Nomenclature of the ARID family of DNA-binding proteins

Deborah Wilskera,1, Loren Probstb,1, Hester M. Wainc, Lois Maltaisd, Philip W. Tuckerb, Elizabeth Moranu,*

aFels Institute for Cancer Research and Molecular Biology, Temple University School of Medicine, Philadelphia, PA 19104, USA
bInstitution for Cellular and Molecular Biology, Department of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, TX 78712-1095, USA
cHUGO Gene Nomenclature Committee (HGNC), Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK
dMouse Genomic Nomenclature Committee (MGNC), Mouse Genome Informatics (MGI), The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

Received 15 March 2005; accepted 31 March 2005
Available online 26 May 2005

Abstract

The ARID (AT-rich interaction domain) is a billion year old DNA-binding domain that is conserved throughout the evolution of higher eukaryotes. The ARID consensus sequence spans about 100 amino acid residues, and structural studies identify the major groove contact site as a modified helix-turn-helix motif. ARID-containing proteins exhibit a range of cellular functions, including participation in chromatin remodeling, and regulation of gene expression during cell growth, differentiation, and development. A subset of ARID family proteins binds DNA specifically at AT-rich sites; the remainder bind DNA nonspecifically. Orthologs to each of the seven distinct subfamilies of mammalian ARID-containing proteins are found in insect genomes, indicating the minimum age for the organization of these higher metazoan subfamilies. Many of these ancestral genes were duplicated and fixed over time to yield the 15 ARID-containing genes that are found in the human, mouse, and dog genomes. This paper describes a nomenclature, recommended by the Mouse Genomic Nomenclature Committee (MGNC) and accepted by the Human Genome Organization (HUGO) Gene Nomenclature Committee, for these mammalian ARID-containing genes that reflects this evolutionary history.

Keywords: ARID family; Nomenclature; DNA-binding proteins; Sequence-specific binding; p270; BAF250; OSA1; SWI1; SMARCF1; Bright; Bdp; Dri; RETN; DRIL1; DRIL2; RBP1; RBBP1L1; SAP180; BCAA; MRF-1; MRF2; Desrt; RBP2; PLU-1; SMCX; SMCY; XE169; jumonji

Introduction

The ARID (AT-rich interaction domain) is a billion year old DNA-binding domain that has been identified in all sequenced higher eukaryotic genomes. The ARID consensus sequence spans about 100 amino acid residues, and structural studies identify the major groove contact site as a modified helix-turn-helix motif [1–4]. The ARID consensus was first identified in the mouse B-cell-specific transcription factor Bright [5] and in the product of the dead ringer (dri, also known as retained, retn) gene of Drosophila melanogaster [6]. DRI and Bright were each isolated in searches designed specifically to identify proteins binding to AT-rich sequences, but neither turned out to contain a previously known DNA-binding domain. Identification of DNA-binding sequences conserved between Bright and DRI defined the parameters of a new DNA-binding domain, whose name was inspired by the interaction of these proteins with AT-rich DNA elements.

Since the discovery of the ARID, many additional proteins containing this domain have been identified. Interestingly though, not all ARID-containing proteins bind...
to DNA in a sequence-specific manner, as discussed further below. The cellular functions of ARID proteins include participation in the regulation of cell growth, differentiation, and development [7,8]. The ARID domain is both ancient and widespread, occurring in (some) protozoa, green algae, higher plants, fungi, and metazoans. No archae- or eubacterial ARID domains have been identified to date.

Sequence relationships reveal seven distinct subfamilies of ARID-containing proteins in metazoans. In mammals these have been given the names ARID1 through ARID5, and JARID1 and JARID2, as shown in Fig. 1 and Tables 1 and 2. Six of these seven subfamilies have been identified in Drosophila melanogaster, which, however, lacks an ARID5-like gene. A putative ARID5 ortholog has been identified in the Apis (honeybee) genome and the predicted protein sequence generates BLAST reciprocal best hits with human and mouse ARID5B proteins. The ARID domain of this protein has a genomic structure that is similar to the ARID5 family of mammalian proteins, possessing introns at the first and third intron positions in ARID5B (see Fig. 3), although there is no intron at the second position. This first shared intron is unique to the ARID5 family. Furthermore, sequences upstream of the ARID domain in ARID5B are also highly conserved in this predicted protein. This upstream conserved region is also present in an Anopheles (mosquito) protein which does not have an identifiable ARID domain in the available genomic sequence within 50 kb of the locus. These findings suggest that an ARID5 ortholog was present in the ancient ancestor to protostomes and deuterostomes, but that this protein was not essential and has since been lost in part or altogether in several descendant lineages.

ARID-containing proteins have also been identified in higher plants and fungi. In the sequenced Arabidopsis genome, eleven ARID-containing proteins that form five subfamilies have been identified. While a JARID1-like protein is clearly evident, orthology between the other metazoan and higher plant ARID-containing subfamilies can not be established based on sequence similarity within the ARID or the presence of common conserved elements outside the ARID.

Four ARID-containing proteins have been identified in Schizosaccharomyces pombe and two in Saccharomyces cerevisiae. Sequence homology within the ARID domain, orthologous binding partners, and the presence of additional conserved elements suggest that these are fungal orthologs of ARID1 and JARID1 and possibly ARID2 and JARID2. These fungal ARID-containing proteins have been identified as both positive and negative regulators of transcription, and as components of nucleosome remodeling complexes.

Fig. 1. The human ARID family of proteins. Genome sequencing reveals 15 ARID-containing proteins in humans. The ARID family proteins can be grouped into subfamilies based on their similarity to each other within the ARID domain. The nomenclature described here reflects this subclassification of the family and clarifies their relationships to each other. A subset of ARID-containing proteins also contains JmjN and JmjC domains, and the proposed nomenclature reflects these relationships as well. Within the proposed subfamilies, members typically have 70 to 85% identity within their ARID sequences, while across subgroups, identity within the ARID sequence drops to about 25 to 30%. The 15 human ARID family proteins are represented by open bars and are aligned according to the position of the ARID sequence (indicated in yellow). The relative positions of other well-characterized domains and motifs are represented by colored symbols identified at the bottom of the figure. The amino acid (aa) length of each protein is shown to the right of the bar. The presence of additional motifs was identified through the Pfam [55] or SMART [56] databases.
The previous nomenclature of the mammalian genes and their products was quite confusing. Most of the proteins had more than one name that was not consistent between species (Tables 1 and 2). Closely related paralogous gene products, such as RBP1 and BRCAA1, had different names that did not indicate their evolutionary relationship. In other cases the previous names suggested a closer relationship than exists. For example, RBP1 and RBP2 were each isolated as pRb-binding proteins, but they have little sequence homology outside of the ARID domain and are not closely evolutionarily related. In at least one case (OSA1), almost identical names had been proposed for two different proteins. We suggest here a common nomenclature system for the mouse and human ARID family genes recommended by the Mouse Genomic

<table>
<thead>
<tr>
<th>Table 1 Human ARID nomenclature</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARID subfamilies</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>ARID1</td>
</tr>
<tr>
<td>ARID1B / ARID1B</td>
</tr>
<tr>
<td>ARID2</td>
</tr>
<tr>
<td>ARID3</td>
</tr>
<tr>
<td>ARID3B / ARID3B</td>
</tr>
<tr>
<td>ARID3C / ARID3C</td>
</tr>
<tr>
<td>ARID4</td>
</tr>
<tr>
<td>ARID4B / ARID4B</td>
</tr>
<tr>
<td>ARID5</td>
</tr>
<tr>
<td>ARID5B / ARID5B</td>
</tr>
<tr>
<td>JARID1</td>
</tr>
<tr>
<td>JARID1B / JARID1B</td>
</tr>
<tr>
<td>JARID1C / JARID1C</td>
</tr>
<tr>
<td>JARID1D / JARID1D</td>
</tr>
<tr>
<td>JARID2</td>
</tr>
</tbody>
</table>

a Many of the protein sequence predictions are not yet absolutely definitive; in general the selected accession numbers represent NCBI RefSeqs from curated series. Some variations are noted in the final column.

<table>
<thead>
<tr>
<th>Table 2 Mouse ARID nomenclature</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARID subfamilies</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>ARID1</td>
</tr>
<tr>
<td>Arid1b / ARID1B</td>
</tr>
<tr>
<td>ARID2</td>
</tr>
<tr>
<td>Arid3a / Arid3A</td>
</tr>
<tr>
<td>Arid3b / Arid3B</td>
</tr>
<tr>
<td>Arid3c / Arid3C</td>
</tr>
<tr>
<td>ARID4</td>
</tr>
<tr>
<td>Arid4b / Arid4B</td>
</tr>
<tr>
<td>ARID5</td>
</tr>
<tr>
<td>Arid5b / Arid5B</td>
</tr>
<tr>
<td>JARID1</td>
</tr>
<tr>
<td>Jarid1b / JARID1B</td>
</tr>
<tr>
<td>Jarid1c / JARID1C</td>
</tr>
<tr>
<td>Jarid1d / JARID1D</td>
</tr>
<tr>
<td>JARID2</td>
</tr>
</tbody>
</table>

a Many of the protein sequence predictions are not yet absolutely definitive; in general the selected accession numbers represent NCBI RefSeqs from curated series. Some variations are noted in the final column.
Results and discussion

Human and mouse ARID family proteins can be grouped into seven subfamilies based on the degree of sequence similarity within the regions of the ARID consensus and across their full-length sequences (Figs. 1–3). The nomenclature described here reflects the natural evolutionary divisions by grouping identifiable paralogous genes. In addition, a major division is recognized for the subset of ARID-containing proteins that also contain the JmjN and JmjC domains, for which orthologs can be identified in plants, fungi, and metazoans. Within each designated subfamily in mammals the degree of amino acid identity within the ARID regions is very high, ranging from 70 to 83%, as shown in Fig. 3. In contrast, amino acid identity within ARID regions between subfamilies is less than 30%. Members within subfamilies generally also show conserved domains and other related elements outside the ARID, as shown in Fig. 1. We describe here each of the proposed subfamilies with emphasis on their relationship to each other and to their apparent Drosophila and yeast counterparts. A short discussion of the DNA-binding properties of the family is included, as well as a discussion of other potential functional motifs, and a brief review of the biological role of each protein as it is currently understood.

ARID1A and ARID1B

The first subfamily is designated ARID1. It contains two members, ARID1A and ARID1B. Each of these proteins has been described in the literature under a variety of different names (see Tables 1 and 2). They are alternative components of mammalian SWI/SNF-related chromatin remodeling complexes [9,10], and ARID1A appears to be silenced in a subset of breast and kidney tumors [11,12]. ARID1A and ARID1B are 80% identical within the ARID and approximately 50% identical across their full-length amino acid sequences, although ARID1A has additional glutamine-rich regions and several LXXLL motifs (presumptive nuclear hormone receptor-binding sites) that are not precisely conserved in ARID1B (see Fig. 1). ARID1A and ARID1B are the apparent mammalian counterparts of Drosophila OSA and S. cerevisiae Swi1p, which are ARID-containing components of Drosophila and yeast SWI/SNF-type complexes, respectively. The ARID1 subfamily is one of two (the other is JARID1, below) that has an identifiable ortholog in budding yeast. ARID1A and ARID1B both have broad tissue distributions [9,13–16].

ARID2

The second subfamily is designated ARID2, and has only one member. ARID2 is an apparent ortholog of the Drosophila ARID protein BAP170 (aliases: BCDNA:GH12174 or CG3274). Both proteins contain an RFX domain, which is an additional DNA-binding domain. The RFX domain is named after Regulatory Factor X, a protein that binds to the X-box of MHC class II genes and is important for their expression (reviewed in [17]). Human ARID2, under the designation DFKZ p686G052, has been identified as a novel partner of the cyclin A1-CDK1 complex through a yeast triple-hybrid approach [18]. This study reports broad expression of ARID2, with a particularly high level in testis. The Drosophila ARID2 ortholog BAP170 was identified recently as a component of the SWI/SNF-like complex, PBAP [19]. The PBAP complex is distinguished from the prototypical Drosophila SWI/SNF-type complex, called BAP, in part by its lack of OSA, and was therefore thought to lack an ARID-containing component. The finding of BAP170 in PBAP is further evidence that ARID-containing subunits are important components of SWI/SNF-related chromatin-remodeling complexes. The mutually exclusive presence of different ARID-containing proteins in the complexes suggests that the ARID proteins contribute to functional specificity.

ARID3A, ARID3B, and ARID3C

The third mammalian ARID subfamily is designated ARID3, and contains three members. The members of the
ARID3 subfamily are the most direct mammalian counterparts of *Drosophila* DRI. ARID3A (originally identified as Bright) and ARID3B are similar in size (75 and 61 kDa, respectively) and are 80% identical in their ARID sequences. ARID3C was identified more recently and is less well studied. The total length of the human protein is predicted to be 412 amino acids. ARID3C is approximately 80% identical to ARID3A and ARID3B in its ARID sequence, and the three proteins are also closely related across the remaining amino acid sequence. The orthology of the members of the mammalian ARID3 subfamily to *Drosophila* DRI is supported by the degree of conservation within the ARID domain as well as by the presence of additional conserved regions both N and C terminal of the ARID which form alpha helices in the protein structure. These N and C terminal helices have together been termed the extended ARID (eARID). The sequence of the eARID regions is more than 70% identical between DRI and mammalian ARID3 proteins, but does not occur in other ARID family members. The C-terminal eARID region has been shown to contact DNA [20,21]. ARID3 subfamily members vary in their tissue distribution. ARID3A is highly expressed in mature B cells [5,22], where it can displace the homeoprotein CUX (CUTL1) to activate the immunoglobulin heavy chain intronic enhancer, E\textsubscript{A} [23]. Human ARID3A binds the pRb-controlled transcription factor E2F1 and rescues Ras-induced senescence in primary murine fibroblasts [24]. In contrast to ARID3A, ARID3B has a broad tissue distribution. The pattern of ARID3C expression has not yet been reported.

ARID4A and ARID4B

The fourth mammalian ARID subfamily is designated ARID4, and contains two members, ARID4A and ARID4B. These proteins are 74% identical within their ARID domains and 40 to 50% identical across the full length of each protein. While both proteins contain a Tudor domain...
and a chromodomain, ARID4A alone contains an LXCXE (pRB-binding) motif. Chromodomains bind to methylated lysine residues, and are therefore implicated in transcriptional repression. Consistent with this observation, ARID4A has been characterized as a repressor of E2F transcription recruited by pRb [25–27]. Both ARID4A and ARID4B have been found in association with the mSin3-histone deacetylase complex and can repress expression of Gal-4 reporter constructs when fused to a Gal-4 DNA-binding domain [27,28]. The ARID4 subfamily proteins appear to be orthologs of the Drosophila protein CG7274, which likewise contains an ARID domain and a Tudor domain. The function of the Tudor domain is not yet clear, but its name derives from its presence in the Drosophila Tudor protein and it is present in several RNA-binding proteins. ARID4A is broadly expressed. ARID4B expression is tightly restricted in normal tissue; it is abundant only in testis, but was first observed as a tumor marker expressed frequently in human carcinomas [29].

**ARID5A and ARID5B**

The fifth ARID subfamily, ARID5, contains two members, ARID5A and ARID5B. Both proteins were characterized by their ability to bind similar AT-rich sequences in the transcriptional modulator of the human cytomegalovirus major immediate-early promoter and to repress transcription from this promoter. Accordingly, they were originally named Modulator Recognition Factors 1 and 2 [30]. The proteins are more than 70% identical within their ARID sequences but are not similar outside the ARID. They are grouped together because their ARID sequences share more similarity to each other than to other members of the ARID family. Expression profiles for ARID5A have not been reported, but ARID5B is expressed broadly [31]. Mouse knockout studies show that targeted disruption of Arid5b does not affect embryonic growth or birth rate, but does result in a high rate of neonatal mortality in homozygotes. This is associated with severe postnatal reduction of lipid accumulation in the Arid5b-null mice, suggesting that Arid5B is essential for accumulation of lipid stores [32]. Homozygotes also have a high incidence of abnormalities of both male and female reproductive organs [31].

**JARID1A, JARID1B, JARID1C, and JARID1D**

The sixth ARID subfamily is designated JARID1 to reflect the presence in these proteins of a feature that defines another recognized protein family, the JmjN and JmjC domains. JARID1 is the largest ARID subfamily. It contains four highly homologous members. JmjN and JmjC domains are present together in at least three other human proteins that do not include ARID domains, while JmjC alone is present in many other proteins. In the fission yeast protein Epe1, a JmjC domain is essential for an activity that counteracts transcriptional silencing by destabilizing heterochromatin [33]. The four JARID1 proteins have more than 80% amino acid identity within the ARID and are highly related across their full sequences. All four have conserved PHD domains, zinc finger-C5HC2 domains and multiple LXXLL motifs, aside from their ARID and Jmj domains. However, only JARID1A contains an LXCXE motif, the presence of which is consistent with the original isolation of JARID1A as a pRb-binding protein. JARID1A enhances nuclear hormone receptor-mediated transcription in reporter assays [34]. The apparent Drosophila ortholog of the JARID1 subfamily is the LID protein. JARID1 is the second subfamily (together with ARID1) that has an apparent counterpart among the two ARID proteins of S. cerevisiae. All of the JARID1 subfamily proteins have broad tissue distribution with the exception of JARID1B (previously designated PLU-1), which is normally tightly restricted to testis and is frequently up-regulated in breast cancer tissue [35]. JARID1B has transcriptional repression activity that may be restricted to certain meiotic stages [36]. Though broadly expressed, JARID1C is particularly abundant in brain, and functional loss of JARID1C results in an X-linked mental retardation syndrome [37].

**JARID2**

The seventh ARID subfamily, JARID2, has only one member. Although JARID2 has JmjN, JmjC, and zinc finger-C5HC2 domains in common with the JARID1 subfamily, it has been assigned to a separate subfamily because the members of JARID1 are more similar to each other than to JARID2. Within the ARID domain, JARID2 is only about 25% identical to members of the JARID1 group, which are more than 80% identical to each other. The Drosophila ortholog of JARID2 is a gene product known as CG3654. JARID2 was first isolated in a mouse gene-trap strategy. In the original study, mutant Jarid2 was linked with formation of an abnormal cruciform-shaped neural groove [38], inspiring the name jumonji (a Japanese term referring to the cruciform shape). Most of our current knowledge of the function of JARID2 derives from developmental studies of the Jarid2 knockout mice. Only recently are the molecular mechanisms beginning to be examined. Jarid2 expression is developmentally important in liver and cardiac cells, and is required for repression of cyclin D1 in cardiac myocytes, an activity through which it may regulate cellular proliferation [39–44].

**DNA-binding activity of ARID-containing proteins**

The ARID was originally identified in ARID3 subfamily proteins as an AT-rich interactive domain [5,6]. ARID5 subfamily proteins were soon shown to bind preferentially to AT-rich sites as well [30,45]. However, when ARID1 subfamily proteins were cloned in Drosophila and humans, they showed DNA-binding activity, but no preference for specific DNA sequences [13,46]. JARID2 also fails to show
a clear preference for AT-rich sequences [43]. A survey of the seven mammalian ARID subfamilies indicates that only ARID3 and ARID5 interact preferentially with AT-rich sites; the remaining subfamilies bind with varying degrees of affinity, and no clear sequence preference [21]. The subfamilies can therefore be classified into two groups: AT-rich specific or nonspecific.

Solution structures have been obtained for two AT-rich-specific ARIDs (the ARID3 ortholog DRI, and ARID5B) and for a nonspecific ARID (ARID1A). Analysis of DRI ARID structure in complex with DNA [2,20] indicates the presence of eight alpha helices (H0 through H7) separated by turns or loops. The H4-Loop2-H5 region is a modified helix-turn-helix motif and contacts the major groove (Fig. 4). Contacts outside the major groove are formed as well; these involve primarily the Loop 1 region (which is actually a beta sheet in DRI) and the C-terminal eARID region. The ARID structures of ARID5B [1,47] and ARID1A [4] are similar and make similar contacts, although the ARID1A interaction is not sequence specific. The major differences in ARID1A are that a region immediately N-terminal to H0 makes DNA contact, and Loop 2 does not interact directly with the major groove. A comparison of DNA contact sites in the three ARID structures can be seen in [4].

Site-specific mutagenesis supports the structural indication of likely DNA interaction sites, but does not reveal precise determinants for sequence specificity or lack of it within the ARID family; more likely binding specificity is determined by multiple interacting differences across the entire ARID structure [21]. A similar situation appears to hold for the distinction between sequence-specific and sequence nonspecific DNA binding in high mobility group (HMG) domain proteins [48,49].

ARID3 subfamily proteins act as oligomers. Oligomerization is required for high-affinity binding of ARID3A to its AT-rich consensus site, and is mediated by a conserved
element in the C-terminus of the protein outside of the ARID and eARID regions (P. Tucker, unpublished data). The ARID3 ortholog DRI likewise forms oligomers dependent on a sequence outside of the ARID and eARID regions [50]. However, the identified oligomerization sequence is not strictly required for DNA binding or sequence specificity in DRI [6, 20, 21, 50].

Sequence specificity, or lack of it, within any particular subfamily appears to remain consistent in divergent organisms including *D. melanogaster*, *C. elegans*, and *S. cerevisiae* [6, 21, 46, 51, 52]. The four *C. elegans* ARID proteins align with ARID1, ARID2, ARID3, and JARID1. Within this smaller panel, an ARID3 representative is still present, and its preference for AT-rich binding sites has been verified experimentally [51]. The two *S. cerevisiae* ARID proteins and the four *S. pombe* ARID proteins are harder to align with specific human subfamilies, but they appear to correlate best with the ARID1 and/or ARID2 subfamilies and the JARID1 subfamily (there may be two representatives of JARID1 in *S. pombe*). These patterns suggest that the DNA-binding preference of the ARID domain was initially sequence nonspecific and the property of sequence specificity was gained in certain gene lineages as the domain evolved.

**Requirement for the ARID in the biological role of the protein**

The precise function of all the mammalian ARID proteins is not known. Members of the AT-specific ARID3 and ARID5 subfamilies appear to be conventional sequence-specific transcription factors with recognized promoter-targeting functions and important roles in development and differentiation [5, 6, 30–32, 53]. Among the sequence nonspecific ARID proteins, several appear to participate in general transcription and chromatin remodeling functions, as discussed above. To date, only the DRI ARID has actually been shown to be required for the physiological function of its cognate protein; an in-frame deletion of approximately seven amino acids within the ARID consensus is unable to rescue the *dri* phenotype during embryonic development of the fly [50]. The ARID of the *S. cerevisiae* protein Swi1p appears dispensable for complementation of the SW1 phenotype [54] but transient reporter assays suggest the ARID is required for a SWI/SNF-type complex-mediated transactivation function in human ARID1B [9]. More physiological experiments are needed. A possible function for the ARID in an ARID1 class protein is suggested by the ability of a 233 residue ARID-containing portion of *Drosophila* OSA to enhance or partially rescue the *osa* phenotype in vivo, when fused to the VP16 activation domain or the repressor domain of Engrailed, respectively, and targeted to the developing notum [46]. A similar functional assay was based on the two repression domains of ARID4A [25, 26], which can act as a repressor of E2F-dependent transcription. One repression domain (R1) maps to a 211 residue ARID-containing portion of the protein. A fusion of R1 alone to a Gal4-DNA-binding region is sufficient to repress expression of a CAT reporter under the control of the Gal4 minimal herpesvirus thymidine kinase promoter. A fusion protein with either R1 or repression domain 2 (R2) deleted is still able to repress, but the absence of both domains abrogates the repression function.

**Summary**

The ARID is an ancient DNA-binding domain that is conserved throughout the evolution of higher eukaryotes. ARID-containing proteins exhibit a wide range of cellular functions, including participation in chromatin remodeling, and regulation of gene expression, cellular differentiation, and growth. Orthologs to each of the seven distinct subfamilies of mammalian ARID-containing proteins are found in insect genomes, indicating the minimum age for the organization of these higher metazoan subfamilies. Many of these ancestral genes were subsequently duplicated and fixed over time to yield the 15 ARID-containing genes that are found in the human, mouse, and dog genomes. This paper describes a nomenclature for these mammalian ARID-containing genes that reflects this evolutionary history.

**Materials and methods**

**Sequencing**

New sequence obtained for this study has been deposited under Accession Numbers: AY727870 (human ARID2), AJ884581 (human ARID3C), and AJ884580 (mouse Arid3C).

**Chromosomal assignments**


**Identification of protein motifs**

The presence of protein motifs was identified through the Pfam [55] or SMART [56] databases.

**Sequence alignment**

ARID domain amino acid sequences were derived by conceptual translation of mouse and human cDNA sequences and aligned manually. Genomic sequences were compared with cDNA-derived sequences to identify the probable splice junctions. In each case, a conventional GT-AG splice junction could account for the reported expressed gene sequence, with the exception of the second intron of JARID1B, where in mouse a GC-AG intron is used. For each dog gene, the genomic locus was identified using.
BLAST and exons were aligned with the human and mouse exons. In each case, a GT-AG splice junction could be identified that aligned precisely with the mouse/human splice junctions. All exons within the dog genome for the ARID domains could be identified unambiguously with the exception of the second ARID2 exon, for which no sequence could be found, and the JARID1D gene, found on the Y chromosome in mouse and human, a chromosome that has not yet been sequenced in dog.

Dendrogram

DNA sequences corresponding to the amino acid sequence shown in Fig. 3 were analyzed using the MrBayes software program [57] to create an unrooted consensus tree.

Structure

Structure representations were generated using the Swiss-Pdb viewer [58], version Deep View-spdbv 3.7 to view the .pdb file (1KQQ) of the Drosophila DRI/RETN ARID domain bound to DNA. A ribbon structure for the protein was created and residues determined to be involved in protein-DNA contact [20] were shaded red. A molecular surface for the double-stranded DNA was calculated and those nucleotides involved in protein-DNA contact [20] were shaded pink or blue. ARID structures derived from the following .pdb files, 1RYU (ARID1A) [4], 1IG6 (ARIDS5B) [47], and 1KN5 (Swi1p) [3], were aligned in SwissPdb viewer with the DRI/RETN structure. All structures were exported to POV-ray (http://www.povray.org/) and rendered. The resulting images were composited in Adobe Illustrator (http://www.adobe.com/), and the contact DNA nucleotides were outlined in yellow. For each structure, the ‘best model’ from the NCBI’s MMDB [59] was downloaded and used in the preparation of the figure.

Acknowledgments

We thank Yuan Chen, Antonia Patsialou, and Albert Lai for helpful comments in the preparation of the manuscript. We thank Josephine Curcio for providing the human and mouse cDNA sequences for ARID3C. This work was supported by PHS Grants CA53592 (E.M.), CA31534 (P.W.T.), and HG00330 (L.M.) from the NIH. D.W. is the recipient of DOD BCRP fellowship DAMD-17-01-1-0407 and a Daniel Swern Fellowship from Temple University. L.P. is a trainee under an IGERT grant from the National Science Foundation (DGE-0114387).

References

T. Takeuchi, Y. Yamazaki, Y. Katoh-Fukui, R. Tsuchiya, S. Kondo, J.
L.R. Jensen, et al., Mutations in the JARID1C gene, which is involved

P.J. Lu, et al., A novel gene (PLU-1) containing highly conserved

S.W. Chan, W. Hong, Retinoblastoma-binding protein 2 (Rbp2)

M.H. Lahoud, et al., Gene targeting of Desrt, a novel ARID class

J. Cao, T. Gao, E.J. Stanbridge, R. Irie, RBP1L1, a retinoblastoma-

T.C. Fleischer, U.J. Yun, D.E. Ayer, Identification and characterization

A. Lai, et al., RBP1 recruits both histone deacetylase-dependent and -

D.S. Peeper, A. Shivarts, T. Brummelkamp, D. Wilsker et al. / Genomics 86 (2005) 242 – 251

J. Chen, J.B. Anderson, C. DeWeese-Scott, N.D. Fedorova, L.Y. Geer,

J.P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of

A. Bateman, et al., The Pfam protein families database, Nucleic Acids Res.

S. Shaham, C.I. Bargmann, Control of neuronal subtype identity by the

D. Wilsker, A. Patsialou, S.D. Zumbrun, S. Kim, Y. Chen, P.B. Dallas,

R.T. Collins, T. Furukawa, N. Tanese, J.E. Treisman, Bright, is not expressed in all human B

D. Wilsker, A. Patsialou, S.D. Zumbrun, S. Kim, Y. Chen, P.B. Dallas,

R.T. Collins, T. Furukawa, N. Tanese, J.E. Treisman, Bright, is not expressed in all human B


Y.C. Yuan, R.H. Whitson, K. Itakura, Y. Chen, Resonance assign-


Y.C. Yuan, R.H. Whitson, K. Itakura, Y. Chen, Resonance assign-


Y.C. Yuan, R.H. Whitson, K. Itakura, Y. Chen, Resonance assign-


Y.C. Yuan, R.H. Whitson, K. Itakura, Y. Chen, Resonance assign-