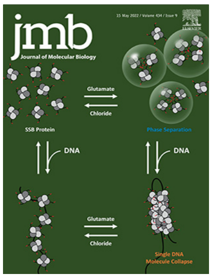
Research Article



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 to Atlanta G. Cook:** atlanta.cook@ed.ac.uk (A.G. Cook), @GauriDeak (G. Deák), @WTcell (A.G. Cook)*

https://doi.org/10.1016/j.jmb.2022.167529
Edited by Anna Panchenko

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.

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 and are likely associated with disease.[1,2] This has aided large-scale assessments, for example, of genetic causality in autism spectrum disorder[3] and inherited cardiomyopathies.[4] The increase in statistical power afforded by the size of these datasets has allowed genes to be ranked for their importance to human health based on their intolerance to variation.

Beyond the effects of loss-of-function variants on a gene, constraint analyses can be further narrowed to focus on 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.[5] Rare mutations occurring in such regions are likely to be pathogenic.[6,7] 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.[8,9] 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.[10]

While calculating variant distributions along protein sequences can help to identify essential domains, it fails to consider the arrangement of variants in 3D space.[6] This has been addressed by manual mapping of variants[11] 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.[12] In an alternative approach, Tang et al. introduced PSCAN, which scores the spatial dispersion of missense variants onto structures,[13] based on a previous finding that neutral variants tend to be dispersed, while pathogenic variants cluster.[14] In a further approach, the MISCAST suite provides an approach to connect probabilities of loss of function with primary sequence level features.[15]

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 (Figure 1(A)).

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, 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 (Figure 1(A)). 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 (Figure 1(B)). This family was selected

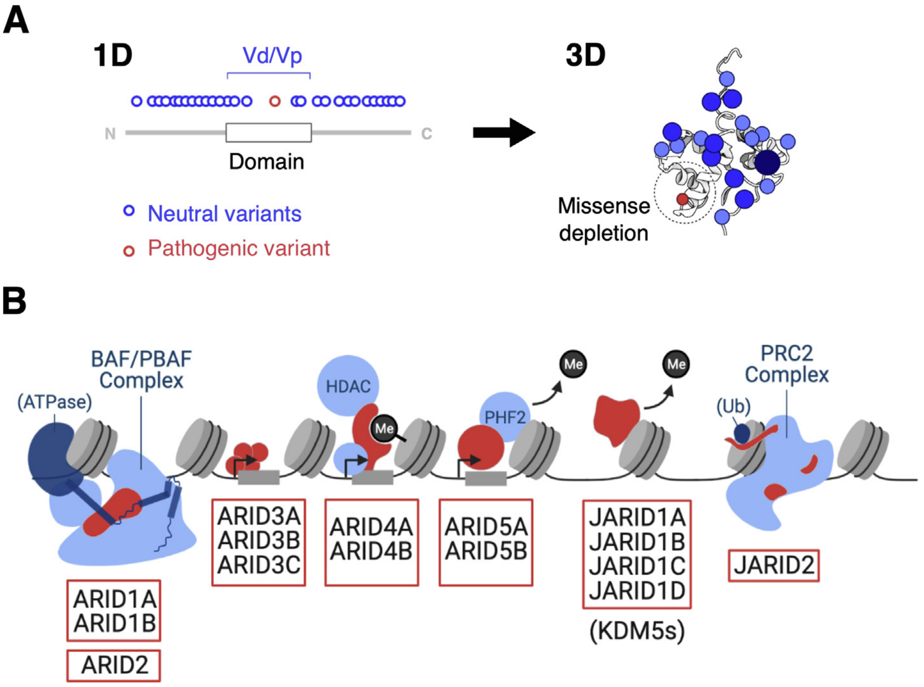


Figure 1. (A) Schematic of the 1D-to-3D approach of mapping missense variants onto protein structures. **(B)** Schematic illustration of the interactions of ARID family members (red) with nucleosomes (grey) and other chromatin-binding proteins (light blue); (P)BAF = (Polybromo-associated) BRG1/BRM-associated factor, HDAC = histone deacetylase, PHF2 = PHD finger 2 (a lysine-specific demethylase), PRC2 = Polycomb Repressive Complex 2.

due to the clinical significance of its members in multiple cancer types[16–18] and rare developmental disorders such as Coffin-Siris syndrome.[19] 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.[20] Despite their common domain, the family members regulate distinct sets of genes via diverse molecular mechanisms (Figure 1(B)). Members of the ARID1 sub-family and ARID2 are core subunits of large nucleosome remodeling complexes while JARID2 is an accessory subunit of a transcription repressor complex.[21] In contrast, the ARID3 proteins are transcription factors,[22] while ARID4s and ARID5s are adapter proteins that recruit other transcriptional regulators.[23–25] ARID5A is thought to be an RNA-binding protein.[26] Finally, the four JARID1 proteins are enzymes that mediate transcriptional changes by removal of histone H3K4 di-/tri-methylation marks.[17]

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).[27] Multiple sequence alignments were generated using MAFFT (RRID: SCR_011811)[28] and pairwise alignments were generated with EMBOSS Needle.[29] Alignments were then visualized in JalView 2.11.1.4 (RRID:SCR_006459).[30] Evolutionary relationships between paralogs in ARID subfamilies were verified using TreeFam (RRID:SCR_013401).[31] A structural sequence alignment for the ARID5B BAH domain was created using the DALI server (RRID:SCR_013433).[32] Functional domains of each family member were annotated using InterPro[33] or based on experimental data (the ARID1A/1B core binding regions,[34] ARID4A/B R2 region,[24] and JARID1A[35] and JARID1B[36] 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.[37] 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”.[37] All of the structures were visualized in PyMOL (Schrödinger Inc.). Surface electrostatics scores for the ARID4A and ARID4B hybrid Tudor domains were calculated using the Adaptive Poisson-Boltzmann Solver (RRID: SCR_008387)[38] 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).[39] 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.[40] All sequence alignments can be accessed at: https://doi.org/10.7488/ds/3190.

Isoform expression analysis

Isoform-specific expression data for ARID5B was obtained from ISOexpresso.[41] 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 individuals.

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.[1,42] 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.

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.[43,44] 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.[43] All variant accession codes are available in Supplementary Table 1.

1D plots and the Vd/Vp ratio

Filtered missense variants were mapped onto protein sequences of ARID family members using Plot Protein.[45] The Plot Protein R script was modified to allow for color manipulation and domain diagram alterations of the output graphs in Inkscape. 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 allele frequency. 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 allele frequencies are summed. The allele frequency values for each position are compressed using a base 10 log scale and the positions are sorted into 6 bins (allele frequency <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/3190.

Results

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 scores (Figure 2(A)) as well as pLi and RVIS scores (Supplementary Table 1).[1,42] The lower the LOEUF, the fewer the variants observed than expected, indicating negative selection against loss-of-function variation. All ARID family genes except ARID3B/3C and JARID1B/1D can be classified as loss-of-function-intolerant (Figure 2(A)). 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).

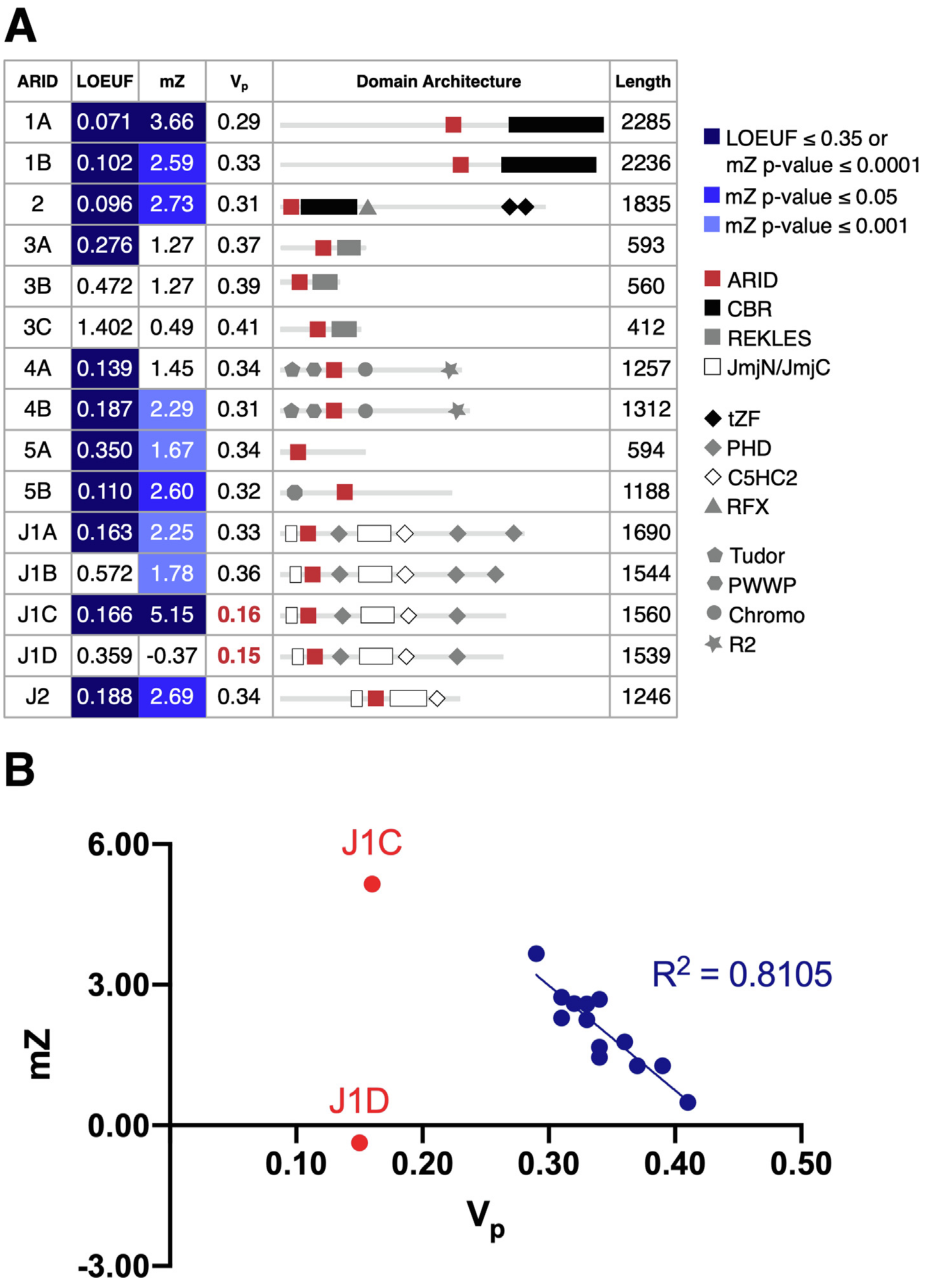
The missense Z score represents the deviation of the observed from the expected number of missense variants (single amino acid substitutions), for a given gene, where positive scores indicate missense depletion and negative scores indicate missense enrichment.[2] ARID1A and JARID1C have missense Z P-values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

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 missense Z scores (Figure 2(B)). However, JARID1C and JARID1D do not fit the observed correlation between missense Z and V_p values (Figure 2(B)). 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.[1,46] This analysis suggests that Vp is a good proxy for missense Z 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 (Figure 2 (A)). To bridge this gap, we developed the '1D-to-3D' approach (Figure 1(A)). 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 allele frequencies of variants at each residue position, with visual discrimination of allele frequencies using increasing sphere size and shade of blue (Figure 1(A)). In line with larger-scale analyses,[5,12] 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 (Figure 2(A)) and 57 % sequence identity in humans. ARID1A and ARID1B are mutually-exclusive core subunits of the BRG1/BRM-associated factors (BAF)



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 2. Genetic constraint in the ARID family (A) An overview of constraint metrics, Vp values, 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 variants in a given gene. The recommended threshold to segregate loss-of-function-intolerant and loss-of-function-tolerant genes is 0.35. The missense Z scores outside of ±3.09 correspond to a recommended P-value threshold of 0.001. **(B)** Correlation between the missense Z scores and Vp values for this dataset (blue). Values for JARID1C (J1C) and JARID1D (J1D) are red.

chromatin remodeling complex.[47] BAF complexes modulate transcription through ATPase-dependent nucleosome sliding/ejection or the recruitment of other regulators.[34] They comprise three modules: a catalytic ATPase (BRG1/BRM), Actin-Related Protein module, and a base module that scaffolds the ATPase, nucleosome, and other regulators (Figure 3(A)).[34] ARID1A or ARID1B are incorporated into the base module through a C-terminal Core Binding Region made up of conserved core binding region A and core binding region B segments connected by intrinsically disordered loops.[48]

Mutations in the BAF complex promote tumorigenesis in multiple cancer types reviewed in.[16] A recent analysis of missense cancer mutations in the BAF complex revealed that several

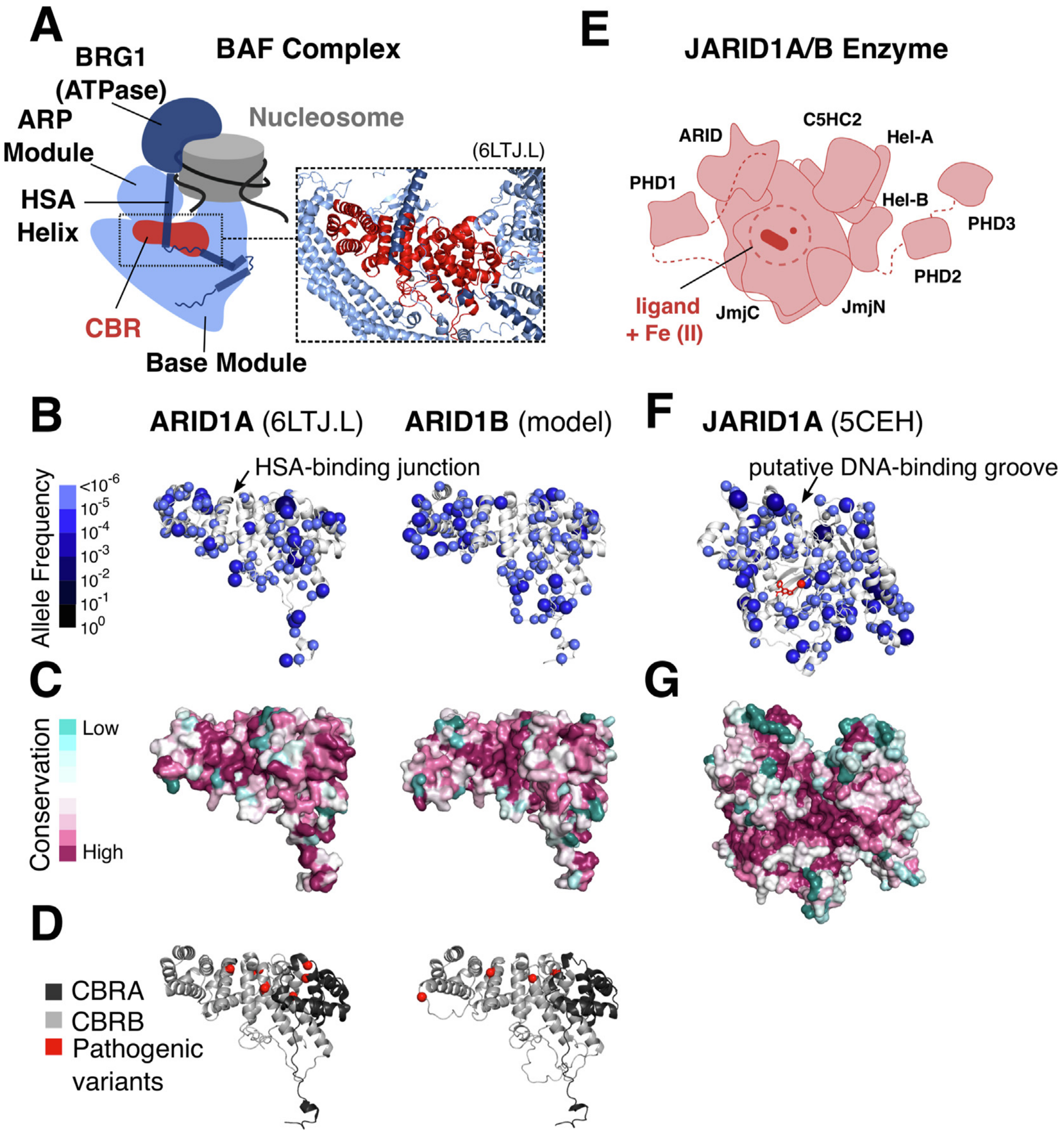


Figure 3. Validation of the 1D-to-3D approach (A) Position of the ARID1A/1B core binding region in the BAF complex**.** Solved structure of the ARID1A core binding region (PDBid 6LTJ, chain L) and modelled structure of the ARID1B core binding region annotated with missense variants **(B),** sequence 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).**

mutations cluster at a junction between the ARID1A/B core binding region and the helicase-SANT-associated helix of the ATPase (Figure 3 (A)).[49] We find that this junction is depleted of neutral missense variants (Figure 3(B)), and this correlates well with evolutionary surface conservation of the core binding region in vertebrates (Figure 3(C)). We also found that pathogenic variants associated with Coffin-Siris Syndrome and non-syndromic intellectual disability variants map in proximity to the junction (Figure 3(D)). ARID2 plays a functionally analogous role to the ARID1 family in the related Polybromo-associated BAF (PBAF) nucleosome remodeling complex.[48] Even though ARID2 is distantly related to the ARID1 family, we found a similar depletion of missense mutations around the putative helix binding site in a model of the ARID2 core binding region (Figure S1). 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 (Figure 3(E)).

We report that the catalytic site in the JmjC domain is visibly missense depleted (Figure 3(F)), correlating with evolutionary conservation (Figure 3(G)). 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). A similar depletion around the active site is observed for JARID1B (Figure S2), but is less pronounced, consistent with the higher LOEUF and lower missense Z metrics reported for JARID1B (Figure 2(A)).

Apart from the catalytic site, depletion of missense variants in JARID1A is also observed in a groove between the ARID and C5HC2 domains, which was hypothesized to accommodate double-stranded DNA (Figure 3(F); Movie S1).[35] However, it should be noted that there has been no experimental evidence to confirm this hypothesis. The second site of depletion could represent a different kind of functional surface, such as a protein-protein interaction or other ligand interaction site. Overall, our findings validate the 1D-to-3D 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 (Figure 4(A)–(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 (Figure 4(B)). 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 (Figure 4(B)).

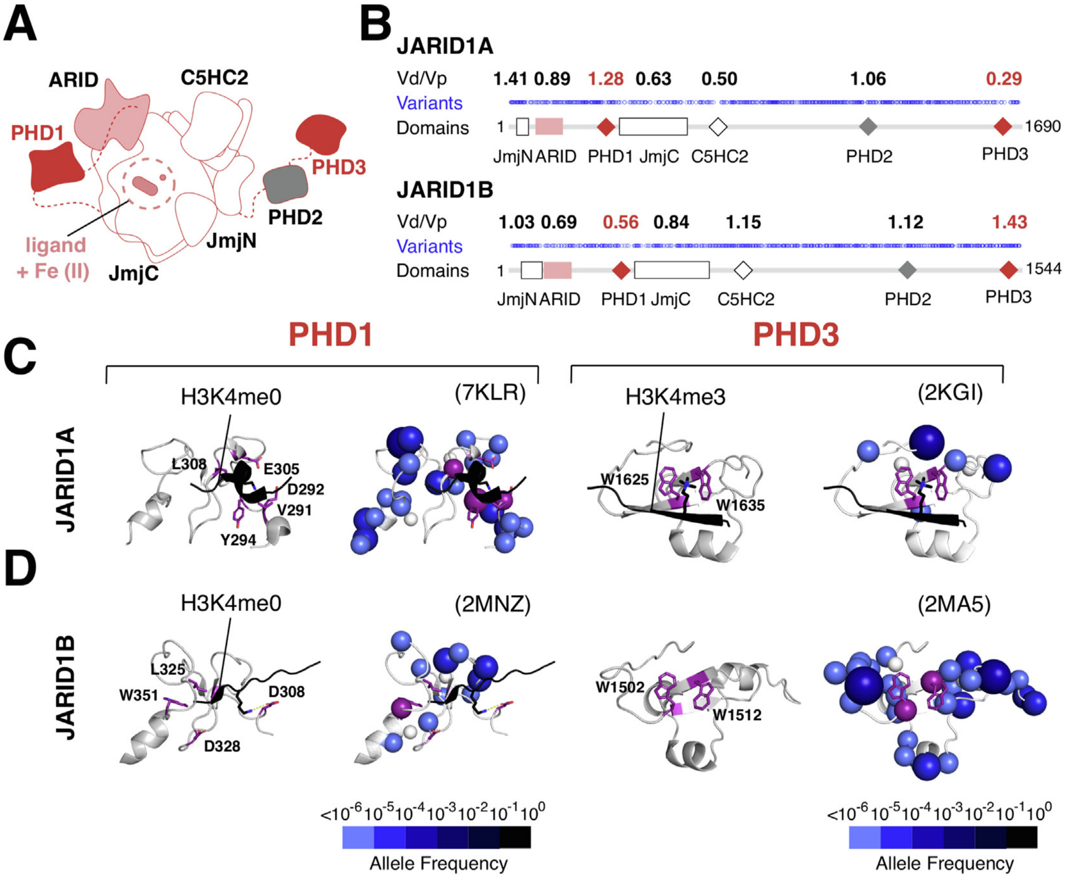


Figure 4. Comparative Analysis of Domains Using the Vd/Vp Ratio (A) Schematic diagram of the JARID1A and JARID1B enzymes (based on PDB structures 5CEH and 5FUP). **(B)** 1D plots of missense variants in JARID1A/1B and with Vd/Vp ratios calculated for their functional domains, shown above the plot. **(C)** 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.

In JARID1A and JARID1B, PHD1 preferentially binds to unmethylated histone H3K4, leading to increased affinity of the JmjC domain for its methylated H3K4 substrates.[50–52] In contrast, PHD3 is thought to be a reader of tri-methylated H3K4 marks.[53,54] 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,[54,55] are affected by neutral missense variants (Figure 4(C)–(D)). Conversely, residues responsible for histone peptide binding in PHD3 of JARID1A (Figure 4(C))[53,55] and PHD1 of JARID1B (Figure 4(D))[52,56] 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.[53] Furthermore, mutations in PHD1, but not PHD3 of JARID1B decreased the regulatory effects on cell migration in a model of triple negative breast cancer.[54] In summary, functional differences in the PHD fingers correlate with different Vd/Vp ratio patterns in JARID1A and JARID1B, suggesting that the differences in 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 underestimated. 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 (Figure 5).

The ARID4 subfamily comprises ARID4A and ARID4B (Figure 2(A)), two paralogous, multidomain adapter proteins that recruit transcriptional regulators such as the retinoblastoma protein, androgen receptor, and the mSin3A histone deacetylase complex to gene promoters.[25,57] Given the presence of an aromatic cage, the chromobarrel domain was hypothesized to bind histone methyl marks.[24] However, independent nuclear magnetic resonance and isothermal titration calorimetry experiments with the ARID4A chromobarrel domain and methylated histone peptides produced conflicting results.[24,58]

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 (Figure 5(A)) and a model of the ARID4B chromobarrel domain (Figure 5(B)). 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 subfunctionalization

ARID4A and ARID4B have diverged only in vertebrate lineages[31] and share identical domain architecture (Figure 2(A)). They participate in the same molecular pathways, including the recruitment of the mSin3A repressive complex to gene promoters[59] and co-activation of the androgen receptor in regulation of male fertility.[57] However, ARID4A knockout mice are viable whereas ARID4B knockouts show early embryonic lethality.[60] Moreover, ARID4B is necessary for spermatogenesis

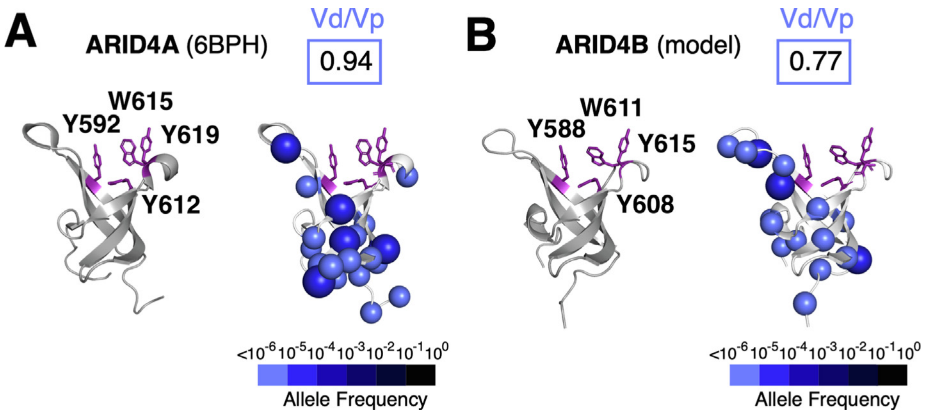


Figure 5. Limitations of the Vd/Vp Ratio. The ARID4A **(A)** and ARID4B **(B)** chromobarrel domains shown without (left) and with (right) missense variants. Putative histone methyl mark-binding residues are indicated in purple.

while ARID4A is not.[57] 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 (Figure 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 (Figure 6(A)). Unlike structurally analogous Tudor domains, ARID4 HTDs lack an aromatic cage associated with histone binding in HTD-2 and have instead been shown, experimentally, to bind DNA through HTD-1[61,62] (Figure 6(A)). A conserved, structurally important glycine in the ARID4B Tudor domain is associated with a developmental disorder variant (Figure 6(A)).

We find that HTD-1 of ARID4B is missense depleted, while HTD-1 of ARID4A is not (Figure 6 (B)). The depleted site corresponds to previously identified DNA-binding residues (Figure 6(B)) and a positively-charged DNA-binding surface of the domain (Figure 6(C)).[62] We also note that an RGR motif, recently found to enhance the DNA-binding affinity of ARID4B,[62] tolerates missense variation (Figure 6(A)). Overall, our findings indicate that the ARID4B Tudor domain likely contributes to the functional differences in DNA binding 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 (Figure 2(A)). We note that the Vd/Vp ratios of all three REKLES domains are >1.00 (Figure 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.

The ARID domain binds to AT-rich promoter sequences and is essential for ARID3 protein function.[22] 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) (Figure 7 (A)). A solution structure of the *Drosophila* Dead ringer in complex with DNA[63] and nuclear magnetic resonance titration experiments of the ARID1A,[64] ARID5B,[65] and JARID1A[66] ARID domains suggest a common mode of DNA 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[63–66](Figure 7 (B)).

The Vd/Vp ratios of ARID domains of the three human ARID3 paralogs are lowest in ARID3A and highest in ARID3C (Figure 7(C)–(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.[67] The domain also harbors a developmental disorder variant,[43] supporting its functional importance (Figure 7

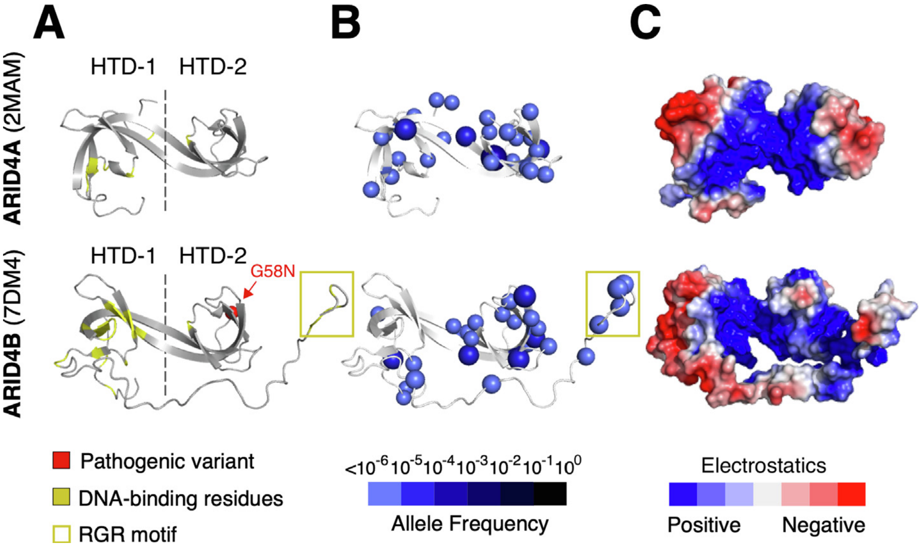


Figure 6. The ARID4 Subfamily. The ARID4A and ARID4B hybrid Tudor domain (HTD) shown with DNA binding residues **(A)**, missense variants **(B)**, and surface electrostatics **(C)**.

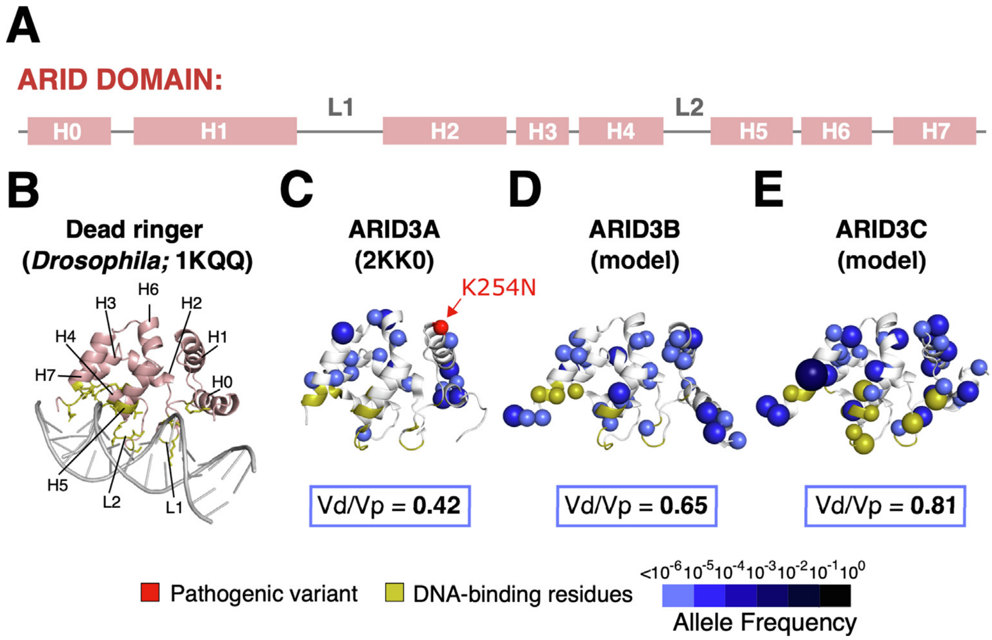


Figure 7. The ARID3 Subfamily (A) Secondary structure of the ARID domain where helices are denoted with H and loops are denoted with L. The ARID3 subfamily has two additional flanking helices H0 and H7. **(B)** Solution structure of the Dead ringer ARID domain in complex with DNA. Human ARID3A **(C)**,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

(C)). 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 (Figure 7 (D)). Similarly, in ARID3C, several residues typically involved in DNA-binding are found to have reported variants (Figure 7(E)). 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 (Figure 2(A)) and a key regulator of liver metabolism, chondrogenesis, and adipogenesis.[23,68,69] ARID5B is known to target the H3K9me2 demethylase PHF2 to gene promoters via its ARID domain.[23,70]

ARID5B has two isoforms, 1 and 2 (Figure 8(A)). In *Xenopus,* isoform expression is spatially segregated during embryonic development, where isoform 1 shows higher abundance than isoform 2.[71] Isoform 1 also shows higher expression in healthy adult human tissues.[41] Isoform1 has an additional N-terminal region that is highly conserved in a subset of vertebrate species (Figure S5) and predicted to form a bromo-adjacent homology (BAH) domain[72]; 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 (Figure S6). All three homologs read histone methyl marks (Figure 8(B)).[73–75] 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 (Figure 8(C)). 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 (Figure 8(C)). 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 (Figure 8(D)). When mapped to this extended AlphaFold model, we note that the missense variants are depleted on the highly conserved, positively-charged C-

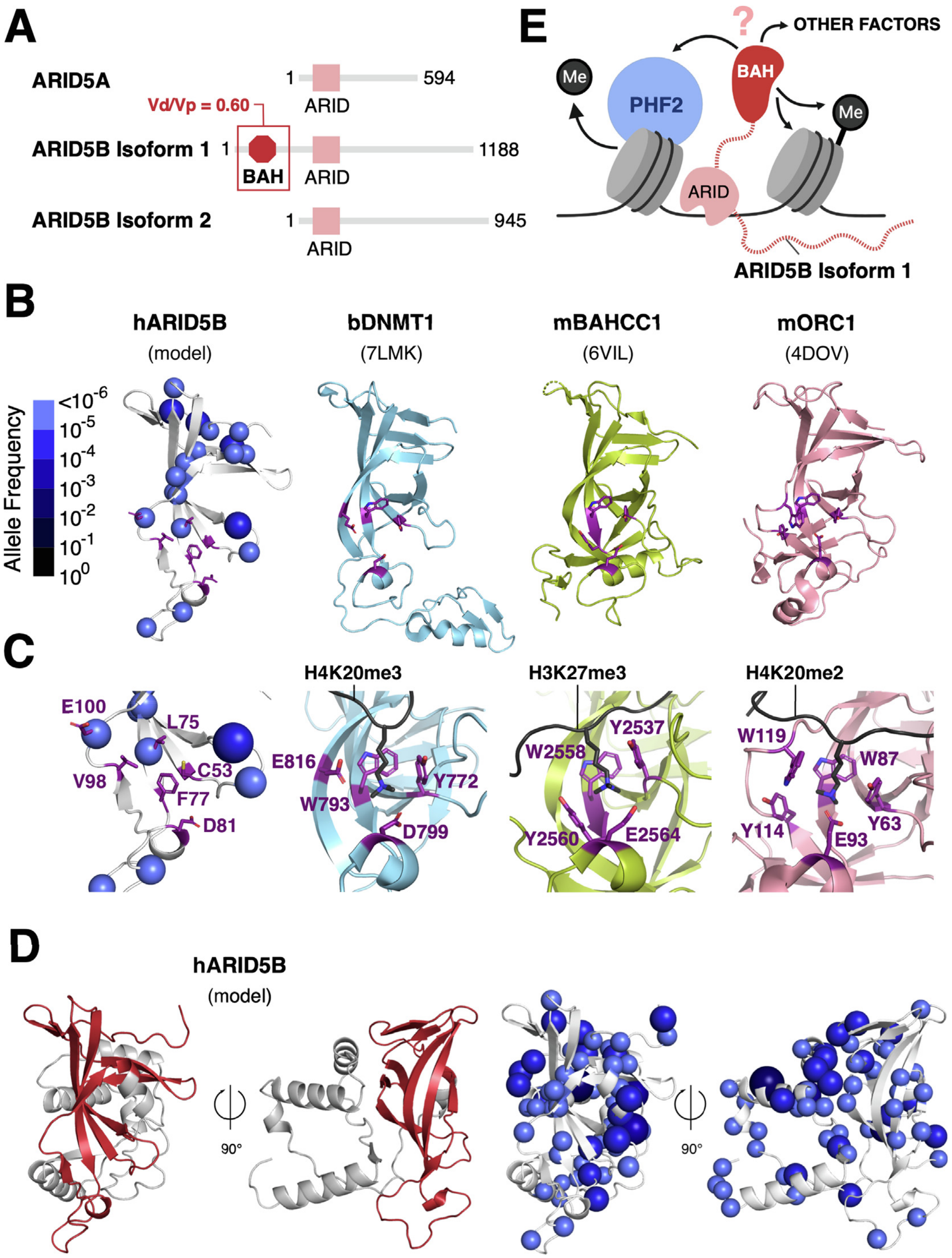


Figure 8. Loss and gain of function in ARID5B (A) Domain architecture of the human ARID5 subfamily. **(B)** Variant-annotated model of the human ARID5B BAH domain compared to solved structures of the bovine DNMT1, mouse BAHCC1, and mouse ORC1 BAH domains (histone methyl mark-binding residues shown in 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.

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 (Figure 8(E)). These predictions call for experimental evidence. However, our analysis demonstrates the power of missense variation in screening for functional features together with structural data.

Discussion

We provide a convenient set of tools for mapping missense variants onto primary and tertiary structures of proteins. Our initial analysis of variant depletion in proteins showed that Vp is a good proxy for other scores such as missense Z scores, which reveal resistance to variation (Figure 2). This also allowed us to identify paralogs that were not suitable for further analysis based on limited available information, such as for JARID1C and JARID1D, which are encoded on sex chromosomes. Vp, and the related Vd/Vp ratio, have the advantage that they are easy to understand in relation to the available data.

Our approach further 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. Our method focused on essential human proteins because our data comprised human variants. However, we showed that it is complementary to phylogenetic conservation analysis (Figure 3) and it is likely that the conserved functional surfaces we characterized are relevant to homologues in other organisms as well. In cases where recent mammalian paralogs have a single homologue in distantly related organisms, our approach also revealed insights into paralog sub-functionalization of protein domains (Figures 6 and 7).

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.[76,77] However, we also note that Vd/Vp ratios do not always provide sufficient information to determine whether small domains are under selective pressure. In the example of the ARID4B chromobarrel domain (Figure 5), mapping variants onto the 3D structure revealed a likely functional site that could not have been supported by the Vd/Vp ratio alone. Therefore, mapping variants onto structural models is a useful complementary application of the tool.

Some additional 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.[78] 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.

Conclusions

Our findings build on previous studies showing that depletion of missense variants can serve to identify functionally important protein sites.[5,12] 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 protein functions.

CRediT authorship contribution statement

Gauri Deák: Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Atlanta G. Cook:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Acknowledgements

The authors would like to thank Sander Granneman and Philipp Voigt for critical reading of the manuscript. This work was supported by 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/3190.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2022.167529.

Received 9 December 2021;
Accepted 1 March 2022;
Available online 4 March 2022

***Keywords*:**
exome;
genetic variation;
transcription factors;
humans;
nucleosomes

***Abbreviations*:**

ARID, AT-rich interactive domain; BAF, BRG1/BRM-associated factors; BAH, bromo-adjacent homology; BAHCC1, BAH domain and coiled-coil containing 1; DNMT1, DNA (cytosine-5)-methyltransferase 1; gnomAD, genome aggregation database; HDAC, histone deacetylase; HTD, hybrid tudor domain; Jmj, jumonji; LOEUF, loss-of-function observed/expected upper bound fraction; NUP98, nucleoporin 98; ORC1, origin recognition complex subunit 1; PBAF, polybromo-associated BRG1/BRM-associated factors; PHD, plant homeodomain; PHF2, PHD Finger 2; pLI, probability of loss-of-function intolerance; PRC2, Polycomb Repressive Complex 2; RVIS, residual variance intolerance score

References

1. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., et al., (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443.
2. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., et al., (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291.
3. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., et al., (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genet.* **46**, 944–950.
4. Walsh, R., Thomson, K.L., Ware, J.S., Funke, B.H., Woodley, J., McGuire, K.J., et al., (2017). Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med.* **19**, 192–203.
5. MacGowan, A., Madeira, F., Britto-Borges, T., Schmittner, M., Cole, C., Barton, G.J., (2017). Human Missense Variation is Constrained by Domain Structure and Highlights Functional and Pathogenic Residues. *bioRxiv*, 127050.
6. Havrilla, J.M., Pedersen, B.S., Layer, R.M., Quinlan, A.R., (2019). A map of constrained coding regions in the human genome. *Nature Genet.* **51**, 88–95.
7. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnel-Luria, A.H., Pierce-Hoffman, E., MacArthur, D. G., et al., (2017). Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*.
8. Brenan, L., Andreev, A., Cohen, O., Pantel, S., Kamburov, A., Cacchiarelli, D., et al., (2016). Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. *Cell Rep.* **17**, 1171–1183.
9. Majithia, A.R., Tsuda, B., Agostini, M., Gnanapradeepan, K., Rice, R., Peloso, G., et al., (2016). Prospective functional classification of all possible missense variants in PPARG. *Nature Genet.* **48**, 1570–1575.
10. Iqbal, S., Perez-Palma, E., Jespersen, J.B., May, P., Hoksza, D., Heyne, H.O., et al., (2020). Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc. Natl. Acad. Sci. USA* **117**, 28201–28211.
11. Bai, D., Wang, J., Li, T., Chan, R., Atalla, M., Chen, R.C., et al., (2021). Differential Domain Distribution of gnomAD- and Disease-Linked Connexin Missense Variants. *Int. J. Mol. Sci.* **22**, 7832.
12. Hicks, M., Bartha, I., di Iulio, J., Venter, J.C., Telenti, A., (2019). Functional characterization of 3D protein structures informed by human genetic diversity. *Proc. Natl. Acad. Sci. USA* **116**, 8960–8965.
13. Tang, Z.Z., Sliwoski, G.R., Chen, G., Jin, B., Bush, W.S., Li, B., et al., (2020). PSCAN: Spatial scan tests guided by protein structures improve complex disease gene discovery and signal variant detection. *Genome Biol.* **21**, 217.
14. Sivley, R.M., Dou, X., Meiler, J., Bush, W.S., Capra, J.A., (2018). Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures. *Am. J. Hum. Genet.* **102**, 415–426.
15. Iqbal, S., Hoksza, D., Perez-Palma, E., May, P., Jespersen, J.B., Ahmed, S.S., et al., (2020). MISCAST: MIssense variant to protein StruCture Analysis web SuiTe. *Nucleic Acids Res.* **48**, W132–W139.
16. Mittal, P., Roberts, C.W.M., (2020). The SWI/SNF complex in cancer - biology, biomarkers and therapy. *Nature Rev. Clin. Oncol.* **17**, 435–448.
17. Plch, J., Hrabeta, J., Eckschlager, T., (2019). KDM5 demethylases and their role in cancer cell chemoresistance. *Int. J. Cancer* **144**, 221–231.
18. Wu, R.C., Young, I.C., Chen, Y.F., Chuang, S.T., Toubaji, A., Wu, M.Y., (2019). Identification of the PTEN-ARID4B-PI3K pathway reveals the dependency on ARID4B by PTEN-deficient prostate cancer. *Nature Commun.* **10**, 4332.
19. Bogershausen, N., Wollnik, B., (2018). Mutational Landscapes and Phenotypic Spectrum of SWI/SNF-Related Intellectual Disability Disorders. *Front. Mol. Neurosci.* **11**, 252.
20. Wilsker, D., Probst, L., Wain, H.M., Maltais, L., Tucker, P. W., Moran, E., (2005). Nomenclature of the ARID family of DNA-binding proteins. *Genomics* **86**, 242–251.

21. Blackledge, N.P., Klose, R.J., (2021). The molecular principles of gene regulation by Polycomb repressive complexes. *Nature Rev. Mol. Cell Biol.* **22**, 815–833.
22. Shandala, T., Kortschak, R.D., Saint, R., (2002). The Drosophila retained/dead ringer gene and ARID gene family function during development. *Int. J. Dev. Biol.* **46**, 423–430.
23. Baba, A., Ohtake, F., Okuno, Y., Yokota, K., Okada, M., Imai, Y., et al., (2011). PKA-dependent regulation of the histone lysine demethylase complex PHF2-ARID5B. *Nature Cell Biol.* **13**, 668–675.
24. Gong, W., Zhou, T., Mo, J., Perrett, S., Wang, J., Feng, Y., (2012). Structural insight into recognition of methylated histone tails by retinoblastoma-binding protein 1. *J. Biol. Chem.* **287**, 8531–8540.
25. Lai, A., Kennedy, B.K., Barbie, D.A., Bertos, N.R., Yang, X. J., Theberge, M.C., et al., (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.
26. Masuda, K., Ripley, B., Nishimura, R., Mino, T., Takeuchi, O., Shioi, G., et al., (2013). Arid5a controls IL-6 mRNA stability, which contributes to elevation of IL-6 level in vivo. *Proc. Natl. Acad. Sci. USA* **110**, 9409–9414.
27. Altenhoff, A.M., Train, C.M., Gilbert, K.J., Mediratta, I., Mendes de Farias, T., Moi, D., et al., (2021). OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res.* **49** D373-D9.
28. Katoh, K., Standley, D.M., (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780.
29. Needleman, S.B., Wunsch, C.D., (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
30. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., Barton, G.J., (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191.
31. Li, H., Coghlan, A., Ruan, J., Coin, L.J., Heriche, J.K., Osmotherly, L., et al., (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34**, D572–D580.
32. Holm, L., (2020). DALI and the persistence of protein shape. *Protein Sci.* **29**, 128–140.
33. Blum, M., Chang, H.Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al., (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354.
34. He, S., Wu, Z., Tian, Y., Yu, Z., Yu, J., Wang, X., et al., (2020). Structure of nucleosome-bound human BAF complex. *Science* **367**, 875–881.
35. Vinogradova, M., Gehling, V.S., Gustafson, A., Arora, S., Tindell, C.A., Wilson, C., et al., (2016). An inhibitor of KDM5 demethylases reduces survival of drug-tolerant cancer cells. *Nature Chem. Biol.* **12**, 531–538.
36. Johansson, C., Velupillai, S., Tumber, A., Szykowska, A., Hookway, E.S., Nowak, R.P., et al., (2016). Structural analysis of human KDM5B guides histone demethylase inhibitor development. *Nature Chem. Biol.* **12**, 539–545.
37. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al., (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*.
38. Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L.E., et al., (2018). Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **27**, 112–128.
39. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., et al., (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350.
40. Zahn-Zabal, M., Dessimoz, C., Glover, N.M., (2020). Identifying orthologs with OMA: A primer. *F1000Res.* **9**, 27.
41. Yang, I.S., Son, H., Kim, S., Kim, S., (2016). ISOexpresso: a web-based platform for isoform-level expression analysis in human cancer. *BMC Genom.* **17**, 631.
42. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., Goldstein, D.B., (2013). Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709
43. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., et al., (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533.
44. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., et al., (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067.
45. Turner, T., (2013). Plot protein: visualization of mutations. *J. Clin. Bioinforma* **3**, 14.
46. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., et al., (2020). Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688.
47. Centore, R.C., Sandoval, G.J., Soares, L.M.M., Kadoch, C., Chan, H.M., (2020). Mammalian SWI/SNF Chromatin Remodeling Complexes: Emerging Mechanisms and Therapeutic Strategies. *Trends Genet.* **36**, 936–950.
48. Mashtalir, N., D'Avino, A.R., Michel, B.C., Luo, J., Pan, J., Otto, J.E., et al., (2018). Modular Organization and Assembly of SWI/SNF Family Chromatin Remodeling Complexes. *Cell.* **175** 1272–88 e20.
49. Mashtalir, N., Suzuki, H., Farrell, D.P., Sankar, A., Luo, J., Filipovski, M., et al., (2020). A Structural Model of the Endogenous Human BAF Complex Informs Disease Mechanisms. *Cell.* **183** 802-17 e24.
50. Longbotham, J.E., Chio, C.M., Dharmarajan, V., Trnka, M. J., Torres, I.O., Goswami, D., et al., (2019). Histone H3 binding to the PHD1 domain of histone demethylase KDM5A enables active site remodeling. *Nature Commun.* **10**, 94.
51. Torres, I.O., Kuchenbecker, K.M., Nnadi, C.I., Fletterick, R. J., Kelly, M.J., Fujimori, D.G., (2015). Histone demethylase KDM5A is regulated by its reader domain through a positive-feedback mechanism. *Nature Commun.* **6**, 6204.
52. Zhang, Y., Yang, H., Guo, X., Rong, N., Song, Y., Xu, Y., et al., (2014). The PHD1 finger of KDM5B recognizes unmodified H3K4 during the demethylation of histone H3K4me2/3 by KDM5B. *Protein Cell.* **5**, 837–850.
53. Wang, G.G., Song, J., Wang, Z., Dormann, H.L., Casadio, F., Li, H., et al., (2009). Haematopoietic malignancies caused by dysregulation of a chromatin-binding PHD finger. *Nature* **459**, 847–851.
54. Klein, B.J., Piao, L., Xi, Y., Rincon-Arano, H., Rothbart, S. B., Peng, D., et al., (2014). The histone-H3K4-specific

demethylase KDM5B binds to its substrate and product through distinct PHD fingers. *Cell Rep.* **6**, 325–335.

55. Longbotham, J.E., Kelly, M.J.S., Fujimori, D.G., (2021). Recognition of Histone H3 Methylation States by the PHD1 Domain of Histone Demethylase KDM5A. *ACS Chem. Biol.*.
56. Hassan, F., Ramelot, T.A., Yang, Y., Cort, J.R., Janjua, H., Kohan, E., et al., (2013). Solution NMR structure of PHD type Zinc finger domain of Lysine-specific demethylase 5B (PLU-1/JARID1B) from Homo sapiens. *Northeast Structural Genomics Consortium (NESG) Target HR7375C*.
57. Wu, R.C., Jiang, M., Beaudet, A.L., Wu, M.Y., (2013). ARID4A and ARID4B regulate male fertility, a functional link to the AR and RB pathways. *Proc. Natl. Acad. Sci. USA* **110**, 4616–4621.
58. Lei, M., Feng, Y., Zhou, M., Yang, Y., Loppnau, P., Li, Y., et al., (2018). Crystal structure of chromo barrel domain of RBBP1. *Biochem. Biophys. Res. Commun.* **496**, 1344–1348.
59. Fleischer, T.C., Yun, U.J., Ayer, D.E., (2003). Identification and characterization of three new components of the mSin3A corepressor complex. *Mol. Cell Biol.* **23**, 3456–3467.
60. Wu, M.Y., Tsai, T.F., Beaudet, A.L., (2006). 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.
61. Gong, W., Wang, J., Perrett, S., Feng, Y., (2014). Retinoblastoma-binding protein 1 has an interdigitated double Tudor domain with DNA binding activity. *J. Biol. Chem.* **289**, 4882–4895.
62. Ren, J., Yao, H., Hu, W., Perrett, S., Gong, W., Feng, Y., (2021). Structural basis for the DNA-binding activity of human ARID4B Tudor domain. *J. Biol. Chem.*, 100506.
63. Iwahara, J., Iwahara, M., Daughdrill, G.W., Ford, J., 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.
64. Kim, S., Zhang, Z., Upchurch, S., Isern, N., 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.
65. Cai, S., Zhu, L., Zhang, Z., Chen, Y., (2007). Determination of the three-dimensional structure of the Mrf2-DNA complex using paramagnetic spin labeling. *Biochemistry* **46**, 4943–4950.
66. Tu, S., Teng, Y.C., Yuan, C., Wu, Y.T., Chan, M.Y., Cheng, A.N., et al., (2008). The ARID domain of the H3K4 demethylase RBP2 binds to a DNA CCGCCC motif. *Nature Struct. Mol. Biol.* **15**, 419–421.
67. Liu, G., Huang, Y.J., Xiao, R., Wang, D., Acton, T.B., Montelione, G.T., (2010). Solution NMR structure of the ARID domain of human AT-rich interactive domain-containing protein 3A: a human cancer protein interaction network target. *Proteins* **78**, 2170–2175.
68. Hata, K., Takashima, R., Amano, K., Ono, K., Nakanishi, M., Yoshida, M., et al., (2013). Arid5b facilitates chondrogenesis by recruiting the histone demethylase Phf2 to Sox9-regulated genes. *Nature Commun.* **4**, 2850.
69. Yamakawa, T., Whitson, R.H., Li, S.L., Itakura, K., (2008). Modulator recognition factor-2 is required for adipogenesis in mouse embryo fibroblasts and 3T3-L1 cells. *Mol. Endocrinol.* **22**, 441–453.
70. Garton, J., Barron, M.D., Ratliff, M.L., Webb, C.F., (2019). New Frontiers: ARID3a in SLE. *Cells* **8**
71. Le Bouffant, R., Cunin, A.C., Buisson, I., Cartry, J., Riou, J. F., Umbhauer, M., (2014). Differential expression of arid5b isoforms in Xenopus laevis pronephros. *Int. J. Dev. Biol.* **58**, 363–368.
72. Yang, N., Xu, R.M., (2013). Structure and function of the BAH domain in chromatin biology. *Crit. Rev. Biochem. Mol. Biol.* **48**, 211–221.
73. Fan, H., Lu, J., Guo, Y., Li, D., Zhang, Z.M., Tsai, Y.H., et al., (2020). BAHCC1 binds H3K27me3 via a conserved BAH module to mediate gene silencing and oncogenesis. *Nature Genet.* **52**, 1384–1396.
74. Kuo, A.J., Song, J., Cheung, P., Ishibe-Murakami, S., Yamazoe, S., Chen, J.K., et al., (2012). The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier-Gorlin syndrome. *Nature* **484**, 115–119.
75. Ren, W., Fan, H., Grimm, S.A., Kim, J.J., Li, L., Guo, Y., et al., (2021). DNMT1 reads heterochromatic H4K20me3 to reinforce LINE-1 DNA methylation. *Nature Commun.* **12**, 2490.
76. Duan, D., Systemic, A.A.V., (2018). Micro-dystrophin Gene Therapy for Duchenne Muscular Dystrophy. *Mol. Ther.* **26**, 2337–2356.
77. Tillotson, R., Selfridge, J., Koerner, M.V., Gadalla, K.K.E., Guy, J., De Sousa, D., et al., (2017). Radically truncated MeCP2 rescues Rett syndrome-like neurological defects. *Nature* **550**, 398–401.
78. Mugal, C.F., Ellegren, H., (2011). Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. *Genome Biol.* **12**, R58.