Available online at www.sciencedirect.com

SCIENCE DIRECT

Genomics 86 (2005) 242 – 251

GENOMICS

www.elsevier.com/locate/ygeno

Nomenclature of the ARID family of DNA-binding proteins☆

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

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

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

Abstract

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

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

Introduction

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

Since the discovery of the ARID, many additional proteins containing this domain have been identified. Interestingly though, not all ARID-containing proteins bind

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

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

[1] These authors contributed equally.

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

to DNA in a sequence-specific manner, as discussed further below. The cellular functions of ARID proteins include participation in the regulation of cell growth, differentiation, and development [7,8]. The ARID domain is both ancient and widespread, occurring in (some) protozoa, green algae, higher plants, fungi, and metazoans. No archae- or eubacterial ARID domains have been identified to date.

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

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

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

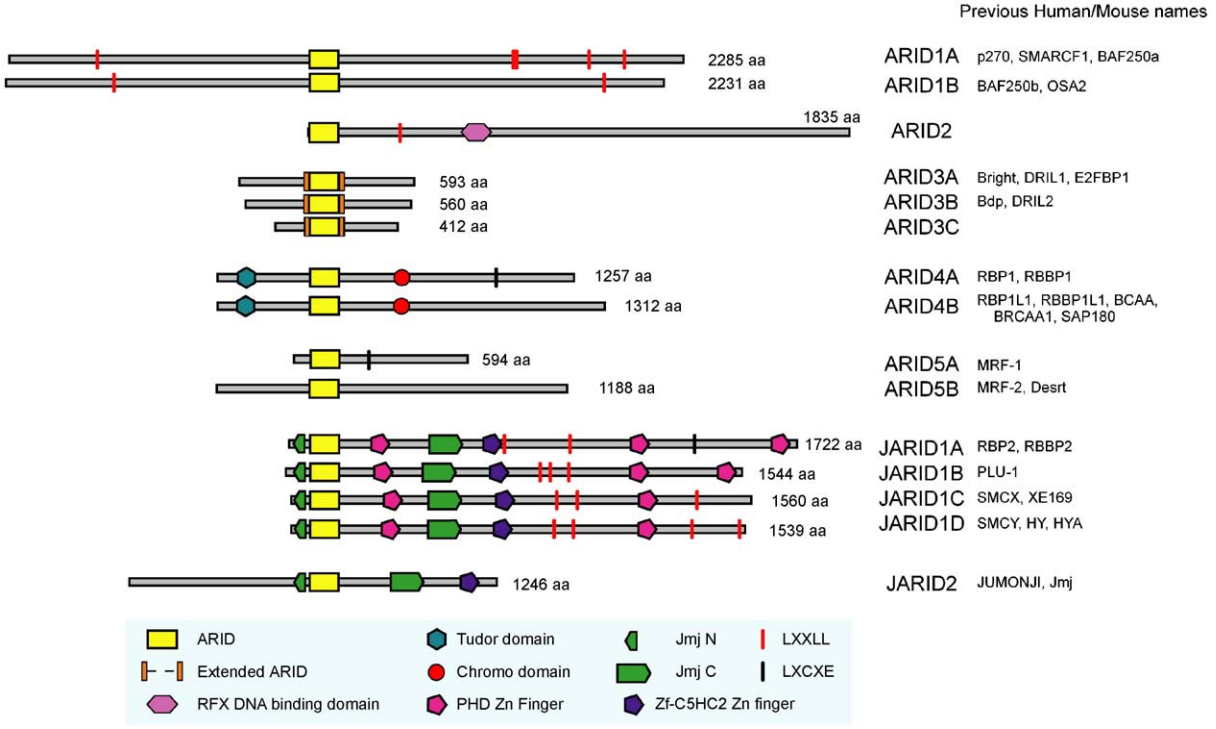

Fig. 1. The human ARID family of proteins. Genome sequencing reveals 15 ARID-containing proteins in humans. The ARID family proteins can be grouped into subfamilies based on their similarity to each other within the ARID domain. The nomenclature described here reflects this subclassification of the family and clarifies their relationships to each other. A subset of ARID-containing proteins also contains JmjN and JmjC domains, and the proposed nomenclature reflects these relationships as well. Within the proposed subfamilies, members typically have 70 to 85% identity within their ARID sequences, while across subgroups, identity within the ARID sequence drops to about 25 to 30%. The 15 human ARID family proteins are represented by open bars and are aligned according to the position of the ARID sequence (indicated in yellow). The relative positions of other well-characterized domains and motifs are represented by colored symbols identified at the bottom of the figure. The amino acid (aa) length of each protein is shown to the right of the bar. The presence of additional motifs was identified through the Pfam [55] or SMART [56] databases.