Missense variants reveal functional insights into the human ARID family of gene regulators

Gauri Deák and Atlanta G. Cook*
Wellcome Centre for Cell Biology, University of Edinburgh, Michael Swann Building, Max Born Crescent, Edinburgh EH9 3BF, United Kingdom

*Correspondence should be addressed to Atlanta Cook (atlanta.cook@ed.ac.uk)

Running title: ARIDs and missense variation

Abstract

Missense variants are alterations to protein coding sequences that result in amino acid substitutions. They can be deleterious if the amino acid is required for maintaining structure or/and function, but are likely to be tolerated at other sites. Consequently, missense variation within a healthy population can mirror the effects of negative selection on protein structure and function, such that functional sites on proteins are often depleted of missense variants. Advances in high-throughput sequencing have dramatically increased the sample size of available human variation data, allowing for population-wide analysis of selective pressures. In this study, we developed a convenient set of tools, called 1D-to-3D, for visualizing the positions of missense variants on protein sequences and structures. We used these tools to characterize human homologues of the ARID family of gene regulators. ARID family members are implicated in multiple cancer types, developmental disorders, and immunological diseases but current understanding of their mechanistic roles is incomplete. Combined with phylogenetic and structural analyses, our approach allowed us to characterise sites important for protein-protein interactions, histone modification recognition, and DNA binding by the ARID proteins. We find that comparing missense depletion patterns among paralogs can reveal sub-functionalization at the level of domains. We propose that visualizing missense variants and their depletion on structures can serve as a valuable tool for complementing evolutionary and experimental findings.

Keywords

Exome; Genetic Variation; Transcription Factors; Humans; nucleosomes; Coffin-Siris syndrome; intellectual disability; PHD zinc fingers; demethylation; DNA binding proteins; Cancer

Introduction

Advances in high-throughput sequencing have led to a sweeping expansion in genetic variation data of human protein-coding genes. With nearly 15 million curated exome variants made available by the international Genome Aggregation Database (gnomAD)(1), statistical analyses have identified genes that are intolerant to loss-of-function (LOF) and are likely associated with disease (1, 2). This has aided large-scale assessments of genetic causality in autism spectrum disorder (3) and inherited cardiomyopathies (4), and has also led to novel frameworks for evaluating drug toxicity or efficacy through consideration of the genetic constraint on target genes (5, 6). The increase in statistical power has allowed intolerance to variation to be used for ranking genes in relation to their importance to human health.

Beyond the effects of LOF variants on a gene, constraint analyses can also be narrowed to missense variants at the level of a protein domain. We define a domain as an independent folding unit or a conserved sequence block that is likely to approximate an independent folding unit. Depletion of missense variants in whole domains, or specific segments within domains, has been found to correlate with evolutionary conservation (7). Rare mutations occurring in such regions are likely to be pathogenic (8, 9). In agreement, saturation mutagenesis studies have shown that mutation-intolerant regions map to conserved protein domains, particularly at residues that are involved in DNA, protein, or ligand binding and are associated with pathogenic variants (10, 11). In another study, domain-level definitions improved disease prediction scores compared to random segmentation of sequences (12). These findings indicate that, like whole genes, functionally important regions in proteins are subject to negative selection. Patterns in population-wide missense variation could therefore be harnessed to gain insights into protein function(13).

While calculating variant distributions along protein sequences can help to identify essential domains, it fails to consider the arrangement of variants in 3D space (8). This has been addressed by manual mapping of variants (14) or computational mapping of variant depletion scores directly onto protein structures. For example, Hicks et al., developed a '3D Tolerance Score' which compares an observed and expected number of missense variants in 5 Å-radius spheres around individual atoms in a 3D structure (15). In an alternative approach, Tang et al. introduced PSCAN, which scores the spatial dispersion of missense variants onto structures (16), based on a previous finding that neutral variants tend to be dispersed, while pathogenic variants cluster (17). In a further approach, the MISCAST suite provides an approach to connect probabilities of loss of function with primary sequence level features (18).

Collectively, the above studies show that surfaces of proteins where important functional sites are located are depleted of missense variants. Statistical approaches enable sorting and/or predicting functional sites using proteome-wide approaches but may not necessarily enable non-specialists to inspect an individual protein of their choice. We developed two programs to allow easy visual inspection of variants on primary sequences (1D) and tertiary structures (3D) from the gnomAD database (Fig. 1A).

First, we generated a program to calculate the average density of missense variants in a protein domain (Vd) to the average density of missense variants in the whole protein (Vp). We show that plotting variants in 1D together with the ratio of Vd/Vp helps identify domains that are missense depleted. The standalone values of Vp for the ARID family members exhibit good correlation with the missense Z score (mZ), typically used in gnomAD (2).

Second, we developed a simple, convenient program called 1D-to-3D to map the same variant data onto any 3D structure, representing variants as spheres of increasing size and color intensity with increasing allele frequency (AF) (Fig. 1A). This allows users to find surfaces of the protein that are depleted for missense variants and to compare this with complementary information such as surface conservation.

Using 1D-to-3D, we performed a comprehensive analysis of missense variation in the "AT-rich interactive domain" (ARID) family of gene regulators (Fig. 1B). This family was selected due to the clinical significance of its members in multiple cancer types (19-23), rare developmental disorders such as Coffin-Siris syndrome (24), and immunological diseases including systemic lupus erythematosus, type II diabetes, and atherosclerosis (25-27). In humans, the ARID family comprises 15 proteins that can further be categorized into 7 subfamilies, all of which have an ARID DNA binding domain (28). Despite their common domain, the family members regulate distinct sets of genes via diverse molecular mechanisms (Fig. 1B). ARID1 and ARID2 are core subunits of the BRG1/BRM-associated factors (BAF) family of nucleosome remodeling complexes while JARID2 is an accessory subunit of the Polycomb Repressive Complex 2 (PRC2) (29, 30). In contrast, ARID3 proteins are transcription factors (31), while ARID4s and ARID5s are adapter proteins that recruit other transcriptional regulators such as the mSin3A-histone deacetylase (HDAC) complex and PHD Finger 2 (PHF2) histone demethylase respectively (32-34). ARID5A is thought to be an RNA-binding protein (35, 36). Finally, the four JARID1 proteins are enzymes that mediate transcriptional changes by removal of histone H3K4 di-/tri-methylation marks (21).

Here we provide a comprehensive analysis of the ARID family as a whole, including domain architecture mapping, phylogenetic analysis and searching for known pathogenic variants. We complemented these analyses with our 1D-to-3D approach to identify surfaces of proteins that are depleted (or not) of missense variants, to provide a deeper annotation of functional sites within these proteins.

Materials and Methods

Sequence Alignments and Domain Annotations

Sequences of ARID family orthologs were selected using the Oma orthology database (RRID:SCR_016425) (37). Multiple sequence alignments were generated using MAFFT (RRID:SCR_011811) (38) and pairwise alignments were generated with EMBOSS Needle (39). Alignments were then visualized in JalView 2.11.1.4 (RRID:SCR_006459) (40). Evolutionary relationships between paralogs in ARID subfamilies were verified using TreeFam (RRID:SCR_013401) (41). A structural sequence alignment for the ARID5B BAH domain was created using the DALI server (RRID:SCR_013433) (42). Functional domains of each family member were annotated using InterPro (43) or based on experimental data (the ARID1A/1B core binding regions (44), ARID4A/B R2 region (33), and JARID1A (45) and JARID1B (46) domains).

Structures and Models

A complete list of analyzed structures is in Supplementary Table 1. For domains with no available structure, we used models from the AlphaFold protein structure database (47). Only regions with predicted local-distance difference test (pLDDT) scores > 70 were considered. The pLDDT score is a confidence measure that reflects the validity of local inter-atomic distances in a predicted structure. A cut-off of > 70 is considered a “generally correct backbone prediction” (48). All of the structures were visualized in PyMOL (49). Surface electrostatics scores for the ARID4A and ARID4B hybrid Tudor domains were calculated using the Adaptive Poisson-Boltzmann Solver (RRID:SCR_008387) (50) in PyMol.

Surface Conservation

Surface conservation was mapped onto the structures of ARID1A (PDB ID 6LTJ, chain L), ARID1B (AlphaFold model, UniProt ID: Q8NFD5, amino acids (aa) 1593-1699 and 1905-2236), ARID2 (AlphaFold model, UniProt ID: Q68CP9, aa 155-464), JARID1A (PDB ID: 5CEH), and JARID1B (PDB ID: 5FUP) using ConSurf (RRID:SCR_002320)(51). For ARID1A/1B, MAFFT alignments of 70 Oma group vertebrate orthologs were submitted to the server. For ARID2 and JARID1A/1B, MAFFT alignments of 165 and 80 Oma group metazoan orthologs were submitted respectively. We selected Oma groups because they exclude paralogs and include only one co-ortholog if several are found for a given species. This yields a collection of non-redundant sequences that can be filtered at specific taxonomic levels (52). All sequence alignments can be accessed at: https://doi.org/10.7488/ds/3128.

Isoform expression analysis

Isoform-specific expression data for ARID5B was obtained from ISOexpresso (53). The ratio of Isoform 1 (uc001jlt.2) to Isoform 2 (uc001jlu.2) expression levels was compared across 735 samples from 22 human tissue types of healthy donors.

Constraint Metrics

Established constraint metrics including pLi, LOEUF, missense Z, and RVIS scores for each ARID family member were obtained from the official gnomAD and Genic Intolerance web browsers (1, 54). P-values corresponding to missense Z-scores were calculated using the Excel NORM.S.DIST function (the output for positive Z-scores was subtracted from 1). A complete list of collected metrics is available in Supplementary Table 1.

Variant Data Processing

7,652 non-synonymous variants associated with UniProt canonical sequences of the 15 ARID family proteins were extracted from the gnomAD v2.1.1 dataset (GRCh37/hg19) (1). The dataset is publicly available and contains variants from 125,748 quality-controlled exomes of unrelated, adult individuals not affected by severe pediatric disease (1). To ensure our analysis was restricted to neutral missense variants, only variants with the Variant Effect Predictor annotation 'missense' were considered, and variants with the ClinVar annotation 'pathogenic', 'likely pathogenic', 'conflicting interpretations of pathogenicity', and 'uncertain significance' were filtered out. To perform this filtering, we developed a Python program called 1D-to-3D.py, which processes variants from csv files downloaded directly from gnomAD (1) (further details in "3D Visualization"). After filtering, 7,540 variants were used for further analyses. All raw and processed gnomAD data can be accessed in supporting information Supplementary Table 2 and Supplementary Table 3 respectively.

1D Plots and the Vd/Vp Ratio

Filtered missense variants were mapped onto protein sequences of ARID family members using Plot Protein (55). The Plot Protein R script was modified to allow for color manipulation and domain diagram alterations of the output graphs in Inkscape (56). Vd/Vp ratios of functional domains were calculated using the following formula:

$$\frac{Vd}{Vp} = \frac{number\ of\ variable\ residue\ positions\ in\ domain}{total\ number\ of\ residues\ in\ domain} \Big/ \frac{number\ of\ variable\ residue\ positions\ in\ protein}{total\ number\ of\ residues\ in\ protein}$$

To automate the process, we developed a Python program called VdVp_Calculator.py. Like 1D-to-3D, the Vd/Vp Calculator processes csv files downloaded directly from gnomAD (1). It requires a user-defined text file with domain boundaries and calculates the Vd/Vp ratios of all functional domains in the protein of interest. The program script and user instructions are accessible at GitLab (https://git.ecdf.ed.ac.uk/cooklab/deak).

3D Visualization

To visualize missense variation in 3D, the 1D-to-3D program uses filtered data from gnomAD (1) and generates a PyMOL script that maps the variants onto protein structures. The variants appear as spheres at the Cα of the associated residue and increase in size and shade of blue with increasing AF. In the case of multiallelic sites, the program applies the addition rule for disjoint events, i.e. if multiple variants occur at the same residue position, their AFs are summed. The AF values for each position are compressed using a base 10 log scale and the positions are sorted into 6 bins (AF <10^{-6}-10^{-5}, 10^{-5}-10^{-4}...10^{-1}-10^{0}). Variants in each bin are visualized as spheres of different size and shade of blue. Further details regarding user inputs and numerical handling can be found in the user instructions and program script, accessible at GitLab (https://git.ecdf.ed.ac.uk/cooklab/deak). We used the 1D-to-3D program to annotate 11 solved and 6 modelled structures of the ARID family members with missense variants. A list of these structures, PyMOL selection names, and start/end residues can be found in Supplementary Table 1. The PyMOL scripts can be accessed at: https://doi.org/10.7488/ds/3128.

Pathogenic Variants

A family-wide search for pathogenic missense variants was performed using ClinVar (RRID:SCR_006169) and DECIPHER (RRID:SCR_006552), two publicly-accessible databases of clinical variants and their phenotypes (57, 58). Using ClinVar, we identified variants with the search criteria 'missense' and 'pathogenic' or 'likely pathogenic.' In DECIPHER, we searched for 'research variants' from the Deciphering Developmental Disorders Study, which collected variants from ~14,000 UK children with undiagnosed developmental disorders (57). All variant accession codes are available in Supplementary Table 1.

Results and Discussion

Genetic Constraint in the ARID Family

As a preliminary assessment of variation in the ARID family, we collated existing constraint metrics for each member from gnomAD. These included “loss-of-function observed/expected upper bound fraction” (LOEUF) (1) and missense Z (mZ) scores (Fig. 2A) as well as pLi and RVIS scores (Supplementary Table 1) (1, 54). The lower the LOEUF, the fewer the variants observed than expected, indicating negative selection against LOF variation. All ARID family genes except ARID3B/3C and JARID1B/1D can be classified as LOF-intolerant (Fig. 2A). This indicates that they are subject to strong purifying selection and are statistically likely to have disease associations, a higher number of protein-protein interaction partners and broad tissue expression (1). Intolerance to variants in the ARID family is also supported by pLI and RVIS scores (Supplementary Table 1).

The mZ score represents the deviation of the observed from the expected number of missense variants (single amino acid substitutions), for a given gene, where positive scores indicate missense depletion and negative scores indicate missense enrichment (2). ARID1A and JARID1C have mZ P-values of < 0.001 and nearly all members have positive scores, indicating that specific residue positions are also under selective constraint (Fig. 2A).

We calculated Vp values that denote the average density of missense variants for each protein in our dataset. As expected, lower Vp values correlate with higher mZ scores (Fig. 2B). However, JARID1C and JARID1D do not fit the observed correlation between mZ and V_p values (Fig. 2B). Rather than missense depletion, this is likely to arise from low data availability because JARID1C and JARID1D are encoded on the X and Y chromosomes respectively (1, 59). This analysis suggests that Vp is a good proxy for mZ and that values outside of the range of 0.29-0.42 may indicate when insufficient data are available for analysis. We excluded JARID1C and JARID1D from subsequent analyses.

These constraint metrics demonstrate that individual ARID family members are under significant selective pressure, yet they are less informative on missense depletion at the level of domains or smaller functional regions (Fig. 2A). To bridge this gap, we developed the '1D-to-3D' approach (Fig. 1A). In '1D', primary protein sequences are annotated with neutral or pathogenic missense variants and Vd/Vp ratios are calculated to compare the linear distribution of missense variants in functional domains. In '3D', the missense variants are mapped onto solved or modelled protein structures. This allows integration of the AFs of variants at each residue position, with visual discrimination of AFs using increasing sphere size and shade of blue (Fig. 1A). In line with larger-scale analyses (7, 15), we hypothesized that the annotated structures would reveal functionally important 3D sites depleted of neutral variants and enriched in pathogenic variants.

Validating the 1D-to-3D Approach on Known ARID Complex Structures

To verify that the 1D-to-3D approach can be used to identify sites depleted of missense variants for this family, we analyzed the structurally well-characterized ARID1 and JARID1 subfamilies. The ARID1 subfamily comprises ARID1A and ARID1B, two vertebrate paralogs with identical domain architecture (Fig. 2A) and 57 % sequence identity in humans. ARID1A and ARID1B are mutually-exclusive core subunits of the BAF chromatin remodeling complex (23). BAF complexes modulate transcription through ATPase-dependent nucleosome sliding/ejection or the recruitment of other regulators (44). They comprise three modules: a catalytic ATPase (BRG1/BRM), actin-related protein (ARP), and a base module that scaffolds the ATPase, nucleosome, and other regulators (Fig. 3A) (44). ARID1A or ARID1B are incorporated into the base module through a C-terminal Core Binding Region (CBR) made up of conserved CBRA and CBRB segments connected by intrinsically disordered loops (30).

Mutations in the BAF complex promote tumorigenesis in multiple cancer types (reviewed in (20)). A recent analysis of missense cancer mutations in the BAF complex revealed that several mutations cluster at a junction between the ARID1A/B CBR and the helicase-SANT-associated (HSA) helix of the ATPase (Fig. 3A) (60). We find that the HSA-CBR junction is depleted of neutral missense variants (Fig. 3B), and this correlates well with evolutionary surface conservation of the CBR in vertebrates (Fig. 3C). We also found that pathogenic variants associated with Coffin-Siris Syndrome and non-syndromic intellectual disability variants map in proximity to the HSA-CBR junction (Fig. 3D). ARID1B is ranked as a top diagnostic gene for both Coffin-Siris syndrome and non-syndromic intellectual disabilities (61). Similar results were obtained for ARID2 (Fig. S1), which is not paralogous to ARID1A and ARID1B, yet plays a functionally analogous role in the related Polybromo-associated BAF (PBAF) nucleosome remodeling complex (30). The distribution of missense variants therefore serves as a valuable, additional layer of information for investigating key protein-protein interaction interfaces in multi-subunit complexes.

Next, we tested whether our approach could identify the catalytic site in members of the JARID1 subfamily. JARID1A/1B are Fe(II)- and 2-oxoglutarate (2-OG)-dependent dioxygenases that catalyse the demethylation of di- or tri-methylated histone H3K4 via their JmjC domain (Fig. 3E) (62). We report that the catalytic site in the JmjC domain is visibly missense depleted (Fig. 3F), correlating with evolutionary conservation (Fig. 3G). Similar but less pronounced results are observed for JARID1B (Fig. S2), consistent with the higher LOEUF and lower mZ metrics reported for JARID1B (Fig. 2A).

Apart from the catalytic site, depletion of missense variants in JARID1A is also observed in a groove between the ARID and C5HC2 domains, which is thought to accommodate double-stranded DNA (Fig. 3F; Movie S1) (45). The variants exhibit clear spatial surface segregation, with a lower abundance of variants on the face containing the catalytic site and higher abundance on the far surface of the enzyme (Movie S1). These findings validate our approach and demonstrate that missense variants complement the use of conservation data to identify surfaces involved in macromolecular interactions and enzyme active sites.

Comparative Analysis of Domains Using the Vd/Vp Ratio

To further investigate the utility of missense variants at the domain level, we compared JARID1A and JARID1B. In addition to the JmjN-JmjC, ARID, and C5HC2 zinc finger domains, they also comprise three PHD fingers (Fig. 4A-B). Mapping missense variants on JARID1A and JARID1B sequences and calculating the Vd/Vp ratios for each domain revealed differences in missense variant depletion in each PHD finger (Fig. 4B). These differences may indicate paralog sub-functionalization at the level of individual domains. In JARID1A, PHD3 but not PHD1 is missense depleted, while in JARID1B, PHD1 but not PHD3 is missense depleted. Neither protein shows missense depletion in PHD2 (Fig. 4B).

In JARID1A and JARID1B, PHD1 preferentially binds to unmethylated histone H3K4, leading to increased affinity of the JmjC domain for its methylated H3K4 substrates (63-65). In contrast, PHD3 is thought to be a reader of tri-methylated H3K4 marks (66, 67). These two domains tolerate sequence variation, suggesting that their ability to bind histone tails is not crucial for function. In particular, residues D292/Y294/L308 in JARID1A and W1502/W512 in JARID1B, which are required for histone peptide binding (67, 68), are affected by neutral missense variants (Fig. 4C-D). Conversely, residues responsible for histone peptide binding in PHD3 of JARID1A (Fig. 4C) and PHD1 of JARID1B (Fig. 4D) are under selective constraint, indicating that they are important in targeting or enhancing the activity of the JARID1 enzymes at their appropriate genomic locations. Consistent with their functional importance, PHD3 in JARID1A was shown to be critical in driving the oncogenic effects of a JARID1A-NUP98 fusion protein in acute myeloid leukemia (66). Furthermore, mutations in PHD1, but not PHD3 of JARID1B decreased the regulatory effects on cell migration in a model of triple negative breast cancer (67). In summary, functional differences in the PHD fingers correlate with different Vd/Vp ratio patterns in JARID1A and JARID1B, suggesting that the differences missense variant depletion we observe represent domain-level functional differences between the two paralogs.

Limitations of the Vd/Vp Ratio

Despite its utility for comparing domains, it should be noted that Vd/Vp ratios are dependent on the user's definition of domain boundaries. Where domain boundaries are not clearly defined, it is possible that Vd/Vp ratios might be over- or under-estimated. Furthermore, small sites that are missense depleted may also be overlooked. For example, the Vd/Vp ratios calculated for the chromobarrel domain of the ARID4 subfamily proteins are relatively high, yet mapping missense variants in 3D reveals depletion of variants in a conserved Tyr-Tyr-Trp-Tyr aromatic cage (Fig. 5).

The ARID4 subfamily comprises ARID4A and ARID4B (Fig. 2A), two paralogous, multi-domain adapter proteins that recruit transcriptional regulators such as the retinoblastoma protein, androgen receptor, and the mSin3A histone deacetylase complex to gene promoters (34, 69). Given the presence of an aromatic cage, the chromobarrel domain was hypothesized to bind histone methyl marks (33). However, independent NMR and ITC titration experiments with the ARID4A chromobarrel domain and methylated histone peptides produced conflicting results (33, 70).

Mapping of missense variants shows that residues that form the aromatic cage are depleted of missense variants in both the solved structure of the ARID4A chromobarrel domain (Fig. 5A) and a model of the ARID4B chromobarrel domain (Fig. 5B). This suggests that methyl-lysine recognition is intact. This observation highlights the importance of mapping missense variants onto 3D structures, especially in the case of small domains, where the Vd/Vp ratio may not provide sufficient resolution to detect functionally important sites.

Comparison of Missense Depletion Between Paralogs Indicates Sites of Sub-Functionalization

ARID4A and ARID4B have diverged only in vertebrate lineages (41) and share identical domain architecture (Fig. 2A). They participate in the same molecular pathways, including the recruitment of the mSin3A repressive complex to gene promoters (71) and co-activation of the androgen receptor in regulation of male fertility (69). However, ARID4A knockout mice are viable whereas ARID4B knockouts show early embryonic lethality (72). Moreover, ARID4B is necessary for spermatogenesis while ARID4A is not (69, 73). These findings indicate that of the two paralogs, ARID4B is likely a more critical determinant of cell fate decisions and cell cycle progression.

Since all domains of ARID4A and ARID4B are structurally related, we investigated if missense depletion could reveal functional differences between the two paralogs. We found domain-wide missense depletion in 1D to be more pronounced in ARID4B (Fig. S3) and observed a difference in the 3D distribution of missense variants on their hybrid Tudor domains (HTDs). The HTDs comprise two sub-domains HTD-1 and HTD-2 (Fig. 6A). Unlike structurally analogous Tudor domains, ARID4 HTDs lack an aromatic cage associated with histone binding in HTD-2 and instead are reported to exhibit DNA binding through HTD-1 (74, 75) (Fig. 6A). A conserved, structurally important glycine of ARID4B is associated with a developmental disorder variant (Fig. 6A).

We find that HTD-1 of ARID4B is missense depleted, while HTD-1 of ARID4A is not (Fig. 6B). The depleted site corresponds to previously identified DNA-binding residues (Fig. 6B) and a positively-charged DNA-binding surface of the domain (Fig. 6C)(75). We also note that an RGR motif, recently found to enhance the DNA-binding affinity of ARID4B (75), tolerates missense variation (Fig. 6A). Overall, our findings indicate that the ARID4B Tudor domain likely contributes to the functional differences between ARID4A and ARID4B.

We next investigated if missense variants could also reveal paralog sub-functionalization in the ARID3 subfamily. The ARID3 proteins are transcription factors comprising the ARID domain and a C-terminal oligomerization domain called REKLES (Fig. 2A). The ARID domain binds to AT-rich promoter sequences and is essential for ARID3 protein function (31, 76). In *Drosophila* and mice, knock out of the ARID3 orthologs, Dead ringer or Bright, respectively, is lethal (31, 77). In humans, the ARID3 subfamily has three members, where the ARID3A ARID domain shares 70 % sequence identity with the ARID domain of ARID3B and 87% identity with the ARID domain of ARID3C. Given this high degree of sequence identity, we hypothesized that differences in missense variation likely reflect paralog sub-functionalization.

The ARID domain is built from six core helices (H1-H6) and two larger loops L1 (between helices H1-H2) and L2 (between helices H4-H5) (Fig.7A). A solution structure of the *Drosophila* Dead ringer in complex with DNA (78) and NMR titration experiments of the ARID1A (79), ARID5B (80), and JARID1A (81) ARID domains suggest a general mechanism of binding. This includes non-sequence specific contacts with the phosphate backbone by L1 and sequence-specific contacts with the major groove by a non-canonical helix-turn-helix motif formed by H4-L2-H5 (78-81)(Fig. 7B).

The Vd/Vp ratios of ARID domains of the three human ARID3 paralogs are lowest in ARID3A and highest in ARID3C (Fig. 7C-E). In ARID3A, the proposed DNA binding residues in L1 and H4-L2-L5, inferred from sequence conservation with Dead ringer, are clear of missense variants. The domain also harbors a developmental disorder variant (57), supporting its functional importance (Fig. 7C). ARID3B has a higher Vd/Vp ratio and some missense variants map to residues typically involved in DNA binding in the ARID family. This suggests that DNA binding activity is less likely to be of functional importance in this paralog (Fig. 7D). Similarly, in ARID3C, several residues typically involved in DNA-binding are found to have reported variants (Fig. 7E). We also note that the Vd/Vp ratios of all three REKLES domains are > 1.00 (Fig. S4). While this suggests that the oligomerization is less likely to be required for ARID3 activity, these domains are short motifs with no available structural data, so may not be suitable for this analysis. Missense variation can therefore also be leveraged to filter structurally similar domains in paralogs for functional importance.

The ARID5B BAH Domain Shows Likely Loss and Gain of Function

Finally, we used 1D-to-3D to give insights on novel structure-guided functional predictions in ARID5B. ARID5B is a highly constrained gene (Fig. 2A) and a key regulator of liver metabolism, chondrogenesis, and adipogenesis (32, 82, 83). ARID5B is known to target the H3K9me2 demethylase PHF2 to gene promoters via its ARID domain (25, 32).

ARID5B has two isoforms, 1 and 2 (Fig. 8A). In *Xenopus,* isoform expression is spatially segregated during embryonic development, where isoform 1 shows higher abundance than isoform 2 (84). Isoform 1 also shows higher expression in healthy adult human tissues (53). Isoform1 has an additional N-terminal region that is highly conserved in a subset of vertebrate species (Fig. S5) and predicted to form a bromo-adjacent homology (BAH) domain (42); this likely defines the functional differences between the isoforms. We compared an AlphaFold model of the ARID5B BAH domain to experimentally determined BAH domain structures and used 1D-to-3D to map missense variants onto the domain.

We identified bovine DNMT1, mouse BAHCC1, and mouse ORC1 BAH domains as the closest structural homologs to ARID5B (Fig. S6). All three homologs read histone methyl marks (Fig. 8B) (85-87). The lower lobe of each BAH fold contains a conserved aromatic cage and acidic residues that bind to methylated lysine through cation-pi and electrostatic/hydrogen bonding interactions, respectively (Fig. 8C). The lower lobe of the ARID5B BAH domain does not have an aromatic cage: one aromatic residue, F77, is present but two positions normally occupied by aromatic residues are replaced with small hydrophobic residues (C53 and L75). Two acidic residues, D81 and E100, are present but both E100 and the aliphatic L75 tolerate missense variation (Fig. 8C). Collectively, these amino acids changes, compared with classic BAH domains, and their tolerance of variation suggest that ARID5B BAH domain is unlikely to bind methyl-lysine marks at this site.

The AlphaFold prediction for the ARID5B BAH domain differs from those in DNMT1, BAHCC1 and ORC1 in that an additional, conserved segment C-terminal to the BAH domain forms part of the fold of this domain (Fig. 8D). When mapped to this extended AlphaFold model, we note that the missense variants are depleted on the highly conserved, positively-charged C-terminal helix, rather than the classic peptide binding interface of BAH domains. This suggests that the C-terminal domain extension could serve as a protein-protein or protein-nucleic acid interaction module within a chromatin context (Fig. 8E). These predictions call for experimental evidence. However, our analysis demonstrates the power of missense variation in screening for functional features together with structural data.

Conclusion

We provide a convenient set of tools for mapping missense variants onto primary and tertiary structures of proteins. This approach allowed us to visually locate regions of proteins that are depleted of population variants, indicative of negative selection pressure. Using this approach, we demonstrated that mapping missense variants onto 3D structures in the context of a large family of proteins reveals functional insights. This method is complementary to phylogenetic conservation analysis and could be useful where insufficient phylogenetic data are available, for example when analyzing recent paralogs.

Using Vd/Vp ratios for individual domains may have a particular utility for researchers working on multidomain proteins where the goal is to identify which domains contribute essential functions. Ranking by Vd/Vp ratio could help prioritize which domains to delete in functional assays. This approach could also be useful for researchers seeking to provide minimal functional constructs of a protein for gene therapy approaches, where limiting the length of the protein, and therefore its coding sequence, can be critical for packaging into a virus (88-90).

Some limitations are noted. First, the sample size of the gnomAD dataset does not achieve mutational saturation (1). This sets limits on interpreting constraint in smaller domains/linear motifs and prevented us from analyzing proteins encoded on the sex chromosomes. Second, while our approach allows allele frequency to be visualized, it does not normalize for codon mutability, where nucleotide sequence composition skews missense enrichment in protein sequences. For example, methylated CpG dinucleotides are known to be hypermutable, resulting in an over-representation of variants in CpG rich and/or heavily methylated codons (91). Finally, some of our analyses were based on calculated models rather than experimentally-determined structures. As missense mapping depends on positional information, validation of these models is essential to confirm any interpretations.

Our findings build on previous studies showing that depletion of missense variants can serve to identify functionally important protein sites (7, 15). We demonstrate how 1D and 3D mapping approaches complement existing findings, provide context to understand the impact of pathogenic variants, functionally differentiate structurally similar domains in paralogs, and support formulation of novel mechanistic hypotheses. Although we focused on proteins with catalytic, DNA binding and epigenetic roles, our approach is applicable to a broad range of human protein functions.

Figure Legends

Figure 1: (A) Schematic of the 1D-to-3D approach of mapping missense variants onto protein structures. **(B)** Schematic illustration of the interactions of ARID family members (red) with nucleosomes (grey) and other chromatin-binding proteins (light blue).

Figure 2: Genetic constraint in the ARID family (A) An overview of constraint metrics, Vps, and domain architecture in the ARID family. LOEUF is the upper boundary of the 90% confidence interval of the observed/expected ratio of loss-of-function (LOF) variants in a given gene. The recommended threshold to segregate LOF-intolerant and LOF-tolerant genes is 0.35. The mZ scores outside of ±3.09 correspond to a recommended P-value threshold of 0.001. **(B)** Correlation between the mZ scores and Vps for this dateset (blue). Values for JARID1C (J1C) and JARID1D (J1D) are red.

Figure 3: Validation of the 1D-to-3D approach (A) Position of the ARID1A/1B CBR in the BAF complex**.** Solved structure of the ARID1A CBR (PDBid 6LTJ, chain L) and modelled structure of the ARID1B CBR annotated with missense variants **(B),** sequence conservation in vertebrates **(C)**, and pathogenic variants **(D)**. **(E)** Schematic illustration of the JARID1A/1B enzymes (based on the PDB structures 5CEH and 5FUP). Solved structure of JARID1A annotated with missense variants (ligand shown in red) **(F)** and sequence conservation in metazoa **(G).**

Figure 4: Comparative Analysis of Domains Using the Vd/Vp Ratio (A) Schematic diagram of the JARID1A and JARID1B enzymes (based on PDB structures 5CEH and 5FUP). **(B)** 1D plots of missense variants in JARID1A/1B and with Vd/Vp ratios calculated for their functional domains, shown above the plot. **(C)** JARID1A and **(D)** JARID1B PHD1 and PHD3 domains shown without (left) and with (right) missense variants; histone peptides are shown in black, peptide-binding residues in purple, and zinc ions in white.

Figure 5: Limitations of the Vd/Vp Ratio. The ARID4A **(A)** and ARID4B **(B)** chromobarrel domains shown without (left) and with (right) missense variants. Putative histone methyl mark-binding residues are indicated in purple.

Figure 6: The ARID4 Subfamily. The ARID4A and ARID4B hybrid Tudor domain (HTD) shown with DNA binding residues **(A)**, missense variants **(B)**, and surface electrostatics **(C)**.

Figure 7 The ARID3 Subfamily (A) Secondary structure of the ARID domain where helices are denoted with H and loops are denoted with L. The ARID3 subfamily has two additional flanking helices H0 and H7. **(B)** Solution structure of the Dead ringer ARID domain in complex with DNA. Human ARID3A **(C)**, ARID3B **(D)**, and ARID3C **(E)** ARID domains shown with missense variants.

Figure 8: Loss and gain of function in ARID5B (A) Domain architecture of the human ARID5 subfamily. **(B)** Variant-annotated model of the human ARID5B BAH domain compared to solved structures of the bovine DNMT1, mouse BAHCC1, and mouse ORC1 BAH domains (histone methyl mark-binding residues shown in purple). **(C)** A closer view of the methyl mark-binding sites (methylated histone peptides shown in black). **(D)** AlphaFold prediction of the ARID5B BAH domain with a C-terminal extension: two orientations, related by a 90 degree rotation, of the BAH domain

(red) and the extension (grey) are shown in the left panel. Corresponding missense variants are shown on the same models in the right panel. **(E)** Potential binding interactions of the ARID5B BAH domain.

Acknowledgements:

The authors would like to thank Sander Granneman and Philipp Voigt for critical reading of the manuscript. This work was supported by a Wellcome Trust with a Senior Research Fellowship to AGC (200898) and a Wellcome Centre core grant (203149). We thank Edward Wallace for help with phylogenetic analysis. We thank Hannah Wapenaar and Hayden Burdett for help with testing code.

Author contributions:

AGC conceived of and directed the research. GD executed research, developed metrics and wrote the code. GD and AGC co-wrote the MS.

Conflict of interest:

The authors declare that they have no conflicts of interest.

Data are available at:
University of Edinburgh GitLab: https://git.ecdf.ed.ac.uk/cooklab/deak
University of Edinburgh DataShare: https://doi.org/10.7488/ds/3128.

Figure 1

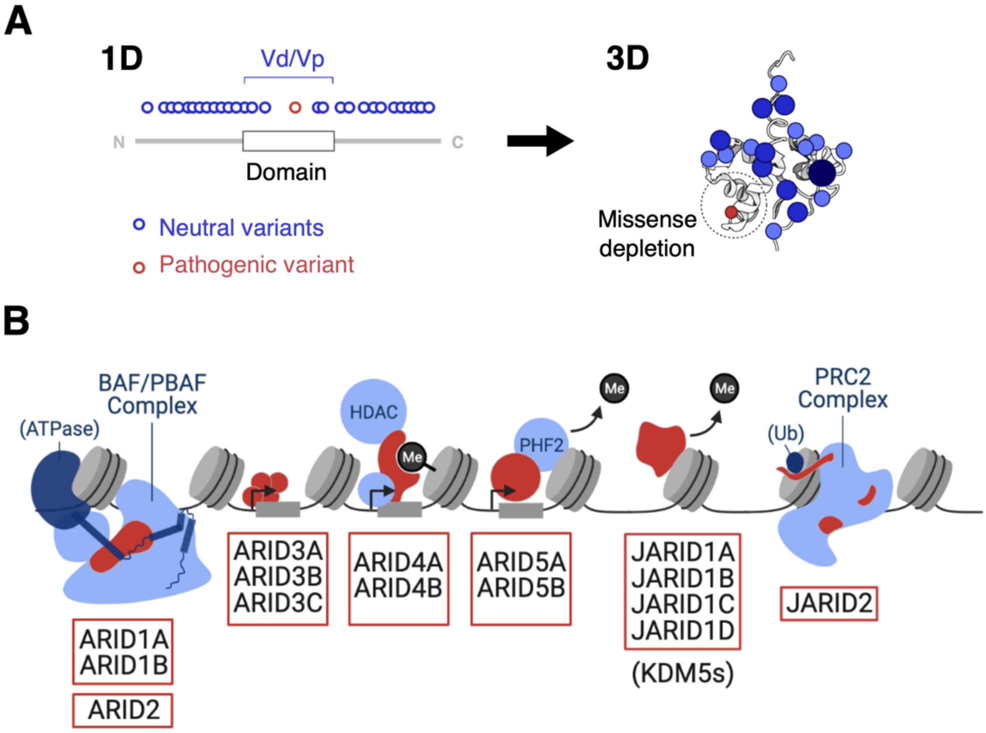
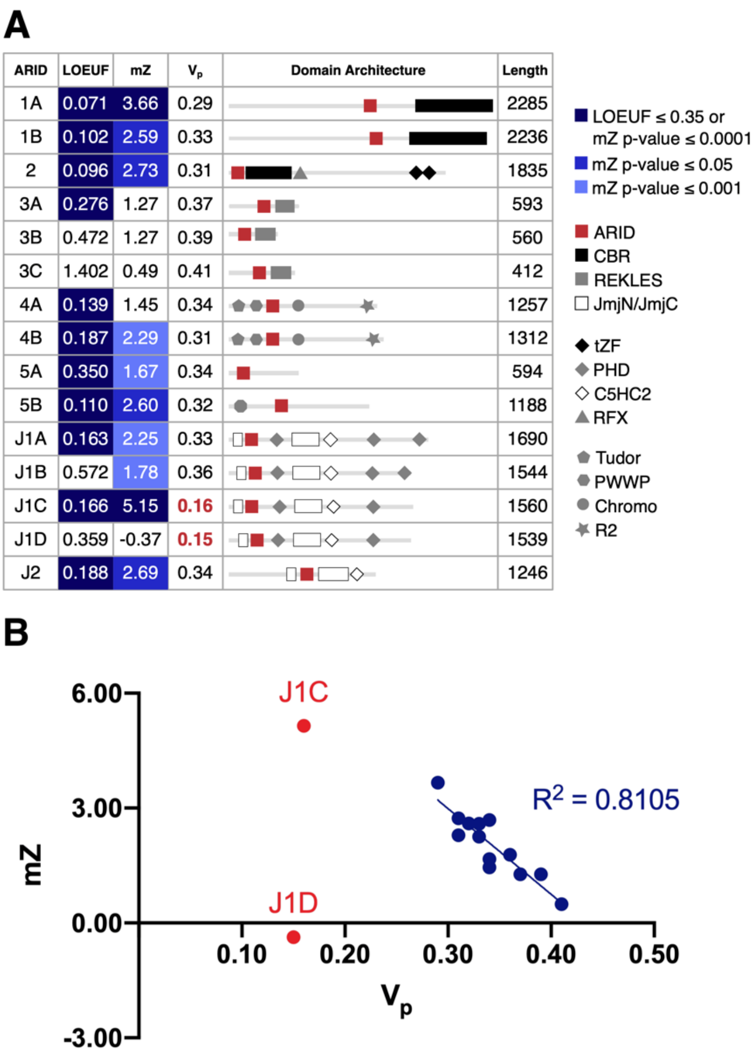

A
1D
Vd/Vp
N
C
Domain
○ Neutral variants
○ Pathogenic variant
3D
Missense depletion
B
BAF/PBAF Complex
(ATPase)
HDAC
Me
PHF2
Me
Me
PRC2 Complex
(Ub)
ARID1A
ARID1B
ARID2
ARID3A
ARID3B
ARID3C
ARID4A
ARID4B
ARID5A
ARID5B
JARID1A
JARID1B
JARID1C
JARID1D
(KDM5s)
JARID2

Figure 2



ARID	LOEUF	mZ	V_p	Domain Architecture	Length
1A	0.071	3.66	0.29		2285
1B	0.102	2.59	0.33		2236
2	0.096	2.73	0.31		1835
3A	0.276	1.27	0.37		593
3B	0.472	1.27	0.39		560
3C	1.402	0.49	0.41		412
4A	0.139	1.45	0.34		1257
4B	0.187	2.29	0.31		1312
5A	0.350	1.67	0.34		594
5B	0.110	2.60	0.32		1188
J1A	0.163	2.25	0.33		1690
J1B	0.572	1.78	0.36		1544
J1C	0.166	5.15	0.16		1560
J1D	0.359	-0.37	0.15		1539
J2	0.188	2.69	0.34		1246

Figure 3

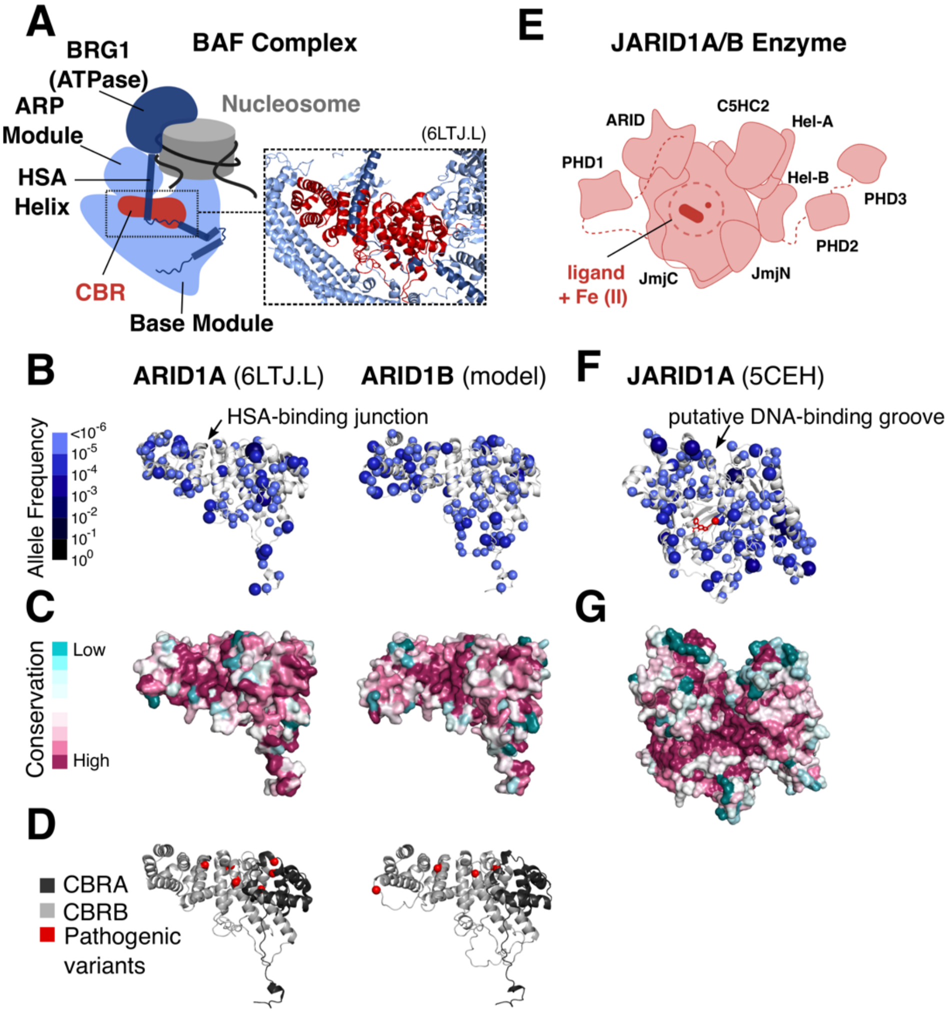

A
BRG1
(ATPase)
BAF Complex
ARP
Module
Nucleosome
(6LTJ.L)
HSA
Helix
CBR
Base Module
E
JARID1A/B Enzyme
ARID
C5HC2
Hel-A
PHD1
Hel-B
PHD3
PHD2
ligand
+ Fe (II)
JmjC
JmjN
B
ARID1A (6LTJ.L)
ARID1B (model)
F
JARID1A (5CEH)
HSA-binding junction
putative DNA-binding groove
Allele Frequency
<10^-6
10^-5
10^-4
10^-3
10^-2
10^-1
10^0
C
G
Conservation
Low
High
D
CBRA
CBRB
Pathogenic
variants

Figure 4

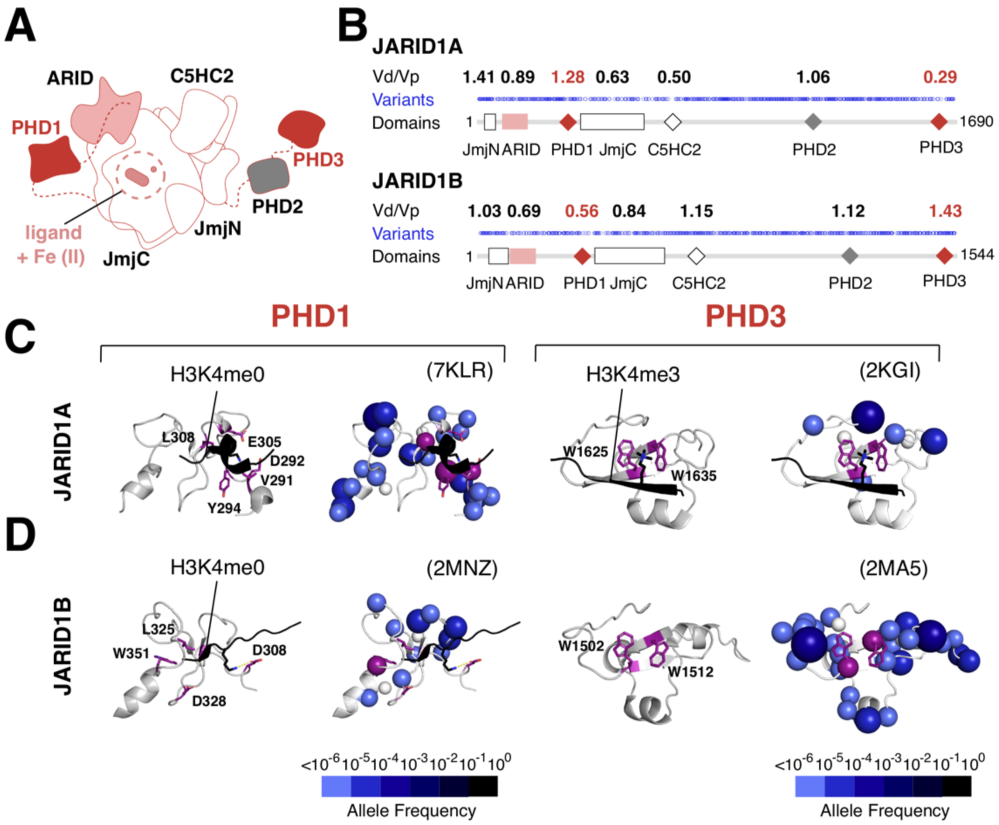

A
ARID
C5HC2
PHD1
PHD3
PHD2
ligand
+ Fe (II)
JmjN
JmjC
B
JARID1A
Vd/Vp 1.41 0.89 1.28 0.63 0.50 1.06 0.29
Variants
Domains 1 1690
JmjN ARID PHD1 JmjC C5HC2 PHD2 PHD3
JARID1B
Vd/Vp 1.03 0.69 0.56 0.84 1.15 1.12 1.43
Variants
Domains 1 1544
JmjN ARID PHD1 JmjC C5HC2 PHD2 PHD3
C
PHD1
PHD3
H3K4me0
(7KLR)
H3K4me3
(2KGI)
JARID1A
L308
E305
D292
V291
Y294
W1625
W1635
D
H3K4me0
(2MNZ)
(2MA5)
JARID1B
L325
W351
D308
D328
W1502
W1512
<10^-6 10^-5 10^-4 10^-3 10^-2 10^-1 10^0
Allele Frequency
<10^-6 10^-5 10^-4 10^-3 10^-2 10^-1 10^0
Allele Frequency

Figure 5

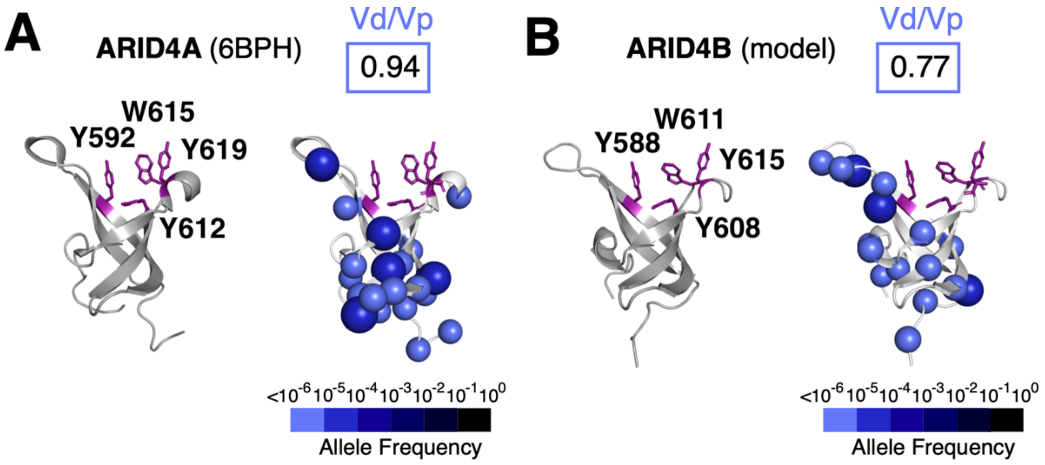

A
ARID4A (6BPH)
Vd/Vp
0.94
W615
Y592
Y619
Y612
<10^-6 10^-5 10^-4 10^-3 10^-2 10^-1 10^0
Allele Frequency
B
ARID4B (model)
Vd/Vp
0.77
W611
Y588
Y615
Y608
<10^-6 10^-5 10^-4 10^-3 10^-2 10^-1 10^0
Allele Frequency

Figure 6

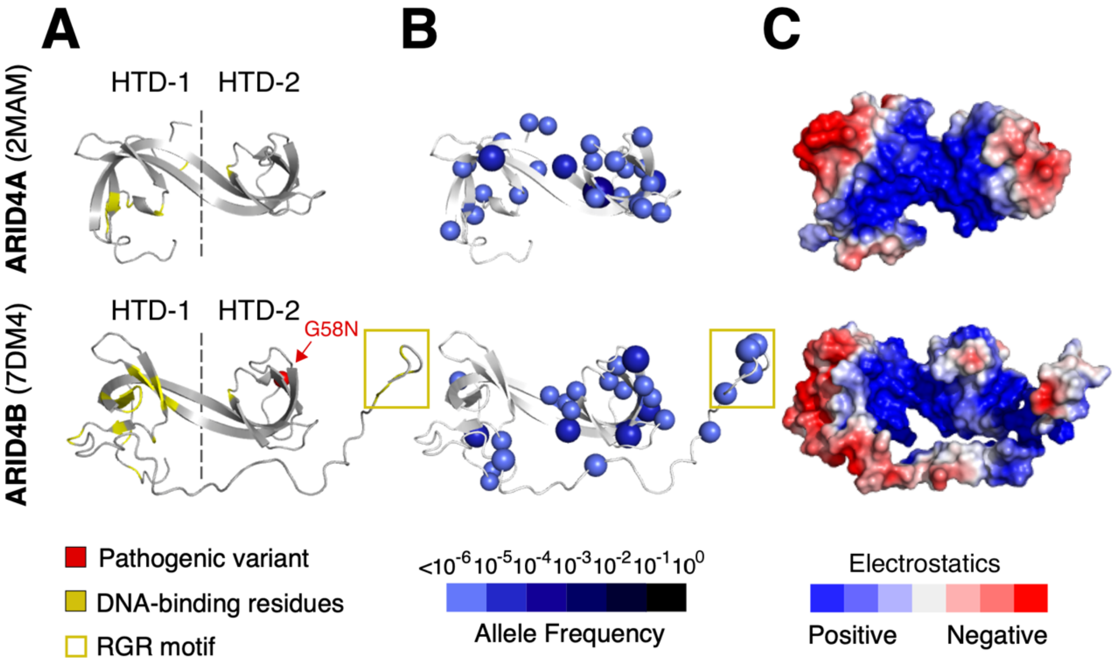

A
B
C
ARID4A (2MAM)
ARID4B (7DM4)
HTD-1
HTD-2
HTD-1
HTD-2
G58N
🟥 Pathogenic variant
🟨 DNA-binding residues
□ RGR motif
<10^-6 10^-5 10^-4 10^-3 10^-2 10^-1 10^0
Allele Frequency
Electrostatics
Positive
Negative

Figure 7

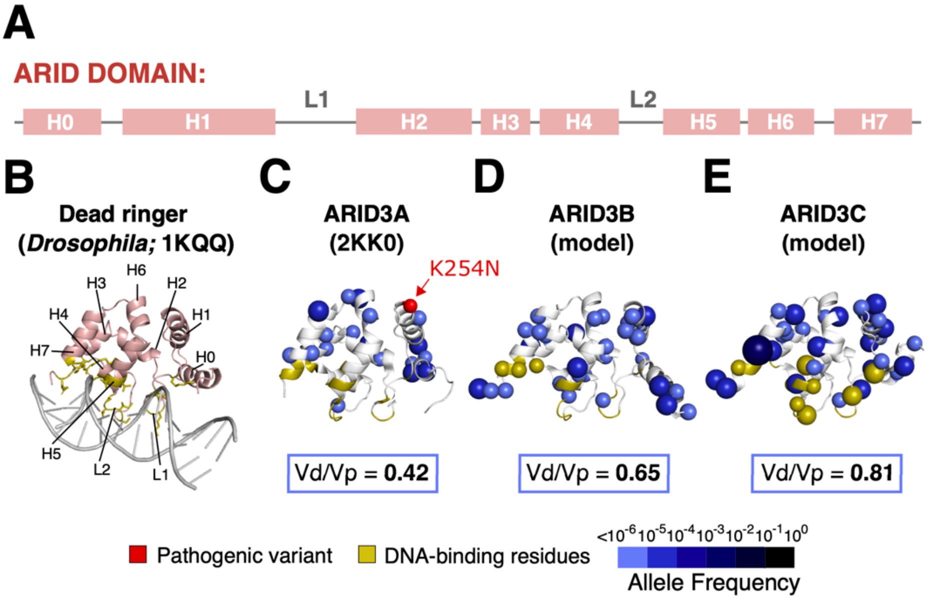
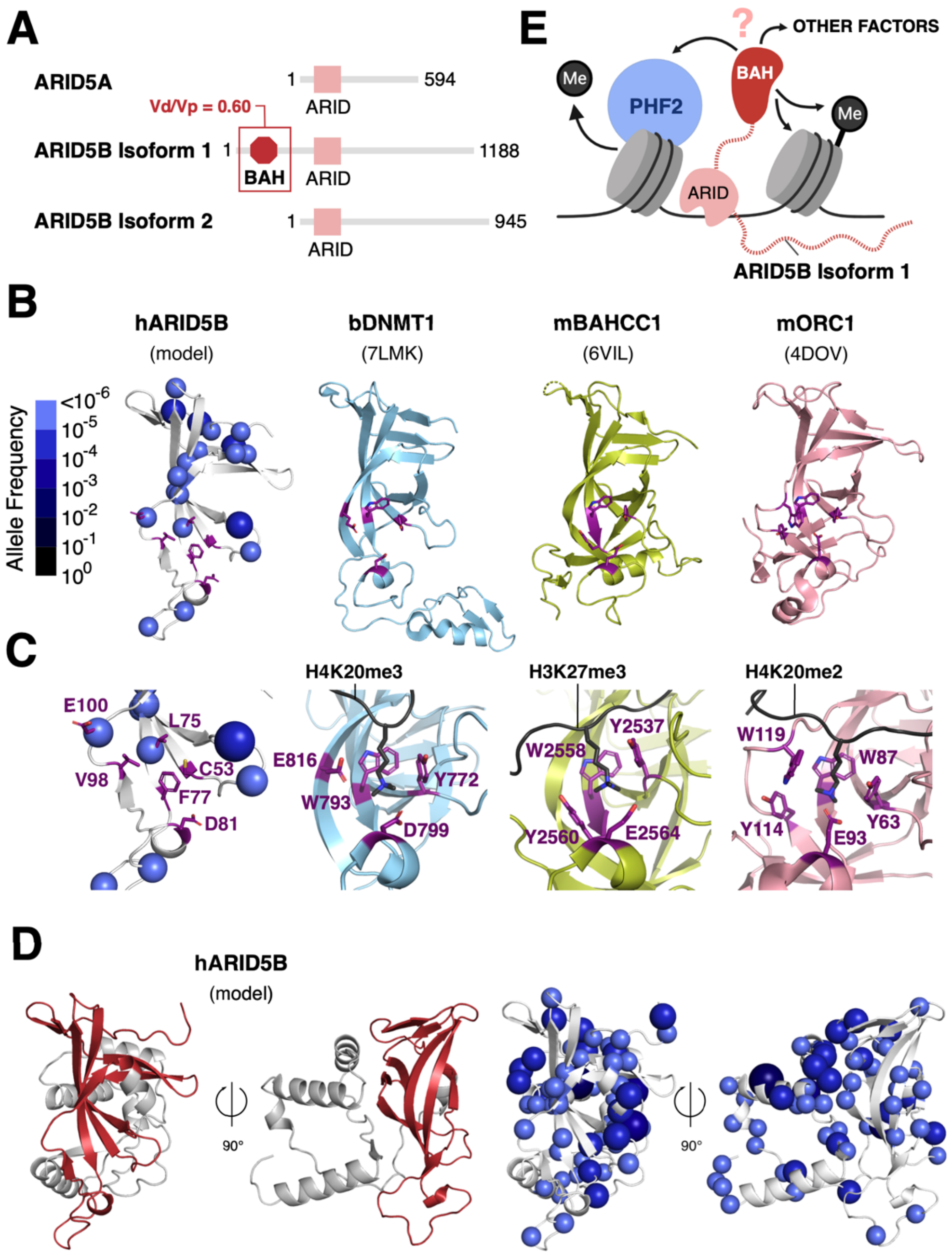

A
ARID DOMAIN:
H0
H1
L1
H2
H3
H4
L2
H5
H6
H7
B
Dead ringer
(*Drosophila;* 1KQQ)
H6
H3
H2
H4
H1
H7
H0
H5
L2
L1
C
ARID3A
(2KK0)
K254N
Vd/Vp = **0.42**
D
ARID3B
(model)
Vd/Vp = **0.65**
E
ARID3C
(model)
Vd/Vp = **0.81**
🟥 Pathogenic variant
🟨 DNA-binding residues
<10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^{0}
Allele Frequency

Figure 8


A
ARID5A
1
594
ARID
Vd/Vp = 0.60
ARID5B Isoform 1
1
BAH
ARID
1188
ARID5B Isoform 2
1
945
ARID
E
?
OTHER FACTORS
Me
PHF2
BAH
Me
ARID
ARID5B Isoform 1
B
hARID5B
(model)
bDNMT1
(7LMK)
mBAHCC1
(6VIL)
mORC1
(4DOV)
Allele Frequency
<10^-6
10^-5
10^-4
10^-3
10^-2
10^-1
10^0
C
E100
L75
C53
V98
F77
D81
H4K20me3
E816
Y772
W793
D799
H3K27me3
Y2537
W2558
Y2560
E2564
H4K20me2
W119
W87
Y114
E93
Y63
D
hARID5B
(model)
90°
90°

References

1. K. J. Karczewski *et al.*, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
2. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).
3. K. E. Samocha *et al.*, A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-950 (2014).
4. R. Walsh *et al.*, Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med* **19**, 192-203 (2017).
5. E. V. Minikel *et al.*, Evaluating drug targets through human loss-of-function genetic variation. *Nature* **581**, 459-464 (2020).
6. N. Whiffin *et al.*, The effect of LRRK2 loss-of-function variants in humans. *Nat Med* **26**, 869-877 (2020).
7. A. MacGowan *et al.*, Human Missense Variation is Constrained by Domain Structure and Highlights Functional and Pathogenic Residues. *bioRxiv* **127050** (2017).
8. J. M. Havrilla, B. S. Pedersen, R. M. Layer, A. R. Quinlan, A map of constrained coding regions in the human genome. *Nat Genet* **51**, 88-95 (2019).
9. K. E. Samocha *et al.*, Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* (2017).
10. L. Brenan *et al.*, Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. *Cell Rep* **17**, 1171-1183 (2016).
11. A. R. Majithia *et al.*, Prospective functional classification of all possible missense variants in PPARG. *Nat Genet* **48**, 1570-1575 (2016).
12. A. B. Gussow, S. Petrovski, Q. Wang, A. S. Allen, D. B. Goldstein, The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol* **17**, 9 (2016).
13. S. Iqbal *et al.*, Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc Natl Acad Sci U S A* **117**, 28201-28211 (2020).
14. D. Bai *et al.*, Differential Domain Distribution of gnomAD- and Disease-Linked Connexin Missense Variants. *International Journal of Molecular Sciences* **22**, 7832 (2021).
15. M. Hicks, I. Bartha, J. di Iulio, J. C. Venter, A. Telenti, Functional characterization of 3D protein structures informed by human genetic diversity. *Proc Natl Acad Sci U S A* **116**, 8960-8965 (2019).
16. Z. Z. Tang *et al.*, PSCAN: Spatial scan tests guided by protein structures improve complex disease gene discovery and signal variant detection. *Genome Biol* **21**, 217 (2020).
17. R. M. Sivley, X. Dou, J. Meiler, W. S. Bush, J. A. Capra, Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures. *Am J Hum Genet* **102**, 415-426 (2018).
18. S. Iqbal *et al.*, MISCAST: MIssense variant to protein StruCture Analysis web SuiTe. *Nucleic Acids Res* **48**, W132-W139 (2020).
19. M. H. Bailey *et al.*, Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034-1035 (2018).

20. P. Mittal, C. W. M. Roberts, The SWI/SNF complex in cancer - biology, biomarkers and therapy. *Nat Rev Clin Oncol* **17**, 435-448 (2020).
21. J. Plch, J. Hrabeta, T. Eckschlager, KDM5 demethylases and their role in cancer cell chemoresistance. *Int J Cancer* **144**, 221-231 (2019).
22. R. C. Wu *et al.*, Identification of the PTEN-ARID4B-PI3K pathway reveals the dependency on ARID4B by PTEN-deficient prostate cancer. *Nat Commun* **10**, 4332 (2019).
23. R. C. Centore, G. J. Sandoval, L. M. M. Soares, C. Kadoch, H. M. Chan, Mammalian SWI/SNF Chromatin Remodeling Complexes: Emerging Mechanisms and Therapeutic Strategies. *Trends Genet* **36**, 936-950 (2020).
24. N. Bogershausen, B. Wollnik, Mutational Landscapes and Phenotypic Spectrum of SWI/SNF-Related Intellectual Disability Disorders. *Front Mol Neurosci* **11**, 252 (2018).
25. J. Garton, M. D. Barron, M. L. Ratliff, C. F. Webb, New Frontiers: ARID3a in SLE. *Cells* **8** (2019).
26. Y. Liu *et al.*, Blood monocyte transcriptome and epigenome analyses reveal loci associated with human atherosclerosis. *Nat Commun* **8**, 393 (2017).
27. G. Wang *et al.*, Associations of variations in the MRF2/ARID5B gene with susceptibility to type 2 diabetes in the Japanese population. *J Hum Genet* **57**, 727-733 (2012).
28. D. Wilsker *et al.*, Nomenclature of the ARID family of DNA-binding proteins. *Genomics* **86**, 242-251 (2005).
29. V. Kasinath *et al.*, JARID2 and AEBP2 regulate PRC2 in the presence of H2AK119ub1 and other histone modifications. *Science* **371** (2021).
30. N. Mashtalir *et al.*, Modular Organization and Assembly of SWI/SNF Family Chromatin Remodeling Complexes. *Cell* **175**, 1272-1288 e1220 (2018).
31. T. Shandala, R. D. Kortschak, R. Saint, The Drosophila retained/dead ringer gene and ARID gene family function during development. *Int J Dev Biol* **46**, 423-430 (2002).
32. A. Baba *et al.*, PKA-dependent regulation of the histone lysine demethylase complex PHF2-ARID5B. *Nat Cell Biol* **13**, 668-675 (2011).
33. W. Gong *et al.*, Structural insight into recognition of methylated histone tails by retinoblastoma-binding protein 1. *J Biol Chem* **287**, 8531-8540 (2012).
34. A. Lai *et al.*, RBP1 recruits the mSIN3-histone deacetylase complex to the pocket of retinoblastoma tumor suppressor family proteins found in limited discrete regions of the nucleus at growth arrest. *Mol Cell Biol* **21**, 2918-2932 (2001).
35. N. Amatya *et al.*, IL-17 integrates multiple self-reinforcing, feed-forward mechanisms through the RNA binding protein Arid5a. *Sci Signal* **11** (2018).
36. K. Masuda *et al.*, Arid5a controls IL-6 mRNA stability, which contributes to elevation of IL-6 level in vivo. *Proc Natl Acad Sci U S A* **110**, 9409-9414 (2013).
37. A. M. Altenhoff *et al.*, OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res* **49**, D373-D379 (2021).
38. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).
39. S. B. Needleman, C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-453 (1970).
40. A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, G. J. Barton, Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191 (2009).

41. H. Li *et al.*, TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* **34**, D572-580 (2006).
42. L. Holm, DALI and the persistence of protein shape. *Protein Sci* **29**, 128-140 (2020).
43. M. Blum *et al.*, The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, D344-D354 (2021).
44. S. He *et al.*, Structure of nucleosome-bound human BAF complex. *Science* **367**, 875-881 (2020).
45. M. Vinogradova *et al.*, An inhibitor of KDM5 demethylases reduces survival of drug-tolerant cancer cells. *Nat Chem Biol* **12**, 531-538 (2016).
46. C. Johansson *et al.*, Structural analysis of human KDM5B guides histone demethylase inhibitor development. *Nat Chem Biol* **12**, 539-545 (2016).
47. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* 10.1038/s41586-021-03819-2 (2021).
48. K. Tunyasuvunakool *et al.*, Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590-596 (2021).
49. Anonymous (The PyMOL Molecular Graphics System. (Schrödinger, LLC).
50. E. Jurrus *et al.*, Improvements to the APBS biomolecular solvation software suite. *Protein Sci* **27**, 112-128 (2018).
51. H. Ashkenazy *et al.*, ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **44**, W344-350 (2016).
52. M. Zahn-Zabal, C. Dessimoz, N. M. Glover, Identifying orthologs with OMA: A primer. *F1000Res* **9**, 27 (2020).
53. I. S. Yang, H. Son, S. Kim, S. Kim, ISOexpresso: a web-based platform for isoform-level expression analysis in human cancer. *BMC Genomics* **17**, 631 (2016).
54. S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, D. B. Goldstein, Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9**, e1003709 (2013).
55. T. Turner, Plot protein: visualization of mutations. *J Clin Bioinforma* **3**, 14 (2013).
56. Anonymous (Inkscape Project v.1.0.2.
57. H. V. Firth *et al.*, DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* **84**, 524-533 (2009).
58. M. J. Landrum *et al.*, ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067 (2018).
59. A. D. Yates *et al.*, Ensembl 2020. *Nucleic Acids Res* **48**, D682-D688 (2020).
60. N. Mashtalir *et al.*, A Structural Model of the Endogenous Human BAF Complex Informs Disease Mechanisms. *Cell* **183**, 802-817 e824 (2020).
61. C. F. Wright *et al.*, Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med* **20**, 1216-1223 (2018).
62. J. Christensen *et al.*, RBP2 belongs to a family of demethylases, specific for tri-and dimethylated lysine 4 on histone 3. *Cell* **128**, 1063-1076 (2007).
63. J. E. Longbotham *et al.*, Histone H3 binding to the PHD1 domain of histone demethylase KDM5A enables active site remodeling. *Nat Commun* **10**, 94 (2019).
64. I. O. Torres *et al.*, Histone demethylase KDM5A is regulated by its reader domain through a positive-feedback mechanism. *Nat Commun* **6**, 6204 (2015).
65. Y. Zhang *et al.*, The PHD1 finger of KDM5B recognizes unmodified H3K4 during the demethylation of histone H3K4me2/3 by KDM5B. *Protein Cell* **5**, 837-850 (2014).
66. G. G. Wang *et al.*, Haematopoietic malignancies caused by dysregulation of a chromatin-binding PHD finger. *Nature* **459**, 847-851 (2009).

67. B. J. Klein *et al.*, The histone-H3K4-specific demethylase KDM5B binds to its substrate and product through distinct PHD fingers. *Cell Rep* **6**, 325-335 (2014).
68. J. E. Longbotham, M. J. S. Kelly, D. G. Fujimori, Recognition of Histone H3 Methylation States by the PHD1 Domain of Histone Demethylase KDM5A. *ACS Chem Biol* 10.1021/acschembio.0c00976 (2021).
69. R. C. Wu, M. Jiang, A. L. Beaudet, M. Y. Wu, ARID4A and ARID4B regulate male fertility, a functional link to the AR and RB pathways. *Proc Natl Acad Sci U S A* **110**, 4616-4621 (2013).
70. M. Lei *et al.*, Crystal structure of chromo barrel domain of RBBP1. *Biochem Biophys Res Commun* **496**, 1344-1348 (2018).
71. T. C. Fleischer, U. J. Yun, D. E. Ayer, Identification and characterization of three new components of the mSin3A corepressor complex. *Mol Cell Biol* **23**, 3456-3467 (2003).
72. M. Y. Wu, T. F. Tsai, A. L. Beaudet, Deficiency of Rbbp1/Arid4a and Rbbp1l1/Arid4b alters epigenetic modifications and suppresses an imprinting defect in the PWS/AS domain. *Genes Dev* **20**, 2859-2870 (2006).
73. R. C. Wu, Y. Zeng, I. W. Pan, M. Y. Wu, Androgen Receptor Coactivator ARID4B Is Required for the Function of Sertoli Cells in Spermatogenesis. *Mol Endocrinol* **29**, 1334-1346 (2015).
74. W. Gong, J. Wang, S. Perrett, Y. Feng, Retinoblastoma-binding protein 1 has an interdigitated double Tudor domain with DNA binding activity. *J Biol Chem* **289**, 4882-4895 (2014).
75. J. Ren *et al.*, Structural basis for the DNA-binding activity of human ARID4B Tudor domain. *J Biol Chem* 10.1016/j.jbc.2021.100506, 100506 (2021).
76. A. Patsialou, D. Wilsker, E. Moran, DNA-binding properties of ARID family proteins. *Nucleic Acids Res* **33**, 66-80 (2005).
77. C. F. Webb *et al.*, The ARID family transcription factor bright is required for both hematopoietic stem cell and B lineage development. *Mol Cell Biol* **31**, 1041-1053 (2011).
78. J. Iwahara, M. Iwahara, G. W. Daughdrill, J. Ford, R. T. Clubb, The structure of the Dead ringer-DNA complex reveals how AT-rich interaction domains (ARIDs) recognize DNA. *EMBO J* **21**, 1197-1209 (2002).
79. S. Kim, Z. Zhang, S. Upchurch, N. Isern, Y. Chen, Structure and DNA-binding sites of the SWI1 AT-rich interaction domain (ARID) suggest determinants for sequence-specific DNA recognition. *J Biol Chem* **279**, 16670-16676 (2004).
80. S. Cai, L. Zhu, Z. Zhang, Y. Chen, Determination of the three-dimensional structure of the Mrf2-DNA complex using paramagnetic spin labeling. *Biochemistry* **46**, 4943-4950 (2007).
81. S. Tu *et al.*, The ARID domain of the H3K4 demethylase RBP2 binds to a DNA CCGCCC motif. *Nat Struct Mol Biol* **15**, 419-421 (2008).
82. K. Hata *et al.*, Arid5b facilitates chondrogenesis by recruiting the histone demethylase Phf2 to Sox9-regulated genes. *Nat Commun* **4**, 2850 (2013).
83. T. Yamakawa, R. H. Whitson, S. L. Li, K. Itakura, Modulator recognition factor-2 is required for adipogenesis in mouse embryo fibroblasts and 3T3-L1 cells. *Mol Endocrinol* **22**, 441-453 (2008).
84. R. Le Bouffant *et al.*, Differential expression of arid5b isoforms in Xenopus laevis pronephros. *Int J Dev Biol* **58**, 363-368 (2014).
85. H. Fan *et al.*, BAHCC1 binds H3K27me3 via a conserved BAH module to mediate gene silencing and oncogenesis. *Nat Genet* **52**, 1384-1396 (2020).
86. A. J. Kuo *et al.*, The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier-Gorlin syndrome. *Nature* **484**, 115-119 (2012).

87. W. Ren *et al.*, DNMT1 reads heterochromatic H4K20me3 to reinforce LINE-1 DNA methylation. *Nat Commun* **12**, 2490 (2021).
88. D. Duan, Systemic AAV Micro-dystrophin Gene Therapy for Duchenne Muscular Dystrophy. *Mol Ther* **26**, 2337-2356 (2018).
89. R. Tillotson *et al.*, Radically truncated MeCP2 rescues Rett syndrome-like neurological defects. *Nature* **550**, 398-401 (2017).
90. W. Zhang, L. Li, Q. Su, G. Gao, H. Khanna, Gene Therapy Using a miniCEP290 Fragment Delays Photoreceptor Degeneration in a Mouse Model of Leber Congenital Amaurosis. *Hum Gene Ther* **29**, 42-50 (2018).
91. C. F. Mugal, H. Ellegren, Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. *Genome Biol* **12**, R58 (2011).

Supplementary figures for:

Missense variants reveal functional insights into the human ARID family of gene regulators

Gauri Deák and Atlanta G. Cook*
Wellcome Centre for Cell Biology, University of Edinburgh, Michael Swann Building, Max Born Crescent, Edinburgh EH9 3BF, United Kingdom

Figure S1: Modelled structure of the ARID2 equivalent of the ARID1A/1B CBRB (annotated with missense variants **(A)** and sequence conservation in metazoa **(B)**

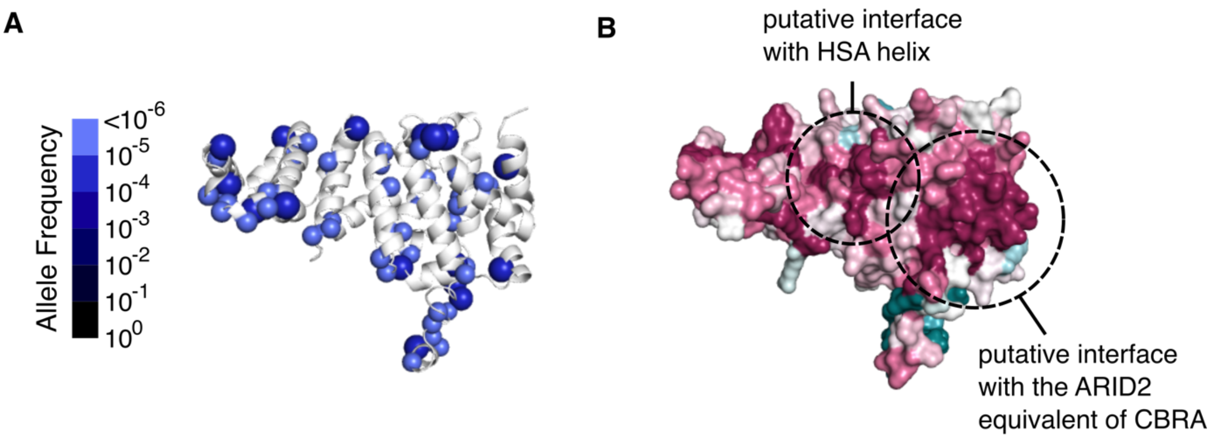

A
Allele Frequency
<10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
B
putative interface with HSA helix
putative interface with the ARID2 equivalent of CBRA

Figure S2: Solved structure of JARID1B (5FUP) annotated with missense variants **(A)** and sequence conservation in metazoa **(B)**

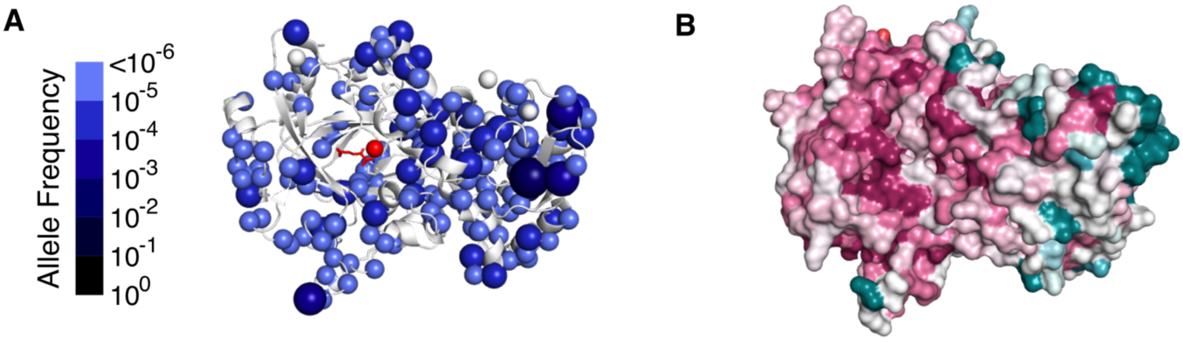


Figure S3: 1D plots of missense variants in ARID4A **(A)** and ARID4B **(B)** and the V_d/V_p ratios calculated for their functional domains

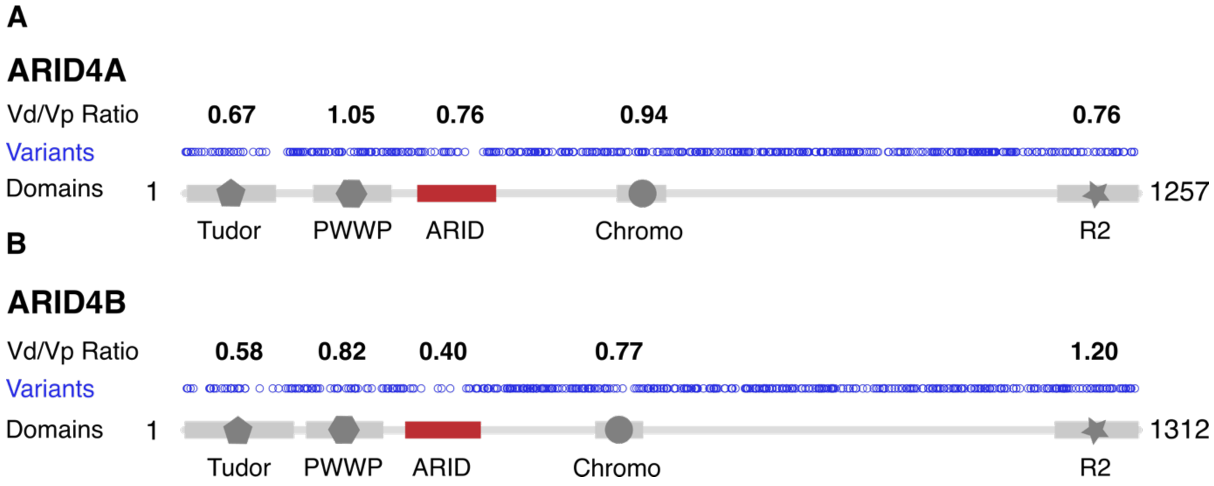


Figure S4: 1D plots of missense variants in ARID3A **(A)**, ARID3B **(B)**, and ARID3C **(C)** and the V_d/V_p ratios calculated for their functional domains

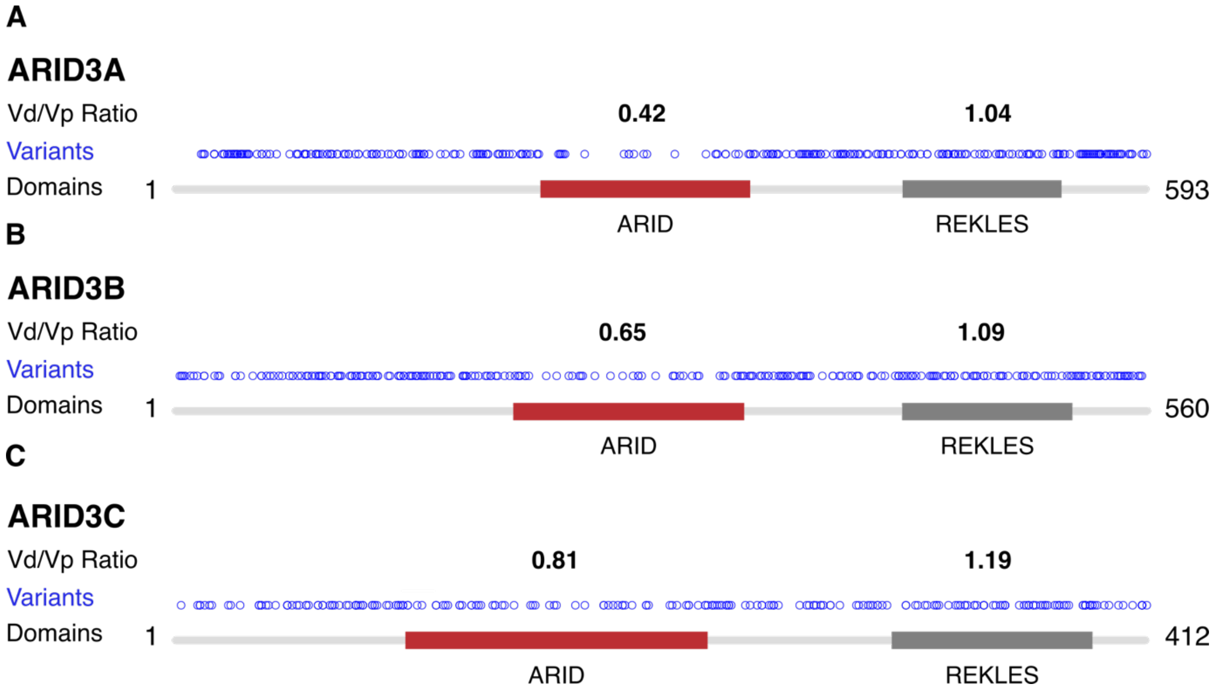


Figure S5: A vertebrate multiple sequence alignment of the N-terminal region of ARID5B isoform 1; The end of the BAH domain and the start of the C-terminal extension is marked by a vertical line at human aa 107. Darker blue represents higher percentage identity.

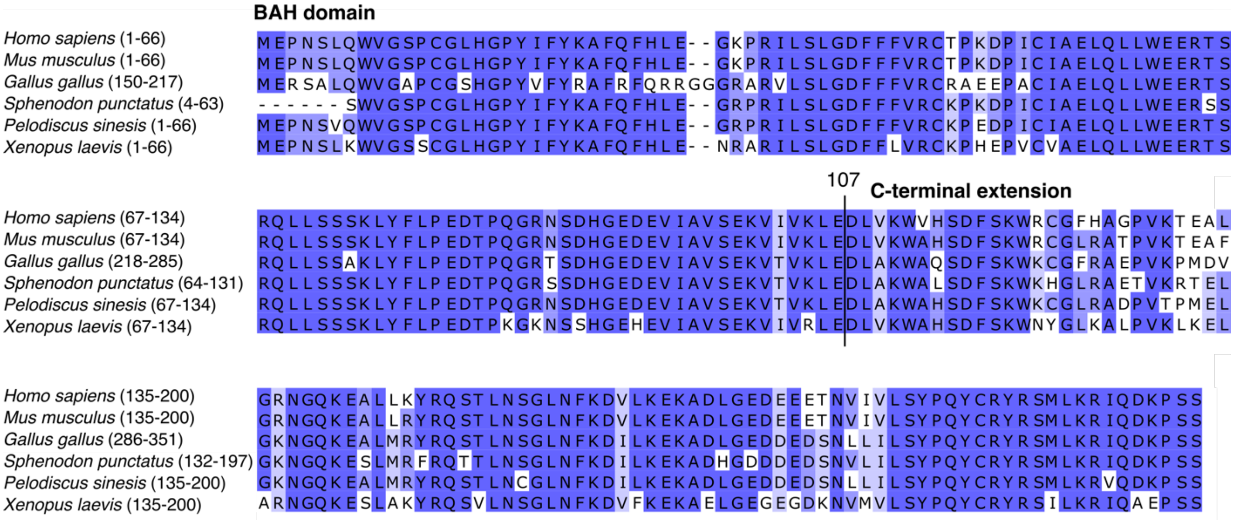
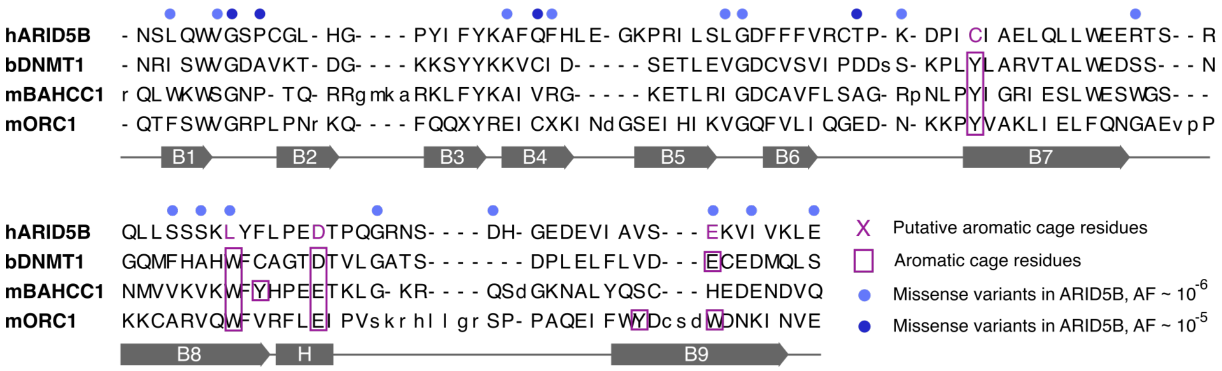

BAH domain
Homo sapiens (1-66) MEPNSLQWVGSPCGLHGPYIFYKAFQFHLE--GKPRILSLGDFFFVRCTPKDPICIAELQLLWEERTS
Mus musculus (1-66) MEPNSLQWVGSPCGLHGPYIFYKAFQFHLE--GKPRILSLGDFFFVRCTPKDPICIAELQLLWEERTS
Gallus gallus (150-217) MERSALQWVGAPCGSHGPYVFYRAFRFQRRGGGRARVLSLGDFFFVRCRAEEPACIAELQLLWEERTS
Sphenodon punctatus (4-63) ------SWVGSPCGLHGPYIFYKAFQFHLE--GRPRILSLGDFFFVRCKPKDPICIAELQLLWEERSS
Pelodiscus sinesis (1-66) MEPNSVQWVGSPCGLHGPYIFYKAFQFHLE--GRPRILSLGDFFFVRCKPEDPICIAELQLLWEERTS
Xenopus laevis (1-66) MEPNSLKWVGSSCGLHGPYIFYKAFQFHLE--NRARILSLGDFFLVRCKPHEPVCVAELQLLWEERTS
107
C-terminal extension
Homo sapiens (67-134) RQLLSSSKLYFLPEDTPQGRNSDHGEDEVIAVSEKVIVKLEDLVKWVHSDFSKWRCGFHAGPVKTEAL
Mus musculus (67-134) RQLLSSSKLYFLPEDTPQGRNSDHGEDEVIAVSEKVIVKLEDLVKWAHSDFSKWRCGLRATPVKTEAF
Gallus gallus (218-285) RQLLSSAKLYFLPEDTPQGRTSDHGEDEVIAVSEKVTVKLEDLAKWAQSDFSKWKCGFRAEPVKPMDV
Sphenodon punctatus (64-131) RQLLSSSKLYFLPEDTPQGRSSDHGEDEVIAVSEKVTVKLEDLAKWALSDFSKWKHGLRAETVKRTEL
Pelodiscus sinesis (67-134) RQLLSSSKLYFLPEDTPQGRNSDHGEDEVIAVSEKVTVKLEDLAKWAHSDFSKWKCGLRADPVTPMEL
Xenopus laevis (67-134) RQLLSSSKLYFLPEDTPKGKNSSHGEHEVIAVSEKVIVRLEDLVKWAHSDFSKWNYGLKALPVKLKEL
Homo sapiens (135-200) GRNGQKEALLKYRQSTLNSGLNFKDVLKEKADLGEDEEETNVIVLSYPQYCRYRSMLKRIQDKPSS
Mus musculus (135-200) GRNGQKEALLRYRQSTLNSGLNFKDVLKEKADLGEDEEETNVIVLSYPQYCRYRSMLKRIQDKPSS
Gallus gallus (286-351) GKNGQKEALMRYRQSTLNSGLNFKDILKEKADLGEDDEDSNLLILSYPQYCRYRSMLKRIQDKPSS
Sphenodon punctatus (132-197) GKNGQKESLMRFRQTTLNSGLNFKDILKEKADHGDDDEDSNVLILSYPQYCRYRSMLKRIQDKPSS
Pelodiscus sinesis (135-200) GKNGQKEALMRYRQSTLNCGLNFKDILKEKADLGEDDEDSNLLILSYPQYCRYRSMLKRVQDKPSS
Xenopus laevis (135-200) ARNGQKESLAKYRQSVLNSGLNFKDVFKEKAELGEGEGDKNVMVLSYPQYCRYRSILKRIQAEPSS

Figure S6: Structural alignment of the ARID5B BAH domain**;** h = human, b = bovine, m = murine


hARID5B -NSLQWWGSPCGL-HG---PYIFYKAFQFHLE-GKPRILSLGDFFFVRCTP-K-DPICIAELQLLWEERTS--R
bDNMT1 -NRISWWGDAVKT-DG---KKSYYKKVCID-----SETLEVGDCVSVIPDDsS-KPLYLARVTALWEDSS---N
mBAHCC1 rQLWKWSGNP-TQ-RRgmkaRKLFYKAIVRG-----KETLRIGDCAVFLSAG-RpNLPYIGRIESLWESWGS---
mORC1 -QTFSWWGRPLPNrKQ---FQQXYREICXKINdGSEIHIKVGQFVLIQGED-N-KKPYVAKLIELFQNGAEvpP
B1
B2
B3
B4
B5
B6
B7
hARID5B QLLSSSKLYFLPEDTPQGRNS----DH-GEDEVIAVS---EKVIVKLE
bDNMT1 GQMFHAHWFCAGTDTVLGATS-------DPLELFLVD---ECEDMQLS
mBAHCC1 NMVVKVKWFYHPEETKLG-KR----QSdGKNALYQSC---HEDENDVQ
mORC1 KKCARVQWFVRFLEIPVskrhIlgrSP-PAQEIFWYDcsdWDNKINVE
B8
H
B9
X Putative aromatic cage residues
□ Aromatic cage residues
● Missense variants in ARID5B, AF ~ 10^-6
● Missense variants in ARID5B, AF ~ 10^-5