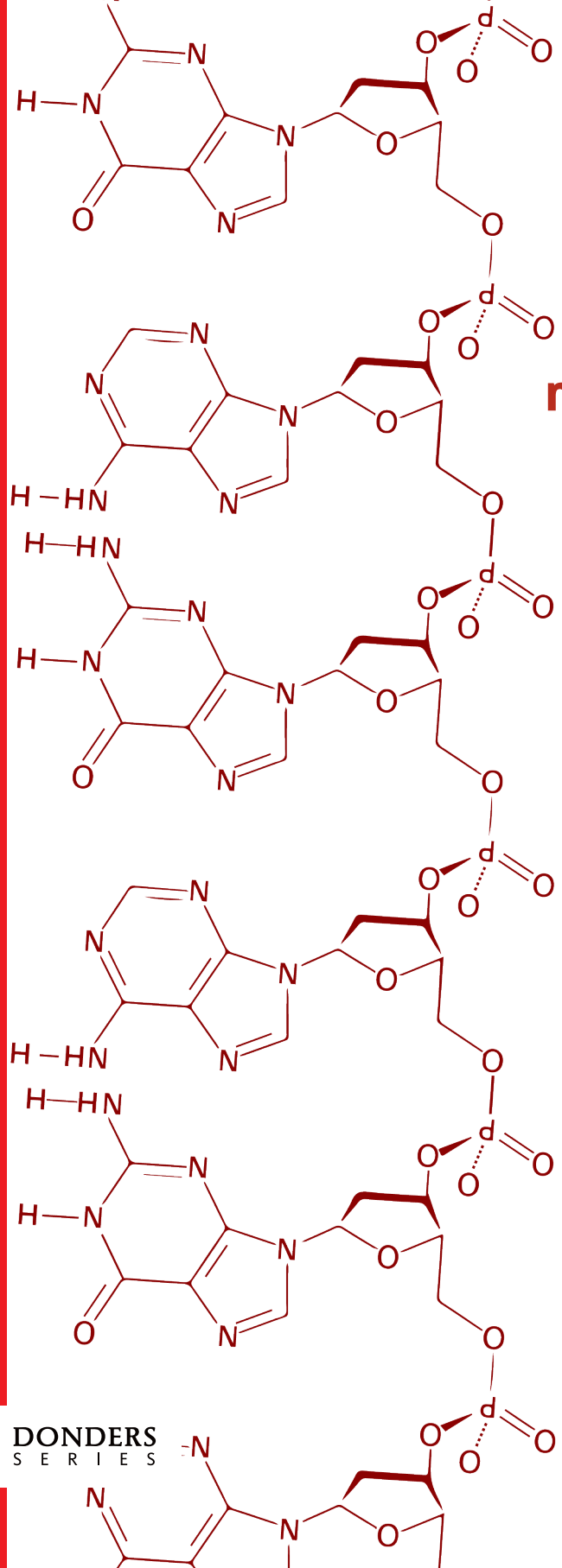


Clinical epigenetics of Mendelian neurodevelopmental disorders

Dmitrijs Rots

DONDERS SERIES

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Clinical epigenetics of Mendelian neurodevelopmental disorders

Dmitrijs Rots

The research presented in this thesis was carried out at the Radboudumc department of human genetics and within the Donders Institute for Brain, Cognition and Behaviour graduate school.

Author: Dmitrijs Rots
Title: Clinical epigenetics of Mendelian neurodevelopmental disorders

Radboud Dissertations Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Proefschrift AIO | Guus Gijben
Cover: Proefschrift AIO | Guntra Laivacuma
Printing: DPN Rikken/Pumbo

ISBN: 9789493296831
DOI: 10.54195/9789493296831
Free download at: www.boekenbestellen.nl/radboud-university-press/dissertations

RADBOUD
UNIVERSITY
PRESS

Clinical epigenetics of Mendelian neurodevelopmental disorders

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 25 november 2024
om 10.30 uur precies

door

Dmitrijs Rots

geboren op 23 december 1994

te Riga (Letland)

Promotoren:
Prof. dr. T. Kleefstra (Erasmus MC)
Prof. dr. H.G. Brunner
Prof. dr. L.E.L.M. Vissers

Manuscriptcommissie:
Prof. dr. B.P.C. van de Warrenburg
Prof. dr. M.A. Mannens (Amsterdam UMC)
Prof. dr. M. Meuwissen (UZ Antwerpen, België)

Clinical epigenetics of Mendelian neurodevelopmental disorders

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Monday, November 25, 2024
at 10.30 am

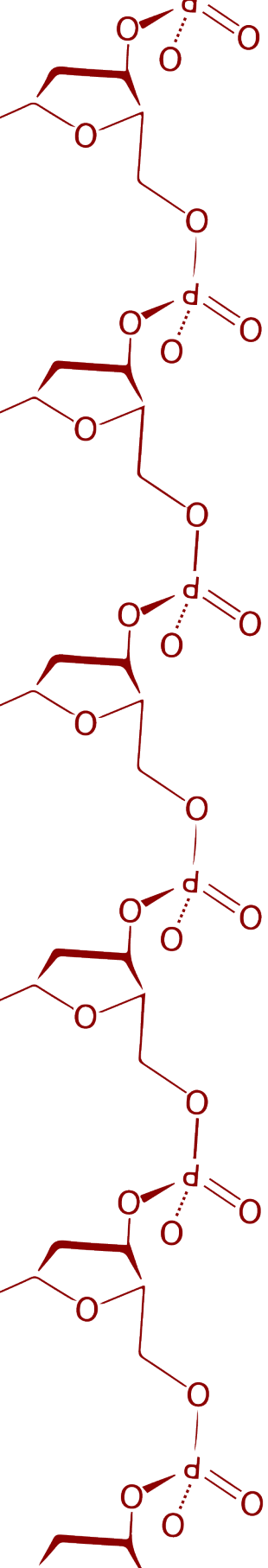
by

Dmitrijs Rots
born on December 23, 1994
in Riga (Latvia)

Supervisors:
Prof. dr. T. Kleefstra (Erasmus MC)
Prof. dr. H.G. Brunner
Prof. dr. L.E.L.M. Vissers

Manuscript Committee:
Prof. dr. B.P.C. van de Warrenburg
Prof. dr. M.A. Mannens (Amsterdam UMC)
Prof. dr. M. Meuwissen (UZ Antwerp, Belgium)

Table of contents



Chapter 1:

General introduction

What are Mendelian neurodevelopmental disorders?

Neurodevelopmental disorders (NDDs) encompass a heterogeneous group of conditions with onset during brain development and are characterized by significant impairments or deficits in cognitive, social, personal, occupational, or academic functions[1]. While the underlying causes of NDDs are diverse, a subset is caused by a pathogenic DNA variant affecting a single gene or locus and are transmitted in accordance with Mendel's principles are called as Mendelian or monogenic NDDs[2-4]. Over 1500 Mendelian NDDs are currently known, and many are likely yet to be discovered[5], so Mendelian NDDs are considered the most genetically and clinically heterogeneous monogenic conditions. Mendelian NDDs generally present with mild to severe intellectual disability (ID), autism spectrum disorders (ASD), developmental delays (DD), and/or other neurodevelopmental conditions (e.g., attention deficit and hyperactivity disorder [ADHD]), as well as congenital anomalies, growth abnormalities, and specific or nonspecific facial dysmorphic features[6]. However, Mendelian NDDs usually exhibit vast clinical heterogeneity with variable expressivity[7]. While most Mendelian NDDs are extremely rare, collectively they are common and represent ~2%–4% of the general population[6,8,9].

Identification of (dominant) NDD-associated genes: past and present

In the past two decades, advances in genetic analyses and molecular technologies have resulted in an explosion of novel disease-associated or morbid gene discovery, many of which are implicated in Mendelian NDD development[3,5].

The discovery of genes for specific, clinically recognized phenotypes, such as Floating-Harbor syndrome, relied on clinically diagnosed cohorts[10]. However, the molecular basis of most well-known clinically recognizable conditions is now known, with only a handful remaining unsolved, e.g., Gomez-Lopez-Hernandes syndrome and Hallermann-Streiff syndrome[11]. Importantly, the majority of NDDs, are rare, and certain disorders are hardly observed twice by the same clinician, which, in combination with the lack of a specific phenotype for the majority of NDDs, makes the phenotype-driven approach extremely challenging[12,13].

With the decreasing cost and increasing availability of (next-generation) sequencing, a genotype-driven approach with reverse phenotyping has become the most commonly used in recent years[14]. The *de novo* paradigm for NDDs revolutionized the field and led to the discovery of the majority of novel dominant genes[3]. In the genotype-driven approach, genome-wide testing, such as chromosomal microarray or trio exome/genome sequencing, is performed on an individual or cohort of individuals with NDDs, irrespective of their specific

clinical presentation. Individuals with overlapping potentially functional variants in the same candidate gene are then compared to identify a possible phenotypic overlap (reverse phenotyping) and define a novel NDD[14] as the occurrence of (*de novo*) variants in the same biologically relevant gene by chance in individuals with similar phenotype is extremely low[3]. It, however, commonly requires international collaborations and large cohorts to identify such (*de novo*) variants in the same gene to establish novel disease–gene associations[5,12]. Importantly, because of the absence of clinical specificity[15], additional alternative approaches are required to confirm causality, e.g., the use of model organisms, statistical enrichment of variants among affected individuals[5,16], or identification of shared molecular phenotypes.

Diagnosis of individuals with Mendelian NDDs: past and present

Similar to the discovery of Mendelian NDD genes, there has been a notable shift in the application of genetic testing from a phenotype-first to a genotype-first approach[13,17]. Traditionally, phenotype-first testing focused on targeted testing of a limited set of genes or chromosomal loci based on the observed symptoms and suspected clinical diagnosis with known genetic etiology in affected individuals (e.g., with Sanger sequencing or FISH, respectively)[13]. Consequently, our knowledge about certain recognizable syndromes is based mainly on the presentations of the most typical or similar features of the initially described individuals.

Genotype-first testing allows the identification of pathogenic causative variants genome-wide without a clear clinical diagnosis and irrespective of the clinical presentation[17]. The genotype-first approach has gained prominence because of 1) the discovery of a myriad of novel (nonspecific) morbid genes and syndromes, 2) the low diagnostic yield of phenotype-driven testing, and 3) the widespread availability of genome-wide testing methods such as chromosomal microarray and (trio) exome/genome sequencing. This approach has allowed not to only increase the diagnostic yield (and identify novel genes implicated in NDD), but also to diagnose individuals with atypical presentation who otherwise would not be recognized and to identify novel genotype-phenotype correlations[17]. As a result, (trio) exome/genome sequencing has become a first- or second-line test in many countries for individuals with suspected Mendelian NDDs, irrespective of their presentation[18,19]. Therefore, with international collaboration, it is currently possible to expand the clinical and genotypic spectrum of many well-known, clinically recognizable conditions. For example, truncating *ARID1B* variants were initially described as a cause of the clinically recognizable and well-known Coffin-Siris syndrome[20], although the analysis of individuals with pathogenic *ARID1B* variants identified through a genotype-first

approach revealed a broad clinical spectrum of the condition, with approximately half of affected individuals presenting with a nonsyndromic ID[21].

Diagnostic testing outcomes

Using the best current diagnostic genome-wide (exome or genome) sequencing methods allows for the direct identification of a molecular cause in approximately one third to one half of individuals with suspected Mendelian NDD – with the majority being unsolved[19,22,23]. In comparison, a molecular diagnosis was identified in only about 7% of individuals in the phenotype-first era[13,24]. A correct diagnosis is required for tailored care of the individuals, as well as the whole family[18]. While limitations of current technologies and bioinformatic tools are responsible for a large proportion of "missed" diagnoses, one of the main reasons for the limited diagnostic yield is our limited ability to correctly interpret genetic data[25,26]. Variant interpretation is based on multiple criteria (e.g., frequency in populational databases, inheritance, and *in silico* predictions), but is hugely based on the available knowledge about gene-related phenotype and genotype spectra[27], for example, if an investigated variant's type, position or genotype does not fit the currently described inheritance pattern for the known disease-gene relationship, the diagnostic testing results will likely be inconclusive, or if a individual's phenotype does not fit the currently described phenotype spectrum for a gene, the diagnostic testing results will also likely be inconclusive. In fact, a variant of uncertain significance (VUS) is reported in about one third of individuals[22,23].

Once more evidence becomes available, about one third of the reclassified VUS will become (likely) pathogenic and two thirds – (likely) benign[28]. Systematic re-analysis performed (with or without resequencing) over time has shown that a molecular diagnosis can still be identified in approximately half of the cases[23,25,26] (**Figure 1**). Novel identified morbid genes or novel genotype-phenotype correlations, together with improved bioinformatic tools and pipelines, are responsible for the majority of the increased diagnostic yield and reclassification of VUS[23,25,26]. Importantly, additional information allows us to reclassify variants that were initially misinterpreted as pathogenic[26]. It is only possible to reclassify a minority of VUS by using additional (functional) testing, e.g., metabolic testing or RNA analysis[23,25], because no suitable (functional) follow-up testing is available in clinical settings for the majority of genes. The expanding number of known NDD-associated genes, common atypical clinical presentations, as well as the increase in exome and genome sequencing performed result in an increasing number of reported VUS. Therefore, there is an urgent need for additional phenotypic, computational, and functional evidence to support variant classification[27,29-31].

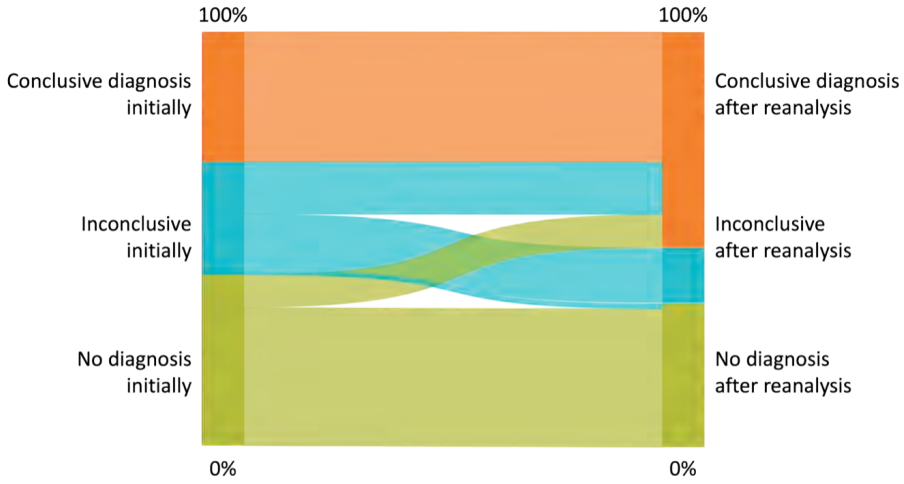


Figure 1. Exome or genome testing results: initially and after timely reanalysis.

Diagnostic testing result distribution (on the Y axis) with respective changes over time after reanalysis (on the X axis).

Epigenetics: basics and principles

Modern definition of the term "epigenetics":

- "changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence"[32];
- "a layer of information that exists beyond that encoded in the DNA sequence, thereby making the genome function distinctively in different cell types"[33].

The genes currently identified to be implicated in Mendelian NDDs have shed light on the molecular pathways crucial for normal brain development and are mainly involved in synaptic function or gene expression regulation, including genes encoding the epigenetic machinery[34,35].

Precise, and dynamic temporal and spatial control of gene expression is required for normal neuronal and non-neuronal cell differentiation, brain development, and cognitive processes. Factors that perturb these developmental processes can cause an NDD[36]. Notably, pathogenic variants in genes involved in epigenetic gene expression regulation are one of the most common causes of Mendelian NDDs[35] that belong to a disease group named Mendelian disorders of the epigenetic machinery (MDEM)[37]. Such genes play a critical role in differentiation and expression regulation across multiple cell and tissue types[35,37]. Consequently, individuals with MDEMs typically present with a dominant, syndromic Mendelian NDD with other comorbidities and congenital anomalies, reflecting the broad impact of dysregulated epigenetic processes on multiple levels of tissue development and function[37,38].

Epigenetic information regulates gene expression without changes in the DNA sequence itself[32]. It encompasses a wide range of molecular mechanisms, including 1) DNA methylation; 2) histone posttranslational modifications; 3) chromatin remodeling, etc. Such epigenetic "marks" can act as "on" or "off" switches, either allowing/promoting or prohibiting/repressing target gene expression, respectively[39,40] (**Figure 6)**. The epigenetic "marks" allow spatiotemporal gene expression regulation and are vital for the production of specialized cell (and tissue) types with different functions, morphologies, and gene expression[40] despite having largely the same genetic information within a single organism. Additionally, it plays an important role in the orchestration of various cellular processes throughout life as a response to environmental stimuli[39,41].

Epigenetic machinery

Epigenetic modifications are deposited, maintained, and erased by a myriad of proteins collectively described as the "epigenetic machinery" (**Figure 2**)[35,37]. The key players of the epigenetic machinery are 1) "writers", which add covalent modifications to histones or DNA; 2) "erasers", which remove these modifications; 3) "readers", which identify and bind to these modification and facilitating recruitment of other proteins; and 4) "remodelers", which modify chromatin structure by changing the content or positions of histones in the nucleosome[37].

The maintenance of epigenetic marks requires the simultaneous deposition and removal of various histone and DNA modifications, as well as complex crosstalk among multiple epigenetic machinery proteins and transcription factors[39]. Moreover, this complex interplay commonly requires a single epigenetic machinery protein having multiple enzymatic and nonenzymatic functions. This phenomenon, known as protein moonlighting[42], also highlights the diverse roles and functions of epigenetic machinery proteins beyond epigenetic regulation[42-44]. For instance, the KMT2C protein (**Figure 3**), a component of the COMPASS-like complex, is best known for its enzymatic histone-3 lysine-4 (H3K4) methylation function, which is performed by a "writer" SET domain[45,46]. In addition, it has an AT-hook, DHHC-type zinc finger, and eight "reader" PHD-type zinc finger domains - all of which are required for binding to DNA and chromatin[47], a WIN motif for binding to WDR5, as well as SET-domain regions required for interactions with other proteins (e.g., ASH2L and RBBP5[48]), FYRC/FYRN domains of unknown function, and an HMG box for binding and nuclear stabilization of KDM6A/B[49]. Furthermore, other nonenzymatic KMT2C functions have been suggested[50], and likely others are yet to be identified.

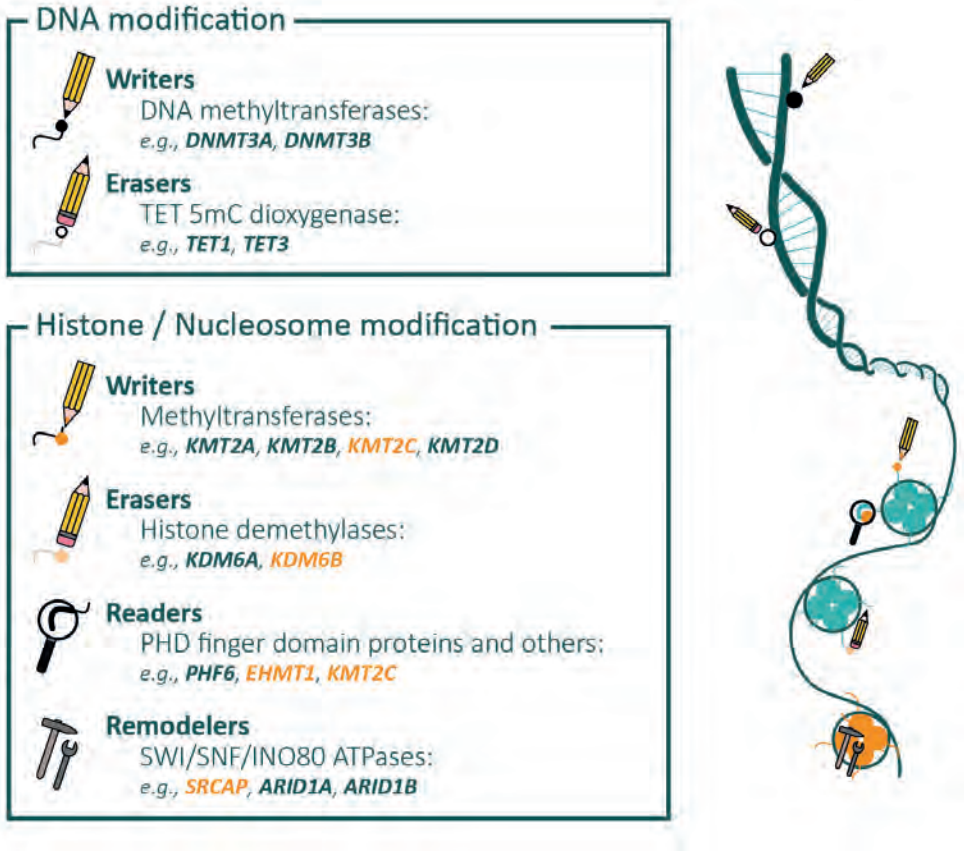


Figure 2. Epigenetic machinery main components and functions.

Genes written in orange have been investigated in this thesis, covering all main classes of the epigenetic machinery.

Similarly, multiple nonenzymatic and nonepigenetic functions of various histone modifiers have recently been identified and are actively investigated[50], indicating that moonlighting proteins are likely common within the epigenetic machinery.

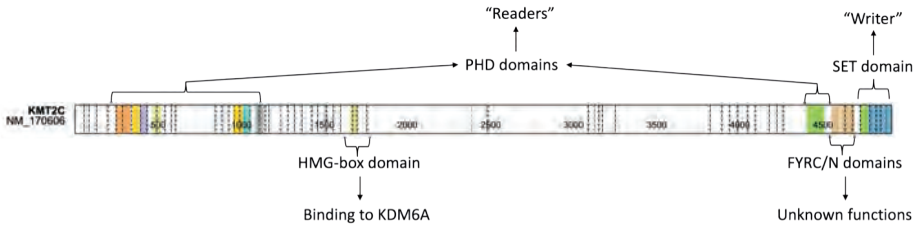


Figure 3. *KMT2C* structure and functions by domain.

Epigenetic marks

Epigenetic marks are controlled by complex interplay with other modifications and transcriptional activity, as well as by the local genome sequence, common and rare genetic variants, cell and tissue types, environmental factor exposure, and stochasticity[51].

DNA methylation is the most studied epigenetic modification. DNA methylation predominantly occurs at cytosine-guanine dinucleotide (CpG) sites, where a methyl group is covalently added to the cytosine (5mC)[52]. Cytosines are methylated at the majority of "solitary" CpG sites, while CpG-rich regions, known as CpG islands (CGIs), remain largely unmethylated[52,53]. CGIs overlap 70% of human promoters and play a crucial a role in gene expression regulation[53]. CpG island hypermethylation in promoter regions can lead to gene transcriptional silencing, while hypomethylation leads to the opposite effect[53,54]. Additionally, in human (and other mammalian) cells, non-CpG methylation (i.e., CpH) also occurs, but they are less abundant, and their functions are different from those of CpG methylation but are not yet fully understood[55].

DNA methylation at CpG sites is tightly regulated and maintained, remaining stable over time (both within cells and in stored in freezers). During early embryonic development, all DNA methylation marks in the human genome are erased, resetting the epigenetic landscape and allowing for subsequent cell differentiation[56]. Afterward, *de novo* DNA methyltransferases, such as DNMT3A and DNMT3B, methylate specific CpG sites (producing 5mC) in different cells, helping cell differentiation, adapting to environmental cues, and establishing cell-specific methylation patterns[57]. Next, DNA methylation at CpG sites is maintained over each cell division by a maintenance DNA methyltransferase - DNMT1[58]. Consequently, specialized cells could "inherit" certain DNA methylation patterns formed during early embryonic events[59]. In contrast, non-CpG methylation occurs on a single DNA strand and, therefore, cannot be maintained throughout cell division, making it abundant only in nondividing cells (such as neurons)[55]. DNA is demethylated via two pathways: 1) passively (e.g., by prevention of methylation after cell division) and 2) actively by oxidation of 5mC to 5hmC by TET enzymes with further repair by the DNA repair system[59] (**Figure 4**).

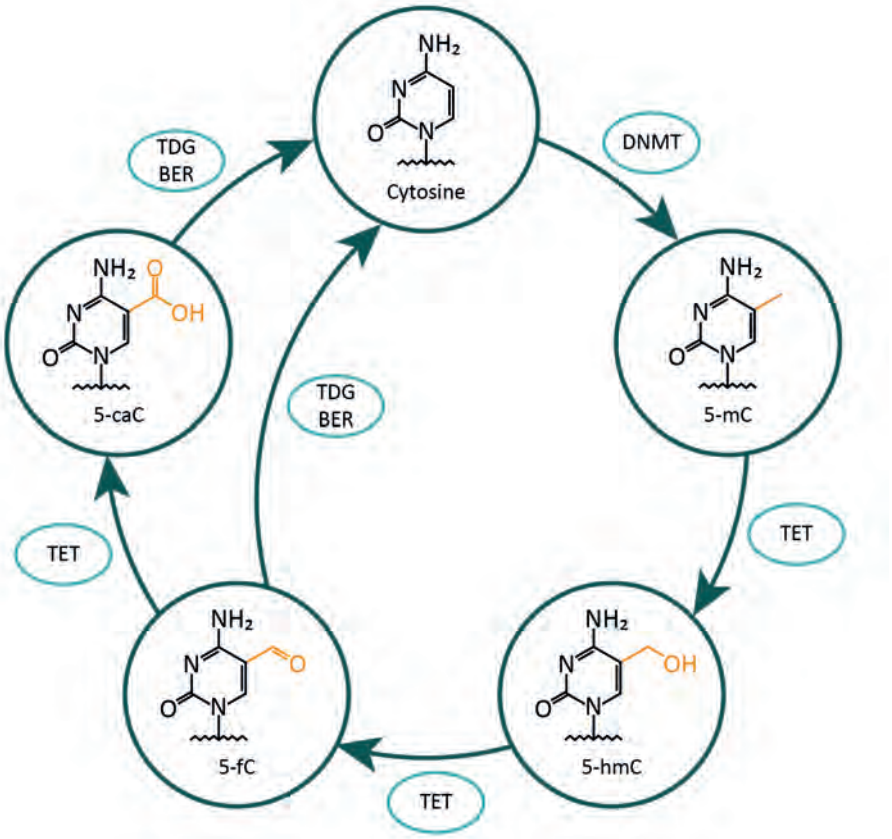


Figure 4. Biochemistry of DNA methylation and de-methylation.

DNMT = DNA methyltransferase; TET = ten eleven translocation enzyme; TDG = thymine DNA glycosylase enzyme; BER = base excision repair; 5-mC = 5-methylcytosine; 5-hmC = 5-hydroxymethylcytosine; 5-fC = 5-flucytosine; 5-caC = 5-carcoxylcytosine.

In contrast to DNA modifications, there is a myriad of known histone posttranslational modifications involving different covalent modifications (e.g., methylations, acetylation, ubiquitination, etc.), affecting different histone proteins and different positions within a histone (**Figure 5**)[41]. These modifications are also involved in the regulation of gene expression and the chromatin state, but the exact effects are known only for a handful of modifications[41,60]. For example, histone-3 lysine-4 di- and trimethylation (H3K4me2-3) is usually deployed by COMPASS and COMPASS-like complexes at gene promoters and is associated with active gene expression[40,45]. By contrast, H3K27me3 is associated with transcriptional repression and is deployed by the PRC2 complex[45,61]. For example, transcriptional repression

is associated with H3K27 trimethylation by the PRC2 complex and H3K4me3 demethylation by "eraser" KDM1A and DNA methylation nearby by recruitment of DNMT3A by PRC2[61,62], while transcriptional activation – with the opposite marks (**Figure 6**). However, the effects of epigenetic modifications are very complex and are cell and context dependent[54,60,63]. Additionally, epigenetic modifications usually occur together, involving and recruiting multiple proteins and creating a very complex "epigenetic code"[60,63].

DNAm signatures

Molecularly, many MDEMs are characterized by the presence of specific genome-wide DNA methylation (DNAm) changes called DNAm signatures[64] or episignatures[65]. Pioneering work by Grafodatskaya et al. (2013) and Choufani et al. (2015) identified DNAm signatures in blood-derived DNA of individuals with pathogenic variants in a histone modifier encoding *KDM5C*[66] and *NSD1*[67], respectively, and showed their utility for investigating the underlying "biology" of a condition and the ability to reclassify VUS[67,68]. This resulted in explosive interest in the field with >100 different genes and/or disorders having a described DNAm signature[64,65] in the last decade, with the numbers increasing every year. Consequently, the DNAm signature-based test ("EpiSign") is now available in diagnostic settings in many countries[69,70].

DNAm signature analysis is relatively affordable (comparable with exome sequencing), available, and efficient, so it has many applications in both diagnostics and research. In diagnostics, DNAm signature testing currently is mainly used for testing VUS in genes with known signatures[69,71]. It is also sometimes used to confirm a clinical diagnosis with "negative" genetic testing results[69,70]. In research, it additionally allows the investigation of the underlying "biology" of a condition, the identification of subgroups within a disorder that is clinically hard to distinguish, the analysis of whether conditions are molecularly related, and the discrimination of a novel disorder with many more scenarios being actively developed[72,73]. However, the DNAm signatures are not well understood yet, i.e., their causative mechanism, sensitivity, and specificity for different variant types, ages, and tissue types all are unknown. These factors not only limit the applications of DNAm signatures but could also lead to misinterpretation of their results and incorrect diagnoses[64,74].

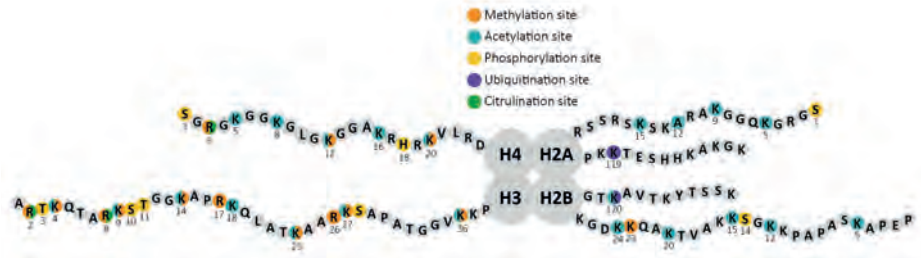


Figure 5. Examples of possible histone modifications and their positions within nucleosome.

Nucleosome protein core consisting of the four different histone proteins (H2A, H2B, H3 and H4; shown in grey). Histone protein tails are used for various posttranslational modifications on various positions (shown with respective colors) by the epigenetic machinery, for example, histone 3 lysine 4 (H3K4) can be mono-, di- or tri- methylated, or acetylated. Altogether, these modifications (also known as marks) create an epigenetic "histone code".

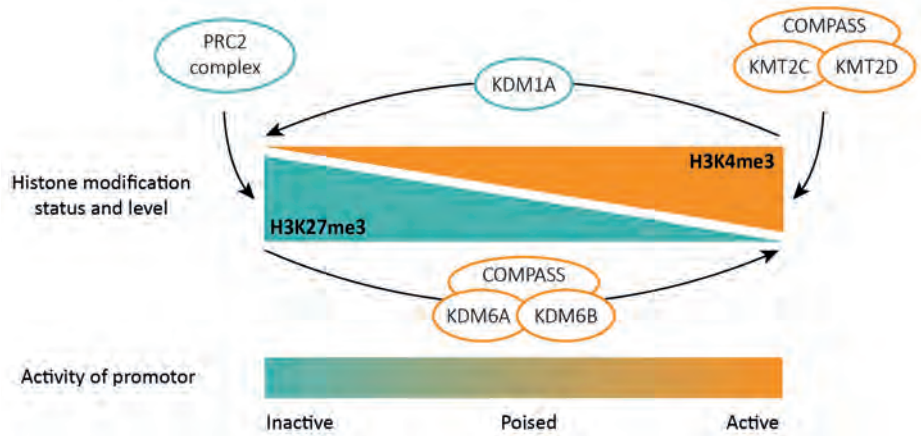


Figure 6. Interplay between H3K4me3 and H3K27me3 and their effects.

H3K4me3 is associated with the active promoter state, while H3K27me3 is associated with the inactive promoter. Usually, these two marks are mutually exclusive. Promoter activity can be upregulated by the "writing" activation mark H3K4me3 (e.g., by KMT2D-COMPASS or KMT2C-COMPASS complexes) and by (simultaneously) "erasing" the inhibition mark H3K27me3 (e.g., by KDM6A or KDM6B genes within COMPASS complexes). Promoter activity can be downregulated by "writing" the inhibition mark H3K27me3 (e.g., by the PRC2 complex) and by (simultaneously) erasing the activation mark H3K4me3 (e.g., by KDM1A).

Aims and scope of this thesis

Mendelian NDDs are extremely heterogeneous and rare, so only limited clinical and genetic information is available for the vast majority of known conditions, and many conditions are yet to be discovered. This poses major challenges for establishing a correct diagnosis (in terms of both variant interpretation and the evaluation of clinical phenotype) and tailoring of prognosis and care to affected individuals and their families. Moreover, a lack of understanding of disease mechanisms limits our ability to develop and investigate suitable treatment options in the future. Therefore, in this thesis, we **characterized the clinical, molecular, and "epigenetic" spectrum and features of several Mendelian NDDs focusing on genes coding the epigenetic machinery**:

- **Chapter 2** describes a novel Mendelian NDD with a specific DNAm signature caused by truncating variants within a specific location of *SRCAP*, previously associated with Floating-Harbor syndrome. We show that the location of truncating variants within *SRCAP* determines the phenotype and DNAm signature, highlighting the functions of different SRCAP regions.

- **Chapter 3** describes the clinical and molecular spectrum of *KDM6B*-related NDD, which was previously named in the OMIM as a "Neurodevelopmental disorder with coarse facies and mild distal skeletal abnormalities" based on a small, initially described cohort. We show that a large cohort analysis allows us to identify many typical clinical features of the disorder that were missed in the initial description and that some of the initially described common symptoms, like coarse facies and distal skeletal abnormalities, are in fact rare and lead to misleading naming of the condition. Additionally, we provide *Drosophila* models that recapitulate the human neurodevelopmental phenotype and allow *KDM6B* VUS testing.

- **Chapter 4** delineates the clinical and molecular spectrum of *KMT2C*-related NDD, which was initially designated by the OMIM as "Kleefstra syndrome 2" based on a small, initially described cohort. We show that *KMT2C*-related NDD has specific clinical, facial, molecular, and DNAm signature that is distinct from Kleefstra and Kabuki syndrome. We also show that disruption of a single protein domain can have similar effects to *KMT2C* haploinsufficiency, highlighting the importance of different *KMT2C* protein functions and/or domains.

- **Chapter 5** describes the clinical and molecular spectrum of pathogenic *EHMT1* variants. We show that the clinical and molecular spectrum of Kleefstra syndrome is much broader than previously published, identify novel genotype-phenotype correlations, and discover novel *EHMT1* genotype subgroups resulting in atypical phenotypes. Detailed analysis and functional testing of various variants reveal that the loss of the enzymatic EHMT1 activity is not the "driver" of the DNAm signature and Kleefstra syndrome pathogenesis, but rather the loss of the other functions, raising questions about biological role of the EHMT1 enzymatic activity.

- **Chapter 6** describes the clinical and DNAm methylation features of 9q34.3 duplications containing *EHMT1*, highlighting the effects of an increased *EHMT1* copy number on neurodevelopment. We also show that *EHMT1* duplications do not mirror deletions clinically, nor on DNAm, as has been described for several deletion/duplication syndromes.

References

1. Diagnostic and statistical manual of mental disorders : DSM-5. (2013). 5th ed. ed. American Psychiatric Association.
2. Rasmussen, S.A., Hamosh, A., Amberger, J., Arnold, C., Bocchini, C., O'Neill, M.J.F., Stumpf, A., and for the, O.c. (2020). What's in a name? Issues to consider when naming Mendelian disorders. Genetics in Medicine *22*, 1573-1575. 10.1038/s41436-020-0851-0.
3. Vissers, L.E., Gilissen, C., and Veltman, J.A. (2016). Genetic studies in intellectual disability and related disorders. Nat Rev Genet *17*, 9-18. 10.1038/nrg3999.
4. Parenti, I., Rabaneda, L.G., Schoen, H., and Novarino, G. (2020). Neurodevelopmental Disorders: From Genetics to Functional Pathways. Trends Neurosci *43*, 608-621. 10.1016/j.tins.2020.05.004.
5. Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., et al. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature *586*, 757-762. 10.1038/s41586-020-2832-5.
6. Prevalence and architecture of de novo mutations in developmental disorders. (2017). Nature *542*, 433-438. 10.1038/nature21062.
7. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet *385*, 1305-1314. 10.1016/s0140-6736(14)61705-0.
8. Anderson, L.L., Larson, S.A., MapelLentz, S., and Hall-Lande, J. (2019). A Systematic Review of U.S. Studies on the Prevalence of Intellectual or Developmental Disabilities Since 2000. Intellect Dev Disabil *57*, 421-438. 10.1352/1934-9556-57.5.421.
9. Maulik, P.K., Mascarenhas, M.N., Mathers, C.D., Dua, T., and Saxena, S. (2011). Prevalence of intellectual disability: a meta-analysis of population-based studies. Res Dev Disabil *32*, 419-436. 10.1016/j.ridd.2010.12.018.
10. Hood, R.L., Lines, M.A., Nikkel, S.M., Schwartzentruber, J., Beaulieu, C., Nowaczyk, M.J., Allanson, J., Kim, C.A., Wieczorek, D., Moilanen, J.S., et al. (2012). Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. Am J Hum Genet *90*, 308-313. 10.1016/j.ajhg.2011.12.001.
11. Zurek, B., Ellwanger, K., Vissers, L., Schüle, R., Synofzik, M., Töpf, A., de Voer, R.M., Laurie, S., Matalonga, L., Gilissen, C., et al. (2021). Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. Eur J Hum Genet *29*, 1325-1331. 10.1038/s41431-021-00859-0.
12. Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2012). Disease gene identification strategies for exome sequencing. Eur J Hum Genet *20*, 490-497. 10.1038/ejhg.2011.258.
13. Jansen, S., Vissers, L., and de Vries, B.B.A. (2023). The Genetics of Intellectual Disability. Brain Sci *13*. 10.3390/brainsci13020231.
14. Wilczewski, C.M., Obasohan, J., Paschall, J.E., Zhang, S., Singh, S., Maxwell, G.L., Similuk, M., Wolfsberg, T.G., Turner, C., Biesecker, L.G., and Katz, A.E. (2023). Genotype first: Clinical genomics research through a reverse phenotyping approach. American journal of human genetics *110*, 3-12. 10.1016/j.ajhg.2022.12.004.
15. Vissers, L., Kalvakuri, S., de Boer, E., Geuer, S., Oud, M., van Outersterp, I., Kwint, M., Witmond, M., Kersten, S., Polla, D.L., et al. (2020). De Novo Variants in CNOT1, a Central Component of the CCR4-NOT Complex Involved in Gene Expression and RNA and Protein Stability, Cause Neurodevelopmental Delay. American journal of human genetics *107*, 164-172. 10.1016/j.ajhg.2020.05.017.

16. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469-476. 10.1038/nature13127.

17. Vissers, L., van Nimwegen, K.J.M., Schieving, J.H., Kamsteeg, E.J., Kleefstra, T., Yntema, H.G., Pfundt, R., van der Wilt, G.J., Krabbenborg, L., Brunner, H.G., et al. (2017). A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. Genetics in medicine : official journal of the American College of Medical Genetics *19*, 1055-1063. 10.1038/gim.2017.1.

18. Manickam, K., McClain, M.R., Demmer, L.A., Biswas, S., Kearney, H.M., Malinowski, J., Massingham, L.J., Miller, D., Yu, T.W., and Hisama, F.M. (2021). Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). Genet Med *23*, 2029-2037. 10.1038/s41436-021-01242-6.

19. Srivastava, S., Love-Nichols, J.A., Dies, K.A., Ledbetter, D.H., Martin, C.L., Chung, W.K., Firth, H.V., Frazier, T., Hansen, R.L., Prock, L., et al. (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. Genet Med *21*, 2413-2421. 10.1038/s41436-019-0554-6.

20. Santen, G.W., Aten, E., Sun, Y., Almomani, R., Gilissen, C., Nielsen, M., Kant, S.G., Snoeck, I.N., Peeters, E.A., Hilhorst-Hofstee, Y., et al. (2012). Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome. Nat Genet *44*, 379-380. 10.1038/ng.2217.

21. van der Sluijs, P.J., Jansen, S., Vergano, S.A., Adachi-Fukuda, M., Alanay, Y., AlKindy, A., Baban, A., Bayat, A., Beck-Wödl, S., Berry, K., et al. (2019). The ARID1B spectrum in 143 patients: from nonsyndromic intellectual disability to Coffin-Siris syndrome. Genet Med *21*, 1295-1307. 10.1038/s41436-018-0330-z.

22. van der Sanden, B., Schobers, G., Corominas Galbany, J., Koolen, D.A., Sinnema, M., van Reeuwijk, J., Stumpel, C., Kleefstra, T., de Vries, B.B.A., Ruiterkamp-Versteeg, M., et al. (2023). The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. Eur J Hum Genet *31*, 81-88. 10.1038/s41431-022-01185-9.

23. Schobers, G., Schieving, J.H., Yntema, H.G., Pennings, M., Pfundt, R., Derks, R., Hofste, T., de Wijs, I., Wieskamp, N., van den Heuvel, S., et al. (2022). Reanalysis of exome negative patients with rare disease: a pragmatic workflow for diagnostic applications. Genome Med *14*, 66. 10.1186/s13073-022-01069-z.

24. Neveling, K., Feenstra, I., Gilissen, C., Hoefsloot, L.H., Kamsteeg, E.J., Mensenkamp, A.R., Rodenburg, R.J., Yntema, H.G., Spruijt, L., Vermeer, S., et al. (2013). A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. Human mutation *34*, 1721-1726. 10.1002/humu.22450.

25. Liu, P., Meng, L., Normand, E.A., Xia, F., Song, X., Ghazi, A., Rosenfeld, J., Magoulas, P.L., Braxton, A., Ward, P., et al. (2019). Reanalysis of Clinical Exome Sequencing Data. N Engl J Med *380*, 2478-2480. 10.1056/NEJMc1812033.

26. Wright, C.F., McRae, J.F., Clayton, S., Gallone, G., Aitken, S., FitzGerald, T.W., Jones, P., Prigmore, E., Rajan, D., Lord, J., et al. (2018). Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. Genet Med *20*, 1216-1223. 10.1038/gim.2017.246.

27. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). 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 *17*, 405-424. 10.1038/gim.2015.30.

28. Harrison, S.M., and Rehm, H.L. (2019). Is 'likely pathogenic' really 90% likely? Reclassification data in ClinVar. Genome Med *11*, 72. 10.1186/s13073-019-0688-9.
29. Wortmann, S.B., Oud, M.M., Alders, M., Coene, K.L.M., van der Crabben, S.N., Feichtinger, R.G., Garanto, A., Hoischen, A., Langeveld, M., Lefeber, D., et al. (2022). How to proceed after "negative" exome: A review on genetic diagnostics, limitations, challenges, and emerging new multiomics techniques. J Inherit Metab Dis *45*, 663-681. 10.1002/jimd.12507.
30. Hack, J.B., Horning, K., Juroske Short, D.M., Schreiber, J.M., Watkins, J.C., and Hammer, M.F. (2023). Distinguishing Loss-of-Function and Gain-of-Function SCN8A Variants Using a Random Forest Classification Model Trained on Clinical Features. Neurol Genet *9*, e200060. 10.1212/nxg.0000000000200060.
31. Dingemans, A.J.M., Hinne, M., Truijen, K.M.G., Goltstein, L., van Reeuwijk, J., de Leeuw, N., Schuurs-Hoeijmakers, J., Pfundt, R., Diets, I.J., den Hoed, J., et al. (2023). PhenoScore quantifies phenotypic variation for rare genetic diseases by combining facial analysis with other clinical features using a machine-learning framework. Nature genetics *55*, 1598-1607. 10.1038/s41588-023-01469-w.
32. Dupont, C., Armant, D.R., and Brenner, C.A. (2009). Epigenetics: definition, mechanisms and clinical perspective. Semin Reprod Med *27*, 351-357. 10.1055/s-0029-1237423.
33. Greally, J.M. (2018). A user's guide to the ambiguous word 'epigenetics'. Nat Rev Mol Cell Biol *19*, 207-208. 10.1038/nrm.2017.135.
34. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568-584.e523. 10.1016/j.cell.2019.12.036.
35. Ciptasari, U., and van Bokhoven, H. (2020). The phenomenal epigenome in neurodevelopmental disorders. Hum Mol Genet *29*, R42-r50. 10.1093/hmg/ddaa175.
36. Khodosevich, K., and Sellgren, C.M. (2023). Neurodevelopmental disorders-high-resolution rethinking of disease modeling. Mol Psychiatry *28*, 34-43. 10.1038/s41380-022-01876-1.
37. Fahrner, J.A., and Bjornsson, H.T. (2019). Mendelian disorders of the epigenetic machinery: postnatal malleability and therapeutic prospects. Hum Mol Genet *28*, R254-r264. 10.1093/hmg/ddz174.
38. Kleefstra, T., Schenck, A., Kramer, J.M., and van Bokhoven, H. (2014). The genetics of cognitive epigenetics. Neuropharmacology *80*, 83-94. 10.1016/j.neuropharm.2013.12.025.
39. Soshnev, A.A., Josefowicz, S.Z., and Allis, C.D. (2016). Greater Than the Sum of Parts: Complexity of the Dynamic Epigenome. Mol Cell *62*, 681-694. 10.1016/j.molcel.2016.05.004.
40. Gerrard, D.T., Berry, A.A., Jennings, R.E., Birket, M.J., Zarrineh, P., Garstang, M.G., Withey, S.L., Short, P., Jiménez-Gancedo, S., Firbas, P.N., et al. (2020). Dynamic changes in the epigenomic landscape regulate human organogenesis and link to developmental disorders. Nat Commun *11*, 3920. 10.1038/s41467-020-17305-2.
41. Millán-Zambrano, G., Burton, A., Bannister, A.J., and Schneider, R. (2022). Histone post-translational modifications - cause and consequence of genome function. Nat Rev Genet *23*, 563-580. 10.1038/s41576-022-00468-7.
42. Jeffery, C.J. (1999). Moonlighting proteins. Trends Biochem Sci *24*, 8-11. 10.1016/s0968-0004(98)01335-8.
43. Yuan, A.H., and Moazed, D. (2024). Minimal requirements for the epigenetic inheritance of engineered silent chromatin domains. Proc Natl Acad Sci U S A *121*, e2318455121. 10.1073/pnas.2318455121.

44. Morgan, M.A.J., and Shilatifard, A. (2023). Epigenetic moonlighting: Catalytic-independent functions of histone modifiers in regulating transcription. Sci Adv *9*, eadg6593. 10.1126/sciadv.adg6593.

45. Cenik, B.K., and Shilatifard, A. (2021). COMPASS and SWI/SNF complexes in development and disease. Nat Rev Genet *22*, 38-58. 10.1038/s41576-020-0278-0.

46. Hu, D., Gao, X., Morgan, M.A., Herz, H.M., Smith, E.R., and Shilatifard, A. (2013). The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. Mol Cell Biol *33*, 4745-4754. 10.1128/mcb.01181-13.

47. Stroynowska-Czerwinska, A.M., Klimczak, M., Pastor, M., Kazrani, A.A., Misztal, K., and Bochtler, M. (2023). Clustered PHD domains in KMT2/MLL proteins are attracted by H3K4me3 and H3 acetylation-rich active promoters and enhancers. Cell Mol Life Sci *80*, 23. 10.1007/s00018-022-04651-1.

48. Xue, H., Yao, T., Cao, M., Zhu, G., Li, Y., Yuan, G., Chen, Y., Lei, M., and Huang, J. (2019). Structural basis of nucleosome recognition and modification by MLL methyltransferases. Nature *573*, 445-449. 10.1038/s41586-019-1528-1.

49. Rickels, R., Wang, L., Iwanaszko, M., Ozark, P.A., Morgan, M.A., Piunti, A., Khalatyan, N., Soliman, S.H.A., Rendleman, E.J., Savas, J.N., et al. (2020). A small UTX stabilization domain of Trr is conserved within mammalian MLL3-4/COMPASS and is sufficient to rescue loss of viability in null animals. Genes Dev *34*, 1493-1502. 10.1101/gad.339762.120.

50. Aubert, Y., Egolf, S., and Capell, B.C. (2019). The Unexpected Noncatalytic Roles of Histone Modifiers in Development and Disease. Trends Genet *35*, 645-657. 10.1016/j.tig.2019.06.004.

51. Feinberg, A.P. (2018). The Key Role of Epigenetics in Human Disease Prevention and Mitigation. N Engl J Med *378*, 1323-1334. 10.1056/NEJMra1402513.

52. Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209-213. 10.1038/321209a0.

53. Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev *25*, 1010-1022. 10.1101/gad.2037511.

54. de Mendoza, A., Nguyen, T.V., Ford, E., Poppe, D., Buckberry, S., Pflueger, J., Grimmer, M.R., Stolzenburg, S., Bogdanovic, O., Oshlack, A., et al. (2022). Large-scale manipulation of promoter DNA methylation reveals context-specific transcriptional responses and stability. Genome Biol *23*, 163. 10.1186/s13059-022-02728-5.

55. Patil, V., Ward, R.L., and Hesson, L.B. (2014). The evidence for functional non-CpG methylation in mammalian cells. Epigenetics *9*, 823-828. 10.4161/epi.28741.

56. Guo, F., Li, X., Liang, D., Li, T., Zhu, P., Guo, H., Wu, X., Wen, L., Gu, T.P., Hu, B., et al. (2014). Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. Cell Stem Cell *15*, 447-459. 10.1016/j.stem.2014.08.003.

57. Chen, B.F., and Chan, W.Y. (2014). The de novo DNA methyltransferase DNMT3A in development and cancer. Epigenetics *9*, 669-677. 10.4161/epi.28324.

58. Moore, L.D., Le, T., and Fan, G. (2013). DNA methylation and its basic function. Neuropsychopharmacology *38*, 23-38. 10.1038/npp.2012.112.

59. Wu, X., and Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev Genet *18*, 517-534. 10.1038/nrg.2017.33.

60. Lukauskas, S., Tvardovskiy, A., Nguyen, N.V., Stadler, M., Faull, P., Ravnsborg, T., Özdemir Aygenli, B., Dornauer, S., Flynn, H., Lindeboom, R.G.H., et al. (2024). Decoding chromatin states by proteomic profiling of nucleosome readers. Nature *627*, 671-679. 10.1038/s41586-024-07141-5.

61. Deevy, O., and Bracken, A.P. (2019). PRC2 functions in development and congenital disorders. Development *146*. 10.1242/dev.181354.

62. Manzo, M., Wirz, J., Ambrosi, C., Villaseñor, R., Roschitzki, B., and Baubec, T. (2017). Isoform-specific localization of DNMT3A regulates DNA methylation fidelity at bivalent CpG islands. Embo j *36*, 3421-3434. 10.15252/embj.201797038.

63. Policarpi, C., Munafò, M., Tsagkris, S., Carlini, V., and Hackett, J.A. (2022). Systematic Epigenome Editing Captures the Context-dependent Instructive Function of Chromatin Modifications. bioRxiv, 2022.2009.2004.506519. 10.1101/2022.09.04.506519.

64. Awamleh, Z., Goodman, S., Choufani, S., and Weksberg, R. (2024). DNA methylation signatures for chromatinopathies: current challenges and future applications. Hum Genet *143*, 551-557. 10.1007/s00439-023-02544-2.

65. Levy, M.A., McConkey, H., Kerkhof, J., Barat-Houari, M., Bargiacchi, S., Biamino, E., Bralo, M.P., Cappuccio, G., Ciolfi, A., Clarke, A., et al. (2022). Novel diagnostic DNA methylation episignatures expand and refine the epigenetic landscapes of Mendelian disorders. HGG Adv *3*, 100075. 10.1016/j.xhgg.2021.100075.

66. Grafodatskaya, D., Chung, B.H., Butcher, D.T., Turinsky, A.L., Goodman, S.J., Choufani, S., Chen, Y.A., Lou, Y., Zhao, C., Rajendram, R., et al. (2013). Multilocus loss of DNA methylation in individuals with mutations in the histone H3 lysine 4 demethylase KDM5C. BMC Med Genomics *6*, 1. 10.1186/1755-8794-6-1.

67. Choufani, S., Cytrynbaum, C., Chung, B.H., Turinsky, A.L., Grafodatskaya, D., Chen, Y.A., Cohen, A.S., Dupuis, L., Butcher, D.T., Siu, M.T., et al. (2015). NSD1 mutations generate a genome-wide DNA methylation signature. Nat Commun *6*, 10207. 10.1038/ncomms10207.

68. Butcher, D.T., Cytrynbaum, C., Turinsky, A.L., Siu, M.T., Inbar-Feigenberg, M., Mendoza-Londono, R., Chitayat, D., Walker, S., Machado, J., Caluseriu, O., et al. (2017). CHARGE and Kabuki Syndromes: Gene-Specific DNA Methylation Signatures Identify Epigenetic Mechanisms Linking These Clinically Overlapping Conditions. Am J Hum Genet *100*, 773-788. 10.1016/j.ajhg.2017.04.004.

69. Sadikovic, B., Levy, M.A., Kerkhof, J., Aref-Eshghi, E., Schenkel, L., Stuart, A., McConkey, H., Henneman, P., Venema, A., Schwartz, C.E., et al. (2021). Clinical epigenomics: genome-wide DNA methylation analysis for the diagnosis of Mendelian disorders. Genet Med *23*, 1065-1074. 10.1038/s41436-020-01096-4.

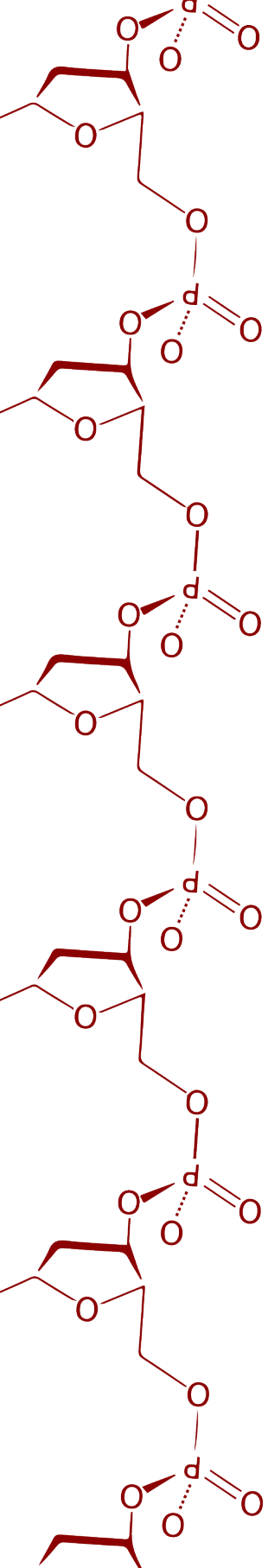
70. Kerkhof, J., Rastin, C., Levy, M.A., Relator, R., McConkey, H., Demain, L., Dominguez-Garrido, E., Kaat, L.D., Houge, S.D., DuPont, B.R., et al. (2024). Diagnostic utility and reporting recommendations for clinical DNA methylation episignature testing in genetically undiagnosed rare diseases. Genetics in medicine : official journal of the American College of Medical Genetics *26*, 101075. 10.1016/j.gim.2024.101075.

71. Mannens, M., Lombardi, M.P., Alders, M., Henneman, P., and Bliek, J. (2022). Further Introduction of DNA Methylation (DNAm) Arrays in Regular Diagnostics. Front Genet *13*, 831452. 10.3389/fgene.2022.831452.

72. Levy, M.A., Relator, R., McConkey, H., Pranckeviciene, E., Kerkhof, J., Barat-Houari, M., Bargiacchi, S., Biamino, E., Palomares Bralo, M., Cappuccio, G., et al. (2022). Functional correlation of genome-wide DNA methylation profiles in genetic neurodevelopmental disorders. Hum Mutat *43*, 1609-1628. 10.1002/humu.24446.

73. Sadikovic, B., Levy, M.A., and Aref-Eshghi, E. (2020). Functional annotation of genomic variation: DNA methylation episignatures in neurodevelopmental Mendelian disorders. Hum Mol Genet *29*, R27-r32. 10.1093/hmg/ddaa144.

74. Chater-Diehl, E., Goodman, S.J., Cytrynbaum, C., Turinsky, A.L., Choufani, S., and Weksberg, R. (2021). Anatomy of DNA methylation signatures: Emerging insights and applications. Am J Hum Genet *108*, 1359-1366. 10.1016/j.ajhg.2021.06.015.



Chapter 2:

Truncating *SRCAP* variants outside the Floating-Harbor syndrome locus cause a distinct neurodevelopmental disorder with a specific DNA methylation signature

Published: American Journal of Human Genetics. 2021 Jun 3; 108(6): 1053–1068.

Authors

Dmitrijs Rots*, Eric Chater-Diehl*, Alexander J.M. Dingemans, Sarah Goodman, Michelle Siu, Cheryl Cytrynbaum, Sanaa Choufani, Ny Hoang, Susan Walker, Zain Awamleh, Joshua Charlow, Stephen Meyn, Rolph Pfundt, Tuula Rinne, Thatjana Gardeitchik, Bert B.A. de Vries, A. Chantal Deden, Erika Leenders, Michael Kwint, Constance T.R.M. Stumpel, Servi J.C. Stevens, Jeroen R. Vermeulen, Jeske V.T. van Harssel, Danielle G.M. Bosch, Koen L.I. van Gassen, Ellen van Binsbergen, Christa M. de Geus, Hein Brackel, Maja Hempel, Davor Lessel,Jonas Denecke,Anne Slavotinek , Jonathan Strober, Amy Crunk, Leandra Folk,Ingrid M. Wentzensen, Hui Yang, Fanggeng Zou, Francisca Millan, Richard Person, Yili Xie, Shuxi Liu, Lilian B. Ousager, Martin Larsen, Laura Schultz-Rogers, Eva Morava, Eric W. Klee, Ian R. Berry, Jennifer Campbell, Kristin Lindstrom, Brianna Pruniski, Ann M. Neumeyer, Jessica A. Radley, Chanika Phornphutkul, Berkley Schmidt,William G. Wilson, Katrin Õunap, Karit Reinson, Sander Pajusalu,Arie van Haeringen, Claudia Ruivenkamp, Roos Cuperus, Fernando Santos-Simarro, María Palomares-Bralo, Marta Pacio-Míguez, Alyssa Ritter, Elizabeth Bhoj, Elin Tønne, Kristian Tveten, Gerarda Cappuccio, Nicola Brunetti-Pierri, Leah Rowe, Jason Bunn, Margarita Saenz, Konrad Platzer, Mareike Mertens, Oana Caluseriu, Małgorzata J.M. Nowaczyk, Ronald D. Cohn, Peter Kannu, Ebba Alkhunaizi, David Chitayat, Stephen W. Scherer, Han G. Brunner, Lisenka E.L.M. Vissers, Tjitske Kleefstra, David A. Koolen**, Rosanna Weksberg**

*,** These authors contributed equally to this work

Abstract

Truncating variants in exons 33 and 34 of the SNF2-related CREBBP activator protein (*SRCAP*) gene cause the neurodevelopmental disorder (NDD) Floating Harbor syndrome (FLHS), characterized by short stature, speech delay and facial dysmorphism. Here, we present a cohort of 33 individuals with clinical features distinct from FLHS and truncating (mostly *de novo*) *SRCAP* variants either proximal (n=28) or distal (n=5) to the FLHS locus. Detailed clinical characterization of the proximal *SRCAP* individuals identified shared characteristics: developmental delay with or without intellectual disability, behavioral and psychiatric problems, non-specific facial features, musculoskeletal issues, and hypotonia. Because FLHS is known to be associated with a unique set of DNA methylation (DNAm) changes in blood, a DNAm signature, we investigated whether there was a distinct signature associated with our cases. A machine-learning model, based on the FLHS DNAm signature, negatively classified all our tested cases. Comparing proximal cases with typically developing controls, we identified a DNAm signature distinct from the FLHS signature. Based on the DNAm and clinical data, we refer to the condition as "non-FLHS *SRCAP*-related NDD". All five distal variants classified negatively using the FLHS DNAm model while two classified positively using the proximal model. This suggests divergent pathogenicity of these variants, though clinically the distal group presented with NDD, similarly to the proximal *SRCAP* group. In sum, for *SRCAP*, there is a clear relationship between variant location, DNAm profile, and clinical phenotype. These results highlight the power of combined epigenetic, molecular, and clinical studies to identify and characterize new genotype-epigenotype-phenotype correlations.

Keywords:

SRCAP, Floating-Harbor syndrome, Speech delay, Intellectual disability, DNA methylation signature, Neurodevelopmental disorders

Introduction

Chromatin remodelers and other epigenetic regulators play a central role in numerous neurodevelopmental processes[1], and, therefore, pathogenic variants in the encoding genes commonly result in neurodevelopmental disorders (NDD)[2]. The *SRCAP* gene encodes the SNF2-related CREBBP activator protein (SRCAP; [OMIM: 611421]), which is an important component of the SRCAP chromatin remodeling complex which regulates transcription of various target genes by incorporating H2AZ-H2B dimers into nucleosomes[3]. Truncating variants in *SRCAP* have been identified as the cause of Floating-Harbor syndrome (FLHS; MIM: 136140)[4]. To date, all FLHS-causing variants have been mapped to the last two exons (exons 33-34) of *SRCAP*, upstream of the AT-hooks of the gene, in a locus further referred as the "FLHS locus"[4-7]. FLHS is a well-recognizable syndrome characterized by a clinical triad consisting of 1) typical craniofacial features (including triangular face, deep-set eyes, broad nose with bulbous tip, low-hanging columella, short philtrum and thin lips), 2) expressive and receptive speech and language delay, and 3) proportionate short stature with delayed bone age. In addition, most individuals show some degree of developmental delay (DD), or intellectual disability (ID)[8]. Although *SRCAP* is widely studied in the context of FLHS[3], the consequences of *SRCAP* variants located outside of the FLHS-causing locus are unknown, despite their increasing identification through next generation sequencing

Different pathogenic variant types and locations within the same gene can be associated with distinct neurodevelopmental disorders[9-13]. DNA methylation (DNAm) is an emerging functional tool that can be used to identify and characterize such disorders when the pathogenic variants occur in epigenetic regulatory genes by the identification of specific patterns of genome-wide DNAm in peripheral blood which we call *DNAm signatures*[14]. To date, we and others have described >50 DNAm signatures associated with epigenetic regulatory genes[14-24], which are particularly useful for classifying variants of uncertain significance (VUS) in these genes as pathogenic or benign as they provide a functional readout of pathogenicity. DNAm signatures can also help to discriminate between related disorders in a differential diagnosis[18,21,23]. Truncating variants associated with FLHS are known to be associated with a DNAm signature in blood[24]; however, DNAm patterns in truncating variants in other genomic regions of *SRCAP* have never been examined.

In this study, we investigate the effect of variants located in *SRCAP* proximal or distal to the FLHS locus on DNAm and clinical phenotype. We report that proximal variants are distinguished by an overlapping, but distinct DNAm signature from

FLHS. CpGs in both signatures map to genes relevant to SRCAP molecular function. We also show that individuals with proximal or distal truncating *SRCAP* variants are clinically distinct from FLHS, showing DD/ID, normal stature, hypotonia, behavioral and psychiatric problems, non-specific dysmorphic features, and musculoskeletal issues but lacking the classic FLHS triad.

Subjects and Methods

Cohort recruitment

To characterize the effects of *SRCAP* variants outside of the FLHS locus (NP_006653.2:p.2329-2748), we collected 33 unrelated individuals with truncating variants outside the FLHS locus. We describe the location of the variants based on their position relative to the FLHS locus: 28 individuals had truncating variants upstream of the 33rd exon (further referred as proximal *SRCAP* group) and 5 individuals had truncating variants downstream of the locus and the first AT-hook domain (further referred as distal *SRCAP* group). We focused on cases with truncating variants because *SRCAP* is intolerant only to loss-of-function variants (pLI=1) but not to missense variants (Z=2.1). The individuals with *SRCAP* truncating variants outside the FLHS locus were identified and recruited through a collaborative network of research and diagnostic centers, the Dutch Genome Diagnostic Laboratories (VKGL) variant sharing database25, and by using MSSNG[26], GeneMatcher[27], DECIPHER[28], and the Simons Simplex Collection (SSC)[29]. Clinical and molecular data were provided by the individuals' clinicians which were compared to data previously described in the literature from cohorts of individuals with FLHS6-8. This study was approved by the institutional review board 'Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen' under number 2011/188.

Variant identification

The *SRCAP* variants were identified by sequencing of large gene panels or whole exome sequencing in diagnostic settings in clinical laboratories or in research settings. The sequencing data analysis was performed as described previously[30-39]. Variants in *SRCAP* were annotated using the GRCh37 reference genome and NM_006662.3 transcript. For all individuals, a truncating *SRCAP* variant (25/33 *de novo*) was considered to be the most likely cause of the individuals' phenotype. A summary of other molecular findings are provided in **Table S1**. Variants are reported in ClinVar (accession SCV001477310 - SCV001477340).

DNAm research participants

Informed consent for methylation analysis was obtained from all research participants according to the protocol approved by the Research Ethics Board of SickKids hospital (REB# 1000038847). DNAm analysis was performed using a subset of case samples and age- and sex-matched typically developing controls (**Table S2**). The non-FLHS cases were recruited as described above. Additional samples (FLHS, Rubinstein-Taybi, and Menke-Hennekam syndrome individuals) used to compare the signature were obtained from the SickKids diagnostic lab. Additionally, DNA samples obtained from individuals with ID or multiple congenital anomalies and various missense variants in *SRCAP* (n=4) were included (**Table S2**). Control samples were obtained from the Province of Ontario Neurodevelopmental Disorders (POND) Network, The Hospital for Sick Children, and The University of Michigan (Dr. Gregory Hanna)[40]. "Typically developing" was defined as healthy and developmentally normal by using formal cognitive/behavioral assessments (samples from POND and The University of Michigan) or via physician/parental screening questionnaires (Hospital for Sick Children). Blood DNA samples were available for eight individuals with FLHS, nine with proximal *SRCAP* variants, and all five individuals with distal *SRCAP* variants. Case samples harboring *SRCAP* variants were split into discovery (n=4/8 FLHS group and n=5/9 proximal *SRCAP* group) and validation cohorts (**Table S2**). Due to low sample numbers, a DNAm signature was not defined for the distal *SRCAP* samples.

DNAm microarray data processing

DNAm microarray data processing was performed as pervious described[19,41]. Briefly, whole blood DNA samples were bisulfite converted using the EpiTect Bisulfite Kit (EpiTect PLUS Bisulfite Kit, QIAGEN). The sodium bisulfite converted DNA was then hybridized to the Illumina Infinium Human MethylationEPIC BeadChip to interrogate over 850,000 CpG sites in the human genome at The Center for Applied Genomics (TCAG), SickKids Research Institute, Toronto, Ontario, Canada. Sample groups were run in five batches, with balanced cases and controls in each batch and on each chip, randomly assigned a chip position. The *minfi* Bioconductor package in *R* was used to preprocess data including quality control, Illumina normalization and background subtraction, followed by extraction of β values. Standard quality control metrics in *minfi* were used, including median intensity QC plots, density plots, and control probe plots: one FLHS sample (EX0537) and one distal *SRCAP* sample (distal *SRCAP* #5) had low median channel intensity values than recommended by *minfi* standards, so neither was used for signature derivation. Probes with detection flaws (n=644), probes near SNPs with minor allele frequencies above 1% (n=29,958), cross-reactive probes[20] (n=41,975), probes with

raw beta= 0 or 1 in > 0.25 % of samples (n=16), non-CpG probes (n=2,925), X and Y chromosome probes (n=57,969) were removed, so a total of n=774,580 probes remained for differential methylation analysis.

DNA methylation signatures

To assess DNAm patterns, we identified differentially methylated sites in whole-blood DNA: one for FLHS variants and one for the proximal *SRCAP* variants, by comparing samples from cases to typically developing controls. We were not able to generate a signature for the distal *SRCAP* cases due to the low sample number, so we opted to classify these samples using the FLHS and proximal *SRCAP* DNAm signatures. For all samples, we applied the blood cell-type proportion estimation tool in *minfi* based on Illumina EPIC array data from FACS sorted blood cells[42]. As there is a substantial effect of age on DNAm[43], we used only DNA samples from cases and controls older than 18 months of age to generate each signature. To match the age and sex distributions of each case group, a distinct (but overlapping) control cohort was used for each comparison (**Table S2**). We identified differentially methylated sites using *limma* regression covaried by age, sex, batch, and five of the six predicted cell types (i.e. all but neutrophils). For the FLHS DNAm signature, we compared cases with genetically and clinically confirmed FLHS (n=4) with matched control samples (n=35). This identified 464 probes with a Benjamini-Hochberg adjusted *p*-value<0.05 and a $|\Delta\beta|>0.20$. For the proximal *SRCAP* signature, we compared cases with a proximal truncating variant (n=5) with matched control samples (n=32). This identified 347 probes with a Benjamini-Hochberg adjusted *p*-value<0.05 and a $|\Delta\beta|>0.20$ (20% methylation difference).

Machine learning classification models

We developed two machine learning models: one using each DNAm signature. Using the *R* package *caret*, CpG sites with correlations equal to or greater than 90% to other signature CpGs were removed as we previously described[18]. This led to a set of n=175 non-redundant CpGs from the proximal *SRCAP* signature and n=255 from the FLHS signature. Next, we developed two support vector machine (SVM) models with linear kernel trained on the non-redundant CpG sites. Each model was trained using the methylation values for the discovery cases compared to their matched discovery controls i.e. proximal *SRCAP* SVM model: proximal *SRCAP* cases (n=5) vs. controls (n=32), FLHS model: FLHS cases (n=4) vs. controls (n=35). The models were set to "probability" mode to generate SVM scores ranging between 0 and 1 (0-100%), classifying samples as "positive" (score>0.5) or "negative" (score<0.5). To test model specificity, EPIC array data from additional typically developing controls (n=97) were scored. To test model sensitivity, validation samples

(FLHS n=4; proximal *SRCAP* n=4) were classified. We also classified distal *SRCAP* cases (n=5) as well as pathogenic CREB-binding protein (*CREBBP*; [OMIM: 600140]) and E1A-Binding Protein, 300-KD (*EP300*; [OMIM: 602700]) variants from individuals clinically diagnosed with Rubinstein-Taybi syndrome ([OMIM: 180849]; n=10) or Menke-Hennekam syndrome 1 ([OMIM: 618332]; n=1) to assess the specificity of the model because of their clinical similarity to FLHS and known interactions between the SRCAP protein with CREBBP/EP300 proteins.

Gene ontology analysis

The lists of CpG positions comprising each DNAm signature were submitted to GREAT (Genomic Regions Enrichment of Annotations Tool) for gene ontology (GO) enrichment analysis[44]. Enrichment of each GO term within the gene list was calculated using a foreground/background hypergeometric test over genomic regions; using the set of CpG sites after *minfi* probe quality control (n=774,580) as a background set. Overlapping genes were mapped using default GREAT settings with the following exceptions: the cut-off to annotate a CpG as overlapping with a gene ("distal gene mapping" setting) was set to 10 kb, and enriched terms with two or more gene hits were reported.

Quantitative Facial Phenotyping

Individuals' faces were analyzed from provided frontal photos using the hybrid model reported previously[45,46]. This model combines two algorithms used for facial recognition (OpenFace[47] and Clinical Face Phenotype Space[48]) to create a 468-dimensional vector of an individuals' facial features. These vectors are used to calculate the clustering impact factor (CIF) of an analyzed group, which is a measurement of how a group of individuals cluster within a group of controls. The controls used for the analysis are age-, sex-, and ethnicity-matched individuals with ID, as reported before[45]. The Mann-Whitney U test is used to determine if the CIF is statistically significantly higher than expected from a random chance. This was performed for all the individuals with FLHS and proximal *SRCAP* variants, as the number of available individuals with distal *SRCAP* variants was not sufficient for the analysis. A p-value <0.05 was considered significant. Further, we tested whether the facial photos of individuals with proximal and distal *SRCAP* variants clustered with the FLHS individuals or controls.

Results

Study cohort

We identified and collected clinical data from 33 unrelated individuals with truncating *SRCAP* variants outside of the FLHS locus. We defined the boundaries of this locus based on reported FLHS-causing variants, from the most proximal variant located in the 33rd exon (NM_006662.3:c.6985C>T p.(Arg2329*)) to the most distal variant, located in the 34th exon upstream to the AT-hooks (NM_006662.3:c.8242C>T p.(Arg2748*); **Figure 1**). Genetic investigations were performed for these individuals based on a presentation of neurodevelopmental and/or musculoskeletal issues. None of these individuals had a clinical diagnosis of FLHS. We identified 28 individuals with truncating variants upstream of the 33rd exon (proximal *SRCAP* group) and 5 with truncating variants downstream of the FLHS locus and the first AT-hook domain (distal *SRCAP* group; **Figure 1**). Most of the variants were *de novo* (23/28 proximal and 2/5 distal *SRCAP*). In two individuals, a proximal *SRCAP* variant was inherited from a healthy parent who was mosaic for the variant and for three individuals in each group, inheritance could not be established because one or both parental samples were unavailable. Unlike FLHS, for which two recurrent variants have been identified in the majority of cases (NM_006662.3:c.7303C>T p.(Arg2435*) or NM_006662.3:c.7330C>T p.(Arg2444*)), almost all of the variants in our cohort (31/33) were unique to each individual. The only recurrent variant was NM_006662.3:c.5633dup p.(Pro1879Thrfs*21). Of the 31 unique truncating variants, 24 are frameshift (including all five distal *SRCAP* variants), five are nonsense and two are splice acceptor-site variants (**Table S1**).

Two distinct DNAm signatures associated with SRCAP

DNAm analysis was performed using a subset of case samples (**Table S2**). Blood DNA samples were available for eight individuals with FLHS, nine with proximal *SRCAP* variants, and all five individuals with distal *SRCAP* variants. Case samples harboring *SRCAP* variants were split into discovery cohorts to define the signatures (n=4 FLHS and n=5 proximal *SRCAP*) and the remaining samples (n=4 FLHS and n=4 proximal *SRCAP*) were used as a validation cohort. We first sought to characterize our cohort using the FLHS DNAm signature to determine if the individuals with different position of truncating *SRCAP* variants had a comparable DNAm profile. To do this, we generated genome-wide DNAm profiles on FLHS cases and analyzed them against matched controls using *limma* regression (**Table S3**). We identified an FLHS DNAm signature of 464 differentially methylated CpG sites ($q<0.05$, $|\Delta\beta|>0.20$). We then characterized the DNAm profile of our proximal and distal *SRCAP* cohorts at the FLHS signature sites using principal components analysis (PCA). None of our

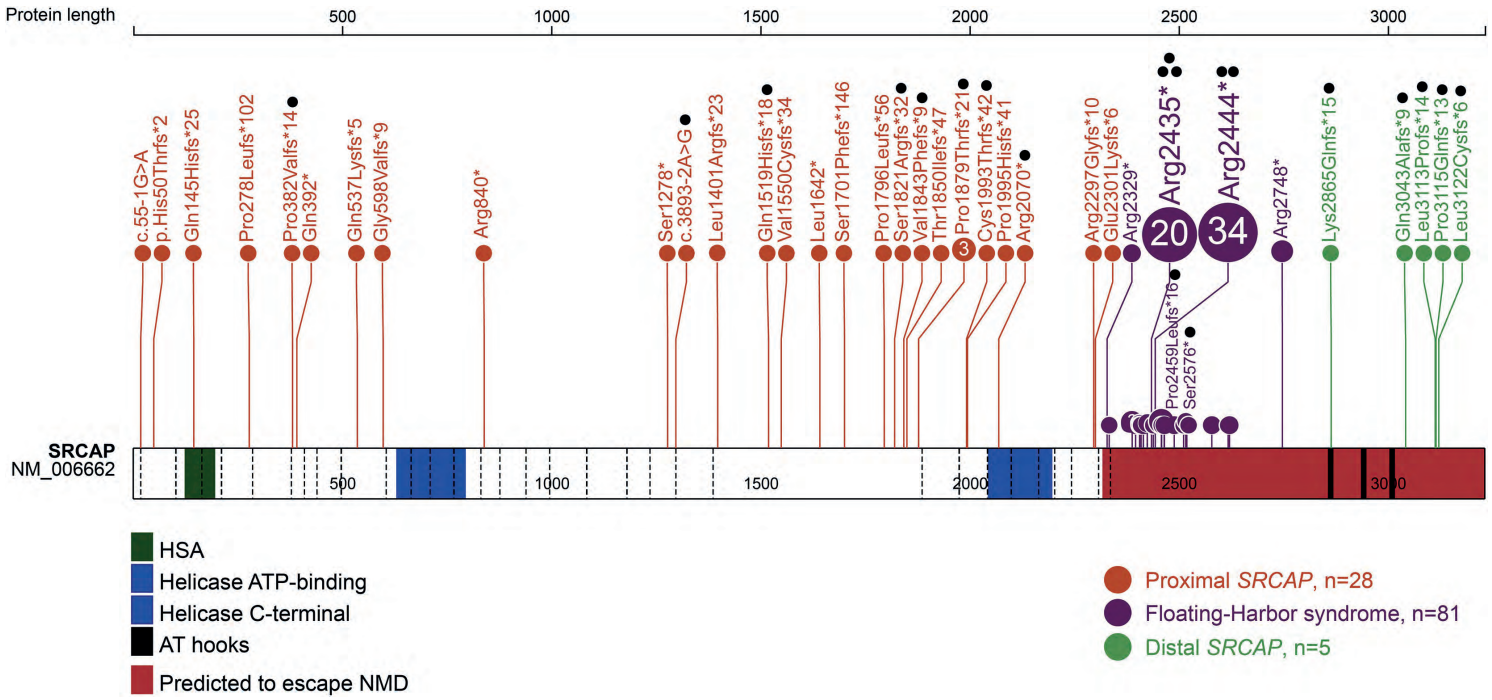


Figure 1. Spectrum and location of the SRCAP truncating variants.

*Schematic representation of the SRCAP protein (NP_006653.2), its functional domains, and variants used in this study. Exon structure, based on NM_006662.3 is provided by dashed lines. Green: HSA-domain (124-196), Blue: Helicase ATP-binding (630-795aa) and C-terminal (2044-2197aa), Black: AT-hooks (2857-2869aa; 2936-2948aa; 3004-3016aa). Locus, predicted to escape NMD (<55bp from the last exon/intron junction), is shown in red. Proximal and distal truncating SRCAP variants identified in this study are shown in orange and green, respectively. Floating-Harbor syndrome-causing variants are depicted in purple (recurrent and the most distant variants are specified). Black dots indicate samples used for DNA methylation analysis.*Two distinct DNAm signatures associated with *SRCAP*

non-FLHS *SRCAP* cohort clustered with FLHS cases using DNAm values at these FLHS signature sites (**Figure 2A**). Notably, all tested proximal SRCAP cases (n=9) and two of the five distal *SRCAP* cases clustered together, and intermediately between the FLHS cases and controls, while the remaining three distal *SRCAP* cases clustered with controls (**Figure 2A**). The intermediate classification of the two cases was also evident using hierarchical clustering (**Figure 2B**). This intermediate profile remained when we clustered all samples using CpG sites made available from a previously published FLHS signature[24] (**Figure S1**). In other disorders of the epigenetic machinery, a case group clustering out from controls using one signature indicates the possible existence of a second overlapping but distinct signature for these cases[18,22,49]. We hypothesized that the intermediate clustering of the proximal *SRCAP* cases was indicative of a unique DNAm signature associated with these cases, overlapping that of FLHS.

To test this hypothesis, we then compared a discovery cohort of proximal *SRCAP* cases (n=5) with age- and sex-matched typically developing controls (n=32) to identify whether proximal *SRCAP* variants are indeed associated with a distinct DNAm signature. From this analysis, we identified a proximal *SRCAP* DNAm signature of 347 differentially methylated CpG sites ($q<0.05$, $|\Delta\beta|>0.20$; **Table S4**). Clustering of all samples at the proximal SRCAP DNAm signature sites showed that all proximal *SRCAP* validation cases (n=4) clustered clearly with proximal *SRCAP* discovery cases, using PCA (**Figure 2C**) and hierarchical clustering (**Figure 2D**). Given the limited sample size and divergent/disparate clustering of the distal *SRCAP* cases, we were not able to generate a signature for this group.

Shared and distinct genomic features of *SRCAP* signatures

We identified a subset of 77 differentially methylated CpG sites shared by the FLHS and proximal *SRCAP* DNAm signatures (**Table S5**). The FLHS signature is composed of both hypo- and hypermethylated sites, with 255/464 (55%) of sites being hypermethylated compared with mean control methylation. In contrast, the proximal *SRCAP* DNAm signature is predominately composed of hypermethylated sites, with 344/347 (99%) of sites having increased DNAm in cases compared with controls. Both DNAm signatures are characterized by clusters of differentially methylated CpGs overlapping the same genomic regions in the same direction of methylation change. Correlated differentially methylated CpGs which are co-localized in regulatory regions have been shown to have direct impacts on gene expression[50]. In the proximal *SRCAP* DNAm signature, 106/347 (30%) signature CpGs are within 100 bp of another signature CpG and 159/347 (46%) are within 1 kb of another signature CpG. In the FLHS signature, 125/464 (27%) signature CpGs are within 100 bp of

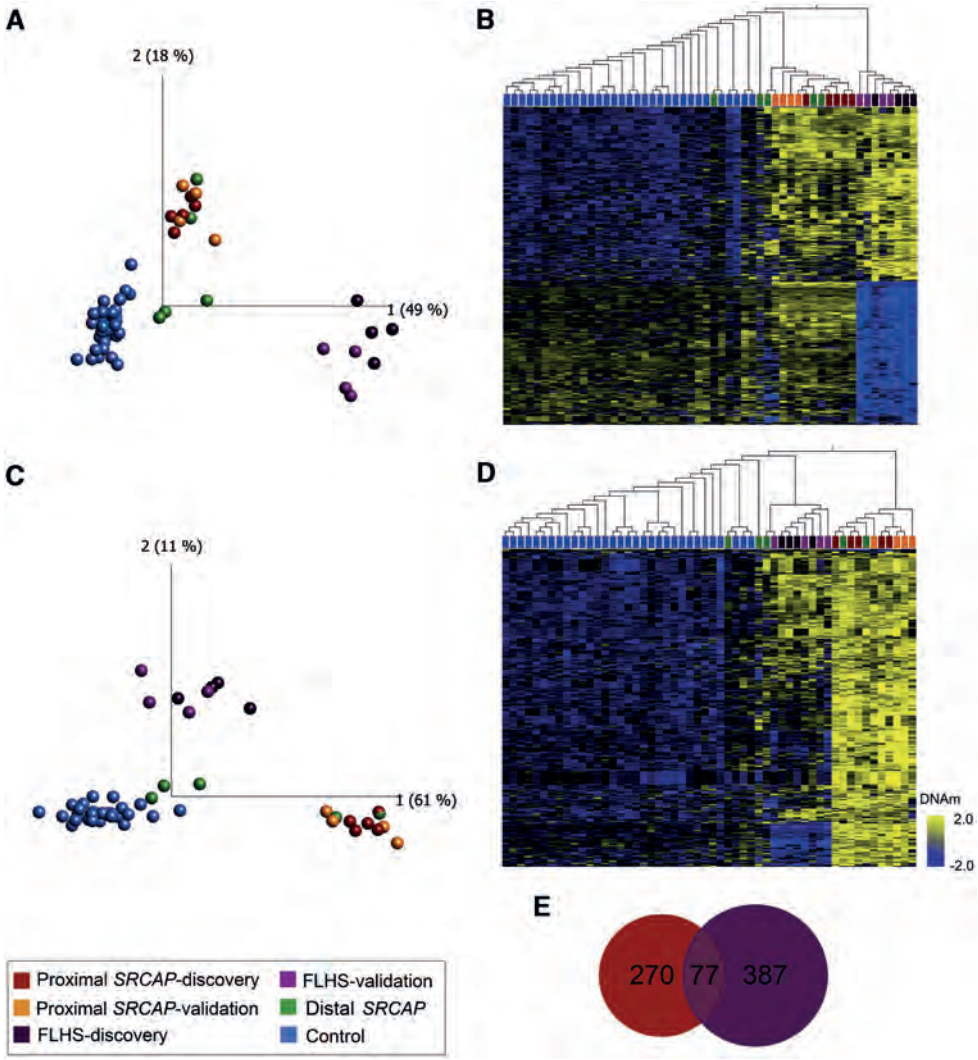


Figure 2. Loss of function variants in *SRCAP* are associated with two distinct but overlapping DNAm signatures.

*FLHS DNAm signature: **A**. Principal components analysis (PCA) and **B**. heatmap showing clustering of FLHS discovery cases (n=4; dark purple), FLHS validation cases (n=4; light purple), proximal SRCAP discovery cases (n=5; dark orange), proximal SRCAP validation cases (n=4; light orange), distal SRCAP cases (green) and discovery controls (n=35; blue) using DNAm values at 464 CpG sites in the FLHS DNAm signatures. FLHS cases clearly segregate from all other samples, while all proximal, and some distal SRCAP cases cluster intermediately. Proximal SRCAP DNAm signature: **C**. PCA and **D**. heatmap showing clustering of the same cases from A and B and matched controls (n=32; blue) using DNAm values at 347 CpG sites in the proximal SRCAP DNAm signature. Proximal SRCAP discovery and validation cases and some distal cases clearly separate from controls, with FLHS cases clustering intermediately. The heatmap color gradient indicates the normalized DNAm value ranging from -2.0 (blue) to 2.0 (yellow). Euclidean distance metric is used in the heatmap clustering dendrograms. **E**. Venn diagram showing the CpG sites are shared and distinct between the proximal SRCAP and FLHS signatures.*

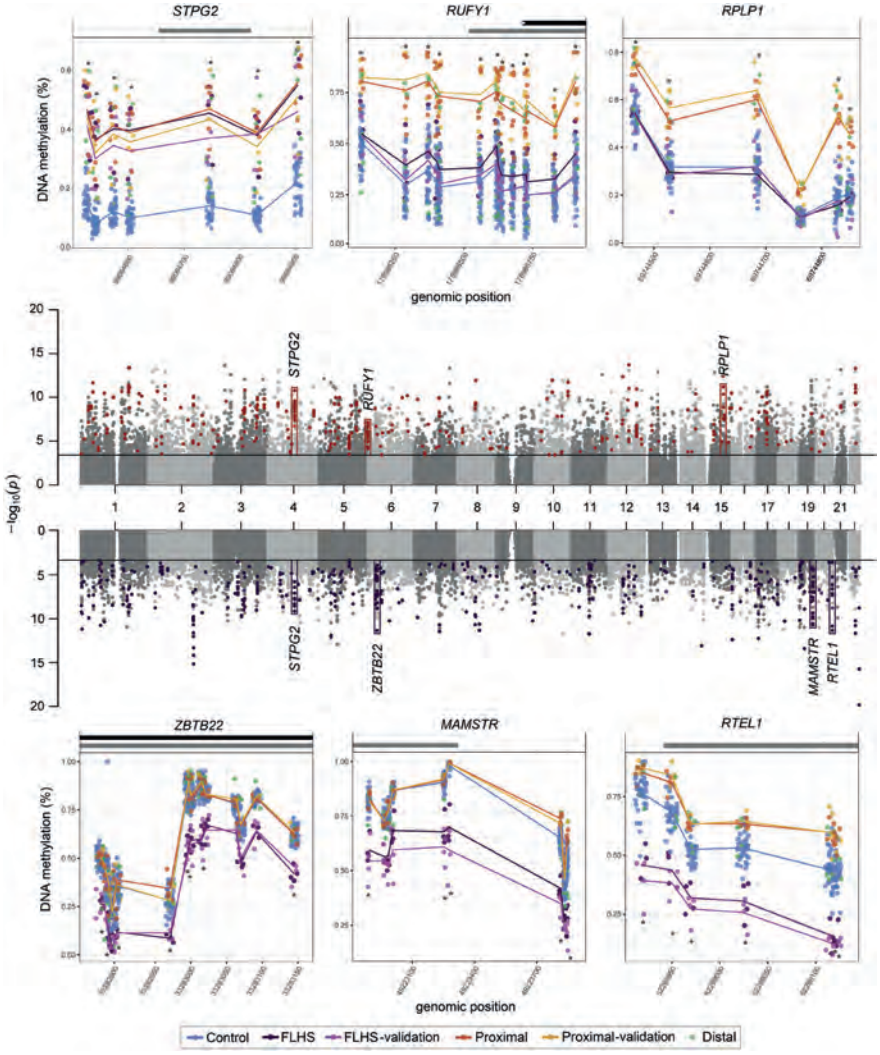


Figure 3. Distribution of differentially methylated regions in *SRCAP* DNAm signatures.

The Manhattan plots (center) show the CpG site p-values for each DNAm signature discovery comparison; sites that meet the |Δβ|>0.20 cut off in the proximal SRCAP signature (upper) are colored orange, those that meet the cut off in the FLHS signature (lower) are colored purple. Six differentially methylated regions are boxed on the Manhattan, with the full plots for sample groups shown above and below. These six illustrate different patterns of DNAm between the proximal SRCAP and FLHS groups. Each is named for the gene body/ promoter to which they map. Grey bars above the plots indicate CpG islands, black represent gene bodies. β values for each sample are indicated with mean lines for groups (except the distal group, since two classified positively and three negatively). The CpGs mapping to STGP2 are present in both signatures, RUFY1 and RPLP1 are in the proximal SRCAP signature, ZBTB2, MAMSTR and RTEL1 are in the FLHS signature. There is increased DNAm at the RTEL1 sites in the Proximal group, though not a large enough Δβ to be in the proximal SRCAP DNAm signature. In all cases, β values in the discovery and validation cohorts are concordant. RPLP1, RUFY1, and RTEL1 are encoded on the plus strand, STPG2, ZBTB22, and MAMSTR are on the minus strand.

another signature CpG and 220/464 (47%) are within 1 kb of another signature CpG. All these neighboring differentially methylated CpGs display the same direction of methylation change between cases and controls (**Table S4**). Furthermore, intervening CpG sites that are not in the signature did meet statistical significance (q<0.05 cutoff) but not the stringent effect size cut-off (|Δβ|>0.20) used in signature derivation. This is illustrated by plotting the β values for six selected examples of these contiguous regions (**Figure 3**). This plot also illustrates the different DNAm patterns in each group. For example, seven probes in a CpG island/enhancer region upstream of *STPG2* show a similar level of increased DNAm in both groups (**Figure 3**). In contrast, differentially methylated CpG sites in the proximal *SRCAP* signature at *RUFY1* (OMIM: 610327) and *RPLP1* (OMIM: 180520) have DNAm levels overlapping controls in FLHS individuals. Finally, hypomethylated *RTEL1* (OMIM:608833) CpG sites in the FLHS signature have increased DNAm in proximal *SRCAP* cases, but these changes do not meet statistical cutoffs for the proximal *SRCAP* signature. In summary, these findings show that there is a more complex relationship between the two signatures than is captured by simply comparing the number of overlapping CpG sites with stringent cut-off criteria.

Machine learning classification of samples using *SRCAP* DNAm signatures

In order to robustly classify each sample, we trained two support vector machine (SVM) models on the DNAm data from the proximal *SRCAP* and FLHS DNAm signatures, respectively (**Figure 4**). We then used these models to classify the remaining samples: FLHS validation (n=4), proximal *SRCAP* validation (n=4), distal *SRCAP* (n=5), and control validation (n=97). We also obtained samples from individuals with pathogenic variants in *CREBBP* and *EP300* to use for classification. Each model generated a probability of pathogenicity score from 0-1 for each sample to which it is applied, with 0.5 being the boundary between a positive and a negative classification (**Table S6, Table S7**). We validated the FLHS SVM model using FLHS validation cases (n=4) which classified positively demonstrating 100% model sensitivity, and additional controls (n=97), all of which classified negatively demonstrating 100% model specificity (**Figure 4A**). The FLHS model also clearly negatively classified all proximal *SRCAP* cases (n=9). To assess the proximal *SRCAP* SVM model sensitivity and specificity, we tested the model using proximal *SRCAP* validation cases (n=4), all of which classified positively, demonstrating 100% model sensitivity, and additional controls (n=97) all of which classified negatively, demonstrating 100% specificity. The proximal *SRCAP* model also clearly classified all FLHS samples negatively (n=8), despite the intermediate clustering described above (**Figure 4B**).

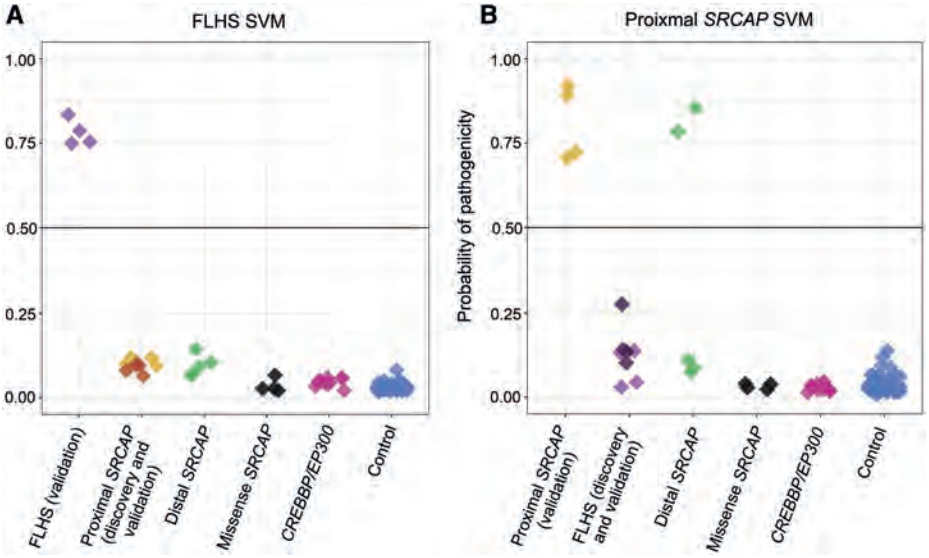


Figure 4. Classification of samples using SVM machine learning models based on each DNAm signature.

*Sample groups were scored using the **A**. FLHS support vector machine (SVM) model and **B**. the proximal SRCAP SVM model. FLHS validation cases (n=4) classified positively using the FLHS model; similarily proximal SRCAP validation cases (n=4) classified positively using the proximal SRCAP model, demonstrating 100% sensitivity of both models. Using the FLHS model, proximal SRCAP cases (n=9) and validation controls (n=97) classified negatively, demonstrating 100% specificity of the model. Using the proximal model, FLHS cases (n=8) and validation controls (n=97) classified negatively demonstrating 100% specificity of the model. SRCAP missense variants (n=4) classified negatively using both models, suggesting them to be benign. Distal SRCAP cases (n=5) all classified negatively on the FLHS signature, suggesting these cases do not have FLHS. Two distal SRCAP cases classified positively on the proximal SRCAP model (distal SRCAP individual #1 and #2) demonstrating concordant DNAm profiles of these cases with the proximal SRCAP cases, while three classified negatively (distal SRCAP individual #3, #4, and #5). Cases with a pathogenic variant in CREBBP (n=10) or EP300 (n=1) all classified negatively using both models.*

Next, we classified all distal *SRCAP* cases (n=5) using both signature models. Using the FLHS SVM model, all five distal *SRCAP* cases classified negatively, demonstrating that these cases are not FLHS (**Figure 4**). Using the proximal *SRCAP* SVM model, two distal *SRCAP* cases classified positively and three classified negatively (**Figure 4**). This demonstrates that these two distal *SRCAP* cases have the same DNAm profile as the proximal *SRCAP* cases at the proximal *SRCAP* DNAm signature sites, possibly indicating that these variants result in the same disorder as the proximal *SRCAP* variants. Clinical features of the distal *SRCAP* group are non-specific and do not differentiate the positive from negative classifying cases (see below clinical features section). A negative classification typically indicates a benign variant for the condition tested, although it is possible that these variants have another signature distinct from the individuals with proximal *SRCAP* variants and FLHS. Additionally,

all four (**Table S2**) obtained samples from individuals with different *SRCAP* missense variants with intellectual disability or multiple congenital anomalies classified negatively using both models (**Figure 4**). Finally, we classified pathogenic *CREBBP* and *EP300* variants from individuals clinically diagnosed with Rubinstein-Taybi syndrome (n=10) or Menke-Hennekam syndrome 1 (n=1), given the similar clinical features to FLHS and the known interactions of the SRCAP complex and CREBBP/EP300; all samples classified negatively using both models (**Figure 4**).

Gene ontology of *SRCAP* DNAm signatures

Finally, we assessed the ontology of the genes overlapping the CpGs in each signature using GREAT[44]. There were 124 genes overlapping the proximal *SRCAP* DNAm signature and 148 for the FLHS signature. Both gene lists were characterized by multiple enriched GO terms related to regulation of chromosome structure and DNA repair (**Tables S8-S13**) which are related to SRCAP molecular function. The top biological process for the proximal *SRCAP* DNAm signature sites is "DNA recombination" (**Table S8**). Of the 22 CpGs in this term, 11 are within 400 bp of the transcriptional start site of *EID3* (EP300 Interacting Inhibitor of Differentiation 3; [OMIM: 612986]). *EID3* encodes a transcriptional repressor that is predicted to function by interfering with CREBBP-dependent transcription factors[51]. Several of the top terms were related to other DNA metabolism processes like translation (**Table S8)**. For the FLHS signature, several similar terms were identified but with different genes implicated. The top biological processes for the FLHS group were related to regulation of telomeres and neural development (**Table S11**), largely due to six CpGs in the promoter of *RTEL1* (Regulator of Telomere Elongation Helicase 1; [OMIM: 608833]). In summary, both signatures map to distinct genes related to SRCAP molecular function.

SRCAP truncating variants proximal to the FLHS locus result in a neurodevelopmental disorder with clinical features distinct from FLHS

We undertook a detailed clinical characterization of our cohort (n=33; **Table S1**). We considered the individuals with proximal *SRCAP* variants (n=28) to be a distinct group, based on the DNAm results, which we refer to as non-FLHS *SRCAP*-related NDD. We compared features of the individuals with the non-FLHS *SRCAP*-related NDD with those with FLHS diagnosis reported in literature and individuals with the distal *SRCAP* variants (**Table 1**).

Table 1. Phenotype comparison between the non-FLHS *SRCAP*-related NDD (proximal *SRCAP*) and distal *SRCAP* groups with the Floating-Harbor syndrome reported in the literature.

Features	**Literature (reported/ observed)**				**This study (reported/ observed)**	
References	FLHS features reported by Le Goff et al., 2013[7]	FLHS features reported by Nikkel et al., 2013[8]	FLHS features reported by Seifert et al., 2014[6]	Total frequency of FLHS features	Non-FLHS *SRCAP*-related NDD	Distal *SRCAP* variants
Individuals included	N=6	N=52	N=5	N=63	N=28	N=5
FLHS facial gestalt	6/6	52/52	5/5	63/63 (100%)	0/28 (0%)	1/5 (20%)
Delayed bone age	6/6	23/25	4/4	33/35 (94%)	0/25 (0%)	0/5 (0%)
Short stature (< -2SD)	6/6	41/52	5/5	52/63 (83%)	0/25 (0%)	0/5 (0%)
HC < -2SD	3/6	9/43	0/5	12/54 (22%)	1/24 (4%)	0/5 (0%)
Macrocephaly	0/6	0/43	0/5	0/54 (0%)	2/24 (8%)	0/5 (0%)
Speech delay	6/6	52/52	5/5	63/63 (100%)	24/25 (96%)	4/5 (80%)
ID /Borderline IQ/ Special education	1/6	37/41	4/5	42/52 (81%)	17/24 (71%)	5/5 (100%)
Schizophrenia/ Psychoses	NR	NR	NR	Not typical	4/25 (16%)	0/5 (0%)
Behavioral problems	NR	9/32	3/5	8/30 (27%)	16/25 (64%)	2/5 (40%)
ASD	NR	NR	NR	Not typical	10/24 (42%)	2/5 (40%)
Seizures	1/6	6/52	0/5	7/63 (11%)	3/27 (11%)	0/5 (0%)
Joint hypermobility/ Musculoskeletal problems	NR	NR	NR	Not typical	13/27 (48%)	4/5 (80%)
Hypotonia	NR	NR	NR	Not typical	16/25 (64%)	2/5 (40%)
Broad thumbs, broad fingertips, brachydactyly	5/6	10/17	5/5	20/28 (72%)	0/28 (0%)	0/5 (0%)
Hearing loss	NR	9/52	2/5	9/52 (17%)	3/26 (12%)	0/5 (0%)
Myopia/ Hypermetropia	NR	5/43	NR	5/43 (12%)	11/26 (42%)	2/5 (40%)
Strabismus	NR	7/43	NR	7/43 (16%)	3/25 (12%)	0/5 (0%)
Cryptorchidism	0/2	5/24	NR	5/26 (19%)	0/13 (0%)	1/4 (25%)
Genitourinary malformations	1/6	7/52	0/5	8/63 (13%)	4/28 (14%)	2/5 (40%)
High pitched voice	NR	8/11	NR	8/11 (73%)	0/26 (0%)	0/5 (0%)

NR = not reported, FLHS = Floating-Harbor syndrome; HC = head circumference; ID = intellectual disability; ASD = autism spectrum disorder.

Some of the features were no assessed, or not possible to assess in all patients, so the frequency of the features is reported among the number of individuals in whom a feature was evaluated. Most of the individuals presented with neurodevelopmental and behavioral issues. Speech and motor delay were reported in 24/25 individuals, and ID was reported in 13/24 individuals. Among individuals with ID, mostly mild ID was reported although there were a few individuals with moderate or severe ID. Additionally, learning difficulties were reported in four individuals with normal IQ. Autism spectrum disorder (ASD) was present in approximately half of the individuals with the non-FLHS *SRCAP*-related NDD (10/24). Although DD/ID are also common among individuals with FLHS, ASD is not typically reported as a feature of FLHS (**Table 1**). Additionally, in the non-FLHS *SRCAP*-related NDD group, 16/25 individuals had various behavioral problems other than ASD including challenging behavior, anger, anxiety, attention deficit and hyperactivity disorder (ADHD). Tics, including Tourette's syndrome, and psychoses/schizophrenia were each reported in four non-FLHS *SRCAP*-related NDD patients.

Unlike individuals with FLHS, the individuals with non-FLHS *SRCAP*-related NDD have normal or tall stature and do not have brachydactyly, broad thumbs, or fingertips. A range of other non-specific skeletal and connective tissue features not typical for FLHS patients were commonly present. Joint hypermobility is reported in 7/27 individuals while chronic musculoskeletal pain was reported in three adult individuals (proximal *SRCAP* individuals #10, #11 and #16). In fact, two individuals (proximal *SRCAP* individuals #10 and #16) were evaluated by rheumatologists regarding the chronic pain without a conclusive diagnosis. As pain is present only in adults, it may develop gradually over time. Additionally, features such as *pectus excavatum* or *carinatum* were reported in five individuals, and scoliosis in three. Notably, these issues were one of the main reasons for the genetic investigation of three individuals.

In general, the affected organ systems and severity of the non-FLHS *SRCAP* group phenotype is variable. For example, seizures, and genitourinary anomalies were each reported in 3/27 and 3/28 individuals, respectively. Noticeably, three individuals were diagnosed at the neonatal or infant age. At infant age, hypotonia, gastro-esophageal reflux disease, and tracheo/laryngomalacia were reported.

Although the number of individuals with distal *SRCAP* truncating variants is small (n=5), they seem to have a similar phenotype to the individuals with the non-FLHS *SRCAP*-related NDD harboring proximal *SRCAP* truncating variants. They have developmental delay with mild ID reported in 4/5 individuals. ASD and other behavioral problems were reported in 3/5 and 2/5 individuals, respectively.

Additionally, musculoskeletal issues (scoliosis, pectus anomalies, joint hypermobility and pain) were reported in 4/5 individuals. Importantly, none of the individuals with distal *SRCAP* variant have short stature and delayed bone age and other typical FLHS features. It is impossible to clinically distinguish the two individuals with distal *SRCAP* variants who classified positively on the proximal *SRCAP* DNAm signature (distal *SRCAP* individual #1 and #2) from the three that classified negatively, because the phenotype is non-specific.

Craniofacial dysmorphic features

Facial photos of eight individuals with proximal *SRCAP* variants and three individuals with distal *SRCAP* variants (one proximal *SRCAP* DNAm signature positive and two negative) are shown (**Figure 5)**. Most individuals presented with a long face and long philtrum, prominent forehead and thin vermilion upper lip with everted lower vermilion and wide mouth. Other typical features were narrow) palpebral fissures, epicanthal folds, periorbital fullness, wide nasal bridge, prominent ears, as well as retro- or prognathia. Despite the fact that most individuals presented with some dysmorphic features, these are nonspecific and variable. Notably, none of the individuals that were positive on the proximal *SRCAP* DNAm signature have a facial gestalt characteristic of FLHS.

To objectively evaluate the facial phenotype similarities and differences, we compared clinical photos of 16 individuals with FLHS from the literature and, 14 proximal and individuals with age, sex and ethnicity matched control individuals with various non-specific NDDs (one matched individual per tested individual) by utilizing a facial feature recognition algorithm as described previously[46]. First, we validated that individuals with FLHS cluster together (p=0.001) and have a significantly different gestalt than the general NDD cohort, thus confirming that FLHS has a specific and recognizable gestalt. Next, we compared individuals with proximal (n=14) *SRCAP* variants with the NDD controls. The analysis was not able to identify specific facial features for the group and did not discriminate cases from controls (p=0.698). This supports our observation that the individuals with proximal *SRCAP* variants do not have a typical gestalt. As the number of the distal *SRCAP* individuals was not sufficient (n=3), we were not able to test whether these individuals have different facial gestalt from the controls. Lastly, we compared individuals with proximal and distal *SRCAP* variants with the individuals with FLHS diagnosis. One individual with distal *SRCAP* variant (distal *SRCAP* individual #4) was classified as FLHS. Indeed, this individual has facial features suggestive of with FLHS, such as a triangular face with prominent forehead and prominent nose, but he also presented with a Marfanoid habitus and does not have other typical FLHS

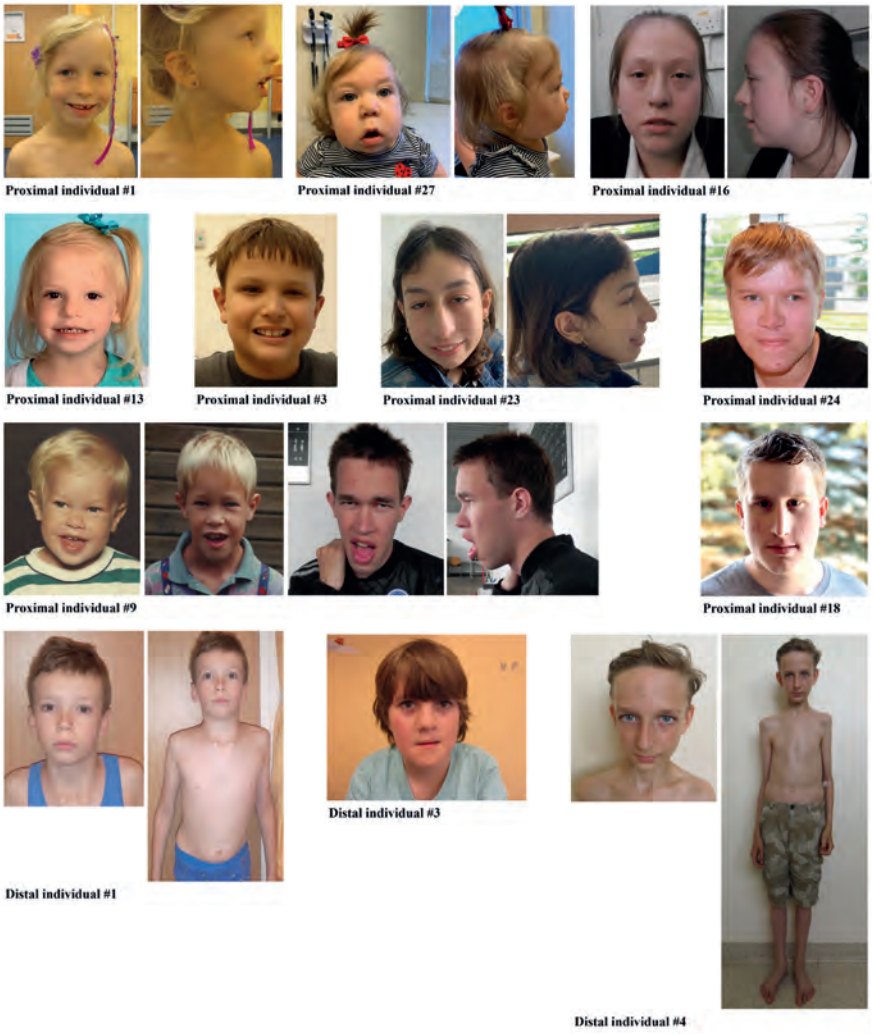


Figure 5. Facial features of individuals with proximal and distal truncating *SRCAP* variants.

Phenotype of nine individuals with the proximal and three individuals with the distal truncating SRCAP gene variants. Photos at age of 2, 8 and 25 years are available for the proximal SRCAP individual #9. Shared facial (non-specific) phenotypic features of the proximal SRCAP group individuals are seen: long face and long, wide philtrum, prominent forehead, thin upper lip vermilion and everted lower vermilion, wide mouth, typical (narrow) palpebral fissures, epicanthal folds, periorbital fullness, wide nasal bridge, prominent ears, and retro- or prognathia. Distal SRCAP individual #4 has some FLHS facial features although he does not have short stature or other typical FLHS features and has Marfanoid habitus with pectus excavatum (similarly to distal SRCAP individual #1).

features. Therefore, we did not classify this individual as FLHS. Surprisingly, one individual with the proximal *SRCAP* variant (proximal *SRCAP* individual #18) was also classified by the tool as FLHS, though our clinical observation, both regarding facial and other features, was not in agreement with this. Although facial feature recognition tools can be useful, this shows that they are not always well suited as a diagnostic tools in isolation, and should be used in conjunction with other tools and information.

Since the clinical and dysmorphic features of patients with proximal *SRCAP* truncating *SRCAP* variants are various and rather non-specific, it is also not possible to phenotypically distinguish proximal *SRCAP* individuals from the distal *SRCAP* ones (regardless of on the proximal *SRCAP* DNAm signature status).

Discussion

Truncating variants clustering within the last two exons of the *SRCAP* gene are a known cause of FLHS, but the molecular and phenotypic consequences of variants outside this locus have not been reported. In this study, we show that truncating variants proximal to the FLHS locus result in a non-FLHS *SRCAP*-related NDD with behavioral and psychiatric problems, non-specific dysmorphic features, musculoskeletal problems, hypotonia. In addition, these individuals demonstrate a specific DNAm signature that can be used to positively identify affected individuals and distinguish them from FLHS. These findings support the non-FLHS *SRCAP*-related NDD as representing a distinct disorder from FLHS. We also show that truncating variants located distally to the known FLHS-locus do not cause Floating-Harbor syndrome, and they seem to result in an NDD different from FLHS, but further collection of additional cases is needed to confirm this.

FLHS is characterized clinically by a typical facial gestalt, short stature with delayed bone age and developmental delay (especially expressive speech delay) with or without mild to moderate ID[8], and characterized molecularly by a specific DNAm signature[24]. We found that none of the 14 tested individuals with a truncating variant outside of the FLHS locus in our cohort with DNAm assessed were positive for the FLHS DNAm signature. This led us to identify a distinct DNAm signature in our cases with proximal *SRCAP* truncating variants, which demonstrates some overlap with the FLHS signature. The SVM model derived from this signature provided clear positive or negative classifications for all samples (**Figure 3**). This model positively classified all the proximal *SRCAP* cases, and 2/5 distal *SRCAP*

cases, while none of them were positively classified using our FLHS SVM model. These clear binary categories (positive or negative) demonstrate very different DNAm profiles for each condition, and strongly suggest they are distinct disorders. Importantly, a sample from the SSC cohort (13857.p1) with a proximal *SRCAP* truncating variant (NM_006662.3:c.6409_6419del p.(Asp2137GlufsTer25)) but limited clinical information, classified positively using the proximal *SRCAP* SVM model and negative on the FLHS model. Without detailed clinical examination, it is not possible to confirm the diagnosis, but given the recruitment criteria for SSC (children 4-18 years old with confirmed ASD and without severe neurological deficit and with negative ASD family history[29]) and DNAm classification we expect that this individual has the non-FLHS *SRCAP*-related NDD. We also found that individuals with pathogenic *CREBBP*, and *EP300* variants classified negatively using both (the proximal *SRCAP* and FLHS) models, demonstrating the ability of these models to discriminate the non-FLHS *SRCAP*-related NDD, and FLHS, from clinically and molecularly related disorders.

We found that none of the individuals in our cohort have the typical FLHS clinical features. Although DD/ID are seen in both FLHS and our study cohort, the individuals from our cohort (both with proximal *SRCAP* and distal *SRCAP* truncating variants) do not have a recognizable facial gestalt, nor short stature with delayed bone age and therefore, are clinically distinguishable from individuals with FLHS. Moreover, while uncommon among individuals with FLHS, the non-FLHS *SRCAP*-related NDD individuals commonly have mild to severe psychiatric and behavioral issues, reported as one of the main challenges for these individuals and their families. Additionally, a high proportion of these individuals have various musculoskeletal problems (e.g. pectus anomalies and scoliosis) as well as joint hypermobility/instability reported at younger age, and joint pain presenting in adulthood. Coupled with the DNAm data, these clinical differences support the view that the non-FLHS *SRCAP*-related NDD and FLHS are distinct conditions.

SRCAP is a chromatin remodeler, and it activates transcription of various genes by depositing the H2A.Z histone in the promoter regions, targeting ~10% of promoters[52]. Deposition of H2A.Z is an important step for the DNA-end resection required for the repair[53]. Both sets (proximal *SRCAP* and FLHS) of the identified signature CpGs map to sets of genes relevant to SRCAP functions. Both were enriched for GO terms related to chromosome structure and DNA repair. “DNA recombination” was the top biological processes hit for the proximal *SRCAP* signature, in part due to the gene *EID3*, a transcriptional repressor expected to interfere with CREBBP-dependent transcription factors, acting in opposition

to SRCAP[51]. Hypermethylation and, therefore, possible silencing of this gene in the context of *SRCAP* haploinsufficiency may further impair CREB-mediated transcription. The FLHS signature was enriched for different genes regulating chromosome structure including *RTEL1*, which encodes a telomeric DNA helicase which appears to be important during early brain development[54]. Some DNA-replication-regulating genes were present in both signatures, e.g. *BRCA1*. Clusters of signature CpGs have been observed in other signatures, but the number and length in the two *SRCAP*-associated signatures is notable. The DNAm values of CpGs exclusive to each signature can demonstrate DNAm values overlapping controls in the other signature, (e.g. *RPLP1*, *MAMSTR*) while some can demonstrate values distinct from controls (e.g. *RTEL1*; **Figure 4**).

We show that the position of truncating variants in the *SRCAP* gene determines the phenotype, which likely occurs via different molecular mechanisms. Based on the available evidence, we hypothesize that the phenotype of the individuals with proximal *SRCAP* variants could be explained by haploinsufficiency. First, all reported variants do not cluster at any specific domain and spread from intron 3 to exon 32 and, therefore, they are expected to undergo nonsense-mediate decay (NMD) or to result in a loss of a significant functional part of the protein, if escaping NMD. Secondly, the gene is intolerant to the loss-of-function variants (pLI=1), which is suggestive of the gene being haploinsufficient[55]. In fact, there are 11 rare LoF variant in gnomAD v2.1.1., but 7/11 variants have skewed allele balance (20-35%), which suggests somatic origin of the variants, and additional two frameshift variants seems to be a single complex indel. Therefore, this confirms that individuals with truncating *SRCAP* variants are depleted from the population database of adults without severe pediatric disorders. Third, Gerundino *et al.* reported a case with overlapping features with our cohort with a *de novo* 186kb 16p11.2 microdeletion that encompasses *SRCAP* and eight other genes[56]. This individual demonstrates a partial clinical overlap with our proximal *SRCAP* cohort including facial features, global developmental delay and normal IQ, behavioral problems (ADHD, inadequate social skills) although the individual's features of microcephaly and short stature do not overlap[56]; however, these additional features could be the result of the other gene deletion in the region. There is currently no evidence that *SRCAP* missense variants can be pathogenic (based on the low Z score, as well as different phenotype and negative DNAm data of the tested individuals); however, more variants should be tested to reach a definitive conclusion. Based on all these data, we hypothesize that truncating variants upstream of the FLHS locus cause a distinct NDD via *SRCAP* haploinsufficiency. This would make FLHS and the non-FLHS *SRCAP*-related NDD analogous to Marshall-Smith syndrome (OMIM: 602535) caused by variants in *NFIX*

(OMIM: 164005) that escape NMD, and Sotos syndrome 2 (OMIM: 614753) caused by variants leading to *NFIX* haploinsufficiency[57]. These conditions have distinct clinical features and are not considered part of a single disorder spectrum.

Most of the FLHS-causing variants are recurrent stop-gain variants which cluster at the 3'-end of the gene predicted to result in a truncated protein that lacks AT-hooks by escaping nonsense-mediated mRNA decay[8]. The AT-hooks are necessary for the direct DNA-binding by epigenetic regulators[58]. It has been shown that the AT-hooks can also serve as nuclear localization signals[59]. Therefore, it is hypothesized that FLHS is caused by the (trans) dominant negative effect (i.e., antimorph) of the truncated SRCAP protein which competes with the wild-type protein to form the SRCAP complex, which can result in a mislocalization of the complex[3,4]. However, there are currently no functional data available supporting this hypothesis. For the *MECP2* gene, truncating variants resulting in a complete loss of the second AT-hook do not cause mislocalization of the protein but rather an impaired chromatin binding and altered chromatin conformation, while only mild reduction of activity was shown for a truncating variant downstream to the AT-hook[60]. Therefore, we propose that a gain-of-function (i.e., neomorph) mechanism should be investigated as the FLHS causal mechanism.

Various mechanisms of distal truncating variant pathogenesis are possible. Our DNAm data suggest that different distal *SRCAP* variants have differing effects on the protein function and resulting phenotype. At this time, the DNAm data cannot definitively determine whether the two positive distal *SRCAP* individuals have the same disorder as the individuals with proximal *SRCAP* variants. The phenotype of the individuals with the distal truncating *SRCAP* variants (both positive and negative on the proximal *SRCAP* DNAm signature) is similar to the individuals with the proximal *SRCAP* variants; however, the overlapping clinical features cannot be used to conclude whether individuals with proximal and distal truncating *SRCAP* variants are affected with the same disorder because the phenotype of the non-FLHS *SRCAP*-related NDD is non-specific. Identification of a distal *SRCAP* DNAm signature that positively classifies the proximal *SRCAP* variants will be necessary to determine if they truly share the same signature and are the same disorder. The three individuals with distal *SRCAP* variants who classified negatively on the proximal *SRCAP* DNAm signature, are not phenotypically distinct from other non-FLHS cases. Genetically, there is a notable pattern: the three negative cases are the most distal, i.e. nearest the end of the gene, while the two positive cases are closer to the FLHS region and AT-hook domains. It may be that the three negative distal *SRCAP* frameshift variants escape NMD and, therefore, with the AT-hooks intact,

result in a functional protein[12]. Unfortunately, no suitable three-dimensional SRCAP complex structure is currently available to evaluate the role of the distal part of the protein. If a functional protein is produced, the clinical features seen in these individuals might be caused by another yet unknown variant. For all these reasons, we currently classify these variants as VUS.

In addition to the utility of the DNAm data described here to discriminate between *SRCAP*-related conditions, they demonstrate the robustness and complexities of DNAm signatures. Previous work has suggested that there may be genotype-epigenotype-phenotype correlations for disorders of the epigenetic machinery, i.e. differences in variant location, DNAm signature, and clinical phenotype are mirrored in each other. The degree of overlap between signatures can reflect the degree of clinical overlap between conditions, demonstrating epigenotype-phenotype correlations[14,18,19,23]. Genotype-epigenotype correlations have also been reported in ADNP syndrome[61], with variant location reflecting changes in DNAm signatures; however, this study did not find any correlation with differences in clinical features. Recent work in *SMARCA2* shows correlations between variant location, DNAm signature and clinical phenotype[12]. The work presented here illustrates a clear genotype-epigenotype-phenotype correlation in one gene: different truncating variant locations within *SRCAP* are associated with distinct clinical presentations and DNAm signatures.

Similar to other recently described novel NDDs, the phenotype of the individuals with non-FLHS *SRCAP* pathogenic variants is non-specific and clinically is not recognizable[12,62]. Thus, shared typical phenotypic features, historically used as the basis of the novel syndrome discoveries, cannot be used as the main evidence that individuals are indeed affected by the same disorder. DNA methylation signatures can provide clarity in these cases, demonstrating, as we did here for *SRCAP*, that individuals with proximal *SRCAP* truncating variants have the same condition, which is distinct from FLHS. The DNAm data also raise questions for further study, namely the pathogenicity of the distal *SRCAP* variants. This work illustrates that the partnership of comprehensive clinical assessment and DNAm signature research have great power both to identify new conditions and to utilize molecular data to support syndrome delineation and stratification.

Supplemental Information

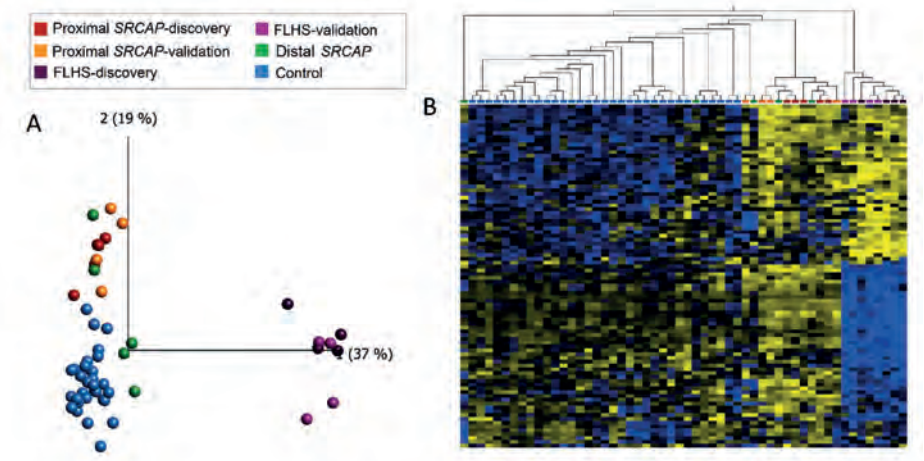


Figure S1. Clustering of *SRCAP* cohort and control samples using previous DNAm signature.

All plots show the discovery, validation, and test samples used in this study. ***A****. PCA using the DNAm values at n=99 CpG sites making up the FLHS DNAm signature from Hood et al., (2016) applied to the samples from the present study. Clustering of all samples is very similar to that of the FLHS signature from this study shown in Figure 1.* ***B****. The heatmap shows hierarchical clustering of DNAm values at the FLHS signature sites from Hood et al., (2016) for all SRCAP samples and discovery controls from the present study. A similar pattern of DNAm to the FLHS signature described in the present study is evident across all samples, including the intermediate, hypermethylated profile of the proximal-SRCAP cases. Euclidian distance metric is used for the clustering dendrogram.*

Tables S1-S13 supporting the findings of this study are available online in the Supplementary material of this article at: DOI: https://doi.org/10.1016/j.ajhg.2021.04.008

Acknowledgments

We are grateful to all the study participants and their families and to the many clinicians who recruited them into this study. This work was supported by Canadian Institutes of Health Research (CIHR) grants (IGH-155182 and MOP-126054) the Province of Ontario Neurodevelopmental Disorders (POND) network (IDS-11-02), and McLaughlin Center (MC 2015-16) grants to R.W. by the Dutch Research Council grant to T.K. (015.014.036) and L.E.L.M.V. (015014066) and Netherlands Organization for Health Research and Development grant to T.K. (91718310) and L.E.L.M.V. (843002608, 846002003), by Donders Junior Researcher Grant 2019 to L.E.L.M.V. and B.B.A.d.V., by Estonian Research Council grants to K.Õ. and K.R. (PRG471) and S.P. (PUTJD827, MOBTP175), by a grant of the Raregenomics network, financed by the Dirección de General de Universidades e Investigación de la Comunidad de Madrid (S2017 / BMD-3721) to M. P.-M., by Deutsche Forschungsgemeinschaft to D.L. (LE4223/1-1). This work is contributed towards the goals of the Solve-RD project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 779257 (to H.B. and L.E.L.M.V.). Several authors of this publication are members of the European Reference Network on congenital malformations and rare intellectual disability (ITHACA). We also acknowledge the technical assistance of Khadine Wiltshire, Youliang Lou, and Chunhua Zhao. Thank you as well to Dr. Greg Hanna for contributing blood DNA samples from typically developing control individuals who had undergone cognitive/behavioral assessments. We also acknowledge Dr. Marleen E. H. Simon, Dr. Alanna Strong, Dr. Femke Tammer and Dr. Bregje van Bon for providing DNA samples of individuals with FLHS and *SRCAP* missense variants.

Declaration of Interests

The authors declare no competing interests.

Web resources

DECIPHER https://decipher.sanger.ac.uk/
GenBank, https://www.ncbi.nlm.nih.gov/genbank/
GeneMatcher https://genematcher.org/
GEO, https://www.ncbi.nlm.nih.gov/geo/
GREAT, http://great.stanford.edu/public/html/
OMIM, https://www.omim.org/
ProteinPaint, https://proteinpaint.stjude.org
Simons Simplex
Collection https://www.sfari.org/resource/simons-simplex-collection/

Data Availability

All genotype and phenotype data supporting findings of this study are available within manuscript and supplement files. The reported variants are available at the ClinVar database with accession number [accession SCV001477310 - SCV001477340]. The DNAm datasets are not publicly available due to institutional ethics restrictions.

References

1. Kleefstra, T., Schenck, A., Kramer, J.M., and van Bokhoven, H. (2014). The genetics of cognitive epigenetics. Neuropharmacology *80*, 83-94. 10.1016/j.neuropharm.2013.12.025.
2. Fahrner, J.A., and Bjornsson, H.T. (2019). Mendelian disorders of the epigenetic machinery: postnatal malleability and therapeutic prospects. Hum Mol Genet *28*, R254-r264. 10.1093/hmg/ddz174.
3. Messina, G., Atterrato, M.T., and Dimitri, P. (2016). When chromatin organisation floats astray: the Srcap gene and Floating-Harbor syndrome. Journal of medical genetics *53*, 793-797. 10.1136/jmedgenet-2016-103842.
4. Hood, R.L., Lines, M.A., Nikkel, S.M., Schwartzentruber, J., Beaulieu, C., Nowaczyk, M.J., Allanson, J., Kim, C.A., Wieczorek, D., Moilanen, J.S., et al. (2012). Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. Am J Hum Genet *90*, 308-313. 10.1016/j.ajhg.2011.12.001.
5. Kehrer, M., Beckmann, A., Wyduba, J., Finckh, U., Dufke, A., Gaiser, U., and Tzschach, A. (2014). Floating-Harbor syndrome: SRCAP mutations are not restricted to exon 34. Clinical genetics *85*, 498-499. 10.1111/cge.12199.
6. Seifert, W., Meinecke, P., Kruger, G., Rossier, E., Heinritz, W., Wusthof, A., and Horn, D. (2014). Expanded spectrum of exon 33 and 34 mutations in SRCAP and follow-up in patients with Floating-Harbor syndrome. BMC medical genetics *15*, 127. 10.1186/s12881-014-0127-0.
7. Le Goff, C., Mahaut, C., Bottani, A., Doray, B., Goldenberg, A., Moncla, A., Odent, S., Nitschke, P., Munnich, A., Faivre, L., and Cormier-Daire, V. (2013). Not all floating-harbor syndrome cases are due to mutations in exon 34 of SRCAP. Human mutation *34*, 88-92. 10.1002/humu.22216.
8. Nikkel, S.M., Dauber, A., de Munnik, S., Connolly, M., Hood, R.L., Caluseriu, O., Hurst, J., Kini, U., Nowaczyk, M.J., Afenjar, A., et al. (2013). The phenotype of Floating-Harbor syndrome: clinical characterization of 52 individuals with mutations in exon 34 of SRCAP. Orphanet journal of rare diseases *8*, 63. 10.1186/1750-1172-8-63.
9. Cuvertino, S., Stuart, H.M., Chandler, K.E., Roberts, N.A., Armstrong, R., Bernardini, L., Bhaskar, S., Callewaert, B., Clayton-Smith, J., Davalillo, C.H., et al. (2017). ACTB Loss-of-Function Mutations Result in a Pleiotropic Developmental Disorder. Am J Hum Genet *101*, 1021-1033. 10.1016/j.ajhg.2017.11.006.
10. Menke, L.A., van Belzen, M.J., Alders, M., Cristofoli, F., Ehmke, N., Fergelot, P., Foster, A., Gerkes, E.H., Hoffer, M.J., Horn, D., et al. (2016). CREBBP mutations in individuals without Rubinstein-Taybi syndrome phenotype. Am J Med Genet A *170*, 2681-2693. 10.1002/ajmg.a.37800.
11. Martinez, F., Marín-Reina, P., Sanchis-Calvo, A., Perez-Aytés, A., Oltra, S., Roselló, M., Mayo, S., Monfort, S., Pantoja, J., and Orellana, C. (2015). Novel mutations of NFIX gene causing Marshall-Smith syndrome or Sotos-like syndrome: one gene, two phenotypes. Pediatric Research *78*, 533. 10.1038/pr.2015.135.
12. Cappuccio, G., Sayou, C., Tanno, P.L., Tisserant, E., Bruel, A.L., Kennani, S.E., Sá, J., Low, K.J., Dias, C., Havlovicová, M., et al. (2020). De novo SMARCA2 variants clustered outside the helicase domain cause a new recognizable syndrome with intellectual disability and blepharophimosis distinct from Nicolaides-Baraitser syndrome. Genet Med. 10.1038/s41436-020-0898-y.
13. Banka, S., Sayer, R., Breen, C., Barton, S., Pavaine, J., Sheppard, S.E., Bedoukian, E., Skraban, C., Cuddapah, V.A., and Clayton-Smith, J. (2019). Genotype-phenotype specificity in Menke-Hennekam syndrome caused by missense variants in exon 30 or 31 of CREBBP. Am J Med Genet A *179*, 1058-1062. 10.1002/ajmg.a.61131.

14. Choufani, S., Cytrynbaum, C., Chung, B.H., Turinsky, A.L., Grafodatskaya, D., Chen, Y.A., Cohen, A.S., Dupuis, L., Butcher, D.T., Siu, M.T., et al. (2015). NSD1 mutations generate a genome-wide DNA methylation signature. Nat Commun *6*, 10207. 10.1038/ncomms10207.
15. Aref-Eshghi, E., Bend, E.G., Colaiacovo, S., Caudle, M., Chakrabarti, R., Napier, M., Brick, L., Brady, L., Carere, D.A., Levy, M.A., et al. (2019). Diagnostic Utility of Genome-wide DNA Methylation Testing in Genetically Unsolved Individuals with Suspected Hereditary Conditions. The American Journal of Human Genetics *104*, 685-700. 10.1016/j.ajhg.2019.03.008.
16. Aref-Eshghi, E., Bend, E.G., Hood, R.L., Schenkel, L.C., Carere, D.A., Chakrabarti, R., Nagamani, S.C.S., Cheung, S.W., Campeau, P.M., Prasad, C., et al. (2018). BAFopathies' DNA methylation epi-signatures demonstrate diagnostic utility and functional continuum of Coffin-Siris and Nicolaides-Baraitser syndromes. Nature communications *9*, 4885-4815. 10.1038/s41467-018-07193-y.
17. Aref-Eshghi, E., Rodenhiser, D.I., Schenkel, L.C., Lin, H., Skinner, C., Ainsworth, P., Paré, G., Hood, R.L., Bulman, D.E., Kernohan, K.D., et al. (2018). Genomic DNA Methylation Signatures Enable Concurrent Diagnosis and Clinical Genetic Variant Classification in Neurodevelopmental Syndromes. American journal of human genetics *102*, 156-174. 10.1016/j.ajhg.2017.12.008.
18. Butcher, D.T., Cytrynbaum, C., Turinsky, A.L., Siu, M.T., Inbar-Feigenberg, M., Mendoza-Londono, R., Chitayat, D., Walker, S., Machado, J., Caluseriu, O., et al. (2017). CHARGE and Kabuki Syndromes: Gene-Specific DNA Methylation Signatures Identify Epigenetic Mechanisms Linking These Clinically Overlapping Conditions. Am J Hum Genet *100*, 773-788. 10.1016/j.ajhg.2017.04.004.
19. Chater-Diehl, E., Ejaz, R., Cytrynbaum, C., Siu, M.T., Turinsky, A., Choufani, S., Goodman, S.J., Abdul-Rahman, O., Bedford, M., Dorrani, N., et al. (2019). New insights into DNA methylation signatures: SMARCA2 variants in Nicolaides-Baraitser syndrome. 1-14. 10.1186/s12920-019-0555-y.
20. Chen, Y.A., Choufani, S., Grafodatskaya, D., Butcher, D.T., Ferreira, J.C., and Weksberg, R. (2012). Cross-reactive DNA microarray probes lead to false discovery of autosomal sex-associated DNA methylation. Am J Hum Genet *91*, 762-764. 10.1016/j.ajhg.2012.06.020.
21. Siu, M.T., Butcher, D.T., Turinsky, A.L., Cytrynbaum, C., Stavropoulos, D.J., Walker, S., Caluseriu, O., Carter, M., Lou, Y., Nicolson, R., et al. (2019). Functional DNA methylation signatures for autism spectrum disorder genomic risk loci: 16p11.2 deletions and CHD8 variants. . Clin Epigenetics *In press*.
22. Choufani, S., Gibson, W.T., Turinsky, A.L., Chung, B.H.Y., Wang, T., Garg, K., Vitriolo, A., Cohen, A.S.A., Cyrus, S., Goodman, S., et al. (2020). DNA Methylation Signature for EZH2 Functionally Classifies Sequence Variants in Three PRC2 Complex Genes. Am J Hum Genet. 10.1016/j.ajhg.2020.03.008.
23. Aref-Eshghi, E., Kerkhof, J., Pedro, V.P., Barat-Houari, M., Ruiz-Pallares, N., Andrau, J.-C., Lacombe, D., Van-Gils, J., Fergelot, P., Dubourg, C., et al. (2020). Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in 42 Mendelian Neurodevelopmental Disorders. The American Journal of Human Genetics *106*, 356-370. https://doi.org/10.1016/j.ajhg.2020.01.019.
24. Hood, R.L., Schenkel, L.C., Nikkel, S.M., Ainsworth, P.J., Pare, G., Boycott, K.M., Bulman, D.E., and Sadikovic, B. (2016). The defining DNA methylation signature of Floating-Harbor Syndrome. Sci Rep *6*, 38803. 10.1038/srep38803.
25. Fokkema, I., van der Velde, K.J., Slofstra, M.K., Ruivenkamp, C.A.L., Vogel, M.J., Pfundt, R., Blok, M.J., Lekanne Deprez, R.H., Waisfisz, Q., Abbott, K.M., et al. (2019). Dutch genome diagnostic laboratories accelerated and improved variant interpretation and increased accuracy by sharing data. Hum Mutat *40*, 2230-2238. 10.1002/humu.23896.

26. RK, C.Y., Merico, D., Bookman, M., J, L.H., Thiruvahindrapuram, B., Patel, R.V., Whitney, J., Deflaux, N., Bingham, J., Wang, Z., et al. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci *20*, 602-611. 10.1038/nn.4524.
27. Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat *36*, 928-930. 10.1002/humu.22844.
28. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. American journal of human genetics *84*, 524-533. 10.1016/j.ajhg.2009.03.010.
29. Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192-195. 10.1016/j.neuron.2010.10.006.
30. Neveling, K., Feenstra, I., Gilissen, C., Hoefsloot, L.H., Kamsteeg, E.J., Mensenkamp, A.R., Rodenburg, R.J., Yntema, H.G., Spruijt, L., Vermeer, S., et al. (2013). A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. Human mutation *34*, 1721-1726. 10.1002/humu.22450.
31. Guillen Sacoto, M.J., Tchasovnikarova, I.A., Torti, E., Forster, C., Andrew, E.H., Anselm, I., Baranano, K.W., Briere, L.C., Cohen, J.S., Craigen, W.J., et al. (2020). De Novo Variants in the ATPase Module of MORC2 Cause a Neurodevelopmental Disorder with Growth Retardation and Variable Craniofacial Dysmorphism. American journal of human genetics *107*, 352-363. 10.1016/j.ajhg.2020.06.013.
32. Frederiksen, A.L., Larsen, M.J., Brusgaard, K., Novack, D.V., Knudsen, P.J., Schrøder, H.D., Qiu, W., Eckhardt, C., McAlister, W.H., Kassem, M., et al. (2016). Neonatal High Bone Mass With First Mutation of the NF-κB Complex: Heterozygous De Novo Missense (p.Asp512Ser) RELA (Rela/p65). J Bone Miner Res *31*, 163-172. 10.1002/jbmr.2590.
33. Holla Ø, L., Busk Ø, L., Tveten, K., Hilmarsen, H.T., Strand, L., Høyer, H., Bakken, A., Skjelbred, C.F., and Braathen, G.J. (2015). Clinical exome sequencing – Norwegian findings. Tidsskr Nor Laegeforen *135*, 1833-1837. 10.4045/tidsskr.14.1442.
34. Haer-Wigman, L., van Zelst-Stams, W.A., Pfundt, R., van den Born, L.I., Klaver, C.C., Verheij, J.B., Hoyng, C.B., Breuning, M.H., Boon, C.J., Kievit, A.J., et al. (2017). Diagnostic exome sequencing in 266 Dutch patients with visual impairment. European journal of human genetics : EJHG *25*, 591-599. 10.1038/ejhg.2017.9.
35. Cappuccio, G., Pinelli, M., Torella, A., Alagia, M., Auricchio, R., Staiano, A., Nigro, V., and Brunetti-Pierri, N. (2017). Expanding the phenotype of DST-related disorder: A case report suggesting a genotype/phenotype correlation. Am J Med Genet A *173*, 2743-2746. 10.1002/ajmg.a.38367.
36. Hempel, M., Cremer, K., Ockeloen, C.W., Lichtenbelt, K.D., Herkert, J.C., Denecke, J., Haack, T.B., Zink, A.M., Becker, J., Wohlleber, E., et al. (2015). De Novo Mutations in CHAMP1 Cause Intellectual Disability with Severe Speech Impairment. American journal of human genetics *97*, 493-500. 10.1016/j.ajhg.2015.08.003.
37. van der Sluijs, P.J., Aten, E., Barge-Schaapveld, D., Bijlsma, E.K., Bökenkamp-Gramann, R., Donker Kaat, L., van Doorn, R., van de Putte, D.F., van Haeringen, A., Ten Harkel, A.D.J., et al. (2019). Putting genome-wide sequencing in neonates into perspective. Genetics in medicine : official journal of the American College of Medical Genetics *21*, 1074-1082. 10.1038/s41436-018-0293-0.
38. Terhal, P.A., Vlaar, J.M., Middelkamp, S., Nievelstein, R.A.J., Nikkels, P.G.J., Ross, J., Créton, M., Bos, J.W., Voskuil-Kerkhof, E.S.M., Cuppen, E., et al. (2020). Biallelic variants in POLR3GL cause endosteal hyperostosis and oligodontia. European journal of human genetics : EJHG *28*, 31-39. 10.1038/s41431-019-0427-0.

39. Pajusalu, S., Kahre, T., Roomere, H., Murumets, Ü., Roht, L., Simenson, K., Reimand, T., and Õunap, K. (2018). Large gene panel sequencing in clinical diagnostics-results from 501 consecutive cases. Clinical genetics *93*, 78-83. 10.1111/cge.13031.

40. Hanna, G.L., Liu, Y., Isaacs, Y.E., Ayoub, A.M., Torres, J.J., O'Hara, N.B., and Gehring, W.J. (2016). Withdrawn/Depressed Behaviors and Error-Related Brain Activity in Youth With Obsessive-Compulsive Disorder. Journal of the American Academy of Child and Adolescent Psychiatry *55*, 906-913.e902. 10.1016/j.jaac.2016.06.012.

41. Goodman, S.J., Cytrynbaum, C., Chung, B.H.-Y., Chater-Diehl, E., Aziz, C., Turinsky, A.L., Kellam, B., Keller, M., Ko, J.M., Caluseriu, O., et al. (2020). EHMT1 pathogenic variants and 9q34.3 microdeletions share altered DNA methylation patterns in patients with Kleefstra syndrome. Journal of Translational Genetics and Genomics *4*, [Online First]. 10.20517/jtgg.2020.23.

42. Salas, L.A., Koestler, D.C., Butler, R.A., Hansen, H.M., Wiencke, J.K., Kelsey, K.T., and Christensen, B.C. (2018). An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol *19*, 64. 10.1186/s13059-018-1448-7.

43. Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol *14*, R115. 10.1186/gb-2013-14-10-r115.

44. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology *28*, 495-501. 10.1038/nbt.1630.

45. van der Donk, R., Jansen, S., Schuurs-Hoeijmakers, J.H.M., Koolen, D.A., Goltstein, L., Hoischen, A., Brunner, H.G., Kemmeren, P., Nellåker, C., Vissers, L., et al. (2019). Next-generation phenotyping using computer vision algorithms in rare genomic neurodevelopmental disorders. Genetics in medicine : official journal of the American College of Medical Genetics *21*, 1719-1725. 10.1038/s41436-018-0404-y.

46. Diets, I.J., van der Donk, R., Baltrunaite, K., Waanders, E., Reijnders, M.R.F., Dingemans, A.J.M., Pfundt, R., Vulto-van Silfhout, A.T., Wiel, L., Gilissen, C., et al. (2019). De Novo and Inherited Pathogenic Variants in KDM3B Cause Intellectual Disability, Short Stature, and Facial Dysmorphism. Am J Hum Genet *104*, 758-766. 10.1016/j.ajhg.2019.02.023.

47. Baltrusaitis, T., Zadeh, A., Lim, Y., and Morency, L.-P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit 10.1109/FG.2018.00019.

48. Ferry, Q., Steinberg, J., Webber, C., FitzPatrick, D.R., Ponting, C.P., Zisserman, A., and Nellåker, C. (2014). Diagnostically relevant facial gestalt information from ordinary photos. Elife *3*, e02020. 10.7554/eLife.02020.

49. Turinsky, A.L., Choufani, S., Lu, K., Liu, D., Mashouri, P., Min, D., Weksberg, R., and Brudno, M. (2020). EpigenCentral: Portal for DNA methylation data analysis and classification in rare diseases. Human mutation. 10.1002/humu.24076.

50. Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet *13*, 484-492. 10.1038/nrg3230.

51. Båvner, A., Matthews, J., Sanyal, S., Gustafsson, J.A., and Treuter, E. (2005). EID3 is a novel EID family member and an inhibitor of CBP-dependent co-activation. Nucleic Acids Res *33*, 3561-3569. 10.1093/nar/gki667.

52. Wong, M.M., Cox, L.K., and Chrivia, J.C. (2007). The chromatin remodeling protein, SRCAP, is critical for deposition of the histone variant H2A.Z at promoters. The Journal of biological chemistry *282*, 26132-26139. 10.1074/jbc.M703418200.

53. Dong, S., Han, J., Chen, H., Liu, T., Huen, M.S.Y., Yang, Y., Guo, C., and Huang, J. (2014). The human SRCAP chromatin remodeling complex promotes DNA-end resection. Curr Biol *24*, 2097-2110. 10.1016/j.cub.2014.07.081.

54. Le Guen, T., Jullien, L., Touzot, F., Schertzer, M., Gaillard, L., Perderiset, M., Carpentier, W., Nitschke, P., Picard, C., Couillault, G., et al. (2013). Human RTEL1 deficiency causes Hoyeraal-Hreidarsson syndrome with short telomeres and genome instability. Hum Mol Genet *22*, 3239-3249. 10.1093/hmg/ddt178.

55. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443. 10.1038/s41586-020-2308-7.

56. Gerundino, F., Marseglia, G., Pescucci, C., Pelo, E., Benelli, M., Giachini, C., Federighi, B., Antonelli, C., and Torricelli, F. (2014). 16p11.2 de novo microdeletion encompassing SRCAP gene in a patient with speech impairment, global developmental delay and behavioural problems. European journal of medical genetics *57*, 649-653. 10.1016/j.ejmg.2014.09.009.

57. Malan, V., Rajan, D., Thomas, S., Shaw, A.C., Louis Dit Picard, H., Layet, V., Till, M., van Haeringen, A., Mortier, G., Nampoothiri, S., et al. (2010). Distinct effects of allelic NFIX mutations on nonsense-mediated mRNA decay engender either a Sotos-like or a Marshall-Smith syndrome. American journal of human genetics *87*, 189-198. 10.1016/j.ajhg.2010.07.001.

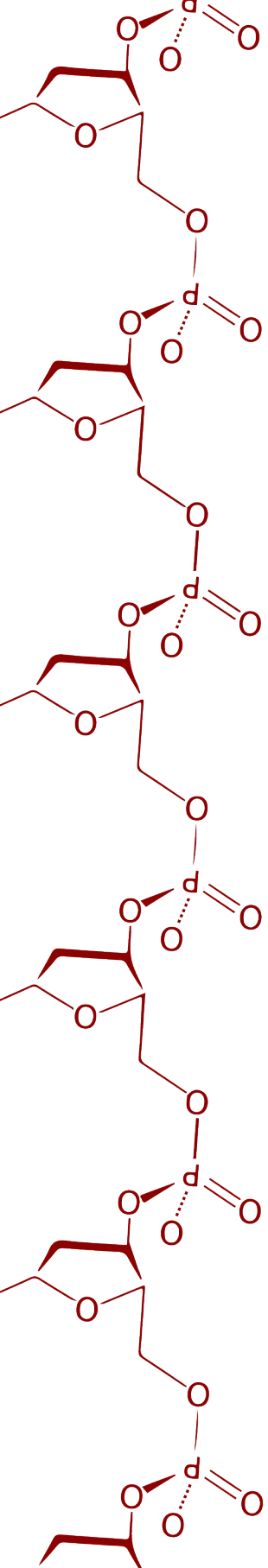
58. Rodríguez, J., Mosquera, J., Couceiro, J.R., Vázquez, M.E., and Mascareñas, J.L. (2015). The AT-Hook motif as a versatile minor groove anchor for promoting DNA binding of transcription factor fragments. Chem Sci *6*, 4767-4771. 10.1039/c5sc01415h.

59. Cattaruzzi, G., Altamura, S., Tessari, M.A., Rustighi, A., Giancotti, V., Pucillo, C., and Manfioletti, G. (2007). The second AT-hook of the architectural transcription factor HMGA2 is determinant for nuclear localization and function. Nucleic acids research *35*, 1751-1760. 10.1093/nar/gkl1106.

60. Baker, S.A., Chen, L., Wilkins, A.D., Yu, P., Lichtarge, O., and Zoghbi, H.Y. (2013). An AT-hook domain in MeCP2 determines the clinical course of Rett syndrome and related disorders. Cell *152*, 984-996. 10.1016/j.cell.2013.01.038.

61. Bend, E.G., Aref-Eshghi, E., Everman, D.B., Rogers, R.C., Cathey, S.S., Prijoles, E.J., Lyons, M.J., Davis, H., Clarkson, K., Gripp, K.W., et al. (2019). Gene domain-specific DNA methylation episignatures highlight distinct molecular entities of ADNP syndrome. Clin Epigenetics *11*, 64. 10.1186/s13148-019-0658-5.

62. Vissers, L., Kalvakuri, S., de Boer, E., Geuer, S., Oud, M., van Outersterp, I., Kwint, M., Witmond, M., Kersten, S., Polla, D.L., et al. (2020). De Novo Variants in CNOT1, a Central Component of the CCR4-NOT Complex Involved in Gene Expression and RNA and Protein Stability, Cause Neurodevelopmental Delay. American journal of human genetics *107*, 164-172. 10.1016/j.ajhg.2020.05.017



Chapter 3:

The clinical and molecular spectrum of the *KDM6B*-related neurodevelopmental disorder

Published: American Journal of Human Genetics. 2023 Jun 1;110(6):963-978.

Authors

Dmitrijs Rots*, Taryn E. Jakub*, Crystal Keung, Adam Jackson, Siddharth Banka, Rolph Pfundt, Bert B.A. de Vries, Richard H. van Jaarsveld, Saskia M. J. Hopman, Ellen van Binsbergen, Irene Valenzuela, Maja Hempel, Tatjana Bierhals, Fanny Kortüm, Francois Lecoquierre, Alice Goldenberg, Jens Michael Hertz, Charlotte Brasch Andersen, Maria Kibæk, Eloise J. Prijoles, Roger E. Stevenson, David B. Everman, Wesley G. Patterson, Linyan Meng, Charul Gijavanekar, Karl De Dios, Shenela Lakhani, Tess Levy, Matias Wagner, Dagmar Wieczorek, Paul J. Benke, María Soledad Lopez Garcia, Renee Perrier, Sergio B. Sousa, Pedro M. Almeida, Maria José Simões, Bertrand Isidor, Wallid Deb, Andrew A. Schmanski, Omar Abdul-Rahman, Christophe Philippe, Ange-Line Bruel, Laurence Faivre, Antonio Vitobello, Christel Thauvin, Jeroen J. Smits, Livia Garavelli, Stefano G. Caraffi, Francesca Peluso, Laura Davis-Keppen, Dylan Platt, Erin Royer, Lisette Leeuwen, Margje Sinnema, Alexander P. A. Stegmann, Constance T.R.M. Stumpel, George E. Tiller, Daniëlle G.M. Bosch, Stephanus T. Potgieter, Shelagh Joss, Miranda Splitt, Simon Holden, Matina Prapa, Nicola Foulds, Sofia Douzgou, Kaija Puura, Regina Waltes, Andreas G. Chiocchetti, Christine M. Freitag, F. Kyle Satterstrom, Silvia De Rubeis, Joseph Buxbaum, Bruce D. Gelb, Aleksic Branko, Itaru Kushima, Jennifer Howe, Stephen W. Scherer, Alessia Arado, Chiara Baldo, Olivier Patat, Demeer Bénédicte, Diego Lopergolo, Filippo M. Santorelli, Tobias B. Haack, Andreas Dufke, Miriam Bertrand, Ruth J. Falb, Angelika Rieß, Peter Krieg, Stephanie Spranger, Maria Francesca Bedeschi, Maria Iascone, Sarah Josephi-Taylor, Tony Roscioli, Michael F. Buckley, Jan Liebelt, Aditi I. Dagli, Emmelien Aten, Anna C.E. Hurst, Alesha Hicks, Mohnish Suri, Ermal Aliu, Sunil Naik, Richard Sidlow, Juliette Coursimault, Gaël Nicolas, Hanna Küpper, Florence Petit, Veyan Ibrahim, Deniz Top, Francesca Di Cara, Genomics England Research Consortium, Raymond J. Louie, Elliot Stolerman, Han G. Brunner, Lisenka E.L.M. Vissers, Jamie M. Kramer**, Tjitske Kleefstra**

*,** These authors contributed equally to this work

Abstract

De novo variants are a leading cause of neurodevelopmental disorders (NDDs), but because every monogenic NDD is different and usually extremely rare, it remains a major challenge to understand the complete phenotype and genotype spectrum of any morbid gene. According to OMIM, heterozygous variants in *KDM6B* cause "neurodevelopmental disorder with coarse facies and mild distal skeletal abnormalities". Here, by examining the molecular and clinical spectrum of 85 new individuals with mostly *de novo* (likely) pathogenic *KDM6B* variants, we demonstrate that this description is inaccurate and potentially misleading. Cognitive deficits are seen consistently in all individuals, but the overall phenotype is highly variable. Notably, coarse facies and distal skeletal anomalies, as defined by OMIM, are rare in this expanded cohort while other features are unexpectedly common (e.g., hypotonia, psychosis etc.). Using 3D-protein structure analysis and an innovative dual *Drosophila* gain of function assay, we demonstrated a disruptive effect of 11 missense/in-frame indels located in or near the enzymatic JmJC or Zn-containing domain of KDM6B. Consistent with the role of *KDM6B* in human cognition, we demonstrated a role for the *Drosophila KDM6B* ortholog in memory and behavior. Taken together, we accurately define the broad clinical spectrum of the *KDM6B*-related NDD, introduce an innovative functional testing paradigm for the assessment of *KDM6B* variants, and demonstrate a conserved role for KDM6B in cognition and behavior. Our study demonstrates the critical importance of international collaboration, sharing of clinical data, and rigorous functional analysis of genetic variants to ensure correct disease diagnosis for rare disorders.

Introduction

The development of the brain is a complex process requiring precise control of gene expression by epigenetic regulators[1], including proteins involved in enzymatic modification of histone tails, ATP-dependent chromatin remodeling, and DNA methylation. Dysfunction of epigenetic regulators frequently results in neurodevelopmental disorders (NDDs)[2]. Pathogenic variants in genes encoding epigenetic regulators, including histone methylases and demethylases, are a common cause of monogenic NDDs[3,4].

The complex of proteins associated with Set1 (COMPASS) and COMPASS-like complexes are important components of the epigenetic machinery[5]. The COMPASS complexes are highly conserved among species, including *Drosophila* and yeast, and their main function is to promote gene expression by methylating histone H3 on lysine 4 (H3K4) with enzymes containing a SET domain, and demethylating histone H3 on lysine 27 (H3K27) through the enzymatic activity of the KDM6A and KDM6B demethylases[5].

KDM6A and KDM6B demethylate di- and trimethylated H3K27 through the catalytic activity of the iron-containing jumonji C domain (JmJC), which is common to different histone demethylases[6]. KDM6B can act independently or as a component of a COMPASS-like complex[7]. KDM6B can also influence transcription independent of its enzymatic activity, although the non-demethylase function of KDM6B is poorly understood[8]. KDM6B dysfunction has also been implicated in various disorders, including cancer, immunologic, and developmental disorders[9].

Recently, Stolerman et al. reported a cohort (n=12) of individuals with *de novo* *KDM6B* variants, suggesting that haploinsufficiency of *KDM6B* may result in a novel syndromic NDD with multisystem involvement[10]. However, current knowledge regarding the molecular and clinical spectrum of the *KDM6B*-related NDD is limited and the function of KDM6B in neurons remains undefined. OMIM currently classifies this *KDM6B*-related NDD as "neurodevelopmental disorder with coarse facies and mild distal skeletal abnormalities" (NEDCFSA). Here, we further characterized the clinical and molecular spectrum of this disorder, based on a large cohort (n=85) of individuals with (likely) pathogenic *KDM6B* variants. In addition, we developed *Drosophila* models to assess the impact of identified *KDM6B* variants and to examine the role of KDM6B in regulating cognitive function and behavior. Our results elucidate a more complete clinical and molecular spectrum for the *KDM6B*-related NDD and indicate an urgent need to reassess the current OMIM description for KDM6B. These findings highlight the challenges in defining rare NDDs in general.

Materials and Methods

Cohort recruitment

We have collected genetic and clinical data from 85 individuals with rare heterozygous (mostly *de novo*) variants in the *KDM6B* gene. The variants were annotated using the GRCh37 reference and NM_001080424.2/ENST00000254846.9 transcript. The individuals were recruited from the Radboudumc in-house diagnostic laboratory, international collaborators, individuals registered in GeneMatcher[11], and individuals included in various research cohorts, such as the Simons Simplex collection (SSC)[12], Deciphering Developmental Delay (DDD)[13], 100,000 Genome Project[14], Pediatric Cardiac Genomics Consortium (PCGC)[15], Autism Sequencing Consortium (ASC)[16,17], and MSSNG[18]. The variants were identified by performing exome or genome sequencing in diagnostic or research settings using standard laboratory methods[19-27]. For individuals identified through the DDD study, a complementary analysis project (CAP #83) was approved which filtered for variants in chromatin remodeling genes throughout the entire cohort. This list was then filtered for rare *de novo* variants with damaging *in silico* predictions. Panel-agnostic re-analysis of locally unsolved cases from the DDD Study (CAP #147) was also performed as previously described[4,28]. For the 100,000 Genome Project, tiered variants from the 3rd September 2020 data release were accessed/filtered via LabKey. Variants were filtered for *de novo* inheritance and clinicians contacted through the AirLock.

After collecting all evidence, we re-interpreted all identified variants according to ACMG variant classification guidelines[29]: variants in 73 individuals were classified as (likely) pathogenic, but variants in 12 individuals were classified as variants of unknown significance (VUS) due to limited or controversial evidence. Clinical features of only individuals with (likely) pathogenic variants were further analyzed. An overview of the study design is shown in **Figure 1A**.

Detailed descriptions of 73 individuals with (likely) pathogenic *KDM6B* variants, their molecular findings, and corresponding study type (clinical or research) are provided in **Table S2**. In most individuals (52/73), variants occur *de novo*, but nine truncating variants were inherited from a mildly affected or unaffected parent; and for 12 individuals, the inheritance was unknown. To provide a more precise description of different clinical feature frequencies, we aggregated our individual data with the previously published 12 individuals (resulting in cohort of 85 individuals with (likely) pathogenic *KDM6B* variants). Incomplete data across individuals was corrected for when calculating total feature frequency, as well

as one individual (#17) with a pathogenic *KDM6B* variant in combination with pathogenic *HNRNPU* variant (#MIM:602869) was not included in the clinical feature frequency calculations (**Table 1**), to minimize possible effects from additional genetic variants. Detailed clinical and molecular descriptions of 12 individuals with *KDM6B* VUS are provided in the **Table S3**. All variants identified in this study were deposited to the ClinVar database (ClinVar accession numbers: SCV002570417 - SCV002570487).

Out of 85 individuals with (likely) pathogenic *KDM6B* variants, 12 had protein altering variants (PAVs) and 73 – protein truncating variants (PTVs). Fisher's exact test was used to compare the frequency of clinical features between the individuals with (likely) pathogenic PAVs vs. PTVs (**Table 1**). Bonferroni correction was used to account for multiple testing.

Ethics

This study was approved by the institutional review board "*Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen*" under number 2011/188. The study participants or their caregivers gave informed consent to participate in the research, of whom 21 consented also to photo publishing (15 individuals with pathogenic *KDM6B* variants and 6 with VUS). Sample data obtained from contributing sites was based on their original ethics protocols referenced in the methods.

Protein structure analysis

The solved three-dimensional (3D) crystal structure of the KDM6B protein was used for analysis of the possible effects of identified variants on the protein. Possible effects of PAVs were predicted based on the wild-type amino acid position and interactions (with other amino acids, other proteins or ligands) and biophysical differences with the mutant amino acid, similarly as described previously[30]. The detailed description of the predicted effects is provided in the **Table S1**. The protein structure used (PDB:5OY3) contains the C-terminal part of the protein (p.1141-1643) with the JmJC and Zn-containing domains required for the H3 tail binding and H3K27-specific demethylation together with ligands and co-factors[6]. KDM6A JmJC domain structure (PDB:3AVR) was used for comparison with KDM6B[31]. For interpretation of the variants in KDM6B protein regions without solved structure (mostly N-terminal part), UniProt (ID: O15054)[32] provided information (e.g. about disordered regions or modified residues), as well as AlphaFold[33] provided O15054 *ab initio* model was used[34].The analysis and visualization were performed using YASARA Structure software[35].

Variant clustering analysis

Significance of variant clustering was calculated separately for PAVs and PTVs as described before [36]. Shortly, geometric mean distance on linear protein structure for the observed variants was compared with randomly permutated variants, performing 1,000,000 permutations and Bonferroni correction for two experiments by using SpatialClustering tool.

Drosophila strains and culture

Flies were reared on standard cornmeal-agar media at 25°C with a 12h/12h light/dark cycle in 70% humidity. The mushroom body (MB) driver, *R14H06-Gal4* (stock #48667), *UAS-Utx-RNAi* (Utx^{RNAi1}; stock #34076), *UAS-mCherry-RNAi* (control^{1}; stock #35785), ubiquitous driver *Act5C-Gal4/CyO* (Act-Gal4; stock #4414), and wing-specific driver *MS1096-Gal4* (stock #8860) fly lines were obtained from the Bloomington Drosophila stock center. A second *UAS-Utx-RNAi* (Utx^{RNAi2}; stock #37664) and its genetic background control line (control^{2}; stock #60000) were obtained from Vienna Drosophila Research Center. Control^{1} was used as a control because it shares a common genetic background with Utx^{RNAi1} and expresses a non-targeting double stranded RNA that controls for any nonspecific effects of a RNAi in general. Null mutations in *Utx* are known to cause lethality in flies[37]. In agreement, expression of Utx^{RNAi} lines with a ubiquitous Act-Gal4 driver resulted in lethality, suggesting that the RNAi lines are effective at inducing knockdown.

UAS-KDM6B transgenic flies were generated through Gateway cloning of the reference *KDM6B* cDNA from KIAA0346 in *pENTR3C-KDM6B* (a gift from Professor Kristian Helin[38]) into *pGW-HA.attB* (GenBank #KC896838) to create *pGW-KDM6B.attB* (*UAS-KDM6B^{ref}*). Nineteen different PAVs (**Figure 2B**) were introduced into *pGW-KDM6B.attB* using PCR-based site directed mutagenesis. The JmJC domain deletion of KDM6B (*UAS-KDM6B$^{\Delta JmJC}$*) was completed using the ligation method and primers adapted from Xiang et al.[39]. A similar ligation method was used to generate domain deletions of the N-terminus (*UAS-KDM6B$^{\Delta Nterm}$;* p.Met1_Pro1100), the Zn containing domain (*UAS-KDM6B$^{\Delta Zndom}$*; p.Tyr1563_Leu1619) and the C-terminal region encompassing Zn containing and linker domains (*UAS-KDM6B$^{\Delta Cterm}$*; p.Thr1489_Arg1682). All 23 *UAS-KDM6B* constructs were validated by Sanger sequencing and inserted into the third chromosome *attP2* landing site through phiC31-mediated transgenesis at Genome Prolab (Sherbrooke, Quebec, Canada).

Drosophila memory, activity, and sleep assays

Utx^{RNAi} and genetic control fly lines were crossed to *R14H06-Gal4*, and the resulting progeny were analyzed in memory, activity and sleep assays. Short-term memory

(STM) and long-term memory (LTM) were assessed on male flies aged 5 days using courtship conditioning, as previously described[40]. Briefly, for each fly pair a courtship index (CI) was calculated, which is the proportion of time spent courting over 10 minutes. A minimum of 30 flies were assayed for each genotype in short-term and long-term memory assays. Within each genotype, naïve flies were compared to trained flies using Kruskal-Wallis test with uncorrected Dunnett's test for multiple comparisons. To assess naïve courting behavior, CI from naïve short- and long-term experiments were pooled for each genotype. Utx^{RNAi} flies were compared to their genetic control using Kruskal-Wallis test with uncorrected Dunnett's test for multiple comparisons.

Total activity and sleep of flies were monitored as previously described[41]. Briefly, a total of 32 flies for each phenotype, males aged 1-4 days, were loaded into activity monitor chambers (Trikinetics, MA, USA). After 1 day of acclimation, fly locomotion was recorded over a 48-hour period of 12h/12h light/dark cycle, and averaged for each fly to reveal typical 24-hour locomotion patterns. Total beam breaks/day were compared between genetic control and corresponding Utx^{RNAi} using t-test with two tailed distribution and unequal variance. A 5-minute period of no activity is defined as 'sleep'[42,43]. Total minutes of sleep over a 48-hour period were averaged for each fly to reveal typical 24-hour sleep patterns and compared using t-test with two tailed distributions and unequal variance.

Drosophila mushroom body morphology

To determine whether morphological defects could be responsible for observed memory and behavioral phenotypes, we visualized the structural morphology of the *Drosophila* mushroom body. Utx^{RNAi} and genetic control fly lines were crossed to *R14H06-Gal4* and males and females aged 2-5 days of the resulting progeny were examined for mushroom body morphology. Brains were dissected in PBS, fixed with 4% paraformaldehyde for 45 minutes at room temperature, and mounted in Vectashield (Vector Laboratories). Brains were imaged using a Zeiss LSM800 confocal microscope at 200X magnification. Confocal stacks were processed using ImageJ software[44]. Gross mushroom body morphology was assessed qualitatively and were consistent between at least 10 brains for each genotype.

Drosophila Gain of Function assays

Gain of function (GoF) phenotypes were observed upon ectopic expression of *UAS-KDM6B*^{ref} using the ubiquitous Act-Gal4 driver and the *MS1096-Gal4* wing driver. Ubiquitous overexpression of *KDM6B*^{ref} causes lethality (**Figure 2B**). *Act-Gal4/CyO* flies were crossed to all *UAS-KDM6B* variants, and the percentage lethality was

calculated by comparing the number of progeny receiving *Act-Gal4* to those receiving the *CyO* balancer chromosome (% lethality = (1 – # Act-Gal4/# CyO) x 100%). For each ubiquitous *UAS-KDM6B* variant cross, between 50-350 progeny were assessed. Percent lethality of *UAS-KDM6B* variant transgenes was compared to *UAS-KDM6B*^{ref} using a Chi-square two-sample test for equality of proportions. Expression of *UAS-KDM6B*^{ref} with *MS1096-Gal4* wing driver causes a defect in the formation of the L5 vein in the posterior compartment of the wing. This leads to splitting of L5 at the distal end and results in the appearance of an extra vein protruding into the third posterior cell (**Figure 2B**). A range of 18-35 male flies, aged 2-5 days, were analyzed in wing-specific overexpression for each UAS-*KDM6B* variant. Fly wings were mounted in glycerol and imaged using Nikon SMZ800N stereo microscope under 40X magnification. The length of the extra vein was measured and quantified using ImageJ software[44]. Statistical comparison of wing vein length was performed using ANOVA with Dunnett's test for multiple comparisons.

Since the *UAS-KDM6B*$^{\Delta JmJC}$ and pathogenic PAVs cannot induce gain of function effects in either assay (i.e. no lethality and no extra vein protrusion (**Fig. 2B**), only PAVs that showed effects on both GoF assays were interpreted as functionally disruptive.

Results

The spectrum of identified KDM6B variants

To define the clinical and molecular spectrum of the *KDM6B*-related neurodevelopmental disorder, we summarized clinical and genetic information from 85 individuals presenting with a NDD with rare heterozygous (mostly *de novo*) variants in *KDM6B*. When combined with the previously reported 12 individuals[10], the total cohort of 97 individuals included information about 81 unique variants in *KDM6B* (**Figure 1**). The vast majority of variants are present in a single individual, with only seven variants (three nonsense, three frameshifts and one in-frame indel) being recurrent in two or three unrelated individuals (as *de novo* and/or inherited). The variants included 60 PTVs (32 frameshift, 21 nonsense, and 7 canonical splicing variants) and 21 PAVs, including 18 missense variants and 3 in-frame indels. Based on ACMG guidelines, all PTVs (n=60) were classified as likely pathogenic (**Figure 1**). For PAVs (n=21), classification was refined with protein structure data and functional analysis in *Drosophila* models with 11 being classified as (likely) pathogenic and 10 as variants of uncertain significance (VUS) (see below).

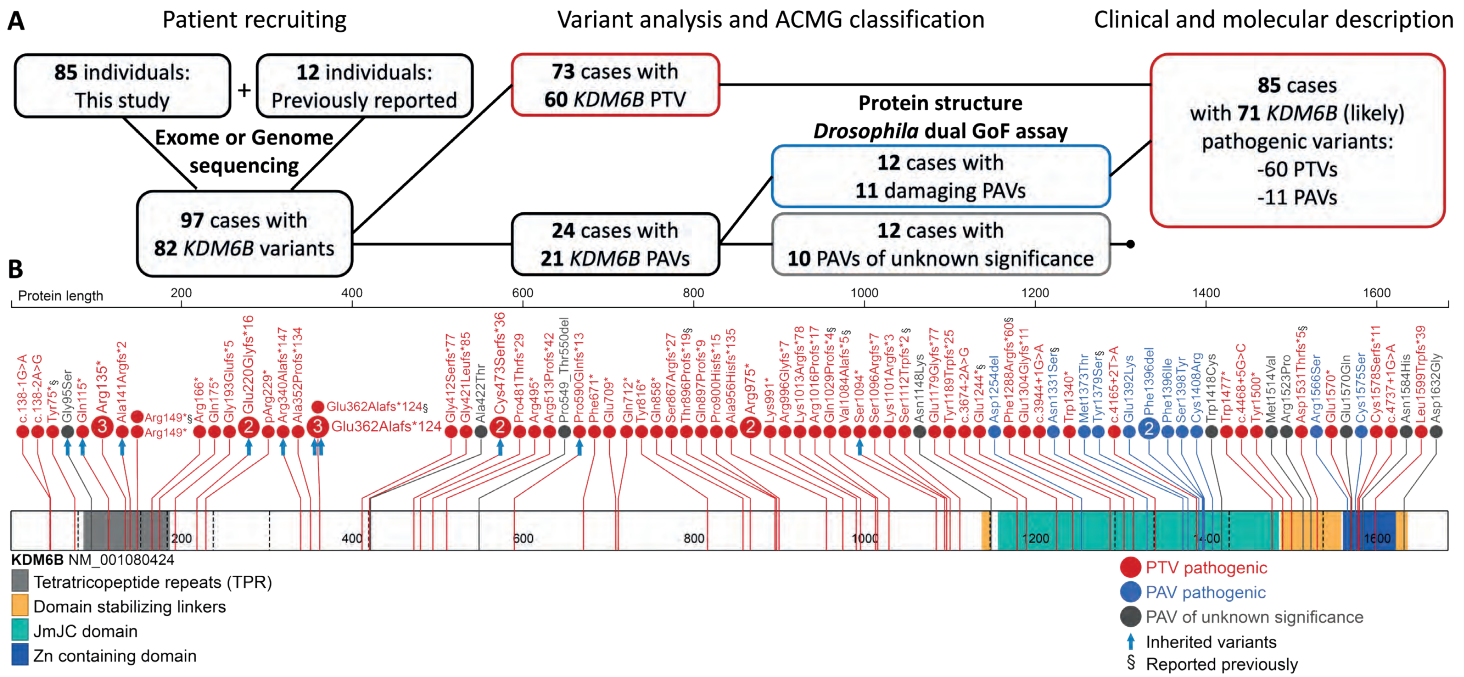


Figure 1. Overview of the study design and the identified *KDM6B* variants.

***A**. Schematic illustration of the study design. **B**. Identified KDM6B variants in independent families and their position.JmJC = Jumonji C domain; Zn=Zinc; PTV = protein truncating (nonsense, frameshift, canonical splice) variants; PAV = protein altering (missense and in-frame indel) variants*

Predicted effect of PAVs on 3D protein structure

KDM6B has two known functional domains, the JmJC domain (p.1157–1485), which is required for the enzymatic activity of the protein (demethylating H3K27me3/2) and a Zn-containing domain (p.1563–1620). The Zn-containing domain is structurally similar to GATA-type zinc fingers, but, unlike zinc fingers, it is responsible for the specific binding to the H3 tail[6]. The interaction between the Zn-domain and H3 ensures the specificity of KDM6B to demethylate H3K27. Adjacent to the Zn-containing and JmJC domains, there are two linker regions (p.1490–1558 and 1623–1635) that interact with and stabilize the domains[6]. While PTVs are scattered throughout the gene (including the last and penultimate exons) (corrected p value = 1 for clustering), 18 of 21 identified PAVs significantly cluster at the C-terminus of the protein, in or near the JmJC and Zn-containing domains (**Figure 1B**) (corrected p value = $2.0x10^{-6}$), at positions that are predicted to be intolerant to missense variation (**Figure S1**). The remaining 3 PAVs (c.283G>A p.(Gly95Ser); c.1264G>A p.(Ala422Thr); c.1645_1650del p.(Pro549_Thr550del)) are located outside of the defined KDM6B protein structure: c.1264G>A p.(Ala422Thr) and c.1645_1650del p.(Pro549_Thr550del) are located in a disordered (without structure) region of the protein, without known functions, while p.(Gly95Ser) is located in the tetratricopeptide repeat (TPR) domain (predicted based on sequence with AlphaFold model and amino acid homology with other TPR domains), whose function in KDM6B is currently also unknown.

Available KDM6B and KDM6A protein structures for the C-terminal region spanning amino acids 1147-1640[6,31] were used to predict the effect of PAVs on KDM6B function. Of the 21 identified KDM6B PAVs, 18 are located in this region of the protein, allowing for prediction of the effect of each amino acid substitution on protein function. Based on the analysis, 16 of the PAVs are predicted to have a disruptive effect on KDM6B protein structure: nine variants are predicted to disrupt the JmJC domain structure or active site; three PAVs are predicted to disrupt the Zn-containing domain structure or H3 binding; and four stabilizing domain variants were predicted to have a significant impact on the structure of the linker or local structure of the C-terminal region, while the remaining two stabilizing linker variants (c.3444T>G p.(Asn1148Lys) and c.4895A>G p.(Asp1632Gly)) are predicted to have a minimal effect (**Figure 2A**). **Table S1** contains a detailed description of the structural predictions (in addition to ACMG classification) for all 21 KDM6B PAVs.

Experimental testing of KDM6B PAVs using a dual Drosophila Gain of Function assay

A robust dual GoF assay using *Drosophila melanogaster* was developed to experimentally assess the damaging effect of KDM6B PAVs on protein function. A *UAS-KDM6B*^{ref} transgene was generated which revealed that overexpression of human KDM6B in different tissues in flies can induce highly consistent GoF phenotypes. Expression of *UAS-KDM6B*^{ref} with the ubiquitous *Actin-Gal4* driver results in lethality, while expression in the wing with *MS1096-Gal4* results in the formation of an extra vein protruding into the third posterior cell (**Figure 2B**). Overexpression of KDM6B mutants lacking the enzymatic JmJC domain (*UAS-KDM6B*$^{\Delta JmJC}$) and/or the Zn-containing domain (*UAS-KDM6B*$^{\Delta Zndom}$) did not induce lethality or the formation of an extra wing vein. This demonstrates that these GoF phenotypes can be used to detect loss-of-function related to absence of KDM6B enzymatic activity (mediated by the JmJC domain), and/or histone binding (mediated by the Zn-containing domain). In contrast, expression of a KDM6B construct with a deletion of the entire N-terminal region of the protein(*UAS-KDM6B*$^{\Delta Nterm}$) was able to induce lethality (similar to wild-type construct *UAS-KDM6B*^{ref}), and extra wing vein formation was only mildly reduced when compared to the KDM6B reference protein. This shows that our GoF assay can robustly assess the KDM6B functionality related to the C-terminal portion of the protein including the JmJC domain and the Zn-containing domain, while the role of the KDM6B N-terminal part is either dispensable or cannot be reliably assessed using current assay.

We used our dual GoF assay to assess the functional effects of all 18 identified KDM6B PAVs that were present in the C-terminal region of the protein. One benign variant: c.1244C>T p.(Pro415Gln), was used as a negative control based on a frequent occurrence in the gnomAD database. Expression of *UAS-KDM6B*^{P415Q} induced lethality and extra wing vein protrusions similar to *UAS-KDM6B*^{ref}, indicating a functional protein as expected (**Figure 2B**). Of the 18 PAVs tested, eight of the nine variants located in the JmJC domain failed to induce the GoF phenotypes or showed significantly reduced magnitude of the phenotypes compared to the controls, indicating a clear loss of KDM6B function associated with these variants. Variants located in the Zn-containing domain and the domain stabilizing linkers showed less consistent loss of function phenotypes. Variants in the Zn-containing domain showed diverse effects from complete loss of function for (c.4724G>C p.Cys1575Ser) to reduced function (c.4696C>A p.(Arg1566Ser) and c.4708G>C p.(Glu1570Gln)), or no loss of function (c.4750A>C p.(Asn1584His)). However, c.4708G>C p.(Glu1570Gln) show only effects on lethality, but not on the wing vein suggesting a different (or mild) effect. Variants in the domain stabilizing linkers

showed little effect on KDM6B function, with only the wing phenotypes being moderately reduced for some alleles. Importantly, the only JmJC domain variant that did not show KDM6B function loss (c.4254G>T p.(Trp1418Cys)) is predicted to disrupt a linker region, and not the JmJC domain (**Table S1 and Figure 2A**) which is consistent with the GoF assay results. These data highlight the sensitivity of the JmJC and Zn-containing domain to missense variants and suggest that the stabilizing function of the linkers is more tolerant to missense variation in general.

Taken together, these data provide evidence for 11 KDM6B PAVs (found in 12 individuals) to be classified as (likely) pathogenic, based on the ACMG guidelines. The remaining 10 PAVs are classified as VUS, including the untested 3 PAVs located in the KDM6B N-terminal region, and 6 alleles that did not show reduced functionality in the dual GoF assay. Details of the ACMG classification for all PAVs are provided in **Table S1**.

The Drosophila KDM6B ortholog Utx is required in neurons for normal cognition and behavior

We aimed to understand the potential fundamental role of KDM6B in the brain using a *Drosophila* LoF model. *Drosophila* has a single ortholog of *KDM6A* and *KDM6B*, *Utx*, which is ubiquitously expressed in the fly brain. Germline loss of *Utx* is lethal[37], and so RNAi knockdown was used to deplete *Utx* in *Drosophila* memory neurons of the mushroom body (MB). MB-specific RNAi knockdown was achieved using the *R14H06-Gal4* driver line, which is highly specific for post-mitotic MB neurons in the adult and larval fly brain[45]. *Utx* MB knockdown flies were assessed for memory, courtship behavior, activity, and sleep. Two unique *Utx* RNAi lines both caused defects in short- and long-term memory and overall courtship behavior (**Figures 3A and 3B**). In addition, overall daily activity and sleep were affected modestly, with one of two RNAi lines showing significant increase in sleep and a significant reduction in activity (**Figures 3C and 3D**).

The *Drosophila* MB undergoes extensive postmitotic morphological remodeling during fly development, and disruption of these morphogenic processes could underly the observed memory and behavior defects. However, confocal imaging of *Utx* RNAi knockdown flies showed normal morphology (**Figure 3E**) suggesting that Utx-dependent memory and behavior defects are not caused by disrupted MB morphogenesis. These findings reflect the broad behavioral effects of the *KDM6B*-haploinsufficiency described in our cohort, confirming a role for this protein family in regulating cognition.

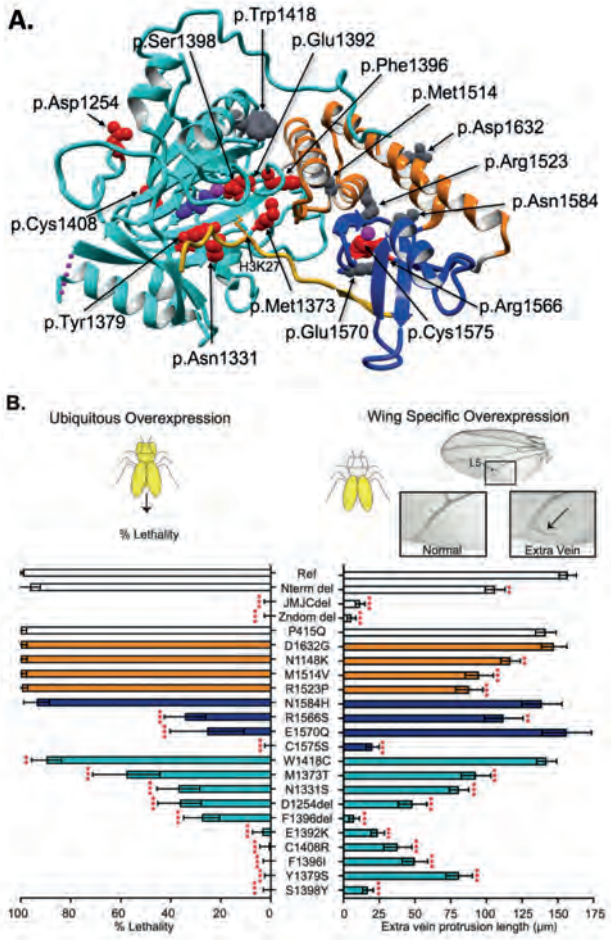


Figure 2. Analysis of *KDM6B* PAVs using protein 3D structure analysis and a dual *Drosophila* gain-of-function assay.

A. *KDM6B fragment (PDB: 5OY3, p.1157-1639) bound to the H3 tail fragment (p.17-33). The JmJC domain is shown in cyan with 2-oxoglutaric acid (purple) bound with an Fe ion (magenta) which are necessary for the enzymatic demethylation of H3K27. The Zn-containing domain is shown in blue with Zn ion (magenta). Two out of three JmJC and Zn-containing domain stabilizing linkers are also visible in the structure (orange). The H3 tail with K27 residue positioned into the active center of the JmJC domain is shown in yellow. Amino acids affected by missense or in-frame indels are shown as balls (pathogenic - red, VUS - grey), affecting all shown domains, as well as binding to H3 tail (see Table S1 for more details on specific variants).* ***B.*** *A dual Drosophila gain-of-function assay was used to assess the disruptive potential of KDM6B PAVs. Ubiquitous overexpression (left) of KDM6B^{ref} using the UAS/Gal4 system results in complete lethality. Percent lethality assessed for KDM6B$^{\Delta Nterm}$, KDM6B$^{\Delta JmJC}$, KDM6B$^{\Delta Zndom}$, KDM6B$^{\Delta Cterm}$, KDM6B^{P415Q} as a benign control, and 18 KDM6B variants were compared to KDM6B^{ref} (Chi-squared test). N=50-230 flies for each genotype. Wing-specific overexpression (right) of KDM6B^{ref} in the fly wings results in the formation of an extra vein protruding off the L5 vein. The length of the extra vein was compared to KDM6B^{ref} (Dunnet's test). N = 18-35 flies for each sample. PAVs are colored based on the domain (JmJC – cyan, Zn-containing – Blue, Stabilizing linkers – Orange; no domain – white, same as in Fig 1B and 2A). ** p<0.01, *** p<0.0001.*

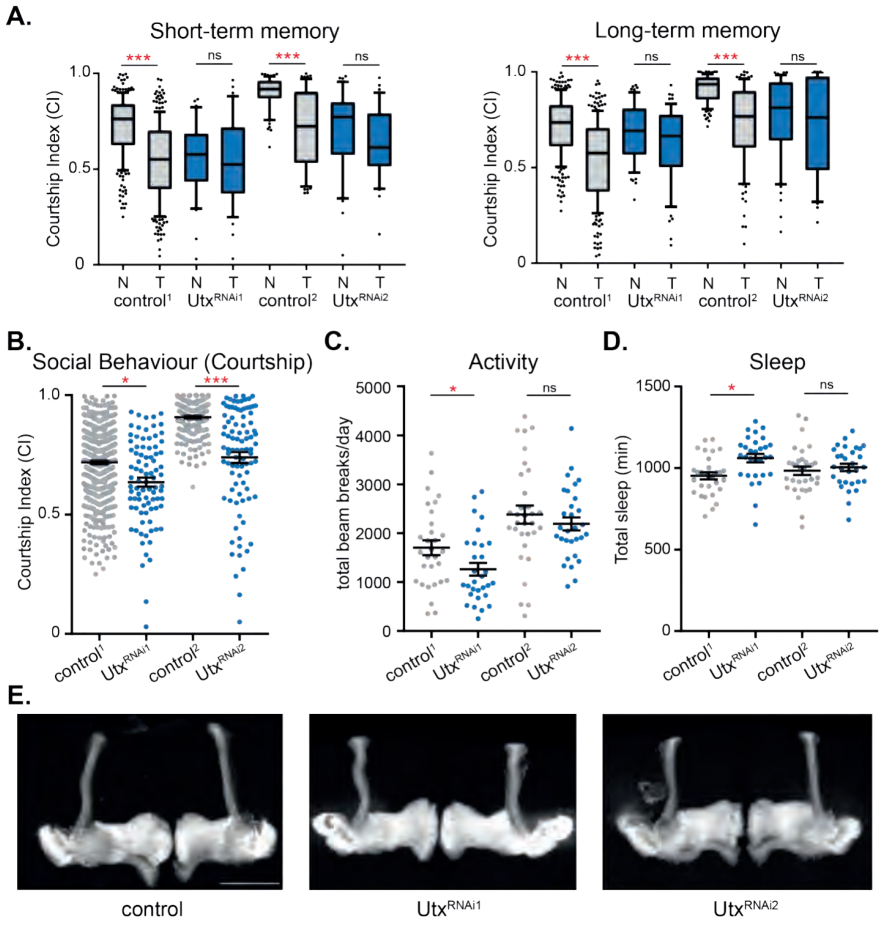


Figure 3. The *Drosophila KDM6B* ortholog, *Utx*, is required in neurons for normal memory and behavior.

A. *Short term (STM) and long term (LTM) courtship memory was assessed upon MB-specific expression (R14H06-Gal4) of two independent Utx RNAi lines (Utx^{RNAi1} and Utx^{RNA2}) and their genetic controls ($control^{1}$ and $control^{2}$). Boxplots show the distribution of courtship indices (CI) for naïve (N) and trained (T) male flies aged 5 days. Memory was observed when a significant reduction in CI occurred between naïve and trained conditions of the same genotype (Kruskal Wallis Test). All controls show a significant reduction in courtship in trained vs naive groups, while Utx RNAi knockdown flies did not. At least 30 flies were tested per condition.* ***B.*** *Naïve courting behavior was pooled from short- and long-term memory assays and compared between MB-specific Utx RNAi knockdown flies and their genetic controls. At least 60 male flies aged 5 days were tested. MB specific Utx RNAi knockdown caused* ***C.*** *reduced daily activity and* ***D.*** *increased sleep compared to genetic controls, but these differences were only significant for Utx^{RNAi1} (t-test). N=32 flies for each genotype.* ***E.*** *No morphological defects were found following MB specific knockdown of Utx compared to their genetic controls. MB morphology was consistent in at least 10 brains for each genotype. Scale bar= 50 µm. * p<0.05, ** p<0.01, *** p<0.0001.*

The clinical spectrum of the KDM6B-related NDD

In total, the genetic and functional analyses presented in this study confirmed the likely pathogenic effect of 71 different *KDM6B* variants identified in 85 individuals (**Figure 1A**). For 64/85 (75%) of these individuals, *de novo* occurrence of the variant was demonstrated, but nine of the individuals inherited the variant (five maternal, four paternal) from a mildly affected (DD, learning problems, ASD) or clinically unaffected parent. All variants, their classifications, and their location within the protein are shown in **Figure 1B**.

For comprehensive characterization of the *KDM6B*-related NDD clinical spectrum, we assessed clinical information only from 85 individuals with *KDM6B* variants that we classified as (likely) pathogenic: 73 described in this study and 12 published previously[10]. For 16/73 newly identified individuals recruited from large cohort research studies, only limited clinical data were available. Proportional statistics along with detailed clinical information for the individuals with the (likely) pathogenic variants are provided in **Table 1** and **Table S2**, respectively. To focus on the *KDM6B* phenotype, one of 73 individuals was not included in the calculation of the clinical features because of an additional pathogenic variant in *HNRNPU* gene. As the effects of the identified VUS are currently unknown, the clinical characteristics of these individuals are provided separately (**Table S3**) and were not included in the clinical feature description or frequency calculations.

Neurodevelopmental abnormalities were present in all individuals with likely pathogenic *KDM6B* variants. Developmental delay (speech-language, motor, or global) was the most common feature, present in all except two individuals. However, at age ≥5 years, when more objective developmental parameters can be assessed, neurodevelopmental problems were less prevalent. Most individuals had intellectual disability (ID), autism spectrum disorder (ASD), or both. The level of ID was mostly mild and was present in 63% of individuals. Importantly, severe ID was reported in only two individuals, with one of the individuals having a second diagnosis due to a pathogenic *HNRNPU* variant (NM_031844.2:c.970A>G p.(Arg324Gly))[46]. ASD was reported in 61% of individuals, and other behavioral problems were reported commonly (60%). A psychotic disorder was present in 20% (4/20) of individuals ≥12 years of age.

A significant proportion of the individuals showed various neurological abnormalities, including hypotonia (57%), sleep disturbances (32%), seizures (13%), and movement disorders (24%), including gait abnormalities, dystonia-like movement, spasticity, and hypertonia with toe walking. In several individuals, these

movement disorders were the main presenting feature and reason for performing genetic testing. Movement disorders resolved over time in two individuals, while one individual required treatment with botulinum toxin injections due to spasticity. Similarly, in several individuals, severe hypotonia was the main presenting feature, leading to muscle biopsy and/or muscle disorder gene panel sequencing to exclude primary myopathies.

Approximately one third (30%) of individuals with *KDM6B* variants displayed features of postnatal overgrowth, with tall stature reported in 8% of individuals, and macrocephaly in 26%. Increased weight was reported in 14% of individuals, and 16% (10/63) of individuals had increased birth weight, of whom 7/10 showed overgrowth features later in life. None of the individuals had short stature and the majority have normal growth parameters.

Gastrointestinal issues were common and sometimes severe with a significant impact on the individuals' care. Neonatal feeding difficulties, or gastroesophageal reflux was present in half (51%). For several patients, severe neonatal feeding difficulties required nasogastric tube feeding or even resulted in admission to neonatal intensive care. Constipation, often chronic, was reported in 18% of individuals and, in some, was the major health concern requiring active treatment and regular follow-ups.

Congenital anomalies of different organ systems were also seen in this cohort. Congenital heart disease affected 13% of individuals. Other congenital abnormalities included cleft lip and/or palate, affecting 4%, and congenital genitourinary system anomalies, observed in 10% of individuals. Musculoskeletal system and limb abnormalities were relatively common but mild and variable, with cutaneous (II, III and sometimes IV toe) syndactyly reported in 9% of individuals, spine curvature abnormalities in 13%, short fingers and/or toes present in 9%, and broad fingers and hands or broad toes and feet in 20%.

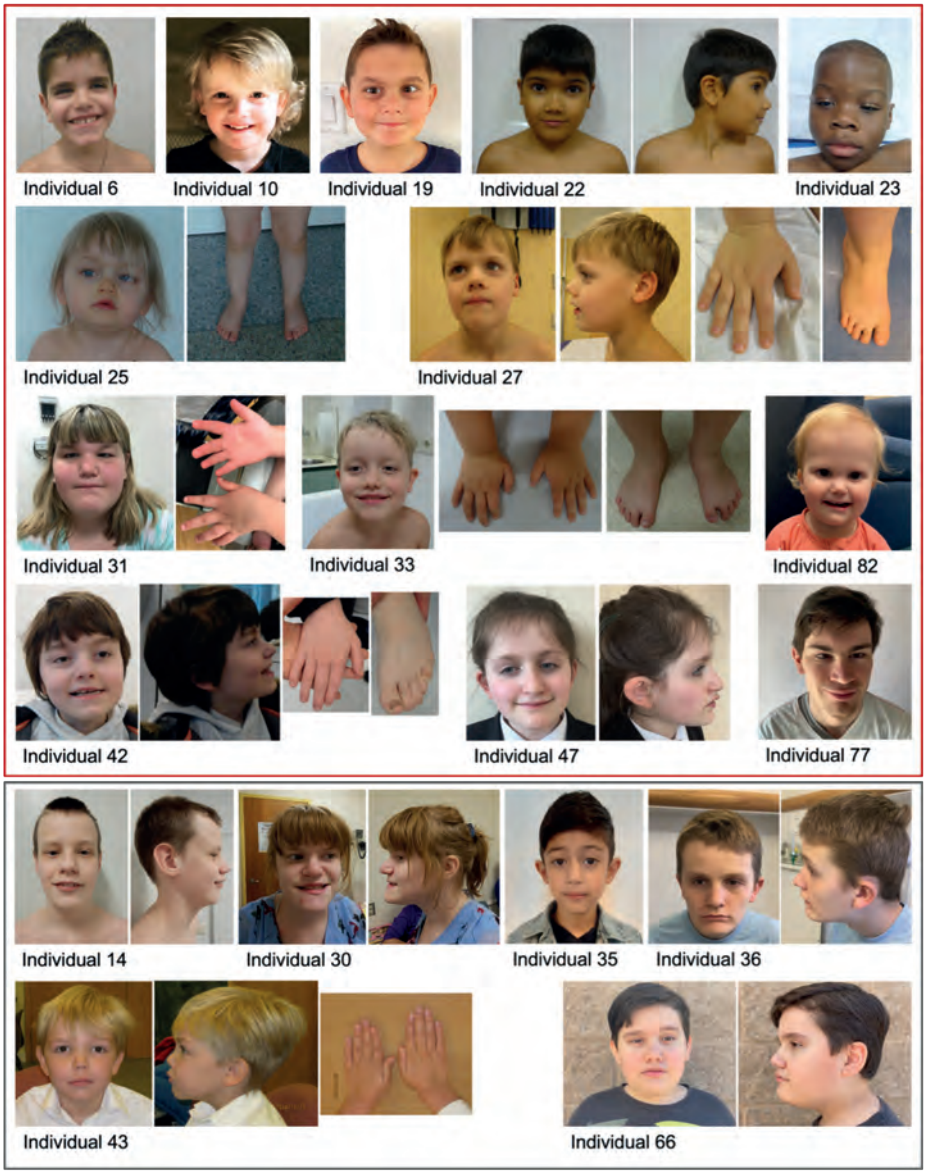


Figure 4. Photos of individuals with identified *KDM6B* variants.

Individuals with (likely) pathogenic variants are shown in the red box above, and those with a VUS are shown in the grey box.

Table 1. Main clinical features among individuals with (likely) pathogenic *KDM6B* variants.

Feature	PTVs N=73*	PAVs N=12	P value (PTVs vs. PAVs)	Total N=85* (%)
Sex (Males/Total)	51/72	10/12	0.50	61/84 (73%)
Growth				
Increased birth weight [>2SD]	10/50	0/8	0.33	10/58 (17%)
Increased weight [>2SD]	9/56	0/8	0.59	9/64 (14%)
Tall stature [>2SD]	5/58	0/8	1.0	5/66 (8%)
Macrocephaly [>2SD]	15/57	2/8	1.0	17/65 (26%)
At least one feature of overgrowth	18/59	3/10	1.0	21/69 (30%)
Neurodevelopmental and psychiatric issues				
Language/speech delay	63/66	9/11	0.15	72/77 (94%)
Motor delay	56/63	10/11	1.0	66/74 (89%)
ID or learning problems	37/56	3/8	0.14	40/64 (63%)
ASD	38/65	8/11	0.51	46/76 (61%)
Behavior problems, non-ASD	40/65	4/8	0.70	44/73 (60%)
Psychotic disorders [≥12 y.o.]	4/18	0/2	1.0	4/20 (20%)
Neurological issues				
Seizures	9/62	0/7	0.58	9/69 (13%)
Sleep disturbances	21/59	0/7	0.09	21/66 (32%)
Movement disorder/ gait disturbances/hypertonia/ataxia	15/59	1/8	0.67	16/67 (24%)
Hypotonia	36/63	4/6	1.0	40/70 (57%)
Gastrointestinal issues				
Neonatal feeding difficulties or gastroesophageal reflux	29/58	4/7	1.0	33/65 (51%)
Constipation	10/55	1/6	1.0	11/61 (18%)

Table 1. Continued

Feature	PTVs N=73*	PAVs N=12	P value (PTVs vs. PAVs)	Total N=85* (%)
Congenital anomalies				
Congenital heart disease	8/57	0/7	0.58	8/64 (13%)
Cleft lip/palate/uvula	1/60	2/7	0.03**	3/67 (4%)
Genitourinary abnormalities	6/55	0/7	1.0	6/62 (10%)
Musculoskeletal and limb abnormalities				
Joint hypermobility	24/56	2/6	1.0	26/62 (42%)
Scoliosis/kyphosis/lordosis	8/57	0/7	0.58	8/64 (13%)
Syndactyly	4/58	2/8	0.15	6/66 (9%)
Short fingers or toes	6/57	0/7	1.0	6/64 (9%)
Broad fingers/fingertips/hands/toes/feet	12/58	1/7	1.0	13/65 (20%)
Sensory issues				
Myopia/Amblyopia	20/54	0/7	0.08	20/61 (33%)
Strabismus	8/57	0/7	0.58	8/64 (13%)
Hearing loss	1/55	0/7	1.0	1/62 (2%)
Recurrent ear infections	7/54	0/5	1.0	7/59 (12%)

*PTVs = protein truncating variants; PAVs = protein altering variants (only (likely) pathogenic variants included); ID – intellectual disability; ASD – autism spectrum disorder; SD – standard deviation*Single individual with a second pathogenic variant in HNRNPU gene was not included in the calculations.**Not significant after correction for multiple testing.*

Dysmorphic facial features were noted for most of the individuals (**Figure 4**) and included anteverted nares with depressed nasal bridge, deep-set eyes with down-slanting and narrow palpebral fissures, and prominent forehead. Additionally, some individuals presented with flat face, synophrys, and overfolded helices. Coarse facial features were uncommon and were reported only in four individuals from this study.

KDM6B-related clinical features described by OMIM (#618505: Neurodevelopmental disorder with coarse facies and mild distal skeletal abnormalities; NEDCFSA), were found to be rare in our large cohort with only 4% incidence of coarse facies, and 9% incidence of very mild distal skeletal abnormalities identified. Overall, the phenotype of the affected individuals was extremely variable, without clear genotype-phenotype correlation for PAVs vs. PTVs (**Table 1**), ranging from isolated developmental delay or neuropsychiatric problems with normal IQ to severe NDD associated with severe ID and/or multiple affected organ systems.

Discussion

In this study, we describe the molecular and clinical spectrum of the *KDM6B*-related NDD using a large cohort of individuals possessing heterozygous PTVs and PAVs in *KDM6B*. Analyses of KDM6B 3D-protein structure combined with an innovative dual GoF assay in *Drosophila*, proved effective for classifying *KDM6B* PAVs. Pathogenic variants in *KDM6B* result in loss of one allele, likely reducing the enzymatic demethylation function of the protein. Our large cohort analysis redefines the *KDM6B*-related clinical spectrum, which includes ID, ASD, facial dysmorphisms, macrocephaly, various neurological and gastrointestinal problems, congenital anomalies, and a relatively high prevalence of psychotic disorders among adult individuals. The neurodevelopmental phenotype observed in our cohort was recapitulated in a *Drosophila* neuronal knockdown model, confirming a conserved role of this family of histone demethylases in cognition and behavior and providing a system to further elucidate the underlying molecular mechanisms.

Analysis of KDM6B protein structure largely explains the effect of the identified PAVs on protein function. The C-terminal region of the protein contains a JmJC and Zn-containing domain in addition to domain-stabilizing linker regions, which are required for normal catalytic activity[6]. According to ACMG variant interpretation guidelines[29], pathogenic moderate (PM1) criteria can be applied for variants located in a functional domain and/or mutational hot-spot. Not surprisingly, pathogenic

de novo KDM6B PAVs significantly cluster at the KDM6B C-terminal region (hot-spot), disrupting the catalytic JmJC domain and the Zn-containing domain, which is required for interaction with histones. Therefore, variants located in these regions could result in loss of H3K27 demethylase activity. To test this, we developed a GoF overexpression assay that can detect loss of KDM6B function associated with its JmJC and Zn-containing domains. Experimental testing of identified PAVs using this *Drosophila* dual GoF assay confirmed the detrimental effects on KDM6B for PAVs occurring in the JmJC domain. However, variable effects were observed for PAVs in the Zn-containing domain and the predicted domain stabilizing linkers. Our assay identified a strong loss of function for the c.4724G>C p.(Cys1575Ser) variant, which directly effects a cysteine in the Zn-containing domain that directly bind the Zn ion. In contrast, moderate loss of function was observed for PAVs predicted to stabilize loops in the Zn-containing domain (c.4696C>A p.(Arg1566Ser) and c.4708G>C p.(Glu1570Gln)), while no loss of function was observed for a PAV present on the surface of the Zn-containing domain that interacts with the domain stabilizing linkers. These results suggests that core amino acids of the domain are more important for protein function than those playing a role in domain stabilization. Indeed, for all four PAVs located in the domain stabilizing linkers, we observed only a minimal (if any) effect on KDM6B function. These results raise doubts about the pathogenicity of these variants, which we classify as VUS. One limitation of our dual GoF assay is that it might be not sensitive to defects in the N-terminal region of the protein as we did not have any clearly pathogenic variant to validate the assay and the N-terminal deletion showed effect only on wing vein. Since this region of the protein is disordered, it was not possible to classify PAVs near the N-terminus of KDM6B. On one hand, N-terminal variants are expected to be benign because this region is mostly disordered, its deletion did affect lethality in our Drosophila GoF assay, it does not have a known tertiary structure (except for a short predicted TPR domain), and is predicted to be mostly tolerant to missense variants on a populational level (**Figure S1**)[47]. However, we cannot exclude that this region is important for the KDM6B non-enzymatic activity (e.g. binding to other proteins) and the variants may act by a different mechanism.

Truncating variants in the last and penultimate exons are usually interpreted with caution as they can escape nonsense-mediated decay (NMD), resulting in translation of a truncated protein. In this study, three individuals with likely pathogenic truncating variants predicted to escape NMD were identified, which were predicted to result in a protein lacking the critical (Zn-containing) domain[48] and predicted to result in functional loss. Supporting this, our GoF assay results show loss of KDM6B function due to deletion of the Zn-containing domain (**Figure 2B**).

Individuals with pathogenic *KDM6B* variants display a wide spectrum of symptoms with variable expressivity. Developmental delay was present in almost all individuals being the most common clinical feature. In total, most reported individuals (~90%) had ID, ASD, or both. This highlights that cognitive deficits are the main consistent clinical feature resulting from *KDM6B* pathogenic variants. Other frequent clinical features included behavioral and psychiatric problems, features of overgrowth, neonatal feeding difficulties, constipation, hypotonia, or movement disorders (spasticity, hypertonia, or ataxia). These features occurred in ~20-60% of individuals and showed variable expressivity. Additionally, most of the individuals also presented with some facial dysmorphism, but most of the dysmorphic features are mild and variable among the individuals and the condition does not have a recognizable facial gestalt. Considering highly variable expressivity, it is not surprising that 9/85 (11%) of individuals had a pathogenic variant inherited from a mildly affected or clinically unaffected parent. Taken together, this condition is unlikely to be recognized based on clinical and dysmorphic features alone. This is similar to other conditions recently described where the phenotype including facial dysmorphism is very broad, requiring additional evidence to prove causality[41,49,50].

Psychotic disorders were reported in four individuals in our study. While it may seem to be a rare feature, it corresponds to 20% of individuals older than 12 years, an age threshold used because psychotic disorders rarely manifest before that age[51]. For the same reason, its true frequency may be underestimated, since only a minority (20/85) of the reported individuals are ≥ 12 years. These findings are in line with other monogenic NDDs that manifest with DD/ID in childhood and psychotic disorders in adolescence/adulthood, e.g., *SETD1A*[41], *KMT2C*[52], *SRCAP*[49], and *EHMT1*[53]. Additionally, *de novo* and rare *KDM6B* gene variants were recently found to be associated with schizophrenia at false discovery rate <5% by Singh et al. 2022, especially among cases with developmental delay, confirming findings from our individual-based cohort[51]. Psychotic disorders are not only complicated to identify and diagnose in such individuals, but they can also have therapeutic implications, as has been recently shown for Kleefstra syndrome, caused by *EHMT1* haploinsufficiency[53]. Therefore, being aware of such risk is critical for accurately diagnosing and providing appropriate treatment and care for these individuals.

There was no genotype-phenotype correlation observed that could explain the variable expressivity of clinical phenotypes. Even though PAVs had different effect sizes in the *Drosophila* GoF assays, individuals with different *KDM6B* variant types (PAVs or PTVs) did not display clear differences in the phenotype. Therefore, we hypothesized that the main differences are likely explained by other genetic or

environmental factors. For example, we have observed a significant sex bias in the cohort, with ~¾ of the affected individuals being males. While the gnomAD database[54], which does not contain individuals with severe pediatric disorders, is depleted of protein-truncating variants (pLI=1, LOEUF=0.14), 13 individuals with high quality truncating variants are still present, with no sex bias (six females and seven males). These data suggest that the previously described[55,56] female protective effect is also at play for the *KDM6B* pathogenic variants. It is likely that in addition to genetic background, rare variants in other NDD genes also contribute to this variability[57,58], as we have observed for one individual with a severe phenotype with two pathogenic variants (in *KDM6B* and *HNRNPU*) (**Table S2**).

Recently, *KDM6B* was independently described to be significantly enriched with *de novo* variants among ~31 thousand individuals with NDD[59] and ~11 thousand individuals with ASD[16]. These observations prove the causality of pathogenic heterozygous *KDM6B* variants in the development of different NDDs and suggests that it is a common NDD cause. Interestingly, in the study by Satterstrom et al., *KDM6B* is categorized as "ASD predominant" by virtue of having a higher frequency of disruptive *de novo* variants in ASD-ascertained probands than in NDD-ascertained probands[16]. However, in our cohort, collected from various sources (including both studies described above) developmental delay is present in almost all individuals, while ASD is present in about two thirds of the individuals. This shows the importance of gathering detailed clinical data to evaluate the findings of meta-analyses conducted by large consortia.

Based on the initial description of 12 individuals with *KDM6B* variants by Stolerman et al., the disorder has been named a "Neurodevelopmental disorder with coarse facies and mild distal skeletal abnormalities" by OMIM (#618505). However, after analyzing this large cohort of individuals with pathogenic *KDM6B* variants, we see that coarse facies, as well as mild distal skeletal abnormalities, are rare and not typical. As it currently stands, such designation could be misleading to professional and patient communities therefore urgent redefinition is required. Based on the wide array of symptoms caused by pathogenic *KDM6B* variants in our cohort, we propose that the name "*KDM6B*-related NDD" would better describe this condition.

While the literature on the role of KDM6B in development is extensive, most studies are limited to cell culture models and often KDM6B haploinsufficiency is used as the control condition in these cell culture studies[60,61]. This suggests that KDM6B haploinsufficiency has a limited effect on cell development and differentiation. This is consistent with the low penetrance of developmental/morphological phenotypes

observed in our cohort (**Table 1, Table S2**). Our analysis of *Utx* in *Drosophila* memory, shows that *Utx* is required for the normal function of adult memory neurons, post development (**Figure 3**). Consistent with this, postnatal knockout of KDM6B in excitatory neurons of mice impairs learning and memory through regulation of dendritic spine formation in the adult mouse brain[62]. Interestingly, a recent study identified autistic-like behavioral deficits in *KDM6B* haploinsufficient mice[63] which seem to replicate some autistic and cognitive features seen on our cohort. Taken together these functional studies suggest that KDM6B mediated demethylation of H3K27 may have an evolutionarily conserved role in adult brain function, which could underly the primary cognitive deficits observed in our cohort.

There are some limitations in this study. First, the cohort has ascertainment bias toward individuals who were genetically tested due to the neurodevelopmental disorders and likely underrepresents mildly affected individuals. There is evidence that some individuals are mildly affected, e.g., 13 individuals with *KDM6B* PTVs are present in the gnomAD database that excluded cases with severe pediatric disorders. Additionally, for the majority of the parents with *KDM6B* pathogenic variants, an NDD phenotype (such as speech delay, learning problems, ASD) was reported, but they were not deeply phenotyped and could not be included in the study. However, they likely represent a milder spectrum of the condition. For better representation of the phenotypic spectrum of the *KDM6B*-related NDD and to reduce bias, we recruited a large cohort of individuals from various sources, both diagnostic testing and research cohorts focusing on various phenotypes. Next, even though we used a dual GoF assay, which allowed us to accurately evaluate the effects of multiple PAVs, the assay might be not sensitive to loss of the N-terminal region function, but the role and functions of this region are unknown. Lastly, *Utx* is the only H3K27 demethylase present in flies, with strong homology to both KDM6B and KDM6A. Behavioral results using a *Utx* loss model are, therefore, relevant to both KDM6A and KDM6B. Disorders cause by mutations in these two protein are both characterized by cognitive deficits, however, there are differences in other clinical phenotypes (e.g., KDM6A pathogenic variants result in Kabuki syndrome type 2, with specific facial features and short stature/microcephaly, while KDM6B-related NDD is associated with overgrowth)[64]. It is unclear whether the different clinical effect of KDM6A and KDM6B are due to different molecular functions of the two proteins, or different expression during human development. There is evidence that KDM6A and KDM6B have some redundancy, but also have unique roles, and differing expression patterns have been observed for KDM6A and KDM6B. A recent study found that in the adult mouse brain KDM6B is specifically expressed in neurons, while KDM6A is expressed in neurons and different types of glia. While the

expression of the two proteins in humans is not well studies, the detailed analysis of mouse expression suggests that clinical differences (for example high prevalence of ASD in our cohort, but not in individuals with KDM6A pathogenic variants) may arise from differing expression patterns. It will be interesting to compare future functional analysis of mouse *Kdm6a, Kdm6b*, and fly *Utx* in the context of cognitive function to understand the true level of redundancy and evolutionary conservation in the brain.

Our study demonstrates the critical importance of international collaboration, sharing of genomic data, and rigorous functional analysis of genetic variants for an unbiased, accurate, and comprehensive definition of rare genetic disorders. Clinically, the *KDM6B*-related NDD is a neurodevelopmental disorder characterized by cognitive defects with broad clinical features of variable expressivity, requiring a molecular diagnosis.

Supplemental information

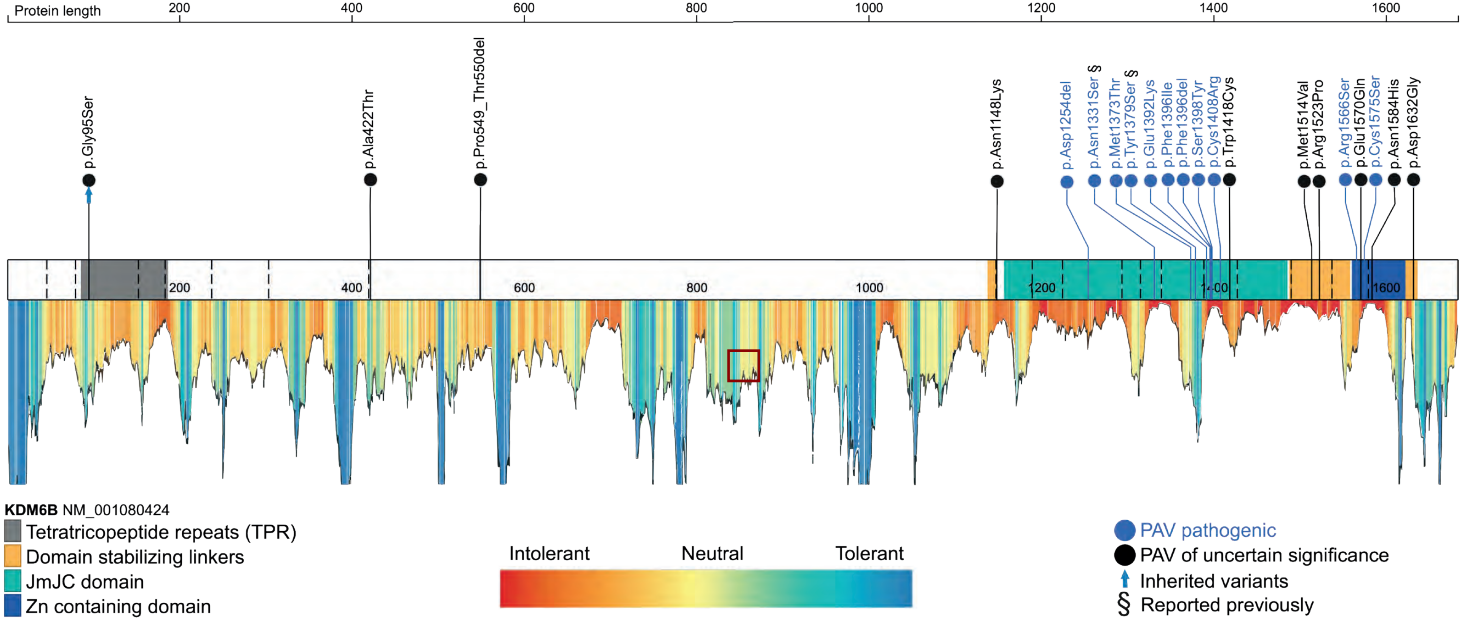


Figure S1. Identified missense and in-frame variant distribution and positional tolerance of missense variants in the *KDM6B* gene.

JmJC=Jumonji C domain; Zn=zinc; PTV=protein truncating (nonsense, frameshift, canonical splice) variants; PAV=protein altering (missense and in-frame indel) variants.

Tables S1-S3 supporting the findings of this study are available online in the Supplementary material of this article at: DOI: 10.1016/j.ajhg.2023.04.008

Acknowledgments

We would like to thank Simons Simplex collection (SSC)(34), Deciphering Developmental Delay (DDD)(35), 100,000 Genome Project(36), Pediatric Cardiac Genomics Consortium (PCGC)(37), Autism Sequencing Consortium (ASC)(29, 38), and Autism Speaks MSSNG(39) projects, as well as the Biobank of the Laboratory of Human Genetics, IRCCS Istituto G. Gaslini (Genoa, Italy, member of the Network Telethon of Genetic Biobanks), for providing cases and samples for the study, as well as GeneDx for providing molecular diagnoses and connecting to the referral physicians. The diagnosis of some patients was made possible through access to the data generated by the 2025 French Genomic Medicine Initiative.

Part of this research was made possible through access to the data and findings generated by the 100KGP. The 100KGP is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100KGP is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100KGP uses data provided by patients and collected by the National Health Service as part of their care and support.

We also would like to thank the Bloomington Drosophila stock center for providing fly stocks and Anastasia Mereshchuk for technical assistance.

Funding

This work was financially supported by:

- Aspasia grant of the Dutch Research Council (015.014.036 to T.K.)
- Netherlands Organization for Health Research and Development (91718310 to T.K.)
- NSERC PGSA grant to T.E.J.
- CIHR Project Scheme grant to J.M.K.
- Canadian Institutes of Health Research Project Grant (PJT 178220) to D.T.
- U.S. National Institutes of Health (HL153009) to B.D.G.
- Autism Speaks, Hospital for Sick Children Foundation and the University of Toronto McLaughlin Centre to S.W.S.
- European Union's Horizon 2020 research and innovation program via The Solve-RD project under grant agreement No 779257 to A.J., S.B., H.G.B. and L.E.L.M.V.
- French Ministry of Health (PHRC-I 18-38, REDIA study) grant and the European Union and Région Normandie in the context of Recherche Innovation Normandie (RIN2018) grant via the European Regional Development Fund (ERDF) to F.L., A.G., J.C., G.N.

- In2Genome project CENTRO-01-0247-FEDER-017800 and GenomePT project POCI-01-0145-FEDER-022184 by the European Regional Development Fund (ERDF) to M.J.S.
- CIHR grant FRN-178220 to D.T. and V.I.
- Research grants from the Japan Agency for Medical Research and Development (AMED) under Grant Nos. JP19km0405216, JP22tm0424222 and JP21wm0425007 and the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Nos. 21K07543, 21H00194.
- European commission grant (AIMS-2-TRIALS – 777394) to C.M.F.

This work has been partly generated within the European Reference Network on Rare Congenital Malformations and Rare Intellectual Disability (ERN-ITHACA). ERN-ITHACA is funded by the EU4Health program of the European Union, under the grant agreement number 101085231.

Competing interests

S.W.S. is a scientific consultant of Population Bio and the King Abdullaziz University, and Athena Diagnostics has licensed intellectual property from his work held by the Hospital for Sick Children, Toronto. All other authors declare they have no competing interests.

Data and materials availability

All data are available in the main text or the supplemental materials. All identified variants and their classification are submitted to ClinVar (accession numbers: SCV002570417 - SCV002570487).

References

1. Gerrard, D.T., Berry, A.A., Jennings, R.E., Birket, M.J., Zarrineh, P., Garstang, M.G., Withey, S.L., Short, P., Jiménez-Gancedo, S., Firbas, P.N., et al. (2020). Dynamic changes in the epigenomic landscape regulate human organogenesis and link to developmental disorders. Nat Commun *11*, 3920. 10.1038/s41467-020-17305-2.
2. Ciptasari, U., and van Bokhoven, H. (2020). The phenomenal epigenome in neurodevelopmental disorders. Hum Mol Genet *29*, R42-r50. 10.1093/hmg/ddaa175.
3. Kleefstra, T., Schenck, A., Kramer, J.M., and van Bokhoven, H. (2014). The genetics of cognitive epigenetics. Neuropharmacology *80*, 83-94. 10.1016/j.neuropharm.2013.12.025.
4. Faundes, V., Newman, W.G., Bernardini, L., Canham, N., Clayton-Smith, J., Dallapiccola, B., Davies, S.J., Demos, M.K., Goldman, A., Gill, H., et al. (2018). Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorders. American journal of human genetics *102*, 175-187. 10.1016/j.ajhg.2017.11.013.
5. Cenik, B.K., and Shilatifard, A. (2021). COMPASS and SWI/SNF complexes in development and disease. Nat Rev Genet *22*, 38-58. 10.1038/s41576-020-0278-0.
6. Jones, S.E., Olsen, L., and Gajhede, M. (2018). Structural Basis of Histone Demethylase KDM6B Histone 3 Lysine 27 Specificity. Biochemistry *57*, 585-592. 10.1021/acs.biochem.7b01152.
7. De Santa, F., Totaro, M.G., Prosperini, E., Notarbartolo, S., Testa, G., and Natoli, G. (2007). The histone H3 lysine-27 demethylase Jmjd3 links inflammation to inhibition of polycomb-mediated gene silencing. Cell *130*, 1083-1094. 10.1016/j.cell.2007.08.019.
8. Meng, Y., Li, H., Liu, C., Zheng, L., and Shen, B. (2018). Jumonji domain-containing protein family: the functions beyond lysine demethylation. J Mol Cell Biol *10*, 371-373. 10.1093/jmcb/mjy010.
9. Zhang, X., Liu, L., Yuan, X., Wei, Y., and Wei, X. (2019). JMJD3 in the regulation of human diseases. Protein Cell *10*, 864-882. 10.1007/s13238-019-0653-9.
10. Stolerman, E.S., Francisco, E., Stallworth, J.L., Jones, J.R., Monaghan, K.G., Keller-Ramey, J., Person, R., Wentzensen, I.M., McWalter, K., Keren, B., et al. (2019). Genetic variants in the KDM6B gene are associated with neurodevelopmental delays and dysmorphic features. Am J Med Genet A *179*, 1276-1286. 10.1002/ajmg.a.61173.
11. Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat *36*, 928-930. 10.1002/humu.22844.
12. Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192-195. 10.1016/j.neuron.2010.10.006.
13. Prevalence and architecture of de novo mutations in developmental disorders. (2017). Nature *542*, 433-438. 10.1038/nature21062.
14. Smedley, D., Smith, K.R., Martin, A., Thomas, E.A., McDonagh, E.M., Cipriani, V., Ellingford, J.M., Arno, G., Tucci, A., Vandrovcova, J., et al. (2021). 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med *385*, 1868-1880. 10.1056/NEJMoa2035790.
15. Jin, S.C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S.R., Zeng, X., Qi, H., Chang, W., Sierant, M.C., et al. (2017). Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nature genetics *49*, 1593-1601. 10.1038/ng.3970.
16. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568-584.e523. 10.1016/j.cell.2019.12.036.

17. Buxbaum, J.D., Daly, M.J., Devlin, B., Lehner, T., Roeder, K., and State, M.W. (2012). The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. Neuron *76*, 1052-1056. 10.1016/j.neuron.2012.12.008.

18. RK, C.Y., Merico, D., Bookman, M., J, L.H., Thiruvahindrapuram, B., Patel, R.V., Whitney, J., Deflaux, N., Bingham, J., Wang, Z., et al. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci *20*, 602-611. 10.1038/nn.4524.

19. Lelieveld, S.H., Reijnders, M.R., Pfundt, R., Yntema, H.G., Kamsteeg, E.J., de Vries, P., de Vries, B.B., Willemsen, M.H., Kleefstra, T., Lohner, K., et al. (2016). Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat Neurosci *19*, 1194-1196. 10.1038/nn.4352.

20. Guillen Sacoto, M.J., Tchasovnikarova, I.A., Torti, E., Forster, C., Andrew, E.H., Anselm, I., Baranano, K.W., Briere, L.C., Cohen, J.S., Craigen, W.J., et al. (2020). De Novo Variants in the ATPase Module of MORC2 Cause a Neurodevelopmental Disorder with Growth Retardation and Variable Craniofacial Dysmorphism. American journal of human genetics *107*, 352-363. 10.1016/j.ajhg.2020.06.013.

21. Brunet, T., Jech, R., Brugger, M., Kovacs, R., Alhaddad, B., Leszinski, G., Riedhammer, K.M., Westphal, D.S., Mahle, I., Mayerhanser, K., et al. (2021). De novo variants in neurodevelopmental disorders-experiences from a tertiary care center. Clinical genetics *100*, 14-28. 10.1111/cge.13946.

22. Lecoquierre, F., Bonnevalle, A., Chadie, A., Gayet, C., Dumant-Forest, C., Renaux-Petel, M., Leca, J.B., Hazelzet, T., Brasseur-Daudruy, M., Louillet, F., et al. (2019). Confirmation and further delineation of the SMG9-deficiency syndrome, a rare and severe developmental disorder. Am J Med Genet A *179*, 2257-2262. 10.1002/ajmg.a.61317.

23. Husson, T., Lecoquierre, F., Cassinari, K., Charbonnier, C., Quenez, O., Goldenberg, A., Guerrot, A.M., Richard, A.C., Drouin-Garraud, V., Brehin, A.C., et al. (2020). Rare genetic susceptibility variants assessment in autism spectrum disorder: detection rate and practical use. Transl Psychiatry *10*, 77. 10.1038/s41398-020-0760-7.

24. Falb, R.J., Müller, A.J., Klein, W., Grimmel, M., Grasshoff, U., Spranger, S., Stöbe, P., Gauck, D., Kuechler, A., Dikow, N., et al. (2021). Bi-allelic loss-of-function variants in KIF21A cause severe fetal akinesia with arthrogryposis multiplex. Journal of medical genetics. 10.1136/jmedgenet-2021-108064.

25. Pezzani, L., Marchetti, D., Cereda, A., Caffi, L.G., Manara, O., Mamoli, D., Pezzoli, L., Lincesso, A.R., Perego, L., Pellicioli, I., et al. (2018). Atypical presentation of pediatric BRAF RASopathy with acute encephalopathy. Am J Med Genet A *176*, 2867-2871. 10.1002/ajmg.a.40635.

26. Hertz, J.M., Svenningsen, P., Dimke, H., Engelund, M.B., Nørgaard, H., Hansen, A., Marcussen, N., Thiesson, H.C., Bergmann, C., and Larsen, M.J. (2022). Detection of DZIP1L mutations by whole-exome sequencing in consanguineous families with polycystic kidney disease. Pediatr Nephrol. 10.1007/s00467-022-05441-4.

27. Schobers, G., Schieving, J.H., Yntema, H.G., Pennings, M., Pfundt, R., Derks, R., Hofste, T., de Wijs, I., Wieskamp, N., van den Heuvel, S., et al. (2022). Reanalysis of exome negative patients with rare disease: a pragmatic workflow for diagnostic applications. Genome Med *14*, 66. 10.1186/s13073-022-01069-z.

28. Jackson, A., Banka, S., Stewart, H., Robinson, H., Lovell, S., and Clayton-Smith, J. (2021). Recurrent KCNT2 missense variants affecting p.Arg190 result in a recognizable phenotype. Am J Med Genet A *185*, 3083-3091. 10.1002/ajmg.a.62370.

29. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). 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 *17*, 405-424. 10.1038/gim.2015.30.

30. Snijders Blok, L., Verseput, J., Rots, D., Venselaar, H., Innes, A.M., Stumpel, C., Õunap, K., Reinson, K., Seaby, E.G., McKee, S., et al. (2023). A clustering of heterozygous missense variants in the crucial chromatin modifier WDR5 defines a new neurodevelopmental disorder. HGG Adv *4*, 100157. 10.1016/j.xhgg.2022.100157.

31. Sengoku, T., and Yokoyama, S. (2011). Structural basis for histone H3 Lys 27 demethylation by UTX/KDM6A. Genes Dev *25*, 2266-2277. 10.1101/gad.172296.111.

32. The UniProt Consortium (2016). UniProt: the universal protein knowledgebase. Nucleic acids research *45*, D158-D169. 10.1093/nar/gkw1099.

33. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583-589. 10.1038/s41586-021-03819-2.

34. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res *50*, D439-d444. 10.1093/nar/gkab1061.

35. Krieger, E., and Vriend, G. (2014). YASARA View—molecular graphics for all devices—from smartphones to workstations. Bioinformatics *30*, 2981-2982. 10.1093/bioinformatics/btu426.

36. Lelieveld, S.H., Wiel, L., Venselaar, H., Pfundt, R., Vriend, G., Veltman, J.A., Brunner, H.G., Vissers, L., and Gilissen, C. (2017). Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes. American journal of human genetics *101*, 478-484. 10.1016/j.ajhg.2017.08.004.

37. Copur, Ö., and Müller, J. (2013). The histone H3-K27 demethylase Utx regulates HOX gene expression in Drosophila in a temporally restricted manner. Development *140*, 3478-3485. 10.1242/dev.097204.

38. Agger, K., Cloos, P.A., Christensen, J., Pasini, D., Rose, S., Rappsilber, J., Issaeva, I., Canaani, E., Salcini, A.E., and Helin, K. (2007). UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. Nature *449*, 731-734. 10.1038/nature06145.

39. Xiang, Y., Zhu, Z., Han, G., Lin, H., Xu, L., and Chen, C.D. (2007). JMJD3 is a histone H3K27 demethylase. Cell Res *17*, 850-857. 10.1038/cr.2007.83.

40. Siegel, R.W., and Hall, J.C. (1979). Conditioned responses in courtship behavior of normal and mutant Drosophila. Proc Natl Acad Sci U S A *76*, 3430-3434. 10.1073/pnas.76.7.3430.

41. Kummeling, J., Stremmelaar, D.E., Raun, N., Reijnders, M.R.F., Willemsen, M.H., Ruiterkamp-Versteeg, M., Schepens, M., Man, C.C.O., Gilissen, C., Cho, M.T., et al. (2020). Characterization of SETD1A haploinsufficiency in humans and Drosophila defines a novel neurodevelopmental syndrome. Mol Psychiatry. 10.1038/s41380-020-0725-5.

42. Shaw, P.J., Cirelli, C., Greenspan, R.J., and Tononi, G. (2000). Correlates of sleep and waking in Drosophila melanogaster. Science *287*, 1834-1837. 10.1126/science.287.5459.1834.

43. Huber, R., Hill, S.L., Holladay, C., Biesiadecki, M., Tononi, G., and Cirelli, C. (2004). Sleep homeostasis in Drosophila melanogaster. Sleep *27*, 628-639. 10.1093/sleep/27.4.628.

44. Schindelin, J., Rueden, C.T., Hiner, M.C., and Eliceiri, K.W. (2015). The ImageJ ecosystem: An open platform for biomedical image analysis. Mol Reprod Dev *82*, 518-529. 10.1002/mrd.22489.

45. Chubak, M.C., Nixon, K.C.J., Stone, M.H., Raun, N., Rice, S.L., Sarikahya, M., Jones, S.G., Lyons, T.A., Jakub, T.E., Mainland, R.L.M., et al. (2019). Individual components of the SWI/SNF chromatin remodelling complex have distinct roles in memory neurons of the Drosophila mushroom body. Dis Model Mech *12*. 10.1242/dmm.037325.

46. Bramswig, N.C., Lüdecke, H.J., Hamdan, F.F., Altmüller, J., Beleggia, F., Elcioglu, N.H., Freyer, C., Gerkes, E.H., Demirkol, Y.K., Knupp, K.G., et al. (2017). Heterozygous HNRNPU variants cause early onset epilepsy and severe intellectual disability. Hum Genet *136*, 821-834. 10.1007/s00439-017-1795-6.

47. Wiel, L., Baakman, C., Gilissen, D., Veltman, J.A., Vriend, G., and Gilissen, C. (2019). MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. Human mutation *40*, 1030-1038. 10.1002/humu.23798.

48. Abou Tayoun, A.N., Pesaran, T., DiStefano, M.T., Oza, A., Rehm, H.L., Biesecker, L.G., and Harrison, S.M. (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Human mutation *39*, 1517-1524. 10.1002/humu.23626.

49. Rots, D., Chater-Diehl, E., Dingemans, A.J.M., Goodman, S.J., Siu, M.T., Cytrynbaum, C., Choufani, S., Hoang, N., Walker, S., Awamleh, Z., et al. (2021). Truncating SRCAP variants outside the Floating-Harbor syndrome locus cause a distinct neurodevelopmental disorder with a specific DNA methylation signature. American journal of human genetics *108*, 1053-1068. 10.1016/j.ajhg.2021.04.008.

50. Vissers, L., Kalvakuri, S., de Boer, E., Geuer, S., Oud, M., van Outersterp, I., Kwint, M., Witmond, M., Kersten, S., Polla, D.L., et al. (2020). De Novo Variants in CNOT1, a Central Component of the CCR4-NOT Complex Involved in Gene Expression and RNA and Protein Stability, Cause Neurodevelopmental Delay. American journal of human genetics *107*, 164-172. 10.1016/j.ajhg.2020.05.017.

51. Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J.D., Bass, N., Bigdeli, T.B., Breen, G., Bromet, E.J., et al. (2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature *604*, 509-516. 10.1038/s41586-022-04556-w.

52. Howrigan, D.P., Rose, S.A., Samocha, K.E., Fromer, M., Cerrato, F., Chen, W.J., Churchhouse, C., Chambert, K., Chandler, S.D., Daly, M.J., et al. (2020). Exome sequencing in schizophrenia-affected parent-offspring trios reveals risk conferred by protein-coding de novo mutations. Nat Neurosci *23*, 185-193. 10.1038/s41593-019-0564-3.

53. Vermeulen, K., Staal, W.G., Janzing, J.G., van Bokhoven, H., Egger, J.I.M., and Kleefstra, T. (2017). Sleep Disturbance as a Precursor of Severe Regression in Kleefstra Syndrome Suggests a Need for Firm and Rapid Pharmacological Treatment. Clin Neuropharmacol *40*, 185-188. 10.1097/wnf.0000000000000226.

54. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443. 10.1038/s41586-020-2308-7.

55. Jacquemont, S., Coe, B.P., Hersch, M., Duyzend, M.H., Krumm, N., Bergmann, S., Beckmann, J.S., Rosenfeld, J.A., and Eichler, E.E. (2014). A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. American journal of human genetics *94*, 415-425. 10.1016/j.ajhg.2014.02.001.

56. Wigdor, E.M., Weiner, D.J., Grove, J., Fu, J.M., Thompson, W.K., Carey, C.E., Baya, N., van der Merwe, C., Walters, R.K., Satterstrom, F.K., et al. (2022). The female protective effect against autism spectrum disorder. Cell Genomics *2*, 100134. https://doi.org/10.1016/j.xgen.2022.100134.

57. Niemi, M.E.K., Martin, H.C., Rice, D.L., Gallone, G., Gordon, S., Kelemen, M., McAloney, K., McRae, J., Radford, E.J., Yu, S., et al. (2018). Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature *562*, 268-271. 10.1038/s41586-018-0566-4.

58. Parenti, I., Rabaneda, L.G., Schoen, H., and Novarino, G. (2020). Neurodevelopmental Disorders: From Genetics to Functional Pathways. Trends Neurosci *43*, 608-621. 10.1016/j.tins.2020.05.004.

59. Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., et al. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature *586*, 757-762. 10.1038/s41586-020-2832-5.

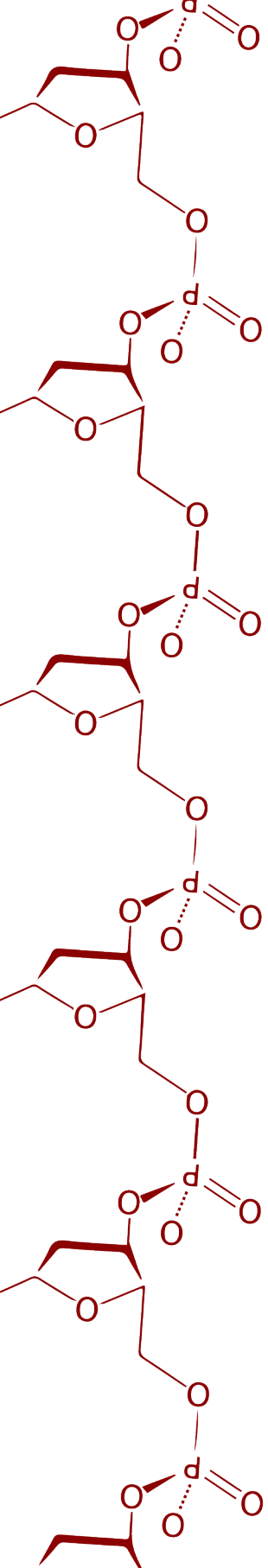
60. Guo, T., Han, X., He, J., Feng, J., Jing, J., Janečková, E., Lei, J., Ho, T.V., Xu, J., and Chai, Y. (2022). KDM6B interacts with TFDP1 to activate P53 signaling in regulating mouse palatogenesis. Elife *11*. 10.7554/eLife.74595.

61. Wang, W., Cho, H., Lee, J.W., and Lee, S.K. (2022). The histone demethylase Kdm6b regulates subtype diversification of mouse spinal motor neurons during development. Nat Commun *13*, 958. 10.1038/s41467-022-28636-7.

62. Wang, Y., Khandelwal, N., Liu, S., Zhou, M., Bao, L., Wang, J.E., Kumar, A., Xing, C., Gibson, J.R., and Wang, Y. (2022). KDM6B cooperates with Tau and regulates synaptic plasticity and cognition via inducing VGLUT1/2. Mol Psychiatry *27*, 5213-5226. 10.1038/s41380-022-01750-0.

63. Gao, Y., Aljazi, M.B., and He, J. (2022). Kdm6b Haploinsufficiency Causes ASD/ADHD-Like Behavioral Deficits in Mice. Frontiers in behavioral neuroscience *16*, 905783. 10.3389/fnbeh.2022.905783.

64. Faundes, V., Goh, S., Akilapa, R., Bezuidenhout, H., Bjornsson, H.T., Bradley, L., Brady, A.F., Brischoux-Boucher, E., Brunner, H., Bulk, S., et al. (2021). Clinical delineation, sex differences, and genotype-phenotype correlation in pathogenic KDM6A variants causing X-linked Kabuki syndrome type 2. Genetics in medicine : official journal of the American College of Medical Genetics *23*, 1202-1210. 10.1038/s41436-021-01119-8.



Chapter 4:

Pathogenic variants in *KMT2C* result in a neurodevelopmental disorder distinct from Kleefstra and Kabuki syndromes

Published in: American Journal of Human Genetics. 2024 Aug 8;111(8):1626-1642.

Authors

Dmitrijs Rots*, Sanaa Choufani*, Victor Faundes*, Alexander J.M. Dingemans, Shelagh Joss, Nicola Foulds, Elizabeth A. Jones, Sarah Stewart, Pradeep Vasudevan, Tabib Dabir, Soo-Mi Park, Rosalyn Jewell, Natasha Brown, Lynn Pais, Sébastien Jacquemont, Khadijé Jizi,Conny M.A. van Ravenswaaij-Arts, Hester Y. Kroes, Constance T. R. M. Stumpel, Charlotte W. Ockeloen, Illja J. Diets, Mathilde Nizon, Marie Vincent, Benjamin Cogné, Thomas Besnard, Marios Kambouris, Emily Anderson, Elaine H Zackai, Carey McDougall, Sarah Donoghue, Anne O'Donnell-Luria, Zaheer Valivullah, Melanie O'Leary, Siddharth Srivastava, Heather Byers, Nancy Leslie, Sarah Mazzola, George E. Tiller, Moin Vera, Joseph J. Shen, Richard Boles, Vani Jain, Elise Brischoux-Boucher, Esther Kinning, Brittany N. Simpson, Jacques C. Giltay, Jacqueline Harris, Boris Keren, Anne Guimier, Pierre Marijon, Bert B.A. de Vries, Constance S. Motter, Bryce A. Mendelsohn, Samantha Coffino, Erica H. Gerkes, Alexandra Afenjar, Paola Visconti, Elena Bacchelli, Elena Maestrini, Andree Delahaye-Duriez, Catherine Gooch, Yvonne Hendriks, Hieab Adams, Christel Thauvin-Robinet, Sarah Josephi-Taylor, Marta Bertoli, Michael J. Parker, Julie W. Rutten, Oana Caluseriu, Hilary J. Vernon, Jonah Kaziyev, Jia Zhu, Jessica Kremen, Zoe Frazier, Hailey Osika, David Breault, Sreelata Nair, Suzanne M. E. Lewis, Fabiola Ceroni, Marta Viggiano, Annio Posar, Helen Brittain, Traficante Giovanna, Gori Giulia, Lina Quteineh, Russia Ha-Vinh Leuchter, Evelien Zonneveld-Huijssoon, Cecilia Mellado, Isabelle Marey, Alicia Coudert, Mariana Inés Aracena Alvarez, Milou G. P. Kennis, Arianne Bouman, Maian Roifman, María Inmaculada Amorós Rodríguez, Juan Dario Ortigoza-Escobar, Vivian Vernimmen, Margje Sinnema, Rolph Pfundt, Han G. Brunner, Lisenka E.L.M. Vissers, Tjitske Kleefstra**, Rosanna Weksberg**, Siddharth Banka**

*,** These authors contributed equally to this work

Abstract

Trithorax-related H3K4 methyltransferases, KMT2C and KMT2D, are critical epigenetic modifiers. Haploinsufficiency of *KMT2C* was only recently recognised as a cause of neurodevelopmental disorder (NDD), so the clinical and molecular spectrum of the *KMT2C*-related (NDD) (now designated as Kleefstra syndrome 2) are largely unknown. We ascertained 98 individuals with rare *KMT2C* variants, including 75 with protein truncating variants (PTVs). Notably, ~15% *KMT2C* PTVs were inherited. Although the most highly expressed *KMT2C* transcript consists of only the last four exons, pathogenic PTVs were found in almost all the exons of this large gene. We found that *KMT2C* variant interpretation can be challenging due to segmental duplications and clonal hematopoesis induced artefacts. Using samples from 27 affected individuals, divided into discovery and validation cohorts, we generated a moderate strength disorder-specific *KMT2C* DNAm signature and demonstrate its utility in classifying non-truncating variants. Based on 81 individuals with pathogenic/likely pathogenic variants, we demonstrate that the *KMT2C*-related NDDis characterized by developmental delay, intellectual disability, behavioral and psychiatric problems, hypotonia, seizures, short stature, and other comorbidities. Using facial module of PhenoScore on photographs of 34 affected individuals, we show that the *KMT2C*-related facial gestalt is significantly different from general neurodevelopmental disorder population. Finally, using PhenoScore and DNAm signatures, we demonstrate that the *KMT2C*-related NDDis clinically and epigenetically distinct from Kleefstra and Kabuki syndromes. Overall, we define the clinical features, the molecular spectrum, and the DNAm signature of the *KMT2C*-related NDD and prove them as distinct from Kleefstra and Kabuki syndromes highlighting the need to rename this condition.

Introduction

Mendelian disorders of epigenetic machinery are amongst the most common forms of neurodevelopmental disorders (NDDs)[1]. Humans possess six histone-3 lysine-4 (H3K4) methyltransferases that are divided into three sub-groups, including the trithorax-related subgroup which consists of KMT2C and KMT2D. These proteins are important components of the epigenetic machinery and are involved in spatiotemporal gene expression regulation[2,3]. KMT2C or KMT2D, together with WDR5, RBBP5, ASH2L, DPY30 (i.e., WRAD subunit), KDM6A, and other proteins, form the COMPASS complex[2] that mono- (H3K4me1)[4] and trimethylates H3K4 (H3K4me3)[5] at active chromatin sites of gene enhancers and promoters, respectively.

Heterozygous loss-of-function *KMT2D* (MIM: 602113) variants were identified to cause Kabuki syndrome type 1 (MIM: 147920) in 2010[6]. However, the consequences of germline *KMT2C* (MIM: 606833) variants in humans have been identified more recently. In a phenotype-led study we identified *de novo* loss-of-function *KMT2C* variants in individuals with clinical characteristics overlapping Kleefstra syndrome[3,7]. Kleefstra syndrome (MIM: 610253) is caused by haploinsufficiency of euchromatin histone methyltransferase 1 (*EHMT1*; MIM #610253) and is characterized by intellectual disability (ID), autism spectrum disorder (ASD), characteristic facial dysmorphisms, and other variable clinical features[8]. In an independent genotype-led study, we studied variants in histone lysine methyltransferases (KMTs) and demethylases (KDMs) in the Deciphering Developmental Disorders (DDD) cohort and discovered that *KMT2C* loss-of-function variants result in a neurodevelopmental phenotype with occasional physical anomalies[9]. However, the clinical and molecular spectrum of the *KMT2C*-related disorder is largely unknown, as only a few such individuals have been reported. Based on our experience, we hypothesized that the *KMT2C*-related NDD is a unique entity that is clinically and molecularly different from Kleefstra and Kabuki syndromes.

In this study, we provide comprehensive clinical, molecular and DNA methylation (DNAm) data for the *KMT2C*-related NDD based on a large previously unreported cohort of individuals, as well as demonstrate that this disorder is different from the molecularly-related Kleefstra and Kabuki type 1 syndromes.

Materials and Methods

Cohort recruitment

This study was approved by the institutional review boards of the South Manchester NHS REC, Radboudumc and the Hospital for Sick Children (Research Ethics Committee Approvals 11/H1003/3/AM02, 2011/188, and #1000038847, respectively). Individuals with rare heterozygous pathogenic (P) or likely pathogenic (LP) variants or variants of uncertain significance (VUS) in *KMT2C* were identified in clinical diagnostic settings (using standard chromosomal microarray, exome, or genome sequencing[10]) or from large NDD research cohorts (the Deciphering Developmental Disorders (DDD) study[11], 100,000 Genomes project[12], Simons Simplex collection (SSC)[13], and MSSNG[14]). Individuals with rare reported single nucleotide variants (SNVs), indels, and copy number variants (CNVs) in the *KMT2C* gene were included. However, CNVs were limited only to those with deletions <1Mb which did not affect other predicted haploinsufficient genes. Individuals with additional pathogenic variants in other genes were also included in the study but were excluded from the clinical feature frequency analysis.

All included individuals or their caregivers/legal representatives consented to participate in this research. Genetic and clinical data from individuals were collected via a customized proforma.

Variant analysis

The variants were annotated using the *KMT2C* MANE select[15] transcript (GenBank: NM_170606.3, GRCh37). All identified variants were re-classified according to the ACMG guidelines 2015[16] based on the clinical and molecular evidence obtained during this study. The variants and their interpretations were submitted to the ClinVar database (ClinVar accession numbers: SCV005044911-SCV005044983 and SCV005044991-SCV005045000). Only individuals with LP/P *KMT2C* variants without known pathogenic variants in other genes were included for further clinical feature analyses.

DNA methylation analysis

Sample processing, DNA methylation (DNAm) signature derivation and analysis were performed similarly as described before[17-19]. Briefly, DNA samples underwent bisulfite conversion using the EpiTect Bisulfite Kit (EpiTect PLUS Bisulfite Kit, QIAGEN) following the manufacturer's protocol. The converted DNAs were then analyzed on the Illumina Infinium Human Methylation EPIC V1 BeadChip (with ~850,000 CpG sites) at The Center for Applied Genomics (TCAG), Hospital for

Sick Children Research Institute, Toronto, Ontario, Canada. The affected individuals' and controls' samples were randomly positioned on the arrays.

DNAm analysis was performed using the minfi R package[20]. Briefly, the minfi package was used for data preprocessing, quality control, normalization, transformation to β values and blood cell composition estimation by Housman's method. Probes with a detection p-value >0.05 in >25% of the samples (911 probes), probes located near common polymorphic variants with minor allele frequencies >1% (166,596 probes), non-specific probes (35,613 probes), probes with raw β value equal to 0 or 1 in >25% of samples (240 probes), non-CpG probes (2,377 probes), and chrX and chrY probes (17,485 probes) were removed from the analysis. Additionally, 705 probes behaving like single nucleotide variants were removed using the MethylToSNP package[21]. In total, 645,199 CpG sites remained for the differential methylation analysis. All samples passed the quality control and were suitable for the analysis.

For the signature derivation, the remaining 645,199 CpG sites' β values were transformed to M values, and differentially methylated CpGs were identified by using a linear regression with monocyte count, batch and second principal components as covariates using the limma R package[22]. Only differentially methylated CpGs, with a 10% methylation difference ($|\Delta\beta|>0.10$) and false discovery rate (FDR)-adjusted p-values <0.05 were selected for the analysis. To remove false positive sites, we further excluded 6 CpGs with methylation β values in cases and controls that followed a SNP-like pattern. This resulted in a DNAm signature of 51 CpGs with $|\Delta\beta|>0.10$. The results were visualized through principal component analysis (PCA) and heatmap plots using the Qlucore Omics software.

Machine learning models and classifications

For the sample classification, we have developed a machine learning model – support vector machine (SVM) with linear kernel using *caret* R package[23] as described before[17]. The model was trained on the signature's CpG sites for the discovery cohorts (16 *KMT2C*-related NDD and 50 control samples). Because infants have noticeably different blood DNAm, 3 infant control samples were added to the controls during the model training, to increase model's specificity across different ages. Receiver operating characteristic analysis was used to select the optimal model using the largest value. The SVM model was set to the "probability" mode, so the model generated scores ranging between 0 and 1, where scores <0.25 were interpreted as "negative", >0.5 as "positive", and 0.25-0.5 as "intermediate".

The model's sensitivity and specificity were evaluated using the *KMT2C*-related NDD and control validation samples, as well as Kabuki type 1 and Kleefstra syndromes samples. Finally, the model was used to classify the 22 testing samples with *KMT2C* VUS.

To evaluate the overlap between *KMT2C*-related NDD DNAm changes with DNAm changes for other disorders of the epigenetic machinery, we analyzed all *KMT2C* samples on 17 other available DNAm signatures deployed on EpigenCentral portal[24], as described before[25]. To evaluate specificity of the *KMT2C*-related NDD DNAm signature, we have utilized six molecularly-confirmed Kabuki syndrome type 1 and six Kleefstra syndrome individuals, as well as 165 healthy controls.

Protein 3D structure analysis

KMT2C is a large (4911 amino acids) multidomain protein, so there is no solved protein structure of the whole KMT2C protein available. Therefore, for the analysis of possible missense variant effects, we have used: 1) solved structures of HMG box (PDB:2YUK), extended PHD6 (PDB:6MLC[26]), and SET (PDB:6KIW[27], 5F6K, 5F59[28]) domains; 2) homology models for FYR (based on PDB:2WZO[29]), and extended PHD2 (based on PDB:4NN2[30], 6U04[31], 7MJU[32], 7D8A[33]) domains; 3) *ab initio* protein models for all missense positions. The homology modeling, analysis, and visualization were performed using YASARA Structure software[34]. The *ab initio* protein models were generated by AlphaFold2[35] (multiple modeled overlapping protein fragments of maximal size 1500 amino acids) and downloaded from the AlphaFold Protein Structure Database[36]. The predicted effect was assessed by at least two different protein structures/models, where available (e.g., solved structure and *ab initio*; two different *ab initio* structures).

Clinical features analyses

Consent acquisition, as well as clinical and molecular description of the recruited individuals were provided by their physicians using a standardized proforma (**Table S1**). However, for statistical analyses, only individuals with LP/P *KMT2C* variants and without confirmed additional pathogenic variants in other genes were considered. For any feature, we excluded individuals with "UNKNOWN" coding as previously described[37]. WHO growth standards per age and sex groups were used to evaluate the growth. Absolute and relative frequencies (expressed as n[%]) were used for describing categorical variables, whereas medians (m) and interquartile ranges (IQR) were used for describing continuous variables. Chi-square/Fisher's exact and Mann–Whitney tests were applied to study categorical and continuous variables, respectively, and their associations with sex (females vs. males), variant

group (protein-altering [PAV] vs. protein-truncating variants [PTV]), and inheritance (inherited vs. *de novo* variants), respectively. Two-tailed, Bonferroni-adjusted p-value <0.05 was considered significant for all statistical analyses, which were carried out using the IBM SPSS Version 29 software.

Facial photo analysis

To identify whether *KMT2C*-related NDD has a facial gestalt, facial 2D photos from 29 out of the 34 individuals with a LP/P *KMT2C* variant were compared against 29 sex-, age- and ethnicity- matched individuals with NDDs as controls (sampling different controls 5 times) by using PhenoScore as described before[38]. Not all 34 individuals included in this study could be included in the PhenoScore analyses, because a matched control was not found for five individuals.

Briefly, the facial module of PhenoScore utilizes a state-of-the-art convolutional neural network used in facial recognition (QMagFace[39]) that recognizes facial features and allows for objectively evaluating and statistically comparing different NDD facial gestalts.

Similarly, by using PhenoScore, the photos from these *KMT2C*-related NDD individuals were compared against the sex, age, and ethnicity matched individuals with either *EHMT1* or *KMT2D* pathogenic variants to objectively evaluate whether the *KMT2C*-related NDD facial gestalt significantly differs from those with Kleefstra and Kabuki type 1 syndromes, respectively.

Results

KMT2C-related NDD cohort curation and identification of variant spectrum

To systematically analyze the genetic and clinical spectrum of the *KMT2C*-related NDD, we ascertained 98 individuals with rare reported *KMT2C* variants, irrespective of their phenotypes through research databases, international collaborations, and previously published individuals[3,7,9].

75 out of 98 individuals from 67 families had 61 distinct heterozygous *KMT2C* variants predicted to be protein truncating (PTVs) (21 nonsense, 27 frameshift, 5 splice site and 8 large deletions). These PTVs were classified as pathogenic (P) or likely pathogenic (LP) (**Figure 1**), because loss of function is the currently accepted disease mechanism[3,7,9] and the gene is intolerant to loss-of-function variants in

population (pLI=1; o/e = 0.08 (0.06 - 0.12))[40]. Of note, 52 of these P/LP variants were *de novo*, 11 were inherited (6 maternal and 5 paternal), and in 12 individuals the inheritance was unknown. Importantly, *KMT2C* is a large gene consisting of 59 exons (NM_170606.3). Ensembl documents 70 transcripts for *KMT2C*, 31 of which are annotated as protein coding: ranging in amino acid length from 4968 (ENST00000682283.1) to 83 (ENST00000684278.1). The MANE select transcript (canonical isoform NM_170606.3/ENST00000262189.11) encodes a protein of 4911 amino acids with mean expression across all adult tissues in the GTEx database of only 0.83 transcripts per million (TPM), while the most highly expressed transcript (5.08 TPM) is ENST00000485655.2 consisting of only the last four exons of the gene[41]. This, however, might be related to the 3' bias of the short-read GTEx data[42]. In our cohort, we found the PTVs are distributed throughout the gene, with the most 5' and the most distal PTVs being located in the 3rd and 57th *KMT2C* exons, suggesting that in spite of apparently relatively low expression of the canonical (MANE) transcript in adult tissues, PTVs across the gene are likely to be pathogenic. We also did not observe specific clustering of PTVs in the gnomAD database (**Figure S1**), but, based on the skewed variant allele frequency, we noted that some of the observed PTVs are either artifacts due to segmental duplications (e.g., c.2710C>T p.(Arg904Ter) and c.1173C>A p.(Cys391Ter)), or somatic variants (e.g., c.6415C>T p.(Arg2139Ter)), as *KMT2C* is known clonal hematopoiesis driver gene[43].

We observed a relatively high proportion of individuals with pathogenic CNVs in our cohort (n=11, 14%; 5/11 are *de novo*). We, therefore, examined the genomic architecture and found that the *KMT2C* contains several segmental duplications with high homology (>98%) to sequences elsewhere in the genome (**Figure 1B**) making this region a "hot-spot" for structural rearrangements[44]. We examined the Database for Genomic Variants (DGV)[45] and gnomAD[40] for structural variants affecting this coding part of the *KMT2C* in the control population in the UCSC genome browser[46] and identified in total 26 partial or intragenic gains and only a single deletion. This suggests that the partial/intragenic gains are unlikely to be pathogenic (if not leading to frameshift). We, however, cannot exclude frequent occurence of artifacts due to the limitations of current technologies in the regions affected by segmental duplications.

In contrast, non-PTVs have not been classified as pathogenic, so 23 out of 98 individuals from 22 families had 22 VUS (18 missense and 2 splice variants and 2 large duplications). Of those, 10 were *de novo*, 6 were inherited, and 6 were of unknown inheritance.

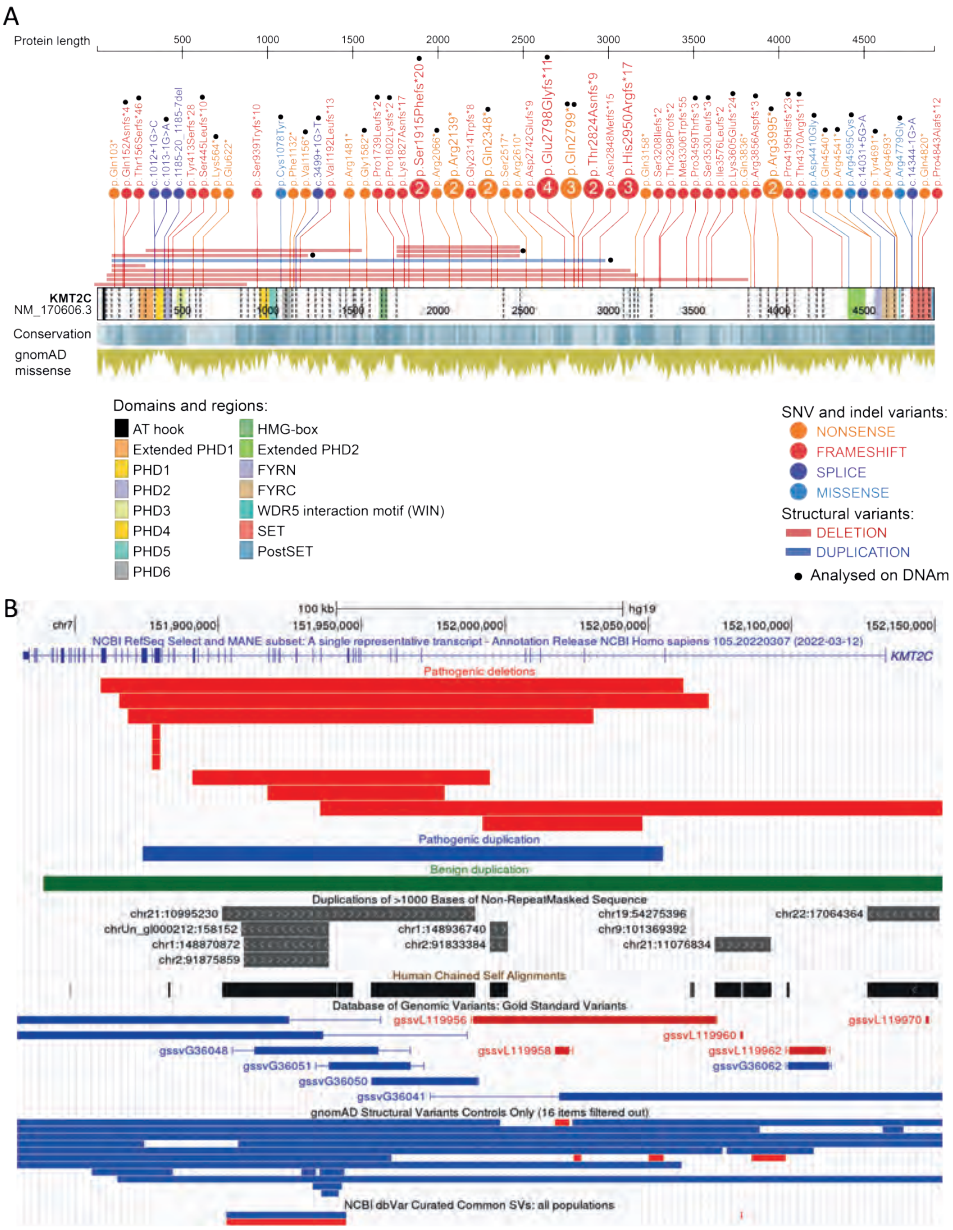


Figure 1. *KMT2C* variant spectrum among the recruited individuals and population.

A. *Likely pathogenic and pathogenic point and copy number variants identified in this study are shown on the linear structure of the KMT2C protein.* ***B.*** *Likely pathogenic and pathogenic, as well as benign copy number variants identified in this study in comparison shown on the KMT2C genomic region with segmental duplications (grey), as well as populational deletions and duplications from DGV and gnomAD databases.*

Collectively, these data show that P/LP truncating *KMT2C* variants are spread across the gene, affecting only MANE or multiple transcripts, and that they can be inherited in some cases. Additionally, clinical classification of *KMT2C* non-truncating variants is currently challenging.

Protein truncating *KMT2C* variants result in a DNAm signature in peripheral blood

We and others have previously shown that pathogenic variants in genes encoding components of epigenetic regulators are associated with genome-wide DNAm changes[47] from which disease-associated DNAm signatures can be derived. Therefore, we performed genome-wide methylation screening on peripheral blood-derived DNA from 16 *KMT2C*-related NDD individuals (discovery cohort) with pathogenic *KMT2C* PTVs and 50 controls using Illumina methylation Epic arrays. We identified 51 significant differentially methylated CpG sites at $|\Delta\beta|>0.10$, and FDR-corrected $p<0.05$ representing the DNAm signature of the *KMT2C*-related NDD (for simplicity, further called *KMT2C* DNAm signature) (**Table S3**). Most of the signature's CpGs were hypomethylated (42/51, 82%). Interestingly, 4/51 (8%) of the signature's CpGs sites mapped to two CpG islands of the WT1 gene (#MIM 607102), which were mostly hypomethylated (**Figure 2C**).

Based on the derived *KMT2C* DNAm signature, we were able to discriminate the discovery *KMT2C* cases from healthy controls based on PCA and heatmap plots (**Figures 2A and 2B**, respectively). Next, we trained an SVM model based on the discovery cohort and tested the sensitivity and specificity of the *KMT2C* DNAm signature using 165 controls without known developmental disorders, and 11 additional validation cases with P/LP PTV *KMT2C* variants. On this SVM model, all controls were classified as negative (SVM values <0.25,100% specificity), and 9/11 validation cases were classified as positive (SVM values >0.5, 82% sensitivity) (**Figure 2D**). Two validation cases were classified negatively: individual #48 presented with typical clinical features and has a *de novo* pathogenic frameshift variant c.13107_13108dup p.(Thr4370Argfs*11); individual #52 is mildly affected with multiple affected children and has a small likely pathogenic *KMT2C* exon 36 and 37 deletion resulting in frameshift. However, we cannot exclude that these cases are high-level mosaics in blood.

Next, to test whether *KMT2C*-related NDD affected individuals share the DNAm signature with other conditions, we analyzed all available samples with a LP/P *KMT2C* variant on 17 available DNAm signatures deployed at EpigenCentral[24]. All *KMT2C*-related NDD samples were classified negatively on all signatures **(Figure 2E)**.

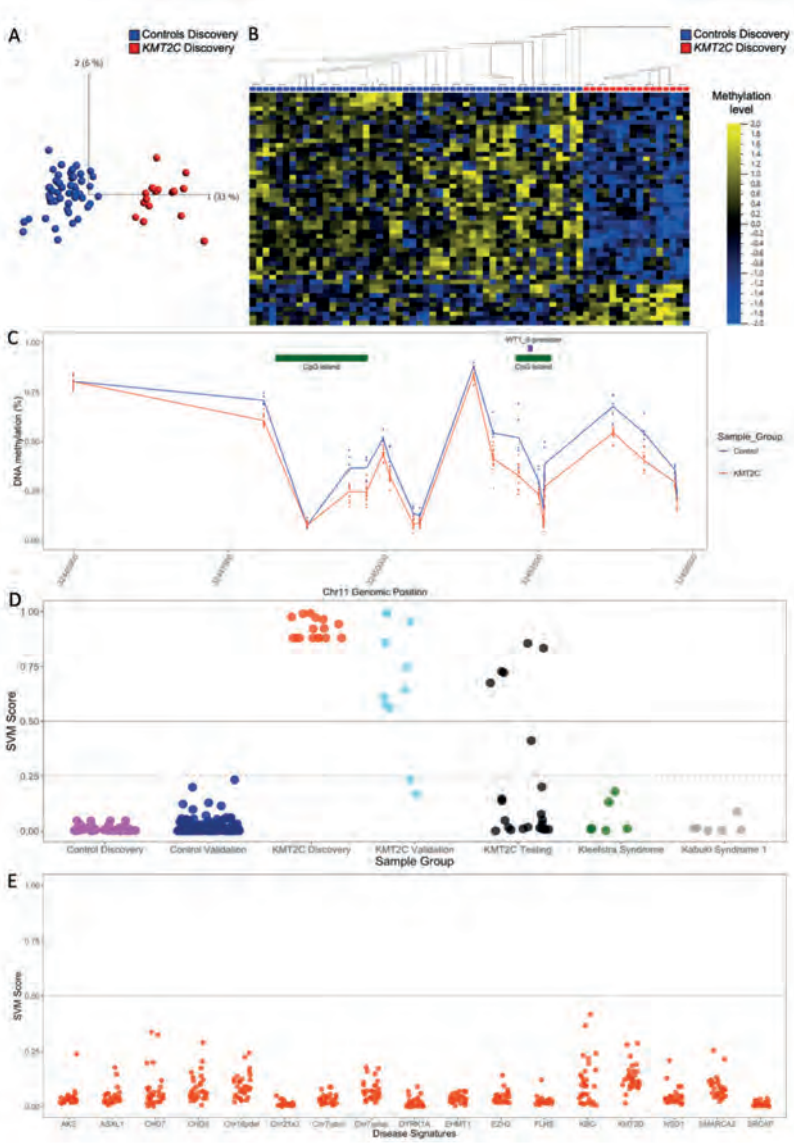


Figure 2. DNAm signature classification results.

A. *PCA plot for the discovery cohort of 16 KMT2C-related NDD cases (red) and 50 controls (blue).* ***B.*** *Heatmap plot with hierarchical clustering for the discovery cohort of 16 KMT2C-related NDD cases (red) and 50 controls (blue), with hypermethylated sites shown in yellow and hypomethylated ones in blue.* ***C.*** *Differentially methylated region which maps to several WT1 CpG islands (green) and one of the promoters (purple) (data obtained from the UCSC genome browser) where each dot shows methylation level at a CpG site in the region for each discovery cohort sample, and the line depicts the mean methylation for controls (blue) and cases (red).* ***D.*** *SVM model classification results for different groups: 16 discovery KMT2C-related NDD cases (red), 50 controls (magenta); validation KMT2C-related NDD cases (light blue) and their matched controls (dark blue); KMT2C VUS (black); Kleefstra syndrome individuals (green) and Kabuki syndrome individuals (grey).* ***E.*** *KMT2C-related NDD classification results against 17 other DNAm signatures.*

These results show that significant changes in methylation patterns of DNA derived from the peripheral blood of individuals with P/LP *KMT2C* variants exist, and that these changes can be used to generate a moderate-effect DNAm signature that does not overlap with the other known disorder-specific DNAm signatures.

DNAm signature can be used to re-interpret *KMT2C* VUS

Next, we explored if DNAm signature could enable classification of the *KMT2C* VUS by testing DNA samples from 22/23 individuals (20 unrelated and one proband-father pair) with 21 distinct *KMT2C* VUS on the *KMT2C* DNAm signature-derived SVM model (one individual with VUS c.14501T>C p.(Val4834Ala) was not available for testing). The SVM model resulted in classification of 5/21 VUS as positive for the *KMT2C* DNAm signature. One out of 21 VUS was classified as intermediate (SVM score 0.25-0.5), and the rest were classified as negative on the *KMT2C* DNAm signature. These results supported re-classification of 6/21 variants as LP and 13/21 variants as benign (described below), while 2/21 variants remained in the VUS category (one additional VUS was unavailable for testing).

Next, we examined the basis of pathogenicity of the PAVs in this cohort. The details for each variant classification and the criteria applied utilizing various evidence are described in **Table S4**. All six variants reclassified as LP/P were classified as positive (or intermediate) on the DNAm signature. Out of the six VUS reclassified as LP/P, two (*de novo*) variants were (re)interpreted as PTVs: duplication of exons 3-38 which is predicted to result in frameshift, if present in tandem, so based on the DNAm signature classification we concluded that the duplication most likely is indeed in tandem; one variant was initially reported from trio exome sequencing as *de novo* missense c.3499G>A p.(Asp1167Asn), located outside of known domains, but the signature analysis aided the reanalysis of the variant which was later confirmed by Sanger sequencing as c.3499+1G>T p.?. The remaining four LP PAVs (c.3233G>A p.(Cys1078Tyr), c.13229A>G p.(Asp4410Gly), c.13783C>T p.(Arg4595Cys) and c.14335C>G p.(Arg4779Gly)) are absent from gnomAD, predicted as pathogenic by *in silico* tools (REVEL[48] > 0.65 and AlphaMissense[49] > 0.56), and all are located in well-known functional domains of the protein (PHD6, ePHD2, FYRN and SET domains, respectively). Moreover, these variants are predicted to disrupt the 3D structure of the respective domains (**Figure S3**), except for the c.13229A>G p.(Asp4410Gly) for which no reliable 3D structure or model was available.

In contrast, all 11/13 variants classified as benign are missense and are negative on the DNAm signature, the majority are not classified as pathogenic by *in silico* tools and are not predicted to affect protein 3D structure and/or located outside

of the known functional domain's (largely in disordered regions) (**Table S4**). Only three benign PAVs occurred *de novo*. Additionally, one splicing variant c.1735+2dup was inherited from father, predicted to result in inframe skipping of exon 12 that encodes a disordered region of the protein and was classified as negative by the DNAm signature, so the variant was classified as benign. Similarly, one benign partial *KMT2C* duplication (1-55 exons) with unknown inheritance was absent from population databases, but similar *KMT2C* 5' partial duplications are common in the general population (**Figure 1B**), and it is also classified as negative on the DNAm signature.

Finally, three missense VUS remained classified as VUS due conflicting or insufficient evidence. One of the missense variants (c.14501T>C p.(Val4834Ala)) is located in the SET domain, is predicted as pathogenic by *in silico* tools (REVEL = 0.8 and AlphaMissense = 0.99) and was predicted to disrupt the domain's 3D structure (**Table S2**), but the sample was not available for DNAm testing. Two missense variants (c.9773A>C p.(His3258Pro) and c.13298C>T p.(Ala4433Val)) were classified negatively on the DNAm signature but had additional significant evidence for pathogenicity: both variants are *de novo*, absent in gnomAD and predicted to be possibly disruptive by *in silico* tools and/or 3D protein analysis (**Table S2, S4**).

These results show that the *KMT2C* DNAm signature can be used to reclassify most, but not all VUS in the gene and that PAVs disrupting functional domains are more likely to be P/LP.

KMT2C loss-of-function variants result in a phenotypically heterogenous NDD syndrome

In total, we identified 81 individuals from 73 families with P/LP *KMT2C* variants. We gathered detailed clinical information on these individuals to understand the clinical spectrum of *KMT2C*-related NDD. Out of 81 individuals, 6 had a second molecular diagnosis due to a pathogenic variant in another NDD-associated gene. To understand the clinical consequences of P/LP *KMT2C* variants, we analyzed the clinical features of 75 individuals with P/LP *KMT2C* variants without any additional known P/LP variants in other NDD genes (**Table 1, S1**).

In this cohort, 45.3% (n =34) were females, and the median age at last examination was 11 years (with IQR 5.17;20). The median weight at the last reported time was -1.01 SD (-1.94;-0.21), the median height was -1.56 SD (-2.32;-1.19), and the median head circumference (HC) was -0.44 SD (-1.65;0.51). Importantly, short stature (≤-2SD) was reported in ~54% of individuals, and 5 of them have received

recombinant human growth hormone (rhGH) with an apparently good response. Also, 9/53 (~17%) of individuals have displayed microcephaly (≤-2SD) at least on one available measurement. These findings suggest *KMT2C*-related NDD is mainly an undergrowth condition.

Neurodevelopmental problems are the most common features in this cohort. Although gross motor delay was described in 87% of individuals, all individuals older than 3.5 years of age had achieved independent walking. Speech delay was reported in 80% of individuals, and ~15% developed mutism at some point in their lifetime. Developmental regression was reported in 6 individuals. ID was reported in ~86% of individuals with half being classified as having moderate or severe ID. Importantly, a proportion of individuals had normal cognitive abilities (~14%; all with *de novo* variants) and some of them were tested due to non-NDD phenotypes (e.g., short stature, seizures). However, we cannot exclude mosaicism in these cases. Furthermore, behavioral, and psychiatric problems were reported in most individuals: features or diagnosis of ASD or ADHD were reported in ~79% and 61% of individuals, respectively; aggressive (both hetero- and self-) behavior was reported in 18% and obsessive/compulsive behavior – in ~15%.

Abnormalities affecting the central nervous system (~34.5%; e.g., ventriculomegaly, white-matter anomalies, syringomyelia) and the cardiovascular system (~24%; e.g., septal defects, valvular anomalies) were frequent. Refractive errors (33%) and hearing loss (~29%) were the most frequent sensory system problems. Recurrent infections (~26%), central/obstructive sleep apnea (~17%), seizures (~15%), kyphosis and/or scoliosis (~16%), and constipation (~16%) were some of the other most significant and frequently encountered medical issues (**Table 1**). Feeding difficulty was the most frequent neonatal finding (49%), followed by neonatal hypotonia (~29%).

Next, we compared the frequencies of clinical features between male and female individuals, type of variants (PAVs vs. PTVs), and their inheritance (inherited vs. *de novo*). None of the features was significant after correction for multiple testing ($p_{adj}>0.05$).

Table 1. Frequencies or distributions of clinical findings of *KMT2C*-related NDD.

Clinical finding [a]	**Frequency or distribution**
Sex (Males/Females, n[%]) (n=75)	41(54.7)/34 (45.3)
Age at last examination (years, m[IQR]) (n=71)	11.33 (5.17;20)
Craniofacial dysmorphisms (n[%]) (n=70)	63 (90)
Antenatal and neonatal features	
Abnormal pregnancy findings (n[%]) (n=61)	14 (23)
Small / Large for gestational age (n[%]) (n=48)	9 (18.8) / 2 (4.2)
Primary Microcephaly/Primary Macrocephaly (n[%]) (n=23)	3 (13) / 1 (4.3)
Neonatal hypotonia (n[%]) (n=52)	15 (28.8)
Neonatal feeding difficulties (n[%]) (n=51)	25 (49)
Neurological and developmental features	
Gross motor delay (n[%]) (n=69)	60 (87)
Fine motor delay (n[%]) (n=48)	38 (79.2)
Speech delay (n[%]) (n=65)	52 (80)
Mutism (n[%]) (n=47)	7 (14.9)
Developmental regression (n[%]) (n=45)	6 (13.3)
Cognitive impairment (n[%]) (n=69) - Mild / Moderate / Severe (n[%])(n=48)	59 (85.5) 24 (50) / 12 (25) / 12 (25)
Features or diagnosis of ASD (n[%]) (n=61)	48 (78.9)
Features or diagnosis of ADHD (n[%]) (n=51)	31 (60.8)
Seizures history (n[%]) (n=66)	10 (15.2)
Hypotonia (n[%]) (n=57)	19 (33.3)
CNS anomalies/abnormalities (n[%]) (n=29)	10 (34.5)
Other neurological issues (n[%]) (n=61) - Abnormal gait (n[%]) - Other (n[%])	14 (22.9) 8 (13) 9 (15)
Other behavioral/psychiatric problems (n[%]) (n=61)- Hetero/self-aggressive behaviour (n[%]) - Obsessive/compulsive behaviour (n[%]) - Other (n[%])	39 (63.9) 11 (18) 9 (14.8) 28 (45.9)
Endocrine anomalies (n[%]) (n=71) - Short stature* (n[%]) - Hypothyroidism (n[%])	40 (56.3) 39 (54.9) 3 (4)
Gastrointestinal/Nutritional anomalies (n[%]) (n=43)- Constipation(n[%]) - Other (n[%])	23 (53.5) 6 (14) 21 (48.8)
Ophthalmological problems (n=60) - Refractive error (n[%]) - Strabismus (n[%]) - Other (n[%])	32 (53.3) 20 (33.3) 12 (20) 4 (6.7)
Hearing Impairment (n[%]) (n=58)	17 (29.3)

Table 1. Continued

Clinical finding [a]	Frequency or distribution
Musculoskeletal anomalies (n[%]) (n=64)-	32 (50)
Kyphosis and/or Scoliosis (n[%])	10 (15.6)
- Joint hypermobility (n[%])	5 (7.8)
- Other (n[%])	25 (39.1)
Ectodermal anomalies (n[%]) (n=41)	16 (39)
- Hypertrichosis	4 (9.8)
- Other (n[%])	12 (29.3)
Immunological anomalies (n[%]) (n=43)	15 (34.8)
- Recurrent infections (n[%])	11 (25.6)
- Other (n[%])	4 (9.3)
Respiratory anomalies (n[%]) (n=42)	14 (33.3)
- Central/obstructive sleep apnea (n[%])	7 (16.7)
- Asthma (n[%])	6 (14.3)
- Other (n[%])	2 (4.8)
Palate anomalies (n[%]) (n=37)	12 (32.4)
- High/narrow palate (n[%])	11 (29.7)
- Cleft lip/palate (n[%])	1 (2.7)
Cardiovascular anomalies (n[%]) (n=54)	13 (24.1)
- Septal defects (n[%])	6 (11.1)
- Conduction disorder (n[%])	6 (11.1)
- Valvular anomalies (n[%])	4 (7.4)
- Other (n[%])	3 (5.6)
Genitourinary anomalies (n[%]) (n=39)	8 (20.5)
- Uni/bilateral inguinal hernia (n[%])	2 (5.1)
- Other (n[%])	7 (17.9)
Dental anomalies (n[%]) (n=70)	13 (18.8)
- Dental crowding (n[%])	6 (8.6)
- Prominent upper incisors (n[%])	6 (8.6)
- Other (n[%])	1 (1.4)

[a] *The number of responders are detailed for every feature, and their frequencies/distributions were calculated according to that number. Individuals with another NDD were excluded.*

** at any momemt documented <-2SD.ADHD=Attention Deficit/Hyperactivity Disorder; ASD=Autism Spectrum Disorder; CNS=Central nervous system; IQR=Interquartile range; m=median; SD=Standard deviation*

Craniofacial dysmorphisms were described in 90% of individuals with P/LP *KMT2C* variants in our cohort. Photographs of 34 individuals were available in this cohort, with 13 consenting for publishing (**Figure 3**). We used these photographs to analyze the facial features of individuals with *KMT2C*-related NDD. The most frequent findings were frontal bossing/prominent-broad forehead, thick/prominent eyebrows, synophrys, down-slanted palpebral fissures, deep-set eyes, epicanthus, hypertelorism, midface retrusion, anteverted nares, thin vermillion of the upper lip, micrognathia, low-set ears, thick ear helices and posteriorly rotated ears. Although

all individuals had craniofacial dysmorphism, overall, we did not identify a clearly recognizable *KMT2C* facial gestalt.

Next, to objectively evaluate whether the *KMT2C*-related NDD has a specific facial gestalt, we used PhenoScore and compared the facial photographs of individuals with the *KMT2C*-related NDD with facial photographs of individuals in a general ID cohort. PhenoScore identified a specific *KMT*2C-related facial gestalt (AUC = 0.91, p<0.001) (**Figure S2**), indicating that the facial features the 29 individuals with a LP/P *KMT2C* variant for whom an age, sex and ethnicity matched NDD control was available are distinguishable from the general NDD population. The PhenoScore identified periorbital and nasal regions as the most different (**Figure S2**), which fit the most common dysmorphic features described above.

Collectively, these observations show that *KMT2C*-related NDD is clinically a highly heterogeneous disorder characterized by intellectual disability, short stature, congenital anomalies, recurrent infections and craniofacial dysmorphism.

KMT2C-related NDD is distinct from Kleefstra and Kabuki syndromes

OMIM designates the *KMT2C*-related NDD as 'Kleefstra syndrome 2' (MIM: 617768) which may be interpreted that the *EHMT1* and *KMT2C* related disorders have significant clinical overlap. However, this overlap has not been examined previously. Also, *KMT2C* and *KMT2D* are part of the same gene family, but their overlap has not been examined previously. For this, firstly, we compared the facial features of individuals with *KMT2C*-related NDD to the photographs of 24 and 10 matched individuals with Kleefstra, and Kabuki type 1 syndromes, respectively. PhenoScore revealed significant differences between the *KMT2C*-related NDD facial gestalt and the facial gestalt of Kleefstra syndrome (AUC=1, p<0.001) and Kabuki syndrome type 1 (AUC=0.8, p=0.04).

DNAm signatures can also be used to delineate epigenetically distinct disorders[50,51]. Pathogenic loss-of-function variants in the *EHMT1*[52] or *KMT2D*[6] result in unique DNAm signatures[51,53]. We analyzed DNA samples from 6 individuals with Kleefstra and 6 individuals with Kabuki 1 syndromes on the *KMT2C* DNAm signature. Samples from Kleefstra and Kabuki type 1 syndromes were classified negatively by the SVM (**Figure 2D**;**Table S5**) on the *KMT2C* DNAm signature. Similarly, all tested *KMT2C*-related NDD individuals were classified negatively on the Kleefstra and Kabuki syndrome signatures (**Figure 2E**), highlighting the DNAm differences among the three conditions.

These findings confirm that the *KMT2C*-related NDD is clinically and epigenetically distinct from both Kleefstra and Kabuki syndromes.

Figure 3. Photos of unpublished individuals with the *KMT2C*-related NDD.

Discussion

In this study, we have demonstrated that heterozygous loss-of-function *KMT2C* variants result in a disorder that is characterized by variable combinations of developmental delay, intellectual disability, short stature, congenital anomalies, craniofacial dysmorphism, recurrent infections, and a moderate strength DNAm signature. We also demonstrate that the *KMT2C*-related NDD is clinically and epigenetically distinct from (*EHMT1*-related) Kleefstra syndrome type 1 and (*KMT2D*-related) Kabuki type 1 syndromes.

Our data show that despite the complexities of the transcripts encoded by *KMT2C,* the PTVs in the gene have similar clinical or epigenetic consequences irrespective of their location in the gene. This observation suggests four possible explanations: 1) the canonical isoform is biologically and disease relevant in spite of its apparent low level of expression in tissues, 2) the developmental expression profile of *KMT2C* may be different from its known adult expression profile, 3) the *KMT2C* transcript expression pattern may not yet be fully known, or 4) the regulatory consequences of loss-of-function variants in the longer transcripts may have complex consequences on the transcriptional network. Additionally, we cannot exclude that the low expression of the canonical transcript in GTEx is a technical artifact of the short-read based RNA sequencing, displaying 3' bias and short-isoform preferences[42]. Our analysis also shows that *KMT2C* is located in a hot-spot for both benign and pathogenic structural rearrangements, and, therefore, CNVs in this gene should be interpreted with caution. In our study, we also showed that rare missense variants that affect conserved residues and disrupt the 3D structure of the PHD Zn-fingers, the FYR domain, or the catalytic SET also cause the same disorder. Interestingly, pathogenic Kabuki syndrome Type 1 causing missense variants in *KMT2D,* a paralogue of *KMT2C,* cluster significantly in several similar functional domains[54]. These findings collectively indicate haploinsufficiency and functional haploinsufficiency as likely mechanisms that underlie *KMT2C*-related NDD.

We also identify a moderate-effect DNAm signature for *KMT2C*-NDD with current sensitivity of ~82%. To identify a reliable signature, we had to utilize a relatively large sample size cohort (of different ages and sex) as initial signatures derived from smaller cohort sizes lacked sensitivity and specificity. It is possible that variable expressivity is typical not only for the clinical features, but also for the DNAm signature. It has been shown previously that the level of DNAm changes correlate with the age of onset of dystonia in the *KMT2B*-related dystonia[55] (MIM: 617284), and the Weaver-syndrome (MIM: 277590) signature was found to be

weaker in mildly affected individuals with a pathogenic *EZH2* (MIM: 601573) variant even within a single family[18]. Therefore, it is possible that mildly affected individuals can be classified as intermediate or negative on the DNAm signature despite the presence of pathogenic variants. For example, one mildly affected individual with a pathogenic *KMT2C* variant (exons 36-37 deletions) classified as negative by the DNAm signature is the mother of three affected children. This individual is mildly affected which might explain the DNAm classification results. Unfortunately, DNA from the children was not available for testing. However, the second individual with pathogenic variant c.13107_13108dup p.(Thr4370Argfs*11) was also classified negatively despite having typical *KMT2C*-related NDD clinical features. However, we were not able to be exclude mosaicism in either case, which may explain the negative classification results. In the future, larger studies will be required to further elucidate the biological basis of such correlations, which may have prognostic and management implications.

A large proportion of the identified *KMT2C* signature's CpG sites (8%) mapped to the *WT1* gene's CpG islands, which are hypomethylated in *KMT2C*-related NDD individuals. Hypomethylation of CpG islands, especially those located in the promoter of a gene, is usually associated with upregulation of gene expression[56], which was previously shown for *WT1 in vitro*[57]. WT1 is not a known interactor of *KMT2C*, so it is unclear whether the hypomethylation at the *WT1* CpG islands is a primary effect of KMT2C disruption or secondary to disrupted interactions of multiple genes. Loss-of-function or dominant-negative *WT1* variants are associated with Wilms tumor (MIM: 194070) or Denys-Drash syndrome (MIM: 194080), respectively, but to date there is no human phenotype associated with *WT1* overexpression. Until now, individuals with *KMT2C*-related NDD have not displayed any tumors or anomalies of sex development that are reported for the *WT1*-related disorders. Recently, a role for WT1 in brain and neuron development was also described[58-60]. Mariotinni et al.[61], proved that Wt1 is important for synaptic plasticity and learning in mice, and that it functions as a long-term memory suppressor. Additionally, they showed that Wt1 overexpression causes a reduction of memory retention[61]. In another study, Ji et al., showed that Wt1 brain-specific loss in mice resulted in depressive-like behavior, but the effects of overexpression were not investigated[60]. At this time, the relevance of *WT1* methylation changes to the phenotype of *KMT2C*-related NDD is not known.

Several neurodevelopmental phenotypes were present in the majority of the individuals in the study cohort, including developmental delay, intellectual disability (mostly mild, but ranging from borderline to severe), ASD, and ADHD. Additionally,

the individuals often have other psychiatric issues, including hetero- or self-aggressive behavior, obsessive-compulsive behavior, and selective mutism. In fact, pathogenic or *de novo KMT2C* variants have also been found to be enriched in individuals with schizophrenia[62], bipolar disorder[63] and other psychiatric disorders[64]. This suggests that *KMT2C*-related NDD is associated with multiple, overlapping neurodevelopmental and psychiatric presentations.

Individuals with *KMT2C*-related NDD present with a broad spectrum of clinical features, but the expressivity of the features is highly variable among individuals. For example, ~14% of the individuals in the entire cohort did not have intellectual or learning disabilities. The variable expressivity likely explains the relatively large proportion (11/75, ~15%) of inherited cases in this cohort, as most of the parents with pathogenic variants were mildly affected.

Individuals with the *KMT2C*-related NDD had common dysmorphic features, including frontal bossing or prominent forehead, downslanted palpebral fissures, deep-set eyes, and hypertelorism, among others. These features are mild and non-specific, and therefore, the facial gestalt was unrecognizable to clinicians, unlike Kleefstra and Kabuki 1 syndromes. However, using PhenoScore[38], we were able to show that the *KMT2C*-related NDD had a typical facial gestalt that is significantly different from Kleefstra and Kabuki 1 syndromes.

KMT2C-related NDD, Kleefstra, and Kabuki 1 syndromes are clearly different conditions despite some clinical and molecular overlap, but it is currently impossible to objectively compare the frequencies of specific features among these conditions head-to-head. Kleefstra and Kabuki syndromes are clinically well-known and well-recognized, and most of the published individuals are diagnosed through a phenotype-first approach[8,65]. This results in biased results with only the most typical affected individuals being initially described. However, with the recent widespread application of exome or genome sequencing in clinical diagnostics, it is possible to also identify non-specific and/or mildly affected individuals[66,67], as well as novel morbid entities within the same gene[50], including *KMT2D*[68]. *KMT2C*-related NDD is not clinically recognizable, so all individuals were diagnosed through a genotype-first approach and this study represents a clinical expansion and delineation of this condition. Therefore, studies on the phenotype of Kleefstra and Kabuki 1 syndromes diagnosed through a genotype-first approach would be necessary to better understand their clinical and molecular spectrum, as well as the differences among the three conditions.

Although *KMT2C*-related NDD, Kleefstra and Kabuki 1 syndrome are caused by haploinsufficiency of methyltransferases, where PTVs are the most frequent type of variant, the proportion of inherited variants is considerably different among them. While Kleefstra and Kabuki 1 syndrome mostly occur by *de novo* loss-of-function variants, with only a handful of inherited cases described so far[69,70] , ~15% of *KMT2C*-related NDD individuals inherited their variants from similarly affected parents. Albeit this may be explained by a biased approach to diagnosis, our study also showed that the expressivity of the *KMT2C*-related NDD is highly variable, and individuals with few manifestations can be expected.

Several individuals in our study were primarily investigated due to short stature - a feature present in more than a half of the *KMT2C*-related NDD individuals reported in this study. Short stature is also common in Kleefstra and Kabuki 1 syndromes[8,65]. The causes of short stature in these syndromes are often unknown but likely complex[65]. However, short stature in individuals with monogenic NDDs is rarely investigated in-depth in routine clinical practice as it is usually considered to be part of the syndrome. While it is known that growth hormone deficiency is present in approximately a third of individuals with Kabuki 1 syndrome[71], this can partially explain the presence of short stature. However, treatment with rhGH has proven to be highly effective and safe[72], and therefore, these individuals should be screened for this deficiency. In this cohort, 40/75 indidivuals with endocrine anomalies had proven short stature, and 6/40 had received rhGH with apparently good results. This should prompt further studies on the prevalence of GH deficiency and the utility of rhGH therapy also for individuals with *KMT2C*-related NDD and short stature.

While various psychiatric disorders are seen in *KMT2C*-related NDD, this is not frequent in Kabuki 1 syndrome[73]. Also, although psychiatric disorders such as schizophrenia[74] and developmental regression[75-77] have been frequently reported in Kleefstra syndrome, nearly all individuals with Kleefstra syndrome manifest intellectual disability that ranges from moderate to severe[8,77], which contrasts with our finding that ~15% of individuals with *KMT2C*-related NDD demonstrate no cognitive impairment. Again, albeit this may be explained by a bias in our diagnostic approach, our study shows that the penetrance and severity of this finding is consistently reduced in the *KMT2C*-related NDD.

In summary, pathogenic variants in the *KMT2C* gene result in a distinct syndromic neurodevelopmental disorder, which is different from Kleefstra and Kabuki 1 syndromes. Therefore, the OMIM designation of *KMT2C*-related NDD as Kleefstra syndrome 2 should be reconsidered, as it may be misleading to clinicians and patients' families.

Supplemental information

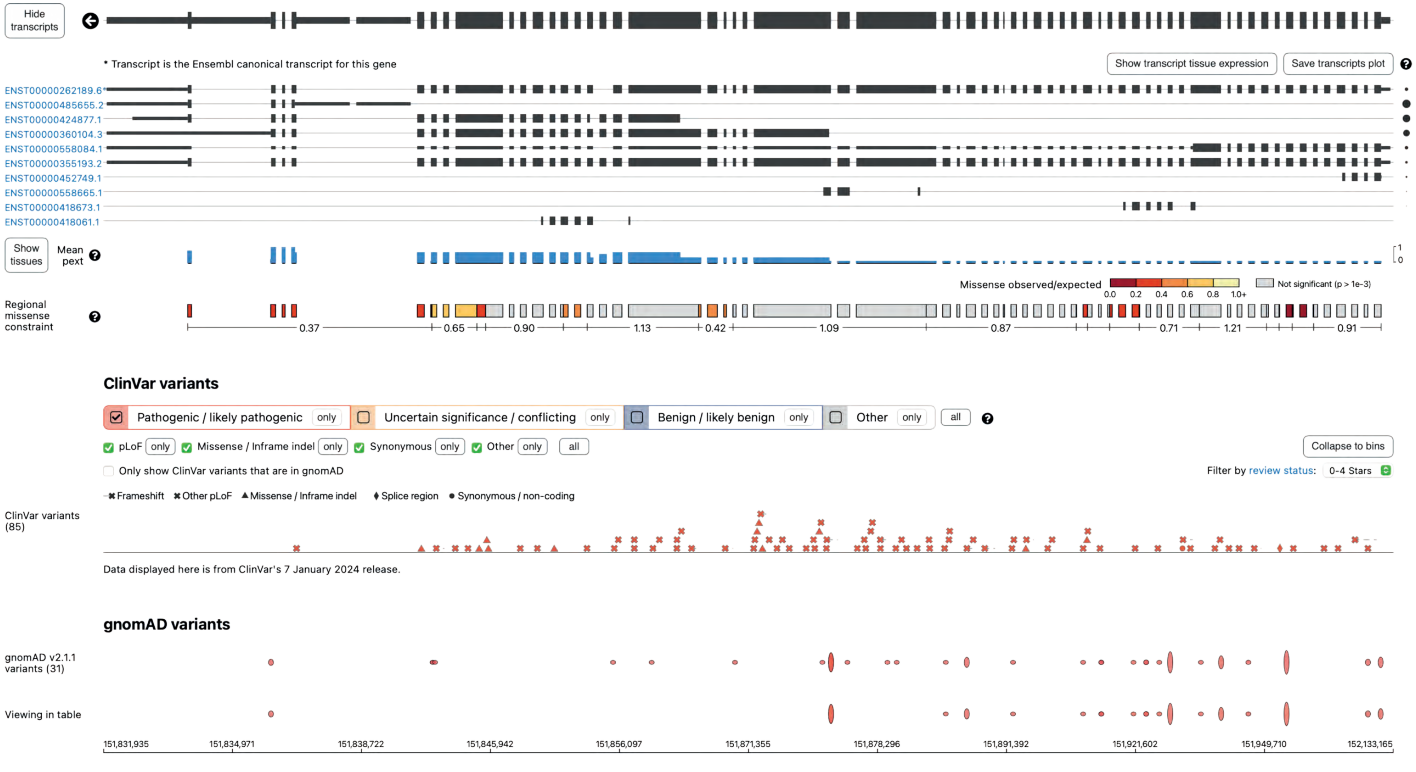


Figure S1. Distribution of the PTVs across *KMT2C* transcripts with their relative expression in the gnomAD database.

Figure S2. *KMT2C*-related NDD facial gestalt heatmap showing significant facial areas most different from controls.

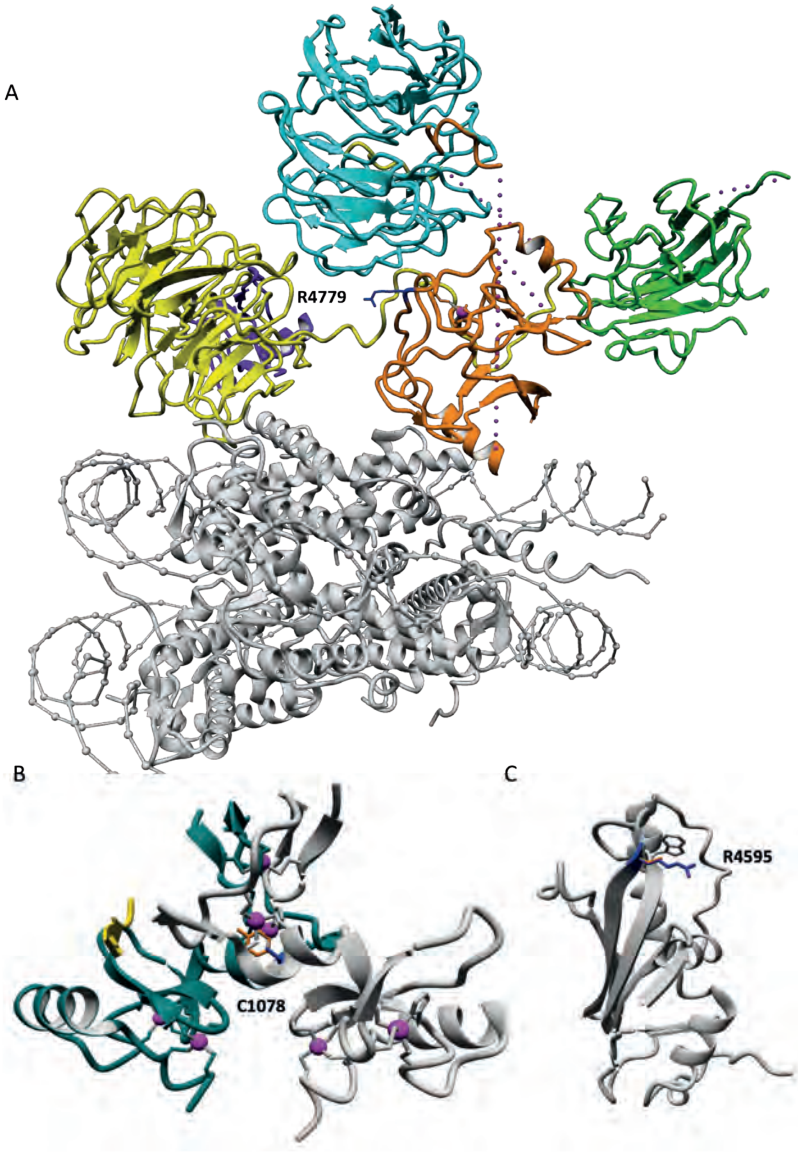


Figure S3. Protein 3D structure effect predictions of the pathogenic PAVs.

A. *Arg4779 is located on a surface of the SET domain that interacts with WDR5 (PDB:5F6K). Arg change to Gly is a radical change and will likely affect the interaction that is required for the KMT2C activity. Previous study (25561738) showed that Arg4779Pro affected the methyltransferase activity of KMT2C.* ***B.*** *Cys1078 is binding to Zn in ePHD6 to ensure the correct folding and structure of the domain (PDB:6MLC); a change to Tyr would disrupt the Zn binding and, therefore, domain structure.* ***C.*** *Arg4595 is located on a surface of the FYR domain (AlphaFold2 model) and interacts with Trp4987, so change to smaller Cys would affect local aminoacid packing and interactions, and as result might destabilize the domain or affect binding to other proteins or binding between FYRN and FYRC domain parts.*

Tables S1-S3, S5 supporting the findings of this study are available online in the Supplementary material of this article at: DOI: 10.1016/j.ajhg.2024.06.009.

Table S4. *KMT2C* VUS classification.

Variant Class	Variant (NM_170606.3)	Occurence in gnomAD v2.1.1.	Inheritance	*In silico* tool predictions (REVEL/SpliceAI)
PAVs	c.2861C>T p.(Thr954Ile)	0	DN	0.59
	c.3233G>A p.(Cys1078Tyr)	0	DN	0.72
	c.5108C>T p.(Ala1703Val)	0	I	0.66
	c.6355C>T p.(Pro2119Ser)	0	U	0.2
	c.8065C>T p.(Leu2689Phe)	0	DN	0.47
	c.8710G>T p.(Asp2904Tyr)	0	U	0.34
	c.8974G>A p.(Gly2992Arg)	11	I	0.62
	c.9773A>C p.(His3258Pro)	0	DN	0.68
	c.12539A>G p.(Tyr4180Cys)	0	DN	0.19
	c.12898T>C p.(Ser4300Pro)	51	I	0.47
	c.13229A>G p.(Asp4410Gly)	0	U	0.82
	c.13298C>T p.(Ala4433Val)	0	DN	0.62
	c.13783C>T p.(Arg4595Cys)	0	DN	0.91
	c.14023G>A p.(Ala4675Thr)	1	U	0.19
	c.14057A>G p.(Asn4686Ser)	98	I	0.13
	c.14153A>C p.(His4718Pro)	0	I	0.58
	c.14335C>G p.(Arg4779Gly)	0	U	0.74
	c.14501T>C p.(Val4834Ala)	0	DN	0.81
Splice	c.1735+2dup p.?	0	I	0.99
	c.3499+1G>T p.?	0	DN	0.98
Dup	duplication exons 3-38	0	DN	-
	duplication exons 1-55	0	U	-

DN = de novo; I = inherited; U = unknown

REVEL > 0.64 or < 0.29 for PAVs; SpliceAI > 0.2 or <0.1 for splice

Blue = evidence for pathogenicity; Green = benign evidence; Grey – unknown or uncertain significance

	***In silico* tool prediction (AlphaMissense)**	**Domain location**	**Predicted 3D protein effect**	**DNAm SVM results**	**Final classification**
	0.59	Dis	N	0.02	B
	1	ePHD6	D	0.85	P
	0.91	HMG	U	0.14	B
	0.07	Dis	N	0.01	B
	0.46	Dis	N	0.01	B
	0.19	Dis	N	0.02	B
	0.26	Dis	N	0.00	B
	1	Coil	U	0.00	U
	0.07	Dis	N	0.20	B
	0.1	Dis	N	0.01	B
	1	ePHD2	-	0.73	P
	0.94	ePHD2	U	0.05	U
	0.94	FYRN	U	0.83	P
	0.12	FYRC	-	0.07	B
	0.08	Dis	N	0.01	B
	0.07	Dis	N	0.05	B
	0.8	SET	D	0.72	P
	0.99	SET	D	-	U
	-	Dis	N	0.03	B
	-	-	-	0.41	P
	-	-	U	0.67	P
	-	-	-	0.15	B
	Alphamissense > 0.56 or < 0.34	Dis = no domains, disordered	N = no predicted effect; D = damaging ; U = unclear	Blue = positive SVM; Grey = intermediate; Green = negative	P = pathogenic; B = benign; U = variant of uncertain significance

Declaration of interest

RW is a consultant (equity) for Alamya Health. The rest of the authors declare no competing interests.

Acknowledgements

This work was supported by a Canadian Institutes of Health Research (CIHR) grant to RW (PJT-178315) and the Ontario Brain Institute (Province of Ontario Neurodevelopmental Disorders (POND) network (IDS11-02)) grants to RW. This work was also supported by a grant from SFARI (887172 for RW); Aspasia grant of the Dutch Research Council (015.014.036 to TK), the Netherlands Organization for Health Research and Development (91718310 to TK); National Human Genome Research Institute grants UM1HG008900 (with additional support from the National Eye Institute, and the National Heart, Lung and Blood Institute) and U01HG011755 (to AODL) and by the R01HG009141 grant, Chan Zuckerberg Initiative Donor-Advised Fund at the Silicon Valley Community Foundation (https://doi.org/10.37921/236582yuakxy) grant 2019-199278 (funder DOI 10.13039/100014989), the Manton Center for Orphan Disease Research, and Boston Children's Hospital Children's Hospital CRDC (to AODL). Special thanks to Biobanc de l'Hospital Infantil Sant Joan de Déu per a la Investigació, which is integrated into the Spanish Biobank Network of ISCIII, and BioNER: Biobank of the Institute of Rare Diseases Research for providing samples.

These results contribute to the overall goals of the Solve-RD project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257 (LV; TK; HB; SB). Several authors of this publication are members of the European Reference Network on Rare Congenital Malformations and Rare Intellectual Disability ERN-ITHACA. Sequencing and analysis of one individual were provided by the Broad Institute Center for Mendelian Genomics (Broad CMG).

SB acknowledges the support of the NIHR Manchester Biomedical Research Centre (NIHR203308) and the MRC Epigenomics of Rare Diseases Node (MR/Y008170/1). We thank the Simons Simplex Collection (SSC), Deciphering Developmental Delay (DDD), the 100,000 Genome Project, the Autism Sequencing Consortium (ASC) and the Autism Speaks MSSNG project for providing the individuals, samples, and/or molecular diagnostic data. We thank Michael Kwint for the technical assistance.

Web resources

https://epigen.ccm.sickkids.ca

Data availability

Recruited individuals' with pathogenic *KMT2C* variants clinical details are provided in the **Table S1**. Recruited individuals' with benign or VUS *KMT2C* variants details are provided in the **Table S2**. The variants and their interpretations were submitted to the ClinVar database (ClinVar accession numbers: SCV005044911-SCV005044983 and SCV005044991-SCV005045000).

The *KMT2C* DNAm signature is available in supplemental material (**Table S3**). The raw datasets supporting the current study have not been deposited in a public repository due to institutional ethics restrictions. All software and R packages used in the study are publicly available as described in the methods section. The *KMT2C* methylation classifier will be made available through EpigenCentral at https://epigen.ccm.sickkids.ca.

References

1. Kleefstra, T., Schenck, A., Kramer, J.M., and van Bokhoven, H. (2014). The genetics of cognitive epigenetics. Neuropharmacology *80*, 83-94. 10.1016/j.neuropharm.2013.12.025.
2. Cenik, B.K., and Shilatifard, A. (2021). COMPASS and SWI/SNF complexes in development and disease. Nat Rev Genet *22*, 38-58. 10.1038/s41576-020-0278-0.
3. Koemans, T.S., Kleefstra, T., Chubak, M.C., Stone, M.H., Reijnders, M.R.F., de Munnik, S., Willemsen, M.H., Fenckova, M., Stumpel, C., Bok, L.A., et al. (2017). Functional convergence of histone methyltransferases EHMT1 and KMT2C involved in intellectual disability and autism spectrum disorder. PLoS Genet *13*, e1006864. 10.1371/journal.pgen.1006864.
4. Hu, D., Gao, X., Morgan, M.A., Herz, H.M., Smith, E.R., and Shilatifard, A. (2013). The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. Mol Cell Biol *33*, 4745-4754. 10.1128/mcb.01181-13.
5. Rampias, T., Karagiannis, D., Avgeris, M., Polyzos, A., Kokkalis, A., Kanaki, Z., Kousidou, E., Tzetis, M., Kanavakis, E., Stravodimos, K., et al. (2019). The lysine-specific methyltransferase KMT2C/MLL3 regulates DNA repair components in cancer. EMBO Rep *20*. 10.15252/embr.201846821.
6. Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet *42*, 790-793. 10.1038/ng.646.
7. Kleefstra, T., Kramer, J.M., Neveling, K., Willemsen, M.H., Koemans, T.S., Vissers, L.E., Wissink-Lindhout, W., Fenckova, M., van den Akker, W.M., Kasri, N.N., et al. (2012). Disruption of an EHMT1-associated chromatin-modification module causes intellectual disability. Am J Hum Genet *91*, 73-82. 10.1016/j.ajhg.2012.05.003.
8. Willemsen, M.H., Vulto-van Silfhout, A.T., Nillesen, W.M., Wissink-Lindhout, W.M., van Bokhoven, H., Philip, N., Berry-Kravis, E.M., Kini, U., van Ravenswaaij-Arts, C.M., Delle Chiaie, B., et al. (2012). Update on Kleefstra Syndrome. Mol Syndromol *2*, 202-212. 10.1159/000335648.
9. Faundes, V., Newman, W.G., Bernardini, L., Canham, N., Clayton-Smith, J., Dallapiccola, B., Davies, S.J., Demos, M.K., Goldman, A., Gill, H., et al. (2018). Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorders. American journal of human genetics *102*, 175-187. 10.1016/j.ajhg.2017.11.013.
10. Schobers, G., Schieving, J.H., Yntema, H.G., Pennings, M., Pfundt, R., Derks, R., Hofste, T., de Wijs, I., Wieskamp, N., van den Heuvel, S., et al. (2022). Reanalysis of exome negative patients with rare disease: a pragmatic workflow for diagnostic applications. Genome Med *14*, 66. 10.1186/s13073-022-01069-z.
11. Prevalence and architecture of de novo mutations in developmental disorders. (2017). Nature *542*, 433-438. 10.1038/nature21062.
12. Smedley, D., Smith, K.R., Martin, A., Thomas, E.A., McDonagh, E.M., Cipriani, V., Ellingford, J.M., Arno, G., Tucci, A., Vandrovcova, J., et al. (2021). 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med *385*, 1868-1880. 10.1056/NEJMoa2035790.
13. Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192-195. 10.1016/j.neuron.2010.10.006.
14. Trost, B., Thiruvahindrapuram, B., Chan, A.J.S., Engchuan, W., Higginbotham, E.J., Howe, J.L., Loureiro, L.O., Reuter, M.S., Roshandel, D., Whitney, J., et al. (2022). Genomic architecture of autism from comprehensive whole-genome sequence annotation. Cell *185*, 4409-4427.e4418. 10.1016/j.cell.2022.10.009.

15. Morales, J., Pujar, S., Loveland, J.E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C.M., et al. (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature *604*, 310-315. 10.1038/s41586-022-04558-8.

16. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). 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 *17*, 405-424. 10.1038/gim.2015.30.

17. Choufani, S., McNiven, V., Cytrynbaum, C., Jangjoo, M., Adam, M.P., Bjornsson, H.T., Harris, J., Dyment, D.A., Graham, G.E., Nezarati, M.M., et al. (2022). An HNRNPK-specific DNA methylation signature makes sense of missense variants and expands the phenotypic spectrum of Au-Kline syndrome. Am J Hum Genet *109*, 1867-1884. 10.1016/j.ajhg.2022.08.014.

18. Choufani, S., Gibson, W.T., Turinsky, A.L., Chung, B.H.Y., Wang, T., Garg, K., Vitriolo, A., Cohen, A.S.A., Cyrus, S., Goodman, S., et al. (2020). DNA Methylation Signature for EZH2 Functionally Classifies Sequence Variants in Three PRC2 Complex Genes. Am J Hum Genet *106*, 596-610. 10.1016/j.ajhg.2020.03.008.

19. Awamleh, Z., Choufani, S., Cytrynbaum, C., Alkuraya, F.S., Scherer, S., Fernandes, S., Rosas, C., Louro, P., Dias, P., Neves, M.T., et al. (2023). ANKRD11 pathogenic variants and 16q24.3 microdeletions share an altered DNA methylation signature in patients with KBG syndrome. Hum Mol Genet *32*, 1429-1438. 10.1093/hmg/ddac289.

20. Fortin, J.P., Triche, T.J., Jr., and Hansen, K.D. (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. Bioinformatics *33*, 558-560. 10.1093/bioinformatics/btw691.

21. LaBarre, B.A., Goncearenco, A., Petrykowska, H.M., Jaratlerdsiri, W., Bornman, M.S.R., Hayes, V.M., and Elnitski, L. (2019). MethylToSNP: identifying SNPs in Illumina DNA methylation array data. Epigenetics Chromatin *12*, 79. 10.1186/s13072-019-0321-6.

22. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res *43*, e47. 10.1093/nar/gkv007.

23. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software *28*, 1 - 26. 10.18637/jss.v028.i05.

24. Turinsky, A.L., Choufani, S., Lu, K., Liu, D., Mashouri, P., Min, D., Weksberg, R., and Brudno, M. (2020). EpigenCentral: Portal for DNA methylation data analysis and classification in rare diseases. Hum Mutat *41*, 1722-1733. 10.1002/humu.24076.

25. Awamleh, Z., Goodman, S., Kallurkar, P., Wu, W., Lu, K., Choufani, S., Turinsky, A.L., and Weksberg, R. (2022). Generation of DNA Methylation Signatures and Classification of Variants in Rare Neurodevelopmental Disorders Using EpigenCentral. Curr Protoc *2*, e597. 10.1002/cpz1.597.

26. Liu, Y., Qin, S., Chen, T.Y., Lei, M., Dhar, S.S., Ho, J.C., Dong, A., Loppnau, P., Li, Y., Lee, M.G., and Min, J. (2019). Structural insights into trans-histone regulation of H3K4 methylation by unique histone H4 binding of MLL3/4. Nat Commun *10*, 36. 10.1038/s41467-018-07906-3.

27. Xue, H., Yao, T., Cao, M., Zhu, G., Li, Y., Yuan, G., Chen, Y., Lei, M., and Huang, J. (2019). Structural basis of nucleosome recognition and modification by MLL methyltransferases. Nature *573*, 445-449. 10.1038/s41586-019-1528-1.

28. Li, Y., Han, J., Zhang, Y., Cao, F., Liu, Z., Li, S., Wu, J., Hu, C., Wang, Y., Shuai, J., et al. (2016). Structural basis for activity regulation of MLL family methyltransferases. Nature *530*, 447-452. 10.1038/nature16952.
29. García-Alai, M.M., Allen, M.D., Joerger, A.C., and Bycroft, M. (2010). The structure of the FYR domain of transforming growth factor beta regulator 1. Protein Sci *19*, 1432-1438. 10.1002/pro.404.
30. Liu, Z., Li, F., Ruan, K., Zhang, J., Mei, Y., Wu, J., and Shi, Y. (2014). Structural and functional insights into the human Börjeson-Forssman-Lehmann syndrome-associated protein PHF6. J Biol Chem *289*, 10069-10083. 10.1074/jbc.M113.535351.
31. Klein, B.J., Cox, K.L., Jang, S.M., Côté, J., Poirier, M.G., and Kutateladze, T.G. (2020). Molecular Basis for the PZP Domain of BRPF1 Association with Chromatin. Structure *28*, 105-110.e103. 10.1016/j.str.2019.10.014.
32. Klein, B.J., Deshpande, A., Cox, K.L., Xuan, F., Zandian, M., Barbosa, K., Khanal, S., Tong, Q., Zhang, Y., Zhang, P., et al. (2021). The role of the PZP domain of AF10 in acute leukemia driven by AF10 translocations. Nat Commun *12*, 4130. 10.1038/s41467-021-24418-9.
33. Zheng, S., Bi, Y., Chen, H., Gong, B., Jia, S., and Li, H. (2021). Molecular basis for bipartite recognition of histone H3 by the PZP domain of PHF14. Nucleic Acids Res *49*, 8961-8973. 10.1093/nar/gkab670.
34. Krieger, E., and Vriend, G. (2014). YASARA View - molecular graphics for all devices - from smartphones to workstations. Bioinformatics *30*, 2981-2982. 10.1093/bioinformatics/btu426.
35. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583-589. 10.1038/s41586-021-03819-2.
36. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res *50*, D439-d444. 10.1093/nar/gkab1061.
37. Faundes, V., Goh, S., Akilapa, R., Bezuidenhout, H., Bjornsson, H.T., Bradley, L., Brady, A.F., Brischoux-Boucher, E., Brunner, H., Bulk, S., et al. (2021). Clinical delineation, sex differences, and genotype-phenotype correlation in pathogenic KDM6A variants causing X-linked Kabuki syndrome type 2. Genetics in medicine : official journal of the American College of Medical Genetics *23*, 1202-1210. 10.1038/s41436-021-01119-8.
38. Dingemans, A.J.M., Hinne, M., Truijen, K.M.G., Goltstein, L., van Reeuwijk, J., de Leeuw, N., Schuurs-Hoeijmakers, J., Pfundt, R., Diets, I.J., den Hoed, J., et al. (2023). PhenoScore quantifies phenotypic variation for rare genetic diseases by combining facial analysis with other clinical features using a machine-learning framework. Nature genetics *55*, 1598-1607. 10.1038/s41588-023-01469-w.
39. Terhorst, P., Ihlefeld, M., Huber, M., Damer, N., Kirchbuchner, F., Raja, K.B., and Kuijper, A. (2021). QMagFace: Simple and Accurate Quality-Aware Face Recognition. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 3473-3483.
40. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443. 10.1038/s41586-020-2308-7.
41. The Genotype-Tissue Expression (GTEx) project. (2013). Nat Genet *45*, 580-585. 10.1038/ng.2653.
42. Cummings, B.B., Karczewski, K.J., Kosmicki, J.A., Seaby, E.G., Watts, N.A., Singer-Berk, M., Mudge, J.M., Karjalainen, J., Satterstrom, F.K., O'Donnell-Luria, A.H., et al. (2020). Transcript expression-aware annotation improves rare variant interpretation. Nature *581*, 452-458. 10.1038/s41586-020-2329-2.

43. Pich, O., Reyes-Salazar, I., Gonzalez-Perez, A., and Lopez-Bigas, N. (2022). Discovering the drivers of clonal hematopoiesis. Nat Commun *13*, 4267. 10.1038/s41467-022-31878-0.

44. Sahoo, T., Theisen, A., Rosenfeld, J.A., Lamb, A.N., Ravnan, J.B., Schultz, R.A., Torchia, B.S., Neill, N., Casci, I., Bejjani, B.A., and Shaffer, L.G. (2011). Copy number variants of schizophrenia susceptibility loci are associated with a spectrum of speech and developmental delays and behavior problems. Genet Med *13*, 868-880. 10.1097/GIM.0b013e3182217a06.

45. MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic acids research *42*, D986-992. 10.1093/nar/gkt958.

46. Nassar, L.R., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., et al. (2023). The UCSC Genome Browser database: 2023 update. Nucleic Acids Res *51*, D1188-d1195. 10.1093/nar/gkac1072.

47. Chater-Diehl, E., Goodman, S.J., Cytrynbaum, C., Turinsky, A.L., Choufani, S., and Weksberg, R. (2021). Anatomy of DNA methylation signatures: Emerging insights and applications. Am J Hum Genet *108*, 1359-1366. 10.1016/j.ajhg.2021.06.015.

48. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. American journal of human genetics *99*, 877-885. 10.1016/j.ajhg.2016.08.016.

49. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science *381*, eadg7492. 10.1126/science.adg7492.

50. Rots, D., Chater-Diehl, E., Dingemans, A.J.M., Goodman, S.J., Siu, M.T., Cytrynbaum, C., Choufani, S., Hoang, N., Walker, S., Awamleh, Z., et al. (2021). Truncating SRCAP variants outside the Floating-Harbor syndrome locus cause a distinct neurodevelopmental disorder with a specific DNA methylation signature. American journal of human genetics *108*, 1053-1068. 10.1016/j.ajhg.2021.04.008.

51. Butcher, D.T., Cytrynbaum, C., Turinsky, A.L., Siu, M.T., Inbar-Feigenberg, M., Mendoza-Londono, R., Chitayat, D., Walker, S., Machado, J., Caluseriu, O., et al. (2017). CHARGE and Kabuki Syndromes: Gene-Specific DNA Methylation Signatures Identify Epigenetic Mechanisms Linking These Clinically Overlapping Conditions. Am J Hum Genet *100*, 773-788. 10.1016/j.ajhg.2017.04.004.

52. Kleefstra, T., Brunner, H.G., Amiel, J., Oudakker, A.R., Nillesen, W.M., Magee, A., Geneviève, D., Cormier-Daire, V., van Esch, H., Fryns, J.P., et al. (2006). Loss-of-function mutations in euchromatin histone methyl transferase 1 (EHMT1) cause the 9q34 subtelomeric deletion syndrome. American journal of human genetics *79*, 370-377. 10.1086/505693.

53. Goodman, S., Cytrynbaum, C., Chung, B., Chater-Diehl, E., Aziz, C., Turinsky, A., Kellam, B., Keller, M., Ko, J.M., Caluseriu, O., et al. (2020). EHMT1 pathogenic variants and 9q34.3 microdeletions share altered DNA methylation patterns in patients with Kleefstra syndrome. Journal of Translational Genetics and Genomics *4*, 144-158. 10.20517/jtgg.2020.23.

54. Faundes, V., Malone, G., Newman, W.G., and Banka, S. (2019). A comparative analysis of KMT2D missense variants in Kabuki syndrome, cancers and the general population. J Hum Genet *64*, 161-170. 10.1038/s10038-018-0536-6.

55. Mirza-Schreiber, N., Zech, M., Wilson, R., Brunet, T., Wagner, M., Jech, R., Boesch, S., Škorvánek, M., Necpál, J., Weise, D., et al. (2022). Blood DNA methylation provides an accurate biomarker of KMT2B-related dystonia and predicts onset. Brain *145*, 644-654. 10.1093/brain/awab360.

56. Garg, P., Jadhav, B., Rodriguez, O.L., Patel, N., Martin-Trujillo, A., Jain, M., Metsu, S., Olsen, H., Paten, B., Ritz, B., et al. (2020). A Survey of Rare Epigenetic Variation in 23,116 Human Genomes Identifies Disease-Relevant Epivariations and CGG Expansions. Am J Hum Genet *107*, 654-669. 10.1016/j.ajhg.2020.08.019.

57. Hamatani, H., Sakairi, T., Ikeuchi, H., Kaneko, Y., Maeshima, A., Nojima, Y., and Hiromura, K. (2019). TGF-β1 alters DNA methylation levels in promoter and enhancer regions of the WT1 gene in human podocytes. Nephrology (Carlton) *24*, 575-584. 10.1111/nep.13411.

58. Schnerwitzki, D., Perry, S., Ivanova, A., Caixeta, F.V., Cramer, P., Günther, S., Weber, K., Tafreshiha, A., Becker, L., Vargas Panesso, I.L., et al. (2018). Neuron-specific inactivation of Wt1 alters locomotion in mice and changes interneuron composition in the spinal cord. Life Sci Alliance *1*, e201800106. 10.26508/lsa.201800106.

59. Schnerwitzki, D., Hayn, C., Perner, B., and Englert, C. (2020). Wt1 Positive dB4 Neurons in the Hindbrain Are Crucial for Respiration. Front Neurosci *14*, 529487. 10.3389/fnins.2020.529487.

60. Ji, F., Wang, W., Feng, C., Gao, F., and Jiao, J. (2021). Brain-specific Wt1 deletion leads to depressive-like behaviors in mice via the recruitment of Tet2 to modulate Epo expression. Mol Psychiatry *26*, 4221-4233. 10.1038/s41380-020-0759-8.

61. Mariottini, C., Munari, L., Gunzel, E., Seco, J.M., Tzavaras, N., Hansen, J., Stern, S.A., Gao, V., Aleyasin, H., Sharma, A., et al. (2019). Wilm's tumor 1 promotes memory flexibility. Nat Commun *10*, 3756. 10.1038/s41467-019-11781-x.

62. Rees, E., Han, J., Morgan, J., Carrera, N., Escott-Price, V., Pocklington, A.J., Duffield, M., Hall, L.S., Legge, S.E., Pardiñas, A.F., et al. (2020). De novo mutations identified by exome sequencing implicate rare missense variants in SLC6A1 in schizophrenia. Nat Neurosci *23*, 179-184. 10.1038/s41593-019-0565-2.

63. Nishioka, M., Kazuno, A.A., Nakamura, T., Sakai, N., Hayama, T., Fujii, K., Matsuo, K., Komori, A., Ishiwata, M., Watanabe, Y., et al. (2021). Systematic analysis of exonic germline and postzygotic de novo mutations in bipolar disorder. Nat Commun *12*, 3750. 10.1038/s41467-021-23453-w.

64. Li, K., Ling, Z., Luo, T., Zhao, G., Zhou, Q., Wang, X., Xia, K., Li, J., and Li, B. (2021). Cross-Disorder Analysis of De Novo Variants Increases the Power of Prioritising Candidate Genes. Life (Basel) *11*. 10.3390/life11030233.

65. Schott, D.A., Blok, M.J., Gerver, W.J., Devriendt, K., Zimmermann, L.J., and Stumpel, C.T. (2016). Growth pattern in Kabuki syndrome with a KMT2D mutation. Am J Med Genet A *170*, 3172-3179. 10.1002/ajmg.a.37930.

66. Rosina, E., Pezzani, L., Pezzoli, L., Marchetti, D., Bellini, M., Pilotta, A., Calabrese, O., Nicastro, E., Cirillo, F., Cereda, A., et al. (2022). Atypical, Composite, or Blended Phenotypes: How Different Molecular Mechanisms Could Associate in Double-Diagnosed Patients. Genes (Basel) *13*. 10.3390/genes13071275.

67. Dyment, D.A., Tétreault, M., Beaulieu, C.L., Hartley, T., Ferreira, P., Chardon, J.W., Marcadier, J., Sawyer, S.L., Mosca, S.J., Innes, A.M., et al. (2015). Whole-exome sequencing broadens the phenotypic spectrum of rare pediatric epilepsy: a retrospective study. Clin Genet *88*, 34-40. 10.1111/cge.12464.

68. Cuvertino, S., Hartill, V., Colyer, A., Garner, T., Nair, N., Al-Gazali, L., Canham, N., Faundes, V., Flinter, F., Hertecant, J., et al. (2020). A restricted spectrum of missense KMT2D variants cause a multiple malformations disorder distinct from Kabuki syndrome. Genet Med *22*, 867-877. 10.1038/s41436-019-0743-3.

69. Bögershausen, N., Gatinois, V., Riehmer, V., Kayserili, H., Becker, J., Thoenes, M., Simsek-Kiper, P., Barat-Houari, M., Elcioglu, N.H., Wieczorek, D., et al. (2016). Mutation Update for Kabuki Syndrome Genes KMT2D and KDM6A and Further Delineation of X-Linked Kabuki Syndrome Subtype 2. Hum Mutat *37*, 847-864. 10.1002/humu.23026.

70. Rots, D., Bouman, A., Yamada, A., Levy, M., Dingemans, A.J.M., de Vries, B.B.A., Ruiterkamp-Versteeg, M., de Leeuw, N., Ockeloen, C.W., Pfundt, R., et al. (2024). Comprehensive EHMT1 variants analysis broadens genotype-phenotype associations and molecular mechanisms in Kleefstra syndrome. American journal of human genetics *111*, 1605-1625. 10.1016/j.ajhg.2024.06.008.

71. Schott, D.A., Gerver, W.J., and Stumpel, C.T. (2016). Growth Hormone Stimulation Tests in Children with Kabuki Syndrome. Horm Res Paediatr *86*, 319-324. 10.1159/000449221.

72. van Montfort, L., Gerver, W.J.M., Kooger, B.L.S., Plat, J., Bierau, J., Stumpel, C., and Schott, D.A. (2021). Follow-Up Study of Growth Hormone Therapy in Children with Kabuki Syndrome: Two-Year Treatment Results. Horm Res Paediatr *94*, 285-296. 10.1159/000519963.

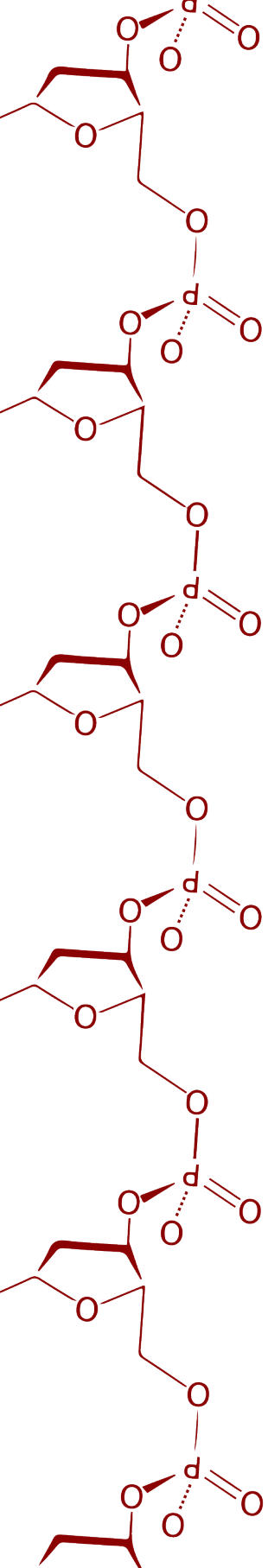
73. Barry, K.K., Tsaparlis, M., Hoffman, D., Hartman, D., Adam, M.P., Hung, C., and Bodamer, O.A. (2022). From Genotype to Phenotype-A Review of Kabuki Syndrome. Genes (Basel) *13*. 10.3390/genes13101761.

74. Vermeulen, K., Staal, W.G., Janzing, J.G., van Bokhoven, H., Egger, J.I.M., and Kleefstra, T. (2017). Sleep Disturbance as a Precursor of Severe Regression in Kleefstra Syndrome Suggests a Need for Firm and Rapid Pharmacological Treatment. Clin Neuropharmacol *40*, 185-188. 10.1097/wnf.0000000000000226.

75. Vermeulen, K., de Boer, A., Janzing, J.G.E., Koolen, D.A., Ockeloen, C.W., Willemsen, M.H., Verhoef, F.M., van Deurzen, P.A.M., van Dongen, L., van Bokhoven, H., et al. (2017). Adaptive and maladaptive functioning in Kleefstra syndrome compared to other rare genetic disorders with intellectual disabilities. Am J Med Genet A *173*, 1821-1830. 10.1002/ajmg.a.38280.

76. Verhoeven, W.M., Egger, J.I., Vermeulen, K., van de Warrenburg, B.P., and Kleefstra, T. (2011). Kleefstra syndrome in three adult patients: further delineation of the behavioral and neurological phenotype shows aspects of a neurodegenerative course. Am J Med Genet A *155a*, 2409-2415. 10.1002/ajmg.a.34186.

77. Morison, L.D., Kennis, M.G.P., Rots, D., Bouman, A., Kummeling, J., Palmer, E., Vogel, A.P., Liegeois, F., Brignell, A., Srivastava, S., et al. (2024). Expanding the phenotype of Kleefstra syndrome: speech, language and cognition in 103 individuals. Journal of Medical Genetics, jmg-2023-109702. 10.1136/jmg-2023-109702.



Chapter 5:

Comprehensive *EHMT1* variants analysis broadens genotype-phenotype associations and molecular mechanisms in Kleefstra syndrome

Published in: American Journal of Human Genetics. 2024 Aug 8;111(8):1605-1625.

Authors

Dmitrijs Rots*, Arianne Bouman*, Ayumi Yamada*, Michael Levy, Alexander J.M. Dingemans, Bert B.A. de Vries, Martina Ruiterkamp-Versteeg, Nicole de Leeuw, Charlotte W. Ockeloen, Rolph Pfundt, Elke de Boer, Joost Kummeling, Bregje van Bon, Hans van Bokhoven, Nael Nadif Kasri, Hanka Venselaar, Marielle Alders, Jennifer Kerkhof, Haley McConkey, Alma Kuechler, Bart Elffers, Rixje van Beeck Calkoen, Susanna Hofman, Audrey Smith, Maria Irene Valenzuela, Siddharth Srivastava, Zoe Frazier, Isabelle Maystadt, Carmelo Piscopo, Giuseppe Merla, Meena Balasubramanian, Gijs W.E. Santen, Kay Metcalfe, Soo-Mi Park, Laurent Pasquier, Siddharth Banka, Dian Donnai, Daniel Weisberg, Gertrud Strobl-Wildemann, Annemieke Wagemans, Maaike Vreeburg, Diana Baralle, Nicola Foulds, Ingrid Scurr, Nicola Brunetti-Pierri, Johanna M. van Hagen, Emilia K. Bijlsma, Anna H. Hakonen, Carolina Courage, David Genevieve, Lucile Pinson, Francesca Forzano, Charu Deshpande, Maria L. Kluskens, Lindsey Welling, Astrid S. Plomp, Els K. Vanhoutte, Louisa Kalsner, Janna A. Hol, Audrey Putoux, Johanna Lazier, Pradeep Vasudevan, Elizabeth Ames, Jessica O'Shea, Damien Lederer, Julie Fleischer, Mary O'Connor, Melissa Pauly, Georgia Vasileiou, André Reis, Catherine Kiraly-Borri, Arjan Bouman, Chris Barnett, Marjan Nezarati, Lauren Borch, Gea Beunders, Kübra Özcan, Stéphanie Miot,Catharina M.L. Volker-Touw, Koen L.I. van Gassen,Gerarda Cappuccio, Katrien Janssens, Nofar Mor, Inna Shomer, Dan Dominissini, Matthew L. Tedder, Alison M. Muir, Bekim Sadikovic, Han G. Brunner, Lisenka E.L.M. Vissers, Yoichi Shinkai**, Tjitske Kleefstra**

*,** These authors contributed equally to this work

Abstract

The shift to a genotype-first approach in genetic diagnostics has revolutionized our understanding of neurodevelopmental disorders, expanding both their molecular and phenotypic spectra. Kleefstra syndrome (KLEFS1) is caused by *EHMT1* haploinsufficiency and exhibits broad clinical manifestations. *EHMT1* encodes euchromatic histone methyltransferase-1 – a pivotal component of the epigenetic machinery.

We have recruited 209 individuals with a rare *EHMT1* variant and performed comprehensive molecular *in silico* and *in vitro* testing alongside DNA methylation (DNAm) signature analysis for the identified variants. We (re)classified the variants as likely pathogenic/pathogenic (and confirming Kleefstra syndrome molecularly) in 191 individuals.

We provide an updated and broader clinical and molecular spectrum of Kleefstra syndrome, including individuals with normal intelligence and familial occurrence. Analysis of the *EHMT1* variants reveals a broad range of molecular effects and their associated phenotypes, including distinct genotype-phenotype associations. Notably, we showed that disruption of the "reader" function of the ankyrin repeat domain by a protein altering variant (PAV) results in a KLEFS1 specific DNAm signature and milder phenotype, while disruption of only "writer" methyltransferase activity of the SET domain – does not result in KLEFS1 DNAm signature or typical KLEFS1 phenotype. Similarly, N-terminal truncating variants result in a mild phenotype without the DNAm signature. We demonstrate how comprehensive variant analysis can provide insights into pathogenesis of the disorder and DNAm signature.

In summary, this study presents a comprehensive overview of KLEFS1 and EHMT1, revealing its broader spectrum and deepening our understanding of its molecular mechanisms, thereby informing accurate variant interpretation, counseling, and clinical management.

Introduction

With the recent shift from phenotype-first to genotype-first approach for the diagnostics of genetic disorders, both the molecular and phenotypic spectrum of multiple neurodevelopmental disorders (NDDs) has expanded[1,2]. This has broadened the phenotypic spectra for some of the "well-recognized" syndromes or even identifying novel subgroups of these conditions[3-5]. Additionally, the widespread application of exome, and genome sequencing also significantly increased the number of identified variants of uncertain significance (VUS)[1,6] and revealing novel molecular mechanisms[3,7]. We experienced all these facets through almost two decades follow up on the Kleefstra syndrome (KLEFS1; OMIM #610253) – a clinically recognizable, autosomal dominant neurodevelopmental condition caused by haploinsufficiency of *EHMT1*[8] (OMIM #607001). *EHMT1* encodes euchromatic histone methyltransferase 1 which is an important epigenetic machinery component[9]. EHMT1 (also known as G9a-Like protein, GLP) in heterodimeric complex with EHMT2 (G9a) and along with other proteins represses gene expression by mono- and di-methylating histone 3 lysine 9 tail (H3K9me1-2)[10-12]. Additionally, it promotes DNA methylation by recruiting DNA methyltransferase 3A (DNMT3A) to the repressed chromatin – providing a link between histone and DNA modifications[13].

Key characteristics of KLEFS1 are moderate to severe intellectual disability (ID) and/or global developmental delay (GDD), autism spectrum disorder (ASD), characteristic facial gestalt and multisystemic involvement[14]. The most prominent facial features are microcephaly, synophrys, mildly upslanted palpebral fissures, midface hypoplasia, coarse facies, protruding tongue and relative prognathism[14]. In addition, multiple organ systems may be involved, and the condition is associated with various clinical conditions and congenital anomalies e.g. endocrine abnormalities including hypothyroidism; feeding difficulties during infancy; obesity; cardiac defects; and skeletal anomalies including scoliosis[14,15]. Additionally, KLEFS1 is characterized by a presence of a specific DNA methylation (DNAm) signature[16,17].

As there is extensive interest in the role of EHMT1/GLP in cell biology[10,18,19] and cancer[20], the increase in *EHMT1* variants identified in individuals with NDD not only allows further accurate characterization of both the phenotypic and the molecular aspects, but also to a broader understanding of EHMT1 functions. To achieve this, we focused on different variant classes with different effects: the protein altering variants (PAVs) (including missense and inframe indel variants), the N-terminal truncating and "classical" loss-of-function variants (both intragenic and multigene). PAVs are often interpreted as VUS. These variants are spread throughout the gene,

both within and outside of the EHMT1 domains. *EHMT1* consist of the "reader" ankyrin repeat (ANKR) domain, the catalytic SET domain which is also involved in heterodimer formation with EHMT2[10] and a recently-identified RING-like domain of unknown function[21]. Understanding of consequent biological aspects and protein function is relevant not only to obtain insights in gene function, but also to allow correct variant interpretation, proper genetic counseling, and adequate clinical management.

To comprehensively investigate the KLEFS1 molecular pathomechanisms in relation to EHMT1 functions, we utilized: 1) *in silico* and *in vitro* functional analyzes; 2) DNA methylation signature analyzes; 3) systematically collected data from a large clinical KLEFS1 cohort with various *EHMT1* variant types. Based on our findings, we provide the current state of knowledge about genotype-phenotype correlations and pathomechanisms of the Kleefstra syndrome.

Methods

Recruitment

Firstly, to provide details on unbiased molecular diagnostics of Kleefstra syndrome, we collected coded data of all individuals with likely pathogenic/pathogenic (LP/P) *EHMT1* variants (including point variants and structural variants like 9q34.3 deletions) from 2004 to 2022 at the genetic laboratory of the KLEFS1 local expertise center. Individuals with 9q34.3 duplications affecting the entire *EHMT1* were not included, as they are not associated with the *EHMT1* haploinsufficiency or KLEFS1[22]. The KLEFS1 diagnostics was provided for the local, as well as (inter-)national individuals with suspected KLEFS1 (and other NDDs). Patients' sex, age at diagnosis, genetic variant and diagnostic method(s) were available for all individuals.

Additionally, to evaluate the prevalence of KLEFS1, we used exome sequenced (ES)-diagnosed individuals since 2013, when (trio) ES with CNV calling became the first-line diagnostic test for individuals with NDD and compared it to the total NDD ES requests at our center. Details are provided in the Supplemental note: KLEFS1 prevalence calculation.

Lastly, to interrogate the *EHMT1* pathogenic variant and clinical spectrum, we collected detailed clinical and molecular data from individuals with rare *EHMT1* variants (classified as VUS or LP/P) who were seen at the expertise center, as well as from the international and national collaborations, who have contacted our KLEFS1

expertise center. As the majority of the known pathogenic variants are truncating, we additionally focused on collecting various PAVs and additionally searched and contacted physicians/researchers regarding the individuals with VUS and/or LP/P *EHMT1* PAVs or N-terminal truncating variants in the DECIPHER[23,24], Dutch clinical laboratory VKGL datashare[25], and ClinVar[26] databases. The clinical information and facial images were provided by their respective physicians and registered in a Castor database.

Protein 3D structure analysis

The ANKR domain was analyzed using PDB:6BY9 and PDB:3B95[27] structures. The SET domain (with pre-SET and post-SET domains) was analyzed using PDB:2RFI[28]. Additional potential domains were searched using a full-length AlphaFold2 structure[29,30]. An identified region with tertiary structure (p.510-650) mapping to the region previously described as cysteine-rich region was aligned using BLAST against the PDB[31] and Swissprot[32] databases. The only proteins containing similar sequence were human and mouse EHMT1/GLP and EHMT2/G9a; the same region in EHMT2 has recently been solved structurrally as the RING-like domain PDB:6MM1[21]. This structure was similar to the AlphaFold2 predicted respective region of the EHMT1 model (RMSD=1.1 A). The AlphaFold2 model was filled with cofactors using AlphaFill[33].

Nonsense mediated decay and translation initiation start site predictions

The regions that are predicted to escape from nonsense mediated decay (NMD) (when containing premature termination codons) were predicted using NMDetective[34]. For the N-terminal truncating variants, possible alternative AUG start codons for the MANE transcript NM_024757.5 were predicted using an overlap between two translation initiation site predictors that preserved the reading frames: ATGpr[35] (reliability score >0.2) and TISpredictor (Kozak similarity score >0.7)[36].

EpiSign

Blood-derived DNA from individuals with *EHMT1* PAVs or other VUS were run on Infinium MethylationEPIC V1 array and further analyzed using KLEFS1 signature on EpiSign V4, as described previously[17,37,38]. This test provides PCA, heatmap, as well as SVM-based score ranging 0-1 allowing to classify a case as having Kleefstra signature or not. Additionally, the samples were classified over 90 other signatures.

In vitro EHMT1 variant effect analysis

To confirm the results of the PAV interpretation and investigate their molecular effects on EHMT1, we selected different *EHMT1* benign, VUS and LP/P PAVs from each genotype-subgroup, similarly as described before[39]. As controls, we included previously tested disruptive ANKR and SET variants p.Pro809Leu, p.Cys1073Tyr, p.Arg1197Trp, p.Cys1203Ala[12,39]. We evaluated mutant EHMT1 binding function to EHMT2, binding function to H3K9me2, as well as enzymatic activity to methylate H3K9me2 and the protein's (thermo)stability.

In brief, for the EHMT2 binding assay, FLAG tagged EHMT1 and GFP tagged EHMT2 vectors were transfected into 293T cells, and immunoprecipitated by FLAG beads. EHMT2 bound with FLAG-EHMT1 was analyzed by Western blotting. For the methyltransferase assay, binding assay to H3K9me2 and thermostability assay, recombinant MBP-EHMT1(635-1298aa)-His proteins were produced in bacteria and purified with Ni-NTA agarose beads. The binding assay to H3K9me2 was performed as described previously with minor modification[11]. To examine pure binding affinity of the ANK repeat domain, UNC0642 (1 μM) was added to the binding reaction (0.42 M NaCl condition) to exclude the H3K9 affinity by SET domain. Tycho NT.6 (NanoTemper) was used for thermostability assay. The methyltransferase assay was performed using recombinant MBP-EHMT1(635-1298aa)-His proteins as previously described[39]. Purified recombinant wild-type and mutant EHMT1 proteins were incubated with recombinant Histone H3 protein and SAM for 2 hours at 30˚C. Methylated Histone H3 was analyzed by Western blotting using anti-H3K9me2 (clone #6D11) and pan H3 antibodies.

For H3K9me2 recovery assay *in vivo*, FLAG tagged full length wild-type and mutant EHMT1 proteins were expressed in *EHMT1* KO HeLa cell, and H3K9me2 status was examined by Western blottings.

Variant reclassification

The identified variants were mapped to the latest genome and *EHMT1* MANE transcript versions (GRCh38 and NM_024757.5, respectively). The variants were reclassified using the ACMG guidelines[40], utilizing: 1) a patient's clinical feature fitness to KLEFS1 spectrum evaluated by our team; 2) protein 3D structure analysis; 3) DNAm classification on Kleefstra syndrome DNA methylation (DNAm) signature using EpiSign[17]; 4) *in vitro* EHMT1 functions evaluation (for selected variants); 5) up-to-date information about variant population frequency, inheritance, recurrence and location within gene, and *in silico* tools predictions (REVEL[41] and AlphaMissense[42] for PAVs and SpliceAI[43] for splicing effect predictions). The variants

and their interpretations were submitted to the ClinVar database (ClinVar accession numbers: SCV005045804-SCV005045896, SCV005049504, and SCV005049572-SCV005049646). Only individuals with LP/P *EHMT1* variants were included in the further KLEFS1 phenotype analysis.

Genotype-phenotype analysis

To assess genotype-phenotype correlations, we conducted descriptive analyzes and utilized the *PhenoScore* software[44], which is a recently-developed machine learning model and includes both facial image and clinical Human Phenotype Ontology (HPO) terms[45]. Six subgroups based on the underlying LP/P genetic variant causing KLEFS1 were determined in advance for the correlation analyzes: 1) multigene copy number variants (CNVs) that include *EHMT1* and at least one additional protein coding gene; 2) isolated *EHMT1* null (haploinsufficiency) variants (nonsense, frameshift, splice-site variants, as well as intragenic deletions and deletions that do not affect other protein-coding genes); 3) Protein altering variants (PAVs) in ANKR domain; 4) PAVs in SET domain; 5) N-terminal truncating variants; 6) Large inframe duplication. The individuals who do not fit these groups were not included in the analysis due to a small sample size. Individuals with a known dual molecular diagnosis were excluded from the analysis to focus on only KLEFS1 features.

Facial and HPO data from group 2 were compared to those of every other group and participants were sex- and age-matched. Linear growth data was compared to the Centers for Disease Control and Prevention (CDC) sex- and age-matched references. A short stature was defined as a height below the 3rd percentile, while a tall stature was defined as height exceeding the 97th percentile. Overweight was defined as BMI ranging between Z-score 1 and 2, whereas obesity was characterized as BMI ≥2 SD.

Quantitative genotype-phenotype correlation analysis was performed using descriptive analyzes for clinical symptoms with a prevalence over 25%. The Fisher-Freeman-Halton test and Bonferroni correction were used, and a two-sided p value <0.05 was considered as significant.

Qualitative genotype-phenotype correlation was analyzed using PhenoScore and was performed using Python 3.9, statistically significant differences were determined using the Wilcoxon signed-rank test, Brier scores, and area under de curve (AUC).

Ethics

Ethical approval for clinical and molecular studies was obtained from the Ethical Committee Arnhem-Nijmegen (#2018-4540). Molecular and clinical data from participants was provided by their clinicians or included in the Radboudumc biobank *Genetics and Rare disease*. This biobank is approved by the ethical committee (#2018-4985) and is part of the Radboudumc biobank initiative. Explicit permission to publish photographs was received from all participants concerned. International participants gave written consent for data sharing to their treating physicians.

Results

Recruited cohort

In total, we recruited 209 individuals, from 200 independent families, with a rare genetic variant involving *EHMT1*. Most individuals (79/209) had deletions (65 multigene involving *EHMT1* and at least one additional gene and 14 affecting only *EHMT1*); three had large inframe gains; one had a balanced translocation disrupting *EHMT1*; 67 individuals had a null point variant (nonsense, splice variant, frameshift); 55 had PAVs (including missense and short inframe deletions <50bp); 4 had (predicted) silent variants. To provide an accurate characterization of the Kleefstra syndrome genotypic and phenotypic spectrum, we sought to investigate all identified variants in detail.

Variant effect prediction on the protein structure

EHMT1 protein contains three known domains: 1) SET domain (including pre-SET and post-SET); 2) ANKR domain and 3) RING-like domain, while the rest of the protein is largely disordered (**Figure 1 and 2**).

Truncating variants

The majority of truncating variants and deletions were predicted to result in *EHMT1* haploinsufficiency (via deletion or nonsense mediated decay) (**Figure 1**). One *de novo* frameshift variant c.3703_3704del p.(Gly1235Argfs*6) is predicted to escape NMD (**Figure S1**), but would result in a loss of a significant part of the SET domain (**Figure 1**) likely leading to a functional loss.

Interestingly, we also identified a sub-group of individuals with atypically mild KLEFS1 phenotype (described in more detail in the KLEFS1 genotype-phenotype correlations section) with N-terminal truncating variants clustering in the first three exons that are predicted to evade NMD (**Figure S1**). Strong alternative start

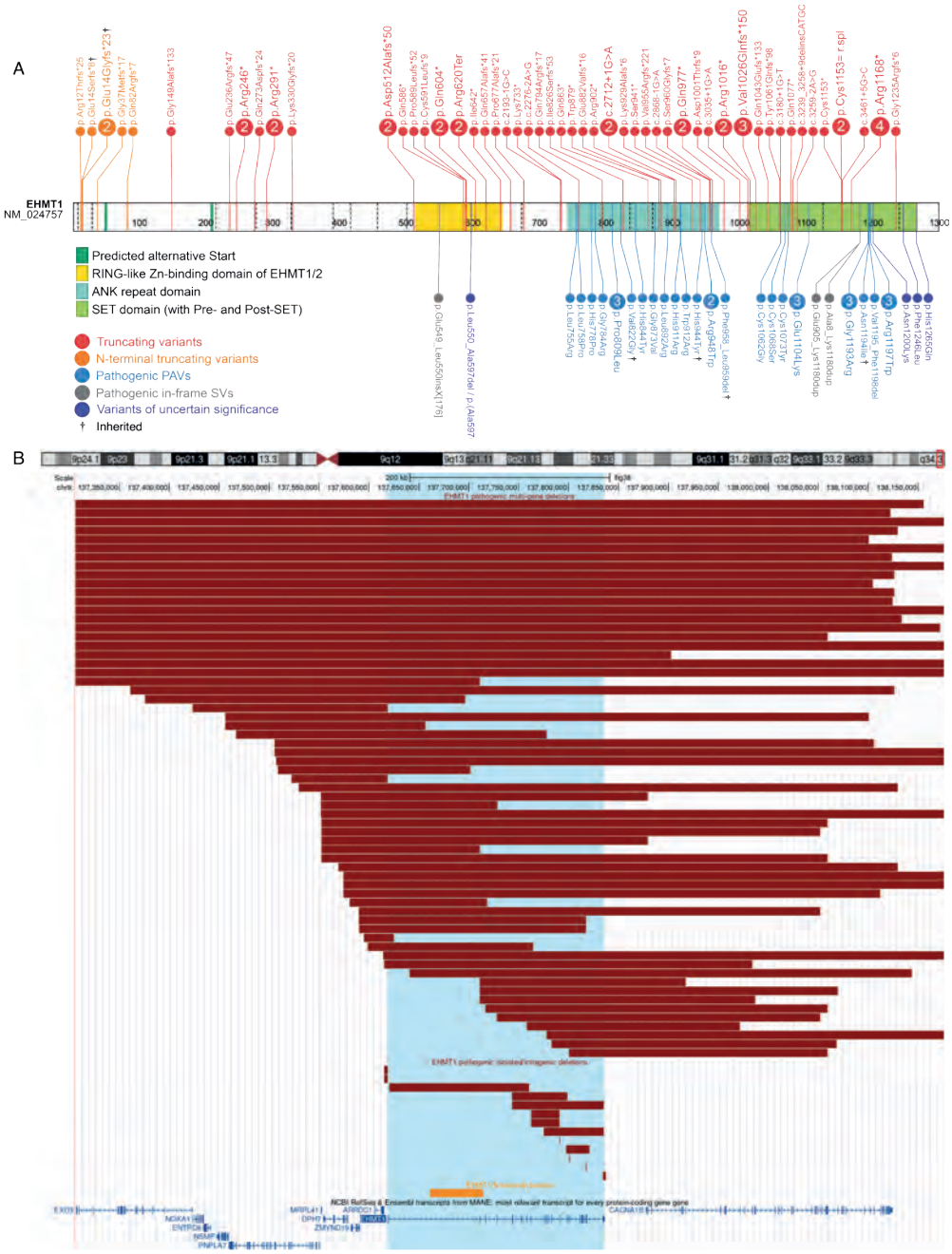


Figure 1. *EHMT1* LP/P variants and VUS identified in this study.

A. *Intragenic location of the EHMT1 point and structural non-deletion variant that were classified as LP/P or VUS;* ***B.*** *Genomic location of the identified 9q34.3 deletions affecting EHMT1 (region highlighted in blue).*

codons were predicted at p.Met48 and p.Met207 [35,36]. Additionally, two protein coding transcripts (NM_001354612.2 and NM_001354259.2) utilize p.Met32 (of the canonical transcript) as a start codon. Therefore, such N-terminal truncating variants are predicted to result in a shorter protein lacking a disordered N-terminal region without known functions.

Synonymous variants

We identified three *de novo* synonymous variants in four unrelated individuals. The c.3459C>T p.(Cys1153=) variant is predicted with high confidence to result in a cryptic donor gain (DG) resulting in a frameshift (SpliceAI DG=0.95 and Pangolin = 0.7 -2bp); the c.1791G>A p.(Ala597=) has low confidence prediction of causing a cryptic donor gain in intron (SpliceAI DG=0.17, 114bp) or loss of the wild-type donor site (Pangolin = 0.25, 0bp), which would result in a skipping of an inframe exon and disrupting the RING-like domain. Lastly, the c.3612G>A p.(Glu1204=) is not predicted to cause a splicing defect (SpliceAI and Pangolin <0.1).

Protein altering variants

We identified 38 unique (mostly *de novo*) PAVs: 35 missense and 2 small inframe deletions (**Figure S2**). Eight N-terminal missense variants (c.589G>A p.(Asp197Asn), c.623C>T p.(Pro208Leu), c.750A>C p.(Leu250Phe), c.823G>A p.(Ala275Thr), c.905A>G p.(Lys302Arg), c.945G>A p.(Met315Ile), c.1231G>A p.(Gly411Ser), c.2159T>G p.(Leu720Trp)) are located outside of domains, in disordered regions, so these variants are not predicted to affect the EHMT1 structure. The majority of the ANKR domain variants (**Figure 2A**) are predicted to disrupt the domain's structure (c.2273T>C p.(Leu758Pro), c.2333A>C p.(His778Pro), c.2350G>A p.(Gly784Arg), c.2426C>T p.(Pro809Leu), c.2465T>G p.(Val822Gly), c.2530C>T p.(His844Tyr), c.2618G>T p.(Gly873Val), c.2675T>G p.(Leu892Arg), c.2732A>G p.(His911Arg), c.2830C>T p.(His944Tyr), c.2873_2878del p.(Phe958_Leu959del)), as well as to disrupt the binding with the H3 tail (c.2734T>C p.(Trp912Arg), c.2842C>T p.(Arg948Trp)); two variants are located on the domain's surface, outside of known interaction surfaces, and are predicted to be neutral (c.2264T>G p.(Leu755Arg), c.2405G>C p.(Cys802Ser)). The majority of the SET domain PAVs (**Figure 2B**) (c.3184T>G p.(Cys1062Gly), c.3203G>C p.(Cys1068Ser), c.3218G>A p.(Cys1073Tyr), c.3310G>A p.(Glu1104Lys), c.3577G>A p.(Gly1193Arg), c.3577G>C p.(Gly1193Arg), c.3581A>T p.(Asn1194Ile), c.3583_3594del p.(Val1195_Phe1198del), c.3589C>T p.(Arg1197Trp)) are located in the core of the domain and are predicted to disrupt the domain's structure and its functions. Four PAVs are located on a surface of the SET domain without known interaction or active sites (c.3284A>G p.(Asn1095Ser), c.3392A>G p.(Tyr1131Cys), c.3401G>A p.(Arg1134Gln), c.3595A>G p.(Ile1199Val)), so these variants are not predicted to disrupt the domain's

structure or functions. Interestingly, we also identified three variants (c.3600C>G p.(Asn1200Lys), c.3736T>C p.(Phe1246Leu), c.3795C>G p.(His1265Gln)) that were located in or near the active center (SAH binding pocket) and were predicted to disrupt the domain's enzymatic activity only, but possibly leaving the domain's structure largely intact (**Figure 2B**). Detailed description of the predicted PAV effects on the protein 3D structure are provided in the **Table S1**.

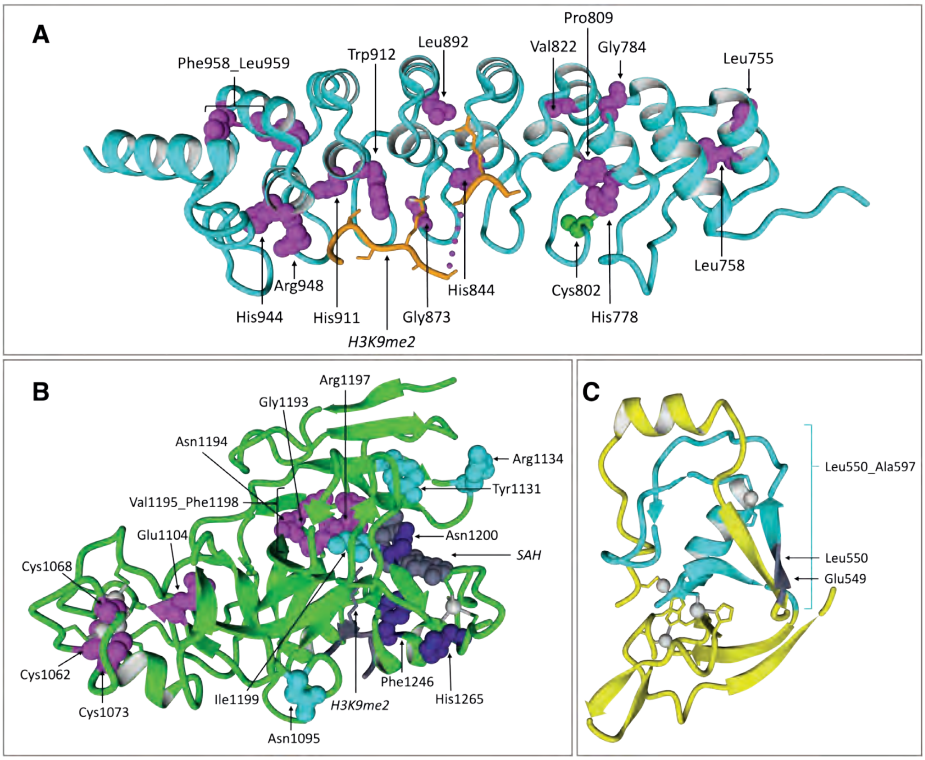


Figure 2. EHMT1 protein altering variant position on the protein 3D structure.

A. *Ankyrin repeat domain (cyan) with H3 tail (orange) (PDB:6BY9; 3B95). Amino Acids affected by variants predicted to disrupt the domain's structure or binding to the H3 tail are shown in magenta and neutral variants – in green.* ***B.*** *SET domain (with pre-SET and Post-SET domains)(green) with H3 tail and SAH (in dark grey) and Zn atoms (grey) (PDB:2RFI). Amino acids affected by variants predicted to disrupt the domain's structure are shown in magenta, variants that are predicted to disrupt the enzymatic activity only – in purple and neutral variants – in cyan.* ***C.*** *RING-like domain (yellow) with Zn atoms (grey). The region removed by skipping the inframe exon due to c.1791G>A is shown in cyan, while the region of insertion c.1647_1648ins643_1170 is shown in dark grey.*

Inframe duplications

In addition, we identified three unique large intragenic gains that are not predicted to disrupt the reading frame (**Figure 1A**). The first resulted in tandem duplication of

the majority of the gene from exon 2 to 25 c.22-?_3540+?dup p.Ala8_Lys1180dup. The next gain involved exons from 19 to 25, so is predicted to result in c.2713-?_3540+?dup p.(Glu905_Lys1180dup), if in tandem. Finally, the third gain was initially identified by exome sequencing and involved exons 4 to 6, although genome sequencing revealed that it was not in tandem, but, in fact, inserted between exons 10 and 11, thereby likely resulting in c.1647_1648ins643_1170 p.(Glu549_Leu550insX[176]) and disrupting the RING-like domain (**Figure 2C**). While the functional consequences of the identified gains are unknown, we hypothesize that these gains could reduce EHMT1 stability and/or affect interactions across the EHMT1 domains.

DNAm signature testing

We utilized the robustness of KLEFS1 DNAm signature to simultaneously classify an extensive cohort of individuals with VUS and non-truncating variants: 42 individuals with 33 different PAVs, 3 individuals with unique inframe duplications, 4 individuals with 3 different synonymous variants and 7 individuals with 5 different N-terminal truncating variants (**Table S2 and Table S3**).

Truncating variants

The *EHMT1* haploinsufficiency-causing variants were used as the "baseline" for the classifications. Two out of nine individuals with the most distal N-terminal truncating variants (c.109delinsAT p.(Gly37Metfs*17) and c.244del p.(Gln82Argfs*7)) were not available for testing. Out of the remaining seven individuals with N-terminal variants, only one (c.34_35insC p.(Arg12Thrfs*25)) tested "intermediate" on the KLEFS1 DNAm signature, while the remaining six individuals were classified "negatively". These findings support a different functional effect of this variants group in comparison to the haploinsufficiency-causing variants.

Synonymous variants

Two unrelated individuals with the same *de novo* variant c.3459C>T p.(Cys1153=) were classified "positively" for the KLEFS1 DNAm signature, supporing the predicted splicing defect related to this variant. The other two individuals with *de novo* synonymous variants c.1791G>A p.(Ala597=) and c.3612G>A p.(Glu1204=) were classified "negatively".

Protein altering variants

In total, 42/55 individuals with 33 unique PAVs were available for testing. Twenty nine out of 42 tested individuals, with 20 different PAVs were classified "positively" on the KLEFS1 DNAm signature, while 13/42 individuals with 13 unique PAVs were

classified as "negative" on the KLEFS1. All PAVs classified as KLEFS1 were located within the ANKR or SET domains. The DNAm signature classification results have high concordance with the predicted variant effects for 28 out of 33 PAVs. To our surprise, the three SET domain PAVs (p.(Asn1200Lys), p.(Phe1246Leu) and p.(His1265Gln)) that are predicted to disrupt the enzymatic activity only (**Table S1**) were classified as "negative" on the KLEFS1 DNAm signature. This suggests that these variants may have a different functional effect compared to the other SET domain and haploinsufficiency-causing pathogenic *EHMT1* variants and do not result in KLEFS1 aberrant DNA methylation. DNAm testing was not performed for 13/55 individuals with PAVs, but 5/13 untested individuals have a recurrent variant and at least one other individual with the same variant was tested; for 2/13 untested individuals – their relatives with the same variant were tested and for 1/13 individual with the c.3577G>A p.(Gly1193Arg) variant – two individuals with another variant affecting the same nucleotide were tested (c.3577G>C p.(Gly1193Arg)). Therefore, the DNAm testing results were not available only for 5/39 of the unique PAVs (p.(Pro208Leu), p.(Ala275Thr), p.(Gly873Val), p.(Asn1095Ser), p.(Val1195_Phe1198del)).

Inframe duplications

Two out of three individuals with large inframe duplications of exons 2-25 and 19-25 (p.Ala8_Lys1180dup and p.(Glu905_Lys1180dup), respectively) were classified "positively" on the signature. The third individual with the non-tandem inframe duplication of exons 4-6 p.(Glu549_Leu550insX[176]) was classified as "intermediate".

Other signatures

All individuals who classified as "positive" or "intermediate" on the KLEFS1 DNAm signature, were classified "negatively" on all other EpiSign V4 available signatures (N=90). However, one individual with two inherited (benign) *EHMT1* variants in compound heterozygous state p.[(Arg1134Gln)];[(Ile1199Val)], who classified "negatively" on the KLEFS1 DNAm signature, classified "positively" on the MRD23 (OMIM #615761) DNAm signature. However, no pathogenic variants in the *SETD5* (OMIM #615743), *ANKRD11* (OMIM #611192) or other genes were identified.

In vitro variant effect analysis

Synonymous variants

Due to the contradictory splicing effect predictions, we tested the effect of the *de novo* c.1791G>A p.(Ala597=) variant in patient-derived fibroblasts. The RNA analysis showed an inframe skipping of exon 11 (data not shown), likely resulting in p.Leu550_Asn597del, which would disrupt the RING-like domain (**Figure 2C**).

N-terminal truncating variants

To explore the effect of the N-terminal truncating variants, we tested the patient-derived fibroblasts with the variant c.21+42335_86-2217del p.(Ala8Argfs*60). Protein immunoblot showed a single band of ~180kDa in the control line, while in proband and paternal cells additional presence of a shorter band of similar intensity was evident (**Figure S3**). This supports the predicted effect evading NMD and the use of an alternative start site to initiate translation.

Protein altering variants

To confirm the predicted effect of PAVs, we comprehensively analyzed various EHMT1 functions and stability for different selected PAVs. We were able to confirm the variant effect predictions on the protein structure and identified additional insights (**Table 1**).

Firstly, we examined histone methyltransferase activity using recombinant ANKR+SET domain proteins (aa 635-1298) (**Figure S4A**). We used previously reported p.Cys1073Tyr and p.Arg1197Trp variants as positive controls, which lost both enzymatic activity, as well as binding ability to EHMT2[39]. We found that the two variants p.Asn1200Lys and p.Phe1246Cys, which were classified negatively on the DNAm signature, disrupt the domain's enzymatic activity (**Figure 3A**). ANKR domain mutants and p.His1265Gln as well as benign variant p.Tyr1131Cys did not affect the histone methyltransferase activity (**Figure 3A**).

Next, we performed the EHMT2 binding assay by expression of EHMT1 and EHMT2 full length proteins in 293T cells. In contrast to p.Cys1073Tyr and p.Arg1197Trp, all ANKR examined variants, as well as the SET domain variants p.Asn1200Lys, p.Phe1246Cys and p.His1265Gln can bind to EHMT2 (**Figure 3B**). In the regulation of EHMT1-EHMT2-dependent H3K9 methylation (H3K9 di-methylation) in cells, the heterodimer formation is essential but EHMT1 enzymatic activity has been shown to be dispensable if EHMT2 enzymatic activity is functional[12]. It has been reported that H3K9 methylation in cells is catalyzed by the EHMT2 enzymatic activity, but not EHMT1 activity. Therefore, p.Asn1200Lys and p.Phe1246Cys, which lost enzymatic activity but kept EHMT2 binding ability, could restore the H3K9 dimethylation level in EHMT1 knockout HeLa cells (**Figure 3C**), suggesting that loss of EHMT1 enzymatic activity is not sufficient to generate the typical KLEFS1 DNAm signature or phenotype. The ankyrin repeat domain of EHMT1/2 has been reported to bind methylated H3K9[27], so we analyzed whether ANKR domain mutations affect H3K9 binding ability. As we expected, p.Pro809Leu, p.Trp912Arg and p.Arg948Trp showed reduced binding affinity for methylated H3K9 peptide (**Figure S4B**), indicating that these mutations decreased EHMT1's "reader" function and, subsequently, "reader" function alone is sufficient to cause Kleefstra

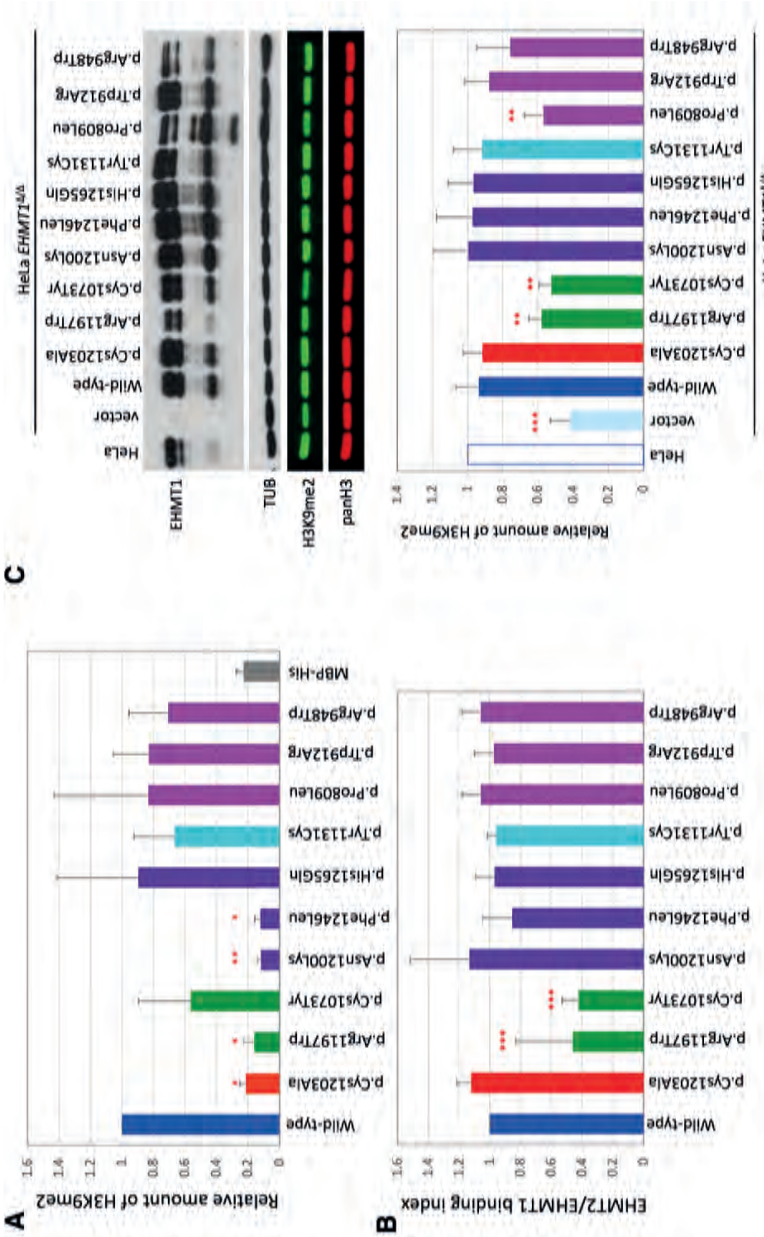


Figure 3. *In vitro* assay results for the *EHMT1* protein altering variants.

A. *In vitro methylation assay was performed using recombinant MBP-EHMT1(635-1298aa)-His proteins. Methylated histone H3 was examined by Western blottings. Graph showed relative amount of H3K9me2 (n=3 independent experiments, means±SD, *p<0.05, **p<0.01, One way ANOVA analysis, Dunnett's multiple comparison test).* ***B.*** *FLAG tagged EHMT1 proteins were expressed together with GFP tagged EHMT2 in 293T cells. EHMT2 protein bound to EHMT1 were precipitated using FLAG M2 beads from cell lysates. Graph showed relative amount of co-precipitated EHMT2 with EHMT1 (n=3 independent experiments, means±SD, ***p<0.001, One way ANOVA analysis, Dunnett's multiple comparison test).* ***C.*** *FLAG tagged EHMT1 wild-type and mutants were stably expressed in EHMT1 knockout HeLa cells. Histone H3 dimethylation level was analyzed by Western blotting. Graph showed relative level of H3K9me2 (means±SD, **p<0.01, ***p<0.001).*

syndrome's DNAm signature and phenotype. In addition to reduced H3K9 binding affinity, p.Pro809Leu and p.Arg948Trp showed thermal instability compared to wild-type or other mutants. It has been already suggested that p.Pro809Leu led to protein misfolding[46]. Our data also showed that the p.Pro809Leu mutant protein is unstable (**Figure S4C**), and expression of p.Pro809Leu could not restore H3K9 di-methylation level in *EHMT1* knockout cell (**Figure 3C**). The variant p.His1265Gln, which was also classified negatively on the DNAm signature, did not show any effect in any of the performed assays and behaved similarly to the wild-type and the benign control variant p.(Tyr1131Cys).

Table 1. Summary of protein altering variant functional effect evaluation.

Variant	Domain	EHMT2 binding	HMT activity	H3 tail peptide binding	Protein stability		H3K9me2 recovery in cells
					Stability in cells	Thermo-stability	
WT	NA	+	+	Baseline	Stable	Baseline	+
p.Cys1203Ala	SET	+	-	NA	Stable	Unchanged	+
p.Arg1197Trp	SET	-	-	NA	Unstable	Unchanged	-
p.Cys1073Tyr	SET	-	weak	NA	Stable	Unchanged	-
p.Asn1200Lys	SET	+	-	NA	Stable	Reduced	+
p.Phe1246Leu	SET	+	-	NA	Stable	Unchanged	+
p.His1265Gln	SET	+	+	NA	Stable	Unchanged	+
p.Tyr1131Cys	SET	+	+	NA	Stable	Unchanged	+
p.Prp809Leu	ANKR	+	+	Decreased	Stable	Reduced	-
p.Trp912Arg	ANKR	+	+	Decreased	Stable	Unchanged	+
p.Arg948Trp	ANKR	+	+	Decreased	Stable	Reduced	+

The variants p.Cys1203Ala, p.Arg1197Trp and p.Cys1073Tyr have been used as a positive controls with previously proven functional effects[39]. The p.Tyr1131Cys was classified as benign, so was used as a negative controls. WT = Wild-type; + = present; - = absent; NA = not applicable; HMT = histone methyltransferase; H3K9 = histone 3 lysine 9.

EHMT1 variant (re-)classification

To provide an accurate Kleefstra syndrome description, we have reinterpreted all identified variants based on the ACMG guidelines[40] utilizing the latest data, as well as evidence obtained during this study (**Table S3**).

Truncating variants

All deletions (N=76, found among 77 individuals) and truncating variants (N=46, found among 69 individuals) which are expected to result in *EHMT1* haploinsufficiency, as well as one balanced translocation disrupting *EHMT1* were classified as pathogenic,

because these variants are absent in gnomAD V2.1.1.[47] and haploinsufficiency is a well-known KLEFS1 disease mechanism[14]. Only three haploinsufficiency-causing variants were inherited from a parent: two – from a parent with a mosaic variant and one – from an affected parent (described below).

Importantly, we also identified a sub-group of 9 individuals from 7 families with a mild KLEFS1 phenotype with seven different N-terminal frameshift variants (one deletion of the second exon and 6 small indels) that are predicted to evade NMD with likely reinitiation of translation. Therefore, they were classified as a separate group (**Table S3**). In two families, the variant occurred *de novo* (c.38_39insA p.(Glu14Glyfs*23) and c.244del p.(Gln82Argfs*7)). In another family, a proband with an alternative condition (*de novo* pathogenic truncating *ASXL3* variant) was identified to also carry the deletion of the second exon p.(Ala8Argfs*60) which was inherited from an unaffected heterozygous father who inherited the variant from an unaffected grandmother who was shown to be mosaic for this variant. In two families the inheritance is unknown (c.34_35insC p.(Arg12Thrfs*25) and c.109delinsAT p.(Gly37Metfs*17)), and in two other families – the variants were inherited from an affected or unaffected parent (c.40dup p.(Glu14Glyfs*23), c.40del p.(Glu14Serfs*8), respectively), but the grandparents were unavailable. Importantly, in 5/7 families the facial phenotype was consistent with the diagnosis of Kleefstra syndrome. Finally, none of the tested individuals classified "positively" on the KLEFS1 DNAm signature and one – as "intermediate". Therefore, we currently classify these variants as LP with expected hypomorphic effect and incomplete penetrance, based on the typical facial phenotype, DNAm testing results and *de novo* occurrence or segregation of the variant in the vast majority of the identified families.

Synonymous variants

Three *de novo* variants were predicted as silent, but two of those are (likely) affecting splicing. The variant c.3459C>T p.(Cys1153=) was identified as *de novo* in two independent individuals with KLEFS1 clinical phenotype, predicted to disrupt splicing, and both individuals were classified as Kleefstra syndrome on DNAm. Therefore, this variant was classified as pathogenic and splicing variant c.3459C>T r.spl.

Next, the variant c.1791G>A p.(Ala597=) was identified in an individual with a phenotype not consistent with Kleefstra syndrome and was classified as control on the DNAm signature. However, the variant occurred *de novo* and was confirmed to result in an inframe skipping of an exon in the individual's fibroblasts-derived mRNA, likely resulting in p.Leu550_Asn597del and disrupting the RING-like domain. Therefore, given the contradictory evidence, we classified this variant as VUS.

Finally, the variant c.3612G>A p.(Glu1204=) was identified in an individual with mild DD and some dysmorphic features as *de novo*. This variant was not predicted to disrupt splicing and was classified negatively on the DNAm signature. Additionally, the individual has a frameshift variant in *SHANK2* (OMIM #603290), which recently has been associated with a non-syndromic NDD[48]. Therefore, *EHMT1* c.3612G>A variant was reclassified as benign.

Protein altering variants

Out of 39 analyzed PAVs, 23 PAVs identified in 39 individuals were reclassified as LP/P, while 13 – as benign and 3 as VUS (**Figure S2**). All pathogenic PAVs were located exclusively in the ANKR (n=14) or SET (n=10) domain and being predicted and/or shown functionally to disrupt functions of these domains, as well as were identified in individuals with a phenotype consistent with Kleefstra syndrome. The benign variants were either located outside domains, in the N-terminal part, or were located in the ANKR or SET domain, but were either too common in gnomAD V2.1.1.[47] and/or located on the domain's surface without predicted effects on its structure.

Finally, three SET domain PAVs (identified in 4 individuals from 3 families) were classified as VUS due to contradictory evidence: the variants were classified as controls by DNAm signature but were predicted to disrupt the SET domain's enzymatic activity, while preserving its structure. This was functionally confirmed *in vitro* for the two variants with unknown inheritance (p.Asn1200Lys and p.Phe1246Leu). While the third variant did not display any functional effects *in vitro*, but was *de novo* (p.His1265Gln). The phenotype was not typical of KLEFS1 for all four individuals. Therefore, we can neither exclude nor confirm the pathogenicity of these three variants, because they have a different functional effect than the other KLEFS1 causing pathogenic *EHMT1* PAVs.

Inframe duplications

The three large inframe duplications were classified as pathogenic as 2/3 were proven as *de novo* and in one individual – the inheritance was unknown, the individuals fit the mild clinical spectrum of Kleefstra syndrome (described below) and were all classified as Kleefstra syndrome on DNAm.

Kleefstra syndrome genotype spectrum

After variant reclassification, *EHMT1* variants were classified as LP/P in 191 individuals (**Figure 1**). In 86 individuals the variant occurred *de novo*, 6 were inherited from parents with a mosaic variant, 8 from an affected parent, 3 from a parent with

a balanced translocation and the inheritance of the remaining 88 individuals is unknown. The study encompasses six different genetic variant groups: 1) 9q34.3 deletions affecting *EHMT1* and at least (part of) one additional protein-coding gene (further referred as multigene deletions for simplicity) (65/191), 2) protein truncating variants and deletions resulting in haploinsufficiency of only *EHMT1* (further referred as intragenic null variants) (75/191), 3) ANKR domain PAVs (23/191), 4) SET domain PAVs (16/191), 5) N-terminal protein truncating variants (9/191), and 6) large inframe duplications (3/191). However, the distribution of the variants in the unbiased cohort from the KLEFS1 expertise center is different (**Figure S5** and Supplemental note: KLEFS1 diagnostic experience), because we have focused on recruiting individuals with non-truncating *EHMT1* variants.

Kleefstra syndrome clinical spectrum

In five individuals out of 191 a second molecular diagnosis likely contributing to the phenotype was identified, so they were excluded from the further phenotype analysis, resulting in a study cohort of 186 individuals. Detailed medical information reported by medical professionals was available for 125 of them (**Figure 4**). The majority of the individuals originated from the Netherlands (58/125, 46%) (**Table S2**). The KLEFS1 clinical spectrum is shown in **Table 2** and described in detail in the Supplemental note: description of KLEFS1 clinical features.

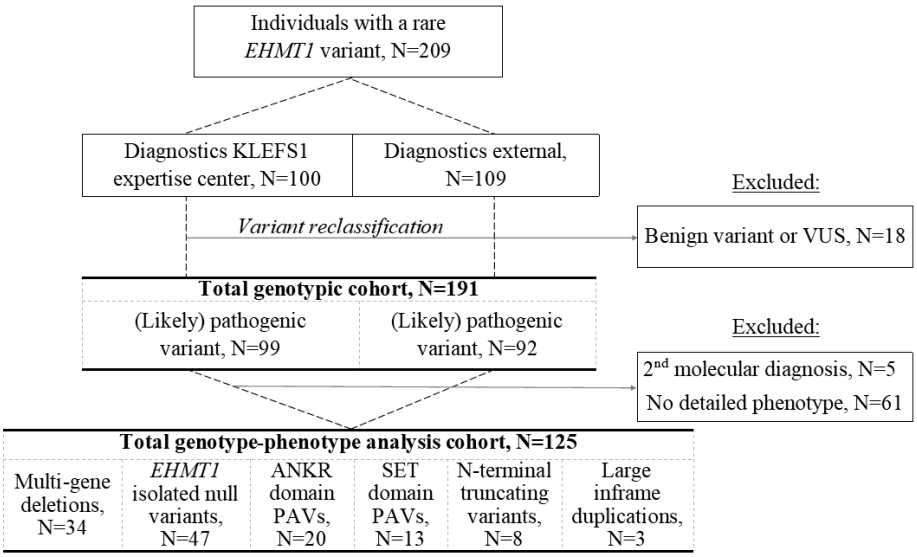


Figure 4. Flowchart of the individuals included and excluded in this study.

Table 2. Kleefstra syndrome clinical spectrum and genotype-phenotype comparison.

Feature	Total cohort	Multi gene deletions	Isolated *EHMT1* null variants
General features	**N=125**	**N=34**	**N=47**
Median age, years (IQR)	14 (8-26)	16 (7-26)	12 (6-26)
Male (%)	44 (35%)	13 (38%)	17 (36%)
Gestation and birth			
Gestational age – a term	85% (106)	85% (27)	79% (38)
Small for gestational age	10% (12)	12% (4)	6% (3)
Large for gestational age	15% (19)	9% (3)	13% (6)
Cesarean section	18% (23)	24% (8)	23% (11)
Nervous system			
ID or learning difficulties	90% (112)	100% (34)	94% (44)
GDD	91% (114)	94% (32)	98% (46)
Epilepsy	27% (34)	35% (12)	26% (12)
Sleep-wake disorders	46% (58)	53% (18)	45% (21)
Insomnia (unspecified)	22% (27)	21% (7)	30% (14)
Sleep onset insomnia	7% (9)	18% (6)	4% (2)
Terminal insomnia	5% (6)	3% (1)	6% (3)
Maintenance insomnia	16% (20)	9% (3)	26% (12)
Sleep apnea	6% (8)	9% (3)	9% (4)
Snoring	3% (4)	0	7% (3)
Motor restlessness during the sleep	2% (3)	6% (2)	2% (1)
Sensory system			
Hearing impairment	29% (36)	38% (13)	30% (14)
Overall visual/eye problems	62% (77)	62% (21)	62% (29)
Strabismus	19% (24)	15% (5)	23% (11)
Hypermetropia	33% (41)	36% (12)	40% (19)
Myopia	7% (9)	9% (3)	9% (4)
Astigmatism	13% (16)	21% (7)	13% (5)
Amblyopia	3% (4)	0	6% (3)
Glaucoma	1% (1)	3% (1)	0
Cerebral visual impairment	2% (3)	3% (1)	4% (2)
Cardiovascular system			
Abnormal heart morphology	31% (39)	47% (16)	32% (15)
Abnormal cardiac septum	19% (24)	25% (12)	15% (7)
Valve insufficiency	2% (2)	3% (1)	2% (1)
Valve stenosis	8% (10)	12% (4)	9% (4)

	ANKR domain PAVs	**SET domain PAVs**	**N-terminal trunc. variants**	***P* value**
	N=20	**N=13**	**N=8**	**-**
	13 (4-36)	15 (11-27)	21 (13-28)	
	8 (40%)	4 (31%)	1 (13%)	
	81% (19)	85% (11)	85% (9)	
	15% (3)	8% (1)	0	
	20% (4)	23% (3)	25% (2)	
	15% (3)	8% (1)	0	
	85% (17)	92% (12)	50% (4)	0.001
	95% (19)	100% (13)	50% (4)	0.001
	20% (4)	38% (5)	13% (1)	0.548
	50% (10)	46% (6)	38% (3)	0.838
	20% (4)	15% (2)	0	
	5% (1)	0	0	
	5% (1)	8% (1)	0	
	15% (3)	15% (2)	0	
	0	8% (1)	0	
	5% (1)	0	0	
	0	0	0	
	10% (2)	23% (3)	50% (4)	0.130
	60% (12)	54% (7)	63% (5)	0.993
	25% (5)	8% (1)	25% (2)	
	20% (4)	23% (3)	25% (2)	0.130
	0	0	13% (1)	
	5% (1)	15% (2)	0	
	0	0	13% (1)	
	0	0	0	
	0	0	0	
	10% (2)	23% (3)	13% (1)	0.042
	5% (1)	8% (1)	13% (1)	
	0	0	0	
	5% (1)	8% (1)	0	

Table 2. Continued

Feature	Total cohort	Multi gene deletions	Isolated *EHMT1* null variants
Ventricular hypertrophy	2% (2)	3% (1)	2% (1)
Cardiac rhythm disturbance	12% (15)	12% (4)	15% (7)
Palpitations	2% (2)	3% (1)	0
Atrial fibrillation	2% (3)	3% (1)	4% (2)
Tachycardia	5% (6)	6% (2)	4% (2)
Respiratory system			
Pulmonary disease	24% (30)	26% (9)	23% (11)
Asthma	3% (4)	0	4% (2)
Tracheomalacia	4% (5)	9% (3)	2% (1)
Digestive system			
(Neonatal) tube feeding	7% (9)	12% (4)	6% (3)
Feeding difficulties in infancy	32% (40)	41% (14)	43% (20)
Constipation	47% (59)	71% (24)	49% (23)
Gastroesophageal reflux disease	23% (29)	32% (11)	30% (14)
Musculoskeletal system			
Pes planus	31% (39)	38% (13)	32% (15)
Hypotonia	42% (52)	47% (16)	45% (21)
Scoliosis	26% (32)	32% (11)	28% (13)
Immune system			
Recurrent infections	46% (58)	53% (18)	57% (27)
Endocrine system			
Endocrine/metabolic disease	18% (22)	21% (7)	11% (5)
Hypothyroidism	14% (17)	21% (7)	9% (4)
Miscellaneous			
Abnormality of the skin	14% (18)	15% (5)	17% (8)
Orofacial cleft	2% (3)	0	2% (1)
Congenital kidney/urinal abnormalities	20% (25)	32% (11)	21% (10)
Behavioral and psychiatric disorders, age ≥5	**N=98**	**N=27**	**N=36**
Behavioral and psychiatric disorders	91% (89)	96% (26)	89% (32)
Autism spectrum disorder	72% (71)	89% (24)	72% (26)
Anxiety disorder	28% (27)	30% (8)	19% (7)
Regression	36% (35)	37% (10)	36% (13)
Attention-deficit/hyperactivity disorders	13% (13)	7% (2)	14% (5)
Schizophrenia Spectrum and other Psychotic Disorders	16% (16)	22% (6)	11% (4)

	ANKR domain PAVs	SET domain PAVs	N-terminal trunc. variants	*P* value
	0	0	0	
	5% (1)	23% (3)	0	
	0	8% (1)	0	
	0	0	0	
	5% (1)	8% (1)	0	
	25% (5)	31% (4)	0	
	10% (2)	0	0	
	0	0	0	
	10% (2)	0	0	
	15% (3)	23% (5)	0 (0%)	0.036
	40% (8)	31% (4)	0 (0%)	0.002
	10% (2)	15% (2)	0	
	25% (5)	31% (4)	25% (2)	0.890
	35% (7)	38% (5)	13% (1)	0.452
	20% (4)	23% (3)	13% (1)	0.815
	30% (6)	23% (3)	25% (2)	0.056
	25% (5)	31% (4)	13% (1)	
	15% (3)	23% (3)	0	
	15% (3)	8% (1)	13% (1)	
	0	15% (2)	0	
	0	15% (2)	25% (2)	
	N=14	**N=11**	**N=8**	
	100% (14)	82% (9)	100% (8)	0.361
	71% (10)	55% (6)	50% (4)	0.088
	21% (3)	36% (4)	50% (4)	0.396
	50% (7)	36% (4)	13% (1)	0.566
	14% (2)	0	38% (3)	
	21% (3)	18% (2)	13% (1)	

Table 2. Continued

Feature	Total cohort	Multi gene deletions	Isolated *EHMT1* null variants
Bipolar and related disorders	7% (7)	11% (3)	6% (2)
Depressive disorders	14% (14)	19% (5)	11% (4)
Obsessive-Compulsive traits	9% (9)	7% (2)	14% (5)
Aggressive behavior	28% (27)	41% (11)	28% (10)
Autoaggression	11% (11)	22% (6)	11% (4)
Irritability, restlessness, agitation	12% (12)	7% (2)	19% (7)
Apathy, diminished motivation, passivity	19% (18)	22% (6)	14% (5)
Growth parameters	**N=110**	**N=29**	**N=43**
Median height percentile (IQR)	60 (25-90)	30 (5-58)	69 (38-93)
Small stature	9% (10)	17% (5)	5% (2)
Tall stature	12% (13)	3% (1)	16% (7)
BMI	**N=109**	**N=29**	**N=43**
Overweight [Z-score >1 and <2]	36% (39)	34% (10)	37% (16)
Obesity [Z-score ≥2]	12% (13)	7% (2)	9% (4)
Median BMI percentile (IQR)	81 (57-94)	73 (43-92)	79 (62-95)
Head circumference	**N=90**	**N=20**	**N=38**
Microcephaly	20% (25)	38% (13)	13% (6)
Median HC percentile (IQR)	22 (3-54)	2 (1-18)	26 (10-49)

Feature prevalence in total cohort and in five genotypic cohorts are shown in %, with number of individuals with the feature shown in brackets. Individuals with large inframe duplications were not included separately and were not analysed in the genotype-phenotype correlation analysis, due to the small group size (n=3). P value was calculated only for features with prevalence ≥25%.

PAVs = protein altering variants; trunc. = truncating; ID = intellectual disability; GDD = global developmental delay; BMI = body mass index; HC = head circumference; IQR = interquartile range.

Kleefstra syndrome genotype-phenotype correlations

Truncating variants

Analyzing differences in symptom prevalence (for clinical symptoms with a prevalence >25%), significant genotype-phenotype associations were observed for ID/learning difficulties (p=0.001), GDD (p=0.001), constipation (p=0.002), feeding difficulties in infancy (p=0.036), and abnormal heart morphology (p=0.042) with ID/learning difficulties, GDD, and constipation being significant also after Bonferroni correction (**Table 2 and Figure S6**). Next, the nominally significant symptoms were compared across the groups to the *EHMT1* null variant group as the baseline. The majority of individuals with a *EHMT1* null variant had ID/learning difficulties (44/47, 94%), and GDD (46/47, 98%). In this group, the mean IQ was 54, based on available

	ANKR domain PAVs	SET domain PAVs	N-terminal trunc. variants	*P* value
	7% (1)	0	13% (1)	
	21% (3)	0	25% (2)	
	0	9% (1)	13% (1)	
	29% (4)	9% (1)	13% (1)	0.327
	7% (1)	0	0	
	21% (3)	0	0	
	36% (5)	18% (2)	0	
	N=17	**N=12**	**N=6**	-
	71 (36-93)	67 (18-78)	62 (35-98)	
	6% (1)	17% (2)	0	
	18% (3)	0	33% (2)	
	N=17	**N=11**	**N=6**	-
	47% (8)	36% (4)	17% (1)	0.803
	12% (2)	18% (2)	50% (3)	
	87 (64-94)	88 (60-98)	81 (57-94)	
	N=16	**N=9**	**N=4**	-
	15% (3)	15% (2)	13% (1)	
	23 (7-54)	50 (3-79)	46 (4-86)	

data from 12 individuals. Constipation was present in half of the cohort (23/47, 49%), and feeding difficulties in infancy were present in 43% (20/47). Recurrent infections were found in the majority of individuals (27/47, 57%). Structural heart defects were prevalent in 32% of individuals (15/47), with atrial septal defects being the most common (5/47). In this cohort, the mean height is at the 64th percentile, with two individuals having a short stature (<p3) for their age and sex (2/43, 5%).

Multigene deletions

For the 9q34.3 multigene deletions, the IQ score was only known for a minority of individuals (3/34). ID was identified as severe in 36% of individuals (8/22), contrasting with the 9% (3/34) of individuals having severe ID in the group with *EHMT1* null

variants (p=0.01) (**Table 2 and Figure S6**). A higher prevalence of constipation was found in this group (24/34 vs. 23/47, p=0.05). Furthermore, height was significantly shorter (30th vs. 69th percentile, p=0.001), with 17% of the individuals exhibiting short stature (5/29). Structural heart defects were most prevalent in multigene deletions, though not statistically significant (16/34 vs. 15/47, p=0.17). Remarkably, all four identified individuals with an atrioventricular septal defect had a multigene deletion larger than 1 Mb. Surprisingly, no other significant differences were found in the prevalence of clinical symptoms, including epilepsy, and growth parameters when comparing deletions <1 Mb to deletions >1 Mb.

Protein altering variants

Next, we have compared the ANKR and SET PAVs to the intragenic *EHMT1* null variant group (**Table 2 and Figure S6**). The eight individuals with a PAV in the ANKR domain had a higher IQ (mean IQ 65 vs. 54, p=0.03). Furthermore, feeding difficulties were significantly less prevalent (3/20 vs. 20/47, p=0.03), as well as structural heart defects (2/20 vs. 15/47, p=0.06), and recurrent infections had a lower prevalence (6/20 vs. 27/47, p=0.04). Similarly, the five individuals with a PAV in the SET domain also had a higher IQ (Mean IQ 66 vs. 54, p=0.04) and lower prevalence of recurrent infections (3/13 vs. 27/47, p=0.03) (**Figure S6**).

N-terminal truncating variants

Individuals with an N-terminal truncating variant presented with a milder phenotype: they had a significantly higher IQ (Mean IQ 75 vs. 54, p=0.02), and a lower prevalence of ID (4/8 vs. 44/47, p=0.0004) and GDD (4/8 vs. 46/47, p <0.0001). Additionally, constipation was not observed (0/8 vs. 23/47, p=0.009), and feeding difficulties in infancy were not present (0/8 vs. 20/47, p=0.02) in this group (**Figure S6**).

Inframe duplications

Three individuals with large inframe duplications were identified. Their median age at phenotypic assessment was 12.75 years (range 4.75-14.00), and one of them was male (1/3, 33%). ID was present in one of these individuals (1/3, 33%), GDD and speech delay were observed in none. Compared to those with a *EHMT1* null variant, the individuals in this group had a significantly lower prevalence of ID (1/3 vs. 44/47, p=0.0005) and GDD (0/3 vs. 47/47, p <0.0001) (**Table 2**). Septal heart defects were observed (ASD in 1/3, VSD in 1/3), and recurrent infections were common (2/3, 66%). These three individuals did not exhibit feeding difficulties in infancy, constipation, or short stature. Furthermore, they did not have epilepsy, hearing impairment, strabismus, gastroesophageal reflux disease, abnormalities of the genitourinary system, scoliosis, or endocrinological problems.

PhenoScore genotype-phenotype analysis

Next, we have compared the *EHMT1* genotypic groups using PhenoScore. For the PhenoScore analyzes, 9/125 individuals were excluded due to the inclusion of only one individual per family (n=8), and the information for one individual was not available at time of PhenoScore analyzes. The average number of HPO terms reported and used for the analysis per individual was 21 (range 4-50).

Correlation analysis included both HPO, facial, and combined datasets from the five genotypic subgroups (as in **Table 2**) using intragenic null variants as the "baseline" for the comparisons (**Table S4).** Large inframe duplications were excluded from the analyzes due to a small sample size. The number of individuals included in the comparison was equal to the number of individuals in the smallest included cohort. Since the number of individuals in each group varied, correlation analyzes were rerun with resampling for each group until all individuals from the largest group were included. In individual comparison, no statistical significant differences were found for HPO terms with PhenoScore analyses between the included genotypic cohorts. In the resampling analysis, a statistical difference was found for HPO terms only when comparing the individuals with *EHMT1* null variants to those with SET domain variants (p=0.03) and N-terminal protein truncating variants (p <0.0001) (**Table S4 and Figure S7**).

PhenoScore correlation analysis was also conducted to examine a potential difference between extragenic deletions >1 Mb and those <1 Mb. No difference in either HPO terms or facial data were observed based on the size of deletion. However, a significant difference (p=0.05) was found in facial analyses for individuals with deletions >1 Mb compared to those <1 Mb.

Additionally, a comparison based on sex was conducted. No significant differences were found between sexes for HPO terms (p=0.44) or facial features (p=0.49). However, a significant difference between sexes was identified when both HPO terms and facial features were included in the analyses (p=0.04)(**Table S4 and Figure S7**).

Inherited Kleefstra syndrome cases

During this study, we identified multiple families in whom a KLEFS1 causing pathogenic *EHMT1* variant in a KLEFS1 proband (n=10) was inherited from a non-mosaic parent (N=7) (**Figure S8**). In all families, the parent was diagnosed after diagnosis in the offspring. All affected parents presented with mild ID and/or GDD, except for one without ID/GDD. In the minority, psychiatric disorders or somatic features were also reported (details provided in the Supplemental note: Familial

KLEFS1 case reports). In all cases, the PAVs were initially classified as VUS, but in one case the variant was thought to be benign after segregating the variant in the mildly affected mother.

Seven families were identified, with four out of the seven families having a PAV in the ANKR or SET domain (p.(Phe958_Leu959del), p.(Val822Gly), p.(His944Tyr), p.(Asn1194Ile)), while only a single family presented with a *EHMT1* null variant (ch r9:g.137797535_137820858del). Two families presented with a familial inherited frameshift variant in the N-terminal domain (p.(Glu14Glyfs*23), p.(Glu14Serfs*8)).

Additionally, an eighth family was identified with two siblings affected p.(Gly37Metfs*17), but parents were unavailable for testing. In a ninth family, a proband with ID/GDD, but was also diagnosed with a pathogenic *ASXL3* (OMIM #615115) variant had an inherited EHMT1 variant from unaffected father (p.(Ala8Argfs*60). The father inherited the variant from an unaffected grandmother who was mosaic for this *EHMT1* variant.

EHMT1 enzymatically deficient protein altering variants

Three individuals from two families with ID were identified with a missense variant in the SET domain that were functionally shown to disrupt SET domain methyltransferase activity, but preserved the ability to bind to EHMT2 (**Table 1 and Figure 3**): p.Asn1200Lys and p.Phe1246Lys. Additionally, a *de novo* variant p.His1265Gln was predicted to have a similar consequence but failed to confirm the effect on the performed functional assays. These individuals presented with ID/GDD, congenital anomalies, and other clinical features (including facial dysmorphism) that only partially align with the KLEFS1 phenotypic spectrum and do not have the typical KLEFS1 phenotype (details provided in the Supplemental note: case reports of individuals with EHMT1 enzymatically-deficient protein altering variants). Currently, these variants are interpreted as VUS and their role in the patient's phenotype is unclear, as well as whether the resulting phenotype is spectrum of KLEFS1 or is a novel *EHMT1* related NDD.

Discussion

In this study, we provide novel insights into Kleefstra syndrome (KLEFS1) pathogenesis and genotype-phenotype correlations in the context of the *EHMT1* functions. We systematically analyzed the largest cohort to date of individuals with KLEFS1 and utilized comprehensive functional *in vitro* and *in silico* variant

analysis. The large cohort analysis allowed us to broaden the phenotypic and genotypic spectrum, as well as identify novel genotype-phenotype correlations. Additionally, we share our experience with KLEFS1 and *EHMT1* diagnostics over time and re-calculated the prevalence of KLEFS1 to be approximately 1:36,000 (Supplemental note). Our study highlights how large cohort analysis, powered by international collaborative effort, can provide novel insights not only into the disorder phenotypic and genotypic spectrum, but also into its pathogenesis, DNA methylation signature origins, and fundamental gene functions.

EHMT1 encodes an epigenetic regulator with multiple domain-specific functions[9]. First, the ANK repeat domain acts as a "reader" and binds to the H3K9me tail and allows the EHMT1-EHMT2 complex to bind to chromatin[10,27]. Especially the "reader function" of EHMT1 is important for efficient establishment of H3K9 methylation by its complex[49]. Second, the SET domain is known as a "writer" with methyltransferase function, but it also ensures binding to EHMT2 to form the protein complex[10,13,50]. Finally, the functions of the disordered N-terminal part and RING-like domain are currently unknown[21]. It is known that heterozygous loss of *EHMT1* results in KLEFS1[8]. However, loss of which specific EHMT1 function (or their combination) drives the development of the syndrome is unclear. Initially, the identification of two pathogenic SET missense variants p.Cys1073Tyr and p.Arg1197Trp in individuals with KLEFS1 led to the hypothesis that loss of the "writer" function drives the phenotype development[14]. However, we observed clustering of pathogenic PAVs within SET, as well as ANKR domains, suggesting more mechanisms are at play. In fact, we observed that PAVs are responsible for ~10% of all KLEFS1 diagnoses in our unbiased expertise center diagnostic cohort (Supplemental note).

The majority of the identified pathogenic ANKR PAVs were predicted to disrupt the domain's structure, thereby affecting not only its function, but also the stability of the whole protein and, therefore, possibly resulting in haploinsufficiency. Indeed, the identified recurrent p.Pro809Leu variant was proven *in vitro* to affect the domain's "reader" function, as well as stability of the protein. However, we have identified two pathogenic *de novo* ANKR missense variants p.Trp912Arg and p.Arg948Trp that were predicted and confirmed *in vitro* to affect only the binding to the H3K9 tail, while preserving other functions. The individuals with these two variants presented with typical KLEFS1 features clinically and had a "positive" Kleefstra syndrome DNAm signature. These results show that the loss of "reader" function alone is sufficient to cause KLEFS1 and its associated DNAm changes. However, on the genotype group level, we observed that the individuals with ANKR domain PAVs typically present with a milder phenotype than the individuals

with intragenic null variants: IQ was found to be higher, with a lower prevalence of recurrent infections, feeding difficulties in infancy, and structural heart defects. Moreover, three out of 20 unrelated ANKR domain's PAVs grouped individuals inherited the variant from a mildly affected parent. In contrast, only a single inherited case was reported among 81 individuals with a truncating *EHMT1* variant or deletion.

Similarly, we identified two SET domain PAV groups: 1) PAVs disrupting the whole domain's structure and all functions and 2) PAVs affecting the enzymatic activity only. The initially described pathogenic SET PAVs p.Cys1073Tyr and p.Arg1197Trp disrupted the whole domain's structure resulting in loss of its EHMT2 binding, as well as its "writer" functions[39]. On the group level, individuals with this type of pathogenic SET PAVs presented with the typical KLEFS1 phenotype, but significantly different and milder from the truncating variants, similarly as individuals with pathogenic PAVs in the ANKR domain. In contrast, individuals with the variants p.Asn1200Lys and p.Phe1246Leu (and possibly p.His1265Gln), that disrupt SET domain's methyltransferase activity while preserving protein's ability to bind EHMT2, presented with a NDD with features not typical of KLEFS1 and did not have the KLEFS1 DNAm signature. These results highlight that the loss of the "writer" activity alone is not sufficient to cause typical KLEFS1 phenotype, or DNAm changes. In fact, we observed that the two enzymatically-inactive EHMT1 proteins due to the p.Asn1200Lys or p.Phe1246Leu variants can restore the H3K9me2 levels in EHMT1 knock-out cells, which is likely driven by the EHMT2 enzymatic activity. Our results are consistent with the previous findings *in vitro* showing that only EHMT2, and not EHMT1, is indispensable for their H3K9me1-2 methyltransferase activity[12] and questions the role of EHMT1 as an epigenetic "writer". However, EHMT1 still may have an epigenetic "writer" activity, e.g., Yamada et al., Ea et al., and others have previously suggested that enzymatic activity of EHMT1 may contribute to H3K9me3 at specific loci[39,51,52], as well as product specificity of other H3 positions, e.g., H3K23m3[53]. Additionally, it has been shown that EHMT1 has a prominent role of non-histone methylation, unlike EHMT2[51,54].

We also identified a group of KLEFS1 individuals with normal intelligence or mild ID/GDD, with a milder phenotype, and without the KLEFS1 DNAm signature due to N-terminal truncating *EHMT1* variants. Their IQ scores was significantly higher (mean 75), and constipation and feeding difficulties in infancy were not present. These variants are predicted to result in a protein without an N-terminal part, due evading NMD and utilization of an alternative start codon (e.g., p.Met48 or p.Met207). We were able to confirm the presence of a shorter protein band of

similar expression level to the wild-type protein for one affected family which was absent in control. The N-terminal part of the protein is disordered and does not currently have known functions. It is known that the disordered regions in chromatin remodelers are widely used for protein-protein interactions and phase separation [55]. In fact, EHMT1 is known to interact not only with H3K9 and EHMT2, but also with NFKB1, SPOP, MPP8, WIZ and other proteins[13,50,52,56]. Additionally, the N-terminal part of EHMT1 undergoes post-translational modifications (e.g., N-acetylation or phosphorylation at positions p.Ala2 and p.Ser38, respectively). While the exact role of these modifications is unknown, it has been previously shown that other post-translational modifications of EHMT1 can have an effect on its function: e.g., Besschetnova A. et al., has recently shown that demethylation of lysine residues at positions p.Lys450 and p.Lys451 significantly increased chromatin binding capacity of EHMT1 in prostate cancer cells, which resulted change of transcriptional landscape[57]. Our findings suggest that the EHMT1 N-terminal part is important for the protein functioning, but require further investigations, while N-truncating variants likely represent hypomorphic *EHMT1* variants with milder phenotype.

In contrast, we found that individuals with multigene deletions (affecting *EHMT1* and at least one additional gene) have more severe phenotype. These findings are consistent with previous observations[14,15]. Although feeding difficulties have previously been associated with large multigene deletions[14], we did not observe a higher prevalence of feeding difficulties in this cohort. Finally, constipation was found to be strongly associated with multigene deletions, which has not been described before.

Loss of a certain EHMT1 single function can be sufficient to cause Kleefstra syndrome. We argue that the phenotype of typical Kleefstra syndrome is driven by a complex loss of multiple EHMT1 functions, including its "reader" and "writer" functions, as well as its binding function to other proteins, and, possibly, other yet unknown functions. The more severe phenotype of the multigene deletions is likely modified by a loss of flanking genes. Further functional and clinical evidence is necessary to understand the functions of the EHMT1 enzymatic activity, functions of the RING-like domain and protein N-terminal part and their contribution to the KLEFS1 phenotype. Currently, the splice variant disrupting RING-like domain c.1791G>A and the variants disrupting the SET enzymatic activity p.Asn1200Lys, p.Phe1246Leu (and possibly p.His1265Gln) remain classified as VUS. Additionally, we also showed that haploinsufficiency of *KMT2C* (OMIM #606833), which is associated with a neurodevelopmental condition currently named in OMIM as "Kleefstra syndrome 2" (OMIM #606833), results in a distinct disorder from Kleefstra syndrome[58].

In this study, we utilized the DNAm signatures to simultaneously classify VUS in 42 individuals, which would be nearly impossible to analyze using *in vitro* functional tests due to their laboriousness. While the DNAm signature testing has classified the vast majority of the variants consistently with other available evidence showing high sensitivity and specificity for the KLEFS1 DNAm signature, we also identified two variant groups (N-terminal truncating and SET domain enzymatically-inactive) with functional effect on the protein, but that were classified negatively on the signature. This is not surprising as the signature was developed based on the typical KLEFS1 individuals. Therefore, "negative" signature results could represent that a tested individual has a variant with a different molecular effect than the variants used for the signature derivation[17,59]. To conclude, the DNAm signatures represent a molecular and epigenetic "phenotype" of a disorder rather than the functional test of a variant, so should be used with caution, only in combination with the other evidence[59,60].

Further, our findings highlight that the origin of DNAm signatures (for different syndromes) are unknown and could provide unexpected results. For example, we observed that loss of the "reader" *EHMT1* activity due to an ANKR disruptive PAV is already sufficient for the presence of the KLEFS1 signature, similarly to the *EHMT1* haploinsufficiency. EHMT1 forms a heterodimer with EHMT2 which acts as a repressive complex together with other proteins, including DNMT3A, resulting in simultaneous deposition of two repressive marks – H3K9me1-2 and DNA methylation[10,13].Therefore, we hypothesize that the mostly hypomethylated KLEFS1 DNAm signature is caused by a loss of DNMT3A activity at the EHMT1 repressed loci either due to haploinsufficiency, or by disrupting binding to the EHMT2 (e.g., via SET domain disruption) or to the H3K9 methylated chromatin (e.g., via ANKR domain disruption). This would also explain why the N-terminally truncated variants do not result in a typical KLEFS1 DNAm signature. However, further functional evidence is necessary to identify the "origins" of the DNAm signature for KLEFS1 and other syndromes.

Finally, the current up-to-date cohort analysis allowed us to calculate the precise frequency of KLEFS1 clinical features (**Table 2**), identify previously poorly recognized features, as well as demonstrate a surprising variety in the clinical and neuropsychiatric spectrum of KLEFS1. We also identified multiple familial KLEFS1 cases with parentally inherited pathogenic variants. In contrast, the previously described clinical spectrum was derived from the phenotype-first cohorts, which primarily included only individuals with typical KLEFS1[14,61]. Interestingly, intellectual capacities ranged from normal to severe ID, while before they were

mostly associated with moderate to severe ID[14,61]. We also showed a variety of behavioral and psychiatric disorders existing in KLEFS1, including existence in childhood, while so far mainly described in (young) adolescence[61-64], and contrary to a previous study which has suggested none to mild behavioral symptoms in children with KLEFS1[65]. Additionally, we found that some of the features have higher prevalence than recognized before, including constipation (47%), and cerebral atrophy or hypoplasia (29%). We emphasize the significant prevalence of recurrent infections (46%) and sleep-wake disorders (46%), which corresponds with the patient-reported findings[66].

The description of multiple familial KLEFS1 occurrences highlights the heterogeneity of the KLEFS1 associated phenotype (even within one family). These findings highlight the importance of knowing the phenotypic spectrum of a disorder and the necessity of deep parental phenotyping for the correct inherited variant interpretation[40]. Otherwise, such variants can be missed and/or misinterpreted as benign during trio exome/genome sequencing.

Conclusion

In this study, we provided up-to-date clinical and molecular information about the KLEFS1 syndrome and *EHMT1*, after almost two decades since its discovery. Our study demonstrates that in-depth analysis of patient-identified variants broadens both molecular and clinical understanding, aiding in further unraveling of the underlying KLEFS1 pathogenesis.

Supplemental information

KLEFS1 diagnostic experience

To estimate different *EHMT1* pathogenic variant frequency in an unbiased way, we analyzed the results of the in-house diagnostic laboratory from the KLEFS1 expertise center in the period 2003-2023 (**Figure S5**). Here, the first diagnosis of KLEFS1 was registered in 2003. In total, 92 pathogenic variants in 98 individuals have been identified. Within this diagnostic cohort, 55% (54/98) of the individuals were female. Approximately ~50% of the KLEFS1 cases are caused by a CNV and ~50% by SNV/indel, with only one – by a structural variant (balanced translocation) affecting *EHMT1* (**Figure S5C**). As (trio) WES or other next generation sequencing-based methods are now widely implemented, individuals diagnosed with KLEFS are increasingly referred by physicians from other clinical centers.

KLEFS1 prevalence calculation

Prior to the implementation of WES diagnostics in 2013, most diagnoses of KLEFS1 were established using either combination of MLPA and Sanger sequencing and/or FISH, and/or chromosomal microarray. Since 2016, when the first diagnostic method to apply for a patient with NDD in our center has become (trio) exome sequencing with CNV calling allowing to identify most of the KLEFS1 causes and shifting to the genotype-first approach, 15 KLEFS1 individuals have been identified among 8038 individuals referred to exome sequencing due to a NDD (only data from 2016-2022 were available and used). Therefore, 0.19% 95%CI (0.10-0.31%) of all NDD cases are attributable to KLEFS1. Based on the Dutch population prevalence of ID of 1.45% [67], the calculated prevalence of KLEFS1 in the general population is 0.0028% 95%CI (0.0015-0.0045%), which is roughly 1 in 36,000 95%CI (22,000-67,000) individuals. However, since 2016, eight cases were still diagnosed using the targeted approach (Array, MLPA and Sanger sequencing) due to a clinical diagnosis in a case with specific KLEFS1 phenotype, so the prevalence could be even higher.

Description of KLEFS1 clinical features

Neurodevelopmental and neurological features

Neurodevelopmental disorders were identified in nearly all individuals, with the most prevalent conditions being intellectual disability (112/125, 90%) and global developmental delay (114/125, 91%). When selected for cases age of 5 years and above, autism spectrum disorder was found in 72% (71/98), and attention-deficit/hyperactivity disorder in 13% (13/98).

The total IQ score was known for 30 individuals, with a median IQ score of 61 (IQR 55-68, range 42-81), indicating severe ID to normal IQ. The degree of ID was classified as mild, moderate, severe, or profound, which was determined for 78 individuals. No ID was observed in 11 individuals with KLFS. Mild ID was identified in 45 individuals (approximately IQ range from 50-69), moderate ID was present in 19 individuals (IQ 36-49), severe ID was found in 14 individuals (IQ 20-35). No difference in IQ was identified between males and females (66 vs. 60, p=0.28). The absence of speech was reported in 10 individuals, and was found to be equally distributed among the various genotypic groups (p=0.76) and between genders (p=0.61).

Epilepsy was observed in 27% of individuals (34/125), with no differences across genotypic groups and genders. Furthermore, within the group of individuals with intragenic deletions, the prevalence of epilepsy was comparable between deletions including gene *CACNA1B* and those not including this gene (p=0.87). Epilepsy type was known for 15 of these individuals: bilateral tonic-clonic seizures were the most common subtype (11/15); absences occurred in 5 individuals (5/15) and they often co-occur with tonic-clonic seizures (3/15); other observed epilepsy subtypes were atonic seizures (2/15), focal seizures (3/15), and nocturnal seizures (1/15). In six additional cases, febrile seizures occurred during infancy or childhood, yet epilepsy is not present.

Cerebral imaging was conducted in 45 individuals, revealing cerebral anomalies in 22/45 individuals. These anomalies included cerebral atrophy in 5/22 individuals (with frontal lobe atrophy noted in three individuals), cerebral hypoplasia in 8/22 individuals (involving the corpus callosum in six, brainstem in two, pons in one, and medulla in one individual). Nonspecific white matter abnormalities were detected in 5/22 individuals. Cerebral cysts were observed in 3/22 individuals (including pituitary cysts in one, cysts in the germinal layer in one, and frontal lobe cysts in one individual).

Behavioral and psychiatric disorders in individuals aged 5 years and above were categorized based on DSM-V classifications (Diagnostic and Statistical Manual of Mental disorders) (**Table 3**). Most of the individuals were diagnosed with one or more disorder (89/98, 91%) with the majority having more than one disorder (69/98, 70%). With age, the prevalence of psychiatric conditions, specifically depression and psychotic disorders, slightly increases. Beyond the age of 15 years, the respective prevalence rates are 93% (52/56), 20% (11/56), and 25% (14/56).

Above the age of 5 years, observed behavioral and psychiatric disorders included anxiety (27/98, 28%), aggressive behavior (27/98, 28%), schizophrenia spectrum and other psychotic disorders (16/98, 16%), bipolar disorder (7/98, 7%), depressive disorders (14/98, 14%), and obsessive-compulsive and related disorders (9/98, 9%). Less common symptoms in KLEFS1 include agitation (12/98, 12%), apathy (18/98, 18%), and self-harm (11/98, 11%).

Cognitive and/or developmental regression was reported in 36% of individuals (35/98). Interestingly, regression was reported predominantly in females (27/35). The majority of participants received antipsychotic treatment during the regression period (28/35, 80%).

Sensory system

Both vision and hearing impairment were frequently present in this cohort. Abnormalities of refraction were reported in nearly half of the individuals (55/122, 45%). The most prevalent refractive disorder was hypermetropia (40/122, 33%), followed by astigmatism (16/122, 13%), and myopia (8/122, 7%). Cerebral vision disorders were reported in three individuals (3/122, 2%). Hearing impairment was observed in 36/122 individuals (30%), and occurred predominantly bilateral (33/36, 92%). Mixed, sensorineural and conductive hearing impairment were reported.

Congenital anomalies

Structural heart defects were observed in 31% of the individuals (39/125), with cardiac septum abnormalities being the most common (24/125, 19%), particularly atrial septal defects (13/125), followed by ventricular septal defects (7/125), and atrioventricular septal defects (4/125). Cardiac rhythm disturbance was reported in 12% (15/125).

Kidney and urogenital abnormalities were observed in 19% (24/125). Among these, the most common were renal insufficiency (2/125), hydronephrosis (4/125), ureteropelvic junction stenosis (2/125), renal cysts (2/125), cryptorchidism (3/125), and inguinal hernia (5/125).

A variety of neuromuscular and skeletal conditions were observed among all genotypic groups (94/125). (Neonatal) hypotonia was most frequently observed (52/125, 42%). Skeletal conditions that were observed multiple times included pectus excavatum, joint laxity, hip dysplasia, pes planus, pes equinovarus, clinodactyly, and camptodactyly.

Growth

A short stature was present in 10 individuals (9%, 10/110), while tall stature was found in 12% of the individuals (13/110). Individuals with a short stature were older at the time of examination than those with a tall stature (19.2 vs. 14.1 years). There was no significant difference in height between the genders (p=0.32).

Overweight and obesity were frequently present, with a prevalence of 48% (52/125). Overall, the BMI percentile interquartile range was within the 57th and 94th percentiles. Interestingly, females showed a significantly higher BMI percentile compared to males (76th vs. 65th percentile, p=0.04).

Immunology

Recurrent infections were identified in 46% (58/125) of individuals, often mentioned at young age. These infections were specified in 21 individuals, with recurrent otitis being the most prevalent (8/21), followed by recurrent respiratory tract infections (6/21), and recurrent urinary tract infections (6/21). Lower respiratory tract infections were observed in 7/21 individuals, with four presenting with recurrent pneumonia and three individuals with bronchitis. Additionally, osteomyelitis was reported in one individual, and acute infectious thyroiditis (described in more details by Bouman A. et al.[15]). Apart from one case with inflammatory bowel disease and one autoimmune thyroiditis, no other autoimmune diseases were documented.

Supplemental note: Familial KLEFS1 case reports.

1. In the first family (**Figure S8A**), a 13-year old male was diagnosed with a maternally inherited missense variant in the ANKR domain (c.2465T>G p.(Val822Gly)). The proband is mothers' only child. His father is not tested. Both affected individuals have mild ID, GDD, and major depressive disorder. The mother's brother shows features suggestive of KLEFS1, but he has not been tested.
2. In the second family (**Figure S8B**), a paternally inherited missense variant in the ANKR domain was identified in a 15-years old female (c.2830C>T p.(His944Tyr)). Her sibling is tested, but is not affected. The variant occurred *de novo* in the father. Both affected individuals have mild ID, GDD, and ASD. The proband experienced a developmental regression period during adolescence, which improved on antipsychotics.
3. In the third family (**Figure S8C**), three probands were diagnosed with a maternally inherited inframe deletion in the ANKR domain (c.2873_2878del p.(Phe958_Leu959del)), so the variant was interpreted as likely benign/ VUS. Methylation analyzes were performed and suggestive for KLEFS1 in all

affected individuals. The parents of the probands are cousins. Father is not available. Mother never went to school, but could read, write, and manage her own household. The probands have had developmental regression over time, with severe ID at time of examination. Two healthy sibs have undergone genetic testing, but they are unaffected. One sib died shortly after birth.

4. In the fourth family (**Figure S8D**), a maternally inherited missense variant in the SET domain was identified in a 3-years old female (c.3581A>T p.(Asn1194Ile)). Mother had learning difficulties and speech delay. The proband has moderate ID and mild GDD. The grandparents were not tested.
5. In the fifth family (**Figure S8E**), a maternally inherited intragenic deletion was identified resulting in a frameshift (deletion exons 17-25; GRCh38(chr9): g.137797535_137820858del). This family includes an affected mother and her two affected sons. The variant occurred *de novo* in the mother. Mother has mild ID and GDD, she went to a mainstream school without additional help. Younger son has GDD and mild ID; no info is available for the second son. She lives independently with her partner and 4 children. Two other siblings of the proband are unaffected.

In four families, an inherited frameshift variant was identified in the N-terminal part:

6. In the sixth family (**Figure S8F**), in a 22-year old female a maternally inherited N-truncating frameshift variant was identified (c.40dup p.(Glu14Glyfs*23)). The inheritance of the variant in the mother is unknown. Mother has mild ID and has several somatic and psychiatric issues. The proband has mild ID and GDD.
7. In the seventh family (**Figure S8G**), another maternally inherited N-truncating frameshift variant (c.40del p.(Glu14Serfs*8)) was identified in a 19-years old female. She has no ID or GDD, but has generalized seizures, hypermetropia and anxiety. Her mother is not affected. The grandparents were not tested.
8. In the eight family (**Figure S8H**), a proband was diagnosed with a familial, paternally inherited N-terminal frameshift variant due to deletion of the 2nd exon of *EHMT1* (p.(Ala8fs*). Her sib, father and grandmother all carried the variant and are unaffected and did not display prominent dysmorphic features. Grandmother was mosaic for the *EHMT1* variant. The proband showed ID/GDD but was also diagnosed with a pathogenic ASXL3 variant.
9. Two siblings were diagnosed with an intragenic deletion in the N-term terminal domain which results in a frameshift (c.109delinsAT p.Gly37Metfs*17). Both siblings were adopted out of the family. They presented with mild ID, GDD and KLEFS1 dysmorphic features. The biological mother was suspected

having KLEFS1 clinically. While full familial recurrence of KLEFS1 is expected, mosaicism cannot be ruled out in the mother, as she was unavailable for testing (**Figure S8I**).

Case reports of individuals with *EHMT1* enzymatically-deficient protein-altering variants

A. NM_024757.5:c.3600C>G p.Asn1200Lys *EHMT1* variant (inheritance unknown) was identified in a 27-year-old male who exhibits severe ID and motor developmental delay. He uses single words and short sentences. Additionally, he has constipation, joint hyperlaxity, epilepsy (both generalized and focal seizures), autism spectrum disorder, and aggressive behavior. Facial features are overlapping, but not specific for consistent with the KLEFS1. An MRI conducted in the first year of life revealed delayed myelination, a wide sylvian fissure, and a retrocerebellar arachnoid cyst.

B. NM_024757.5:c.3736T>C p.Phe1246Leu *EHMT1* variant (inheritance unknown) was identified in a 6-years-old girl with additional pathogenic 16p11.2 deletion (GRCh38 chr16:28,832,615-29,043,898). She was born with a cesarean section and was small for gestational age (-2.69 SD). At age 5, she is overweight (p94), shows motor and speech developmental delay and some facial features (round, flat face with mild midface retrusion, small upturned nose, thin lips) consistent with a diagnosis of KLEFS1. She had an atrial septal defect that was closed surgically. The MRI did not reveal any brain anomalies, and there are no indications of epilepsy or mental health disorders, aside from frequent temper tantrums. She also experiences sleep onset and sleep maintenance insomnia. She has an older sister with the same *EHMT1* variant who has mild ID and seizures but without KLEFS1 facial features. However, she does not carry the 16p11.2 deletion. Both of the sisters are adopted out.

C. NM_024757.5:c.3795C>G p.His1265Gln *de novo* variant was identified in a 24-years-old male. His parents are consanguineous. This individual has intellectual disability, motor developmental delay, and severe speech/language delay and dysmorphic features (microcephaly, brachycephaly, deep set eyes, arched eyebrows, short philtrum) consistent with KLEFS1. He also currently has obesity (BMI=30.84 kg/m2), microcephaly (-2.18 SD), retinitis pigmentosa, and pes planus. There is no evidence of epilepsy or behavioral abnormalities. A cerebral MRI conducted in infancy showed wide perivascular spaces with no other abnormalities.

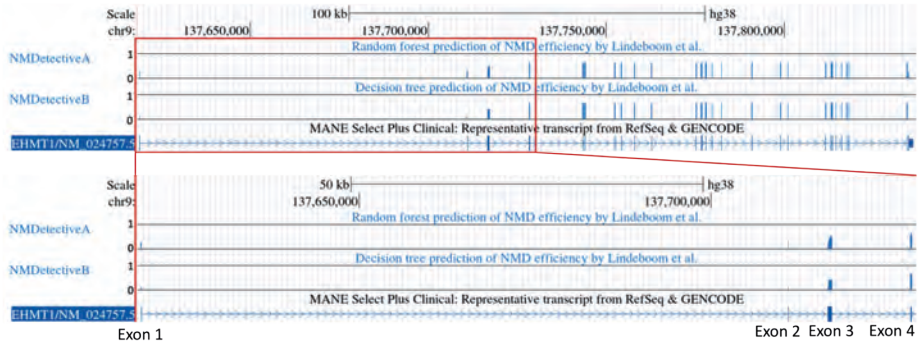


Figure S1. Predictions of escape from the NMD for *EHMT1* by NMDetective.

NMDetective (ranging from 0 to 1) showing decreased predictions for NMD (<0.5) for the first two and in part third EHMT1 (NM_024757.5) exons.

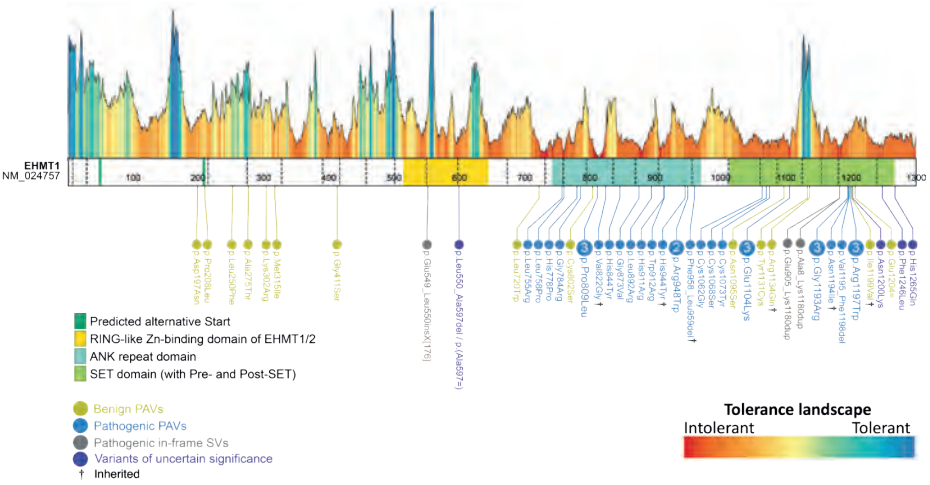


Figure S2. *EHMT1* missense variant tolerance landscape with all evaluated PAVs in this study.

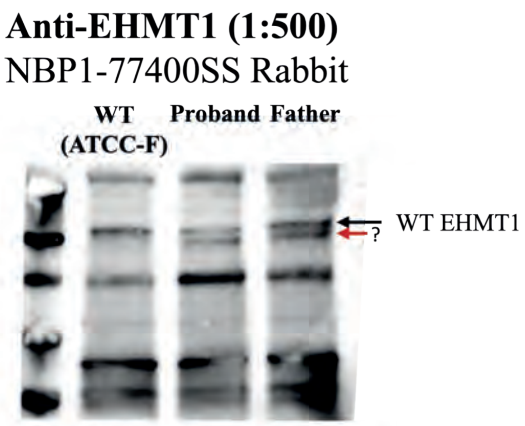


Figure S3. Protein immunoblot for the deletion of the exon 2 c.21+42335_86-2217del p.(Ala8Argfs*60).
*The immunoblot showing a shorter band in a proband and her father with the deletion of the exon 2 c.21+42335_86-2217del p.(Ala8Argfs*60) variant which is absent in control, confirming the escape from NMD and production of a shorter protein isoform.*

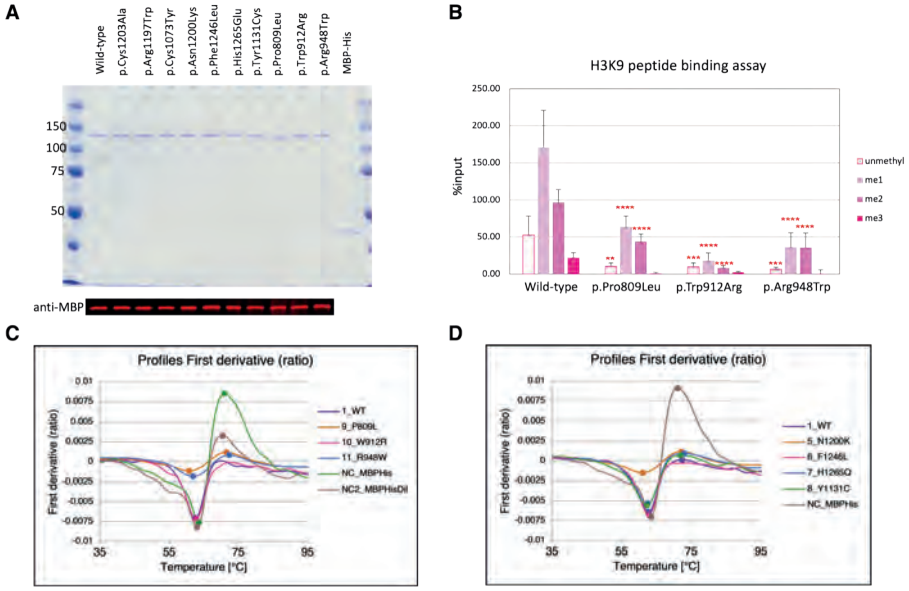


Figure S4. *In vitro* variant testing supplemental results.
A. *Shows expression of purified rMBP-hEHMT1(635-1298)-His (mutant) proteins.* ***B.*** *Shows binding ability to the H3K9 peptide (nonmethylated and methylated) for the ANKR domain mutant proteins.* ***C.*** *and* ***D.*** *shows mutant protein thermostability for the ANKR and SET domain variants, respectively.WT = Wild-type*

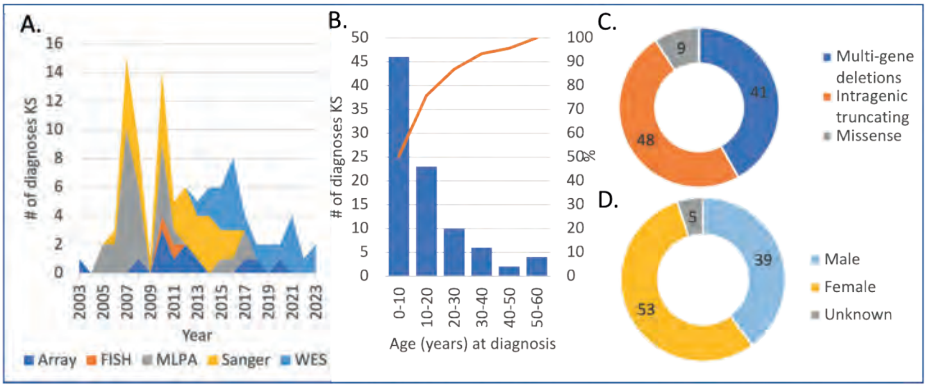


Figure S5. Diagnostics of Kleefstra syndrome at Radboudumc KLEFS1 expertise center: 2003-2023.

A. *Shows the diagnostic techniques used that identified a pathogenic EHMT1 variant.* ***B.*** *Shows age at diagnosis (in years) with cumulative % of the individuals (orange line).* ***C.*** *Shows the variant type distribution by size.* ***D.*** *Shows the sex distribution of the identified individuals.*

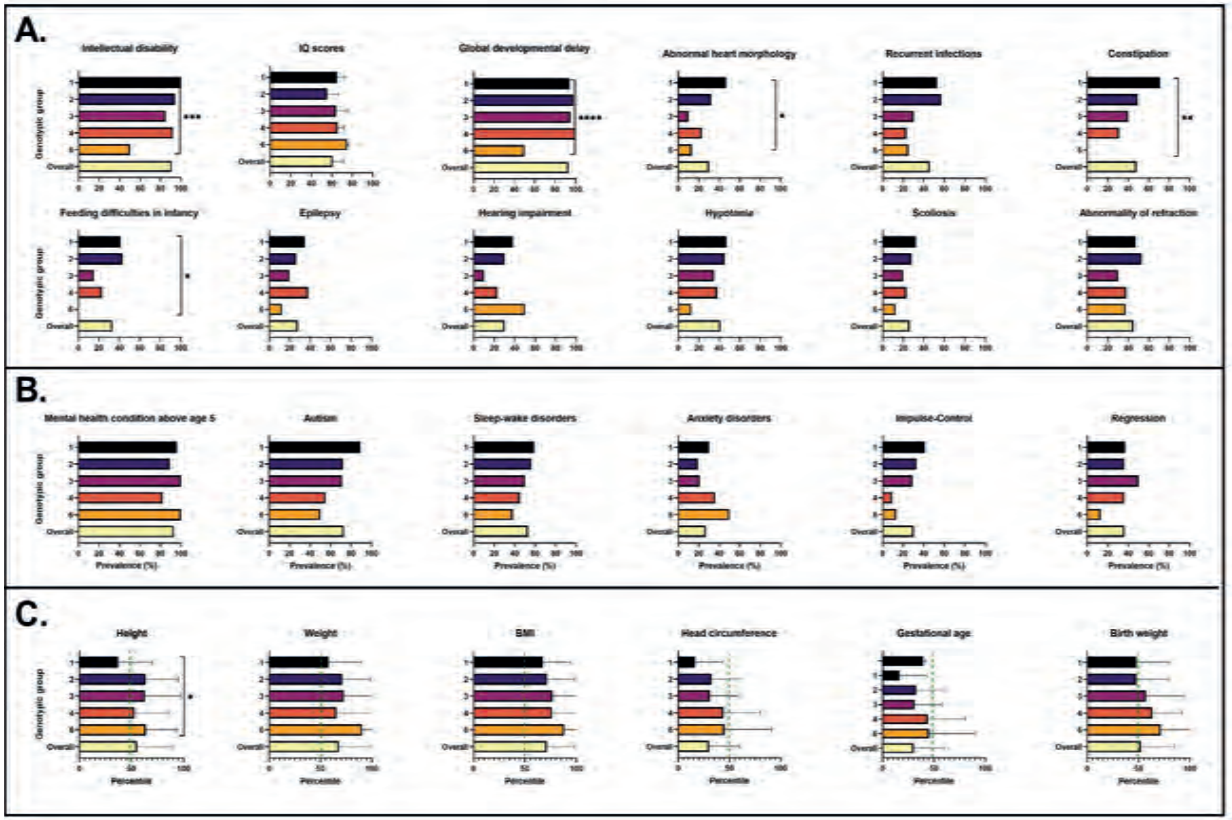


Figure S6. Kleefstra syndrome genotype-phenotype correlations.

Genotypic and phenotypic information was available for 125 individuals. ***A.*** *For clinical symptoms, ANOVA showed a significant difference among genotypic groups for intellectual disability, global developmental delay, abnormal heart morphology, constipation, feeding difficulties in infancy, and height. The first three conditions remain significant after Bonferroni correction.* ***B.*** *Mental disorders are common among individuals from all genotypic groups (91%) and were not significantly different.* ***C.*** *For growth, individuals with multigene CNVs showed a significantly lower height compared to individuals with EHMT1 null variants. Overall, overweight/obesity was commonly observed (48%). Groups are coded as 1 = multigene copy number variants that include EHMT1 and at least one additional protein coding gene; 2 = isolated EHMT1 null variants (nonsense, frameshift, splice-site variants, as well as intragenic deletions and deletions that do not affect other protein-coding genes); 3 = Protein altering variants (PAVs) in ANKR domain; 4 = PAVs in SET domain; 5 = N-terminal truncating variants.*

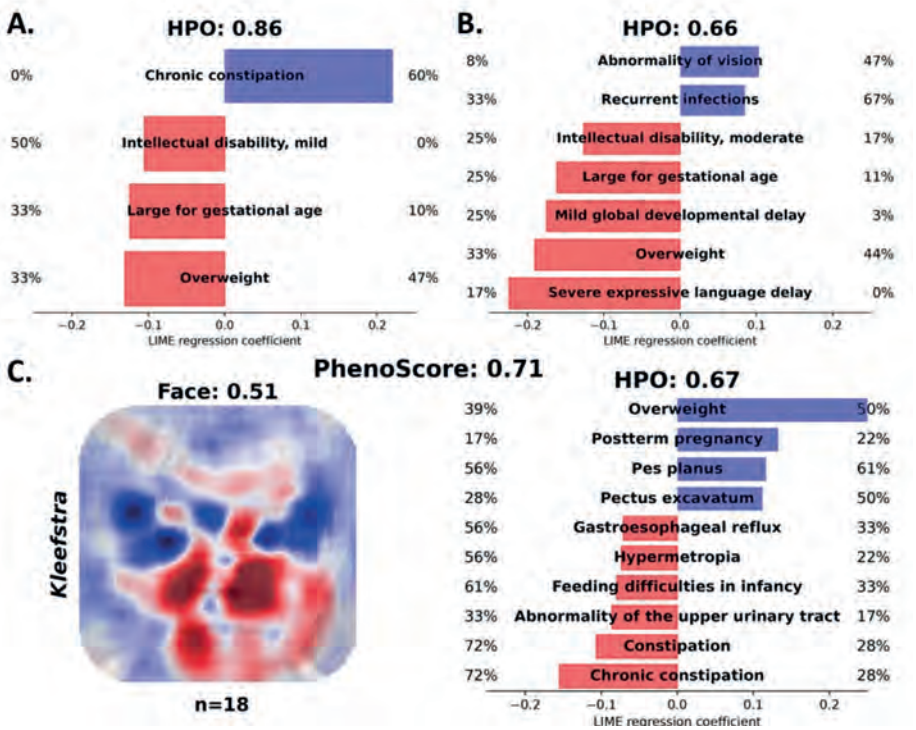


Figure S7. Kleefstra syndrome significant genotype-phenotype and sex differences calculated using PhenoScore.

LIME regression analysis shows significant differences in clinical features based on HPO data by comparing ***A.*** *the cohort of EHMT1 intragenic null variants (purple) vs. SET domain PAVs (red);* ***B.*** *EHMT1 null variants (purple) vs. N-terminal protein truncating variants (red), as well significant differences between (****C****) females vs. males on combined HPO and facial analysis.*

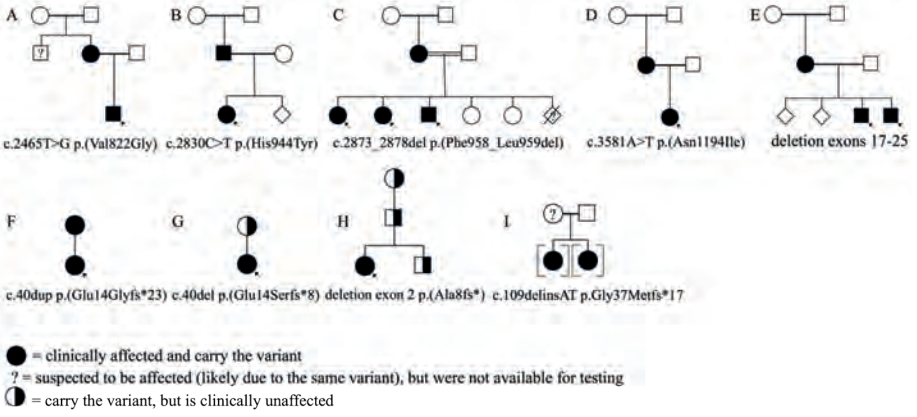


Figure S8. Pedigrees of familial KLEFS1 cases.

Tables S1, S2, and S4 supporting the findings of this study are available online in the Supplementary material of this article at: DOI: 10.1016/j.ajhg.2024.06.008.

Table S3. *EHMT1* variant of uncertain significance reclassification evidence and results.

Variant Class	Variant NM_024757.5	Occurence in the cohort	Occurence in gnomAD v2.1.1	Inheritance	*In silico* tool predictions (REVEL/SpliceAI)
PAVs	c.589G>A p.(Asp197Asn)	1	16	DN	0.11
	c.623C>T p.(Pro208Leu)	1	3	-	0.21
	c.750A>C p.(Leu250Phe)	1	0	DN	0.13
	c.823G>A p.(Ala275Thr)	1	0	DN	0.07
	c.905A>G p.(Lys302Arg)	1	52	-	0.17
	c.945G>A p.(Met315Ile)	1	0	DN	0.15
	c.1231G>A p.(Gly411Ser)	1	1	-	0.23
	c.2159T>G p.(Leu720Trp)	1	0	-	0.52
	c.2264T>G p.(Leu755Arg)	1	0	DN	0.46
	c.2273T>C p.(Leu758Pro)	1	0	M	0.94
	c.2333A>C p.(His778Pro)	1	0	DN	0.86
	c.2350G>A p.(Gly784Arg)	1	0	DN	0.86
	c.2405G>C p.(Cys802Ser)	1	6	-	0.46
	c.2426C>T p.(Pro809Leu)	3	0	DN	0.77
	c.2465T>G p.(Val822Gly)	1	0	IA	0.85
	c.2530C>T p.(His844Tyr)	1	0	DN	0.77
	c.2618G>T p.(Gly873Val)	1	0	DN	0.85
	c.2675T>G p.(Leu892Arg)	1	0	DN	0.93
	c.2732A>G p.(His911Arg)	1	0	DN	0.93
	c.2734T>C p.(Trp912Arg)	1	0	-	0.77
	c.2830C>T p.(His944Tyr)	1	0	IA	0.78
	c.2842C>T p.(Arg948Trp)	3	0	DN	0.49
	c.2873_2878del p.(Phe958_Leu959del)	1	0	IA	-
	c.3184T>G p.(Cys1062Gly)	1	0	DN	0.98
	c.3203G>C p.(Cys1068Ser)	1	0	DN	0.87
	c.3218G>A p.(Cys1073Tyr)	1	0	DN	0.93
	c.3284A>G p.(Asn1095Ser)	1	0	-	0.28
	c.3310G>A p.(Glu1104Lys)	3	0	DN	0.94
	c.3392A>G p.(Tyr1131Cys)	1	16	-	0.7
	c.3401G>A p.(Arg1134Gln)	1	6	IU	0.12
	c.3577G>A p.(Gly1193Arg)	1	0	DN	0.97
	c.3577G>C p.(Gly1193Arg)	2	0	DN	0.97
	c.3581A>T p.(Asn1194Ile)	1	0	IA	0.92

	***In silico* tool predictions (AlphaMissense)**	**Domain location**	**Predicted 3D protein effect**	**EpiSign results**	***In vitro* functional testing**	**KLEFS1 phenotype**	**Final classification**
	0.09	N	N	N	-	N	B
	0.48	N	N	-	-	-	B
	0.08	N	N	N	-	-	B
	0.19	N	N	-	-	-	B
	0.18	N	N	N	-	-	B
	0.34	N	N	N	-	N	B
	0.12	N	N	N	-	N	B
	0.97	N	N	N	-	N	B
	0.99	A	N	KS	-	Y	P
	1	A	D	KS	-	Y	P
	0.99	A	D	KS	-	Y	P
	0.99	A	D	KS	-	-	P
	0.51	A	N	N	-	-	B
	0.99	A	D	KS	D	Y	P
	0.99	A	D	KS	-	Y	P
	0.99	A	D	KS	-	-	P
	0.99	A	D	-	-	Y	P
	0.99	A	D	KS	-	Y	P
	1	A	D	KS	-	Y	P
	1	A	D	KS	D	Y	P
	0.99	A	D	KS	-	Y	P
	0.99	A	U	KS	D	Y	P
	-	A	D	KS	-	Y	P
	0.99	S	D	KS	-	Y	P
	0.99	S	D	KS	-	Y	P
	0.99	S	D	KS	D	Y	P
	0.09	S	N	-	-	N	B
	0.99	S	D	KS	-	Y	P
	0.4	S	N	N	N	N	B
	0.06	S	N	N	-	N	B
	0.99	S	D	-	-	Y	P
	0.99	S	D	KS	-	Y	P
	0.99	S	D	KS	-	Y	P

Table S3. Continued

Variant Class	Variant NM_024757.5	Occurence in the cohort	Occurence in gnomAD v2.1.1	Inheritance	*In silico* tool predictions (REVEL/SpliceAI)
	c.3583_3594del p.(Val1195_Phe1198del)	1	0	M	-
	c.3589C>T p.(Arg1197Trp)	3	0	DN	0.92
	c.3595A>G p.(Ile1199Val)	1	0	IU	0.63
	c.3600C>G p.(Asn1200Lys)	1	0	-	0.91
	c.3736T>C p.(Phe1246Leu)	1	0	-	0.89
	c.3795C>G p.(His1265Gln)	1	0	DN	0.34
Silent	c.1791G>A p.(Ala597=) r.spl	1	0	DN	0.17
	c.3459C>T p.(Cys1153=) r.spl	2	0	DN	0.95
	c.3612G>A p.(Glu1204=)	1	0	DN	0
N-terminal truncating	c.21+42335_86-2217del p.(Ala8Argfs*60)	1	0	M	-
	c.34_35insC p.(Arg12Thrfs*25)	1	0	-	-
	c.40del p.(Glu14Serfs*8)	1	0	IU	-
	c.40dup p.(Glu14Glyfs*23)	1	1	IA	-
	c.38_39insA p.(Glu14Glyfs*23)	1	0	IA	-
	c.109delGinsAT p.(Gly37Metfs*17)	1	0	-	-
	c.244del p.(Gln82Argfs*7)	1	0	DN	-
Inframe dups	Exon 2-25 dup p.(Ala8_Lys1180dup)	1	0	DN	-
	Exon 4-6 dup-ins p.(Glu549_Leu550insX[176])	1	0	DN	-
	Exon 19-25 dup p.(Glu905_Lys1180dup)	1	0	-	-

DN = *de novo*; IA = inherited from affected parent; IU =inherited from unaffected; M = inherited from a mosaic parent; - = unknown

REVEL > 0.644 or < 0.29 for missense; SpliceAI > 0.2 or <0.1 for silent variants

Blue = evidence for pathogenicity; Green = benign evidence; Grey – unclear significance

	In silico tool predictions (AlphaMissense)	Domain location	Predicted 3D protein effect	EpiSign results	*In vitro* functional testing	KLEFS1 phenotype	Final classification
	-	S	D	-	-	Y	P
	0.97	S	D	KS	D	Y	P
	0.18	S	U	N	-	N	B
	0.99	S	D	N	D	U	VUS
	0.99	S	D	N	D	U	VUS
	0.8	S	U	N	N	U	VUS
	-	R	U	N	U	N	VUS
	-	-	-	KS	-	Y	P
	-	-	-	N	-	U	B
	-	N	-	N	-	N	P
	-	N	-	U	-	U	P
	-	N	-	N	-	N	P
	-	N	-	N	-	Y	P
	-	N	-	N	-	Y	P
	-	N	-	-	-	Y	P
	-	N	-	-	-	Y	P
	-	S	-	KS	-	Y	P
	-	R	U	U	-	U	P
	-	A	-	KS	-	Y	P

AlphaMissense >0.564 or <0.34

N = no domain, disordered; A = ANKR; S = SET; R = Ring-like

N = no predicted effect; D = damaging; U = unclear

KS = KLEFS1 episignature positive; N = negative

N = no effect; D = damaging; U = unclear

Y = yes; N = no; U = unclear

P = pathogenic; B = benign; VUS = uncertain significance

5

Acknowledgements

This work was supported by Aspasia grant of the Dutch Research Council (015.014.036 to TK), the Netherlands Organization for Health Research and Development grants (10250022110003 to TK and AB and 91718310 to TK).

These results contribute to the overall goals of the Solve-RD project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 779257 (LV; TK; HB; SB). Several 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).

SB acknowledges the support of the NIHR Manchester Biomedical Research Centre (NIHR203308) and the MRC Epigenomics of Rare Diseases Node (MR/Y008170/1). We thank the GeneDx, Deciphering Developmental Delay (DDD) for providing the individuals, samples, and/or molecular diagnostic data. We thank Michael Kwint for the technical assistance. We thank Wendy Hocking for reaching to the KLEFS1 community.

Conflicts of interest

AMM is an employee of GeneDx, LLC. B.S. is a shareholder in EpiSign Inc., a biotechnology company involved in commercialization of EpiSign™ technology. The rest of the authors declare no competing interests.

Data Availability

Recruited individuals' identified *EHMT1* variants and clinical details are provided in the **Table S2**. The variants and their interpretations were submitted to the ClinVar database (ClinVar accession numbers: SCV005045804-SCV005045896, SCV005049504, and SCV005049572-SCV005049646).

References

1. Jansen, S., Vissers, L., and de Vries, B.B.A. (2023). The Genetics of Intellectual Disability. Brain Sci *13*. 10.3390/brainsci13020231.
2. Wilczewski, C.M., Obasohan, J., Paschall, J.E., Zhang, S., Singh, S., Maxwell, G.L., Similuk, M., Wolfsberg, T.G., Turner, C., Biesecker, L.G., and Katz, A.E. (2023). Genotype first: Clinical genomics research through a reverse phenotyping approach. American journal of human genetics *110*, 3-12. 10.1016/j.ajhg.2022.12.004.
3. Rots, D., Chater-Diehl, E., Dingemans, A.J.M., Goodman, S.J., Siu, M.T., Cytrynbaum, C., Choufani, S., Hoang, N., Walker, S., Awamleh, Z., et al. (2021). Truncating SRCAP variants outside the Floating-Harbor syndrome locus cause a distinct neurodevelopmental disorder with a specific DNA methylation signature. American journal of human genetics *108*, 1053-1068. 10.1016/j.ajhg.2021.04.008.
4. Cuvertino, S., Hartill, V., Colyer, A., Garner, T., Nair, N., Al-Gazali, L., Canham, N., Faundes, V., Flinter, F., Hertecant, J., et al. (2020). A restricted spectrum of missense KMT2D variants cause a multiple malformations disorder distinct from Kabuki syndrome. Genet Med *22*, 867-877. 10.1038/s41436-019-0743-3.
5. Menke, L.A., van Belzen, M.J., Alders, M., Cristofoli, F., Ehmke, N., Fergelot, P., Foster, A., Gerkes, E.H., Hoffer, M.J., Horn, D., et al. (2016). CREBBP mutations in individuals without Rubinstein-Taybi syndrome phenotype. Am J Med Genet A *170*, 2681-2693. 10.1002/ajmg.a.37800.
6. Vissers, L., van Nimwegen, K.J.M., Schieving, J.H., Kamsteeg, E.J., Kleefstra, T., Yntema, H.G., Pfundt, R., van der Wilt, G.J., Krabbenborg, L., Brunner, H.G., et al. (2017). A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. Genetics in medicine : official journal of the American College of Medical Genetics *19*, 1055-1063. 10.1038/gim.2017.1.
7. Snijders Blok, L., Verseput, J., Rots, D., Venselaar, H., Innes, A.M., Stumpel, C., Õunap, K., Reinson, K., Seaby, E.G., McKee, S., et al. (2023). A clustering of heterozygous missense variants in the crucial chromatin modifier WDR5 defines a new neurodevelopmental disorder. HGG Adv *4*, 100157. 10.1016/j.xhgg.2022.100157.
8. Kleefstra, T., Brunner, H.G., Amiel, J., Oudakker, A.R., Nillesen, W.M., Magee, A., Geneviève, D., Cormier-Daire, V., van Esch, H., Fryns, J.P., et al. (2006). Loss-of-function mutations in euchromatin histone methyl transferase 1 (EHMT1) cause the 9q34 subtelomeric deletion syndrome. American journal of human genetics *79*, 370-377. 10.1086/505693.
9. Fahrner, J.A., and Bjornsson, H.T. (2019). Mendelian disorders of the epigenetic machinery: postnatal malleability and therapeutic prospects. Hum Mol Genet *28*, R254-r264. 10.1093/hmg/ddz174.
10. Sanchez, N.A., Kallweit, L.M., Trnka, M.J., Clemmer, C.L., and Al-Sady, B. (2021). Heterodimerization of H3K9 histone methyltransferases G9a and GLP activates methyl reading and writing capabilities. The Journal of biological chemistry *297*, 101276. 10.1016/j.jbc.2021.101276.
11. Shirai, A., Kawaguchi, T., Shimojo, H., Muramatsu, D., Ishida-Yonetani, M., Nishimura, Y., Kimura, H., Nakayama, J.I., and Shinkai, Y. (2017). Impact of nucleic acid and methylated H3K9 binding activities of Suv39h1 on its heterochromatin assembly. Elife *6*. 10.7554/eLife.25317.
12. Tachibana, M., Matsumura, Y., Fukuda, M., Kimura, H., and Shinkai, Y. (2008). G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription. Embo j *27*, 2681-2690. 10.1038/emboj.2008.192.
13. Chang, Y., Sun, L., Kokura, K., Horton, J.R., Fukuda, M., Espejo, A., Izumi, V., Koomen, J.M., Bedford, M.T., Zhang, X., et al. (2011). MPP8 mediates the interactions between DNA methyltransferase Dnmt3a and H3K9 methyltransferase GLP/G9a. Nat Commun *2*, 533. 10.1038/ncomms1549.

14. Willemsen, M.H., Vulto-van Silfhout, A.T., Nillesen, W.M., Wissink-Lindhout, W.M., van Bokhoven, H., Philip, N., Berry-Kravis, E.M., Kini, U., van Ravenswaaij-Arts, C.M., Delle Chiaie, B., et al. (2012). Update on Kleefstra Syndrome. Mol Syndromol *2*, 202-212. 10.1159/000335648.

15. Bouman, A., Geelen, J.M., Kummeling, J., Schenck, A., van der Zwan, Y.G., Klein, W.M., and Kleefstra, T. (2023). Growth, body composition, and endocrine-metabolic profiles of individuals with Kleefstra syndrome provide directions for clinical management and translational studies. Am J Med Genet A. 10.1002/ajmg.a.63472.

16. Goodman, S.J., Cytrynbaum, C., Chung, B.H.-Y., Chater-Diehl, E., Aziz, C., Turinsky, A.L., Kellam, B., Keller, M., Ko, J.M., Caluseriu, O., et al. (2020). EHMT1 pathogenic variants and 9q34.3 microdeletions share altered DNA methylation patterns in patients with Kleefstra syndrome. Journal of Translational Genetics and Genomics *4*, 144-158. 10.20517/jtgg.2020.23.

17. Aref-Eshghi, E., Kerkhof, J., Pedro, V.P., Barat-Houari, M., Ruiz-Pallares, N., Andrau, J.C., Lacombe, D., Van-Gils, J., Fergelot, P., Dubourg, C., et al. (2020). Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in 42 Mendelian Neurodevelopmental Disorders. American journal of human genetics *106*, 356-370. 10.1016/j.ajhg.2020.01.019.

18. Demond, H., Hanna, C.W., Castillo-Fernandez, J., Santos, F., Papachristou, E.K., Segonds-Pichon, A., Kishore, K., Andrews, S., D'Santos, C.S., and Kelsey, G. (2023). Multi-omics analyses demonstrate a critical role for EHMT1 methyltransferase in transcriptional repression during oogenesis. Genome Res *33*, 18-31. 10.1101/gr.277046.122.

19. Pang, K.K.L., Sharma, M., and Sajikumar, S. (2019). Epigenetics and memory: Emerging role of histone lysine methyltransferase G9a/GLP complex as bidirectional regulator of synaptic plasticity. Neurobiol Learn Mem *159*, 1-5. 10.1016/j.nlm.2019.01.013.

20. Nachiyappan, A., Gupta, N., and Taneja, R. (2022). EHMT1/EHMT2 in EMT, cancer stemness and drug resistance: emerging evidence and mechanisms. Febs j *289*, 1329-1351. 10.1111/febs.16334.

21. Kerchner, K.M., Mou, T.C., Sun, Y., Rusnac, D.V., Sprang, S.R., and Briknarová, K. (2021). The structure of the cysteine-rich region from human histone-lysine N-methyltransferase EHMT2 (G9a). J Struct Biol X *5*, 100050. 10.1016/j.yjsbx.2021.100050.

22. Rots, D., Rooney, K., Relator, R., Kerkhof, J., McConkey, H., Pfundt, R., Marcelis, C., Willemsen, M.H., van Hagen, J.M., Zwijnenburg, P., et al. (2024). Refining the 9q34.3 microduplication syndrome reveals mild neurodevelopmental features associated with a distinct global DNA methylation profile. Clinical genetics. 10.1111/cge.14498.

23. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. American journal of human genetics *84*, 524-533. 10.1016/j.ajhg.2009.03.010.

24. Prevalence and architecture of de novo mutations in developmental disorders. (2017). Nature *542*, 433-438. 10.1038/nature21062.

25. Fokkema, I., van der Velde, K.J., Slofstra, M.K., Ruivenkamp, C.A.L., Vogel, M.J., Pfundt, R., Blok, M.J., Lekanne Deprez, R.H., Waisfisz, Q., Abbott, K.M., et al. (2019). Dutch genome diagnostic laboratories accelerated and improved variant interpretation and increased accuracy by sharing data. Hum Mutat *40*, 2230-2238. 10.1002/humu.23896.

26. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research *46*, D1062-d1067. 10.1093/nar/gkx1153.

27. Collins, R.E., Northrop, J.P., Horton, J.R., Lee, D.Y., Zhang, X., Stallcup, M.R., and Cheng, X. (2008). The ankyrin repeats of G9a and GLP histone methyltransferases are mono- and dimethyllysine binding modules. Nat Struct Mol Biol *15*, 245-250. 10.1038/nsmb.1384.

28. Wu, H., Min, J., Lunin, V.V., Antoshenko, T., Dombrovski, L., Zeng, H., Allali-Hassani, A., Campagna-Slater, V., Vedadi, M., Arrowsmith, C.H., et al. (2010). Structural biology of human H3K9 methyltransferases. PLoS One *5*, e8570. 10.1371/journal.pone.0008570.

29. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583-589. 10.1038/s41586-021-03819-2.

30. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res *50*, D439-d444. 10.1093/nar/gkab1061.

31. consortium, w. (2018). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic acids research *47*, D520-D528. 10.1093/nar/gky949.

32. The UniProt Consortium (2016). UniProt: the universal protein knowledgebase. Nucleic acids research *45*, D158-D169. 10.1093/nar/gkw1099.

33. Hekkelman, M.L., de Vries, I., Joosten, R.P., and Perrakis, A. (2023). AlphaFill: enriching AlphaFold models with ligands and cofactors. Nat Methods *20*, 205-213. 10.1038/s41592-022-01685-y.

34. Lindeboom, R.G.H., Vermeulen, M., Lehner, B., and Supek, F. (2019). The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. Nature genetics *51*, 1645-1651. 10.1038/s41588-019-0517-5.

35. Salamov, A.A., Nishikawa, T., and Swindells, M.B. (1998). Assessing protein coding region integrity in cDNA sequencing projects. Bioinformatics *14*, 384-390. 10.1093/bioinformatics/14.5.384.

36. Gleason, A.C., Ghadge, G., Chen, J., Sonobe, Y., and Roos, R.P. (2022). Machine learning predicts translation initiation sites in neurologic diseases with nucleotide repeat expansions. PLoS One *17*, e0256411. 10.1371/journal.pone.0256411.

37. Levy, M.A., McConkey, H., Kerkhof, J., Barat-Houari, M., Bargiacchi, S., Biamino, E., Bralo, M.P., Cappuccio, G., Ciolfi, A., Clarke, A., et al. (2022). Novel diagnostic DNA methylation episignatures expand and refine the epigenetic landscapes of Mendelian disorders. HGG Adv *3*, 100075. 10.1016/j.xhgg.2021.100075.

38. Rooney, K., and Sadikovic, B. (2022). DNA Methylation Episignatures in Neurodevelopmental Disorders Associated with Large Structural Copy Number Variants: Clinical Implications. Int J Mol Sci *23*. 10.3390/ijms23147862.

39. Yamada, A., Shimura, C., and Shinkai, Y. (2018). Biochemical validation of EHMT1 missense mutations in Kleefstra syndrome. J Hum Genet *63*, 555-562. 10.1038/s10038-018-0413-3.

40. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). 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 *17*, 405-424. 10.1038/gim.2015.30.

41. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. American journal of human genetics *99*, 877-885. 10.1016/j.ajhg.2016.08.016.

42. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science *381*, eadg7492. 10.1126/science.adg7492.

43. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535-548.e524. 10.1016/j.cell.2018.12.015.

44. Dingemans, A.J.M., Hinne, M., Truijen, K.M.G., Goltstein, L., van Reeuwijk, J., de Leeuw, N., Schuurs-Hoeijmakers, J., Pfundt, R., Diets, I.J., den Hoed, J., et al. (2023). PhenoScore quantifies phenotypic variation for rare genetic diseases by combining facial analysis with other clinical features using a machine-learning framework. Nature genetics *55*, 1598-1607. 10.1038/s41588-023-01469-w.

45. Gargano, M.A., Matentzoglu, N., Coleman, B., Addo-Lartey, E.B., Anagnostopoulos, A.V., Anderton, J., Avillach, P., Bagley, A.M., Bakštein, E., Balhoff, J.P., et al. (2024). The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic acids research *52*, D1333-d1346. 10.1093/nar/gkad1005.

46. Blackburn, P.R., Tischer, A., Zimmermann, M.T., Kemppainen, J.L., Sastry, S., Knight Johnson, A.E., Cousin, M.A., Boczek, N.J., Oliver, G., Misra, V.K., et al. (2017). A Novel Kleefstra Syndrome-associated Variant That Affects the Conserved TPLX Motif within the Ankyrin Repeat of EHMT1 Leads to Abnormal Protein Folding. The Journal of biological chemistry *292*, 3866-3876. 10.1074/jbc.M116.770545.

47. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv, 531210. 10.1101/531210.

48. Caumes, R., Smol, T., Thuillier, C., Balerdi, M., Lestienne-Roche, C., Manouvrier-Hanu, S., and Ghoumid, J. (2020). Phenotypic spectrum of SHANK2-related neurodevelopmental disorder. European journal of medical genetics *63*, 104072. 10.1016/j.ejmg.2020.104072.

49. Liu, N., Zhang, Z., Wu, H., Jiang, Y., Meng, L., Xiong, J., Zhao, Z., Zhou, X., Li, J., Li, H., et al. (2015). Recognition of H3K9 methylation by GLP is required for efficient establishment of H3K9 methylation, rapid target gene repression, and mouse viability. Genes Dev *29*, 379-393. 10.1101/gad.254425.114.

50. Bian, C., Chen, Q., and Yu, X. (2015). The zinc finger proteins ZNF644 and WIZ regulate the G9a/GLP complex for gene repression. Elife *4*. 10.7554/eLife.05606.

51. Chopra, A., Cho, W.C., Willmore, W.G., and Biggar, K.K. (2020). Hypoxia-Inducible Lysine Methyltransferases: G9a and GLP Hypoxic Regulation, Non-histone Substrate Modification, and Pathological Relevance. Front Genet *11*, 579636. 10.3389/fgene.2020.579636.

52. Ea, C.K., Hao, S., Yeo, K.S., and Baltimore, D. (2012). EHMT1 protein binds to nuclear factor-κB p50 and represses gene expression. The Journal of biological chemistry *287*, 31207-31217. 10.1074/jbc.M112.365601.

53. Vinson, D.A., Stephens, K.E., O'Meally, R.N., Bhat, S., Dancy, B.C.R., Cole, R.N., Yegnasubramanian, S., and Taverna, S.D. (2022). De novo methylation of histone H3K23 by the methyltransferases EHMT1/GLP and EHMT2/G9a. Epigenetics Chromatin *15*, 36. 10.1186/s13072-022-00468-1.

54. Tsusaka, T., Kikuchi, M., Shimazu, T., Suzuki, T., Sohtome, Y., Akakabe, M., Sodeoka, M., Dohmae, N., Umehara, T., and Shinkai, Y. (2018). Tri-methylation of ATF7IP by G9a/GLP recruits the chromodomain protein MPP8. Epigenetics Chromatin *11*, 56. 10.1186/s13072-018-0231-z.

55. Tsang, B., Pritišanac, I., Scherer, S.W., Moses, A.M., and Forman-Kay, J.D. (2020). Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. Cell *183*, 1742-1756. 10.1016/j.cell.2020.11.050.

56. Zhang, J., Gao, K., Xie, H., Wang, D., Zhang, P., Wei, T., Yan, Y., Pan, Y., Ye, W., Chen, H., et al. (2021). SPOP mutation induces DNA methylation via stabilizing GLP/G9a. Nat Commun *12*, 5716. 10.1038/s41467-021-25951-3.

57. Besschetnova, A., Han, W., Liu, M., Gao, Y., Li, M., Wang, Z., Labaf, M., Patalano, S., Venkataramani, K., Muriph, R.E., et al. (2023). Demethylation of EHMT1/GLP Protein Reprograms Its Transcriptional Activity and Promotes Prostate Cancer Progression. Cancer Res Commun *3*, 1716-1730. 10.1158/2767-9764.Crc-23-0208.

58. Rots, D., Choufani, S., Faundes, V., Dingemans, A.J.M., Joss, S., Foulds, N., Jones, E.A., Stewart, S., Vasudevan, P., Dabir, T., et al. (2024). Pathogenic variants in KMT2C result in a neurodevelopmental disorder distinct from Kleefstra and Kabuki syndromes. American journal of human genetics *111*, 1626-1642. 10.1016/j.ajhg.2024.06.009.

59. Kerkhof, J., Rastin, C., Levy, M.A., Relator, R., McConkey, H., Demain, L., Dominguez-Garrido, E., Kaat, L.D., Houge, S.D., DuPont, B.R., et al. (2024). Diagnostic utility and reporting recommendations for clinical DNA methylation episignature testing in genetically undiagnosed rare diseases. Genetics in medicine : official journal of the American College of Medical Genetics *26*, 101075. 10.1016/j.gim.2024.101075.

60. Awamleh, Z., Goodman, S., Choufani, S., and Weksberg, R. (2023). DNA methylation signatures for chromatinopathies: current challenges and future applications. Hum Genet. 10.1007/s00439-023-02544-2.

61. Vermeulen, K., de Boer, A., Janzing, J.G.E., Koolen, D.A., Ockeloen, C.W., Willemsen, M.H., Verhoef, F.M., van Deurzen, P.A.M., van Dongen, L., van Bokhoven, H., et al. (2017). Adaptive and maladaptive functioning in Kleefstra syndrome compared to other rare genetic disorders with intellectual disabilities. Am J Med Genet A *173*, 1821-1830. 10.1002/ajmg.a.38280.

62. Vermeulen, K., Staal, W.G., Janzing, J.G., van Bokhoven, H., Egger, J.I.M., and Kleefstra, T. (2017). Sleep Disturbance as a Precursor of Severe Regression in Kleefstra Syndrome Suggests a Need for Firm and Rapid Pharmacological Treatment. Clin Neuropharmacol *40*, 185-188. 10.1097/wnf.0000000000000226.

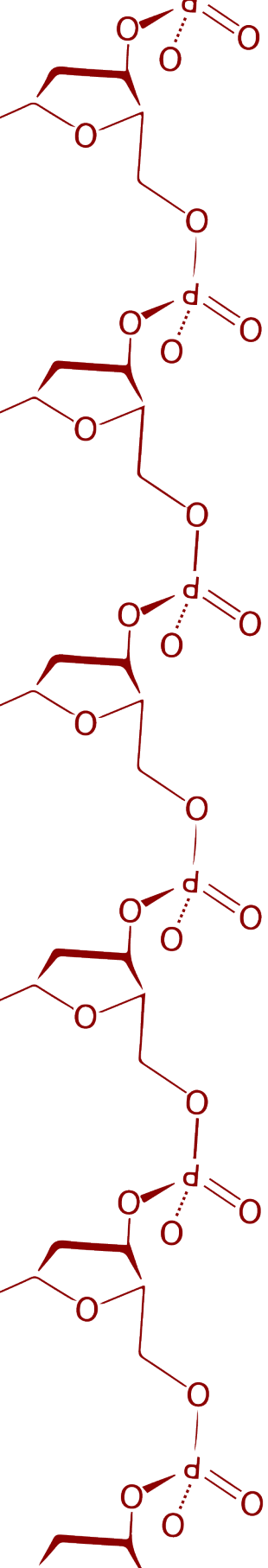
63. Kleefstra, T., van Zelst-Stams, W.A., Nillesen, W.M., Cormier-Daire, V., Houge, G., Foulds, N., van Dooren, M., Willemsen, M.H., Pfundt, R., Turner, A., et al. (2009). Further clinical and molecular delineation of the 9q subtelomeric deletion syndrome supports a major contribution of EHMT1 haploinsufficiency to the core phenotype. Journal of medical genetics *46*, 598-606. 10.1136/jmg.2008.062950.

64. Verhoeven, W.M., Egger, J.I., Vermeulen, K., van de Warrenburg, B.P., and Kleefstra, T. (2011). Kleefstra syndrome in three adult patients: further delineation of the behavioral and neurological phenotype shows aspects of a neurodegenerative course. Am J Med Genet A *155a*, 2409-2415. 10.1002/ajmg.a.34186.

65. Haseley, A., Wallis, K., and DeBrosse, S. (2021). Kleefstra syndrome: Impact on parents. Disabil Health J *14*, 101018. 10.1016/j.dhjo.2020.101018.

66. Morison, L.D., Kennis, M.G.P., Rots, D., Bouman, A., Kummeling, J., Palmer, E., Vogel, A.P., Liegeois, F., Brignell, A., Srivastava, S., et al. (2024). Expanding the phenotype of Kleefstra syndrome: speech, language and cognition in 103 individuals. Journal of medical genetics. 10.1136/jmg-2023-109702.

67. Cuypers, M., Tobi, H., Naaldenberg, J., and Leusink, G.L. (2021). Linking national public services data to estimate the prevalence of intellectual disabilities in The Netherlands: results from an explorative population-based study. Public Health *195*, 83-88. 10.1016/j.puhe.2021.04.002.



Chapter 6:

Refining the 9q34.3 microduplication syndrome reveals mild neurodevelopmental features associated with a distinct global DNA methylation profile

Published: Clinical Genetics. 2024 Jun;105(6):655-660.

Authors
Dmitrijs Rots*, Kathleen Rooney*, Raissa Relator, Jennifer Kerkhof, Haley McConkey, Rolph Pfundt, Carlo Marcelis, Marjolein H. Willemsen, Johanna M. van Hagen, Petra Zwijnenburg, Marielle Alders, Katrin Õunap, Tiia Reimand, Olga Fjodorova, Siren Berland, Eva Liahjell, Ognjen Bojovic, Marjolein Kriek, Claudia Ruivenkamp, Maria Teresa Bonati, Han G. Brunner, Lisenka E.L.M. Vissers, Bekim Sadikovic**, Tjitske Kleefstra**

*,** These authors contributed equally to this work

Abstract

Precise regulation of gene expression is important for correct neurodevelopment. 9q34.3 deletions affecting the *EHMT1* gene result in a syndromic neurodevelopmental disorder named Kleefstra syndrome. In contrast, duplications of the 9q34.3 locus encompassing *EHMT1* have been suggested to cause developmental disorders, but only limited information has been available.

We have identified 15 individuals from 10 unrelated families, with 9q34.3 duplications <1.5Mb in size, encompassing *EHMT1* entirely. Clinical features included mild developmental delay, mild intellectual disability or learning problems, autism spectrum disorder, and behavior problems. The individuals did not consistently display dysmorphic features, congenital anomalies, or growth abnormalities. DNA methylation analysis revealed a weak DNAm profile for the cases with 9q34.3 duplication encompassing *EHMT1*, which could segregate the majority of the affected cases from controls.

This study shows that individuals with 9q34.3 duplications including *EHMT1* gene present with mild non-syndromic neurodevelopmental disorders and DNA methylation changes different from Kleefstra syndrome.

Keywords:
EHMT1, neurodevelopmental disorder, 9q34.3 duplication, DNA methylation

Introduction

Brain and neuronal development are complicated processes requiring precise regulation of gene expression [1,2]. Therefore, disrupted genes encoding factors that modify chromatin (so-called epigenetic machinery), are a common cause of monogenic neurodevelopmental disorders (NDDs)[2].

EHMT1 is a member of the epigenetic machinery. Through the complex with EHMT2 and other proteins, it is involved in gene expression and chromatin structure regulation by histone-3 lysine-9 methylation [3,4]. *EHMT1* haploinsufficiency results in syndromic NDD named Kleefstra syndrome (KLEFS1), previously known as 9q34.3 deletion syndrome (MIM: 610253) [4,5]. KLEFS1 is mainly characterized by moderate-severe intellectual disability, recognizable facial features, hypotonia, microcephaly, short stature, and congenital anomalies [5]. Additionally, individuals with KLEFS1 have specific DNA methylation changes [6,7].

Recently, a large cohort study shown that the majority of haploinsufficient genes are predicted to also be triplosensitive, including *EHMT1* [8]. While haploinsufficiency of *EHMT1* is well-known, only several individuals with 9q34.3 duplications of variable size have been described [9]. Therefore, in this study, we aimed to describe clinical, molecular and DNA methylation features of individuals with small 9q34.3 microduplications entirely encompassing *EHMT1*.

Methods

For this study, we have focused on individuals with small 9q34.3 microduplications (<1.5 Megabases in size) entirely encompassing *EHMT1* without other known haploinsufficient or triplosensitive genes. We have collected clinical and molecular data of 15 cases from 10 unrelated families via the Radboudumc expertise center, international collaborations, and literature. All individuals consented for the study. Two of the cases have been published previously by Bonati et al. [9]. The duplications were identified in diagnostic settings using chromosomal microarrays or exome sequencing. To confirm breakpoints, PCR-free genome sequencing was performed for two individuals, as described before [10].

For 11/15 included cases, blood-derived DNA was available for methylation analysis (**Table S1**). The analysis of global DNA methylation profile was performed based on our laboratory's previously described methods [6,11].

Supplemental methods

The analysis of global DNA methylation profile was performed using Illumina Infinium EPIC v1 BeadChip arrays based on our laboratory's workflow and EpiSign software™. In short, methylation analysis was performed in RStudio (v.1.4.1106) with R (v.4.1.1). Data was normalized using the Illumina method and background correction performed using the minfi package (v.1.40.0). After removal of probes located on chrX and chrY (n=19,681), cross-reacting probes (n=72,487), probes containing single nucleotide polymorphisms at or near the CpG sites (n=31,647), probes suggested for removal by Illumina after manufacturing change (n=1,967), probes with beta values of 0 or 1, and the top 1% most variable probes, 694,434 probes remained for the analysis. To assess for outliers and assess batch structure, principal component analysis was performed. The MatchIt package (v.4.3.4) was utilized to randomly select age, sex and array type matched controls from the EpiSign Knowledge Database at a ratio of 1:5. One case (Bonati et al., P1) was excluded from the discovery cases, because it constantly was classified as control in the exploratory analysis. In total, 10/11 cases and 50 controls were utilized for the final analysis. The excluded case has the most distal duplication with the least genes included.

To identify differentially methylated probes (DMPs), beta values were converted to M-values and linear regression performed using the limma package (v.3.50.0). Probe selection was performed in three stages: firstly, 800 probes were retained with the highest product of absolute methylation differences between cases and controls and the negative of the logarithm of p-values. Next, 400 probes with the highest area under the receiver's operating characteristic curve were retained. Lastly, using Pearson's correlation coefficients, all probes with a pair-wise correlation >0.7 were eliminated. Scaling of the pair-wise Euclidean distance was performed using the ggplot2 package (v.3.3.5) and visualized in heatmap and multidimensional scaling (MDS) plots. A support vector machine (SVM) classifier was constructed using the e1071 package (v.1.7-9) and tested using leave-one-out cross validation method as previously described [11]. The SVM model was trained using the 261 DMPs and 75% of controls and other neurodevelopmental disorder samples on EpiSign™. The 25% remaining are used as testing samples (grey). SVM hyperparameters, such as kernel and class weights, were determined using grid search. Finally, analysis was also performed to identify differentially methylated regions (DMRs) using DMRcate package (v.2.8.3) with regions defined to have least three contiguous probes within 1000bp, and 5% minimum methylation difference between cases and controls. Additionally, we have classified the samples against the other available episignatures (**Figure 2C**):

ADCADN: Cerebellar ataxia deafness and narcolepsy syndrome; AUTS18: Susceptibility to autism 18; BEFAHRS: Beck-Fahrner syndrome: BFLS: Borjeson–Forssman–Lehmann syndrome; BISS: Blepharophimosis-intellectual disability SMARCA2 syndrome; CdLS: Cornelia de Lange syndrome; CHARGE: CHARGE syndrome; Chr16p11.2del: Chromosome 16p11.2 deletion syndrome; CSS: Coffin–Siris syndrome; CSS4: Coffin-Siris syndrome 4; CSS9: Coffin–Siris syndrome 9; Down: Down syndrome; Dup7: 7q11.23 duplication syndrome; DYT28: Dystonia 28; EEOC: Epileptic encephalopathy-childhood onset; FLHS: Floating-Harbour syndrome; GTPTS: Genitopatellar syndrome; HMA: Hunter-McAlpine craniosynostosis syndrome; HVDAS: Helsmoortel–van der Aa syndrome; ICF: Immunodeficiency-centromeric instability-facial anomalies syndrome; IDDSELD: Intellectual developmental disorder with seizures and language delay; Kabuki: Kabuki syndrome; KDVS: Koolen-De Vries syndrome; Kleefstra: Kleefstra syndrome; LLS: Luscan-Lumish syndrome; MKHK: Menke-Hennekam syndrome; MLASA2: Myopathy lactic acidosis and sideroblastic anemia 2; MRD23: Intellectual developmental disorder 23; MRD51: Intellectual developmental disorder 51; MRX93: Intellectual developmental disorder X-linked 93; MRX97: Intellectual developmental disorder X-linked 97; MRXSA: Intellectual developmental disorder X-linked syndromic Armfield type; MRXSCH: Intellectual developmental disorder X-linked syndromic Christianson type; MRXSCJ: Intellectual developmental disorder X-linked syndromic Claes-Jensen type; MRXSN: Intellectual developmental disorder X-linked syndromic Nascimento type; MRXSSR: Intellectual developmental disorder X-linked syndromic Snyder–Robinson type; PHMDS: Phelan–McDermid syndrome; PRC2: PRC2 complex (Weaver and Cohen-Gibson) syndrome; RENS1: Renpenning syndrome; RMNS: Rahman syndrome; RSTS: Rubinstein–Taybi syndrome; SBBYSS: Ohdo syndrome; Sotos: Sotos syndrome; TBRS: Tatton–Brown–Rahman syndrome; WDSTS: Wiedemann–Steiner syndrome; WHS: Wolf-Hirschhorn syndrome; Williams: Williams syndrome.

Publicly available DNAm datasets are deposited in GEO, and include data for various developmental syndromes searchable by the disorder name (e.g., Kabuki syndrome, Sotos syndrome, CHARGE syndrome, immunodeficiency-centromeric instability-facial anomalies (ICF) syndrome, Williams-Beuren syndrome, Chr7q11.23 duplication syndrome, BAFopathies, Down syndrome).

Data in the EpiSign Knowledge Database, including the 9q34.3 cohort, are not available due to Research Ethics Board and institutional restrictions. EpiSign™ is proprietary software and is not publicly available.

Results

To define the clinical and molecular spectrum of individuals with 9q34.3 duplications, we have recruited 15 individuals from 10 families (**Table 1** and **Table S1**). In 4/15 cases, the duplication has been confirmed to occur *de novo*, while in 8/15 cases, it was inherited from a mildly affected parent and in 3/15, the inheritance was unknown.

Table 1. Main clinical features among individuals with 9q34.3 copy number variants.

Feature	Frequency among 9q34.3 duplication cohort N=15 (%)
Inheritance	4/15 *de novo* (27%) (6 paternal/2 maternal/3 unknown)
Sex (Males/Total)	10/15 (67%)
Growth	
Low birth weight	1/9 (11%)
Overweight or obesity[KLEFS1]	2/13 (15%)
Short stature[KLEFS1]	1/13 (8%)
Abnormal head circumference[KLEFS1]	0/10 (0%)
Neurodevelopmental and psychiatric issues	
Language/speech delay[KLEFS1]	12/13 (92%)
Motor delay[KLEFS1]	8/12 (67%)
Intellectual disability[KLEFS1] or learning problems	10/13 (77%)
Autism spectrum disorder[KLEFS1]	5/13 (38%)
Behavior problems, not autism spectrum	8/14 (57%)
Psychoses or Schizophrenia[KLEFS1]	0/15 (0%)
Neurological issues	
Seizures[KLEFS1]	2/15 (13%)
Hypotonia[KLEFS1]	2/14 (14%)
Sleep disturbances[KLEFS1]	4/10 (40%)
Congenital anomalies	
Congenital heart disease[KLEFS1]	0/15 (0%)
Cleft lip or palate	0/15 (0%)
Genitourinary abnormalities[KLEFS1]	0/8 (0%)

KLEFS1= features typical for Kleefstra syndrome

Most of the identified duplications (8/10) were <1Mb in size (~0.3 to 1.4Mb), containing from 3 to 53 protein-coding genes. The duplication positions with genic content are depicted in **Figure 1** and provided in **Table S1**. Only *EHMT1* and *ARRDC1* genes overlap all duplications. To confirm and specify the (*de novo*) duplications in

two cases (P2 and P1 from Bonati et al.), genome sequencing was performed which confirmed the duplications contain full-length *EHMT1* and are in tandem.

The most prevalent feature among the individuals was mild developmental delay (DD) (present in 92%) and mild intellectual disability (ID) or learning problems (77%). Additionally, these individuals commonly presented with autism spectrum disorder (38%) and/or other behavior problems (57%) like aggression, anxiety etc. Similar to KLEFS1, sleeping issues were commonly reported (40%), but were not associated with psychoses (**Table 1** and **Table S1**).

Figure 1. Genomic view of the 9q34.3 regions with microduplications.

Genes are colored based on the pTriplo (from 0 to 1; light blue to dark blue, accordingly) and pHaplo scores (from 0 to 1; light grey to black, accordingly). pTriplo or pHaplo score values are shown after the gene names with higher score representing higher level of predicted dosage sensitivity.

These individuals did not display recognizable or prominent dysmorphic features, nor had any congenital anomalies. One adult individual (P8, brother of P7), was reported to have no neurodevelopmental or neurological symptoms, suggesting incomplete penetrance, but we cannot exclude presence of mild symptoms.

To determine whether 9q34.3 duplications encompassing *EHMT1* (further named "EHMT1dup" for simplicity) would cause DNA methylation changes, we compared 10 affected cases against controls. We identified 261 differentially methylated probes (**Table S2**) associated with EHMT1dup, but did not identify any significant DMRs. We demonstrated that the selected DMPs were capable of segregating the majority of the affected EHMT1dup cases from controls using unsupervised clustering (**Figure 2A-B**). Using the constructed SVM classifier (**Figure 2C**), all EHMT1dup positive cases showed a methylation variant pathogenicity (MVP) score close to 1 compared with the negative cases close to 0. We observed no elevation in MVP score for any KLEFS1 cases. One case (P1 from Bonati et al.) did not share the same DNA methylation changes and was classified as "negative" (**Figure 2**). Additionally, leave-one-out cross validation suggested low sensitivity of the identified methylation profile (**Figure S1**).

Additionally, the classifier did not show complete specificity for the 9q34.3 duplications, as two of the ADCADN (MIM: 604121) training, and two testing samples, and one case each from the CSS9 (MIM: 615866), MRXSCJ (MIM: 300534) and WHS (MIM: 194190) cohorts, had elevated MVP scores. When compared EHMT1dup separately with the ADCADN episignature, we observed that the methylation profiling can distinguish between the two conditions (**Figure S2 A-B**). Therefore, the identified DNAm changes are mild and not fully sensitive and specific, so we call them as EHMT1dup DNA methylation "profile" rather than "episignature".

Finally, we compared the EHMT1dup samples with KLEFS1 samples [6] (**Figure S2 C-D**). We observed that the KLEFS1 cases do not overlap with EHMT1dup. We observed that the EHMT1dup DNA methylation profile contained 162 (162/261, 62%) hypomethylated CpGs compared with 132 (132/136, 97%) in the KLEFS1 profile. Therefore, both cohorts' DMPs are predominantly hypomethylated, with no DMPs overlap between the two cohorts but different DMPs overlapping the same gene: *TRAPPC9* (MIM: 611966). *TRAPPC9* is associated with autosomal recessive intellectual developmental disorder (MIM: 613192).

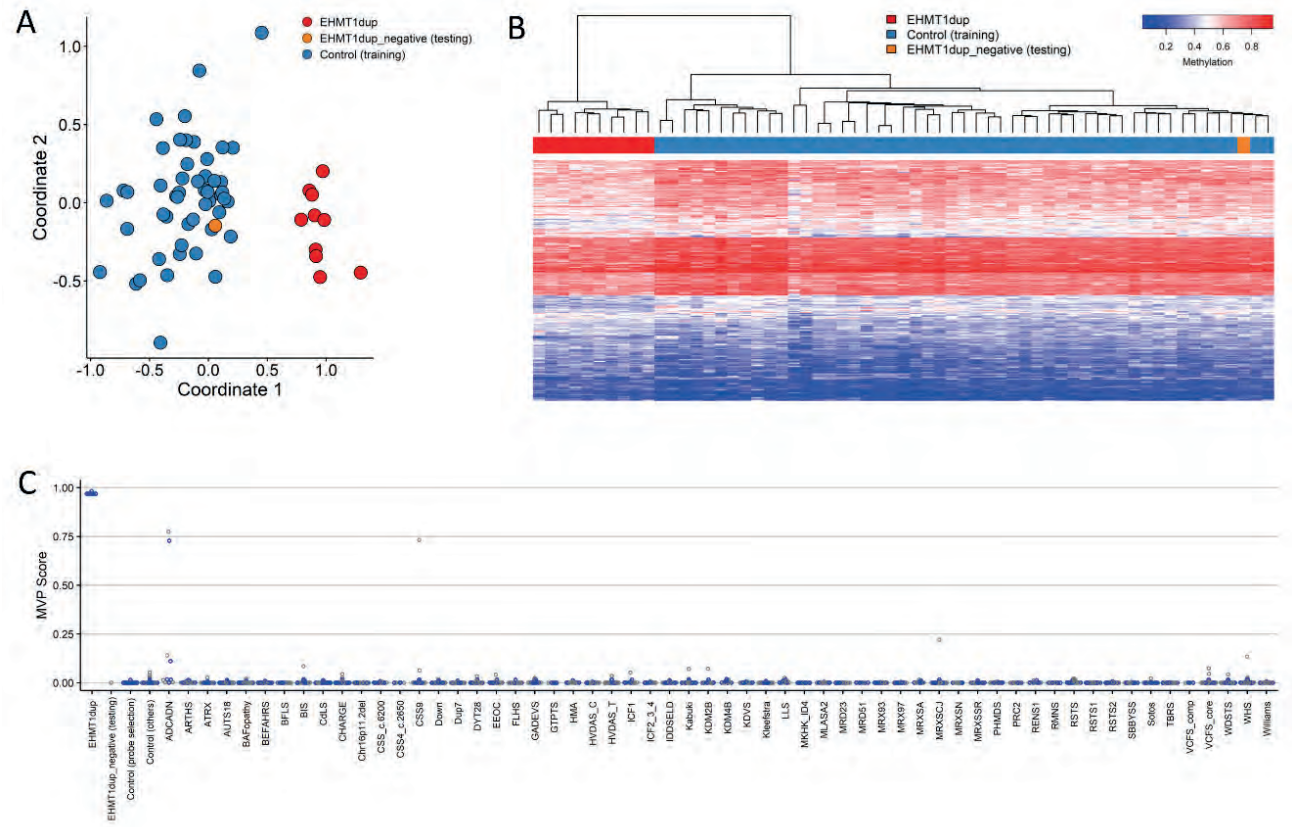


Figure 2. 9q34.3 duplication encompassing *EHMT1* (EHMT1dup) DNA methylation profile

A. *Multidimensional scaling plot shows clustering of the EHMT1dup cases (red) together and away from controls (blue) and one negative case (orange);* ***B.*** *Heatmap indicates clear separation of the EHMT1dup cases from controls using the identified 261 DMPs;* ***C.*** *The classification results using EHMT1dup SVM model.*

Discussion

In this study, we further define that individuals with 9q34.3 duplications encompassing *EHMT1* share non-syndromic NDD with mild DD/ID, autism spectrum disorder and behavioral problems, as well as a common DNA methylation profile. Importantly, 9q34.3 duplications involving *EHMT1* and nearby genes are not found in healthy control cohorts [8,12] and contain several genes predicted to be triplosensitive (pTriplo >0.9), notably, including *EHMT1* [8]. While the clinical features are non-specific and variable, common genetic findings and DNA methylation profile likely confirm the 9q34.3 microduplications as the main cause of the NDD among the described individuals.

Though *EHMT1* is a compelling candidate for being the main 9q34.3 triplosensitivity driver[9], it is almost impossible to have an isolated duplication of a single gene in this region because of the high gene density in 9q34.3. All duplications described here include at least one additional gene that is predicted to be triplosensitive, e.g. *ZMYND19* (pTriplo=0.91) or *CACNA1B* (pTriplo=1)[8]. *ZMYND19* has not been associated with a human phenotype, but *CACNA1B* is associated with a recessive NDD with epilepsy[13]. To reduce the bias and possible contribution by other genes, individuals with large duplications were not included in the study.

Clinical features, as well as DNA methylation changes associated with the duplications are distinct from those found among KLEFS1 individuals [5]. Cellular effects of the *EHMT1* overexpression vs. loss are also different: e.g., *EHMT1* is overexpressed in many cancers, and it induces cell proliferation and resistance to treatment, while *EHMT1* loss results in cell apoptosis or reduced proliferation [14,15]. In *Drosophil*a, both the *EHMT1* ortholog overexpression and loss resulted in reduced learning and memory but were more prominent for loss [16]. This indicates that *EHMT1* expression is highly dosage sensitive for correct neurodevelopment.

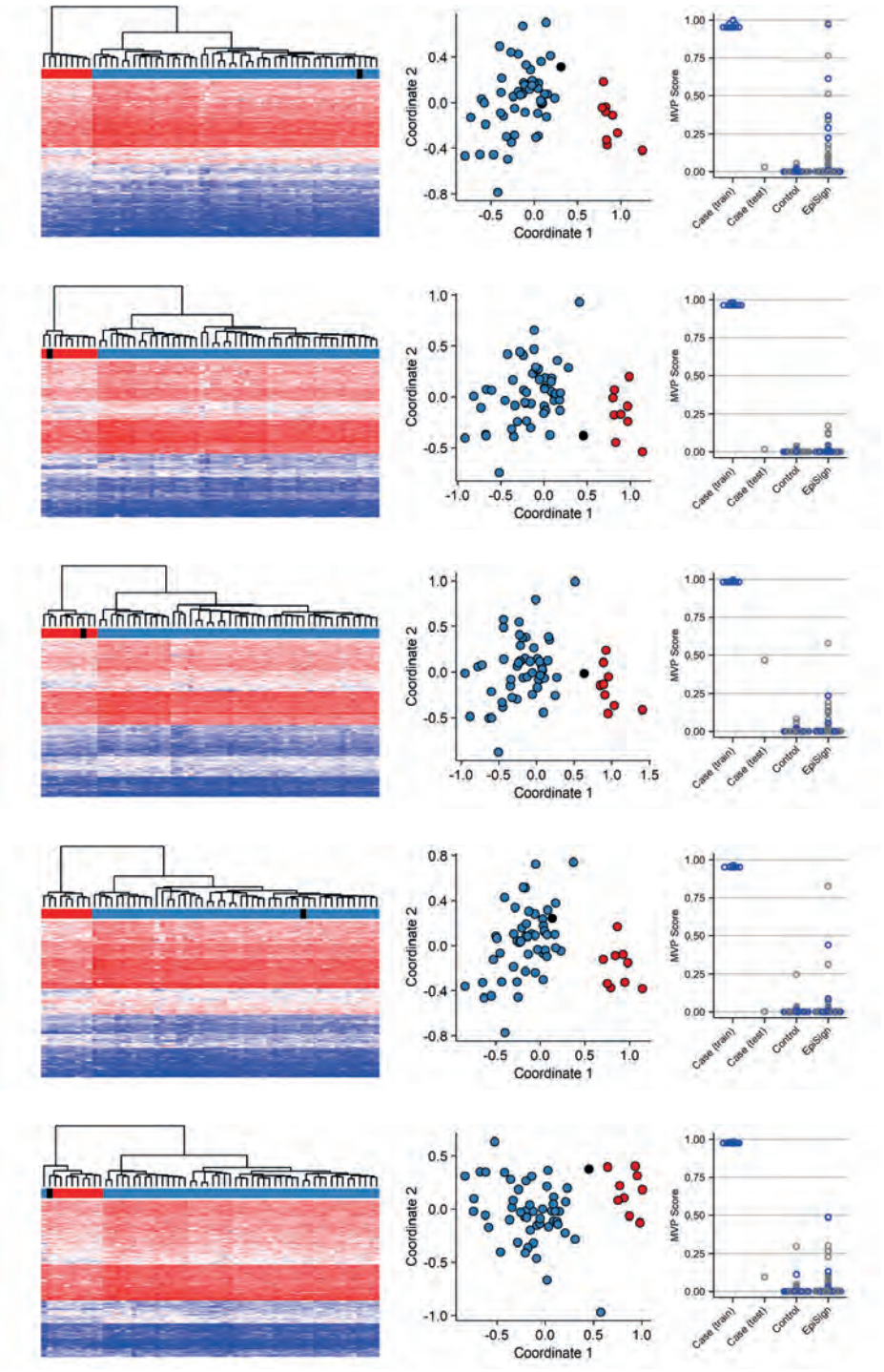
Interestingly, the 9q34.3 duplication DNA methylation profile is not "mirroring" or overlapping the KLEFS1 episignature, as has been demonstrated for some reciprocal microdeletion/microduplication syndromes [17]. It is possible that the difference in cellular effects of the overexpression vs. loss results in different DNA methylation changes. It is also possible that the profile identified among the 9q34.3 microduplication cases results from overexpression of several genes or another gene (e.g. *ZMYND19*), as has been shown for 22q13.3 deletion [17,18]. Noticeably, the individual who was classified negatively on the DNA methylation profile (P1 from Bonati et al.), has the most distal duplication with the least genes included (besides

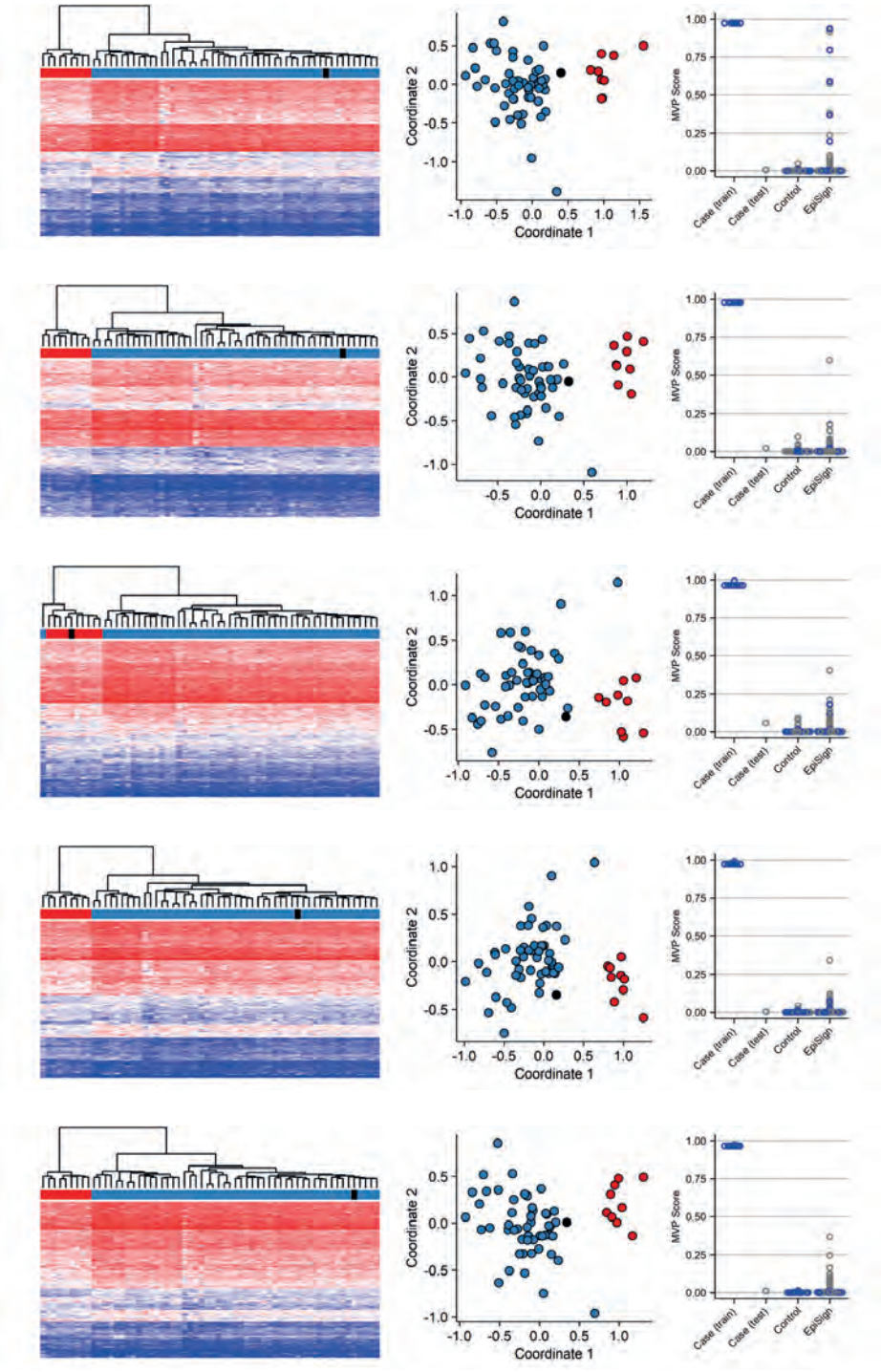
EHMT1, includes only *CACNA1B* and *ARRDC1*), which might explain the negative results. By using genome sequencing the duplication in this case was confirmed to include the entire *EHMT1* gene, as well as previously was shown to result in *EHMT1* overexpression [9].

While EHMT1 is one of the main 9q34.3 triplosensitive genes, a combination of duplicated genes likely contributes to the 9q34.3 microduplication phenotype and, possibly, to the DNA methylation changes. The identified DNA methylation profile is mild and more cases with different breakpoints are necessary to clarify the main drivers of the methylation changes and triplosensitivity of the 9q34 region.

Supplemental information

Figure S1: Leave-one-out cross validation results.




Coordinate 2
Coordinate 1
MVP Score
Case (train)
Case (test)
Control
EpiSign

In each cross-validation set, a single test case sample (dark blue) is used as testing. The remaining EHMT1dup cases used for episignature training are shown in red and control training samples shown in blue in the heatmap and MDS plots. The last plots demonstrate the MVP scores of the Support Vector Machine (SVM) classifier model that was trained using the selected affected EHMT1dup episignature probes from the training cases, 75% of controls and other EpiSign™ samples (blue). The remaining 25% of controls and other disorder samples were used as testing.

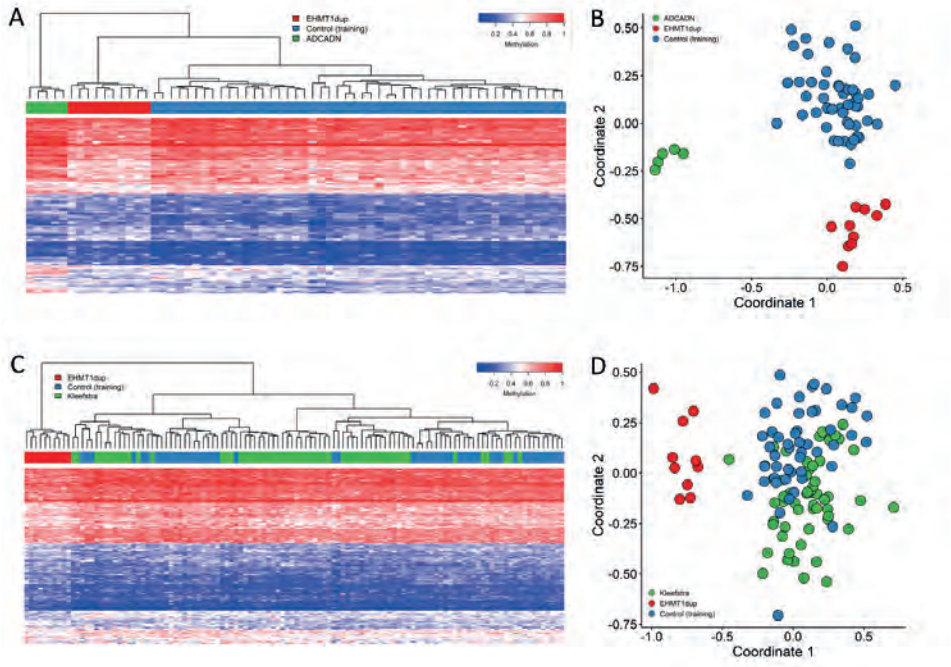


Figure S2: Comparison of the EHMT1dup DNA methylation profile with ADCADN.

A. *Heatmap indicates separate clustering of ADCADN cases (green) from EHMT1dup cases (red) and controls (blue). Each row represents one of the 261 differentially methylated probes (DMPs) selected as the DNA methylation profile and each column represents either an affected EHMT1dup or ADCADN case or a control;* ***B.*** *Multidimensional scaling plot also shows distinct clustering of the EHMT1dup cohort from ADCADN cases and controls;* ***C.*** *Heatmap indicates separate clustering of Kleefstra cases (green) with controls and separate from EHMT1dup cases;* ***D.*** *MDS also shows distinct clustering of the EHMT1dup cohort from Kleefstra cases that are closer to controls.*

Tables S1, S2 supporting the findings of this study are available online in the Supplementary material of this article at: DOI: 10.1111/cge.14498

Acknowledgments

This work was financially supported by Aspasia grant of the Dutch Research Council (015.014.036) and Netherlands Organization for Health Research and Development (91718310) awarded to T.K.; K.O. and T.R. are supported by Estonian Research Council grant PRG471. Funding for this study is provided in part by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-188) to B.S.

This work has been partly generated within the ERN-ITHACA which is funded by the European Union's EU4Health program (101085231) and with support of the Solve-RD project (779257) to H.G.B. and L.E.L.M.V.

We thank Michael Kwint for technical assistance and Elke de Boer for the help with data analysis.

Conflict of Interest

B.S. is a shareholder in EpiSign Inc., a biotechnology company involved in commercialization of EpiSign™ technology.

Data Availability

All data analysed in the manuscript are available in the main text or supplemental materials. All identified duplications are submitted to ClinVar (accession numbers: SCV004027852-SCV004027862). Raw DNA methylation data are not available due to institutional and ethics restrictions.

Ethics

This study was conducted in accordance with the regulations of the Western University Research Ethics Board (REB116108; REB106302) and the ethics committee of Arnhem-Nijmegen (Nr.2018-4540).

References

1. Mossink, B., Negwer, M., Schubert, D., and Nadif Kasri, N. (2021). The emerging role of chromatin remodelers in neurodevelopmental disorders: a developmental perspective. Cell Mol Life Sci *78*, 2517-2563. 10.1007/s00018-020-03714-5.
2. Ciptasari, U., and van Bokhoven, H. (2020). The phenomenal epigenome in neurodevelopmental disorders. Hum Mol Genet *29*, R42-r50. 10.1093/hmg/ddaa175.
3. Bian, C., Chen, Q., and Yu, X. (2015). The zinc finger proteins ZNF644 and WIZ regulate the G9a/GLP complex for gene repression. Elife *4*. 10.7554/eLife.05606.
4. Kleefstra, T., Brunner, H.G., Amiel, J., Oudakker, A.R., Nillesen, W.M., Magee, A., Geneviève, D., Cormier-Daire, V., van Esch, H., Fryns, J.P., et al. (2006). Loss-of-function mutations in euchromatin histone methyl transferase 1 (EHMT1) cause the 9q34 subtelomeric deletion syndrome. American journal of human genetics *79*, 370-377. 10.1086/505693.
5. Willemsen, M.H., Vulto-van Silfhout, A.T., Nillesen, W.M., Wissink-Lindhout, W.M., van Bokhoven, H., Philip, N., Berry-Kravis, E.M., Kini, U., van Ravenswaaij-Arts, C.M., Delle Chiaie, B., et al. (2012). Update on Kleefstra Syndrome. Mol Syndromol *2*, 202-212. 10.1159/000335648.
6. Aref-Eshghi, E., Kerkhof, J., Pedro, V.P., Barat-Houari, M., Ruiz-Pallares, N., Andrau, J.C., Lacombe, D., Van-Gils, J., Fergelot, P., Dubourg, C., et al. (2020). Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in 42 Mendelian Neurodevelopmental Disorders. American journal of human genetics *106*, 356-370. 10.1016/j.ajhg.2020.01.019.
7. Goodman, S.J., Cytrynbaum, C., Chung, B.H.-Y., Chater-Diehl, E., Aziz, C., Turinsky, A.L., Kellam, B., Keller, M., Ko, J.M., Caluseriu, O., et al. (2020). EHMT1 pathogenic variants and 9q34.3 microdeletions share altered DNA methylation patterns in patients with Kleefstra syndrome. Journal of Translational Genetics and Genomics *4*, 144-158. 10.20517/jtgg.2020.23.
8. Collins, R.L., Glessner, J.T., Porcu, E., Lepamets, M., Brandon, R., Lauricella, C., Han, L., Morley, T., Niestroj, L.M., Ulirsch, J., et al. (2022). A cross-disorder dosage sensitivity map of the human genome. Cell *185*, 3041-3055.e3025. 10.1016/j.cell.2022.06.036.
9. Bonati, M.T., Castronovo, C., Sironi, A., Zimbalatti, D., Bestetti, I., Crippa, M., Novelli, A., Loddo, S., Dentici, M.L., Taylor, J., et al. (2019). 9q34.3 microduplications lead to neurodevelopmental disorders through EHMT1 overexpression. Neurogenetics *20*, 145-154. 10.1007/s10048-019-00581-6.
10. van der Sanden, B., Schobers, G., Corominas Galbany, J., Koolen, D.A., Sinnema, M., van Reeuwijk, J., Stumpel, C., Kleefstra, T., de Vries, B.B.A., Ruiterkamp-Versteeg, M., et al. (2022). The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. European journal of human genetics : EJHG. 10.1038/s41431-022-01185-9.
11. Levy, M.A., McConkey, H., Kerkhof, J., Barat-Houari, M., Bargiacchi, S., Biamino, E., Bralo, M.P., Cappuccio, G., Ciolfi, A., Clarke, A., et al. (2022). Novel diagnostic DNA methylation episignatures expand and refine the epigenetic landscapes of Mendelian disorders. HGG Adv *3*, 100075. 10.1016/j.xhgg.2021.100075.
12. MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic acids research *42*, D986-992. 10.1093/nar/gkt958.
13. Gorman, K.M., Meyer, E., Grozeva, D., Spinelli, E., McTague, A., Sanchis-Juan, A., Carss, K.J., Bryant, E., Reich, A., Schneider, A.L., et al. (2019). Bi-allelic Loss-of-Function CACNA1B Mutations in Progressive Epilepsy-Dyskinesia. American journal of human genetics *104*, 948-956. 10.1016/j.ajhg.2019.03.005.

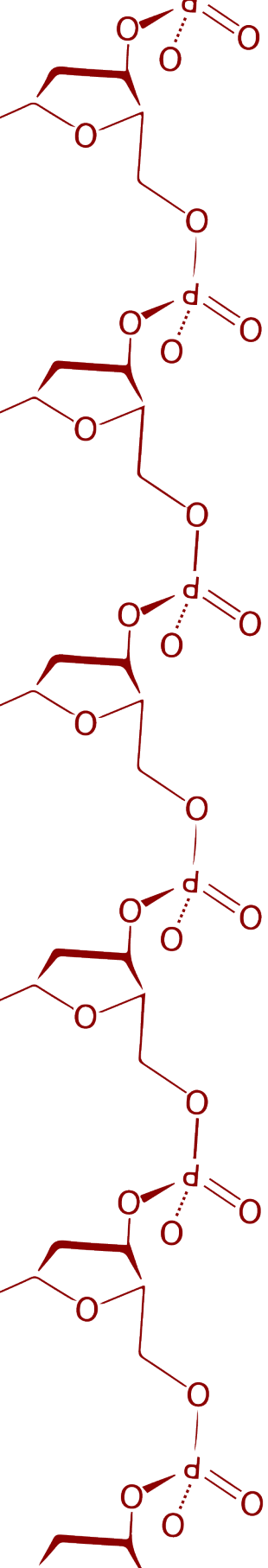
14. Nachiyappan, A., Gupta, N., and Taneja, R. (2022). EHMT1/EHMT2 in EMT, cancer stemness and drug resistance: emerging evidence and mechanisms. Febs j *289*, 1329-1351. 10.1111/febs.16334.

15. Zhang, J., Gao, K., Xie, H., Wang, D., Zhang, P., Wei, T., Yan, Y., Pan, Y., Ye, W., Chen, H., et al. (2021). SPOP mutation induces DNA methylation via stabilizing GLP/G9a. Nat Commun *12*, 5716. 10.1038/s41467-021-25951-3.

16. Kramer, J.M., Kochinke, K., Oortveld, M.A., Marks, H., Kramer, D., de Jong, E.K., Asztalos, Z., Westwood, J.T., Stunnenberg, H.G., Sokolowski, M.B., et al. (2011). Epigenetic regulation of learning and memory by Drosophila EHMT/G9a. PLoS Biol *9*, e1000569. 10.1371/journal.pbio.1000569.

17. Rooney, K., and Sadikovic, B. (2022). DNA Methylation Episignatures in Neurodevelopmental Disorders Associated with Large Structural Copy Number Variants: Clinical Implications. Int J Mol Sci *23*. 10.3390/ijms23147862.

18. Schenkel, L.C., Aref-Eshghi, E., Rooney, K., Kerkhof, J., Levy, M.A., McConkey, H., Rogers, R.C., Phelan, K., Sarasua, S.M., Jain, L., et al. (2021). DNA methylation epi-signature is associated with two molecularly and phenotypically distinct clinical subtypes of Phelan-McDermid syndrome. Clin Epigenetics *13*, 2. 10.1186/s13148-020-00990-7.



Chapter 7:

General discussion

Mendelian NDDs: next steps after discovery

With the implementation of CMA and ES, the number of identified novel Mendelian NDDs has skyrocketed in the past decade. Currently, more than 1500 genes have been implicated in NDD development[1], and hundreds of novel NDDs are waiting to be discovered[2]. However, the rarity of such conditions limits the initial description to a small cohort or several individuals, restricting the amount of information (including genotypic and phenotypic spectrum) available on any novel condition[3]. Therefore, this constrains the recognition and correct diagnosis and care of individuals with such disorders. With a detailed clinical description of a disorder, we can identify its typical features, understand the evolution of symptoms over time and/or at different ages, and provide tailored care. Therefore, **the next crucial step after the discovery of novel Mendelian NDD is their detailed characterization**. Such an effort requires broad international collaboration, data sharing, and extensive time to identify and collect enough novel individuals to comprehensively define each Mendelian NDD. For example, *KMT2C* was first suggested as a cause of syndromic NDD in 2012[4] and confirmed as a Mendelian NDD in 2017[5] with the description of five affected individuals. More than a decade after the original report, we described in **chapter 4** the phenotypic, genotypic, and DNA methylation features and spectrum of this condition, based on a cohort of 80 affected individuals, allowing a more precise definition of the *KMT2C*-related NDD.

This thesis is focused on a detailed characterization of Mendelian NDDs caused by disruption of epigenetic machinery components, specifically those affecting *SRCAP* (**chapter 2**), *KDM6B* (**chapter 3**), *KMT2C* (**chapter 4**), and *EHMT1* (**chapters 5 and 6**) genes. I show that systematic and in-depth analysis of the clinical and molecular spectrum of extended cohorts of affected individuals allows us to achieve multiple clinically and biologically relevant purposes:

1) improve rare NDD diagnostics and care by providing information for better interpretation of clinical and genetic data for individuals with variants in the described genes;
2) identify novel genotype-phenotype correlations and/or phenotype subgroups;
3) improve our understanding of disease pathogenesis and affected protein functions;
4) identify gene domains/regions/transcripts important for correct protein functioning.

Beyond the *de novo* paradigm

The majority of severe NDDs are caused by pathogenic *de novo* variants[6-8]. However, with the description of a large cohort, we identify that the majority of NDDs represent a broad spectrum of severity. For example, Kleefstra syndrome is typically associated with moderate to severe ID[9], but we and others have also identified affected individuals without an intellectual disability. Therefore, we and others[10,11] observe an increased number of identified individuals with pathogenic variants inherited from a mildly affected or clinically unaffected parent and is likely associated with the frequency of mild presentation of a condition. In **chapter 3**, we identified 9/85 individuals with inherited *KDM6B*-related NDD; in **chapter 4** - 10/81 individuals with inherited *KMT2C*-related NDD; **in chapter 5** - 7/191 individuals with inherited Kleefstra syndrome; and in **chapter 6** - 3/10 families with an inherited 9q34.3 duplication. These observations mean that if a variant in a gene associated with severe NDD is inherited from a parent, the variant should not be considered benign automatically but should be thoroughly interpreted using all available evidence (including the detailed parental phenotype and further segregation in the family)[12]. Additionally, filtering out inherited variants during trio exome/genome analysis could result in a missed diagnosis. Identifying such inherited causes has a high recurrence risk and could have an important impact on family planning. To avoid missing inherited cases, even trio exome/genome analysis should begin with a thorough interpretation of variants within an *in silico* gene panel (or panels) associated with the individual's phenotype requested by an expert clinician and only in the case of "negative" findings should proceed with an "open" exome/genome analysis of a limited variant set (*de novo*, biallelic and X-linked), as suggested by the ESHG guidelines[13].

Variant effect interpretation

In vivo and *in vitro* functional analyses have been considered the gold standard for the evaluation of variant functional effects[14,15]. However, such functional analyses have multiple limitations: 1) they are usually laborious, 2) they are rarely available in diagnostic settings, 3) they typically assess only a single function, and 4) they only try to imitate human disease conditions[14,16]. Therefore, **the best possible model organism for human disease is – human**[7]. I argue that with comprehensive clinical and molecular information in combination with "downstream" genetic variant evidence (e.g., DNAm signature, differential RNA expression, or metabolomic analysis) on human samples, the same or similar conclusions can be drawn as

with model organisms, but they would accurately and directly reflect the human condition. However, *in vitro* and *in vivo* functional analyses are still highly valuable[16], especially for conditions or variants with limited information and/or without available "downstream" analyses. For example, in **chapter 3**, we failed to identify a DNAm signature for the *KDM6B*-related NDD, so variant testing using a *Drosophila* dual gain-of-function assay allowed multiple *KDM6B* variant testing, while in **chapter 5**, functional *in vitro* analyses allowed us to confirm the functional effects of *EHMT1* variants predicted from protein 3D structure and DNAm signature analysis. For a comprehensive interpretation of the variant effects, I have utilized multiple levels of evidence, in addition to the standard ACMG criteria[12], focusing on:

1) the phenotype as functional effect read-out;
2) protein structure analysis for variant functional effect prediction;
3) DNAm signature analysis.

Phenotype as functional effect read-out of a genetic variant

Genetic variants with similar functional (d)effects result in similar phenotypes[17-19]. Therefore, for clinically recognizable Mendelian NDDs, the **phenotype of an individual can be utilized as a functional read-out of a variant effect**[18,20] **and the subsequent mechanism of a disorder**. For example, individuals with deletions of the *EHMT1* gene have the same phenotypic features as truncating variants within *EHMT1*[9], confirming the same functional effect of both variants and leading to loss of protein functions. In **chapter 2**, striking phenotypic differences between the individuals with truncating variants located proximal and distally in *SRCAP* (condition now named DEHMBA) in comparison to the well-known Floating-Harbor syndrome caused by truncating variants in the specific FLHS locus in the *SRCAP* gene[21] led to the identification of a novel NDD and shed light on the different functional effects of the truncating variants, based on their position within *SRCAP*. In **chapter 3**, we showed that individuals with PAVs disrupting the enzymatic activity of the KDM6B JmJC domain have the same phenotype as individuals with truncating variants, suggesting that the loss of JmJC enzymatic activity drives the phenotypic feature development of NEDCFSA. By contrast, in **chapter 5**, we showed that the loss of only EHMT1 enzymatic "writer" activity - does not result in typical KLEFS1, while the loss of "reader" activity results in a milder phenotypic spectrum of KLEFS1. This suggests that the typical KLEFS1 phenotype is a result of the loss of multiple EHMT1 functions, both enzymatic and nonenzymatic. Finally, in **chapter 6**, we observed that *EHMT1* duplications do not mirror the phenotype of *EHMT1* deletions, as has been shown for some deletion/duplication syndromes[22], but instead result in mild nonsyndromic NDD[23], suggesting that *EHMT1* overexpression

has different functional effects on the phenotype. While functional tests typically read-out effects on a certain protein function, we show that the phenotype can be used in addition to functional tests to assess the effects of a variant agnostically.

Furthermore, how well the phenotype fit a disorder can be objectively evaluated independently on clinical expertise with the disorder based on facial photos only (e.g., using DeepGestalt[24]) or by incorporating the clinical phenotype with facial photos (e.g., the PhenoScore tool[18]). By using PhenoScore and similar tools, we were able to objectively prove phenotypic differences between disorders in **chapter 2** (DEHMBA vs. FLHS), **chapter 4** (*KMT2C*-related NDD vs. Kleefstra and Kabuki syndromes), and between genotypic groups of the same disorder in **chapter 5** (EHMT1 ANKR-domain PAVs vs. truncating variants).

Protein structure analysis for functional effect prediction of a genetic variant

Protein structure analysis is a powerful approach that has been used in research for many decades for variant effect prediction and/or hypothesis generation[25-29]. The possibility and quality of such predictions depend on the available protein structures and their quality[30]. However, the recently developed AlphaFold2[31] tool allows the generation of high-quality protein models for the majority of known human proteins, and AlphaFold 3 takes it further with protein-protein and protein-DNA or RNA complex predictions[32], increasing the number of proteins and protein complexes available for variant interpretation[33,34]. Therefore, prediction of the variant effects on protein 3D structure 1) has proven its utility in numerous research projects and 2) is now available for the majority of proteins but still is not included in routine variant interpretation in diagnostics. A study from Exeter was the first to show the clinical utility of protein structural analysis in the diagnostic laboratory, which aided the reclassification of approximately half of all evaluated "hot" VUS[35]. Additionally, protein structure analysis could also help understand variant effects even for a variant already functionally tested, e.g., by saturation genome editing[36] and could complement and verify such large-scale tests or provide evidence for discordant classification between the saturation genome editing and genetic evidence[37].

Protein structure analysis was used to evaluate variants in all the genes studied in this thesis. Importantly, in **chapter 5**, we identified a subgroup of missense variants within the EHMT1 SET domain that were predicted to disrupt enzymatic activity but were classified negatively by the Kleefstra syndrome DNA methylation signature analysis. In comparison, other SET domain missense variants were classified positively by the signature. Further comparison of such "negative" and "positive"

SET domain missense variants showed that they result in different phenotypes, as well as have different predicted effects on the domain: "negative" missense variants were predicted to disrupt enzymatic activity only, whereas "positive" variants disrupt the structure of the whole domain. Driven by the structural predictions, we were able to confirm these effects using *in vitro* functional assays and shed light on the pathomechanisms causing Kleefstra syndrome.

ACMG variant interpretation guidelines do not specifically incorporate protein structure predictions as evidence for variant interpretation[12], but the ACGS 2024 criteria suggest that such information can be used as supporting pathogenic evidence[38]. I argue that **protein structure analysis should be included in the routine variant interpretation workflow** and could reduce uncertainty and the number of VUS, as well as identify novel disease subgroups.

Moonlighting by histone-modifying enzymes

Histone-modifying enzymes are known to perform specific histone modifications, e.g., KMT2C is known to methyltransferase H3K4 and EHMT1 – H3K9[39]. However, to perform their function, these proteins are typically multitasking (moonlighting), at least by binding to other proteins to access the chromatin at the right place and time, but little is known about the nonenzymatic functions of these proteins[40]. Understanding proteins and their domain functions is crucial to understanding disease pathophysiology, variant effects, and potential treatment development[26,30,32,35].

Histone-modifying enzyme functions are typically studied by knocking out the whole protein and/or deleting parts of it[40,41]. However, this way, all or multiple functions can be disrupted and, as a result, a "downstream" effect could be misjudged as the main effect of such knock-out[41]. For example, deleting the EHMT1 (pre-)SET methyltransferase domain leads to loss of H3K9 methylation, suggesting a role for this domain in H3K9 methylation[42]. However, the EHMT1 (pre-) SET domain is also involved in forming a heterodimer with EHMT2, which, in fact, performs H3K9 methylation, and EHMT1 itself is dispensable for this function[42-44]. I argue that **knocking out a protein to study its (dys)function provides only a partial landscape and/or can lead to the identification of "downstream" effects from a cascade of events.** Similarly, studying only one function, as often done for saturation genome editing could miss important functional (d)effects of a variant[37]. I propose that **variants identified among affected individuals (as well as somatic mutations enriched in cancers) have a high probability of being functional, so by comprehensively studying these variants alongside the whole protein or its part deletions could help pinpoint specific protein functions and determine**

variant effects[45]. To compare with other rare diseases, only in-depth analysis of individuals with inborn errors of immunity helped identify and confirm different immune cell subtypes (T and B lymphocytes) and their functions[46].

In **chapter 5** we identified individuals with NDD likely due to the EHMT1 SET domain's methyltransferase function being disrupted by a PAV (while maintaining its capacity to bind to EHMT2) but lacking the typical Kleefstra syndrome phenotype and DNAm signature. Other SET domain PAVs caused typical Kleefstra syndrome by disrupting both enzymatic and EHMT2-binding functions. Similarly, PAVs disrupting the ANKR domain and/or its ability to bind to the H3 tail also resulted in Kleefstra syndrome. Based on our findings and previous literature[42,44,47], we speculate that loss of EHMT1 enzymatic activity causes NDD by disrupting nonhistone protein methylation, but this requires further investigation.

Similarly, in **chapter 3** we showed a significant clustering of *de novo* PAVs in the C-terminal part of *KDM6B*. These PAVs were predicted to disrupt the KDM6B enzymatic domain JmJC, the Zn-containing domain required for H3-tail binding, and the stabilizing linkers required for the structural integrity of the whole C-terminal part[48]. However, a disruptive effect in the Drosophila dual gain-of-function assay was only shown for PAV disrupting *KDM6B* enzymatic activity (including binding ability to the H3 tail required for demethylation) but not for variants disrupting stabilizing linkers. Clustering of *de novo* PAVs and local intolerance in the general population in these linkers suggests that these PAVs disrupt an alternative yet unknown KDM6B function (e.g., binding to other proteins).

DNA methylation signatures

Why are DNAm signatures present?

The mechanism responsible for DNAm signatures is largely unknown, but they are expected to result from disrupted regulation of the epigenetic machinery[49,50]. Two interplay mechanisms between the epigenetic machinery and DNA methylation are proposed: 1) direct recruitment of DNMT or TET enzymes by epigenetic machinery proteins to regions of interest, or 2) activation and recruitment (or repression) of DNMT enzymes due to specific histone modifications or the chromatin state[51]. Additionally, we cannot exclude the presence of DNAm signatures due to "downstream" changes in cellular processes due to a dysfunction of a certain protein. Based on the functional variety of the discovered DNAm signatures[52], I speculate that for different genes, different mechanisms (or combinations of those) are likely at play.

Currently, most of the signatures are derived for disorders of the epigenetic machinery[49,50,53]. However, there are known DNAm signatures for the genes not directly known to be a part of the epigenetic machinery, transcription factors being the largest group (e.g., *YY1, ANKRD11, SOX11, ZNF711, SIN3A,* etc.)[53]. However, it is known that transcription factors and other DNA-binding proteins could interact with the epigenetic machinery and promote both DNA methylation and histone code changes[54,55] for the precise regulation of gene expression. DNAm signatures have also been recently described for splicing-regulating proteins like *PQBP1*[56] and cell-cycle regulating proteins like *CDK13* and *CCNK*[57]. While these proteins are not known to directly interact with DNA or chromatin, these proteins form quaternary complexes with other proteins that could regulate the chromatin state (e.g., both PQBP1 and CDK13 can bind to RNA polymerase II[58,59], which is known to modify chromatin during transcription[60,61]. Finally, there are also described signatures for genes without any known association with chromatin (e.g., SMS encoding spermine synthase or *SLC32A1* encoding a solute carrier protein) suggesting a complex downstream cellular dysfunction causing DNAm signatures[53]. However, moonlighting of such proteins with a direct epigenetic regulator function cannot be excluded.

In contrast, no signatures have been possible to identify (yet) for some of the disorders with a clear association with epigenetic regulation, including *MECP2* (associated with Rett syndrome)[62] or *KDM6B*[62] (associated with *KDM6B*-related NDD [**chapter 3**]). Similarly, haploinsufficiency of *KMT2C*, a very important epigenetic regulator, unexpectedly resulted in only a moderate-effect DNAm signature that we were able to identify only with a relatively large sample size (**chapter 4**).

In summary, this suggests that epigenetic regulation and interplay are extremely complex, with a myriad of involved proteins (directly or indirectly), and, therefore, the range of Mendelian errors of the epigenetic machinery is much larger than we had anticipated. Additionally, this suggests that DNAm signatures could also be identified for many genes without a direct association with the epigenetic machinery.

Why are DNAm signatures present in the way they are?

In a single cell, DNA methylation on a CpG site can occur on a single allele (50% methylated), on both alleles (100% methylated), or neither allele (0%) (**Figure 7 top panel**). However, when analyzing DNA derived from a bulk of many thousand cells and different cell types (e.g., blood), the average DNA methylation of all cells and all cell types is evaluated. A typical DNAm signature consists of differentially methylated CpG sites with the methylation difference usually being less than

20%[49,52]. Therefore, a DNA methylation level of ~20% could represent methylation of a single allele in 40% of cells or methylation of both alleles in 20% of cells in a sample, as these changes could be isolated to a specific blood lineage or present in all blood cells in approximately equal levels (**Figure 1 bottom panel**). For example, age-related clock DNAm changes are largely associated with the activation of T and NK cells[63]. However, the distribution of methylated CpGs across cells and tissues for DNAm signatures is largely unknown and, in theory, could be different for different conditions and signatures (e.g., depending on which ontogenesis point the gene-of-interest is active and regulates chromatin modification).

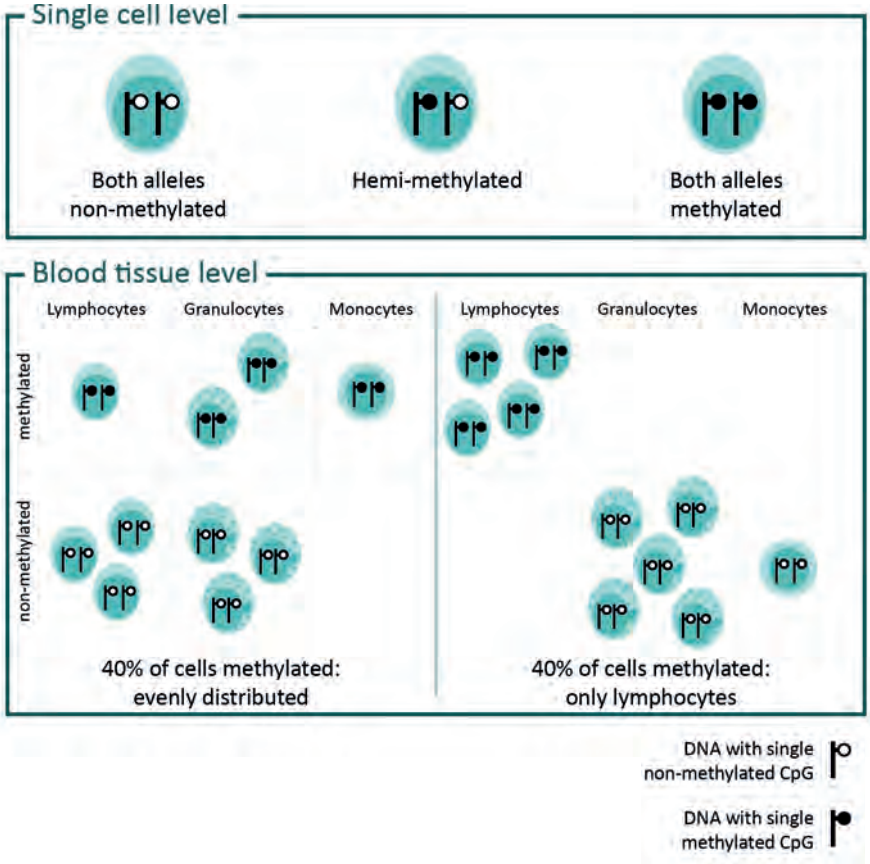


Figure 1. Possible representation of CpG differential methylation in single cell and blood.

The timing of DNAm signature occurrence is also unknown. Almost all DNA methylation in the genome is erased during embryonic development and established anew during every ontogenesis[64]. Further, DNA methylation at CpG

sites is maintained over cell divisions by DNMT1[65]. Therefore, it can be assumed that the DNA methylation signatures already occur during embryonic development when epigenetic regulators dynamically regulate gene expression, and so they are present across different cell lineages[66,67]. Alternatively, during differentiation, cells establish a cell-specific gene expression profile and must react to environmental stimuli, which also require a change in DNA methylation in the cell's genome (and, therefore, signature generation) later in ontogenesis[64]. For example, many epigenetic regulators (e.g., COMPASS complex genes like *KMT2C, KMT2D, SETD1A, KDM6B, KDM6A,* as well as *EHMT1,* etc.) are required for dynamic activation of immune responses in leukocytes, and loss of these proteins can result in immune dysregulation/deficiency[68,69]. I speculate that both these **early embryonic and later ontogenesis points in time contribute to the DNAm signature generation that we observe in blood** and represent early, as well as late, cellular processes. This would also mean that different tissues only partially share the signature. This is supported by the fact that DNAm signatures derived from blood fail to correctly classify DNA extracted from other tissues when compared to blood-derived DNAs but could allow the grouping of cases vs. controls within a single tissue[67,70].

DNAm signature applications

Currently, the DNAm signatures are derived and tested mostly on methylation arrays using blood-derived DNA, as it is widely available and stable over time[49,50]. Usually, leftovers from DNA testing are used, so it does not require additional sampling or invasive procedures. The DNA samples are bisulfite-converted, which allows the identification of CpG methylation levels (without discriminating 5mC and 5hmC)[50,71].

Since the discovery of DNAm signatures, their use for VUS interpretation has become widespread [50,72,73]. However, it should be noted that **DNAm signatures can be applied for multiple purposes**:

1) Affected individual testing;
2) To delineate and compare conditions;
3) To study the underlying biology of a disorder.

DNAm signatures for VUS testing: applications

In **chapters 2, 4,** and **5,** we actively used the respective signatures to robustly classify multiple VUS in the *SRCAP*, *KMT2C*, and *EHMT1* genes, respectively. While the DNAm signatures allowed us to simultaneously and quickly test dozens of variants and reduce the uncertainty of VUS, we have also encountered multiple challenges.

Firstly, the DNAm signature effect size, as well as the sensitivity and specificity, can vary greatly between conditions[49,52]. For example, Sotos syndrome has a signature that affects most of the genome and is highly sensitive and specific[67,74]; BAF-complex genes have a shared signature, so individual genes have low specificity[53]; *KMT2C*-related NDD or *EHMT1* duplication-associated DNAm changes are of moderate effect (**chapter 4** and **6**, respectively) and have classified several individuals with pathogenic variants as "negative," so they have lower sensitivity. Interestingly, it is unknown if such false-negative cases represent technically or biologically relevant features. It has been observed that for some conditions, individuals with a milder phenotype could have less prominent DNA methylation changes[66,75,76]. Both false negatively classified *KMT2C*-related NDD individuals in **chapter 4** are mildly affected, so we speculate that this classification, if true, could represent incomplete penetrance for the DNAm signature, but requires further validation.

Secondly, DNAm signature sensitivity for PAV testing is unknown, because signatures are typically derived and validated on one variant type (e.g., truncating variants) and then tested on different ones (e.g., PAVs)[53]. A PAV (or a truncating variant escaping NMD) could target a single domain's function and have a different effect than the loss of the whole protein[50]. Therefore, "negative" classification by a signature does not exclude variant pathogenicity and should not be used to classify a variant as benign[12,73]. We showed in **chapter 5** that pathogenic *EHMT1* variants with functional effects other than haploinsufficiency have a "negative" Kleefstra syndrome DNAm signature. We also highlight that the actual cause of a signature for a disorder (loss of a single or multiple functions of a protein) is unknown (**Figure 2**). Similarly, due to signature overlap between different conditions, a "positive" result could reflect a pathogenic variant in a different gene rather than a tested VUS and does not automatically mean that the tested VUS is pathogenic[77]. Additionally, some technical biases could also result in false positive, e.g., in **chapter 4,** we have encountered that the DNA of newborns has a different DNA methylation profile[78] that could produce false-positive results. Therefore, **DNAm signature results should never be used in isolation to classify a variant but rather be used as a piece of evidence within the variant classification framework**.

Finally, there are currently no accepted standards or guidelines for applications of DNAm signatures for variant interpretation, and the field must agree on the DNAm signature validation and application standards within the variant interpretation workflow. Depending on the disorder, the respective signature, and how thoroughly it has been validated (including on different variant types and their locations within

the gene), different evidence levels could be applied to the DNAm signature results: from supporting to moderate or even strong[16,38].

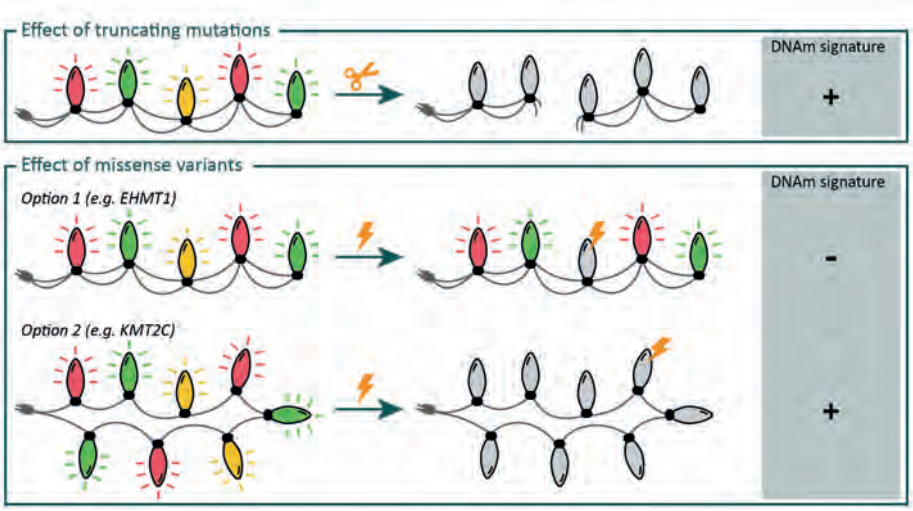


Figure 2. Possible predicted outcomes of protein altering variants on protein function and DNAm signature presence.

*Epigenetic machinery proteins (represented by Christmas lights) commonly consist of multiple domains (represented by lightbulbs) connected via flexible disordered regions (represented as cords). Protein truncating variants (represented by scissors) disrupt the whole protein with all its domains, resulting in the presence of a haploinsufficiency-associated DNAm signature (top panel). Missense variants, however, (represented by lightning) could have different effects on a protein, likely depending on the protein structure and functions (bottom panel). For example, if disruption of a single domain results in disruption of all the protein's domains and functions, the same effects are expected as per haploinsufficiency, including the presence of the DNAm signature (**Option 2**). However, if disruption of a single domain preserves the functions of the rest of the domains, the expected effects are likely different from haploinsufficiency, including the absence of a haploinsufficiency DNAm signature (**Option 1**). This way, pathogenic disruptive missense or other protein altering variants could be misclassified as controls using DNAm classifiers.*

DNAm signature testing: functional test or phenotyping

DNAm signatures represent the result of gene product dysfunction. Therefore, some argue that the signatures can be used as a functional test for a variant[53]. However, signatures test the function of a gene product, rather than the variant itself, and one can identify a signature in individuals without a known genetic variant[62,73]. Additionally, only DNMT and TET proteins are directly involved in DNA methylation[65], so the other DNAm signature-associated genes are involved in DNA methylation changes indirectly[51]. Therefore, I argue that **DNAm signatures are rather a part of the MDEM (molecular) phenotype** in the same way that hyperphenylalaninemia is a part of the phenylketonuria phenotype. Guidelines suggest that "Functional evidence

from patient-derived material best reflects the organismal phenotype" and does not test a single variant[16].

This distinction has important consequences for the clinical applications of the signatures. First, a positive DNAm signature can confirm a clinical diagnosis, but does not necessarily confirm the pathogenicity of a variant itself. Next, if signatures are considered as a functional test, they could be used to apply functional evidence of the pathogenicity of a variant (PS3) or of the variant being benign (BS3)[12,16,38]. However, if the signatures are considered as the phenotype, a "positive" signature would satisfy the application of phenotype specificity criteria (PP4), but a "negative" signature result would not provide evidence for benign variant classification[12]. Additionally, the clinical phenotype could be used together with the DNAm signature results to evaluate the specificity of the phenotype and to modify the criterion strength accordingly[38].

DNAm signatures: to delineate and compare NDDs

The majority of recently described NDDs do not have a specific phenotype and are not clinically recognizable[2]. Therefore, additional evidence is required to prove causality of variants within a single gene to a given novel NDD[7,79]. **The presence of a shared DNAm signature is proof of a shared specific molecular (epi-)phenotype.** In **chapter 2**, we failed to identify clinical or facial features specific to the "proximal" *SRCAP* truncating variants - either clinically or by using the facial feature analysis tool. However, we have utilized a DNAm signature to show that truncating *SRCAP* variants outside of the known FLHS locus (both "proximal" and "distal") could result in the same NDD with a specific DNAm signature. Similarly, in **chapter 6**, we show that 9q34 duplications result in a nonsyndromic, nonspecific NDD with common DNAm changes. In **chapter 4**, we showed that individuals with truncating *KMT2C* variants share the same signature irrespective of their location within the gene or transcript, despite some variants being present only in a longer and less expressed transcript. This allowed us to conclude that haploinsufficiency of the longest (canonical) transcript causes the phenotype despite low expression of the transcript in tissues at an adult age and aids variant interpretation irrespective of the location or transcript affected.

Additionally, DNAm signatures could help to differentiate between different conditions within the same gene or different genes (or the contrary – define them into a single NDD). Genes with similar functions or pathways have similar (or identical) DNAm signatures[52,66]. For example, *KMT2D* and *KDM6A* are both associated with Kabuki syndrome[80]. Molecularly, the signature of Kabuki syndrome is similar

irrespective of the causative gene[50,81], showing that these two conditions are related not only clinically but also molecularly, despite having opposite enzymatic functions (KMT2D is a histone-3 methylase and KDM6A is a demethylase). In contrast, in **chapter 2**, we showed that FLHS and "proximal" *SRCAP* have different, but partially overlapping DNAm signatures. This shows that the two signatures are related (caused by dysfunction of the same gene), but highlights the differences between the two disorders. Similarly, in **chapter 4**, we utilized the DNAm signatures to convincingly prove that *KMT2C*-related NDD (previously described as Kleefstra syndrome 2) is distinct not only clinically, but also molecularly from Kleefstra syndrome type 1 and, therefore, should be renamed to avoid further confusions between the two conditions.

Finally, DNAm signature analysis could help to identify disease subgroups. Helsmoortel-van der Aa syndrome is considered a single clinical condition with two different DNAm signatures depending on the variant location within the gene[82,83]. However, only a recent analysis using a novel AI tool (PhenoScore) allowed the differentiation of phenotypic differences associated with the two signatures[18]. In **chapter 5**, we identified that N-terminal truncating *EHMT1* variants result in a mild phenotype different from classical Kleefstra syndrome and showed that such individuals also do not have a classical Kleefstra syndrome signature, confirming different effects of these variants on the molecular level, similar to those on the phenotype level.

DNAm signatures: studying the underlying biology of an NDD

DNAm signatures also represent a biological "read-out" of certain gene dysfunctions. For example, DNMT3A performs CpG methylation[65], so *DNTM3A* haploinsufficiency results in a large, hypomethylated DNAm signature[84]; *TET3* is involved in CpG demethylation[65,85], so *TET3* loss expectedly results in a hypermethylated signature[86].

EHMT1 is known to recruit DNMT3A to H3K9me sites for DNA methylation[43]. In **chapter 5**, the loss of EHMT1 or its function of binding to the H3 tail produces a DNAm signature that is largely hypomethylated[52,62,87], likely representing an inability to recruit DNMT3A. By contrast, loss of EHMT1 enzymatic function (while preserving the ability to bind the H3 tail and EHMT2) does not result in the Kleefstra syndrome DNAm signature, likely representing the preserved ability to recruit DNMT3A and that EHMT1 enzymatic activity is not involved in signature generation.

In **chapter 2**, we showed that analysis of the involved CpG sites also reflects the functions of the gene. For example, the differentially methylated CpGs map to genes related to SRCAP function, and gene ontology analysis showed "DNA recombination" and "DNA repair", which are known functions of the SRCAP protein[88,89], as one of the most significant biological processes for the "proximal" *SRCAP* DNAm signature sites. Interestingly, in **chapter 4**, we found that ~10% of *KMT2C*-related NDD DNAm signature map to a differentially methylated region within *WT1* gene CpG islands, suggesting epigenetic interaction between *KMT2C* and *WT1*, which has not been described previously. WT1 is known to be important for correct neurodevelopment[90-92], so this epigenetic interplay could also be related to the phenotypic features of the disorder and warrants further studies.

Future perspectives

Importance of data sharing

Rare disease research requires proactive collaboration among multiple centers to collect enough data about a condition[93,94]. The GeneMatcher platform[95], developed over a decade ago, allowed the identification and description of hundreds of novel disorders[96] simply by connecting physicians and/or researchers to exchange information about patients. Similarly, the Human Disease Genes website series allows the collection of data on individuals for known disorders[97]. However, these platforms require proactive participation and registration and would require widespread adaptation of both research and clinical institutes to allow the collection of large, rare disease cohorts. Alternatively, such databases as ClinVar[98] or the Dutch Clinical Laboratory Collaboration[99] allow public sharing of genetic variant data in a nonidentifiable way but do not provide any clinical information. Currently, the collection of large cohorts is extremely time-consuming and includes a search of in-house-diagnosed individuals; contact initiation via GeneMatcher, ClinVar, VKGL, and similar databases; and a search of the literature, as well as personal contacts by indviduals' relatives or physicians seeking a consult with an expertise centrum. For example, 15 out of 44 *EHMT1* VUS (34%) and PAVs described in **chapter 5** and 8 out of 10 families with *EHMT1* duplications (80%) in **chapter 6** were collected via personal communication with physicians seeking a consult with Prof. T. Kleefstra at the Kleefstra syndrome expertise centrum.

Additionally, the sharing of bioinformatic pipelines and tools is necessary to obtain reproducible results across different laboratories[100], which is crucial for independent validations of the results and standardization of test procedures[101]. Currently, the

DNAm signature analysis is provided via the EpiSign test[72] or the open-access platforms EpigenCentral[102,103] and Heidelberg Molecular Neuropathology DNAm-based classifier of tumors[104]. However, these platforms do not share the pipeline and data used to generate the classifiers and are either entirely closed or provide only limited interaction. An independent evaluation showed low sensitivity and unstable performance of multiple published signatures, highlighting the importance of data and pipeline sharing and independent validation of the results to provide a reliable molecular test in diagnostic settings[74].

Widespread application of data sharing (genetic and clinical) that minimizes private information misuse must be adopted by the human genetics community to understand and provide detailed information for all ultrarare genetic disorders, as well as provide data for the development of next-generation precise AI tools[105,106]. This would require the development of novel genetic data sharing, storing, and encoding approaches and broad discussions among different stakeholders. Otherwise, many individuals with rare disorders will likely remain underdiagnosed and understudied, without personalized care.

Mendelian NDDs: Next developments in diagnostics

Currently, only ~30%–50% of individuals with (suspected) Mendelian NDD are "solved" using standard genomic testing (exome or genome sequencing)[8,107,108]. As the next step, multiple omics technologies (e.g., transcriptome, methylome, metabolome) are being applied in the hope of pinpointing the molecular cause of a disorder[109-113]. In addition, continuously improving genome sequencing methods are being applied, aiming to identify genetic variants missed by a previous technique (e.g., exome sequencing, followed by improved exome sequencing[8], followed by short-read genome sequencing[108], followed by long-read genome sequencing[114,115]). While both approaches can provide a genetic diagnosis for previously unsolved cases and identify novel molecular mechanisms, most individuals remain unsolved. This suggests that the main limiting factor is our inability to interpret genomic data, rather than missing causative genetic variants in the data. Therefore, I believe that **improving our understanding of variant effects and variant interpretation will be one of the main challenges in clinical genetics for the next decade,** requiring novel tools, techniques, and skills.

The "third generation" of sequencing ("next-next-generation sequencing") in the last years not only improved drastically in sequencing quality, but also decreased in price[116]. "Third-generation" sequencing can already provide high-quality, long-read genome sequencing covering regions and identifying variants inaccessible by short-read genome sequencing[117,118]. Importantly, both PacBio and Oxford

Nanopore "third-generation" long-read sequencing provides DNA methylation calls simultaneously with DNA sequencing to identify genetic as well as "epigenetic" disorders (e.g., imprinting disorders like Prader-Willi syndrome) in the same test[112,117]. Additionally, this could provide opportunities to perform DNAm signature analysis from genomic data right away, reducing the uncertainty of VUS or identifying nongenetic disorders (like fetal alcohol syndrome[119] or exposure to some other toxic/terratogenic agents). Moreover, CpG and non-CpG site methylation and 5hmC calling in addition to 5mC (e.g., provided by nanopore sequencing) could help to identify novel signatures or improve the accuracy of existing ones. Therefore, **I argue that "third-generation" sequencing with further improvements in the techniques, bioinformatic tools, and decreasing costs will become the first and last all-in-one test an individual with a suspected (epi-) genetic disorder would need**, requiring only reinterpretation of the data, as our knowledge grows, until the case is solved. However, this will also bring new challenges, e.g., the protection and use of private data, because, in addition to genetic data, it will contain environmental or acquired condition data encoded within DNAm (e.g., smoking[120] or certain drug exposure[121]).

Independently, the discovery of novel disease mechanisms and subsequent development of new (artificial intelligence-powered) tools is required to make sense of the genomic data on both the variant level, and on the individual level. On the variant level, the splice-effect prediction tool SpliceAI[122] has already proven to be highly accurate in predicting the effects of variants (including noncoding ones) on splicing[123] and is currently being incorporated into clinical diagnostics, as well as guidelines[124]. Similarly, AlphaFold2/3 can allow the development of new tools for PAV interpretation (as AlphaMissense[125]), based on 3D structure effects. However, such accurate tools for noncoding variants that do not affect splicing are unfortunately missing[126]. On the individual level, an objective way to identify or confirm a clinical diagnosis is crucial for correct variant interpretation[12]. Currently, different tools are available, e.g., PhenoScore allows a case to be classified based on HPO and/or facial features[18], and DNA methylation signatures allow classification, based on DNA methylation changes[53]. However, both tools only work on a handful of syndromic NDDs. In **chapter 4**, we showed the utility of both tools for delineating and differentiating *KMT2C*-related NDD from Kleefstra and Kabuki syndromes, but limited individual number limited the accuracy of the tools and our ability to classify all identified *KMT2C* variants. Further development of such AI-powered tools will be largely dependent on the availability of large "truth datasets"[127,128] and, therefore, will require widespread data sharing by scientific and clinical (and patient) communities to provide sufficient training data from rare disorders.

Implications for treatment

Improved diagnostics and understanding of Mendelian disorder pathogenesis resulted in the first successful orphan treatment options for patients[129,130]. Currently, the IRDiRC has the ambitious goal of having 1000 novel therapies approved for rare diseases within a decade[131,132]! Further, an understanding of natural disease history (including that of different variant types, where relevant) and the availability of objectively measurable, clinically relevant endpoints (or biomarkers of such) would be crucial to proving the effect of a novel treatment[133].

For example, a truncating variant leading to NMD that completely disrupts protein function could be avoided by skipping one or multiple exons and preserving the reading frame with antisense oligonucleotides[134]. However, if such skipping were to produce a protein lacking an important domain, the effect of such therapy would be questionable, so it is crucial to understand the critical regions of a gene or protein. As an example, SRCAP, KDM6B, KMT2C, and EHMT1 proteins have large, disordered regions with no known function. In **chapters 2, 3, 4, and 5**, respectively, we showed that variants disrupting a single functional domain (or region) of these proteins could result in a phenotype comparable to haploinsufficiency, whereas variants occurring in disordered regions largely have no effect. In **chapter 3**, we described a benign inherited *KMT2C* canonical splice variant c.1735+2dup that is predicted to result in in-frame exon skipping that encodes a disordered protein region. This suggests that only skipping an exon that encodes a disordered region with no known function could, in theory, restore the functions of these proteins.

Another approach that has been attempted for Kabuki syndrome 1 (caused by haploinsufficiency of H3K4 methyltransferase *KMT2D*)[135]: pharmacological inhibition of KDM1A, an H3K4 demethylase, is suggested to counterbalance the effects of loss of H3K3 methylation. While this approach is more universal and is not influenced by the variant type, such an intervention would only counterbalance a loss of the histone-modifying function of a protein. However, multiple histone modifiers have been described as moonlighting proteins with nonenzymatic and non-histone modifying functions[136-138], so the success of such therapies would depend on which functions of the protein are crucial and drive phenotype development, and which functions are dispensable. For example, in **chapter 5**, we show that loss of EHMT1 enzymatic activity alone does not result in typical KLEFS1 but is still a likely cause of severe NDD, possibly via loss of non-histone modifying activity, suggesting the importance of EHMT1 moonlighting. Therefore, counterbalancing the loss of H3K9 methyltransferase activity would not address an important pathomechanism of EHMT1 loss. To conclude, **an in-depth understanding of disease mechanisms**

is required for the development of a novel therapy option[139], and analyses of large cohorts can provide the crucial information needed to offer an appropriate option.

References

1. https://www.ebi.ac.uk/gene2phenotype. (2024).
2. Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., et al. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature *586*, 757-762. 10.1038/s41586-020-2832-5.
3. Lelieveld, S.H., Reijnders, M.R., Pfundt, R., Yntema, H.G., Kamsteeg, E.J., de Vries, P., de Vries, B.B., Willemsen, M.H., Kleefstra, T., Lohner, K., et al. (2016). Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat Neurosci *19*, 1194-1196. 10.1038/nn.4352.
4. Kleefstra, T., Kramer, J.M., Neveling, K., Willemsen, M.H., Koemans, T.S., Vissers, L.E., Wissink-Lindhout, W., Fenckova, M., van den Akker, W.M., Kasri, N.N., et al. (2012). Disruption of an EHMT1-associated chromatin-modification module causes intellectual disability. Am J Hum Genet *91*, 73-82. 10.1016/j.ajhg.2012.05.003.
5. Koemans, T.S., Kleefstra, T., Chubak, M.C., Stone, M.H., Reijnders, M.R.F., de Munnik, S., Willemsen, M.H., Fenckova, M., Stumpel, C., Bok, L.A., et al. (2017). Functional convergence of histone methyltransferases EHMT1 and KMT2C involved in intellectual disability and autism spectrum disorder. PLoS Genet *13*, e1006864. 10.1371/journal.pgen.1006864.
6. Vissers, L.E., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., et al. (2010). A de novo paradigm for mental retardation. Nat Genet *42*, 1109-1112. 10.1038/ng.712.
7. Vissers, L.E., Gilissen, C., and Veltman, J.A. (2016). Genetic studies in intellectual disability and related disorders. Nat Rev Genet *17*, 9-18. 10.1038/nrg3999.
8. Schobers, G., Schieving, J.H., Yntema, H.G., Pennings, M., Pfundt, R., Derks, R., Hofste, T., de Wijs, I., Wieskamp, N., van den Heuvel, S., et al. (2022). Reanalysis of exome negative patients with rare disease: a pragmatic workflow for diagnostic applications. Genome Med *14*, 66. 10.1186/s13073-022-01069-z.
9. Willemsen, M.H., Vulto-van Silfhout, A.T., Nillesen, W.M., Wissink-Lindhout, W.M., van Bokhoven, H., Philip, N., Berry-Kravis, E.M., Kini, U., van Ravenswaaij-Arts, C.M., Delle Chiaie, B., et al. (2012). Update on Kleefstra Syndrome. Mol Syndromol *2*, 202-212. 10.1159/000335648.
10. van der Spek, J., den Hoed, J., Snijders Blok, L., Dingemans, A.J.M., Schijven, D., Nellaker, C., Venselaar, H., Astuti, G.D.N., Barakat, T.S., Bebin, E.M., et al. (2022). Inherited variants in CHD3 show variable expressivity in Snijders Blok-Campeau syndrome. Genet Med *24*, 1283-1296. 10.1016/j.gim.2022.02.014.
11. van der Sluijs, P.J., Alders, M., Dingemans, A.J.M., Parbhoo, K., van Bon, B.W., Dempsey, J.C., Doherty, D., den Dunnen, J.T., Gerkes, E.H., Millller, I.M., et al. (2021). A Case Series of Familial ARID1B Variants Illustrating Variable Expression and Suggestions to Update the ACMG Criteria. Genes (Basel) *12*. 10.3390/genes12081275.
12. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). 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 *17*, 405-424. 10.1038/gim.2015.30.
13. Matthijs, G., Souche, E., Alders, M., Corveleyn, A., Eck, S., Feenstra, I., Race, V., Sistermans, E., Sturm, M., Weiss, M., et al. (2016). Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet *24*, 2-5. 10.1038/ejhg.2015.226.
14. Rodenburg, R.J. (2018). The functional genomics laboratory: functional validation of genetic variants. J Inherit Metab Dis *41*, 297-307. 10.1007/s10545-018-0146-7.

15. Maia, N., Nabais Sá, M.J., Melo-Pires, M., de Brouwer, A.P.M., and Jorge, P. (2021). Intellectual disability genomics: current state, pitfalls and future challenges. BMC Genomics *22*, 909. 10.1186/s12864-021-08227-4.

16. Brnich, S.E., Abou Tayoun, A.N., Couch, F.J., Cutting, G.R., Greenblatt, M.S., Heinen, C.D., Kanavy, D.M., Luo, X., McNulty, S.M., Starita, L.M., et al. (2019). Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med *12*, 3. 10.1186/s13073-019-0690-2.

17. Galer, P.D., Ganesan, S., Lewis-Smith, D., McKeown, S.E., Pendziwiat, M., Helbig, K.L., Ellis, C.A., Rademacher, A., Smith, L., Poduri, A., et al. (2020). Semantic Similarity Analysis Reveals Robust Gene-Disease Relationships in Developmental and Epileptic Encephalopathies. Am J Hum Genet *107*, 683-697. 10.1016/j.ajhg.2020.08.003.

18. Dingemans, A.J.M., Hinne, M., Truijen, K.M.G., Goltstein, L., van Reeuwijk, J., de Leeuw, N., Schuurs-Hoeijmakers, J., Pfundt, R., Diets, I.J., den Hoed, J., et al. (2023). PhenoScore quantifies phenotypic variation for rare genetic diseases by combining facial analysis with other clinical features using a machine-learning framework. Nature genetics *55*, 1598-1607. 10.1038/s41588-023-01469-w.

19. van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G., and Leunissen, J.A.M. (2006). A text-mining analysis of the human phenome. European Journal of Human Genetics *14*, 535-542. 10.1038/sj.ejhg.5201585.

20. Rosina, E., Pezzani, L., Pezzoli, L., Marchetti, D., Bellini, M., Pilotta, A., Calabrese, O., Nicastro, E., Cirillo, F., Cereda, A., et al. (2022). Atypical, Composite, or Blended Phenotypes: How Different Molecular Mechanisms Could Associate in Double-Diagnosed Patients. Genes (Basel) *13*. 10.3390/genes13071275.

21. Hood, R.L., Lines, M.A., Nikkel, S.M., Schwartzentruber, J., Beaulieu, C., Nowaczyk, M.J., Allanson, J., Kim, C.A., Wieczorek, D., Moilanen, J.S., et al. (2012). Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. Am J Hum Genet *90*, 308-313. 10.1016/j.ajhg.2011.12.001.

22. Rooney, K., and Sadikovic, B. (2022). DNA Methylation Episignatures in Neurodevelopmental Disorders Associated with Large Structural Copy Number Variants: Clinical Implications. Int J Mol Sci *23*. 10.3390/ijms23147862.

23. Bonati, M.T., Castronovo, C., Sironi, A., Zimbalatti, D., Bestetti, I., Crippa, M., Novelli, A., Loddo, S., Dentici, M.L., Taylor, J., et al. (2019). 9q34.3 microduplications lead to neurodevelopmental disorders through EHMT1 overexpression. Neurogenetics *20*, 145-154. 10.1007/s10048-019-00581-6.

24. Pantel, J.T., Hajjir, N., Danyel, M., Elsner, J., Abad-Perez, A.T., Hansen, P., Mundlos, S., Spielmann, M., Horn, D., Ott, C.E., and Mensah, M.A. (2020). Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals With and Without a Genetic Syndrome: Diagnostic Accuracy Study. J Med Internet Res *22*, e19263. 10.2196/19263.

25. Ittisoponpisan, S., Islam, S.A., Khanna, T., Alhuzimi, E., David, A., and Sternberg, M.J.E. (2019). Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? J Mol Biol *431*, 2197-2212. 10.1016/j.jmb.2019.04.009.

26. Sen, N., Anishchenko, I., Bordin, N., Sillitoe, I., Velankar, S., Baker, D., and Orengo, C. (2022). Characterizing and explaining the impact of disease-associated mutations in proteins without known structures or structural homologs. Brief Bioinform *23*. 10.1093/bib/bbac187.

27. Gerasimavicius, L., Livesey, B.J., and Marsh, J.A. (2022). Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. Nat Commun *13*, 3895. 10.1038/s41467-022-31686-6.

28. Venselaar, H., Camilli, F., Gholizadeh, S., Snelleman, M., Brunner, H.G., and Vriend, G. (2013). Status quo of annotation of human disease variants. BMC Bioinformatics *14*, 352. 10.1186/1471-2105-14-352.

29. Venselaar, H., Te Beek, T.A., Kuipers, R.K., Hekkelman, M.L., and Vriend, G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics *11*, 548. 10.1186/1471-2105-11-548.

30. Pak, M.A., and Ivankov, D.N. (2022). Best templates outperform homology models in predicting the impact of mutations on protein stability. Bioinformatics *38*, 4312-4320. 10.1093/bioinformatics/btac515.

31. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583-589. 10.1038/s41586-021-03819-2.

32. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 10.1038/s41586-024-07487-w.

33. Akdel, M., Pires, D.E.V., Pardo, E.P., Jänes, J., Zalevsky, A.O., Mészáros, B., Bryant, P., Good, L.L., Laskowski, R.A., Pozzati, G., et al. (2022). A structural biology community assessment of AlphaFold2 applications. Nat Struct Mol Biol *29*, 1056-1067. 10.1038/s41594-022-00849-w.

34. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res *50*, D439-d444. 10.1093/nar/gkab1061.

35. Caswell, R.C., Gunning, A.C., Owens, M.M., Ellard, S., and Wright, C.F. (2022). Assessing the clinical utility of protein structural analysis in genomic variant classification: experiences from a diagnostic laboratory. Genome Med *14*, 77. 10.1186/s13073-022-01082-2.

36. Radford, E.J., Tan, H.-K., Andersson, M.H.L., Stephenson, J.D., Gardner, E.J., Ironfield, H., Waters, A.J., Gitterman, D., Lindsay, S., Abascal, F., et al. (2023). Saturation genome editing of DDX3X clarifies pathogenicity of germline and somatic variation. Nature Communications *14*, 7702. 10.1038/s41467-023-43041-4.

37. Kennis, M.G.P., et al. DDX3X-related neurodevelopmental disorder in males – presenting a new cohort of 18 males and a literature review. Submitted.

38. Miranda Durkie, E.-J.C., Ian Berry, Martina Owens, Clare Turnbull, Richard H Scott, Robert W Taylor, Zandra C Deans, Sian Ellard, Emma L Baple, Dominic J McMullan (2024). ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2024. ACGS https://www.acgs.uk.com/quality/best-practice-guidelines.

39. Park, J., Lee, K., Kim, K., and Yi, S.J. (2022). The role of histone modifications: from neurodevelopment to neurodiseases. Signal Transduct Target Ther *7*, 217. 10.1038/s41392-022-01078-9.

40. Morgan, M.A.J., and Shilatifard, A. (2023). Epigenetic moonlighting: Catalytic-independent functions of histone modifiers in regulating transcription. Sci Adv *9*, eadg6593. 10.1126/sciadv.adg6593.

41. Policarpi, C., Munafò, M., Tsagkris, S., Carlini, V., and Hackett, J.A. (2022). Systematic Epigenome Editing Captures the Context-dependent Instructive Function of Chromatin Modifications. bioRxiv, 2022.2009.2004.506519. 10.1101/2022.09.04.506519.

42. Tachibana, M., Matsumura, Y., Fukuda, M., Kimura, H., and Shinkai, Y. (2008). G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription. Embo j *27*, 2681-2690. 10.1038/emboj.2008.192.

43. Chang, Y., Sun, L., Kokura, K., Horton, J.R., Fukuda, M., Espejo, A., Izumi, V., Koomen, J.M., Bedford, M.T., Zhang, X., et al. (2011). MPP8 mediates the interactions between DNA methyltransferase Dnmt3a and H3K9 methyltransferase GLP/G9a. Nat Commun *2*, 533. 10.1038/ncomms1549.

44. Yamada, A., Shimura, C., and Shinkai, Y. (2018). Biochemical validation of EHMT1 missense mutations in Kleefstra syndrome. J Hum Genet *63*, 555-562. 10.1038/s10038-018-0413-3.

45. Proietti Onori, M., and van Woerden, G.M. (2021). Role of calcium/calmodulin-dependent kinase 2 in neurodevelopmental disorders. Brain Res Bull *171*, 209-220. 10.1016/j.brainresbull.2021.03.014.

46. Gitlin, A.D., and Nussenzweig, M.C. (2015). Immunology: Fifty years of B lymphocytes. Nature *517*, 139-141. 10.1038/517139a.

47. Tsusaka, T., Kikuchi, M., Shimazu, T., Suzuki, T., Sohtome, Y., Akakabe, M., Sodeoka, M., Dohmae, N., Umehara, T., and Shinkai, Y. (2018). Tri-methylation of ATF7IP by G9a/GLP recruits the chromodomain protein MPP8. Epigenetics Chromatin *11*, 56. 10.1186/s13072-018-0231-z.

48. Jones, S.E., Olsen, L., and Gajhede, M. (2018). Structural Basis of Histone Demethylase KDM6B Histone 3 Lysine 27 Specificity. Biochemistry *57*, 585-592. 10.1021/acs.biochem.7b01152.

49. Chater-Diehl, E., Goodman, S.J., Cytrynbaum, C., Turinsky, A.L., Choufani, S., and Weksberg, R. (2021). Anatomy of DNA methylation signatures: Emerging insights and applications. Am J Hum Genet *108*, 1359-1366. 10.1016/j.ajhg.2021.06.015.

50. Awamleh, Z., Goodman, S., Choufani, S., and Weksberg, R. (2024). DNA methylation signatures for chromatinopathies: current challenges and future applications. Hum Genet *143*, 551-557. 10.1007/s00439-023-02544-2.

51. Cedar, H., and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet *10*, 295-304. 10.1038/nrg2540.

52. Levy, M.A., Relator, R., McConkey, H., Pranckeviciene, E., Kerkhof, J., Barat-Houari, M., Bargiacchi, S., Biamino, E., Palomares Bralo, M., Cappuccio, G., et al. (2022). Functional correlation of genome-wide DNA methylation profiles in genetic neurodevelopmental disorders. Hum Mutat *43*, 1609-1628. 10.1002/humu.24446.

53. Levy, M.A., McConkey, H., Kerkhof, J., Barat-Houari, M., Bargiacchi, S., Biamino, E., Bralo, M.P., Cappuccio, G., Ciolfi, A., Clarke, A., et al. (2022). Novel diagnostic DNA methylation episignatures expand and refine the epigenetic landscapes of Mendelian disorders. HGG Adv *3*, 100075. 10.1016/j.xhgg.2021.100075.

54. Kooy, R.F. (2022). ZNF711 puts a spell on DNA. Eur J Hum Genet *30*, 396-397. 10.1038/s41431-022-01048-3.

55. Soshnev, A.A., Josefowicz, S.Z., and Allis, C.D. (2016). Greater Than the Sum of Parts: Complexity of the Dynamic Epigenome. Mol Cell *62*, 681-694. 10.1016/j.molcel.2016.05.004.

56. Haghshenas, S., Foroutan, A., Bhai, P., Levy, M.A., Relator, R., Kerkhof, J., McConkey, H., Skinner, C.D., Caylor, R.C., Tedder, M.L., et al. (2023). Identification of a DNA methylation signature for Renpenning syndrome (RENS1), a spliceopathy. Eur J Hum Genet *31*, 879-886. 10.1038/s41431-023-01313-z.

57. Rouxel, F., Relator, R., Kerkhof, J., McConkey, H., Levy, M., Dias, P., Barat-Houari, M., Bednarek, N., Boute, O., Chatron, N., et al. (2022). CDK13-related disorder: Report of a series of 18 previously unpublished individuals and description of an epigenetic signature. Genet Med *24*, 1096-1107. 10.1016/j.gim.2021.12.016.

58. Berg, S.K. (2020). Characterization of chromatin-bound interactome of cyclin-dependent kinase 12 (CDK12). Norwegian University of Life Sciences, Ås http://creativecommons.org/licenses/by-nc-nd/4.0/deed.no

59. Okazawa, H. (2018). PQBP1, an intrinsically disordered/denatured protein at the crossroad of intellectual disability and neurodegenerative diseases. Neurochem Int *119*, 17-25. 10.1016/j.neuint.2017.06.005.

60. Kulaeva, O.I., Gaykalova, D.A., and Studitsky, V.M. (2007). Transcription through chromatin by RNA polymerase II: histone displacement and exchange. Mutat Res *618*, 116-129. 10.1016/j.mrfmmm.2006.05.040.

61. Penagos-Puig, A., Claudio-Galeana, S., Stephenson-Gussinye, A., Jácome-López, K., Aguilar-Lomas, A., Chen, X., Pérez-Molina, R., and Furlan-Magaril, M. (2023). RNA polymerase II pausing regulates chromatin organization in erythrocytes. Nat Struct Mol Biol *30*, 1092-1104. 10.1038/s41594-023-01037-0.

62. Aref-Eshghi, E., Kerkhof, J., Pedro, V.P., Barat-Houari, M., Ruiz-Pallares, N., Andrau, J.C., Lacombe, D., Van-Gils, J., Fergelot, P., Dubourg, C., et al. (2020). Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in 42 Mendelian Neurodevelopmental Disorders. American journal of human genetics *106*, 356-370. 10.1016/j.ajhg.2020.01.019.

63. Jonkman, T.H., Dekkers, K.F., Slieker, R.C., Grant, C.D., Ikram, M.A., van Greevenbroek, M.M.J., Franke, L., Veldink, J.H., Boomsma, D.I., Slagboom, P.E., et al. (2022). Functional genomics analysis identifies T and NK cell activation as a driver of epigenetic clock progression. Genome Biol *23*, 24. 10.1186/s13059-021-02585-8.

64. Monk, D., Mackay, D.J.G., Eggermann, T., Maher, E.R., and Riccio, A. (2019). Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. Nat Rev Genet *20*, 235-248. 10.1038/s41576-018-0092-0.

65. Moore, L.D., Le, T., and Fan, G. (2013). DNA methylation and its basic function. Neuropsychopharmacology *38*, 23-38. 10.1038/npp.2012.112.

66. Choufani, S., Gibson, W.T., Turinsky, A.L., Chung, B.H.Y., Wang, T., Garg, K., Vitriolo, A., Cohen, A.S.A., Cyrus, S., Goodman, S., et al. (2020). DNA Methylation Signature for EZH2 Functionally Classifies Sequence Variants in Three PRC2 Complex Genes. Am J Hum Genet *106*, 596-610. 10.1016/j.ajhg.2020.03.008.

67. Choufani, S., Cytrynbaum, C., Chung, B.H., Turinsky, A.L., Grafodatskaya, D., Chen, Y.A., Cohen, A.S., Dupuis, L., Butcher, D.T., Siu, M.T., et al. (2015). NSD1 mutations generate a genome-wide DNA methylation signature. Nat Commun *6*, 10207. 10.1038/ncomms10207.

68. Greulich, F., Wierer, M., Mechtidou, A., Gonzalez-Garcia, O., and Uhlenhaut, N.H. (2021). The glucocorticoid receptor recruits the COMPASS complex to regulate inflammatory transcription at macrophage enhancers. Cell Rep *34*, 108742. 10.1016/j.celrep.2021.108742.

69. Martínez-Cano, J., Campos-Sánchez, E., and Cobaleda, C. (2019). Epigenetic Priming in Immunodeficiencies. Front Cell Dev Biol *7*, 125. 10.3389/fcell.2019.00125.

70. Awamleh, Z., Choufani, S., Wu, W., Rots, D., Dingemans, A.J.M., Nadif Kasri, N., Boronat, S., Ibañez-Mico, S., Cuesta Herraiz, L., Ferrer, I., et al. (2024). A new blood DNA methylation signature for Koolen-de Vries syndrome: Classification of missense KANSL1 variants and comparison to fibroblast cells. Eur J Hum Genet *32*, 324-332. 10.1038/s41431-024-01538-6.

71. Wang, T., Loo, C.E., and Kohli, R.M. (2022). Enzymatic approaches for profiling cytosine methylation and hydroxymethylation. Mol Metab *57*, 101314. 10.1016/j.molmet.2021.101314.

72. Sadikovic, B., Levy, M.A., Kerkhof, J., Aref-Eshghi, E., Schenkel, L., Stuart, A., McConkey, H., Henneman, P., Venema, A., Schwartz, C.E., et al. (2021). Clinical epigenomics: genome-wide DNA methylation analysis for the diagnosis of Mendelian disorders. Genet Med *23*, 1065-1074. 10.1038/s41436-020-01096-4.

73. Kerkhof, J., Rastin, C., Levy, M.A., Relator, R., McConkey, H., Demain, L., Dominguez-Garrido, E., Kaat, L.D., Houge, S.D., DuPont, B.R., et al. (2024). Diagnostic utility and reporting recommendations for clinical DNA methylation episignature testing in genetically undiagnosed rare diseases. Genetics in medicine : official journal of the American College of Medical Genetics *26*, 101075. 10.1016/j.gim.2024.101075.

74. Husson, T., Lecoquierre, F., Nicolas, G., Richard, A.C., Afenjar, A., Audebert-Bellanger, S., Badens, C., Bilan, F., Bizaoui, V., Boland, A., et al. (2024). Episignatures in practice: independent evaluation of published episignatures for the molecular diagnostics of ten neurodevelopmental disorders. Eur J Hum Genet *32*, 190-199. 10.1038/s41431-023-01474-x.

75. Oexle, K., Zech, M., Stühn, L.G., Siegert, S., Brunet, T., Schmidt, W.M., Wagner, M., Schmidt, A., Engels, H., Tilch, E., et al. (2023). Episignature analysis of moderate effects and mosaics. Eur J Hum Genet *31*, 1032-1039. 10.1038/s41431-023-01406-9.

76. Mirza-Schreiber, N., Zech, M., Wilson, R., Brunet, T., Wagner, M., Jech, R., Boesch, S., Škorvánek, M., Necpál, J., Weise, D., et al. (2022). Blood DNA methylation provides an accurate biomarker of KMT2B-related dystonia and predicts onset. Brain *145*, 644-654. 10.1093/brain/awab360.

77. Marwaha, A., Costain, G., Cytrynbaum, C., Mendoza-Londono, R., Chad, L., Awamleh, Z., Chater-Diehl, E., Choufani, S., and Weksberg, R. (2022). The utility of DNA methylation signatures in directing genome sequencing workflow: Kabuki syndrome and CDK13-related disorder. Am J Med Genet A *188*, 1368-1375. 10.1002/ajmg.a.62650.

78. Merid, S.K., Novoloaca, A., Sharp, G.C., Küpers, L.K., Kho, A.T., Roy, R., Gao, L., Annesi-Maesano, I., Jain, P., Plusquin, M., et al. (2020). Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age. Genome Med *12*, 25. 10.1186/s13073-020-0716-9.

79. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469-476. 10.1038/nature13127.

80. Adam, M.P., Banka, S., Bjornsson, H.T., Bodamer, O., Chudley, A.E., Harris, J., Kawame, H., Lanpher, B.C., Lindsley, A.W., Merla, G., et al. (2019). Kabuki syndrome: international consensus diagnostic criteria. J Med Genet *56*, 89-95. 10.1136/jmedgenet-2018-105625.

81. Butcher, D.T., Cytrynbaum, C., Turinsky, A.L., Siu, M.T., Inbar-Feigenberg, M., Mendoza-Londono, R., Chitayat, D., Walker, S., Machado, J., Caluseriu, O., et al. (2017). CHARGE and Kabuki Syndromes: Gene-Specific DNA Methylation Signatures Identify Epigenetic Mechanisms Linking These Clinically Overlapping Conditions. Am J Hum Genet *100*, 773-788. 10.1016/j.ajhg.2017.04.004.

82. Bend, E.G., Aref-Eshghi, E., Everman, D.B., Rogers, R.C., Cathey, S.S., Prijoles, E.J., Lyons, M.J., Davis, H., Clarkson, K., Gripp, K.W., et al. (2019). Gene domain-specific DNA methylation episignatures highlight distinct molecular entities of ADNP syndrome. Clin Epigenetics *11*, 64. 10.1186/s13148-019-0658-5.

83. Breen, M.S., Garg, P., Tang, L., Mendonca, D., Levy, T., Barbosa, M., Arnett, A.B., Kurtz-Nelson, E., Agolini, E., Battaglia, A., et al. (2020). Episignatures Stratifying Helsmoortel-Van Der Aa Syndrome Show Modest Correlation with Phenotype. Am J Hum Genet *107*, 555-563. 10.1016/j.ajhg.2020.07.003.

84. Smith, A.M., LaValle, T.A., Shinawi, M., Ramakrishnan, S.M., Abel, H.J., Hill, C.A., Kirkland, N.M., Rettig, M.P., Helton, N.M., Heath, S.E., et al. (2021). Functional and epigenetic phenotypes of humans and mice with DNMT3A Overgrowth Syndrome. Nat Commun *12*, 4549. 10.1038/s41467-021-24800-7.

85. Wu, X., and Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev Genet *18*, 517-534. 10.1038/nrg.2017.33.

86. Levy, M.A., Beck, D.B., Metcalfe, K., Douzgou, S., Sithambaram, S., Cottrell, T., Ansar, M., Kerkhof, J., Mignot, C., Nougues, M.C., et al. (2021). Deficiency of TET3 leads to a genome-wide DNA hypermethylation episignature in human whole blood. NPJ Genom Med *6*, 92. 10.1038/s41525-021-00256-y.

87. Goodman, S., Cytrynbaum, C., Chung, B., Chater-Diehl, E., Aziz, C., Turinsky, A., Kellam, B., Keller, M., Ko, J.M., Caluseriu, O., et al. (2020). EHMT1 pathogenic variants and 9q34.3 microdeletions share altered DNA methylation patterns in patients with Kleefstra syndrome. Journal of Translational Genetics and Genomics *4*, 144-158. 10.20517/jtgg.2020.23.

88. Messina, G., Prozzillo, Y., Delle Monache, F., Santopietro, M.V., Atterrato, M.T., and Dimitri, P. (2021). The ATPase SRCAP is associated with the mitotic apparatus, uncovering novel molecular aspects of Floating-Harbor syndrome. BMC Biol *19*, 184. 10.1186/s12915-021-01109-x.

89. Dong, S., Han, J., Chen, H., Liu, T., Huen, M.S.Y., Yang, Y., Guo, C., and Huang, J. (2014). The human SRCAP chromatin remodeling complex promotes DNA-end resection. Curr Biol *24*, 2097-2110. 10.1016/j.cub.2014.07.081.

90. Schnerwitzki, D., Perry, S., Ivanova, A., Caixeta, F.V., Cramer, P., Günther, S., Weber, K., Tafreshiha, A., Becker, L., Vargas Panesso, I.L., et al. (2018). Neuron-specific inactivation of Wt1 alters locomotion in mice and changes interneuron composition in the spinal cord. Life Sci Alliance *1*, e201800106. 10.26508/lsa.201800106.

91. Schnerwitzki, D., Hayn, C., Perner, B., and Englert, C. (2020). Wt1 Positive dB4 Neurons in the Hindbrain Are Crucial for Respiration. Front Neurosci *14*, 529487. 10.3389/fnins.2020.529487.

92. Ji, F., Wang, W., Feng, C., Gao, F., and Jiao, J. (2021). Brain-specific Wt1 deletion leads to depressive-like behaviors in mice via the recruitment of Tet2 to modulate Epo expression. Mol Psychiatry *26*, 4221-4233. 10.1038/s41380-020-0759-8.

93. Julkowska, D., Austin, C.P., Cutillo, C.M., Gancberg, D., Hager, C., Halftermeyer, J., Jonker, A.H., Lau, L.P.L., Norstedt, I., Rath, A., et al. (2017). The importance of international collaboration for rare diseases research: a European perspective. Gene Ther *24*, 562-571. 10.1038/gt.2017.29.

94. Zurek, B., Ellwanger, K., Vissers, L., Schüle, R., Synofzik, M., Töpf, A., de Voer, R.M., Laurie, S., Matalonga, L., Gilissen, C., et al. (2021). Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. Eur J Hum Genet *29*, 1325-1331. 10.1038/s41431-021-00859-0.

95. Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat *36*, 928-930. 10.1002/humu.22844.

96. Hamosh, A., Wohler, E., Martin, R., Griffith, S., Rodrigues, E.D.S., Antonescu, C., Doheny, K.F., Valle, D., and Sobreira, N. (2022). The impact of GeneMatcher on international data sharing and collaboration. Hum Mutat *43*, 668-673. 10.1002/humu.24350.

97. Dingemans, A.J.M., Stremmelaar, D.E., Vissers, L., Jansen, S., Nabais Sá, M.J., van Remortele, A., Jonis, N., Truijen, K., van de Ven, S., Ewals, J., et al. (2021). Human disease genes website series: An international, open and dynamic library for up-to-date clinical information. Am J Med Genet A *185*, 1039-1046. 10.1002/ajmg.a.62057.

98. Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., et al. (2020). ClinVar: improvements to accessing data. Nucleic Acids Res *48*, D835-d844. 10.1093/nar/gkz972.

99. Fokkema, I., van der Velde, K.J., Slofstra, M.K., Ruivenkamp, C.A.L., Vogel, M.J., Pfundt, R., Blok, M.J., Lekanne Deprez, R.H., Waisfisz, Q., Abbott, K.M., et al. (2019). Dutch genome diagnostic laboratories accelerated and improved variant interpretation and increased accuracy by sharing data. Hum Mutat *40*, 2230-2238. 10.1002/humu.23896.

100. Djaffardjy, M., Marchment, G., Sebe, C., Blanchet, R., Bellajhame, K., Gaignard, A., Lemoine, F., and Cohen-Boulakia, S. (2023). Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems. Comput Struct Biotechnol J *21*, 2075-2085. 10.1016/j.csbj.2023.03.003.

101. Roy, S., Coldren, C., Karunamurthy, A., Kip, N.S., Klee, E.W., Lincoln, S.E., Leon, A., Pullambhatla, M., Temple-Smolkin, R.L., Voelkerding, K.V., et al. (2018). Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. J Mol Diagn *20*, 4-27. 10.1016/j.jmoldx.2017.11.003.

102. Turinsky, A.L., Choufani, S., Lu, K., Liu, D., Mashouri, P., Min, D., Weksberg, R., and Brudno, M. (2020). EpigenCentral: Portal for DNA methylation data analysis and classification in rare diseases. Hum Mutat *41*, 1722-1733. 10.1002/humu.24076.

103. Awamleh, Z., Goodman, S., Kallurkar, P., Wu, W., Lu, K., Choufani, S., Turinsky, A.L., and Weksberg, R. (2022). Generation of DNA Methylation Signatures and Classification of Variants in Rare Neurodevelopmental Disorders Using EpigenCentral. Curr Protoc *2*, e597. 10.1002/cpz1.597.

104. Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D.E., et al. (2018). DNA methylation-based classification of central nervous system tumours. Nature *555*, 469-474. 10.1038/nature26000.

105. van der Velde, K.J., Singh, G., Kaliyaperumal, R., Liao, X., de Ridder, S., Rebers, S., Kerstens, H.H.D., de Andrade, F., van Reeuwijk, J., De Gruyter, F.E., et al. (2022). FAIR Genomes metadata schema promoting Next Generation Sequencing data reuse in Dutch healthcare and research. Sci Data *9*, 169. 10.1038/s41597-022-01265-x.

106. Wohler, E., Martin, R., Griffith, S., Rodrigues, E.D.S., Antonescu, C., Posey, J.E., Coban-Akdemir, Z., Jhangiani, S.N., Doheny, K.F., Lupski, J.R., et al. (2021). PhenoDB, GeneMatcher and VariantMatcher, tools for analysis and sharing of sequence data. Orphanet J Rare Dis *16*, 365. 10.1186/s13023-021-01916-z.

107. Srivastava, S., Love-Nichols, J.A., Dies, K.A., Ledbetter, D.H., Martin, C.L., Chung, W.K., Firth, H.V., Frazier, T., Hansen, R.L., Prock, L., et al. (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. Genet Med *21*, 2413-2421. 10.1038/s41436-019-0554-6.

108. van der Sanden, B., Schobers, G., Corominas Galbany, J., Koolen, D.A., Sinnema, M., van Reeuwijk, J., Stumpel, C., Kleefstra, T., de Vries, B.B.A., Ruiterkamp-Versteeg, M., et al. (2023). The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. Eur J Hum Genet *31*, 81-88. 10.1038/s41431-022-01185-9.

109. Colin, E., Duffourd, Y., Tisserant, E., Relator, R., Bruel, A.L., Tran Mau-Them, F., Denommé-Pichon, A.S., Safraou, H., Delanne, J., Jean-Marçais, N., et al. (2022). OMIXCARE: OMICS technologies solved about 33% of the patients with heterogeneous rare neuro-developmental disorders and negative exome sequencing results and identified 13% additional candidate variants. Front Cell Dev Biol *10*, 1021785. 10.3389/fcell.2022.1021785.

110. Wortmann, S.B., Oud, M.M., Alders, M., Coene, K.L.M., van der Crabben, S.N., Feichtinger, R.G., Garanto, A., Hoischen, A., Langeveld, M., Lefeber, D., et al. (2022). How to proceed after "negative" exome: A review on genetic diagnostics, limitations, challenges, and emerging new multiomics techniques. J Inherit Metab Dis *45*, 663-681. 10.1002/jimd.12507.

111. Kopajtich, R., Smirnov, D., Stenton, S., Loipfinger, S., Meng, C., Scheller, I., Freisinger, P., Baski, R., Berutti, R., Behr, J., et al. (2021). Integration of proteomics with genomics and transcriptomics increases the diagnostic rate of Mendelian disorders. medRxiv.

112. Cheung, W., Johnson, A., Rowell, W., Farrow, E., Hall, R., Cohen, A.S.A., Means, J., Zion, T., Portik, D., Saunders, C., et al. (2022). Direct haplotype-resolved 5-base HiFi sequencing for genome-wide profiling of hypermethylation outliers in a rare disease cohort. medRxiv.

113. Liu, N., Xiao, J., Gijavanekar, C., Pappan, K.L., Glinton, K.E., Shayota, B.J., Kennedy, A.D., Sun, Q., Sutton, V.R., and Elsea, S.H. (2021). Comparison of Untargeted Metabolomic Profiling vs Traditional Metabolic Screening to Identify Inborn Errors of Metabolism. JAMA Netw Open *4*, e2114155. 10.1001/jamanetworkopen.2021.14155.

114. Pauper, M., Kucuk, E., Wenger, A.M., Chakraborty, S., Baybayan, P., Kwint, M., van der Sanden, B., Nelen, M.R., Derks, R., Brunner, H.G., et al. (2021). Long-read trio sequencing of individuals with unsolved intellectual disability. Eur J Hum Genet *29*, 637-648. 10.1038/s41431-020-00770-0.

115. Steyaert, W., Sagath, L., Demidov, G., Yépez, V.A., Esteve-Codina, A., Gagneur, J., Ellwanger, K., Derks, R., Weiss, M., Ouden, A.d., et al. (2024). Unravelling undiagnosed rare disease cases by HiFi long-read genome sequencing. medRxiv, 2024.2005.2003.24305331. 10.1101/2024.05.03.24305331.

116. Mantere, T., Kersten, S., and Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. Front Genet *10*, 426. 10.3389/fgene.2019.00426.

117. Sanford Kobayashi, E., Batalov, S., Wenger, A.M., Lambert, C., Dhillon, H., Hall, R.J., Baybayan, P., Ding, Y., Rego, S., Wigby, K., et al. (2022). Approaches to long-read sequencing in a clinical setting to improve diagnostic rate. Sci Rep *12*, 16945. 10.1038/s41598-022-20113-x.

118. Kucuk, E., van der Sanden, B., O'Gorman, L., Kwint, M., Derks, R., Wenger, A.M., Lambert, C., Chakraborty, S., Baybayan, P., Rowell, W.J., et al. (2023). Comprehensive de novo mutation discovery with HiFi long-read sequencing. Genome Med *15*, 34. 10.1186/s13073-023-01183-6.

119. Lussier, A.A., Morin, A.M., MacIsaac, J.L., Salmon, J., Weinberg, J., Reynolds, J.N., Pavlidis, P., Chudley, A.E., and Kobor, M.S. (2018). DNA methylation as a predictor of fetal alcohol spectrum disorder. Clin Epigenetics *10*, 5. 10.1186/s13148-018-0439-6.

120. Li, J.L., Jain, N., Tamayo, L.I., Tong, L., Jasmine, F., Kibriya, M.G., Demanelis, K., Oliva, M., Chen, L.S., and Pierce, B.L. (2024). The association of cigarette smoking with DNA methylation and gene expression in human tissue samples. Am J Hum Genet *111*, 636-653. 10.1016/j.ajhg.2024.02.012.

121. Melka, M.G., Castellani, C.A., Rajakumar, N., O'Reilly, R., and Singh, S.M. (2014). Olanzapine-induced methylation alters cadherin gene families and associated pathways implicated in psychosis. BMC Neurosci *15*, 112. 10.1186/1471-2202-15-112.

122. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535-548.e524. 10.1016/j.cell.2018.12.015.

123. Smith, C., and Kitzman, J.O. (2023). Benchmarking splice variant prediction algorithms using massively parallel splicing assays. bioRxiv. 10.1101/2023.05.04.539398.

124. Walker, L.C., Hoya, M., Wiggins, G.A.R., Lindy, A., Vincent, L.M., Parsons, M.T., Canson, D.M., Bis-Brewer, D., Cass, A., Tchourbanov, A., et al. (2023). Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup. Am J Hum Genet *110*, 1046-1067. 10.1016/j.ajhg.2023.06.002.

125. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science *381*, eadg7492. 10.1126/science.adg7492.

126. Ellingford, J.M., Ahn, J.W., Bagnall, R.D., Baralle, D., Barton, S., Campbell, C., Downes, K., Ellard, S., Duff-Farrier, C., FitzPatrick, D.R., et al. (2022). Recommendations for clinical interpretation of variants found in non-coding regions of the genome. Genome Med *14*, 73. 10.1186/s13073-022-01073-3.

127. de Hond, A.A.H., Leeuwenberg, A.M., Hooft, L., Kant, I.M.J., Nijman, S.W.J., van Os, H.J.A., Aardoom, J.J., Debray, T.P.A., Schuit, E., van Smeden, M., et al. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med *5*, 2. 10.1038/s41746-021-00549-7.

128. Decherchi, S., Pedrini, E., Mordenti, M., Cavalli, A., and Sangiorgi, L. (2021). Opportunities and Challenges for Machine Learning in Rare Diseases. Front Med (Lausanne) *8*, 747612. 10.3389/fmed.2021.747612.

129. Anguela, X.M., and High, K.A. (2019). Entering the Modern Era of Gene Therapy. Annu Rev Med *70*, 273-288. 10.1146/annurev-med-012017-043332.

130. Bick, D., Bick, S.L., Dimmock, D.P., Fowler, T.A., Caulfield, M.J., and Scott, R.H. (2021). An online compendium of treatable genetic disorders. Am J Med Genet C Semin Med Genet *187*, 48-54. 10.1002/ajmg.c.31874.

131. Hechtelt Jonker, A., Hivert, V., Gabaldo, M., Batista, L., O'Connor, D., Aartsma-Rus, A., Day, S., Sakushima, K., and Ardigo, D. (2020). Boosting delivery of rare disease therapies: the IRDiRC Orphan Drug Development Guidebook. Nat Rev Drug Discov *19*, 495-496. 10.1038/d41573-020-00060-w.

132. Monaco, L., Zanello, G., Baynam, G., Jonker, A.H., Julkowska, D., Hartman, A.L., O'Connor, D., Wang, C.M., Wong-Rieger, D., and Pearce, D.A. (2022). Research on rare diseases: ten years of progress and challenges at IRDiRC. Nat Rev Drug Discov *21*, 319-320. 10.1038/d41573-022-00019-z.

133. May, M. (2023). Rare-disease researchers pioneer a unique approach to clinical trials. Nat Med *29*, 1884-1886. 10.1038/s41591-023-02333-4.

134. Dzierlega, K., and Yokota, T. (2020). Optimization of antisense-mediated exon skipping for Duchenne muscular dystrophy. Gene Ther *27*, 407-416. 10.1038/s41434-020-0156-6.

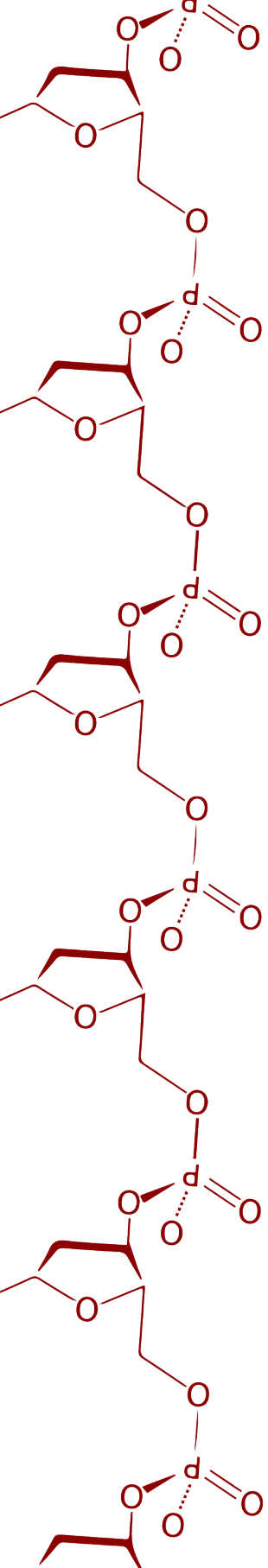
135. Zhang, L., Pilarowski, G., Pich, E.M., Nakatani, A., Dunlop, J., Baba, R., Matsuda, S., Daini, M., Hattori, Y., Matsumoto, S., et al. (2021). Inhibition of KDM1A activity restores adult neurogenesis and improves hippocampal memory in a mouse model of Kabuki syndrome. Mol Ther Methods Clin Dev *20*, 779-791. 10.1016/j.omtm.2021.02.011.

136. Aubert, Y., Egolf, S., and Capell, B.C. (2019). The Unexpected Noncatalytic Roles of Histone Modifiers in Development and Disease. Trends Genet *35*, 645-657. 10.1016/j.tig.2019.06.004.

137. Jeffery, C.J. (1999). Moonlighting proteins. Trends Biochem Sci *24*, 8-11. 10.1016/s0968-0004(98)01335-8.

138. Yuan, A.H., and Moazed, D. (2024). Minimal requirements for the epigenetic inheritance of engineered silent chromatin domains. Proc Natl Acad Sci U S A *121*, e2318455121. 10.1073/pnas.2318455121.

139. Koch, P.J., and Koster, M.I. (2021). Rare Genetic Disorders: Novel Treatment Strategies and Insights Into Human Biology. Front Genet *12*, 714764. 10.3389/fgene.2021.714764.



Appendix

Summary of the thesis

Neurodevelopmental disorders (NDDs) constitute a broad spectrum of rare conditions arising from various genetic and environmental factors. A subset of NDDs are caused by pathogenic variants in a single gene or locus and are termed **Mendelian NDDs**. These disorders present with diverse symptoms and exhibit significant clinical heterogeneity. In recent years, advancements in genetic technologies and transitioning from phenotype-first to genotype-first approaches have driven a surge in novel gene and Mendelian NDD discoveries, identifying the epigenetic machinery encoding genes as one of the major gene groups associated with Mendelian NDDs. This highlighted that the epigenetic gene regulation is essential for the normal neurodevelopment and can result in Mendelian NDD, when disrupted by a pathogenic genetic variant. Despite technological advancements, interpreting genetic data remains challenging, so a large proportion of the variants are being classified as variants of uncertain significance (VUS) (**Figure 1 in chapter 1**).

Because of the rarity of these conditions and diagnostic challenges, **the true clinical and molecular spectrum of the most Mendelian NDDs largely remain unknown,** despite increasing application of the next-generation sequencing. As result, this complicates 1) genetic variant interpretation, as well as 2) diagnosed patient care, requiring further research to characterize these conditions.

Therefore, **this thesis is aimed to comprehensively characterize the clinical, molecular, and DNA methylation spectrum and features of several Mendelian NDDs, focusing on the disorders of the epigenetic machinery**.

DNA methylation signature testing offers promising avenues for diagnostics and research, enabling the identification of mechanisms associated with NDDs, as well as providing complementary genetic testing, aiding in VUS interpretation, and/or confirming clinical diagnoses. However, DNAm signatures can be applied for multiple purposes:

1) Affected individual testing;
2) To delineate and compare conditions;
3) To study the underlying biology of a disorder.

The crosstalk between the epigenetic machinery components results in a complex and simultaneous depositions of DNA, histone, and nucleosome modifications resulting in changes of chromatin state and gene expression. Consequently,

NDDs caused by disruptions in genes encoding the epigenetic machinery can be associated with disorder-specific DNA methylation signatures. We have broadly utilized the DNA methylation signatures in all research chapters, except for **chapter 3**, because (unexpectedly) we were not able to identify a DNA methylation signature for the *KDM6B*-related NDD, despite the gene's clear role in epigenetic regulations. in **chapters 2, 4, and 5,** we utilized DNA methylation signatures to robustly test multiple VUS in the *SRCAP*, *KMT2C*, and *EHMT1* genes, respectively. In conjunction with protein 3D and other *in silico* and *in vitro* analyses, we were able to reclassify the majority of VUS in these genes. Finally, we have shown how the analysis of a large individual cohort in conjunction with DNA methylation analysis can be utilized to analyze protein functions. Notably, **the exact etiology of DNA signatures, as well as their sensitivity and specificity, remains unknown**, limiting their clinical applications.

In **chapter 2,** we described a novel Mendelian NDD with a specific DNA methylation signature caused by pathogenic (truncating) variants in the *SRCAP*, which was previously only associated with Floating-Harbor syndrome. We utilized DNA methylation signatures to prove that individuals with unspecific NDD features with a truncating *SRCAP* variant outside of the Floating-Harbor locus are indeed all affected by the same condition. This also showed that this novel NDD is different from Floating-Harbor syndrome not only clinically, but also molecularly.

In **chapters 3 to 6,** we expanded and delineated the clinical and genotypic spectrum of NDDs associated with the *KDM6B*, *KMT2C*, and *EHMT1* genes, as well as 9q34.3 duplications, respectively, by collecting and analyzing large cohorts of affected individuals. In these chapters, we demonstrate that a comprehensive analysis of the **clinical features** of a condition based on large cohorts of the affected individuals can provide 1) the more precise clinical spectrum and identify rare features of these conditions and 2) delineate these conditions from the similar conditions. For example, the *KDM6B*-related NDD was initially named in OMIM as "Neurodevelopmental disorder with coarse facies and mild distal skeletal abnormalities" based on the initial cohort of 12 individuals, but we have shown in a larger cohort (**chapter 3**) that coarse facies, nor distal skeletal abnormalities are not typical for this condition, therefore, having a misleading name in OMIM. As result, it has been recently renamed in OMIM as "Stolerman neurodevelopmental syndrome". Similarly, the *KMT2C*-related NDD was initially described in a patient with clinical features overlapping Kleefstra syndrome, which is caused by *EHMT1* haploinsufficiency, resulting in naming this condition in OMIM as Kleefstra syndrome 2. However, in **chapter 4**, by analysing 80 individuals with the *KMT2C*-related NDD, we

show that it is a unique clinical entity, distinct from the Kleefstra syndrome. We also demonstrate that a comprehensive analysis of the **variants** identified in affected individuals can provide not only the genetic spectrum of a condition, but also its pathomechanisms and biological functions of a gene. For example, disruption of a single PHD "reader" or the SET "writer" domain of KMT2C (see also **Figure 3 in chapter 1**) both results in the same condition, with the same DNA methylation signature, despite the presence of multiple PHD domains. In contrast, in **chapter 5**, we discovered that disruption of EHMT1 "reader", but not "writer" activity results in typical Kleefstra syndrome with the typical DNA methylation signature. based on the comprehensive *in silico*, *in vitro* and DNA methylation analysis of the identified individuals with *EHMT1 de novo* variants. Therefore, we demonstrate the need for and utility of large cohort analyses for understanding such rare conditions and providing optimal care and diagnostics, as well as possible therapeutic strategies. These efforts currently require laborious manual curation from multiple sources, so they would require broad and widespread data sharing to describe the clinical and molecular spectrum across all known Mendelian NDDs in the future.

In summary, we have demonstrated that **large cohort analyses together with DNAm signature testing and comprehensive variant analysis can provide important information about rare Mendelian NDD clinical and molecular spectra with important implications for patient care, disease mechanisms and the biological functions of a gene** and, as result, provide important information for the potential development of treatment strategies. We also demonstrated **the utility of DNAm signatures not only to robustly test various VUS, but also to delineate and compare different Mendelian NDDs conditions**, while also highlighting their limitations. However, international collaboration with broad and widespread data sharing, as well as broader DNA methylation testing, would be necessary to enable such analyses for a noticeable number of rare disorders. Additionally, further studies are necessary to determine the etiology and variability of DNAm signatures to improve their clinical applications.

Samenvatting van het proefschrift

Neurobiologische ontwikkelingsstoornissen (Engels "NDDs") vormen een breed spectrum van zeldzame aandoeningen die voortkomen uit verschillende genetische en omgevingsfactoren. Een subset van NDDs wordt veroorzaakt door pathogene varianten in een enkel gen of locus en wordt "**Mendeliaanse NDDs**" genoemd. Deze stoornissen vertonen diverse symptomen en aanzienlijke klinische heterogeniteit. In de afgelopen jaren hebben vooruitgangen in genetische technologieën en de overgang van fenotype-eerst naar genotype-eerst benaderingen geleid tot een toename van nieuwe genen en beschrijvingen van Mendeliaanse NDD's. Hierbij werden de voor zogenaamde epigenetische machinerie coderende-genen geïdentificeerd als een van de belangrijkste genengroepen geassocieerd met Mendeliaanse NDD's. Dit benadrukt dat de epigenetische genregulatie essentieel is voor de normale neurologische ontwikkeling en kan resulteren in Mendeliaanse NDD, wanneer dit verstoord wordt door een pathogene variant in een van de betreffende genen. Ondanks technologische vooruitgangen, blijft het interpreteren van genetische varianten een uitdaging, waardoor een groot deel wordt geclassificeerd als varianten van onzekere betekenis (Engels "VUS") (**Figuur 1** in **hoofdstuk 1**).

Vanwege de zeldzaamheid van deze aandoeningen en diagnostische uitdagingen, **blijft het klinische en moleculaire spectrum van de meeste Mendeliaanse NDD's grotendeels onbekend**, ondanks de toename van het diagnostische gebruik van "next-generation sequencing" technologies. Dit bemoeilijkt 1) de interpretatie van genetische varianten en 2) de zorg voor gediagnosticeerde patiënten. Daarom is een verder onderzoek vereist om deze aandoeningen beter te karakteriseren.

Het doel van **dit proefschrift is daarom het uitgebreid karakteriseren van het klinische, moleculaire en DNA-methyleringsspectrum en de kenmerken van meerdere Mendeliaanse NDD's, met de nadruk op de stoornissen van de epigenetische machinerie**.

DNA-methyleringssignatuur analyse biedt veelbelovende mogelijkheden voor diagnostiek en onderzoek. Hiermee kunnen mechanismen die geassocieerd zijn met NDDs bestudeerd worden en aanvullende genetische testen worden uitgevoerd, die helpen bij de interpretatie van VUS en/of het bevestigen van klinische diagnoses. DNA-methyleringssignaturenkunnen echter voor meerdere doeleinden worden gebruikt:

1) Om aangedane individuen te testen;
2) Om aandoeningen af te bakenen en te vergelijken;
3) Om de onderliggende biologie van een stoornis te bestuderen.

De interactie tussen de componenten van de epigenetische machinerie resulteert in complexe en gelijktijdige afzettingen van DNA-, histon- en nucleosoommodificaties die leiden tot veranderingen van de chromatine en genexpressie. Omdat NDDs veroorzaakt kunnen worden door verstoringen in genen die coderen voor de epigenetische machinerie, kunnen ze aandoeningpecifieke DNA-methyleringssignaturen hebben. We hebben DNA- methyleringssignaturen gebruikt in alle hoofdstukken, behalve in **hoofdstuk 3**, omdat we (onverwacht) geen DNA-methyleringssignatuur konden identificeren voor de *KDM6B*-gerelateerde NDD, ondanks de bekende rol van het gen in epigenetische reguleringen. In de **hoofdstukken 2**, **4 en 5**, hebben we DNA-methyleringssignaturen gebruikt om meerdere VUSs in de *SRCAP*, *KMT2C* en *EHMT1* genen robuust te testen. In combinatie met de 3D-eiwit en andere *in silico* en *in vitro* analyses, konden we de meerderheid van de VUS in deze genen te herclassificeren. Tenslotte, we laten zien hoe de analyse van genetische varianten verzameld vanuit relatief grote patienten cohorten , in combinatie met de DNA- methyleringssignatuuranalyse kan worden gebruikt om eiwitfuncties te bestuderen. **De exacte etiologie van DNA-signaturen, evenals hun sensitiviteit en specificiteit blijven echter onbekend, waardoor hun klinische toepassing beperkt is**.

In **hoofdstuk 2**, beschrijven we een nieuwe Mendeliaanse NDD met een specifieke DNA-methyleringssinatuur veroorzaakt door pathogene (truncerende) varianten in *SRCAP*, datvoorheen alleen geassocieerd was met Floating-Harbor syndroom. We gebruikten een DNA- methyleringssinatuuranalyse om aan te tonen dat personen met niet-specifieke NDD-kenmerken met een truncerende *SRCAP*-variant buiten het Floating-Harbor locus inderdaad allemaal een vergelijkbaar fenotype hebben. Dit toonde ook aan dat deze nieuwe NDD niet alleen klinisch, maar ook moleculair verschillend van het Floating-Harbor syndroom is.

In de **hoofdstukken 3 tot en met 6** hebben we het klinische en genotypische spectrum van NDD's geassocieerd met respectievelijk de *KDM6B*, *KMT2C* en *EHMT1* genen en 9q34.3 duplicaties uitgebreid en afgebakend door het verzamelen en analyseren van grote cohorten van getroffen individuen. In deze hoofdstukken, laten we zien dat een uitgebreide analyse van de klinische kenmerken van een aandoening op basis van relatief grotere cohorten kan zorgen voor 1) nauwkeuriger het klinische spectrum kan vaststellen en zeldzame kenmerken van deze aandoeningen kan

identificeren en 2) deze aandoeningen kan onderscheiden van vergelijkbare/ gerelateerde aandoeningen. Bijvoorbeeld, de *KDM6B*-gerelateerde NDD werd aanvankelijk in OMIM genoemd als "Neurodevelopmental disorder with coarse facies and mild distal skeletal abnormalities" op basis van het initiële cohort van 12 individuen van Stolerman et al., maar we hebben in een groter cohort (**hoofdstuk 3**) aangetoond dat “coarse facies”, net als “distal skeletal abnormalities” niet typisch zijn voor deze aandoening. Dus de OMIM naam was eigenlijk misleidend, maar nu is het onlangs hernoemd als "Stolerman neurodevelopmental syndrome". Op dezelfde manier, werd de *KMT2C*-gerelateerde NDD in eerste instantie beschreven bij een patiënt met klinische kenmerken die overlappen met het Kleefstra-syndroom, dat wordt veroorzaakt door *EHMT1* haploinsufficiëntie. Dit leidde tot het noemen van deze aandoening in OMIM als Kleefstra-syndroom 2. In **hoofdstuk 4,** laten we zien door analyse van 80 individuen met *KMT2C*-gerelateerde NDD dat het een unieke klinische entiteit is, verschillend van het Kleefstra syndroom. We laten ook zien dat een uitgebreide analyse van de **varianten** geïdentificeerd bij de aangedane individuen niet alleen het genetische spectrum van een aandoening kan definieren, maar ook de ziektemechanismen en biologische functies van een gen. Bijvoorbeeld, verstoring van een enkele PHD "reader" of de SET "writer" domeinen van *KMT2C* (zie ook **Figuur 3** in **hoofdstuk 1**) resulteert beide in dezelfde aandoening, met dezelfde DNA-methyleringssignatuur, ondanks de aanwezigheid van meerdere PHD-domeinen. Op basis van de uitgebreide *in silico*, *in vitro* en DNA-methyleringssignatuuranalyse van de geïdentificeerde individuen met *EHMT1 de novo* varianten, ontdekten ook we in **hoofdstuk 5** dat verstoring van de EHMT1 "reader", maar niet "writer" activiteit resulteert in het typische Kleefstra-syndroom met de typische DNA-methyleringssingatuur. Daarom tonen we de noodzaak en bruikbaarheid aan van de studie van grote cohorten om het volle spectrum van de respectievelijkezeldzame aandoeningen te begrijpen. Dit is een voorwaarde voorhet bieden van optimale zorg en diagnostiek, evenals toepassen en ontwikkelen van mogelijke therapeutische strategieën. Deze inspanningen vereisen momenteel moeizame handmatige curatie van meerdere bronnen, dus het brede en wijdverspreide delen van gegevens s nodig om het klinische en moleculaire spectrum van alle bekende Mendeliaanse NDD's in de toekomst te beschrijven.

Samenvattend, hebben we aangetoond dat **grote cohortstudies samen met DNA-methyleringssignaturen en uitgebreide variantanalyses belangrijke informatie kunnen bieden over de klinische en moleculaire spectra van zeldzame Mendeliaanse NDD's. Dit heeft belangrijke implicaties voor patiëntenzorg, ziektemechanismen en de biologische functies van betreffende genen te begrijpen. Bovendien kunnen we hiermee belangrijke informatie genereren**

voor de mogelijke ontwikkeling van behandelingsstrategieën. We hebben ook **de bruikbaarheid van DNA-methyleringssignaturen aangetoond, niet alleen om verschillende VUSs robuust te testen, maar ook om verschillende Mendeliaanse NDD-aandoeningen af te bakenen en te vergelijken, terwijl we ook hun tekortkomingen benadrukken**. Internationale samenwerking met het breed en wijdverspreid delen van gegevens, evenals bredere toepassing van DNA-methylatietesten, zijn echter nodig.. Daarnaast zijn verdere studies nodig om de etiologie en variabiliteit van DNA-methyleringssignaturen te bepalen om hun klinische toepassingen te verbeteren.

Research data management

Ethics

This thesis is based on the results of research involving human participants. The thesis and its **chapters 2 to 6** were conducted in accordance with the Declaration of Helsinki and local regulations. The studies described in **chapters 2 to 6** were performed based on the approvals granted by the Ethical Committee Arnhem-Nijmegen #2011/188 or #2018-4540. Included individuals or their legal representatives consented to inclusion in the studies and the publication of unidentifiable information. A specific consent was obtained for publishing identifiable data such as facial photographs, where applicable. The consents were obtained by recruiting clinicians/researchers and stored locally per local regulations.

Data collection, privacy, and availability

Standard clinical and genetic variant data were collected using standardized proforma completed by referral clinicians/researchers.

The individuals were pseudonymized for publication. The key, as well as non-coded and collected identifiable data are stored at the human genetics department of Radboudumc with access restricted to the members of the involved research groups. These data will be stored for at least 15 years.

The collected clinical and genetic variant data from **chapters 2 to 6** are published and available within the respective published manuscripts/chapters as supplemental information. The generated DNA methylation signatures described in **chapters 2, 4, and 6** are also provided within the published chapters as supplemental information. The identified genetic variants and their respective classification were additionally submitted to the open-access ClinVar database with reference to the respective study (accession numbers reported in the respective chapters). Therefore, the secondary data and results generated during the studies are published and are publicly available in line with the FAIR principles.

The (raw) primary next-generation sequencing data sets (like FASTQ/BAM files of exome or genome sequencing) were generated in diagnostic and/or research settings in Radboudumc and by respective collaborating institution and are stored locally and are not available due to the sensitive nature of the data. The (raw) primary DNA methylation array data sets (like .idat files) were generated by

collaborating institutions and are stored locally in respective institution and are not available publicly due to the sensitive nature of the data.

Funding

The project was funded by the Netherlands Organization for Health Research and Development grants (ZonMW VIDI grant nr.:91718310 to T.K.) and the Dutch Research Council (NWO Aspasia grant nr.:015.014.036 to T.K.). Further funding sources and conflicts of the interest from the research **chapters 2 to 6** are provided within the respective chapters.

Part of this PhD training was funded by the:

1) Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) Ter Meulen grant (to D.R.) for visit to R.Weksberg group at SickKids, Toronto, ON, Canada;
2) European Joint Programme on Rare Diseases (EJP RD) Mobility fellowship (to D.R.) to visit S.Banka group at Manchester University NHS Foundation Trust, Manchester, UK.

Curriculum vitae

Dmitrijs was born on 23rd December 1994 in Riga, Latvia (NL: Lettland). He graduated from Riga Ukrainian School in 2012 and continued his education at Riga Stradins University, Faculty of Medicine, where he obtained his first experience and interest in the field of human genetics. Dmitrijs obtained medical doctor degree in 2018.

After graduation, Dmitrijs began his PhD studies at the Department of Human genetics, Radboudumc, Nijmegen, The Netherlands under supervision of Prof. T. Kleefstra, Prof. L.E.L.M. Vissers, and Prof. H.G. Brunner. During his PhD trajectory, Dmitrijs focused on rare neurodevelopmental disorders and developed an interest in the epigenetics of rare disorders, with DNA methylation signatures emerged as a novel tool in the field. To deepen his understanding of DNA methylation analysis and result interpretation, he applied for two grants (EJP-RD and KNAW Ter Meulen) to visit leading research groups in the field: Prof.S.Banka's group in Manchester, UK, and Prof. R.Weksberg's group in Toronto, Canada, respectively. With both applications approved, Dmitrijs was able not only to learn and obtain practical experience of DNA methylation analysis, and establish collaborations for further projects, but also immerse himself into different educational, research and clinical practices/culture.

After completing his projects at Radboudumc, Dmitrijs was able to continue to study DNA methylation and genetics of rare disorders as a researcher in the group of prof. T. Kleefstra at Erasmus MC, Rotterdam, The Netherlands. Additionally, he works part-time as a clinical laboratory geneticist in Riga, Latvia, utilizing the knowledge and experience obtained during his time in Nijmegen and while visiting other centres.

During his PhD journey, Dmitrijs was fortunate not only to participate in various research meetings, learn from field experts, and get familiar with different research and healthcare organizational systems, as well as cultures, but also to be close to his friends and family, and marry loving and charming Adele, who gifted him the greatest joy in his life – their son, Dāvids.

PhD portfolio

Department: **Human Genetics, Radboudumc**
PhD period: **01/12/2018 – 31/03/2023**
PhD Supervisor(s): **Prof. T. Kleefstra, prof. L.E.L.M. Vissers, prof. H.G. Brunner**

Training activity	
Courses	**Year**
Introduction days (DGS; Radboudumc)	2019
Scientific integrity course (DGS)	2019
Graduate school day (DGS) x2	2019; 2021
Basic Course on Regulations and Organization for clinical investigators (NFU)	2021
Dutch language courses A0-B2 (Instituut Jeroen Bosch; Radboud in'to Languages)	2019-2021
Language Development for Academic Writing (RU)	2019
Statistics for PhD candidates by using SPSS (RU)	2019
Introduction in using R (Radboudumc)	2019
Grant writing and presenting (RU)	2019
8th European course in clinical dysmorphology & Eurodysmoclub (UCBM)	2020
Understanding proteins in 3D (RU)	2021
Scientific writing for PhD candidates (RU)	2021
Design and illustration (RU)	2021
Presenting and poster pitching (RU)	2021
The art of finishing up (RU)	2021
Analytic Storytelling (RU)	2021
9th international Workshop on cancer genetics and cytogenetics diagnostics (Radboudumc)	2022
(Inter)national lectures, symposia and conferences	**Year**
European Society of Human Genetics (ESHG) conference (Online): Oral presentation	2020
European Society of Human Genetics (ESHG) conference (Vienna, Austria): Poster presentation	2022
The Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) and the International Human Epigenome Consortium (IHEC) conference (Esterel, Quebec, Canada): Oral presentation	2022
American Society of Human Genetics (ASHG) conference (Los Angeles, CA, USA): Attendance	2022
Manchester Dysmorphology Conference (Manchester): Oral presentation	2023
EuroNDD workshop (Lisbon, Portugal): Oral presentation	2024

Seminars	**Year**
Theme discussions Human genetics department (weekly): attendance and presentations x2	2018-2023
Research meetings clinical genetics section (monthly): attendance and presentations x2	2018-2023
Radboud Research rounds NDD theme: attendance and presentation x1	2018-2022
Others	**Year**
European Joint Programme on Rare Diseases (EJP RD) Mobility fellowship for research visit prof. S.Banka group (Manchester University NHS Foundation Trust, Manchester, UK)	2021
Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) Ter Meulen grant for research visit prof. R.Weksberg group (SickKids, Toronto, ON, Canada)	2022
Best spoken presentation by an early career researcher at Manchester Dysmorphology Conference (Manchester, UK)	2023

Acknowledgements

I never planned on leaving Latvia, but after learning the Bertinoro courses and discovering what Nijmegen is doing, I could not imagine a better place to learn genetics. And I am eternally thankful for the opportunity and trust granted to me to pursue a PhD at Radboudumc and to all people who spent their time teaching and helping me. As I approach my thirties, my memory is not as sharp anymore, so I apologize if I forgotten to mention you personally.

I have had the privilege of working with the best supervision team possible, and I learned far more than I could have ever imagined.

Tjitske, thank you for always finding time – either answering my countless questions, manually going through the data, or reading manuscripts. I am incredibly grateful for your support and help to pursue novel ideas and for making them better. What you do (and achieve) for research, patients (as well as entire department nowadays) is truly inspiring.

Lisenka, thank you for teaching me how to better structure my thoughts and plans (and how to draw diagrams). While I did not master this art, I do hope I have at least improved significantly. Your guidance with so many things has been invaluable – both within and outside of genetics or academia. Thank you for showing masterclass of how innovations and translational research should be performed. However, what I am the especially grateful for is your help with non-work-related matters – whether it was just an advice, a friendly conversation, or literally bringing ironing-board to my new apartment. As a young and naïve expat, this meant a lot to me.

Han, thank you for your honesty and for sharing your knowledge with me. I enjoyed our discussions during meetings and learned a lot. I still remember one of our first progress meetings, when discussing the first results, you asked me if I have read the new paper by Lupski et al., about the *de novo* copy number variant mechanisms, which, of course, I was not familiar with. The paper was one week "old" at that moment. After that meeting, I realized – I should read much more new literature if I want to keep up at least with the meetings. This turned into a habbit, which served me immensely during my PhD studies.

I was blessed by the opportunity to learn also from other brilliant researchers and clinicians (apparently, I needed to learn too much and needed a lot of guidance).

David, I did not have any experience with clinical genetics when we have started working to finish a "small project" about *SRCAP*. Thank you for your time, patience, and experience guiding and helping me – I remember coming to you with a lot with questions. At the end, this "small project" was my first published project and ended up being fundamental to the rest of my thesis.

Sid, thank you for hosting me in Manchester, which allowed me to obtain my first hands-on experience with DNA methylation analysis. Thank you for collaborating on so many projects. Your ability to explain and structure complex topics – whether in a meeting or in a manuscript - is truly impressive, and I hope I have picked up at least some of those skills from you. And finally, thank you for introducing me to the wonders of Indian cuis

Rosanna and **Sanaa**, thank you for hosting me at SickKids. We have collaborated on so many projects and hopefully more to come! **Rosanna**, your commitment to quality and integrity is an example to us all. **Sanaa**, your boundless enthusiasm for research is just contagious! I am deeply grateful for your openess, your willingness to share your knowledge, and for teaching me nearly everything I know about DNA methylation. I enjoyed our discussion on so many different topics (from cancer to imprinting).

Eric, **Zain** and **Sarah**, it was a pleasure getting to know and working with you. Thank you for all your advice and feedback while I was in Toronto, as well as for many great collaborations.

I was lucky to work and getting to know so many great people from human genetics.

Alex and Christian, you were among the first people from Nijmegen I met (and from whom I got first impression about the amazing work being done at Radboudumc). Thank you for helping bringing me there. Thank you for sharing your knowledge and insights into the novel (or old) molecular or bioinformatic methods with me and with the whole genetics community. A separate thank you, Alex, for yout presentations tip (about memorising the first slides) - it has helped me a lot.

Joost, Elke, Arianne, Lex, Lot, Jet, Karolis, Juliet and all my other colleagues - it has been a pleasure to get to get to know you, as well as to work with you and having lunch discussions. It is remarkable how everyone of you became an expert in your respective field, while working at the same place, so I had the opportunity to ulitize your knowledge (whether clinical, molecular or AI) on so many of my

projects! Also thank you for bringing your questions - somehow those were more challenging than I had to solve for my projects, which pushed me to investigate harder and learn more.

Thank you for making the office life fun, as well helping me with Dutch and showing the most important spots in Nijmegen (like the best kapsalon places).

Bert, Rolph, Nicole and Erik-Jan, thank you for sharing your knowledge with me (and others) and always helping and explaining either clinical, or molecular questions about variants and diagnostics. Finally, thank you for contributing to my projects. I have learned a lot from you. I also have thoroughly enjoyed the ID meetings, so thanks for the opportunity.

Michael, thank you for guiding me through the lab, even though you had to overtake things at the end. Your positive attitude is infectious, and I am very grateful for your help, whcih has been vital for every project in this thesis.

Hanka, this would have been a different thesis without protein 3D analysis, so understanding proteins in 3D is one of the most important skills that I have learned during my PhD. Thank you for teaching me and helping with so many projects.

Tjitske, I owe you another huge thank you for bringing me into a new chapter of my life at **Erasmus MC** and for building a wonderful team, which does inspiring things. **Tjakko, Herma, Mark, Niko, Rachel, Virginie, Laura, Daphne, Fede** and everyone else in the department, it is a pleasure to work with you and to learn from you. **Fede**, thank you for your endless enthusiam and for agreeing to do all the crazy, but also boring (but necessary) projects together!

Thank you to all the collaborators and families for contributing to the projects laid out in this thesis or those yet to be finished. **Jamie, Taryn, Ayumi, Yoichi** and **Victor**, I truly appreciate your hard work and for co-authoring the manuscripts - it has been a pleasure collaborating with you!

Mani dārgie **BKUS** kolēģi, esmu ļoti pateicīgs un lepns par iespēju strādāt tik augsta līmeņa, profesionālā komandā tepat, Latvijā. Jūs esat fantastiski!

Madara, kaut mums var nesakrīst viedokļi, bet jau 10 gadus (kopš 2.kursa, starp citu!) es Tevi, profesore, apbrīnoju un Tu mani nebeidz iedvesmot. Paldies, ka palīdzēji man kļūt par to, kas es esmu (jeb pati vainīga, ja kaut kas nepatīk).

Linda, paldies par atbalstu jau kopš studiju laikiem!

Dārgie **draugi**, paldies par jūsu palīdzību, atbalstu un humoru. Kur lai es nenokļūtu, nekad nejūtos vientuļi, pateicoties jums.

Dārgā **mamma un tēti**, es pat nevaru iedomāties, ko es darītu, bez jūsu mūžīgā atbalsta. Jūs vienmēr man palīdzējāt, ļāvat darīt to, ko es gribu, pieņēmat Adeli kā savu meitu un esmu bazgala pateicīgs jums par to visu!

Mīļā **Adele** un **Dāvid**, jūs esat "ģenerālsponsori" tam, kāpēc es jūtos laimīgs un kāpēc es baudu katru dienu, kā arī tam, kāpēc šī disertācija tapa tik ilgi. Nezinu, kā esmu pelnījis tik fantastisku ģimeni, bet es esmu bezgala pateicīgs par šādu laimi.

Donders graduate school

For a successful research Institute, it is vital to train the next generation of scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School in 2009. The mission of the Donders Graduate School is to guide our graduates to become skilled academics who are equipped for a wide range of professions. To achieve this, we do our utmost to ensure that our PhD candidates receive support and supervision of the highest quality.

Since 2009, the Donders Graduate School has grown into a vibrant community of highly talented national and international PhD candidates, with over 500 PhD candidates enrolled. Their backgrounds cover a wide range of disciplines, from physics to psychology, medicine to psycholinguistics, and biology to artificial intelligence. Similarly, their interdisciplinary research covers genetic, molecular, and cellular processes at one end and computational, system-level neuroscience with cognitive and behavioural analysis at the other end. We ask all PhD candidates within the Donders Graduate School to publish their PhD thesis in de Donders Thesis Series. This series currently includes over 600 PhD theses from our PhD graduates and thereby provides a comprehensive overview of the diverse types of research performed at the Donders Institute. A complete overview of the Donders Thesis Series can be found on our website: https://www.ru.nl/donders/donders-series

The Donders Graduate School tracks the careers of our PhD graduates carefully. In general, the PhD graduates end up at high-quality positions in different sectors, for a complete overview see https://www.ru.nl/donders/destination-our-former-phd. A large proportion of our PhD alumni continue in academia (>50%). Most of them first work as a postdoc before growing into more senior research positions. They work at top institutes worldwide, such as University of Oxford, University of Cambridge, Stanford University, Princeton University, UCL London, MPI Leipzig, Karolinska Institute, UC Berkeley, EPFL Lausanne, and many others. In addition, a large group of PhD graduates continue in clinical positions, sometimes combining it with academic research. Clinical positions can be divided into medical doctors, for instance, in genetics, geriatrics, psychiatry, or neurology, and in psychologists, for instance as healthcare psychologist, clinical neuropsychologist, or clinical psychologist. Furthermore, there are PhD graduates who continue to work as researchers outside academia, for instance at non-profit or government organizations, or in pharmaceutical companies. There are also PhD graduates who work in education, such as teachers in high school, or as lecturers in higher education. Others continue

in a wide range of positions, such as policy advisors, project managers, consultants, data scientists, web- or software developers, business owners, regulatory affairs specialists, engineers, managers, or IT architects. As such, the career paths of Donders PhD graduates span a broad range of sectors and professions, but the common factor is that they almost all have become successful professionals.

For more information on the Donders Graduate School, as well as past and upcoming defences please visit:

http://www.ru.nl/donders/graduate-school/phd/