66–80 *Nucleic Acids Research, 2005, Vol. 33, No. 1*
doi:10.1093/nar/gki145

DNA-binding properties of ARID family proteins

Antonia Patsialou, Deborah Wilsker and Elizabeth Moran*

Fels Institute for Cancer Research and Molecular Biology, Temple University School of Medicine, Philadelphia PA 19140, USA

Received July 30, 2004; Revised October 28, 2004; Accepted December 3, 2004

ABSTRACT

The ARID (*A*–T *R*ich *I*nteraction *D*omain) is a helix–turn–helix motif-based DNA-binding domain, conserved in all eukaryotes and diagnostic of a family that includes 15 distinct human proteins with important roles in development, tissue-specific gene expression and proliferation control. The 15 human ARID family proteins can be divided into seven subfamilies based on the degree of sequence identity between individual members. Most ARID family members have not been characterized with respect to their DNA-binding behavior, but it is already apparent that not all ARIDs conform to the pattern of binding AT-rich sequences. To understand better the divergent characteristics of the ARID proteins, we undertook a survey of DNA-binding properties across the entire ARID family. The results indicate that the majority of ARID subfamilies (i.e. five out of seven) bind DNA without obvious sequence preference. DNA-binding affinity also varies somewhat between subfamilies. Site-specific mutagenesis does not support suggestions made from structure analysis that specific amino acids in Loop 2 or Helix 5 are the main determinants of sequence specificity. Most probably, this is determined by multiple interacting differences across the entire ARID structure.

INTRODUCTION

The ARID (*A*–T *R*ich *I*nteraction *D*omain) is a helix–turn–helix motif-based DNA-binding domain, conserved in all eukaryotes and diagnostic of a family that comprises 15 distinct human proteins. ARID proteins, although diverse in function, all appear to play important roles in development, tissue-specific gene expression and cell growth regulation [reviewed in (1,2)]. The ARID consensus sequence, which spans about 100 residues, was first identified as a DNA-binding domain in the mouse B cell-specific transcription factor, Bright (3), and in the Dead ringer protein (Dri) of *Drosophila melanogaster* (4). Dri and Bright were each isolated in searches designed to detect proteins binding selectively to AT-rich sequences. Recognition of the Bright/Dri consensus defined the parameters of a new DNA-binding domain, and the properties of Bright and Dri inspired its name. MRF-1 and MRF2, which bind the CMV enhancer and repress its activity, are also ARID-containing proteins that bind selectively to AT-rich sites (5,6).

Although the first studied ARID-containing proteins bind preferentially to AT-rich sites, this behavior does not appear to be an intrinsic feature of the domain. Most ARID family members have not been characterized with respect to their DNA-binding behavior, but it has become apparent that not all ARIDs conform to the pattern of binding AT-rich sequences. p270 is a human ARID-containing protein that is an integral member of SWI/SNF-related chromatin remodeling complexes (7–10). p270 contains a complete ARID consensus and binds DNA with an affinity similar to Dri, but is unable to select oligonucleotides of any preferred sequence from a random pool (9,11). Lack of sequence specificity has been shown independently for the ARID family member, Osa, the closest *Drosophila* counterpart of p270 (12).

The ARID structures for MRF2, Dri, p270 and its yeast counterpart SWI1 have been studied by NMR (13–16). Despite the high degree of conservation in the domain, at least three distinct structural patterns are recognized: MRF2 and SWI1 both have six helices and two loops. Dri has one more helix on each end formed by sequences outside the consensus and a β-sheet instead of a flexible loop between Helix 1 and Helix 2. p270 has an additional short N-terminal helix, but no C-terminal helix or any β-sheets. The structures of the MRF2, Dri and p270 ARIDs have also been solved in complex with DNA (15,17,18). All studies agree that the ARID binds DNA via both the major and the minor grooves, and that major groove contacts are made through residues in Loop 2 and/or Helix 5.

The human ARID family can be divided into seven subfamilies based on the degree of sequence identity between individual members (Figures 1 and 2). The diverse

*To whom correspondence should be addressed. Tel: +215 707 7313; Fax: +215 707 6989; Email: betty@temple.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

```
                  H0            H1        Loop1 or β-sheet      H2        H3       H4     Loop2       H5          H6          H7
ARID1A  SSSSTTTNEKITKLYELGGEPERKMWVDRYLAFTEEKAMGMTNLPAVGRKPLDLYRLYVSVKEIGGLTQVNKNKKWRELATNLNVG-TSSSAASSLKKQYIQCLYAFECKIER
ARID1B  SSSSTTTGEKITKVYELGNEPERKLWVDRYLTFMEERGSPVSSLPAVGKKPLDLFRLYVCVKEIGGLAQVNKNKKWRELATNLNVG-TSSSAASSLKKQYIQYLFAFECKIER

ARID2           MANSTGKAPPDERRKGLAFLDELRQFHHSRGSPFKKIPAVGGKELDLHGLYTRVTTLGGFAKVSEKNQWGEIVEEFNFPRSCSNAAFALKQYYLRYLEKYEKVHHF

ARID3A  DWTYEEQFK---QLYELDGDPKRKEFLDDLFSFMQKRGTPVNRIPIMAKQVLDLFMLYVLVTEKGGLVEVINKKLWREITKGLNLPTSITSAAFTLRTQYMKYLYPYECEKRGLSNPNELQAAIDSNRREGRRQ
ARID3B  GGRGREISRDFAKLYELDGDPERKEFLDDLFVFMQKRGTPINRIPIMAKQILDLYMLYKLVTEKGGLVEIINKKIWREITKGLNLPTSITSAAFTLRTQYMKYLYAYECEKKALSSPAELQAAIDGNRREGRRP
ARID3C        MTPILPQLYELDADPKRKEFLDDLFSFMQKRGTPVNRVPIMAKQVLDLYALFRLVTAKGGLVEVINRKVWREVTRGLSLPTTITSAAFTLRTQYMKYLYPYECETRALSSPGELQAAIDSNRREGRRQ

ARID4A  EKEKEAKKTEEEVPEEELDPEERDNFLQQLYKFMEDRGTPINKPPVLGYKDLNLFKLFRLVYHQGGCDNIDSGAVWKQIYMDLGIPILNSAASYNLKTAYRKYLYGFEEYCRS
ARID4B  DEKEKEDNSSEEEEEIEPFPEERENFLQQLYKFMEDRGTPINKRPVLGYRNLNLFKLFRLVHKLGGFDNIESGAVWKQVYQDLGIPVLNSAAGYNVKCAYKKYLYGFEEYCRS

ARID5A  PISLEDSPEAGGEREEEQEREEEQAFLVSLYKFMKERHTPIERVPHLGFKQINLWKIYKAVEKLGAYELVTGRRLWKNVYDELGGSPGSTSAATCTRRHYERLVLPYVRHLKG
ARID5B  KVSNEEKPKVAIG---EECRADEQAFLVALYKYMKERKTPIERIPYLGFKQINLWTMFQAAQKLGGYETITARRQWKHIYDELGGNPGSTSAATCTRRHYERLILPYERFIK

JARID1A VKSFRFTPRVQRLNELEAMTRVRLDFLDQLAKFWELQGS-TLKIPVVERKILDLYALSKIVASKGGFEMVTKEKKWSKVGSRLGYLPGK-GTGSLLKSHYERILYPYELFQSG
JARID1B VDKLHFTPRIQRLNELEAQTRVKLNFLDQIAKYWELQGS-TLKIPHVERKILDLFQLNKLVAEEGGFAVVCKDRKWTKIATKMGFAPGK-AVGSHIRGHYERILNPYNLFLSG
JARID1C VDNFRFTPRIQRLNELEAQTRVKLNYLDQIAKFWEIQGS-SLKIPNVERRILDLYSLSKIVVEEGGYEAICKDRRWARVAQRLNYPPGK-NIGSLLRSHYERIVYPYEMYQSG
JARID1D VDNFRFTPRVQRLNELEAQTRVKLNYLDQIAKFWEIQGS-SLKIPNVERKILDLYSLSKIVIEEGGYEAICKDRRWARVAQRLHYPPGK-NIGSLLRSHYERIIYPYEMFQSG

JARID2  NDEMRFVTQIQHIHKLGRRWGPNVQRLACIKKHLKSQGITMDELPLIGGCELDLACFFRLINEMGGMQQVTELKKWNKLSDMLRIPKTAQERLAKLQEAYCQYILSYDSLSPE

Dri     GWSFEEQFKQVRQLYEINDDPKRKEFLDDLFSFMQKRGTPINRLPIMAKSVLDLYELYNLVIARGGLVDVINKKLWQEIIKGLHLPSSITSAAFTLRTQYMKYLYPYECEKKNLSTPAELQAAIDGNRREGRRS

SWI1                    NNKQYELFMKSLIENCKKRNMPLQSIPEIGNRKINLFYLYMLVQKFGGADQVTRTQQWSMVAQRLQI-----SDYQQLESIYFRILLPYERHMIS
```

Figure 1. Alignment of human ARID sequences. The amino acid sequences of the ARID region of the 15 human ARID family members are shown. The shading indicates the boundaries of the α-helices where the structural data are known (13,15,17). The sequences of the ARIDs of *D.melanogaster* Dri, which has an ARID3-class sequence and *S.cerevisieae* SWI1, are shown as well because structural data is also available for them (14,16,18). Helices are labeled at the top from H0 to H7. The location of Loops 1 and 2 and the β-sheet (which so far is found only in the ARID3-class sequence) are also shown. The five invariant residues of the ARID region are shown in red. Part of the 'extended ARID' that characterizes the ARID3 subfamily is shown to indicate the degree of homology in this region. The ARID2 and ARID3C sequences begin at the initial methionine of the protein. The sequences were aligned using the Clustal W 1.8 multiple sequence alignment program (50). The computer-generated alignment was modified slightly to reflect higher level structural data.

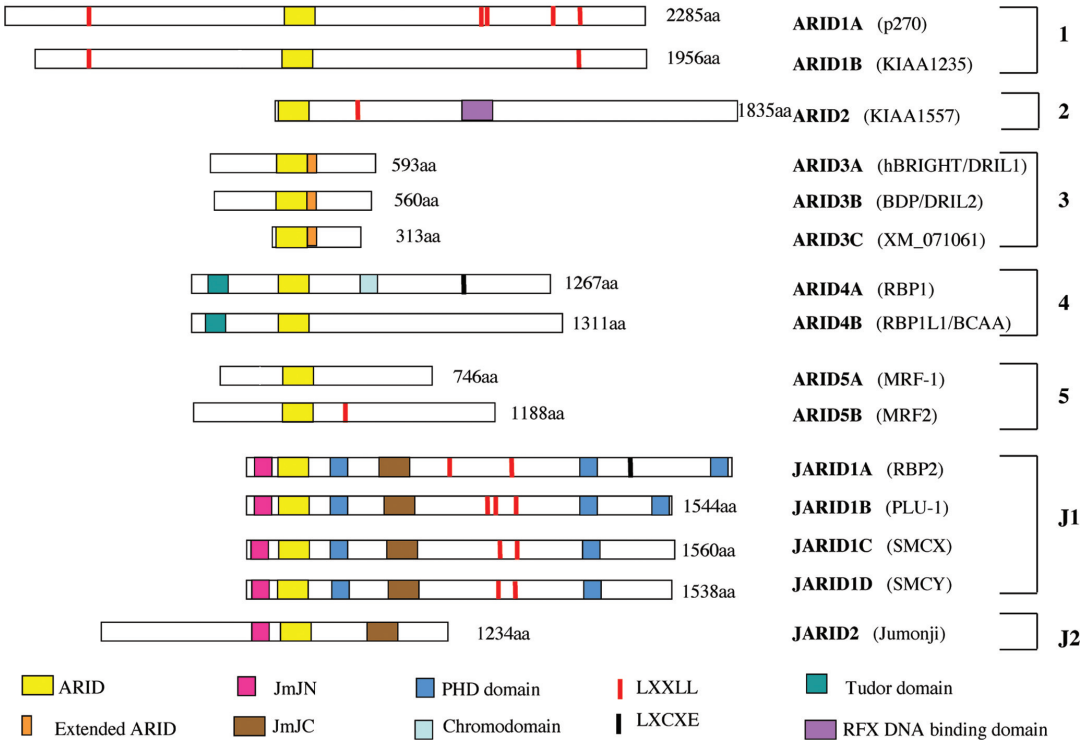

Figure 2. The human ARID family of proteins. Genome sequencing reveals 15 ARID-containing proteins in humans. The ARID family proteins can be grouped into subfamilies based on their similarity to each other within the ARID domain. The nomenclature described here reflects this subclassification of the 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 subgroups of the ARID family, members typically have 70–85% identity within their ARID sequences, while across subgroups identity within the ARID sequence drops to ~25–30%. 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 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 to the right of the bar. The presence of additional motifs was identified through the Pfam database (51).

characteristics of the ARID proteins studied so far prompted a survey of DNA-binding properties across the entire ARID family. The results indicate that the majority of ARID subfamilies (i.e. five out of seven) bind DNA without obvious sequence preference. DNA-binding affinity also varies somewhat between subfamilies. Site-specific mutagenesis does not support suggestions made from structure analysis that specific amino acids in Loop 2 or Helix 5 are the main

determinants of sequence specificity. Most probably, this is determined by multiple interacting differences across the entire ARID structure.

MATERIALS AND METHODS

Plasmids

GST fusion constructs. The p270 fusion protein experiments were originally performed with the product of plasmid pNDX (9). For the mutagenesis studies, a shorter expression construct designated pNDB8 was generated to be more comparable in size with the Dri fusion peptide. The NDB8 plasmid expresses amino acids 958–1188 of p270 (numbering is according to the sequence at accession number NP_006006). The Dri fusion protein is the product of p410 (4), which was kindly provided by R. Saint (University of Adelaide, Australia) and expresses Dri residues 258–410 (according to accession number AAB05771).

A series of plasmids was assembled expressing GST-fusion proteins containing the ARID regions of representative members of ARID subfamilies. The MRF2 fusion protein is the product of pMRF2-GST, which was constructed by ligating a BamHI/SalI restriction fragment from the insert of plasmid MRF2pQE30 [(13); kindly provided by Yuan Chen at the Beckman Institute, City of Hope, Duarte, CA] into the pGEX4T vector (Pharmacia Biotech). A construct containing the ARID domain of human RBP1, called GST–ARID (19), was provided Dr Philip Branton (McGill University, Montreal, Canada). The ARID2 sequence was generated by RT–PCR from HepG2 cells using oligonucleotides ARID2-F (5′-ATAATGGCAAACTCGACGGGGAAG) and ARID2-R (5′-CACCCCGGCATTAGCAAGTAGTAA) to yield a 630 bp fragment that encodes amino acids 1–209 according to accession number XP_350876. The fragment was cloned into pCR2.1-TOPO vector (Invitrogen) to make ARID2–TOPO. The EcoRI fragment of ARID2–TOPO was sub-cloned into the EcoRI site of pGEX-4T1 (Pharmacia Biotech) to make pARID2-pGEX. The PLU-1 sequence was generated by RT–PCR from MCF-7 cells using oligonucleotides PLU-1 For (5′-TTCGCGGACCCCTTCGCTTTCA) and PLU-1 Rev short (5′-AATATTCATGGCCTCTGCTCTC). The reaction generated a 597 bp fragment extending from nucleotide 213 to 810 (according to accession number AJ132440.1), which was cloned into the pCR2.1-TOPO vector to create pPLU-1-TOPO. A PLU-1 containing BstXI restriction fragment was released from the vector, blunted with T4 DNA polymerase, and ligated with SmaI digested pGEX-4T1 to generate pPLU-1-GST. pPLU-1-GST generates a GST-fusion protein containing amino acids 42–241 of PLU-1 according to accession number CAB43532. An RBP2 sequence-containing PCR fragment was generated with primers RBP2-F-Xho (5′-AGACTCGAGTTCACAGATCCGCTCAGCTTTATC) and RBP2-R-Xho (5′-AGACTCGAGTTTAGGACACCTCCAGTCTCCTTT) from the plasmid template pCMV-HA-RBP2 (provided by Philip Branton), and cloned into the pCR2.1-TOPO vector to create pRBP2-TOPO. An XhoI restriction fragment from the RBP2-TOPO insert was blunted with Klenow polymerase and ligated with SmaI-digested pGEX-4T3 to create the plasmid pRBP2-pGEX. This construct produces a GST-fusion protein containing RBP2 amino acids 29–339 (accession number NP_005047). The jumonji fragment was amplified by PCR from a murine brain cDNA library in a vector backbone of pACT-2 (Clontech) that was kindly provided by Dr Premkumar Reddy (Fels Institute, Temple University School of Medicine, Philadelphia, PA). jumonji-TOPO was generated by PCR using the primers jumonji For (5′-AGAGAATTCTGTGAAAATCGTTCTACCTCGCAA) and jumonji Rev (5′-AGACTCGAGATGACAGTCCTTCTCTT-CCACTAA) to generate a 1030 bp fragment extending from nucleotide 1750 to 2780 according to accession number D31967. The PCR fragment was cloned into the pCR2.1-TOPO vector to create pjumonji-TOPO, excised from the vector with EcoRI and XhoI, and cloned into EcoRI/XhoI-digested pGEX-4T1 to create pjumonji-pGEX-4T1. This construct creates a GST-fusion protein containing amino acids 519–858 of jumonji (accession number NP_068678). This is the only case where a murine sequence was used in the subfamily constructs, but the mouse and human proteins are 92% identical across the coding span of the insert.

In vitro translation constructs. The p270 pNE9-B2 *in vitro* expression plasmid and the Dri *in vitro* expression plasmid pDriT2 were described previously (11).

Generation of amino acid substitution mutations

All mutations were generated using the QuikChange (Stratagene) system according to the manufacturer's instructions. The forward primers used to generate amino acid substitutions in pNE9-B2 or pDriT2 are as follows (the substituted bases are underlined):

p270.P1042A: GCATGACAAATCTGGCTGCTGTGGG-TAGGAAACC
p270.W1073A: GGTCAACAAGAACAAAAAAGCGCG-GGAACTTGCAACC
p270.Y1096A: CCTTGAAAAAGCAGGCTATCCAGTG-CTCTATGC
Dri.P306A: CCGATCAATCGGCTGGCGATAATGGCC-AAATCGG
Dri.W337A: CAACAAGAAGCTGGCGCAGGAGATCA-TCAAGGGGC
Dri.Y361A: CCCTGCGCACCCAAGCCATGAAGTATC-TGTACCCG

The remaining mutations were generated in p410 or pNDB8. The forward primers used to generate amino acids substitutions are as follows (the substituted bases are underlined):

Dri.SSS: GCCCTCCAGCATCTCCAGTGCCGCCTCCT-CCCTGCGCACCC
p270.TFT: GTGGGCACATCAACCAGTGCTGCCTTCA-CCTTGAAAAAGCAG

Deletion mutants were generated by a loop-out technique using a primer designed to form a junction between residues at the borders of the deletion. The sequence of the forward primers used to generate the deletions is as follows (the nucleotides that mark the boundaries of the loop-out are underlined):

Dri.ΔC: GAATCTGAGCACGCAGATGCCGATGACG
p270.ΔN: GGGACACCCAAGACAGAAATCACCAAGT-TGTATGAGCTG

Chimera mutants were made by first looping-out the sequence to be replaced, and then looping-in the desired sequence. The sequence of the forward primers used to generate the deletions is as follows (the nucleotides that mark the boundaries of the loop-out are underlined):

L2.H5.out: CGGGAACTTGCAACCAACCTCTTGAAA-AAGCAGTATATCCAG
H4.out: GGATTGACTCAGGTCAACAAGAACAAACT-CCACCTGCCCTCCAGC

The sequence of the forward primers used to generate the insertions is as follows (the nucleotides that form the loop-in are underlined):

L2.H5.in: GCAACCAACCTCCACCTGCCCTCCAGCA-TCACCAGTGCCGCCTTCACCTTGAAAAAGCAG
H4.in: GTCAACAAGAACAAACTGTGGCAGGAGAT-CATCAAGGGGCTCCACCTGCC

The sequence changes and the integrity of the surrounding sequences for all mutants were verified by DNA sequencing.

Sequence-specific selection of DNA

GST-fusion proteins were used in pull-down assays with a pool of Lambda DNA restriction fragments. The assay was performed as described previously (11,12). Restriction fragments were filled in with [α-^{32}P]dATP. Labeled DNA (0.8 μg) was incubated with 50 ng of GST-fusion protein bound to glutathione–agarose beads for 1 h at 4°C in Lambda DNA-binding buffer [20 mM HEPES (pH 7.6), 1 mM EDTA (pH 8), 10 mM $(NH_4)_2SO_4$, 0.2% Tween-20, 1 mM DTT, 25 μg/ml BSA and 25 μg/ml poly(dI–dC)] plus varying amounts of KCl, as indicated in the text. The beads were washed three times with Lambda DNA-binding buffer minus DTT, BSA and poly(dI–dC). Bound DNA was eluted by boiling in Formamide loading buffer (90% formamide, 1× TBE, 0.04% bromophenol blue and 0.04% xylene cyanol), separated on a 6% sequencing gel and visualized by autoradiography.

For the oligonucleotide competition assays, 10 ng of ^{32}P-end-labeled double-stranded oligonucleotide was incubated with 100 ng of GST-fusion protein bound to glutathione beads in the Lambda DNA-binding buffer containing 50 mM KCl, 100 μg of salmon sperm DNA and varying amounts of unlabeled double-stranded competitor oligonucleotide, as indicated in the text. The beads were washed and the bound DNA was eluted and visualized as described above.

In vitro translation and DNA cellulose chromatography

The wild-type and mutant plasmid constructs were used to generate ^{35}S-methionine-labeled polypeptides using the TNT-coupled reticulocyte system (Promega). *In vitro* translated proteins were diluted in one bed volume (0.5 ml) of Column loading buffer [10 mM potassium phosphate (pH 6.2), 0.5% NP40, 10% glycerol, 1 mM DTT, aprotinin (1 mg/ml), pepstatin (1 mg/ml), leupeptin (1 mg/ml)], and applied to native DNA cellulose columns (Pharmacia). The protein sample was passed through the column twice. The columns used were Poly-Prep Chromatography Columns (Bio Rad catalog number 731-1550). Unbound material is designated flow-through (FT). The columns were then washed multiple times with 1.0 bed volume column-loading buffer containing 50 mM NaCl (these are the 50 mM wash fractions), and eluted stepwise with column-loading buffer adjusted to contain increasing concentrations of NaCl from 100 to 800 mM, as indicated in the text. Fractions were analyzed by SDS–PAGE. The signal on the dried gel was quantified using a phosphorimager (Fuji) and associated software.

RESULTS

ARID subfamilies vary in sequence specificity and DNA-binding affinity

Human and mouse ARID-containing proteins can be classified into seven subfamilies: ARID1, ARID2, ARID3, ARID4, ARID5, JARID1 and JARID2. Within each designated subfamily, the degree of identity within the ARID regions is very high, ranging from 70 to 83% (Figure 1). In contrast, identity between ARID regions across subfamilies is <30%. Members within subfamilies generally also show clear relationships outside the ARID, as shown in Figure 2. These subclassifications are the basis for the current nomenclature of the ARID family, which has recently been accepted by the HUGO Gene Nomenclature Committee (HGNC) and the Mouse Genomic Nomenclature Committee (MGNC).

The DNA-binding properties of only a few ARID family members have been reported. *Drosophila* Dri and its murine ortholog Bright (ARID3A), as well as human MRF2 (ARID5B) all bind AT-rich sites selectively (3,4,6). However, human p270 (ARID1A), the closely related human protein ARID1B, and their apparent *Drosophila* and yeast counterparts, Osa and SWI1, all bind DNA without sequence specificity (9,11,12,20). A better understanding of the biological role of the ARID family will require a more thorough understanding of the distribution of sequence-specific DNA-binding properties among the individual members. We therefore undertook a survey designed to determine the DNA-binding properties of at least one member of each ARID subfamily.

Because amino acid identity within the ARID consensus is so high within subfamilies, originally a single member of each subfamily was selected to test for sequence specificity. Recombinant GST-fusion proteins were constructed using sequences that include the ARID domain of each protein examined. The sequence specificity of each protein was then examined in a DNA pull-down assay. This assay allows each protein access to a pool of Lambda DNA restriction fragments of varying size and sequence. As shown in Figure 3, Dri (the *Drosophila* counterpart of ARID3A) and MRF2 (ARID5B) bind in a sequence-specific manner in this assay, selectively binding to some fragments and not others. Selectivity for specific fragments becomes more pronounced in more stringent binding conditions (i.e. increased salt concentrations). Slight differences in the selected fragments between Dri and MRF2 probably reflect the fact that the two proteins select slightly different consensus sites *in vitro* (3,4,6). The major bands consistently selected by Dri in this assay are indicated by markers to the right of the Dri panel in Figure 3. In contrast to Dri, p270 (ARID1A) binds in a

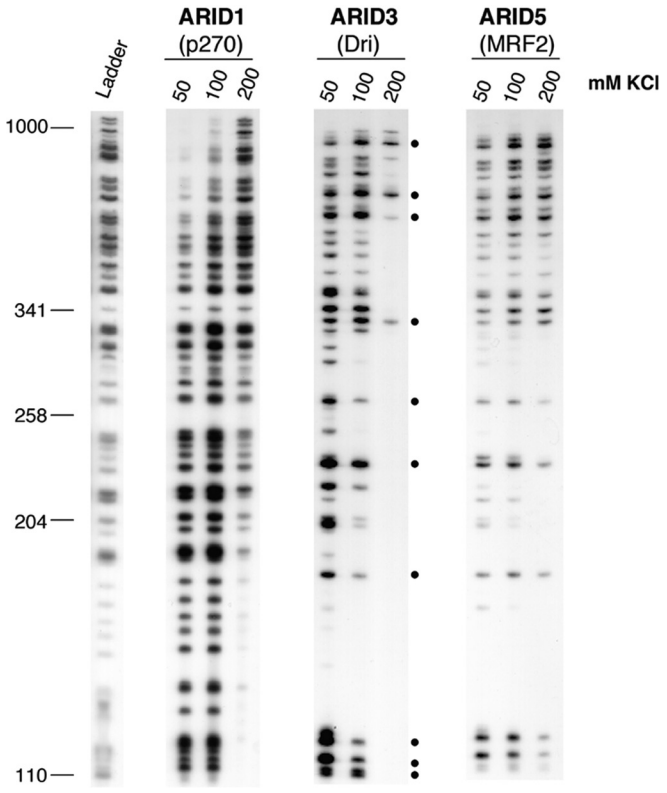

Figure 3. DNA-binding properties of the ARID family. Lambda phage DNA was digested with EcoRI, HindIII and Sau3AI to generate a large DNA oligonucleotide pool predicted to contain 128 fragments ranging in size from 12 to 2225 bp. The fragments were filled in with [^{32}P]dATP, incubated with GST-fusion proteins containing the ARID regions of each representative subfamily member as indicated, pulled down with glutathione beads, and analyzed by PAGE. Lane 1 shows the unselected pool of DNA fragments. Remaining lanes show the fragments selected in Lambda DNA-binding buffer with increasing KCl concentrations as indicated for p270, Dri and MRF2. Each subfamily is indicated at the top with the particular representative subfamily member assayed indicated directly below. The dots on the right of the Dri panel designate the major bands that are consistently selected by Dri in at least 10 repeats of this assay.

non-specific manner, binding to all fragments offered to it, showing selectivity only for longer fragments (>200 bp) at higher salt concentrations, presumably because longer fragments offer multiple binding sites. Despite the differences in sequence specificity, all three ARID proteins show similar affinities for DNA. These patterns have been documented previously (4,6,9,11), and are shown here for ease of comparison and as controls for the assay. The ARID1B protein has also been compared directly with p270, and found to bind without specificity (20).

The assay was used to examine the sequence specificity and DNA-binding affinity of representative members from each of the four remaining ARID subfamilies (Figure 4). ARID2 is the only member of its subfamily. A full-length human cDNA has not been reported thus far, but Genbank sequences predict an ARID consensus sequence at the N-terminus of the *ARID2* gene product. Isolation of N-terminal cDNA sequence by RT–PCR from the human liver cell line HepG2 confirms the presence of the ARID in the transcript (accession number AY727870.1). Studies on mammalian ARID2 have not yet been reported, but the protein is an apparent ortholog of the *Drosophila* ARID protein BCDNA:GH12174 (CG3274). Both proteins contain an RFX domain, which is an additional DNA-binding domain [reviewed in (21)]. Interestingly, the protein product of *Drosophila* BCDNA:GH12174 was recently found to be a component of the SWI/SNF-like complex PBAP, and was designated BAP170 (22). This complex is distinguished from the BAP SWI/SNF-like complex in part by its lack of Osa. This finding extends the role of ARID-containing subunits as components of SWI/SNF-related chromatin-remodeling complexes. Analysis of ARID2 in the DNA pull-down assay (Figure 4) indicates that it binds DNA without sequence specificity, like all other known ARID-containing components of SWI/SNF-related complexes.

ARID4 subfamily DNA-binding activity is represented here by RBP1 (ARID4A). Amino acid identity within the ARID consensus is 75% between RBP1 (ARID4A) and RBP1L1 (ARID4B), the only other member of this class. The assay shown in Figure 4 indicates that RBP1 also binds DNA without sequence specificity. RBP1 has been characterized as a repressor of E2F-dependent transcription recruited by the retinoblastoma protein (pRb) and can recruit histone deacetylase (19,23,24). RBP1L1 (syn.: SAP180) is also able to repress transcription, at least when tethered to DNA through the Gal DNA-binding domain (25). Both RBP1 and RBP1L1/SAP180 have been found in association with the mSIN3-histone deacetylase complex (19,25).

JARID1 is the largest ARID subfamily. It contains four highly homologous members. RBP2 (JARID1A) can enhance nuclear hormone receptor transactivation in reporter assays (26). PLU-1 is highly expressed in breast cancers, and in reporter assays has transcriptional repressor properties (27). SMCX (JARID1C) and SMCY (JARID1D) are thought to be regulators of minor histocompatability antigen (28,29). The four JARID1 proteins share 83% amino acid identity within the ARID and are highly related across their full sequences. This subfamily, in common with JARID2, contains highly conserved JmJN and JmJC domains. The proposed nomenclature reflects these relationships. The function of JmJN and JmJC domains is not yet clear, but they exist in proteins other than ARID family members (30,31). Two representatives of the JARID1 subfamily were chosen for analysis. A second subfamily member was included for two reasons. First, amino acid identity between PLU-1 (JARID1B) and the other three members of this subfamily varies more than is typical within subfamilies in the Loop 2 and Helix 5 region of the ARID sequence (see Figure 1). PLU-1 has a histidine within the Helix 5 region at a position where the other members of this subfamily contain a leucine. This region is the major groove interaction site in other ARID members and could be expected to play an important role in sequence recognition. Second, PLU-1 is expressed in a highly tissue restricted manner in contrast to other JARID1 members, which are broadly expressed [reviewed in (1)]. RBP2 (JARID1A) was chosen to represent the more typical members of this subfamily. As shown in Figure 4, both RBP2 (JARID1A) and PLU-1 (JARID1B) bind DNA with little or no discernible sequence specificity.

The panel was completed by testing jumonji (JARID2), the only member of its subfamily. jumonji is developmentally

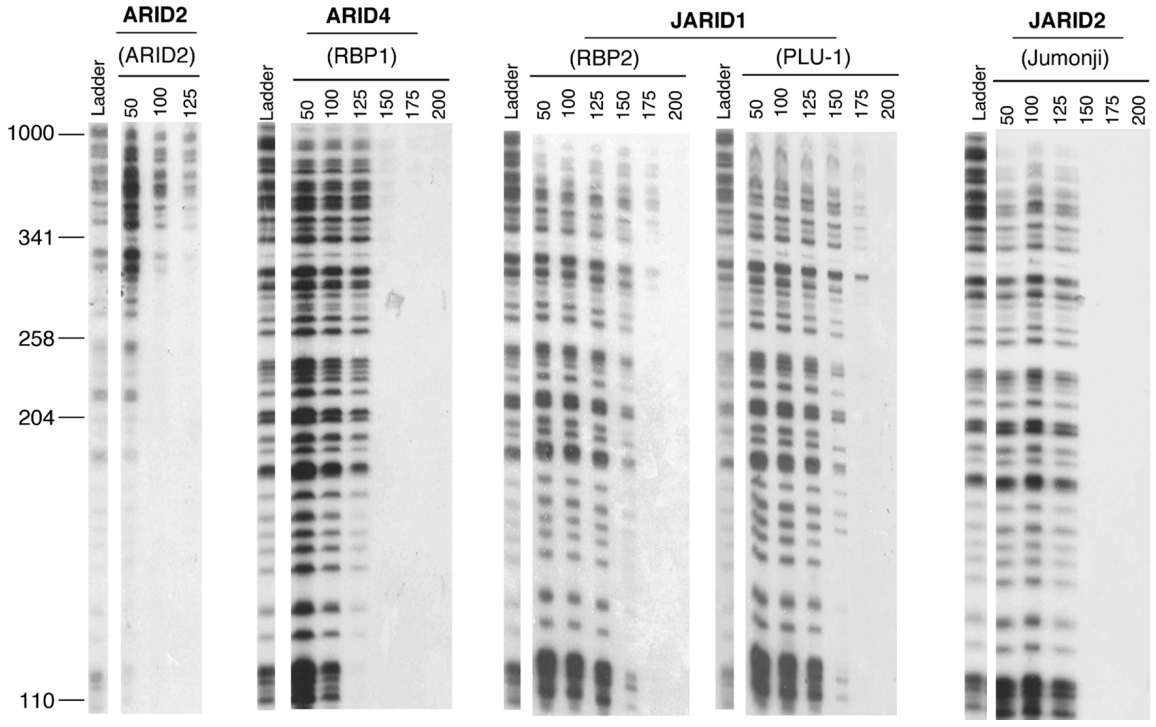

Figure 4. DNA-binding properties of the ARID family. Representative members of the remaining subfamilies were assayed as indicated in the legend to Figure 3. The unselected pool of DNA fragments is labeled Ladder and is shown for each individual experiment.

important in diverse organs (32,33), and can act as a transcriptional repressor in a reporter assay (34). Although jumonji has JmJN and JmJC domains in common with the JARID1 subfamily, the members of JARID1 are more similar to each other than to jumonji. Within the ARID domain, jumonji is only about 25% identical to members of the JARID1 group, which are 83% identical to each other. The jumonji ARID domain binds DNA in the pull-down assay without detectable sequence specificity (Figure 4). jumonji does show more of a tendency than other ARID family members to retain binding to lower molecular weight (<200 bp) DNA fragments even at high stringency, suggesting it does not disassociate as rapidly from DNA. This survey indicates that five of the seven ARID subfamilies bind DNA with no obvious sequence specificity. These results are summarized in Table 1.

The domains in the ARID1, ARID3 and ARID5 subfamilies retain DNA-binding affinity up to at least 200 mM KCl concentration (Figure 3). DNA affinity columns likewise indicate that p270 and Dri have similar DNA-binding affinities [(11) and Figure 8]. ARID1B, which is closely related overall to p270, retains DNA binding up to about 175 mM KCl [(20) and additional data not shown]. JARID1 subfamily domains also retain binding to at least 175 mM KCl (Figure 4). However, the data in Figure 4 indicate that ARID domains of the ARID2, ARID4 and JARID2 subfamilies have relatively low DNA-binding affinity. While this assay is not a direct measure of affinity, the results suggest that there are three distinguishable categories in the ARID family with regard to DNA-binding: sequence non-specific with low affinity, sequence non-specific with high affinity and sequence specific with high affinity. Previously, we showed that *Saccharomyces cerevisieae* SWI1 has relatively low affinity DNA-binding behavior that correlates with atypical sequence in the Loop 2 and Helix 5 region (11). Current results indicate that DNA-binding affinity of ARID family members can be low for reasons not easily apparent from inspection of the ARID sequence.

Table 1. Categorization of ARID subfamilies according to sequence specificity

HUGO nomenclature	Aliases	Tissue specificity
AT-specific		
ARID3A	Bright, DRIL1, E2FBP1	Restricted (mature B cells and testes) (3)
ARID3B	Bdp, DRIL2	Broad (52)
ARID3C	XM_071061	Not reported
ARID5A	MRF-1	Not reported
ARID5B	MRF2	Broad w/some specialization [high in brain, kidney, lung (39)]
Sequence non-specific		
ARID1A	p270, BAF250a, hOsa1, OSA1, B120, hSWI1, p250, SMARCF1	Broad (9,41,42,53,54)
ARID1B	pKIAA1235, BAF250b, p250R, hOsa2, held/OSA1	Broad (41,42,55,56)
ARID2	pKIAA1557	Broad[a]
ARID4A	RBP1	Broad (57)
ARID4B	RBP1L1, BCAA1, SAP180	Restricted (testes) (58)
JARID1A	RBP2	Broad (57)
JARID1B	PLU-1	Restricted (testes) (59)
JARID1C	SMCX, XE169	Broad (28)
JARID1D	SMCY, KIAA0234	Broad (28)
JARID2	jumonji	Specialized [brain, heart, skeletal muscles, kidney, thymus (33)]

[a]http://www.kazusa.or.jp/huge/gfpage/KIAA1557

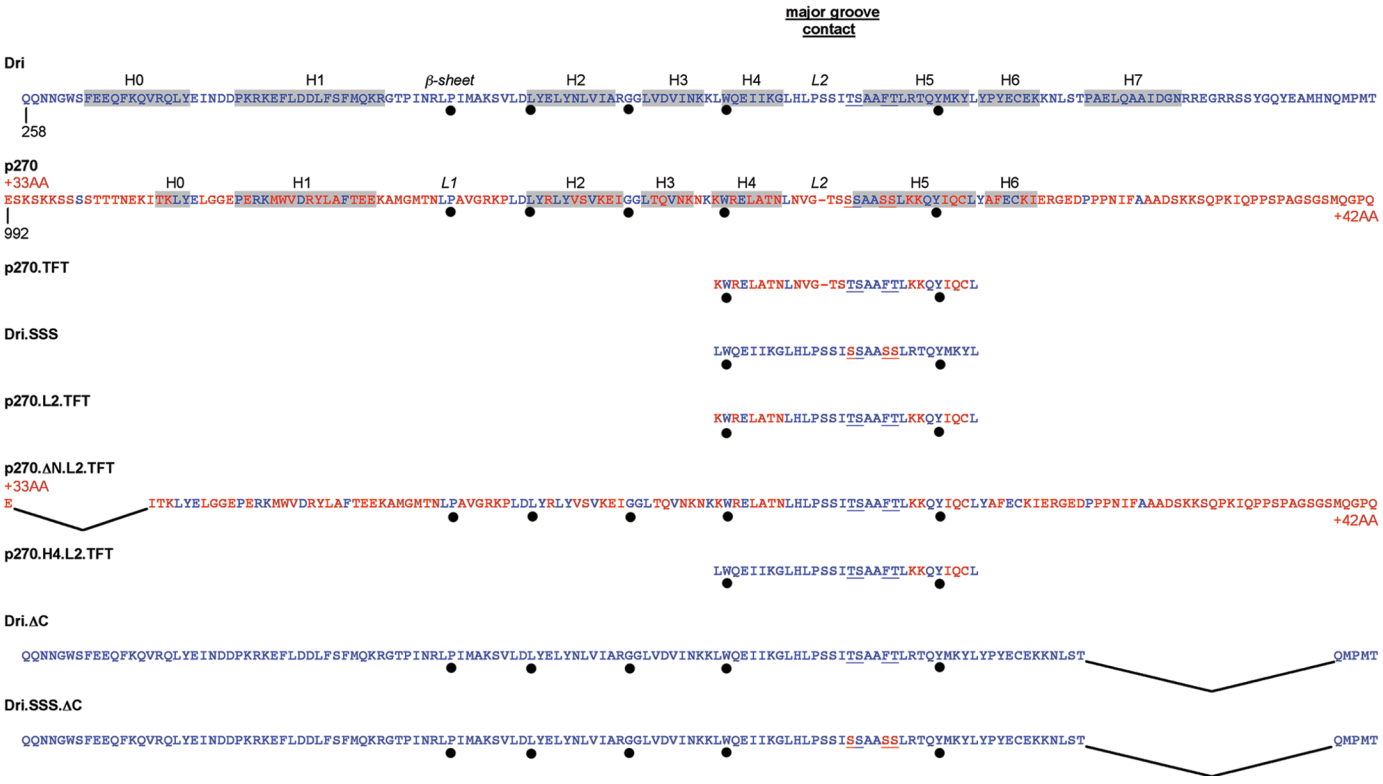

Figure 5. Sequences of the p270 and Dri mutants. The amino acid sequences of the wild-type p270 and Dri ARIDs are aligned. The Dri sequence is shown in blue print. The p270 sequence is shown in red print, with identity to Dri shown in blue. The sequence shown for Dri is the expression product of plasmid p410. The p270 peptide used in these assays is the expression product of plasmid pNDB8, which is longer on each end than the Dri peptide. The number of additional amino acids on each side is indicated in the figure. For each sequence, the first residue is given a number corresponding to its position in the full-length protein (accession numbers p270: NP_006006, Dri: AAB05771). The five residues that are invariant among all known ARID sequences are indicated by dots. The α-helices determined from NMR studies are indicated by grey shading and are numbered (from H0 to H7) above each sequence, along with the loops (L1 and L2) and β-sheet. Both Dri and p270 ARIDs have been studied by NMR in complex with DNA (15,18).The Helix4–Loop2–Helix5 region is the helix–turn–helix motif that contacts the major groove in both proteins, although in p270 Loop 2 contacts seem to contribute less than in Dri. The proteins also contact the adjacent minor grooves. In Dri, this happens through the β-sheet and the end of Helix 7. p270 contacts the minor groove via the Loop 1 region that corresponds to the β-sheet, but the C-terminal area of the p270 ARID does not seem to contribute to DNA binding as much as this region does in Dri. Additionally, a region of about 15 amino acids upstream of p270 Helix 0 interacts with DNA, but a comparable contact site does not exist in Dri. According to the structural study of the Dri–DNA complex, four residues in the Loop2–Helix5 region, two threonines (T), one serine (S) and one phenylalanine (F), make base-specific contacts. These residues and the corresponding residues in p270, all serines (S), are indicated by underlining in the figure. The mutant peptides p270.TFT, Dri.SSS, p270.L2.TFT and p270.H4.L2.TFT have changes only in the Helix4–Loop2–Helix5 region and therefore only that region is shown. For the in-frame deletion mutants, the whole sequence is shown, with the boundaries between deleted sequences shown by the solid lines. For the p270.ΔN.L2.TFT mutant, residues from S993 to K1007 were deleted. For the Dri.ΔC and the Dri.SSS.ΔC mutants, residues from P378 to N405 were deleted.

Sequence specificity does not depend solely on the identity of the specific major groove contact residues of Dri

Inspection of the sequences in Figure 1 does not reveal any obvious distinction between sequence-specific and sequence-non-specific ARIDs. However, the structures of the MRF2, Dri and p270 ARIDs have been solved in complex with DNA (15,17,18). Each study agrees that a portion of the region encompassing Loop 2 and/or Helix 5 lies within the major groove (see Figure 5), and that regions upstream and/or downstream of the junction of Loop 2 and Helix 5 contact the minor groove. The results have generated some ideas about the basis for sequence specificity, but these have not yet been tested empirically. Iwahara *et al*. (18) studied the Dri ARID by NMR, and identified four residues in Loop 2 and Helix 5 of Dri that make base-specific interactions in the major groove of the DNA. These residues included two threonines (T), a serine (S) and a phenylalanine (F), and are underlined in the Dri ARID sequence shown in Figure 5. p270 has a serine (S) at each of the corresponding positions. This study suggested that the lack of the threonines and of a non-polar residue at the phenylalanine position underlies the lack of sequence specificity in p270.

We undertook a site-specific mutagenesis study to compare the role of individual elements of the ARID in the determination of sequence specificity. The structures of the Dri and p270 ARIDs are the best characterized among their respective types in regard to DNA interactions, so these domains were chosen for comparison. To test whether the presence of the threonine and phenylalanine residues in Helix 5 is sufficient to confer sequence specificity, we generated a mutant variant in which threonines and a phenylalanine were introduced into the appropriate positions in p270. The sequence of the resultant mutant, designated p270.TFT, is shown in Figure 5.

The behavior of the p270.TFT mutant was examined in the Lambda DNA pull-down assay. The results (Figure 6A) show

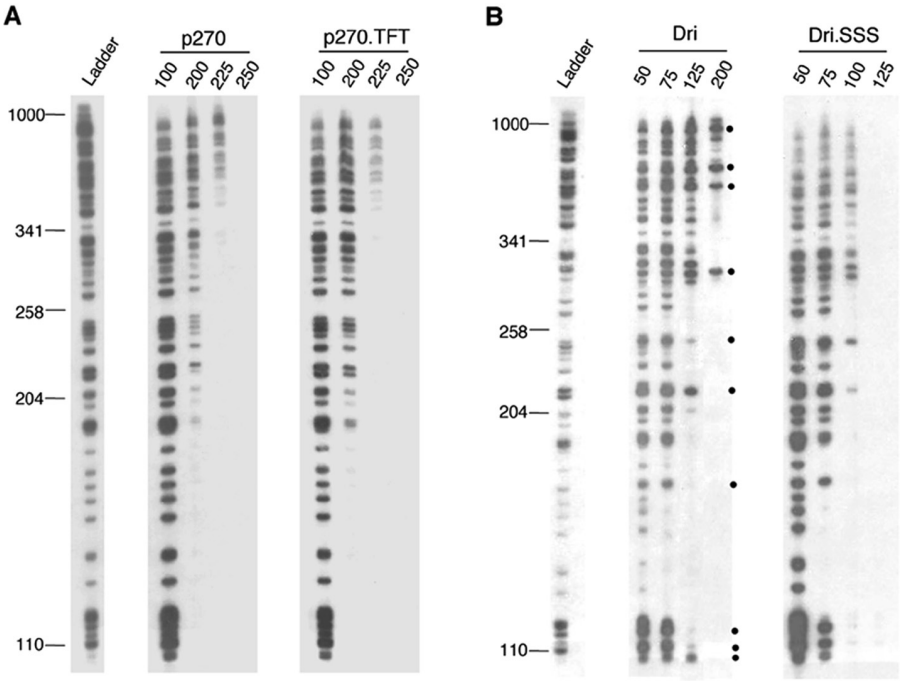

Figure 6. Assay of substitutions in the Helix 5 contact residues. The sequence specificity of the p270.TFT (**A**) and Dri.SSS (**B**) mutant peptides was tested in the Lambda DNA pull-down assay as described in the legend to Figure 3. The profiles of the wild-type p270 and Dri peptides are shown for reference. The dots on the right of the Dri panel designate the major bands that are consistently selected by Dri in this assay.

the mutant behaves exactly like wild-type p270. The substitutions do not confer any detectable capacity for sequence-specific binding, even at the highest stringency.

We also generated the reverse substitution in the Dri construct, replacing the presumptive base-specific contact residues with serines. The sequence of the resultant mutant, designated Dri.SSS, as shown in Figure 5, and the DNA-binding behavior is shown in Figure 6B. Strikingly, the Dri.SSS variant maintains a clear capacity for sequence-specific binding, selecting a pattern of DNA fragments very similar to those selected by wild-type Dri. It is apparent, though, from the KCl titration in Figure 6B, that the Dri.SSS variant has reduced overall affinity for DNA. No DNA binding is detected at this exposure in the 125 mM lane, while wild-type Dri consistently shows detectable binding in similar assays to at least 200 mM KCl (Figures 3 and 6B). The most direct explanation for these results is that these positions in Dri do make significant DNA contacts that are important for affinity, but which are not major determinants of sequence specificity. The DNA-binding affinity of p270 is strong despite the presence of serines at these positions, implying that the role of individual positions is not neccesarily directly comparable between different ARIDs.

The role of the helix–turn–helix motif

The NMR-derived p270 structure was reported earlier this year (15) and compared directly with the MRF2 structure (17). The structure indicates that Helix 5 of p270 lies within the major groove. However, these authors determined, by assessment of changes in the dynamics of the complex, that the shorter Loop 2 of p270 is less flexible than the corresponding loop in MRF2. This suggested a 'folding upon binding' mechanism of sequence recognition, in which the shorter length and/or less flexible composition of Loop 2 of p270 in comparison to MRF2 and Dri affects orientation of the major groove contact residues, and is thus responsible for lack of sequence-specific contact.

p270 Loop 2 does not appear to contact DNA directly (15), but to evaluate the possibility that Loop 2 affects the orientation of Helix 5 within the major groove, a p270 chimera was constructed in which the Loop 2 sequence of Dri was placed in the p270.TFT construct. This chimera, designated p270.L2.TFT, contains the major groove contact residues of Dri as well as a Loop 2 sequence derived entirely from Dri, which, therefore, should be sufficiently long and flexible to permit proper orientation of the DNA contact residues within the major groove. Nonetheless, when tested in the DNA pull-down assay, the p270.L2.TFT chimera shows no greater tendency to sequence specificity than wild-type p270, and, indeed, has slightly less overall affinity for DNA (Figure 7). Since the TFT substitution alone did not affect p270 DNA-binding affinity, it is likely that the introduction of the exogenous Loop 2 sequence created a distortion that interferes with the overall strength of DNA contact in the p270 ARID.

A significant difference between the way the p270 and Dri ARIDs interact with DNA is that the p270 ARID has an additional large minor groove interaction site just upstream of Helix 0 (15). We considered the possibility that this region of 15 amino acids interacts strongly and non-specifically with DNA in a way that masks potential sequence-specificity in p270. We therefore deleted this segment in the p270.L2.TFT chimera to generate a new construct designated p270. ΔL2.TFT. The DNA-binding affinity of this fragment is

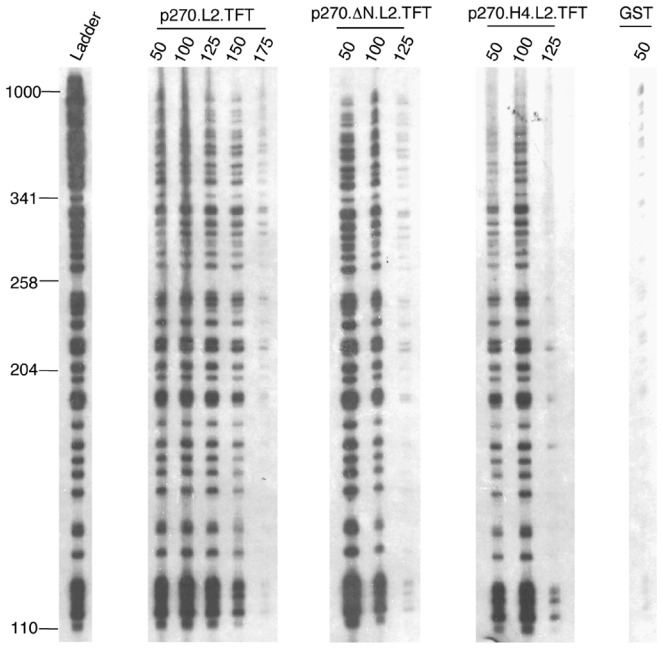

Figure 7. Assay of p270/Dri chimeras. The sequence specificity of the p270.L2.TFT, p270.ΔN.L2.TFT and p270.H4.L2.TFT mutant peptides was tested with the Lambda DNA pull-down assay as described in the legend to Figure 3. The GST DNA-binding profile is shown as a control.

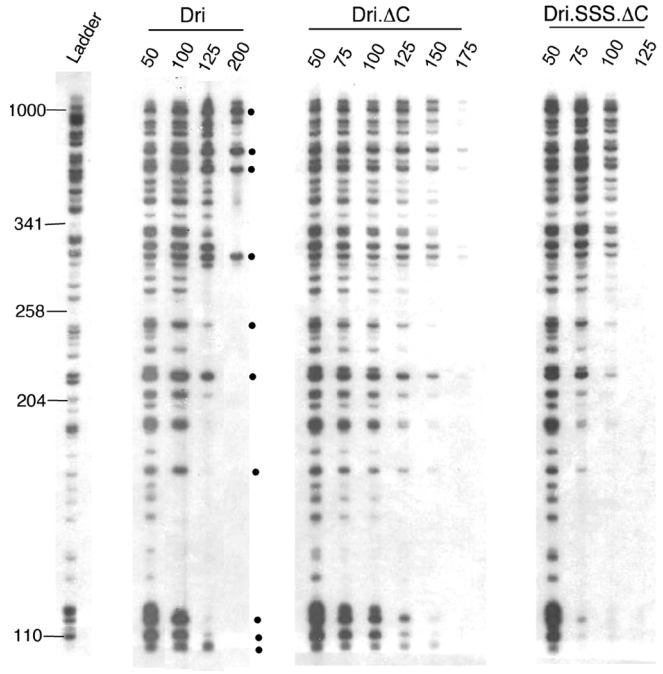

Figure 8. Deletion of the ARID3 extended sequence. The sequence specificity of the Dri.ΔC and Dri.SSS.ΔC mutant peptides was tested with the Lambda DNA pull-down assay as described in the legend to Figure 3. The profile of the wild-type Dri, along with the markers for the major bands selected by Dri, are also shown for reference.

reduced still further, confirming that the N-terminal region contributes significantly to DNA contact. However, the peptide still shows little or no selection for specific fragments (Figure 7). This argues against the possibility that sequence selectivity was transferred by the introduction of the Loop 2 and Helix 5 residues of Dri, but was masked by the unique N-terminal contact region of p270.

The ARID is categorized as a modified helix–turn–helix motif-based DNA-binding domain, in which the second helix of the motif (Helix 5) is the recognition helix. To test the possibility that the first helix of the motif (Helix 4) influences the orientation of Loop 2 and the recognition helix, the p270.L2.TFT construct was further modified such that the entire region from the beginning of Helix 4 to the last major groove contact residue consists of contiguous Dri sequence. The name of this mutant is p270.H4.L2.TFT. This construct still binds to DNA without any clear sequence selectivity (Figure 7). Moreover, affinity is reduced below that of the p270.L2.TFT variant, supporting the suggestion above, that introduction of exogenous sequence creates distortions that interfere with the overall strength of DNA contact in the p270 ARID. Individual elements are not directly exchangeable between different ARIDs. Together, these results indicate that sequence specificity in the ARID does not depend solely on the specific amino acid composition in the major groove contact region.

Contribution of the extended ARID region

Members of the ARID3 subfamily in all species studied are characterized by the presence of an 'extended' ARID sequence, a region of very high identity (>70% identity across ~35 residues) immediately following the core ARID consensus (see Figure 1). This region includes Helix 7, which is so far unique to the ARID3 subfamily, and extends beyond it. The extended ARID region has been identified as a DNA contact region in Dri (18). The extended ARID sequence is not present in the ARID5 subfamily, so cannot be a required determinant of sequence specificity. However, a corresponding position C-terminal to the core ARID consensus in MRF2 has been identified as a DNA contact region (17). In contrast, the corresponding region in p270 does not appear to make significant DNA contact (15). To assess the contribution of this region of Dri to sequence specificity, an in-frame deletion of sequence encoding 28 amino acids was generated in this region. The resulting mutant is designated Dri.ΔC. The DNA pull-down assay indicates that this construct retains a considerable measure of sequence selectivity (Figure 8). The construct shows slightly reduced DNA-binding affinity, consistent with the interpretation that the deleted region includes a DNA contact site.

The deletion of the extended ARID was also engineered in the Dri.SSS background. The resulting mutant is designated Dri.SSS.ΔC. In the DNA pull-down assay, the Dri.SSS.ΔC construct (Figure 8) has the same reduced DNA-binding affinity as was seen in the Dri.SSS mutant in Figure 6B. The sequence selectivity is further reduced, but a weak selectivity is still evident.

The lambda DNA restriction fragment pool was used as the target DNA in order to offer a wide range of sequence possibilities to the ARID proteins used in this survey. This allowed for the possibility that some family members, or some mutant variants, would show sequence preference, but for a previously unrecognized sequence. However, a disadvantage

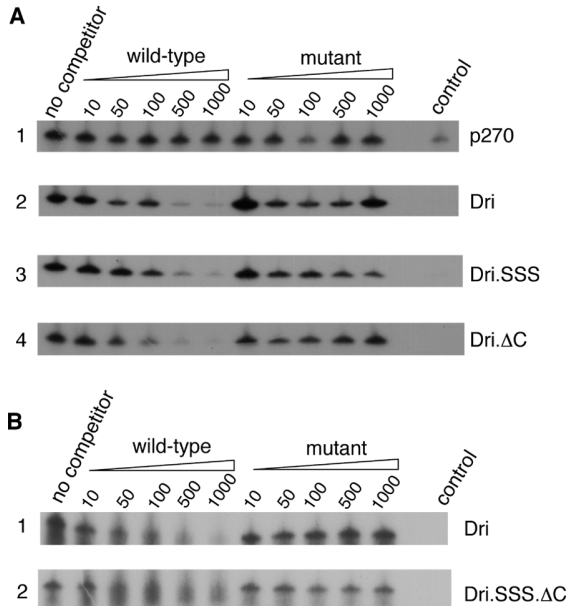

Figure 9. Competition assay with the Dri mutants. The competition assays show binding of the GST-fusion peptides to a labeled oligonucleotide containing a consensus sequence for Dri. All reactions contain a 10 000-fold excess of salmon sperm DNA. Binding to the consensus sequence was competed with increasing amounts of unlabeled specific competitor, either the consensus sequence (wild-type) or an altered sequence (mutant) in a 10-fold, 50-fold, 100-fold, 500-fold or 1000-fold excess. A separate reaction was used in each experiment with just glutathione–agarose beads as a control.

of the lambda DNA restriction fragment pool is that the complex restriction pattern precludes the identification of individual restriction fragments, or the actual sequence of the selected fragments. To obtain a more quantitative measurement of sequence specificity, selected mutants were probed in an oligonucleotide competition assay, where their affinity for a Dri consensus binding site (CCAATTAATCCC) was compared with their affinity for an altered consensus site (CCAATTGCTCCC). The consensus sites were synthesized as three tandem repeats. This assay was performed in low salt (50 mM) conditions, so that the effect of increasing salt concentrations on the conformation of the protein would not be a factor in the assay. The results are shown in Figure 9A. The Dri peptide shows a clear preference for its identified consensus site in this assay, as reported previously (4). A 500-fold excess of cold competitor with the correct consensus sequence competes effectively with the labeled probe, while the altered sequence, even at 1000-fold excess, shows little ability to displace the peptide (Figure 9A, panel 2). In contrast, the AT-rich consensus site does not compete for p270 binding any better than the mutant oligonucleotide (Figure 9A, panel 1).

The Dri.SSS and Dri.ΔC variants were both tested in this assay. The results show that each has a higher preference for the AT-rich consensus site than for the altered oligonucleotide (Figure 9A, panels 3 and 4), meaning that they clearly retain sequence-specific binding. This is consistent with the behavior they showed in Figures 6B and 8.

When we attempted this assay with the Dri.SSS.ΔC construct, we found that the peptide bound poorly to the oligonucleotide even at 50 mM salt. Because the ARID proteins show a generally higher affinity for longer pieces of DNA, we attempted the assay with a longer oligonucleotide, containing eight repeats of the consensus sequence rather than three. The Dri.SSS.ΔC peptide bound well to this probe (Figure 9B). Wild-type Dri showed the same behavior on this probe as on the shorter one: it was displaced more readily by the true consensus sequence than by the altered sequence (Figure 9B, panel 1). The Dri.SSS.ΔC peptide showed less specificity than wild-type Dri, but a weak selectivity was still evident (Figure 9B, panel 2), consistent with the behavior seen in Figure 8.

The DNA-binding phenotype of the double mutant, Dri.SSS.ΔC, implies that the region C-terminal to the core ARID consensus, and amino acid identity at the major groove contact site, contribute to the presence of sequence specificity in ARID3 subfamily proteins, but do not support a conclusion that small amino acid differences, such as the identity of residues at the junction of Loop 2 and Helix 5, or the length of Loop 2, are the principal determinants of sequence specificity. Rather, the results suggest that overall differences in the three-dimensional structure of individual ARID subfamilies determines the presence of sequence specificity. A similar situation appears to hold for the distinction between sequence-specific and sequence-non-specific DNA binding in HMG domain proteins, considered further in the Discussion.

p270 and Dri differ in their ability to tolerate mutations in the aromatic scaffold

The potential for differences in the overall structure of the p270 and Dri ARIDs was probed by introducing changes into the aromatic scaffold of the two domains. Within the core consensus sequence, there are five invariant amino acids that are almost identically spaced across each ARID. These are indicated by red text in Figure 1 and dots in Figure 4, and include a tryptophan (W) in Helix 4, a tyrosine (Y) in Helix 5 and a proline (P) in Loop 1. The presence of a series of invariant aromatic residues has been recognized as a structural scaffold in other helix–turn–helix motifs, including the DNA-binding motif in c-Myb and the homeodomain (35,36).

To test the contribution of the invariant residues in the ARID structure, specific invariant residues were changed to the small neutral residue alanine in the ARIDs of both Dri and p270. The resultant wild-type and mutant peptides were translated *in vitro*, and their DNA-binding affinity was assessed using a sensitive and quantitative DNA affinity column chromatography assay described previously (9,11). Because the DNA is in large excess, the assay is unbiased with respect to sequence specificity. The results are shown in Figure 10.

The interaction of the wild-type p270 ARID-containing peptide with DNA is as strong as that of the wild-type Dri ARID-containing peptide. In both wild-type proteins, 80–90% of the signal is retained on the columns. The remainder comes off in the flow-through and the first wash, and presumably represents a fraction of peptide that did not bind due to impaired folding. The proline-to-alanine substitution has very little effect on the elution profile of either p270 or Dri, suggesting that this residue, though invariant, is not by itself critical for the maintenance of structural integrity in the domain.

On the other hand, the Helix 4 tryptophan-to-alanine substitution seriously impaired binding to native DNA in both

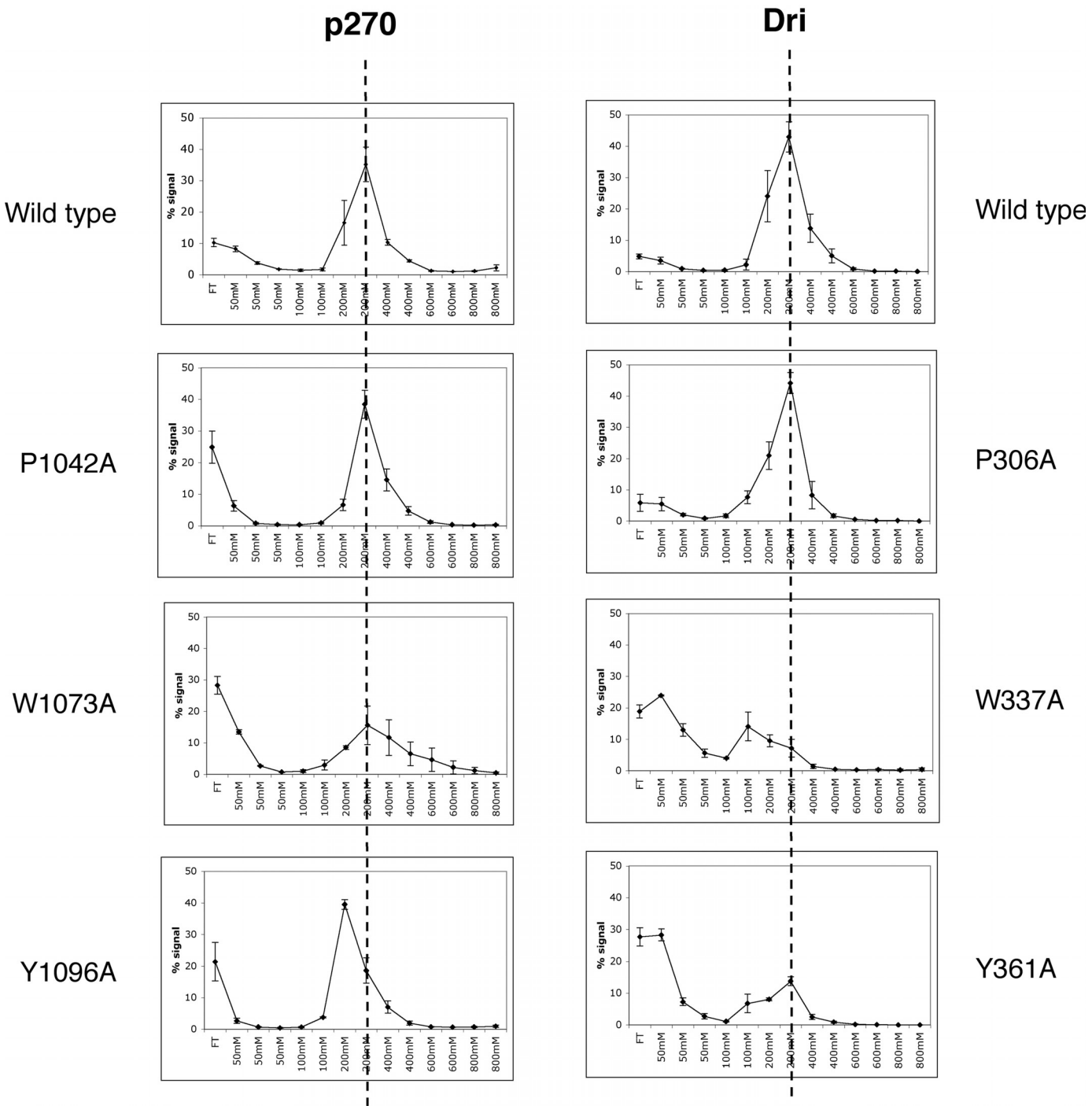

Figure 10. Substitution of invariant residues. The strength of the interaction of the wild-type p270 and Dri peptides, as well as peptides where invariant residues were substituted by an alanine, was tested by DNA affinity chromatography. *In vitro* translated ^{35}S-methionine-labeled peptides were applied to a native DNA cellulose column as described in Materials and Methods. Bound protein was eluted stepwise with loading buffer adjusted to contain increasing concentrations of NaCl from 100 to 800 mM, as indicated in the figure. Fractions were separated by SDS–PAGE and the p270 signal in each fraction was quantified by phosphorimaging. The results are plotted as the percentage of signal in each fraction relative to the entire signal recovered. Each experiment was performed at least twice and the error bars represent the average deviation. Graphs are aligned for ease of comparison. The dashed line indicates the second 200 mM fraction for reference.

ARIDs. Much of the mutant protein (about 45–50% in either p270 or Dri) fails to bind to the column and is recovered in the flow-through and wash fractions. The strongly deleterious effect of the tryptophan substitution suggests that the invariable tryptophan plays a critical role in maintaining the overall integrity of the ARID structure in both sequence-specific and sequence-non-specific representatives of the family.

p270 and Dri showed a different tolerance to the third mutation, a tyrosine-to-alanine substitution in Helix-5. The elution profile of the p270 mutant peptide is similar to that

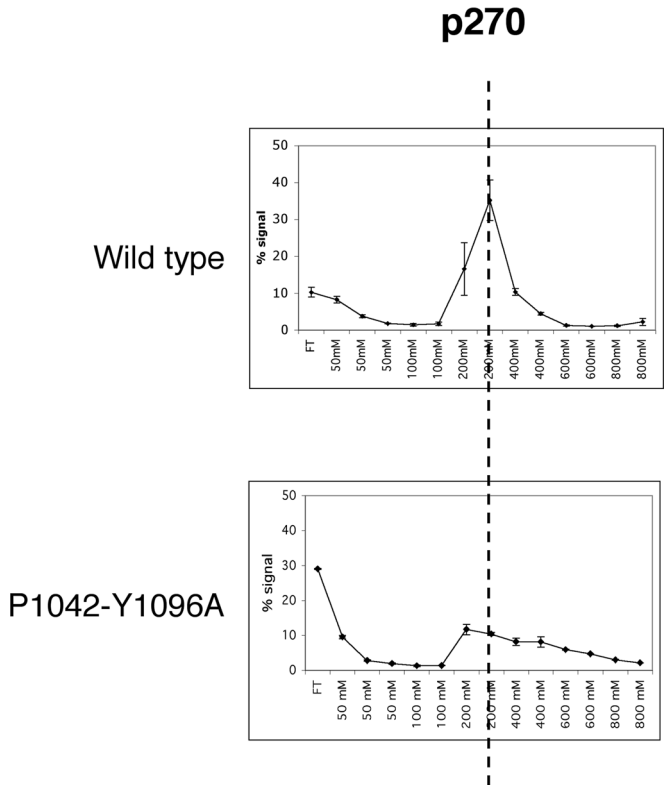

Figure 11. Combination of the proline and tyrosine substitutions can act synergistically to impair p270 ARID binding to DNA. The strength of the interaction of the combined substitution mutant was tested by DNA affinity chromatography as described above. The elution profile of the wild-type p270 is repeated in this panel and the graphs are aligned for ease of comparison. The dashed line indicates the second 200 mM fraction for reference.

of the wild-type peptide. Approximately the same amount of signal is retained on the column, although the shift in the elution peak from the second to the first 200 mM fraction indicates a weakening of affinity. This type of elution profile suggests that the substitution causes loss of one or more DNA contact sites, but does not suggest that protein folding is grossly affected. In contrast, the corresponding substitution in Dri is as deleterious as the tryptophan substitution, with 40–50% of the signal failing to bind to the column, implying that this residue is critical in the Dri ARID for maintaining proper structure.

To probe further the role of the Helix 5 tyrosine in the p270 ARID, the Y1096A substitution was combined with the P1042A substitution. The effect of the combined mutations was highly synergistic (Figure 11), generating a DNA-binding profile almost as defective as that seen with the W1073A substitution. This confirms that the Helix 5 tyrosine is important to structural integrity in the p270 ARID, but the results suggest that the p270 ARID is more able than the Dri ARID to tolerate changes in its aromatic scaffold. Thus, there appear to be fundamental differences in the ARID structures of p270 and Dri that go beyond simple differences at specific amino acid positions. This is consistent with the detrimental effects observed above of exchanging presumably analogous sequences between the two proteins. The mutagenesis studies argue against a conclusion that specific amino acids in Loop 2 or Helix 5 are the main determinants of sequence specificity. Most probably, this is determined by multiple interacting differences across the entire ARID structure.

DISCUSSION

The overall conclusion from the survey described here is that the majority of ARID subfamily domains bind DNA without regard to sequence specificity. Thus, the acronym is somewhat of a misnomer, although it is a well established and useful descriptor for a domain whose parameters are well-defined. This survey did not probe the behavior of every single member of the human ARID family. The proteins that have not been tested directly, here or elsewhere, among the subfamilies now designated as sequence non-specific are RBP1L1 (ARID4B), SMCX and SMCY (JARID1C and JARID1D). Each of these shows at least 75% identity and even greater similarity to the tested members of its subfamily. In addition, a mention of data not shown in a report on RBP1L1 (syn:SAP180) notes that a high-affinity consensus binding site could not be found in DNA-binding site selection experiments (25). Among the subfamilies now designated as AT-specific, only DRIL2 (ARID3B) and ARID3C have not been tested empirically, but again, there is at least 75% identity and more than 90% similarity between these ARID sequences and the AT-specific prototypes Bright and Dri. There is a potential conflict between our conclusions and a report suggesting that an ARID-containing fusion peptide of jumonji (JARID2) may have general selectivity for AT-rich sequences, since a majority of sequences selected by jumonji from a pool of random oligonucleotides were AT rich (34). However, several sequences that jumonji bound with equally high affinity in that study were not AT rich, and a precise consensus site could not be identified. The present survey is concerned with the properties inherent in the ARID sequence from each subfamily. As such, it was conducted with fusion proteins expressing the respective ARID sequences separate from the context of the native proteins. It remains possible that the endogenous proteins acquire a degree of sequence-specific binding behavior in physiological conditions.

The emergence of specific ARID subfamilies appears to have occurred early in evolution. *S.cerevisieae* encodes two ARID proteins. The ARID sequences do not correlate closely with any particular human subfamily, but overall the proteins seem most similar to the ARID1 and JARID1 subfamilies. *Schizosaccharomyces pombe* encodes four ARID proteins, two that are members of chromatin remodeling complexes and two that share similarity to the JARID1 subfamily. *Ceanorhabditis elegans* encodes four ARID proteins, aligning with human subfamilies ARID1, ARID2, ARID3 and JARID1, thus including a single AT-specific subfamily representative (for an excellent review of ARID evolution see http://www.lifesci.utexas.edu/research/tuckerlab/bright/evolution/). The ARID protein CFI-1 is the only identified member of an ARID3-type subfamily within *C.elegans* and prefers the same AT-rich consensus sequence as Dri in a competition assay (37). *Drosophila melanogaster* encodes six ARID proteins, one aligning with each subfamily except the second AT-rich specific subfamily ARID5. These patterns suggest that

ARIDs probably began as sequence non-specific and gained the property of sequence specificity through evolution.

The precise function of all the human ARID proteins is not known. Members of the AT-specific ARID3 and ARID5 subfamilies are sequence-specific transcription factors with recognized promoter targeting functions and important roles in development and differentiation (3,4,5,38–40). Among the sequence-non-specific ARID proteins, several appear to participate in general transcription and chromatin remodeling functions. ARID1A and ARID1B are mutually alternative members of human SWI/SNF-related complexes (20,41,42) and ARID1A (p270) is implicated in the tumor suppressor activity of the complexes (43). Human ARID2 is uncharacterized, but the *Drosophila* ortholog of ARID2 is a member of a SWI/SNF-like complex (22). ARID4A and ARID4B can associate with the mSIN3-histone deacetylase complex (19,25). Members of the JARID1 and JARID2 subfamilies show transcription activation and/or repression functions (26,27,34). To date, only the Dri and Bright (ARID3A) ARIDs have actually been shown to be required for the physiological function of their cognate proteins (44,45). The ARID of the *S.cerevisieae* protein SWI1 appears dispensable for complementation of the SWI1 phenotype (46), but transient reporter assays suggest the ARID is required for a transactivation function in human ARID1B (41). More physiological experiments are needed.

Site-specific mutagenesis has not revealed any precise determinants for sequence specificity or lack of it within the ARID family. Most probably, this is determined by multiple interacting differences across the entire ARID structure. A similar situation appears to hold for the distinction between sequence-specific and sequence-non-specific DNA binding in high mobility group (HMG) domain proteins. HMG domain containing proteins bind DNA through contacts in the minor groove. They recognize DNA structures such as four-way junctions, distorted cisplatin-kinked DNA and supercoiled DNA, and generally have the ability to bend DNA. One HMG protein subfamily consists of transcription factors like LEF-1 (lymphoid enhancer factor-1) and SRY (mammalian sex determining gene) that bind sequence specifically to AT-rich sequences in enhancer and promoter regions. Members of this subfamily contain one copy of the HMG domain and are tissue specific. Another subfamily comprises chromosomal proteins such as HMG1 and HMG2 that bind DNA in a sequence-non-specific manner. These proteins generally contain two or more HMG domains (47,48). There is a high degree of sequence similarity and structural characteristics between the sequence-specific and the sequence-non-specific HMG domains in complex with DNA. Some highly conserved residues have been identified as very important in sequence specificity of HMG domains, but these residues alone are not the sole determinant of sequence specificity [reviewed in (47)]. Rather, sequence-specificity appears to be a combination of effects of residues on the domain's positioning, affinity, its stability in complex with DNA, the number of interactions on the protein–DNA interface and the number of base-specific contacts (49). Studies of the HMG domain indicate the difference between sequence-specific and sequence-non-specific members of the same family is generally more complex than the simple substitution of contact residues for neutral residues.

ACKNOWLEDGEMENTS

We thank Michael Van Scoy and Dina Halegua for excellent technical assistance, and Robert Saint, Philip Branton and E. Premkumar Reddy for gifts of plasmids and other reagents. We are also grateful to Yuan Chen, Lois Maltais, Charles Grubmeyer, Phil Tucker, Loren Probst, Dale Haines, Scott Shore, Carmen Sapienza, Xavier Graña-Amat, and to members of our lab for advice and discussions. This work was supported by PHS grant CA53592 (E.M.) from the NIH, and a shared resources grant to the Fels Institute, CA88261. D.W. is the recipient of a DOD BCRP fellowship DAMD-17-01-1-0407 and a Daniel Swern Fellowship from Temple University. A.P. is the recipient of a DOD BCRP fellowship DAMD-17-02-1-0578 and a Daniel Swern Fellowship from Temple University. Funding the Open Access publication charges for this article was provided by PHS grant CA53592 (E.M.).

REFERENCES

1. Wilsker,D., Patsialou,A., Dallas,P.B. and Moran,E. (2002) ARID proteins: a diverse family of DNA binding proteins implicated in the control of cell growth, differentiation, and development. *Cell Growth Differ.*, **13**, 95–106.
2. Kortschak,R.D., Tucker,P.W. and Saint,R. (2000) ARID proteins come in from the desert. *Trends Biochem. Sci.*, **25**, 294–299.
3. Herrscher,R.F., Kaplan,M.H., Lelsz,D.L., Das,C., Scheuermann,C. and Tucker,P.W. (1995) 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**, 3067–3082.
4. Gregory,S.L., Kortschak,R.D., Kalionis,B. and Saint,R. (1996) Characterization of the dead ringer gene identifies a novel, highly conserved family of sequence-specific DNA binding proteins. *Mol. Cell. Biol.*, **16**, 792–799.
5. Huang,T.H., Oka,T., Asai,T., Okada,T., Merrills,B.W., Gertson,P.N., Whitson,R.H. and Itakura,K. (1996) Repression via differentiation-specific factor of the human cytomegalovirus enhancer. *Nucleic Acids Res.*, **24**, 1695–1701.
6. Whitson,R.H., Huang,T. and Itakura,K. (1999) The novel Mrf-2 DNA-binding domain recognizes a five-base core sequence through major and minor-groove contacts. *Biochem. Biophys. Res. Commun.*, **258**, 326–331.
7. Dallas,P.B., Yaciuk,P. and Moran,E. (1997) Characterization of monoclonal antibodies raised against p300: both p300 and CBP are present in intracellular TBP complexes. *J. Virol.*, **71**, 1726–1731.
8. Dallas,P.B., Cheney,I.W., Liao,D., Bowrin,V., Byam,W., Pacchione,S., Kobayashi,R., Yaciuk,P. and Moran,E. (1998) p300/CREB binding protein-related protein p270 is a component of mammalian SWI/SNF complexes. *Mol. Cell. Biol.*, **18**, 3596–3603.
9. Dallas,P.B., Pacchione,S., Wilsker,D., Bowrin,V., Kobayashi,R. and Moran,E. (2000) The human SWI/SNF complex protein, p270, is an ARID family member with nonsequence-specific DNA binding activity. *Mol. Cell. Biol.*, **20**, 3137–3146.
10. Nie,Z., Xue,Y., Yang,D., Zhou,S., Deroo,B.J., Archer,T.K. and Wang,W. (2000) A specificity and targeting subunit of a human SWI/SNF family-related chromatin-remodeling complex. *Mol. Cell. Biol.*, **20**, 8879–8888.
11. Wilsker,D., Patsialou,A., Zumbrun,S.D., Kim,S., Chen,Y., Dallas,P.B. and Moran,E. (2004) The DNA-binding properties of the ARID-containing subunits of yeast and mammalian SWI/SNF complexes. *Nucleic Acids Res.*, **32**, 1345–1353.
12. Collins,R.T., Furukawa,T., Tanese,N. and Treisman,J.E. (1999) Osa associates with the Brahma chromatin remodeling complex and promotes the activation of some target genes. *EMBO J.*, **18**, 7029–7040.
13. Yuan,Y.C., Whitson,R.H., Liu,Q., Itakura,K. and Chen,Y. (1998) A novel DNA-binding motif shares structural homology to DNA replication and repair nucleases and polymerases. *Nature Struct. Biol.*, **5**, 959–964.

14. Iwahara,J. and Clubb,R.T. (1999) Solution structure of the DNA binding domain from Dead ringer, a sequence-specific AT-rich interaction domain (ARID). *EMBO J.*, **18**, 6084–6094.
15. Kim,S., Zhang,Z., Upchurch,S., Isern,N. and Chen,Y. (2004) Structure and DNA-binding sites of the SWI1 AT-rich interaction domain (ARID) suggest determinants for sequence-specific DNA recognition. *J. Biol. Chem.*, **279**, 16670–16676.
16. Tu,X., Wu,J., Xu,Y. and Shi,Y. (2001) 1H, 13C and 15N resonance assignments and secondary structure of ADR6 DNA-binding domain. *J. Biomol. NMR*, **21**, 187–188.
17. Zhu,L., Hu,J., Lin,D., Whitson,R., Itakura,K. and Chen,Y. (2001) Dynamics of the Mrf-2 DNA binding domain free and in complex with DNA. *Biochemistry*, **40**, 9142–9150.
18. Iwahara,J., Iwahara,M., Daughdrill,G.W., Ford,J. and Clubb,R.T. (2002) The structure of the Dead ringer–DNA complex reveals how AT-rich interaction domains (ARIDs) recognize DNA. *EMBO J.*, **21**, 1197–1209.
19. Lai,A., Kennedy,B.K., Barbie,D.A., Bertos,N.R., Yang,X.J., Theberge,M.C., Tsai,S.C., Seto,E., Zhang,Y., Kuzmichev,A., Lane,W.S., Reinberg,D., Harlow,E. and Branton,P.E. (2001) 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**, 2918–2932.
20. Wang,X., Nagl,N.G.Jr., Wilsker,D., Van Scoy,M., Pacchione,S., Dallas,P.B. and Moran,E. (2004) Two related ARID family proteins are alternative subunits of human SWI/SNF complexes. *Biochem. J.*, **383**, 319–325.
21. Emery,P., Durand,B., Mach,B. and Reith,W. (1996) RFX proteins, a novel family of DNA binding proteins conserved in the eukaryotic kingdom. *Nucleic Acids Res.*, **24**, 803–807.
22. Mohrmann,L., Langenberg,K., Krijgsveld,J., Kal,A.J., Heck,A.J.R. and Verrijzer,C.P. (2004) Differential targeting of two distinct SWI/SNF-related Drosophila chromatin-remodeling complexes. *Mol. Cell. Biol.*, **24**, 3077–3088.
23. Lai,A., Marcellus,R.C., Corbell,H.B. and Branton,P.E. (1999) RBP1 induces growth arrest by repression of E2F-dependent transcription. *Oncogene*, **18**, 2091–2100.
24. Lai,A., Lee,J.M., Yang,W.M., DeCaprio,J.A., Kaelin,W.G.Jr., Seto,E. and Branton,P.E. (1999) RBP1 recruits both histone deacetylase-dependent and -independent repression activities to retinoblastoma family proteins. *Mol. Cell. Biol.*, **19**, 6632–6641.
25. Fleischer,T.C., Yun,U.J. and Ayer,D.E. (2003) Identification and characterization of three new components of the mSin3A corepressor complex. *Mol. Cell. Biol.*, **23**, 3456–3467.
26. Chan,S.W. and Hong,W. (2001) Retinoblastoma-binding protein 2 (Rbp2) potentiates nuclear hormone receptor-mediated transcription. *J. Biol. Chem.*, **276**, 28402–28412.
27. Tan,K., Shaw,A.L., Madsen,B., Jensen,K., Taylor-Papadimitriou,J. and Freemont,P.S. (2003) Human PLU-1 has transcriptional repression properties and interacts with the developmental transcription factors BF-1 and PAX9. *J. Biol. Chem.*, **278**, 20507–20513.
28. Agulnik,A.I., Mitchell,M.J., Lerner,J.L., Woods,D.R. and Bishop,C.E. (1994) A mouse Y chromosome gene encoded by a region essential for spermatogenesis and expression of male-specific minor histocompatibility antigens. *Hum. Mol. Genet.*, **3**, 873–878.
29. Agulnik,A.I., Mitchell,M.J., Mattei,M.G., Borsani,G., Avner,P.A., Lerner,J.L. and Bishop,C.E. (1994) A novel X gene with a widely transcribed Y-linked homologue escapes X-inactivation in mouse and human. *Hum. Mol. Genet.*, **3**, 879–884.
30. Balciunas,D. and Ronne,H. (2000) Evidence of domain swapping within the jumonji family of transcription factors. *Trends Biochem. Sci.*, **25**, 274–276.
31. Clissold,P.M. and Ponting,C.P. (2001) JmjC: cupin metalloenzyme-like domains in jumonji, hairless and phospholipase A2beta. *Trends Biochem. Sci.*, **26**, 7–9.
32. Takeuchi,T., Yamazaki,Y., Katoh-Fukui,Y., Tsuchiya,R., Kondo,S., Motoyama,J. and Higashinakagawa,T. (1995) Gene trap capture of a novel mouse gene, jumonji, required for neural tube formation. *Genes Dev.*, **9**, 1211–1222.
33. Lee,Y., Song,A.J., Baker,R., Micales,B., Conway,S.J. and Lyons,G.E. (2000) Jumonji, a nuclear protein that is necessary for normal heart development. *Circ. Res.*, **86**, 932–938.
34. Kim,T.G., Kraus,J.C., Chen,J. and Lee,Y. (2003) JUMONJI, a critical factor for cardiac development, functions as a transcriptional repressor. *J. Biol. Chem.*, **278**, 42247–42255.
35. Saikumar,P., Ramachadran,M. and Reddy,P. (1990) Role of the tryptophan repeats and flanking amino acids in Myd-DNA interactions. *Proc. Natl Acad. Sci. USA*, **87**, 8452–8456.
36. Subramaniam,V., Jovin,T.M. and Rivera-Pomar,R.V. (2001) Aromatic amino acids are critical for stability of the bicoid homeodomain. *J. Biol. Chem.*, **276**, 21506–21511.
37. Shaham,S. and Bargmann,C.I. (2002) Control of neuronal subtype identity by the *C. elegans* ARID protein CFI-1. *Genes Dev.*, **16**, 972–983.
38. Kaplan,M.H., Zong,R.T., Herrscher,R.F., Scheuermann,R.H. and Tucker,P.W. (2001) Transcriptional activation by a matrix associating region-binding protein. contextual requirements for the function of bright. *J. Biol. Chem.*, **276**, 21325–21330.
39. Lahoud,M.H., Ristevski,S., Venter,D.J., Jermiin,L.S., Bertoncello,I., Zavarsek,S., Hasthorpe,S., Drago,J., de Kretser,D., Hertzog,P.J. and Kola,I. (2001) Gene targeting of Desrt, a novel ARID class DNA-binding protein, causes growth retardation and abnormal development of reproductive organs. *Genome Res.*, **11**, 1327–1334.
40. Whitson,R.H., Tsark,W., Huang,T.H. and Itakura,K. (2003) Neonatal mortality and leanness in mice lacking the ARID transcription factor Mrf-2. *Biochem. Biophys. Res. Commun.*, **312**, 997–1004.
41. Inoue,H., Furukawa,T., Giannakopoulos,S., Zhou,S., King,D.S. and Tanese,N. (2002) Largest subunits of the human SWI/SNF chromatin-remodeling complex promote transcriptional activation by steroid hormone receptors. *J. Biol. Chem.*, **277**, 41674–41685.
42. Nie,Z., Yan,Z., Chen,E.H., Sechi,S., Ling,C., Zhou,S., Xue,Y., Yang,D., Murray,D., Kanakubo,E., Cleary,M.L. and Wang,W. (2003) Novel SWI/SNF chromatin-remodeling complexes contain a mixed-lineage leukemia chromosomal translocation partner. *Mol. Cell. Biol.*, **23**, 2942–2952.
43. Wang X., Nagl,N.G.Jr., Flowers S., Zweitzig D., Dallas P.B. and Moran E. (2004) Expression of p270 (ARID1A), a component of human SWI/SNF complexes, in human tumors. *Int. J. Cancer*, **112**, 636–642.
44. Shandala,T., Kortschak,R.D. and Saint,R. (2002) The Drosophila retained/dead ringer gene and ARID gene family function during development. *Int. J. Dev. Biol.*, **46**, 423–430.
45. Nixon,J.C., Rajaiya,J. and Webb,C.F. (2004) Mutations in the DNA-binding domain of the transcription factor, Bright, act as dominant negative proteins and interfere with immunoglobulin transactivation. *J. Biol. Chem.*, **279**, 52465–52472.
46. Prochasson,P., Neely,K.E., Hassan,A.H., Li,B. and Workman,J.L. (2003) Targeting activity is required for SWI/SNF function *in vivo* and is accomplished through two partially redundant activator-interaction domains. *Mol. Cell*, **12**, 983–990.
47. Murphy,F.V.,IV and Churchill,M.E. (2000) Nonsequence-specific DNA recognition: a structural perspective. *Structure Fold Des.*, **8**, R83–R89.
48. Thomas,J.O., Travers,A.A. (2001) HMG1 and 2, and related 'architectural' DNA-binding proteins. *Trends Biochem Sci.*, **26**, 167–174.
49. Hsiao,N.W., Samuel,D., Liu,Y.N., Chen,L.C., Yang,T.Y., Jayaraman,G. and Lyu,P.C. (2003) Mutagenesis study on the zebra fish SOX9 high-mobility group: comparison of sequence and non-sequence specific HMG domains. *Biochemistry*, **42**, 11183–11193.
50. Depiereux,E., Baudoux,G., Briffeuil,P., Reginster,I., DeBolle,X., Vinals,C. and Feytmans,E. (1997) Match-Box_server: a multiple sequence alignment tool placing emphasis on reliability. *Comput. Appl. Biosci.*, **13**, 249–256.
51. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L., Studholme D.J., Yeats C. and Eddy S.R. (2004) The Pfam Protein Families Database. *Nucleic Acids Res.*, **32**, D138–D141.
52. Numata,S., Claudio,P.P., Dean,C., Giordano,A. and Croce,C.M. (1999) Bdp, a new member of a family of DNA-binding proteins, associates with the retinoblastoma gene product. *Cancer Res.*, **59**, 3741–3747.
53. Takeuchi,T., Nicole,S., Misaki,A., Furihata,M., Iwata,J., Sonobe,H. and Ohtsuki,Y. (2001) Expression of SMARCF1, a truncated form of SWI1, in neuroblastoma. *Am. J. Pathol.*, **158**, 663–672.
54. Kozmik,Z., Machon,O., Kralova,J., Kreslova,J., Paces,J. and Vlcek,C. (2001) 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**, 140–148.
55. Nagase,T., Ishikawa,K., Kikuno,R., Hirosawa,M., Nomura,N. and Ohara,O. (1999) Prediction of the coding sequences of unidentified human genes. XV. The complete sequences of 100 new cDNA

clones from brain which code for large proteins in vitro. *DNA Res*., **6**, 337–345.
56. Hurlstone,A.F.L., Olave,I.A., Barker,N., van Noort,M. and Clevers,H. (2002) Cloning and characterization of hELD/OSA1, a novel BRG1 interacting protein. *Biochem. J*., **364**, 255–264.
57. Fattaey,A.R., Helin,K., Dembski,M.S., Dyson,N. and Harlow,E., Vuocolo,G.A., Hanobik,M.G., Haskell,K.M., Oliff,A., Defeo-Jones,D. and Jones,R.E. (1993) Characterization of the retinoblastoma binding proteins RBP1 and RBP2. *Oncogene*, **8**, 3149–3156.
58. Cao,J., Gao,T., Stanbridge,E.J. and Irie,R. (2001) RBP1L1, a retinoblastoma-binding protein-related gene encoding an antigenic epitope abundantly expressed in human carcinomas and normal testis. *J. Natl Cancer Inst*., **93**, 1159–1165.
59. Lu,P.J., Sundquist,K., Baeckstrom,D., Poulsom,R., Hanby,A., Meier-Ewert,S., Jones,T., Mitchell,M., Pitha-Rowe,P., Freemont,P. and Taylor-Papadimitriou,J. (1999) A novel gene (PLU-1) containing highly conserved putative DNA/chromatin binding motifs is specifically up-regulated in breast cancer. *J. Biol. Chem*., **274**, 15633–15645.