A comparative analysis of KMT2D missense variants in Kabuki syndrome, cancers and the general population

DOI:
10.1038/s10038-018-0536-6

Document Version
Accepted author manuscript

A comparative analysis of *KMT2D* missense variants in Kabuki syndrome, cancers and the general population

Víctor Faundes[1,2], Geraldine Malone[3], William G. Newman[1,3], Siddharth Banka[1,3].

1. Division of Evolution & Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester M13 9WL, UK.

2. Laboratorio de Genética y Enfermedades Metabólicas, Instituto de Nutrición y Tecnología de los Alimentos (INTA), Universidad de Chile, Santiago 7830490, Chile.

3. Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University NHS Foundation Trust, Manchester M13 9WL, UK.

Running title: Comparative analysis of *KMT2D* missense variants.

The authors declare no conflicts of interest.

Correspondence

Dr Siddharth Banka

Manchester Centre for Genomic Medicine

St Mary's Hospital

Manchester M13 9WL

United Kingdom

Tel: +44 (0) 161 70 10980; Fax: +44 (0) 161 27 66145

siddharth.banka@manchester.ac.uk.

Abstract

Determining the clinical significance of germline and somatic *KMT2D* missense variants (MVs) in Kabuki syndrome (KS) and cancers can be challenging. We analysed 1 920 distinct *KMT2D* MVs that included 1 535 germline MVs in controls (Control-MVs), 584 somatic MVs in cancers (Cancer-MVs) and 201 MV in individuals with KS (KS-MVs). The proportion of MVs likely to affect splicing was significantly higher for Cancer-MVs and KS-MVs than in Control-MVs ($p=0.000018$). Our analysis identified significant clustering of Cancer-MVs and KS-MVs in the PHD#3 and #4, RING#4 and SET domains. Areas of enrichment restricted to just Cancer-MVs (FYR-C and between amino acids 3 043-3 248) or KS-MVs (Coiled–coil#5, FYR-N and between amino acids 4 995-5 090) were also found. Cancer-MVs and KS-MVs tended to affect more conserved residues (lower BLOSUM scores, $p<0.001$ and $p=0.007$). KS-MVs are more likely to increase the energy for protein folding (higher ELASPIC $\Delta\Delta G$ scores, $p=0.03$). Cancer-MVs are more likely to disrupt protein interactions (higher StructMAn scores, $p=0.019$). We reclassify several presumed pathogenic MVs as benign or as variants of uncertain significance. We raise the possibility of as yet unrecognised 'non-KS' phenotype(s) associated with some germline pathogenic *KMT2D* MVs. Overall, this work provides insights into the disease mechanism of *KMT2D* variants and can be extended to other genes, mutations in which also cause developmental syndromes and cancer.

Keywords: *KMT2D*; Kabuki syndrome; missense variant; protein domain

1 Introduction

Histone lysine methylation defects are an important cause for developmental disorders and cancers (1, 2). *KMT2D* (formerly known as *MLL2* and *ALR*) encodes lysine (K)-specific methyltransferase 2D, which catalyses the mono-, di- and trimethylation of the lysine 4 on histone 3 (H3K4), promoting the expression of its target genes (3). Germline deleterious heterozygous *KMT2D* variants cause Kabuki syndrome type 1 (KS, MIM# 147920), a rare congenital disorder characterized by intellectual disability, growth retardation, distinctive facial features and structural anomalies (4-7). Somatic deleterious *KMT2D* variants have been described in a spectrum of cancers including leukaemias, gastrointestinal and central nervous system tumours (8, 9).

Correct interpretation of *KMT2D* variants is crucial for diagnosis in KS and disease progression in cancers (10, 11). About 80% of deleterious germline *KMT2D* variants are predicted to result in a truncated protein (5) (Figure S1). Germline pathogenic missense *KMT2D* variants are also frequently encountered in KS (4, 5, 12-31). In contrast, only 35% of somatic *KMT2D* variants in cancers are predicted to be protein truncating (Figure S1). Approximately 50% of the somatic variants found in cancers are missense, and the remaining are in-frame insertions/deletions and synonymous variants (32) (Figure S1).

Although limited functional analysis of *KMT2D* variants is now possible, determining the consequences of *KMT2D* missense variants (MVs) in diagnostic setting remains challenging because parental segregation is not always possible, and especially due to incomplete understanding of KMT2D protein structure and its interactions (33-36). Notably, the three-dimensional structure of only the SET domain of the protein is available (PDB entries 4z4p and 4erq) (37). A systematic study of *KMT2D* MVs can, therefore, have significant clinical

benefits and help to distinguish pathogenic from benign germline variants and driver somatic variants from passenger ones. Additionally, this may provide insights into the structure and function of this important protein. Furthermore, the consequences of disease-causing germline and somatic variants can be different. For example, some activating somatic *BRAF* variants cause malignant melanoma (38), while other activating germline *BRAF* variants cause cardiofaciocutaneous syndrome (MIM #115150) (39). Somatic loss-of-function *SMARCA4* variants cause hypercalcemic type small cell carcinoma of the ovary (40) and postulated activating germline *SMARCA4* variants are associated with Coffin-Siris syndrome (MIM #614609) (41). However, germline and somatic *KMT2D* MVs have not previously been systematically compared. Likewise, loss-of-function, dominant negative or activating germline MVs in the same gene can cause different phenotypes or diseases (42-45). Although, all KS-causing *KMT2D* variants are presumed to be loss-of-function, the possibility of other phenotypes resulting from a different spectrum of germline *KMT2D* variants has not been examined. Similarly, loss-of-function, dominant negative or activating somatic MVs can have different consequences (46). However, this aspect has not been explored for *KMT2D* previously. For all these reasons, we performed a comprehensive systematic study of *KMT2D* MVs.

2 Methods

The study design is summarised in Figure 1. The databases and tools used in this study are summarised in Tables S1 and S2.

2.1 Compilation and interpretation of *KMT2D* MVs

93 *KMT2D* MVs reported in control population (Control-MVs) were compiled from the Exome
94 Aggregation Consortium (47) (ExAC, Version 0.3.1) database, the 1000 Genomes (1K-G)
95 Project (48), Database of Single Nucleotide Polymorphisms (dbSNP) (49) and the NHLBI-
96 GO Exome Sequencing Project (ESP) (50). The ExAC data was accessed via
97 http://exac.broadinstitute.org/ and the other data were obtained from the Ensembl version 80-
98 GRCh37. For ExAC, only high-quality and non-flagged sites were included. For analyses, we
99 assumed that Control-MVs did not result in any phenotype.
100
101 *KMT2D* MVs annotated as being identified only in somatic tissue (Cancer-MVs) were
102 compiled from the Catalogue of Somatic Mutations in Cancer (COSMIC) (32) database,
103 version 77.
104
105 *KMT2D* MVs reported in KS (KS-MVs) were obtained from literature (and cross-checked
106 with Human Gene Mutation Database Professional® [HGMD]) (51), ClinVar (52) and our
107 in-house database for Kabuki syndrome test results. Of note, the Manchester Centre for
108 Genomic Medicine has offered diagnostic *KMT2D* genotyping by sequencing since 2012.
109
110 All the Control-MVs, Cancer-MVs and KS-MVs were assessed by the Ensembl Variant
111 Effect Predictor (VEP) (53) to obtain their minor allele frequencies and to identify the
112 variants that were likely to disrupt splicing. EX-SKIP tool (54) was used to identify
113 substitutions that may result in exon skipping in mature transcripts. All MVs predicted not to
114 disrupt splicing were mapped with their frequencies on KMT2D protein domains, regions and
115 motifs (according to UniProt accession number O14686) using the Mutation Mapper tool
116 from the cBio Cancer Genomics Portal (55, 56). For purpose of our analysis, we divided the

regions of the protein sequence that are not part of a specific domain or motif into 19 'no domain' regions (Figure S2).

Next, for all MVs that were predicted not to significantly affect splicing, we generated the Blocks Substitution Matrix Series 62 (BLOSUM62) (57) scores for evolutionary conservation analyses, the Ensemble Learning Approach for Stability Prediction of Interface and Core mutations (ELASPIC) algorithm $\Delta\Delta G$ values (58) for changes to the thermodynamic properties resulting from substitutions, the Structural Mutation Annotation (StructMAn) score (59) for calculating the impact of MVs on the interaction of KMT2D with other proteins and ligands, obtaining the probability-of-disruption scores when possible. The PDB file for the longest chain reported for KMT2D as part of a complex was downloaded from the Protein Data Bank in Europe (37) (PDB entry 4erq) in order to support the analyses given by ELASPIC and StructMAn.

2.2 Statistical Analysis

To study the association between the type of the phenotype and the location of MVs, the likelihood ratio chi-square test was applied. The Z-test with the Bonferroni correction was used to compare the proportion of MVs on each location according to the phenotype. The Kruskal-Wallis test with multiple comparisons was applied to compare the BLOSUM62 scores, ELASPIC $\Delta\Delta G$ and StructMAn interaction scores amongst the phenotypes, which were also described using the median and interquartile range. For all statistical analyses, the IBM SPSS® version 22 programme was used and a two-sided, exact p-value <0.05 was considered as significant.

3 Results

3.1 Compilation of variants

In total we identified 1 920 distinct MVs, which included 1 535 *KMT2D* Control-MVs, 584 *KMT2D* Cancer-MVs and 201 KS-MVs (Table S3). Of note, six MVs were reported in all three groups, 85 were reported in both Cancer-MVs and Control-MVs groups, 83 were reported in both KS-MVs and Control-MVs groups, and 23 were reported in both Cancer-MVs and KS-MVs groups (Figure S3) (Table S3-1).

The MAFs for 1 211/1 535 (78.9%) Control-MVs were <1/10 000, and for 53/1 535 (3.5%) Control-MVs was >1/1 000 (Table S3). The Arg5048 was the most frequently altered amino acid in the Cancer-MVs group (7/584, 1.2%), followed by Arg3582 and Arg3727 (each 5/584, 0.9%) (Table S3). The Arg5179 was the most frequently altered amino acid in the KS-MVs group (8/201, 4%), followed by the Arg5048 and Arg5432 amino acids (each 7/201, 3.5%) (Table S3).

16/1 535 Control-MVs, 14/584 Cancer-MVs and 11/201 KS-MVs were predicted to significantly affect splicing (two of these variants were present in both Control-MVs and KS-MVs groups, and one in both Cancer-MVs and KS-MVs groups) (Table S3-2) (Figure 2). As these variants are likely to result in loss of function by introduction of frameshift, they were excluded from subsequent analyses that were performed on 1,519 Control-MVs, 570 Cancer-MVs and 190 KS-MVs. The proportion of presumed MVs predicted to affect splicing is significantly higher for KS-MVs and Cancer-MVs in comparison with Control-MVs (χ^2=21.88, df=2, p=0.000018). Of these 41 variants that are predicted to disrupt splicing, 6/16 (37.5%) in controls, 8/14 (57.1%) in cancer and 7/11 (63.6%) in KS affect either the first or

last bases of exons, demonstrating a further enrichment of canonical splice-donor and splice-acceptor sites in cancer and KS (Table S3-2) (Figure 2). EX-SKIP tool analysis showed that out of these six Control-MVs, two (c.50C>T and c.5188G>A) did not increase the probability of exon-skipping when compared against wild-type (WT) and the remaining four (c.4131G>C, c.4419G>T, c.4693G>T, c.4694C>T) variants were predicted to result in in-frame exon skipping.

3.2 Location of MVs

We identified several regions of constraint for Control-MVs (Figure 3; Tables 1 and 2). Cancer-MVs clustered in PHD#3, PHD#4, RING#4, FYR-C, and SET domains in comparison with Control-MVs ($p<0.05$) (Tables 1 and 2). Cancer-MVs also clustered specifically between amino acid numbers 3 043-3 248 (No Domain #8 in Figure S2) when compared with Control-MVs and KS-MVs ($p<0.05$) (Table 2). KS-MVs clustered in PHD#3, PHD#4, Coiled–coil#5, RING#4, FYR-N and SET domains when compared with Control-MVs ($p<0.05$) (Tables 1 and 2). KS-MVs also clustered specifically between amino acid numbers 4 995-5 090 (No Domain #16 in Figure S2) when compared with Control-MVs and Cancer-MVs ($p<0.05$) (Table 2).

3.3 Consequences on protein properties

The median BLOSUM score for Control-MVs was -1 (-2;1), for Cancer-MVs was -1 (-2;0), and for KS-MVs was -1 (-2;0) (Figure 4). Overall, the BLOSUM scores for Cancer-MVs and KS-MVs were significantly lower when compared to Control-MVs ($p<0.001$ and $p=0.007$, respectively) (Figure 4).

The ELASPIC $\Delta\Delta G$ score for Control-MVs was 0.76 (0.25;1.07), for Cancer-MVs was 0.89 (0.4;1.46), and for KS-MVs was 0.98 (0.34;2.17) (Figure 4). The ELASPIC $\Delta\Delta G$ scores for KS-MVs were significantly higher when compared to Control-MVs (p=0.03) (Figure 4). No other pairwise comparisons were significant (Figure 4).

The StructMAn score for Control-MVs was 0.17 (0.14;0.26), for Cancer-MVs was 0.32 (0.15;0.42), and for KS-MVs was 0.21 (0.14;0.34) (Figure 4). The StructMAn scores for Cancer-MVs were significantly higher when compared to Control-MVs (p=0.019). No other pairwise comparisons were significant (Figure 4).

4 Discussion

We present a comprehensive analysis of *KMT2D* MVs reported in control populations, cancers and KS. Rare *KMT2D* MVs are frequent in the general population as nearly 80% of Control-MVs have a MAF <1/10 000 (Table S3). Hence, the rarity of a *KMT2D* variant is not a reliable indicator of pathogenicity. This compilation highlights five arginine residues in KMT2D that are recurrently substituted in cancer (Arg5048, Arg3582 and Arg3727) and KS (Arg5048, Arg5179 and Arg5432) (Table S3). Interestingly, Arg5048 is amongst the most frequently mutated residues in both cancer and in KS. Arg5048 and Arg5432 are located outside any recognized domains of the protein (No domain #16 and #18, respectively in Figure S2). The Arg5432Trp substitution has been shown to disrupt the interaction of KMT2D with RBBP5 and ASH2L, and result in loss of its catalytic activity (60). Arg5179 is located in the FYR-N domain, which is a region of around 50-100 amino acids enriched in phenylalanine (F) and tyrosine (Y) found in chromatin-associated proteins (61). Arg3582 and Arg3727 are located in the coiled-coils #3 and #4, respectively. Coiled-coils are a type of

212 secondary structure composed of two or more alpha helices which pack together like a cable.
213 These structures help to position catalytic activities at fixed distance (62).
214
215 Intriguingly, we found that six *KMT2D* MVs have been described in controls, cancers and
216 KS; 85 in cancer and controls; and 83 in KS and controls (Tables S3-1). Several possibilities
217 could account for these MVs being observed in control and disease cohorts. Overlap between
218 controls and cancer MVs could be explained by incorrect curation of germline variants as
219 somatic-only in the COSMIC database or wrongly curated somatic variants as germline
220 benign variants in controls. Overlap between controls and KS-MVs could be explained by
221 incorrect interpretation of pathogenicity of these benign variants in KS. Alternatively, these
222 variants may be causing KS with reduced penetrance. However, incomplete penetrance has
223 never been reported in KS. Notably, in other disorders, somatic mosaicism of truly
224 pathogenic variants in healthy controls has been described (e.g. in Bohring-Opitz syndrome)
225 (63) and this could be another explanation for some overlap observed between KS-MVs and
226 Control-MVs. 65/83 of the overlapping KS-MVs and Control-MVs are located outside the
227 regions of enrichment in KS-MVs, therefore, they are more likely to be benign variants
228 (Table S3-1).
229
230 MVs predicted to alter splicing, those affecting canonical splice-donor and splice-acceptor
231 sites were significantly more frequent in cancer and KS, which is consistent with the loss-of-
232 function mechanism associated with these two disorders (Table S3-2) (Figure 2). These
233 variants in cancer and KS should be more appropriately reclassified as splicing variants.
234
235 Of note, the six Control-MVs affecting the first or last nucleotide of exons are all located at
236 the first half of the gene (exons 2, 13, 16, 17, 18, 21; Figure 2), which should allow the

expression of an alternative protein coding transcript (ENST00000526209.1). The protein encoded by this alternative transcript includes the catalytic SET and Post-SET domains without the PHD-type and RING-type zinc fingers, the SPPPEPEA region, the HMG Box, coiled-coils, the LXXLL motifs and the FYR-N and –C domains (Figure S2). This observation points towards the potential redundancy of the N-terminus of KMT2D, which is consistent with previous observations and may indicate the compensatory capacity of the alternative transcript for normal development (60, 64, 65). Interestingly, 11/16 (68%) *KMT2D* protein-truncating variants (PTVs) reported in ExAC are located in the first half of the gene (from residue 1 to 2,768). This is in contrast with *KMT2D* PTVs in HGMD and COSMIC, where 39% of KS-PTVs, and 53% of Cancer-PTVs are in this region.

We demonstrate significant clustering of Cancer-MVs and KS-MVs in the PHD-type zinc fingers #3 and #4, RING-type zinc finger #4 and SET domains, reflecting the importance of these domains in the function of KMT2D. The PHD (plant homeodomain) fingers are domains of 50–80 amino acids containing a zinc-binding motif that appears in many chromatin-associated proteins, which recognise methylated H3K4 (66). The RING-type zinc fingers are composed of 40-60 amino acids that bind two atoms of zinc, and may mediate protein-protein interactions (67). The SET (Su(var)3-9, Enhancer-of-zeste, and Trithorax) domain is composed of 130-140 amino acids in which resides the methyltransferase activity and the substrate-binding sites (60, 68). This similarity of clustering seen in Cancer-MVs and KS-MVs is strongly suggestive that these variants result in loss-of-function.

We found significant clustering of Cancer-MVs in the FYR-C domain and between residue numbers 3 043-3 248 (No domain #8 in Figure S1). The FYR-C domains have the features similar to those of FYR-N domains (61). Notably, these regions were not enriched for KS-

MVs. The lack of KS-MVs in these regions could be due to the lack of power of our study. Alternatively, these variants may result in dominant-negative or gain-of-function effects, specific to some cancers. We, therefore, specifically looked at the type of cancers reported with Cancer-MVs in the FYR-C domain and between residues 3 043-3 248 (No domain #8). This showed that 87% and 82.1% of the variants detected in the FYR-C domain and No domain #8 regions came from solid cancers, but there was no enrichment for a specific type of cancer (Table S3). Another possibility is that germline MVs in this region may result in a condition different from KS, which has yet to be delineated.

460/570 (80.7%) Cancer-MVs were outside the regions of the protein with statistically significant clustering. Interestingly, 84/460 Cancer-MVs are part of set of overlapping Cancer-MVs and Control-MVs in comparison with only 7/110 Cancer-MVs in the cancer-enriched regions of KMT2D (Table S3-1). Overall, this analysis suggests that a substantial number of these Cancer-MVs, which lie outside the cancer-enriched regions of KMT2D, may not be driver variants but passengers ones.

For KS-MVs we detected significant clustering in the Coiled–coil#5 and FYR-N domains, and in between residue numbers 4 995-5 090 (No Domains #16 in Figure S1), but we did not identify significant clustering in these regions for Cancer-MVs. As MVs in these three regions are likely to result in loss-of-function, the lack of Cancer-MVs in these regions is likely to be due to lack of statistical power.

120/190 KS-MVs were outside the regions of the protein with statistically significant clustering. Of note, 75/120 KS-MVs were also seen in Control-MVs in comparison with only 12/70 KS-MVs in the KS-enriched regions of KMT2D (Table S3-1). Furthermore, 107/120

MVs were either inherited from an apparently unaffected parent or the information on inheritance was unavailable. Taken together, 75 KS-MVs can be classed as benign or variants of uncertain significance when classified according to the American College of Medical Genetics guidelines (69). Finally, the misdiagnosis of KS in some patients might also explain that their phenotypes do not match with their genetic findings, which may be benign. Unfortunately, many KS-MVs were got from sources without a comprehensive individual delineation of the syndrome, and most of those patients were just described as suffering from KS (e.g. ClinVar, Hannibal et al. (12); Van Laarhoven et al. (30)). Therefore, we could not filter patients with a true KS phenotype from those without it.

22 MVs were seen in both KS and cancers (Table S3-1). Of note, 21 of these were present in KS-enriched and/or Cancer-enriched regions. The unique MV that was not part of any of these enriched regions, the p.Arg5340Leu substitution, may abolish the interaction between KMT2D and WDR5 resulting in the complete loss of the H3K4 dimethylation activity of the complex (33, 34). Thus, all the overlapping KS-MVs and Cancer-MVs are highly likely to be pathogenic.

We did not find clustering of pathogenic MVs in a number of recognised domains and motifs in KMT2D such as the SPPPEPEA region, the HMG Box, most coiled-coils (except coiled-coil#5), the LXXLL motifs and the Post-SET domains. The SPPPEPEA region is a poorly characterised sequence of repeats composed by the amino acids Serine (S), Proline (P), Glutamic acid (E) and Alanine (A) (70). The HMG (High mobility group) Box is a sequence of ~75 amino acids that binds DNA (71). The LXXLL (L, Leucine; X, any amino acid) motifs are necessary to activate nuclear receptors, and therefore, to activate transcription (72). The Post-SET domain also contributes to the methyltransferase activity of KMT2D (68). Our

312 results suggest that these regions of KMT2D are more tolerant to variations or that there may
313 be as yet unrecognised phenotypes associated with variants in these regions.
314
315 We found that the Cancer-MVs and KS-MVs tend to affect more conserved residues, KS-
316 MVs increase the energy that the protein needs for folding/interacting, and that Cancer-MVs
317 have a greater probability of disrupting protein interactions. We did not identify significant
318 difference in the ELASPIC $\Delta\Delta G$ scores or StructMAn scores of Cancer-MVs or KS-MVs
319 against Control-MVs, respectively (Figure 4B and C), which could be due to limited
320 available information on dynamics and interaction sites of KMT2D. This is reflected by our
321 observations that the ELASPIC $\Delta\Delta G$ scores and StructMAn interaction scores could be
322 generated for only 222/2 279 MVs and 92/2 279 MVs, respectively. This also limited the
323 analysis of scores according to the locations (e.g. the enriched regions) as most of these
324 values were given for the catalytic, PHD-1 and PHD-2 Zinc fingers domains only (Table S3).
325
326 Although this approach needs confirmation by large-scale functional analyses, which are
327 being described just recently (73), and a better characterisation of the protein structure of
328 KMT2D, a recent study about functional consequences of some MVs in this gene confirms
329 our methodology. Cocciadiferro et al. (34) demonstrated that MVs detected in patients with
330 KS and located on PHD-type zinc fingers #3 and #4 (p.Glu1391Lys, pMet1417Val,
331 p.Ile1428Thr, p.Ser1476Cys), RING-type zinc finger #4 (p.Thr5098Pro), FYR-N
332 (p.Gly5189Arg, p. Trp5217Met) and SET (p.Arg5471Met, p.Glu5425Lys, p. Arg5471Met,
333 p.Tyr5510Asp) domains, and in between residue numbers 4 995-5 090 (No Domain #16;
334 p.Phe5034Val, p.His5059Pro) decreased catalytic activity and/or disrupt the interaction of
335 KMT2D with ASH2L/RbBP5. These are exactly the same regions and domains that our study
336 found to be enriched in KS-MVs when compared to Control-MVs. Two exceptions are PHD-

type zinc finger #5 and Coiled–coil#5 domains. While the p.Gln1522Arg MV in the former also disrupted enzymatic activity and interaction with ASH2L/RbBP5, this domain was not detected to be enriched in KS-MVs in our analysis. This may be explained by the lack of enough MVs detected in patients with KS in this domain. Inversely, no MVs in Coiled–coil#5 were studied by Cocciadiferro et al. (34), which cannot discard this domain as relevant for the function of KMT2D.

Similarly, few Cancer-MVs have been characterised functionally and those findings are also concordant with our results. Zhang et al. (74) demonstrated that MVs detected in patients with lymphomas and located on RING-type zinc finger #4 (p.Cys5092Ser, p.Cys5092Tyr), FYR-C (p.Asp5257Val) and SET (p.Arg5432Trp, p.Asn5437Ser, p.Gly5467Asp) domains decreased catalytic activity of KMT2D. These three domains were found to be enriched in Cancer-MVs when compared to Control-MVs. Other relevant MVs that decreased KMT2D activity in lymphomas were p.Arg5027Leu and p.Leu5056, which are located between residue numbers 4 995-5 090 (No Domain #16). This region was not detected to be enriched in Cancer-MVs in our analysis, which may be explained by the type of cancer studied. Inversely, no MVs in PHD-type zinc fingers #3 and #4, and between residue numbers 3 043-3 248 (No domain #8) were studied by Zhang et al. (74), which cannot discard these domains as relevant for the function of KMT2D.

In conclusion, this compilation can aid analysis of *KMT2D* MVs in diagnostic laboratories. We show that rarity of *KMT2D* variants has limited value in determination of their pathogenicity. We have identified a set of recurrent *KMT2D* MVs in cancer and KS. We show that some presumed *KMT2D* MVs are in fact likely to result in loss of function by introduction of frameshift. This work leads to reclassification of a set of presumed pathogenic

MVs as benign variants or as VUS. We identify regions of the KMT2D protein that demonstrate significant clustering of MVs in cancer and KS within and outside the known domains and regions of the protein. We establish that the mechanism of most pathogenic *KMT2D* Cancer-MVs is loss of function, although other possibilities cannot be ruled out for some atypical Cancer-MVs. We raise the possibility of as yet unrecognised 'non-KS' phenotypes associated with some germline pathogenic MVs. Finally, this work provides insights into the disease mechanism of cancers driven by *KMT2D* mutations and of KS1 (Kabuki syndrome type 1). Future work will be needed to understand the impact of the MVs that could not be examined by the described *in-silico* programmes. Similar analyses in other genes, mutations in which also cause developmental syndromes and cancer, should also be carried out in the future (1, 2).

5 Acknowledgments

Víctor Faundes acknowledges to CONICYT, Chile's National Commission for Scientific and Technological Research, for its scholarship support (grant number 72160007). All the authors acknowledge to the Kabuki Research Fund at Manchester University NHS Foundation Trust.

6 Disclosure statement

The authors declare no conflict of interest.

7 References

1. Faundes V, Newman WG, Bernardini L, Canham N, Clayton-Smith J, Dallapiccola B, et al. Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorders. American journal of human genetics. 2018;102(1):175-87.

2. Copeland RA. Protein methyltransferase inhibitors as precision cancer therapeutics: a decade of discovery. Philos Trans R Soc Lond B Biol Sci. 2018;373(1748).

3. Issaeva I, Zonis Y, Rozovskaia T, Orlovsky K, Croce CM, Nakamura T, et al. Knockdown of ALR (MLL2) reveals ALR target genes and leads to alterations in cell adhesion and growth. Molecular and cellular biology. 2007;27(5):1889-903.

4. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature genetics. 2010;42(9):790-3.

5. Banka S, Veeramachaneni R, Reardon W, Howard E, Bunstone S, Ragge N, et al. How genetically heterogeneous is Kabuki syndrome?: MLL2 testing in 116 patients, review and analyses of mutation and phenotypic spectrum. European journal of human genetics : EJHG. 2012;20(4):381-8.

6. Banka S, Howard E, Bunstone S, Chandler KE, Kerr B, Lachlan K, et al. MLL2 mosaic mutations and intragenic deletion-duplications in patients with Kabuki syndrome. Clinical genetics. 2013;83(5):467-71.

7. Bogershausen N, Wollnik B. Unmasking Kabuki syndrome. Clinical genetics. 2013;83(3):201-11.

8. Ford DJ, Dingwall AK. The cancer COMPASS: navigating the functions of MLL complexes in cancer. Cancer genetics. 2015;208(5):178-91.

9. Rao RC, Dou Y. Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. Nat Rev Cancer. 2015;15(6):334-46.

10. Guo C, Chen LH, Huang Y, Chang CC, Wang P, Pirozzi CJ, et al. KMT2D maintains neoplastic cell proliferation and global histone H3 lysine 4 monomethylation. Oncotarget. 2013;4(11):2144-53.

11. Kim JH, Sharma A, Dhar SS, Lee SH, Gu B, Chan CH, et al. UTX and MLL4 coordinately regulate transcriptional programs for cell proliferation and invasiveness in breast cancer cells. Cancer research. 2014;74(6):1705-17.

12. Hannibal MC, Buckingham KJ, Ng SB, Ming JE, Beck AE, McMillin MJ, et al. Spectrum of MLL2 (ALR) mutations in 110 cases of Kabuki syndrome. American journal of medical genetics Part A. 2011;155A(7):1511-6.

13. Li Y, Bogershausen N, Alanay Y, Simsek Kiper PO, Plume N, Keupp K, et al. A mutation screen in patients with Kabuki syndrome. Human genetics. 2011;130(6):715-24.

14. Micale L, Augello B, Fusco C, Selicorni A, Loviglio MN, Silengo MC, et al. Mutation spectrum of MLL2 in a cohort of Kabuki syndrome patients. Orphanet journal of rare diseases. 2011;6:38.

15. Paulussen AD, Stegmann AP, Blok MJ, Tserpelis D, Posma-Velter C, Detisch Y, et al. MLL2 mutation spectrum in 45 patients with Kabuki syndrome. Human mutation. 2011;32(2):E2018-25.

16. Kokitsu-Nakata NM, Petrin AL, Heard JP, Vendramini-Pittoli S, Henkle LE, dos Santos DV, et al. Analysis of MLL2 gene in the first Brazilian family with Kabuki syndrome. American journal of medical genetics Part A. 2012;158A(8):2003-8.

17. Tanaka R, Takenouchi T, Uchida K, Sato T, Fukushima H, Yoshihashi H, et al. Congenital corneal staphyloma as a complication of Kabuki syndrome. American journal of medical genetics Part A. 2012;158A(8):2000-2.

18. Zarate YA, Zhan H, Jones JR. Infrequent Manifestations of Kabuki Syndrome in a Patient with Novel MLL2 Mutation. Molecular syndromology. 2012;3(4):180-4.

19. Ito N, Ihara K, Tsutsumi Y, Miyake N, Matsumoto N, Hara T. Hypothalamic pituitary complications in Kabuki syndrome. Pituitary. 2013;16(2):133-8.

20. Miyake N, Koshimizu E, Okamoto N, Mizuno S, Ogata T, Nagai T, et al. MLL2 and KDM6A mutations in patients with Kabuki syndrome. American journal of medical genetics Part A. 2013;161A(9):2234-43.

21. Makrythanasis P, van Bon BW, Steehouwer M, Rodriguez-Santiago B, Simpson M, Dias P, et al. MLL2 mutation detection in 86 patients with Kabuki syndrome: a genotype-phenotype study. Clinical genetics. 2013;84(6):539-45.

22. Cheon CK, Sohn YB, Ko JM, Lee YJ, Song JS, Moon JW, et al. Identification of KMT2D and KDM6A mutations by exome sequencing in Korean patients with Kabuki syndrome. J Hum Genet. 2014;59(6):321-5.

23. Giordano P, Lassandro G, Sangerardi M, Faienza MF, Valente F, Martire B. Autoimmune haematological disorders in two Italian children with Kabuki syndrome. Italian journal of pediatrics. 2014;40:10.

24. Micale L, Augello B, Maffeo C, Selicorni A, Zucchetti F, Fusco C, et al. Molecular analysis, pathogenic mechanisms, and readthrough therapy on a large cohort of Kabuki syndrome patients. Human mutation. 2014;35(7):841-50.

25. Dentici ML, Di Pede A, Lepri FR, Gnazzo M, Lombardi MH, Auriti C, et al. Kabuki syndrome: clinical and molecular diagnosis in the first year of life. Archives of disease in childhood. 2015;100(2):158-64.

26. Lin JL, Lee WI, Huang JL, Chen PK, Chan KC, Lo LJ, et al. Immunologic assessment and KMT2D mutation detection in Kabuki syndrome. Clinical genetics. 2015;88(3):255-60.

27. Lindsley AW, Saal HM, Burrow TA, Hopkin RJ, Shchelochkov O, Khandelwal P, et al. Defects of B-cell terminal differentiation in patients with type-1 Kabuki syndrome. The Journal of allergy and clinical immunology. 2015.

28. Liu S, Hong X, Shen C, Shi Q, Wang J, Xiong F, et al. Kabuki syndrome: a Chinese case series and systematic review of the spectrum of mutations. BMC medical genetics. 2015;16:26.

29. Morgan AT, Mei C, Da Costa A, Fifer J, Lederer D, Benoit V, et al. Speech and language in a genotyped cohort of individuals with Kabuki syndrome. American journal of medical genetics Part A. 2015;167(7):1483-92.

30. Van Laarhoven PM, Neitzel LR, Quintana AM, Geiger EA, Zackai EH, Clouthier DE, et al. Kabuki syndrome genes KMT2D and KDM6A: functional analyses demonstrate critical roles in craniofacial, heart and brain development. Human molecular genetics. 2015;24(15):4443-53.

31. Yuen RK, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N, et al. Whole-genome sequencing of quartet families with autism spectrum disorder. Nature medicine. 2015;21(2):185-91.

32. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic acids research. 2015;43(Database issue):D805-11.

33. Shinsky SA, Hu M, Vought VE, Ng SB, Bamshad MJ, Shendure J, et al. A non-active-site SET domain surface crucial for the interaction of MLL1 and the RbBP5/Ash2L heterodimer within MLL family core complexes. Journal of molecular biology. 2014;426(12):2283-99.

34. Cocciadiferro D, Augello B, De Nittis P, Zhang J, Mandriani B, Malerba N, et al. Dissecting KMT2D missense mutations in Kabuki syndrome patients. Human Molecular Genetics. 2018:ddy241-ddy.

35. Aref-Eshghi E, Schenkel LC, Lin H, Skinner C, Ainsworth P, Pare G, et al. The defining DNA methylation signature of Kabuki syndrome enables functional assessment of genetic variants of unknown clinical significance. Epigenetics-Us. 2017;12(11):923-33.

36. Sobreira N, Brucato M, Zhang L, Ladd-Acosta C, Ongaco C, Romm J, et al. Patients with a Kabuki syndrome phenotype demonstrate DNA methylation abnormalities. European journal of human genetics : EJHG. 2017;25(12):1335-44.

37. Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, et al. PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. Nucleic acids research. 2016;44(D1):D385-95.

38. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature. 2002;417(6892):949-54.

39. Sarkozy A, Carta C, Moretti S, Zampino G, Digilio MC, Pantaleoni F, et al. Germline BRAF mutations in Noonan, LEOPARD, and cardiofaciocutaneous syndromes: molecular diversity and associated phenotypic spectrum. Human mutation. 2009;30(4):695-702.

40. Ramos P, Karnezis AN, Craig DW, Sekulic A, Russell ML, Hendricks WP, et al. Small cell carcinoma of the ovary, hypercalcemic type, displays frequent inactivating germline and somatic mutations in SMARCA4. Nature genetics. 2014;46(5):427-9.

41. Tsurusaki Y, Okamoto N, Ohashi H, Kosho T, Imai Y, Hibi-Ko Y, et al. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. Nature genetics. 2012;44(4):376-8.

42. Reijnders MRF, Ansor NM, Kousi M, Yue WW, Tan PL, Clarkson K, et al. RAC1 Missense Mutations in Developmental Disorders with Diverse Phenotypes. American journal of human genetics. 2017;101(3):466-77.

43. Menke LA, study DDD, Gardeitchik T, Hammond P, Heimdal KR, Houge G, et al. Further delineation of an entity caused by CREBBP and EP300 mutations but not resembling

Rubinstein-Taybi syndrome. American journal of medical genetics Part A. 2018;176(4):862-76.

44. Cuvertino S, Stuart HM, Chandler KE, Roberts NA, Armstrong R, Bernardini L, et al. ACTB Loss-of-Function Mutations Result in a Pleiotropic Developmental Disorder. American journal of human genetics. 2017;101(6):1021-33.

45. Riviere JB, van Bon BW, Hoischen A, Kholmanskikh SS, O'Roak BJ, Gilissen C, et al. De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome. Nature genetics. 2012;44(4):440-4, S1-2.

46. Seiler M, Peng S, Agrawal AA, Palacino J, Teng T, Zhu P, et al. Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. Cell reports. 2018;23(1):282-96 e4.

47. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.

48. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

49. Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome research. 1999;9(8):677-9.

50. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493(7431):216-20.

51. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular

genetics, diagnostic testing and personalized genomic medicine. Human genetics. 2014;133(1):1-9.

52. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic acids research. 2016;44(D1):D862-8.

53. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome biology. 2016;17(1):122.

54. Raponi M, Kralovicova J, Copson E, Divina P, Eccles D, Johnson P, et al. Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6. Human mutation. 2011;32(4):436-44.

55. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery. 2012;2(5):401-4.

56. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling. 2013;6(269):pl1.

57. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America. 1992;89(22):10915-9.

58. Berliner N, Teyra J, Colak R, Garcia Lopez S, Kim PM. Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. PloS one. 2014;9(9):e107353.

59. Gress A, Ramensky V, Buch J, Keller A, Kalinina OV. StructMAn: annotation of single-nucleotide polymorphisms in the structural context. Nucleic acids research. 2016;44(W1):W463-8.

60. Li Y, Han J, Zhang Y, Cao F, Liu Z, Li S, et al. Structural basis for activity regulation of MLL family methyltransferases. Nature. 2016;530(7591):447-52.

61. Garcia-Alai MM, Allen MD, Joerger AC, Bycroft M. The structure of the FYR domain of transforming growth factor beta regulator 1. Protein Sci. 2010;19(7):1432-8.

62. Truebestein L, Leonard TA. Coiled-coils: The long and short of it. Bioessays. 2016;38(9):903-16.

63. Carlston CM, O'Donnell-Luria AH, Underhill HR, Cummings BB, Weisburd B, Minikel EV, et al. Pathogenic ASXL1 somatic variants in reference databases complicate germline variant interpretation for Bohring-Opitz Syndrome. Human mutation. 2017;38(5):517-23.

64. Dhar SS, Lee SH, Kan PY, Voigt P, Ma L, Shi XB, et al. Trans-tail regulation of MLL4-catalyzed H3K4 methylation by H4R3 symmetric dimethylation is mediated by a tandem PHD of MLL4. Gene Dev. 2012;26(24):2749-62.

65. Ali M, Hom RA, Blakeslee W, Ikenouye L, Kutateladze TG. Diverse functions of PHD fingers of the MLL/KMT2 subfamily. Biochimica et biophysica acta. 2014;1843(2):366-71.

66. Sanchez R, Zhou MM. The PHD finger: a versatile epigenome reader. Trends in biochemical sciences. 2011;36(7):364-72.

67. Borden KL, Freemont PS. The RING finger domain: a recent example of a sequence-structure family. Current opinion in structural biology. 1996;6(3):395-401.

68. Zhang Y, Mittal A, Reid J, Reich S, Gamblin SJ, Wilson JR. Evolving Catalytic Properties of the MLL Family SET Domain. Structure. 2015;23(10):1921-33.

69. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular

Pathology. Genetics in medicine : official journal of the American College of Medical Genetics. 2015;17(5):405-24.

70. UniProt Consortium. O14686 (KMT2D_HUMAN) 1998 [updated 17 February 2016. Available from: http://www.uniprot.org/uniprot/O14686.

71. Malarkey CS, Churchill ME. The high mobility group box: the ultimate utility player of a cell. Trends in biochemical sciences. 2012;37(12):553-62.

72. Savkur RS, Burris TP. The coactivator LXXLL nuclear receptor recognition motif. J Pept Res. 2004;63(3):207-12.

73. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018.

74. Zhang J, Dominguez-Sola D, Hussein S, Lee JE, Holmes AB, Bansal M, et al. Disruption of KMT2D perturbs germinal center B cell development and promotes lymphomagenesis. Nature medicine. 2015;21(10):1190-8.

8 Figure Legends

8.1 Figure 1: Study design

Summary of steps followed for compilation and analysis of missense variants (MV) in *KMT2D*.

8.2 Figure 2: Presumed *KMT2D* MVs that are likely to disrupt splicing are enriched in Kabuki syndrome and cancer

Variants affecting the first or last three bases of exons (first/last in red, second/second last in orange and third/third last in green) are depicted. Variants seen in Kabuki syndrome are denoted with *, variants seen in cancer are denoted with #, and are placed above the transcript (ENST00000301067.11), whereas control variants are placed below the transcript. The proportion of presumed MVs predicted to affect splicing is significantly higher for KS-MVs and Cancer-MVs in comparison with Control-MVs (χ^2=21.88, df=2, p=0.000018). Within the variants predicted to disrupt splicing, a further enrichment of canonical splice-donor and splice-acceptor sites can be found in cancer and KS (variants in red). Interestingly, the six Control-MVs affecting the canonical splice-donor and splice-acceptor sites either do not increase the probability of exon-skipping or are predicted to result in in-frame exon skipping.

8.3 Figure 3: Specific regions of the KMT2D protein are enriched for missense variants in Kabuki syndrome and cancer

Distributions of *KMT2D* missense variants (MV) seen in (A) control population, (B) cancers, and (C) Kabuki syndrome (KS) is shown. The X-axis shows the length of the KMT2D protein and the location of its domains and regions. The domains and regions are color-coded

and the legend is provided at the bottom of the figure. The enriched regions/domains in cancers or in Kabuki syndrome are highlighted in red brackets in the respective panels. The Y-axis in (A) shows minor allele frequencies of controls and in (B and C) the number of times a specific Cancer-MV or KS-MV was seen in our cohort. (D) Proportion of *KMT2D* missense variants grouped according to domains and regions.

8.4 Figure 4: Cancer and Kabuki syndrome MVs affect more conserved residues, increase KMT2D delta-delta free energy and may disrupt its interaction with other proteins.

Global comparisons of (A) BLOSUM62, (B) ELASPIC ΔΔG and(C) StructMAn scores of missense variants (MV) seen in control population, cancers and Kabuki syndrome. When compared to Control-MVs, Cancer-MVs and KS-MVs have both significantly lower BLOSUM scores, KS-MVs have significantly higher ELASPIC ΔΔG scores, and Cancer-MVs have significantly higher StructMAn scores.

Compilation of missense variants (MV) in *KMT2D*

1K-G
dbSNP
ESP
ExAC

COSMIC

Literature
HGMD
ClinVar
In-house database

Control
n=1535

Cancer
n=584

Kabuki syndrome
n=201

***In-Silico* Analysis**

VEP:
MAF, BLOSUM62
& Variants affecting splicing

Protein Properties:
ELASPIC
StructMAn
UniProt & InterPro Locations

Variants affecting splicing
n=41

Control
n=1519

Cancer
n=570

Kabuki syndrome
n=190

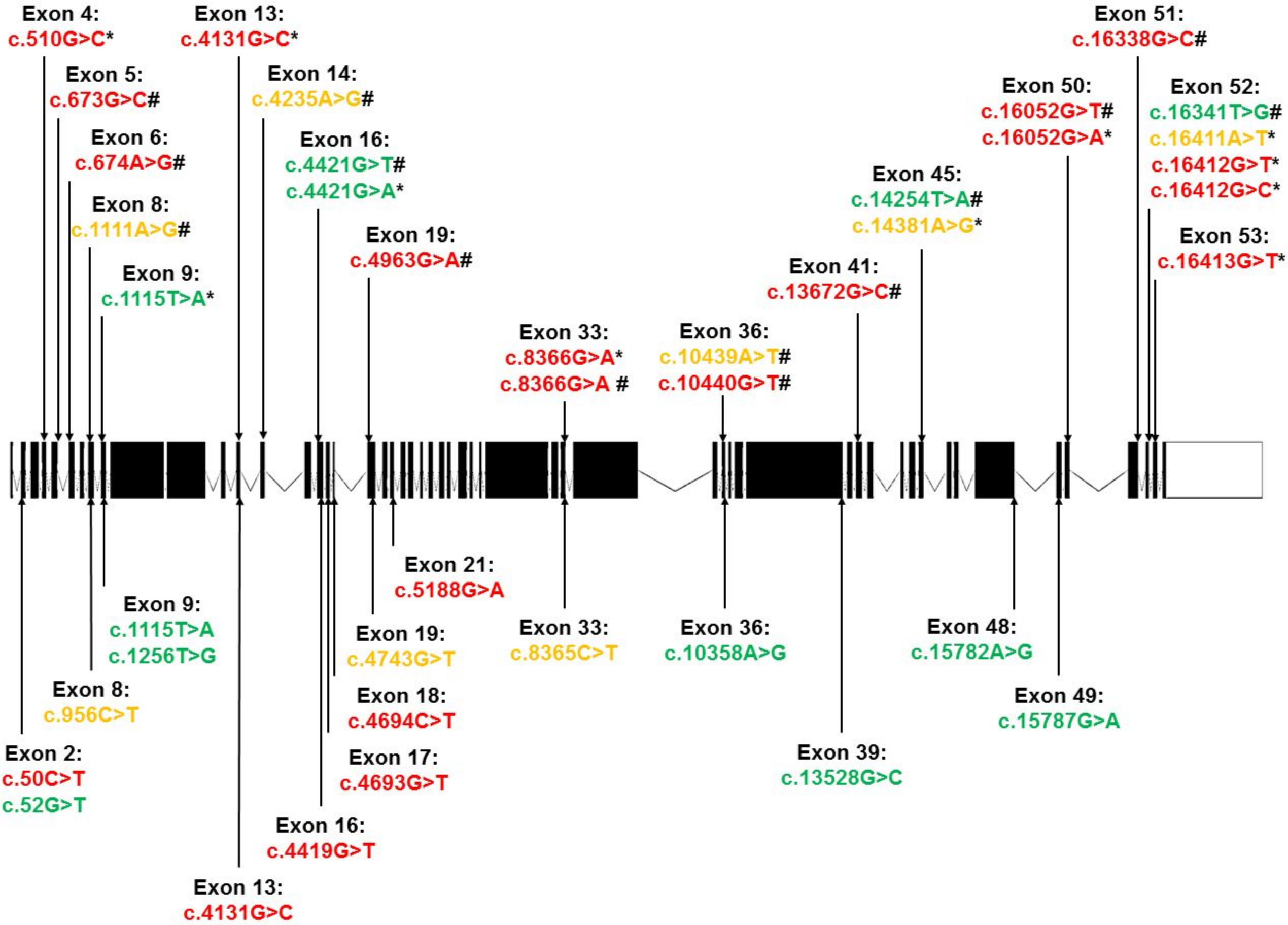
Comparison

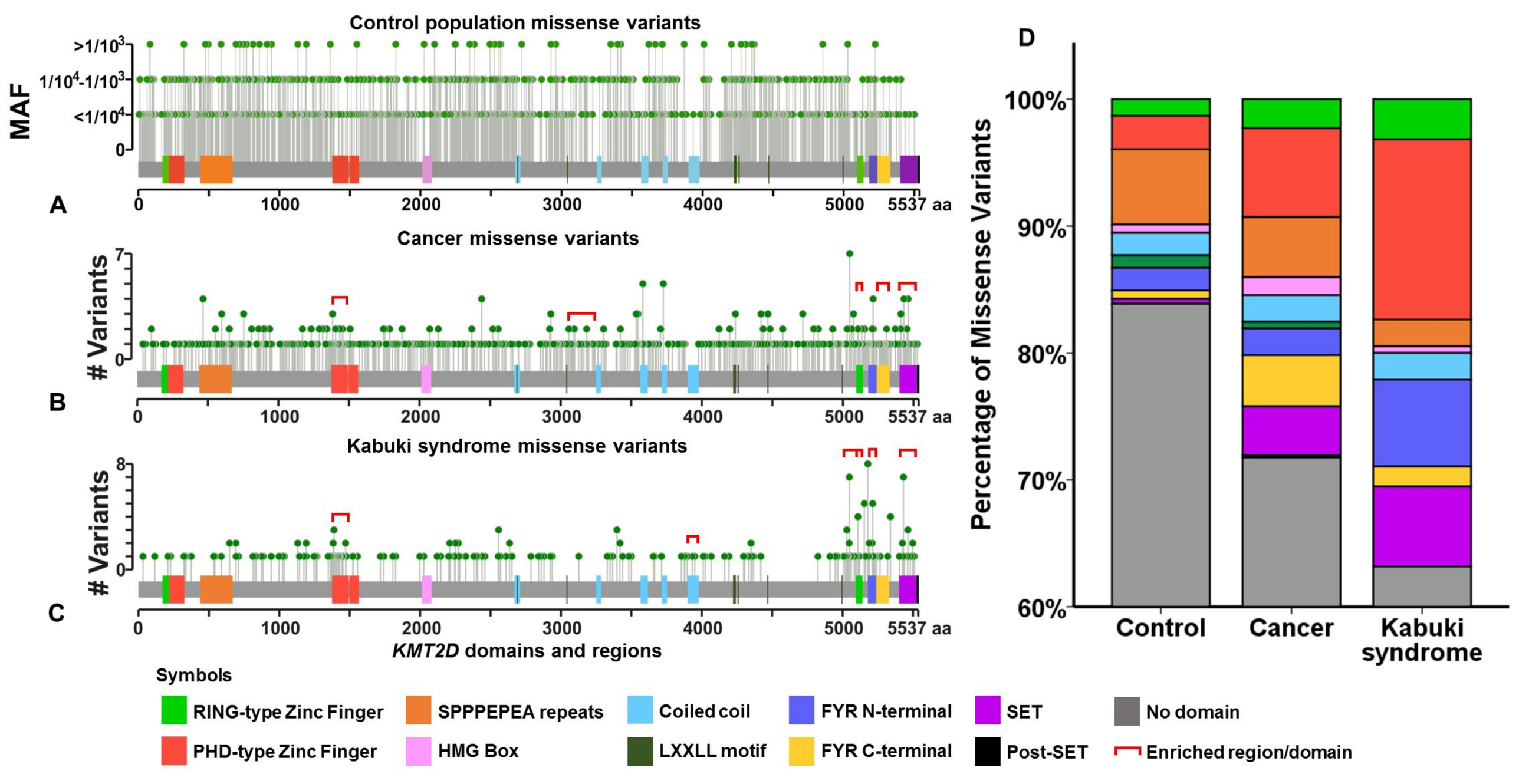
Protein Location
UniProt & InterPro

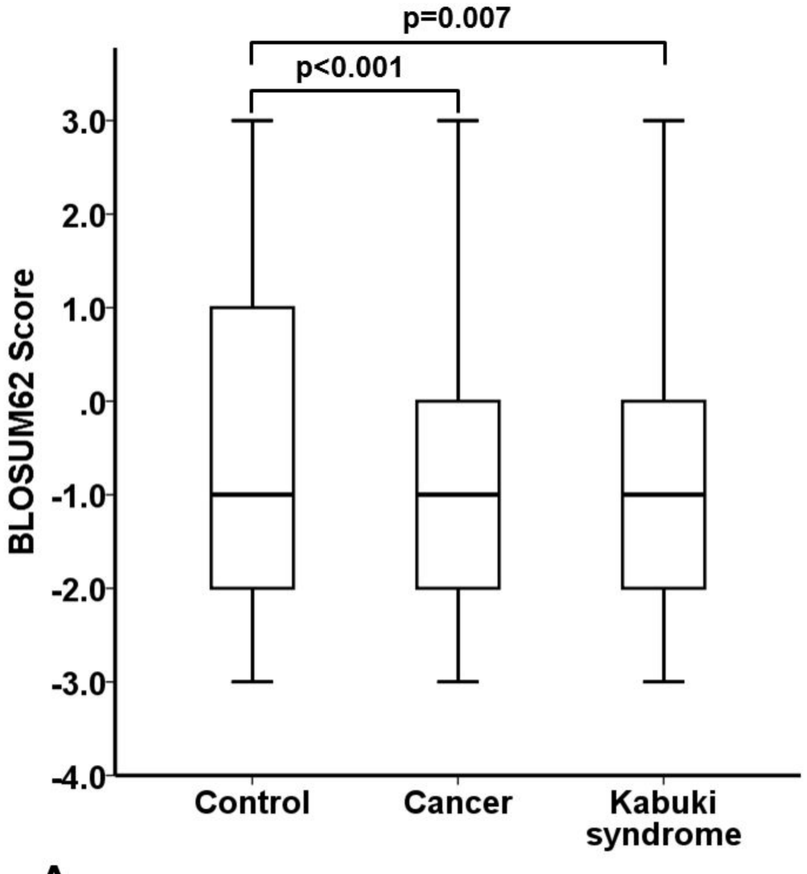
Evolutionary Conservation:
BLOSUM 62 Scores

Thermodynamic Properties
ELASPIC Scores

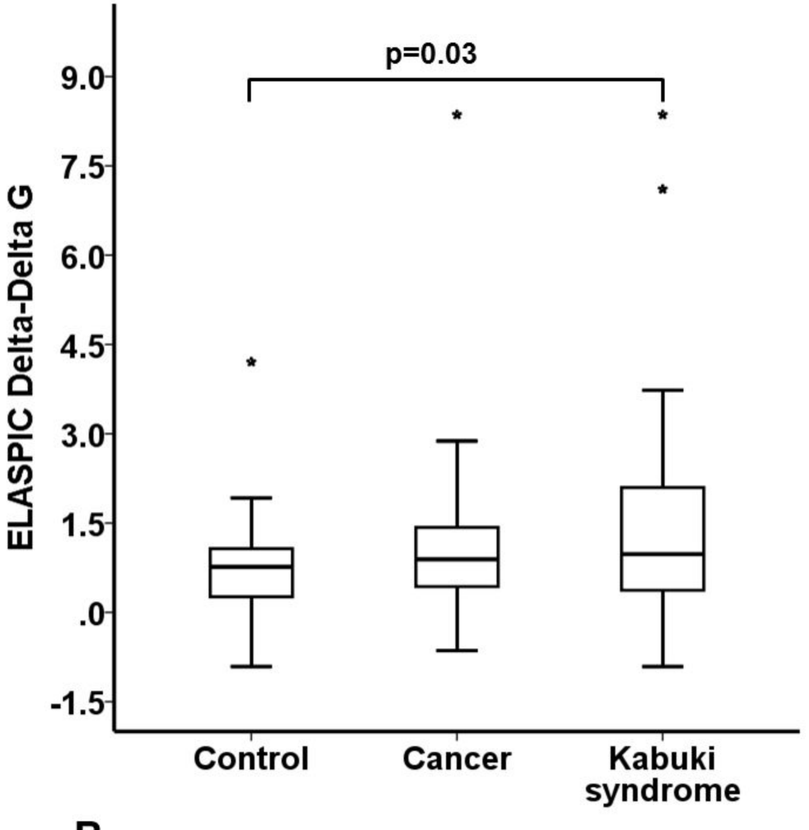
Interaction:
StructMAn Scores


Exon 4:
c.510G>C*
Exon 5:
c.673G>C#
Exon 6:
c.674A>G#
Exon 8:
c.1111A>G#
Exon 9:
c.1115T>A*
Exon 13:
c.4131G>C*
Exon 14:
c.4235A>G#
Exon 16:
c.4421G>T#
c.4421G>A*
Exon 19:
c.4963G>A#
Exon 33:
c.8366G>A*
c.8366G>A #
Exon 36:
c.10439A>T#
c.10440G>T#
Exon 41:
c.13672G>C#
Exon 45:
c.14254T>A#
c.14381A>G*
Exon 50:
c.16052G>T#
c.16052G>A*
Exon 51:
c.16338G>C#
Exon 52:
c.16341T>G#
c.16411A>T*
c.16412G>T*
c.16412G>C*
Exon 53:
c.16413G>T*
Exon 2:
c.50C>T
c.52G>T
Exon 8:
c.956C>T
Exon 9:
c.1115T>A
c.1256T>G
Exon 13:
c.4131G>C
Exon 16:
c.4419G>T
Exon 17:
c.4693G>T
Exon 18:
c.4694C>T
Exon 19:
c.4743G>T
Exon 21:
c.5188G>A
Exon 33:
c.8365C>T
Exon 36:
c.10358A>G
Exon 39:
c.13528G>C
Exon 48:
c.15782A>G
Exon 49:
c.15787G>A

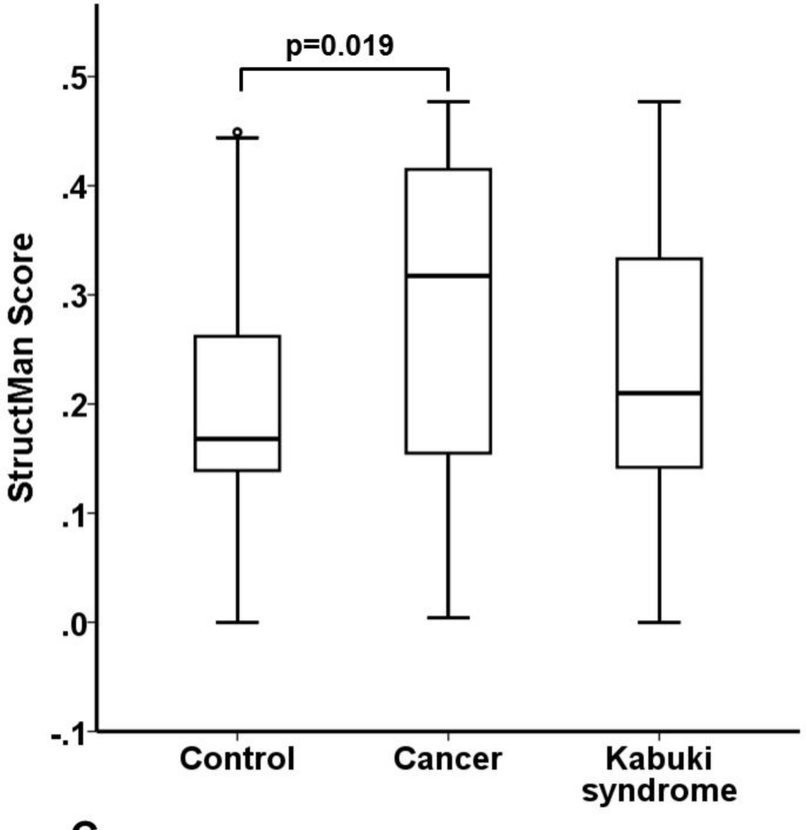

Control population missense variants
MAF
>1/10^3
1/10^4-1/10^3
<1/10^4
0
A
0
1000
2000
3000
4000
5000
5537 aa
Cancer missense variants
Variants
7
0
B
Kabuki syndrome missense variants
8
C
KMT2D domains and regions
D
Percentage of Missense Variants
100%
90%
80%
70%
60%
Control
Cancer
Kabuki syndrome
Symbols
RING-type Zinc Finger
PHD-type Zinc Finger
SPPPEPEA repeats
HMG Box
Coiled coil
LXXLL motif
FYR N-terminal
FYR C-terminal
SET
Post-SET
No domain
Enriched region/domain


p=0.007
p<0.001
BLOSUM62 Score
3.0
2.0
1.0
.0
-1.0
-2.0
-3.0
-4.0
Control
Cancer
Kabuki syndrome

A


p=0.03
ELASPIC Delta-Delta G
9.0
7.5
6.0
4.5
3.0
1.5
.0
-1.5
Control
Cancer
Kabuki syndrome

B


p=0.019
StructMan Score
.5
.4
.3
.2
.1
.0
-.1
Control
Cancer
Kabuki syndrome

C

Table 1. Comparison of proportions of missense variants seen in control population, cancer and Kabuki syndrome according to their grouped locations.

Protein Domain[a]	**Control Population** **n (%)** **(n=1 519)**	**Cancer** **n (%)** **(n=570)**	**Kabuki syndrome** **n (%)** **(n=190)**
RING-type Zinc Finger	20 (1.3)	13 (2.3)	6 (3.2)
PHD-type Zinc Finger	*40 (2.6)*	*40 (7.0)*[b]	*27 (14.2)*[c]
SPPPEPEA region	90 (5.9)	27 (4.7)	4 (2.1)
HMG Box	10 (0.7)	8 (1.4)	1 (0.5)
Coiled coil	27 (1.8)	12 (2.1)	4 (2.1)
LXLL motif	15 (1)	3 (0.5)	0 (0)
FYR-N Terminal	*27 (1.8)*	*12 (2.1)*	*13 (6.8)*[b]
FYR-C Terminal	*10 (0.7)*	*23 (4)*[b]	*3 (1.6)*
SET	*6 (0.4)*	*22 (3.9)*[b]	*12 (6.3)*[b]
Post-SET	0 (0)	1 (0.2)	0 (0)
No Domain	*1 274 (83.9)*	*409 (71.8)*[d]	*120 (63.2)*[d]

[a] In order of location; domains with significantly different proportions amongst the phenotypes (p-adjusted<0.05) are depicted in italic.

[b] Proportion significantly higher than controls.

[c] Proportion significantly higher than the other two groups.

[d] Proportion significantly lower than controls.

Table 2. Comparison of proportions of missense variants seen in control population, cancer and Kabuki syndrome according to every significantly different location.

Protein Domain[a]	**Length of region (delimiting amino acids)**	**Control Population n (%) (n=1 519)**	**Cancer n (%) (n=570)**	**Kabuki syndrome n (%) (n=190)**
PHD-type Zinc Finger #3	57 (1 374-1 430)	5 (0.3)	13 (2.3)[b]	12 (6.3)[c]
PHD-type Zinc Finger #3 & #4	11 (1 420-1 430)	1 (0.1)	3 (0.5)	4 (2.1)[b]
PHD-type Zinc Finger #4	61 (1 420-1 480)	5 (0.3)	13 (2.3)[b]	6 (3.2)[b]
No Domain #5	447 (1 565-2 011)	122 (8.0)	41 (7.2)	5 (2.6)[d]
No Domain #6	588 (2 081-2 668)	232 (15.3)	47 (8.2)[d]	25 (13.2)
No Domain #8	206 (3 043-3 248)	26 (1.7)	28 (4.9)[c]	1 (0.5)
Coiled Coil #5	79 (3 897-3 975)	1 (0.1)	0 (0)	3 (1.6)[c]
No Domain #15	22 (4 468-4 989)	159 (10.5)	55 (9.6)	6 (3.2)[e]
No Domain #16	96 (4 995-5 090)	25 (1.6)	13 (2.3)	15 (7.9)[c]
RING-type Zinc Finger #4	47 (5 091-5 137)	9 (0.6)	11 (1.9)[b]	5 (2.6)[b]
FYR N-Terminal	61 (5 175-5 235)	27 (1.8)	12 (2.1)	13 (6.8)[c]

FYR C-Terminal	92 (5 236-5 327)	10 (0.7)	23 (4.0)[b]	3 (1.6)
SET	123 (5 397-5 519)	6 (0.4)	22 (3.9)[b]	12 (6.3)[b]

[a] In order of location; only domains with significantly different proportions amongst the phenotypes (p-adjusted<0.05) are depicted. For visualisation of these regions, see Figure S2.

[b] Proportion significantly higher than controls.

[c] Proportion significantly higher than the other two groups.

[d] Proportion significantly lower than controls.

[e] Proportion significantly lower than the other two groups.