












## SHORT REPORT

# DNA methylation signature classification of rare disorders using publicly available methylation data

Mathis Hildonen<sup>1</sup>  | Marco Ferilli<sup>2</sup>  | Tina Duelund Hjortshøj<sup>3</sup>  |  
Morten Dunø<sup>3</sup>  | Lotte Risom<sup>3</sup>  | Mads Bak<sup>3</sup>  | Jakob Ek<sup>3</sup>  |  
Rikke S. Møller<sup>4,5</sup>  | Andrea Ciolfi<sup>2</sup>  | Marco Tartaglia<sup>2</sup>  | Zeynep Tümer<sup>1,6</sup> 

<sup>1</sup>Kennedy Center, Department of Clinical Genetics, Copenhagen University Hospital, Rigshospitalet, Glostrup, Denmark

<sup>2</sup>Molecular Genetics and Functional Genomics, Ospedale Pediatrico Bambino Gesù, IRCCS, Rome, Italy

<sup>3</sup>Department of Clinical Genetics, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark

<sup>4</sup>Department of Epilepsy Genetics and Personalized Treatment, The Danish Epilepsy Centre, Dianalund, Denmark

<sup>5</sup>Department of Regional Health Research, University of Southern Denmark, Odense, Denmark

<sup>6</sup>Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

## Correspondence

Zeynep Tümer, Kennedy Center, Department of Clinical Genetics, Copenhagen University Hospital, Rigshospitalet, Gl. Landevej 7, 2600 Glostrup, Denmark.

Email: [zeynep.tumer@regionh.dk](mailto:zeynep.tumer@regionh.dk)

## Funding information

A.P. Møller Fonden, Grant/Award Number: 20-L-0371; Familien Hede Nielsens Fond, Grant/Award Number: 2021- 0010; Italian Ministry of Health, Grant/Award Numbers: Ricerca Corrente 2021, 5x1000 2021, RCR-2020-23670068\_001; Italian Ministry of Research, Grant/Award Number: FOE 2019; OUH & RH fælles forskningspulje, Grant/Award Number: A4734

## Abstract

Disease-specific DNA methylation patterns (DNAm signatures) have been established for an increasing number of genetic disorders and represent a valuable tool for classification of genetic variants of uncertain significance (VUS). Sample size and batch effects are critical issues for establishing DNAm signatures, but their impact on the sensitivity and specificity of an already established DNAm signature has not previously been tested. Here, we assessed whether publicly available DNAm data can be employed to generate a binary machine learning classifier for VUS classification, and used variants in *KMT2D*, the gene associated with Kabuki syndrome, together with an existing DNAm signature as proof-of-concept. Using publicly available methylation data for training, a classifier for *KMT2D* variants was generated, and individuals with molecularly confirmed Kabuki syndrome and unaffected individuals could be correctly classified. The present study documents the clinical utility of a robust DNAm signature even for few affected individuals, and most importantly, underlines the importance of data sharing for improved diagnosis of rare genetic disorders.

## KEYWORDS

epigenetics, epismutation, Kabuki syndrome, *KMT2D*, Mendelian disorders, rare disorders, VUS classification

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Clinical Genetics* published by John Wiley & Sons Ltd.

## 1 | INTRODUCTION

During the recent years, analysis of disease/gene-specific DNA methylation (DNAm) patterns has emerged as a highly informative complementary tool to classify the pathogenicity of gene variants for the diagnosis of rare genetic disorders (RGDs).<sup>1,2</sup> Machine learning classifiers (MLCs) which recognize these patterns (DNAm signatures) were initially developed to evaluate genetic variants of uncertain clinical significance (VUS) using methylation arrays. Subsequently, they were used to solve clinically ambiguous cases with a neurodevelopmental phenotype but without a genetic diagnosis,<sup>1–6</sup> to predict the functional consequences of the variants (loss- or gain-of function),<sup>3</sup> and to distinguish the molecular subtypes of a given disorder, for example, Phelan-McDermid syndrome.<sup>7</sup> The number of disorders with DNAm signatures keeps growing, but training MLCs for effective VUS testing for each gene can be rather challenging taking into consideration the large number of RGDs and only few individuals with a given RDG would be identified in a single diagnostic laboratory. Indeed, depending on the effect size of the DNAm changes, a sample size of minimum 10 individuals affected by a given RGD together with age and sex-matched controls are generally necessary to train MLCs,<sup>8</sup> and bona fide representative DNA samples for individual disorders are not always available. A potential way to circumvent this issue is to take advantage of publicly available DNAm datasets and known CpG sites characterizing the disease specific DNAm signature to train the MLC. It is, however, uncertain whether batch effects and datasets originating from different methylation array platforms can influence the ability of the classifier to differentiate between pathogenic and benign variants.

In this study, we investigated whether we could correctly predict the pathogenicity of an unselected panel of pathogenic *KMT2D* variants identified in seven individuals with clinical diagnosis of Kabuki syndrome (OMIM #147920) referred to our department for genetic diagnosis, using publicly available DNAm data (GSE97362),<sup>9</sup> and the previously established DNAm signature for *KMT2D* and Kabuki syndrome defined by 153 CpG sites.<sup>3</sup>

## 2 | MATERIALS AND METHODS

### 2.1 | Subjects, methylation array analysis and public methylation data

Three different datasets, where genome-wide DNAm data was generated using peripheral blood, were included in the study (Table 1):

The internal cohorts 1 and 2 (IC1 and IC2) and public cohort (PC). The affected individuals in IC1 and IC2 were clinically diagnosed with Kabuki syndrome and a pathogenic or likely pathogenic *KMT2D* variant, as evaluated according to the ACMG criteria,<sup>10</sup> was detected in each individual (Supplementary Table 1). DNA of the affected individuals and healthy controls of IC1 and IC2 was bisulfite converted, and genome-wide DNAm levels were quantified using Infinium MethylationEPIC arrays (850K) (Illumina, San Diego, CA). The PC comprised samples from 19 individuals with pathogenic or likely pathogenic *KMT2D* variants and 57 age and sex-matched healthy controls, where genome-wide DNAm data was generated using Infinium 450K arrays (Illumina) and was publicly available (GEO: GSE97362).<sup>9</sup> The present study has been conducted in accordance with the tenets of the Declaration of Helsinki and was approved by the ethical committees of the Capital Region of Denmark (H-22050775) and Ospedale Pediatric Bambino Gesù (ref. 1702\_OPBG\_2018). Detailed methods can be found in Supporting information.

## 3 | RESULTS AND DISCUSSION

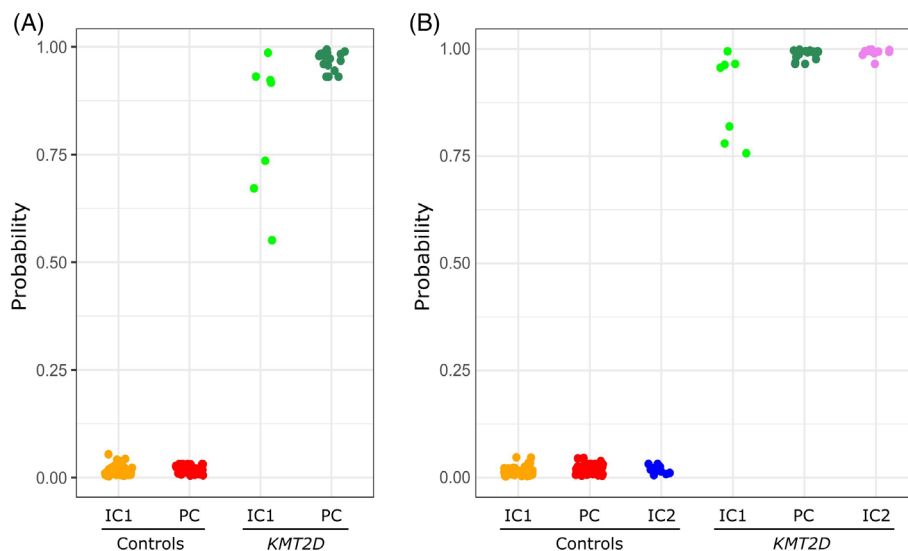
To investigate whether publicly available DNAm data could be used to train a machine-learning model directed to classify a VUS based on DNAm patterns, a public dataset consisting of 19 individuals with *KMT2D* variants and 57 controls (GSE97362, Table 1)<sup>9</sup> was used to train an MLC to classify *KMT2D* variants leading to Kabuki syndrome. The 153 CpG sites used for the classifier were previously published as sites defining the DNAm signature for Kabuki syndrome.<sup>3</sup> A DNAm dataset from an internal testing cohort (IC1, Table 1) was then used to validate the MLC. As different array platforms were used (450K and 850K arrays were used for the public and internal datasets, respectively), the data were merged into a single dataset and normalized to correct the different distribution of methylation values between the two different array platforms prior to the establishment of the classifier (Supplementary Figure 1).

Following a SuperLearner ranking of the tested machine-learning approaches (Supplementary Table 2), a support vector machine (SVM) classifier was trained to classify *KMT2D* variants, and then applied to the testing cohort. Samples receiving a probability score above 0.5 were classified as having a DNAm pattern matching the Kabuki syndrome DNAm signature. Of note, the SVM was able to correctly predict the pathogenicity of all the variants in our testing cohort, as well as all the controls, even though training and testing samples were

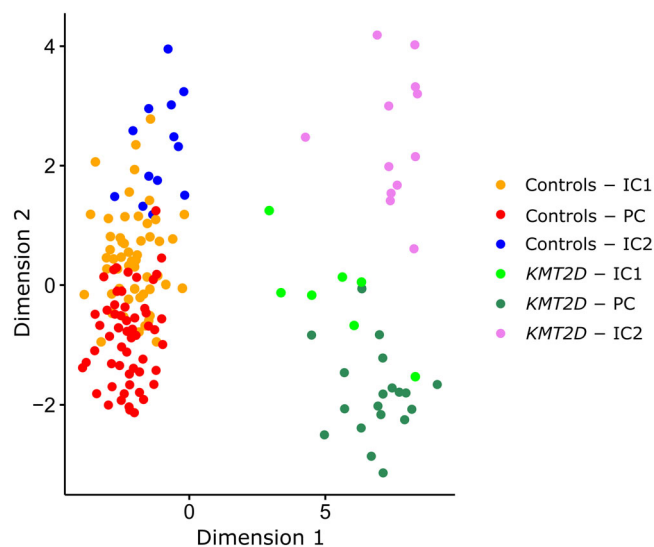
**TABLE 1** Cohorts included in the study.

| Cohort                  | Individuals with pathogenic <i>KMT2D</i> variants (n) | Controls (n) | Array platform | Type            |
|-------------------------|-------------------------------------------------------|--------------|----------------|-----------------|
| Internal cohort 1 (IC1) | 7                                                     | 55           | 850K           | Testing cohort  |
| Internal cohort 2 (IC2) | 12                                                    | 13           | 850K           | Training cohort |
| Public cohort (PC)      | 19                                                    | 57           | 450K           | Training cohort |

Abbreviations: 450K, Infinium 450K arrays covering 450 000 CpG sites (Illumina); 850K, Infinium MethylationEPIC arrays covering approx. 850 000 CpG sites (Illumina); n, number of individuals.



**FIGURE 1** SVM prediction scores. (A) Probability scores of the MLC, which was trained using the public cohort. All the samples from affected individuals in the testing cohort (IC1) received probability scores above 0.5 and their variants were thus correctly classified as pathogenic. (B) The probability scores of the MLC, which was trained using both the public cohort and the IC2 as well as an oversampling step (SMOTE) to balance the data. Probability scores, and thus the prediction confidence for all the samples in the testing cohort increased to above 0.75, supporting the predictions from using only the public cohort for MLC training. IC1, internal cohort 1; IC2, internal cohort 2; PC, public cohort. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]



**FIGURE 2** Multidimensional scaling of public and internal data. Samples from individuals with KMT2D variants clearly separates from the control cluster in the first dimension. There can be seen some separation of samples from the different cohorts in the second dimension, although this separation does not consistently coincide with the array platform the samples were analyzed on. IC1, internal cohort 1; IC2, internal cohort 2; PC, public cohort. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

analyzed using different array platforms (Illumina 450K and 850K arrays, respectively) (Figure 1A).

To assess the impact of the training cohort size on the sensitivity and prediction confidence (probability), MLCs were trained using

smaller subsets of the public cohort (nine or 14 KMT2D samples included from the public cohort and all the 57 PC-controls, Supplementary Table 3). To minimize the impact of sample selection on the results, the training and testing were repeated 10 times for each subset size with random combinations of KMT2D samples, and the results were averaged. Furthermore, an MLC supplemented with additional samples (IC2, Table 1) was trained, and finally a synthetic oversampling step (SMOTE)<sup>11</sup> was added to balance the ratio of affected individuals and controls in the dataset before MLC training. As expected, the prediction confidence increased with the size of the training cohort from an average probability of 0.74 including nine KMT2D samples to 0.84 when all 19 samples from PC was supplemented with the IC2 data (Supplementary Table 3). Balancing the dataset by SMOTE further improved the prediction confidence to an average probability of 0.89 for all samples (Supplementary Table 3, Figure 1B). None of the MLCs misclassified any of the control samples. Surprisingly, the average sensitivity (proportion of pathogenic variants classified correctly) of the MLCs using nine KMT2D samples was higher (0.93) than the average of the classifiers using 14 samples (0.91) (Supplementary Table 3). This indicates that when employing MLCs on samples including more heterogeneous DNAm patterns, the individual DNAm patterns of the samples used in the training cohort are important for the sensitivity of the MLC, suggesting caution in sample selection for small training datasets (e.g., <15 samples).

Multi-dimensional scaling (MDS) analysis was employed to further verify the SVM predictions. Following model evaluation (GOF = 0.50,  $R^2 = 0.92$ ), we used DNAm values from the 153 CpGs to check sample clustering in the new lower-dimensional space. Our analysis

showed that the samples from individuals with *KMT2D* variants were clearly clustered separately from the controls by the first dimension (Figure 2). This means that the largest source of variation at the 153 CpG-sites in this dataset was due to the methylation differences between individuals with Kabuki syndrome and controls. Batch effects, which slightly separated the internal cohort 1, internal cohort 2, and the public cohort in the second dimension further highlighted the importance of a careful evaluation of sources of bias that could affect the sensitivity and specificity of DNAm signatures, as previously reported.<sup>2</sup> Despite the slight separation between IC1 (850K arrays), IC2 (850K arrays) and the public cohort (450K arrays) clusters, the samples did not form consistent clusters by which array type they were analyzed with. This indicates that the main differences between the cohorts came from other batch effects rather than the samples being analyzed on different array platforms, and that following proper normalization of the data, cross-platform MLCs for variant classification can be a useful option.

Furthermore, it was tested whether the classification results would still be valid when the ratio of samples analyzed with different array types in the dataset was unbalanced, as it is a common situation that only a few individuals with a specific RGD are diagnosed in a clinical laboratory. For this purpose, data from one sample with a *KMT2D* variant and one control (generated with 850K arrays) was normalized together with the data from the 76 samples of the public cohort (generated with 450K arrays). The MLC prediction confidence for the classification of the *KMT2D* sample was similar to the confidence when the full testing cohort dataset (IC1: 7 *KMT2D* and 55 control samples) was included (probability scores of 0.54 and 0.55, respectively). The samples were also correctly clustered using MDS analysis (Supplementary Figure 2). This suggests that even with a small number of samples DNAm signatures could be an effective tool for testing the pathogenicity of gene variants when a robust DNAm signature and an adequate size of training dataset are available. This result advocates for the importance of making data publicly available to improve genetic diagnosis. It is though not possible to define the minimum number of samples necessary for MLC training as the individual DNAm patterns of the samples both in the training and testing cohort would affect the ability of the MLC to recognize and classify according to a given DNAm signature.

## 4 | CONCLUSION

In this study, we trained a machine-learning classifier by using a publicly available DNAm dataset and a previously generated disease-specific DNAm signature for variant classification in Kabuki syndrome. The trained classifier could correctly predict the pathogenicity of *KMT2D* variants identified in seven Kabuki syndrome patients. An increasing number of DNAm signatures are being identified<sup>2,3,5</sup> but this method cannot be established in clinical laboratories where only a few individuals are diagnosed with a given RGD. When DNAm data is publicly available, even if different methylation array platforms are employed, it is possible to overcome this hindrance enabling a higher diagnostic yield for a range of rare genetic disorders.

## AUTHOR CONTRIBUTIONS

Conceptualization: Mathis Hildonen, Zeynep Tümer. Methodology: Mathis Hildonen, Andrea Ciolfi, Marco Tartaglia, Zeynep Tümer. Genetic and clinical investigations: Marco Ferilli, Tina Duelund Hjortshøj, Morten Dunø, Lotte Risom, Mads Bak, Jakob Ek, Rikke S. Møller, Andrea Ciolfi, Marco Tartaglia. Data analysis: Mathis Hildonen, Marco Ferilli, Andrea Ciolfi. Project administration: Mathis Hildonen, Zeynep Tümer. Visualization: Mathis Hildonen. Preparation of the original draft: Mathis Hildonen. Finalization of the manuscript: Mathis Hildonen, Andrea Ciolfi, Marco Tartaglia, Zeynep Tümer.

## ACKNOWLEDGMENTS

We acknowledge Darci T. Butcher and colleagues for making the methylation data available for public use. Several of the authors of this publication are members of the European Reference Network on Rare Congenital Malformations and Rare Intellectual Disability ERN-ITHACA (EU Framework Partnership Agreement ID: 3HP-HP-FPA ERN-01-2016/739516). This work was supported, in part, by the Italian Ministry of Health (Ricerca Corrente 2021, to Andrea Ciolfi; 5x1000 2021 and RCR-2020-23670068\_001, to Marco Tartaglia), Italian Ministry of Research (FOE 2019, to Marco Tartaglia), A.P. Møller Fonden (20-L-0371, to Mathis Hildonen), Familien Hede Nielsens Fond (2021-0010, to Mathis Hildonen), and OUH & RH fælles forskningspulje (A4734, to Zeynep Tümer).

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

The raw methylation data from the public cohort and internal cohort 1 used in this study can be accessed at Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>: GSE97362 and GSE218186, respectively). Processed M-values from the 153 CpG sites used in this study can be found in Supplementary Table 5.

## ORCID

Mathis Hildonen  <https://orcid.org/0000-0002-2016-3670>

Marco Ferilli  <https://orcid.org/0000-0002-9883-311X>

Tina Duelund Hjortshøj  <https://orcid.org/0000-0002-3045-8990>

Mads Bak  <https://orcid.org/0000-0003-2762-1002>

Rikke S. Møller  <https://orcid.org/0000-0002-9664-1448>

Andrea Ciolfi  <https://orcid.org/0000-0002-6191-0978>

Marco Tartaglia  <https://orcid.org/0000-0001-7736-9672>

Zeynep Tümer  <https://orcid.org/0000-0002-4777-5802>

## REFERENCES

1. 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;102:156-174. doi:10.1016/j.ajhg.2017.12.008
2. Sadikovic B, Levy MA, Kerkhof J, et al. Clinical epigenomics: genome-wide DNA methylation analysis for the diagnosis of Mendelian disorders. *Genet Med.* 2021;23:1065-1074. doi:10.1038/s41436-020-01096-4

3. Aref-Eshghi E, Kerkhof J, Pedro VP, et al. Evaluation of DNA methylation epigenatures for diagnosis and phenotype correlations in 42 Mendelian neurodevelopmental disorders. *Am J Hum Genet.* 2020; 106:356-370. doi:[10.1016/j.ajhg.2020.01.019](https://doi.org/10.1016/j.ajhg.2020.01.019)
4. Ciolfi A, Aref-Eshghi E, Pizzi S, et al. Frameshift mutations at the C-terminus of HIST1H1E result in a specific DNA hypomethylation signature. *Clin Epigenet.* 2020;12:7. doi:[10.1186/s13148-019-0804-0](https://doi.org/10.1186/s13148-019-0804-0)
5. Levy MA, McConkey H, Kerkhof J, et al. Novel diagnostic DNA methylation epigenatures expand and refine the epigenetic landscapes of Mendelian disorders. *Hum Genet Genomics Adv.* 2022;3:100075. doi:[10.1016/j.xhgg.2021.100075](https://doi.org/10.1016/j.xhgg.2021.100075)
6. Aref-Eshghi E, Bend EG, Colaiacovo S, et al. Diagnostic utility of genome-wide DNA methylation testing in genetically unsolved individuals with suspected hereditary conditions. *Am J Hum Genet.* 2019; 104:685-700. doi:[10.1016/j.ajhg.2019.03.008](https://doi.org/10.1016/j.ajhg.2019.03.008)
7. Schenkel LC, Aref-Eshghi E, Rooney K, et al. DNA methylation epigenature is associated with two molecularly and phenotypically distinct clinical subtypes of Phelan-McDermid syndrome. *Clin Epigenet.* 2021;13:2. doi:[10.1186/s13148-020-00990-7](https://doi.org/10.1186/s13148-020-00990-7)
8. Chater-Diehl E, Goodman SJ, Cytrynbaum C, Turinsky AL, Choufani S, Weksberg R. Anatomy of DNA methylation signatures: emerging insights and applications. *Am J Hum Genet.* 2021;108:1359-1366. doi:[10.1016/j.ajhg.2021.06.015](https://doi.org/10.1016/j.ajhg.2021.06.015)
9. Butcher DT, Cytrynbaum C, Turinsky AL, et al. CHARGE and Kabuki syndromes: gene-specific DNA methylation signatures identify epigenetic mechanisms linking these clinically overlapping conditions. *Am J Hum Genet.* 2017;100:773-788. doi:[10.1016/j.ajhg.2017.04.004](https://doi.org/10.1016/j.ajhg.2017.04.004)
10. Richards S, Aziz N, Bale S, et al. ACMG laboratory quality assurance committee 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. *Genet Med.* 2015;17:405-424. doi:[10.1038/gim.2015.30](https://doi.org/10.1038/gim.2015.30)
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;16: 321-357. doi:[10.1613/jair.953](https://doi.org/10.1613/jair.953)

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Hildonen M, Ferilli M, Hjortshøj TD, et al. DNA methylation signature classification of rare disorders using publicly available methylation data. *Clinical Genetics.* 2023;103(6):688-692. doi:[10.1111/cge.14304](https://doi.org/10.1111/cge.14304)