Functional correlation of genome-wide DNA methylation profiles in genetic neurodevelopmental disorders

Michael Levy[1], Raissa Relator[1], Haley McConkey[1], Erinija Prankeviciene[1], Jennifer Kerkhof, Mouna Barat-Houari[2], Sara Bargiacchi[3], Elisa Biamino[4], María Palomares Bralo[5], Gerarda Cappuccio[6,7], Andrea Ciolfi[8], Angus Clarke[9], Barbara R. DuPont[10], Mariet W. Elting[11], Laurence Faivre[12,13], Timothy Fee[10], Robin S Fletcher[10], Cherick Florian[14,15], Aidin Foroutan[16], Michael J. Friez[10], Cristina Gervasini[17], Sadegheh Haghshenas[16], Benjamin A. Hilton[10], Zandra Jenkins[18], Simranpreet Kaur[19], Suzanne Lewis[20], Raymond J. Louie[10], Silvia Maitz[21], Donatella Milani[22], Angela T. Morgan[23], Renske Oegema[24], Elsebet Østergaard[25,26], Nathalie Ruiz Pallares[2], Maria Piccione[27], Simone Pizzi[8], Astrid S Plomp[28], Cathryn Poulton[29], Jack Reilly[16], Raissa Relator[1], Rocio Rius[30,31], Stephen Robertson[18], Kathleen Rooney[1,16], Justine Rousseau [32], Gijs W. E. Santen[33], Fernando Santos-Simarro[5], Josephine Schijns[34], Gabriella Maria Squeo[35], Miya St John[23], Christel Thauvin-Robinet[12,13,36,37], Giovanna Traficante[3], Pleuntje J. van der Sluijs[33], Samantha A. Vergano[38,39], Niels Vos[40], Kellie K. Walden[10], Dimitar Azmanov[41], Tugce Balci[42,43], Siddharth Banka[44,45], Jozef Gecz[46,47], Peter Henneman[28], Jennifer A. Lee[10], Marcel M.A.M. Mannens[28], Tony Roscioli[48,49,50,51], Victoria Siu[42,43], David J. Amor[23], Gareth Baynam[29,52,53], Eric G. Bend[54], Kym Boycott[55,56], Nicola Brunetti-Pierri[6,7], Philippe M. Campeau[32], John Christodoulou[19], David Dyment[57], Natacha Esber[58], Jill A. Fahrner[59], Mark D. Fleming[60], David Genevieve[15], Kristin D. Kerrnohan[55,61], Alisdair McNeill[62], Leonie A. Menke[34], Giuseppe Merla[35,63], Paolo Prontera64, Cheryl Rockman-Greenberg[65], Charles Schwartz[10], Steven A. Skinner[10], Roger E. Stevenson[10], Antonio Vitobello[12,36], Marco Tartaglia[8], Matthew L. Tedder[10] , Marielle Alders[28], Bekim Sadikovic[1,16]

1 Verspeeten Clinical Genome Centre; London Health Sciences Centre, London, ON N6A 5W9, Canada.
2 Autoinflammatory and Rare Diseases Unit, Medical Genetic Department for Rare Diseases and Personalized Medicine, CHU Montpellier, Montpellier, France.
3 Medical Genetics Unit, "A. Meyer" Children's Hospital of Florence, Florence, Italy.
4 Department of Pediatrics, University of Turin, Italy.
5 Institute of Medical and Molecular Genetics (INGEMM), Hospital Universitario La Paz, IdiPAZ, CIBERER, ISCIII, Madrid, Spain.
6 Department of Translational Medicine, Federico II University of Naples, Italy.
7 Telethon Institute of Genetics and Medicine, Pozzuoli, Italy.
8 Genetics and Rare Diseases Research Division, Ospedale Pediatrico Bambino Gesù, IRCCS, 00146 Rome, Italy.
9 Cardiff University School of Medicine, Cardiff, United Kingdom.
10 Greenwood Genetic Center, Greenwood, SC, 29646, USA.
11 Department of Clinical Genetics, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands.
12 INSERM-Université de Bourgogne UMR1231 GAD « Génétique Des Anomalies du Développement », FHU-TRANSLAD, UFR Des Sciences de Santé, Dijon, France.
13 Centre de Référence Maladies Rares «Anomalies du Développement et Syndromes Malformatifs », Centre de Génétique, FHU-TRANSLAD, CHU Dijon Bourgogne, Dijon, France.
14 Genetic medical center, CHU Clermont Ferrand, France.
15 Montpellier University, Reference Center for Rare Disease, Medical Genetic Department for Rare Disease and Personalize Medicine, Inserm Unit 1183, CHU Montpellier, Montpellier, France.
16 Department of Pathology and Laboratory Medicine, Western University, London, ON N6A 3K7, Canada.

17 Division of Medical Genetics, Department of Health Sciences, Università degli Studi di Milano, Milan, Italy.
18 Dunedin School of Medicine, University of Otago, Dunedin, New Zealand.
19 Brain and Mitochondrial Research Group, Murdoch Children's Research Institute and Department of Paediatrics, University of Melbourne, Melbourne, Australia.
20 BC Children's and Women's Hospital and Department of Medical Genetics, Faculty of Medicine, University of British Columbia.
21 Clinical Pediatric Genetics Unit, Pediatrics Clinics, MBBM Foundation, Hospital San Gerardo, Monza, Italy.
22 Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy.
23 Murdoch Children's Research Institute and Department of Paediatrics, University of Melbourne, Melbourne, Australia.
24 Department of Genetics, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands.
25 Department of Clinical Genetics, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark.
26 Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark.
27 Medical Genetics Unit Department of Health Promotion, Mother and Child Care, Internal Medicine and Medical Specialties, University of Palermo, Palermo, Italy.
28 Amsterdam UMC, University of Amsterdam, Department of Human Genetics, Amsterdam Reproduction and Development Research Institute, Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands.
29 Undiagnosed Diseases Program, Genetic Services of Western Australia, King Edward Memorial Hospital, Perth, Australia.
30 Brain and Mitochondrial Research Group, Murdoch Children's Research Institute, Melbourne, Australia.
31 Department of Paediatrics, University of Melbourne, Melbourne, Australia.
32 CHU Sainte-Justine Research Center, University of Montreal, Montreal, QC, H3T 1C5.
33 Department of Clinical Genetics, LUMC, Leiden, The Netherlands.
34 Department of Pediatrics, Emma Children's Hospital, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands.
35 Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Via S. Pansini 5, 80131 Naples, Italy.
36 Unité Fonctionnelle d'Innovation Diagnostique des Maladies Rares, FHU-TRANSLAD, France Hospitalo-Universitaire Médecine Translationnelle et Anomalies du Développement (TRANSLAD), Centre Hospitalier Universitaire Dijon Bourgogne, CHU Dijon Bourgogne, Dijon, France.
37 Centre de Référence Déficiences Intellectuelles de Causes Rares, Hôpital D'Enfants, CHU Dijon Bourgogne, 21000, Dijon, France.
38 Division of Medical Genetics and Metabolism, Children's Hospital of The King's Daughters, Norfolk VA, USA.
39 Department of Pediatrics, Eastern Virginia Medical School, Norfolk, VA, USA.
40 Department of Clinical Genetics, Amsterdam UMC, University of Amsterdam, Amsterdam Reproduction and Development Research Institute, Meibergdreef 9, Amsterdam, The Netherlands.
41 Department of Diagnostic Genomics, PathWest Laboratory Medicine, QEII Medical Centre, Perth, Australia.
42 Department of Pediatrics, Division of Medical Genetics,Western University, London, ON N6A 3K7.
43 Medical Genetics Program of Southwestern Ontario, London Health Sciences Centre and Children's Health Research Institute, London, ON N6A5W9, Canada.

44 Division of Evolution, Infection & Genomics, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, United Kingdom.
45 Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University NHS Foundation Trust, Health Innovation Manchester, Manchester, United Kingdom.
46 School of Medicine, Robinson Research Institute, University of Adelaide, Adelaide, SA 5005, Australia.
47 South Australian Health and Medical Research Institute, Adelaide, SA 5005, Australia.
48 Neuroscience Research Australia (NeuRA), Sydney, Australia.
49 Prince of Wales Clinical School, Faculty of Medicine, University of New South Wales, Sydney, Australia.
50 New South Wales Health Pathology Randwick Genomics, Prince of Wales Hospital, Sydney, Australia.
51 Centre for Clinical Genetics, Sydney Children's Hospital, Sydney, Australia.
52 Undiagnosed Diseases Program, Genetic Services of Western Australia, King Edward Memorial Hospital, Perth, Australia.
53 Division of Paediatrics and Telethon Kids Institute, Faculty of Health and Medical Sciences, Perth, Australia.
54 PreventionGenetics, Marshfield, WI, USA.
55 Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, ON, Canada.
56 Department of Genetics, Children's Hospital of Eastern Ontario, Ottawa, ON, Canada.
57 Children's Hospital of Eastern Ontario, Ottawa, Canada.
58 KAT6A Foundation.
59 Departments of Genetic Medicine and Pediatrics, Johns Hopkins University, Baltimore, MD, 21205, USA.
60 Department of Pathology, Boston Children’s Hospital.
61 Newborn Screening Ontario, Children's Hospital of Eastern Ontario, Ottawa, Canada.
62 Department of Neuroscience, University of Sheffield, UK, and Sheffield Children's Hospital NHS Foundation Trust.
63 Laboratory of Regulatory and Functional Genomics, Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo (Foggia), Italy.
64 Medical Genetics Unit, University of Perugia Hospital SM della Misericordia, Perugia, Italy.
65 Dept of Pediatrics and Child Health, Rady Faculty of Health Sciences, University of Manitoba and Program in Genetics and Metabolism, Shared Health MB, Winnipeg, MB.

Corresponding Author:

Bekim Sadikovic PhD DABMG FACMG
Program Head, Molecular Diagnostics, Pathology and Laboratory Medicine, London Health Sciences Centre and St. Joseph’s Health Care London
Professor, Pathology and Laboratory Medicine, Western University
Scientific and Clinical Director, Verspeeten Clinical Genome Centre, London Health Sciences Centre
Email bekim.sadikovic@lhsc.on.ca
Tel. 519 685 8500 x53074

Abstract

An expanding range of hereditary genetic syndromes are characterized by genome-wide disruptions in DNA methylation profiles referred to as episignatures. Episignatures, detectible in peripheral blood, are distinct, highly sensitive and specific biomarkers that have recently been applied in clinical diagnosis of genetic syndromes. Episignatures are contained within the broader disorder-specific genome-wide DNA methylation changes which can share significant overlap amongst different episignature conditions. In this study we perform functional genomic assessment and comparison of disorder-specific and overlapping genome-wide DNA methylation changes related to 65 genetic syndromes with previously described diagnostic episignatures. We demonstrate evidence of disorder-specific and recurring genome wide differentially methylated probes (DMPs) and regions (DMRs). Overall distribution of DMPs and DMRs across majority of the neurodevelopmental genetic syndromes shows substantial enrichment in gene promoters and CpG islands, and under representation of the more variable intergenic regions. Overrepresentation analysis shows significant enrichment of the recurring DMPs and DMRs in gene pathways and networks related to neurodevelopment, including neuronal generation and differentiation and axon guidance. We demonstrate a strong correlation between the molecular function of the affected genes and the relatedness of the consequent DNA methylation profiles as evidence of functional roles of DNA methylation episignatures in the etiology of genetic neurodevelopmental disorders. This study builds on our understanding of DNA methylation, expanding beyond the diagnostic utility of DNA methylation episignatures as a key functional element in the molecular etiology of genetic neurodevelopmental disorders.

Introduction

DNA methylation is a fundamental aspect of mammalian development, and changes in DNA methylation are closely related to variation in the underlying genome [PMID 31399642, 33931130]. A rapidly growing number of genes causing neurodevelopmental syndromes has been shown to be associated with distinct changes in DNA methylation in patients affected with these disorders [PMID 30456829]. These methylation changes may be a direct cause of the disruption of the gene function as in the chromatin remodeling, DNA methyltransferase, and histone modification genes [PMID 34608297, 30875234]. More recent work has shown that genome-wide changes in DNA methylation are also found in patients with pathogenic variants in genes that have no known direct role in DNA methylation or chromatin remodeling [V3 paper]. Such indirect changes may be caused by perturbations in the interconnected molecular pathways, including transcriptional regulation and protein signaling. Genetic variants that are inherited or that occur at the earliest stages of development can therefore have wide reaching impact on DNA methylation throughout development. Such changes can be propagated through cell differentiation and tissues development. Hence, easily accessible tissue such as peripheral blood can be used to demonstrate changes in DNA methylation and develop biomarkers of specific syndromes that occurred at early stages of development [PMID 32109418, V3 paper]. More than sixty syndromes genetic neurodevelopmental syndromes have now been identified which exhibit such changes DNA methylation, and the patterns of DNA methylation changes, referred to as episignatures, are now being used to as a diagnostic clinical biomarker [33547396].

We have previously described development of DNA methylation episignatures as highly sensitive and specific diagnostic biomarkers in over sixty genetic neurodevelopmental disorders [PMID 32109418, V3 paper]. We demonstrated that a genetic syndrome may have more than one episignature depending on the location and/or functional consequence of the underlying genetic variant. Conversely, similar

syndromes, such as those caused by pathogenic variants in genes from the same gene family or a protein network, may share a common episignature. As molecular biomarkers, episignatures are optimized for clinical diagnostics and commonly represent only a fraction of the totality of the DNA methylation change in any given disorders.

We have previously described 57 distinct diagnostic episignatures encompassing 65 neurodevelopmental syndromes caused by pathogenic variants in 61 genes [AJHG 2019, AJHG, 2020, AJHG 2021, GIM 2021, and V3 paper]. In this study we expand on this work by investigating the broader context of changes in DNA methylation by performing functional genomic assessment and comparison of disorder-specific and overlapping genome-wide DNA methylation changes in these syndromes. We aimed to map disorder-specific and recurring genome wide differentially methylated probes (DMPs) and regions (DMRs) in relation to the functional genomic elements including gene promoters and CpG islands and intergenic regions. We explore the functional impact of these changes in relation to the corresponding gene pathways and transcriptional networks. By using various correlation analyses we assess relatedness of the genetic etiology and the consequent DNA methylation profiles in the etiology of genetic neurodevelopmental disorders.

Materials and Methods

Patient cohorts

The case cohorts consisted of 1381 samples from patients who were diagnosed with one of the 65 neurodevelopmental conditions and who had a positive Episign result for one of the corresponding 58 episignatures. Mean, median, minimum, and maximum case cohort sizes were 25, 14, 3, and 191 (Table X, Table SX). The control cohort consisted of 4231 samples: 2701 unaffected controls and 1530

unresolved samples. Unaffected controls are from individuals with no specific neurodevelopmental phenotype and no known pathogenic or suspected pathogenic variant in any of the episignature-related genes. These controls included a mix of samples from publicly available databases indicated to be "control", "wildtype", or similar, and new samples from patients clinically assessed as not having a neurodevelopmental phenotype. Unresolved samples are from patients with a neurodevelopmental phenotype but who were assessed as negative for all current episignatures [ref V3 paper].

Sample processing

Peripheral blood DNA was extracted using standard techniques. Bisulfite conversion was performed with 500 ng of genomic DNA using the Zymo EZ-96 DNA Methylation Kit (D5004), and bisulfite-converted DNA was used as input to the Illumina Infinium HumanMethylation450 (450K array) or MethylationEPIC BeadChip array (EPIC array). Array data were generated according to the manufacturer's protocol. Sample quality control was performed using the R minfi package version 1.38.0.

Methylation probe processing and selection

The data analysis pipeline was adapted from previously described [PMID 32109418, V3 paper]. IDAT files containing methylated and unmethylated signal intensity were imported into R 4.1.0 for analysis. Normalization was performed using the Illumina normalization method with background correction using the minfi package. Probes with a detection p-value > 0.01, probes located on the X and Y chromosomes, probes which contained SNPs at the CpG interrogation or single nucleotide extension sites, and probes which are known to cross-react with other genomic locations were removed [PMID 23314698, 27717381]. For each cohort, a set of controls was chosen using the R package matchit version

4.2.0, matched for age, sex, and array type. For each case sample, one to ten controls were used (case:control ratio of 1:1 to 1:10). Mean, median, minimum, and maximum control cohort sizes were 60, 56, 30, and 191 (Table SXX).

Methylation levels (beta values) were used for linear regression modeling using the limma package version 3.48.0 [PMID 25605792]. Estimated blood cell proportions [PMID 22568884] were added to the model matrix as confounding variables. The generated p-values were moderated using the eBayes function. To facilitate comparisons between samples processed using 450K and EPIC arrays, only probes found on both arrays were used for analysis. Probes which had a mean methylation difference of less than 5% between the case and control samples were removed and Benjamini-Hochberg adjusted p values were calculated for the remaining probes. Probes with an adjusted p value less than 0.01 were selected as DMPs for analysis, except for cohorts KDM4B and CSS4_c.2650 which had too few probes and probes with a non-adjusted p value < 0.001 were used.

Identification of differentially methylated regions

Genome-wide DMR analysis was also implemented to determine regions that are significantly differentiated between cohort cases and matched controls. Methylation beta values equal to 0 or 1 were initially shifted by a very small value (1e-10) to avoid infinite M-values during conversion implemented using minfi [minfi]. DMR analysis on the matrix of M-values were identified using the DMRcate package [dmrcate], where regions were defined to have at least five CpG probes within 1000 bp of each other. Minimum absolute mean methylation difference between cohort cases and controls was set to 0.1 and significant results were chosen using a Fisher p-value cut-off of 0.01.

Cohort comparisons and data visualizations

Circos-style plots were made using the R package circlize version 0.4.14. Differentially methylated regions (DMRs) and DMPs were annotated in relation to CpG islands (CGI) and genes using the R

package annotatr version 1.18.1 with AnnotationHub version 3.0.0 and annotations hg19_cpgs, hg19_basicgenes, hg19_genes_intergenic, and hg19_genes_intronexonboundaries. CGI annotations included CGI shores from 0-2kb on either side of CGIs, CGI shelves from 2-4kb on either side of CGI, and inter-CGI regions encompassing all remaining regions. For gene annotations, promoters were up to 1Kb upstream of the transcription start site (TSS) and promoter+ the region 1-5Kb upstream of the TSS. Annotations to untranslated regions (5' UTR and 3' UTR), exons, introns, and exon/intron boundaries were combined into the coding sequence (CDS) category. Heatmaps were made using the R package pheatmap version 1.0.12.

Tree and leaf dendrograms were made by first aggregating probe methylation levels using their median value across samples with the same condition. Euclidean clustering using Ward's method on the distances was then implemented on the combined values. Our initial analysis showed that the number of DMPs affect the clustering results, so final analysis was implemented by initially selecting the top 500 DMPs ranked by p-values for each cohort before aggregation of beta values. For cohorts with fewer than 500 DMPs all DMPS were used. This resulted to 20904 distinct probes across all groups. Clustering results were visualized as a binary tree using the R package TreeAndLeaf [Ref/version treeandleaf] to incorporate additional information such as global mean methylation difference and total number of DMPs selected for each cohort.

Two-dimensional and three-dimensional representations of the topological structure of the entire cohort database were analyzed using unified mapping approximation and projection (UMAP). The global structure approximated by UMAP [Ref umap] was obtained by using 210 probes that were most differentiating across all cohorts selected by random forest feature importance as described below. The UMAP parameter for the number of nearest neighbors was set to 10 and minimum distance in final layout set to 0.99, and results were visualized using 2D and 3D scatter plots [ggplot2, plot3D].

Selection of most differentiating probes across all cohorts

Probes evaluated to be most discriminating of the 56 cohorts were identified by random forest using the R package randomforest version XX [Ref randomforest]. Feature importance was computed for the selected 20904 probes as previously described. Random forest multiclassification models were trained using the 20904 probes and 1381 samples, and variable importance measured as mean accuracy decrease was computed for each probe. Due to the randomness of the model, we repeated the procedure 1000 times and summed all variable importance values across all sets. Finally, probes who ranked in the top one percentile were selected (210 probes). For each trial, 100 trees were fitted using 145 (default: sqrt(# of features)) randomly sampled probes at each split. Downsampling was incorporated to account for the sample imbalance across cohorts, and the number of samples drawn per group at each split was set to the minimum number of samples among all cohorts to ensure an identical value.

Functional annotation of genes overlapping selected DMPs and DMRs

Gene Ontology (GO) and KEGG pathways associated with DMPs and DMRs were identified using the R package clusterProfiler version XX, as well as human phenotype ontology using enrichR [Ref enrichR]. DMPs were first converted to gene IDs using the R package missMethyl version X. Overrepresentation analysis (ORA) was performed using either all DMPs, or using DMPs that were found in at least five cohorts. For DMRs, the CpG probes within each DMR were converted to gene IDs. Then ORA was performed using either all probes within all DMRs, or using probes that were found in at least two DMRs. The background lists of annotations were generated using all probes used for differential methylation analysis of all signatures (post-filtering).

Network diagrams

A network diagram was made by first determining the number of shared probes between each pair of cohorts. Each probe in a pair of cohorts was categorized depending on the direction of the probe's change in methylation: hyper-hyper (probe had increased methylation in both cohorts), hypo-hypo (probe had decreased methylation in both cohorts), hypo-hyper or hyper-hypo (probe was increased in one cohort and decreased in the other). The obtained data matrix was visualized using Cytoscape version 3.9 [PMID 25199793] in which nodes represent cohorts, the edges connecting the nodes represent hyper or hypo methylated probes and the weight of the edge is proportional to the absolute count of the probes shared by the two nodes.

Results

Detection of differentially methylated probes and regions

Using the above describe method we generated lists of DMPs for each cohort. CSS4_c.2650 which only has three samples had zero significant probes, and KDM4B which has 6 samples but more mild methylation changes had only 77 DMPs. For these two cohorts we therefore used a non-adjusted p value to attain 464 and 279 DMPs for all subsequent analysis. The 56 cohorts therefore ranged minimum of 279 DMPs for KDM4B to 151848 DMPs for ADCADN, with a mean of 13427 and a median of 5272 (Table 1, Figure S1A).

We next searched for DMPs that are found in more than one cohort. The 56 cohorts include a total of 253431 unique DMPs. 113911 (55%) are unique, meaning they are found in only one cohort, while 139520 (45%) are found in two or more cohorts. Most of the unique DMPs were found in the cohorts with the largest number of total DMPs: ADCADN and ICF2_3_4 accounted for 85% of the unique probes (Figure S1B). All other cohorts shared at least 85% of their DMPs with at least one other cohort, with all 1015 BAFopathy DMPs found in at least one other cohort (Figure S1B). The cohorts with the largest

number of DMPs generally also had the largest number of shared DMPs (Figure 1, Figures S1A,B). Among the 139520 DMPs found in two or more cohorts 46635 (33%) are found in exactly two cohorts while one DMP is found in each of 27 and 28 cohorts (Figure S1C).

The lists of DMPs were then used to identify differentially methylated regions (DMRs). 48 cohorts returned significant results ranging from one to 1384 DMRs, with the median of the DMR counts at eight and mean of 89 (Supplementary Table XX). Eight cohorts did not have any significant DMR detected: AUTS18, BAFopathy, CSS_c.6200, CSS4_c.2650, CSS9, Kabuki, KDM4B, MRX93. Two of the cohorts with no DMRs, CSS4_c.2650 and KDM4B, have the fewest number of DMPs when using an adjusted p-value cut-off, explaining the lack of identified DMRs. Most cohorts with no significant DMRs either have mild changes in methylation (Kabuki, MRD51, MRX93) or relatively small number of identified DMPs (AUTS18), or both (BAFopathy, CSS9). Therefore, as expected, cohorts that were highly hypo/hypermethylated and with relatively large ratio of DMPs also have the highest number of identified DMRs, such as ADCADN (1384 DMRs), ICF2_3_4 (851), Sotos (809) and ICF1 (514). [To add supplementary figure (scatter plot) showing relationship between # of DMPs, # of DMRs, and mean methylation changes]

Genomic context of differentially methylated probes

We next examined the genomic locations of the probes. First, we assessed locations in relation to CpG islands (CGI). CpG annotations were available for 3,137,161,264 nucleotides divided into CGI (0.7%), CGI shores (3.2%), CGI shelves (2.8%), and inter-CGI regions (93.3%). However, since CGI are often the

location of DNA methylation they are over-represented on the DNA methylation microarrays. After initial filtering to remove chromosome X and Y and certain other probes as described in the Methods, 18.6% of microarray probes overlapped with CGI. This represents the “background” or “default” distribution of probes on the microarray (Figure 2A). 44 of the 56 cohorts (78.6%) had probes enriched for CGI, meaning they had greater than 18.6% of their DMPs in CGI, while 12 cohorts have less than 18.6% of their DMPs in CGI. CGI and nearby regions (shelves and shores) account for 43.6% of probes in the default distribution. Nearly all (54/56) of the cohorts were enriched for these CGI and near-CGI regions. Only ICF2_3_4 and WHS are enriched for inter-CGI regions (Figure 2A).

Similar analysis was then performed for the 49 cohorts which had at least one DMR. Since the microarrays contain probes and not DMRs a default distribution for DMRs cannot be generated. However, since DMRs require several probes within a limited region it is expected that DMRs even by random chance will be more often found in CGIs. 64% of the 5221 total DMRs overlapped CGI, and 38 of the 49 DMR cohorts (79%) had 50% or more of their DMRs overlapping CGI. There was variability in results between cohorts with several having all DMRs overlapping CGI and the lowest (cohort WHS) having only 1/13 (7.7%) of its DMRs overlapping CGI (Figure 2B).

DMPs and DMRs were then annotated in relation to genes. The default distribution of probes found 22.4% at promoters, 4.6% promoter+, 49.9% in CDS, and 23.1% intergenic (Figure 2C). 43/56 cohorts (76.8%) had probes enriched for promoters, meaning they had greater than 22.4% of their DMPs in promoters. Promoters and promoters+ account for 27.0% of probes in the default distribution. Nearly all (55/56) of the cohorts were enriched for this extended promoter region. Only ICF2_3_4 at 25.8% was not enriched for these promoter regions (Figure 2C). 58.1% of the 5221 total DMRs overlapped the extended promoter regions, and 29 of the 49 DMR cohorts (59.2%) had 50% or more of their DMRs overlapping the extended promoter regions (Figure 2D).

We performed ORA on CpG probes in both DMRs and DMPs. Due to the large number of selected DMPs, as well as DMRs, returned in our analysis for ADCADN, we implemented two analyses: the first one tests for enriched terms for all results in both DMRs and DMPs; the second one tests for enriched terms in repeatedly selected probes for the purpose of determining which GO terms are most likely associated to the cohorts as a whole. Duplication counts were set to minimum of 2 for DMRs, i.e., probes in the range of two or more DMRs, and 6 for DMPs, or probes were selected in 6 or more cohorts. When using all DMRs, 515 GO terms were found to enrich the overlapping genes. However, only 20 terms were returned by our analysis using genes overlapping duplicated probes in the DMRs, 18 of which were also in the 515 GO terms in the first analysis. The top ten most significant terms for the DMR results in both tests are shown in Table 2. On the other hand, in the probe level, all DMPs were enriched by 18 GO terms while duplicated DMPs were enriched by 586 GO terms, 15 of which are common to both analyses. The top ten most significant terms are shown in Table 3. Interestingly, results show that the genes associated with the DMPs have functions related to biological processes such as nervous system development, developmental process, neurogenesis, generation of neurons and neuron differentiation. Subsequent ORA for KEGG pathways using duplicated overlaps with all DMR results revealed enrichment of genes related to the neuroactive ligand-receptor interaction pathway (p.adjust = 0.018) and the arachidonic acid metabolism pathway (p.adjust = 0.046), while using duplicated DMPs indicated enrichment in pathways in calcium signaling (p.adjust = 5.53e-6), axon guidance (p.adjust = 1.53e-4), focal adhesion (p.adjust = 1.53e-4), MAPK signaling (p.adjust = 1.60e-4), and cancer (p.adjust = 0.001) [data not shown]. Additionally, enrichment tests using the Human Phenotype Ontology database on the same gene lists only returned one significant result: autosomal dominant inheritance (p.adjust < 0.001) [data not shown].

Relationships between cohorts

All DMPs were used to calculate mean and median values for each cohort to identify overall trends in hypo- and hypermethylation. 37 (66.1%) cohorts had mean hypomethylation and 19 (33.9%) cohorts had mean hypermethylation. Using a stricter cutoff of at least a 5% change in mean methylation there were 12 (21.4%) hypomethylated cohorts and 10 (17.9%) hypermethylated (Figure 3A).

To investigate relationships across all cohorts without bias caused by the number of DMPs selected, clustering analysis was performed on the combined top n DMPs for each cohort as detailed in the Methods section, and visualized using a binary tree as illustrated in Figure 3B. The nodes or leaves of the tree is colored based on the global mean methylation difference for the corresponding cohort, while the size is scaled to the number of significant DMPs identified. The 56 cohorts can ultimately be clustered into two groups: one group along the branch of ADCADN (upper left), and the rest of the tree branches as the second group. Sub-clustering of the second group is also evident. At first glance, some patterns are evident in the clustering as most of the highly hypo/hypermethylated cohorts are close together. Furthermore, for the other sub-clusters, cohorts on the same branch are either in the same range of mean methylation difference or number of DMPs due to the similarities in either node size or node color. We also see groupings consistent with our previous analysis of these cohorts where conditions sharing similarities, phenotypically or genetically, were clustered together: such as Sotos, ICF, RMNS, BFLS and TBRS [V2], and RSTS1 and RSTS2 [V3?]. Cohort pairs were also observed generating terminal branches suggesting high level of similarity. Some of these cohort couples include BAFopathy and CSS9, which are both included in the BAF complex, ARTHS and SBBYSS, which are caused by mutations in KAT6 genes, and RSTS1 and RSTS2. To visualize global structure, we analyzed all cohorts using the most differentiating probes identified by random forest feature selection. Topological structures were approximated by UMAP and projected into two-dimensional and three-dimensional spaces as seen in Figure 4. Results of this analysis were concordant with the clustering analysis. Cohorts that are more

alike are closer together, such as RSTS1 and RSTS2, and ARTHS and GTPTS, which is also associated with a KAT6 gene. While we can see a large degree of overlap for several cohorts in the 2D projection, we also observe locally condensed independent groupings of the same cohorts in the 3D projection. This demonstrates the level of complexity of the overall structure of the data and the effectiveness of a small set of probes to distinguish them to a certain degree.

The network diagram of the probes shared between the cohorts illustrates several important details. First, the probes unique to the cohort (indicated by the self-loops) for the majority of the cohorts are hyper-methylated, while the probes, shared between the cohorts are hypomethylated practically in all cases, except of BEFAHRS (source node in the network), which shares probes mixed hyper-hypo status with MRD51, MRX93, GADEVS, BISS and DYT28 and except of Chr16p11.2del (source node) which shares mixed hyper-hypo status probes with DYT28. Mixed status of the shared probe is when the probe is hypermethylated in the first cohort, but hypomethylated in the other cohort and vice versa. Second, the ADCADN cohort that has a largest number of differentially methylated probes is not sharing of a significant proportion of probes with any other syndrome as other syndromes with high number of differentially methylated probes do. Third, although this is fully connected network in which each cohort shares at least one probe with all other syndromes, it is easy to distinguish groups of cohorts that share substantial number of the DM probes with each other. One such “triangle” is Sotos, RMNS and TBRS. While sharing a small number of probes by the cohorts can happen by chance, a substantial number of shared DM probes may indicate an underlying biological process that is common to all cohorts.

Discussion

Significant overlap in differentially methylated probes between disorders

Episignatures are used as clinical biomarkers and can act as a screen for patients undergoing first-tier diagnostic testing, or as a reflex test for patients with a variant of unknown significance or no variant identified [Jen Italy paper (under revision) & PMID 33547396). These methylation profiles are sensitive and specific to each disorder, and at times are also gene, region, or even variant specific (V3 paper ref). To achieve this specificity in 56 episignatures so far, DNA methylation profiles are optimized for use as a diagnostic biomarker by selecting the most differentially methylated probes and training against all other episignature samples [V3 paper] to generate the disorder classifier. In this study, we sought to look at methylation profiles of these disorders from a biological perspective, assessing all differentially methylated probes in each condition. The overwhelming observation is that there is significant overlap between all syndromes, highlighting the importance of training episignature classifiers against other disorders to allow for the proper detection of a specific methylation profile. Episignature detection would not be possible if only the most significantly differentiated probes for one disorder were considered as these probes would be present in many other disorder methylation profiles. This overlap is not necessarily surprising as the majority of these conditions are neurodevelopmental and often display similar or ambiguous clinical presentation that results in multiple disorders being considered in the differential diagnosis for a given patient. This highlights the importance of generating highlight specific episignatures; the usefulness of these biomarkers in the clinic depends on the ability to use a supervised algorithm that considers all detectable episignatures concurrently to avoid misclassification [PMID 32109418, V3 paper].

The cohorts that share a large numbers probes with many other disorders are those with a high number of significantly differentially methylated probes. When assessing the heatmap in figure 1, there are rows that are darker in colour when compared to the rest, indicating that that disorder shares a high percentage of probes with many of the disorders listed in the columns. Some examples include ADCADN, BEFAHRS, RMNS, Sotos, ICF1, and TBRS. These disorders are in the top 10 cohorts with highest

number of probes, but are also all involved in chromatin remodeling through DNA methylation (ADCADN BEFARHS, ICF1, TBRS), histone methylation (Sotos), or linker histones (RMNS). ADCADN also demonstrated the most unique probes, a consequence of the sheer number of detected differentially methylated probes. Many other disorders involving chromatin remodeling genes demonstrated significant overlap and a high number of differentially methylated probes, including FLHS, ICF_2_3_4, HVDAS_T, DYT28, and BFLS. Copy number variant disorders that include chromatin remodeling genes within the deletion and duplications (HMA, Sotos, Dup7, Williams) also demonstrate high degrees of overlap and number of differentially methylated probes, and reciprocating deletion and duplication syndromes (HMA vs Sotos and Dup7 vs Williams) show some overlap in probes but are dissimilar in the both the UMAP clustering, as well as lying on different branches in the leaf and tree diagram.

Sotos, TBRS, and RMNS overgrowth disorders show high overlap in probes (AUST18 and PCR2 show overlap)

As observed in both Figure 1 and 5 three overgrowth disorders, Sotos (caused by mutations in *NSD1)*, RNMS (caused by mutations in *HIST1H1E*), and TBRS (caused by mutations in DNMT3A), show significant overlap in differentially methylated probes. As mentioned previously, a "hypomethylation triangle" between these three syndromes. All three genes contribute to overgrowth phenotypes in patients and also are involved in chromatin remodeling directly (PMID: 28475857). NSD1 is a histone methyltransferase and functional studies have shown that loss of NSD1 results in redistribution of DNMT3A and reduced methylation at the expected regions (PMID: 31485078). Therefore, hypomethylation at shared probes may be a consequence of either lack of NSD1 recruitment of DNMT3A or the loss of DNMT3A altogether. *HIST1H1E*, a linker histone, has key roles in chromatin accessibility and regulation of gene expression, aligning its functionality with histone methyltransferase NSD1 and DNA methyltransferase. Additionally, a recent study found that these 3 genes, as well as 3 others (*CHD8, EED,* and *EZH2*), accounted for the mutations in 44% of patients in a large cohort of

assessing molecular etiology for overgrowth and intellectual disability (PMID: 28475857). Clinically variants in these genes present with a similar phenotype and molecularly they are involved in chromatin organization and gene expression. Therefore, the observed overlap in all probes, as well as their relatedness when assessing only top 500 probes, is another layer of functional evidence indicating these syndromes may have very similar pathological mechanisms. We also assessed the methylation patterns for the 3 epigenetic genes assessed in the study by Tatton-Brown *et al.* EZH2 and EED are components of the polycomb repressive complex 2 (PCR2) and mutations in both genes are included in the assessed PCR2 cohort. CHD8 is an ATP-dependent chromatin-remodeling factor and variants in this gene cause AUST18. Both the PCR2 and AUST18 cohorts exhibited small numbers of definitely methylated probes (less than 3000), however a large number of their differentially methylated probes (between 38 and 74%) are present in TBRS, Sotos, or RMNS probe lists, indicating common regions are impacted in these overgrowth syndromes.

Differences in methylation profiles in paralogous genes

Of the 56 cohorts assessed, 2 sets of paralogous genes are involved in multiple syndromes. Firstly, KAT6A and KAT6B are paralogous lysine acetyltransferases that form a complex with other proteins to control gene expression by histone acetylation (PMID: 33130515). Truncating mutations in the C-terminal transactivation domain of *KAT6A* cause ARTHS (PMID: 25728777), while truncating mutations in the proximal portion of the last exon its paralog, *KAT6B*, lead to a protein with no transactivation domain and cause GTPTS (PMID: 22715153). KAT6B mutations can also lead to another syndrome SBBYS. SBBYS mutations can result in nonsense-mediated decay, or more distally in the last exon (PMID: 22715153). SBBYS and ARTHS cluster more closely on the leaf and tree diagram, while GTPTS is a few branches away. On recent paper suggests that truncating mutations in the proximal portion of *KAT6B* lead to a gain of function in the protein (PMID: 22715153) and that this possible gain of function causes the phenotypes present in GTPTS but not in SBBYSS. This provides a possible reason as to why ARTHS

and SBBYS group more closely when compared to GTPTS. Mutations causing GTPTS and ARTHS fall in similar regions in the two genes (*KAT6B* and *KAT6A*, respectively), however the two genes only share 60% sequence. Further investigations assessing the protein changes caused by variants may provide further insight as to why ARTHS and SBBYSSS are more similar to each other than GTPTS, and if there truly is a gain of function within GTPTS variants that creates this dissimilarity.

Two other paralogs, *CREBBP* and *EP300,* are associated to 3 assessed cohorts. CREBBP and EP300 are transcriptional coactivators and histone acetyltransferases that interact with over 400 interacting proteins (PMID: 20110770). Mutations in *CREBBP* cause RSTS1 and mutations in *EP300* cause RSTS2, whereas variants in exon 30 and 31 of either gene can cause MKHK 1 and 2, respectively. Our cohort contains mutations in both genes that fall in the intrinsically disordered linker (ID4) region of these proteins (MKHK_ID4). One hypothesis is that missense mutations observed in MKHK patients result in gain of function within the proteins, resulting in a different phenotype compared to the loss of function observed in RSTS. Our data provides further functional evidence that these two syndromes have different pathological mechanisms with RSTS 1 and 2 showing high similarity to each other and mean hypomethylation and MKHK_ID4 exhibits overall hypermethylation and dissimilarity to RSTS1 and 2, as observed by the tree and leaf diagram. Gene expression analysis and functional assessment MKHK variants will provide more insight on the molecular mechanisms of these two syndromes.

Hypomethylated probes are most commonly shared between disorders

The conditions with the most differentially methylated probes also had the most unique probes, as outlined in the network diagram (Figure 5). The overlap in differentially methylated probes among all disorders is clear, however, the vast majority of overlapping probes between conditions are hypomethylated. Epigenetic changes to both DNA and histones, both transient and inherited, are essential to proper development, allowing for proper DNA expression that is cell-specific and temporal. Hypomethylation is indicative of gene activation, leading to the idea that overlapping probes may be

involved in genes that may be expressed inappropriately. Alternatively, hypermethylated probes tend to be unique to a given syndrome as observed by the red loop coming from a disorder node in Figure 5. Given that hypermethylation is associated with gene silencing, these disorder-specific probes may provide insight in genes that are repressed in the wrong cell type or at the wrong time and could contribute to a given phenotype.

Gene ontology analysis identifies enrichment in developmental and neurological pathways

The cohorts assessed in this study are Mendelian genetic disorders with established episignatures. The motivation of using methylation profiles as clinical biomarkers comes in part with the shared and non-specific clinical presentations exhibited in these syndromes, including the spectrum of neurodevelopmental delays and dysmorphic features (PMID: 29214565). Given many of these cohorts exhibit intellectual disability and developmental delay, it was not surprising that many of the top gene ontology terms for both DMPs and DMRs were involved in neurologic processes, such as chemical synaptic transmission, trans-synaptic signaling, synapse assembly, and glutamatergic synaptic transmission. Additionally, terms involved in developmental pathways and morphology, such as anatomical structure morphogenesis, cell-cell adhesion pathways, nervous system development, were enriched. Differential methylation within these genes, or near their promoters, implies possible alterations on gene expression, and with enrichment of DMPs and DMRs at CpG islands and promoters, these pathways may be impacted by abberant gene expression.

Though our DNA methylation data was generated from peripheral blood samples, this enrichment of neurodevelopmental pathways points towards the possible inappropriate expression of genes required for proper cortical development. Spatial and temporal control of gene expression through DNA methylation is a highly dynamic process during development and many of the cohorts studied are involved its regulation. A recent review highlights the importance of DNA methylation in neuronal development within a set of neurodevelopmental syndromes, many of which are represented by our

cohorts(https://doi.org/10.3389/fnins.2021.776809). Further analysis of specific genes impacted, as well as direction of methylation change in the context of a given disorder, will provide further insight into possible underlying genetic pathways that may contribute to a given syndrome phenotype. Gene expression analysis would also further solidify the impact of these methylation changes on the genes in question.

Conclusions:

References

[To do]

Figure titles and legends

Figure 1: Differentially methylated probes found shared between multiple cohorts. A. Percent of probes that are shared between each pair of cohorts. For each pair the colors indicate the percent of the top/bottom cohort's probes that are also found in the left/right cohort's probes. **B.** Probes that are shared between each pair of cohorts. Each labelled sector represents one cohort. The thickness of connecting line represents the number of probes shared between the two cohorts.

Figure 2: DMPs and DMRs annotated in the context of CpG islands and genes. A. DMPs annotated in the context of CpG islands. **B.** DMRs annotated in the context of CpG islands. **C.** DMPs annotated in the context of genes. **D.** DMRs annotated in the context of genes. For CpG plots: Island, CpG islands; Shore, within 0-2kb of a CpG island boundary; Shelf, within 2-4kb of a CpG island boundary; Inter_CGI, all other regions in the genome. For gene context plots: Promoter, 0-1kb upstream of the transcription start site; Promoter+, 1-5kb upstream of the transcription start site; CDS, coding sequence. For DMP plots, the

Probes column represents the “background” or “default” distribution of all 450K array probes after initial filtering and used as input for DMP analysis. For DMR analysis, the numbers above each bar indicate the number of DMRs identified for each cohort. The following cohorts had no detected DMRs: AUTS18, BAFopathy, CSS4_c.2650, CSS9, Kabuki, KDM4B, MRX93.

Figure 3: Relationships between cohorts. A. Methylation differences of all differentially methylated probes for each cohort, sorted by mean methylation. Each circle represents one probe. Red lines indicate mean methylation, yellow lines indicate median methylation. **B.** Tree and leaf visualization of Euclidean clustering of 56 cohorts using the top n DMPs for each group, where n = min (# of DMPs, 500). Cohort samples were aggregated using the median value of each probe within a group. A leaf node represents a cohort, with node sizes illustrating relative scales of the number of selected DMPs for the corresponding cohort, and node colors are indicative of the global mean methylation difference.

Figure 4: UMAP visualization of 56 cohorts using the most differentiating probes. A-C. UMAP results projected to 3-dimensional space and snapshot from different perspectives. **D.** UMAP results projected to 2-dimensional plane.

Figure 5: Differentially methylated probe sharing between the 56 cohorts. Network diagram shows cohorts connected by edges representing probes shared between them. Edges represent hyper-hyper, hyper-hypo, hypo-hyper and hypo-hypo type of connections in which an edge’s width is proportional to the total number of probes shared (in the range from 1 to 138,727). Probes unique to the cohort are represented by the self-loop. The total number of the differentially methylated probes in the cohort is simultaneously coded by a color and a height of the ellipse representing the cohort. The cohorts with a substantial number of probes are color-coded by an increasing gradient of purple; while cohorts with a small number of differentially methylated probes are color-coded by a yellow-white gradient.

Tables

Table 1: List of cohorts.

Syndrome	Signature Abbreviation	Underlying gene or region	OMIM	Samples	Probes	Category
X-linked alpha-thalassemia/mental retardation syndrome (ATRX)	ATRX	*ATRX*	301040	30	8666	SWI/SNF chromatin remodeling
Arboleda-Tham syndrome (ARTHS)	ARTHS	*KAT6A*	616268	18	4487	Histone acetyltransferase
Autism, susceptibility to, 18 (AUTS18)	AUTS18	*CHD8*	615032	28	2319	Transcription factor
Beck-Fahrner syndrome (BEFAHRS)	BEFAHRS	*TET3*	618798	16	30391	DNA demethylase
Blepharophimosis Intellectual disability SMARCA2 Syndrome	BISS	*SMARCA2*	619293	12	10186	SWI/SNF chromatin remodeling
Börjeson-Forssman-Lehmann syndrome (BFLS)	BFLS	*PHF6*	301900	14	12321	Transcription factor
Cerebellar ataxia, deafness, and narcolepsy, autosomal dominant (ADCADN)	ADCADN	*DNMT1*	604121	5	151848	DNA methyltransferase
CHARGE syndrome	CHARGE	*CHD7*	214800	74	840	Transcription factor
Chr16p11.2 deletion syndrome, 593-KB	Chr16p11.2del	Chr16p11.2del	611913	18	10105	CNV
Coffin-Siris syndrome-1,2 (CSS1,2)	CSS_c.6200*	*ARID1B* *ARID1A*	135900 614607	4	3451	SWI/SNF chromatin remodeling
Coffin-Siris syndrome-1,2,3,4; Nicolaides-Baraitser syndrome (CSS12,3,4; NCBRS)	BAFopathy	*ARID1B* *ARID1A* *SMARCB1* *SMARCA4* *SMARCA2*	135900 614607 614608 614609 601358	124	1015	SWI/SNF chromatin remodeling
Coffin-Siris syndrome-4 (CSS4)	CSS4_c.2656*	*SMARCA4*	614609	3	464	SWI/SNF chromatin remodeling
Coffin-Siris syndrome-9 (CSS9)	CSS9	*SOX11*	615866	13	430	Transcription factor
Cohen-Gibson syndrome; Weaver syndrome (COGIS; WVS)	PRC2	*EED* *EZH2*	617561 277590	8	2444	Histone deacetylase Histone methyltransferase
Cornelia de Lange syndromes 1,2,3,4 (CDLS1,2,3,4)	CdLS	*NIPBL* *SMC1A* *SMC3* *RAD21*	122470 300590 610759 614701	70	3623	Chromosome cohesion/condensation; DNA repair (RAD21)
Down syndrome	Down	Chr21 trisomy	190685	40	24712	CNV
Dystonia 28, childhood-onset (DYT28)	DYT28	*KMT2B*	617284	10	25260	Histone methyltransferase
Epileptic encephalopathy, childhood-onset (EEOC)	EEOC	*CHD2*	615369	9	5284	Transcription factor
Floating Harbour syndrome (FLHS)	FLHS	*SRCAP*	136140	21	26811	SWI/SNF chromatin remodeling
Gabriele-de Vries syndrome (GADEVS)	GADEVS	*YY1*	617557	10	4380	Transcription factor
Genitopatellar syndrome (see also Ohdo syndrome, SBBYSS variant) (KAT6B)	GTPTS	*KAT6B*	606170	4	3008	Histone acetyltransferase
Helsmoortel-van der Aa syndrome (HVDAS)	HVDAS_C*	*ADNP*	615873	14	6986	Transcription factor
Helsmoortel-van der Aa syndrome (HVDAS)	HVDAS_T*	*ADNP*	615873	21	16756	Transcription factor
Hunter McAlpine craniosynostosis syndrome	HMA	Chr5q35-qter dup	601379	8	17948	CNV

Immunodeficiency-centromeric instability-facial anomalies syndrome 1 (ICF1)	ICF_1	*DNMT3B*	242860	8	38656	DNA methyltransferase
Immunodeficiency-centromeric instability-facial anomalies syndromes 2,3,4 (ICF2,3,4)	ICF_2_3_4	*ZBTB2* *CDCA7* *HELLS*	614069 616910 616911	7	66568	Transcription factor c-Myc responsive gene SWI/SNF chromatin remodeling
Intellectual developmental disorder with seizures and language delay (IDDSELD)	IDDSELD	*SETD1B*	619000	11	5264	Histone methyltransferase
Kabuki syndromes 1,2 (KABUK1,2)	Kabuki	*KMT2D* *KDM6A*	147920 300867	191	3749	Histone methyltransferase Histone demethylase
KDM2B-related syndrome	KDM2B	*KDM2B*	unofficial	9	3632	Histone demethylase
Autosomal dominant intellectual developmental disorder-65 (MRD65)	KDM4B	*KDM4B*	619320	6	279	Histone demethylase
Kleefstra syndrome 1 (KLEFS1)	Kleefstra	*EHMT1*	610253	32	4124	Histone methyltransferase
Koolen de Vreis syndrome (KDVS)	KDVS	*KANSL1*	610443	16	6490	Histone acetylation
Luscan-Lumish syndrome (LLS)	LLS	*SETD2*	616831	4	2405	Histone methyltransferase
Menke-Hennekam syndromes 1,2 (MKHK1,2)	MKHK_ID4*	*CREBBP* *EP300*	618332 618333	13	2570	Histone acetyltransferase
Intellectual developmental disorder, X-linked, syndromic, Armfield type (MRXSA)	MRXSA	*FAM50A*	300261	6	4618	mRNA splicing
Mental retardation, autosomal dominant 23 (MRD23)	MRD23	*SETD5*	615761	25	2795	Histone methyltransferase
Mental retardation, autosomal dominant 51 (MRD51)	MRD51	*KMT5B*	617788	7	19803	Histone methyltransferase
Intellectual developmental disorder, X-linked 93 (MRX93)	MRX93	*BRWD3*	300659	11	16894	Transcription factor
Intellectual developmental disorder, X-linked 97 (MRX97)	MRX97	*ZNF711*	300803	18	3770	Transcription factor
Intellectual developmental disorder, X-linked syndromic, Nascimento-type (MRXSN)	MRXSN	*UBE2A*	300860	4	6065	Enzyme
Intellectual developmental disorder, X-linked, Snyder-Robinson type (MRXSSR)	MRXSSR	*SMS*	309583	17	4062	Enzyme
Intellectual developmental disorder, X-linked, syndromic, Claes-Jensen type (MRXSCJ)	MRXSCJ	*KDM5C*	300534	58	5013	Histone demethylase
Myopathy, lactic acidosis, and sideroblastic anemia 2 (MLASA2)	MLASA2	*YARS2*	613561	11	2304	tRNA synthesis
Ohdo syndrome, SBBYSS variant (SBBYSS)	SBBYSS	*KAT6B*	603736	9	1956	Histone acetyltransferase
Phelan-McDermid syndrome (PHMDS)	PHMDS	Chr22q13.3del	606232	11	17581	CNV
Rahman syndrome (RMNS)	RMNS	*HIST1H1E*	617537	9	26101	Linker histone
Renpenning syndrome (RENS1)	RENS1	*PQBP1*	309500	8	5228	mRNA splicing
Rubinstein-Taybi syndrome 1 (RSTS1)	RSTS1	*CREBBP*	180849	37	5279	Histone acetyltransferase
Rubinstein-Taybi syndrome 2 (RSTS2)	RSTS2	*EP300*	613684	29	7998	Histone acetyltransferase
Sotos syndrome 1 (SOTOS1)	Sotos	*NSD1*	117550	69	43022	Histone methyltransferase
Tatton-Brown-Rahman syndrome (TBRS)	TBRS	*DNMT3A*	615879	30	35130	DNA methyltransferase

Velocardiofacial syndrome (VCFS)	VCFS	Chr22q11.2del	192430	47	4134	CNV
Wiedemann-Steiner syndrome (WDSTS)	WDSTS	*KMT2A*	605130	52	4777	Histone methyltransferase
Williams-Beuren deletion syndrome (WBS)	Williams	Chr7q11.23del	194050	22	13131	CNV
Williams-Beuren duplication syndrome (Chr7q11.23 duplication syndrome)	Dup7	Chr7q11.23dup	609757	13	6963	CNV
Wolf-Hirschhorn syndrome (WHS)	WHS	Chr4p16.13del	194190	17	7838	CNV

* Episignatures which encompass a specific region or variant within a gene.

Table 2: Top 10 most significant GO terms from enrichment analysis of DMRs.

Using all CpG sites in the selected DMRs of all signatures.

Ontology	ID	Description	GeneRatio	Adjusted p value
BP	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	77/2272	2.281E-22
BP	GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	100/2272	2.4689E-20
BP	GO:0009653	anatomical structure morphogenesis	482/2272	2.0273E-12
BP	GO:0009887	animal organ morphogenesis	222/2272	1.7703E-11
BP	GO:0007268	chemical synaptic transmission	160/2272	3.7073E-11
BP	GO:0098916	anterograde trans-synaptic signaling	160/2272	3.7073E-11
BP	GO:0003002	regionalization	94/2272	3.7073E-11
BP	GO:0022610	biological adhesion	289/2272	4.2794E-11
BP	GO:0099537	trans-synaptic signaling	160/2272	5.9064E-11
BP	GO:0007155	cell adhesion	286/2272	1.0448E-10

Using duplicated CpG sites in the selected DMRs of all signatures.

Ontology	ID	Description	GeneRatio	Adjusted p value
BP	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	46/543	1.9945E-27
BP	GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	54/543	4.8572E-25
BP	GO:0016339	calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	11/543	0.00010667
BP	GO:0007416	synapse assembly	19/543	0.00134764
BP	GO:0035249	synaptic transmission, glutamatergic	12/543	0.02574892
BP	GO:0048232	male gamete generation	34/543	0.03108224
BP	GO:0007276	gamete generation	39/543	0.03735384
CC	GO:0034702	ion channel complex	23/562	0.01432234
CC	GO:0034703	cation channel complex	19/562	0.01432234
CC	GO:1902495	transmembrane transporter complex	24/562	0.01432234

Table 3: Top 10 most significant GO terms from enrichment analysis of DMPs.

Using all selected DMPs in all signatures.

Ontology	ID	Description	GeneRatio	Adjusted p value
BP	GO:0016043	cellular component organization	5977/16538	6.9304E-05
BP	GO:0071840	cellular component organization or biogenesis	6164/16538	6.9304E-05
BP	GO:0007399	nervous system development	2237/16538	0.00149203
BP	GO:0051179	localization	6254/16538	0.01290062
BP	GO:0022008	neurogenesis	1546/16538	0.01352787
BP	GO:0032502	developmental process	6000/16538	0.02559775
BP	GO:0030182	neuron differentiation	1294/16538	0.02559775
BP	GO:0051234	establishment of localization	4851/16538	0.02559775
BP	GO:0048699	generation of neurons	1435/16538	0.02636818
BP	GO:1901564	organonitrogen compound metabolic process	6131/16538	0.02727857

Using DMPs selected in more than five signatures.

Ontology	ID	Description	GeneRatio	Adjusted p value
BP	GO:0007399	nervous system development	1482/9330	9.017E-23
BP	GO:0048856	anatomical structure development	3431/9330	3.9082E-21
BP	GO:0032502	developmental process	3678/9330	2.9784E-20
BP	GO:0007275	multicellular organism development	3148/9330	2.2164E-18
BP	GO:0048731	system development	2840/9330	3.0198E-18
BP	GO:0048468	cell development	1279/9330	3.0198E-18
BP	GO:0009653	anatomical structure morphogenesis	1645/9330	3.6223E-17
BP	GO:0022008	neurogenesis	1025/9330	1.8849E-15
BP	GO:0048699	generation of neurons	956/9330	3.9628E-15
BP	GO:0030182	neuron differentiation	868/9330	8.1588E-15