Available online at www.sciencedirect.com

SCIENCE DIRECT

Genomics 86 (2005) 242 – 251

GENOMICS

www.elsevier.com/locate/ygeno

Nomenclature of the ARID family of DNA-binding proteins☆

Deborah Wilsker[a,1], Loren Probst[b,1], Hester M. Wain[c], Lois Maltais[d], Philip W. Tucker[b], Elizabeth Moran[a,*]

[a]*Fels Institute for Cancer Research and Molecular Biology, Temple University School of Medicine, Philadelphia, PA 19104, USA*
[b]*Institution for Cellular and Molecular Biology, Department of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, TX 78712-1095, USA*
[c]*HUGO Gene Nomenclature Committee (HGNC), Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK*
[d]*Mouse 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 is an ancient 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.
© 2005 Elsevier Inc. All rights reserved.

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

☆ Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession Nos.

* Corresponding author. Fax: +1 215 707 7313.
E-mail address: betty@temple.edu (E. Moran).

[1] These authors contributed equally.

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.03.013

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.

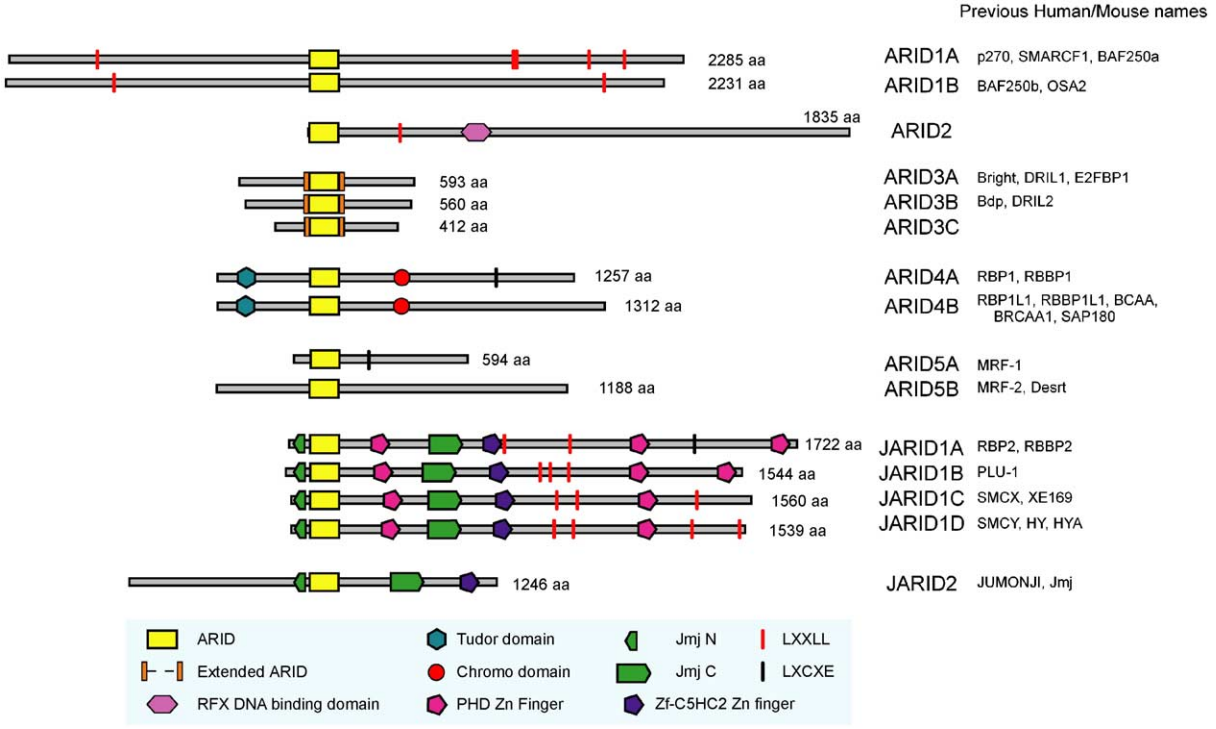

Fig. 1. The human ARID family of proteins. Genome sequencing reveals 15 ARID-containing proteins in humans. The ARID family proteins can be grouped into subfamilies based on their similarity to each other within the ARID domain. The nomenclature described here reflects this subclassification of the family 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.

Table 1
Human ARID nomenclature

ARID subfamilies	HUGO nomenclature (gene / protein)	Previous human names	Chromosomal location	Accession number (selected)[a]	Notes
ARID1	*ARID1A* / ARID1A	p270, SMARCF1, BAF250a, B120, OSA1, hOsa1, p250, C1orf4, hSWI1	1p35.3	NP_006006.3	An alternative splice variant of ARID1B lacking one exon directly 5′ to the ARID, is described in NM_175863.2. The optional exon is not present in ARID1A.
	ARID1B / ARID1B	pKIAA1235, BAF250b, p250R, hOsa2, hELD/OSA1	6q25.1	NP_059989.1	
ARID2	*ARID2* / ARID2	pKIAA1557, DKFZp686G052	12q12	NM_152641.2	
ARID3	*ARID3A* / ARID3A	Bright, DRIL1, E2FBP1	19p13.3	NP_005215.1	
	ARID3B / ARID3B	Bdp, DRIL2	15q24	NP_006456.1	
	ARID3C / ARID3C	LOC138715: similar to E2F binding protein, XM_071061, Bright-like	9p13.2	NM_001017369	
ARID4	*ARID4A* / ARID4A	RBP1, RBBP1	14q23.1	NP_002883.2	
	ARID4B / ARID4B	RBP1L1, BCAA, BRCAA1, SAP180, RBBP1L1	1q42.1–q43	NP_057458.4	
ARID5	*ARID5A* / ARID5A	MRF-1	2q11.2	NP_997646.1	
	ARID5B / ARID5B	MRF2	10q21.2	Q14865	
JARID1	*JARID1A* / JARID1A	RBP2, RBBP2	12p11	NP_005047.1	
	JARID1B / JARID1B	PLU-1, PUT1, RBBP2H1A	1q32.1	NP_006609.3	
	JARID1C / JARID1C	SMCX, XE169, DXS1272E	Xp11.22–p11.21	NP_004178.1	
	JARID1D / JARID1D	SMCY, HY, HYA, KIAA0234	Yq11	NP_004644.2	
JARID2	*JARID2* / JARID2	jumonji	6p24–p23	NP_004964.2	

[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.

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 2
Mouse ARID nomenclature

ARID subfamilies	MGNC nomenclature (gene / protein)	Previous murine names	Chromosomal location	Accession number (selected)[a]	Notes
ARID1	*Arid1a* / ARID1A	1110030E03Rik, Osa1, Smarcf1	4 D3	NP_291044.1	*Arid1a* is incomplete at 5′ end.
	Arid1b / ARID1B	B230217J03Rik	17 A1	XP_139711.5	*Arid1b* sequence diverges from human at 5′ end and may be incorrect.
ARID2	*Arid2* / ARID2	4432409D24Rik	15F1	NP_780460.1	ARID2 is incomplete at 5′ end.
ARID3	*Arid3a* / ARID3A	Bright, Dri1	10 C1	NP_031906.1	
	Arid3b / ARID3B	Bdp, Dri2	9C	NP_062663.1	
	Arid3c / ARID3C		4B1	NM_001017362	
ARID4	*Arid4a* / ARID4A	Rbbp1, A630009N03	12 C2	XP_354675.2	
	Arid4b / ARID4B	BCAA, BRCAA1, Rbp1l1, SAP180, RBBP1_1, 6330417L24Rik, 6720480E17Rik	13 A1	NP_919238.1	
ARID5	*Arid5a* / ARID5A	D430024K22Rik, Mrf1	1 B	NP_666108.2	
	Arid5b / ARID5B	Mrf2, Desrt, 4930580B11, 530435G07Rik	10 B5.1	AAM93269.1	
JARID1	*Jarid1a* / JARID1A	RBP2, Rbbp2, MGC11659	6 F1	XP_359326.2	
	Jarid1b / JARID1B	Plu1, 2010009J12Rik	1 E4	NP_690855.1	
	Jarid1c / JARID1C	Smcx	X F2–F4	NP_038696.1	
	Jarid1d / JARID1D	Smcx	Y A1	NP_035549.1	
JARID2	*Jarid2* / JARID2	Jmj	13 A5	NP_068678.1	

[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.

Nomenclature Committee (MGNC) and accepted by the Human Genome Organization (HUGO) Gene Nomenclature Committee. The nomenclature of the ARID family genes can also be viewed at the Human Genome Database website (http://gdbwww.gdb.org/ accession ID GDB:11511844).

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.

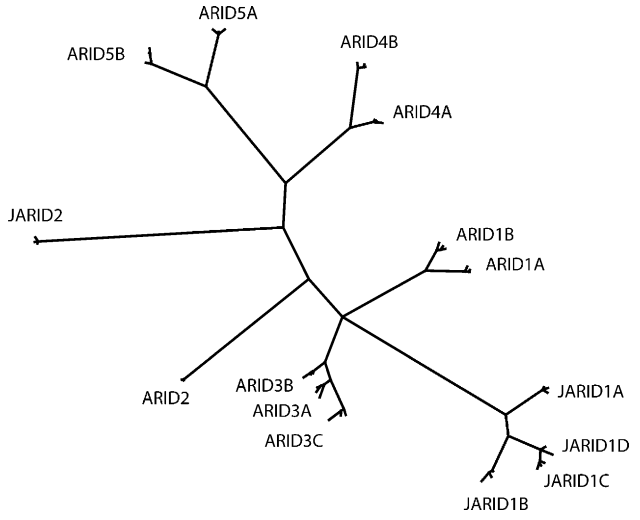

Fig. 2. Dendrogram of mouse, dog, and human ARID domains. The relationships between the different ARID subfamilies can be seen in this dendrogram, which was created from DNA sequences corresponding to the protein segments in Fig. 3 using the MrBayes software [57] to perform the analysis to create this unrooted consensus tree. As can be seen, the domains are naturally grouped with orthologs falling near the tips of the branches, indicating their similarity. In most cases domain conservation between species is so extensive that the three orthologs at the tips of the branches do not resolve clearly in the figure. The orthology relationships are confirmed by the presence of additional common conserved elements outside the ARID domain.

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 *R*egulatory *F*actor *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 DKFZ 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

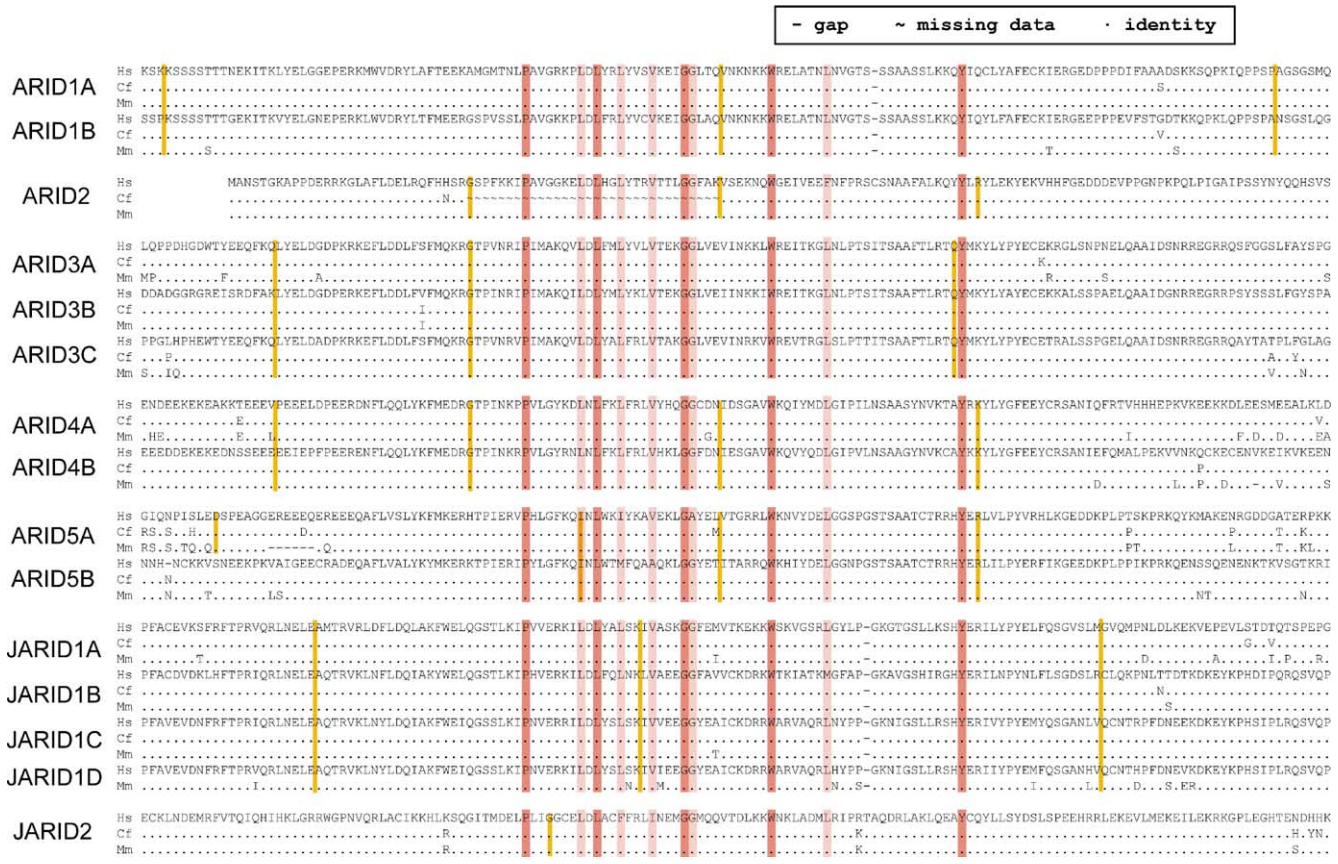

Fig. 3. Alignment of the mouse, dog, and human ARID domains. The alignment was created from translation of reported mouse (*Mus musculus*) and human (*Homo sapiens*) cDNA sequences, as well as from predicted ARID-containing gene sequences available from the dog (*Canis familiaris*) genome. In each of these organisms there is a one to one relationship between each ARID-containing ortholog (with the exception of the dog JARID1D gene, predicted to lie on the as yet unsequenced Y chromosome). The human sequences are shown in their entirety; dots indicate identity at the equivalent positions in the other species. In a few instances, small gaps were introduced to maximize the alignment; these are indicated by hyphens. Residue positions showing near or complete identity are highlighted in light and dark red, respectively. Putative exon-exon junctions are indicated by vertical yellow bars. Genomic structural conservation in addition to the sequence conservation confirms the paralogous gene subfamilies. The ARID of ARID2 is at the immediate N-terminus of the protein. The reported sequence for canine ARID2 is still incomplete; missing sequence is indicated by sim (tilde) symbols in the figure. The boundaries of sequence conservation defining the ARID across evolution encompass approximately 100 amino acids centered on the highly conserved positions highlighted in the figure; a larger portion of protein sequence is shown here to include flanking exon junctions.

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, Eμ [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

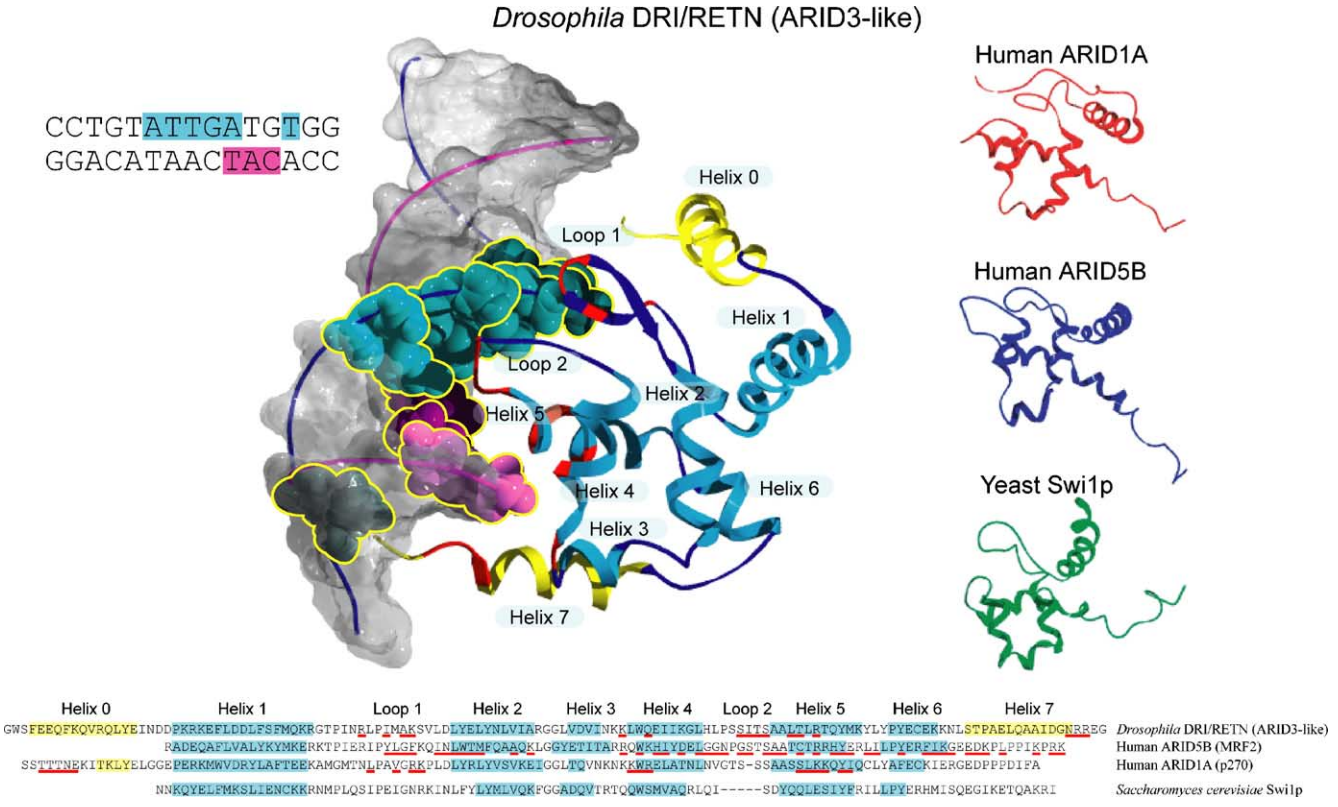

Fig. 4. Solution structures of the ARID domains of *Drosophila melanogaster* DRI/RETN (ARID3 ortholog), human ARID1A and ARID5B, and *Saccharomyces cerevisiae* Swi1p. The solved NMR structure of a peptide encompassing the ARID of the *Drosophila dri/retn* gene product complexed to a DNA sequence derived from its native binding site in a silencer within the *zen* gene locus identifies regions of the protein that are important for both major and minor groove contact by this ARID domain [20]. The core ARID helices are shown in light blue, and the extended ARID helices in yellow. Dark blue regions in the ribbon diagram indicate loops or turns between helices; the dark blue arrow in the structure indicates the beta sheet in the DRI ARID. Residues on the protein that contact DNA have been highlighted with red on the ribbon structure, and nucleotides within the DNA duplex that make protein contact are shown in blue or pink [20]. The 3′-most T nucleotide in the upper strand is shaded gray in the structure to indicate that this is on the reverse side of the molecule. Similarity of the *Drosophila* DRI/RETN ARID structure can be seen in comparison to the ARID structures of human ARID1A [4], human ARID5B [47], and *S. cerevisiae* Swi1p [3]. The peptide sequences used to generate the structures are aligned below. The amino acid sequences are wild type in all cases except DRI/RETN, where investigators [20] introduced a F > L substitution in Helix 5 (AAL*T). Helices indicated in the .pdb files are indicated by shading. The core ARID helices are shaded in light blue, and the extended ARID helices in yellow, and correspond with the respectively colored regions in the ribbon diagram. Regions that are predicted to contact DNA, based on structural analysis (DRI/RETN) [20] or chemical shift data (ARID1A and ARID5B) [4, 47, and Y. Chen, personal communication], are indicated by red underlining. The ARID structure of *S. cerevisiae* Swi1p was not solved in complex with DNA. Swi1p binds DNA with relatively low affinity, likely due in part to the short Loop 2 and the absence of basic residues (K or R) at the appropriate positions in Helix 5 [52].

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 non-specific 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 SWI 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 positive 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

Chromosomal assignments were based on information available at Entrez_Gene (http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene).

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 (ARID5B) [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

[1] Y.C. Yuan, R.H. Whitson, Q. Liu, K. Itakura, Y. Chen, A novel DNA-binding motif shares structural homology to DNA replication and repair nucleases and polymerases, Nat. Struct. Biol. 5 (1998) 959–964.

[2] J. Iwahara, R.T. Clubb, Solution structure of the DNA binding domain from Dead ringer, a sequence-specific AT-rich interaction domain (ARID), EMBO J. 18 (1999) 6084–6094.

[3] X. Tu, J. Wu, Y. Xu, Y. Shi, 1H, 13C and 15N resonance assignments and secondary structure of ADR6 DNA-binding domain, J. Biomol. NMR 21 (2001) 187–188.

[4] S. Kim, Z. Zhang, S. Upchurch, N. Isern, Y. Chen, Structure and DNA-binding sites of the SWI1 AT-rich interaction domain (ARID) suggest determinants for sequence-specific DNA recognition, J. Biol. Chem. 279 (2004) 16670–16676.

[5] R.F. Herrscher, M.H. Kaplan, D.L. Lelsz, C. Das, C. Scheuermann, P.W. Tucker, The immunoglobulin heavy-chain matrix-associating regions are bound by Bright: a B cell-specific trans-activator that describes a new DNA-binding protein family, Genes Dev. 9 (1995) 3067–3082.

[6] S.L. Gregory, R.D. Kortschak, B. Kalionis, R. Saint, Characterization of the dead ringer gene identifies a novel, highly conserved family of sequence-specific DNA binding proteins, Mol. Cell. Biol. 16 (1996) 792–799.

[7] D. Wilsker, A. Patsialou, P.B. Dallas, E. Moran, ARID proteins: a diverse family of DNA binding proteins implicated in the control of cell growth, differentiation, and development, Cell Growth Differ. 13 (2002) 95–106.

[8] R.D. Kortschak, P.W. Tucker, R. Saint, ARID proteins come in from the desert, Trends Biochem. Sci. 25 (2000) 294–299.

[9] H. Inoue, T. Furukawa, S. Giannakopoulos, S. Zhou, D.S. King, N. Tanese, Largest subunits of the human SWI/SNF chromatin-remodeling complex promote transcriptional activation by steroid hormone receptors, J. Biol. Chem. 277 (2002) 41674–41685.

[10] X. Wang, N.G. Nagl Jr., D. Wilsker, M. Van Scoy, S. Pacchione, P.B. Dallas, E. Moran, Two related ARID family proteins are alternative subunits of human SWI/SNF complexes, Biochem. J. 383 (2004) 319–325.

[11] M.F. Decristofaro, B.L. Betz, C.J. Rorie, D.N. Reisman, W. Wang, B.E. Weissman, Characterization of SWI/SNF protein expression in human breast cancer cell lines and other malignancies, J. Cell. Physiol. 186 (2001) 136–145.

[12] X. Wang, N.G. Nagl Jr., S. Flowers, D. Zweitzig, P.B. Dallas, E. Moran, Expression of p270 (ARID1A), a component of human SWI/SNF complexes, in human tumors, Int. J. Cancer 112 (2004) 636–642.

[13] P.B. Dallas, S. Pacchione, D. Wilsker, V. Bowrin, R. Kobayashi, E. Moran, The human SWI/SNF complex protein, p270, is an ARID family member with nonsequence -specific DNA binding activity, Mol. Cell. Biol. 20 (2000) 3137–3146.

[14] Z. Nie, Y. Xue, D. Yang, S. Zhou, B.J. Deroo, T.K. Archer, W. Wang, A specificity and targeting subunit of a human SWI/SNF family-related chromatin-remodeling complex, Mol. Cell. Biol. 20 (2000) 8879–8888.

[15] Z. Kozmik, O. Machon, J. Kralova, J. Kreslova, J. Paces, C. Vlcek, Characterization of mammalian orthologues of the Drosophila osa gene: cDNA cloning, expression, chromosomal localization, and direct physical interaction with Brahma chromatin-remodeling complex, Genomics 73 (2001) 140–148.

[16] A.F. Hurlstone, I.A. Olave, N. Barker, M. van Noort, H. Clevers, Cloning and characterization of hELD/OSA1, a novel BRG1 interacting protein, Biochem. J. 364 (2002) 255–264.

[17] P. Emery, B. Durand, B. Mach, W. Reith, RFX proteins, a novel family of DNA binding proteins conserved in the eukaryotic kingdom, Nucleic Acids Res. 24 (1996) 803–807.

[18] S. Diederichs, N. Baumer, P. Ji, S.K. Metzelder, G.E. Idos, T. Cauvet, W. Wang, M. Moller, S. Pierschalski, J. Gromoll, M.G. Schrader, H.P. Koeffler, W.E. Berdel, H. Serve, C. Muller-Tidow, Identification of interaction partners and substrates of the cyclin A1-CDK2 complex, J. Biol. Chem. 279 (2004) 33727–33741.

[19] L. Mohrmann, K. Langenberg, J. Krijgsveld, A.J. Kal, A.J.R. Heck, C.P. Verrijzer, Differential targeting of two distinct SWI/SNF-related Drosophila chromatin-remodeling complexes, Mol. Cell. Biol. 24 (2004) 3077–3088.
[20] J. Iwahara, M. Iwahara, G.W. Daughdrill, J. Ford, R.T. Clubb, The structure of the Dead ringer-DNA complex reveals how AT-rich interaction domains (ARIDs) recognize DNA, EMBO J. 21 (2002) 1197–1209.
[21] A. Patsialou, D. Wilsker, E. Moran, DNA binding properties of ARID family proteins, Nucleic Acids Res. 33 (2005) 66–80.
[22] J.C. Nixon, J.B. Rajaiya, N. Ayers, S. Evetts, C.F. Webb, The transcription factor, Bright, is not expressed in all human B lymphocyte subpopulations, Cell. Immunol. 228 (2004) 42–53.
[23] Z. Wang, A. Goldstein, R.T. Zong, D. Lin, E.J. Neufeld, R.H. Scheuermann, P.W. Tucker, Cux/CDP homeoprotein is a component of NF-muNR and represses the immunoglobulin heavy chain intronic enhancer by antagonizing the bright transcription activator, Mol. Cell. Biol. 19 (1999) 284–295.
[24] D.S. Peeper, A. Shvarts, T. Brummelkamp, S. Douma, E.Y. Koh, G.Q. Daley, R. Bernards, A functional screen identifies hDRIL1 as an oncogene that rescues RAS-induced senescence, Nat. Cell Biol. 4 (2002) 148–153.
[25] A. Lai, R.C. Marcellus, H.B. Corbell, P.E. Branton, RBP1 induces growth arrest by repression of E2F-dependent transcription, Oncogene 18 (1999) 2091–2100.
[26] A. Lai, et al., RBP1 recruits both histone deacetylase-dependent and -independent repression activities to retinoblastoma family proteins, Mol. Cell. Biol. 19 (1999) 6632–6641.
[27] A. Lai, et al., RBP1 recruits the mSIN3-histone deacetylase complex to the pocket of retinoblastoma tumor suppressor family proteins found in limited discrete regions of the nucleus at growth arrest, Mol. Cell. Biol. 21 (2001) 2918–2932.
[28] T.C. Fleischer, U.J. Yun, D.E. Ayer, Identification and characterization of three new components of the mSin3A corepressor complex, Mol. Cell. Biol. 23 (2003) 3456–3467.
[29] J. Cao, T. Gao, E.J. Stanbridge, R. Irie, RBP1L1, a retinoblastoma-binding protein-related gene encoding an antigenic epitope abundantly expressed in human carcinomas and normal testis, J. Natl. Cancer Inst. 93 (2001) 1159–1165.
[30] T.H. Huang, et al., Repression via differentiation-specific factor of the human cytomegalovirus enhancer, Nucleic Acids Res. 24 (1995) 1695–1701.
[31] M.H. Lahoud, et al., Gene targeting of Desrt, a novel ARID class DNA-binding protein, causes growth retardation and abnormal development of reproductive organs, Genome Res. 11 (2001) 1327–1334.
[32] R.H. Whitson, W. Tsark, T.H. Huang, K. Itakura, Neonatal mortality and leanness in mice lacking the ARID transcription factor Mrf-2, Biochem. Biophys. Res. Commun. 312 (2003) 997–1004.
[33] N. Ayoub, K. Noma, S. Isaac, T. Kahan, S.I. Grewal, A. Cohen, A novel jmjC domain protein modulates heterochromatization in fission yeast, Mol. Cell. Biol. 23 (2003) 4356–4370.
[34] S.W. Chan, W. Hong, Retinoblastoma-binding protein 2 (Rbp2) potentiates nuclear hormone receptor-mediated transcription, J. Biol. Chem. 276 (2001) 28402–28412.
[35] P.J. Lu, et al., A novel gene (PLU-1) containing highly conserved putative DNA/chromatin binding motifs is specifically up-regulated in breast cancer, J. Biol. Chem. 274 (1999) 15633–15645.
[36] B. Madsen, M. Tarsounas, J.M. Burchell, D. Hall, R. Poulsom, J. Taylor-Papadimitriou, PLU-1, a transcriptional repressor and putative testis-cancer antigen, has a specific expression and localisation pattern during meiosis, Chromosoma 112 (2003) 124–132.
[37] L.R. Jensen, et al., Mutations in the JARID1C gene, which is involved in transcriptional regulation and chromatin remodeling, cause X-Linked mental retardation, Am. J. Hum. Genet. 76 (2005) 227–236.
[38] T. Takeuchi, Y. Yamazaki, Y. Katoh-Fukui, R. Tsuchiya, S. Kondo, J. Motoyama, T. Higashinakagawa, Gene trap capture of a novel mouse gene, jumonji, required for neural tube formation, Genes Dev. 9 (1995) 1211–1222.
[39] Y. Lee, A.J. Song, R. Baker, B. Micales, S.J. Conway, G.E. Lyons, Jumonji, a nuclear protein that is necessary for normal heart development, Circ. Res. 86 (2000) 932–938.
[40] K. Kitajima, M. Kojima, S. Kondo, T. Takeuchi, A role of jumonji gene in proliferation but not differentiation of megakaryocyte lineage cells, Exp. Hematol. 29 (2001) 507–514.
[41] H. Anzai, A. Kamiya, H. Shirato, T. Takeuchi, A. Miyajima, Impaired differentiation of fetal hepatocytes in homozygous jumonji mice, Mech. Dev. 120 (2003) 791–800.
[42] M. Toyoda, et al., Jumonji downregulates cardiac cell proliferation by repressing cyclin D1 expression, Dev. Cell. 5 (2003) 85–97.
[43] T.G. Kim, J.C. Kraus, J. Chen, Y. Lee, JUMONJI, a critical factor for cardiac development, functions as a transcriptional repressor, J. Biol. Chem. 278 (2003) 42247–42255.
[44] T. Ohno, K. Nakajima, M. Kojima, M. Toyoda, T. Takeuchi, Modifiers of the jumonji mutation downregulate cyclin D1 expression and cardiac cell proliferation, Biochem. Biophys. Res. Commun. 317 (2004) 925–929.
[45] R.H. Whitson, T. Huang, K. Itakura, The novel Mrf-2 DNA-binding domain recognizes a five-base core sequence through major and minor-groove contacts, Biochem. Biophys. Res. Comm. 258 (1999) 326–331.
[46] R.T. Collins, T. Furukawa, N. Tanese, J.E. Treisman, Osa associates with the Brahma chromatin remodeling complex and promotes the activation of some target genes, EMBO J. 18 (1999) 7029–7040.
[47] Y.C. Yuan, R.H. Whitson, K. Itakura, Y. Chen, Resonance assignments of the Mrf-2 DNA-binding domain, J. Biomol. NMR 11 (1998) 459–460.
[48] F.V. Murphy IV, M.E. Churchill, Nonsequence-specific DNA recognition: a structural perspective, Struct. Fold Des. 8 (2000) R83–R89.
[49] J.O. Thomas, A.A. Travers, HMG1 and 2, and related 'architectural' DNA-binding proteins, Trends Biochem. Sci. 26 (2001) 167–174.
[50] T. Shandala, R.D. Kortschak, R. Saint, The Drosophila retained/dead ringer gene and ARID gene family function during development, Int. J. Dev. Biol. 46 (2002) 423–430.
[51] S. Shaham, C.I. Bargmann, Control of neuronal subtype identity by the *C. elegans* ARID protein CFI-1, Genes Dev. 16 (2002) 972–983.
[52] D. Wilsker, A. Patsialou, S.D. Zumbrun, S. Kim, Y. Chen, P.B. Dallas, E. Moran, The DNA-binding properties of the ARID-containing subunits of yeast and mammalian SWI/SNF complexes, Nucleic Acids Res. 32 (2004) 1345–1353.
[53] M.H. Kaplan, R.T. Zong, R.F. Herrscher, R.H. Scheuermann, P.W. Tucker, Transcriptional activation by a matrix associating region-binding protein. contextual requirements for the function of bright, J. Biol. Chem. 276 (2001) 21325–21330.
[54] P. Prochasson, K.E. Neely, A.H. Hassan, B. Li, J.L. Workman, Targeting activity is required for SWI/SNF function in vivo and is accomplished through two partially redundant activator-interaction domains, Mol. Cell. 12 (2003) 983–990.
[55] A. Bateman, et al., The Pfam protein families database, Nucleic Acids Res. 32 (2004) D138–D141.
[56] I. Letunic, et al., SMART 4.0: towards genomic data integration, Nucleic Acids Res. 32 (2004) D142–D144.
[57] J.P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogeny, Bioinformatics 17 (2001) 754–755.
[58] N. Guex, M.C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling, Electrophoresis 18 (1997) 2714–2723.
[59] J. Chen, J.B. Anderson, C. DeWeese-Scott, N.D. Fedorova, L.Y. Geer, S. He, D.I. Hurwitz, J.D. Jackson, A.R. Jacobs, C.J. Lanczycki, C.A. Liebert, C. Liu, T. Madej, A. Marchler-Bauer, G.H. Marchler, R. Mazumder, A.N. Nikolskaya, B.S. Rao, A.R. Panchenko, B.A. Shoemaker, V. Simonyan, J.S. Song, P.A. Thiessen, S. Vasudevan, Y. Wang, R.A. Yamashita, J.J. Yin, S.H. Bryant, MMDB: Entrez's 3D-structure database, Nucleic Acids Res. 31 (2003) 474–477.