Blood DNA methylation provides an accurate biomarker of *KMT2B*-related dystonia and predicts onset

Nazanin Mirza-Schreiber,[1,2,†] Michael Zech,[1,3,†] Rory Wilson,[4] Theresa Brunet,[1,3] Matias Wagner,[1,3] Robert Jech,[5] Sylvia Boesch,[6] Matej Škorvánek,[7,8] Ján Necpál,[9] David Weise,[10,11] Sandrina Weber,[1,12] Brit Mollenhauer,[12] Claudia Trenkwalder,[12] Esther M. Maier,[13] Ingo Borggraefe,[13] Katharina Vill,[13] Annette Hackenberg,[14] Veronika Pilshofer,[15] Urania Kotzaeridou,[16] Eva Maria Christina Schwaibold,[17] Julia Hoefele,[3] Melanie Waldenberger,[4] Christian Gieger,[4,18] Annette Peters,[18,19,20] Thomas Meitinger,[3] Barbara Schormair,[1,3] Juliane Winkelmann[1,3,21,22] and Konrad Oexle[1,2,3]

[†]**These authors contributed equally to this work.**

Abstract

Dystonia is a prevalent, heterogeneous movement disorder characterized by involuntarily abnormal postures. Biomarkers of dystonia are notoriously lacking.

Here, a biomarker is reported for histone lysine methyltransferase (KMT2B)-deficient dystonia, a leading subtype among the individually rare monogenic dystonias. It was derived by applying a support vector machine to an episignature of 113 DNA CpG sites which, in blood cells, showed significant epigenome-wide association with KMT2B deficiency and at least 1x log-fold change of methylation. This classifier was accurate both when tested on the general population and on samples with various other deficiencies of the epigenetic machinery, thus allowing for definitive evaluation of variants of uncertain significance and identifying patients who may profit from deep brain stimulation, a highly successful treatment in KMT2B-deficient dystonia.

Methylation was increased in KMT2B deficiency at all 113 CpG sites. The coefficients of variation of the normalized methylation levels at these sites also perfectly classified the samples with KMT2B-deficient dystonia. Moreover, the mean of the normalized methylation levels correlated well with the age at onset of dystonia ($p = 0.003$) – being lower in samples with late or incomplete penetrance – thus serving as a predictor of disease onset and severity.

Similarly, it may also function in monitoring the recently envisioned treatment of KMT2B deficiency by inhibition of DNA methylation.

Author affiliations:

1 Institute of Neurogenomics (ING), Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany

2 Neurogenetic Systems Analysis Group, Institute of Neurogenomics (ING), Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany

3 Institute of Human Genetics, Technical University of Munich, School of Medicine, 81675 Munich, Germany

4 Research Unit Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany

5 Department of Neurology, Charles University, 1st Faculty of Medicine and General University Hospital in Prague, 121 08 Prague, Czech Republic

6 Department of Neurology, Anichstrasse 35, 6020 Innsbruck, Austria

7 Department of Neurology, P. J. Safarik University, 04011 Kosice, Slovakia

8 Department of Neurology, University Hospital L. Pasteur, 04011 Kosice, Slovakia

9 Department of Neurology, Zvolen Hospital, 96001 Zvolen, Slovakia.

10 Department of Neurology, Asklepios Fachklinikum Stadtroda, 07646 Stadtroda, Germany

11 Department of Neurology, University of Leipzig, 04103 Leipzig, Germany

12 University Medical Center Goettingen, Department of Neurology and Paracelsus-Elena-Klinik, 34128 Kassel, Germany

13 Dr. von Hauner Children's Hospital, Ludwig-Maximilians-University, 80337 Munich, Germany

14 Department of Pediatric Neurology, University Children's Hospital, 8032 Zürich, Switzerland

15 Ordensklinikum Linz, Barmherzige Schwestern, 4010 Linz, Austria

16 Department of Child Neurology and Metabolic Medicine, Center for Pediatric and Adolescent Medicine, University Hospital Heidelberg, 69120 Heidelberg, Germany

17 Institute of Human Genetics, Heidelberg University, 69120 Heidelberg, Germany

18 German Center for Diabetes Research (DZD), 85764 Neuherberg, Germany

19 Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany

20 Chair of Epidemiology, Institute for Medical Information Processing, Biometry and Epidemiology, Medical Faculty, Ludwig-Maximilians-Universität München, 81377 Munich, Germany

21 Chair of Neurogenetics, Technical University of Munich, School of Medicine, 81675 Munich, Germany

21 Munich Cluster for Systems Neurology (SyNergy), 81377 Munich, Germany.

Correspondence to: Prof. Dr. Konrad Oexle

Institute of Neurogenomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstr. 1, D-85764 Munich-Neuherberg, Germany

E-mail: konrad.oexle@helmholtz-muenchen.de

Running title: DNAme as biomarker in KMT2B deficiency

Keywords: dystonia; KMT2B; episignature; age at onset; mode of inheritance

Abbreviations: ACMG = American College of Medical Genetics and Genomics; AAO = Age at onset; CpG site = DNA sequence consisting of a cytosine followed by a guanine; CV(z) = coefficient of variation of the normalized methylation levels; DD = developmental delay; ID = intellectual disability; EWAS = epigenome-wide association study; H3K4 = lysine at position 4 of histone H3; KMT2B = histone lysine methyl transferase 2B; KORA = Cooperative Health Research in the Augsburg Region; LoF = loss of function; logFC = log-fold change; mean(z) = mean of the normalized methylation levels; SNP = single nucleotide polymorphism; SVM = support vector machine; VUS = variant of uncertain significance; WES = whole exome sequencing

Introduction

Dystonia is characterized by involuntary abnormal posturing and movements. About 20% of the patients in movement disorder clinics present with dystonia.[1] Their phenotypes are rather heterogeneous with respect to clinical manifestation, presenting comorbidity, and underlying etiology. One or more body regions may be affected, and dystonia may occur as an isolated feature, in combination with other movement disorders, or as a complex disease together with other neurological dysfunctions such as developmental delay (DD).[2–4] Genetic studies have unveiled a large number of different monogenic subtypes. Whole exome sequencing (WES) of 708 dystonia index patients recently indicated 78 distinct monogenic causes in 19% of them.[5] Among the individually rare monogenic causes, heterozygous missense and loss-of-function (LoF) variants in *KMT2B* were found to be the most frequent subtype in this cohort, accounting for 9% of the monogenic diagnoses. *KMT2B*-related disease is characterized by early-onset generalized dystonia, usually associated with intellectual disability (ID) or other neurodevelopmental comorbidities.[6]

KMT2B encodes a widely expressed epigenetic regulator that catalyzes the transfer of methyl groups to the lysine residue (K) at position 4 of histone H3 (H3K4) and, as such, belongs to the "writers" within the epigenetic regulatory machinery. Pathogenic variants of genes that encode for epigenetic writers, readers, erasers, or remodelers of the chromatin have been implicated in a wide range of diseases, frequently including neurodevelopmental defects.[7]

Histone modifications overlap and interact with methylation of CpG sites in the DNA sequence. Specifically, there is a strong anti-correlation between H3K4 trimethylation and DNA methylation.[8,9] Recent studies have examined the effect of mutations affecting the epigenetic machinery on genome-wide DNA methylation, including mutations of writer enzymes as encoded by *KMT2D* and *NSD1*, for instance, which cause the syndromic neurodevelopmental disorders Kabuki syndrome and Sotos syndrome, respectively.[10–18] For about 40 Mendelian disorders, robust disease-specific DNA methylation patterns ("episignatures") have been identified thereby. Each episignature involves a specific set of CpG sites across the genome whose weighted DNA methylation pattern is tightly associated with the respective disease.

As they affect another writer enzyme, we predicted that LoF variants of *KMT2B* might also leave traces in the genome-wide DNA methylation profile. Indeed, we have derived a highly sensitive and specific episignature associated with KMT2B deficiency. Thereby, we provide a much needed, accurate biomarker in the field of dystonias where biomarkers are notoriously lacking.[1] Special relevance arises from the fact that *KMT2B*-related dystonia is exceptionally responsive to deep brain stimulation.[19]

Materials and methods

Study participants

74 patients with dystonia and 115 patients with ID/DD entered the methylation analysis. They were examined by movement-disorder neurologists and human geneticists, and diagnosed in keeping with the respective diseases' clinical consensus definitions.[2] Molecular evaluation by WES of peripheral blood DNA and variant interpretation according to the guidelines of the American College of Medical Genetics and Genomics (ACMG) were performed as detailed elsewhere.[5] 110 ID/DD patient samples passed the quality control after methylation analysis (see below) as well as 69 dystonia patient samples, including 13 samples with pathogenic/likely pathogenic variants in *KMT2B* and 4 samples with variants of uncertain significance (VUS) in *KMT2B*. For details of these patients' ages, sexes, and diagnoses see Supplementary Table 1. The study also included 2413 adult individuals from the general population where common diseases prevail; including 1209 individuals with Restless Legs Syndrome and 1204 participants of the KORA study on the Bavarian population.[20] 2278 of these population samples passed the quality control. The study participants were not related to each other and not known to receive medication with substantial impact on DNA methylation.

Methylation analysis, EWAS, and selection of CpG sites for the episignature

Methylation analysis of peripheral blood DNA was performed by Illumina MethylationEPIC BeadChip (San Diego, CA) according to the manufacturer's protocol. The EPIC probe array interrogates the methylation status at more than 850,000 CpG sites across the genome. Methylation intensities were determined and analyzed using R version 3.6.1 (R Core Team 2020). Probes with detection p-value > 0.01, probes interrogating the sex chromosomes, probes at known single nucleotide polymorphisms (SNPs), cross-reactive probes, and probes with a call rate < 0.95% were excluded from downstream analyses. Samples with mean detection p-value > 0.05 or call rate < 95% were excluded due to poor sample quality. Background correction and normalisation (preprocess Quantile)[21] were done using the minfi package.[22] The methylation level of each CpG site was assessed as beta value (β), indicating the percentage of methylation, and as *M*-value which indicates the log2 ratio of the intensities of methylated and unmethylated states (Du et al 2010),

$$M = \log_2\left(\beta_k/(1-\beta_k)\right) \quad (1)$$

An epigenome-wide association study (EWAS) was performed to discover differentially methylated CpG sites in 12 patients with WES-identified pathogenic or likely pathogenic *KMT2B* variants (sparing one for unbiased verification) versus 8 patients with a genetically proven diagnosis of *SGCE*-associated myoclonus-dystonia (Supplementary Table 1b). We chose the latter as controls because the gene product encoded by *SGCE* (ε-sarcoglycan) is a transmembraneous protein of the cell membrane and as such unlikely to influence the epigenetic machinery. The EWAS using the limma package[23] was a linear regression of *M*-values on *KMT2B* mutation status, sex, age and Houseman estimates of white blood cell type composition, yielding coefficients and eBayes moderated *p*-values.[24] Because mutation status was coded as 0 and 1 in controls and cases, respectively, and because the covariates had little influence, the coefficient of the mutation status approximately equalled the difference of the averages of *M* in the *n* cases (*k*) and in *m* controls (*s*). With

$$\Sigma_k^n M_k/n - \Sigma_s^m M_s/m = \log_2\left(\Pi_k^n\left(\beta_k/(1-\beta_k)\right)^{1/n}\Big/\Pi_s^m\left(\beta_s/(1-\beta_s)\right)^{1/m}\right) \quad (2)$$

we took that coefficient as the level of log-fold change (*logFC*), in keeping with the name ("lfc") of the threshold function in limma. Overall 113 CpG sites with $p < 5x10^{-08}$ according to genome-wide Bonferroni correction and *logFC* > 1 were considered to be sufficiently effective for inclusion in the *KMT2B* episignature (Supplementary Table 2). Analysis of the receiver operating characteristic on mutation carriers vs non-carriers showed an area under the curve (AUC) larger than 0.94 for all 113 CpG sites, with AUC = 1 in 96% of them.

Unsupervised classification

The selected significant 113 CpG sites were examined for their ability to distinguish between carriers of disease-causing *KMT2B* variants and dystonia patients without any suspicious *KMT2B* variant by hierarchical clustering (see Fig.2) using "ward.D" minimization of Euclidean distances as provided by the gplots R package[25] and by multidimensional scaling (MDS) of the pairwise Euclidean distances between samples using limma.[23] Further we applied centroid-based Kmeans clustering[26] to the *KMT2B* mutation carriers and the population cohort by defining successively increasing number of clusters (k = 2 - 7).

SVM classification model for disease-causing *KMT2B* variants

Sets for training and testing the support vector machine (SVM) classifier were selected as follows. Among the 12 WES-diagnosed dystonia patients used in the EWAS, we randomly selected 3 for the test set and the rest for the training set. The test set was augmented by one more patient with a disease-causing *KMT2B* variant who was not part of the EWAS and therefore served as an independent unbiased sample. As controls we used 52 dystonia patients not carrying any suspicious *KMT2B* variant, randomly selecting 14 for the test set and the rest for the training set (Supplementary Table 1b).

Based on the 113 CpG sites determined by the EWAS, the classifier was trained and tested on the *M*-values using the SVM provided in the e1071 R package with linear kernel (cost-C) and 10-fold cross-validation. In keeping with Aref-Eshghi et al.[12], SVM generated decision values were converted to probability scores using Platt's scaling method.[27] Diagnostic application of the classifier included 4 cases with VUS in *KMT2B*.

Independent validation of the classification model

To further validate its specificity, we applied the classifier to the *M*-values of 110 ID/DD samples that had been analysed on the same EPIC array, comprising 72 samples with WES-diagnosed disease-causing variants, including variants in genes affecting the epigenetic machinery, 15 samples with a VUS in some of these genes, and 23 samples without any candidate variant (Supplementary Table 1c). Furthermore, the SVM classifier was applied to the 2278 population samples analysed in 3 separate EPIC array batches.

Training and testing of this classifier implied two decisions: We used only non-KMT2B dystonia patients as controls in the training step and among the KMT2B dystonia patients in the test set there was only one who was not already part of the EWAS. We modified these decisions in case of an alternative classifier which was trained with ID/DD patients as negative controls and was tested on a set in which three of four KMT2B dystonia patients had not already been part of the EWAS (with the EWAS case sample and power thus being somewhat smaller; for details see legend to Supplementary Fig.2).

Mean and coefficient of variation of normalized methylation

We derived the normalized methylation levels of the 113 CpG sites of the episignature, and calculated their mean and coefficient of variation (CV) in each individual. To do so, we first calculated the *z*-values (z_{ij}) for each individual *i* at each site *j*, and then determined

$$mean_i = \sum_j z_{ij}/113 \quad (3)$$

and

$$CV_i = SD_i/|mean_i| = \sqrt{\left(\sum_j \left(z_{ij} - mean_i\right)^2/112\right)}\Big/|mean_i| \quad (4)$$

across the 113 sites in each individual. As usual, z_{ij} indicated the normalized aberration from the mean methylation of site *j* in controls (i.e., the difference from that mean divided by the standard deviation in the controls). Assessing *z*-values compensates for natural differences in methylation and methylation variation between sites. For the same reason, the calculation used the *M*-values of methylation (see above) as they show less heteroscedasticity than the *β*-values.[28] Calculations were performed batch-wise to avoid batch effects. For the batch of

patients with dystonia or ID/DD, we used the same controls as in case of the EWAS (see above). For the batches with the population samples, the calculation of *z*-values was based on all individuals since outliers were not to be expected.

Data availability

Information on the dystonia and ID/DD samples used in this study, and on the 113 CpG sites selected for the episignature is available in the Supplementary material.

Results

EWAS, selection of CpG sites, and unsupervised clustering

The EWAS on 12 cases and 8 controls revealed 3171 CpG sites (Fig.1) associated with disease-causing *KMT2B* variants at an FDR < 0.01. These sites are annotated to 1326 of 25293 genes according to the annotation file "MethylationEPIC_v-1-0_B4.csv" of the supplier. Seven of them overlapped with the set of 78 genes previously found to carry disease-relevant variants in a cohort of 708 dystonia index patients.[5] The overlap was not significant (p = 0.114, 1-sided Fisher's exact test). Regression analysis adjusted for sex, age and blood cell type composition, detected 113 CpG sites that passed the twofold threshold of genome-wide significance and abs(*logFC*) > 1 (Fig.1 and Supplementary Table 2). Hierarchical clustering built on these 113 CpG sites exposed a specific methylation profile, separating carriers of disease-causing variants from non-carriers (Fig.2). A clear separation was also achieved by Kmeans clustering (Supplementary Fig.1). All 113 CpG sites were hypermethylated in carriers relative to non-carriers (Fig.1, 2, 4). 82 (73%) were located inside or nearby one of 69 different genes and 93 (89%) inside or nearby a CpG island (Supplementary Table 2).

13 of the 113 sites overlapped the 3643 CpG sites comprising the episignatures which Aref-Eshghi et al.[12] derived for 34 Mendelian disorders (without reporting each episignature individually). The overlap was significant ($p = 4\text{x}10^{-15}$, Fisher's exact test), but the 100 non-overlapping sites were sufficient to clearly separate disease-causing *KMT2B* variants from other mutations of the epigenetic machinery (see below).

SVM classification model for disease-causing *KMT2B* variants

For each individual, the SVM model together with Platt's scaling[27] generated a probability score between 0 and 1 of having a methylation profile typical of disease-causing *KMT2B* variants. Mean ± standard deviation of the probability score in the training set were 0.916 ± 0.053 in the individuals with *KMT2B* mutation and 0.016 ± 0.006 in those without. In the test set, all 4 individuals with *KMT2B* mutation also ranged well above 0.5 (0.932 ± 0.028) while those without *KMT2B* mutation, including carriers of disease-causing variants in genes affecting the epigenetic machinery such as *DNMT1* and *MORC2*,[7,17] scored well below 0.5 at 0.093 ± 0.185, confirming the high accuracy of the classifier.

Having trained and tested the classifier for specific recognition of the effects of KMT2B deficiency, we performed further assessment in four independent sets. These sets did not overlap with the samples used for EWAS determination of the episignature's CpG sites or for training and testing the classifier. The first set consisted of 110 samples with a diagnosis of DD/ID and a disease-causing mutation in various genes affecting the epigenetic machinery, including epigenetic writers such as *KMT2A*, *KMT2D*, and *NSD1*,[7] samples with VUS in some of these genes, and samples without WES-detectable candidate variants (Supplementary Table 1c). These 110 samples were analysed within the same EPIC array batch as the training and test samples. None of them received a probability score larger than 0.23 and most of them were close to zero (Fig.3) which further confirmed the excellent specificity of the classifier.

The other three cohorts together contained 2278 population samples. These samples were analysed in three independent EPIC array batches. According to their origin we did not expect positive episignature scores. In fact, except one KORA sample, all samples had scores well below 0.5. See below for a detailed discussion of the outlier.

All dystonia, ID/DD and population samples kept their positive or negative classifications when we applied an alternative SVM classifier which was trained with ID/DD samples as negative controls and which was tested on a set that included three (instead of one) among four KMT2B dystonia samples that were entirely unbiased for not being part of the EWAS already (see Supplementary Fig.2).

Specificity and sensitivity of the classifier

Even if assuming that the KORA sample with a score > 0.5 was a false-positive, the classifier still had a specificity of almost 100% (99.96% with Wilson's 95%-CI being [99.72%, 100%]). Unbiased assessment of the sensitivity was restricted by the number of available samples with disease-causing *KMT2B* variants. One such sample in the test set was not part of the EWAS. It was clearly recognized with a score of 0.925. For the four *KMT2B* mutant samples in the test set, the sensitivity was estimated to be 100% with Wilson's 95%-CI being [51%, 100%]. For training and test set combined, the same estimate had a 95%-CI of [77%, 100%].

Mean and CV of normalized methylation

Because KMT2B deficiency implied hypermethylation of all CpG sites in the episignature, we aimed for representing this hypermethylation in a single quantity that accounts for the fact that methylation levels and their variations differ between sites. Therefore, we normalized the methylation level at each site and calculated each individual's $mean(z) \pm SD(z)$ of these normalized levels. Moreover, since KMT2B deficiency should affect all associated sites - while other deficiencies likely affect only subsets of these sites - we also calculated their coefficient of variation $CV(z) = SD(z)/|mean(z)|$. We chose $CV(z)$ instead of $SD(z)$ because natural processes tend to be Poisson like with $SD(z) \approx |mean(z)|$. As shown in Fig.4, $mean(z)$ and $CV(z)$ perfectly separated the samples with disease-causing *KMT2B* variants from those without ($p < 2.2 \times 10^{-16}$).

Four samples with Sotos syndrome due to *NSD1* mutation included in the study were found to be strongly hypomethylated ($mean(z) \pm SD(mean(z)) = -3.83 \pm 0.27$) when compared to the other non-*KMT2B* samples (0.39 ± 0.99) examined on the same array ($p = 1.7 \times 10^{-06}$, 2-sided t-test, Fig.4). This finding is in keeping with the previous observation that genome-wide 99.3% of all CpG sites are hypomethylated in NSD1 deficiency.[15]

Correlation of mean normalized methylation with age at onset

In patients with disease-causing *KMT2B* variants, the age at onset (AAO) was significantly correlated ($p = 0.003$, linear regression) with $mean(z)$ as shown in Fig.5. We also observed

that *mean*(*z*) tended to be lower in patients with missense variants. Accordingly, we confirmed the correlation between type of variant and AAO which has been reported recently.[19] AAO was 5.3 ± 2.6 years (mean ± SD) in patients with LoF variants vs 12.0 ± 5.6 years in patients with missense variants (p = 0.03, one-sided rank test; Fig.5).

Mean(*z*) explained r^2 = 57% of the AAO variance in KMT2B-deficient dystonia patients. With transformation Z'=arctanh(r) and bias correction (R-function pwr.r.test), this predicts that replication of the finding at a significance level of 0.05 and with a power of 80% would require 11 observations.

As compared to patients with complex KMT2B-deficient dystonia, patients with isolated KMT2B-deficient dystonia tended to have a lower mean of the normalized CpG site methylation but the difference was not significant (average of 5.57 vs 6.18, p = 0.3, one-sided rank test).

Application to individual cases

We applied the classifier to 4 patients with VUS in *KMT2B* (Fig.3, 4, Table 1). The first was a 9-year-old boy with generalized isolated dystonia since the age of 6 years. WES revealed 3 heterozygous missense variants in this patient. Two variants (c.659G>A, p.Arg220Gln, paternal; and c.2368G>A, p.Glu790Lys, maternal) were also found in gnomAD controls and mapped to a N-terminal protein region in which no disease-causing variants have been reported thus far.[19] By contrast, the third variant, c.7693C>G, p.Arg2565Gly, was not observed in any genomic control database and mapped to a C-terminal mutational hotspot associated with *KMT2B*-related dystonia (CADD score of 25.1). This variant was inherited from an apparently unaffected father, resulting in a classification as a VUS according to the ACMG standards. We applied the SVM classifier to this patient and his unaffected parents in order to verify or exclude a disease-causing effect of that variant. Interestingly, probability score = 0.80 and *CV*(*z*) = 0.54 were in perfect accordance with those of other patients with pathogenic or likely pathogenic *KMT2B* variants. Moreover, the carrier father also showed values in that range, i.e., 0.75 and 0.51, respectively, while they were entirely normal in the mother (0.046 and 1.40, respectively). *Mean*(*z*) values were also similarly elevated in father and son (4.33 and 4.44, respectively, while completely normal, i.e., 1.11, in the mother), and thus appeared to be somewhat too low for the AAO in the affected patient and too high in the

unaffected father when compared to other patients (Fig.5). Assuming that the father remains unaffected, this variant shows incomplete phenotypic penetrance as has been observed before in patients with missense variants.[19]

The next patient was a 31-year-old female presenting with a combined dystonia-parkinsonism syndrome since the age of 20 years. Two WES-identified deletions in *PRKN* were compatible with a diagnosis of juvenile Parkinson disease type 2 (MIM:600116), but we wondered whether a novel heterozygous VUS in her *KMT2B* gene (c.364-2A>G, p.?) also contributed to the phenotype. Of note, the *KMT2B* c.364-2A>G variant affected a canonical splice site, suggestive of a loss-of-function allele. However, probability score = 0.037, *CV*(*z*) = 2.34, and *mean*(*z*) = 0,88 were all normal (Fig.4). We therefore concluded that this variant was unlikely to be a major driver of the observed disease phenotype and re-classified it as "likely benign".

In a 52-year-old female with generalized isolated dystonia and AAO = 7, WES had revealed the heterozygous *KMT2B* variant c.5336G>A, p.Arg1779Gln.[4] The relevance of this VUS has remained inconclusive.[4] The available pedigree was not informative. The probability score (0.126) was clearly below 0.5, and *CV*(*z*) = 0.93 was substantially larger than those of patients with obvious *KMT2B* etiology. *Mean*(*z*) = 1.73 appeared to be too low for the early AAO (Fig.5). While these values did not indicate a causative effect of this variant, they were conspicuously high when compared to the non-*KMT2B* patients on the same array (97% quantile), especially in view of the prior probability due to phenotype and WES finding, so that a doubt remained about the benign nature of the variant.

The heterozygous *KMT2B* variant c.4622C>T, p.Ala1541Val in a 60-year-old male with generalized isolated dystonia and AAO = 43 was also observed in his 2 unaffected daughters[4], but due to the possibility of incomplete penetrance it may be the cause of the patient's dystonia. The DNA methylation profile revealed a probability score (0.339) clearly below 0.5 and a *CV*(*z*) = 0.93 substantially larger than those of patients with obvious *KMT2B* etiology. Comparatively low *mean*(*z*) = 2.73 was in keeping with the late AAO. Moreover, these values were conspicuously high when compared to the non-*KMT2B* patients on the same array (99% quantile of the probability scores derived on the same array). With the phenotype being consistent with *KMT2B*-related dystonia and the rare *KMT2B* variant being the only suspicious WES finding, the variant may thus be causative, having late or incomplete penetrance.

Three other samples on the same array showed a probability score below 0.5 but above the 97%-quantile (Fig.3). Their phenotypes did not include dystonia and they did not show potentially causative variants in *KMT2B*.

In the KORA population cohort[20] we identified a 65-year-old female with probability score = 0.812 and $CV(z) = 0.50$, indicating KMT2B deficiency, and with comparatively low $mean(z) = 2.60$ suggesting late or incomplete penetrance. Indeed, the she did not report a movement problem and had not visited a doctor during the last 12 months. She selected "moderate" when asked about strong, moderate, minimal or absent problems with concentration, and when asked about memory problems and problems with understanding TV programs she selected "a little more" from 3 possible answers ("none", "a little more" or "much more" problems than in the past) in both categories. 12% of her age group (60 – 68 years) indicated at least her level of deficit in all three categories. WES revealed neither any potentially pathogenic *KMT2B* variant nor any likely causative variant in any other gene. This, however, did not exclude the possibility of a non-coding variant with dysregulatory effect.

Discussion

KMT2B functions as a writer in the epigenetic machinery, catalyzing the transfer of methyl groups to the lysine at position 4 of histone 3 (H3K4).[7] Heterozygous LoF of *KMT2B* affects genes essential for central nervous development and function, and has been found to be one of the most frequent causes of monogenic dystonia.[5,6,19,29] Since DNA methylation and histone H3K4 methylation show marked anti-correlation[8,9], we predicted that inborn deficiency of *KMT2B* should result in an aberrant DNA methylation profile that would be useful as biomarker. Indeed, when we performed an EWAS on KMT2B deficiency, we derived an episignature of 113 differentially methylated CpG sites with genome-wide significance and at least one log-fold change that allowed the training of a SVM classifier of excellent sensitivity and specificity. In keeping with the anti-correlation of DNA and H3K4 methylation, all 113 sites showed increased methylation in KMT2B-deficient patients.

Episignatures of peripheral blood cell DNA have been derived for several distinct syndromes caused by deficiencies of the epigenetic machinery.[10–16,18] Most of these episignatures, including those related to *KMT2C* and *KMT2D*, were derived in a recent large study.[12] Thirteen of the CpG sites found in these episignatures overlapped with the 113 CpG sites of

the *KMT2B*-associated episignature presented here. Therefore, we assessed DNA methylation profiles of patients with various deficiencies of the epigenetic machinery, including samples from patients with deficiencies of *KMT2C* or *KMT2D* (Fig.3). None of them were classified as having a disease-causing *KMT2B* variant, thus highlighting the outstanding specificity of the *KMT2B*-associated classifier which was further confirmed when we applied the classifier to 2278 individuals from the general population in whom pathogenic *KMT2B* variants were not to be expected. The sensitivity of the episignature also appeared to be excellent. All patients with pathogenic or likely pathogenic *KMT2B* variants showed closely similar scores well above 0.5 (Fig.3, 4).

Besides the probability score of the SVM classifier, we produced two other useful quantifiers. *Mean*(*z*) of the normalized methylation levels at the episignature's CpG sites proportionally integrates the effect of KMT2B deficiency on them, while their coefficient of variation *CV*(*z*) aims for the homogeneity of this effect, which at first approximation should be independent of the degree of KMT2B deficiency. Both quantifiers properly separated the positive from the negative cases, the *CV*(*z*) even as accurately as the SVM classifier, indicating that a rather simple assessment of disease-associated CpG sites may already function sufficiently in diagnostics and in cost-efficient screening of dystonia patients.

Mean(*z*) turned out to correlate significantly ($p = 0.003$) with AAO, being larger if AAO was early. Together with the tendency of *mean*(*z*) to be lower in samples with LoF of *KMT2B* ($p = 0.056$), these observations complied well with the recently reported correlation of AAO with LoF in *KMT2B*.[19] As a predictor of AAO, *mean*(*z*) can contribute to predictive and precise patient care. We assume that similar predictors should be achievable for other diseases caused by enzyme deficiencies within the epigenetic machinery.

Episignatures have proven useful in definitive classification of VUS.[17] This was also true for VUS in *KMT2B*-associated dystonia. In one patient we could thus show that the VUS did not contribute to the patient's dystonia-parkinsonism. In another patient, where a VUS was still questionable because the phenotypically unaffected father also carried it, index patient and father were classified as affected, thus validating the pathogenicity of the VUS. The case of the father confirmed the recent observation of *KMT2B* variants with late or incomplete penetrance.[19] Two other VUS (p.Arg1779Gln and p.Ala1541Val) could not be classified definitively, however, even after examining the episignature. They were classified as negative

but they scored conspicuously high (≥ 97 percentile of the individuals without *KMT2B* variant). At least p.Ala1541Val may possibly be causative with late/incomplete penetrance. Indeed, the rather late AAO of the patient was compatible with her comparatively low *mean*(*z*). If the training of a classifier has not included very late-onset and mild phenotypes, samples with negative scores (< 0.5) but conspicuously high levels as compared to controls, need to be scrutinized by assessing further evidence. This caveat likely applies to all episignature classifiers.

Interestingly, one among 2278 population samples was classified as positive. At age 65 this woman appeared to be unaffected, in keeping with her comparatively low *mean*(*z*). Negative WES analysis suggested a non-coding, potentially regulatory variant acting *in cis* or *in trans*. The patient's subjective report of moderate impairment of memory and ability to concentrate raised the question whether mild KMT2B deficiency may contribute to cognitive decline at older age. Indeed, mice with conditional knockout of *kmt2b* in adult forebrain excitatory neurons show impaired hippocampus-dependent memory function.[30] However, the report put her at the 12 %-quantile of her age group which does not allow for definitive conclusions.

The approximately linear anti-correlation between KMT2B activity as represented by *mean*(*z*) and disease severity as represented by AAO, also sheds some light on the mode of inheritance of *KMT2B*-associated dystonia which is autosomal-dominant. This is at odds with inborn errors of metabolic enzymes that usually have a recessive mode of inheritance since the relation between metabolic intoxication and residual enzyme activity is hyperbolic and not linear. Of interest in this respect, molecular analyses recently suggested that KMT2B alias MLL2 protects developmental genes from repression by repelling polycomb repressive complex and DNA methylation machineries, while H3K4me3 is not underlying the transcriptional regulation of MLL2 targets.[31] Consequently, the authors of that study predicted KMT2B deficiency could be potentially rescued by inhibiting DNA methylation. If so, the assessment of the KMT2B-dependent DNA methylation level by *mean*(*z*) might also serve for monitoring such methylation-modifying therapy.

In summary, we have detected a highly specific and sensitive episignature and derived SVM classifier score, *mean*(*z*), and *CV*(*z*) as the first biomarkers of *KMT2B*-associated dystonia. We demonstrated their utility in identifying, confirming or rejecting the diagnosis, especially in cases where sequencing analysis fails or remains inconclusive, and the utility of *mean*(*z*) to

predict AAO which is a proxy of disease severity. Both utilities are of great importance because *KMT2B*-associated dystonia is relatively frequent among monogenic dystonia cases and can be treated successfully by deep brain stimulation.[19]

Funding

MZ, JW, and BS receive research support from the German Research Foundation (DFG 458949627; ZE 1213/2-1; WI 1820/14-1; SCHO 1644/4-1). SB is a member of the European Reference Network for Rare Neurological Diseases – Project ID No 739510. This study was funded by institutional funding from Technische Universität München, Munich, Germany, Helmholtz Zentrum München, Munich, Germany, and Charles University, Prague, Czech Republic (PROGRES Q27). This study was also funded by the Czech Ministry of Education under grant AZV: NV19-04-00233 and under the frame of EJP RD, the European Joint Programme on Rare Diseases (EJP RD COFUND-EJP N° 825575), as well as the Slovak Grant and Development Agency under contract APVV-18-0547 and the Operational Programme Integrated Infrastructure, funded by the ERDF under No. ITMS2014+: 313011V455.

The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research has been supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Balint B, Mencacci NE, Valente EM, et al. Dystonia. *Nat Rev Dis Prim*. 2018;4(1). doi:10.1038/s41572-018-0023-6

2. Albanese A, Bhatia K, Bressman SB, et al. Phenomenology and classification of dystonia: A consensus update. *Mov Disord*. 2013;28(7). doi:10.1002/mds.25475

3. Fung VSC, Jinnah HA, Bhatia K, Vidailhet M. Assessment of patients with isolated or combined dystonia: An update on dystonia syndromes. *Mov Disord*. 2013;28(7):889-898. doi:10.1002/mds.25549

4. Zech M, Boesch S, Jochim A, et al. Clinical exome sequencing in early-onset generalized dystonia and large-scale resequencing follow-up. *Mov Disord*. 2017;32(4):549-559. doi:10.1002/mds.26808

5. Zech M, Jech R, Boesch S, et al. Monogenic variants in dystonia: an exome-wide sequencing study. *Lancet Neurol*. 2020;19(11):908-918. doi:10.1016/S1474-4422(20)30312-4

6. Zech M, Boesch S, Maier EM, et al. Haploinsufficiency of KMT2B , Encoding the Lysine-Specific Histone Methyltransferase 2B, Results in Early-Onset Generalized Dystonia. *Am J Hum Genet*. 2016;99(6):1377-1387. doi:10.1016/j.ajhg.2016.10.010

7. Bjornsson HT. The Mendelian disorders of the epigenetic machinery. *Genome Res*. 2015;25(10). doi:10.1101/gr.190629.115

8. Ooi SKT, Qiu C, Bernstein E, et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*. 2007;448(7154):714-717. doi:10.1038/nature05987

9. Meissner A, Mikkelsen TS, Gu H, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*. 2008;454(7205):766-770. doi:10.1038/nature07107

10. Aref-Eshghi E, Schenkel LC, Lin H, et al. The defining DNA methylation signature of

Kabuki syndrome enables functional assessment of genetic variants of unknown clinical significance. *Epigenetics*. 2017;12(11). doi:10.1080/15592294.2017.1381807

11. Aref-Eshghi E, Rodenhiser DI, Schenkel LC, et al. Genomic DNA Methylation Signatures Enable Concurrent Diagnosis and Clinical Genetic Variant Classification in Neurodevelopmental Syndromes. *Am J Hum Genet*. 2018;102(1). doi:10.1016/j.ajhg.2017.12.008

12. Aref-Eshghi E, Kerkhof J, Pedro VP, et al. Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in 42 Mendelian Neurodevelopmental Disorders. *Am J Hum Genet*. 2020;106(3). doi:10.1016/j.ajhg.2020.01.019

13. Bend EG, Aref-Eshghi E, Everman DB, et al. Gene domain-specific DNA methylation episignatures highlight distinct molecular entities of ADNP syndrome. *Clin Epigenetics*. 2019;11(1). doi:10.1186/s13148-019-0658-5

14. Breen MS, Garg P, Tang L, et al. Episignatures Stratifying Helsmoortel-Van Der Aa Syndrome Show Modest Correlation with Phenotype. *Am J Hum Genet*. 2020;107(3). doi:10.1016/j.ajhg.2020.07.003

15. Choufani S, Cytrynbaum C, Chung BHY, et al. NSD1 mutations generate a genome-wide DNA methylation signature. *Nat Commun*. 2015;6(1). doi:10.1038/ncomms10207

16. Hood RL, Schenkel LC, Nikkel SM, et al. The defining DNA methylation signature of Floating-Harbor Syndrome. *Sci Rep*. 2016;6(1):38803. doi:10.1038/srep38803

17. Sadikovic B, Levy MA, Kerkhof J, et al. Clinical epigenomics: genome-wide DNA methylation analysis for the diagnosis of Mendelian disorders. *Genet Med*. Published online February 5, 2021. doi:10.1038/s41436-020-01096-4

18. Schenkel LC, Aref-Eshghi E, Skinner C, et al. Peripheral blood epi-signature of Claes-Jensen syndrome enables sensitive and specific identification of patients and healthy carriers with pathogenic mutations in KDM5C. *Clin Epigenetics*. 2018;10(1):21. doi:10.1186/s13148-018-0453-8

19. Cif L, Demailly D, Lin J-P, et al. *KMT2B* -related disorders: expansion of the phenotypic spectrum and long-term efficacy of deep brain stimulation. *Brain*. 2020;143(11). doi:10.1093/brain/awaa304

20. Holle R, Happich M, Löwel H, Wichmann H. KORA - A Research Platform for Population Based Health Research. *Das Gesundheitswes*. 2005;67(S 01):19-25. doi:10.1055/s-2005-858235

21. Touleimat N, Tost J. Complete pipeline for Infinium ® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*. 2012;4(3):325-341. doi:10.2217/epi.12.21

22. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10). doi:10.1093/bioinformatics/btu049

23. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47-e47. doi:10.1093/nar/gkv007

24. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13(1):86. doi:10.1186/1471-2105-13-86

25. Warnes GR, Bolker B, Bonebakker L, et al. Data, gplots: Various R Programming Tools for Plotting. Published 2020. https://cran.r-project.org/package=gplots

26. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Published 2020. https://www.r-project.org/

27. Platt JC. Probabilities for SV Machines. In: Smola AJ, Bartlett PL, Schölkopf B, Schuurmans D, eds. *Advances in Large Margin Classifiers*. MIT Press; 2000:61-74. https://www.semanticscholar.org/paper/Probabilistic-Outputs-for-SVMs-and-Comparisons-to-Platt-Karampatziakis/18a72c64859a700c16685386514c30d70765a63e

28. Du P, Zhang X, Huang C-C, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11(1):587. doi:10.1186/1471-2105-11-587

29. Meyer E, Carss KJ, Rankin J, et al. Mutations in the histone methyltransferase gene KMT2B cause complex early-onset dystonia. *Nat Genet*. 2017;49(2):223-237. doi:10.1038/ng.3740

30. Kerimoglu C, Agis-Balboa RC, Kranz A, et al. Histone-Methyltransferase MLL2 (KMT2B) Is Required for Memory Formation in Mice. *J Neurosci*. 2013;33(8):3452-3464. doi:10.1523/JNEUROSCI.3356-12.2013

31. Douillet D, Sze CC, Ryan C, et al. Uncoupling histone H3K4 trimethylation from developmental gene expression via an equilibrium of COMPASS, Polycomb and DNA methylation. *Nat Genet*. 2020;52(6):615-625. doi:10.1038/s41588-020-0618-1

Figure legends

Figure 1 Volcano plot of EWAS results. Blue dots depict CpG sites with genome-wide significance ($p < 5x10^{-8}$), orange dots sites with a log-fold change abs(*logFC*) > 1, and pink dots the 113 sites with both.

Figure 2 Hierarchical clustering of the SVM training set (columns) consisting of 9 individuals with pathogenic or likely pathogenic *KMT2B* variants (right cluster) and 39 without any suspicious *KMT2B* variant (left cluster), using the 113 differentially methylated CpG sites (rows) selected for the classifier. DNA methylation *M*-values ranged from -5 to 3.5 and indicated the hypermethylation (darker colour) in samples with *KMT2B* mutation.

Figure 3 SVM probability scores for disease-causing *KMT2B* variants. The vertical axis represents scores between 0 and 1, with a score over 0.5 indicating that a sample is more likely to carry a dystonia-causing variant in *KMT2B* than not to carry such a variant. The horizontal axis lists the different genes affected by disease-causing variants or VUS, a category for patients without WES-detectable candidate variant, and a category for individuals representing the general population. Each point represents an individual sample.

Figure 4 Histograms of individual mean and coefficient of variation (CV) of the episignature's normalized methylation levels as compared to the histogram of the SVM classifier's probability score. Especially the CV (displayed as 1/CV) also allowed to identify the 13 samples with dystonia-causing *KMT2B* variant (red) among other samples (grey) with monogenic or suspected monogenic defects, including other deficiencies of the epigenetic machinery. The framed purple samples carried the four *KMT2B*-VUS discussed in the text, i.e., from left to right, c.364-2A>G, p.?; c.5336G>A, p.Arg1779Gln; c.4622C>T, p.Ala1541Val; and 7693C>G, Arg2565Gly. The last one was verified as disease-causing variant by the probability score as well as the CV. The four grey samples with strongly negative mean normalized methylation of less than -3 had NSD1 deficiency.

Figure 5 Pairwise comparisons of age at dystonia onset, type of disease-causing *KMT2B* variant, and mean of the normalized methylation levels. The samples are the same as the samples with disease-causing KMT2B variants displayed in Fig.4. Calculation of the individual mean of the normalized blood cell methylation level at the 113 CpG sites of the

KMT2B episignature is described in the Material and Methods section. Boxplots contain median, interquartile interval (IQR), a lower whisker indicating the smallest observation greater than or equal to lower hinge minus 1.5 * IQR, and an analogous upper whisker.

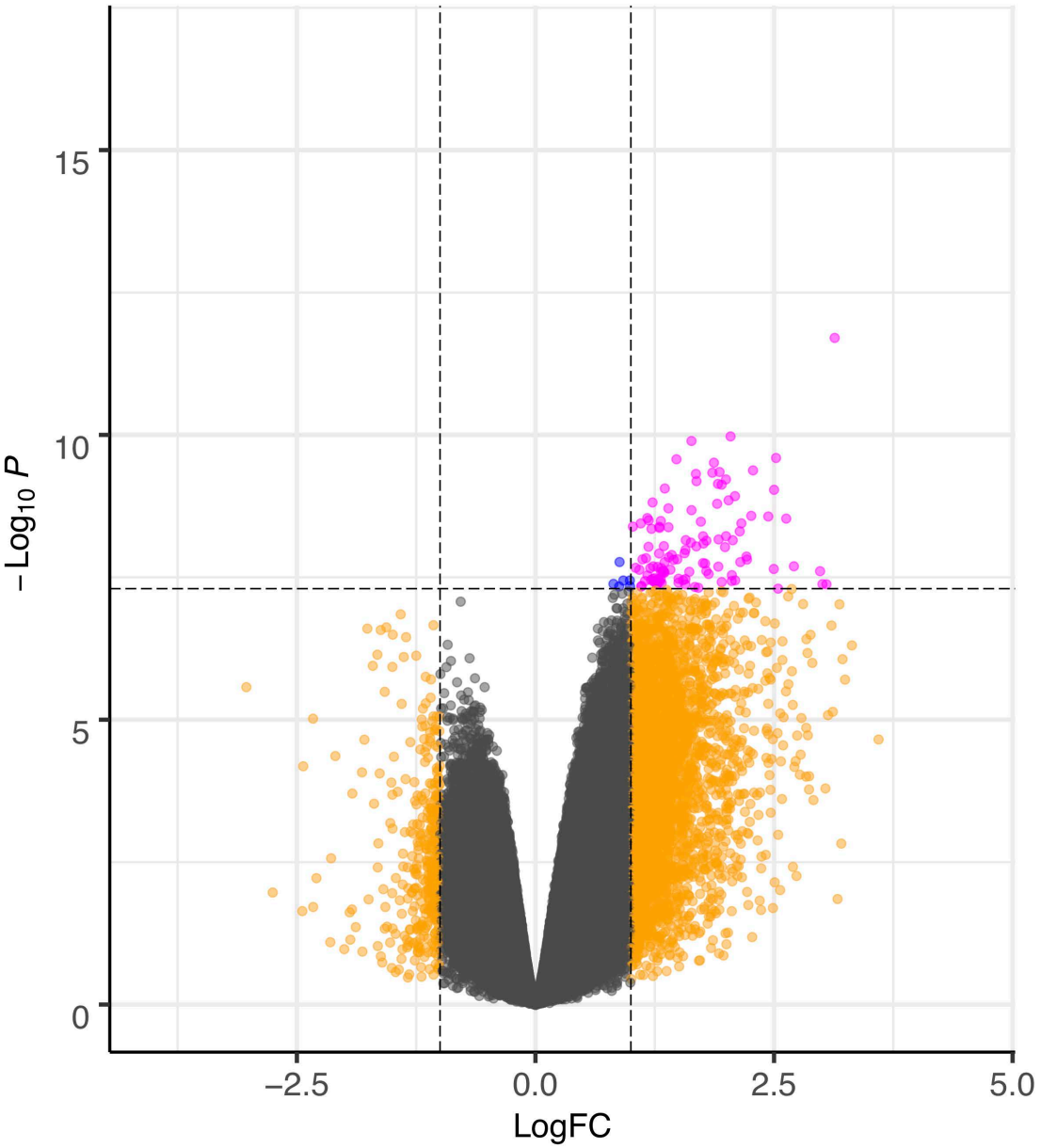

FIGURE 1

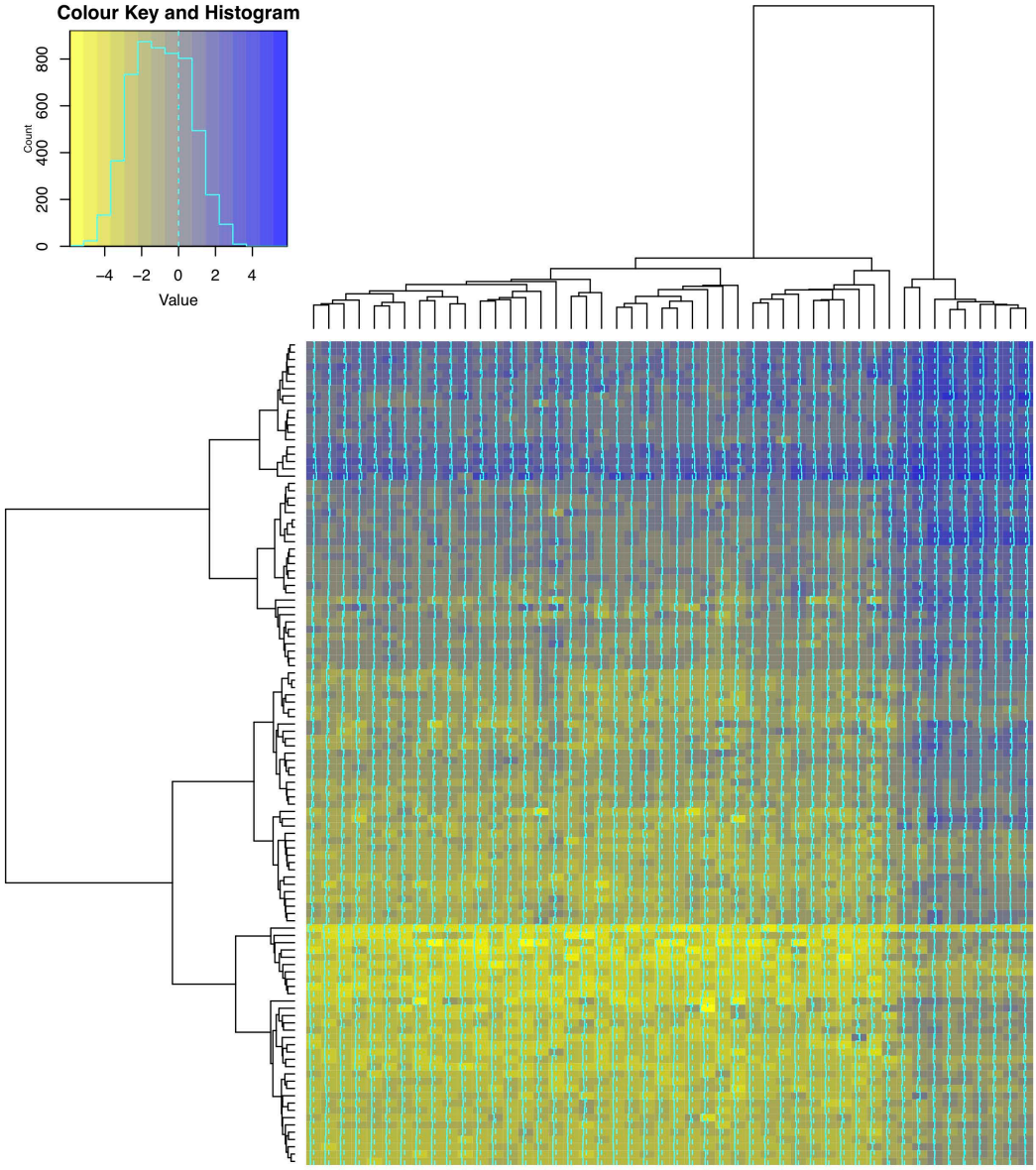

FIGURE 2

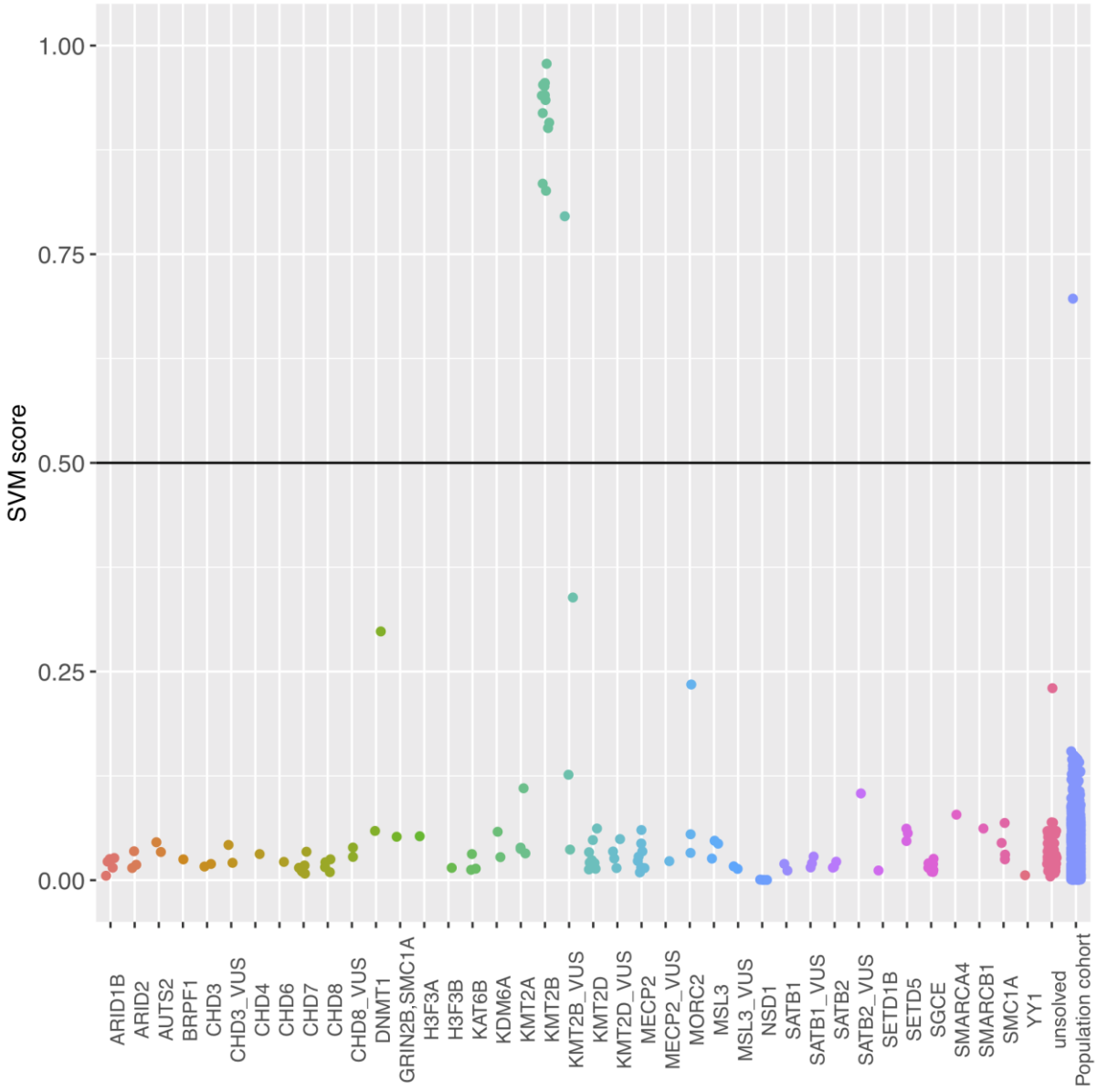

FIGURE 3

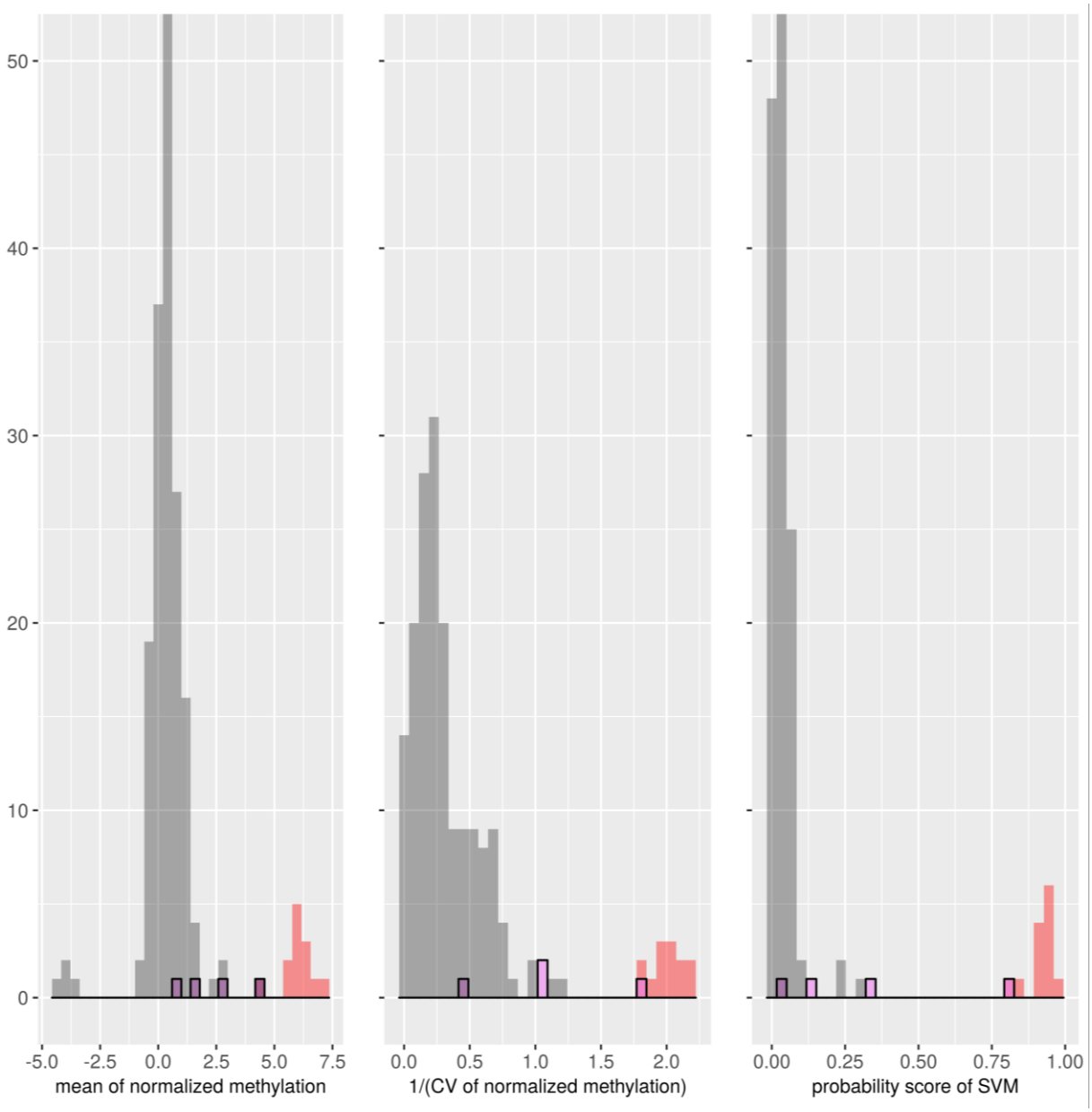

FIGURE 4

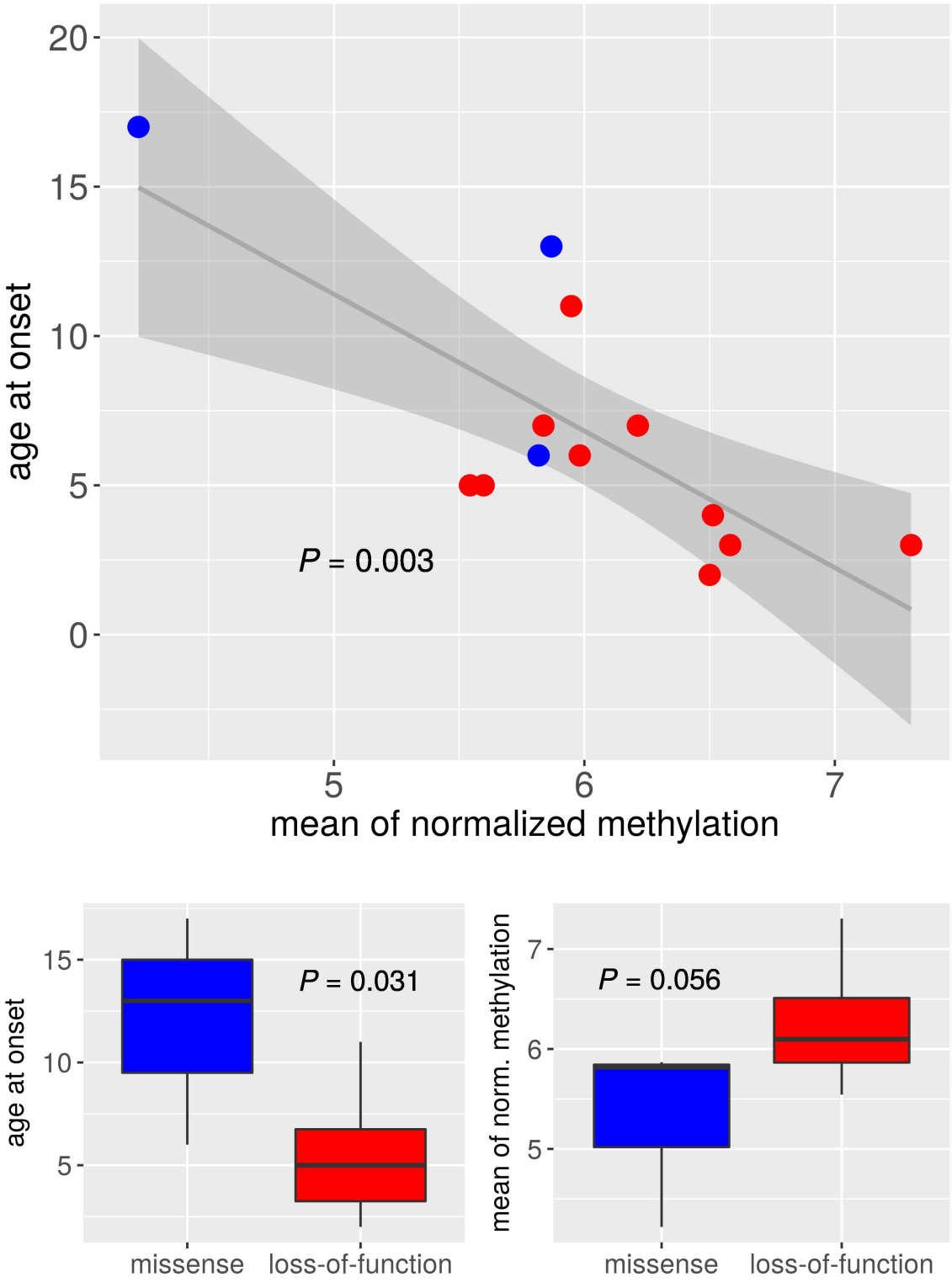

FIGURE 5

Table 1 Overview of different *KMT2B* mutation states

State	*KMT2B* variant	Phenotype	Age at onset, years	SVM score	1/*CV*(*z*)	*Mean*(*z*)	Interpretation
Causative variant	LoF	Mostly complex dystonia	5.3 ± 2.6	0.93 ± 0.04	2.00 ± 0.13	6.20 ± 0.54	Early penetrance
Causative variant	Missense	Isolated/complex dystonia	12.0 ± 5.6	0.89 ± 0.05	2.01 ± 0.12	5.30 ± 0.94	Later penetrance
VUS_1	p.Arg2565Gly	Isolated dystonia	6	0.80	1.85	4.44	Pathogenic with in-complete penetrance or variable AAO
Father of VUS_1		Unaffected	>35	0.75	1.96	4.33	
Mother of VUS_1	None	Unaffected	-	0.05	0.71	1.11	-
VUS_2	c.364-2A>G, p.?	Dystonia + Parkinson's disease	20	0.04	0.43	0.88	Not pathogenic; phenotype due to *PRKN* mutation
VUS_3	p.Arg1779Gln	Isolated dystonia	7	0.13 (=97% quantile)	1.08	1.73	Possibly irrelevant; *mean*(*z*) too low for early AAO
VUS_4	p.Ala1541Val	Isolated dystonia	43	0.34 (=99% quantile)	1.08	2.73	Possibly pathogenic in spite of SVM score; *mean*(*z*) fits late AAO
KORA population case	Not identifiable by WES	Decline of memory and concentration	>64	0.81	2.00	2.60	Non-coding effect in cis or trans; *mean*(*z*) fits non-penetrance
Normal *KMT2B*	None in WES	Various	-	0.03 ± 0.04	0.30 ± 0.24	0.34 ± 0.93	-

Ranges indicated by " ± " represent the standard deviation.