PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/191908

Please be advised that this information was generated on 2020-03-05 and may be subject to change.
Identification of rare de novo epigenetic variations in congenital disorders

Mafalda Barbosa1,2,3, Ricky S. Joshi1, Paras Garg1, Alejandro Martin-Trujillo1, Nihir Patel1, Bharati Jadhav1, Corey T. Watson1, William Gibson1, Kelsey Chetnik4, Chloe Tessereau1, Hui Mei5,14, Silvia De Rubeis3,6, Jennifer Reichert3,6, Fatima Lopes7, Lisenga E.L.M. Vissers8, Tjitske Kleefstra8, Dorothy E. Grice9,10, Lisa Edelmann5, Gabriela Soares11, Patricia Maciel7, Han G. Brunner8,12, Joseph D. Buxbaum1,3,6,10, Bruce D. Gelb1,13 & Andrew J. Sharp1,2

Certain human traits such as neurodevelopmental disorders (NDs) and congenital anomalies (CAs) are believed to be primarily genetic in origin. However, even after whole-genome sequencing (WGS), a substantial fraction of such disorders remain unexplained. We hypothesize that some cases of ND–CA are caused by aberrant DNA methylation leading to dysregulated genome function. Comparing DNA methylation profiles from 489 individuals with ND–CAs against 1534 controls, we identify epivariations as a frequent occurrence in the human genome. De novo epivariations are significantly enriched in cases, while RNAseq analysis shows that epivariations often have an impact on gene expression comparable to loss-of-function mutations. Additionally, we detect and replicate an enrichment of rare sequence mutations overlapping CTCF binding sites close to epivariations, providing a rationale for interpreting non-coding variation. We propose that epivariations contribute to the pathogenesis of some patients with unexplained ND–CAs, and as such likely have diagnostic relevance.
Epi

mutations represent a class of mutational event where the epigenetic status of a genomic locus deviates significantly from the normal state, and can be classified into two main types: primary epimutations are thought to represent stochastic errors in the establishment or maintenance of an epigenetic state, while secondary epimutations are downstream events related to an underlying change in the DNA sequence. Both secondary and primary epimutations that originate in the germline will be constitutive events found in all cells. In contrast, primary epimutations that occur post-fertilization may result in somatic mosaicism. Constitutive (i.e., non-mosaic) epimutations are known to underlie several genetic disorders that can be identified in blood-derived DNA: 5–15% of patients with hereditary nonpolyposis colon cancer present with constitutional MLHI promoter methylation, and fragile X syndrome, the most common cause of inherited intellectual disability, results from a secondary epimutation in which hypermethylation of an expanded CGG repeat at the FMR1 promoter causes transcriptional silencing.

With the recent dramatic advances in genomic technologies, genome-wide surveys of cohorts of patients with neurodevelopmental disorders (NDs) and congenital anomalies (CAs) (ND–CAs) for point mutations and structural variations have greatly advanced our understanding of their genetic etiologies. However, even after whole genome sequencing (WGS), no causative mutation can be identified in many such cases. We hypothesized that some cases of ND–CA that remain refractory to conventional sequence-based analysis harbor rare epigenetic aberrations (termed epivariations), which are associated with dysregulation of normal genome function, and that these would be missed by the conventional sequencing approaches. We identify rare epigenetic changes that are absent in thousands of controls in ~20% of patients with ND–CA. From large-scale sequencing, population and expression studies, we conclude that epivariations are: (i) frequently associated with extreme outlier and mono-allelic gene expression; (ii) generally conserved across multiple tissues within an individual, validating the use of blood DNA to study ND–CA; (iii) sometimes occur secondary to cis-linked regulatory mutations, providing a rationale for interpreting non-coding genetic variants; (iv) can occur sporadically with a remarkably high de novo rate, and (v) a subset exhibit non-Mendelian inheritance, suggesting they are often being reset between generations by epigenetic reprogramming. We propose that epivariations likely contribute to the pathogenesis of some patients with unexplained ND–CAs, and suggest that epigenome profiling represents a promising method for the study of human disease that complements sequence-based approaches.

Results

Identification of epivariations in cases and controls. We studied a cohort comprising 489 individuals with ND–CA: most had been previously tested by copy number variation (CNV) microarray, all had undergone exome sequencing, and some had undergone WGS, yet no putatively pathogenic mutations had been identified. Almost 90% of the patients had an ND, 50% were classified as having an autism spectrum disorder, 16% had an epilepsy/seizure phenotype; 65% also had multiple CA, including congenital heart defects (CHD) (36%), facial dysmorphism (29%), growth anomalies (22%), and micro/macrocephaly (13%) (full details in Supplementary Data 1). We hypothesized that this cohort represented an optimal population in which to search for novel pathogenic epivariations since an underlying genomic abnormality was suspected, but many common environmental and genetic causes of ND–CA had been excluded. Methylation profiling in ND–CA samples was performed with the Illumina Infinium Human Methylation 450 BeadChip (450k array). Profiles in each ND–CA sample were compared individually against a control cohort comprising 1534 unrelated individuals from four publicly available datasets (GSE36064, GSE40279, GSE42861, and GSE53045). We also searched for epivariations in two cohorts of population controls by comparison against this same set of 1534 individuals: 117 families (GSE56105) were used to assess the inheritance of epivariations in controls (Supplementary Data 2); 2711 unrelated individuals (GSE55763) were used to assess the frequency of epivariations in the general population (Supplementary Data 3). We utilized a sliding window approach to identify epivariations in each sample, defined as 1 kb regions containing ≥3 probes showing rare outlier methylation absent in the set of 1534 common control individuals (see Supplementary Fig. 1 and Methods section). After stringent quality control, including removal of loci with clusters of poorly hybridizing probes and extensive manual curation to remove technical and batch effects, we identified a total of 143 epivariations in 114 ND–CA samples (i.e., 23% of the probands tested). Twenty percent of the ND–CA cohort carried one epivariation (n = 98), while 3% of the individuals tested presented two or more epivariations (n = 16) (Supplementary Data 4 and Supplementary Fig. 2).

Using PCR/bisulfite sequencing, we attempted orthogonal confirmation for 70 epivariations. We observed concordant changes in methylation for 55 of the 58 assays that provided useful data, yielding a 95% true positive rate for differentially methylated regions (DMRs) detected by array (Supplementary Data 5). Allelic analysis demonstrated that these epivariations represent large methylation changes specifically on one allele, consistent with the hypothesis that epivariations represent allelic events. In most cases, we observed two clusters of largely methylated and unmethylated reads occurring in approximately equal proportions (Fig. 1), although in some instances the interpretation of validation experiments was made complex due to highly biased allelic representation, presumably reflecting preferential PCR amplification of one allele (Supplementary Fig. 3). In addition to searching for epivariations in samples with ND–CA, we also screened two large cohorts of population controls, identifying a total of 719 DMRs in the 3326 control samples analyzed (Supplementary Data 2 and 3). Thus, epivariations are a relatively common occurrence in the human genome, and are not always associated with any discernable clinical phenotype. Twenty-four of the epivariations identified in our cases with ND–CA were also found in one or more of these controls, therefore indicating that either these DMRs are unrelated to the patient phenotype, or perhaps are associated with incomplete penetrance. However, we observed a 1.2-fold enrichment in the frequency of epivariations in the 489 ND–CA samples when compared to 2711 population controls (Supplementary Fig. 2), although this does not reach statistical significance (p = 0.058, two-sided Fisher’s exact test).

Using a combination of 450k arrays and bisulfite PCR/sequencing assays, we were able to assess the inheritance of 57 DMRs identified in our patients with ND–CA: 33 of the 57 epivariations tested were also present in apparently unaffected parents, and thus represent inherited events. However, 42% (n = 24) of the epivariations identified in patients with ND–CA were absent in both parental samples, and thus occurred as de novo events. When compared to epivariations found in 117 control pedigrees (Supplementary Data 2), this represents a 2.8-fold enrichment in the rate of de novo epivariations in cases compared to controls (p = 0.007, two-sided Fisher’s exact test) (Fig. 2). Thus, while it is currently unclear whether many of the epivariations identified contribute to the phenotypes of the patients in our study, the paradigm of de novo mutational events...
echoes that observed for other classes of genetic mutation (copy number and single nucleotide variation (SNV)) deemed pathogenic in ND and CHD cohorts.\textsuperscript{9, 10}

In addition to their de novo nature, recurrence of mutations found in unrelated patients with a similar phenotype is commonly used as a way of assigning significant evidence for the involvement of a specific gene or locus in disease. We identified 12 recurrent epivariations (Supplementary Fig. 4), i.e., the same methylation change was identified in multiple unrelated probands. Of these, two epivariations encompassed the promoters of genes known to show altered methylation in congenital disease (MEG3 and FMR1)\textsuperscript{6, 11}, showing that our approach successfully detects pathogenic epivariations. The two males identified with hypermethylation at FMR1 had phenotypes consistent with a diagnosis of fragile X, primarily intellectual disability (ID) and behavioral anomalies. While both had previously been tested by PCR and reported as normal, subsequent Southern blot testing confirmed the presence of the classical FMR1 triplet repeat expansion, although in one case this was an apparent mosaic event. A third recurrent epivariation coincides with a fragile site containing a hypermethylated triplet repeat expansion (\textit{FRA10AC1})\textsuperscript{12} although this, and four other recurrent epivariations detected in our disease cohort, was also identified in population controls, suggesting that they are unlikely to be pathogenic. One of the novel recurrent epivariations detected only in our patient cohort was found in two patients with CHD (Probands 22 and 117), and represents a recurrent hypomethylation defect at the promoter—5' UTR—first exon of MOVIOLI, a gene with an embryonic heart-specific isoform that interacts with the master cardiac transcription factor NKX2.5\textsuperscript{13} (Fig. 1). One patient with this epivariation at MOVIOLI presented with double outlet right ventricle, hypoplastic left ventricle, asplenia, and short stature, while the second presented with pulmonary stenosis, laryngo-bronchio-tracheomalacia, and foot polydactyly. Finally, using less stringent criteria for identifying DMRs (see Methods), we detected methylation defects in 11 probands at 10 imprinted loci (Supplementary Fig. 5 and Supplementary Table 1), 90% of which occurred de novo. Of note, we observed loss of methylation at two known imprinted loci that have no prior disease associations (\textit{NAA60/ZNF597} in Probands 6 and 62, and \textit{L3MBTL1} in Proband 308), although in both cases similar losses of methylation were also observed in population controls, making the pathogenic significance of loss of imprinting at these loci uncertain.

Regulatory mutations underlie some epivariations. Based on previous studies\textsuperscript{15, 16}, we hypothesized that some epivariations might occur secondarily to an underlying regulatory sequence mutation. In order to identify mutations disrupting regulatory elements (e.g., transcription factor binding sites) that might underlie the methylation changes observed in our cohort, we performed high-resolution array comparative genomic hybridization (CGH) and targeted DNA sequencing of 50 DMRs and their flanking sequences. We detected rare sequence mutations that

![Fig. 1](image-url)
that co-segregated with epivariations and potentially impact regulatory elements at 24% of the loci tested: six CNVs (Fig. 3 and Supplementary Fig. 6) and seven SNVs (Supplementary Data 4, 6, and 7). Where inheritance data from parental samples were available, we found that all of these rare CNVs and SNVs segregated with the presence of the DMR, suggesting that the epivariations occurred secondarily to the underlying sequence mutation.

Of the rare segregating SNVs detected at DMRs, three were SNVs within the canonical binding sites for CTCF (CCCTC-binding factor), a transcription factor with roles in chromatin organization (Fig. 4), including a de novo SNV that disrupts a CTCF binding motif in association with a de novo epivariation (Proband 70) (Supplementary Fig. 7). In each case, the disrupted CTCF motif was either overlapping or very close to (separation <1 kb) the DMR. This represents a significant enrichment for rare SNVs disrupting CTCF binding sites in the vicinity of epivariations when compared to the same regions in other samples in whom we performed targeted sequencing, but who did not carry epivariations at these loci (p = 0.0015, two-sided Fisher’s exact test), strongly implicating rare cis-linked variants in regulatory sequence as a causative factor underlying some epivariations. Furthermore, given the low frequency of de novo SNVs and epivariations in the genome, it is highly unlikely that a de novo SNV and a de novo epivariation would co-occur at the same locus in an individual by chance, providing additional support that some epivariations represent secondary causes evoked by disruption of CTCF binding. Using paired methylation and sequence data from 90 individuals studied by the 1000 Genomes Project (Supplementary Data 8), we replicated this enrichment for rare SNVs disrupting CTCF binding motifs around epivariations (p = 0.049, two-sided Fisher’s exact test), identifying two rare CTCF-disrupting SNVs, one of which co-segregates with the presence of an epivariation in multiple unrelated individuals. Though readily detectable by WGS, there is considerable difficulty in interpreting the functional significance of variants outside of coding regions. Thus, we propose that the use of epigenome profiling represents a complementary approach that can provide a rationale for interpreting non-coding genetic variation.

**Functional consequences of epivariations.** In order to provide insight into the biology and functional consequences of epivariations, we performed studies of gene expression, inheritance, and tissue conservation using datasets of DNA methylation (Supplementary Data 9), gene expression (Supplementary Data 10), and genotype data derived from population controls. Using paired RNAsseq and DNA methylation data in 90 samples from the 1000 Genomes Project, we verified that epivariations encompassing gene promoters were often associated with large changes in gene expression, with hypomethylation leading to increased expression and hypermethylation to transcriptional repression, consistent with the known repressive effects of promoter DNA methylation (p = 9.2 × 10^-5, Wilcoxon Rank-Sum test) (Fig. 5, Supplementary Data 10). We also observed that many hypermethylated epivariations at promoters are associated with complete silencing of one allele (Supplementary Fig. 8). While these observations were made in a control cohort, this suggests that some epivariations have an impact comparable to that of loss-of-function coding mutations.

**Epivariations are generally present in multiple tissues.** While epigenetic profiles can vary substantially between cell types, it is unclear whether similar cell-specific variability exists for epivariations. To address this, we analyzed cohorts in which methylation profiles were available from multiple different tissues. In samples from the GenCord population, in which methylation data from fibroblasts, B cells, and T cells sampled from dozens of newborns are available, by first identifying DMRs in T cells, we observed a very strong concordance for outlier methylation at the same locus in fibroblasts derived from the same individual (Spearman rank correlation of 0.75, p = 1.2 × 10^-27, Wilcoxon Rank-Sum test) (Fig. 6, Supplementary Data 9). Similar concordance for outlier methylation at epivariations was also observed between fibroblasts and B cells.

A similar trend for conservation of epivariations across multiple different post-mortem tissues was also observed in a second cohort. Here, epivariations found in blood were nearly all visible in multiple other somatic tissues sampled from the same individual (Supplementary Fig. 9). Thus, we conclude that the majority of epivariations are constitutive events found in multiple tissues. This provides confidence that epivariations of relevance for ND-CA can be detected using DNA extracted from readily available sources such as peripheral blood leukocytes.

**Evidence for non-Mendelian inheritance of epivariations.** Despite strong evidence that some of the epivariations we observed are secondary events related to the presence of an underlying sequence change (Figs. 3 and 4), we were unable to detect cis-linked sequence mutations associated with the majority of epivariations in our cohort, suggesting that these might instead represent primary epivariations that arose sporadically. As the mammalian genome undergoes several rounds of demethylation and remethylation during gametogenesis, embryonic and somatic development, theoretically there is considerable potential for primary epivariations to be reset to the default state. We therefore assessed how often epivariations are stably transmitted between parents and their offspring. Using a large control cohort comprising 117 nuclear families, we studied the heritability of epivariations between generations, identifying 47 epivariations segregating within these pedigrees. We observed a marked deviation from the expectations of Mendelian inheritance, with only 32 instances of parent–child transmission in 95 informative meioses; significantly fewer than the Mendelian expectation of 47.5 transmissions (p = 0.027, two-sided Fisher’s exact test) (Supplementary Data 2). Therefore, this apparent reduction in heritability indicates that primary epivariations often exhibit non-Mendelian inheritance, and suggests they are frequently reset between generations by epigenetic reprogramming.
Discussion

In this study, we set out to investigate the prevalence, causes, and consequences of epigenetic defects in the human genome, and to study their potential role in the etiology of ND–CA disorders. By performing epigenome profiling in a large cohort of 489 patients with diverse ND–CA, all of whom had previously undergone microarray testing and/or exome or genome sequencing, in addition to analyzing >5000 population controls, we demonstrate that epivariations are a relatively frequent occurrence in the human population. Subsequent analysis of cell lines showed that the presence of epivariations is often associated with large changes in the expression of cis-linked genes, indicating functional consequences on the genome. Furthermore, we demonstrated that epivariations are generally conserved across multiple tissues, validating the use of peripheral blood for the studies of ND–CA.

We observed both recurrent epivariations in cases that were absent in thousands of population controls, and a significant excess of de novo epivariations in cases compared to controls, both of which are hallmarks often associated with pathogenic variants. Despite this, the pathogenic significance of many of the epivariations we identified remains uncertain. For example, in many cases, epivariations were inherited from apparently unaffected parents, suggesting that they are unlinked to the observed phenotype. While it is likely that the presence of some epivariations we observed in ND–CA cases is unrelated to patient phenotype, we note that not all inherited events are benign, and there are many examples of rare inherited sequence variants that show variable penetrance.

Studies of the inheritance of epivariations in families showed that they can occur de novo at very high frequency (up to 42%), yet also show significantly reduced heritability compared to Mendelian expectations, suggesting that they are often reset during meiosis. However, in contrast to this dynamic process of frequent gain and loss, targeted sequencing and array CGH identified segregating rare sequence variants that disrupt
Fig. 4 Targeted sequencing of epivariation loci identifies a significant enrichment of rare SNVs within the CTCF canonical binding motif.

a) Hypermethylation in NOS3 (chr7:150704999-150706354) in Proband 103 (outlier in green); in the lower UCSC Genome Browser view, the DMR location is shown as a green bar, and a rare SNV that lies within the CTCF binding motif (blue region within black bar) in this same individual is shown in red. b) CTCF motif according to ENCODE Factorbook repository. Rare SNVs overlapping this CTCF binding motif were identified in four DMR carriers: (1) Proband 103: SNV (chr7:150705968 C>T), (2) Proband 70: SNV (chr19:295321 C>T), (3) Proband 176: 1bp deletion (chr20:36793857 delT), (4) HapMap samples NA19239, NA19184; NA20296: rs116767319 (chr5:177707147 C>T). c) 450k array analysis identified a DMR in NA19239, and a rare SNV (rs116767319) at that CTCF-binding motif in cis. We tested two other carriers of rs116767319 (NA19184 and NA20296) using allele-specific bisulfite sequencing, and found that both showed methylation on the T allele, thus confirming segregation of the epivariation with SNV. In contrast, a sample (NA20375) homozygous for the reference C allele is unmethylated.

In addition to mutations that disrupt regulatory elements such as TF binding sites, expansions of GC-rich tandem repeats can also result in local DNA hypermethylation. Indeed, we identified multiple individuals with gains of methylation at known CGG repeats, including FMR1/FRA3 (two cases), XLT1/FRA16A (one case)34, FRA10A/C11/FRA10 (two cases and nine controls)35, and DIP2B/FRA12 (one control)35. Both individuals with FMR1 hypermethylation were shown to carry the classic CGG expansion that causes Fragile X, and although not tested, it is likely that the gains of methylation observed at the other three known fragile sites are also caused by similar repeat expansions. While our targeted sequencing experiments of other epivariations did not identify any novel tandem repeat expansions, this class of mutation is difficult to detect with short-read sequencing. Thus, it remains possible that some of the hypermethylated epivariations we observed might be caused by this mechanism.

Previous studies have made attempts to investigate the prevalence of epigenetic changes in patients with ND–CA. Kolarova et al.34 compared the DNA methylation profiles of 82 patients against 19 controls, identifying a total of 157 DMRs. Consistent with our own findings, Kolarova et al.34 also identified patients with hypomethylation defects of MEG3 and MOV10L1. While loss of imprinting at MEG3 is a known cause of Temple syndrome, the recurrent observation of hypomethylation of MOV10L1 is significant (total n = 3 of 571 cases versus 0 of 4878 controls, p < 0.001, Fisher’s Exact test), and provides additional evidence implicating this locus with developmental defects. Similarly Aref-Eshghi et al.35 assessed methylation patterns in a total of 528 samples, identifying altered methylation levels at several imprinted loci that were absent in controls. Consistent with our own study, their cohort included two patients with hypermethylation of HM13, thus demonstrating this as another recurrent alteration found in patients with ND–CA, and suggesting that this may be an imprinting disorder.

Our study shows that epivariations are a relatively common feature in the human genome, that some are associated with
Informed consent was obtained from subjects where required. Some samples were
obtained as residual DNA remaining after clinical testing, and after anonymization obtained as residual DNA remaining after clinical testing, and after anonymization with an average age of 10 years (range: newborn to 54 years), comprising 32% females and 68% males, were enrolled in the study (Supplementary Table 1).

Inclusion criteria entailed the patient having undergone previous exome sequencing, with no pathogenic findings identified. Many samples tested had also undergone a number of locus-specific tests for common causes of ND—CA, such as Fragile X testing, genomic microarray (Affymetrix 250k SNP array, or Agilent array CGH), and/or WGS. This cohort results from a collaborative effort of multiple centers/groups, namely: 163 trios from The Seaver Autism Center for Research and Treatment at the Icahn School of Medicine at Mount Sinai (USA), 155 trios from the Pediatric Cardiac Genomics Consortium under an approved ancillary study of the 1000 Genomes Project, and combined these with SNP genotypes and RNAseq data from a total of 462 samples to measure quantitative and allelic effects of epivariations on gene expression.

In Fig. 5 we show these changes in the local gene expression, and raises the possibility that they may be implicated in the etiology of developmental disorders. In an era when WGS is being applied to many thousands of human genomes, epivariations represent a class of genetic variation that remains undetectable by purely sequence-based approaches. We anticipate that future studies exploring the relationship between sequence variation and epigenetic state will further illuminate the regulatory architecture of the human genome, providing novel insight into the consequences of non-coding mutations.

**Methods**

**Patients.** A total of 489 patients with idiopathic sporadic NDs and/or multiple CAs with an average age of 10 years (range: newborn to 54 years), comprising 32% females and 68% males, were enrolled in the study (Supplementary Table 1). Inclusion criteria entailed the patient having undergone previous exome sequencing, with no pathogenic findings identified. Many samples tested had also undergone a number of locus-specific tests for common causes of ND—CA, such as Fragile X testing, genomic microarray (Affymetrix 250k SNP array, or Agilent array CGH), and/or WGS. This cohort results from a collaborative effort of multiple centers/groups, namely: 163 trios from The Seaver Autism Center for Research and Treatment at the Icahn School of Medicine at Mount Sinai (USA), 155 trios from the Pediatric Cardiac Genomics Consortium under an approved ancillary study of the 1000 Genomes Project, and combined these with SNP genotypes and RNAseq data from a total of 462 samples to measure quantitative and allelic effects of epivariations on gene expression. We identified epivariations in 90 lymphoblastoid cell lines studied as part of the 1000 Genomes Project, and combined these with SNP genotypes and RNAseq data from a total of 462 samples to measure quantitative and allelic effects of epivariations on gene expression.

**Fig. 5** Epivariations are frequently associated with large changes in gene expression. We identified epivariations in 90 lymphoblastoid cell lines studied as part of the 1000 Genomes Project, and combined these with SNP genotypes and RNAseq data from a total of 462 samples to measure quantitative and allelic effects of epivariations on gene expression. **a** An individual with hypermethylation of the **UBE2T** promoter (solid green line) compared to 89 other individuals (dashed gray lines) presented **b** the lowest gene expression (green dot on the boxplot) of the cohort. **c** Using heterozygous SNPs within RNAseq reads, we observed monoallelic expression of **UBE2T** in the epivariation carrier (outlier highlighted in green). **d** An individual with hypomethylation of the **GTSF1** promoter (solid green line) presents **e** the highest level of expression (green dot on the boxplot). **f** Violin plots show that individuals with hypomethylated epivariations at gene promoters show significantly increased expression of that gene, whereas individuals with hypermethylated promoter epivariations show significantly reduced expression of that gene (p = 9.2 × 10^-5, Wilcoxon Rank-Sum test). In box plots (**b** and **e**), the center line shows the median; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles; outliers are shown as individual points. In the violin plot (**f**), the white dots show the median; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles.

![Fig. 5](image)

**Methods**

**Patients.** A total of 489 patients with idiopathic sporadic NDs and/or multiple CAs with an average age of 10 years (range: newborn to 54 years), comprising 32% females and 68% males, were enrolled in the study (Supplementary Table 1). Inclusion criteria entailed the patient having undergone previous exome sequencing, with no pathogenic findings identified. Many samples tested had also undergone a number of locus-specific tests for common causes of ND—CA, such as Fragile X testing, genomic microarray (Affymetrix 250k SNP array, or Agilent array CGH), and/or WGS. This cohort results from a collaborative effort of multiple centers/groups, namely: 163 trios from The Seaver Autism Center for Research and Treatment at the Icahn School of Medicine at Mount Sinai (USA), 155 trios from the Pediatric Cardiac Genomics Consortium under an approved ancillary study of the 1000 Genomes Project, and combined these with SNP genotypes and RNAseq data from a total of 462 samples to measure quantitative and allelic effects of epivariations on gene expression.

**Fig. 5** Epivariations are frequently associated with large changes in gene expression. We identified epivariations in 90 lymphoblastoid cell lines studied as part of the 1000 Genomes Project, and combined these with SNP genotypes and RNAseq data from a total of 462 samples to measure quantitative and allelic effects of epivariations on gene expression. **a** An individual with hypermethylation of the **UBE2T** promoter (solid green line) compared to 89 other individuals (dashed gray lines) presented **b** the lowest gene expression (green dot on the boxplot) of the cohort. **c** Using heterozygous SNPs within RNAseq reads, we observed monoallelic expression of **UBE2T** in the epivariation carrier (outlier highlighted in green). **d** An individual with hypomethylation of the **GTSF1** promoter (solid green line) presents **e** the highest level of expression (green dot on the boxplot). **f** Violin plots show that individuals with hypomethylated epivariations at gene promoters show significantly increased expression of that gene, whereas individuals with hypermethylated promoter epivariations show significantly reduced expression of that gene (p = 9.2 × 10^-5, Wilcoxon Rank-Sum test). In box plots (**b** and **e**), the center line shows the median; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles; outliers are shown as individual points. In the violin plot (**f**), the white dots show the median; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles.

**Methods**

**Patients.** A total of 489 patients with idiopathic sporadic NDs and/or multiple CAs with an average age of 10 years (range: newborn to 54 years), comprising 32% females and 68% males, were enrolled in the study (Supplementary Table 1). Inclusion criteria entailed the patient having undergone previous exome sequencing, with no pathogenic findings identified. Many samples tested had also undergone a number of locus-specific tests for common causes of ND—CA, such as Fragile X testing, genomic microarray (Affymetrix 250k SNP array, or Agilent array CGH), and/or WGS. This cohort results from a collaborative effort of multiple centers/groups, namely: 163 trios from The Seaver Autism Center for Research and Treatment at the Icahn School of Medicine at Mount Sinai (USA), 155 trios from the Pediatric Cardiac Genomics Consortium under an approved ancillary study of the 1000 Genomes Project, and combined these with SNP genotypes and RNAseq data from a total of 462 samples to measure quantitative and allelic effects of epivariations on gene expression.

In Fig. 5 we show these changes in the local gene expression, and raises the possibility that they may be implicated in the etiology of developmental disorders. In an era when WGS is being applied to many thousands of human genomes, epivariations represent a class of genetic variation that remains undetectable by purely sequence-based approaches. We anticipate that future studies exploring the relationship between sequence variation and epigenetic state will further illuminate the regulatory architecture of the human genome, providing novel insight into the consequences of non-coding mutations.

**Methods**

**Patients.** A total of 489 patients with idiopathic sporadic NDs and/or multiple CAs with an average age of 10 years (range: newborn to 54 years), comprising 32% females and 68% males, were enrolled in the study (Supplementary Table 1). Inclusion criteria entailed the patient having undergone previous exome sequencing, with no pathogenic findings identified. Many samples tested had also undergone a number of locus-specific tests for common causes of ND—CA, such as Fragile X testing, genomic microarray (Affymetrix 250k SNP array, or Agilent array CGH), and/or WGS. This cohort results from a collaborative effort of multiple centers/groups, namely: 163 trios from The Seaver Autism Center for Research and Treatment at the Icahn School of Medicine at Mount Sinai (USA), 155 trios from the Pediatric Cardiac Genomics Consortium under an approved ancillary study of the 1000 Genomes Project, and combined these with SNP genotypes and RNAseq data from a total of 462 samples to measure quantitative and allelic effects of epivariations on gene expression.

In Fig. 5 we show these changes in the local gene expression, and raises the possibility that they may be implicated in the etiology of developmental disorders. In an era when WGS is being applied to many thousands of human genomes, epivariations represent a class of genetic variation that remains undetectable by purely sequence-based approaches. We anticipate that future studies exploring the relationship between sequence variation and epigenetic state will further illuminate the regulatory architecture of the human genome, providing novel insight into the consequences of non-coding mutations.
analyzed a cohort with a large age range as the goal was to understand epigenetics of human aging, GSE42861 enrolled individuals with rheumatoid arthritis as well as healthy individuals, and GSE53045 included adults who smoked as well as non-smokers. In total, our control cohort had 60% females and 40% males with an average age of 56 years, range: 1–101 years old. From the samples included in the GEO entries, we excluded a small number of outlier individuals based on principal component analysis (PCA) of autosomal probes. By sequentially comparing each control sample against the remainder using our DMR calling pipeline (see below), we also removed those samples that reported >10 DMRs. A final list of the 1534 controls utilized is available on request.

In order to compare the rate of de novo epivariations in probands to a cohort of healthy individuals, we used the dataset GSE56105 from GEO, which comprised 614 healthy individuals from 117 families (mother, father, each with 2–4 children). In order to assess the frequency of epivariations in the general population, we used the dataset GSE57673 from GEO, which comprised 2711 unrelated individuals (40% Type 2 diabetes cases, 60% population controls). Each profile was generated using the DNA extracted from peripheral blood followed by hybridization to the illumina 450k array.8

Methylation array. Genome-wide DNA methylation profiling was performed using Human Methylation 450k BeadChips (Illumina Inc., San Diego, CA, USA) according to the manufacturer’s recommended protocol40. Patient samples were processed in six different batches using three different facilities: New York Genome Center (three batches), Genomics Core of the Icahn School of Medicine at Mount Sinai (one batch), and Genomics lab at Northeast Health (one batch). We observed no significant differences in the number of DMRs called per sample, or the rates of secondary validation based on array batch or processing center.

Quality control and normalization. Raw data files with β values, color, intensity, and detection p-values per probe were obtained from the genomics facilities. Quality control steps entailed performing a gender check, a screen for potential regions of homozygous deletion, PCA plots, and density plots of M values. Beta values from autosomes and β values from chromosome X were processed and analyzed separately.

We inferred patient gender by calculating both the mean p-value of chromosome Y probes, and the mean β value of chromosome X probes per sample, and compared these to the patient records. A mismatch between the array-inferred gender and the reported patient gender was detected in four samples; these samples were removed from further analysis and published GEO entries.

As the 450k array utilizes hybridization of DNA to infer a methylation value per probe, regions of homozygous deletion where probes have no target DNA to hybridize often yield very low signal. In such regions of homozygous deletion, reported β values are often highly erratic, and thus can easily appear as outlier β values when compared to the rest of the population, yielding potential false-positive DMRs in our analysis. We assessed the presence of putative homozygous deletion regions by identifying clusters of probes with failed detection p-values. For a high-quality 450k array hybridization, typically >0.1% of autosomal probes yield detection p-values >0.01. Thus, clusters of multiple probes in any one sample with failed detection p-values (p >0.01) should be extremely rare. The probable arguments were set to flag 3 kb windows where three or more probes yielded detection p >0.01 in each sample. In total, across the 489 probands, we detected 182 putative regions of homozygous deletion (Supplementary Data 1). By comparing these putative CNVs with publicly available data24 (minor allele frequency (MAF) ≥5%), we observed that 81% of clusters of probes with failed detection p-values corresponded to common CNV loci or intergenic space, validating this as an appropriate method. Regions identified as putative homozygous deletion in each sample were then excluded from the list of DMR calls.

PCA and density plots of M values were obtained with the lumi and methylumi R packages42,43. We excluded one sample from downstream analysis because it was a clear outlier by PCA. Two additional samples (Proband 488 and Preband 489) were removed from the analysis of chromosome X because they were clear outliers by PCA based on chromosome X data.

After filtering on failed methylation values (p >0.01) in each sample, we performed color correction, background correction, and quantile normalization of β values across all probands control samples using lumi and methylumi. Finally, we performed normalization to account for the different data distributions of Infinium type I and type II probes using the BMIQ package44.

Identification of candidate DMRs. Identification of putative DMRs was performed using a custom perl script (available at https://github.com/AndyMSSMLab/Scripts/tree/master). We utilized a sliding window approach to individually compare the methylation profile in each proband against the entire control cohort, detecting regions of robust outlier methylation represented by multiple independent probes with extreme methylation values, including at least one probe with methylation values well outside that observed in any control sample. This was done as single probes can present outlier values due to technical array artifacts, hybridization artifacts such as the presence of underlying sequence variants, or the presence of C-T mutations at the CpG being assayed. Stringent thresholds for calling DMRs were set as follows:

- **Hypomethylation:** the proband presents, in a 1-kb window, probes that fulfill both of the following criteria:
  (i) At least 3 probes in each have β values above the 99.9th percentile of the control distribution for that probe, and are ≥0.15 above the control mean.
  (ii) At least 1 probe with a β value ≥0.1 above the maximum observed in controls for that probe.

- **Hypomethylation:** the proband presents, in a 1-kb window, probes that fulfill both of the following criteria:
  (i) At least 3 probes in each have β values below the 0.1th percentile of the control distribution for that probe, and are ≥0.15 below the control mean.
  (ii) At least 1 probe with a β value ≥0.1 below the minimum observed in controls for that probe.

All DMRs were manually curated to remove the loci that were deemed false-positive calls. Despite performing probe-level filtering and multiple rounds of normalization of the array data, such measures are imperfect and do not remove all probes that show aberrant signals due to underlying technical or biological effects. We observed both systematic batch effects, and also sporadic false positives in single samples that were filtered, as follows:

  (i) Batch effects, i.e., technical differences due to arrays being processed in separate groups, were sometimes observed between cases and controls. Here, it was usually observed that there was either a systematic shift in β values reported by one or more probes within a region between arrays processed in different batches. In some cases, the mean of each batch was significantly different, with every sample showing a shift, whereas in other cases the means of the two populations remained similar, but a subset of the samples in one batch showed a gradient of deviations, with the β values of multiple cases lying in the extreme tail of the control distribution.

  (ii) In some cases, while 3 probes within a 1-kb region were identified as outliers, the outlier probes were not in a contiguous block as would be expected for a true methylation change, and were interspersed with other probes that showed no difference compared to the control population. We interpreted these signals as likely random groupings of individual probes that each yielded outlier beta values for some other reason, e.g., rare underlying variations that influenced probe performance, or poor hybridization performance of individual probes. Indeed, we identified that many such cases were due to regions of homozygous deletion as indicated by clusters of probes with failed array detection p-values (Supplementary Data 11).

We observed that 98.6% of the probands presented less than 10 DMRs. Due to a clear increase in the rate of false-positive DMRs as assessed by manual curation, samples with >10 DMRs were excluded (Supplementary Fig. 10). Only the 143 epivariations positive were deemed true positives by both researchers, and were kept for downstream analysis (Supplementary Fig. 2 and Supplementary Data 4). All genomic coordinates are in build GRCh37/hg19. We identified 12 recurrent epivariations (Supplementary Fig. 4). Methylation profiles of 614 healthy individuals from 117 families (GSE56105) were subject to the same analysis, comparing individual methylation profiles against 1534 controls for search of outlier DMRs (Supplementary Data 2). Taking into account the pedigree information, assessment of epivariation inheritance was performed by inspection of data plots of all family members at the DMR locus; 2711 unrelated individuals (GSE57673) were subject to the same pipeline for identification of DMRs (Supplementary Data 3).

Epivariation calling at imprinted loci. To identify epivariations affecting imprinted loci, a total of 763 45K array probes mapping to 50 imprinted loci were selected from Monk et al.44 and Joshi et al.14. The mean methylation level was calculated in each proband and control by averaging β values for all probes within each imprinted locus. For each proband, methylation changes were considered as epivariations when either the mean methylation level showed a difference greater than 3 standard deviations from the mean of controls, or when the mean β value was ≥0.08 or <0.92. Epivariations identified in imprinted loci are listed in Supplementary Table 1.

Bisulfite sequencing. In order to assess the accuracy of our identification of putative epivariations from array data, we performed secondary validation experiments using bisulfite PCR amplicon sequencing. Validation studies were performed in both the proband and parental DNAs (where available) to determine if the putative epivariations identified by bisulfitetreated DNA were validated by PCR, followed by next-generation sequencing (NGS) (Illumina MiSeq). Sample preparation for MiSeq was performed on a Fluidigm AccessArray high-throughput PCR machine with sample bar-codes incorporated in a second round of PCR. Allele-specific methylation was assessed where coverage was >100 reads. Libraries prepared using this method yielded on average 50 million MiSEQ (2 × 130 bp) platform, which scores net methylation in each amplicon based on the ratio of C to T bases at CpG positions. For each amplicon and sample, methylation
percentages (Methylated reads/Total reads) averaged across the covered CpGs (≥100X) were provided.

Seventy epivariations were selected for validation with this methodology (Supplementary Data 5). Eleven assays failed to work. Of the remainder, we verified that 33 epivariations were inherited (67% maternal and 33% paternal), 24 were de novo, and 2 were true positives but their inheritance was not assessed (Supplementary Data 4). In some instances, the interpretation of validation experiments was made complex due to highly biased allelic representation, presumably reflecting preferential PCR amplification of one allele (Supplementary Fig. 3).

To confirm the epivariations at imprinted loci, 2 μg of DNA was treated with sodium bisulfite and purified using the EpiTect Bisulfite kit (Quagen, Germantown MD). Bisulfite PCR was performed on bisulfite-treated DNA using HotStarTaq DNA Polymerase (Quagen, Germantown MD) and specific primers. After amplification, PCR products were cloned into TOPO TA vector (Invitrogen, Carlsbad, CA) and transformed into chemically competent TOP10 cells (Invitrogen, Carlsbad, CA) for subsequent sequencing using M13R primers (Supplementary Fig. 5).

Agilent custom designed array. In order to identify CNVs overlapping or adjacent to epivariations that could underlie the methylation change, we designed a custom ultra-high-resolution CGH array specifically targeting 28 DMRs and ≥600 kb of their flanking sequences. At each epivariation locus to be assayed for CNVs, we selected a mean density of 1 probe per ~150 bp to ≥ 100 bp of each DMR, and more sparse coverage (mean density of 1 probe every ~600 bp) extending a further 500 kb upstream and downstream. Normalization probes spaced throughout the autosomes and additional control probes on chromosomes X and Y were also included in the array. The custom array was designed and ordered through Agilent’s online portal (https://earray.chem.agilent.com/earray/). DNA samples were processed in the Cytogenetics and Cytogenomics Laboratory of the Icahn School of Medicine at Mount Sinai. The epivariations assayed for CNVs with this method are listed in Supplementary Data 6.

Targeted sequencing of epivariations. In order to identify SNVs within DMRs or their flanking sequences that could underlie methylation changes, we performed targeted sequencing of 36 DMRs (Supplementary Data 4), including an additional ±75 kb of sequence from each flank, using a custom sequence capture assay and a HiSeq2500 instrument. Briefly, library preparation entailed shearing DNA using a Bioanalyzer (Diagene, Denyel NL) to a mean fragment size of 300 bp, use of the KAPA Prep kit and barcoding the DNA with NextFLEX (Bio Scientific, Austin TX). Capture was performed with a custom oligonucleotide DNA capture kit (Nimblegen, Madison WI). NGS was performed using paired-end 150 bp reads generated with an Illumina HiSeq 2500. Thirty-six samples were multiplexed per sequencing lane, and sequencing was performed on the Genomics Core at New York University. Paired-end reads were mapped against the human reference genome (hg19) using BWA-MEM45 (https://github.com/lb3/bwa, v0.7.12) with default parameters. Duplicate reads were marked using Sambambl46 (https://github.com/lomereiter/sambamba, v0.5.5). Samblaster was used to generate variant call data (https://github.com/hall-lab/svtyper, v0.0.4). SVs with allele balance (AB) <0.1 were filtered out.

Variant discovery was performed using GATK’s n+1 joint genotyping protocol (https://software.broadinstitute.org/gatk/documentation/article.php?id = 3893). The protocol involves multiple steps. Initially, the gVCF file was created individually for each sample using the haplotype caller utility. Next, joint genotyping was performed on all samples together and a single vcf file generated. In all steps, we restricted genotyping to the targeted loci. The resulting variants call were annotated with CADP scores (http://cadd.gs.washington.edu/, v1.3), MAF from the 1000 Genomes Project (phase 3) using Annovar51 (http://annovar.openbioinformatics.org/en/latest/, v2016Feb01). Rare variants (100kbD > 1000) with MAF ≤ 1% were further annotated with transcription factor (TF)-binding sites predicted by the CENTIPEDE algorithm52, and DNase hypersensitivity sites from the ENCODE project53 (Supplementary Data 7). Where parental genotypes were available, concordance in the other available tissues (Supplementary Fig. 9). For each DMR called in these 90 samples, we extracted SNVs located within ±50 kb of each DMR, yielding a total of 20,398 DMR-SNV pairs. Variants were then filtered to retain only those with MAF ≤ 0.1% in the total 1000 Genomes population (Supplementary Data 8), and annotated for the overlapping TFBS using the ENCODE/Factorbook database, as described above. TFBS enrichment analysis utilized the 89 individuals without a DMR as background for each test.

Epivariations in B cells, T cells, and fibroblasts. We obtained raw DNA methylation data generated using the Illumina 450k Human Methylation BeadChip from the Genord cohort from the EDMR-EBI European Genome–Phenome Archive (https://www.ebi.ac.uk/ega/) under accession number EGAS00001000446, representing 107 fibroblast cultures, 66 T-cell cultures, and 111 immortalized B-cell cultures derived from a cohort of newborns21. As described above, we performed lumi and BMFQ normalization. Given that we lacked a large set of control samples of each cell type for comparison, DMRs in each sample were identified as outliers relative to the rest of the population, using a 1-kb sliding window. In each window, we required at least 3 probes with β value ≥ 0.15 the maximum, or ≥ 0.15 below the minimum, of the observed in other 89 individuals. Based on the β values of each probe located within the 1 kb DMR loci, we calculated the population rank of each individual carrying an epivariation defined in T cells and fibroblast methylation profiles (Supplementary Data 9).

Gene expression studies in 90 lymphoblastoid cell lines. We used normalized methylation data for a filtered set of 443,498 probes in 90 samples analyzed as part of the 1000 Genomes Project for which both variant calls and RNAseq data were also available ( GEO GSE9672523).20 DMRs were called using an outlier approach using a 1-kb sliding window, with at least 3 probes with β value ≥ 0.15 the maximum, or ≥ 0.15 below the minimum, of the observed in the other 89 individuals. RNAseq data (http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1-analysis_results/) and SNV data24 were obtained and used to measure total gene expression, and allele expression levels based on heterozygous transcribed SNVs. For the latter, at each transcribed SNP position with at least 7 overlapping reads, we made counts of the number of reads containing reference and alternate alleles. In total, there were 300,111 sites within Refseq gene annotations with at least one individual carrying a heterozygous SNP with read depth ≥7. We linked each DMR with associated genes based on physical overlap with gene promoter regions, defined as ≥2 kb from the annotated transcription start site (TSS) (Supplementary Data 10).

Analysis of SNVs around DMRs in 90 lymphoblastoid cell lines. We downloaded RNAseq data for 90 controls from the 1000 Genomes Project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phasedRelease53/SV/mapping/) and RNAseq data from GEO (GSE84872) the methylation profiles generated using the Illumina 450k array from five deceased individuals, each profiled in six different tissues (peripheral blood, liver, skeletal muscle, pancreas, omental fat, and spleen)24. Data for each tissue were filtered and normalized separately, following the same approach as used for the previously described data53. For each DMR called in each of these 90 samples, we extracted SNVs located within ±50 kb of each DMR, yielding a total of 20,398 DMR-SNV pairs. Variants were then filtered to retain only those with MAF ≤ 0.1% in the total 1000 Genomes population (Supplementary Data 8), and annotated for the overlapping TFBS using the ENCODE/Factorbook data, as described above. TFBS enrichment analysis utilized the 89 individuals without a DMR as background for each test.

Conservation of epivariations across multiple tissues. We downloaded from GEO (GSE48472) the methylation profiles generated using the Illumina 450k array from five deceased individuals, each profiled in six different tissues (peripheral blood, liver, skeletal muscle, pancreas, omental fat, and spleen)24. Data for each tissue were filtered and normalized separately, following the same approach as used for the previously described data53. For each tissue, we then quantiled normalized β values with those for the 1534 controls. DMRs were then called in each of the six blood samples using the same approach as described above (see “Methylation array”). For each DMR observed in blood, we generated plots of these loci and manually curated for concordance in the other available tissues (Supplementary Fig. 9).

Code availability. Computer code used in this study is available on GitHub: https://github.com/AndyMSSMLab/Scripts/tree/master

Data availability. The methylation array data used in this publication have been deposited in Gene Expression Omnibus (GEO) under accession GSE89353 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE89353).

Received: 29 August 2017 Accepted: 8 May 2018
Published online: 25 May 2018

References
1. Horsthemke, B. Epimutations in human disease. Curr. Top. Microbiol. Immunol. 310, 45–59 (2006).
2. Castillo, E. A. et al. Prediction of MLHI constitutional epimutations as a cause of Lynch syndrome in unselected versus selected consecutive series of patients with colorectal cancer. J. Med. Genet. 52, 498–502 (2015).
3. Willemsen, R., Levengra, J., & Oostra, B. CGG repeat in the FMR1 gene: size matters. Clin. Genet. 80, 214–225 (2011).
4. Bamshad, M. J. et al. Exome sequencing as a tool for Mendelian disease gene discovery - supplementary information. Nat. Rev. Genet. 12, 745–755 (2011).
5. Miller, D. T. et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am. J. Hum. Genet. 89, 1856–1870 (2011).
6. Gilissen, C. et al. Genome sequencing identifies major causes of severe intellectual disability. Nature 511, 344–347 (2014).
7. McRae, A. F. et al. Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biol. 15, R73 (2014).
8. Ishbe, B. et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biol. 16, 37 (2015).
9. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. Nat. Rev. Genet. 13, 565–575 (2012).
10. Honny, J. et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science 350, 1262–1266 (2015).
11. SaraDou, T. et al. Folate-sensitive fragile site FRA10A is due to an expansion of a CGG repeat in a novel gene, FRA10AC1, encoding a nuclear protein. Genomics 84, 69–81 (2004).
12. Kagami, M. et al. Deletions and epimutations affecting the human 14q32.2 region in patients with Silver-Russell syndrome and growth retardation. J. Biol. Chem. 287, 28750–28757 (2003).
13. Joshua, R. S. et al. Methylation profiling of uniparental disomy subjects provides a map of parental epigenetic bias in the human genome. Am. J. Hum. Genet. 99, 555–566 (2016).
14. Gronskov, K. et al. Deletions and rearrangements of the H19/IGF2 enhancer region in patients with Silver-Russell syndrome and growth retardation. J. Med. Genet. 48, 308–311 (2011).
15. Lichtenberg, M. et al. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. Nat. Genet. 41, 112–117 (2009).
16. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303, 3997v1 (2013).
17. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 30, 2503–2505 (2014).
18. Wang, T. et al. Genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
19. Van der Auwer, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 11, 11.10.1–11.10.33 (2014).
20. Wang, T. et al. The human genome variation server: a database for storing high-quality DNA variation data. Nucleic Acids Res. 38, 137–141 (2010).
21. Pique-Regi, R., Degner, J. & Pai, A. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 21, 447–455 (2011).
22. Yen, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science 356, eaaf2239 (2017).
23. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nat. Genet. 50, 506–511 (2018).
24. Kutter-Rietschel, M. et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. elife 3, 1–13 (2014).
25. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. 13, 484–492 (2012).
26. Li, H. & Handsaker, B. E. SAMtools and Genome-wide Association studies. Nat. Genet. 46, 274–276 (2014).
27. Li, H. et al. An improved map of human genomic variation from short-read sequencing. Genome Res. 21, 366–371 (2011).
28. Zhu, H. et al. Computational identification of human intergenic regulatory elements using a large-scale dataset of ChIP-seq experiments. Genome Res. 23, 1517–1530 (2013).
29. Wang, T. et al. A method for genome-wide association analysis that recognizes functional variation in humans. Genome Res. 19, 17 (2009).
30. Rhee, S. Y. et al. ANNOVAR: functional annotation of genome variants. Nucleic Acids Res. 38, e124 (2010).
31. Wang, T. et al. Genome-wide association studies toward the genetic diagnoses of complex human diseases. Nat. Rev. Genet. 12, 486–492 (2011).
32. Wang, T. et al. The Broad Institute of Harvard and MIT: a Wiki database for transcription factor-binding data generated by the ENCODE consortium. Nat. Biotechnol. 41, 171–176 (2013).
33. Layer, R. et al. Chiar, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15, R84 (2014).
34. Wang, T. et al. Factobook.org: a Wiki database for transcription factor-binding data generated by the ENCODE consortium. Nat. Biotechnol. 41, 171–176 (2013).
35. Li, H. & Handsaker, B. E. SAMtools and Genome-wide Association studies. Nat. Genet. 46, 274–276 (2014).
36. Wang, T. et al. A method for genome-wide association analysis that recognizes functional variation in humans. Genome Res. 19, 17 (2009).
37. Wang, T. et al. The Broad Institute of Harvard and MIT: a Wiki database for transcription factor-binding data generated by the ENCODE consortium. Nat. Biotechnol. 41, 171–176 (2013).
38. Layer, R. et al. Chiar, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15, R84 (2014).
39. Wang, T. et al. Factobook.org: a Wiki database for transcription factor-binding data generated by the ENCODE consortium. Nat. Biotechnol. 41, 171–176 (2013).
40. Layer, R. et al. Chiar, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15, R84 (2014).
41. Chuang, C. et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat. Methods 12, 966–968 (2015).
42. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).
43. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. Trends Genet. 29, 575–584 (2013).

Acknowledgements

The authors are grateful to the patients and families who participated in this study and to the collaborators who supported patient recruitment. This work was supported by NIH grants 1R01GM111600 and research grant 6-FY13-92 from the March of Dimes to A.J.S., grant HL098123 to B.D.G. and A.J.S., Gulbenkian Programme for Advanced Medical Education and the Portuguese Foundation for Science and Technology (SFRH/BD1NT/51549/2011, P3IC/IC/83026/2007, 3IC/IC/83021/2007, SFRH/BD/90167/2012, Portugal) to P.M., F.L., and M.B., by the Northern Portugal Regional Operational Programme (NORTE 2020), under the 2020 Partnership Agreement, through the European Regional Development Fund (FEDER) (NORTE-01-01-FEDER-00003) to P.M., a Beitru of Pinos Postdoctoral Fellowship to R.S.J. (2011BP-A00515), and a Seaver Foundation Grant to S.D.R. The views expressed are those of the authors and do not necessarily reflect those of the National Heart, Lung, and Blood Institute or the National Institutes of Health. Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S100D01852. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.
Author contributions
M.B., R.S.J., P.G., H.G.R., J.D.B., B.D.G., and A.J.S. were leading contributors to the design and analysis of this study; M.B., T.K., D.E.G., G.S., P.M., H.G.R., J.D.B., and B.D.G. contributed with samples of probands and relatives; M.B., D.E.G., S.D.R., J.R., F.L., P.M., L.V., T.K., and G.S. contributed with patient clinical/genetic information; P.G., N.P., B.J., C.T.W., and K.C. wrote and performed bioinformatic analysis; A.M.T. analyzed and validated the methylation profiles of imprinted loci; W.G. performed library preparation and capture for targeted sequencing; C.T. contributed for Agilent custom designed aCGH; H.M. and L.E. processed the Agilent custom designed aCGH; M.B. and A.J.S. wrote the manuscript, all authors commented on it.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04540-x.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018