Immunogenetics
DOI 10.1007/s00251-007-0268-9

ORIGINAL PAPER

# Evolutionarily conserved and divergent regions of the Autoimmune Regulator (*Aire*) gene: a comparative analysis

**Mark Saltis · Michael F. Criscitiello · Yuko Ohta · Matthew Keefe · Nikolaus S. Trede · Ryo Goitsuka · Martin F. Flajnik**

Received: 17 August 2007 / Accepted: 5 December 2007


**Abstract** During T cell differentiation, medullary thymic epithelial cells (MTEC) expose developing T cells to tissue-specific antigens. MTEC expression of such self-antigens requires the transcription factor autoimmune regulator (Aire). In mammals, defects in *aire* result in multi-tissue, T cell-mediated autoimmunity. Because the T cell receptor repertoire is randomly generated and extremely diverse in all jawed vertebrates, it is likely that an *aire*-dependent T cell tolerance mechanism also exists in nonmammalian vertebrates. We have isolated *aire* genes from animals in all gnathostome classes except the cartilaginous fish by a combination of molecular techniques and scanning of expressed sequence tags and genomic databases. The deduced amino acid sequences of Aire were compared among mouse, human, opossum, chicken, *Xenopus*, zebrafish, and pufferfish. The first of two plant homeodomains (PHD) in human Aire and regions associated with nuclear and cytoplasmic localization are evolutionarily conserved, while other domains are either absent or divergent in one or more vertebrate classes. Furthermore, the second zinc-binding domain previously named Aire PHD2 appears to have greater sequence similarity with Ring finger domains than to PHD domains. Point mutations in defective human *aire* genes are generally found in the most evolutionarily conserved regions of the protein. These findings reveal a very rapid evolution of certain regions of *aire* during vertebrate evolution and support the existence of an *aire*-dependent mechanism of T cell tolerance dating back at least to the emergence of bony fish.

**Keywords** Comparative immunology · Autoimmunity · Transcription factors · Autoimmune regulator

**Electronic supplementary material** The online version of this article (doi:10.1007/s00251-007-0268-9) contains supplementary material, which is available to authorized users.

M. Saltis · M. F. Criscitiello · Y. Ohta · M. F. Flajnik (✉)
Department of Microbiology and Immunology,
University of Maryland, Baltimore,
660 West Redwood Street, HH324,
Baltimore, MD 21201, USA
e-mail: MFlajnik@som.umaryland.edu

N. S. Trede
Division of Pediatrics, The Huntsman Cancer Institute,
University of Utah,
2000 Circle of Hope,
Salt Lake City, UT 84112, USA

M. Keefe
Division of Molecular Biology, University of Utah,
2000 Circle of Hope,
Salt Lake City, UT 84112, USA

R. Goitsuka
Research Institute for Biological Sciences,
Tokyo University of Science,
2669 Yamazaki,
Noda, Chiba 278-0022, Japan

## Introduction

As T cells develop in the thymus, they undergo a series of rearrangements of the T cell receptor variable (V) diversity (D), and joining (J) genes, resulting in the formation of a randomly generated antigen receptor repertoire with virtually unlimited diversity. In one form of negative selection in the thymus, medullary thymic epithelial cells (MTEC) expose developing T cells to a broad range of tissue-specific self-antigens (reviewed in Kyewski and Klein 2006). The expression of many of these antigens in the thymus requires the transcription factor autoimmune

regulator (Aire; Anderson et al. 2005). In mammals, *aire* is expressed primarily in subsets of MTEC and thymic dendritic cells, although in very low amounts in other tissues as well (Heino et al. 1999). Aire upregulates expression of genes encoding tissue-specific self-antigens in MTEC (e.g. insulin, retinal antigens); T cell recognition of these self-antigens either results in deletion of autoreactive cells or in selection of tissue-specific T-regulatory cells that afford protection from destructive autoimmunity (Mathis and Benoist 2007). The mechanism of Aire transcriptional regulation is unknown, but tightly clustered genes in many regions of the genome seem to be activated under its influence suggesting a remarkable type of global control of gene expression (Johnnidis et al. 2005). Autoimmune polyglandular syndrome type 1 (APS1, also known as autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy) is the result of an autosomal recessive defect in *aire* (The Finnish-German APECED Consortium 1997), characterized by autoimmunity of endocrine organs, ectodermal dystrophies, insulin-dependent diabetes, gonadal atrophy, hypothyroidism, and/or pernicious anemia (Ahonen et al. 1990).

The functional regions described for human Aire (545 amino acids) include a deoxyribonucleic acid (DNA)-binding "human Sp100, Aire1, NucP41/P75, and *Drosophila* DEAF1 domain" (SAND), four LXXLL nuclear receptor boxes (NRB), a homogeneously staining region (HSR), two zinc finger-binding plant homeodomains (PHD1 and PHD2), a proline-rich region (PRR), a nuclear localization signal (NLS), and a C-terminal domain (CTD; Fig. 1). Nuclear matrix localization and nuclear transport are affected by mutations in the NRB and NLS, respectively (Ilmarinen et al. 2006; Pitkanen et al. 2001), while the HSR functions in homodimerization (Meloni et al. 2005). The SAND domain interacts with DNA in a nonspecific manner (Bottomley et al. 2001). The Aire PHD1 domain is a Zn finger DNA-binding domain, possibly having ubiquitin ligase activity (Uchida et al. 2004).

As originally proposed by Kyewski and Klein, *aire*-dependent T cell-negative selection likely arose simultaneously with T/B cell-dependent adaptive immunity (Kyewski and Klein 2006), which originated when jawed vertebrates emerged some 500 million years ago. Therefore, we investigated the presence of *aire* orthologues in non-mammalian vertebrates. We isolated *aire* genes and complementary DNAs (cDNAs) from representatives of three nonmammalian vertebrate classes, an amphibian (*Xenopus tropicalis*), a bird (*Gallus gallus*), and a bony fish (*Danio rerio*). Partial expressed sequence tags (EST) or genomic sequences were also obtained from other species. We compared the deduced amino acid sequences and genomic organizations to published human and mouse Aire proteins. Through these comparative analyses, conserved and divergent domains and regions were revealed, and we found that some parts of *aire* evolved rapidly. Furthermore, we propose a reclassification of the domain previously described as Aire-PHD2 as a Ring-finger domain (Ring) based on phylogenetic analysis.

## Materials and methods

### Blast search EST and genomic sequences

All BLAST searches (www.ncbi.nlm.nih.gov) used the following ascension numbers and sequences: human Aire: CAA08759, NM_000383 and mouse: CAB36909, BC103511. Genomic or EST databases of each respective species were searched using BLASTx, tBLASTn, and tBLASTx with BLOSSUM 45 matrix. The following

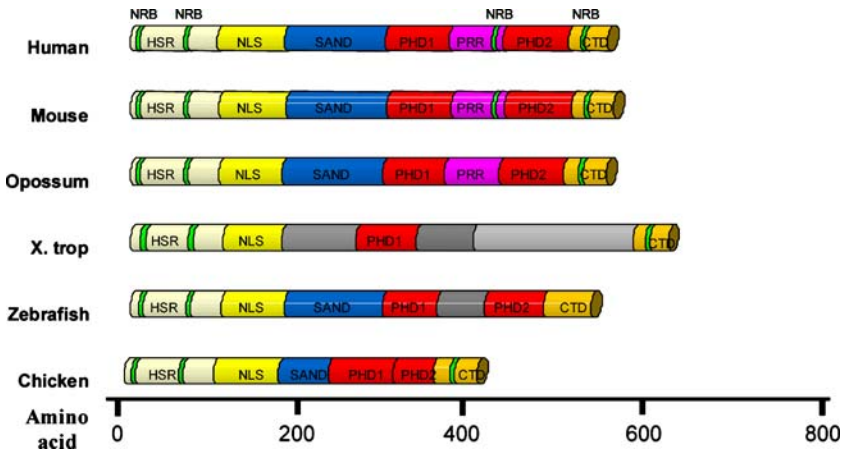



**Fig. 1** Aire domain composition in vertebrates. Comparative view of the Aire protein from human, mouse, opossum, zebrafish, frog, and chicken. The number of amino acid residues is shown at the *bottom*. The figure is to scale, and domain assignment is based on the deduced amino acid sequence compared to human Aire-1 (see Fig. 2 for accession numbers). *Light* and *dark gray shades* represent unique domains. *Green*, NRB; *tan*, HSR; *yellow*, NLS; *blue*, SAND; *red*, PHD1 and PHD2; *orange*, CTD; *purple* PRR

sequences were obtained: frog (*X. tropicalis*): DT431622, DT431623, BX709411, CR566920, genomic scaffold 55 (www.genome.JGI-psf.org v4.1), Japanese pufferfish (*Takifugu rubripes*): genomic scaffold 73 (www.genome.JGI-psf.org, v4.0), green-spotted pufferfish (*Tetraodon nigroviridis*): chromosome 15 (www.genoscope.cns.fr), Turkey (*Meleagris gallopavo*): AY235111, opossum: ENSMODG00000011425 (www.ensembl.org).

### Isolation of total RNA from frog, chicken, and zebrafish

Multiple organs were collected and pooled from 20 *X. tropicalis* and 20 *X. laevis* frogs approximately 2 months postmetamorphosis. Thymi from multiple adult zebrafish were obtained and pooled. Total ribonucleic acid (RNA) was prepared from the thymus of a single leghorn chicken (*G. gallus*) using TRI reagent (Sigma, MO) as per manufacturer's protocol and the Trizol Reagent (Invitrogen, CA) for all frog and zebrafish tissue samples.

### cDNA preparation from total RNA

First-strand cDNA was made from 1 μg total RNA using the SuperScript III First-Strand Synthesis System (Invitrogen) as per manufacturer's instructions.

### RT-PCR amplification of *aire* from chicken, frog, and zebrafish

Chicken and frog (*X. tropicalis*) polymerase chain reaction (PCR) primers for amplification were designed from the HSR to PHD1 and from PHD1 to the CTD, based on homology to human *aire* based on tBLASTn comparison. The initial primers for zebrafish partial *aire* were designed according to short homologous sequences using tBLASTn comparison of the zebrafish genome aligned to human *aire*. Partial sequences were obtained with the SMART RACE cDNA Amplification Kit (Clonetech, CA); primers were then designed to amplify the full-length *aire* by reverse transcriptase (RT)-PCR. Frog primers used in this experiment are described in Supplemental Figs. 1 and 2: *X. tropicalis* primers: 5′-GCACTGAGATAGCTGTGGCCGT-3′, 5′-GAGTTAATATGTTGCGATGGATG-3′, 5′-CATC CATCGCAACATATTAACTC-3′, 5′-ATGTTCTGA AATGCGCATTGCAG-3′; chicken primers: 5′-GCCGTG CCATGCTCTGGAATGC-3′, 5′-TGCTGAAACTG CACCGCACGGAGATCG-3′, 5′-CACGCATGAGCTG CATTGCCACGTC-3′, 5′-CGACCACGAGGATGAG TGTGCAGTGT-3′; zebrafish primers: 5′-ATGTCTAAGG TGGAGAGTTTTGAAGAGT-3′, 5′-GTGAACTTCATTGG AAATAGCCTTGGG-3′. All PCR reactions were performed using *Taq* polymerase (Invitrogen) and 1 μl cDNA with the following settings: 94°C/5 min, 35 cycles of 94°C/30 s, 50–60°C/30 s, 72°C/2–4 min. PCR products were cloned into the pCR2.1 vector, and sequences were verified by the University of Maryland, Baltimore Genomics Core facility. The following sequences were deposited in the National Center for Biotechnology Information database: *X. tropicalis*: EU004201, *X. laevis*: EU042188, EU042189: zebrafish; EU042187, and chicken: EU030003–EU030008.

### Northern blot analysis and RT-PCR of *aire* transcripts in frog

Approximately 20 μg total RNA from thymus and other tissues was loaded on agarose gel and electrophoresed for 18 h at 20 V. The RNA was blotted onto a nitrocellulose membrane. A DNA probe (bold print, Supplemental Fig. 1) was PCR-labeled with $^{32}$P-deoxycytidine triphosphate (Mertz and Rashtchian 1994) and hybridized under high stringency conditions as previously described (Bartl et al. 1997). The blot was exposed to film for 3 weeks. *aire* transcripts were detected from cDNA from a multi-tissue panel from *X. tropicalis* as described for cDNA amplification for sequencing and primers as described above.

### cDNA library construction and screening from *X. tropicalis*

Total RNA, containing ~20% thymus RNA, from *X. tropicalis* was obtained as described above. Poly(A) messenger RNA (mRNA) was isolated by using a PolyATtract mRNA isolation system (Promega, WI). The Uni-Zap cDNA library was constructed from 5 μg mRNA using the Zap-cDNA Gigapak III Gold Cloning Kit (Stratagene, CA). A $^{32}$P-labeled probe, as used in Northern blot analysis, was used to screen the library under high stringency conditions (Bartl et al. 1997).

### Comparative and phylogenetic analysis

Deduced amino acid sequences of *aire* exons as well as PHD and related RING domains were aligned with default Clustal W (Thompson et al. 1997) parameters (including gap opening penalty of 10 and gap extension penalty of 0.05). Accession numbers used were: human Aire NP_000374, mouse Aire AAI03512, human WSTF AAC97879, human KAP1 AAB37341, human MEKK1 AAC97073, human KSHV-K3 AAB62674, and human C-MIR NP_001002266. Bioedit (Hall 1999) was used to create pairwise sequence identity matrices for Table 1. The PHYLIP (Felsenstein 1989) suite of programs were employed for the phylogenetic analysis of the PHD domains. After CLUSTAL W alignment, multiple datasets for bootstrap analysis were randomly generated in Seqboot, and Prodist calculated distances from those using maximum likelihood estimates. Neighbor-joining trees were created for each with Neighbor, and

**Table 1** Deduced amino acid percent identity of each exon to human *aire*

| Human exon | Mouse | Opossum | Chicken | *X. tropicalis* | Zebrafish |
|---|---|---|---|---|---|
| 1 | 89 | 60 | 66 | 44 | 43 |
| 2 | 86 | 71 | 62 | 38 | 43 |
| 3 | 65 | 23[a] | 42 | 25 | 4 |
| 4 | 68 | 60[b] | 40 | 36 | 36 |
| 5 | 87 | 68 | 8 | 9 | 63 |
| 6 | 76 | 66 | NA | 14 | 30 |
| 7 | 48 | 50 | 27 | 30 | 22 |
| 8 | 95 | 85 | 78 | 79 | 76 |
| 9 | 56 | 32 | 15 | 21 | 18 |
| 10 | 48 | 34 | NA | 8[c] | 13 |
| 11 | 75 | 45 | NA | NA | 22 |
| 12 | 51 | 22 | 22 | NA | 19 |
| 13 | 81 | 39 | 39 | 30 | 25 |
| 14 | 73 | 44 | 63 | 39[d] | 37 |
| Overall | 74 | 53 | 46 | 27 | 37 |

SAND domains is contained within human exons 5–7; PHD1 contained within human exons 8 and 9; RING/PHD2 is contained within human exons 11 and 12; CTD is within human exons 13 and 14. *NA* No analogous exon, therefore, subsequent exons numbers shift forward by one

[a] Opossum exon 3 plus 5′ end of exon 4

[b] Opossum 3′ end of exon 4

[c] *X. tropicalis* exon 10 plus 5′ end of 11 (rest of 11 has no nonfrog analogue)

[d] Chicken exon 11, frog exon 13, see Supplemental Fig. 4 for detailed alignment

Consense was used to draw the consensus tree topology. Default settings were used to create this tree, including Dayhoff's PAM matrix, no outgroup rooting, and Majority Rule (extended). One thousand bootstrap replications were analyzed. Trees drawn with the Fitch algorithm gave similar topology to the neighbor-joining tree shown.

## Results and discussion

### Sequences of various vertebrate classes

The different domains and regions of Aire, as delineated in mouse and human, are displayed in Fig. 1. A predicted Aire protein sequence for opossum was obtained by a BLASTx search of the genome, and it was aligned to mouse and human Aire. Other than the deletion of one NRB in opossum, all of the regions aligned well in the Aire sequences from the three mammalian species. Sequences exhibited low identity between the mammalian species, particularly within the PRR (Fig. 2, Supplemental Fig. 4, Table 1).

### Analysis of frog and chicken Aire reveals poor domain homology within the tetrapods

Our original plan was to isolate *aire* from *Xenopus* for use in functional studies of tolerance during ontogeny (Kyewski and Klein 2006). Portions of the *X. tropicalis aire* sequence were obtained using tBLASTx searches with the human sequence as bait on the *X. tropicalis* genomic database. On scaffold 55, a sequence was found that was similar to the human and mouse Aire HSR, NRB, PHD1, and CTD (Fig. 1). PCR amplification from thymus cDNA was then performed to obtain partial *aire* cDNA sequences; additionally, several partial sequences were acquired that overlapped with the PCR clones in the *X. tropicalis* EST databases (Supplemental Fig. 1). We then screened a *X. tropicalis* thymus/spleen/intestine cDNA library with these probes and isolated one full-length *aire* cDNA clone (Supplemental Fig. 1). Unexpectedly, frog Aire is quite divergent from human and mouse, both in sequence and domain/region composition (Figs. 1 and 2). The second PHD domain is not present in the frog protein; frog *aire* contains one large exon in a similar position that encodes a sequence with little similarity to any of the known *aire* genes, notably lacking the characteristic Zn-binding residues (C/HXXC) in the deduced amino acid sequence (Fig. 2). This unique region was confirmed by numerous PCR-amplified and cDNA library-derived clones as well as three EST sequences. Furthermore, we verified this unusual *aire* sequence in the closely related *X. laevis* (Supplemental Fig. 1). Together, these data suggest that certain *aire* domains are poorly conserved between mammals and frog.

### Frog aire is predominantly expressed within the thymus

A panel of RNA from numerous tissues was extracted from postmetamorphic *X. tropicalis*. A single band of approximately 2,400 bp was observed on a Northern blot only in the thymus (Fig. 3a). A faster migrating band was identified in the brain, which may represent an alternatively spliced transcript. A RT-PCR also was performed, and amplification of *aire* was most prominent in the thymus (Fig. 3c). Low levels of *aire* expression were seen in the testes and brain, consistent with a study in humans (Klamp et al. 2006). These data show that despite the poor conservation in sequence between mammals and frog, like in mammals, *aire* expression is highest in the thymus of amphibians and the gene is likely involved in tolerance induction.

### Chicken *aire* shows further domain diversity within the tetrapods

Because of the low similarity between frog and mammalian Aire, we wanted to study its structure in representatives of other vertebrate classes. Partial sequences of chicken (class Aves) *aire* were obtained through BLAST searches of the chicken genome database using human and mouse sequences as queries. With no ESTs in the chicken databases, chicken *aire* PCR primers were then designed in the

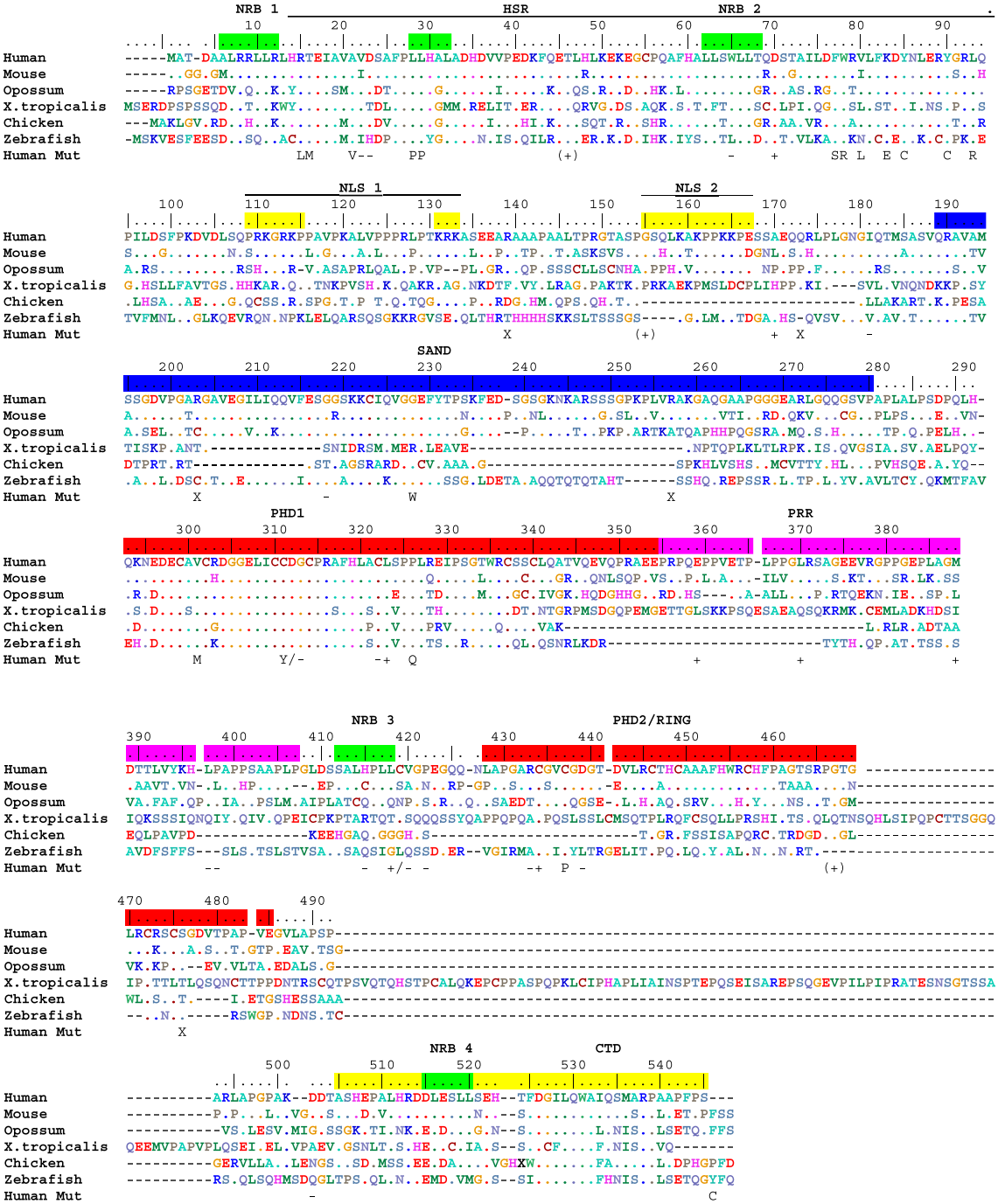



**Fig. 2** Aire amino acid sequence alignment. The deduced amino acid sequence of Aire for each species was aligned using ClustalW followed by manual manipulation for each domain. *Dots* indicate identical amino acid residues, *dashes* represent absence of an amino acid, and domain designation is color coded and represented above the human sequence. *Notation on the bottom line* denote mutations observed in human APS1 patients: *cross*, mutation resulting in premature stop codon; *minus sign*, deletion of specified residues resulting in frameshift; *plus sign*, insertion of residues resulting in frameshift; *plus sign enclosed in parentheses*, insertions in introns that result in disruption of splicing; point mutations are noted by the single letter amino acid substitution. Residues 418 and 311 are represented by more than one mutation separated by *slash*. Sequences for alignment were derived from the following accession numbers or sequence databases: human Aire: CAA08759, NM_000383; mouse: CAB36909, BC103511; *X. tropicalis*: EU004201, DT431622, DT431623, BX709411, CR566920, genomic scaffold 55 (www.genome.JGI-psf.org v4.1); zebrafish: EU042187, chicken: EU030003-EU030008; opossum: ENSMODG00000011425 (www.ensembl.org)

putative exons 1 and 14 (based on the human sequence) in the hopes of amplifying the majority of the cDNA. PCR amplification of thymus cDNA uncovered four major splice variants or potential isoforms (Supplemental Fig. 2). The sequences obtained by PCR were then aligned to the chicken genomic database. The chicken genome project is

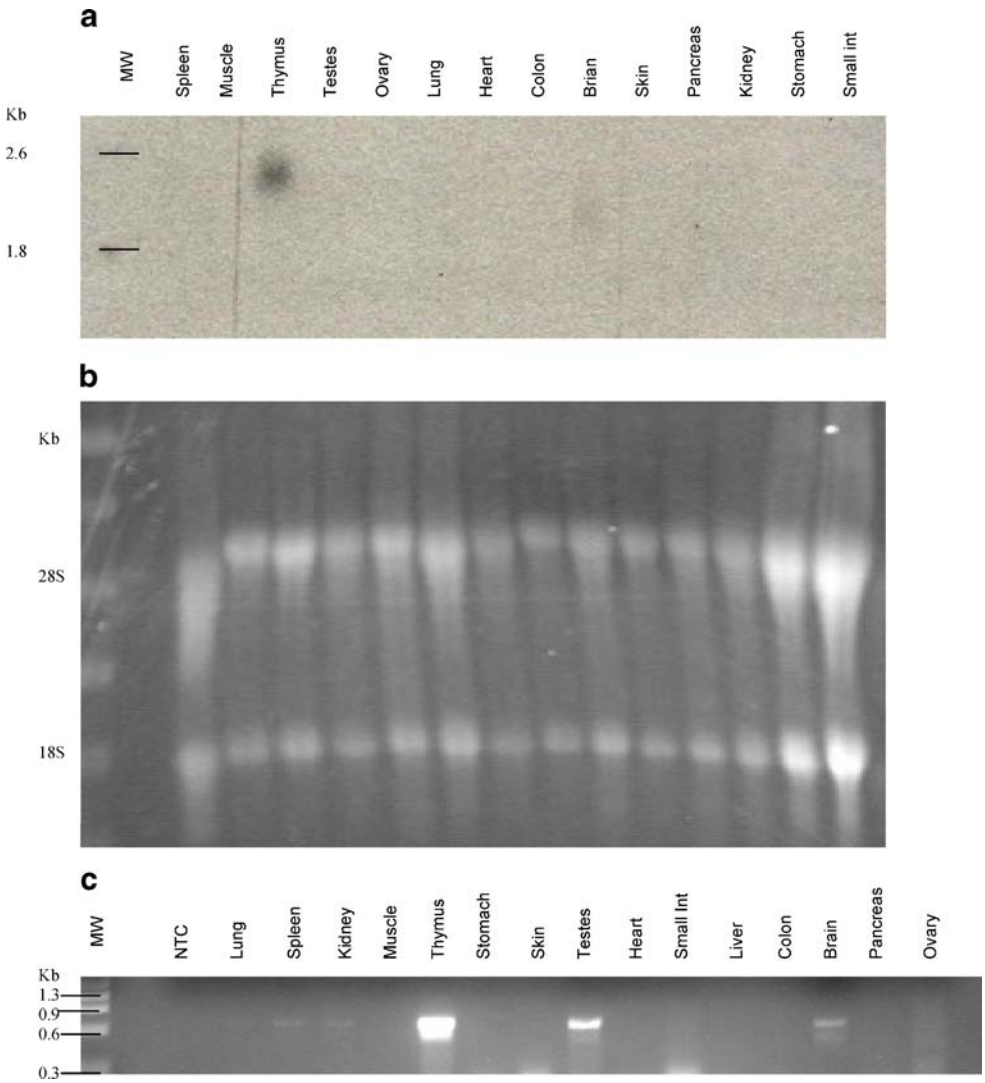



**Fig. 3** Detection of *aire* mRNA expression in various frog tissues. **a** Northern blot analysis of tissue array from *X. tropicalis*. 20 μg of total RNA was used, and the blot was probed with a DNA fragment extending from PHD 1 to the CTD (Supplemental Fig. 1). Tissues are noted above each lane. **b** An image of ribosomal 18S and 28S RNA bands was used as a loading control. **c** RT-PCR analysis of the tissue array from young frogs aged 2 months, using 35 cycles, primers from the HSR to upstream of PHD1 and cDNA made from 1 μg of RNA. All size markers are noted in the *leftmost lane*

only partially completed, and large gaps within the genomic region for *aire* were evident (Shiina et al. 2007). Using the same specific primers as were used for cDNA amplification, the chicken *aire* was amplified from genomic DNA. Consistent with the compact nature of the chicken genome, chicken *aire* spans only ~3,500 bp. The *aire* cDNA was then aligned to the genomic DNA sequences and further compared to a partial turkey genome sequence (Supplemental Fig. 2); the deduced amino acid sequences were compared to the human, chicken, and frog sequences (Figs. 1 and 2). Unlike frog Aire, certain conserved residues of the SAND and PHD2 domains are evident. However, two of the PHD2 ion-binding residue pairs (C/HxxC) are absent (Fig. 2 and Supplemental Fig. 4), and parts of the exon encoding PHD2 are fused to the second exon of PHD1 (equivalent to exons 8 and 10 in human), thus eliminating the PRR. From these data, *aire* has clearly undergone significant diversification in more than one class of tetrapods. However, like the frog protein, chicken Aire retains several domains found in the human—HSR, 3 NRB, NLS, PHD1, and CTD—which are presumed to be indispensable for function.

## Aire from teleost fishes is more similar to mammals than to frog or chicken

After observing large variations in the domain composition of Aire in tetrapods, we isolated *aire* from the zebrafish.

The full-length *aire* cDNA was sequenced using primers designed from the 5′ and 3′ untranslated region, and the translated sequence was compared to those of other species. In addition to HSR, NLS, PHD1, and CTD found in all of the tetrapods analyzed, zebrafish Aire also contains the PHD2 and SAND domains (Fig. 1). No sequence similarity was observed between the human PRR and the zebrafish sequence; however, a segment of the zebrafish and frog Aire contains a serine-rich region downstream of the PHD1, which might serve the same function. The sequences were further compared for amino acid identity (Fig. 2). Our results indicate that the Aire domain composition is more highly conserved between the fishes and mammals than representatives of other tetrapod classes. *aire* transcription was determined by RT-PCR of a multiple-tissue cDNA panel, and expression was only detected in the thymus (data not shown). Deduced amino acid sequence comparisons between zebrafish Aire and the two species of pufferfish were obtained by Clustal W with minor manual adjustment. These alignments revealed gaps within the SAND domain between pufferfish and zebrafish, suggesting variance even among teleosts (Supplemental Fig. 3). Additionally, a sequence was identified from the Elephant shark genome homologous to a small portion of the teleost PHD1 (data not shown). However, at this time, we are unable to verify that this fragment is in fact *aire* from cartilaginous fish.

## Analysis of protein translations

### The NRB, NLS, and HSR are conserved

We analyzed the N-terminal sequence of Aire containing the HSR, NRB, and NLS for all species (Fig. 2). Numerous residues of the HSR are involved in dimerization (Meloni et al. 2005), which is essential for the function of Aire in humans; APS1 is observed in numerous patients with HSR mutations (Fig. 2). The human HSR has a predicted tertiary structure containing four alpha helices and encompasses two NRB, and this feature is present in all vertebrate classes studied. The conserved nature of the HSR in other vertebrates when compared to humans suggests that this domain is also required for dimerization in other vertebrates.

NRB are common in transcription factors and are essential for proper nuclear localization. The NRB is an α-helical motif required for ligand-dependent binding of coactivators of transcription to nuclear receptor proteins. The N-terminal and C-terminal NRB, as well as the inverted NRB at amino acid position 27–33, are conserved in all species (Fig. 2, Supplemental Fig. 4). NRB bind accessory factors 1 or 2 complexes with variable specificity, and they have been organized into three major classes (Chang et al. 1999). Two general consensus sequences within the Aire NRB were detected that differ from the three described NRB classes, a charged (D/E)L(R/D)XLL and uncharged A(L/I)LXXLL or ALXXLL (Fig. 2, Supplemental Fig. 4). These data show that the Aire NRB comprise a unique subset of evolutionarily conserved NRB (with unknown ligands).

Potential NLS are evident in all species in which *aire* has been isolated (Fig. 1, yellow box). Our analysis of human Aire shows two potential NLS (Fig. 2, amino acid residues 110–133 and 155–167). Residues 113–133 have been determined to function as a monopartite NLS that interacts with the minor binding site of importin-α family proteins, including α3 and α5 and a lesser extent to α1 (Ilmarinen et al. 2006). These previous studies of the Aire NLS were based on mutagenesis of residues 113, 114, and 131–133. Loss of NLS activity was evident only after mutation of residues 131–133, indicating a likely monopartite NLS. In contrast to these findings, minor-site binding is typically consistent with bipartite NLS, while monopartite NLS bind to the major binding site. These findings did not take into consideration upstream residues 110–111 that may replace the deleted residues at 113–114; our data show that an equivalent to the human Aire at residues 110–111 or 113–114 (R or K) is conserved in all species, as well as residues corresponding to human Aire 131–133. Based on the conserved nature of the NLS and previous data showing minor groove binding, we theorize that the Aire NLS at residues 110–133 still potentially functions as a bipartite NLS.

A second potential NLS is found downstream of this first NLS at 159–167. A consensus sequence of GXXX KXPPKK(D/E) is observed, and similarity to a subset of monopartite NLS exists. However, NLS typically conform loosely to consensus sequences, which are mostly identified by amino acid composition and confirmed by mutagenesis assays (Nair et al. 2003). Further study is needed to analyze the function of these conserved residues in relation to nuclear localization. However, the conserved nature of the residues 159–167 suggests that this motif is a candidate for investigation as a second Aire NLS.

### The SAND domain is poorly conserved

As mentioned, the SAND domain has been implicated in direct DNA binding. The presence of a SAND domain has been reported previously in human Aire (Kumar et al. 2001). We aligned the SAND domains from each species (Supplemental Fig. 4) and found that conserved residues in this domain are present in zebrafish and mammalian Aire. However, a few of these residues are observed in chicken or frog suggesting either a different function of the SAND domain or that the absence of this domain does not impair Aire function.

The PHD1 domain is highly conserved between species

The Aire PHD1 domain is well conserved in all vertebrates examined. Aire PHD1 has been predicted to function in protein/protein interactions because the 3D structure shows overall negatively charged residues on the surface (Bottomley et al. 2005). These negatively charged residues are also conserved in all species. Moreover, the region between the third and fourth H/CXXC Zn-binding motif has been found to be crucial for the PHD domain binding affinity. Based on these findings, it is likely that Aire PHD1 binds a similar or the same substrate in the sequences analyzed. The chicken *aire PHD1* exon exhibits a unique genomic arrangement where the second portion of it is fused on the same exon as *PHD2* (Supplemental Fig. 2). At this time, we can only speculate that this is a consequence of the general compaction of the chicken genome and this potentially has no functional significance (Schmutz and Grimwood 2004).

The proline-rich region is characteristic only of mammalian Aire

Alignment of the region downstream of the first PHD domain shows that the PRR is present only in mammals and that even among mouse, human, and opossum, it is divergent in sequence (Table 1 and Supplemental Fig. 4). The "homologous" region in the frog and fish is a serine-rich section, and chicken totally lacks this portion of the protein, as the *PHD1*- and *PHD2*-coding regions are joined in exon 8. There are no obvious PRR repetitive patterns in

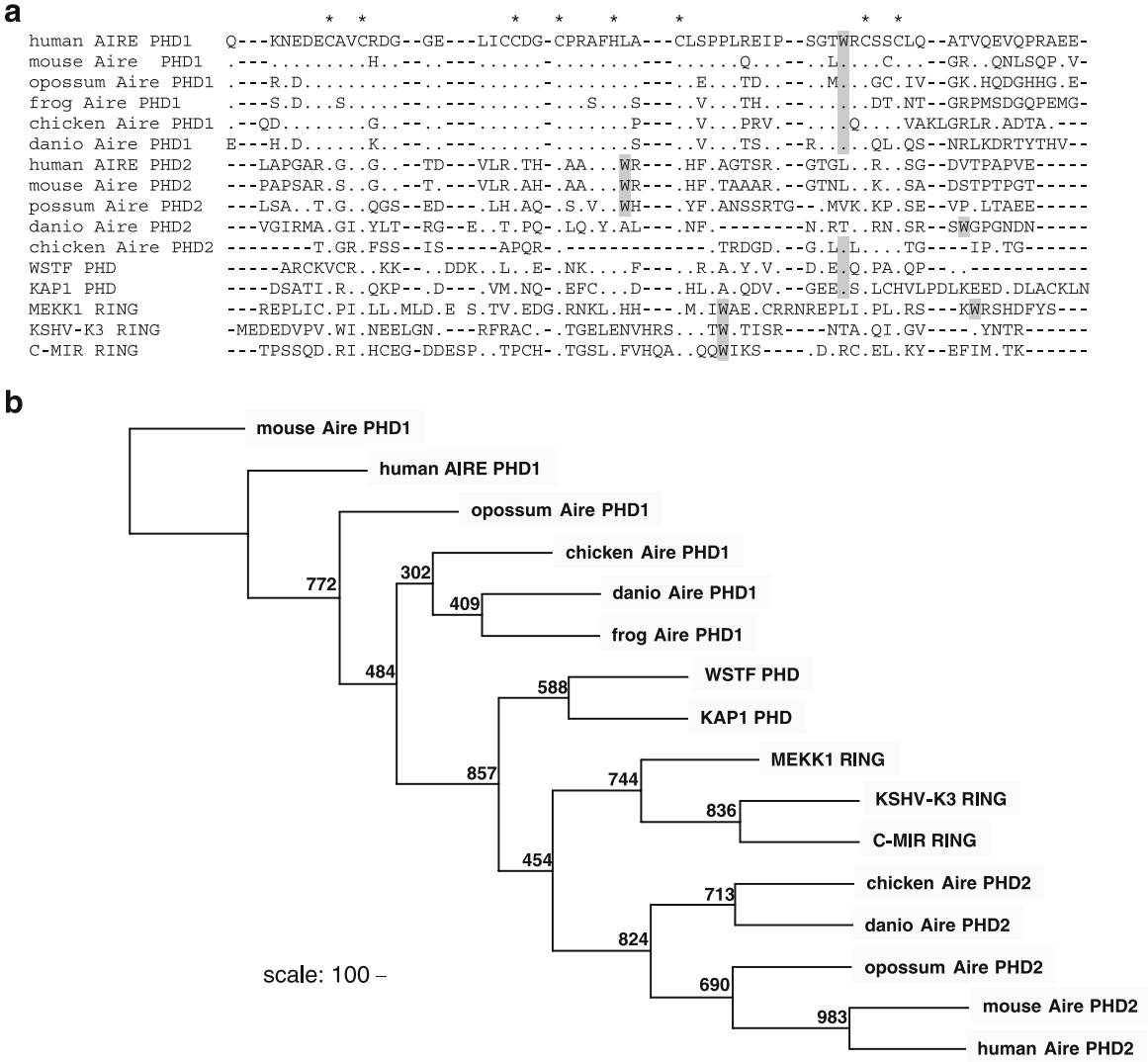



**Fig. 4** Amino acid alignment of Aire PHD domains with RING domains and phylogenetic tree. **a** PHD domains of Aire from the discussed species were aligned with non-Aire PHD domains and RING finger domains. *Hyphens* show gaps introduced in the sequences, and *dots* show identity with top human AIRE sequence. *Asterisks* mark Zn-binding residues. *Shaded boxes* denote conserved tryptophan residues. **b** Neighbor-joining tree drawn from alignment in **a**. *Bar at the bottom* shows genetic distance. Support out of 1,000 bootstrap replications is shown at *nodes*

Aire as is observed in other proteins with PRR (Williamson 1994). In human and mouse, there exists a potential NRB (ALXXLL) within the PRR, but it is unclear whether this sequence is functional. Based on these data, we hypothesize that the PRR is likely to be structural in nature because it is neither represented by missense mutations in human disease (Fig. 2) nor conserved evolutionarily.

## PHD2 aligns with ring finger domains instead of PHD domains

The Aire PHD2 is a putative Zn-binding domain with negative charges on the predicted exposed surfaces (Bottomley et al. 2005). The frog completely lacks PHD2, and the first half of the chicken PHD2 is fused on the same exon as the second half of PHD1. A phylogenetic analysis was performed including the Aire PHD domains and several other proteins containing Ring finger domains (Fig. 4); note that PHD and Ring domains are members of the "treble class family" of Zn-binding domains, but there is only weak structural similarity between them (Grishin 2001). The Aire PHD2 from all species groups more closely with ring finger domains (Ring). Additionally, the PHD1 and other PHD domains contain a conserved tryptophan, but the position of the PHD2 tryptophan is consistent with those of Ring but not PHD domains. Based on these observations, we propose that Aire PHD2 of all species is better classified as a Ring domain.

## The Aire C-terminal domain is well conserved

The CTD contains a NRB in all species except zebrafish (Fig. 2) and a highly conserved area of charged residues. There is no correlation between this "motif" and domains of known function. The CTD is the third most highly conserved region of Aire (Table 1), and mutation or deletion of this domain results in APS1 in humans. We theorize that the CTD provides an essential function in the Aire protein, yet this function remains unknown.

## The *aire* genomic organization is not evolutionarily conserved

*aire* from mouse, chicken, opossum, and zebrafish were compared to the human orthologue (Fig. 5). The first four exons of *aire* are relatively conserved in both their amino acid translation and exon size among evolutionarily distant animals. The PHD1 is encoded by exon 8 in human or the equivalent for each species, and human exon 13 and 14 (or similar 3′ exons) encoding the CTD are also highly conserved. Other exons are deleted, joined, or of various sizes when comparing the vertebrate species: *SAND*, *PRR*, *PHD2*, and two of the four *NRB*. These data show that the genomic arrangement of *aire* is relatively divergent between humans and other vertebrates. Furthermore, the more conserved regions of *aire* correlate with greater exon conservation (Table 1).

## Comparative analysis and human disease

Point mutations in *aire* alleles of APS1 patients (Heino et al. 2001; Podkrajsek et al. 2005) correlate well with evolutionarily conserved residues. Of the 17 point mutations that result in amino acid substitutions, 14 occur in residues that are 100% conserved in the species analyzed (Fig. 2). Conversely, the insertions and deletions that result in human APS1 occur predominantly within the exons that are the least evolutionarily conserved. Eight of 20 insertions and deletions within the coding sequence of Aire occur in exon 10 alone; this exon encoding the PRR is the least conserved domain within *aire* and is flanked by an intronic microsatellite DNA sequence in Aves (Supplemental Fig. 3). Numerous microsatellite DNA sequences flank human *aire* on chromosome 21q22.3 (Chen et al. 1998). Because point mutations occur somewhat randomly and insertions and deletions occur predominantly at areas of genomic instability, we theorize that certain areas of the *aire* coding region, exon 10 for instance, span regions of greater genomic instability.

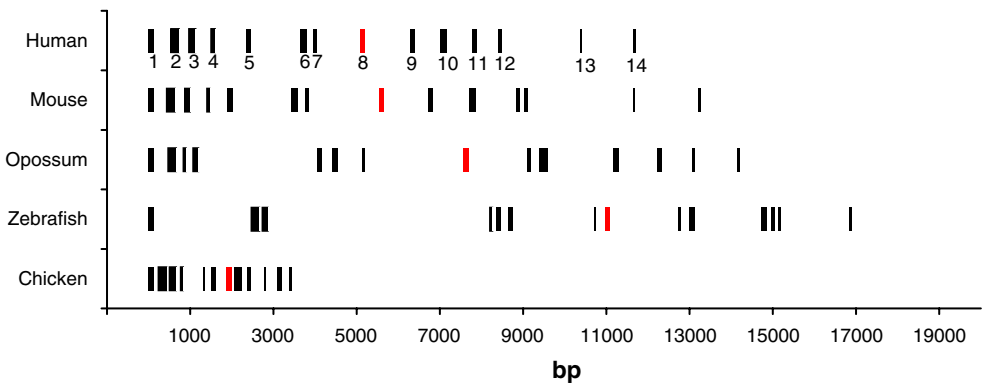



**Fig. 5** Genomic arrangement of the *aire* coding region in human (14 exons), mouse (14), opossum (14), chicken (12), and zebrafish (14). The frog is omitted because the genome seems to be misassembled in several introns. The figure is to scale, and nucleotide positions are shown on the *x*-axis. Exon 8, encoding a portion of PHD1, is in *red* for reference

## Conclusions

Comparative analysis of Aire among different vertebrate classes revealed that the most conserved domains are the NLS, NRB, HSR, PHD1, and CTD. With the exception of the PHD1, the most highly conserved regions of Aire are involved in nuclear localization, transport, and dimerization. Together, this suggests that Aire interacts with a major cofactor or cofactors through the PHD1, which is conserved through evolution, and potentially links to other coactivators that are not conserved. Conversely, the SAND domain or other DNA interaction domains have evolved rapidly and perhaps act in a more indiscriminate manner with binding partners. We show that the PHD2 domain shows greater sequence similarity with Ring than to other PHD domains. Furthermore, our findings show that, not unexpectedly, evolutionarily conserved residues correlate with mutations in human *aire* that result in autoimmunity.

**Acknowledgments** This work has been funded by the NIH (R01AI27877).

## References

Ahonen P, Myllarniemi S, Sipila I, Perheentupa J (1990) Clinical variation of autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy (APECED) in a series of 68 patients. N Engl J Med 322:1829–1836

Anderson MS, Venanzi ES, Chen Z, Berzins SP, Benoist C, Mathis D (2005) The cellular mechanism of Aire control of T cell tolerance. Immunity 23:227–239

Bartl S, Baish MA, Flajnik MF, Ohta Y (1997) Identification of class I genes in cartilaginous fish, the most ancient group of vertebrates displaying an adaptive immune response. J Immunol 159:6097–6104

Bottomley MJ, Collard MW, Huggenvik JI, Liu Z, Gibson TJ, Sattler M (2001) The SAND domain structure defines a novel DNA-binding fold in transcriptional regulation. Nat Struct Biol 8:626–633

Bottomley MJ, Stier G, Pennacchini D, Legube G, Simon B, Akhtar A, Sattler M, Musco G (2005) NMR structure of the first PHD finger of autoimmune regulator protein (AIRE1). Insights into autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy (APECED) disease. J Biol Chem 280:11505–11512

Chang C, Norris JD, Gron H, Paige LA, Hamilton PT, Kenan DJ, Fowlkes D, McDonnell DP (1999) Dissection of the LXXLL nuclear receptor–coactivator interaction motif using combinatorial peptide libraries: discovery of peptide antagonists of estrogen receptors alpha and beta. Mol Cell Biol 19:8226–8239

Chen QY, Lan MS, She JX, Maclaren NK (1998) The gene responsible for autoimmune polyglandular syndrome type 1 maps to chromosome 21q22.3 in US patients. J Autoimmun 11:177–183

Felsenstein J (1989) PHYLIP—phylogeny inference package (version 3.2). Cladistics 5:164–166

Grishin NV (2001) Treble clef finger—a functionally diverse zinc-binding structural motif. Nucleic Acids Res 29:1703–1714

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98

Heino M, Peterson P, Kudoh J, Nagamine K, Lagerstedt A, Ovod V, Ranki A, Rantala I, Nieminen M, Tuukkanen J, Scott HS, Antonarakis SE, Shimizu N, Krohn K (1999) Autoimmune regulator is expressed in the cells regulating immune tolerance in thymus medulla. Biochem Biophys Res Commun 257:821–825

Heino M, Peterson P, Kudoh J, Shimizu N, Antonarakis SE, Scott HS, Krohn K (2001) APECED mutations in the autoimmune regulator (AIRE) gene. Hum Mutat 18:205–211

Ilmarinen T, Melen K, Kangas H, Julkunen I, Ulmanen I, Eskelin P (2006) The monopartite nuclear localization signal of autoimmune regulator mediates its nuclear import and interaction with multiple importin alpha molecules. FEBS J 273:315–324

Johnnidis JB, Venanzi ES, Taxman DJ, Ting JP, Benoist CO, Mathis DJ (2005) Chromosomal clustering of genes controlled by the aire transcription factor. Proc Natl Acad Sci USA 102:7233–7238

Klamp T, Sahin U, Kyewski B, Schwendemann J, Dhaene K, Tureci O (2006) Expression profiling of autoimmune regulator AIRE mRNA in a comprehensive set of human normal and neoplastic tissues. Immunol Lett 106:172–179

Kumar PG, Laloraya M, Wang CY, Ruan QG, voodi-Semiromi A, Kao KJ, She JX (2001) The autoimmune regulator (AIRE) is a DNA-binding protein. J Biol Chem 276:41357–41364

Kyewski B, Klein L (2006) A central role for central tolerance. Annu Rev Immunol 24:571–606

Mathis D, Benoist C (2007) A decade of AIRE. Nat Rev Immunol 7:645–650

Meloni A, Fiorillo E, Corda D, Perniola R, Cao A, Rosatelli MC (2005) Two novel mutations of the AIRE protein affecting its homodimerization properties. Hum Mutat 25:319

Mertz LM, Rashtchian A (1994) Nucleotide imbalance and polymerase chain reaction: effects on DNA amplification and synthesis of high specific activity radiolabeled DNA probes. Anal Biochem 221:160–165

Nair R, Carter P, Rost B (2003) NLSdb: database of nuclear localization signals. Nucleic Acids Res 31:397–399

Pitkanen J, Vahamurto P, Krohn K, Peterson P (2001) Subcellular localization of the autoimmune regulator protein. Characterization of nuclear targeting and transcriptional activation domain. J Biol Chem 276:19597–19602

Podkrajsek KT, Bratanic N, Krzisnik C, Battelino T (2005) Autoimmune regulator-1 messenger ribonucleic acid analysis in a novel intronic mutation and two additional novel AIRE gene mutations in a cohort of autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy patients. J Clin Endocrinol Metab 90:4930–4935

Schmutz J, Grimwood J (2004) Genomes: fowl sequence. Nature 432:679–680

Shiina T, Briles WE, Goto RM, Hosomichi K, Yanagiya K, Shimizu S, Inoko H, Miller MM (2007) Extended gene map reveals tripartite motif, C-type lectin, and Ig superfamily type genes within a subregion of the chicken MHC-B affecting infectious disease. J Immunol 178:7162–7172

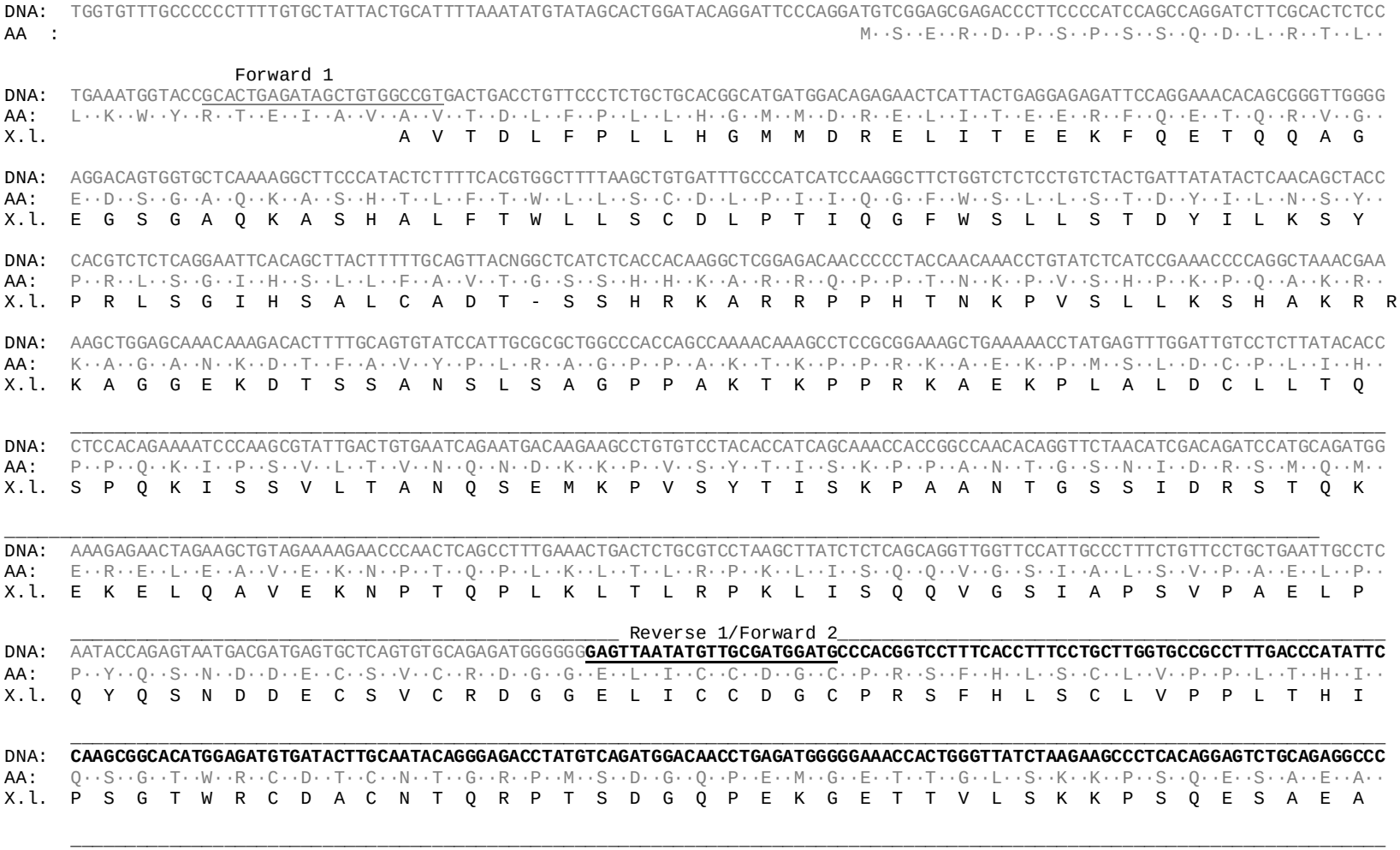
The Finnish-German APECED Consortium (1997) An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. Nat Genet 17:399–403

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Uchida D, Hatakeyama S, Matsushima A, Han H, Ishido S, Hotta H, Kudoh J, Shimizu N, Doucas V, Nakayama KI, Kuroda N, Matsumoto M (2004) AIRE functions as an E3 ubiquitin ligase. J Exp Med 199:167–172

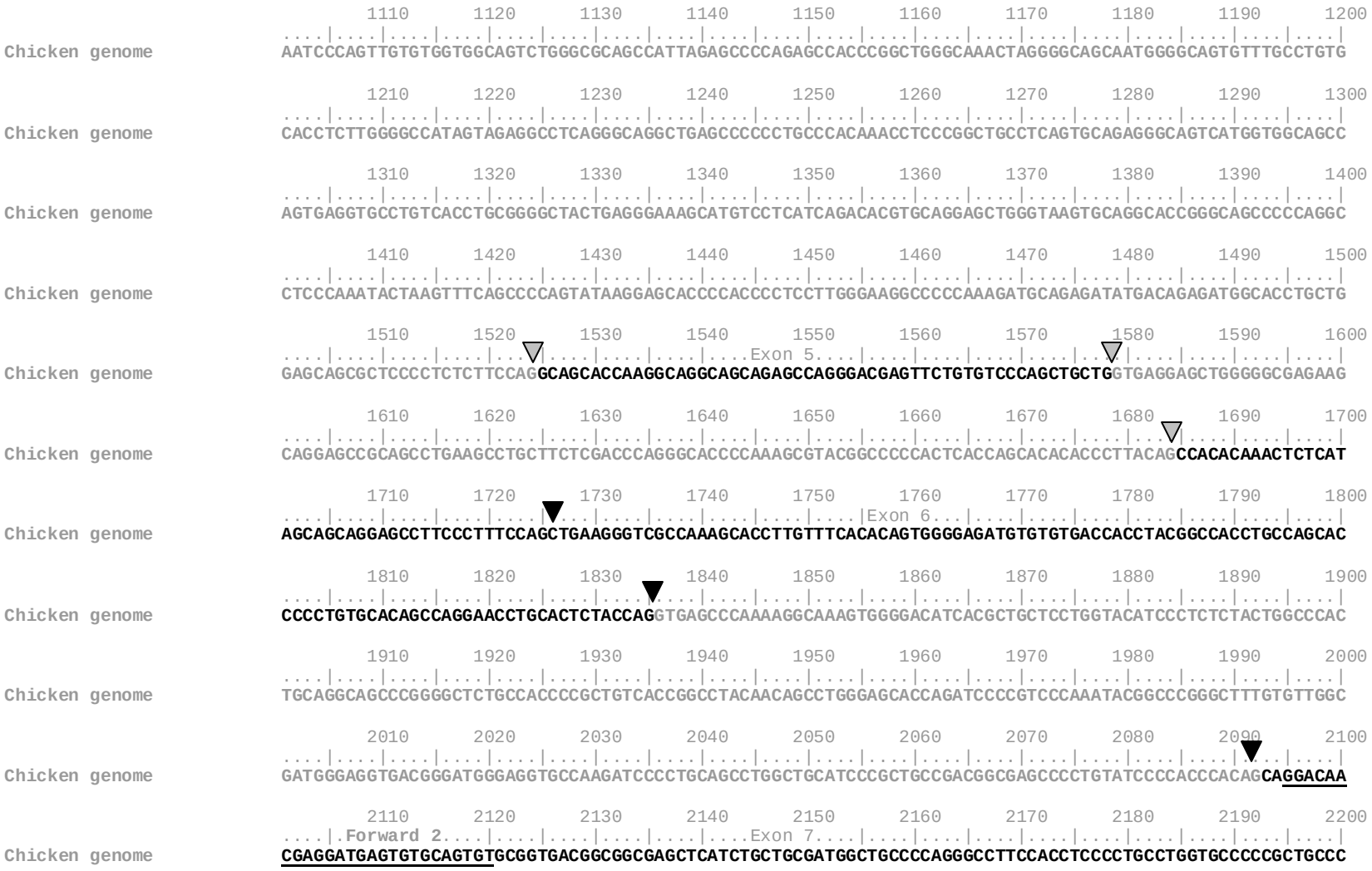
Williamson MP (1994) The structure and function of proline-rich regions in proteins. Biochem J 297(Pt 2):249–260

## Supplemental Figure 1

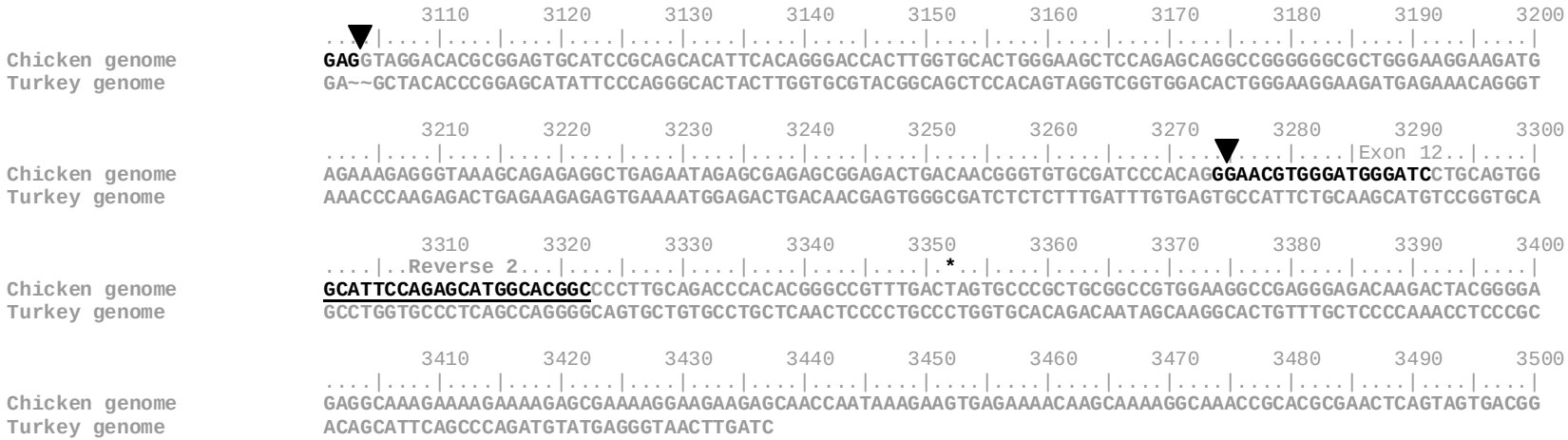



AA: Q··S··Q··K··R··M··K··V··C··E··M··L··A··D··K··H··D··S··I··I··Q··K··S··S··S··I··Q··N··Q··I··Y··P··Q··I··V··A··Q··P··E··I·
X.l. Q S Q K R M K V C E M P A D K H D S I I Q K S N T I R T L N N P Q L V A Q P V I
DNA: GCCCAAAACCTACAGCCAGAACACAGACCTGCTCACAACAACAAAGCTCTTACCAAGCTCCCCCTCAGCCTCAGGCCTGCCCCCAGTCACTTTCAAGCCTCTGCATGTCTCAGACACCCT
AA: C··P··K··P··T··A··R··T··Q··T··C··S··Q··Q··Q··S··S··Y··Q··A··P··P··Q··P··Q··A··C··P··Q··S··L··S··S··L··C··M··S··Q··T··P··
X.l. C P K P T G R T Q N C S Q Q Q S S F Q M L P Q P Q A C P R S L S N L C I S Q A P
DNA: AGGCAGTTTTGCTCGCAACTTTTGCCCAGGTCTCATATTAGAACCTCACATCAGCTGCAGACCAACTCTCAACATTTGTCTATTCCACAACCTTGTTCCACATCTGGTGGTCAGATAC
AA : L··R··Q··F··C··S··Q··L··L··P··R··S··H··I··R··T··S··H··Q··L··Q··T··N··S··Q··H··L··S··I··P··Q··P··C··S··T··S··G··G··Q··I·
X.l. V R Q I C S Q F L P R S Q I R T S Q Q P Q A N P Q H L S T P Q P C S T S L P P I
DNA: CTTGTACTACTCTCACCCTTCAGTCCCAGAACTGCACTACTCCACCAGACAATACACGCTCCTGCCAAACACCTAGCGTCCAAACTCAGCATAGCACACCCTGTGCTTTACAGAAAGAGC
AA : P··C··T··T··L··T··L··Q··S··Q··N··C··T··T··P··P··D··N··T··R··S··C··Q··T··P··S··V··Q··T··Q··H··S··T··P··C··A··L··Q··K··E··
X.l. P C T T L T F P S Q N Y P T P A D N T C T S Q T P S V Q T Q H S K P C A L Q K E
DNA: CTTGTCCTCCAGCATCACCACAACCTAAACTCTGCATTCCACATGCGCCCTTAATTGCCATTAACTCTCCCACTGAACCCCAGTCTGAGATCAGTGCAAGGGAACCTTCCCAGGGCGAAG
AA : P··C··P··P··A··S··P··Q··P··K··L··C··I··P··H··A··P··L··I··A··I··N··S··P··T··E··P··Q··S··E··I··S··A··R··E··P··S··Q··G··E·
X.l. P C P L A S L E P T L C I P L V P L N G I N S A T E P Q F E V S P R P P V Q G E
DNA: TGCCCATCCTGCCAATACCCAGGGCAACAGAATCCAACTCTGGTACCTCCTCAGCTCAAGAAGAGATGGTACCTGCACCAGTACCTCTTCAATCTGAAATAGGAGAGCTCAAGGTTCCAG
AA : V··P··I··L··P··I··P··R··A··T··E··S··N··S··G··T··S··S··A··Q··E··E··M··V··P··A··P··V··P··L··Q··S··E··I··G··E··L··K··V··P·
X.l. V P I L P I P G T T E S N S G T S S A P E E M V P A P V A L Q A E I G E L K V P
Reverse 2
DNA: CAGAAGTCGCAGGAAGCAACTTGACATTGAGTCGACATGAACTCGAATGCCTTATTGCTGAGAGCTCCTTTGATTGCTTCCTGCAATGGGCATTTCAGAACATTTCCCGTCCTGTGCAGTGAC
AA : A··E··V··A··G··S··N··L··T··L··S··R··H··E··L··E··C··L··I··A··E··S··S··F··D··C··F··L··Q··W··A··F··Q··N··I··S··R··P··V··Q··*·
X.l. A D A A G S N L T L S R H E L E C L I A E S S F D C F L Q W A F Q N I S
DNA3’ CATAATCCTGTATGAGGATAGCATATTCCAGGGCACCTTGACTGTGATATTCCTAAAACACCAATGGAGGGAGTAACTGCTGAGGAATATAGAGGAATTCCATATAGTAATTTTAAAAACGTG
TTTTATCAGAATTTAACCCATAAGCATATAATATAAACATAAACCACATTTGTGATTGCTTAATAAAGTTATTTCCTGCTCATCAAAAAAAAAAAAAAAAA

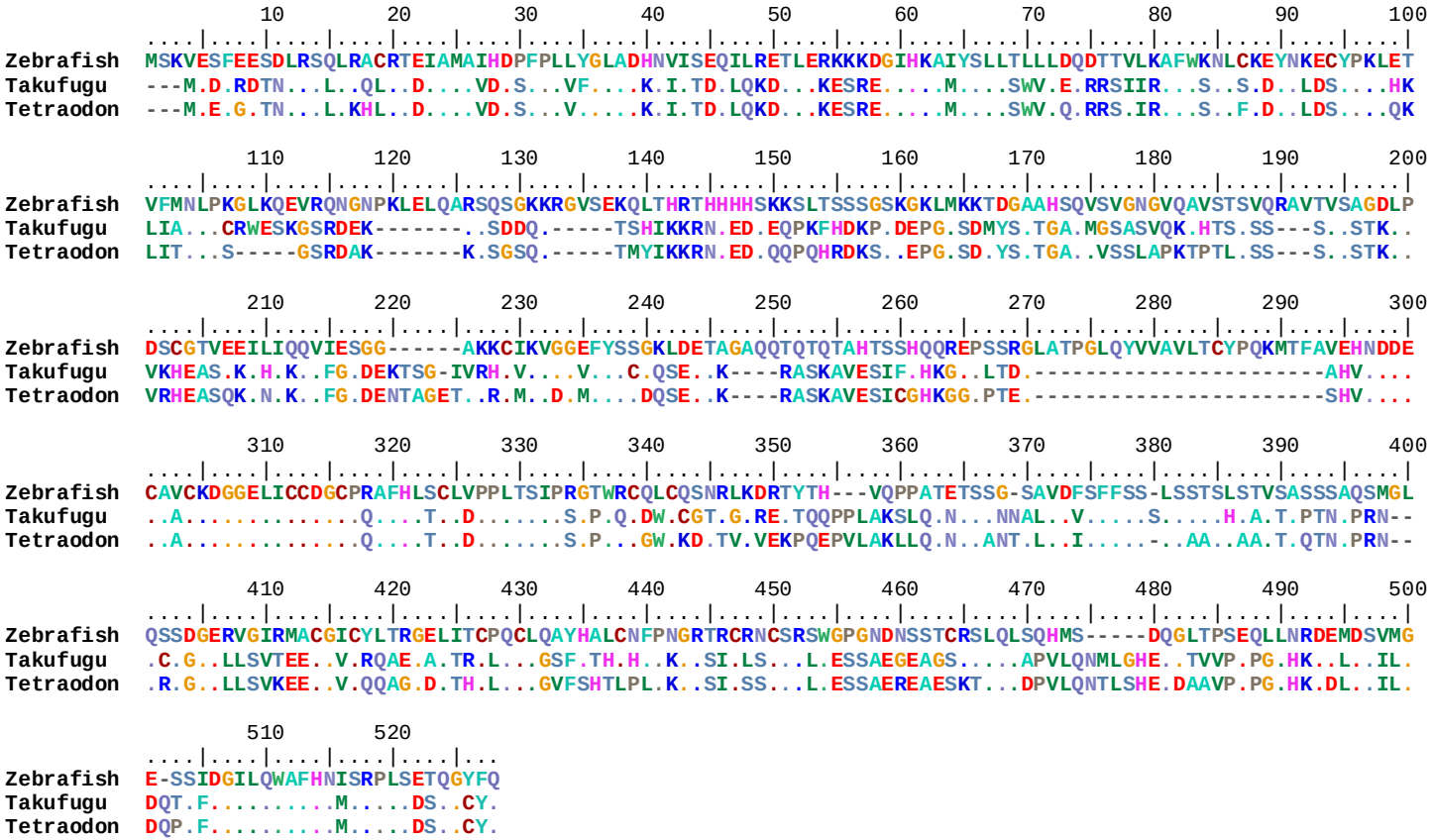
## Supplemental Figure 2


1110 1120 1130 1140 1150 1160 1170 1180 1190 1200
Chicken genome AATCCCAGTTGTGTGGTGGCAGTCTGGGCGCAGCCATTAGAGCCCCAGAGCCACCCGGCTGGGCAAACTAGGGGCAGCAATGGGGCAGTGTTTGCCTGTG
1210 1220 1230 1240 1250 1260 1270 1280 1290 1300
Chicken genome CACCTCTTGGGGCCATAGTAGAGGCCTCAGGGCAGGCTGAGCCCCCCTGCCCACAAACCTCCCGGCTGCCTCAGTGCAGAGGGCAGTCATGGTGGCAGCC
1310 1320 1330 1340 1350 1360 1370 1380 1390 1400
Chicken genome AGTGAGGTGCCTGTCACCTGCGGGGCTACTGAGGGAAAGCATGTCCTCATCAGACACGTGCAGGAGCTGGGTAAGTGCAGGCACCGGGCAGCCCCCAGGC
1410 1420 1430 1440 1450 1460 1470 1480 1490 1500
Chicken genome CTCCCAAATACTAAGTTTCAGCCCCAGTATAAGGAGCACCCCACCCCTCCTTGGGAAGGCCCCCAAAGATGCAGAGATATGACAGAGATGGCACCTGCTG
1510 1520 1530 1540 1550 1560 1570 1580 1590 1600
Exon 5
Chicken genome GAGCAGCGCTCCCCTCTCTTCCAGGCAGCACCAAGGCAGGCAGCAGAGCCAGGGACGAGTTCTGTGTCCCAGCTGCTGGTGAGGAGCTGGGGGCGAGAAG
1610 1620 1630 1640 1650 1660 1670 1680 1690 1700
Chicken genome CAGGAGCCGCAGCCTGAAGCCTGCTTCTCGACCCAGGGCACCCCAAAGCGTACGGCCCCCACTCACCAGCACACACCCTTACAGCCACACAAACTCTCAT
1710 1720 1730 1740 1750 1760 1770 1780 1790 1800
Exon 6
Chicken genome AGCAGCAGGAGCCTTCCCTTTCCAGCTGAAGGGTCGCCAAAGCACCTTGTTTCACACAGTGGGGAGATGTGTGTGACCACCTACGGCCACCTGCCAGCAC
1810 1820 1830 1840 1850 1860 1870 1880 1890 1900
Chicken genome CCCCTGTGCACAGCCAGGAACCTGCACTCTACCAGGTGAGCCCAAAAGGCAAAGTGGGGACATCACGCTGCTCCTGGTACATCCCTCTCTACTGGCCCAC
1910 1920 1930 1940 1950 1960 1970 1980 1990 2000
Chicken genome TGCAGGCAGCCCGGGGCTCTGCCACCCCGCTGTCACCGGCCTACAACAGCCTGGGAGCACCAGATCCCCGTCCCAAATACGGCCCGGGCTTTGTGTTGGC
2010 2020 2030 2040 2050 2060 2070 2080 2090 2100
Chicken genome GATGGGAGGTGACGGGATGGGAGGTGCCAAGATCCCCTGCAGCCTGGCTGCATCCCGCTGCCGACGGCGAGCCCCTGTATCCCCACCCACAGCAGGACAA
2110 2120 2130 2140 2150 2160 2170 2180 2190 2200
Forward 2
Exon 7
Chicken genome CGAGGATGAGTGTGCAGTGTGCGGTGACGGCGGCGAGCTCATCTGCTGCGATGGCTGCCCCAGGGCCTTCCACCTCCCCTGCCTGGTGCCCCCGCTGCCC

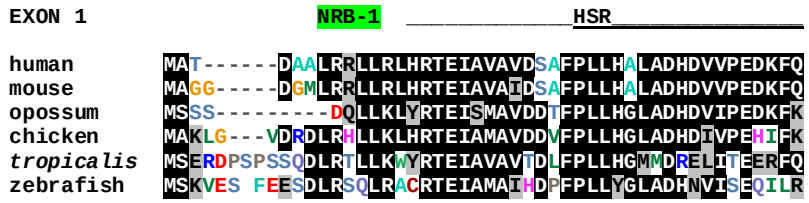

2210 2220 2230 2240 2250 2260 2270 2280 2290 2300
Chicken genome CGTGTCCCGAGGTGAGCCCCGCGCTGCCGTGGGGCCCAGCTGTACCTCATGTCCCCACTGCCCCTCACTGCTGTGCTGGGGTTGTGCAGCGGGACGTGGC
2310 2320 2330 2340 2350 2360 2370 2380 2390 2400
Exon 8
Chicken genome AATGCAGCTCATGCGTGGCCAAGCTGGGCCGGCTGCGGGAGGCAGACACGGCTGCAGAGCAGCTCCCTGCAGTCCCAGACAAGGAGGAGCACGGTGCCCA
2410 2420 2430 2440 2450 2460 2470 2480 2490 2500
Chicken genome GCCGGGAGGAGGCCACGGCAGCACCTGTGGCCGCTGCTTCTCCAGCATCTCTGCACCCCAACGCTGCCCTACTCGTGATGGGGACCCTGGGTGAGTCCCC
Turkey genome ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~GATCTCCGCACCCCAACGCTGCCCCACGCGTGATGGAGACCCCGGGTGAGTCCCC
2510 2520 2530 2540 2550 2560 2570 2580 2590 2600
Chicken genome GCCGCCTTTCCCCACATCACACCCCACACCTCCCCGTTCCTTCCAACTGCTCAGAAGGCTTCAAGACACCTCGTTGGCTTGCAGAGGGCTGTGGCTTTGC
Turkey genome TCCGCCTGCCCCCACGTCACAGCCCAAGCCTCTCCATTCCTTCCTGCCGCTCGGAGCGCTCGCAAATACCTCGTTGGCTTGCAGAGGGCTGTGGCTTTGC
2610 2620 2630 2640 2650 2660 2670 2680 2690 2700
Exon 9
Chicken genome AGCTCCTGCACGGGCATCCCAGAAACAGGCAGCCACGAGAGCAGTGCAGCTGCTGGAGAACGCGTGCTCCTGGCAGCAAAGGTTGGTGCACCCAGGGCCG
Turkey genome GGCTCCTGCGTGAGCACCCCAGAAGCAGGCAGCTGGGAGAGCACCGCAGCTGCTGCAAACCACGTGCTTCAGGCAGCAAAGGTTGGTACAGCCAGGGCCG
2710 2720 2730 2740 2750 2760 2770 2780 2790 2800
Chicken genome GTACCGGTGCGGGGAACGCAGGCGGTGCTGGGTAAAGGCGCCTGGAGGAACACCAAAAGAGAGAGAGAGGGGGGGGGAAGGGAGAGGGAGAGGAGGCCGA
Turkey genome GAACCGGTGCGCAGAATGCAGGCGGTGCTGGGTGAAGGCGCTTGGAGGAACAGCCAGAGAGACAGAGAGACAGAGAGACAGAGAGAGAGAGA
2810 2820 2830 2840 2850 2860 2870 2880 2890 2900
Exon 10
Chicken genome GGGACAAGGAGTGGCCGCAGGGGGAGGGACGCGGCAGCGCGAGATGGAGGCGATGCTGAAGCTGTCGTGCAGGAGGAAGAGCACAGCCCAGGGGCTGCGA
Turkey genome GACAGAGAGAGAGAGAGAGACAGAGACAGAGAGAGAGAGAGAAAGAGAGAGGAGGCCGAGGGACAGGGAATGGTTGCTGGGCCGCGGGCACGGCACCTAC
2910 2920 2930 2940 2950 2960 2970 2980 2990 3000
Chicken genome CCCCGGCCTCCGGTCCGTGCAGGACCGAAAAGG~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~AGCCCACGGNCC
Turkey genome ACATGGAGATGTTGCTGAAGATGTACTGGAGGGGGAAGAGCACAGCCCAGGGGCTGTGACTGCGGCCTGCGGGCCGTGAAGGACCGAAACGAGCGCATGA
3010 3020 3030 3040 3050 3060 3070 3080 3090 3100
exon 11
Chicken genome AGAAGGTAGGGGGTGACGCATCTCCTCCCACAGCTAAANAATGGCTCAACCAGCAGCGTCCCCATGTCCAGCAGGGAAGAGCTCGACGCCC~TCCTGAGT
Turkey genome CCAGAGGTGGGGGTGACACATCTTCCCCCACAGTTGGAGAACGGCTCAGCGAGCACTGACCCCATGTCCTGCAGGGAAGAGCTCAAATGCCCTCCTGAGT


3110 3120 3130 3140 3150 3160 3170 3180 3190 3200
Chicken genome GAGGTAGGACACGCGGAGTGCATCCGCAGCACATTCACAGGGACCACTTGGTGCACTGGGAAGCTCCAGAGCAGGCCGGGGGGCGCTGGGAAGGAAGATG
Turkey genome GA~~GCTACACCCGGAGCATATTCCCAGGGCACTACTTGGTGCGTACGGCAGCTCCACAGTAGGTCGGTGGACACTGGGAAGGAAGATGAGAAACAGGGT
3210 3220 3230 3240 3250 3260 3270 3280 3290 3300
Exon 12
Chicken genome AGAAAGAGGGTAAAGCAGAGAGGCTGAGAATAGAGCGAGAGCGGAGACTGACAACGGGTGTGCGATCCCACAGGGAACGTGGGATGGGATCCTGCAGTGG
Turkey genome AAACCCAAGAGACTGAGAAGAGAGTGAAAATGGAGACTGACAACGAGTGGGCGATCTCTCTTTGATTTGTGAGTGCCATTCTGCAAGCATGTCCGGTGCA
3310 3320 3330 3340 3350 3360 3370 3380 3390 3400
Reverse 2
Chicken genome GCATTCCAGAGCATGGCACGGCCCCTTGCAGACCCACACGGGCCGTTTGACTAGTGCCCGCTGCGGCCGTGGAAGGCCGAGGGAGACAAGACTACGGGGA
Turkey genome GCCTGGTGCCCTCAGCCAGGGGCAGTGCTGTGCCTGCTCAACTCCCCTGCCCTGGTGCACAGACAATAGCAAGGCACTGTTTGCTCCCCAAACCTCCCGC
3410 3420 3430 3440 3450 3460 3470 3480 3490 3500
Chicken genome GAGGCAAAGAAAAGAAAAGAGCGAAAAGGAAGAAGAGCAACCAATAAAGAAGTGAGAAAACAAGCAAAAGGCAAACCGCACGCGAACTCAGTAGTGACGG
Turkey genome ACAGCATTCAGCCCAGATGTATGAGGGTAACTTGATC

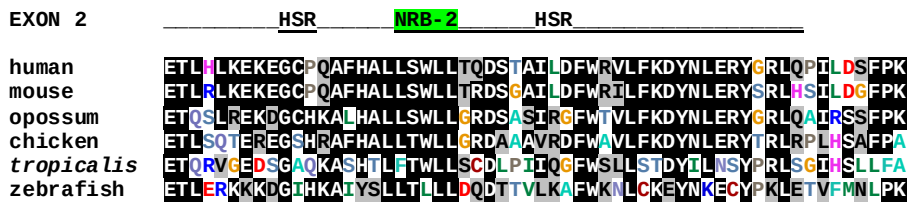
## Supplemental Figure 3



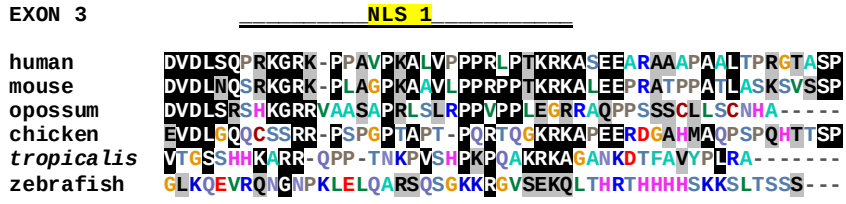
## Supplemental Figure 4




| | human | mouse | opossum | chicken | *tropicalis* | zebrafish |
|---|---|---|---|---|---|---|
| human | ID | 89% | 60% | 66% | 44% | 43% |
| mouse | 89% | ID | 58% | 64% | 42% | 45% |
| opossum | 60% | 58% | ID | 62% | 46% | 45% |
| chicken | 66% | 64% | 62% | ID | 48% | 51% |
| *tropicalis* | 44% | 43% | 46% | 48% | ID | 42% |
| zebrafish | 43% | 45% | 45% | 51% | 42% | ID |



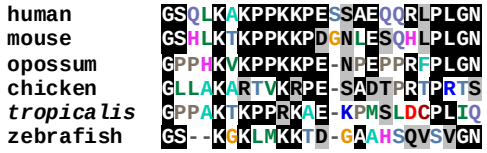

| | human | mouse | opossum | chicken | *tropicalis* | zebrafish |
|---|---|---|---|---|---|---|
| human | ID | 86% | 71% | 62% | 38% | 43% |
| mouse | 86% | ID | 66% | 60% | 40% | 40% |
| opossum | 71% | 66% | ID | 64% | 45% | 41% |
| chicken | 62% | 60% | 64% | ID | 41% | 38% |
| *tropicalis* | 38% | 40% | 45% | 41% | ID | 31% |
| zebrafish | 43% | 40% | 41% | 38% | 31% | ID |




| | human | mouse | opossum | chicken | *tropicalis* | zebrafish |
|---|---|---|---|---|---|---|
| human | ID | 65% | 23% | 42% | 25% | 4% |
| mouse | 65% | ID | 26% | 38% | 21% | 6% |
| opossum | 23% | 26% | ID | 13% | 17% | 6% |
| chicken | 42% | 38% | 13% | ID | 21% | 4% |
| *tropicalis* | 25% | 21% | 17% | 21% | ID | 0% |
| zebrafish | 4% | 6% | 6% | 4% | 0% | ID |

opossum exon 3 plus 5' end of exon 4

**EXON 4** **NLS 2**




| | human | mouse | opossum | chicken | *tropicalis* | zebrafish |
|---|---|---|---|---|---|---|
| human | ID | 68% | 60% | 40% | 36% | 36% |
| mouse | 68% | ID | 56% | 20% | 40% | 36% |
| opossum | 60% | 56% | ID | 33% | 50% | 29% |
| chicken | 40% | 20% | 33% | ID | 21% | 17% |
| *tropicalis* | 36% | 40% | 50% | 21% | ID | 17% |
| zebrafish | 36% | 36% | 29% | 17% | 17% | ID |

opossum 3' end of exon 4

**EXON 5** SAND


| | human | mouse | opossum | chicken | *tropicalis* | zebrafish |
|---|---|---|---|---|---|---|
| human | ID | 87% | 68% | 8% | 9% | 63% |
| mouse | 87% | ID | 71% | 8% | 9% | 63% |
| opossum | 68% | 71% | ID | 8% | 9% | 53% |
| chicken | 8% | 8% | 8% | ID | 6% | 8% |
| *tropicalis* | 9% | 9% | 9% | 6% | ID | 14% |
| zebrafish | 63% | 63% | 53% | 8% | 14% | ID |

**EXON 6** SAND


| | human | mouse | opossum | chicken | *tropicalis* | zebrafish |
|---|---|---|---|---|---|---|
| human | ID | 76% | 66% | 0% | 14% | 30% |
| mouse | 76% | ID | 64% | 0% | 16% | 28% |
| opossum | 66% | 64% | ID | 0% | 18% | 34% |
| chicken | 0% | 0% | 0% | ID | 0% | 0% |
| *tropicalis* | 14% | 16% | 18% | 0% | ID | 8% |
| zebrafish | 30% | 28% | 34% | 0% | 8% | ID |

```
EXON 7          _________SAND_________

human           --G--------GGEARLGQQGSVPAPLALPSDPQLHQ
mouse           --G--------RDEQKVGQQCGVPPLPSLPSEPQVNQ
opossum         --GS-------RAEMQLSQHGSVPATPAQPPELHLHQ
chicken         AEGSPKHLVSHSGEMCVTTYGHLPAPPVHSQEPALYQ
tropicalis      ----------VGSIALS----VPA--ELPQ----YQ
zebrafish       -----------GLATPGLQYVVAVLTCYPQKMTFA-
```

| | human | mouse | opossum | chicken | *tropicalis* | zebrafish |
|---|---|---|---|---|---|---|
| human | ID | 48% | 50% | 27% | 30% | 22% |
| mouse | 48% | ID | 36% | 22% | 19% | 19% |
| opossum | 50% | 36% | ID | 30% | 25% | 7% |
| chicken | 27% | 22% | 30% | ID | 16% | 5% |
| *tropicalis* | 30% | 19% | 25% | 16% | ID | 15% |
| zebrafish | 22% | 19% | 7% | 5% | 15% | ID |

chicken exon 6

```
EXON 8          _________________PHD-1_________________

human           --KNEDECAVCRDGGELICCDGCPRAFHLACLSPPLREIP-
mouse           --KNEDECAVCHDGGELICCDGCPRAFHLACLSPPLQEIP-
opossum         --RNDDECAVCRDGGELICCDGCPRAFHLACLEPPLTDIPS
chicken         -QDNEDECAVCGDGGELICCDGCPRAFHLPCLVPPLPRVPS
tropicalis      --SNDDECSVCRDGGELICCDGCPRSFHLSCLVPPLTHIP-
zebrafish       VEHNDDECAVCKDGGELICCDGCPRAFHLSCLVPPLTSIPS
```

| | human | mouse | opossum | chicken | *tropicalis* | zebrafish |
|---|---|---|---|---|---|---|
| human | ID | 95% | 85% | 78% | 79% | 76% |
| mouse | 95% | ID | 82% | 78% | 76% | 76% |
| opossum | 85% | 82% | ID | 78% | 82% | 83% |
| chicken | 78% | 78% | 78% | ID | 73% | 78% |
| *tropicalis* | 79% | 76% | 82% | 73% | ID | 80% |
| zebrafish | 76% | 76% | 83% | 78% | 80% | ID |

chicken exon 7

```
EXON 9          _____________PHD-1_____________><_____________PRR_____________

human           SGTWRCSSCL--QATVQEVQPR----AEEPRPQEPPVETP--------------------------
mouse           SGLWRCSCCL--QGRVQQNLSQ----PEVSRPPELPAETP--------------------------
opossum         GMWRCGCCI--VGKVHQDGHH----GEERDPHS---ETA--------------------------
chicken         GTWQCSSCVAKLGRLREADTAAEQLPAVPDKEEHGAQPGGGHGSTCGRCFSSISAPQRCPTRDGDPG
tropicalis      GTWRCDTCN--TGRPMSDGQ-----PEMGETTGLSKK----------------------------
zebrafish       GTWRCQLCQ--SNR-LKDR----------TYTHVQ------------------------------
```

| | human | mouse | opossum | chicken | *tropicalis* | zebrafish |
|---|---|---|---|---|---|---|
| human | ID | 56% | 32% | 15% | 21% | 18% |
| mouse | 56% | ID | 38% | 16% | 29% | 18% |
| opossum | 32% | 38% | ID | 9% | 27% | 19% |
| chicken | 15% | 16% | 9% | ID | 12% | 9% |
| *tropicalis* | 21% | 29% | 27% | 12% | ID | 30% |
| zebrafish | 18% | 18% | 19% | 9% | 30% | ID |

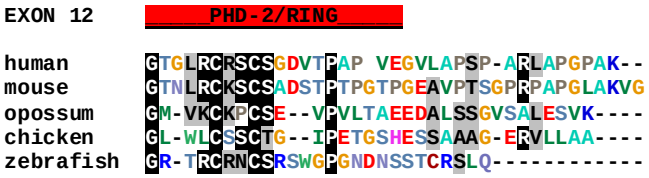
chicken exon 8


| | human | mouse | opossum | *tropicalis* | zebrafish |
|---|---|---|---|---|---|
| human | ID | 48% | 34% | 8% | 13% |
| mouse | 48% | ID | 21% | 6% | 13% |
| opossum | 34% | 21% | ID | 10% | 15% |
| *tropicalis* | 8% | 6% | 10% | ID | 4% |
| zebrafish | 13% | 13% | 15% | 4% | ID |

no chicken analog, *tropicalis* exon 10 plus 5' end of 11 (rest of 11 has no non *tropicalis* analog)

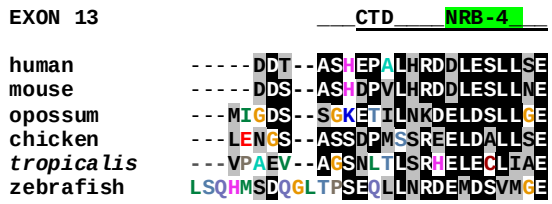

| | human | mouse | opossum | zebrafish |
|---|---|---|---|---|
| human | ID | 75% | 45% | 22% |
| mouse | 75% | ID | 33% | 22% |
| opossum | 45% | 33% | ID | 23% |
| zebrafish | 22% | 22% | 23% | ID |

no chicken or *tropicalis* analogous exon



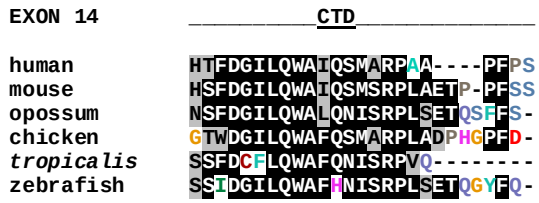

| | human | mouse | opossum | chicken | zebrafish |
|---|---|---|---|---|---|
| human | ID | 51% | 22% | 22% | 19% |
| mouse | 51% | ID | 18% | 18% | 15% |
| opossum | 22% | 18% | ID | 19% | 18% |
| chicken | 22% | 18% | 19% | ID | 12% |
| zebrafish | 19% | 15% | 18% | 12% | ID |

chicken exon 9, no *tropicalis* analog




| | human | mouse | opossum | chicken | *tropicalis* | zebrafish |
|---|---|---|---|---|---|---|
| human | ID | 81% | 39% | 39% | 30% | 25% |
| mouse | 81% | ID | 39% | 43% | 30% | 21% |
| opossum | 39% | 39% | ID | 30% | 26% | 36% |
| chicken | 39% | 43% | 30% | ID | 35% | 18% |
| *tropicalis* | 30% | 30% | 26% | 35% | ID | 18% |
| zebrafish | 25% | 21% | 36% | 18% | 18% | ID |

chicken exon 10, *tropicalis* exon 12




| | human | mouse | opossum | chicken | *tropicalis* | zebrafish |
|---|---|---|---|---|---|---|
| human | ID | 73% | 44% | 63% | 39% | 37% |
| mouse | 73% | ID | 67% | 59% | 42% | 56% |
| opossum | 44% | 67% | ID | 46% | 50% | 73% |
| chicken | 63% | 59% | 46% | ID | 35% | 50% |
| *tropicalis* | 39% | 42% | 50% | 35% | ID | 50% |
| zebrafish | 37% | 56% | 73% | 50% | 50% | ID |