ARTICLE

DNA Methylation Signature for *EZH2* Functionally Classifies Sequence Variants in Three PRC2 Complex Genes

Sanaa Choufani,[1] William T. Gibson,[2,3] Andrei L. Turinsky,[1,4] Brian H.Y. Chung,[5] Tianren Wang,[1] Kopal Garg,[1] Alessandro Vitriolo,[6] Ana S.A. Cohen,[2,3,7] Sharri Cyrus,[2,3] Sarah Goodman,[1] Eric Chater-Diehl,[1] Jack Brzezinski,[1,8] Michael Brudno,[1,4,9] Luk Ho Ming,[10] Susan M. White,[11] Sally Ann Lynch,[12] Carol Clericuzio,[13] I. Karen Temple,[14] Frances Flinter,[15] Vivienne McConnell,[16] Tom Cushing,[17] Lynne M. Bird,[18] Miranda Splitt,[19] Bronwyn Kerr,[20] Stephen W. Scherer,[1,21] Jerry Machado,[22] Eri Imagawa,[23] Nobuhiko Okamoto,[24] Naomichi Matsumoto,[23] Guiseppe Testa,[6,25] Maria Iascone,[26] Romano Tenconi,[27] Oana Caluseriu,[28] Roberto Mendoza-Londono,[29,30] David Chitayat,[29,30,31] Cheryl Cytrynbaum,[1,29,32] Katrina Tatton-Brown,[33] and Rosanna Weksberg[1,29,30,32,34,*]

Weaver syndrome (WS), an overgrowth/intellectual disability syndrome (OGID), is caused by pathogenic variants in the histone methyltransferase *EZH2*, which encodes a core component of the Polycomb repressive complex-2 (PRC2). Using genome-wide DNA methylation (DNAm) data for 187 individuals with OGID and 969 control subjects, we show that pathogenic variants in *EZH2* generate a highly specific and sensitive DNAm signature reflecting the phenotype of WS. This signature can be used to distinguish loss-of-function from gain-of-function missense variants and to detect somatic mosaicism. We also show that the signature can accurately classify sequence variants in *EED* and *SUZ12*, which encode two other core components of PRC2, and predict the presence of pathogenic variants in undiagnosed individuals with OGID. The discovery of a functionally relevant signature with utility for diagnostic classification of sequence variants in *EZH2*, *EED*, and *SUZ12* supports the emerging paradigm shift for implementation of DNAm signatures into diagnostics and translational research.

Introduction

Weaver syndrome (WS [MIM: 277590]), an overgrowth/intellectual disability syndrome (OGID), is characterized by pre- and postnatal overgrowth, accelerated osseous maturation, characteristic craniofacial features, and variable intellect.[1] Sequence variants in *EZH2* (Enhancer of Zeste, *Drosophila*, homolog 2 [MIM: 601573]) are the primary reported cause of WS,[2,3] accounting for more than 90% of individuals. *EZH2* encodes part of the catalytic component

[1]Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada; [2]British Columbia Children's Hospital Research Institute, Vancouver, BC V5Z 4H4, Canada; [3]Department of Medical Genetics, University of British Columbia, Vancouver, BC V6H 3N1, Canada; [4]Centre for Computational Medicine, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada; [5]Pediatrics and Adolescent Medicine, Queen Mary Hospital and Hong Kong Children's Hospital, The University of Hong Kong, 999077 Hong Kong; [6]Department of Oncology and Hemato-oncology, University of Milan, Milan 20122, Italy; Laboratory of Stem Cell Epigenetics, IEO, European Institute of Oncology, IRCCS, Milan 20139, Italy; [7]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [8]Division of Haematology and Oncology, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada; Department of Pediatrics, University of Toronto, Toronto, ON M5G 1X8, Canada; [9]Department of Computer Science, University of Toronto, Toronto, ON M5S 3H5, Canada; [10]Clinical Genetic Service, Department of Health, 999077 Hong Kong; [11]Victorian Clinical Genetics Service, Murdoch Children's Research Institute, Melbourne, VIC 3052, Australia; Department of Paediatrics, University of Melbourne, Melbourne, VIC 3052, Australia; [12]Department of Clinical Genetics, Temple Street Children's University Hospital, Dublin, D01 XD99, Ireland; [13]Pediatric Genetics, University of New Mexico, Albuquerque, NM 87131, USA; [14]Faculty of Medicine, University of Southampton and the Wessex Clinical Genetics Service, University Hospital Southampton NHS Foundation Trust, Southampton SO16 6YD, UK; [15]Department of Clinical Genetics, Guy's and St Thomas' NHS Foundation Trust, London SE1 9RT, UK; [16]Northern Ireland Regional Genetics Service, Belfast City Hospital, Belfast BT9 7AB, UK; [17]Pediatric Genetics, University of New Mexico, Albuquerque, NM 87131, USA; [18]Department of Pediatrics, University of California, San Diego, San Diego, CA 92093, USA; Division of Genetics, Rady Children's Hospital of San Diego, San Diego, CA 92123, USA; [19]Northern Genetics Service, Institute of Genetics Medicine, Newcastle upon Tyne NE1 3BZ, UK; [20]Division of Evolution and Genomic Sciences, School of Biological Sciences, University of Manchester, Manchester M13 9PL, UK; Manchester Centre for Genomic Medicine, Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Sciences Centre, Manchester M13 9WL, UK; [21]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A1, Canada; The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON M5G 1X8 Canada; McLaughlin Centre, University of Toronto, Toronto, ON M5S 1A1, Canada; [22]PreventionGenetics, Marshfield, WI 54449, USA; [23]Department of Human Genetics, Yokohama City University Graduate School of Medicine, Fukuura 3-9, Kanazawa-ku, Yokohama 236-0004, Japan; [24]Department of Medical Genetics, Osaka Medical Center and Research Institute for Maternal and Child Health, Osaka 594-1101, Japan; [25]Human Technopole, Center for Neurogenomics, Via Cristina Belgioioso 171, Milan 20157, Italy; [26]Laboratorio di Genetica Medica, ASST Papa Giovanni XXIII, Piazza OMS 1, 24127 Bergamo, Italy; [27]Dipartimento Pediatria, University of Padova, Via Giustiani 3, 35128 Padova, Italy; [28]Department of Medical Genetics, University of Alberta, Edmonton, AB T6G 2H7, Canada; The Stollery Pediatric Hospital, Edmonton, AB T6G 2H7, Canada; [29]Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada; [30]Department of Pediatrics, University of Toronto, Toronto, ON M5S 1A1, Canada; [31]Prenatal Diagnosis and Medical Genetics Program, Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada; [32]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A1, Canada; [33]St George's University Hospitals NHS Foundation Trust, London SW17 0QT, UK; St George's, University of London, London SW17 0RE, UK; Section of Cancer Genetics, Institute of Cancer Research, Surrey SM2 5NG, UK; [34]Institute of Medical Sciences, University of Toronto, Toronto, ON M5S 1A8, Canada
*Correspondence: rweksb@sickkids.ca
https://doi.org/10.1016/j.ajhg.2020.03.008.
© 2020 The Author(s). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

of the polycomb repressive complex 2 (PRC2), which regulates genome-wide chromatin structure and gene expression through methylation of lysine 27 of histone H3; this mark drives chromatin condensation and transcriptional repression.[4] EZH2 in combination with EED (embryonic ectoderm development) and SUZ12 (SUZ12 polycomb repressive complex 2 subunit) form the core complex of PRC2. Recently, it has been established that pathogenic variants in *EED* (MIM: 605984) and *SUZ12* (MIM: 606245) cause two clinically overlapping OGID syndromes, Cohen-Gibson (MIM: 617561) and SUZ12-related overgrowth syndrome, respectively.[5,6]

The phenotypic consequences of pathogenic variants in *EZH2* are variable; clinical diagnosis is therefore challenging in the absence of the characteristic facial phenotype. The clinical features can be age dependent, and in the largest case series to date, fewer than half (47%) of individuals with *EZH2* pathogenic variants were considered to have the classic Weaver syndrome facial gestalt.[1] Tall stature, with a height at least two standard deviations (SDs) above the mean, is the most consistent EZH2-related clinical feature, present in ~90% of affected individuals. Intellectual disability is also common, present in ~80%, but can be very mild. However, tall stature and intellectual disability are nonspecific and generally common, which limits their utility as discriminating clinical features of WS and its related disorders.[1]

Several studies have shown that WS-associated pathogenic *EZH2* variants cause a loss rather than a gain of enzymatic activity. Cohen et al.[7] found reduced histone methyltransferase activity using an *in vitro* assay. In a mouse model of WS, decreased di- and trimethyl-H3K27 were identified in homozygous and heterozygous embryos, thereby supporting a role for reduced methyltransferase function as the cause of WS.[8]

We and others have recently found that disorders caused by pathogenic sequence variants in epigenes (genes encoding proteins that demonstrate "epigenetic" functions) may cause cells to change their typical DNA methylation (DNAm) and transcriptional profiles. Individuals with pathogenic variants in some epigenes exhibit disorder-specific signatures comprised of genome-wide, multilocus DNA methylation alterations. These signatures can be used to functionally classify sequence variants, with high sensitivity and specificity and to improve our understanding of the pathophysiology of the associated genetic disorder.[9–15] We have also shown that pathogenic sequence variants in clinically overlapping, genetically distinct disorders may lead to similar epigenetic patterns of regulatory disruption. Our group has recently reported this phenomenon in Kabuki syndrome (KS) types 1 and 2, for which sequence variants in different causative genes (*KMT2D* [MIM: 602113], *KDM6A* [MIM: 300128]) share an epigenetic disease signature[10] and in BAF complex-associated disorders, we and others have reported similar findings.[12,15]

In this study, we identified a genome-wide DNAm signature associated with pathogenic variants in *EZH2*. We were successful in using this signature to detect somatic mosaicism for *EZH2* variants and to classify both gain-of-function (GoF) and loss-of-function (LoF or hypomorphic) sequence variants in *EZH2*. Further, this *EZH2* signature can be used to identify and classify pathogenic sequence variants in other genes in the PRC2 core complex, namely *EED* and *SUZ12*. These data provide evidence for convergent molecular signatures for three PRC2 complex genes and highlight the functional relevance of this DNAm signature for predicting variant pathogenicity, detecting somatic mosaicism, and most importantly, for discriminating different functional effects of sequence variants.

Subjects and Methods

Cohort with *EZH2* Variants

The research was approved by the Research Ethics board at The Hospital for Sick Children (REB# 1000038847) and consent was obtained from participating subjects and/or their parents or guardians. Peripheral blood samples from subjects with *EZH2* variants (n = 40) were used in this study, of whom eight unrelated subjects were used as a discovery set with a clinical diagnosis of Weaver syndrome and pathogenic sequence variants in *EZH2*, seven previously published in Gibson et al.[2] and Cohen and Gibson[7] and one unpublished case (see Table S1). Another unrelated cohort with clinical diagnosis of WS consisting of eight previously published WS individuals with pathogenic *EZH2* sequence variant as part of the Childhood Overgrowth Consortium in the UK[1] was used as an independent validation set (Table S2). This *EZH2* variant cohort included a familial case series with a pathogenic *EZH2* sequence variant: GenBank: NM_004456.4; c.466A>G (p.Lys156Glu). The proband has a family history of overgrowth, and the *EZH2* sequence variant segregated with the overgrowth phenotype over three generations. The family included the proband, two affected siblings, the mother, and the maternal grandfather.[3] An additional test cohort for *EZH2* variant classification consisted of 19 samples with *EZH2* variants (see Table S9), collected from PreventionGenetics, USA; from British Columbia Children's Hospital at the University of British Columbia, Vancouver, Canada; from St George's University Hospitals NHS Foundation Trust, London, UK; from the Department of Human Genetics, Yokohama City University Graduate School of Medicine, Fukuura, Japan; from Laboratorio di Genetica Medica, ASST Papa Giovanni XXIII, Piazza OMS, Bergamo, Italy; from the Department of Medical Genetics, University of Alberta, Edmonton, Canada; and from the Victorian Clinical Genetics Services, Murdoch Children's Research Institute, Royal Children Hospital, Victoria, Australia.

Cohort with *EED* Variants

Three individuals with *de novo EED* pathogenic variants were previously reported[5–7] and are described in Table S9.

Cohort with *SUZ12* Variants

Five individuals with *SUZ12* variants were recruited to the study including one previously reported familial case subject comprised of the proband and affected father.[6] Two individuals with missense variants in *SUZ12* recruited from British Columbia Children's Hospital at the University of British Columbia, Vancouver, Canada and from the Department of Pediatrics and Adolescent

Medicine from the University of Hong Kong. One autism-affected case subject with 17q11.2 dup overlapping *SUZ12* identified at the Hospital for Sick Children, Toronto, Canada. Full description of the sequence variants is given in Table S9.

All gene variant annotations as well as *in silico* prediction using PolyPhen-2, SIFT, and MutationTaster were generated using Alamut visual 2.11. CADD scores were obtained using CADD database v.1.4.

Control Cohort

Genomic DNA from peripheral blood was obtained on a set of controls (n = 23) selected as age- and sex-matched neurotypical controls to the *EZH2* discovery set (Table S3). In addition, 148 controls were used to determine the specificity of the DNAm signature. These control samples were obtained from the POND Network, The Hospital for Sick Children, and The University of Michigan (Dr. Greg Hanna).[16] Neurotypical was defined as healthy and developmentally normal on formal cognitive/behavioral assessments (samples from POND and The University of Michigan) or via physician/parental screening questionnaires (Hospital for Sick Children). Additional blood controls were obtained from publicly available datasets taken from the Gene Expression Omnibus. In total, we used data from 718 unrelated subjects from the general population without clinically obvious neurodevelopmental phenotypes, who had DNA extracted from peripheral blood and had undergone profiling with the Illumina 450k array (GEO: GSE54670, GSE54399, GSE51245, GSE89353, GSE36064, GSE128801, GSE53045, GSE40279, GSE42861).[17–25] We restricted subjects to those age <50 years to match our WS cohort and excluded one control sample, as it presented as outlier based on principal component analysis (PCA) of autosomal probes. For detailed information see Table S7 and Figure S5.

DNAm Array Processing

Genome-wide DNAm profiling on control and affected subjects matched for age and sex was performed at The Center for Applied Genomics (TCAG), SickKids Research Institute, Toronto, ON, Canada. Genomic DNA from each subject was sodium bisulfite converted using the EpiTect Bisulfite Kit (EpiTect PLUSBisulfite Kit, QIAGEN), according to the manufacturer's protocol. Modified genomic DNA was then processed and analyzed on the Infinium HumanMethylationEPIC BeadChip (Illumina 850K) according to the manufacturer's protocol.[26] The distribution of the samples on the arrays was randomized for both disease and control samples. All signature-derivation samples (WS and controls) were run in the same batch.

Quality Control and Normalization

The raw IDAT files were converted into β-values, which represent DNAm levels as a percentage (between 0 and 1), using the *minfi* Bioconductor package in *R*. Data preprocessing included filtering out non-specific probes (41,135 probes); probes with detection p value > 0.05 in more than 25% of the samples (824 probes); probes with single-nucleotide polymorphic sites (SNPs) located within 10 bp of the targeted CpG site or a single base extension as well as probes near SNPs with minor allele frequencies above 1% (n = 29,958); probes with raw beta = 0 or 1 in > 0.25% of samples (n = 21); non-CpG probes (n = 2,932); and X and Y chromosome probes (n = 19,627) for a total of 91,343 probes removed and a total of n = 774,516 probes remaining for differential methylation analysis. Standard quality control metrics in *minfi* were used, including median intensity QC plots, density plots, and control probe plots: all samples passed quality control and were included in the study.

Differential DNAm Analysis

The analysis was performed using our previously published protocol.[10] Differential DNAm analysis between WS and controls was performed at 774,516 CpG sites using beta scores, which represent DNAm levels as a percentage (between 0 and 1). The β-value from each sample at the remaining 774,516 CpGs was used for downstream analysis and generation of a DNAm signature. Beta values were logit transformed to M-values using the following equation: log2(beta/(1-beta)). A linear regression modeling using *limma* package[27] was used to identify the differentially methylated probes. We estimated blood cell proportions using Houseman's algorithm and the Bioconductor packages, *minfi* and *FlowSorted.Blood.EPIC*. This method generates proportions of $CD8^+$ T cells, $CD4^+$ T, natural killer, B cells, monocytes, and granulocytes (mainly neutrophils, Neu) (Table S13).[28] These estimated values for each cell component were incorporated into the model matrix of the regression analysis as covariates along with sex and age. The analysis was done on the discovery set of 8 WS and 23 controls per the following regression model: DNAm was regressed against sex+age+CD8T+CD4T+NK+Bcell+Mono+Neu for each CpG site. The generated p values were corrected for multiple testing using the Benjamini-Hochberg method. A significant difference in DNAm between WS and control samples for each CpG site was required to meet the cutoffs of Benjamini-Hochberg adjusted p values < 0.05 and $|\Delta\beta| \geq 0.10$ (10% methylation differences) as previously reported.[9,10]

Generation of Disease Score Classification Model using Correlation Analysis

We used a previously described pipeline for generating disease scores using an established disease-specific DNAm signature.[9,10] At each of the 229 signature CpGs, a median DNAm level was computed across the WS individuals (n = 8) used to generate the signature, resulting in a reference profile. Similarly, a robust median-DNAm reference profile for the signature controls (n = 23) was created. The classification of each additional gene variant or control DNAm sample was based on extracting a vector $BR_{sig}R$ of its DNAm values in the signature CpGs and comparing $BR_{sig}R$ to the two reference profiles computed above. *EZH2* score was defined as: $EZH2\ score = r(BR_{sig}R, WS\ profile) - r(BR_{sig}R, control\ profile))$ where *r* is the Pearson correlation coefficient. A classification model was developed based on scoring each new DNAm sample using the *EZH2* Score: a test sample with a positive score is more similar to the WS reference profile based on the signature CpGs and is therefore classified as "pathogenic" whereas a sample with a negative score is more similar to the control-blood reference profile and is classified as "benign." The classification is implemented in *R*. To test specificity, EPIC array data from 148 additional neurotypical controls were scored and classified. To test sensitivity, eight additional unrelated WS individuals with *EZH2* pathogenic variants were scored and classified.

Generation of Machine Learning Model for Variant Classification

Using the R package caret, probes with very similar methylation patterns with correlation greater or equal to 90% (redundant probes) were removed as we previously described[10] leading to a subset of 119 CpGs. Next, we developed a machine-learning

model, a support vector machine (SVM) model with linear kernel that had been trained on the significant CpG sites from the discovery cohorts after further filtering to remove redundant CpGs. The model was set to the "probability" mode to generate SVM scores ranging between 0 and 1 (or 0% and 100%), thus classifying samples as "WS" (high scores) or "not-WS" (low scores). This SVM model was built as a tool for the classification of variants in *EZH2*, *EED*, and *SUZ12*.

Identification of Differentially Methylated Regions

To identify differentially methylated regions (DMRs) that are associated with *EZH2* variants, we used the bumphunting method[29] which strengthens the detection of regional differences by combining differential-methylation patterns across neighboring CpG sites.[30] The bumphunting design matrix accounted for the potential confounding effects of sex and age and was used to identify regions with consecutive CpGs no more than 0.5 Kb apart and an average regional methylation difference $|\Delta\beta| \geq 10\%$. Statistical significance was established using 1,000 randomized bootstrap iterations, as is recommended in the Bioconductor *bumphunter* package when accounting for confounders. The resulting DMRs were post-filtered to retain only those with p value < 0.05 and a length (number of consecutive CpGs) of at least five CpGs. The analysis was performed on the same sets of case and control subjects used as the discovery cohort.

Genomic and Gene-Set Enrichment Analyses

For genomic enrichment analysis, the list of 229 CpG sites (foreground genomic regions) was submitted to GREAT (Genomic Regions Enrichment of Annotations Tool)[31] using the default settings. We used the set of CpG sites after *minfi* probe quality control (n = 774,516) as the background genomic regions.

Gene-set enrichment analysis was performed using g: Profiler to identify Gene Ontology (GO) Biological Process terms overrepresented in the annotations of genes overlapping the differentially methylated CpGs. The enriched GO terms with Benjamini-Hochberg corrected p values < 0.05 were reported. The redundant GO terms were reduced and visualized as interactive networks using EnrichmentMap app on the Cytoscape platform as previously described.[32]

Enzymatic Activity of EZH2

The luminescence assay to determine EZH2 enzymatic activity for the p.Ala738Thr sequence variant was performed at BPS Bioscience as follows: a 50 μl reaction mix containing 50 μM S-adenosylmethionine, EZH2 enzyme (as part of an artificially assembled PRC2 complex derived from baculovirus expression vectors), and 20 mM phosphate buffer (pH 7.4), 0.05% Tween-20 HMT buffer 2 (BPS #52170) was added to wells coated with the substrate. Incubation was done for 1 h at room temperature, then antibody against methylated lysine 27 (K27) residue of histone H3 was added and incubated for 1 h. Secondary horseradish peroxidase (HRP)-labeled antibody was added and incubated for 30 min. Finally, HRP chemiluminescent substrates were added and luminescence was read in using a microplate chemiluminescence reader. Control replicates were done with two different lot numbers of baculovirus-expressed human EZH2 bearing the canonical protein sequence, and a third iteration of the assay was done with EZH2 bearing a threonine residue at position 738, replacing the alanine.

Molecular Protein Modeling

We applied PRODRG,[33] a tool for high-throughput crystallography of protein-ligand complexes, to build *ad hoc* topologies of S-Adenosyl methionine (SAM) and S-Adenosyl homocysteine residues for Gromacs. We performed homology modeling via Modeler v.9.22,[34] setting the MD refinement value to "refine.very_slow" and leaving the remaining parameters at default. Only the SET and post-SET domains of EZH2 were modeled to conform with (1) the absence of known structures for the post-SET domain and (2) the need to simplify the model also in terms of computational costs and complexity. As a template structure for the modeling of SET domain we used 4MI0.[35]

Energy Minimization and ΔG Measurements

Potential energy minimization was performed on each EZH/SAH complex structures with GROMACS 2019 through a multi-step conjugate gradient algorithm using Amber99-ILDN force-field.[36] The minimization procedure automatically stopped when the resulting structure reached an RMSD threshold of 0.01. Estimates of the Free Energy of binding of each complex was measured via autodockVina.[37]

Screening of Subjects Affected by Syndromic Overgrowth

We tested the DNAm signature generated for *EZH2* against DNAm profiles generated on other syndromic overgrowth cohorts including subjects with Sotos syndrome and pathogenic variants in *NSD1* (n = 49),[9] Tatton-Brown Rahman syndrome and pathogenic variants in *DMNT3A* (n = 5),[22] and susceptibility to autism and pathogenic variants in *CHD8* (n = 10).[14] DNAm profiles at the 229 CpG sites were extracted from each subject and tested using the disease score correlation matrix against the DNAm profiles of controls and WS. DNAm profiles for all these subjects were generated on the Illumina 450k array.

Cohort with Undiagnosed Overgrowth and Intellectual Disability

Samples with undiagnosed overgrowth and intellectual disability for whom previous testing failed to identify *NSD1* and/or *EZH2* coding variants were included in this analysis. These samples were profiled on the Illumina 450k array as part of a large cohort study of overgrowth syndromes at the Weksberg lab with samples collected from the Division of Clinical Genetics at The Hospital for Sick Children, at the Department of Pediatrics and Adolescent Medicine from the University of Hong Kong, and at British Columbia Children's Hospital at the University of British Columbia. The UBC study recruited some samples referred in by international collaborators, some of which were referred because of a known variant in *EZH2*, and others referred in for the purposes of identifying the cause of undiagnosed overgrowth and/or intellectual disability. Samples included 73 DNAm profiles from blood-derived DNA generated on the Illumina 450k array. DNAm profiles at the *EZH2*-specific classification signature were extracted from each subject and tested using the disease score classification model to establish diagnosis in a search of potential individuals with WS. Once a DNAm profile similar to WS was identified in these undiagnosed individuals, next-generation sequencing was performed on these individuals to identify variants in PRC2 complex members.

Whole-Genome Sequencing

Subjects that were identified with DNAm profiles similar to the WS profile were subjected to next-generation sequencing for variant

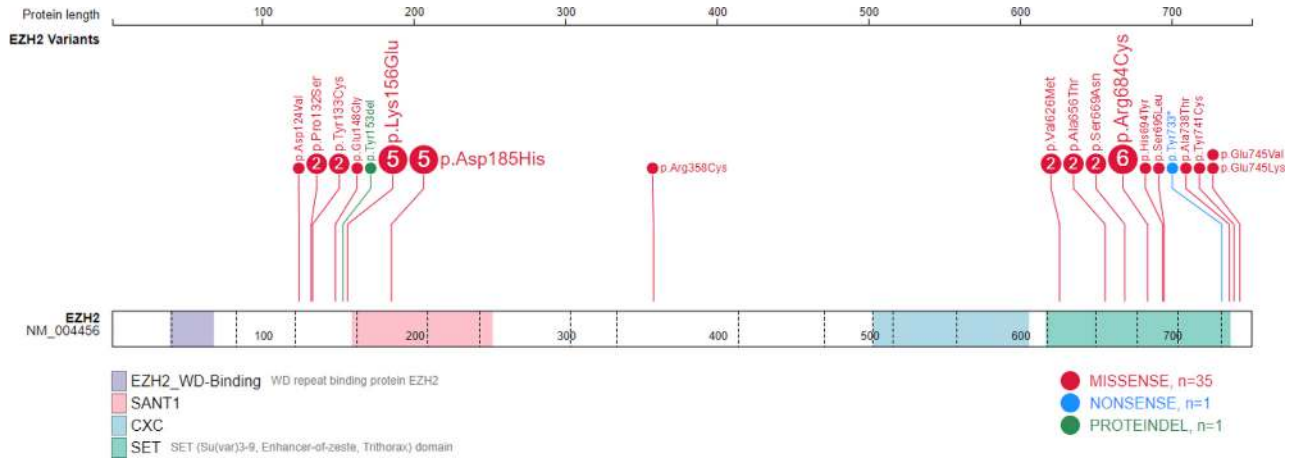

Figure 1. Comprehensive Visualization of *EZH2* Sequence Variants using ProteinPaint
Schematic representation of *EZH2* sequence variants included in the study. Each distinct variant in *EZH2* is represented by a disc sized in proportion to the number of samples and filled with the color representing its class based on the legend. Missense variants which constitute the large proportion of variants in *EZH2* are colored in red, nonsense variants in blue, and indels in green. Sequence variants are positioned by their amino acid coordinates based on *EZH2* (GenBank: NM_004456.4, hg19). The dotted vertical lines inside the protein delineate the boundaries of coding exons and the filled colors within the protein correspond to known protein domains.

identification. About 1 μg of genomic DNA was submitted to TCAG for genomic library preparation and whole-genome sequencing. DNA samples were quantified using Qubit High Sensitivity Assay and sample purity was checked using Nanodrop OD260/280 ratio. 700 ng of DNA was used as input material for library preparation using the Illumina TruSeq PCR-free DNA Library Prep Kit following the manufacturer's recommended protocol. In brief, DNA was fragmented to an average size of 400 bp using sonication on a Covaris LE220 instrument; fragmented DNA was end-repaired, A-tailed, and indexed TruSeq Illumina adapters with overhang-Ts added to the DNA. Libraries were validated on a Bioanalyzer DNA High-Sensitivity chip to check for size and absence of primer dimers and quantified by qPCR using the Kapa Library Quantification Illumina/ABI Prism Kit protocol (KAPA Biosystems). The Validated library was paired-end sequenced on the Illumina HiSeq X platform following Illumina's recommended protocol to generate paired-end reads of 150-bases in length.

Pyrosequencing

Genotyping was performed using quantitative pyrosequencing for *EZH2* variant: GenBank: NM_004456.4; c.2006G>A (p.Ser669Asn) in a father-child pair. Targeted assay was designed using the PyroMark Assay Design Software 2.0 (QIAGEN). Primer set sequences consisted of forward primer-5′-AAGCTGACAGAAGAGG GAAAGTG-3′; reverse primer 5′-TCCCAGCTCTGAAACATAC CA-3′ and sequencing primer 5′-TTCAAGTTGAACAGAAAG-3′. The amplification protocol was developed using a biotinylated universal primer approach. Regions of interest were amplified by PCR and pyrosequencing was carried out using the PyroMark Q24 pyrosequencer (QIAGEN) according to the manufacturer's protocol. Output data were analyzed using PyroMark Q24 Software (QIAGEN), which calculates the allelic percentage for each allele, allowing quantitative comparisons.

Ethics Statement

The study protocol has been approved by the Hospital for Sick Children Research Ethics Board (REB 1000038847). All the participants provided informed consent prior to sample collection. All samples and records were de-identified before any experimental or analytical procedures. The research was conducted in accordance with all relevant ethical regulations.

Results

Identification of DNAm Signature in Weaver Syndrome

To identify an *EZH2* DNAm signature, we generated genome-wide DNAm profiles using Infinium Human MethylationEPIC BeadChip arrays to test DNA from blood samples of individuals with pathogenic sequence variants in *EZH2* and control subjects. A comprehensive map illustrating the specific variants in *EZH2* and the number of affected individuals for each variant is shown in Figure 1. The complete WS cohort included 21 subjects, all of whom had clinical features of WS and *EZH2* pathogenic sequence variants identified in molecular diagnostic laboratories (Tables S1 and S2). The Canadian cohort was used for discovery (n = 8 unrelated individuals) (Table S1) and the UK cohort was used for validation (n = 8 unrelated subjects and n = 5 affected members of the same family) (Table S2). Both cohorts included individuals from international centers. The demographics for the discovery cohort were as follows: for WS there were five males and three females and the mean age ± standard deviation at sample collection was 16 ± 14.7 years (range 1–43 years). The 23 sex- and age-matched control subjects included 15 males and eight females; their mean age at the time of sample collection was 15.9 ± 9.8 years (range 3–39 years) (Table S3).

We used our established pipeline as outlined in the Subjects and Methods for signature derivation. Of the 774,516 CpG sites tested for differential DNAm between WS-affected case subjects and control subjects, we identified 229 statistically significant changes in DNAm across the

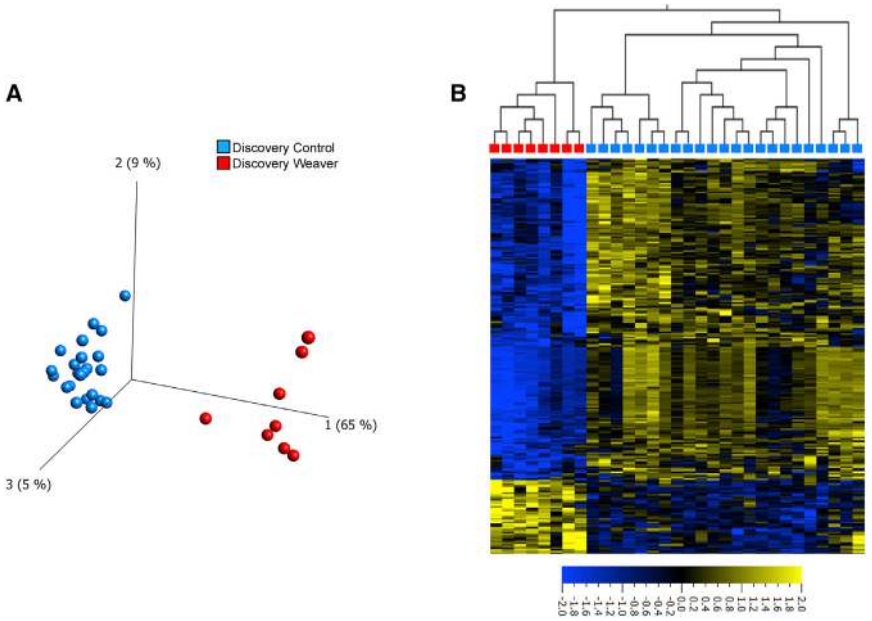

Figure 2. *EZH2*-Specific DNAm Signature

Principal components analysis (PCA) plot (A) and corresponding hierarchical clustering (B) (Eucledian distance metrics) and representative heatmap of 31 samples (n = 8 WS; n = 23 controls) using the differentially methylated CpG sites comprising the *EZH2*-specific DNAm signature (229 CpG sites). In both (A) and (B), samples labeled with red represent Weaver syndrome, blue samples are controls. On the heatmap, yellow indicates high DNAm and blue indicates low DNAm. For the heatmap, data are normalized for visualization (mean = 0, variance = 1).

genome at an FDR adjusted p value < 0.05 and $|\Delta\beta| \geq 0.10$ (Table S4). More than 81% of these sites showed hypomethylation in WS-affected case subjects compared to control subjects and the remaining 19% displayed hypermethylation in WS. Most of these sites were mapped to gene promoter regions within 5 kb of the transcription start site (Figure S1).

The signature CpG sites were examined using principal component analysis (PCA) (Figure 2A) and hierarchical clustering (Figure 2B) to assess their capacity to separate WS-affected case subjects from control subjects. As seen in Figure 2, the significant CpGs could be used to segregate the discovery cohort of WS-affected case subjects from control subjects. Consistent with the analysis of DNAm at single CpG sites, regional DNAm analysis using a bumphunting approach[29] identified several genomic segments that spanned between 5 and 35 CpG sites overlapping important, previously reported targets of *EZH2*[38] (Table S5). One example of such a genomic segment is the *HOXA* gene cluster, which was also identified as significant at the level of single CpG analysis.

Validation of the *EZH2* DNAm Signature

To test the specificity and sensitivity of the *EZH2* DNAm signature, comprised of 229 CpG sites, we generated median-methylation profiles of control subjects and WS-affected case subjects from the discovery cohort, and classified our independent validation cohort of WS-affected individuals (n = 8) and control subjects (n = 148) as either "WS" (positive disease score) or "not-WS" (negative disease score) based on their DNAm profiles using the correlation-based classification model (see Subjects and Methods). All controls showed DNAm profiles similar to the control profile, had negative disease scores, and were classified as "not-WS" demonstrating 100% specificity (Figure 3A). Each case in the WS validation cohort clustered with WS-affected case subjects and not with control subjects and generated positive disease scores. Therefore, the validation cohort demonstrated 100% sensitivity (Figure 3A, Table S6). We also tested the ability of the *EZH2* signature to classify DNAm profiles generated from a three-generation family with a segregating *EZH2* variant: GenBank: NM_004456.4; c.466A>G (p.Lys156Glu). All affected family members clustered with the WS-affected individuals and had positive disease scores (Figure 3A, Table S6).

To further assess the specificity of the *EZH2* signature, we evaluated its performance using a collection of control blood DNAm data extracted from the GEO repository. As there are very limited control data available for the EPIC array, we utilized data from Illumina 450k arrays (n = 718) (Table S7) as well as control 450k array data from our group (n = 80) to compare to a subset of the WS-discovery cohort also generated on the 450k array. For this analysis, we defined 161 CpG sites that overlapped the *EZH2* signature on the EPIC and 450k arrays and used the correlation-based classification model (see Subjects and Methods). As seen in Figure 3B, all 798 control samples had low disease scores for *EZH2* (Table S8), so they were predicted as "not-WS" (i.e., not to have pathogenic variants in *EZH2*), again demonstrating 100% specificity of the signature (Figure 3B). These results highlight the robustness of the *EZH2* signature, as it overcame many sources of variation such as sex, age, batch, and DNA processing methods contained in the GEO cohorts.

Classifying Models for *EZH2* Sequence Variants

Using the highly specific and sensitive *EZH2* signature, we tested 19 independent subjects with *EZH2* sequence variants (see Table S9) using the correlation-based model for variant classification (Table S6, Figure 4A). We found that ten subjects had positive disease scores and were therefore classified as "WS" and nine had negative disease scores and were classified as "not-WS."

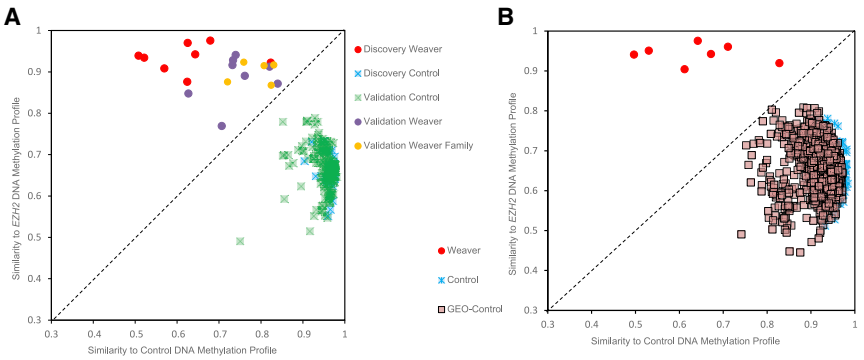

Figure 3. Testing the Sensitivity and Specificity of the *EZH2*-Specific Signature

(A) Plot representing the median-methylation profiles of WS-affected individuals (y axis) and control subjects (x axis) using the *EZH2* DNAm signature. The dashed line is set to represent the decision boundary for which individuals above the dashed lines have DNAm profiles more similar to *EZH2* signature and below the dashed line have DNAm profiles more similar to control subjects. A set of independent WS-affected individuals (validation cohort, purple circles, n = 8) as well as a WS-affected family (light orange circles, n = 5 affected members) with *EZH2* pathogenic variants were classified as "WS" (i.e., all individuals classified as more similar to the *EZH2* signature than control subjects) indicating high accuracy of the *EZH2* DNAm signature. The specificity of *EZH2* signature was estimated on an independent control validation set of 148 control samples (green crossed boxes); all subjects classified as more similar to control subjects (specificity 100%).

(B) Performance of the *EZH2* signature on data generated on 450k array including overlapping WS-affected subjects from the discovery cohort n = 7 (red circles), controls n = 80 (blue squares), and GEO controls n = 718 (brown squares) generated on 450k array. All control subjects had DNAm profiles more similar to the control profile and were therefore classified as "not-WS."

WS, Weaver syndrome.

Next, we tested the support vector machine-learning model (SVM), trained on significant CpG sites from the discovery cohorts, on the validation sets of WS-affected case subjects and control subjects. This model generated scores between 0% and 100%, with high scores classified as "WS" and low scores classified as "not-WS." We found that it correctly predicted the classification of all WS-affected case subjects and control samples with 100% accuracy (Table S10). Then, we used the SVM model on the test cohort of 19 subjects with *EZH2* variants (Table S9, Figure 4B). Ten out of the 19 variants were classified as pathogenic or "WS" (scores between 70%–95%) and eight variants were classified as benign or "not-WS" (scores between 0%–17%). One subject (described below) had an intermediate SVM score of 49%.

The correlation and SVM models were concordant for predicting the pathogenicity of *EZH2* variants in this testing cohort (Figures 4A and 4B) with one exception. Both models classified ten of the test cohort samples as pathogenic and eight as benign. Those classified as pathogenic included eight with missense variants in *EZH2* (Figure 4, Table S9); five of these were de-identified samples from PreventionGenetics, each with an associated clinical diagnosis of WS (PG-55, PG-61, PG-75, PG-87, and PG-45). The other three missense variants included a clinically affected father-child pair (EX0079 and EX0080, respectively) and one individual (MDL#76455) with an *EZH2* variant: GenBank: NM_004456.4; c.2006G>A (p.Ser669-Asn) inherited from a tall but clinically unaffected father (MDL#67485). The other two variants predicted to be pathogenic were copy number variants (CNVs) including a previously published *EZH2* deletion GenBank: NM_004456.4; *EZH2* (partial [exon20] deletion) (25.42 kb deletion involving *EZH2* and *CUL1*) associated with WS[6] and an *EZH2* duplication: GenBank: NM_004456.4; c.2196-2_2211dupAGATACAGCCAGGCTGAT.[1] The latter CNV was identified in an individual diagnosed with WS at birth (S126694). Although this sample had a positive disease score, it did not cluster with the majority of WS samples. Review of the clinical findings for this individual revealed an atypical presentation, i.e., more severe ID than most individuals with WS, as well as seizures and contractures, suggesting the possibility of a dual or more complex genetic diagnosis.

The sample for which discordant results were obtained was MDL#67485, the father of MDL#76455. Both father and son were identified to have the same *EZH2* variant; however, the child had a clinical diagnosis of WS while the father presented with tall stature but no other features of WS. The affected son (MDL#76455) had an SVM score of 82% and a positive disease score of 0.04 using the correlation-based model. The SVM score in the father was in the intermediate range (49%) but he had a negative disease score of −0.12. Considering the discordant predictions for the same variant, we investigated the possibility of somatic mosaicism for the *EZH2* variant in the father's blood-derived genomic DNA. Using quantitative pyrosequencing, we genotyped the *EZH2* variant in both the father and the affected child and found that the percentage of the variant allele was 46% in the blood of the child, but only 38% in the blood of the father, suggesting somatic mosaicism.

The eight samples that were classified by both models as benign included five subjects with undiagnosed OGID who inherited an *EZH2* variant: GenBank: NM_004456.4; c.553G>C (p.Asp185His). None of these individuals had the typical features of WS. Limited family history information was available; in one case the variant was known to be inherited from a phenotypically normal parent (Table S9).

Two subjects with *EZH2 de novo* missense variants and clinical findings atypical for WS were classified as benign using both classification models (Figures 4A and 4B). The subject with the *EZH2* variant GenBank: NM_004456.4; c.897+5G>A had a height SD +2.0, head circumference,

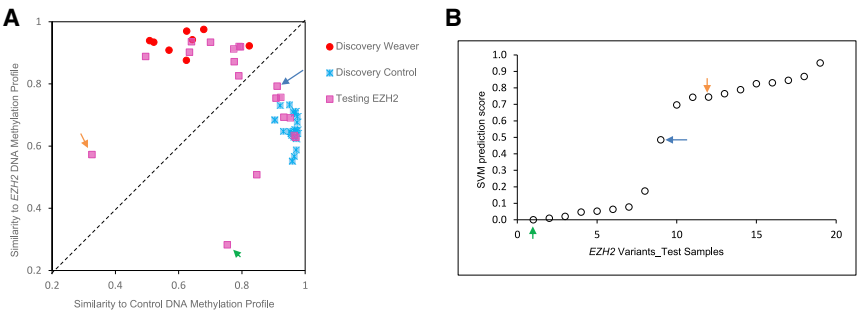

Figure 4. Testing the Ability of the *EZH2*-Specific Signature to Classify *EZH2* Variants

(A) Plot representing the median-methylation profiles of WS-affected (y axis) and control subjects (x axis) using the *EZH2* DNAm signature. A set of independent individuals with *EZH2* sequence variants (pink squares, n = 19) were classified using the *EZH2* signature. Of the 19 variants, ten classified as more similar to the *EZH2* DNAm profile than controls. The remaining nine variants classified as more similar to the control profile.

(B) Plot representing the Support Vector machine (SVM) scores (y axis). The SVM prediction model was used to predict pathogenicity of *EZH2* variants based on the DNAm signature. All ten variants predicted as pathogenic in (A) had also very high SVM scores > 70% and the remaining nine variants had very low SVM scores < 20% except one variant with an SVM score of 49%. Blue arrow represents sample MDL#67845 (p.Ser669Asn). Orange arrow represents sample S126694 (*EZH2*_c.2196-2_2211dupAGATACAGCCAGGCTGAT). Green arrow represents sample A1646 (p.Ala738Thr).
WS, Weaver syndrome.

OFC SD +1.5 (age of assessment, 6 years), and moderate intellectual disability with autism spectrum disorder (ASD). This variant had a negative correlation-based disease score (−0.3) and low SVM score of 1% and was therefore classified as benign or "not-WS." The individual with *EZH2* variant GenBank: NM_004456.4; c.1072C>T (p.Arg358Cys) was 6'3" at 21 years of age with normal intellect. This variant had a negative disease score of −0.16 and an SVM score of 17% and was classified as "not-WS."

Next, we tested an interesting *de novo EZH2* variant identified in an individual who presented with a phenotype characterized by growth failure. This child was born by Cesarean section at 39 weeks gestation for fetal distress. Birth weight was 2,055 g, length 43 cm, and head circumference 33 cm. There was no family history of WS. She had transient neonatal hypoglycemia and hypotonia. At 9 months of age she had a "clover-leaf" shaped skull, large anterior fontanelle, sparse eyebrows, upslanting palpebral fissures, and small ears. Development was moderately to severely delayed. By age 6 years, her weight was 9 kg, length 90 cm, and head circumference < 3rd centile for age.

The *EZH2* variant GenBank: NM_004456.4; c.2212G>A (p.Ala738Thr) in this subject (A1646) was predicted to be "not-WS" with an SVM score of 0% and a negative disease score of −0.47. Unsupervised hierarchical clustering of this variant showed that it clustered separately from both control subjects and from WS-affected individuals but, interestingly, it generated a DNAm profile *opposite* to that of the WS-affected individuals and very different from control individuals (Figures 5A and S2). This suggested that p.Ala738Thr is a GoF variant rather than a LoF variant for *EZH2*, which is also consistent with the phenotypic presentation (undergrowth rather than overgrowth). To further assess the putative GoF effect of this variant, a luminescence enzymatic assay was done at BPS Bioscience to test the enzymatic activity of EZH2 p.Ala738Thr. As shown in Figure 5B, this assay demonstrated increased enzymatic activity of this variant compared to wild-type EZH2.

Next, we performed a *de novo* modeling of the post-SET domain of EZH2, building a protein structure including the Ala738 residue and associated proteins. We used the S-adenosyl methionine (SAM) and S-adenosyl homocysteine (SAH) structures to obtain the EZH2-SAM and EZH2-SAH complexes, respectively (Figure S3A). We found that the substitution of the Ala738 residue to Thr does not significantly change the affinity for SAM (Figure S3B) whereas the substitution of the Ala738 residue to Thr induces an intermediate SAM/SAH binding mode featuring both the rotation of Gln735 side chain into a SAM-favorable state and a hydrogen-bond network that involves the side chains of Gln735, p.Ala738Thr, and Tyr663, leading to an increased affinity for both cofactors (Figure S3C). Taken together, these data support the hypothesis that this mutant form of EZH2 has a higher affinity for an active intermediate state of the enzyme, leading to the observed increased processivity (i.e., GoF).

Shared Functionality in PRC2 Complex Genes

Pathogenic sequence variants in *EZH2*, *SUZ12*, and *EED*, the core components of the PRC2 complex, have been associated with three clinically overlapping but distinct syndromes: Weaver, SUZ12-related overgrowth, and Cohen-Gibson syndromes, respectively. Therefore, we tested the ability of the *EZH2* signature to predict the pathogenicity of variants in *EED* (n = 3) and *SUZ12* (n = 4), in individuals with features that overlap WS and also in an individual who underwent genome sequencing for investigation of ASD (no other clinical information available), who was identified to have a 1.4 Mb 17q11.2 dup including *NF1* (MIM: 613113), *SUZ12*, and several RefSeq genes (Table S9). We used both the correlation-based and SVM models to classify these variants. All three subjects with *EED* missense variants showed high SVM scores, between 92% and 96%, and had positive disease scores (Figure 6, Tables S6 and S10), suggesting a pathogenic variant. Two subjects with *SUZ12* missense variants (A1765 and EX0066) had high SVM scores (92% and

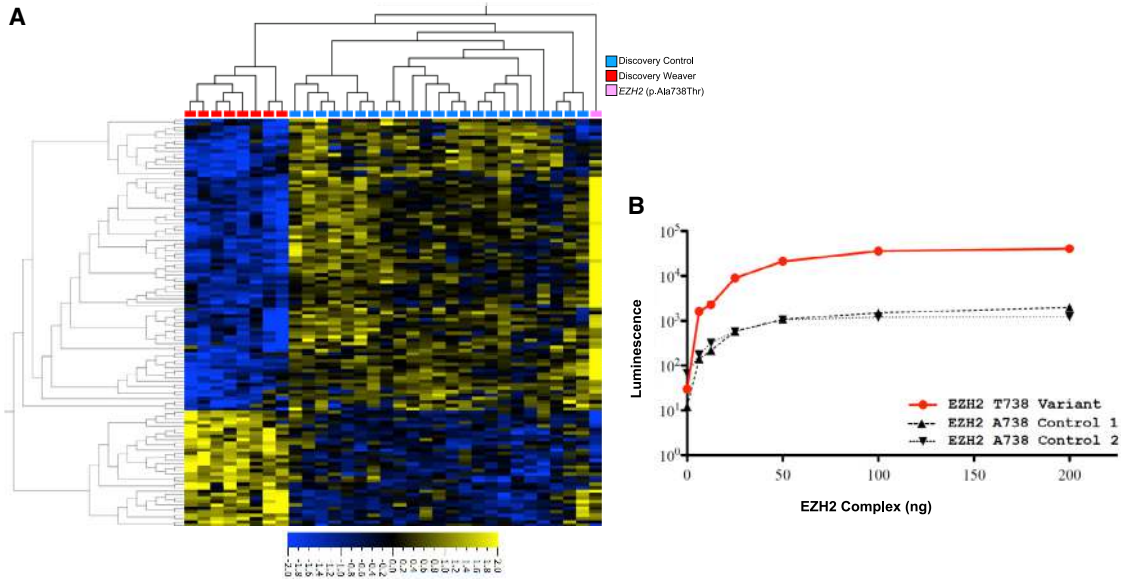

Figure 5. Gain-of-Function Variant in *EZH2* Have Opposite DNA Methylation Profile at the *EZH2* Signature
(A) Heatmap showing the hierarchical clustering of the DNAm profile of WS individuals (n = 8, red) with LoF (hypomorphic) variants in *EZH2*, controls (n = 23, blue) and *EZH2* GoF variant (p.Ala738Thr; pink) using the *EZH2* signature. On the heatmap, yellow indicates high DNAm and blue indicates low DNAm. The subject with an *EZH2* variant in pink display opposite DNAm profile when compared to WS DNAm profiles. For the heatmap, data are normalized for visualization (mean = 0, variance = 1).
(B) Enzymatic activity of *EZH2* GoF variant using an *in vitro* luminescence assay. Mutant *EZH2* (p.Ala738Thr) pre-assembled into PRC2 showed increased EZH2-mediated H3K27 methylation activity.
WS, Weaver syndrome; GoF, gain of function; LoF, loss of function.

96%) and positive disease scores. The third subject (EX0067) is the father of EX0066, who presented with mild clinical OGID features including tall stature, prominent forehead, chin crease, and normal intellect, and was identified to be mosaic for the variant by PCR based deep sequencing. He was found to have 8.4% mosaicism in blood leukocytes and 27% mosaicism in hair.[6] While the child had a 96% SVM score and a disease score of +0.36, the father had an SVM score of 19% and a disease score of −0.17. The low level of variant mosaicism in blood likely accounts for the negative scores and for the benign classification of this variant in the father. The fourth subject (M/R728468(2)) with a *SUZ12* variant of uncertain significance (VUS) had an SVM score of 5% and a disease score of −0.35. He presented with an atypical phenotype and inherited the variant from a clinically normal father, validating its benign classification. The fifth subject (EX0209) who harbored a duplication including *SUZ12* had an SVM score of 2% and a negative disease score of −0.27, suggesting that the duplication of *SUZ12* is unrelated to the OGID phenotype associated with LoF in SUZ12-related overgrowth syndrome.

Classification of Overgrowth Syndromes

Since overgrowth syndromes caused by heterozygous sequence variants in epigenes often share overlapping clinical features, we investigated the DNAm profile generated on the 450k arrays for Sotos syndrome (n = 49 [MIM: 117550]), Tatton-Brown Rahman syndrome (n = 5 [MIM: 615879]), and susceptibility to ASD (n = 10 [MIM: 615032]) caused by pathogenic variants in *NSD1* (MIM: 606681), *DNMT3A* (MIM: 602769), and *CHD8* (MIM: 610528), respectively.[9,14,22] The DNAm profiles of these individuals were compared to 7 WS-affected individuals (with pathogenic *EZH2* variants) from the discovery cohort and 80 control samples generated on the 450k. All individuals with pathogenic variants in *NSD1*, *DNMT3A*, and *CHD8* received a strongly negative disease score (Figure 7 and Table S11) and were therefore classified as "not WS."

Classification of Undiagnosed Overgrowth Syndromes

In order to investigate the potential utility of the *EZH2* signature as a first-tier diagnostic assay, we compared the profiles of 73 methylomes generated in our laboratory using the 450k on blood samples from individuals with OGID that tested negative for targeted sequencing of *NSD1* and *EZH2*. The goal of this analysis was to determine whether disease-specific DNAm signatures could be utilized to improve the diagnostic yield in this patient population. Using the *EZH2* signature, we were able to identify two samples that showed a positive disease score of 0.12 and 0.2 (Figure 8, Table S11). These two subjects were clinically diagnosed as WS but targeted testing of *EZH2* and subsequently *EED* were negative for both individuals. Since the DNAm profiles for these two subjects clustered with WS-affected individuals, we performed next-generation sequencing analysis on a research basis and identified in each subject a *de novo* frameshift variants in *SUZ12*: GenBank: NM_015355.3; c.1878del (p.Phe626Leufs*7);

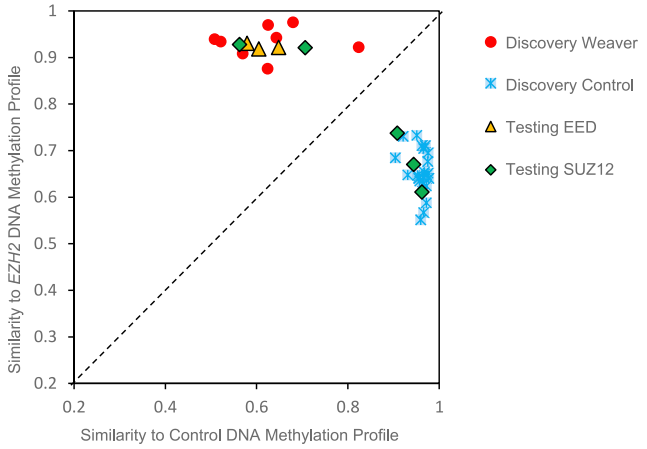

Figure 6. Testing the Utility of the *EZH2* Signature in Classifying Sequence Variants in Other Components of the PRC2 Complex
Using the *EZH2* signature, we compared the DNAm profiles of three subjects with *EED* sequence variants (yellow triangles) and those with *SUZ12* variants (green diamonds) to the DNAm profiles of controls (blue crossed boxes) and *EZH2* pathogenic variants (red circles). All three subjects with pathogenic variants in *EED* had DNAm profiles more similar to the *EZH2* profile than control subjects. Two subjects with *SUZ12* pathogenic variants also classified with Weaver syndrome and the three remaining *SUZ12* variants showed DNAm profiles more similar to controls.

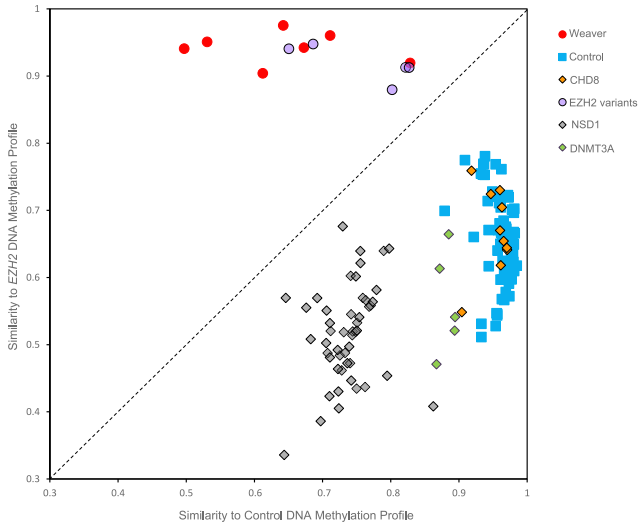

Figure 7. Classification of Subjects with Syndromic Overgrowth using *EZH2* Signature
Plot representing samples with sequence variants in epigenes associated with other overgrowth syndromes. These included subjects with: *NSD1* pathogenic variants associated with Sotos syndrome (n = 49, GEO: GSE74432), *DNMT3A* pathogenic variants associated with Tatton-Brown Rahman syndrome (n = 5; GEO: GSE128801), and *CHD8* pathogenic variants associated with macrocephaly and susceptibility to autism (n = 10; GEO: GSE113967). These data were compared to seven WS-affected subjects from the discovery cohort and five test individuals with pathogenic mutations in *EZH2* which were run on the Illumina 450k. All overgrowth syndrome individuals had DNAm profiles more similar to control subjects and distinguishable from WS.

and c.1715_1716insCA (p.Leu572Phefs*11). These two variants were validated by Sanger sequencing as recently reported.[39] The clinical findings in both individuals were consistent with those reported in other individuals with SUZ12-related overgrowth, which has phenotypic overlap with WS, as recently described.[39] These data provide further evidence for the utility of DNAm profiling as a first-tier diagnostic tool, in addition to its previously recognized proficiency as a second-tier tool for VUS classification.[9,10,40]

Functional Enrichment of the *EZH2* DNAm Signature

Gene-set enrichment analysis was performed using g:Profiler[32] on the 89 genes that overlapped the 229 CpG sites in the *EZH2* DNAm signature. The results demonstrate enrichment for genes with roles in pattern specification processes, skeletal system development, and regulation of morphogenesis such as Homeobox A5 (*HOXA5* [MIM: 142952]), ALX homeobox 4 (*ALX4* [MIM: 605420]), and SIX homeobox 2 (*SIX2* [MIM: 604994]) (Figure S4A, Table S12) (Benjamini-Hochberg corrected p value < 0.01). This enrichment in skeletal development and organ morphogenesis pathways reflects the known roles of the PRC2 core complex in cellular lineage (and subsequent tissue) specification and also reflects some of the cardinal features of WS (i.e., tall stature, advanced osseous maturation, neuronal migration disorders, and developmental delay) (Figure S4B), thereby further validating the utility of this signature to elucidate the functional, biological, and molecular impact of *EZH2* pathogenic variants.

Discussion

We have identified a highly sensitive and specific *EZH2* DNAm signature that can be used to classify missense variants in *EZH2* as pathogenic or benign. Notably, this report offers unique insights about using DNAm signatures to classify GoF sequence variants and detect somatic mosaicism. Further, this signature has utility in classifying sequence variants in two other genes, *EED* and *SUZ12*, that encode proteins that participate with EZH2 in the repressive PRC2 complex. Finally, we demonstrated the first-tier diagnostic capability of this signature, based on its ability to predict the presence of pathogenic variants in a PRC2 complex gene (*SUZ12*) in two individuals with undiagnosed OGID syndrome.

In the development of DNAm signatures, an important consideration is the issue of cell type variation. As DNA methylation varies markedly among all cell types, including hematopoietic cells,[41] it is important to account for inter-individual differences in cell types of whole blood samples when analyzing DNAm data from this tissue.[42] Here, we estimated the cell proportions in the discovery cohort using the Houseman method[28] and included these estimates as covariates when identifying *EZH2*-specific signature. This method uses DNAm measured in purified blood cells from a small number of healthy adults to generate cell proportion estimates in

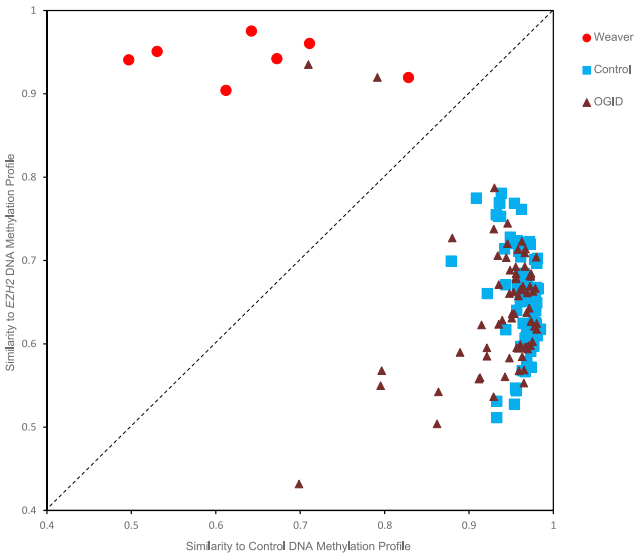

Figure 8. Testing the Ability of the *EZH2* Signature in Classifying Undiagnosed OGID-Affected Subjects Based on Their DNAm Profiles

OGID-affected subjects included in this analysis were previously tested negative for targeted mutations screening in *NSD1* and *EZH2*. Out of the 73 subjects with OGID (brown triangles), we identified that most had DNAm profiles similar to control subjects (blue squares). Interestingly, we identified two subjects with DNAm profiles more similar to the *EZH2* profile than control samples. OGID, overgrowth and intellectual disability.

mixed cell populations, such as whole blood. We found no statistically significant differences in blood cell type composition between the affected individuals and control subjects, which aligned with reports of individuals with EZH2-related overgrowth having typical blood cell composition (see GeneRevies in Web Resources). Future work measuring both complete blood counts and DNAm in the same individuals with WS is needed to confirm the efficacy of the Houseman method in this context.

PRC2 belongs to the Polycomb group (PcG) protein family of chromatin-modifying enzymes that function as repressors of gene expression.[43] In mammalian cells, PRC2 is primarily targeted to the unmethylated CpG islands of inactive developmental genes.[44] PRC2 recruitment and binding to the targeted CpG island is facilitated in part by polycomb-like proteins (PCLs) and JARID2 which link the PRC2 complex to genomic sites enriched in CpG dense regions.[44,45] In the *EZH2* LoF (hypomorphic) state, we observed a loss of DNA methylation at CpG sites enriched in promoter regions of developmental genes, which can lead to transcriptional changes during critical developmental time points resulting in somatic overgrowth characteristic of WS. In the case of the *EZH2* GoF state, we observed a gain of methylation at the same CpG sites which can lead to transcriptional changes during development resulting in growth restriction. Further work is needed to explore the underlying transcriptional dysregulation at the gene promoters overlapping the *EZH2*-specific signature in relevant cell types and critical time points during development.

Our results support the position that the DNA methylation changes observed in individuals carrying germline *EZH2* pathogenic variants represent a downstream cascade of events at the molecular level in response to a genetic change. These DNA methylation changes recur across multiple generations if the genetic change is present, demonstrating the impact of *EZH2* pathogenic variants on epigenetic programming during embryonic development. Evidence for this hypothesis is provided by our genome-wide DNAm findings in a three-generation family in which the *EZH2* variant segregates from a father to his daughter and her three sons. All family members showed the same *EZH2*-specific epigenotype. This suggests that the same epigenotype was re-established downstream of a dysfunctional PRC2 complex in somatic tissues of each affected family member, in each generation. We have previously shown similar findings in a two-generation family where one father and two affected offspring with inherited pathogenic *NSD1* variant all share the same *NSD1*-specific DNA methylation signature.[9]

Constitutional pathogenic variants in *EZH2* are known to cause WS; most of these are missense variants. Thus, predicting pathogenicity of these variants can present significant challenges, which now can be resolved using the *EZH2* signature. Such classifications are particularly useful for individuals with clinical findings atypical for WS who carry *de novo* missense variants in *EZH2*. The *EZH2* signature, derived from constitutional pathogenic sequence variants, is comprised of differentially methylated CpG sites in WS-affected individuals relative to control subjects; the majority (>81%) of these sites are hypomethylated in WS-affected individuals relative to control subjects, whereas 19% are hypermethylated. This suggests that pathogenic *EZH2* sequence variants cause a failure of promoter CpG methylation at CpG sites critically important for normal growth and development. Support for the functional relevance of the DNAm signature comes from the fact that the most enriched CpG sites and genomic regions overlap *HOX* genes, which are known targets of EZH2.[38] In cancer as well, it has been shown that loss of EZH2 contributes to epigenetic-dependent overexpression of the *HOX* genes.[38]

We show that the *EZH2* signature can differentiate the functional effects of missense variants in *EZH2* using two previously validated classification models specifically correlation-based[9,14,15] and machine learning.[10,40,46] We propose that review of the clinical features of subjects carrying sequence variants in disease-associated genes, such as *EZH2*, combined with the application of several methods to visualize the data in different ways as presented here, can improve the utility of this highly sensitive and specific functional assay for variant pathogenicity classification. We expect that this approach will also be useful for testing the functionality of non-coding variants in disease-associated genes. Discordant results using the two different analytic models were of interest in assessing mosaicism in a father with a sequence variant in *EZH2* who presented with

tall stature, normal intellect, and no other features of WS. Pyrosequencing studies identified low-level mosaicism in his blood which is highly relevant for accurate genetic counseling for the family.

In an individual with an *EZH2* missense variant (GenBank: NM_004456.4; c.2212G>A [p.Ala738Thr]) who presented with growth restriction rather than an overgrowth phenotype, we considered the possibility of a GoF variant and used three independent approaches to assess this hypothesis. Utilizing the DNAm signature generated using LoF sequence variants in *EZH2*, we identified DNAm alterations at most CpGs observed in the *EZH2* signature but found that these were "opposite in direction" to the DNAm changes in the *EZH2* signature. Further support for a GoF role for this variant was derived from an *in vitro* luminescence assay showing increased enzymatic activity. Finally, protein modeling showed that the p.Ala738Thr form of EZH2 shows a strong preference for an intermediate SAM/SAH binding state, which helps to explain both the increased processivity observed *in vitro* and to corroborate the GoF inferred from the individual-specific methylation profile and the clinical presentation.

It has been shown that methylation of H3K27 by EZH2 requires the presence of two additional proteins: EED and SUZ12.[47] Inactivation of any one of these three protein subunits severely compromises the enzymatic activity of PRC2 and results in the reduction of H3K27me3.[48–50] Not surprisingly, individuals with *EZH2*, *EED*, and *SUZ12* pathogenic variants present with overlapping phenotypes, including generalized overgrowth, similar craniofacial features, advanced bone maturation, macrocephaly, and variable degrees of intellectual disability.[51] As there are few specific clinical features that distinguish each of these three disorders from each other, it is difficult in many of these individuals, for even the most astute clinician, to define specific genomic diagnosis in a particular case. Here we show that the pathogenic sequence variants in these genes that confer similar phenotypes also confer a common DNAm signature that represents the functional effect of PRC2 perturbation during development. As more individuals with sequence variants in *EED* and *SUZ12* are identified, it will become possible to assess whether specific DNAm "sub-signatures" exist for LoF variants in each gene. This will be of great interest to both clinicians and scientists who struggle with elucidating the boundaries between syndromes both molecularly and clinically (lumping versus splitting of syndromes). Although these three conditions currently carry independent names, some would argue that they constitute a phenotypic spectrum that reflects their integrated molecular actions via a shared functional complex.

We have shown that the *EZH2* signature has 100% specificity, in that it does not misclassify as positive any individuals with OGID and pathogenic variants in *NSD1*, *DNMT3A*, or *CHD8*. This is congruent with what we published previously for other OGID signatures such as *NSD1*.[9] We have also demonstrated that a gene-specific DNAm signature can be more broadly applicable if that gene encodes a protein that participates in a functional complex. This is very important for both first-tier diagnostics and for gene discovery. Our work is distinct from that reported for the BAF complex[12] in which the derivation of the signature involved many genes in the complex. In this paper we were able to use the information that the *EZH2* signature also reflected the presence of pathogenic variants in other PRC2 complex genes to efficiently assign a definitive genomic diagnosis to two individuals with features of OGID. In both individuals, a clinical diagnosis of WS was suspected, but targeted sequencing for *EZH2* and *EED* was negative. The fact that DNAm profiles were positive prompted us to arrange next-generation sequencing for other genes encoding proteins in the PRC2 complex, leading to the identification of pathogenic sequence variants in *SUZ12*. For the many genes that encode proteins participating in functional complexes, our approach provides a valuable paradigm for first-tier diagnostics that could also play an important role in future novel gene discovery.

There is considerable overlap between germline/constitutional sequence variants observed in WS and the acquired somatic *EZH2* sequence variants observed in myeloid malignancies. Two of the sequence variants present in our WS-affected individuals, p.Arg684Cys and p.Tyr733*, have also been detected as somatically acquired sequence variants in myeloid malignancies.[52] These two sequence variants were identified constitutionally in seven unrelated individuals in the current series. None of these individuals have developed malignancies, though it is noteworthy that the oldest of these seven individuals is less than 10 years old. Given that myeloid malignancies associated with somatic *EZH2* sequence variants usually present later in life, it will be important to follow these individuals with WS who could be at increased risk of myeloid malignancies not only in childhood, but also over their lifetime.[3] Long-term clinical data for WS will be valuable for time-to-event analysis that estimates the age-specific malignancy risk in these individuals.

In summary, this study demonstrates that pathogenic LoF variants in *EZH2* are associated with a highly sensitive and specific DNAm signature that has significant diagnostic power to classify pathogenic versus benign variants. This study also provides unique finding wherein a DNAm signature has the ability to distinguish GoF from LoF variants, as well as the capability to detect somatic mosaicism of coding *EZH2* variants. Notably we also show that the *EZH2* DNAm signature can positively classify pathogenic sequence variants in two other genes, *EED* and *SUZ12*, encoding proteins in the core PRC2 complex. Finally, we have demonstrated that DNAm signatures for genes that encode proteins in a functional complex can play an important role not only for elucidating molecular pathophysiology and as a tool for first- and second-tier diagnostics, but also for paradigms for new gene discovery.

Data Availability

Some datasets used in this study are available publicly and may be obtained from gene expression omnibus (GEO) using the following accession numbers: GSE54670, GSE54399, GSE51245, GSE89353, GSE53045, GSE40279, GSE42861, GSE128801, GSE36064, GSE74432, GSE113967, and GSE128801. The remaining datasets are not publicly available due to institutional ethics restrictions.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2020.03.008.

Acknowledgments

We are grateful to all the families with childhood overgrowth conditions who participated in this research and to the many clinicians who recruited them into the study. We also acknowledge the technical assistance of Youliang Lou and Chunhua Zhao. This work was supported by Canadian Institutes of Health Research (CIHR) grants to R.W. (IGH-155182 and MOP-126054), by CIHR Project Grant PJT-148830 to W.T.G., and by a BCCHR intramural IGAP salary award to W.T.G. Bioinformatic analyses were supported in part by the Canadian Centre for Computational Genomics (C3G), part of the Genome Technology Platform (GTP), funded by Genome Canada through Genome Quebec and Ontario Genomics (A.L.T., M.B.), Genome Canada through Ontario Genomics (A.L.T., S. Choufani, M.B., and R.W.), and the Ontario Brain Institute (OBI). OBI is an independent non-profit corporation, funded partially by the Ontario government. Protein Modeling was supported by Telethon Foundation (grant number GEP13105 to G.T.), the EPIGEN Flagship Project of the Italian National Research Council (to G.T.), the European Research Council (grant number 616441-DISEASEAVATARS to G.T.), and Ricerca Corrente granted by the Italian Ministry of Health (to G.T.). The opinions, results, and conclusions are those of the authors, and no endorsement by the Ontario Brain Institute is intended or should be inferred.

Declaration of Interests

The authors declare no competing interests.

Received: December 9, 2019
Accepted: March 10, 2020
Published: April 2, 2020

Web Resources

AnnotationHub, https://bioconductor.org/packages/release/bioc/html/AnnotationHub.html
CADD, https://cadd.gs.washington.edu/snv
g:Profiler, http://biit.cs.ut.ee/gprofiler/
GenBank, https://www.ncbi.nlm.nih.gov/genbank/
GeneReviews, Tatton-Brown, K., and Rahman, N. (1993). EZH2-Related Overgrowth, https://www.ncbi.nlm.nih.gov/books/NBK148820/
GEO, https://www.ncbi.nlm.nih.gov/geo/
GROMACS, https://doi.org/10.1016/j.softx.2015.06.001
GREAT, http://great.stanford.edu/public/html/
OMIM, https://www.omim.org/
ProteinPaint, https://proteinpaint.stjude.org

References

1. Tatton-Brown, K., Murray, A., Hanks, S., Douglas, J., Armstrong, R., Banka, S., Bird, L.M., Clericuzio, C.L., Cormier-Daire, V., Cushing, T., et al.; Childhood Overgrowth Consortium (2013). Weaver syndrome and EZH2 mutations: Clarifying the clinical phenotype. Am. J. Med. Genet. A. *161A*, 2972–2980.
2. Gibson, W.T., Hood, R.L., Zhan, S.H., Bulman, D.E., Fejes, A.P., Moore, R., Mungall, A.J., Eydoux, P., Babul-Hirji, R., An, J., et al.; FORGE Canada Consortium (2012). Mutations in EZH2 cause Weaver syndrome. Am. J. Hum. Genet. *90*, 110–118.
3. Tatton-Brown, K., Hanks, S., Ruark, E., Zachariou, A., Duarte, Sdel.V., Ramsay, E., Snape, K., Murray, A., Perdeaux, E.R., Seal, S., et al.; Childhood Overgrowth Collaboration (2011). Germline mutations in the oncogene EZH2 cause Weaver syndrome and increased human height. Oncotarget *2*, 1127–1133.
4. Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science *298*, 1039–1043.
5. Cohen, A.S., Tuysuz, B., Shen, Y., Bhalla, S.K., Jones, S.J., and Gibson, W.T. (2015). A novel mutation in EED associated with overgrowth. J. Hum. Genet. *60*, 339–342.
6. Imagawa, E., Higashimoto, K., Sakai, Y., Numakura, C., Okamoto, N., Matsunaga, S., Ryo, A., Sato, Y., Sanefuji, M., Ihara, K., et al. (2017). Mutations in genes encoding polycomb repressive complex 2 subunits cause Weaver syndrome. Hum. Mutat. *38*, 637–648.
7. Cohen, A.S., and Gibson, W.T. (2016). EED-associated overgrowth in a second male patient. J. Hum. Genet. *61*, 831–834.
8. Lui, J.C., Barnes, K.M., Dong, L., Yue, S., Graber, E., Rapaport, R., Dauber, A., Nilsson, O., and Baron, J. (2018). Ezh2 Mutations Found in the Weaver Overgrowth Syndrome Cause a Partial Loss of H3K27 Histone Methyltransferase Activity. J. Clin. Endocrinol. Metab. *103*, 1470–1478.
9. 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. 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.
11. 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.
12. 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. Nat. Commun. *9*, 4885.
13. 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.; Care4Rare Canada Consortium (2018). Genomic DNA Methylation Signatures Enable Concurrent Diagnosis and Clinical Genetic Variant Classification in Neurodevelopmental Syndromes. Am. J. Hum. Genet. *102*, 156–174.
14. 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 *11*, 103.
15. 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. BMC Med. Genomics *12*, 105.
16. 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. J. Am. Acad. Child Adolesc. Psychiatry *55*, 906–913.e2.
17. Accomando, W.P., Wiencke, J.K., Houseman, E.A., Nelson, H.H., and Kelsey, K.T. (2014). Quantitative reconstruction of leukocyte subsets using DNA methylation. Genome Biol. *15*, R50.
18. Montoya-Williams, D., Quinlan, J., Clukay, C., Rodney, N.C., Kertes, D.A., and Mulligan, C.J. (2018). Associations between maternal prenatal stress, methylation changes in IGF1 and IGF2, and birth weight. J. Dev. Orig. Health Dis. *9*, 215–222.
19. Friemel, C., Ammerpohl, O., Gutwein, J., Schmutzler, A.G., Caliebe, A., Kautza, M., von Otte, S., Siebert, R., and Bens, S. (2014). Array-based DNA methylation profiling in male infertility reveals allele-specific DNA methylation in PIWIL1 and PIWIL2. Fertil. Steril. *101*, 1097–1103.e1.
20. Barbosa, M., Joshi, R.S., Garg, P., Martin-Trujillo, A., Patel, N., Jadhav, B., Watson, C.T., Gibson, W., Chetnik, K., Tessereau, C., et al. (2018). Identification of rare de novo epigenetic variations in congenital disorders. Nat. Commun. *9*, 2064.
21. Alisch, R.S., Barwick, B.G., Chopra, P., Myrick, L.K., Satten, G.A., Conneely, K.N., and Warren, S.T. (2012). Age-associated DNA methylation in pediatric populations. Genome Res. *22*, 623–632.
22. Jeffries, A.R., Maroofian, R., Salter, C.G., Chioza, B.A., Cross, H.E., Patton, M.A., Dempster, E., Temple, I.K., Mackay, D.J.G., Rezwan, F.I., et al. (2019). Growth disrupting mutations in epigenetic regulatory molecules are associated with abnormalities of epigenetic aging. Genome Res. *29*, 1057–1066.
23. Dogan, M.V., Shields, B., Cutrona, C., Gao, L., Gibbons, F.X., Simons, R., Monick, M., Brody, G.H., Tan, K., Beach, S.R., and Philibert, R.A. (2014). The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics *15*, 151.
24. Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol. Cell *49*, 359–367.
25. Kular, L., Liu, Y., Ruhrmann, S., Zheleznyakova, G., Marabita, F., Gomez-Cabrero, D., James, T., Ewing, E., Lindén, M., Górniewicz, B., et al. (2018). DNA methylation as a mediator of HLA-DRB1*15:01 and a protective variant in multiple sclerosis. Nat. Commun. *9*, 2397.
26. Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density DNA methylation array with single CpG site resolution. Genomics *98*, 288–295.
27. 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.
28. 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.
29. Jaffe, A.E., Murakami, P., Lee, H., Leek, J.T., Fallin, M.D., Feinberg, A.P., and Irizarry, R.A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int. J. Epidemiol. *41*, 200–209.
30. Li, D., Xie, Z., Pape, M.L., and Dye, T. (2015). An evaluation of statistical methods for DNA methylation microarray data analysis. BMC Bioinformatics *16*, 217.
31. 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. Nat. Biotechnol. *28*, 495–501.
32. Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat. Protoc. *14*, 482–517.
33. Schüttelkopf, A.W., and van Aalten, D.M. (2004). PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. Acta Crystallogr. D Biol. Crystallogr. *60*, 1355–1363.
34. Webb, B., and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. Curr. Protoc. Bioinformatics *54*, 1–, 37.
35. Wu, H., Zeng, H., Dong, A., Li, F., He, H., Senisterra, G., Seitova, A., Duan, S., Brown, P.J., Vedadi, M., et al. (2013). Structure of the catalytic domain of EZH2 reveals conformational plasticity in cofactor and substrate binding sites and explains oncogenic mutations. PLoS ONE *8*, e83737.
36. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O., and Shaw, D.E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins *78*, 1950–1958.
37. Trott, O., and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. *31*, 455–461.
38. Xu, F., Liu, L., Chang, C.K., He, Q., Wu, L.Y., Zhang, Z., Shi, W.H., Guo, J., Zhu, Y., Zhao, Y.S., et al. (2016). Genomic loss of EZH2 leads to epigenetic modifications and overexpression of the HOX gene clusters in myelodysplastic syndrome. Oncotarget *7*, 8119–8130.
39. Cyrus, S.S., Cohen, A.S.A., Agbahovbe, R., Avela, K., Yeung, K.S., Chung, B.H.Y., Luk, H.M., Tkachenko, N., Choufani, S., Weksberg, R., et al.; C.A.U.S.E.S. Study (2019). Rare SUZ12 variants commonly cause an overgrowth phenotype. Am. J. Med. Genet. C. Semin. Med. Genet. *181*, 532–547.

40. 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. Am. J. Hum. Genet. *104*, 685–700.
41. Reinius, L.E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S.E., Greco, D., Söderhäll, C., Scheynius, A., and Kere, J. (2012). Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS ONE *7*, e41361.
42. Jaffe, A.E., and Irizarry, R.A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. *15*, R31.
43. Schuettengruber, B., Bourbon, H.M., Di Croce, L., and Cavalli, G. (2017). Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. Cell *171*, 34–57.
44. Li, H., Liefke, R., Jiang, J., Kurland, J.V., Tian, W., Deng, P., Zhang, W., He, Q., Patel, D.J., Bulyk, M.L., et al. (2017). Polycomb-like proteins link the PRC2 complex to CpG islands. Nature *549*, 287–291.
45. Youmans, D.T., Schmidt, J.C., and Cech, T.R. (2018). Live-cell imaging reveals the dynamics of PRC2 and recruitment to chromatin by SUZ12-associated subunits. Genes Dev. *32*, 794–805.
46. 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.
47. Guglielmelli, P., Biamonte, F., Score, J., Hidalgo-Curtis, C., Cervantes, F., Maffioli, M., Fanelli, T., Ernst, T., Winkelman, N., Jones, A.V., et al. (2011). EZH2 mutational status predicts poor survival in myelofibrosis. Blood *118*, 5227–5234.
48. Pasini, D., Bracken, A.P., Jensen, M.R., Lazzerini Denchi, E., and Helin, K. (2004). Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. EMBO J. *23*, 4061–4071.
49. Faust, C., Schumacher, A., Holdener, B., and Magnuson, T. (1995). The eed mutation disrupts anterior mesoderm production in mice. Development *121*, 273–285.
50. Shen, X., Liu, Y., Hsu, Y.J., Fujiwara, Y., Kim, J., Mao, X., Yuan, G.C., and Orkin, S.H. (2008). EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency. Mol. Cell *32*, 491–502.
51. Imagawa, E., Albuquerque, E.V.A., Isidor, B., Mitsuhashi, S., Mizuguchi, T., Miyatake, S., Takata, A., Miyake, N., Boguszewski, M.C.S., Boguszewski, C.L., et al. (2018). Novel SUZ12 mutations in Weaver-like syndrome. Clin. Genet. *94*, 461–466.
52. Ernst, T., Chase, A.J., Score, J., Hidalgo-Curtis, C.E., Bryant, C., Jones, A.V., Waghorn, K., Zoi, K., Ross, F.M., Reiter, A., et al. (2010). Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. Nat. Genet. *42*, 722–726.