Evolutionary analysis reveals the role of a non-catalytic domain of peptidyl arginine deiminase 2 in transcriptional regulation

José Luis Villanueva-Cañas‡[1,12], Narcis Fernandez-Fuentes‡[2], Dominik Saul[3,4,] Robyn Laura Kosinsky[5], Catherine Teyssier[6], Malgorzata Ewa Rogalska[1], Ferran Pegenaute Pérez[7,8], Baldomero Oliva[8,9], Cedric Notredame[1,8], Miguel Beato[1,8], Priyanka Sharma*[10,11]

[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]Division of Endocrinology, Mayo Clinic, Rochester, MN 55905, USA
Robert and Arlene Kogod Center on Aging, Mayo Clinic, Rochester, MN 55905, USA

[4] Department of Trauma and Reconstructive Surgery, BG Clinic, University of Tübingen, Tübingen, Germany

[5]Robert Bosch Center for Tumor Diseases, Stuttgart, Germany.

[6]Institut de Recherche en Cancérologie de Montpellier (IRCM), INSERM U1194, Université de Montpellier, Institut du Cancer de Montpellier (ICM), F-34298 Montpellier, France

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

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

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

[10]Institut de Pharmacologie et de Biologie Structurale (IPBS), Université de Toulouse, Centre national de la recherche scientifique (CNRS)

[11]Université Toulouse III - Paul Sabatier (UPS), Toulouse, France.

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

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

‡ These authors contributed equally

Abstract

Peptidyl arginine deiminases (PADIs) catalyze protein citrullination, a post-translational conversion of arginine to citrulline. The most widely expressed member of this family, PADI2, regulates cellular processes that impact several diseases. We hypothesized that we could gain new insights into PADI2 function through a systematic evolutionary and structural analysis. Here, we identify 20 positively selected PADI2 residues, 16 of which are structurally exposed and maintain PADI2 interactions with cognate proteins. Many of these selected residues reside in non-catalytic regions of PADI2. We validate the importance of a prominent loop in the middle domain that encompasses PADI2 L162, a residue under positive selection. This site is essential for interaction with the transcription elongation factor (P-TEFb) and mediates active transcription of the oncogenes *c-MYC*, and *CCNB1,* as well as impacting cellular proliferation. These insights could be key to understanding and addressing the role of the PADI2 *c-MYC* axis in cancer progression.

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

Significance Statement

Here we use a systematic evolutionary analysis to identify positively selected residues in the non-catalytic domain of PADI2 and link the positive selection of key residues to a role in transcription. Specifically, a structurally exposed loop in the PADI2 middle domain encompasses the positively selected residue L162 which is linked to transcription and cellular proliferation. This loop contributes to PADI2 interaction with the P-TEFb complex and cellular proliferation. Our results showcase the utility of combining evolutionary and experimental approaches to dissect the evolution of essential functional processes.

Introduction

Members of the peptidyl arginine deiminase (PADI) family catalyze post-translational calcium-dependent citrullination of arginines, converting them into the non-coded amino acid citrulline [1–4]. Arginine citrullination is a widespread post-translational modification (PTM) that increases the hydrophobicity of proteins and can contribute to fine-tuning physiological processes. For instance, citrullination of core histones weakens histone–nucleic acid interactions, impacting both chromatin organization and transcription [5]. Citrullination can also affect protein folding and consequently protein function [6–9]. The five members of the PADI family have distinct tissue-specific expression patterns: 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 ovarian egg cells, embryonic tissues, 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, including autoimmune disorders [10] and neurological diseases [11,12], as well as with several cancers (e.g., breast, cervical, liver, lung, ovarian, thyroid, gastric, and prostate cancer) [13–22]. The link between PADI2 and 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 [23]. All family members are Ca^{2+}-dependent, but PADI2 also relies on ordered calcium binding to its active site for substrate binding and catalysis, as demonstrated in the deimination of arginine 26 in histone H3 (H3R26) [24–27]. Importantly, H3R26 citrullination is critical for interaction with the chromatin remodeller SMARCAD1 (SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A containing DEAD/H box 1) and hence links PADI2 to gene activation in the context of naive pluripotency [28]. PADI2 also contributes to chromatin modification that promotes the differentiation of oligodendrocyte precursors and efficient myelination, processes required for motor and cognitive functions [29]. Recently, PADI2 was found to citrullinate MEK1, thereby promoting signaling by extracellular signal-regulated protein kinase 1/2 (ERK1/2) signaling in endometrium cancer [22].

Recently, we identified another unique function of PADI2: the deimination of arginine 1810 (R1810) on repeat 31 of the C-terminal domain (CTD) of the large subunit of RNA polymerase 2 (RNAPII). This modification potentiates the interaction between the RNAPII-CTD and the positive transcription elongation factor b complex (P-TEFb) and consequently facilitates gene expression that is essential for cell identity [30]. Indeed, we reported that citrullination of R1810 in the CTD of

RNAPII facilitates its interaction with the P-TEFb kinase complex and promotes the recruitment of CDK9 to transcription start sites (TSS) [30]. This contributes to overcoming the RNAPII pausing barrier, highlighting the functional connection between PADI2 and the P-TEFb kinase complex. These studies support the notion that PADI2 can selectively 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, worms, and flies. A recent study focusing on the comprehensive identification of *PADI* homologs unveiled the evolutionary trajectory of *PADIs* within the animal lineage [31]. *PADIs* appear to have been introduced from cyanobacteria into animals by horizontal gene transfer (HGT) [31], supporting the previous hypothesis of HGT as a mechanism for introducing new genetic material into vertebrate genomes [32–35]. Enzymes mediating citrullination 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 lamblia* ADI, an extended form of the free L-arginine deiminase gADI found in this human parasite [36,37]. Previous small-scale phylogenetic analyses of the PADI family have 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) [38–41]. Of note, PADI1 and PADI3, and PADI4 and PADI6, are more closely related in evolutionary terms than PADI2 and any other family member [6]. This finding suggests that PADI2 is most closely related to the ancestral PADI enzyme and that the other PADI enzymes arose through gene duplication more recently in evolution.

The enzymes responsible for essential PTMs (including phosphorylation, acetylation, and glycosylation) are found across all domains of life, suggesting that they were present in the Last Universal Common Ancestor (LUCA) [42,43]. Similarly, the PADI_C domain was present in the LUCA, indicating an ancient origin of citrullination. The PADI_M domain, which encompasses the sequential calcium-binding sites and maintains allosteric communication with PADI_C [25], is present in cyanobacteria. Indeed, recent work supports the degeneration of the PADI_N domain in cyanobacteria and demonstrates the existence of catalytically active PADI proteins in cyanobacteria [31]. The relatively recent appearance of the NTD during the PADI family evolution suggests it has a functional relevance that is unique to higher organisms.

Previously, we found that PADI2 interacts with P-TEFb to maintain the expression of actively transcribed genes [30,44]. The P-TEFb complex comprises cyclin-dependent kinase 9 (CDK9) and its regulatory partner cyclin T1 (CCNT1) [45]. Activation of the P-TEFb complex is required early in transcription to overcome RNAPII pausing and promote the productive phase of transcription elongation [45–48]. Lack of recruitment of the P-TEFb complex and associated components has been linked to several disease conditions [49–51], highlighting the importance of this process in healthy physiology. While PADI2 could be a therapeutic target that specifically prevents p-TEFb-mediated expression of genes that promote disease states, more needs to be understood.

Here, we analyze 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 belong to the non-catalytic NTD or to the middle domain of PADI2. These positions suggest that they can participate 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 impact on PADI2 interactions with the P-TEFb complex. We found that leucine 162 (L162), in the exposed loop of the middle domain of PADI2, underwent positive selection during the initial divergence of primates. Indeed, this residue contributes to interactions between PADI2 and the P-TEFb complex. These results establish a link between the evolutionary selection of key residues in PADI2 and their role in regulating the protein-protein interactions required for functional transcriptional regulation.

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 human PADI2 in the OMA (Orthologous MAtrix) database [52] and collected all the available sequences in mammals (see Materials and Methods). After close inspection, we discarded partial sequences and low-quality genomes, leaving 185 PADI orthologs that represent 25 mammalian species and three bird species (**S1 Fig**). We used these sequences to build a multiple sequence alignment (MSA, **S1 File**) and reconstructed a phylogenetic tree using the software RaxML, rooting the tree using the bird sequences as an outgroup (**S2 Fig,** see Materials and Methods).

The tree we built is consistent with five PADI genes that resulted from duplication of a common ancestral sequence. 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 known species relationships being broadly recapitulated within each of these clusters (**S2 Fig)**. Robust bootstrap supports (82-100, **S3 Fig**) for this tree strongly support the hypothesis that the genes within each cluster are orthologous. We observed clustering between *PADI1* and *PADI3* and between *PADI4* and *PADI6*, with *PADI2* being the most distant ortholog among family members. Notably, PADI2 also has shorter branch lengths in comparison to other family members, indicating a higher level of conservation among family members in this cluster. PADI6 family members have the longest branch lengths, indicating this cluster is the most divergent and suggesting these genes arose from an old duplication event. 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.

Two bird genes (i.e., ANAPL06996 and ANAPL06995 in duck) in the unpainted cluster (**S2 Fig**) did not exhibit clear orthology to specific *PADI* genes in either Ensembl or the OMA, though some were annotated as *PADI1* orthologs. Our data suggest that there was a bird-specific duplication, possibly from *PADI1,* which is the closest gene in the synteny analysis (**S4A Fig**). Our species comparison also revealed an inversion of the entire *PADI* locus that appears to have occurred in the common ancestor of humans and gibbons (Hominoidea; **S4B Fig**). Most mammalian species have the same gene order as that found in mice (**S4C Fig**). Overall, our analyses revealed key evolutionary changes at the *PADI* locus, including gene re-arrangements and duplications within mammals.

PADI2 is highly conserved across species

Since PADI2 is likely to be the ancestral copy, we decided to focus on the most recent changes occurring in its sequence. By focusing on a specific phylogenetic selection of species, we aimed to identify changes that have occurred since the common amniote ancestor that occur with good resolution in different mammals. Our dataset contains *PADI2* representatives in all three major extant groups in class Mammalia and the principal orders or infraclasses: primates, rodents, carnivores, cetartiodactyla, lagomorphs, proboscidea, and marsupials. We built a multiple sequence alignment with all 28 *PADI2* sequences (see Materials and Methods**, S1 File)** 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 [53]. 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 [53].

We calculated the dN/dS ratio for all internal branches and leaves in the tree using CodeML (**Fig 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 supports the idea that PADI2 is a highly conserved protein, suggesting that it performs critical functions in the organism (**Fig 1**). However, we observed a relatively high dN/dS in the common ancestor of all primates, possibly indicating relaxation of selection or the fixation of a few beneficial mutations due to positive selection at this point in evolution. The dN/dS ratio requires a high number of changes to detect a selection effect and our data indicate 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. However, this analysis suggests that intriguing selection processes operated during PADI2 evolution.

Detection of positively selected residues in PADI2

We used an orthogonal approach, the branch site positive selection test, to surpass the limitations of dN/dS analysis and identify specific residues that have undergone selection [54,55]. 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. 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 [54]. 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 positive selection test on 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 the alignment of a 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 Fig 2**.

Positively selected residues in PADI2 may mediate protein-protein interactions

We next integrated our evolutionary insights into PADI2 with known structural information to study the properties of positively selected residues. To pursue this analysis, we mapped all the positively selected amino acids onto the structure of PADI2. 12 of the 20 positively selected amino acids mapped onto the PADI2_N and PADI2_M domains, while eight mapped to the PADI2_C domain. The latter contains the catalytic pocket and is responsible for the citrullination of arginine residues. (**Table 1, Fig 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, File S2, 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, approximately 50% of the amino acids are located in loops, while the others are either in beta strands or alpha 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 binding partners. These amino acids are fully exposed, and their chemical properties correspond with classical protein interfaces [56,57]. We thus characterized these positively selected amino acids using Multi-VORFFIP (MVORFFIP), a tool that predicts protein-, peptide-, DNA-, and RNA-binding sites [56]. Notably, the MVORFFIP predictions for all except V201 and L25 yielded very high scores (> 0.7, on a scale of 0 to 1), supporting the hypothesis that they are interface residues (**Fig 3C-D, Table 1**). S263 is 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) may be to prevent distortion of a nearby beta-sandwich. Indeed, a larger hydrophobic amino acid at this position would change the packing of the beta-strands.

We found that three positively selected residues in the PADI2_C domain, L342, F380, and S401, are close to the catalytic pocket. These residues are likely to play a role in substrate specificity (L342) and in the dynamics of catalysis (F380 and S401). K452, S536, and V538, however, are located in a helix that is distant from the active site pocket. Judging from the structural microenvironment, these residues could also play a role in protein-protein interactions, in line with the high scores assigned by MVORFFIP (**Fig 3D, Table 1)** [56]. Finally, D507 and M663 are both located in helix-capping positions. Conservation in helix capping is important for the stability and

integrity of the helix. Residue M663 may also play a role in the packing of the C-terminal tail. A CAPS Database search [58] showed that other helices with the same capping structure present a conserved small hydrophobic patch in the same position by including M, V, or L residues. It is noteworthy that M663 is indeed strictly conserved across all human PADIs. These observations suggest that most of these positively selected amino acids may function in facilitating and stabilizing protein-protein interactions.

The middle domain of PADI2 maintains interactions with the P-TEFb complex

We next tested whether residues that have been positively selected in evolution but are not linked to the catalytic domain of PADI2, might interact with other proteins. For this analysis, we examined the P-TEFb kinase complex, which we previously found to interact with PADI2 [30]. The structures of PADI2 [25] and CDK9-CCNT1 [59] are known individually, and therefore it was possible to derive a structural model of the PADI2–CDK9/CCNT1 complex by protein docking. After deriving the docking ensemble of PADI2-CDK9/CCNT, we ranked the models. Importantly, we used a swarm-based approach (among other metrics) to select the putative model of interaction, reasoning that determining the region of PADI2 would be relevant to its interaction with CDK9/CCNT1 (**S4D Fig, File S3, Movie S2**). Moreover, we derived the structural model using AlphaFold-Multimer [60,61]. As shown in **S4E Fig (File S4, Movie S3)**, interacting regions in the structural model from AlphaFold-Multimer overlap substantially with those identified by docking (**S4D Fig**). More specifically, the predicted interface between PADI2 and CDK9-CCNT1 contained PADI2_M and the hinge region was overrepresented with the number of positively selected residues. Importantly, L162 is present on the highly exposed loop (**Fig 4A**).

While AlphaFold-Multimer generates a single model, docking is represented by a range of docking poses. This range of models allowed us to identify regions of the predicted interface, i.e., PADI2_M and PADI2_C (**Fig 4A**), that are over-represented in the 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 over-represented **(Fig 4B, File S5, 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. We therefore selected these residues 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 a positively selected residue, suggesting that it is important for maintaining this interaction.

The loop encompassing positively selected Leu162 is important for cell proliferation and maintaining PADI2 interactions with the P-TEFb complex

Since the P-TEFb kinase complex is involved in transcription elongation and cell proliferation [62–64], we experimentally tested whether the L162-encompassing loop in the middle domain of PADI2 affects cell proliferation. In addition to L162, we tested its neighboring amino acids, W161 and T159, which also ranked highly in our evolutionary and interaction prediction analysis 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 (**S5A-B Fig**). We monitored 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 (**Fig. 5A, S5C Fig).** 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 PADI2.

Therefore, we focused on the L162A, L162A/W161A, and L162A/W161A/T159A mutants along with WT PADI2 for further analysis. 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 (**Fig 5B).** These results confirmed the functional role of the positively selected L162 in maintaining PADI2 interactions with the 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 validated the reduced expression levels of the *c-MYC*, *CENPQ*, and *CCNB1* genes in cells expressing the single, double, or triple PADI2 mutant (**Fig 5C**). Of note, the levels of PADI2 expression did not differ significantly. 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 *c-MYC*, and *CCNB1, CENPQ* without affecting the levels of a control gene (the low-expressed *LRRC39* gene). Next, we examined if the L162A mutant in PADI2 can mediate CDK9 recruitment to the promoter region of the target genes. We performed a chromatin immunoprecipitation (ChIP) assay using a CDK9-specific antibody in HeLa cells specifically expressing either WT PADI2, the L162A mutant, or the double and triple mutants. We found that CDK9 occupancy decreases in the presence of mutants compared to the WT PADI2 control (**Fig 5D)**, suggesting that the positively selected L162 residue, along with W161A/T159A, is important to maintaining PADI2 interaction with the P-TEFb complex.

Given the fact that the L162-encompassing loop is important to maintaining PADI2 interaction with the P-TEFb complex, we next explored the effects of the PADI2 triple-mutant (L162A/W161A/T159A) on the transcriptome using single-cell RNA sequencing experiments. We compared the single-cell transcriptomics atlas of PADI2 triple-mutant to that of WT PADI2-expressing HeLa cells (**S6A-B Fig).** After quality control, 49,503 cells were subjected to downstream analysis, of which, 17,139 and 32,364 cells were derived from the WT and triple mutant PADI2 samples, respectively. We did not observe any significant differences in cell populations through unsupervised clustering and dimensionality reduction. TSNE plots organized the data into three clusters (**S6A Fig).** However, these clusters are not significantly different in triple mutants in comparison to WT (**S6B Fig).** Among the differentially expressed genes, we found *c-MYC* (**Fig 6A-B)**. Within the top genes that are differentially expressed, we also found *CCNB1* and *CENPQ,* which are related to cell cycle regulation (**Fig 6A-B)**. Remarkably, in global differential expression (DEseq) analysis, the PADI2 triple mutant (L162A/W161A/T159A) affected the expression of over 469 genes (Log2FC>0.58 or Log2FC<-0.58, p-value <0.05). Of these, 162 genes were downregulated and 307 upregulated (**Table S2).** Gene ontology analysis of the top down-regulated genes (Log2FC<-1.0, p-value <0.01) revealed enrichment in the cell cycle, RNA binding, and chromosome organization. The up-regulated genes were instead enriched for genes linked to the inflammatory response, cell-cell signaling, and cell adhesion processes (Log2FC>1.0, p-value <0.01) (**Fig 6C and Table S3)**. We also investigated the regulatory gene sets using the gene ontology tool. The promoter regions of downregulated genes were significantly enriched for GTF2A2 (UniProt: P52657), GTF2E2 (UniProt: P29084), FOXE1 (UniProt: O00358), MZF1(UniProt: P28698), ZFHX3_(UniProt: Q15911) binding motifs (**Fig 6D and Table S4**). Meanwhile, the upregulated genes showed significant enrichment in the promoter regions for motifs that can bind ZNF92 (UniProt: Q03936), METTL14 (UniProt: Q9HCE5), GFI1 and CTTTGA motif (**Fig 6E and Table S4**). This analysis highlights a connection between regulatory elements and differentially expressed (WT vs PADI2 triple mutant) genes.

.

Importantly, *c-MYC, CCNB1,* and *CENPQ* overexpression have been linked to oncogenesis [65–70]. These results support the functional relevance of the positively selected L162-encompassing loop in the middle domain of PADI2 in the expression of important oncogenic genes. Modulation of *c-MYC* expression by epigenetic regulation in cancer cells has been suggested to increase heterogeneity in the transcriptional landscape that promotes tumorigenesis [71]. These observations suggest a potential function for the L162-encompassing loop of PADI2 in the regulation of the *c-MYC /CCNB1/CENPQ* axis which has been linked to tumorigenesis.

Discussion

We characterized the evolution of the PADI gene family using the most complete dataset of mammalian orthologous sequences to date. The three major extant groups in class Mammalia 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 [72], there are 5416 mammalian species divided into 29 orders. Most known species belong to the 7 biggest orders (4865 species), all of which have at least one representative in our phylogenetic analysis.

The functional relevance of PADI2 is highlighted by its high sequence conservation across species and 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 sequences 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 **(Figs 1-2, S1-S4 Figs)**. Since these selected residues could have a strong functional impact on PADI2, we studied them further using the human sequence.

By examining the location of the positively selected residues in the structure of human PADI2, modeled in a complex with R1810 in the CTD of RNAPII, we determined that the majority of positively selected PADI2 residues were structurally highly exposed in the non-catalytic domains of PADI2 (**Fig 3, Table 1**). Notably, the highly exposed position and chemical properties of these positively selected residues suggested they play important functions in cellular processes. In addition, our MVORFFIP analysis revealed that these interface residues may play a significant role in protein-protein interactions (**Fig 3D**). Therefore, it could be postulated that the appearance of these positively selected residues contributed to the evolution of PADI2 function by modulating its interaction with an essential cognate set of 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 of RNAPII-R1810 on transcription [30]. Hence, we used structure modeling approaches to derive the tertiary structure of the PADI2–CCNT1/CDK9 complex. To identify the PADI2 residues that contribute to forming the PADI2–CCNT1/CDK9 complex, we analyzed structural models and identified a highly exposed loop in the PADI2_M domain as a potential interacting region. Strikingly, this loop includes the positively selected residue L162, along with T159 and W161. This observation supports the notion that positively evolved residues tend to coincide with the important residues on the protein surface, and that positively selected structural clusters are important for cellular function (**S6C Fig**).

Our data in HeLa cells overexpressing the GFP-tagged WT and a mutant PADI2 also unveiled an essential function for the L162-containing loop in interaction with the P-TEFb kinase complex and cell proliferation. Notably, mutation of 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 one limitation of our study was that we could not mutate the endogenous PADI2 gene due to the hyperpolyploid nature of the HeLa genome. Therefore, we based our analysis on overexpression of GFP-tagged wild-type and mutant versions of PADI2 in HeLa cells.

We focused on the interactions between PADI2 and the P-TEFb complex, specifically with the positively selected L162-encompassing loop. However, in the future, it would be interesting to investigate the functional role of other positively selected PADI2 amino acids in cellular processes. Nonetheless, our analysis shows the importance of using a multi-disciplinary approach. Here, we used comparative genomics, evolutionary analysis, structural modeling, and genetic perturbations to analyze the functional relevance of a recently evolved non-catalytic domain in a highly conserved enzyme family. Overall, this evolutionary approach led to the identification of a PADI2 structural loop that is under positive selection and specifically supports an important role for PADI2 in the modulation of transcription elongation.

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 [73]. The selected plasmid sequence was verified by 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**GCA**CCC***GCC*GCA**CCCAAGGAGGACTGCCGTGATG -3′

R3 5′- CATCACGGCAGTCCTCCTTGGG**TGC*GGC***GGG**TGC**CTCTCGGTCACAGTTCACCA-3′.

The corresponding fragments were generated by each pair of primers using oligonucleotide assembly and introduced by Gibson assembly [73]. 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 the plasmid being tested 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 and 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 using Flow Jo 10. 6. 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 Lipofectamine 3000 (Invitrogen) in 96-well plates. 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 microscopy images using the Incucyte software algorithm to generate a proliferation index corresponding to the change in confluence for each well. These measurements are the 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 the 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 [74]. 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**.

Single-cell samples, library preparation, and sequencing

HeLa cells (0.6×10^6) were transfected with either wild-type PADI2 (WT) or the PADI2 triple mutant (with L642A/W161A/T159A) using Lipofectamine 3000 (Invitrogen) in 60mm plates. GFP-positive singlet cells and live cells were suspended in 1ml of PBS+BSA 0.05%. Cells were centrifuged at 400 rcf for 5 min at 4°C in order to bring the cell concentration to 300-1000 cells/µl. Cell concentration and viability were determined by manual counting using a Neubauer chamber and staining the cells with Trypan blue. Cells were partitioned into Gel Bead-In-Emulsions (GEMs) using the Chromium Controller system (10X Genomics), with a target recovery of 5000 total cells per sample. cDNA sequencing libraries were prepared using the Next GEM Single Cell 3' Reagent Kits v3.1 (10X Genomics, PN-1000268), following the manufacturer's instructions. Briefly, after GEM-RT clean-up, cDNA was amplified for 11 cycles, and cDNA QC and quantification were performed on an Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies). cDNA libraries were indexed by PCR using the PN-1000215 Dual Index Kit TT Set A Plate. Size distribution and concentration of 3' cDNA libraries were verified on an Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies). Finally, sequencing of cDNA libraries was carried out using the Illumina NovaSeq 6000.

Single-cell RNA sequencing data processing and analysis

Transcriptome-wide analysis of human L162A/W161A/T159A-PADI2 mutant and WT cells was performed at single-cell resolution using NovaSeq 6000 pipeline. Sequencing data were aligned to the human reference genome Grch38. Data with at least 500 unique molecular identifiers (UMIs), log10 genes per UMI >0.8, >250 genes per cell and a mitochondrial ratio of less than 20% were extracted, normalized, and integrated using the Seurat package v4.0.3 in R4.0.2. After quality control and integration, we performed a modularity-optimized Louvain clustering. After quality control, 17,139 WT and 32,364 PADI2mut cells remained. Altogether, we used two biological replicates (PADI2_mut1: 16,177 cells, PADI2mut2: 16,187 cells, WT1: 8,568 cells, WT2: 8,571 cells).

Gene Ontology (GO) and Regulatory Gene Set analysis

GO Annotation and regulatory gene set analysis were performed using the online tool Gene Set Enrichment Analysis (GSEA, http://software.broadinstitute.org/gsea/index.jsp) collection database v5 [75,76]. The significant cut-off p-value and FDR q-value < 0.05. Plots were done with the use of Prism (GraphPad Prism 10.0.3 for MacOS), Cytoscape (version 3.9.1) and R4.0.2.

ChIP-qPCRs: For ChIP assays [30], 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 an anti-CDK9 antibody (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. This step was followed by incubation with 50µl of Dynabeads® M-280 sheep anti-mouse IgG (11201D, Thermo Scientific) for 3 hours. Beads were washed 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 with 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 at 4°C. Subsequently, crosslinks were reversed at 65º C overnight, followed by RNase treatment for 1.5 hours, and bound DNA was purified by Phenol-Chloroform extraction. The resulting 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 protease 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 the 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 [52]. 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 bird species (chicken, turkey, and duck) were used as outgroups. The protein sequences were aligned using the software PRANK [77] and a pruned mammalian guide tree with branch distances [78]. 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 [79]. Local realignments of two regions were done using MAFFT [80] within AliView [81]. Another MSA using only one-to-one orthologous PADI2 sequences was also built. TrimAI [82] 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 [83]. The MSAs generated and the trees used are available as Supplementary Material S1.

Phylogenetic reconstruction

We reconstructed the PADI family tree using RaxML [84] 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 [85]. For each branch and leaf in the tree, we performed a branch-site test of positive selection [54], implemented in the Phylogenetic Analysis by Maximum Likelihood (PAML) software package [85] and available in the environment for tree exploration *(*ETE3*)* framework as bsA/bsA1 [86]. 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 [56]. 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 interface 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, PADI2-CDK9, was derived using protein docking as follows. The crystal structure of PADI2 (PDB code: 4n2a) [25] and CDK9/CCNT1 (PDB code: 3blr) [59] were available. The protein docking was performed using VD^{2}OCK [87], generating the ensemble of over 10000 docking poses. Docking conformers were clustered using the GROMACS [88] and sorted by the ES3DC potential [89]. The top 200 poses were selected to identify the top 10 putative interface residues: for this, a composite score was calculated using the docking information, the MVORFFIP score and sequence conservation. Among the top 200 docking poses, as discussed 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 [90].

AlphaFold-Multimer modeling of PADI2-CDK9/CCNT1 complex

Structural models of the trimer PADI2-CDK9/CCNT1 were generated using AlphaFold-Multimer [60] as follows. The sequences were retrieved as mentioned above, PADI2 (PDB code: 4n2a) [25] and CDK9/CCNT1 (PDB code: 3blr) [59]. The structure was derived using model parameters corresponding to "model_1_multimer".

Data Availability

The OMA is an open-source database for the inference of orthologs among complete genomes https://omabrowser.org/oma/home/. ETE3 was used for the reconstruction, analysis, and visualization of phylogenetic trees and is 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. File S2.pse, File S3.pse, File S4.pse, and 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 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. Single-cell RNA sequencing data performed and used in this study have been deposited at GEO under the accession code GSE246420.

Acknowledgment

We thank Fátima Gebauer and François Le Dily from CRG for their constructive criticism and advice on this manuscript. We acknowledge Veronica A. Raker for manuscript editing. We thank the CNAG (National Center for Genomic Analysis, Barcelona) Genomic facility and CRG flow Cytometry unit for all technical support.

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 French National Research Agency (ANR) Young Investigator grant (ANR-21-CE12-0010), Cancer Research Foundation (ARCPJA22021050003683) National Institute of Health and Medical Research (INSERM) Young Recruitment Support (U1194SHA), Cancéropôle Grand Sud-Ouest collaboration grant (R21031FF), Centre national de la recherche scientifique (CNRS), Université Toulouse III - Paul Sabatier (UPS), Toulouse, France. This work was also supported by a grant to M.B. from the European Research Council Synergy Grant "4DGenome" (609989).

Declaration of Interests

The authors declare no competing financial interests.

Supplementary Data

Supplementary data are available online.

References

1. van Venrooij WJ, Pruijn GJM. Citrullination: A small change for a protein with great consequences for rheumatoid arthritis. Arthritis Res. 2000;2: 249–251. doi:10.1186/ar95
2. Wang S, Wang Y. Peptidylarginine deiminases in citrullination, gene regulation, health and pathogenesis. Biochim Biophys Acta - Gene Regul Mech. 2013;1829: 1126–1135. doi:10.1016/j.bbagrm.2013.07.003
3. Fuhrmann J, Clancy KW, Thompson PR. Chemical Biology of Protein Arginine Modifications in Epigenetic Regulation. Chem Rev. 2015;115: 5413–5461. doi:10.1021/acs.chemrev.5b00003
4. Christophorou MA, Sharma P, Zhang X, Christophorou MA. Citrullination : new tricks for an old mod. 2023; 4–7.
5. Christophorou MA. The virtues and vices of protein citrullination. R Soc Open Sci. 2022;9: 79–95. doi:10.1098/rsos.220125
6. Vossenaar ER, Zendman AJW, Van Venrooij WJ, Pruijn GJM. PAD, a growing family of citrullinating enzymes: Genes, features and involvement in disease. BioEssays. 2003;25: 1106–1118. doi:10.1002/bies.10357
7. Tanikawa C, Ueda K, Suzuki A, Iida A, Nakamura R, Atsuta N, et al. Citrullination of RGG Motifs in FET Proteins by PAD4 Regulates Protein Aggregation and ALS Susceptibility. Cell Rep. 2018;22: 1473–1483. doi:10.1016/j.celrep.2018.01.031
8. Christophorou MA, Castelo-Branco G, Halley-Stott RP, Oliveira CS, Loos R, Radzisheuskaya A, et al. Citrullination regulates pluripotency and histone H1 binding to chromatin. Nature. 2014;507: 104–108. doi:10.1038/nature12942
9. Sharma P, Azebi S, England P, Christensen T, Møller-Larsen A, Petersen T, et al. Citrullination of Histone H3 Interferes with HP1-Mediated Transcriptional Repression. PLoS Genet. 2012;8: 1–15. doi:10.1371/journal.pgen.1002934
10. Chang X, Xia Y, Pan J, Meng Q, Zhao Y, Yan X. PADI2 is significantly associated with rheumatoid arthritis. PLoS One. 2013;8. doi:10.1371/journal.pone.0081259
11. Ishigami A, Ohsawa T, Hiratsuka M, Taguchi H, Kobayashi S, Saito Y, et al. Abnormal accumulation of citrullinated proteins catalyzed by peptidylarginine deiminase in

hippocampal extracts from patients with Alzheimer's disease. J Neurosci Res. 2005;80: 120–128. doi:10.1002/jnr.20431
12. Arif M, Kato T. Increased expression of PAD2 after repeated intracerebroventricular infusions of soluble Aβ25-35 in the Alzheimer's disease model rat brain: Effect of memantine. Cell Mol Biol Lett. 2009;14: 703–714. doi:10.2478/s11658-009-0029-x
13. Cherrington BD, Zhang X, Mcelwee JL, Morency E, Anguish LJ, Coonrod SA. Potential Role for PAD2 in Gene Regulation in Breast Cancer Cells. PLoS One. 2012;7: e41242. doi:10.1371/journal.pone.0041242
14. Mohanan S, Cherrington BD, Horibata S, Mcelwee JL, Thompson PR, Coonrod SA. Potential Role of Peptidylarginine Deiminase Enzymes and Protein Citrullination in Cancer Pathogenesis. Biochem Res Int. 2012;2012: 895343. doi:10.1155/2012/895343
15. Guo W, Zheng Y, Xu B, Ma F, Li C, Zhang X, Wang Y CX. Investigating the expression , effect and tumorigenic pathway of PADI2 in tumors. Onco Targets Ther. 2017;10: 1475–1485.
16. Wang L, Song G, Zhang X, Feng T, Pan J, Chen W, et al. PADI2-mediated citrullination promotes prostate cancer progression. Cancer Res. 2017;77: 5755–5768. doi:10.1158/0008-5472.CAN-17-0150
17. Horibata S, Rogers KE, Sadegh D, Anguish LJ, Mcelwee JL, Shah P, et al. Role of peptidylarginine deiminase 2 (PAD2) in mammary carcinoma cell migration. BMC Cancer. 2017;17: 378. doi:10.1186/s12885-017-3354-x
18. Song S, Yu Y. Progression on citrullination of proteins in gastrointestinal cancers. Front Oncol. 2019;9: 1–6. doi:10.3389/fonc.2019.00015
19. Yuzhalin AE. Citrullination in cancer. Cancer Res. 2019;79: 1274–1284. doi:10.1158/0008-5472.CAN-18-2797
20. Beato M, Sharma P. Peptidyl arginine deiminase 2 (PADI2)-mediated arginine citrullination modulates transcription in cancer. Int J Mol Sci. 2020;21: 1–16. doi:10.3390/ijms21041351
21. Gao B shan, Rong C shu, Xu H mei, Sun T, Hou J, Xu Y. Peptidyl Arginine Deiminase, Type II (PADI2) Is Involved in Urothelial Bladder Cancer. Pathol Oncol Res. 2020;26: 1279–1285. doi:10.1007/s12253-019-00687-0
22. Xue T, Liu X, Zhang M, Qiukai E, Liu S, Zou M, et al. PADI2-Catalyzed MEK1 Citrullination Activates ERK1/2 and Promotes IGF2BP1-Mediated SOX2 mRNA Stability in Endometrial Cancer. Adv Sci. 2021;8: 1–17. doi:10.1002/advs.202002831
23. Darrah E 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. 2012;71: 92–98. doi:10.1136/ard.2011.151712.Peptidylarginine
24. Dreyton CJ, Knuckley B, Jones JE, Lewallen DM, Thompson PR. Mechanistic studies of protein arginine deiminase 2: Evidence for a substrate-assisted mechanism. Biochemistry. 2014;53: 4426–4433. doi:10.1021/bi500554b
25. Slade DJ, Fang P, Dreyton CJ, Zhang Y, Fuhrmann J, Rempel D, et al. Protein arginine deiminase 2 binds calcium in an ordered fashion: Implications for inhibitor design. ACS Chem Biol. 2015;10: 1043–1053. doi:10.1021/cb500933j
26. Zhang X, Bolt M, Guertin MJ, Chen W, Zhang S, Cherrington BD, et al. Peptidylarginine deiminase 2-catalyzed histone H3 arginine 26 citrullination facilitates estrogen receptor α target gene activation. Proc Natl Acad Sci U S A. 2012;109: 13331–13336. doi:10.1073/pnas.1203280109
27. Guertin MJ, Zhang X, Anguish L, Kim S, Varticovski L, Lis JT, et al. Targeted H3R26 Deimination Specifically Facilitates Estrogen Receptor Binding by Modifying Nucleosome Structure. PLoS Genet. 2014;10: 1–12. doi:10.1371/journal.pgen.1004613
28. Xiao S, Lu J, Sridhar B, Cao X, Yu P, Zhao T, et al. SMARCAD1 Contributes to the Regulation of Naive Pluripotency by Interacting with Histone Citrullination. Cell Rep. 2017;18: 3117–3128. doi:10.1016/j.celrep.2017.02.070
29. Falcão AM, Meijer M, Scaglione A, Rinwa P, Agirre E, Liang J, et al. PAD2-Mediated

Citrullination Contributes to Efficient Oligodendrocyte Differentiation and Myelination. Cell Rep. 2019;27: 1090-1102.e10. doi:10.1016/j.celrep.2019.03.108
30. Sharma P, Lioutas A, Fernandez-Fuentes N, Quilez J, Carbonell-Caballero J, Wright RHG, et al. Arginine Citrullination at the C-Terminal Domain Controls RNA Polymerase II Transcription. Mol Cell. 2019;73: 84-96.e7. doi:10.1016/j.molcel.2018.10.016
31. Cummings TFM, Gori K, Sanchez-Pulido L, Gavriilidis G, Moi D, Wilson AR, et al. Citrullination Was Introduced into Animals by Horizontal Gene Transfer from Cyanobacteria. Mol Biol Evol. 2022;39. doi:10.1093/molbev/msab317
32. Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. Genome Biol. 2015;16: 1–13. doi:10.1186/s13059-015-0607-3
33. Emamalipour M, Seidi K, Zununi Vahed S, Jahanban-Esfahlan A, Jaymand M, Majdi H, et al. Horizontal Gene Transfer: From Evolutionary Flexibility to Disease Progression. Front Cell Dev Biol. 2020;8. doi:10.3389/fcell.2020.00229
34. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: Building the web of life. Nat Rev Genet. 2015;16: 472–482. doi:10.1038/nrg3962
35. Husnik F, McCutcheon JP. Functional horizontal gene transfer from bacteria to eukaryotes. Nat Rev Microbiol. 2018;16: 67–79. doi:10.1038/nrmicro.2017.137
36. Goulas T, Mizgalska D, Garcia-Ferrer I, Kantyka T, Guevara T, Szmigielski B, et al. Structure and mechanism of a bacterial host-protein citrullinating virulence factor, Porphyromonas gingivalis peptidylarginine deiminase. Sci Rep. 2015;5: 1–17. doi:10.1038/srep11969
37. Touz MC, Rópolo AS, Rivero MR, Vranych CV, Conrad JT, Svard SG, et al. Arginine deiminase has multiple regulatory roles in the biology of Giardia lamblia. J Cell Sci. 2008;121: 2930–2938. doi:10.1242/jcs.026963
38. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49: D412–D419. doi:10.1093/nar/gkaa913
39. Asaga H, Ishigami A. Protein deimination in the rat brain after kainate administration: Citrulline-containing proteins as a novel marker of neurodegeneration. Neurosci Lett. 2001;299: 5–8. doi:10.1016/S0304-3940(00)01735-3
40. Rogers G, Winter B, McLaughlan C, Powell B, Nesci T. Peptidylarginine deiminase of the hair follicle: Characterization, localization, and function in keratinizing tissues. J Invest Dermatol. 1997;108: 700–707. doi:10.1111/1523-1747.ep12292083
41. Rus'd AA, Ikejiri Y, Ono H, Yonekawa T, Shiraiwa M, Kawada A, et al. Molecular cloning of cDNAs of mouse peptidylarginine deiminase type I, type III and type IV, and the expression pattern of type I in mouse. Eur J Biochem. 1999;259: 660–669. doi:10.1046/j.1432-1327.1999.00083.x
42. Beltrao P, Bork P, Krogan NJ, Van Noort V. Evolution and functional cross-talk of protein post-translational modifications. Mol Syst Biol. 2013;9: 1–13. doi:10.1002/msb.201304521
43. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34: 1812–1819. doi:10.1093/molbev/msx116
44. Corden JL. An Arginine Nexus in the RNA Polymerase II CTD. Mol Cell. 2019;73: 3–4. doi:10.1016/j.molcel.2018.12.013
45. Marshall NF, Price DH. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. J Biol Chem. 1995;270: 12335–12338. doi:10.1074/jbc.270.21.12335
46. Adelman K, Lis JT. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet. 2012;13: 720–731. doi:10.1038/nrg3293
47. Jonkers I, Lis JT. Getting up to speed with transcription elongation by RNA polymerase II. Nat Rev Mol Cell Biol. 2015;16: 167–177. doi:10.1038/nrm3953

48. Core L, Adelman K. Promoter-proximal pausing of RNA polymerase II: A nexus of gene regulation. Genes Dev. 2019;33: 960–982. doi:10.1101/gad.325142.119
49. Chen FX, Smith ER, Shilatifard A. Born to run: Control of transcription elongation by RNA polymerase II. Nat Rev Mol Cell Biol. 2018;19: 464–478. doi:10.1038/s41580-018-0010-5
50. Martin RD, Hébert TE, Tanny JC. Therapeutic targeting of the general RNA polymerase II transcription machinery. Int J Mol Sci. 2020;21. doi:10.3390/ijms21093354
51. Muniz L, Nicolas E, Trouche D. RNA polymerase II speed: a key player in controlling and adapting transcriptome composition. EMBO J. 2021;40: 1–21. doi:10.15252/embj.2020105740
52. Altenhoff AM, Glover NM, Train CM, Kaleb K, Warwick Vesztrocy A, Dylus D, et al. The OMA orthology database in 2018: Retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res. 2018;46: D477–D485. doi:10.1093/nar/gkx1019
53. Yang Z, Bielawski JR. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 2000;15: 496–503. doi:10.1016/S0169-5347(00)01994-7
54. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005;22: 2472–2479. doi:10.1093/molbev/msi237
55. Kosiol C, Vinař T, Da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, et al. Patterns of positive selection in six mammalian genomes. PLoS Genet. 2008;4. doi:10.1371/journal.pgen.1000144
56. Segura J, Jones PF, Fernandez-Fuentes N. A holistic in silico approach to predict functional sites in protein structures. Bioinformatics. 2012;28: 1845–1850. doi:10.1093/bioinformatics/bts269
57. Jones S, Heyningen P Van, Berman HM, Thornton JM. Arcturus_Summer_Hba. 1999.
58. Segura J, Oliva B, Fernandez-Fuentes N. CAPS-DB: A structural classification of helix-capping motifs. Nucleic Acids Res. 2012;40: 479–485. doi:10.1093/nar/gkr879
59. Baumli S, Lolli G, Lowe ED, Troiani S, Rusconi L, Bullock AN, et al. The structure of P-TEFb (CDK9/cyclin T1), its complex with flavopiridol and regulation by phosphorylation. EMBO J. 2008;27: 1907–1918. doi:10.1038/emboj.2008.121
60. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A GT, Žídek A, Bates R, Blackwell S, Yim J, Ronneberger O BS, Zielinski M, Bridgland A, Potapenko A, Cowie A, Tunyasuvunakool K JR, Clancy E, Kohli P JJ and DHD. Protein complex prediction with AlphaFold-Multimer. bioRxiv. 2022. Available: https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2
61. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589. doi:10.1038/s41586-021-03819-2
62. Lin FR, Niparko JK, Ferrucci and L. 基因的改变NIH Public Access. Bone. 2014;23: 1–7. doi:10.1146/annurev-biochem-052610-095910.RNA
63. Kohoutek J. P-TEFb- The final frontier. Cell Div. 2009;4: 19. doi:10.1186/1747-1028-4-19
64. Manuscript A. multi-tasking P-TEFb. 2009;20: 334–340. doi:10.1016/j.ceb.2008.04.008.The
65. Gabay M, Li Y, Felsher DW. MYC activation is a hallmark of cancer initiation and maintenance. Cold Spring Harb Perspect Med. 2014;4: 1–14. doi:10.1101/cshperspect.a014241
66. Mollaoglu G, Guthrie MR, Böhm S, Brägelmann J, Can I, Ballieu PM, et al. MYC Drives Progression of Small Cell Lung Cancer to a Variant Neuroendocrine Subtype with Vulnerability to Aurora Kinase Inhibition. Cancer Cell. 2017;31: 270–285. doi:10.1016/j.ccell.2016.12.005
67. Qiu X, Boufaied N, Hallal T, Feit A, Polo A De, Luoma AM, et al. MYC drives aggressive prostate cancer by disrupting transcriptional pause release at androgen receptor targets. Nat Commun. 2022;13: 2559. Available: https://pubmed.ncbi.nlm.nih.gov/35562350/

68. Cui Z, Xiao L, Chen F, Wang J, Lin H, Li D, et al. High mRNA Expression of CENPL and Its Significance in Prognosis of Hepatocellular Carcinoma Patients. Dis Markers. 2021;2021. doi:10.1155/2021/9971799
69. Zhang H, Zhang X, Li X, Meng WB, Bai ZT, Rui SZ, et al. Effect of CCNB1 silencing on cell cycle, senescence, and apoptosis through the p53 signaling pathway in pancreatic cancer. J Cell Physiol. 2018;234: 619–631. doi:10.1002/jcp.26816
70. Chen EB, Qin X, Peng K, Li Q, Tang C, Wei YC, et al. HnRNPR-CCNB1/CENPF axis contributes to gastric cancer proliferation and metastasis. Aging (Albany NY). 2019;11: 7473–7491. doi:10.18632/aging.102254
71. Liu C, Kudo T, Ye X, Gascoigne K. Cell-to-cell variability in Myc dynamics drives transcriptional heterogeneity in cancer cells. Cell Rep. 2023;42: 112401. doi:10.1016/j.celrep.2023.112401
72. Jones KE, Bielby J, Cardillo M, Fritz SA, O'Dell J, Orme CDL, et al. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology. 2009;90: 2648–2648. doi:10.1890/08-1494.1
73. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods. 2009;6: 343–345. doi:10.1038/nmeth.1318
74. Vicent GP, Zaurin R, Nacht AS, Li A, Font-Mateu J, Le Dily F, et al. Two chromatin remodeling activities cooperate during activation of hormone responsive promoters. PLoS Genet. 2009;5. doi:10.1371/journal.pgen.1000567
75. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102: 15545–15550. doi:10.1073/pnas.0506580102
76. Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34: 267–273.
77. Veidenberg A, Medlar A, Loytynoja A. Wasabi: An integrated platform for evolutionary sequence analysis and data visualization. Mol Biol Evol. 2016;33: 1126–1130. doi:10.1093/molbev/msv333
78. Villanueva-Cañas JL, Ruiz-Orera J, Agea MI, Gallo M, Andreu D, Albà MM. New genes and functional innovation in mammals. Genome Biol Evol. 2017;9: 1886–1900. doi:10.1093/gbe/evx136
79. Villanueva-Cañas JL, Laurie S, Albà MM. Improving genome-wide scans of positive selection by using protein isoforms of similar length. Genome Biol Evol. 2013;5: 457–467. doi:10.1093/gbe/evt017
80. Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30: 3059–3066. doi:10.1093/nar/gkf436
81. Larsson A. AliView: A fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014;30: 3276–3278. doi:10.1093/bioinformatics/btu531
82. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25: 1972–1973. doi:10.1093/bioinformatics/btp348
83. Suyama M, Torrents D, Bork P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34: 609–612. doi:10.1093/nar/gkl315
84. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313. doi:10.1093/bioinformatics/btu033
85. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24: 1586–1591. doi:10.1093/molbev/msm088

86. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016;33: 1635–1638. doi:10.1093/molbev/msw046
87. Segura J, Marín-López MA, Jones PF, Oliva B, Fernandez-Fuentes N. VORFFIP-driven dock: V-D2OCK, a fast, accurate protein docking strategy. PLoS One. 2015;10: 1–12. doi:10.1371/journal.pone.0118107
88. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics. 2013;29: 845–854. doi:10.1093/bioinformatics/btt055
89. Feliu E, Aloy P, Oliva B. On the analysis of protein-protein interactions via knowledge-based potentials for the prediction of protein-protein docking. Protein Sci. 2011;20: 529–541. doi:10.1002/pro.585
90. Pei J, Grishin N V. AL2CO: Calculation of positional conservation in a protein sequence alignment. Bioinformatics. 2001;17: 700–712. doi:10.1093/bioinformatics/17.8.700

Figures and Tables

Fig 1 (related to S1-S4 figs and File S1) PADI2 conservation across species. (A) Diagram showing the main steps involved in 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, as calculated by ete3.

Fig 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 columns with positions that are detected as being 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.

Fig 3 (related to Table 1, 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 according to 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.

Fig 4 (related to File S5, Movie S4, S4 Fig). 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 shown as a blue sphere. (**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.

Fig 5 (related to S5). Positively selected L162 encompassing loop 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 ≥ 3 biological experiments for all plots in the figure. Two-tailed unpaired Student's t-test was used to determine the 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.

Fig 6 (related to S6). The L162 enclosed loop regulates the transcription regulation. (**A**). Dotplot depicting the top 55 up- and downregulated genes in L162A/W161A/T159A-PADI2 mutants vs. WT. Selected were the p_{adj} <0.05 and highest FC (fold change) according to the Find Markers function in Seurat (4.3.0, log2FC.threshold=0.25, min. pct=0.1). (**B**). Dot plots depicting the expression of c-MYC and CCNB1 in L162A/W161A/T159A-PADI2 mutants vs. WT. The circle size depicts the percentage of cells expressing the respective gene, while the color encodes the average expression. (**C**). Gene set enrichment analysis (GSEA) for molecular functions and biological processes. Representative processes are presented (*Upper* -Downregulated genes, *Lower*- Up-regulated genes). The X-axis shows the -log10 transformed p-values. GO, Gene Ontology. (**D-E**). The connection map of the target genes from the representative regulatory elements in WT/ (L162A/W161A/T159A) PADI2 (**D**) upregulated genes (**E**) downregulated genes.

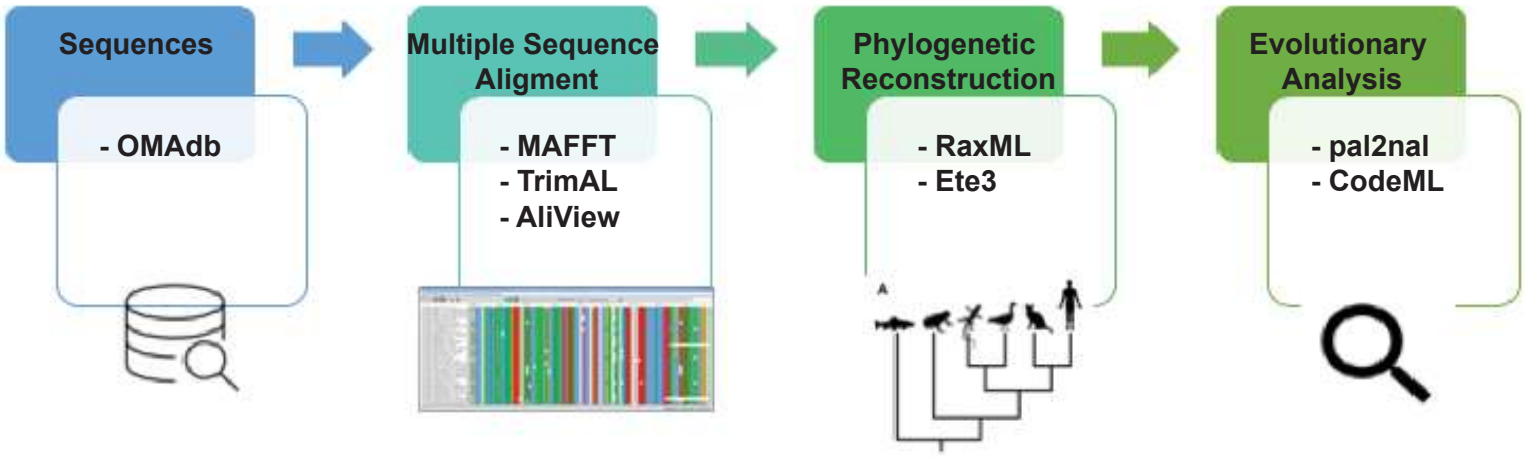
Table 1. Positively selected sites across species under the branch site (CodeML) and mapped onto the PADI2 human structure. The position of the amino acid in the domain, MSA, and PDB file (Supplementary material file S2) are shown in the domain, sequence position, and PDB number, respectively. The flanking residues are shown with the given positively selected amino acid depicted in bold. The type of secondary structure and solvent accessibility (i.e., exposure) is shown in the secondary structure and accessibility column. Finally, the MVORFFIP scores (see Materials and 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	WH**M**VP	*0.7*	Partially buried

Fig1

A


Sequences
- OMAdb
Multiple Sequence Aligment
- MAFFT
- TrimAL
- AliView
Phylogenetic Reconstruction
- RaxML
- Ete3
Evolutionary Analysis
- pal2nal
- CodeML

B

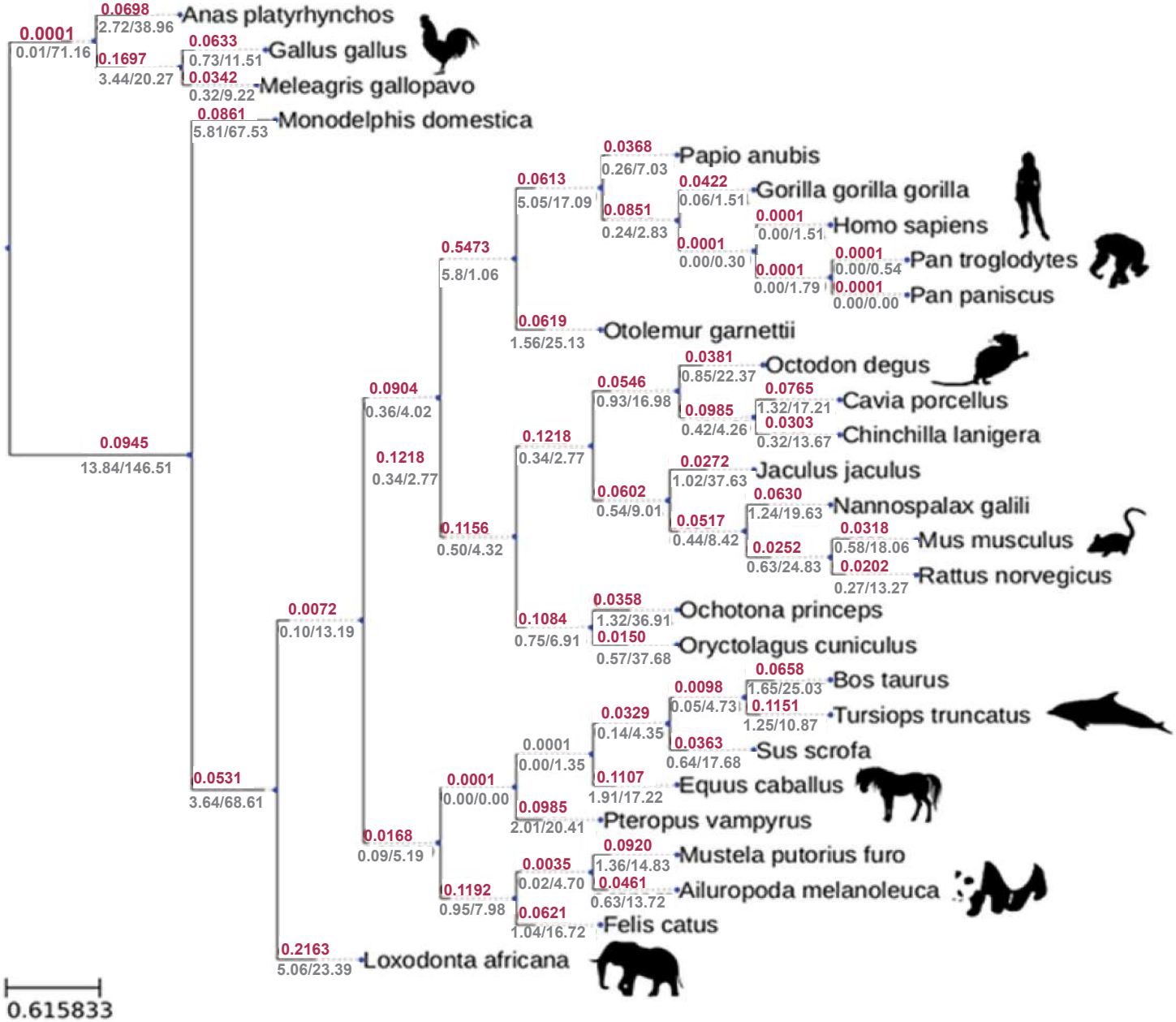

Anas platyrhynchos
Gallus gallus
Meleagris gallopavo
Monodelphis domestica
Papio anubis
Gorilla gorilla gorilla
Homo sapiens
Pan troglodytes
Pan paniscus
Otolemur garnettii
Octodon degus
Cavia porcellus
Chinchilla lanigera
Jaculus jaculus
Nannospalax galili
Mus musculus
Rattus norvegicus
Ochotona princeps
Oryctolagus cuniculus
Bos taurus
Tursiops truncatus
Sus scrofa
Equus caballus
Pteropus vampyrus
Mustela putorius furo
Ailuropoda melanoleuca
Felis catus
Loxodonta africana
0.615833

Fig 2

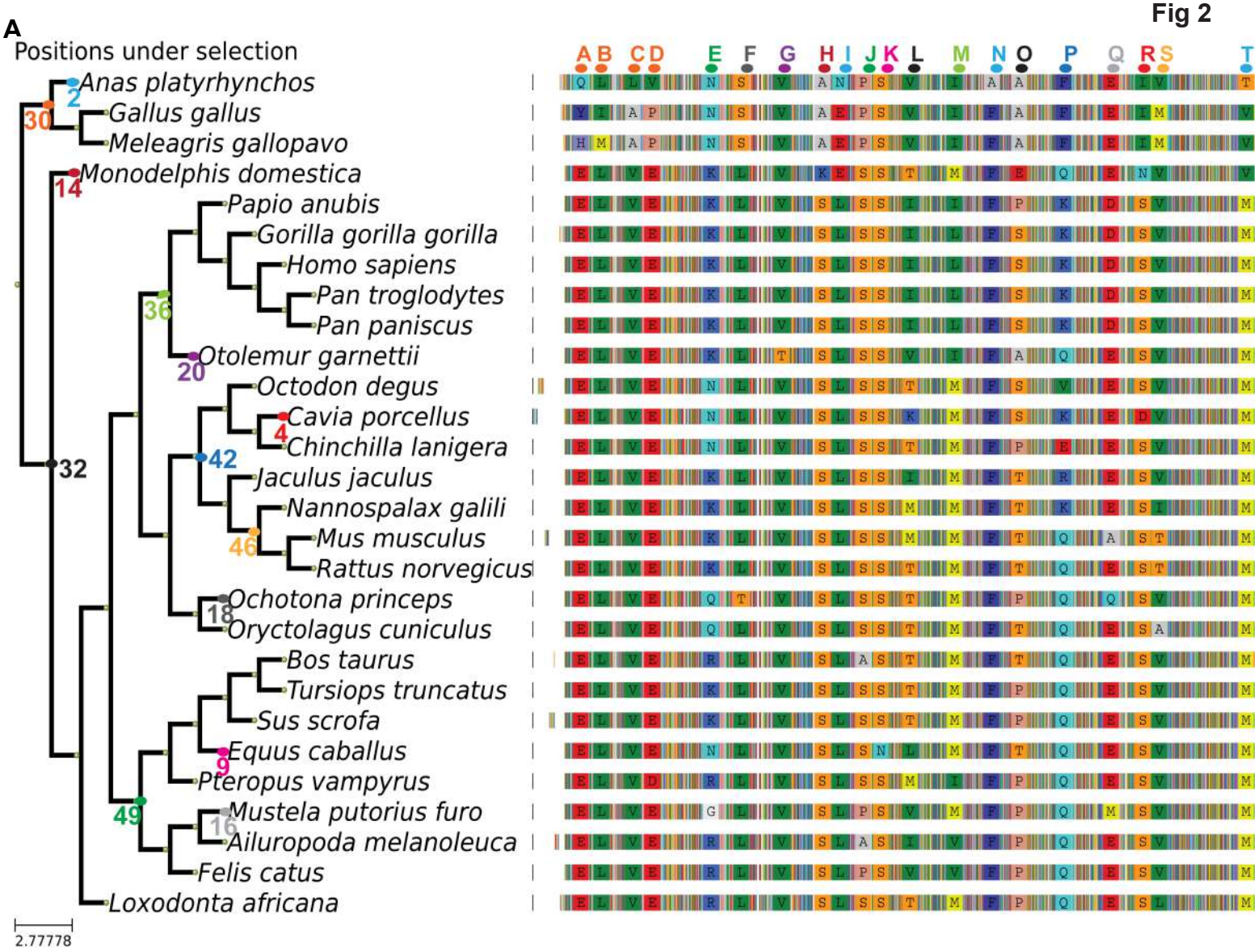

A
Positions under selection
A B C D E F G H I J K L M N O P Q R S T
Anas platyrhynchos
Gallus gallus
Meleagris gallopavo
Monodelphis domestica
Papio anubis
Gorilla gorilla gorilla
Homo sapiens
Pan troglodytes
Pan paniscus
Otolemur garnettii
Octodon degus
Cavia porcellus
Chinchilla lanigera
Jaculus jaculus
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

Fig 3

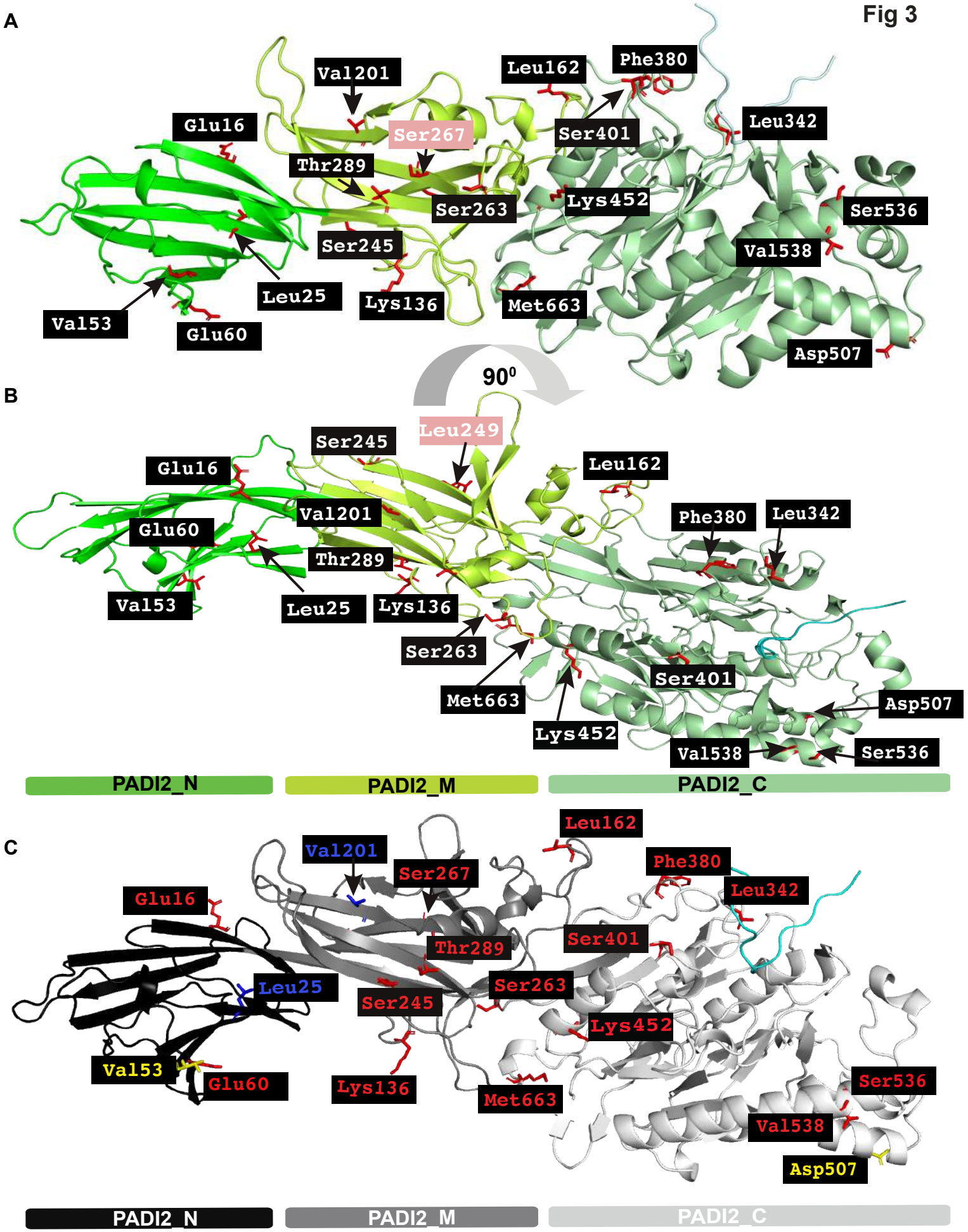

A
Val201
Leu162
Phe380
Glu16
Ser267
Ser401
Leu342
Thr289
Ser263
Lys452
Ser536
Ser245
Val538
Leu25
Lys136
Met663
Val153
Glu60
Asp507
90⁰
B
Leu249
Ser245
Leu162
Glu16
Val201
Phe380
Leu342
Glu60
Thr289
Val153
Leu25
Lys136
Ser263
Ser401
Met663
Asp507
Lys452
Val538
Ser536
PADI2_N
PADI2_M
PADI2_C
C
Leu162
Val201
Ser267
Phe380
Leu342
Glu16
Ser401
Thr289
Leu25
Ser263
Ser245
Lys452
Val153
Glu60
Lys136
Met663
Ser536
Val538
Asp507
PADI2_N
PADI2_M
PADI2_C
D
Glu16
Leu25
Val153
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)

Fig 4

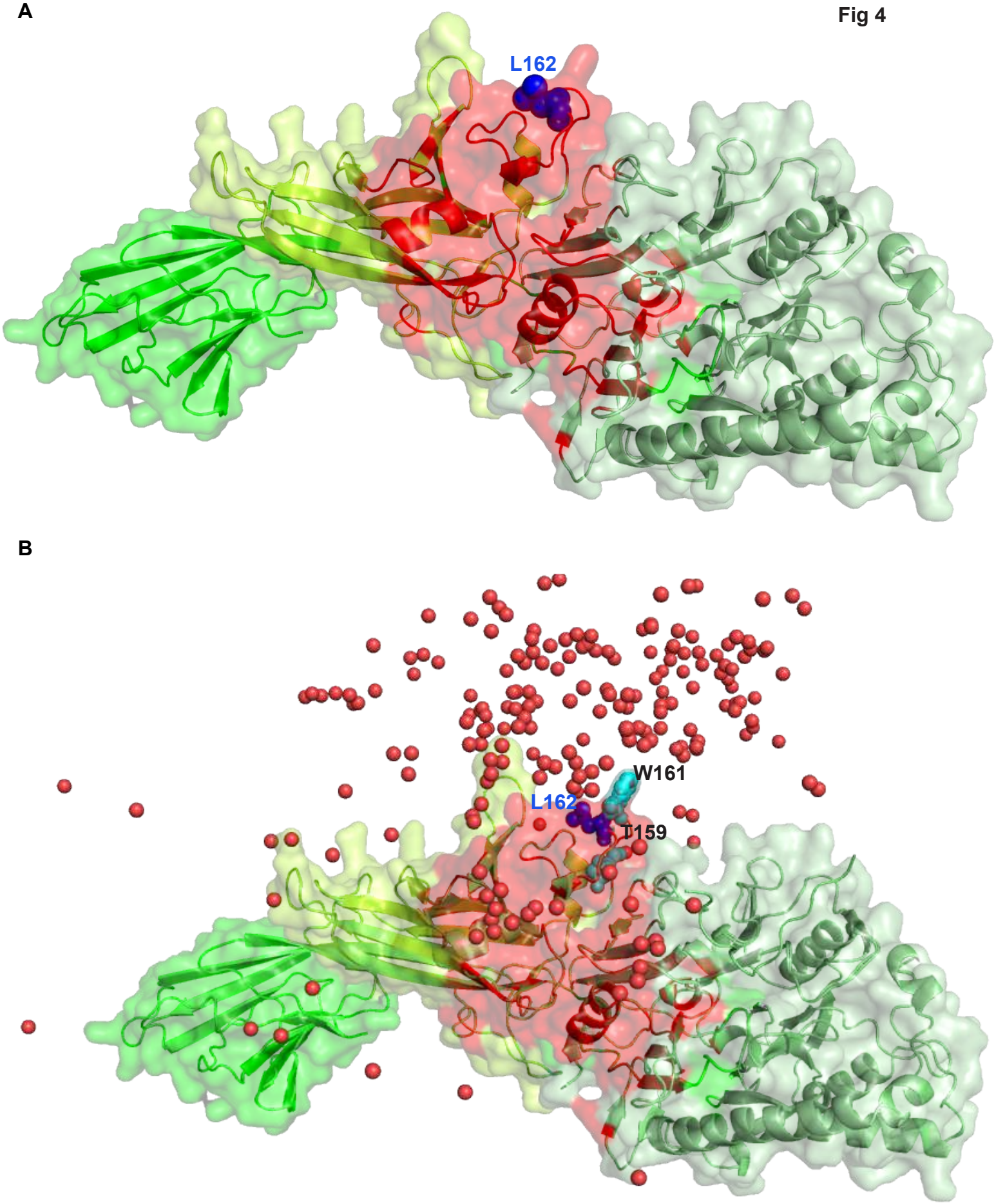
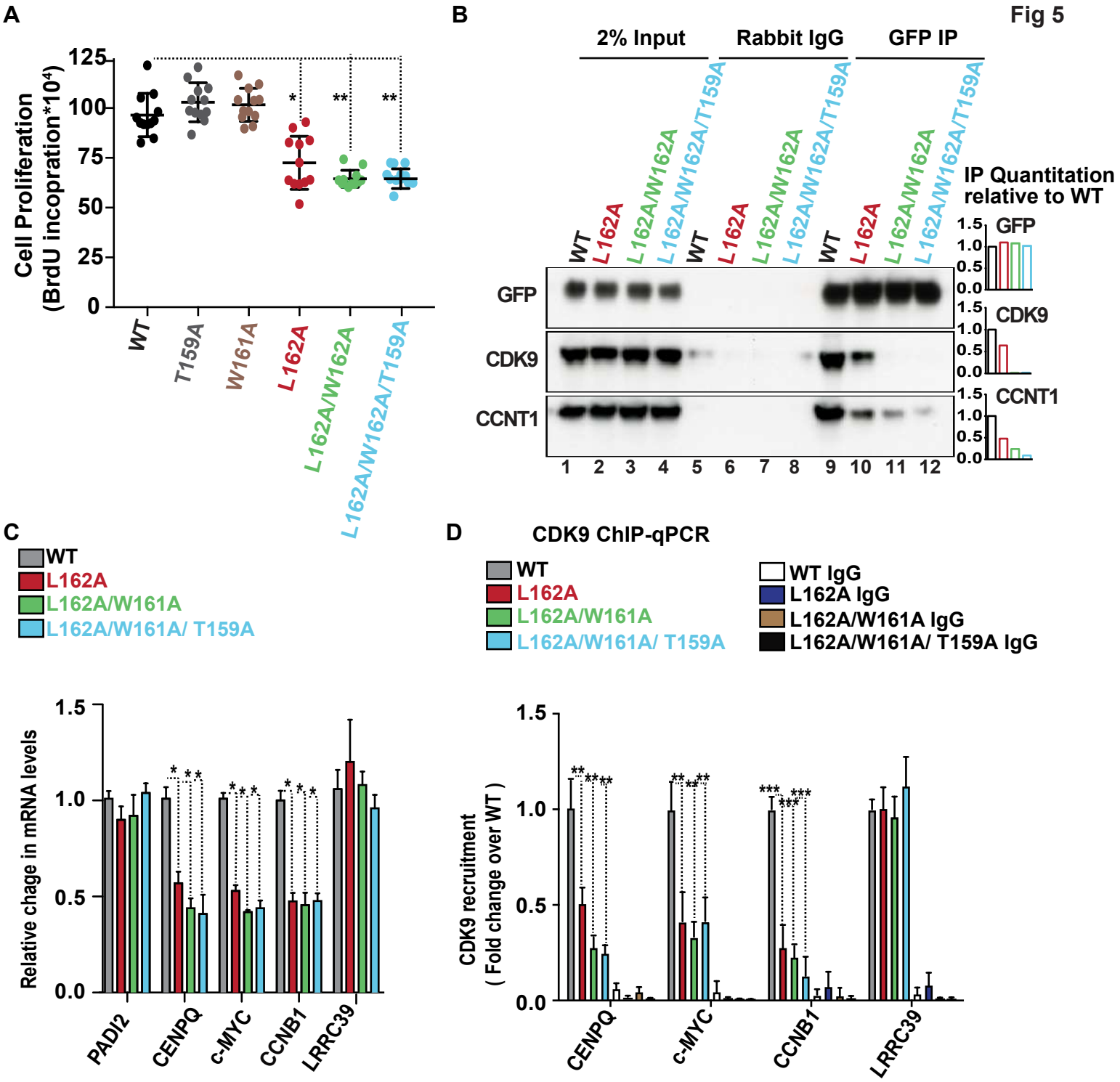


Fig 5


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
GFP
CDK9
CCNT1
1 2 3 4 5 6 7 8 9 10 11 12
IP Quantitation relative to WT
1.5
1.0
0.5
0.0
C
WT
L162A
L162A/W161A
L162A/W161A/ T159A
Relative chage in mRNA levels
PADI2
CENPQ
c-MYC
CCNB1
LRRC39
D
CDK9 ChIP-qPCR
WT
L162A
L162A/W161A
L162A/W161A/ T159A
WT IgG
L162A IgG
L162A/W161A IgG
L162A/W161A/ T159A IgG
CDK9 recruitment (Fold change over WT)
CENPQ
c-MYC
CCNB1
LRRC39

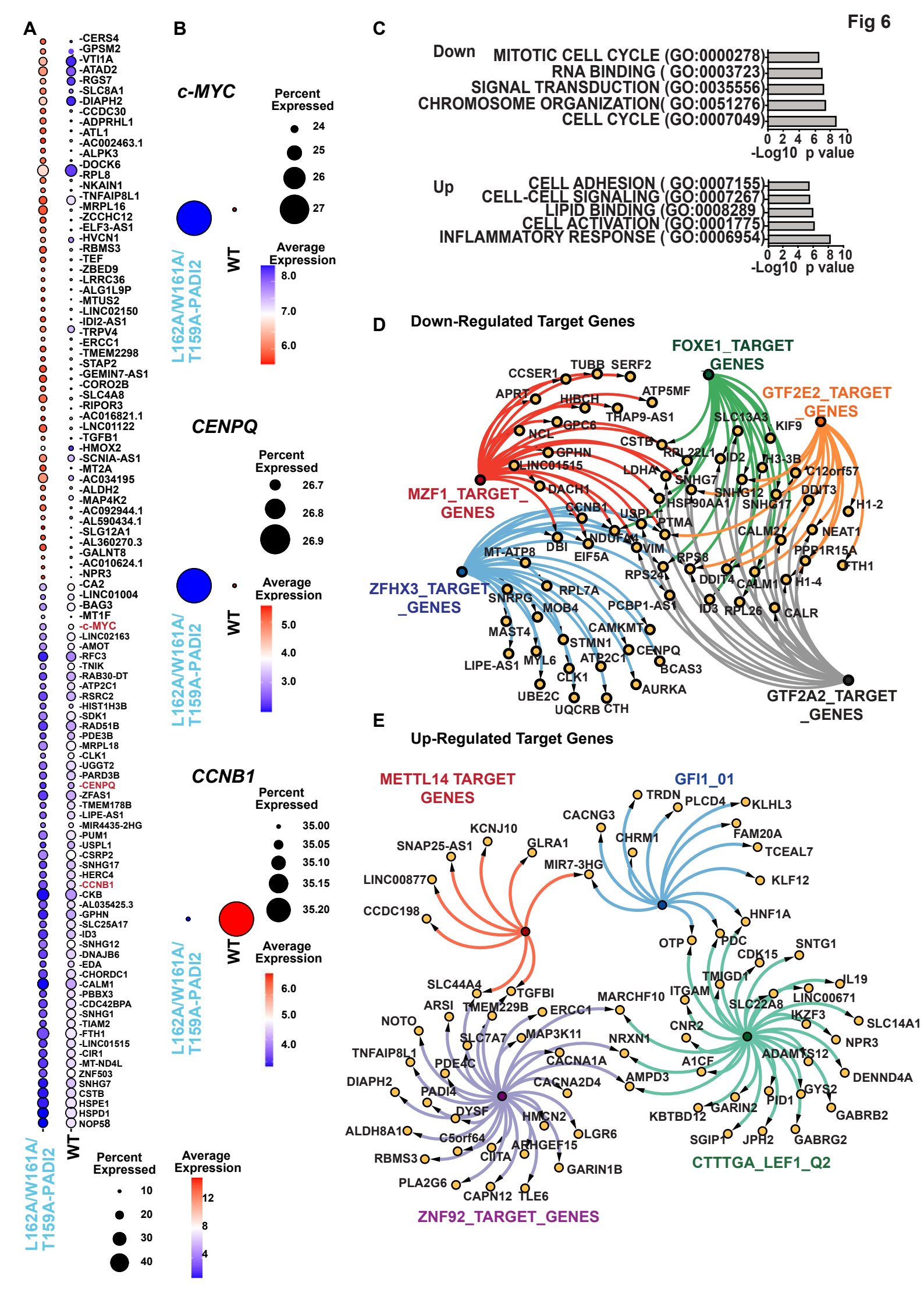


Fig 6
A
B
C
D
E
c-MYC
CENPQ
CCNB1
Percent Expressed
Average Expression
L162A/W161A/ T159A-PADI2
WT
Down
MITOTIC CELL CYCLE (GO:0000278)
RNA BINDING (GO:0003723)
SIGNAL TRANSDUCTION (GO:0035556)
CHROMOSOME ORGANIZATION(GO:0051276)
CELL CYCLE (GO:0007049)
Up
CELL ADHESION (GO:0007155)
CELL-CELL SIGNALING (GO:0007267)
LIPID BINDING (GO:0008289)
CELL ACTIVATION (GO:0001775)
INFLAMMATORY RESPONSE (GO:0006954)
-Log10 p value
Down-Regulated Target Genes
FOXE1_TARGET GENES
GTF2E2_TARGET _GENES
MZF1_TARGET_ GENES
ZFHX3_TARGET _GENES
GTF2A2_TARGET _GENES
Up-Regulated Target Genes
METTL14 TARGET GENES
GFI1_01
CTTTGA_LEF1_Q2
ZNF92_TARGET_GENES