Frequent branch-specific changes of protein substitution rates in closely related lineages

Neel Prabh[1,*] and Diethard Tautz[1]

1 Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön, Germany

* Corresponding author: prabh@evolbio.mpg.de

Running title : Molecular clock is a property of gene aggregates

Keywords : Protein divergence, Molecular clock, Rapid evolution, Lineage-specific

Abstract

Since its inception, the investigation of protein divergence has revolved around a more or less constant rate of sequence information decay that led to the formation of the molecular clock model for sequence evolution. We use here the classical approach of amino acid sequence comparisons to examine the overall divergence of proteins and the possibility of lineage-specific acceleration. By generating and analysing a high-confidence dataset of 13,160 syntenic orthologs from four ape species, including humans, we found that only less than 1% of the ortholog families are entirely in line with the clock model in each of their branches. The most common departure from the expected decay rate involves higher than expected substitutions on just one or two branches of the individual families. However, when taken as aggregate, even a small set of families conform well with the clock assumptions. We identified ADCYAP1 as the most divergent human protein-coding gene with 10% human-specific substitutions. Such lineage-specific highly accelerated genes were not limited to humans but appear as a general pattern that accompanies the formation of species. Our analysis uncovers a much more dynamic history of substitution rate changes in most protein families than usually assumed. Such fluctuations can result in bursts of rapid acceleration followed by periods of strong conservation that effectively cancel each other. Although this gives an impression of a long-term constant rate, the actual history of protein sequence evolution appears to be more complicated.

Introduction

The idea of a constant divergence of proteins over time has existed since the initial investigations into protein divergence, which started with the examination of serological evidence followed by the analysis of haemoglobin homologs (Zuckerkandl and Pauling 1962; Nuttall 1904). Refinement of this idea then led to the formulation of the molecular clock hypothesis of a more or less constant decay of sequence information in genes over evolutionary time (Zuckerkandl and Pauling 1965; Takahata 2007). The effect was ascribed to the assumption that changes in amino acids of a protein would be mostly in functionally nearly neutral positions and this fostered the development of the neutral theory of evolution (Kimura 1968). However, examples that violated the molecular clock pattern were also identified early on, initially in haemoglobin itself (Goodman et al. 1975). Based on an extended sampling, Goodman et al. noted ".. in contradistinction to conclusions on the constancy of evolutionary rates, the haemoglobin genes evolved at markedly non-constant rates", pointing out that phases

of adaptation can lead to a lineage-specific change of substitution rates. Still, in cumulative studies across many genes, the molecular clock pattern is often supported and is nowadays systematically used to compile divergence times for the tree of life (Kumar et al. 2017). To study selection patterns in proteins, the comparison of the number of nonsynonymous substitutions (dN, amino acid altering) to the number of synonymous substitutions (dS, silent) has become customary, with the assumption that dS evolves under neutral rates (Schmid and Aquadro 2001; Clark et al. 2003; Albà and Castresana 2005; Magalhães et al. 2007; Larracuente et al. 2008; Cai and Petrov 2010; Toll-Riera et al. 2011; Johnson et al. 2014; Prabh and Rödelsperger 2016; Prabh et al. 2018; Guillén et al. 2019). However, several studies have raised doubt on this assumption (Parmley and Hurst 2007; Chamary et al. 2006; Hurst 2009, 2011; Plotkin and Kudla 2011; Wang et al. 2011).

Our study revisits the original approach, which utilised amino acid sequence comparisons to study the question of the overall divergence rates of proteins and the possibility of lineage-specific acceleration. Despite the tremendous growth of the databases in the past decades, this fundamental analysis, which drove the concept discussions more than 50 years ago, has been mostly neglected. With the availability of large datasets, alignments of protein-sequences became automatised to handle such comparisons efficiently, while accepting that this creates noise in case of suboptimal alignments around indels or highly diverged regions (Cantarel et al. 2006; Talavera and Castresana 2007; Lunter et al. 2008; Wong et al. 2008; Markova-Raina and Petrov 2011; Ebersberger and von Haeseler). Hence, any attempt to go back to the original approach requires alignment optimisation to make them credible. Further, full genome data have shown that the problem of misalignments between duplicated copies of the genes can be a major impediment and needs to be systematically addressed (Forslund et al. 2018; Glover et al. 2019).

We present here a highly curated dataset using four species from the ape phylogeny, including humans to revisit the question of decay rate constancy for a large part of the known coding sequences of these genomes. Surprisingly, we find that at least among the shallow branches in this phylogeny, there is a branch-specific departure from rate constancy for most of the analysed genes. Only 117 out of 12,046 diverging families are fully in line with a clock model, although when taken in aggregate, they conform well with clock assumptions. We conclude that

there is a high probability of acceleration and deceleration of substitution rates for most genes, even at short evolutionary time scales.

Results:

Synteny guided ortholog identification

To study lineage-specific divergence at the amino acid residues level, we needed to identify orthologous proteins in the extant species. To ascertain this, we chose four ape species: human, chimpanzee, gorilla and gibbon (Figure 1a). They have a well-documented evolutionary history, and their overall genome divergence is sufficiently small to ensure unambiguous alignments of proteins. Furthermore, the human genome is among the best-curated genomes and serves, therefore as a reliable reference for comparisons. Orthologs of the genes annotated in humans were identified by a combination of reciprocal best BLAST hits and the pairwise analysis of gene order (Figure 1b). Among the 19,976 annotated human genes in the study, we found 15,560 syntenic with chimpanzee, 15,007 with gorilla, and 14,162 with white-cheeked gibbon (Figure 1c). Note that there are large scale chromosomal rearrangements in the gibbon genome, but at the smaller-scale, it is largely comparable with the other apes (Carbone et al. 2014). In total, we found 13,163 ortholog gene families shared between the four species, which represents about two-thirds of the annotated human genes. Note that the failure to identify definite orthologs for the remainder of the genes is mostly due to duplication patterns that could not be fully resolved based on our strict filtering criteria (see Methods). Still, this constitutes the most abundant gene set comparison analysed for these species so far.

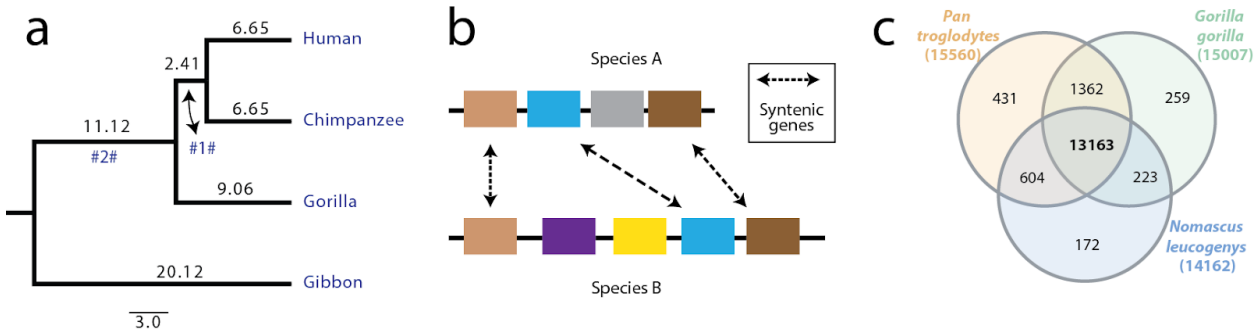

Figure 1: Syntenic orthologs. a) Phylogeny of apes adapted from www.timetree.org (Kumar et al. 2017), branch length is divergence time in mya. #1# and #2# are the two internal branches. b) Collinear or syntenic genes between two species A and B. c) Venn diagram depicting the overlap of *Homo sapiens* genes syntenic with the other three apes.

Optimised alignment

The protein divergence can be due to amino acid substitutions, but also changes in the length of the reading frames, either due to new start/stop codons or inclusion/exclusion of exons. Therefore, we have analysed how far these latter factors influence our gene set by examining the length variation between the longest and shortest orthologs from each family (Figure 2a). One-third of the ortholog families (N=4,397) had no length variation with each having the exact same length, another one-third (N=4,446) had longest orthologs that were less than 5% longer than the shortest orthologs (Figure 2b). This confirmed that the majority of ortholog families in our analysis were made from proteins that do not show considerable variation in their lengths.

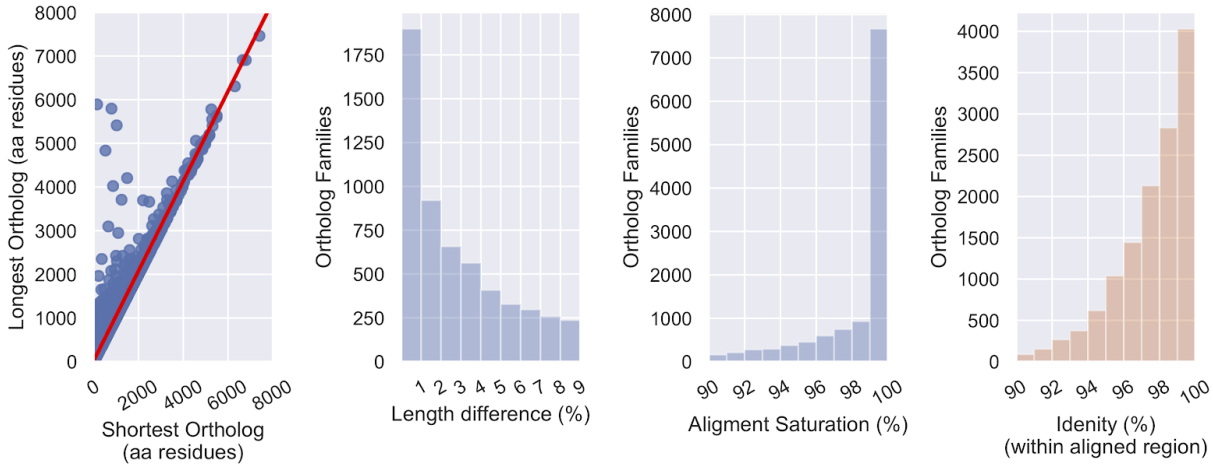

Figure 2: Ortholog families. a) Scatter plot of the shortest and the longest orthologs of each ortholog family. The regression line is drawn in red. b) Histogram showing ortholog family distribution for maximum length difference per 100 residues of the shortest ortholog. c) Histogram showing ortholog family distribution for alignment saturation. d) Histogram showing ortholog family distribution for identical sites per 100 aligned sites. Lower limits were excluded from the bins in panels b, c, and d.

The overall length similarity of the orthologs allowed us to employ stringent criteria (zero gap tolerance and low contiguous substitution threshold) for creating multiple sequence alignments from these families. Only three ortholog families did not show any overlap in the final alignment due to the presence of truncated genes (Supplemental file); we removed these from further analysis. Thirty-eight ortholog families shared less than 50 residue overlap, but we retained these. The presence of non-overlapping families suggested that the rigour of our alignment

protocol could potentially have led to the filtering of a large number of sites. So, we checked whether or not the alignments stretched across the entire length of the shortest ortholog. We estimated the completeness of our alignments by calculating the alignment saturation level, which represents the fraction of the smallest ortholog that remained in the final alignment. Our results show that nearly half of the ortholog families (N=6029) had 100% alignment saturation, and only 11% of all ortholog families had less than 90% alignment saturation (Figure 2c). The mean saturation level stood at 96.5%. The observation which further bolstered the confidence in the quality of our alignments was that 87% of all ortholog families shared over 95% identity with all four species within the aligned region (Figure 2d).

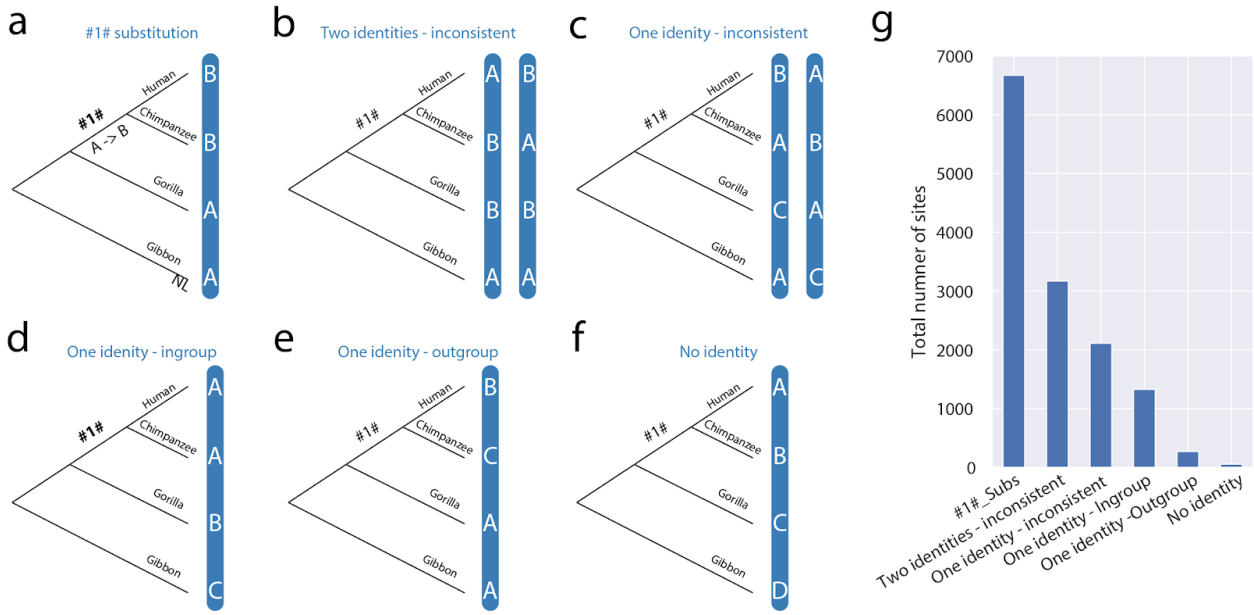

Figure 3: Sites substituted in more than one species. a-f) Different types of substitution patterns identified from the alignments. Letters in vertical blue bars represent residues on an aligned site. g) Bar plot representing the total number of sites based on substitution types.

Substitution patterns and the molecular clock

Of the aligned 7,609,284 amino acid residues, 97.6% were identical in all four species. Among the 183,424 sites with at least one substitution, 92.6% were species-specific substitutions, i.e. they were substituted in only one of the four species. Moreover, 55.7% of the substituted sites were only substituted in gibbons, here the substitutions on the #2# branch were implicitly included.

133 13,557 sites in 5,669 protein families were substituted in more than one species, and we divided
134 them into six categories (Figure 3a-f), based on the phylogeny. About half of these sites
135 (N=6,668) fell onto the phylogenetically consistent branch #1# leading to human/chimpanzee
136 (Figure 3g). 5,285 sites were phylogenetically inconsistent in different combinations (Figure
137 3b,c). Since convergent substitutions are unlikely in this dataset, this testifies the influence of
138 incomplete lineage sorting and/or introgression effects in this shallow phylogeny. The remaining
139 sites constitute multiple substitutions (Figure 3d-f), including the category 'No identity', which
140 covered 52 sites with a different amino acid in each branch (Figure 3f) and each site was in a
141 different ortholog family.

142 An overall molecular clock pattern

143 The branch-specific substitutions, including the #1# branch, added up to 96.2% of all
144 substitutions and we relate the further analysis only to these sites. They can be used to obtain
145 branch-specific substitution rates, and when scaled to the branch length from timetree, they all
146 conform to a close range of 0.036-0.043 substitutions per site per Mya (Table 1). Of course,
147 given that the timetree branches were also calculated from molecular data, this consistency is
148 not surprising per se (Kumar et al. 2017). Still, it shows that it is possible to use the overall
149 divergence as the basis for a hypothesis of clock-like divergence.

150 **Table 1**: Branch-specific substitution rates

	Branch timetree (Mya)	Total subs	% subs per site	Average % subs per site from subsampling	SD from subsampling	% subs per site per Mya from average
Human	6.65	20718	0.27	0.27	0.003	0.041
Chimpanzee	6.65	19721	0.26	0.27	0.004	0.041
#1#	2.41	6668	0.09	0.09	0.002	0.037
Gorilla	9.06	27174	0.36	0.37	0.004	0.041
Gibbon	31.24	102254	1.34	1.41	0.004	0.045

151 To test whether the mean branch-specific substitutions per site are a reliable measure of the
152 overall substitution rate along the branch, for each branch we subsampled half of the ortholog
153 families 5,000 times and each time calculated the mean % substitutions per site of the branch.
154 The respective values were normally distributed (Supplemental Figure 1) with a low standard

deviation estimate of the mean (Table 1). Thus, the cumulative substitution calculations across a large sample of families are reasonably consistent.

Average rate departures

As shown above, the random sampling across a broad set of gene families leads to a random distribution of sampling variances, approximated by normal distributions. However, the distribution of individual rates per family shows a significant departure from the normal distribution (Figure 4a). There was a long tail with 1,117 gene families showing no substitution at all. Also, there were several hundred gene families for which the values exceed the expectations from a normal distribution, i.e. above the expected line - note that exact cutoffs cannot be defined for such an extremely skewed distribution.

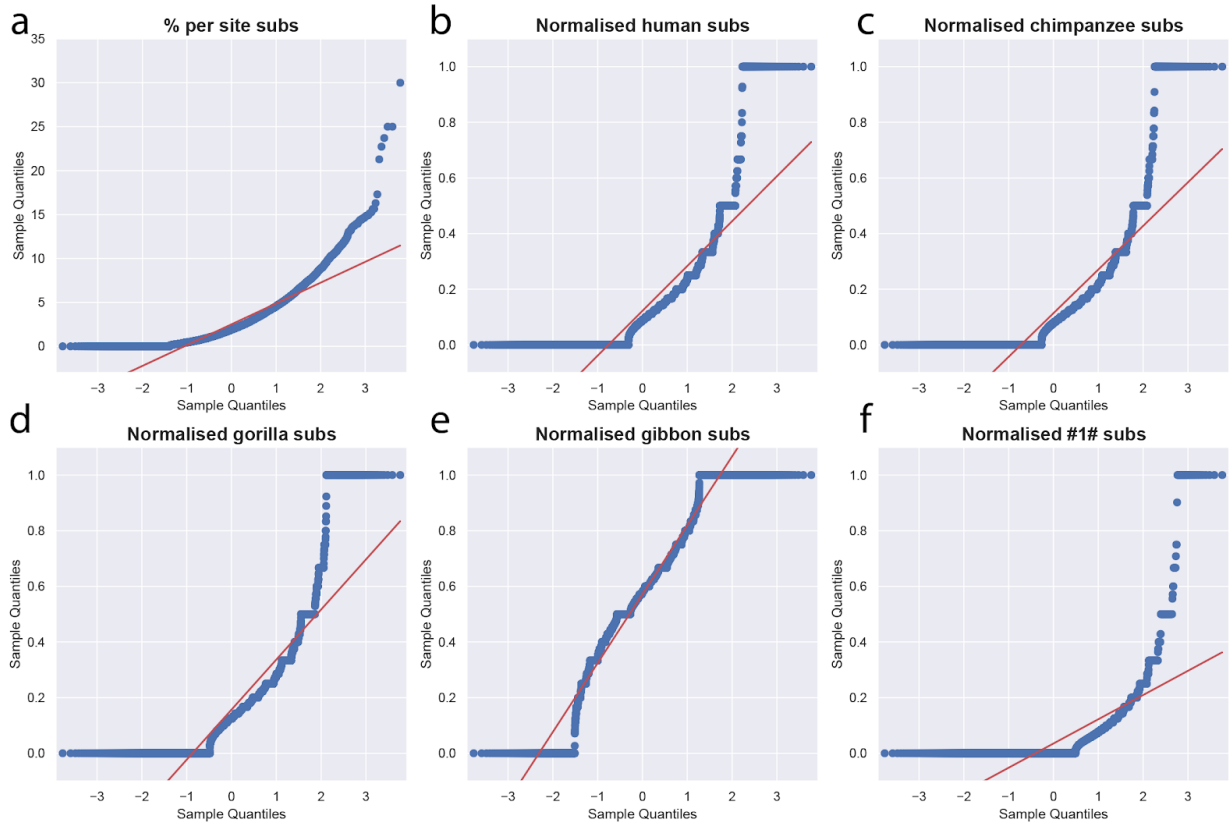

Figure 4: qq-plot analysis values distributions for (A) averages of branch-specific substitutions per site for all families, (B-F) branch-specific substitutions per site per family for the five branches in the tree.

This overall pattern can be interpreted as families of genes that are either subject to high constraints or subject to a generally faster evolution. Alternatively, gene families could have different rates in different branches, which would violate a generally constant clock assumption. For the gene families with zero substitutions overall, we would need further comparisons with more distant taxa to test this, i.e. we removed all 1,117 gene families with zero substitutions across the branches from the following analysis. We retained a total of 12,046 families with at least one diverging gene. For these, we calculated branch-specific substitutions per site, normalised for each family (see methods). Under a constant clock model, these should approximate a normal distribution for each branch. Still, the corresponding qq-plots again show major deviations from a normal distribution for all branches, except the long branch leading to Gibbons (Figure 4B-F). Most interestingly, the gene families falling into these extremes on one branch are not necessarily the same that fall into the extremes on another branch.

To better quantify this, we use an ad hoc cutoff to identify branches with a higher than expected number of substitutions (Supplemental Table 1). We derive this from the average that we observe at the internal branch #1# (see methods). According to the set maximum cutoff, only 334 families did not have any branches with larger than expected substitutions (Table 2). More than half of the families (N=6093) had more than expected substitutions on just one branch, followed by 4,919 families that had two branches with more than expected substitutions. 675 families had three branches with a greater than expected number of substitutions. We examined the distributions of species sets when either two or three branches had more than expected substitutions. Both distributions did not show a statistically significant departure from normality, which also suggests that the branch-specific substitutions are independent.

Overall, we only found 117 families, where all five branches were within their range of minimum and maximum threshold (Figure 5). Suspecting that our set cutoff was perhaps too stringent, we subsampled 117 families 5,000 times and calculated mean normalised substitutions for each branch. 4,044 subsamples had substitutions within the expected range on all five branches (Figure 5). Thus, even with a small subset of families, an overwhelming majority of subsamples tend to conform with the molecular clock; still, 89% of the families show higher than expected substitution on at least one branch

Table 2: Branches with higher than expected substitutions

Ortholog family branches with a higher than expected number of substitutions	Ortholog families	Ortholog family branches with a higher than expected number of substitutions	Ortholog families
Gibbon	2797	#1#	310
Gorilla	1180	Human, #1#	286
Human	977	Human, Chimpanzee, Gorilla	238
Chimpanzee	829	Chimpanzee, #1#	216
Human, Gorilla	718	Chimpanzee, Gorilla, #1#	112
Human, Chimpanzee	679	Human, Gorilla, #1#	108
Human, Gibbon	607	Human, Chimpanzee, #1#	97
Chimpanzee, Gorilla	576	Human, Chimpanzee, Gibbon	38
Gorilla, Gibbon	562	Chimpanzee, Gibbon, #1#	37
Chimpanzee, Gibbon	553	Human, Gibbon, #1#	35
Gibbon, #1#	380	Human, Chimpanzee, Gorilla, #1#	25
Gorilla, #1#	342	Gorilla, Gibbon, #1#	9
None	334	Chimpanzee, Gorilla, Gibbon	1

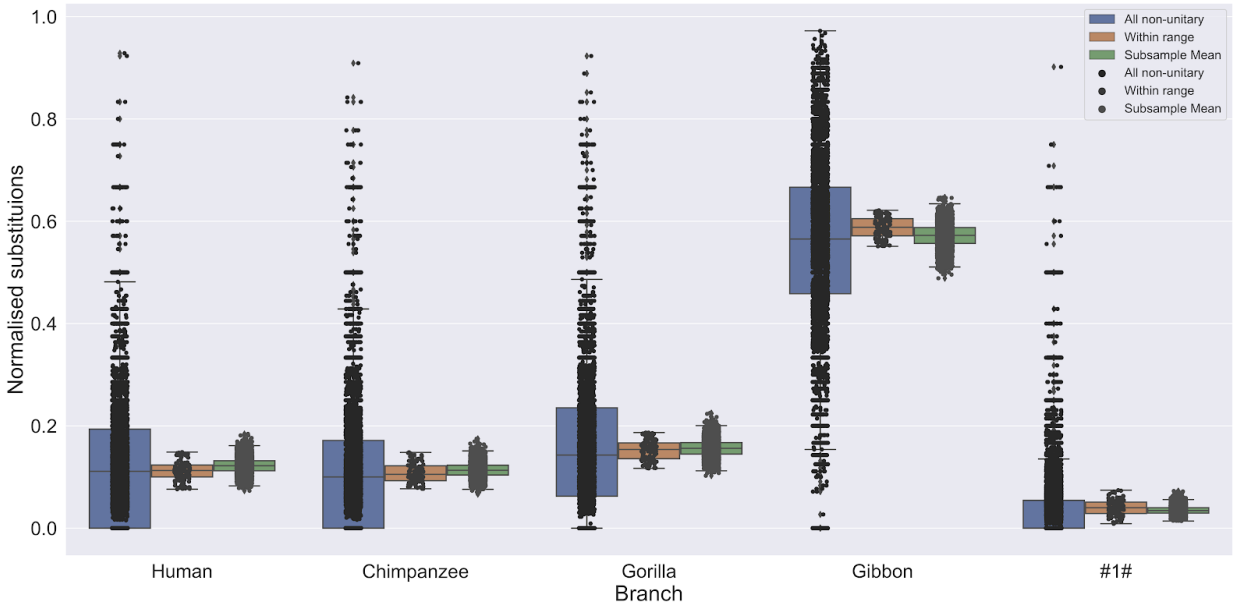

Figure 5: Distribution of normalised substitutions. All non-unitary includes gene families that have substitutions on more than one branch.

Table 3: Genes diverging rapidly on a particular branch

Branch	Branch related transcript	Ortholog family / Human transcript	Gene	Description	Branch-specific % subs per site	Norm branch subs	Align overlap	Align sat	Most recent duplication in the last common ancestors of
Human	ENST00000450565	ENST00000450565	ADCYAP1	adenylate cyclase activating polypeptide 1	7.60	0.93	171	97.16	Jawed vertebrates
Human	ENST00000369951	ENST00000369951	OPN1LW	opsin 1, long wave sensitive	3.76	0.83	266	87.21	Apes and Old World monkeys
Human	ENST00000637878	ENST00000637878	PVALEF	parvalbumin like EF-hand containing	3.76	0.50	133	99.25	-
Human	ENST00000259845	ENST00000259845	PSORS1C 2	psoriasis susceptibility 1 candidate 2	2.94	0.57	136	100.00	-
Human	ENST00000454136	ENST00000454136	BTNL2	butyrophilin like 2	2.88	0.41	243	89.67	Jawed vertebrates
Chimp	ENSPTRT00000079998	ENST00000542996	CALU	calumenin	6.25	1.00	320	98.77	Bilateral animals
Chimp	ENSPTRT00000107370	ENST00000296280	MASP1	mannan binding lectin serine peptidase 1	6.18	0.84	518	74.00	Jawed vertebrates
Chimp	ENSPTRT00000055069	ENST00000380041	SCML1	Scm polycomb group protein like 1	5.21	0.38	326	98.79	Bilateral animals
Chimp	ENSPTRT00000039999	ENST00000371633	LCN12	lipocalin 12	3.23	0.42	155	88.07	Mammals
Chimp	ENSPTRT00000029628	ENST00000343470	LYAR	Ly1 antibody reactive	2.65	0.42	378	99.74	-
#1#	ENST00000274520	ENST00000274520	IL9	interleukin 9	2.78	0.40	144	100.00	-
#1#	ENST00000299191	ENST00000299191	C16orf78	chromosome 16 open reading frame 78	2.27	0.22	264	100.00	-
#1#	ENST00000433976	ENST00000433976	ZNF778	zinc finger protein 778	1.34	0.21	598	98.36	Primates
#1#	ENST00000295622	ENST00000295622	TEX55	testis expressed 55	1.34	0.16	449	99.78	Chordates
#1#	ENST00000392027	ENST00000392027	ALPP	alkaline phosphatase, placental	1.18	0.19	509	96.22	Homininae
Gorilla	ENSGGOT00000061260	ENST00000255499	-	ring finger protein 128	7.26	0.85	317	78.86	Bilateral animals
Gorilla	ENSGGOT00000043555	ENST00000432056	PRAC2	PRAC2 small nuclear protein	5.15	0.41	136	97.14	-
Gorilla	ENSGGOT00000003429	ENST00000340083	-	docking protein 7	5.08	0.45	295	69.09	Gorilla
Gorilla	ENSGGOT00000048499	ENST00000449199	SLC66A1 L	solute carrier family 66 member 1 like	4.48	0.38	134	100.00	Animals and Fungi
Gorilla	ENSGGOT00000014985	ENST00000331952	C14orf180	chromosome 14 open reading frame 180	3.57	0.38	168	98.25	-
Gibbon	ENSNLET00000016102	ENST00000513010	S100Z	S100 calcium binding protein Z	15.22	0.93	92	93.88	Vertebrates
Gibbon	ENSNLET00000048260	ENST00000345088	DPPA3	developmental pluripotency associated 3	14.97	0.96	147	92.45	Placental mammals
Gibbon	ENSNLET00000000523	ENST00000398462	-	sperm associated antigen 11B	13.43	0.90	67	62.04	Placental mammals
Gibbon	ENSNLET00000005490	ENST00000393330	TSPAN8	tetraspanin 8	12.66	0.81	237	100.00	Bilateral animals
Gibbon	ENSNLET00000017025	ENST00000397301	-	troponin T3, fast skeletal type	12.57	1.00	167	97.09	Bilateral animals

The fastest evolving genes

We manually curated a list of five most divergent proteins on each branch (Table 3). Three human genes, ADCYAP1, PSORS1C2, and BTNL2, were associated with neuronal phenotypes

such as schizophrenia (Hashimoto et al. 2007; Amare et al. 2020; Gusev et al. 2019) and OPN1LW was associated with vision (Supplemental Table 2). Thus, at least four of the top five human genes have known association with the brain. OPN1LW has the most closely related paralog among these genes, and the duplication event was traced to the common ancestor of apes and old world monkeys. Thus the genes appear to be fast diverging even in the absence of recent duplication (Yates et al. 2020). This also stands true for the top genes on the other branches. Calmuenin, Ring finger protein 128, S 100 calcium binding protein z, and Interleukin 9 were the most divergent genes on the chimpanzee, gorilla, gibbons, and #1# branch respectively. Interestingly, three out of the five topmost diverging genes on the #1# branch (IL9, C16orf78, and ALPP) had less than expected number of substitutions in both human and chimpanzee lineage (Table 4). In fact, IL9 the fastest diverging gene on the #1# branch did not have a single human- or chimpanzee-specific substitution. Thus, we have identified three genes that were ancestrally under rapid divergence regime, but have been conserved in both human and chimpanzee lineages.

Table 4: Genes with complex evolutionary history

Gibbon transcript	Ortholog family / Human transcript	Gene	Description	Branch-specific substitutions					Align overlap	Align sat
				Human	Chimp	Gorilla	Gibbon	#1#		
-	ENST00000274520	IL9	interleukin 9	0	0	0	6	4	144	100.00
-	ENST00000299191	C16orf78	chromosome 16 open reading frame 78	1	1	3	16	6	264	100.00
-	ENST00000392027	ALPP	alkaline phosphatase, placental	1	2	5	18	6	509	96.22
ENSNLET00000049004	ENST00000268655	ZNF174	zinc finger protein 174	2	2	3	0	1	407	100
ENSNLET00000059391	ENST00000335021	-	defensin beta 107A	2	1	2	0	1	70	100
ENSNLET00000032475	ENST00000340934	BROX-201	BRO1 domain and CAAX motif containing	1	1	1	0	1	343	100
ENSNLET00000036362	ENST00000409110	SAG	S-antigen visual arrestin	1	2	1	0	1	405	100
ENSNLET00000018060	ENST00000440890	BMPR1B	bone morphogenetic protein receptor type 1B	2	1	1	0	1	532	100
ENSNLET00000001874	ENST00000634080	DNAJB6	DnaJ heat shock protein family (Hsp40) member B6	1	1	1	0	1	300	92.02
ENSNLET00000024142	ENST00000635031	-	Putative MORF4 family-associated protein 1-lik...	1	1	2	0	1	119	100

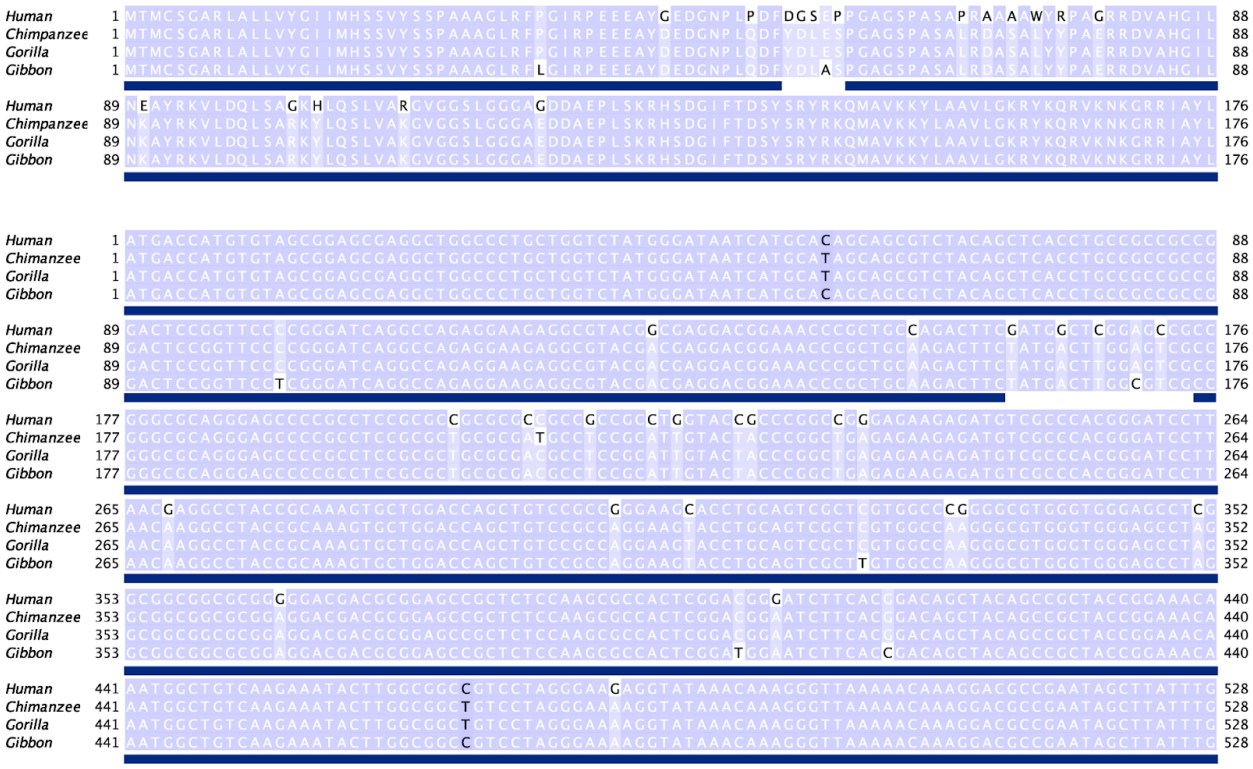

Figure 6: Protein and CDS multiple sequence alignments of the ADCYAP1 orthologs. Sites retained in the final alignment are underlined by the blue blocks.

ADCYAP1 was the most divergent human gene in our analysis. A previous study has shown that the gene went through accelerated adaptive evolution (Wang et al. 2005), however in the absence of genome sequence from other species they could not make any conclusion about its evolutionary rate with respect to other genes. The gene encodes a neuropeptide Pituitary adenylate cyclase-activating polypeptide (PACAP). PACAP, along with its receptor PAC1 (ADCYAP1R1), plays a crucial role in the regulation of behavioural and cellular stress response (Ressler et al. 2011). PACAP is known to stimulate adenylate cyclase in pituitary cells and also promotes neuron projection (Emery et al. 2013). ADCYAP1 has biased expression in appendix, brain, gall bladder, testis, and nine other tissues (Fagerberg et al. 2014). Within the sites retained in the final alignment there were 13 substitutions along the human branch and one substitution on the gibbon branch, at the nucleotide level, there were 20 human-specific and seven other substitutions within the same region (Figure 6). It is important to note here that all human-specific substitutions are A/T to G/C. A comparison of both amino acid and nucleotide alignment revealed that the five amino acid residues stretch not included in the final alignment also included only five nucleotide substitutions, but they resulted in five contiguous substitutions

at the protein level that led to the removal of these residues from the final alignment. This stretch had four human-specific amino acid residues as well as nucleotide substitutions and including these five residues raised the human-specific % substitutions per site to 9.7.

Conserved only on the gibbon branch

We also found seven families where the otherwise most extended branch (gibbon) had zero substitutions, while all the other branches had higher than expected substitutions (Table 4). Six out of the seven genes show 100% alignment saturation (Align sat), which confirms that any substituted gibbon sites have not been removed as a part of a rapidly diverging block. Although, not a single branch has more than three substitutions in any of these genes, still, the lack of substitution only on the longest branch is unlikely to be a chance event.

Discussion:

Using the classic approach to study protein-sequence divergence rates across time, we find a surprisingly frequent departure from rate equality for individual branches within a given phylogeny. At the same time, we see an overall molecular clock pattern when aggregating the data. Given that the branch-specific substitution rate changes include both, higher divergence and higher conservation, it appears that these effects cancel each other out in the aggregate data. This interpretation could reconcile the many previous reports on molecular clock departure patterns in individual protein families with the currently generally accepted notion that molecular data can be reliably used to derive splitting time estimates of taxa.

Technical considerations

To arrive at our conclusions, we needed a carefully curated dataset. The first challenge in the estimation of protein divergence was to establish a set of orthologous gene families. It was essential that we identified true orthologs derived from the same ancestral gene in the last common ancestor of the extant species. Identification of definite orthologs even among closely related species is made complicated by evolutionary processes such as deletion, duplication, and gene-conversion. Our reliance on synteny to identify a set of collinear genes verified that these positional orthologs are not just homologous but are also situated at loci with conserved gene order. Despite using stringent criteria, we identified two-third of the possible ortholog families. Three findings validated our confidence in this approach. First, most ortholog families did not show considerable variation in their protein lengths. Second, even though we did not

allow any gaps in our alignments, only 38 ortholog families shared less than 50 residues overlap, and the mean alignment saturation was 96.5%. Third, there was 97.6% identity within the 7.6 million aligned sites, and substituted sites were heavily enriched with branch-specific substitutions. Thus, we created a large and high-confidence set of ortholog families that allowed a comprehensive analysis of protein divergence in the four ape species.

A constant rate of sequence divergence predicted by the molecular clock should result in similar substitution rates on all branches of the species tree. We estimated branch-specific substitution rates by only employing the phylogenetically consistent substitutions, which make 96.2% of all substitutions, and when scaled to the timetree, they confirmed that the substitutions per site per Mya falls within a close range of 0.036-0.043 on all branches. Given that the timetree branch lengths were also estimated from the molecular data, the conformity of substitution rates was expected, however, whether this constant rate of decay was also maintained at the individual family level was not clear. To facilitate comparison of substitution rates among families, we calculated % substitution per site, normalised for the overlap length of each family. Through subsampling, we established that the branch-specific mean % substitution per site was also consistent with the uniform molecular clock hypothesis. Nonetheless, 89% of families had at least one branch with more substitution than what was expected if the substitution rates were uniform across all families. The most common departure from the expected decay rate involves higher than expected substitutions on just one or two branches.

We also observe higher than expected conservation patterns in branches, but since this is mostly reflected in zero substitutions only, it is difficult to quantify it in our shallow phylogeny. We have therefore included this analysis only for the longest branch and show some evidence for this. However, the current evidence could also be interpreted as an acceleration in the other four branches. While this would not be the most parsimonious explanation, it would still be possible, given the large number of accelerations we observe in the dataset. Evidently, the higher than expected conservation issue needs, therefore, to be further studied in deeper phylogenies. The suggestion that the acceleration and deceleration effects could balance each other out over time is also suggested by the comparisons of the distribution of branch-specific rates (Figure 4). While the rate distributions are highly skewed for each of the shorter branches, the qq-plot pattern approaches normality for the long branch (compare Fig. 4a-d with Fig. 4e). Since the latter reflects an accumulation of possibly different rate histories for each protein family, it could

represent an integration of the effects, at least partially. This integration was shown by the three genes that were rapidly evolving at the ancestral #1# branch but are conserved in both human and chimpanzee lineages.

Extreme genes

Given that the rate constancy of substitution rates is usually ascribed to neutral evolution, one can ascribe departures from this to positive or negative selection. While population demography and introgression effects could also contribute to departures from the neutral expectation, these effects are less relevant after extended splitting time, which is the case in our comparisons. Hence, it seems to designate genes with extreme changes in substitutions rates as genes that may have had a specific adaptive relevance for the respective taxon in whos branch they occur.

We have focused our analysis on the five fastest diverging genes on each branch and seven genes that are most conserved in the gibbon branch, but we emphasise that the list could easily be extended. Among the 25 highly divergent genes, five at every branch, only one was recently duplicated, confirming that rapid sequence divergence can take place even in the absence of duplication. In humans, we found that the rapidly evolving genes are involved in essential biological processes such as vision and cognition, and these genes were associated with disease phenotypes such as schizophrenia, autism, and colourblindness. We identified ADCYAP1 as the most divergent human protein-coding gene. It encodes a 176 amino acid residue protein that contains 17 human-specific substitutions, which estimates to a substitution frequency of 10%. Despite the high divergence, PACAP, the product of ADCYAP1, remains a key mediator of the behavioural and cellular stress response and the gene shows biased expression in appendix, brain, gall bladder, testis, and nine other tissues. The biological relevance of this gene coupled with the lack of recent duplicates affirms that the accelerated divergence did not result from functional redundancy. To our knowledge, no other protein has been shown to have such a high rate of human-specific divergence. It is necessary to emphasise that each of the lineages includes such lineage-specific highly accelerated genes, i.e. it is not special for humans to find such cases, but it is evidently a general pattern that accompanies the formation of species.

Conclusion

Our analysis reveals an overall surprisingly dynamic history of substitution rate changes in most protein families. The dynamics appear to be so high that acceleration and deceleration may cancel each other out over extended times. While this could give the impression of a long-term constant rate, which is often assumed as a null model for protein evolution, the actual history of the evolution of a given protein sequence could be much more complex.

Methods

One to one orthologs

For each species, we downloaded the CDS fasta file and gff file from the Ensembl ftp server (release-98) (Yates et al. 2020). We extracted the fasta sequence for the CDS of the longest isoform of each gene categorised as “biotype:protein-coding” in the gff file for further analysis. We translated the extracted CDS fasta sequences to obtain their corresponding protein sequence. To detect homologous genes, for each pair of species in our analysis, we ran all vs all BLASTP (Albà and Castresana 2007). The BLASTP result file and gff files of each species pair were provided as an input to MCScanX for synteny ascertainment (Wang et al. 2012). MCScanX was allowed to call a collinear block if a minimum of three collinear genes were found for the species pair with a maximum gap of two genes in between (Figure 2a). Several recent studies also relied on collinearity to establish orthologous relationships (Vakirlis et al. 2020; Zhao and Schranz 2019; Zhang et al. 2019; Rödelsperger et al. 2017; Heger and Ponting 2007; Lu et al. 2017; Sieriebriennikov et al. 2018).

We parsed the collinear gene pairs obtained from the MCScanX using the following method:

1. Both protein sequences from each collinear gene pair were aligned with Stretcher (Madeira et al. 2019).
2. If both proteins had 85% or more sequence identity, then this syntenic gene pair was retained.
    a. Else, we checked if either gene has a better BLASTP match, based on BLASTP score, with another gene from the other species. If so, drop the gene pair.
3. If a gene was present in more than one syntenic pair, then we retained the pair with the best reciprocal BLASTP score.

a. If two such gene pairs had the same BLASTP score, then we retained the pair which was within the larger syntenic block (based on the number of genes within each block).

Thus, at the end, we were left with a list of 1:1 orthologs for the given species pair (Figure 1c).

To determine the ortholog gene families, we overlapped the list of 1:1 orthologs of human genes with chimpanzee, gorilla, and gibbon genes. We retained only those gene families that have orthologs of the human genes in all three lists (Figure 1b). The length variation within the ortholog families was calculated by subtracting the length of the shortest ortholog from the longest ortholog.

Multiple sequence alignment of orthologous gene families

To investigate protein sequence divergence caused by single nucleotide substitution, we need to align amino acid residues that are derived from the same site of their last common ancestor. Given that most gene families are of comparable length, we set out to create alignments with fewer gaps. Hence, MAFFT was run with a gap opening penalty of 3 (Katoh and Standley 2013). Then to remove unreliable columns from the alignment, we used Gblocks with the following parameters (Talavera and Castresana 2007):

1. Minimum Number Of Sequences For A Conserved Position: 4
2. Minimum Number Of Sequences For A Flanking Position: 4
3. Maximum Number Of Contiguous Nonconserved Positions: 2
4. Minimum Length Of A Block: 10
5. Allowed Gap Positions: 0

Thus, in the end, we were left with the concatenation of all the conserved blocks identified by Gblocks. These blocks were free of any gap and at most contained two substituted sites contiguously. Given the phylogenetic proximity of all four species under investigation, we assumed that it was unlikely that many instances of three or more contiguous amino acid substitutions would result from independent point substitution events. Therefore, to avoid the inclusion of insertion or deletion events within the alignment block, we have removed any gaps or contiguous substitution of three and more residues.

Since gaps are not allowed in our alignment, their maximum length is limited by the shortest ortholog. We use this qualification to measure the completeness of the alignment of each ortholog family. If the overall alignment length is equal to the length of the shortest ortholog, then this family is considered to have attained 100% alignment saturation. The 'Alignment Saturation' level is calculated as per the following formula:

$$Alignment\ Saturation = \frac{Alignment\ overlap}{Length\ of\ the\ shortest\ sequence} \times 100\ \%$$

$$Alignment\ overlap = Number\ of\ aligned\ sites$$

% substitutions per site

The % substitutions per site were calculated as per the following formula:

$$\%\ substitutions\ per\ site = \frac{Number\ of\ substituted\ sites}{Alignment\ overlap} \times 100\ \%$$

For total % substitutions per site, all substituted sites were used in the above formula. But for branch-specific substitution frequencies, only branch-specific sites were used. The branch-specific sites were identified as sites with species-specific substitution i.e. only one substitution in any given column that is specific to one of the four species sequences. #1#-specific substitutions were identified as columns with human-chimpanzee identity and gorilla-gibbon identity, as shown in Figure 3a.

Distribution of total and branch-specific substitution rates were independently obtained by randomly subsampling half of the ortholog families 5000 times and calculating the respective mean % substitutions per site each time. A two-tailed chi-squared probability hypothesis test was conducted on each distribution, to check for departure from normality (D'agostino and Pearson 1973). The mean and the standard deviation of the mean % substitutions per site for each branch was obtained from their respective subsample distribution.

The normalisation of branch-specific substitutions and cutoff determination

To compare branches across the ortholog families, we normalised each branch of every ortholog family by using the following formula:

$$Normalised\ substitutions\ of\ Branch\ A = \frac{Substitutions\ on\ Branch\ A}{\Sigma(Substitutions\ on\ all\ branches)}$$

To test the conformity of the individual ortholog family trees with the tree generated by the entire data (Figure 5a), we estimated a permissive normalised substitution range for each branch (Supplemental Table 1). The maximum value of this range was calculated by adding the normalised substitutions of the #1# branch form the overall tree to that of the given branch, and the minimum value was calculated by subtracting the same. Thus, only the #1# branch was allowed to have a lower limit of zero. Subsampling of 117 ortholog families was done 5,000 times to obtain the distribution of their mean normalised substitutions.

Identification and analysis of most divergent genes and genes conserved on the gibbon branch

We identified the most divergent outliers on each species-specific branch by selecting ortholog families with normalised substitution higher than the maximum cutoff of the branch and minimum of three substitutions on the given branch. The top five candidates on each branch were manually curated after sorting the outliers according to their % substitutions per site and normalised substitutions. Further, we visually inspected the alignments and removed candidates that had multiple substitution within short alignment blocks. We also removed candidates that had tandem duplicates. Such filtered candidates are flagged in the ‘Comment’ section of the ‘NormDf.tsv’ file. The top 5 human candidates were validated with tissue-specific human expression data and Ensembl CDS alignment (Fagerberg et al. 2014). Due to lack of tissue-specific expression data in other species, we only retained the top candidates on other branches that were supported by their respective Ensembl CDS alignment. The duplicates of candidate genes were identified based on reciprocal BLASTP hit within our dataset and the paralogue information from the Ensembl database (Yates et al. 2020). The nucleotide alignment of ADCYAP1 was manually created.

We identified the genes conserved only on the gibbon branch as the genes with zero gibbon-specific substitutions but more than expected substitution on every other branch according to the set maximum threshold.

Statistical analysis

Statistical analysis and tests were done using custom python codes. Statistical significance was always tested at the p-value threshold of 0.05. To test departure from normality we conducted two-tailed chi-squared probability hypothesis test (D'agostino and Pearson 1973). Subsampling was done by random selection of samples with replacements.

Data Access

All raw data used in this study were acquired from ENSEMBL ftp server (release-98). Online supplementary material, including codes, are publicly available at https://github.com/neelduti/ApeDivergence.

Supplementary Information

Supplementary text, figure, and tables are available in the supplemental file.

Acknowledgements

The authors would like to thank Dr. Chen Xie for proofreading the manuscript. We thank Tautz lab members and Dr. Julien Dutheil for discussion. This work was supported by institutional funding through the Max Planck Society to DT.

Author contributions

NP designed the study with support from DT. NP performed the analysis. NP and DT wrote the manuscript.

Disclosure declaration

Authors declare no competing interest.

References:

Albà MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol Biol Evol* **22**: 598–606.

Albà MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. *BMC Evol Biol* **7**: 53.

Amare AT, Vaez A, Hsu Y-H, Direk N, Kamali Z, Howard DM, McIntosh AM, Tiemeier H, Bültmann U, Snieder H, et al. 2020. Bivariate genome-wide association analyses of the broad depression phenotype combined with major depressive disorder, bipolar disorder or schizophrenia reveal eight novel genetic loci for depression. *Mol Psychiatry* **25**: 1420–1429.

Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol Evol* **2**: 393–409.

Cantarel BL, Morrison HG, Pearson W. 2006. Exploring the relationship between sequence similarity and accurate phylogenetic trees. *Mol Biol Evol* **23**: 2090–2100.

Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**: 195–201.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7**: 98–108.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**: 1960–1963.

D'agostino R, Pearson ES. 1973. Tests for departure from normality. Empirical results for the distributions of b2 and √b1. *Biometrika* **60**: 613–622.

Ebersberger I, von Haeseler A. 28 Exploring phylogenomic data. *Deep Metazoan Phylogeny: The Backbone of the Tree of Life*. http://dx.doi.org/10.1515/9783110277524.595.

Emery AC, Eiden MV, Mustafa T, Eiden LE. 2013. Rapgef2 connects GPCR-mediated cAMP signals to ERK activation in neuronal and endocrine cells. *Sci Signal* **6**: ra51.

Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, et al. 2014. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* **13**: 397–406.

Forslund K, Pereira C, Capella-Gutierrez S, da Silva AS, Altenhoff A, Huerta-Cepas J, Muffato M, Patricio M, Vandepoele K, Ebersberger I, et al. 2018. Gearing up to handle the mosaic nature of life in the quest for orthologs. *Bioinformatics* **34**: 323–329.

Glover N, Dessimoz C, Ebersberger I, Forslund SK, Gabaldón T, Huerta-Cepas J, Martin M-J,

Muffato M, Patricio M, Pereira C, et al. 2019. Advances and Applications in the Quest for Orthologs. *Mol Biol Evol* **36**: 2157–2164.

Goodman M, Moore GW, Matsuda G. 1975. Darwinian evolution in the genealogy of haemoglobin. *Nature* **253**: 603–608.

Guillén Y, Casillas S, Ruiz A. 2019. Genome-Wide Patterns of Sequence Divergence of Protein-Coding Genes Between Drosophila buzzatii and D. mojavensis. *J Hered* **110**: 92–101.

Gusev FE, Reshetov DA, Mitchell AC, Andreeva TV, Dincer A, Grigorenko AP, Fedonin G, Halene T, Aliseychik M, Filippova E, et al. 2019. Chromatin profiling of cortical neurons identifies individual epigenetic signatures in schizophrenia. *Transl Psychiatry* **9**: 256.

Hashimoto R, Hashimoto H, Shintani N, Chiba S, Hattori S, Okada T, Nakajima M, Tanaka K, Kawagishi N, Nemoto K, et al. 2007. Pituitary adenylate cyclase-activating polypeptide is associated with schizophrenia. *Mol Psychiatry* **12**: 1026–1032.

Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res* **17**: 1837–1849.

Hurst LD. 2009. Evolutionary genomics and the reach of selection. *J Biol* **8**: 12.

Hurst LD. 2011. Molecular genetics: The sound of silence. *Nature* **471**: 582–583.

Johnson KP, Allen JM, Olds BP, Mugisha L, Reed DL, Paige KN, Pittendrigh BR. 2014. Rates of genomic divergence in humans, chimpanzees and their lice. *Proc Biol Sci* **281**: 20132174.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* **217**: 624–626.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34**: 1812–1819.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in Drosophila. *Trends Genet* **24**: 114–123.

Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. 2008. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res* **18**: 298–309.

Lu T-C, Leu J-Y, Lin W-C. 2017. A Comprehensive Analysis of Transcript-Supported De Novo Genes in Saccharomyces sensu stricto Yeasts. *Mol Biol Evol* **34**: 2823–2838.

Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, et al. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* **47**: W636–W641.

Magalhães JP de, de Magalhães JP, Church GM. 2007. Analyses of human–chimpanzee orthologous gene pairs to explore evolutionary hypotheses of aging. *Mechanisms of Ageing and Development* **128**: 355–364. http://dx.doi.org/10.1016/j.mad.2007.03.004.

Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Res* **21**: 863–874.

Nuttall GHF. 1904. Blood Immunity and Blood-Relationships.

Parmley JL, Hurst LD. 2007. How common are intragene windows with KA > KS owing to purifying selection on synonymous mutations? *J Mol Evol* **64**: 646–655.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics* **12**: 32–42. http://dx.doi.org/10.1038/nrg2899.

Prabh N, Rödelsperger C. 2016. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? *BMC Bioinformatics* **17**: 226.

Prabh N, Roeseler W, Witte H, Eberhardt G, Sommer RJ, Rödelsperger C. 2018. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in nematodes. *Genome Res* **28**: 1664–1674.

Ressler KJ, Mercer KB, Bradley B, Jovanovic T, Mahan A, Kerley K, Norrholm SD, Kilaru V, Smith AK, Myers AJ, et al. 2011. Post-traumatic stress disorder is associated with PACAP and the PAC1 receptor. *Nature* **470**: 492–497.

Rödelsperger C, Meyer JM, Prabh N, Lanz C, Bemm F, Sommer RJ. 2017. Single-Molecule Sequencing Reveals the Chromosome-Scale Genomic Architecture of the Nematode Model Organism Pristionchus pacificus. *Cell Rep* **21**: 834–844.

Schmid KJ, Aquadro CF. 2001. The evolutionary analysis of “orphans” from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes. *Genetics* **159**: 589–598.

Sieriebriennikov B, Prabh N, Dardiry M, Witte H, Röseler W, Kieninger MR, Rödelsperger C, Sommer RJ. 2018. A Developmental Switch Generating Phenotypic Plasticity Is Part of a Conserved Multi-gene Locus. *Cell Rep* **23**: 2835–2843.e4.

Takahata N. 2007. Molecular clock: an anti-neo-Darwinian legacy. *Genetics* **176**: 1–6.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**: 564–577.

Toll-Riera M, Laurie S, Albà MM. 2011. Lineage-specific variation in intensity of natural selection in mammals. *Mol Biol Evol* **28**: 383–398.

Vakirlis N, Carvunis A-R, McLysaght A. 2020. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife* **9**. http://dx.doi.org/10.7554/eLife.53500.

Wang D, Liu F, Wang L, Huang S, Yu J. 2011. Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. *Biology Direct* **6**: 13. http://dx.doi.org/10.1186/1745-6150-6-13.

Wang Y-Q, Qian Y-P, Yang S, Shi H, Liao C-H, Zheng H-K, Wang J, Lin AA, Cavalli-Sforza LL, Underhill PA, et al. 2005. Accelerated evolution of the pituitary adenylate cyclase-activating polypeptide precursor gene during human origin. *Genetics* **170**: 801–806.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee T-H, Jin H, Marler B, Guo H, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**: e49.

Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. *Science* **319**: 473–476.

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al. 2020. Ensembl 2020. *Nucleic Acids Res* **48**: D682–D688.

Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, et al. 2019. Rapid evolution of protein diversity by de novo origination in Oryza. *Nat Ecol Evol*. http://dx.doi.org/10.1038/s41559-019-0822-5.

Zhao T, Schranz ME. 2019. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proc Natl Acad Sci U S A* **116**: 2165–2174.

Zuckerkandl E, Pauling L. 1965. Evolutionary Divergence and Convergence in Proteins. *Evolving Genes and Proteins* 97–166. http://dx.doi.org/10.1016/b978-1-4832-2734-4.50017-6.

Zuckerkandl E, Pauling L. 1962. *Molecular Disease, Evolution, and Genic Heterogeneity*.