HGG Advances

ARTICLE

Fine mapping regulatory variants by characterizing native CpG methylation with nanopore long-read sequencing

Yijun Tian,[1,5,*] Shannon K. McDonnell,[2] Lang Wu,[3,4] Nicholas B. Larson,[2] and Liang Wang[1,*]

Summary

5-Methylcytosine (5mC) is the most common DNA modification in the human genome. Bisulfite conversion combined with short-read sequencing captures this modification at single-nucleotide resolution but introduces PCR duplication bias and limits co-methylation analysis between distant cytosines. To resolve these limitations, we used nanopore long-read sequencing to profile human methylation and performed long-range co-methylation analysis with native DNA modification information. We analyzed the nanopore demo data in the adaptive sampling sequencing targeting the CpG islands and applied the linkage disequilibrium (LD) R^2 to identified methylation haplotype blocks (MHBs). We found that the cancer genome exhibited significantly smaller MHBs, higher CpG density, and a lower methylation LD R^2 value compared to normal cells. Additionally, we demonstrated the superiority of long-read sequencing in capturing large MHBs compared with short-read sequencing. By profiling the methylation changes near the JASPAR motif and actual chromatin immunoprecipitation sequencing (ChIP-seq) peaks, we also studied the epigenetic changes related to protein binding. Based on adaptive sampling technology, we conducted nanopore sequencing targeting regions with methylation quantitative trait loci (mQTLs) and genome-wide association study (GWAS) risk variants in the 22Rv1 cell line. After analyses, we inspected the closest haplotype-specific methylated region near the variant and identified allele-specific methylated regions with allele-specific accessibility signals in the ATAC-seq data. This study demonstrates the feasibility of nanopore sequencing for methylome profiling while preserving haplotype information, offering an innovative approach to elucidate the epigenetic changes driven by noncoding variants in the human genome.

Introduction

Over the past decades, enormous phenotypical connections have been identified between SNPs and complex disease risk via genome-wide association studies (GWASs).[1,2] To help us understand the underlying biological meaning, expression quantitative trait locus (eQTL) analysis[3,4] has been used to determine variants associated with the RNA expression of susceptible genes. Beyond the productivity of these association findings, challenges arise regarding prioritizing the causal (regulatory) variants and the target genes, which are not easily recognizable due to linkage disequilibrium.[5–8] Methylation quantitative trait locus (mQTL) analysis is more mechanistically oriented, focusing on detecting associations between genetic variation and DNA methylation levels.[9,10] This approach inherently suggests that DNA sequence variations may directly influence the DNA base modification at methylation sites, offering deeper insight into the regulatory role of SNPs. To distinguish between methylated and unmethylated cytosines, bisulfite conversion is commonly used with microarray or next-generation sequencing to profile the feature at the genome-wide scale. However, due to technical limitations, the conversion-based approach presents several challenges for methylation detection and mQTL interpretation. First and foremost, the conversion eliminates allele information for SNPs containing cytosine, which hampers haplotype identification for each sequencing read.

Additionally, when measuring co-methylation profiles, next-generation sequencing (NGS) methods face difficulties in evaluating distant CpG sites, primarily due to read length constraints. Moreover, PCR amplification during NGS library preparation can result in duplicated methylation patterns and introduce biases in 5mC detection. For microarray-based methylation profiling, only 1.5%–3.0% of CpG sites in the human genome are covered,[11] with the data provided as a summarized percentage rather than as precise base modification calls. Importantly, the existing mQTL experimental design depends on associations across samples between genotypes and methylation levels of several million SNPs and CpG sites. This approach generates numerous correlations influenced by genotype allele linkage disequilibrium (LD), leading to mQTL findings that are broadly distributed throughout the genome. Consequently, these

[1]Department of Tumor Microenvironment and Metastasis, Moffitt Cancer Center, Tampa, FL 33612, USA; [2]Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA; [3]Division of Cancer Epidemiology, Population Sciences in the Pacific Program, University of Hawai'i Cancer Center, University of Hawai'i at Mānoa, Honolulu, HI 96813, USA; [4]Department of Interdisciplinary Oncology and Department of Genetics, LSU-LCMC Health Cancer Center, School of Medicine, Louisiana State University Health Sciences Center, New Orleans, LA 70112, USA
[5]Lead contact
*Correspondence: yijun.tian@moffitt.org (Y.T.), liang.wang@moffitt.org (L.W.)
https://doi.org/10.1016/j.xhgg.2025.100532.
© 2025 Published by Elsevier Inc. on behalf of American Society of Human Genetics.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

findings may not be particularly useful for prioritizing functional variants or regulatory elements, especially in robust large-scale studies involving over 1,000 samples. In 4,170 whole-blood samples, Huan et al.[12] identified 4.7 million *cis*-mQTLs affecting 121.6K methylation probes, roughly 25% of the Illumina Infinium HumanMethylation450K array. Bonder et al.[13] found 272,037 independent *cis*-mQTLs covering 34.4% of all 405,709 CpG sites tested. More recently, a meta-analysis[14] using genotype LD clumping identified 248,607 independent *cis*-mQTL associations accounting for 45% of the HumanMethylation450K array. To accurately assess the functionality of an mQTL, a more nuanced understanding of the DNA methylome is needed to better model the proximity between the candidate variant and the differentially methylated CpG sites.

To address these concerns, we aimed to identify a distinct sequencing approach to detect DNA base modifications and genetic variations at the single-molecule level. Ultimately, we decided to leverage nanopore long-read technology to develop a more robust framework to elucidate the allele-specific methylation in the human genome. Using this technology, we directly identified 5mCpG and 5hmCpG base modifications with single-nucleotide resolution over kilobases-long reads. This technology enables PCR-free genome-scale observations of the methylome over unprecedented genomic distances. With the shift from linear regression models to directly comparing differentially methylated regions between haplotypes, the requirements for sample size have become less stringent. Moreover, the continuous 5mC profiling with nanopore technology enhances the *de novo* identification of differentially methylated regions (DMRs), providing more accurate insights into the functionality of target regions and benefitting the interpretation of their underlying biological significance.

In this project, we explored the potential of nanopore long-read technology for profiling the methylome of the human genome and developed a computational pipeline to detect allele-specific methylation using adaptive sampling in prostate cells. We demonstrated the pipeline's effectiveness in identifying functional variants and methylated regions.

Material and methods

Nanopore open data retrieval

Adaptive sampling (AS) is an advanced technique used in nanopore sequencing to selectively enrich or deplete specific regions of interest from a genomic sample in real time. To evaluate the AS performance, we accessed the reduced representation methylation sequencing (RRMS) result from the Oxford Nanopore Open Data S3 bucket (s3://ont-open-data/rrms_2022.07/), which aimed to target 310 Mb of the human genome regions highly enriched for CpG sites, including ~28,000 CpG islands, ~50,600 shores, and ~42,700 shelves as well as ~21,600 promoter regions (https://community.nanoporetech.com/attachments/7599/download). The Binary Alignment Map (BAM) files were retrieved using the Amazon Web Service (AWS) command line interface. The RRMS AS run includes genomic sequencing run from COLO829 (melanoma), COLO829BL (parental lymphoblast), and HCC1395 and HCC1395BL (parental lymphoblast). Since the BAM file included alignment information to the hg38 genome and per-read base modification predictions for the CpG context 5mC call, we directly used it for MONOD2 co-methylation analysis. The reduced representation bisulfite sequencing data were retrieved from the same S3 bucket under the bisulfite directory (s3://ont-open-data/rrms_2022.07/bisulfite/alignment/). These alignments were generated by Diagenode Premium RRBS Kit v.2 from the match sample and deduplicated based on unique molecular identifier (UMI). We directly used the RRBS BAM file for MONOD2 co-methylation analysis.

Co-methylation analysis of nanopore long reads with MONOD2

MONOD2 is a toolkit for co-methylation analysis in bisulfite sequencing data. To fit the nanopore base modification information to MONOD2 requirements, we developed a Perl script (MMMLparse.pl) to generate a bisulfite-emulative BAM file based on the base modification tag. After the *in silico* bisulfite conversion, we generated a methylation haplotype-retrieving Perl script (getHaplo_SE_cgOnly.pl) for single-end reads adapted from the MONOD2 script (getHaplo_PE_cgOnly.pl). After methylation haplotype pattern retrieval, methylation haplotype blocks (MHBs) were defined with the cghap2mhbs.sh MONOD2 script. Since the co-methylation profiling estimation was based on methylation LD, we applied a minimal methylation frequency of 30% and a minimal read depth of 50 for calculating the R^2 to avoid bias caused by low-variability sites or poor read coverage. We used a minimal R^2 of 0.1 for MHB determination. After the MHB determination, we used the "smoothScatter" function from the R graphics package to visualize the distribution of R^2 and the distance for the CpG sites in the same MHB.

5mCpG methylation profiling near transcription factor binding motifs

To characterize the 5mCpG methylation levels surrounding known transcription factor binding sites, we used the modkit pileup function (https://github.com/nanoporetech/modkit) to generate a bedGraph file for each CpG site and then converted them to a bigwig file with the bedGraphToBigWig program (https://www.encodeproject.org/software/bedgraphtobigwig/). We downloaded the *Homo sapiens* (hg38) familial binding site collection from the JASPAR database and generated a control BED file for each motif from the human genome with BEDtools.[15] We then used deeptools[16] to compute and visualize cumulative methylation to each motif and its surrounding genome. From the intermediate signal matrix, we extracted the cumulative methylation percentages from 10 adjacent bins centered on the transcription factor (TF) motif. These values were compared between TF-bound and matched control regions using the Wilcoxon rank test and the Kolmogorov-Smirnov test. p values were subsequently adjusted within each sample using the Benjamini-Hochberg (BH) procedure to control the false discovery rate.

Target design of AS

After evaluating the performance of nanopore RRMS demo data, we aimed to design a customized AS BED file to enrich sequences

at GWAS and mQTL variant loci. We downloaded prostate cancer risk variants reported by the GWAS Catalog (https://www.ebi.ac.uk/gwas/home). We then used LDlink[17] to retrieve variants highly correlated with the tag SNPs (genotype LD $R^2 \geq 0.5$) in any of these populations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR) or South Asian (SAS). As a result, we obtained a total of 68,436 SNPs with prostate cancer risk association.

We also downloaded mQTL summary statistics of the TCGA prostate cancer cohort[18] and filtered the associations by the regression coefficient ($R \geq 0.5$ or ≤ -0.5). As a result, we found 137,825 SNPs with strong mQTL effect. To generate an AS BED file, we expanded 4,000 bp on both sides of each SNP locus and merged the expanded intervals. As an outcome, the BED file covered 250,689,518 bp with 5,569 intervals, roughly 8.36% of the human genome.

AS nanopore sequencing in 22Rv1 prostate cells

The 22Rv1 (RRID: CVCL_1045) cell line was obtained from the ATCC and grown in RPMI 1640 medium supplemented with 10% fetal bovine serum (FBS). All cell lines were verified with short tandem repeat (STR) profiling before use. After harvesting the cells during the log phase, we used the Puregene Cell Kit (QIAGEN, 158043) to extract high-quality genomic DNA with RNase A treatment included. The genomic DNA was sheared into 8 kb fragments with g-TUBE (Covaris, 520079). The sheared DNA was used for PCR-free nanopore sequencing library preparation using the Ligation Sequencing Kit (Oxford Nanopore Technology, SQK-LSK110). After library preparation, 100 fmol library solution was loaded to an R9.4.1 flow cell. With the target BED file and the matched reference FASTA plugged in, we started the data acquisition in MinKNOW and kept it running until the flow cell efficiency was fully depleted. The details of the adaptive sequencing summary can be found in Table S1.

Nanopore sequencing base-calling and read-level phasing

After the sequencing finished, we used dorado basecaller on an NVIDIA graphics processing unit (GPU) to infer DNA sequence and CpG site base modification based on the model dna_r9.4.1_e8_sup@v3.3_5mCG_5hmCG&rdquo. The base-calling step generated hg38 alignments with per-read base modification information for downstream phasing and methylation analysis. To evaluate the enrichment performance of the AS run, we used the "mosdepth" program[19] to summarize the coverage for the genome and the target region. To resolve the SNP calling difficulty in homopolymer regions and improve the accuracy of SNP calling for nanopore long-read data, we applied a pipeline called "PEPPER-Margin-DeepVariant"[20] to provide state-of-the-art variant calling for the 22Rv1 genome. We then used the candidate variants for read-level phasing and haplotype tagging in LongPhase.[21]

Haplotype-specific methylation calling and *de novo* DMR analysis

To perform DMR statistical analysis, we used the strand-specific methylation levels on either haplotype as a *bona fide* replicate observation for each CpG site. More technical, we used the modkit pileup function to call 5mCpG methylation with haplotype partitioning and strand-specific output. The output bedGraph files were used to identify *de novo* DMRs between the two haplotypes in the "metilene" software. To improve sensitivity and fit better with the long-read sequencing data, we increased the "–maxdist" parameter (allowed distance between two CpGs within a DMR) to 1,000 bp and lower the "–mincpgs" parameter (minimum number of CpGs in a DMR) to 5. The metilene program provides the adjusted and unadjusted p value for the Mann-Whitney U test (p-MWU). Considering the limited sample size, we used the unadjusted p-MWU and the methylation difference (delta) to define significant DMRs.

Correlation between chromatin accessibility and CpG methylation

To quantify the chromatin accessibility in the 22Rv1 cell line, we calculated the Tn5 insertion coverage by normalizing the transposase cut event count in each peak to the peak width. We also calculated the average methylation percentage for each peak. The paired profiling was plotted with the "smoothScatter" function from the R graphics package for visualization.

Data accession and source code repository

The source code generated in this study can be accessed through the GitHub repository (https://github.com/Yijun-Tian/Nanopore). The 22Rv1 ATAC-seq data were obtained from the GEO (GEO: GSE264518).[22] The eQTL data of prostate tissue were retrieved from the GTEx Consortium.[23] The prostate cancer mQTL data were downloaded from the Pancan-meQTL database.[18] The base-called AS nanopore sequencing data were submitted to the NCBI BioProject repository (NCBI: PRJNA1158791; https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1158791).

Ethics approval

Human prostate tissue samples were analyzed under protocols approved by the Moffitt Cancer Center IRB (MCC50468).

Results

Long reads extend the scope of co-methylation profiling

The co-methylation analysis intends to study the correlation between different CpG sites to identify genomic regions showing coordinated epigenetic changes. With bisulfite-emulated nanopore data, we used the methylation LD R^2 from the MONOD2 method[24] to measure the non-random association of the methylation status between two CpG sites (Figure 1A), in which an R^2 value close to 1.0 indicates strong co-methylation, while a value near 0 indicates uncorrelated methylation. After defining the MHBs in each RRMS sample, we examined the methylation LD R^2 profiles within these blocks in relation to CpG site distances. In cancer cell lines (COLO829 for melanoma and HCC1395 for breast cancer), the methylation LD R^2 values were predominantly clustered between 0.0 and 0.2, indicating weak coordination of methylation states (Figure 1B). In contrast, the corresponding parental lymphoblastoid cell lines (COLO829BL and HCC1395BL) exhibited a broader distribution of LD R2 values, ranging from 0.1 to 0.6, suggesting more variable or coordinated methylation patterns in these blocks (Figure 1C). We also compared the genomic size and CpG density of the MHBs

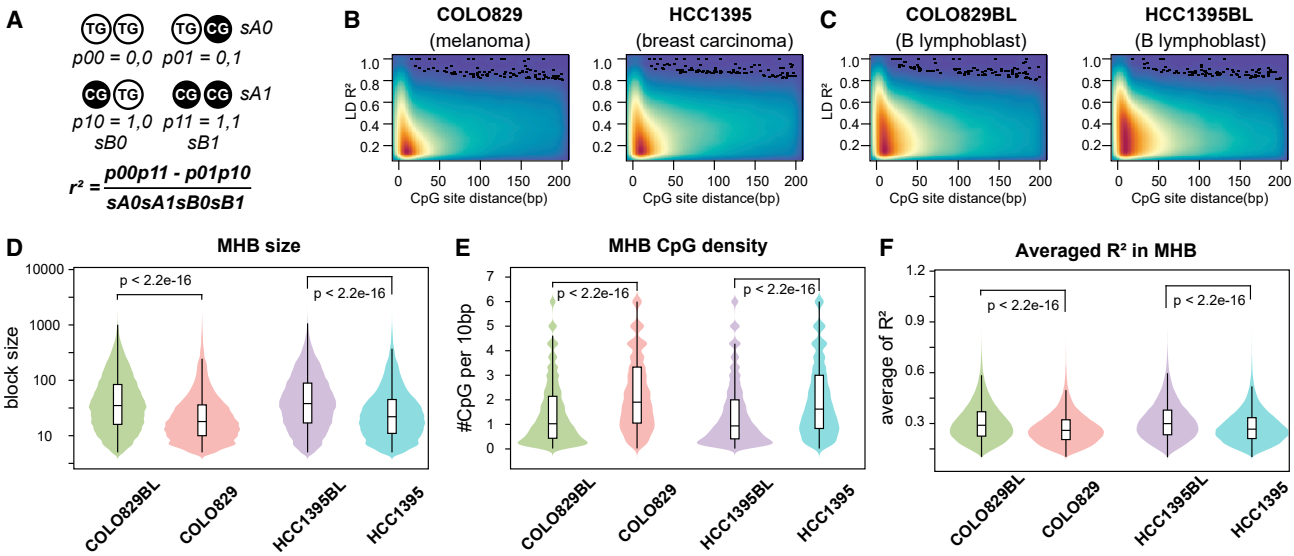

Figure 1. Characterization of CpG methylation haplotype block profiling with RRMS data

(A) Explanation of using LD R^2 calculation for co-methylation analysis.

(B and C) Smooth scatterplot of co-methylation profiles for CpG sites located in the methylation haplotype block (MHB) in cancerous COLO829 and HCC1395 (B) cells and parental normal COLO829BL and HCC1395BL (C) cells.

(D–F) Comparison of MHB metrics, such as block size (D), CpG density (E), and averaged R^2 value (F), in the four RRMS samples. Boxplots display the distribution of values across groups: the horizontal line indicates the median, the box spans the interquartile range (25th–75th percentile), and whiskers extend to values within 1.5 × interquartile range (IQR). Kolmogorov–Smirnov test was used to compare the MHB metric difference.

between the cancer and normal RRMS samples. Cancer cells exhibited significantly smaller MHBs (Figure 1D) and higher CpG density (Figure 1E) within these regions compared to normal cells. Furthermore, consistent with the observed patterns in the density scatterplots, the average methylation LD R^2 values within cancer MHBs were significantly lower than those in normal cells (Figure 1F).

Since the AS demo repository also included reduced representation bisulfite sequencing (RRBS) data for matched samples, we conducted further benchmarking to evaluate the performance of nanopore sequencing in quantifying DNA methylation. Specifically, we assessed the correlation between CpG methylation levels measured by nanopore and RRBS across varying read coverages. We observed that the correlation coefficient increased with read depth and plateaued once coverage reached approximately 20× (Figure S1A), indicating reliable methylation quantification by nanopore sequencing at this threshold. Although CpG-level methylation measurements were consistent between nanopore and bisulfite sequencing, the co-methylation patterns revealed by bisulfite data differed substantially, likely due to limited read length. Interestingly, a subset of CpG sites with strong pairwise correlations disproportionately shaped the co-methylation profiles observed in the RRBS data (Figure S1B). In contrast, this highly correlated CpG population was largely absent in the RRMS-derived profiles (Figure S1C), suggesting that the limited read span of RRBS may artificially inflate localized co-methylation signals. We also analyzed MHB metrics using RRBS data. Although the differences were still statistically significant (Figures S1D–S1F), we highly suspect they can be attributed to the truncated co-methylation metrics from limited fragment length. As a result, we consider these conclusions less robust compared to those derived from long-read data. To better understand the differences between the two technologies, we selected two representative MHBs with sufficient regional coverage in both RRBS and RRMS datasets (Figures S2A and S2B). In the RRMS data, the same region exhibited intact and continuous co-methylation blocks, whereas the RRBS data showed only fragmented and scattered co-methylation signals.

Nanopore native sequencing identifies associations of CpG methylation with human TF binding sites

Leveraging the comprehensive allelic methylation coverage provided by nanopore native sequencing, we profiled CpG methylation surrounding known human TF motifs. When comparing the methylation difference for each motif, we generated a matched control BED file by "bedtools shuffle," which randomly selects control intervals with the same length and contig for each TF binding interval and excludes a 3,000 bp region centered on each TF binding site. With this strategy, we scanned a total of 805 TF binding motifs and identified 57 TFs in COLO829BL, 66 TFs in COLO829, 63 TFs in HCC1395BL, and 76 TFs in HCC1395 cells that showed significantly altered methylation patterns (Wilcoxon false discovery rate [FDR] ≤ 0.05 and absolute mean difference $\geq 10\%$) (Figure 2A). Among these, we highlighted NRF1 (Figure 2B) and RARA (Figure 2C) as representative TFs associated with hypo- and hypermethylation, respectively.

Given the well-established relationship between CTCF binding and DNA methylation, we further investigated

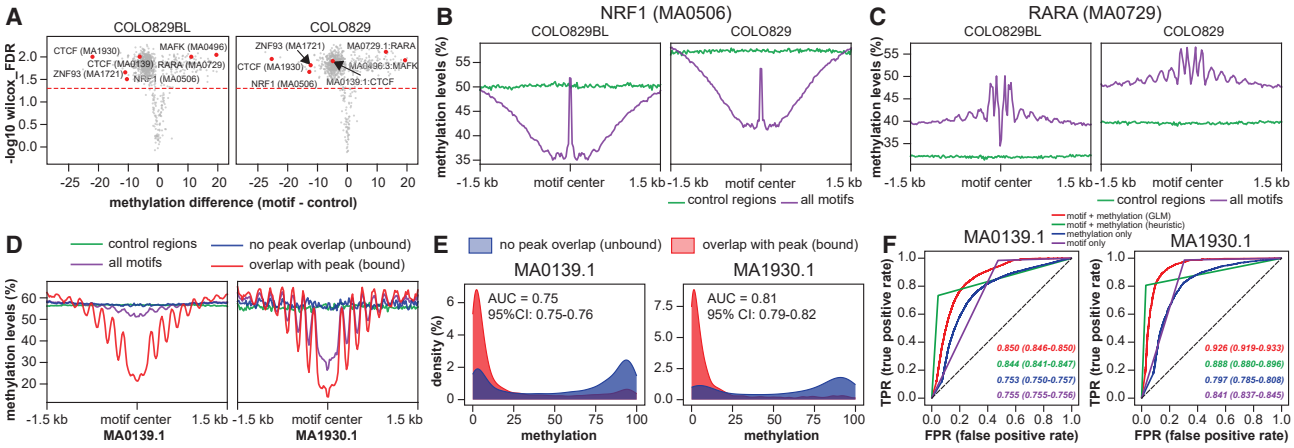

Figure 2. Characterization of CpG methylation profiling with RRMS data on JASPAR TF binding sites

(A) Volcano plots showing the motif and control region methylation difference (10 bins surrounding the motif center, 200 bp) in COLO829BL and COLO829 cells.

(B and C) Averaged CpG methylation levels grouped by NRF1 (B) and RARA (C) factor motifs or respective control regions.

(D) Averaged CpG methylation levels grouped by different CTCF motifs and control regions.

(E) Distribution of the average methylation levels in the CTCF binding motif.

(F) AUC curve of predicting CTCF binding status with methylation levels and/or sequence motifs.

CpG methylation changes surrounding CTCF motifs, stratified by overlap with chromatin immunoprecipitation sequencing (ChIP-seq) peaks. This allowed us to distinguish between bound and unbound CTCF motif instances. Notably, CTCF motifs overlapping ChIP-seq peaks exhibited significantly lower methylation compared to those without overlap (Figure 2D). Since CTCF binding is known to be influenced by methylation within its core motif,[25,26] we examined CpG methylation levels specifically within the CTCF motif to determine their role in modulating binding. We found that bound CTCF motifs (overlapping ChIP-seq peaks) consistently exhibited less than 25% methylation, while unbound motifs displayed a broader range of methylation levels (Figures 2E and S3A). To pinpoint critical CpG sites, we mapped all potential CpG positions within the CTCF motif and assessed their methylation separation power. Although all CpG sites showed significant methylation differences between bound and unbound motifs (Figure S3B), the C2 position displayed the most pronounced separation, with 0% IQR overlap compared to C12, suggesting a decisive role of the C2 position in mediating CTCF binding (Figure S3C). Based on this evidence, we integrated CpG methylation profiles within the CTCF motif with sequence-based motif scores to predict CTCF binding in COLO829 cells. The combined models outperformed models using either methylation or motif information alone, achieving the highest receiver operating characteristic (ROC) area under the curve (AUC) (Figure 2F). To account for the class imbalance in our dataset, we evaluated classifier performance using precision-recall AUC (PR AUC). While ROC AUC values were generally high, PR AUC provided a more stringent measure given the low prevalence of positives. As shown for motif MA0139.1 (Figure S4A), the combined model achieves ROC AUC = 0.850 and PR AUC = 0.352, both outperforming predictions by either methylation or sequence motif individually. Although PR AUC values were much smaller than ROC AUC values, they are 2–9 times greater than the random baseline, confirming that the models greatly enrich for true positives. The same trend has been found for motif MA01930.1 (Figure S4B). These findings demonstrate that incorporating both motif and methylation features substantially improves precision in identifying true positives under class imbalance. This result highlights the potential of using a single DNA sequencing assay to simultaneously capture genetic and epigenetic features for predicting TF binding. Additionally, we observed similar methylation trends of the abovementioned factor binding sites in HCC1395BL and HCC1395 cells (Figures S5A–S5E). The full list of TF methylation profiling can be accessed in Table S1.

Nanopore AS sequencing increases sequence coverage at regions of interest and facilitates methylation haplotype analysis

To selectively observe the DNA methylation patterns related to germline risk variants linking to prostate cancer phenotypes, we examined the GWAS Catalog[27] (https://www.ebi.ac.uk/gwas/home) and TCGA prostate cancer mQTL[18] databases and identified a total of 68,436 SNPs with prostate cancer risk association and 137,825 SNPs with strong mQTL effect (regression coefficient $R \geq 0.5$ or ≤ -0.5) (Figure 3A). Interestingly, only 2,539 SNPs overlapped between the two groups. As expected, compared to genome background, the AS enriched 3.88- and 4.76-fold more sequences at mQTL and GWAS SNP regions, respectively (Figure 3B). Across the 22 autosomal chromosomes, we found a significant negative correlation between the target proportion (proportion of target region length to the belonging chromosome) and the fold of enrichment

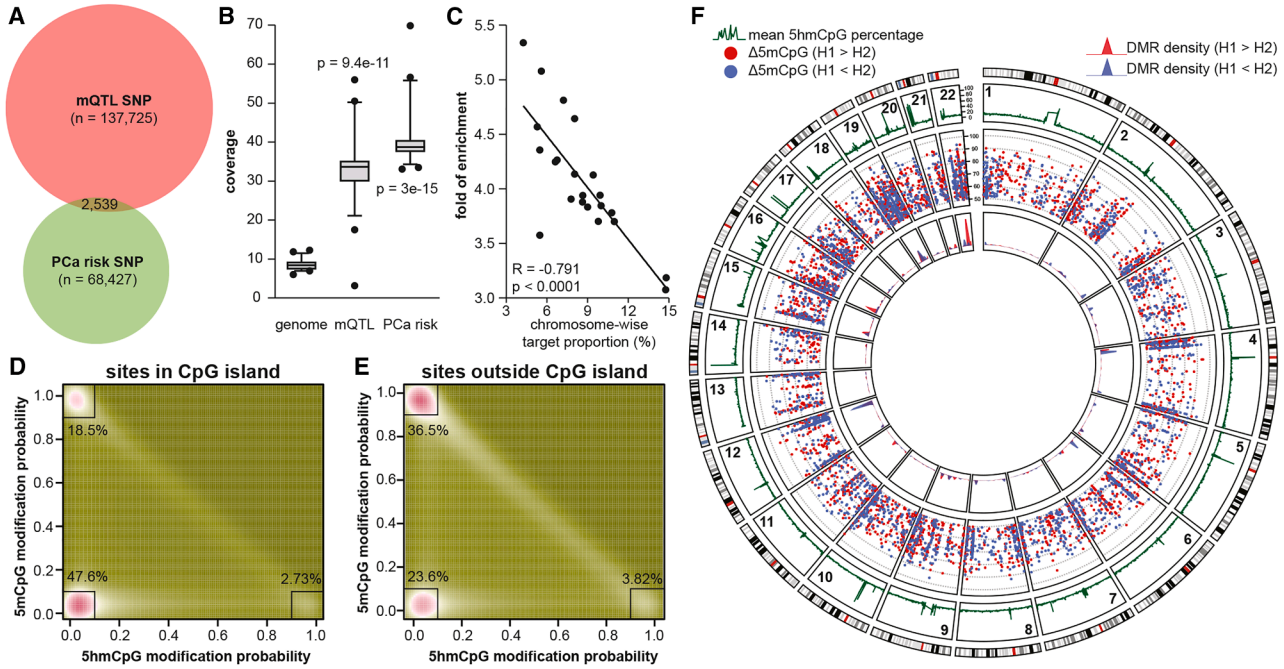

Figure 3. Adaptive sampling nanopore long-read sequencing targeting GWAS and mQTL loci
(A) Proportional Venn diagram of the target SNP with TCGA mQTL signals and GWAS risk significance.
(B) Sequencing coverage summarized by chromosomes with AS run in the 22Rv1 genome.
(C) Correlation between the target proportion and enrichment fold across the 22 autosomes.
(D and E) Distribution of dual methylation calling for the CpG sites in (D) or outside of (E) the CpG island in chromosome 19. The percentages highlight the proportion of highly confident modified sites within the nearby squares.
(F) Circular plot of 5hmC modification percentage (green line, outermost circle), DMR methylation difference (red and blue scatters, middle circle), and the DMR density (red and blue peaks, innermost circle).

(Figure 3C). By grouping the methylation calls by their location, we observed the expected hypomethylation for CpG sites located in CpG islands (Figure 3D) and hypermethylation for CpG sites located outside of CpG islands (Figure 3E). After a PEPPER-Margin-DeepVariant and LongPhase pipeline run, we found 72.8% of heterozygous variants phased (Figure S6A) and 30,725 phase blocks (Figure S6B) for the 22Rv1 genome. These phase blocks include 73.6 variants (Figure S6C) and span 62433.67 bp in the 22Rv1 genome (Figure S6D) on average. To find haplotype-specific methylated regions, we performed *de novo* DMR analysis between the two haplotypes of the 22Rv1 genome, treating the methylation levels reported by each DNA strand as replicates, since most 5mCG exist symmetrically at CpG dinucleotides. This analysis identified 6,390 DMRs that showed haplotype-specific 5mCpG modification in the 22Rv1 genome (p-MWU $\leq 1 \times 10^{-5}$ and delta $\geq 50\%$). Since the differential methylation was identified between the haplotypes, one can visualize those clustered DMRs to identify potential overlapping loci. Using the "genomicDensity" function of the circlize R package,[28] we identified DMR clusters demonstrating extensive haplotype-specific methylation across the genome (Figure 3F). The full list of DMRs can be found in Table S3. Furthermore, we also found that some DMRs are caused by allele-specific methylation, offering strong functional hints for the target variant.

Nanopore native sequencing confirms allele-specific methylation associated with GWAS or mQTL variants

Previously, using bisulfite amplicon sequencing, we reported a functional prostate cancer risk variant causing allele-specific methylation in the 22Rv1 genome.[7] In nanopore sequencing data, we found that the risk SNP rs7247241 (NC_000019.10:38256631:T>C) was adjacent to a 162-bp co-methylated region (chr19:38256705–38256867), with the T allele consistently hypermethylated relative to the C allele (Figure 4A). Additionally, the ATAC-seq result demonstrated that the hypomethylated C allele tended to be more accessible than the hypermethylated T allele (Figure 4A). Although no mQTLs are being reported for this locus, we found that the risk allele (T) was significantly associated with elevated PPP1R14A gene expression (Figure 4B), identified as an oncogene in our previous work. In another locus mapping to the KCNIP3 gene, we found that the genotype of mQTL SNP rs2113417 (NC_000002.12:95334107:G>A) was associated with the methylation level of a 710-bp region (chr2:95333760-95334470) containing this SNP (Figure 4C). This polymorphic region showed strong allele-specific chromatin accessibility (Figure 4C) and a strong eQTL signal in the GTEx prostate cohort (Figure 4D). Interestingly, of the three mQTLs reported for rs2113417, only CpG site cg03595348 is located within

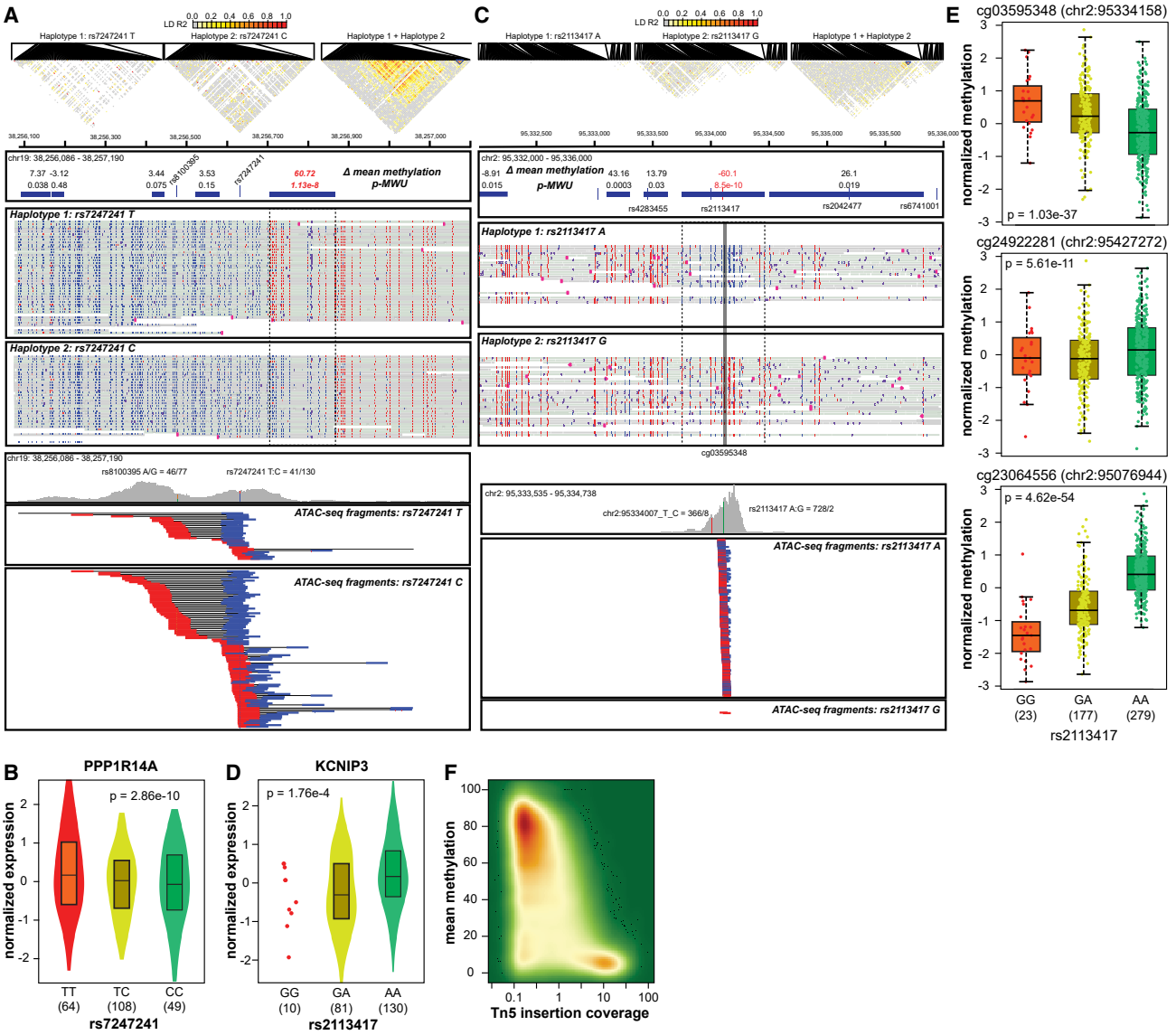

Figure 4. Haplotype specific methylation regions co-localized with allelic accessible region at GWAS and mQTL loci in 22Rv1 cells

(A) Exemplary differential methylated region (DMR) identified near the GWAS risk variant rs7247241 and chromatin accessibility profiling with the rs7247241 alleles in 22Rv1 cells.

(B) GTEx eQTL results for rs7247241 genotype and PPP1R14A expression in prostate tissue.

(C) Exemplary DMR identified near the mQTL variant rs2113417 and chromatin accessibility profiling with the rs2113417 alleles in 22Rv1 cells.

(D) GTEx eQTL results for rs2113417 genotype and KCNIP3 expression in prostate tissue.

(E) TCGA mQTL results for rs2113417 in prostate cancer tissue.

(F) Correlation between Tn5 insertion coverage and methylation percentage in the chromatin-accessible (ATAC-seq peak) regions in 22Rv1 cells.

Boxplots display the distribution of values across groups: the horizontal line indicates the median, the box spans the interquartile range (25th–75th percentile), and whiskers extend to values within 1.5 × IQR. Simple linear regression test was used to compare the gene expression and methylation difference between genotypes.

the same DMR as the SNP, although its *p* value did not show the best (Figure 4E). For the above two SNPs, we also validated the allele-specific methylation pattern by sequencing four prostate tumor tissue samples (Figures S7A and S7B). To describe the relationship between the chromatin accessibility and the methylation level at the genome scale, we calculated the Tn5 insertion coverage and mean methylation percentage for each ATAC-seq peak and found that two profiles separated the peaks into two main clusters, with the major one showing high methylation (~80%) and low chromatin accessibility and the minor one showing low methylation (less than 10%) and high chromatin accessibility (Figure 4F).

Discussion

The technical limitation of bisulfite sequencing has been extensively discussed over the last decade, particularly with regard to its tendency to fragment DNA and introduce PCR amplification bias during sample preparation. In the co-methylation analysis based on whole-genome bisulfite sequencing (WGBS) with short-read NGS, Guo et al.[24] reported that nearly 90% of CpG sites within the same MHB exhibit strong methylation LD ($R^2 \geq 0.9$) regardless of the cell type.

However, in our analysis using both PCR-free nanopore long-read sequencing and UMI-deduplicated RRBS data, we did not observe such a high frequency of strongly co-methylated CpG pairs. One possible explanation is that PCR may over-amplify sporadically co-methylated molecules and artificially inflating methylation LD. Another possibility is that the relatively short read lengths of NGS limit the observation window to small MHBs, often composed of nearby CpG sites, which may lead to an overestimation of genome-wide co-methylation.

From a statistical perspective, population-level metrics such as R^2 can be influenced by several confounding factors, including minimum read depth and the methylation allele frequencies of the two CpG sites being evaluated. As discussed in previous studies,[29–31] R^2 values are sensitive to the allele frequency distribution and can yield artificially high LD (e.g., $R^2 = 1.0$) when both CpG sites exhibit very high or very low methylation frequencies. To mitigate these effects and ensure robust co-methylation estimates, we applied a minimum read depth threshold of 50 and required a minor allele frequency $\geq$0.3 at each CpG site when calculating co-methylation R^2 values.

Our analysis of methylation profiling on binding motifs provided a comprehensive epigenetic landscape for known human TFs. The observation of methylation changes at the CTCF motif is consistent with previous studies,[32,33] suggesting that CTCF is a major factor in directing localized DNA hypermethylation. With integrated DNA methylation profiling and sequence-based motif prediction, a single nanopore DNA sequencing run is expected to accurately predict 80%–90% of true CTCF binding events. Other TFs reported to be sensitive to or associated with DNA methylation, such as NRF1[34] and RARA,[35] also demonstrated consistent adjacent methylation changes in our analysis. In our supplemental analysis, we further examined the methylation levels of C2 and C12 sites in the core CTCF motif. Notably, methylation at C2 was far more discriminative than at C12, aligning with previous findings,[25,26] which showed that C2 methylation reduces CTCF binding affinity by approximately 23-fold, whereas C12 methylation does not significantly affect binding.

With the increasing availability of powerful GPU resources, an increasing number of applications use nanopore AS to enrich targets of interest or deplete unwanted genomes.[36,37] The computational approach utilizes GPU-based fast base-calling to perform raw electrical signal mapping and make real-time "read until" decisions to select specific DNA molecules. Here, we successfully applied the AS approach to enrich germline variant regions, aiming to obtain high PCR-free coverage to identify allele-associated epigenetic changes. However, due to the high error rate in nanopore reads, the alignments are frequently filled with small insertions or deletions (indels), especially in the homopolymer regions, which poses significant challenges for the subsequent variant calling and haplotype phasing analysis. To this end, we applied a neural network tool called DeepVariant[20,38] to polish the candidate variants to achieve the best calling accuracy. Based on the continuity nature of the CpG methylome, we implemented a *de novo* DMR selection procedure to identify clusters of CpG sites that showed modification differences between the two haplotypes. Considering that the haplotype-specific DMRs do not inform the potential causal variants, we included additional information of the closest phasing SNP to each DMR and also annotated whether the SNP is a GWAS or mQTL variant. Under the assumption that a putative functional SNP could change its nearby genome methylation profiling, we anticipate this methylation-based annotation will help uncover the SNP functionality from previous epidemiology and QTL findings.

From our findings, we provided two candidate variants with known or potential functions in prostate cells. For example, the risk variant rs7247241 has been found to be associated with nearby genome methylation in bisulfite sequencing, and the association is progressively stronger in prostate cancer cells than in normal tissue.[7] For another SNP, rs2113417, we successfully applied the nanopore data to discover the most related allelic DMR, which further revealed that the biological origin of the corresponding mQTL signals was from the CpG site cg03595348. This result suggests that the haplotype-specific DMR approach may help remove the mQTL redundancy and improve causal association discovery. Additionally, considering the minor allele frequency of rs2113417, the nanopore DMR approach also provides valuable information to evaluate the epigenetic effect driven by those rare alleles. One shortcoming of applying this approach to a population cohort is that the read-level variant calling and phasing are currently limited to each diploid individual, which will generate different haplotype phasing sets for different persons. This could challenge the between-haplotype DMR identification. To expand this approach to a multiple-sample design, a genotype-based reference would be helpful to further harmonize and standardize the haplotype information so all samples can be compared and analyzed together.

The current allele-specific methylation detection relies on haplotype identification through variant calling, phasing, and haplotype tagging steps for read group classifications. In each of these steps, we aimed to use

applications with the most cutting-edge initiatives and the best adaptability in order to provide optimal accuracy and compatibility for the current study. For instance, in the variant calling procedure, we chose to apply the PEPPER-Margin-DeepVariant pipeline[20] for the benefit of adjusting systematic errors in homopolymer regions. For phasing and haplotype tagging, we selected LongPhase software[21] for better computational efficiency and compatibility with large structural variations. Notably, specialized tools such as MethPhaser[39] have been developed to utilize methylation signals from Oxford Nanopore Technologies to extend SNP-based phasing. Moreover, Akbari et al.[40] combined single-cell template strand sequencing (Strand-seq) with imprinted DMR (iDMR) profiling to achieve chromosome-scale haplotype phasing and parent-of-origin assignment, underscoring the value in enhancing phasing accuracy and parental origin resolution. Most importantly, during our inspections, we noted that putative functional variants could directly influence their nearby genome methylation and therefore contribute to proximal differential methylation regions in the read-length scale for functionality inference. Based on these observations, we would envision an approach to bypass the haplotype phasing procedures to directly compare the methylation difference between reads carrying the reference and alternative allele for a potential regulatory domain related to a candidate SNP.

Taking together, we spotlighted multiple advantages of the nanopore long-read sequencing for methylation analysis and envisioned an innovative concept with this technology to address the technical issues in determining functional allele-specific methylation. The analytical framework offers a unique perspective and uncovers previously unknown insights in the field of genetics.

Data and code availability

The datasets generated during this study are available under BioProject accession PRJNA1158791 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1158791). The code used for data analysis during this study is available at the GitHub repository (https://github.com/Yijun-Tian/Nanopore).

Acknowledgments

This study was supported by the National Institutes of Health (R01CA250018 and R01CA212097 to L. Wang and R01CA263494 to L. Wu). The funders had no role in study design, data collection and analysis, publication decision, or manuscript preparation.

Author contributions

Conception and design, Y.T., S.M., N.L., L. Wu, and L. Wang; development of methodology, Y.T., S.M., N.L., L. Wu, and L. Wang; acquisition of data, Y.T.; analysis and interpretation of data, Y.T., L. Wu, S.M., and N.L.; writing, review, and revision of the manuscript, Y.T., S.M., N.L., L. Wu, and L. Wang; study supervision, Y.T., N.L., L. Wu, and L. Wang.

Declaration of interests

L. Wu has provided consulting services to Pupil Bio Inc. and reviewed manuscripts for Gastroenterology Report, not related to this study, and received honoraria.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2025.100532.

Web resources

bedGraphTo, https://www.encodeproject.org/software/bedgraphtobigwig/
Code repository, https://github.com/Yijun-Tian/Nanopore
Data repository, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1158791
GWAS Catalog, https://www.ebi.ac.uk/gwas/home
MODKIT, https://github.com/nanoporetech/modkit
RRMS, https://community.nanoporetech.com/attachments/7599/download

Received: July 2, 2025
Accepted: October 13, 2025

References

1. Laitinen, V.H., Rantapero, T., Fischer, D., Vuorinen, E.M., Tammela, T.L.J., PRACTICAL Consortium, Wahlfors, T., and Schleutker, J. (2015). Fine-mapping the 2q37 and 17q11.2-q22 loci for novel genes and sequence variants associated with a genetic predisposition to prostate cancer. Int. J. Cancer *136*, 2316–2327. https://doi.org/10.1002/ijc.29276.
2. Gu, F., Zhang, H., Hyland, P.L., Berndt, S., Gapstur, S.M., Wheeler, W., Ellipse Consortium, T., Amos, C.I., Bezieau, S., Bickeböller, H., et al. (2017). Inherited variation in circadian rhythm genes and risks of prostate cancer and three other cancer sites in combined cancer consortia. Int. J. Cancer *141*, 1794–1802. https://doi.org/10.1002/ijc.30883.
3. Liu, D., Zhu, J., Zhou, D., Nikas, E.G., Mitanis, N.T., Sun, Y., Wu, C., Mancuso, N., Cox, N.J., Wang, L., et al. (2022). A transcriptome-wide association study identifies novel candidate susceptibility genes for prostate cancer risk. Int. J. Cancer *150*, 80–90. https://doi.org/10.1002/ijc.33808.
4. Macauda, A., Piredda, C., Clay-Gilmour, A.I., Sainz, J., Buda, G., Markiewicz, M., Barington, T., Ziv, E., Hildebrandt, M.A.T., Belachew, A.A., et al. (2021). Expression quantitative trait loci of genes predicting outcome are associated with survival of multiple myeloma patients. Int. J. Cancer *149*, 327–336. https://doi.org/10.1002/ijc.33547.
5. Nelson, S., Carrick, D., Daee, D., Fingerman, I., and Gillanders, E. (2024). Variation to Biology: Optimizing Functional Analysis of Cancer Risk Variants. J. Natl. Cancer Inst. *116*, 1882–1889. https://doi.org/10.1093/jnci/djae173.
6. Tian, Y., Dong, D., Wang, Z., Wu, L., Park, J.Y., Wei, G.H., Wang, L.; and PRACTICAL/ELLIPSE consortium (2023). Combined CRISPRi and proteomics screening reveal a cohesin-CTCF-bound allele contributing to increased expression of

RUVBL1 and prostate cancer progression. Preprint at bioRxiv. https://doi.org/10.1101/2023.01.18.524405.
7. Tian, Y., Soupir, A., Liu, Q., Wu, L., Huang, C.C., Park, J.Y., and Wang, L. (2022). Novel role of prostate cancer risk variant rs7247241 on PPP1R14A isoform transition through allelic TF binding and CpG methylation. Hum. Mol. Genet. *31*, 1610–1621. https://doi.org/10.1093/hmg/ddab347.
8. IGVF Consortium (2024). Deciphering the impact of genomic variation on function. Nature *633*, 47–57. https://doi.org/10.1038/s41586-024-07510-0.
9. Corradi, C., Lencioni, G., Felici, A., Rizzato, C., Gentiluomo, M., Ermini, S., Archibugi, L., Mickevicius, A., Lucchesi, M., Malecka-Wojciesko, E., et al. (2024). Potential association between PSCA rs2976395 functional variant and pancreatic cancer risk. Int. J. Cancer *155*, 1432–1442. https://doi.org/10.1002/ijc.35046.
10. Kloth, M., Goering, W., Ribarska, T., Arsov, C., Sorensen, K.D., and Schulz, W.A. (2012). The SNP rs6441224 influences transcriptional activity and prognostically relevant hypermethylation of RARRES1 in prostate cancer. Int. J. Cancer *131*, E897–E904. https://doi.org/10.1002/ijc.27628.
11. Fennell, L.J., Hartel, G., McKeone, D.M., Bond, C.E., Kane, A., Leggett, B.A., Patch, A.M., and Whitehall, V.L.J. (2022). Comparative analysis of Illumina Mouse Methylation BeadChip and reduced-representation bisulfite sequencing for routine DNA methylation analysis. Cell Rep. Methods *2*, 100323. https://doi.org/10.1016/j.crmeth.2022.100323.
12. Huan, T., Joehanes, R., Song, C., Peng, F., Guo, Y., Mendelson, M., Yao, C., Liu, C., Ma, J., Richard, M., et al. (2019). Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. Nat. Commun. *10*, 4267. https://doi.org/10.1038/s41467-019-12228-z.
13. Bonder, M.J., Luijk, R., Zhernakova, D.V., Moed, M., Deelen, P., Vermaat, M., van Iterson, M., van Dijk, F., van Galen, M., Bot, J., et al. (2017). Disease variants alter transcription factor levels and methylation of their binding sites. Nat. Genet. *49*, 131–138. https://doi.org/10.1038/ng.3721.
14. Min, J.L., Hemani, G., Hannon, E., Dekkers, K.F., Castillo-Fernandez, J., Luijk, R., Carnero-Montoro, E., Lawson, D.J., Burrows, K., Suderman, M., et al. (2021). Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nat. Genet. *53*, 1311–1321. https://doi.org/10.1038/s41588-021-00923-x.
15. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.
16. Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165. https://doi.org/10.1093/nar/gkw257.
17. Machiela, M.J., and Chanock, S.J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics *31*, 3555–3557. https://doi.org/10.1093/bioinformatics/btv402.
18. Gong, J., Wan, H., Mei, S., Ruan, H., Zhang, Z., Liu, C., Guo, A.Y., Diao, L., Miao, X., and Han, L. (2019). PancanmeQTL: a database to systematically evaluate the effects of genetic variants on methylation in human cancer. Nucleic Acids Res. *47*, D1066–D1072. https://doi.org/10.1093/nar/gky814.
19. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868. https://doi.org/10.1093/bioinformatics/btx699.
20. Shafin, K., Pesout, T., Chang, P.C., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Kolmogorov, M., Eizenga, J.M., Miga, K.H., et al. (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat. Methods *18*, 1322–1332. https://doi.org/10.1038/s41592-021-01299-w.
21. Lin, J.H., Chen, L.C., Yu, S.C., and Huang, Y.T. (2022). LongPhase: an ultra-fast chromosome-scale phasing algorithm for small and large variants. Bioinformatics *38*, 1816–1822. https://doi.org/10.1093/bioinformatics/btac058.
22. Tian, Y., Wu, L., Huang, C.C., and Wang, L. (2024). Identify Regulatory eQTLs by Multiome Sequencing in Prostate Single Cells. Preprint at bioRxiv. https://doi.org/10.1101/2024.06.19.599704.
23. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.
24. Guo, S., Diep, D., Plongthongkum, N., Fung, H.L., Zhang, K., and Zhang, K. (2017). Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat. Genet. *49*, 635–642. https://doi.org/10.1038/ng.3805.
25. Monteagudo-Sanchez, A., Noordermeer, D., and Greenberg, M.V.C. (2024). The impact of DNA methylation on CTCF-mediated 3D genome organization. Nat. Struct. Mol. Biol. *31*, 404–412. https://doi.org/10.1038/s41594-024-01241-6.
26. Hashimoto, H., Wang, D., Horton, J.R., Zhang, X., Corces, V.G., and Cheng, X. (2017). Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. Mol. Cell *66*, 711–720.e3. https://doi.org/10.1016/j.molcel.2017.05.004.
27. Cerezo, M., Sollis, E., Ji, Y., Lewis, E., Abid, A., Bircan, K.O., Hall, P., Hayhurst, J., John, S., Mosaku, A., et al. (2025). The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. Nucleic Acids Res. *53*, D998–D1005. https://doi.org/10.1093/nar/gkae1070.
28. Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. Bioinformatics *30*, 2811–2812. https://doi.org/10.1093/bioinformatics/btu393.
29. VanLiere, J.M., and Rosenberg, N.A. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. Theor. Popul. Biol. *74*, 130–137. https://doi.org/10.1016/j.tpb.2008.05.006.
30. Eberle, M.A., Rieder, M.J., Kruglyak, L., and Nickerson, D.A. (2006). Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. PLoS Genet. *2*, e142. https://doi.org/10.1371/journal.pgen.0020142.
31. Wray, N.R. (2005). Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. Twin Res. Hum. Genet. *8*, 87–94. https://doi.org/10.1375/1832427053738827.
32. Damaschke, N.A., Gawdzik, J., Avilla, M., Yang, B., Svaren, J., Roopra, A., Luo, J.H., Yu, Y.P., Keles, S., and Jarrard, D.F.

(2020). CTCF loss mediates unique DNA hypermethylation landscapes in human cancers. Clin. Epigenet. *12*, 80. https://doi.org/10.1186/s13148-020-00869-7.

33. Butcher, D.T., Mancini-DiNardo, D.N., Archer, T.K., and Rodenhiser, D.I. (2004). DNA binding sites for putative methylation boundaries in the unmethylated region of the BRCA1 promoter. Int. J. Cancer *111*, 669–678. https://doi.org/10.1002/ijc.20324.
34. Domcke, S., Bardet, A.F., Adrian Ginno, P., Hartl, D., Burger, L., and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature *528*, 575–579. https://doi.org/10.1038/nature16462.
35. Hassan, H.M., Kolendowski, B., Isovic, M., Bose, K., Dranse, H. J., Sampaio, A.V., Underhill, T.M., and Torchia, J. (2017). Regulation of Active DNA Demethylation through RAR-Mediated Recruitment of a TET/TDG Complex. Cell Rep. *19*, 1685–1697. https://doi.org/10.1016/j.celrep.2017.05.007.
36. Kovaka, S., Fan, Y., Ni, B., Timp, W., and Schatz, M.C. (2021). Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nat. Biotechnol. *39*, 431–441. https://doi.org/10.1038/s41587-020-0731-9.
37. Loose, M., Malla, S., and Stout, M. (2016). Real-time selective sequencing using nanopore technology. Nat. Methods *13*, 751–754. https://doi.org/10.1038/nmeth.3930.
38. Poplin, R., Chang, P.C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. *36*, 983–987. https://doi.org/10.1038/nbt.4235.
39. Fu, Y., Aganezov, S., Mahmoud, M., Beaulaurier, J., Juul, S., Treangen, T.J., and Sedlazeck, F.J. (2024). MethPhaser: methylation-based long-read haplotype phasing of human genomes. Nat. Commun. *15*, 5327. https://doi.org/10.1038/s41467-024-49588-0.
40. Akbari, V., Hanlon, V.C.T., O'Neill, K., Lefebvre, L., Schrader, K.A., Lansdorp, P.M., and Jones, S.J.M. (2023). Parent-of-origin detection and chromosome-scale haplotyping using long-read DNA methylation sequencing and Strand-seq. Cell Genom. *3*, 100233. https://doi.org/10.1016/j.xgen.2022.100233.