Direct haplotype-resolved 5-base HiFi sequencing for genome-wide profiling of hypermethylation outliers in a rare disease cohort

Warren A Cheung[1], William J Rowell[2], Emily Farrow[1], Adam F Johnson[1], Richard Hall[2], Ana SA Cohen[3], John Means[1], Tricia Zion[1], Daniel M Portik[2], Christopher T Saunders[2], Boryana Koseva[1], Chengpeng Bi[1], Carl Schwendinger-Schreck[1], Byunggil Yoo[1], Jeffrey J Johnston[1], Margaret Gibson[1], Isabelle Thiffault[3], Scott T Younger[1], Thomas Curran[4], Aaron M Wenger[2], Elin Grundberg[1] and Tomi Pastinen[1*]

[1] Department of Pediatrics, Genomic Medicine Center, Children's Mercy Kansas City, Kansas City, MO
[2] Pacific Biosciences, Menlo Park, CA
[3] Department of Pathology and Laboratory Medicine, Children's Mercy Kansas City, Kansas City, MO
[4] Children's Mercy Research Institute, Kansas City, MO
*Correspondence to tpastinen@cmh.edu

Introductory paragraph

Long-read HiFi genome sequencing (GS) allows for accurate detection and direct phasing of single nucleotide variants (SNV), indels and structural variants (SV) [1]. Recent algorithmic development enables simultaneous detection of CpG methylation (mCpG) for analysis of regulatory element (RE) activity in HiFi-GS. We generated a comprehensive haplotype-resolved HiFi-GS dataset from a rare disease cohort of 276 samples in 152 families to identify rare (~0.5%) hyper-mCpG events. We found that 81% of these events are allele-specific and predicted to cause loss of RE (LRE). We demonstrated heritability of extreme hyper-mCpG including rare *cis* SNVs and SVs causing short (~200bp) and large hyper-mCpG events (>1 kb), respectively. We identified novel repeat expansions in proximal promoters predicting allelic gene silencing via hyper-mCpG and demonstrated allelic transcriptional events downstream. On average 30-40 LREs overlapped rare disease genes per patient providing indications for variation prioritization. LRE led to a previously undiagnosed pathogenic allele in *DIP2B* causing global developmental delay. We propose that HiFi-GS in unsolved rare disease cases will allow detection of unconventional diseases alleles due to LRE.

Main text

Short-read exome (srES) or srGS is the tool of choice for the detection of SNVs in most human genetic applications. However, even with strict case selection for rare genetic disease in a clinical trial, srGS achieved only a ~30% diagnostic rate [2], leaving most rare disease cases remain unsolved. On the other hand, 3rd generation long-read platforms, such as PacBio's Single Molecule, Real-Time (SMRT) HiFi-GS technology (12 kb – 16 kb) has been demonstrated to produce not only high quality SNV in difficult-to-map regions but also accurately detects SVs genome-wide [3].

Within our large pediatric rare disease program, Genomic Answers for Kids (GA4K), collecting genomic data and health information for families with a suspected genetic disorder, we have integrated an enhanced sequencing pipeline using HiFi-GS into the routine follow-up of unsolved cases. Our pilot phase of ~1000 families revealed that incorporating SVs from HiFi-GS added up to 13% of new diagnoses in previously unsolved cases and HiFi-GS increased discovery rate with >4-fold more rare coding SVs than srGS [1]. However, the relative impact on undiagnosed rare disease due to incomplete interpretation of detected variants such as those mapping to non-coding regions remains unknown.

We have shown that non-coding SNVs can have pervasive effect on genome function and RE activity from splicing variation [4,5] to transcript levels in tissues and primary cells [6,7] as well as chromatin-states [8] and mCpG levels [9,10]. We exploited heterozygosity of functional alleles to augment the detection of differential RE activity [11] in the case of rare disease and for common population SNVs as well as quantifying other allele silencing effects, such as imprinting or X-inactivation measurable in chromatin, mCpG and gene expression data [12-14]. We also previously showed that allelic RE hyper-mCpG reflects allelic regulatory/gene silencing and have shown that genetic effects altering methylation are more likely to be shared across tissues [15,16].

Established hypermethylation signatures are linked to several monogenic diseases [17,18] including imprinting disorders [19,20], and "epimutations", where missing rare disease alleles were non-coding leading to hypermethylation and promoter inactivation are emerging at locus specific investigations [21]. In most known cases of disease impactful hypermethylation the effect is restricted to one allele but even higher resolution techniques with short-reads for DNA (mCpG) methylation (WGBS) [16] lack resolution for genome-wide analysis of allele-specific effects.

A recent HiFi-GS production pipeline update (Sequel IIe system release v11, PacBio, Menlo Park, CA) combines a deep learning model that integrates sequencing kinetics and base context, to generate mCpG profiles genome-wide from standard sequencing libraries. The augmented 5-base HiFi-GS platform allows single-molecular resolution of mCpG together with phasing from long contiguous accurate reads, with the potential to detect allele-specific events at an order of magnitude increased efficiency as compared to WGBS.

Leveraging the GA4K program, we have generated a comprehensive mCpG dataset from 1367 enrolled participants (**Supplemental Table 1**) using a combination of WGBS (N=1184) and HiFi-GS (N=276; **Supplemental Table 2**) of which a subset (N=93) has corresponding measurements. We used the latter set to extract mCpG levels for paired sites (>20x coverage, 16.9M CpGs) for sample-wise correlations showing remarkable consistency across the two methods (median Pearson R=0.90, **Supplemental Table 3**). In addition, we extracted the top 500 most variable autosomal CpGs based on a subset of the HiFi-GS samples sequenced at high depth (N=139) for CpG-based correlations. Similarly, we noted high correlations compared to randomly permuted values (median Pearson R=0.86, **Supplemental Figure 1**).

We first tested using positive controls (i.e. individuals with unstable repeat disorders) how genome-wide mCpG data by WGBS performs with known disease relevant regions demonstrating hyper-mCpG footprints in *FMR1* and *FXN* (**Supplemental Figure 2**) and also two newly identified FRDA1 (*FXN* intron 1 expansion) carriers. However, these WGBS data were

not capable of reading into the pathogenic repeat or resolve carrier methylation from homozygou. We then applied HiFi-GS, where the haplotype-resolved HiFi-GS coupled with simultaneous high resolution mCpG profiling allowed us both to detect a *DMPK* repeat expansion and its associated hyper-mCpG signature [18] in a congenital case of Myotonic dystrophy type 1 (DM1) (**Figure 1**). We further demonstrate that haplotype resolved mCpG across differentially methylated regions (DMRs) can be broadly utilized for the diagnosis of imprinting disorders [19]. We exemplify this resolution at causal regions linked to Albright hereditary osteodystrophy [22] (*GNAS-AS1*; **Supplemental Figure 3**), transient neonatal diabetes mellitus [23] (*PLAGL1*; **Supplemental Figure 4**), Schaaf-Yang syndrome [24] (*MAGEL2*; **Supplemental Figure 5**) and Temple Syndrome [25] (*MEG3/DLK1*; **Supplemental Figure 6**).

As shown for unstable repeat disorders, hyper-CpG can induce transcriptional silencing of disease genes via inactivation or loss of RE (LRE). Thus, we hypothesized that screening for outlying hyper-mCpG events genome-wide may aid the identification of non-coding, functional rare SNVs and SVs in unsolved cases. We first ranked each CpG based on the mCpG distribution to catalog extreme hyper-mCpGs (**Methods**) and then classified as a hyper-CpG tiles (200bp) if two or more hyper-CpGs in the tile were extreme. The hyper-CpG tile was classified as rare if it was only present in two or less unrelated individuals. Rare hyper-CpG tiles were then filtered reporting only those where the average z-score of all the CpGs of the tile was two or more (**Figure 2**). Using these criteria, we found 25,543 extreme hyper-mCpGs tiles (**Supplemental Table 4)** of which 81% was haploid specific. The magnitude of haploid dependency of extreme hyper-mCpGs tiles uniquely captured by HiFi-GS likely explains the lower abundance of similar calculations using WGBS: the average number of extreme hyper-mCpGs tiles per individual was eight while HiFi-GS yielded on average 117 extreme hyper-mCpGs per individual (**Supplemental Table 5)**. However, these conservative measures of extreme hyper-mCpGs tiles reduces true discovery at dynamic regions where mCpGs show variation in the population. To analyze extreme hyper-mCpGs outliers within these regions, we allowed the hyper-mCpGs tiles to be present in three or more unrelated individuals (removed the rarity criteria), but only kept instances where the average z-score of the CpGs in the hyper-mCpG tile was five or greater. These additional criteria retained another 31,726 hyper-mCpGs tiles (**Supplemental Table 6)**.

We then focused our analyses to the HiFi-GS subset carried out in affected patients (173 patients from 152 families**; Supplemental Table 1**) to provide the first genome-wide characterization of rare hyper-mCpG events in a rare disease cohort which included a total of 36,196 hyper-mCpGs (autosomes only; **Supplemental Table 7**). We observed 3.1-fold more sharing of same hypermethylated tiles among related patients than expected by chance ($P_{perm}<0.05$) indicating potential genetic underpinnings of rare hyper-mCpG events. We found that, using an empirical q value of less than 10%, causal rare SNVs (minor allele frequency, MAF, <0.5%) and rare SVs (MAF<1%) linked to rare hyper-mCpG events are located within ~1-2kb and <10kb for SNVs for SVs, respectively. (**Figure 3A-B**). To bolster the potential causality for local rare genetic variation we exploited the haploid subset of hyper-mCpGs coupled with phased rare SNVs from same reads (**Figure 3C-D**) and showed statistically significant enrichment for local rare *cis*-variants extending up to 1kb from the hyper-CpG event. Similarly, using orthogonal hyper-mCpG discovery in additional patient samples by WGBS indicates that proximal rare *cis* variants among unrelated individuals can lead to recurrence of hyper-mCpG tile in unrelated samples

(**Supplemental Figure 7**). However, the local SVs and IN/DEL variation in reads is not phased by WhatsHap (**Methods**) and required manual curation. For 40 randomly chosen unphased short DEL variants called by DeepVariant and verified in read data we observed 31/40 (binomial P= 0.0007) in *cis* suggesting similar behavior as for SNVs with respect to allelic and distance distribution.

Various hyper-mCpGs-signatures for local, rare genetic variation in *cis* are evident in our data, from SNVs (**Supplemental Figure 8**) and SVs (**Supplemental Figure 9**) to discovery of new rare hypermethylated repeat expansions (**Supplemental Figure 10-12**), signatures for duplications (**Supplemental Figure 13**), insertions (**Supplemental Figure 14**) and deletions (**Supplemental Figures 15-17**). Finally, most hyper-mCpG tiles are typically restricted to few hundred bp (**Figure 4**) in the immediate vicinity of detected hyper-mCpG tile. However, 1494/36,196 (4%) of the tiles are linked to larger mCpG perturbation (2 or more tiles) and among such extended signals the local SVs (<1kb) are significantly more common than expected based on their overall distribution: 129/783 (16%) versus 96/1057 (9%) local (<1kb) SVs or *cis*-SNVs are observed with larger hyper-mCpG footprint, respectively.

Since we focused on rare hyper-mCpG it is anticipated that the large magnitude of outliers can predominantly be found in the relatively hypomethylated, dynamic [26] REs of the human genome. Consequently, the rare outlier may contribute to RE silencing and impact gene expression. We demonstrated the functional translation of mCpG in a subset of analyses from patients by co-occurrence of allelic rare hyper-mCpG events in proximal RE and concurrent allelic mRNA silencing (**Supplemental Figure 18**). Notably, this was achieved in non-blood cells (iPSCs) from a subset of patients (**Methods**) indicating that some variation exhibits relative tissue independence.

Having genetically and functionally characterized rare hyper-mCpGs tiles we next queried their relationship with rare (OMIM) disease genes to explore candidate functional changes for unsolved disease. Among the patients we observed a total of 5,438 hyper-mCpG tiles across 4,400 OMIM genes (~30 per patient). To focus the candidate set further we selected OMIM hyper-mCpG tiles close to the proximal promoter of the gene (+/- 1kb) this yielded 1,341 regions. We then further restricted to hyper-mCpGs with local *cis*-rare SNV or SV which yielded 66 regions. Manual curation among these identified allelic hyper-mCpG of recessive disease genes (**Figure 5**). Four hypermethylation events (in three patients) were flagged for follow-up one tile in *NSD1,* one in *SET* and two adjacent hyper-mCpGs tiles in *DIP2B.* Follow-up revealed that the hypermethylation of proximal regulatory region of *DIP2B* is immediately adjacent to previously undiagnosed [1] repeat expansion within pathogenic range (~280-300 repeats, >250 considered pathogenic). This diagnostic finding was clinically validated by repeat PCR (**Methods**). Within our large HiFi-GS sample we had additional unsolved patient specimens with expanded *DIP2B* repeats (**Figure 6**), however all other expansions were below pathogenic length and remained hypomethylated – exemplifying the augmented power of long repeat resolving reads coupled with mCpG detection.

Rare disease, and particularly unsolved genetic diseases are a translational challenge in the era of advanced molecular diagnostics. Better data sharing and increasing understanding of expressivity through larger molecularly classified rare disease cohorts have the capacity to discover new genes and assign function to variants of unknown significance (VUS), respectively. However,

structural genetic variation and non-coding variation with disease causing potential are areas of potential unrealized diagnostic yield due to technological and analytical hurdles. Suggestions for closing technology gaps include coupling next-generation sequencing (NGS) with additional assays gathering either transcriptomic [27] or epigenomic [28] data. However, integrated genome-wide, high-resolution assessment of any two data modalities coupled with high haploid distinction has not been attempted to date. In our rare disease cohort, we demonstrate several key aspects of mCpG HiFi-GS for exploration of disease variation. Known disease-linked -and parent-of-origin mCpG is recovered alongside with full GS. Rare novel variation in epigenome is heritable and can be shown to be physically linked to rare non-coding genetic variation in long reads. A substantial fraction of outliers' signatures is caused by complex, local rare SV known to be challenging for NGS. We demonstrate predicted silencing of REs represented by hyper-mCpG outliers propagating to transcriptional silencing in same patient but in distinct lineage cells, showing that blood-based 5-base HiFi-GS can have multi-tissue relevance. Employing an analytical approach to identifying population variation, we can limit the search to reasonable sets of candidate alleles for prioritization of manual curation where a stringent filter already identified previously missed diagnosis for a rare disease involving *DIP2B*. However, the potential discoveries by this platform, even within this dataset with follow-up validation, will expand to novel disease alleles and demonstrates a new ability to link non-coding variation to clinical evaluation in rare diseases.

Methods

Study Cohort

The study cohort described includes 1243 affected patients from 1078 families, with a total of 1367 individuals (detailed in **Supplemental Table 1**) enrolled in the Genomic Answers for Kids program [1]. Patients with a suspected rare disease were referred from multiple different specialties, with the largest proportion nominated by Clinical Genetics, followed by Neurology. A continuum of pediatric conditions is represented, ranging from congenital anomalies to more subtle neurological and neurobehavioral clinical presentations later in childhood. Of the 1243 affected patients, 141 had a known genetic diagnosis at the initiation of the study.

Sample collection and preparation

Whole blood was obtained from each study participant in EDTA and sodium heparin collection tubes for DNA and peripheral blood mononuclear cell isolation (PBMC), respectively. DNA was isolated using the chemagic™ 360 automated platform (PerkinElmer) and stored in -80°C. PBMCs were isolated using a RoboSep-S (StemCell) and the EasySep Direct Human PBMC Isolation Kit (StemCell 19654RF). After automated separation, the enriched cell fraction was centrifuged at 300 rcf for 8 minutes then resuspended in 1 mL of ACK Lysing Buffer (Thermo Fisher A1049201) and incubated at room temperature for 5 minutes. The cell suspension was diluted with 13 mL of PBS (Thermo Fisher 10010023) supplemented with 2% FBS (Cytiva SH30088.03HI) then centrifuged at 300 rcf for 8 minutes. The cell pellet was resuspended in 1 mL of PBS + 2% FBS, and cell count and viability were assessed using a Countess II automated cell counter (Thermo Fisher). Cells were centrifuged at 300 rcf for 8 minutes and resuspended in 1 mL of cold CryoStor CS10 (StemCell 07930) then transferred to a cryogenic storage vial. Cells were frozen slowly in a Corning CoolCell FTS30 placed at -80°C overnight then transferred to liquid nitrogen vapor the following day for long-term storage. Human patient-specific induced pluripotent stem cells (iPSCs) were generated from a subset of the patient's PBMCs using episomal vectors. In brief, PBMCs were maintained in StemSpan SFEM II (STEMCELL Tech, 09605) media plus 1X Antibiotic-Antimycotic (ThermoFisher, 15240062) for 3-5 days, followed by nucleofection of the following episomal plasmids: pCXLE-hOCT3/4-shp53-F (Addgene, 27077), pCXLE-hSK (Addgene, 27078), pCXWB-EBNA1 (Addgene, 37624) and pCXLE-hUL (Addgene, 27080). After nucleofection, transfected PBMCs were plated onto 35-mm matrigel (Corning, 354277) coated tissue culture dishes in Stemspan SFEM II media supplemented with 10 μM Y-27632 (Tocris, 1254). Two days later, ReproTESR (STEMCELL Tech, 05926) was added and plates centrifuged at 50xg for 30 minutes. Every other day fresh ReproTESR was added until iPSC colonies were visible. Once colonies were visible ReLeSR (STEMCELL Tech, 05872) was used to isolate iPSC colonies. iPSC colonies were maintained in mTeSR1 complete media (STEMCELL Tech, 85850) on matrigel coated tissue culture plates and refed every other day or as needed until ready. Cells were cultured at 37 °C in 5% CO2.

PacBio HiFi long-read genome sequencing and analysis

In total of ~5 ug of DNA per sample was sheared to a target size of 14 kb using the Diagenode Megaruptor3 (Diagenode, Liege, Belgium). SMRTbell libraries were prepared with the SMRTbell Express Template Prep Kit 2.0 (100-938-900, Pacific Biosciences, Menlo Park, CA) following the manufacturer's standard protocol (101-693-800) with some modifications as described previously [1]. Libraries were sequenced on the Sequel IIe Systems using the Sequel II

Binding Kit 2.0 (101-842-900) or 2.2 (102-089-000) and Sequel II Sequencing Kit 2.0 (101-820-200) with 30 hr movies/SMRT cell. Samples were sequenced to a target of >10X coverage. Circular consensus reads were generated with ccs v6.3 (https://github.com/PacificBiosciences/ccs) using the "--hifi-kinetics" option to generate consensus kinetics tags, and primrose v1.1 (https://github.com/PacificBiosciences/primrose) was used to predict 5mC modification of each CpG motif and generate base modification ("MM") and base modification probability ("ML") BAM tags. HiFi Read mapping, variant calling, and genome assembly were performed using a Snakemake workflow (https://github.com/PacificBiosciences/pb-human-wgs-workflow-snakemake). HiFi reads were mapped to GRCh38 (GCA_000001405.15) with pbmm2 v1.9 (https://github.com/PacificBiosciences/pbmm2). Structural variants were called with pbsv v2.8 (https://github.com/PacificBiosciences/pbsv) with "--hifi --tandem-repeats human_GRCh38_no_alt_analysis_set.trf.bed" options to pbsv discover and "--hifi -m 20" options to pbsv call. Small variants were called with DeepVariant v1.3 following DeepVariant best practices for PacBio reads (https://github.com/google/deepvariant/blob/r1.3/docs/deepvariant-pacbio-model-case-study.md) and locally phased with WhatsHap v1.0 [29]. Local phase haplotype tags ("HP") were added to the aligned BAM by WhatsHap v1.0. Pileup-based consensus methylation sites and probabilities were generated by the script "aligned_bam_to_cpg_scores.py" from pb-CpG-tools v1.1.0 (https://github.com/PacificBiosciences/pb-CpG-tools/) with the "-q 1 -m denovo -p model -c 10" options.

Whole-genome bisulfite sequencing and analysis

The same source of DNA used for PacBio HiFi long-read genome sequencing was used for whole-genome bisulfite sequencing. Whole-genome sequencing libraries were generated from 1ug of genomic DNA spiked with 0.1% (w/w) unmethylated λ DNA (Promega) previously fragmented to 300–400 bp peak sizes using the Covaris focused-ultrasonicator E210. Fragment size was controlled on a Bioanalyzer DNA 1000 Chip (Agilent) and the KAPA High Throughput Library Preparation Kit (KAPA Biosystems) was applied. End repair of the generated dsDNA with 3′- or 5′-overhangs, adenylation of 3′-ends, adaptor ligation and clean-up steps were carried out as per KAPA Biosystems' recommendations. The cleaned-up ligation product was then analyzed on a Bioanalyzer High Sensitivity DNA Chip (Agilent) and quantified by PicoGreen (Life Technologies). Samples were then bisulfite converted using the EZ DNA Methylation-Lightning Gold Kit (Zymo), according to the manufacturer's protocol. Bisulfite-converted DNA was quantified using OliGreen (Life Technologies) and, based on quantity, amplified by 9–12 cycles of PCR using the Kapa Hifi Uracil+DNA polymerase (KAPA Biosystems), according to the manufacturer's protocol. The amplified libraries were purified using Ampure Beads and validated on Bioanalyzer High Sensitivity DNA Chips and quantified by PicoGreen. Libraries were sequenced on the Illumina NovaSeq6000 system using 150-bp paired-end sequencing. Whole genome bisulfite sequence reads were pre-processed with fastp to trim adapter and low-quality bases, then alignment was performed with Illumina's DRAGEN aligner followed by post-processing with samtools, Picard, Bismark [30] and Bis-SNP [31] for marking of duplicates, methylation calling and SNV calling. To avoid potential biases in downstream analyses, we applied our benchmark filtering criteria as follows; ≥ 5 total reads, no overlap with SNPs (dbSNP 137), ≤ 20% methylation difference between strands, no overlap with DAC Blacklisted Regions (DBRs) or Duke Excluded Regions (DERs) generated by the ENCODE project:

(http://hgwdev.cse.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability). Methylation values at each site were calculated as total (forward and reverse) non-converted C-reads over total (forward and reverse) reads. CpGs were counted once per location combining both strands together.

Classification of extreme hypermethylation outliers

The high-depth HiFi-GS data set was used to calculate mean, standard deviation (SD) and 95th quantile (Q95) for each of the 28,195,690 CpGs (**Supplemental Table 2)**. Then, using the HiFi-GS data set, extreme hyper mCpG tiles (200bp) were identified if two or more CpGs per tile each had mCpG deviating three SD than the Q95 for that CpG. Tiles were further filtered based on the average z-score of each mCpG value mapping to each tile. For comparison, we similarly calculated mean, SD and 99th quantile (Q99) for 27,097,911 profiled by WGBS. Extreme mCpG tiles were identified when two or more CpGs per tile each had mCpG deviating one or more SD than the Q99 for that CpG. The hyper-CpG tile was further classified as rare in two ways: 1) present in two or less unrelated individuals and the average z-score of all the CpGs of the tile was two or more and 2) present in three or more unrelated individuals, but only kept instances where the average z-score of the CpGs in the hyper-mCpG tile was five or greater.

Mapping genetic variation to extreme hypermethylation outliers

The distance from the extreme bin to the closest gnomAD [32] high confidence variant in the sample with the extreme bin was compared to the distance from the same bin coordinates to the closest gnomAD high confidence variant in all other samples (with a variant on that chromosome) to obtain a percentile rank. The distance to the closest phased (via whatshap) heterozygote gnomAD high confidence variant on hap1 in the sample was also compared with the extreme bin against the distance from that bin coordinate to the closest phased hap1 gnomAD variant in all other samples to obtain a percentile rank, and similarly for the closest phased heterozygote variant on hap2.

PacBio HiFi long-read transcript (isoform) sequencing and analysis

RNA was isolated from 17 iPSC samples using a RNeasy Mini Kit (Qiagen, Cat. No. 74104), following the Quick Start Protocol associated with the kit. The RNA concentration for each sample was determined with a Qubit RNA BR Assay Kit (ThermoFisher, Q10210) and the RIN was determined with a RNA ScreenTape (Agilent, Cat. No. 5067-5577 and 5067-5576) on the TapeStation platform. A maximum of 300ng of RNA with a RIN score greater than 7.0 was aliquoted from each sample to be used as an input into cDNA synthesis for Iso-Seq library preparation. If the concentration of RNA was too low for there to be 300ng of RNA in 7µL, 300ng of RNA was aliquoted into a 1.5mL tube and then the sample was concentrated using a vacuum concentration system without heating until the sample was less than or equal to 7µL in volume. The RNA underwent cDNA synthesis with a NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (NEB, Cat. No. E6421S) and an Iso-Seq Express Oligo Kit (PacBio, Cat. No. 101-737-500) as described in the Iso-Seq Express Template Preparation for Sequel and Sequel II Systems protocol from PacBio. The samples were not amplified with barcoded primers during the cDNA amplification steps since the downstream libraries would not be multiplexed for sequencing. The amplified cDNA was purified for long transcripts greater than 3kb in length after the first cDNA amplification step. The quantity of cDNA was determined with a Qubit dsDNA HS Assay Kit (ThermoFisher, Cat. No. Q32854) and if the

cDNA yield was less than 160ng, the required input for the Sequel II system, the samples were reamplified following the procedure described in Appendix 1 of the Iso-Seq protocol. NEBNext High-Fidelity 2X PCR Master Mix (NEB, Cat. No. M0541S) used in the PCR reamplification master mix and after reamplification, the cDNA was bead-cleaned following the low-yielded sample procedure. The cDNA was quantified again to ensure that there was adequate yield for SMRTbell library preparation. In total, 160 to 500ng of cDNA from each sample was used as an input into the SMRTbell library preparation protocol without pooling the cDNA. The final concentration of the libraries was determined with a Qubit dsDNA HS Assay Kit and the library size was determined with a High Sensitivity D5000 ScreenTape (Agilent, Cat. No. 5067-5592 and 5067-5593) on the TapeStation platform. Libraries were sequenced on the Sequel IIe Systems using the Sequel II Binding Kit 2.0 (101-842-900) or 2.1 (101-820-500) and Sequel II Sequencing Kit 2.0 (101-820-200) with 30 hr movies/SMRT cell at 90pM loading with 1 cell/sample. The IsoSeq3 pipeline (https://github.com/PacificBiosciences/IsoSeq) was used to generate full-length non-concatemer (FLNC) reads which were then aligned to the reference genome (GRCh38) using minimap2 with the argument -ax splice:hq. Reads overlapping a CCDS region and informative of a heterozygous SNV were kept and each allele was count for each read covered by at least 5X.

Clinical Validation

"Short" PCR amplification of the normal allele of the FRA12A/*DIP2B* repeat– containing region was performed with the aid of 2X Failsafe buffer J (received with FailSafe enzyme) with use of primers derived from the sequences flanking the repeat (DIP2B-CGG-F primer, 5' - GTCTTC[1]AGCCTGACTGGGCTGG-3' , and DIP2B-CGG-R, primer 5'- CCGG[1]CGACGGCTCCAGGCCTCG-3'. 95°C-3'; (95°C-1';60°C-1'-0.5/cycle; 72°C-1')x10; (95°C-1'; 55°C-1';72-1')x25; 72°C-1', 4°C ∞). The PCR products were electrophoresed on a 1.5% agarose gel. Triplet-repeat PCR was also performed by primed PCR according to the principles described previously [33,34] to detect CAG-repeat expansion.
Three different primers were added to the PCR mixture: a single forward fluorescently labeled primer (DIP2B-CGG-F) and a combination of two reverse primers (P4, 5' - TACGCATCCCAGTTTGAGACGGCCGCCGCCGCCGCCGC-3', and P3, 5' - TACGCATCCCAGTTTGAGACG-3') in a 1:10 ratio. The reverse primer P4 anneals at different sites of the CGG repeat, which produces PCR products of different lengths that differ from each other by a multiple of three residues. After depletion of the P4 primer, the P3 primer takes over and amplifies the PCR products of different lengths. TP-PCR conditions (95-7' (95-1", 63+0.5/cycle- 1", 72-1') x10, (95- 30", 56- 30", 72-1'+10"/cycle) x25, 72- 10',4C-∞)
Diluted PCR products (1:5) were mixed with 1200LIZ and Hi-Di Formamide, PCR products were size fractionated on a Prism ABI 3500 DNA sequencer (Applied Biosystems).

Data availability

All data are deposited to dbGAP (https://www.ncbi.nlm.nih.gov/gap/) under the accession number phs002206.v3.p1.

Code availability

Only publicly available tools were used in data analysis as described wherever relevant in Methods.

References

1. Cohen, A.S.A. *et al.* Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes. *Genet Med* **24**, 1336-1348 (2022).
2. Petrikin, J.E. *et al.* The NSIGHT1-randomized controlled trial: rapid whole-genome sequencing for accelerated etiologic diagnosis in critically ill infants. *NPJ Genom Med* **3**, 6 (2018).
3. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* **30**, 1291-1305 (2020).
4. Lalonde, E. *et al.* RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res* **21**, 545-54 (2011).
5. Kwan, T. *et al.* Tissue effect on genetic control of transcript isoform variation. *PLoS Genet* **5**, e1000608 (2009).
6. Grundberg, E. *et al.* Population genomics in a disease targeted primary cell model. *Genome Res* **19**, 1942-52 (2009).
7. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-9 (2012).
8. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414 e24 (2016).
9. Cheung, W.A. *et al.* Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. *Genome Biol* **18**, 50 (2017).
10. Grundberg, E. *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet* **93**, 876-90 (2013).
11. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11**, 533-8 (2010).
12. Morcos, L. *et al.* Genome-wide assessment of imprinted expression in human cells. *Genome Biol* **12**, R25 (2011).
13. Light, N. *et al.* Interrogation of allelic chromatin states in human cells by high-density ChIP-genotyping. *Epigenetics* **9**, 1238-51 (2014).
14. Cotton, A.M. *et al.* Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol* **14**, R122 (2013).
15. Busche, S. *et al.* Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation. *Genome Biol* **16**, 290 (2015).
16. Allum, F. *et al.* Dissecting features of epigenetic variants underlying cardiometabolic risk using full-resolution epigenome profiling in regulatory elements. *Nat Commun* **10**, 1209 (2019).
17. Jin, P. & Warren, S.T. Understanding the molecular basis of fragile X syndrome. *Hum Mol Genet* **9**, 901-8 (2000).
18. Barbe, L. *et al.* CpG Methylation, a Parent-of-Origin Effect for Maternal-Biased Transmission of Congenital Myotonic Dystrophy. *Am J Hum Genet* **100**, 488-505 (2017).

19. Monk, D. *et al.* Recommendations for a nomenclature system for reporting methylation aberrations in imprinted domains. *Epigenetics* **13**, 117-121 (2018).
20. Butler, M.G. Imprinting disorders in humans: a review. *Curr Opin Pediatr* **32**, 719-729 (2020).
21. Oussalah, A. *et al.* Epimutations in both the TESK2 and MMACHC promoters in the Epi-cblC inherited disorder of intracellular metabolism of vitamin B12. *Clin Epigenetics* **14**, 52 (2022).
22. Cederbaum, S.D. & Lippe, B.M. Probable autosomal recessive inheritance in a family with Albright's hereditary osteodystrophy and an evaluation of the genetics of the disorder. *Am J Hum Genet* **25**, 638-45 (1973).
23. Temple, I.K. & Shield, J.P. Transient neonatal diabetes, a disorder of imprinting. *J Med Genet* **39**, 872-5 (2002).
24. Schaaf, C.P. *et al.* Truncating mutations of MAGEL2 cause Prader-Willi phenotypes and autism. *Nat Genet* **45**, 1405-8 (2013).
25. Temple, I.K., Cockwell, A., Hassold, T., Pettay, D. & Jacobs, P. Maternal uniparental disomy for chromosome 14. *J Med Genet* **28**, 511-4 (1991).
26. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
27. Fresard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med* **25**, 911-919 (2019).
28. Barbosa, M. *et al.* Identification of rare de novo epigenetic variations in congenital disorders. *Nat Commun* **9**, 2064 (2018).
29. Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. *BioRxiv* (2016).
30. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-2 (2011).
31. Gao, S. *et al.* BS-SNPer: SNP calling in bisulfite-seq data. *Bioinformatics* **31**, 4006-8 (2015).
32. Gudmundsson, S. *et al.* Addendum: The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **597**, E3-E4 (2021).
33. Warner, J.P. *et al.* A general method for the detection of large CAG repeat expansions by fluorescent PCR. *J Med Genet* **33**, 1022-6 (1996).
34. Winnepenninckx, B. *et al.* CGG-repeat expansion in the DIP2B gene is associated with the fragile site FRA12A on chromosome 12q13.1. *Am J Hum Genet* **80**, 221-31 (2007).

Acknowledgment

We would like to thank all families for participating in our study. This work was made possible by the generous gifts to Children's Mercy Research Institute and Genomic Answers for Kids program at Children's Mercy Kansas City. We also would like to thank Nick Nolte, Dan Louiselle and Rebecca Biswell for their work in sample processing; Laura Puckett and Adam Walters for their work in library preparation and sequencing and the GA4K coordination team led by Bradley Belden for their work in clinical coordination. T.P holds the Dee Lyons/Missouri Endowed Chair in Pediatric Genomic Medicine and E.G holds the Roberta D. Harding & William F. Bradley, Jr. Endowed Chair in Genomic Research.

Competing interests

W.J.R., R.J.H., D.M.P., C.T.S., and A.M.W. are current or past employees of Pacific Biosciences. All other authors have no competing interests.

Ethical declarations

The study was approved by the Children's Mercy Institutional Review Board (IRB) (Study # 11120514). Informed written consent was obtained from all participants prior to study inclusion.

Figure Legends

Figure 1. 5-base HiFi-GS in DM1 patient with DMPK repeat. Left panel shows raw reads and variant calls identifying "soft-clipped" reads (i.e. reads with red ends), 45bp deletion and variable sized insertion (4,823bp-5,190) at repeat coordinates. Right panel illustrates the same reads now phased (maternal top, paternal bottom under the dotted line) and with CpG modification staining (blue indicating non-methylated prediction and red indicating methylated prediction from primrose algorithm). With phasing and CpG methylation staining the soft-clipped and insertion (expansion) spanning reads show consistent hypermethylation flanking the disease allele.

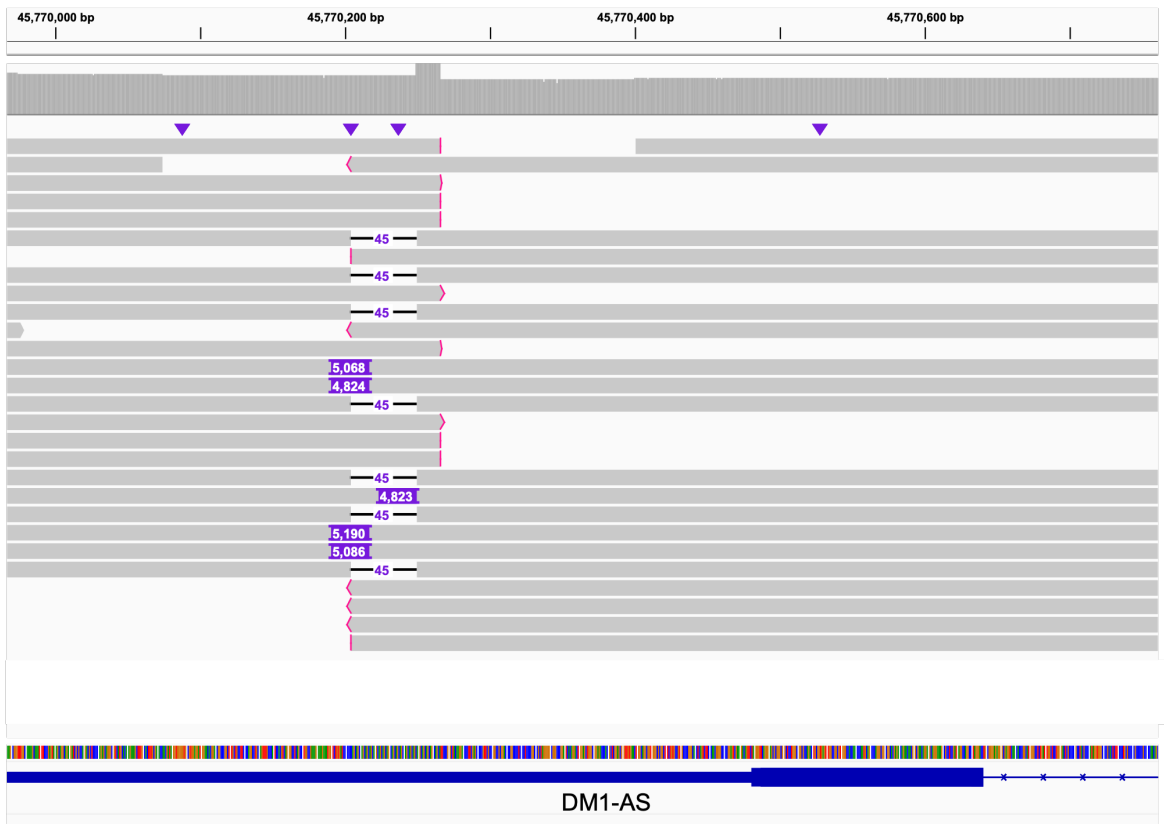
Figure 2. Detection of hypermethylation events in rare disease cohort. Three CpG methylation scores per individual (all reads, Hap1 or Hap2 phased reads) are compared to population mean to extract z-scores. Tiles spanning at least three CpGs with consistently positive z-scores are averaged to extract hyper mCpG-tile z-score, where rare extreme z-scores are tallied for each individual – majority of these scores in 5-base HiFi-GS are stemming from single haplotype.

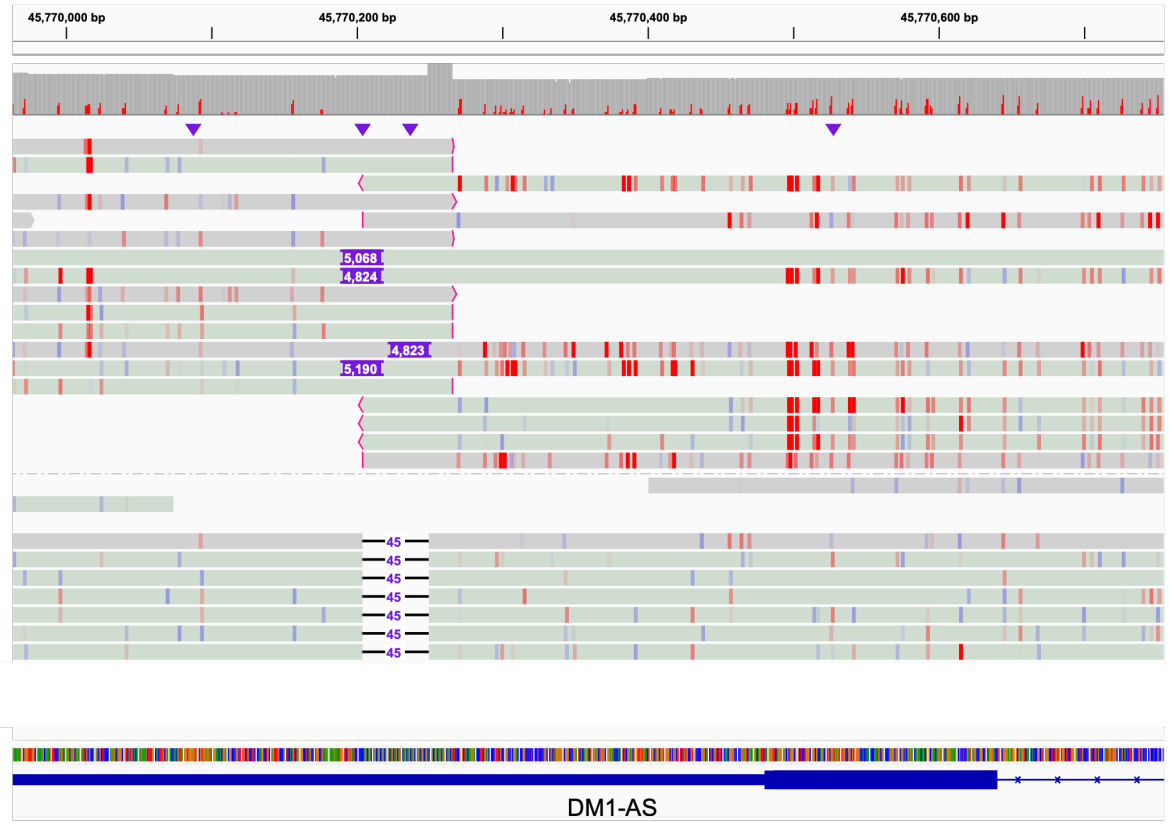
Figure 3. Rare hyper-mCG tiles show enrichment to local *cis*-variation. A) Each rare hyper-mCpG tile was queried for closest rare SNV (<0.5% MAF gnomAD) and distances recorded (x-axis, log10(bp)). B) Same process was repeated for SVs (<1% MAF in HiFi-GS population data), in parallel for each hyper-mCpG tile the distances from all other tile/rare variants were recorded to deem that q value = 0.1 corresponds to ~1000bp for SNVs and 10kb for SVs. The rare SNVs at close proximity (C and D) show significant cis association with rare CpG methylation in same reads as expected.

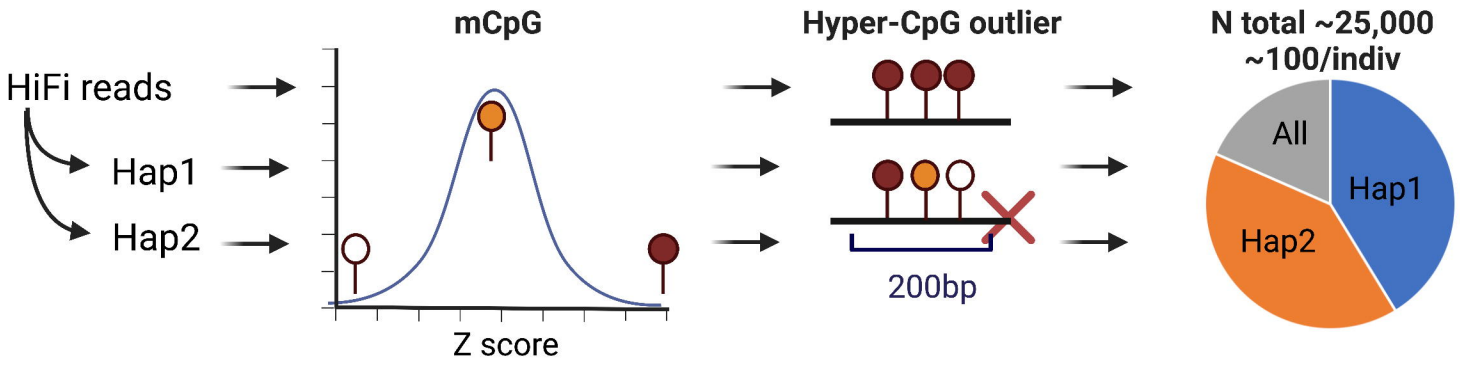
Figure 4. Statistical determination of hyper-mCpG tile landscape. To measure if the marginal extreme CpG methylation window call linked to local variant spans larger region we fetched 5-base HiFi-GS reads overlapping candidate *cis*-variant (here chr10:113,854,859 A to C transversion, boxed just 5' to autosomal recessive disease gene *NHLRC2* shown in track F). Track A shows p-value (-log10P) from 2-by-2 Fisher's exact test for each CpG site methylated/unmethylated state in reads with rare C-allele (track D) versus methylated/unmethylated states from common A-allele (track E) carrying reads. Track B summarizes average methylation level across all rare variant carrying reads and Track C summarizes average methylation level from the population prevalent A-allele. Hypermethylation shows statistical significance for C-allele across ~300bp – relatively typical SNV-linked hyper-mCpG landscape often corresponding to single hypomethylated regulatory element.

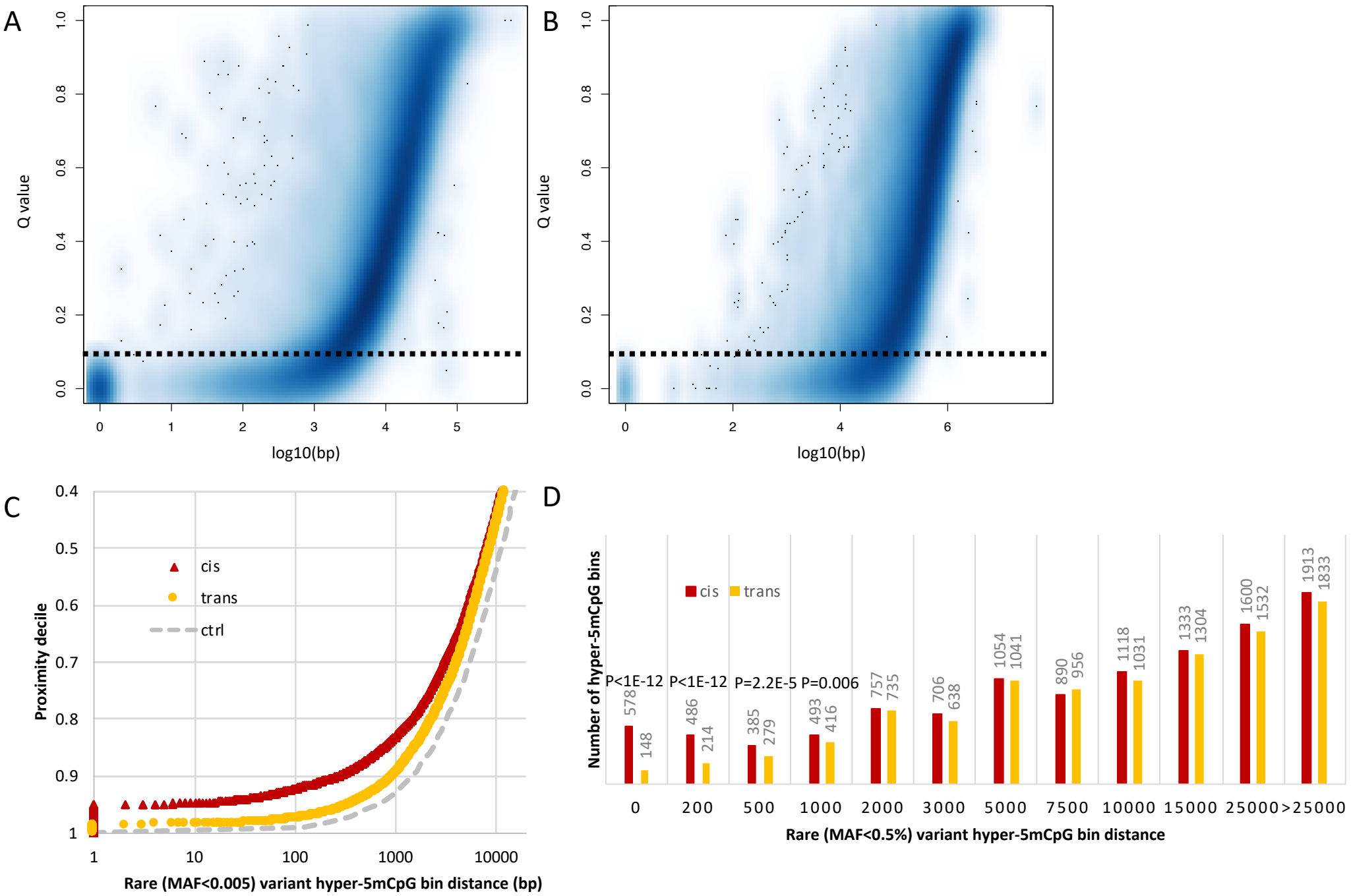
Figure 5. Candidate OMIM gene with rare hypermethylation event. Most 5' proximal SNV or SV linked hypermethylation events in our dataset occurred in autosomal recessive gene promoters. An example shown here where paternally (track E are reads from father) inherited rare 5'UTR SNV is flanked by several hundred base pairs of hypermethylation transmitted to proband (track C), maternal alleles are hypomethylated (tracks D and F). Probands methylation for Hap1 (paternal) and Hap 2 (maternal) are averaged in tracks A and B, respectively. In this case the patient only carries the rare hyper-mCpG causing variant in *MTHFR* – no other rare or functional variant are observed.

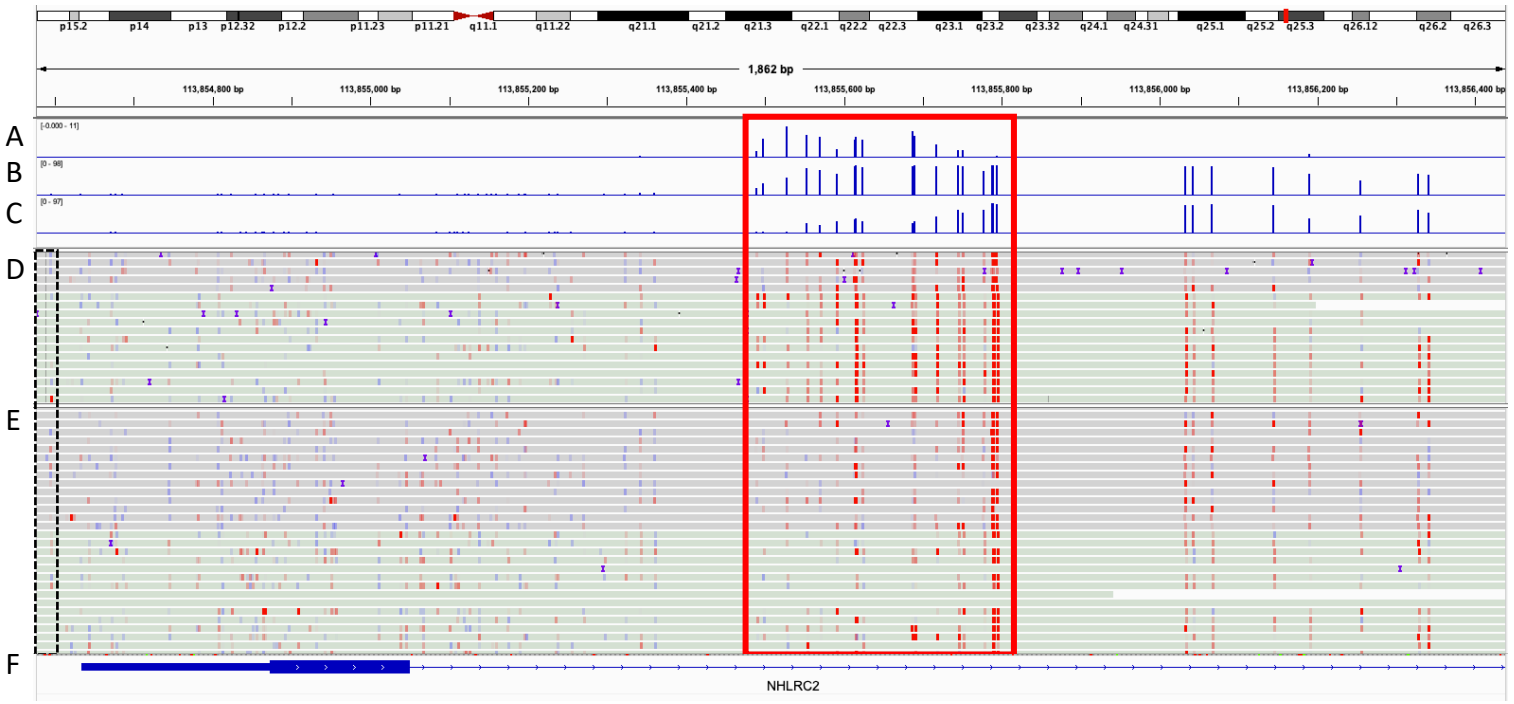
Figure 6. 5-base HiFi-GS detects pathogenic allele in DIP2B causing global developmental delay. Left panel shows haplotype-resolved HiFi-GS reads (maternal top, paternal bottom under the dotted lines) across three individuals detecting large (top two panels), expanded (middle two panels) and no repeats (bottom two panels) at the *DIP2B* locus. Right panel illustrates the same reads now with CpG modification staining (blue indicating non-methylated prediction and red indicating methylated prediction from primrose algorithm). With phasing and methylation staining the large (280-300 repeats) expansion (top two panels) show consistent hypermethylation and becomes diagnostic (pathogenic) resolving previously undiagnosed proband with global developmental delay. The expanded (but not pathogenic) repeat in second unaffected individual (middle two panels) remained hypomethylated (and not diagnostic) like the other unaffected individual without the repeat expansion (bottom two panels).

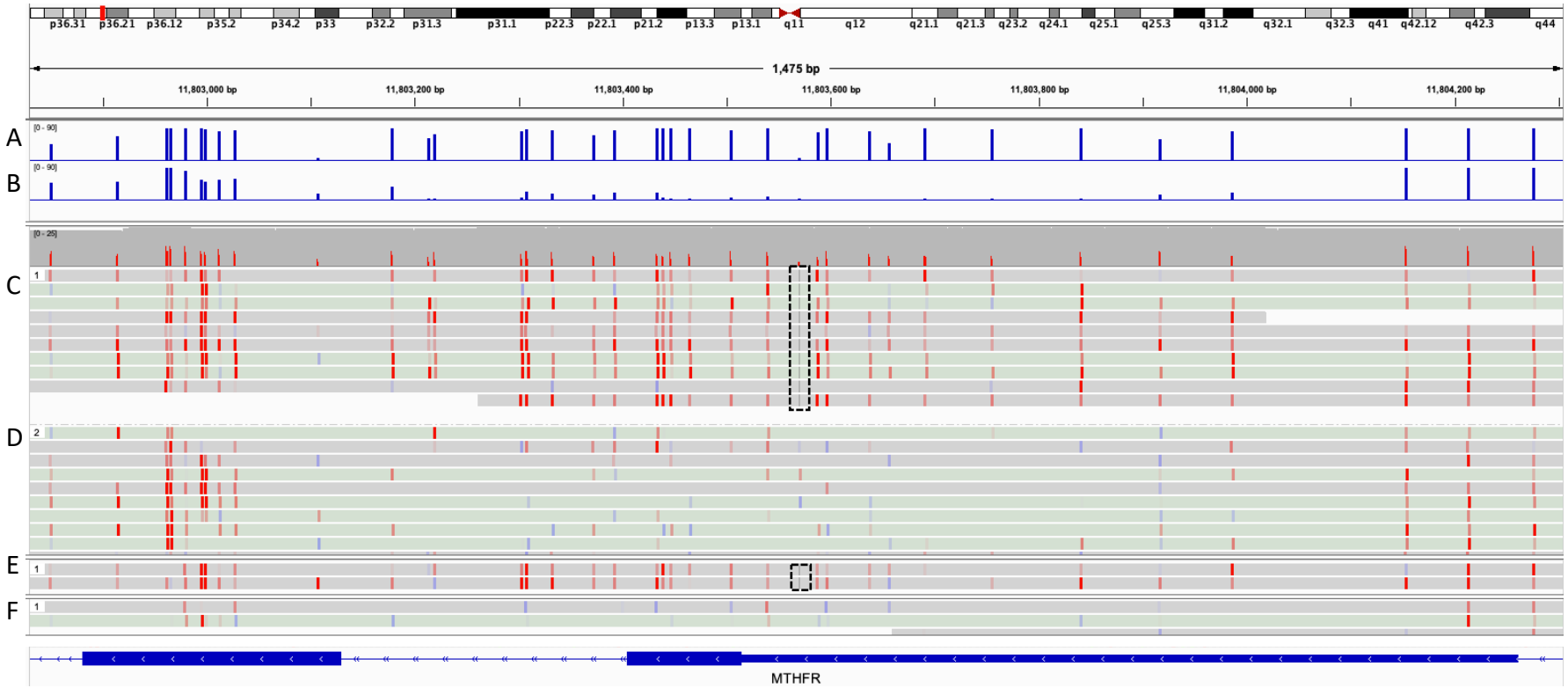

45,770,000 bp
45,770,200 bp
45,770,400 bp
45,770,600 bp
45
5,068
4,824
4,823
5,190
5,086
DM1-AS

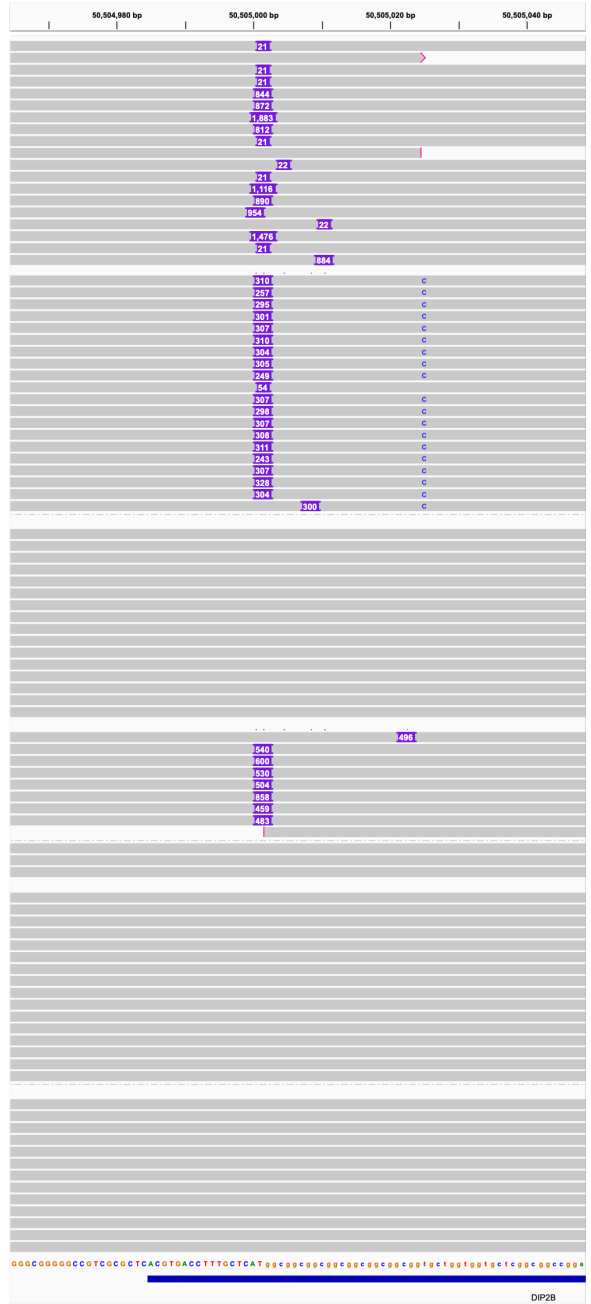

45,770,000 bp
45,770,200 bp
45,770,400 bp
45,770,600 bp
5,068
4,824
4,823
5,190
45
DM1-AS

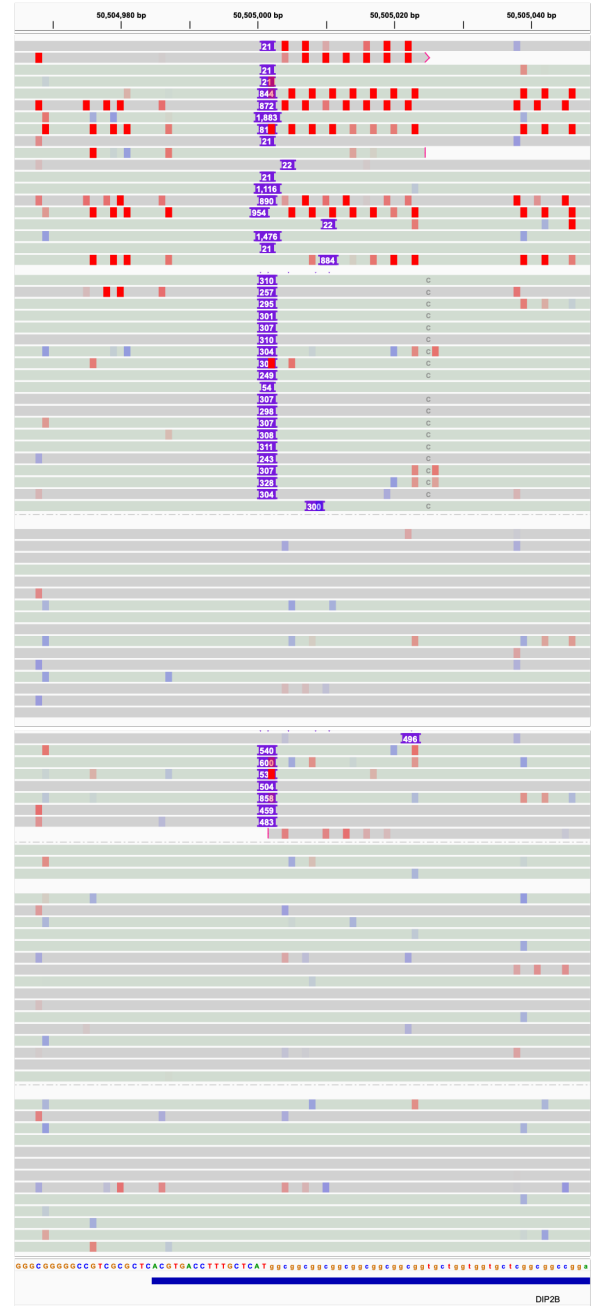

HiFi reads
Hap1
Hap2
mCpG
Z score
Hyper-CpG outlier
200bp
N total ~25,000
~100/indiv
All
Hap1
Hap2


A
Q value
log10(bp)
B
Q value
log10(bp)
C
cis
trans
ctrl
Proximity decile
Rare (MAF<0.005) variant hyper-5mCpG bin distance (bp)
D
cis
trans
Number of hyper-5mCpG bins
P<1E-12
P<1E-12
P=2.2E-5
P=0.006
578
148
486
214
385
279
493
416
757
735
706
638
1054
1041
890
956
1118
1031
1333
1304
1600
1532
1913
1833
0
200
500
1000
2000
3000
5000
7500
10000
15000
25000
>25000
Rare (MAF<0.5%) variant hyper-5mCpG bin distance


p15.2
p14
p13
p12.32
p12.2
p11.23
p11.21
q11.1
q11.22
q21.1
q21.2
q21.3
q22.1
q22.2
q22.3
q23.1
q23.2
q23.32
q24.1
q24.31
q25.1
q25.2
q25.3
q26.12
q26.2
q26.3
1,862 bp
113,854,800 bp
113,855,000 bp
113,855,200 bp
113,855,400 bp
113,855,600 bp
113,855,800 bp
113,856,000 bp
113,856,200 bp
113,856,400 bp
A
B
C
D
E
F
NHLRC2


p36.31 p36.21 p36.12 p35.2 p34.2 p33 p32.2 p31.3 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.3 q41 q42.12 q42.3 q44
1,475 bp
11,803,000 bp
11,803,200 bp
11,803,400 bp
11,803,600 bp
11,803,800 bp
11,804,000 bp
11,804,200 bp
A
B
C
D
E
F
MTHFR


50,504,980 bp
50,505,000 bp
50,505,020 bp
50,505,040 bp
DIP2B


50,504,980 bp
50,505,000 bp
50,505,020 bp
50,505,040 bp
DIP2B