000686 | 0.666667 | 0.006329 | |
| 6400 | ORF1a | 2046 | 12 | 0.005974 | 0.008641 | 0.032641 | |
| 6460 | ORF1a | 2066 | 2 | 3.89E-06 | 1.39E-06 | 0.000171 | |
| 6475 | ORF1a | 2071 | 1 | 0.005981 | 0.666667 | 0.032538 | |
| 6511 | ORF1a | 2083 | 4 | 0.000549 | 0.003735 | 0.005421 | |
| 6562 | ORF1a | 2100 | 1 | 0.004435 | 0.00227 | 0.02602 | |
| 6589 | ORF1a | 2109 | 1 | 0.001766 | 0.666667 | 0.012629 | |
| 6739 | ORF1a | 2159 | 1 | 5.52E-06 | 0.666667 | 0.000219 | |
| 6742 | ORF1a | 2160 | 1 | 0.00052 | 0.666667 | 0.005334 | |
| 6802 | ORF1a | 2180 | 1 | 0.000191 | 0.666667 | 0.002793 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 6817 | ORF1a | 2185 | 1 | 0.000172 | 0.666667 | 0.0026 | |
| 20022 | ORF1b | 2185 | 3 | 0.006914 | 0.666667 | 0.03667 | |
| 6883 | ORF1a | 2207 | 2 | 0.00074 | 0.001363 | 0.006635 | |
| 6886 | ORF1a | 2208 | 2 | 0.019377 | 0.666667 | 0.082223 | |
| 6997 | ORF1a | 2245 | 1 | 0.001779 | 0.666667 | 0.012651 | |
| 7003 | ORF1a | 2247 | 1 | 0.009594 | 0.666667 | 0.047338 | |
| 20277 | ORF1b | 2270 | 1 | 2.67E-05 | 0.666667 | 0.000608 | |
| 20283 | ORF1b | 2272 | 1 | 0.008112 | 0.007402 | 0.041305 | |
| 20301 | ORF1b | 2278 | 1 | 7.03E-05 | 0.666667 | 0.001356 | |
| 20385 | ORF1b | 2306 | 1 | 0.0084 | 0.666667 | 0.0421 | |
| 7186 | ORF1a | 2308 | 1 | 0.000669 | 0.666667 | 0.006355 | |
| 20445 | ORF1b | 2326 | 1 | 0.006341 | 0.666667 | 0.034203 | |
| 7333 | ORF1a | 2357 | 2 | 0.000219 | 5.11E-05 | 0.003164 | |
| 20622 | ORF1b | 2385 | 3 | 0.001966 | 0.666667 | 0.013456 | |
| 7579 | ORF1a | 2439 | 1 | 0.028218 | 0.666667 | 0.112961 | |
| 7582 | ORF1a | 2440 | 1 | 0.012728 | 0.666667 | 0.058705 | |
| 21168 | ORF1b | 2567 | 2 | 0.040317 | 0.035717 | 0.148763 | |
| 21276 | ORF1b | 2603 | 1 | 0.001259 | 0.666667 | 0.009957 | |
| 8077 | ORF1a | 2605 | 2 | 3.54E-05 | 0.666667 | 0.000751 | |
| 8080 | ORF1a | 2606 | 1 | 0.000231 | 0.666667 | 0.003269 | |
| 8083 | ORF1a | 2607 | 3 | 0.023362 | 0.666667 | 0.096873 | |
| 8086 | ORF1a | 2608 | 1 | 0.010977 | 0.666667 | 0.052731 | |
| 8101 | ORF1a | 2613 | 1 | 0.011703 | 0.666667 | 0.055176 | |
| 21534 | ORF1b | 2689 | 1 | 0.030048 | 0.029845 | 0.118793 | |
| 21543 | ORF1b | 2692 | 2 | 0.001176 | 0.014442 | 0.009476 | |
| 21546 | ORF1b | 2693 | 1 | 0.000684 | 0.666667 | 0.00636 | |
| 8449 | ORF1a | 2729 | 1 | 3.21E-05 | 0.666667 | 0.000692 | |
| 8452 | ORF1a | 2730 | 1 | 1.59E-05 | 0.666667 | 0.000431 | |
| 8455 | ORF1a | 2731 | 1 | 0.012151 | 0.666667 | 0.056868 | |
| 8554 | ORF1a | 2764 | 2 | 0.000134 | 0.666667 | 0.00222 | |
| 8557 | ORF1a | 2765 | 1 | 0.004055 | 0.666667 | 0.024581 | |
| 8614 | ORF1a | 2784 | 2 | 1.68E-05 | 0.666667 | 0.000437 | |
| 8653 | ORF1a | 2797 | 4 | 0.004313 | 0.666667 | 0.025416 | |
| 8656 | ORF1a | 2798 | 2 | 0.002419 | 0.166279 | 0.015871 | |
| 8920 | ORF1a | 2886 | 2 | 0.00141 | 0.666667 | 0.010877 | |
| 9934 | ORF1a | 3224 | 3 | 0.000547 | 0.666667 | 0.005441 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 9937 | ORF1a | 3225 | 1 | 0.016766 | 0.666667 | 0.073092 | |
| 9940 | ORF1a | 3226 | 1 | 0.033013 | 0.666667 | 0.128913 | |
| 10054 | ORF1a | 3264 | 1 | 0.005691 | 0.666667 | 0.03136 | |
| 10618 | ORF1a | 3452 | 1 | 0.00156 | 0.666667 | 0.011753 | |
| 10621 | ORF1a | 3453 | 3 | 0.037569 | 0.14509 | 0.143619 | |
| 10630 | ORF1a | 3456 | 1 | 0.038032 | 0.008006 | 0.14452 | |
| 10633 | ORF1a | 3457 | 4 | 0.002512 | 0.000787 | 0.016313 | |
| 10891 | ORF1a | 3543 | 3 | 0.041422 | 0.029466 | 0.152399 | |
| 10987 | ORF1a | 3575 | 2 | 0.038846 | 0.666667 | 0.146739 | |
| 11080 | ORF1a | 3606 | 5 | 0.022712 | 0.008496 | 0.094793 | |
| 11176 | ORF1a | 3638 | 1 | 1.05E-05 | 0.666667 | 0.000342 | |
| 11281 | ORF1a | 3673 | 1 | 0.030057 | 0.030085 | 0.11846 | |
| 11284 | ORF1a | 3674 | 1 | 0.002782 | 0.000645 | 0.01762 | |
| 11287 | ORF1a | 3675 | 0 | 0.00064 | 0.000132 | 0.006124 | |
| 11290 | ORF1a | 3676 | 2 | 0.013009 | 0.046357 | 0.059356 | |
| 11635 | ORF1a | 3791 | 1 | 0.043122 | 0.039802 | 0.157291 | |
| 11833 | ORF1a | 3857 | 1 | 0.006485 | 0.666667 | 0.034686 | |
| 12340 | ORF1a | 4026 | 1 | 2.31E-05 | 0.666667 | 0.000536 | |
| 12706 | ORF1a | 4148 | 1 | 9.82E-05 | 0.666667 | 0.001811 | |
| 13066 | ORF1a | 4268 | 2 | 0.003095 | 0.666667 | 0.019312 | |
| 13345 | ORF1a | 4361 | 1 | 0.039414 | 0.666667 | 0.14757 | |
| 13441 | ORF1a | 4393 | 2 | 1.16E-06 | 0.666667 | 7.35E-05 | |
| 13444 | ORF1a | 4394 | 3 | 1.70E-05 | 0.666667 | 0.000432 | |
| 13447 | ORF1a | 4395 | 1 | 0.013224 | 0.666667 | 0.059908 | |
| 13453 | ORF1a | 4397 | 1 | 0.0005 | 0.666667 | 0.005215 | |

D.S3 Table. List of sites found to be selected differentially along internal branches between BA.1 and background sequences (FDR ≤ 0.2) using the Contrast-FEL method. Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome. Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. Ratio of omega (BA.1: reference): the ratio of site-level omega estimates for the two sets of branches. If this ratio is > 1, then selection on BA.1 is stronger. These values are highly imprecise and should be viewed as qualitative measures. p-value: the p-value for the likelihood ratio test that omega ratios between the internal branches of the two clades are different. This is not corrected for multiple testing. q-value: multiple-test corrected q-value (Benjamini-Hochberg).

| **Coordinate (SARS-CoV-2)** | **Gene/ORF** | **Codon (in gene/ORF)** | **Ratio of omega; (BA.1-01022022 : reference)** | **p** | **q** |
|---|---|---|---|---|---|
| 286 | ORF1a | 8 | NaN | 0.009257 | 0.833127 |
| 712 | ORF1a | 150 | NaN | 0.035673 | 0.917317 |
| 1189 | ORF1a | 309 | NaN | 0.003527 | 1 |
| 11275 | ORF1a | 3671 | NaN | 0.022739 | 1 |
| 11290 | ORF1a | 3676 | NaN | 0.012339 | 0.894541 |
| 12169 | ORF1a | 3969 | NaN | 0.035978 | 1 |
| 16974 | ORF1b | 1169 | NaN | 0.004807 | 1 |
| 21814 | S | 85 | NaN | 0.002092 | 0.078337 |
| 21823 | S | 88 | NaN | 0.002741 | 0.094314 |
| 21832 | S | 91 | NaN | 0.012487 | 0.244555 |
| 21901 | S | 114 | NaN | 0.031158 | 0.374192 |
| 21973 | S | 138 | NaN | 0.009673 | 0.195453 |
| 22066 | S | 169 | NaN | 0.007004 | 0.174833 |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | Ratio of omega; (BA.1-01022022 : reference) | p | q |
|---|---|---|---|---|---|
| 22198 | S | 213 | NaN | 0.017001 | 0.313651 |
| 22591 | S | 344 | NaN | 0.021268 | 0.297513 |
| 23116 | S | 519 | NaN | 0.034775 | 0.413725 |
| 23584 | S | 675 | NaN | 0.000996 | 0.04372 |
| 24925 | S | 1122 | NaN | 0.019468 | 0.339492 |
| 25695 | ORF3a | 102 | NaN | 0.031753 | 0.970229 |
| 26061 | ORF3a | 224 | NaN | 0.019072 | 1 |
| 26457 | E | 72 | NaN | 0.028905 | 1 |

D.S4 Table. List of sites found to be evolving under directional selection in the entire tree, using a FUBAR-like implementation of the DEPS [35] method. The BA.1 tree was rooted on the genome reference sequence for this analysis. Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. Target amino-acid: which amino-acids have statistical support (Bayes Factor ≥ 100) for accelerated evolution towards them.

| **Coordinate (SARS-CoV-2)** | **Gene/ORF** | **Codon (in gene/ORF)** | **Target amino-acids** |
|---|---|---|---|
| 1384 | ORF1a | 374 | Y |
| 3094 | ORF1a | 944 | L |
| 4180 | ORF1a | 1306 | S |
| 5923 | ORF1a | 1887 | I |
| 6400 | ORF1a | 2046 | L |
| 6514 | ORF1a | 2084 | I |
| 7726 | ORF1a | 2488 | F |
| 10027 | ORF1a | 3255 | I |
| 15520 | ORF1b | 685 | Y |
| 21574 | S | 5 | F |
| 21616 | S | 19 | R |
| 22027 | S | 156 | G |
| 22033 | S | 158 | G |
| 22090 | S | 177 | V |
| 22138 | S | 193 | L |
| 22195 | S | 212 | L |
| 22597 | S | 346 | K |
| 22897 | S | 446 | S |
| 22990 | S | 477 | N |
| 22993 | S | 478 | K |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | Target amino-acids |
|---|---|---|---|
| 23011 | S | 484 | A |
| 23047 | S | 496 | S |
| 23116 | S | 519 | L |
| 23200 | S | 547 | T |
| 23443 | S | 628 | K |
| 23596 | S | 679 | N |
| 23662 | S | 701 | V |
| 24409 | S | 950 | N |
| 24802 | S | 1081 | V |
| 25611 | ORF3a | 74 | F |
| 27381 | ORF6 | 61 | D |

D.S5 Table. Pairs of sites (BA.1) found to have epistatic (co-evolving) substitution patterns by BGM method. Gene/ORF: which gene or ORF does this site belong to. Codon 1/2 (in gene/ORF): the location of the two interacting codons in the corresponding Gene/ORF. Posterior probability of non-independence: estimated posterior probability that substitutions which occur on the interior branches of the BA.1 clade are not independent.

| **Gene/ORF** | **Codon 1 (in gene/ORF)** | **Codon 2 (in gene/ORF)** | **Posterior probability of non-independence** |
|---|---|---|---|
| ORF1b | 2183 | 2254 | 0.963173 |
| ORF1b | 1600 | 1601 | 0.724754 |
| ORF1b | 1600 | 1602 | 0.598017 |
| ORF1b | 1601 | 1602 | 0.636395 |
| ORF1b | 1913 | 1914 | 0.901897 |
| ORF1b | 1913 | 1916 | 0.515523 |
| ORF1b | 1914 | 1916 | 0.516807 |
| ORF1b | 1938 | 1953 | 0.914167 |
| ORF1b | 1955 | 1960 | 0.922513 |
| ORF1b | 1520 | 1522 | 0.552822 |
| ORF1b | 1522 | 1523 | 0.874259 |
| ORF1a | 148 | 149 | 0.755009 |
| M | 36 | 37 | 0.864317 |
| M | 36 | 38 | 0.869714 |
| M | 136 | 187 | 0.543369 |
| M | 140 | 170 | 0.893662 |
| M | 165 | 170 | 0.767712 |
| M | 165 | 171 | 0.713331 |
| M | 205 | 207 | 0.92213 |
| M | 206 | 209 | 0.705248 |
| M | 206 | 210 | 0.627488 |
| M | 209 | 210 | 0.626832 |
| M | 219 | 222 | 0.819452 |
| ORF1a | 909 | 910 | 0.876319 |
| ORF1a | 971 | 2105 | 0.755581 |

| Gene/ORF | Codon 1 (in gene/ORF) | Codon 2 (in gene/ORF) | Posterior probability of non-independence |
|---|---|---|---|
| ORF1a | 1004 | 2153 | 0.878498 |
| ORF1a | 1140 | 1141 | 0.887164 |
| ORF1a | 1361 | 2255 | 0.879833 |
| ORF1a | 1667 | 1668 | 0.604395 |
| ORF1a | 1667 | 1670 | 0.57555 |
| ORF1a | 1672 | 1673 | 0.730503 |
| ORF1a | 1813 | 1814 | 0.880839 |
| ORF1a | 1852 | 1929 | 0.598971 |
| ORF1a | 1852 | 1931 | 0.633561 |
| ORF1a | 1920 | 1921 | 0.883353 |
| ORF1a | 1929 | 1931 | 0.620828 |
| ORF1a | 2035 | 2036 | 0.749219 |
| ORF1a | 2060 | 2061 | 0.880444 |
| ORF1a | 2287 | 2307 | 0.884525 |
| ORF1a | 2881 | 2908 | 0.961669 |
| ORF1a | 3674 | 3675 | 0.50831 |
| ORF1a | 3675 | 3676 | 0.572833 |
| ORF6 | 58 | 59 | 0.689751 |
| ORF7a | 1 | 3 | 0.622113 |
| ORF7a | 1 | 6 | 0.593798 |
| ORF7a | 3 | 6 | 0.599717 |
| ORF7a | 54 | 61 | 0.86735 |
| ORF7a | 119 | 121 | 0.874435 |
| S | 30 | 191 | 0.881577 |
| S | 81 | 500 | 0.58242 |
| S | 81 | 502 | 0.572048 |
| S | 85 | 114 | 0.680537 |
| S | 88 | 93 | 0.93889 |
| S | 89 | 98 | 0.997748 |
| S | 90 | 108 | 0.720441 |
| S | 92 | 101 | 0.674095 |
| S | 99 | 109 | 0.514272 |
| S | 119 | 120 | 0.501016 |
| S | 119 | 121 | 0.512053 |
| S | 167 | 168 | 0.582037 |
| S | 167 | 170 | 0.637733 |

| Gene/ORF | Codon 1 (in gene/ORF) | Codon 2 (in gene/ORF) | Posterior probability of non-independence |
|---|---|---|---|
| S | 168 | 170 | 0.621658 |
| S | 172 | 173 | 0.529886 |
| S | 177 | 517 | 0.567614 |
| S | 211 | 212 | 0.509865 |
| S | 236 | 240 | 0.748729 |
| S | 238 | 239 | 0.615233 |
| S | 320 | 628 | 0.880251 |
| S | 405 | 408 | 0.781827 |
| S | 406 | 523 | 0.890215 |
| S | 410 | 411 | 0.885203 |
| S | 452 | 478 | 0.748739 |
| S | 496 | 498 | 0.655958 |
| S | 500 | 502 | 0.577658 |
| S | 517 | 519 | 0.708631 |
| S | 679 | 681 | 0.998578 |
| S | 684 | 1199 | 0.885016 |
| S | 721 | 722 | 0.883027 |
| S | 777 | 996 | 0.886557 |
| S | 849 | 939 | 0.877465 |
| S | 1251 | 1252 | 0.885291 |

D.S6 Table. List of sites found to be under pervasive negative selection by FEL ($p \leq 0.05$) along internal branches in BA.1. Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome. Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. Synonymous rate: Site estimate for the synonymous substitution rate (alpha). These values are highly imprecise and should be viewed as qualitative measures. Non-synonymous rate: Site estimate for the non-synonymous substitution rate (beta). These values are highly imprecise and should be viewed as qualitative measures. p-value: the p-value for the likelihood ratio test that beta / alpha $\neq$ 1. q-value: multiple-test corrected q-value (Benjamini-Hochberg).

| **Coordinate (SARS-CoV-2)** | **Gene/ORF** | **Codon (in gene/ORF)** | **Synonymous rate** | **Non-synonymous rate** | **p-value** | **q-value** |
|---|---|---|---|---|---|---|
| 21562 | S | 1 | 10000 | 0.185601 | 3.62E-05 | 0.015367 |
| 25392 | ORF3a | 1 | 463.998 | 7.365598 | 0.000298 | 0.094947 |
| 27201 | ORF6 | 1 | 33.93005 | 1.87503 | 0.021585 | 1 |
| 27899 | ORF8 | 3 | 2.321653 | 0 | 0.02104 | 1 |
| 27902 | ORF8 | 4 | 3.246806 | 0 | 0.024544 | 1 |
| 27402 | ORF7a | 4 | 1.685688 | 0 | 0.02754 | 1 |
| 21592 | S | 11 | 31.64784 | 0 | 3.10E-07 | 0.000394 |
| 27255 | ORF6 | 19 | 55.92456 | 0.954167 | 0.022489 | 1 |
| 27279 | ORF6 | 27 | 25.50938 | 0 | 0.010421 | 0.82913 |
| 26642 | M | 41 | 3.084009 | 0 | 0.041997 | 1 |
| 25524 | ORF3a | 45 | 186.4884 | 0 | 0.005221 | 0.511259 |
| 26678 | M | 53 | 3.683198 | 0 | 0.014247 | 1 |
| 21733 | S | 58 | 4.282356 | 0 | 0.047224 | 1 |
| 27378 | ORF6 | 60 | 2.683319 | 0 | 0.004557 | 0.527331 |
| 27591 | ORF7a | 67 | 2.607672 | 0 | 0.046379 | 1 |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | Synonymous rate | Non-synonymous rate | p-value | q-value |
|---|---|---|---|---|---|---|
| 21763 | S | 68 | 14.56174 | 2.300288 | 0.017517 | 1 |
| 27600 | ORF7a | 70 | 1.919624 | 0 | 0.045943 | 1 |
| 26744 | M | 75 | 249.9906 | 0.874662 | 0.000725 | 0.153844 |
| 26756 | M | 79 | 3.653489 | 0 | 0.030511 | 1 |
| 21811 | S | 84 | 7.910205 | 0 | 0.026215 | 1 |
| 27684 | ORF7a | 98 | 2.231685 | 0 | 0.046926 | 1 |
| 21871 | S | 104 | 217.6417 | 3.732462 | 0.023821 | 1 |
| 26846 | M | 109 | 162.9927 | 0.718412 | 0.005075 | 0.538423 |
| 28235 | ORF8 | 115 | 2.1218 | 0 | 0.035736 | 1 |
| 28250 | ORF8 | 120 | 1.304367 | 0.067552 | 0.00815 | 0.691659 |
| 25773 | ORF3a | 128 | 93.36818 | 0 | 0.01077 | 0.80651 |
| 25782 | ORF3a | 131 | 104.5756 | 3.045737 | 0.020346 | 1 |
| 26924 | M | 135 | 5.083786 | 0 | 0.023029 | 1 |
| 22018 | S | 153 | 214.7278 | 0.453261 | 0.018243 | 1 |
| 27044 | M | 175 | 2.831371 | 0 | 0.049442 | 1 |
| 25968 | ORF3a | 193 | 327.1658 | 0 | 0.002536 | 0.403509 |
| 14164 | ORF1b | 233 | 10000 | 0 | 0.000472 | 0.120284 |
| 22567 | S | 336 | 5.088154 | 0 | 0.035761 | 1 |
| 23188 | S | 543 | 4.20759 | 0 | 0.048863 | 1 |
| 24259 | S | 900 | 8925.374 | 1.395566 | 0.000782 | 0.142258 |
| 3034 | ORF1a | 924 | 30.27936 | 0 | 0.015346 | 1 |
| 24646 | S | 1029 | 10000 | 0 | 0.005877 | 0.534414 |
| 24997 | S | 1146 | 28.73682 | 0 | 5.47E-06 | 0.003482 |
| 25270 | S | 1237 | 805.0741 | 1.390289 | 0.002998 | 0.424073 |
| 18981 | ORF1b | 1838 | 1284.942 | 2.185727 | 0.003248 | 0.413468 |
| 19836 | ORF1b | 2123 | 10.74522 | 0 | 0.036059 | 1 |
| 20160 | ORF1b | 2231 | 11.72924 | 0 | 0.045801 | 1 |
| 20997 | ORF1b | 2510 | 9.269929 | 0 | 0.034442 | 1 |
| 8080 | ORF1a | 2606 | 142.8995 | 0.386153 | 0.016421 | 1 |
| 21540 | ORF1b | 2691 | 6.954534 | 0 | 0.041656 | 1 |
| 12025 | ORF1a | 3921 | 24.60216 | 0 | 0.019187 | 1 |
| 12253 | ORF1a | 3997 | 359.2106 | 3.65119 | 0.016763 | 1 |

D.S7 Table. BUSTED[S] selection results on the BA.1 SARS-CoV-2 clade across segments. Segment corresponds to the gene or ORF does under analysis. Omega1 refers to the first omega rate class, p1 refers to proportion of sites which fit this rate class. Omega2 refers to the first omega rate class, p2 refers to proportion of sites which fit this rate class. Omega3 refers to the first omega rate class which captures the episodic diversifying features, p3 refers to proportion of sites which fit this rate class. P-value, the p-value for the likelihood ratio test. Q-value refers to the multiple-test corrected q-value (Benjamini-Hochberg).

| **Segment** | **O1** | **p1** | **O2** | **p2** | **O3** | **p3** | **p-value** | **q-value** |
|---|---|---|---|---|---|---|---|---|
| exonuclease | 0 | 0.977627 | 0 | 0.013168 | 67.36418 | 0.009205 | 0 | 0 |
| M | 0 | 0.809655 | 0 | 0.11217 | 8.956224 | 0.078175 | 0 | 0 |
| nsp3 | 0 | 0.042594 | 0.509816 | 0.956992 | 1548.363 | 0.000414 | 0 | 0 |
| ORF3a | 0.03361 | 0.921932 | 0.039946 | 0.025221 | 20.52543 | 0.052847 | 0 | 0 |
| RdRp | 0.039359 | 0.159409 | 0.154944 | 0.836359 | 97.84895 | 0.004232 | 0 | 0 |
| S | 0.748525 | 0.989975 | 1 | 0.008125 | 368.1641 | 0.0019 | 0 | 0 |
| nsp4 | 0 | 0.896621 | 0 | 0.0814 | 25.37751 | 0.021979 | 2.78E-16 | 6.74E-16 |
| nsp2 | 0 | 0.583828 | 0 | 0.374363 | 20.50532 | 0.041809 | 2.68E-13 | 5.69E-13 |
| nsp6 | 0.021029 | 0.925471 | 0.161762 | 0.048648 | 21.76419 | 0.025881 | 3.79E-10 | 7.15E-10 |
| helicase | 0 | 0.716279 | 0 | 0.261369 | 21.2443 | 0.022352 | 1.65E-09 | 2.80E-09 |
| endornase | 0.02557 | 0.018136 | 0.212065 | 0.973433 | 41.591 | 0.008432 | 4.01E-09 | 6.20E-09 |
| leader | 0 | 0.633048 | 0 | 0.285303 | 13.93084 | 0.081648 | 1.27E-06 | 1.79E-06 |
| ORF7a | 0.005064 | 0.901847 | 0.005064 | 0.02768 | 11.595 | 0.070473 | 1.19E-05 | 1.56E-05 |

| Segment | O1 | p1 | O2 | p2 | O3 | p3 | p-value | q-value |
|---|---|---|---|---|---|---|---|---|
| nsp9 | 0.37549 | 0.935559 | 0.58565 | 0.064351 | 2362.953 | 9.04E-05 | 0.000113 | 0.000137 |
| ORF6 | 2.13E-05 | 0.169278 | 0.001357 | 0.736372 | 10.43698 | 0.09435 | 0.000354 | 0.000401 |
| nsp8 | 0.017784 | 0.21553 | 0.230485 | 0.754762 | 8.288697 | 0.029708 | 0.208269 | 0.221286 |
| nsp7 | 0.443394 | 0.831769 | 0.541215 | 0.168231 | 3.306423 | 0 | 0.5 | 0.5 |

D.S8 Table. List of sites found to be under diversifying positive selection by MEME ($p \leq 0.05$) along internal branches in BA.5, as well as biochemical properties that are important at this site (via the PRIME method). Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome. Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. % Of branches with omega>1: the fraction of tree branches (internal branches BA.1 clade) that have evidence of diversifying positive selection at this site (100%—pervasive selection, <100% –episodic selection). p-value: the p-value for the likelihood ratio test that a non-zero fraction of branches have omega > 1 (i.e., episodic diversifying selection at this branch). This is not corrected for multiple testing; the MEME test is generally conservative on real data. q-value: multiple-test corrected q-value (Benjamini-Hochberg). Properties: which, if any, of the five compositive biochemical properties are conserved or changed at this site.

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| 21562 | S | 1 | 1 | 0.000169 | 0.019526 | |
| 27201 | ORF6 | 1 | 1 | 0.010275 | 0.327002 | |
| 21568 | S | 3 | 2 | 0.00732 | 0.266223 | |
| 26528 | M | 3 | 4 | 0.014307 | 0.379428 | |
| 26268 | E | 9 | 4 | 0.02109 | 0.497166 | |
| 301 | leader | 13 | 1 | 0.048054 | 0.679694 | |
| 25431 | ORF3a | 14 | 1 | 0.005009 | 0.212552 | |
| 21616 | S | 19 | 2 | 0.00516 | 0.205285 | |
| 325 | leader | 21 | 1 | 0.038926 | 0.635289 | |
| 25458 | ORF3a | 23 | 3 | 0.019663 | 0.481359 | |
| 21640 | S | 27 | 3 | 0.003239 | 0.164939 | |
| 28369 | N | 33 | 0 | 0.014453 | 0.360751 | |
| 394 | leader | 44 | 1 | 0.030184 | 0.573491 | |
| 25536 | ORF3a | 49 | 2 | 0.039795 | 0.625426 | |
| 28043 | ORF8 | 51 | 1 | 0.048722 | 0.674158 | |
| 27354 | ORF6 | 52 | 1 | 0.019718 | 0.473601 | |
| 27357 | ORF6 | 53 | 1 | 0.039089 | 0.629883 | |
| 27375 | ORF6 | 59 | 1 | 0.002566 | 0.15554 | |
| 27381 | ORF6 | 61 | 7 | 0.031117 | 0.574087 | |
| 28471 | N | 67 | 1 | 0.026593 | 0.564209 | |
| 26445 | E | 68 | 2 | 0.004124 | 0.181031 | |
| 21766 | S | 69 | 1 | 0.00032 | 0.033954 | |
| 27609 | ORF7a | 73 | 2 | 0.00389 | 0.176836 | |
| 26460 | E | 73 | 0 | 0.028819 | 0.573231 | |
| 27624 | ORF7a | 78 | 2 | 0.010901 | 0.330392 | |
| 511 | leader | 83 | 3 | 0.001529 | 0.114489 | |
| 517 | leader | 85 | 1 | 0.000132 | 0.016772 | |
| 535 | leader | 91 | 2 | 0.026612 | 0.55536 | |
| 25713 | ORF3a | 108 | 2 | 0.014423 | 0.367221 | |
| 28250 | ORF8 | 120 | 3 | 0.00609 | 0.234934 | |
| 27753 | ORF7a | 121 | 1 | 0.003558 | 0.16775 | |
| 21985 | S | 142 | 2 | 0.030324 | 0.567684 | |
| 28723 | N | 151 | 1 | 0.026528 | 0.572376 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| 814 | nsp2 | 184 | 1 | 0.029086 | 0.569634 | |
| 28876 | N | 202 | 2 | 0.002951 | 0.163335 | |
| 883 | nsp2 | 207 | 1 | 0.048221 | 0.674566 | |
| 22198 | S | 213 | 2 | 0.000448 | 0.040759 | |
| 26058 | ORF3a | 223 | 2 | 0.013378 | 0.362344 | |
| 1168 | nsp2 | 302 | 1 | 0.049512 | 0.670525 | |
| 22576 | S | 339 | 5 | 1.58E-05 | 0.005044 | |
| 22597 | S | 346 | 2 | 0.028716 | 0.580253 | |
| 1312 | nsp2 | 350 | 1 | 0.043398 | 0.649955 | |
| 22672 | S | 371 | 6 | 1.56E-07 | 0.000199 | |
| 22675 | S | 372 | 1 | 4.51E-06 | 0.002873 | |
| 22678 | S | 373 | 3 | 0.000627 | 0.053186 | |
| 22684 | S | 375 | 2 | 0.02958 | 0.570545 | |
| 22687 | S | 376 | 3 | 0.000932 | 0.074181 | |
| 22774 | S | 405 | 3 | 0.011208 | 0.324271 | |
| 22783 | S | 408 | 0 | 0.03929 | 0.625209 | |
| 29509 | N | 413 | 2 | 0.023869 | 0.552454 | |
| 29518 | N | 416 | 1 | 0.000324 | 0.031693 | |
| 22810 | S | 417 | 1 | 0.001809 | 0.127956 | |
| 29524 | N | 418 | 1 | 0.00259 | 0.149895 | |
| 29527 | N | 419 | 2 | 2.69E-05 | 0.00489 | |
| 22879 | S | 440 | 6 | 1.78E-05 | 0.004523 | |
| 22915 | S | 452 | 3 | 0.000103 | 0.014594 | |
| 22990 | S | 477 | 2 | 0.001915 | 0.128275 | |
| 22993 | S | 478 | 3 | 0.005044 | 0.207128 | |
| 23011 | S | 484 | 2 | 0.034668 | 0.596378 | |
| 23017 | S | 486 | 2 | 0.011172 | 0.330734 | |
| 23053 | S | 498 | 2 | 0.00774 | 0.27368 | |
| 14959 | RdRp | 498 | 1 | 0.035422 | 0.601232 | |
| 23074 | S | 505 | 4 | 1.37E-05 | 0.0058 | |
| 1819 | nsp2 | 519 | 1 | 0.042066 | 0.637494 | |
| 15358 | RdRp | 631 | 1 | 0.024459 | 0.546256 | |
| 23524 | S | 655 | 2 | 0.010612 | 0.329492 | |
| 2230 | nsp2 | 656 | 1 | 0.046147 | 0.660053 | |
| 23533 | S | 658 | 1 | 0.024236 | 0.550931 | |
| 23596 | S | 679 | 3 | 0.00298 | 0.158085 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| 15520 | RdRp | 685 | 1 | 0.031567 | 0.56598 | |
| 23851 | S | 764 | 6 | 1.79E-05 | 0.003799 | |
| 23947 | S | 796 | 1 | 0.013333 | 0.368981 | |
| 2788 | nsp3 | 842 | 2 | 0.038116 | 0.630146 | |
| 24421 | S | 954 | 2 | 0.031708 | 0.560621 | |
| 24466 | S | 969 | 3 | 0.003245 | 0.158861 | |
| 16614 | helicase | 1049 | 2 | 0.002476 | 0.157582 | |
| 25045 | S | 1162 | 6 | 2.74E-05 | 0.004361 | |
| 16974 | helicase | 1169 | 1 | 0.040299 | 0.618074 | |
| 3889 | nsp3 | 1209 | 2 | 0.027994 | 0.574785 | |
| 17205 | helicase | 1246 | 1 | 0.040033 | 0.621492 | |
| 25333 | S | 1258 | 2 | 0.045533 | 0.666251 | |
| 17409 | helicase | 1314 | 1 | 0.048932 | 0.669788 | |
| 17676 | helicase | 1403 | 1 | 0.025045 | 0.549693 | |
| 17745 | helicase | 1426 | 3 | 0.012181 | 0.344587 | |
| 19974 | endornase | 2169 | 0 | 0.009098 | 0.304786 | |
| 6937 | nsp3 | 2225 | 1 | 0.006549 | 0.245218 | |
| 6994 | nsp3 | 2244 | 1 | 0.046019 | 0.665701 | |
| 21294 | methyltransferase | 2609 | 2 | 0.009392 | 0.306559 | |
| 21516 | methyltransferase | 2683 | 1 | 0.035554 | 0.595524 | |
| 21522 | methyltransferase | 2685 | 1 | 0.032396 | 0.564929 | |
| 21534 | methyltransferase | 2689 | 1 | 0.008337 | 0.286836 | |
| 9343 | | 3027 | 2 | 0.044409 | 0.657349 | |
| 12028 | nsp7 | 3922 | 2 | 0.014321 | 0.372045 | |
| 13462 | RdRp | 4400 | 1 | 0.031198 | 0.567366 | |

D.S9 Table. List of sites found to be under diversifying positive selection by MEME ($p \leq 0.05$) along all branches in BA.5, as well as biochemical properties that are important at this site (via the PRIME method). Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome. Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. % Of branches with omega>1: the fraction of tree branches (internal branches BA.1 clade) that have evidence of diversifying positive selection at this site (100%—pervasive selection, <100% –episodic selection). p-value: the p-value for the likelihood ratio test that a non-zero fraction of branches have omega > 1 (i.e., episodic diversifying selection at this branch). This is not corrected for multiple testing; the MEME test is generally conservative on real data. q-value: multiple-test corrected q-value (Benjamini-Hochberg). Properties: which, if any, of the five compositive biochemical properties are conserved or changed at this site.

| **Coordinate (SARS-CoV-2)** | **Gene/ORF** | **Codon (in gene/ORF)** | **# of selected branches** | **p-value** | **p-value/int** | **q-value** | **Properties** |
|---|---|---|---|---|---|---|---|
| 21562 | S | 1 | 1 | 7.55E-06 | 0.000169 | 0.003204 | |
| 25392 | ORF3a | 1 | 1 | 0.006289 | 0.666667 | 0.117729 | |
| 27201 | ORF6 | 1 | 1 | 0.012678 | 0.010275 | 0.189866 | |
| 25398 | ORF3a | 3 | 2 | 0.005659 | 0.666667 | 0.11257 | |
| 26528 | M | 3 | 1 | 0.018366 | 0.014307 | 0.233801 | |
| 25401 | ORF3a | 4 | 1 | 0.001146 | 0.666667 | 0.040528 | |
| 25404 | ORF3a | 5 | 2 | 0.02685 | 0.666667 | 0.313577 | |
| 21574 | S | 5 | 5 | 0.042636 | 0.666667 | 0.437703 | |
| 25407 | ORF3a | 6 | 1 | 0.018172 | 0.666667 | 0.23605 | |
| 26268 | E | 9 | 12 | 0.0129 | 0.02109 | 0.184517 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value/int | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 25431 | ORF3a | 14 | 5 | 0.002944 | 0.005009 | 0.078069 | |
| 21616 | S | 19 | 7 | 0.040306 | 0.00516 | 0.417146 | |
| 27450 | ORF7a | 20 | 2 | 0.004938 | 0.666667 | 0.106547 | |
| 27453 | ORF7a | 21 | 1 | 0.023153 | 0.666667 | 0.280698 | |
| 28369 | N | 33 | 1 | 0.003594 | 0.014453 | 0.087994 | |
| 27354 | ORF6 | 52 | 1 | 0.039394 | 0.019718 | 0.411057 | |
| 25554 | ORF3a | 55 | 5 | 0.028947 | 0.076926 | 0.331976 | |
| 27375 | ORF6 | 59 | 2 | 0.004601 | 0.002566 | 0.106495 | |
| 27570 | ORF7a | 60 | 4 | 0.001541 | 0.666667 | 0.053007 | |
| 28471 | N | 67 | 2 | 0.046673 | 0.026593 | 0.457037 | |
| 28094 | ORF8 | 68 | 3 | 0.017978 | 0.279123 | 0.235939 | |
| 26445 | E | 68 | 2 | 0.020882 | 0.004124 | 0.258086 | |
| 21766 | S | 69 | 0 | 0.009251 | 0.00032 | 0.14907 | |
| 26460 | E | 73 | 2 | 0.007646 | 0.028819 | 0.128065 | |
| 490 | ORF1a | 76 | 5 | 0.014883 | 0.666667 | 0.203715 | |
| 21787 | S | 76 | 4 | 0.046259 | 0.054279 | 0.456498 | |
| 517 | ORF1a | 85 | 1 | 1.37E-05 | 0.000132 | 0.002905 | |
| 535 | ORF1a | 91 | 6 | 0.03772 | 0.026612 | 0.403513 | |
| 544 | ORF1a | 94 | 4 | 0.01273 | 0.666667 | 0.188428 | |
| 27696 | ORF7a | 102 | 3 | 0.036435 | 0.333835 | 0.393063 | |
| 27708 | ORF7a | 106 | 2 | 0.010908 | 0.666667 | 0.169336 | |
| 28244 | ORF8 | 118 | 2 | 0.006889 | 0.320082 | 0.123518 | |
| 28247 | ORF8 | 119 | 4 | 3.93E-05 | 0.067939 | 0.004171 | |
| 28250 | ORF8 | 120 | 6 | 0.031603 | 0.00609 | 0.352897 | |
| 27753 | ORF7a | 121 | 2 | 0.012971 | 0.003558 | 0.183473 | |
| 26894 | M | 125 | 5 | 0.043476 | 0.666667 | 0.442764 | |
| 28672 | N | 134 | 1 | 0.000888 | 0.666667 | 0.035337 | |
| 28723 | N | 151 | 4 | 0.004964 | 0.026528 | 0.105324 | |
| 25854 | ORF3a | 155 | 3 | 0.045252 | 0.050578 | 0.453588 | |
| 28735 | N | 155 | 2 | 0.048077 | 0.666667 | 0.467189 | |
| 27038 | M | 173 | 2 | 6.23E-05 | 0.666667 | 0.006104 | |
| 814 | ORF1a | 184 | 2 | 0.044696 | 0.029086 | 0.451568 | |
| 895 | ORF1a | 211 | 1 | 0.0183 | 0.666667 | 0.235309 | |
| 28903 | N | 211 | 2 | 0.048092 | 0.666667 | 0.463799 | |
| 901 | ORF1a | 213 | 1 | 0.005562 | 0.666667 | 0.112387 | |
| 22198 | S | 213 | 6 | 0.005794 | 0.000448 | 0.111748 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value/int | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 904 | ORF1a | 214 | 1 | 0.004495 | 0.666667 | 0.105955 | |
| 910 | ORF1a | 216 | 1 | 0.002959 | 0.666667 | 0.076882 | |
| 27170 | M | 217 | 1 | 9E-05 | 0.666667 | 0.007637 | |
| 952 | ORF1a | 230 | 1 | 0.008969 | 0.666667 | 0.146378 | |
| 1033 | ORF1a | 257 | 1 | 0.029308 | 0.666667 | 0.333115 | |
| 1042 | ORF1a | 260 | 1 | 0.000721 | 0.666667 | 0.031644 | |
| 1111 | ORF1a | 283 | 1 | 0.006268 | 0.666667 | 0.119088 | |
| 22534 | S | 325 | 1 | 0.000191 | 0.666667 | 0.01155 | |
| 22576 | S | 339 | 8 | 0.004841 | 1.58E-05 | 0.106242 | |
| 22597 | S | 346 | 9 | 0.038932 | 0.028716 | 0.40959 | |
| 1309 | ORF1a | 349 | 1 | 0.000126 | 0.666667 | 0.008915 | |
| 22642 | S | 361 | 1 | 2.8E-06 | 0.666667 | 0.003565 | |
| 29353 | N | 361 | 2 | 0.006655 | 0.107086 | 0.12103 | |
| 22672 | S | 371 | 9 | 0.001108 | 1.56E-07 | 0.040293 | |
| 22675 | S | 372 | 1 | 5.43E-06 | 4.51E-06 | 0.003458 | |
| 22678 | S | 373 | 4 | 0.012293 | 0.000627 | 0.186291 | |
| 22684 | S | 375 | 4 | 0.007301 | 0.02958 | 0.127319 | |
| 22687 | S | 376 | 4 | 0.000583 | 0.000932 | 0.029702 | |
| 29467 | N | 399 | 3 | 0.028442 | 0.666667 | 0.329147 | |
| 14674 | ORF1b | 403 | 1 | 0.0055 | 0.666667 | 0.112933 | |
| 22783 | S | 408 | 0 | 0.007224 | 0.03929 | 0.127721 | |
| 29518 | N | 416 | 1 | 0.000171 | 0.000324 | 0.010876 | |
| 22810 | S | 417 | 5 | 0.01635 | 0.001809 | 0.221417 | |
| 29524 | N | 418 | 1 | 0.02195 | 0.00259 | 0.268676 | |
| 29527 | N | 419 | 2 | 3.11E-05 | 2.69E-05 | 0.0044 | |
| 1546 | ORF1a | 428 | 1 | 9.11E-06 | 0.666667 | 0.002319 | |
| 14752 | ORF1b | 429 | 1 | 0.004257 | 0.666667 | 0.102245 | |
| 22879 | S | 440 | 13 | 0.002459 | 1.78E-05 | 0.069567 | |
| 14791 | ORF1b | 442 | 1 | 0.012798 | 0.666667 | 0.187262 | |
| 22915 | S | 452 | 10 | 3.26E-05 | 0.000103 | 0.004151 | |
| 23017 | S | 486 | 8 | 0.001886 | 0.011172 | 0.06002 | |
| 23053 | S | 498 | 5 | 0.012832 | 0.00774 | 0.185625 | |
| 23068 | S | 503 | 1 | 0.000105 | 0.666667 | 0.008354 | |
| 23074 | S | 505 | 4 | 0.007387 | 1.37E-05 | 0.125377 | |
| 1891 | ORF1a | 543 | 1 | 0.004668 | 0.666667 | 0.104263 | |
| 15112 | ORF1b | 549 | 1 | 0.025418 | 0.666667 | 0.299605 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value/int | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 23293 | S | 578 | 1 | 0.020743 | 0.666667 | 0.258879 | |
| 23524 | S | 655 | 6 | 0.03105 | 0.010612 | 0.349795 | |
| 15520 | ORF1b | 685 | 2 | 0.024822 | 0.031567 | 0.295314 | |
| 15682 | ORF1b | 739 | 1 | 0.000889 | 0.666667 | 0.034304 | |
| 23851 | S | 764 | 0 | 0.034942 | 1.79E-05 | 0.383458 | |
| 2698 | ORF1a | 812 | 1 | 0.002891 | 0.666667 | 0.080013 | |
| 15970 | ORF1b | 835 | 1 | 0.001579 | 0.666667 | 0.052886 | |
| 24217 | S | 886 | 1 | 0.002169 | 0.666667 | 0.064209 | |
| 3181 | ORF1a | 973 | 2 | 0.005759 | 0.666667 | 0.112788 | |
| 3544 | ORF1a | 1094 | 2 | 2.87E-05 | 0.666667 | 0.004564 | |
| 24928 | S | 1123 | 2 | 0.006389 | 0.666667 | 0.117871 | |
| 16896 | ORF1b | 1143 | 3 | 0.000755 | 0.666667 | 0.032052 | |
| 25045 | S | 1162 | 10 | 0.011015 | 2.74E-05 | 0.168948 | |
| 3805 | ORF1a | 1181 | 1 | 0.000234 | 0.666667 | 0.013514 | |
| 3814 | ORF1a | 1184 | 1 | 1.39E-05 | 0.666667 | 0.002527 | |
| 3949 | ORF1a | 1229 | 1 | 0.00235 | 0.666667 | 0.067978 | |
| 4021 | ORF1a | 1253 | 1 | 0.000597 | 0.666667 | 0.029245 | |
| 4423 | ORF1a | 1387 | 1 | 0.016365 | 0.666667 | 0.219297 | |
| 4426 | ORF1a | 1388 | 1 | 0.003082 | 0.666667 | 0.078459 | |
| 4639 | ORF1a | 1459 | 1 | 0.000549 | 0.666667 | 0.029119 | |
| 4660 | ORF1a | 1466 | 1 | 0.002113 | 0.666667 | 0.064036 | |
| 18039 | ORF1b | 1524 | 2 | 7.81E-06 | 0.666667 | 0.002485 | |
| 18393 | ORF1b | 1642 | 6 | 0.048946 | 0.13215 | 0.468479 | |
| 19320 | ORF1b | 1951 | 1 | 0.010906 | 0.666667 | 0.171397 | |
| 19326 | ORF1b | 1953 | 2 | 7.31E-05 | 0.666667 | 0.006648 | |
| 6355 | ORF1a | 2031 | 1 | 0.003321 | 0.666667 | 0.082895 | |
| 19620 | ORF1b | 2051 | 1 | 0.000169 | 0.666667 | 0.011356 | |
| 20088 | ORF1b | 2207 | 4 | 0.045423 | 0.666667 | 0.451741 | |
| 6937 | ORF1a | 2225 | 1 | 0.00208 | 0.006549 | 0.064594 | |
| 7054 | ORF1a | 2264 | 1 | 0.000648 | 0.666667 | 0.029456 | |
| 7057 | ORF1a | 2265 | 2 | 0.002901 | 0.666667 | 0.078565 | |
| 20541 | ORF1b | 2358 | 1 | 0.019681 | 0.666667 | 0.248065 | |
| 7348 | ORF1a | 2362 | 1 | 3.64E-05 | 0.666667 | 0.004213 | |
| 20646 | ORF1b | 2393 | 1 | 0.000966 | 0.666667 | 0.036163 | |
| 20658 | ORF1b | 2397 | 2 | 0.004613 | 0.666667 | 0.104865 | |
| 20814 | ORF1b | 2449 | 1 | 0.001875 | 0.666667 | 0.061191 | |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | p-value/int | q-value | Properties |
|---|---|---|---|---|---|---|---|
| 20817 | ORF1b | 2450 | 1 | 0.014465 | 0.666667 | 0.200151 | |
| 21384 | ORF1b | 2639 | 1 | 0.035507 | 0.666667 | 0.386328 | |
| 21516 | ORF1b | 2683 | 1 | 0.024268 | 0.035554 | 0.291448 | |
| 21522 | ORF1b | 2685 | 2 | 0.033227 | 0.032396 | 0.36781 | |
| 21534 | ORF1b | 2689 | 2 | 0.009964 | 0.008337 | 0.158549 | |
| 21543 | ORF1b | 2692 | 1 | 0.000803 | 0.666667 | 0.032992 | |
| 8920 | ORF1a | 2886 | 2 | 0.007327 | 0.666667 | 0.126046 | |
| 10051 | ORF1a | 3263 | 1 | 0.000119 | 0.666667 | 0.008908 | |
| 10885 | ORF1a | 3541 | 1 | 0.017717 | 0.666667 | 0.234935 | |
| 11068 | ORF1a | 3602 | 2 | 0.005112 | 0.666667 | 0.106687 | |
| 12667 | ORF1a | 4135 | 2 | 0.013506 | 0.666667 | 0.188941 | |
| 12676 | ORF1a | 4138 | 1 | 0.038587 | 0.666667 | 0.409339 | |
| 12679 | ORF1a | 4139 | 1 | 0.007947 | 0.666667 | 0.131387 | |
| 13441 | ORF1a | 4393 | 1 | 0.00061 | 0.666667 | 0.028757 | |
| 13456 | ORF1a | 4398 | 1 | 0.000289 | 0.666667 | 0.016003 | |

D.S10 Table. List of sites found to be selected differentially along internal branches between BA.5 and background sequences (FDR ≤ 0.2) using the Contrast-FEL method. Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome. Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. Ratio of omega (BA.1: background): the ratio of site-level omega estimates for the two sets of branches. If this ratio is > 1, then selection on BA.1 is stronger. These values are highly imprecise and should be viewed as qualitative measures. p-value: the p-value for the likelihood ratio test that omega ratios between the internal branches of the two clades are different. This is not corrected for multiple testing. q-value: multiple-test corrected q-value (Benjamini-Hochberg).

| **Coordinate (SARS-CoV-2)** | **Gene/ORF** | **Codon (in gene/ORF)** | **Ratio of ω (BA.5-08092022 : reference)** | **p** | **q** |
|---|---|---|---|---|---|
| 910 | ORF1a | 216 | NaN | 0.005141 | 1 |
| 8656 | ORF1a | 2798 | NaN | 0.047495 | 1 |
| 8692 | ORF1a | 2810 | NaN | 0.047353 | 1 |
| 9751 | ORF1a | 3163 | NaN | 0.049352 | 1 |
| 9928 | ORF1a | 3222 | NaN | 0.001606 | 0.803102 |
| 18414 | ORF1b | 1649 | NaN | 0.003789 | 1 |
| 19689 | ORF1b | 2074 | NaN | 0.043348 | 0.937398 |
| 22204 | S | 215 | NaN | 0.004135 | 0.478538 |
| 22939 | S | 460 | NaN | 0.048163 | 1 |
| 23590 | S | 677 | NaN | 0.003931 | 0.50038 |
| 24466 | S | 969 | NaN | 0.019797 | 1 |
| 29353 | N | 361 | NaN | 0.011 | 1 |

D.S11 Table. List of sites found to be evolving under directional selection in the entire tree, using a FUBAR-like implementation of the DEPS method. The BA.5 tree was rooted on the genome reference sequence for this analysis. Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. Target amino-acid: which amino-acids have statistical support (Bayes Factor ≥ 100) for accelerated evolution towards them.

| **Coordinate (SARS-CoV-2)** | **Gene/ORF** | **Codon (in gene/ORF)** | **Target amino-acids** |
|---|---|---|---|
| 535 | ORF1a | 91 | K |
| 4183 | ORF1a | 1307 | S |
| 17676 | ORF1b | 1403 | M |
| 21574 | S | 5 | F |
| 22426 | S | 289 | I |
| 22783 | S | 408 | S |
| 23602 | S | 681 | P |
| 23671 | S | 704 | L |
| 25887 | ORF3a | 166 | L |
| 26268 | E | 9 | I |
| 28142 | ORF8 | 84 | L |

D.S12 Table. Pairs of sites (BA.5) found to have epistatic (co-evolving) substitution patterns by BGM method. Gene/ORF: which gene or ORF does this site belong to. Codon 1/2 (in gene/ORF): the location of the two interacting codons in the corresponding Gene/ORF. Posterior probability of non-independence: estimated posterior probability that substitutions which occur on the interior branches of the BA.1 clade are not independent.

| **Gene/ORF** | **Codon 1 (in gene/ORF)** | **Codon 2 (in gene/ORF)** | **Posterior probability of non-independence** |
|---|---|---|---|
| ORF1a | 15 | 16 | 0.946007 |
| ORF1b | 2683 | 2684 | 0.679627 |
| ORF1b | 2683 | 2685 | 0.611989 |
| ORF1b | 2684 | 2685 | 0.664562 |
| ORF1b | 2688 | 2689 | 0.964216 |
| N | 361 | 368 | 0.918809 |
| N | 413 | 418 | 0.510817 |
| N | 413 | 419 | 0.503234 |
| N | 418 | 419 | 0.803379 |
| ORF1a | 842 | 1307 | 0.842686 |
| ORF1a | 930 | 1209 | 0.832483 |
| ORF1a | 953 | 1693 | 0.923044 |
| ORF1a | 1102 | 1786 | 0.927006 |
| ORF1a | 1382 | 1452 | 0.928894 |
| ORF1a | 1428 | 2285 | 0.779728 |
| ORF1a | 2284 | 2285 | 0.79289 |
| ORF1a | 3675 | 3676 | 0.973533 |
| ORF1a | 3753 | 3758 | 0.932013 |
| ORF6 | 52 | 53 | 0.909923 |
| ORF6 | 58 | 59 | 0.773791 |
| ORF6 | 59 | 60 | 0.755288 |
| ORF7a | 54 | 83 | 0.877268 |
| ORF8 | 5 | 6 | 0.911477 |
| ORF8 | 11 | 51 | 0.760622 |
| ORF1b | 631 | 690 | 0.93087 |
| S | 2 | 3 | 0.835832 |
| S | 67 | 829 | 0.930761 |

| Gene/ORF | Codon 1 (in gene/ORF) | Codon 2 (in gene/ORF) | Posterior probability of non-independence |
|---|---|---|---|
| S | 69 | 452 | 0.594335 |
| S | 153 | 1258 | 0.659871 |
| S | 184 | 968 | 0.924075 |
| S | 372 | 373 | 0.921997 |
| S | 375 | 376 | 0.529954 |
| S | 440 | 452 | 0.834062 |
| S | 477 | 478 | 0.962286 |
| S | 498 | 501 | 0.633729 |
| S | 679 | 1128 | 0.615697 |
| S | 954 | 969 | 0.992706 |

D.S13 Table. List of sites found to be under pervasive negative selection by FEL ($p \leq 0.05$) along internal branches in BA.5. Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome. Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. Synonymous rate: Site estimate for the synonymous substitution rate (alpha). These values are highly imprecise and should be viewed as qualitative measures. Non-synonymous rate: Site estimate for the non-synonymous substitution rate (beta). These values are highly imprecise and should be viewed as qualitative measures. p-value: the p-value for the likelihood ratio test. "q-value": multiple-test corrected q-value (Benjamini-Hochberg).

| **Coordinate (SARS-CoV-2)** | **Gene/ORF** | **Codon (in gene/ORF)** | **Synonymous rate** | **Non-synonymous rate** | **p-value** | **q-value** |
|---|---|---|---|---|---|---|
| 28273 | N | 1 | 10000 | 0 | 6.3E-06 | 0.008014 |
| 27893 | ORF8 | 1 | 247.1457 | 1.3186 | 2.37E-05 | 0.015089 |
| 26522 | M | 1 | 559.3321 | 0 | 0.001515 | 0.642693 |
| 27393 | ORF7a | 1 | 152.2261 | 0 | 0.003268 | 1 |
| 27486 | ORF7a | 32 | 5.692068 | 0 | 0.046657 | 1 |
| 26678 | M | 53 | 6.057653 | 0 | 0.022252 | 1 |
| 26439 | E | 66 | 2.88519 | 0 | 0.040565 | 1 |
| 27603 | ORF7a | 71 | 3.70275 | 0 | 0.046734 | 1 |
| 27627 | ORF7a | 79 | 3.745723 | 0 | 0.022945 | 1 |
| 26846 | M | 109 | 90.93501 | 1.717621 | 0.024722 | 1 |
| 26855 | M | 112 | 6.794666 | 0 | 0.015218 | 1 |
| 27750 | ORF7a | 120 | 4.781123 | 0 | 0.02637 | 1 |
| 11848 | ORF1a | 3862 | 91.50479 | 2.364265 | 0.038229 | 1 |
| 12634 | ORF1a | 4124 | 223.8636 | 0 | 0.015438 | 1 |

D.S14 Table. List of sites found to be under diversifying positive selection by MEME ($p \leq 0.05$) along internal branches in B.1.621, as well as biochemical properties that are important at this site (via the PRIME method). Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome. Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. % of branches with omega > 1: the fraction of tree branches (internal branches BA.1 clade) that have evidence of diversifying positive selection at this site (100%—pervasive selection, <100% –episodic selection). p-value: the p-value for the likelihood ratio test that a non-zero fraction of branches have omega > 1 (i.e., episodic diversifying selection at this branch). This is not corrected for multiple testing; the MEME test is generally conservative on real data. q-value: multiple-test corrected q-value (Benjamini-Hochberg). Properties: which, if any, of the five compositive biochemical properties are conserved or changed at this site.

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| 27210 | ORF6 | 4 | 2 | 7.99E-06 | 0.000479 | [] |
| 13516 | ORF1b | 17 | 1 | 0.033615 | 0.168077 | [] |
| 21640 | S | 27 | 1 | 0.000824 | 0.024726 | [] |
| 21844 | S | 95 | 2 | 0.006738 | 0.086636 | [] |
| 25707 | ORF3a | 106 | 2 | 0.041085 | 0.145005 | [] |
| 28253 | ORF8 | 121 | 0 | 0.030038 | 0.180231 | [] |
| 28675 | N | 135 | 1 | 0.032275 | 0.170868 | [] |
| 21991 | S | 144 | 5 | 2.43E-06 | 0.000438 | [] |
| 21994 | S | 145 | 4 | 0.000198 | 0.008931 | [] |
| 21997 | S | 146 | 3 | 0.000599 | 0.021563 | [] |
| 22000 | S | 147 | 3 | 0.011012 | 0.1166 | [["Overall", 0.00980911 0281853162 ,null],["secondary",0.03 0792390885 97696,15]] |
| 22003 | S | 148 | 1 | 0.022124 | 0.165933 | [] |
| 25887 | ORF3a | 166 | 2 | 0.034565 | 0.163731 | [] |
| 28873 | N | 201 | 1 | 0.046029 | 0.138086 | [] |
| 14122 | ORF1b | 219 | 1 | 0.036867 | 0.170155 | [] |
| 29023 | N | 251 | 1 | 0.042523 | 0.141743 | [] |
| 26157 | ORF3a | 256 | 4 | 0.020334 | 0.159134 | [] |
| 26160 | ORF3a | 257 | 2 | 0.001986 | 0.051066 | [["charge",0 .000661388 0399832794 ,4.37365524 5819593]] |
| 14407 | ORF1b | 314 | 2 | 0.023961 | 0.172519 | [] |
| 14470 | ORF1b | 335 | 1 | 0.033901 | 0.164921 | [] |
| 14530 | ORF1b | 355 | 1 | 0.041647 | 0.144161 | [] |
| 29443 | N | 391 | 1 | 0.032574 | 0.167522 | [] |
| 22810 | S | 417 | 3 | 0.003099 | 0.055774 | [] |
| 1546 | ORF1a | 428 | 1 | 0.049605 | 0.133267 | [] |
| 14767 | ORF1b | 434 | 1 | 0.038462 | 0.161005 | [] |
| 14785 | ORF1b | 440 | 1 | 0.038325 | 0.168257 | [["charge",0 ,14.9663834 4453895]] |
| 23062 | S | 501 | 2 | 0.007021 | 0.084247 | [] |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| 15001 | ORF1b | 512 | 1 | 0.045799 | 0.139725 | [] |
| 1819 | ORF1a | 519 | 2 | 0.012283 | 0.110548 | [] |
| 23275 | S | 572 | 2 | 0.004679 | 0.070188 | [] |
| 15406 | ORF1b | 647 | 1 | 0.038318 | 0.17243 | [] |
| 2230 | ORF1a | 656 | 3 | 0.031613 | 0.172433 | [] |
| 16075 | ORF1b | 870 | 1 | 0.045212 | 0.140312 | [] |
| 2944 | ORF1a | 894 | 1 | 0.024669 | 0.158588 | [] |
| 16260 | ORF1b | 931 | 1 | 0.030981 | 0.17989 | [] |
| 24409 | S | 950 | 6 | 3.36E-06 | 0.000302 | [["Overall", 0.02493895 424621928, null]] |
| 16911 | ORF1b | 1148 | 1 | 0.041725 | 0.141708 | [] |
| 17025 | ORF1b | 1186 | 1 | 0.024148 | 0.167182 | [] |
| 25333 | S | 1258 | 3 | 0.008528 | 0.095937 | [] |
| 25336 | S | 1259 | 2 | 0.015739 | 0.134903 | [] |
| 17820 | ORF1b | 1451 | 2 | 0.039264 | 0.160624 | [] |
| 17976 | ORF1b | 1503 | 2 | 0.011316 | 0.107201 | [] |
| 18129 | ORF1b | 1554 | 1 | 0.039544 | 0.158177 | [] |
| 18327 | ORF1b | 1620 | 1 | 0.039596 | 0.15494 | [] |
| 18393 | ORF1b | 1642 | 1 | 0.046184 | 0.136281 | [] |
| 18411 | ORF1b | 1648 | 1 | 0.044393 | 0.140188 | [] |
| 18525 | ORF1b | 1686 | 1 | 0.047999 | 0.132921 | [] |
| 18960 | ORF1b | 1831 | 1 | 0.046205 | 0.134145 | [] |
| 19479 | ORF1b | 2004 | 2 | 0.003583 | 0.058639 | [["volume", 0,9.8470271 04931275],["charge",0.0 0020934376 61232489,-9.90317353 3212716]] |
| 19482 | ORF1b | 2005 | 1 | 0.047439 | 0.133424 | [] |
| 19548 | ORF1b | 2027 | 1 | 0.047055 | 0.134442 | [] |
| 6535 | ORF1a | 2091 | 1 | 0.024659 | 0.164392 | [] |
| 20550 | ORF1b | 2361 | 1 | 0.006451 | 0.08932 | [] |
| 21234 | ORF1b | 2589 | 1 | 0.049474 | 0.134929 | [] |
| 21540 | ORF1b | 2691 | 1 | 0.002993 | 0.059864 | [["bipolar",2 .775557561 |

| Coordinate (SARS-CoV-2) | Gene/ORF | Codon (in gene/ORF) | # of selected branches | p-value | q-value | Properties |
|---|---|---|---|---|---|---|
| | | | | | | 562891e-15,14.65946104227764],["volume",0,11.02121929945064],["charge",0.018715457146 06748,-15]] |
| 8578 | ORF1a | 2772 | 1 | 0.019802 | 0.162018 | [["volume",0.000030536145002590 94,14.28300343026133]] |
| 8614 | ORF1a | 2784 | 1 | 0.040048 | 0.153376 | [] |
| 8659 | ORF1a | 2799 | 1 | 0.044183 | 0.142017 | [["charge",0,15]] |
| 8728 | ORF1a | 2822 | 1 | 0.038334 | 0.164289 | [] |
| 9139 | ORF1a | 2959 | 1 | 0.011134 | 0.111345 | [] |
| 9244 | ORF1a | 2994 | 1 | 0.040214 | 0.150801 | [] |
| 9424 | ORF1a | 3054 | 1 | 0.040251 | 0.147862 | [] |
| 9472 | ORF1a | 3070 | 1 | 0.040304 | 0.145094 | [] |
| 9478 | ORF1a | 3072 | 1 | 0.031055 | 0.174687 | [] |
| 9892 | ORF1a | 3210 | 1 | 0.043196 | 0.14137 | [] |
| 10027 | ORF1a | 3255 | 2 | 0.001996 | 0.044903 | [] |
| 11080 | ORF1a | 3606 | 1 | 0.028793 | 0.178713 | [] |

D.S15 Table. List of sites found to be evolving under directional selection in the entire tree, using a FUBAR-like implementation of the DEPS method. The B.1.621 tree was rooted on the genome reference sequence for this analysis. Coordinate (SARS-CoV-2): the starting coordinate of the codon in the NCBI reference SARS-CoV-2 genome Gene/ORF: which gene or ORF does this site belong to. Codon (in gene/ORF): the location of the codon in the corresponding Gene/ORF. Target amino-acid: which amino-acids have statistical support (Bayes Factor ≥ 100) for accelerated evolution towards them.

| **Coordinate (SARS-CoV-2)** | **Gene/ORF** | **Codon (in gene/ORF)** | **Target amino-acids** |
|---|---|---|---|
| 10027 | ORF1a | 3255 | I |
| 21574 | S | 5 | F |
| 22810 | S | 417 | N |
| 23011 | S | 484 | E |
| 26157 | ORF3a | 256 | S |