Evolutionary insights reveal a new role of PADI2 in transcriptional elongation

José Luis Villanueva-Cañas‡[1,7], Narcis Fernandez-Fuentes‡[2], Catherine Teyssier[3], Malgorzata Ewa Rogalska[1], Ferran Pegenaute Pérez[4,5], Baldomero Oliva[5,6], Cedric Notredame[1,5], Miguel Beato[1,5], Priyanka Sharma*[3]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88,08003 Barcelona, Spain.

[2]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, Ceredigion, United Kingdom

[3]Institut de Recherche en Cancérologie de Montpellier, INSERM, Université de Montpellier, Institut Régional du Cancer de Montpellier, Montpellier, France

[4]Live-Cell Structural Biology Laboratory, Department of Medicine and Life Sciences, Barcelona E-08005, Spain.

[5]Universitat Pompeu Fabra (UPF), Barcelona, Spain.

[6]Structural Bioinformatics Laboratory (GRIB-IMIM), Department of Medicine and Life Sciences, Barcelona E-08003, Spain.

[7]Present Address: [José Luis Villanueva-Cañas], Molecular Biology CORE (CDB), Hospital Clínic de Barcelona, Barcelona, Spain

*Priyanka Sharma
Email:priyanka.sharma@inserm.fr

‡ These authors contributed equally

Abstract

Protein citrullination (deimination) is the post-translational modification of arginine to the non-coded amino acid citrulline and is catalyzed by the peptidyl arginine deiminase (PADI) enzyme family. The most widely expressed of the PADIs family, PADI2, regulates cellular processes that impact several diseases. How PADI2 has evolved in mammals to gain its fundamental function is not fully understood. By performing a systematic evolutionary analysis, we now identify 20 positively selected residues in PADI2, 16 of which are structurally exposed mainly on the N-terminal and middle domains of PADI2. Our integrated evolutionary and structural analyses suggest that these residues have roles in maintaining the PADI2 interactions with cognate proteins. We experimentally demonstrate that one of the loops in the middle domain participates in the interactions of PADI2 with the positive-transcription elongation factor (P-TEFb), which is essential for active transcription and cellular proliferation. This interpretation is supported by our finding that L162 within this loop evolved under positive selection. Our work demonstrates the power of combining sequence-based phylogenetic methods with structural information for investigating the selective function of the middle domain of PADI2 in modulating transcription. This in-depth knowledge could be key to understanding the role of PADI2 overexpression in disease and points to potential targetable regions of the protein.

Keywords: Arginine citrullination, peptidyl arginine deiminase 2, integrative structure and evolution analysis, protein-protein interactions, positive transcription elongation factor b.

Significance Statement

A systematic evolutionary analysis identified positively selected residues in the non-catalytic domain of PADI2. We established a link between the positive evolution of key residues in the PADI2 and their role in transcription. Specifically, we identified the structurally exposed loop encompassing the positively selected L162 in the PADI2 middle domain and its role in transcription and cellular proliferation. This loop contributes to the PADI2 interaction with the P-TEFb complex and cellular proliferation. Our results showcase the use of combining evolutionary and experimental approaches to dissect the dynamic of evolutionary processes.

Introduction

The peptidyl arginine deiminase (PADI) family members catalyze calcium-dependent citrullination of protein-embedded arginines and convert them into the non-coded amino acid citrulline (1–3). Citrullination of arginine residues is a widespread post-translational modification (PTM) that increases the hydrophobicity of proteins and can contribute to fine-tuning physiological processes. For instance, when it occurs in core histones, it weakens the histone–nucleic acid interactions, impinging on both chromatin organization and transcription (4). Citrullination can also affect protein folding and consequently protein function (5–8). The five members of the PADI family exhibit different tissue expressions: PADI1 is present only in the epidermis and uterus; PADI3, in the epidermis and hair follicles; PADI4, in immune cells, brain, uterus, and bone marrow; PADI6, in the ovarian egg cells, embryo, and testicles; and PADI2, in brain, uterus, spleen, breast, pancreas, skin, and skeletal muscles (2, 3). Notably, PADI2 has been associated with multiple pathological states, such as autoimmune disorders (9) and neurological diseases (10, 11), as well as with several cancers (e.g., breast, cervix, liver, lung, ovary, thyroid, gastric, and prostate cancer) (12–21). This association of PADI2 with various diseases underscores an unmet need to elucidate its molecular mechanisms of action and to understand its pathophysiological function.

PADI family members have distinct substrates and target diverse arginine residues in different proteins (22). All family members require Ca^{2+} ions for their function, but PADI2 also relies on an ordered calcium binding to its active site for substrate binding and catalysis, as was shown for the deimination of arginine 26 in histone H3 (H3R26) (23–26). The PADI2-mediated H3R26 citrullination mark interacts with SMARCAD1 (SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A containing DEAD/H box 1) to regulate naive pluripotency (27), highlighting PADI2´s function in gene activation. PADI2 contributes to chromatin modification that promotes the differentiation of oligodendrocyte precursors and efficient myelination, which are required for motor and cognitive functions (28). Recently, PADI2 was found to citrullinate MEK1, thereby promoting extracellular signal-regulated protein kinases 1/2 (ERK1/2) signaling in endometrium cancer (21). Additionally, we have identified another unique function of PADI2, namely the deimination of arginine 1810 (R1810) on the repeat 31 of the C-terminal domain (CTD) of the large subunit of RNA polymerase 2 (RNAP2). This modification potentiates the interaction of the CTD with the positive transcription elongation factor b complex (P-TEFb) and consequently facilitates the expression of genes that are essential for cell identity (29). These studies support the

notion that PADI2 has a selective ability to citrullinate arginine residues in specific proteins, which may explain its relevance to several pathophysiological conditions.

The *PADI* genes are ubiquitous in vertebrates but are absent from yeast, worm, and flies. A recent study focusing on the comprehensive identification of *PADI* homologs unveiled the evolutionary trajectory of *PADIs* within the animal lineage (30). *PADIs* appear to have been introduced from cyanobacteria into animals by horizontal gene transfer (HGT) (30), supporting the previous hypothesis of HGT as a mechanism for the introduction of new genetic material into the vertebrate genomes (31–34). The citrullinating enzymes found in human parasites and microbes are highly divergent in sequence and have different substrate specificities. These include pPAD, which is an extended agmatine deiminase found in *Porphyromonas gingivalis*, and giardia ADI, which is an extended form of the free L-arginine deiminase gADI found in the human parasite *Giandia lamblia* (35, 36). Previous small scale phylogenetic analysis of the PADI family has shown that PADI2 is the most conserved family member. Mammalian PADIs comprise three structural domains; the N-terminal domain (NTD) (PADI_N, Pfam annotation: PF08526), the middle domain (PADI_M, Pfam annotation: PF08527), and the catalytic C-terminal domain (PADI_C, Pfam annotation: PF03068) (37–40). Of note, there is a closer relationship between PADI1 and PADI3, and between PADI4 and PADI6, than between any member and PADI2 (5), suggesting that PADI2 could be the original founder within the PADI family and that the others arose by gene duplication.

The enzymes responsible for some of the essential PTMs (including phosphorylation, acetylation, and glycosylation) are known to be present across all domains of life, suggesting that they were present in the Last Universal Common Ancestor (LUCA) (41, 42). Similarly, the PADI_C domain was also present in the LUCA, indicating an ancient origin of citrullination. The PADI_M domain that encompasses the sequential calcium-binding sites, which maintains the allosteric communication with PADI_C (24), is present in cyanobacteria. Indeed, recent work supports the appearance of a degenerated PADI_N domain in cyanobacteria and demonstrates the existence of catalytically active PADI proteins in the cyanobacteria (30). The relatively recent appearance of the NTD within the PADI family evolution suggests it has a unique functional relevance in higher organisms.

Previously, we found that PADI2 interacts with the positive transcription elongation factor b complex (P-TEFb) to maintain the expression of the actively transcribing genes (29, 43). The P-TEFb complex comprises cyclin-dependent kinase 9 (CDK9) and its regulatory partner cyclin T1 (CCNT1) (44). During the early phase of transcription initiation, the activation of the P-TEFb complex is

required to overcome RNAP2 pausing and to continue with a productive phase of transcription elongation (44–47). The absence of proper recruitment of the P-TEFb complex along with associated components of transcription activation complexes has been linked to several disease conditions (48–50), highlighting its functional implication in maintaining the proper transcription output.

In this work, we characterized the recent evolution of the PADI protein family. We performed a comparative genomics analysis of PADI2 and identified 20 putative amino acid substitutions in different species, that might be important for PADI2 structure and functionality in mammals. The majority of the selected amino acids are exposed in the three-dimensional (3D) structure, and they belong to the non-catalytic NTD or to the middle domain of PADI2, suggesting that they could participitate in protein-protein interactions. We investigated the functional relevance of positively selected exposed amino acids in the NTD and middle domain of PADI2 by analyzing their implication in modulating PADI2 interactions with the P-TEFb complex. We found that positively selected leucine 162 (L162), in the exposed loop of the middle domain of PADI2, is positively selected at the basis of all primates, and contributes to the interaction with the P-TEFb complex. These results established a regulatory link between the favorable evolution of key residues in the PADI2 and their role in promoting the protein-protein interactions required for functional transcription regulation.

Materials and Methods

Cell lines

HeLa cells (ATCC CCL-2) were grown in DMEM with 10% FBS and 100U/ml penicillin-streptomycin according to the ATCC's recommendations. Cells were transfected using Lipofectamine 3000 (Invitrogen) according to the manufacturer's instructions.

PADI2-GFP plasmids and fluorescence-activated cell sorting (FACS) sorting

PADI2 was cloned into the pCPR0032 GFP-tagged vector using the forward primer (F) 5´-AGAACCTGTACTTCCAATCCATGCTGCGCGAGCGGAC-3´ and the reverse primer (R) 5´-GATCCGTATCCACCTTTACTTTAGGGCACCATGTGCCACC-3´ using Gibson assembly (51). The selected plasmid sequence was verified by the Sanger sequencing. Next, PADI2 wild-type (WT) was used to generate a single mutant (T159A or W161A or L162A), a double mutant (L162A/W161A), and a triple mutant (L162A/ W161A/ T159A), using the following primers:

PADI2- T159A

F1 5´- TGGTGAACTGTGACCGAG**AGG**CACCCTGGTTGCCCAAGGAGGACTGCCGTGATG -3´

R1 5´- CATCACGGCAGTCCTCCTTGGGCAACCAGGGTG**CCT**CTCGGTCACAGTTCACCA -3´.

PADI2- W161A

F1 5´- TGGTGAACTGTGACCGAGAGACACC**CGC**GTTGCCCAAGGAGGACTGCCGTGATG -3´

R1 5´- CATCACGGCAGTCCTCCTTGGGCAAC**GCG**GGTGTCTCTCGGTCACAGTTCACCA -3´.

PADI2- L162A

F1 5´- TGGTGAACTGTGACCGAGAGACACCCTGG**GCA**CCCAAGGAGGACTGCCGTGATG-3´

R1 5´-CATCACGGCAGTCCTCCTTGGG**TGC**CCAGGGTGTCTCTCGGTCACAGTTCACCA-3´.

PADI2- L162A/ W161A

F2 5´- TGGTGAACTGTGACCGAGAGACACCC***GCC*GCA**CCCAAGGAGGACTGCCGTGATG -3´

R2 5´- CATCACGGCAGTCCTCCTTGGG**TGC*GGC***GGGTGTCTCTCGGTCACAGTTCACCA -3´.

PADI2- L162A/W161A/T159A

F3 5´- TGGTGAACTGTGACCGAGAG**<u>GCA</u>**CCC***GCC*GCA**CCCAAGGAGGACTGCCGTGATG -3´

R3 5´- CATCACGGCAGTCCTCCTTGGG**TGC*GGC***GGG**<u>TGC</u>**CTCTCGGTCACAGTTCACCA-3´.

The corresponding fragments were generated by each pair of primers per oligonucleotide assembly and introduced by Gibson assembly (51). All the generated mutants were confirmed by Sanger sequencing.

For transfection, 2×10^6 HeLa cells were seeded in 10-cm plates, and 4 ug of each of the plasmid was transfected using Lipofectamine 3000 (Invitrogen) for 24 hours according to the manufacturer´s instructions. Cells were trypsinized, and GFP-positive live cells were sorted using BD influx (Becton and Dickinson, San Jose, CA). Briefly, cells were stained with 1µg/mL concentration of DAPI (4′, 6-diamidino-2-phenylindole) before FACS sorting. A SSC-H (side scatter height) versus FSC-H (forward scatter height), morphological-related parameters dot-plot was used to exclude debris by gating cells; doublets were then excluded using a FSC-H versus FSC-A (forward scatter area) by gating singlets followed by dead cells were excluded using DAPI versus FSC-A dot-plot by gating living cells. GFP-positive cells were identified and isolated using a G4 gate in GFP versus autofluorescence (AF) dot-plot. Obtained data were analyzed by using the Flow Jo 10. 6. The GFP-positive cells were used for experiments. Cells were centrifuged and stored as pellets at –80°C prior to RNA extraction and immunoprecipitation experiments.

Cell Proliferation assay

BrdU (5´-bromo-2´-deoxyuridine) cell proliferation assay: HeLa cells (0.3×10^3) were transfected with either wild-type PADI2 (WT) or mutant PADI2 (with L642A or T159 or W161A or L642A/W161A, or L642A/W161A/T159A) using the Lipofectamine 3000 (Invitrogen) in 96-well plate. The cell proliferation ELISA BrdU (5´-bromo-2´-deoxyuridine) colorimetric assay (Roche,11647229001) was performed as per the manufacturer's instructions. The experiments were performed on at least eight biological replicates.

Incucyte ® Proliferation Assays for Live-Cell Analysis

HeLa cells seeded in 96-well plates with 300 cells per well, transfected with either wild-type PADI2 (WT) or mutant PADI2 (with L642A or T159 or W161A or L642A/W161A, or L642A/W161A/T159A). After 24hours of transfection, imaging was performed using the IncuCyte live cell imaging system (Essen BioScience). Scans at 4× magnification were taken every 8 hours for 5 days. Cell confluence was calculated from the microscopy images by the Incucyte software algorithm to generate a proliferation index corresponding to the change in confluence for each well. These measurements are a mean ± SEM of at least six replicates.

RNA extraction and RT-qPCRs

RNA from HeLa cells transfected with either wild-type PADI2 (WT) or mutant PADI2 (with L642A, or L642A/W161A, or L642A/W161A/T159A) was extracted using RNeasy (Qiagen) according to manufacturer´s instructions. Purified RNA (1µg) was used for DNase treatment (Thermo Scientific), and quantified with a Qubit 3.0 Fluorometer (Life Technologies).

Reverse transcription of RNA was performed using a qScript™ cDNA Synthesis Kit (Quanta Bioscience 95047-100) according to the manufacturer´s instructions. Complementary DNA was quantified by qPCR using Roche Lightcycler (Roche), as previously described (52). For each gene product, relative RNA abundance was calculated using the standard curve method and expressed as relative RNA abundance after normalizing against the human *GAPDH* gene level. All gene expression data generated by RT-qPCR represented the average and ± SEM of at least three biological replicates. Primers used for RT-qPCR are listed in **Table S1**.

ChIP-qPCRs: For ChIP assays (29), 4 x 10^6 of FACS sorted GFP-positive HeLa cells (WT-PADI2, L162A-PADI2, L162A/W161A-PADI2, L162A/W161A/T159A-PADI2) were cross-linked for 10 min with 1% formaldehyde at 37ºC. The chromatin lysate was sonicated to a DNA fragment size range of 100-200bp using a Biorupter sonicator (Diagenode). CDK9 was immunoprecipitated with 15µg of CDK9 (D-7, sc-13130, lot no # B1422) or control mouse IgG (12-371, Merck) in IP Buffer with 2X SDS buffer (100mM NaCl, 50mM Tris-HCl, pH8, 5mM EDTA and 0.5% SDS) and 1X Triton buffer (100mM Tris-HCl, pH8.8, 100mM NaCl, 5mM EDTA and 5% Triton-X) with protease inhibitors (11836170001, Roche) for 16 hours at 4º C. Followed by incubating with 50µl of Dynabeads® M-280 sheep anti-mouse IgG (11201D, Thermo Scientific) for 3 hours. Beads were washed with 3 times with low salt buffer (140mM NaCl, 50mM HEPES, pH 7.4, 1% Triton-X 100), 2 times with high salt buffer (500 mM NaCl, 50mM HEPES, pH 7.4, 1% Triton-X 100) followed by single wash of LiCl Buffer (10mM Tris HCl pH 8.0, 250 mM LiCl, 1% NP-40, 1% sodium deoxycholic acid and 1mM EDTA) and 1× TE buffer in cold room. Subsequently, crosslinks were reversed at 65º C overnight, followed by RNAase treatment for 1.5 hours, and bound DNA was purified by Phenol-Chloroform extraction. The resultant eluted DNA was quantified by Qubit 3.0 Fluorometer (Life Technologies) and followed by real-time qPCR analysis. Data are represented as fold-change over input fraction from at least 3 biological replicate experiments. Primers used for qPCR are listed in **Table S1**.

GFP-tagged PADI2 Immunoprecipitation, western blot

Briefly, 3 × 10^6 FACS sorted GFP-positive (WT-PADI2, L162A-PADI2, L162A/W161A-PADI2, L162A/W161A/T159A-PADI2) cells were lysed on ice for 30 min in lysis buffer (1% Triton X-100 in 50mM Tris pH 7.4–7.6, 130 mM NaCl) containing proteases inhibitors (11836170001, Roche) with rotation, followed by sonication for 7 min with every 30 sec on / 30 sec off. After centrifugation at 4ºC and 13,000 rpm for 10 min, extracts were used for protein quantitation. For immunoprecipitation (IP) assay, 2mg of extract was incubated for 12 hours with 100µl Dynabeads Protein A (10002D, Thermo Scientific). The monoclonal antibody anti-GFP (polyclonal rabbit, A-11122, Invitrogen) or a control rabbit IgG (2729S, Cell Signaling) was coupled with Dynabeads before incubation with extract at 4ºC. The samples were washed 10 times with lysis buffer and boiled for 5 min in SDS gel sample buffer. Proteins were visualized by 4% to 12% SDS-PAGE gels and western blotting, using anti-GFP (11814460001, Roche), anti-CDK9 (sc-13130, Santa Cruz), or anti-CCNT1 (A303-499A, Bethyl Labs) were used for western blots.

Multiple sequence alignments

All of the homolog PADI sequences available in the OMA database were downloaded (53). The species for the study were selected using three criteria: (i) a good representation of the different mammalian lineages, (ii) good sequence quality, and (iii) the presence of a full PADI2 sequence. Three outgroups of bird species (chicken, turkey, and duck) were also used. The protein sequences were aligned using the software PRANK (54) and a pruned mammalian guide tree with branch distances (55). This program uses an evolutionary model to place insertions and deletions, minimizing the over-alignment of non-homologous regions, and has been shown to improve dN/dS estimates (56). Local realignments of two regions were done using MAFFT (57) within AliView (58). Another MSA using only one-to-one orthologous PADI2 sequences was also built. TrimAI (59) was applied to the PADI family MSA, enabling the -automated1 option, as recommended before a phylogenetic reconstruction. The PADI2 protein MSA was then converted into a nucleotide alignment using the script Pal2Nal. The corresponding cDNA sequences for each species were gathered from the OMA database and converted using the script Pal2Nal (60). The MSAs generated and the trees used are available as Supplementary Material S1.

Phylogenetic reconstruction

The PADI family reconstruction was done using RaxML (61) with 200 bootstrap values (-N 200), with an optimal amino acid substitution model chosen automatically (-m PROTGAMMAAUTO), and a rapid bootstrap algorithm (-f a) along with reproducible seed values (-p 12345, -x 345). Two reconstructions were performed: one with the original PADI family MSA and another one with the trimmed version of the alignment.

Estimation of dN/dS and positive selection

Estimates were made for both the number of non-synonymous substitutions per non-synonymous site (dN) and the number of synonymous substitutions per synonymous site (dS) using the free-ratio model in CodeML (62). For each branch and leaf in the tree, we performed a branch-site test of positive selection (63), implemented in the Phylogenetic Analysis by Maximum Likelihood (PAML) software package (62) and available in the environment for tree exploration *(*ETE V3*)* framework as bsA/bsA1 (64). A likelihood ratio (LRT) was calculated as 2*(L1-L0), where L1 and L0 are the maximum likelihood value for the alternative hypothesis and the null hypothesis respectively. A chi-squared distribution with 1 degree of freedom was used to calculate the p-values. Only the positions for which the p-value (< 0.05) is significant for positive selection are reported.

Characterization of positively selected residues using MVORFFIP

The prediction of interface residues was done using MVORFFIP (65). MVORFFIP is a structure-based prediction method that identifies protein-, peptide-, RNA- and DNA interfaces based on a range of structural, evolutionary, experimental, and energy-based information integrated by a Random Forest classifier The structure of PADI2 was submitted to the MVORFFIP server (http://www.bioinsilico.org/MVORFFIP) and protein interfaces scores were assigned to individual amino acids.

Structural modeling of PADI2-CDK9/CCNT1 and selection of putative interface residues.

The structure of the trimer complex the PADI2-CDK9 was derived using protein docking as follows. The crystal structure of PADI2 (PDB code: 4n2a) (24) and CDK9/CCNT1 (PDB code: 3blr) (66) were available. The protein docking was performed using VD^2OCK (67), generating the ensemble of over 10000 docking poses. Docking conformers were clustered using the GROMACS (68) and sorted by the ES3DC potential (69). The top 200 poses were selected to identify the top 10 putative interface residues as shown next. For this, a composite score was calculated using the docking information, the MVORFFIP score and sequence conservation. Among the top 200 docking poses as shown above a normalized score (Z score) was calculated for each exposed residue in PADI2. The mean and standard deviations to calculate the Z score were computed over the entire docking space, i.e. considering all docking poses. The MVORFFIP score was also computed as shown above. Finally, the conservation of each residue was computed using al2co (70).

AlphaFold-Multimer modeling of PADI2-CDK9/CCNT1 complex

Structural models of the trimer PADI2-CDK9/CCNT1 were generated using AlphaFold-Multimer (71) as follow. The sequences were retrieved as mentioned above, PADI2 (PDB code: 4n2a) (24) and CDK9/CCNT1 (PDB code: 3blr) (66). The structure was derived using the model parameters corresponding to “model_1_multimer”.

Results

Evolutionary analysis of the PADI family

To obtain a global picture of evolutionary relations within the PADI family, we first searched for all PADI family members with sequence homology to the human PADI2 in the OMA (Orthologous MAtrix) database (53) and gathered all the available sequences in mammals (see Materials and Methods). After close inspection, we discarded partial sequences and low-quality genomes and selected 185 confident sequences originating from 25 mammalian species and 3 bird

species (**Supplementary Figure S1**). With these sequences, we built a multiple sequence alignment (MSA, **Supplementary Material File S1**) and reconstructed a phylogenetic tree using the software RaxM, rooting the tree using the bird sequences as outgroups (**Supplementary Figure S2,** see Materials and Methods).

The tree we built is consistent with the 5 genes being related and resulting from duplication. Indeed, each of the five well-defined clusters in the reconstructed tree corresponds to the five different *PADI* genes (*PADI1*, green; *PADI2* purple; PADI*3*, blue; *PADI4*, red; and *PADI6*, orange) with the species tree being broadly recapitulated within each of these clusters (**Supplementary Figure S2)**. The tree comes along with robust bootstrap supports (82-100, **Supplementary Figure S3**) and therefore constitutes strong evidence for the genes within each cluster to be orthologous to one another. We observed clustering between *PADI1* and *PADI3*, and between *PADI4* and *PADI6*, with PADI2 being the most distant copy among members. Notably, PADI2 also has shorter branch lengths, indicating a higher level of conservation than other family members. PADI6 is the member with the largest branch lengths, placing it as the most divergent and suggesting a recent duplication (72). Interestingly, there are only three PADI genes in each bird species, and only one set of clusters within the PADI2 family, suggesting that one or more duplication events took place after the divergence from birds. PADI2 is also the closest subfamily to the remaining PADI outgroup species clusters, indicative of an ancestral position in this family.

The two bird copies (i.e., ANAPL06996 and ANAPL06995 in duck) in the unpainted cluster in Supplementary Figure S2 did not exhibit a clear orthology assignation in either Ensembl or the OMA, with some annotated as *PADI1* copies. Our data suggest that there was a bird-specific duplication, possibly from *PADI1,* which is the closest gene in the synteny analysis (**Supplementary Figure S4A**). We also observed an inversion of the whole region that appears to have happened in the common ancestor of humans and gibbon (Hominoidea), as compared with other species (**Supplementary Figure S4B**). Most mammalian species have the same gene order as the one found in mice (**Supplementary Figure S4C**). These analyses helped us to characterize gene re-arrangements and duplications within mammals in the *PADI* locus.

PADI2 is highly conserved

As PADI2 is likely to be the ancestral copy, we decided to focus on the most recent changes occurring in its sequence. With the phylogenetic selection of species, we aimed to identify changes that have occurred since the common amniote ancestor, with a good resolution in different mammals.

Our dataset contains representatives in all the three extant major groups in the Mammalia class and the principal orders or infraclasses: primates, rodents, carnivores, cetartiodactyla, lagomorphs, proboscidea, and marsupials. We built a multiple sequence analysis with all 28 PADI2 sequences (see Materials and Methods**, Supplementary Material File S1)** and computed the ratio of substitution rates at non-synonymous sites and synonymous sites (dN/dS; also termed Ka/Ks) to infer the direction and magnitude of natural selection acting on protein-coding genes (73). The difficulty of estimating the dN/dS ratio grows with the phylogenetic distance between sequences. Maximum likelihood methods were used to correct for multiple substitutions in the same site, and we estimated important parameters, such as the divergence between sequences or the transition/transversion ratio by deducing the most likely values to produce the input data (73).

We calculated the dN/dS ratio for all internal branches and leaves in the tree using CodeML (**Figure 1**). We observed low dN/dS values, whereby most of the branches had a value of < 0.2. This was consistent with negative selection; in other words, changes in this genetic sequence were actively selected against. This analysis showed that PADI2 is a highly conserved protein, suggesting that it performs critical functions in the organism (**Figure 1**). However, we observed a relatively high dN/dS in the common ancestor of all primates, indicating either relaxation of selection at this moment or perhaps the fixation of a few beneficial mutations due to positive selection.

Detection of positively selected residues in PADI2

The dN/dS ratio requires a high number of changes to detect the effect of selection and we have already seen that dN/dS values are very low in PADI2. Another limitation of this method is that it does not distinguish amino acid substitutions that are chemically similar from others with very different properties. We therefore used the branch-site positive selection test to detect specific residues under selection (63, 74). This test compares a null model in which codon-based dN/dS for all branches can only be ≤ 1; in the alternative model, the labeled foreground branch may include codons evolving at dN/dS > 1 (63). This test can discriminate between the relaxation of selection or positive selection and can detect individual residues that are under positive selection in a particular lineage. After running the branch-site test of positive selection in every branch of the PADI2 tree, we detected 20 individual relevant substitutions in different parts of the tree. We removed a few candidates for positively selected amino acids detected by CodeML in *Loxodonta africana* due to alignment of the non-homologous region (238-278), possibly due to an assembly error in that species. The residues identified in the different species or branches are shown in **Table 1 and Figure 2**.

Positively selected residues at the NTD and middle domain of PADI2 are likely to be involved in mediating protein-protein interactions.

We next explored the structural characterization of the detected residues specific to PADI2 in the different species, to integrate the evolutionary analysis with the known structural information and to study the properties of positively selected residues. For this, we mapped all the positively selected amino acids onto the structure of PADI2, with the aim of further understanding its potential role based on structural information. Twelve of the 20 positively selected amino acids mapped onto the: PADI2_N and PADI2_M domains, and 8 of them, onto the PADI2_C domain. The latter contains the catalytic pocket and is responsible for the citrullination of arginine residues. (**Table 1, Figure 3A-B**). Of note, only one of the positively selected amino acids (L25) is fully buried (amino acids are numbered as shown in **Table 1, Supplementary File 2, Supplementary Movie S1**); most of the other amino acids are fully exposed except for L342, V538, and M663, which are partially buried. In the secondary structure, around 50% of the amino acids are located in loops, while the remaining are either in beta strands or alfa helices (including helix-capping regions) (**Table 1**).

After evaluating the features of PADI2_N and PADI2_M domains and analyzing each particular amino acid and its potential role based on structural data, we concluded that the likely roles of E16, V53, E60, K136, L162, V201, S245, S249, S267, and T289 are to mediate interactions with putative partners. These amino acids are fully exposed, and their chemical properties correspond with classical protein interfaces (65, 75). We thus characterized these amino acids using Multi-VORFFIP (MVORFFIP), a tool to predict protein-, peptide-, DNA-, and RNA-binding sites for these positively selected residues (65). Notably, the MVORFFIP predictions for all except V201and L25 yielded very high scores (> 0.7, on a scale of 0 to 1), therefore predicting them to be interface residues (**Figure 3C-D, Table 1**). S263 was located in a long loop close to the interface between domains, suggesting that it plays a role in hinge motions and geometrical orientation between domains. Given its position in the structure, the likely role of L25 (the only positively selected amino acid that is completely buried) is to prevent the distortion of the packing on the beta-sandwich. Indeed, a larger hydrophobic amino acid at this position would change the packing of the beta-strands.

Concerning the positively selected position in the PADI2_C domain, L342, F380, and S401appeared close to the catalytic pocket. These residues are likely to play a role in substrate specificity (L342) and in the dynamics of the catalysis (F380 and S401). In turn, K452, S536, and V538 are located in the helix and quite far away from the active pocket. Judging by the structural microenvironment, these residues could also play a role in protein-protein interactions, in line with the high scores assigned by MVORFFIP (**Figure 3D, Table 1)** (65). Finally, D507 and M663 are

both located in helix cappings. Conservation in helix capping is important for the stability and integrity of the helix; in the case of M663, it also might play a role in the packing of the C-terminal tail. A search on the CAPS Database (76) showed that other helices with the same capping structure present a conserved small hydrophobic in the same position by including M, V, or L. It is noteworthy that M663 was indeed strictly conserved across all human PADIs. These observations suggested that most of these positively selected amino acids most likely have key functions in facilitating and stabilizing the protein -protein interactions.

The middle domain of PADI2 contributes to maintaining its interactions with the P-TEFb complex

We next addressed how evolutionary positively selected residues at the non-catalytic domain of PADI2 could be involved in PADI2's interactions with other proteins. For this, we targeted the P-TEFb kinase complex, which we previously found to interact with PADI2 (29). Indeed, we reported that citrullination of R1810 at the CTD of RNAP2 facilitates its interaction with the P-TEFb kinase complexes and promotes chromatin recruitment of the CDK9 to the transcription start sites (TSS) (29). This contributes to overcoming the RNAP2 pausing barrier, highlighting the functional connection of PADI2 with the P-TEFb kinase complex.

The structures of the PADI2 (24) and CDK9-CCNT1 (66) are known individually, and therefore it was possible to derive the structural model of the PADI2–CDK9/CCNT1 complex by protein docking. After deriving the docking ensemble of PADI2-CDK9/CCNT, we devised a ranking of models. An important aspect of the selection of the putative model of the interaction was the swarm-based approach (among other metrics), i.e., determining the region of PADI2 would be more relevant to its interaction with CDK9/CCNT1 (**Figure 4A, Supplementary File 3, Supplementary Movie S2**). Moreover, we derived the structural model by using AlphaFold-Multimer (71, 77). As shown in **Figure 4B (Supplementary File 4, Supplementary Movie S3)**, the structural model of AlphaFold-Multimer overlaps substantially with the same regions identified by docking (**Figure 4A**). Namely, the predicted interface between PADI2 and CDK9-CCNT1 contained PADI2_M and the hinge region with PADI2_C domain, and in particular the regions with a number of positively selected residues (the most important of which is L162) (**Figure 5A**).

While AlphaFold-Multimer generates a single model, docking is represented by a range of docking poses that allow us to identify which region of the predicted interface, i.e PADI2_M and

PADI2_C (**Figure 5A**) is over-represented among docking space. The distribution of the top 200 docking poses (represented using a unique point depicting the center of mass of CDK9/CCNT1) revealed that the region around the hinge between the PADI2_M and PADI2_C was overly represented **(Figure 5B, Supplementary File 5, Supplementary Movie S4)**. This particular region includes L162, a residue overrepresented in the docking conformers, along with T159 and W161. These three residues are located in a highly exposed region, making them candidates for experimental validation. Note that L162 in PADI2 was present among the top-ranking interface residues likely to meditate the interaction with CDK9/CCNT1 and is also one of the positively selected residues, suggesting that it is important for maintaining this interaction.

Positively selected Leu162 is important for cell proliferation and maintaining PADI2 interactions with the P-TEFb complex

Considering that the P-TEFb kinase complex is involved in transcription elongation and cell proliferation (78–80), we experimentally tested whether the L162-encompassing loop in the middle domain of PADI2 affects cell proliferation. In addition to L162, we also tested its neighboring amino acids, W161 and T159, which also ranked highly, and are located in a highly exposed loop region structurally close to the boundary between the PADI2_M and the catalytic domain. We selectively expressed the green fluorescence protein (GFP)-tagged wild-type (WT) PADI2 as well as the single (T159A or W161A or L162A), double (L162A/W161A), or triple (L162A/W161A/T159A) mutant of PADI2. The GFP-positive cells were sorted by FACS to ensure positive cell selection (**Supplementary Figure S5A, B**). We monitored the cell proliferation in HeLa cells expressing the GFP-tagged PADI2 WT and mutants (single, T159A or W161A or L162A, double, L162A/W161A, and triple, L162A/W161A/T159A). Strikingly, we observed that among the three single PADI2 mutants, L162A significantly decreased cell proliferation compared to WT, highlighting the functional relevance of L162 in the middle domain of PADI2 (**Figure 6A, Figure S5C).** In the same manner, the PADI2 double mutant L162A/W161A, as well as the triple mutant L162A/W161A/T159A, showed a further significant reduction in cell proliferation. These observations highlight the functional role of L162A in this remarkably exposed loop in the middle domain of the PADI2.

As a result, we then focused on the L162A, L162A/W161A, and L162A/W161A/T159A mutants along with WT PADI2. In the immunoprecipitation assay, using a GFP-tagged antibody, we observed that the P-TEFb complex (containing CCNT1 and CDK9) from HeLa cells was efficiently immunoprecipitated with GFP-WT PADI2 but not with GFP-L162A-PADI2 (**Figure 6B).** These

results confirmed the functional role of the positively selected L162 in maintaining PADI2´s interactions with P-TEFb complex. Likewise, the P-TEFb complex was only weakly immunoprecipitated with either the double or triple PAD12 mutant, highlighting the function of the highly exposed region in the middle domain of PADI2 (PADI2_M) in this protein-protein interaction. Considering that proper recruitment of the P-TEFb complex is required for efficient transcription of highly expressed genes relevant to cell proliferation, we analyzed the expression levels of the *FOS*, *c-MYC*, and *CCND1* genes in cells expressing the single, double, or triple PADI2 mutant. Of note, the levels of PADI2 expression did not differ significantly (**Figure 6C).** We observed that the L162A single mutant (as well as double and triple mutants) significantly and specifically reduced the expression of highly expressed genes, including *FOS*, *c-MYC*, and *CCND2,* without affecting the levels of a control gene (the low-expressed *LRRC39* gene). To examine if the L162A exposed loop of PADI2 is important for CDK9 recruitment to the promoter region of the target genes, we performed chromatin immunoprecipitation (ChIP) assay with CDK9 specific antibody in HeLa cells specifically expressing WT, L162A single mutant, as well as double and triple mutants. We found that CDK9 occupancy decreases in the presence of mutants in comparison to the WT PADI2, suggesting that the L162 loop is important to maintaining PADI2 interaction with the P-TEFb complex. These results support the functional relevance of the positively selected L162 encompassing loop present in the middle domain of PADI2 in transcription regulation.

Discussion

We characterized the evolution of the PADI gene family by using the most complete dataset of mammalian orthologous sequences to date. There are three extant major groups in the Mammalia class, which are organized into two subclasses: Prototheria which includes monotremes (platypus and echidnas), and Theria, which includes the infraclasses Methatheria (marsupials), and Eutheria (placental mammals). According to Jones et al. 2009 (81), there are 5416 species divided into 29 orders. Most of the species described belong to the 7 biggest orders (4865 species), and all of them have at least one representative in our used phylogeny.

The functional relevance of PADI2 is highlighted by its high sequence conservation across species, with very low dN/dS values. We took advantage of the natural experiments performed by evolution and applied sensitive methodologies such as the branch-site test of positive selection. Comparing the amino acid sequence of 25 mammalian species, we identified 20 positively selected residues predominantly located in the non-catalytic domain of the PADI2 in different species and phylogenetic branches **(Figures 1 and 2, Supplementary Figures S1-S4**). As these selected residues

could have a strong functional impact on PADI2, and therefore we studied them further, using the human sequence.

By examining the location of the positively selected residues in the structure of the human version of PADI2, modeled in a complex with R1810 in the CTD of RNAP2, we determined that the majority of the PADI2 unique residues were structurally highly exposed in the non-catalytic domains of PADI2 (**Figure 3, Table 1**). Notably, the highly exposed nature and chemical properties of the positively selected residue on the non-catalytic domain of the PADI2 suggested their possible function in cellular processes. In addition, our MVORFFIP analysis revealed a possible significance of these interface residues in protein-protein interactions (**Figure 3D**). Therefore, it could be postulated that the evolutionary appearance of these positively selected residues across the evolutionary lineages contributed to the function of PADI2 by modulating its interaction with the essential cognate set of the proteins. Further work will be required to investigate these possibilities.

Specifically, the P-TEFb complex, comprising CCNT1 and CDK9, interacts with PADI2 and facilitates the effects of citrullination at R1810 in RNAP2 on transcription (29). Hence, we used the structure modeling approaches to derive the tertiary structure of the PADI2–CCNT1/CDK9 complex. To identify the PADI2 residues that contribute to generating the PADI2–CCNT1/CDK9 complex, we analyzed the structural models and identified a highly exposed loop at the PADI2_M domain as a potential interacting region. Strikingly, this loop includes the positively evolved residue L162, along with T159 and W161. This observation supports the notion that positively evolved residues tend to coordinate with the important residues on the protein surface, and that positively selected structural clusters are important for cellular function (**Figure S5D**).

Our data in HeLa cells overexpressing the GFP-tagged WT and a mutant PADI2 also unveiled an essential function of the positively evolved L162 encircling loop in the middle domain of PADI2 in supporting its interaction with the P-TEFb kinase complex and cell proliferation. Notably, mutation of the L162, alone or in combination with mutated W161 and/or T159, reduced (i) cell proliferation, (ii) PADI2 interaction with P-TEFb kinase complex, and (iii) reduced the PADI2-dependent expression of highly expressed genes that are relevant for cell growth. Note that, as we were not able to mutate the endogenous PADI2 gene due to the hyperpolyploidy nature of the HeLa genome, therefore, one limitation of our study is that our data are based on overexpression of GFP-tagged wild-type and mutant versions of PADI2 in HeLa cells.

We focused on the interactions of PADI2 with the P-TEFb complex, but we cannot exclude other additional steps in gene expression that are affected by the positively selected amino acids in PADI2. Nonetheless, our analysis shows the importance of using a multi-disciplinary approach here, comparative genomics, evolutionary analysis, structural modeling, and genetic perturbations to analyze the functional relevance of a recently evolved non-catalytic domain of enzyme families. Overall, this evolutionary approach led us to identify a selective PADI2 structural loop that specifically confers PADI2 with an important basic role in the modulation of transcription elongation.

Data Availability

The OMA is an open-source available database for the inference of orthologs among complete genomes https://omabrowser.org/oma/home/. The ETE3 was used for the reconstruction, analysis, and visualization of phylogenetic trees is freely available at http://etetoolkit.org. Other open-source programs used here were GROMACS, VD2OCK, MVORFFIP and al2co, which are available at http://www.gromacs.org, http://www.bioinsilico.org/VD2OCK, http://www.bioinsilico.org/MVORFFIP, and http://prodata.swmed.edu/download/pub/al2co/, respectively. The Supplementary File S2.pse, Supplementary File S3.pse, Supplementary File S4.pse, and Supplementary File S5.pse correspond to PyMOL (http://pymol.org) sessions created to produce Supplementary Movie S1.mpg, Supplementary Movie S2.mpg, Supplementary Movie S3.mpg, and Supplementary Movie S4.mpg respectively. The PyMOL sessions also give access to the coordinates of the PADI2 and structural model of the PADI2/CDK9-CCNT1 complex.

Supplementary Data

Supplementary data are available online.

Author Contributions: Conceptualization, P.S.; Methodology, J.L.V., N.F.F., P.S.; Investigation, J.L.V., N.F.F., B.O., M.B., and P.S.; Formal Analysis, J.L.V., N.F.F., and P.S.; Data Curation, J.L.V., N.F.F., M.E.R., C.T., F.P.P., B.O., C.N., P.S.; Project Administration, P.S., Writing-Original Draft, P.S. Writing-Review & Editing, J.L.V., N.F.F., B.O., M.E.R., C.T., F.P.P., M.B., C.N., P.S.; Funding Acquisition, P.S.; Supervision, P.S.

Funding

This work was supported by grants to P.S. from the National Institute of Health and Medical Research (INSERM) Young Recruitment Support (U1194SHA), Cancéropôle Grand Sud-Ouest collaboration grant (R21031FF), and the French National Research Agency (ANR) Young Investigator grant

(ANR-21-CE12-0010). This work was also supported by a grant to M.B. from the European Research Council Synergy Grant "4DGenome" (609989), Spanish Ministry of Innovation (PID2019-105173RBI00 and PID2019-110384GB-C219).

Declaration of Interests

The authors declare no competing financial interests.

Acknowledgment

We thank Fátima Gebauer and François Le Dily from CRG, Stéphan Jalaguier and Vincent Cavaillès from IRCM, for their constructive criticism and advice on this manuscript. We acknowledge Veronica A. Raker for manuscript editing.

References

1. van Venrooij,W.J. and Pruijn,G.J.M. (2000) Citrullination: A small change for a protein with great consequences for rheumatoid arthritis. *Arthritis Research*, **2**, 249–251.
https://doi.org/10.1186/ar95
http://www.ncbi.nlm.nih.gov/pubmed/11094435

2. Wang,S. and Wang,Y. (2013) Peptidylarginine deiminases in citrullination, gene regulation, health and pathogenesis. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, **1829**, 1126–1135.
https://doi.org/10.1016/j.bbagrm.2013.07.003
http://www.ncbi.nlm.nih.gov/pubmed/23860259

3. Fuhrmann,J., Clancy,K.W. and Thompson,P.R. (2015) Chemical Biology of Protein Arginine Modifications in Epigenetic Regulation. *Chemical Reviews*, **115**, 5413–5461.
https://doi.org/10.1021/acs.chemrev.5b00003
http://www.ncbi.nlm.nih.gov/pubmed/25970731

4. Christophorou,M.A. (2022) The virtues and vices of protein citrullination. *Royal Society Open Science*, **9**, 79–95.
https://doi.org/10.1098/rsos.220125

5. Vossenaar,E.R., Zendman,A.J.W., Van Venrooij,W.J. and Pruijn,G.J.M. (2003) PAD, a growing family of citrullinating enzymes: Genes, features and involvement in disease. *BioEssays*, **25**, 1106–1118.
https://doi.org/10.1002/bies.10357
http://www.ncbi.nlm.nih.gov/pubmed/14579251

6. Tanikawa,C., Ueda,K., Suzuki,A., Iida,A., Nakamura,R., Atsuta,N., Tohnai,G., Sobue,G., Saichi,N., Momozawa,Y., *et al.* (2018) Citrullination of RGG Motifs in FET Proteins by PAD4 Regulates Protein Aggregation and ALS Susceptibility. *Cell Reports*, **22**, 1473–1483.
https://doi.org/10.1016/j.celrep.2018.01.031
http://www.ncbi.nlm.nih.gov/pubmed/29425503

7. Christophorou,M.A., Castelo-Branco,G., Halley-Stott,R.P., Oliveira,C.S., Loos,R., Radzisheuskaya,A., Mowen,K.A., Bertone,P., Silva,J.C.R., Zernicka-Goetz,M., *et al.* (2014) Citrullination regulates pluripotency and histone H1 binding to chromatin. *Nature*, **507**, 104–108.
https://doi.org/10.1038/nature12942
http://www.ncbi.nlm.nih.gov/pubmed/24463520

8. Sharma,P., Azebi,S., England,P., Christensen,T., Møller-Larsen,A., Petersen,T., Batsché,E. and Muchardt,C. (2012) Citrullination of Histone H3 Interferes with HP1-Mediated Transcriptional Repression. *PLoS Genetics*, **8**, 1–15.
https://doi.org/10.1371/journal.pgen.1002934
http://www.ncbi.nlm.nih.gov/pubmed/23028349

9. Chang,X., Xia,Y., Pan,J., Meng,Q., Zhao,Y. and Yan,X. (2013) PADI2 is significantly associated with rheumatoid arthritis. *PLoS ONE*, **8**.
https://doi.org/10.1371/journal.pone.0081259
http://www.ncbi.nlm.nih.gov/pubmed/24339914

10. Ishigami,A., Ohsawa,T., Hiratsuka,M., Taguchi,H., Kobayashi,S., Saito,Y., Murayama,S., Asaga,H., Toda,T., Kimura,N., *et al.* (2005) Abnormal accumulation of citrullinated proteins catalyzed by peptidylarginine deiminase in hippocampal extracts from patients with Alzheimer's disease. *Journal of Neuroscience Research*, **80**, 120–128.
https://doi.org/10.1002/jnr.20431
http://www.ncbi.nlm.nih.gov/pubmed/15704193

11. Arif,M. and Kato,T. (2009) Increased expression of PAD2 after repeated intracerebroventricular infusions of soluble Aβ25-35 in the Alzheimer's disease model rat brain: Effect of memantine. *Cellular and Molecular Biology Letters*, **14**, 703–714.
https://doi.org/10.2478/s11658-009-0029-x
http://www.ncbi.nlm.nih.gov/pubmed/19641855

12. Cherrington,B.D., Zhang,X., Mcelwee,J.L., Morency,E., Anguish,L.J. and Coonrod,S.A. (2012) Potential Role for PAD2 in Gene Regulation in Breast Cancer Cells. *PLoS One*, **7**, e41242.
https://doi.org/10.1371/journal.pone.0041242

13. Mohanan,S., Cherrington,B.D., Horibata,S., Mcelwee,J.L., Thompson,P.R. and Coonrod,S.A. (2012) Potential Role of Peptidylarginine Deiminase Enzymes and Protein Citrullination in Cancer Pathogenesis. *Biochemistry Research International*, **2012**, 895343.
https://doi.org/10.1155/2012/895343

14. Guo W, Zheng Y, Xu B, Ma F, Li C, Zhang X, Wang Y,C.X. (2017) Investigating the expression , effect and tumorigenic pathway of PADI2 in tumors. *OncoTargets and therapy*, **10**, 1475–1485.

15. Wang,L., Song,G., Zhang,X., Feng,T., Pan,J., Chen,W., Yang,M., Bai,X., Pang,Y., Yu,J., *et al.* (2017) PADI2-mediated citrullination promotes prostate cancer progression. *Cancer Research*, **77**, 5755–5768.
https://doi.org/10.1158/0008-5472.CAN-17-0150
http://www.ncbi.nlm.nih.gov/pubmed/28819028

16. Horibata,S., Rogers,K.E., Sadegh,D., Anguish,L.J., Mcelwee,J.L., Shah,P., Thompson,P.R.

and Coonrod,S.A. (2017) Role of peptidylarginine deiminase 2 (PAD2) in mammary carcinoma cell migration. *BMC Cancer*, **17**, 378.
https://doi.org/10.1186/s12885-017-3354-x

17. Song,S. and Yu,Y. (2019) Progression on citrullination of proteins in gastrointestinal cancers. *Frontiers in Oncology*, **9**, 1–6.
https://doi.org/10.3389/fonc.2019.00015

18. Yuzhalin,A.E. (2019) Citrullination in cancer. *Cancer Research*, **79**, 1274–1284.
https://doi.org/10.1158/0008-5472.CAN-18-2797
http://www.ncbi.nlm.nih.gov/pubmed/30894374

19. Beato,M. and Sharma,P. (2020) Peptidyl arginine deiminase 2 (PADI2)-mediated arginine citrullination modulates transcription in cancer. *International Journal of Molecular Sciences*, **21**, 1–16.
https://doi.org/10.3390/ijms21041351
http://www.ncbi.nlm.nih.gov/pubmed/32079300

20. Gao,B. shan, Rong,C. shu, Xu,H. mei, Sun,T., Hou,J. and Xu,Y. (2020) Peptidyl Arginine Deiminase, Type II (PADI2) Is Involved in Urothelial Bladder Cancer. *Pathology and Oncology Research*, **26**, 1279–1285.
https://doi.org/10.1007/s12253-019-00687-0
http://www.ncbi.nlm.nih.gov/pubmed/31267364

21. Xue,T., Liu,X., Zhang,M., Qiukai,E., Liu,S., Zou,M., Li,Y., Ma,Z., Han,Y., Thompson,P., *et al.* (2021) PADI2-Catalyzed MEK1 Citrullination Activates ERK1/2 and Promotes IGF2BP1-Mediated SOX2 mRNA Stability in Endometrial Cancer. *Advanced Science*, **8**, 1–17.
https://doi.org/10.1002/advs.202002831

22. Darrah,E. 2012 (2012) Peptidylarginine deiminase 2, 3 and 4 have distinct specificities against cellular substrates: Novel insights into autoantigen selection in rheumatoid arthritis Erika. *Ann Rheum Dis.*, **71**, 92–98.
https://doi.org/10.1136/ard.2011.151712.Peptidylarginine
http://www.ncbi.nlm.nih.gov/pubmed/1000000221

23. Dreyton,C.J., Knuckley,B., Jones,J.E., Lewallen,D.M. and Thompson,P.R. (2014) Mechanistic studies of protein arginine deiminase 2: Evidence for a substrate-assisted mechanism. *Biochemistry*, **53**, 4426–4433.
https://doi.org/10.1021/bi500554b
http://www.ncbi.nlm.nih.gov/pubmed/24989433

24. Slade,D.J., Fang,P., Dreyton,C.J., Zhang,Y., Fuhrmann,J., Rempel,D., Bax,B.D., Coonrod,S.A., Lewis,H.D., Guo,M., *et al.* (2015) Protein arginine deiminase 2 binds calcium in an ordered fashion: Implications for inhibitor design. *ACS Chemical Biology*, **10**, 1043–1053.
https://doi.org/10.1021/cb500933j
http://www.ncbi.nlm.nih.gov/pubmed/25621824

25. Zhang,X., Bolt,M., Guertin,M.J., Chen,W., Zhang,S., Cherrington,B.D., Slade,D.J., Dreyton,C.J., Subramanian,V., Bicker,K.L., *et al.* (2012) Peptidylarginine deiminase 2-catalyzed histone H3 arginine 26 citrullination facilitates estrogen receptor α target gene

activation. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 13331–13336.
https://doi.org/10.1073/pnas.1203280109
http://www.ncbi.nlm.nih.gov/pubmed/22853951

26. Guertin,M.J., Zhang,X., Anguish,L., Kim,S., Varticovski,L., Lis,J.T., Hager,G.L. and Coonrod,S.A. (2014) Targeted H3R26 Deimination Specifically Facilitates Estrogen Receptor Binding by Modifying Nucleosome Structure. *PLoS Genetics*, **10**, 1–12.
https://doi.org/10.1371/journal.pgen.1004613
http://www.ncbi.nlm.nih.gov/pubmed/25211228

27. Xiao,S., Lu,J., Sridhar,B., Cao,X., Yu,P., Zhao,T., Chen,C.C., McDee,D., Sloofman,L., Wang,Y., *et al.* (2017) SMARCAD1 Contributes to the Regulation of Naive Pluripotency by Interacting with Histone Citrullination. *Cell Reports*, **18**, 3117–3128.
https://doi.org/10.1016/j.celrep.2017.02.070
http://www.ncbi.nlm.nih.gov/pubmed/28355564

28. Falcão,A.M., Meijer,M., Scaglione,A., Rinwa,P., Agirre,E., Liang,J., Larsen,S.C., Heskol,A., Frawley,R., Klingener,M., *et al.* (2019) PAD2-Mediated Citrullination Contributes to Efficient Oligodendrocyte Differentiation and Myelination. *Cell Reports*, **27**, 1090-1102.e10.
https://doi.org/10.1016/j.celrep.2019.03.108
http://www.ncbi.nlm.nih.gov/pubmed/31018126

29. Sharma,P., Lioutas,A., Fernandez-Fuentes,N., Quilez,J., Carbonell-Caballero,J., Wright,R.H.G., Di Vona,C., Le Dily,F., Schüller,R., Eick,D., *et al.* (2019) Arginine Citrullination at the C-Terminal Domain Controls RNA Polymerase II Transcription. *Molecular Cell*, **73**, 84-96.e7.
https://doi.org/10.1016/j.molcel.2018.10.016
http://www.ncbi.nlm.nih.gov/pubmed/30472187

30. Cummings,T.F.M., Gori,K., Sanchez-Pulido,L., Gavriilidis,G., Moi,D., Wilson,A.R., Murchison,E., Dessimoz,C., Ponting,C.P. and Christophorou,M.A. (2022) Citrullination Was Introduced into Animals by Horizontal Gene Transfer from Cyanobacteria. *Molecular Biology and Evolution*, **39**.
https://doi.org/10.1093/molbev/msab317
http://www.ncbi.nlm.nih.gov/pubmed/34730808

31. Crisp,A., Boschetti,C., Perry,M., Tunnacliffe,A. and Micklem,G. (2015) Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biology*, **16**, 1–13.
https://doi.org/10.1186/s13059-015-0607-3
http://www.ncbi.nlm.nih.gov/pubmed/25785303

32. Emamalipour,M., Seidi,K., Zununi Vahed,S., Jahanban-Esfahlan,A., Jaymand,M., Majdi,H., Amoozgar,Z., Chitkushev,L.T., Javaheri,T., Jahanban-Esfahlan,R., *et al.* (2020) Horizontal Gene Transfer: From Evolutionary Flexibility to Disease Progression. *Frontiers in Cell and Developmental Biology*, **8**.
https://doi.org/10.3389/fcell.2020.00229

33. Soucy,S.M., Huang,J. and Gogarten,J.P. (2015) Horizontal gene transfer: Building the web of life. *Nature Reviews Genetics*, **16**, 472–482.

https://doi.org/10.1038/nrg3962
http://www.ncbi.nlm.nih.gov/pubmed/26184597

34. Husnik,F. and McCutcheon,J.P. (2018) Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology*, **16**, 67–79.
https://doi.org/10.1038/nrmicro.2017.137
http://www.ncbi.nlm.nih.gov/pubmed/29176581

35. Goulas,T., Mizgalska,D., Garcia-Ferrer,I., Kantyka,T., Guevara,T., Szmigielski,B., Sroka,A., Millan,C., Uson,I., Veillard,F., *et al.* (2015) Structure and mechanism of a bacterial host-protein citrullinating virulence factor, Porphyromonas gingivalis peptidylarginine deiminase. *Scientific Reports*, **5**, 1–17.
https://doi.org/10.1038/srep11969
http://www.ncbi.nlm.nih.gov/pubmed/26132828

36. Touz,M.C., Rópolo,A.S., Rivero,M.R., Vranych,C.V., Conrad,J.T., Svard,S.G. and Nash,T.E. (2008) Arginine deiminase has multiple regulatory roles in the biology of Giardia lamblia. *Journal of Cell Science*, **121**, 2930–2938.
https://doi.org/10.1242/jcs.026963
http://www.ncbi.nlm.nih.gov/pubmed/18697833

37. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: The protein families database in 2021. *Nucleic Acids Research*, **49**, D412–D419.
https://doi.org/10.1093/nar/gkaa913
http://www.ncbi.nlm.nih.gov/pubmed/33125078

38. Asaga,H. and Ishigami,A. (2001) Protein deimination in the rat brain after kainate administration: Citrulline-containing proteins as a novel marker of neurodegeneration. *Neuroscience Letters*, **299**, 5–8.
https://doi.org/10.1016/S0304-3940(00)01735-3
http://www.ncbi.nlm.nih.gov/pubmed/11166924

39. Rogers,G., Winter,B., McLaughlan,C., Powell,B. and Nesci,T. (1997) Peptidylarginine deiminase of the hair follicle: Characterization, localization, and function in keratinizing tissues. *Journal of Investigative Dermatology*, **108**, 700–707.
https://doi.org/10.1111/1523-1747.ep12292083
http://www.ncbi.nlm.nih.gov/pubmed/9129218

40. Rus'd,A.A., Ikejiri,Y., Ono,H., Yonekawa,T., Shiraiwa,M., Kawada,A. and Takahara,H. (1999) Molecular cloning of cDNAs of mouse peptidylarginine deiminase type I, type III and type IV, and the expression pattern of type I in mouse. *European Journal of Biochemistry*, **259**, 660–669.
https://doi.org/10.1046/j.1432-1327.1999.00083.x
http://www.ncbi.nlm.nih.gov/pubmed/10092850

41. Beltrao,P., Bork,P., Krogan,N.J. and Van Noort,V. (2013) Evolution and functional cross-talk of protein post-translational modifications. *Molecular Systems Biology*, **9**, 1–13.
https://doi.org/10.1002/msb.201304521
http://www.ncbi.nlm.nih.gov/pubmed/24366814

42. Kumar,S., Stecher,G., Suleski,M. and Hedges,S.B. (2017) TimeTree: A Resource for

Timelines, Timetrees, and Divergence Times. *Molecular biology and evolution*, **34**, 1812–1819.
https://doi.org/10.1093/molbev/msx116
http://www.ncbi.nlm.nih.gov/pubmed/28387841

43. Corden,J.L. (2019) An Arginine Nexus in the RNA Polymerase II CTD. *Molecular Cell*, **73**, 3–4.
https://doi.org/10.1016/j.molcel.2018.12.013
http://www.ncbi.nlm.nih.gov/pubmed/30609390

44. Marshall,N.F. and Price,D.H. (1995) Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *Journal of Biological Chemistry*, **270**, 12335–12338.
https://doi.org/10.1074/jbc.270.21.12335
http://www.ncbi.nlm.nih.gov/pubmed/7759473

45. Adelman,K. and Lis,J.T. (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews. Genetics*, **13**, 720–731.
https://doi.org/10.1038/nrg3293
http://www.ncbi.nlm.nih.gov/pubmed/22986266

46. Jonkers,I. and Lis,J.T. (2015) Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, **16**, 167–177.
https://doi.org/10.1038/nrm3953
http://www.ncbi.nlm.nih.gov/pubmed/25693130

47. Core,L. and Adelman,K. (2019) Promoter-proximal pausing of RNA polymerase II: A nexus of gene regulation. *Genes and Development*, **33**, 960–982.
https://doi.org/10.1101/gad.325142.119
http://www.ncbi.nlm.nih.gov/pubmed/31123063

48. Chen,F.X., Smith,E.R. and Shilatifard,A. (2018) Born to run: Control of transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, **19**, 464–478.
https://doi.org/10.1038/s41580-018-0010-5
http://www.ncbi.nlm.nih.gov/pubmed/29740129

49. Martin,R.D., Hébert,T.E. and Tanny,J.C. (2020) Therapeutic targeting of the general RNA polymerase II transcription machinery. *International Journal of Molecular Sciences*, **21**.
https://doi.org/10.3390/ijms21093354
http://www.ncbi.nlm.nih.gov/pubmed/32397434

50. Muniz,L., Nicolas,E. and Trouche,D. (2021) RNA polymerase II speed: a key player in controlling and adapting transcriptome composition. *The EMBO Journal*, **40**, 1–21.
https://doi.org/10.15252/embj.2020105740
http://www.ncbi.nlm.nih.gov/pubmed/34254686

51. Gibson,D.G., Young,L., Chuang,R.Y., Venter,J.C., Hutchison,C.A. and Smith,H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, **6**, 343–345.
https://doi.org/10.1038/nmeth.1318
http://www.ncbi.nlm.nih.gov/pubmed/19363495

52. Vicent,G.P., Zaurin,R., Nacht,A.S., Li,A., Font-Mateu,J., Le Dily,F., Vermeulen,M., Mann,M. and Beato,M. (2009) Two chromatin remodeling activities cooperate during activation of hormone responsive promoters. *PLoS Genetics*, **5**.
https://doi.org/10.1371/journal.pgen.1000567
http://www.ncbi.nlm.nih.gov/pubmed/19609353

53. Altenhoff,A.M., Glover,N.M., Train,C.M., Kaleb,K., Warwick Vesztrocy,A., Dylus,D., De Farias,T.M., Zile,K., Stevenson,C., Long,J., *et al.* (2018) The OMA orthology database in 2018: Retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Research*, **46**, D477–D485.
https://doi.org/10.1093/nar/gkx1019
http://www.ncbi.nlm.nih.gov/pubmed/29106550

54. Veidenberg,A., Medlar,A. and Loytynoja,A. (2016) Wasabi: An integrated platform for evolutionary sequence analysis and data visualization. *Molecular Biology and Evolution*, **33**, 1126–1130.
https://doi.org/10.1093/molbev/msv333
http://www.ncbi.nlm.nih.gov/pubmed/26635364

55. Villanueva-Cañas,J.L., Ruiz-Orera,J., Agea,M.I., Gallo,M., Andreu,D. and Albà,M.M. (2017) New genes and functional innovation in mammals. *Genome Biology and Evolution*, **9**, 1886–1900.
https://doi.org/10.1093/gbe/evx136
http://www.ncbi.nlm.nih.gov/pubmed/28854603

56. Villanueva-Cañas,J.L., Laurie,S. and Albà,M.M. (2013) Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biology and Evolution*, **5**, 457–467.
https://doi.org/10.1093/gbe/evt017

57. Katoh,K., Misawa,K., Kuma,K.I. and Miyata,T. (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.
https://doi.org/10.1093/nar/gkf436
http://www.ncbi.nlm.nih.gov/pubmed/12136088

58. Larsson,A. (2014) AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, **30**, 3276–3278.
https://doi.org/10.1093/bioinformatics/btu531
http://www.ncbi.nlm.nih.gov/pubmed/25095880

59. Capella-Gutiérrez,S., Silla-Martínez,J.M. and Gabaldón,T. (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
https://doi.org/10.1093/bioinformatics/btp348
http://www.ncbi.nlm.nih.gov/pubmed/19505945

60. Suyama,M., Torrents,D. and Bork,P. (2006) PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, **34**, 609–612.
https://doi.org/10.1093/nar/gkl315
http://www.ncbi.nlm.nih.gov/pubmed/16845082

61. Stamatakis,A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
https://doi.org/10.1093/bioinformatics/btu033
http://www.ncbi.nlm.nih.gov/pubmed/24451623

62. Yang,Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.
https://doi.org/10.1093/molbev/msm088
http://www.ncbi.nlm.nih.gov/pubmed/17483113

63. Zhang,J., Nielsen,R. and Yang,Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*, **22**, 2472–2479.
https://doi.org/10.1093/molbev/msi237
http://www.ncbi.nlm.nih.gov/pubmed/16107592

64. Huerta-Cepas,J., Serra,F. and Bork,P. (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, **33**, 1635–1638.
https://doi.org/10.1093/molbev/msw046
http://www.ncbi.nlm.nih.gov/pubmed/26921390

65. Segura,J., Jones,P.F. and Fernandez-Fuentes,N. (2012) A holistic in silico approach to predict functional sites in protein structures. *Bioinformatics*, **28**, 1845–1850.
https://doi.org/10.1093/bioinformatics/bts269
http://www.ncbi.nlm.nih.gov/pubmed/22563069

66. Baumli,S., Lolli,G., Lowe,E.D., Troiani,S., Rusconi,L., Bullock,A.N., Debreczeni,J.É., Knapp,S. and Johnson,L.N. (2008) The structure of P-TEFb (CDK9/cyclin T1), its complex with flavopiridol and regulation by phosphorylation. *EMBO Journal*, **27**, 1907–1918.
https://doi.org/10.1038/emboj.2008.121
http://www.ncbi.nlm.nih.gov/pubmed/18566585

67. Segura,J., Marín-López,M.A., Jones,P.F., Oliva,B. and Fernandez-Fuentes,N. (2015) VORFFIP-driven dock: V-D2OCK, a fast, accurate protein docking strategy. *PLoS ONE*, **10**, 1–12.
https://doi.org/10.1371/journal.pone.0118107
http://www.ncbi.nlm.nih.gov/pubmed/25763838

68. Pronk,S., Páll,S., Schulz,R., Larsson,P., Bjelkmar,P., Apostolov,R., Shirts,M.R., Smith,J.C., Kasson,P.M., Van Der Spoel,D., *et al.* (2013) GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **29**, 845–854.
https://doi.org/10.1093/bioinformatics/btt055
http://www.ncbi.nlm.nih.gov/pubmed/23407358

69. Feliu,E., Aloy,P. and Oliva,B. (2011) On the analysis of protein-protein interactions via knowledge-based potentials for the prediction of protein-protein docking. *Protein Science*, **20**, 529–541.
https://doi.org/10.1002/pro.585
http://www.ncbi.nlm.nih.gov/pubmed/21432933

70. Pei,J. and Grishin,N. V. (2001) AL2CO: Calculation of positional conservation in a protein

sequence alignment. *Bioinformatics*, **17**, 700–712.
https://doi.org/10.1093/bioinformatics/17.8.700
http://www.ncbi.nlm.nih.gov/pubmed/11524371

71. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A,G.T., Žídek A, Bates R, Blackwell S, Yim J, Ronneberger O,B.S., Zielinski M, Bridgland A, Potapenko A, Cowie A, Tunyasuvunakool K,J.R. and Clancy E, Kohli P,J.J. and D.H.D. (2022) Protein complex prediction with AlphaFold-Multimer. *bioRxiv*.

72. Pegueroles,C., Laurie,S. and Albà,M.M. (2013) Accelerated evolution after gene duplication: A time-dependent process affecting just one copy. *Molecular Biology and Evolution*, **30**, 1830–1842.
https://doi.org/10.1093/molbev/mst083
http://www.ncbi.nlm.nih.gov/pubmed/23625888

73. Yang,Z. and Bielawski,J.R. (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, **15**, 496–503.
https://doi.org/10.1016/S0169-5347(00)01994-7
http://www.ncbi.nlm.nih.gov/pubmed/11114436

74. Kosiol,C., Vinař,T., Da Fonseca,R.R., Hubisz,M.J., Bustamante,C.D., Nielsen,R. and Siepel,A. (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genetics*, **4**.
https://doi.org/10.1371/journal.pgen.1000144
http://www.ncbi.nlm.nih.gov/pubmed/18670650

75. Jones,S., Heyningen,P. Van, Berman,H.M. and Thornton,J.M. (1999) Arcturus_Summer_Hba.
http://www.ncbi.nlm.nih.gov/pubmed/10222198

76. Segura,J., Oliva,B. and Fernandez-Fuentes,N. (2012) CAPS-DB: A structural classification of helix-capping motifs. *Nucleic Acids Research*, **40**, 479–485.
https://doi.org/10.1093/nar/gkr879
http://www.ncbi.nlm.nih.gov/pubmed/22021380

77. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
https://doi.org/10.1038/s41586-021-03819-2
http://www.ncbi.nlm.nih.gov/pubmed/34265844

78. Lin,F.R., Niparko,J.K. and Ferrucci, and L. (2014) 基因的改变NIH Public Access. *Bone*, **23**, 1–7.
https://doi.org/10.1146/annurev-biochem-052610-095910.RNA
http://www.ncbi.nlm.nih.gov/pubmed/1000000221

79. Kohoutek,J. (2009) P-TEFb- The final frontier. *Cell Division*, **4**, 19.
https://doi.org/10.1186/1747-1028-4-19

80. Manuscript,A. (2009) multi-tasking P-TEFb. **20**, 334–340.
https://doi.org/10.1016/j.ceb.2008.04.008.The

81. Jones,K.E., Bielby,J., Cardillo,M., Fritz,S.A., O'Dell,J., Orme,C.D.L., Safi,K.,

Sechrest,W., Boakes,E.H., Carbone,C., *et al.* (2009) PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, **90**, 2648–2648.
https://doi.org/10.1890/08-1494.1

Figures and Tables

Figure 1 (related to Supplementary Figures S1-S4 and Supplementary Material File S1) PADI2 conservation across species. (A) Diagram showing the main steps for the phylogenetic reconstruction and evolutionary analysis. Each box contains a list of the tools used. (**B**) The PADI2 gene tree showing the dN/dS ratios (red), for every branch as well as dN and dS (grey), for every branch. The common branch to all primates is highlighted in blue. Branch lengths are proportional to nucleotide substitutions.

Figure 2 (related to Table 1) Summary of the branch site test results for internal and terminal branches. (**A**) Summary of the multiple sequence alignment, showing only the columns of the positions that are detected under positive selection with the branch site test. The colored-blurred positions in between represent the rest of the alignment (illustrated in Supplementary File 1). The node number correspondence can be found in Supplementary Figure S1.

Figure 3 (related to Table 1, Supplementary Material File S2 and Movie S1) Mapping of positively selected amino acids onto the PADI2 structure. **A-B.** Ribbon representation of PADI2 showing the N-terminal domain (PADI2_N), the middle domain (PADI2_M), and the C-terminal domain (PADI2_C) in different shades of green (in green, bright green, and pale green, respectively). (**B)** The same as (A) but applying a 90-degree rotation over the longitudinal axis. Pink highlighted residues are specific to the respective orientation. PyMol session of PADI2 and the highlighted amino acids are available in Supplementary Material File S2. (**C**) Cartoon representation of PADI2 with positively selected amino acids colored by the MVORFFIP scores as shown in the heatmap in panel D; PADI2_N, PADI2_M and PADI2_C domains colored in black, grey, and white, respectively. (**D**) The heatmap is based on the MVORFFIP scores as mentioned in Table 1.

Figure 4 (related to Supplementary Material File S3, S4 and Movie S2, S3) Modeled structure of PADI2 with P-TEFb complex. Structural model of PADI2-CDK9/CCNT1 complex derived by: **(A)** Docking, showing one of the top dockings poses. (**B)** Using AlphaFold-Multimer; In both panels, PADI2 is shown in the cartoon representation using the same color scheme as in Figure 3, and CDK9 and CCNT1 are shown in the cartoon representation in cyan and magenta, respectively; L162 is shown in blue using a sphere representation. The PyMol session of PADI2 with the P-TEFb complex with docking and AlphaFold-Multimer is available in Supplementary Material File S3 and S4, respectively**.**

Figure 5 (related to Supplementary Material File S5 and Movie S4). The predicted interface between PADI2 and the P-TEFb complex. (A) Surface representation of PADI2 with the same color scheme as in Figure 3 with the consensus interface with P-TEFb complex derived from docking and AlphaFold-Multimer structural models highlighted in red and L162 in sphered and blue. (**B**) The same representation as in (A) including the top 200 docking poses represented as red spheres depicting the center of mass of CDK9/CCNT1; T159 and W161 are also shown in cyan as sphere representation. The PyMol session of PADI2 with the P-TEFb complex with docking and AlphaFold-Multimer shared region is available in Supplementary Material File S5.

Figure 6 (related to Supplementary Figure S5). Positively selected L162 contributes to its interaction with the P-TEFb complex. (A) Cell proliferation of HeLa cells specifically expressing

the WT or mutant PAD12 (single, T159A or W161A or L162A; double, L162A/W161A or triple, L162A/W161A/T159A). Data represent mean ± SEM of at least eight biological experiments. *p value < 0.05; **p value < 0.01. **(B)** Immunoprecipitation with GFP-specific antibody or non-immune rabbit IgG of GFP positive Hela cells nuclear extracts expressing wild-type (WT)or mutant (single, L162A; double, L162A/W161A or triple, L162A/W161A/T159A) PADI2 followed by western blot with the indicated antibodies. The relative quantification is shown as a bar plot. **(C)** Quantitative RT-qPCR validation in HeLa cells selectively expressing the WT or mutant PAD12 (as in B). Changes in mRNA levels were normalized to *GAPDH* mRNA. Data represent the mean ± SEM of $n \geq 3$ biological experiments for all plots in the figure. Two-tailed unpaired Student's t-test was used to determine statistical significance between the groups. (**D**) ChIP-qPCR assay performed in HeLa cells selectively expressing the WT or mutant PAD12 (as in B) with CDK9 antibody. Non-immune IgG was used as a negative control. Y-axis: fold change over the input samples. Data represent mean ± SEM of three biological experiments, *p-value < 0.05; **p value < 0.01.

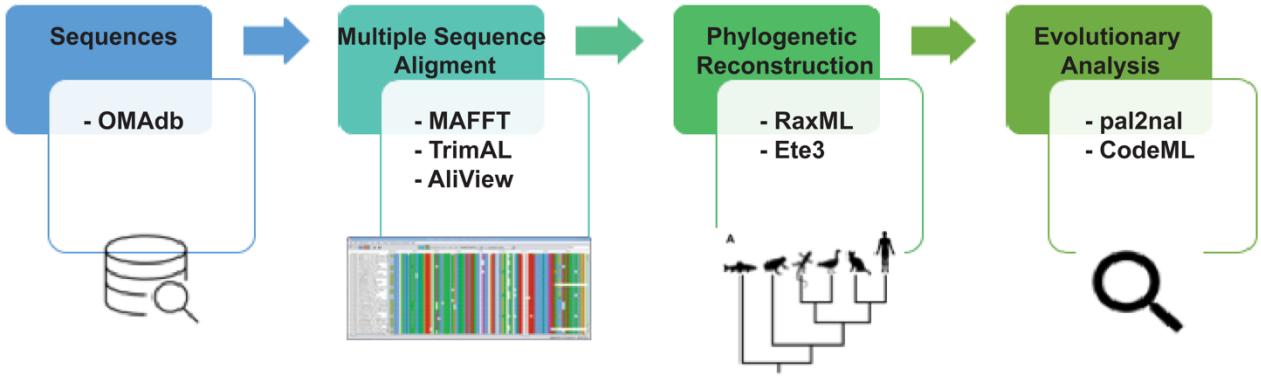
Table 1. Selected sites across the species under the branch site for positive selection (CodeML) and mapped onto PADI2 human structure. The position of the amino acid in the domain, MSA, and PDB file (Supplementary material file S2) are shown in the domain, sequence position, and PDB number, respectively. The flanking residues are shown with the given positively selected amino acid depicted in bold. The type of secondary structure and accessibility to the solvent (i.e., exposure) is shown in the secondary structure and accessibility column. Finally, the MVORFFIP scores (see Materials amd Methods) are also indicated.

Domain	Alignment position	Sequence position (human)	Secondary structure	Flanking residues	MVORFFIP scores (0-1)	Accessibility
PADI_N	70	E16	Beta strand	RV**E**AV	0.8	Exposed
PADI_N	79	L25	Beta strand	TY**L**WT	0.0	Buried
PADI_N	107	V53	Beta strand	EV**V**RD	0.5	Exposed
PADI_N	114	E60	Loop	AE**E**VA	0.7	Exposed
PADI_N	190	K136	Loop	NP**K**KA	0.6	Exposed
PADI_N	216	L162	Loop	PW**L**PK	*0.7*	Exposed
PADI_N	260	V201	Beta strand	EI**V**LY	*0.2*	Exposed
PADI_N	304	S245	Loop	GG**S**AE	0.6	Exposed
PADI_N	308	L249	Beta strand	EL**L**FF	*0.8*	Exposed
PADI_N	322	S263	Loop	GF**S**GL	*0.7*	Exposed
PADI_N	326	S267	Beta strand	LV**S**IH	*0.7*	Exposed
PADI_N	350	T289	Beta strand	TD**T**VI	*0.6*	Exposed
PADI_C	401	L342	Loop	QY**L**NR	*0.7*	Partially buried
PADI_C	439	F380	Loop	KD**F**PV	*0.8*	Exposed
PADI_C	460	S401	Loop	FE**S**VT	*0.8*	Exposed
PADI_C	511	K452	Helix	FL**K**AQ	*0.7*	Exposed
PADI_C	567	D507	Helix capping	QK**D**GH	*0.5*	Exposed
PADI_C	596	S536	Helix	NE**S**LV	*0.7*	Exposed
PADI_C	598	V538	Helix	SL**V**QE	*0.8*	Partially buried

PADI_C	723	M663	Helix capping	WHMVP	*0.7*	Partially buried

Figure 1

A



B

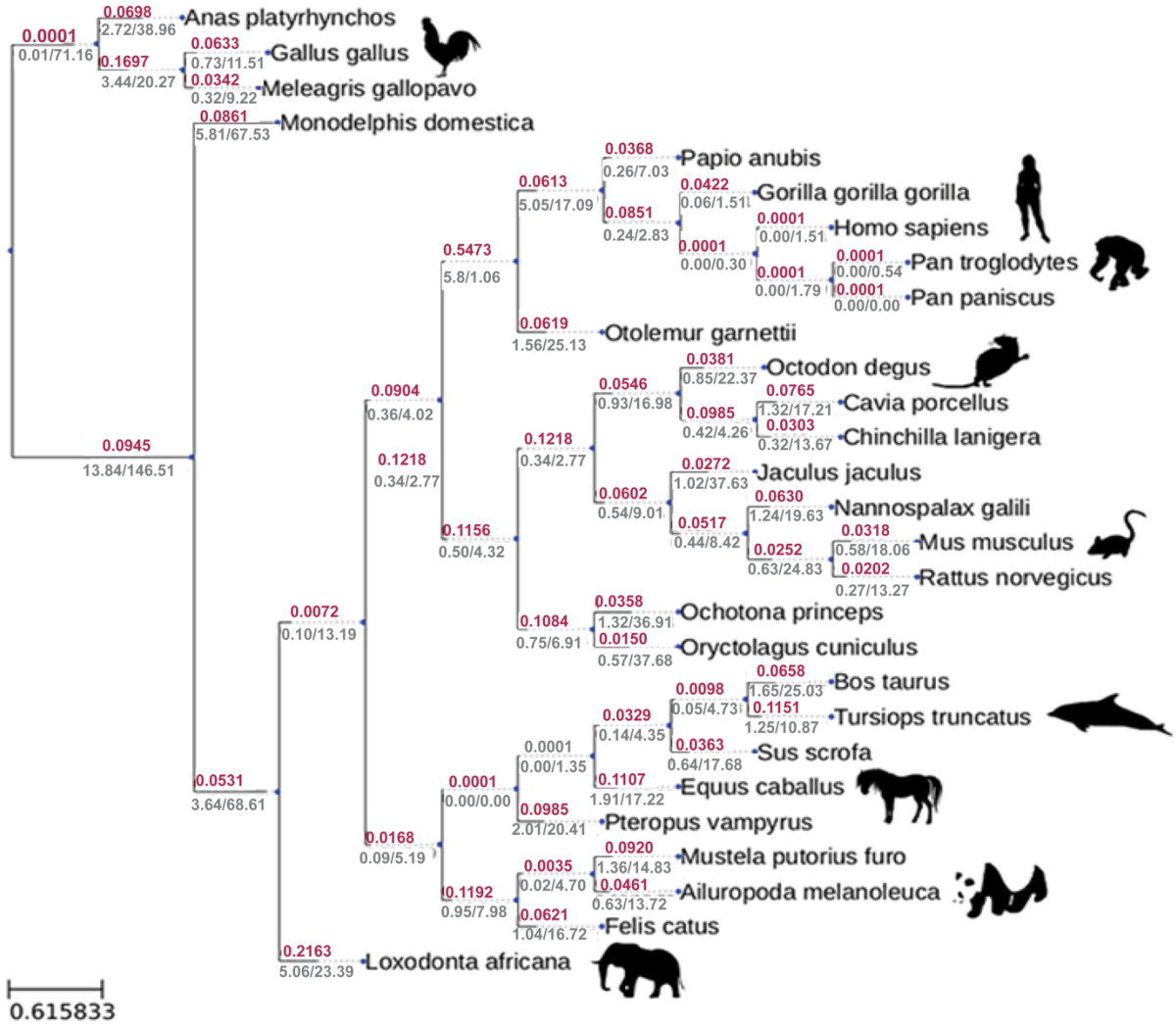


Figure 2

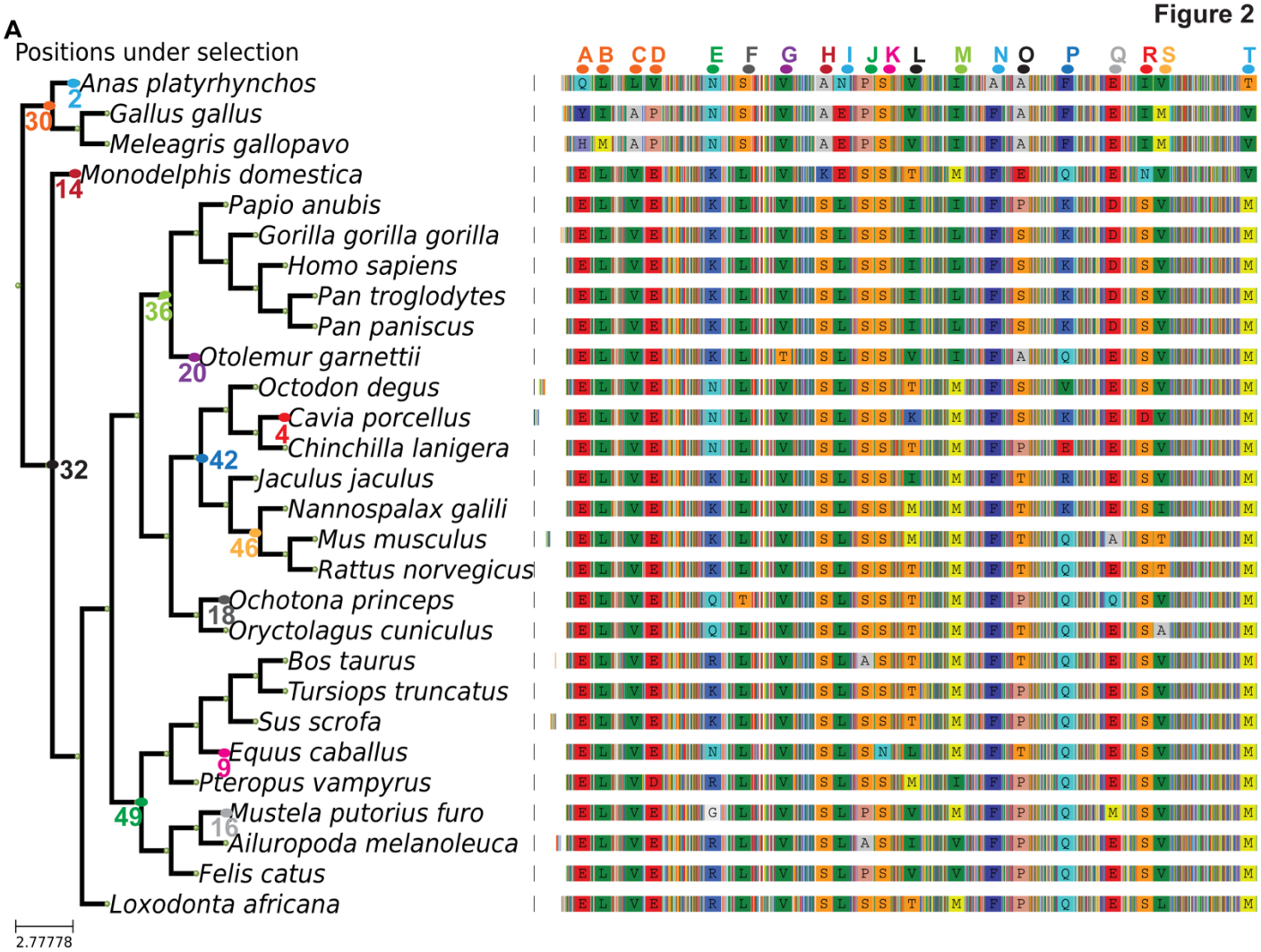
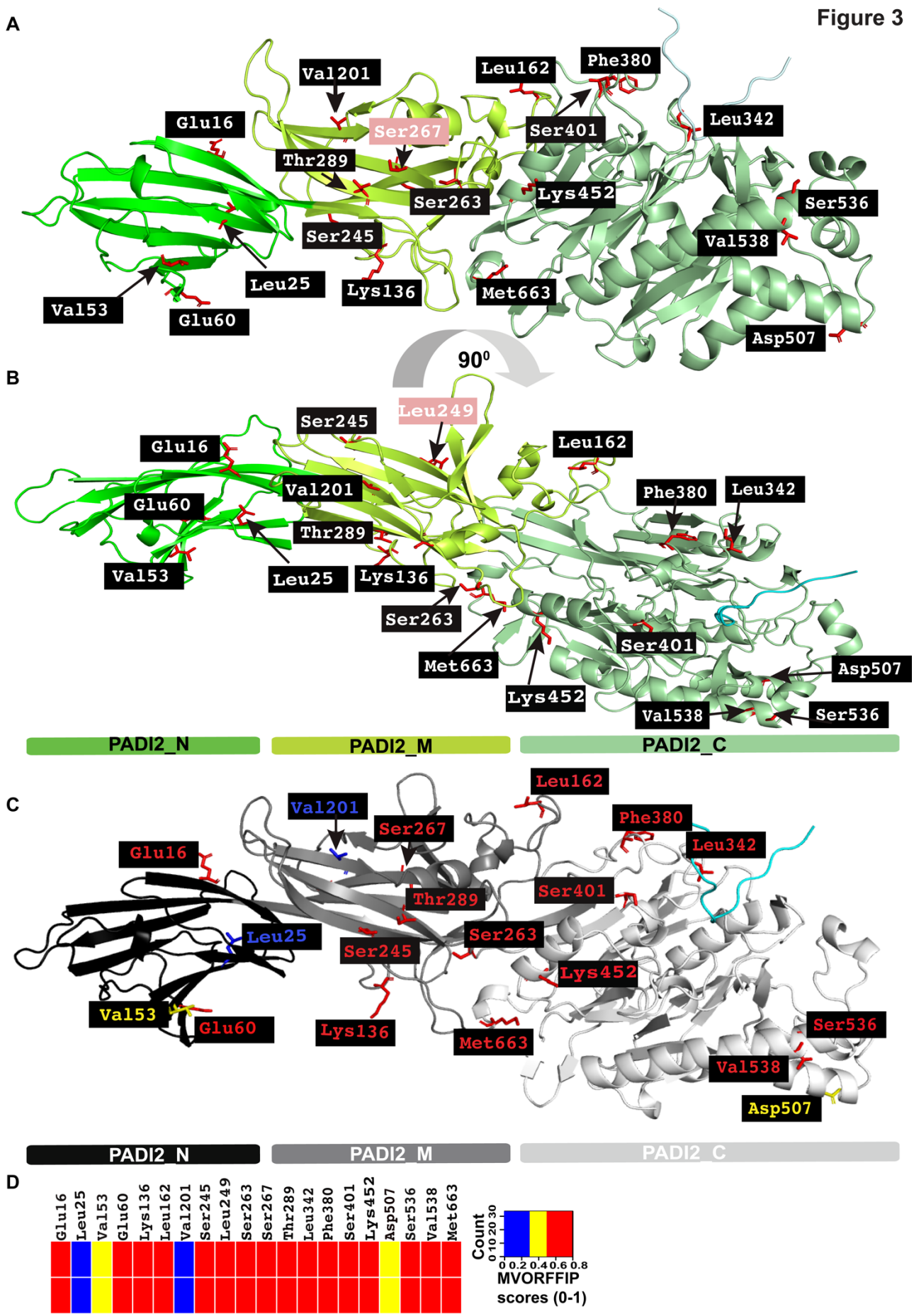
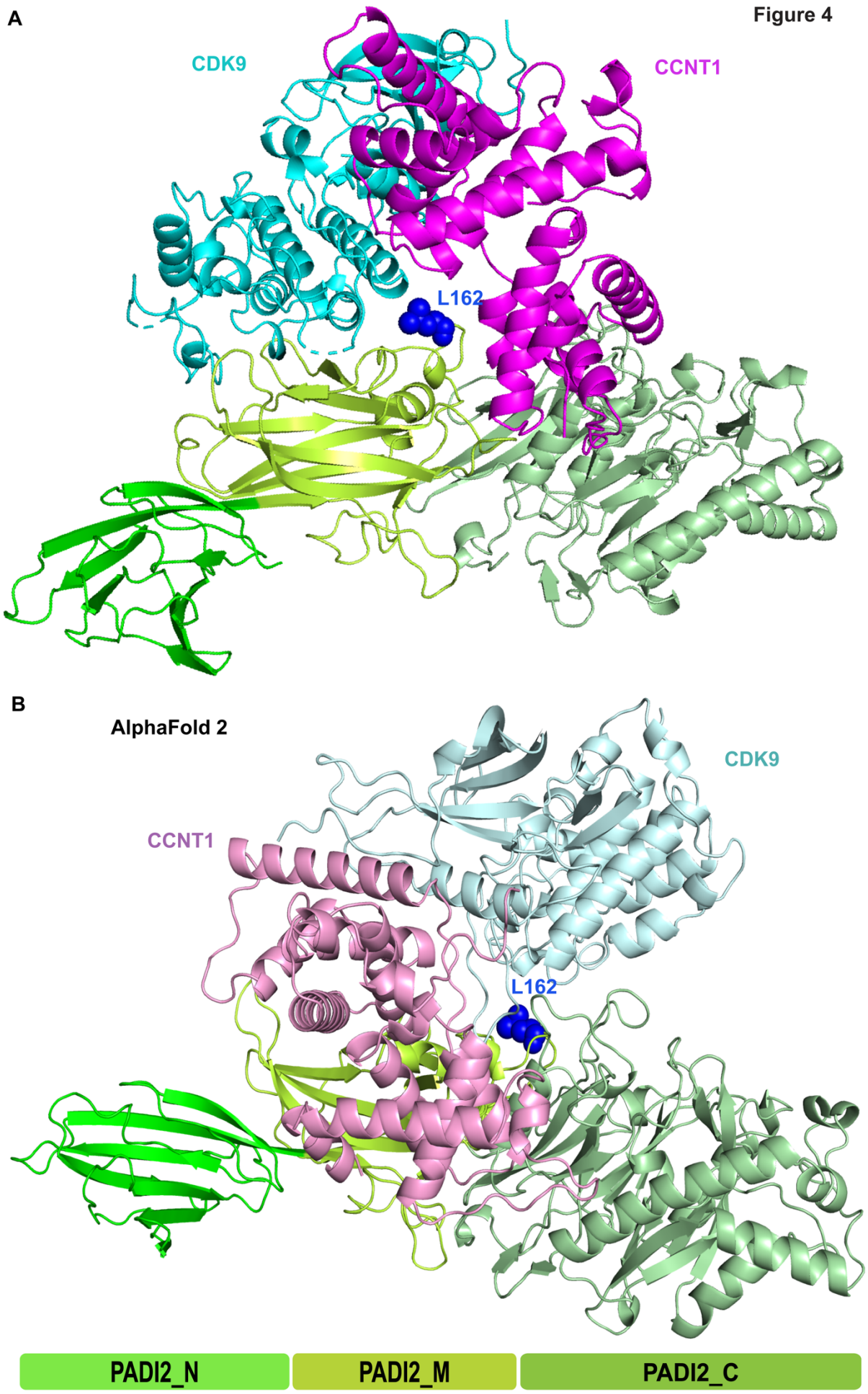

A
Positions under selection
A B C D E F G H I J K L M N O P Q R S T
Anas platyrhynchos
Gallus gallus
Meleagris gallopavo
Monodelphis domestica
Papio anubis
Gorilla gorilla gorilla
Homo sapiens
Pan troglodytes
Pan paniscus
Otolemur garnettii
Octodon degus
Cavia porcellus
Chinchilla lanigera
Jaculus jaculus
Nannospalax galili
Mus musculus
Rattus norvegicus
Ochotona princeps
Oryctolagus cuniculus
Bos taurus
Tursiops truncatus
Sus scrofa
Equus caballus
Pteropus vampyrus
Mustela putorius furo
Ailuropoda melanoleuca
Felis catus
Loxodonta africana
2
30
14
36
20
32
4
42
46
18
9
49
16
2.77778

Figure 3


A
Val201
Leu162
Phe380
Glu16
Ser267
Ser401
Leu342
Thr289
Ser263
Lys452
Ser536
Ser245
Val538
Leu25
Lys136
Met663
Val53
Glu60
Asp507
90⁰
B
Leu249
Ser245
Leu162
Glu16
Val201
Phe380
Leu342
Glu60
Thr289
Val53
Leu25
Lys136
Ser263
Ser401
Met663
Asp507
Lys452
Val538
Ser536
PADI2_N
PADI2_M
PADI2_C
C
Leu162
Val201
Phe380
Ser267
Leu342
Glu16
Ser401
Thr289
Leu25
Ser263
Ser245
Lys452
Val53
Glu60
Lys136
Met663
Ser536
Val538
Asp507
PADI2_N
PADI2_M
PADI2_C
D
Glu16
Leu25
Val53
Glu60
Lys136
Leu162
Val201
Ser245
Leu249
Ser263
Ser267
Thr289
Leu342
Phe380
Ser401
Lys452
Asp507
Ser536
Val538
Met663
Count
0 10 20 30
0 0.2 0.4 0.6 0.8
MVORFFIP
scores (0-1)


A
Figure 4
CDK9
CCNT1
L162
B
AlphaFold 2
CDK9
CCNT1
L162
PADI2_N
PADI2_M
PADI2_C

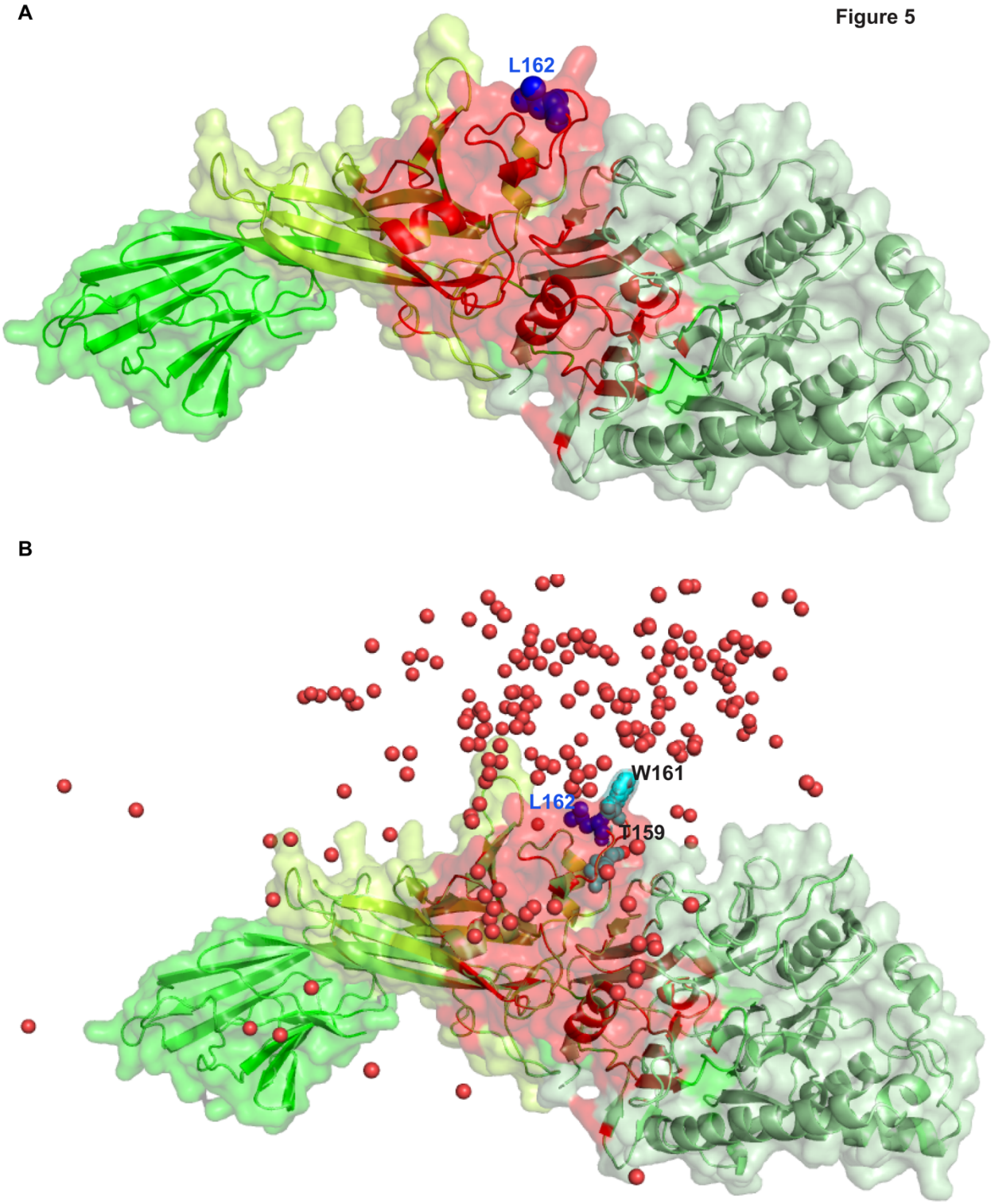
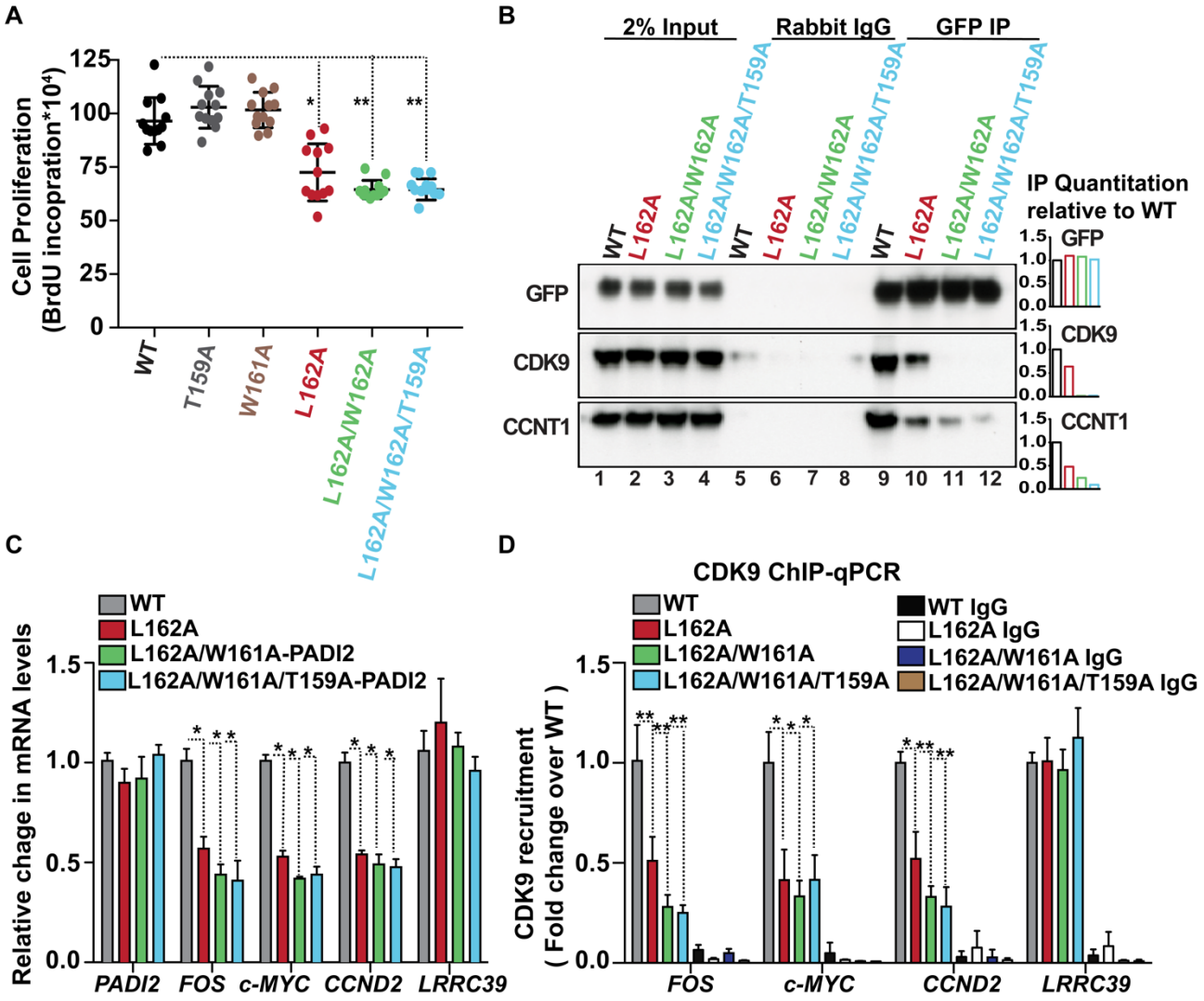

A
Figure 5
L162
B
W161
L162
T159

Figure 6


A
Cell Proliferation (BrdU incopration*10⁴)
125
100
75
50
25
0
WT
T159A
W161A
L162A
L162A/W162A
L162A/W162A/T159A
B
2% Input
Rabbit IgG
GFP IP
WT
L162A
L162A/W162A
L162A/W162A/T159A
IP Quantitation relative to WT
GFP
CDK9
CCNT1
1 2 3 4 5 6 7 8 9 10 11 12
C
Relative chage in mRNA levels
WT
L162A
L162A/W161A-PADI2
L162A/W161A/T159A-PADI2
PADI2
FOS
c-MYC
CCND2
LRRC39
D
CDK9 ChIP-qPCR
CDK9 recruitment (Fold change over WT)
WT
L162A
L162A/W161A
L162A/W161A/T159A
WT IgG
L162A IgG
L162A/W161A IgG
L162A/W161A/T159A IgG
FOS
c-MYC
CCND2
LRRC39