Review

ARID Proteins: A Diverse Family of DNA Binding Proteins Implicated in the Control of Cell Growth, Differentiation, and Development

Deborah Wilsker, Antonia Patsialou, Peter B. Dallas, and Elizabeth Moran

Fels Institute for Cancer Research and Molecular Biology, Temple University School of Medicine, Philadelphia, Pennsylvania 19140 [D. W., A. P., E. M.], and TWN Telethon Institute for Child Health Research and the Center for Child Health Research, University of Western Australia, Subiaco, Western Australia 6008 [P. B. D.]

Abstract
The ARID family of DNA binding proteins was first recognized ~5 years ago. The founding members, murine Bright and Drosophila dead ringer (Dri), were independently cloned on the basis of their ability to bind to AT-rich DNA sequences, although neither cDNA encoded a recognizable DNA binding domain. Mapping of the respective binding activities revealed a shared but previously unrecognized DNA binding domain, the consensus sequence of which extends across ~100 amino acids. This novel DNA binding domain was designated AT-rich interactive domain (ARID), based on the behavior of Bright and Dri. The consensus sequence occurs in 13 distinct human proteins and in proteins from all sequenced eukaryotic organisms. The majority of ARID-containing proteins were not cloned in the context of DNA binding activity, however, and their features as DNA binding proteins are only beginning to be investigated. The ARID region itself shows more diversity in structure and function than the highly conserved consensus sequence suggests. The basic structure appears to be a series of six α-helices separated by β-strands, loops, or turns, but the structured region may extend to an additional helix at either or both ends of the basic six. It has also become apparent that the DNA binding activity of ARID-containing proteins is not necessarily sequence specific. What is consistent is the evidence that family members play vital roles in the regulation of development and/or tissue-specific gene expression. Inappropriate expression of ARID proteins is also increasingly implicated in human tumorigenesis. This review summarizes current knowledge about the structure and function of ARID family members, with a particular focus on the human proteins.

Introduction
About 5 years ago, a new class of DNA binding proteins, defined by a novel DNA binding domain, was recognized. The two proteins in which this domain was originally defined are murine Bright and Drosophila dead ringer. Bright is a B cell-specific transactivator cloned in a search for proteins binding to immunoglobulin heavy-chain matrix-associating regions. Matrix attachment regions are AT-rich sequences, and the Bright protein was indeed found to bind preferentially to AT-rich DNA sequences in an oligonucleotide selection and enhancement protocol (1). At the same time the Drosophila gene product, dead ringer (dri), was cloned in a search for novel proteins associating with homebox domains. Homeobox domains are also AT-rich sequences, and the Dri protein was likewise found to bind preferentially to AT-rich DNA sequences in a similar oligonucleotide selection and enhancement protocol (2).

What distinguished both of these proteins at the time was the lack of a recognizable DNA binding domain. When these investigators mapped the DNA binding regions in their respective proteins and realized they had identified highly related sequences, the parameters of a previously unrecognized DNA binding domain became apparent. The degree of conservation in the respective domains is remarkable, given that these proteins were cloned from distantly related organisms and that the proteins are not otherwise similar.

This novel DNA binding domain was designated ARID, based on the shared features of Bright and Dri. The derivation of a Bright/Dri consensus sequence led to the recognition of other ARID-containing proteins already cloned or subsequently added to the database. About a dozen distinct human ARID proteins have been recognized, as well as six Drosophila members of the family. ARID proteins or open reading frames are also apparent in yeast, Arabidopsis, and Caenorhabditis elegans. The majority of ARID-containing proteins were not cloned in the context of DNA binding activity, and their features as DNA binding proteins are only beginning to be investigated.

As awareness of the family has grown, a tighter consensus sequence has emerged. This consensus extends across ~100...
residues, of which ∼39 are highly conserved with regard to both identity and spacing. As a point of comparison, the homeodomain consensus spans 60 residues, of which ∼20 are highly conserved (reviewed in Ref. 3). The Bright/Dri homology extends ∼40 residues past the ARID consensus. This “extended ARID” sequence now appears to be characteristic of just one subfamily. Outside of the ARID region, proteins of the ARID family show diversity of sequence, structure, size, and function, although subgroups are readily discernible. The ARID region itself has proved to be more diverse in structure and function than the highly conserved consensus sequence suggests. The ARID protein family has been reviewed recently (4), but new information has since emerged, particularly in the realization that not all ARID proteins show sequence specificity in their DNA binding activity. The latter point is striking, given the high order of structure and degree of conservation of ARID regions. In general, such features are linked with increasing specificity as seen, for example, in homeodomains. ARID proteins are also becoming increasingly implicated in human tumorigenesis. This review summarizes what is currently known about the salient features of members of the ARID protein family. Our main focus is the mammalian proteins, although the Drosophila and yeast proteins are considered where relevant for comparison. Much of the information discussed in the text is summarized in Tables 1 and 2 and Figs. 1 and 2.

### Table 1 Functions of human ARID proteins*

<table>
<thead>
<tr>
<th>Yeast</th>
<th>Drosophila</th>
<th>Mouse</th>
<th>Human</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWI1</td>
<td>Osa (eld)</td>
<td>p270</td>
<td>p270</td>
<td>p270 is a component of human SWI/SNF complexes (11, 12, 14) and is deficient in some breast and ovarian cancer lines (60). osa associates with the Drosophila brahma (SWI/SNF-related) complex (5, 6), modifies E2F (47), and is an antagonist of wingless (46).</td>
</tr>
<tr>
<td>CG7274</td>
<td>KIAA1235</td>
<td>RBP1 (RBBP1)</td>
<td>Retinoblastoma binding protein-1 (20); represses E2F-dependent transcription (22–24).</td>
<td></td>
</tr>
<tr>
<td>ORF YMR716w (Ecm5p)</td>
<td>Lid (CG9088)</td>
<td>RBP1L1 (BCAA)</td>
<td>Retinoblastoma-binding protein-1-like 1. Highly expressed in cancers of various tissue origins but restricted in normal tissue (25).</td>
<td></td>
</tr>
<tr>
<td>SMCY</td>
<td>An evolutionarily conserved protein encoded on the Y chromosome (35).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMCX (XE169)</td>
<td>The X-chromosome homologue of SMCY; SMCX escapes X-inactivation (26).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLU-1</td>
<td>Up-regulated in breast cancer (27).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG3654</td>
<td>jumonji (JMJ)</td>
<td>Developmentally important in the nervous system, liver, spleen, thymus, and heart (28–30).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRF-1</td>
<td>Modulator recognition factor-1; represses CMV enhancer activity (43, 67).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dri</td>
<td>Modulator recognition factor-2; represses CMV enhancer activity (43, 67).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bright</td>
<td>Bdp (DRIL2)</td>
<td>Bright and dead ringer homologous protein (32).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCDNA: GH12174</td>
<td>BCDNA:GH12174 does not appear to have a direct counterpart in human cells.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The 13 human ARID proteins are grouped with their closest Drosophila and S. cerevisiae counterparts. Alternate protein and gene names for some proteins are indicated in parentheses. Where the murine gene product has been published under a different name, that name is indicated in column 3.

Number and Diversity of ARID Proteins

The extensive sequence data now available ensure that the predicted protein repertoires of the well-sequenced organisms are largely known. Two ARID-containing proteins have been revealed in budding yeast. The best known is SWI1, a component of the SWI/SNF complex, a multicomponent complex involved in chromatin remodeling and broad aspects of transcription regulation. The other is an open reading frame homologous to the RBP2/PLU-1/SMCX/SMCY subgroup in humans,
discussed below. Six ARID-containing proteins are apparent in the Drosophila genome. Osa is structurally related to SWI1, and associates with the brahma complex, which is the Drosophila equivalent of the SWI/SNF complex (5, 6). Little imaginal discs (lid) was cloned recently in a screen for new trithorax group genes. It is recognizably similar to the second yeast protein and is closely related to human Bdp (7). Dri acts as a coactivator or corepressor at specific transcription sites (8, 9) and has no apparent orthologue in yeast. Its closest mammalian counterpart is Bright and Bdp. The Drosophila BCDNA:GH12174 open reading frame has an ARID sequence close to Dri and Bright, but without the extended ARID sequences. Two other open reading frames containing the ARID consensus are apparent in the Drosophila genome. They are designated CG7274 and CG3654, and outside their ARID regions they are related to the human proteins RBP1 and jumonji, respectively.

The number and forms of ARID proteins broaden further in mammalian cells. Eleven distinct human ARID-containing open reading frames were counted in the Celera human genome sequence (10), but 13 are apparent in public databases. The human ARID-containing proteins vary in size from human Bright (DRIL1) and Bdp, which contain just <600 amino acids, to p270, which contains >2000. p270 is the most direct structural and functional orthologue of SWI1 and Osa, p270 shows ~80% identity with Osa in the well-conserved COOH-terminal region and is stably associated with human SWI/SNF (hSWI/SNF) complexes (11, 12). Four distinct human ARID proteins are very similar to the second ARID protein of yeast. These are RBP2, SMCY, SMCX, and Plu-1, discussed individually below. The most direct human orthologues of Dri are human Bright (DRIL-1) and Bdp (DRIL-2). Human Bright and Bdp are similar in size ($M_r$ 75,000 and $M_r$ 61,000, respectively) and almost iden-

<table>
<thead>
<tr>
<th>Protein</th>
<th>$M_r$ (observed/predicted)</th>
<th>$r$</th>
<th>Chromosome</th>
<th>Tissue distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>p270 (SMARCF1)</td>
<td>270,000/2285 aa</td>
<td>1p36.1-p35</td>
<td>Broad, Northern Blots show similar levels of expression in the full range of tissues tested: spleen, thymus, prostate, testis, ovary, small intestine, colon, peripheral blood lymphocytes, heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas (12).</td>
<td></td>
</tr>
<tr>
<td>KIAA1235</td>
<td>245,000/1711 aa</td>
<td>6q25.1-q25.3</td>
<td>Broad (18).</td>
<td></td>
</tr>
<tr>
<td>RBP1 (RBBP1)</td>
<td>200,000/143,000 (observed/predicted)</td>
<td>14q22.3</td>
<td>Broad with some specialization. RBP1 is expressed in all tissues examined by Northern blot, although the level of expression among different tissues is not constant (results with specific tissues were not reported; Ref. 20).</td>
<td></td>
</tr>
<tr>
<td>RBP1L1 (BCAA)</td>
<td>1311 aa</td>
<td>1q42.1-q43</td>
<td>Restricted. Among normal tissues, RBP1L1 is well expressed only in testis, but high expression was seen in all cancer tissues examined of breast ovary, lung, colon, and pancreatic origin (25).</td>
<td></td>
</tr>
<tr>
<td>RBP2 (RBBP2)</td>
<td>195,000/1722 aa</td>
<td>12p11</td>
<td>Broad with some specialization as indicated for RBP1 (20).</td>
<td></td>
</tr>
<tr>
<td>SMCY</td>
<td>1538 aa</td>
<td>Yq11</td>
<td>Specific to males, but RT-PCR indicates similar levels of expression in the full range of tissues tested: brain, kidney, liver, lung, muscle, spleen, and heart (35).</td>
<td></td>
</tr>
<tr>
<td>SMCX (XE169)</td>
<td>1560 aa</td>
<td>Xp11.22-p11.21</td>
<td>RT-PCR indicates similar levels of expression in the full range of tissues tested: brain, kidney, liver, muscle, spleen, and heart (35).</td>
<td></td>
</tr>
<tr>
<td>PLU-1</td>
<td>1544 aa</td>
<td>1q32.1</td>
<td>Restricted. In normal tissues, Plu-1 is well expressed only in testis, but it is consistently up-regulated in breast cancers (27).</td>
<td></td>
</tr>
<tr>
<td>jumonji (JMJ)</td>
<td>160,000/1266 aa</td>
<td>6p24-p23.</td>
<td>Specialized. Abundant in brain, heart, skeletal muscle, kidney, and thymus but hard to detect in lung, liver, or spleen (29).</td>
<td></td>
</tr>
<tr>
<td>MRF-1</td>
<td>Unknown</td>
<td>2p11.1</td>
<td>Not reported.</td>
<td></td>
</tr>
<tr>
<td>MRF2</td>
<td>83,000/743 aa in mouse</td>
<td>10q11.22</td>
<td>The expression profile of MRF2 is not reported. Expression of Desrt (murine MRF2) is broad with some specialization. A Northern blot shows abundant expression in brain, kidney, and lung; moderate expression in heart, small intestine, and muscle; and no detectable signal in liver, spleen, large intestine, or skin (33).</td>
<td></td>
</tr>
<tr>
<td>Bdp (DRIL2)</td>
<td>61,000/560 aa</td>
<td>15q24</td>
<td>RNA was detected in a broad range of tissues but was abundant in placenta, testis, and leukocytes (32).</td>
<td></td>
</tr>
<tr>
<td>Bright (DRIL1)</td>
<td>75,000 (observed)</td>
<td>19p13.3</td>
<td>Restricted. A ribonuclease protection assay shows message accumulation in mature B cells but not in T cells or immature B cells; in mouse tissues, expression was detected in testis but not in brain, kidney, lung, liver, spleen, or thymus (1).</td>
<td></td>
</tr>
</tbody>
</table>

a The total number of amino acids (aa) of the major form of each of the 13 human ARID-containing proteins is shown here. Where endogenous full-length protein has been observed, the relative migration rate ($M_r$) is indicated.
tical in their ARID sequences. Their sequences are not highly similar outside the ARID region, but their relationship to Dri is apparent in the conservation of an extra sequence of \(~30\) amino acids that extends directly COOH-terminal from the ARID consensus. This extended ARID sequence is \(>75\)% identical in Dri, Bright, and Bdp but does not occur in other ARID-containing proteins. Sequences of five other ARID-containing proteins have been identified in the human genome. Each of the human ARID-containing proteins is introduced here briefly and discussed further under specific topics below.

**Fig. 1.** Schematic representation of the human ARID family proteins. The 13 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 differently colored bars or boxes in the appropriate protein structures and identified at the bottom of the figure. The amino acid (aa) length of each protein is shown at the right of the bar. The length of MRF2 is estimated from the corresponding murine product. The full-length sequence of MRF1 is not yet reported. In cases where alternative splice forms are predicted from Genbank sequences, the most complete form is depicted.

**Fig. 2.** Schematic representation of the Drosophila ARID family proteins. The six Drosophila 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 differently colored bars or boxes in the appropriate protein structures and identified at the bottom of the figure. The amino acid (aa) length of each protein is shown at the right of the corresponding bar.

p270. p270 was first recognized and cloned through its shared antigenic specificity with p300 and CBP (11, 13). Immune complex analysis revealed that p270 is an integral member of human SWI/SNF complexes (11, 12). Recently, independent cloning of a band designated BAF250 in hSWI/SNF complexes reaffirmed that BAF250 is indeed p270 (14). p270 was also cloned independently in a screen for expressed sequences containing trinucleotide repeats, although the cDNA sequence reported by these authors contains a frame-shift that results in a predicted molecular weight of only \(M_r 120,000\) (15,

***KIAA1235***. KIAA1235 was identified in a human fetal brain library in a search for large expressed cDNA sequences (18). The KIAA1235 gene product is very closely related to p270 (>60% identical across its entire sequence), although it is clearly the product of a distinct gene mapping to a different chromosome. Curiously, virtually all known functional motifs, including the ARID sequence, are altered in the KIAA1235 product relative to p270 in ways that suggest the proteins have distinct functions. Analysis with antibodies capable of distinguishing the KIAA1235 protein and p270 indicates that the endogenous KIAA1235 protein migrates at \( M_r \approx 245,000 \) in vivo and does associate with hSWI/SNF complexes.4

**RBP1.** RBP1 was cloned in a search for pRb binding partners soon after pRb was identified as a negative regulator of E2F (19, 20). RBP1 contains the LXCXE motif first identified as a pRb binding motif in DNA tumor virus oncoprotein products such as the adenovirus E1A proteins (reviewed in Ref. 21). RBP1 received relatively little attention until a recent series of reports established that RBP1 acts as a repressor of E2F-dependent transcription and can recruit histone deacetylase activity to pRb/E2F complexes (22–24).

**RBP1L1.** RBP1-like protein 1 was identified through antibodies to an epitope expressed frequently in human carcinomas. Cloning of the epitope-encoding cDNA revealed a protein that is 40–50% identical to RBP1, although RBP1L1 does not contain an LXCXE motif. RBP1L1 expression is tightly restricted to testis in normal tissues, but expression is abundant in many carcinomas (25).

**RBP2.** RBP2 was cloned in the same initial screen as RBP1 and also contains an LXCXE (pRb-binding) motif in addition to the ARID consensus (20). The ARID sequence and the LXCXE motif are shared features, but RBP2 is not otherwise related to RBP1. Rather, RBP2 is closely related across its entire length to the SMC proteins and Plu-1. RBP2, Plu-1, and the SMC proteins also share specific sequence motifs with jumonji.

**SMCY/SMCX.** SCMY was cloned while looking for genes involved in the expression of the minor histocompatibility antigen H-Y. SCMY is encoded on the Y chromosome, and SMCX is the X-chromosome homologue of SCMY. SMCX is one of the few X-chromosome-encoded genes known to escape X inactivation (26). The SMC proteins are closely related to RBP2 but do not contain an LXCXE pRb binding motif.

**Plu-1.** Plu-1 was identified by differentially screening a fetal brain library with cDNAs prepared from a human mammary epithelial cell line overexpressing c-ErbB2 in a probe for genes up-regulated in breast cancer. Plu-1 is closely related to RBP2 and the SMC protein, but like the SMC proteins, Plu-1 does not contain the LXCXE pRb binding motif (27). Most human ARID proteins are rather broadly expressed, but Plu-1 expression in normal adult tissue is tightly restricted to testis. In agreement with the method of its isolation, however, Plu-1 is consistently expressed in breast cancers.

**jumonji.** jumonji was first isolated in a mouse gene trap strategy. In the original study, the mutant jumonji gene was linked with formation of an abnormal cruciform-shaped nucleolar groove (“jumonji” translates as “cruciform” in Japanese; Ref. 28). jumonji has since been described as developmentally important in the liver, spleen, thymus, and heart as well as the nervous system (29). The human jumonji sequence is also available (30). In addition to the ARID consensus, jumonji shows significant homology to RBP2 and the SMC proteins in two regions of about 40 and 127 residues. These regions have been respectively designated jmjN and jmjC in a recent report discussing evolutionary relationships among such jumonji domain-containing proteins (31).

**Bright (DRIL1).** Murine Bright is a B cell-specific transactivator cloned in a search for proteins binding to immunoglobulin heavy-chain matrix-associating regions. Bright and Bdp (see below) are the closest mammalian orthologues of Drosophila dead ringer. In addition to the 94-residue ARID consensus common to the entire family, both Bright and Bdp share with dead ringer a highly conserved sequence of ~30 additional residues COOH-terminally extended from the core ARID consensus. The human gene product DRIL-1 (Dri-like protein-1) is 80% identical to murine Bright. Outside of the ARID and the extended ARID sequence, Bright and Bdp are not closely related to each other or to Dri. However, they are distinguished among ARID proteins by their relatively small sizes, with apparent molecular weights in the range of \( M_r, 60,000–75,000 \).

**Bdp (DRIL-2).** Bdp was cloned from a human testis library as part of a directed search for a potential tumor suppressor gene (32). Bdp does not appear to be the tumor suppressor sought in the study, but the presence of the ARID consensus prompted further characterization of the protein. Bdp is similar to human Bright in its overall structure but not closely related to Bright outside of the ARID and extended ARID sequences. Bdp and Bright have distinct, only partially overlapping, tissue distribution profiles.

**MRF-1.** MRF-1 and MRF2 were cloned by virtue of their ability to bind to similar sequences in the transcriptional modulator of the human cytomegalovirus major immediate-early promoter. Although MRF-1 and MRF2 are more closely related to each other in their ARID sequences than they are to other members of the family, they are not related outside of the ARID domain. The entire sequence and molecular weight are not known for either human protein, although the size and sequence of MRF2 can be approximated closely from its mouse counterpart, as discussed below.

**MRF2.** MRF2 was cloned in the same approach used for MRF-1. The full-length human MRF2 sequence is not yet represented in public databases, but the full-length murine sequence is available as the gene product desert (33). Despite the lack of full-length sequence, MRF2 is one of the best studied of the human ARID proteins in terms of DNA binding activity. A PCR selection and amplification approach shows that the ARID region of MRF2 binds preferentially to an AT-rich core sequence that is similar but not identical to the core sequences recognized by Dri and Bright (34). The three-dimensional structure of the MRF2 ARID region has been

---

solved by solution NMR. It differs from that of Dri in several important aspects, discussed further below.

**Tissue-specific versus Broad Range Expression of Human ARID Proteins**

Expression profiles of human ARID proteins range from broad to very narrow. Northern blots indicate that p270 is well expressed in all 16 tissues probed (12). Reverse transcription-PCR results indicate that the p270-related KIAA1235 gene product is also widely expressed in normal tissues (18). Fattaey *et al.* (20) report that RBP1 is expressed in all tissues examined by Northern blot, although the level of expression among different tissues is not constant (results with specific tissues were not reported). Expression of the related protein RBP1L1 in normal human tissues is sharply restricted and abundant only in testis, although it is also abundant in many types of carcinomas. Among the subgroup comprising RBP2, PLU-1, SMCX, and SMCY, three members (RBP2, SMCX, and SMCY) appear to be widely expressed (20, 35), whereas PLU-1 expression is tightly restricted to testis (27). Plu-1 is consistently expressed in breast cancers, however, as discussed further below. Jumonji expression is variable. It is highly expressed in some tissues, such as brain and heart, but not as apparent in others (29). Bright expression is highly restricted. RNase protection assays show an accumulation of Bright message in mature B cells but not in T cells or immature B cells (1). Bdp RNA was detected by Northern blot in a broad range of tissues but was more abundant in placenta, testis, and leukocytes (32). Expression profiles of MRF-1 and MRF2 have not been reported, but Northern blot analysis of Desrt, the murine counterpart of MRF2, shows variable expression across a wide range of tissues (33). Expression patterns for the mammalian ARID proteins are compiled in Table 2.

**Structure and DNA Binding Activity of ARID Domains**

**Not All ARID Proteins Prefer AT-rich Sites.** Bright, Bdp, and *Drosophila* Dri are >80% identical in their ARID sequences. The DNA binding behaviors of both Bright and Dri have been well characterized. Oligonucleotide selection and amplification shows that murine Bright has a preference for AT-rich sites similar to those found in the matrix attachment regions that served as the probe for isolation of the protein. The selection technique yielded a core hexamer consensus of (A/G)AT(T/A)AA. The selected sequences also consistently showed ATC runs containing AT dimers, features characteristic of matrix attachment region recognition sites. Nucleotide changes affecting any of these features significantly impacted DNA binding (1). Oligonucleotide selection and amplification analysis of Dri yielded a core hexamer consensus almost identical to that selected by Bright: (A/G)ATTAA (2). This is consistent with the consensus engrafted homeodomain binding site sequence (TCATTAATGTA) used to isolate Dri. Bdp is able to bind similar matrix attachment sequences as Bright, although the DNA binding activity of Bdp has not yet been explored further (32). Most likely Bright and Bdp perform similar sequence-specific DNA binding functions in different subsets of tissues.

MRF-1 and MRF2 were isolated as proteins binding to AT-rich target sequences in the CMV major immediate-early promoter. The preference of MRF2 for AT-rich sites has been demonstrated directly by carboxyoxyl interference and in an oligonucleotide selection and amplification assay (34). The consensus preferred binding site is AATA(C/T). The interference assays indicate that MRF2 can distinguish among several similar AT-rich hexamers within the probe, suggesting that other factors in addition to an AT-rich recognition site determine binding specificity. MRF-1 is closely related to MRF2 across the entire ARID region, and unpublished results (cited in Ref. 34) indicate that MRF-1 protects the same sequence as MRF2 in the interference assay.

Exploration of the DNA binding behavior of p270 and Osa has added a new dimension to ARID functions. Oligonucleotide selection and amplification reveals no preference in p270 for AT-rich sequences, and indeed, no identifiable sequence preference at all (12), although in an electromobility shift assay, p270 does appear to interact preferentially with an unusual pyrimidine-rich promoter element in comparison with synthetic oligonucleotides of normal purine/pyrimidine content (14). The general lack of sequence-specific binding in p270 is consistent with a similar finding with Osa in a restriction digest fragment selection assay (6). The behavior of this subset of ARID proteins expands the repertoire of ARID functions, although the physiological role of the nonsequence-specific DNA binding activity in these human and *Drosophila* SWI/SNF complex-associated ARID proteins has not yet been elucidated.

The amino acid sequence of the ARID regions gives no clue to the basis for DNA binding specificity or lack of it. The sequence preferences of Bright and Dri might derive partly from the extended ARID region. However, the activity of MRF2, which does not share the extended ARID sequence, indicates either that the basic ARID consensus is sufficient to specify a preference for AT-rich interactive sites or that nonconserved sequences near the ARID consensus contribute to specificity. The sequence-specific binding activity of Dri and MRF2 is apparent in ARID-containing fragments as small as 152 or 108 amino acid residues, respectively, whereas p270 and Osa fragments as large as 418 or 233 residues, respectively, do not bind DNA specifically.

The DNA binding activity of the remaining ARID family proteins has not been examined in any detail. Thus, we do not really know the full dimensions of the ARID-based DNA binding function. Among the unexplored mammalian proteins, some are highly tissue specific, which may suggest that they, like Bright, are sequence-specific DNA binding proteins. Others, like p270, are widely expressed. A fuller understanding of the family will require a more systematic characterization of individual DNA binding behaviors.

**ARID Regions Contact Both Major and Minor Grooves.** DNA binding proteins generally recognize their target sequences through base-specific contacts in the major groove (36). For a significant minority, however, sequence recognition occurs primarily through minor groove contacts (37). Because the major and minor groove surfaces of the bases present different chemical substituents, the mechanisms of recognition in each case may be fundamentally different.
Various lines of evidence suggest that ARID proteins make both major and minor groove contacts. Several matrix attachment region protein interactions are sensitive to competition by the minor groove-binding antibiotic distamycin A. This sensitivity extends to Bright, indicating that the Bright ARID requires a minor groove interaction to bind its target sequences (1). Distamycin sensitivity suggests that MRF2 also requires a minor groove interaction. However, substitutions in the MRF2-selected core pentamer sequence designed to affect base structure only in the major groove, weaken binding at four of the five positions, suggesting that the nucleotide sequence at most of the core positions is recognized through major groove contacts (34). Required interactions in the minor groove may occur outside the pentamer core.

**ARID Structure.** The ARID is a highly structured \( \alpha \)-helix-based DNA binding domain. Helix-loop-helix and helix-turn-helix motifs each have a two- or three-helix structure in which one helix (the recognition helix) contacts the major groove. Homeodomains have a third helix supporting the alignment of the recognition helix. Other DNA binding protein families that contain helix-turn-helix motifs often have one to three additional conserved helices around the basic motif (reviewed in Ref. 38). A few proteins contact the major groove via \( \beta \)-sheets. Computer algorithms predict that ARID regions consist of a series of at least six \( \alpha \)-helices. Nuclear magnetic resonance solution structures have been obtained for two ARID sequences, Dri (39) and MRF2 (40). The structures are similar but differ in important features. MRF2 has six helices (H1 to H6); Dri has these six and one more on each end (H0 and H7) extending beyond the consensus (Fig. 3). MRF2 has a loop between H1 and H2, whereas Dri has a \( \beta \)-sheet located in the analogous position.

Helices H2-H6 form a similar three-dimensional structure in both domains, although they do not superimpose completely. Both structures predict that DNA contact and sequence recognition is made through H5 and its preceeding turn interacting with the major groove of DNA, whereas other residues contact the minor groove or phosphate backbone. Contact with the minor groove is believed to involve the loop between H1 and H2 of MRF2, or the \( \beta \)-sheet in the analogous position in Dri. The flexible COOH terminus of MRF2 or the equivalent helical structure of H7 in Dri may form additional important contacts with the minor groove or phosphate backbone (41). This is similar in many aspects to the homeo-domain interaction with DNA. The recognition helix of the homeodomain makes sequence-specific contact with the major groove, whereas the adjacent minor grooves are contacted, respectively, by a flexible arm and a loop located between the other two helices of the domain (3, 40). However, homeodomains all recognize the same core motif (ATTA; reviewed in Ref. 42), whereas ARIDs do not. In particular, p270 contains a well-conserved ARID consensus, including the predicted recognition helix, but is largely non-specific in its DNA binding activity (Figs. 3 and 4).

Five residues are absolutely invariable among the human, *Drosophila* and *Saccharomyces cerevisiae* ARID sequences (Fig. 4). The invariable proline (P), tryptophan (W), and tyrosine (Y) residues have ring-structured side chains that presumably contribute rigidity to the structure. Basic or polar residues appear at conserved intervals in the consensus sequence and may be instrumental in DNA contact. A few mutagenesis studies have been initiated on the ARID sequence. Deletion of a seven-amino acid stretch that includes the invariable tryptophan in Dri abrogates DNA binding and acts in a dominant-negative manner to impair the ability of wild-type Dri to rescue the lethal Dri-null phenotype. A truncation that eliminates H0 and part of H1 impairs the ability of Bright to bind DNA in an electromobility shift assay (1). A truncation that eliminates sequences NH2- to the predicted H1-H2 loop in p270 seriously impairs binding of the p270 ARID to native DNA cellulose columns, as does a combined substitution of the invariable tryptophan and tyrosine residues (12).

**Regulation of ARID DNA Binding Activity.** There are indications that the DNA binding activity of ARID proteins may be regulated by other cellular processes. MRF2 was isolated as a repressor of the hCMV enhancer, which is repressed in undifferentiated Tera 2 and THP-1 cells. Reti-
ARID Proteins

ARID Protein Functions

Development. Studies of several ARID family members have revealed their importance during development and gene expression. Homozygous null mutants for three of the Drosophila ARID family members have been generated, and all are lethal at early stages. Dri-deficient embryos are defective in hindgut and muscle development and in embryonic patterning (9). Dri has been identified specifically as a component of a complex required for dorsal-mediated repression (8). Dri binds AT-rich sites in the 5′ region of the Drosophila zen gene, which is called VRR (ventral repression region). Through this specific binding, Dri directs dorsal to this site, and this complex then recruits the repressor, groucho, resulting in ventral repression of zen. Of the other Drosophila ARID proteins, Osa is required for embryonic segmentation and affects patterning of the wing and imaginal eye disc as well as neuronal differentiation (46). Osa antagonizes signaling by wingless, a Wnt family member, during development (6, 46). Osa is a component of the broma chromatin remodeling complex (5) and is linked genetically with E2F-mediated transcriptional regulation (47). lid is a member of the trithorax group of genes and therefore is predicted to help maintain the expression pattern of homeotic genes at the chromatin level during development (7).

jumonji was the first mammalian ARID gene to be examined in a knockout mouse. jumonji homozygous knockouts are embryonic lethal by day E15.5, and the mutant embryos show severe neural tube defects. LacZ expression in Mmj transgenic heterozygotes is strong at specific locations in the brain during embryonic development. Postnatally, expression is seen in Purkinje cells and eventually all granule cells. Outside the neural tube, expression is relatively weak throughout development (28).

Desrt homozygous knockout mice show reduced viability. Approximately 50% of homozygotes die in utero or within a few hours after birth. Survivors are growth retarded at birth and after, attaining ~69% the weight of their wild-type littermates. Desrt heterozygotes have abnormalities in the male and female reproductive organs as well as in the adrenal gland (33). Whole mount in situ hybridization shows expression of Desrt in the limb bud and interdigital tissue, suggesting a potential role during limb patterning. Desrt is also expressed transiently during development in several distinct locales, such as the otic vesicles, endolymphatic diverticulum, auditory meatus, premigratory neural crest, liver denticulum, lung buds, and lining of the oral cavity, suggesting a role in organogenesis (48).

Expression of Osa1, the mouse orthologue of p270, has been examined during mouse development by in situ hybridization. Osa1 is ubiquitous in early development but in time becomes more restricted to the limb buds, eye lens, neural tube, and brain (17). However, adult tissue Northern blots indicate that p270 is expressed similarly in all tissues examined (see Table 2).

Bright is primarily expressed in B lymphocytes in adults. Its expression appears to be regulated during fetal development, where it is expressed in pre-B cells and activated mature B lymphocytes. Bright can be detected in the fetal liver, thymus, and brain by reverse transcription-PCR at day

Fig. 4. ARID sequence alignments. The amino acid sequences of the ARID regions of the 13 human ARID-containing proteins were aligned using the Clustal W 1.8 multiple sequence alignment program (65). The black boxes indicate residues identical in at least 7 of the 13 proteins. Gray shading indicates positions where at least seven residues are closely related but not identical. Five residues are invariable; these are indicated by underlining. The consensus sequence extends across 94 residues, of which 39 are highly conserved with regard to both identity and spacing. (The consensus is defined here as identity at a specific position in at least 7 of the 13 human proteins). The Bright (DRIL1/Bdp (DRIL2) homology (which is also a feature of Drosophila Dri) extends for 35–40 residues past the ARID consensus and appears to be characteristic of one subfamily within the ARID family. Accession numbers for the human gene products used in the alignment are given in parentheses: p270 (AF265208), KIAA1235 (BAA86549), DRIL1 (NP_005215), DRIL2 (NP_006456), MRF-1 (M62324), MRF2 (M73837), RBPI (P29374), RBPI (NP 057458), Jumonji (Q92833), SMCX (L25270), SMCY (NP_004644), RBP2 (S66431), and PLU-1 (CAB43532).
16 of gestation (49). Expression of other mammalian ARID family members has not been examined during development, but their adult tissue distribution patterns are discussed in “Tissue-specific versus Broad Range Expression of Human ARID Proteins” and Table 2. Overall, ARID proteins appear to be widely expressed during development and in some cases are crucial to survival.

**Gene Expression.** p270 is a component of human SWI/SNF complexes (11). These are ATP-dependent chromatin remodeling complexes first described in yeast. p270 is likely to be involved in recruiting SWI/SNF to specific promoters through interactions with nuclear hormone receptors via its LXXLL motifs. p270 has been shown to bind the glucocorticoid receptor and to activate transcription from a reporter plasmid with glucocorticoid response elements in a hormone-dependent manner (14).

RBP1 associates with the pocket region of pRb and can repress transcription from E2F-dependent promoters. Two repression domains (R1 and R2) have been mapped within RBP1. R2 is COOH-terminal and can associate with the mSin3-HDAC complex (24). Rpb1 recruits this complex to the pRb pocket and can repress E2F-mediated transcription in an HDAC-dependent manner. R1 maps to a region that includes the ARID domain and represses transcription in an HDAC-independent manner (23).

RBP2, like p270, is part of a subset of ARID proteins that contains LXXLL motifs and are therefore predicted to bind nuclear hormone receptors. RBP2 has been found recently to bind the glucocorticoid, estrogen, vitamin D, and retinoic acid receptors in vitro and to bind the estrogen receptor in vivo. Overexpression of RBP2 can enhance transcription induced by each of these hormones in reporter assays. Addition of pRb further enhanced estrogen-induced transactivation in this assay (50).

SMCX is located on the Y chromosome. Y chromosome genes are generally needed only for male-restricted functions and are expressed only in testes. SMCY, however, is expressed ubiquitously (35). SMCX is located on the X chromosome. X-inactivation in females is believed to be the mechanism by which gene expression from the X-chromosome is equalized between male and female. However, SMCX is one of the few X-chromosome genes known to escape X-inactivation (26). This suggests that the SMCX and SMCY genes are functional homologues that are largely interchangeable, and that the dose of transcript is important for the function of these genes (51). Although the biological roles of these proteins is not yet clear, SMCY does encode several H-Y antigen epitopes (51).

PLU-1 is the fourth member of the subset of closely related ARID proteins that includes RBP2, SMXC, and SMCY. Little is known about the biological role of PLU-1, but its expression pattern is intriguing. It is tightly restricted to testis in normal tissue blots but up-regulated with high frequency in breast cancer lines (27), as discussed further in “ARID Proteins and Human Tumorigenesis.”

Data are beginning to emerge about the biological role of Jumonji. It may be involved in negative regulation of cell growth. Overexpression of jmj in COS and NIH3T3 cells results in decreased cell proliferation (52). Likewise, megakaryocyte progenitor cells from jmj --/-- mice show increased cell proliferation but not increased differentiation (53).

MRF2 is a repressor of the hCMV enhancer. It binds the enhancer, which is repressed in undifferentiated Tera 2 and THP-1 cells. Binding activity is markedly reduced in differentiated cells, and enhancer activity is restored (43). Presumably, MRF2 plays a role in the repression of inappropriate differentiation-specific gene expression, but such an activity has not yet been demonstrated directly.

The role of Bright in the regulation of immunoglobulin heavy chain expression has been reviewed recently (45). Although Bright is known to increase immunoglobulin transcription 3–7-fold in antigen-activated B cells, the mechanism of activation is not clear. Bright binds to MARs in the intronic enhancer region of the immunoglobulin heavy chain gene as a tetramer. Bright is able to form bends in DNA of 80–90 degrees and may be able to facilitate long-range interactions in the enhancer region. Bright associates with specific nuclear matrix proteins and may affect chromatin configuration and nuclear sublocalization.

**Other Functional Motifs and Domains in ARID Proteins.** ARID proteins can be divided into several subgroups based on the presence of additional structural features in the proteins. For the smaller ARID proteins, the ARID consensus is the dominant feature of the protein. Many of the larger ARID proteins have a more complex array of recognized protein domains (see Fig. 1). Jumonji contains two conserved regions that have been recognized in an overlapping family of proteins. These domains are designated JmjN and JmjC, denoting their relative positions within the jumonji protein. The combination of JmjN and the ARID domain occurs in Drosophila and more distantly related proteins as well. JmjN and JmjC domains occur together in the four ARID proteins of the RBP2/PLU-1/SMCX/SMCY group, and in a few other human proteins linked with transcription, but without ARID domains (31). JmjC domains may occur alone in a much wider group of proteins (54). The functions of the Jmj domains are not yet known, although these authors predict on the basis of certain structural similarities, that JmjC domains may be enzymatically active domains related to cupin metalloenzyme domains. The RBP2/PLU-1/SMCX/SMCY group is linked further by the presence of multiple PHD-type zinc-finger domains. PHD domains are found in >60 human proteins (10).

The RBP2/PLU-1/SMCX/SMCY group also contains LXXLL motifs, which generally serve as binding sites for liganded nuclear hormone receptors (55, 56). These motifs are present in p270 as well. MRF2 and the KIAA1235 protein each have one such motif. The significance of the single motifs is not clear, because functional LXXLL motifs usually occur as multiples. However, MRF2 binding to the human cytomegalovirus enhancer in Tera 2 cells is dependent on whether the cells are treated with retinoic acid (43), implying that the DNA binding activity of MRF2 may be regulated by its binding to the retinoic acid receptor. Binding to nuclear hormone receptors has been demonstrated directly in p270, which binds the glucocorticoid receptor in vitro and in vivo (14), and in RBP2, which binds several receptors in vitro and the estrogen receptor in vivo (50). RBP2 and p270 can both increase hormone-responsive activation of reporter plasmids, as discussed above.
RBP1 and RBP1L1 are related across their entire length and share a Tudor domain near each NH₂ terminus. Nine Tudor domain-containing proteins have been identified in the human genome (10). The Tudor domain is found in many proteins that colocalize with ribonucleoprotein or single-strand DNA-associated complexes in the nucleus, the mitochondrial membrane, or at kinetochores. One of these is the SMN gene, defects in which cause spinal muscular atrophy. The Tudor domain mediates binding of the SMN protein to spliceosomal core proteins (57).

Although generally similar, RBP1 and RBP1L1 differ in important aspects. Unique motifs in RBP1 include a partial (69% complete) chromodomain and an LXCXE motif, which specify pRb binding. An LXCXE motif is, however, present in RBP2, which was isolated as a pRb binding protein. Bdp may also contain a pRb binding activity. Bdp does not have an LXCXE motif, and the pRb binding activity observed was impaired by substitution of the invariable proline or tryptophan in the ARID sequence, implying it is dependent on the integrity of the ARID structure (32). This interaction has not been demonstrated in vivo.

ARID Proteins and Human Tumorigenesis

 Emerging data indicate that aberrant expression of ARID proteins is a fairly common feature of mammalian tumor cells. Plu-1 was cloned directly as a product that is specifically up-regulated in breast tumor cells (27). Plu-1 is well expressed in at least five of four common breast cancer cell lines examined but poorly expressed in at least six of eight colon cancer cell lines (27). RBP1L1 was also cloned directly as a tumor antigen. Similar to Plu-1, its expression in normal human tissue is abundant only in testis, but RBP1L1 expression is abundant in all types of carcinomas screened: breast, ovary, lung, colon, and pancreatic. A link between high expression in human cancers and in normal testis has been noted before and is discussed in Cao et al. (25).

Components of human SWI/SNF complexes are frequently lost or altered in tumor cells (58, 59), and this pattern is now known to include the ARID-containing p270 as well. Reduced expression of p270 has been observed in 3 of 21 common breast cancer lines screened as well as in C33A cervical carcinoma cells (14, 60). Another example of ARID loss in tumorigenesis is the SMCY gene, which is lost with high frequency in prostate tumor samples (61).

Two different ARID-containing proteins (RBP1 and RBP2) were cloned through association with pRb. Although their expression has not been screened in tumor cells, they clearly have the potential to contribute to control of cell proliferation. Bdp also shows some evidence of a pRb binding function (32). DRIL1 binds to the E2F transcription factor (62) and has very recently been shown to rescue Ras-induced senescence in primary murine fibroblasts and cause them to become oncogenic (63). No link has yet been made between tumorigenesis and the PLU-1 subfamily protein SMCX or the remaining ARID proteins, jumonji, MRF-1, and MRF2.

Overview

The ARID region appears as a single-copy motif in 13 human proteins, ranging in size from human Bright (DRIL1) and Bdp, which contain just <600 amino acids, to p270, which contains >2000. Most of the human ARID proteins can be viewed as a series of pairs or a group of four, with distinct but clearly related sequences and/or additional shared motifs that distinguish them from other members of the ARID family. The ARID consensus spans ~100 residues, of which 30–40 are highly conserved in proteins from the full range of eukaryotic species. The domain has a more complex structure than most other α-helix-based DNA binding domains. The two ARID structures thus far available predict that major groove contact is made through a specific α-helix in a core structure similar to the helix-turn-helix motif in homeodomains. However, the ARID sequence can form at least six α-helices (as in the case of MRF2) and as many as eight as seen in Dri. Moreover, Dri also contains a two-stranded β-sheet, which in this case appears to contact the minor groove. p270 and Drosophila Osa each contain a well-conserved ARID consensus, including the predicted recognition helix, but both appear to be largely nonspecific in their DNA binding activity.

Although the ARID family is smaller than most other families of DNA binding proteins, it nevertheless encompasses both ubiquitously expressed members and members whose expression is highly restricted. Gene regulation activities among the ARID proteins also cover a broad range. For example, Bright plays a particular role at matrix attachment regions, RBP1 is part of a deacetylase-associated repressor complex, and p270 is a component of nucleosome remodeling complexes. The ARID sequence has apparently been adopted by a range of proteins with diverse DNA binding needs, but this sense of broad adaptation contrasts with the highly conserved sequence of the motif and with the conservation of the full range of ARID protein structure and function from fly to human. Continued analysis of ARID proteins should shed new light on the nature of DNA protein interactions.

Acknowledgments

We thank Yuan Chen, Robert Whitson, and Takahiro Nagase for communication of unpublished results.

References