DNA methylation episignature in Gabriele-de Vries syndrome

Florian Cherik, Jack Reilly, Jennifer Kerkhof, Michael Levy, Haley Mcconkey, Mouna Barat-Houari, Kameryn M Butler, Christine Coubes, Jennifer A Lee, Gwenael Le Guyader, et al.

DNA methylation episignature in Gabriele-de Vries Syndrome

Florian Cherik [1]*, Jack Reilly[2]*, Jennifer Kerkhof[3], Michael Levy[3], Haley McConkey[3], Mouna Barat-Houari[4], Kameryn M. Butler[5], Christine Coubes[6], Jennifer A Lee[5], Gwenael Le Guyader[7], Raymond J. Louie[5], Wesley G. Patterson[5], Matthew L. Tedder[5], Mads Bak[8], Trine Bjørg Hammer[8-9], William Craigen[10], Florence Démurger[11], Christelle Dubourg[12-13], Mélanie Fradin[14], Rachel Franciskovich[10,15], Eirik Frengen[16], Jennifer Friedman[17], Nathalie Ruiz Palares[4], Maria Iascone[18], Doriana Misceo[16], Pauline Monin[19], Sylvie Odent[20], Christophe Philippe[21], Flavien Rouxel[6], Veronica Saletti[22], Petter Strømme[23], Perla Cassayre Thulin[24], Bekim Sadikovic[2-3&], David Genevieve[6&]

1. Service de génétique clinique, Centre de Référence Maladies Rares « Anomalies du Développement et syndromes malformatifs du Sud-Est", CHU de Clermont-Ferrand, Clermont-Ferrand, France.
2. Department of Pathology and Laboratory Medicine, Western University, London, ON N6A 3K7, Canada.
3. Molecular Diagnostics Program and Verspeeten Clinical Genome Centre, London Health Sciences and Saint Joseph's Healthcare, London, ON, Canada.
4. Autoinflammatory and Rare Diseases Unit, Medical Genetic Department for Rare Diseases and Personalized Medicine, Centre Hospitalier Universitaire de Montpellier, 34090 Montpellier, France.
5. Greenwood Genetic Center, JC Self Research Institute of Human Genetics, Greenwood, SC, USA.
6. Département de Génétique Médicale, Maladies Rares et Médecine Personnalisée, Montpellier University, CHU de Montpellier, CLAD ASOOR Montpellier, France.

7. Service de Génétique Clinique, CHU de Poitiers, Poitiers, France.
8. Clinical genetic department, Righospitalet, Copenhagen, Denmark.
9. Epilepsy Genetics and Personalized Medicine, Danish Epilepsy Centre, Dianalund, Denmark.
10. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA.
11. Service de génétique Médicale, Centre hospitalier Bretagne-Atlantique, Vannes, France.
12. Service de Génétique Moléculaire et Génomique, CHU, Rennes, F-35033, France.
13. Univ Rennes, CNRS, IGDR, UMR 6290, Rennes, F-35000, France.
14. Service de Génétique Clinique, Centre Référence Anomalies du Développement CLAD Ouest, Centre Hospitalier Universitaire Rennes, 16 boulevard de Bulgarie, F-35203, Rennes, France.
15. Texas Children's Hospital.
16. Department of Medical Genetics, Oslo University Hospitals and University of Oslo, Oslo, Norway.
17. UCSD Departments of Neurosciences and Pediatrics; Rady Children's Hospital Division of Neurology; Rady Children's Institute for Genomic Medicine.
18. Laboratorio Genetica Medica, ASST Papa Giovanni XXIII, 24127 Bergamo, Italy.
19. Département de génétique médicale, service de génétique clinique, hôpital femme-mère-enfant, Hospices Civils de Lyon.
20. Service de Génétique Clinique, Centre de référence anomalies du développement CLAD-Ouest, CHU Rennes, ERN ITHACA, CNRS UMR 6290 IGDR « Institut de Génétique et développement de Rennes », Université de Rennes, Rennes, France.

21. Functional Unit of Innovative Diagnosis for Rare Diseases, Dijon Bourgogne University Hospital, Dijon, France.
22. Developmental Neurology Unit, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy.
23. Division of Pediatric and Adolescent Medicine, Oslo University Hospital, and University of Oslo, Oslo, Norway.
24. Department of Neurology, University of Utah.

*Participated equally

& Co-corresponding authors

Pr Geneviève David

Département de Génétique Médicale, CHRU Arnaud de Villeneuve

371 avenue du doyen Gaston Giraud, 34000 Montpellier, France

Phone: +33 (0)4 67 33 65 64

Fax: +33 (0)4 67 33 60 52

Email: d-genevieve@chu-montpellier.fr

Dr Bekim Sadikovic

Department of Pathology and Laboratory Medicine, London Health Sciences Centre

800 Commissioners Rd. East, N6A5W9 London, Canada

Phone: +1 (519) 685-8122

Email: bekim.sadikovic@lhsc.on.ca

ABSTRACT

PURPOSE: Gabriele-de Vries syndrome (GADEVS) is a rare genetic disorder characterized by developmental delay, and/or intellectual disability, hypotonia, feeding difficulties and distinct facial features. In order to refine the phenotype and to better understand the molecular basis of the syndrome, we analyzed the clinical data and performed genome-wide DNA methylation analysis of a series of individuals carrying an *YY1* variant.

METHODS: Clinical data were collected for 13 individuals not yet reported through an international call for collaboration. DNA was collected for 11 of these individuals and 2 individuals previously reported in an attempt to delineate a specific DNA methylation signature in GADEVS.

RESULTS: Phenotype in most individuals overlapped with the previously described features. We also describe one individual with atypical phenotype, heterozygous for a missense variant in a domain usually not involved in individuals with *YY1* pathogenic missense variations. We described a specific peripheral blood DNA methylation profile associated with *YY1* variants.

CONCLUSION: We report a distinct DNA methylation episignature in GADEVS. We expand the clinical profile of GADEVS to include thin/sparse hair and cryptorchidism. We highlight the utility of DNA methylation episignature analysis for classification of variants of unknown clinical significance.

KEYWORDS: Gabriele-de Vries syndrome, YY1, epigenetics, episignature, DNA methylation, intellectual disability, feeding disorders

INTRODUCTION

Alteration of proteins involved in chromatin regulation is a well-established cause of many neurodevelopmental disorders. Among these conditions, Gabriele-de Vries syndrome (GADEVS, MIM 617557) is a rare congenital disorder characterized by variable intellectual disability (ID), various neurological disorders (hypotonia, abnormal movements, behavioral disorders, brain abnormalities), feeding difficulties, ophthalmological abnormalities, significant but not specific facial features, and more rarely cardiac or renal malformations[1–5]. GADEVS is mainly caused by pathogenic missense variants in Yin Yang 1 Transcription Factor gene (*YY1*, MIM 600013) and less frequently by truncating variants or whole gene deletions, suggesting haploinsufficiency as the underlying mechanism[1].

YY1 encodes the Yin Yang 1 Transcription factor, which is a ubiquitously expressed transcription factor in mammals. Its name comes from its ability to be both an activator and a repressor of transcription[6]. YY1 is characterized by four highly conserved C2H2 Zinc fingers located in its C-terminal domain. The N-terminal region corresponds to the transcriptional activation domain. A transcriptional repression domain, including the REPO domain allowing the recruitment of the polycomb complex, is located between N-terminal region and Zinc fingers domain[7–9].

It has been demonstrated that genetic disorders involving genes related to chromatin regulatory functions exhibit specific DNA methylation signature, referred to as an episignature[10–12]. DNA methylation episignature analysis has recently been implemented as the diagnostic clinical genomic DNA methylation test EpiSign, in patients with rare disorders, providing strong evidence for clinical utility including the ability to provide conclusive diagnostic findings in the majority of subjects tested[13]. In this study, we describe the clinical phenotype of 13 previously unpublished individuals heterozygous for a pathogenic or likely

pathogenic variant or a complete deletion of *YY1*, as well as a specific epigenetic signature associated with GADEVS.

MATERIALS AND METHODS

SUBJECTS AND SERIES

We contacted clinicians about 19 individuals heterozygous for a pathogenic or likely pathogenic variant or a deletion of *YY1* through clinical networks (Groupe DI France, AnDDI-RARES (http://anddi-rares.org/), ERN ITACHA (https://ern-ithaca.eu/) and GeneMatcher (http://www.genematcher.org)[14]. We collected clinical and molecular data, DNA samples, brain MRI and neuropsychological assessment data of individuals from this series, when available. Referring physicians provided the data by filling in a standardized table.

This study was approved by the Institutional Review Board of Montpellier University Hospital (IRBMTP_2020_05_202000459, ClinicalTrial.gov identifier: NCT04381715) and the Western University Research Ethics Board (REB 106302). We obtained informed written consent from all individuals or their legal guardians to participate in the study and to publish their photographs. All samples and records were de-identified. The research was conducted in accordance with the Declaration of Helsinki.

MOLECULAR STUDIES

Diagnostic laboratories performed genetic tests on DNA from blood samples using next-generation sequencing or chromosomal microarrays. The pathogenicity of point variants was verified according to American College of Medical Genetics and Genomics (ACMG) and Association for molecular pathology (AMP) classification criteria[15], using the varsome interface (https://varsome.com/)[16]. The visualization of the variants on the protein sequence was performed with the ProteinPaint tool (https://proteinpaint.stjude.org/), using the canonical isoform NM_003403.

STATISTICAL ANALYSIS

To describe the continuous variables (SD of growth parameters, age of milestones acquisition), we calculated medians, minimums, maximums and interquartile ranges in order to construct corresponding boxplots. We also included data from the literature in these graphs.

METHYLATION ARRAY AND QUALITY CONTROL

DNA methylation analysis and episignature classifier development was performed using a previously established protocol[11,12,17,18]. Genomic DNA was extracted from peripheral blood samples using standard techniques and followed by bisulfite conversion and hybridization to the Illumina infinium methylation EPIC bead chip arrays, according to manufacturer's protocol. Idat files, containing methylated and unmethylated signal intensity plots (beta values) were produced from these microarrays, and used for analysis in R 4.0.2. Normalization was performed using the Illumina Infinium methylation EPIC array with background correction from the minfi package[19]. Previously defined exclusion criteria[12,17] were used to exclude probes with detection p values >0.01, probes on the X and Y chromosomes, probes known to contain SNPs at the site of CpG interrogation or single nucleotide extension, and probes known to cross react with chromosomal locations other than their target regions. All samples were examined for genome-wide methylation distribution and those deviating from a bimodal distribution were excluded. Factor analysis using a principal component analysis (PCA) was performed to examine batch effect and identify outliers.

DNA METHYLATION PROFILING

Probe methylation levels (beta values), were calculated as the ratio of signal intensity in methylated probes versus total sum of unmethylated and methylated probes, resulting in

values ranging from zero to one. To allow for linear regression modeling, beta values were logit transformed using the limma package[20], allowing for the identification of differentially methylated probes. Data were adjusted for the blood cell type composition as per Houseman et al[21]. Estimated blood cell proportion was added to the model matrix of the linear models as a confounding variable[22]. Using the eBayes function in the limma package[23], p values were moderated and corrected for multiple testing using the Benjamini Hochberg method. Probes with the most significant methylation differences were selected using two items from this dataset: the level of methylation difference (relative methylation signal intensity), and the probability that an observed difference is due to random chance (p values). Evaluation of this interaction was carried out by multiplying the absolute methylation difference between affected cases and controls by the negative value of the log transformed p values, and ranking the top 1000 probes with the highest values from this transformation. Next, receiver operating characteristic analysis (ROC) was performed on each probe, to measure the pairwise correlation coefficient between probes. Probes with low area under curve values from ROC analysis were removed, as well as highly correlated probes, eliminating probes with low sensitivity and specificity, and probes with highly correlated characteristics using Pearson's correlation coefficient. This ensures that the final probeset contains the most differentiating, non-redundant probes that are not influenced by random data structures. Only probes with a methylation difference greater than 5% were included in this analysis. This probe filtering process was designed to avoid reporting of probes with low effect size, and those influenced by technical or random variations as conducted in previous studies[11,12].

SELECTION OF MATCHED CONTROLS FOR METHYLATION PROFILING

For episignature characterization, mapping of probes and feature selection, matched controls were randomly selected from the LHSC EpiSign Knowledge Database (EKD)[12]. All of the samples were assayed, therefore all the controls selected for episignature identification were analyzed using the same array type. Samples were matched by age, sex and batch using the MatchIt package. A 4:1 ratio of controls to cases was deemed optimal for this analysis, as previously described[11]. PCA analysis was performed after each attempt at matching to detect outliers and determine data structures for the presence of batch effect. Outlier samples, and those with highly aberrant data structures were removed, and subsequent matching trials were performed until consistent iterations with no outliers in the first two components of the PCA were derived. No such samples were identified for removal in this cohort.

CLUSTERING AND DIMENSION REDUCTION

Hierarchical clustering and multidimensional scaling were used after each iteration of analysis to examine the data structure of the identified episignature. Hierarchical clustering was performed using Ward's method on Euclidean distance by the ggplot package[24,25]. Multi-dimensional scaling provides a visual representation of sample methylation profile similarity based on the scaling of the pairwise Euclidean distances between each sample.

DISCOVERY/TRAINING COHORT SELECTION

Identification of disease-specific episignatures was performed using a randomly selected sub-set of the database, on a 75:25 ratio of discovery:training, using the caTools package in R. Testing samples were used to assess the performance of the classification model developed later in the study. For every disease group in the discovery cohort, a sex and age-matched control group with a sample size at least four times larger was selected from the

reference control group using the MatchIT package, and methylation profiles were compared between the two.

CROSS VALIDATION

For each round of validation, one of the 13 selected samples was removed from probe selection, alongside matched controls. The remaining samples were designated as testing samples, and all three groups were modeled using multidimensional scaling to determine how they cluster/segregate with one another. This process was repeated with different combinations of assigned training and testing samples until all cases had been removed from probe selection and used for testing once (see Figure S2).

CLASSIFICATION MODEL

Specificity of the episignature was assessed using the Methylation Variant Pathogenicity (MVP) score, using all the identified probes. A support vector machine (SVM) used a linear kernel for training on GADEVS cases and controls. Once again, a 4:1 ratio of controls to cases was used to divide both the case and control samples previously matched and used for probe selection into training and testing cohorts for the SVM. Furthermore, the remaining unselected samples from the EKD were also divided similarly (75% training, 25% testing) to allow for comparison and testing of signature robustness against all of the samples in the EKD. Using the e1071 R package, we performed 10-fold cross validation to determine hyperparameters optimal for episignature classification. In this process, the training set was divided into ten folds by random assignment, where the first nine are used for training, and the last used for testing the accuracy of the model. The mean accuracy over all rounds was then calculated, and hyperparameters with the best performance by this metric were selected. The model provides a score ranging from 0-1 for each subject, representing the model's confidence in predicting whether the subject has a DNA methylation profile similar

to the GADEVS probe set or not. Conversion of these SVM decision values was done using Platt's scaling method[26], and the class obtaining the greatest score determined the predicted phenotype. A classification as GADEVS was made when a sample received the greatest score for that class (normally greater than 0.5). Finally, the model was applied to both a training set of a large cohort of individuals with clinical and molecular diagnoses of neurodevelopmental disorders, as well as a group of healthy controls to determine its effective specificity.

VALIDATION OF CLASSIFICATION

To ensure the model is not susceptible to the batch structure of the methylation experiment, the classifier was applied to samples assayed on the same batch as the cases used for training. Using methylation data from individuals without a confirmed diagnosis of GADEVS within the EKD and assayed on the same microarray chip as case samples, methylation profiles were modeled to ensure the classifier was not confounded by technical artifacts unique to the given microarray. Specificity was determined by supplying a large number of DNA methylation arrays from unaffected subjects to the model. To further assess the specificity of the GADEVS classifier relative to other neurodevelopmental disorders, we applied it to cases with other patient cohorts exhibiting distinct episignatures within the EKD.

RESULTS

CONSTITUTION OF THE SERIES

We contacted the referring clinicians of 19 individuals heterozygous for a pathogenic or likely pathogenic variant in *YY1*. We excluded three individuals either because they refused to participate in the study or because neither clinical data nor a DNA sample was available.

Another individual was excluded because the *YY1* variant was inherited from a healthy parent. We therefore included 15 individuals in this study. For details see Table S1.

Among these 15 individuals, 13 were not previously reported; these 13 individuals were labeled YY1-1 to YY1-13 and constituted the clinical series that allowed us to refine the phenotypic data related to *YY1* pathogenic variants. The two remaining individuals, "individual 5" and "individual 8," were initially reported by Gabriele et al, 2017[1]. Regarding the episignature series, we used DNA samples from 11 individuals of the clinical series (samples from individuals YY1-8 and YY1-9 were not available) along with DNA samples from "individual 5" and "individual 8." The episignature series is detailed in Table S2.

CLINICAL SPECTRUM ASSOCIATED WITH YY1 PATHOGENIC VARIANTS

Clinical data were collected for the 13 individuals (YY1-1 to YY1-13) not previously reported. Detailed clinical data are available in the Table S3.

Among this series, 12 individuals had a phenotype overlapping with that previously described in the literature. Unfortunately, individual YY1-6 (father of individual YY1-7) died accidentally before being clinically assessed. The only data concerning individual YY1-6 is the presence of ID. Because a clearly unusual phenotype and genotype, individual YY1-10 could not be considered to have GADEVS.We therefore chose to describe him separately.

Table 1 summarize the clinical data from this series and the literature.

The 12 individuals with phenotype overlapping with the literature presented with variable ID and/or developmental delay. All these individuals presented with craniofacial features among which the most frequent were long face, broad forehead, simple ears, malar hypoplasia and full nasal tip. They also frequently had thin and/or sparse hair. (figure 1A).

We also observed various neurological disorders such as hypotonia, behavioral disorders (ASD, low frustration tolerance, anxiety, self-harm, ADHD), and abnormal movement

(dystonia). Feeding disorders were present in 10/10 individuals. Frequent additional features include skeletal abnormalities, ophthalmologic abnormalities, and cryptorchidism.

Overall distribution (including data from literature) of ages of growth parameters and milestones achievement is represented in figure 1 (respectively B and C).

Individual YY1-10 was considered to have an unusual *YY1* phenotype because of overgrowth and obesity (BMI=41kg/m²), slight macrocephaly (HC at 53cm [+2.3SD]) and moderate craniofacial features (See figure 1A).

Full clinical features of individual YY1-10 is detailed in Table S3.

YY1 VARIANTS SPECTRUM

We collected molecular data from 13 unpublished individuals including a father-son pair (individuals YY1-6 and YY1-7). Except for this pair, all variants were *de novo*. The variants of the series and from the literature are represented on the YY1 protein sequence in Figure S1. Among these 13 individuals, 12 carried a pathogenic or likely pathogenic sequencing variant (10 were missense and two were truncating variants). All missense variants were located in Zinc finger domain except for individuals YY1-10. The variant p.(Gly176Asp) from individual YY1-10 was located in the transcriptional repression domain. Missense variants located in this domain has never been previously described in the literature to our knowledge. The last individual YY1-3 had a microdeletion encompassing *YY1*, *WARS1* and the 3' end of *EML1*.

DETECTION AND VERIFICATION OF AN EPISIGNATURE FOR YY1/GADEVS

DNA methylation profiles from 13 individuals peripheral blood samples, which all had confirmed molecular variants in the *YY1* gene, were used to establish a DNA methylation episignature for this disorder. Overall methylation patterns in all 13 patients were assessed for several key features, including sample quality, and similarity of the sample methylation profiles to case samples versus controls. Of these, one sample, YY1-10 segregated

consistently with controls, exhibiting methylation patterns more similar to age and sex matched control samples than the rest of the disorder cohort, and was removed from probe selection. Comparisons were carried out, matching GADEVS samples with age, sex and batch-matched controls at a ratio of 4:1 (4 matched controls for each case sample). When compared to controls, significant differences in methylation patterns across 487 probes, which are visualized using a volcano plot (Figure S2) were detected. Selected probes had a minimum methylation difference of 10%, and a multiple testing corrected p value of <0.01 (limma multivariable regression modeling).

VISUALIZATION OF METHYLATION PROFILES INDICATES DISTINCT CLUSTERING PATTERNS OF YY1 CASES

Hierarchical clustering was used to visualize methylation differences based on the selected probes, and was plotted using Ward's method alongside 56 age and sex-matched control samples (see figure 2A). This model demonstrated a clear separation of the control and case samples, with the exception of the YY1-10 sample. This sample grouped with control samples in all iterations of the model, indicating that the associated variant in this sample results in a methylation profile more similar to control samples than the other cases with confirmed *YY1* variants. The location and characteristics of the variant are atypical : this is a missense variant within the transcriptional repression domain of YY1, and reported presentation of overgrowth characteristics. Multiple dimensional scaling (MDS) showed similar findings, with cases grouping tightly together away from the control cohorts (see figure 2B). Cross validation using GADEVS samples was performed, showing in the majority of cases that the remaining testing samples clustered with the other GADEVS samples, and segregated from the controls (see figure 2D). In three cases, samples YY1-6, YY1-7 and YY1-11, cross validation showed less specific clustering along with lowered MVP scores,

suggesting a level of signal heterogeneity and further data structure within the observed common episignature. However, all samples consistently segregated with the case cohort in hierarchical clustering and multidimensional scaling plots, and received high MVP scores when provided to the finalized SVM classifier (see Figure 2C and Figure S2).

MVP SCORE DEMONSTRATES SENSITIVITY AND SPECIFICITY OF GADEVS EPISIGNATURE

Samples were provided to a support vector machine binary classifier with a linear kernel to assess the sensitivity and specificity, and the ability of the selected probe set to classify samples. For each sample, the classifier provides a methylation variant pathogenicity (MVP) score between 0 and 1. When plotted against control samples, all GADEVS samples received high scores (>0.8) close to 1, while the control scores remained near 0, indicating the classifier has a high sensitivity for the detection of the GADEVS episignature (see Figure 2C). Furthermore, specificity of the classifier was tested by providing it with a large number of subjects with a confirmed diagnosis of a neurodevelopmental disorder of various types with existing episignatures within the EKD. 75% of both case and control samples from other syndromes in the EKD were used for training, with the additional 25% reserved for testing. Case samples scored >0.8, while the remaining non-GADEVS cases scored very low, with no case exceeding a score of 0.5 to be classified as a GADEVS sample, indicating a very high level of specificity for the selected probe set.

DISCUSSION

We describe the phenotype of 12 new individuals heterozygous for a pathogenic or likely pathogenic variant in *YY1*, proposed to lead to YY1 loss-of-function as reported by Gabriele et al, 2017[1]. In addition, missense variants in the zinc finger domains and truncating variants both lead to an overall decrease in the occupancy of YY1 on the genome and a loss of H3K27 acetylation at the active enhancers linked by YY1, and consequently to a differential

expression of target genes[1]. It was therefore postulated that YY1 could have an impact on DNA methylation especially since YY1 has been demonstrated to have the ability to recruit the Polycomb complex[27,28] known to be involved in the control of DNA methylation[29].

We observed a similar phenotype in our series to that described in the literature, such as variable ID and developmental delay, behavioral and abnormal movement disorders, skeletal abnormalities, and ophthalmological abnormalities, associated with craniofacial features and feeding difficulties with a consequent low BMI in individuals with the classical variants.

We also observed some differences including additional clinical features not previously described in the literature. These included thin and/or sparse hair (6/10). Looking at pictures from the literature it seems that collectively, 15/22 (68%) of *YY1* individuals had this clinical feature.

In addition, congenital malformations and cardiac malformations seem less frequent in this series. We also observed cryptorchidism in 3/7 males, whereas this feature has been described only once in the literature. However, *YY1*-related disorders are very rare, and it is difficult to make conclusions on such a small number of individuals. Indeed, despite an international call for recruitment we could only identify 13 new individuals with *YY1* pathogenic or likely pathogenic variants according to ACMG/AMP classification criteria. This condition is probably still underdiagnosed since the involvement of *YY1* in neurodevelopmental disorders only became recognized in 2017.

As the severity of ID seems to be variable in GADEVS, we wanted to study the neurocognitive profile of individuals carrying a *YY1* variant in order to highlight a possible specific pattern. However, data from neuropsychological assessments were largely insufficient, because the data were incomplete or uninterpretable. Additional studies should be performed to this point.

In addition to the clinical features, here we demonstrate the first evidence of a peripheral blood DNA methylation episignature, as a common molecular phenotype in patients presenting with classical features of GADEVS. All samples provided evidence of a common methylation profile for GADEVs, with of limited signal heterogeneity within the cross-validation model for 3 samples (YY1-6, YY1-7, and YY1-11) which received more moderate scores compared to the rest of the cohort. These findings, alongside the atypical sample (YY1-10) indicate the possibility of additional data structure, or sub-signatures, associated with variants in the YY1 sequence, similar to what is observed in some other genetic conditions[17,18]. Further research with larger sample size will be necessary to study this hypothesis.

Individual YY1-3 carrying a deletion encompassing *YY1* plus two other genes (*WARS1* [MIM 191050] and *EML1* [MIM 602033]) has a similar epigenetic signature to that observed in individuals with pathogenic missense variants, suggesting that his phenotype can be at least partially attributed to *YY1* haploinsufficiency. In addition to dystonia previously described by Gabriele et al, 2017, Carminho-Rodrigues et al, 2020[3] and Zorzi et al, 2021[4], individual YY1-3 also has severe spasticity, as well as short stature (-3.9 SD). One of the other two genes, *WARS1*, could explain the additional neurological feature as this gene is associated with a dominant distal motor neuropathy phenotype[30–32].

Regarding individual with atypical localization of missense variant (individual YY1-10 with p.(Gly176Asp), located in in the transcriptional repression domain), we observed some major differences in phenotype than GADEVS, i.e. overgrowth, obesity and macrocephaly. Moreover, his DNA methylation profile is not specific and does not fit with the GADEVS episignature. The p.(Gly176Asp) variant was initially considered as likely pathogenic according to the ACMG/AMP classification criteria (*de novo* variant absent from gnomAD

exomes and genomes) but the result of the DNA methylation analysis has allowed us to reclassify this variant to unknown significance related to GADEVS. However, whether this variant is likely benign is not certain, given the possibility of yet to be defined alternate episignatures or lack thereof. The utility of EpiSign analysis in the reclassification of variants of uncertain clinical significance has been recently demonstrated in the clinical setting in a large number of Mendelian disorders with established episignatures[13]. Several studies have been published from our lab thus far involving additional substratification of episignatures [17,18] further highlighting the importance of methylation profiling in elucidating complex presentations of phenotype that remain unexplained by genetic diagnosis alone.

Considering the phenotype of overgrowth in individual YY1-10, the pathophysiological mechanism could be the selective alteration of the transcriptional repression function. However functional analysis or additional individuals with the same p.(Gly176Asp) *YY1* variant should be necessary to definitively ruled out or to confirm this variant to be responsible for a novel YY1-related disorder.

In conclusion, we describe 12 novel individuals with Gabriele-de Vries syndrome. We identified novel features (i.e., thin and/or sparse hair and cryptorchidism in males). We also describe for the first time a highly sensitive and specific DNA methylation episignature for GADEVS and demonstrate the utility of EpiSign in the clinical assessment of variants of uncertain clinical significance. Additional research is necessary to support the expanded clinical spectrum and genotype-phenotype correlations in GADEVS.

DATA AVAILABILITY

The summarized, anonymized data for each subject are described in the study. The raw anonymized DNA methylation data are available from the authors upon request. Software used in this study is publicly available (https://topepo.github.io/caret/, https://bioconductor.org/packages/release/bioc/html/limma.html https://bioconductor.org/packages/release/bioc/html/minfi.html, https://ggplot2.tidyverse.org/, https://www.rdocumentation.org/packages/MatchIt/versions/4.3.2, https://cran.r-project.org/web/packages/caTools/index.html, https://cran.r-project.org/web/packages/e1071/index.html) and detailed analytical methodology is as previously reported[10].

ACKNOWLEDGEMENTS

Funding for this study was provided, in part, by the London Health Sciences Molecular Diagnostics Development Fund and Genome Canada Genomic Applications Partnership Program Grant (Beyond Genomics: Assessing the Improvement in Diagnosis of Rare Diseases using Clinical Epigenomics in Canada, EpiSign-CAN) awarded to B.S.

We thank the families for their participation in this study.

AUTHOR INFORMATION

Conceptualization: FC, JR, PM, DG, BS; Data curation: FC, JR, ML, DG; Formal analysis: FC, JR, ML, JK, MB, MBH, KMB, JAL, RJL, MLT, CD, NRP, MI, CP, EW, DJ; Investigation: FC, JR, JK, KB, CC, GLD, WGP, TBH, FD, WC, MF, RF, EF, JF, HG, DM, PM, SO, FR, VS, PS, PT, DG, BS; Methodology: DG, BS; Project administration: DG, BS; Supervision: DG, BS; Validation: FC, JR, DG; Visualization: FC, JR; Writing – original draft: FC, JR, DG, BS; Writing – review & editing: DG, BS

ETHICS DECLARATION

This study was approved by the Institutional Review Board of Montpellier University Hospital (IRBMTP_2020_05_202000459, ClinicalTrial.gov identifier: NCT04381715) and the Western University Research Ethics Board (REB 106302). We obtained informed written consent from all individuals or their legal guardians to participate in the study and to publish their photographs. All samples and records were de-identified. The research was conducted in accordance with the Declaration of Helsinki.

REFERENCES

1. Gabriele M, Vulto-van Silfhout AT, Germain PL, et al. YY1 Haploinsufficiency Causes an Intellectual Disability Syndrome Featuring Transcriptional and Chromatin Dysfunction. *The American Journal of Human Genetics*. 2017;100(6):907-925. doi:10.1016/j.ajhg.2017.05.006

2. Morales-Rosado JA, Kaiwar C, Smith BE, Klee EW, Dhamija R. A case of *YY1* -associated syndromic learning disability or Gabriele-de Vries syndrome with myasthenia gravis. *Am J Med Genet Part A*. Published online December 14, 2018:ajmg.a.40626. doi:10.1002/ajmg.a.40626

3. Carminho-Rodrigues MT, Steel D, Sousa SB, et al. Complex movement disorder in a patient with heterozygous YY1 mutation (Gabriele-de Vries syndrome). *Am J Med Genet*. 2020;182(9):2129-2132. doi:10.1002/ajmg.a.61731

4. Zorzi G, Keller Sarmiento IJ, Danti FR, et al. *YY1* -Related Dystonia: Clinical Aspects and Long-Term Response to Deep Brain Stimulation. *Mov Disord*. Published online February 27, 2021:mds.28547. doi:10.1002/mds.28547

5. Tan L, Li Y, Liu F, et al. A 9-month-old Chinese patient with Gabriele-de Vries syndrome due to novel germline mutation in the *YY1* gene. *Mol Genet Genomic Med*. Published online December 23, 2020. doi:10.1002/mgg3.1582

6. Shi Y, Seto E, Chang LS. Transcriptional Repression by YY1, a Human GLI-Kr ippel-Related Protein, and Relief of Repression by Adenovirus EIA Protein. :12.

7. Sarvagalla S, Kolapalli SP, Vallabhapurapu S. The Two Sides of YY1 in Cancer: A Friend and a Foe. *Front Oncol*. 2019;9:1230. doi:10.3389/fonc.2019.01230

8. He Y, Casaccia-Bonnefil P. The Yin and Yang of YY1 in the nervous system. *Journal of Neurochemistry*. 2008;106(4):1493-1502. doi:10.1111/j.1471-4159.2008.05486.x

9. Atchison ML, Basu A, Zaprazna K, Papasani M. Mechanisms of Yin Yang 1 in Oncogenesis: The Importance of Indirect Effects. *Crit Rev Oncog*. 2011;16(3-4):143-161. doi:10.1615/CritRevOncog.v16.i3-4.20

10. Aref-Eshghi E, Kerkhof J, Pedro VP, et al. Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in 42 Mendelian Neurodevelopmental Disorders. *The American Journal of Human Genetics*. 2020;106(3):356-370. doi:10.1016/j.ajhg.2020.01.019

11. 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(4):685-700. doi:10.1016/j.ajhg.2019.03.008

12. 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):156-174. doi:10.1016/j.ajhg.2017.12.008

13. 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

14. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. *Human Mutation*. 2015;36(10):928-930. doi:10.1002/humu.22844

15. Richards S, Aziz N, Bale S, 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. *Genet Med*. 2015;17(5):405-423. doi:10.1038/gim.2015.30

16. Kopanos C, Tsiolkas V, Kouris A, et al. VarSome: the human genomic variant search engine. Wren J, ed. *Bioinformatics*. 2019;35(11):1978-1980. doi:10.1093/bioinformatics/bty897

17. 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):64. doi:10.1186/s13148-019-0658-5

18. Aref-Eshghi E, Bend EG, Hood RL, et al. BAFopathies' DNA methylation epi-signatures demonstrate diagnostic utility and functional continuum of Coffin–Siris and Nicolaides–Baraitser syndromes. *Nat Commun*. 2018;9(1):4885. doi:10.1038/s41467-018-07193-y

19. 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):1363-1369. doi:10.1093/bioinformatics/btu049

20. 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. doi:10.1093/nar/gkv007

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

22. Reinius LE, Acevedo N, Joerink M, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One*. 2012;7(7):e41361. doi:10.1371/journal.pone.0041361

23. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article3. doi:10.2202/1544-6115.1027

24. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. Second edition. Springer; 2016.

25. Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*. 1963;58(301):236-244. doi:10.1080/01621459.1963.10500845

26. Smola AJ, Bartlett PJ. *Advances in Large Margin Classifiers*. MIT Press; 2000.

27. Atchison L. Transcription factor YY1 functions as a PcG protein in vivo. *The EMBO Journal*. 2003;22(6):1347-1358. doi:10.1093/emboj/cdg124

28. Wilkinson FH, Park K, Atchison ML. Polycomb recruitment to DNA in vivo by the YY1 REPO domain. *Proceedings of the National Academy of Sciences*. 2006;103(51):19296-19301. doi:10.1073/pnas.0603564103

29. Viré E, Brenner C, Deplus R, et al. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature*. 2006;439(7078):871-874. doi:10.1038/nature04431

30. Li JQ, Dong HL, Chen CX, Wu ZY. A novel WARS mutation causes distal hereditary motor neuropathy in a Chinese family. *Brain*. 2019;142(9):e49. doi:10.1093/brain/awz218

31. Tsai PC, Soong BW, Mademan I, et al. A recurrent WARS mutation is a novel cause of autosomal dominant distal hereditary motor neuropathy. *Brain*. 2017;140(5):1252-1266. doi:10.1093/brain/awx058

32. Wang B, Li X, Huang S, et al. A novel WARS mutation (p.Asp314Gly) identified in a Chinese distal hereditary motor neuropathy family. *Clin Genet*. 2019;96(2):176-182. doi:10.1111/cge.13563

FIGURE AND TABLE LEGENDS

Figure 1: Representation of some clinical features related to *YY1*. A: Front and lateral view of individuals from this series. Common facial features are long face, broad forehead, simple ears, malar hypoplasia, full nasal tip and sparse hair. B: Boxplots showing distribution of ages at sitting alone, walking alone and first words in standard deviation. C: Boxplots showing distribution of height, head circumference (HC) and BMI, in standard deviation.

Figure 2: GADEVS has a specific DNA methylation episignature. A. DNA methylation signal intensity plot for 13 patients with identified *YY1* variants sorted by hierarchical clustering. Cases in red represent GADEVS cases, those in blue indicate cases with no phenotypic or genotypic presentation of GADEVS, including samples with confirmed presentation of other syndromes, and the final case in orange, refers to sample YY1-10, which was removed from probe selection following segregation with control samples.

B. Multidimensional scaling plot representing the dimensions of variation in methylation signal intensity at informative CpG identified for GADEVS. Represents similarity of methylation profiles of GADEVS patients, marked in red.

C. SVM classifier model for GADEVS. Each sample receives scores for the probability of having a DNA methylation profile similar to cases as compared to samples with a confirmed Episignature in the EKD. Higher value on Y-axis indicates that a sample presents a methylation profile more similar to cases compared to the methylation profiles of patients with other disorders. Thirty-six other syndromes with confirmed Episignatures from the EKD are plotted based on this relative scale of similarity to indicate probeset specificity for the case disorder.

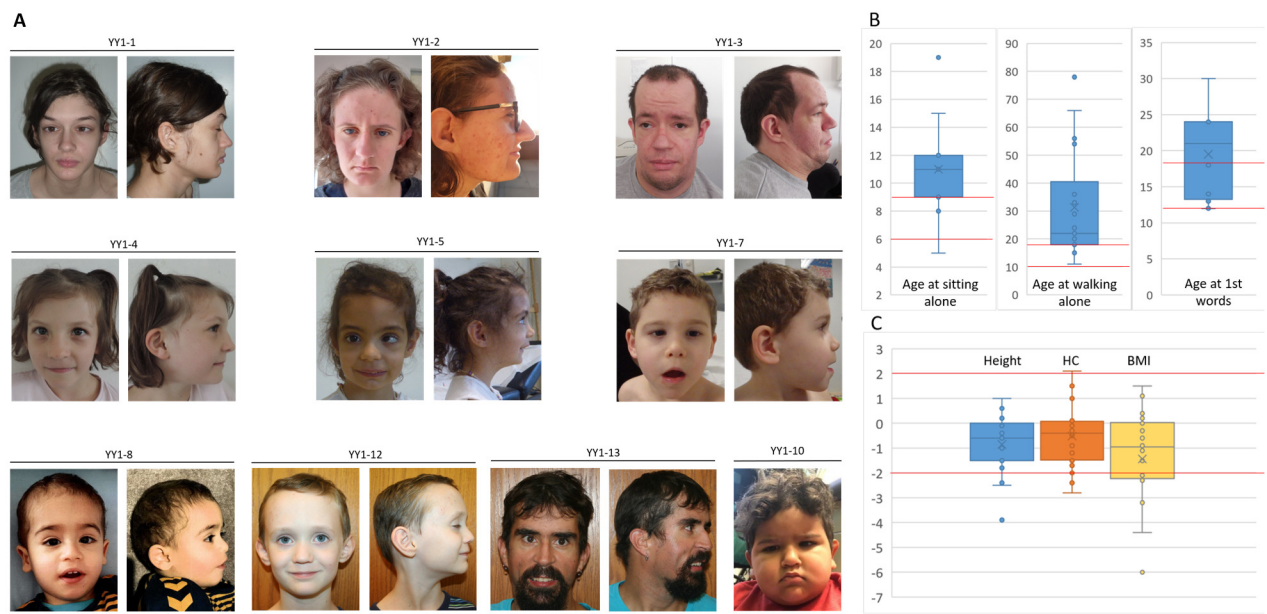
D. Cross Validation summary representing the MVP scores received for each sample during their respective testing round. Case samples are marked in red, while the remaining samples

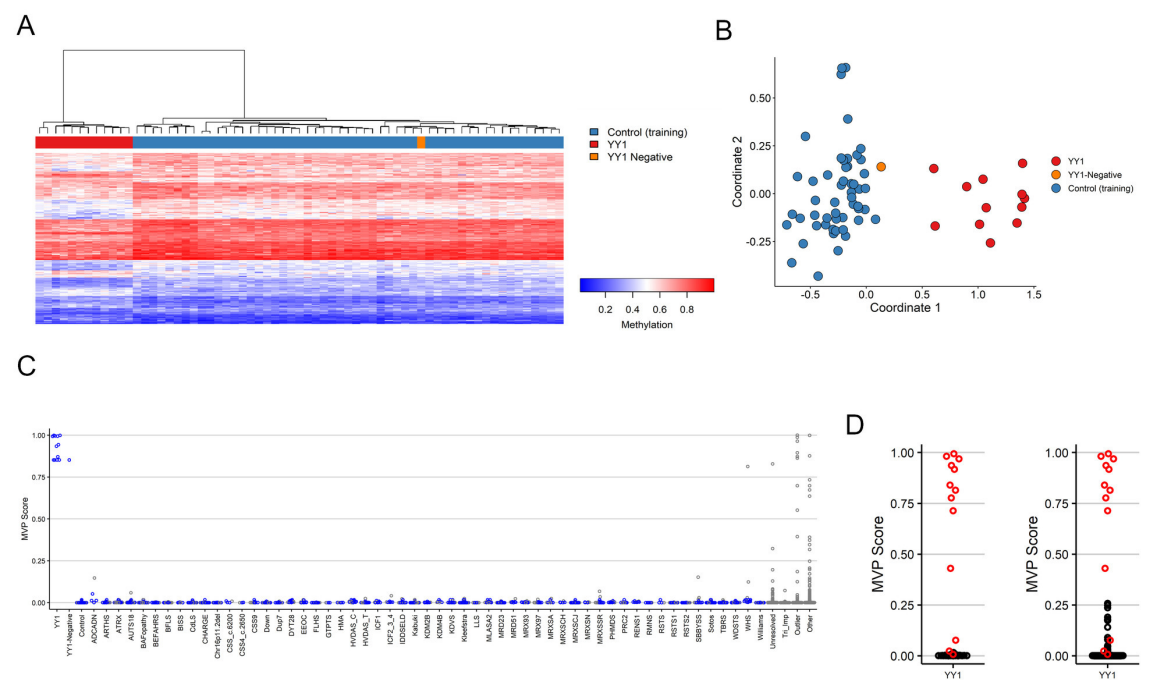
from the EKD are marked in black. Left side plot contains MVP scores for the EKD samples following training the SVM against controls, while the right contains MCP scores for EKD samples following training the SVM against all samples within the EKD.

Table 1: Summary of clinical features of individuals carrying a pathogenic variant of *YY1* (this series and the literature[1–5]). Individual YY1-6 is not included in this table due to lack of clinical data. Individual YY1-10 is described separately because of its atypical variant and phenotype. +: feature present; -: feature absent; NR: not reported; NK: not known. a: dystonia, dyskinesia; b: camptodactyly, joint hyperlaxity, scoliosis, plagiocephaly, turricephaly; c: Hyperopia, superficial punctatae keratitis, nystagmus, strabismus (5/12), astigmatism, myopia, cortical vision abnormalities. d: frequencies based on our own interpretation of the pictures available in literature.

CONFLICT OF INTEREST NOTIFICATION PAGE

The authors declare no conflict of interest


A
YY1-1
YY1-2
YY1-3
YY1-4
YY1-5
YY1-7
YY1-8
YY1-12
YY1-13
YY1-10
B
Age at sitting alone
Age at walking alone
Age at 1st words
C
Height
HC
BMI


A
Control (training)
YY1
YY1 Negative
0.2
0.4
0.6
0.8
Methylation
B
Coordinate 2
0.50
0.25
0.00
-0.25
-0.5
0.0
0.5
1.0
1.5
Coordinate 1
YY1
YY1-Negative
Control (training)
C
MVP Score
1.00
0.75
0.50
0.25
0.00
YY1
YY1-Negative
Control
ADCADN
ARTHS
ATRX
AUTS1B
BAFopathy
BEFAHRS
BRS
CdLS
CHARGE
Chr16p11.2del
CSS_c.6200
CSS4_c.2660+
CSS9
Down
DupYY1
DYT28
EEOC
FLHS
GTPT5
HMA
HVDAS_C
HVDAS_T
ICF
ICF2_3_4
IDDSELD
Kabuki
KDM3B
KDM4B
KDV5
Koolstra
LIS
MLASA2
MRD23
MRD51
MRX93
MRX97
MRX9A
MRX9CH
MRX9CJ
MRX9DS
MRX9R
PHM25
PRC2
RENS1
RNNS
RSTS
RSTS1
RTS2
SBBYSS
Sore
TBRS
WGSTS
WHS
Williams
Unresolved
Th_Het
Other
D
MVP Score
1.00
0.75
0.50
0.25
0.00
YY1
MVP Score
1.00
0.75
0.50
0.25
0.00
YY1

	Classical YY1 phenotype			Atypical phenotype
	Present study (n=11)	Literature (n=14)	Total (n=26)	YY1-10
Growth				
IUGR	1/9	4/13	5/22 (23%)	-
Short stature	2/11	2/14	4/25 (16%)	Overgrowth
BMI < -2SD	4/10	3/10	7/20 (35%)	Obesity
Microcephaly	2/10	1/12	3/22 (14%)	Macrocephaly
Development				
Motor delay	8/11	11/14	19/25 (76%)	+
Language delay	10/11	10/12	20/23 (87%)	+
ID	10/11	11/12	21/23 (91%)	+
Neurological features				
Hypotonia	5/11	5/13	10/24 (42%)	+
Behavioral disorders	10/11	7/12	17/23 (74%)	+
Abnormal movement	4/11[a]	7/12	11/23 (48%)	-
Abnormal brain MRI	4/8	8/13	12/21 (57%)	+
Miscellaneous				
Cardiac abnormalities	1/9	4/11	5/20 (25%)	-
Cryptorchidism	3/7	1/5	4/12 (33%)	-
Skeletal abnormalities	9/10[b]	8/13	17/23 (74%)	-
Feeding disorders	10/10	12/13	22/23 (96%)	-
Constipation	4/11	NR	4/11 (36%)	-
Sparse hair	6/10	NR (9/12)[d]	15/22 (68%)	-
Endocrine abnormalities	2/9	3/14	5/22 (16%)	-
Recurrent infections	2/10	3/14	5/24 (21%)	-
Ophthalmologic abnormalities	9/10[c]	7/13	16/23 (70%)	+
Deafness	1/10	NR	1/10 (10%)	-
Morphological features				
Long face	8/11	NR (7/12)[d]	15/23 (65%)	-
Facial asymmetry	3/10	9/14	12/24 (50%)	-
Broad forehead	9/11	14/14	23/25 (92%)	-
Ears abnormality	11/11	12/12	23/23 (100%)	-
Upslanting palpebral fissures	4/10	1/11	5/21 (24%)	-
Downslanting palpebral fissures	2/10	6/11	8/21 (38%)	-
Full nasal tip	8/10	11/13	19/23 (83%)	+
Malar hypoplasia	6/10	11/13	17/23 (74%)	-
Smooth philtrum	3/9	NR (2/12)[d]	5/21 (24%)	Deep
Thin upper lip	5/10	NR (1/12)[d]	6/22 (27%)	Thick
Thick lower lip	2/10	5/13	7/21 (33%)	+
Pointed chin	3/10	5/12	8/22 (36%)	-
Micrognathia	3/10	NR (3/12)[d]	6/22 (27%)	-