The American Journal of Human Genetics, Volume *107*

Supplemental Data

Episignatures Stratifying Helsmoortel-Van Der Aa Syndrome Show Modest Correlation with Phenotype

Michael S. Breen, Paras Garg, Lara Tang, Danielle Mendonca, Tess Levy, Mafalda Barbosa, Anne B. Arnett, Evangeline Kurtz-Nelson, Emanuele Agolini, Agatino Battaglia, Andreas G. Chiocchetti, Christine M. Freitag, Alicia Garcia-Alcon, Paola Grammatico, Irva Hertz-Picciotto, Yunin Ludena-Rodriguez, Carmen Moreno, Antonio Novelli, Mara Parellada, Giulia Pascolini, Flora Tassone, Dorothy E. Grice, Daniele Di Marino, Raphael A. Bernier, Alexander Kolevzon, Andrew J. Sharp, Joseph D. Buxbaum, Paige M. Siper, and Silvia De Rubeis

Figure S1

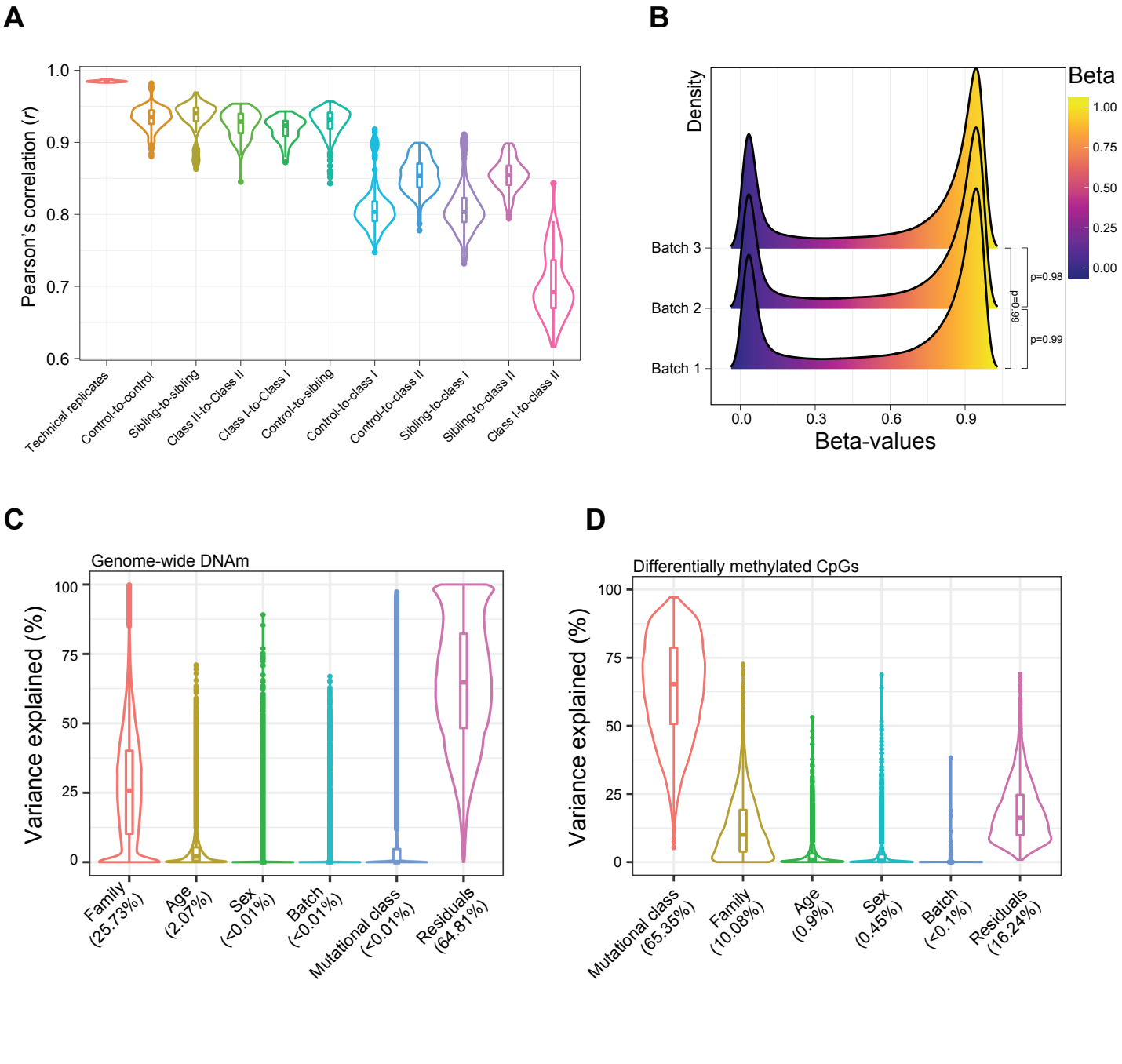
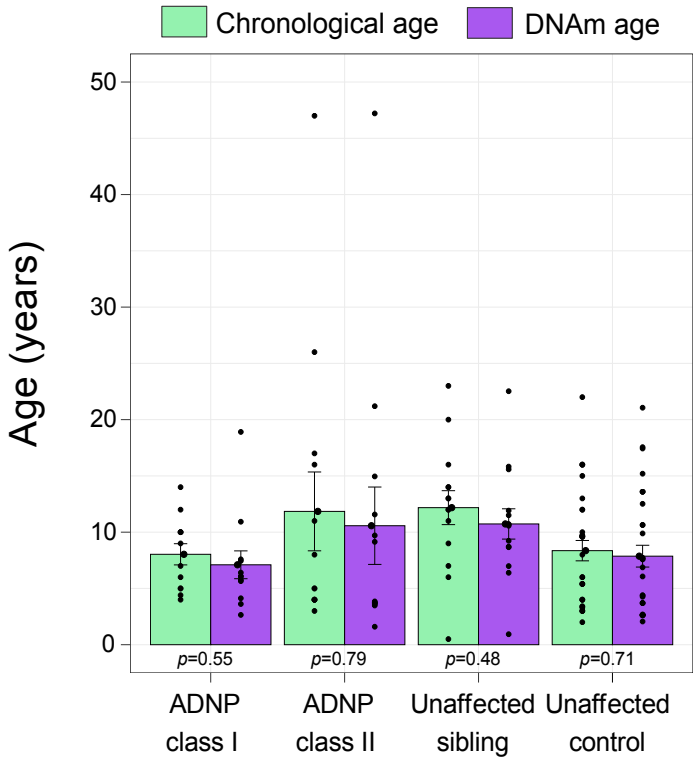

A
Pearson's correlation (r)
1.0
0.9
0.8
0.7
0.6
Technical replicates
Control-to-control
Sibling-to-sibling
Class II-to-Class II
Class I-to-Class I
Control-to-sibling
Control-to-class I
Control-to-class II
Sibling-to-class I
Sibling-to-class II
Class I-to-class II
B
Density
Beta
1.00
0.75
0.50
0.25
0.00
Batch 3
Batch 2
Batch 1
p=0.98
p=0.99
0.0
0.3
0.6
0.9
Beta-values
C
Genome-wide DNAm
Variance explained (%)
100
75
50
25
0
Family (25.73%)
Age (2.07%)
Sex (<0.01%)
Batch (<0.01%)
Mutational class (<0.01%)
Residuals (64.81%)
D
Differentially methylated CpGs
Variance explained (%)
100
75
50
25
0
Mutational class (65.35%)
Family (10.08%)
Age (0.9%)
Sex (0.45%)
Batch (<0.1%)
Residuals (16.24%)

Figure S2

A


Chronological age
DNAm age
Age (years)
0
10
20
30
40
50
p=0.55
p=0.79
p=0.48
p=0.71
ADNP class I
ADNP class II
Unaffected sibling
Unaffected control

B

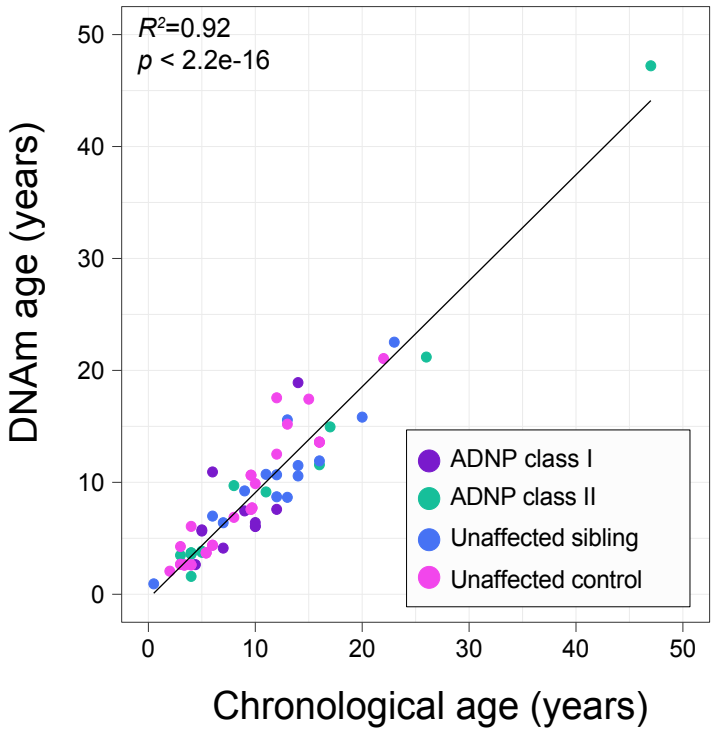

R²=0.92
p < 2.2e-16
DNAm age (years)
Chronological age (years)
0
10
20
30
40
50
ADNP class I
ADNP class II
Unaffected sibling
Unaffected control

Figure S3

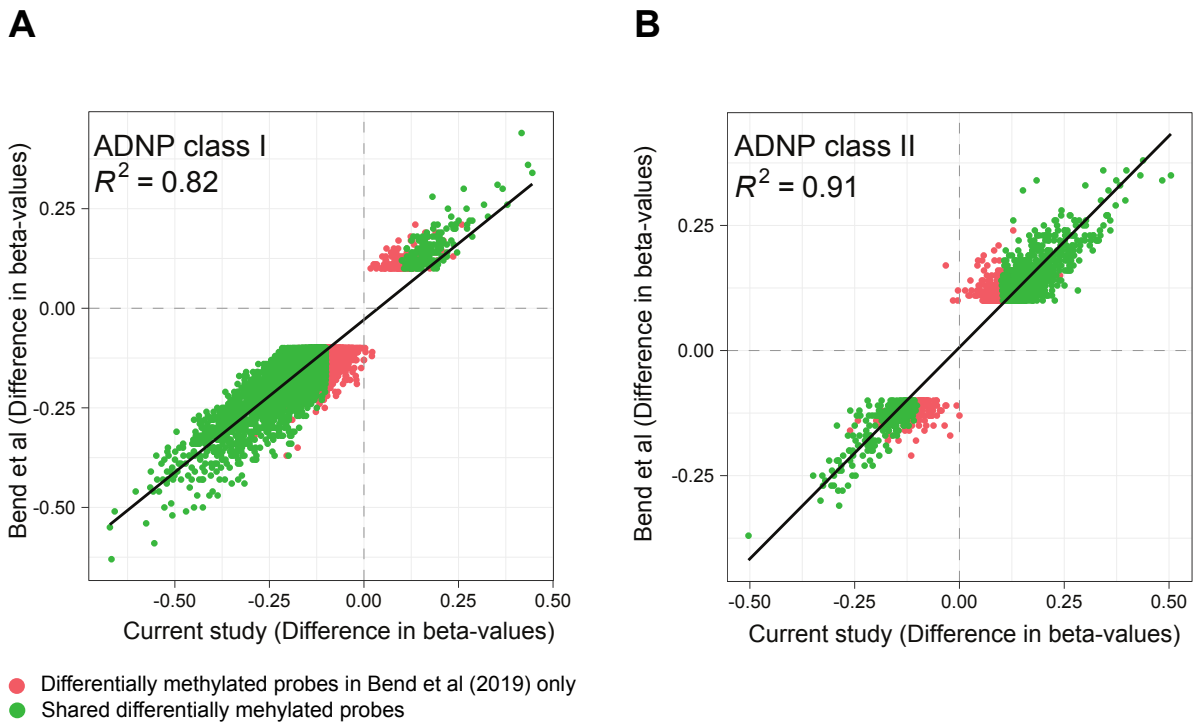
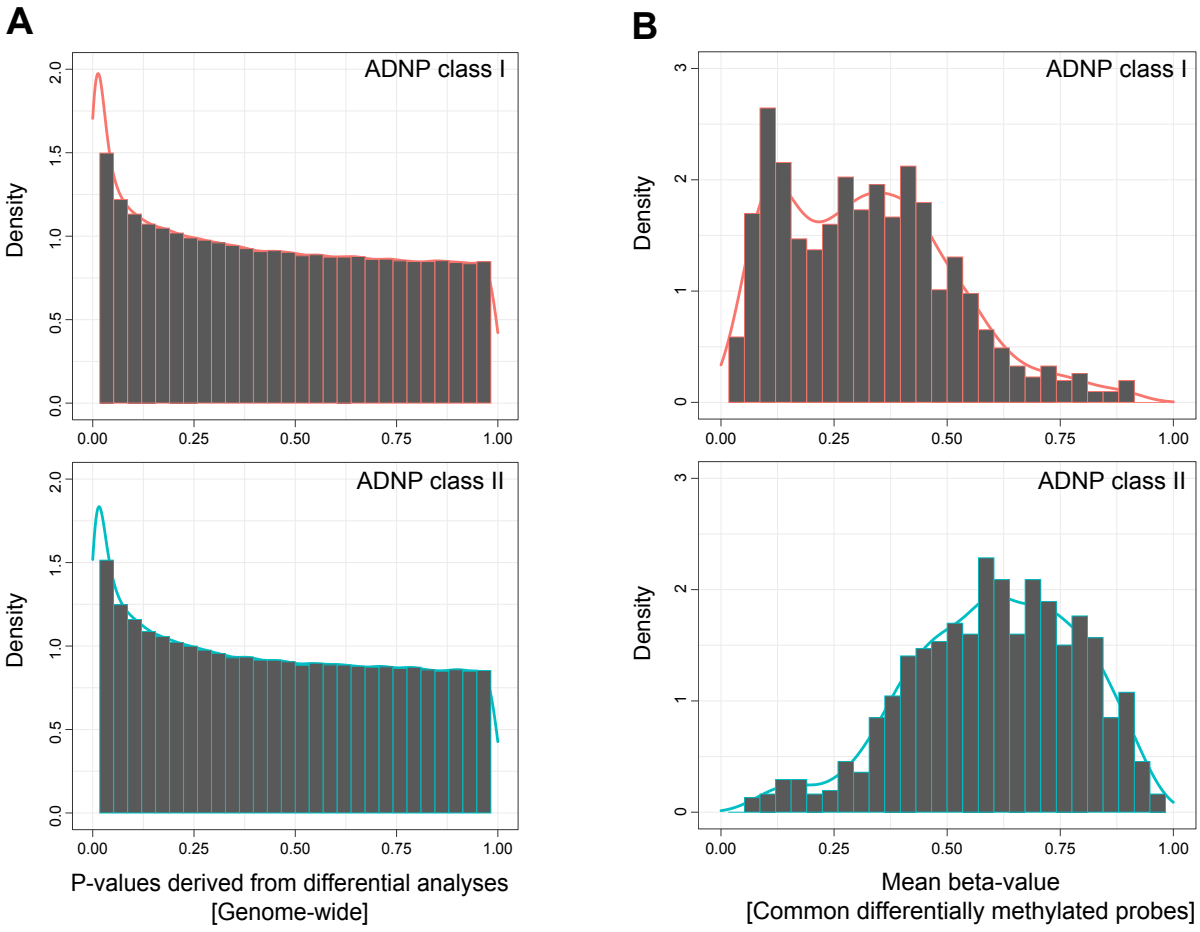

A
ADNP class I
R^2 = 0.82
Bend et al (Difference in beta-values)
Current study (Difference in beta-values)
0.25
0.00
-0.25
-0.50
-0.50
-0.25
0.00
0.25
0.50
B
ADNP class II
R^2 = 0.91
Bend et al (Difference in beta-values)
Current study (Difference in beta-values)
0.25
0.00
-0.25
-0.50
-0.25
0.00
0.25
0.50
Differentially methylated probes in Bend et al (2019) only
Shared differentially mehylated probes

Figure S4


A
ADNP class I
ADNP class II
Density
P-values derived from differential analyses
[Genome-wide]
B
ADNP class I
ADNP class II
Density
Mean beta-value
[Common differentially methylated probes]

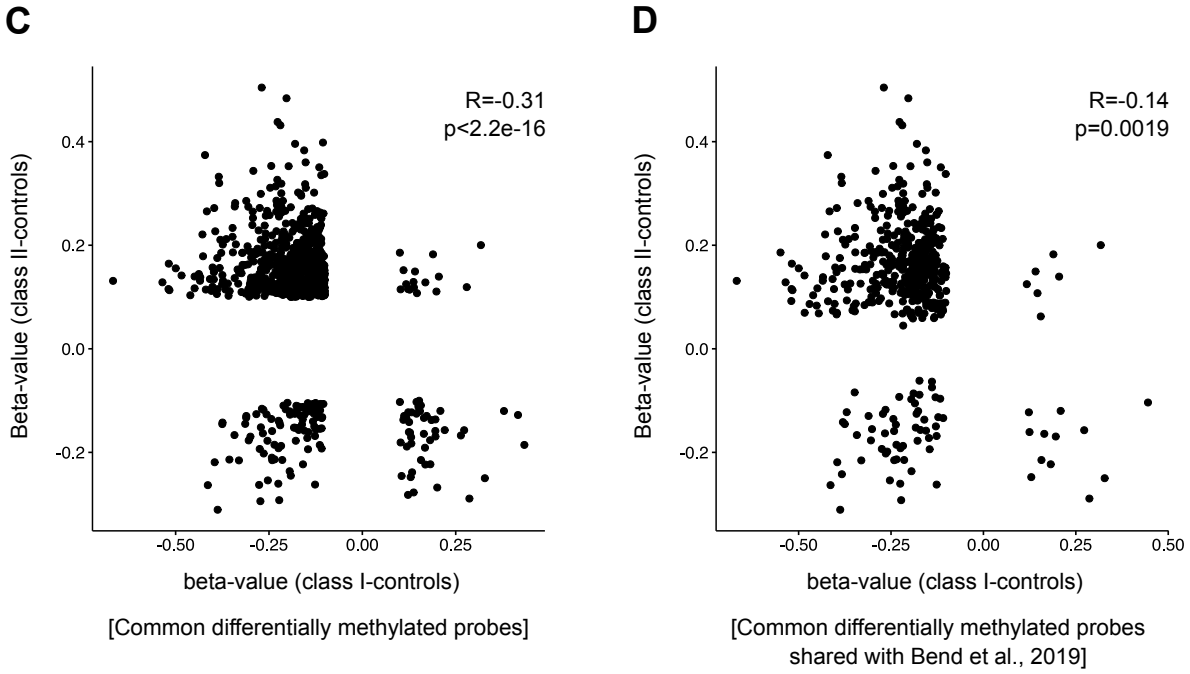

C
R=-0.31
p<2.2e-16
Beta-value (class II-controls)
beta-value (class I-controls)
[Common differentially methylated probes]
D
R=-0.14
p=0.0019
Beta-value (class II-controls)
beta-value (class I-controls)
[Common differentially methylated probes
shared with Bend et al., 2019]

Figure S5

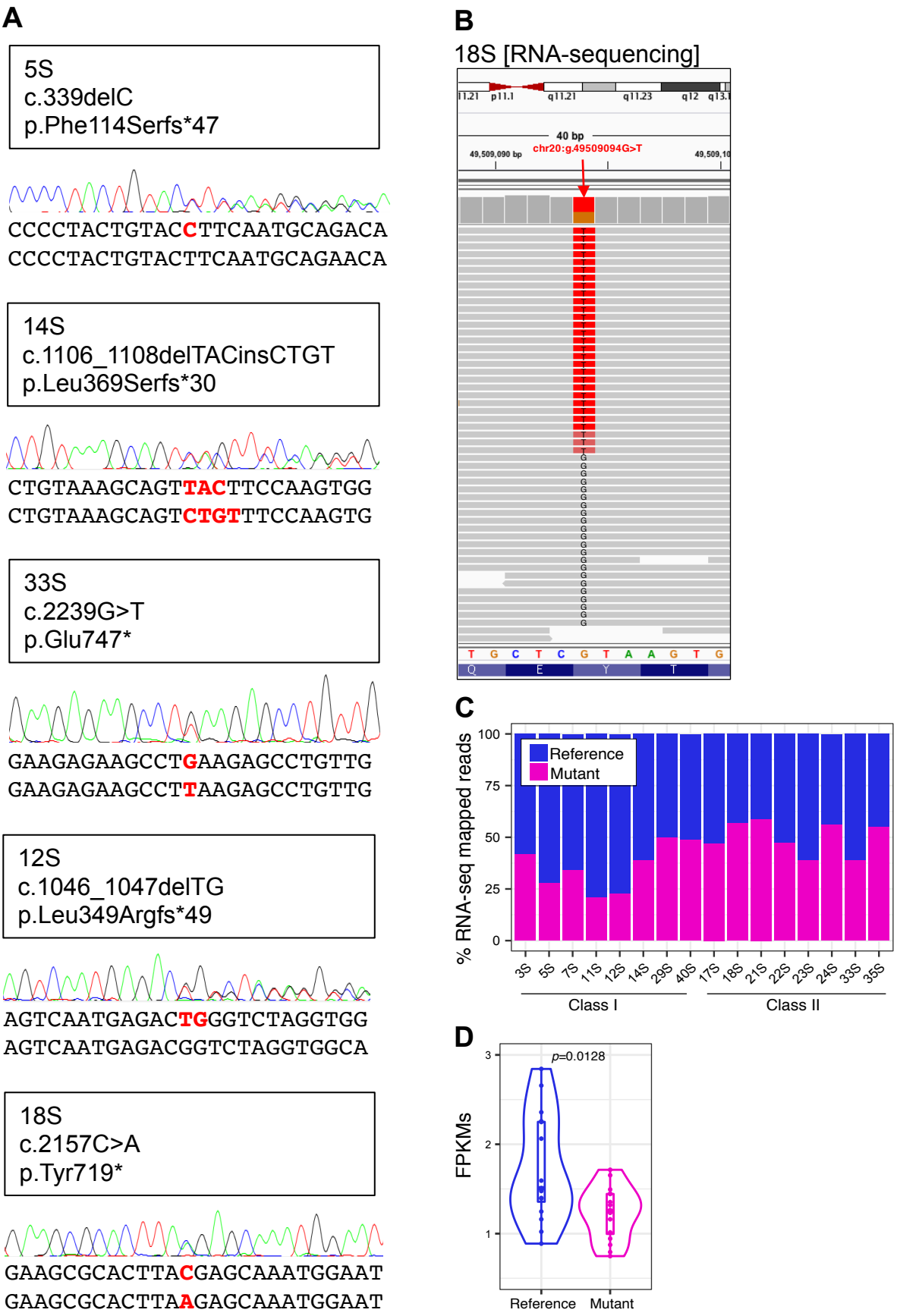
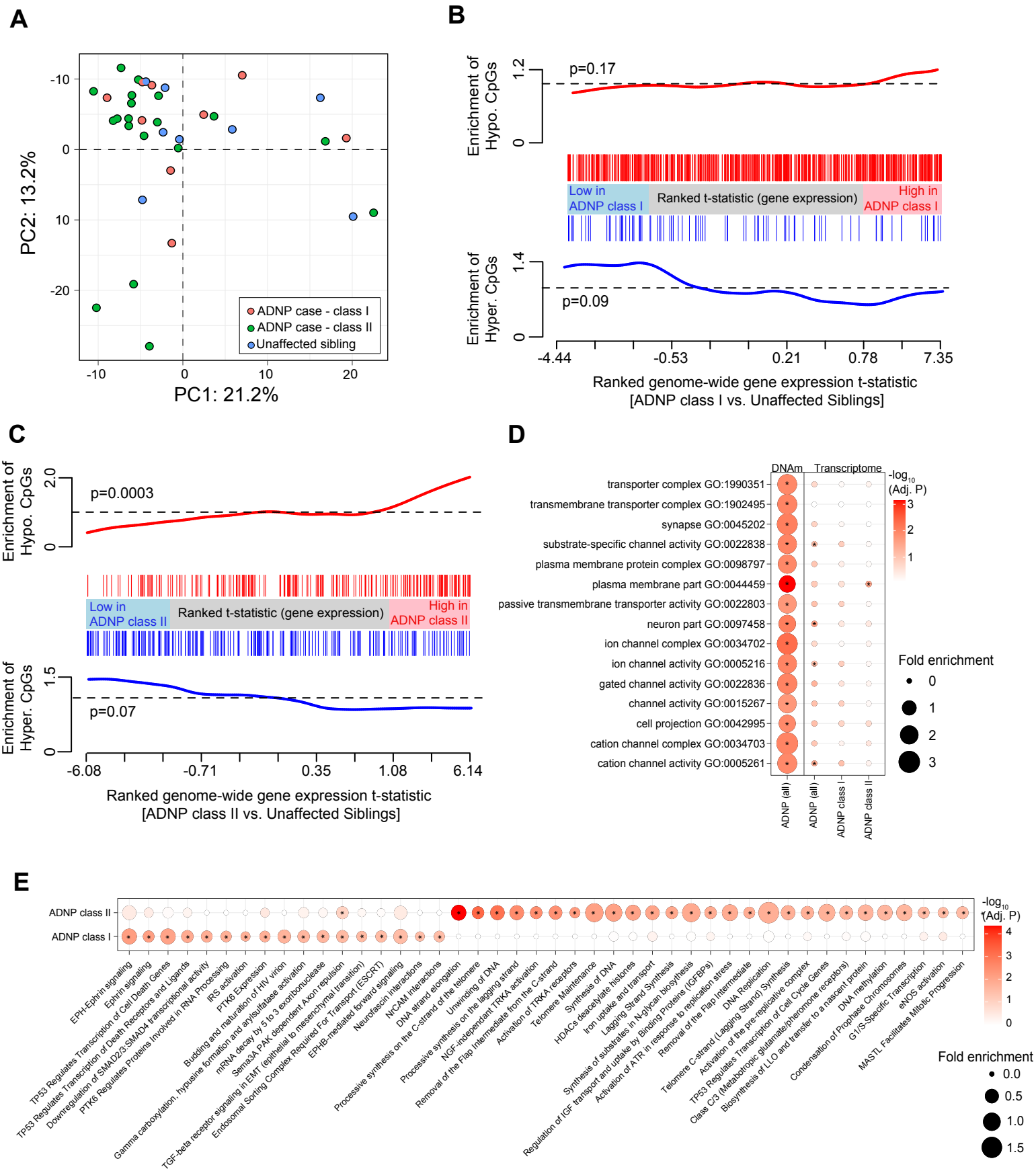

A
5S
c.339delC
p.Phe114Serfs*47
CCCCTACTGTACCTTCAATGCAGACA
CCCCTACTGTACTTCAATGCAGAACA
14S
c.1106_1108delTACinsCTGT
p.Leu369Serfs*30
CTGTAAAGCAGTTACTTCCAAGTGG
CTGTAAAGCAGTCTGTTTCCAAGTG
33S
c.2239G>T
p.Glu747*
GAAGAGAAGCCTGAAGAGCCTGTTG
GAAGAGAAGCCTTAAGAGCCTGTTG
12S
c.1046_1047delTG
p.Leu349Argfs*49
AGTCAATGAGACTGGGTCTAGGTGG
AGTCAATGAGACGGTCTAGGTGGCA
18S
c.2157C>A
p.Tyr719*
GAAGCGCACTTACGAGCAAATGGAAT
GAAGCGCACTTAAGAGCAAATGGAAT
B
18S [RNA-sequencing]
40 bp
chr20:g.49509094G>T
C
% RNA-seq mapped reads
Reference
Mutant
Class I
Class II
D
FPKMs
p=0.0128
Reference
Mutant

Figure S6


A
PC2: 13.2%
PC1: 21.2%
ADNP case - class I
ADNP case - class II
Unaffected sibling
B
Enrichment of Hypo. CpGs
p=0.17
Low in ADNP class I
Ranked t-statistic (gene expression)
High in ADNP class I
Enrichment of Hyper. CpGs
p=0.09
Ranked genome-wide gene expression t-statistic [ADNP class I vs. Unaffected Siblings]
C
Enrichment of Hypo. CpGs
p=0.0003
Low in ADNP class II
Ranked t-statistic (gene expression)
High in ADNP class II
Enrichment of Hyper. CpGs
p=0.07
Ranked genome-wide gene expression t-statistic [ADNP class II vs. Unaffected Siblings]
D
DNAm
Transcriptome
transporter complex GO:1990351
transmembrane transporter complex GO:1902495
synapse GO:0045202
substrate-specific channel activity GO:0022838
plasma membrane protein complex GO:0098797
plasma membrane part GO:0044459
passive transmembrane transporter activity GO:0022803
neuron part GO:0097458
ion channel complex GO:0034702
ion channel activity GO:0005216
gated channel activity GO:0022836
channel activity GO:0015267
cell projection GO:0042995
cation channel complex GO:0034703
cation channel activity GO:0005261
ADNP (all)
ADNP (all)
ADNP class I
ADNP class II
-log10 (Adj. P)
Fold enrichment
E
ADNP class II
ADNP class I
EPH-Ephrin signaling
Ephrin signaling
TP53 Regulates Transcription of Cell Death Genes
TP53 Regulates Transcription of Death Receptors and Ligands
Downregulation of SMAD2/3:SMAD4 transcriptional activity
PTK6 Regulates Proteins Involved in RNA Processing
IRS activation
PTK6 Expression
Budding and maturation of HIV virion
Gamma carboxylation, hypusine formation and arylsulfatase activation
mRNA decay by 5 to 3 exoribonuclease
Sema3A PAK dependent Axon repulsion
TGF-beta receptor signaling in EMT (epithelial to mesenchymal transition)
Endosomal Sorting Complex Required For Transport (ESCRT)
EPHB-mediated forward signaling
Neurotrascin interactions
NrCAM interactions
DNA strand elongation
Processive synthesis on the C-strand of the telomere
Unwinding of DNA
Processive synthesis on the lagging strand
NGF-independant TRKA activation
Removal of the Flap Intermediate from the C-strand
Activation of TRKA receptors
Telomere Maintenance
Synthesis of DNA
HDACs deacetylate histones
Iron uptake and transport
Lagging Strand Synthesis
Synthesis of substrates in N-glycan biosynthesis
Regulation of IGF transport and uptake by Binding Proteins (IGFBPs)
Activation of ATR in response to replication stress
Removal of the Flap Intermediate
DNA Replication
Telomere C-strand (Lagging Strand) Synthesis
Activation of the pre-replicative complex
TP53 Regulates Transcription of Cell Cycle Genes
Class C/3 (Metabotropic glutamate/pheromone receptors)
Biosynthesis of LLO and transfer to a nascent protein
Condensation of Prophase Chromosomes
DNA methylation
G1/S-Specific Transcription
eNOS activation
MASTL Facilitates Mitotic Progression
-log10 (Adj. P)
Fold enrichment

Supplemental Figures and Tables legend

Figure S1. Strong reproducibility across methylation batches. (A) Sample-to-sample Pearson's correlation across the three methylation batches. **(B)** Frequency distribution of β-values for all probes across the three batches. Significance was tested with a Wilcoxon rank sum test. **(C)** The linear mixed modeling frame work of the variancePartition R package was used to partitioned the fraction of genome-wide methylation variance explained by known factors in the dataset, including family (for probands and siblings), age, sex, methylation batch, and mutational class (class I vs. class II). Median variance explained in displayed below each factor in brackets. **(D)** Fraction of variance explained on differentially methylated probes that can be attributed to key variables in the dataset, including family (for probands and siblings), age, sex, methylation batch, and mutational class (class I vs. class II).

Figure S2. Strong correlation between chronological age and methylation-inferred (DNAm) age. (A) Chronological (green) and DNAm age (purple) across the four sample groups: *ADNP* class I mutations, *ADNP* class II mutations, unaffected siblings, and unaffected age-matched controls. **(B)** Correlation between chronological and DNAm age across the four sample groups: *ADNP* class I cases (purple), *ADNP* class II cases (green), unaffected siblings (blue), and unaffected age-matched controls (pink).

Figure S3. Robust convergence between the episignatures found in this study and in a previous study. (A) Scatterplot showing the methylation difference (mean $\beta\text{-values}_{cases}$ – mean $\beta\text{-values}_{controls}$) detected by Bend and colleagues (2019) [1] (Y-axis) and by this study (X-axis) for the probes found differentially methylated in class I individuals only by Bend and colleagues (2019) [1] (red) or by both studies (green). **(B)** Scatterplot showing the methylation difference (mean $\beta\text{-values}_{cases}$ – mean $\beta\text{-values}_{controls}$) detected by Bend and colleagues (2019) [1] (Y-axis) and by this study (X-axis) for the probes found differentially methylated in class II individuals only by Bend and colleagues [1] (red) or by both studies (green).

Figure S4. (A) Frequency distribution of the p-values for all the probes detected in class I cases (top panel) and class II cases (bottom panel). **(B)** Frequency distribution of the mean β-values for the 888 differentially methylated CpGs shared between the two classes for individuals with class I mutations (top panel) or class II mutations (bottom panel). **(C)** Correlation between (mean $\beta\text{-values}_{cases}$ – mean $\beta\text{-values}_{controls}$) class I (X-axis) and (mean $\beta\text{-values}_{cases}$ – mean $\beta\text{-values}_{controls}$) class II (Y-axis) cases for the 888 probes found shared between class I and class II in this study. **(D)** Correlation between (mean $\beta\text{-values}_{cases}$ – mean $\beta\text{-values}_{controls}$) in class I (X-axis) and (mean $\beta\text{-values}_{cases}$ – mean $\beta\text{-values}_{controls}$) in class II (Y-axis) for the 472 probes found shared between class I and class II in both this study and Bend and colleagues (2019) [1].

Figure S5. Biallelic expression of *ADNP* mutant mRNAs for (**A**) three class I mutations (p.Phe114Serfs*47, p.Leu349Argfs*49, and pLeu369Serfs*30) and two class II mutations (p.Tyr719* and p.Glu747*). The Sanger sequencing traces, alongside the cDNA sequence for reference (top) and mutant (bottom) alleles are shown. (**B**) RNA-sequencing pile-up of class II mutation p.Tyr719* (sample 18S) displaying coverage and number of mutant reads (red) and reference (*i.e.* healthy) reads (grey). (**C**) RNA-sequencing quantified the total fraction of *ADNP* mutant alleles and reference alleles present for each sample. Note, sample 1S was not included in this analysis (Table S1). (**D**) The distribution of expression (FPKM) for all mutated reads aggregated together relative to reference reads of *ADNP* across all samples. We observed lower expression of the mutant allele (Mann-Whitney U Test, P=0.01).

Figure S6. (**A**) Principal component analysis (PCA) on the RNAseq data from 17 *ADNP* cases (class I, red; class II green) and 19 unaffected sibling controls (blue), showing no distinct separation between mutational classes or controls. PCA was constructed using only genes harboring significant differentially methylated CpGs within their respective gene body. Competitive gene-set enrichment analysis examined over- and under-expression of genes harboring differentially methylated CpGs in their respective promoters. The analysis is partitioned to evaluate hypomethylated CpGs relative to over-expressed genes and hypermethylated CpGs relative to under-expressed genes, respectively for (**B**) class I cases and (**C**) and class II cases. Hypomethyated CpGs mapped to over-expressed genes in individuals with class II mutations (P=0.0003). No other significant associations were detected. (**D**) Supervised GO enrichment of pathways found to be significantly enriched for differentially methylated CpGs (labeled as DNAm) assessed three ways using

transcriptome data: *i)* using all cases versus controls (all); *ii)* class I mutations versus controls (class I); and, *iii)* class II mutations versus controls (class II). (**E**) Exploratory GO enrichment analysis and the top enrichment terms for individuals with class I and class II mutations.

Table S1. Genetic and demographic information about the 43 *ADNP* individuals included in the study. For each patient, the position of the genetic mutation on cDNA, protein and genomic DNA, coding exon location, effect, inheritance, classification of pathogenicity, mutation class, data availability, and the methylation batch are shown. Demographic information includes sex and country of origin.

Table S2. Differentially methylated CpGs, differentially methylated regions (defined as three or more differentially methylated CpGs within 1 kb of each other), and differentially expressed genes in individuals with class I and class II *ADNP* mutations.

Table S3. Extended version of Table 1, with separate and combined analyses of cohorts S and W across all clinical measures.

Supplemental Subjects and Methods

Cohorts, patients and genetic information

The study comprises four cohorts. Cohort A was collected as part of the Autism Sequencing Consortium [2; 3]. Cohort R was collected at the Medical Genetics Laboratory of Ospedale San Camillo-Forlanini [4] or Ospedale Pediatrico Bambino Gesù, Rome, Italy. Cohort S was collected at the Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai. Cohort W was collected at the University of Washington. Mutations in Cohort A were identified through research-based whole-exome sequencing [2; 3] and validated by Sanger sequencing. Mutations in Cohort R, S, and W were identified and validated in Clinical Laboratory Improvement Amendments (CLIA)-certified laboratories. Participation was approved by the Institutional Review Boards of participating sites. All caregivers provided informed written consent and assent was obtained when appropriate. The methylation analyses were performed on cohorts A, R and part of S because of DNA sample availability. The transcriptomic analyses were performed on part of S because this was the only cohort with RNA samples available. The phenotypic analyses were performed on cohorts S and W because they were the only two collected prospectively. Columns M-O of Table S1 reports the analyses performed on each of the 43 individuals. Specifically, we used 24 cases (cohorts A, R, and part of S) for methylation analyses; 30 cases (S and W) for phenotypic analyses; and, 17 cases (part of cohort S) for RNA-sequencing analysis. The 19 unaffected age-matched and 14 unaffected sibling controls were from cohort S. All mutations are described according to the Human Genome Variation Society (HGVS) guidelines for mutation nomenclature. The cDNA and amino acid positions are annotated according to RefSeq mRNA and protein sequence (NM_015339.4 and NP_056154.1). Nucleotide numbering referring to cDNA uses +1 as the A of the ATG translation initiation codon in the reference sequence, with the initiation codon as codon 1. *ADNP* contains five exons, of which three are coding (Fig. 1A, Table S1). Our dataset includes a p.Gln67His missense mutation (individual 3S, see Table 1) that was clinically evaluated as likely pathogenic by Ambry Genetics. In addition to being *de novo*, absent in population databases, and located at a conserved amino acid and nucleotide position, Ambry's Translational Genomics laboratory pursued RNA studies that showed that the variant impacts splicing and causes the in-frame skipping of coding exon 2, leading to a p.Glu37_Gln67del in-frame deletion.

Annotation of ADNP protein domains

To verify that the ADNP domains annotated in the literature have strong support and probe the existence of additional domains, we used the ELM resource [5] to predict linear motifs. We found that the proteins contain eight zinc finger domains instead of the nine originally annotated using Pfam in 2001 [6]. All eight zinc fingers in ADNP are of the cysteine-cysteine-histidine-histidine (C2H2) type. Also, ELM re-mapped the boundaries of the homeobox domain (see table below) and identified eight low complexity regions. Information on the re-annotation used in Figure 1A is in the table below:

Domain	Start	End	Method
Zinc Finger 1	74	97	ELM prediction
Zinc Finger 2	107	129	ELM prediction
Low Complexity Region	130	141	ELM prediction
Zinc Finger 3	165	188	ELM prediction
Zinc Finger 4	221	244	ELM prediction
Low Complexity Region	393	418	ELM prediction
Low Complexity Region	424	441	ELM prediction
Zinc Finger 5	447	469	ELM prediction
Zinc Finger 6	489	510	ELM prediction
Zinc Finger 7	512	535	ELM prediction
Low Complexity Region	584	597	ELM prediction

Zinc Finger 8	622	647	ELM prediction
Nuclear Localization Sequence (NLS)	714	733	Literature [7]
Low Complexity Region	725	749	ELM prediction
Homeobox	769	810	ELM prediction
HP1-binding site	820	824	Literature [7; 8]
Low Complexity Region	871	889	ELM prediction
Low Complexity Region	953	964	ELM prediction
Low Complexity Region	1006	1017	ELM prediction

Clinical evaluation

Prospective clinical and psychological characterization was completed for 22 individuals seen at the Seaver Autism Center (cohort S) and 10 individuals seen at the University of Washington (cohort W). A battery of standardized assessments was used to examine ASD, intellectual functioning, adaptive behavior, language, motor skills, and sensory processing (see below). The medical evaluation included psychiatric, neurological, and clinical genetics examinations, and medical record review.

ASD phenotype. Gold-standard ASD diagnostic testing included the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) [9], the Autism Diagnostic Interview-Revised (ADI-R) [10], and a clinical evaluation to assess Diagnostic and Statistical Manual for Mental Disorders, Fifth Edition (DSM-5) criteria for ASD [11]. A consensus diagnosis was determined based on results from the ADOS-2, ADI-R, and the clinical evaluation. The ADOS-2 and ADI-R were administered and scored by research reliable raters and the psychiatric evaluation was completed by a board-certified child and adolescent psychiatrist or licensed psychologist. The ADOS-2 is a semi-structured observational assessment that provides scores in the domains of Social Affect, Restricted and Repetitive Behavior, and a total score. A comparison score ranging from 1-10, with higher scores reflecting a greater number of symptoms, was calculated to examine symptom severity within each ADOS-2 domain and in total [12]. Twenty-one individuals received Module 1 of the ADOS, for children who are nonverbal or communicate using single words. Eight individuals received Module 2, for individuals who communicate using phrase speech. One individual received Module 3, for children who are verbally fluent. The ADI-R is a structured caregiver interview that assesses ASD symptomatology within the domains of socialization, communication, and repetitive and restricted interests and behavior. A consensus diagnosis was determined for each participant based on results from the ADOS-2, ADI-R and clinical evaluation using DSM-5.

Intellectual functioning. Global cognitive ability was measured using the Mullen Scales of Early Learning [13], the Stanford Binet Intelligence Scales, Fifth Edition [14], or the Differential Ability Scales, Second Edition (DAS-II) [15], depending on age and verbal ability. The Mullen is validated for children from birth to 68 months, but is commonly used for older individuals with ID [16]. Developmental quotients were calculated using age equivalents divided by chronological age as has been done in previous studies [17]. For example, a nonverbal developmental quotient was computed by dividing the mean age equivalents on the visual reception and fine motor scales by the child's chronological age and then multiplying by 100. The DAS-II is a measure of cognitive functioning that assesses a child's verbal reasoning, nonverbal reasoning, and spatial abilities. A general conceptual ability index can be calculated to assess overall intellectual functioning. The Stanford-Binet Intelligence Scales, Fifth Edition is an intelligence test that produces a nonverbal intellectual quotient (IQ), verbal IQ, and full scale IQ based on performance across five scales: fluid reasoning, knowledge, quantitative reasoning, visual-spatial, and working memory.

Adaptive behavior. The Vineland Adaptive Behavior Scales, Second Edition, Survey Interview Form (Vineland-II) [18] and the Vineland Adaptive Behavior Scales, Third Edition, Comprehensive Interview Form (Vineland-3) are clinician-administered interviews that assesses adaptive behavior in the domains of communication, daily living skills, socialization, and motor skills. The Vineland-II was completed for 13 individuals. The Vineland-3 was completed for 19 individuals. The motor domain is intended for children ages six years and under, but was

assessed in 19 individuals given significant motor delays in this population. The Vineland-II and Vineland-3 were also used in conjunction with cognitive testing to identify the presence and severity of ID.

Language skills. Language milestones were assessed during the ADI-R and the psychiatric evaluation. Current expressive and receptive language abilities were assessed using the Mullen, Vineland-II, MacArthur-Bates Communicative Development Inventories [19], Peabody Picture Vocabulary Test, Fourth Edition [20], and Expressive Vocabulary Test [21] as appropriate.

Motor skills. Motor milestones were assessed during the ADI-R and the psychiatric evaluation. Current motor skills were assessed using the Vineland-II and Mullen fine and gross motor skills domains. The Beery Visual-Motor Integration Test, 6th Edition [22] was completed when appropriate.

Caregiver questionnaires. To further assess everyday behavior, caregiver questionnaires were completed in the domains of sensory processing (Sensory Profile [23]), aberrant behavior (Aberrant Behavior Checklist [24]), repetitive behavior (Repetitive Behavior Scale-Revised [25]), socialization (Social Responsiveness Scale, 2nd Edition [26]), and coordination (Developmental Coordination Disorder Questionnaire [27]).

Genome-wide methylation arrays and data quality control

For the methylation analyses, genomic DNA was isolated from the peripheral blood of 24 *ADNP* cases, 19 unaffected age-matched, and 14 unaffected siblings from cohorts S, A and R. DNA methylation analysis was performed using the Ilumina EPIC 850K methylation arrays, according to the manufacturer's protocol. Whole-blood genomic DNA samples for methylation analyses were prepared at each collecting site of the four cohorts described above and then further harmonized at the Seaver Autism Center. The methylation assays were performed in three batches, each including cases from multiple cohorts and controls. Column P of Table S1 reports the batch number for each sample. Batches 1 and 2 were run at the New York Genome Center, while batch 3 was run at NXT-Dx. In each batch, we included a number of controls present in previous batches as internal controls. All the methylation data in this paper are available at the Gene Expression Omnibus under accession number GSE152428. To ensure that batch effects have minimal influence on the current analysis, we performed a standardized multistep analytic approach (Figure S1). First, we computed genome-wide pairwise Pearson correlation coefficients across all possible pairs of samples and confirm that the highest correlations are amongst technical replicates included in separate batches ($R > 0.98$) (Figure S1A). Second, we evaluated whether any large shifts in the distribution of normalized β-values were observed across batches, and found no significant differences in the distributions of β-values across batches (Figure S1B). Third, we applied a mixed linear model from the variancePartition R package [28] to compute the fraction of genome-wide methylation variance explained by differences in batch, mutational class, family status, sex and age (Figure S1C). Collectively, these factors explained ~35% of methylation variation, with family as a repeated measure having the largest genome-wide effect that explained a median 25.7% of the observed variation. The remaining factors, including batch, had a limited effect on methylation variance. Fourth, we analyzed all batches together for the joint analysis presented in this study and included batch as a covariate in our regression model. By properly covarying for known sources of methylation variation, it is possible to (partially) correct for some variables. Finally, following differential methylation analyses, we once again applied a mixed linear model to quantify variance explained using only differentially methylated probes that were found to be significantly associated with *ADNP* mutational classes. As expected, the differences in mutational classes had the largest effect and explained a median 65.35% of the variance, whereas batch had a minor contribution to the variance on these probes (< 0.001 median variance) (Fig. S1D).

The sex of the individuals included in the study was inferred both by the number of probes on chromosome Y (chrY) with detection p>0.01, and the mean β-value of probes on chromosome X (chrX) per sample. Samples with high failure rate for chrY probes and high mean β-value for chrX probes were inferred as females, while samples with low failure rate for chrY probes and low mean β-value for chrX probes were inferred as males.

Epigenetic age (DNAm age) was predicted using the online tool published by Horvath, 2013 [29]. The chronological and DNAm age for cases and controls are as follows:

Group	Chronological age (yrs)	Age inferred from methylation data (DNAm age) (yrs)
ADNP class I cases	8.03 ± 0.94	7.10 ± 1.24
ADNP class II cases	11.85 ± 3.50	10.57 ± 3.44
Unaffected siblings	12.18 ± 1.51	10.73 ± 1.35
Unaffected controls	8.36 ± 0.91	7.87 ± 0.97

The fraction of circulating peripheral blood cell types were predicted for each sample using the method of Houseman et al. 2012 [30]. Predictions for the following cell types were obtained: CD4+ T cells, CD8+ T cells, natural killer cells, B lymphocytes, monocytes and granulocytes.

Data from *ADNP* cases and controls were normalized as described in previous studies [31; 32]. In short, 862,927 probe sequences (50-mer oligonucleotides) were remapped to the reference human genome hg19 (NCBI37) using BSMAP, allowing up to 2 mismatches and 3 gaps, to retain uniquely mapped autosomal probes. We removed any probe that overlapped SNPs with MAF ≥5% identified by the 1000 Genomes Project within 5 bp upstream of the targeted CpG. Further, we removed probes with a detection p>0.01 in each individual. After filtering, we retained 820,167 autosomal probes, which were subjected to background correction, two color channel normalization and quantile normalization using the *lumi* package in R [33]. The distributions of Infinium I and Infinium II probes were adjusted using *BMIQ* [34]*.* Probes were then annotated based on their position relative to RefSeq genes using BEDTools v2.17 [35]. We defined promoter regions as ±2 kb from transcriptional start sites, gene body regions as transcription start to transcription end, and intergenic regions not annotated by the preceding categories. We also annotated individual CpG based on their overlap with CpG islands based on annotations in the UCSC genome browser, CpG shore (±2 kb of island), CpG shelf (±2 kb of shore), and CpG sea (regions outside the previous three categories).

Identification of an episignature in *ADNP* cases

We performed linear regression using age, sex, predicted blood cell composition and *ADNP* mutation status as independent variables (Test Model: Methylation ~ Disease status + Age + Sex + CD4T + natural killer cells + B cells + Monocytes + Granulocytes + Batch). Regression analysis was completed separately for class I and class II *ADNP* mutations. We did not include CD8T cell composition in the model due to very low abundance across all samples. We selected probes associated with disease status at 1% FDR and with minimum β-value difference between *ADNP* cases and controls ≥0.1. Principal component analysis and unsupervised clustering of episignatures was performed on methylation data following Combat batch adjustment [36] to remove systematic sources of variability related to batches without introducing false signal. The differentially methylated probes are listed in Table S2. To identify differentially methylated regions, we have selected differentially methylated regions with three or more probes within 1 kb of each other. These differentially methylated regions are also listed in Table S2.

Comparison of episignatures found here and in Bend et al (2019).

We extracted the 5,987 and 1,374 differentially methylated CpGs (class I and class II, respectively) in Bend et al (2019) [1] and crossed with our lists of differentially methylated CpGs. Of the 6,448 sites we detected as differentially methylated in *ADNP* cases in class I, 4,143 were also found differentially methylated in Bend et al (2019) [1]. Of the 2,582 sites we detected as differentially methylated in *ADNP* cases in class II, 1,007 were also found differentially methylated in Bend et al (2019) [1].

Also, the differentially methylated probes overlapping between our study and Bend et al (2019) [1] completely agree in terms of directionality of change. Of the 4,143 class I probes in common between the two analyses,

3,974 were hypomethylated in both studies and the remaining were hypermethylated in both studies. Of the 1,007 overlapping probes for class II, 771 were found hypermethylated in both studies and the remaining were hypomethylated in both studies.

Further, we examined the number of differentially methylated probes that lie in the gene promoter (± 2 kb from the transcriptional start sites) or in the gene body (transcription start to transcription end) across the two studies. The stratification is as follows:

Mutational class	Genomic location	Bend et al (2019)	This study	Overlapping probes
Class I	Promoter	1,662 (27.7%)	1,839 (28.5%)	1,144 (27.6%)
	Gene body	2,953 (49.3%)	3,282 (50.9%)	2,089 (50.4%)
Class II	Promoter	449 (32.7%)	897 (34.7%)	322 (31.9%)
	Gene body	856 (62.3%)	1,558 (60.3%)	647 (64.3%)

RNA isolation, library preparation, and quantification of gene expression

Blood was collected for 17 *ADNP* cases and 19 unaffected siblings using PAXgene RNA tubes (Qiagen, Valencia, CA, USA) and total RNA was extracted and purified in accordance with the PAX gene RNA kit per manufacturer's instructions. Globin mRNA was depleted from samples using the GLOBINclear Human Kit (Life Technologies, Carlsbad, CA, USA). The quantity of purified RNA was measured on a Nanodrop 2000 Spectrophotomerter (Thermo Scientific; 61.4 ± 24.1 ng μl^{-1}) and RNA integrity numbers (RIN) measured with the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA; 8.31 ± 0.68). The Illumina TruSeq Total RNA kit (Illumina, San Diego, CA, USA) was used for library preparation accordingly to manufacturer instructions without any modifications. The 36 indexed RNA libraries were pooled and sequenced using long paired-end chemistry (2x150 bp) at an average read depth of 10M reads per sample using the Illumina HiSeq2500. Illumina adapter sequences were trimmed from all fragmented reads using TrimGalore (options –paired –illumina) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). All high-quality trimmed reads were mapped to UCSC *Homo sapiens* reference genome (build hg37) using default STAR v2.4.0 parameters [37]. Samtools was used to convert bamfiles to samfiles and featureCounts [38] was used to quantify gene expression levels for each individual sample using default paired-end parameters.

RNA-seq data quality control

Raw count data measured 56,632 genes across 36 subjects. Unspecific filtering removed lowly expressed genes that did not meet the requirement of a minimum of 1 count per million (cpm) in at least 8 subjects (40% of subjects). A total of 20,491 genes were retained, then subjected to edgeR VOOM normalization [39], a variance-stabilization transformation method. Normalized data were inspected for outlying samples using unsupervised hierarchical clustering of subjects (based on Pearson coefficient and average distance metric) and principal component analysis to identify potential outliers outside two standard deviations from these averages. No outliers were present in these data.

Mutant allele abundance

We took measures to reduce the likelihood of false positives and biases in quantification of relative allele abundances at each mutation location. RNA-seq reads were filtered and mapped to retain only uniquely mapped reads, using STAR [35] and SAMtools, and then further removed potential PCR duplicates using the MarkDuplicates function in Picard tools with default parameters (http://broadinstitute.github.io/picard/faq.html). Allelic counts were computed at each single-nucleotide variant position in each sample using the SAMtools mpileup function. Next, we recorded the number of overlapping sequences containing the mutant and reference allele. The mutant allele frequency was defined as the number of covering RNA-seq reads

containing the mutant allele at that position, divided by the total number of RNA-seq reads overlapping that position.

Differential gene expression and concordance with methylation changes

A moderated *t*-test, implemented through the *limma* package [39], assessed differential gene expression between unaffected siblings and *ADNP* cases with class I and class II mutations, respectively. The analysis covaried for the possible influence of sex, gender, RIN and sequencing batch on gene expression differences. Significance threshold was set to a Benjamini-Hochberg multiple test corrected *P*-value <0.05. Correlation adjusted mean rank (CAMERA) gene set enrichment [40] was performed using the two sets of resulting summary statistics for *ADNP* cases with class I and class II mutations. CAMERA performs a competitive gene set rank test to assess whether the genes in a given set are highly ranked in terms of differential expression relative to genes that are not in the set. The test ranks gene expression differences in *ADNP* cases with class I and class II mutations relative to unaffected siblings, respectively, to test whether gene sets are over-represented towards the extreme ends of these ranked lists. It uses *limma*'s linear model framework and accommodates the observational-level weights from *voom* in the testing procedure. After adjusting the variance of the resulting gene set test statistic by a variance inflation factor that depends on the gene-wise correlation (which we set to default parameters, 0.01) and the size of the set, a *P*-value is returned and adjusted for multiple testing. We used this function to test whether *ADNP*-associated changes in gene expression were indeed enriched for genes with significant differentially methylated CpGs. We specifically focused on the enrichment of *i)* hyper-methylated CpGs among the reported under-expressed genes and *ii)* hypo-methylated CpGs among the reported over-expressed genes in *ADNP* cases. In this fashion, we can further gauge the distribution of differentially methylated CpGs across a ranked list of differentially expressed genes.

Enrichment analyses

We assessed enrichment of differentially methylated CpG and differentially expressed genes using five gene lists: *a)* 914 genes implicated in developmental delay/intellectual disability (DD/ID), based on the Developmental Disorders Genotype-Phenotype Database (DDG2P) (https://decipher.sanger.ac.uk/info/ddg2p); *b)* 102 ASD risk genes defined by the Autism Sequencing Consortium [3]; *c)* 145 risk genes for CHD built in house; *d)* Gene ontology (GO) molecular functions and cellular processes; and *e)* REACTOME pathway gene sets. For DNA methylation analyses, we utilized a background gene set comprising 25,141 genes that overlapped 606,904 probes covered by the Illumina 850K EPIC methylation arrays. For transcriptome analyses, we utilized a background gene set of 20,491 genes that were defined as 'expressed' in the current data set (described above). For each comparison between the differentially methylated and differentially expressed genes in the input set, we first constructed the empirical distribution by randomly sampling the same number of genes as in the input set from the background gene set 10,000 times, using a custom script in R. Enrichment *P*-values were computed by calculating the number of sampled gene lists that had at least as many overlapping genes with the target sets as the input set, divided by 10,000 permutations.

Data availability

All methylation and RNA-sequencing raw data files are made available at the Gene Expression Omnibus (GEO) under SuperSeries accession number GSE152428.

RT-PCR and Sanger sequencing

Reverse transcription was performed on DNAseI-treated 50ng of RNA using High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, California, USA) according to the manufacturer's protocol. PCR was performed using FastStart PCR Master (Roche, Mannheim, Germany) or Phusion High-Fidelity PCR Kit (New England BioLabs, Ipswich, Massachusetts), using primers indicated below:

Sample	Forward primer	Reverse primer
5S	ACGAAAAACCAGGACTATCGGA	AAACAGCTTGCTCTACACTGTCA
12S	ATCGGTTCCCTTGCTTCTGG	TGGCCCGATGAGAGAGAAGA
14S	TGCAGCAGAACAACTATGGAGT	CTGCAGCAGGTTTGGAACTG
18S	ATACCAGCAACATGACCGCC	TGGTGGGATAGGGCTGTTTG
33S	CACCCTCTCGGCTTAATCAGT	TAAACTGGCTGCTAGCTTCTCAA

PCR products were purified using a MultiScreen PCR filter plate (Millipore Sigma, Burlington, Massachusetts) and submitted for Sanger Sequencing at Genewiz with the primers indicated above.

Supplemental references

1. Bend, E.G., Aref-Eshghi, E., Everman, D.B., Rogers, R.C., Cathey, S.S., Prijoles, E.J., Lyons, M.J., Davis, H., Clarkson, K., Gripp, K.W., et al. (2019). Gene domain-specific DNA methylation episignatures highlight distinct molecular entities of ADNP syndrome. Clin Epigenetics 11, 64.
2. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515, 209-215.
3. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell 180, 568-584 e523.
4. Pascolini, G., Agolini, E., Majore, S., Novelli, A., Grammatico, P., and Digilio, M.C. (2018). Helsmoortel-Van der Aa Syndrome as emerging clinical diagnosis in intellectually disabled children with autistic traits and ocular involvement. Eur J Paediatr Neurol 22, 552-557.
5. Gouw, M., Michael, S., Samano-Sanchez, H., Kumar, M., Zeke, A., Lang, B., Bely, B., Chemes, L.B., Davey, N.E., Deng, Z., et al. (2018). The eukaryotic linear motif resource - 2018 update. Nucleic Acids Res 46, D428-D434.
6. Zamostiano, R., Pinhasov, A., Gelber, E., Steingart, R.A., Seroussi, E., Giladi, E., Bassan, M., Wollman, Y., Eyre, H.J., Mulley, J.C., et al. (2001). Cloning and characterization of the human activity-dependent neuroprotective protein. J Biol Chem 276, 708-714.
7. Cappuyns, E., Huyghebaert, J., Vandeweyer, G., and Kooy, R.F. (2018). Mutations in ADNP affect expression and subcellular localization of the protein. Cell Cycle 17, 1068-1075.
8. Mosch, K., Franz, H., Soeroes, S., Singh, P.B., and Fischle, W. (2011). HP1 recruits activity-dependent neuroprotective protein to H3K9me3 marked pericentromeric heterochromatin for silencing of major satellite repeats. PLoS One 6, e15894.
9. Lord, C., Rutter, M., DiLavore, P.S., Risi, S., Gotham, K., and Bishop, D. (2012). Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2) Manual (Part I): Modules 1-4.(Torrance, CA: Western Psychological Services).
10. Lord, C., Rutter, M., and Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord 24, 659-685.
11. American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders, 5th edition, text revision.(Washington, DC: American Psychiatric Association).
12. Hus, V., Gotham, K., and Lord, C. (2014). Standardizing ADOS domain scores: separating severity of social affect and restricted and repetitive behaviors. J Autism Dev Disord 44, 2400-2412.
13. Mullen, E.M. (1995). Mullen Scales of Early Learning.(Circle Pines, MN: American Guidance Services).
14. Roid, G.H. (2003). Stanford Binet Intelligence Scales (5th edition).(Itasca, IL: Riverside Publishing).
15. Elliot, C.D. (2007). Differential Ability Scales—Second edition: Introductory and technical manual.(San Antonio, TX: Harcourt Assessment).
16. Bishop, S.L., Guthrie, W., Coffing, M., and Lord, C. (2011). Convergent validity of the Mullen Scales of Early Learning and the differential ability scales in children with autism spectrum disorders. American journal on intellectual and developmental disabilities 116, 331-343.
17. Akshoomoff, N. (2006). Use of the Mullen Scales of Early Learning for the assessment of young children with Autism Spectrum Disorders. Child Neuropsychol 12, 269-277.
18. Sparrow, S.S., Cicchetti, D.V., and Balla, D.A. (2005). Vineland Adaptive Behavior Scales: Second edition (Vineland II), Survey Interview Form/Caregiver Rating Form.(Livonia, MI: Pearson Assessments).
19. Fenson, L., Marchman, V.A., Thal, D.J., Dale, P.S., Reznick, J.S., and Bates, E. (2007). MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual, Second Edition.(Baltimore, MD: Brookes Publishing Co.).
20. Dunn, L.M., and Dunn, D.M. (2007). PPVT-4: Peabody picture vocabulary test.(Bloomington, MN: Pearson Assessments).
21. Williams, K.T. (2007). The Expressive Vocabulary Test (2nd edition).(Circle Pines, MN: AGS Publishing).
22. Beery, K.E., Buktenica, N.A., and Beery, N.A. (2010). Beery-Buktenica Developmental Test of Visual-Motor Integration, Sixth Edition.(Minneapolis, MN: Pearson).

23. Dunn, W., and Westman, K. (1997). The Sensory Profile: The performance of a national sample of children without disabilities. American Journal of Occupational Therapy 51, 25-34.
24. Aman, M.G., and Singh, N.N. (1986). Aberrant Behavior Checklist (ABC) Manual.(East Aurora: Slosson Educational Publications, Inc).
25. Lam, K.S., and Aman, M.G. (2007). The Repetitive Behavior Scale-Revised: independent validation in individuals with autism spectrum disorders. Journal of autism and developmental disorders 37, 855-866.
26. Constantino, J.N., and Gruber, C.P. (2012). Social Responsiveness Scale, Second Edition (SRS-2).(Torrance: Western Psychological Services).
27. Wilson, B.N., Dewey, D., and Campbell, A. (1998). Developmental coordination disorder questionnaire (DCDQ).(Canada: Alberta Children's Hospital Research Center).
28. Hoffman, G.E., and Schadt, E.E. (2016). variancePartition: interpreting drivers of variation in complex gene expression studies. BMC bioinformatics 17, 483.
29. Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol 14, R115.
30. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., and Kelsey, K.T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics 13, 86.
31. 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. Nature communications 9, 2064.
32. Garg, P., and Sharp, A.J. (2019). Screening for rare epigenetic variations in autism and schizophrenia. Hum Mutat.
33. Du, P., Kibbe, W.A., and Lin, S.M. (2008). lumi: a pipeline for processing Illumina microarray. Bioinformatics 24, 1547-1548.
34. Teschendorff, A.E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 29, 189-196.
35. Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics 47, 11 12 11-34.
36. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28, 882-883.
37. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.
38. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923-930.
39. 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 research 43, e47.
40. Wu, D., and Smyth, G.K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic acids research 40, e133.