Citrullination was introduced into animals by horizontal gene transfer from cyanobacteria

Thomas F. M. Cummings[1†*], Kevin Gori[2], Luis Sanchez-Pulido[1], Gavriil Gavriilidis[1], David Moi[3,4], Abigail R. Wilson[1], Elizabeth Murchison[2], Christophe Dessimoz[3,4,5], Chris P. Ponting[1] and Maria A. Christophorou[1,6* **].

[1] MRC Human Genetics Unit, The Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, United Kingdom

[2] Transmissible Cancer Group, Department of Veterinary Medicine, Madingley Road, Cambridge CB3 0ES, United Kingdom

[3] Department of Computational Biology, and Center for Integrative Genomics, University of Lausanne, Genopode, 1015 Lausanne, Switzerland

[4] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

[5] Department of Genetics Evolution and Environment, and Department of Computer Science, University College London, Darwin Building, Gower Street, London, WC1E 6BT

[6] The Babraham Institute, Cambridge, CB22 3AT, United Kingdom

[†] Current address: MRC Protein Phosphorylation & Ubiquitylation Unit, School of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, UK

* Correspondence and requests for materials: maria.christophorou@babraham.ac.uk; tfmcummings@gmail.com

** Lead contact: maria.christophorou@babraham.ac.uk

Abstract

Protein post-translational modifications (PTMs) add great sophistication to biological systems. Citrullination, a key regulatory mechanism in human physiology and pathophysiology, is enigmatic from an evolutionary perspective. Although the citrullinating enzymes peptidylarginine deiminases (PADIs) are ubiquitous across vertebrates, they are absent from yeast, worms and flies. Based on this distribution *PADIs* were proposed to have been horizontally transferred, but this has been contested. Here, we map the evolutionary trajectory of *PADIs* into the animal lineage. We present strong phylogenetic support for a clade encompassing animal and cyanobacterial *PADIs* that excludes fungal and other bacterial homologues. The animal and cyanobacterial PADI proteins share functionally relevant primary and tertiary synapomorphic sequences that are distinct from a second PADI type present in fungi and actinobacteria. Molecular clock calculations and sequence divergence analyses using the fossil record estimate the last common ancestor of the cyanobacterial and animal *PADIs* to be less than one billion years old. Additionally, under an assumption of vertical descent, PADI sequence change during this evolutionary time frame is anachronistically low, even when compared to products of likely endosymbiont gene transfer, mitochondrial proteins and some of the most highly conserved sequences in life. The consilience of evidence indicates that *PADIs* were introduced from

cyanobacteria into animals by horizontal gene transfer (HGT). The ancestral cyanobacterial PADI is enzymatically active and can citrullinate eukaryotic proteins, suggesting that the *PADI* HGT event introduced a new catalytic capability into the regulatory repertoire of animals. This study reveals the unusual evolution of a pleiotropic protein modification.

Introduction

Post-translational modifications (PTMs) allow for temporal and spatial control of protein function in response to cellular and environmental signals and comprise an integral part of cellular and organismal life. The development of ever more sensitive and quantitative analytical methods has made possible the identification of PTMs within cells and has enhanced our understanding of the molecular and cellular functions they regulate. This has led to renewed interest in studying previously known, as well as newly identified modifications. Although PTMs have been classically studied in eukaryotic organisms, an increasing number of them are also found in bacteria (Koonin 2010). Some PTMs, such as phosphorylation, acetylation and glycosylation are ubiquitous across all domains of life suggesting that the enzymes that catalyse them existed in the Last Universal Common Ancestor (LUCA) (Beltrao et al. 2013). In other cases, such as protein ubiquitylation, this is less clear. Although ubiquitin itself is absent from eubacteria and archaea, other ubiquitin-like domains have been identified and shown to be added and removed from proteins in a similar manner in bacteria (Iyer et al. 2008; Pearce et al. 2008; Hochstrasser 2009; Koonin 2010; Macek et al. 2019).

Citrullination is the post-translational conversion of a protein arginine residue to the non-coded amino acid citrulline and is catalysed by PADI enzymes in a calcium-dependent manner (Sugawara et al. 1982; Wang and Wang 2013). Although citrullination involves a small mass change of only 0.98Da, the removal of a positive charge from the arginine side chain can lead to profound biochemical changes and is known to alter protein structure, sub-cellular localisation and affinity to other proteins and nucleic acids (Tanikawa et al. 2009; Guo and Fast 2011; Stadler et al. 2013; Christophorou et al. 2014; Snijders et al. 2015; Tanikawa et al. 2018; Sharma et al. 2019). Via these alterations PADIs regulate fundamental physiological processes such as gene expression, chromatin compaction and the innate immune response to bacterial infection (Wang et al. 2009; Wang and Wang 2013; Christophorou et al. 2014). Notably, deregulation of PADIs is strongly implicated in the aetiology of a host of pathologies including autoimmunity (rheumatoid arthritis, ulcerative colitis, psoriasis and type I diabetes), neurodegeneration (multiple sclerosis, Alzheimer's and prion diseases) and metastatic cancer (Suzuki et al. 2003; Musse et al. 2008; Zhang et al. 2011; Wang and Wang 2013; Yuzhalin et al.

2018), while loss of PADI activity compromises neurodevelopment, fertility and embryo development (Christophorou et al. 2014; Xu et al. 2016; Falcão et al. 2019).

In an evolutionary context, *PADIs* are puzzling. Orthologues of the human *PADIs* are ubiquitous in bony fish, birds, reptiles, amphibians and mammals, but are unexpectedly missing from many eukaryotes including plants, yeast, worms and insects. The *PADI* gene is widely thought to have appeared first in the last common ancestor of teleosteans and mammals (Balandraud et al. 2005; Wang and Wang 2013; Nicholas and Bhattacharya 2014), with duplications in subsequent lineages resulting in five mammalian paralogues. Therefore, citrullination seemingly defies the perception that PTMs are of ancient origin (Beltrao et al. 2013).

Mammalian PADIs consist of three structural domains, the N-terminal (PAD_N, Pfam annotation: PF08526), middle (PAD_M, Pfam annotation: PF08527) and catalytic C-terminal domain (PAD_C, Pfam annotation: PF03068). Although PADI proteins are widely considered to be specific to vertebrates, their crystal structures (Arita et al. 2004; Slade et al. 2015) hint at a possibly more ancient origin as they reveal that the catalytic (PAD_C) domain adopts the same pentein fold as a variety of other widely distributed proteins that otherwise show little similarity in terms of amino acid conservation (Shirai et al. 2001; Linsky and Fast 2010) (Figure S1). The pentein-fold containing group of proteins comprises a broad family of guanidino-group (the functional group of the side chain of arginine and agmatine) modifying enzymes that possess hydrolase, dihydrolase and amidinotransferase catalytic activity, sharing a catalytic core of a Cys, His and two polar guanidine binding residues – Asp or Glu (Linsky and Fast 2010). Two such proteins with citrullinating activity are known among some bacteria and eukaryotes: pPAD, an extended agmatine deiminase found in *Porphyromonas gingivalis* and giardiaADI, an extended form of the free *L*-arginine deiminase gADI, found in the human parasite *Giardia Lamblia* (Touz et al. 2008; Goulas et al. 2015). These enzymes contain a distant PAD_C domain but lack PAD_N and PAD_M domains, are highly divergent in sequence and have different substrate specificities. In addition, mammalian genomes encode two distant homologues of the PAD_C domain: N(G),N(G)-dimethylarginine dimethylaminohydrolase [DDAH] and Glycine amidinotransferase [AGAT] (Linsky and Fast 2010). Both DDAH and AGAT are divergent in sequence, also lack PAD_N and PAD_M domains and do not appear to catalyse citrullination. The presence of this ancient fold and catalytic triad within PAD_C suggests that it may have been present early in cellular life, but the evolutionary provenance of the animal PADI enzymes has remained unclear.

A 2015 study by Crisp *et al.,* identified possible *PADI* homologues in some bacterial species. Based on the finding that a possible homologue could be identified in prokaryotes but

not in multiple *Drosophila* and *Caenorhabditis* species, the authors included *PADIs* among a list of 145 genes proposed to have been transferred into the genome of a vertebrate ancestor of extant mammals by horizontal gene transfer (HGT, also known as lateral gene transfer) (Crisp et al. 2015). HGT is the non-heritable transmission of genetic material from one organism to another, often via a virus or mobile genetic element and involving endosymbiotic or commensal relationships between donor and recipient (Boto 2014; Soucy et al. 2015). HGT is widespread among prokaryotes and is recognised as a mechanism that shapes the evolution and adaptive potential of bacteria, for example in the acquisition of antibiotic resistance (Ochman et al. 2000; Koonin et al. 2001). Although many cases of horizontal transfer have been reported between bacteria and unicellular eukaryotes, fewer bacteria-to-animal HGT events have been reported to date (Keeling and Palmer 2008; Dunning Hotopp 2011; Boto 2014). The majority of cases involve transfer into an invertebrate host, such as an insect or worm (Gladyshev et al. 2008; Moran and Jarvik 2010; Chou et al. 2015; Lacroix and Citovsky 2016; Dunning Hotopp 2018). Moreover, it has been proposed that HGT into animals with specialised germline cells is very rare (Jensen et al. 2016). These few accounts of bacteria-to-animal HGT have been the topic of intense debate (Stanhope et al. 2001; Crisp et al. 2015; Martin 2017; Salzberg 2017; Husnik and McCutcheon 2018; Leger et al. 2018). The genome-wide approach employed by Crisp *et al.* to search for possible HGT events in vertebrates was disputed by Salzberg, and 45 of the highest confidence candidates were re-analyzed and rebutted on a case-by-case basis. In the instance of the *PADI* gene, this reanalysis showed that a *PADI* can also be identified in *Priapulus caudatus* (a marine worm) and therefore that the lack of *PADI* in at least *Drosophila spp.* must be explained by gene loss (Salzberg 2017). Salzberg additionally recalculated the HGT index for many of the possible HGT candidates, including the *PADIs*, in light of additional sequences that can be identified showing that they no longer pass the original parametric criterion for HGT proposed by Crisp *et al.* (Salzberg 2017). Individual claims of HGT should be considered carefully and tested against the alternative hypothesis of widespread independent gene losses (Salzberg 2017). In light of the absence of *PADI* homologues in most invertebrate animals, *PADI* evolution requires detailed consideration.

Results

Comprehensive Identification of PADI homologues

In order to understand the distribution and evolution of citrullination we sought to identify all *PADI* homologues from across life. We started by collecting orthologous *PADIs* using the EggNOG database, employing an unsupervised clustering algorithm of all proteins contained in 2031 genomes across cellular life (Huerta-Cepas et al. 2016). To expand on this list, we used HMMER searches to identify all sequences in current sequence databases that contain a

PAD_C domain, as defined by having significant sequence similarity (E-value < $1x10^{-3}$), and assessed these for the presence of critical substrate-binding and calcium-binding residues annotated to human PADIs (Slade et al. 2015). This was supplemented by additional iterative jackhmmer searches as well as tblastn and Position-Specific Iterated BLAST (PSI-BLAST) searches of genomic databases.

The taxonomic distribution of *PADIs* and proportion of species that harbour a PADI orthologue are presented in Table 1. PADIs are not ubiquitous across the metazoa but are present across major branches of vertebrates, including jawless fish, sharks and rays, bony fish, amphibians, reptiles, birds and mammals. Out of all species whose genomes have been sequenced to date, the earliest diverging invertebrate animals with a *PADI* gene are *Priapulus caudatus* (an ecdysozoan), *Saccoglossus kowalevskii* (a hemichordate), and *Branchiostoma belcheri* (a cephalochordate). In addition, we identified a number of PADI sequences with conservation of substrate and calcium-binding residues in bacteria and fungi. *PADIs* are also not ubiquitous across bacteria (found in fewer than 1% of bacterial species), and are most prevalent within cyanobacteria (found in 11% of cyanobacteria). No eukaryotes diverging before opisthokonta have a detectable *PADI* homologue. Our searches also returned two outliers, one in archaea and one in viruses. However, upon closer inspection, both hits were determined to be due to misattribution (Figures S2, S3; *see also* Methods) and were therefore not included in further analyses. This taxonomic distribution could suggest an evolutionary model in which *PADI* genes were lost independently in many separate lineages. In this scenario, gene loss occurred in all early-branching lineages leading to at least 306 non-opisthokont eukaryotes and in other lineages, for example those leading to *Drosophila* and *Caenorhabditis*.

To explore the relationship of PADIs to other distantly related sequences, we aligned fungal, bacterial and animal PADIs with sequences possessing significant HMMER similarity to pPAD and gADI and conducted phylogenetic analysis under a time-reversible model (Figure S4). Bacterial, fungal and animal PADIs form a single outgroup that excludes both pPAD and gADI enzyme types, showing that each of the three types of protein is phyletically distinct. The pPAD and gADI type proteins can therefore be excluded from further consideration of the evolutionary origin of animal PADIs.

A strongly supported clade contains cyanobacterial and animal but not fungal PADIs

Firstly, we used HMMER to obtain all PADI sequences in the UniProtKB rp55 database and performed phylogenetic analysis using MrBayes and IQTree, recovering a clade of animal and bacterial PADIs distinct from fungal and other bacterial PADIs (Figure S5a,b). We then repeated the phylogenetic analyses on a subset of 150 sequences, ensuring the length of the

alignment of PADI sequences (495 columns) was at least three times the number of taxa considered in the tree, to limit “rough likelihood surface” issues that may arise with datasets of relatively few sites and many taxa (Stamatakis et al. 2020) (Figure S5c). To avoid possible biases in subsampling, we took all bacterial PADI sequences contained within the Pathosystems Resource Integration Center (PATRIC) database for analysis (82 sequences). We then included 35 fungal sequences that cover the broadest span in HMMER sequence similarity to the human sequence (E-values between $5x10^{-26}$ and $1.4x10^{-46}$). Finally, we subsampled metazoan sequences to maximise lineage representation in species maintaining a PADI (the 5 paralogues in *Homo sapiens*, *Pongo abelii* and in *Mus musculus*, the 3 paralogues found in *Gallus gallus, Chelonia mydas* and *Alligator mississipiensis*, and the single paralogue found in *Xenopus laevis, Takifugu rubripes, Tetraodon nigroviridis, Astyanax mexicanus, Danio rerio*, *Oncorhynchus mykiss, Callorhinchus milii, Branchiostoma floridae* and *Priapulus caudatus*). Amino acid sequences were used as this enables more reliable alignment among widely divergent taxa. This is especially important as PADI sequences span across bacteria, fungi and metazoa. All sequences, intermediate alignments and trees are provided in Supplementary Files 1-5. Very strong bootstrap support (>95%) was obtained for a clade restricted to certain cyanobacterial and animal PADIs that excludes a fully supported outgroup clade containing fungal, actinobacterial and proteobacterial sequences (Figure S5a-c). With full branch support, the fungal and actinobacterial sequences were recovered as clades and found to be sister taxa in the tree. This tree topology, whereby animal sequences have closer affinity to those in cyanobacteria than to other eukaryotic (fungal) sequences is surprising because it is inconsistent with the known species tree.

Phylogenetic tree inferences, in particular those obtained from single genes, are subject to errors. It is possible that the observed topology represents the failure of phylogenetic inference in the case of this individual gene, such that an artefact (e.g. model misspecification) might explain the affinity of the separate eukaryotic *PADIs* to different bacterial *PADI* types. For instance, using a fixed rate matrix of amino acid substitutions to produce the tree (Jones et al. 1992; Whelan and Goldman 2001; Kalyaanamoorthy et al. 2017) can be inappropriate if there is evolutionary rate variation over different parts of the tree or deviation from typical protein substitution rates. In particular, attention has been drawn previously to heterotachous evolution, where the evolutionary substitution rate of a given site may change over time (Lopez et al. 2002). Heterotachy is particularly plausible in the case of the *PADI* gene tree because *PADI* is found in species across the tree of life (animals, fungi, cyanobacteria, actinobacteria). This could be detected if the tree topology was found to vary under different models of rate variation.

In order to analyse whether our phylogenetic tree may be subject to model violation, we undertook more parameter-rich analyses on 50 sequences that were subsampled from the larger tree, and assessed their topological congruence and node support. We removed multiple paralogues in metazoa using the basal paralogue PADI2 and removed sequences with close branches such that we were able to maintain the maximum sequence diversity in the tree (9 fungi, 13 metazoa, 29 bacteria). We then performed the same fixed empirical rate matrix phylogenetic analysis on the smaller set of sequences to check for congruence, before undertaking a number of phylogenetic analyses (Figure 1 and Supplementary File 6). This included a Bayesian approach that samples over different fixed empirical rate matrices (Ronquist et al. 2012); a maximum likelihood approach using a mixture model of 20 different fixed amino acid rate matrices (C20) (Quang et al. 2008); a Bayesian approach that allows for infinite mixture model categories sampled from the alignment by making use of a Dirichlet process prior (CAT-GTR) (Lartillot and Philippe 2004); and a maximum likelihood approach, designed specifically for heterotachous datasets, that allows different branch length classes across the tree (GHOST model) (Crotty et al. 2020). In addition, we produced maximum likelihood trees where eukaryotic sequences were constrained to be monophyletic under the best performing models (Trees 8 and 9, Figure 1b and Supplementary File 7).

All of the above analyses recovered a single topology in support of a clade of cyanobacterial and animal sequences to the exclusion of a clade of fungal and actinobacterial sequences (Figure 1a, clades Ai, Aii, Bi and Bii). Posterior probabilities or bootstrap values for this topology were high, approaching 100% for each of the diverse methods (Figure 1b). The analysis was repeated using additional bootstrap algorithms, including the full non-parametric bootstrap, obtaining full support (Felsenstein 1985; Hoang et al. 2018). Topology constraint tests rejected a number of randomly generated trees, which confirmed the high branch support values. These alternative trees and the constrained trees for the expected model where eukaryotic PADIs are restricted to a monophyletic group were all significantly rejected ($p<0.001$) by multiple statistical tests including the AU-test (Shimodaira 2002; Strimmer and Rambaut 2002; Susko 2014) (Figure 1b). We conclude that the topology of a clade of cyanobacterial and animal PADI sequences to the exclusion of fungal and actinobacterial sequences is robust to differently specified models.

Cyanobacterial and animal PADIs share unique synapomorphies

The high bootstrap values and congruent topologies across a wide variety of methods lend strong support to our tree topology. Nevertheless, we sought to identify features of the protein sequence that may independently validate the phylogenetic topology.

Firstly, we examined how the PADI protein domain architecture is distributed across orthologues using Pfam annotations, which are powered by HMMER searches (Finn et al. 2015). As mentioned above, all metazoan PADIs possess the three PADI domains, PAD_N, PAD_M and PAD_C (Figure S1). The cyanobacterial PADIs closest to mammalian PADIs (from *SPM* and *NX* cyanobacteria) appear to possess two Pfam-annotated domains: a PAD_M domain and a PAD_C domain, but not a PAD_N domain. By contrast, other bacterial and fungal PADIs are only annotated with the PAD_C domain. To identify domains that might have been overlooked by Pfam, we carried out more sensitive profile-to-profile HMM searches (Söding 2005; Zimmermann et al. 2018) (Figure S6a). We made a multiple sequence alignment firstly of cyanobacterial species contained in the clade of metazoan sequences (Figure 1a, Clade Ai), and secondly of the remaining bacterial and fungal sequences (Figure 1a, sequences outside of Clade Aii). Regions corresponding to each of the PAD_N, PAD_M and PAD_C domains from human PADI2 were extracted and searched against a database of profiles of all domains contained in Pfam. This revealed that the bacterial and fungal sequences outside Clade Aii possess a divergent version of the PAD_M domain, but do not possess a PAD_N domain: the PAD_N region is completely absent from those fungal and bacterial orthologues, including cyanobacteria diverging earlier than *SPM/NX*. By contrast, the cyanobacterial homologues contained within Clade Ai (diverging after *SPM* and *NX* clades) possess all three domains including a degenerate metazoan PAD_N cupredoxin type domain (PAD_N domain: E-value<$1x10^{-7}$). We then identified the cyanobacterial sequence that is predicted to adopt the PAD_N secondary structure using PsiPred and aligned this with animal PAD_N sequences. The predicted cyanobacterial PAD_N sequence aligns well with the human PAD_N domain, as determined experimentally using PADI2 crystal structure data (Slade et al. 2015) (Figure 2a), confirming that the Clade Ai cyanobacterial PADIs possess a degenerate PAD_N domain.

Secondly, we analysed representative fungal, actinobacterial, cyanobacterial, and metazoan PADI sequences for the conservation of calcium-binding and active site residues (Figure 2b). The allosteric binding of up to six calcium ions allows formation of the PADI2 active site cleft and is an absolute requirement for catalytic activity (Slade et al. 2015). All catalytic residues and substrate binding residues are fully conserved among all PADI homologues (Figure 2b). In addition, calcium-binding sites 3 and 1 appear to be fully conserved, while calcium site 5 is also likely conserved. Calcium binding site 6 is likely to be conserved functionally, as the substitution of D125 to N and E131 to D, which are present in both actinobacterial and fungal sequences, are expected to preserve ion binding. Intriguingly, however, calcium sites 2 and 4 appear to be exclusive to Clade Ai (late diverging cyanobacterial and metazoan) sequences. The fungal and actinobacterial sequences diverge from binding sites 2 and 4 to a different amino acid motif. Critically, only Clade Ai PADI sequences conserve the calcium switch residue D389 (residues: 369-389). In actinobacterial and fungal sequences

this residue is substituted to glycine and therefore incompetent for metal coordination (Slade et al. 2015) (Figure 2b). This indicates that the ordered, sequential calcium binding in the PAD_M domain, which is responsible for the allosteric communication between PAD_M and the catalytic PAD_C domain in human PADI2 (Slade et al. 2015) is likely to be conserved only in Clade Ai PADIs. As a result, a potentially different mode of calcium regulation operates in the fungal and actinobacterial PADIs.

Additionally, we find that fungal and actinobacterial sequences share features that are not present in the Clade Ai PADIs. This includes a conserved region within calcium binding sites 3-5 that is absent from the metazoan and cyanobacterial sequences (Figure 2b: amino acids 155-180, where differences conserved between fungal and actinobacterial sequences are highlighted in yellow). Also of interest is a highly conserved 10 amino acid beta sheet that connects the PAD_M and PAD_C domains (Figure 2b: amino acids 292-302). This region is conserved closely in fungal and actinobacterial sequences, but to a different 10 amino acid sequence containing a distinctive triple histidine motif (Figure 2b: amino acids 300-302).

We therefore find primary and tertiary amino acid sequences that are specific to either the cyanobacterial/metazoan or the actinobacterial/fungal PADIs. It is implausible that blocks of sequence of up to ten amino acids were derived convergently and independently in these two groups of PADIs. Thus these sequence features are indicative of a common ancestry of actinobacterial and fungal PADIs that is distinct from the ancestry of cyanobacterial and metazoan PADIs and constitute synapomorphies. The phylogenetic topology presented in Fig. 1 is consistent whether built with or without the above synapomorphic sequence features and PAD_N domain (Figure S6b). As these features occur at the level of the amino acid sequence and at the level of a whole protein domain (Figure 2c), they are robust to differences in rate variation across the tree and to saturated sequence artefacts (Doolittle 1994; Zhang and Kumar 1997; Bazykin et al. 2007; Baalsrud et al. 2018). These features therefore provide strong additional support of the phylogenetic topology presented in Figure 1.

The PADI sequence divergence between cyanobacteria and animals is anachronistically low

The remarkably high similarity of Clade Ai cyanobacterial and animal PADIs prompted us to examine the rate of sequence change between them in more detail. To do this, we firstly sought to understand the extent of change of PADIs relative to other highly conserved proteins in the species that bridge the closest PADI homologues. We therefore analysed a large number of the most conserved proteins in life to approximate a mean minimum extent of accumulated genetic divergence (AGD), represented by sequence change, occurring between *Cyanothece*

sp. 8801 and *Branchiostoma belcheri* and compared this to the divergence of the PADI sequence between these two species (Figure 3a). As a negative control, we compared the difference in bitscore density (Δbitscore) for these conserved proteins and for the PADI sequence between *Branchiostoma belcheri* and *Homo sapiens* (Figure 3b). We also analysed 19 proteins of likely endosymbiont gene transfer (EGT) origin and 10 proteins encoded in mitochondrial genomes. The approach used to calculate the AGD of a given protein between its homologues in *Homo sapiens*, *Branchiostoma* and *Cyanothece* is described in Figure S7. Since mitochondrial and EGT-derived proteins in metazoa may be closer to their bacterial homologues than might be expected for other vertically inherited genes, the accumulated genetic divergence for these classes of protein may be even lower than the AGD for highly conserved ribosomal proteins. We reasoned that the accumulated genetic divergence of mitochondrial and EGT-derived proteins may therefore mimic that of an anciently horizontally transferred gene into eukaryotes and the AGD calculated for PADIs may be even lower than these classes of proteins if it was acquired more recently than the mitochondrion (as we hypothesise for the PADI gene). Indeed, EGT and mitochondrially encoded proteins have an average AGD that is significantly lower than that of vertically acquired proteins between *Cyanothece sp. 8801* and *Branchiostoma belcheri* (Figure 3a), but not between *Branchiostoma belcheri* and *Homo sapiens* (Figure 3b). We find that the AGD of PADI falls below that calculated for vertically transferred protein sequences, as assessed over the same timescale (Figure 3a), falling 6 standard deviations below that of vertically transferred protein sequences, but behaves as expected between *Branchiostoma belcheri* and *Homo sapiens* (Figure 3b). PADIs show less sequence change than all proteins individually analysed over this timescale and less even than ribosomal RNA (Methods). Indeed, they fall 2 standard deviations below even the mean of EGT candidate genes (Figure 3a). Finally, we calculated the AGD for each mitochondrially encoded protein as compared to its own closest bacterial homologue (as opposed to the homologue from *Cyanothece sp 8801*). PADIs exhibit a lower AGD than any of the individual mitochondrially encoded proteins relative to each of their nearest bacterial homologues. With a p value of 0.0073 (*see* Methods), we reject the null hypothesis that PADIs fall within the normal distribution of AGD values calculated for mitochondrially encoded proteins relative to their closest bacterial homologue. A model of vertical descent of PADIs from bacteria, or PADI acquisition via EGT, requires that, across lineages where PADIs cannot be observed in modern genomes, in addition to the large number of independent gene losses, PADIs would have been under greater constraint than any other known sequence in life (Isenbarger et al. 2008).

We then used a Bayesian phylogenetic approach to predict the divergence time between Ai Clade cyanobacterial and animal PADI sequences under a strict molecular clock model and under an uncorrelated lognormal (UCLN) relaxed clock model, using known fossil ages of

metazoans as calibrations (Drummond et al. 2006; Drummond and Suchard 2010; Bouckaert et al. 2014). In the relaxed UCLN clock model, distinct rates are given along each branch with rates drawn at random from a lognormal distribution. Under a model of descent from bacteria, or under a model of EGT, these predictions are expected to be at least as old as the last eukaryotic ancestor, since horizontal transfer is known to be common in bacteria and archaea (Betts et al. 2018). In general, the prediction of the divergence time of a node derived from analysis of a single gene would be significantly greater than the global estimate, as evolutionary rates for a single gene may be greater than the minimum in either lineage.

We performed parallel analysis on the median gene from our EGT candidates above (enolase or ENO) to provide an internal comparison for the divergence time predicted by PADI sequences and calibrations from fossil ages. Our analysis yielded an estimate of less than one billion years for the age of the root of the tree as estimated by PADI sequences (Figure 3c-e). Under all approaches, the divergence times were not congruent with the geologically-defined divergence and were found to be 1.7 billion years (strict clock) or 1.3 billion years (UCLN relaxed clock) lower than that predicted by the ENO gene (Figure 3c-e). The upper bound of our divergence times (95% credible interval) was found to be below the lower bound of the range of globally and geographically defined estimates for the date of the LUCA (>3,900 Ma), the date for eukaryogenesis (1,866–1,679 or 1,842–1,210 Ma), and the date of the symbiotic origin of mitochondria (2,053–1,210 Ma). The use of ENO as a control is likely to be conservative as seen from its AGD, which is lower than any individual ribosomal protein (Figure 3a). These divergence time estimates are therefore inconsistent with vertical descent of metazoan PADIs from bacteria or with descent via EGT and are instead consistent with a horizontal acquisition event that is more recent than the acquisition of the mitochondrion by eukarya. The divergence times predicted by these clock models are approximately dated at the time of divergence of the last common ancestor of PADI-harbouring metazoa.

The cyanobacterial PADI protein is catalytically active

Considering the high degree of similarity between Clade Ai cyanobacterial and metazoan PADIs, including all necessary catalytic residues and calcium binding residues, we hypothesised that the ancestral cyanobacterial protein is likely to be catalytically active and calcium dependent (Fig. 4a). To test this, we prepared a recombinant version of the three-domain PADI from *Cyanothece sp. 8801* (here referred to as “cyanoPADI”) and assayed its catalytic activity alongside human PADI4. Analogously to the human enzyme, cyanoPADI can citrullinate multiple proteins in mouse cell lysates (Fig. 4b). In addition, cyanoPADI shows absolute dependence on calcium for activity. This demonstrates that the calcium-dependent

regulation found in mammalian PADIs is also a feature of the ancestral cyanobacterial protein and suggests that the conserved calcium-binding sites, which were used in the evolutionary analysis as signifiers of synapomorphy, are functional (Fig. 2b and Fig. 4). Remarkably, and despite the absence of histones from bacteria, cyanoPADI catalyses citrullination of histone H3 (Fig. 4c), which is a known target of mammalian PADI4. The enzyme is additionally active at a physiologically relevant temperature for cyanobacteria (Fig. 4c). Thus cyanoPADI is a *bona fide* calcium-dependent peptidylarginine deiminase with sufficient similarity or promiscuity to catalyse citrullination of mammalian substrates.

Discussion

It has been hypothesised that very few protein modification types existed in the LUCA and these have been diversified to give rise to the >200 PTMs known today (Beltrao et al. 2013). We sought to map the evolutionary origin of citrullination, which is implicated in the regulation of a variety of physiological and pathological processes in humans. Our analyses of PADI homologues across life reveal the existence of two clearly discernible PADI types: one containing three structural domains and sharing functionally relevant sequence features and one containing two structural domains and divergent sequence features. The taxonomic distribution of these two types of homologues is highly unusual, in that three-domain PADIs are present in animal and late-diverging cyanobacteria, while two-domain PADIs are present in fungi and all other bacteria (Figures 1, 2, S6). This evidence can be reconciled with vertical evolutionary descent if the last eukaryotic common ancestor (LECA) harboured two paralogous *PADI* genes which underwent widespread and mutually exclusive losses throughout evolution: firstly, the three-domain *PADI* present in late-diverging cyanobacteria and metazoa was lost from lineages leading to every other species in life; and secondly, the two-domain *PADI* present in fungi, actinobacteria and proteobacteria must be separately accounted for in independent gene losses in lineages leading to all other species. In lineages that harbour no *PADI,* the two paralogues must have been lost independently (Figure S8). It is notable that no species is observed to possess both PADI types.

The above scenario, although highly unparsimonious, would be supported if rates of *PADI* sequence evolution across a species phylogeny were consistent with respect to geologically defined timings and with genes well known to have been inherited vertically from bacteria or by EGT from the LECA. Our analyses of sequence divergence provide evidence to the contrary. In absolute terms, the similarity of cyanobacterial and branchiostomal PADIs to human PADIs is almost identical: 70.20% vs 70.90% respectively by pairwise amino acid similarity. However, a much greater amount of time has elapsed since the cyanobacterial and human genes have shared a last common ancestor than the genes from the other species pair

(branchiostoma and humans). Even under assumptions of heterotachy, where rates of evolution may differ between different lineages, a minimal amount of nearly neutral genetic divergence nonetheless accumulates over evolutionary timescales in all lineages (Takahata and Kimura 1994; Isenbarger et al. 2008). Under the assumption of vertical descent, the observed *PADI* sequence changes are anachronistically low even compared to the most highly conserved genomic sequences in life, including ribosomal proteins and even EGT candidates and genes encoded in the mitochondrion.

The explanation for the observation of such little sequence change is more mundane under the assumption of horizontal transfer (Figure 5). A HGT event from late-diverging *SPM/NX* clade cyanobacteria to a last common ancestor within the animal lineage, although ancient, would have occurred much more recently than the LUCA and also more recently than the mitochondrion. HGT can therefore fully account for the phylogenetic distribution, as well as the slow rates of evolution observed. The two lines of evidence are complementary and independent. The timing of transfer (neoproterozoic: 1000-542MYA) is consistent with the presence of marine nitrogen fixing cyanobacteria with specialised arginine catabolic pathways (Schriek et al. 2007), and with the emergence of metazoa in the cyanobacterial habitat (Erwin et al. 2011; Yuan et al. 2011; Sánchez-Baracaldo et al. 2014). A second HGT event, from actinobacterial species that are known to be fungal pathogens, most parsimoniously explains the existence of the two-domain fungal PADI (Clade Bi in Figures 1 and S6; Figures 2 and 5). This is consistent with the absence of a *PADI* gene either in eukaryotic species diverging before opisthokonts or in early diverging fungi such as yeast.

Closer examination of *PADI* phylogeny in bacteria provides additional support for HGT and indicates the directionality of horizontal transfer (Figure S9 and S10). Firstly, strong support is found for bacterial PADIs that form an outgroup to both the two-domain and three-domain PADI sequences (Figure S5 and S10). These bacterial outgroup sequences suggest that PADIs were not horizontally acquired by bacteria. Secondly, the fact that the metazoan-type three-domain PADI only emerges in the late-diverging *SPM* and *NX* clades of cyanobacteria, and the cyanobacterial PADI phylogeny mirrors the expected species tree (Uyeda et al. 2016) (Figure S9), indicates that the three-domain PADI did not exist in the LUCA. The existence of cyanobacterial outgroup sequences, with a discernable origin within bacterial evolution, specifically implies the direction of HGT of the three-domain PADI was from cyanobacteria into metazoa and not in reverse (Figure S9).

All but one metazoan PADI sequence identified by our comprehensive searches in genomic and proteomic databases were found in deuterostomes – the exception being found in the *Priapulus caudatus* genome, a protostome. This suggests that the HGT took place either at

the root of the deuterostomes, or possibly at the root of bilateria. Note that this part of the tree of life remains poorly resolved, with an extremely short branch between the bilaterian common ancestor and the deuterostomes (Philippe et al. 2019).

Biochemical analyses of the ancestral three-domain PADI (cyanoPADI) show that it is competent for catalysis (Figure 4), while a recent study has identified catalytically active PADI homologues in the thermotolerant fungi *Emericella dentata* and *Aspergillus nidulans* (El-Sayed et al. 2019). The discovery of catalytically active PADI orthologues in bacteria and fungi offers fertile ground for investigation of the roles of citrullination in these organisms.

Our finding that the cyanoPADI can citrullinate mammalian substrates (Figure 4) indicates that a novel catalytic capability was added to the regulatory repertoire of metazoan cells by HGT. The newly acquired regulatory function is likely to have enhanced biochemical diversity in animals. Fish genomes contain a single *PADI* gene, but duplications resulted in five tandem repeated paralogues in mammalian genomes (Chavanas et al. 2004) (Figure S10). The fact that these duplicated genes were retained across many animal genomes suggests that they were unlikely to be functionally redundant. In the course of vertebrate evolution, citrullination was thus expanded in scope and adapted to a variety of cellular contexts, ranging from neutrophil extracellular trap release to stem cell potency, and from oligodendrocytes to bone marrow and keratinocytes (Nicholas and Bhattacharya 2014). The emerging physiological roles of the vertebrate PADIs, such as in the regulation of pluripotency and embryonic development (Brahmajosyula and Miyake 2013; Christophorou et al. 2014; Xu et al. 2016; Xiao et al. 2017), and the newly described role of the fish PADI in tissue regeneration (Golenberg et al. 2020), point to possible selective advantages conferred to metazoans by PADIs and offer a possible explanation for the fact that PADIs were retained so widely (Huang 2013). In a similar vein, it is interesting to consider our findings in light of the proposal that genes with a role in antimicrobial defence are amenable to co-option by eukaryotic innate immune systems (Chou et al. 2015). The extent to which the molecular mechanisms that regulate the human PADIs were also conserved from cyanobacteria or were newly co-opted in vertebrates remains an intriguing open question.

It is notable that no citrullination-reversing enzyme has been identified in any species to date. The evolutionary analysis of PADIs presented here adds extra complexity as to whether the reverse catalytic process might have also arisen or been propagated. It has been postulated that “toolkits” of PTM writer, eraser and reader enzymes may have evolved in a coordinated fashion and this has been studied formally in the context of protein phosphorylation (Lim and Pawson 2010). In this context, the investigation into potential reverse catalysis for citrullination should be extended to include bacterial and fungal enzymes.

A related consideration is prompted by the known role of PADIs in autoimmunity. It has been proposed that the exogenous citrullinating activity of pPAD at sites of periodontal infection is an initiating event in the development of rheumatoid arthritis, by predisposing individuals with prior periodontal infection to the development of autoantibodies against citrullinated endogenous proteins (Anti-Citrullinated Protein Antibodies, ACPAs) (Mikuls et al. 2012). It is therefore of note that pPAD and gADI genes are more widespread than previously thought (Figure S4) and that the PADIs described in this paper can be found in a number of human pathogens and in *Stachybotrys chlorohalonata* (black mold). A re-evaluation of the initiating events responsible for citrullination-specific breaks in immune tolerance may therefore be warranted.

This work reveals the remarkable evolutionary trajectory of the *PADI* gene family and uncovers the origin of a protein modification with diverse functions in human physiology and disease. In combination, the pieces of evidence presented above comprise a compelling case of ancient horizontal transfer of a bacterial gene into animals.

Author contributions

TFMC and MAC conceived the idea for the project and wrote the manuscript. TFMC performed phylogenetic, conservation, domain architecture, time divergence, and structural analyses analyses. KG performed phylogenetic and conservation analyses. LS-P performed structure-informed multiple sequence alignments. ARW generated the vector for expression of recombinant cyanoPADI. TFMC and GG performed protein expression and purification and carried out biochemical assays. CD and DM advised on aspects of taxonomy and phylogeny. CPP advised on aspects of structural and evolutionary biology. CD and CPP helped edit the manuscript.

Acknowledgements

This work was funded by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 105642/A/14/Z) and a Medical Research Council/University of Edinburgh Chancellor's Fellowship to MAC. CPP and LS-P were funded by the Medical Research Council (MC_UU_00007/15). DM and CD were funded by Swiss National Science Foundation Grant 183723. We thank M. Reijns for the gift of the pGEX-His plasmid, and G. Abrusán, G. Slodkowicz, N.D. Hastie, B.W. Turner and members of the Christophorou laboratory for critical discussions of the work.

References

Arita K, Hashimoto H, Shimizu T, Nakashima K, Yamada M, Sato M. 2004. Structural basis for Ca2+-induced activation of human PAD4. Nat Struct Mol Biol. 11(8):777–783. doi:10.1038/nsmb799.

Baalsrud HT, Tørresen OK, Solbakken MH, Salzburger W, Hanel R, Jakobsen KS, Jentoft S. 2018. De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. Mol Biol Evol. doi:10.1093/molbev/msx311.

Balandraud N, Gouret P, Danchin EGJ, Blanc M, Zinn D, Roudier J, Pontarotti P. 2005. A rigorous method for multigenic families' functional annotation: The peptidyl arginine deiminase (PADs) proteins family example. BMC Genomics. doi:10.1186/1471-2164-6-153.

Bazykin GA, Kondrashov FA, Brudno M, Poliakov A, Dubchak I, Kondrashov AS. 2007. Extensive parallelism in protein evolution. Biol Direct. doi:10.1186/1745-6150-2-20.

Beltrao P, Bork P, Krogan NJ, Van Noort V. 2013. Evolution and functional cross-talk of protein post-translational modifications. Mol Syst Biol. doi:10.1002/msb.201304521.

Betts HC, Puttick MN, Clark JW, Williams TA, Donoghue PCJ, Pisani D. 2018. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nat Ecol Evol. doi:10.1038/s41559-018-0644-x.

Boto L. 2014. Horizontal gene transfer in the acquisition of novel traits by metazoans. Proc R Soc B Biol Sci. doi:10.1098/rspb.2013.2450.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol. doi:10.1371/journal.pcbi.1003537.

Brahmajosyula M, Miyake M. 2013. Role of peptidylarginine deiminase 4 (PAD4) in pig parthenogenetic preimplantation embryonic development. Zygote. doi:10.1017/S0967199412000160.

Chavanas S, Méchin MC, Takahara H, Kawada A, Nachat R, Serre G, Simon M. 2004. Comparative analysis of the mouse and human peptidylarginine deiminase gene clusters reveals highly conserved non-coding segments and a new human gene, PADI6. Gene. doi:10.1016/j.gene.2003.12.038.

Chou S, Daugherty MD, Peterson SB, Biboy J, Yang Y, Jutras BL, Fritz-Laylin LK, Ferrin MA, Harding BN, Jacobs-Wagner C, et al. 2015. Transferred interbacterial antagonism genes augment eukaryotic innate immune function. Nature. doi:10.1038/nature13965.

Christophorou M a, Castelo-Branco G, Halley-Stott RP, Oliveira CS, Loos R, Radzisheuskaya A, Mowen K a, Bertone P, Silva JCR, Zernicka-Goetz M, et al. 2014. Citrullination regulates pluripotency and histone H1 binding to chromatin. Nature. 507(7490):104–8. doi:10.1038/nature12942. http://www.ncbi.nlm.nih.gov/pubmed/24463520.

Christophorou MA, Castelo-Branco G, Halley-Stott RP, Oliveira CS, Loos R, Radzisheuskaya A, Mowen KA, Bertone P, Silva JCR, Zernicka-Goetz M, et al. 2014. Citrullination regulates pluripotency and histone H1 binding to chromatin. Nature. 507(7490):104–108. doi:10.1038/nature12942.

Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. 2015. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. Genome Biol. doi:10.1186/s13059-015-0607-3.

Crotty SM, Minh BQ, Bean NG, Holland BR, Tuke J, Jermiin LS, Haeseler A Von. 2020. GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments. Syst Biol. doi:10.1093/sysbio/syz051.

Doolittle RF. 1994. Convergent evolution: the need to be explicit. Trends Biochem Sci. doi:10.1016/0968-0004(94)90167-8.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. doi:10.1371/journal.pbio.0040088.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biol. doi:10.1186/1741-7007-8-114.

Dunning Hotopp JC. 2011. Horizontal gene transfer between bacteria and animals. Trends Genet. doi:10.1016/j.tig.2011.01.005.

Dunning Hotopp JC. 2018. Grafting or pruning in the animal tree: Lateral gene transfer and gene loss? BMC Genomics. doi:10.1186/s12864-018-4832-5.

El-Sayed ASA, Shindia AA, AbouZaid AA, Yassin AM, Ali GS, Sitohy MZ. 2019. Biochemical characterization of peptidylarginine deiminase-like orthologs from thermotolerant Emericella dentata and Aspergillus nidulans. Enzyme Microb Technol. doi:10.1016/j.enzmictec.2019.02.004.

Erwin DH, Laflamme M, Tweedt SM, Sperling EA, Pisani D, Peterson KJ. 2011. The Cambrian conundrum: Early divergence and later ecological success in the early history of animals. Science. doi:10.1126/science.1206375.

Falcão AM, Meijer M, Scaglione A, Rinwa P, Agirre E, Liang J, Larsen SC, Heskol A, Frawley R, Klingener M, et al. 2019. PAD2-Mediated Citrullination Contributes to Efficient Oligodendrocyte Differentiation and Myelination. Cell Rep. doi:10.1016/j.celrep.2019.03.108.

Felsenstein J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution (N Y). doi:10.2307/2408678.

Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR. 2015. HMMER web server: 2015 Update. Nucleic Acids Res. doi:10.1093/nar/gkv397.

Gladyshev EA, Meselson M, Arkhipova IR. 2008. Massive horizontal gene transfer in bdelloid rotifers. Science. doi:10.1126/science.1156407.

Golenberg N, Squirrell JM, Bennin DA, Rindy J, Pistono PE, Eliceiri KW, Shelef MA, Kang J, Huttenlocher A. 2020. Citrullination regulates wound responses and tissue regeneration in zebrafish. J Cell Biol. doi:10.1083/jcb.201908164.

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. Sci Rep. doi:10.1038/srep11969.

Guo Q, Fast W. 2011. Citrullination of Inhibitor of Growth 4 (ING4) by Peptidylarginine Deminase 4 (PAD4) disrupts the interaction between ING4 and p53. J Biol Chem. 286(19):17069–17078. doi:10.1074/jbc.M111.230961.

Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the ultrafast bootstrap approximation. Mol Biol Evol. doi:10.1093/molbev/msx281.

Hochstrasser M. 2009. Origin and function of ubiquitin-like proteins. Nature. doi:10.1038/nature07958.

Huang J. 2013. Horizontal gene transfer in eukaryotes: The weak-link model. BioEssays. doi:10.1002/bies.201300007.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, et al. 2016. EGGNOG 4.5: A hierarchical orthology framework with

improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. doi:10.1093/nar/gkv1248.

Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. Nat Rev Microbiol. doi:10.1038/nrmicro.2017.137.

Isenbarger TA, Carr CE, Johnson SS, Finney M, Church GM, Gilbert W, Zuber MT, Ruvkun G. 2008. The most conserved genome segments for life detection on earth and other planets. Orig Life Evol Biosph. doi:10.1007/s11084-008-9148-z.

Iyer LM, Burroughs AM, Aravind L. 2008. Unraveling the biochemistry and provenance of pupylation: A prokaryotic analog of ubiquitination. Biol Direct. doi:10.1186/1745-6150-3-45.

Jensen L, Grant JR, Laughinghouse HD, Katz LA. 2016. Assessing the effects of a sequestered germline on interdomain lateral gene transfer in Metazoa. Evolution. doi:10.1111/evo.12935.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics. doi:10.1093/bioinformatics/8.3.275.

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods. doi:10.1038/nmeth.4285.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. doi:10.1038/nrg2386.

Koonin E V. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. Genome Biol. doi:10.1186/gb-2010-11-5-209.

Koonin E V., Makarova KS, Aravind L. 2001. Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. Annu Rev Microbiol. doi:10.1146/annurev.micro.55.1.709.

Lacroix B, Citovsky V. 2016. Transfer of DNA from bacteria to eukaryotes. MBio. doi:10.1128/mBio.00863-16.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. doi:10.1093/molbev/msh112.

Leger MM, Eme L, Stairs CW, Roger AJ. 2018. Demystifying Eukaryote Lateral Gene Transfer. BioEssays. doi:10.1002/bies.201700242.

Lim WA, Pawson T. 2010. Phosphotyrosine Signaling: Evolving a New Cellular Communication System. Cell. doi:10.1016/j.cell.2010.08.023.

Linsky T, Fast W. 2010. Mechanistic similarity and diversity among the guanidine-modifying members of the pentein superfamily. Biochim Biophys Acta - Proteins Proteomics. doi:10.1016/j.bbapap.2010.07.016.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol. doi:10.1093/oxfordjournals.molbev.a003973.

Macek B, Forchhammer K, Hardouin J, Weber-Ban E, Grangeasse C, Mijakovic I. 2019. Protein post-translational modifications in bacteria. Nat Rev Microbiol. doi:10.1038/s41579-019-0243-0.

Martin WF. 2017. Too Much Eukaryote LGT. BioEssays. doi:10.1002/bies.201700115.

Mikuls TR, Thiele GM, Deane KD, Payne JB, O'Dell JR, Yu F, Sayles H, Weisman MH, Gregersen PK, Buckner JH, et al. 2012. Porphyromonas gingivalis and disease-related autoantibodies in individuals at increased risk of rheumatoid arthritis. Arthritis Rheum. doi:10.1002/art.34595.

Moran NA, Jarvik T. 2010. Lateral transfer of genes from fungi underlies carotenoid production

in aphids. Science. doi:10.1126/science.1187113.

Musse AA, Li Z, Ackerley CA, Bienzle D, Lei H, Poma R, Harauz G, Moscarello MA, Mastronardi FG. 2008. Peptidylarginine deiminase 2 (PAD2) expression in a transgenic mouse leads to specific central nervous system (CNS) myelin instability. Dis Model Mech. 1(4):229–240.

Nicholas AP, Bhattacharya SK. 2014. Protein deimination in human health and disease.

Ochman H, Lawrence JG, Grolsman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. doi:10.1038/35012500.

Pearce MJ, Mintseris J, Ferreyra J, Gygi SP, Darwin KH. 2008. Ubiquitin-like protein involved in the proteasome pathway of Mycobacterium tuberculosis. Science. doi:10.1126/science.1163885.

Philippe H, Poustka AJ, Chiodin M, Hoff KJ, Dessimoz C, Tomiczek B, Schiffer PH, Müller S, Domman D, Horn M, et al. 2019. Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria. Curr Biol. doi:10.1016/j.cub.2019.04.009.

Quang LS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. doi:10.1093/bioinformatics/btn445.

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol. doi:10.1093/sysbio/sys029.

Salzberg SL. 2017. Horizontal gene transfer is not a hallmark of the human genome. Genome Biol. doi:10.1186/s13059-017-1214-2.

Sánchez-Baracaldo P, Ridgwell A, Raven JA. 2014. A neoproterozoic transition in the marine nitrogen cycle. Curr Biol. doi:10.1016/j.cub.2014.01.041.

Schriek S, Rückert C, Staiger D, Pistorius EK, Michel KP. 2007. Bioinformatic evaluation of L-arginine catabolic pathways in 24 cyanobacteria and transcriptional analysis of genes encoding enzymes of L-arginine catabolism in the cyanobacterium Synechocystis sp. PCC 6803. BMC Genomics. doi:10.1186/1471-2164-8-437.

Sharma P, Lioutas A, Fernandez-Fuentes N, Quilez J, Carbonell-Caballero J, Wright RHG, 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. Mol Cell. 73(1):84-96.e7. doi:10.1016/j.molcel.2018.10.016.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. doi:10.1080/10635150290069913.

Shirai H, Blundell TL, Mizuguchi K. 2001. A novel superfamily of enzymes that catalyze the modification of guanidino groups. Trends Biochem Sci. doi:10.1016/S0968-0004(01)01906-5.

Slade DJ, Fang P, Dreyton CJ, Zhang Y, Fuhrmann J, Rempel D, Bax BD, Coonrod SA, Lewis HD, Guo M, et al. 2015. Protein arginine deiminase 2 binds calcium in an ordered fashion: Implications for inhibitor design. ACS Chem Biol. 10(4):1043–1053. doi:10.1021/cb500933j.

Snijders AP, Hautbergue GM, Bloom A, Williamsom JC, Minshull TC, Phillips HL, Mihaylov SR, Gjerde DT, Hornby DP, Wilson SA, et al. 2015. Arginine methylation and citrullination of splicing factor proline- and glutamine-rich (SFPQ/PSF) regulates its association with mRNA. RNA. 21(3):347–359. doi:10.1261/rna.045138.114.

Söding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics. doi:10.1093/bioinformatics/bti125.

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: Building the web of life. Nat Rev Genet. doi:10.1038/nrg3962.

Stadler SC, Vincent CT, Fedorov VD, Patsialou A, Cherrington BD, Wakshlag JJ, Mohanan S, Zee BM, Zhang X, Garcia BA, et al. 2013. Dysregulation of PAD4-mediated citrullination of nuclear GSK3β activates TGF-β signaling and induces epithelialto-mesenchymal transition in breast cancer cells. Proc Natl Acad Sci U S A. doi:10.1073/pnas.1308362110.

Stamatakis A, Kozlov AM, Kozlov A. 2020. Efficient Maximum Likelihood Tree Building Methods. In: Phylogenetics in the Genomic Era.

Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR. 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature. doi:10.1038/35082058.

Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc R Soc B Biol Sci. doi:10.1098/rspb.2001.1862.

Sugawara K, Oikawa Y, Ouchi T. 1982. Identification and properties of peptidylarginine deiminase from rabbit skeletal muscle. J Biochem. doi:10.1093/oxfordjournals.jbchem.a133755.

Susko E. 2014. Tests for two trees using likelihood methods. Mol Biol Evol. doi:10.1093/molbev/msu039.

Suzuki A, Yamada R, Chang X, Tokuhiro S, Sawada T, Suzuki M, Nagasaki M, Nakayama-Hamada M, Kawaida R, Ono M, et al. 2003. Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat Genet. 34(4):395–402. doi:10.1038/ng1206.

Takahata N, Kimura M. 1994. The Neutral Theory of Molecular Evolution. Princ Med Biol. doi:10.1016/B978-1-55938-802-3.50013-4.

Tanikawa C, Ueda K, Nakagawa H, Yoshida N, Nakamura Y, Matsuda K. 2009. Regulation of protein citrullination through p53/PADI4Network in DNA damage response. Cancer Res. 69(22):8761–8769. doi:10.1158/0008-5472.CAN-09-2280.

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 Rep. 22(6):1473–1483. doi:10.1016/j.celrep.2018.01.031.

Touz MC, Rópolo AS, Rivero MR, Vranych CV, Conrad JT, Svard SG, Nash TE. 2008. Arginine deiminase has multiple regulatory roles in the biology of Giardia lamblia. J Cell Sci. doi:10.1242/jcs.026963.

Uyeda JC, Harmon LJ, Blank CE. 2016. A comprehensive study of cyanobacterial morphological and ecological evolutionary dynamics through deep geologic time. PLoS One. doi:10.1371/journal.pone.0162539.

Wang S, Wang Y. 2013. Peptidylarginine deiminases in citrullination, gene regulation, health and pathogenesis. Biochim Biophys Acta - Gene Regul Mech. 1829(10):1126–1135. doi:10.1016/j.bbagrm.2013.07.003.

Wang Y, Li M, Stadler S, Correll S, Li P, Wang D, Hayama R, Leonelli L, Han H, Grigoryev SA, et al. 2009. Histone hypercitrullination mediates chromatin decondensation and neutrophil extracellular trap formation. J Cell Biol. doi:10.1083/jcb.200806072.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. doi:10.1093/oxfordjournals.molbev.a003851.

Xiao S, Lu J, Sridhar B, Cao X, Yu P, Zhao T, Chen CC, McDee D, Sloofman L, Wang Y, et al. 2017. SMARCAD1 Contributes to the Regulation of Naive Pluripotency by Interacting with Histone Citrullination. Cell Rep. 18(13):3117–3128. doi:10.1016/j.celrep.2017.02.070.

Xu Y, Shi Y, Fu J, Yu M, Feng R, Sang Q, Liang B, Chen B, Qu R, Li B, et al. 2016. Mutations in PADI6 Cause Female Infertility Characterized by Early Embryonic Arrest. Am J Hum Genet. doi:10.1016/j.ajhg.2016.06.024.

Yuan X, Chen Z, Xiao S, Zhou C, Hua H. 2011. An early Ediacaran assemblage of macroscopic and morphologically differentiated eukaryotes. Nature. doi:10.1038/nature09810.

Yuzhalin AE, Gordon-Weeks AN, Tognoli ML, Jones K, Markelc B, Konietzny R, Fischer R, Muth A, O'Neill E, Thompson PR, et al. 2018. Colorectal cancer liver metastatic growth depends on PAD4-driven citrullination of the extracellular matrix. Nat Commun. 9(1). doi:10.1038/s41467-018-07306-7.

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol. doi:10.1093/oxfordjournals.molbev.a025789.

Zhang X, Gamble MJ, Stadler S, Cherrington BD, Causey CP, Thompson PR, Roberson MS, Kraus WL, Coonrod SA. 2011. Genome-Wide analysis reveals PADI4 cooperates with Elk-1 to activate C-Fos expression in breast cancer cells. PLoS Genet. 7(6). doi:10.1371/journal.pgen.1002112.

Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. doi:10.1016/j.jmb.2017.12.007.

Figure Legends

Figure 1: Phylogeny of the PADI sequence. a) Consensus topology for all phylogenetic methods with branch lengths from Bayesian phylogenetic inference with MrBayes. Solid circles indicate consensus node support of >95%. **b)** Summary table of the different phylogenetic analyses performed corresponding to trees shown in full in supplementary data. Ultrafast bootstrap 2 values with 1000 replicates for trees 1, 3 , 4, 5, 6, 7; Felsenstein bootstrap values with 100 replicates for tree 2; or posterior probabilities for trees 10 and 11 are presented in the table for the nodes labelled in the tree that are critical to different evolutionary scenarios. Log likelihoods and the Bayesian information criterion are presented for all maximum likelihood trees. In addition, maximum likelihood constraint trees 8 and 9 were constructed where opisthokonta were constrained to be monophyletic under the maximum likelihood models used for tree 1 and tree 5. Trees were concatenated and analysed using the AU-test with 10,000 replicates. Nomenclature for the different models is as used in IQtree 1.6.12. The best supported maximum likelihood tree and the Bayesian trees are shown in bold.

Figure 2: Synapomorphic features among PADI orthologues. a) Alignment of putative PAD_N domains from *SPM/NX* clade cyanobacterial PADI sequences with the PAD_N domain from human PADI paralogues and *Rhincodon typus* (whale shark). The colouring scheme indicates the average BLOSUM62 scores of each alignment column: red (>3.5), violet (between 3.5 and 2) and light yellow (between 2 and 0.5). Peach arrows shown below the cyanobacterial sequences indicate PsiPred predicted secondary structure (beta sheets). Green arrows (beta sheets) correspond to the known secondary structure of the PAD_N domain of human PADI2. **b)** Analysis of synapomorphic regions, representing six PADI sequences from each of metazoa, cyanobacteria, actinobacteria and fungi. Consensus sites across the six species are shown with standard single letter amino acid abbreviations. “nc” (non-nonserved) represents the absence of consensus conservation to one or two amino acids across the six species. The numbering given above the alignment and corresponds to the ungapped site of human PADI2 such that residues can be compared to Slade *et al.* Sites showing conservation across all four domains are coloured in green; sequence features common to metazoan and cyanobacterial PADIs that are excluded from fungal/actinobacterial sequences are coloured in purple; sequence features common to fungal and actinobacterial PADIs that are excluded from metazoanand cyanobacterial sequences are coloured in yellow. The existence of both purple and yellow sequence features is indicative of synapomorphic primary sequence features. **c)** Crystal structure of human PADI2 presented with PAD_N domain coloured in black, PAD_M domain in grey and PAD_C domain in white. Synapomorphic regions are coloured in cyan and calcium ions are shown as yellow spheres.

Figure 3: Sequence divergence analyses. a,b) Analysis of the sequence divergence of 26 vertically transferred proteins, 19 candidate EGT proteins, and 10 proteins encoded in the mitochondrial genome. **a)** Box and whisker plot showing the calculated accumulated genetic divergence (AGD) between *Cyanothece sp. PCC 8801* and *Branchiostoma floridae* relative to *Homo sapiens*. **b)** Box and whisker plot showing the normalised Δbitscore between *Branchiostoma floridae* and *Homo sapiens*. The cross represents the mean. All protein values are plotted with outliers exceeding 1.5X the interquartile range shown. The null hypothesis that PADIs fall within the normal distribution of each set of proteins was rejected with $p<0.0001$ denoted as ***; or $p<0.05$ denoted as *. **c,d)** Estimated divergence time of late diverging *SPM/NX* clade cyanobacteria and metazoa based on their PADI sequences, as calibrated using geologically defined constraints from the fossil record. Metazoan and *SPM/NX* DNA sequences were used for Bayesian phylogenetic analysis in BEAST2 under the strict clock and the uncorrelated lognormal (UCLN) clock models. A calibrated Yule model was used as the tree prior using a GTR model with 5 gamma distributed rate categories. Divergence times from the fossil record were used as normally distributed node age priors centered on the median ages of

six different nodes from metazoa with a sigma value covering the uncertainty of the estimate. The marginal posterior distribution of the age of the root of the whole tree was used to estimate the divergence time. **c)** Box and whisker plot for the estimate divergence time from each analysis showing two independent runs per analysis. **d)** Kernel density estimate for each analysis showing two independent runs per analysis. **e)** Table of summary statistics for the estimated divergence time.

Figure 4: Biochemical analyses of the cyanobacterial PADI enzyme from *Cyanothece sp. 8801* (cyanoPADI). a) The citrullination reaction results in the converison of a positively charged peptidyl arginine residue to a neutral peptidyl citrulline and it is carried out by PADI enzymes in a calcium-dependent manner. **b,c)** Immunoblot analyses of citrullination assays using GST-His-tagged recombinant enzymes. **b)** Whole cell lysates from mouse embryonic stem cells were used as substrate and the presence of citrullination in a protein sequence-independent manner was assessed using the ModCit antibody. Nucleophosmin (NPM1) is used as a loading control. **c)** Recombinant human histone H3 was used as substrate and citrullination of H3 arginine 2 was assessed. Total histone H3 is used as loading control.

Figure 5: Proposed model of PADI evolution. Domain architecture is denoted in the figure legend. Horizontal transfer of the 3-domain sequence from cyanobacteria to metazoa denoted by a black arrow, likely horizontal transfer of the 2-domain sequence from actinobacteria to fungi denoted by a dark gray arrow and transfer of the mitochondrion to the LECA denoted by a light gray arrow. Proposed origin for the PADI sequence is within bacterial evolution and emergence of the 3-domain PADI is within the *SPX/NM* cyanobacterial clade. Gene losses observed in various metazoan lineages after the HGT are indicated with a narrow dashed line.

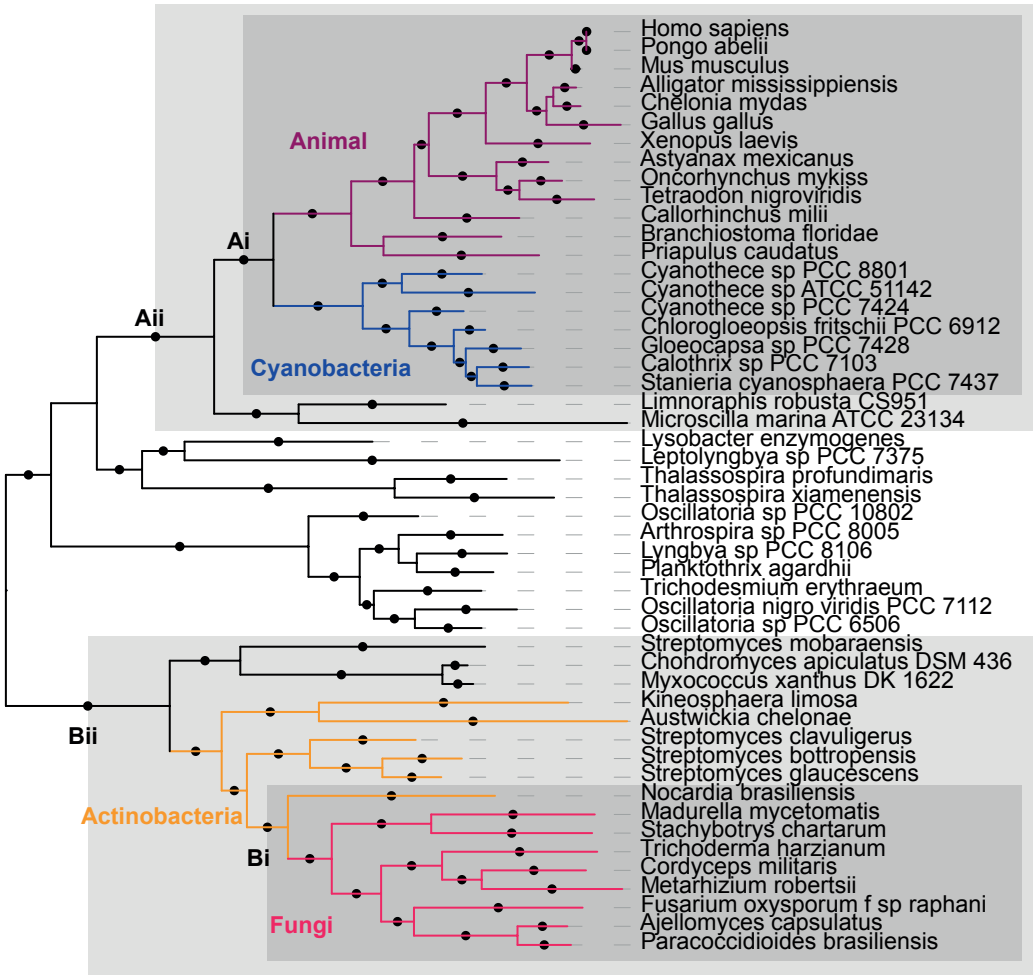
Table 1

Group	NCBI Taxonomy ID	Unique species with a PADI	Species with proteomes in UniprotKB	Percentage of species with a PADI
Bacteria	2	295	38842	0.76
Cyanobacteria	1117	56	506	11.07
Actinobacteria	201174	136	4870	2.79
Proteobacteria	1224	69	16196	0.43
Eukaryotes	2759	406	2241	18.12
Animals (Metazoa)	33208	229	612	37.42
Insects	50557	0	142	0.00
Worms (Annelida)	6340	0	2	0.00
Fungi	4751	177	1098	16.12
Yeast (Ascomycota)	4890	176	760	23.16
Yeast (Saccharomyces)	4930	0	13	0.00
Plants (Viridiplantae)	33090	0	244	0.00
Opisthokonta (metazoa and fungi)	33208 & 4751	406	1710	23.74
Pre-opisthokonta (Eukarya, not metazoa or fungi)	2759 & NOT (33208 | 4751)	0	531	0.00
Archaea	2157	1	2107	0.05
Viruses	10239	1	99210	0.001

Table 1: The number and proportion of species harbouring a putative PADI orthologue. HMM searches (https://www.ebi.ac.uk/Tools/hmmer) for similarity to the vertebrate PAD_C domain from human PADI2, were carried out using HmmerWeb version 2.41.1 against the UniProtKB (v.2019_09) database. Unique species with significant sequence similarity (E-value < $1x10^{-3}$) are presented. Proportions are given relative to the total number of species in within UniProtKB, for each group.

Figure 1

a



b

Tree Number	Inference	Method	Ai	Aii	Bi	Bii	ln(L)	BIC	AU test
1	ML	WAG+R5+FO	99	100	99	100	-51921.3	104711.0	0.463
2	ML	WAG+R5+FO_FB	95	100	91	100	-51923.2	104714.7	0.572
3	ML	LG+R6+FO	98	100	98	100	-51908.9	104700.2	0.462
4	ML	C20+FO	96	100	100	100	-52910.4	106780.2	0.186
5	**ML**	**WAG+F+C20+R5**	**96**	**100**	**100**	**100**	**-51663.5**	**104335.4**	**0.574**
6	ML	WAG+FO+H4	99	100	100	100	-51605.0	106081.3	0.557
7	ML	WAG+FO*H4	94	100	99	100	-51380.8	106032.1	0.462
8	CT by ML	WAG+R5+FO	–	–	–	–	-52362.8	105593.9	<0.001
9	CT by ML	WAG+F+C20+R5	–	–	–	–	-52068.8	105146.1	<0.001
10	**Bayesian**	**MrBayes**	**100**	**100**	**100**	**100**	**–**	**–**	**0.463**
11	**Bayesian**	**PhyloBayes**	**99**	**100**	**100**	**100**	**–**	**–**	**0.486**

Figure 2

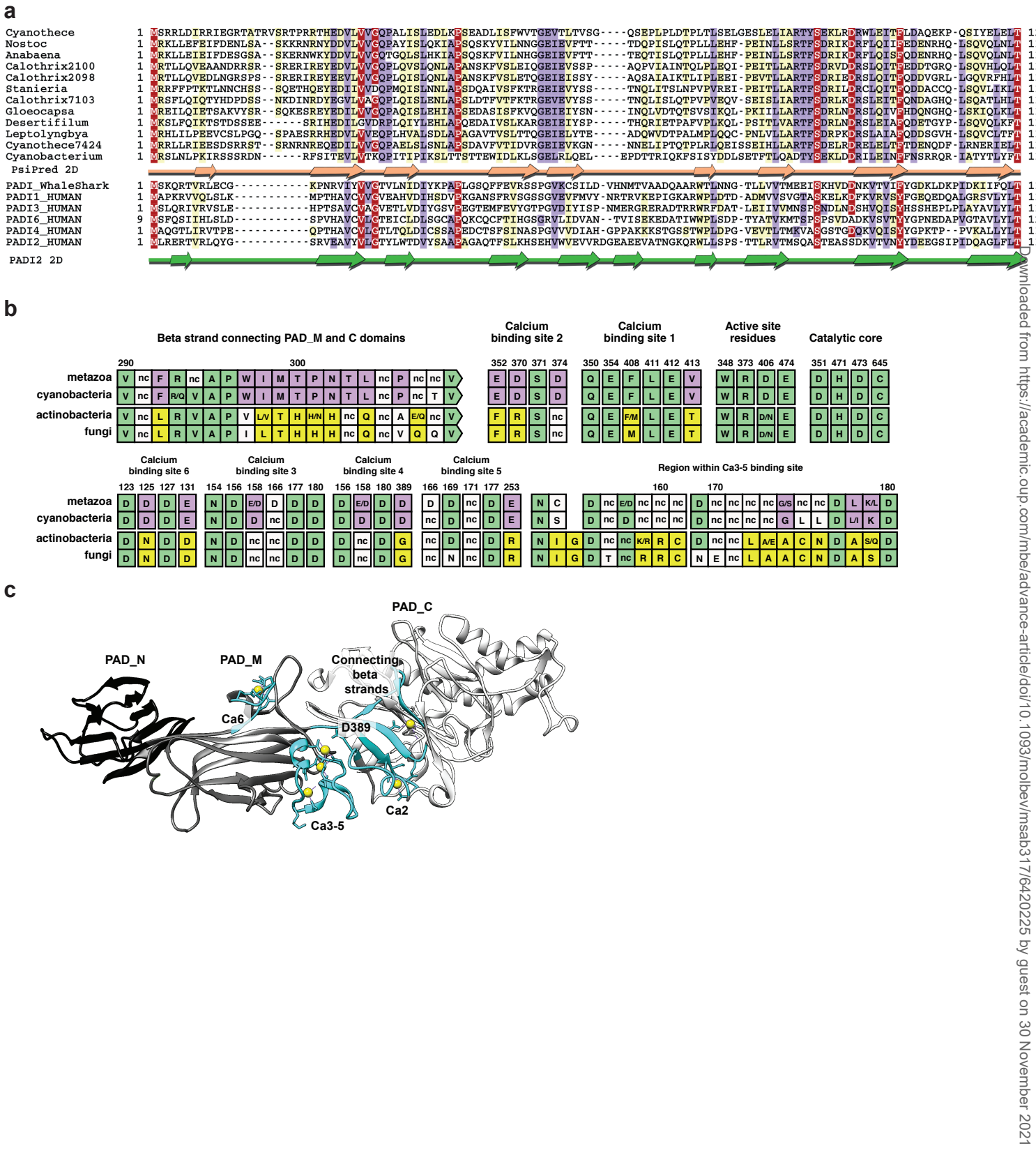
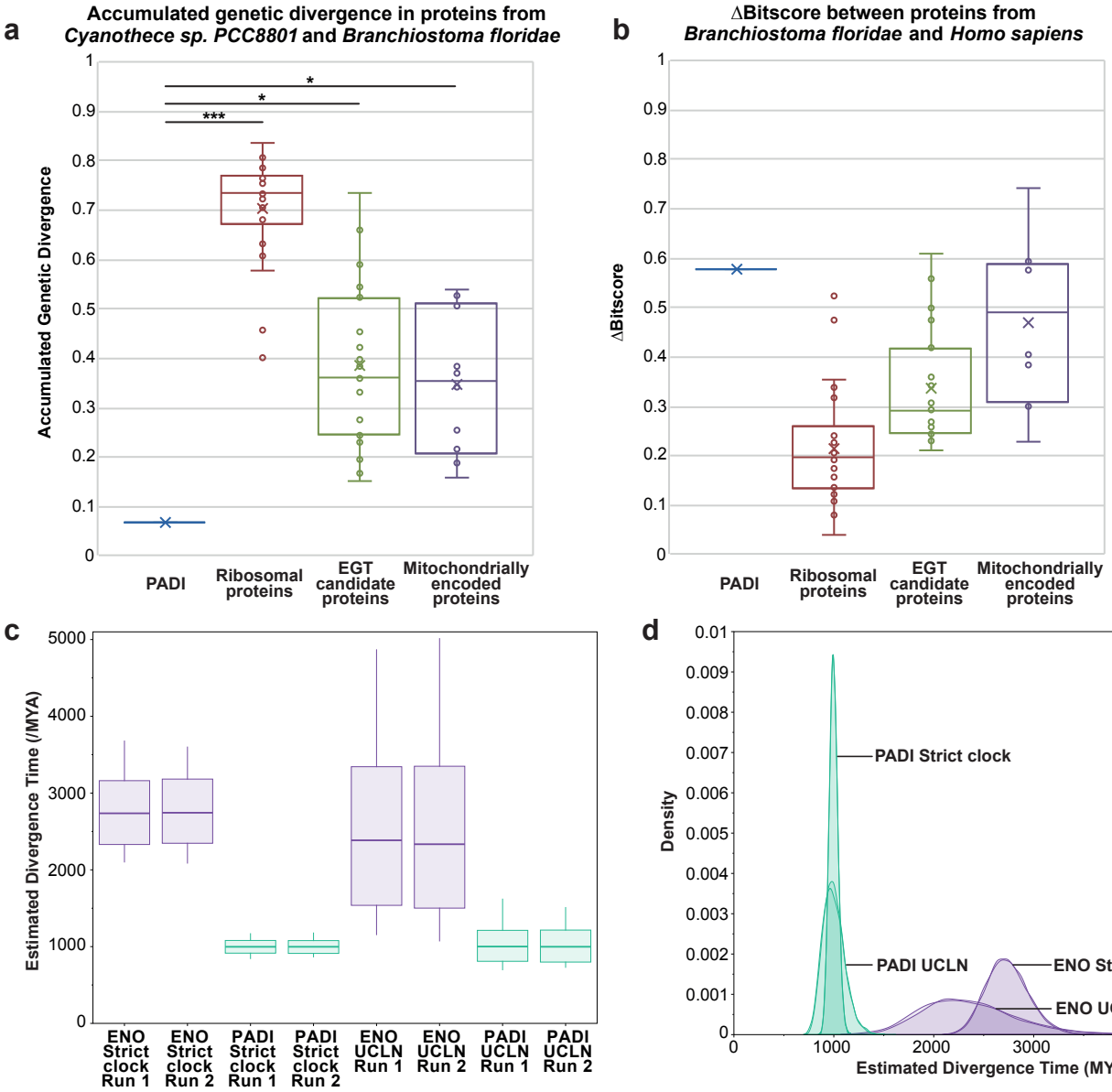


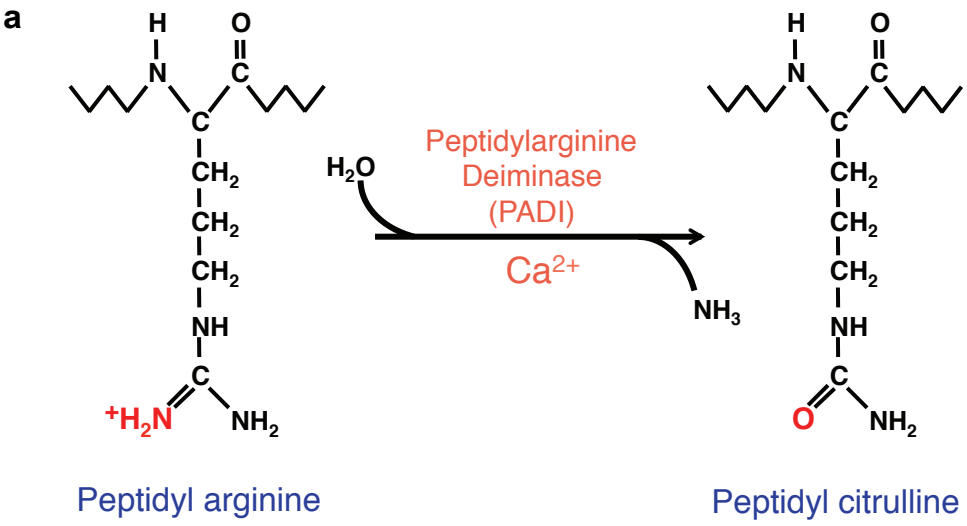
Figure 3

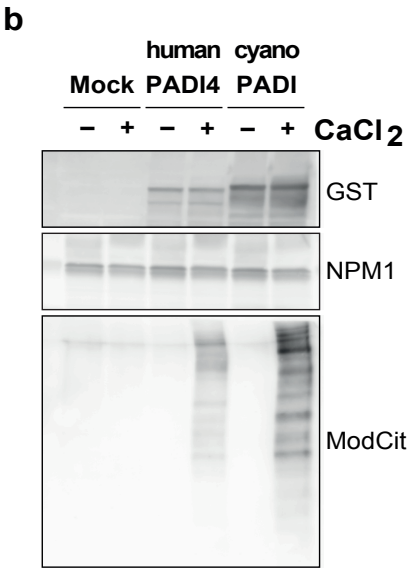


e

	ENO Strict Clock Run 1	ENO Strict Clock Run 2	PADI Strict Clock Run 1	PADI Strict Clock Run 2	ENO UCLN Run 1	ENO UCLN Run 2	PADI UCLN Run 1	PADI UCLN Run 2
Mean	2736.2896	2743.845	997.6094	997.7531	2384.8546	2333.8561	1000.6955	998.1353
Stderr of mean	3.1995	3.1779	0.5399	0.5713	20.8487	22.4966	4.7137	4.7518
Stdev	212.8374	216.4931	42.5362	42.6549	476.3007	480.1344	106.1321	110.6886
Median	2727.0506	2729.1837	996.3539	996.2449	2330.6141	2288.6189	992.2871	987.4645
95% HPD interval (LB)	2330.2552	2346.3226	914.7292	911.5626	1537.3756	1501.2355	808.4079	798.3852
95% HPD interval (HB)	3161.8364	3182.1895	1079.8871	1078.4124	3343.3161	3349.4396	1211.971	1215.4161
Effective sample size (ESS)	4424.6	4640.4	6206.1	5573.4	521.9	455.5	506.9	542.5

Figure 4





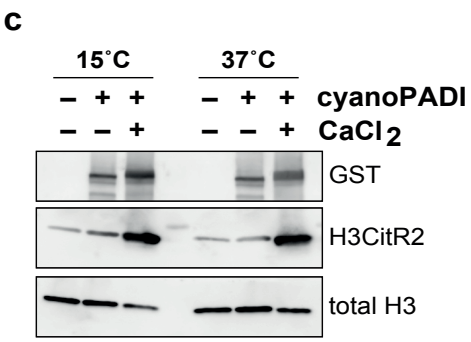


Figure 5