# Exploring Protein Interactions Through the Development of *in silico* Methodologies and Genetic Analysis

**A PhD Thesis for the Degree of**

**Doctor of Philosophy**

**in**

**Computational Biology**

**Faculty of Sciences, School of Biosciences.**

**University of Kent**

**Jake McGreig**

**2021**

# Declaration

No part of this thesis has been submitted in support of an application for any degree of other qualification of the University of Kent, or any other University in the Institution of learning.

Name: Jake McGreig

Date: 1st November 2021

# Abstract

Significant advances in sequencing technologies over the past twenty years have drastically changed scientific approaches to solving biological problems. A shift towards big data driven research and computational predictions has been made possible by vast sequence and structural databases, populated by sequences determined by modern sequencing methods. This thesis focuses on the use of these resources to predict protein function in the form of ligand-binding sites, in the determination of positions in proteins that are pivotal in virus pathogenesis, and in the elucidation of amino acids in myosin proteins which are adapted to the mass of mammalian species.

Firstly, a tool developed to predict protein-ligand interactions is introduced that utilises up-to-date data resources, tools, and machine learning algorithms to identify, and assign confidence to, candidate ligand-binding positions in proteins. This work builds on 3DLigandSite in the first major update since its release. A new webserver is presented in this work for users to query their protein sequences for probable ligand-binding residues.

Sequence analysis of conventional myosin sequences identified positions associated with increased body mass and no association with clade in the second area explored in this thesis. In beta-cardiac myosin, nine positions were found to adjust contraction velocity when mutated, highlighting the predictive power from alignments combined with a measurable phenotype. This further shows the time- and cost-saving benefits of combined computational and experimental methods.

Finally, differentially conserved positions were identified between SARS-CoV and SARS-CoV-2, determining probable variants for the source of phenotypic differences between these viruses. A webserver for other researchers to perform these analyses was developed, enabling users to query coronavirus sequences for these variants. Our analysis included all sequences available from GISAID at the time of writing. Some of the DCPs identified using these methods are shown to be found in the ACE2 binding site of the spike protein, and in close vicinity to serine protease binding sites, which are essential for virus entry into the host.

# Acknowledgements

Firstly, I would like to thank my primary supervisor **Mark Wass** for his continued support and insight throughout my PhD which has been invaluable to this research and my time working on these projects. He has given me lots of opportunities throughout my PhD and facilitated a great environment to work in.

I would like to thank all members of the Wass-Michaelis group, particularly **Stuart Masterson**, **Helen Grimsley**, **Magdalena Antczak**, and **Henry Martell**. **Stuart** and **Henry** for all the strange lunchtime conversations and insights into maize-based snacks. **Helen** and **Henry** for their competitiveness and countless games of racquet sports that kept me sane throughout these four years. **Magda** for her advice and our walks that really helped when working from home.

I would like thank **Liam McCarthy** who probably has the best-worst sense of humour that always puts a smile on my face and has been a close friend since I started.

I would like to thank **Ginny** who has always been there for me on our endless adventures and in saving me from PhD stresses.

I would like to thank **my parents** who have always supported me and helped me every step of the way for which I will always be grateful. I would also like to thank my brothers **Sam** and **Henry** for all their support.

# Table of Contents

# List of Figures

**Chapter 1**

**Chapter 2**



**Chapter 3**

# List of Tables

**Chapter 1**



**Chapter 2**



**Chapter 3**

# List of Abbreviations

| | |
|---|---|
| ACE2 | Angiotensin-converting enzyme 2 |
| ADP | Adenosine diphosphate |
| ATP | Adenosine triphosphate |
| BLAST | Basic Local Alignment Search Tool |
| BLOSUM | Blocks Substitution Matric |
| CAFA | Critical Assessment of Functional Annotation |
| CASP | Critical Assessment of protein Structure Prediction |
| CATH | Class Architecture Topology Homology |
| COVID-19 | Coronavirus Disease 2019 |
| CPE | Cytopathogenic Effect |
| DCM | Dilated cardiomyopathy |
| DCP | Differentially Conserved Position |
| DNA | Deoxyribonucleic Acid |
| GDT | Global Distance Test |
| GISAID | Global Initiative on Sharing Avian Influenza Data |
| gnomAD | Genome Aggregation Database |
| GRCH37 | Genome Reference Consortium Human Reference 37 |
| GRCH38 | Genome Reference Consortium Human Reference 38 |
| HCM | Hypertrophic cardiomyopathy |
| HMM | Hidden Markov Model |
| MERS | Middle Eastern Respiratory Syndrome |
| Mg | Magnesium |
| MSA | Multiple Sequence Alignment |
| NCBI | National Center for Biotechnology Information |
| NGS | Next Generation Sequencing |
| NMR | Nuclear Magnetic Resonance |
| NSP | Non-Structural Protein |
| ORF | Open Reading Frame |
| PANGOLIN | Phylogenetic Assignment of Named Global Outbreak Lineages |
| PCR | Polymerase Chain Reaction |
| PDB | Protein DataBank |
| RNA | Ribonucleic Acid |
| SARS-CoV | Severe Acute Respiratory Syndrome Coronavirus |
| SARS-CoV-2 | Severe Acute Respiratory Syndrome Coronavirus 2 |
| SASA | Solvent Accessible Surface Area |
| SCOP | Structural Classification Of Proteins |
| SDP | Specificity Determining Position |
| SINE | Short interspersed nuclear element |
| WGS | Whole Genome Sequence |
| WHO | World Health Organisation |

# Chapter 1: Introduction

This thesis focuses on the development and use of computational tools to predict the functional regions of proteins. This has been applied by i) creating webservers to predict ligand-binding positions in proteins, ii) by identifying sequence positions in coronaviruses that are differentially conserved, and iii) in the analysis of evolution in myosin-II protein isoforms. This thesis is composed of a series of papers illustrating this work. One of these papers outlines the functionalities and performance of the 3DLigandSite webserver, and details how to use it. One section details statistical techniques implemented to identify positions in myosin-II proteins from a range of mammalian clades that are adapted to body mass and heart rate. A final paper informs on the application of a differentially conserved position pipeline to decipher positions in coronaviruses that are likely to be associated with the transmission and pathogenicity of the disease.

## 1.1 Sequencing

Advances in sequencing technologies have rapidly increased our understanding of biological organisms through data-orientated analyses. These analyses are made possible by the vast increases in data availability over the past few decades brought about by the reduction in the speed and cost of sequencing (Leinonen, Sugawara, and Shumway 2011; Sayers et al. 2020), as shown in **Figure 1.1.** The cost of sequencing the human genome has

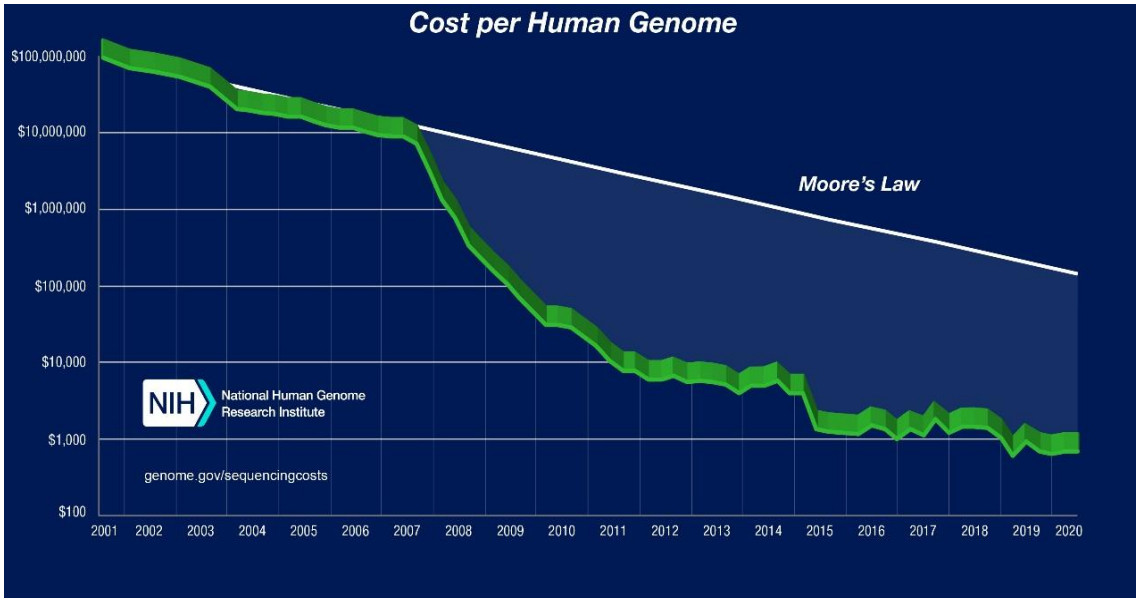



**Figure 1.1.** The cost of sequencing the human genome. Figure extracted from NIH, *Wetterstrand, no date.*

fallen by >99.99% since 2001, with the price currently at ~$1000 (Wetterstrand, n.d.), making sequencing of human and many other genomes much more accessible. One main reason for the reduction in the price of sequencing is the amount of money invested into genome projects, which has improved sequencing technologies.

### 1.1.1 Genome Projects

The Human Genome Project started in 1990, and after 13 years and between 500 million and 1 billion US dollars of funding, the human genome was successfully mapped, and a platform for large-scale sequencing projects was established. A key finding of this project was that the number of genes encoded by the human genome of roughly 20,500 is considerably less than the estimates of the time (Abdellah et al. 2004), thought to be between 50,000 to 100,000 (Kanavakis and Xaidara 2001).

In 2008, the 1000 Genomes Project was launched to catalogue human variation data, and it was completed in 2015. In this project, data obtained from over 2500 individuals and 26 populations enabled the characterisation of ~88 million variants (Auton and Salcedo 2015) in human DNA. In addition, genomic representation from broader human populations was gained, and the identification of disease- and disease-causing variants by the filtering of neutral variants (Devuyst 2015) developed our knowledge of human diseases. Further, the NHS funded a sequencing project in 2012 titled the 100,000 Genomes Project, aimed at sequencing 100,000 genomes from 85,000 NHS patients affected by rare diseases and cancer. This data is vital to advancing personalised medicine (Genomics England Ltd 2017).

These projects, and many more, have enabled a plethora of human sequences and sequence variation to be identified, to which novel sequencing data can be compared. Efforts to update and improve the human reference genome are ongoing, and the version currently widely in use is the 38$^{th}$ edition (GRCH38) (Ballouz, Dobin, and Gillis 2019; Marx 2013). It is much improved since the first version, which contained ~150,000 gaps, far greater than the 250 gaps in GRCH37 (“E Pluribus Unum” 2010). The scale of genome projects has encouraged innovation in the methods and technologies used to sequence. It has paved the way for the much more cost-effective sequencing of a wide range of organisms (Wetterstrand, n.d.), culminating in the ever-increasing sequences available in popular resource banks today, such as UniProt (Consortium 2017) and NCBI

(NCBI Resource Coordinators 2017). A number of different techniques can be used to sequence these genomes, Sanger Sequencing (Sanger, Nicklen, and Coulson 1977) and Next-Generation Sequencing (NGS) being two of the most commonly utilised and each have various advantages and disadvantages.

### 1.1.2 Sanger Sequencing

This technique is the 'gold standard' for clinical research sequencing and is often used to confirm NGS data. It is a highly accurate technique with an accuracy of circa 99.99%. However, it has a high cost in comparison to NGS when performed on genomes. This is a result of Sanger sequencing only being able to sequence a single DNA fragment at a time. Sanger Sequencing is termed the chain termination method, where a single-stranded sequence of DNA is identified by the complementary synthesis of polynucleotide chains, where chain-terminating dideoxynucleotides (ddNTPs) are incorporated into the chain (Sanger, Nicklen, and Coulson 1977; Heather and Chain 2016). In the next step, gel electrophoresis, negatively charged bases move towards an anode, then fluorescently tagged ddNTPs are read and identified in the sequence. It can produce long reads at high precision, making it suitable for sequencing single genes and verifying the findings of NGS data.

### 1.1.3 Next Generation Sequencing

NGS is a modern high-throughput sequencing technique that describes a number of different modern methods that can sequence DNA much more quickly than Sanger sequencing. One such example is Illumina sequencing. In this process, reads with lengths of 100-150bp are generated. PCR is performed on these reads to amplify them, creating spots with many copies of the same read, which are then separated into single strands to be sequenced. The slides are flooded with fluorescently labelled bases each with a terminator, so only one base is added at a time. Images of the slide are then taken at each read location, then the terminators and fluorescent signals are removed, and the process is repeated. Programs are then used to identify the base in each image of each read, and these are used to construct a sequence. These reads are aligned to a reference sequence to construct a contig. This process takes around four hours to complete, with an error rate of 0.1-1% (Salk, Schmitt, and Loeb 2018).

Nanopore sequencing is a third generation NGS technology that is commonplace for monitoring outbreaks of viruses due to its low cost, ease of mobility, and high-throughput (Greninger et al. 2015). Portable analysis of samples using MinION and Flongle devices (J. Li et al. 2020; Lu, Giordano, and Ning 2016) enable samples to be sequenced in real-time, and benchtop devices scale up this process. In this technology, Nanopores are embedded in an electro-resistant membrane, where an electric current that flows across it is measured. This sequencing method works by decoding the disruption caused by a molecule passing through a nanopore, using base-calling algorithms to determine the DNA/RNA sequence. When first introduced, the raw accuracy of this technique is much lower than other methods described in this thesis, at around 85% (Jain et al. 2018), however continuous improvements since then have resulted in >95 % accuracy now being achieved (Oxford Nanopore Technologies 2020) . However, the portability, scalability, direct analysis of samples, and real time data analysis capabilities of this method make it a valuable tool for monitoring viruses (Quick et al. 2016), and in the rapid identification of viral pathogens (Greninger et al. 2015).

### 1.1.4 Sequencing and Data Explosion

The development of affordable platforms for these sequencing methods has led to many sequences being deposited in public data repositories. Furthermore, this increase in the number of sequences is not limited to humans but includes a substantial increase in the number of different species sequenced (O'Leary et al. 2016) as indicated in **Figure 1.2**.

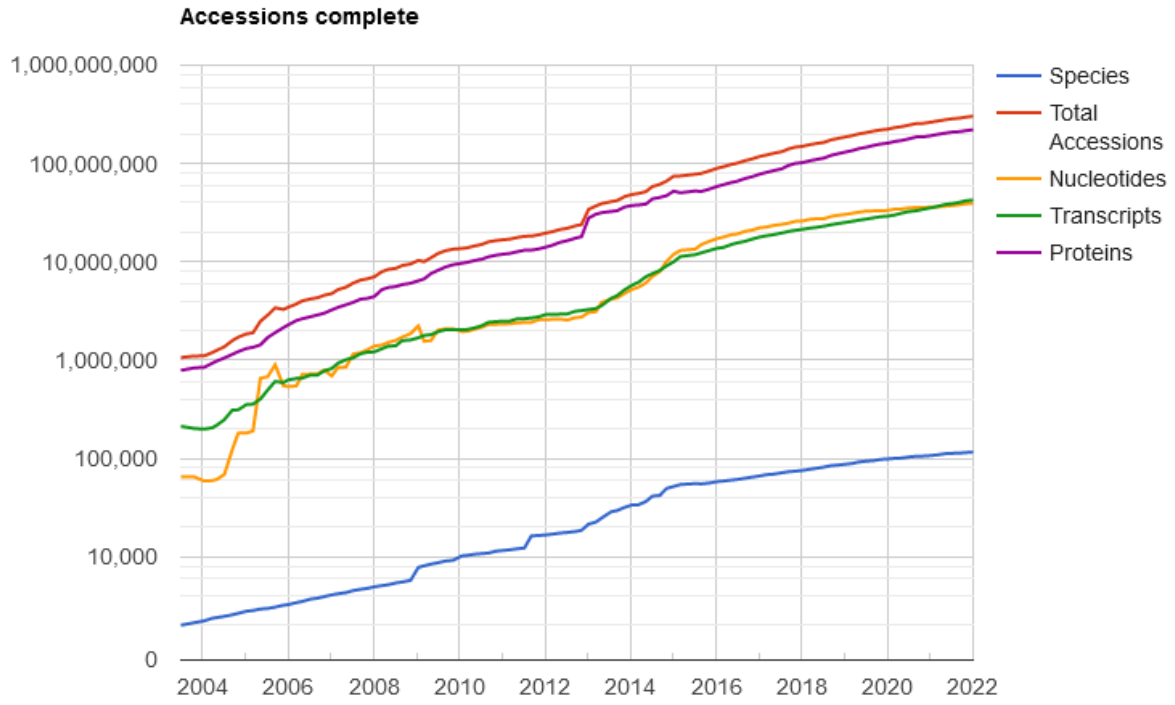



**Figure 1.2**. The rate of increase in sequence records in the RefSeq databank. Figure extracted from RefSeq (O'Leary et al., 2016) growth statistics page 06/02/2022.

Data from a wide range of organisms and clades is vital for phylogenetic inference and understanding. It also furthers the global efforts to characterise the genomes of a wide range of organisms. An area of major interest since the 100,000 Genomes Project has been sequencing representative species from each eukaryotic taxonomic family (Lewin et al. 2018). In 2018, 2500 eukaryotic species had had their genomes sequenced, and this project aims to sequence 10-15 million eukaryotic species' genomes. A collection of data of this breadth and detail is crucial for obtaining fundamental biological insights. These include the understanding of biological systems, which can be useful in exploiting biological mechanisms to our benefit, such as in the production of biotherapeutics and as model systems. Further insights can be gained from the identification of relationships between pathogenic organisms and hosts, enabling rapid development of preventative measures. In

addition, the identification of new sequences, and identifying evolutionary relationships between organisms can all advance the scientific knowledge base and provide a greater view of biodiversity. Due to the large investment in this field and other large-scale sequencing projects, the amount and diversity of sequence datasets will likely continue to rise.

Despite these innovations and commitments, one vital area of understanding sequence content requires attention. The ratio of functionally annotated to unannotated or poorly annotated sequences remain high. The roles of many of the genes and sequences deposited are not experimentally validated, leading to ~99% of functional annotations of data being assigned through automated annotation methods. As of 12/02/2020, there are ~208-million TrEMBL sequences and ~564,000 SwissProt entries in UniProt (Consortium 2017), highlighting the vast number of unreviewed sequences that require function prediction and verification. Similarly, the NCBI database shows a drastic increase in the number of whole-genome sequences (WGS) after 2003 and a steady increase in GenBank records (**Figure 1.3**) (Sayers et al. 2020). This broad dataset has prompted large-scale

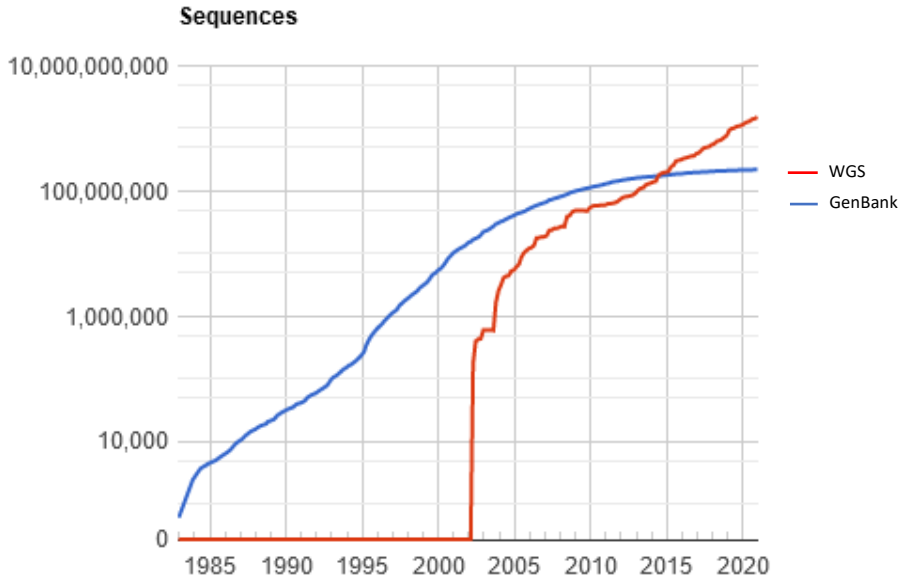



**Figure 1.3.** The increase in records deposited in the NCBI data resource from 1985-2020. Whole genome sequences and GenBank records are shown.

assessments of sequences that have led to the development of powerful tools, techniques, and databases, such as sequence homology algorithms and phylogenetic inference resources. The use of sequence homology to group genes and infer functionalities forms

the basis of many bioinformatics programs, which aim to identify the function of a gene by comparing it to similar sequences. Clustering UniProt at 50% and 90% sequence identity (UniRef50 and UniRef90) reduces the sequence numbers to ~49 million and ~131 million, respectively. Clustering is a method that uses a sequence similarity threshold to reduce the number duplicated or sufficiently similar records in a dataset. CD-HIT (Fu et al. 2012) is one such clustering algorithm that groups sequences starting with the longest sequence, then iterating through the sequence data from the longest sequence to shortest sequence and either assigning the sequence as a redundant or representative sequence. The dramatic decrease in the number of sequences as a result of clustering shows that although the total number of sequences is high, redundancy is also substantial. By identifying comparable sequences to a query sequence, homologous genes can be used to predict protein functions, evolutionary relationships between organisms and characterise protein families.

## 1.2 Inferring Protein Function

To combat the large amount of sequencing data for which little information is known, automated annotation methods are used to ascertain the likely biological function of the proteins. UniProt utilises InterPro (Zdobnov and Apweiler 2001) to classify sequences into functional domains and important sites. A protein domain is an independently stable, short region in a polypeptide chain (Xu and Nussinov 1998) that can form a functional unit and often has a conserved sequence and/or structure. UniProt uses a range of member databases to predict features, intrinsic disorder, homologous superfamilies, and domains in a protein sequence. These annotations provide insight for unreviewed sequences in UniProt. Additionally, the identification of the subcellular compartment to which a protein is targeted is a key step in understanding a protein's role, which can be determined by identifying signalling peptides, minimising the interactions in which it can participate and narrowing down a likely protein function.

Protein function can be inferred by the use of alignment tools which identify sequences with high similarity, and one of the most significant advances in these comparative analyses was the development of BLAST (basic local alignment search tool) (NCBI Resource Coordinators 2017). BLAST takes query sequences and compares them to large databases, returning the best matches. Since its development, BLAST has been updated extensively to cope with the ever-increasing sequence database sizes whilst

retaining accurate and time-efficient comparisons, and it is one of the most powerful tools for inferring a protein's function. BLAST utilises BLOSUM (Henikoff and Henikoff 1992) matrices to judge the quality of sequence alignments. BLOSUM (Blocks Substitution Matrix) was determined by local alignments of conserved regions in protein families. The frequencies of amino acids and at each position were counted and used to score the substitution probabilities of all 20 standard amino acids. The higher a BLOSUM score, the more likely the corresponding amino acid substitution is.

More recent methods employ profile-HMMs to identify sequence similarities. These are probabilistic models that capture position-specific conservation information for each amino acid in an alignment (Eddy 1998). This enables the identification of more distantly related sequences to a query. One such example that utilises HMMs to conduct sequence searches is HHsearch (Söding 2005). In this method, a profile-HMM is built for the query sequence and compared to a database of HMMs. These profiles contain far more information than a single sequence, and as such are far more sensitive to more divergent sequences. A ranked list of matches are generated in this method, displaying the metrics of similarity between the query match regions and their corresponding hits in the HMM database.

There are, however, limitations in assigning protein function, mostly arising from the rapid increase in sequences. Poorly or incorrectly annotated protein domains lead to a propagation of error effect when predicting protein function, whereby errors in the initial labelling of the domains are subsequently used to annotate proteins, resulting in these errors being passed on (Friedberg, 2006). Proteomic data is susceptible to error propagation, and strategies to combat the misleading data have been being implemented for over a decade (Schnoes et al. 2009), including the introduction of evidence codes (SwissProt) and a more conservative approach to assigning function. More recently, machine learning has been implemented to discover incorrect function annotations (Nakano, Lietaert, and Vens 2019).

Another fundamental method of distinguishing correctly annotated proteins from mislabelled data is the use of orthologs. Databanks of orthologs comprise genes in different species that evolved from a common ancestor and retain the same functionality, e.g. EggNOG (Huerta-Cepas et al. 2016) and PANTHER (Mi et al. 2021). The development of

tools that compare annotated sequences to families of orthologs enables the sequence's protein function to be verified and the assignment of likely evolutionarily related organisms.

Though identification of the likely subcellular location and domains in a protein are useful for predicting function, much more information can be acquired through analysis of the protein structure (Hvidsten et al. 2009).

### 1.2.1 Protein Structure Prediction

Protein structures can be classified based on their sub-structures and function. The CATH (Class Architecture Topology Homology) database (Dawson et al. 2017) uses manual curation to place protein structures into hierarchies at four levels. The Class level is based on the protein domain secondary structures, namely the frequency of alpha-helices (Class 1) or beta-sheets (Class 2), whilst Class 3 comprises a combination of both. Each class is split into Architectures which identifies domains with similar arrangements of secondary structures. At the T-level, the connectivity of the secondary structures is considered, and at the H-level, domain homologs at the sequence, structural and functional level are incorporated. The sequences at this level are also clustered at 35, 60, 95 and 100 % sequence identity to generate more detailed levels of classification (Sillitoe et al. 2013).

An alternate method of protein structure classification, SCOP (Structural Classification Of Proteins) (Andreeva et al. 2020; Murzin et al. 1995), sorts proteins into classes, folds, superfamilies and families and derives them based on expert knowledge. Classes are based on the secondary structure elements; folds on the arrangement and topological conditions of the secondary structures; superfamilies are based on low-level sequence identity and functional features; and families are based on a higher level of sequence identity and functional evidence. Around 70 % of domain definitions overlap between SCOP and CATH (Csaba, Birzele, and Zimmer 2009), indicating a high level of similarity between them, and both are the gold standards in protein domain classification.

CATH also incorporates functional families to capture the functional diversity within superfamilies. Functional families (FunFams) are subclusters of protein domains within homologous superfamilies which have related structures and functions, based purely on sequence patterns. They are identified by hierarchical clustering of the domains within

superfamilies (D. A. Lee, Rentzsch, and Oreng 2009), followed by optimisation using FunFHMMer (S. Das et al. 2015), which uses conservation metrics and specificity-determining positions (SDPs, described in more detail in **Chapter 1.4.4**) to categorise clusters into those likely to share function (Dawson et al. 2017). These FunFams are important for assigning functional information to proteins, and at least 50 % of human proteins are functionally annotated with FunFams, far greater than the ~8 % that are experimentally characterised (S. Das et al. 2015).

The number of protein superfamilies is far fewer than the number of protein structures, indicating that there are preferential folds for proteins to exist in. This lends itself to protein structure prediction, which can utilise the large amount of protein structural information to predict likely folds for sequences with unknown structures.

Experimentally, protein structures are typically solved using X-ray crystallography or Nuclear Magnetic Resonance (NMR), which can produce high-resolution models. Due to the time-expensive nature and difficulty performing these techniques for many proteins, the number of protein sequences with associated protein structures is low. Compared to the number of protein structures, the number of sequences remains high, as showcased in **Figure 1.4** (Berman et al. 2000). Moreover, whilst the number of sequences available is

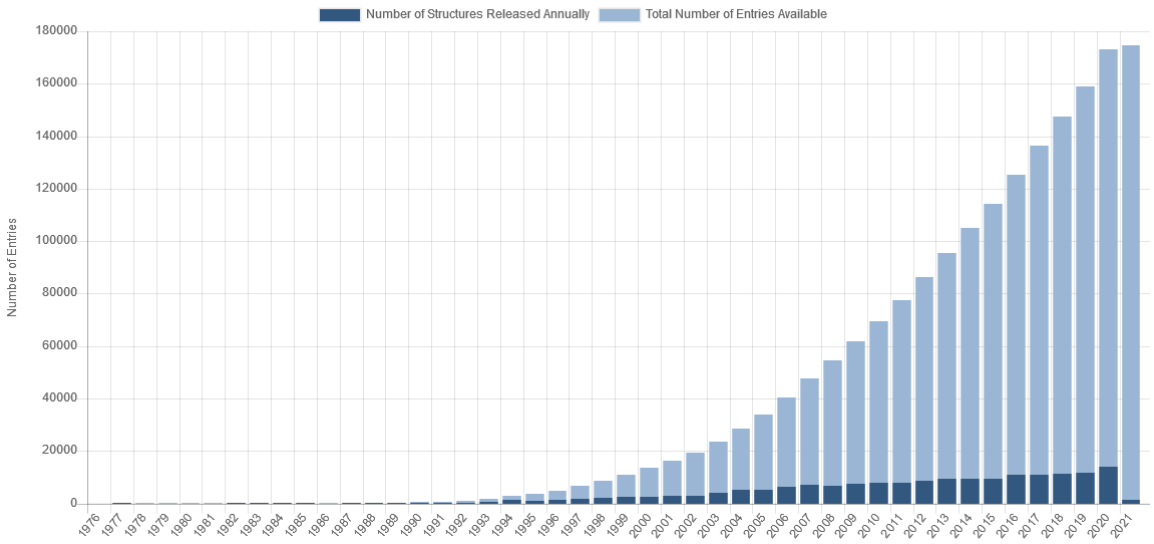



**Figure 1.4**. Protein structures released yearly by the PDB, and the total number of entries available. Figure extracted from PDB Statistics.

growing exponentially, there is a relatively smaller increase in the PDB structures deposited each year, with around 180,000 entries to date. This pales compared to the ~230 million sequences currently available in GenBank (Sayers et al. 2020). By observing the non-redundant datasets for each resource, we can estimate that far fewer sequences currently have bona-fide structures. The gulf between these values has arisen from a combination of several reasons. Firstly, there are proteins, particularly membrane proteins, which are difficult to isolate and produce in the quantities required for X-ray crystallography. Secondly, the cost, time, and skill it takes to generate viable, accurate, and high-resolution protein models is expensive, meaning it is only performed where knowing the protein model is vital. Finally, the proteins which are of keen research and economic interest are preferentially modelled, and subsequently, the size of the non-redundant databank grows more slowly as the number of these unmodelled structures of interest reduces.

Further, an area of research interest since the Human Genome Project is the structural modelling of the human proteome. Protein sequences are able to fold into a three-dimensional structure based on the amino acids of which they are comprised. In the protein folding process, the secondary structures alpha helices and beta sheets form rapidly as they are stabilised by intramolecular hydrogen bonds, followed by hydrophilic and hydrophobic interactions reorientating the secondary structures to face the aqueous or hydrophobic core environments, respectively. This is followed by disulphide bridge formation between cysteines. This forms the tertiary structure of the protein. In some proteins, quaternary structures are formed of multiple protein chains. The native state of the protein is its free energy minimum, where it is most stable. Here the unfolded states have the most energy, and through energy minimisation, the protein folds into its native state (Dobson 2004).

Understanding the structure of human proteins leads to a greater amount of detail that can be acquired when analysing disease-causing variants, metabolic pathways, and the roles of proteins. However, despite 70 % of human proteins having a known or homologous protein structure (Somody, MacKinnon, and Windemuth 2017), around 50 % of human protein structures are modelled with < 60 % coverage. Many of these proteins are likely implicated in diseases. By solving more protein structures, better quality and more targeted therapies for human diseases will arise due to improved understanding of protein

interactions. To help bridge the gap to complete structural coverage of the human proteome, methods to predict proteins structures have been developed.

The use of computational tools to predict protein structure can drastically reduce the cost and time required to identify how proteins fold, and how sequence changes affect a 3-dimensional protein model. The use of homology modelling to generate probable structures has been the most widely used and reliable method since the early 2000s. However, the use of neural networks to determine 3D representations of sequences without reference structures is becoming a powerful tool. The continued improvement of protein structure prediction is assessed in the Critical Assessment of Structural Proteins (CASP) (Kryshtafovych et al. 2019a). These assessments measure the model quality of predicted structures against protein structures solved using X-ray techniques, protein NMR, and cryo-electron microscopy that are previously unseen. The accuracy of the models is assessed using a metric termed the Global Distance Test (GDT), which is calculated by superimposing the predicted structure against the target. The amount of amino acid alpha carbon atoms within a distance to the corresponding match on the target is used to generate a score between 0 and 100, with a score above 90 typically being considered close to experimentally correct. Protein flexibility and the folding of a protein in different conditions can affect the folding of the protein, consequently, a small margin of error can be tolerated. The varying methodologies used to generate structures using these methods are discussed below.

#### 1.2.1.1 Phyre2 and homology-based methods

Phyre2 (Kelly et al. 2015) utilises profile HMMs to identify homologous protein structures to a query sequence. It then uses the backbone coordinates as a template, and models sidechains of the query onto this backbone. Confidence, coverage, and sequence identity scores are output to users, where regions of interest or the whole protein can be assessed to ascertain if the predicted structure accurately represents the query sequence.

Initially, an HHBlits (Remmert et al. 2012) search is performed against a non-redundant database of the sequences of protein structures, generating a multiple-sequence alignment. This alignment is then used to construct the secondary structure using PSIPRED (Buchan et al. 2010), and an HMM for the query is generated. The query HMM is then scanned against a database of known protein structures, and the best scoring results are

used to build the backbone of the structure. Insertions and deletions are then corrected using loop modelling, and side chains are added to the model, which is then assigned a confidence, sequence identity, and coverage score.

The ease of use and consistently high performance in the CASP evaluations make Phyre2 (Kelly et al. 2015) a valuable resource for predicting protein structure, as well as the performance scores associated with each prediction, which allows for the reliability of the results to be assessed. However, it is limited by currently being unable to predict multimeric structures, as well as in predicting difficult CASP targets due to the limitations of homology modelling.

Though Phyre2 is the method utilised in Chapters 2 and 4 of this thesis, several other strongly performing homology-based methods of predicting protein structure are available including RaptorX (Källberg et al. 2012) and IntFold (Mcguffin et al. 2019). RaptorX offers both a homology-modelling based and deep-learning-based generation of protein structures to its users. Where similar templates are insufficient to generate reliable results, structures can be generated in a template-free method using neural networks. It differs from Phyre2 by implementing a different strategy for models with little homology to known structures, improving predictions for poorly understood proteins. IntFold incorporates ligand-binding, domain, and disorder prediction providing users with functional insights into the predicted model (Mcguffin et al. 2019).

I-TASSER (Iterative Threading Assembly Refinement) (Roy, Kucukural, and Zhang 2010) is another tool the can be used to predict protein structure and function. It performs consistently well in CASP assessments, ranking in the top two spots since 2006 (Kryshtafovych et al. 2019b; Moult et al. 2018; Mariani et al. 2011). Initially, similar folds are identified in the PDB using LOMETS (S. Wu and Zhang 2007) template-based structure prediction. This is followed by the reassembly of fragments excised from these PDB templates into full-length models and selecting near-native models by clustering these decoys. A molecular dynamics simulation then refines the structure, and functional annotations are provided by COACH using substructure comparison and sequence profile alignments (Yang, Roy, and Zhang 2013b).

Rosetta (Raman et al. 2009) is a common protein structure prediction tool that performs consistently strongly in CASP, and offers the functionality of to explore protein backbone, side-chain, and sequence space. Several methods to predict protein structures can be used, with template-based, sequence-based, and specific protocols for membrane- and metallo-proteins (Barth, Wallner, and Baker 2009).

#### 1.2.1.3 AlphaFold

In the past few years, deep learning algorithms to model protein structures without solved protein templates have been developed and performed well. These breakthrough algorithms have been of great value in solving the protein folding problem. This problem, Levinthal's Paradox in 1969 (Levinthal 1969), highlights the dilemma that an unfolded polypeptide chain has an estimated $10^{300}$ possible conformations in three-dimensional space. Producing this number of models to identify the likely native pose would be far too computationally expensive to be considered a viable approach. However, proteins themselves can fold spontaneously within milliseconds, highlighting that a method exists to determine the configuration of a protein.

AlphaFold (Senior et al. 2020) uses neural networks trained on a comprehensive set of protein structures and sequences to predict the distances between pairs of amino acids and the angles of the chemical bonds that connect them. A metric that estimates the likeness to the native structure is also trained using a separate network. Gradient descent is then used to optimise the scores and produce predicted structures. This methodology has performed well on CASP13 and CASP14 targets (Alquraishi 2019), even where homologous templates are not available. The strong performance of AlphaFold in these assessments shows that effective protein structure prediction can now be achieved for a wide range of proteins. Models for Coronavirus proteins predicted by this tool were made publicly available during the global Coronavirus pandemic in 2020 (John Jumper, Kathryn Tunyasuvunakool, Pushmeet Kohli, Demis Hassabis 2020). These predictions were verified by similar models produced experimentally several months after the prediction release (John Jumper, Kathryn Tunyasuvunakool, Pushmeet Kohli, Demis Hassabis 2020).

AlphaFold has recently been used to predict the proteomes of model organisms, including humans, resulting in confident structural predictions for 58% of residues in the human

proteome (Tunyasuvunakool et al. 2021). This increases the structural coverage for humans by 41% (Tunyasuvunakool et al. 2021), dramatically increasing the knowledge base in the structural field. Further, the tool is now available for users to implement on their protein sequences – enabling widespread use of AlphaFold for proteins which can produce highly accurate predictions even when homology to the given protein is low (Jumper et al. 2021). The availability and high levels of accuracy demonstrated by AlphaFold (Jumper et al. 2021; OpenAI 2020) is a landmark in protein structure prediction and is likely to improve with future versions.

AlphaFold is able to draw on a large library of single domain protein structures, enabling predictions to be made for almost any domain, and is expected to perform well on multidomain and multimeric structures due to the way it recognises patterns in residue connectivity (Skolnick et al. 2021). One main reason that AlphaFold is so successful is being able to draw on databases with a high level of structural completeness (PDB) (Skolnick et al. 2021), and includes a huge amount of sequence data (~2.2 TB). An area in which AlphaFold can be improved is in the identification of biochemical or phenotypical effects. One method that seeks to provide this information is AlphaFill (Hekkelman et al. 2021) which uses sequence and structural similarity to AlphaFold models to transplant missing small molecules onto the predicted structures.

Improvements in deep learning and homology modelling resources will remain an area of keen focus with advances in machine learning. In addition, groups identifying novel methods to improve their algorithms, assessing new features associated with improved GDT, and building on and implementing different strategies utilised by competitors in the CASP assessments will further the method success. Using homology methods, large-scale predictions of protein structures will drastically increase the sequences with probable structural models, from which protein function can be derived. Furthermore, the application of the state-of-the-art methods described will improve the structural coverage of the human proteome and many other organisms.

### 1.2.2 Protein-Ligand Interactions

A protein structure can be used to identify likely ligand-binding positions and regions. The interaction partners of proteins are pivotal in determining the mechanism by which

proteins elicit a function. Proteins can interact with other proteins, ligands, and other small molecules, and determining these partners narrows down the number of locations, functions, and pathways in which the target protein is involved.

#### 1.2.2.1 Ligand Types

In this thesis, we define a ligand as any biologically relevant small molecule that interacts with a protein. We separate ligands into two groups, metal and non-metal ligands. Between 30 % and 40 % of proteins require metal ions to perform biological functions (Andreini et al. 2008), emphasising the key functional and structural roles of metals in proteins. Currently, only 18 cognate metal ligands are found in the PDB, determined cognate by FireDB (Maietta et al. 2014). This is a subset of the 31 metal ligands found, with the additional 13 being either non-cognate or ambiguous in their biological relevance.

Non-metal ligands encompass a much larger dataset. There are currently 636 cognate, 51 ambiguous, and ~32000 non-cognate ligands (Maietta et al. 2014), highlighting the vast number of protein structures solved with ligands that lack biological relevance, or those solved with synthetic inhibitors. These non-metal ligands can be further split into groups: organic, ionic, peptide, and polynucleotide, which each may exhibit different binding profiles to biological targets. A key observation in the last ligand binding assessment in CASP10 (Gallo Cassarino, Bordoli, and Schwede 2014) was that any following assessments of ligand binding would benefit from classifying the ligands into these categories.

#### 1.2.2.2 Features and assessment of small molecule binding

Regions of a protein that bind to a ligand or an interaction partner are usually in close vicinity in three-dimensional space yet are not necessarily adjacent in the sequence. Protein folding increases the complexity of assigning ligand binding sites as it enables primary structure amino acids to fold into similar positions to other amino acids in the protein chain. For this reason, a protein structure of the target sequence is required for accurate predictions of small molecule binding.

Ligand-binding regions can be predicted using homologous structures and sequences by determining the solvent-accessible surface area (SASA) (B. Lee and Richards 1971) and by use of conservation metrics to identify probable active site regions within a protein (Lopez et al. 2011). SASA is a key determinate of small molecule binding as ligands bind into

pockets on the protein surface to elicit their function, and this metric distinguishes surface from buried residues.

Conservation of a sequence can be inferred from sequence alignments using Jensen-Shannon Divergence to estimate the importance of a given position. In this method, each alignment position is compared to a precomputed background set of amino acids under no evolutionary pressure. Positions that vary significantly to the background set are then predicted to be functionally important or constrained. In addition, positions in the sequence within three positions to the alignment position also impact the score, the more conserved they are, the higher the score for the alignment position (Capra and Singh 2007a). Spatial neighbours would likely improve the accuracy of this method, however structural information is often not available where sequence data is.

A proteins' binding regions can be determined by comparing a query structure with unknown binding sites to a library of structures with known bound biologically relevant ligands, thus using homology to infer interactions with small molecules.

CASP8 through CASP10 (Gallo Cassarino, Bordoli, and Schwede 2014; López, Ezkurdia, and Tress 2009; Schmidt et al. 2011) evaluated the performance of ligand-binding methods on a total set of 70 protein structures which had no prior ligand-binding information known. These assessments used the Matthews Correlation Coefficient (Matthews 1975) to determine the accuracy of the predicted binding residues compared to the native bound residues of the target. These comparisons can be made using structural superposition programs such as TM-align (Y. Zhang and Skolnick 2005) and MAMMOTH (Lupyan, Leo-Macias, and Ortiz 2005). TM-align aligns the backbone $C_\alpha$ coordinates of protein structures and performs well on proteins with large differences in sequence length and therefore is useful for capturing structural information from a range of PDB structures. It works by initially aligning the secondary structures, followed by the gapless matching of the two structures, ranked by the TM-Score metric. The alignments are then refined using a heuristic iterative algorithm (Y. Zhang and Skolnick 2005). MAMMOTH identifies structural matches that are least likely to occur by chance by creating a motif alignment of residues which share properties between the query and target. The criteria for matching is the local structure overlap, the secondary structure, the structural superposition, and the ordering in the primary sequence. These are used to compute an E-value, and the tool performs well

at superposing remote homologs (Ortiz, Strauss, and Olmea 2009). These are underlying methods used to infer ligand-binding through structural similarity.

Several different methods to predict ligand- and metal-binding residues in proteins are described below.

#### 1.2.2.1 METALPDB

MetalPDB (Andreini et al. 2013) was developed to showcase minimal functional sites that bind to metals. Its database is built from all PDB structures that contain metal atoms. The minimal functional site is defined as: the metal atoms, ligands within a distance threshold, and the ligand neighbours, and they describe the local environment around the metal (Andreini et al. 2013). This tool facilitates the investigation of metal-binding sites in proteins and allows users to examine probable metal-binding regions in proteins.

#### 1.2.2.2 Firestar

Firestar (Lopez et al. 2011) uses PSI-BLAST to generate profiles for an input sequence or structure. Functional residues are then mapped onto the alignments, and residue conservation and reliability scores are calculated. The conservation of binding site amino acids is a key indicator of a residue involvement in binding (Tress, Jones, and Valencia 2003). Finally, predicted functional residues are output to the user.

#### 1.2.2.3 COACH-D

COACH-D (Y. Cui et al. 2019) combines protein-ligand binding information from four template-based methods, considering only biologically relevant ligands from BioLiP (Yang, Roy, and Zhang 2013a). Sequence conservation and structural geometry are identified using ConCavity (Capra et al. 2009), and the likely binding residues from this and the template predictions are combined to produce a consensus prediction. The ligands identified by these analyses are then docked into the binding site using AudoDock Vina (Trott and Olson 2010) to identify the optimal binding conformation.

#### 1.1.2.4 3DLigandSite (2010 Release)

3DLigandSite (Wass, Kelley, and Sternberg 2010a) uses MAMMOTH (Lupyan, Leo-Macias, and Ortiz 2005) to perform a structural search of a query protein structure to identify structurally similar proteins in the PDB that contain ligands from a curated database of biologically relevant ligands. It aligns identified structures to the query structure using

TMAlign (Y. Zhang and Skolnick 2005). This superimposes the ligands from the PDB structures onto the structure of the query proteins. Ligands are then clustered using single-linkage clustering and used to predict ligand-binding amino acids, and the probable ligands for a binding site.

### 1.2.5 Machine Learning

There has been a movement towards the implementation of machine learning algorithms to predicting biological systems and mechanisms. These algorithms, though complex, are usually fundamentally categorised as supervised or unsupervised learning. In supervised learning, data is separated into feature and target variables, where the features of a training set are fitted using a classifier, such as Random Forest (T. K. Ho 1995) or Support Vector Machine (SVM) (Cortes and Vapnik 1995). The results can then be used to predict the target variable of the test and validation datasets. A broad range of features can be determined from a protein sequence or structure, and here the operator can choose which are required for predicting their target variable.

The SVM classifier is a method that focuses on points that are difficult to decipher between classes, reasoning that if it can perform well at distinguishing these challenging data points, then less challenging data points should be easy to classify.

Random Forest and ExtraTrees are classifiers built on a large number of decision trees. Decision trees have a series of nodes representing decisions, where a pure node dictates that the data in it belongs to a single class. As the number of nodes increase, the depth of the tree increases, which leads to increased accuracy, to a point. However, runtime and accuracy can decrease when there are too many node splits, therefore the depth parameter of a decision tree classifier should be explored when making machine learning models (Kamiński, Jakubczyk, and Szufel 2018). Random Forest and Extra Trees classifiers work by considering the prediction of every tree and using a majority vote to determine the classification. Random forest uses bootstrap replicates to subsample the input data, whereas Extra Trees does not. Extra Trees also selects the cut points for splitting nodes randomly, reducing variance, whereas the Random Trees classifier finds the optimal cut point. This results in Extra Trees having a faster run time as it does not have to compute

the optimal cut points, but the two classifiers typically perform similarly (Geurts, Ernst, and Wehenkel 2006).

Logistic Regression (Cox 1959) is a fast and simple classification method. It works fitting a sigmoid curve with input data to distinguish to categories. It performs well on simple datasets when the input data is linearly separable. It struggles however to identify complex relationships between the dependant and independent variables.

In unsupervised learning, a common use is exploratory data analysis which identifies hidden patterns and groupings in a dataset that does not contain labelled responses. This can reduce the complexity of large datasets (De Ridder, De Ridder, and Reinders 2013; Sommer and Gerlich 2013) and cluster data. In a biological context, these learning methods facilitate phenotype profiling (Perlman et al., 2004), enable unknown phenotypes to be explored (Wang et al., 2008), and cluster large datasets for faster processing and identifying patterns.

### 1.2.6 Methods for Phylogenetic Analysis

Phylogenetics is the study of the reconstruction of evolutionary relationships between organisms. It can be applied to several biological fields, including the identification of related genes and in species classification. Phylogenetic relationships can be inferred from sequences, at the protein or nucleotide level, yet nucleotide data yields more accurate trees due to the capturing of both synonymous and non-synonymous mutations. Phylogenetic trees can be constructed using a range of methods. Two methods utilised in this thesis are Bayesian and maximum likelihood approaches.

#### 1.2.6.1 Bayesian Inference

Bayesian inference methods use prior probability distributions, these being the probability of an event occurring before additional data is collected, to calculate the posterior probability, which is the likelihood of an event based on previous events occurring. In Bayesian-based approaches, the prior probability of a phylogeny is combined with the tree likelihood to produce a posterior probability (J. P. Huelsenbeck et al. 2001). The best estimate of the phylogenetic relationship between the data is the tree with the highest posterior probability (Douady 2003). Several tools are available to conduct Bayesian

inference of phylogeny on biological data, including BEAST (Bouckaert et al. 2019) and MrBayes (John P. Huelsenbeck and Ronquist 2001).

#### 1.2.6.2 Maximum Likelihood

Maximum likelihood models for phylogenetic assessments utilise more sequence information than Bayesian methods. Probabilities at each alignment position are calculated, and the likelihood of these changes and their positional variability are also considered. This method produces trees which are ranked by the highest compound probability, the tree with the maximum likelihood. It relies on assumptions about sequence evolution such as equal mutation rates and the independent evolution of binding sites (Golding and Felsenstein 1990). It is a computational intense method that requires far more compute power than Bayesian methods (Douady 2003).

#### 1.2.6.3 Phylogenetic Independent Contrasts

Phylogenetic Independent Contrasts (PIC) (Felsenstein 1985) is a statistical method for incorporating trait information into phylogenetic trees. This method enables shared evolutionary traits to be distinguished from traits under a different selection pressure. The hypothesis of a trait being evolutionary correlated with another if the contrasts in one trait covaries significantly with contrasts in another trait (Quader et al. 2004).

The processes and tools outlined in Section 1.2 can be applied to most research projects which can guide and prioritise experiments to accelerate research findings. Here, we apply these techniques to protein-ligand prediction software development, coronavirus, and myosin to increase knowledge in these fields and provide information and tools for the wider scientific community to implement.

## 1.3 Myosin

Myosins are motor proteins that are vital in many cell processes for force and translocation. They are the power behind both voluntary and involuntary movements, functioning to convert chemical energy into motion at the cost of ATP. There are at least 20 myosin classes of distinct function and structure (Krendel and Mooseker 2005), and each have their specialisms. The myosin-II class, or conventional myosin, is the focus of Chapter 3, where the isoforms expressed in different tissues in mammals are investigated for sequence changes across a range of organisms of different masses. Myosin-II is a large protein (circa 500 kDa (Ojima 2019)) responsible for muscle contraction in muscle cells and the formation of stress fibres in non-muscle cells.

### 1.3.1 Myosin-II Structure

Myosin-II proteins are organised into two heavy chains and four light chains. The heavy chains comprise an N-terminal head (or motor) domain, a C-terminal coiled-coil tail domain, and a neck region connecting them, which binds to the myosin light chains. The complete myosin-II protein is very large, and as such an accurate complete structural model of it does not exist. However, high-resolution models of the motor, in both rigor (PDBID: 2MYS) and flexed PDBID: 1ATN) states, the neck, and regions of the tail domain are available in the PDB (Kabsch et al. 1990; Rayment et al. 1993), enabling detailed analysis of structural interactions.

#### 1.3.1.1 The Motor Domain

The N-terminal half of the heavy chain and one of each light chain form a single globular motor domain of ~80 kDa (**Figure 1.5**). The motor domain contains an actin-binding region (residues 655-677 and 757-771 (Consortium 2017)), which forms cross-bridges between the myosin and actin filaments, in addition to an ATP binding site where hydrolysis of ATP allows filament sliding. This hydrolysis is activated by actin, ensuring that ATPase activity is fastest when the motor domain is bound to actin. In the absence of actin, a decrease in the rate of hydrolysis is observed. The motor domain functions to couple ATP hydrolysis with motion (Harvey et al. 2000) and it is connected to the tail domain by a linker neck domain. The neck domain contains between one and seven IQ motifs which form an alpha helix that binds to calmodulin, a key stage in enabling contraction to begin (Krendel and Mooseker 2005).

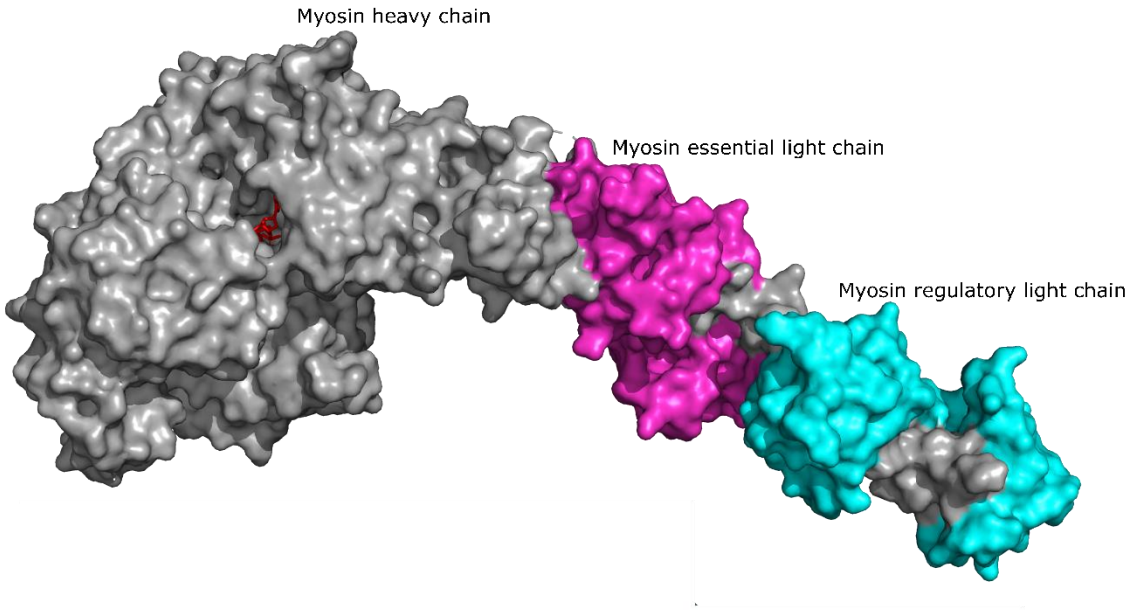



**Figure 1.5. Structure of myosin motor domain.** The protein structure of myosin motor domain with MgADP bound (PDBID: 1B7T). The regulatory light chain is shown in cyan, the essential light chain is shown in magenta, and the myosin heavy chain is shown in grey.

#### 1.3.1.2 The Tail Domain

The two carboxyl halves of the heavy chain dimerize to form a coiled-coil tail, enabling the myosin heavy chains to dimerise and form two-headed motors (Shu et al. 1999; Harvey et al. 2000). In myosin-II, multiple tail domains associate to form thick filaments, which function to organise the skeletal muscle so that motor domains are positioned at each end of the filament, with a bare zone located centrally (Harvey et al. 2000). In addition, the tail region contains conserved protein domains. The organisation of the tail domain is typically similar within myosin classes. The activity of the protein occurs primarily in the motor domain, but structurally the tail is vital for protein function and is also implicated in the stability of the inactivated state of myosin (Woodhead et al. 2005; Blankenfeldt et al. 2006).

### 1.3.2 Myosin-II Function

Myosin-II plays a vital role in muscle contraction. This occurs through a mechanism described in the sliding-filament model (Hanson and Huxley 1953; Huxley and Niedergerke 1954; Cooke 2004) (**Figure 1.6)** where myosin undergoes a series of events during movement, and causes the myosin to slide relative to the thin filament. The myosin exists

in a number of conformation states during each cycle: a prehydrolysis state where it is not bound to actin, an ADP-$P_i$ state where it is bound to actin, and a post-power stroke state (Lodish et al. 2000). In this process, the myosin motor is bound tightly to actin in a rigor state. The binding of ATP to myosin in the prehydrolysis state opens a cleft near the actin-binding site on myosin, causing dissociation from the actin. ATPase activity cleaves the inorganic phosphate from the ATP molecule, and partially closes the cleft, trapping the hydrolysis products and restoring the actin-binding site, allowing myosin to bind a new actin subunit. Large rotations around the neck region prepare the complex for the power stroke, where the motor pivots and moves the actin filament as the $P_i$ is released, and this restores the rigor conformation of the motor domain. ADP is then dissociated from the binding site (Lodish et al. 2000).

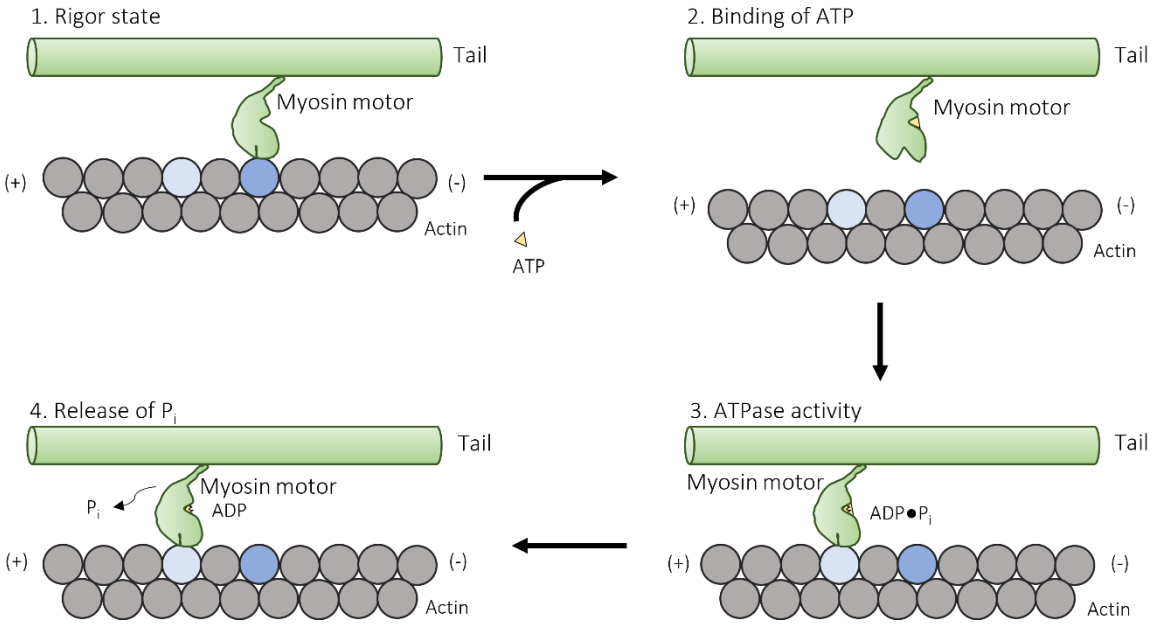



**Figure 1.6. ATP hydrolysis coupled to movement of myosin along an actin filament.**

**1**. The initial, prehydrolysis rigor state of myosin bound to the actin filament. **2.** Upon binding of ATP to the ATP binding site on the myosin motor domain, a cleft opens, causing the dissociation of the thick (myosin) from thin (actin) filament. **3.** The motor domain ATPase activity hydrolyses the ATP into ADP and $P_i$, causing the closure of the cleft, preventing the hydrolysis products from release, and restoring the actin-binding site. **4.** The myosin is then primed for the power stroke, where $P_i$ is released, and the actin filaments slide towards the middle of the thick filament (in myosin-II). Figure generated using Lodish et al., 2000, Figure 18-25 as a template.

This process drives movement and is fundamental for life as we know it today. The investigation of conventional myosin proteins in close detail is vital to understand how the protein is adapted to the demands of a diverse range of organisms needs.

### 1.3.3 Myosin II Isoforms

In mammalian organisms, a range of different myosin isoforms are expressed. The characterisation of the functional classes of these myosins is based on the motor domains of the sequences, from which phylogenetic relationships are inferred (**Figure 1.7**) (Berg, Powell, and Cheney 2001). It is hypothesised that myosin I and myosin II were the earliest members of the myosin superfamily to evolve (Thompson and Langford 2002), which is likely the reason for the greatest divergence of isoforms in these classes (**Figure 1.7**). These classes of myosin are highly important for cellular processes – myosin I functions in vesicle transport, it is a ubiquitous cellular protein and vital for cell survival, and myosin II is responsible for regulating cell morphology (Newell-Litwa, Horwitz, and Lamers 2015) and the muscle contraction on which many multicellular organisms rely. Unsurprisingly, a large number of these proteins are found in humans. The human genome encodes for 40 myosin genes, covering 12 classes of myosin (Krendel and Mooseker 2005). Of these, there are 14 known conventional myosin genes including two cardiac, two non-muscle, a smooth muscle, and six skeletal muscle genes (Berg, Powell, and Cheney 2001).

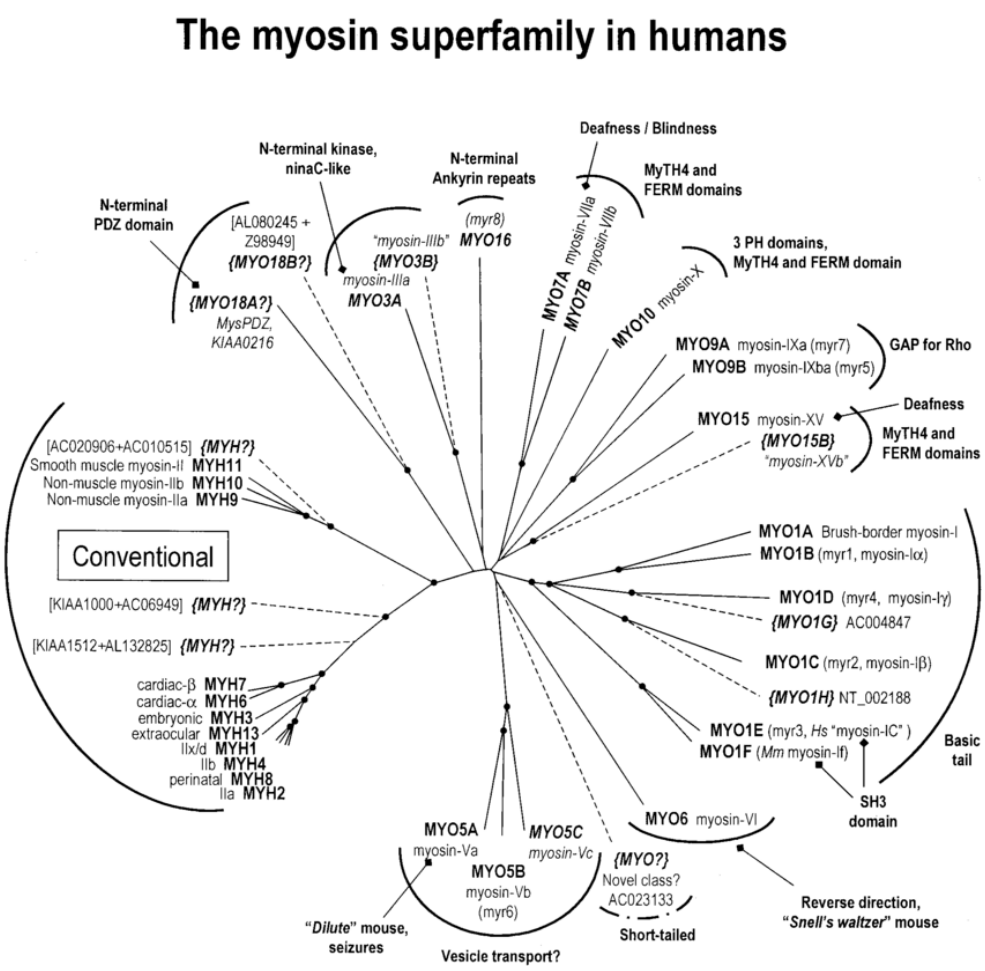



**Figure 1.7. An unrooted phylogenetic tree of the myosin superfamily in humans.** The phylogenetic relationship between myosin classes, and the intra-class distribution of class isoforms is depicted in this figure. Figure extracted from Berg, Powell and Cheney, 2001, Figure 1.

### 1.3.3.1 Myosin-II Cardiac Isoforms

Myosin-II cardiac isoforms include the alpha (α, MYH6) and beta (β, MYH7) genes. These are located in tandem on chromosome 14 (Marín-García, Goldenthal, and Moe 2007) and play key roles in heart muscle contraction. The alpha-cardiac isoform exhibits a higher rate of ATPase activity than the beta-cardiac isoform, and the amount of expression of each isoform contributes to the contraction velocity and force generation in the heart (Nakao et al. 1997; Mahdavi, Chambers, and Nadal-Ginard 1984). The alpha-cardiac isoform is primarily expressed in the atria (Gelb and Chin 2012), while the beta-cardiac isoform is expressed in humans in both the embryonic heart and the atria of adults (Marín-García, Goldenthal, and Moe 2007). Mutations in these isoforms are implicated in diseases, principally hypertrophic cardiomyopathies (HCM), as well as dilated cardiomyopathies

(DCM) (Francine Parker and Peckham 2020). Most of these mutations are missense – 920/1000 in MYH7 and 128/145 in MYH6 (Francine Parker and Peckham 2020). The mutations are classified based on their prevalence in patients with cardiomyopathies, however, this likely has led to a large number of passenger mutations reported.

#### 1.3.3.2 Myosin-II Non-muscle isoforms

In mammals, there are three non-muscle myosin II isoforms of the heavy chain that contribute to cell adhesion, protrusion, and polarity (Vicente-Manzanares et al. 2009). These isoforms are non-muscle A (NMA) encoded by MYH9, non-muscle B (NMB) (MYH10) and non-muscle C (NMC) (MYH14). Cell types appear to express up to two of the non-muscle isoforms and never three under normal conditions (Betapudi 2014). For example, MYH9 and MYH10 are expressed at comparable levels in epithelial and endothelial cells, whilst in lung tissue MYH14 is predominantly expressed (Betapudi 2014).

#### 1.3.3.3 Myosin-II Skeletal Muscle Isoforms

There are six skeletal muscle myosin-II isoforms which form fast/type-II fibres: IIa (MYH2), IIb (MYH4), IIx (MYH1), extraocular (MYH13), embryonic (MYH3), and perinatal (MYH8) (Resnicow et al. 2010). Embryonic and perinatal myosin are expressed early on in muscle development and thus are termed developmental myosin heavy chains (MyHC). Expression levels of these isoforms decrease rapidly shortly after birth, only continuing expression in specialised muscles and regenerating dystrophic muscles (Resnicow et al. 2010). Extraocular MyHC is highly specialised and only expressed in the extraocular and laryngeal muscles, and is one of the fastest myosins in terms of shortening velocity (Sciote et al. 2002). The adult-fast MyHCs (IIa, IIb, and IIx) are the main isoforms expressed in skeletal muscles, in varying levels based on several factors including age, disease, and exercise.

#### 1.3.3.4 Myosin-II Smooth Muscle Isoform

There are four smooth muscle isoforms encoded for by the MYH11 gene by alternative splicing, and they are differentially expressed as muscle cells mature (Kwartler et al. 2014). Two of these isoforms are expressed in the aorta (Babu, Warshaw, and Periasamy 2000), and all function in muscle contraction.

#### 1.3.3.5 Other

SlowT (MYH7b) is a slow twitch isoform discovered in mammals in 2002. It is more highly conserved than either of the cardiac isoforms (Warkman et al. 2012), and appears vital for normal cardiac function, as MyH7b knockouts and mutations can cause HCM in mice (Warkman et al. 2012).

These isoforms are summarised in **Table 1.1**, where the isoform, gene name, and tissue in which the isoform is primarily expressed are shown.

| Type | Isoform | Gene | Tissue |
|---|---|---|---|
| Cardiac | α | MYH6 | Heart, skeletal |
| | β | MYH7 | Heart |
| Fast Twitch | IIA | MYH2 | Skeletal, tongue |
| | IIB | MYH4 | Skeletal |
| | IIX | MYH1 | Skeletal |
| Developmental | Embryonic | MYH3 | Oesophagus, prostate, skeletal |
| | Perinatal | MYH8 | Oesophagus, skeletal, tongue |
| Non-Muscle | Non-muscle A | MYH9 | N/A |
| | Non-muscle B | MYH10 | N/A |
| | Non-muscle C | MYH14 | Intestine, skeletal |
| Smooth Muscle | Smooth muscle | MYH11 | Smooth muscle, urinary bladder |
| Other | SlowT | MYH7B | Heart, skeletal |
| | Extraocular | MYH13 | Placenta, skeletal, stomach |

**Table 1.1**. Table summarising the isoforms and tissue expression of active myosin isoforms. Tissue expression information obtained from the Human Protein Atlas (Uhlén et al. 2015).

### 1.3.4 Distribution across Mammalian Clades

Phylogenetically, humans are placed in the mammalian class. This class comprises over 6450 species (Burgin et al. 2018), with organisms ranging in sizes from the Etruscan shrew at ~2 grams (Jürgens et al. 1996; Brecht et al. 2011) to the largest terrestrial mammal, the African bush elephant, at 10400 kilograms (maximum) (Larramendi 2016) and the heaviest mammal the blue whale at >150000 kilograms (Árnason et al. 2018). There are ~5400 living mammalian species (Foley, Springer, and Teeling 2016) inhabiting a range of niches and

biomes, dictating their adaptations in size and function. Mammalian masses are frequently found between these mass extremes (**Figure 1.8**) (Blackburn and Gaston 1998), highlighting the diversity within this group and inviting research questions as to how muscle cell architecture and, in particular, myosin proteins are adapted to mammals at a range of masses.

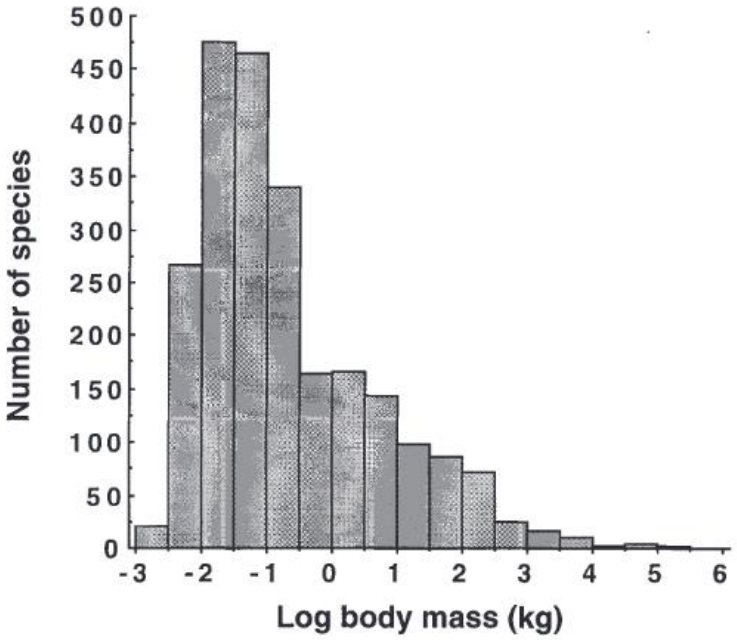



**Figure 1.8. Histogram depicting the distribution of mammalian masses**. This figure is based on mammalian species mass values known at the publication date. Figure extracted from Blackburn and Gaston, 1998, Figure 1.

Mammalia can be subdivided into four superordinal groups: Euarchontoglires, Laurasiatheria, Afrotheria, and Xenarthra (Foley, Springer, and Teeling 2016). Sequencing data is mostly available for the first three groups, which also encompasses most living mammals. Euarchontoglires comprises placental mammals, including rodents, primates, dermopterans, lagomorphs and scandentians (V. Kumar, Hallström, and Janke 2013). There are over 150 genomes available from members of this clade in GenBank (X. Song et al. 2021; Sayers et al. 2020). However, there are many more without available sequencing data, as there are over 1500 species of rodent alone (Tibbetts 2017). Despite this, a broad range of sizes of these mammals are covered by the sequencing data – from small rodents such as

mice up to large primates such as gorillas, covering both the Glires (rabbits, hares and rodents) and Euarchonta (tree shrews, flying lemurs, and primates). Euarchontaglires descends from the Boreoeutheria clade along with Laurasiatheria (Springer et al. 2011).

Laurasiatheria contains six orders however their phylogenetic placements within the superorder are a topic of much debate. The orders are Eulipotyphla (hedgehogs, moles and shrews), Perissodactyla (horses, rhinoceroses), Cetartidactyla (hoofed placental mammals, cetaceans), Chiroptera, Carnivora, and Pholidota (pangolins) (Hu, Zhang, and Yu 2012). This incredibly diverse superorder ranges from whales to bats, yet the placements of its orders phylogenetically remains a challenge due to the divergence of major laurasiatherian lineages occurring over a short period (1-4 million years) (M. Y. Chen, Liang, and Zhang 2017). Sequence data from all major lineages of Laurasiatheria are available (M. Y. Chen, Liang, and Zhang 2017), providing a good data resource for investigating the divergence of myosin proteins in this thesis.

Afrotheria is a more divergent superordinal group. This encompasses placental mammals, including the Tubulidentata, Macroscelidea, Chrysochloridae, Tenrecoidea, Sirenia, Hyracoidea, and Proboscidea (Seiffert 2007). The phylogenetic basis for this grouping is strongly supported by a range of genomic data, where indels, nuclear and mitochondrial DNA sequences (Amrine-Madsen et al. 2003), and SINEs (Nikaido et al. 2003) lend credit to the clade. It has the longest stem divergence of around 25 million years (Seiffert 2007), and as such contains a highly varied group of organisms.

### 1.3.5 Motivation of research

Placental mammals have adapted to most biological niches on the planet and have evolved into a plethora of different shapes and sizes. Each niche presents its own challenges: an elephant and a mouse would exhibit vastly different muscle fibres to achieve their movements. By investigating sequence changes in placental mammal myosin-II isoforms, we can begin to understand the different demands that size, and other factors, can place on muscle contraction. The primary focus of research performed in Chapter 3 is based on the beta-cardiac myosin-II isoform.

## 1.4 Coronaviruses

Coronavirinae is a virus family comprising four genera of positive-sense, single-stranded RNA viruses: α-, β-, γ-, and δ-coronaviruses. Of these, the α- and β-coronaviruses are known to cause disease in humans, and Severe Acute Respiratory Syndrome (SARS-CoV), Middle Eastern Respiratory Syndrome (MERS), and Severe Acute Respiratory Syndrome (SARS-CoV-2) have all caused severe outbreaks in the past twenty years (**Table 1.2**).

| **Species** | **Cases** | **Deaths** |
|---|---|---|
| SARS-CoV | 8098 | 774 |
| MERS | 2566 | 866 |
| SARS-CoV-2 | >111 million | >2.46 million |

**Table 1.2**. Wide-spread betacoronavirus outbreak cases and deaths. Outbreak data was taken from the WHO.

Though not fully elucidated, the reservoir of these viruses appears to be of bat origin (Latinne et al. 2020). Transmission of MERS to humans is most common through infected dromedary camels (WHO) and was originally transmitted from bats to camels in the distant past (Reusken, Haagmans, and Koopmans 2014). The natural reservoir of SARS-CoV is likely to be horseshoe bats (L. F. Wang and Eaton 2007) and spill over from this population to humans occurred through intermediate hosts such as palm civets (M. Wang et al. 2005). Phylogenetic analyses suggest that the zoonotic source of SARS-CoV-2 is horseshoe bats (RaTG13 ) (Z. Zhu et al. 2020), though pangolins have also been proposed as the natural reservoir of this virus due to having the closest sequence similarity to the RatG13 coronavirus and SARS-CoV-2 (T. Zhang, Wu, and Zhang 2020; P. Zhou et al. 2020; Lam et al. 2020). Of the β-coronaviruses, SARS-CoV and SARS-Cov-2 are markedly more similar, sharing 80.21% genomic sequence identity and 97.71% sequence identity in the spike protein, whilst the MERS spike protein and SARS-CoV-2 spike protein share 32.79% sequence identity (Kaur et al. 2021). The similarities between SARS-CoV and SARS-CoV-2 have prompted comparison between the two viruses to better understand each virus phenotype and pathogenicity.

### 1.4.1 Virus Phenotype and Pathogenesis

The SARS outbreak originated in China in 2003 and spread to four other countries, resulting in 8098 cases and 774 deaths, a mortality rate of 9.6%. It typically infects the lower respiratory tract causing flu-like symptoms, fever, chills and malaise (Gu and Korteweg 2007), and is spread through droplets and contact with contaminated surfaces (Olsen et al. 2003). The genome of this virus is organised into 11 ORFs that can produce 23 proteins (Ruan et al. 2003).

In December 2019 in the Wuhan food market in China, an outbreak of novel coronavirus SARS-CoV-2 spread rapidly across the globe. As of 24/09/2021 there have been 219 million cases and 4.55 million deaths, a mortality rate of 2.1%, though many more cases are expected in the asymptomatic population that do not factor into the mortality rate. This virus infects the respiratory tract and is transmitted from host-to-host through direct contact, by aerosols, and in exhaled breath (Ma et al. 2020; Lednicky et al. 2020; Q. Li et al. 2020), particularly in poorly ventilated areas (Meyerowitz et al. 2021). The virus genome encodes 11 genes (Khailany, Safdar, and Ozaslan 2020), including a polyprotein that is

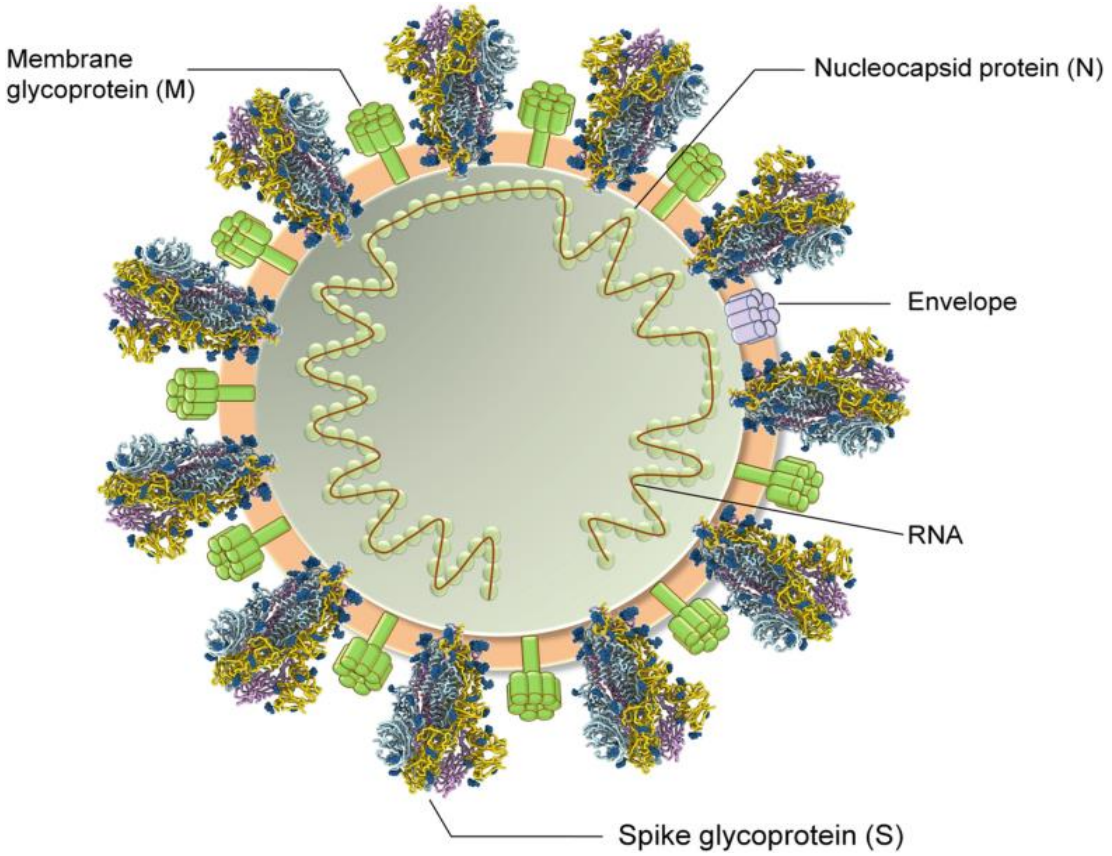



**Figure 1.9.** Structural protein organisation for SARS-CoV-2. Figure extracted from (Kumar et al., 2020)

cleaved into 16 non-structural proteins, structural proteins (**Figure 1.9**), and several accessory proteins encoded by ORFs (S. Kumar et al. 2020).

The mechanism of cellular entry is similar for these SARS-CoV and SARS-CoV-2. The dominant host receptor for infection is the angiotensin-converting enzyme 2 (ACE2), which interacts with the viral spike protein to mediate cell entry. The spike protein must be primed by the serine protease TMPRSS2 and cysteine proteases cathepsin B and cathepsin L (Hoffmann, Kleine-Weber, et al. 2020), which generates the S1/S2 fragment and S2' fragment of the spike protein (Hussain et al. 2020). The proteolytic cleavage of the spike protein by TMPRS22 occurs at positions R685-S686 and R815-S816 (Hussain et al. 2020), producing S1 and S2 fragments. The S1 domain of spike protein binds to ACE2, altering the conformation of the S2 domain and allowing fusion between the viral envelope and plasma membrane of the target cell (Xia et al. 2020; Tang et al. 2020) (**Figure 1.10**).

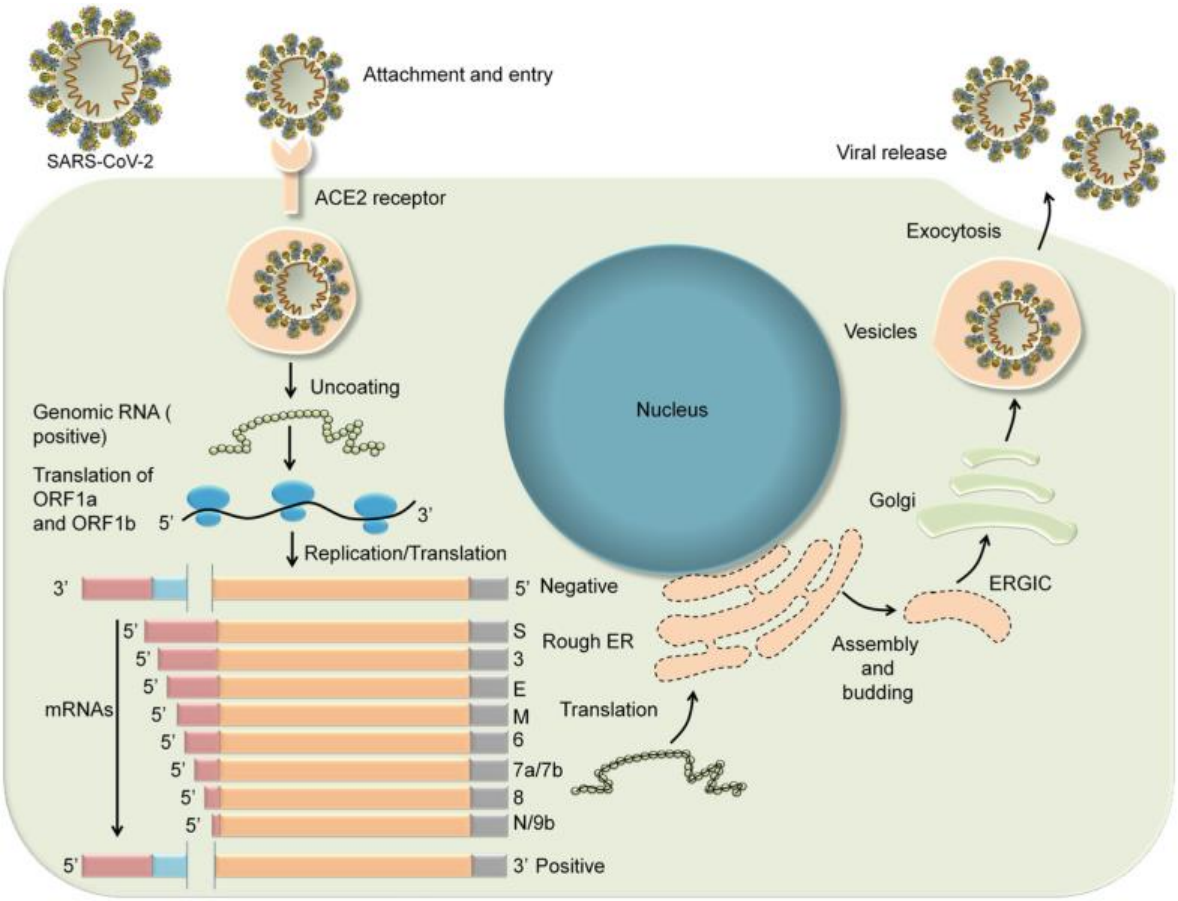



**Figure 1.10**. Overview of the Replication of SARS-CoV-2. Figure extracted from Kumar et al.

The SARS-CoV-2 spike protein binds to ACE2 with a greater affinity than the SARS-CoV Spike (Wrapp et al. 2020), resulting in an increased transmission rate. Furthermore, SARS-CoV-2 exhibits a longer incubation period, is infectious more quickly after symptom onset, and has a higher proportion of mild illness in hosts (Petersen et al. 2020). These make for a disease that is highly effective in transmission and difficult to contain.

### 1.4.2 Coronavirus Proteins

Despite the genomic similarities of SARS-CoV and SARS-CoV-2, the transmission and severity of the viruses are clearly distinct. The differing phenotypes of the viruses must therefore be a result of differences in the proteins that the viruses encode.

One protein of interest is the spike glycoprotein. The spike protein is a large homotrimeric protein, and each chain is made up of 1273 amino acids. It is made up of two functionally distinct regions, the S1 region is involved in the attachment of the virus, and the S2 subunit is involved in the entry of the virus. The main function of this protein is to facilitate virus entry into host cells through the binding of the ACE2 receptor. After internalisation into endosomes, this protein undergoes structural changes and allows the entry of nucleocapsids that can start replication (Gonçalves et al. 2020).

The envelope protein (E) functions in the assembly and release of the virus and is required for pathogenesis but not replication (Nieto-Torres et al. 2014). The membrane protein (M) is the most abundant structural protein and functions to give the virion its shape, promote membrane curvature, and bind to the nucleocapsid. The interaction of the nucleocapsid protein with M and NSP3 functions to package the viral genome into viral particles. This protein binds to RNA using its N-terminal or C-terminal domain, with optimal binding utilising both (Chang et al. 2006).

ORF1a and ORF1b produce two large polyproteins by proteolytic cleavage (Yoshimoto 2020). Protease enzymes (NSP3 and NSP5) cleave the polyproteins into NSP1-11 and 1-16, respectively (Astuti and Ysrafil 2020). Most of these proteins function in replicase–transcriptase complex assembly and are involved in RNA replication and transcription (Fehr and Perlman 2015).

### 1.4.3 SARS-CoV-2 Phylogeny and nomenclature

SARS-CoV-2 has spread rapidly, infecting a large proportion of the human population. A range of phenotypes of the virus are observed in different regions and display different mortality rates and transmissibility. This could be due to mutations in the coronavirus genome or ecological factors such as population density and accessibility to better quality healthcare. Additional factors such as government guidelines on virus containment, identification of COVID-19 cases through testing procedures, and the definitions of COVID-19 related deaths are all important considerations when comparing cases across countries. Through sequence analysis of the virus, it was possible to identify the mutations in the virus and subsequently track their spread. Variants advantageous to transmissibility and viral spread quickly became dominant as the pandemic grew (Le Page 2021).

Deposition of SARS-CoV-2 genome sequences from around the world into data banks invited the determination of phylogenetic clades and groupings of sequences. Several different nomenclatures have been used to dictate clade – typically coined by the sequence deposition groups. GISAID (Global Initiative on Sharing Avian Influenza Data) (GISAID 2020), PANGOLIN (Phylogenetic Assignment of Named Global Outbreak Lineages) (O'Toole et al. 2021), and Nextstrain (Hadfield et al. 2018) each name the sequence groups generated by their phylogenetic analyses differently. Comparison of the clades named by these resources are shown in **Figure 1.11** below (Alm et al. 2020). For the purposes of this thesis, we will use the GISAID nomenclature. GISAID currently separates SARS-CoV-2 sequences into nine groups. The naming is based on the mutation that leads to a branching and novel clade introduction. The clades S and L were prevalent at the start of the pandemic but now constitute a small fraction of sequences compared to other clades. The L clade split into clades G and V, then G further split into GR, GH and GV. GR subsequently split into GRY, and the branching of these clades are illustrated in **Figure 1.11**.

Of particular interest, the D614G clade arose in February 2020 (Isabel et al. 2020) and quickly became the dominant clade in regions it entered. It was accompanied by mutations in ORF1ab in proteins involved in replication, possibly affecting the replication rate of the virus (Koyama, Platt, and Parida 2020). Subsequent mutations within this branch resulted in additional clades evolving, which must be monitored to anticipate virus adaptations. In particular, the GK clade (delta coronavirus) poses a significant threat to the containment of

COVID-19 as it exhibits a lower sensitivity to antibodies (Planas et al. 2021) and has become dominant in numerous countries after its identification in India in late 2020. This variant and 13 others are considered variants of concern by Public Health England (Public Health England 2020; 2021) and must continue to be tracked and investigated. To aid in public understanding of current coronavirus variants, the WHO has developed a simple labelling system for coronavirus variants of interest, using letters of the Greek alphabet (WHO 2021).

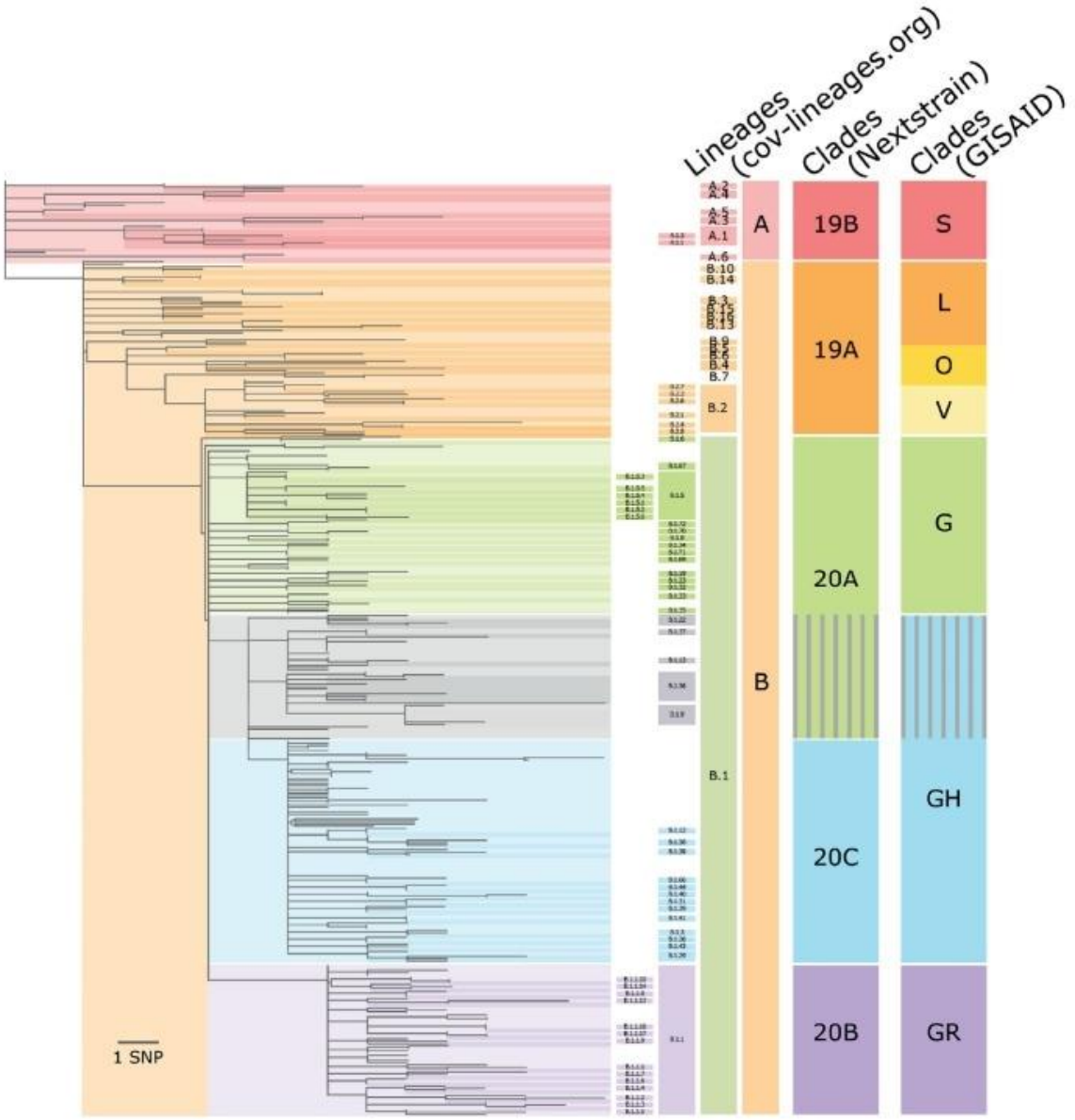



**Figure 1.11. Clade representation by sequence deposition and annotation resources.** The strains of SARS-CoV-2 as represented by GISAID, PANGOLIN, and Nextstrain. Figure extracted from (Alm et al., 2020).

### 1.4.4 Differentially Conserved Positions

Differentially Conserved Positions (DCPs), also known as Specificity Determining Positions (SDPs), are positions in the sequences of two groups of proteins that are conserved within their groups and exclusive to their group (Sloutsky and Naegle 2016). This method was originally used to identify functionally important residues in proteins and to investigate interactions at protein interfaces (Rausell et al. 2010; Casari, Sander, and Valencia 1995). DCPs are calculated from multiple sequence alignments (MSAs) between two related groups of sequences (see **Figure 1.12**). Proteins are under evolutionary pressures, and

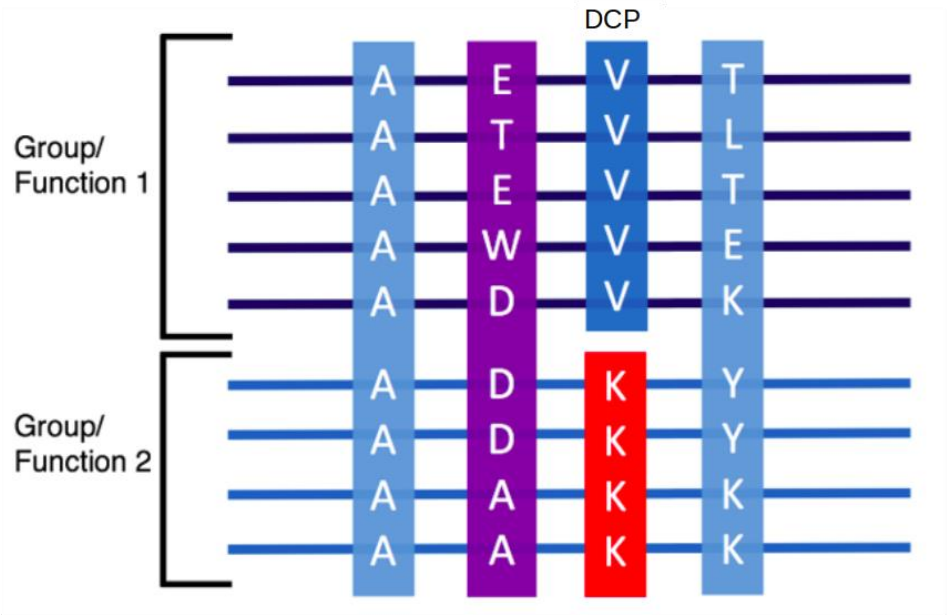



**Figure 1.12.** Example identification of DCPs. Position three in the alignment would be considered a DCP.

random mutations can give rise to a number of different phenotypes. However, for a protein to maintain its core functionality, large sections of the protein must remain unchanged. Further, variable positions are less important biochemically and are more structurally conserved, with several amino acids being interchangeable in these positions. The features of amino acids can be conserved, be that the size of the amino acid or its functional groups, and often this feature conservation can be indicative of a structural or functional advantage. Remaining positions that are conserved between groups of evolutionarily related organisms are candidates for determining specificity and affecting

divergent phenotypes of the protein. Scoring functions for characterising DCPs differ depending on the program used, but in each only highly conserved positions that differ between groups can be classed as DCPs.

There are, however, limitations of MSAs. Alignment algorithms assign gaps by a range of methods and penalties for introducing single gaps and multiple gaps vary. The penalties for the introduction of a single gap versus multiple gaps differ (Edgar 2004; Sievers and Higgins 2014a; Katoh and Standley 2013). Though this is unlikely to be a frequent occurrence between groups of high sequence similarity, it can be addressed by using multiple alignment algorithms and assigning each DCP a confidence score based on the number of different alignments in which it is a DCP.

Further, the method is only applicable when sequence similarity between the two sets is sufficiently conserved, as a high level of divergence results in a very large set of DCPs. Sequences with sequencing errors and poorly sequenced regions can give rise to false positive or false negative calling of DCPs, but methods have now been developed to reduce this uncertainty (Castresana 2000). In addition, the size of each group should be representative of the species. Sequences from an isolated event may result in poor coverage of the species, leading to biased output and less informative DCPs identified. The effective number of sequences can be calculated from the sequence length, the number of sequences in the MSA and the sequence identities between them (C. Zhang et al. 2020).

Sequence alignments with sufficient data quality and quantity can be investigated to decipher candidate positions responsible for a change in phenotype between proteins of related organisms. Furthermore, DCPs can assign more specific functions to unannotated proteins (Hannenhalli and Russell 2000). S3det (Muth et al. 2012) is one such tool that identifies DCPs and maps them to protein structures to gauge the biological relevance of the DCP.

This technique is particularly applicable for viruses, which encode few genes, mutate rapidly, and elicit differing phenotypes often. One such example of the effectiveness of DCPs in identifying changes associated with phenotype is in Ebolavirus. Ebolaviruses are negative strand RNA viruses that form part of the filoviridae family and have a length of ~19000bp. There are six known species of Ebolavirus with varying degrees of pathogenicity.

Zaire, or EBOV, is the most lethal and has a mortality rate ranging from 25%-90%. In the 2014 outbreak in West Africa, for which EBOV was responsible, over 11,323 people died of this virus. Its high mortality rate, transmission through direct contact with bodily fluids, and a lengthy incubation period of between 2-21 days facilitated a widespread pandemic. Three additional Ebolavirus species – Sudan virus, Tai Forest virus, and Bundibugyo virus – have all shown to be pathogenic in humans, however, the onset of disease from Reston Ebolavirus has yet to occur. This indicates that it is non-pathogenic in humans. By identifying DCPs between pathogenic and non-pathogenic strains of the virus, a subset of 189 amino acid differences between the two groups was collected, which likely contain the cause of the phenotypic differences. Further structural analysis narrowed this set to 47 DCPs (Pappalardo et al. 2016) and fewer present in regions of interest, enabling predictions about the changes required to make Reston Ebolavirus pathogenic. Additional examinations using DCPs have indicated that the recently discovered novel Ebolavirus, Bombali virus, is likely non-pathogenic in humans (Martell et al. 2019). This highlights the predictive application of this method which can be applied to other viruses with distinct phenotypes.

## 1.7 Scope and Outline of this Thesis

This thesis focuses on the use of computational tools to predict and assign protein function from sequence and structural analysis.

Chapter 1 introduces and provides background for the topics explored in this thesis and explains fundamental methodologies on which much of this thesis relies.

Chapter 2 describes the development of 3DLigandSite tool and webserver and its performance on validation data. This work is set to be sent out for review. My role in this research was design and creation of the webservers, the 3DLigandSite Pipeline, and in the drafting of the manuscript.

Chapter 3 describes sequence analysis of myosin-II isoforms and the altering of human beta-cardiac myosin-II to behave like rat myosin-II. This work is published in PLOS Biology (C. A. Johnson et al. 2021). My contribution to this research includes all bioinformatics works performed and the drafting of the manuscript.

Chapter 4 describes the application of Coronavirus Analysis Tool (CAT) to Coronavirus Proteins and the possible implications of DCPs identified. This work is published in Bioinformatics. My role in this project was creating a pipeline to determine DCPs found between coronaviruses, in the analysis on the DCPs identified, and in the drafting of the manuscript. I also developed the CAT webserver and its functionality in predicting DCPs.

Chapter 5 summarises the work from previous chapters and explores the implications of the research and future work for building on the findings.

Further research performed during my PhD and not included in this thesis includes:

- Investigation of DCPs in the recently discovered Bombali Ebolavirus to identify whether it is likely pathogenic (Martell et al. 2019).
- Deposition of 3DLigandSite results as part of the functional annotations of PDB data in PDBe-KB (Varadi et al. 2020).
- Investigation of Aprotinin as an inhibitor of SARS-CoV-2 replication (Bojkova et al. 2020).

# Chapter 2: 3DLigandSite: Structure-based prediction of protein-ligand binding sites

Jake E McGreig, Hannah Uri, Magdalena Antczak, Michael JE Sternberg, Martin Michaelis, Mark N Wass

My contribution to this work is as follows:

1. Wrote all scripts for preparing data and databases that 3DLigandSite uses.
2. Wrote the 3DLigandSite program.
3. Created a webserver to host 3DLigandsite.
4. Created the front and backend of processing for the webserver.
5. Analysed the results.
6. Wrote the manuscript, along with Mark Wass.

## 2.1 Abstract

3DLigandSite is a web tool for the prediction of ligand-binding sites in proteins. It allows users to submit either a query protein sequence or structure. Results are displayed in multiple formats to enable users to investigate the predicted binding sites, including an interactive Mol* molecular visualisation of the protein and the predicted binding site. Here, we report a significant update since the first release of 3DLigandSite in 2010. The overall methodology remains the same, with 3DLigandSite inferring candidate binding sites in proteins using known binding sites in related protein structures as templates. The tool now uses HHSearch to identify template structures. Moreover, we introduced a machine learning element as the final prediction step, which improves the accuracy of predictions and provides a confidence score for each residue in a predicted binding site. Validation of 3DLigandSite on a set of 6416 binding sites obtained recall of 0.72 at 0.75 precision. 3DLigandSite is available at https://www.wass-michaelislab.org/3dligandsite.

## 2.2 Introduction

Elucidation of protein function remains a difficult and important task, with many millions of proteins present in UniProt (Bateman et al. 2021) and only a small fraction of them functionally annotated (N. Zhou et al. 2019), making automated sequence annotation tools essential. Small molecules that bind to proteins are intimately related to protein function; they can be substrates or products of an enzyme reaction, cofactors (Mukhopadhyay et al. 2019) that play an essential role in catalysis or have important structural or regulatory roles (Torrance, MacArthur, and Thornton 2008).

Methods for predicting ligand-binding sites (reviewed in (Zhao, Cao, and Zhang 2020)) use a range of different approaches, including sequence conservation (Capra and Singh 2008), structural approaches such as identifying pockets on the protein surface, the combined analysis of sequence and structural information (Capra et al. 2009), and machine learning (Krivák and Hoksza 2018; Jendele et al. 2019; Santana et al. 2020; Jiménez et al. 2017). 3DLigandSite and methods such as firestar (Lopez et al. 2011), FINDSITE (Skolnick and Brylinski 2009; Feinstein and Brylinski 2014), COACH-D (Q. Wu et al. 2018) and FunFOLD2 (Roche, Buenavista, and McGuffin 2013) utilise knowledge of existing binding sites in solved protein structures present in the Protein Databank (PDB) (Armstrong et al. 2020). 3DLigandSite, FINDSITE, FunFOLD2 and COACH-D combine the modelling of protein structure with the identification of homologous proteins in the PDB that have ligands bound to them. These binding sites in identified homologues are then used to infer binding sites in the query protein. By contrast, firestar uses FireDB (Maietta et al. 2014), a database of ligand-binding residues extracted from protein structures in the PDB and also catalytic residues extracted from the catalytic site atlas (Ribeiro et al. 2018).

Here, we present the first major update to the 3DLigandSite webserver. 3DLigandSite (Wass, Kelley, and Sternberg 2010a) was first developed in 2010 to automate an approach that was successfully used in the ligand-binding site experiment in the 8th Critical Assessment of protein Structure Prediction (CASP) (Wass and Sternberg 2009; López, Ezkurdia, and Tress 2009). Over the last eleven years, 3DLigandSite has become widely used, attracting on average 125,000 submissions per annum. 3DLigandSite binding site predictions have been used for a diverse range of purposes, including genome annotation (Nishiyama et al. 2018; Antczak, Michaelis, and Wass 2019), antiviral screening (Kuhlmann

et al. 2017), the analysis of single nucleotide variants associated with disease (Bernkopf et al. 2014; O'Grady et al. 2016; Martell et al. 2017; Chambers et al. 2011) and the development of fluorescent sensors (C. H. Ho and Frommer 2014). Over the last two years, 3DLigandSite binding site prediction has been incorporated into the PDBe Knowledgebase (Varadi et al. 2020), making binding site predictions for protein structures in the PDBe (Armstrong et al. 2020) widely available.

The basic algorithm remains the same in the new version, but it now makes use of the latest sequence searching methods and incorporates machine learning as the final step in the prediction process, which improves prediction accuracy and associates a confidence score with each individual residue predicted to be part of a binding site. This is combined with a new webserver that offers improved functionality for users to investigate the predicted binding sites.

## 2.3 The 3DLigandSite Method

A summary of the 3DLigandSite methodology is outlined in **Figure 2.1**. Users submit either a protein sequence in FASTA format or a protein structure (in PDB format). Where a sequence is submitted, Phyre2 (Kelley et al. 2015) is used to perform template-based modelling, or AlphaFold (Jumper et al. 2021) is used to predict the protein structure. Where the sequence is identical to a protein structure in the AlphaFold model organisms' resource or the PDB, these structures are used. The next step focuses on the identification of ligand-bound structures that are homologous to the query protein. Originally, 3DLigandSite used MAMMOTH (Lupyan, Leo-Macias, and Ortiz 2005) to perform a structural search of the query structure against a structural library of proteins from the PDB, which was a time-consuming step, typically taking between 40-80 minutes. This has been replaced by a sequence-based search using HHSearch (Söding 2005) to screen a sequence library of ligand-bound proteins in the PDB (detailed below), which only takes a few minutes to run. A structural search is provided as an advanced option that the user can choose to perform. Here, TM-Align (Y. Zhang and Skolnick 2005) is used to align each ligand-binding structural database protein structure onto the query structure, where alignments with TM-Scores of 0.6 or above are considered.

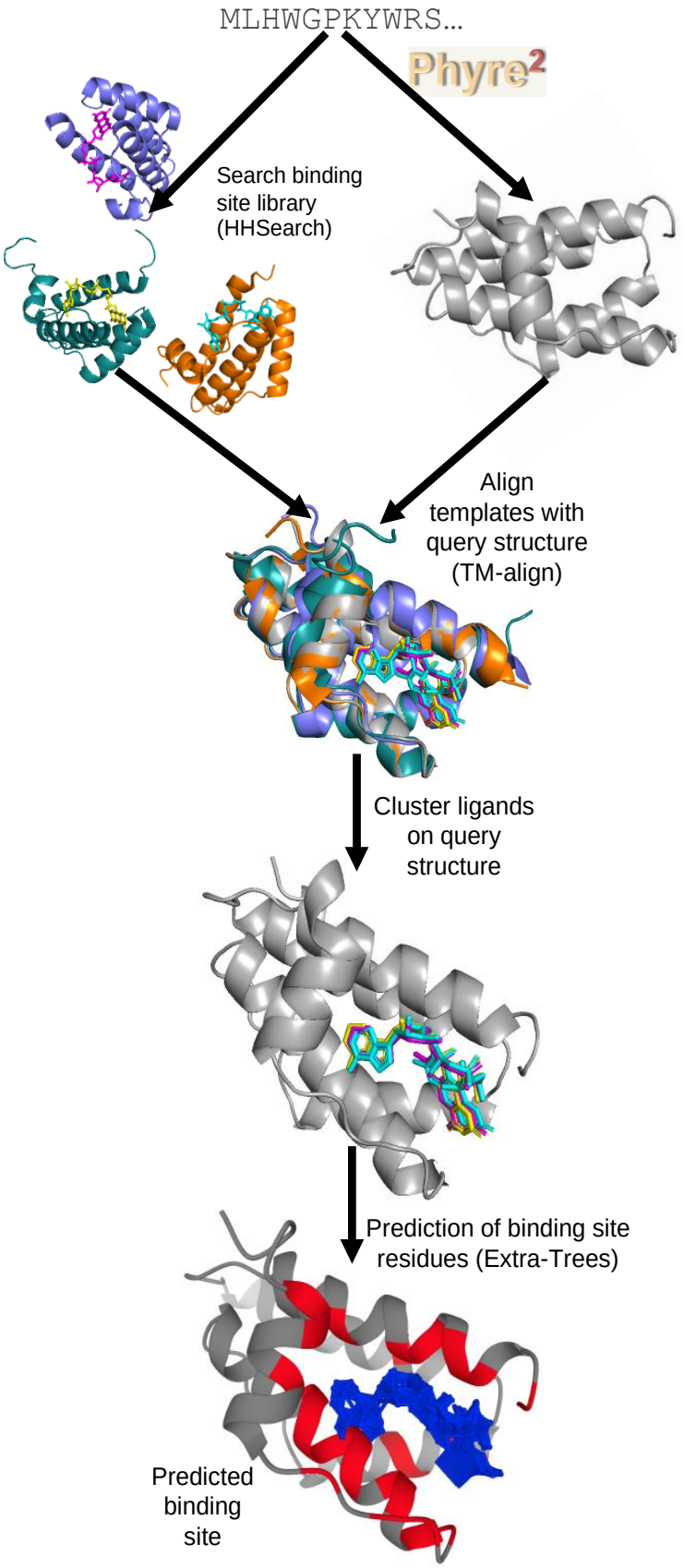

MLHWGPKYWRS…
Phyre2
Search binding site library (HHSearch)
Align templates with query structure (TM-align)
Cluster ligands on query structure
Prediction of binding site residues (Extra-Trees)
Predicted binding site

**Figure 2.1.** An overview of the 3DLigandSite method. Users submit either a protein sequence or structure. Where sequences are submitted Phyre2 is used to model the 3D structure. HHsearch is used to search a sequence library of protein structures with ligands bound. Hits from this search are aligned with the structure of query protein, the ligands present are clustered. Each cluster of ligands represents a potential binding site in the protein. A machine learning classifier is used to predict which of the residues around the cluster are likely to form part of a binding site.

All matches from the sequence search with an HHSearch probability score greater than 80 % are retained, and their protein structures are aligned to the query structure using TM-Align (Y. Zhang and Skolnick 2005). The user can reduce the HHSearch probability cut off if they would like to use less confident matches to the query sequence.

The alignment of the library structures onto the query structure superimposes the ligands present in the library structures onto the query structure. These ligands are then clustered. Originally 3DLigandSite used single linkage clustering to cluster both metal and non-metal ligands, resulting in some very large clusters of ligands. To avoid this, we now generate clusters such that 50 % of each ligand must overlap with at least one of the other ligands in the same cluster. Moreover, metal and non-metal ligands are now separately clustered to make individual predictions of metal and non-metal binding sites.

The prediction of residues that form the binding site based on each cluster of ligands is the final step in the prediction process. Each cluster may contain multiple different ligands or many instances of the same (or similar) ligands in different poses. We also do not know what ligand actually binds to the query protein at this location. To predict the residues that form a binding site at this location, 3DLigandSite originally predicted any residue within 0.8 Å of at least 25 % of the ligands in a cluster to be part of the binding site. We have now introduced a Logistic Regression classifier (detailed below) to perform this final prediction step. This produces accurate predictions and a confidence score associated with each residue in the predicted binding site.

### 2.2.1 Generation of the library of ligand-bound protein structures

To generate the library of biologically relevant protein binding sites, protein structures were extracted from the PDB and filtered to retain only those containing ligands classed as

cognate by FireDB (Maietta et al. 2014). The protein structures were clustered, and the ligands from proteins in each cluster mapped onto a representative structure to reduce search time. The amino acid sequences of the retained structures were clustered using CD-HIT (Fu et al. 2012) using an 80 % sequence identity threshold. The protein models in each cluster identified were then aligned against the cluster representative (obtained from CD-HIT) using TM-align (Y. Zhang and Skolnick 2005), and the ligands were superimposed onto the representative structure and retained. An HHSearch (Steinegger et al. 2019) sequence database was then built from the representative sequences for searching user-submitted protein sequences against.

### 2.2.2 Calculating residue conservation

To calculate residue conservation, HHBlits (Remmert et al. 2012) was used to search the query sequence against the UniClust30 database (Mirdita et al. 2017). The multiple sequence alignment was then used to calculate the Jensen-Shannon Divergence (Capra and Singh 2007b) conservation score.

### 2.2.3 Machine learning-based prediction of binding site residues

The machine learning step was introduced to accurately predict which residues are most likely to be part of the binding site around a cluster of ligands. An equal number of binding and non-binding residues on the query protein were used for training and testing. For each of these residues, a set of features was extracted and converted to a 0-1 range (**Table 2.1**). A number of features were considered for best determining binding propensity. The features include distance measurements to the ligand cluster, residue conservation and amino acid properties such as charge, hydrophobicity and van der Waals volume (**Table 2.1**). Solvent accessibility scores were obtained from ProAct2 (Tjelvar S.G. Olsson, Antony M.E. Churchill, William R. Pitt 2012). Distance-based features were calculated such as the minimum, maximum, and average distance of each residue to ligands in the cluster, and the percentage of ligands in the cluster within 0.8 Å + Van der Waals radii of the amino acid. Univariate feature selection was used to identify ligand contacts, the three distance features, residue conservation in metals and non-metals as the most informative features for predicting ligand binding, with the negatively charged feature included for metals as well. A single distance metric was selected to avoid overtraining on a similar feature, resulting in the ligand contacts, minimum ligand distance, negatively charged, and residue

conservation as our chosen features to predict ligand binding. The negatively charged amino acid feature showed a preference for metal-binding and is in keeping with the literature (Al-Karadaghi n.d.; Barber-Zucker, Shaanan, and Zarivach 2017). Solvent accessibility is not considered for the machine learning step as it is not considered in commonly used ligand binding calculations; however, it may be useful to the user to identify easily targetable positions, and therefore this information is retained for the user.

| Feature | Value range |
|---|---|
| JS Divergence Score (conservation) | 0-1 |
| ***Amino acid properties*** | |
| Hydrophobicity | 0-1 |
| Polar uncharged | 0/1 (1 if polar uncharged, 0 otherwise) |
| Isoelectric point | 0-1 |
| Aromatic | 0/1 (1 if aromatic, 0 otherwise) |
| Van der Waals volume | 0-1 |
| Positive | 0/1 (1 if positive, 0 otherwise) |
| Negative | 0/1 (1 if negative, 0 otherwise) |
| Amino acid | Each amino acid 1 if present, if not 0. I.e., Is tyrosine? 1, Is alanine? 0 |
| ***3DLigandSite features*** | |
| Min ligand distance | 0-1 (Value/10, any value greater than 1 is scored as 1) |
| Max ligand distance | 0-1 (Value/10, any value greater than 1 is scored as 1) |
| Average ligand distance | 0-1 (Value/10, any value greater than 1 is scored as 1) |
| Ligand Contacts | 0-1 (Percentage of ligands that the residue is within 0.8/0.4Å + VDW of/100) |

**Table 2.1. Features used in machine learning.**

The scikit-learn Python package was used to train Support Vector Machines (SVMs) (Cortes and Vapnik 1995), Extra-Trees, Logistic Regression, and Random Forest classifiers.

The data were then fitted with optimum parameters from 100 random iterations and three cross-validation steps using GridSearchCV within scikit-learn. A randomly generated 80:20 train-test split was used to fit the models.

The training and test sets comprise monomers with cognate ligands bound. These structures were identified by filtering the PDB, clustering their sequences using MMseqs2 (Steinegger and Söding 2017) at a maximum sequence identity of 40 %. This resulted in 5223 metal and 4995 non-metal binding sites. A subset of 1600 metal and 1573 non-metal binding sites were randomly selected for testing and training from ~2000 protein structures. The remaining binding sites were used as a validation set to evaluate performance on the trained classifiers (that had not been used in training) **(Figure 2.2**). The split of the data in this way was to ensure there was enough data to accurately validate this method, whilst also providing lots of structures for the classifiers to learn from. The PDB identifiers and chains of all sequences used are provided (**Supplementary Table 2.1**).

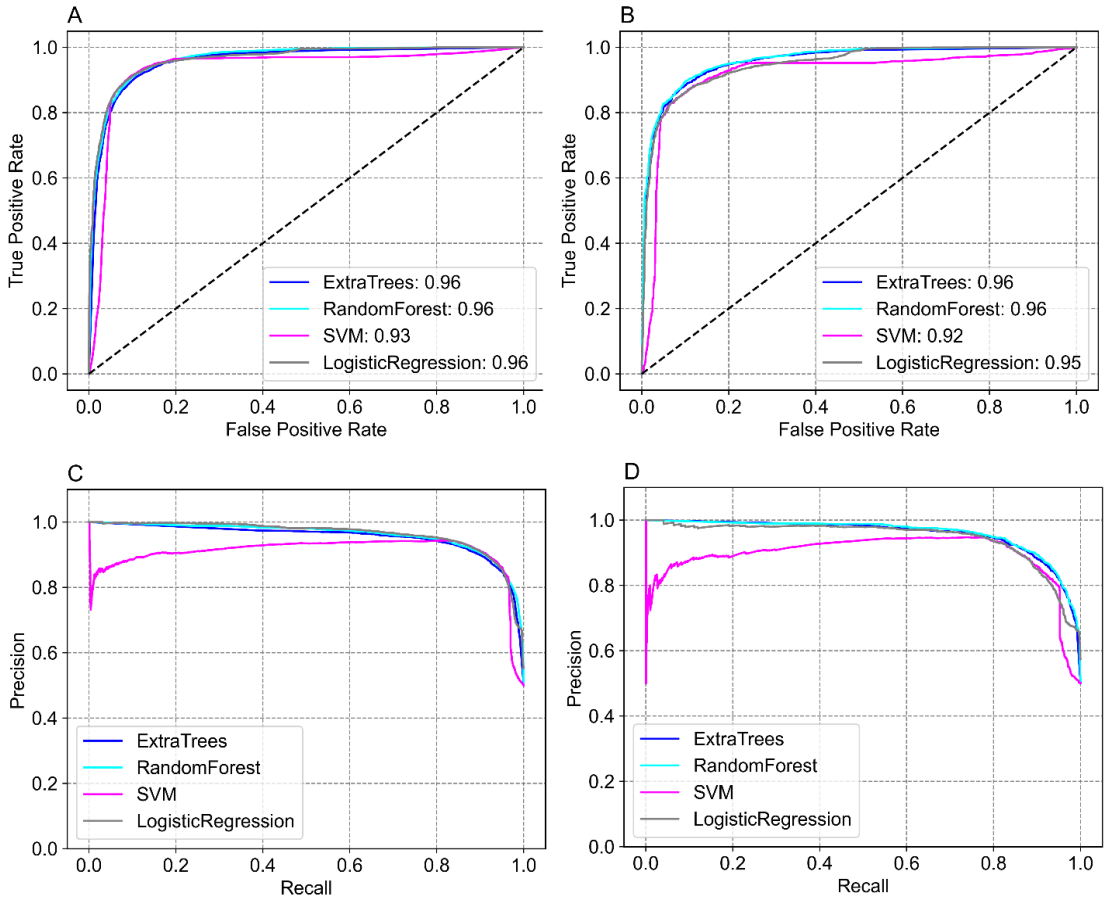



**Figure 2.2. Benchmarking of 3DLigandSite on the cross-validation training-testing data**. Receiver operator characteristic (ROC) curves and Precision-Recall graphs are shown for the prediction of binding sites of non-metal (A and C) and metal (B and D) ligands.

The residues used for each binding site were obtained by selecting all residues of the query protein. From this set, binding residues were classed as all residues within VDW radii + 0.8 Å of the ligand present in the PDB structure, with all other residues classed as non-binding. This resulted in 1976 and 6950 metal and non-metal binding residues, respectively, and an equal number of randomly selected non-binding residues (**Table 2.2**), providing the positive and negative examples required for training the machine learning classifiers.

## 2.4 Evaluating 3DLigandSite Performance

The performance of 3DLigandSite was assessed using the validation set (see methods section), which contained 59203 and 16166 (**Table 2.2**) non-metal and metal-binding residues, respectively, that had not been used in the testing or training of the classifiers. Performance was assessed using multiple measures of precision, sensitivity (recall), and a Receiver Operator Characteristic (ROC) graph.

| | **Number of binding sites** | **Number of binding residues** | **Number of non-binding residues** |
|---|---|---|---|
| **Metal binding sites** | | | |
| Train/test | 1600 | 1976 | 1976 |
| Validation | 2889 | 16166 | 825376 |
| | | | |
| **Non-metal binding sites** | | | |
| Train/test | 1573 | 6950 | 6950 |
| Validation | 3527 | 59203 | 1044947 |

**Table 2.2. Testing, training, and validation dataset sizes.** The training/test set was used for five-fold cross validation using an 80:20 split, with 80% of the data used for training and testing on the remaining 20%.

The Logistic Regression classifier performed best on the non-metal binding sites, with a AUROC of 0.99, though a similar performance is observed for ExtraTrees and RandomForest classifiers (**Figure 2.3A**, **Table 2.3**). For metals, the Logistic Regression classifier performed best with the computed area under the receiver operating characteristic curve (ROCAUC) score of 0.99 (**Figure 2.3C**, **Table 2.2**). These results demonstrate that 3DLigandSite makes accurate predictions of binding site residues. For example, it is able to obtain 92 % recall at 75 % precision for non-metal binding sites (**Figure 2.3C**) and 52 % recall at 75 % precision for metal-binding sites (**Figure 2.3D**). The performance of the metal-binding predictor was

worse than that of non-metals. This is likely due to a number of reasons: metals typically bind to fewer residues than non-metals due to the size difference between the two, leading to comparatively more false positives when variation in the atomic positions of the metals in clustering results in more amino acids labelled as binding. In addition, the validation data set includes proteins that are difficult to predict due to the lack of sequentially homologous proteins, possibly indicating the metal binding sites are more structurally conserved than sequentially. Another reason for the reduced performance in metal-binding prediction than non-metal-binding prediction could be that non-metal clusters of ligands occur more frequently, lowering the ranking of the metal-binding site, and therefore resulting in a worse performance in the validation set.

**A. Non-metal ligands**

| | Test | | | Validation Seq-Search | | | Validation Struc-Search | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Model*** | **Precision** | **Recall** | **AUC *** | **Precision** | **Recall** | **AUC** | **Precision** | **Recall** | **AUC** |
| ***ET*** | 0.9 | 0.9 | 0.96 | 0.85 | 0.92 | 0.97 | 0.83 | 0.89 | 0.92 |
| ***RF*** | 0.9 | 0.9 | 0.96 | 0.86 | 0.92 | 0.98 | 0.84 | 0.89 | 0.93 |
| ***SVM*** | 0.91 | 0.91 | 0.93 | 0.9 | 0.93 | 0.94 | 0.87 | 0.89 | 0.9 |
| ***LogR*** | 0.91 | 0.91 | 0.95 | 0.91 | 0.92 | 0.99 | 0.88 | 0.89 | 0.95 |

*Average from 5-Fold CV

**B. Metal ligands**

| | Test | | | Validation | | | Validation Struc-Search | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Model*** | **Precision** | **Recall** | **AUC *** | **Precision** | **Recall** | **AUC** | **Precision** | **Recall** | **AUC** |
| ***ET*** | 0.89 | 0.89 | 0.96 | 0.71 | 0.89 | 0.96 | 0.71 | 0.78 | 0.84 |
| ***RF*** | 0.9 | 0.9 | 0.96 | 0.71 | 0.89 | 0.96 | 0.72 | 0.78 | 0.85 |
| ***SVM*** | 0.88 | 0.88 | 0.92 | 0.79 | 0.89 | 0.91 | 0.73 | 0.78 | 0.7 |
| ***LogR*** | 0.88 | 0.88 | 0.95 | 0.79 | 0.89 | 0.98 | 0.73 | 0.78 | 0.9 |

*Average from 5-Fold CV

**Table 2.3. Benchmarking machine learning performance.** The performance of four classifiers on datasets are summarised here. ET = Extra-Trees, RF = Random Forest, SVM = Support Vector Machine, LogR = LogisticRegression. Results for **A)** Non-metal ligands and **B)** Metal ligands.

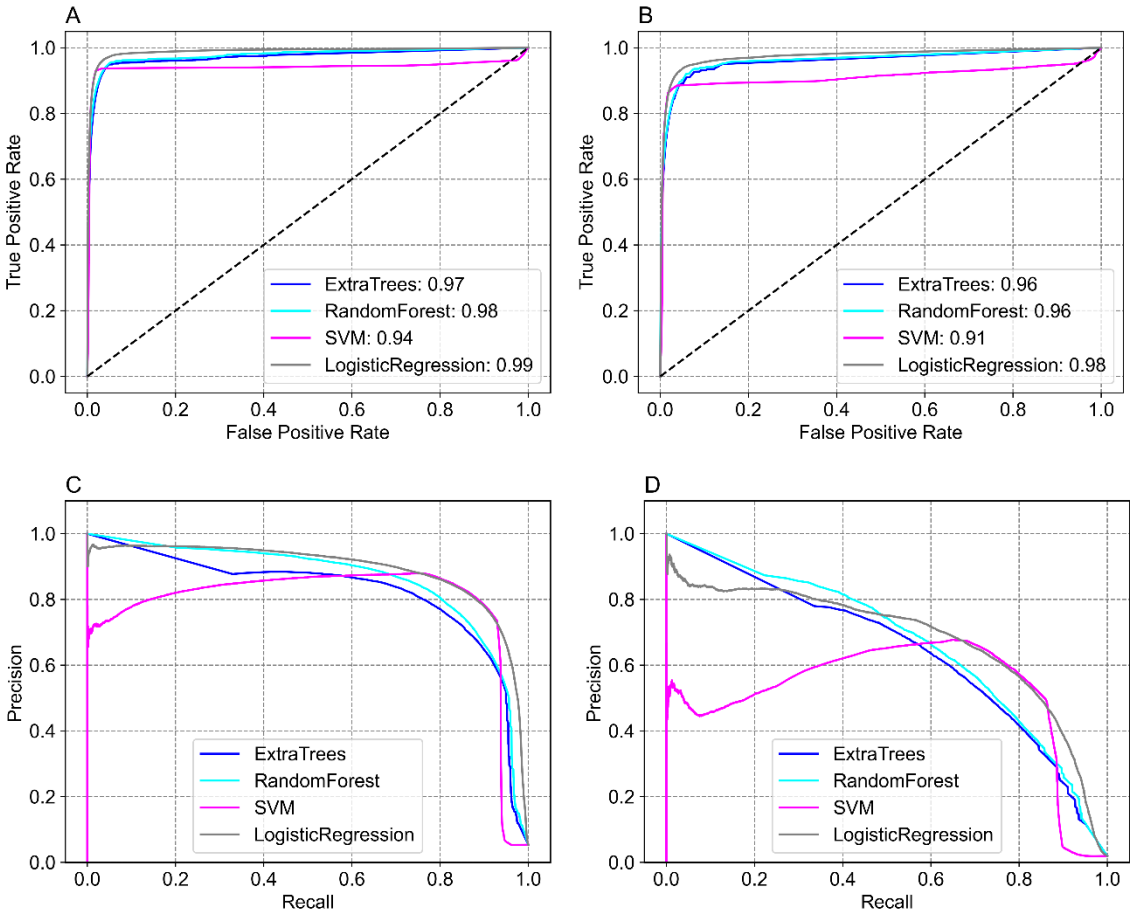



**Figure 2.3. Benchmarking the 3DLigandSite machine learning classifier.** Receiver operator characteristic (ROC) curves and Precision-Recall graphs are shown for the prediction of binding sites of non-metal (A and C) and metal (B and D) ligands.

The ability of the classifiers to distinguish binding and non-binding positions was plotted showing the distribution of probability scores assigned to binding and non-binding residues in the validation set for metals and non-metals (**Figure 2.4**). We evaluated the performance of 3DLigandSite on the CASP8, CASP9 and CASP10 datasets, totalling 70 targets. The sequence-based homology search resulted in MCC, precision and recall scores of 0.73, 0.65, and 0.85, respectively, for 70 targets. Using the structural-search option resulted in MCC, precision and recall scores of 0.72, 0.67, and 0.8, respectively, for the 70 targets. Structural search results at a range of TM-Scores is shown in **Table 2.4**. This was surprising given recent studies that have shown that structural searches are better at identifying related protein structures (J. Chen et al. 2018). Given, the extra time taken to perform the

structural search (circa 4 hours per submission) and the slightly poorer prediction, we generally recommend using the default sequence-based search.

| Sequence-based | MCC | Precision | Recall | Targets |
|---|---|---|---|---|
| HHSearch Prob 75% | 0.73 | 0.65 | 0.85 | 70 |
| **Structure Based** | | | | |
| TMScore 0.5 | 0.65 | 0.62 | 0.71 | 70 |
| TMScore 0.6 | 0.72 | 0.67 | 0.8 | 70 |
| TMScore 0.7 | 0.71 | 0.68 | 0.77 | 70 |
| TMScore 0.8 | 0.69 | 0.65 | 0.77 | 70 |

**Table 2.4. CASP Assessment**. The performance of the sequence-based and structure-based 3dligandsite tool on the CASP dataset.

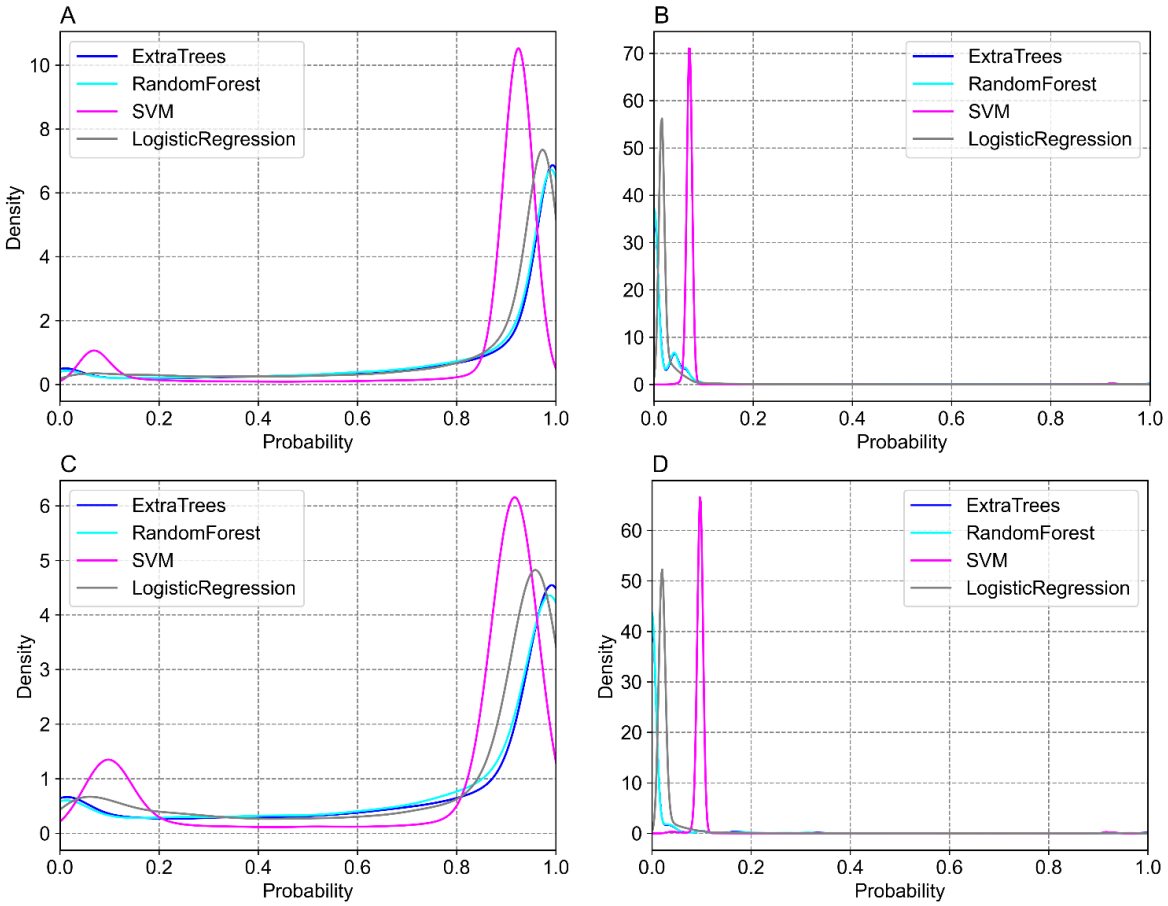



**Figure 2.4. Assessment of the average probability scores assigned to binding and non-binding residues in the metal and non-metal residues in the validation set.** Non-metal binding residues (A), non-metal not binding residues (B), metal binding residues (C), metal not binding residues (D).

## 2.5 The 3DLigandSite Web Server

The 3DLigandSite web server is available at https://www.wass-michaelislab.org/3dligandsite. The web server is free to all without a login requirement. Users can select to submit either a protein sequence (in FASTA format) or a protein structure (in PDB format). Where a sequence is submitted, the first step of the prediction process is to model the protein structure using Phyre2 (Kelley et al. 2015) or AlphaFold (Jumper et al. 2021). The runtime for such submissions is longer for these methods as modelling protein structures is a time-consuming process. Any sequences submitted that are equivalent to the sequences of proteins structures present in the PDB or AlphaFold model organisms' dataset are automatically matched to these structures and these structures are used for the prediction. Where users submit a protein structure, the runtime is typically less than five minutes using default settings. Users who provide an email address receive an email upon submission and also once the results are ready for viewing. The web server includes a help section that provides recordings that work users through both the submission process and also interpretation of data in the results pages.

### 2.4.1 Results Output

3DLigandSite results pages are split into three main sections. Results are initially presented as a sequence view (**Figure 2.5A**), which shows the amino acid sequence of the submitted protein, residue conservation, and a row for each of the clusters of ligands that have been identified as potential binding sites (**Figure 2.5**). This provides users with an easily interpretable view of the predicted binding sites.

The second section of results shows the cluster table, which includes details of the clusters identified, the number of ligands present in each cluster and the number of structures that these ligands come from. The ligands are represented by the three-letter codes from the mmCIF dictionary (Adams et al. 2019) and are linked to the small molecule details in the PDB (**Figure 2.5**). Clusters are sorted according to the number of ligands present in the cluster. There is greater confidence that a cluster represents a binding site when there is evidence for this from multiple protein structures. The second table in this section contains a tab for each ligand cluster and lists all of the residues predicted to be in the binding site along with the conservation score, solvent accessibility, and the probability calculated by the classifiers.

The final section of the results page contains a Mol* molecular viewer (www.molstar.org) (Sehnal et al. 2018) that by default displays the protein structure in a cartoon format along with the ligands in the top-ranked cluster, highlighting the predicted binding site residues for this cluster in red (**Figure 2.5**). The Mol* viewer enables users to inspect the predicted binding sites within the protein structure and offers multiple features for exploring the structure. The 3DLigandSite control panel to the right of the viewer provides easy to use functions such as changing the colour or format of the display of the ligands and the protein structure. Further functionality is available via the Mol* built-in options shown on the top right of the viewer. The control panel also includes a button enabling users to generate publication-quality images of the current display in the viewer.

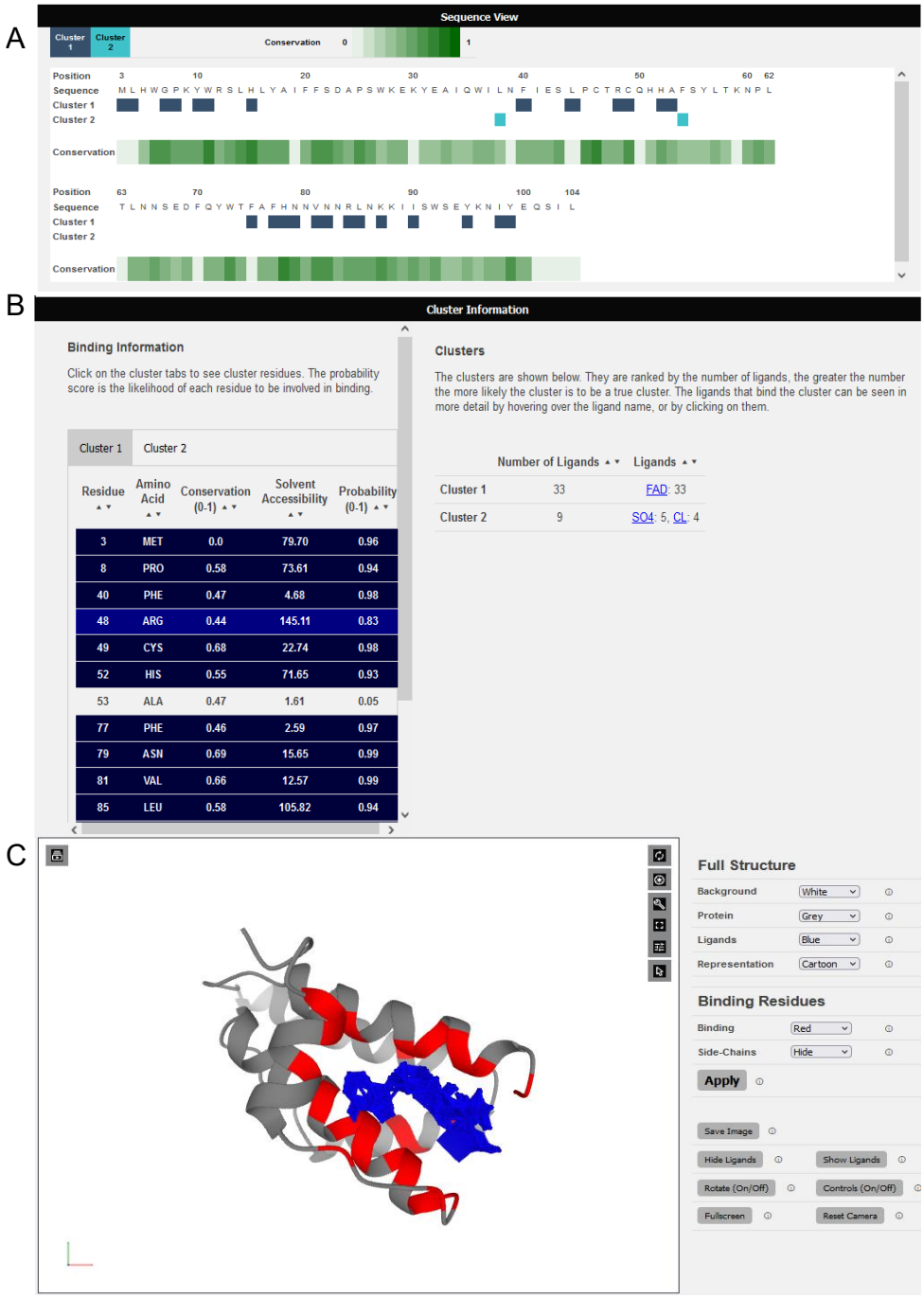



**Figure 2.5.** Viewing results on the 3DLigandSite web server of T0396 from the CASP assessments, PDBID: 1JR8. Results are presented in 3 main sections: A) a sequence view, which maps sequence conservation and the different clusters identified onto the protein sequence. B) Details of the clusters, including the number of ligands and type of ligand are displayed as well as a table listing the residues predicted to form the binding site for each cluster. C) The structural analysis section includes a Mol* molecular viewer to visualise the protein, the predicted binding site and the clusters used to make the predictions. A separate control panel (on the right) enables users to easily modify the display.

## 2.6 Use Cases

As set out in the introduction, 3DLigandSite predictions have been widely used for a range of different biological and biomedical purposes (Wass, Kelley, and Sternberg 2010a; López, Ezkurdia, and Tress 2009; Wass and Sternberg 2009; Antczak, Michaelis, and Wass 2019; Nishiyama et al. 2018; Kuhlmann et al. 2017; Chambers et al. 2011) (20-26). For example, with widespread use of sequencing technologies, there is extensive interest in the analysis of non-synonymous single nucleotide variants (nsSNVs), specifically to identify those nsSNVs that may alter protein structure and function and be associated with a phenotype such as a disease. Thus, 3DLigandSite has been applied to analyse such nsSNVs for a range of diseases, from liver disease to cardiomyopathies.

One application has been to study nsSNVs present in individuals with cystinuria, which is caused by variants in two genes, SLC7A9 and SLC3A1, that encode a dimeric amino acid transporter (Thomas et al. 2014). Cystinuria is caused by variants that affect the ability of the transporter to transport cystine into cells, which results in the formation of kidney stones. In a recent study (Martell et al. 2017), 3DLigandSite was used to model the structure and ligand binding sites of the two encoded proteins and to analyse how the set of nsSNVs observed in a cohort of patients may affect transporter function and be linked with the severity of the disease that patients experienced.

## 2.7 Concluding Remarks

The 3DLigandSite web server provides free access to an easy-to-use resource for modelling small molecule binding sites in proteins. This widely used resource has been extensively updated to provide improved functionality and to reduce the run time of user submissions. Our benchmarking demonstrates that 3DLigandSite can obtain high recall with high precision, therefore accurately predicting binding sites for users to apply to the proteins they are researching.

# Chapter 3: Identification of sequence changes in myosin II that adjust muscle contraction velocity

Chloe A. Johnson*, Jake E. McGreig* et al.

"Identification of sequence changes in myosin II that adjust muscle contraction velocity". PLOS, https://doi.org/10.1371/journal.pbio.3001248

My contribution to this work was as follows:

1. Wrote all the scripts for calculating mass vs percentage identity.
2. Identified myosin sequences to be used in these analyses.
3. Adapted the scripts for residue transition plots.
4. Generated figures of bioinformatics analyses.
5. Contributed to writing the manuscript.
6. Analysed results.

## 3.1 Abstract:

The speed of muscle contraction is related to body size; muscles in larger species contract at slower rates. Since contraction speed is a property of the myosin isoform expressed in a muscle, we investigated how sequence changes in a range of muscle myosin II isoforms enable this slower rate of muscle contraction. We considered 798 sequences from 13 mammalian myosin II isoforms to identify any adaptation to increasing body mass. We identified a correlation between body mass and sequence divergence for the motor domain of the four major adult myosin II isoforms (β/Type I, IIa, IIb, IIx), suggesting that these isoforms have adapted to increasing body mass. In contrast, the non-muscle and developmental isoforms show no correlation of sequence divergence with body mass. Analysis of the motor domain sequence of β-myosin (predominant myosin in Type-I/slow and cardiac muscle) from 67 mammals from two distinct clades identifies 16 sites, out of 800, associated with body mass ($p_{adj}<0.05$) but not with the clade ($p_{adj}>0.05$). Both clades change the same small set of amino acids, in the same order from small to large mammals, suggesting a limited number of ways in which contraction velocity can be successfully manipulated. To test this relationship, the nine sites that differ between human and rat were mutated in the human β-myosin to match the rat sequence. Biochemical analysis revealed that the rat-human β-myosin chimera functioned like the native rat myosin with a two-fold increase in both motility and in the rate of ADP release from the actin.myosin cross-bridge (the step that limits contraction velocity). Thus, these sequence changes indicate adaptation of β-myosin as species mass increased to enable a reduced contraction velocity and heart rate.

## 3.2 Introduction:

Proteins can adapt over time, tuning their function to the specific needs of the organisms in which they are expressed. That is, the same protein expressed in different organisms (orthologues) may have distinct properties to better suit the needs of each organism. This is well established for muscle contraction where the maximum shortening velocity, $V_0$, is a property of the myosin isoform expressed in the tissue. $V_0$ varies more than 5-fold for muscles expressing the same myosin isoform in different mammals. For example, the $V_0$ of rat and human Type I/β-cardiac myosin (hereafter referred to as β-myosin) are 1.42 and 0.33 μm $s^{-1}$ half $sarcomere^{-1}$ respectively (Pellegrino et al. 2003). Across a range of mammals, $V_0$ for each muscle myosin isoform is inversely related to the size of the mammal; larger mammals have slower contracting muscles (Luana Toniolo et al. 2004; Pellegrino et al. 2003). This can most commonly be observed in cardiac tissue where heart rate (related to the contraction velocity, see below) varies widely and is slower for larger mammals (Savage et al. 2007).

Since $V_0$ is a property of the myosin isoform expressed in the tissue, there are expected to be changes in the myosin sequence that have resulted in a myosin with the needed velocity. However, changes in sequence may be occurring for reasons other than adjusting the velocity. Here we ask if it is possible to define which sequence changes tune $V_0$ for an individual myosin isoform.

Several attempts have been made to identify the sequence changes associated with changes in the myosin isoforms and the associated changes in shortening velocity. These used the relatively small numbers of sequences available at the time or considered the differences between paralogues of muscle myosin IIs. The studies focussed on non-conservative substitutions in sequence (Bottinelli and Reggiani 2000) or on the variable surface loops (Schiaffino and Reggiani 2011; Luana Toniolo et al. 2004) and identified areas of interest, but did not locate specific residues nor how they might influence contraction velocity. Large numbers of mammalian myosin sequences are now available from various genome projects. Using this large data set, we test the hypothesis that muscle myosin-II isoforms from mammals have adaptations in protein sequence associated with mean body mass and $V_0$.

We have examined 798 mammalian myosin II sequences, from 13 myosin II isoforms and find that the myosin isoforms found in adult sarcomeric muscle (Type II fast muscle isoforms IIa, IIb and IIx and β; see **Table 3.1** for definition of isoforms) have a much higher variation in sequence than the non-muscle myosins (NMA and NMB). This is consistent with an adaptation of contraction velocity to body mass occurring only in the adult muscle myosins and not in the non-muscle cellular myosin which are unaffected by overall species size.

| Gene name | Heavy Chain | Muscle Type | Isoform | Short name | No complete sequences | Motor | | Tail | |
|---|---|---|---|---|---|---|---|---|---|
| **Major adult sarcomeric muscle myosins** | | | | | | $R^2$ | Gradient (%/log(kg)) | $R^2$ | Gradient (%/log(kg)) |
| *MYH1* | MyHC-1 | Fast | Skeletal IIb | IIb | 51 | 0.48 | -0.52 ± 0.20 | 0.58 | -0.75 ± 0.12 |
| *MYH2* | MyHC-2 | Fast | Skeletal IIa | IIa | 77 | 0.44 | -0.53 ± 0.10 | 0.35 | -0.43 ± 0.11 |
| *MYH4* | MyHC-4 | Fast | Skeletal IID/X | IIx | 41 | 0.84 | -0.84 ± 0.11 | 0.85 | -0.68 ± 0.09 |
| *MYH6* | MyHC-6 | Cardiac | α cardiac | α | 65 | 0.3 | -0.36 ± 0.29 | 0.35 | -0.31 ± 0.11 |
| *MYH7* | MyHC-7 | Cardiac/Slow | β-cardiac | β | 67 | 0.63 | -0.72 ± 0.11 | 0.03 | -0.06 ± 0.16 |
| **Specialist adult sarcomeric myosins** | | | | | | | | | |
| *MYH13* | MyHC-13 | Fast | Extraocular | EXOC | 60 | 0.18 | -0.32 ± 0.10 | 0.21 | -0.72 ± 0.2 |
| *MYH7b* | MyHC-7b | Slow | Slow Tonic | SlowT | 72 | 0.05 | -0.08 ± 0.06 | 0.03 | -0.063 ± 0.06 |
| **Developmental muscle isoforms** | | | | | | | | | |
| *MYH3* | MyHC-3 | Developmental | Embryonic | EMB | 60 | 0.13 | -0.09 ± 0.04 | 0.22 | -0.31 ± 0.12 |
| *MYH8* | MyHC-8 | Developmental | Perinatal | PERI | 54 | 0.01 | -0.04 ± 0.11 | 0.21 | -0.28 ± 0.12 |
| **Non sarcomeric myosins** | | | | | | | | | |
| *MYH9* | MyHC-9 | Non-muscle | Non-muscle A | NMA | 59 | 0.03 | -0.033 ± 0.036 | 0.14 | -0.14 ± 0.06 |
| *MYH10* | MyHC-10 | Non-muscle | Non-muscle B | NMB | 69 | 0.11 | 0.042 ± 0.022 | 0.09 | -0.08 ± 0.05 |
| *MYH11* | MyHC-11 | Smooth muscle | Smooth Muscle | SM | 71 | 0.52 | -0.37 ± 0.07 | 0.08 | -0.17 ± 0.09 |
| *MYH14* | MyHC-14 | Non-muscle | Non-muscle C | NMC | 52 | 0.10 | -0.21± 0.11 | 0.203 | -0.65± 0.35 |

**Table 3.1. Myosin II Isoforms considered.** The myosin isoforms, number of sequences, and overview of mass vs sequence identity results. MyHC 13 & 7b are labelled *specialized* because unlike the other sarcomeric forms they are not found alone in a specific muscle type but only in combination with other isoforms.

We go on to examine in greater detail the sequence differences within the more widely studied mammalian β-myosin (MYH7, expressed in slow Type I and cardiac muscle) to establish if it is possible to define a relationship between amino acid sequence and velocity of contraction. We examine 67 mammalian β-myosin sequences from 2 clades and identify a group of 16 amino acids which have the strongest association with the size of the mammal and not the clade. Four are in the hypervariable regions (Loop2 and the N-terminus). Of the 12 remaining amino acids, nine differ between human and rat β-myosin. We test our hypothesis that these nine residues influence $V_0$ through the construction and subsequent biochemical characterisation of a rat-human β-myosin chimera (hereafter referred to as chimera) in which the nine rat residues are exchanged into the human β-myosin.

## 3.3 Materials and Methods

Protein sequences were extracted from RefSeq (Pruitt et al. 2014) and UniProt (Consortium 2017) as listed in **Table S3.1**. To ensure that each sequence corresponded to the specific myosin isoform we used both the UniProt and eggNOG (Huerta-Cepas et al. 2016) annotations and for the model species available in APPRIS (Rodriguez et al. 2018) selected the principal isoform. Isoform determination was also checked with the gene tree (**Figure S3.6**) of all sequences. There were 15 sequences in which eggNOG gave a different assignment to UniProt (six NMA, seven α and one IIa); these are described in the supplementary information.

Incomplete sequences were excluded from our analysis because sequence gaps could have a major effect on the sequence comparison for such closely related isoforms.

For each isoform, the protein sequences were divided into the Motor (1-800, β-myosin numbering) and Tail (842 – 1936) regions. For each region of each myosin, the sequences were aligned using Clustal Omega (Sievers and Higgins 2014a); the resulting multiple sequence alignment was used to construct a percentage identity matrix between the species. Sequence identity was used rather than sequence similarity as we are considering small changes (>93% identity, >98% similarity) within the isoform and substitutions that would normally be classed as similar (e.g. aspartate to glutamate) may be relevant.

The masses of each species were extracted from a wide range of information sources and are listed in **Table S3.2**. Where the literature provides a range of adult body masses the arithmetic mean was selected. To compare sequence divergence against either evolutionary time or animal body mass, the relevant matrices were plotted against each other. Amino acid sequences for the β-myosin motor from 67 different mammalian species comprised organisms from the clades Euarchontoglires (32), Laurasiatheria (30), Metatheria (4) and Afrotheria (1).

Protein sequence divergence was plotted against the species mass. A robust regression was fitted to reduce the weighting of the outliers. This was done by minimizing absolute difference rather than squared distance which should reduce the amount of under- and over-estimation in value difference caused by the square.

Residues associated with cardiomyopathies were obtained from Parker & Peckham (Francine Parker and Peckham 2020) and compared to the residues associated with body mass by considering exact matches to residues associated with cardiomyopathy. These were grouped into hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM), and other.

### 3.3.1 Phylogenetic Independent Contrasts

Two statistical methods were used to create phylogenetic trees, maximum likelihood (ML) and Bayesian inference methods, using programs Randomized Axelerated Maximum Likelihood (RAxML) (Stamatakis 2014) and Bayesian Evolutionary Analysis Sampling Trees (BEAST) and its corresponding user interface BEAUtI (Drummond et al. 2012), to increase reliability as species sequence identity is high. The RaxML trees were generated using the CIPRES Science Gateway (M. A. Miller, Pfeiffer, and Schwartz 2010). TreeAnnotator (Drummond et al. 2012) was then used to generate a consensus tree for each set of 10000 trees produced by BEAST. *Monodelphis domestica* (opossum) was selected as an outgroup. The ML and Bayesian trees were compared using Compare2Trees (Nye, Liò, and Gilks 2006), with black nodes on the trees indicating branches where the two trees disagree. Trees were drawn using FigTree (Rambaut 2014).

Phylogenetically independent contrasts (Felenstein, 1985) were performed using the ape (Paradis, Claude, and Strimmer 2004) and phytools (Revell 2012) packages in R. This analysis used species trees generated from TimeTree from the species present in our analyses for the β, α, IIa, IIb, and IIx isoforms.

### 3.3.2 Statistical analysis:

For each position in the alignment that had more than one amino acid present, species masses were compared between the two highest frequency amino acids at that position using the Mann-Whitney U test, a nonparametric two-sample test. Multiple testing was accounted for by applying the Bonferroni correction (Bonferroni 1936). Bonferroni correction is used to correct the error rate when multiple tests are being performed simultaneously, as such is the case with this data. It reduces Type 1 errors and increases the confidence in null hypothesis rejection. To avoid very imbalanced comparisons, the analysis was not run if the frequency of the second amino acid was less than 10% of the

frequency of the most frequent one. Where more than two amino acids were present at an alignment position, only the two most frequent amino acids were considered. See **Figure S3.3** for details of sites with more than 2 amino acids.

Alignment positions were divided into three groups: those with an adjusted p-value ($p_{adj}$) less than 0.01 (with Bonferroni correction applied this is equivalent to a p-value of $p= 9.50x10^{-04}$), those with $0.01 < p_{adj} < 0.05$ (5 % significance threshold equivalent to $p=9.6x10^{-4}$), and those with a $p_{adj} > 0.05$. In addition, the two highest frequency amino acids were coded as 0 and 1 and a logistic regression model was fitted with log(mass) as the explanatory variable (**Figure 3.3** & **Figure S3.2**). In order to overlay these residue plots, as the coding of the amino acids as 0 and 1 was arbitrary, it was done in such a way that the slope of the fitted logistic regression line was positive (**Figure 3.3**). The value of mass at which the two amino acids were predicted to be equally likely to occur was estimated from the regression line.

For each alignment position, a 2x2 table was constructed classifying the species by amino acids present (most frequent and second most frequent) and clade. Fisher's exact test was used to test whether these two factors were associated. The residue and $-\log_{10}$ of the P-value from the Fisher's exact test were plotted to identify residues for which the amino acid variation was likely to have resulted from clade associated changes. The residue and $-\log_{10}$ of the p-value from the Mann-Whitney U test were also plotted to determine when residue variation was likely attributed to mass changes. Finally, the $-\log_{10}$ p-values obtained from both tests were plotted against each other. For each of these plots, lines at positions of the Bonferroni adjusted p-values 0.01, and 0.05 were added to assess the confidence in each residues association with mass or clade.

All statistical analyses were run in R (R Core team 2018).

### 3.3.3 Molecular Biology of the chimera

A pUC19 plasmid containing the human β-myosin motor domain gene was digested with NsiI and NgoMIV to excise DNA encoding for residues 310 – 599 of the human β-myosin. This region was replaced with a complementary pair of synthetic oligos encoding for the same region, but with the nine amino-acid substitutions listed (Ala326Ser, Ser343Pro, Leu366Gln, Ile421Ala, Thr424Ile, Ala430Ser, Arg434Lys, Phe553Tyr, Pro573Gln). The

subsequent clone was confirmed by sequencing. This chimera gene was cloned into a pShuttle CMV vector to allow recombinant replication-deficient adenovirus production, as previously described (Deacon et al 2012).

### 3.3.4 Protein purification:

The chimera and the human β-myosin motor domains (known as subfragment 1 or S1) were expressed and purified as described previously (Resnicow et al. 2010). Briefly, the adenoviruses were used to infect $C_2C_{12}$ myotubes in culture and resulted in overexpression of recombinant myosin proteins. The heavy chains (residues 1-842) were co-expressed in $C_2C_{12}$ myotubes with His-tagged human ventricular essential light chain. The recombinant proteins also carried the endogenous mouse regulatory light chain (MLC3). This is homologous to subfragment 1, S1, generated by proteolytic digestion of myosin. For motility assays the heavy chain was additionally tagged with an eight residue (RGSIDTWV) C-terminal extension. Cell pellets were homogenized in a low salt buffer and centrifuged, and the supernatants were purified by affinity chromatography using a HisTrap HP 1 ml column. The proteins were then dialyzed into the low salt experimental buffer (25 mM KCl, 20 mM MOPS, 5 mM $MgCl_2$, 1 mM DTT, pH 7.0).

The SNAP-PDZ18 affinity tag used for *in vitro* motility measurements were purified as described in (Huang, Jin. Nagy, Stanislav. Koide, Akiko. Rock, Ronald. Koide 2009; Aksel et al. 2015). SNAP-PDZ18 was expressed through a pHFT2 expression vector, and the plasmid transformed into *E. coli* BL21 DE3 cells. The protein was purified using nickel-affinity chromatography, and dialyzed in PBS.

Actin was prepared from rabbit muscle as described by (Spudich and Watt 1971). The actin was labelled with pyrene at Cys-374 as described in (Criddle, Geeves, and Jeffries 1985). When used at sub-micromolar concentrations the actin was stabilized by incubation in a 1:1 mixture with phalloidin.

Rat β-myosin S1 was prepared from soleus muscle which was dissected immediately post-mortem and stored on ice. The muscle was homogenised into Guba-Straub buffer and left to stir for 30 minutes. After centrifugation at 4600 RPM for 30 minutes at 4 °C, the supernatant was subject to myosin precipitation as described in (Margossian, SS. Lowey 1982). The resulting myosin was digested with 0.1 mg chymotrypsin per ml of solution and

left to stir for 10 mins exactly, at room temperature. To stop the digestion, 0.5 mM phenylmethylsulfonyl fluoride (PMSF) was added and the solution left to stir for 10 minutes. The digested myosin solution was dialysed into the low salt experimental buffer overnight (25 mM KCl, 20 mM MOPS, 5 mM $MgCl_2$, 1 mM DTT, pH 7.0). Precipitated myosin and light meromyosin was pelleted and removed via centrifugation at 12,000 RPM for 10 minutes, with the supernatant containing the purified soleus S1. SDS-Gels of the purified protein were run and compared to the expressed human β-myosin and chimera S1.

### 3.3.5 Stopped-flow spectroscopy

Kinetic measurements for S1 of chimera, human β-myosin and rat soleus myosin were performed as described previously (M. Bloemink et al. 2013; Walklate et al. 2016; Resnicow et al. 2010). Solutions were buffered with 25 mM KCl, 20 mM MOPS, 5 mM $MgCl_2$, 1 mM DTT at pH 7.0, and measurements were conducted at 20 °C on a High-Tech Scientific SF-61 DX2 stopped-flow system. Traces were analysed in Kinetic Studio (TgK Scientific) and Origin.

### 3.3.6 *In vitro* motility assay

Motility assays were performed essentially as described previously (Adhikari et al. 2016; Aksel et al. 2015). Briefly, flow chambers were constructed with coverslips coated with nitrocellulose mounted on glass slides. Reagents were loaded in the following order: 1) SNAP-PDZ18 affinity tag; 2) BSA to block the surface from non-specific binding; 3) S1 of human β-myosin or the chimera with an eight amino acid C-terminal affinity clamp; 4) BSA to wash the chamber; 5) rhodamine-phalloidin-labelled rabbit actin; 6) an oxygen-scavenging system consisting of 5 mg/ml glucose, 0.1 mg/ml glucose oxydase and 0.02 mg/ml catalase 7; 2 mM ATP. Partially inactivated myosin heads in S1 preparations were removed by incubating with a 10-fold molar excess of actin and 2 mM ATP for 15 minutes, then sedimentation at 100,000 RPM for 15 minutes. Supernatant was collected containing active myosin heads. All solutions were diluted into 25 mM imidazole, 25 mM KCl, 4 mM $MgCl_2$, 1 mM EGTA, 1 mM DTT, pH 7.5. Actin filaments were detected using a widefield fluorescence imaging system (described in (M. Johnson, East, and Mulvihill 2014)) with UAPON 100XOTIRF NA lens (Olympus) and QuantEM emCCD camera (Photometrics). The system was controlled and data analysed using Metamorph software (Molecular Devices, Sunnyvale, USA). Assays were performed at 20 °C and were repeated with three fresh protein preparations, with at least three movies of 30 second duration, recorded at a rate

of 0.46 sec per frame. Individual velocities were determined from motile filaments that demonstrated a smooth consistent movement over 10 frames (4.6 sec). 100 individual measured velocities were used to calculate the mean velocity for each recombinant myosin.

## 3.4 Results

All myosin II isoforms contain a motor-domain, and a tail-domain. The N-terminal globular motor domain (approximately 800 amino acids) contains all of the requirements for motor activity, while the C-terminal tail domain (~1200 amino acids, almost entirely a single α helix) drives dimerization and assembly into myosin filaments. Here we focus on the motor domain but also report a summary of a similar analysis of the tail domain.

Our analysis considered 13 different mammalian myosin II isoforms (**Table 1**), the five main adult sarcomeric muscle isoforms (three fast muscle – IIa, IIb, IIx and two cardiac isoforms, α and β, also known as the slow skeletal isoform), two relatively rare adult sarcomeric forms (extraocular and slow tonic), two developmental isoforms (perinatal and embryonic), a smooth muscle isoform and three non-muscle isoforms (non-muscle A, B and C). The non-muscle isoforms provide a negative control, as they act only at the cellular level and are therefore unlikely to be influenced by body mass. We identified all available complete myosin II sequences of these 13 myosin isoforms (see methods; **Table S3.1**). This resulted in a total of 798 sequences, with an average of 61 sequences per isoform (range 41-74 sequences; **Table 3.1**).

### 3.4.1 Adaptation of the myosin II motor domain to increasing body mass

To consider how the motor domain of myosin II isoforms may have adapted to increasing body mass, we investigated whether there was a correlation between sequence divergence and body mass. Mass values were collected from the literature, and we report the arithmetic mean of the range of values reported for each species. The range of masses covers more than six orders of magnitude from 6 g to 10,000 kg (**Table S3.1**). Being one of the smallest species, the mouse was selected as a reference. The sequence identity of each other species with the mouse sequence was plotted against the species body mass (see **Figure 3.1, Table 3.1**, **Figure S3.1**). The analysis can alternatively be done using the myosin sequence from largest species, but this is not always the same for each myosin II data set.

We therefore used the commonly available mouse sequences throughout. Sequence identity rather than conservation was used because within each myosin isoform a low level of divergence was expected (>95 % identity). As such, even conservative changes of amino acids that tune the function of the protein may be relevant to adaptation to body mass. For example, a 2-fold change in a rate constant that controls shortening velocity, would according to transition state theory require only a change of the order of 1 kcal/mol in the activation energy. This is comparable to a single hydrogen bond or van der Waals interaction (see note in Supp information and (Frauenfelder and Wolynes 1985; Gutfreund 1995)).

In **Figure 3.1** the sequence identity vs mass is plotted for four representative myosin motor domains, the slow/cardiac β-myosin, the fast muscle IIx, the non-muscle IIa and the embryonic isoforms. Plots for the other 9 isoforms are presented in Figure S3.1. The plots in **Figure 3.1** show that both adult sarcomeric forms β and IIx have a strong correlation with body mass ($R^2$ = 0.63 and 0.838 respectively) with gradients of -0.72 and -0.84 percent divergence per log kg. The data for all 13 isoforms is summarised in **Figure 3.1E** and **Table 3.1**. Based on the rate of divergence with body mass, the 13 myosin II isoforms form three groups. The first group with gradients > 0.5 % per log kg body mass contains four of the five main adult sarcomeric muscle isoforms, IIb, β, IIa and IIx isoforms with the IIb isoform plots showing a wider scatter in data as indicated in the 20-40 % error in the gradient value. For a myosin motor domain of ~800 amino acids, a gradient of >0.5 % per log kg body mass means >4 amino acids change for each 10-fold increase in body mass. At the other extreme, five of the isoforms (slow tonic, embryonic, perinatal, non-muscle A and non-muscle B) exhibit little divergence of sequence and no correlation with body mass (slope < 0.1, $R^2$ < 0.13) i.e. < 1 amino acid per log kg increase in mass.

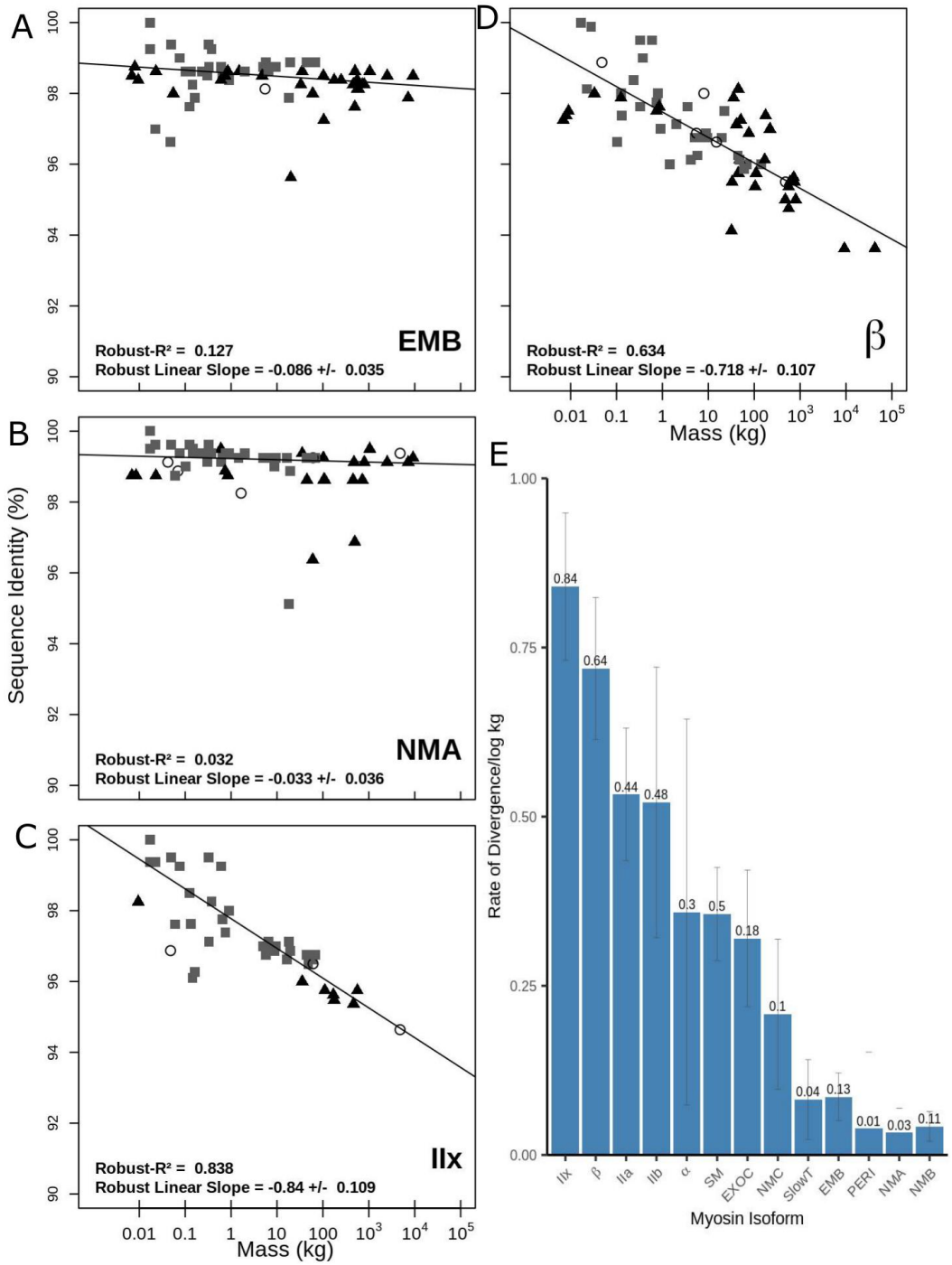



**Figure 3.1**. **Sequence Identity (%ID) vs Mass (kg) for myosin II motor domains.**

A-D) Sequence Identity (%ID) vs Mass (kg) for the motor domains of four myosin II isoforms; embryonic (EMB), non-muscle A (NMA), β-myosin, and IIx. Symbols indicate the clade that each species belongs to: grey squares (Euarchontoglires), black triangles (Laurasiatheria) and clear circles (Afrotheria and Metatheria). Each plot has been fitted with a robust linear regression. Sequence identity is pairwise to the mouse. The $R^2$ value and slope are shown on each plot.

The remaining four isoforms (α-myosin, smooth muscle and extraocular and NMC) are intermediate between the two groups with a lower rate of sequence divergence with body mass (-0.21 – -0.36) with only the smooth muscle isoform having an $R^2 > 0.3$ (0.498). Note the very large scatter in the data for α-cardiac myosin (**Figure S3.1D** gradient -0.36 ± 0.29)

A similar range of gradients (-0.75 to -0.05) and $R^2$ values (0.85 to 0.016) were observed for the tail domains but only the adult sarcomeric myosins had $R^2$ values greater than 0.3 (**Table 3.1**, **Figure S3.1**) and for these sarcomeric myosins, the gradients and $R^2$ values were similar for the motor and tail domains. An intriguing exception was the β-myosin isoform, where there was no correlation between size and sequence variation in the tail (gradient -0.06 %/log(kg), $R^2$ 0.03) showing that the tail is far more highly conserved than the motor and more highly conserved than for other sarcomeric myosin tail domains.

Given that there may be some error associated with the species body mass, we tested what effect this may have on the results of this analysis. To do this we randomly modified the body mass of each species such that they were varied in the range 80 %-120 % of the average value being used. This was repeated 1000 times each time reperforming the analysis. Our results demonstrate that the results are stable to such errors (**Table S3.1**).

Species and gene trees were generated for the sarcomeric myosin isoforms (**Figure S3.6**). To exclude the possibility that the observations for adult sarcomeric myosins could simply be the result of phylogenetic relationships between the sequences we used Felsenstein's phylogenetically independent contrasts (PIC; (Felsenstein 1985)) method to exclude the phylogenetic relationship as a factor for the correlation between sequence divergence and body mass. For β-myosin we still observe a significant correlation ($p=0.06 \times 10^{-4}$). For the other three main adult isoforms (IIa, IIb, IIx) where a significant correlation had been observed, only IIx was significant (p=0.009) while IIa and IIb were not (p=0.54 and 0.1 respectively). We also performed the equivalent analysis for the tail domain data (**Table S3.1**) and observed a p-value of 0.29 for the β-myosin, indicating that variation in the tail domain is explained by phylogenetic relationships.

### 3.4.2 Adaptation of the β-myosin motor domain to reduce contraction velocity as species size increased.

Of the adult sarcomeric myosin isoforms, β-myosin is unique as it is the only slow muscle myosin and it is expressed in few tissues, primarily slow, Type 1 muscle and in cardiac muscle (it is the dominant isoform expressed in the ventricle of the heart of mammals larger than 1 kg) and as result of both of these it performs the same specific function in different species. The other striated muscle isoforms are expressed in multiple tissues and may be involved in multiple different functions.

The negative correlation between species heart rate and body mass is well established across a wide range of species (Savage et al. 2007). Given our observation that β-myosin shows a strong association of sequence change with body mass, it is possible that there has been selection pressure on this isoform to enable a change in contraction velocity which is associated with the change in heart rate as body mass has increased. However, the observation of small correlation between sequence and body mass for the α isoform implies that this relationship between body mass and sequence if, present, is less dominant for α than for β. It is possible that the role of β as the only slow skeletal muscle isoform is more significant in defining the relationship to body mass than its role in cardiac muscle. This would make it similar to the fast skeletal muscle isoforms (IIa, IIb, IIx) which show strong dependence of sequence on body mass.

In addition, the rate constant controlling ADP release is easily measured for β-myosin and limits contraction velocity for this isoform **(Table 3.2** and refs therein). Therefore, all further analyses focus on variation in β-myosin.

| *Slow muscle/β cardiac myosin* | ***Measured*** | | ***Predicted τ (sec)*** | | |
|---|---|---|---|---|---|
| | $k_{-ADP}$ ($s^{-1}$) | $V_0$ (μm/sec/half sarcomere) | $\tau_{ADP} = 1/k_{-ADP}$ (msec) | $\tau_{Vo} = d/V_0$ (msec) | Ratio $\tau_{ADP}/\tau_{Vo}$ |
| ***Rat*** | *119* | *1.42* | *8.4* | *7.04* | *1.19* |
| ***Rabbit*** | *63* | *0.67* | *15.9* | *14.9* | *1.06* |
| ***Human*** | *30** | *0.33* | *33* | *30.3* | *1.08* |
| ***Cow*** | *27* | *0.27* | *37* | *37.0* | *1.00* |

*Experimental data was collected at 100 mM KCl and 12 °C.*

*$k_{-ADP}$ values for bovine, rabbit and human are from Deacon et al* (Deacon et al. 2012), *the rat from this study. NB the value for human $k_{-ADP}$ at 12 °C was estimated from an Arrhenius plot of values between 20 and 10 °C. These values are consistent with rat and porcine β data collected using the two headed myosin fragment, heavy meromyosin carried out at 100 mM KCl and 15 °C.*

*$V_o$ data for rat, rabbit & human are from Pellegrino et al* (Pellegrino et al. 2003)*, bovine from Toniolo et al* (L. Toniolo et al. 2005).

**Table 3.2. The relationship between the predicted and measured parameters for four slow/beta cardiac myosin isoforms**.

In terms of the actin myosin cross bridge cycle, the dominant model proposes that the maximum velocity ($V_0$) is limited by the lifetime of the strongly attached force holding state (τ) the "detachment limited model" ((Siemankowski, Wiseman, and White 1985), $V_0 = d/\tau$ where d is the working stroke of the cross bridge; assumed here to be 5 nm (7). For the mammalian, β- myosin isoform, it is well established that τ is defined by the rate constant controlling ADP release $k_{-ADP} = 1/\tau$ (Deacon et al. 2012; Nyitrai and Geeves 2004; Marieke J. Bloemink and Geeves 2011) . Thus, values of $k_{-ADP}$ measured using myosin motor domains isolated from β-cardiac/slow muscle of a mammal predict remarkably accurately the maximum shortening velocity of a muscle fibre taken from the same tissues.

The location of sequence changes in the motor and tail domains were examined to evaluate if particular structural features of the domains are especially variable (**Figure 3.2**). For each residue we considered the frequency of the consensus amino acid (black lines) and the number of different amino acids present at each position (red lines). In the β-myosin motor

domain, of the 800 amino acids, 632 were totally conserved and a further 114 sites were highly conserved (i.e., fewer than 4/67 species have a different amino acid; for 69 of these sites only one species had a different amino acid). These changes occurred in so few species that no conclusions could be drawn about the driver for these changes. For 52 positions the consensus amino acid was present in less than 90 % of the sequences, thus these positions with more frequent alternate amino acids are of greater interest (highlighted in **Figure 3.2** by crossing the dotted line).

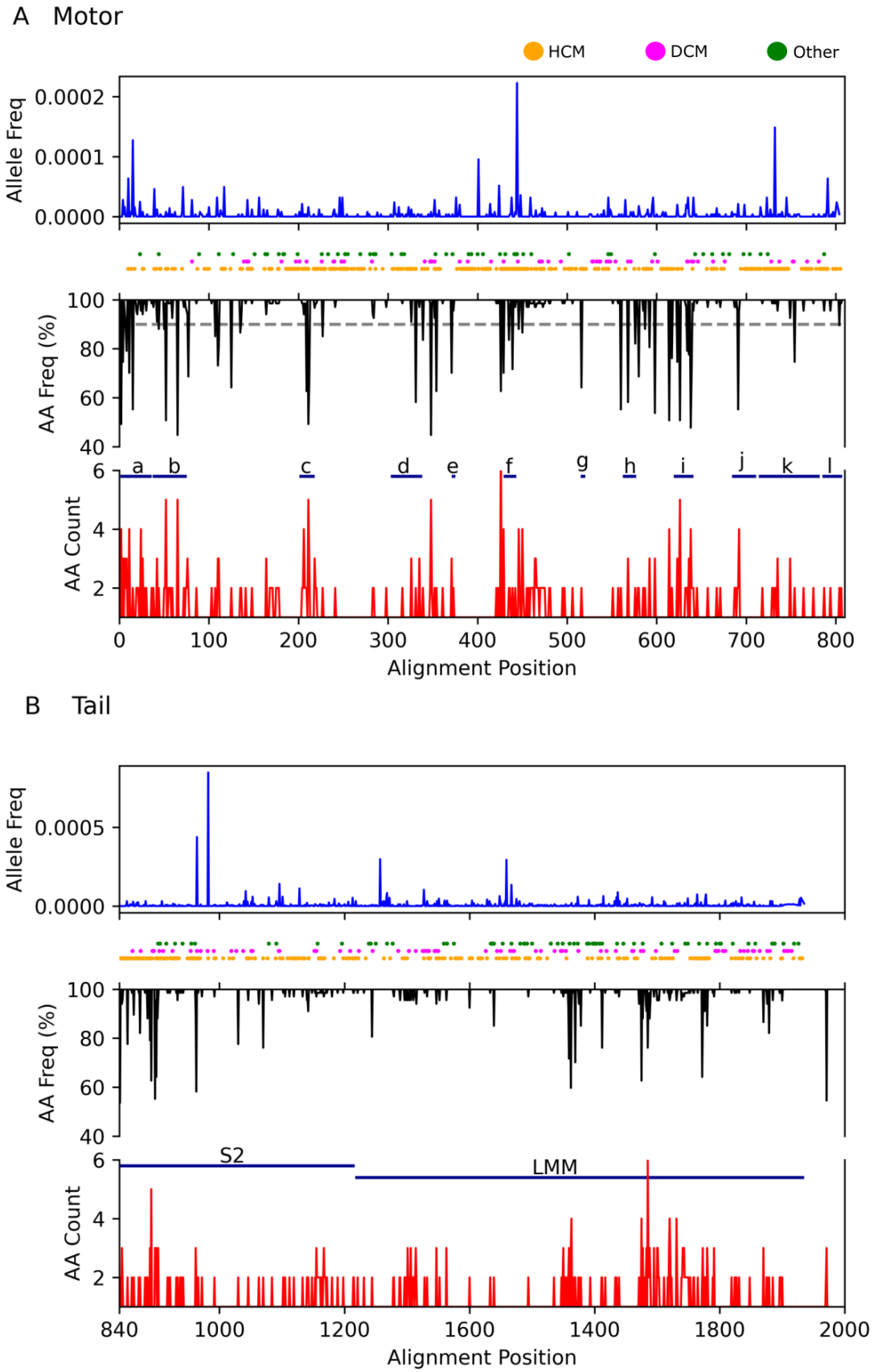

A Motor
HCM
DCM
Other
Allele Freq
0.0002
0.0001
0.0000
AA Freq (%)
100
80
60
40
a
b
c
d
e
f
g
h
i
j
k
l
AA Count
6
4
2
0
100
200
300
400
500
600
700
800
Alignment Position
B Tail
Allele Freq
0.0005
0.0000
AA Freq (%)
100
80
60
40
S2
LMM
AA Count
6
4
2
840
1000
1200
1600
1400
1800
2000
Alignment Position

**Figure 3.2. Location of human variation, cardiomyopathy associated variants and non-conserved residues for the β isoform.**

Blue lines show the frequency of missense variation present in the Genome Aggregation Database (gnomAD). Circles indicate the residue position of variants associated with cardiomyopathy (HCMs in orange, DCMs in magenta, and other in green (Francine Parker and Peckham 2020). Black lines show the number of times the consensus amino acid occurs at each residue position in the sequence (maximum equals the no of sequences used). Red lines show the number of different amino acids occurring at that position in the sequence (minimum 1). The dashed grey lines in **A** (in the section with the y-axis label AA Freq (%)) indicate the position at which more than 10% of sequences contain an alternate amino acid. Key functional areas of the sequence are as listed: Motor domain: a. N-terminus 1-36, b. SH3 domain 37-75, c. Loop 1 201- 215, d. Drosophila exon 7 region 300-335, e. Loop 4 368-372, f. Helix-O 426-440, g.Relay helix/loop, h. Loop 3 558-573, i. Loop 2 615-637, j. SH helices 680-707, k. Converter 710-778, l. IQ 781-810. B. Tail. Tail: IQ - neck region containing the IQ light chain binding motif, S2 - Sub fragment 2 coiled-coil, LMM - Light Meromyosin filament forming region. Raw data files are available at Figshare.

The sequence variations are scattered throughout the motor both within and outside the major functional regions of the motor domain with no identifiable pattern to the location of the changes (**Figure 3.2**). High levels of variation are found in the N-terminal domain (1-60) and near the surface loops, Loop 1 (near residue 210) and Loop 2 (near 630). These loops are known to be hypervariable across the larger myosin family.

The 52 common sites of variation between species were of interest to compare with amino acids in human β-myosin implicated in myopathies (Hypertrophic Cardiomyopathies, Dilated Cardiomyopathies, and other cardiomyopathies as orange, magenta, and green dots in **Figure 3.2** (Francine Parker and Peckham 2020)). More than half the sites in the motor domain (~500/800) have been found to have mutations linked to myopathies but only 20 out of the 52 sites identified to differ amongst mammals are common between the two groups. Of these only one E44D shares the same mutation; in all other cases the substituted amino acid is different. Thus, there is no simple correlation between the two groups of positions with changed sequences. This is illustrated by the converter domain (710-778, region k in **Figure 3.2**) which is enriched in myopathy mutations (59 reported mutations at 35 sites in the 68 residues of the converter domain) but there are very few

sequence changes between species in this region with only one position that differs in more than two species (Gly747, replaced by Ser in seven species).

Further, analysis of gnomAD (Homburger et al. 2016) revealed that there is very little variation in β-myosin within healthy human populations (**Figure 3.2**), thus suggesting that the sequence is highly constrained within human but varies between species. This constraint in humans for the MYH7 gene has been analysed in several papers and in databases (Freeman, Nakao, and Leinwand 2001).

### 3.4.3 Distinguishing between variation due to clade and body mass in β-myosin.

Change in sequence can be driven by several things. Here we identified positions of variation that are associated with body mass and also compared this to variation that is clade-specific between the two main clades present in the data set (*Euarchontoglires* and *Laurasiatheria*). Of the 67 complete sequences examined, about half were *Euarchontoglires* (32, e.g., rodents and primates) with a mass range of 0.011 - 140 kg and half were *Laurasiatheria* (30, e.g., bats, ungulates, cetaceans) with a mass range of 0.0069 - 42500 kg.

The 52 sites of variation that occur in >10 % of species, were analysed to distinguish between changes that correlated with clade and those that correlated with body mass as illustrated in **Figure 3.3** for four sites with the remaining plots in Supplementary Information (**Figure S3.2**). At most positions (41/52), only two residues were observed at each specific site; in the small number of positions (11/52) with multiple amino acids, only the two most frequent residues were considered. The identity of the two most frequent amino acids were coded as 0 and 1 (with the amino acid more frequently present in smaller species as 0) and a logistic regression model was fitted with log(mass) as the explanatory variable (**Figure 3.3**, **Figure S3.2**; see Methods) to model the transition between residues. For the four positions present in **Figure 3.3**, two had a strong correlation with species body mass (the amino acid common in small mammals is given first) A326P and I349M; ($p_{adj}$ << 0.01. Note the adjusted one percent significance threshold is $p=1.92x10^{-4}$ and the 5 % significance threshold is $p=9.62x10^{-4}$. $p_{adj}$ will be used to indicate the adjusted significance threshold), and each has a distinct midpoint mass for the transition between the two amino acids. In contrast, I125V has a low association with mass ($p_{adj} > 0.05$) and M77L has an intermediate association ($p_{adj}$ ~0.05), however both M77L and V125I have a strong

association with the clade (**Figure 3.4**). This is highlighted in **Figure 3.3A** where the four sequence positions are mapped against the species tree; L77 and I125 are found almost exclusively in *Laurasiatheria*. In contrast it is clear from **Figure 3.3A** that the variation at positions 326 and 349 is not explained by the phylogenetic relationships, with the two different amino acids spread across the two clades.

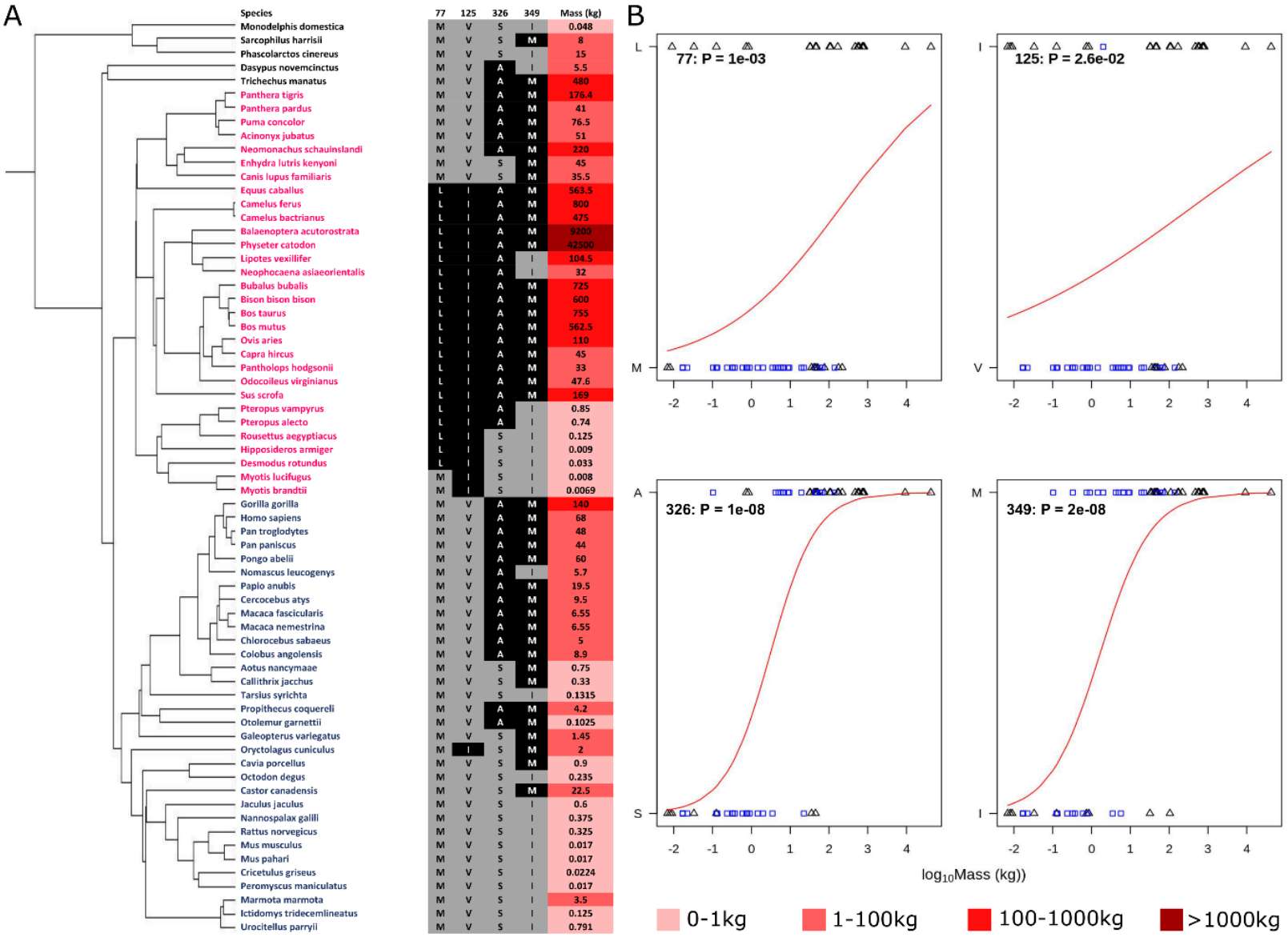

A
Species
77
125
326
349
Mass (kg)
Monodelphis domestica 0.048
Sarcophilus harrisii 8
Phascolarctos cinereus 15
Dasypus novemcinctus 5.5
Trichechus manatus 480
Panthera tigris 176.4
Panthera pardus 41
Puma concolor 76.5
Acinonyx jubatus 51
Neomonachus schauinslandi 220
Enhydra lutris kenyoni 45
Canis lupus familiaris 35.5
Equus caballus 563.5
Camelus ferus 800
Camelus bactrianus 475
Balaenoptera acutorostrata 9200
Physeter catodon 42500
Lipotes vexillifer 104.5
Neophocaena asiaeorientalis 32
Bubalus bubalis 725
Bison bison bison 600
Bos taurus 755
Bos mutus 562.5
Ovis aries 110
Capra hircus 45
Pantholops hodgsonii 33
Odocoileus virginianus 47.6
Sus scrofa 169
Pteropus vampyrus 0.85
Pteropus alecto 0.74
Rousettus aegyptiacus 0.125
Hipposideros armiger 0.009
Desmodus rotundus 0.033
Myotis lucifugus 0.008
Myotis brandtii 0.0069
Gorilla gorilla 140
Homo sapiens 68
Pan troglodytes 48
Pan paniscus 44
Pongo abelii 60
Nomascus leucogenys 5.7
Papio anubis 19.5
Cercocebus atys 9.5
Macaca fascicularis 6.55
Macaca nemestrina 6.55
Chlorocebus sabaeus 5
Colobus angolensis 8.9
Aotus nancymaae 0.75
Callithrix jacchus 0.33
Tarsius syrichta 0.1315
Propithecus coquereli 4.2
Otolemur garnettii 0.1025
Galeopterus varlegatus 1.45
Oryctolagus cuniculus 2
Cavia porcellus 0.9
Octodon degus 0.235
Castor canadensis 22.5
Jaculus jaculus 0.6
Nannospalax galili 0.375
Rattus norvegicus 0.325
Mus musculus 0.017
Mus pahari 0.017
Cricetulus griseus 0.0224
Peromyscus maniculatus 0.017
Marmota marmota 3.5
Ictidomys tridecemlineatus 0.125
Urocitellus parryii 0.791
B
77: P = 1e-03
125: P = 2.6e-02
326: P = 1e-08
349: P = 2e-08
log10Mass (kg))
0-1kg
1-100kg
100-1000kg
>1000kg

**Figure 3.3**. **Residue-mass transition plots for four representative amino acid sites**.

A) Species tree of organisms used in this analysis is shown, with the amino acids present in each species at the four sites and the mass of the organism displayed adjacent to it. Darker reds in the mass column are indicative of a greater mass value. In this tree, species from the Euarchontoglires clade are highlighted blue, and species from Laurasiatheria are highlighted pink. B) Binomial regression mapping the transition of the most frequent amino acid at positions in the motor region of β-myosin to the second most frequent amino acid at that position. The residue numbering is that of the human β-cardiac myosin. The blue squares are Euarchontoglires, and the triangles are Laurasiatheria. The P-value with each plot arises from a test of the null hypothesis that amino acid type is unrelated to mass. Raw data files are available at Figshare.

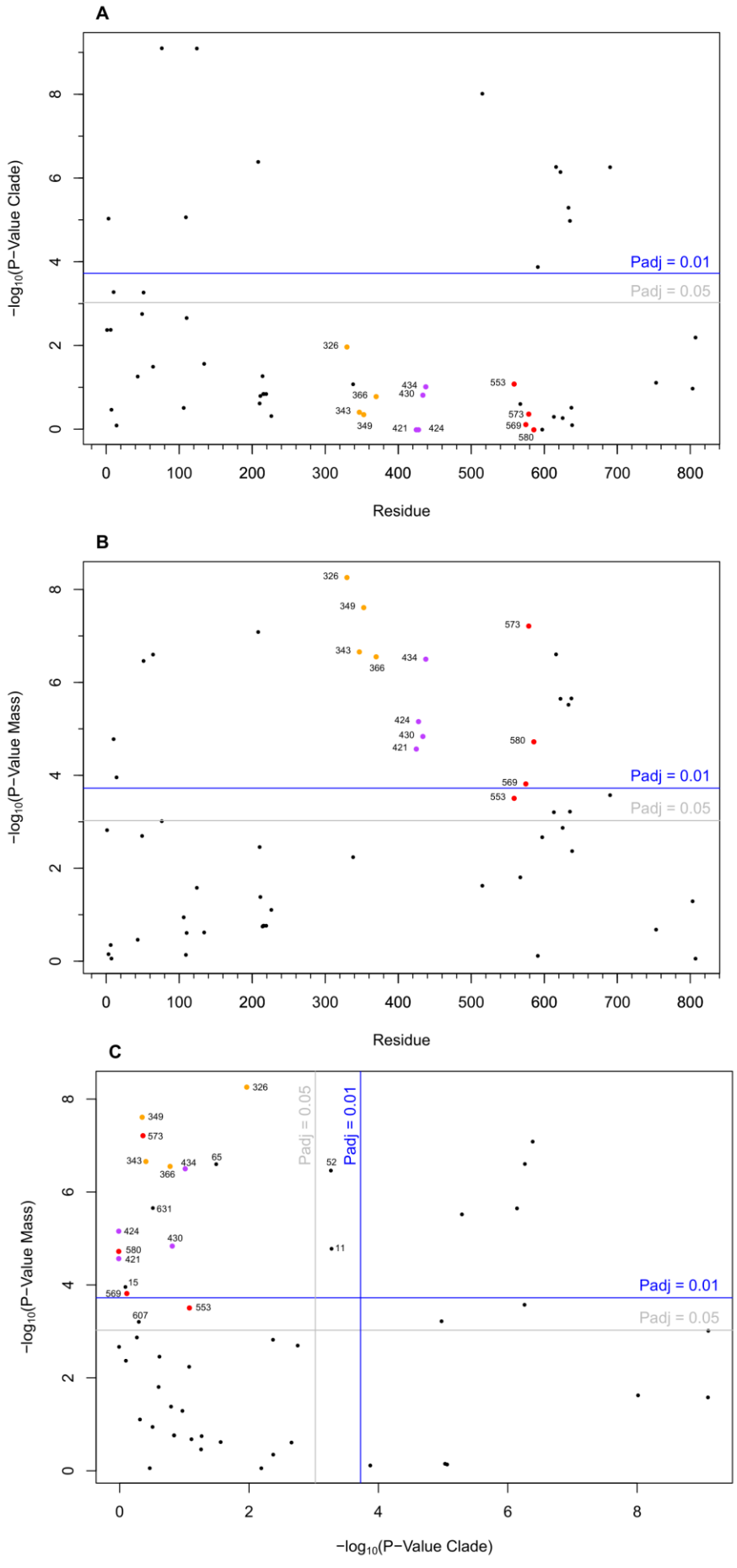

A
−log10(P−Value Clade)
Residue
Padj = 0.01
Padj = 0.05
326
366
343
349
434
430
421
424
553
573
569
580
B
−log10(P−Value Mass)
Residue
C
−log10(P−Value Mass)
−log10(P−Value Clade)
Padj = 0.05
Padj = 0.01
65
52
631
11
15
607

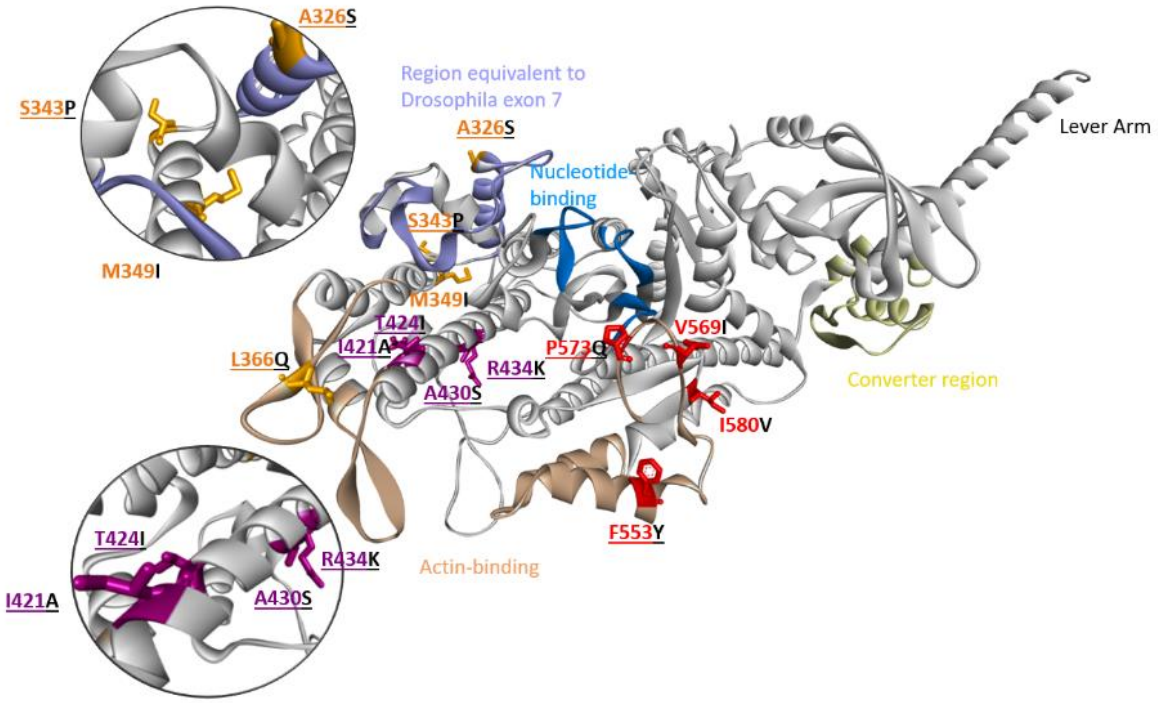
**Figure 3.4. Residue mass change association with mass, clade, and each other.**

The association of each residue with clade (A), mass (B), and the association of the P-values of clade vs mass (C) are plotted. 1 % and 5 % significance thresholds are shown with the Bonferroni adjusted lines (blue and grey). The three groups of residues investigated are highlighted in orange, purple and red, and labelled with the human residue numbering in each plot. Raw data files are available at Figshare.

Overall, only 12 sites had a very strong association with clade ($p_{adj} \leq 0.01$) and a further two were significant at the five percent level ($p_{adj} \leq 0.05$; Figure 3.4A). Some of these residues occur in two groups; one group in the N terminal region (4, 11, 52, 77, 110, 125) and four residues near surface loop 2 (610, 616, 627, 629). The remaining four are at D208E, E509T, T585I and I684M. Twenty positions were strongly associated ($p_{adj} \leq 0.01$) with body mass and a further four were associated at $p_{adj} \leq 0.05$ (**Figure 3.4B**). Nine positions were associated with both clade and body mass (**Figure 3.4C**), which is likely to represent that the very largest mammals (body mass > 500 kg) in the data set are all *Laurasiatheria* (Figure **S3.5**).

Twelve of the 24 sites associated with body mass occur in the known hypervariable regions, four in the N terminal region (**11**,15,**52**,65, bold **residues** also occur in the clade list), one in loop 1 (**D208E**), and a further six occur in or near loop 2 (607,**610,616,627,629**, 631)) and one at **I684M**. The remaining 12 sites group into three sets of four; most with $p_{adj} \leq 0.01$ (coloured in **Figures 3.4** & **3.5**). Comparing the strength of association between clade and body mass, these 12 sites, are strongly associated with body mass but not with clade (Figure 3.4C). Hence, we propose that these 12 positions are likely to be important in determining the β-myosin velocity of contraction. At eight of the 12 positions only two amino acids are observed, one position contains three amino acids, although the third is only present once (residue 366). Multiple amino acids (4-7) were observed at the remaining three positions. For two of the positions, 421 and 424, this reflects a subset of the species from one clade having an alternate amino acid in some of the larger species (see **Figure S3.3**).

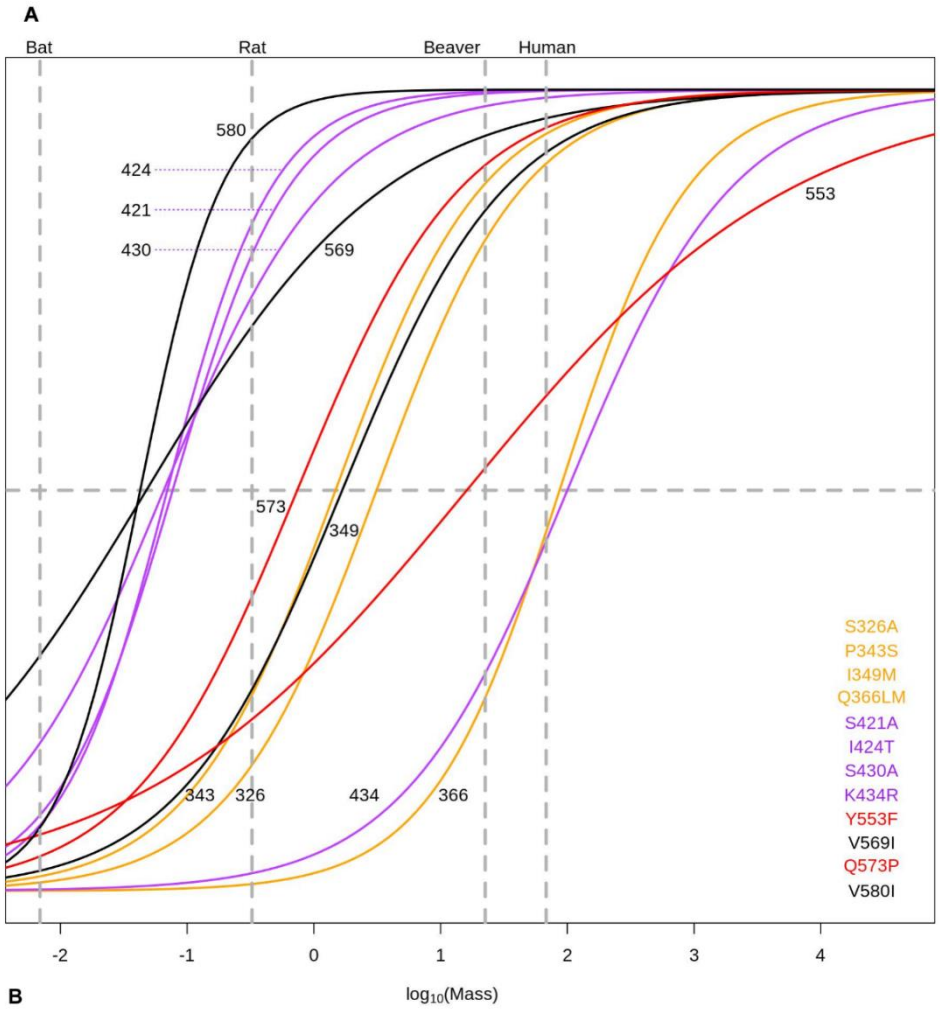
A




B

| Large/small | Mammal change | Residue | A326S | S343P | M349I | L366Q | I421A | T424I | A430S | R434K | F553Y | V569I | P573Q | I580V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P clade | | | 0.01 | 0.38 | 0.43 | 0.16 | 1 | 1 | 0.15 | 0.09 | 0.08 | 0.75 | 0.42 | 1 |
| P mass | | | 5.37E-09 | 2.14E-07 | 2.39E-08 | 2.73E-07 | 2.63E-05 | 6.77E-06 | 1.41E-05 | 3.07E-07 | 3.03E-04 | 1.48E-04 | 5.95E-08 | 1.84E-05 |
| Log (mass, kg) | Mid point | | 0.49 | 0.17 | 0.22 | 1.94 | -1.10 | -1.15 | -1.15 | 2.00 | 1.21 | -1.3 | -0.13 | -1.40 |
| | Range | 10-90% | 2.57 | 2.57 | 2.87 | 2.25 | 2.04 | 1.83 | 2.89 | 2.89 | 5.74 | 4.21 | 2.88 | 1.47 |

**Figure 3.5**. **Location of residues switched in the chimera**. Structure of human β-myosin (homology model based on Protein Databank ID 6FSA). The actin-binding site is highlighted in brown, exon 7 in blue, the nucleotide binding site in marine blue, and the converter region in yellow. The three groups of residues investigated are highlighted and labelled in orange (326, 343, 349, 366), purple (421, 424, 430, 434), and red (553, 569, 573, 580) in each plot, and those that were switched are underlined. The structure shown represents one of the known conformations adopted by myosin during the turnover of ATP and is shown to illustrate the relative position of the residues of interest in relation to e.g., actin and ADP binding sites. Since the residues highlighted are not directly coupled to ADP binding, it is likely that the stability of specific conformations and/or the transition (activation barrier) between conformations are affected by the sequence changes. For a review of the myosin motor domain structure see (Robert-Paganin, Auguin, and Houdusse 2018; Geeves and Holmes 1999) and for a broader discussion of activation barriers and conformational changes see the Supplementary Information and (Schmid and Hugel 2020). The Table indicates the details about the three groups of variable residues, the adjusted probability ($p_{adj}$) of association with clade or Mass, the Log(mass) at the midpoint of the transition of the regression line between the two amino acids, and the range (10 – 90 %) over which the transition occurs.




| Species | Mass (kg) | 434 | 366 | 553 | 326 | 349 | 343 | 573 | 421 | 424 | 430 | 569 | 580 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bat | 0.0069 | K | Q | Y | S | I | P | Q | S | I | S | V | V |
| Mouse | 0.017 | K | Q | Y | S | I | P | Q | S | I | S | V | V |
| Rat | 0.325 | K | Q | Y | S | I | P | Q | A | I | S | I | I |
| Rabbit | 2 | K | Q | Y | S | M | L | P | A | T | A | I | I |
| Beaver | 22.5 | K | Q | F | S | M | S | P | A | T | A | I | I |
| Human | 68 | R | L | F | A | M | S | P | I | T | A | I | I |
| Cow | 755 | R | L | F | A | M | T | P | V | K | A | I | I |

**Figure 3.6**. **Residue mass transition plots**. Overlapping binomial regression mapping the transition of the most frequent amino acid at positions in the motor region of β-myosin to the second most frequent amino acid at that position for the 12 residues of interest. The three groups of residues investigated are highlighted and labelled in orange (326, 343, 366), purple (421, 424, 430, 434), and red (553, 573) in each plot. The residues in black are residues of interest, but were not mutated (349, 569, 580). The table shows the mass and residues for species at different weights, where the bat and cow are from the clade Laurasiatheria, and the other species are from Euarchontoglires clade. Raw data files are available at Figshare.

The first group of residues is in a region 331-371 (orange residues in **Figure 3.4 – 3.6**). One residue 366, is just before the actin binding loop known as the Cardiomyopathy Loop. The remaining three residues are adjacent to the region of Drosophila myosin II coded by exon 7 (and the "linker region"). This region is one of four exons in the single myosin II gene of *Drosophila* which are alternately spliced to generate all isoforms of myosin II in *Drosophila* (Bernstein and Milligan 1997). We have previously shown (B. Miller et al. 2007) that the alternatively spliced forms of this region alter ADP release in the *Drosophila* myosin. The second set (426-439; residues Magenta in **Figure 3.4 – 3.6**) is in a long helix (Helix-O) in the upper 50 kDa domain that links the Cardiomyopathy loop to the nucleotide binding pocket (via switch 2). The third region (560 - 587; residues Red in **Figure 3.4** and **Figure 3.5**) is in the lower 50 kDa domain and residue 553 forms part of the Helix-Loop-Helix actin binding motif while 573 and 569 are part of actin binding Loop 3. The last residue 580 is on the β-strand that follows Loop 3. Thus, four of the twelve residues are in or close to actin binding loops, three are close to a region known to influence ADP binding (the region coded by exon 7) and four are on the helix that links the actin binding site to the nucleotide pocket. It is therefore possible that many of these residues could influence how actin binding leads to displacement of ADP.

Comparing the myopathy mutations listed by (Parker and Peckham, (Francine Parker and Peckham 2020)) six mutations occur in the 16 sites we found to be associated with mass ($p_{adj}$ <0.05) but not with clade. Each of these myopathy sites have a different substitution to the one found across mammals except R434K. Of the six sites, ClinVar (Landrum et al. 2018) lists only A326P as a potential cardiomyopathy mutation. The others are classified as benign or of uncertain significance.

### 3.4.4 Experimental testing of the computational predictions:

We have previously expressed the motor domain of human β-myosin in mouse $C_2C_{12}$ muscle cells and isolated the protein using His tags attached to the co-expressed human light chain. This is currently the only way to express mammalian striated muscle myosin motors domains but is complex and time consuming and yields just a few mg of protein (B. M. Miller et al. 2007; Q. Wang, Moncman, and Winkelmann 2003; Srikakulam and Winkelmann 2004). To test the hypothesis that the highlighted group of 12 residues is

responsible for a significant part of the adjustment of ADP release rate constant, we generated a chimeric human-rat β-myosin motor domain in which the region containing the nine positions (of the 12) that vary between human and rat, was replaced with the amino acid sequence present in rat (A326S, S343P, L366Q, I421A, T424I, A430S, R434K, F553Y, P573Q – human residue number and amino acid listed first). The other three positions are the same in rat and human (349, 569 & 580). At residue 421 we replaced Ile with Ala as present in rat, although Ser is present in most of the smallest mammals (See **Figure S3.5**).

As shown in **Table 2**, the velocity of contraction for β cardiac/Type I slow muscle fibres in rat and humans differ by a factor of ~4. Given that these residues have a range of transition masses (see **Figure 3.6** and Discussion), the hypothesis is that each of these nine residues will contribute a fraction of the difference between the rat and human β-myosin ADP release rate constant and hence, velocity. With all nine residues changed, our prediction was that the differences in the rate constant would be large enough to be easily detectable.

The S1 fragment of human β-myosin and the chimera were expressed in $C_2C_{12}$ cells and purified with the human essential light chain attached. Few details of the kinetic characterisation of the rat β-myosin S1 have been published (Freeman, Nakao, and Leinwand 2001). The rat β-myosin S1 was therefore purified from rat soleus muscle to use as a comparator for the chimera. The supplementary data include the SDS PAGE of all three proteins used in this study and demonstrates that all three proteins are pure and contain the appropriate light chains (**Figure S3.4**).

As a test of the functions of the chimeric protein, the ATP-induced dissociation of the chimera from pyrene labelled actin was monitored and compared to the recombinant human and the native rat S1. A typical transient is presented (inset in **Figure 3.7B**) and the observed amplitude of the signal change was the same for all three proteins. The similarity of observed amplitudes of the pyrene signal changes for the chimera, human and native rat proteins indicates that the chimera binds actin and releases it on ATP binding as for the human and rat S1. This is consistent with the chimera being a fully folded and active protein. A plot of the observed rate constant ($k_{obs}$) vs [ATP] gives a straight line which defines the apparent 2$^{nd}$ order rate constant for the reaction (**Figure 3.7B**, **Table 3.3**) and

appears the same for all three proteins. The observed rate constant of this reaction has been defined for many myosins and has two components, $k_{obs} = [ATP] K'_1 k'_{+2}$. The reaction is sensitive to both the affinity of ATP for the complex ($K'_1$) and the efficiency with which ATP induced a major conformational change in the myosin ($k'_{+2}$). This involves the closure of switches 1 and 2 onto the ATP and the opening of the major cleft in the actin binding site of myosin. The absence of any change in $K'_1 k'_{+2}$ is consistent with a well-preserved nucleotide pocket and a preserved communication pathway between the ATP binding pocket and the actin binding site.

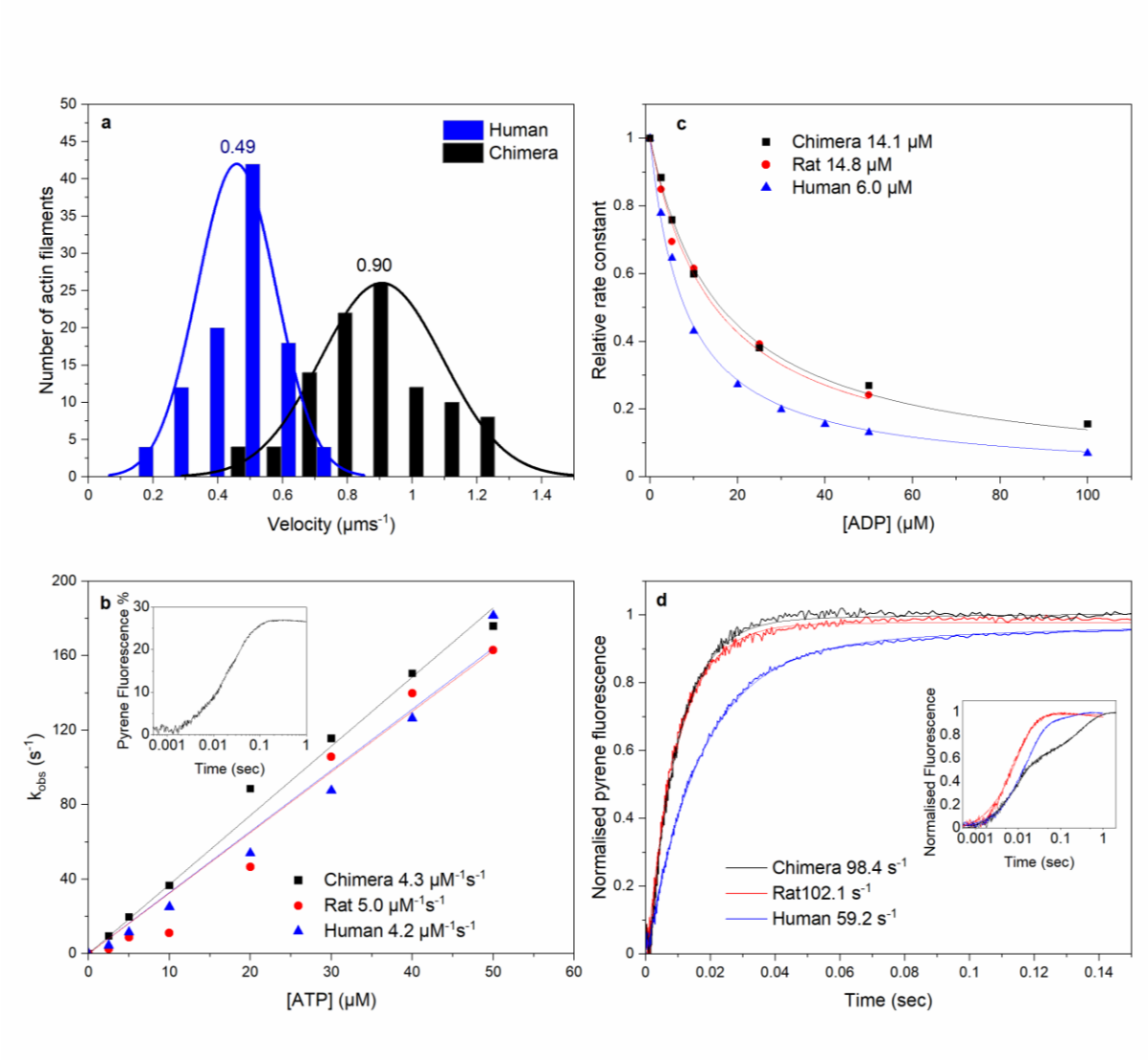


**Figure 3.7. In vitro analysis of the chimera, rat and human β-S1 proteins. A**. Histogram of in vitro velocity of 100 rhodamine-labelled-phalloidin actin filaments moving over human β-S1 or chimera S1. The solid line shows the data fitted to a single Gaussian curve. The mean velocity for the human β-S1 was 0.49 ± 0.028 μm s-1 and for the chimera 0.90 ± 0.015 μm s-1. **B**. The effect of ATP concentration on $k_{obs}$ for ATP-induced dissociation of pyrene-actin.S1. The gradient generates a second order rate constant of ATP binding; the values for the 3 proteins are highlighted next to the plot. Inset shows an example transient of 50 nM pyrene actin-chimera S1 mixed with 20 μM ATP, resulting in a fluorescence change of 26%. **C**. Plot of $k_{obs}$ dependence on [ADP] for the ATP induced dissociation of pyrene-actin.S1. 50 nM pyrene-actin.S1 was mixed with 10 μM ATP and varying [ADP] (0-100 μM). Numbers indicate the values of ADP affinity for acto.S1, $K_{ADP}$, for the 3 proteins. **D**. To measure $k_{+ADP}$, ADP is displaced from pyrene-actin.S1.ADP complex by an excess of ATP. 2 mM ATP was mixed with 250 nM S1 which was pre-incubated with 500 nM pyrene-actin and 100 μM ADP. $k_{+ADP}$ values for the 3 proteins are given 7D. Inset showing data on a longer log time scale showing the slow phase components of the transients. The average values from 3 independent measurements for experiments shown in B, C and D are summarised in Table 3. Raw data files are available at Figshare. The inset shown in **Figure 3.7D**; a complication of the ADP displacement measurement is that ADP displacement from human β-myosin occurs in two phases (fast and slow). The fast phase corresponds with ADP released at the end of the normal ATPase cycle while the slow phase is a trapped ADP which is released much slower and at a much slower rate than the overall cycling. This is therefore a dead-end side branch of the pathway commonly seen in slow muscle & non muscle myosins (Resnicow et al. 2010; Srikakulam and Winkelmann 2004; Nyitrai and Geeves 2004). The fraction of ADP trapped in this way is characteristic of each myosin. The rat β-myosin S1 has no apparent slow phase, the human has ~10 % of ADP released in the slow phase while the chimera has a larger fraction (~40 %) of the total ADP released in the slow phase. The role of the substituted amino acids in the slow phase requires further study, but the reader is referred to the literature for a broader study of this phenomena.

| | *Rat native S1* | *Chimera S1* | *Human S1* | *Chimera/Rat ratio* | *Chimera/Human ratio* |
|---|---|---|---|---|---|
| ***ATP binding to A.S1 ($\mu M^{-1}s^{-1}$)*** | *5 ± 0.16* | *4.5 ± 0.2* | *4.4 ± 0.3* | *0.9* | *1.02* |
| ***ADP affinity to A.S1 ($\mu M$)*** | *14 ± 0.09* | *14 ± 0.14* | *6.1 ± 0.7* | *1* | *2.3* |
| ***ADP release from A.S1.ADP ($s^{-1}$)*** | *107.2 ± 7.3* | *100.7 ± 4.2* | *59 ± 3.3* | *0.94* | *1.71* |
| ***Motility ($\mu m.s^{-1}$)*** | *NA* | *0.90 ± 0.221* | *0.49 ± 0.129* | *NA* | *1.84* |

*Data are from three separate preparation of protein from either different cell pellets (expressed protein) or different soleus muscles from the rat. Experimental conditions were 25 mM KCl, 20 °C.*

**Table 3.3**. **Comparison of ATP and ADP binding parameters of native rat S1, and the $C_2C_{12}$ cell expressed human β-myosin and chimera S1.**

Errors reported are SEM, except for motility which is the HWHM of the normal distribution.

The affinity of ADP for actin.S1 was measured in a competition assay with ATP (**Figure 3.7C**) and the affinity of ADP for the rat actin.S1 complex (14 µM) was 2.3-fold weaker than for the human WT protein (6.3 µM). These values are consistent with published values (M. J. Bloemink et al. 2007). The chimera was distinct from the human S1 and indistinguishable from the rat S1. To confirm this result the ADP release rate constant was measured directly by displacing ADP from actin.S1.ADP through addition of an excess of ATP. The results (**Figure 3.7D**) for human and rat S1 are again consistent with published values with ADP leaving the rat complex at ~ twice the rate of the human complex (107 vs 59 $s^{-1}$). The

chimera was indistinguishable from the rat S1. As predicted, the amino acids introduced into the human β-myosin motor domain weaken ADP affinity for actin.myosin by accelerating ADP release to make the human β-myosin S1 behave like the rat β-myosin S1.

Motor activity of the recombinant human β-myosin S1 and the chimera protein was measured using an *in vitro* motility assay (**Figure 3.7A**, **Supplementary movie S1**). This assay determines the myosin-mediated velocities of fluorescent actin-filaments moving on a nitrocellulose-coated slide surface. The human WT β-myosin moved actin at a velocity of 0.49 $\mu m.s^{-1}$, at 20 °C. For the chimera, the mean filament velocity was almost 2-fold faster than human WT at 0.9 $\mu m.s^{-1}$, which is consistent with the ~ 2-fold increase in ADP release-rate data. Our human S1 velocity was similar to the 0.612 $\mu m.s^{-1}$ value reported by Ujfalusi et al, which was measured at 23 °C (Ujfalusi et al. 2018). A similar velocity of 0.378 $\mu m.s^{-1}$ was reported for full length human-β myosin at 25 °C and a velocity of 0.624 $\mu m.s^{-1}$ for the rat. This gives a rat/human ratio of 1.65, very similar to our chimera/human ratio (1.84). The motility assay was not performed for the rat S1 as we do not have an expression system for the protein. The native rat S1 has only a single light chain and lacks a tag to attach the protein to the surface. The rat protein will not therefore give a valid comparable measurement. However, it is known from the literature (**Table 3.2** references therein) that the rat protein moves 3-5 times faster than the human protein, depending upon the exact measurement conditions.

## 3.5 Discussion

The well-established observation that the muscles of large mammals contract at slower speeds than small mammals (Lindstedt 1987) and that the maximum contraction velocity is a property of the myosin isoforms expressed in the muscle (Pellegrino et al. 2003) led us to ask if it is possible to define which amino acid changes are responsible for the change in contraction velocity. Our analysis of a large number of myosin II sequences indicates that adult sarcomeric myosins have a larger variation in sequence than two of the non-muscle myosins (or developmental striated muscle myosin isoforms) which are expected to be independent of species size. Furthermore, there appears to be a correlation between size difference and sequence difference.

Examination of 67 mammalian β-myosin sequences identified sequence changes at 12 sites which had a high correlation with species size and little correlation with clade. **Figure 3.6** plots the probability of transition vs species mass for each of these sites and indicates that there is a distinct mid-point of transition for each site as shown in **Figure 3.5B**. Reading the mass of a given species on the x-axis allows the probability of each site having the amino acid associated with small or large species to be estimated. The trend in midpoint values implies there is a distinct order in which the amino acids at the 12 sites change as the species mass increases from the smallest to the largest. Furthermore, the order in which the sequence changes is similar for both clades. This is illustrated for seven species in **Figure 3.6B** where the seven sites are shown in the order of the mid-point mass of the transition. The sites change from yellow to blue as the species mass increases in the order expected. A similar plot for all 67 species is given in the Supp information (**Figure S3.5**) and shows the same phenomena but, as might be expected, with more scatter in the data.

Our analysis suggests that there are reasons to believe that the sequence of the motor domain is changing with the mass of the species. Before considering the sequence changes in the motor domain of β-myosin in more detail it is useful to consider why contraction velocity and specifically the maximum velocity is an important factor in defining muscle performance.

### 3.5.1 The central role of $V_0$ in muscle physiology

The maximum contraction velocity of a muscle, $V_0$, is a key attribute of muscle contraction that plays a significant role in defining both the Force-Velocity relationship, power output and the efficiency of muscle contraction (Bottinelli and Reggiani 2000; Schiaffino and Reggiani 2011). $V_{0,}$ one part of the Force-Velocity relationship of a muscle, as empirically defined by Hill in 1938, is a fundamental property which underpins both the power output (Power = Force x Velocity) and the contraction velocity at which power and efficiency of contraction are maximum; normally considered the optimal working conditions for a muscle. Power and efficiency are fundamental parameters which define the movement of an animal and crucially the output of the heart. In contrast to $V_{0,}$ which varies more than 10-fold for different myosin isoforms (Canepari et al. 2010), the maximum force a myosin can generate varies relatively little between myosin isoforms when assessed at the single molecule level or in the muscle sarcomere (Canepari et al. 2010; Schiaffino and Reggiani 2011). Maximum velocity is therefore a central parameter that plays a significant role in defining both the Force-Velocity relationship, power output and the efficiency of muscle contraction.

There are strong theoretical arguments and experimental evidence that ADP release is the event in the crossbridge cycle that defines the maximum velocity of contraction in β-myosin (**Table 3.2**). We need to understand how the rate of ADP release is adjusted to match the physiological requirements.

### 3.5.2 How does the location of the 12 residues in the motor domain influence ADP release and hence $V_0$.

The sites identified are not directly linked to the ADP binding pocket and this is not a surprise since the same pocket binds ATP and compromising ATP binding is likely to be detrimental to myosin function. This illustrates the problem in modifying myosin, the mutation should affect only ADP release and not any other event in the cycle which it is assumed are optimised to make efficient use of the energy of ATP hydrolysis.

Some loops on myosin's surface are known to be variable across myosin in general and surface loop 1 has been shown to modulate ADP release in some myosins (Scallop muscle myosin II and smooth muscle myosin II have an alternate splice in this region (Kurzawa-

Goertz et al. 1998; Canepari et al. 2010), and smooth muscle myosin, Dictyostelium myosin II and human myosin 1b have been modified to explore the role of variations in this region (Haase and Morano 1996; Sweeney et al. 1998; Murphy and Spudich 1998). Loop 1 does show some variations here but this has no correlation with size. Note that the Rat and Cow share identical sequences in this region.

Myosin is a coordinated mechanical system with each part of the structure able to sense allosteric changes across the structure as a whole – as illustrated by the many mutations in β-myosin that have been explored. Each mutation can have multiple effects on the cycle as a whole. Thus, in principle, mutations anywhere in the motor domain could influence ADP release but the system is tightly constrained because ADP release must be modulated without a negative effect elsewhere in the cycle. There will be a limited number of ways in which this can be achieved. The evidence from **Figure 3.6** is that a solution to this problem exists and both clades have found the same solution.

*Drosophila* has multiple isoforms of muscle myosin II (flight, cardiac, leg, embryonic muscle etc) and all of these isoforms are encoded by a single gene. Alternate splicing of four regions of the motor domain results in the expression of each different isoform. One of these regions is coded by exon 7 which has been shown to modulate ADP release and this region overlaps with some of the mutations highlighted here. Thus, different routes to modulating ADP release are possible and it is intriguing that Drosophila and mammals use the same region. To define exactly how the set of 12 sites identified here can modulate ADP release and $V_0$ will require more detailed molecular structures and detailed molecular dynamics.

One question about the relationship between myosin sequence and mammal size is whether such a correlation is seen with any other sarcomeric protein. It is of course impossible to be definitive about this but examination of the other major sarcomeric proteins actin, tropomyosin and the three troponin components (I, T & C) reveals no such correlation. In fact, actin and tropomyosin are among the most conserved eukaryotic proteins showing little variation among mammals or indeed vertebrates (Barua 2013; Barua et al. 2012). Troponins do have a higher degree of variation than actin or tropomyosin but we found no evidence of a correlation between sequence and mammal size. There has been a report of a correlation in the size of repetitive Pro Ala rich regions in myosin binding

protein C (Shaffer and Harris 2009) and in myosin light chain 1 (Bicer and Reiser 2007) but these have to-date only considered a small number of species and include a broader data set of vertebrates.

In conclusion, we have used a computational approach to identify variation in β-myosin that is associated with increasing body mass, which would have a role in reducing heart rate. The experimental characterisation of the chimeric human-rat β-myosin demonstrates that these residues do indeed control the rate of ADP, the rate limiting step in the myosin cycle, and thus are likely to have adapted to enable the slower heart rate required in larger animals as they have evolved from small to large (Copes Law; (Hone and Benton 2005; Stanley 1973)).

# Chapter 4: Differentially conserved amino acid positions may reflect differences in SARS-CoV-2 and SARS-CoV behaviour

Denisa Bojkova[#], Jake E. McGreig[#], Katie-May McLaughlin[#], Stuart G. Masterson[#] et al. "Differentially conserved amino acid positions may reflect differences in SARS-CoV-2 and SARS-CoV behaviour". *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btab094

My contribution to this work was as follows:

1. Wrote all the scripts for processing the genomic data.
2. Wrote the pipeline for determining differentially conserved positions.
3. Performed the mapping DCPs onto relevant protein structures and assessed likely impacts, along with Katie May-McLaughlin.
4. Produced Figures 4.2A, 4.2B, 4.2C, 4.2D, along with Mark Wass.
5. Created the back- and front-end of a webserver to host the DCP tool used in this Chapter. Website URL: https://www.wass-michaelislab.org/cat/.
6. Involved in drafting and approval of the final manuscript.

## 4.1 Abstract

**Motivation:** SARS-CoV-2 is a novel coronavirus currently causing a pandemic. Here, we performed a combined in-silico and cell culture comparison of SARS-CoV-2 and the closely related SARS-CoV.

**Results:** Many amino acid positions are differentially conserved between SARS-CoV-2 and SARS-CoV, which reflects the discrepancies in virus behaviour, i.e., more effective human-to-human transmission of SARS-CoV-2 and higher mortality associated with SARS-CoV. Variations in the S protein (mediates virus entry) were associated with differences in its interaction with ACE2 (cellular S receptor) and sensitivity to TMPRSS2 (enables virus entry via S cleavage) inhibition. Anti-ACE2 antibodies more strongly inhibited SARS-CoV than SARS-CoV-2 infection, probably due to a stronger SARS-CoV-2 S-ACE2 affinity relative to SARS-CoV S. Moreover, SARS-CoV-2 and SARS-CoV displayed differences in cell tropism. Cellular ACE2 and TMPRSS2 levels did not indicate susceptibility to SARS-CoV-2. In conclusion, we identified genomic variation between SARS-CoV-2 and SARS-CoV that may reflect the differences in their clinical and biological behaviour.

## 4.2 Introduction

In December 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a novel betacoronavirus, was identified that causes a respiratory disease and pneumonia called coronavirus disease 19 (COVID-19) (Gorbalenya, A.E., Baker 2020; N. Zhu et al. 2020). As of 22nd of December 2020, 77,801,721 confirmed COVID-19 cases and 1,713,109 COVID-19 deaths have been reported (Dong, Du, and Gardner 2020). Since 2002, SARS-CoV-2 is the third betacoronavirus, after severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), that has caused a substantial outbreak associated with significant mortality (A. Wu et al. 2020).

SARS-CoV-2 is closely related to SARS-CoV (Gorbalenya, A.E., Baker 2020; A. Wu et al. 2020). Entry of both viruses is mediated via interaction of the viral Spike (S) protein with the cellular receptor ACE2, and both viruses depend on S activation by cellular proteases, in particular by TMPRSS2 (Wrapp et al. 2020; Wan et al. 2020; Yan et al. 2020; Hoffmann, Kleine-Weber, et al. 2020; A. Wu et al. 2020; Walls et al. 2020; J. Cui, Li, and Shi 2019). Despite these similarities, the diseases caused by SARS-CoV-2 (COVID-19) and SARS-CoV (SARS) differ. According to WHO, the SARS-CoV outbreak resulted in 8,098 confirmed and suspected cases and 774 deaths, equalling a mortality rate of 9.6 % (www.who.int). Estimated mortality rates for SARS-CoV-2 are below 1 % (Borges do Nascimento et al. 2020) SARS-CoV was only spread by symptomatic patients with severe disease (Cheng et al. 2013). In contrast, SARS-CoV-2 has been reported to be transmitted by individuals who are asymptomatic during the incubation period or who do not develop symptoms at all (Rivett et al. 2020).

We have developed an approach to identify sequence-associated phenotypic differences between related viruses based on the identification of differentially conserved amino acid sequence positions (DCPs) and in silico modelling of protein structures (Pappalardo et al. 2016; Martell et al. 2019). Conserved amino acid positions are likely to be of functional relevance, and differential conservation may indicate functional differences and they have been widely used for the analysis of protein families (Rausell et al. 2010; K. M. Das et al. 2015). Here, we used this method to identify differentially conserved positions that may explain phenotypic differences between SARS-CoV-2 and SARS-CoV. These data were combined with data derived from virus-infected cells.

## 4.3 Methods

### 4.3.1 Structural analysis

Sequences for each of the SARS-CoV-2 proteins were obtained from the GISAID resource. The protein sequences were then filtered for sequences from human hosts with high coverage, and sequences with spans of X's were removed. The number of sequences retained after filtering for each protein is shown in **Supplementary Table S4.4**. Fifty-three SARS-CoV genome sequences derived from human hosts were downloaded from VIPR (Pickett et al. 2012). Open Reading Frames (ORFs) were extracted using EMBOSS getorf (Rice, Longden, and Bleasby 2000) and matched to known proteins using BLAST. Fragments and mismatches were discarded. To match the ORF1ab non-structural proteins, a BLAST database of the sequences from the SARS non-structural proteins was generated and the SARS-CoV-2 ORF1ab searched against it. The sequences for each protein were then aligned using ClustalO (Sievers and Higgins 2014b) with default settings.

Conserved positions were identified by calculating the Jensen-Shannon divergence score (Capra and Singh 2007b) for each position in the multiple sequence alignment of virus proteins. Differing alignment positions with conservation score >0.8 for both species were considered as differentially conserved positions (DCPs).

SARS-CoV-2 and SARS-CoV protein structures were downloaded from the Protein Databank (PDB; **Supplementary Table S4.1**) (Armstrong et al. 2020). Where structures were not available, they were modelled using Phyre2 (Kelly et al. 2015)(**Supplementary Table S4.2**). Ligand binding sites were modelled using 3DLigandSite (Wass, Kelley, and Sternberg 2010b). DCPs were mapped onto protein structures using PyMOL. Exposed (solvent-accessible) and buried (solvent-inaccessible) residues were identified using Python module *findSurfaceResidues* with default parameters. Amino acid changes at DCPs were manually analysed for their potential impact on protein structure and function based on presence or absence of hydrogen bonding, changes in hydrogen bonding capacity and changes in charge in SARS-CoV compared with SARS-CoV-2 proteins. Where models were unavailable, mutagenesis was performed within PyMOL to assess the potential impact of the amino acid changes. Here, different rotamers of the mutated amino acid were cycled through to identify the rotamer with least clashes. Future investigation into these positions should focus on the use of molecular dynamics or minimisation to identify the effect of each

mutant on the protein structure. The structural analysis grouped DCPs into six different categories based on the effect that they were proposed to have. These include 'unlikely', 'possible' and 'likely'. The possible and likely categories were split into three and two subgroups respectively depending on the type of effect (**Supplementary Table S4.3**).

### 4.3.2 Cell culture

The Caco2 cell line was obtained from DSMZ (Braunschweig, Germany). The cells were grown at 37 °C in minimal essential medium (MEM) supplemented with 10 % foetal bovine serum (FBS), 100 IU/ml penicillin, and 100 µg/mL of streptomycin. 293 cells (PD-02-01; Microbix Bisosystems Inc.) and 293/ACE2 cells (Kamitani et al. 2006)(kindly provided by Shinji Makino, UTMB, Galveston, Texas) were cultured in Dulbecco's modified Eagle medium (DMEM) supplemented with 10 % FBS, 50 IU/ mL penicillin, and 50 µg/ mL streptomycin. Selection of 293/ACE2 cells constitutively expressing human angiotensin-converting enzyme 2 (ACE2) was performed by addition of 12 µg/ mL blasticidin. All culture reagents were purchased from Sigma (Munich, Germany). Cells were regularly authenticated by short tandem repeat (STR) analysis and tested for mycoplasma contamination.

### 4.3.3 Virus infection

The isolate SARS-CoV-2/1/Human/2020/Frankfurt (Hoehl et al. 2020) was cultivated in Caco2 cells as previously described for SARS-CoV strain FFM-1 (J. Cinatl et al. 2004), as they are highly permissive to infection with SARS-CoV. Virus titres were determined as TCID50/ml in confluent cells in 96-well microtitre plates (J. Cinatl et al. 2003; Jindrich Cinatl et al. 2005).

### 4.3.4 Western blot

Western blotting was performed as previously described (Schneider et al. 2017). Briefly, cells were lysed using Triton-X-100 sample buffer, and proteins were separated by SDS-PAGE. Proteins were blotted on a nitrocellulose membrane (Thermo Scientific). Detection occurred by using specific antibodies against β-actin (1:2500 dilution, Sigma-Aldrich, Munich, Germany), ACE2, and TMPRSS2 (both 1:1000 dilution, abcam, Cambridge, UK) followed by incubation with IRDye-labeled secondary antibodies (LI-COR Biotechnology, IRDye®800CW Goat anti-Rabbit, 926-32211, 1:40,000) according to the manufacturer's

instructions. Protein bands were visualised by laser-induced fluorescence using infrared scanner for protein quantification (Odyssey, Li-Cor Biosciences, Lincoln, NE, USA).

### 4.3.5 Receptor blocking experiments

SARS-CoV/ SARS-CoV-2 receptor blocking experiments were adapted from Cinatl et al (2004). Caco2 cells were pre-treated for 30 min at 37 °C with goat antibodies directed against the human ACE2 or DDP4 ectodomain (R&D Systems, Wiesbaden-Nordenstadt, Germany). Then, cells were washed three times with PBS and infected with SARS-CoV-2 at MOI 0.01. Cytopathogenic effects were monitored 48 hours post infection. Cytopathogenic effect (CPE) was assessed visually by light microscopy by two independent laboratory technicians 48 hours after infection (J. Cinatl et al. 2003).

### 4.3.6 Antiviral assay

Confluent cell cultures were infected with SARS-CoV-2 or SARS-CoV in 96-well plates at MOI 0.01 in the absence or presence of drug. Cytopathogenic effect (CPE) was assessed visually by light microscopy by two independent laboratory technicians 48h post infection (J. Cinatl et al. 2003).

### 4.3.7 Viability assay

Cell viability was determined by 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) assay modified after Mosmann (Mosmann 1983), as previously described (Onafuye et al. 2019).

### 4.3.8 qPCR

SARS-CoV-2 and SARS-CoV RNA was isolated from cell culture supernatants using AVL buffer and the QIAamp Viral RNA Kit (Qiagen) according to the manufacturer's instructions. RNA was subjected to OneStep qRT-PCR analysis using the SYBR green based Luna Universal One-Step RT-qPCR Kit (New England Biolabs) and a CFX96 Real-Time System, C1000 Touch Thermal Cycler. Primers were adapted from the WHO protocol (Corman et al. 2020) targeting the open reading frame for RNA-dependent RNA polymerase (RdRp) of both SARS-CoV-2 and SARS-CoV: RdRP_SARSr-F2 (GTGARATGGTCATGTGTGGCGG) and RdRP_SARSr-R1 (CARATGTTAAASACACTATTAGCATA) using 0.4 µM per reaction. RNA copies/ml were determined by standard curves which were using plasmid DNA (pEX-A128-

RdRP) harbouring the corresponding amplicon regions for SARS-CoV-2 RdRP target sequence (GenBank Accession number NC_045512). For each condition, three biological replicates were used. Mean and standard deviation were calculated for each group.

## 4.4 Results

### 4.4.1 Determination of differentially conserved positions (DCPs)

Coronavirus genomes harbour single-stranded positive sense RNA (+ssRNA) of about 30 kilobases in length, which contain six or more open reading frames (ORFs) (J. Cui, Li, and Shi 2019; A. Wu et al. 2020). The SARS-CoV-2 genome has a size of approximately 29.8 kilobases and was annotated to encode 14 ORFs and 27 proteins (A. Wu et al. 2020). Two ORFs at the 5'-terminus (ORF1a, ORF1ab) encode the polyproteins pp1a and pp1b, which comprise 15 non-structural proteins (nsps), the nsps 1 to 10 and 12-16 (A. Wu et al. 2020). Additionally, SARS-CoV-2 encodes four structural proteins (S, E, M, N) and eight accessory proteins (3a, 3b, p6, 7a, 7b, 8b, 9b, orf14)(A. Wu et al. 2020). This set-up resembles that of SARS-CoV. The 8a protein in SARS-CoV is absent in SARS-CoV-2. 8b is longer in SARS-CoV-2 (121 amino acids) than in SARS-CoV (84 amino acids), and 3b is shorter in SARS-CoV-2 (22 amino acids) than in SARS-CoV (154 amino acids) (A. Wu et al. 2020).

To identify genomic differences between SARS-CoV-2 and SARS-CoV that may affect the structure and function of the encoded virus proteins, we identified differentially conserved amino acid positions (DCPs) (Rausell et al. 2010) and determined their potential impact by in silico modelling (Pappalardo et al. 2016; Martell et al. 2019).

In the reference sequences of the 22 SARS-CoV-2 virus proteins that could be compared with SARS-CoV, 1393 positions encoded different amino acids. 891 (64 %, 9 % of all SARS-CoV-2 genome residues) of these positions were DCPs (**Supplementary Table S4.3**). Most of the amino acid substitutions at DCPs appear to be fairly conservative as demonstrated by the average BLOSUM substitution score of 0.32 (median 0; **Figure S4.1**) and with 69 % of them having a score of 0 or greater (the higher the score the more frequently such amino acid substitutions are observed naturally in evolution). 46 % of DCPs represent conservative changes where amino acid properties are retained (e.g. change between two hydrophobic amino acids), 18 % represented polar - hydrophobic substitutions, and <10 % changes between charged amino acids **(Supplementary Table S4.3**).

Six of the SARS-CoV-2 proteins have a higher proportion of DCPs, S, 3a, p6, nsp2, nsp3 (papain-like protease), and nsp4 with 14.82 %, 11.68 %, 9.52 %, 21.38 %, 17.9 %, and 10.8 % of their residues being DCPs, respectively (**Supplementary Table S4.4**). Very few DCPs were observed in the envelope (E) protein and most of remaining non-structural proteins encoded by ORF1ab. For example, no residues in the helicase and <4 % of residues in the RNA-directed RNA polymerase, 2'-O-Methyltransferase, nsp8 and nsp9 are DCPs (Table S4.1). We were able to map 572 DCPs onto protein structures (**Supplementary Figure S4.2**, **Supplementary Tables S4.3** and **S4.4**). Nearly all of the mapped DCPs occur on the protein surface (86 %), with only 34 DCPs buried within the protein, primarily in S and the papain-like protease (nsp3) (**Supplementary Table S4.3**). We propose that 49 DCPs are likely to result in structural/ functional differences between SARS-CoV and SARS-CoV-2 proteins. A further 259 could result in some change. The remaining 264 DCPs seem unlikely to have a substantial functional impact (**Table 4.1**).

| Effect | Reason |
|---|---|
| Unlikely | Conservative changes (between residues with the same polarity/charge) which do not affect ability to form hydrogen bonds with equivalent residues in SARS-CoV and SARS-CoV-2 |
| Possible – conformational change | Changes which could affect the ability of a sidechain of a residue in a given position to form hydrogen bonds with equivalent residues in SARS-CoV and SARS-CoV-2 (e.g. gain/loss of polarity, substitution for larger/smaller sidechain) but no such effects are visible, or conservative changes (between residues with the same polarity/charge) which appear in the model to result in gain/loss of hydrogen bonding between equivalent residues in SARS-CoV and SARS-CoV-2 (but mutagenesis suggests hydrogen bonding is possible with sidechain rotation) |
| Possible – alteration of sidechain/ligand interactions | Changes which result in gain of charge/alter the charge of a sidechain for a residue in a given position |
| Possible – conformational change and alteration of sidechain/ligand interactions | Changes which affect the ability of a sidechain of a residue in a given position to form hydrogen bonds with equivalent residues in SARS-CoV and SARS-CoV-2 (e.g. gain/loss of polarity, substitution for larger/smaller sidechain) but no such effects are visible, and changes which result in gain of charge/alter the charge of a sidechain for a residue in a given position |
| Likely – conformational change | Changes which result in visible alteration in the conformation of a protein at a given location (e.g. through loss of hydrogen bonding between equivalent residues in SARS-CoV and SARS-CoV-2) and/or which result in the loss of capacity for hydrogen bonding |
| Likely – conformational change (and possible alteration of sidechain/ligand interactions) | Changes which result in visible alteration in the conformation of a protein at a given location (e.g. through loss of hydrogen bonding between equivalent residues in SARS-CoV and SARS-CoV-2, and/or which result in the loss of capacity for hydrogen bonding and which result in gain of charge/alter the charge of a sidechain for a residue in a given position) |

**Table 4.1.** Criteria used for classifying proposed effect on protein structure and function within the structural analysis.

### 4.4.2 Differentially conserved positions (DCPs) in interferon antagonists

At least 10 SARS-CoV proteins have roles in interferon antagonism (Totura and Baric 2012). Two of these proteins, p6 and the papain-like protease (nsp3), contain many DCPs, two have very few DCPs (nsp7 and nsp16), five have intermediate proportions of DCPs (nsp14, nsp1, nsp15, N and M), while p3b is not encoded by SARS-CoV-2. Initial studies have identified a difference in the interferon inhibition between SARS-CoV and SARS-CoV-2 (Lokugamage, Schindewolf, and Menachery 2020). Thus, it is possible that especially the DCPs in p6 and the papain-like protease may have an effect on interferon inhibition.

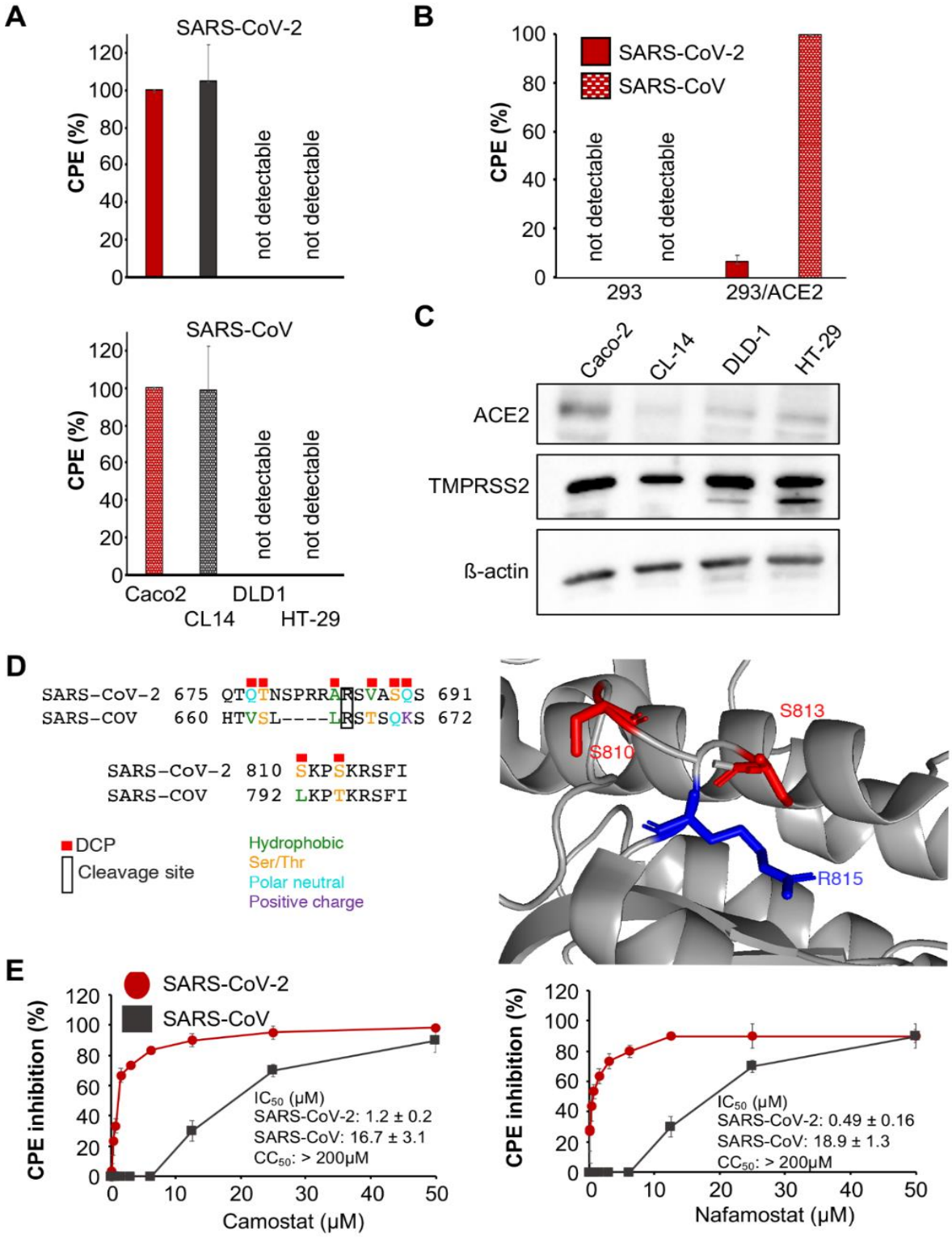

A
SARS-CoV-2
CPE (%)
not detectable
not detectable
SARS-CoV
Caco2
DLD1
CL14
HT-29
B
SARS-CoV-2
SARS-CoV
293
293/ACE2
C
Caco-2
CL-14
DLD-1
HT-29
ACE2
TMPRSS2
ß-actin
D
SARS-CoV-2 675 QTQTNSPRRARSVASQS 691
SARS-COV 660 HTVSL----LRSTSQKS 672
SARS-CoV-2 810 SKPSKRSFI
SARS-COV 792 LKPTKRSFI
DCP
Cleavage site
Hydrophobic
Ser/Thr
Polar neutral
Positive charge
S813
S810
R815
E
SARS-CoV-2
SARS-CoV
CPE inhibition (%)
IC50 (µM)
SARS-CoV-2: 1.2 ± 0.2
SARS-CoV: 16.7 ± 3.1
CC50: > 200µM
Camostat (µM)
IC50 (µM)
SARS-CoV-2: 0.49 ± 0.16
SARS-CoV: 18.9 ± 1.3
CC50: > 200µM
Nafamostat (µM)

**Figure 4.1. SARS-CoV-2 and SARS-CoV replication in cell culture**. A) Cytopathogenic effect (CPE) formation 48h post infection in MOI 0.01-infected Caco2, CL14, DLD-1, and HT29 cells. Representative images showing immunostaining for double-stranded RNA (indicates virus replication) and quantification of virus genomes by qPCR are presented in **Supplementary Figure S3**. B) CPE formation in SARS-CoV and SARS-CoV-2 (MOI 0.01)-infected ACE2-negative 293 cells and 293 cells stably expressing ACE2 cells (293/ACE2) 48h post infection. Immunostaining for double-stranded RNA and quantification of virus genomes by qPCR is shown in **Figure S4**. C) Western blots indicating cellular ACE2 and TMPRSS2 protein levels in uninfected cells. Uncropped blots are provided in **Figure S5**. D) A sequence view of the DCPs in the vicinity of the S two cleavage sites and an image of the R815 cleavage site and closely located DCPs. S is cleaved and activated by TMPRSS2. E) Concentration-dependent effects of the TMPRSS2 inhibitors camostat and nafamostat on SARS-CoV-2- and SARS-CoV-induced cytopathogenic effect (CPE) formation determined 48h post infection in Caco2 infected at an MOI of 0.01. Similar effects were observed in CL14 cells (**Supplementary Figure S6**).

### 4.4.3 Differences in cell tropism between SARS-CoV-2 and SARS-CoV

Next, we elucidated whether the substantial number of DCPs results in different phenotypes in cell culture, using the cell lines Caco2, CL14 (susceptible to SARS-CoV infection), HT-29, and DLD-1 (non-susceptible) (J. Cinatl et al. 2004). Analogously to SARS-CoV infection, SARS-CoV-2 replication was detected in Caco2 and CL14 cells, but not in HT-29 or DLD-1 cells, as shown by cytopathogenic effects (CPE) (**Figure 4.1A**), staining for double-stranded RNA (**Supplementary Figure S4.3A**), and viral genomic RNA levels (**Supplementary Figure S4.3B**).

However, ACE2-expressing 293 cells differed in their susceptibility to SARS-CoV-2 and SARS-CoV (**Figure 4.1B, Supplementary Figure S4.4**). ACE2 has been identified as a cellular receptor for both SARS-CoV-2 and SARS-CoV (J. Cui, Li, and Shi 2019; A. Wu et al. 2020; Hoffmann, Kleine-Weber, et al. 2020; Walls et al. 2020; Wan et al. 2020; Wrapp et al. 2020; Yan et al. 2020). Unmodified 293 cells are not susceptible to SARS-CoV infection due to a lack of ACE2 expression. However, 293 cells that stably express ACE2 (293/ACE2) support SARS-CoV infection (Kamitani et al. 2006). As expected, infection of 293 cells with SARS-CoV or SARS-CoV-2 did not result in detectable cytopathogenic effect (CPE) (**Figure 4.1B**), but a SARS-CoV-induced CPE was detected in 293/ACE2 cells (**Figure 4.1B**). In contrast,

293/ACE2 cells displayed limited permissiveness to SARS-CoV-2 infection (**Figure 4.1B**). Staining for double-stranded RNA (**Figure S4A**) and detection of viral genomic RNA copies (**Figure S4B**) confirmed these findings. Hence, the ACE2 status does not reliably predict cell sensitivity to SARS-CoV-2. Indeed, CL-14 was characterised by lower ACE2 levels than DLD-1 and HT29 (**Figure 4.1C**).

SARS-CoV-2 and SARS-CoV cell entry depends on S cleavage by transmembrane serine protease 2 (TMPRSS2) (Y. Zhou et al. 2015; Hoffmann, Kleine-Weber, et al. 2020). However, the non-SARS-CoV-2 susceptible and susceptible cell lines displayed similar TMPRSS2 levels (**Figure 4.1C**). Thus, cellular TMPRSS2 levels do also not reliably predict cell susceptibility to SARS-CoV-2.

### 4.4.4 Differences between SARS-CoV-2 and SARS-CoV S (Spike) protein cleavage sites and sensitivity to protease inhibitors

R667 and R797 are the critical cleavage sites in SARS-CoV S that are recognised by TMPRSS2 (Simmons et al. 2013; Y. Zhou et al. 2015). These cleavage sites are conserved in SARS-CoV-2 (R685 and R815) (**Figure 4.1D**). However, there is a four amino acid insertion in SARS-CoV-2 S prior to R685 and many of the residues close to R685 are DCPs (V663=Q677, S664=T678, T669=V687, Q671=S689, K672=Q690 DCPs are represented by the SARS CoV residue followed by the SARS-CoV-2 residue) (**Figure 4.1D**). The R815 cleavage site has two DCPs in close proximity (L792=S810, T795=S813) (**Figure 4.1D**). Around the R685 cleavage site two DCPs retain polar side chains (S664=T678, Q671=S689), while the others represent larger changes between hydrophobic and polar side chains (V663=Q677, T669=V687) and one changes from a positive charge to a polar side chain (K672=Q690). While around the R815 cleavage site, one substitution is conservative (T795=S813) and the other is a hydrophobic to polar change (L792=S810).

These changes are likely to impact on TMPRSS2-mediated S cleavage. Indeed, SARS-CoV-2 was more sensitive than SARS-CoV to inhibition by the serine protease inhibitors camostat and nafamostat (**Figure 4.1E**, **Supplementary Figure S4.6**), which are known to inhibit TMPRSS2-mediated S cleavage and virus entry (Y. Zhou et al. 2015; Hoffmann, Kleine-Weber, et al. 2020; Hoffmann, Schroeder, et al. 2020). This confirms that the observed differences in the amino acid sequence of S have functional consequences.

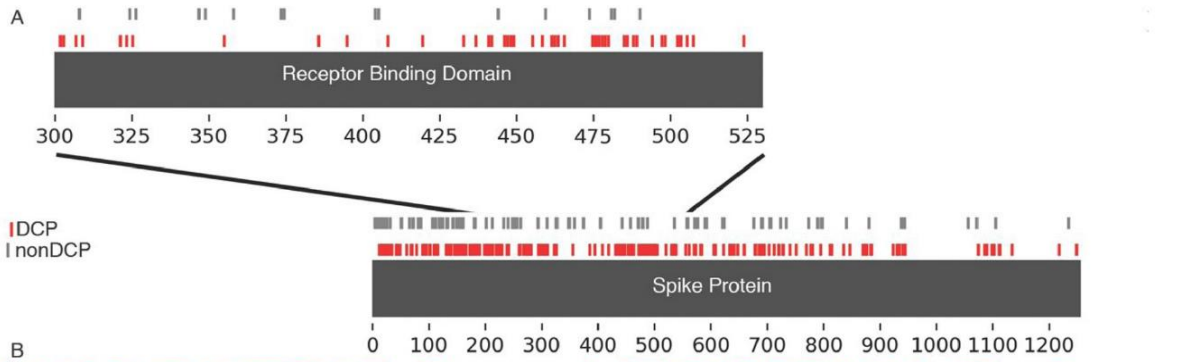
### 4.4.5 Differences in between SARS-CoV-2 and SARS-CoV S interaction with ACE2

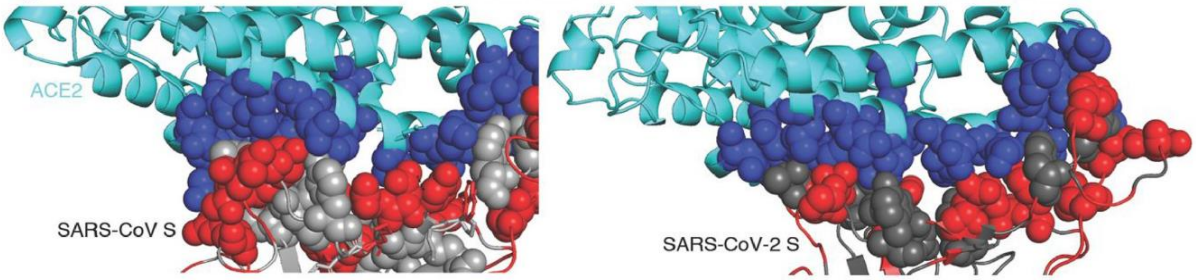
Our computational analysis detected further interesting changes in the S protein. SARS-CoV-2 S is 77.46 % sequence identical to the SARS-CoV S and most of the remaining positions are DCPs (186 residues) (**Supplementary Table S4.1**).

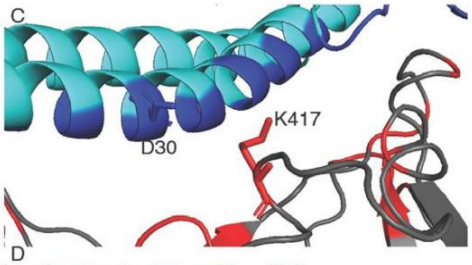
The SARS-CoV S receptor binding domain (residues 306-527, equivalent to 328-550 in SARS-CoV-2) is enriched in DCPs, containing 43 DCPs (19 % of residues). Nine of the 24 SARS-CoV S residues in direct contact with ACE2 were DCPs (**Figure 4.2A**, **Table S4.4**). Five of these DCPs represent conservative substitutions in amino acid (hydrophobic – hydrophobic or polar-polar), two hydrophobic polar substitutions, one positive charge to polar change, while the ninth is substitution between a hydrophobic and positively charged amino acid (**Supplementary Table S4.4**).

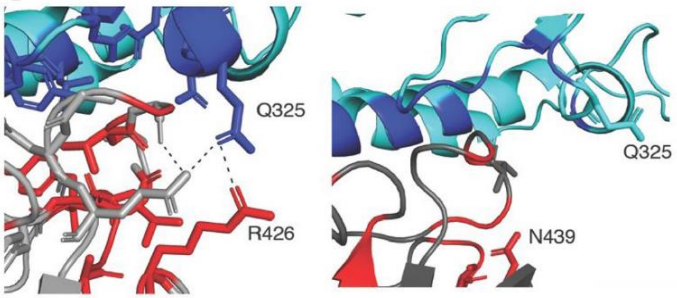
Analysis of the DCPs using the SARS-CoV and SARS-CoV-2 S protein complexes with ACE2 (W. Song et al. 2018; Yan et al. 2020) identified runs of DCPs (A430-T433, F460-A471) in surface loops forming part of the S-ACE2 interface and resulted in different conformations in SARS-CoV-2 S compared to SARS-CoV S (**Figure 4.2A**, **4.2B**). Two DCPs remove intramolecular hydrogen bonding within the spike protein in SARS-CoV-2 (**Supplementary Table S4**) and three DCPs (R426=N439, N479=QQ493, Y484=Q498) are residues that form hydrogen bonds with ACE2. For two of these positions, hydrogen bonding with ACE2 is present with both S proteins, but for R426=N439 hydrogen bonding with ACE2 is only observed with SARS-CoV S. N439 in SARS-CoV-2 S is not present in the interface and the sidechain points away from the interface. Further, analysis of the SARS-CoV-2 S-ACE2 complex highlighted important roles of the V404=K417 DCP, where K417 in SARS-CoV-2 S is able to form a salt bridge with ACE2 D30 (**Figure 4.2C**, **4.2D**) (Yan et al. 2020).

Alanine scanning (Chakraborti et al. 2005) and adaptation experiments (Wan et al. 2020) have identified 16 SARS-CoV S residues impacting on the binding affinity with ACE2. For all five residues identified from adaptation studies and four of the 11 identified by alanine scanning experiments, different amino acids are present in SARS-CoV-2 S (**Figure 4.2E**), highlighting the difference in the interaction with ACE2.



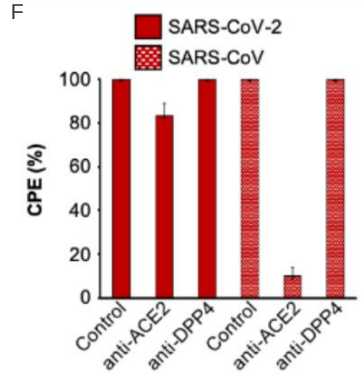




E

| SARS-CoV residue | Mutated to | SARS-CoV-2 residue |
|---|---|---|
| Alanine Scanning | | |
| K390 | A | **R403** |
| R426 | A | **N439** |
| D429 | A | D451 |
| T431 | A | **K444** |
| I455 | A | I477 |
| N473 | A | N496 |
| F483 | A | F506 |
| Q492 | A | Q515 |
| Y494 | A | Y517 |
| R495 | A | R518 |
| | | |
| Adaptation | | |
| Y442 | F increases | **L455** |
| L472 | F increases | **F486** |
| N479 | R is equivalent, K reduces) | **Q493** |
| D480 | G reduces | **S494** |
| T487 | S reduces | **N501** |



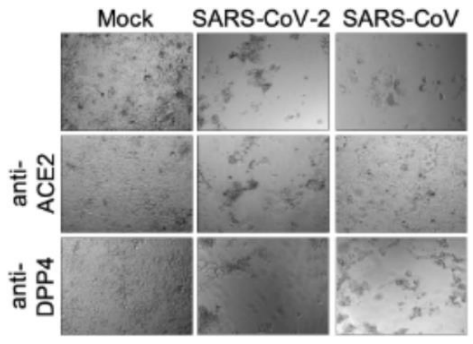



**Figure 4.2.** SARS-CoV-2 and SARS-CoV S interaction with ACE2. A-D) Differentially conserved positions in the Spike protein. A) A sequence view of the DCPs present in the Spike protein, with an inset showing the receptor binding domain. B) The S interface with ACE2 (cyan). The ACE2 interface is shown in blue spheres, DCPs in red. C) The V404=K417 DCP. D) The R426=N439 DCP, the left image shows SARS-CoV S R426, the image on the right show the equivalent N439 in SARS-CoV-2 S. E) SARS-CoV residues associated with altering ACE2 affinity and the residues at these positions in SARS-CoV-2 S. F) Cytopathogenic effect (CPE) formation in SARS-CoV-2 and SARS-CoV (MOI 0.01)-infected Caco2 cells in the presence of antibodies directed against ACE2 or DPP4 (MERS-CoV receptor) 48 hours post infection.

In agreement with our structural analysis, we detected differences in the effects of an anti-ACE2 antibody on SARS-CoV-2 and SARS-CoV infection. Antibodies directed against ACE2 were previously shown to inhibit SARS-CoV replication (W. Li et al. 2003). In line with this, an anti-ACE2 antibody inhibited SARS-CoV infection in Caco2 cells (**Figure 4.2F**). In contrast, the anti-ACE2 antibody displayed limited activity against SARS-CoV-2 infection (**Figure 4.2F**). This shows that it is more difficult to antagonise SARS-CoV-2 infection with anti-ACE2 antibodies and supports previous findings indicating a stronger binding affinity of SARS-CoV-2 S to ACE2 compared to SARS-CoV S (Walls et al. 2020; Wrapp et al. 2020). As anticipated, antibodies directed against DPP4, the MERS-CoV receptor (De Wit et al. 2016; J. Cui, Li, and Shi 2019), did not interfere with SARS-CoV or SARS-CoV-2 infection (**Figure 4.4E**).

## 4.5 Discussion

Here, we performed an in-silico analysis of the effects of differentially conserved amino acid positions (DCPs) between SARS-CoV-2 and SARS-CoV proteins on virus protein structure and function in combination with a comparison of wild-type SARS-CoV-2 and SARS-CoV in cell culture.

We identified 891 DCPs, which represents 64 % of the amino acid positions that differ between SARS-CoV-2 and SARS-CoV and nearly 9 % of all residues encoded by the SARS-CoV genome. 49 of these DCPs are likely to have a structural and functional impact. The DCPs are not equally distributed between the proteins. DCPs are enriched in S, 3a, p6, nsp2, papain-like protease, and nsp4, but very few DCPs are present in the envelope (E) protein and most of the remaining non-structural proteins encoded by ORF1ab. This indicates that the individual proteins differ in their tolerance to sequence changes and/or their exposure to selection pressure exerted by the host environment.

The large proportion of DCPs reflects the differences in the clinical behaviour of SARS-CoV-2 and SARS-CoV. Mortality associated with SARS-CoV is higher than that associated with SARS-CoV-2 (Borges do Nascimento et al. 2020; J. Cui, Li, and Shi 2019). SARS-CoV causes a disease of the lower respiratory tract. Infected individuals are only contagious when they experience symptoms (De Wit et al. 2016). SARS-CoV-2 is present in the upper respiratory tract and can be readily transmitted prior to the onset of symptoms. Mild but infectious cases may substantially contribute to its spread (Rivett et al. 2020).

Although further research will be required to elucidate in detail, which DCPs are responsible for which differences in virus behaviour, our analysis has already provided important clues. Both viruses use ACE2 as a receptor and are activated by the transmembrane serine protease TMPRSS2 (W. Li et al. 2003; J. Cui, Li, and Shi 2019; Hoffmann, Kleine-Weber, et al. 2020; Walls et al. 2020; Wan et al. 2020; Wrapp et al. 2020; Yan et al. 2020). Our results show, however, that the ACE2 and the TMPRSS2 status are not sufficient to predict cells susceptibility to SARS-CoV-2 or SARS-CoV. The cell line CL14 supported SARS-CoV-2 replication, although it displayed lower ACE2 levels and similar TMPRSS2 levels to non-susceptible DLD-1 and HT29 cells. Thus, attempts to identify SARS-

CoV-2 target cells based on the ACE2 status (Luan et al. 2020; Qiu et al. 2020) need to be considered with caution.

As previously described (Kamitani et al. 2006), ACE2 expression rendered SARS-CoV non-permissive 293 cells susceptible to SARS-CoV. However, ACE2 expression had a substantially lower impact on SARS-CoV-2 infection. This suggests the presence of further host cell factors that determine SARS-CoV-2 susceptibility. Based on our sequence analysis, DCPs in the viral interferon antagonists may contribute to the differences observed in the cellular tropism of SARS-CoV-2 and SARS-CoV.

Our computational analysis detected DCPs in the ACE2-binding domain of S, which are likely to impact S-ACE2 binding. In agreement, an anti-ACE2 antibody displayed higher efficacy against SARS-CoV than against SARS-CoV-2, illustrating the differences between SARS-CoV-2 S and SARS-CoV S interaction with ACE2. This probably reflects an increased SARS-CoV-2 S affinity to ACE2 compared to SARS-CoV S (Wrapp et al. 2020), which may be more difficult to antagonise.

To mediate virus entry, S needs to be cleaved by host cell proteases, in particular by TMPRSS2 (Y. Zhou et al. 2015; Hoffmann, Kleine-Weber, et al. 2020; Hoffmann, Schroeder, et al. 2020). The S cleavage sites are conserved between SARS-CoV-2 and SARS-CoV. However, we found DCPs in close vicinity to the S cleavage sites, which are likely to affect S cleavage by host cell enzymes and/ or the activity of protease inhibitors on S cleavage. Indeed, the serine protease inhibitors camostat and nafamostat, which interfere with S cleavage (Hoffmann, Kleine-Weber, et al. 2020; Hoffmann, Schroeder, et al. 2020), displayed increased activity against SARS-CoV-2 infection than against SARS-CoV infection, confirming the functional relevance of the DCPs.

In conclusion, our in-silico study revealed a substantial number of differentially conserved amino acid positions in the SARS-CoV-2 and SARS-CoV proteins. In agreement, cell culture experiments indicated differences in the cell tropism of these two viruses and showed that cellular ACE2 and TMPRSS2 levels do not reliably indicate cell susceptibility to SARS-CoV-2. Moreover, we identified DCPs in S that are associated with differences in the interaction with ACE2 and increased SARS-CoV-2 sensitivity to the protease inhibitors camostat and nafamostat relative to SARS-CoV.

# Chapter 5: General Discussion

This thesis explores themes of elucidating protein function and specific protein properties related to select amino acids. This chapter comments on the relevance of these works, their development, and the research implications on future projects.

## 5.1 Advances in Protein-Ligand Binding

Despite significant advances in the last few decades, both in terms of cost and speed, in determining sequence information, the link between sequence and its function still requires research. Current costs of determining protein function experimentally are compounded by labour intensive and often difficult procedures that are a strongly limiting factor in performing function determination experiments. Further, due to the vast number of genes with unknown functions, genes that appear less interesting are unlikely to be investigated. The need for automatic annotation of these sequences is vital to improving our understanding of pathways and systems, with the mammoth aim of elucidating the role of all coding DNA.

The research performed in Chapter 2 showcases the methodology we have developed, 3DLigandsite, and its new features, which should increase the amount of functionally annotated data and improve the accuracy of protein-ligand binding predictions. By hosting this tool as a webserver, users are able to identify likely binding regions, probable ligands, surface binding residues, and establish a priority system for experimentally identifying binding residues as guided by the probability scores we associate to each amino acid.

The method we describe does have its limitations. For example, a lack of homologous structures to a given input query will likely return no results. However, we have minimised the effect of this limitation by firstly allowing users to reduce the HHSearch probability threshold, thus allowing for more distantly related proteins to be captured in the initial homology scan. This works automatically on the webserver when there are no results available at a user set threshold, reducing the probability score by 5 % until a minimum of 60 % is reached. In addition, the option of a structural search is provided. A structural search compares structures containing biologically relevant ligands to the query structure.

Structures can be more conserved than sequences, and therefore may identify homologs where a sequence search does not. However, it is a much slower process than the sequence search.

Another limitation is the clustering of likely irrelevant ligands in the context of the query structure. In solved PDB structures, often ligands are added to aid in crystallisation and are subsequently retained in the structures when the models are deposited in data banks. Despite our database only recognising biologically relevant ligands, there is some overlap between these ligands and common stabilising agents. Through a propagation effect, these ligands carry over onto our predictions. Submissions with low homology to other structures can show a high ranked binding site that is not likely. For most predictions, these effects are diminished due to the presence of true binding sites, so the propagated ligands are lost or shown as a low confidence binding site.

Further, the dynamic nature of proteins, and the conformational rearrangement that often occurs on ligand binding, may result in a query structure that is not arranged in a bound state, and therefore predicted residues may differ. The difference will likely be minimal, but a method of optimising the structure of the protein to a cluster of ligands could be implemented in future releases. Comparison of the binding site in the query structure and homologous protein structures from the PDB or from AlphaFold to determine the best pose for ligand-binding could be implemented. Further, the non-metal ligands could be docked into the binding site to optimise the interaction.

These limitations highlight that improvements are still necessary to fulfil the characterisation of binding sites on protein structures; however, the tool we have developed addresses most key statements outlined in the last assessment of ligand binding predictions in CASP10. Metal and non-metal ligands are now clustered separately – allowing for more refined clusters are preventing multiple metal ligands spanning across a non-metal binding site. A probability score is now assigned for each amino acid in the binding site – based on key attributes of the amino acid and its position, and this feature can aid researchers in prioritising residues to investigate. In addition, solvent accessibility is now included for the user to aid in establishing which positions are likely targetable.

These features, and a clear user interface, will reduce the time that experimental procedures take by providing a basis for what an interaction between a query protein and ligands may look like. Improvements will continue to take place in this field. The recent publication of predicted structures for the human proteome and other model organisms released by AlphaFold has drastically increased our understanding of human coding sequences and is a landmark in structural biology and bioinformatics. Development of methods to identify the functions of these predicted structures are key, and tools such as that described in Chapter 2 will play a key role in this, however steps towards homology independent determination of ligand binding should be explored.

## 5.2 Elucidation of sequence positions associated with muscle contraction velocity

Public access to vast data and sequencing resources drove the research in Chapter 3. Sequence data for enough mammalian species to perform inter-and intra-clade comparisons of myosin-II proteins would not have previously been possible. The utilisation of these resources, combined with a measurable phenotypic feature (mass), enabled the identification of sequence positions that likely correlated with the contraction velocity of the heart.

Though the function of mammalian conventional myosins is well studied, the sequential components that exert control over the mechanisms behind the nuances of muscle contraction require further understanding. Methods of determining the role of amino acids in a given sequence are required to fully elucidate the protein's function and decipher positions with structural roles, binding propensity, and regulation, amongst other functions. Typical experimental characterisation of these sites is performed by associating variants to disease, such as in cardiomyopathies in beta-cardiac myosin or by mutagenesis experiments. We have demonstrated that the alignment of these proteins and statistical methods that assess variation at alignment positions can be a powerful tool for investigating the relevance of these positions in proteins. Applying this methodology to proteins participating in mechanisms with a measurable phenotypic variable could shine a light on the function of sequence positions and narrow down the possible positions

involved in said function. The guidance of experimental work by computational methods is vital for allowing research to be performed, making expensive techniques cheaper by narrowing the scope and reducing the number of tests that may otherwise need to be performed.

The computational method performed in Chapter 3 is limited by the availability of sequencing and phenotypic data available for all organisms in the study. Lack of reliable sources for species masses could impact the binomial regression analysis, though it likely has a minimal impact as a log scale is utilised, and only a large divergence in mass would drastically affect results. Further, as more species of mammalian clades are sequenced as part of global sequencing initiatives, supporting information could be incorporated into the analyses we performed. This could elucidate additional positions that transition to a different amino acid at mass thresholds. Akin to this, only mammalian clades are included in this study, and it would be interesting to identify if similar trends are observed in distinct clades.

It is clear that sequence positions distinct from the ATP binding site in beta cardiac myosin influence the rate at which ATP is hydrolysed in mammals. By changing just nine sites in human beta cardiac myosin to those of the rat equivalent, the behaviour of the ATPase was affected. The combination of dry lab and wet lab practices expounded the role of key sequence positions in beta-cardiac myosin. Additional experiments to identify the contribution of each sequence position mutated in the chimera to the ADP release rate would further increase our understanding of the mechanism behind the altered release rate.

## 5.3 Functional Impact of DCPs in Viruses

The impact of viruses on the human population has never been more in the public eye and the tracking of virus changes is of great import. The pandemic caused by SARS-CoV-2 has shown how quickly a new virus can ravage populations across continents. Perhaps the more worrying aspect of this virus is its similarity to the much more lethal SARS-CoV, making epidemiology essential for predicting, managing, and preventing disease outbreak.

Coronavirus is one of the greatest challenges of this generation, yet it has galvanised the global scientific community to better understand the virus, how to control its spread. Widespread sequencing of SARS-CoV-2 genomes and their subsequent deposition into data archives, such as GISAID, enable real-time tracking of variants in the coronavirus genome and the classification of variants into phylogenetic groupings. Identification of proteins encoded by the genome of SARS-CoV-2 and subsequent solving of the protein structures provided a platform for the effects of variants to be assessed. Where experimental and traditional homology modelling methods failed to produce structural models, novel deep learning-based methods provided protein structures. Remarkably, due to the vast amount of information gained from worldwide studies on coronavirus, many groups were able to mass produce viable vaccines in the fastest conception-to-mass-dispersion of a vaccine seen in history, where at least 17 COVID-19 vaccines have been developed. This consorted effort demonstrates the effectiveness of modern science when pointed towards a common purpose.

Early indications of a coronavirus pandemic raised fears that it could carry a significantly higher mortality rate in humans than we now know. The SARS pandemic in 2002/2004 resulted in an 11 % mortality rate (Chan-Yeung and Xu 2003), with 10-20 % of cases culminating in the need for mechanical ventilators (Chu et al. 2008) . With the greater capability for transmission, the SARS-CoV-2 pandemic could have been far worse if the virus displayed a similar severity of the disease. Sharing 80 % sequence similarity with the SARS-CoV virus, SARS-CoV-2 variants must continue to be monitored and assessed to prevent further cases and deaths.

In Chapter 4, a pipeline to identify differentially conserved positions between SARS-CoV and SARS-CoV-2 is introduced. In these works, we identified 186 positions in the spike protein that are likely responsible for the phenotypic differences between the viruses. By assessing the impact of these DCPs and identifying important functional areas near the DCPs, key insights into the differences between the two viruses can be made. An altered mode of binding to the ACE2 receptor was identified as shown by DCPs in the binding site and several other studies. DCPs were also found close to the S cleavage site, which is targeted by serine protease drugs camostat and nafamostat. These drugs showed increased activity against SARS-CoV-2 than SARS-CoV, highlighting the relevance of DCPs in

virus function. The method described in Chapter 4 is available as a webserver that allows users to submit SARS-CoV and SARS-CoV-2 sequences and identifies DCPs. This tool is useful for the ongoing analysis of coronavirus variants, where the loss or gain of DCPs may directly relate to the phenotype of the virus. In addition, it will allow researchers to easily categorise differences between their chosen input sequences, which has already proven a valuable tool in the analysis of viral species.

The works performed in this thesis demonstrate the value of combined computational and experimental analysis and the vital role that bioinformatics has in streamlining research. The continued development of computational methods to predict and assign function to biological systems will increase our knowledge and enable more targeted, detailed research to occur. We already see the great impact that machine learning has had on the field of biology, and it is likely its impact will only grow.

# References

Abdellah, Zahra, Alireza Ahmadi, Shahana Ahmed, Matthew Aimable, Rachael Ainscough, Jeff Almeida, Claire Almond, et al. 2004. “Finishing the Euchromatic Sequence of the Human Genome.” *Nature* 431 (7011): 931–45. https://doi.org/10.1038/nature03001.

Adams, Paul D., Pavel V. Afonine, Kumaran Baskaran, Helen M. Berman, John Berrisford, Gerard Bricogne, David G. Brown, et al. 2019. “Announcing Mandatory Submission of PDBx/MmCIF Format Files for Crystallographic Depositions to the Protein Data Bank (PDB).” *Acta Crystallographica Section D: Structural Biology*. https://doi.org/10.1107/S2059798319004522.

Adhikari, Arjun S., Kristina B. Kooiker, Saswata S. Sarkar, Chao Liu, Daniel Bernstein, James A. Spudich, and Kathleen M. Ruppel. 2016. “Early-Onset Hypertrophic Cardiomyopathy Mutations Significantly Increase the Velocity, Force, and Actin-Activated ATPase Activity of Human β-Cardiac Myosin.” *Cell Reports* 17 (11): 2857–64. https://doi.org/10.1016/j.celrep.2016.11.040.

Aksel, Tural, Elizabeth ChoeYu, Shirley Sutton, Kathleen M. Ruppel, and James A. Spudich. 2015. “Ensemble Force Changes That Result from Human Cardiac Myosin Mutations and a Small-Molecule Effector.” *Cell Reports* 11 (6): 910–20. https://doi.org/10.1016/j.celrep.2015.04.006.

Al-Karadaghi, Salam. n.d. “Basic Characteristics of the 20 Amino Acids: Hydrophobic, Hydrophilic, Polar and Charged.” Accessed February 6, 2022. https://proteinstructures.com/structure/amino-acids/.

Alm, Erik, Eeva K. Broberg, Thomas Connor, Emma B. Hodcroft, Andrey B. Komissarov, Sebastian Maurer-Stroh, Angeliki Melidou, et al. 2020. “Geographical and Temporal Distribution of SARS-CoV-2 Clades in the WHO European Region, January to June 2020.” *Eurosurveillance*. https://doi.org/10.2807/1560-7917.ES.2020.25.32.2001410.

Alquraishi, Mohammed. 2019. “AlphaFold at CASP13.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz422.

Amrine-Madsen, Heather, Klaus Peter Koepfli, Robert K. Wayne, and Mark S. Springer. 2003. “A New Phylogenetic Marker, Apolipoprotein B, Provides Compelling Evidence for Eutherian Relationships.” In *Molecular Phylogenetics and Evolution*. https://doi.org/10.1016/S1055-7903(03)00118-0.

Andreeva, Antonina, Eugene Kulesha, Julian Gough, and Alexey G. Murzin. 2020. “The SCOP Database in 2020: Expanded Classification of Representative Family and Superfamily Domains of Known Protein Structures.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz1064.

Andreini, Claudia, Ivano Bertini, Gabriele Cavallaro, Gemma L. Holliday, and Janet M. Thornton. 2008. “Metal Ions in Biological Catalysis: From Enzyme Databases to General Principles.” *Journal of Biological Inorganic Chemistry*. https://doi.org/10.1007/s00775-008-0404-5.

Andreini, Claudia, Gabriele Cavallaro, Serena Lorenzini, and Antonio Rosato. 2013.

"MetalPDB: A Database of Metal Sites in Biological Macromolecular Structures." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks1063.

Antczak, Magdalena, Martin Michaelis, and Mark N. Wass. 2019. "Environmental Conditions Shape the Nature of a Minimal Bacterial Genome." *Nature Communications*. https://doi.org/10.1038/s41467-019-10837-2.

Armstrong, David R., John M. Berrisford, Matthew J. Conroy, Aleksandras Gutmanas, Stephen Anyango, Preeti Choudhary, Alice R. Clark, et al. 2020. "PDBe: Improved Findability of Macromolecular Structure Data in the PDB." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz990.

Árnason, Úlfur, Fritjof Lammers, Vikas Kumar, Maria A. Nilsson, and Axel Janke. 2018. "Whole-Genome Sequencing of the Blue Whale and Other Rorquals Finds Signatures for Introgressive Gene Flow." *Science Advances*. https://doi.org/10.1126/sciadv.aap9873.

Astuti, Indwiani, and Ysrafil. 2020. "Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An Overview of Viral Structure and Host Response." *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* 14 (4): 407–12. https://doi.org/10.1016/j.dsx.2020.04.020.

Auton, Adam, and Tovah Salcedo. 2015. "The 1000 Genomes Project." In *Assessing Rare Variation in Complex Traits: Design and Analysis of Genetic Studies*. https://doi.org/10.1007/978-1-4939-2824-8_6.

Babu, Gopal J., David M. Warshaw, and Muthu Periasamy. 2000. "Smooth Muscle Myosin Heavy Chain Isoforms and Their Role in Muscle Physiology." *Microscopy Research and Technique*. https://doi.org/10.1002/1097-0029(20000915)50:6<532::AID-JEMT10>3.0.CO;2-E.

Ballouz, Sara, Alexander Dobin, and Jesse A. Gillis. 2019. "Is It Time to Change the Reference Genome?" *Genome Biology*. https://doi.org/10.1186/s13059-019-1774-4.

Barber-Zucker, Shiran, Boaz Shaanan, and Raz Zarivach. 2017. "Transition Metal Binding Selectivity in Proteins and Its Correlation with the Phylogenomic Classification of the Cation Diffusion Facilitator Protein Family." *Scientific Reports* 7 (1): 1–12. https://doi.org/10.1038/s41598-017-16777-5.

Barnes, Michael R., and Ian C. Gray. 2003. *Bioinformatics for Geneticists*. Wiley.

Barth, P., B. Wallner, and D. Baker. 2009. "Prediction of Membrane Protein Structures with Complex Topologies Using Limited Constraints." *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0808323106.

Barua, Bipasha. 2013. "Periodicities Designed in the Tropomyosin Sequence and Structure Define Its Functions." *BioArchitecture*. https://doi.org/10.4161/bioa.25616.

Barua, Bipasha, Donald A. Winkelmann, Howard D. White, and Sarah E. Hitchcock-DeGregori. 2012. "Regulation of Actin-Myosin Interaction by Conserved Periodic Sites of Tropomyosin." *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1212754109.

Bateman, Alex, Maria Jesus Martin, Sandra Orchard, Michele Magrane, Rahat Agivetova, Shadab Ahmad, Emanuele Alpi, et al. 2021. “UniProt: The Universal Protein Knowledgebase in 2021.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa1100.

Berg, J. S., B. C. Powell, and R. E. Cheney. 2001. “A Millennial Myosin Census.” *Molecular Biology of the Cell*. https://doi.org/10.1091/mbc.12.4.780.

Berman, H M, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. 2000. “The Protein Data Bank.” *Nucleic Acids Research* 28 (1): 235–42. https://doi.org/10.1093/nar/28.1.235.

Bernkopf, Marie, Gerald Webersinke, Chanakan Tongsook, Chintan N. Koyani, Muhammad A. Rafiq, Muhammad Ayaz, Doris Müller, et al. 2014. “Disruption of the Methyltransferase-like 23 Gene METTL23 Causes Mild Autosomal Recessive Intellectual Disability.” *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddu115.

Bernstein, Sanford I., and Ronald A. Milligan. 1997. “Fine Tuning a Molecular Motor: The Location of Alternative Domains in the Drosophila Myosin Head.” *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.1997.1160.

Betapudi, Venkaiah. 2014. “Life without Double-Headed Non-Muscle Myosin II Motor Proteins.” *Frontiers in Chemistry*. https://doi.org/10.3389/fchem.2014.00045.

Bicer, Sabahattin, and Peter J. Reiser. 2007. “Variations in Apparent Mass of Mammalian Fast-Type Myosin Light Chains Correlate with Species Body Size, from Shrew to Elephant.” *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*. https://doi.org/10.1152/ajpregu.00098.2006.

Blackburn, Tim M., and Kevin J. Gaston. 1998. “The Distribution of Mammal Body Masses.” *Diversity and Distributions*. https://doi.org/10.1046/j.1365-2699.1998.00015.x.

Blankenfeldt, Wulf, Nicolas H. Thomä, John S. Wray, Mathias Gautel, and Ilme Schlichting. 2006. “Crystal Structures of Human Cardiac β-Myosin II S2-Δ Provide Insight into the Functional Role of the S2 Subfragment.” *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0606741103.

Bloemink, M., J. Deacon, D. Resnicow, LA. Leinwand, and MA. Geeves. 2013. “The Superfast Human Extraocular Myosin Is Kinetically Distinct from the Fast Skeletal IIa, IIb, and IId Isoforms.” *The Journal of Biological Chemistry* 288 (38): 27469–79. https://doi.org/10.1074/jbc.M113.488130.

Bloemink, M. J., N. Adamek, C. Reggiani, and M. A. Geeves. 2007. “Kinetic Analysis of the Slow Skeletal Myosin MHC-1 Isoform from Bovine Masseter Muscle.” *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2007.08.050.

Bloemink, M.J, C.M Dambacher, A.F Knowles, G.C Melkani, M.A Geeves, and S.I Bernstein. 2009. “Alternative Exon 9-Encoded Relay Domains Affect More than One Communication Pathway in the Drosophila Myosin Head.” *Journal of Molecular Biology* 389 (4): 707–21. https://doi.org/10.1016/J.JMB.2009.04.036.

Bloemink, M, and MA Geeves. 2011. “Shaking the Myosin Family Tree: Biochemical Kinetics Defines Four Types of Myosin Motor.” *Seminars in Cell and Developmental Biology*. https://doi.org/10.1016/j.semcdb.2011.09.015.

Bloemink, Marieke J., and Michael A. Geeves. 2011. “Shaking the Myosin Family Tree: Biochemical Kinetics Defines Four Types of Myosin Motor.” *Seminars in Cell and Developmental Biology*. https://doi.org/10.1016/j.semcdb.2011.09.015.

Bloemink, MJ, GC Melkani, CM Dambacher, SI Bernstein, and MA Geeves. 2011. “Two Drosophila Myosin Transducer Mutants with Distinct Cardiomyopathies Have Divergent ADP and Actin Affinities.” *The Journal of Biological Chemistry* 286 (32): 28435–43. https://doi.org/10.1074/JBC.M111.258228.

Bojkova, Denisa, Marco Bechtel, Katie May McLaughlin, Jake E. McGreig, Kevin Klann, Carla Bellinghausen, Gernot Rohde, et al. 2020. “Aprotinin Inhibits SARS-CoV-2 Replication.” *Cells*. https://doi.org/10.3390/cells9112377.

Bonferroni, C. E. 1936. “1936 Teoria Statistica Delle Classi e Calcolo Delle Probabilità.” *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commerciali Di Firenze*.

Borges do Nascimento, Israel Júnior, Nensi Cacic, Hebatullah Mohamed Abdulazeem, Thilo Caspar von Groote, Umesh Jayarajah, Ishanka Weerasekara, Meisam Abdar Esfahani, et al. 2020. “Novel Coronavirus Infection (COVID-19) in Humans: A Scoping Review and Meta-Analysis.” *Journal of Clinical Medicine*. https://doi.org/10.3390/jcm9040941.

Bottinelli, R., and C. Reggiani. 2000. “Human Skeletal Muscle Fibres: Molecular and Functional Diversity.” *Progress in Biophysics and Molecular Biology*. https://doi.org/10.1016/S0079-6107(00)00006-7.

Bouckaert, Remco, Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, et al. 2019. “BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis.” *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1006650.

Brecht, Michael, Robert Naumann, Farzana Anjum, Jason Wolfe, Martin Munz, Carolin Mende, and Claudia Roth-Alpermann. 2011. “The Neurobiology of Etruscan Shrew Active Touch.” *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2011.0160.

Buchan, D W A, S M Ward, A E Lobley, T C O Nugent, K Bryson, and D T Jones. 2010. “Protein Annotation and Modelling Servers at University College London.” *Nucleic Acids Research* 38 (Web Server issue): W563-8. https://doi.org/10.1093/nar/gkq427.

Burgin, Connor J., Jocelyn P. Colella, Philip L. Kahn, and Nathan S. Upham. 2018. “How Many Species of Mammals Are There?” *Journal of Mammalogy*. https://doi.org/10.1093/jmammal/gyx147.

Canepari, M., M. A. Pellegrino, G. D’Antona, and R. Bottinelli. 2010. “Skeletal Muscle Fibre Diversity and the Underlying Mechanisms.” *Acta Physiologica*. https://doi.org/10.1111/j.1748-1716.2010.02118.x.

Capra, John A., Roman A. Laskowski, Janet M. Thornton, Mona Singh, and Thomas A. Funkhouser. 2009. “Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure.” *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1000585.

Capra, John A., and Mona Singh. 2007a. “Predicting Functionally Important Residues from Sequence Conservation.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btm270.

———. 2008. “Characterization and Prediction of Residues Determining Protein Functional Specificity.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btn214.

Capra, John A, and Mona Singh. 2007b. “Predicting Functionally Important Residues from Sequence Conservation.” *Bioinformatics (Oxford, England)* 23 (15): 1875–82. https://doi.org/10.1093/bioinformatics/btm270.

Casari, Georg, Chris Sander, and Alfonso Valencia. 1995. “A Method to Predict Functional Residues in Proteins.” *Nature Structural Biology*. https://doi.org/10.1038/nsb0295-171.

Castresana, Jose. 2000. “Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis.” *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

Chakraborti, Samitabh, Ponraj Prabakaran, Xiaodong Xiao, and Dimiter S. Dimitrov. 2005. “The SARS Coronavirus S Glycoprotein Receptor Binding Domain: Fine Mapping and Functional Characterization.” *Virology Journal*. https://doi.org/10.1186/1743-422X-2-73.

Chambers, J. C., W. Zhang, J. S. Sehmi, X. Li, M. N. Wass, P. Van der Harst, H. Holm, et al. 2011. “Genome-Wide Association Study Identifies Loci Influencing Concentrations of Liver Enzymes in Plasma.” *Nature Genetics*. https://doi.org/10.1038/ng.970.

Chan-Yeung, Moira, and Rui Heng Xu. 2003. “SARS: Epidemiology.” *Respirology*. https://doi.org/10.1046/j.1440-1843.2003.00518.x.

Chang, Chung Ke, Shih Che Sue, Tsan Hung Yu, Chiu Min Hsieh, Cheng Kun Tsai, Yen Chieh Chiang, Shin Jye Lee, et al. 2006. “Modular Organization of SARS Coronavirus Nucleocapsid Protein.” *Journal of Biomedical Science*. https://doi.org/10.1007/s11373-005-9035-9.

Chen, Junjie, Mingyue Guo, Xiaolong Wang, and Bin Liu. 2018. “A Comprehensive Review and Comparison of Different Computational Methods for Protein Remote Homology Detection.” *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbw108.

Chen, Meng Yun, Dan Liang, and Peng Zhang. 2017. “Phylogenomic Resolution of the Phylogeny of Laurasiatherian Mammals: Exploring Phylogenetic Signals within Coding and Noncoding Sequences.” *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evx147.

Cheng, Vincent C.C., Jasper F.W. Chan, Kelvin K.W. To, and K. Y. Yuen. 2013. “Clinical Management and Infection Control of SARS: Lessons Learned.” *Antiviral Research*.

https://doi.org/10.1016/j.antiviral.2013.08.016.

Chu, Dachen, Ran Chou Chen, Chia Yu Ku, and Pesus Chou. 2008. “The Impact of SARS on Hospital Performance.” *BMC Health Services Research*. https://doi.org/10.1186/1472-6963-8-228.

Cinatl, J., G. Hoever, B. Morgenstern, W. Preiser, J. U. Vogel, W. K. Hofmann, G. Bauer, M. Michaelis, H. F. Rabenau, and H. W. Doerr. 2004. “Infection of Cultured Intestinal Epithelial Cells with Severe Acute Respiratory Syndrome Coronavirus.” *Cellular and Molecular Life Sciences*. https://doi.org/10.1007/s00018-004-4222-9.

Cinatl, J., B. Morgenstern, G. Bauer, P. Chandra, H. Rabenau, and H. W. Doerr. 2003. “Glycyrrhizin, an Active Component of Liquorice Roots, and Replication of SARS-Associated Coronavirus.” *Lancet*. https://doi.org/10.1016/S0140-6736(03)13615-X.

Cinatl, Jindrich, Martin Michaelis, Birgit Morgenstern, and Hans Wilhelm Doerr. 2005. “High-Dose Hydrocortisone Reduces Expression of the pro-Inflammatory Chemokines CXCL8 and CXCL10 in SARS Coronavirus-Infected Intestinal Cells.” *International Journal of Molecular Medicine*. https://doi.org/10.3892/ijmm.15.2.323.

Consortium, UniProt. 2017. “UniProt: The Universal Protein Knowledgebase.” *Nucleic Acids Research* 45 (D1): D158–69. https://doi.org/10.1093/nar/gkw1099.

Cooke, Roger. 2004. “The Sliding Filament Model: 1972–2004.” *Journal of General Physiology*.

Corman, Victor M., Olfert Landt, Marco Kaiser, Richard Molenkamp, Adam Meijer, Daniel K.W. Chu, Tobias Bleicker, et al. 2020. “Detection of 2019 Novel Coronavirus (2019-NCoV) by Real-Time RT-PCR.” *Eurosurveillance*. https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” *Machine Learning*. https://doi.org/10.1023/A:1022627411411.

Cox, D. R. 1959. “The Regression Analysis of Binary Sequences.” *Journal of the Royal Statistical Society: Series B (Methodological)*. https://doi.org/10.1111/j.2517-6161.1959.tb00334.x.

Criddle, A H, M A Geeves, and T Jeffries. 1985. “The Use of Actin Labelled with N-(1-Pyrenyl)Iodoacetamide to Study the Interaction of Actin with Myosin Subfragments and Troponin/Tropomyosin.” *The Biochemical Journal* 232 (2): 343–49.

Csaba, Gergely, Fabian Birzele, and Ralf Zimmer. 2009. “Systematic Comparison of SCOP and CATH: A New Gold Standard for Protein Structure Analysis.” *BMC Structural Biology*. https://doi.org/10.1186/1472-6807-9-23.

Cui, Jie, Fang Li, and Zheng Li Shi. 2019. “Origin and Evolution of Pathogenic Coronaviruses.” *Nature Reviews Microbiology*. https://doi.org/10.1038/s41579-018-0118-9.

Cui, Yifeng, Qiwen Dong, Daocheng Hong, and Xikun Wang. 2019. “Predicting Protein-Ligand Binding Residues with Deep Convolutional Neural Networks.” *BMC Bioinformatics*. https://doi.org/10.1186/s12859-019-2672-1.

Das, Karuna M., Edward Y. Lee, Suhayla E. Al Jawder, Mushira A. Enani, Rajvir Singh, Leila Skakni, Nizar Al-Nakshabandi, Khalid AlDossari, and Sven G. Larsson. 2015. “Acute Middle East Respiratory Syndrome Coronavirus: Temporal Lung Changes Observed on the Chest Radiographs of 55 Patients.” *American Journal of Roentgenology*. https://doi.org/10.2214/AJR.15.14445.

Das, Sayoni, David Lee, Ian Sillitoe, Natalie L. Dawson, Jonathan G. Lees, and Christine A. Orengo. 2015. “Functional Classification of CATH Superfamilies: A Domain-Based Approach for Protein Function Annotation.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv398.

Dawson, Natalie L., Tony E. Lewis, Sayoni Das, Jonathan G. Lees, David Lee, Paul Ashford, Christine A. Orengo, and Ian Sillitoe. 2017. “CATH: An Expanded Resource to Predict Protein Function through Structure and Sequence.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1098.

Deacon, John C, Marieke J Bloemink, Heresh Rezavandi, Michael A Geeves, and Leslie A Leinwand. 2012. “Erratum to: Identification of Functional Differences between Recombinant Human α and β Cardiac Myosin Motors.” *Cellular and Molecular Life Sciences : CMLS* 69 (24): 4239–55. https://doi.org/10.1007/s00018-012-1111-5.

Devuyst, Olivier. 2015. “The 1000 Genomes Project: Welcome to a New World.” *Peritoneal Dialysis International*. https://doi.org/10.3747/pdi.2015.00261.

Dobson, Christopher M. 2004. “Principles of Protein Folding, Misfolding and Aggregation.” In *Seminars in Cell and Developmental Biology*, 15:3–16. Elsevier Ltd. https://doi.org/10.1016/j.semcdb.2003.12.008.

Dong, Ensheng, Hongru Du, and Lauren Gardner. 2020. “An Interactive Web-Based Dashboard to Track COVID-19 in Real Time.” *The Lancet Infectious Diseases*. https://doi.org/10.1016/S1473-3099(20)30120-1.

Douady, C. J. 2003. “Comparison of Bayesian and Maximum Likelihood Bootstrap Measures of Phylogenetic Reliability.” *Molecular Biology and Evolution* 20 (2): 248–54. https://doi.org/10.1093/molbev/msg042.

Drummond, Alexei J., Marc A. Suchard, Dong Xie, and Andrew Rambaut. 2012. “Bayesian Phylogenetics with BEAUti and the BEAST 1.7.” *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/mss075.

“E Pluribus Unum.” 2010. *Nature Methods*. https://doi.org/10.1038/nmeth0510-331.

Eddy, Sean R. 1998. “Profile Hidden Markov Models.” *Bioinformatics*. Oxford University Press. https://doi.org/10.1093/bioinformatics/14.9.755.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

Fehr, Anthony R., and Stanley Perlman. 2015. “Coronaviruses: An Overview of Their Replication and Pathogenesis.” In *Coronaviruses: Methods and Protocols*. https://doi.org/10.1007/978-1-4939-2438-7_1.

Feinstein, Wei P., and Michal Brylinski. 2014. “EFindSite: Enhanced Fingerprint-Based Virtual Screening against Predicted Ligand Binding Sites in Protein Models.” *Molecular Informatics*. https://doi.org/10.1002/minf.201300143.

Felsenstein, J. 1985. “Phylogenies and the Comparative Method.” *American Naturalist*. https://doi.org/10.1086/284325.

Foley, Nicole M., Mark S. Springer, and Emma C. Teeling. 2016. “Mammal Madness: Is the Mammal Tree of Life Not yet Resolved?” *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2015.0140.

Frauenfelder, H, and P G Wolynes. 1985. “Rate Theories and Puzzles of Hemeprotein Kinetics.” *Science (New York, N.Y.)* 229 (4711): 337–45. https://doi.org/10.1126/science.4012322.

Freeman, Kalev, Koichi Nakao, and Leslie A. Leinwand. 2001. “Low Sequence Variation in the Gene Encoding the Human β-Myosin Heavy Chain.” *Genomics*. https://doi.org/10.1006/geno.2001.6573.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts565.

Gallo Cassarino, Tiziano, Lorenza Bordoli, and Torsten Schwede. 2014. “Assessment of Ligand Binding Site Predictions in CASP10.” *Proteins: Structure, Function and Bioinformatics*. https://doi.org/10.1002/prot.24495.

Geeves, M. A., and K. C. Holmes. 1999. “Structural Mechanism of Muscle Contraction.” *Annual Review of Biochemistry*. https://doi.org/10.1146/annurev.biochem.68.1.687.

Gelb, Bruce D., and Stephanie E. Chin. 2012. “Genetics of Congenital Heart Disease.” In *Muscle*. https://doi.org/10.1016/B978-0-12-381510-1.00034-X.

Genomics England Ltd. 2017. “Genomics England and the 100,000 Genomes Project.” *Genomics England Website*.

Geurts, Pierre, Damien Ernst, and Louis Wehenkel. 2006. “Extremely Randomized Trees.” *Machine Learning*. https://doi.org/10.1007/s10994-006-6226-1.

GISAID. 2020. “GISAID Initiative.” *Advances in Virus Research*.

Golding, Brian, and Joe Felsenstein. 1990. “A Maximum Likelihood Approach to the Detection of Selection from a Phylogeny.” *Journal of Molecular Evolution*. https://doi.org/10.1007/BF02102078.

Gonçalves, Ricardo Lemes, Túlio César Rodrigues Leite, Bruna de Paula Dias, Camila Carla da Silva Caetano, Ana Clara Gomes de Souza, Ubiratan da Silva Batista, Camila Cavadas Barbosa, Arturo Reyes-Sandoval, Luiz Felipe Leomil Coelho, and Breno de Mello Silva. 2020. “SARS-CoV-2 Mutations and Where to Find Them: An in Silico Perspective of Structural Changes and Antigenicity of the Spike Protein.” *BioRxiv*. https://doi.org/10.1101/2020.05.21.108563.

Gorbalenya, A.E., Baker, S.C. et al. 2020. “Coronaviridae Study Group of the International

Committee on Taxonomy of Viruses." *Nat Microbiol*.

Greninger, Alexander L., Samia N. Naccache, Scot Federman, Guixia Yu, Placide Mbala, Vanessa Bres, Doug Stryke, et al. 2015. "Rapid Metagenomic Identification of Viral Pathogens in Clinical Samples by Real-Time Nanopore Sequencing Analysis." *Genome Medicine*. https://doi.org/10.1186/s13073-015-0220-9.

Gu, Jiang, and Christine Korteweg. 2007. "Pathology and Pathogenesis of Severe Acute Respiratory Syndrome." *American Journal of Pathology*. https://doi.org/10.2353/ajpath.2007.061088.

Gutfreund, H. 1995. *Kinetics for the Life Sciences : Receptors, Transmitters, and Catalysts*. Cambridge University Press.

Haase, Hannelore, and Ingo Morano. 1996. "Alternative Splicing of Smooth Muscle Myosin Heavy Chains and Its Functional Consequences." *Journal of Cellular Biochemistry*. https://doi.org/10.1002/(SICI)1097-4644(19960315)60:4<521::AID-JCB8>3.0.CO;2-U.

Hadfield, James, Colin Megill, Sidney M. Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A. Neher. 2018. "NextStrain: Real-Time Tracking of Pathogen Evolution." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty407.

Hannenhalli, Sridhar S., and Robert B. Russell. 2000. "Analysis and Prediction of Functional Sub-Types from Protein Sequence Alignments." *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.2000.4036.

Hanson, Jean, and Hugh E. Huxley. 1953. "Structural Basis of the Cross-Striations in Muscle." *Nature*. https://doi.org/10.1038/172530b0.

Harvey, Lodish, Berk Arnold, Zipursky S Lawrence, l Matsudaira Pau, Baltimore David, and Darnel James. 2000. *Molecular Cell Biology. 4th Edition*. *Journal of the American Society for Mass Spectrometry*. https://doi.org/10.1016/j.jasms.2009.08.001.

Heather, James M., and Benjamin Chain. 2016. "The Sequence of Sequencers: The History of Sequencing DNA." *Genomics*. https://doi.org/10.1016/j.ygeno.2015.11.003.

Hekkelman, Maarten L, Ida De Vries, Robbie P Joosten, and Anastassis Perrakis. 2021. "AlphaFill: Enriching the AlphaFold Models with Ligands and Co-Factors." *BioRxiv*, November, 2021.11.26.470110. https://doi.org/10.1101/2021.11.26.470110.

Henikoff, S., and J. G. Henikoff. 1992. "Amino Acid Substitution Matrices from Protein Blocks." *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.89.22.10915.

Ho, Cheng Hsun, and Wolf B. Frommer. 2014. "Fluorescent Sensors for Activity and Regulation of the Nitrate Transceptor CHL1/NRT1.1 and Oligopeptide Transporters." *ELife*. https://doi.org/10.7554/eLife.01917.

Ho, Tin Kam. 1995. "Random Decision Forests." In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. https://doi.org/10.1109/ICDAR.1995.598994.

Hoehl, Sebastian, Holger Rabenau, Annemarie Berger, Marhild Kortenbusch, Jindrich Cinatl, Denisa Bojkova, Pia Behrens, et al. 2020. “Evidence of SARS-CoV-2 Infection in Returning Travelers from Wuhan, China.” *New England Journal of Medicine*. https://doi.org/10.1056/nejmc2001899.

Hoffmann, Markus, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S. Schiergens, et al. 2020. “SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor.” *Cell*. https://doi.org/10.1016/j.cell.2020.02.052.

Hoffmann, Markus, Simon Schroeder, Hannah Kleine-Weber, Marcel A. Müller, Christian Drosten, and Stefan Pöhlmann. 2020. “Nafamostat Mesylate Blocks Activation of SARS-CoV-2: New Treatment Option for COVID-19.” *Antimicrobial Agents and Chemotherapy*. https://doi.org/10.1128/AAC.00754-20.

Homburger, Julian, Eric Green, Colleen Caleshu, Margaret Sunitha, Rebecca Taylor, Kathleen Ruppel, Raghu Metpally, et al. 2016. “Multi-Dimensional Structure Function Relationships in Human β-Cardiac Myosin from Population Scale Genetic Variation.” *Multi-Dimensional Structure Function Relationships in Human β-Cardiac Myosin from Population Scale Genetic Variation.* https://doi.org/10.1101/039321.

Hone, David W.E., and Michael J. Benton. 2005. “The Evolution of Large Size: How Does Cope’s Rule Work?” *Trends in Ecology and Evolution*. https://doi.org/10.1016/j.tree.2004.10.012.

Hu, Jing Yang, Ya Ping Zhang, and L. Yu. 2012. “Summary of Laurasiatheria (Mammalia) Phylogeny.” *Dong Wu Xue Yan Jiu = Zoological Research / “Dong Wu Xue Yan Jiu” Bian Ji Wei Yuan Hui Bian Ji*. https://doi.org/10.3724/sp.j.1141.2012.e05-06e65.

Huang, Jin. Nagy, Stanislav. Koide, Akiko. Rock, Ronald. Koide, Shohei. 2009. “A Peptide Tag System for Facile Purification and Single-Molecule Immobilization.” *Biochemistry* 48: 11834–36. https://doi.org/10.1021/bi901756n.A.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. “Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology.” *Science*. https://doi.org/10.1126/science.1065889.

Huelsenbeck, John P., and Fredrik Ronquist. 2001. “MRBAYES: Bayesian Inference of Phylogenetic Trees.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/17.8.754.

Huerta-Cepas, Jaime, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, et al. 2016. “EGGNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1248.

Hussain, Mushtaq, Nusrat Jabeen, Anusha Amanullah, Ayesha Ashraf Baig, Basma Aziz, Sanya Shabbir, Fozia Raza, and Nasir Uddin. 2020. “Molecular Docking between Human Tmprss2 and Sars-Cov-2 Spike Protein: Conformation and Intermolecular Interactions.” *AIMS Microbiology*. https://doi.org/10.3934/microbiol.2020021.

Huxley, A. F., and R. Niedergerke. 1954. “Structural Changes in Muscle during Contraction: Interference Microscopy of Living Muscle Fibres.” *Nature*. https://doi.org/10.1038/173971a0.

Hvidsten, Torgeir R., Astrid Lægreid, Andriy Kryshtafovych, Gunnar Anderson, Krzysztof Fidelis, and Jan Komorowski. 2009. “A Comprehensive Analysis of the Structure-Function Relationship in Proteins Based on Local Structure Similarity.” *PLoS ONE*. https://doi.org/10.1371/journal.pone.0006266.

Isabel, Sandra, Lucía Graña-Miraglia, Jahir M. Gutierrez, Cedoljub Bundalovic-Torma, Helen E. Groves, Marc R. Isabel, Ali Reza Eshaghi, et al. 2020. “Evolutionary and Structural Analyses of SARS-CoV-2 D614G Spike Protein Mutation Now Documented Worldwide.” *Scientific Reports*. https://doi.org/10.1038/s41598-020-70827-z.

Jain, Miten, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, et al. 2018. “Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads.” *Nature Biotechnology*. https://doi.org/10.1038/nbt.4060.

Jendele, Lukas, Radoslav Krivak, Petr Skoda, Marian Novotny, and David Hoksza. 2019. “PrankWeb: A Web Server for Ligand Binding Site Prediction and Visualization.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz424.

Jiménez, J., S. Doerr, G. Martínez-Rosell, A. S. Rose, and G. De Fabritiis. 2017. “DeepSite: Protein-Binding Site Predictor Using 3D-Convolutional Neural Networks.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btx350.

John Jumper, Kathryn Tunyasuvunakool, Pushmeet Kohli, Demis Hassabis, AlphaFold Team. 2020. “Computational Predictions of Protein Structures Associated with COVID-19.” 2020. https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19.

Johnson, Chloe A., Jake E. McGreig, Sarah T. Jeanfavre, Jonathan Walklate, Carlos D. Vera, Marta Farre, Daniel P. Mulvihill, et al. 2021. “Identification of Sequence Changes in Myosin II That Adjust Muscle Contraction Velocity.” *PLoS Biology*. https://doi.org/10.1371/journal.pbio.3001248.

Johnson, Matthew, Daniel A. East, and Daniel P. Mulvihill. 2014. “Formins Determine the Functional Properties of Actin Filaments in Yeast.” *Current Biology* 24 (13): 1525–30. https://doi.org/10.1016/j.cub.2014.05.034.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature*, July, 1–11. https://doi.org/10.1038/s41586-021-03819-2.

Jürgens, Klaus D., Roger Fons, Thomas Peters, and Susanne Sender. 1996. “Heart and Respiratory Rates and Their Significance for Convective Oxygen Transport Rates in the Smallest Mammal, the Etruscan Shrew Suncus Etruscus.” *Journal of Experimental Biology*. https://doi.org/10.1242/jeb.199.12.2579.

Kabsch, W, H G Mannherz, D Suck, E F Pai, and K C Holmes. 1990. “Atomic Structure of the

Actin:DNase I Complex [See Comments].” *Nature*.

Källberg, Morten, Haipeng Wang, Sheng Wang, Jian Peng, Zhiyong Wang, Hui Lu, and Jinbo Xu. 2012. “Template-Based Protein Structure Modeling Using the RaptorX Web Server.” *Nature Protocols* 7 (8): 1511–22. https://doi.org/10.1038/nprot.2012.085.

Kamiński, Bogumił, Michał Jakubczyk, and Przemysław Szufel. 2018. “A Framework for Sensitivity Analysis of Decision Trees.” *Central European Journal of Operations Research* 26 (1): 135–59. https://doi.org/10.1007/s10100-017-0479-6.

Kamitani, Wataru, Krishna Narayanan, Cheng Huang, Kumari Lokugamage, Tetsuro Ikegami, Naoto Ito, Hideyuki Kubo, and Shinji Makino. 2006. “Severe Acute Respiratory Syndrome Coronavirus Nsp1 Protein Suppresses Host Gene Expression by Promoting Host MRNA Degradation.” *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0603144103.

Kanavakis, E., and A. Xaidara. 2001. “The Human Genome Project.” *Archives of Hellenic Medicine*. National Institute on Alcohol Abuse and Alcoholism. https://doi.org/10.5005/jp/books/10279_22.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/mst010.

Kaur, Navpreet, Rimaljot Singh, Zahid Dar, Rakesh Kumar Bijarnia, Neelima Dhingra, and Tanzeer Kaur. 2021. “Genetic Comparison among Various Coronavirus Strains for the Identification of Potential Vaccine Targets of SARS-CoV2.” *Infection, Genetics and Evolution*. https://doi.org/10.1016/j.meegid.2020.104490.

Kelley, Lawrence A, Stefans Mezulis, Christopher M Yates, Mark N Wass, and Michael J E Sternberg. 2015. “The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis.” *Nature Protocols* 10 (6): 845–58. https://doi.org/10.1038/nprot.2015.053.

Kelly, L.A., S. Mezulis, C. Yates, M. Wass, and M. Sternberg. 2015. “The Phyre2 Web Portal for Protein Modelling, Prediction, and Analysis.” *Nature Protocols* 10 (6): 845–58. https://doi.org/10.1038/nprot.2015-053.

Khailany, Rozhgar A., Muhamad Safdar, and Mehmet Ozaslan. 2020. “Genomic Characterization of a Novel SARS-CoV-2.” *Gene Reports*. https://doi.org/10.1016/j.genrep.2020.100682.

Koyama, Takahiko, Daniel Platt, and Laxmi Parida. 2020. “Variant Analysis of SARS-Cov-2 Genomes.” *Bulletin of the World Health Organization*. https://doi.org/10.2471/BLT.20.253591.

Krendel, Mira, and Mark S. Mooseker. 2005. “Myosins: Tails (and Heads) of Functional Diversity.” *Physiology*. https://doi.org/10.1152/physiol.00014.2005.

Krivák, Radoslav, and David Hoksza. 2018. “P2Rank: Machine Learning Based Tool for Rapid and Accurate Prediction of Ligand Binding Sites from Protein Structure.” *Journal of Cheminformatics*. https://doi.org/10.1186/s13321-018-0285-8.

Kryshtafovych, Andriy, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult.

2019a. “Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XIII.” *Proteins: Structure, Function and Bioinformatics*. https://doi.org/10.1002/prot.25823.

———. 2019b. “Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XIII.” *Proteins: Structure, Function and Bioinformatics*. John Wiley and Sons Inc. https://doi.org/10.1002/prot.25823.

Kuhlmann, F. Matthew, John I. Robinson, Gregory R. Bluemling, Catherine Ronet, Nicolas Fasel, and Stephen M. Beverley. 2017. “Antiviral Screening Identifies Adenosine Analogs Targeting the Endogenous DsRNA Leishmania RNA Virus 1 (LRV1) Pathogenicity Factor.” *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1619114114.

Kumar, Swatantra, Rajni Nyodu, Vimal K. Maurya, and Shailendra K. Saxena. 2020. “Morphology, Genome Organization, Replication, and Pathogenesis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2).” In . https://doi.org/10.1007/978-981-15-4814-7_3.

Kumar, Vikas, Björn M. Hallström, and Axel Janke. 2013. “Coalescent-Based Genome Analyses Resolve the Early Branches of the Euarchontoglires.” *PLoS ONE*. https://doi.org/10.1371/journal.pone.0060019.

Kurzawa-Goertz, S. E., C. L. Perreault-Micale, K. M. Trybus, A. G. Szent-Györgyi, and M. A. Geeves. 1998. “Loop I Can Modulate ADP Affinity, ATPase Activity, and Motility of Different Scallop Myosins. Transient Kinetic Analysis of S1 Isoforms.” *Biochemistry*. https://doi.org/10.1021/bi972844+.

Kwartler, Callie S., Jiyuan Chen, Dhananjay Thakur, Shumin Li, Kedryn Baskin, Shanzhi Wang, Zhao V. Wang, et al. 2014. “Overexpression of Smooth Muscle Myosin Heavy Chain Leads to Activation of the Unfolded Protein Response and Autophagic Turnover of Thick Filament-Associated Proteins in Vascular Smooth Muscle Cells.” *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M113.499277.

Lam, Tommy Tsan Yuk, Na Jia, Ya Wei Zhang, Marcus Ho Hin Shum, Jia Fu Jiang, Hua Chen Zhu, Yi Gang Tong, et al. 2020. “Identifying SARS-CoV-2-Related Coronaviruses in Malayan Pangolins.” *Nature*. https://doi.org/10.1038/s41586-020-2169-0.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2018. “ClinVar: Improving Access to Variant Interpretations and Supporting Evidence.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx1153.

Larramendi, Asier. 2016. “Shoulder Height, Body Mass, and Shape of Proboscideans.” *Acta Palaeontologica Polonica*. https://doi.org/10.4202/app.00136.2014.

Latinne, Alice, Ben Hu, Kevin J. Olival, Guangjian Zhu, Libiao Zhang, Hongying Li, Aleksei A. Chmura, et al. 2020. “Origin and Cross-Species Transmission of Bat Coronaviruses in China.” *Nature Communications*. https://doi.org/10.1038/s41467-020-17687-3.

Lednicky, John A., Michael Lauzard, Z. Hugh Fan, Antarpreet Jutla, Trevor B. Tilly, Mayank Gangwar, Moiz Usmani, et al. 2020. “Viable SARS-CoV-2 in the Air of a Hospital Room

with COVID-19 Patients." *International Journal of Infectious Diseases*. https://doi.org/10.1016/j.ijid.2020.09.025.

Lee, B., and F. M. Richards. 1971. "The Interpretation of Protein Structures: Estimation of Static Accessibility." *Journal of Molecular Biology*. https://doi.org/10.1016/0022-2836(71)90324-X.

Lee, David A., Robert Rentzsch, and Christine Oreng. 2009. "GeMMA: Functional Subfamily Classification within Superfamilies of Predicted Protein Structural Domains." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkp1049.

Leinonen, Rasko, Hideaki Sugawara, and Martin Shumway. 2011. "The Sequence Read Archive." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq1019.

Levinthal, Cyrus. 1969. "How to Fold Graciously." *Mössbauer Spectroscopy in Biological Systems Proceedings*.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018. "Earth BioGenome Project: Sequencing Life for the Future of Life." *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1720115115.

Li, Jingjing, Weipeng Quan, Shuge Yan, Shuangju Wu, Jianhu Qin, Tingting Yang, Fan Liang, Depeng Wang, and Yu Liang. 2020. "Rapid Detection of SARS-CoV-2 and Other Respiratory Viruses by Using LAMP Method with Nanopore Flongle Workflow." *BioRxiv*. https://doi.org/10.1101/2020.06.03.131474.

Li, Qun, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, et al. 2020. "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia." *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa2001316.

Li, Wenhui, Michael J. Moore, Natalya Vasllieva, Jianhua Sui, Swee Kee Wong, Michael A. Berne, Mohan Somasundaran, et al. 2003. "Angiotensin-Converting Enzyme 2 Is a Functional Receptor for the SARS Coronavirus." *Nature*. https://doi.org/10.1038/nature02145.

Lindstedt, Stan L. 1987. *Allometry: Body Size Constraints in Animal Design*. *Drinking Water and Health, Volume 8: Pharmacokinetics in Risk Assessment*. https://doi.org/10.17226/1015.

Lodish, Harvey, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. 2000. "Myosin: The Actin Motor Protein." https://www.ncbi.nlm.nih.gov/books/NBK21724/.

Lokugamage, Kumari G., Craig Schindewolf, and Vineet D. Menachery. 2020. "SARS-CoV-2 Sensitive to Type I Interferon Pretreatment." *BioRxiv*.

López, Gonzalo, Lakes Ezkurdia, and Michael L. Tress. 2009. "Assessment of Ligand Binding Residue Predictions in CASP8." *Proteins: Structure, Function and Bioinformatics*. https://doi.org/10.1002/prot.22557.

Lopez, Gonzalo, Paolo Maietta, Jose Manuel Rodriguez, Alfonso Valencia, and Michael L.

Tress. 2011. “Firestar - Advances in the Prediction of Functionally Important Residues.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr437.

Lu, Hengyun, Francesca Giordano, and Zemin Ning. 2016. “Oxford Nanopore MinION Sequencing and Genome Assembly.” *Genomics, Proteomics and Bioinformatics*. https://doi.org/10.1016/j.gpb.2016.05.004.

Luan, Junwen, Yue Lu, Xiaolu Jin, and Leiliang Zhang. 2020. “Spike Protein Recognition of Mammalian ACE2 Predicts the Host Range and an Optimized ACE2 for SARS-CoV-2 Infection.” *Biochemical and Biophysical Research Communications*. https://doi.org/10.1016/j.bbrc.2020.03.047.

Lupyan, Dmitry, Alejandra Leo-Macias, and Angel R. Ortiz. 2005. “A New Progressive-Iterative Algorithm for Multiple Structure Alignment.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bti527.

Ma, Jianxin, Xiao Qi, Haoxuan Chen, Xinyue Li, Zheng Zhang, Haibin Wang, Lingli Sun, et al. 2020. “COVID-19 Patients in Earlier Stages Exhaled Millions of SARS-CoV-2 per Hour.” *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*. https://doi.org/10.1093/cid/ciaa1283.

Mahdavi, V, A P Chambers, and B Nadal-Ginard. 1984. “Cardiac Alpha- and Beta-Myosin Heavy Chain Genes Are Organized in Tandem.” *Proc Natl Acad Sci U S A*.

Maietta, Paolo, Gonzalo Lopez, Angel Carro, Benjamin J. Pingilley, Leticia G. Leon, Alfonso Valencia, and Michael L. Tress. 2014. “FireDB: A Compendium of Biological and Pharmacologically Relevant Ligands.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt1127.

Margossian, SS. Lowey, S. 1982. “Preparation of Myosin and Its Subfragments from Rabbit Skeletal Muscle.” *Methods in Enzymology* 85: 55–71.

Mariani, Valerio, Florian Kiefer, Tobias Schmidt, Juergen Haas, and Torsten Schwede. 2011. “Assessment of Template Based Protein Structure Predictions in CASP9.” *Proteins* 79 Suppl 10: 37–58. https://doi.org/10.1002/prot.23177.

Marín-García, José, Michael J. Goldenthal, and Gordon W. Moe. 2007. *Post-Genomic Cardiology*. *Post-Genomic Cardiology*. https://doi.org/10.1016/B978-0-12-373698-7.X5000-1.

Martell, Henry J., Stuart G. Masterson, Jake E. McGreig, Martin Michaelis, and Mark NWass. 2019. “Is the Bombali Virus Pathogenic in Humans?” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz267.

Martell, Henry J., Kathie A. Wong, Juan F. Martin, Ziyan Kassam, Kay Thomas, and Mark N. Wass. 2017. “Associating Mutations Causing Cystinuria with Disease Severity with the Aim of Providing Precision Medicine.” *BMC Genomics*. https://doi.org/10.1186/s12864-017-3913-1.

Marx, V. 2013. “A Star Is Born: The Updated Human Reference Genome.” *Nature Methods*, 2013.

Matthews, B. W. 1975. “Comparison of the Predicted and Observed Secondary Structure

of T4 Phage Lysozyme." *BBA - Protein Structure*. https://doi.org/10.1016/0005-2795(75)90109-9.

Mcguffin, Liam J., Recep Adiyaman, Ali H.A. Maghrabi, Ahmad N. Shuid, Danielle A. Brackenridge, John O. Nealon, and Limcy S. Philomina. 2019. "IntFOLD: An Integrated Web Resource for High Performance Protein Structure and Function Prediction." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz322.

Meyerowitz, Eric A., Aaron Richterman, Rajesh T. Gandhi, and Paul E. Sax. 2021. "Transmission of SARS-CoV-2: A Review of Viral, Host, and Environmental Factors." *Annals of Internal Medicine*. https://doi.org/10.7326/M20-5008.

Mi, Huaiyu, Dustin Ebert, Anushya Muruganujan, Caitlin Mills, Laurent Philippe Albou, Tremayne Mushayamaha, and Paul D. Thomas. 2021. "PANTHER Version 16: A Revised Family Classification, Tree-Based Classification Tool, Enhancer Regions and Extensive API." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa1106.

Miller, Becky, Marieke J Bloemink, Michael Geeves, Mik;os Nyitrai, and Sandford Bernstein. 2007. "A Variable Domain Near the ATP Binding Site in Drosophila Muscle Myosin Is Part of the Communication Pathway between the Nucleotide and Actin-Binding Sites." *Journal of Molecular Biology* 368 (4): 1051–66.

Miller, Becky M., Marieke J. Bloemink, Miklós Nyitrai, Sanford I. Bernstein, and Michael A. Geeves. 2007. "A Variable Domain near the ATP-Binding Site in Drosophila Muscle Myosin Is Part of the Communication Pathway between the Nucleotide and Actin-Binding Sites." *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2007.02.042.

Miller, Mark A., Wayne Pfeiffer, and Terri Schwartz. 2010. "Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees." In *2010 Gateway Computing Environments Workshop, GCE 2010*. https://doi.org/10.1109/GCE.2010.5676129.

Mirdita, Milot, Lars Von Den Driesch, Clovis Galiez, Maria J. Martin, Johannes Soding, and Martin Steinegger. 2017. "Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1081.

Mosmann, Tim. 1983. "Rapid Colorimetric Assay for Cellular Growth and Survival: Application to Proliferation and Cytotoxicity Assays." *Journal of Immunological Methods*. https://doi.org/10.1016/0022-1759(83)90303-4.

Moult, John, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. 2018. "Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XII." *Proteins: Structure, Function and Bioinformatics* 86 (March): 7–15. https://doi.org/10.1002/prot.25415.

Mukhopadhyay, Abhik, Neera Borkakoti, Lukáš Pravda, Jonathan D. Tyzack, Janet M. Thornton, and Sameer Velankar. 2019. "Finding Enzyme Cofactors in Protein Data Bank." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz115.

Murphy, Coleen T., and James A. Spudich. 1998. "Dictyostelium Myosin 25-50K Loop Substitutions Specifically Affect ADP Release Rates." *Biochemistry*.

https://doi.org/10.1021/bi972903j.

Murzin, Alexey G., Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. 1995. “SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures.” *Journal of Molecular Biology*. https://doi.org/10.1016/S0022-2836(05)80134-2.

Muth, Thilo, Juan A. García-martín, Antonio Rausell, David Juan, Alfonso Valencia, and Florencio Pazos. 2012. “JDet: Interactive Calculation and Visualization of Function-Related Conservation Patterns in Multiple Sequence Alignments and Structures.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btr688.

Nakano, Felipe Kenji, Mathias Lietaert, and Celine Vens. 2019. “Machine Learning for Discovering Missing or Wrong Protein Function Annotations.” *BMC Bioinformatics*. https://doi.org/10.1186/s12859-019-3060-6.

Nakao, Koichi, Wayne Minobe, Robert Roden, Michael R. Bristow, and Leslie A. Leinwand. 1997. “Myosin Heavy Chain Gene Expression in Human Heart Failure.” *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI119776.

NCBI Resource Coordinators. 2017. “Database Resources of the National Center for Biotechnology Information.” *Nucleic Acids Research* 45 (D1): D12–17. https://doi.org/10.1093/nar/gkw1071.

Newell-Litwa, Karen A., Rick Horwitz, and Marcelo L. Lamers. 2015. “Non-Muscle Myosin II in Disease: Mechanisms and Therapeutic Opportunities.” *DMM Disease Models and Mechanisms*. https://doi.org/10.1242/dmm.022103.

Nieto-Torres, Jose L., Marta L. DeDiego, Carmina Verdiá-Báguena, Jose M. Jimenez-Guardeño, Jose A. Regla-Nava, Raul Fernandez-Delgado, Carlos Castaño-Rodriguez, et al. 2014. “Severe Acute Respiratory Syndrome Coronavirus Envelope Protein Ion Channel Activity Promotes Virus Fitness and Pathogenesis.” *PLoS Pathogens*. https://doi.org/10.1371/journal.ppat.1004077.

Nikaido, Masato, Hidenori Nishihara, Yukio Hukumoto, and Norihiro Okada. 2003. “Ancient SINEs from African Endemic Mammals.” *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msg052.

Nishiyama, Tomoaki, Hidetoshi Sakayama, Jan de Vries, Henrik Buschmann, Denis Saint-Marcoux, Kristian K. Ullrich, Fabian B. Haas, et al. 2018. “The Chara Genome: Secondary Complexity and Implications for Plant Terrestrialization.” *Cell*. https://doi.org/10.1016/j.cell.2018.06.033.

Nye, Tom M W, Pietro Liò, and Walter R. Gilks. 2006. “A Novel Algorithm and Web-Based Tool for Comparing Two Alternative Phylogenetic Trees.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bti720.

Nyitrai, M, and MA Geeves. 2004. “Adenosine Diphosphate and Strain Sensitivity in Myosin Motors.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 359 (1452): 1867–77.

O’Grady, Gina L., Heather A. Best, Tamar E. Sztal, Vanessa Schartner, Myriam Sanjuan-Vazquez, Sandra Donkervoort, Osorio Abath Neto, et al. 2016. “Variants in the

Oxidoreductase PYROXD1 Cause Early-Onset Myopathy with Internalized Nuclei and Myofibrillar Disorganization." *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2016.09.005.

O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. "Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1189.

O'Toole, Áine, Emily Scher, Anthony Underwood, Ben Jackson, Verity Hill, John T McCrone, Rachel Colquhoun, et al. 2021. "Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool." *Virus Evolution*, July. https://doi.org/10.1093/ve/veab064.

Ojima, Koichi. 2019. "Myosin: Formation and Maintenance of Thick Filaments." *Animal Science Journal*. https://doi.org/10.1111/asj.13226.

Olsen, Sonja J., Hsiao-Ling Chang, Terence Yung-Yan Cheung, Antony Fai-Yu Tang, Tamara L. Fisk, Steven Peng-Lim Ooi, Hung-Wei Kuo, et al. 2003. "Transmission of the Severe Acute Respiratory Syndrome on Aircraft." *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa031349.

Onafuye, Hannah, Sebastian Pieper, Dennis Mulac, Jindrich Cinatl, Mark N. Wass, Klaus Langer, and Martin Michaelis. 2019. "Doxorubicin-Loaded Human Serum Albumin Nanoparticles Overcome Transporter-Mediated Drug Resistance in Drug-Adapted Cancer Cells." *Beilstein Journal of Nanotechnology*. https://doi.org/10.3762/bjnano.10.166.

OpenAI. 2020. "AlphaFold : A Solution to a 50-Year-Old Grand Challenge in Biology." *Https://Deepmind.Com/Blog*.

Ortiz, Angel R., Charlie E.M. Strauss, and Osvaldo Olmea. 2009. "MAMMOTH (Matching Molecular Models Obtained from Theory): An Automated Method for Model Comparison." *Protein Science*. https://doi.org/10.1110/ps.0215902.

Oxford Nanopore Technologies. 2020. "New Research Algorithms Yield Accuracy Gains for Nanopore Sequencing." Oxford Nanopore. Community News. 2020. https://nanoporetech.com/about-us/news/new-research-algorithms-yield-accuracy-gains-nanopore-sequencing.

Page, Michael Le. 2021. "What Are the New Coronavirus Variants?" *New Scientist*. https://doi.org/10.1016/s0262-4079(21)00081-6.

Pappalardo, Morena, Miguel Julia, Mark J. Howard, Jeremy S. Rossman, Martin Michaelis, and Mark N. Wass. 2016. "Conserved Differences in Protein Sequence Determine the Human Pathogenicity of Ebolaviruses." *Scientific Reports*. https://doi.org/10.1038/srep23743.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. "APE: Analyses of Phylogenetics and Evolution in R Language." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btg412.

Parker, F, and M Peckham. 2020. "Disease Mutations in Striated Muscle Myosins."

*Biophysical Reviews* 12 (4): 887–94. https://doi.org/10.1007/S12551-020-00721-5.

Parker, Francine, and Michelle Peckham. 2020. “Disease Mutations in Striated Muscle Myosins.” *Biophysical Reviews*. https://doi.org/10.1007/s12551-020-00721-5.

Pellegrino, M. A., M. Canepari, R. Rossi, G. D’Antona, C. Reggiani, and R. Bottinelli. 2003. “Orthologous Myosin Isoform and Scaling of Shortening Velocity with Body Size in Mouse, Rat, Rabbit and Human Muscles.” *Journal of Physiology*. https://doi.org/10.1113/jphysiol.2002.027375.

Petersen, Eskild, Marion Koopmans, Unyeong Go, Davidson H. Hamer, Nicola Petrosillo, Francesco Castelli, Merete Storgaard, Sulien Al Khalili, and Lone Simonsen. 2020. “Comparing SARS-CoV-2 with SARS-CoV and Influenza Pandemics.” *The Lancet Infectious Diseases*. Lancet Publishing Group. https://doi.org/10.1016/S1473-3099(20)30484-9.

Pickett, Brett E., Eva L. Sadat, Yun Zhang, Jyothi M. Noronha, R. Burke Squires, Victoria Hunt, Mengya Liu, et al. 2012. “ViPR: An Open Bioinformatics Database and Analysis Resource for Virology Research.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr859.

Planas, Delphine, David Veyer, Artem Baidaliuk, Isabelle Staropoli, Florence Guivel-Benhassine, Maaran Michael Rajah, Cyril Planchais, et al. 2021. “Reduced Sensitivity of SARS-CoV-2 Variant Delta to Antibody Neutralization.” *Nature*. https://doi.org/10.1038/s41586-021-03777-9.

Pruitt, Kim D., Garth R. Brown, Susan M. Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M. Farrell, et al. 2014. “RefSeq: An Update on Mammalian Reference Sequences.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt1114.

Public Health England. 2020. “Coronavirus (COVID-19): Guidance.” Public Health England. 2020.

———. 2021. “Variants: Distribution of Case Data, 11 June 2021 - GOV.UK.” 2021. https://www.gov.uk/government/publications/covid-19-variants-genomically-confirmed-case-numbers/variants-distribution-of-case-data-11-june-2021.

Qiu, Ye, Yuan Bo Zhao, Qiong Wang, Jin Yan Li, Zhi Jian Zhou, Ce Heng Liao, and Xing Yi Ge. 2020. “Predicting the Angiotensin Converting Enzyme 2 (ACE2) Utilizing Capability as the Receptor of SARS-CoV-2.” *Microbes and Infection*. https://doi.org/10.1016/j.micinf.2020.03.003.

Quader, S., K. Isvaran, R. E. Hale, B. G. Miner, and N. E. Seavy. 2004. “Nonlinear Relationships and Phylogenetically Independent Contrasts.” *Journal of Evolutionary Biology*. https://doi.org/10.1111/j.1420-9101.2004.00697.x.

Quick, Joshua, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, et al. 2016. “Real-Time, Portable Genome Sequencing for Ebola Surveillance.” *Nature*. https://doi.org/10.1038/nature16996.

R Core team. 2018. “R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL Https://Www.R-

Project.Org/.” R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing , Vienna, Austria. ISBN 3-900051-07-0, URL Http://Www.R-Project.Org/. 2018. https://doi.org/10.2788/95827.

Raman, Srivatsan, Robert Vernon, James Thompson, Michael Tyka, Ruslan Sadreyev, Jimin Pei, David Kim, et al. 2009. “Structure Prediction for CASP8 with All-Atom Refinement Using Rosetta.” *Proteins: Structure, Function and Bioinformatics*. https://doi.org/10.1002/prot.22540.

Rambaut, Andrew. 2014. “FigTree.” FigTree. 2014. https://doi.org/10.1046/j.1471-8286.2003.00511.x.

Rausell, Antonio, David Juan, Florencio Pazos, and Alfonso Valencia. 2010. “Protein Interactions and Ligand Binding: From Protein Subfamilies to Functional Specificity.” *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0908044107.

Rayment, Ivan, Wojciech R. Rypniewski, Karen Schmidt-Bäse, Robert Smith, Diana R. Tomchick, Matthew M. Benning, Donald A. Winkelmann, Gary Wesenberg, and Hazel M. Holden. 1993. “Three-Dimensional Structure of Myosin Subfragment-1: A Molecular Motor.” *Science*. https://doi.org/10.1126/science.8316857.

Remmert, Michael, Andreas Biegert, Andreas Hauser, and Johannes Söding. 2012. “HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment.” *Nature Methods*. https://doi.org/10.1038/nmeth.1818.

Resnicow, Daniel I., John C. Deacon, Hans M. Warrick, James A. Spudich, and Leslie A. Leinwand. 2010. “Functional Diversity among a Family of Human Skeletal Muscle Myosin Motors.” *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0913527107.

Reusken, Chantal E.M., Bart L. Haagmans, and Marion P.G. Koopmans. 2014. “[Dromedary Camels and Middle East Respiratory Syndrome: MERS Coronavirus in the ’Ship of the Desert’].” *Nederlands Tijdschrift Voor Geneeskunde*.

Revell, Liam J. 2012. “Phytools: An R Package for Phylogenetic Comparative Biology (and Other Things).” *Methods in Ecology and Evolution*. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Ribeiro, António J.M., Gemma L. Holliday, Nicholas Furnham, Jonathan D. Tyzack, Katherine Ferris, and Janet M. Thornton. 2018. “Mechanism and Catalytic Site Atlas (M-CSA): A Database of Enzyme Reaction Mechanisms and Active Sites.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx1012.

Rice, Peter, Lan Longden, and Alan Bleasby. 2000. “EMBOSS: The European Molecular Biology Open Software Suite.” *Trends in Genetics*. https://doi.org/10.1016/S0168-9525(00)02024-2.

Ridder, Dick De, Jeroen De Ridder, and Marcel J.T. Reinders. 2013. “Pattern Recognition in Bioinformatics.” *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbt020.

Rivett, Lucy, Sushmita Sridhar, Dominic Sparkes, Matthew Routledge, Nick K. Jones, Sally Forrest, Jamie Young, et al. 2020. “Screening of Healthcare Workers for SARS-CoV-2

Highlights the Role of Asymptomatic Carriage in COVID-19 Transmission." *ELife*. https://doi.org/10.7554/eLife.58728.

Robert-Paganin, Julien, Daniel Auguin, and Anne Houdusse. 2018. "Hypertrophic Cardiomyopathy Disease Results from Disparate Impairments of Cardiac Myosin Function and Auto-Inhibition." *Nature Communications*. https://doi.org/10.1038/s41467-018-06191-4.

Roche, Daniel B., Maria T. Buenavista, and Liam J. McGuffin. 2013. "The FunFOLD2 Server for the Prediction of Protein-Ligand Interactions." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt498.

Rodriguez, Jose Manuel, Juan Rodriguez-Rivas, Tomás Di Domenico, Jesús Vázquez, Alfonso Valencia, and Michael L Tress. 2018. "APPRIS 2017: Principal Isoforms for Multiple Gene Sets." *Nucleic Acids Research* 46 (D1): D213–17. https://doi.org/10.1093/nar/gkx997.

Roy, Ambrish, Alper Kucukural, and Yang Zhang. 2010. "I-TASSER: A Unified Platform for Automated Protein Structure and Function Prediction." *Nature Protocols* 5 (4): 725–38. https://doi.org/10.1038/nprot.2010.5.

Ruan, Yi Jun, Chia Lin Wei, Ling Ai Ee, Vinsensius B. Vega, Herve Thoreau, Se Thoe Su Yun, Jer Ming Chia, et al. 2003. "Comparative Full-Length Genome Sequence Analysis of 14 SARS Coronavirus Isolates and Common Mutations Associated with Putative Origins of Infection." *Lancet* 361 (9371): 1779–85. https://doi.org/10.1016/S0140-6736(03)13414-9.

Salk, Jesse J., Michael W. Schmitt, and Lawrence A. Loeb. 2018. "Enhancing the Accuracy of Next-Generation Sequencing for Detecting Rare and Subclonal Mutations." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2017.117.

Sanger, F, S Nicklen, and A.R Coulson. 1977. "DNA Sequencing with Chain-Terminating." *Proc Natl Acad Sci USA*.

Santana, Charles A., Sabrina A. de Silveira, João P.A. Moraes, Sandro C. Izidoro, Raquel C. de Melo-Minardi, António J.M. Ribeiro, Jonathan D. Tyzack, Neera Borkakoti, and Janet M. Thornton. 2020. "GRaSP: A Graph-Based Residue Neighborhood Strategy to Predict Binding Sites." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa805.

Savage, Van M., Andrew P. Allen, James H. Brown, James F. Gillooly, Alexander B. Herman, William H. Woodruff, and Geoffrey B. West. 2007. "Scaling of Number, Size, and Metabolic Rate of Cells with Body Size in Mammals." *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0611235104.

Sayers, Eric W., Mark Cavanaugh, Karen Clark, James Ostell, Kim D. Pruitt, and Ilene Karsch-Mizrachi. 2020. "GenBank." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz956.

Schiaffino, Stefano, and Carlo Reggiani. 2011. "Fiber Types in Mammalian Skeletal Muscles." *Physiological Reviews*. https://doi.org/10.1152/physrev.00031.2010.

Schmid, S, and T Hugel. 2020. “Controlling Protein Function by Fine-Tuning Conformational Flexibility.” *ELife* 9 (e57180). https://doi.org/10.7554/ELIFE.57180.

Schmidt, Tobias, Jürgen Haas, Tiziano Gallo Cassarino, and Torsten Schwede. 2011. “Assessment of Ligand-Binding Residue Predictions in CASP9.” *Proteins: Structure, Function and Bioinformatics*. https://doi.org/10.1002/prot.23174.

Schneider, Constanze, Thomas Oellerich, Hanna Mari Baldauf, Sarah Marie Schwarz, Dominique Thomas, Robert Flick, Hanibal Bohnenberger, et al. 2017. “SAMHD1 Is a Biomarker for Cytarabine Response and a Therapeutic Target in Acute Myeloid Leukemia.” *Nature Medicine*. https://doi.org/10.1038/nm.4255.

Schnoes, Alexandra M., Shoshana D. Brown, Igor Dodevski, and Patricia C. Babbitt. 2009. “Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies.” *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1000605.

Sciote, James J., Terence J. Morris, Michael J. Horton, Carla A. Brandon, and Clark Rosen. 2002. “Unloaded Shortening Velocity and Myosin Heavy Chain Variations in Human Laryngeal Muscle Fibers.” *Annals of Otology, Rhinology and Laryngology*. https://doi.org/10.1177/000348940211100203.

Sehnal, D, A S Rose, J Koča, S K Burley, and S Velankar. 2018. “Mol*: Towards a Common Library and Tools for Web Molecular Graphics.” In *Workshop on Molecular Graphics and Visual Analysis of Molecular Data*.

Seiffert, Erik R. 2007. “A New Estimate of Afrotherian Phylogeny Based on Simultaneous Analysis of Genomic, Morphological, and Fossil Evidence.” *BMC Evolutionary Biology*. https://doi.org/10.1186/1471-2148-7-224.

Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, et al. 2020. “AlphaFold.” *Nature*.

Shaffer, Justin F., and Samantha P. Harris. 2009. “Species-Specific Differences in the Pro-Ala Rich Region of Cardiac Myosin Binding Protein-C.” *Journal of Muscle Research and Cell Motility*. https://doi.org/10.1007/s10974-010-9207-8.

Shu, Shi, Randall J. Lee, Janine M. LeBlanc-Straceski, and Taro Q.P. Uyeda. 1999. “Role of Myosin II Tail Sequences in Its Function and Localization at the Cleavage Furrow in Dictyostelium.” *Journal of Cell Science*. https://doi.org/10.1242/jcs.112.13.2195.

Siemankowski, R F, M O Wiseman, and H White. 1985. “ADP Dissociation from Actomyosin Subfragment 1 Is Sufficiently Slow to Limit the Unloaded Shortening Velocity in Vertebrate Muscle.” *Proceedings of the National Acadamy of Science of the United States of America* 82 (February): 658–62.

Sievers, Fabian, and Desmond G. Higgins. 2014a. “Clustal Omega.” *Current Protocols in Bioinformatics*. https://doi.org/10.1002/0471250953.bi0313s48.

———. 2014b. “Clustal Omega.” *Current Protocols in Bioinformatics*. https://doi.org/10.1002/0471250953.bi0313s48.

Sillitoe, Ian, Alison L. Cuff, Benoit H. Dessailly, Natalie L. Dawson, Nicholas Furnham, David

Lee, Jonathan G. Lees, et al. 2013. “New Functional Families (FunFams) in CATH to Improve the Mapping of Conserved Functional Sites to 3D Structures.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks1211.

Simmons, Graham, Pawel Zmora, Stefanie Gierer, Adeline Heurich, and Stefan Pöhlmann. 2013. “Proteolytic Activation of the SARS-Coronavirus Spike Protein: Cutting Enzymes at the Cutting Edge of Antiviral Research.” *Antiviral Research*. https://doi.org/10.1016/j.antiviral.2013.09.028.

Skolnick, Jeffrey, and Michal Brylinski. 2009. “FINDSITE: A Combined Evolution/Structure-Based Approach to Protein Function Prediction.” *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbp017.

Skolnick, Jeffrey, Mu Gao, Hongyi Zhou, and Suresh Singh. 2021. “AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function.” *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/acs.jcim.1c01114.

Sloutsky, Roman, and Kristen M. Naegle. 2016. “High-Resolution Identification of Specificity Determining Positions in the LacI Protein Family Using Ensembles of Sub-Sampled Alignments.” *PLoS ONE*. https://doi.org/10.1371/journal.pone.0162579.

Söding, Johannes. 2005. “Protein Homology Detection by HMM-HMM Comparison.” *Bioinformatics (Oxford, England)* 21 (7): 951–60. https://doi.org/10.1093/bioinformatics/bti125.

Sommer, Christoph, and Daniel W. Gerlich. 2013. “Machine Learning in Cell Biology-Teaching Computers to Recognize Phenotypes.” *Journal of Cell Science*. https://doi.org/10.1242/jcs.123604.

Somody, Joseph C., Stephen S. MacKinnon, and Andreas Windemuth. 2017. “Structural Coverage of the Proteome for Pharmaceutical Applications.” *Drug Discovery Today*. https://doi.org/10.1016/j.drudis.2017.08.004.

Song, Wenfei, Miao Gui, Xinquan Wang, and Ye Xiang. 2018. “Cryo-EM Structure of the SARS Coronavirus Spike Glycoprotein in Complex with Its Host Cell Receptor ACE2.” *PLoS Pathogens*. https://doi.org/10.1371/journal.ppat.1007236.

Song, Xuhao, Tingbang Yang, Xinyi Zhang, Ying Yuan, Xianghui Yan, Yi Wei, Jun Zhang, and Caiquan Zhou. 2021. “Comparison of the Microsatellite Distribution Patterns in the Genomes of Euarchontoglires at the Taxonomic Level.” *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2021.622724.

Springer, Mark S., Robert W. Meredith, Jan E. Janecka, and William J. Murphy. 2011. “The Historical Biogeography of Mammalia.” *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2011.0023.

Spudich, J A, and S Watt. 1971. “The Regulation of Rabbit Skeletal Muscle Contraction. I. Biochemical Studies of the Interaction of the Tropomyosin-Troponin Complex with Actin and the Proteolytic Fragments of Myosin.” *The Journal of Biological Chemistry* 246 (15): 4866–71.

Srikakulam, Rajani, and Donald A. Winkelmann. 2004. “Chaperone-Mediated Folding and

Assembly of Myosin in Striated Muscle." *Journal of Cell Science*. https://doi.org/10.1242/jcs.00899.

Stamatakis, Alexandros. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." *Bioinformatics* 30 (9): 1312–13. https://doi.org/10.1093/bioinformatics/btu033.

Stanley, Steven M. 1973. "AN EXPLANATION FOR COPE'S RULE." *Evolution*. https://doi.org/10.1111/j.1558-5646.1973.tb05912.x.

Steinegger, Martin, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J. Haunsberger, and Johannes Söding. 2019. "HH-Suite3 for Fast Remote Homology Detection and Deep Protein Annotation." *BMC Bioinformatics*. https://doi.org/10.1186/s12859-019-3019-7.

Steinegger, Martin, and Johannes Söding. 2017. "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets." *Nature Biotechnology*. https://doi.org/10.1038/nbt.3988.

Sweeney, H. Lee, Steven S. Rosenfeld, Fred Brown, Lynn Faust, Joe Smith, Jun Xing, Leonard A. Stein, and James R. Sellers. 1998. "Kinetic Tuning of Myosin via a Flexible Loop Adjacent to the Nucleotide Binding Pocket." *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.273.11.6262.

Tang, Tiffany, Miya Bidon, Javier A. Jaimes, Gary R. Whittaker, and Susan Daniel. 2020. "Coronavirus Membrane Fusion Mechanism Offers a Potential Target for Antiviral Development." *Antiviral Research*. Elsevier B.V. https://doi.org/10.1016/j.antiviral.2020.104792.

Thomas, Kay, Kathie Wong, John Withington, Matthew Bultitude, and Angela Doherty. 2014. "Cystinuria - A Urologist's Perspective." *Nature Reviews Urology*. https://doi.org/10.1038/nrurol.2014.51.

Thompson, Reid F., and George M. Langford. 2002. "Myosin Superfamily Evolutionary History." *Anatomical Record*. https://doi.org/10.1002/ar.10160.

Tibbetts, Paul. 2017. " Evolution of Nervous Systems. Second Edition. Volume 1: The Evolution of the Nervous Systems in Nonmammalian Vertebrates . Editor-in-Chief: Jon H. Kaas; Volume Editor: Georg Striedter. Academic Press. Amsterdam (The Netherlands) and Boston (Massachusetts." *The Quarterly Review of Biology*. https://doi.org/10.1086/693610.

Tjelvar S.G. Olsson, Antony M.E. Churchill, William R. Pitt, John E. Ladbury & Mark A. Williams. 2012. "ProACT2 – Analysis of Solvent Accessibility, Cavities and Contacts in Proteins and Their Complexes."

Toniolo, L., L. Maccatrozzo, M. Patruno, F. Caliaro, F. Mascarello, and C. Reggiani. 2005. "Expression of Eight Distinct MHC Isoforms in Bovine Striated Muscles: Evidence for MHC-2B Presence Only in Extraocular Muscles." *Journal of Experimental Biology*. https://doi.org/10.1242/jeb.01904.

Toniolo, Luana, Marco Patruno, Lisa Maccatrozzo, Maria A. Pellegrino, Monica Canepari, Rosetta Rossi, Giuseppe D'Antona, Roberto Bottinelli, Carlo Reggiani, and Francesco

Mascarello. 2004. “Fast Fibres in a Large Animal: Fibre Types, Contractile Properties and Myosin Expression in Pig Skeletal Muscles.” *Journal of Experimental Biology*. https://doi.org/10.1242/jeb.00950.

Torrance, James W., Malcolm W. MacArthur, and Janet M. Thornton. 2008. “Evolution of Binding Sites for Zinc and Calcium Ions Playing Structural Roles.” *Proteins: Structure, Function and Genetics*. https://doi.org/10.1002/prot.21741.

Totura, Allison L., and Ralph S. Baric. 2012. “SARS Coronavirus Pathogenesis: Host Innate Immune Responses and Viral Antagonism of Interferon.” *Current Opinion in Virology*. https://doi.org/10.1016/j.coviro.2012.04.004.

Tress, Michael L., David Jones, and Alfonso Valencia. 2003. “Predicting Reliable Regions in Protein Alignments from Sequence Profiles.” *Journal of Molecular Biology*. https://doi.org/10.1016/S0022-2836(03)00622-3.

Trott, O, and A J Olson. 2010. “AutoDock Vina.” *J. Comput. Chem.* https://doi.org/10.1002/jcc.21334.

Tsiavaliaris, Georgios, Setsuko Fujita-Becker, Renu Batra, Dmitrii I Levitsky, F Jon Kull, Michael A Geeves, and Dietmar J Manstein. 2002. “Mutations in the Relay Loop Region Result in Dominant-Negative Inhibition of Myosin II Function in Dictyostelium.” *EMBO Reports* 3 (11): 1099–1105. https://doi.org/10.1093/embo-reports/kvf214.

Tunyasuvunakool, Kathryn, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, et al. 2021. “Highly Accurate Protein Structure Prediction for the Human Proteome.” *Nature*, July, 1–9. https://doi.org/10.1038/s41586-021-03828-1.

Uhlén, Mathias, Linn Fagerberg, Björ M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. “Tissue-Based Map of the Human Proteome.” *Science*. https://doi.org/10.1126/science.1260419.

Ujfalusi, Zoltan, Carlos D. Vera, Srbolujub M. Mijailovich, Marina Svicevic, Elizabeth Choe Yu, Masataka Kawana, Kathleen M. Ruppel, James A. Spudich, Michael A. Geeves, and Leslie A. Leinwand. 2018. “Dilated Cardiomyopathy Myosin Mutants Have Reduced Force-Generating Capacity.” *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.RA118.001938.

Varadi, Mihaly, John Berrisford, Mandar Deshpande, Sreenath S. Nair, Aleksandras Gutmanas, David Armstrong, Lukas Pravda, et al. 2020. “PDBE-KB: A Community-Driven Resource for Structural and Functional Annotations.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz853.

Vicente-Manzanares, Miguel, Xuefei Ma, Robert S. Adelstein, and Alan Rick Horwitz. 2009. “Non-Muscle Myosin II Takes Centre Stage in Cell Adhesion and Migration.” *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm2786.

Walklate, J., C. Vera, M.J. Bloemink, M.A. Geeves, and L. Leinwand. 2016. “The Most Prevalent Freeman-Sheldon Syndrome Mutations in the Embryonic Myosin Motor Share Functional Defects.” *Journal of Biological Chemistry* 291 (19).

https://doi.org/10.1074/jbc.M115.707489.

Walls, Alexandra C., Young Jun Park, M. Alejandra Tortorici, Abigail Wall, Andrew T. McGuire, and David Veesler. 2020. “Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.” *Cell*. https://doi.org/10.1016/j.cell.2020.02.058.

Wan, Yushun, Jian Shang, Rachel Graham, Ralph S. Baric, and Fang Li. 2020. “Receptor Recognition by the Novel Coronavirus from Wuhan: An Analysis Based on Decade-Long Structural Studies of SARS Coronavirus.” *Journal of Virology*. https://doi.org/10.1128/jvi.00127-20.

Wang, L. F., and B. T. Eaton. 2007. “Bats, Civets and the Emergence of SARS.” *Current Topics in Microbiology and Immunology*. https://doi.org/10.1007/978-3-540-70962-6_13.

Wang, Ming, Meiying Yan, Huifang Xu, Weili Liang, Biao Kan, Bojian Zheng, Honglin Chen, et al. 2005. “SARS-CoV Infection in a Restaurant from Palm Civet.” *Emerging Infectious Diseases*. https://doi.org/10.3201/eid1112.041293.

Wang, Qun, Carole L. Moncman, and Donald A. Winkelmann. 2003. “Mutations in the Motor Domain Modulate Myosin Activity and Myofibril Organization.” *Journal of Cell Science*. https://doi.org/10.1242/jcs.00709.

Warkman, Andrew S., Samantha A. Whitman, Melanie K. Miller, Robert J. Garriock, Catherine M. Schwach, Carol C. Gregorio, and Paul A. Krieg. 2012. “Developmental Expression and Cardiac Transcriptional Regulation of Myh7b, a Third Myosin Heavy Chain in the Vertebrate Heart.” *Cytoskeleton*. https://doi.org/10.1002/cm.21029.

Wass, Mark N., Lawrence A. Kelley, and Michael J E Sternberg. 2010a. “3DLigandSite: Predicting Ligand-Binding Sites Using Similar Structures.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq406.

Wass, Mark N., and Michael J.E. Sternberg. 2009. “Prediction of Ligand Binding Sites Using Homologous Structures and Conservation at CASP8.” *Proteins: Structure, Function and Bioinformatics*. https://doi.org/10.1002/prot.22513.

Wass, Mark N, Lawrence A Kelley, and Michael J E Sternberg. 2010b. “3DLigandSite: Predicting Ligand-Binding Sites Using Similar Structures.” *Nucleic Acids Research* 38 (Web Server issue): W469-73. https://doi.org/10.1093/nar/gkq406.

Wetterstrand, KA. n.d. “DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: Www.Genome.Gov/Sequencingcostsdata. Accessed [16/02/2021].”

WHO. 2021. “WHO Announces Simple, Easy-to-Say Labels for SARS-CoV-2 Variants of Interest and Concern.” 2021. https://www.who.int/news/item/31-05-2021-who-announces-simple-easy-to-say-labels-for-sars-cov-2-variants-of-interest-and-concern.

Wit, Emmie De, Neeltje Van Doremalen, Darryl Falzarano, and Vincent J. Munster. 2016. “SARS and MERS: Recent Insights into Emerging Coronaviruses.” *Nature Reviews Microbiology*. https://doi.org/10.1038/nrmicro.2016.81.

Woodhead, John L., Fa Qing Zhao, Roger Craig, Edward H. Egelman, Lorenzo Alamo, and Raúl Padrón. 2005. “Atomic Model of a Myosin Filament in the Relaxed State.” *Nature*. https://doi.org/10.1038/nature03920.

Wrapp, Daniel, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching Lin Hsieh, Olubukola Abiona, Barney S. Graham, and Jason S. McLellan. 2020. “Cryo-EM Structure of the 2019-NCoV Spike in the Prefusion Conformation.” *Science*. https://doi.org/10.1126/science.aax0902.

Wu, Aiping, Yousong Peng, Baoying Huang, Xiao Ding, Xianyue Wang, Peihua Niu, Jing Meng, et al. 2020. “Genome Composition and Divergence of the Novel Coronavirus (2019-NCoV) Originating in China.” *Cell Host and Microbe*. https://doi.org/10.1016/j.chom.2020.02.001.

Wu, Qi, Zhenling Peng, Yang Zhang, and Jianyi Yang. 2018. “COACH-D: Improved Protein-Ligand Binding Sites Prediction with Refined Ligand-Binding Poses through Molecular Docking.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky439.

Wu, Sitao, and Yang Zhang. 2007. “LOMETS: A Local Meta-Threading-Server for Protein Structure Prediction.” *Nucleic Acids Research* 35 (10): 3375–82. https://doi.org/10.1093/nar/gkm251.

Xia, Shuai, Qiaoshuai Lan, Shan Su, Xinling Wang, Wei Xu, Zezhong Liu, Yun Zhu, Qian Wang, Lu Lu, and Shibo Jiang. 2020. “The Role of Furin Cleavage Site in SARS-CoV-2 Spike Protein-Mediated Membrane Fusion in the Presence or Absence of Trypsin.” *Signal Transduction and Targeted Therapy*. Springer Nature. https://doi.org/10.1038/s41392-020-0184-0.

Xu, Dong, and Ruth Nussinov. 1998. “Favorable Domain Size in Proteins.” *Folding and Design*. https://doi.org/10.1016/S1359-0278(98)00004-2.

Yan, Renhong, Yuanyuan Zhang, Yaning Li, Lu Xia, Yingying Guo, and Qiang Zhou. 2020. “Structural Basis for the Recognition of SARS-CoV-2 by Full-Length Human ACE2.” *Science*. https://doi.org/10.1126/science.abb2762.

Yang, Jianyi, Ambrish Roy, and Yang Zhang. 2013a. “BioLiP: A Semi-Manually Curated Database for Biologically Relevant Ligand-Protein Interactions.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks966.

———. 2013b. “Protein-Ligand Binding Site Recognition Using Complementary Binding-Specific Substructure Comparison and Sequence Profile Alignment.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btt447.

Yoshimoto, Francis K. 2020. “The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19.” *Protein Journal*. https://doi.org/10.1007/s10930-020-09901-4.

Zdobnov, E M, and R Apweiler. 2001. “InterProScan - an Integration Platform for the Signature-Recognition Methods in InterPro.” *Bioinformatics* 17 (9): 847–48. https://doi.org/10.1093/bioinformatics/17.9.847.

Zhang, Chengxin, Wei Zheng, S. M. Mortuza, Yang Li, and Yang Zhang. 2020. “DeepMSA: Constructing Deep Multiple Sequence Alignment to Improve Contact Prediction and

Fold-Recognition for Distant-Homology Proteins." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz863.

Zhang, Tao, Qunfu Wu, and Zhigang Zhang. 2020. "Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak." *Current Biology*. https://doi.org/10.1016/j.cub.2020.03.022.

Zhang, Yang, and Jeffrey Skolnick. 2005. "TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gki524.

Zhao, Jingtian, Yang Cao, and Le Zhang. 2020. "Exploring the Computational Methods for Protein-Ligand Binding Site Prediction." *Computational and Structural Biotechnology Journal*. https://doi.org/10.1016/j.csbj.2020.02.008.

Zhou, Naihui, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, et al. 2019. "The CAFA Challenge Reports Improved Protein Function Prediction and New Functional Annotations for Hundreds of Genes through Experimental Screens." *Genome Biology*. https://doi.org/10.1186/s13059-019-1835-8.

Zhou, Peng, Xing Lou Yang, Xian Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao Rui Si, et al. 2020. "A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin." *Nature*. https://doi.org/10.1038/s41586-020-2012-7.

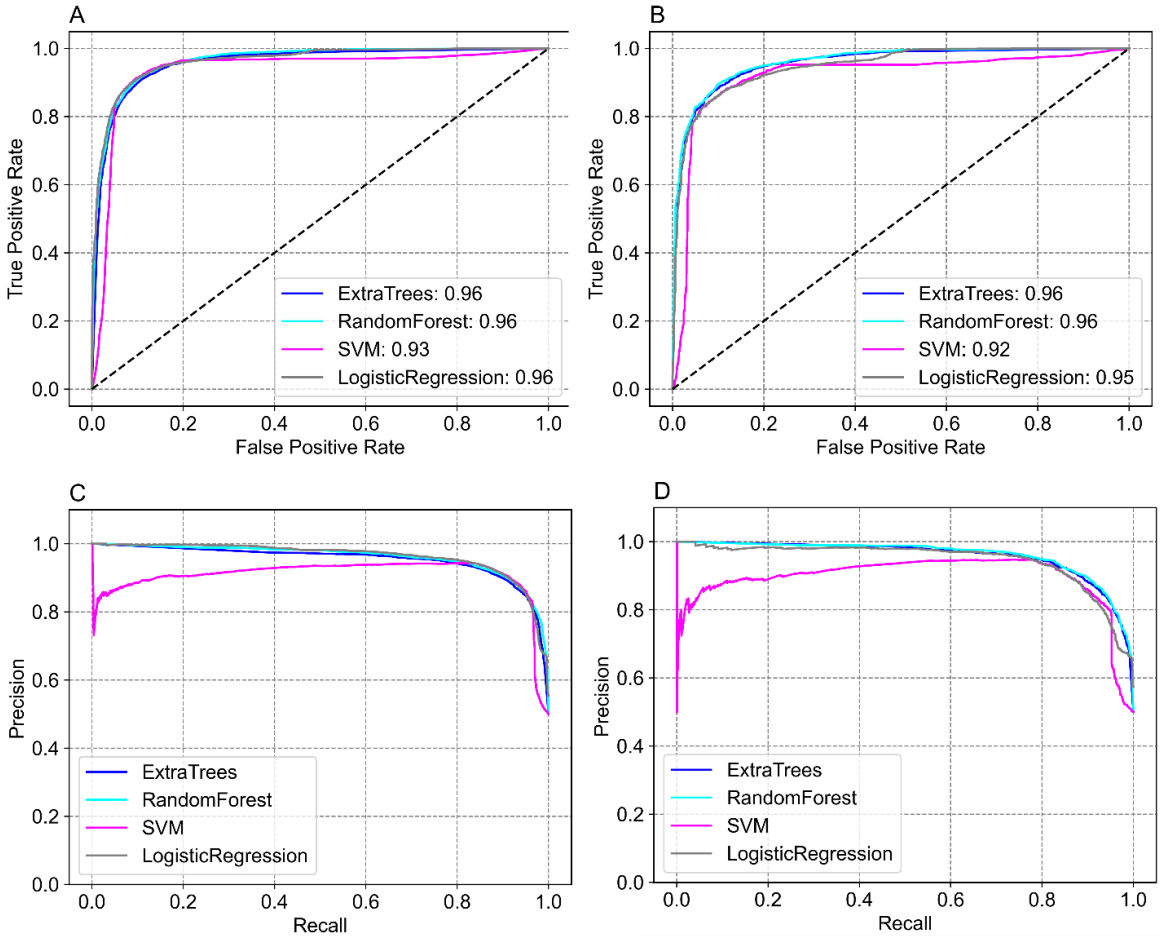
Zhou, Yanchen, Punitha Vedantham, Kai Lu, Juliet Agudelo, Ricardo Carrion, Jerritt W. Nunneley, Dale Barnard, et al. 2015. "Protease Inhibitors Targeting Coronavirus and Filovirus Entry." *Antiviral Research*. https://doi.org/10.1016/j.antiviral.2015.01.011.

Zhu, Na, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, et al. 2020. "A Novel Coronavirus from Patients with Pneumonia in China, 2019." *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa2001017.

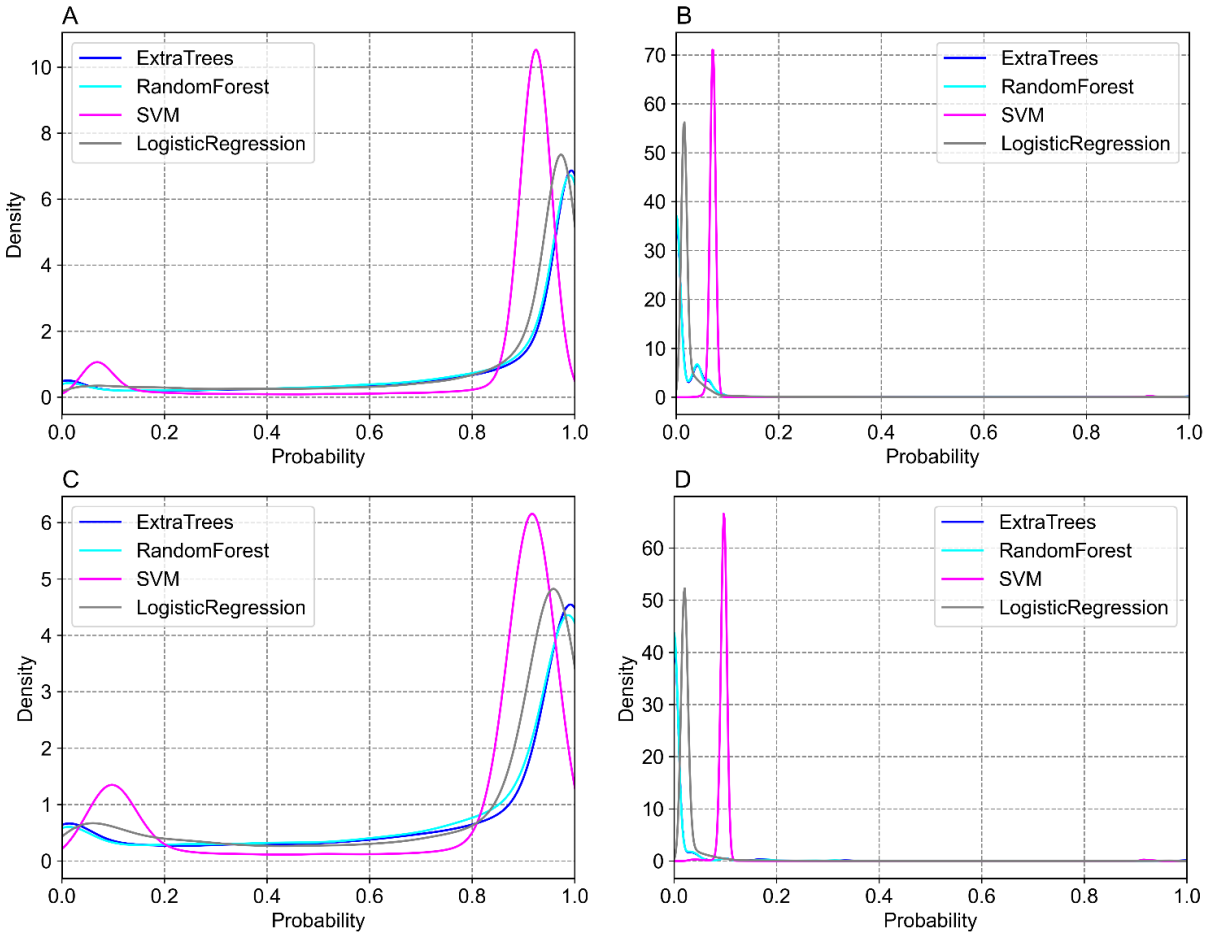
Zhu, Zhixing, Xihua Lian, Xiaoshan Su, Weijing Wu, Giuseppe A. Marraro, and Yiming Zeng. 2020. "From SARS and MERS to COVID-19: A Brief Summary and Comparison of Severe Acute Respiratory Infections Caused by Three Highly Pathogenic Human Coronaviruses." *Respiratory Research*. https://doi.org/10.1186/s12931-020-01479-w.

# Appendix 1. Chapter 2 Supplementary Material

**Supplementary Figure 2.1.** Benchmarking of 3DLigandSite on the cross-validation training-testing data. Receiver operator characteristic (ROC) curves and Precision-Recall graphs are shown for the prediction of biding sites of metal (A and B) and non-metal (C and D) ligands.



**Supplementary Figure 2.2. Assessment of the average probability scores assigned to binding and non-binding residues in the metal and non-metal residues in the validation set.** Non-metal binding residues (A), non-metal not binding residues (B), metal binding residues (C), metal not binding residues (D).



**Supplementary Table 2.1. PDB and chain Identifiers of protein structures.** The PDB and chain Identifiers of protein structures used in the training, testing, and validation of 3DLigandSite machine learning performance.

https://github.com/jmcgreig/3DLigandSite_Data

# Appendix 2. Chapter 3 Supplementary Material

**Assignment of isoforms to myosin sequences**

Our assignments of the isoforms are based on the data in UNIPROT. We have checked all of the sequences against several data bases including NCBI, and eggnog. All agree on the isoforms except eggnog in which 15 of the sequences had been assigned to a different isoform. In eggnog six NMB sequences have been assigned to NMA but this gives these species two copies of NMA and no NMB sequence. Similarly, seven α sequences have been assigned to β with the same results – two β sequences and no α. One IIa sequence was assigned to IIb. Given these anomalies we have used the UNIPROT assignments. Furthermore, we have generated a phylogenetic tree (**Supplementary figure S6**) in which sequences for each isoform group into the expected isoform.

**Why are conservative amino acid changes potentially important in tuning muscle contraction velocity?**

The velocity of contraction is 2-3-fold faster for a rat muscle expressing β-myosin than for the same isoform in human muscle. Before considering the role of conservative changes in sequence it is constructive to examine the free energy changes involved in such an acceleration of velocity.

The relationship between the rate constant defining a chemical/biochemical reaction and temperature was defined empirically by Arrhenius as $k = A\ e^{-Ea/R.T}$ where R is the gas constant, T is absolute temperature, A is an arbitrary constant and Ea is the activation energy – the energy barrier between the reactants and products. Essentially the ratio of the activation barrier (Ea) to thermal energy (R.T) defines the fraction of molecules with sufficient kinetic energy from the environment (at temperature, T) to pass over the barrier.

ADP release from the actin-myosin complex is the event thought to limit the velocity of contraction (see Table 3.2 and Figure 3.6). The activation energy for ADP release from A.M.ADP in human β-myosin has been measured as ~ 90 kJ $mol^{-1}$ (Deacon et al. 2012), while at room temperature (298 K) R.T is 2.48 kJ $mol^{-1}$. This means that few molecules at any one time will react. The rate of reaction can be accelerated by increasing the temperature or lowering the activation energy. The activation energy is measured from the temperature dependence of the rate constant and an activation energy of 90 kJ $mole^{-1}$ is typical of many

protein reactions. For an Ea of this size an increase of temperature of 10 K (or ~3% of 298 K) will increase the rate constant by ~3 fold. Changing the sequence of a protein can increase the rate of a reaction by reducing the activation energy. From the Arrhenius equation it is simple demonstrate that to increase k by a factor of 3 (at constant T and A) requires a decrease in Ea of the order of 2.7 kJ. $mol^{-1}$ . i.e. $k/k' = 3 = A\ e^{-Ea/RT} / A\ e^{-E'a/RT}$ or $\ln 3 = (E'a- Ea)/R.T.$ For a temperature of 298 K and R = 8.314 $J.mol^{-1}\ K^{-1}$ , then E'a- Ea = 2.7 $kJ.mol^{-1}$.

The decrease in activation energy is therefore small compared to Ea itself (2.7 vs 90 kJ $mol^{-1}$), small compared to a weak side chain interaction (e.g., a hydrogen bond is ~ 20 kJ mol-1) and the increase in k is equivalent to that induced by a 10 K temperature rise. In the case of the human-rat chimera we show that a 2-3-fold acceleration in the rate constant of ADP release and the velocity in the motility assay can be induced by a set of 9 amino acid substitutions some conservative (K434R, Y553F) others not (P343S, Q573P). The contribution of each side chain is impossible to assess from our work, we assume each makes a small contribution to the overall change. If we assume, for the sake of argument, that each amino acid substitution makes a similar size contribution to the change in Ea then each would contribute 2.7/9 or 0.30 kJ $mole^{-1}$ the equivalent of a 1 K temperature rise. From this calculation it would be unwise to assume a conservative amino acid substitution cannot make a contribution to the change in Ea. The slow change of contraction velocity over time then occurs by the accumulation or multiple very small changes in sequence each conferring a marginal advantage.

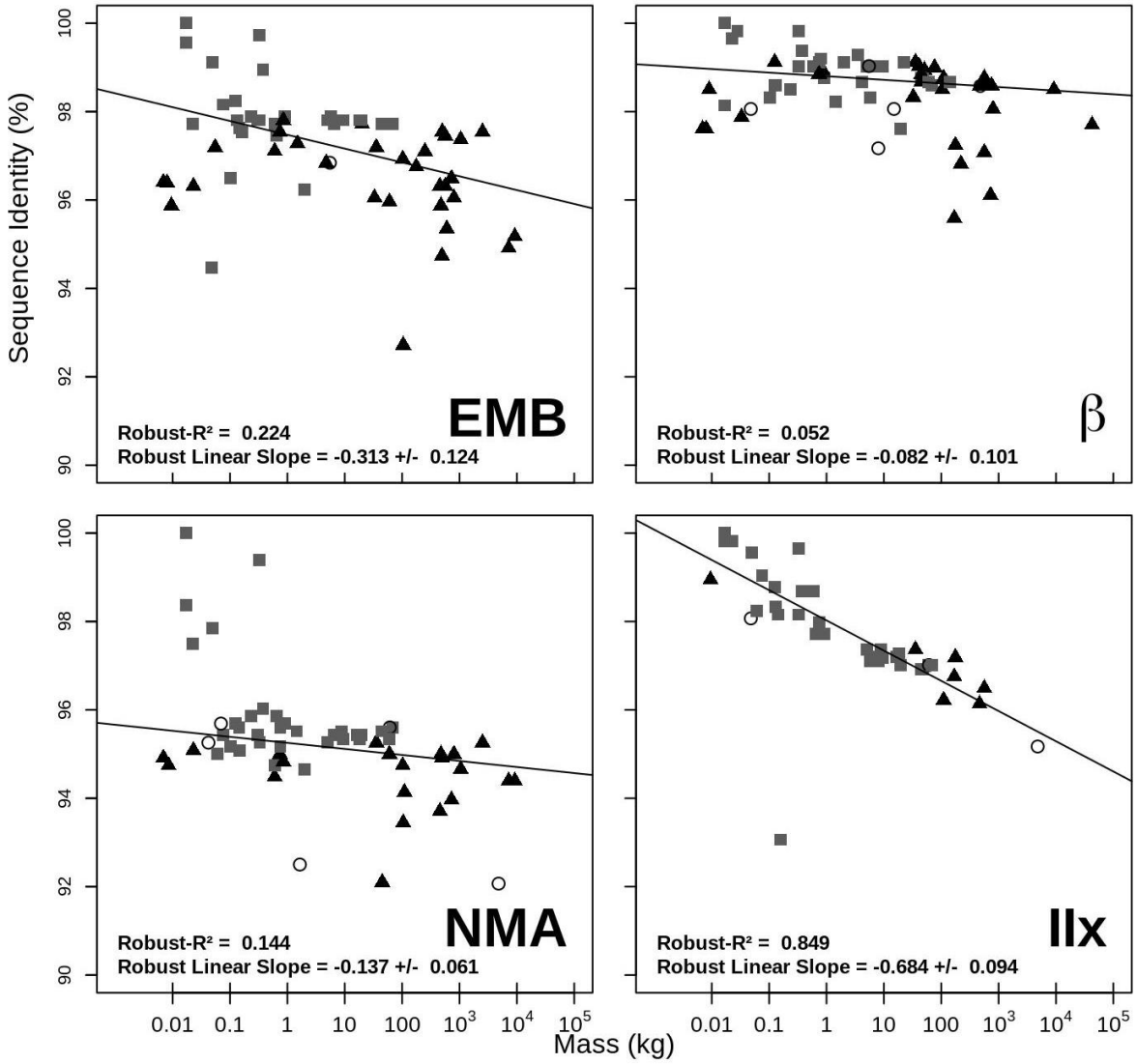
Conserved amino acids are defined as having similar properties (charge, hydrophobicity etc) but they are not identical. Small changes considered conservative Asp/Glu, Lys/Arg or Val/Lue can have a significant effect on activity. For example, the well-known serine protease family has a catalytic triad of Ser/His/Asp which is invariant, the Asp for example cannot in general be replaced by Glu (Barnes and Gray 2003). In the case of the myosin ~ 1000 mutations in human MyHC7 have been reported which are linked to familial cardiomyopathies of these ~60 % are in the motor domain (reviewed in Parker & Peckham 2020(F Parker and Peckham 2020). These myosins, containing a mutation are for the most part functional but have a hyper- or hypo- contractile phenotype often resulting in disease

in adulthood. Although most of the mutations would be considered non-conservative, many are not e.g., Asp/Glu (positions 554, 497), Val/Leu (pos 186, 216), Arg/Lys (pos 207, 721) Leu or Val/Met (338, 427).

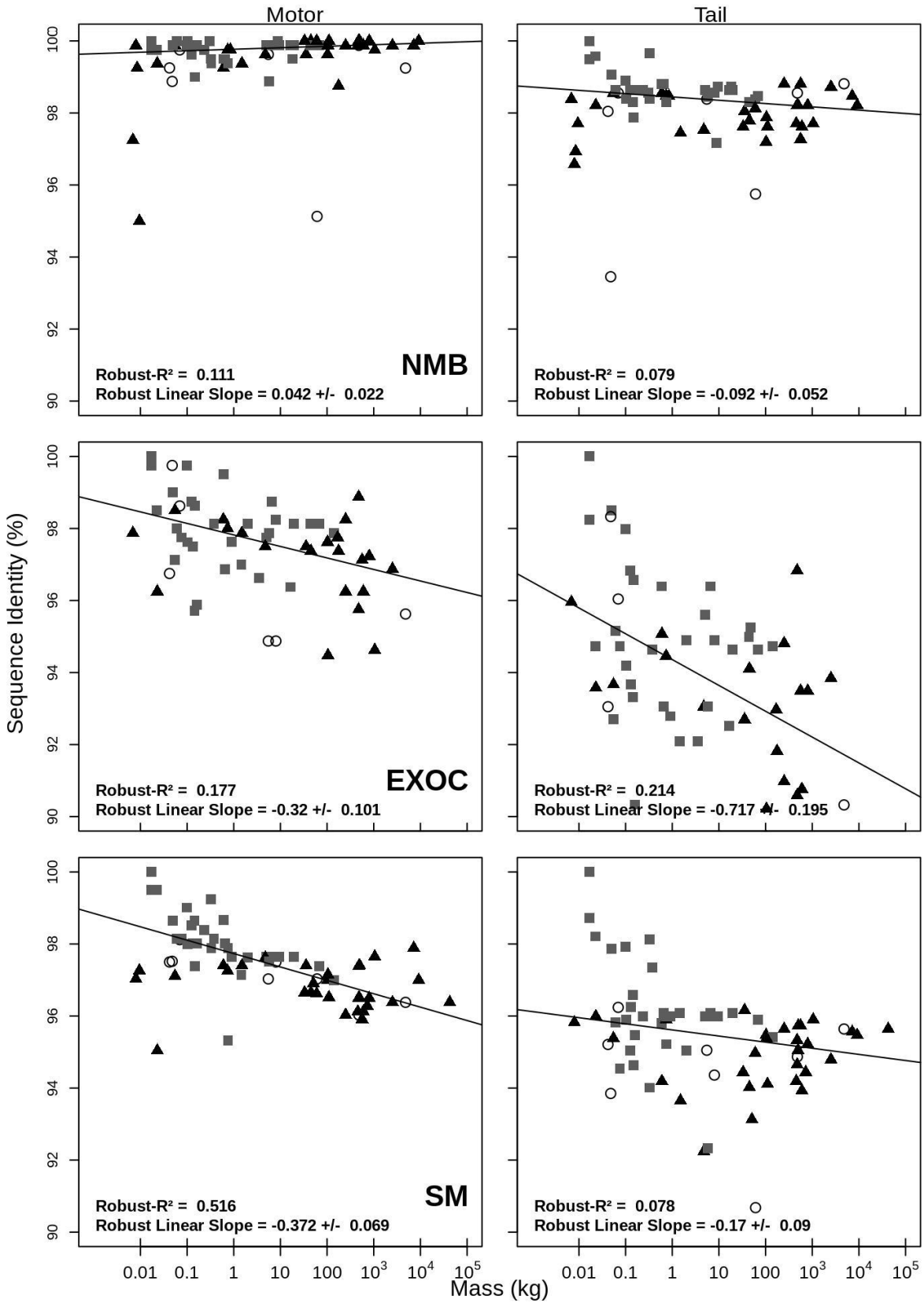
In addition to being important in specific interactions in protein catalysis, in ligand recognition and protein stability, amino acid side chains are important in conformational flexibility - the ability of proteins to access the optimal conformational states linked to function(Schmid and Hugel 2020). Since the sites we have identified are not directly involved in ADP binding we suspect conformational flexibility to be affected since mutations often affect the ability of the remote actin and ADP binding sites to communicate(M.J Bloemink et al. 2009; MJ Bloemink et al. 2011; Tsiavaliaris et al. 2002). In fact, the thermodynamic coupling between actin and ADP binding defines different types of myosin motor activity (M Bloemink and Geeves 2011).

**Supplementary Figure 3.1. Mass vs sequence identity plots.** The motor and tail domains have been analysed separately. The grey squares are Euarchontoglires, the black triangles are Laurasiatheria and the open circles are the Afrotheria and Metatheria groups. Each plot has been fitted with a robust linear regression. Sequence identity is pairwise to the mouse. The $R^2$ value and slope gradient are shown on each plot. Raw data files are available at Figshare.

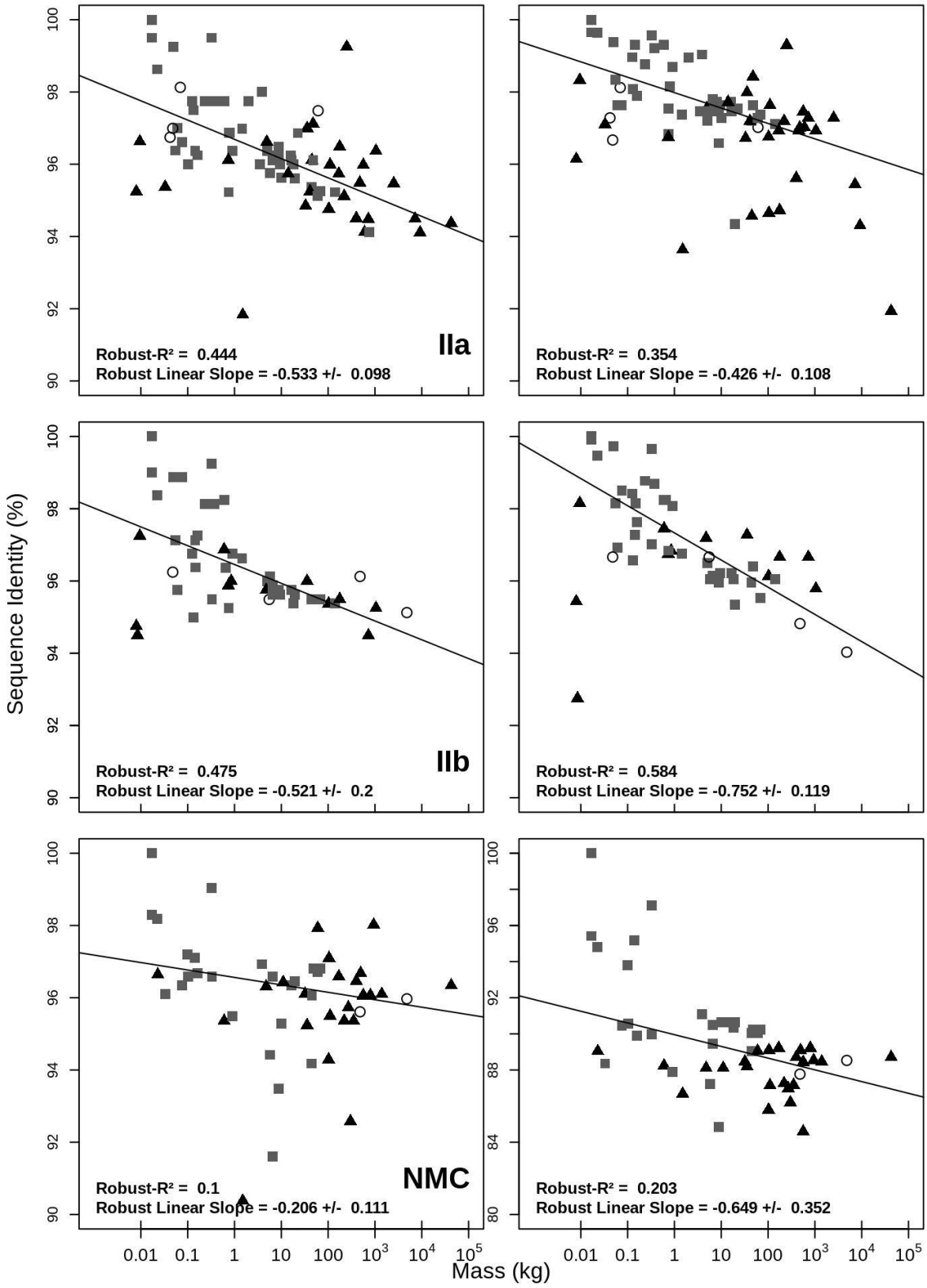
**A – Tail domains for EMB, β, NMA, IIx**



**B – Motor and Tail domains for NMB, EXOC and SM.**

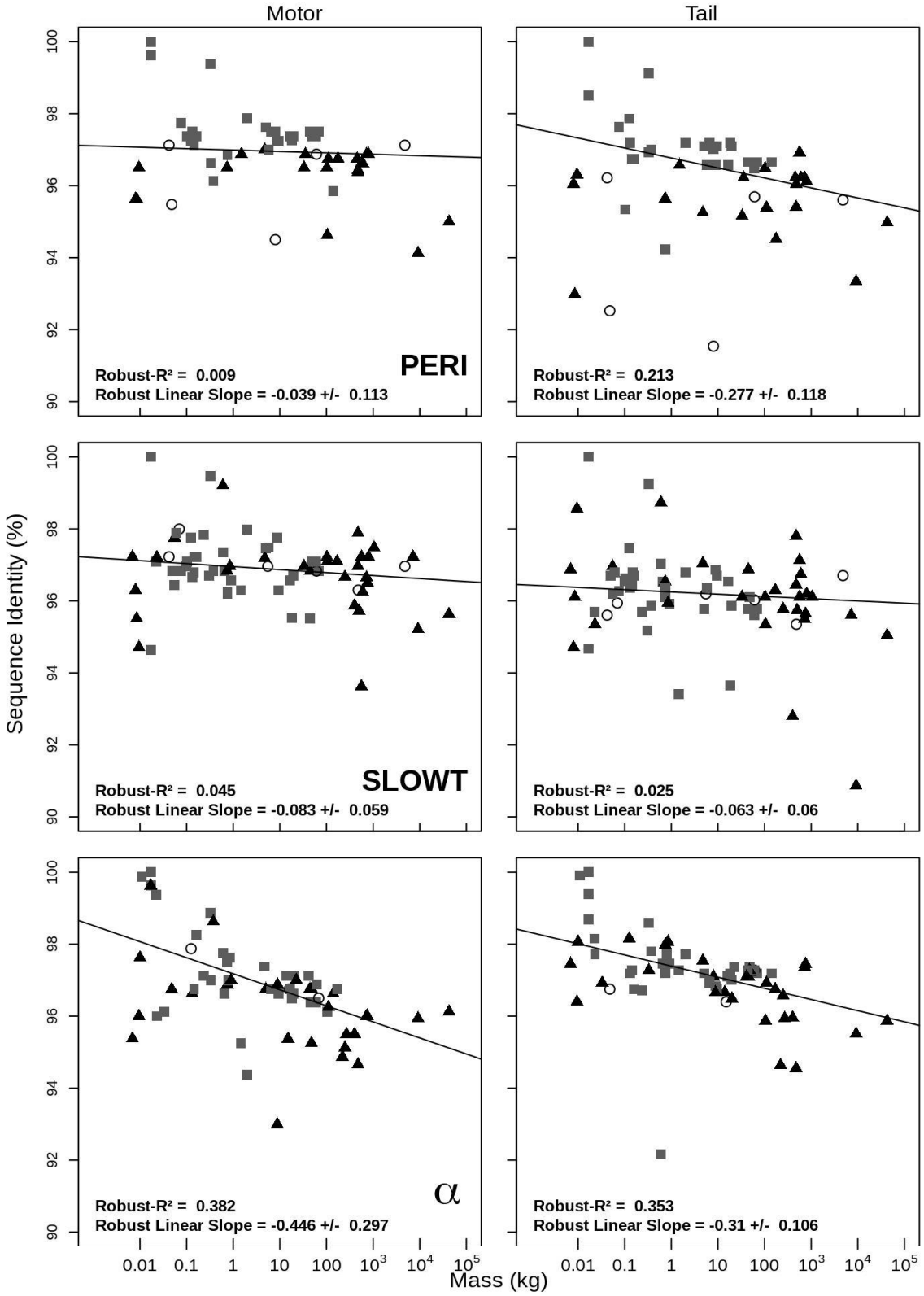


**C – Motor and Tail domains for IIa, IIb and NMC**

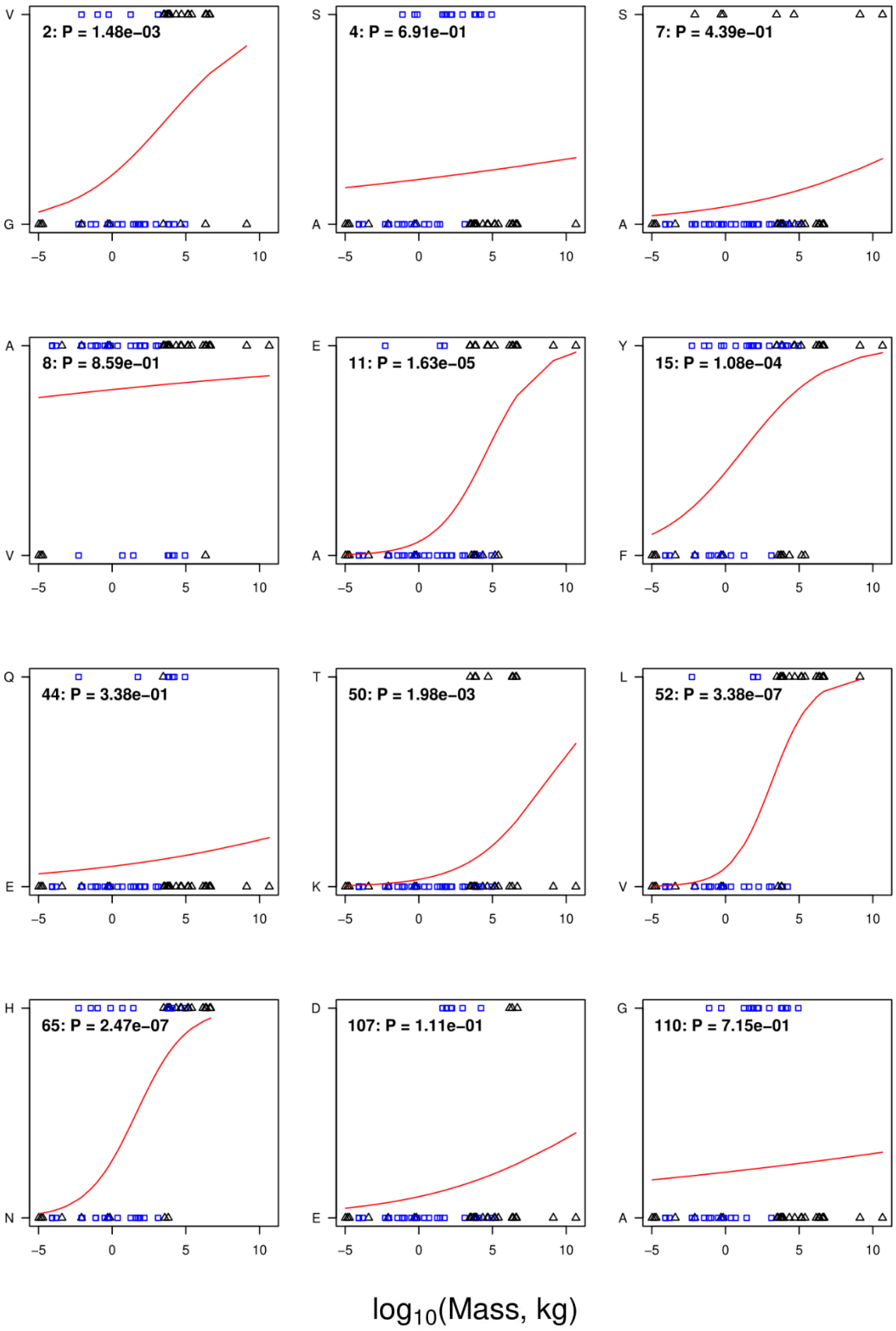


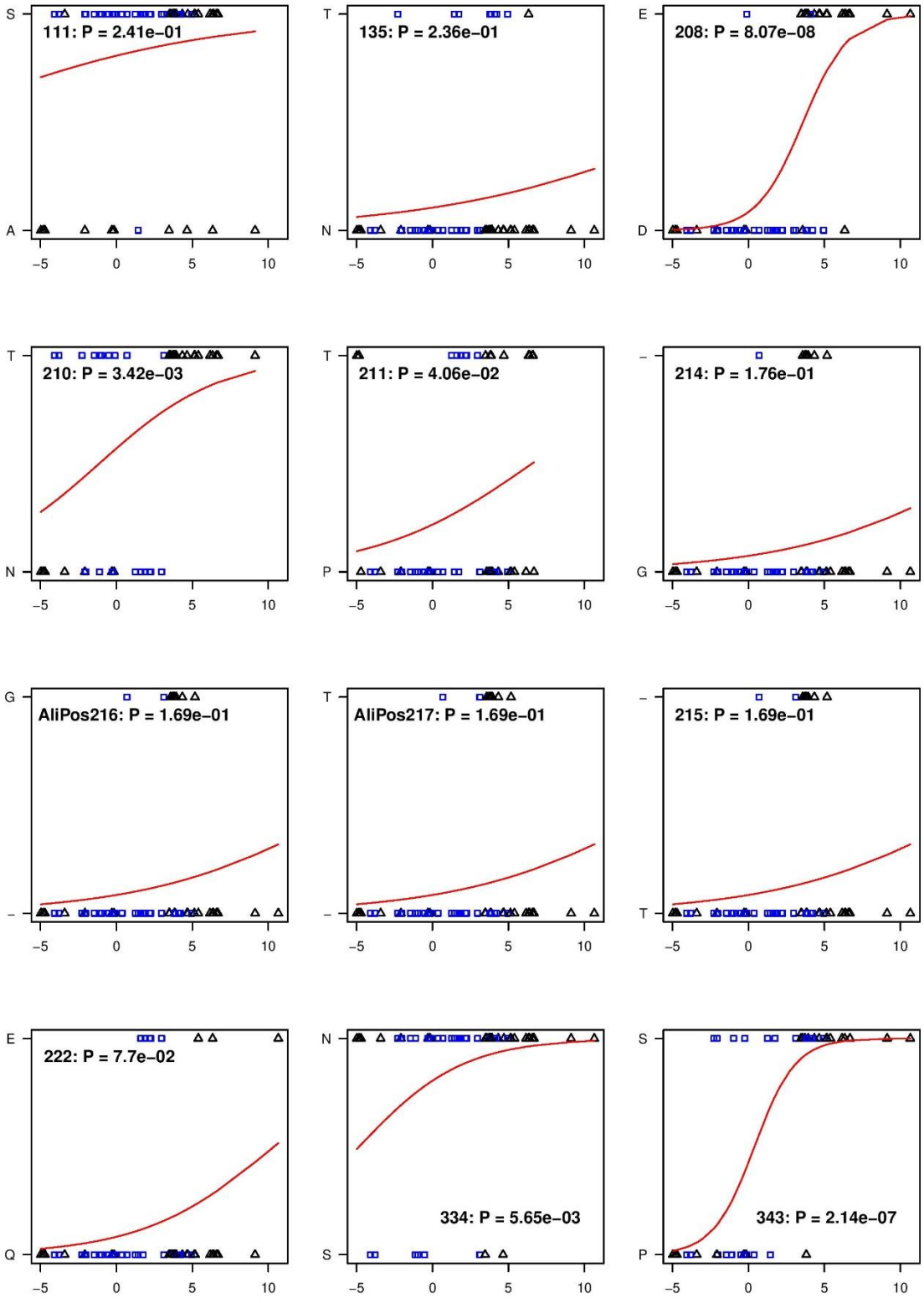

NB Note change of y-scale for NMC to accommodate the broader spread of values for the NMC tail

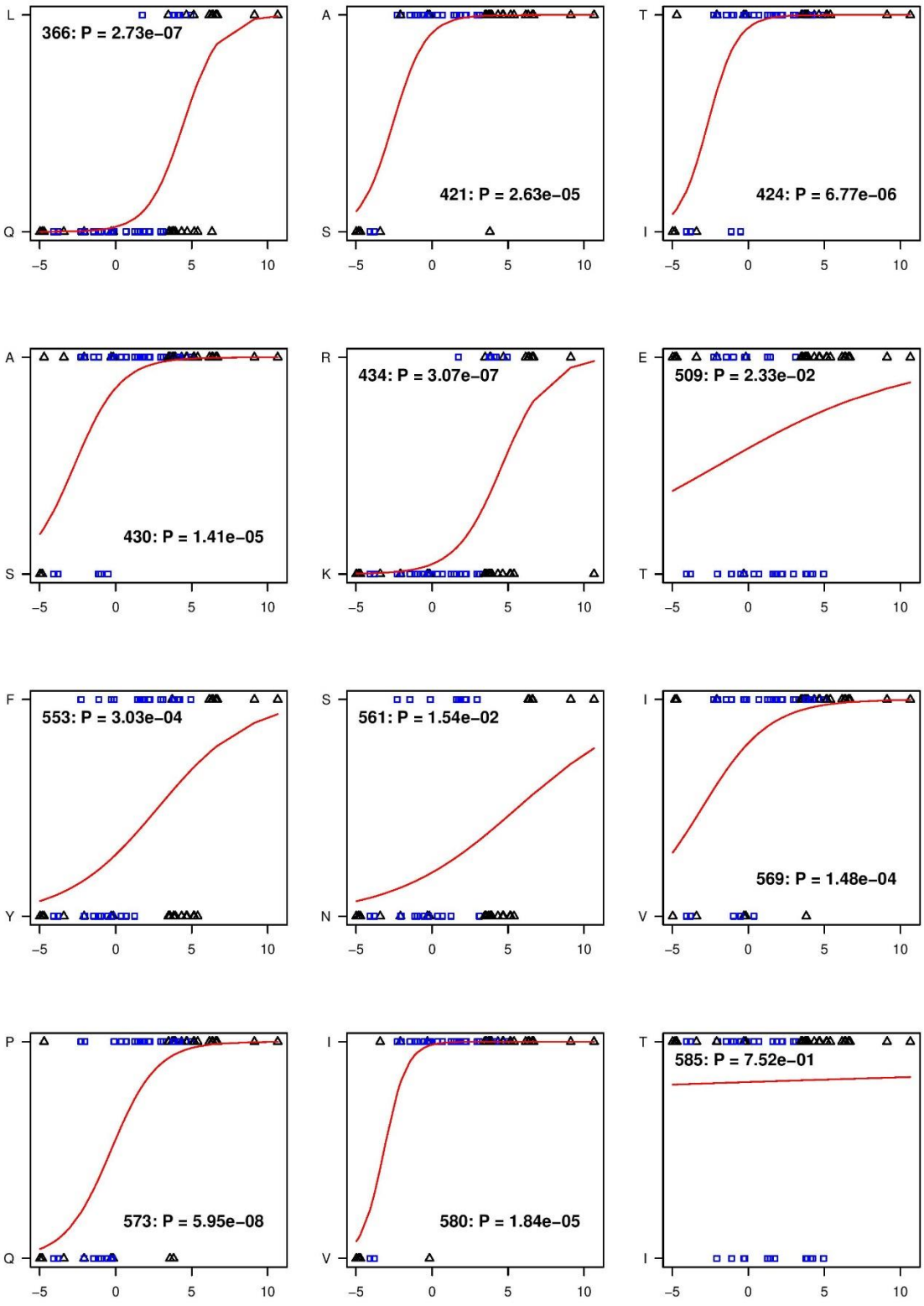
## D – Motor and Tail domains for PERI, SlowT, and α

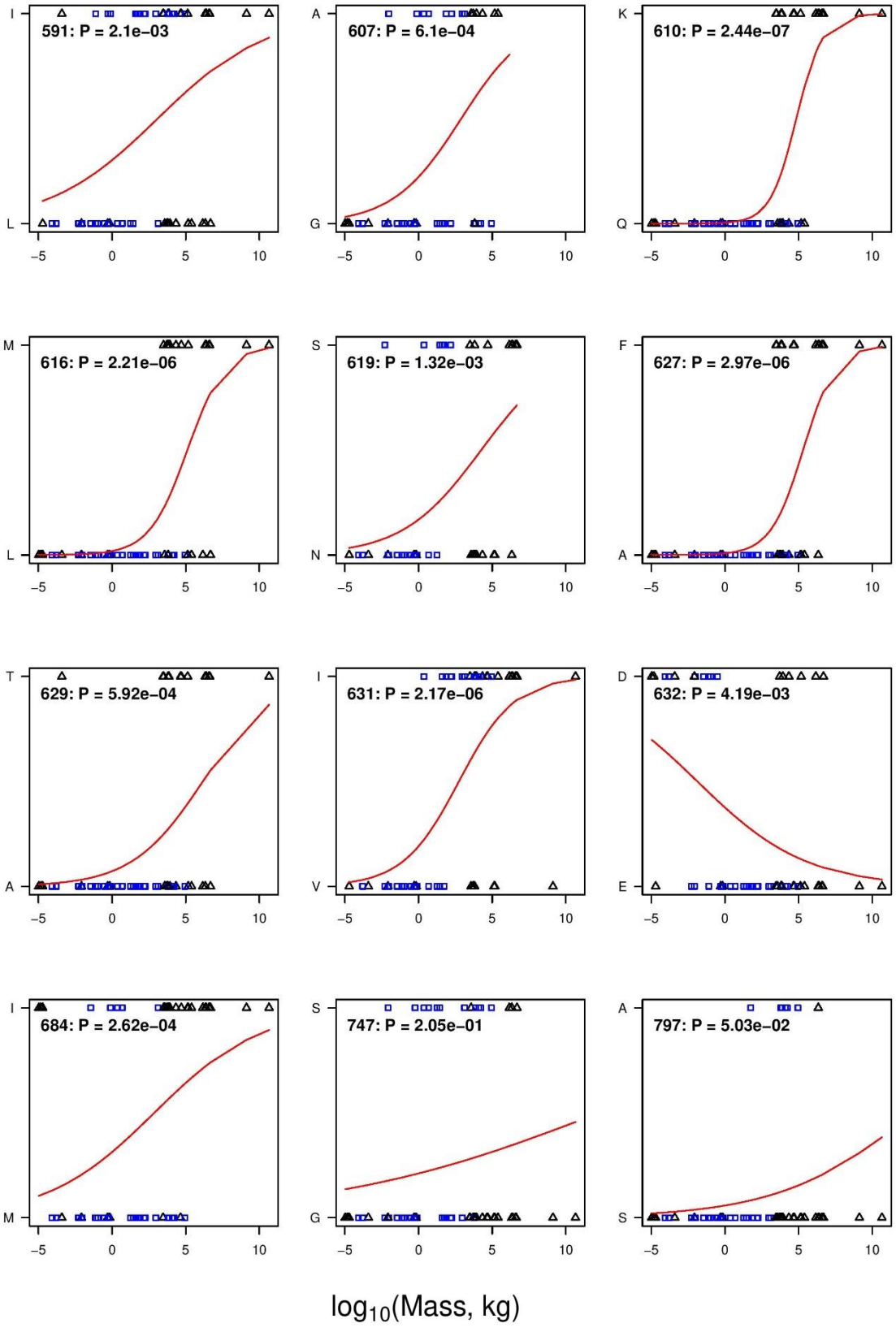


**Supplementary Figure 3.2. Residue-mass transition plots.** Binomial regression mapping the transition of the most frequent amino acid at positions in the motor region of β-myosin to the second most frequent amino acid at that position. The residue numbering is that of the human β-myosin, as oppose to the alignment position. The black squares are Euarchontoglires, and the triangles are Laurasiatheria. The P-value with each plot indicate the probability that the transition of the amino acids is not a result of change in mass. AliPos refers to positions in the sequence alignment that are not present in human β-myosin. Raw data files are available at Figshare.

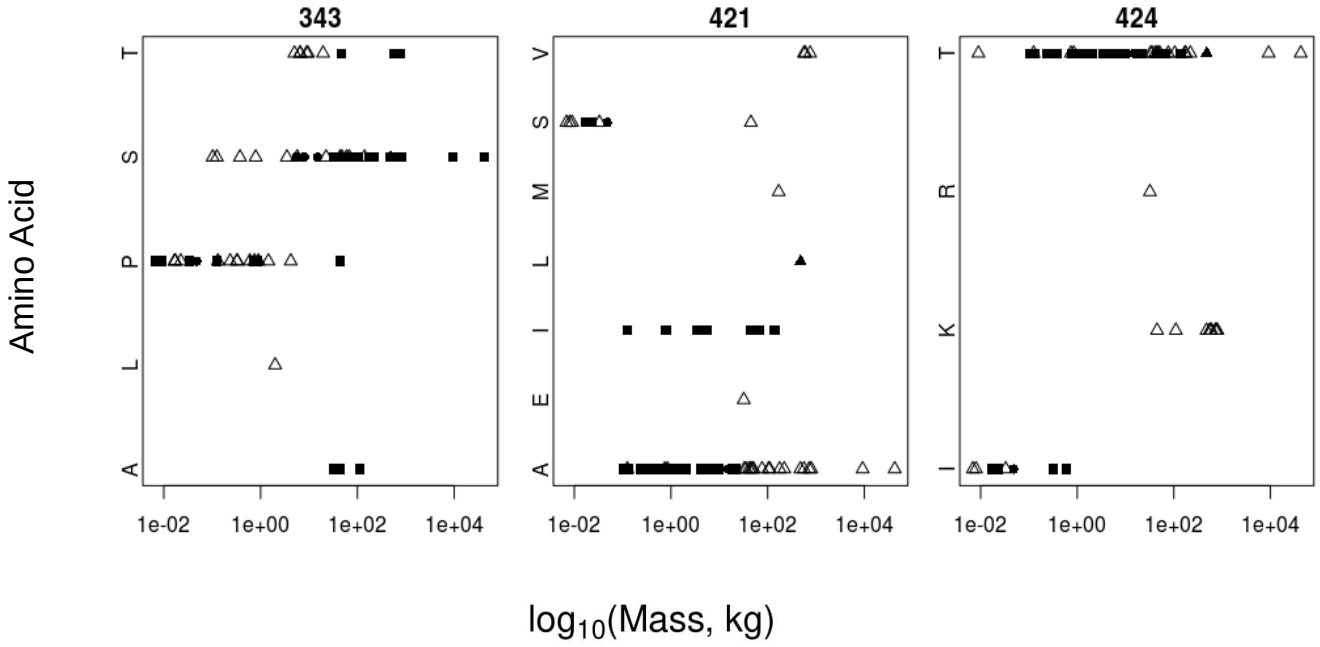

V
G
2: P = 1.48e−03
S
A
4: P = 6.91e−01
S
A
7: P = 4.39e−01
A
V
8: P = 8.59e−01
E
A
11: P = 1.63e−05
Y
F
15: P = 1.08e−04
Q
E
44: P = 3.38e−01
T
K
50: P = 1.98e−03
L
V
52: P = 3.38e−07
H
N
65: P = 2.47e−07
D
E
107: P = 1.11e−01
G
A
110: P = 7.15e−01
−5
0
5
10
log10(Mass, kg)


111: P = 2.41e−01
135: P = 2.36e−01
208: P = 8.07e−08
210: P = 3.42e−03
211: P = 4.06e−02
214: P = 1.76e−01
AliPos216: P = 1.69e−01
AliPos217: P = 1.69e−01
215: P = 1.69e−01
222: P = 7.7e−02
334: P = 5.65e−03
343: P = 2.14e−07
log10(Mass, kg)

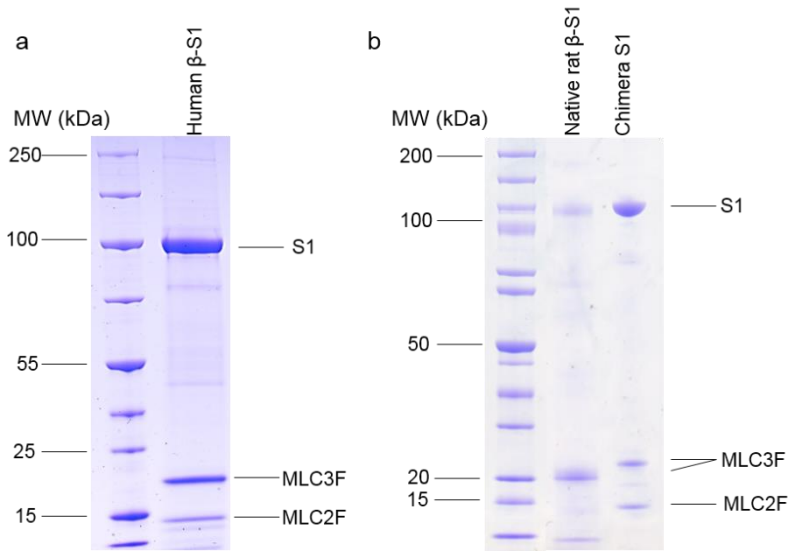

L
Q
366: P = 2.73e−07
A
S
421: P = 2.63e−05
T
I
424: P = 6.77e−06
A
S
430: P = 1.41e−05
R
K
434: P = 3.07e−07
E
T
509: P = 2.33e−02
F
Y
553: P = 3.03e−04
S
N
561: P = 1.54e−02
I
V
569: P = 1.48e−04
P
Q
573: P = 5.95e−08
I
V
580: P = 1.84e−05
T
I
585: P = 7.52e−01
−5
0
5
10
log10(Mass, kg)


591: P = 2.1e−03
607: P = 6.1e−04
610: P = 2.44e−07
616: P = 2.21e−06
619: P = 1.32e−03
627: P = 2.97e−06
629: P = 5.92e−04
631: P = 2.17e−06
632: P = 4.19e−03
684: P = 2.62e−04
747: P = 2.05e−01
797: P = 5.03e−02
log10(Mass, kg)

**Supplementary Figure 3.3. Highly variable residue mass vs amino acid frequencies.** Residues which had more than two sites of variation, with the third most frequent amino acid being close in frequency to the second most common amino acid. The black squares are Euarchontoglires, and the triangles are Laurasiatheria. The residue numbering is that of human β-cardiac myosin. Raw data files are available at Figshare.



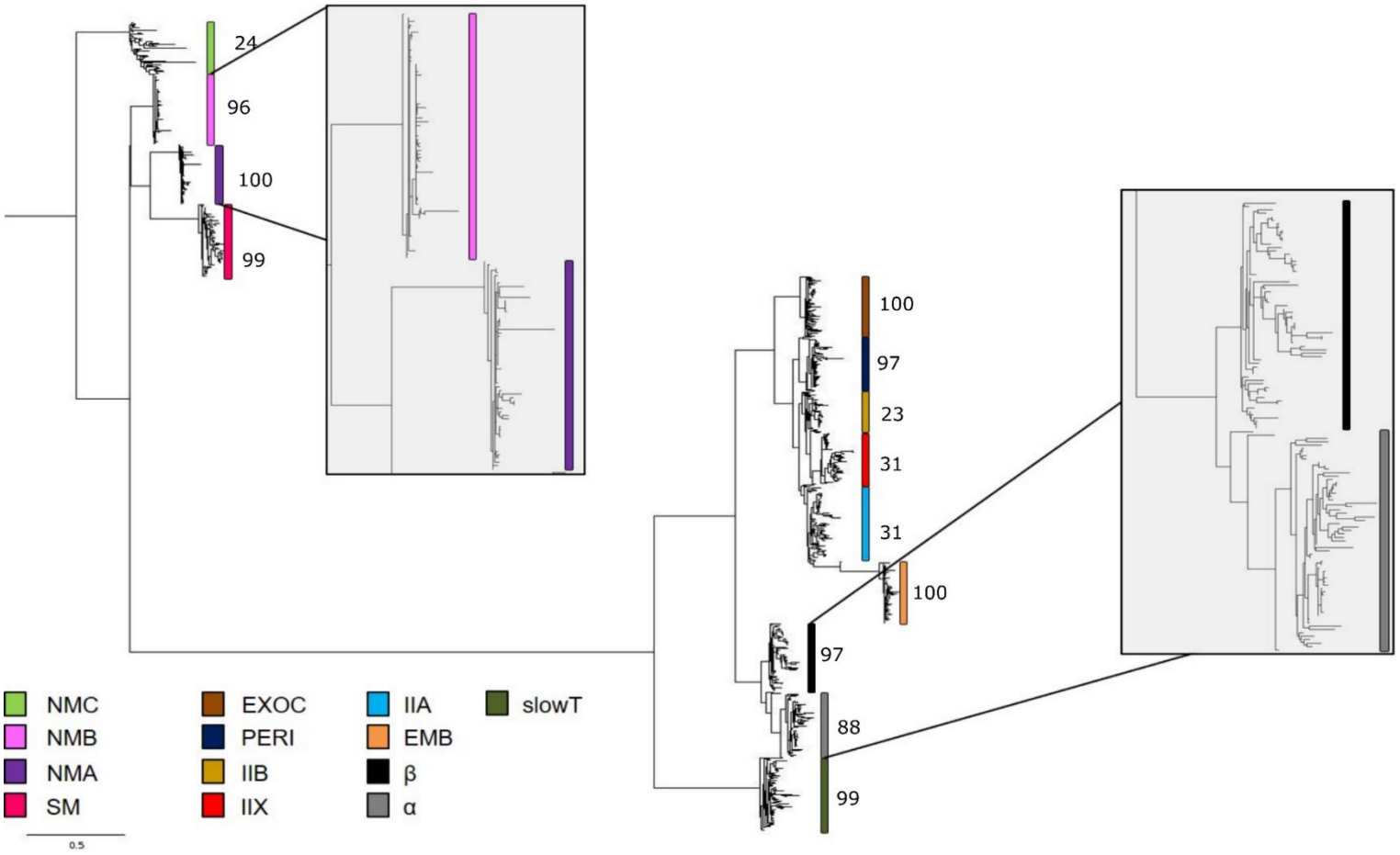
**Supplementary Figure 3.4. SDS-PAGE of the three protein preparations.** A. Recombinant human β-S1 with 2 light chains. B. Native rat β-S1 and recombinant chimera S1 with 2 light chains.



**Supplementary Figure 3.5. Distribution of amino acids at twelve positions.** This table shows the amino acids that occurs in each of the twelve residues, those predicted to be assocaited with mass and that are of interest, in each species, sorted by mass. Yellow background is the predominant amino acid in small mammals. Blue background is the predominant amino acid in large mammals. The clades Laurasiatheria, Euarchontoglires, Metatheria, and Afrotheria are represented by the letter L, E, M and A respectively.

| Species | Clade | 366 | 434 | 326 | 553 | 343 | 349 | 573 | 421 | 424 | 430 | 569 | 580 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Myotis_brandtii | L | Q | K | S | Y | P | I | Q | S | I | S | V | V |
| Myotis_lucifugus | L | Q | K | S | Y | P | I | Q | S | I | S | I | V |
| Hipposideros_armiger | L | Q | K | S | Y | P | I | P | S | T | A | I | V |
| Peromyscus_maniculatus_bairdii | E | Q | K | S | Y | P | I | Q | S | I | S | V | V |
| Mus_musculus | E | Q | K | S | Y | P | I | Q | S | I | S | V | V |
| Mus_pahari | E | Q | K | S | Y | P | I | Q | S | I | S | V | V |
| Cricetulus_griseus | L | Q | K | S | Y | P | I | Q | S | I | S | V | V |
| Desmodus_rotundus | M | Q | K | S | Y | P | I | Q | S | I | A | V | I |
| Monodelphis_domestica | E | Q | K | S | Y | P | I | Q | S | I | S | I | I |
| Otolemur_garnettii | E | Q | K | A | F | S | M | P | A | T | A | I | I |
| Rousettus_aegyptiacus | L | Q | K | S | Y | P | I | Q | A | T | A | I | I |
| Ictidomys_tridecemlineatus | E | Q | K | S | Y | S | I | Q | I | T | A | I | I |
| Tarsius_syrichta | E | Q | K | S | Y | P | I | P | A | T | A | I | I |
| Octodon_degus | E | Q | K | S | Y | P | I | Q | A | T | A | I | I |
| Rattus_norvegicus | E | Q | K | S | Y | P | I | Q | A | I | S | I | I |
| Callithrix_jacchus | E | Q | K | S | F | P | M | Q | A | T | A | I | I |
| Nannospalax_galili | E | Q | K | S | Y | S | I | Q | A | T | S | V | I |
| Jaculus_jaculus | E | Q | K | S | Y | P | I | Q | A | I | S | V | I |
| Pteropus_alecto | E | Q | K | A | Y | P | I | Q | A | T | A | V | I |
| Aotus_nancymaae | E | Q | K | S | F | P | M | Q | A | T | A | I | I |
| Urocitellus_parryii | E | Q | K | S | Y | S | I | Q | I | T | A | I | I |
| Pteropus_vampyrus | L | Q | K | A | Y | P | I | Q | A | T | A | V | V |
| Cavia_porcellus | E | Q | K | S | F | P | M | P | A | T | A | I | I |
| Galeopterus_variegatus | E | M | K | S | Y | P | M | P | A | T | A | V | I |
| Oryctolagus_cuniculus | E | Q | K | S | Y | L | M | P | A | T | A | I | I |
| Marmota_marmota_marmota | E | Q | K | S | Y | S | I | P | I | T | A | I | I |
| Propithecus_coquereli | E | Q | K | A | F | P | M | P | A | T | A | I | I |
| Chlorocebus_sabaeus | E | Q | K | A | F | T | M | P | A | T | A | I | I |
| Dasypus_novemcinctus | M | Q | K | A | Y | S | I | P | A | T | A | I | I |
| Nomascus_leucogenys | E | L | R | A | F | S | I | P | I | T | A | I | I |
| Macaca_fascicularis | E | Q | K | A | F | T | M | P | A | T | A | I | I |
| Macaca_nemestrina | E | Q | K | A | F | T | M | P | A | T | A | I | I |
| Sarcophilus_harrisii | M | Q | K | S | Y | S | M | P | A | T | S | I | I |
| Colobus_angolensis_palliatus | E | Q | K | A | F | T | M | P | A | T | A | I | I |
| Cercocebus_atys | E | Q | K | A | F | T | M | P | A | T | A | I | I |
| Phascolarctos_cinereus | M | L | K | S | F | S | I | P | A | T | A | I | I |
| Papio_anubis | E | Q | K | S | F | S | M | P | A | T | A | I | I |
| Castor_canadensis | E | Q | K | S | F | S | M | P | A | T | A | I | I |
| Neophocaena_asiaeorientalis_asiaeorientalis | L | L | K | A | Y | S | I | P | E | R | A | I | I |
| Pantholops_hodgsonii | L | Q | R | A | Y | A | M | P | A | T | A | I | I |
| Canis_lupus_familiaris | L | Q | K | S | Y | S | M | Q | A | T | A | I | I |
| Panthera_pardus | L | Q | K | A | F | S | M | P | A | T | A | I | I |
| Pan_paniscus | E | L | R | A | F | S | M | P | I | T | A | I | I |
| Capra_hircus | L | Q | R | A | Y | A | M | P | A | K | A | I | I |
| Enhydra_lutris_kenyoni | L | Q | K | S | Y | P | M | Q | S | T | A | V | I |
| Odocoileus_virginianus_texanus | L | Q | R | A | Y | T | M | P | A | T | A | I | I |
| Pan_troglodytes | E | L | R | A | F | S | M | P | I | T | A | I | I |
| Acinonyx_jubatus | L | Q | K | A | Y | S | M | P | A | T | A | I | I |
| Pongo_abelii | E | L | R | A | F | S | M | P | I | T | A | I | I |
| Homo_sapiens | E | L | R | A | F | S | M | P | I | T | A | I | I |
| Puma_concolor | L | Q | K | A | Y | S | M | P | A | T | A | I | I |
| Lipotes_vexillifer | L | L | K | A | Y | S | I | P | A | T | A | I | I |
| Ovis_aries | L | Q | R | A | Y | A | M | P | A | K | A | I | I |
| Gorilla_gorilla_gorilla | E | L | R | A | F | S | M | P | I | T | A | I | I |
| Sus_scrofa | L | L | K | A | Y | S | M | P | M | T | A | I | I |
| Panthera_tigris_altaica | L | Q | K | A | Y | S | M | P | A | T | A | I | I |
| Neomonachus_schauinslandi | L | Q | K | A | Y | S | M | P | A | T | A | I | I |
| Camelus_bactrianus | L | L | R | A | F | S | M | P | A | K | A | I | I |
| Trichechus_manatus_latirostris | A | L | R | A | F | S | M | P | L | T | A | I | I |
| Bos_mutus | L | L | R | A | F | T | M | P | V | K | A | I | I |
| Equus_caballus | L | Q | R | A | F | S | M | P | A | K | A | I | I |
| Bison_bison_bison | L | L | R | A | F | T | M | P | V | K | A | I | I |
| Bubalus_bubalis | L | L | R | A | F | T | M | P | A | K | A | I | I |
| Bos_taurus | L | L | R | A | F | T | M | P | V | K | A | I | I |
| Camelus_ferus | L | L | R | A | F | S | M | P | A | K | A | I | I |
| Balaenoptera_acutorostrata_scammoni | L | L | R | A | F | S | M | P | A | T | A | I | I |
| Physeter_catodon | L | L | K | A | F | S | M | P | A | T | A | I | I |

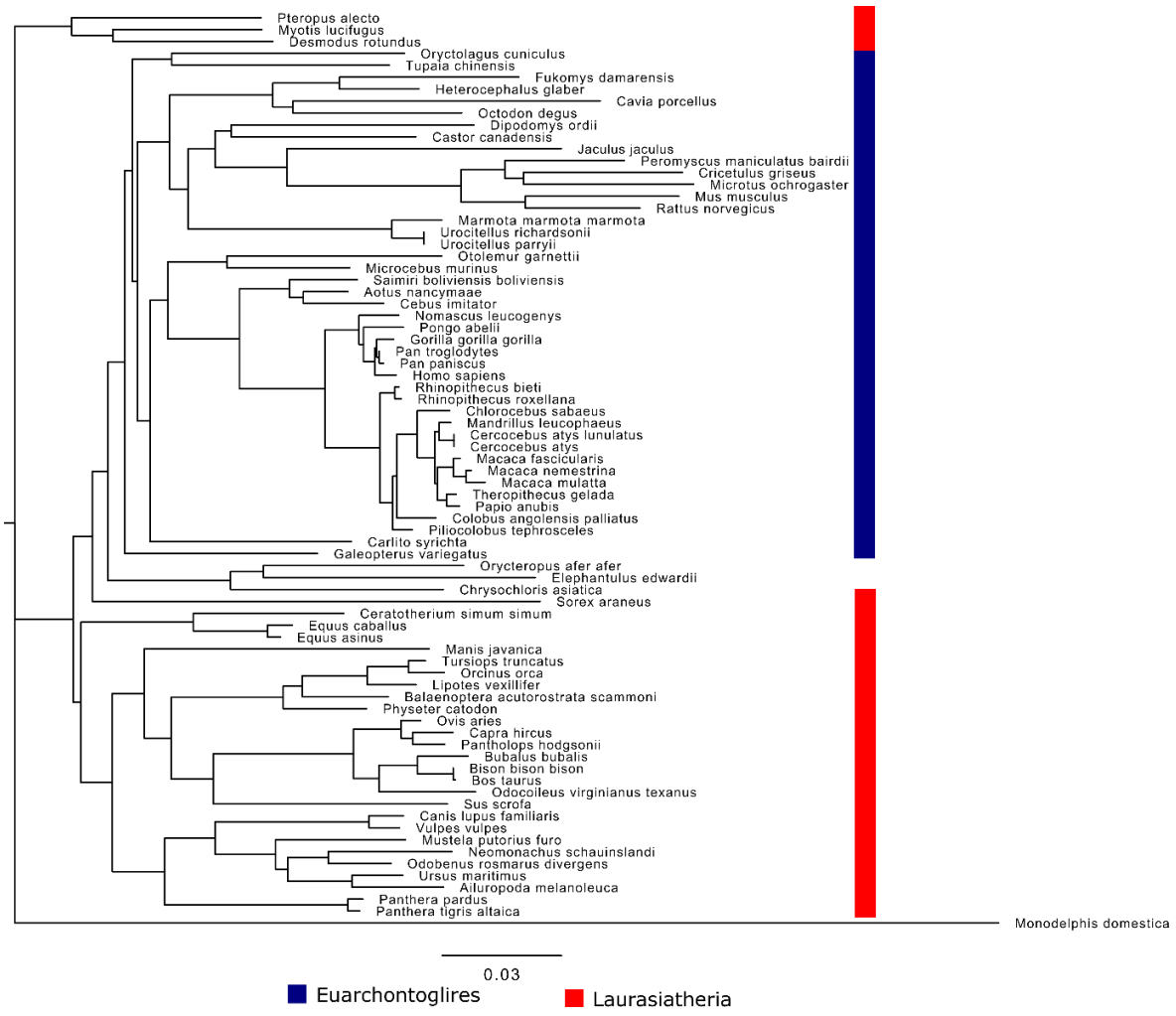
**Supplementary Figure 3.6. Tree of all myosin isoforms.** Maximum-Likelihood phylogenetic tree generated for all myosin isoforms. Isoforms are labelled according to the key. Zoomed in sections show the α and β regions, and the NMA and NMB regions. Bootstrap values for the isoform branches are shown adjacent to the bars identifying each isoform.


24
96
100
99
100
97
23
31
31
100
97
88
99
NMC
NMB
NMA
SM
EXOC
PERI
IIB
IIX
IIA
EMB
β
α
slowT
0.5

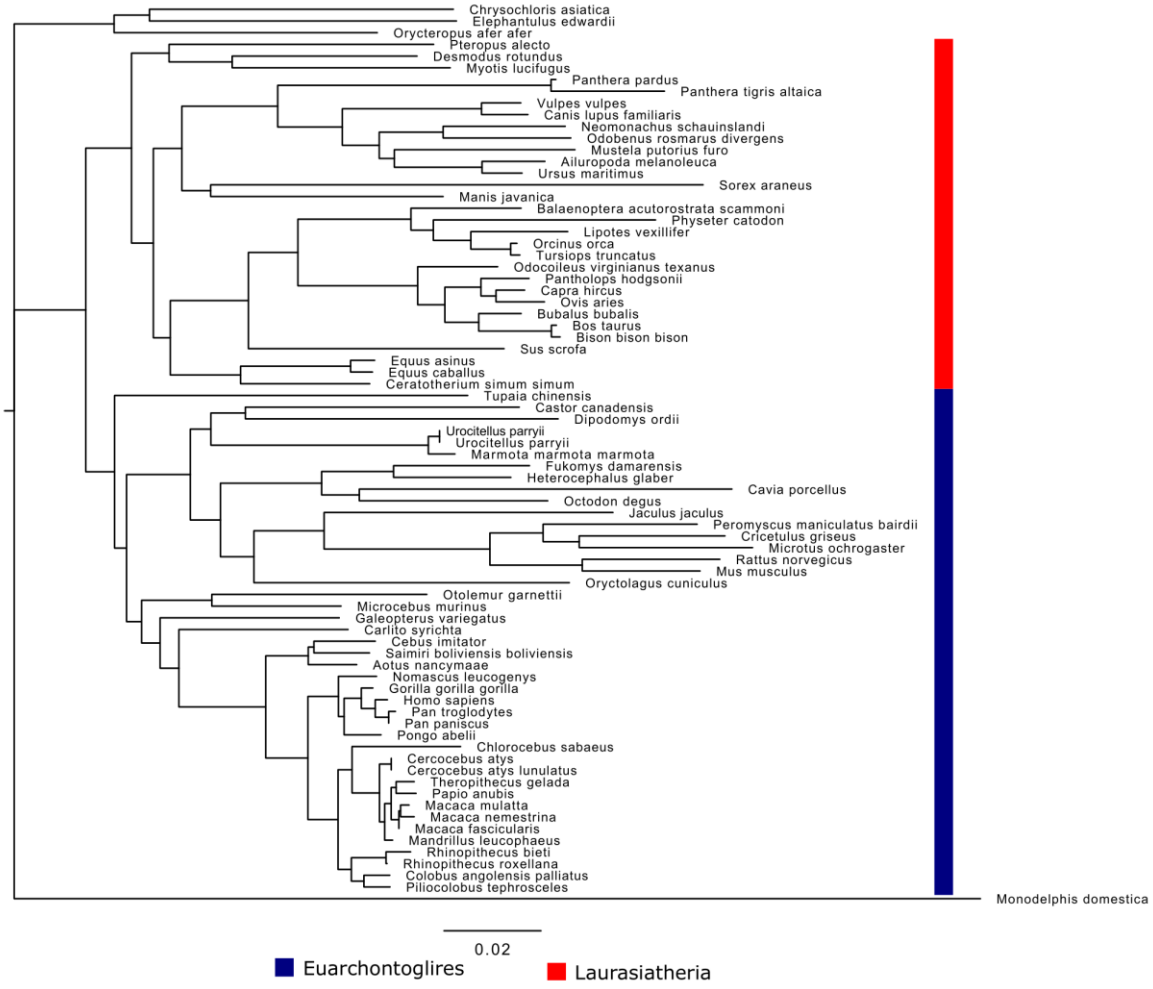
**Supplementary Figure 3.7. Phylogenetic (Gene) Trees for sarcomeric isoforms.**

Phylogenetic trees for IIa, IIb, IIx, α and β isoforms motor and tail domains.

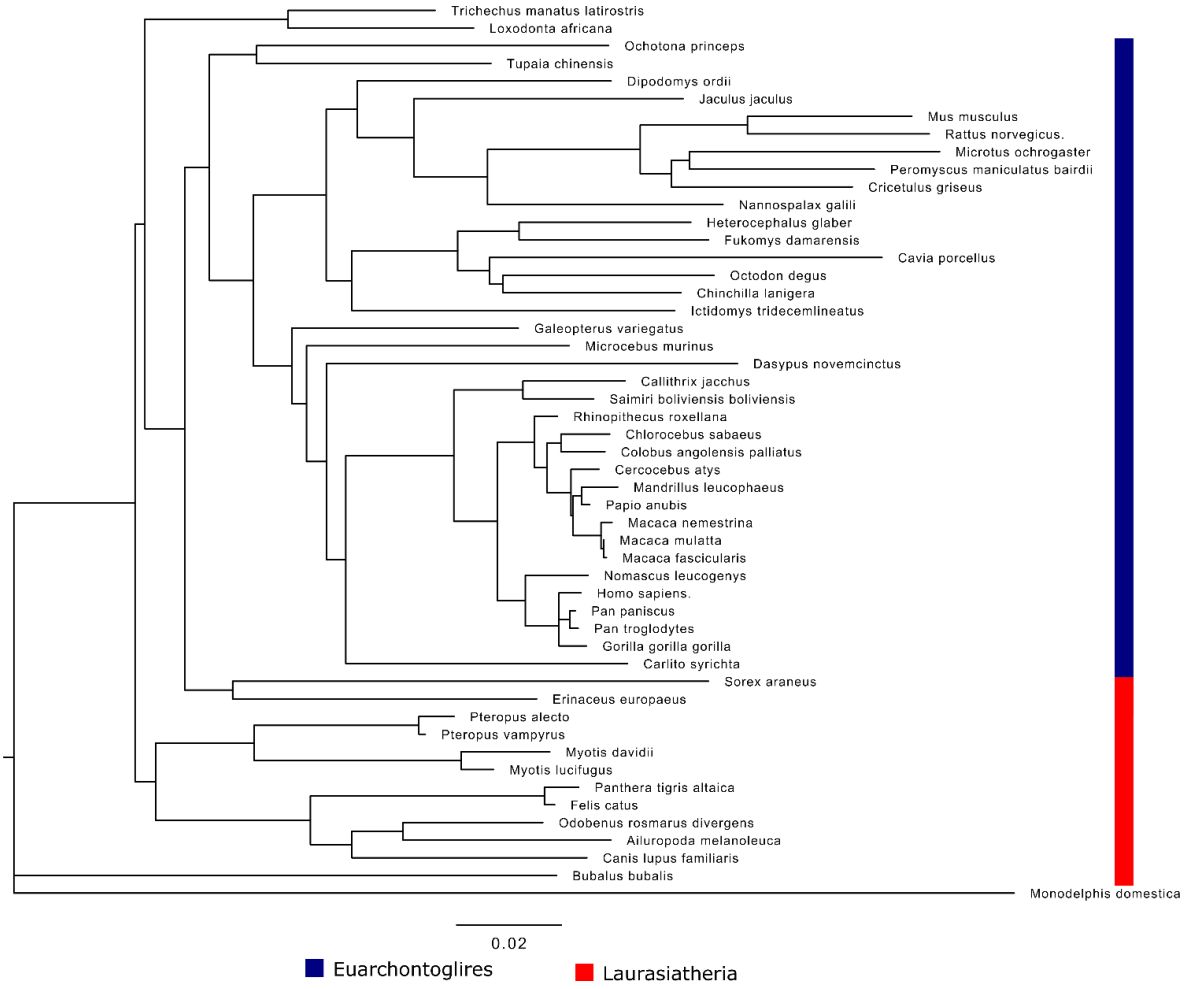
**A** IIa Motor



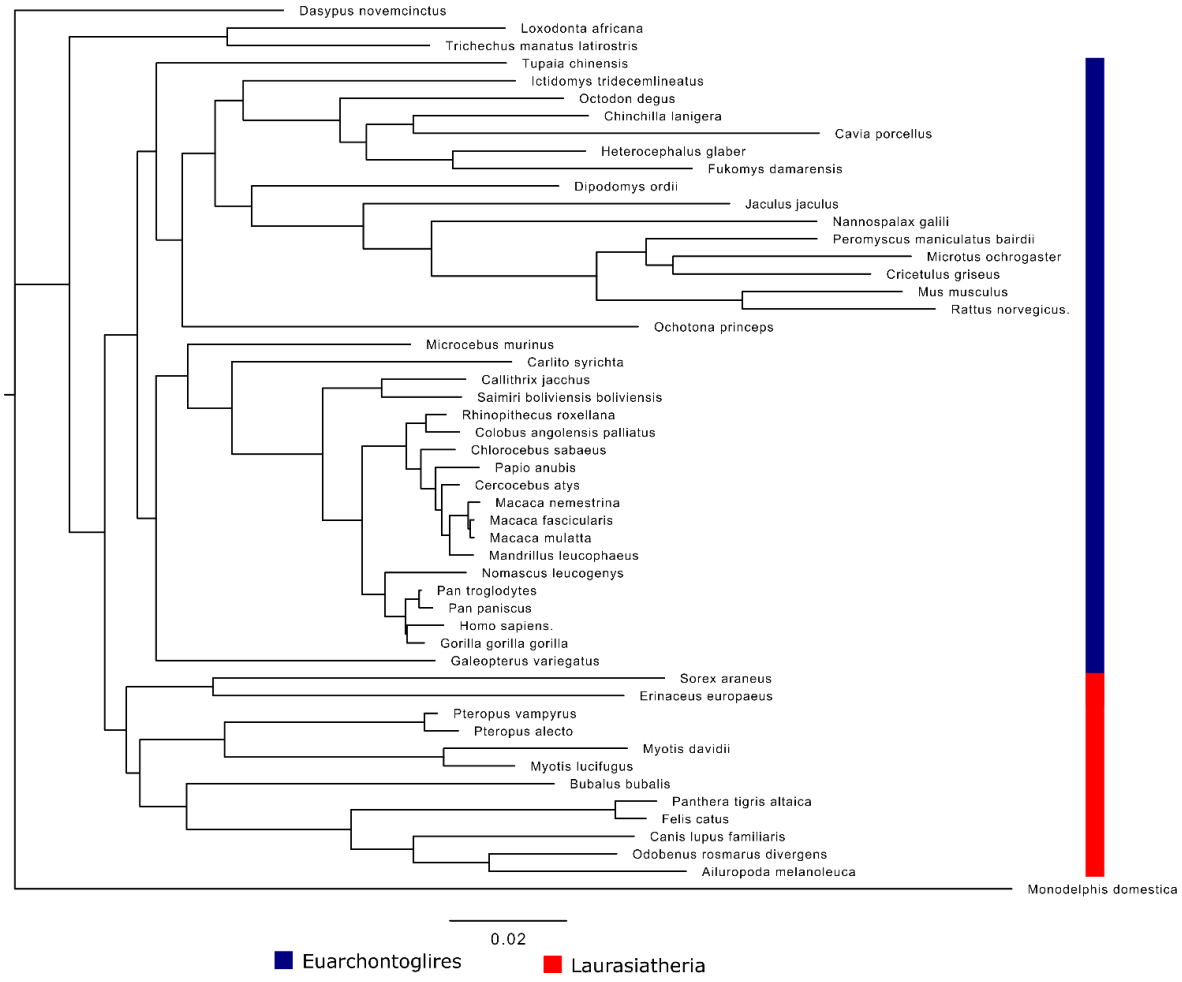
**B** IIa Tail


Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Pteropus alecto
Desmodus rotundus
Myotis lucifugus
Panthera pardus
Panthera tigris altaica
Vulpes vulpes
Canis lupus familiaris
Neomonachus schauinslandi
Odobenus rosmarus divergens
Mustela putorius furo
Ailuropoda melanoleuca
Ursus maritimus
Sorex araneus
Manis javanica
Balaenoptera acutorostrata scammoni
Physeter catodon
Lipotes vexillifer
Orcinus orca
Tursiops truncatus
Odocoileus virginianus texanus
Pantholops hodgsonii
Capra hircus
Ovis aries
Bubalus bubalis
Bos taurus
Bison bison bison
Sus scrofa
Equus asinus
Equus caballus
Ceratotherium simum simum
Tupaia chinensis
Castor canadensis
Dipodomys ordii
Urocitellus parryii
Urocitellus parryii
Marmota marmota marmota
Fukomys damarensis
Heterocephalus glaber
Cavia porcellus
Octodon degus
Jaculus jaculus
Peromyscus maniculatus bairdii
Cricetulus griseus
Microtus ochrogaster
Rattus norvegicus
Mus musculus
Oryctolagus cuniculus
Otolemur garnettii
Microcebus murinus
Galeopterus variegatus
Carlito syrichta
Cebus imitator
Saimiri boliviensis boliviensis
Aotus nancymaae
Nomascus leucogenys
Gorilla gorilla gorilla
Homo sapiens
Pan troglodytes
Pan paniscus
Pongo abelii
Chlorocebus sabaeus
Cercocebus atys
Cercocebus atys lunulatus
Theropithecus gelada
Papio anubis
Macaca mulatta
Macaca nemestrina
Macaca fascicularis
Mandrillus leucophaeus
Rhinopithecus bieti
Rhinopithecus roxellana
Colobus angolensis palliatus
Piliocolobus tephrosceles
Monodelphis domestica
0.02
Euarchontoglires
Laurasiatheria

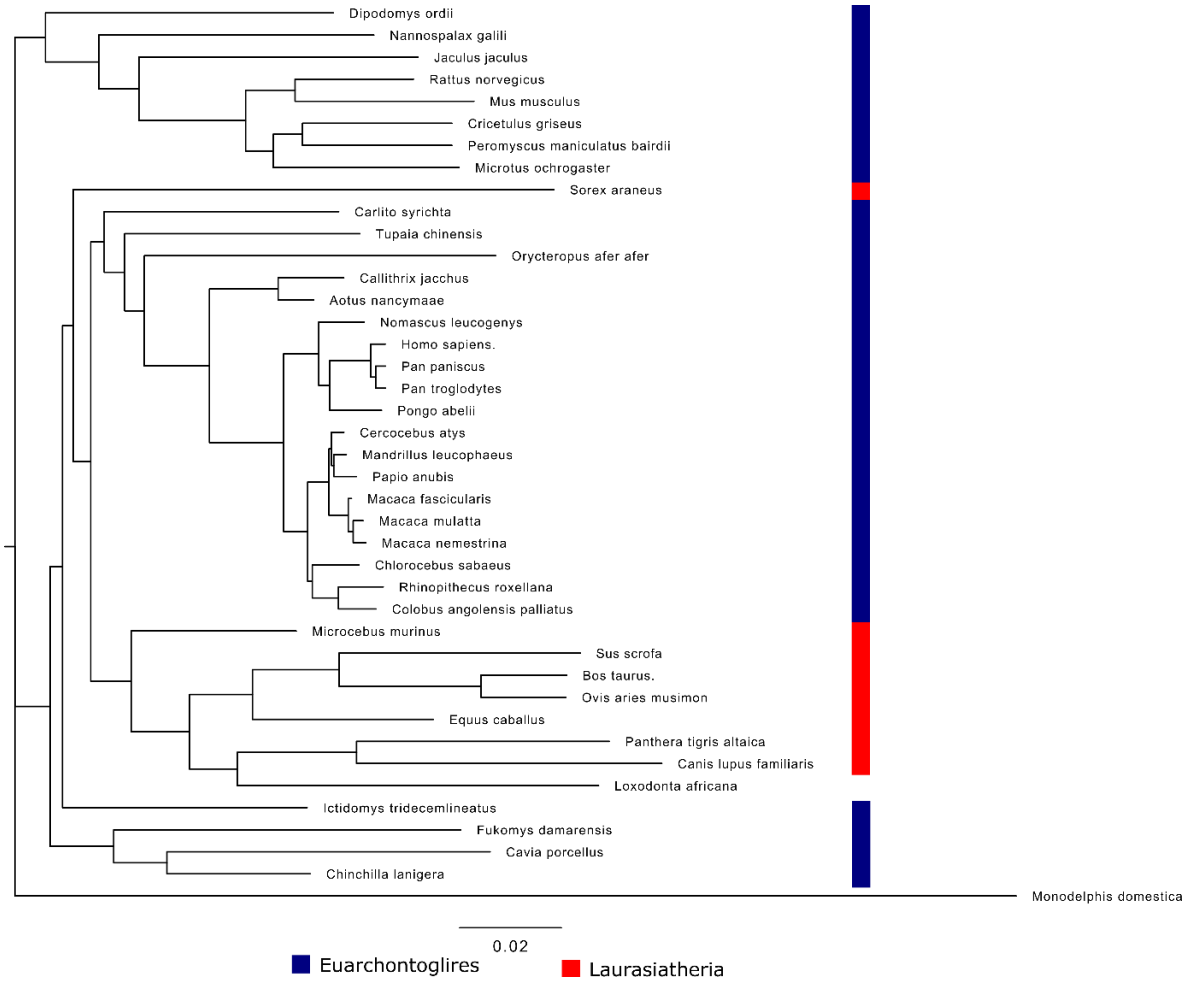
**C** IIb Motor


Trichechus manatus latirostris
Loxodonta africana
Ochotona princeps
Tupaia chinensis
Dipodomys ordii
Jaculus jaculus
Mus musculus
Rattus norvegicus.
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Nannospalax galili
Heterocephalus glaber
Fukomys damarensis
Cavia porcellus
Octodon degus
Chinchilla lanigera
Ictidomys tridecemlineatus
Galeopterus variegatus
Microcebus murinus
Dasypus novemcinctus
Callithrix jacchus
Saimiri boliviensis boliviensis
Rhinopithecus roxellana
Chlorocebus sabaeus
Colobus angolensis palliatus
Cercocebus atys
Mandrillus leucophaeus
Papio anubis
Macaca nemestrina
Macaca mulatta
Macaca fascicularis
Nomascus leucogenys
Homo sapiens.
Pan paniscus
Pan troglodytes
Gorilla gorilla gorilla
Carlito syrichta
Sorex araneus
Erinaceus europaeus
Pteropus alecto
Pteropus vampyrus
Myotis davidii
Myotis lucifugus
Panthera tigris altaica
Felis catus
Odobenus rosmarus divergens
Ailuropoda melanoleuca
Canis lupus familiaris
Bubalus bubalis
Monodelphis domestica
0.02
Euarchontoglires
Laurasiatheria

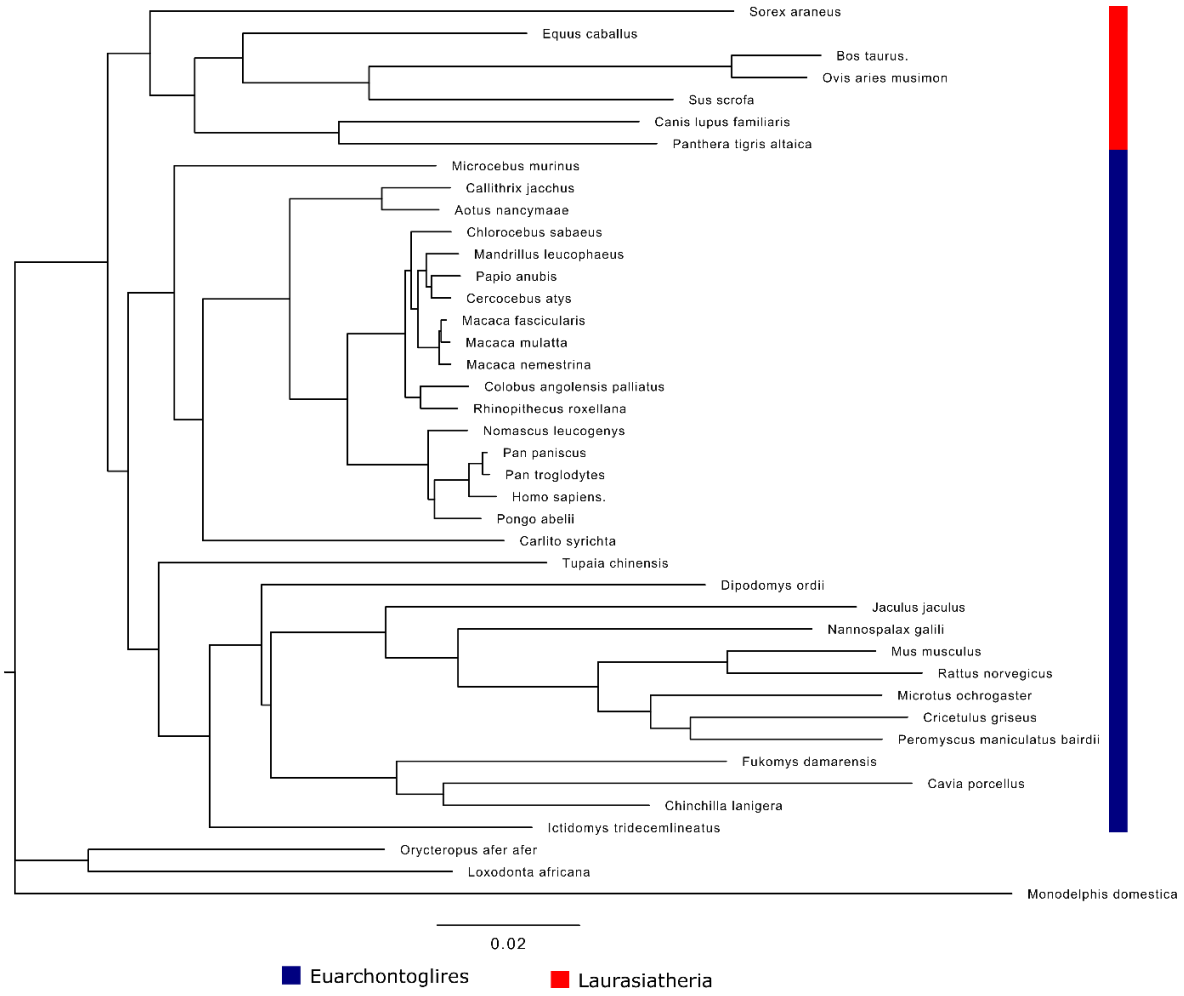
**D** IIb Tail


Dasypus novemcinctus
Loxodonta africana
Trichechus manatus latirostris
Tupaia chinensis
Ictidomys tridecemlineatus
Octodon degus
Chinchilla lanigera
Cavia porcellus
Heterocephalus glaber
Fukomys damarensis
Dipodomys ordii
Jaculus jaculus
Nannospalax galili
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus.
Ochotona princeps
Microcebus murinus
Carlito syrichta
Callithrix jacchus
Saimiri boliviensis boliviensis
Rhinopithecus roxellana
Colobus angolensis palliatus
Chlorocebus sabaeus
Papio anubis
Cercocebus atys
Macaca nemestrina
Macaca fascicularis
Macaca mulatta
Mandrillus leucophaeus
Nomascus leucogenys
Pan troglodytes
Pan paniscus
Homo sapiens.
Gorilla gorilla gorilla
Galeopterus variegatus
Sorex araneus
Erinaceus europaeus
Pteropus vampyrus
Pteropus alecto
Myotis davidii
Myotis lucifugus
Bubalus bubalis
Panthera tigris altaica
Felis catus
Canis lupus familiaris
Odobenus rosmarus divergens
Ailuropoda melanoleuca
Monodelphis domestica
0.02
Euarchontoglires
Laurasiatheria

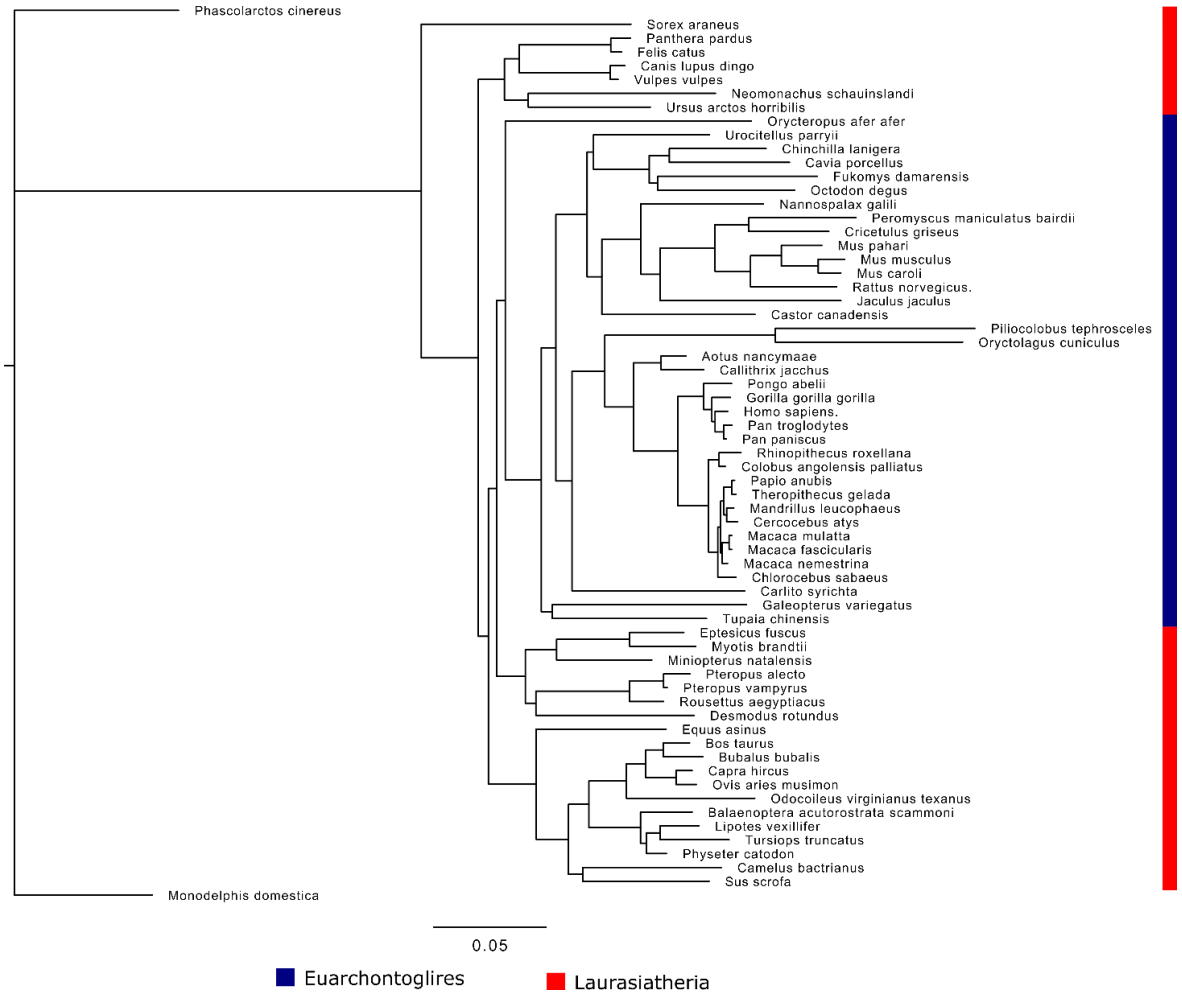
**E** IIx Motor


Dipodomys ordii
Nannospalax galili
Jaculus jaculus
Rattus norvegicus
Mus musculus
Cricetulus griseus
Peromyscus maniculatus bairdii
Microtus ochrogaster
Sorex araneus
Carlito syrichta
Tupaia chinensis
Orycteropus afer afer
Callithrix jacchus
Aotus nancymaae
Nomascus leucogenys
Homo sapiens.
Pan paniscus
Pan troglodytes
Pongo abelii
Cercocebus atys
Mandrillus leucophaeus
Papio anubis
Macaca fascicularis
Macaca mulatta
Macaca nemestrina
Chlorocebus sabaeus
Rhinopithecus roxellana
Colobus angolensis palliatus
Microcebus murinus
Sus scrofa
Bos taurus.
Ovis aries musimon
Equus caballus
Panthera tigris altaica
Canis lupus familiaris
Loxodonta africana
Ictidomys tridecemlineatus
Fukomys damarensis
Cavia porcellus
Chinchilla lanigera
Monodelphis domestica
0.02
Euarchontoglires
Laurasiatheria

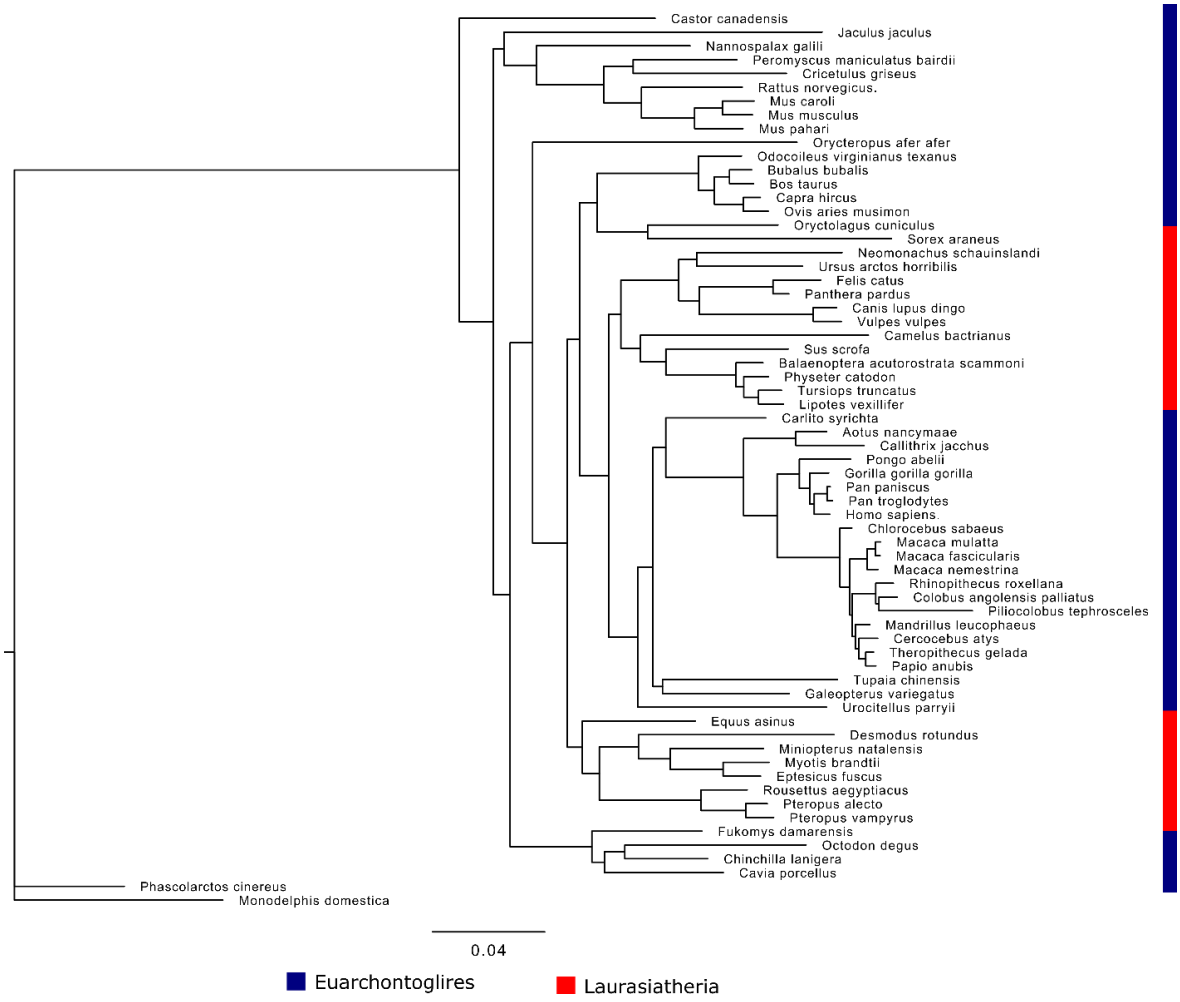
**F** Ilx Tail


Sorex araneus
Equus caballus
Bos taurus.
Ovis aries musimon
Sus scrofa
Canis lupus familiaris
Panthera tigris altaica
Microcebus murinus
Callithrix jacchus
Aotus nancymaae
Chlorocebus sabaeus
Mandrillus leucophaeus
Papio anubis
Cercocebus atys
Macaca fascicularis
Macaca mulatta
Macaca nemestrina
Colobus angolensis palliatus
Rhinopithecus roxellana
Nomascus leucogenys
Pan paniscus
Pan troglodytes
Homo sapiens.
Pongo abelii
Carlito syrichta
Tupaia chinensis
Dipodomys ordii
Jaculus jaculus
Nannospalax galili
Mus musculus
Rattus norvegicus
Microtus ochrogaster
Cricetulus griseus
Peromyscus maniculatus bairdii
Fukomys damarensis
Cavia porcellus
Chinchilla lanigera
Ictidomys tridecemlineatus
Orycteropus afer afer
Loxodonta africana
Monodelphis domestica
0.02
Euarchontoglires
Laurasiatheria

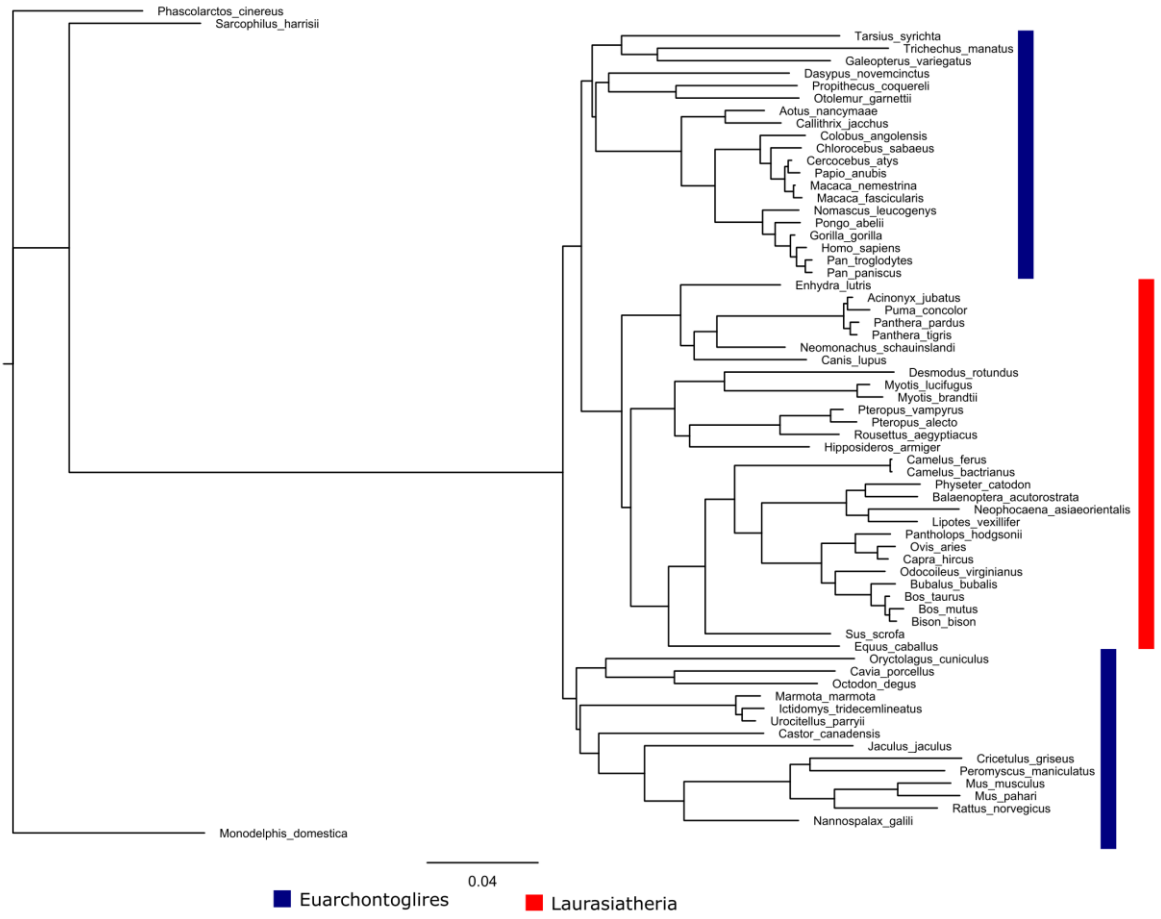
**G** α Motor


Phascolarctos cinereus
Sorex araneus
Panthera pardus
Felis catus
Canis lupus dingo
Vulpes vulpes
Neomonachus schauinslandi
Ursus arctos horribilis
Orycteropus afer afer
Urocitellus parryii
Chinchilla lanigera
Cavia porcellus
Fukomys damarensis
Octodon degus
Nannospalax galili
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus pahari
Mus musculus
Mus caroli
Rattus norvegicus.
Jaculus jaculus
Castor canadensis
Piliocolobus tephrosceles
Oryctolagus cuniculus
Aotus nancymaae
Callithrix jacchus
Pongo abelii
Gorilla gorilla gorilla
Homo sapiens.
Pan troglodytes
Pan paniscus
Rhinopithecus roxellana
Colobus angolensis palliatus
Papio anubis
Theropithecus gelada
Mandrillus leucophaeus
Cercocebus atys
Macaca mulatta
Macaca fascicularis
Macaca nemestrina
Chlorocebus sabaeus
Carlito syrichta
Galeopterus variegatus
Tupaia chinensis
Eptesicus fuscus
Myotis brandtii
Miniopterus natalensis
Pteropus alecto
Pteropus vampyrus
Rousettus aegyptiacus
Desmodus rotundus
Equus asinus
Bos taurus
Bubalus bubalis
Capra hircus
Ovis aries musimon
Odocoileus virginianus texanus
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Tursiops truncatus
Physeter catodon
Camelus bactrianus
Sus scrofa
Monodelphis domestica
0.05
Euarchontoglires
Laurasiatheria

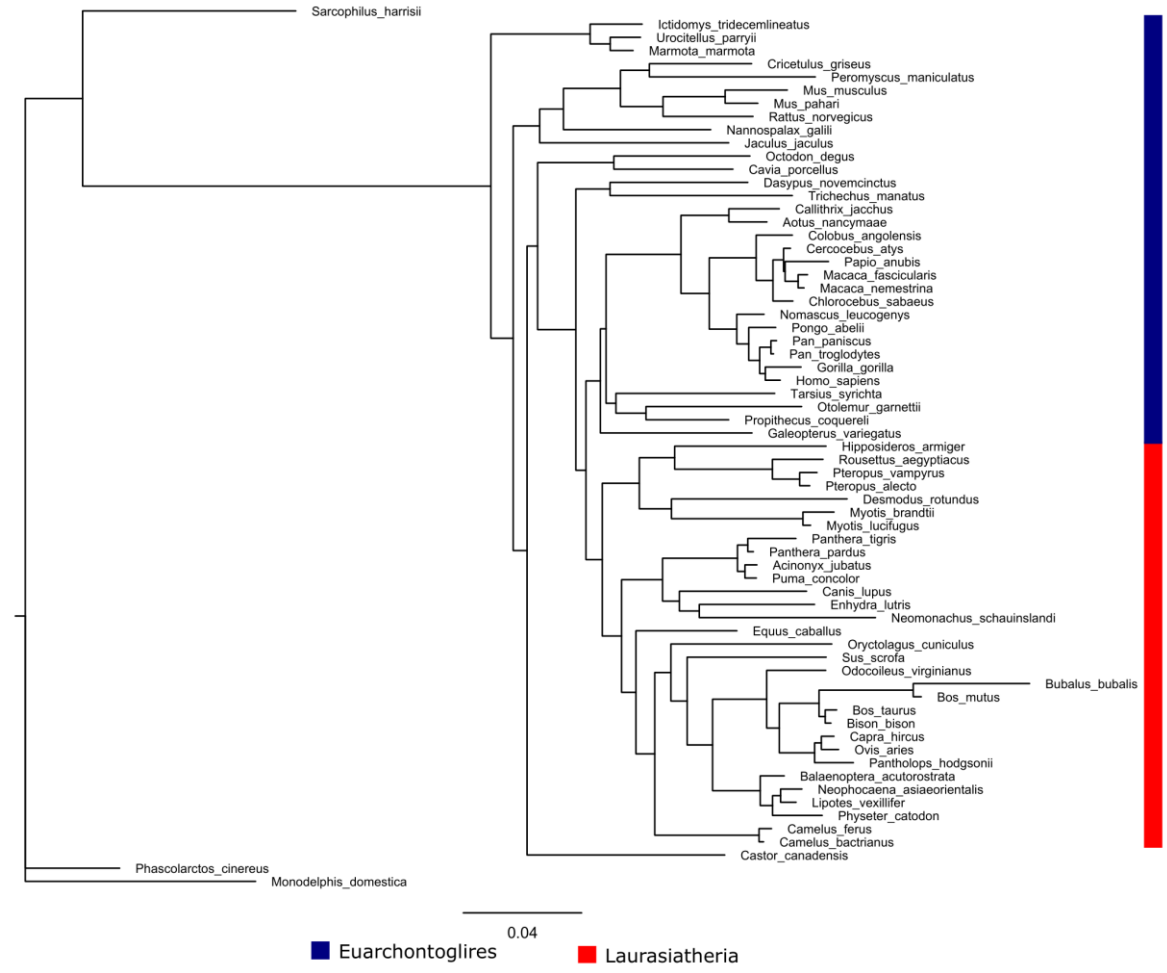
**H** α Tail


Castor canadensis
Jaculus jaculus
Nannospalax galili
Peromyscus maniculatus bairdii
Cricetulus griseus
Rattus norvegicus.
Mus caroli
Mus musculus
Mus pahari
Orycteropus afer afer
Odocoileus virginianus texanus
Bubalus bubalis
Bos taurus
Capra hircus
Ovis aries musimon
Oryctolagus cuniculus
Sorex araneus
Neomonachus schauinslandi
Ursus arctos horribilis
Felis catus
Panthera pardus
Canis lupus dingo
Vulpes vulpes
Camelus bactrianus
Sus scrofa
Balaenoptera acutorostrata scammoni
Physeter catodon
Tursiops truncatus
Lipotes vexillifer
Carlito syrichta
Aotus nancymaae
Callithrix jacchus
Pongo abelii
Gorilla gorilla gorilla
Pan paniscus
Pan troglodytes
Homo sapiens.
Chlorocebus sabaeus
Macaca mulatta
Macaca fascicularis
Macaca nemestrina
Rhinopithecus roxellana
Colobus angolensis palliatus
Piliocolobus tephrosceles
Mandrillus leucophaeus
Cercocebus atys
Theropithecus gelada
Papio anubis
Tupaia chinensis
Galeopterus variegatus
Urocitellus parryii
Equus asinus
Desmodus rotundus
Miniopterus natalensis
Myotis brandtii
Eptesicus fuscus
Rousettus aegyptiacus
Pteropus alecto
Pteropus vampyrus
Fukomys damarensis
Octodon degus
Chinchilla lanigera
Cavia porcellus
Phascolarctos cinereus
Monodelphis domestica
0.04
Euarchontoglires
Laurasiatheria

## I β-Motor


Phascolarctos_cinereus
Sarcophilus_harrisii
Monodelphis_domestica
Tarsius_syrichta
Trichechus_manatus
Galeopterus_variegatus
Dasypus_novemcinctus
Propithecus_coquereli
Otolemur_garnettii
Aotus_nancymaae
Callithrix_jacchus
Colobus_angolensis
Chlorocebus_sabaeus
Cercocebus_atys
Papio_anubis
Macaca_nemestrina
Macaca_fascicularis
Nomascus_leucogenys
Pongo_abelii
Gorilla_gorilla
Homo_sapiens
Pan_troglodytes
Pan_paniscus
Enhydra_lutris
Acinonyx_jubatus
Puma_concolor
Panthera_pardus
Panthera_tigris
Neomonachus_schauinslandi
Canis_lupus
Desmodus_rotundus
Myotis_lucifugus
Myotis_brandtii
Pteropus_vampyrus
Pteropus_alecto
Rousettus_aegyptiacus
Hipposideros_armiger
Camelus_ferus
Camelus_bactrianus
Physeter_catodon
Balaenoptera_acutorostrata
Neophocaena_asiaeorientalis
Lipotes_vexillifer
Pantholops_hodgsonii
Ovis_aries
Capra_hircus
Odocoileus_virginianus
Bubalus_bubalis
Bos_taurus
Bos_mutus
Bison_bison
Sus_scrofa
Equus_caballus
Oryctolagus_cuniculus
Cavia_porcellus
Octodon_degus
Marmota_marmota
Ictidomys_tridecemlineatus
Urocitellus_parryii
Castor_canadensis
Jaculus_jaculus
Cricetulus_griseus
Peromyscus_maniculatus
Mus_musculus
Mus_pahari
Rattus_norvegicus
Nannospalax_galili
0.04
Euarchontoglires
Laurasiatheria

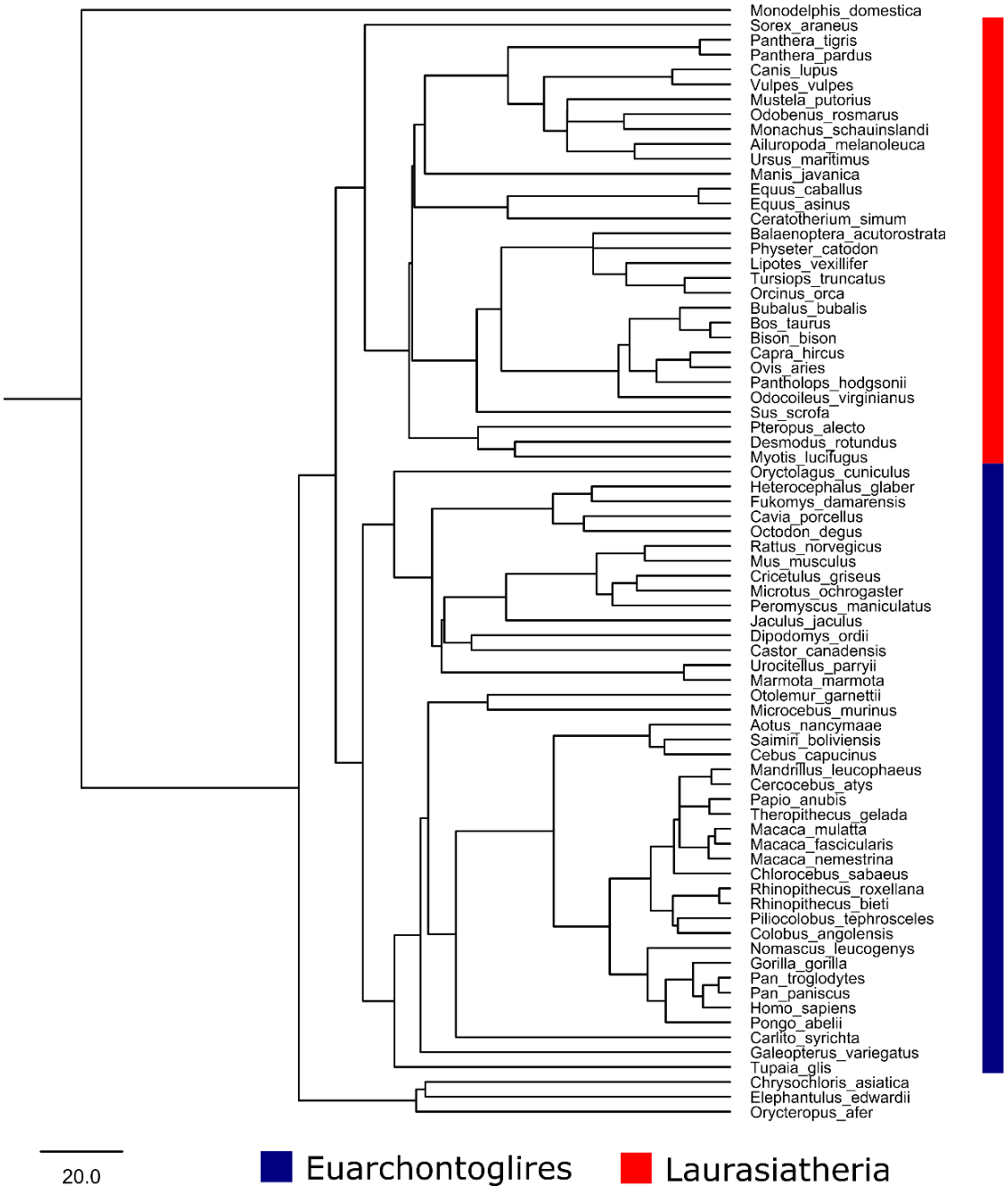
## J β-Tail


Sarcophilus_harrisii
Ictidomys_tridecemlineatus
Urocitellus_parryii
Marmota_marmota
Cricetulus_griseus
Peromyscus_maniculatus
Mus_musculus
Mus_pahari
Rattus_norvegicus
Nannospalax_galili
Jaculus_jaculus
Octodon_degus
Cavia_porcellus
Dasypus_novemcinctus
Trichechus_manatus
Callithrix_jacchus
Aotus_nancymaae
Colobus_angolensis
Cercocebus_atys
Papio_anubis
Macaca_fascicularis
Macaca_nemestrina
Chlorocebus_sabaeus
Nomascus_leucogenys
Pongo_abelii
Pan_paniscus
Pan_troglodytes
Gorilla_gorilla
Homo_sapiens
Tarsius_syrichta
Otolemur_garnettii
Propithecus_coquereli
Galeopterus_variegatus
Hipposideros_armiger
Rousettus_aegyptiacus
Pteropus_vampyrus
Pteropus_alecto
Desmodus_rotundus
Myotis_brandtii
Myotis_lucifugus
Panthera_tigris
Panthera_pardus
Acinonyx_jubatus
Puma_concolor
Canis_lupus
Enhydra_lutris
Neomonachus_schauinslandi
Equus_caballus
Oryctolagus_cuniculus
Sus_scrofa
Odocoileus_virginianus
Bubalus_bubalis
Bos_mutus
Bos_taurus
Bison_bison
Capra_hircus
Ovis_aries
Pantholops_hodgsonii
Balaenoptera_acutorostrata
Neophocaena_asiaeorientalis
Lipotes_vexillifer
Physeter_catodon
Camelus_ferus
Camelus_bactrianus
Castor_canadensis
Phascolarctos_cinereus
Monodelphis_domestica
0.04
Euarchontoglires
Laurasiatheria

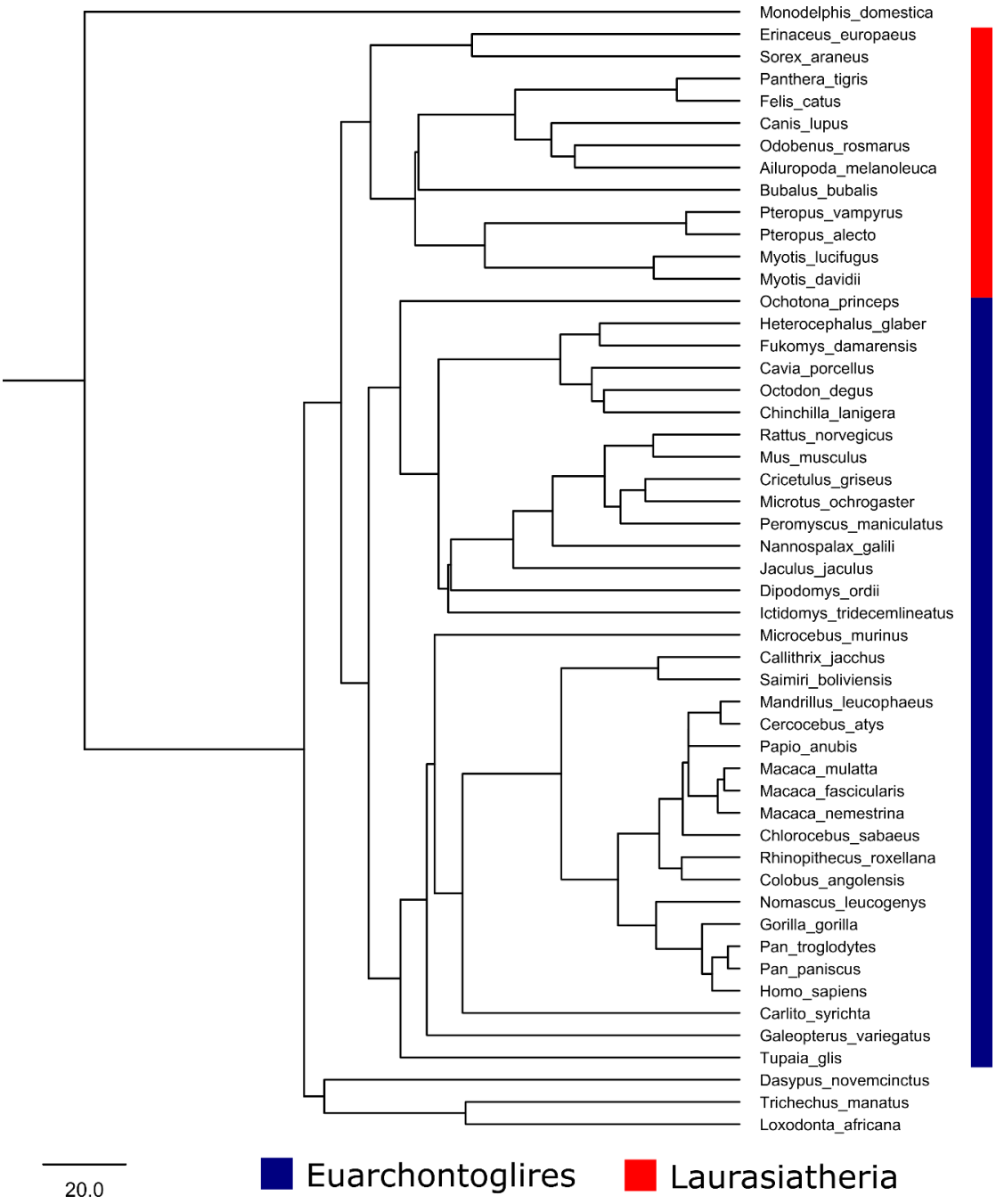
**Supplementary Figure 3.8. Phylogenetic (Species) Trees for sarcomeric isoforms.**

Phylogenetic trees for IIa, IIb, IIx, α and β species generated from TimeTree.

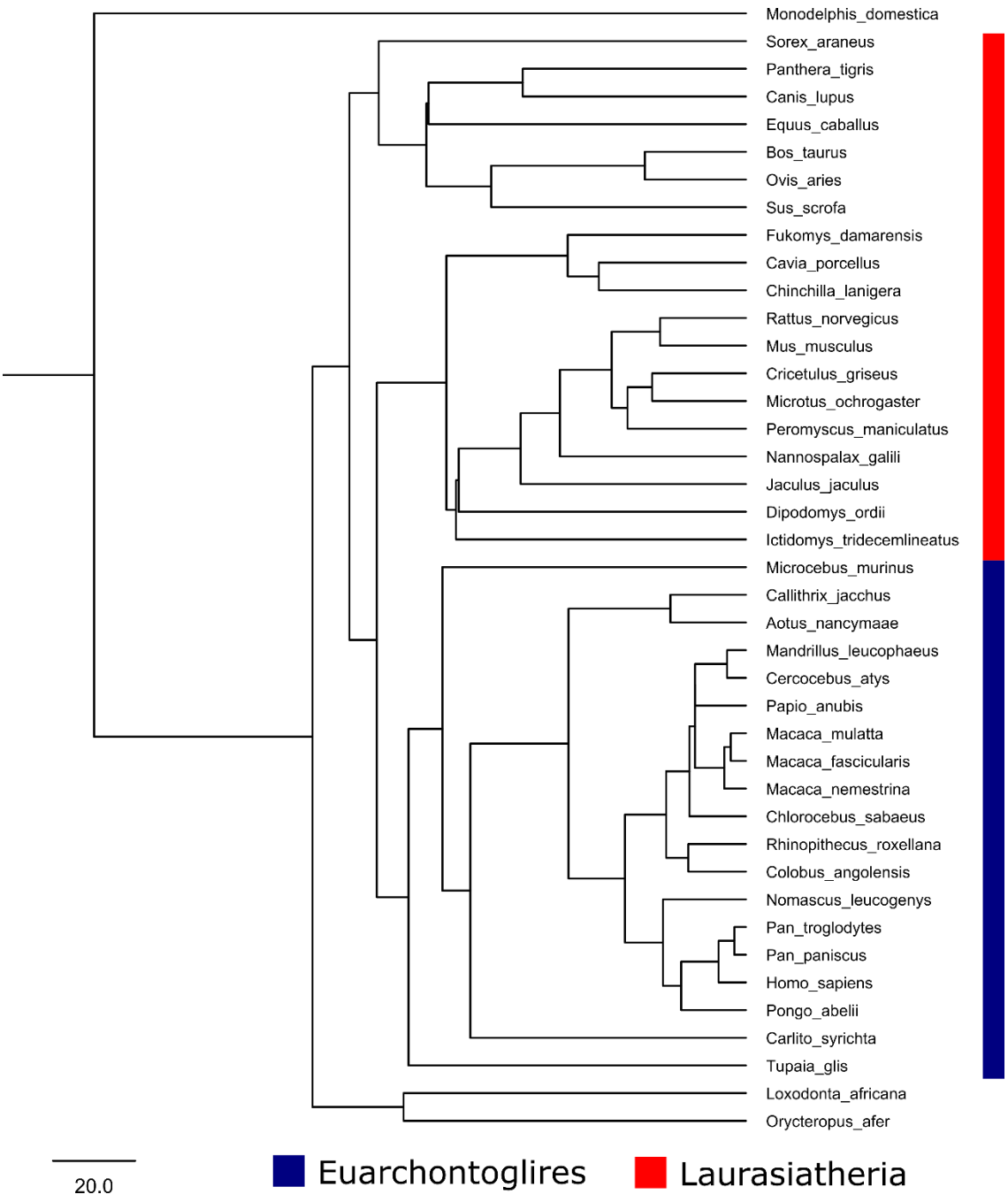
**A IIa**



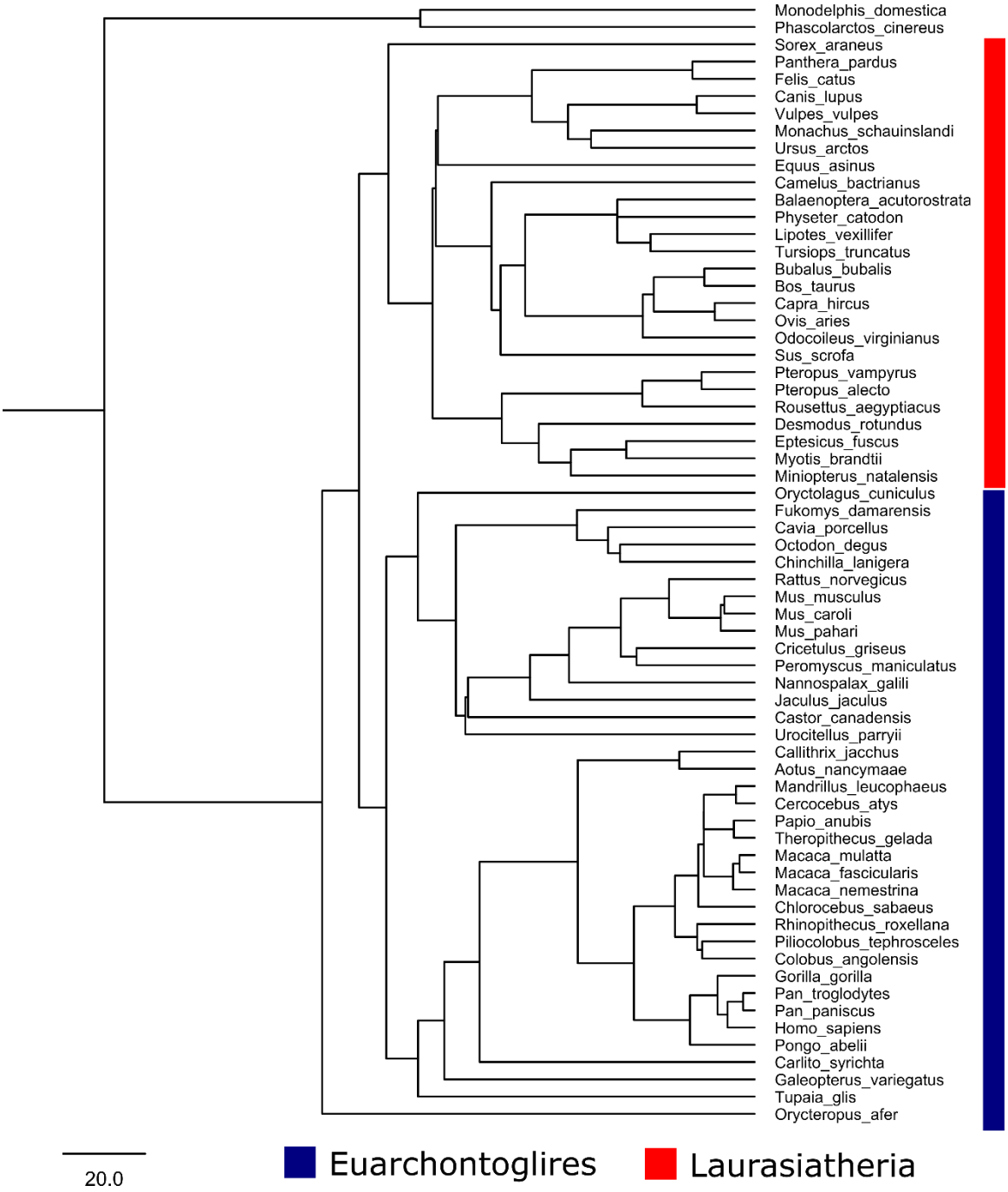
**B IIb**


Monodelphis_domestica
Erinaceus_europaeus
Sorex_araneus
Panthera_tigris
Felis_catus
Canis_lupus
Odobenus_rosmarus
Ailuropoda_melanoleuca
Bubalus_bubalis
Pteropus_vampyrus
Pteropus_alecto
Myotis_lucifugus
Myotis_davidii
Ochotona_princeps
Heterocephalus_glaber
Fukomys_damarensis
Cavia_porcellus
Octodon_degus
Chinchilla_lanigera
Rattus_norvegicus
Mus_musculus
Cricetulus_griseus
Microtus_ochrogaster
Peromyscus_maniculatus
Nannospalax_galili
Jaculus_jaculus
Dipodomys_ordii
Ictidomys_tridecemlineatus
Microcebus_murinus
Callithrix_jacchus
Saimiri_boliviensis
Mandrillus_leucophaeus
Cercocebus_atys
Papio_anubis
Macaca_mulatta
Macaca_fascicularis
Macaca_nemestrina
Chlorocebus_sabaeus
Rhinopithecus_roxellana
Colobus_angolensis
Nomascus_leucogenys
Gorilla_gorilla
Pan_troglodytes
Pan_paniscus
Homo_sapiens
Carlito_syrichta
Galeopterus_variegatus
Tupaia_glis
Dasypus_novemcinctus
Trichechus_manatus
Loxodonta_africana
20.0
Euarchontoglires
Laurasiatheria

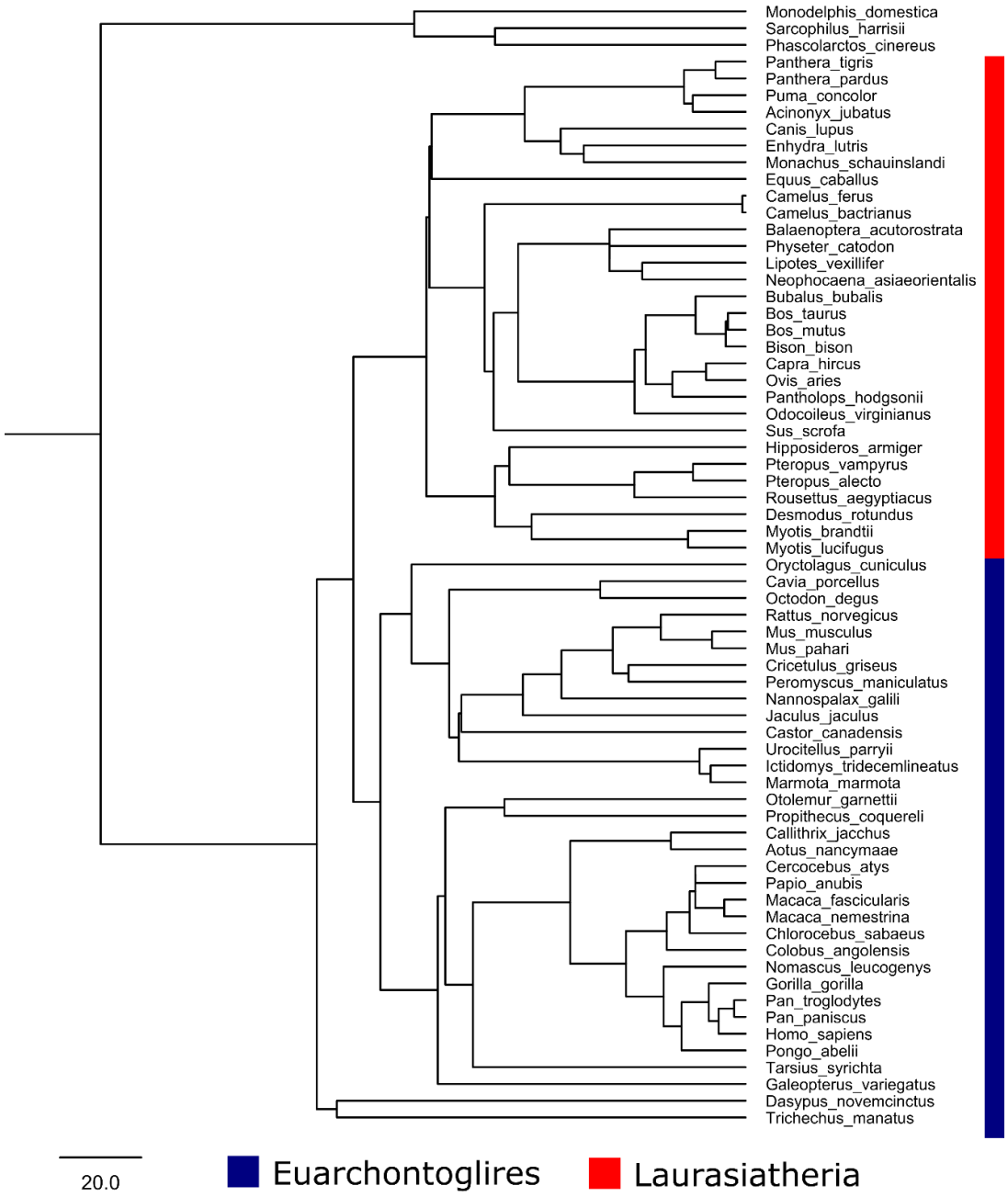
**C IIx**


Monodelphis_domestica
Sorex_araneus
Panthera_tigris
Canis_lupus
Equus_caballus
Bos_taurus
Ovis_aries
Sus_scrofa
Fukomys_damarensis
Cavia_porcellus
Chinchilla_lanigera
Rattus_norvegicus
Mus_musculus
Cricetulus_griseus
Microtus_ochrogaster
Peromyscus_maniculatus
Nannospalax_galili
Jaculus_jaculus
Dipodomys_ordii
Ictidomys_tridecemlineatus
Microcebus_murinus
Callithrix_jacchus
Aotus_nancymaae
Mandrillus_leucophaeus
Cercocebus_atys
Papio_anubis
Macaca_mulatta
Macaca_fascicularis
Macaca_nemestrina
Chlorocebus_sabaeus
Rhinopithecus_roxellana
Colobus_angolensis
Nomascus_leucogenys
Pan_troglodytes
Pan_paniscus
Homo_sapiens
Pongo_abelii
Carlito_syrichta
Tupaia_glis
Loxodonta_africana
Orycteropus_afer
20.0
Euarchontoglires
Laurasiatheria

**D α**


Monodelphis_domestica
Phascolarctos_cinereus
Sorex_araneus
Panthera_pardus
Felis_catus
Canis_lupus
Vulpes_vulpes
Monachus_schauinslandi
Ursus_arctos
Equus_asinus
Camelus_bactrianus
Balaenoptera_acutorostrata
Physeter_catodon
Lipotes_vexillifer
Tursiops_truncatus
Bubalus_bubalis
Bos_taurus
Capra_hircus
Ovis_aries
Odocoileus_virginianus
Sus_scrofa
Pteropus_vampyrus
Pteropus_alecto
Rousettus_aegyptiacus
Desmodus_rotundus
Eptesicus_fuscus
Myotis_brandtii
Miniopterus_natalensis
Oryctolagus_cuniculus
Fukomys_damarensis
Cavia_porcellus
Octodon_degus
Chinchilla_lanigera
Rattus_norvegicus
Mus_musculus
Mus_caroli
Mus_pahari
Cricetulus_griseus
Peromyscus_maniculatus
Nannospalax_galili
Jaculus_jaculus
Castor_canadensis
Urocitellus_parryii
Callithrix_jacchus
Aotus_nancymaae
Mandrillus_leucophaeus
Cercocebus_atys
Papio_anubis
Theropithecus_gelada
Macaca_mulatta
Macaca_fascicularis
Macaca_nemestrina
Chlorocebus_sabaeus
Rhinopithecus_roxellana
Piliocolobus_tephrosceles
Colobus_angolensis
Gorilla_gorilla
Pan_troglodytes
Pan_paniscus
Homo_sapiens
Pongo_abelii
Carlito_syrichta
Galeopterus_variegatus
Tupaia_glis
Orycteropus_afer
20.0
Euarchontoglires
Laurasiatheria

**E β**


Monodelphis_domestica
Sarcophilus_harrisii
Phascolarctos_cinereus
Panthera_tigris
Panthera_pardus
Puma_concolor
Acinonyx_jubatus
Canis_lupus
Enhydra_lutris
Monachus_schauinslandi
Equus_caballus
Camelus_ferus
Camelus_bactrianus
Balaenoptera_acutorostrata
Physeter_catodon
Lipotes_vexillifer
Neophocaena_asiaeorientalis
Bubalus_bubalis
Bos_taurus
Bos_mutus
Bison_bison
Capra_hircus
Ovis_aries
Pantholops_hodgsonii
Odocoileus_virginianus
Sus_scrofa
Hipposideros_armiger
Pteropus_vampyrus
Pteropus_alecto
Rousettus_aegyptiacus
Desmodus_rotundus
Myotis_brandtii
Myotis_lucifugus
Oryctolagus_cuniculus
Cavia_porcellus
Octodon_degus
Rattus_norvegicus
Mus_musculus
Mus_pahari
Cricetulus_griseus
Peromyscus_maniculatus
Nannospalax_galili
Jaculus_jaculus
Castor_canadensis
Urocitellus_parryii
Ictidomys_tridecemlineatus
Marmota_marmota
Otolemur_garnettii
Propithecus_coquereli
Callithrix_jacchus
Aotus_nancymaae
Cercocebus_atys
Papio_anubis
Macaca_fascicularis
Macaca_nemestrina
Chlorocebus_sabaeus
Colobus_angolensis
Nomascus_leucogenys
Gorilla_gorilla
Pan_troglodytes
Pan_paniscus
Homo_sapiens
Pongo_abelii
Tarsius_syrichta
Galeopterus_variegatus
Dasypus_novemcinctus
Trichechus_manatus
20.0
Euarchontoglires
Laurasiatheria

**Supplementary Table 3.1. Mass bootstrapping for each myosin isoform.** The mass of species for each isoform was randomised 1000 times at 0.8-1.2 times the mass value for each given species. This table shows the range of values produced from this analysis. The values for the motors are shown in A, and values for the tails in B.

**A Motor domain**

| *Isoform* | Error Range | Slope Range | $R^2$- Range | Error STD-DEV | Slope STD-DEV | $R^2$- STD-DEV |
|---|---|---|---|---|---|---|
| ***IIb*** | 0.181 - 0.216 | -0.053 | 0.448 - 0.497 | 0.01 | 0.01 | 0.01 |
| ***IIa*** | 0.095 - 0.101 | -0.025 | 0.426 - 0.461 | 0 | 0 | 0 |
| ***IIx*** | 0.083 - 0.138 | -0.068 | 0.796 - 0.869 | 0.01 | 0.01 | 0.01 |
| ***α*** | 0.267 - 0.326 | -0.066 | 0.339 - 0.42 | 0.01 | 0.01 | 0.01 |
| ***β*** | 0.098 - 0.112 | -0.031 | 0.618 - 0.658 | 0 | 0.01 | 0.01 |
| ***EXOC*** | 0.099 - 0.104 | -0.023 | 0.165 - 0.187 | 0 | 0 | 0 |
| ***slowT*** | 0.058 - 0.061 | -0.011 | 0.039 - 0.05 | 0 | 0 | 0 |
| ***EMB*** | 0.034 - 0.036 | -0.007 | 0.118 - 0.136 | 0 | 0 | 0 |
| ***PERI*** | 0.105 - 0.122 | -0.02 | 0.005 - 0.014 | 0 | 0 | 0 |
| ***NMA*** | 0.034 - 0.037 | -0.007 | 0.027 - 0.038 | 0 | 0 | 0 |
| ***NMB*** | 0.021 - 0.023 | 0.04 - 0.043 | 0.103 - 0.12 | 0 | 0 | 0 |
| ***SM*** | 0.066 - 0.073 | -0.015 | 0.479 - 0.514 | 0 | 0 | 0.01 |
| ***NMC*** | 0.107 - 0.117 | -0.022 | 0.095 - 0.113 | 0 | 0 | 0 |

**B Tail domain**

| *Isoform* | Error Range | Slope Range | $R^2$-Range | Error STD-DEV | Slope STD-DEV | $R^2$- STD-DEV |
|---|---|---|---|---|---|---|
| ***IIb*** | 0.112 - 0.126 | -0.041 | 0.558 - 0.607 | 0 | 0.01 | 0.01 |
| ***IIa*** | 0.104 - 0.112 | -0.025 | 0.337 - 0.371 | 0 | 0 | 0.01 |
| ***IIx*** | 0.076 - 0.127 | -0.037 | 0.831 - 0.862 | 0.01 | 0.01 | 0.01 |
| ***α*** | 0.093 - 0.113 | -0.02 | 0.338 - 0.368 | 0 | 0 | 0 |
| ***β*** | 0.156 - 0.194 | -0.037 | 0.059 - 0.121 | 0.01 | 0.01 | 0.01 |
| ***EXOC*** | 0.19 - 0.2 | -0.043 | 0.204 - 0.225 | 0 | 0.01 | 0 |
| ***slowT*** | 0.058 - 0.062 | -0.011 | 0.013 - 0.021 | 0 | 0 | 0 |
| ***EMB*** | 0.12 - 0.128 | -0.024 | 0.208 - 0.237 | 0 | 0 | 0 |
| ***PERI*** | 0.113 - 0.125 | -0.023 | 0.198 - 0.225 | 0 | 0 | 0 |
| ***NMA*** | 0.058 - 0.064 | -0.011 | 0.135 - 0.154 | 0 | 0 | 0 |
| ***NMB*** | 0.051 - 0.054 | -0.011 | 0.069 - 0.088 | 0 | 0 | 0 |
| ***SM*** | 0.099 - 0.105 | -0.018 | 0.127 - 0.144 | 0 | 0 | 0 |
| ***NMC*** | 0.34 - 0.367 | -0.063 | 0.189 - 0.217 | 0 | 0.01 | 0 |

**Supplementary Table 3.2. Accession numbers for isoforms considered.**

https://doi.org/10.1371/journal.pbio.3001248.s011

**Supplementary Movie 3.1. Movie to show the *in vitro* motility of the wild type and chimera beta-cardiac myosin.**

https://figshare.com/articles/media/In_vitro_motility_of_WT_and_chimeric_beta-myosin_/14763676

# Appendix 3. Chapter 4 Supplementary Material

**Figure S4.1. The BLOSUM scores for the amino acid substitutions present in the SDPs**. A graph is plotted that combines all of the proteins combined and one for each of the individual proteins that were analysed.

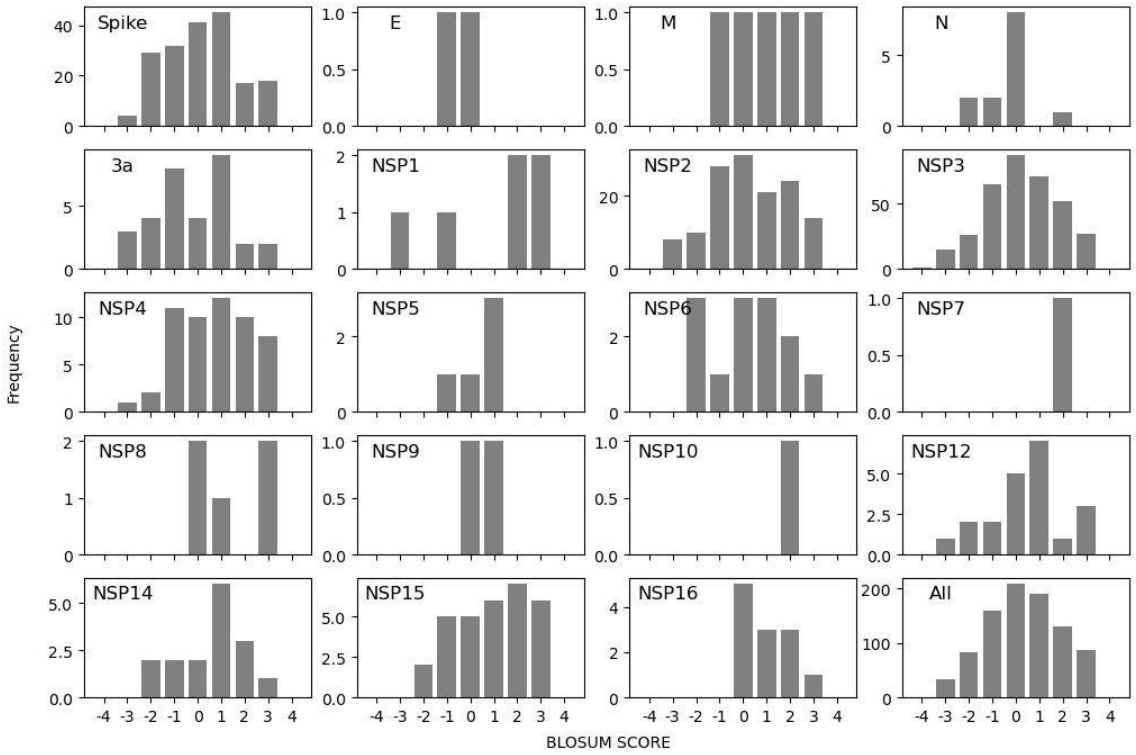
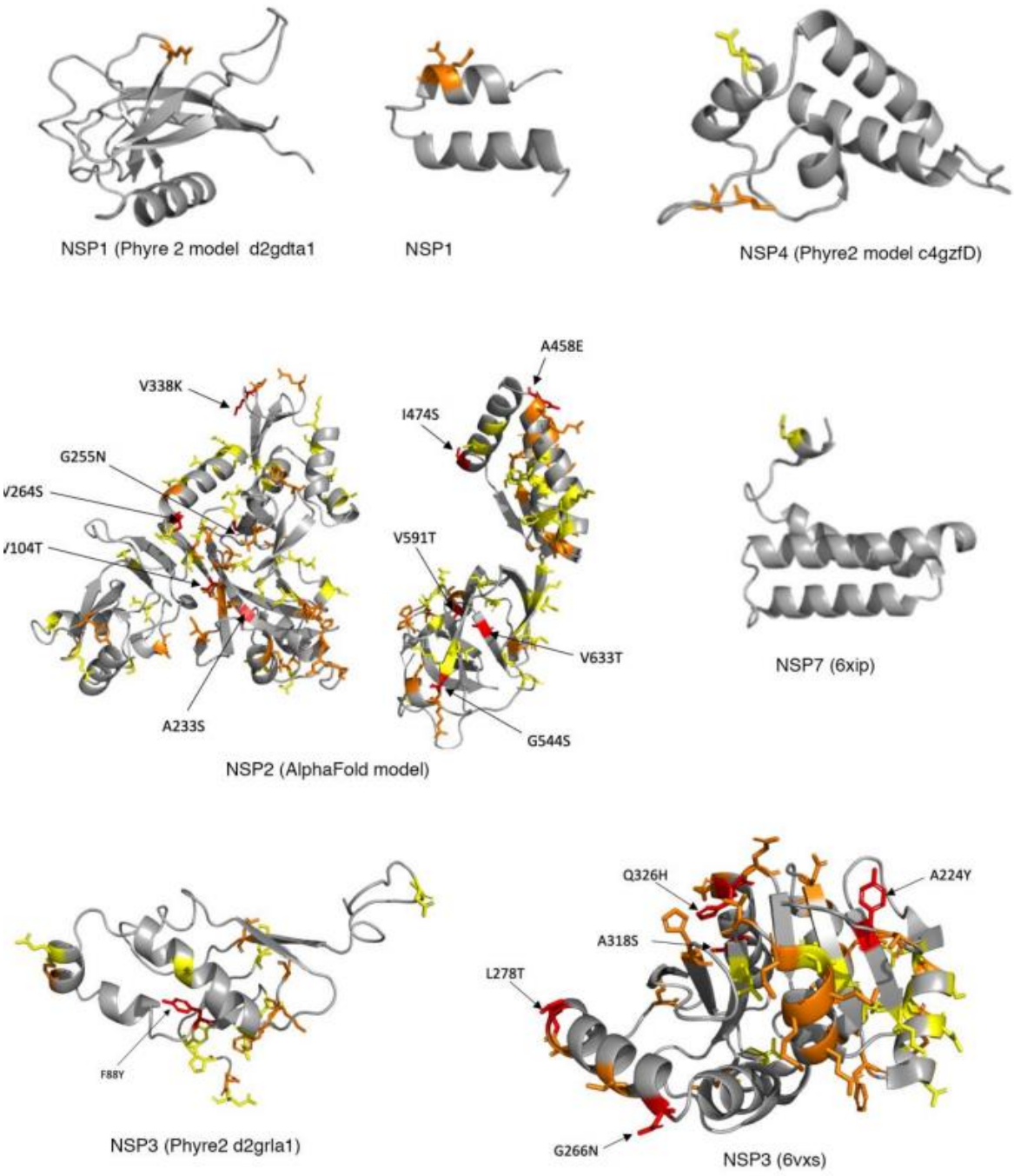
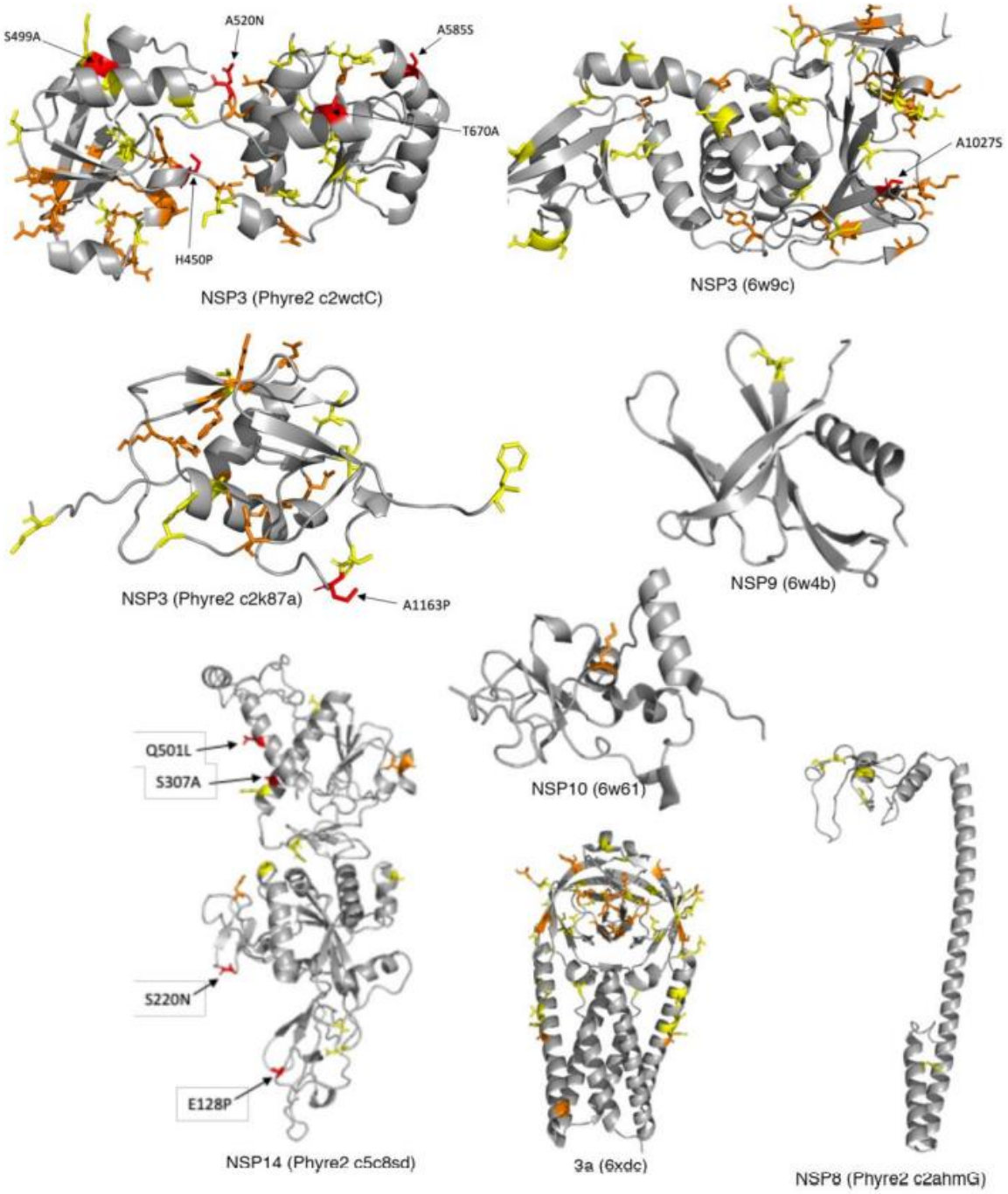
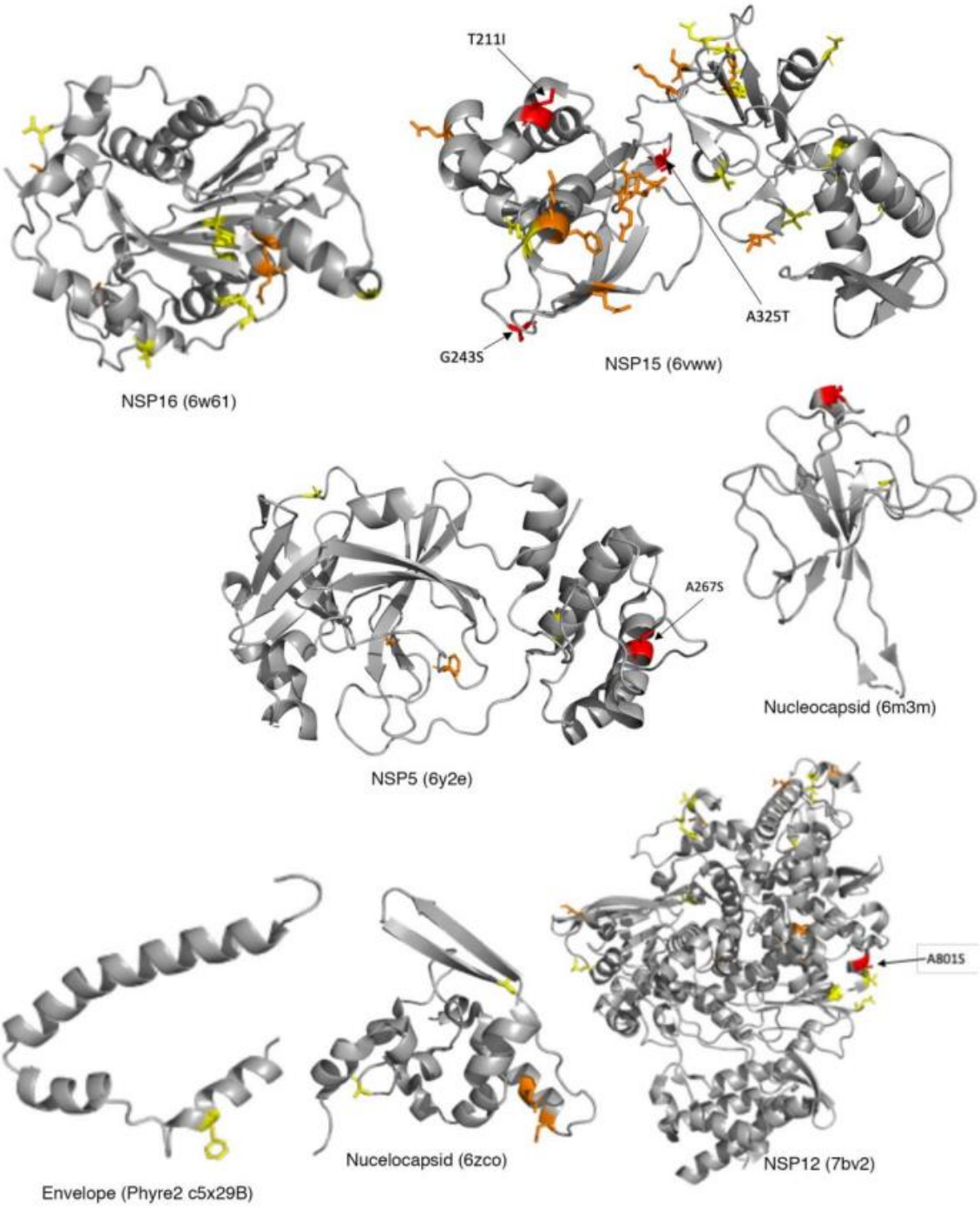


**Figure S4.2.** Overview of modelled DCPs. DCPs with likely functional effects are indicated by arrows and labelled. Structural model shown is indicated in brackets. DCPs likely to have an effect are coloured red; DCPs with a possible effect are shown in orange; and DCPs unlikely to have an effect are coloured yellow. Please refer to Table S6 for full details of structural analysis of each DCP.




S499A
A520N
A585S
T670A
H450P
NSP3 (Phyre2 c2wctC)
A1027S
NSP3 (6w9c)
NSP3 (Phyre2 c2k87a)
A1163P
NSP9 (6w4b)
NSP10 (6w61)
Q501L
S307A
S220N
E128P
NSP14 (Phyre2 c5c8sd)
3a (6xdc)
NSP8 (Phyre2 c2ahmG)


T211I
A325T
G243S
NSP15 (6vww)
NSP16 (6w61)
A267S
Nucleocapsid (6m3m)
NSP5 (6y2e)
A801S
Envelope (Phyre2 c5x29B)
Nucelocapsid (6zco)
NSP12 (7bv2)

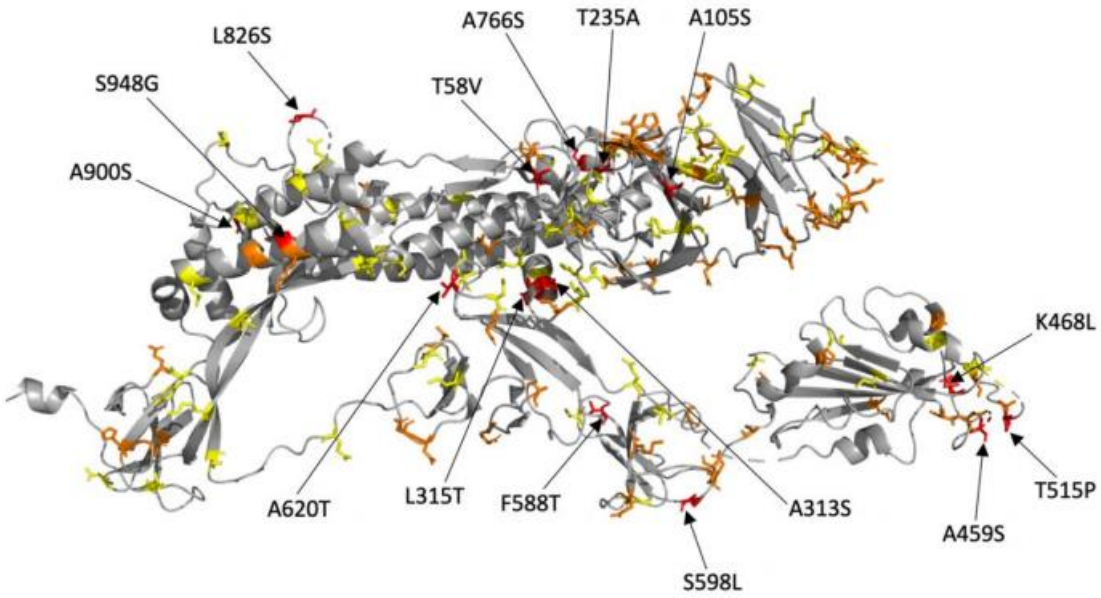

L826S
A766S
T235A
A105S
S948G
T58V
A900S
K468L
T515P
A620T
L315T
F588T
A313S
A459S
S598L

**Figure S4.3.** SARS-CoV-2 and SARS-CoV susceptibility of cell lines. A) Representative images showing MOI 0.01-infected cells immunostained for double-stranded RNA 48h post infection. B) Quantification of virus genomes by qPCR at different time points post infection (p.i.). Values are presented as means ± S.D. (n =3).

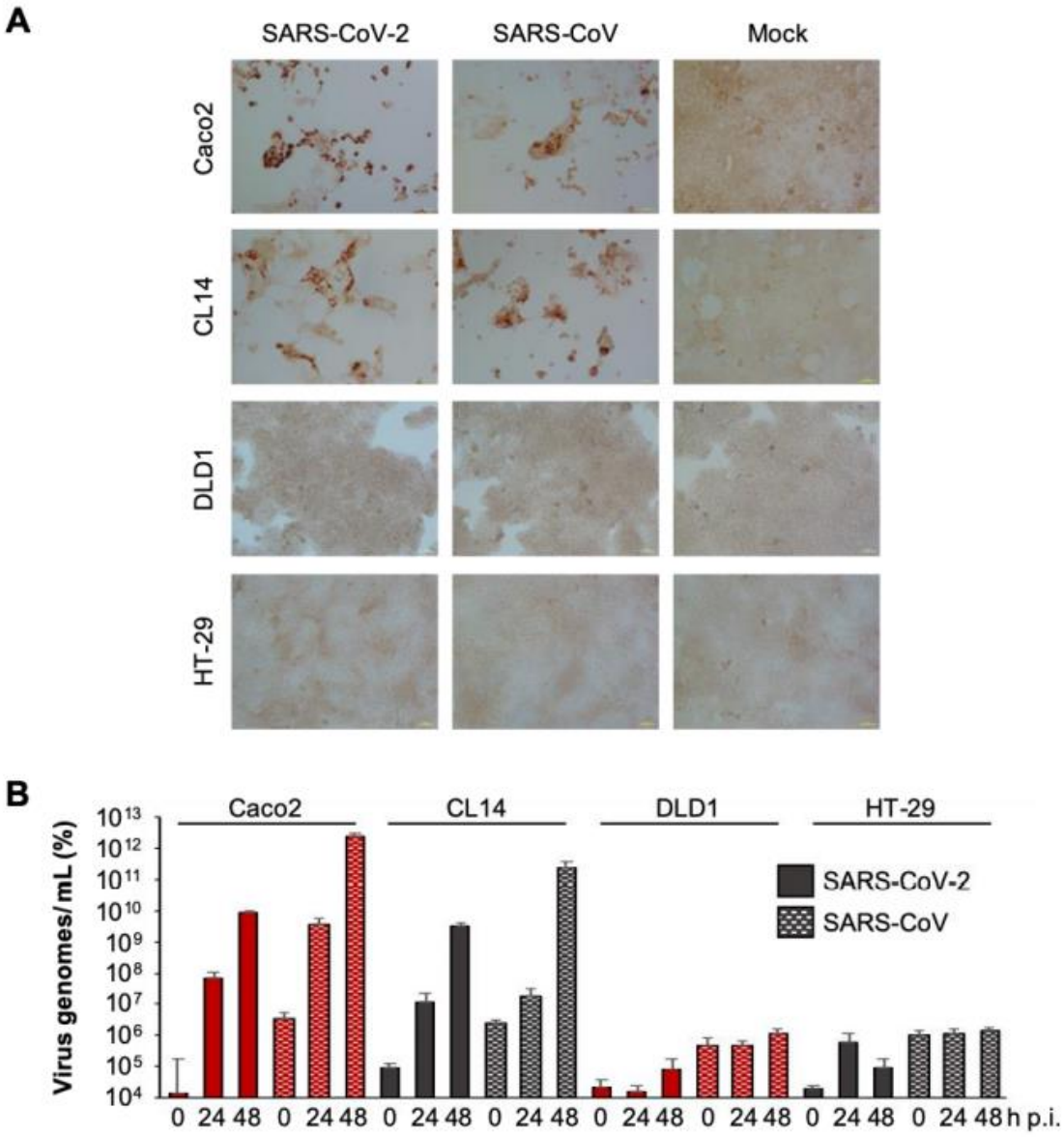


**Figure S4.4.** SARS-CoV-2 and SARS-CoV replication in 293 cells stably expressing ACE2 cells (293/ACE2). A) Immunostaining for double-stranded RNA (indicating virus replication) in SARS-CoV-2 and SARS-CoV (MOI 0.01)-infected 293/ACE2 cells 48h post infection. B) Quantification of virus genomes by qPCR in SARS-CoV-2 and SARS-CoV (MOI 0.01)-infected 293/ACE2 cells 48h post infection. Values are presented as means ± S.D. (n =3).

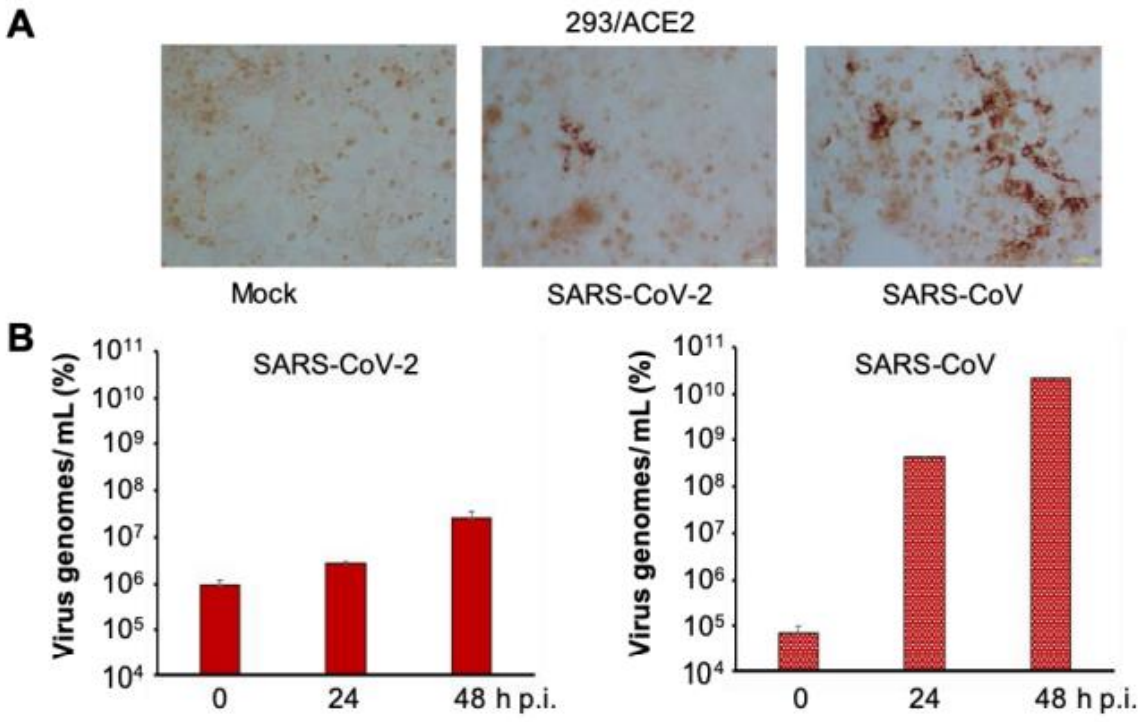


**Figure S4.5.** Uncropped Western blots for Figure 2D. 293/ACE2 cells served as positive control for ACE2. *Protein quantification

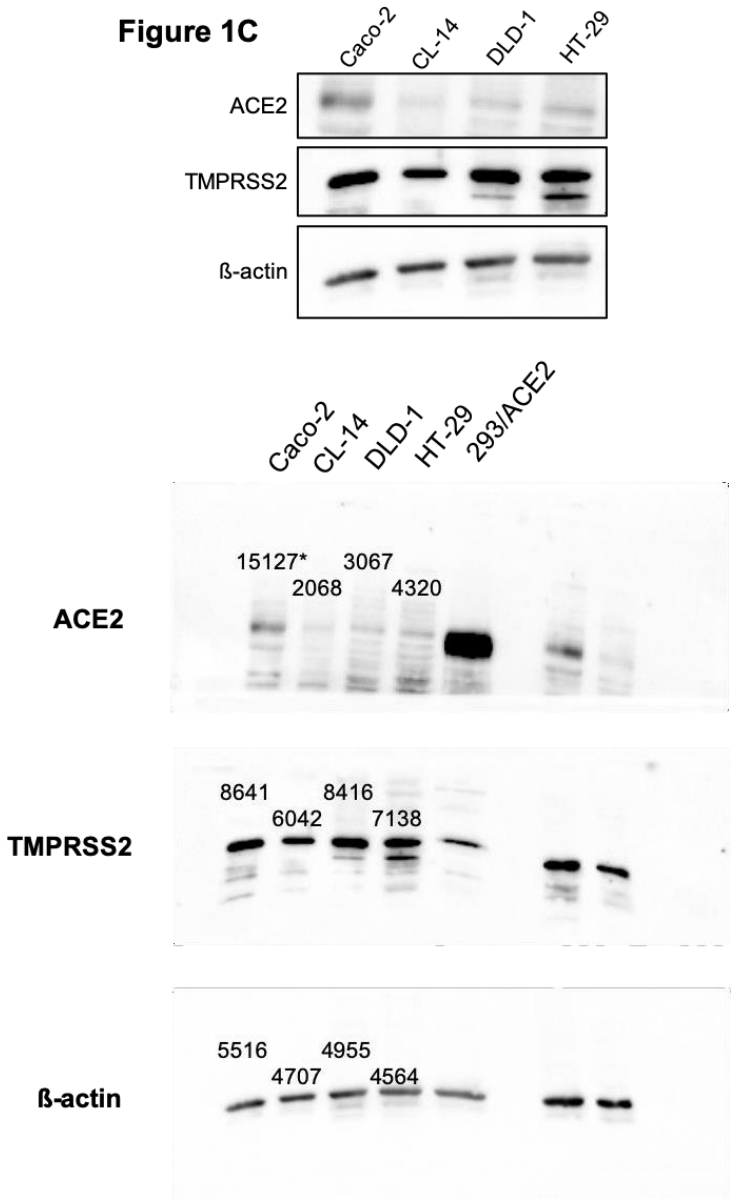


**Figure S4.6**. Role of TMPRSS2-mediated S cleavage in SARS-CoV-2 and SARS-Co-V replication. Concentration-dependent effects of the TMPRSS2 inhibitors camostat and nafamostat on SARS-CoV- 2- and SARS-CoV-induced cytopathogenic effect (CPE) formation determined 48h post infection in CL14 cells infected at an MOI of 0.01. Values are presented as means ± S.D. (n =3).

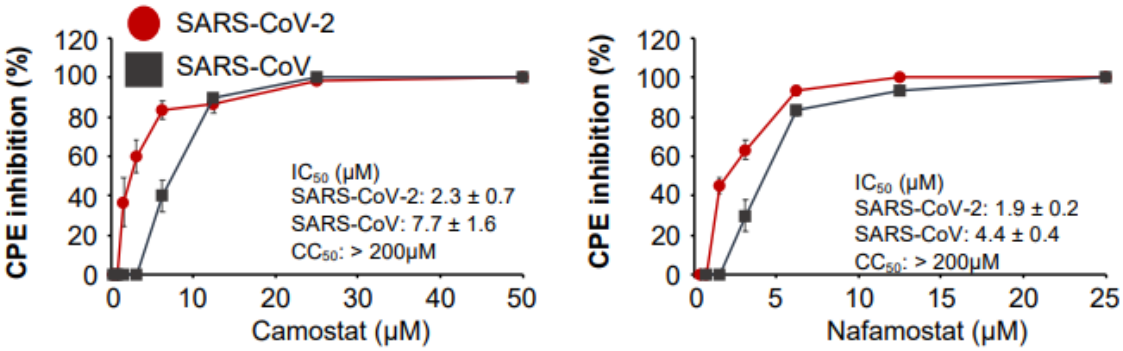


**Table S4.1.** Protein structures used for structural analysis obtained from the Protein Databank.

| SARS | PDB identifier | Protein | Residues |
|---|---|---|---|
| SARS-CoV | 2hsx | NSP1 | 13-127 |
| SARS-CoV | 2gri | NSP3 | 1-111 |
| SARS-CoV | 2fav | NSP3 | 185-354 |
| SARS-CoV-2 | 6vxs | NSP3 | 207-373 |
| SARS-CoV | 2w2g | NSP3 | 389-652 |
| SARS-CoV | 2kaf | NSP3 | 655-720 |
| SARS-CoV | 5y3e | NSP3 | 723-1036 |
| SARS-CoV-2 | 6w9c | NSP3 | 748-1061 |
| SARS-CoV | 2k87 | NSP3 | 1066-1180 |
| SARS-CoV | 2h2z | NSP5 | 1-306 |
| SARS-CoV-2 | 6y2e | NSP5 | 1-306 |
| SARS-CoV | 6nur | NSP7 | 2-71 |
| SARS-CoV-2 | 6xip | NSP7 | 1-70 |
| SARS-CoV | 2ahm | NSP8 | 1-190 |
| SARS-CoV | 2fyg | NSP10 | 10-132 |
| SARS-CoV-2 | 6w61 | NSP10 | 18-132 |
| SARS-CoV | 6nur | NSP12 | 41-819 |
| SARS-CoV-2 | 7bv2 | NSP12 | 31-929 |
| SARS-CoV | 5c8s | NSP14 | 1-525 |
| SARS-CoV | 2h85 | NSP15 | 1-345 |
| SARS-CoV-2 | 6vww | NSP15 | 1-346 |
| SARS-CoV | 2xyq | NSP16 | 1-290 |
| SARS-CoV-2 | 6w61 | NSP16 | 1-299 |
| SARS-CoV | 6acg | S:ACE2 | 18-1119 |
| SARS-CoV-2 | 6m17 | S:ACE2 | 336-518 |
| SARS-CoV | 5xlr | S | 33-1120 |
| SARS-CoV | 5wrg | S | 261-1058 |
| SARS-CoV-2 | 6vsb | S | 27-1146 |
| SARS-CoV-2 | 6xdc | 3a | 40-238 |
| SARS-CoV | 5x29 | E | 8-65 |
| SARS-CoV | 1yo4 | 7a | 16-99 |
| SARS-CoV | 1ssk | N | 49-185 |
| SARS-CoV-2 | 6m3m | N | 48-173 |
| SARS-CoV | 2gib | N | 270-366 |
| SARS-CoV-2 | 6zco | N | 248-364 |

**Table S4.2.** Structural models generated by Phyre2 and used for structural analysis. Where structures were not available from the Protein Databank, the structures were modelled.

| SARS | Template structure | Protein | Residues | Coverage | Confidence | Identity (%) |
|---|---|---|---|---|---|---|
| SARS-CoV-2 | 2gdta1 | NSP1 | 13-127 | 64 | 100 | 86 |
| SARS-CoV | 6zojj | NSP1 | 148-180 | 17 | 99.1 | 76 |
| SARS-CoV-2 | 6zojj | NSP1 | 148-180 | 18 | 99.8 | 100 |
| SARS-CoV-2 | 2gria1 | NSP3 | 2-111 | 5 | 100 | 77 |
| SARS-CoV-2 | 2acfa1 | NSP3 | 207-373 | 8 | 100 | 74 |
| SARS-CoV-2 | 2wctC | NSP3 | 425-676 | 12 | 100 | 76 |
| SARS-CoV-2 | 2fe8B | NSP3 | 745-1058 | 16 | 100 | 82 |
| SARS-CoV-2 | 2k87A | NSP3 | 1089-1203 | 5 | 100 | 82 |
| SARS-CoV-2 | 3gzfD | NSP4 | 403-477 | 18 | 100 | 41 |
| SARS-CoV-2 | 2duca1 | NSP5 | 2-283 | 98 | 100 | 96 |
| SARS-CoV-2 | 2ahmG | NSP8 | 1-175 | 95 | 100 | 97 |
| SARS-CoV-2 | 1uw7A | NSP9 | 1-90 | 100 | 100 | 97 |
| SARS-CoV-2 | 2g9tT | NSP10 | 9-116 | 86 | 100 | 98 |
| SARS-CoV-2 | 6nusA | NSP12 | 118-909 | 87 | 100 | 97 |
| SARS-CoV-2 | 5c8sD | NSP14 | 1-504 | 97 | 100 | 95 |
| SARS-CoV | 6xdcB | 3a | 40-238 | 72 | 100 | 77 |
| SARS-CoV-2 | 5x29B | E | 8-65 | 77 | 99.8 | 91 |
| SARS-CoV-2 | 1yo4A | 7a | 16-98 | 67 | 100 | 91 |

**Table S4.3**. Specificity Determining Positions (DCPs) identified between SARS-CoV and SARS-CoV-2.

| Protein (SARS-CoV) | Protein (SARS-CoV-2) | Sequences in Dataset | Protein Length (SARS-CoV) | DCPs Identified | % of Residues DCPs |
|---|---|---|---|---|---|
| S | S | 73863 | 1255 | 186 | 14.82 |
| 3a | ORF3a | 91214 | 274 | 32 | 11.68 |
| 3b | | n/a | 154 | | |
| E | E | 94787 | 76 | 2 | 2.63 |
| M | M | 93860 | 221 | 15 | 2.26 |
| 6 | 6 | 94935 | 63 | 13 | 9.52 |
| 7a | 7a | 82940 | 122 | 0 | 0 |
| 7b | 7b | n/a | 44 | NA | |
| 8a/8b | 8 | n/a | 39/84 | NA | NA |
| 9b | | n/a | 98 | NA | |
| N | N | 91609 | 422 | 13 | 3.08 |
| | ORF10 | n/a | n/a | | |
| nsp1 | nsp1 | 93621 | 180 | 6 | 3.33 |
| nsp2 | nsp2 | 88288 | 636 | 136 | 21.38 |
| Nsp3 | nsp3 | 75324 | 1922 | 344 | 17.90 |
| nsp4 | nsp4 | 89707 | 500 | 54 | 10.80 |
| nsp5 | nsp5 | 91731 | 306 | 5 | 1.63 |
| nsp6 | nsp6 | 93432 | 290 | 13 | 4.48 |
| nsp7 | nsp7 | 95038 | 83 | 1 | 1.20 |
| nsp8 | nsp8 | 94806 | 198 | 5 | 2.53 |
| nsp9 | nsp9 | 94970 | 113 | 2 | 1.77 |
| nsp10 | nsp10 | 92505 | 139 | 1 | 0.72 |
| nsp12 | nsp12 | 89874 | 932 | 21 | 2.25 |
| nsp13 | nsp13 | 91305 | 601 | 0 | 0 |
| nsp14 | nsp14 | 72306 | 527 | 16 | 3.04 |
| nsp15 | nsp15 | 85595 | 346 | 31 | 8.96 |
| nsp16 | nsp16 | 83565 | 298 | 12 | 4.03 |
| | | | | | |
| Total | | | | 891 | 9.36 |

**Table S4.4.** Analysis of DCPs present in the SARS-CoV and SARS-CoV-2 Spike protein interface with human ACE2.

| SDP | SARS-CoV structural analysis | SARS-CoV-2 structural analysis | Effect? |
|---|---|---|---|
| V404=K417 | V404 is not in the interface | K417 is in the interface and could form a salt bridge with ACE-2 D30 | Likely – new polar interaction within interface |
| R426=N439 | Loss of hydrogen bond to ACE2 Gln325 due to shorter sidechain. N would still be able to form hydrogen bonds | N439 is located away from the interface site and so does not form a hydrogen bond with ACE2. Instead forms a hydrogen bond with S443 (also a DCP – A430=S443) which is likely to stabilise the loop they are both part of. | Likely – Loss of interface hydrogen bond. |
| Y442=L455 | Y422 forms hydrogen bond to backbone of W476 – loss could result in conformational change. The sidechain also contacts the backbone of ACE2 D30 and K31 | L455 remains in interface and contacts ACE2 D30 and H34. | Likely – loss of intramolecular hydrogen bond |
| F460=Y473 | Conservative change. | Introduction of OH group that can form hydrogen bonds. Y473 forms hydrogen bond with backbone of R457 and is closer to ACE2 T27 so potential to form hydrogen bond in interface. | Possible – introduction of hydrogen bond (could be with ACE2) |
| P462=A475 | Located in a loop, could affect this conformation – many DCPs in this loop | Loop has different conformation. | Possible – Conformational change of loop |
| N479=Q493 | Interface hydrogen bond formed with ACE2 H34 backbone. With a shorter sidechain this this may be lost in SARS-CoV-2. | Q493 forms a hydrogen bond with ACE2 E35 in this complex. So hydrogen bond is maintained but also different. | Possible – hydrogen bond with ACE2 retained but to different residue. |
| Y484=Q498 | Y484 can form hydrogen bonds with ACE2 Gln42 (sidechain) and intramolecular H bonds with T433 (backbone), Y436 (sidechain). | Q498 maintains hydrogen bonds with ACE2 Gln42 | Possible – change in residue forming hydrogen bonds with ACE2. |
| T485=P499 | Sidechain points away from interface, loss of hydrogen bond with R426 (also a DCP) backbone in adjacent loop. This hydrogen bond is likely to coordinate the structure between these two loops. There are multiple DCPs present in both loops | Loop conformation similar as for SARS-CoV structure but not coordinated with other loop | Likely - loss of intramolecular hydrogen bond |
| I489=V503 | Conservative change I489 in direct contact with ACE2 Q325 | slightly smaller sidechain is further away from ACE2 Q325. | Unlikely. |