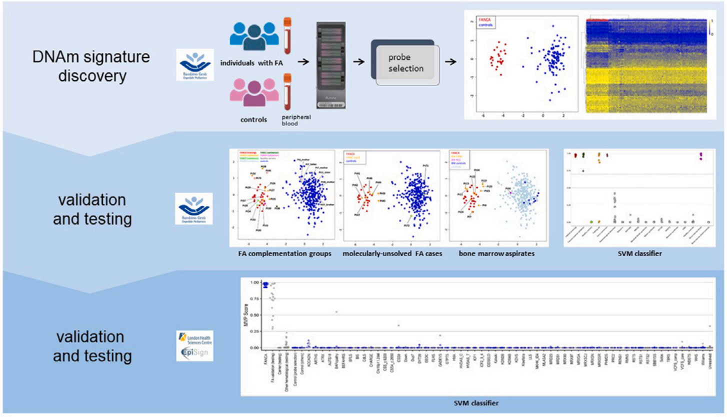
Identification of a robust DNA methylation signature for Fanconi anemia

Graphical abstract



Authors

Daria Pagliara, Andrea Ciolfi, Lucia Pedace, ..., Bekim Sadikovic, Franco Locatelli, Marco Tartaglia

Correspondence

franco.locatelli@opbg.net (F.L.),
marco.tartaglia@opbg.net (M.T.)

Fanconi anemia (FA) is the most common bone marrow failure syndrome. Molecular diagnosis is challenging due to its genetic heterogeneity and wide mutation spectrum. We identify an FA-specific DNA methylation signature that aids in classification of variants and establishing/refuting a clinical diagnosis in molecularly uninformative and revertant cases.

Pagliara et al., 2023, The American Journal of Human Genetics *110*, 1938–1949
November 2, 2023 © 2023 American Society of Human Genetics.
https://doi.org/10.1016/j.ajhg.2023.09.014

Identification of a robust DNA methylation signature for Fanconi anemia

Daria Pagliara,[1,15] Andrea Ciolfi,[2,15] Lucia Pedace,[1,15] Sadegheh Haghshenas,[3,16] Marco Ferilli,[2,16] Michael A. Levy,[3] Evelina Miele,[1] Claudia Nardini,[1] Camilla Cappelletti,[2] Raissa Relator,[3] Angela Pitisci,[1] Rita De Vito,[4] Simone Pizzi,[2] Jennifer Kerkhof,[3] Haley McConkey,[3,5] Francesca Nazio,[1] Sarina G. Kant,[6] Maddalena Di Donato,[7] Emanuele Agolini,[7] Marta Matraxia,[7] Barbara Pasini,[8] Alessandra Pelle,[8] Tiziana Galluccio,[9] Antonio Novelli,[7] Tahsin Stefan Barakat,[6,10] Marco Andreani,[9] Francesca Rossi,[11] Cristina Mecucci,[12] Anna Savoia,[13] Bekim Sadikovic,[3,5,17] Franco Locatelli,[1,14,17,*] and Marco Tartaglia[2,17,*]

Summary

Fanconi anemia (FA) is a clinically variable and genetically heterogeneous cancer-predisposing disorder representing the most common bone marrow failure syndrome. It is caused by inactivating predominantly biallelic mutations involving >20 genes encoding proteins with roles in the FA/BRCA DNA repair pathway. Molecular diagnosis of FA is challenging due to the wide spectrum of the contributing gene mutations and structural rearrangements. The assessment of chromosomal fragility after exposure to DNA cross-linking agents is generally required to definitively confirm diagnosis. We assessed peripheral blood genome-wide DNA methylation (DNAm) profiles in 25 subjects with molecularly confirmed clinical diagnosis of FA (FANCA complementation group) using Illumina's Infinium EPIC array. We identified 82 differentially methylated CpG sites that allow to distinguish subjects with FA from healthy individuals and subjects with other genetic disorders, defining an FA-specific DNAm signature. The episignature was validated using a second cohort of subjects with FA involving different complementation groups, documenting broader genetic sensitivity and demonstrating its specificity using the EpiSign Knowledge Database. The episignature properly classified DNA samples obtained from bone marrow aspirates, demonstrating robustness. Using the selected probes, we trained a machine-learning model able to classify EPIC DNAm profiles in molecularly unsolved cases. Finally, we show that the generated episignature includes CpG sites that do not undergo functional selective pressure, allowing diagnosis of FA in individuals with reverted phenotype due to gene conversion. These findings provide a tool to accelerate diagnostic testing in FA and broaden the clinical utility of DNAm profiling in the diagnostic setting.

Introduction

Fanconi anemia (FA) (MIM: PS227650) is a clinically heterogeneous multisystem, cancer-predisposing disorder representing the most common inherited bone marrow failure (BMF) syndrome.[1–3] Clonal defects of hematopoiesis usually appear in the first decade of life and involve mild to moderate cytopenia. The risk of developing BMF, hematological malignancies (e.g., myelodysplastic neoplasms, acute myeloid leukemia), and other cancers (e.g., squamous cell carcinomas of head and neck) significantly increases with age and characterizes the natural history of the disease.[3–5] Café au lait spots, skeletal anomalies (e.g., deformities of the thumb and radius), and short stature are also common features, while less frequent malformations involve the genitourinary, cardiovascular, gastrointestinal, and cerebral systems.[3] In subjects presenting with absent or subtle congenital anomalies, FA diagnosis is suspected only at the onset of the hematologic condition, which is often indistinguishable from other constitutional or acquired BMF syndromes.

The phenotypic heterogeneity of FA is mirrored by an equally marked genetic heterogeneity. The disorder is caused by inactivating variants in genes encoding proteins of the FA/BRCA pathway, whose function mediates the DNA damage repair process required for maintenance of genomic stability and regulates cell-cycle checkpoints and replication fork remodeling.[6–10] To date, 23 genes

[1]Department of Hematology/Oncology and Cell and Gene Therapy, Bambino Gesù Children's Hospital, IRCCS, 00146 Rome, Italy; [2]Molecular Genetics and Functional Genomics, Bambino Gesù Children's Hospital, IRCCS, 00146 Rome, Italy; [3]Verspeeten Clinical Genome Centre, London Health Sciences Centre, London, ON N6A 5W9, Canada; [4]Department of Laboratories, Bambino Gesù Children's Hospital, IRCCS, 00146 Rome, Italy; [5]Department of Pathology and Laboratory Medicine, Western University, London, ON N6A 3K7, Canada; [6]Department of Clinical Genetics, Erasmus MC University Medical Center, 3015 Rotterdam, the Netherlands; [7]Laboratory of Medical Genetics, Translational Cytogenomics Research Unit, Bambino Gesù Children Hospital, IRCCS, 00146 Rome, Italy; [8]AOU Città della salute e della scienza di Torino, Molinette's Hospital, 10126 Torino, Italy; [9]Laboratory of Transplant Immunogenetics, Department of Hematology/Oncology, Cell and Gene Therapy, IRCCS Bambino Gesù Children's Hospital, 00146 Rome, Italy; [10]ENCORE Expertise Center for Neurodevelopmental Disorders, Erasmus MC University Medical Center, 3015 Rotterdam, the Netherlands; [11]Department of Woman, Child and of General and Specialist Surgery, University of Campania "Luigi Vanvitelli," 80138 Naples, Italy; [12]Institute of Hematology and Center for Hemato-Oncology Research, University and Hospital of Perugia, 06123 Perugia, Italy; [13]Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, 37134 Verona, Italy; [14]Department of Pediatrics, Catholic University of the Sacred Hearth, 00168 Rome, Italy
[15]These authors contributed equally
[16]These authors contributed equally
[17]Senior author
*Correspondence: franco.locatelli@opbg.net (F.L.), marco.tartaglia@opbg.net (M.T.)
https://doi.org/10.1016/j.ajhg.2023.09.014.

have been implicated in FA, with 60%–70% of cases attributed to biallelic loss-of-function variants in *FANCA* (MIM: 607139) (Human Gene Mutation Database; Fanconi Anemia Mutation Database; Leiden Open Variation Database). FA is largely characterized by an autosomal recessive inheritance; exceptions are represented by *RAD51*-related FA (FANCR; MIM: 617244), which is transmitted as an autosomal dominant trait, and *FANCB*-related FA (FANCB; MIM: 300514), which is a recessive X-linked condition.[3]

Molecular diagnosis of FA still represents a challenge due to the genetic heterogeneity characterizing the disease and extremely variable nature of the underlying genetic lesions. As for other genetic diseases, establishing the clinical relevance of variants of uncertain significance (VUSs) remains a major diagnostic issue requiring dedicated functional validation efforts and leaving the diagnosis undetermined in most cases.[11,12] The identification of variants in non-coding regions of the genome further adds to the complexity.[13] Based on the well-known hypersensitivity of FA cells to DNA interstrand crosslink-inducing agents, cytogenetic testing for the increased chromosomal breakages or rearrangements occurring in peripheral blood (PB) cells in presence of diepoxybutane (DEB) or mitomycin C (MMC) is currently used to confirm diagnosis.[14] However, these assays are labor intensive, time consuming, and can provide false-negative results in case of low cellularity and mosaicism.

DNA methylation (DNAm) is an epigenetic mark occurring throughout the genome with key function in controlling gene expression.[15] DNAm patterns vary across tissues and developmental stages and are mitotically heritable with high fidelity through individual cell lineages.[16–18] Unique and stable DNAm signatures (also known as "episignatures"), which are defined as the cumulative DNAm patterns occurring at multiple cytosine-phosphate-guanine (CpG) dinucleotides across the genome, have been described in a growing number of genetic disorders and have been postulated to be a functional consequence of pathogenic variants affecting epigenetic regulators. They are emerging as highly accurate and stable biomarkers.[19,20] Indeed, DNAm profiling has successfully been used to confirm the diagnosis in subjects with strong clinical suspicion who could not be molecularly solved routinely via genomic sequencing.[21–25]

Here, we show that FA is characterized by a distinctive DNAm signature. We provide evidence that the resolved episignature is robust and highly sensitive and specific for classification of FA individuals with respect to subjects with other blood and bone marrow (BM) disorders, other rare disorders with known episignatures, and healthy control individuals. We successfully applied this episignature to confirm or provide evidence against diagnosis of FA in clinically and/or genetically unsolved cases. Finally, we demonstrate that the identified DNAm signature maintains "memory" of FA in subjects with genetic reversions and can be applied using DNA obtained from either PB or BM.

Material and methods

Subjects

The OPBG discovery cohort (n = 25) included individuals with DEB-positive, molecularly confirmed clinical diagnosis of FA (FANCA complementation group [MIM: 227650]). The OPBG validation cohorts included 14 subjects with clinical diagnosis of FA belonging to different complementation groups with molecularly confirmed diagnosis and/or positive DEB testing. This validation cohort also included 6 apparently healthy subjects heterozygous for pathogenic variants associated with FA and two subjects with FA in whom molecular reversion of one of the mutated alleles was observed, resulting in a complete phenotypic rescue in hematopoietic cells. The clinical characteristics of these cohorts are summarized in Table 1. The OPBG cohort also included 201 healthy subjects, 23 hematological acquired/constitutional disorders (14 individuals with refractory cytopenia of childhood [RCC; MIM: 614286], 4 with aplastic anemia [AA; MIM: 609135], 2 with Diamond-Blackfan anemia 6 [DBA6; MIM: 612561], 1 with GATA2-deficiency disorder [GATA2-D; MIM: 614172], 1 with BMF syndrome 2 [BMFS2; MIM: 615715], and 1 with unclassified BM dysplasia) tested negative by DEB assay, as well as 94 individuals belonging to a clinically heterogeneous group of genetic diseases, which were also included in the analyses directed to define and validate the FA DNAm signature. The validation analysis also included BM aspirates obtained from 10 individuals of the discovery cohort along with one individual with RCC and 9 healthy subjects. The testing phase analyzed 9 molecularly unsolved cases with a clinical diagnosis of FA or having features suggestive of the disorder. The molecular and clinical information of the subjects belonging to the FA sub-cohorts or having other hematological conditions and healthy individuals heterozygous for pathogenic variants in FA-associated genes (i.e., healthy carriers) is reported in Tables S1 and S2. Finally, two large cohorts including healthy individuals (n = 795) and individuals affected with one of >60 genetic disorders (n = 1,927) collected in the EpiSign Knowledge Database (EKD) were used for a second independent validation phase to determine specificity of the classifier relative to a range of other previously described episignature disorders.

All subjects had been screened by parallel sequencing using either custom gene panels or clinical exome sequencing. In all individuals with clinical diagnosis of FA (or having a suggestive phenotype), DEB/MMC testing was performed, and mutation scan included all the currently known genes implicated in FA and SNP array/multiplex ligation-dependent probe amplification (MLPA) analysis (supplemental methods).

The study was approved by the Ethical Committees of the Ospedale Pediatrico Bambino Gesù (1702 OPBG 2018) and Western University (REB116108 and REB106302). Clinical data and DNA specimens were collected, stored, and used in accordance with the ethical standards of the declaration of Helsinki protocols, with signed informed consents from the participating subjects/families.

Methylation analysis

PB/BM genomic DNA was extracted using standard techniques. DNAm profiling was performed using the Illumina Infinium MethylationEPIC BeadChip (EPIC) arrays and 500 ng DNA as input material,[22,23] according to the manufacturer's protocol. BeadChip processing was carried out using an Illumina iScan microarray platform. To minimize systematic bias, the 135 newly

Table 1. Clinical characteristics of the discovery and validation FA cohort

	Number or median	Percentage or range
Subjects	39	100
Males	28	72
Females	11	28
Median age, years		
At diagnosis	7	0–26
At sampling	8	2–26
Hematological status at sampling		
BM failure[a]	34	87
Mild	6	18
Moderate	9	26
Severe	19	56
Clonal evolution	2	5
Acute myeloid leukemia	1	2.5
Acute lymphoid leukemia	1	2.5
Absence of hematological alterations	3	8
Transfusion dependency	14	36
BM, cytogenetic abnormalities	1	2
Clinical phenotype		
Skin manifestations	32	82
Ocular defects	17	44
Renal and urinary tract malformations	17	44
Thumb or radial abnormalities	9	23
Congenital heart disease	9	23
Ear abnormalities/conductive deafness	5	13
Other skeletal abnormalities	6	15
Endocrinopathy	4	10

BM, bone marrow.

[a] Reduction of neutrophil count and/or platelet count and/or Hb level below standard-age ranges, classified as mild, moderate, and severe in the presence of at least one of the following criteria: mild, absolute neutrophil count $<1.5\ 10^9$/L, platelet count 150,000–50,000 10^9/L, Hb lower than normal for age but >8 g/dL; moderate, absolute neutrophil count $<1\ 10^9$/L, platelet count <50,000 10^9/L, Hb < 8 g/dL; severe, absolute neutrophil count $<0.5\ 10^9$/L, platelet count <30,000 10^9/L, Hb < 8 g/dL.

processed samples were randomly distributed in different experiments.

Data analysis was performed as previously described.[19,23] IDAT files were imported into R version 4.2.1 for analysis by means of *ChAMP* v.2.26.0[26] and normalized with background correction using the *minfi* package (v.1.44.0).[27] Probes located on X/Y chromosomes or known to cross-react with chromosomal locations other than their target regions, containing SNPs at/near the tested CpG sites, and with a detection p value >0.01 were excluded, resulting in approximately 700,000 high-quality probes that were used in the subsequent analyses. Principal component analysis (PCA)/multidimensional scaling (MDS) was performed to inspect any batch effect and identify outlier samples. For the discovery of the DNAm signature, the *MatchIt* package (v.4.5.4)[28] was used to select the best-matching controls considering an in-house database including >300 samples, considering age and sex as matching variables, providing a control sample size (n = 111, labeled as "training") four times larger than that of tested FANCA cases (n = 25).

Methylation levels (β values) were converted to M values, which were used for linear-regression modeling by means of empirical Bayes moderated t-statistic corrected for false discovery rate (Benjamini-Hochberg's FDR, *limma* package [v.3.54.2]) to identify differentially methylated probes (DMPs).[29] Estimated blood cell type proportions for each sample were added to the model matrix to reduce the bias associated with those confounding variables.[30] The most informative 1,000 probes were identified considering the interaction between the effect size (i.e., absolute mean methylation difference $|\beta| > 0.05$) and FDR threshold <0.05. Receiver's operating curve characteristic (ROC) analysis was performed to identify the top 500 probes. Then, probes with a Pearson's pairwise correlation >0.84 were removed to identify the minimal set of independent probes defining the FANCA-specific DNAm signature. Normalized β values for each sample were compared by means of MDS, considering the pairwise Euclidean distances between samples. Hierarchical clustering was performed using the *ggplots* package (v.3.4.2).[31] Leave-one-out sample cross-validation was evaluated by MDS analysis.

A machine-learning classification model based on the generated DNAm signature was used to categorize samples.[19–23] A support vector machine (SVM) model was trained by using 75% of the set, including the 25 molecularly and clinically confirmed FANCA cases and the corresponding matched controls. This model generates probabilities of pathogenicity score from "0" (control matching) to "1" (FA matching) for each sample. The SVM classifier was trained with linear kernel function with the *e1071* R package (v.1.7) by using the nu-classification option. The remaining 25% of samples represented the test set by which the algorithm calculates the model's best hyperparameters and accuracy, performing a 5-fold cross-validation during the training process. This procedure was repeated four times to verify that each sample was used three times for training and once for testing. An SMOTE (*imbalance* R package v.1.0.2.1) oversampling technique was carried out to overcome class imbalance between affected and control samples in the training process.[32] SVM classifier scores below 0.30 were considered as not matching the DNAm signature, from 0.30 to 0.70 were considered inconclusive, and >0.70 indicated high-confidence occurrence of match.

A second SVM classifier using the EKD dataset was also trained based on the same DNAm signature using a set composed of the 25 FANCA samples and 63 matched control samples of the OPBG cohort plus a portion of samples from EKD (healthy controls and individuals with other rare disorders), which was similarly split in training and test set (75% and 25%, respectively). The detailed process of constructing the classifier has previously been reported.[19,21] This additional classifier, created using a reference cohort comprising thousands of samples, was developed to validate the DNAm signature and maximize the sensitivity and specificity of the model toward FANCA.

Differentially methylated regions (DMRs) were determined by extracting regions containing at least five different CpGs within 1 kb with a minimum methylation difference of 10% and an FDR <0.01 by using the *DMRcate* package (v.2.12.0).[33] DMRs were evaluated for pathways and gene-set enrichment by means of *missMethyl R* package (v.1.32.1),[34] and Manhattan plot was carried out by using *qqman R* package (v.0.1.8).[35] The overlap between DMRs and 127 reference epigenomes from the NIH Roadmap Epigenomics Consortium was evaluated using *GIGGLE*.[36]

Results

DNAm profiling splits FANCA subjects and control subjects with full specificity

EPIC-based PB DNAm profiling of 25 individuals with bona fide diagnosis of FA, complementation group A, was performed to explore the occurrence of a specific DNAm signature in FA. Within this discovery cohort, 24 subjects had a clinical diagnosis of FA, harbored biallelic pathogenic/likely pathogenic variants in *FANCA* (GenBank: NM_000135.4), and had a positive DEB test. These subjects had variants that could be considered as representative of the molecular spectrum of pathogenic *FANCA* lesions, as they included large gene/intragenic (multi-exon) deletions and non-sense, splice site, frameshift, and missense changes (Table S1). All variants were classified according to the American College of Medical Genetics and Genomics (ACMG) criteria.[37] A single subject (individual 13) with clinical features of FA, positive DEB and FANCD2 ubiquitination tests, and showing compound heterozygosity for a pathogenic variant and a VUS affecting a residue adjacent to a codon previously reported altered in FA was also included. In these individuals, the clinical diagnosis of FA was established during childhood (mean age at diagnosis = 7.4 ± 2.8 years) (Table S2). At sampling, these subjects ranged from 2 to 18 years of age with an average of 9.2 ± 3.7 years. Non-FA individuals (n = 111) were selected from an internal database (OPBG, Rome) to generate an age-, sex-, and batch-matched control group including both healthy subjects (n = 69) and individuals affected by a heterogeneous group of rare disorders (n = 42). MDS analysis documented a superimposed distribution pattern of the control samples jointly processed with the discovery cohort and the remaining age- and sex-matched controls, supporting the absence of any significant batch effect (Figure S1). An unbiased evaluation of the distribution of genome-wide p values for differentially methylated sites in FA samples pointed out the involvement of >160,000 CpGs throughout the EPIC array ($\pi_0 = 0.76$) (Figure S2). By using a combination of linear regression modeling and ROC analysis followed by removing the most correlated CpG sites, we selected 82 probes constituting the minimal informative set defining the FANCA-specific DNAm signature (Figure 1A; Table S3). MDS-based leave-one-out sample cross-validation confirmed the robustness of the selected probes (Figure S3), 32% of which were hypomethylated and 68% were hypermethylated compared with controls (Figure 1B).

To validate further the identified episignature and test its specificity and robustness, MDS and unsupervised hierarchical clustering analyses were performed including additional 132 healthy control subjects (age at sampling ranging between 1.4 and 79.2 years, mean = 22.3 ± 17.4), 41 individuals with a genetically heterogeneous group of rare disorders (age at sampling range: 0.8 to 48 years, mean = 13.7 ± 12.6), and 14 children with hematological disorders with clinical features partially overlapping with FA (RCC, n = 7; AA, n = 3; DBA6, n = 2; GATA2-D, n = 1; BMFS2, n = 1) (Tables S1 and S2). Both analyses confirmed a distinct clustering of the *FANCA* samples with respect to all the other subcohorts (Figure 1C), providing evidence that the selected subset of EPIC probes defines a disease-specific signature able to properly classify these individuals with high specificity.

The FANCA DNAm signature has broad specificity for FA

To further assess the specificity and sensitivity of the generated episignature, 12 subjects with clinical diagnosis of FA having biallelic pathogenic/likely pathogenic variants in genes implicated in FA and positive DEB test were analyzed. These subjects belonged to different FA

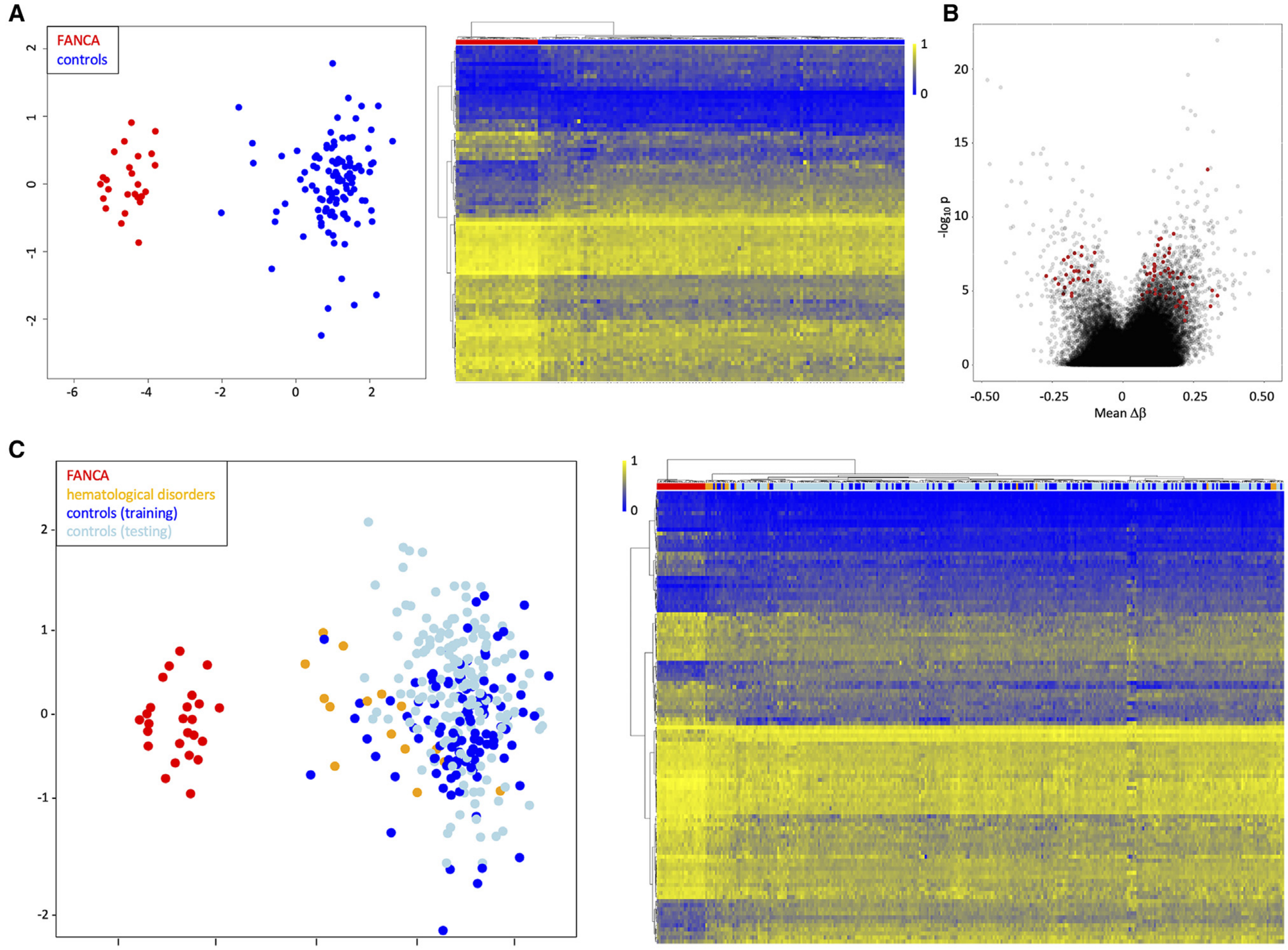


Figure 1. Identification of a DNAm signature for the Fanconi anemia complementation group A

(A) DNAm discovery. MDS (left) and heatmap (right) plots showing clustering of the DNAm profiles of 25 FANCA samples (red) segregating from those of 111 age-, sex-, and batch-matched control samples (blue) using 82 differentially methylated CpG probes defining the FA episignature. Control samples were used for model training and included healthy subjects and individuals with other genetic disorders. The heatmap showing the DNAm levels were clustered by Ward's method with dendrograms representing the Euclidian distances between samples (columns) and individual CpG sites (rows). Sample groups are indicated using color bars above the heatmap.

(B) Volcano plot showing differences in methylation of the tested probes (represented as circles on the plot) between the FANCA and control groups. For each probe, the magnitude (mean methylation difference, x axis) and significance (−log10 adjusted p value, y axis) of DNAm difference between groups was evaluated to identify the most informative probes (red). Negative and positive mean methylation differences reflect decreased methylation (hypomethylation) and increased methylation (hypermethylation) in FA samples compared with controls, respectively.

(C) DNAm signature validation. MDS (left) and unsupervised hierarchical clustering (right) analyses were performed by considering additional 132 healthy controls and 41 individuals affected with other genetic diseases (light blue) and 14 pediatric cases with hematological disorders with clinical features partially overlapping with FA (RCC, AA, DBA6, GATA2-D, and BMFS2) (orange). Sample groups are indicated using color bars above the heatmap.

complementation groups (FANCA, n = 6; FANCC [MIM: 227645], n = 2; FANCG [MIM: 614082], n = 2; FANCL [MIM: 614083], n = 1; FANCP [MIM: 613951], n = 1) (Tables S1 and S2). Two other affected siblings (family 23), who were homozygous for a *FANCA* VUS but DEB test and FANCD2 ubiquitination test positive, were also included. These analyses also considered 6 healthy subjects harboring heterozygous pathogenic/likely pathogenic *FANCA*/*FANCG* variants to test the ability of the approach to discriminate between affected individuals and healthy carriers. MDS and hierarchical clustering analyses properly classified all FANCA, FANCC, FANCG, and FANCP subjects who showed a DNAm pattern consistent with the FA episignature and were plotted apart from the 295 control subjects, including healthy subjects and individuals with different genetic diseases (Figure 2A). Proper classification of these individuals was attained by applying a machine learning-based scoring system (SVM prediction scores >0.7 indicate a high-confidence match) (Figure 2B). Of note, individual 38, representing the only sample belonging to the FANCL complementation group, slightly diverged from the FA cluster, though also showed a supportive SVM score (0.75). In this subject, genotyping performed using DNA obtained from buccal

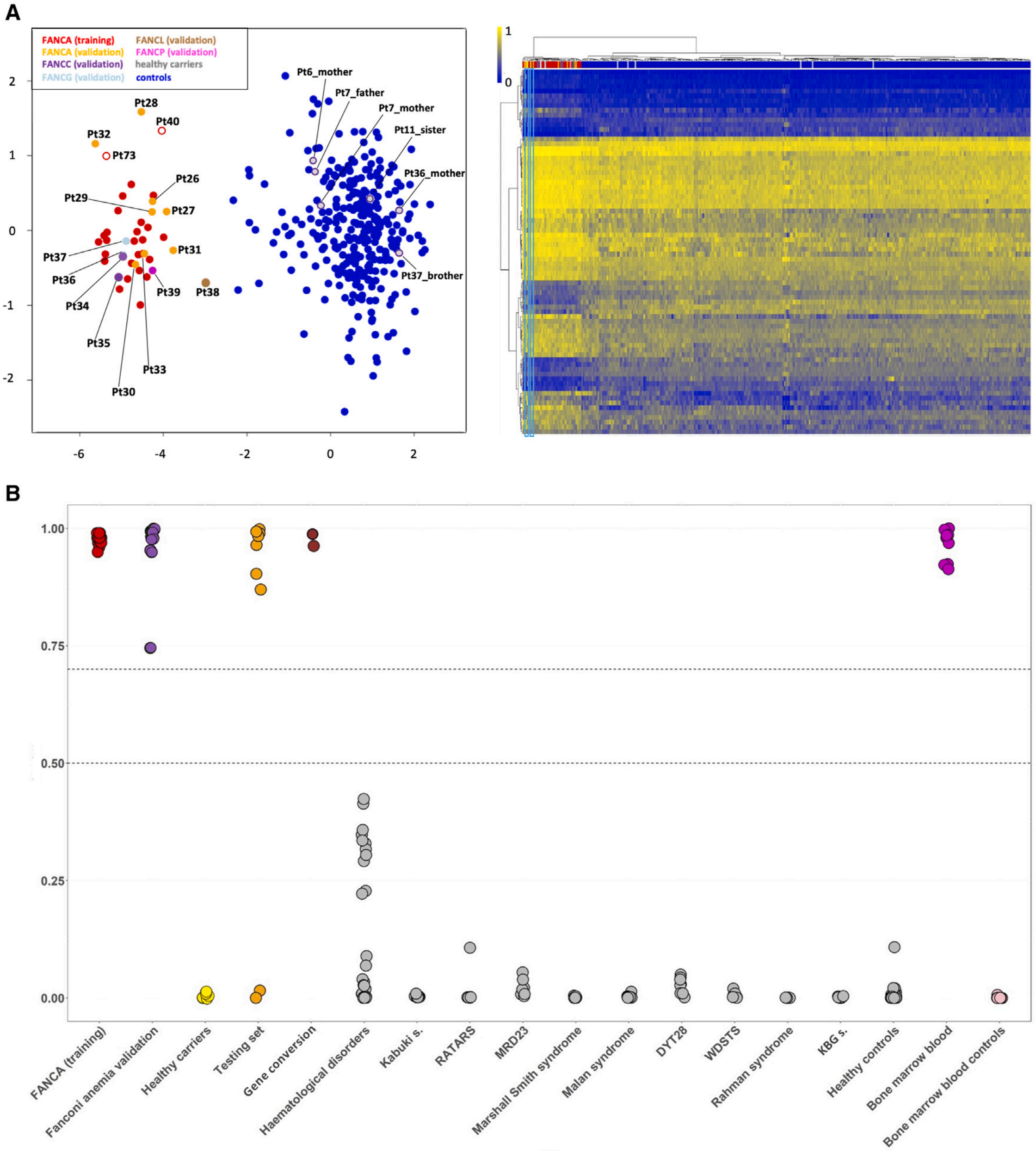


Figure 2. The FANCA DNAm signature shows broad specificity for Fanconi anemia

(A) MDS (left) and heatmap (right) plots showing clustering of the peripheral blood DNAm profiles of 14 subjects with FA belonging to different complementation groups (FANCA, orange; FANCC, purple; FANCG, light blue; FANCL, brown; FANCP, magenta), segregating from those of healthy heterozygous carriers (gray), and healthy controls plus other genetic disorders (blue). FANCA samples used for training are depicted in red; FA subjects with reverted phenotype due to gene conversion (individual 40 and individual 73) are in unfilled red circles. The heatmap showing the DNAm levels were clustered by Ward's method with dendrograms representing the Euclidian distances between samples (columns) and individual CpG sites (rows). Sample groups are indicated using color bars above the heatmap, using the same color code of the left; individual 40 and individual 73 were depicted in yellow, and their profile is highlighted by light-blue box.

(B) Sample classification using the FA DNAm signature. An SVM classification model was trained with FANCA samples used for probe selection (red) and used to classify different cohorts available in an internal database. Showed results are the summary of 4-fold cross-validation when the SVM model is trained using FANCA training samples and 75% of all other control samples in the OPBG database. The FA samples belonging to the different complementation groups (peripheral blood, purple; bone marrow aspirates, magenta), heterozygous healthy carriers (yellow), molecularly and clinically unsolved FA cases (orange), reverted FA cases (maroon), and 25% of controls (healthy subjects and individuals with different genetic diseases) (peripheral blood, gray; bone marrow aspirates, pink) were used for testing. Each sample is plotted on the basis of its scoring by the model. SVM scoring ranges from 0 to 1 (y axis), representing the probability of having a DNAm profile fitting FA. All FA samples showed an SVM score >0.75, while non-FA samples had an SVM score <0.50. RATARS: Radio-Tartaglia syndrome (MIM: 619312); MRD23: intellectual developmental disorder 23 (MIM: 615761); DYT28: dystonia 28 (MIM: 617284); WDSTS: Wiedemann-Steiner syndrome (MIM: 605130); KBG syndrome (MIM: 148050).

swab confirmed the compound heterozygosity for the two pathogenic variants, excluding somatic mosaicism (data not shown). Conversely, all healthy heterozygous carriers were properly classified within the control groups (Figure 2; Table S4). These findings demonstrate that the generated DNAm signature is sufficiently robust to properly weigh FA gene dosage and has broad specificity to accurately classify the different FA complementation groups.

The FA DNAm signature is retained in reverted FA due to gene conversion

Revertant mosaicism occurs in FA.[38] Indeed, genetic instability might be potentially beneficial by increasing the opportunity to correct the constitutional genetic lesions by reversion of the original mutations. In FA, the proliferative advantage of the progeny of the self-corrected precursor cell is expected to result in clonal expansion and gradual replacement of the defective cell population, leading to complete reversal of the hematologic phenotype.[38,39] Though rare events, we had the chance to test the generated DNAm signature in two subjects with revertant mosaicism. The first individual (individual 40) was a 33-year-old male who had been diagnosed with FA when he was 7 years old (Tables S1 and S2). His 1-year-older affected sibling had developed acute leukemia shortly after the diagnosis of BM aplasia. In this subject, BM reversion involving the maternally inherited allele was demonstrated (Figure S4). Nevertheless, the subject showed normal blood cell counts (normal leukocyte differential count, white cells: 5.86 10^3/mmc, red cells: 4.66 10^6/mmc, platelets: 217 10^3/mmc) as well as normal chromosome breakage rates as a consequence of the reversion, and the DNAm pattern of this subject matched the FA-specific DNAm signature, as shown by the MDS and hierarchical clustering analyses (Figure 2A). The second individual (individual 73) was an 18-year-old girl born to a consanguineous family of Kurdish origin. The subject showed radial dysplasia at birth; she had short stature (−4.1 SDS) with normal body proportions, microcephaly (−2.3 SDS), delayed bone age, and hypo-/hyperpigmentation in the neck and left side of the abdomen. MLPA analysis documented a paternally transmitted increased copy number of exons 28 and 29 in both blood- and fibroblast-derived DNA. Scans of the entire coding sequence of *FANCA* did not allow the identification of the maternal variant. The MMC test was positive in fibroblasts but negative in leukocytes, indicating a diagnosis of FA with reverse mosaicism in BM, which was in line with the observed DNAm pattern (Figure 2A). The applied SVM classifier supported a diagnosis of FA with high confidence in both individuals (SVM score >0.95, in both cases) (Figure 2B; Table S4).

The FA DNAm signature properly classifies molecularly and/or clinically unsolved cases

The generated episignature was applied to 9 molecularly unsolved cases, including 7 subjects with clinical features suspicious of FA with inconclusive molecular data and 2 individuals with a complex phenotype with compound heterozygous variants in *FANCI* (MIM: 611360) (individual 71) or having one pathogenic variant in *BRCA1* (MIM: 113705) and one VUS in *LIG4* (MIM: 601837) (individual 72) (Tables S1 and S2). In the first subgroup, four subjects harbored at least one VUS involving *FANCA* (individual 45 and individual 46), *FANCG* (MIM: 602956) (individual 42), and *FANCI* (individual 41), while variants and structural rearrangements associated with FA had not been reported in three individuals by using a specifically designed gene panel or clinical exome sequencing (individual 43, individual 44, and individual 47). In all of these subjects, MDS and hierarchical clustering analyses based on the selected informative FA-specific probes consistently documented a clear clustering with FA samples, while the two other individuals were placed in the control group (Figure 3A). In line with these findings, the SVM classifier unambiguously scored the first 7 samples as matching FA while rejecting FA diagnosis in the remaining two cases (Figure 2B; Table S4). Subsequent molecular reassessment of individual 47 led us to identify a homozygous exon 5 deletion in *FANCA*. While the genetic defects could not be identified in the remaining two subjects, a positive DEB test was found in all 7 samples, confirming the DNAm profiling findings. Similarly, DEB testing performed in individual 72 was negative, in line with the DNAm profiling finding. No additional analyses could be performed in individual 71.

The FA DNAm signature correctly classifies DNA samples obtained from BM aspirates and demonstrates specificity using the EKD

Due to the different cell composition of PB and BM, we then tested whether the generated DNAm signature could be successfully applied to classify genomic DNA samples obtained from BM aspirates of 10 affected individuals with molecularly confirmed clinical diagnosis of FA and 10 age-matched non-FA subjects, the latter including 9 healthy subjects and one individual with RCC (Tables S1 and S2). All samples correctly matched with their relative group in the MDS and unsupervised hierarchical clustering analyses (Figure 3B) and were properly classified by the SVM scoring with high confidence (Figure 2B; Table S4), further providing evidence of robustness.

The specificity of the generated DNAm signature was finally tested using the EKD dataset, the largest EPIC database. A classifier was created by training the 25 FANCA samples against the matched control samples and 75% of control samples and samples with other disorders from the previously published clinical EpiSign v3 classifier within EKD.[20] Based on this SVM classifier, methylation variant pathogenicity (MVP) scores between 0 and 1 were generated to measure the similarity of EKD samples to the identified episignature. The remaining 25% of samples were reserved for testing the model.[20] The model demonstrated high specificity, with all but 3 testing samples from healthy controls and other genetic disorders

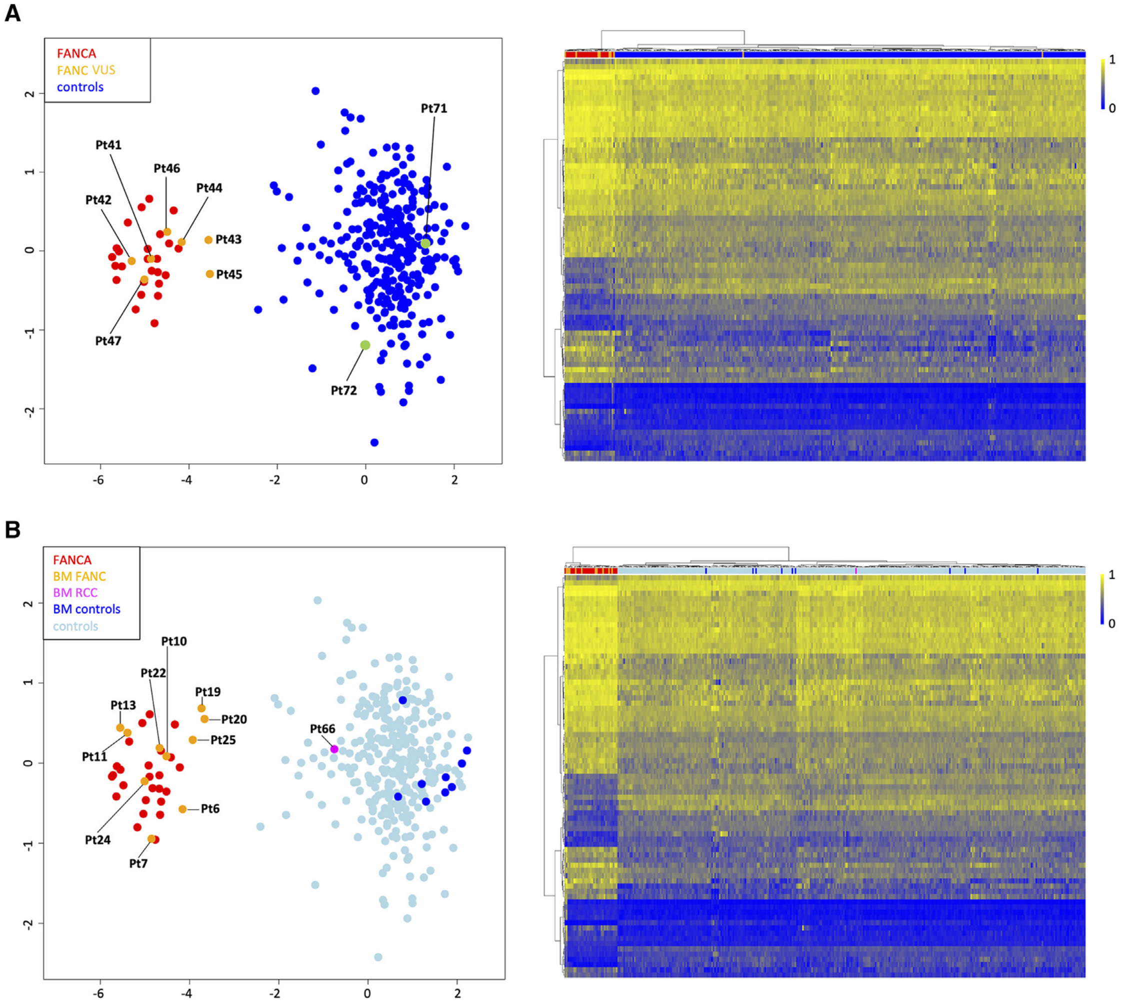


Figure 3. The FA DNAm signature successfully classifies molecularly or clinically unsolved FA cases as well as bone marrow aspirate samples

(A) MDS (left) and heatmap (right) plots showing the clustering of 9 "unsolved" cases who were tested using the FA DNAm signature. Clustering with FA samples is observed for 7 subjects with clinical features fitting or suggestive of FA with inconclusive molecular data (orange), confirming diagnosis of FA. In these cases, DEB testing validated the DNAm analysis. The remaining 2 individuals (green) were found to cluster with the control group (blue), rejecting a diagnosis of FA. DEB testing performed in one of the two cases confirmed the DNAm finding. The FANCA samples used for training are shown in red. Sample groups are indicated using color bars above the heatmap.
(B) MDS (left) and heatmap (right) plots showing the episignature robustness in properly classifying BMA samples from FA cases (orange) and those from healthy individuals (blue) and a subject with RCC (magenta). FANCA and control samples from peripheral blood are depicted in red and light blue, respectively. Sample groups are indicated using color bars above the heatmap.

(n = 1,927) classified as not matching the FA episignature when using a cutoff MVP score of 0.25. Genetic information to verify/exclude occurrence of biallelic FA gene variants in these individuals was not available. The model also exhibited high sensitivity, with all FANCA samples and the majority of samples belonging to different FA complementation groups classified as FA with high confidence (MVP score: 0.50) (Figure 4; Table S5), demonstrating a broader applicability of the signature as part of an established clinical diagnostic classifier.

In silico functional analysis of FA DMRs

Comparison of the DNAm profiles characterizing the FANCA and control cohorts identified 124 FA-specific DMRs, the majority showing DNA hypermethylation (91%) (i.e., higher DNAm levels than controls) (Figure S5). Functional annotation of DMRs did not reveal consistent enrichment of any particular biological processes/pathways using annotated gene sets from Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and gene set enrichment analysis (GSEA) (Table S6). By exploring the genomic distribution of the probes identifying the FA-specific DMRs, a striking enrichment in CpG islands was observed (47% vs. 19%) (Figure S6A). Such enrichment mainly involved gene bodies, which are positively correlated with gene expression when methylated, and was underrepresented in promoters (TSS200/1500) and 1st exon regions, which are generally poorly methylated in actively transcribed genes (Figure S6B). Remarkably, an opposite distribution was observed when considering shelves and open-sea regions.

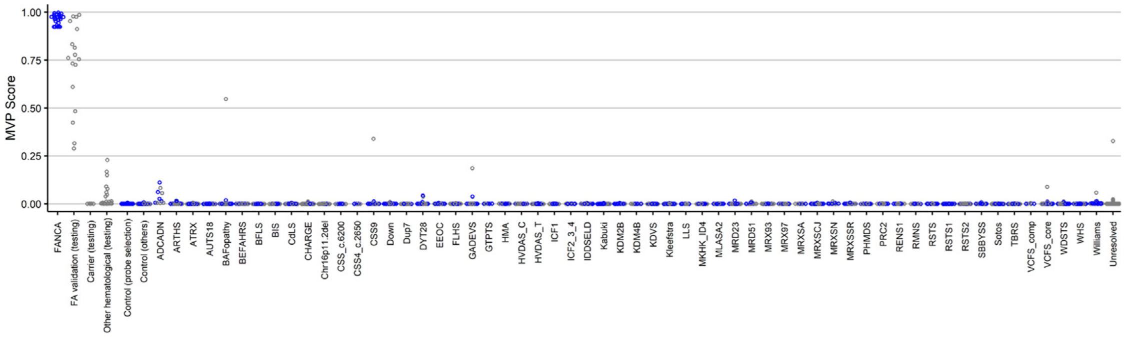


Figure 4. The FA DNAm signature demonstrates high specificity using different datasets from the EpiSign knowledge database
Training samples are depicted in blue, while testing samples are in gray. For each tested sample, the MVP score was generated using an SVM classifier. With only 3 exceptions, all testing samples from controls and other rare genetic disorders (approx. 2,000 samples) received MVP scores <0.25, demonstrating an overall high specificity of the model. Genetic information to verify/exclude the occurrence of biallelic FA gene variants in the three non-FA cases was not available. Additionally, healthy carriers and samples from other hematological disorders also received scores below 0.25, further confirming the model's specificity to FA. All tested FA individuals, including the revertant ones due to gene conversion, received scores above the cut-off value of 0.25, indicating an overall high sensitivity of the model. ADCADN, cerebellar ataxia deafness and narcolepsy syndrome (MIM: 604121); ARTHS, Arboleda-Tham syndrome (MIM: 616268); ATRX, X-linked alpha-thalassemia/impaired intellectual development syndrome (MIM: 300032); AUTS18, susceptibility to autism 18 (MIM: 615032); BEFAHRS, Beck-Fahrner syndrome (MIM: 618798); BFLS, Borjeson-Forssman-Lehmann syndrome (MIM: 301900); BIS, blepharophimosis-intellectual disability *SMARCA2* syndrome (MIM: 619293); CdLS, Cornelia de Lange syndrome (MIM: 122470); CSS_c.6200, Coffin-Siris syndrome 1,2 (MIM: 135900 and 614607), missense variants within the BAF250_C domain; CSS4_c.2650, Coffin-Siris syndrome 4 (MIM: 614609), missense variants within the helicase ATP-binding domain; CSS9, Coffin-Siris syndrome 9 (MIM: 615866); DYT28, dystonia 28 (MIM: 617284); EEOC, epileptic encephalopathy-childhood onset (MIM: 615369); FLHS, Floating-Harbour syndrome (MIM: 136140); GADEVS, Gabriele-de Vries syndrome (MIM: 617557); GTPTS, genitopatellar syndrome (MIM: 606170); HMA, Hunter-McAlpine craniosynostosis syndrome (MIM: 601379); HVDAS, Helsmoortel-van der Aa syndrome (MIM: 615873), central region (C), terminal region (T); ICF, immunodeficiency-centromeric instability-facial anomalies syndrome (MIM: 242860); IDDSELD, intellectual developmental disorder with seizures and language delay (MIM: 619000); KDM2B, KDM2B-related neurodevelopmental disorder; KDM4B, intellectual developmental disorder, autosomal dominant 65 (MIM: 619320); KDVS, Koolen-De Vries syndrome (MIM: 610443); LLS, Luscan-Lumish syndrome (MIM: 616831); MKHK, Menke-Hennekam syndrome 1 and 2 (MIM: 618332 and 618333), ID4 domains (ID4); MLASA2, myopathy lactic acidosis and sideroblastic anemia 2 (MIM: 613561); MRD, intellectual developmental disorder (MRD23 [MIM: 615761], MRD51 [MIM: 617788]); MRX93, intellectual developmental disorder X-linked (MIM: 300659); MRXSA, intellectual developmental disorder X-linked syndromic Armfield type (MIM: 300261); MRXSCJ, intellectual developmental disorder X-linked syndromic Claes-Jensen type (MIM: 300534); MRXSN, intellectual developmental disorder X-linked syndromic Nascimento type (MIM: 300860); MRXSSR, intellectual developmental disorder X-linked syndromic Snyder-Robinson type (MIM: 309583); PHMDS, Phelan-McDermid syndrome (MIM: 606232); PRC2, PRC2 complex (Weaver and Cohen-Gibson) syndrome; RENS1, Renpenning syndrome (MIM: 309500); RMNS, Rahman syndrome (MIM: 617537); RSTS, Rubinstein-Taybi syndrome (MIM: 180849); SBBYSS, Ohdo syndrome (MIM: 603736); TBRS, Tatton-Brown-Rahman syndrome (MIM: 615879); VCFS, velocardiofacial syndrome (MIM: 192430), deletions of chromosome 22q11.2 with the typical deletion range (core), comprehensive (comp), also including proximal deletion range; WDSTS, Wiedemann-Steiner syndrome (MIM: 605130); WHS, Wolf-Hirschhorn syndrome (MIM: 194190).

The genomic distribution of DMRs was also assessed considering the genome-wide maps of histone modifications, chromatin accessibility, DNAm, and mRNA expression across 127 human cell types and tissue provided by the NIH Roadmap Epigenomics Consortium,[40] documenting an enrichment of regions associated with both active (i.e., active transcription start sites [TSSs] and flanking regions of active TSSs) and inactive states (i.e., bivalent poised TSSs, bivalent enhancers, and regions flanking bivalent TSS enhancers) (Figure S7). Notably, both regions are characterized by H3K4me1/3 markers and low percentage of methylated CpGs in healthy individuals.[40] FA-specific DMRs, however, did not display any clear cell lineage-/tissue-specific patterns (Figure S8).

To further explore the occurrence of functional selection of DMRs in FA, we looked for statistically significant DMP enrichments ($p < 10^{-6}$) in genes with roles in biological processes dysregulated in FA (Figure S9). Among the 6 identified genes, a single DMR was identified to involve *ZFPM1* (MIM: 601950) (Table S7), which encodes a well-established transcription factor, FOG1, with a role in the regulation of erythropoiesis.[41,42] The biological significance of the hypermethylated status at this locus requires further investigation.

Discussion

We identify a disease-specific DNAm signature associated with biallelic loss-of-function *FANCA* variants. This episignature demonstrates broad specificity, properly classifying FA cases of different complementation groups and discriminating FA from other clinically overlapping hematological and other genetic disorders. It is robust as it can be applied to categorize DNAm profiles generated from BM aspirates. Notably, the episignature correctly classified two reverted FA cases, indicating that it is based on the use of CpG sites that do not undergo functional selection.

Despite concerted efforts to standardize guidelines for the clinical classification of sequence variants, VUS functional interpretation remains challenging in the clinic.[43] Negative and inconclusive findings also contribute to leaving the diagnosis undetermined in a significant proportion of cases. Both issues apply to FA, which is also characterized by wide phenotypic heterogeneity and has considerable clinical overlap with other BMF diseases, making the diagnosis of this disorder even more challenging in some individuals based on their clinical manifestations, particularly in the absence of pancytopenia. In these subjects, FA is suspected based on the genomic findings. As a consequence, validation of the genetic findings or a clinical hypothesis of FA commonly requires DEB testing, which is considered the gold standard for FA diagnosis. While often used as a first-line test for subjects with clinical suspicion of FA, DEB testing requires appropriate expertise and dedicated biosamples, is labor intensive and time consuming, and might provide false-negative results. Our results support the use of DNAm profiling as a complementary diagnostic tool for FA. The identification of a DNAm signature for FA makes classification of VUSs possible as well as establishing or refuting a clinical diagnosis in molecularly uninformative and revertant cases. Notably, the episignature is sufficiently robust to be successfully applied considering BM aspirates as a tissue source. While it has been generated specifically for subjects with biallelic pathogenic *FANCA* variants who represent the most common FA subgroup, the generated DNAm signature has successfully been applied to subjects belonging to different complementation groups, documenting wide specificity, in principle. Testing of larger validation cohorts representative of the other FA subgroups is, however, a required step to validate its use cross-sectionally and to consider this tool as a first-line diagnostic approach.

Somatic mosaicism in FA can occur from reversion or other compensatory mutations in hematopoietic progenitor cells from which eventually a stem cell population with functional DNA repair capacity emerges.[44] In this case, DEB/MMC testing on PB may provide negative results; in the presence of clinical signs/features highly suggestive of FA, the test is usually performed on other cell lineages (e.g., skin fibroblasts or hair follicles), which might lead to confirmation of mosaic FA. By applying DNAm profiling, we proved the first-tier diagnostic capability of this signature in reverted FA resulting from gene conversion. This finding indicates that the identified DNAm signature sites are not subjected to functional selection and that their methylation status keeps memory of the originally perturbed epigenetic landscape independently from relevant functional changes controlling cellular endophenotypes. In line with these considerations, functional annotation of FA DMRs excluded enrichment of biological processes/pathways and involvement of genes participating in pathways relevant to hematopoiesis and BM failure. Our findings imply that the selected CpG sites can be considered as “passive footprints” of prior DNA damage/repair events associated with the defective use of homologous recombination to repair DNA interstrand crosslinks.[9] In FA, the impaired function of the FA/BRCA pathway causes accumulation of toxic DNA double-strand breaks and the use of alternative error-prone DNA repair pathways (e.g., non-homologous end-joining [NHEJ] repair), which is believed to cause increased susceptibility in structural rearrangements in the genome. Of note, the occurrence of a specific genome-wide DNAm pattern in FA suggests that the repair of DNA double-strand breaks by NHEJ repair may play a role in perturbing the landscape of the DNAm status. Consistent with our findings, a direct involvement of NHEJ repair in rewriting or revising the methylation status of the repaired genomic regions has been suggested as a putative source contributing to the altered methylation patterns characterizing cancer cells.[45] Indeed, deep sequencing analysis of post-repair DNA has documented the occurence of both loss and gain of DNAm in areas flanking the break sites. Based on the consideration that double-strand breaks are largely stochastic events, the identification of shared aberrantly methylated CpG sites suggests that specific genomic regions might be prone to NHEJ repair-associated DNAm rewriting. On the other hand, histone methylation changes are known to occur in response to DNA damage to orchestrate damage-induced chromatin state transition and DNA damage response,[46,47] which is expected to indirectly impact the DNAm status. The finding that double-strand breaks induce DNA hypermethylation[48,49] is in line with the overall hypermethylated status characterizing the DMPs of FA subjects compared with control subjects.

DNAm signature testing has been demonstrated to overcome both technical and interpretative limitations of the diagnostic workflow based on genome scan. To date, over 60 Mendelian disorders have been associated with specific DNAm profiles that have been successfully applied to clinical diagnostics.[20] Mostly characterized by developmental delay/intellectual disability, these genetic conditions constitute an increasingly diverse group of diseases. While originally functionally linked to chromatin accessibility regulation, recent findings have documented that DNAm signatures can also be identified in disorders involving genes that are not directly related to the epigenetic machinery. This work demonstrates that genome-wide DNAm profiling can be used as an informative diagnostic tool in hematological disorders.

Data and code availability

Some of the datasets used in this study are publicly available and may be obtained from Gene Expression Omnibus (GEO) using the following accession numbers: GSE11 6992, GSE66552, GSE74432, GSE97362, GSE116300, GSE 95040, GSE104451, GSE125367, GSE55491, GSE108423, GSE89353, GSE52588, GSE42861, GSE85210, GSE87571, GSE87648, GSE99863, and GSE35069. The remaining generated and analyzed DNAm datasets supporting the study are not publicly available due to privacy/ethical/legal restrictions. Data are available from one of the corresponding authors (M.T.) upon reasonable request and with permission of individual participating subject/legal guardian. The R code used for DNAm signature identification, validation, and testing (OPBG, Rome, Italy) is available at GitHub (https://github.com/Ferix96/EpiMethHub.git).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.09.014.

Acknowledgments

We are grateful to the families who participated in this study. This work was supported, in part, by the Italian Ministry of Health (CCR-2017-23669081, RCR-2022-23682289 and Ricerca Corrente 2019 to M.T.; Ricerca Corrente 2023 to A.C.) and Genome Canada Genomic Applications Partnership Program (to B.S.).

Declaration of interests

Dr. Sadikovic is a shareholder in EpiSign Inc, a software company involved in commercialization of EpiSign Technology.

Received: June 29, 2023
Accepted: September 27, 2023
Published: October 20, 2023

Web resources

ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/
dbSNP, https://www.ncbi.nlm.nih.gov/snp
DMRcate, https://code.bioconductor.org/browse/DMRcate/
EpiSign Knowledge Database, https://episign.lhsc.on.ca/knowledge_database.html#
e1071, https://cran.r-project.org/web/packages/e1071/
Fanconi Anemia Mutation Database, https://www2.rockefeller.edu/fanconi/
Gene Ontology (GO), http://geneontology.org/
ggplot2, https://ggplot2.tidyverse.org
gnomAD, https://gnomad.broadinstitute.org/
Human Gene Mutation Database, https://www.hgmd.cf.ac.uk/ac/index.php
imbalance, http://github.com/ncordon/imbalance
Kyoto Encyclopedia of Genes and Genomes (KEGG), https://www.genome.jp/kegg/
Leiden Open Variation Database, https://www.lovd.nl/
Molecular Signatures Database (GSEA), https://www.gsea-msigdb.org/gsea/msigdb
OMIM, http://www.omim.org/

References

1. Kutler, D.I., Singh, B., Satagopan, J., Batish, S.D., Berwick, M., Giampietro, P.F., Hanenberg, H., and Auerbach, A.D. (2003). A 20-year perspective on the International Fanconi Anemia Registry (IFAR). Blood *101*, 1249–1256.
2. Auerbach, A.D. (2009). Fanconi anemia and its diagnosis. Mutat. Res. *668*, 4–10.
3. Mehta, P.A., and Ebens, C. (2002). Fanconi Anemia. In GeneReviews Seattle (WA), M.P. Adam, G.M. Mirzaa, R.A. Pagon, S.E. Wallace, L.J. Bean, K.W. Gripp, and A. Amemiya, eds. (University of Washington, Seattle), pp. 1993–2023.
4. Schneider, M., Chandler, K., Tischkowitz, M., and Meyer, S. (2015). Fanconi anemia: genetics, molecular biology, and cancer – implications for clinical management in children and adults. Clin. Genet. *88*, 13–24.
5. Dufour, C., and Pierri, F. (2022). Modern management of Fanconi anemia. Hematology. Am. Soc. Hematol. Educ. Program *2022*, 649–657.
6. D'Andrea, A.D., and Grompe, M. (2003). The Fanconi anaemia/BRCA pathway. Nat. Rev. Cancer *3*, 23–34.
7. Kee, Y., and D'Andrea, A.D. (2010). Expanded roles of the Fanconi anemia pathway in preserving genomic stability. Genes Dev. *24*, 1680–1694.
8. Ceccaldi, R., Sarangi, P., and D'Andrea, A.D. (2016). The Fanconi anaemia pathway: new players and new functions. Nat. Rev. Mol. Cell Biol. *17*, 337–349.
9. García-de-Teresa, B., Rodríguez, A., and Frias, S. (2020). Chromosome Instability in Fanconi Anemia: From Breaks to Phenotypic Consequences. Genes *11*, 1528.
10. Badra Fajardo, N., Taraviras, S., and Lygerou, Z. (2022). Fanconi anemia proteins and genome fragility: unraveling replication defects for cancer therapy. Trends Cancer *8*, 467–481.
11. Weck, K.E. (2018). Interpretation of genomic sequencing: variants should be considered uncertain until proven guilty. Genet. Med. *20*, 291–293.
12. Wong, A.K., Sealfon, R.S.G., Theesfeld, C.L., and Troyanskaya, O.G. (2021). Decoding disease: from genomes to networks to phenotypes. Nat. Rev. Genet. *22*, 774–790.
13. D'haene, E., and Vergult, S. (2021). Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. Genet. Med. *23*, 34–46.
14. Auerbach, A.D. (2015). Diagnosis of Fanconi anemia by diepoxybutane analysis. Curr. Protoc. Hum. Genet. *85*, 8.7.1–8.7.17.
15. Greenberg, M.V.C., and Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. Nat. Rev. Mol. Cell Biol. *20*, 590–607.
16. Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nat. Rev. Genet. *14*, 204–220.
17. Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.
18. Fernandez, A.F., Assenov, Y., Martin-Subero, J.I., Balint, B., Siebert, R., Taniguchi, H., Yamamoto, H., Hidalgo, M., Tan, A.C., Galm, O., et al. (2012). A DNA methylation fingerprint of 1628 human samples. Genome Res. *22*, 407–419.
19. Aref-Eshghi, E., Kerkhof, J., Pedro, V.P., Groupe DI France, Barat-Houari, M., Ruiz-Pallares, N., Andrau, J.C., Lacombe, D., Van-Gils, J., Fergelot, P., et al. (2020). Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in 42 Mendelian Neurodevelopmental Disorders. Am. J. Hum. Genet. *106*, 356–370.
20. Levy, M.A., McConkey, H., Kerkhof, J., Barat-Houari, M., Bargiacchi, S., Biamino, E., Bralo, M.P., Cappuccio, G., Ciolfi, A., Clarke, A., et al. (2022). Novel diagnostic DNA methylation

episignatures expand and refine the epigenetic landscapes of Mendelian disorders. HGG Adv. *3*, 100075.

21. 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.
22. Ciolfi, A., Aref-Eshghi, E., Pizzi, S., Pedace, L., Miele, E., Kerkhof, J., Flex, E., Martinelli, S., Radio, F.C., Ruivenkamp, C.A.L., et al. (2020). Frameshift mutations at the C-terminus of HIST1H1E result in a specific DNA hypomethylation signature. Clin. Epigenetics *12*, *7*.
23. Ciolfi, A., Foroutan, A., Capuano, A., Pedace, L., Travaglini, L., Pizzi, S., Andreani, M., Miele, E., Invernizzi, F., Reale, C., et al. (2021). Childhood-onset dystonia-causing KMT2B variants result in a distinctive genomic hypermethylation profile. Clin. Epigenetics *13*, 157.
24. Sadikovic, B., Levy, M.A., Kerkhof, J., Aref-Eshghi, E., Schenkel, L., Stuart, A., McConkey, H., Henneman, P., Venema, A., Schwartz, C.E., et al. (2021). Clinical epigenomics: genome-wide DNA methylation analysis for the diagnosis of Mendelian disorders. Gen. Med. *23*, 1065–1074.
25. Ferilli, M., Ciolfi, A., Pedace, L., Niceta, M., Radio, F.C., Pizzi, S., Miele, E., Cappelletti, C., Mancini, C., Galluccio, T., et al. (2022). Genome-Wide DNA Methylation Profiling Solves Uncertainty in Classifying *NSD1* Variants. Genes *13*, 2163.
26. Tian, Y., Morris, T.J., Webster, A.P., Yang, Z., Beck, S., Feber, A., and Teschendorff, A.E. (2017). ChAMP: updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics *33*, 3982–3984.
27. Fortin, J.P., Triche, T.J., and Hansen, K.D. (2017). Preprocessing, normalization and integration of the Illumina Human Methylation EPIC array with minfi. Bioinformatics *33*, 558–560.
28. Ho, D.E., Imai, K., King, G., and Stuart, E.A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. J Stat Soft *42*. https://doi.org/10.18637/jss.v042.i08.
29. 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.
30. 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 Human Methylation EPIC Bead Array. Genome Biol. *19*, 64.
31. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag).
32. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. *16*, 321–357.
33. Peters, T.J., Buckley, M.J., Statham, A.L., Pidsley, R., Samaras, K., V Lord, R., Clark, S.J., and Molloy, P.L. (2015). De novo identification of differentially methylated regions in the human genome. Epigenet. Chromatin *8*, 6.
34. Maksimovic, J., Oshlack, A., and Phipson, B. (2021). Gene set enrichment analysis for genome-wide DNA methylation data. Genome Biol. *22*, 173.
35. D Turner, S. (2018). Qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J. Open Source Softw. *3*, 731.
36. Layer, R.M., Pedersen, B.S., DiSera, T., Marth, G.T., Gertz, J., and Quinlan, A.R. (2018). GIGGLE: a search engine for large-scale integrated genome analysis. Nat. Methods *15*, 123–126.
37. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.
38. Joenje, H., Arwert, F., Kwee, M.L., Madan, K., and Hoehn, H. (1998). Confounding factors in the diagnosis of Fanconi anaemia. Am. J. Med. Genet. *79*, 403–405.
39. Gross, M., Hanenberg, H., Lobitz, S., Friedl, R., Herterich, S., Dietrich, R., Gruhn, B., Schindler, D., and Hoehn, H. (2002). Reverse mosaicism in Fanconi anemia: natural gene therapy via molecular self-correction. Cytogenet. Genome Res. *98*, 126–135.
40. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
41. Cantor, A.B., and Orkin, S.H. (2002). Transcriptional regulation of erythropoiesis: an affair involving multiple partners. Oncogene *21*, 3368–3376.
42. Mancini, E., Sanjuan-Pla, A., Luciani, L., Moore, S., Grover, A., Zay, A., Rasmussen, K.D., Luc, S., Bilbao, D., O'Carroll, D., et al. (2012). FOG-1 and GATA-1 act sequentially to specify definitive megakaryocytic and erythroid progenitors. EMBO J. *31*, 351–365.
43. Lappalainen, T., and MacArthur, D.G. (2021). From variant to function in human disease genetics. Science *373*, 1464–1468.
44. Fargo, J.H., Rochowski, A., Giri, N., Savage, S.A., Olson, S.B., and Alter, B.P. (2014). Comparison of chromosome breakage in non-mosaic and mosaic patients with Fanconi anemia, relatives, and patients with other inherited bone marrow failure syndromes. Cytogenet. Genome Res. *144*, 15–27.
45. Allen, B., Pezone, A., Porcellini, A., Muller, M.T., and Masternak, M.M. (2017). Non-homologous end joining induced alterations in DNA methylation: A source of permanent epigenetic change. Oncotarget *8*, 40359–40372.
46. Gong, F., and Miller, K.M. (2019). Histone methylation and the DNA damage response. Mutat. Res. Rev. Mutat. Res. *780*, 37–47.
47. Fernandez, A., O'Leary, C., O'Byrne, K.J., Burgess, J., Richard, D.J., and Suraweera, A. (2021). Epigenetic Mechanisms in DNA Double Strand Break Repair: A Clinical Review. Front. Mol. Biosci. *8*, 685440.
48. Cuozzo, C., Porcellini, A., Angrisano, T., Morano, A., Lee, B., Di Pardo, A., Messina, S., Iuliano, R., Fusco, A., Santillo, M.R., et al. (2007). DNA damage, homology-directed repair, and DNA methylation. PLoS Genet. *3*, e110.
49. O'Hagan, H., Mohammad, H.P., and Baylin, S.B. (2008). Double strand breaks can initiate gene silencing and SIRT1-dependent onset of DNA methylation in an exogenous promoter CpG island. PLoS Genet. *4*, e1000155.