An Evaluation of Dobzhansky-Muller Incompatibilities in Protein Evolution

---

A Senior Honors Thesis Presented to the

Faculty of the Department of Biology and

Biochemistry

University of Houston

In Partial Fulfillment of the Requirements for

the Degree Bachelor of Science

By

Hannah Esopenko

May 2022

An Evaluation of Dobzhansky-Muller Incompatibilities in Protein Evolution

---

Hannah Esopenko

APPROVED:

Dr. Ricardo Azevedo
Department of Biology and Biochemistry

Dr. Rebecca Zufall
Department of Biology and Biochemistry

Dr. Ann Cheek
Department of Biology and Biochemistry
Honors Reader

Dr. Dan Wells, Dean
College of Natural Sciences and Mathematics

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude and appreciation to Dr. R. Azevedo for his expertise, patience, and mentorship throughout this past year. Dr. Azevedo has guided me through this process with encouragement and kindness. His profound knowledge and passion for evolutionary biology inspired me in class to pursue this research. He was always there when I needed guidance and I cannot thank him enough. Additionally, I would like to thank my thesis committee members Dr. A. Cheek and Dr. R. Zufall for their commitment to my thesis project. I deeply value their professional input and advice. Lastly, I would like to thank my family and friends for their unwavering encouragement and support, now and always.

An Evaluation of Dobzhansky-Muller Incompatibilities in Protein Evolution

---

An Abstract of a Senior Honors Thesis Presented

to the Faculty of the Department of Biology and

Biochemistry

University of Houston

In Partial Fulfillment of the Requirements for

the Degree Bachelor of Science

By

Hannah Esopenko

May 2022

ABSTRACT

This study is focused on the evolution of Dobzhansky-Muller Incompatibilities (DMIs) and Compensated Pathogenic Deviations (CPDs) in protein evolution. DMIs are genetic differences that occur by post zygotic isolation to reduce the overall fitness of an organism. Meanwhile, CPDs are pathogenic mutations that show no adverse effects to the organism as there is an additional mutation somewhere in the sequence that compensates for the deleterious nature of the mutation. Therefore, studying the nature of DMIs and CPDs provides a deeper understanding as to how deleterious events arise throughout the evolution of species.

A study conducted by Kondrashov *et al.* (2002) addressed DMIs in protein evolution by identifying the occurrence of CPDs when the nonhuman orthologs deviated from the reference human ortholog sequence. Kondrashov *et al.*'s (2002) study was clever in construction, but the methodology was unclear, and the results appeared to be over simplified. To analyze the validity of the Kondrashov *et al.* (2002) paper, a similar study using restricted parameters and modern bioinformatic databases was conducted for this senior thesis project. To do so, 24 primate orthologs of 32 genes responsible for Mendelian diseases were collected and compared to the pathogenic missense data of humans to identify CPDs. Through computational analysis and the visual representation of protein alignments, 26 valid CPD hits were found. The 26 CPD hits presented in four general patterns: single species CPD, single clade CPD with two or more species, convergent evolution of a CPD, and ancestral CPDs. A statistical analysis was performed to determine whether factors such as the length of the protein, the evolutionary distance between sequences, or the number of pathogenic variants played a role in the number of CPDs found. The relationship between the number of CPDs found and the evolutionary distance between sequences and the amount of pathogenic variant data were found to be statistically significantly correlated.

More data and research into primate genomes and the nature of CPDs is required to accurately determine their occurrence. This will help predict how CPDs arise in species and better evaluate the claims made in the Kondrashov *et al.* (2002) paper.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

1.1 The Impact of Evolution

The theory of evolution not only explained how life on earth evolved over time, but it also provided the backbone to future fields of study and revolutionary scientific advancement. Evolution provides an explanation for the development of living organisms from the first single celled micro-organism to the vast diversity of unicellular and multicellular species throughout history. Darwin's ground-breaking publication, *On the Origin of Species* in 1859, encompassed the magnitude of the theory of evolution and proposed the driving forces and limiting factors of natural selection and speciation. Darwin's findings were before modern genetics, and since then have been further supported with comprehensive genome sequencing and bioinformatic data. From Darwin's initial study and subsequent experimental studies conducted worldwide, the generation and divergence of species is most notably accomplished by slight successive mutations or changes to genes or chromosomes that become integrated into populations over time. As Darwin theorized evolution through generalization and inference from morphological data, it can be regarded as one of the most important achievements of morphological biology (Dobzhansky, 1937). The results from Darwin's studies and published works established the theory of evolution and promoted further study in the field that is active to this day.

1.2 Dobzhansky–Muller Incompatibilities

Although Darwin's construct of evolution and natural selection was ground-breaking, it did not explain exactly why maladaptive traits involved in speciation such as hybrid sterility and inviability occur in populations. Darwin was aware that hybrid sterility was not advantageous, and therefore it could not have accumulated by the preservation of beneficial adaptations (Darwin,

1859). Based on that concept, he aimed to show that sterility was not an independently acquired quality, but it was dependent on other genetic differences (Darwin, 1859). Darwin in turn dedicated a chapter of *On the Origin of Species* to further address the nonbeneficial outcomes from the mating of different species.

It was not until nearly half a century later that a comprehensive explanation was provided for the genetic incompatibilities of hybrids. The work done by Theodosius Dobzhansky and Herman Muller took large steps toward solving the species problem, by introducing the idea that hybrid sterility and inviability are caused by interacting complementary genes (Orr, 1996). Both Dobzhansky and Muller focused primarily on postzygotic reproductive isolation when conducting their individual studies. Postzygotic isolation occurs when individuals of different species can mate and produce offspring, in contrast to prezygotic isolation that prevents fertilization of the egg. The offspring from postzygotic isolation have an overall lower fitness compared to either parent species, and as a result are not favored by natural selection. Inviability and hybrid sterility are two forms of postzygotic isolation.

In his early work, Dobzhansky proposed two types of hybrid sterility, the chromosomal type, and the genic type. Chromosomal sterility is a result of differences in chromosome structure and improper alignment during meiosis causing irregular pairing and disjunction, while genetic sterility is due to interactions between complementary genetic factors from both maternal and paternal lineages (Dobzhansky, 1933, 1934). To test genetic hybrid sterility, Dobzhansky conducted an experiment with two sibling species, *Drosophila pseudoobscura* ("Race A") that carried mapped visible markers, and *Drosophila persimilis* ("Race B") (Dobzhansky, 1936). He found that when the F1 progeny of Race A was crossed with Race B it resulted in sterile males and fertile females. He then backcrossed the fertile F1 females with pure males of either Race A or B.

This backcrossed progeny allowed Dobzhansky to identify which combinations of the ancestral elements were necessary to induce sterility, and which caused the individual to be fertile (Dobzhansky, 1936**)**. Based on these results, he postulated that sterility and inviability arise from evolution of separate lineages of alleles at different loci that may increase fitness independently, but when brought together in hybrids the combination results in lower overall fitness (Turelli & Orr, 2000).

Dobzhansky's findings were further supported and elaborated by Muller, who highlighted that natural selection may drive the evolution of postzygotic reproductive isolation (Muller, 1942). Muller also noted that the two genetic changes can occur in the same lineage and do not need to occur in both separate lineages. From compiling their independent works done with *Drosophila* crosses, the genetic incompatibilities underlying hybrid sterility and inviability are now known as Dobzhansky-Muller incompatibilities (DMIs). This model explains a key role of speciation and proposes that deficits in hybrids are the result of negative epistatic interactions between alleles of different loci that have independent genetic backgrounds (Turelli *et al.*, 2001; Wang *et al.*, 2013).

1.3 Compensated Pathogenic Deviations

A central focus in structural and evolutionary biology is studying the changes that arise between human genetic sequences in comparison to their nonhuman orthologs (Barešić *et al.*, 2010). An area of emphasis in both fields is what is known as Compensated Pathogenic Deviations (CPDs). This phenomenon occurs when individual amino acid substitutions are pathogenic or deleterious alone, but when paired with an additional mutation in the sequence, result in a neutral or beneficial effect on the overall fitness of the organism. These compensated mutations in turn

enable the organism to flourish with no adverse effects despite harboring a potentially detrimental mutation (Barešić & Martin, 2011).

Without the accompanying compensatory mutation, the pathogenic substitution may act as a DMI and reduce the overall fitness of the organism. A study conducted by Kondrashov *et al.* (2002) highlighted the relationship between the two concepts using a CPD that they identified in the β-hemoglobin protein (Kondrashov *et al*., 2002). Kondrashov *et al.* (2002) noted that the substitution of Valine (Val) to Glutamate (Glu) at site 20 in the human β-hemoglobin is pathogenic but acts as the wild-type allele in the horse sequence with no adverse effects. The horse sequence has the Val to Glu substitution which is accompanied by a Histidine (His) residue at site 69. This His residue in horses appears to compensate for the pathogenic nature of the Glu mutation. Whereas the human sequence experiences the pathogenic effects of the Glu mutation as it has a Glycine (Gly) residue at site 69 as opposed to His. This uncompensated glutamate mutation is pathogenic in the human sequence and in turn acts as a DMI that is deleterious.

Studying CPDs is relevant as it provides a deeper understanding of the fitness landscape of protein evolution and provides insight about the molecular nature of diseases (Barešić & Martin, 2011). Kimura *et al.* (1986) developed the neutral theory which states molecular evolutionary changes largely result from fixation of neutral or nearly neutral mutations, as opposed to the previous belief of Darwinian evolution with selection towards advantageous mutations (Kimura *et al*., 1986). His theory in turn, assumes that Darwinian selection is primarily involved with phenotypic evolution and controlled by positive selection that produces adaptation of organisms to their environment (Kimura *et al.*, 1986). Kimura's work with compensatory mutations was used as the foundation for many studies investigating CPDs and their prevalence in genomes (Kimura *et al*., 1986). A central focus of these studies revolves around comparing human protein sequences

to nonhuman protein sequences to assess how CPDs bridge the fitness valleys of a population, the extent at which they occur, and the involvement of RNA structures (Knies *et al.*, 2008; Kern & Kondrashov, 2004).

1.4 Kondrashov et al. *(2002) and Thesis Objectives*

As noted previously, Kondrashov *et al*. (2002) conducted a clever study regarding DMIs in protein evolution that sparked further analysis. A key point to highlight was the ability of Kondrashov *et al.* (2002) to identify DMIs without having to physically cross species. He and his team were able to clearly identify DMIs by locating CPDs in nonhuman orthologs (Kondrashov *et al.*, 2002). This elegant construction allows for easier analysis and shows significant advancement in bioinformatic technology and the resources that have become available since Dobzhansky and Muller's time.

Kondrashov *et al.* (2002) focused the study on 32 human proteins responsible for Mendelian disease. The team collected and compared known pathological missense mutations of the human proteins to the amino acid substitutions that occurred at the same sites over the course of evolution in the corresponding animal orthologs. From the Kondrashov *et al.* (2002) data analysed, broad generalizations were made about how often CPDs occur when looking at the overall deviations from the reference sequence. This was done by analysing at all the sites in ortholog sequences that differed from the human reference sequence that contained a known pathogenic variant. The 32 proteins used, and the summary of the data from the Kondrashov *et al.* (2002) study is shown in Table 1.

Table 1. Summary of data and results from Kondrashov *et al.* (2002)

Locus	Known missense	Known nonsense	All missense	All nonsense	Pathogenic missense	Orthologs	CPD
ABCD1	124	30	4,415	236	0.22	3	0
ALPL	83	5	3,147	156	0.82	6	1
AR	212	28	5,446	346	0.48	10	10
ATP7B	118	16	8,829	421	0.35	3	1
BTK	135	82	3,944	297	0.12	3	2
CASR	52	3	6,477	360	0.96	5	0
CBS	74	4	3,295	166	0.93	4	3
CFTR	446	133	8,733	650	0.25	15	70
CYBB	80	47	3,398	218	0.11	7	0
F7	82	6	2,778	166	0.82	3	1
F8	354	74	13,995	996	0.34	3	4
F9	419	55	2,759	211	0.58	9	3
G6PD	103	—	3,114	169	0.5*	15	17
GALT	99	11	2,247	139	0.56	6	7
GBA	94	11	3,060	173	0.48	3	0
GJB1	188	19	1,696	88	0.51	6	4
HBB	152	13	877	34	0.45	218	109
HPRT1	90	6	1,307	81	0.93	8	1
IL2RG	51	27	2,177	148	0.13	3	0
KCNH2	72	10	6,919	282	0.29	4	0
KCNQ1	69	—	4,022	186	0.5*	4	0
L1CAM	52	17	7,524	423	0.17	3	0
LDLR	273	67	5,242	297	0.23	8	15
MPZ	52	6	1,519	85	0.48	6	1
MYH7	62	—	11,620	755	0.5*	26	14
OCA1	71	13	3,155	226	0.39	106	31
PAH	272	23	2,699	193	0.85	6	8
PMM2	53	1	1,473	100	1.00	6	5
RHO	64	—	2,102	105	0.5*	116	10
TP53	80	7	2,345	129	0.63	29	5
TTR	74	—	870	41	0.5*	19	13
VWF	122	8	2,507	122	0.74	125	215

The columns are (left to right): name of a locus/protein; number of known pathogenic missense mutations; number of known pathogenic nonsense mutations; number of all missense mutations; number of all nonsense mutations; estimated fraction of pathogenic mutations among all missense mutations ("*" absence of direct data); number of analyzed nonhuman orthologs; and total number of validated CPDs detected in these orthologs.

To add to the intrigue of his sophisticated identification of DMIs and overall findings from his analysis, the comprehensive methodology and results from Kondrashov *et al.* (2002) have yet to be widely replicated. One study conducted by Kulathinal *et al.* (2004) presented similar findings about the occurrence of CPDs from deviations from the reference sequence but had significantly different methodology (Kulathinal *et al.*, 2004). Kulathinal *et al.* (2004) used much more restricted parameters which focused only on insect genomes differing from *D. melanogaster*. Kulathinal *et al.* (2004) looked at numerous insects which were closely related, the shared homology of insects made for an adequate and relative comparison. This highlights the challenges of the Kondrashov *et al.* (2002) study which chose to assess a vast range of species.

This leads to a subsequent point of interest regarding the accuracy of the results obtained from this thought-provoking study. Due to the large number of species Kondrashov *et al.* (2002) used, it reduced the confidence in the homology between the sequences and potentially allowed for an increase of less accurate CPD identifications. Additionally, the paper did not clearly outline the ways in which the study was carried out. As seen in Table 1, it shows how many orthologs were used for each species but does not specify the species to which those orthologs belonged. This allowed for a certain extent of interpretation to determine how to reconstruct the study under the same conditions. Although the concepts presented in the paper were significant and intriguing in theory, it would be beneficial to take a closer look at how one would replicate this study with the resources available twenty years later. The purpose of this Senior Honors Thesis project was to replicate the Kondrashov *et al.* (2002) study with the bioinformatic databases available in 2022, while narrowing the species parameters to determine whether similar results could be obtained.

METHODS

The research for this study was conducted through a series of Jupyter Notebooks constructed by Dr. Azevedo. To work in these notebooks, I had to install Anaconda (https://www.anaconda.com/) which provided me access to JupyterLab (https://jupyter.org/). From there I was able work in the notebooks, run the codes, and input the necessary data I retrieved from bioinformatic databases to obtain hits for CPDs. To narrow the range of species being analysed, I focused only on primate ortholog sequences which included up to 24 nonhuman genera (which can be seen in the results section). This differs from the Kondrashov *et al.* (2002) study that included all animal ortholog sequences. A comparison between the number of orthologs looked at for each gene analyzed can be seen in Table 2.

2.1 Orthologs and Alignments

I began by creating FASTA files of all the primate ortholog sequences for each human gene used in the Kondrashov *et al.* (2002) paper. Starting with ABCD1, the *Homo sapien*s ABCD1 protein was queried in NCBI (https://www.ncbi.nlm.nih.gov/). From there, all the primate ortholog protein sequences were collected on August 2nd 2021 and processed to remove the redundant sequences from the analysis. NCBI can determine orthology by looking between the genome of focus and a reference genome (most commonly the human sequence) and the two are then tracked as a group (NCBI 2022). These were then made into FASTA files to be used for future steps. This process was repeated for the other 31 proteins and all the data collected was similarly recorded and saved. I then generated alignments of the 32 previously created FASTA files and constructed them into ALN files. Protein sequences were aligned using the MUSCLE program version 3.8.1551 with default parameters (Edgar, 2004). These ALN files were able to be opened in the

Table 2. The number of animal ortholog sequences identified by Kondrashov *et al*. (2002) vs. the primate ortholog sequences from 2021

Protein	Kondrashov Orthologs	Primate Orthologs
ABCD1	3	23
ALPL	6	22
AR	10	24
ATP7B	3	24
BTK	3	18
CASR	5	23
CBS	4	20
CFTR	15	23
CYBB	7	20
F7	3	22
F8	3	22
F9	9	23
G6PD	15	18
GALT	6	24
GBA	3	20
GJB1	6	2
HBB	218	7
HPRT1	8	12
IL2RG	3	23
KCNH2	4	23
KCNQ1	4	23
L1CAM	3	23
LDLR	8	24
MPZ	6	19
MYH7	26	6
OCA1	106	21
PAH	6	18
PMM2	6	21
RHO	116	11
TP53	29	22
TTR	19	20
VWF	125	24

The data was collected from the Kondrashov *et al.* (2002) study and the unique primate ortholog sequences collected in 2021 from NCBI. The primate orthologs shown in the table include the 24 nonhuman genera.

alignment viewer application (https://alignmentviewer.org) to physically see the changes at each location for all the ortholog sequences.

2.2 Human Genetic Variants

Next, I collected the phenotypic effects due to human genetic variants for each of the 32 genes from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) on August 4th, 2021 (Landrum *et al*., 2014). To do so, I exported the single missense variants for each gene. I then sorted through the ClinVar data for each gene and discarded all classifications of clinical significance that were not pathogenic. The clinical significance values are labelled using terminology such as "pathogenic," "likely pathogenic," "uncertain significance," "likely benign," and "benign" to describe the nature of the variants when present in the human sequence (Richards *et al*., 2015). From this list of classifications, I restricted the analysis to include "pathogenic" and "likely pathogenic" variants only. Next, I checked to see whether the nucleotide and variant sites referenced in ClinVar corresponded to the data collected from the primate ortholog sequences from each protein. Any proteins sequences that failed to match their orthologs sequence were flagged and required further analysis. I then dealt with these invalid transcripts that had potential alignment issues and transcripts not matching any human protein sequences. These two types of flagged protein sequences were either fixed or discarded from the records moving forward.

2.3 Finding CPDs

To begin the final analysis, I pulled the transcripts from the revised ClinVar data records and the ortholog sequences in the FASTA files. I then extracted and identified each individual pathogenic human variant identified in the ClinVar data for each of the 32 proteins. Next, I processed each alignment previously generated, identifying the human variant sites known to be

pathogenic to each ortholog protein sequence. Building off this work done, the next step was to identify all possible CPDs. This was achieved when a nonhuman ortholog sequence contained a known pathogenic mutation that was detrimental to humans at the same site in the sequence. This was assumed to be a CPD as the mutation appeared in the recorded ortholog sequence without possessing the deleterious effects the human counterpart would experience. For every CPD identified the name of the gene, site location, transcript number, human wild-type amino acid, and pathogenic mutant amino acid were recorded.

The potential CPDs were screened by eye in the alignment viewer, and then I proceeded to visually validate the putative CPD hits. The next step in the analysis was to test the putative CPDs by investigating the vicinity of the sequence, 10 amino acids on either side of the site. The CPDs were validated if they had no gaps closer than 10 sites to a CPD and a minimum of four exact matches 10 sites before and after (Kondrashov *et al*., 2002). This paralleled the validation criteria used in the Kondrashov *et al*. (2002) study to ensure the homology between the sequences and similarity of the CPDs hits found. If an identified CPD failed this validation criteria, it was not considered to be a valid hit. This validation process was then repeated for each potential CPD hit and recorded and analyzed.

2.4 Identifying Patterns in CPD Hits

The phylogeny of primates was required to identify the nature of the validated CPDs and to better understand how the mutations arose in species evolution. To do so, a phylogenetic tree was constructed from the compiled data collected from two detailed studies focused on the phylogeny of primates (Finstermeier *et al.*, 2013; Perelman *et al.*, 2011). The tree construction in both papers was done using maximum likelihood (ML) and it helped visualize the relationship

between the 25 primate orthologs used in this study. It is important to note that the trees are not scaled to represent the evolutionary distance of the divergence of genera but are representative of the relation between the 25 primate orthologs. In addition, the phylogenetic trees include the genera with redundant sequences, and therefore the number of ortholog species in the trees may differ from the numbers presented in Table 2.

Using the previously constructed phylogenetic tree from the literature, I was able to manipulate the specific mutation and number of genera present for each identified CPD hit in Mesquite (http://www.mesquiteproject.org). The ancestral state reconstruction of the fixed tree was done using Wagner parsimony. This is the simplest and most common case of parsimony that assumes evolutionary changes are reversible and generates the pattern with the least amount of observed evolutionary changes (Felsenstein, 1983). The individualize analysis of each validated CPD was required as they all differed in the variant amino acid and the amount of non redundant primate ortholog sequences. In turn Mesquite allowed for the visualization of the phylogenetic tree tailored to each particular gene and site mutation. This allowed me to approximate where each CPD arose in the primate lineage and determine whether it was ancestral or derived.

2.5 Correlates of the Number of CPDs per Gene

A statistical analysis was conducted to determine whether the length of the sequence and or the number of ClinVar variants recorded had a relation to the number of CPDs found. Dr. Azevedo used the human reference sequence and a sequence from *Macaca mulatta* to compare their sequence length, the number of gaps, changes between the two sequences, and the evolutionary distance between the two species. In addition, the amount of ClinVar data for each protein and the total number of valid CPDs found in this study were compared to the data collected

for the human and *Macaca mulatta* sequence. Dr. Azevedo then calculated the Spearman's rank correlation between the number of CPDs and the possible correlates: length of the protein, evolutionary distance, number of pathogenic variants. This statistical analysis was used to determine whether there was a relationship between the possible correlates and predicting the amount of CPD hits.

RESULTS

3.1 Human Genetic Variants

The number of human genetic variants collected regarding the 32 proteins from ClinVar was slightly different in comparison to the pathogenic data Kondrashov *et al.* (2002) used. In the twenty years between these two studies, the number of mutations classified as pathogenic decreased for 13 proteins but increased for 19 other proteins. A comparison between the amount of pathogenic data identified by Kondrashov *et al.* (2002) and the ClinVar data I collected August 4th, 2021, are shown in Table 3.

The increase of pathogenic variants recorded for each protein was expected as the time between the two studies allowed for more research to be conducted. But the considerable number of proteins for which the number of identified pathogenic variants decrease was unexpected. A possible explanation for this could be an increase of medical knowledge and understanding of the effects of the 32 human protein variant sites. This could potentially change the previously recorded "pathogenic" and "likely pathogenic" clinically significant classifications to "conflicting pathogenicity," "likely benign," and/or one of the other classifications that were not included in the parameters of this study.

Table 3. The number of pathogenic missense mutations identified in 2002 vs. 2021

Protein	Pathogenetic 2002 Data	Pathogenetic 2021 Data
ABCD1	124	134
ALPL	83	76
AR	212	132
ATP7B	118	185
BTK	135	102
CASR	52	109
CBS	74	68
CFTR	446	391
CYBB	80	55
F7	82	26
F8	354	269
F9	419	138
G6PD	103	53
GALT	99	174
GBA	94	103
GJB1	188	111
HBB	152	127
HPRT1	90	46
IL2RG	51	55
KCNH2	72	206
KCNQ1	69	210
L1CAM	52	54
LDLR	273	808
MPZ	52	97
MYH7	62	270
OCA1	71	72
PAH	272	358
PMM2	53	56
RHO	64	105
TP53	80	241
TTR	74	77
VWF	122	149

The data was collected from the Kondrashov *et al.* (2002) study and the ClinVar database in 2021.

3.2 Putative CPDs Found

Of the 32 proteins used, I identified a total of 72 putative CPDs in 18 of the proteins. Of these 18 proteins, some were identified as having one or multiple sites with putative CPDs. The putative CPDs were flagged by having a known pathogenic amino acid variant at a particular site in the protein sequence, opposed to the human wild-type amino acid. The software proceeded to highlight any additional variants or gaps accumulated in that site. A few genes such as ALPL, F9, PMM2, and VWF were found having only one site with a putative CPD. Whereas, the other 12 proteins identified had multiple sites in the protein sequences with the mutant variant. The results of all the identified putative CPDs are shown in Table 4.

Based on the information presented in Table 4, all the hits were further analysed by looking at the protein alignments of all the orthologs in the alignment viewer application. From this visual representation the determination of valid and invalid CPDs and hits that resulted from poor sequence alignment could be made. The issue with a poorly aligned sequence or excessively divergent sequence is that the homology at the site can not be certain. Figure 1 shows the difference between an excessively divergent sequence using ALPL, an invalid CPD using F8, and a valid CPD using GBA. The validation criteria proposed by Kondrashov *et al.* (2002) were followed in this analysis. Any putative CPD hits that had no gaps closer than 10 amino acid sites, and a minimum of four exact matches 10 sites before and after were considered valid CPDs. Any putative CPD failing these criteria was considered invalid. A total of 26 validated CPD were identified using the alignment viewer and the validation criteria. The validated hits are shown in Table 4 with a "*" next to the protein name.

Table 4. Identified putative CPDs

Gene	Transcript	Site	Wild-Type	Mutant
ALPL	NM_000478.6	491	G	R
AR	NM_000044.6	611	N	K
AR	NM_000044.6	788	M	V
ATP7B	NM_000053.4	1178	T	A
ATP7B*	NM_000053.4	969	R	Q
CFTR	NM_000492.4	1	M	R
CFTR*	NM_000492.4	13	S	F
F8	NM_000132.3	2319	P	L
F8	NM_000132.3	2185	L	S
F8	NM_000132.3	2183	M	V
F8*	NM_000132.3	2038	N	S
F8	NM_000132.3	1979	G	V
F8	NM_000132.3	585	I	T
F8	NM_000132.3	584	Q	K
F8	NM_000132.3	494	I	T
F9*	NM_000133.3	75	R	Q
GALT*	NM_000155.4	23	T	A
GALT	NM_000155.4	97	N	D
GALT	NM_000155.4	114	H	L
GALT	NM_000155.4	129	M	L
GALT	NM_000155.4	186	H	Y
GALT*	NM_000155.4	198	I	T
GALT	NM_000155.4	204	R	P
GALT	NM_000155.4	212	Q	P
GALT	NM_000155.4	226	L	P
GALT	NM_000155.4	319	H	Q
GALT	NM_000155.4	329	S	P
GALT	NM_000155.4	333	R	L
GALT	NM_000155.4	363	E	K
GBA*	NM_000157.4	535	R	H
GBA*	NM_000157.4	502	R	C
GBA*	NM_000157.4	416	G	S
GBA*	NM_000157.4	351	W	C
GBA*	NM_000157.4	350	H	R
GBA*	NM_000157.4	324	R	C
GBA*	NM_000157.4	227	N	S

Table 4. Continued

Gene	Transcript	Site	Wild-Type	Mutant
GBA*	NM_000157.4	223	W	R
KCNH2	NM_172056.2	92	R	L
KCNH2	NM_172056.2	72	P	L
KCNH2	NM_172056.2	65	T	P
KCNH2	NM_172056.2	56	R	Q
KCNQ1	NM_181798.1	145	G	D
KCNQ1	NM_000218.2	1	M	K
KCNQ1	NM_000218.2	2	A	G
KCNQ1	NM_000218.2	114	L	P
KCNQ1	NM_000218.2	115	E	K
KCNQ1	NM_000218.2	117	P	S
KCNQ1	NM_000218.2	125	Y	D
LDLR	NM_000527.5	36	D	E
LDLR	NM_000527.5	46	C	S
LDLR*	NM_000527.5	137	G	S
LDLR	NM_000527.5	579	D	G
LDLR	NM_000527.5	583	H	D
LDLR*	NM_000527.4	88	R	K
LDLR*	NM_000527.4	215	R	H
LDLR*	NM_000527.4	470	S	G
LDLR*	NM_000527.4	471	R	K
LDLR*	NM_000527.4	578	V	A
TYR	NM_000372.5	346	G	V
TYR	NM_000372.5	370	M	I
PAH*	NM_000277.3	413	R	H
PAH	NM_000277.3	400	R	S
PAH*	NM_000277.3	176	R	Q
PMM2	NM_000303.3	226	T	S
TP53	NM_000546.5	276	A	G
TP53	NM_000546.5	273	R	G
TTR*	NM_000371.4	142	V	I
TTR*	NM_000371.3	77	G	R
TTR*	NM_000371.3	82	E	K
TTR*	NM_000371.3	127	I	V
TTR	NM_000371.3	131	L	M
TTR	NM_000371.3	136	Y	S
VWF	NM_000552.4	528	N	S

The validated hits are denoted by a "*" next to the protein name. The columns are (left to right): protein name; transcript ID, site of CPD in the sequence; the wild-type human amino acid; the mutant CPD amino acid.

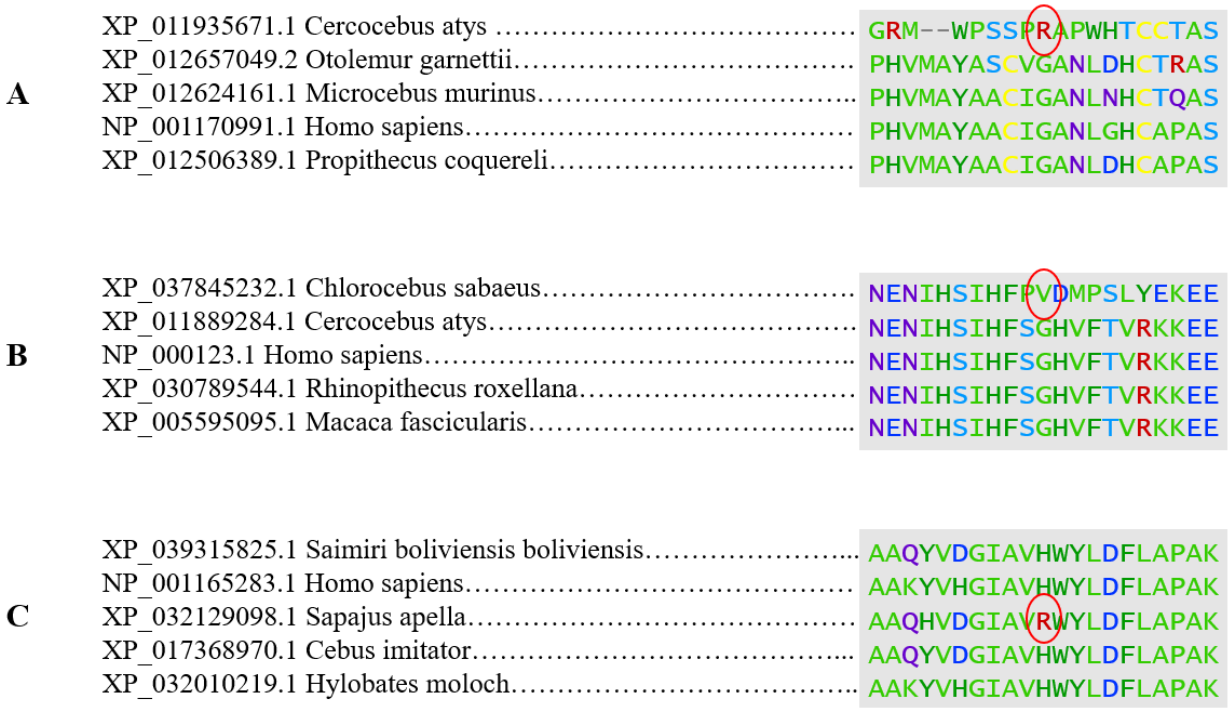

Figure 1. The different types of alignments observed for the identified putative CPDs. It shows 10 amino acids on either side of the CPD circled in red to highlight the validation criteria A.) Represents the identified putative CPD of ALPL at site 491 having a mutant R in the *Cercocebus atys* ortholog sequence opposed to the wild-type G amino acid. The overall sequences show large amounts of variation on either side of the potential CPD and includes gaps within 10 sites, therefore is poorly aligned and an invalid CPD. B.) Represents the identified putative CPD of F8 at site 1979 having a V mutant in the *Chlorocebus sabaeus* ortholog sequence opposed to the wild-type G amino acid. This alignment shows a relatively well conserved sequence but has less than four exact matches to the left of the CPD, therefore we considered it an invalid CPD C.) Represents the identified putative CPD of GBA at site 350 having a mutant R in the *Sapajus apella* ortholog sequence opposed to the wild-type H amino acid. This alignment is well conserved and has no gaps and more than four exact matches on either side of the CPD, therefore is considered a valid CPD.

3.3 Validated CPDs

To confirm the accuracy of the previously validated CPDs, I was able to check all the sites and highlight the different ortholog sequences containing the mutant amino acid. From this, all 26 hits were considered valid and contained a valid CPD hit in one or multiple primate ortholog protein sequences. I then looked at how many changes in the sequence were present and the number of gaps 10 sites before and after the CPD. In some cases, the identified CPD was near the end of the sequence and did not have exactly 10 sites to the right, but it was still considered valid if the left side leading up to the CPD was well conserved. The sites with the lowest divergence from the human protein sequence showed the greatest sequence similarity and the strongest incidence of a true CPD. The overall results can be seen in Table 5.

From Table 5, the similarities and differences between the CPD hits are apparent. The data presented highlights that there were 0 gaps or sites out of range recorded for all the identified CPDs. This indicated true hits and shows the difference in the number of changes in the sequence 10 sites to either side. Based on the results, it is evident that the most common type of the mutant CPD variant recorded was isolated to a single species, which can be seen in 15 out of the 26 validated CPD hits. The 3 potential single species mutations occurring only in *Otolemur* were not included in this total as it unclear whether the origin species contained the mutant variant or the human wild type. This uncertainty of the ancestral state is consistent with both the data from the tables and phylogenetic trees. This is due to the simplistic nature of Wagner parsimony reconstruction of ancestral states as it aims to minimize the evolutionary character changes between the tips of the trees and does not use a statistical model to define uncertainties (Joy *et al.*, 2016). The remaining 8 identified hits showed multiple primate orthologs containing a CPD at a particular site in the sequence.

Table 5. Validated CPDs with all identified primate ortholog sequences

Gene	Protein ID	Species	Site	Changes to sequence	Gaps
ATP7B	XP_030669295.1	*Nomascus leucogenys*	491	0	0
CFTR	XP_003789873.1	*Otolemur garnettii*	13	7	0
F8	XP_035144731.1	*Callithrix jacchus*	2038	0	0
	XP_035144732.1	*Callithrix jacchus*		0	0
	XP_035144729.1	*Callithrix jacchus*		0	0
F9	XP_023373463.1	*Otolemur garnettii*	75	2	0
GALT	XP_003800307.1	*Otolemur garnettii*	23	5	0
GALT	XP_012512307.1	*Propithecus coquereli*	198	1	0
	XP_012512299.1	*Propithecus coquereli*		1	0
GBA	XP_032010219.1	*Hylobates moloch*	535	0	0
GBA	XP_031509384.1	*Papio anubis*	502	0	0
GBA	XP_011932225.1	*Cercocebus atys*	416	1	0
	XP_025237659.1	*Theropithecus gelada*		0	0
	XP_031509384.1	*Papio anubis*		0	0
GBA	XP_011932225.1	*Cercocebus atys*	351	0	0
GBA	XP_032129098.1	*Sapajus apella*	350	3	0
GBA	XP_025237659.1	*Theropithecus gelada*	324	1	0
GBA	XP_011932225.1	*Cercocebus atys*	227	4	0
	XP_025237659.1	*Theropithecus gelada*		4	0
	XP_031509384.1	*Papio anubis*		4	0
GBA	XP_011932225.1	*Cercocebus atys*	223	4	0
	XP_025237659.1	*Theropithecus gelada*		4	0
	XP_031509384.1	*Papio anubis*		4	0
LDLR	XP_021532454.1	*Aotus nancymaae*	137	5	0
LDLR	XP_021573646.1	*Carlito syrichta*	88	2	0
LDLR	XP_021573646.1	*Carlito syrichta*	215	2	0
	XP_023373820.1	*Otolemur garnettii*		2	0
	XP_012665677.1	*Otolemur garnettii*		3	0
	XP_021532454.1	*Aotus nancymaae*		3	0
	XP_035141328.1	*Callithrix jacchus*		3	0
	XP_002761791.1	*Callithrix jacchus*		3	0
	XP_039321728.1	*Saimiri boliviensis boliviensis*		3	0
	XP_037583448.1	*Cebus imitator*		3	0
	XP_032107631.1	*Sapajus apella*		2	0
LDLR	XP_012496802.1	*Propithecus coquereli*	470	0	0

Table 5. Continued

Gene	Protein ID	Species	Site	Changes to sequence	Gaps
LDLR	XP_021532454.1	*Aotus nancymaae*	471	3	0
	XP_035141328.1	*Callithrix jacchus*		3	0
	XP_002761791.1	*Callithrix jacchus*		3	0
	XP_039321728.1	*Saimiri boliviensis boliviensis*		3	0
	XP_037583448.1	*Cebus imitator*		3	0
	XP_032107631.1	*Sapajus apella*		3	0
LDLR	XP_035141328.1	*Callithrix jacchus*	578	2	0
	XP_002761791.1	*Callithrix jacchus*		2	0
PAH	XP_024111838.1	*Pongo abelii*	413	0	0
	XP_024111837.1	*Pongo abelii*		0	0
	XP_035113308.1	*Callithrix jacchus*		0	0
	XP_035113305.1	*Callithrix jacchus*		0	0
	XP_012324641.1	*Aotus nancymaae*		0	0
	XP_003929669.1	*Saimiri boliviensis boliviensis*		0	0
	XP 032155347.1	*Sapajus apella*		0	0
	XP_032155339.1	*Sapajus apella*		0	0
PAH	XP_035113308.1	*Callithrix jacchus*	176	0	0
	XP 035113305.1	*Callithrix jacchus*		0	0
	XP_012324641.1	*Aotus nancymaae*		0	0
	XP_003929669.1	*Saimiri boliviensis boliviensis*		0	0
TTR	XP_007972558.2	*Chlorocebus sabaeus*	142	0	0
TTR	XP_021562314.1	*Carlito syrichta*	77	3	0
	XP_021562313.1	*Carlito syrichta*		3	0
	XP_021562312.1	*Carlito syrichta*		3	0
	XP_008071577.1	*Carlito syrichta*		3	0
TTR	XP_011829762.1	*Mandrillus leucophaeus*	82	0	0
TTR	XP_010333768.1	*Saimiri boliviensis boliviensis*	127	3	0
	XP_032149005.1	*Sapajus apella*		3	0
	XP_012302669.1	*Aotus nancymaae*		2	0
	NP_001254679.1	*Callithrix jacchus*		2	0

The columns are (left to right): name of protein; the protein ID sequence; species name; site of CPD in the sequence; the number of changes between the human and primate ortholog sequence; the number of gaps between the human and primate ortholog sequence.

3.4 Patterns Occurring in the Validated CPD Hits

From looking at the phylogenetic trees and comparing it to the results in Table 5, there were four distinct patterns of CPD occurrence: in a single species, isolated to a single clade with two or more species, a convergent evolution, and an ancestral variant with a subsequent derived human wild type.

i.) A CPD Occurring in a Single Species

The mutation found in a single species was the most common pattern and was observed to occur in over half of the identified CPD hits. An example of the single species mutation is shown in Figure 2 using the F8 protein having a single CPD in the *Callithrix* sequence. Of the 25 primate species, the F8 protein sequence was only available for 23. Therefore, the missing two genera, *Carlito* and *Mandrillus* were excluded from the tree. The constructed phylogenetic tree of F8 demonstrates that the CPD at site 2038 is derived. The human amino acid is ancestral as all the species, including the human sequence contain the wild-type amino acid.

The single CPD present in the *Callithrix* genus of the F8 protein sequence was evaluated further to see if any similar mutations were made in comparison to the human sequence. In the F8 protein sequence there were 10 different sites identified that displayed a similar pattern of a variant in the *Callithrix* genus in comparison to the human sequence. These 10 sites show cases of a single mutation in *Callithrix*, multiple mutations in nearby relatives, and two sites in which the human species carried the differing allele. The 10 sites identified are shown in Table 6.

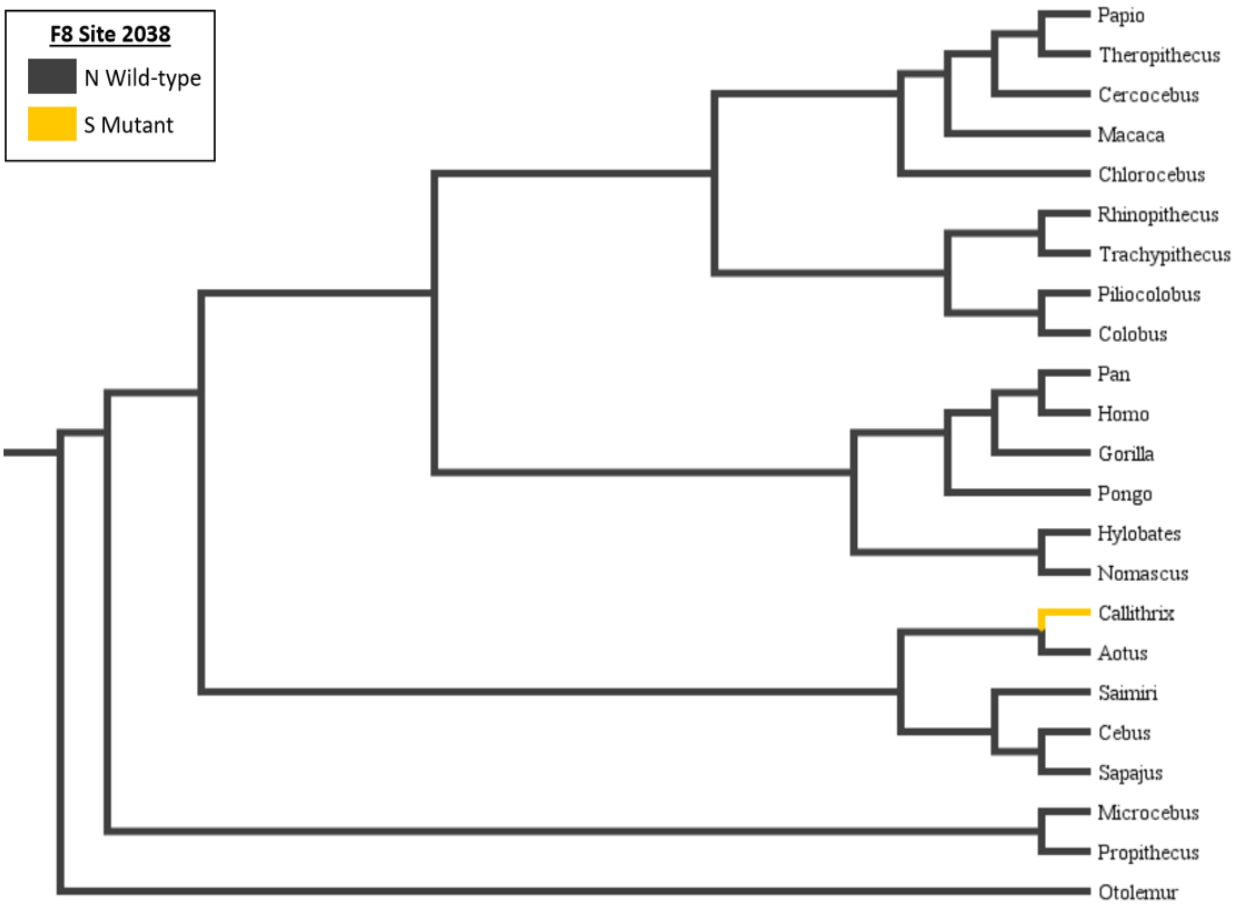

Figure 2. The phylogenetic tree of a single species CPD. Displaying a CPD in the *Callithrix* species of the F8 protein at site 2038.

Table 6. The identified sites in the F8 protein that have variants in the *Callithrix* genus that differ from the human sequence

Site	Genera differing from the human sequence	Human amino acid	Variant amino acid	Description of occurrence
1920	*Callithrix*	A	P	Independent derived variant
	Chlorocebus	A	S	Independent derived variant
1924	*Callithrix/Sapajus* clade	I	V	Derived variant restricted to clade
	Propithecus	I	V	Independent derived variant
1994	All genera excluding the *Pan/Gorilla* clade and *Chlorocebus*	L	V	Ancestral variant, human amino acid restricted to clade and Chlorocebus
2050	All genera excluding the *Pan/Gorilla* clade and *Otolemur/Microcebus* clade	H	R	Ancestral human amino acid, derived variant in multiple clades
2068	*Callithrix*	K	N	Independent derived variant
2098	*Otolemur*	M	K	Independent variant
	Cebus/Sapajus clade and *Callithrix*	M	T	Derived variant restricted to clade and an independent variant
2218	*Callithrix/Sapajus* excluding *Saimiri*	M	I	Derived variant restricted to clade
	Saimiri	M	V	Independent derived mutation
2262	All genera	I	V	Ancestral variant, human amino acid only present in humans
2289	All genera	Q	H	Ancestral variant, human amino acid only present in humans
	Otolemur	Q	R	Independent variant
2294	*Cebus/Sapajus* clade and *Otolemur*	F	S	Derived variant restricted to clade and an independent variant
	Callithrix	F	L	Independent derived variant

The columns are (left to right): The site of the variant; the name of the genus that deviates from the human sequence; the human amino acid; the variant amino acid; how the variant occurs in relation to the other species in the tree.

ii.) CPDs Isolated to a Single Clade with Two or More Species

The second pattern observed was a single clade with two or more species accumulating the CPD. An example of this is shown in Figure 3 using the protein TTR which had five orthologs belonging to the same clade acquire a CPD. The TTR protein sequence was available for all the 25 primate species with no missing genera. Based on the phylogenetic tree constructed for TTR at site 127, it indicates that the CPDs accumulated are derived. This is apparent as the origin species and all other ortholog sequences including the human sequence contain the wild-type amino acid.

Looking closer at the TTR protein, revealed there were 19 sites that contained a different variant in the *Callithirix/Sapajus* clade represented in Figure 3 when compared to the human sequence. Of these sites, 14 showed the same mutation restricted to this clade indicating it was common for this group of species to accumulate a potential deviation from the human sequence. Whereas, the other five sites showed mutations unrestricted and occuring in multiple ortholog sequences, or sharing the ancestral species and the human allele being derived. The 19 sites identified and the species containing the variant can be seen in Table 7.

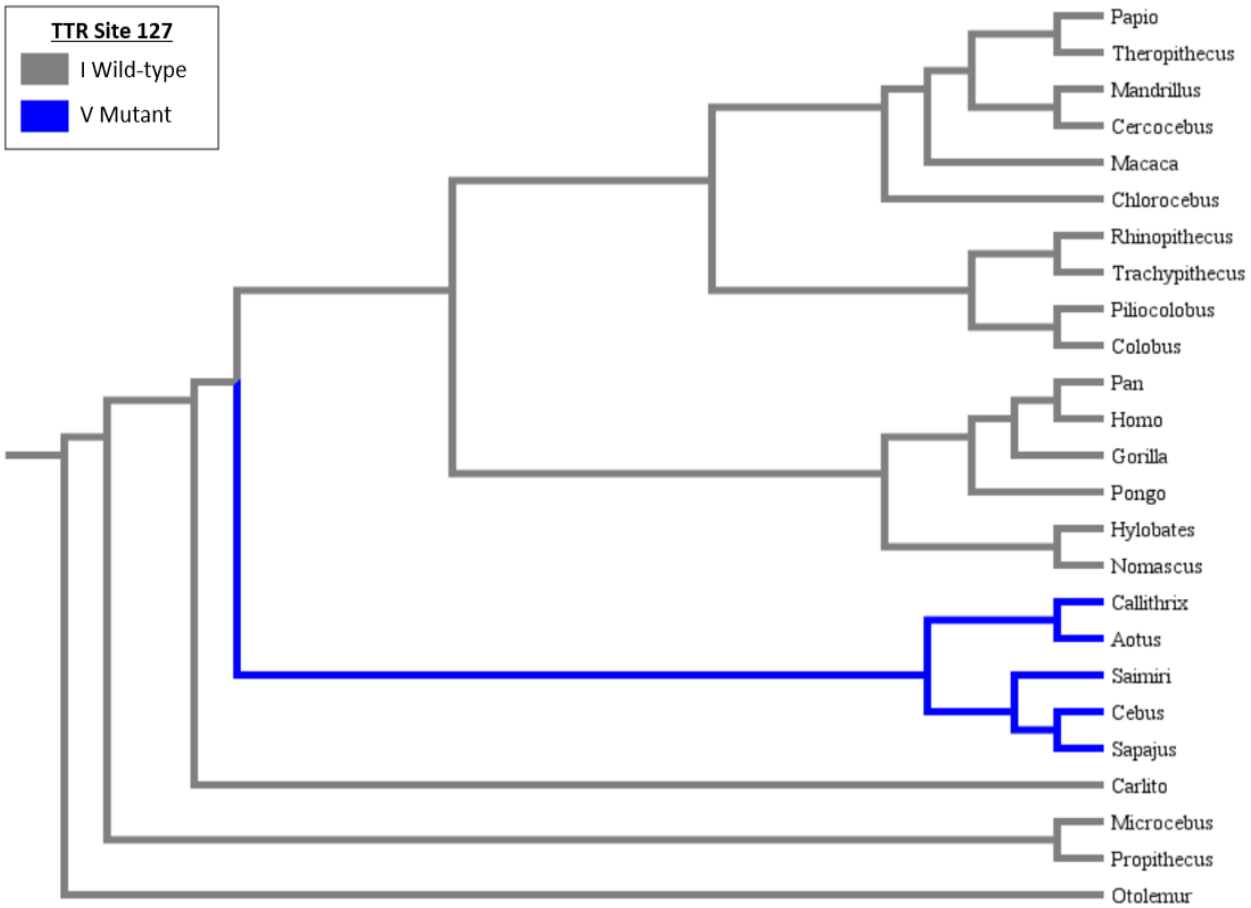

Figure 3. The phylogenetic tree of a derived CPD restricted to a clade. Displaying a CPD in the TTR *Callithrix*/*Sapajus* clade at site 127.

Table 7. The identified sites in the TTR protein that have variants in the *Callithrix/Sapajus* clade that differ from the human sequence

Site	Genera differing from the human sequence	Human amino acid	Variant amino acid	Description of occurrence
5	*Callithrix/Sapajus* clade	R	H	Derived variant restricted to clade
	Microcebus/Otolemur clade	R	G	Ancestral variant restricted to clade
22	*Callithrix/Sapajus* clade	P	H	Derived variant restricted to clade
28	*Callithrix/Sapajus* clade	Y	S	Derived variant restricted to clade
29	*Callithrix/Sapajus* clade	K	S	Derived variant restricted to clade
	Microcebus	K	R	Independent derived variant
41	*Callithrix/Sapajus* clade	R	Q	Derived variant restricted to clade
43	*Callithrix/Sapajus* clade	S	R	Derived variant restricted to clade
51	*Callithrix/Sapajus* clade	H	S	Derived variant restricted to clade
	Microcebus/Otolemur clade	H	K	Derived variant restricted to clade
54	All genera excluding the *Microcebus/Propithecus* clade	R	K	Ancestral variant, human amino acid restricted to clade and Humans
59	All genera excluding *Trachypithecus*	D	E	Ancestral variant, human amino acid only present in Trachypithecus
83	*Callithrix/Otolemur* clade	E	K	Ancestral variant, derived human amino acid
86	*Callithrix/Sapajus* clade	E	K	Derived variant restricted to clade

Table 7. Continued

Site	Genera differing from the human sequence	Human amino acid	Variant amino acid	Description of occurrence
95	*Callithrix/Sapajus* clade	T	S	Derived variant restricted to clade
100	*Callithrix/Sapajus* clade	K	H	Derived variant restricted to clade
	Propithecus	K	T	Independent derived variant
106	*Callithrix/Sapajus* clade	P	S	Derived variant restricted to clade
112	*Callithrix/Sapajus* clade	E	D	Derived variant restricted to clade
124	*Callithrix/Sapajus* clade and *Mandrillus/Cercocebus*	R	H	Derived variant in Callithrix/ Sapajus clade and other genera
126	*Callithrix/Sapajus* clade	T	I	Derived variant restricted to clade
127	*Callithrix/Sapajus* clade	I	V	Derived variant restricted to clade
143	*Callithrix/Sapajus* clade, *Otolemur*, *Colobus*, and *Carlito*	T	S	Ancestral variant and derived in Callithrix/ Sapajus clade And other genera
144	*Callithrix/Sapajus* clade	N	D	Derived variant restricted to clade

The columns are (left to right): The site of the variant; the name of the genera that deviates from the human sequence; the human amino acid; the variant amino acid; how the variant occurs in relation to the other species in the tree.

In addition to TTR, GBA was a gene with multiple validated CPD hits that occurred in one clade. GBA was interesting as it had 8 out of the 26 validated CPD hits identified in this study, and the sites that deviated were relatively well conserved to *Papio, Theropthecus,* and *Cercocebus.* An example of this is shown in Figure 4 using GBA at site 416. There were no GBA protein sequences for *Callithrix*, *Pan*, *Pilicolobus*, and *Mandrillus*, therefore they were not included in the constructed tree. Figure 4 is also representative of sites 227 and 223 as they similarly showed CPDs in those same 3 species. To add to the intrigue of the well conserved clade containing CPDs, 3 of the remaining 5 validated hits showed single variants involving the same 3 species. Site 502 had a CPD in *Papio,* site 351 in *Cercocebus,* and site 324 in *Theropthecus.* Having 6 out of the 8 identified CPDs for GBA all occurring in one clade does not appear to be a coincidence. This sparks the question as to why there were so many observed CPDs, and variants concentrated in the *Papio/Cercocebus* clade and requires further research.

Looking closer at the GBA protein sequence there were 14 sites that *Papio, Theropthecus,* and *Cercocebus* differed from the human sequence. The data highlighted that 5 of the 14 sites had isolated mutations in the 3 species, while the rest of the species including the human sequence had the wild-type amino acid. This represents the highly recorded divergence in these 3 species in the GBA protein sequence. Meanwhile, in the other nine sites the human sequence differed from the rest of the species, or the more recent clades formed after the human sequence contained the mutant allele. The 14 sites identified can be seen more detail in Table 8.

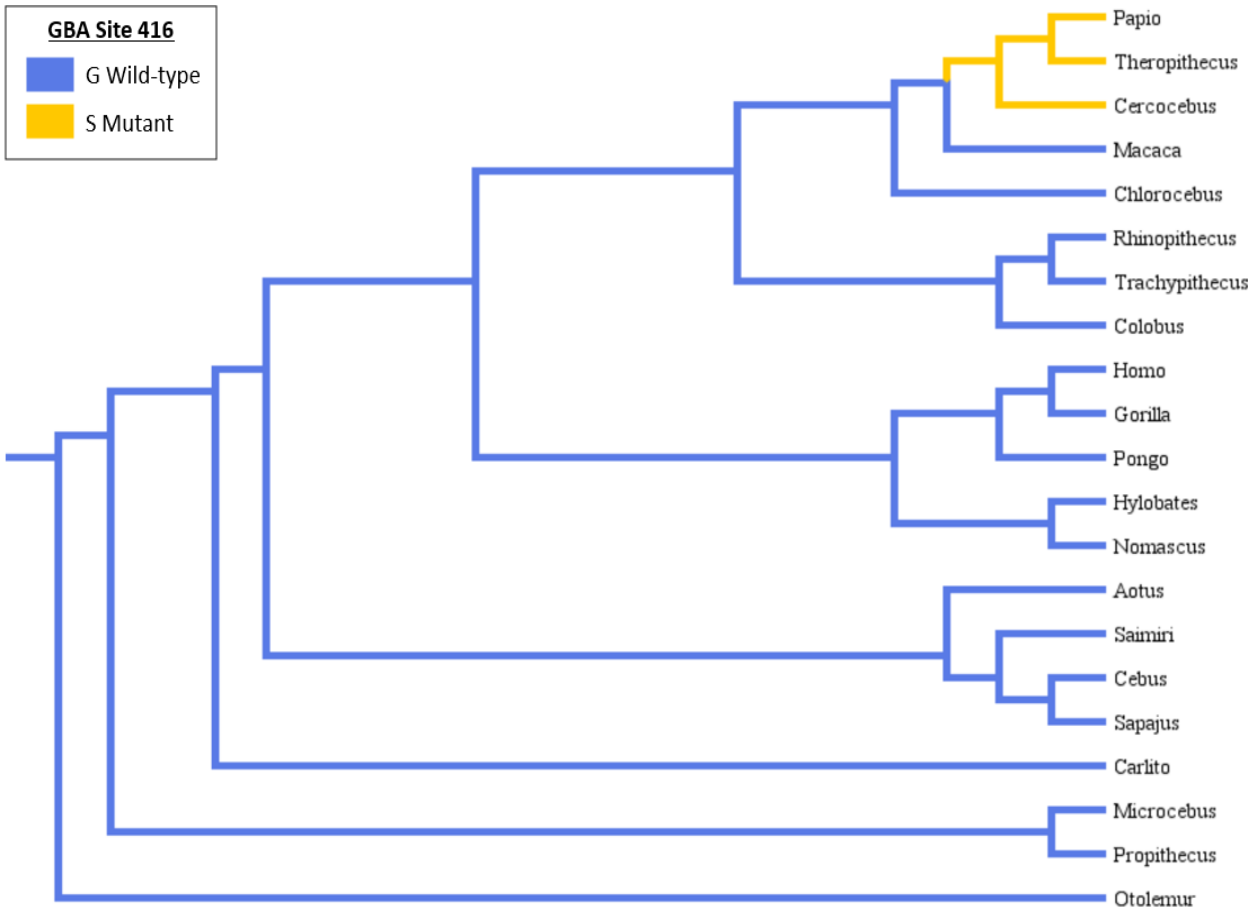

Figure 4. The phylogenetic tree of the GBA protein at site 416. It is also representative of valid CPDs identified at sites 227 and 223.

Table 8. The identified sites in the GBA protein that have variants in the *Papio/Cerocebus* clade that differ from the human sequence

Site	Genera differing from the human sequence	Human amino acid	Variant amino acid	Description of occurrence
23	All genera excluding the *Homo/Nomascus* clade	G	A	Ancestral variant, human amino acid restricted to clade
65	All genera excluding *Gorilla*	F	L	Ancestral variant, human amino acid derived in humans and Gorillas
66	*Papio/Colobus* clade	D	E	Derived variant restricted to clade
68	All genera excluding *Gorilla* and *Nomascus*	P	L	Ancestral variant, human amino acid derived in humans and Gorillas and Nomascus
92	All genera excluding *Gorilla*	M	T	Ancestral variant, human amino acid derived in humans and Gorillas
94	All genera excluding *Hylobates*	P	T	Ancestral variant, human amino acid derived in humans and Hylobates
99	All genera excluding the *Homo/Nomascus* clade and *Otolemur*	H	R	Ancestral variant, derived Homo/ Nomascus clade
205	*Papio/Cerocebus* clade	Q	X	Derived variant restricted to clade
	Microcebus/Otolemur clade	Q	K	Derived variant restricted to clade
216	*Papio/Cerocebus* clade	S	N	Derived variant restricted to clade
226	*Papio/Cerocebus* clade	T	I	Derived variant restricted to clade
230	*Papio/Cerocebus* clade	V	G	Derived variant restricted to clade
234	*Papio/Cerocebus* clade	V	G	Derived variant restricted to clade
441	*Papio/Colobus* clade	I	V	Derived variant restricted to clade
485	All genera excluding *Gorilla*	A	T	Ancestral variant, human amino acid derived in humans and Gorillas

The columns are (left to right): The site of the variant; the name of the genera that deviates from the human sequence; the human amino acid; the variant amino acid; how the variant occurs in relation to the other species in the tree.

iii.) Convergent Evolution of CPDs

The third pattern observed was the convergent evolution of a CPD, which was present in only 1 of the 26 identified CPDs. This was identified in two separate lineages which accumulated a CPD independently of each other. Figure 5 shows PAH having a CPD in the *Callithrix*/*Sapajus* clade and arising independently in *Pongo*. The database did not contain a PAH protein sequence for the *Cebus* genus, therefore *Cebus* was excluded from the tree. Based on the evidence provided by the tree, the convergent generation of CPDs in the PAH protein sequence at site 413 is derived. This is determined as the ancestor species and all the other ortholog sequences including the human sequence contain the wild-type amino acid.

Examining the PAH protein further, there were no derived substitutions with the same convergent distribution including *Pongo.* Once *Pongo* was excluded from the search there were a total of 5 sites that differed between the human sequence and the same clade of species identified at site 413. Of these 5, sites 30 and 363 stood out. At site 30, there is a mutation isolated to the *Callithrix*/*Sapajus* clade with another independent mutation occurring in *Pongo*, while the rest of the species contain the wild-type allele. This is interesting as it is similar to the original CPD involved in independent evolution shown in Figure 5 This highlights the significance of the four genera clade in relation to *Pongo*, a close relative to *Homo sapiens.* The second site 363, similarly found mutations restricted in the *Callithrix*/*Sapajus* clade. The other three sites found variants in relation to the human sequence in the *Pan*/*Hylobates* clade and a couple other previously diverged species. The five sites and the species containing the variant can be seen in Table 9.

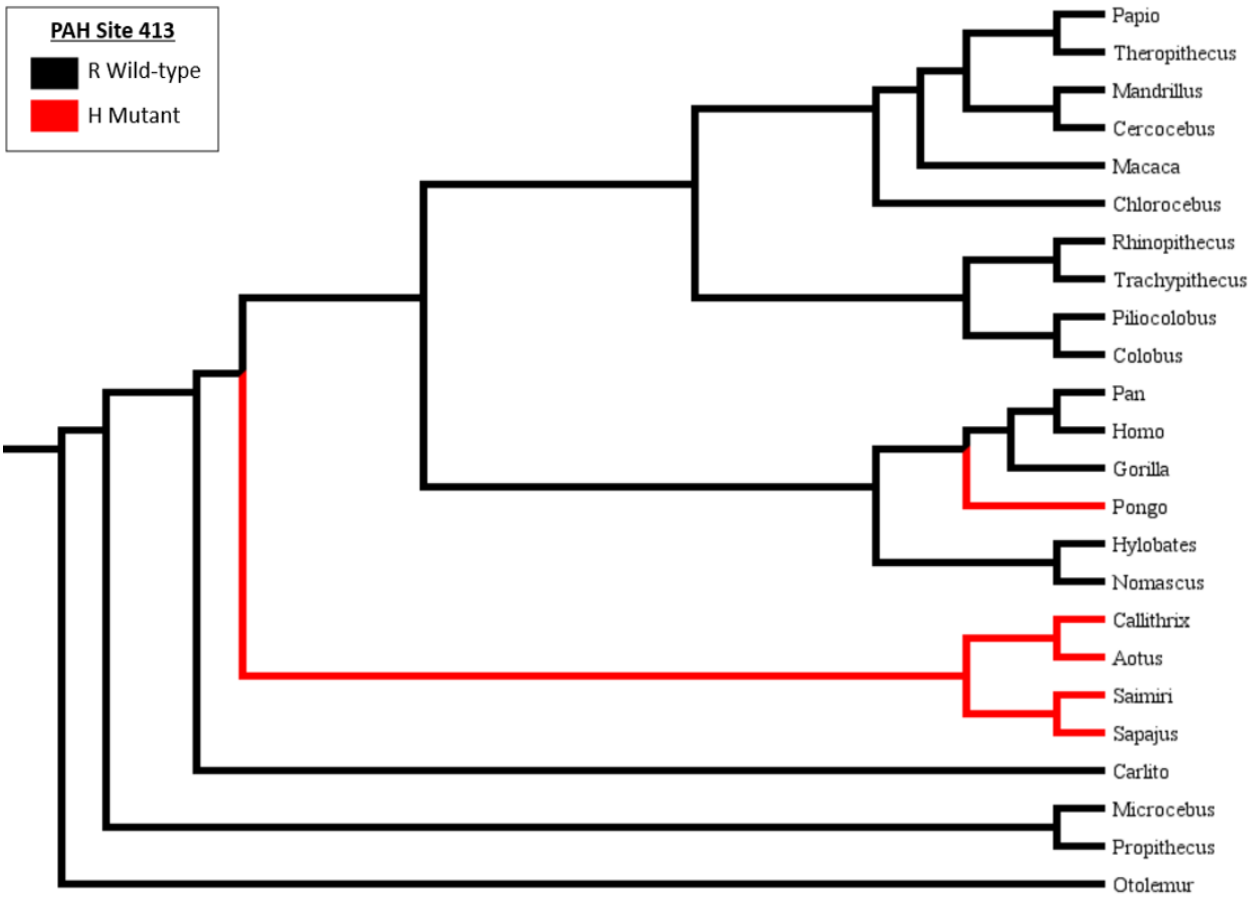

Figure 5. The phylogenetic tree of a derived CPD demonstrating convergent evolution. Displaying identified validated CPDs in the *Callithrix*/*Sapajus* clade and *Pongo* species of the PAH protein at site 413.

Table 9. The identified sites in the PAH protein that have variants in the *Callithrix/Sapajus* clade that differ from the human sequence

Site	Genera differing from the human sequence	Human amino acid	Variant amino acid	Description of occurrence
9	*Callithrix/Sapajus* clade and *Microcebus*	P	R	Derived variant restricted to clade and Microcebus
30	*Callithrix/Sapajus* clade	N	T	Derived variant restricted to clade
	Pongo	N	S	Independent derived variant
180	All genera excluding the *Pan/ Nomascus* clade	M	T	Ancestral variant, human amino acid derived in Pan/ Nomascus clade
363	*Callithrix/Sapajus* clade	M	T	Derived variant restricted to clade
376	All genera excluding the *Pan/ Nomascus* clade	N	K	Ancestral variant, human amino acid derived in Pan/ Nomascus clade

The columns are (left to right): The site of the variant; the name of the genera that deviates from the human sequence; the human amino acid; the variant amino acid; how the variant occurs in relation to the other species in the tree.

iv.) Ancestral CPD

The fourth and final pattern observed showed the ancestral species to contain the CPD variant, while the human wild-type amino acid was later derived. This is shown in Figure 6 using LDLR at site 215 where the origin species contains the variant amino acid, and divergent lineages acquired the human wild-type amino acid. The LDLR protein sequence was available for all the 25 primate genera. The LDLR phylogenetic tree was constructed using ML and the ancestral states of the CPD site were reconstructed using parsimony. By assuming this model is correct it can be assumed that the ancestral state contained the CPD mutant amino acid. To be more confident in the nature of the ancestral state further analysis and a more detailed and statistical based model are required.

The LDLR protein displayed 10 alternative sites in its sequence that had the same ancestral species differing from the human sequence. The most prominent of all the 10 sites was at 552 that showed the *Pan*/*Nomuscus* clade with one variant while the other previously derived species all share the same amino acid at that position. This pattern matches the CPD orientation displayed in Figure 6 The mutant allele being present in the ancestral state indicates that somewhere in the lineage a substitution occurred for an alternative allele to derive the human sequence, but our data set does not provide insight as to what and when that occurred. The other nine sites identified showed variable mutations beginning at the *Pan* genera, spanning multiple previously derived relative species. The 10 sites and the species containing the variant can be seen in Table 10.

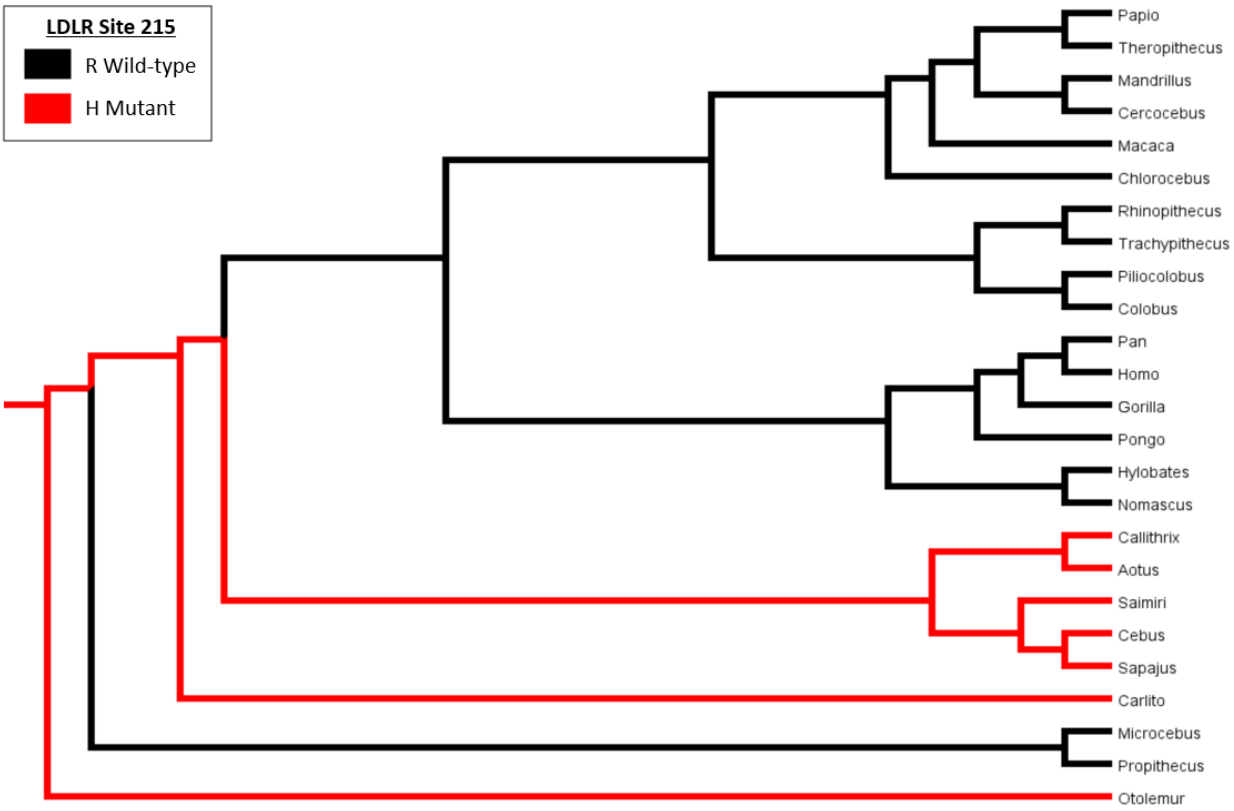

Figure 6. The phylogenetic tree of an ancestral CPD. Displaying the identified validated CPDs in the *Callithrix*/*Sapajus* clade and the *Otolemur* species of the LDLR protein at site 215.

Table 10. The identified sites in the LDLR protein that have variants in the *Callithrix/Carlito* clade and *Otolemur* that differ from the human sequence

Site	Genera differing from the human sequence	Human amino acid	Variant amino acid	Description of occurrence
96	All genera excluding the *Pan/Pongo* clade	D	E	Ancestral variant, human amino acid restricted to clade
328	All genera excluding *Pan/Nomascus* clade and *Rhinopithecus/Colobus* clade	V	I	Ancestral variant, human amino acid restricted to the clades
454	All genera excluding *Pan/Gorilla* clade	C	Y	Ancestral variant, human amino acid restricted to clade
473	All genera excluding *Pan/Gorilla* clade	I	L	Ancestral variant, human amino acid restricted to clade
552	*Callithrix/Otolemur* clade	I	V	Ancestral variant, derived human amino acid
613	All genera excluding *Pan/Gorilla* clade	V	I	Ancestral variant, human amino acid restricted to clade
782	*Callithrix/Otolemur* clade, *Papio*, and *Theropithecus*	K	E	Ancestral variant in clade, Papio, and Theropithecus, derived human amino acid
788	All genera excluding *Pan/Gorilla* clade	V	G	Ancestral variant, human amino acid restricted to clade
806	All other genera	V	A	Ancestral variant, human amino acid only present in humans
817	All genera excluding *Pan/Nomascus* clade and *Microcebus/Propithecus*	S	N	Ancestral variant, human amino acid restricted to the clades

The columns are (left to right): The site of the variant; the name of the genera that deviates from the human sequence; the human amino acid; the variant amino acid; how the variant occurs in relation to the other species in the tree.

3.5 The Correlation of CPDs to Bioinformatic Data

It is not clear what caused the number of CPDs to differ so drastically between genes. It can be hypothesised that factors such as the physiological properties or the rate of evolution occurring in the genes could play a role. To try to determine what factors may be correlated to the occurrence of CPDs, a statistical analysis using sequence data was conducted. By comparing the human reference sequence to a primate ortholog it allowed evolutionary divergence between the two species to be studied and in turn highlights any correlation between CPDs and the data collected in the study. To do so the human reference sequence and *Macaca mulatta* sequence were aligned and the deviance between the two were recorded and further analysed. This involved looking at the length of the sequences, number of gaps between the sequences, the number of changes between the sequences, and the evolutionary distance between the sequences which can be seen in Table 11 This data was then compared to the known ClinVar variants for each of the 32 human proteins and the validated CPD hits found in my study.

The data presented in Table 11 shows that on average the human and *Macaca mulatta* protein sequences deviate at roughly 2.5% of amino acid sites (determined from the distance column). The *Macaca* sequence was selected as the comparison species because it had enough deviations to indicate evolutionary changes from the human sequence, but still shared a large number of homologous sites. As seen in Table 11 the proteins BTK and GJB1 showed zero divergence between the two sequences, which accounts for the lack of CPDs found in these proteins. The higher the "distance" value (the proportion of sites differing between the sequences) of each protein indicated the faster the sequences are evolving from each other. In addition, the distance value, the number of ClinVar variants and length of the sequence provided potential relations with CPDs that required statistical analysis.

Table 11. The correlation between protein sequence data and CPDs

Protein	Length	Gaps	Changes	Distance	ClinVar	CPDs
ABCD1	745	0	18	0.02346	134	0
ALPL	524	0	14	0.025	76	0
AR	920	26	34	0.00984	132	0
ATP7B	**1465**	**0**	**49**	**0.038**	**185**	**1**
BTK	659	0	0	0	102	0
CASR	1078	0	16	0.01694	109	0
CBS	551	0	17	0.03191	68	0
CFTR	**1480**	**0**	**26**	**0.01838**	**391**	**1**
CYBB	570	0	19	0.03577	55	0
F7	444	0	30	0.07105	26	0
F8	**2351**	**0**	**116**	**0.05246**	**269**	**1**
F9	**461**	**0**	**13**	**0.02611**	**138**	**1**
G6PD	545	0	5	0.00794	53	0
GALT	**379**	**0**	**10**	**0.02431**	**174**	**2**
GBA	**536**	**3**	**15**	**0.02636**	**103**	**8**
GJB1	283	0	0	0	111	0
HBB	147	0	8	0.05604	127	0
HPRT1	218	0	1	0.00447	46	0
IL2RG	369	0	10	0.0271	55	0
KCNH2	1159	0	8	0.00728	206	0
KCNQ1	676	0	14	0.01977	210	0
L1CAM	1257	0	7	0.00586	54	0
LDLR	**860**	**0**	**51**	**0.06051**	**808**	**6**
MPZ	248	0	3	0.01051	97	0
MYH7	1935	0	23	0.01163	270	0
OCA1	529	0	22	0.04068	72	0
PAH	**452**	**0**	**9**	**0.0228**	**358**	**2**
PMM2	246	0	4	0.01434	56	0
RHO	348	0	6	0.01605	105	0
TP53	393	0	17	0.0437	241	0
TTR	**147**	**0**	**9**	**0.05843**	**77**	**4**
VWF	2813	0	103	0.03521	149	0

The proteins with validated CPD are bolded in red. The data represented in the Length, Gaps, Changes and Distance columns are comparing the human sequence to the *Macaca mulatta* sequence. The last two columns are the data collected previous in the study. The columns are (left to right): name of the protein; length of the sequences; the number of gaps between the sequences; the number of changes between the sequence; the proportion of sites differing between the sequences; the number of ClinVar variants for each protein; the number of validated CPDs identified in this study

The Spearman's rank correlation was conducted to analyze the number of CPDs and the possible correlates: length of the protein, evolutionary distance, number of pathogenic variants. The results of this indicated two statistically significant correlations for CPDs/ClinVar $\rho = 0.458$ ($p < 0.01$) and CPDs/distance $\rho = 0.383$ ($p < 0.05$). While the correlation between CPDs and length was nonsignificant. The analysis also showed that the ClinVar number is correlated with length, but it is not significantly correlated with distance. Therefore, Clinvar and distance are independent predictors of the number of CPDs.

The impact of the statistically significant correlations can be seen in Table 11 as LDLR and TTR had two of the highest distance values and showed multiple valid CPDs. Yet the distance value is not completely definitive as HBB shows a similarly high distance value but no valid CPDs. Considering the high distance value, the lack of CPD hits in HBB could potentially be a result of the number of gaps present between the human and *Macaca* sequence or the amount of ClinVar variants present. HBB is shown to have 127 pathogenic ClinVar variants which seems relatively average but not significant. Whereas GBA was found to have 808 pathogenic variants and found 8 validated CPDs. In addition, CFTR, PAH, and F8 have the next highest amount of ClinVar variants, and all had valid CPDs found in their sequences. This statistical analysis sheds light on the relationship between CPDs and bioinformatic data but is not completely conclusive, therefore further analysis is required.

DISCUSSION

4.1 How CPDs Evolve

The data collected from this study indicated 26 valid CPDs and showed four prominent patterns of how they arose: in a single species, isolated to a single clade, a convergent evolution, and ancestral variant with a later derived human wild type. The nature of these patterns can be better explained by understanding the different ways in which CPDs are produced in a lineage. There are roughly three scenarios, two with the human sequence amino acid as the ancestral state and one with the variant amino acid as the ancestral state. All three scenarios contain a compensatory mutation somewhere in the sequence that determines whether the change between the human and variant is deleterious. The three scenarios can be seen in Figure 7.

The scenario containing the variant amino acid at the ancestral state is identified as Case A. This scenario occurs when the compensatory mutation occurs after the variant amino acid changes to the human wild-type amino acid. The remaining two cases, B and C, have the human amino acid as the ancestral state, with a compensatory mutation and a change to the variant. Case B occurs when the ancestor has the same amino acid as modern humans such that the change to the variant amino acid is deleterious. This evolution of the variant requires a compensatory change somewhere else in the sequence that will compensate for the deleterious effects. Whereas Case C also carries the human amino acid at the ancestral state, but a change to the variant is not deleterious. In this case the compensatory change occurs in the human lineage and as a result makes the variant mutation deleterious.

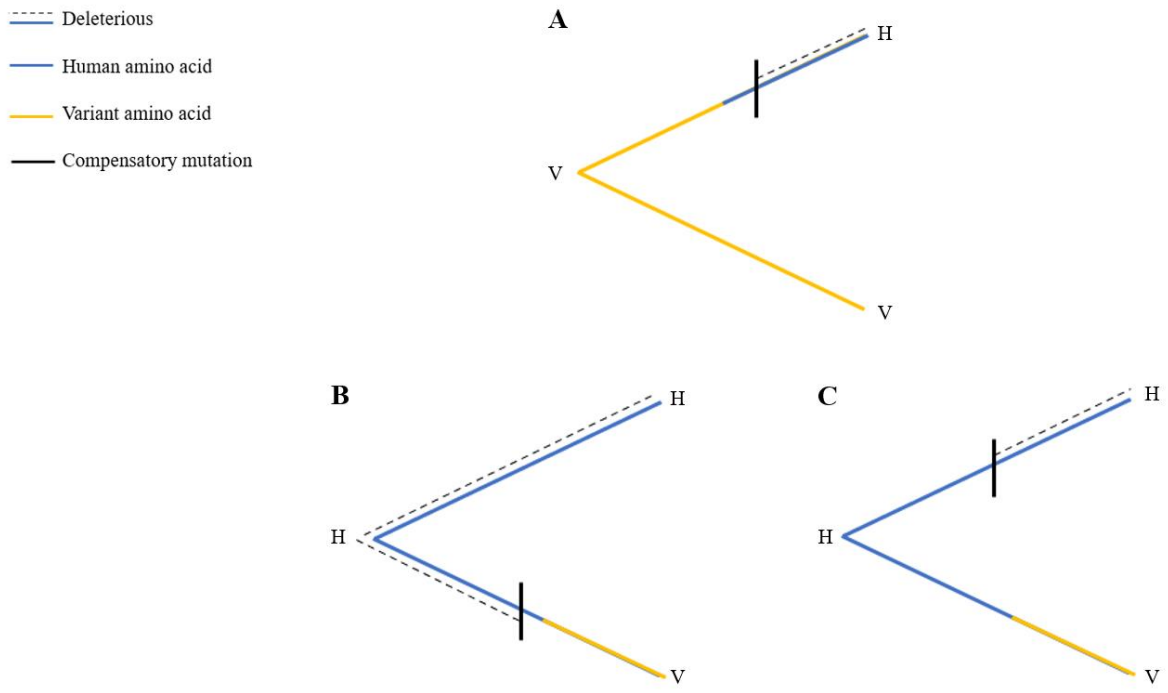

Figure 7. The three scenarios in which CPDs arise. The solid black line indicates where the compensatory mutation occurs. The dashed black line indicates if the variant amino acid were to occur in that same location in the human lineage it would be deleterious (V= variant amino acid, H= human amino acid) A.) In Case A the variant amino acid is at the ancestral state and the compensatory mutation occurs after the change to the human wild-type amino acid occurs B.) In Case B the human wild-type amino acid is in the ancestral state and the compensatory mutation accompanies the change to the variant amino acid C.) in Case C the ancestral state is the human wild-type amino acid, and the compensatory mutation occurs in the human sequence.

The sequence data from the study identifies the CPDs in each species but does not give information as to when in time these mutations and substitutions occurred in the lineage. One example that can be classified as Case A is LDLR. As seen in Figure 6 LDLR at site 215 showed the variant as the ancestral state that had changed in the human sequence. We can observe this change from the data but cannot pinpoint the exact location on the tree where the compensatory mutation occurred. Therefore, it cannot be determined whether the mutation happened before or after the change from the ancestral variant to the human amino acid.

In turn, there are several instances of validated CPDs that follow Case B and C. Specifically Case B, which presumably appears at the tips or ends of the trees. This can be seen in our single species CPDs that are isolated such that the change to the variant amino acid would be deleterious without the compensatory mutation accompanying it. Case C requires more in-depth data analysis to accurately classify, as it is difficult to determine where in the lineage the compensatory mutation exactly occurs. It can be assumed that Case C will lead to easier evolution of the variant site as it is not deleterious without the accompanying compensatory mutation found in the human sequence. The remaining two patterns of CPDs noted in the results can then be categorized as either Case B or Case C, as they have the human sequence as the ancestral state and the variant amino acid is derived later in the primate phylogeny. A more detailed analysis as to when and where the compensatory mutation occurred in each example would be needed to accurately determine the proper case classification for the single clade harboring a CPD and the independent evolution of CPDs from different clades.

4.2 The Degree of Pathogenicity

The 32 genes used in this study were chosen as they are well studied and known to be responsible for Mendelian disease. Therefore, the pathogenic nature of the variant amino acid when uncompensated is significant as it is associated with severe disease and disorders in the human sequence. By merely looking at the gene name abbreviations used in this study, it is unclear as to what these proteins are responsible for in the human body. To address this, the eight proteins found to have valid CPDs were looked at in more detail from the physiological perspective. The full name and the common disease associated with the pathogenic mutation was retrieved from NCBIs (https://www.ncbi.nlm.nih.gov/) and can be seen in Table 12.

The glimpse at the 8 proteins and the severity of the pathogenic mutant in these sequences helps to better understand the significance of these proteins and their direct relation to the human body. The proteins shown in Table 12 a quarter of the total number of proteins reviewed in this study but highlights how detrimental these uncompensated pathogenic mutations are to humans. This shows the importance of understanding the compensatory mutations developed in the primate ortholog sequences as they do not appear to have adverse effects when containing the pathogenic variant. The severity of these mutations could also be a potential explanation as to why certain genes contain CPDs and others do not. It is possible that the genes lacking CPDs indicate that a mutation in the sequence is extremely detrimental and therefore unable to be recorded. This hypothesis would also be applicable to the sequences that show low evolutionary divergence in Table 11, as any divergence would have severe deleterious effects and not allow the organism to survive to be studied. More research into the degree of pathogenicity and the physiological effects of each of these proteins in humans and primates would help explain the nature of CPDs and the frequency at which they occur.

Table 12. The Disease/Disorder associated with a mutation in the human proteins found to have validated CPDs

Protein	Full Name	Associated Disease/Disorder
ATP7B	ATPase copper transporting beta	Wilson disease
CFTR	CF transmembrane conductance regulator	Cystic fibrosis
F8	Coagulation factor VIII	Hemophilia A
F9	Coagulation factor IX	Hemophilia B or Christmas disease
GALT	Galactose-1-phosphate uridylyltransferase	Galactosemia
GBA	Glucosylceramidase beta	Gaucher disease
LDLR	Low density lipoprotein receptor	Familial hypercholesterolemia
TTR	Transthyretin	Amyloid deposition

The columns are (left to right): the protein abbreviation, the full name of the protein, the associated disease/disorder when there is a mutation in the protein.

4.3 Parsimony

The phylogenetic tree was used in this study were developed using ML and the ancestral states of the CPD site were reconstructed using Wagner parsimony. This model was a good starting point for this thesis project, but there are problems with it that can be better addressed using other methods. Therefore, to be able to confidently confirm the conclusions drawn from the data collected from this study, a more complex model is required. Using a model with a statistical component would help to identify the confidence interval of each branch containing the particular amino acid under review in the protein sequence. This would help to confirm the certainty of the origin species and distinguish areas on the tree where the change between the human wild-type and variant amino acid occurs. By doing so it would help clarify the nature of the CPDs that appear to have the variant amino acid as the ancestral state. This would help strengthen the accuracy of the conclusion of this study and provide a solid foundation moving forward and expanding upon the work done.

4.4 Structural Analysis

The structures of PAH and TTR were reviewed to gain insight on the validated CPD sites and neighbouring amino acids in the protein sequences. These two proteins were chosen as they had full length protein structures that were well conserved. The protein structures were retrieved from RCSB Protein Data Bank (http://www.rcsb.org). Figure 8 shows the two protein structures and the location of the each CPD highlighting the neighboring sites that interact with the amino acid at the site. The two proteins had hydrogen bonds with amino acids in close proximity but did not provide any obvious indication of the nature or whereabouts of the potential accompanying compensatory mutation. The candidate substitutions identified in Figure 8 showed parallel

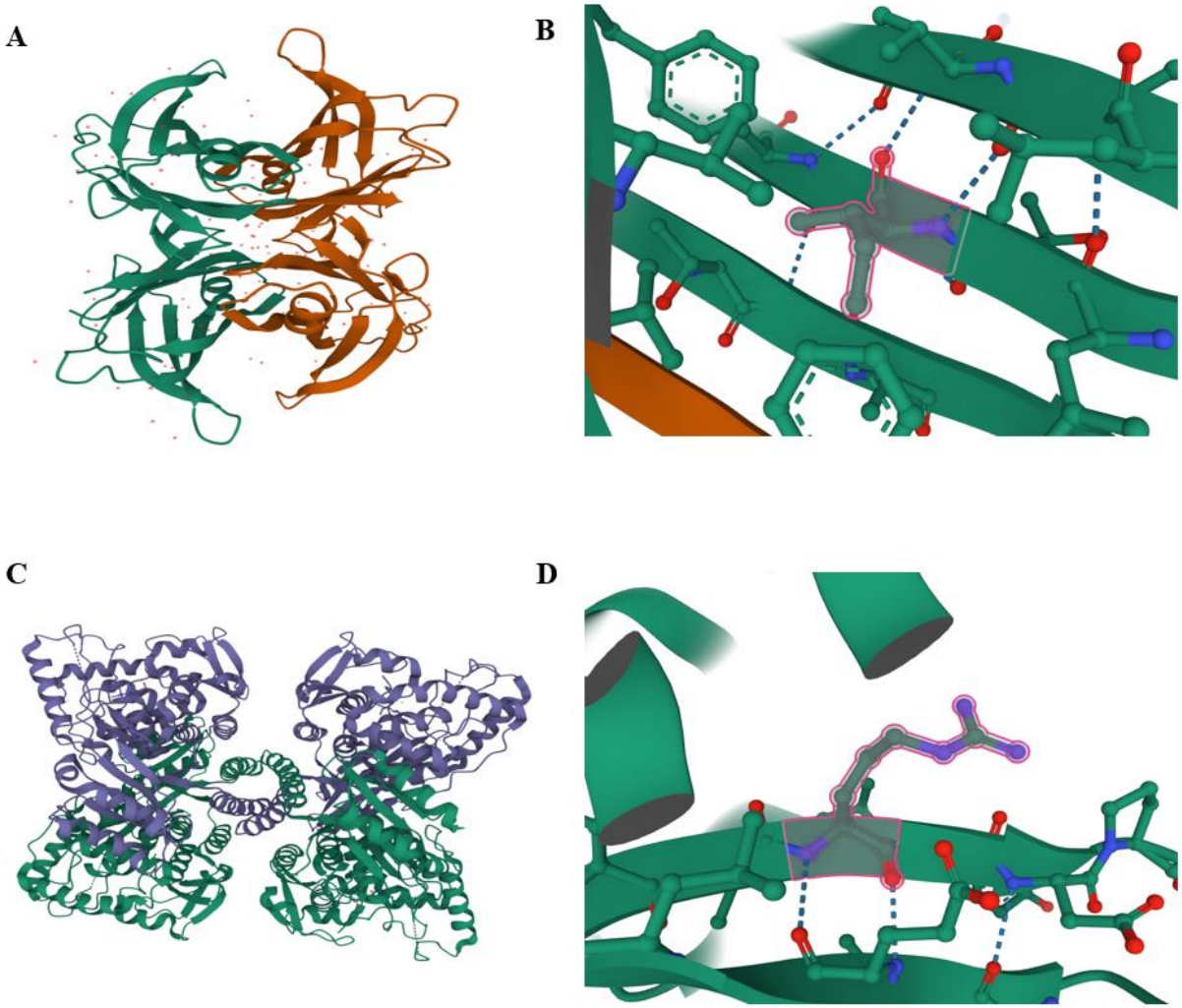

Figure 8. The protein structure of PAH and TTR and their associated sites containing a validated CPD. All the 3D structures and images were retrieved from RCSB PDB (http://www.rcsb.org) and created using Mol* (Sehnal, D., 2021) A.) The full-length protein structure of TTR (PDB ID: 1DVQ) B.) The I wild-type amino acid at site 127 of the TTR protein sequence is outlined in pink. The dashed lines indicate hydrogen bonds attaching the amino acid at site 127 to the amino acids at site 13 and 15 C.) The full-length protein structure of PAH (PDB ID: 6HYC) D.) The R wild-type amino acid at site 413 of the PAH protein sequence is outline in pink. The adjacent dashed lines representing two hydrogens bonds with site 422.

patterns of evolution but were not clearly interacting with the CPD sites. More in-depth research and analysis would be required to identify the region surrounding the CPD site and evaluate the neighbouring amino acid interactions. This issue parallels the case classification analysis, in that the sequence data collected needs to be further expanded on to achieve a more precise explanation and understanding of where and how the compensatory mutations interact with the CPD site.

4.5 Comparison to Kondrashov et al. *(2002)*

The Kondrashov *et al.* (2002) study published a sophisticated analysis of their findings that made the study itself seem relatively simple. This senior thesis project aimed to reproduce the Kondrashov *et al.* (2002) study with restricted parameters and updated bioinformatic databases. In doing so, it became apparent that the presentation of the Kondrashov *et al.* (2002) study was potentially oversimplified as there were certain discrepancies between the two.

The wide range of species Kondrashov *et al.* (2002) used made it challenging for conclusions to be drawn due to the vast diversity of orthologs used. Kondrashov *et al.* (2002) did not clearly outline whether all the animal orthologs were being used to acquire their CPD data, or if some classes of animals were only used to map out phylogenetic distance from humans (this can be seen in Figure 3 of the Kondrashov *et al.* (2002) paper. The opaque explanation of the methods and ways in which the study was conducted impacted the analysis and the main generalization made by Kondrashov *et al.* (2002). This impact can also be seen in the structural analysis conducted in which it was able to identify the compensatory mutation upon basic structural review. Lacking the explanation of how he came to analyse the protein structure, it challenged the nature of his findings and made it harder to replicate. The structural analysis done in this study shows it was not as easily achieved and would require more resources to evaluate and identify the region and amino acids surrounding the CPD site.

To sum up the overall findings and conclusion of the paper, Kondrashov *et al.* (2002) stated that roughly 10% of deviations of nonhuman protein sequences from the human ortholog sequence are CPDs (Kondrashov *et al.*, 2002). This statement is then supposedly applicable to the large range of species he reviewed and in turn can be widely generalized across different populations. As this senior thesis study only looked at primates, it is difficult to make a comparative statement as the parameters differed significantly. This thesis project did find multiple CPDs, while following a similar experimental set up as Kondrashov *et al.* (2002), but to adequately break down this statement and access the accuracy of the findings in the Kondrashov *et al.* (2002) paper further analysis would be required. This could be done by conducting additional studies opening the parameters of species analysis and then being able to compile all the data and see if it is congruent with the work of Kondrashovs *et al*. (2002).

While comparing the data collected in this study and analyzing the Kondrashov *et al.* (2002) paper one central question raised was, can one predict how many CPDs are present by looking at the nature and amount of pathogenic data available for each protein? Both the statistical analysis and the data presented in Table 11 indicates there is a degree of independent correlation between the occurrence of CPDs/evolutionary difference between sequences and CPDs/amount of pathogenic variant data. These statistically significant relationships shed light on the occurrence of CPDs, but do not comprehensively explain the exact nature of the relationship. To accurately answer the question more research must be conducted. It is important to note that the ClinVar database is constantly being updated, therefore the next person to build upon this study will have to compare it to the latest data available in the database. This comparison will help understand any changes that may have occurred.

CONCLUSION

As stated by Kondrashov *et al.* (2002), data on pathogenic mutations provide valuable information on deleterious amino acids and with the addition of structural data it helps to reveal the molecular basis of compensatory substitutions (Kondrashov *et al.*, 2002). This study found 26 valid CPD hits from reviewing 32 human proteins related to Mendelian disease and their primate ortholog sequences. This is significant as the parameters of this study were restricted compared to the Kondrashov *et al.* (2002) paper, and yet a sizable number of CPDs were found to be valid and allowed for analysis. From the results of the study, there were four prominent patterns of how CPDs arose in protein sequences, which paralleled the different Cases (A, B, and C) of CPD origin. To better understand the nature of CPDs, the exact location of where the compensatory mutation occurs is needed. To do this more in-depth research is required. The sequence data collected in this study was sufficient to identify CPD hits in the ortholog protein sequences and in turn start the evaluation of the Kondrashov *et al.* (2002) paper.

To build off the work done in this study, the next potential step would be to widen the parameters of orthologs used and potentially screen more proteins. As mentioned previously, the Kondrashov *et al.* (2002) paper looked at various species within the animal kingdom. Therefore, moving forward from this study, the orthologs analyzed could be expanded to include all mammals. After broadening the parameters, the analysis could be repeated and would allow for better observation of the phylogenetic distance between the different species. This could then be done until the same parameters of the Kondrashov *et al.* (2002) study were achieved, and a true comparison and evaluation could be conducted.

An additional route in following up to this project would be to increase the research regarding primate genome sequencing. The human genome has been studied at length, but the genome sequences of primates is less well-known. By gaining a better understanding of primate's genome it would help to identify where in the sequence the compensatory mutations occurred and provide clarity on the nature of the CPDs identified in this study.

BIBLOGRAPHY

Barešić, A., and A. C. R. Martin. 2011. Compensated pathogenic deviations. BioMolecular Concepts 2:281–292.

Barešić, A., L. E. M. Hopcroft, H. H. Rogers, J. M. Hurst, and A. C. R. Martin. 2010. Compensated pathogenic deviations: analysis of structural effects. Journal of Molecular Biology 396:19–30.

Darwin, C., 1859 On the Origin of Species. John Murray, London

Dobzhansky, T., 1933. On the sterility of the interracial hybrids in *Drosophila pseudoobscura*. The Proceedings of the National Academy of Sciences of the United States of America 19:397–403.

Dobzhansky, T., 1934 Studies on hybrid sterility. I. Spermatogenesis in pure and hybrid *Drosophila pseudoobscura*. Zeitschrift für Zellforschung und Mikroskopische Anatomie 21: 169-221.

Dobzhansky, T., 1936. Studies on hybrid sterility. II. localization of sterility factors in *Drosophila pseudoobscura* Hybrids. Genetics 21:113–135.

Dobzhansky, T., 1937. Genetics and the origin of species. Columbia Univ. Press, New York.

Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Felsenstein, J. 1983. Parsimony in systematics: biological and statistical issues. Annual Review of Ecology & Systematics 14:313–333.

Finstermeier, K., D. Zinner, M. Brameier, M. Meyer, E. Kreuz, M. Hofreiter, and C. Roos. 2013. A mitogenomic phylogeny of living primates. PLoS ONE 8: e69504.

Joy, J. B., R. H. Liang, R. M. McCloskey, T. Nguyen, and A. F. Y. Poon. 2016. Ancestral reconstruction. PLoS Computational Biology 12:e1004763

Kern, A. D., and F. A. Kondrashov. 2004. Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. Nature Genetics 36:1207–1212.

Kimura, M., B. C. Clarke, A. Robertson, and A. J. Jeffreys. 1986. DNA and the neutral theory. Philosophical Transactions of the Royal Society of London. B, Biological Sciences 312:343–354.

Knies, J. L., K. K. Dang, T. J. Vision, N. G. Hoffman, R. Swanstrom, and C. L. Burch. 2008. Compensatory evolution in RNA secondary structures increases substitution rate variation among sites. Molecular Biology and Evolution 25:1778–1787.

Kondrashov, A. S., S. Sunyaev, and F. A. Kondrashov. 2002. Dobzhansky–Muller incompatibilities in protein evolution. The Proceedings of the National Academy of Sciences of the United States of America 99:14878–14883.

Kulathinal, R. J., B. R. Bettencourt, and D. L. Hartl. 2004. Compensated deleterious mutations in insect genomes. Science 306:1553–1554.

Landrum, M. J., J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research 42: D980-985.

Muller, H. J., 1942 Isolating mechanisms, evolution, and temperature. Biological Symposia. 6: 71-125

Orr, H. A. 1996. Dobzhansky, Bateson, and the genetics of speciation. Genetics 144:1331–1335.

Perelman, P., W. E. Johnson, C. Roos, H. N. Seuánez, J. E. Horvath, M. A. M. Moreira, B. Kessing, J. Pontius, M. Roelke, Y. Rumpler, M. P. C. Schneider, A. Silva, S. J. O'Brien, and J. Pecon-Slattery. 2011. A molecular phylogeny of living primates. PLoS Genetics 7: e1001342.

Richards, S., N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H. L. Rehm, and ACMG Laboratory Quality Assurance Committee. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine 17:405–424.

Sehnal, D., S. Bittrich, M. Deshpande, R. Svobodová, K. Berka, V. Bazgier, S. Velankar, S. K. Burley, J. Koča, and A. S. Rose. 2021. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Research 49: W431–W437.

Turelli, M., and H. A. Orr. 2000. Dominance, epistasis, and the genetics of postzygotic isolation. Genetics 154:1663–1679.

Turelli, M., N. H. Barton, and J. A. Coyne. 2001. Theory and speciation. Trends in Ecology & Evolution 16:330–343.

Wang, R. J., C. Ané, and B. A. Payseur. 2013. The evolution of hybrid incompatibilities along a phylogeny. Evolution 67:2905–2922.