

The evolution of the deubiquitinating enzyme superfamily

by

Caitlyn Vlasschaert

Manuscript-based thesis submitted in partial fulfillment of
the M.Sc. program in Biochemistry specializing in Bioinformatics

Department of Biochemistry and Immunology

Faculty of Medicine

University of Ottawa

July 2016

© Caitlyn Vlasschaert, Ottawa, Canada, 2016

ABSTRACT

Multiple versions of a parent gene can function within molecular systems as gene duplicates (paralogs) and alternatively spliced isoforms. Proteins related in this manner often serve redundant roles, though they can be selectively or randomly prescribed unique functions. The present collection of three manuscripts details the evolution of members of the deubiquitinating enzyme superfamily. The first manuscript delineates the chronology of USP4, USP15 and USP11 emergence and concludes that the presumed ancestor, USP11, is in fact a recent duplicate and that, at minimum, one copy of USP4 or USP15 is required for organismal viability. The second determines that the long and short isoforms of mammalian USP4 are maintained by natural selection to occupy discrete spatial roles. The final manuscript broadens the scope and objectively draws the genealogy of all deubiquitinating enzymes, with emphasis on significant points of functional divergence of paralogs within innate immunity and DNA repair pathways.

ACKNOWLEDGEMENTS

I would first like to thank my supervisors, Dr. Doug Gray and Dr. Xuhua Xia for continuously pushing me throughout my Master's research. Doug, thank you immensely for always believing in me and my far-out ideas. I would not have chosen to defer my plans in order to pursue an M.Sc., or a career in research for that matter, had I been supervised by anyone else. Dr. Xia, thank you for introducing me to the world of molecular evolution and for your constant support.

Thank you to my thesis committee members, lab mates (especially Dr. Josée Coulombe from the Gray lab), and adopted lab mates (David Cook) for your input and for the stimulating discussions, and to my best friends Matthieu Laframboise and Erika Schmitz who cheered me along the whole way.

Finally, thank you to my family. To my dad, for allowing me to drone on about USPs and evolution while intently listening the entire time. To my sister, Chelsea, for grounding me and forcing me to cultivate the essential skill of summarizing my work into a short narrative. To my mom, who is both my greatest support and the one who pushes me to reach further the most. I would like to dedicate this thesis to my late grandfather, Lorne Beaucage, who told me during the first year of my undergraduate degree that he would be present at *all* of my graduation ceremonies and whom I continuously strive to make proud.

TABLE OF CONTENTS

Abstract.....	ii
Acknowledgements.....	iii
Table of Contents.....	iv
List of Abbreviations	v
List of Figures and Tables.....	vi
Introduction.....	1
Statement of Objectives	6
Results.....	7
Chapter I – Evolution of the highly networked deubiquitinating enzymes USP4, USP15 and USP11	7
Chapter II – Selection preserves Ubiquitin Specific Protease 4 alternative exon skipping in therian mammals	66
Chapter III – The evolution and functional diversification of the deubiquitinating enzyme superfamily.....	105
Discussion.....	162
References.....	167
<i>Curriculum vitae</i>	169

LIST OF ABBREVIATIONS

BMP	bone morphogenic protein
DUB	deubiquitinating enzyme
JAMM	JAB1/MPN/Mov34 metalloenzyme
K48	Lysine 48
K63	Lysine 63
Ka	non-synonymous substitutions per non-synonymous site
Ks	synonymous substitutions per synonymous site
M1	Methionine 1
MJD	Machado-Josephin domain
OTU	ovarian tumour domain
SSD	small-scale duplication
TGF- β	transforming growth factor beta
Ub	ubiquitin
UBL	ubiquitin-like
UCH	ubiquitin C-terminal hydrolases
UPS	ubiquitin proteasome system
USP	ubiquitin-specific protease
WGD	whole-genome duplication

LIST OF FIGURES AND TABLES

First Manuscript

Figure 1. Comparison of USP4, USP15 and USP11 features.

Figure 2. Phylogenetic distribution of USP4, USP15 and USP11.

Figure 3. Maximum likelihood reconstruction of aligned codon sequences.

Figure 4. Maximum likelihood reconstruction of recoded amino acid sequences.

Figure 5. Shared synteny of USP4 and USP11 loci in Euteleosts.

Figure 6. Bayesian dating of aligned co-duplicate codon sequences.

Figure 7. Summary of signature features of USP4, USP15 and USP11

Figure 8. Synteny of USP4, USP15, and USP11 in anole compared to chicken and human.

Figure 9. Phylogeny of USP genes.

Table 1. List of coding sequences analyzed with corresponding accession numbers.

Table 2. Summary of USP4, USP15 and USP11 interaction partners.

Table 3. Pooled progeny of USP4- and USP15-null mouse crosses.

Second Manuscript

Figure 1. Schema of *Usp4* exonic properties and potential mechanisms of exon skipping.

Figure 2. Characterization of relative BPS strengths.

Figure 3. Length of USP4 introns for 62 bird, fish and mammal species.

Figure 4. Mismatched flanking 5' splice site strengths leads to alternative skipping of exon 7.

Figure 5. Exon skipping does not occur in the USP4 gene of the chicken or the zebrafish.

Figure 6. Long and short isoform production was selected for in therian mammals.

Table 1. Position weight matrix for USP4 5' splice sites

Third manuscript

Figure 1. Evolutionary expansion of the entire eukaryotic deubiquitinating enzyme superfamily.

Figure 2. Tissue specificity of human deubiquitinating enzymes in an evolutionary context.

Figure EV2. Relative abundance of each DUB in human tissues.

Figure 3. Selectivity of deubiquitinating enzyme expression in gametes and embryogenesis.

Figure 4. Evolutionary network of ubiquitin proteasomal system involvement in innate immunity.

Figure 5. Changing intensity of purifying selection, measured by the ratio of nonsynonymous substitution rate over synonymous substitution rate (K_a/K_s), during the evolution and functional diversification of USP18 and ISG15.

Figure 6. Rewiring of the deubiquitinating enzyme network for new roles in DNA repair.

Figure EV6. Evolution of coiled-coil domains within USP16 and USP45.

INTRODUCTION

The dynamic conjugation and removal of ubiquitin regulates virtually every cellular process. Ubiquitination involves the covalent attachment of ubiquitin (Ub), a 76 amino acid protein, to lysine residues on target proteins via its C-terminus. Several Ub and ubiquitin-like units can subsequently be conjugated onto the first to elongate the Ub chain in a process known as polyubiquitination. Several linkage types exist and they are designated based on which internal lysine (K) or methionine (M) residue serves as point of attachment for the next Ub. Ubiquitination is a versatile type of post-translational modification as its effect is dependent upon linkage type. Effectuating the canonical role of Ub in the ubiquitin-proteasome system (UPS), lysine-48 (K48) linkages mark the target protein for proteasomal degradation. Five other minor lysine linkage types, namely K6, K11, K27, K29 and K33, have also been associated with proteolysis¹. Linkages via lysine-63 (K63) as well as linear chains formed via C-terminal attachments to N-terminal Ub methionine residues (M1)² have strict non-proteolytic roles and instead serve as dynamic modulators of cellular signaling as their conjugation often enables protein-protein interactions, much like phosphorylation. In addition, several ubiquitin-like molecules (UBLs), which earn their name from their structural homology to di-Ub, can also be Ub chains constituents and serve degradative and non-degradative purposes³. Thus, ubiquitin and ubiquitin-like molecules directly control protein stability as well as permit various interactions, depending on the Ub chain topology.

Dynamic conjugation and removal of Ub post-translational modifiers are respectively catalyzed by E3 ubiquitin ligases and deubiquitinating enzymes (DUBs).

These antagonistic enzymes act in concert to maintain cellular dynamic ubiquitination needs. Within the superfamily of deubiquitinating enzymes (defined by their common role), there exists 5 families (defined by their catalytic domain). From largest to smallest in number, they are: the ubiquitin-specific proteases (USPs), the ovarian tumour (OTU) proteases, the JAB1/MPN/Mov34 metalloenzyme (JAMM) motif proteases, the Machado-Josephin domain (MJD) proteases, and the Ubiquitin C-terminal Hydrolase (UCH) proteases. The USPs, OTUs, MJDs and UCHs function as classical cysteine proteases while the JAMM motif DUBs are metalloproteases that use coordinating zinc ions to cleave substrates, which are often not limited or specific to ubiquitin.

DUBs often serve overlapping roles, for example in development, cell cycle progression, DNA repair, RNA splicing, and innate immunity. Several homologous DUBs have shared substrates owing largely to their sequence similarity, while the functions of several closely- and distantly-related enzymes also converge to regulate multiple substrates within a common pathway. While this may be protective in anti-viral immunity, where the UPS is often the target of viral hi-jacking⁴, ubiquitination system enzyme redundancy is observed in many other cellular pathways. The genetic catalogue of DUBs has expanded throughout evolution though remains small compared to that of E3 ubiquitin ligases in both the yeast *Saccharomyces cerevisiae* (24 DUBs to 68 ubiquitin ligases) and human (93 DUBs to more than 600 ligases⁵). Given these ratios, DUB redundancy within the proteomic landscape might not be expected.

Enzymes within a genome that are homologous, termed paralogs, are birthed by duplication. Gene duplication is a very common process. Among all methods of gene

genesis, including *de novo* synthesis and domain chimerization, it creates the most genetic fodder⁶. These duplicates, however, more often than not undergo pseudogenization. Several factors influence whether a paralog is retained as a functional gene that contributes to the cellular proteome, lncRNA-ome, miRNA-ome, etc. These factors influence the strength of selective pressure that is exerted upon the gene and they continuously shape the constituents of the genome.

Natural selection acts at the molecular level in the evolution of genes and cellular systems. The selective forces derive from reproductive biases associated with improved biological fitness. The function of a parent gene influences whether a duplicate is selectively retained, for example if it confers gene dosage benefits^{7,8} or if a minor secondary function can be exploited by the paralog⁶. Additionally, the timing of duplication can have an effect on whether a paralog is retained, since selective pressure for it to occupy node(s) within a cellular network is dependent on the environment into which it was birthed. That is, the utility of a genetic product is subjective to the state of networks in which it is embedded.

Purifying/negative selection preserves sequences while positive selection allows beneficial mutations to become fixed, for example in the case of co-evolution with interaction partners. For protein-coding genes, these pressures can be estimated by Ka/Ks values. The number of non-synonymous substitutions (substitutions that alter the coded amino acid) per non-synonymous site (codon positions where a substitution would alter the amino acid) is quantified by Ka, while Ks measures synonymous substitutions per synonymous site. When Ka is much lower than Ks, the sequence undergoes little change

across evolutionary time and purifying selection is inferred to be strong. While a large K_a could indicate directional positive selection, it can also be reflective of strong mutational bias (for example a gene translocated to a high-AT isochore might undergo greater AT-biased coding sequence change due to its genomic position)⁹. Thus, selective pressure can act to retain coding sequences that improve gene fitness. However, substitutional change is also influenced by genetic drift and biased mutational and recombination rates; a combination of these forces determines whether and in what form genes are conserved. Ignoring these factors leads to the evolutionary misconception that natural selection has produced a plethora of organisms that operate in a manner that are perfectly adapted to their environment, both at the physiological and cellular level. In fact, there is a substantial degree of entropy and uncertainty in the makeup of genes and biological systems.

Genetic drift refers the fixation of genetic loci with certain substitutions (i.e., alleles) or deletions within a population simply by chance. It has a substantial effect on gene duplicate fate and evolution as a whole. Genetic drift is conceptually opposite to selective pressure in that all substitutions have equal probability of being fixed in this model, i.e. $K_a = K_s$ if genetic drift were acting alone. That is not to say it selectively creates substitutions that are disadvantageous, but rather permits the fixation of neutral or deleterious substitutions (which the majority of substitutions are¹⁰). Genetic drift operates more strongly when effective population size is small^{11,12}. Thus, genetic loss does not always indicate a lack of evolutionary benefit as selection does not always overcome the deleterious effects of drift.

By the same token, genetic drift would also pervade for potentially useful genes that are not expressed due to lack of promoter, for example genes that are duplicated by retrotransposition. In order for their evolutionary benefit to be recognized within cellular systems and for selection to preserve these retrogenes, they must be expressed.

Acquisition of promoters may paradoxically occur due to genetic drift, though it has been suggested that retrogenes are constitutively expressed in testis by use of their endogenous long terminal repeats (LTRs) as promoters^{13,14} and that their utility is tested in the testes.

The mode of duplication more generally prescribes the predisposition of paralogs to the aforementioned evolutionary forces. Two rounds of whole genome duplication (WGD) in the ancestral vertebrate generated many of the related genes present today¹⁵. Smaller-scale duplications (SSDs), e.g. segmental duplication, also produce genes that usually inherit the coding sequences as well as regulatory elements of their parents. However, retained genes derived from SSD tend to evolve faster, conceivably because only a portion of the genome is duplicated and there is a disturbance in molecular stoichiometry. In sum, gene families are expanded in various lineages of the evolutionary tree by duplication of ancestral genes by means of retrotransposition, WGD, and SSD. The resulting paralogs are either conserved or are lost due to the combined effects of selective pressure and genetic drift.

Paralogs can be viewed as multiple copies of a single gene within a genome that are often divergent in terms of sequence and function. Alternative splicing products can also fit this definition, though these are accompanied by a different set of evolutionary assumptions. 95% of human genes are alternatively spliced^{16,17}, but most of this is

stochastic noise^{18,19}. Multiple isoforms can result from genetic drift at splice sites surrounding non-essential exons. However, the generation of multiple isoforms can be preserved by natural selection if these hold distinct regulatory roles in cellular networks. Careful observation of molecular conservation patterns can help distinguish whether alternative splicing is a selective functional innovation or stochastic²⁰.

In the context of deubiquitinating enzymes, one ancestral DUB can evolve several forms within a genome through duplication and alternative splicing to cleave ubiquitin in several different cellular contexts. The collection of human DUBs represents the retained expansion products from the ancestral set over the course of evolution. Tracing the genealogy of this superfamily of enzymes should yield considerable insight into the level of redundancy of the system.

STATEMENT OF HYPOTHESIS AND OBJECTIVES

The goal of the current work is to gain insights into origins and functional expansion of DUBs, initially focusing on a small subset of three paralogs, USP4, USP15 and USP11, establishing the timing and order of their divergence as well as evaluating the evolutionary and biological significance of USP4 alternative splicing, and finally extrapolating our methods to characterize the expansion and function of all DUBs. We hypothesize that many evolutionarily conserved DUB paralogs should have important, distinct properties, though concede that at least part of the systemic redundancy is owed to stochastic evolutionary processes. We first and foremost seek to objectively understand the evolutionary means that produced the observable diversity of DUBs in various species in order to ascribe biological importance to their roles where it is due.

CHAPTER I

Evolution of the highly networked deubiquitinating enzymes USP4, USP15, and USP11

This paper was published in BMC Evolutionary Biology: Vlasschaert, C., Xia X., Coulombe, J., Gray D. Evolution of the highly networked deubiquitinating enzymes USP4, USP15, and USP11. *BMC Evol Biol.* 2015 Oct 26;15:230. doi: 10.1186/s12862-015-0511-1.

The student performed the bioinformatics analyses, prepared the figures and wrote a first draft of the manuscript, which was edited by the student and her co-supervisors.

Research Article

Evolution of the highly networked deubiquitinating enzymes USP4, USP15, and USP11

Caitlyn Vlasschaert^{1,2,3}, Xuhua Xia^{1,4}, Josée Coulombe³ & Douglas A. Gray^{2,3*}

¹ Department of Biology, University of Ottawa

² Department of Biochemistry, Microbiology and Immunology, University of
Ottawa

³ The Ottawa Hospital Research Institute, Ottawa, Canada

⁴ Ottawa Institute of Systems Biology, Ottawa, Canada

*corresponding author. Centre for Cancer Therapeutics, Ottawa Hospital Research
Institute, 501 Smyth Rd., Ottawa, Ontario, Canada K1H 8L6.

Caitlyn Vlasschaert email: caitlyn.vlasschaert@gmail.com

Xuhua Xia email: Xuhua.Xia@uottawa.ca

Josée Coulombe email: jcoulombe@ohri.ca

Douglas A. Gray email: dgray@ohri.ca

ABSTRACT

Background: USP4, USP15 and USP11 are paralogous deubiquitinating enzymes as evidenced by structural organization and sequence similarity. Based on known interactions and substrates it would appear that they have partially redundant roles in pathways vital to cell proliferation, development and innate immunity, and elevated expression of all three has been reported in various human malignancies. The nature and order of duplication events that gave rise to these extant genes has not been determined, nor has their functional redundancy been established experimentally at the organismal level. We have employed phylogenetic and syntenic reconstruction methods to determine the chronology of the duplication events that generated the three paralogs.

Results: Our analyses indicate that USP4 and USP15 arose from whole genome duplication prior to the emergence of jawed vertebrates. Despite having lower sequence identity USP11 was generated later in vertebrate evolution by small-scale duplication of the USP4-encoding region. While USP11 was subsequently lost in many vertebrate species, all available genomes retain a functional copy of either USP4 or USP15, and through genetic crosses of mice with inactivating mutations we have confirmed that viability is contingent on a functional copy of USP4 or USP15. Loss of ubiquitin-exchange regulation, constitutive skipping of the seventh exon and neural-specific expression patterns are derived states of USP11. Post-translational modification sites differ between USP4, USP15 and USP11 throughout evolution.

Conclusions: In isolation sequence alignments can generate erroneous USP gene phylogenies. Through a combination of methodologies the gene duplication events that gave rise to USP4, USP15, and USP11 have been established. Although it operates in the

same molecular pathways as the other USPs, the rapid divergence of the more recently generated USP11 enzyme precludes its functional interchangeability with USP4 and USP15. Given their multiplicity of substrates the emergence (and in some cases subsequent loss) of these USP paralogs would be expected to alter the dynamics of the networks in which they are embedded.

Keywords: ubiquitin-specific protease, gene duplication, gene phylogeny, synteny, subfunctionalization

BACKGROUND

Protein ubiquitin tags are post-translational modifications that serve to either target substrates for proteasomal degradation or modify their interactive capacities [1]. Protein ubiquitination status is determined by the activities of the ubiquitin ligases that conjugate the ubiquitin moieties and the deubiquitinating enzymes (DUBs) that remove them; the balance of these activities thus affects key cellular processes. Among the most extensively networked [2] DUBs are the ubiquitin-specific protease (USP) paralogs USP4 and USP15, which regulate cell growth, embryonic development and innate immunity via their interactions with TGF- β [3, 4], Wnt/ β -catenin [5] and NF- κ B [6–8] pathway proteins respectively. USP4 and USP15 are also the only catalytic DUBs known to interact with the spliceosome [9–11], with more than eleven splicing factors identified as overlapping substrates [2]. This functional redundancy likely relates to their homology (there is 56.9% amino acid identity in *Homo sapiens* as indicated in Figure 1). One other DUB, USP11, bears considerable, albeit lesser, sequence identity to USP4 (44.5%

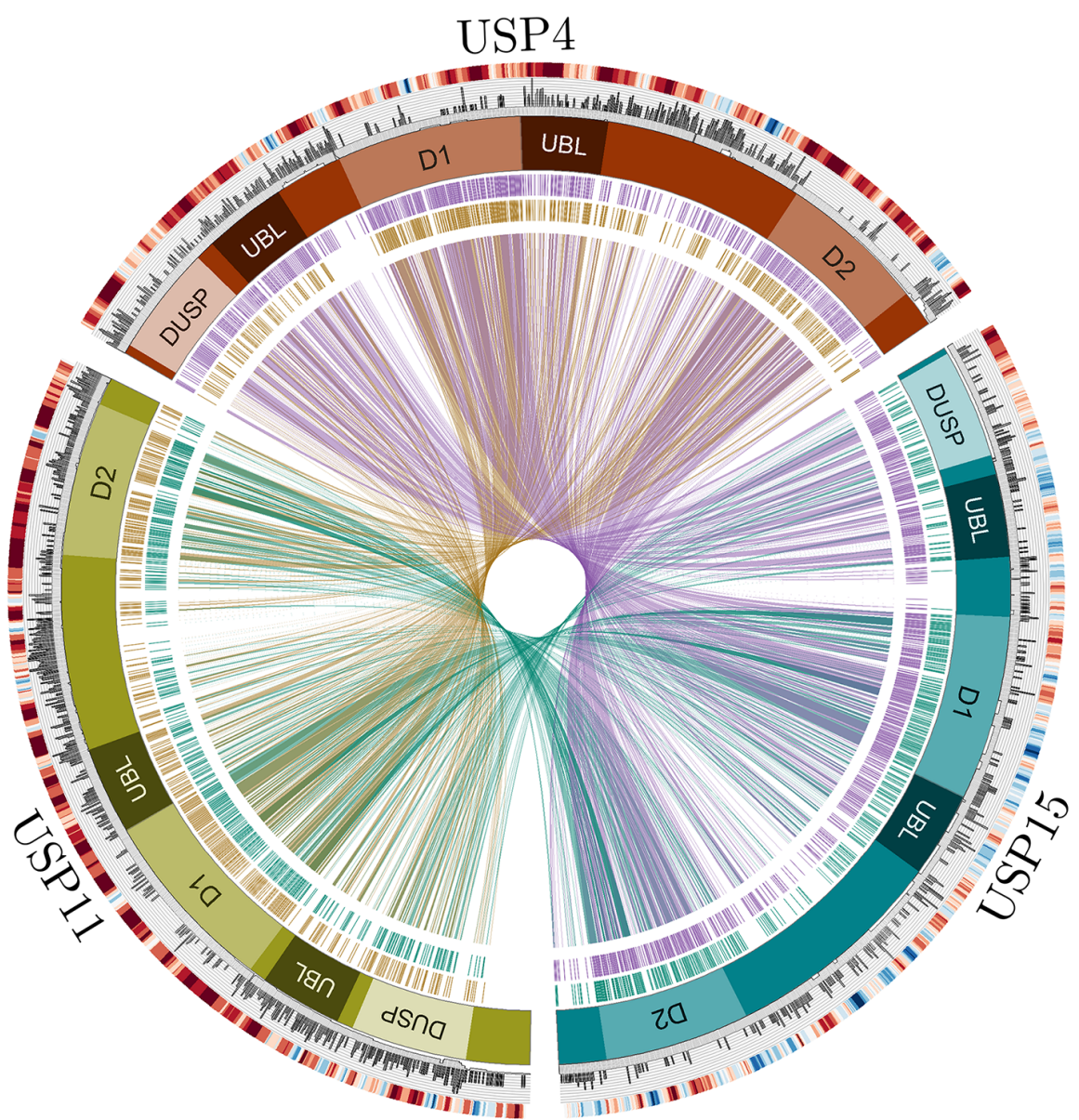


Figure 1. Comparison of USP4, USP15 and USP11 features. The red, blue and green boxes arranged in a circle represent USP4, USP15 and USP11, respectively. Domain structures are marked as follows: DUSP, domain in USP (N-terminal domain specific to these USPs); UBL, ubiquitin-like domain; D1 & D2, bi-part catalytic domain mediating ubiquitin cleavage. The interior of the circle links amino acid identities among paralogs, where each line represents an identical aligned residue. Links are colored as follows: USP4-USP15 purple; USP11-USP15 teal; USP4-USP11 gold. Alignment links are separated into two outer rings to facilitate viewing.

The exterior of the circle features two rings illustrating the following: Inner ring: orthologous protein conservation. The histogram shows site-specific entropy among vertebrate species in black. High entropy reflects high dissimilarity. For comparative measure, the number of species containing the aligned region in question is below the histogram in gray. Low species count indicates amino acid indels. Outer ring: GC content. The heat map indicates relative GC content at the third codon position (GC3), where high GC content is red and low GC content is blue.

identity) and USP15 (43.2%). The three paralogs share a common domain organization, consisting of a DUSP (*domain in USP*), two UBL (*ubiquitin-like*) and a bi-part catalytic domain (Figure 1).

Overexpression of these DUBs has been noted in various human cancers, which may be attributable to their collective regulation of oncogenic proteins. For instance, all three paralogs regulate the type I TGF- β receptor while USP15 and USP11 also regulate several of its downstream effectors [4, 12, 13]. Conversely, whereas USP4 and USP15 target p53-inhibiting ligases ARF-BP1 [14] and MDM2 [15], respectively, USP11 stabilizes p53 [16] as well as several other tumor suppressors including PML [17], BRCA2 [18] and Mre11 complex members MRE11 & RAD50 [2]. In sum, though these paralogs are functionally redundant in some capacities, each appears to have undergone substantial subfunctionalization and neofunctionalization. A summary of their known protein interactions is presented in Table 2.

Functional versions of USP4, USP15 and USP11 are detectable in most branches of the vertebrate lineage including human. Of the three, USP4 and USP15 are most similar in terms of sequence identity (Figure 1) and deubiquitination substrates (Table 2), which is consistent with the (USP11,(USP15,USP4)) branching pattern observed in phylogenetic analyses of these DUBs [19, 20]. This would suggest that the duplication that gave rise to USP4 and USP15 occurred most recently. However, a survey of USP paralogs encoded by metazoan genomes (Figure 2) contradicts this hypothesis: while functional USP4 and USP15 are present in cartilaginous fish at the emergence of the gnathostome branch, USP11 is not identifiable until bony vertebrates make their appearance. What is more, all single-copy USP sequences have most identity with either

Pathway		USP4	ref.	USP15	ref.	USP11	ref.
RNA splicing	Lsm2	+		+		-	
	Lsm4	+		+		-	
	Lsm6	+		+		-	
	Mepce	+		+		-	
	Naa38	+		+		-	
	Ppih	+		+		-	
	Prp3	+	[9]	+		-	
	Prp31	+		+		-	
	Prp4	+		+		-	
	Sart3	+	[9]	+	[10]	-	
TGF- β signalling	Tut1	+		+		-	
	Tgfbr1	+	[13]	+	[4]	+	[12]
	Smad7	-		+	[4]	+	[12]
	Smad1	-		+	[4]	-	
	Smad2	-		+	[4]	-	
	Smad3	-		+	[4]	-	
	Smad4	-		+	[4]	-	
	Smurf1	-		+		-	
	Smurf2	-		+	[4]	-	
Tumor suppression	Bmpr1a	-		+	[81]	-	
	p53	-		-		+	[16]
	Mdm2	-		+	[15]	-	
	Arf-bp1	+	[14]	-		-	
	Brca2	-		-		+	[18]
	Pml	-		-		+	[17]
	Notch1	-		+		+	
Innate immunity	Rb	+	[82]	-		-	
	Keap1	-		+	[83]	+	
	Rig-i	+	[84]	-		-	
	Trim25	-		+	[85]	-	
	Trim21	+	[86]	-		-	
	Rip1	+	[87]	-		-	
	Tak1	+	[6]	-		-	
	Traf6	+	[7]	-		-	
	Traf2	+	[7]	-		-	
	Ikka	-		-		+	[88]
	Ikba	-		+	[8]	+	[85]
Wnt/ β -catenin signalling	Apc	-		+	[90]	-	
	Nlk	+	[5]	-		-	
Other	A2ar	+	[91]	-		-	
	Usp7	-		+		+	[92]

Table 2. Summary of USP4, USP15 and USP11 interaction partners. Interactors were identified in a mass proteomic analysis by Sowa *et al.* [2] and those validated by independent, small-scale studies are shaded gray and referenced. Note that these include K48- and K63-linked substrates and possibly non-substrate binding partners. Red-shaded interactors indicate proteins that have antagonistic roles in the indicated pathway.

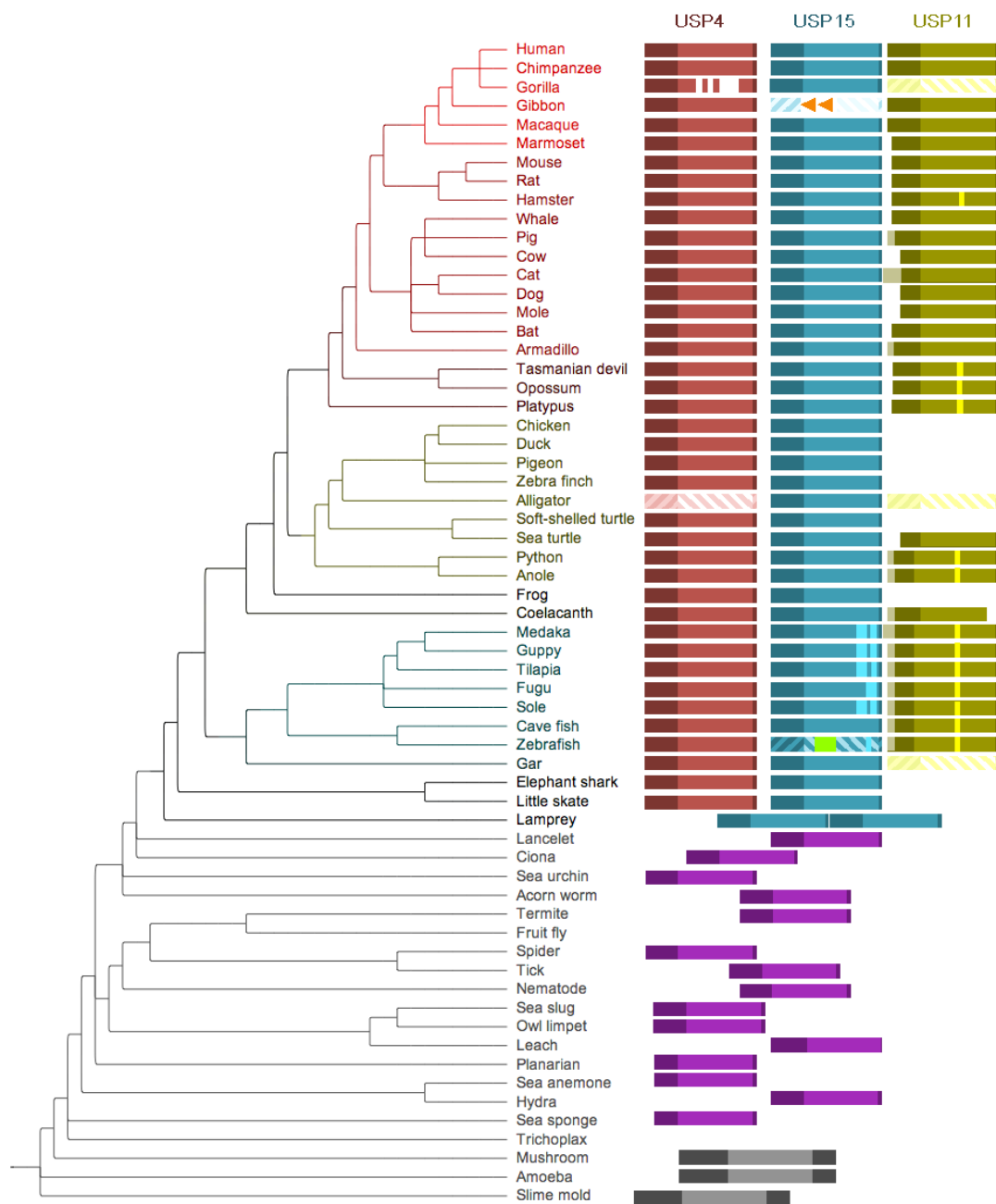


Figure 2. Phylogenetic distribution of USP4, USP15 and USP11. Red, blue, green, purple and black boxes represent USP4, USP15, USP11, ancestral (single copy) USP, and pre-USP ancestor sequences, respectively. Sequences are annotated and aligned according to their Reciprocal Best BLAST Hits (RBBHs), where lateral positioning of ancestral sequences indicates relative identity to human USP sequences. Translucent diagonally striped boxes indicate pseudogenes. Orange arrows indicate disruptive LINE1 element insertions in gibbon USP15 and green arrows indicate potentially disruptive insertion of a repetitive sequence of unknown origin in zebrafish USP15. Highlighted vertical bars indicate poly-glutamate sequences in USP15 and USP11.

USP4 or USP15. It is nevertheless possible that the USP11 duplication occurred earlier though its traces were erased by pseudogenization in deeper-branching species. One phylogeny represents the USP4, USP15 and USP11 relationship as a trifurcation [21], acknowledging its cryptic nature.

To understand the evolutionary changes in sequence, structure, and function among these paralogs, it is very important to know the temporal sequence of duplication. This enables us to determine which are the ancestral states and which are the derived states that potentially represent adaptation in response to an ancient environment. This motivated us to do phylogenetic studies to characterize the branching pattern and the timing of duplication events. An integrative *in silico* approach probing these systematic changes in a comparative genomic framework was employed to trace the duplication and subsequent radiation of USP4, USP15 and USP11. We first quantified and characterized USP paralogs in a set of representative metazoan genomes and delineated their divergence times in reference to known whole genome duplication events. We then evaluated ortholog variability to gain insight into the evolutionary processes that gave rise to the three paralogs observable in humans.

RESULTS

Phylogenetics based on aligned nucleotide and amino acid sequence

Fifty USP4, USP15 and USP11 coding sequences from 23 representative vertebrate and invertebrate genomes (listed in Table 1, Material and Methods) were aligned using MUSCLE [22] with Gblocks cleaning [23], yielding an aligned length of 3981 sites. For phylogenetic reconstruction, we used the maximum likelihood method implemented in DAMBE [24] with the estimated transition/transversion ratio, the F84

model, and *Amphimedon queenslandica* (sea sponge) as the outgroup. The resulting unrooted tree (Figure 3) has drastically different evolutionary rates among different lineages, with USP11 evolving particularly faster than other lineages. We performed a likelihood ratio test of the molecular clock hypothesis with the 50 sequences and the TN93 model, and the clock hypothesis is strongly rejected ($\ln L$ without clock = -17452.3864, $\ln L$ with clock = -17630.0485, likelihood ratio chi-square = 355.3242, $DF = 48$, $p < 0.0001$). We have also tested the clock hypothesis by using the third codon positions only, but the clock hypothesis is still strongly rejected ($\ln L$ with no clock = -4053.1815, $\ln L$ with clock = -4116.3185, Likelihood ratio chi-square = 126.2739, $DF = 48$, $p < 0.0001$). Thus, the paralogous sequences are not appropriate for dating. Indeed, age-calibrated phylogenetic dating of the codon sequences generated a tree that placed the divergence of USP15 vertebrate sequences before that of the single-copy ancestral sequences (Supplementary Figure 1). This erroneous topology reflects a discord between the substitution model's assumptions and the nature of the sequences: USP15 orthologs are situated in low-GC regions in vertebrates (human, mouse, chicken, lizard) while USP11 and USP4 are in moderately high-GC isochores. This bias is reflected in their respective GC3 contents (Figure 1) and thus violates the fundamental assumption of time homogeneity of all practical substitution models. We note that the paralogous genes in vertebrate species are often located in different GC isochores [25, 26]. For this reason, a nucleotide-based or codon-based analysis may bias phylogenetic estimation. To address this problem, we have also analyzed aligned amino acid sequences of the 23 species by the likelihood method. We have adopted the approach recommended by Rodriguez-Ezpeleta et al. [27] by recoding amino acids by size and polarity into four groups: small

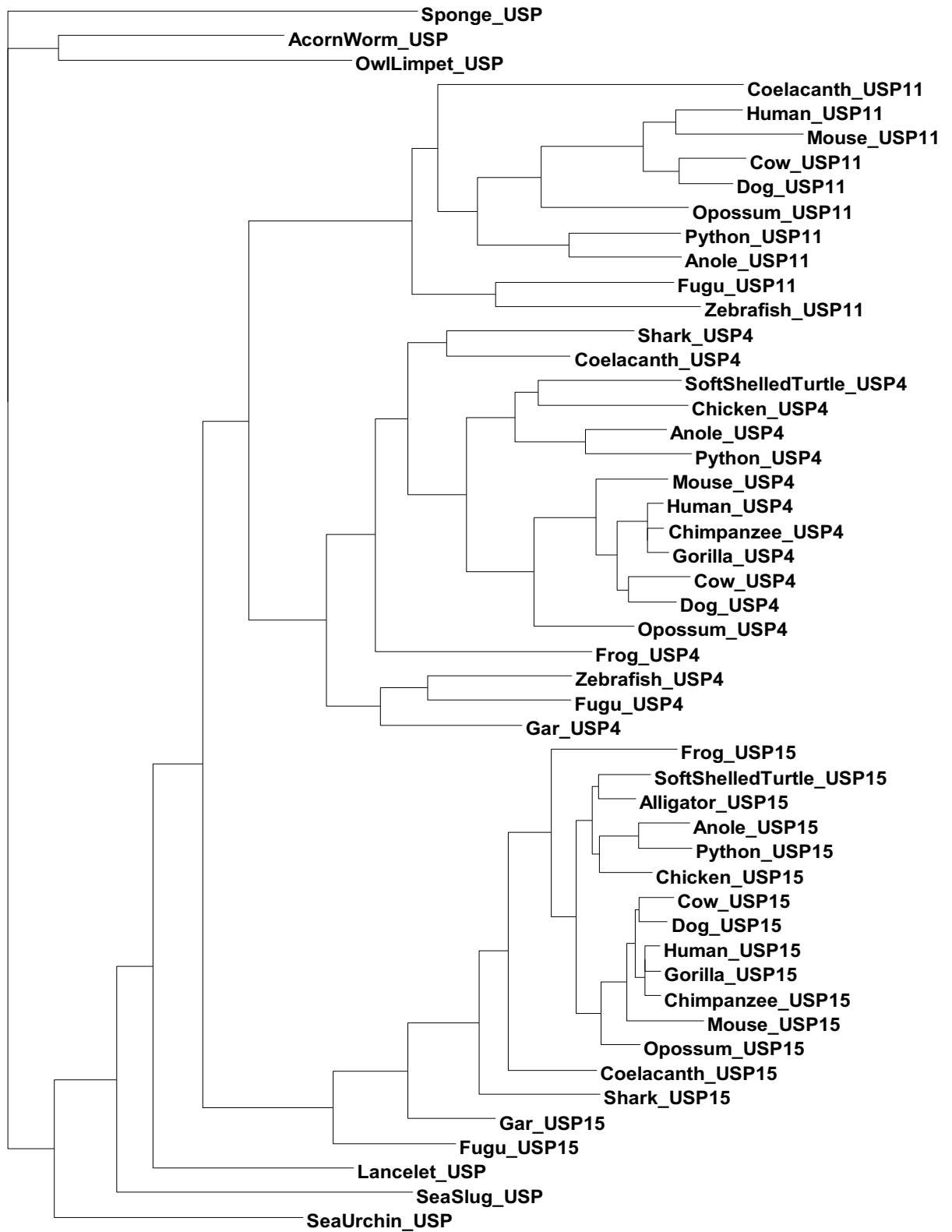


Figure 3. Maximum likelihood reconstruction of aligned codon sequences. A maximum likelihood tree of three paralogous genes from representative vertebrate species is represented together with their orthologs from invertebrate species. The unrooted tree was constructed with the F84 model and the maximum likelihood method implemented in DAMBE.

and polar (SCTND), large and polar (QEKRHY), small and non-polar (PAGV), large and non-polar (ILMFW). This approach not only results in more robust phylogenetic reconstruction, but also dramatically reduces computation time. The resulting tree (Figure 4) is largely concordant with the maximum likelihood tree topology based on aligned nucleotide sequences (Fig. 3), i.e., USP15 splitting first from USP4/USP11, followed by the USP4 and USP11 split, with the primitive species encompassing a single ortholog clustered close to the root.

The branching pattern of Figures 3 and 4 enables us to infer an approximate time for gene duplication events. The USP15 lineage splits from (USP4,USP11) during the period between the divergence of vertebrates from primitive chordates (from 581 to 460.6 millions of years ago, or MYA [28]) and the branching of shark from teleost (462.5 to 421.75 MYA [28]), corresponding to the timing of a known whole genome duplication event [29]. A second gene duplication leads to the USP4/USP11 split which occurred in the common ancestor of bony fishes represented by gar, fugu, zebrafish and coelacanth (421.75 to 416 MYA [28]). USP11 is absent in shark. Given that the shark genome has evolved little [30], we may infer that the absence is ancestral instead of secondary loss, i.e., the USP4/USP11 split occurs after the divergence of shark from the ancestor of teleosts.

Synteny of USP11 and USP4 loci supports duplication in a Euteleostome ancestor

Gene homologs often bear not only high sequence identity to their ancestors, but can also retain their genomic context. Synteny, the linear conservation of physically linked gene clusters within or between genomes, can be revelatory of paralogous or orthologous evolutionary relationships. We thus conducted a comparative analysis of the

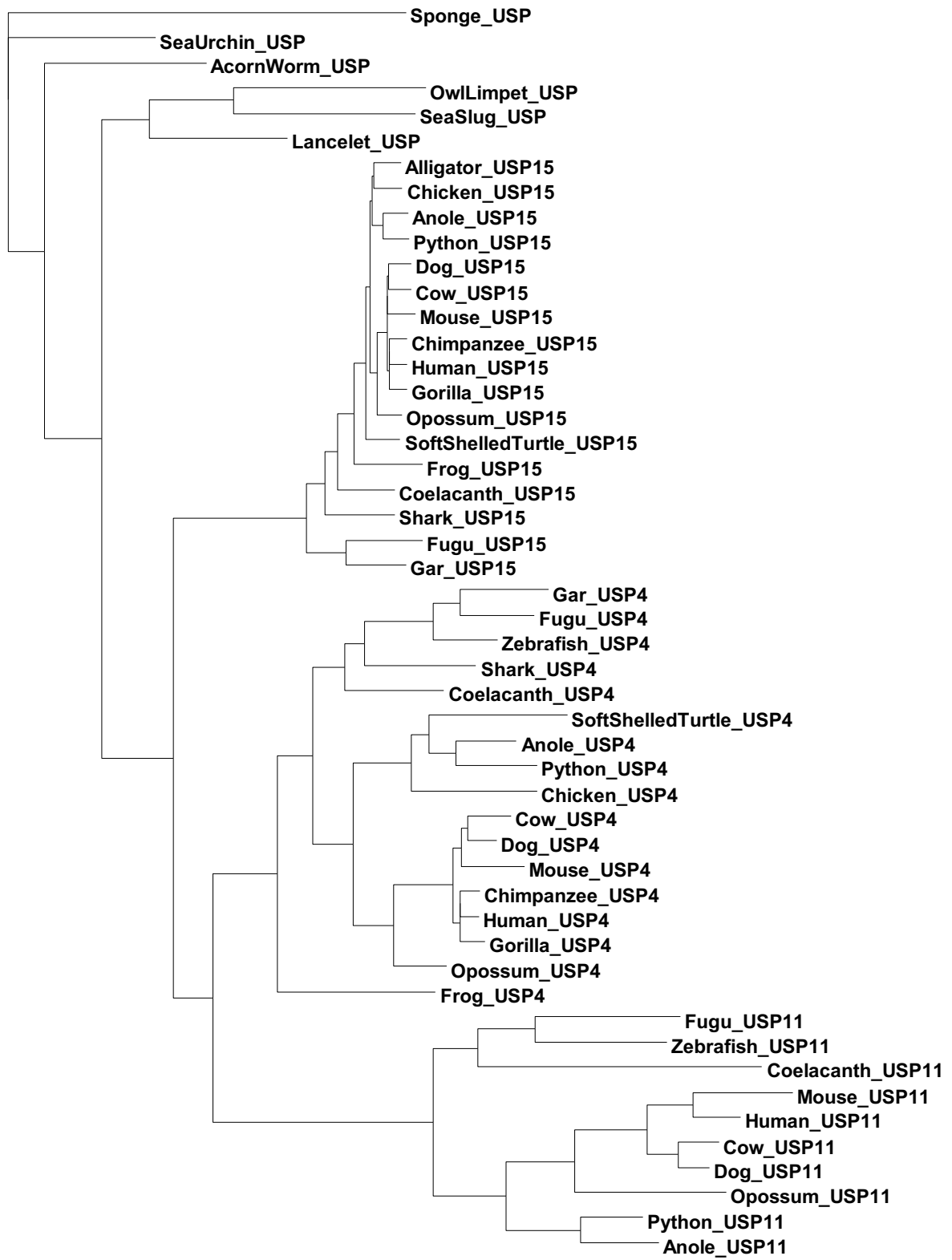


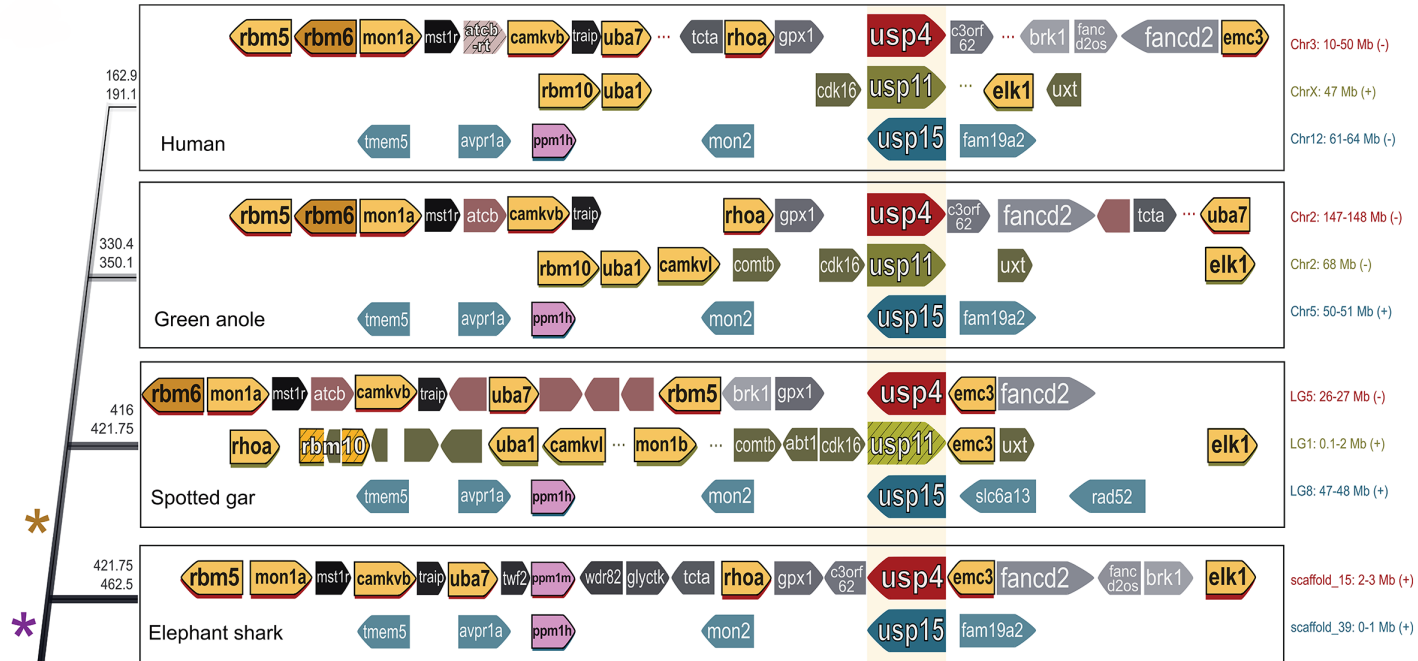
Figure 4. Maximum likelihood reconstruction of recoded amino acid sequences.

Depicted is a maximum likelihood tree of 50 aligned amino sequences after recoding amino acids by size and polarity into four groups: small and polar (SCTND), large and polar (QEKRHY) small and non-polar (PAGV), large and non-polar (ILMFW). The rooted tree was produced using the ProtML method implemented in DAMBE.

genomic region encompassing USP4 in *Callorhincus milii* (elephant shark) and the regions surrounding USP4 and the USP11 pseudogene in *Lepisosteus oculatus* (spotted gar), representing putative pre- and post-duplication loci. We found that several genes adjacent to shark USP4 map physically near to the USP4 orthologs in gar and other higher vertebrates including human and anole (Figure 5). In fact, the synteny of the region is remarkably well conserved after duplication: in addition to USP11, six other functional paralogs of genes surrounding USP4 in shark and in gar can be identified within 1Mb of gar pseudo-USP11, while these co-duplicates are absent from the shark genome. Invertebrate genomes likewise encode only a single copy of these genes. In contrast, no USP11 co-duplicates can be identified at the USP15 locus. This supports our inferred branching patterns (Figs. 3 and 4) and altogether suggests that a concerted duplication of the USP4 chromosomal region median to the divergences of gnathostomes and euteleostomes gave rise to USP11.

As a consequence of significantly different rates of evolution, Bayesian molecular dating of USP4 and USP11 aligned sequences overestimates their divergence time at 583-885 MYA. Three parallel runs of aligned USP4, USP11 and ancestral USP sequences with six calibration points converged at the rooted tree topology shown in Supplementary Figure 2 (note that several of the speciation node patterns and timing are largely inconsistent with known evolutionary relationships). Two sets of identified co-duplicates, RBM5/UBA7 and RBM10/UBA1, are co-localized with USP4 and USP11 respectively throughout vertebrate evolution, and can be used to date the duplication event by proxy. While neither RBM5/RBM10 nor UBA7/UBA1 follow a strict molecular clock, $\Delta\ln L$ of RBM is greatly reduced compared to that of USP4 and USP11 ($\Delta\ln L_{\text{USP}} = 177.6621$,

A



B



Figure 5. Shared synteny of USP4 and USP11 loci in Euteleosts. A) Illustrated comparison of USP loci for elephant shark (*Callorhincus milii*), spotted gar (*Lepisosteus oculatus*), green anole (*Anolis carolinensis*) and human (*Homo sapiens*). Genes are represented by arrows, where black outlines indicate paralogous genes and striping indicates pseudogenes. Paralogs shared by USP4 and USP11 are coloured gold, while those shared by USP4 and USP15 are coloured purple. Genomic location of loci is indicated to the right. Upper and lower estimates of divergence times (in millions of years) indicated to the left for the following clades (in ascending order): jawed vertebrates (incl. shark), euteleosts (incl. gar), tetrapods (incl. anole) and mammals (human). Stars indicate inferred divergence times for USP4-15 (purple) and USP4-USP11 (gold).

B) Schema of paralogous gene collinearity and rearrangement events in (USP4-USP11) and (USP4-USP15) loci.

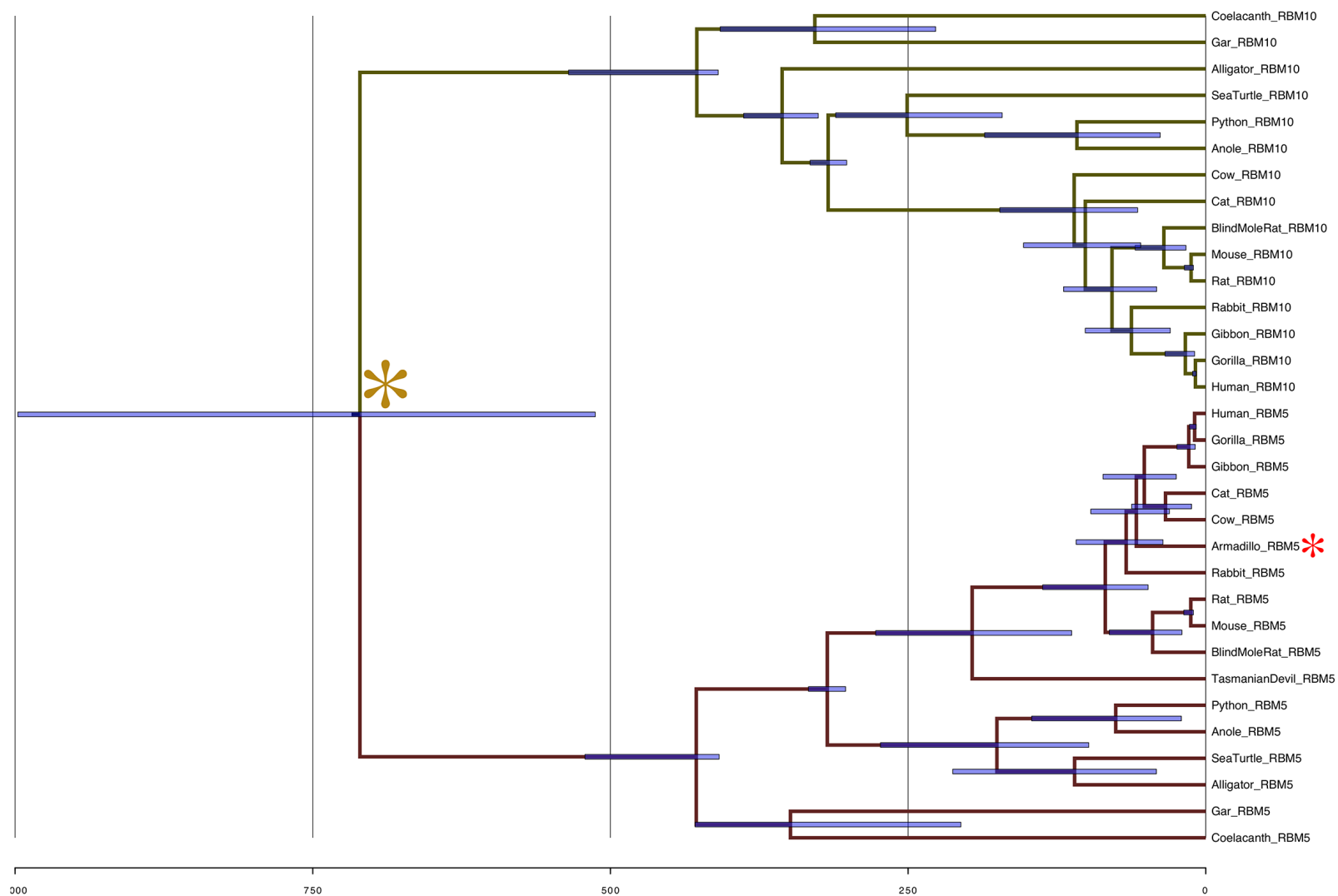


Figure 6. Bayesian dating of aligned co-duplicate codon sequences. Phylogenetic reconstruction and fossil-calibrated dating of aligned codon sequences for RBM5 and RBM10 was generated using BEAST v. 1.8. 95% credible intervals are indicated. Calibration points were obtained from TimeTree. The gold star indicates the inferred RBM5-RBM10 divergence time. Red stars indicate major deviations from true topology.

$\Delta \ln L_{\text{RBM}} = 77.8252$; $\Delta \ln L_{\text{UBA}} = 412.3984$). Figure 6 presents a phylogenetic reconstruction of RBM5 and RBM10 using a relaxed molecular clock; at 512 MYA, the 95% credible interval upper bound of the predicted divergence time for these co-duplicates falls nearer the expected range and thus represents a rough estimate for the timing of duplication of the USP4 loci.

We believe our analyses provide overwhelming evidence in favor of a (USP15,(USP4,USP11)) branching pattern as opposed to the (USP11(USP4,USP15)) pattern that would be inferred based on sequenced relatedness [19, 20]. We posit that USP11 experienced greater coding sequence drift immediately following its duplication, resulting in complete pseudogenization in some species (e.g. gar) while in others a fast-evolving (Figs. 3, 4), subfunctionalized (Table 2) protein emerged that is less similar than its well-conserved ancestors, USP4 and USP15. Adopting this novel understanding of their evolutionary relationship, we next examined the variability among USP homologs.

Signature features of USP paralogs

Four key molecular traits are thought to engender paralog functional radiation: structure-function innovations, distinctive spliced isoforms, altered cellular regulation (via post-translational modification), and specific spatiotemporal expression patterns [31]. While the defining domain architecture presented in Figure 1 is pervasive in all USP4, USP11 and USP15 as well as ancestral (single copy) homologs, divergence among the structured domains and the unstructured linking regions is observed, which has been reported to confer differential enzymatic properties [20, 32]. We herein derive the constitutive evolutionary differences, or molecular signatures, that have defined USP4, USP15 and USP11 from their inception using branching pattern knowledge and ancestral

state reconstruction. Filtering paralog-defining features is more informative than monospecies sequence alignment, which contains intraparalog (species-specific) variations likely to be especially pronounced in the fast-evolving USP11. These conserved molecular signatures, explained below, are summarized in Figure 7.

Structure-function innovations

First let us consider the molecular signatures of structured regions in USP4, USP15, and USP11. The “domain in USP” (DUSP) and “ubiquitin-like” (UBL) structured regions form the N-terminal domain that distinguishes this subgroup of USPs. DUSP-UBL domains mediate some enzyme-substrate interactions [9, 33–35] and confer intrinsic regulatory capacities that have been structurally modeled for mammalian USP4, USP15 and USP11 [20, 32, 35, 36]. For instance, USP4 dimerization occurs in equilibrium through this domain, while neither USP15 nor USP11 are expected to dimerize *in vivo* [20]. The DUSP-UBL domain of USP4 also regulates ubiquitin active-site binding dynamics through its association with the unstructured insert region [32], though the absence of key residues impedes this regulatory function in human USP11 [32, 35]. The enzyme kinetics of USP15 are more similar to that of USP4 [32]. Given our derivation of their duplication chronology, it seems likely that the loss of ubiquitin-exchange regulation in USP11 is a derived and not ancestral state, though structural information is available only for mammalian proteins. Figure 7a presents an alignment of the DUSP-UBL domains of an ancestral USP with the signature sequences of USP4, USP15 and USP11. Lancelet was selected as the ancestral species because it is the closest single-copy relative (Figure 2). In addition, the domain sequence is identical in *Branchiostoma floridae* and the newly sequenced *B. belcheri*, two lancelets that have

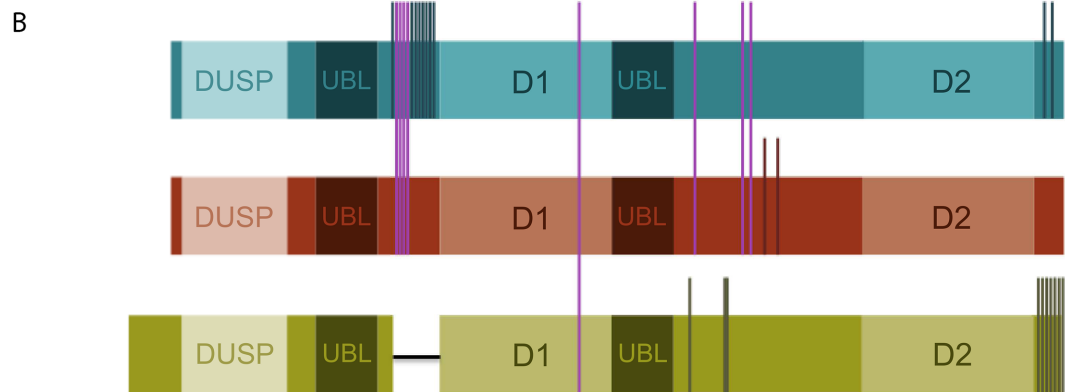


Figure 7. Summary of signature features of USP4, USP15 and USP1. A) Alignment of signature sequences for USP4 (red), USP15 (blue) and USP11 (green) with the single-copy sequence (purple). The DUSP-UBL compound domain is shown and coloured as in Figure 1. B) Schematic illustration of signature phosphorylation sites and loss of alternatively spliced exon in USP11. Vertical bars traverse sequences with shared phosphorylation sites.

experienced a high degree of protein evolution [37], suggesting that it is an accurate depiction of a pre-duplication USP. USP signature sequences indicate residues that are conserved in a majority of members from each phylogenetic clade (elimination of species-specific substitutions). While the key residues are largely conserved in USP4, USP15, and the ancestral USP, disruption of the hydrophobic pocket and shortening of DUSP-UBL linker [32, 35] are signatures of USP11. This derived state implies that USP11 has had a different mode of action throughout time.

The two parts of the structured catalytic domain of these USPs, D1 and D2, are the most highly conserved regions among paralogs and orthologs (Figure 1). Both are required for catalytic activity, and their conservation extends beyond the USPs under current consideration to the entire USP subfamily of deubiquitinating enzymes.

Distinctive spliced isoforms

Whereas the seventh exon (E₇) is alternatively spliced in USP4 and USP15, a corresponding exon is absent in USP11. The flexible linker region separating the DUSP-UBL and catalytic domains is roughly 20 residues long in USP11, its length in USP4 and USP15 short isoforms and the minimal length required for the aforementioned domain interaction [32]. Shark USP4 and the lancelet single-copy USP, ancestral to USP11, contain E₇; what is more, both long and short isoforms have been reported in chondrichthyes. Thus, the “permanent skipping” of E₇ in all USP11 represents a derived state. Alterations in the stoichiometry of USP4 isoforms have been reported for a rare bone disease [38], though the functional consequences of E₇ alternative splicing have not been studied. In all species, the polypeptides encoded by USP4 and USP15 E₇ are serine-rich, and many serve as putative post-translational modification (PTM) sites as identified

in large-scale studies on human proteins[39]. In sum, the loss of E₇ is a signature derived state of USP11 with potential functional or regulatory implications.

Altered cellular regulation

Post-translational modification (PTM) regulation can differ among gene duplicates. Some PTM sites are well conserved while others stably differentiate the USP paralogs in question. For one, Ser445 (a known Akt phosphorylation site [13]), there is conservation in all USP4, USP11, USP15 and ancestral homologs. There are, on the other hand, multiple cases wherein a putative phosphorylation site has been lost or gained in USP11 relative to its ancestor, USP4. Within the insert region, USP4 Ser675 and Ser680 (identified phosphorylation substrates in multiple studies [40–48]) are conserved in USP15 but absent from USP11. Similarly, in an alignment of all USPs, putative phosphorylation site USP4 Tyr539 [39] is conserved in USP4 and USP15 while it is substituted by Phe in USP11. Slightly downstream, USP11 has Tyr551 (a reported phosphorylation site [39]) and Tyr554 whereas His and Phe, respectively, are universally present in USP4 and USP15. Still within the insert region, at positions 607 and 608, there exists in USP11 a pair of tyrosines that have been identified as phosphorylation sites in several large-scale studies [39]. The region in question aligns poorly with other paralogs, though there are two reported, albeit low confidence, serine phosphorylation sites in USP4 and none in USP15. As previously mentioned, the alternatively spliced exon, lost in USP11, contains several reported phosphorylated serines in USP4 and USP15 [39].

The N and C-termini are remarkably different among USP4, USP11 and USP15. The N-terminus of USP11 is longer, more disordered and more hydrophobic (rich in alanine). In addition, the C-termini of all gnathostome USP15 present a segment rich in

aspartic acid, glutamic acid and asparagine (e.g. human: 962-DEDSNDNDNDIENEN-976; shark: 978-DEDCNENDVENEN-990), except those of teleost fish, which instead have C-terminal segment(s) exceptionally rich in glutamate (e.g. zebrafish: 775-EKEEEEEDEDEEDVNDSEQEED-795; tongue sole: 966-DEEDEEEEEEEEGEVEEEDEEEEEEGRE-981, 1015-NEREDEEEEEEEEEEEEEEEQE-1035).

A poly-E repetitive sequence is also found in USP11 of various organisms, including teleost fish, some reptiles, the opossum and the Chinese hamster. These regions are schematically highlighted in Figure 2. Aspartic acid and asparagine residues can be hydroxylated, though it remains to be seen whether any hydroxylation of such residues occurs within the acidic domains of the USPs. In addition, the D- & N- rich C-terminus of non-teleost fish USP15 presents two validated serine phosphorylation sites [39, 50–55], absent from USP4, whereas human USP11 has seven of these sites [39, 50, 55–57] within its final 20 residues that are conserved among mammals. While many of these conserved and differential phosphorylation sites remain to be functionally characterized, most are all located within unstructured regions, namely the insert, linker, and C-terminal regions. This is consistent with reports that disordered region often serve as PTM substrates [58–61] and changes in PTM regulation contributes to the functional divergence of paralogs [31]. In addition to phosphorylation and hydroxylation the disordered regions of the USPs may be subject to a number of other modifications including acetylation, methylation, and/or the addition of peptide moieties such as ubiquitin or SUMO. The contribution of this growing repertoire of PTMs to USP4, USP15, and USP11 regulation has yet to be established.

In sum, each paralog has distinctive signature features that represent common

evolutionary categories namely structure-function innovations, distinctive spliced isoforms and altered cellular regulation. The fourth common differentiating trait, different spatiotemporal expression patterns, will be discussed in a later section.

Variable mechanisms of USP11 loss

As depicted in Figure 2, USP11 has been lost multiple times throughout vertebrate evolution. In select fish, reptile and mammalian genomes, the syntenic loci where USP11 habitually resides hosts USP11 pseudogenes in lieu of functional genes. For instance, among reference primate genomes, USP11 is uniquely pseudogenized in *Gorilla gorilla*. The phylogenetic dispersal of pseudogenization suggests that these events occurred independently. In birds such as chicken, however, the entire syntenic region containing USP11 has been deleted while USP4, USP15 and their respective neighbouring genes are conserved, as illustrated in Figure 8. A chromosomal rearrangement event in the avian ancestor may be responsible for the deletion of the segment containing USP11.

In vivo demonstration of a minimal requirement for USP4 or USP15

Whereas the variable retention or loss of USP11 suggests that it is dispensable, all species contain either or both USP4 and USP15. It is reasonable to speculate that one functional copy from this gene pair is essential for viability, and we tested this hypothesis using mouse strains in which the *Usp4* or *Usp15* gene had been inactivated by the insertion of a retroviral provirus. The TF2497 and TF2834 strains have gene-trap proviruses in the *Usp4* and *Usp15* genes respectively; in both cases the insertion disrupts the gene near the 5' end and precludes expression of a functional gene product (indeed no transcript can be detected by the sensitive method of reverse-transcription/polymerase

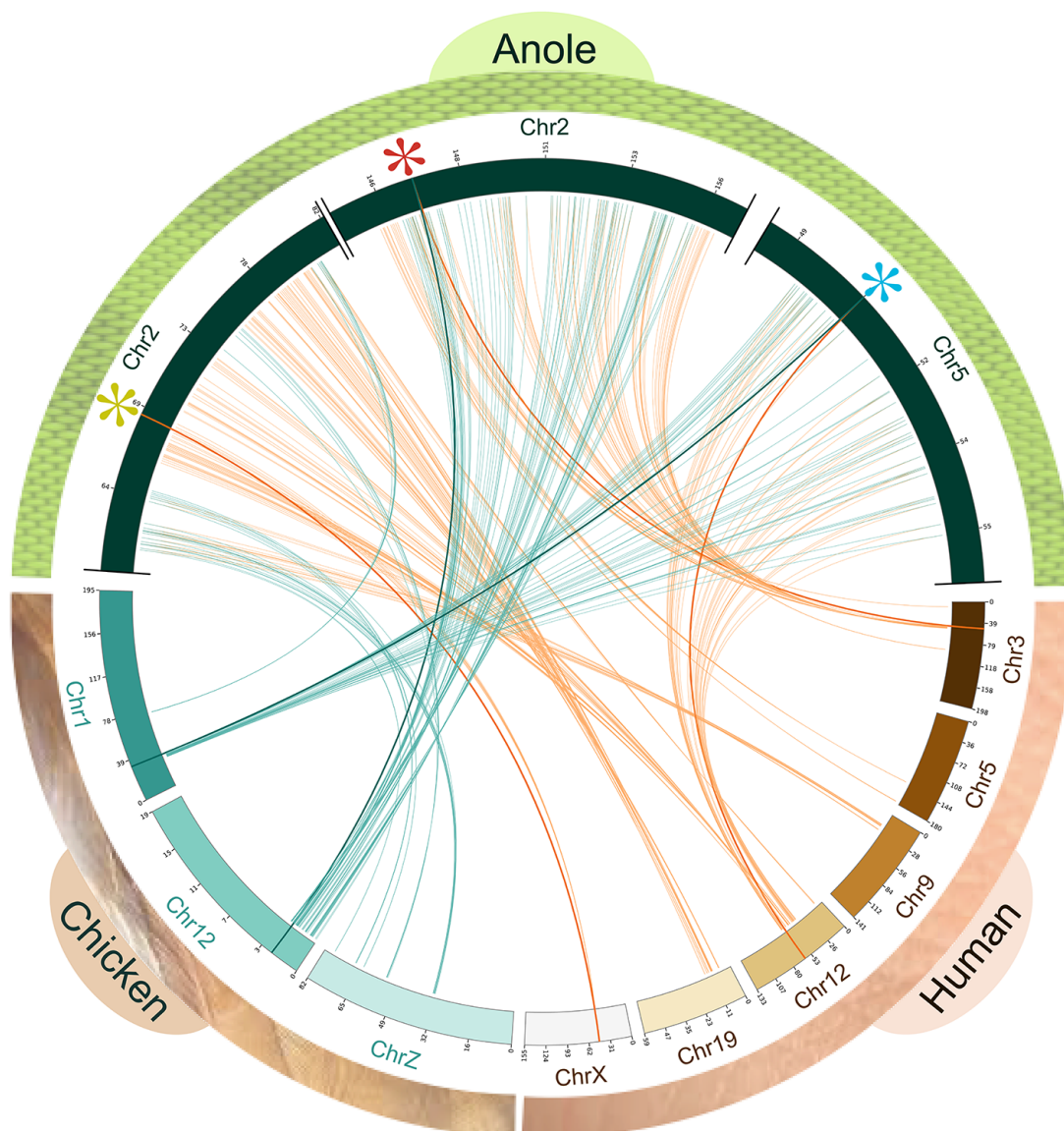


Figure 8. Synteny of USP4, USP15, and USP11 in anole compared to chicken and human. USP11 and surrounding region is absent in chicken. Asterisks represent the positions of anole USP genes (green indicates USP11, red is USP4, and teal is USP15).

chain reaction). Both strains are viable when homozygous for the inactivating insertion, and we have found no evidence of reduced fertility or obvious phenotypic effects. The lack of phenotypic consequences could be explained by functional redundancy between the USP4 and USP15 enzymes, but to determine if this is the case we conducted genetic crosses between mice heterozygous for the two genes (approval for these experiments was provided by the Animal Care Committee, University of Ottawa). As reported in Table 3, of 166 pups born we were unable to identify any progeny that had inactivating mutations in all four alleles, though all other expected genotypes were detected. Given that one of sixteen pups would be expected with the compound null genotype the lethality of this genotype can be asserted with a high level of confidence (from a binomial analysis $p = (15/16)^{166} = 0.000022$). We therefore conclude that USP4 and USP15 have sufficient functional redundancy to rescue inactivating mutations in a reciprocal fashion. The presence of functional USP11 genes is insufficient to rescue pups that are null for both USP4 and USP15. Mice are viable with one functional allele from the USP4/USP15 gene pair, though some apparent deviation from Mendelian ratios suggests that there may be phenotypic consequences of this haploinsufficient state. The nature of these consequences will be explored in future studies.

DISCUSSION

In the present work we have established the duplication chronology of a subgroup of highly networked ubiquitin-specific proteases, USP4, USP15 and USP11, and have characterized their subsequent radiation. According to the widely accepted 2R theory [29], vertebrate genomes have undergone two rounds of whole genome duplication (WGD). While it was conventionally assumed that these WGD events predated the

		USP 15			<i>Total</i>
		wt	het	null	
USP4	wt	23	29	17	69
	het	21	35	11	67
	null	8	22	0	30
<i>Total</i>		52	86	28	166

Table 3. Pooled progeny of USP4- and USP15-null mouse crosses. Mice heterozygous for inactivating mutations of both USP4 and USP15 were mated, and progeny genotyped by polymerase chain reaction analysis. The number of progeny of each genotype is indicated, where wt represents homozygous wild-type, het represents heterozygous, and null represents homozygous null mice. The absence of progeny null for both USP4 and USP15 is statistically significant ($p = 0.000022$).

divergence of jawless and jawed vertebrates [62], recent analysis of the elephant shark genome [30] placed at least one WGD event median to cyclostome and gnathostome divergence. In fact, subsequent studies have suggested that the 2R events occurred independently in cyclostomes and gnathostomes [63, 64], and that the former expansion was further shaped by an additional, lamprey-specific WGD. Thus, the USP15-like duplicate in lampreys is likely not orthologous to gnathostome USP4; rather, USP4 and USP15 appear to be ohnologs derived from WGD in a jawed vertebrate ancestor. In addition, the USP15-likeness of the lamprey version suggests that this paralog is the ancestral sequence, though there is no consensus among invertebrates as to whether their single copy most resembles USP4 or USP15 (Figure 2). The issue cannot be settled by genomic synteny reconstruction, which is considerably more difficult in earlier species due to increased divergence time and the present lack of chromosomal assembly data for many species. In contrast, well-conserved intragenomic synteny points to the emergence of USP11 as the result of a more recent duplication event that does not coincide with any reported WGD event; it is likely the product of a small-scale duplication (SSD). The characterization of USP4/15 and USP4/11 duplications as WGD and SSD, respectively, corroborates well with reported trends for these phenomena: SSD-derived paralog sequences tend to evolve faster and are more functionally divergent [65].

Several notable differences in USP composition exist between and within clades. Primate species appear to have inconsistent USP repertoires: USP15 was inactivated by insertion of a LINE1 element in the gibbon, while erasure of USP11 and reduction of unstructured USP4 domains can both be observed in the gorilla. USP11 was also lost in the avian ancestor, inferred by the consistent absence of its genomic locus in all bird

genomes (Figure 8). Curiously, avian USP4 presents notable deviations from the signatures of this paralog: of the six bird genomes surveyed, all bear mutations in crucial residues for the ubiquitin-exchange mechanism [32], i.e. Arg40 and/or Met24 mutated in all, disruption of DUSP-UBL linker residues (a.a. 88-92) in chicken, QQD box region deleted in duck, and so on. In fact, USP4 is also more divergent in other species where USP11 was lost, which lead for example to the consistently incorrect branching patterns for frog and gorilla USP4 (Figs. 3, 4; Supp. Figs. 1, 2). What is more, avian USP4 adopts some USP11 signatures: these have collectively lost the Ser675 and Ser680 phosphorylation sites, while USP4 of the pigeon and zebra finch have also lost E₇. The loss of these features that define all other USP4 (and USP15) is a derived state of USP11 and may thus represent a homoplastic convergence of avian USP4 toward the USP11 sequence.

Most of the signature features distinguish USP11 from USP4 and USP15, though the divergence of these last two is of practical interest due to their high protein sequence identity (Fig. 1), functional overlap (Table 2), and capacity for reciprocal rescue at the organismal level (Table 3). USP4 and USP15 differ in their codon usage: USP15, located in GC-poor isochores, employs more AT-ending codons than USP4. Low GC content is common in germ-line specific genes [25]. USP15 is in fact expressed at notably elevated levels in mature oocytes [66] (oocytes being the cell type for which its expression is the highest in mice [67]) while USP4 is at low abundance throughout oocyte maturation [68]. USP4 is predominantly expressed in somatic cells, particularly those of the immune system [67]. The distinct spatiotemporal expression patterns of USP4 and USP15 could explain why these redundant proteins have been maintained: vertebrate genomes could

optimally encode two versions of an ancestral protein to accommodate its important roles in germ and somatic cells. While we show that one functional copy of USP4 or USP15 is a minimum requirement for viability (Table 3), the observed departure from Mendelian ratios may arise from a functional deficiency in oocytes haploinsufficient for USP15. Planned experiments (including *in vitro* culture of early embryos) should be informative in this regard. The expansion of TGF- β pathway substrates in USP15 may reflect an enhanced role in the regulation of oocyte development [69–71], while USP4 may have become the USP of greater importance in innate immunity pathways, as reflected by an increased number of substrates (Table 2). Further, an inserted in-frame zebrafish-specific repetitive element has modified the USP15 catalytic domain coding sequence of this species. While it remains to be seen whether the enzymatic activity of USP15 has been altered or inactivated by this insertion, we anticipate that perturbation of USP15 will provide insights into DUB network rewiring in the zebrafish. As a model system that is amenable to the testing of hypotheses through genome manipulation, the zebrafish should be ideal for future investigations of the respective roles and expression patterns of USP4 and USP15. The expression pattern of USP11 is notably distinct: without exception in human, mouse, rat, and pig its expression is predominantly neuronal [67]. In contrast to its paralogs [4], USP11 exerts a protective effect in glioma [17] as it stabilizes many tumor suppressors (Table 2).

While all organisms minimally retain USP4 or USP15 and some have in addition USP11, none have more than these three closely related USPs (including teleost fish and lamprey, which have experienced a third whole genome duplication). Genomes coding for USP4, USP15 and USP11 may thus represent the optimal system, where USP11 is an

optional descendant whose functional contributions remain largely unexplored. Prior to the advent of whole genome sequencing it would have been reasonable to predict that with increasing organismal complexity there would be increasing complexity in molecular systems essential for development and tissue homeostasis, and the machinery relating to ubiquitin conjugation and removal would be high on the list of molecular systems expected to become more elaborate. In the case of deubiquitinating enzymes such a prediction would have been validated by whole genome sequencing: whereas vertebrate genomes encode more than 50 USP enzymes, roughly half this number are encoded by the genome of the fruit fly, and roughly half again by the genome of the budding yeast. In the evolution of complex molecular pathways the additional USP genes generated by WGD or SSD events could have provided substrate material for neofunctionalization or subfunctionalization. One can easily imagine how innovation within an augmented USP repertoire could facilitate innovation in complex signaling cascades (as exemplified by the NF- κ B pathway central to innate immunity, or the TGF- β pathway). While we have restricted our focus to USP4, USP15, and USP11 we believe our analysis of the evolutionary history of this subset of deubiquitinating enzymes has been instructive in a broader sense. It demonstrates, for example, that BLAST alignments, while intuitive, can be misleading in the construction of an accurate USP phylogeny. Sequence similarity alone would not predict the branching pattern summarized in Figure 9, which arose from extensive phylogenetic reconstruction of the DUSP-containing USP family incorporating aligned nucleotide and amino acid sequences, taxonomic distribution and patterns of synteny. We are hopeful that our approach can serve as a template for future studies of USP gene evolution, and will

ultimate lead to a better understanding of the origins of this important gene family.

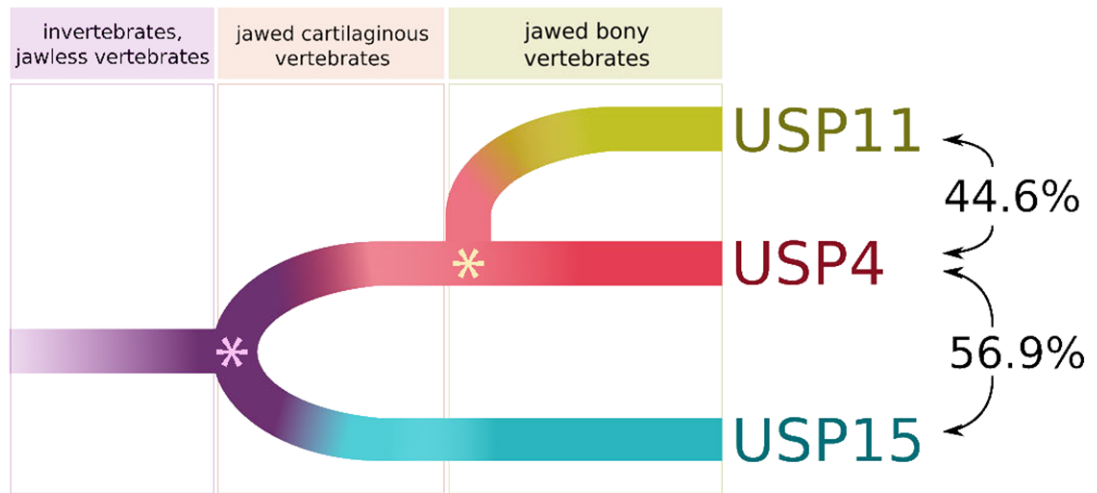


Figure 9. Phylogeny of USP genes. Percentages refer to amino acid identity between indicated USPs after global alignment. The pink asterisk denotes a whole genome duplication event, and the yellow asterisk denotes a small-scale duplication event involving USP4 and surrounding genes.

METHODS

Sequence retrieval

Identification and proper annotation of homologs in an array of species is a first essential step in studying the evolution of duplicated genes. Coding sequences were retrieved from GenBank [72] and from genome project databases [62, 73, 74] using the well-annotated human sequences for USP4, USP15 and USP11 as tBLASTn queries [75]. Reciprocal Best BLAST Hit (RBBH) annotation transfer was applied to unannotated genomes. Accession IDs for all sequences are below.

Species	USP4	USP15	USP11
Human	NC_000003	NC_000012	NC_000023
Gorilla	NC_018427	NC_018436	--
Chimpanzee	NC_006490	NC_006479	NC_006491
Rhesus monkey	NC_007859	NC_007868	NC_007878
Dog	NC_006602	NC_006592	NC_006621
Cat	NC_018724	NC_018729	NC_018741
Cow	AC_000179	AC_000162	AC_000187
Whale	NW_006725543	NW_006713252	NW_006727531
Mouse	NC_000075	NC_000076	NC_000086
Rat	NC_005107	NC_005106	NC_005120
Opossum	NC_008806	NC_008808	NC_008809
Little brown bat	NW_005872009	NW_005871371	NW_005871244
European shrew	NW_004545936	NW_004545859	NW_004545915
Star-nosed mole	NW_004567105	NW_004567135	NW_004567128
Armadillo	NW_004467831	NW_004502972	NW_004483933
Platypus	NW_001794469	NW_001688637	NW_001598857
Mallard	NW_004677124	NW_004676435	--
Zebra finch	NC_011476	NC_011463	--
Chicken	NC_006099	NC_006088	--
Alligator	--	NW_006225048	--
Python	NW_006532620	NW_006535771	NW_006532331
Green sea turtle	NW_006635848	NW_006644513	NW_006577128
Soft-shelled turtle	NW_005853649	NW_005858962	--
Anole	NC_014777	NC_014780	NC_014777
Xenopus	NW_004668239	NW_004668234	--
Coelacanth	NW_005819144	NW_005819645	NW_005821768
Gar	NC_023183	NC_023186	--
Zebrafish	NC_007117	NW_003336534	NC_007119
Zebra mbuna	NW_004531721	NW_004531746	NW_004531844
Guppy	NC_024335	NC_024353	NC_024337
Fugu	NC_018908	NC_018907	NC_018892
Medaka	XM_004068480	NW_004090515	NW_004095165
Shark	NW_006890068	NW_006890092	--
Lancelet	--	NW_003101526	--
Tunicate	--	NC_020166	--
Acorn worm	NW_003149765	NW_003123910	--
Sea urchin	NW_003577258	--	--
Sea slug	NW_004797520	--	--
Hydra	--	NW_004173592	--
Sea sponge	NW_003546314	--	--
Lamprey	lamprey_JL9400, lamprey_JL10812		
Little skate	LS-transcriptB2-ctg62960, LS-transcriptB2-ctg14739		
Owl limpet	gw1.79.7.1		

Table 1. List of coding sequences analyzed with corresponding accession numbers.

Sequence entropy

The site-wise Shannon entropy of aligned vertebrate USP amino acid sequences was calculated using DAMBE [24]. The results were plotted as a histogram using Circos [76].

GC content analysis

The seqinr package in R was employed to generate plots for the GC content of the third codon positions (GC3) of the *Homo sapiens* USP4, USP15 and USP11 coding sequences using a sliding window of width 10. A heatmap of GC3 content was generated using Circos [76].

Species tree reconstruction

A taxonomic phylogeny was generated using PhyloT[77]. Paralog affiliations were attributed as per their RBBH (described in *Sequence retrieval*). Putative non-processed pseudogene loci were confirmed using GenScan [78]. The SynMap function in CoGe [79] enabled comparison of the synteny of USP neighbouring regions in *Anolis carolinensis*, *Homo sapiens*, and *Gallus gallus*, which was visualized using Circos.

Phylogenetic analysis

USP4, USP15 and USP11 codon sequences were aligned using MUSCLE [22]. The maximum-likelihood method using estimated transition/transversion ratio and F84 model, as implemented in DAMBE [24], was used to derive a phylogeny rooted on *Amphimedon queenslandica*. Molecular clock analyses were also conducted using DAMBE [24].

Divergence dating

Codon alignments were produced by the MUSCLE algorithm using default GBlocks parameters as implemented in TranslatorX [22]. AIC and LRT nucleotide substitution model tests in DAMBE [24] designated the Generalized Time Reversible (GTR) as most

appropriate for all alignments. BEAST v.1.8.2 [80], a Bayesian Markov chain Monte Carlo (MCMC)-based phylogenetic dating program, was employed to quantify USP age-calibrated divergence times. All analyses used a log normal relaxed molecular clock. The USP4/USP11 analysis used six calibration points obtained from TimeTree [28] (in millions of years): Dog-Cow[USP4,USP11]: 60, Human-Opossum[USP4]: 112, Human-Anole[USP4, USP11]: 320 & Gar-Human[USP4]: 418. The RBM5/RBM10 analysis employed four pairs of calibration points: Human-Gorilla[RBM5,RBM10]: 8, Mouse-Rat[RBM5,RBM10]: 10.4, Human-Anole[RBM5,RBM10]: 320 & Human-Zebrafish [RBM5,RBM10]: 425. Tracer was used to verify similar convergence after 20 million steps for 3 runs in each case.

Mouse genetic crosses

TF2497 and TF2834 strains were purchased from Taconic Laboratories (Germantown, New York, USA), and were housed in a barrier facility at the University of Ottawa under protocol ME-305, approved by the Animal Care Committee, University of Ottawa. The strains were crossed to obtain mice heterozygous for proviral insertions in both the *Usp4* and *Usp15* genes. Eight pairs of compound heterozygous mice were mated under standard conditions, and progeny were obtained for genotyping. Genotyping was performed at 3-4 weeks of age, using tissue from ear punches. DNA was prepared using the REDExtract-N-Amp™ Tissue PCR Kit (Sigma-Aldrich Canada, Oakville, Ontario) and polymerase chain reaction was performed using the kit reagents. For genotyping of USP4 the forward primer used was derived from the third exon (upstream of the proviral insertion site): 5'- CCAGCAGCCTATTGTCAGAA -3', where reverse primers were derived from the third intron (downstream of the proviral insertion site): 5'-

TCAGTACTTAGGGATCTCTGA -3' or from the neomycin phosphotransferase gene within the provirus: 5' - AACCTGCGTGCAATCCATCT -3'. Amplification conditions for USP4 were as follows: initial denaturation at 95C for 3 min followed by 30 cycles of 95C for 30 sec, 57C for 30 sec and 72C for 60 sec and a final cycle at 72C for 5 min. A PCR product of approximately 250 base pairs was generated from the wild type gene, whereas the disrupted gene generated a product of approximately 1000 bp as detected by ethidium bromide staining of 1% agarose gels. For USP15 a similar strategy was adopted using the forward primer: 5' – GGT TTGAAGGATAACGTAGGC -3', and reverse primers 5' – ATAAACCCTCTTGCAGTTGCATC -3' and 5' - GAGTACCTAACAGGCACTTGAGACG -3'. USP15 PCR conditions were similar except that annealing was done at 55C for 30 seconds and elongation at 72C was reduced to 45 seconds.

ACKNOWLEDGEMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada through grant RGPIN-2015-05879 (awarded to DAG). Caitlyn Vlasschaert was the recipient of a Frederick Banting and Charles Best Scholarship from the Canadian Institutes of Health Research. We thank Dr. Stéphane Aris-Brosou for his time and guidance in troubleshooting the BEAST phylogenetic analyses. We are grateful to David Cook for his help with Circos, and acknowledge the kind assistance of Tim Ramsay in the statistical analysis of mouse crosses.

AVAILABILITY OF SUPPORTING DATA

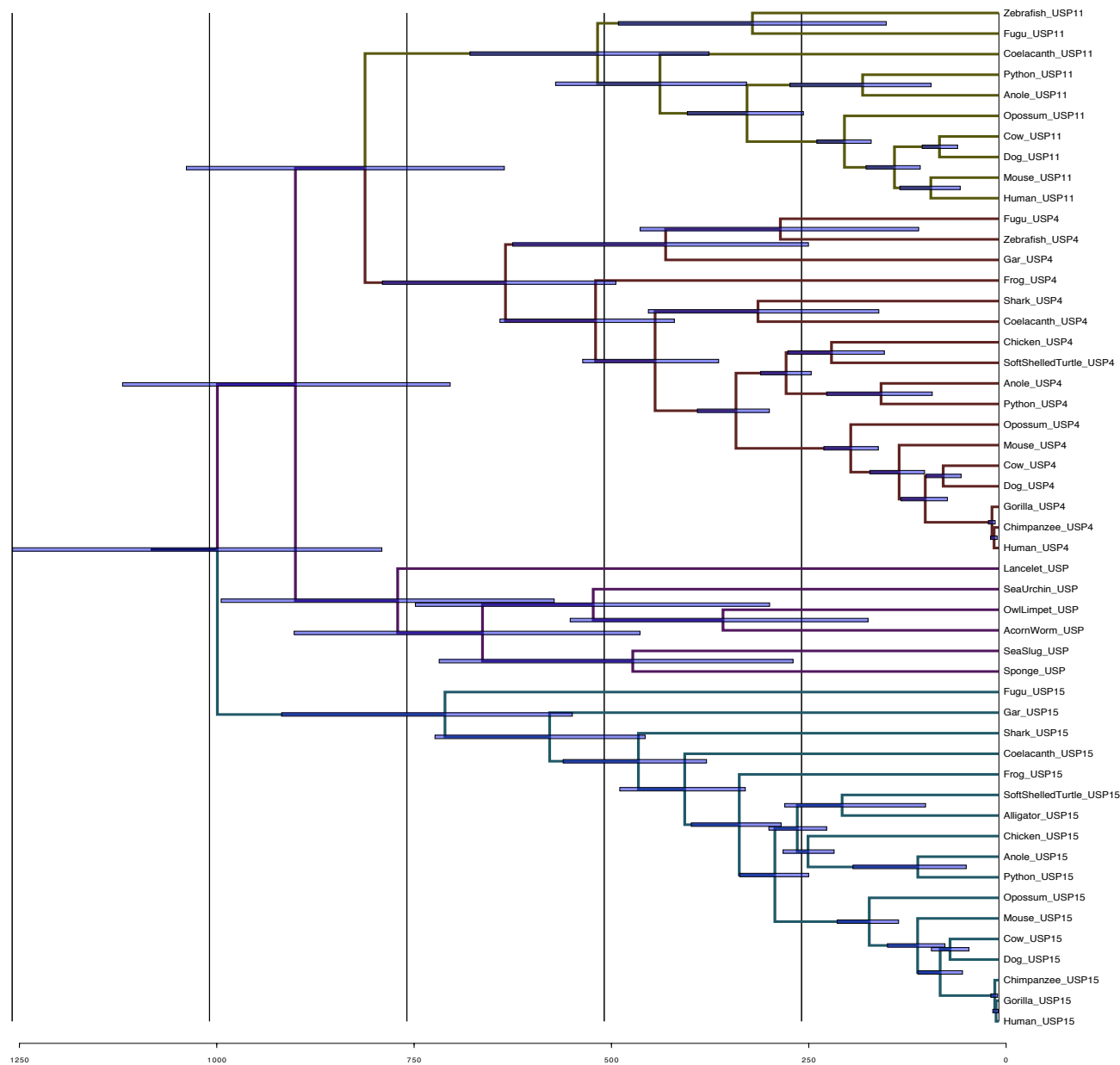
All supporting data are included as additional files in the form of Supplemental Figures 1 and 2.

COMPETING INTERESTS

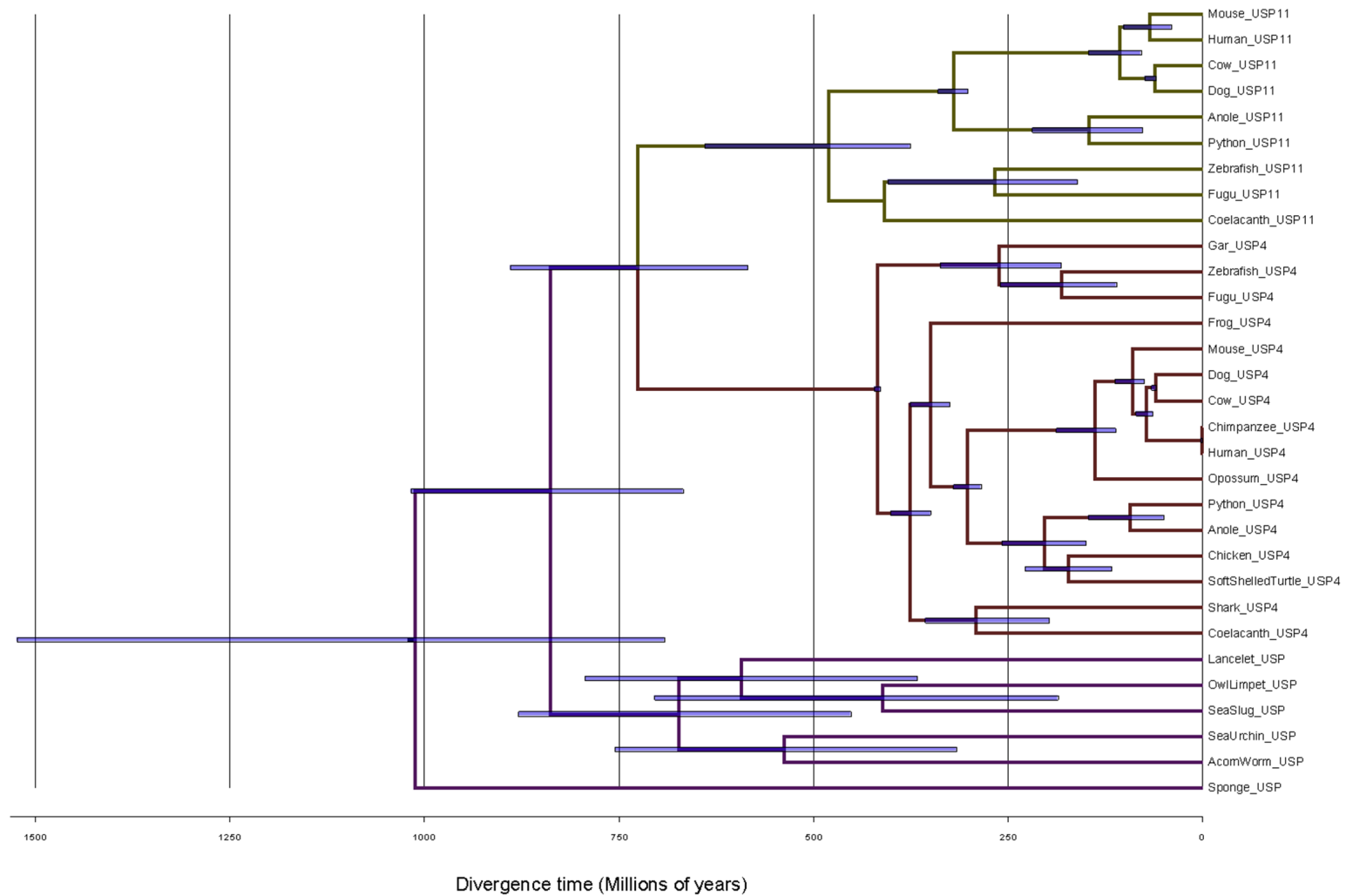
The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

CV contributed to the design of the study, carried out all computational analyses, and wrote the initial draft of the manuscript. XX contributed to the design of the study and advised on computational methodologies and the interpretation of data. JC performed mouse genetic crosses and contributed to the manuscript. DAG conceived of the study, participated in its design, and wrote sections of the manuscript. All authors read and approved the final manuscript.



Supplementary Figure 1. Bayesian dating of aligned codon sequences. Phylogenetic reconstruction and fossil-calibrated dating of aligned codon sequences for USP4, USP15, USP11 and ancestral USP sequences was generated using BEAST v. 1.8. 95% credible intervals are indicated. Calibration points were obtained from TimeTree.



Supplementary Figure 2. Bayesian dating of aligned USP4 and USP11 codon sequences. Phylogenetic reconstruction and fossil-calibrated dating of aligned codon sequences for USP4 and USP11 was generated using BEAST v. 1.8. 95% credible intervals are indicated. Calibration points were obtained from TimeTree. The gold star indicates the inferred USP4-USP11 divergence time. Red stars indicate major deviations from true topology

REFERENCES

1. Grabbe C, Husnjak K, Dikic I: **The spatial and temporal organization of ubiquitin networks.** *Nat Rev Mol Cell Biol* 2011, **12**:295–307.
2. Sowa ME, Bennett EJ, Gygi SP, Harper JW: **Defining the human deubiquitinating enzyme interaction landscape.** *Cell* 2009, **138**:389–403.
3. Aggarwal K, Massagué J: **Ubiquitin removal in the TGF- β pathway.** *Nat Cell Biol* 2012, **14**:656–657.
4. Eichhorn PJA, Rodón L, González-Juncà A, Dirac A, Gili M, Martínez-Sáez E, Aura C, Barba I, Peg V, Prat A, Cuartas I, Jimenez J, García-Dorado D, Sahuquillo J, Bernards R, Baselga J, Seoane J: **USP15 stabilizes TGF- β receptor I and promotes oncogenesis through the activation of TGF- β signaling in glioblastoma.** *Nat Med* 2012, **18**:429–435.
5. Zhao B, Schlesiger C, Masucci MG, Lindsten K: **The ubiquitin specific protease 4 (USP4) is a new player in the Wnt signalling pathway.** *J Cell Mol Med* 2009, **13**:1886–1895.
6. Fan Y-H, Yu Y, Mao R-F, Tan X-J, Xu G-F, Zhang H, Lu X-B, Fu S-B, Yang J: **USP4 targets TAK1 to downregulate TNF α -induced NF- κ B activation.** *Cell Death Differ* 2011, **18**:1547–1560.
7. Xiao N, Li H, Luo J, Wang R, Chen H, Chen J, Wang P: **Ubiquitin-specific protease 4 (USP4) targets TRAF2 and TRAF6 for deubiquitination and inhibits TNF α -induced cancer cell migration.** *Biochem J* 2012, **441**:979–986.
8. Schweitzer K, Bozko PM, Dubiel W, Naumann M: **CSN controls NF- κ B by deubiquitinylation of I κ B α .** *EMBO J* 2007, **26**:1532–1541.
9. Song EJ, Werner SL, Neubauer J, Stegmeier F, Aspden J, Rio D, Harper JW, Elledge SJ, Kirschner MW, Rape M: **The Prp19 complex and the Usp4Sart3 deubiquitinating enzyme control reversible ubiquitination at the spliceosome.** *Genes Dev* 2010, **24**:1434–1447.
10. Long L, Thelen JP, Furgason M, Haj-Yahya M, Brik A, Cheng D, Peng J, Yao T: **The U4/U6 recycling factor SART3 has histone chaperone activity and associates with USP15 to regulate H2B deubiquitination.** *J Biol Chem* 2014, **289**:8916–8930.
11. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock REW, Brinkman FSL, Lynn DJ: **InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation.** *Nucleic Acids Res* 2013, **41**:D1228–D1233.
12. Al-Salihi MA, Herhaus L, Macartney T, Sapkota GP: **USP11 augments TGF β signalling by deubiquitylating ALK5.** *Open Biol* 2012, **2**:120063.
13. Zhang L, Zhou F, Drabsch Y, Snaar-Jagalska E, Mickanin C, Huang H, Sheppard K-A, Lu C, Dijke P ten: **USP4 is regulated by Akt phosphorylation and deubiquitylates TGF-beta type I receptor.** *Nat Preced* 2012.

14. Zhang X, Berger FG, Yang J, Lu X: **USP4 inhibits p53 through deubiquitinating and stabilizing ARF-BP1.** *EMBO J* 2011, **30**:2177–2189.
15. Zou Q, Jin J, Hu H, Li HS, Romano S, Xiao Y, Nakaya M, Zhou X, Cheng X, Yang P, Lozano G, Zhu C, Watowich SS, Ullrich SE, Sun S-C: **USP15 stabilizes MDM2 to mediate cancer-cell survival and inhibit antitumor T cell responses.** *Nat Immunol* 2014, **15**:562–570.
16. Ke J, Dai C, Wu W, Gao J, Xia A, Liu G, Lv K, Wu C: **USP11 regulates p53 stability by deubiquitinating p53.** *J Zhejiang Univ Sci B* 2014, **15**:1032–1038.
17. Wu H-C, Lin Y-C, Liu C-H, Chung H-C, Wang Y-T, Lin Y-W, Ma H-I, Tu P-H, Lawler SE, Chen R-H: **USP11 regulates PML stability to control Notch-induced malignancy in brain tumours.** *Nat Commun* 2014, **5**.
18. Schoenfeld AR, Apgar S, Dolios G, Wang R, Aaronson SA: **BRCA2 Is Ubiquitinated In Vivo and Interacts with USP11, a Deubiquitinating Enzyme That Exhibits Prosurvival Function in the Cellular Response to DNA Damage.** *Mol Cell Biol* 2004, **24**:7444–7455.
19. Clague MJ, Barsukov I, Coulson JM, Liu H, Rigden DJ, Urbé S: **Deubiquitylases From Genes to Organism.** *Physiol Rev* 2013, **93**:1289–1315.
20. Elliott PR, Liu H, Pastok MW, Grossmann GJ, Rigden DJ, Clague MJ, Urbé S, Barsukov IL: **Structural variability of the ubiquitin specific protease DUSP-UBL double domains.** *FEBS Lett* 2011, **585**:3385–3390.
21. Jacq X, Kemp M, Martin NMB, Jackson SP: **Deubiquitylating Enzymes and DNA Damage Response Pathways.** *Cell Biochem Biophys* 2013, **67**:25–43.
22. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792–1797.
23. Castresana J: **Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis.** *Mol Biol Evol* 2000, **17**:540–552.
24. Xia X: **DAMBE5: A Comprehensive Software Package for Data Analysis in.** *Mol Biol Evol* 2013, **30**:1720–1728.
25. Vinogradov AE: **Isochores and tissue-specificity.** *Nucleic Acids Res* 2003, **31**:5212–5220.
26. Rodin SN, Parkhomchuk DV: **Position-associated GC asymmetry of gene duplicates.** *J Mol Evol* 2004, **59**:372–384.
27. Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H: **Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies.** *Syst Biol* 2007, **56**:389–399.
28. Hedges SB, Dudley J, Kumar S: **TimeTree: a public knowledge-base of divergence times among organisms.** *Bioinformatics* 2006, **22**:2971–2972.

29. Dehal P, Boore JL: **Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate.** *PLoS Biol* 2005, **3**.
30. Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, Ohta Y, Flajnik MF, Sutoh Y, Kasahara M, Hoon S, Gangu V, Roy SW, Irimia M, Korzh V, Kondrychyn I, Lim ZW, Tay B-H, Tohari S, Kong KW, Ho S, Lorente-Galdos B, Quilez J, Marques-Bonet T, Raney BJ, Ingham PW, Tay A, Hillier LW, Minx P, Boehm T, et al.: **Elephant shark genome provides unique insights into gnathostome evolution.** *Nature* 2014, **505**:174–179.
31. Nguyen Ba AN, Strome B, Hua JJ, Desmond J, Gagnon-Arsenault I, Weiss EL, Landry CR, Moses AM: **Detecting Functional Divergence after Gene Duplication through Evolutionary Changes in Posttranslational Regulatory Sequences.** *PLoS Comput Biol* 2014, **10**.
32. Clerici M, Luna-Vargas MPA, Faesen AC, Sixma TK: **The DUSP-Ubl domain of USP4 enhances its catalytic efficiency by promoting ubiquitin exchange.** *Nat Commun* 2014, **5**:5399.
33. Zhao B, Velasco K, Sompallae R, Pfirrmann T, Masucci MG, Lindsten K: **The ubiquitin specific protease-4 (USP4) interacts with the S9/Rpn6 subunit of the proteasome.** *Biochem Biophys Res Commun* 2012, **427**:490–496.
34. Hayes SD, Liu H, MacDonald E, Sanderson CM, Coulson JM, Clague MJ, Urbé S: **Direct and indirect control of mitogen-activated protein kinase pathway-associated components, BRAP/IMP E3 ubiquitin ligase and CRAF/RAF1 kinase, by the deubiquitylating enzyme USP15.** *J Biol Chem* 2012, **287**:43007–43018.
35. Harper S, Gratton HE, Cornaciu I, Oberer M, Scott DJ, Emsley J, Dreveny I: **Structure and Catalytic Regulatory Function of Ubiquitin Specific Protease 11 N-Terminal and Ubiquitin-like Domains.** *Biochemistry (Mosc)* 2014, **53**:2966–2978.
36. Harper S, Besong TMD, Emsley J, Scott DJ, Dreveny I: **Structure of the USP15 N-terminal domains: a β -hairpin mediates close association between the DUSP and UBL domains.** *Biochemistry (Mosc)* 2011, **50**:7995–8004.
37. Huang S, Chen Z, Yan X, Yu T, Huang G, Yan Q, Pontarotti PA, Zhao H, Li J, Yang P, Wang R, Li R, Tao X, Deng T, Wang Y, Li G, Zhang Q, Zhou S, You L, Yuan S, Fu Y, Wu F, Dong M, Chen S, Xu A: **Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes.** *Nat Commun* 2014, **5**.
38. Klinck R, Laberge G, Bisson M, McManus S, Michou L, Brown JP, Roux S: **Alternative splicing in osteoclasts and Paget's disease of bone.** *BMC Med Genet* 2014, **15**:98.
39. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M: **PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse.** *Nucleic Acids Res* 2012, **40**(Database issue):D261–D270.

40. Zanivan S, Gnad F, Wickström SA, Geiger T, Macek B, Cox J, Fässler R, Mann M: **Solid tumor proteome and phosphoproteome analysis by high resolution mass spectrometry.** *J Proteome Res* 2008, **7**:5314–5326.
41. Huttlín EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villén J, Haas W, Sowa ME, Gygi SP: **A tissue-specific atlas of mouse protein phosphorylation and expression.** *Cell* 2010, **143**:1174–1189.
42. Yu Y, Yoon S-O, Poulogiannis G, Yang Q, Ma XM, Villén J, Kubica N, Hoffman GR, Cantley LC, Gygi SP, Blenis J: **Phosphoproteomic analysis identifies Grb10 as an mTORC1 substrate that negatively regulates insulin signaling.** *Science* 2011, **332**:1322–1326.
43. Gnad F, Young A, Zhou W, Lyle K, Ong CC, Stokes MP, Silva JC, Belvin M, Friedman LS, Koeppen H, Minden A, Hoeflich KP: **Systems-wide analysis of K-Ras, Cdc42, and PAK4 signaling by quantitative phosphoproteomics.** *Mol Cell Proteomics MCP* 2013, **12**:2070–2080.
44. Lundby A, Secher A, Lage K, Nordsborg NB, Dmytriiev A, Lundby C, Olsen JV: **Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues.** *Nat Commun* 2012, **3**:876.
45. Nishi H, Fong JH, Chang C, Teichmann SA, Panchenko AR: **Regulation of protein-protein binding by coupling between phosphorylation and intrinsic disorder: analysis of human protein complexes.** *Mol Biosyst* 2013, **9**:1620–1626.
46. Trost M, Sauvageau M, Hérault O, Deleris P, Pomiès C, Chagraoui J, Mayotte N, Meloche S, Sauvageau G, Thibault P: **Posttranslational regulation of self-renewal capacity: insights from proteome and phosphoproteome analyses of stem cell leukemia.** *Blood* 2012, **120**:e17–27.
47. Zhou H, Di Palma S, Preisinger C, Peng M, Polat AN, Heck AJR, Mohammed S: **Toward a comprehensive characterization of a human cancer cell phosphoproteome.** *J Proteome Res* 2013, **12**:260–271.
48. Sharma K, D'Souza RCJ, Tyanova S, Schaab C, Wiśniewski JR, Cox J, Mann M: **Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling.** *Cell Rep* 2014, **8**:1583–1594.
49. Iyuke FO, Green JR, Willmore WG: **Active Learning for the Prediction of Asparagine/Aspartate Hydroxylation Sites on Proteins.** ACTAPRESS; 2011.
50. Dephoure N, Zhou C, Villén J, Beausoleil SA, Bakalarski CE, Elledge SJ, Gygi SP: **A quantitative atlas of mitotic phosphorylation.** *Proc Natl Acad Sci U S A* 2008, **105**:10762–10767.
51. Gauci S, Helbig AO, Slijper M, Krijgsveld J, Heck AJR, Mohammed S: **Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach.** *Anal Chem* 2009, **81**:4493–4501.

52. Olsen JV, Vermeulen M, Santamaria A, Kumar C, Miller ML, Jensen LJ, Gnad F, Cox J, Jensen TS, Nigg EA, Brunak S, Mann M: **Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis**. *Sci Signal* 2010, **3**:ra3.
53. Christensen GL, Kelstrup CD, Lyngsø C, Sarwar U, Bøgebo R, Sheikh SP, Gammeltoft S, Olsen JV, Hansen JL: **Quantitative phosphoproteomics dissection of seven-transmembrane receptor signaling using full and biased agonists**. *Mol Cell Proteomics MCP* 2010, **9**:1540–1553.
54. Phanstiel DH, Brumbaugh J, Wenger CD, Tian S, Probasco MD, Bailey DJ, Swaney DL, Tervo MA, Bolin JM, Ruotti V, Stewart R, Thomson JA, Coon JJ: **Proteomic and phosphoproteomic comparison of human ES and iPS cells**. *Nat Methods* 2011, **8**:821–827.
55. Bian Y, Song C, Cheng K, Dong M, Wang F, Huang J, Sun D, Wang L, Ye M, Zou H: **An enzyme assisted RP-RPLC approach for in-depth analysis of human liver phosphoproteome**. *J Proteomics* 2014, **96**:253–262.
56. Van Hoof D, Muñoz J, Braam SR, Pinkse MWH, Linding R, Heck AJR, Mummery CL, Krijgsveld J: **Phosphorylation dynamics during early differentiation of human embryonic stem cells**. *Cell Stem Cell* 2009, **5**:214–226.
57. Kettenbach AN, Schweppe DK, Faherty BK, Pechenick D, Pletnev AA, Gerber SA: **Quantitative phosphoproteomics identifies substrates and functional modules of Aurora and Polo-like kinase activities in mitotic cells**. *Sci Signal* 2011, **4**:rs5.
58. Kurotani A, Tokmakov AA, Kuroda Y, Fukami Y, Shinozaki K, Sakurai T: **Correlations between predicted protein disorder and post-translational modifications in plants**. *Bioinformatics* 2014, **30**:1095–1103.
59. Vuzman D, Hoffman Y, Levy Y: **Modulating protein-DNA interactions by post-translational modifications at disordered regions**. *Pac Symp Biocomput Pac Symp Biocomput* 2012:188–199.
60. Dyson HJ, Wright PE: **Intrinsically unstructured proteins and their functions**. *Nat Rev Mol Cell Biol* 2005, **6**:197–208.
61. Ma B, Nussinov R: **Regulating highly dynamic unstructured proteins and their coding mRNAs**. *Genome Biol* 2009, **10**:204.
62. Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE, Morgan JR, Buxbaum JD, Sachidanandam R, Sims C, Garruss AS, Cook M, Krumlauf R, Wiedemann LM, Sower SA, Decatur WA, Hall JA, Amemiya CT, Saha NR, Buckley KM, Rast JP, Das S, Hirano M, McCurley N, Guo P, Rohner N, et al.: **Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution**. *Nat Genet* 2013, **45**:415–421, 421e1–2.

63. Mehta TK, Ravi V, Yamasaki S, Lee AP, Lian MM, Tay B-H, Tohari S, Yanai S, Tay A, Brenner S, Venkatesh B: **Evidence for at least six Hox clusters in the Japanese lamprey (*Lethenteron japonicum*)**. *Proc Natl Acad Sci* 2013, **110**:16044–16049.
64. Nah GSS, Tay B-H, Brenner S, Osato M, Venkatesh B: **Characterization of the Runx Gene Family in a Jawless Vertebrate, the Japanese Lamprey (*Lethenteron japonicum*)**. *PLoS ONE* 2014, **9**:e113445.
65. Fares MA, Keane OM, Toft C, Carretero-Paulet L, Jones GW: **The Roles of Whole-Genome and Small-Scale Duplications in the Functional Specialization of *Saccharomyces cerevisiae* Genes**. *PLoS Genet* 2013, **9**:e1003176.
66. Kocabas AM, Crosby J, Ross PJ, Otu HH, Beyhan Z, Can H, Tam W-L, Rosa GJM, Halgren RG, Lim B, Fernandez E, Cibelli JB: **The transcriptome of human oocytes**. *Proc Natl Acad Sci* 2006, **103**:14027–14032.
67. Zimmermann P, Hennig L, Gruissem W: **Gene-expression analysis and network discovery using Genevestigator**. *Trends Plant Sci* 2005, **10**:407–409.
68. Chen J, Torcia S, Xie F, Lin C-J, Cakmak H, Franciosi F, Horner K, Onodera C, Song JS, Cedars MI, Ramalho-Santos M, Conti M: **Somatic cells regulate maternal mRNA translation and developmental competence of mouse oocytes**. *Nat Cell Biol* 2013, **15**:1415–1423.
69. Dragovic RA, Ritter LJ, Schulz SJ, Amato F, Thompson JG, Armstrong DT, Gilchrist RB: **Oocyte-secreted factor activation of SMAD 2/3 signaling enables initiation of mouse cumulus cell expansion**. *Biol Reprod* 2007, **76**:848–857.
70. Elvin JA, Yan C, Matzuk MM: **Oocyte-expressed TGF-beta superfamily members in female fertility**. *Mol Cell Endocrinol* 2000, **159**:1–5.
71. Jackowska M, Kempisty B, Woźna M, Piotrowska H, Antosik P, Zawierucha P, Bukowska D, Nowicki M, Jaśkowski JM, Brüssow K-P: **Differential expression of GDF9, TGFβ1, TGFβ2 and TGFβ3 in porcine oocytes isolated from follicles of different size before and after culture in vitro**. *Acta Vet Hung* 2013, **61**:99–115.
72. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2013, **41**(Database issue):D36–42.
73. Simakov O, Marletaz F, Cho S-J, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo D-H, Larsson T, Lv J, Arendt D, Savage R, Osoegawa K, de Jong P, Grimwood J, Chapman JA, Shapiro H, Aerts A, Otillar RP, Terry AY, Boore JL, Grigoriev IV, Lindberg DR, Seaver EC, Weisblat DA, Putnam NH, Rokhsar DS: **Insights into bilaterian evolution from three spiralian genomes**. *Nature* 2013, **493**:526–531.
74. Wyffels J, L. King B, Vincent J, Chen C, Wu CH, Polson SW: **SkateBase, an elasmobranch genome project and collection of molecular resources for chondrichthyan fishes**. *F1000Research* 2014, **3**.

75. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.
76. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: An information aesthetic for comparative genomics.** *Genome Res* 2009.
77. Letunic I, Bork P: **Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.** *Nucleic Acids Res* 2011, **39**(suppl 2):W475–W478.
78. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**:78–94.
79. Lyons E, Pedersen B, Kane J, Freeling M: **The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids.** *Trop Plant Biol* 2008, **1**:181–190.
80. Drummond AJ, Suchard MA, Xie D, Rambaut A: **Bayesian Phylogenetics with BEAUti and the BEAST 1.7.** *Mol Biol Evol* 2012, **29**:1969–1973.
81. Herhaus L, Al-Salihi MA, Dingwell KS, Cummins TD, Wasmus L, Vogt J, Ewan R, Bruce D, Macartney T, Weidlich S, Smith JC, Sapkota GP: **USP15 targets ALK3/BMPRI1A for deubiquitylation to enhance bone morphogenetic protein signalling.** *Open Biol* 2014, **4**:140065.
82. Blanchette P, Gilchrist CA, Baker RT, Gray DA: **Association of UNP, a ubiquitin-specific protease, with the pocket proteins pRb, p107 and p130.** *Oncogene* 2001, **20**:5533–5537.
83. Villeneuve NF, Tian W, Wu T, Sun Z, Lau A, Chapman E, Fang D, Zhang DD: **USP15 negatively regulates Nrf2 through deubiquitination of Keap1.** *Mol Cell* 2013, **51**:68–79.
84. Wang L, Zhao W, Zhang M, Wang P, Zhao K, Zhao X, Yang S, Gao C: **USP4 positively regulates RIG-I-mediated antiviral response through deubiquitination and stabilization of RIG-I.** *J Virol* 2013, **87**:4507–4515.
85. Pauli E-K, Chan YK, Davis ME, Gableske S, Wang MK, Feister KF, Gack MU: **The ubiquitin-specific protease USP15 promotes RIG-I-mediated antiviral signaling by deubiquitylating TRIM25.** *Sci Signal* 2014, **7**:ra3.
86. Di Donato F, Chan EK, Askanase AD, Miranda-Carus M, Buyon JP: **Interaction between 52 kDa SSA/Ro and deubiquitinating enzyme UnpEL: a clue to function.** *Int J Biochem Cell Biol* 2001, **33**:924–934.
87. Hou X, Wang L, Zhang L, Pan X, Zhao W: **Ubiquitin-specific protease 4 promotes TNF- α -induced apoptosis by deubiquitination of RIP1 in head and neck squamous cell carcinoma.** *FEBS Lett* 2013, **587**:311–316.

88. Yamaguchi T, Kimura J, Miki Y, Yoshida K: **The deubiquitinating enzyme USP11 controls an IkappaB kinase alpha (IKKalpha)-p53 signaling pathway in response to tumor necrosis factor alpha (TNFalpha).** *J Biol Chem* 2007, **282**:33943–33948.
89. Sun W, Tan X, Shi Y, Xu G, Mao R, Gu X, Fan Y, Yu Y, Burlingame S, Zhang H, Rednam SP, Lu X, Zhang T, Fu S, Cao G, Qin J, Yang J: **USP11 negatively regulates TNFalpha-induced NF-kappaB activation by targeting on IkappaBalpha.** *Cell Signal* 2010, **22**:386–394.
90. Huang X, Langelotz C, Hetfeld-Pechoc BKJ, Schwenk W, Dubiel W: **The COP9 signalosome mediates beta-catenin degradation by deneddylation and blocks adenomatous polyposis coli destruction via USP15.** *J Mol Biol* 2009, **391**:691–702.
91. Milojevic T, Reiterer V, Stefan E, Korkhov VM, Dorostkar MM, Ducza E, Ogris E, Boehm S, Freissmuth M, Nanoff C: **The ubiquitin-specific protease Usp4 regulates the cell surface level of the A2A receptor.** *Mol Pharmacol* 2006, **69**:1083–1094.
92. Maertens GN, Messaoudi-Aubert S El, Elderkin S, Hiom K, Peters G: **Ubiquitin-specific proteases 7 and 11 modulate Polycomb regulation of the INK4a tumour suppressor.** *EMBO J* 2010, **29**:2553–2565.

CHAPTER II

Selection preserves Ubiquitin Specific Protease 4 alternative exon skipping in therian mammals

This paper was published in Scientific Reports: Vlasschaert, C., Xia X., Gray D.

Selection preserves Ubiquitin Specific Protease 4 alternative exon skipping in therian mammals. *Sci. Rep.* **6**, 20039; doi: 10.1038/srep20039 (2016).

The student performed the majority of the bioinformatics analyses, devised the basis for the wet lab experiments and performed cell transfections for Figure 6, prepared the figures and wrote a first draft of the manuscript, which was edited by the student and her co-supervisors.

**Selection preserves Ubiquitin Specific Protease 4 alternative exon skipping in
therian mammals**

Caitlyn Vlasschaert^{1,3,4}, Xuhua Xia^{2,3,*}, and Douglas A. Gray^{1,4,*}

¹ Ottawa Hospital Research Institute, Ottawa, K1H 8L6, Canada

² Department of Biology, University of Ottawa, Ottawa, K1N 6N5, Canada]

³ Ottawa Institute of Systems Biology, Ottawa, K1H 8M5, Canada

⁴ Department of Biochemistry, Microbiology and Immunology, University of Ottawa,
Ottawa, K1H 8M5, Canada

corresponding authors: xxia@uottawa.ca, dgray@ohri.ca 613-737-7700 ext. 70331

keywords: alternative splicing, exon skipping, genetic drift

ABSTRACT

Ubiquitin specific protease 4 (USP4) is a highly networked deubiquitinating enzyme with reported roles in cancer, innate immunity and RNA splicing. In mammals it has two dominant isoforms arising from inclusion or skipping of exon 7 (E₇). We evaluated two plausible mechanisms for the generation of these isoforms: (A) E₇ skipping due to a long upstream intron and (B) E₇ skipping due to inefficient 5' splice sites (5'SS) and/or branchpoint sites (BPS). We then assessed whether E₇ alternative splicing is maintained by selective pressure or arose from genetic drift. Both transcript variants were generated from a USP4-E₇ minigene construct with short flanking introns, an observation consistent with the second mechanism whereby differential splice signal strengths are the basis of E₇ skipping. Optimization of the downstream 5'SS eliminated E₇ skipping. Experimental validation of the correlation between 5'SS identity and exon skipping in vertebrates pinpointed the +6 site as the key splicing determinant. Therian mammals invariably display a 5'SS configuration favouring alternative splicing and the resulting isoforms have distinct subcellular localizations. We conclude that alternative splicing of mammalian USP4 is under selective maintenance and that long and short USP4 isoforms may target substrates in various cellular compartments.

INTRODUCTION

Ubiquitin-specific protease 4 (USP4) is a deubiquitinating enzyme that can edit or remove ubiquitin chains of various topologies. USP4 can remove both degradative K48- and regulatory K63-linked ubiquitin chains from a growing list of protein targets that include key players in a number of signal transduction pathways. Its substrates include the transforming growth factor- β (TGF- β) I receptor [1,2] and the TGF- β -activated kinase 1 (TAK1) [3], Wnt/ β -catenin pathway transcription factor Nemo-like kinase (NLK) [4], the p53 antagonist ARF-BP1 [5], anti-viral response mediator RIG-I [6], TNF α /NF- κ B inflammatory pathway mediators TRAF-2 & -6 [7,8], and the master growth factor signalling kinase PDK1 [9]. Homeostatic regulation of USP4 deubiquitination targets is provided by the opposing actions of E3 ubiquitin ligase enzymes that promote degradation or complex assembly.

USP4 has both nuclear importing and exporting signal motifs and can shuttle between nucleus and cytoplasm [10]. Subcellular localization is partially controlled by AKT-mediated phosphorylation of the serine residue at position 445, which promotes cytoplasmic retention [2]. Nuclear localization, coupled with the observation that USP4 interacts with at least eleven splicing factors [11] including Prp3 [12], suggests that USP4 regulates spliceosomal activity. This is particularly interesting since multiple spliced variants of USP4 have been reported in human. The two major isoforms differ by the inclusion of Exon 7 (E₇; Figure 1), though neither differential activities nor specificities have been attributed to the resulting long and short isoforms since their discovery [13,14] shortly after that of the gene itself [15]. The two isoforms (with or without E₇) both produce functional USP4 proteins in deubiquitination assays [14]. E₇ forms the major

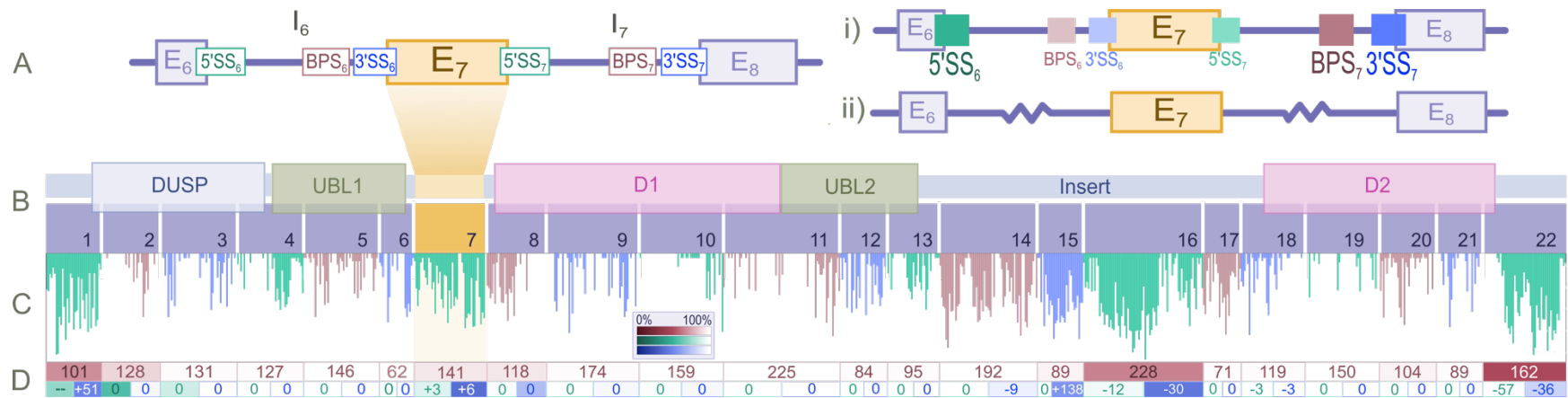


Figure 1. Schema of *Usp4* exonic properties and potential mechanisms of exon skipping.

A) Proposed alternative mechanisms for skipping of exon 7 in USP4:

- i.* differential splice site strengths
- ii.* long flanking introns

B) *Usp4* exon structure relative to encoded protein domains. *D1* and *D2* form the cysteine protease catalytic domain that effectuates ubiquitin cleavage. The regulatory *DUSP-UBL1* domain physically interacts with the unstructured *Insert* region and with *DUSP-UBL1* domains of other USP4 monomers leading to dimerization.

C) USP4 protein sequence entropy. USP4 exon-retained isoform amino acid sequences for 62 species were aligned using MUSCLE [51]. Shannon entropy was calculated using DAMBE [50] and plotted using the ggplot2 [52] package in R.

D) Exon length variability. Mode of length per exon (in nucleotides) for 37 mammals indicated in pink boxes. Mode of 33 birds and 24 fish exon lengths relative to mammalian mode (codon indels) indicated in green and blue boxes, respectively. Shading of boxes indicates fraction of total sequences represented by the mode, where darkening is commensurate with increasing variability (see inset legend above).

part of the flexible linker region that enables the critical interaction of the neighbouring DUSP-UBL1 and insert domains of USP4 (Figure 1B). In the E₇-skipped (short) isoform, the linker is shortened from sixty-seven to twenty amino acids though retains sufficient flexibility to enable comparable domain interaction [16]. Despite the proposed facultative role of this alternatively spliced exon[16], the majority of conserved phosphorylation sites in USP4 are within E₇, suggesting the protein isoforms may be differently regulated [17]. What is more, the longer variant is the principal isoform and the relative amount of exon-skipped transcripts is consistent within but different among tissue types, suggesting alternative splicing may be regulated [14]. The mechanism for E₇ skipping in human USP4 mRNA is not known, nor are the functional implications of this alternative splicing.

Exon skipping is the most frequent form of alternative splicing and accounts for about 40% of all alternative splicing in higher eukaryotes [18-20]. A number of exon skipping mechanisms have been proposed, including:

1. differential strength of splice signals leading to alternative splice site selection (reviewed in Keren et al. [21])
2. excessively long flanking introns [22-27]
3. intermixture of major and minor class introns [28].

As USP4 does not have minor class introns, we chose to examine the first two mechanisms of E₇ skipping, as illustrated in Figure 1A i) and ii). First, if an exon is functionally important, then flanking sequences recognized by splicing factors will be under strong purifying selection to maintain proper spliced end-products. If the inclusion of E₇ is unimportant, then its proximal splice sites (3'SS₆, 5'SS₇, and BPS₆) may

experience greater sequence drift, leading to weakening of these sites relative to 5'SS₆, 3'SS₇ and BPS₇ and exon 7 skipping (Figure 1Ai). Selection can however also act to preserve differential splice signal strengths in order to enable the production of two isoforms. In short, strong 5'SS₆, 3'SS₇ and BPS₇ relative to 5'SS₇, 3'SS₆ and BPS₆, resulting from passive drift or active selection, could permit E₇ skipping. Second, exons flanked by long introns are known to be intermittently skipped. As such, the alternative inclusion of E₇ in USP4 mRNA could be explained by appropriately long flanking introns (Figure 1Aii). We evaluated these alternative hypotheses by constructing a computational evolutionary framework to characterize USP4 splicing patterns in multiple vertebrate lineages and tested our *in silico* predictions in the laboratory to gauge whether the two isoforms of USP4 observed in humans have discrete functional and/or regulatory roles or whether they are the by-product of reduced selection for E₇ retention.

Results

Bioinformatic analysis

To study exon-intron architecture and splice signal conservation, we downloaded GenBank sequences for well-annotated USP4 sequences for 62 vertebrate species covering major vertebrate taxa (see *Supplementary Table S1*).

Sequence and length conservation of USP4 exons

Both long and short isoforms of USP4 comprise a bi-partite catalytic domain and a regulatory DUSP-UBL1 domain, where the seventh exon forms part of the unstructured flexible linker between these two (Figure 1B). Sequence identity (Figure 1C) and length (Figure 1D) of USP4 exons are highly conserved for exons that encode structured functional domains and less well for exons that correspond to unstructured regions. E₇

exhibits greater length variation than its neighbors. Relative to mammals, E₇ of birds and fish generally encodes one and two additional amino acids respectively, while the length of E₇ within each of these clades is variable; this suggests multiple indel events in USP4 during the evolutionary diversification of vertebrates. Sequence conservation among aligned USP4 exons, quantified using Shannon entropy in Figure 1C, reveals that E₇ is located within a highly variable region. The entropic nature of E₇ conforms to the general pattern that alternatively spliced exons exhibit less sequence conservation [29], which can extend to exon-intron boundaries and thus weaken the signals of proximal splice sites. However, high entropy is also observed in all other USP4 exons coding for unstructured regions (i.e., exons 1, 14, 15, 16, 17, 22), which are not alternatively spliced. The disordered nature of the unstructured region encoded by E₇ could simply permit its sequence variability and the splice sites may not be affected. To determine whether E₇ neighboring splice site signals are weakened, we quantified their strengths relative to the optimal sequences recognized by the spliceosome.

Relative strengths of branchpoint site signals

The signal strength of a branch point site (BPS) in vertebrates is generally contingent upon the presence of a YURAY motif located at an optimal distance from 3' end of the intron ($D_{to3'.opt}$) followed by a pyrimidine-rich tract and the AG dinucleotide of the 3'SS signal. $D_{to3'.opt}$ can be revealed by mapping the intronic locations of several BPS motifs and deriving an optimal range of values for these. A previous analysis of 397 human housekeeping genes reported $D_{to3'.opt}$ to be 21–34 nt, representing 83% of introns in these highly expressed genes [30]. In Figure 2A, we have derived a more sensitive $D_{to3'.opt}$ of 20-40 nt based on the locations of consensus BPS sequences (YURAY) within

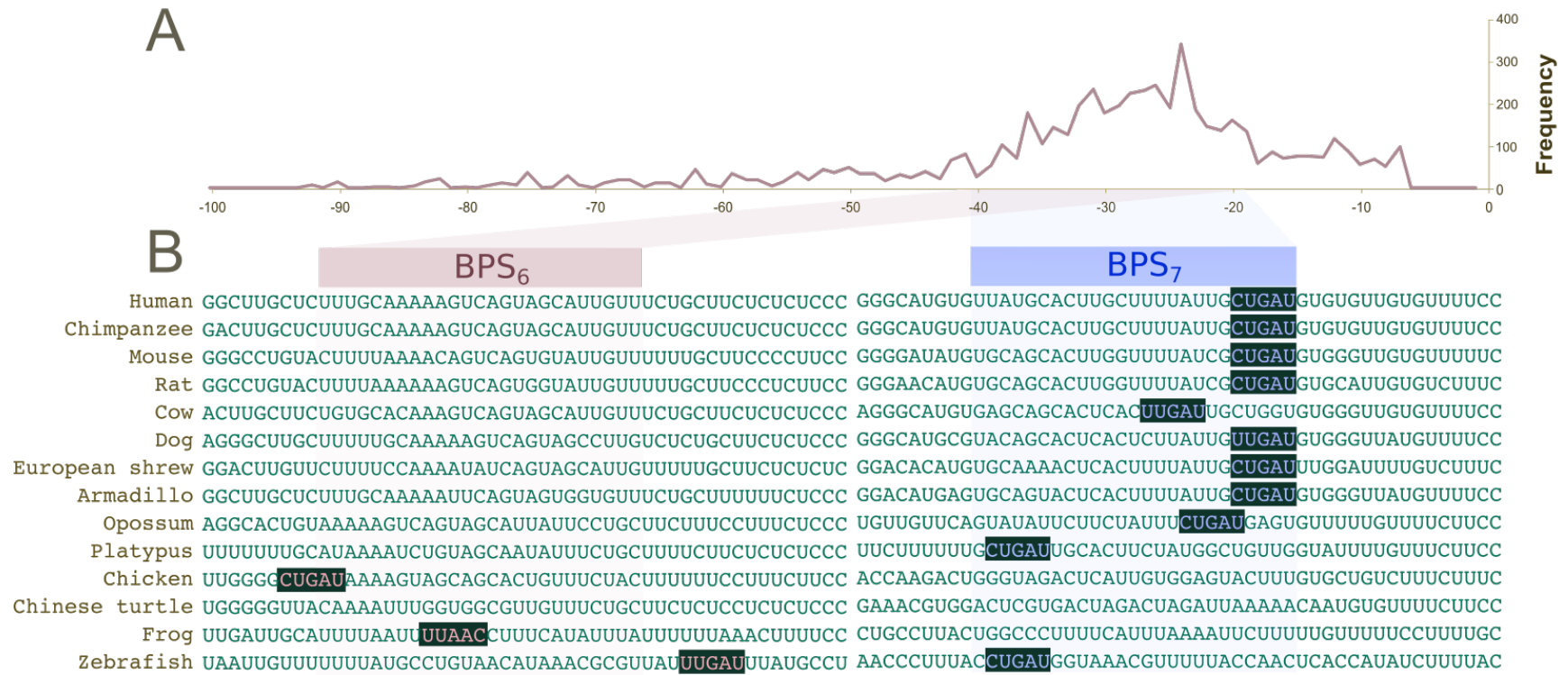


Figure 2. Characterization of relative BPS strengths.

A) Distribution of location of strong BPS sequences ("YURAY" motif with single downstream AG dinucleotide corresponding to the 3'SS) in all introns of human chromosome 22. The X-axis depicts the last 200 nt of the introns. Introns shorter than 200 nt were excluded to avoid statistical bias. $D_{to3'.opt}$ is 20-40 nt.

B) Locations of YURAY sequences within 50nt of 3' ends of introns 6 and 7 of USP4 in 14 vertebrate species.

all 15770 introns of human chromosome 22 protein-coding genes. This slightly larger window for $D_{to3'}.opt$ will reduce the false negative rate in detecting true BPSs in the neighbouring introns of E_7 . If a strong BPS is absent from the upstream intron (I_6), a strong downstream BPS within I_7 could lead to exon skipping (Figure 1A). We therefore evaluated the relative strengths of USP4 BPS₆ and BPS₇ for 14 well-studied species representing the major vertebrate taxa.

The location of YURAY motifs in I_7 (Figure 2B) suggests a strong BPS₇ among all studied mammalian species. A strong BPS₇ is also present in I_7 of zebrafish (*Danio rerio*), but is missing in the frog (*Xenopus tropicalis*), the chicken (*Gallus gallus*), and the Chinese turtle (*Pelodiscus sinensis*). The latter two species retain candidate YURAY motifs which can be eliminated from consideration due to their greater distance from the 3'SS (73nt for the chicken, 85 nt for the Chinese turtle) and to the presence of a downstream non-3'SS AG (non-3'SS AG dinucleotides between a BPS and a 3'SS AG are avoided as the first AG following the branch-point is generally used as the 3'SS AG).

YURAY motifs are absent from the 40 last nucleotides of USP4- I_6 in all mammalian species. Taken together, mammalian species in general have a weak BPS₆ and a relatively strong BPS₇. This lends plausibility to the first hypothesis from Fig. 1A, wherein E_7 splicing results from alternative pairing of 5'SS₆ and BPS₇ (E_7 skipping) or of 5'SS₇ and BPS₇ (E_7 inclusion). In contrast, non-mammalian species tend to have a strong BPS₆ (except for the Chinese turtle). YURAY motifs are located near $D_{to3'}.opt$ in I_6 of the chicken ($D_{to3'} = 47$ nt) and zebrafish ($D_{to3'} = 16$ nt) and frog (three YURAY motifs with $D_{to3'} = 36, 42$ and 53 nt, respectively). This suggests that E_7 skipping should be less likely

in non-mammalian than in mammalian species, though the Chinese turtle may be an exception.

Relative strengths of proximal and distal splice signals

In addition to a weak BPS₆ relative to BPS₇, a stronger 5'SS₆ than 5'SS₇ and/or stronger 3'SS₇ than 3'SS₆ would favor E₇ skipping (Figure 1Ai). Position weight matrices (PWM) measures site-specific nucleotide usage bias in a motif alignment, where the consensus motif typically has the highest PWM score (PWMS). PWMSs are routinely used to characterize the signal strength of splice sites [31,32].

A PWM for the 5'SS sequences of human chromosome 22 introns (Table 1) shows a consensus 5'SS consistent with what has been documented in the literature, i.e., a core motif of AG|GUAAGU, where “|” indicates the exon-intron junction. PWMSs derived from zebrafish and chicken chromosomes are almost identical to the human 5'SS matrix in Table 1A. Thus, this PWM can be used to generate PWMSs as comparable measures of signal strength at 5'SS₆ and 5'SS₇ from the 14 representative vertebrate USP4 sequences, where significantly larger scores indicate stronger splice signal strength. Given Table 1A, the maximal PWMS is 16.9. PWMSs are consistently larger for 5'SS₆ than 5'SS₇ for the ten mammalian USP4 sequences (Table 1B), with the mean PWMS being 9.2314 for 5'SS₆ and 5.455 for 5'SS₇ ($t = 19.759$, $df = 11$, $p < 0.0001$, paired-sample t-test). This lends support for the scenario depicted in Figure 1Ai where a stronger 5'SS₆ favors E₇ skipping in mammals. In contrast, for chicken and zebrafish, PWMS₆ is greater than PWMS₇ while *Pelodiscus sinensis* again conforms to the mammalian pattern PWMS₆ > PWMS₇. There were no significant differences in flanking

Site	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8
A	0.2082	0.1306	0.3720	1.4204	-1.5830	-8.3696	-8.9985	1.0932	1.4771	-1.9859	-0.7113	0.0161	-0.5293
C	0.5104	0.6957	1.0657	-0.9485	-2.3528	-10.4401	-5.2125	-2.4697	-1.0419	-1.4937	0.0111	0.3350	0.6644
G	-0.7793	-0.4697	-0.9095	-1.4947	1.2021	1.4713	-9.4051	0.3050	-1.2529	1.1711	-0.6257	-0.0225	-0.3267
U	0.2021	-0.3378	-1.1008	-0.5682	-1.9475	-10.5499	2.2380	-3.5486	-1.0776	-1.8007	1.0212	-0.3701	0.2039
Species		5'SS ₆				PWMS ₆			5'SS ₇				PWMS ₇
Human		UCAAA GUAAGUGA				8.6625			AGGGG GUAAGAGC				5.0898
Chimpanzee		UCAAA GUAAGUCA				9.0200			AGGGG GUAAGAGC				5.0898
Mouse		UCAAA GUAAGUGA				9.5187			AGGGG GUAAGAAC				5.5747
Rat		UCAAA GUAAGUGA				9.3740			AGGGG GUAAGAAC				5.5747
Cow		UCAAA GUAAGUGA				9.1264			AGGGG GUAAGAGC				5.0898
Dog		UCAAA GUAAGUGA				9.1264			AGGGG GUAAGAGC				5.5361
Shrew		ACAAA GUAAGUGG				9.3351			CGGGG GUAAGCGC				6.5608
Armadillo		CCAAA GUAAGUCA				9.7922			AGGGG GUAAGAGC				5.5361
Opossum		UCCAA GUAAGUGA				9.9648			AGGGG GUAAGAGC				4.6637
Platypus		UCAAA GUAAGGGA				8.3157			CGGGG GUAAGAGC				4.8212
Chicken		GCAAA GUAAGAGU				5.6036			AGGGG GUAAGAGC				5.9924
Zebrafish		UCAAA GUAUGUCC				7.9026			AACAG GUCAGCUU				8.1978
Turtle		ACAAA GUAAGCAU				10.6407			GGGGG GUAAGACC				3.8889
Frog		ACAAA GUGAGUAU				8.5266			AAAGG GUAACUAU				6.4370

Table 1.

- A)** Position weight matrix (PWM) for 15770 exon-intron junctions (5'SS) from human chromosome 22. The consensus sequence (CAG|GUAAGU) is shown in bold. The five exonic nucleotides are labeled as -5 to -1 and the 8 intronic nucleotides as 1 to 8.
- B)** Position weight matrix scores (PWMS₆ and PWMS₇, respectively) for the 5' splice sites of introns 6 and 7 (5'SS₆ and 5'SS₇ in Fig. 1Ai) from 14 vertebrate species according to the scoring matrix in Table 1A.

3'SS PWMSs (as might be expected since 5'SS and BPS are most important for determining exon-intron boundaries [33]).

Length of the introns flanking exon 7

Because exons flanked by long introns tend to be skipped during the splicing process [22-27], we have examined whether E_7 is flanked by long introns. As observed in Figure 3, I_6 varies dramatically in length between clades. The average length of I_6 is the largest amongst all introns for mammals ([2733,17911] nt) and second largest for birds ([3903,15190] nt; second to I_{13}). Whereas I_6 is also relatively long in cartilaginous fish, the earliest-diverging clade presenting USP4 (shark = 4961 nt), it is contrastingly short in bony fish ([96,310] nt). Altogether this suggests a large number of indels during the evolution of different vertebrate lineages, likely the most amongst all introns of USP4.

Although exons with long flanking introns tend to be lost in the final mRNA [22-27], there is discrepancy concerning the relative effect of upstream and downstream introns. The detailed experimental study on CD44 [22] shows that exon skipping occurs only when the exon is flanked on both sides by long introns, and that the effect of the two introns appears to be symmetrical. However, subsequent studies on the relationship between intron length and exon skipping [23,24] are not always consistent with these earlier findings. In particular, it appears as though the upstream intron has a greater effect on exon skipping than the downstream intron [25], which would suggest that the potential contribution of the long I_6 to E_7 skipping in mammals should not be ignored.

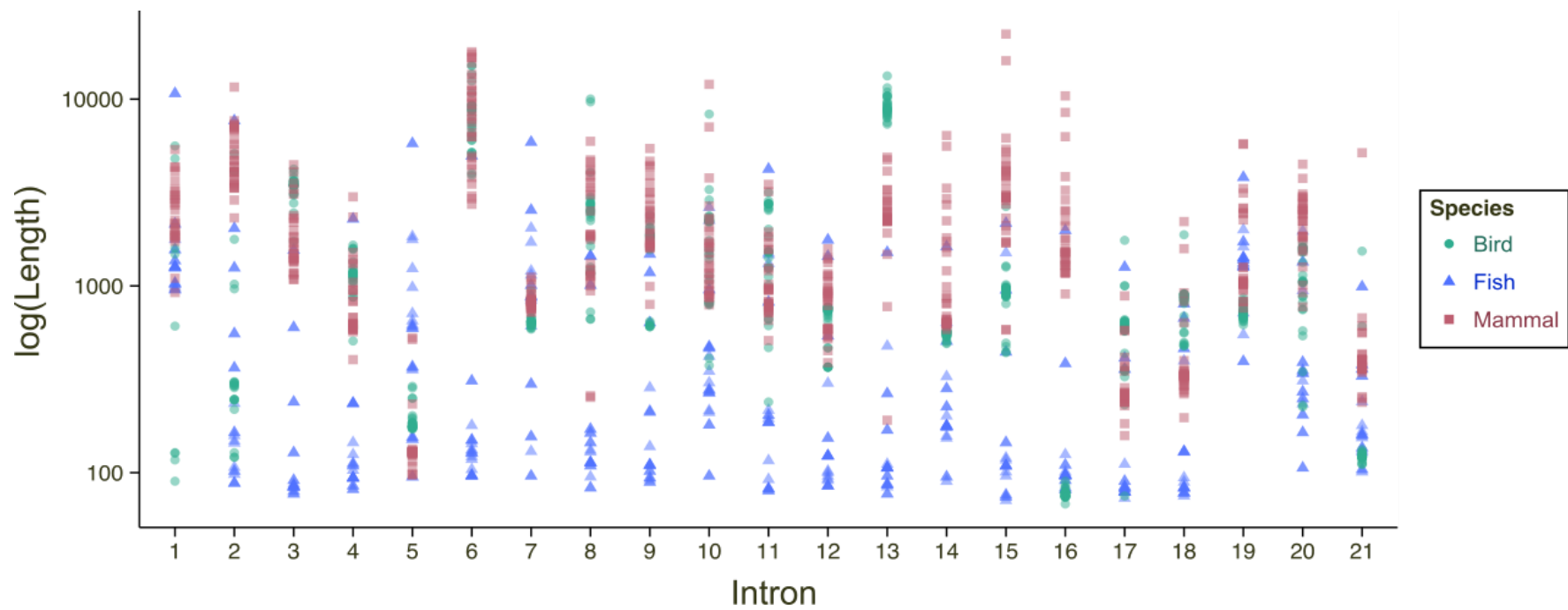


Figure 3. Length of USP4 introns for 62 bird, fish and mammal species. $\log_{10}$ scale used for y-axis. Intron 6 is highly variable in length.

Experimental tests of alternative hypotheses

Determination of splicing mechanism

Although our results are consistent with the hypothesis that E₇ skipping results from BPS₇ strengthening relative to BPS₆ and 5'SS₆ stronger than 5'SS₇, they do not exclude the possibility that the longer intron I₆ or other potentially complicating factors in USP4 pre-mRNA may contribute to E₇ skipping. To test the hypothesis that E₇ skipping is due to differential splice signals at 5'SS and BPS, we created a minigene construct by inserting the human USP4 genomic sequence encompassing E₇ (together with 75 nt at the 3' tail of I₆ and 51 nt at the 5' end of I₇) into the well characterized splicing reporter pXJ41 (the generous gift of Dr. Sushma Grellscheid, Durham University). In the resulting minigene the human E₇ genomic fragment resides in the second intron of the rabbit beta hemoglobin gene (Figure 4A). If E₇ skipping is due to the long upstream I₆, then we should observe no E₇ skipping in this construct with short upstream intron. The minigene mimics the scenario in Fig. 1Ai with the differential signal strength of splice sites: 5'SSa (Fig. 4A) is stronger (PWMS = 7.5358) than 5'SS₇ (PWMS = 5.0898) and BPSb (Fig. 4A) is stronger than BPS₆, with the former having a CUAAC (YURAY) sequence located 35 nt from the 3' end of the intron and the latter having no YURAY at D_{to3'}.opt. If E₇ skipping in its natural USP4 mRNA is due to such differential strength of splice signals, then we should observe E₇ skipping in the minigene mRNA.

When the minigene was expressed in human U2OS osteosarcoma cells, RT-PCR analysis using primers upstream and downstream of the rabbit exons revealed two isoforms of the size predicted for retention and exclusion of E₇ (Figure 4B). Similar results were obtained in the unrelated HeLa cell line (not shown). This finding excludes

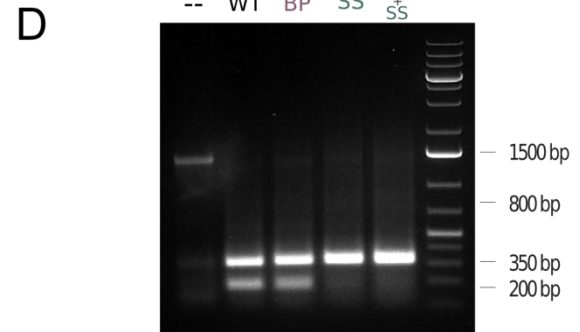
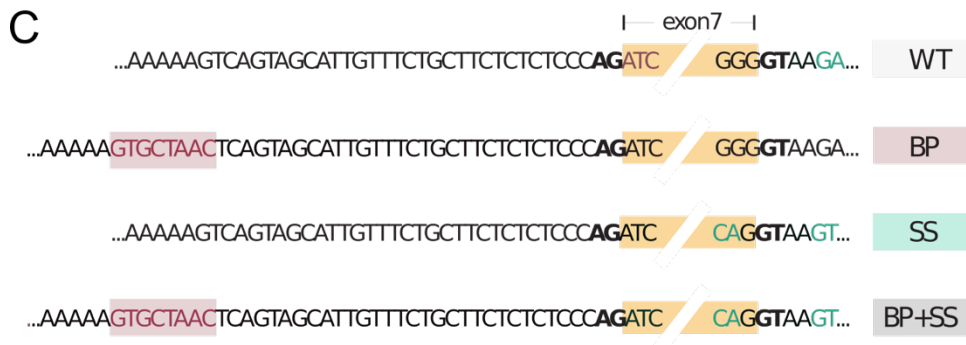
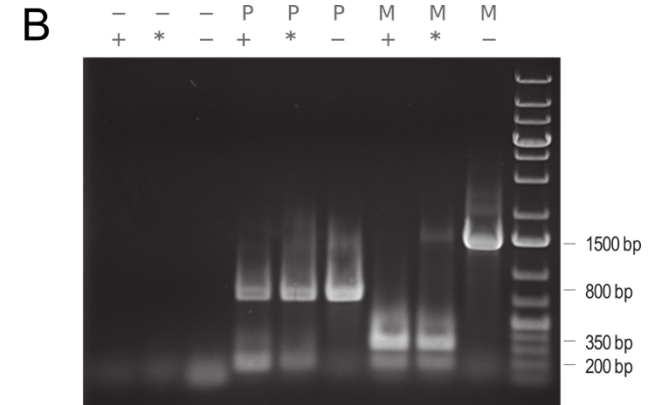
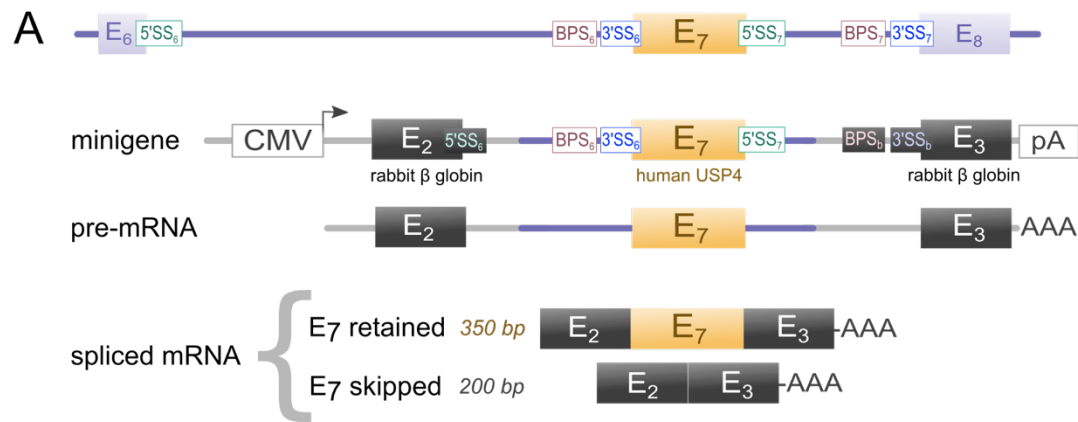


Figure 4. Mismatched flanking 5' splice site strengths leads to alternative skipping of exon 7. **A)** Schematic of the RT-PCR assay construct. Assays were performed using oligonucleotide primers positioned upstream and downstream of the rabbit exons **B).** Exon skipping as detected in minigene RT-PCR assay. Lane labels specify transfected construct and reagents in U2OS cells. First line: -: untransfected control, **P**: empty plasmid (pXJ41), **M**: USP4-E₇ minigene construct (pDG467). Second line: +: oligo dT-primed cDNA reaction, *: random-primed cDNA reaction, -: random primers, no reverse transcriptase (RT). A 200 base pair (bp) product of is visible after primed reactions of pXJ41-transfected cells. Exon-retained and -skipped products are visible as 350 and 200 bp bands, respectively, in pDG467-transfected cells. Products of approximately 800 and 1500 bp were generated pXJ41- and pDG467-transfected conditions, respectively, in the absence of RT and likely arose from amplification of DNA rather than cDNA. **C)** Schematic of site-directed mutations. The wild type (WT) sequences and the boundaries of E₇ (yellow box) are shown. In the branch point (BP) mutant a YURAY sequence (indicated in pink) was inserted 32 residues upstream of E₇. In the splice site (SS) mutants the downstream exon-intron boundary sequence were optimized (substituted residues indicated in green). The combination mutant (BP+SS) incorporated both mutations. **D)** RT-PCR analysis of site-directed mutants. Lane labels correspond to transfected mutant constructs as per Figure 4C. Exon-skipped products are absent from SS and BP+SS mutants.

exon skipping as a consequence of intron length but is consistent with E₇ skipping due to the differential strength of splice signals. We therefore explored the contributions of the BPS₆ and 5'SS₇ elements by site-directed mutagenesis of the E₇ minigene construct, inserting a consensus YURAY sequence at D_{to3'.opt} in Ia (upstream of E₇) and/or engineering an optimized 5'SS₇ element in Ib (depicted as BP and SS respectively in Figure 4C). The mutated versions of the minigene were transfected into U2OS cells and RT-PCR analysis was performed as before. Whereas the introduction of the consensus YURAY sequence had no effect on the ratio of exon retained and exon excluded products, the latter was undetectable in RNA isolated from cells transfected with minigenes in which the 5'SS₇ element had been optimized (Figure 4D).

Phylogenetic distribution of USP4 alternative splicing

In contrast to mammalian species, 5'SS₆s of chicken and zebrafish USP4 are weaker than 5'SS₇s, as indicated by their PWMS values (Table 1). To verify whether (as would be predicted) E₇ skipping does not occur in such species, RT-PCR analysis was performed on RNA isolated from primary chick fibroblast cultures (the generous gift of Dr. J.S Diallo, Ottawa Hospital Research Institute). Primers corresponding to sequences in exons 6 and 8 were used to detect the presence or absence of the seventh exon as depicted in Figure 5A. We detected only the exon-retained version of the transcript (Fig. 5B). However, when the human minigene was introduced into chick embryo fibroblasts by transfection both isoforms were detected (Fig. 5C). The absence of exon skipping in the chicken cells could thus be directly attributed to the primary sequence of the chicken USP4 pre-mRNA. Our data exclude the possibility that E₇ retention occurs in the chicken as a consequence of an altered repertoire of splicing factors in avian versus

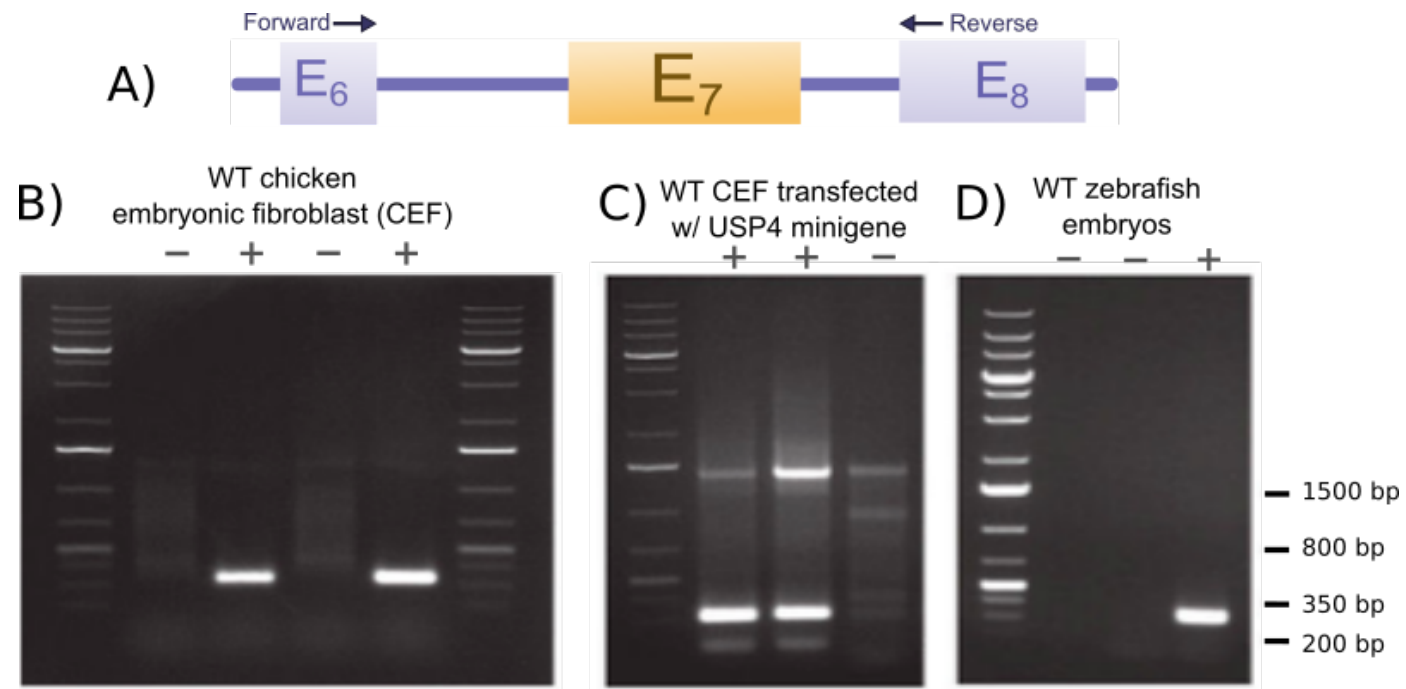


Figure 5. Exon skipping does not occur in the USP4 gene of the chicken or the zebrafish.

A) Schematic of the RT-PCR strategy using primers specific for exons 6 and 8. **B)** RT-PCR analysis of RNA from the endogenous USP4 gene in chicken. Lane labels specify addition (+) or omission (-) of RT (control) to chicken embryo fibroblast (CEF) RNA extracts. When RT was included a single PCR product of predicted size (329 bp) for the exon-retained transcript was detected. **C)** RT-PCR analysis of CEF transfected with the human USP4 minigene construct. Lane labels specify transfection of pDG467 (+) or an irrelevant control plasmid (-) in CEF cells. Both the exon-retained and exon-included products were detected (as in human cells). **D)** RT-PCR analysis of RNA from the endogenous USP4 gene in the zebrafish. Lane labels specify addition (+) or omission (-) of template (lane 2) or RT (lane 3) to larval zebrafish RNA extracts. A single amplification product was detected of the size predicted for the exon-retained cDNA (313 bp).

mammalian cells (see *Discussion*). By similar logic we predict that exon skipping would not occur in the zebrafish gene; RT-PCR analysis of RNA from larval stage zebrafish (the generous gift of Dr. Marc Ekker, University of Ottawa) confirmed the presence of a single exon-retained isoform (Fig. 5D). In support of this, performing a BLASTn of USP4 exons 5-13 against the 600,432 chicken EST sequences recovered four sequences with E₇ but no sequence without E₇. Among the 1,488,339 zebrafish ESTs, seven have E₇ but none are without E₇. In contrast, searching the 8,704,868 human ESTs recovered five sequences with and seven without E₇. The corresponding numbers from the 4,853,570 mouse EST sequences are 20 and 8, respectively. Our conceptual framework based on the relative strengths of 5' splice signals thus correctly predicted splicing propensity in these model organisms, confirmed by both database and experimental analyses.

As is shown in Figure 4D, the optimization of three nucleotides in the 5' splice site downstream of USP4-E₇ according to the consensus sequence, namely -3G→C, -2G→A and +6A→T, proved sufficient to eliminate exon skipping in the human USP4 minigene. Among species observed in Table 1, the nine therian mammal 5'SS₆s feature optimal nucleotides -2A and +6T while suboptimal -3G, -2G and +6A penalize the 5'SS₇ PWMSs of all members of this lineage. These residues are identical in the Chinese turtle and are likely thus responsible for the observed alternative E₇ skipping in this distant relative. In contrast, both flanking splice sites of E₇ in zebrafish feature optimal nucleotides (5'SS₆: -3A, -2A, +6T; 5'SS₇: -3C, -2A, +6C), which preclude E₇ exclusion. Curiously, in chicken, these determinant nucleotides are identical to those of mammals which produce E₇ skipping with the exception of the 5'SS₆ +6N site, which is weak (+6A). The upstream and downstream 5'SSs in chicken, though weak, are equivalent and

prevent exon skipping as in zebrafish. The +6N site may thus be the discriminant factor in E₇ skipping propensity. To verify this, we expanded the scope of our analysis to include all sequenced genomes bearing USP4 to see whether the 5'SS mismatching (in particular +6 site mismatching) predicts splicing proclivity. While direct expansion of our analytical framework is limited by insufficient EST data and biological sample unavailability, we can infer splicing patterns from RNA-seq datasets. Similar to the methodology used for EST mining, we performed a BLASTn of available RNA-seq data from the Sequence Read Archive (SRA) using the USP4 coding sequence with E₇ removed as a query. In the absence of hits crossing the exon 6-8 boundary for multiple, sufficiently large expression datasets, species were deemed to forgo short isoform production.

Figure 6A summarizes USP4 splicing patterns in a phylogenetic context with corresponding flanking 5'SS sequence logos indicated. According to the PWM in Table 1A, +6A and +6G weaken the 5'SS while +6T is optimal and +6C is neutral (weighted consensus illustrated in Figure 6C). For all tetrapods, when the downstream +6 site is stronger than the upstream +6 site, there is alternative splicing of E₇. This correlation is particularly apparent in the avian phylum: chicken and turkey have no E₇ skipping (+6A; +6A), all other birds either exhibit skipping (+6C/T; +6A) or loss of E₇. What is more, some members of sister taxa have lost the ability to produce the long isoform: E₇ is deleted in *Corvus brachyrhynchos* but present in *Corvus cornix cornix*; absent from Adelie penguin but present in Emperor penguin, for example. In contrast to this substantial variability, all mammals retain an optimal E₇ skipping configuration, +6T;+6A (with the exception of the clade root: platypus USP4 has +6G; +6A and,

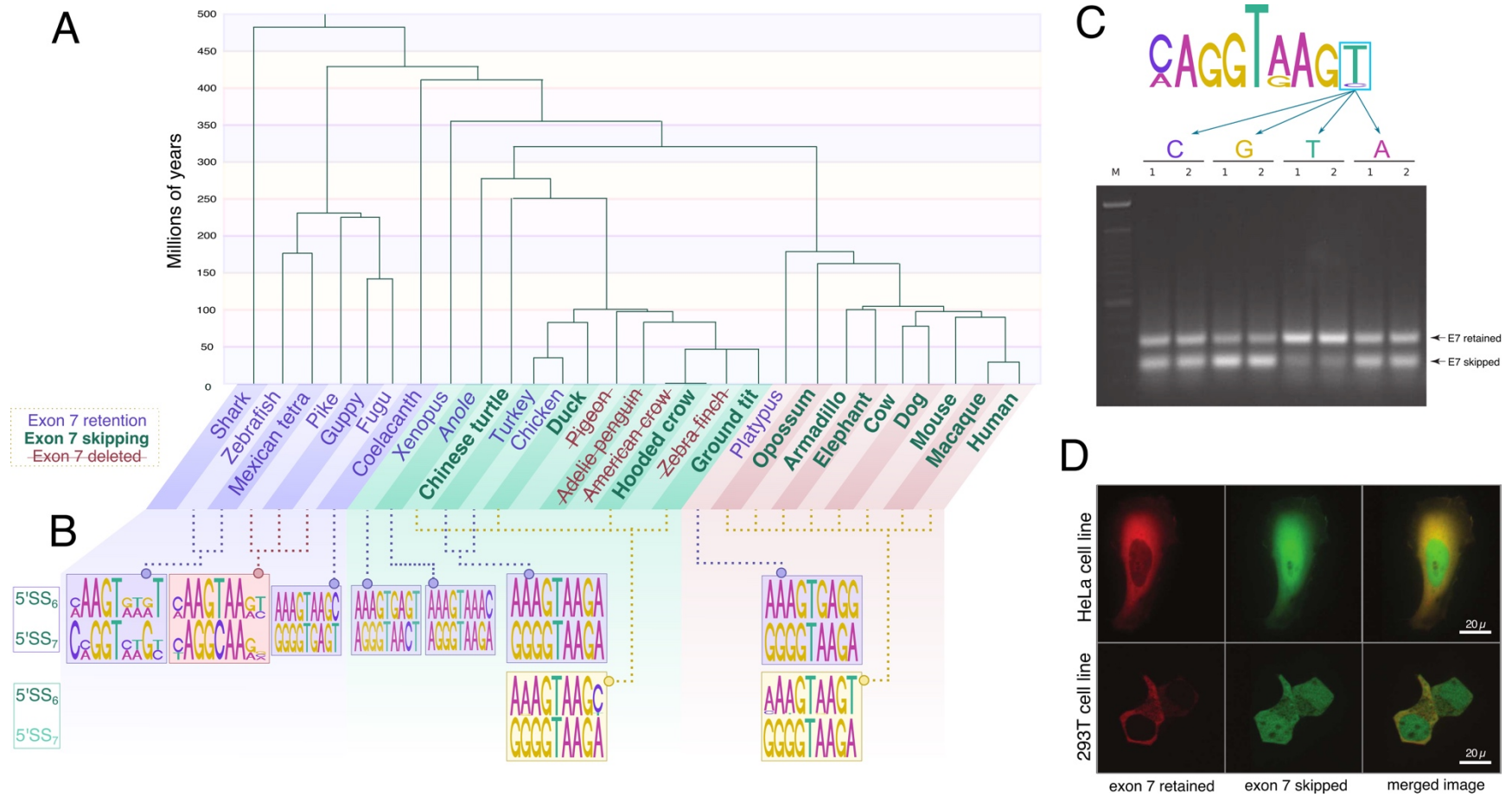


Figure 6. Long and short isoform production was selected for in therian mammals. **A)** USP4 exon 7 skipping propensity throughout the vertebrate phylogeny (see inset Legend). **B)** 5'SS₆ and 5'SS₇ identities for species in A). In B), splice site configurations that lead to constitutive retention of E₇ are in the top row and those that lead to alternative skipping are in the bottom row. **C)** Sequence logo for an optimal 5'SS (from Table 1) and changes in long-to-short isoform ratios (IR_{L/S}) after experimental replacement of the sixth intronic nucleotide (+6 site) of the downstream 5'SS₇ of human USP4 in H1299 cells. IR_{L/S} quantifications are as follows: C = 1.00, G = 0.81, T = 1.52, A (WT) = 1.02. **D)** Subcellular localization of exon 7 skipped and exon 7 retained USP4 isoforms. Long and short isoforms with appended green (GFP) and red (mKATE) fluorescent tags, respectively, were transfected into 293T and HeLA cells.

consistent with our model, does not undergo skipping). In theory, many nucleotides substitutions could disrupt the 5'SS if alternative splicing were the result of drift. Since the same splice site configuration is maintained throughout ~220 million years of mammalian evolution there may be selection for this particular +6 configuration. In Figure 6C and *Supplemental Figure 1*, we show the effects of downstream +6 site point mutation from native +6A to +6T, +6C and +6G in human and mouse cell lines. In each case, the splicing propensity changed in direct relation with the estimated fitness in our PWM: alternative splicing was nearly eliminated in +6T, slightly reduced in +6C and increased in +6G. Therefore, we propose that the highly conserved +6A site within 5'SS₇ is under natural selection to maintain both long and short USP4 isoforms in therian mammals.

Differential localizations and roles of spliced isoforms

The evidence supporting alternative splicing selection in mammalian USP4 is strong; we would consequently expect the two isoforms to have distinct cellular roles. Indeed, we observed distinct subcellular localizations of long and short USP4 isoforms in both single- and double-transfections of HeLa, U2OS, 293T and 3T3 cells (see Figure 6D). While the short isoform was distributed throughout the cell, the localization of the long isoform was largely cytoplasmic in most if not all cells in the four cell lines examined. Potential implications of this observation are discussed below. Altogether, our results suggest that the two major USP4 isoforms generated by alternative skipping of its seventh exon may not be functionally redundant as previously suggested.

DISCUSSION

It has been proposed that mutations that weaken the 5' splice site are responsible for the evolutionary shift from constitutive to alternative splicing in many vertebrate genes, as reviewed in Keren *et al.* [21], and compelling evidence has been presented in support of this hypothesis [34]. While most minor splice variants are attributable to noisy splicing [35,36], USP4 constitutes a rare case wherein selective pressure acts to conserve differential 5'SS strengths leading to exon skipping in therian mammals. The approach we presented here focuses on these cis-acting splice sites, which offers a more basic but more direct framework towards understanding the splicing code. 5' splice sites can recruit trans-acting alternative splicing factors for intrinsic splice regulation. For example, deleterious exon skipping in survival of motor neuron (SMN) pre-mRNA can be attributed to recruitment of splice repressor U2AF⁶⁵ by a weak downstream 5'SS [37].

While trans-acting factors interacting with the 5'SSs of USP4 may similarly regulate E₇ skipping, our model explains USP4-E₇ splicing propensity independent of other cis-regulatory sequences such as exonic splice enhancers (ESEs), which may or may not be selectively co-optimized in USP4 alternative splicing. Our study also highlights the importance of experimental verification of alternative hypotheses. Although the bioinformatics framework alone cannot distinguish between the two mechanisms proposed in Figure 1i and ii, the experimental results demonstrate that relative 5'SS strengths are far better predictors of alternative splicing than BPSs or upstream intron lengths. Further, our combinatorial *in silico* and experimental approach identified the +6 site within the 5'SS as the splicing discriminant. Intronic +6 site mutations have been reported as splicing instigators in other genes such as SMN1

[38,39]. E₇ skipping in SMN1 leads to spinal muscular atrophy (SMA), and SNPs that cause this aberrant skipping have been identified in patients at the downstream 5'SS₇ at the +6 site (+6T→G). SMN1 has a very close paralog, SMN2, that is incapable of rescuing SMN1 deficiency in SMA because its E₇ is also skipped due to a WT nucleotide variant, 5'SS₇ +6G. Thus, +6T at the downstream 5'SS₇ of SMN1/2 promotes upstream exon inclusion while +6G promotes near-total upstream exon skipping. In mammalian WT USP4, 5'SS₆ has +6T while 5'SS₇ has +6A. As reflected in the PWM in Table 1 and observed in Fig. 6C, the strengths of +6 site nucleotides are predicted to be as follows: T > C ≥ A > G, where a stronger 5'SS₆ +6 relative to 5'SS₇ +6 correlates with splicing proclivity. It is curious that +6A ≠ +6G and that the former was selected as the weak downstream nucleotide of USP4. A plausible mechanism for +6A-dependent alternative skipping may involve U1C, a component of the U1 snRNP that preferentially recognizes a 5'SS motif with +6A, GTATAA [40], and can interact with splicing regulator TIA-1 [41,42] to promote exon retention, for example in SMN2 [43,44]. Several genes undergo U1C-dependent alternative splicing [45,46]. Based on linear changes in the relative abundances of the short and long isoforms observed during differentiation of P19 embryonic carcinoma cells (Gray, unpublished) we postulate that E₇ of USP4 may be subject to regulated alternative splicing in therian mammals.

Retention or exclusion of the amino acids encoded by exon 7 does not affect the protease activity of the USP4 enzyme (using a synthetic substrate [14]) and the ubiquitin-exchange regulatory mechanism proceeds equally in both USP4 isoforms [16]. We have nonetheless shown that there is selection for alternative splicing maintenance in mammalian USP4. Establishing the molecular selection driver should be highly

informative. We show that the two isoforms display distinct subcellular localizations, which suggests that (1) propensity and/or (2) capacity for substrate interactions may differ. First, vital cytoplasmic (e.g. TGF- β pathway [1]) and nuclear (e.g. spliceosomal [11,12]) substrates have been reported for USP4 (isoforms specificities not declared). Long and short USP4 isoform production may be advantageous for simultaneous, collective targeting of key substrates in various cellular compartments. On the other hand, the two isoforms almost certainly have some distinct interactors. Cytoplasmic retention of USP4-long is mediated by phosphorylation of a ubiquitously conserved serine (S445); the apparent absence of this regulation in USP4-short may reflect a lack of phosphorylation by Akt [2]. Exon 7 of mammalian USP4 is serine-rich (16 out of 47 residues; human sequence: RSSTAPSRNFTTSPKSSASPYSSVSASLIANGDSTSTCGMHSSGVSRG) and also contains six constitutively charged residues (five positively charged and one negatively charged). While the +4 net charge difference between isoforms likely affects substrate interactions, the serine-rich exon also retains multiple phosphorylation sites (underlined) that are conserved among all mammals. USP4-E₇ may be phosphorylated in conjunction with Ser445 for nuclear exclusion or may be required for interaction with Akt and other substrates. For instance, SART3, a spliceosomal factor and deubiquitination target of both USP4 [12] and its paralog USP15 [47], has been reported to interact with serine-rich (not to be confused with serine/arginine-rich) domains of proteins [48]. Interestingly, USP15 contains an analogous, serine-enriched alternatively spliced seventh exon (SPGASSNFSTLPKISPSSLNNYNNMNNR; reported phosphorylation sites underlined). Splice boundaries and amino acid sequence differ between E₇ of USP4 and USP15, suggesting that alternative splicing arose independently

in these, though they both maintain significant proportions of serines. This may be a case of stabilizing selection acting on clusters of phosphorylation sites [49]. There may be an important feedback loop involving the splicing and subsequent localization of USP4 and USP15, two DUBs that critically interact with the spliceosome. Long and short USP4 production and thus DUB modification of isoform-specific substrates may differ across tissue types. It remains to be seen whether such isoform-specific substrates drove evolutionary conservation of the dual isoforms within placental mammals.

To summarize, we have shown that the long and short isoforms of USP4 have distinct properties and their contributions to cellular networks should be considered separately. Most proteins have more than one reported isoform, and though most may be considered non-essential noise, distinct functional variants, such as in USP4, must not be grouped as one protein. The roles of all significantly expressed minor splice variants should be studied more carefully.

METHODS

Bioinformatic analysis

Well-annotated USP4 sequences for 62 vertebrate species were downloaded from GenBank (See Supplementary Table S2) covering major vertebrate taxa. Coding sequences, exons, introns, and exon-intron junctions (5 nt on the exon side and 12 nt on the intron side) were extracted and analyzed by using DAMBE [50].

Shannon entropy (H) is used as a measure of site-specific sequence variability over a sliding window in Figure 1C.

$$H = - \sum p_i \log_2 p_i$$

N in Eq. (1) is the number of character states ($N = 20$ for protein), and p_i is the frequency of each amino acid.

To identify the optimal distance ($D_{\text{to3'}.opt}$) of branchpoint sites (BPS), we searched for the presence of YURAY motifs without downstream non-3'SS AG dinucleotides in all 15770 introns in human chromosome 22 and compiled $D_{\text{to3'}}$, defined as the number of nucleotides from the beginning of YURAY (human) to the 3' end of the intron. The range that contains the majority of $D_{\text{to3'}}$ values is labeled $D_{\text{to3'}.opt}$. We applied the same approach for chicken and zebrafish introns.

To identify whether USP4 is subject to alternative splicing in diverse species, BLAST searches were performed using putative species-specific short isoforms (i.e., with exon 7 deleted) as a query against RNA-seq datasets from the Sequence Read Archive (SRA). RNA-Seq dataset identifiers are listed in the Supplementary Table S2.

RT-PCR analysis of spliced isoforms

The USP4 exon 7 reporter plasmid pDG467 was generated by insertion of a segment of human USP4 genomic DNA into the minigene reporter pXJ41. A segment of the USP4 gene on chromosome 3 (49348797-49349237) was amplified by polymerase chain reaction from H1299 cell DNA using Phusion high fidelity DNA polymerase (Thermo Scientific, Waltham, MA) with the forward primer

AAAAAAGAATTCATTACAGGCACGAGCCACTG and the reverse primer

AAAAAAGAATTCGCCCATCCCTTCATAAACAA (annealing temperature 55C). The resulting 699 base pair DNA product was digested with EcoRI and gel purified prior to ligation into the MfeI site of pXJ41. The branch point insertion plasmid pDG484 was generated using the Phusion site directed mutagenesis system (Thermo Scientific,

Waltham, MA) with pDG467 as the template, forward primer 5'-GCTAACTCAGTAGCATTGTTTCTGCTTCTC-3', and reverse primer 5'-ACTTTTTGCAAAGAGCAAGCCCTATTTA-3' (68C annealing temperature). The splice donor optimization plasmid pDG485 was generated with pDG467 as the template, forward primer GTAAGTGCAGGTCCTTTCACTCTGCTTC, and reverse primer CTGCTGCTGACACCGGAACTGT (68C annealing temperature). The combination plasmid pDG468 was generated using the same primers and conditions, but using pDG484 as the template. Substitutions at the +6 position of the splice site were generated by polymerase chain reaction with pDG467 as the template. For all such substitutions the reverse primer CCCCTGCTGACACCGGAACTGT was used. For the A to C substitution the forward primer was GTAAGCGCAGGTCCTTTCACTCTGCTTC. For the A to G substitution the forward primer was GTAAGGGCAGGTCCTTTCACTCTGCTTC. For the A to T substitution the forward primer was GTAAGTGCAGGTCCTTTCACTCTGCTTC. The conditions for polymerase chain reaction were as above, but with an annealing temperature of 72C. The engineered mutations in all USP4-derived plasmids were verified before subsequent transfection experiments were performed. For analysis of splicing isoforms human U2OS cells, mouse NIH 3T3 cells or human H1299 cells (ATTC, Manassas, VA) were transfected at 50% confluence in 6 well dishes with 1 µg of plasmid at 3 µl of GeneJuice (EMD Millipore, Billerica, MA) using the manufacturer's protocol. RNA was harvested 24 hours post-transfection using a GeneJET RNA purification system (Thermo Scientific, Waltham, MA). Coupled reverse transcription/polymerase chain reaction was performed for each sample using the MyTaq One Step RT-PCR kit (BioLine, Taunton, MA) with forward primer

GCTCCGGATCGATCCTGAGAACT and reverse primer GCTGCAATAAACAAGTTCTGC (60C annealing). RT-PCR products were analyzed on 1.2% agarose gels. DNA products were stained with Safe-Red (Applied Biological Materials, Inc., Richmond BC) and visualized using a UV gel camera apparatus (UVP, Upland CA).

Microscopy and imaging

To establish the localization of exon-retained and exon skipped isoforms of USP4, cDNAs corresponding to the mouse isoforms were obtained from Origene Technologies Inc. (Rockville, MD, USA) as C-terminal fusions with the red fluorescent protein mKATE or monomeric GFP, respectively. One microgram of each plasmid was introduced into cells cultured on cover slips using the GeneJuice transfection reaction (Millipore Canada, Etobicoke ON) following the supplied protocol. 24 hours post-transfection the cells were fixed for 10 minutes in 0.4% paraformaldehyde and were mounted in Vectashield mounting media (Vector Laboratories Canada, Burlington ON). Images were acquired using a Zeiss Axiovert 200M microscope equipped with the Apotome optical sectioning module.

Acknowledgements

This study was supported by Discovery Grants of the Natural Science and Engineering Research Council of Canada (NSERC) to X.X. and D.A.G. and by a grant from the Cancer Research Society, Inc. to D.A.G.

Author contributions statement

CV, XX, and DAG contributed equally to the design of the study and the preparation of the manuscript. CV and XX were responsible for computational analyses, while DAG was responsible for the experimental aspects of the study.

Competing financial interests

The authors declare that they have no competing interests.

References

1. Aggarwal, K. & Massagué, J. *Ubiquitin removal in the TGF- β pathway. Nat Cell Biol.* **14**, 656–657 (2012).
2. Zhang, L. *et al.* *USP4 is regulated by AKT phosphorylation and directly deubiquitylates TGF- β type I receptor. Nat Cell Biol.* **14**, 717–726 (2012).
3. Fan, Y.-H. *et al.* *USP4 targets TAK1 to downregulate TNF α -induced NF- κ B activation. Cell Death Differ.* **18**, 1547–1560 (2011).
4. Zhao, B., Schlesiger, C., Masucci, M. G. & Lindsten, K. *The ubiquitin specific protease 4 (usp4) is a new player in the wnt signalling pathway. J Cell Mol Med.* **13**, 1886–1895 (2009).
5. Zhang, X., Berger, F. G., Yang, J. & Lu, X. *USP4 inhibits p53 through deubiquitinating and stabilizing ARF-BP1. EMBO J.* **30**, 2177–2189 (2011).
6. Wang, L. *et al.* *USP4 positively regulates RIG-I-mediated antiviral response through deubiquitination and stabilization of RIG-I. J Virol.* **87**, 4507–4515 (2013).
7. Xiao, N. *et al.* *Ubiquitin-specific protease 4 (USP4) targets TRAF2 and TRAF6 for deubiquitination and inhibits TNF α -induced cancer cell migration. Biochem J.* **441**, 979–986 (2012).
8. Zhou, F. *et al.* *Ubiquitin-specific protease 4 mitigates Toll-like/interleukin-1 receptor signaling and regulates innate immune activation. J Biol Chem.* **287**, 11002–11010 (2012).
9. Uras, I. Z., List, T. & Nijman, S. M. B. *Ubiquitin-specific protease 4 inhibits mono-ubiquitination of the master growth factor signaling kinase PDK1. PLoS One* **7**, e31003 (2012).
10. Soboleva, T. A., Jans, D. A., Johnson-Saliba, M. & Baker, R. T. *Nuclear-cytoplasmic shuttling of the oncogenic mouse UNP/USP4 deubiquitylating enzyme. J Biol Chem.* **280**, 745–752 (2005).
11. Sowa, M. E., Bennett, E. J., Gygi, S. P. & Harper, J. W. *Defining the human deubiquitinating enzyme interaction landscape. Cell.* **138**, 389–403 (2009).
12. Song, E. J. *et al.* *The Prp19 complex and the Usp4sart3 deubiquitinating enzyme control reversible ubiquitination at the spliceosome. Genes Dev.* **24**, 1434–1447 (2010).
13. Gupta, K., Chevrette, M. & Gray, D. A. *The Unp proto-oncogene encodes a nuclear protein. Oncogene.* **9**, 1729–1731 (1994).
14. Frederick, A., Rolfe, M. & Chiu, M. I. *The human UNP locus at 3p21.31 encodes two tissue-selective, cytoplasmic isoforms with deubiquitinating activity that have reduced expression in small cell lung carcinoma cell lines. Oncogene.* **16**, 153–165 (1998).
15. Gupta, K., Copeland, N. G., Gilbert, D. J., Jenkins, N. A. & Gray, D. A. *Unp, a mouse gene related to the tre oncogene. Oncogene.* **8**, 2307–2310 (1993).
16. Clerici, M., Luna-Vargas, M. P. A., Faesen, A. C. & Sixma, T. K. *The DUSP-Ubl domain of USP4 enhances its catalytic efficiency by promoting ubiquitin exchange. Nat Commun.* **5**, 5399 (2014).
17. Vlasschaert, C., Xia, X., Coulombe, J. & Gray, D. A. *Evolution of the highly networked deubiquitinating enzymes USP4, USP15, and USP11. BMC Evol Biol.* **15**, 230 (2015).
18. Alekseyenko, A. V., Kim, N. & Lee, C. J. *Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. RNA* **13**, 661–670 (2007).

19. Sugnet, C. W. *et al.* Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput Biol.* **2**, e4 (2006).
20. Sammeth, M., Foissac, S. & Guigó, R. A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol.* **4**, e1000147 (2008).
21. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet.* **11**, 345–355 (2010).
22. Bell, M. V., Cowper, A. E., Lefranc, M. P., Bell, J. I. & Sreaton, G. R. Influence of intron length on alternative splicing of CD44. *Mol Cell Biol.* **18**, 5930–5941 (1998).
23. Roy, M., Kim, N., Xing, Y. & Lee, C. The effect of intron length on exon creation ratios during the evolution of mammalian genomes. *RNA* **14**, 2261–2273 (2008).
24. Kandul, N. P. & Noor, M. A. F. Large introns in relation to alternative splicing and gene evolution: a case study of *Drosophila bruno-3*. *BMC Genet.* **10**, 67 (2009).
25. Fox-Walsh, K. L. *et al.* The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *PNAS* **102**, 16176–16181 (2005).
26. Sterner, D. A., Carlo, T. & Berget, S. M. Architectural limits on split genes. *PNAS* **93**, 15081–15085 (1996).
27. Kim, E., Magen, A. & Ast, G. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.* **35**, 125–131 (2007).
28. Patel, A. A. & Steitz, J. A. Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol.* **4**, 960–970 (2003).
29. Xing, Y. & Lee, C. Assessing the application of Ka/Ks ratio test to alternatively spliced exons. *Bioinformatics.* **21**, 3701–3703 (2005).
30. Gao, K., Masuda, A., Matsuura, T. & Ohno, K. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res.* **36**, 2257–2267 (2008).
31. Zheng, C. L., Fu, X.-D. & Gribskov, M. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA* **11**, 1777–1787 (2005).
32. Dewey, C. N., Rogozin, I. B. & Koonin, E. V. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics.* **7**, 311 (2006).
33. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol.* **6**, 386–398 (2005).
34. Lev-Maor, G. *et al.* The “alternative” choice of constitutive exons throughout evolution. *PLoS Genet.* **3**, e203 (2007).
35. Melamud, E. & Moulton, J. Stochastic noise in splicing machinery. *Nucleic Acids Res.* **37**, 4873–4886 (2009).
36. Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLoS Genet.* **6**, e1001236 (2010).
37. Cho, S. *et al.* Splicing inhibition of U2af65 leads to alternative exon skipping. *PNAS* **112**, 9926–9931 (2015).
38. Wirth, B. *et al.* Quantitative Analysis of Survival Motor Neuron Copies: Identification of Subtle SMN1 Mutations in Patients with Spinal Muscular Atrophy, Genotype-Phenotype Correlation, and Implications for Genetic Counseling. *Am J Hum Genet.* **64**, 1340–1356 (1999).

39. Lorson, C. L., Hahnen, E., Androphy, E. J. & Wirth, B. *A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy*. *PNAS* **96**, 6307–6311 (1999).
40. Du, H. & Rosbash, M. *The U1 snRNP protein U1c recognizes the 5' splice site in the absence of base pairing*. *Nature* **419**, 86–90 (2002).
41. Förch, P., Puig, O., Martínez, C., Šeraphin, B. & Valcárcel, J. *The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5' splice sites*. *EMBO J.* **21**, 6882–6892 (2002).
42. Bauer, W. J., Heath, J., Jenkins, J. L. & Kielkopf, C. L. *Three RNA recognition motifs participate in RNA recognition and structural organization by the pro-apoptotic factor TIA-1*. *J Mol Biol.* **415**, 727–740 (2012).
43. Singh, N. N. *et al.* *TIA1 Prevents Skipping of a Critical Exon Associated with Spinal Muscular Atrophy*. *Mol Cell Biol.* **31**, 935–954 (2011).
44. Klar, J. *et al.* *Welander distal myopathy caused by an ancient founder mutation in TIA1 associated with perturbed splicing*. *Hum Mutat.* **34**, 572–577 (2013).
45. Rösel, T. D. *et al.* *RNA-Seq analysis in mutant zebrafish reveals role of U1c protein in alternative splicing regulation*. *EMBO J.* **30**, 1965–1976 (2011).
46. Rösel-Hillgärtner, T. D. *et al.* *A Novel Intra-U1 snRNP Cross-Regulation Mechanism: Alternative Splicing Switch Links U1c and U1-70k Expression*. *PLoS Genet.* **9**, e1003856 (2013).
47. Long, L. *et al.* *The U4/U6 recycling factor SART3 has histone chaperone activity and associates with USP15 to regulate H2b deubiquitination*. *J Biol Chem.* **289**, 8916–8930 (2014).
48. Harada, K., Yamada, A., Yang, D., Itoh, K. & Shichijo, S. *Binding of a SART3 tumor-rejection antigen to a pre-mRNA splicing factor RNPS1: A possible regulation of splicing by a complex formation*. *Int J Cancer.* **93**, 623–628 (2001).
49. Landry, C. R., Freschi, L., Zarin, T. & Moses, A. M. *Turnover of protein phosphorylation evolving under stabilizing selection*. *Front Genet.* **5** (2014).
50. Xia, X. *DAMBE5: A Comprehensive Software Package for Data Analysis in Mol Biol Evol.* **30**, 1720–1728 (2013).
51. Edgar, R. C. *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
52. Wickham, H. *ggplot2: elegant graphics for data analysis* (Springer: New York, 2009).

CHAPTER III

The evolution and functional diversification of the deubiquitinating enzyme superfamily

This manuscript is under review for the journal *Genome Biology and Evolution*.

The student performed all evolutionary bioinformatics analyses (Figures 1, 3bc, 4, 5, 6, Supplementary Figure 2) while a student co-author performed the transcriptomic analyses (Figure 2, Supplementary Figure 1). The student also prepared all figures and wrote a first draft of the manuscript, which was edited by all authors.

Research Article

The evolution and functional diversification of the deubiquitinating enzyme superfamily

Caitlyn Vlasschaert^{1,2,3}, David Cook^{2,4}, Xuhua Xia^{3,5} & Douglas A. Gray^{1,2*}

¹ Department of Biochemistry, Microbiology and Immunology, University of Ottawa

² The Ottawa Hospital Research Institute, Ottawa, Canada

³ Department of Biology, University of Ottawa

⁴ Department of Cellular and Molecular Medicine, University of Ottawa

⁵ Ottawa Institute of Systems Biology, Ottawa, Canada

*corresponding author. Centre for Cancer Therapeutics, Ottawa Hospital Research Institute, 501 Smyth Rd., Ottawa, Ontario, Canada K1H 8L6.

Caitlyn Vlasschaert email: caitlyn.vlasschaert@gmail.com

David Cook email: dcook082@uottawa.ca

Xuhua Xia email: Xuhua.Xia@uottawa.ca

Douglas A. Gray email: dgray@ohri.ca

ABSTRACT

Ubiquitin and ubiquitin-like molecules are attached to and removed from cellular proteins in a dynamic and highly regulated manner. Deubiquitinating enzymes are critical to this process, and the genetic catalogue of deubiquitinating enzymes expanded greatly over the course of evolution. Extensive functional redundancy has been noted among the 93 members of the human deubiquitinating enzyme (DUB) superfamily. This is especially true of genes that were generated by duplication (termed paralogs) as they often retain considerable sequence similarity. Since complete redundancy in systems should be eliminated by selective pressure we theorized that many overlapping DUBs must have significant and unique spatiotemporal roles that can be evaluated in an evolutionary context. We have determined the evolutionary history of the entire class of deubiquitinating enzymes, including the sequence and means of duplication for all paralogous pairs. To investigate their uniqueness, we have investigated cell-type specificity in developmental and adult contexts, and have investigated the co-emergence of substrates from the same duplication events. Our analysis has revealed examples of DUB gene subfunctionalization, neofunctionalization, and nonfunctionalization.

Keywords:

deubiquitinating enzymes, molecular evolution, innate immunity, DNA repair

INTRODUCTION

Ubiquitin is a molecular tag, a small protein whose attachment to other proteins can influence their fate. The first role established for ubiquitin conjugation was in the regulation of protein stability, but attachment of ubiquitin was later found to influence the subcellular localization of substrates and their proclivity to interact with signalling complexes (reviewed by Varshavsky (Varshavsky 2012)). For some functions ubiquitin remains covalently attached as a monomer, but the full multiplicity of functions derives from the potential of ubiquitin to be assembled into chains of various topologies, generating what has been called the "ubiquitin code" (Komander and Rape 2012). The constellation of potential substrates and the varying conditions in which conjugation to any given substrate may or may not be desirable necessitates a specific and responsive enzymatic machinery for ubiquitin attachment and removal. The initial attachment of ubiquitin and assembly of chains is orchestrated by an understandably large repertoire of ubiquitin conjugases and ligases (acting in concert), and is opposed by a somewhat smaller repertoire of deubiquitinating enzymes (DUBs). The opposing activities of ligases and DUBs can be modulated by post-translational modifications, allowing rapid adjustments to be made through signalling input. In mammalian cells there are several hundred ubiquitin ligases, and in the order of a hundred DUBs (Hutchins et al. 2013). It is a safe assumption that the metazoan ubiquitin system in all its daunting complexity had its evolutionary origins in a simpler system; the unicellular eukaryote *Saccharomyces cerevisiae* reportedly has 68 ubiquitin ligases and 24 DUBs (Hutchins et al. 2013), whereas thermophilic archaea (thought to be the closest living relatives of eukaryotes (Koonin 2015)) have a minimalist ubiquitin "toolkit" (Grau-Bové et al. 2015). As an extreme example the genome of *Candidatus Caldiarchaeum subterraneum*

contains an operon incorporating a single ligase and DUB in tandem with a ubiquitin-like gene (Nunoura et al. 2011).

The focus of the current work is the evolutionary origin of the DUBs. It is likely that gene duplication has provided raw material for the expansion of this part of the toolkit, but the extent to which gene duplication has contributed to the metazoan DUB repertoire has not been previously evaluated, nor have the types of duplication events and the subsequent specialization of duplicated genes been comprehensively explored. The two rounds (2R) of whole genome duplication (WGD) that are purported to have occurred early in vertebrate evolution (Dehal and Boore 2005) would have generated a surfeit of duplicated genes (designated "ohnologs" (Wolfe 2000) in honour of Ohno, who proposed the 2R-WGD mechanism). By its very nature WGD globally preserves molecular stoichiometry, whereas duplication only of chromosomal regions (segmental duplication) disrupts the stoichiometry of unlinked genes with potentially deleterious consequences. DUBs can have many interacting partners (Sowa et al. 2009), and as network hubs could be very sensitive to dosage alterations. Stoichiometry can be restored by silencing a duplicated gene, and silencing is indeed the fate of most gene duplicates over a timeframe of a few million years (Lynch and Conery 2000). The sequence of duplicated genes is initially identical, but with subsequent divergence there is the potential for subfunctionalization (a division of existing molecular functions), which in metazoans may be achieved by dividing duties within a cell or by dividing the pattern of gene expression such that ohnologs are expressed in different cell types or at different developmental stages. There is also the possibility of neofunctionalization (the acquisition of novel functions by one or both duplicates), which may temporally coincide with the emergence of novel molecular pathways or more subtle innovations. These would be adaptive changes, but in the absence of strong selection or in

species with small effective population size genetic drift may promote substitutions culminating in pseudogenization of a gene duplicate. Such occurrences may be informative with respect to the functional redundancy of DUBs (Vlasschaert et al. 2015).

Ubiquitin-mediated proteolysis plays a central role in ancient eukaryotic systems (the cell cycle, for example), but for increasingly elaborate developmental and homeostatic pathways there may have been a requirement for an expanded DUB repertoire. We have sought to determine how many of the vertebrate DUBs can be considered ohnologs, and whether there are clear examples of DUB subfunctionalization and/or neofunctionalization. We chose to concentrate on the roles of DUBs in two pathways that predate 2R-WGD but whose regulation over the course of metazoan evolution has become increasingly baroque: innate immunity and DNA repair.

RESULTS

Evolutionary history of the deubiquitinating enzyme class

We first sought to delineate the large-scale evolutionary history of the entire superfamily of deubiquitinating enzymes, or DUBs. The DUB repertoire at the origin of three major eukaryotic clades is presented in Figure 1A in the form of a Venn diagram with intersecting budding yeast, leaf and sea urchin representing fungi, plants and animals, respectively. Protist repertoires are not explicitly represented due to the group's paraphyletic nature (Schlegel and Hülsmann 2007) and limited genome availability for the individual clades, but they are used to support inferred ancestral ages of DUBs, as described below. Twenty-three DUBs and six DUB-related genes, which have homology with DUBs but do not have isopeptidase properties are inferred to be at the base of the eukaryotic kingdoms. Three additional DUBs are shared among opisthokonts (fungi and animals) and seven are yeast-specific, amounting to a total of 39 fungal

DUBs. This number is greater than the 24 DUBs reported in a Hidden Markov Model scan of the *Saccharomyces cerevisiae* genome (Hutchins et al. 2013) because it includes those not found in the yeast genome but present in other fungi as well as in several protist genomes and because four DUB members were overlooked in the original study (PAN2, ALG13, COPS5 and MPND). An additional nine DUBs are shared between plants and animals though were also likely present in the eukaryotic ancestor since they are also present in protists, which are median to these kingdoms. DUBs that underwent subsequent duplication in plants or yeast and those that are constitutive subunits of complexes are identified in Figure 1A. As the figure indicates, of the three kingdoms considered twenty ancestral DUBs are unique to Animalia, though eight of these are also found in protist genomes. Some DUBs were found to be notably absent in the genomes of several orders of Insecta (including Diptera, which contains *Drosophila melanogaster*) and in some cases were absent from the entire class. As previously noted, all five families of DUBs are ancestrally represented (Hutchins et al. 2013), though DUBs within these families do not all share sufficient similarity to be considered paralogs (Vilella et al. 2009) (Figure 1B).

The working set of 59 DUBS common to Animalia is expansive, and the repertoire is increases by more than half that number in the human genome. Two DUBs with particular cleavage targets, OTULIN (linear ubiquitin chain specificity (Keusekotten et al. 2013)) and USPL1 (SUMO specificity (Schulz et al. 2012)), emerged in bilateria (Figure 1A) while the remainder arose during vertebrate radiation. Figure 1B presents the sequential expansion of DUBs from the bilaterian set to the distinct human genetic repertoire by means of whole genome duplication, small-scale duplication, retrogenic duplication, tandem duplication and chimerization. The first major expansion of DUBs coincides with the emergence of jawed vertebrates approximately 450 million years ago (Venkatesh et al. 2014). Two rounds of whole

genome duplication (WGD) in the basal vertebrate(Dehal and Boore 2005) provided extensive genetic fodder for network rewiring and organismal remodelling in the subsequent radiation of vertebrates. At least twenty-one DUBs were derived and maintained from this WGD event (Figure 1B).

Given the method of duplication, ohnologs should only be located adjacent to one another if significant chromosomal rearrangement occurred. The 3' tail of the *Usp50* gene, which first appears in cartilaginous fish, overlaps with the 3'UTR of its paralog, *Usp8*, in both the human and elephant shark (*Callorhinchus milii*) genomes. USP8 is the ancestral DUB with which USP50 retains highest coding sequence similarity (in shark and human) and thus is most likely to be its parent. Though USP50 arose concordant with WGD timing, its arrangement with USP8 suggests that their relationship may not be ohnologous.

The percent similarity between globally aligned (Edgar 2004) human duplicated DUB protein sequences is indicated on the arrows of Figure 1B. While several pairs have diverged drastically (for example the USP47-USP18 pair with 14% similarity, or the USP8-USP50 pair with 19% similarity), 70% of ohnologous pairs have retained more than 50% sequence similarity. In some of these cases, both equally resemble protein sequences from the single ancestral copies in invertebrate Animalia (acknowledged in the nomenclature used in Figure 1A). High coding sequence similarity does not always equate to limited subfunctionalization. For example, JOSD1 is a membrane-bound DUB that requires allosteric activating ubiquitination and regulates cell motility and endocytotic processes, while its ohnolog JOSD2 (though 62% similar) is cytoplasmic and retains innate deubiquitinase activity (Seki et al. 2013: 1). UCHL1 and UCHL3 are 73% similar but display different patterns of tissue-specific gene expression: an upstream neuron-restrictive silencing element (NRSE) drives neuron-specific expression in the

former (Barrachina et al. 2007) and renders it a critical player in neuronal homeostasis (UCHL1 deficiency in neurodegeneration is not physiologically rescued by UCHL3). Other ohnologous DUB pairs, such as USP4 and USP15 are interchangeable with respect to organismal viability but have subtle yet evolutionarily-stable properties that are distinctive (Vlasschaert et al. 2015). Functional redundancy has however not been evaluated for many WGD-derived pairs, including USP12-USP46 whose similarity is 92%.

Following the WGD mass expansion, small-scale chromosomal duplications of the genomic regions encoding USP2 and USP4 gave rise to USP21 and USP11 respectively in bony vertebrates (Vlasschaert et al. 2015). All other changes contributing to the human DUB repertoire occurred in the mammalian lineage. Several retrogenes, mostly incurred in the mammalian ancestor, were added while the CYLD duplicate (named CYLD-like in Figure 1B, or CYLDL, for convenience) was specifically deleted in all mammals. As might be expected (Drouin 2006)(Potrzebowski et al. 2010), most functional processed retro-pseudogenes are on the X chromosome. USP29 is an autosomal exception, but it is poorly conserved: human and gorilla protein identity is only 80% and the gene is absent in multiple species. Its reported role in stabilizing p53 (Liu et al. 2011) is therefore somewhat counterintuitive. USP17 is a mammalian retrogene with clusters of variably transcribed pseudogenes tandemly arranged on the chromosomes 4 and 8 (Burrows et al. 2005; Burrows et al. 2010). USP9Y is a Y-linked, non-retrogenic copy of USP9X that originated in the common ancestor of euarchontoglires (a clade that includes primates and rodents) and encompasses several proximal pseudogenes. USP32 similarly has an array of neighbouring, non-processed pseudogenes. As with the USP17 and USP32 pseudogenes, the X-chromosome retrogene ATXN3L, the exon-bearing tandem duplicate USP18 and the chimera USP6 were more recently acquired at various times during primate

evolution. The detailed divergence and functional discrepancies for the majority of paralogous pairs in Figure 1B has yet to be formally addressed.

Related deubiquitinating enzymes evolve discrete spatiotemporal occupations

Retained duplicate genes are thought to diverge either in terms of coding sequence to carry out discrete cellular roles resulting from structurofunctional changes or in terms of expression pattern to accomplish similar functions in discrete spatiotemporal contexts²⁴. To illustrate the latter, Figure 2 presents the tissue specificities of all known human DUBs. In Figure 2A, RNA-Seq expression data from 2923 samples of 53 human tissues obtained from the Genotype-Tissue Expression (GTEx) Project²⁵ was used to explore tissue specific expression of DUBs. RPKM expression values for each DUB were standardized (Z-score transformation) to allow for a comparison of tissue enrichment between DUBs. These comparisons should be interpreted with caution as an equal degree of enrichment between two DUBs is not indicative of similar absolute expression. Median expression values for each DUB are represented at the bottom row of the heatmap to assist with these comparisons; similarly, Supplementary Figure 1A is the row-normalized counterpart of Fig 2A that ranks absolute DUB expression for each tissue and provides an estimate of cellular mRNA distributions. The two rows below the DUB names in Fig 2A summarize DUB evolutionary age, as inferred from Figure 1, and chromosomal loci, i.e. whether a DUB is on an autosome or a sex chromosome. Integration of transcriptomic and evolutionary trends in this way permits visualization of correlations between the two.

A few clusters of tissue-specificity are observable in Figure 2A. For example, there is a cluster of DUBs with relative enrichment in brain tissues encompassing known neuron-specific DUBs UCHL1 (Day and Thompson 2010) and USP11 (Vlasschaert et al. 2015) as well as USP43, MPND, OTUB1, USP46, OTUD7A, USP30, USP33, USP22, USP27X, and USP51. In

fact, UCHL1, USP11, USP22 and OTUB1, in that order, are the four DUBs with highest raw expression values in the brain (Supplementary Figure 1B). It is peculiar that USP27X and USP51, X-linked retrogenes derived from a USP22 ancestor, maintain their source gene's expression pattern. There are documented cases of retrogenes derived from aberrant transcripts where the promoter sequence is retained (McCarrey 1987); it remains to be seen whether this is the source of the conserved expression pattern in these paralogs. There exists a second set of DUBs with moderately brain-enriched expression that form a sub-cluster within a large group of DUBs enriched in lymphocytes transformed with Epstein-Barr virus (a standard method lymphocyte immortalization). A third set exists with enrichment in both these immortalized lymphocytes and in the cerebellum. Concordant with reports that EBV transformation results in enriched expression of genes related to cell cycle and immunity (Çalışkan et al. 2011), DUBs with integral roles in these processes are highly enriched in EBV-induced lymphocytes. For example, USP4 regulates the stability of multiple innate immunity proteins (e.g., TAK1 (Fan et al. 2011: 1), TRAFs 2 & 6 (Xiao et al. 2012: 2), RIG-I (Wang et al. 2013)), as well as cell cycle checkpoint regulators pRb (Blanchette et al. 2001) and ARF-BP1 (Zhang et al. 2011), a p53 antagonist. Most DUBs enriched in the immortalized lymphocytes are contrastingly depleted in whole blood cell sample (Fig 2A) from which the immortalized cell lines were derived (Lonsdale et al. 2013).

Several DUBs are enriched in muscle cells as indicated in Figure 2A. All of these except USP25 form a cluster, within which USP28 and COPS5 form a heart-specific subcluster. USP25, the ohnolog of USP28, is also testis-enriched, precluding its integration into the muscle cluster. The molecular basis of this tissue bias is known: USP25 has muscle- and testis-specific isoforms while USP28 has a heart- and brain-specific isoform that is preferentially expressed (Valero et al.

2001). Building off of their established evolutionary relationships (Figure 1), Figure 2B features pairwise plots of tissue specificity differences and absolute expression values for several evolutionarily-related DUBs.

Perhaps most intriguing in Figure 2A is the pervasive enrichment of DUBs in the testis, which defines one large cluster though is also observed in several other DUBs. This group includes all X-chromosomal retrogenes except the two derived from USP22 as well as other DUBs incurred recently during the subspecialisation of primate branches, though is not restricted to these. However, the “young DUBs”, namely USP6, OTUD6A, USP29, ATXN3L and USP26, along with USP50, form a subcluster characterized by virtual absence of expression in other tissues. It is thought that the permissive chromatin architecture in meiotic spermatocytes and post-meiotic spermatids permits widespread genetic expression in these cells (Soumillon et al. 2013); as such, many new genes are birthed with facilitated expression in germ cells, which comprise a large fraction of testicular samples (Baran et al. 2015). Comparison of open chromatin marks at OTUD6B and OTUD6A loci across several mouse tissues suggests DUB retrogenic testis-specificity is conserved (Supplementary Figure 1C).

Deubiquitinating enzymes exhibit clustered temporal expression patterns during embryonic development (Yan et al. 2013), many of which appear to become activated or deactivated at the time of embryonic activation (Figure 3A). Of note, several DUBs from the testis cluster in Figure 2A are also expressed in oocytes, where transcripts for all except ATXN3L and USP26 are detectable at some point during early embryonic development (Figure 3A; “expression” of testis-specific DUBs in the early embryo may however reflect residual male germ cell transcripts (Johnson et al. 2011)). ATXN3L was duplicated from ATXN3 in the simian ancestor (Figures 1B, 3B) and is the most effective ubiquitin cleaver of all Josephin domain-containing proteins,

attributable to the optimization of two amino acid sites (Weeks et al. 2011). The hydrophobic substitutions, S12F and R59L, were acquired in a step-wise manner during the relatively brief evolutionary lifespan of ATXN3L (Figure 3B) and only synonymous substitutions are observed at these sites in organisms diverging after the substitutions for hydrophobic residues. USP26 equally retains significant conservation across mammals. Though silenced in all other cell types including fertilized zygotes, these X-linked genes may be important for testis development as they are subject to evolutionary constraints. USP6, a non-retrogenic germ cell-specific DUB on chromosome 17 (Chr17) in humans, is a chimera derived from the fusion of the N-terminus of a *Tbc1d3* paralog and the C-terminus of a *Usp32* paralog (Paulding et al. 2003). In addition to the protein-coding genes themselves, there are multiple USP32 and TBC1D3 pseudogenes annotated on Chr17 that display varied levels of expression. Of note, several *Tbc1d3* and *Usp32* pseudogene copies have testis-specific expression while others are not expressed (Fig. 3C). Transcription at these genetic loci, whether broadly or specific to the testis, supports the idea that the “pseudogene” label does not necessarily indicate absolute absence of expression.

Continuous re-wiring of the deubiquitinating enzyme system in immunity

The ubiquitin-proteasome system (UPS) plays critical roles in innate immunity, where many deubiquitinating enzymes terminate immune responses to prevent chronic inflammation (Sun 2008). Owing to their extensive modulatory roles, CYLD and TNFAIP3 (A20) are often integrated into immune pathway schematics (Kanehisa et al. 2006) while many others have reported roles in immunity (Fig 4A). Cartilaginous fish represent the earliest diverging clade with an adaptive immune system (Flajnik and Rumfelt 2000; Venkatesh et al. 2007) and incorporate several novel innate immunity genes (Venkatesh et al. 2014). WGD in the gnathostome ancestor concomitantly generated twenty-one DUBs (Figure 1B); the retention of

some of these ohnologs may be driven by immune pathway regulation. Figure 4A depicts a snapshot of DUB immune system regulation with respect to the evolutionary emergence of each gene presented. As indicated in the legend, all components colored yellow or green emerged via WGD (i.e. they first appear in cartilaginous fish genomes) and include the entire antiviral pathway as well as the UPS components regulating it, such as DUBs USP18 (Goldmann et al. 2015), USP4 (Wada et al. 2006; Wang et al. 2013) and UCHL1 (Karim et al. 2013) and E3 ligases SOCS1 (Ungureanu et al. 2002: 1), TRIM21 (Higgs et al. 2008) and TRIM25 (Gack et al. 2007: 25). Innate immune rewiring is apparent in non-ohnologous DUBs (Fig. 1A) such as USP3 (Cui et al. 2014), USP10 (Niu et al. 2013; Wang et al. 2015), USP7 (Zapata et al. 2001; Zaman et al. 2013) and OTUD5 (DUBA) (Kayagaki et al. 2007), as well as in ancestral paralogs that have different roles in immunity than their progeny (for example USP2 (Fig 4B)). While USP2a is an ancestral gene, the two short USP2 isoforms with alternative 5' exons arose over the course of evolution. The appearance of USP2b coincides with the gene duplication event which gave rise to USP21 in bony vertebrates (Fig 4B). The other short USP2 isoform (USP2c) acts to preserve cell viability in inflammation since, contrary to USP2a, it is not inhibited by TRAF2 (Mahul-Mellier et al. 2012) (Fig 4A). The distinct 5' exon of human USP2c is only predicted to be protein-coding in certain other primates and rodents (excluding mouse). USP2c, and its distinct immune functions, may thus represent a novel innovation in the euarchontoglires ancestor. The NF- κ B Essential Modulator (NEMO) is the regulatory subunit (γ) of the IKK complex that activates the NF- κ B pathway. An ohnolog of NEMO, optineurin (OPTN), maintains structural homology though evolved to negatively regulate NEMO signaling competitively (Zhu et al. 2007) and in association with DUBs (Nagabhushana et al. 2011). Conversely, FAM105A is a conserved ohnolog of FAM105B (commonly referred to as OTULIN for "OTU deubiquitinase

with linear linkage specificity”) that bears inactivating substitutions in human. FAM105A may competitively inhibit OTULIN’s roles in immunity, akin to OPTN and NEMO, an hypothesis supported by the fact that FAM105A retains ubiquitin-binding abilities (Oshikawa et al. 2012). The substitutions that lead to the FAM105A inactivation were however only acquired during mammalian radiation (Fig 4C); thus, in other vertebrates, FAM105A may still act as a DUB.

Interferon-stimulated gene 15 (ISG15) is a ubiquitin-like modifier that emerged in gnathostomes (Loeb and Haas 1992) and mediates species-specific roles in immunity. As its name suggests, ISG15 is stimulated by interferon (IFN) signaling, as are the host enzymes that mediate its conjugation and removal from target proteins. In humans, ISG15 increases viral susceptibility and is a critical allosteric regulator of USP18 (a terminator of IFN signaling), while in mice it plays an antiviral role and does not stabilize USP18 (Speer et al. 2016). Five modified amino acids in human ISG15 relative to mice enable the NS1 protein of influenza B to target this ohnolog and is the molecular basis of species-specific infection (Yuan and Krug 2001: 15; Guan et al. 2011). Further reflective of its divergence, coding sequences of zebrafish and human ISG15 are both more similar to each species’ di-ubiquitin gene than to each other, though they are syntenic, suggesting a common origin. However divergent, ISG15 removal is nevertheless exclusively mediated by USP18 in humans, mice, and zebrafish (Chen et al. 2015). We investigated whether there are changes in the selective pressure acting on USP18 correlative to changes in ISG15. We performed a Ka/Ks analysis, which calculates the ratio of non-synonymous nucleotide substitutions per non-synonymous codon site (Ka) to synonymous substitutions per synonymous site (Ks) per for each branch of a phylogenetic tree. Strongly conserved genes have low Ka/Ks values (near 0), since greater ratios indicate that more changes to the coding sequence were retained over the course of evolution. Conventionally, it was

thought that $Ka/Ks = 1$ is indicative of neutral evolution, $Ka/Ks < 1$ is indicative of purifying selection, and $Ka/Ks > 1$ is indicative of positive selection. Although such interpretation has many problems due to mutation bias and amino acid usage bias (Xia and Kumar 2006), the relative change in Ka/Ks ratio along a phylogenetic tree may still shed light on the changing selection intensity of USP18 and ISG15. Inferences as to strength and type of selection can be more reliably drawn for longer tree branches and when the number of sites (length of aligned sequence) is larger since these lower the standard error.

The phylogenetic trees in Figure 5A and B respectively illustrate rates of change in the coding sequences of USP18 and ISG15 over the course of vertebrate evolution. A scale of 0 to 1 (represented by blue-to-red shading of branches) is used to quantify the strength of purifying selective pressure on each branch. Branches with Ka/Ks values above 1 are coloured bright red and their associated values are indicated on the tree.

The USP18 phylogeny in Figure 5A includes 14 more species than the ISG15 tree in Figure 5B because the ISG15 gene was deleted in the ancestor of Archosauria (birds and crocodilians) as well as specifically from the soft-shelled turtle (*Pelodiscus sinensis*) and gibbon (*Nomascus leucogenys*) genomes. Consistent with the hypothesis that USP18 and ISG15 co-evolve, Figure 5A illustrates that there is decreased selective pressure (increased Ka/Ks values) on the USP18 genes of birds relative to mammals. However, the evolution of these genes does not appear to be correlated in the turtle lineage: ISG15, deleted in the soft-shelled turtle, is highly conserved in painted turtles and sea turtles, while USP18 is present in soft-shelled turtles but dramatically altered relative to the inferred ancestor in sea turtles.

USP18 is relatively well conserved among mammals (Figure 5A). USP18 undergoes an appreciably higher amount of coding sequence change during boreoeutheria divergence (Ka/Ks

value of 0.84). In addition, duplication of USP18 in *Homininae* gave rise to USP41 (Fig1B), which maintains expression pattern homology and moderate sequence identity (globally aligned human proteins: 79%) though it also has a high Ka/Ks value (0.96). In contrast, ISG15 experiences relatively high rates of evolutionary change among mammals (Fig 5B). Though this may suggest that selection does not act to preserve USP18-ISG15 concordance, the sequence of ISG15 is much shorter; lower absolute Ka and Ks values increase the representation of non-synonymous mutations on the Ka/Ks ratios coded in the tree. Thus, ISG15 may be one of many interactors influencing the selective pressure acting on USP18 and USP18-derived sequences. For example, while the 14-amino acid tail required for IFNAR2 inhibition (Goldmann et al. 2015) is deleted in USP41, the catalytic residues required for ISG15 cleavage are conserved.

Neofunctionalization of ohnologs in double-strand break repair

Several deubiquitinating enzymes critically regulate the DNA damage response (DDR), including factors involved in the recognition of double-strand breaks (DSBs) and subsequent signaling to arrest the cell cycle for either apoptosis or DNA repair by non-homologous end joining (NHEJ) or homologous repair (HR) (Nishi et al. 2014; Citterio 2015; Kee and Huang 2015). Figure 6 illustrates DUB paralog involvement in DSB repair from an evolutionary perspective. The shaded boxes next to the DUB names indicate whether they are recruited to sites of DNA damage and whether they exert a quantitative effect on DNA repair, as determined by Nishi *et al* (Nishi et al. 2014). The nature of each interaction is summarized the arrow type (as defined in the Legend); these are documented along with their source literature in the Supplementary Information.

Several instances of post-duplicative innovations are apparent in Figure 6. Of note, the interaction between USP4 and the DNA endonuclease CtIP (RDDP8), critical to DSB end

resection in HR (Liu et al. 2015), represents a post-WGD neofunctionalization. Both proteins are novel ohnologs and the interaction domains on USP4 and RDDP8 do not share homology in the ancestor. That is, the insert region of USP4 is significantly different from that of USP15 (Vlasschaert et al. 2015), and the N-terminal domain of RDDP8 (pfam10482) is found only in gnathostomes (post-WGD organisms) while its C-terminal domain (pfam08573) bears homology with many earlier-diverging eukaryotes including yeast. USP4's interaction with HDAC2 likely also emerged after WGD as a small region of USP4 (a.a. 188-302), which encompasses its unique alternatively spliced exon (Vlasschaert et al. 2015; Vlasschaert et al. 2016), interacts with this HDAC1-derived ohnolog.

Other examples of ohnologous innovation include OTUB1/OTUB2, USP16/45, and USP25/28. OTUB1 binds Ubc13 (UBE2N) to inhibit K63-linked-ubiquitination of chromatin in DSB repair, while its ohnolog, OTUB2, shows much lower affinity for Ubc13 (Sato et al. 2012) and instead promotes HR by interacting with another novel ohnolog, L3MBTL1 (Citterio 2015). USP16 pairs with HERC2 to remove H2A K15-linked ubiquitin conjugates and downregulate DSB repair (Zhang et al. 2014) while its ohnolog, USP45, has no known involvements in the DNA damage response. The N-termini of human USP16 and USP45 have high sequence identity (42% for the first 400 residues), but differ markedly in exon 5, which codes for the coiled-coil HERC2 interaction domain in USP16. The USP16/45 ancestor in invertebrates carries traits of both of its progeny, though is predicted to lack the N-terminal coiled-coil domain using the COILS server (Lupas et al. 1991). Consistent with the hypothesis that USP16-HERC2 is a novel WGD-derived interaction, the coiled-coil domain first emerges in gnathostome USP16 (see Supplementary Figure 2). Finally, highly similar ohnologs USP28 and USP25 are each singularly involved in DNA repair and immune responses, respectively. USP28 associates with and

stabilizes 53BP1, Claspin and MDC1 in DSB repair (Zhang et al. 2006) (Fig 6), whereas USP25 is not known to regulate any DDR players (Zhang et al. 2006; Breuer et al. 2013). Contrarily, USP25 downregulates cellular responses to bacterial infection and autoimmunity by removing K63-linked ubiquitin on TRAF5 and TRAF6 when in complex with ACT1A (Fig 4A). USP25 also stabilizes TRAF2 (Zhong et al. 2013), TRAF3 and TRAF6 (Lin et al. 2015) to promote antiviral immunity; USP28 has no known immune interactors (Sowa et al. 2009; Breuer et al. 2013). These ohnolog-specific roles in two important cellular processes could represent subfunctionalization as the USP25/28 common ancestor in invertebrates resembles both its derivatives equally and all of the aforementioned substrates are ancestral and non-duplicated (with the exception of TRAF5, which was derived from TRAF2). Thus, USP4, OTUB2, USP16 and USP28 appear to be WGD-derived paralogs with innovative interactions in DNA repair relative to their ancestors.

DISCUSSION

Tracing the genealogy and radiation of the 93 deubiquitinating enzymes present in the human genome can help answer questions relating to their redundant and distinct roles. The collection of DUBs is indeed a superfamily composed of five functionally-related families, all of which are represented in the eukaryotic common ancestor. Members of these families cooperate to accomplish the common goal of meeting cells' dynamic deubiquitination needs in various processes, such as immune reactions and DNA repair, and often converge to serve functionally redundant roles. Duplications at several points in evolution, most notably the whole genome duplications preceding gnathostome emergence, provided evolutionary fodder for DUB system rewiring. We set out to detangle the DUB network and distinguish redundancy that results from post-duplicative conservation of interaction domains from neofunctionalization events, such as

the WGD-derived USP4-CtIP interaction in DSB repair. We also sought to predict whether novel paralogs whose cellular roles have not yet been extensively studied (e.g. USP41, ATXN3L, FAM105A) might be important nodes in these networks.

Twenty DUBs have been retained in the human genome from the two rounds of whole-genome duplication of the ancestral vertebrate genome, which contained 61 DUBs. This means that more than 160 duplicates have become nonfunctional over the course of evolution: after the second round of WGD there would be 183 new ohnologs (244 minus the original 61), of which 20 were actually retained, meaning 163 were lost. Several selective and stochastic factors determine whether genes are maintained. According to the nearly neutral theory, stochastic mutations, including deleterious ones, can become more easily fixed in species with small effective population sizes. In addition, the rates of different types of mutations vary across different genomic regions and between species. Thus, in addition to selection for beneficial mutations (or against deleterious ones), mutational biases and genetic drift play an appreciable role in losses or changes in gene function over time. These factors may explain why, for example, CYLDL is absent in mammals and ISG15 has been lost in several species including birds, gibbon and soft-shelled turtle (Fig 5B) despite possible benefit to the organism. Loss of nodes within a system due to drift can affect the selective pressure acting on other genes that regulated that node ancestrally, as is seen in the case of USP18 and ISG15 (Figure 5).

Whether they occur by ohnologous, retrogenic, or segmental means, all gene duplication events originate in single organisms. Retention of paralogs along a branch of the tree of life, for example in the mammalian lineage, requires that duplicates be prevalent enough to be fixed within the effective population at its root. Purifying selection against deleterious mutations presumably enables duplicates to attain fixation while retaining function. An evolutionary

conundrum known as Ohno's dilemma (Bergthorsson et al. 2007) consequentially follows: if selective pressure conserves the protein-coding sequence of paralogs, how do new functions emerge? The innovation, amplification and divergence (IAD) model (Bergthorsson et al. 2007) proposes that new paralogs are subject to continuous positive selection to exploit beneficial side-functions of the parent gene. The utility of this side function is "recognized" by the system via amplification of the parent gene, which facilitates fixation of a duplicate with mutations that enhance this side-function. Selection would then operate to maintain both paralogs in their divergent roles. This mechanism may explain the evolution of the USP16-HERC2 interaction (Figure 6, Supplementary Figure 2): the USP16 coil-coiled interaction domain at ~200 a.a. is absent in lancelet USP16/45 gene and in the human copy of USP45, though traces of a coiled-coil domain are present in shark USP45. It is possible that a USP16/45 copy with a rudimentary coil-coiled domain was amplified preceding gnathostome divergence and its capacity to interact with WGD-derived HERC2 was "recognized" as advantageous. Thus, selection for the coil-coiled domain of USP16 may have enabled its fixation and allowed divergence from USP45 over the course of subsequent vertebrate evolution.

Some DUB-derived genes have lost their deubiquitinase function, either ancestrally (italicized in Fig 1A) or during the course of vertebrate evolution (e.g. FAM105A in Figure 4C), and have adopted other cellular roles. They may have also also lent domains to other non-DUBs: USP7 is one of three proteins containing a meprin and TRAF homology (MATH) domain in the ancestral eukaryote (the two others are E3 ubiquitin ligases TRIM37 and SPOP), a domain that is prevalent in many vertebrate immune signalling proteins (Zapata et al. 2001), especially in plants (Liu et al. 2009). Thus, the influence of ubiquitin-conjugating system expansion directly extends beyond that which is portrayed in Figure 1.

An emerging theory in molecular evolution may offer a model of paralog retention that is complementary to the IAD model. The prerequisite amplification step in the IAD model poses an additional hurdle for retrogene retention, since these, by virtue of their method of duplication, would not normally be expressed and systemic “recognition” of their important sub-function would not be possible. In our present work, we noted that several human DUB genes that were derived relatively recently (along the mammalian lineage) are expressed specifically in the testis. The “out of the testis” gene birth model proposed by Kaessman (Kaessmann 2010) hypothesizes that the permissive open chromatin structure in meiotic spermatocytes and post-meiotic spermatids allows the expression of sequences that lack efficient promoters (retrogenes, for example) in these cells. Testis-specific expression is reportedly associated with demethylation of CpG-rich promoters (Soumillon et al. 2013) and/or endogenous retrovirus LTR exaptation (Melé et al. 2015). In the absence of a conventional promoter the potentially deleterious impact of genetic drift on gene expression would be reduced in the testis, and its potential role in neofunctionalization would subsequently be enhanced since mutations are then statistically more likely to be adaptive in this scenario. Testis-specific expression would then substitute for the “amplification” step in the IAD model in some cases, and selection for stronger promoters would enable wider expression profiles for these paralogs. Several DUBs and DUB pseudogenes with testis-specific expression may have been (or may be in the process of being) birthed out of the testis (Figure 2).

The 2R-WGD events proposed by Ohno lie in the distant past, and given the complexity of biological networks deciphering the later subfunctionalization, neofunctionalization, and nonfunctionalization of resulting ohnologs at a global level may be an insurmountable task. We have investigated deubiquitinating enzymes as a microcosm of gene evolution following 2R-

WGD, and in addition to illuminating the evolutionary relationships of DUB genes have identified specific instances in which subfunctionalization, neofunctionalization, and nonfunctionalization has occurred in the context of innate immunity and DNA repair. There is accumulating evidence in the literature that members of the same repertoire of actors play important roles in other cellular systems (developmental pathways involving Wnt or TGF- β , or tumour suppressor mechanisms). These added roles may impart added selective constraints on DUB gene evolution, which in the examples just given would be limited to metazoans in which the pathways are operative. This may be a fruitful direction for future research, but with regard to gene evolution it is clear that within the DUB superfamily there is more to be learned.

MATERIALS AND METHODS

Determination of the DUB repertoires

The Database of Ubiquitinating and Deubiquitinating Enzymes (DUDE-db) v. 1.0 (Hutchins et al. 2013) was used to derive the complete DUB repertoires of several animal, plant, and fungal genomes.

Derivation of homologous relationships

Paralogs are defined as genes within a genome sharing a common duplicative origin (e.g. whole genome duplication (WGD), small-scale duplication, retrogenic duplication). We have inferred paralogous relationships from the time of vertebrate WGD (>480 MYA) onward in Figures 1B, 4 and 5. Genes present in the agnathostome ancestor are labelled “ancestral” (Figure 1A) whereas their derivatives are qualified based on the mode of duplication. The timing of these duplications is approximated by the divergence time of the earliest branching group of animals where the new paralog is present (while verifying its absence in the syntenic region of more earlier diverging animals). DUBs that are predicted to be paralogs by the EnsemblCompara

GeneTrees pipeline (Vilella et al. 2009), including ancestral DUBs, are grouped in Figure 1B. Paralogous pairs within these groups were inferred by reciprocal best BLAST (RBB) of new paralogs to determine their most likely ancestor and their similarity was quantified by pairwise global alignment using MUSCLE (Edgar 2004) (Fig 1B).

Paralog pairs are classified as ohnologs (generated by whole genome duplication) when there is only one copy in animals diverging before the vertebrate WGD events, e.g. lancelet (*Branchiostoma floridae*), sea squirt (*Ciona intestinalis*), and sea lamprey (*Petromyzon marinus*), and two copies in cartilaginous fish (elephant shark (*Callorhinchus milii*), little skate (*Leucoraja erinacea*), small spotted catshark (*Scyliorhinus canicula*) via SkateBase (Wyffels et al. 2014)), the earliest-diverging post-WGD organisms (Venkatesh et al. 2014). We also verified that the two chondrichthyan paralogs are RBB matches with the human paralogs to ensure that they do not represent an independent duplication event in that lineage.

In cases where the parent gene clearly more closely resembles one of its paralogous progeny (on the basis of RBB and comparing domain structure), these are accordingly further classified as “ancestral” and “novel” ohnologs, labels which serve as contextual information in the evaluation of neofunctionalization.

Tissue-specific expression analysis

GTEx Analysis V4 gene RPKM values, along with accompanying sample ID annotations were downloaded from the GTEx portal (<http://www.gtexportal.org/home/>). The data was imported into R (<https://www.R-project.org/>) and DUB expression values were subsetted using a manually curated list of gene IDs. Each of the 2923 sample IDs were assigned to their source tissue using the sample annotation file and the median expression for each DUB, within each

tissue, was calculated. This value was used as the representative expression value for each tissue in downstream analysis.

To determine the tissue specificity of each DUB, the representative expression values from each tissue were standardized using a Z-score transformation. The standardized expression values of each DUB were then combined into a matrix, and tissues (rows) and DUBs (columns) were independently clustered by using complete linkage hierarchical clustering on the pairwise Euclidean distance values. These data were represented in a heatmap (Fig 2A) using the `heatmap.2` function of the `gplots` package for R (<https://CRAN.R-project.org/package=gplots>). Supplementary Figure 1 was produced in a similar manner, except the Z-score standardization was performed across DUBs, but within each tissue to effectively rank DUBs by their expression values.

For pairwise comparison of tissue enrichment between paralogous DUBs (Fig 2B), a tissue specificity score was calculated by dividing the representative expression value of a DUB in a given tissue by the median expression of that DUB across all tissues. The log₂ ratio of the specificity scores of related DUBs was then calculated to produce a log₂ fold enrichment value, representing the extent to which the expression of one DUBs within the pair is specific to a given tissue, compared to its related ohnologue. This was performed for each tissue and each DUB pair.

Embryogenesis expression plots

Single-cell RNA-Seq expression values from human oocytes and embryos at specific stages of development (zygote, 2-cell, 4-cell, 8-cell, morula, and blastocyst) were produced by Yan, *et al.* Processed RPKM values were acquired (GSE36552) and DUB expression values were subsetted using a manually curated list of DUBs. The heatmap in Fig 3A was produced using the same

approach as described above for Fig 2A. To identify clusters of distinct expression patterns throughout embryo development, the genes comprising distinct clusters (see dendrogram above the heatmap in Fig 2A) were pooled and the expression values of each gene were normalized to the median expression value of that gene throughout embryo development. These normalized values were then log2 transformed and the average log2(normalized value) for each stage was plotted.

USP18 and ISG15 phylogenies

We aligned USP18 and ISG15 codon sequences in three steps, first translating the sequences into amino acids, aligning the amino acid sequences with MUSCLE (Edgar 2004) and finally aligning the codon sequences against aligned amino acid sequences. These three steps are automated in DAMBE (Xia, 2013). Phylogenetic trees were reconstructed with PhyML (Guindon and Gascuel 2003) using the TN93 model (Tamura and Nei 1993) as well as the maximum likelihood methods and distance-based methods (with simultaneously estimated MLCompositeTN93 distance) implemented in DAMBE. For subtrees that are not strongly supported with bootstrap values, we also consulted well corroborated species phylogenies in the Tree of Life web project (<http://tolweb.org/Vertebrata>). Our objective is to obtain a well-corroborated species tree to evaluate the synonymous and nonsynonymous substitution rate (designated K_s and K_a , respectively) along the branches of the tree. We reconstructed ancestral sequences for internal nodes and then computed K_a and K_s along the tree for each branch (i.e., between two internal nodes or between a leaf node and an internal node) by using Li's method (Li 1993) implemented in DAMBE.

Figure 1. Evolutionary expansion of the entire eukaryotic deubiquitinating enzyme superfamily.

A) Pre-WGD: Expansion from the eukaryotic common ancestor set to craniata (630

MYA(Hedges et al. 2006)). All ancestral DUBs present in the human genome are categorized in this modified Venn diagram according to when they appeared evolutionarily (based on common ancestor sharing). Intersecting yeast, leaf, and sea urchin represent the three major eukaryotic kingdoms, i.e. fungi, plants, and animals. Single asterisks (*) indicate genes found in several protist genomes. The human genome retains 61 DUBs from the pre-WGD ancestor, eight of which are constitutive proteins complex members (underlined). Six have lost their isopeptidase activity (*italicized*), and another six are notably absent in the genomes of several orders of insects (†). The chronology of evolutionary events that yielded the yeast (*Saccharomyces*

cerevisiae) DUB complement is also shown, where double asterisks (**) demarcate ancestral DUBs that are present in other fungi but have been lost in yeast. Circled genes indicate DUBs that were subsequently duplicated in the fungal (orange) or plant (green) lineages. Human genome DUB nomenclature is used except in the cases of yeast-specific genes. **B) Post-WGD:**

Expansion from the gnathostome ancestor (>480 MYA(Hedges et al. 2006)) to the human genome. Arrows connect ancestral paralogs in the centre of the circle to their duplicates, which are stratified based on age of duplication. Ancestral paralogs are coloured according to Fig. 1A. Percentage similarity of each globally aligned pair is indicated on the arrowheads. The orange stratum represents ohnologs derived by whole genome duplication (WGD) roughly half a billion years ago. Paralogs derived by small-scale duplication (SSD) in the bony vertebrate ancestor, USP11 and USP21, are identified in the green stratum. The purple stratum groups functional retrotransposed DUBs incurred in mammals after the divergence of marsupials. Finally, the blue stratum indicates human DUBs acquired more recently by retrotransposition (ATXN3L),

chimerization (USP6), and other means (USP41, USP9Y). The exterior ring groups the five families within the DUB superfamily: ubiquitin-specific proteases (USPs; grey), ubiquitin C-terminal hydrolases (UCHs; green), Machado-Joseph Disease protein domain proteases (MJDs; yellow), JAMM motif proteases (JAMM; blue), and the ovarian tumour proteases (OTUs; orange). Subgroupings within the USP, MJD and OTU groups indicate paralogous groups as predicted by the EnsemblCompara GeneTrees pipeline (Vilella et al. 2009).

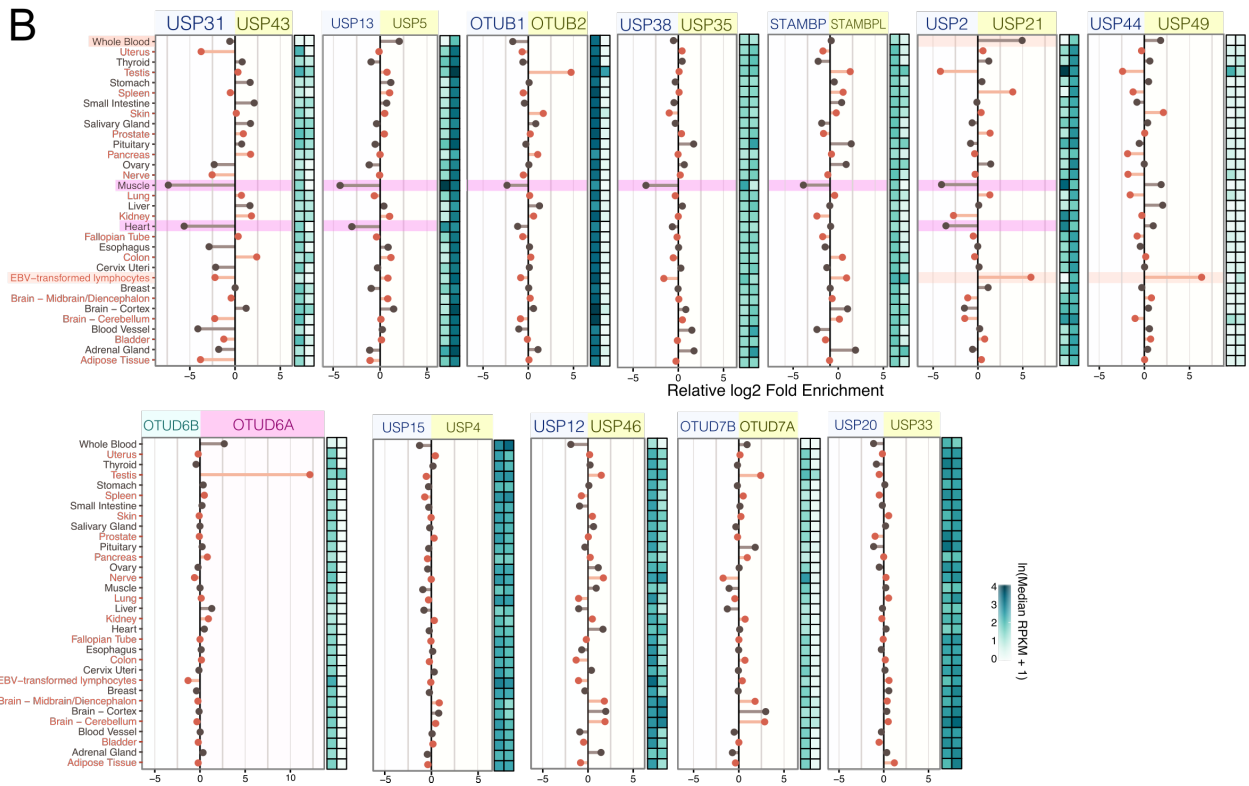
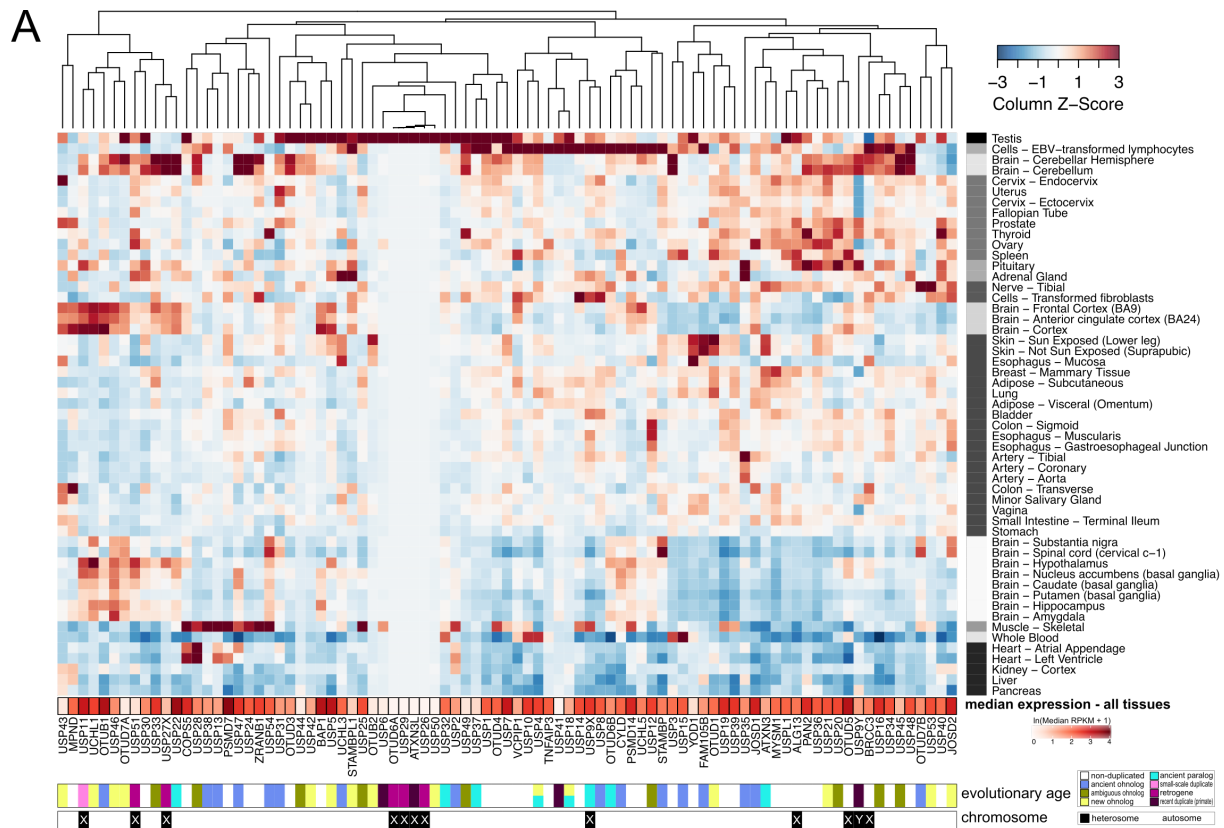
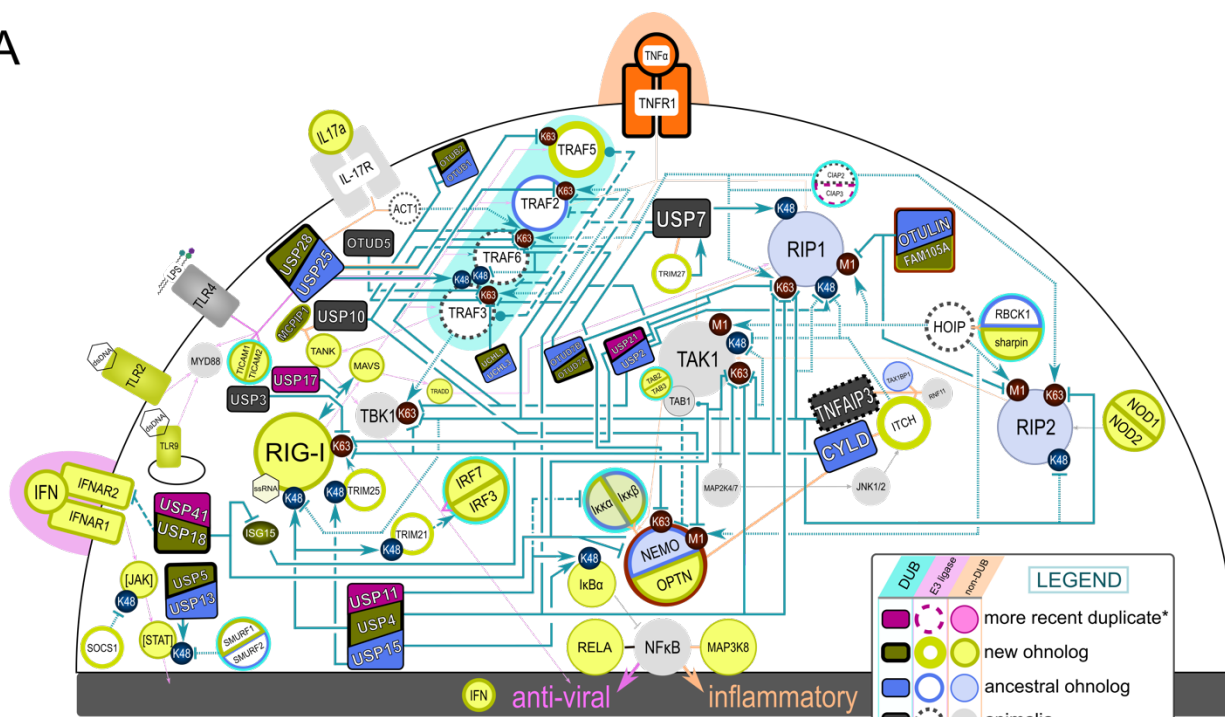


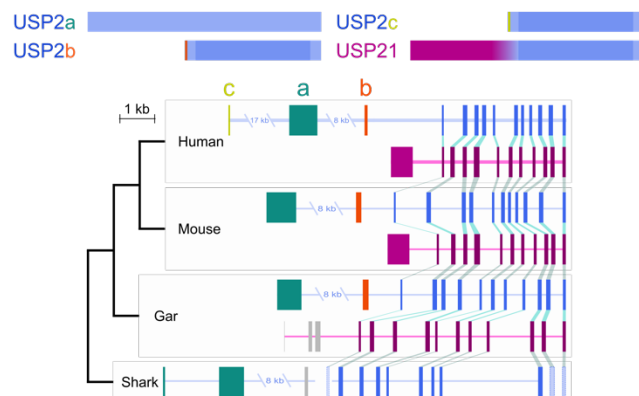
Figure 2. Tissue specificity of human deubiquitinating enzymes in an evolutionary context. **A)** Clustered heatmap of standardized expression values from the GTEx project. The dendrogram above the heatmap represents DUB clustering, while tissue clustering is represented by the shaded boxes to the right of the heatmap. Median expression level for all tissues, evolutionary age (from Figure 1) and chromosomal locus are indicated for each DUB beneath the matrix. **B)** Pairwise comparison of tissue expression enrichment ($\log_2$ Fold Enrichment; see methods for details) and raw expression values (RPKM; bluescale shaded boxes) for select DUB paralogs. Highlighted in pink, many ancestral paralogs display more than 5.6-fold enrichment (2.5 on $\log_2$ scale) in muscle tissues (heart and skeletal muscle) compared to their derivatives.

Figure 3. Selectivity of deubiquitinating enzyme expression in gametes and embryogenesis. **A)** Single-cell RNA-seq data was used to generate a clustered matrix of expression enrichment at different stages of human embryogenesis. Enrichment levels of several clusters drastically shifts around the 4-to-8 cell stage, a time when oocytal mRNA reserves become depleted and embryonic transcription begins in humans. Many of the oocyte-enriched DUBs are also testis-enriched (Fig 2A). **B)** Fitness-enhancing changes in the coding sequence of testis-specific paralog, ATXN3L, throughout primate evolution. As reported by Weeks *et al.*⁴⁰, ATXN3L is the most efficient isopeptidase of the MJD family owing to two hydrophobic acid substitutions. The S12F and R59L substitutions were sequentially fixed in the catarrhine and hominoid ancestors, respectively. **C)** Permissive expression of the chimera USP6, its parents and their pseudogenes in the testis. Domain structures of USP6, TBC1D3 and USP32 are illustrated. Most of the known pseudogenes of *Tbc1d3* (which forms the N-terminal end of the USP6 chimera) as well as multiple *Usp32* pseudogenes are especially expressed in the testis.

A



B



C

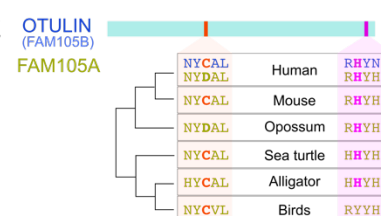


Figure 4. Evolutionary network of ubiquitin proteasomal system involvement in innate immunity. **A)** Comprehensive network of currently known DUB interactions in innate immunity. The roles of E3 ubiquitin ligase enzymes that interregulate with DUBs are also depicted. All components of the schematic are coloured according to their evolutionary age and the nature of their interactions within the system are indicated by line and arrow type (see Legend). Detailed information about the depicted interactions is available in Supplementary Tables 1 and 2. **B)** Evolution and conservation of USP2 isoforms and paralogs. USP2 has 3 isoforms (a, b, c) and one conserved paralog, USP21. Schematic depicts homology of exons throughout evolution and the emergence of new 5' exons, which are unique in each isoform and paralog. USP2a is ancestral; USP2b and USP21 were derived in bony vertebrates while USP2c is present in some but not all euarchontoglires. **C)** FAM105A, an OTULIN ohnolog, retains catalytic residues necessary for linear ubiquitin cleavage in several vertebrates but not humans.

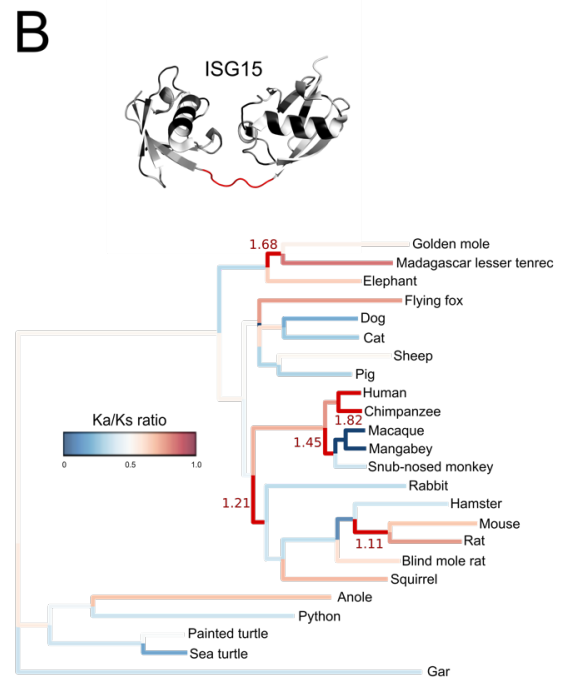
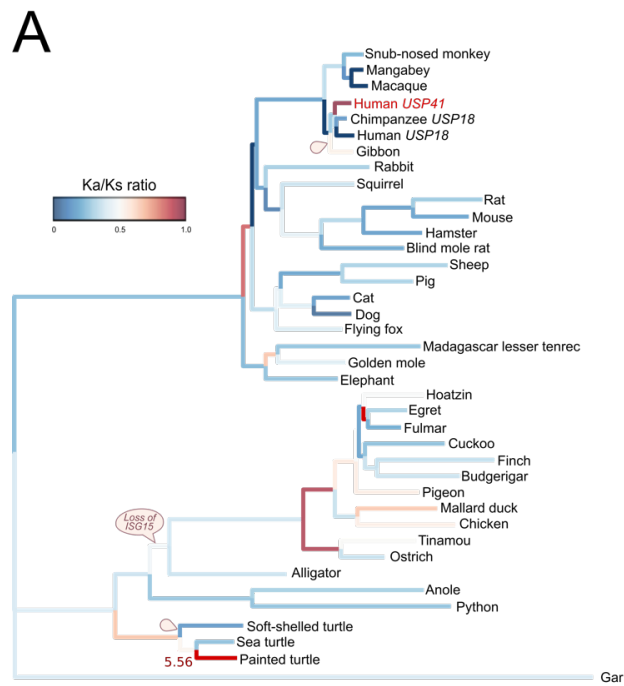


Figure 5. Changing intensity of purifying selection, measured by the ratio of nonsynonymous substitution rate over synonymous substitution rate (K_a/K_s), during the evolution and functional diversification of **A)** USP18 and **B)** ISG15. Branches are coloured according to the inset scale (blue-to- red for 0-to-1), and associated numbers are indicated near the branches when K_a/K_s values exceed 1.

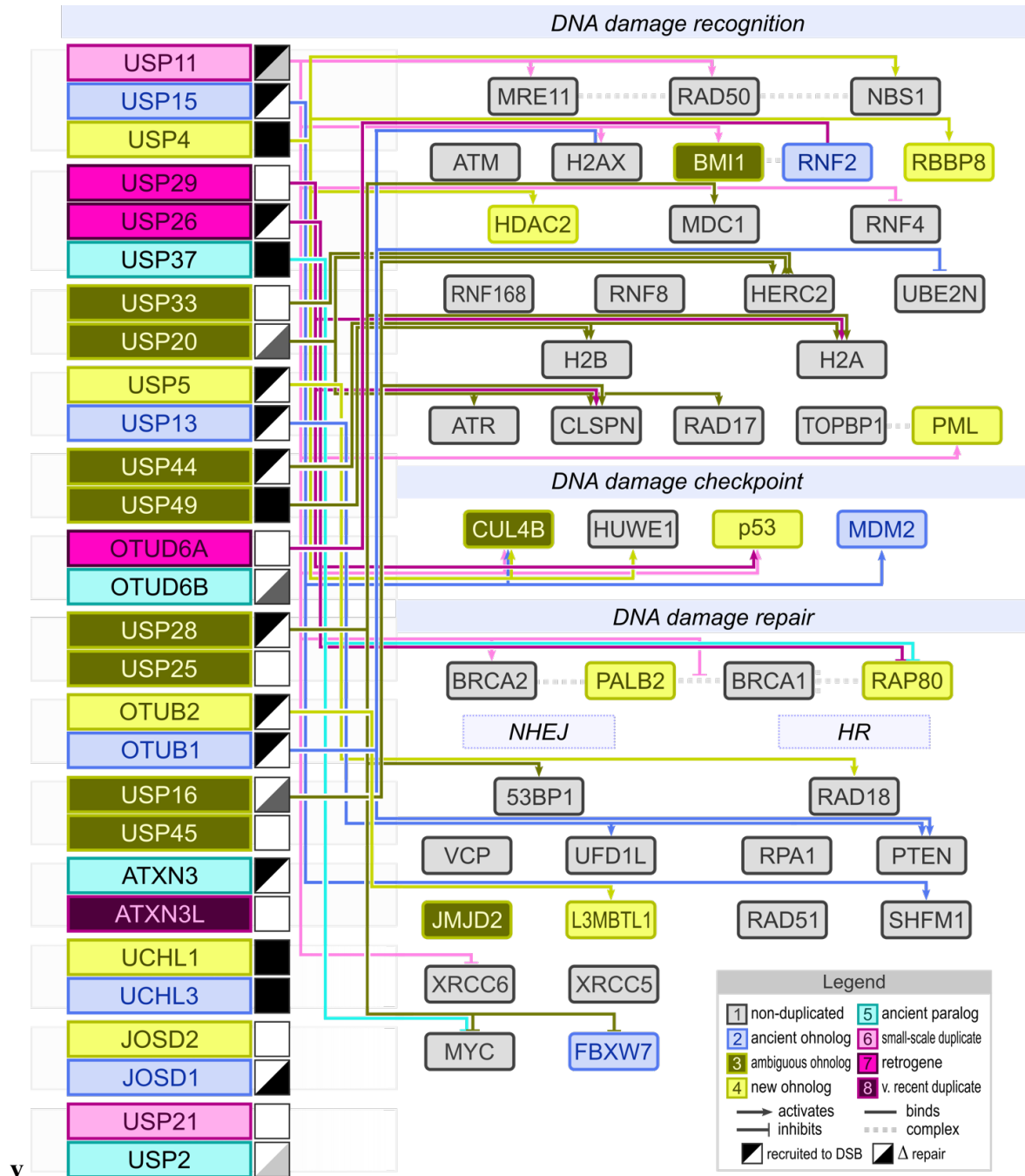


Figure 6. Rewiring of the deubiquitinating enzyme network for new roles in DNA repair.

Groups of paralogous DUBs with known involvements in double-strand break (DSB) repair (Nishi et al. 2014) are shown with their interaction partners in an evolutionary context. Paralogy is indicated by boxes to the left of the figure. Shaded boxes next to each DUB name indicate whether they are recruited to sites of DNA damage and whether they exert a quantitative effect on DNA repair, as indicated in the Legend. Detailed information about the depicted interactions is available in Supplementary Table 3.

References

- Baran Y, Subramaniam M, Biton A, Tukiainen T, Tsang EK, Rivas MA, Pirinen M, Gutierrez-Arcelus M, Smith KS, Kukurba KR, et al. 2015. The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* 25:927–936.
- Barrachina M, Moreno J, Juvés S, Moreno D, Olivé M, Ferrer I. 2007. Target Genes of Neuron-Restrictive Silencer Factor Are Abnormally Up-Regulated in Human Myotilinopathy. *Am. J. Pathol.* 171:1312–1323.
- Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: Evolution of new genes under continuous selection. *Proc. Natl. Acad. Sci.* 104:17004–17009.
- Blanchette P, Gilchrist CA, Baker RT, Gray DA. 2001. Association of UNP, a ubiquitin-specific protease, with the pocket proteins pRb, p107 and p130. *Oncogene* 20:5533–5537.
- Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock REW, Brinkman FSL, Lynn DJ. 2013. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res.* 41:D1228–D1233.
- Burrows JF, McGrattan MJ, Johnston JA. 2005. The DUB/USP17 deubiquitinating enzymes, a multigene family within a tandemly repeated sequence. *Genomics* 85:524–529.
- Burrows JF, Scott CJ, Johnston JA. 2010. The DUB/USP17 deubiquitinating enzymes: A gene family within a tandemly repeated sequence, is also embedded within the copy number variable Beta-defensin cluster. *BMC Genomics* 11:250.
- Çalışkan M, Cusanovich DA, Ober C, Gilad Y. 2011. The effects of EBV transformation on gene expression levels and methylation profiles. *Hum. Mol. Genet.* 20:1643–1652.
- Chen C, Zhang Y-B, Gui J-F. 2015. Expression characterization, genomic structure and function analysis of fish ubiquitin-specific protease 18 (USP18) genes. *Dev. Comp. Immunol.* 52:112–122.
- Citterio E. 2015. Fine-tuning the ubiquitin code at DNA double-strand breaks: deubiquitinating enzymes at work. *Cancer Genet.*:282.
- Cui J, Song Y, Li Y, Zhu Q, Tan P, Qin Y, Wang HY, Wang R-F. 2014. USP3 inhibits type I interferon signaling by deubiquitinating RIG-I-like receptors. *Cell Res.* 24:400–416.
- Day INM, Thompson RJ. 2010. UCHL1 (PGP 9.5): Neuronal biomarker and ubiquitin system protein. *Prog. Neurobiol.* 90:327–362.
- Dehal P, Boore JL. 2005. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biol.* [Internet] 3. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1197285/>

- Drouin G. 2006. Processed Pseudogenes Are More Abundant in Human and Mouse X Chromosomes than in Autosomes. *Mol. Biol. Evol.* 23:1652–1655.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
- Fan Y-H, Yu Y, Mao R-F, Tan X-J, Xu G-F, Zhang H, Lu X-B, Fu S-B, Yang J. 2011. USP4 targets TAK1 to downregulate TNF α -induced NF- κ B activation. *Cell Death Differ.* 18:1547–1560.
- Flajnik MF, Rumfelt LL. 2000. The Immune System of Cartilaginous Fish. In: Pasquier PDL, M.D PGWL, editors. *Origin and Evolution of the Vertebrate Immune System. Current Topics in Microbiology and Immunology.* Springer Berlin Heidelberg. p. 249–270.
- Gack MU, Shin YC, Joo C-H, Urano T, Liang C, Sun L, Takeuchi O, Akira S, Chen Z, Inoue S, et al. 2007. TRIM25 RING-finger E3 ubiquitin ligase is essential for RIG-I-mediated antiviral activity. *Nature* 446:916–920.
- Goldmann T, Zeller N, Raasch J, Kierdorf K, Frenzel K, Ketscher L, Basters A, Staszewski O, Brendecke SM, Spiess A, et al. 2015. USP18 lack in microglia causes destructive interferonopathy of the mouse brain. *EMBO J.* 34:1612–1629.
- Grau-Bové X, Sebé-Pedrós A, Ruiz-Trillo I. 2015. The Eukaryotic Ancestor Had a Complex Ubiquitin Signaling System of Archaeal Origin. *Mol. Biol. Evol.* 32:726–739.
- Guan R, Ma L-C, Leonard PG, Amer BR, Sridharan H, Zhao C, Krug RM, Montelione GT. 2011. Structural basis for the sequence-specific recognition of human ISG15 by the NS1 protein of influenza B virus. *Proc. Natl. Acad. Sci. U. S. A.* 108:13468–13473.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52:696–704.
- Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22:2971–2972.
- Higgs R, Gabhann JN, Larbi NB, Breen EP, Fitzgerald KA, Jefferies CA. 2008. THE E3 UBIQUITIN LIGASE RO52 NEGATIVELY REGULATES IFN- β PRODUCTION POST-PATHOGEN RECOGNITION BY POLYUBIQUITIN-MEDIATED DEGRADATION OF IRF3. *J. Immunol. Baltim. Md 1950* 181:1780–1786.
- Hutchins AP, Liu S, Diez D, Miranda-Saavedra D. 2013. The Repertoires of Ubiquitinating and Deubiquitinating Enzymes in Eukaryotic Genomes. *Mol. Biol. Evol.* 30:1172–1187.
- Johnson GD, Lalancette C, Linnemann AK, Leduc F, Boissonneault G, Krawetz SA. 2011. The sperm nucleus: chromatin, RNA, and the nuclear matrix. *Reproduction* 141:21–36.
- Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20:1313–1326.

- Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34:D354–D357.
- Karim R, Tummers B, Meyers C, Biryukov JL, Alam S, Backendorf C, Jha V, Offringa R, Ommen G-JB van, Melief CJM, et al. 2013. Human Papillomavirus (HPV) Upregulates the Cellular Deubiquitinase UCHL1 to Suppress the Keratinocyte's Innate Immune Response. *PLOS Pathog* 9:e1003384.
- Kayagaki N, Phung Q, Chan S, Chaudhari R, Quan C, O'Rourke KM, Eby M, Pietras E, Cheng G, Bazan JF, et al. 2007. DUBA: a deubiquitinase that regulates type I interferon production. *Science* 318:1628–1632.
- Kee Y, Huang TT. 2015. The Role of Deubiquitinating Enzymes in DNA Repair. *Mol. Cell Biol.*:MCB.00847–15.
- Keusekotten K, Elliott PR, Glockner L, Fiil BK, Damgaard RB, Kulathu Y, Wauer T, Hospenthal MK, Gyrd-Hansen M, Krappmann D, et al. 2013. OTULIN Antagonizes LUBAC Signaling by Specifically Hydrolyzing Met1-Linked Polyubiquitin. *Cell* 153:1312–1326.
- Komander D, Rape M. 2012. The Ubiquitin Code. *Annu. Rev. Biochem.* 81:203–229.
- Koonin EV. 2015. Archaeal ancestors of eukaryotes: not so elusive any more. *BMC Biol.* 13:1–7.
- Lin D, Zhang M, Zhang M-X, Ren Y, Jin J, Zhao Q, Pan Z, Wu M, Shu H-B, Dong C, et al. 2015. Induction of USP25 by viral infection promotes innate antiviral responses by mediating the stabilization of TRAF3 and TRAF6. *Proc. Natl. Acad. Sci. U. S. A.* 112:11324–11329.
- Liu H, Zhang H, Wang X, Tian Q, Hu Z, Peng C, Jiang P, Wang T, Guo W, Chen Y, et al. 2015. The Deubiquitylating Enzyme USP4 Cooperates with CtIP in DNA Double-Strand Break End Resection. *Cell Rep.* 13:93–107.
- Liu J, Chung H-J, Vogt M, Jin Y, Malide D, He L, Dundr M, Levens D. 2011. JTV1 co-activates FBP to induce USP29 transcription and stabilize p53 in response to oxidative stress. *EMBO J.* 30:846–858.
- Liu J, Elmore JM, Coaker G. 2009. Investigating the functions of the RIN4 protein complex during plant innate immune responses. *Plant Signal. Behav.* 4:1107–1110.
- Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* 36:96–99.
- Loeb KR, Haas AL. 1992. The interferon-inducible 15-kDa ubiquitin homolog conjugates to intracellular proteins. *J. Biol. Chem.* 267:7806–7813.

- Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45:580–585.
- Lupas A, Van Dyke M, Stock J. 1991. Predicting coiled coils from protein sequences. *Science* 252:1162–1164.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.
- Mahul-Mellier A-L, Datler C, Pazarentzos E, Lin B, Chaisaklert W, Abuali G, Grimm S. 2012. De-ubiquitinating proteases USP2a and USP2c cause apoptosis by stabilising RIP1. *Biochim. Biophys. Acta* 1823:1353–1365.
- McCarrey JR. 1987. Nucleotide sequence of the promoter region of a tissue-specific human retroposon: comparison with its housekeeping progenitor. *Gene* 61:291–298.
- Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al. 2015. The human transcriptome across tissues and individuals. *Science* 348:660–665.
- Nagabhushana A, Bansal M, Swarup G. 2011. Optineurin Is Required for CYLD-Dependent Inhibition of TNF α -Induced NF- κ B Activation. *PLoS ONE* 6:e17477.
- Nishi R, Wijnhoven P, Sage C le, Tjeertes J, Galanty Y, Forment JV, Clague MJ, Urbé S, Jackson SP. 2014. Systematic characterization of deubiquitylating enzymes for roles in maintaining genome integrity. *Nat. Cell Biol.* 16:1016–1026.
- Niu J, Shi Y, Xue J, Miao R, Huang S, Wang T, Wu J, Fu M, Wu Z-H. 2013. USP10 inhibits genotoxic NF- κ B activation by MCP1P1-facilitated deubiquitination of NEMO. *EMBO J.* 32:3206–3219.
- Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H, Chee G-J, Hattori M, Kanai A, Atomi H, et al. 2011. Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res.* 39:3204–3223.
- Oshikawa K, Matsumoto M, Oyamada K, Nakayama KI. 2012. Proteome-wide identification of ubiquitylation sites by conjugation of engineered lysine-less ubiquitin. *J. Proteome Res.* 11:796–807.
- Paulding CA, Ruvolo M, Haber DA. 2003. The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proc. Natl. Acad. Sci. U. S. A.* 100:2507–2511.
- Potrzebowski L, Vinckenbosch N, Kaessmann H. 2010. The emergence of new genes on the young therian X. *Trends Genet.* 26:1–4.

- Sato Y, Yamagata A, Goto-Ito S, Kubota K, Miyamoto R, Nakada S, Fukai S. 2012. Molecular Basis of Lys-63-linked Polyubiquitination Inhibition by the Interaction between Human Deubiquitinating Enzyme OTUB1 and Ubiquitin-conjugating Enzyme UBC13. *J. Biol. Chem.* 287:25860–25868.
- Schlegel M, Hülsmann N. 2007. Protists – A textbook example for a paraphyletic taxon. *Org. Divers. Evol.* 7:166–172.
- Schulz S, Chachami G, Kozaczekiewicz L, Winter U, Stankovic-Valentin N, Haas P, Hofmann K, Urlaub H, Ovaas H, Wittbrodt J, et al. 2012. Ubiquitin-specific protease-like 1 (USPL1) is a SUMO isopeptidase with essential, non-catalytic functions. *EMBO Rep.* 13:930–938.
- Seki T, Gong L, Williams AJ, Sakai N, Todi SV, Paulson HL. 2013. JosD1, a membrane-targeted deubiquitinating enzyme, is activated by ubiquitination and regulates membrane dynamics, cell motility, and endocytosis. *J. Biol. Chem.* 288:17145–17155.
- Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef S, Gnirke A, et al. 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* 3:2179–2190.
- Sowa ME, Bennett EJ, Gygi SP, Harper JW. 2009. Defining the human deubiquitinating enzyme interaction landscape. *Cell* 138:389–403.
- Speer SD, Li Z, Buta S, Payelle-Brogard B, Qian L, Vigant F, Rubino E, Gardner TJ, Wedeking T, Hermann M, et al. 2016. ISG15 deficiency and increased viral resistance in humans but not mice. *Nat. Commun.* 7:11496.
- Sun S-C. 2008. Deubiquitylation and regulation of the immune response. *Nat. Rev. Immunol.* 8:501–511.
- Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10:512–526.
- Ungureanu D, Saharinen P, Junttila I, Hilton DJ, Silvennoinen O. 2002. Regulation of Jak2 through the ubiquitin-proteasome pathway involves phosphorylation of Jak2 on Y1007 and interaction with SOCS-1. *Mol. Cell. Biol.* 22:3316–3326.
- Valero R, Bayés M, Francisca Sánchez-Font M, González-Angulo O, González-Duarte R, Marfany G. 2001. Characterization of alternatively spliced products and tissue-specific isoforms of USP28 and USP25. *Genome Biol.* 2:research0043.1–research0043.10.
- Varshavsky A. 2012. The Ubiquitin System, an Immense Realm. *Annu. Rev. Biochem.* 81:167–176.
- Venkatesh B, Kirkness EF, Loh Y-H, Halpern AL, Lee AP, Johnson J, Dandona N, Viswanathan LD, Tay A, Venter JC, et al. 2007. Survey Sequencing and Comparative Analysis of the Elephant Shark (*Callorhynchus milii*) Genome. *PLoS Biol* 5:e101.

- Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, Ohta Y, Flajnik MF, Sutoh Y, Kasahara M, et al. 2014. Elephant shark genome provides unique insights into gnathostome evolution. *Nature* 505:174–179.
- Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19:327–335.
- Vlasschaert C, Xia X, Coulombe J, Gray DA. 2015. Evolution of the highly networked deubiquitinating enzymes USP4, USP15, and USP11. *BMC Evol. Biol.* 15:230.
- Vlasschaert C, Xia X, Gray DA. 2016. Selection preserves Ubiquitin Specific Protease 4 alternative exon skipping in therian mammals. *Sci. Rep.* 6:20039.
- Wada K, Tanji K, Kamitani T. 2006. Oncogenic protein UnpEL/Usp4 deubiquitinates Ro52 by its isopeptidase activity. *Biochem. Biophys. Res. Commun.* 339:731–736.
- Wang L, Zhao W, Zhang M, Wang P, Zhao K, Zhao X, Yang S, Gao C. 2013. USP4 positively regulates RIG-I-mediated antiviral response through deubiquitination and stabilization of RIG-I. *J. Virol.* 87:4507–4515.
- Wang W, Huang X, Xin H-B, Fu M, Xue A, Wu Z-H. 2015. TRAF Family Member-associated NF- κ B Activator (TANK) Inhibits Genotoxic Nuclear Factor κ B Activation by Facilitating Deubiquitinase USP10-dependent Deubiquitination of TRAF6 Ligase. *J. Biol. Chem.* 290:13372–13385.
- Weeks SD, Grasty KC, Hernandez-Cuebas L, Loll PJ. 2011. Crystal structure of a Josephin-ubiquitin complex: evolutionary restraints on ataxin-3 deubiquitinating activity. *J. Biol. Chem.* 286:4555–4565.
- Wolfe K. 2000. Robustness--it's not where you think it is. *Nat. Genet.* 25:3–4.
- Wyffels J, L. King B, Vincent J, Chen C, Wu CH, Polson SW. 2014. SkateBase, an elasmobranch genome project and collection of molecular resources for chondrichthyan fishes. *F1000Research* [Internet] 3. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4184313/>
- Xiao N, Li H, Luo J, Wang R, Chen H, Chen J, Wang P. 2012. Ubiquitin-specific protease 4 (USP4) targets TRAF2 and TRAF6 for deubiquitination and inhibits TNF α -induced cancer cell migration. *Biochem. J.* 441:979–986.
- Xia X, Kumar S. 2006. CODON-BASED DETECTION OF POSITIVE SELECTION CAN BE BIASED BY HETEROGENEOUS DISTRIBUTION OF POLAR AMINO ACIDS ALONG PROTEIN SEQUENCES. *Comput. Syst. Bioinforma. Life Sci. Soc. Comput. Syst. Bioinforma. Conf.*:335–340.

- Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, et al. 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20:1131–1139.
- Yuan W, Krug RM. 2001. Influenza B virus NS1 protein inhibits conjugation of the interferon (IFN)-induced ubiquitin-like ISG15 protein. *EMBO J.* 20:362–371.
- Zaman MM-U, Nomura T, Takagi T, Okamura T, Jin W, Shinagawa T, Tanaka Y, Ishii S. 2013. Ubiquitination-deubiquitination by the TRIM27-USP7 complex regulates tumor necrosis factor alpha-induced apoptosis. *Mol. Cell. Biol.* 33:4971–4984.
- Zapata JM, Pawlowski K, Haas E, Ware CF, Godzik A, Reed JC. 2001. A Diverse Family of Proteins Containing Tumor Necrosis Factor Receptor-associated Factor Domains. *J. Biol. Chem.* 276:24242–24252.
- Zhang D, Zaugg K, Mak TW, Elledge SJ. 2006. A Role for the Deubiquitinating Enzyme USP28 in Control of the DNA-Damage Response. *Cell* 126:529–542.
- Zhang X, Berger FG, Yang J, Lu X. 2011. USP4 inhibits p53 through deubiquitinating and stabilizing ARF-BP1. *EMBO J.* 30:2177–2189.
- Zhang Z, Yang H, Wang H. 2014. The Histone H2A Deubiquitinase USP16 Interacts with HERC2 and Fine-tunes Cellular Response to DNA Damage. *J. Biol. Chem.* 289:32883–32894.
- Zhong H, Wang D, Fang L, Zhang H, Luo R, Shang M, Ouyang C, Ouyang H, Chen H, Xiao S. 2013. Ubiquitin-specific proteases 25 negatively regulates virus-induced type I interferon signaling. *PloS One* 8:e80976.
- Zhu G, Wu C-J, Zhao Y, Ashwell JD. 2007. Optineurin negatively regulates TNFalpha- induced NF-kappaB activation by competing with NEMO for ubiquitinated RIP. *Curr. Biol. CB* 17:1438–1443.

SUPPLEMENTARY FILES

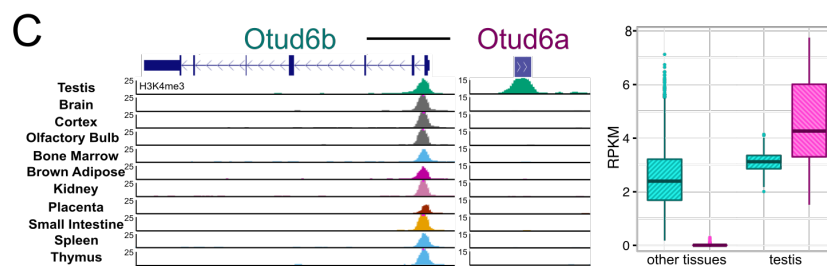
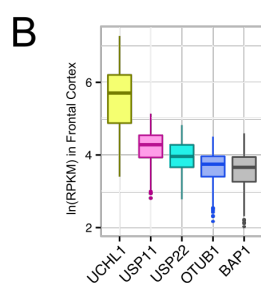
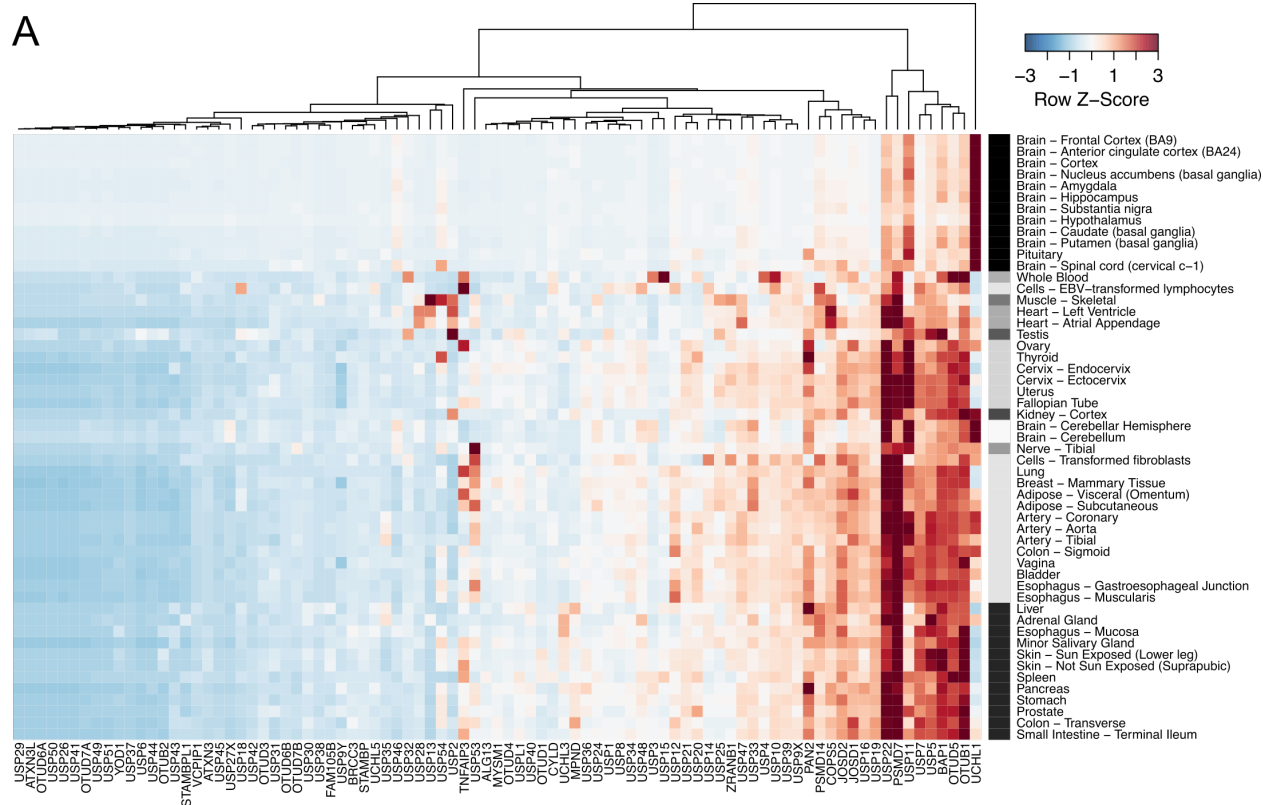
Supplementary Figure 1. Relative abundance of each DUB in human tissues.

Supplementary Figure 2. Evolution of coiled-coil domains within USP16 and USP45.

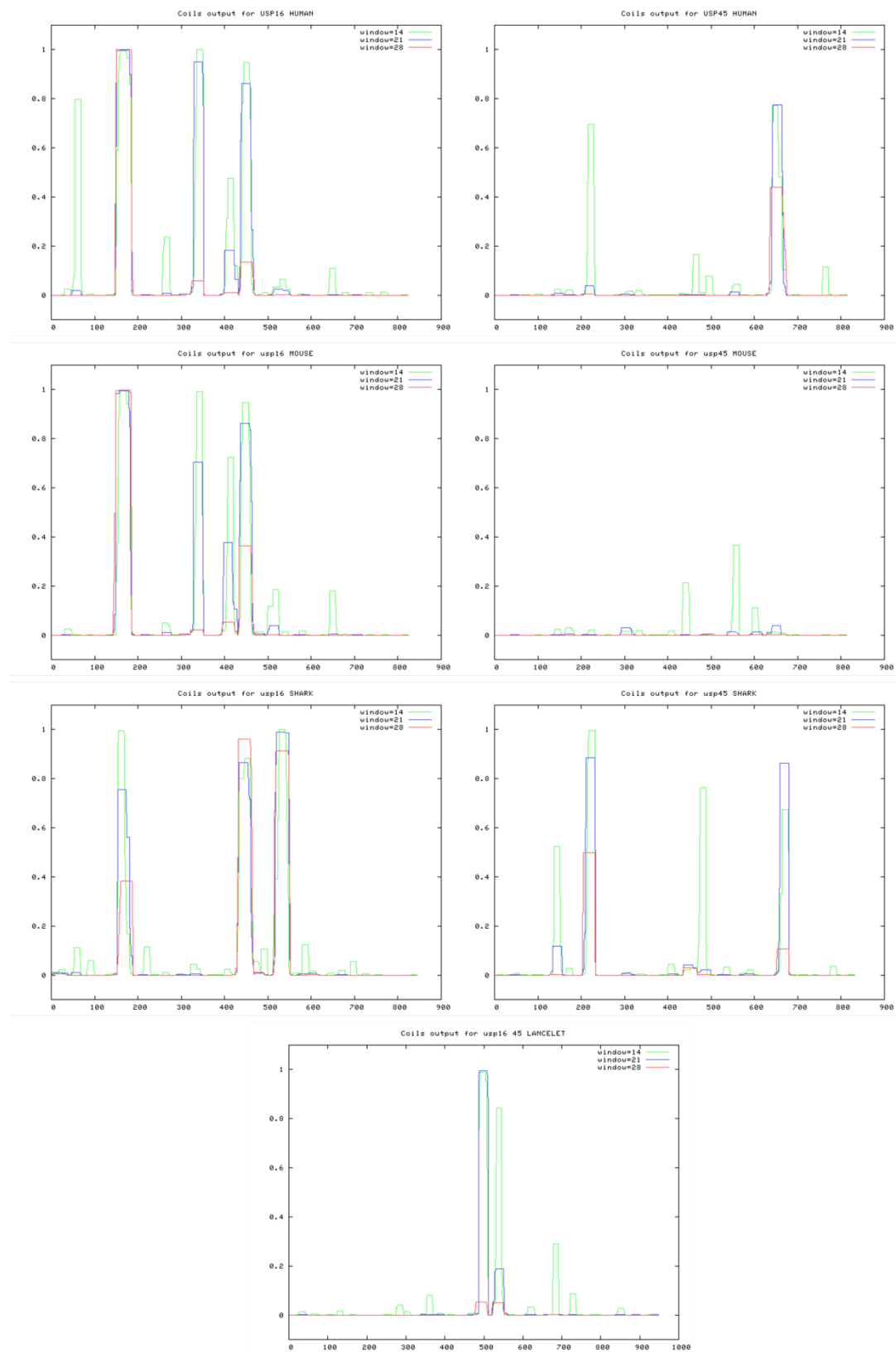
Supplementary Table 1. Detailed deubiquitinating enzyme interactions in innate immunity from Figure 4A.

Supplementary Table 2. Detailed E3 ubiquitin ligase enzyme interactions in innate immunity from Figure 4A.

Supplementary Table 3. Detailed deubiquitinating enzyme interactions in DNA double strand break repair from Figure 6.



Supplementary Figure 1. Relative abundance of each DUB in human tissues. **A)** The GTeX dataset from Figure 1A was processed and visualized using row-normalized Z-scores. Shading indicates stoichiometry of DUB transcripts per in each tissue type. **B)** Raw expression values (RPKM) for the five most abundant DUBs in the frontal cortex. **C)** Comparison of the H3K4me3 histone, modification associated with active promoters, at the genetic loci for murine *Otub6b* and *Otub6a* in several tissues. Open chromatin marks indicate *Otub6a* is selectively expressed in mouse testis, as in human, suggesting an evolutionary conservation of tissue specificity.



Supplementary Figure 2. Evolution of coiled-coil domains within USP16 and USP45. Human USP16 interacts with HERC2 via its coil-coiled domain at ~200a.a. Using the COILS server, coil-coiled domains are predicted at this location in USP16 of human, mouse and shark and also in USP45 of shark, but are absent from the ancestral sequence and USP45 of human and mouse. This suggests that the USP16-HERC2 interaction was derived by whole genome duplication and that USP16 was subsequently subfunctionalized.

The Supplementary Tables are detailed lists of deubiquitinating (DUB) and E3 ubiquitin ligase enzyme interactions in innate immunity and DNA double strand break (DSB) repair as depicted in Figures 4A and 6. Interactors are listed along with their evolutionary ages (Ev. Age) as indicated in the legend below.

Legend	
1	ancestral, no paralogs
2	ohnolog, ancestral copy
3	ohnolog, ambiguous
4	ohnolog, new copy
5	non-ohnolog paralog, ancient copy
6	small-scale duplicate paralog
7	retrogenic paralog
8	recent duplicate (primates)

Supplementary Table 1 – Detailed DUB interactions in innate immunity from Figure 4A.

Interactor 1	Ev. Age	Interactor 2	Ev. Age	Type of Interaction	Reference
USP18	5	IFNAR2	4	inhibition	(Goldmann et al. 2015)
USP18	5	NEMO	2	binding, inhibitory	(Yang et al. 2015)
USP18	5	IKBKA	3	binding, inhibitory	(Yang et al. 2015)
USP18	5	IKBKB	3	binding, inhibitory	(Yang et al. 2015)
USP18	5	TAK1	1	K63-deubiquitination	(Yang et al. 2015)
USP18	5	TAB1	1	binding	(Yang et al. 2015)
USP18	5	ISG15	4	binding, protective	(Zhang et al. 2015)
USP17	7	MAVS	4	binding, protective	(Chen et al. 2010)
USP17	7	RIG-I	4	binding, protective	(Chen et al. 2010)
USP21	6	RIG-I	4	K63-deubiquitination	(Fan et al. 2014)
USP21	6	RIP1	2	K63-deubiquitination	(Xu et al. 2010)
USP21	6	ISG15	4	binding	(Ye et al. 2011)
USP21	6	TRAF6	1	K63-deubiquitination	(Bremm et al. 2010)
USP2	5	TRAF6	1	K63-deubiquitination	(Bremm et al. 2010)
USP2(b)	5	TBK1	1	K63-deubiquitination	(L. Zhang et al. 2014)
USP2(a,c)	5	RIP1	2	K63-deubiquitination	(Mahul-Mellier et al. 2012)
USP2(a,c)	5	TRAF2	2	K63-deubiquitination	(Mahul-Mellier et al. 2012)
USP2(a,c)	5	RIP1	2	K48-deubiquitination	(Mahul-Mellier et al. 2012)
USP11	6	NFKBIA	4	K48-deubiquitination	(Sun et al. 2010)
USP11	6	IKBKA	3	inhibition	(Yamaguchi et al. 2007)
USP4	4,5	RIG-I	4	K48-deubiquitination	(Wang et al. 2013)
USP4	4,5	TRIM21	4	K48-deubiquitination	(Wada et al. 2006)
USP4	4,5	TRAF6	1	K63-deubiquitination	(Xiao et al. 2012)
USP4	4,5	TRAF2	2	K63-deubiquitination	(Xiao et al. 2012)
USP4	4,5	TAK1	1	K63-deubiquitination	(Fan et al. 2011)
USP4	4,5	RIP1	2	K63-deubiquitination	(Hou et al. 2013)
USP15	2	TRIM25	4	K48-deubiquitination	(Pauli et al. 2014)
USP15	2	NFKBIA	4	K48-deubiquitination	(Schweitzer et al. 2007)
UCHL1	4	NEMO	2	inhibition	(Karim et al. 2013)
UCHL1	4	TRAF3	1	K63-deubiquitination	(Karim et al. 2013)
OTUB2	4	TRAF6	1	K63-deubiquitination	(Li et al. 2010)
OTUB2	4	TRAF3	1	K63-deubiquitination	(Li et al. 2010)
OTUB1	2	TRAF3	1	K63-deubiquitination	(Li et al. 2010)
OTUB1	2	TRAF6	1	K63-deubiquitination	(Li et al. 2010)
USP25	2	TRAF5	4	deubiquitination	(Zhong et al. 2012)
USP25	2	TRAF6	1	deubiquitination	(Zhong et al. 2012)
USP25	2	Act1	1	binding	(Zhong et al. 2012)
USP25	2	TLR4	1	binding	(Zhong et al. 2013)
USP25	2	TRAF3	1	K48-deubiquitination	(Zhong et al. 2013)
USP25	2	MyD88	1	binding	(Zhong et al. 2013)
USP13	2	STAT1	4	K48-deubiquitination	(Yeh et al. 2013)
OTUD7B	2	TRAF3	1	K63-deubiquitination	(Hu et al. 2013)
OTUD7B	2	TRAF6	1	K63-deubiquitination	(Luong et al. 2013)
OTUD7B	2	RIP1	2	K63-deubiquitination	(Enesa et al. 2008)
OTULIN	2	RIP1	2	M1-deubiquitination	(Keusekotten et al. 2013)

OTULIN	2	RIP2	2	M1-deubiquitination	(Fiil et al. 2013)
USP3	1	RIG-I	4	K63-deubiquitination	(Cui et al. 2014)
USP7	1	TRAF5	4	binding	(Zapata et al. 2001)
USP7	1	TRAF2	2	inhibition	(Zaman et al. 2013)
USP7	1	TRAF6	1	K63-deubiquitination	(Daubeuf et al. 2009)
USP7	1	TRAF3	1	binding	(Zapata et al. 2001)
USP7	1	NEMO	2	K63-deubiquitination	(Daubeuf et al. 2009)
USP7	1	TRIM27	4	binding (USP7 activated)	(Zapata et al. 2001)
USP7	1	RIP1	2	K48, 27, 29-deUb	(Zapata et al. 2001)
USP10	1	TRAF6	1	K63-deubiquitination	(W. Wang et al. 2015)
USP10	1	NEMO	2	M1-deubiquitination	(Niu et al. 2013)
USP10	1	TANK	4	binding	(W. Wang et al. 2015)
USP10	1	MCPIP1	4	binding	(Niu et al. 2013)
TNFAIP3	1	ITCH	4	binding (complex w. TAX1BP1)	(Shembade et al. 2008)
TNFAIP3	1	RIP2	2	K63-deubiquitination	(Hitotsumatsu et al. 2008)
TNFAIP3	1	RIP1	2	K63-deubiquitination	(Wertz et al. 2004)
TNFAIP3	1	TAK1	2	K63-deubiquitination	(Skaug et al. 2011)
TNFAIP3	1	TRAF6	1	K63-deubiquitination	(Boone et al. 2004)
TNFAIP3	1	TBK1	1	K63-deubiquitination	(Parvatiyar et al. 2010)
TNFAIP3	1	NEMO	2	binding, inhibitory	(Skaug et al. 2011)
CYLD	1	ITCH	4	binding	(Ahmed et al. 2011)
CYLD	1	TAK1	1	K63-deubiquitination	(Ahmed et al. 2011)
CYLD	1	OPTN	4	binding	(Nagabhushana et al. 2011)
CYLD	1	RIP1	1	K63-deubiquitination	(Nagabhushana et al. 2011)
CYLD	1	RIG-I	4	K63-deubiquitination	(Friedman et al. 2008)
CYLD	1	TBK1	1	K63-deubiquitination	(Friedman et al. 2008)
OTUD5	1	TRAF3	1	K63-deubiquitination	(Kayagaki et al. 2007)

Supplementary Table 2 – Detailed E3 ubiquitin ligase enzyme interactions in innate immunity
from Figure 4A.

Interactor 1	Ev. Age	Interactor 2	Ev. Age	Type of Interaction	Reference
SOCS1	4	JAK	4	K48-ubiquitination	(Piganis et al. 2011)
SMURF1	4	STAT	4	K48-ubiquitination	(Yuan et al. 2012)
TRIM25	4	RIG-I	4	K63-ubiquitination	(Gack et al. 2007)
TRIM21	4	IRF3	4	binds, prevents poly-Ub	(Yang et al. 2009)
TRAF3	1	TBK1	1	K63-ubiquitination	(Kim et al. 2014)
CIAP2	5	TRAF3	1	K48-ubiquitination	(Zhong et al. 2013)
TRIM27	4	USP7	1	K63-ubiquitination	(Zaman et al. 2013)
CIAP2	5	RIP1	2	K63-ubiquitination	(Bertrand et al. 2008)
CIAP3	6	RIP1	2	K63-ubiquitination	(Bertrand et al. 2008)
TNFAIP3	1	RIP1	2	K48-ubiquitination	(Wertz et al. 2004)
TNFAIP3	1	RIP2	2	K48-ubiquitination	(Hitotsumatsu et al. 2008)
TNFAIP3	1	TAK1	1	K48-ubiquitination	(Skaug et al. 2011)
TNFAIP3	1	RIG-I	4	K48-ubiquitination	(Lin et al. 2006)
TNFAIP3	1	TRAF6	1	K48-ubiquitination	(Boone et al. 2004)
HOIP	1	RBCK1	2	binding	(Rodgers et al. 2014)
RBCK1	2	sharpin	4	binding	(Rodgers et al. 2014)
HOIP	1	sharpin	4	binding	(Rodgers et al. 2014)
HOIP	1	RIP1	2	M1-ubiquitination	(Rodgers et al. 2014)
HOIP	1	NEMO	2	M1-ubiquitination	(Rodgers et al. 2014)
HOIP	1	TAK1	1	M1-ubiquitination	(Niu et al. 2011)
ITCH	4	RIP1	2	K48-ubiquitination	(Shembade et al. 2008)
ITCH	4	TAK1	1	K48-ubiquitination	(Ahmed et al. 2011)
ITCH	4	TNFAIP3	1	binding (complex)	(Shembade et al. 2008)
ITCH	4	CYLD	1	binding (complex)	(Ahmed et al. 2011)
TRAF2	2	USP2a	2	inhibition	(Mahul-Mellier et al. 2012)
TRIM21	4	USP4	4	K48-ubiquitination	(Wada et al. 2006)

Supplementary Table 3 – Detailed DUB interactions in DSB repair from Figure 6.

Interactor 1	Ev. Age	Interactor 2	Ev. Age	Type of Interaction	Reference
USP11	6	MRE11	1	Binding	(Sowa et al. 2009)
USP11	6	RAD50	1	Binding	(Sowa et al. 2009)
USP11	6	H2AX	1	Deubiquitination	(Yu et al. 2016)
USP11	6	PML	4	De-SUMOylation	(Hendriks et al. 2015)
USP11	6	RNF4	2	Binding; Func. Inhib.	(Hendriks et al. 2015)
USP11	6	p53	4	K48-deubiquitination	(Ke et al. 2014)
USP11	6	PALB2	4	Deubiquitination	(Orthwein et al. 2015)
USP11	6	XRCC6	1	Binding	(Sowa et al. 2009)
USP11	6	BRCA2	1	Binding	(Schoenfeld et al. 2004)
USP11	6	BMI1	3	Binding	(Maertens et al. 2010)
USP11	6	CUL4B	3	Binding	(Bennett et al. 2010)
USP15	2	CUL4B	3	Binding	(Bennett et al. 2010)
USP15	2	MDM2	2	K48-deubiquitination	(Zou et al. 2014)
USP15	2	SHFM1	1	Binding	(Besche et al. 2014)
USP4	4,5	CUL4B	3	Binding	(Bennett et al. 2010)
USP4	4,5	NBS1	1	Binding	(Liu et al. 2015)
USP4	4,5	CtIP	4	Binding	(Liu et al. 2015)
USP4	4,5	RAD50	1	Binding	(Wijnhoven et al. 2015)
USP4	4,5	MRE11	1	Binding	(Wijnhoven et al. 2015)
USP4	4,5	HDAC2	4	deubiquitination	(Li et al. 2016)
USP29	7	p53	4	K48-deubiquitination	(Liu et al. 2011)
USP29	7	CLSPN	1	K48-deubiquitination	(Martín et al. 2015)
USP29	7	H2A	1	deubiquitination	(Mosbech et al. 2013)
USP26	7	RAP80	4	inhibitory	(Typas et al. 2015)
USP37	5	RAP80	4	inhibitory	(Typas et al. 2015)
USP37	5	MYC	1	K48-deubiquitination	(Pan et al. 2015)
USP33	3	HERC2	1	inhibited by HERC2	(Chan et al. 2014)
USP20	3	HERC2	1	inhibited by HERC2	(Yuan et al. 2014)
USP20	3	ATR	1	phosphorylated by ATR	(Yuan et al. 2014)
USP20	3	CLSPN	1	K48-deubiquitination	(Yuan et al. 2014)
USP20	3	RAD17	1	K48-deubiquitination	(Shanmugam et al. 2014)
USP5	4	RAD18	1	binding	(Nakajima et al. 2014)
USP13	2	UFD1L	1	binding	(Chen et al. 2011)
USP13	2	PTEN	1	K48-deubiquitination	(J. Zhang et al. 2013)
USP44	3	H2B	1	deubiquitination	(Wijnhoven et al. 2015)
USP44	3	H2A	1	deubiquitination	(Mosbech et al. 2013)
USP49	3	H2B	1	deubiquitination	(Z. Zhang et al. 2013)
OTUD6A	7	RNF2	2	binding	(Hayes et al. 2012)
OTUD6B	5	No known interactors			
USP28	3	53BP1	1	stabilization	(Zhang et al. 2006)
USP28	3	MDC1	1	K48-deubiquitination	(Zhang et al. 2006)
USP28	3	CLSPN	1	K48-deubiquitination	(Zhang et al. 2006)
USP28	3	MYC	1	K48-deubiquitination	(Popov et al. 2007)
USP28	3	FBXW7	2	binding	(Popov et al. 2007)
USP25	3	No known interactors			

OTUB2	4	L3MBTL1	4	deubiquitination	(Kato et al. 2014)
OTUB1	2	PTEN	1	binding	(G. Wang et al. 2015)
OTUB1	2	UBE2N	1	inhibitory binding	(Nakada et al. 2010)
USP16	3	HERC2	1	interaction	(Z. Zhang et al. 2014)
USP16	3	H2A	1	deubiquitination	(Shanbhag et al. 2010)
ATXN3	5	No known interactors			
ATXN3L	8	No known interactors			
UCHL1	4	No known interactors			
UCHL3	2	No known interactors			
JOSD2	4	No known interactors			
JOSD1	2	No known interactors			
USP21	4	No known interactors			
USP2	2	No known interactors			

DISCUSSION

Not all of nature's variability exemplifies evolutionary adaptation. As an example from a classic evolutionary biology paper by Gould and Lewontin (1979),

"We fault the adaptationist programme for its failure to distinguish current utility from reasons for origin (male tyrannosaurs may have used their diminutive front legs to titillate female partners, but this will not explain why they got so small)."

The authors go on to explain that the relatively small front limbs are likely by-products of increased head and hindlimb size²¹. The importance of decrypting which distinct features shaped by evolution represent adaptations, and which are incidental, is concretized by their analogy with architectural spandrels. The arrangement of arches supporting the dome of Saint Mark's Basilica is such that triangular areas, or pendentives, are created between the vertices of each pair of arches. These pendentives serve as canvases for detailed paintings and mosaics of biblical images. However, they were not created for this purpose; the existence of pendentives is consequential to the chosen architectural style. Designating these structures as "architectural spandrels" does not imply that these are dispensable, but rather that their main observable function, as decorative ornaments, is facultative. Biological spandrels like short tyrannosaurus arms similarly do not represent purposeful evolutionary adaptations but more likely are developmental byproducts of negative allometry²¹, which has since been corroborated by paleontological findings^{22,23}. While adorning artwork and tiny arms are often the main points of focus in descriptions of the pendentives of the Basilica of San Marco and tyrannosaurs, we can intuitively appreciate that the functions of these unique features do not explain their evolutionary purpose. By the same token, an evolutionary-informed approach to

understanding deubiquitinating enzyme redundancy would seek to determine which unique DUB moieties represent selective adaptations by contextualizing these in an evolutionary framework instead of proposing *raisons d'être* for all observed variability. In the present work, my co-authors and I have uncovered the evolutionary roots of all DUBs and began to answer pertinent questions as to whether the tangential branches should be biologically significant. For instance, the existence of long and short isoforms of USP4 was previously known, but these were thought to be redundant²⁴; we uncovered not only that these occupied distinct subcellular compartments (and likely exert discrete spatial roles), but that alternative splicing of USP4 is preserved by Darwinian selection. This finding has empowered ongoing work since the publication of Vlasschaert, *et al.*²⁵, which has identified phosphorylation sites within the alternatively spliced exon of USP15 that prescribe cytoplasmic retention as well as possible kinases responsible for this. Whereas USP4 alternative splicing was previously considered an evolutionary spandrel²⁴, our work supports the importance of findings such as the disturbance of USP4 isoform ratios in Paget's disease²⁶.

Our framework enabled us to identify when and how DUB genetic innovations arose relative to the emergence of their interaction partners, as well as whether these contribute to the proteome networks of different cell types. When properties are highly conserved, this is indicative of functional importance (for example the single nucleotide that permits alternative splicing of USP4); however, the opposite is not always true. We observed that several genes were absent from some organisms' genomes though remained highly conserved in most other lineages, for example USP11 and ISG15 in

birds, CYLDL in mammals, and USP4 in gorillas. It would be inaccurate to assume that these DUBs were lost because they no longer conferred a selective advantage or because they play facultative roles as genetic drift may have a considerable impact in these cases. Genomic position can bias deletion propensity for example in birds wherein extensive chromosome rearrangement is a likely culprit for gene loss²⁷. Studying organisms from which genes have been deleted can however be useful to inspect compensation within its networks. The genome of gorilla should be a useful comparative tool to study compensatory changes in USP15 or USP11, for example. In zebrafish, an in-frame insertion of a *Danio*-specific repetitive element within *usp15* alters this ancestral gene. The Gray lab has since observed that, while the *usp15* coding sequence can be translated, the resulting protein does not appear to cleave ubiquitin. USP15 is critical for proper embryonic development of mammals and zebrafish. In human, USP4, USP15 and USP11 are all key regulators of the TGF- β receptor^{28–30} while USP15 additionally regulates downstream SMAD effectors^{29,31}. Human USP15 also enhances bone morphogenic protein (BMP) signaling through deubiquitination of BMPRII³² while in zebrafish it is one of four DUBs that regulates this pathway³³. The interaction of zebrafish Usp15 with Tsc22d3 effectively blocks transduction from Bmp4 to Smad proteins and impairs early embryogenesis³⁴. Our *Usp4/Usp15* double knock-out mouse breeding experiments demonstrate that one copy of either *Usp4* or *Usp15* is the minimal requirement for survival in mice³⁵; more recent characterization of the developmental stages of these mice suggests that the DUB knock-out is lethal early in development and that deletion of *Usp15* appears to more greatly hinder the development of organs such as the liver. Going

forward, it would be interesting to more precisely establish the role of *Usp15* in murine development, as well as to compare the embryogenic roles of zebrafish *usp15* with that of other fish. Non-catalytic roles have been ascribed to several DUBs, and those of zebrafish *usp15* may be conserved in non-*Danio* species.

A single ancestral gene can be present in many forms within a proteome as paralogs and alternative spliced isoforms. As exemplified in the current work, deciphering which of these multiple versions are biological spandrels as opposed to adaptations (for example paralogs that exploit and fine-tune a minor secondary role of the parental gene⁶) can be facilitated by a combinatory bioinformatics and wet lab approach in an evolutionary context. Tracing their genealogy enables us to appreciate how the five DUB family trees are deeply rooted and how they all together sustain cellular deubiquitination needs, analogous to how a diverse group of trees could commonly provide shade for an area. The directions of growth and bifurcations of tree branches are influenced by weather, availability of nutrients, water and sunlight, and human activity; in an evolutionary genomic context, these are akin to genetic drift, spatiotemporal context and selective pressures. Branches can be selectively cut by arborists if they are intrusive, though these can be subject to random removal if they are broken, for example in thunderstorms or by the weight of snow. Like genetic drift, meteorological events can affect all floras in principle, but whether they have lasting effects is influenced by the strength of branches (effective population size) and emplacement of trees (genomic context). Humans (or other animals) may also take selective measures to protect trees and/or branches that are valuable. For example, we cover aesthetic shrubs from harsh

winters and are careful not to break branches that carry tire swings. Branches that grow in the presence of sunlight and nutrients are useful providers of shade, but can be adapted by birds as sites to build nests. As in the case of duplication and neofunctionalization of genes, the branch does not emerge because a nest needs a place to rest, but once a nest is built, its shape can be affected by erosion, sunlight, and weight of the nest. The evolution of the relationship between DUBs and some of their interactors may be understood through this branch-nest analogy. Many DUBs are critical nodes in innate immunity and DNA repair, and some predate important parts of these systems (e.g. antiviral immunity). DUB branches may have provided extensive support for the development of these “nest-works”. An evolutionary approach, such as we employ, would compare the nest-bearing branch to other nest-bearing branches as well as empty branches in order to understand how branches are shaped by their interactions with nests. It also would look at how nests are adaptively supported in instances where their original branches have been broken off. Applying an comparative evolutionary lens to the study of DUBs has provided insight into the nature of their relationships within biological networks and helped guide future studies so that we do not attempt to attribute purpose to evolutionary spandrels.

REFERENCES

1. Kravtsova-Ivantsiv, Y. & Ciechanover, A. Non-canonical ubiquitin-based signals for proteasomal degradation. *J Cell Sci* **125**, 539–548 (2012).
2. Rieser, E., Cordier, S. M. & Walczak, H. Linear ubiquitination: a newly discovered regulator of cell signalling. *Trends Biochem. Sci.* **38**, 94–102 (2013).
3. Veen, A. G. van der & Ploegh, H. L. Ubiquitin-Like Proteins. *Annu. Rev. Biochem.* **81**, 323–357 (2012).
4. Bailey-Elkin, B. A., van Kasteren, P. B., Snijder, E. J., Kikkert, M. & Mark, B. L. Viral OTU Deubiquitinases: A Structural and Functional Comparison. *PLoS Pathog.* **10**, (2014).
5. Li, W. *et al.* Genome-Wide and Functional Annotation of Human E3 Ubiquitin Ligases Identifies MULAN, a Mitochondrial E3 that Regulates the Organelle's Dynamics and Signaling. *PLoS ONE* **3**, (2008).
6. Bergthorsson, U., Andersson, D. I. & Roth, J. R. Ohno's dilemma: Evolution of new genes under continuous selection. *Proc. Natl. Acad. Sci.* **104**, 17004–17009 (2007).
7. Kondrashov, F. A. in *Evolution after Gene Duplication* (eds. Dittmar, K. & Liberles, D.) 57–76 (John Wiley & Sons, Inc., 2010).
8. Qian, W., Liao, B.-Y., Chang, A. Y.-F. & Zhang, J. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.* **26**, 425–430 (2010).
9. Xia, X. & kumar, S. Codon-based detection of positive selection can be biased by heterogeneous distribution of polar amino acids along protein sequences. *Comput. Syst. Bioinforma. Life Sci. Soc. Comput. Syst. Bioinforma. Conf.* 335–340 (2006).
10. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).
11. Nei, M., Suzuki, Y. & Nozawa, M. The Neutral Theory of Molecular Evolution in the Genomic Era. *Annu. Rev. Genomics Hum. Genet.* **11**, 265–289 (2010).
12. Ohta, T. & Gillespie, J. H. Development of Neutral and Nearly Neutral Theories. *Theor. Popul. Biol.* **49**, 128–142 (1996).
13. Consortium, T. Gte. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
14. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).
15. Dehal, P. & Boore, J. L. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biol.* **3**, (2005).
16. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
17. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
18. Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLoS Genet* **6**, e1001236 (2010).

19. Melamud, E. & Moulton, J. Stochastic noise in splicing machinery. *Nucleic Acids Res.* **37**, 4873–4886 (2009).
20. Chen, L., Tovar-Corona, J. M. & Urrutia, A. O. Alternative Splicing: A Potential Source of Functional Innovation in the Eukaryotic Genome. *Int. J. Evol. Biol.* **2012**, (2012).
21. Gould, S. J. & Lewontin, R. C. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proc. R. Soc. Lond. B Biol. Sci.* **205**, 581–598 (1979).
22. Vargas, A. The evolution of Arm Size in Theropod Dinosaurs: A Developmental Hypothesis. *Not. Mens. Mus. Nac. Hist. Nat. Chile* **338**, 16–19 (1999).
23. Bybee, P. J., Lee, A. H. & Lamm, E.-T. Sizing the Jurassic theropod dinosaur Allosaurus: assessing growth strategy and evolution of ontogenetic scaling of limbs. *J. Morphol.* **267**, 347–359 (2006).
24. Clerici, M., Luna-Vargas, M. P. A., Faesen, A. C. & Sixma, T. K. The DUSP-Ubl domain of USP4 enhances its catalytic efficiency by promoting ubiquitin exchange. *Nat. Commun.* **5**, 5399 (2014).
25. Klinck, R. *et al.* Alternative splicing in osteoclasts and Paget’s disease of bone. *BMC Med. Genet.* **15**, 98 (2014).
26. Völker, M. *et al.* Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome Res.* **20**, 503–511 (2010).

CURRICULUM VITAE for CAITLYN VLASSCHAERT

SUMMARY of QUALIFICATIONS

- Bilingual – English/French
- Dean's List for All Years of B.Sc. (9.94 DGPA) and M.Sc. (10.0 CGPA)

EDUCATION/TRAINING

Doctor of Medicine

September 2015-present

Northern Ontario School of Medicine, Sudbury, ON

Rhodes Scholarship Finalist (2015)

VP Education (2016-2018), Research Committee Representative (2015-2016)

Executive Member of following Interest Groups: Geriatrics (President), Medical and Health Research, American College of Physicians

M.Sc. Biochemistry specializing in Bioinformatics

September 2014-present

University of Ottawa, Ottawa, ON

Recipient of the M.Sc. Syed Sattar Student Award (2016)

Scholarship: Frederick Banting and Charles Best Canada Graduate Scholarship (CGS-M)

Sectional Editor, University of Ottawa Journal of Medicine (2014-present)

B.Sc. Biomedical Sciences – Cellular and Molecular Medicine option

2010-2014

University of Ottawa, Ottawa, ON

Recipient of the Faculty of Science Research Excellence Award for Honour's Project

Scholarships: Maximum Value Entry and Merit Scholarships, Francophone Bursary

Ontario Secondary School Diploma

2010

École secondaire catholique Theriault, Timmins, ON

Recipient of the Governor General award for highest overall average upon graduation (97.8%)

RESEARCH EXPERIENCE

M.Sc. Student – *Evolutionary Molecular Biology*

September 2014 – present

Dr. Douglas Gray, Ottawa Hospital Research Institute; Dr. Xuhua Xia, University of Ottawa

Thesis title: **“The evolution of the deubiquitinating enzyme superfamily”**

Paper-based thesis containing the following 3 first author publications:

1) “Evolution of the highly networked deubiquitinating enzymes USP4, USP15, and USP11”

- Published in *BMC Evolutionary Biology* (DOI: 10.1186/s12862-015-0511-1)
- Earned 2nd place at the Ottawa Institute of Computational Biology Poster Day (2015) and 2nd place at the University of Ottawa BMI Graduate Poster Day (2015)
- Presented at SMBE 2015 by Dr. Douglas Gray

2) “Selection preserves Ubiquitin Specific Protease 4 alternative exon skipping in therian mammals”

- Published in *Scientific Reports* (DOI: 10.1038/srep20039)
- Earned 1st place at the Ottawa Institute of Computational Biology Poster Day (2015)
- Poster presented at the Annual Meeting of the Society of Molecular Biology and Evolution in Vienna, Austria (SMBE 2015)

3) “The evolution and functional diversification of the deubiquitinating enzyme superfamily”

- Manuscript submitted to *Molecular Systems Biology*

Summer Student – *Microbial Bioinformatics*

May – August 2014

Dr. Fiona Brinkman, Simon Fraser University, Burnaby, BC

Title: **“Identification of protein markers for bacteria with atypical diderm membranes”**

- Third author paper in *Nucleic Acids Research* (DOI: 10.1093/nar/gkv1271)
- First author manuscript submitted to *PeerJ*
- Mentored undergraduate student for sub-project (author on *PeerJ* submission)
- Curated InnateDB database maintained by the Brinkman Laboratory (member of the *IMEx International Consortium of Molecular Interaction Databases*)
 - Added over 500 molecular interactions

Undergraduate Honour’s Project – *Molecular Biology*

September 2013-April 2014

Dr. Douglas Gray, Ottawa Hospital Research Institute; Dr. Xuhua Xia, University of Ottawa

Thesis title: **“The regulatory susceptibilities of Ubiquitin Specific Protease 4”**

- Determined the effect of phosphorylation of a pair of evolutionarily conserved serine residues in USP4
 - Created wild-type and mutant fluorescent peptide constructs
 - Performed fluorescent microscopy and flow cytometry of transfected cells

Summer Research Program (SRP) Student – *Behavioral Genetics*

July – August 2013

Dr. Carmen Sandi, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Title: **“The molecular regulation of altered NCAM polysialylation and early-life stress related to development of aggression”**

- Quantified molecular and anatomical perturbations in two mouse models of psychopathy
 - a) Genetic model: measured changes in the expression of genes related to anger, sociability and stress in mice lacking the polysialyltransferase (STX) gene
 - b) Environmental model: used a cryostat-microtome to slice and analyze changes in brains of mice exposed to early life stress (maternal neglect)
- Only Canadian student chosen for 2013 (overall applicant success rate: 3%)

Summer Research Assistant – *Geriatric Immunology*

May - June 2013

Dr. Janet McElhaney, Advanced Medical Research Institute of Canada, Sudbury, ON

- Assisted in the optimization process of a new *Xcelligence* system

Undergraduate Research Opportunity Program Student – *Cellular Signaling*

fall 2012

Dr. Stephen Gee, University of Ottawa, Ottawa, ON

- Characterized the lysine- and proline- rich PAK1 binding site for diacylglycerol kinase
- Presented poster with results at student symposium in April 2013

VOLUNTEER EXPERIENCE

Piano Teacher

October 2013 – Present

Ottawa, ON & Sudbury, ON

- Have provided free weekly lessons to a total of eight elementary school-aged children

Activities Coordinator & Fundraising Assistant

September 2011 – April 2015

Best Buddies (University of Ottawa Chapter), Ottawa, ON

- Maintained long-term friendships with individuals with developmental disabilities
- Planned monthly group activities for Club (100+ members)