Fine Mapping Regulatory Variants by Characterizing Native CpG Methylation with Nanopore Long-Read Sequencing

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

1. Department of Tumor Microenvironment and Metastasis, Moffitt Cancer Center, Tampa, FL 33612, United States
2. Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, United States
3. Division of Cancer Epidemiology, Population Sciences in the Pacific Program, University of Hawaiʻi Cancer Center, University of Hawaiʻi at Mānoa, Honolulu, HI 96813, United States

*__Correspondence to:__

Liang Wang, Department of Tumor Microenvironment and Metastasis, H. Lee Moffitt Cancer Center & Research Institute, Vincent Stabile Research Building 22403, 12902 Magnolia Drive, Tampa 33612, USA. Tel: 813-745-4955 Fax: 813-745-6606 Email: Liang.Wang@moffitt.org

Summary

5-methylcytosine (5mC) is the most common chemical modification occurring on the CpG sites across the human genome. Bisulfite conversion combined with short-read whole genome sequencing can capture and quantify the modification at single nucleotide resolution. However, the PCR amplification process could lead to duplicative methylation patterns and introduce 5mC detection bias. Additionally, the limited read length also restricts co-methylation analysis between distant CpG sites. The bisulfite conversion process presents a significant challenge for detecting variant-specific methylation due to the destruction of allele information in the sequencing reads. To address these issues, we sought to characterize the human methylation profiling with the nanopore long-read sequencing, aiming to demonstrate its potential for long-range co-methylation analysis with native modification call and intact allele information retained. In this regard, we first analyzed the nanopore demo data in the adaptive sampling sequencing run targeting all human CpG islands. We applied the linkage disequilibrium (LD) R^2 to calculate the co-methylation in nanopore data, and further identified 27,875, 50,481, 26,542 and 51,189 methylation haplotype blocks (MHB) in COLO829, COLO829BL, HCC1395 and HCC1395BL cell lines, respectively. Interestingly, while we found that majority of the co-methylation were in a short range (≤200bp), a small portion (1~3%) showed long distance (≥1,000bp), suggesting potential remote regulatory mechanisms across the genome. To further characterize the epigenetic changes related to transcription factor binding, we profiled the 5mC percentage changes surrounding various motif sites in JASPAR collection and found that CTCF and KLF5 binding sites showed reduced methylation, while FOXE1 and ZNF354A sites showed increased methylation. To further investigate the allele-specific 5mCG in the prostate genome, we designed a target region covering methylation quantitative trait loci (mQTL) and genome-wide association study (GWAS) risk germline variants and generated long reads with adaptive sampling run in the 22Rv1 cell line. To identify the allele-specific methylation in the 22Rv1 cell line, we performed long-read based phasing and compared the 5mCG signals between the two haplotypes. As a result, we identified 6,390 haplotype-specific methylated regions in the 22Rv1 cell line (p-MWU ≤ 1e-5 and delta ≥ 50%). By examining haplotype-specific methylated regions near the phasing variants, we identified examples of allele-specific methylated regions that showed allele-specific accessibility in the ATAC-seq data. By further integrating the ATAC-seq data of 22Rv1, we found that methylation levels were negatively correlated with chromatin accessibility at the genome-wide scale. Our study has revealed native methylome profiling while preserving haplotype information, offering a novel approach to uncovering the regulatory mechanisms of the human prostate genome.

Introduction

Over the past decades, enormous phenotypical connections have been identified between single nucleotide polymorphism (SNP) and complex disease risk in the Genome-Wide Association Study (GWAS). To help understand the underpinning biological meaning, the expression Quantitative Trait Locus (eQTL) analysis had 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 the linkage disequilibrium (Consortium, 2024; Nelson et al., 2024; Tian et al., 2023; Tian et al., 2022). The methylation Quantitative Trait Locus (mQTL) analysis is more mechanistically oriented, focusing on detecting associations between genetic variation and DNA methylation levels. This approach inherently suggests that DNA sequence variations may directly influence the DNA base modification of methylation sites, offering deeper insight into the regulatory role of SNPs. To distinguish between methylated and unmethylated cytosines, the bisulfite conversion is commonly used with microarray or next-generation sequencing to profile the feature at the genome-wide scale. However, due to the 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% to 3.0% of CpG sites in the human genome are covered, 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 the genotypes of millions of SNPs and the methylation levels of thousands of CpG sites. This approach generates numerous correlations influenced by allele linkage disequilibrium (LD), leading to mQTL findings that are broadly distributed throughout the genome. Consequently, these 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. (Huan et al., 2019) identified 4.7 million cis-mQTLs affecting 121.6K methylation probes, roughly 25% of Illumina Infinium HumanMethylation450K array. Bonder et al. (Bonder et al., 2017) found 272,037 independent cis-mQTLs covering 34.4% of all 405,709 CpG sites tested. More recently, a meta-analysis (Min et al., 2021) using linkage disequilibrium (LD) clumping identified 248,607 independent cis-mQTL associations accounting for 45% of the 450K array. To accurately assess the functionality of an mQTL, a more nuanced understanding of DNA methylome is needed to better model the proximity between the candidate variant and the differentially methylated CpG sites.

To address these concerns, we aim to identify a novel sequencing approach to detect DNA base modifications and genetic variations at the single-molecule level. Ultimately, we decide 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 have directly identified 5mC and 5hmC 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 benefiting 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.

Methods

Nanopore Open Data retrieval

Adaptive sampling (AS) is an advanced technique used in nanopore sequencing technology to selectively enrich or deplete specific regions of interest from a genomic sample in real-time. To evaluate the adaptive sampling performance, we accessed the Reduced Representation Methylation Sequencing (RRMS) result from 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 AWS command line interface. Since the BAM file included alignment information to the hg38 genome and per-read base modification predictions for CpG context 5mC call, we directly used it 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 requirement, we developed a PERL script (*MMMLparse.pl*) to generate a bisulfite emulative BAM file based on the base modification tag in aware of the indels, mismatch and soft-clip bases in the read. After the in-silico bisulfite conversion, we generated methylation haplotype retrieving PERL script (*getHaplo_SE_cgOnly.pl*) for single-end read adapted from MONOD2 script (*getHaplo_PE_cgOnly.pl*). After the methylation haplotype pattern retrieval, the methylation haplotype block (MHB) was defined with *cghap2mhbs.sh* MONOD2 script. Since the co-methylation profiling estimation was based on linkage disequilibrium, we also applied a minimal methylation frequency of 30% and a minimal read depth 50 for calculating the R^2 to avoid the bias caused by low variability site or poor read coverage. For MHB determination, we used minimal R^2 of 0.05 to visualize all possible co-methylation sites in Figure 1B-1C and used the same R^2 of 0.5 for direct comparison with Guo's result (Guo et al., 2017) in Supplementary Figure 1B-1E. After the MHB determination, we used "*smoothScatter*" function from the R graphics package to visualize the distribution of R^2 and the distance for those CpG sites in the same MHB.

5mCG methylation profiling near transcription factor binding motifs

To characterize the 5mCG 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 convert to bigwig file with 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 (Quinlan and Hall, 2010). We then used deeptools (Ramirez et al., 2016) to compute and visualize cumulative methylation to each motif and its surrounding genome. From the intermediate signal matrix, we extracted the cumulative methylation percentage from the most adjacent bin to compare the transcription factor (TF) motif region and the control region in t-tests.

Target design of adaptive sampling (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 (Machiela and Chanock, 2015) to retrieve variants highly associated with the tag SNPs (LD $R^2 \geq 0.5$) in either of the populations below: 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 TCGA prostate cancer cohort (Gong et al., 2019) and filtered the association 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 4000 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 prostate cell 22Rv1

The 22Rv1 (RRID: CVCL_1045) cell was obtained from the ATCC and grown in RPMI1640 medium supplemented with 10% fetal bovine serum (FBS). After harvesting the cell during the log phase, we used Puregene Cell Kit (QIANGEN, 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 Ligation Sequencing Kit (Oxford Nanopore Technology, SQK-LSK110). After library preparation, 100 fmol library solution was loaded to 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 running until the flow cell efficiency was fully depleted.

Nanopore adaptive sampling sequencing base-calling and read-level phasing

After the sequencing finished, we used dorado basecaller on NVIDIA GPU to infer DNA sequence and CpG site base modification based on the model “dna_r9.4.1_e8_sup@v3.3_5mCG_5hmCG”. 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 (Pedersen and Quinlan, 2018) to summarize the coverage for the genome and the target region. To resolve the SNP calling difficulty in the homopolymer region and improve the accuracy of SNP calling for nanopore long read data, we applied a pipeline called “PEPPER-Margin-DeepVariant” (Shafin et al., 2021) to provide the 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 (Lin et al., 2022).

Haplotype-specific methylation calling and de-novo differential methylated region (DMR) analysis

To generate multiple observations for the same CpG site, we used modkit pileup function to call 5mCG methylation with haplotype-partitioning and strand specific output. The output bedGraph files were used to identify de-nono DMRs between the two haplotypes in “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 1000 bp and lower the “--mincpgs” parameter (minimum number of CpGs in a DMR) to 5.

Correlation between chromatin accessibility and CpG methylation

To quantify the chromatin accessibility in the 22Rv1 cell, we calculated the Tn5 insertion coverage by normalizing the Transposase cut event count in each peak to the peak width. We also calculated the averaged 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 was obtained from the GEO accession GSE264518 (Tian et al., 2024). The eQTL data of prostate tissue was retrieved from GTEx Consortium (Consortium, 2020). The prostate cancer mQTL data was downloaded from Pancan-meQTL database (Gong *et al.*, 2019). The raw data for 22Rv1 sequencing was stored with the GEO accession number PRJNA1158791 (https://dataview.ncbi.nlm.nih.gov/object/PRJNA1158791?reviewer=ovnrt2bnad9muc43sjsa2nc5os).

Results

Co-methylation profiling guided by long-read sequencing

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 LD R^2 from the MONOD2 method 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 indicates strong co-methylation, while a value near 0 indicates random methylation. In the RRMS data (**Figure 1B, upper panel**), we visualized the LD R2 profiling with the CpG site distance and found that most of the LD R^2 in cancerous cell lines (COLO829 and HCC1395) was between 0 to 0.2, while the value in parental lymphoblast cell lines (COLO829BL and HCC1395BL) was between 0.2 to 0.6, suggesting an overall diminished synchronization of the epigenetic profiling in cancer cells. The long read (**Figure 1B, lower panel**) profiling also demonstrated a small proportion of distant CpG sites (≥1500 bp) showed weak-to-median co-methylation ($R^2 \geq 0.2$), suggesting an operation of unison between these sites. As an example (**Supplementary Figure 1A**), we demonstrated a locus with the long-range co-methylation on chromosome 13 in the 4 samples. According to Guo et al., we also applied a more stringent criterion of R^2 for MHB identification to see whether those highly correlated sites exsist in the nanopore data. However, with a minimal R^2 of 0.5, we did not find the perfect coupled sites (**Supplementary Figure 1B-1E**) as observed in the bisulfite sequencing data (Guo *et al.*, 2017).

CpG methylation footprint of human transcription factor (TF)

With the comprehensive AS coverage on the CpG dense region, we were able to estimate the epigenetic profiling in the genome surrounding the known human TF motifs. We scanned a total of 835 TF binding motifs and identified 251, 269, 194, and 148 TFs with significantly altered methylation ($p \leq 0.05$) in COLO829BL, COLO829, HCC1395BL, and HCC1395 cell line, respectively. In COLO829BL and COLO829 cell lines, we highlighted CTCF (**Figure 2A**), NRF1 (**Figure 2B**) and ZNF93 (**Figure 2C**) as the TF associated with hypomethylation in the nearby genome, and RARA (**Figure 2D**) as the TF associated with hypermethylation. Consistently, we observed similar trends in HCC1395BL and HCC1395 cells (**Supplementary Figure 2A-2D**).

Adaptive sampling nanopore sequencing targeting phenotypical variant

To selectively observe the DNA methylation patterns related to germline risk variants linking to prostate cancer phenotypes, we examined GWAS catalogue (https://www.ebi.ac.uk/gwas/home) and TCGA prostate cancer mQTL 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, we only found 2,539 SNPs overlapped between the two groups. Compared to genome background, the AS enriched 3.88 and 4.76 folds more sequences at mQTL and GWAS SNP regions, respectively (**Figure 3B**). Across the 22 chromosomes, we found a significant negative correlation between the target length (proportion of target region length to the belonging chromosome) and the fold of enrichment (**Figure 3C**).i Through grouping the methylation call by their location, we further identified expected hypomethylation for CpG sites located in CpG island (**Figure 3D**), and hypermethylation for CpG sites located outside CpG island (**Figure 3E**). 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. The DMR analysis found 6,390 genomic locations showed haplotype-specific 5mCG modification in the 22Rv1 genome (p-MWU ≤ 1e-5 and delta ≥ 50%). With the help of the "genomicDensity" function, we identified apparent imprinting regions consisting of DMR clusters across the genome in the circle plot (**Figure 3F**). The full list of DMRs can be accessed in **Supplementary Table 1**. Furthermore, we also found that some DMRs are likely caused by allele-specific methylation, offering strong function hints for the target variant.

Allele-specific methylation associated with GWAS or mQTL variant

Previously, we reported a functional prostate cancer risk variant causing allele-specific methylation in the 22Rv1 genome with bisulfite amplicon sequencing (Tian *et al.*, 2022). In nanopore sequencing data, we found that the risk SNP rs7247241 was adjacent to a 162-bp co-methylated region (chr19:38256705-38256867), with the T allele consistently hypermethylated than 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. We also highlighted another locus mapping to the KCNIP3 gene, with the genotype of mQTL SNP rs2113417 associated with the methylation level of a 710-bp region (chr2:95333760-95334470) (**Figure 4C**). Consistently, we identified an even stronger allele-specific chromatin accessibility in the region containing rs2113417 (**Figure 4C**). Interestingly, of the three mQTLs reported for rs2113417, only CpG site cg03595348 is located within the same DMR as the SNP, although its p-value did not show the best (**Figure 4D**). To describe the relationship between the chromatin accessibility and the methylation level at the genome scale, we calculated 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 concern of bisulfite sequencing has been thoroughly discussed over the last decade, particularly regarding its potential to fragment DNA templates and introduce PCR amplification bias during sample preparation. In the co-methylation analysis based on WGBS with short-read NGS, Guo et al. (Guo *et al.*, 2017) identified that over 80% of those CpG sites located in the same methylation haplotype block were with high LD ($R^2 \geq 0.9$), regardless of the cell type. However, in a PCR-free condition of nanopore long-read sequencing, we could not identify these highly associated co-methylation sites. One of the hypotheses is that during the PCR amplification step, co-methylated molecules may have been over-amplified, potentially introducing artificial LD sites. Another possibility is that when applying the population-based statistic to evaluate read-level co-methylation status, several additional metrics must be considered, such as the minimum read depth and minor allele frequency of the two CpG sites. According to previous discussion (Eberle et al., 2006; VanLiere and Rosenberg, 2008; Wray, 2005), R^2 calculation is dependent on allele frequencies of the two loci, which could easily generate perfect LD ($R^2 = 1.0$) when both CpG sites are with extremely high and low methylation frequencies. Therefore, we applied a minimal read depth of 50 and a minor allele frequency of 0.3 for the co-methylation R^2 calculation, aiming to focus on those sites of high confidentiality in the PCR-free condition.

The long-read technology also provided a broader vision to study epigenetic co-regulation. Ideally, sequencing the human genome at the telomere-to-telomere scale would reveal a full spectrum of possible long-range interactions based on epigenetic modification. However, the technical issue is that as the read length increases, the nanopore sequencing throughput are reduced due to the complex DNA secondary structure or heavy modification genomic region obstruction, which explains why the RRMS demo data used 8kb fragmented DNA for maximum read coverage. Based on the observations of the distant co-methylation (**Supplementary Figure 1A**), the application successfully reaching high coverage with extended read lengths may provide greater insight into these longer-range interactions. Our analyses of the methylation profiling on known binding motifs also described a unique epigenetic footprint for human TFs. The observation of methylation changes of CTCF motif is consistent with previous studies (Damaschke et al., 2020), suggesting that the CTCF is a major in directing localized DNA hypermethylation. Other TFs reported to be sensitive to or associated with DNA methylation, such as NRF1 (Domcke et al., 2015), ZNF93 (Fukuda et al., 2022) and RARA (Hassan et al., 2017), also demonstrated consistent adjacent methylation changes in our analysis.

With the increasing availability of powerful GPU resources, more and more applications use nanopore adaptive sampling (AS) to enrich target-of-interest or deplete unwanted genome (Kovaka et al., 2021; Loose et al., 2016). The computational approach utilizes GPU-based fast basecalling to perform raw electrical signal mapping to make real-time “Read Until” decisions to select specific DNA molecules. Here, we successfully applied the AS approach to enrich the germline variant regions, aiming to obtain high PCR-free coverage to identify allele-associated epigenetic changes. However, due to the relatively high error rate in nanopore reads, the alignments are frequently filled with small 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 (Poplin et al., 2018; Shafin *et al.*, 2021) to polish the candidate variants to achieve the best calling accuracy. Based on the continuity nature of CpG methylome, we implemented a de-nono DMR selection procedure to identify clusters of CpG sites that showed modification differences between the two haplotypes. Considering that the haplotype-specific DMRs don’t inform the potential causal variants, we included additional information of the closest phasing SNP to each DMR and also annotated further whether the SNP is an GWAS or mQTL variant. From the results, we identified 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 our DMR analysis reproduced this finding. For another SNP rs2113417, we successfully applied the nanopore data to discover the most related DMR between the alleles, which further revealed that the biological origin is from the CpG site cg03595348 among the mQTL signals. 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 for observing the epigenetic effect driven by those rare alleles. One shortcoming of using the utility to a population cohort is that the read-level variant calling and phasing based on the nanopore long-read data is currently limited to each diploid individual, which will generate different haplotype phase set information according to different persons. This could challenge the between-haplotype DMR identification. To further expand this utility to a multiple-sample design, a genotype-based reference would be helpful for further harmonizing and standardizing the haplotype information so that all samples could be compared and analyzed together.

Taken together, we spotlighted multiple advantages of the nanopore long-read sequencing for methylation analysis and envisioned an innovative utility with this technology to address the technical issue in determining allele-specific methylation. The analytical framework introduced in this study for long-read DNA methylation analysis offers a new perspective and uncovers previously unknown insights in the field of genetics.

Acknowledgements

This study has been 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 Statement

Conception and design: Y. Tian, S. McDonnell, N. Larson, L. Wu, L. Wang
Development of methodology: Y. Tian, S. McDonnell, N. Larson, L. Wu, L. Wang
Acquisition of data: Y. Tian
Analysis and interpretation of data: Y. Tian, L. Wu, S. McDonnell, N. Larson
Writing, review, and revision of the manuscript: Y. Tian, S. McDonnell, N. Larson, L. Wu, L. Wang
Study supervision: Y. Tian, N. Larson, L. Wu, L. Wang

Conflicts of Interest Statement

L.Wu. provided consulting service to Pupil Bio Inc., and reviewed manuscripts for Gastroenterology Report, not related to this study, and received honorarium. No potential conflicts of interest were disclosed for other authors.

Figure Legends

Figure 1. A. Explanation of using linkage disequilibrium R^2 calculation for co-methylation analysis. Smooth scatter plot of co-methylation profiles for CpG sites located in the methylation haplotype block (MHB) in the normal-tumor pair of COLO829BL and COLO829 (**B**) cells, and HCC1395BL and HCC1395 (**C**) cells.

Figure 2. In COLO829BL and COLO829 cells, CpG methylation status surrounding known transcription factor binding motifs, such as CTCF (**A**), NRF1 (**B**), ZNF93 (**C**) and RARA (**D**).

Figure 3. A. Proportional Venn diagram of the target SNP with TCGA mQTL signals and GWAS risk significance. **B.** Sequencing coverage summarized by chromosomes with adaptive sampling run in 22Rv1 genome. **C.** Correlation between the target length and enrichment fold across the 22 autosomes. Distribution of dual methylation calling for the CpG sites in (**D**) or outside (**E**) 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 4. A. Exemplary differential methylated region (DMR) identified near the GWAS risk variant rs7247241 and the 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 differential methylated region (DMR) identified near the mQTL variant rs2113417 and the chromatin accessibility profiling with the rs2113417 alleles in 22Rv1 cells. **D.** GTEx eQTL results for rs2113417 genotype and KCNIP3 expression in prostate tissue. **F.** TCGA mQTL results for rs2113417 in prostate cancer tissue.

References

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. 10.1038/ng.3721.

Consortium, G.T. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318-1330. 10.1126/science.aaz1776.

Consortium, I. (2024). Deciphering the impact of genomic variation on function. Nature *633*, 47-57. 10.1038/s41586-024-07510-0.

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 Epigenetics *12*, 80. 10.1186/s13148-020-00869-7.

Domcke, S., Bardet, A.F., Adrian Ginno, P., Hartl, D., Burger, L., and Schubeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature *528*, 575-579. 10.1038/nature16462.

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. 10.1371/journal.pgen.0020142.

Fukuda, K., Makino, Y., Kaneko, S., Shimura, C., Okada, Y., Ichiyanagi, K., and Shinkai, Y. (2022). Potential role of KRAB-ZFP binding and transcriptional states on DNA methylation of retroelements in human male germ cells. Elife *11*. 10.7554/eLife.76822.

Gong, J., Wan, H., Mei, S., Ruan, H., Zhang, Z., Liu, C., Guo, A.Y., Diao, L., Miao, X., and Han, L. (2019). Pancan-meQTL: a database to systematically evaluate the effects of genetic variants on methylation in human cancer. Nucleic Acids Res *47*, D1066-D1072. 10.1093/nar/gky814.

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. 10.1038/ng.3805.

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. 10.1016/j.celrep.2017.05.007.

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. 10.1038/s41467-019-12228-z.

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. 10.1038/s41587-020-0731-9.

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. 10.1093/bioinformatics/btac058.

Loose, M., Malla, S., and Stout, M. (2016). Real-time selective sequencing using nanopore technology. Nat Methods *13*, 751-754. 10.1038/nmeth.3930.

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. 10.1093/bioinformatics/btv402.

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. 10.1038/s41588-021-00923-x.

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. 10.1093/jnci/djae173.

Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867-868. 10.1093/bioinformatics/btx699.

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. 10.1038/nbt.4235.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842. 10.1093/bioinformatics/btq033.

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*, W160-165. 10.1093/nar/gkw257.

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. 10.1038/s41592-021-01299-w.

Tian, Y., Dong, D., Wang, Z., Wu, L., Park, J.Y., Wei, G.H., Wang, L., and consortium, P.E. (2023). Combined CRISPRi and proteomics screening reveal a cohesin-CTCF-bound allele contributing to increased expression of RUVBL1 and prostate cancer progression. bioRxiv. 10.1101/2023.01.18.524405.

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. 10.1093/hmg/ddab347.

Tian, Y., Wu, L., Huang, C.C., and Wang, L. (2024). Identify Regulatory eQTLs by Multiome Sequencing in Prostate Single Cells. bioRxiv. 10.1101/2024.06.19.599704.

VanLiere, J.M., and Rosenberg, N.A. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. Theor Popul Biol *74*, 130-137. 10.1016/j.tpb.2008.05.006.

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. 10.1375/1832427053738827.

Figure 1

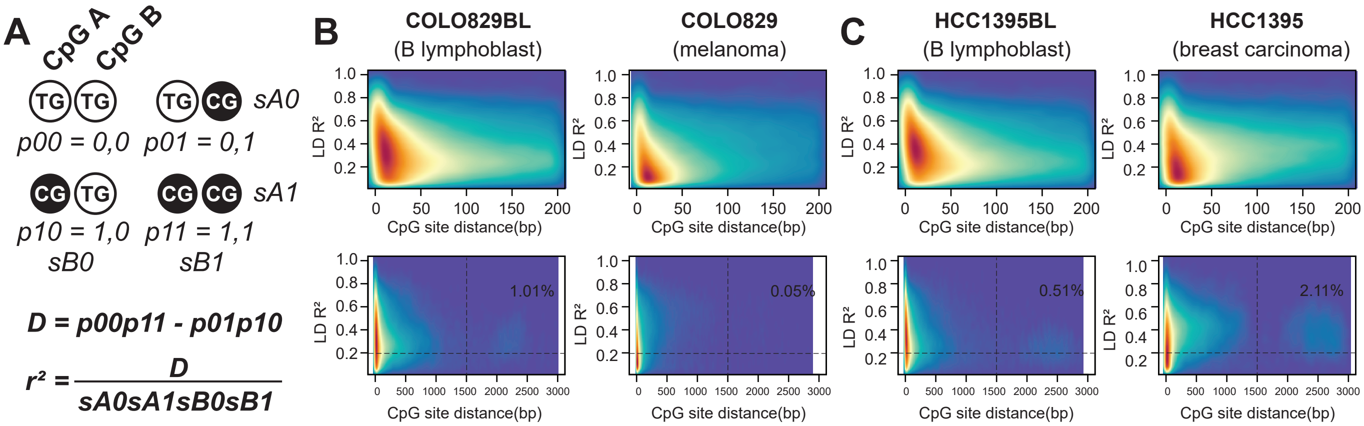

A
CpG A
CpG B
TG TG
TG CG
sA0
p00 = 0,0
p01 = 0,1
CG TG
CG CG
sA1
p10 = 1,0
p11 = 1,1
sB0
sB1
D = p00p11 - p01p10
r² = D / sA0sA1sB0sB1
B
COLO829BL
(B lymphoblast)
COLO829
(melanoma)
C
HCC1395BL
(B lymphoblast)
HCC1395
(breast carcinoma)
LD R²
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
CpG site distance(bp)
500
1000
1500
2000
2500
3000
1.01%
0.05%
0.51%
2.11%

Figure 2

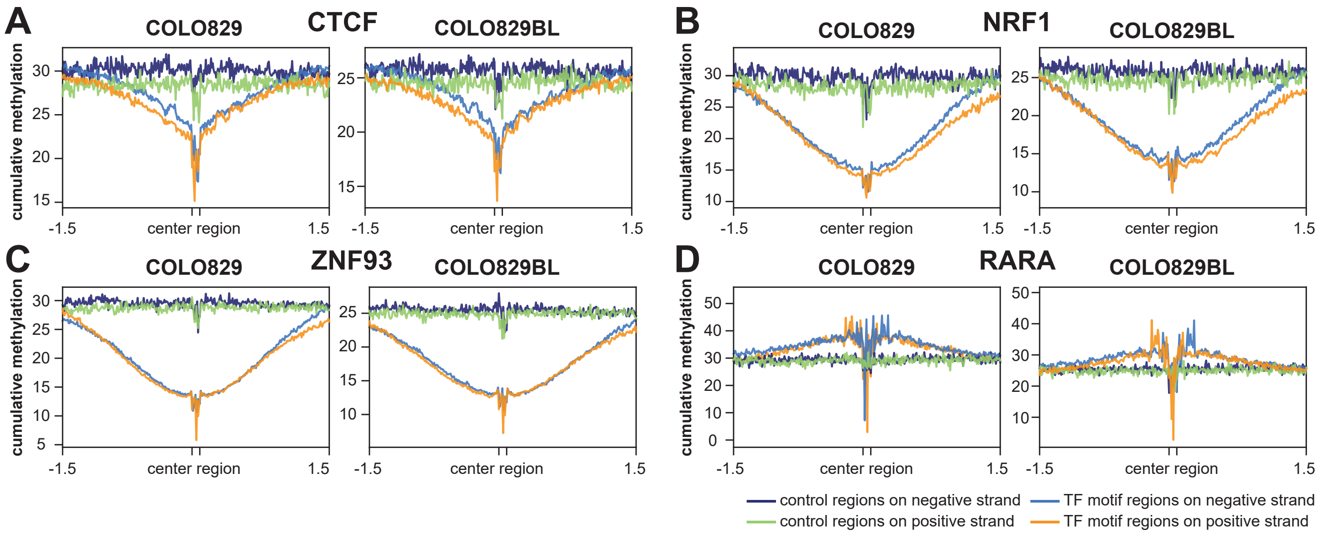

A
CTCF
COLO829
COLO829BL
cumulative methylation
30
25
20
15
25
20
15
-1.5
center region
1.5
B
NRF1
COLO829
COLO829BL
cumulative methylation
30
25
20
15
10
25
20
15
10
C
ZNF93
COLO829
COLO829BL
cumulative methylation
30
25
20
15
10
5
25
20
15
10
D
RARA
COLO829
COLO829BL
cumulative methylation
50
40
30
20
10
0
50
40
30
20
10
control regions on negative strand
control regions on positive strand
TF motif regions on negative strand
TF motif regions on positive strand

Figure 3

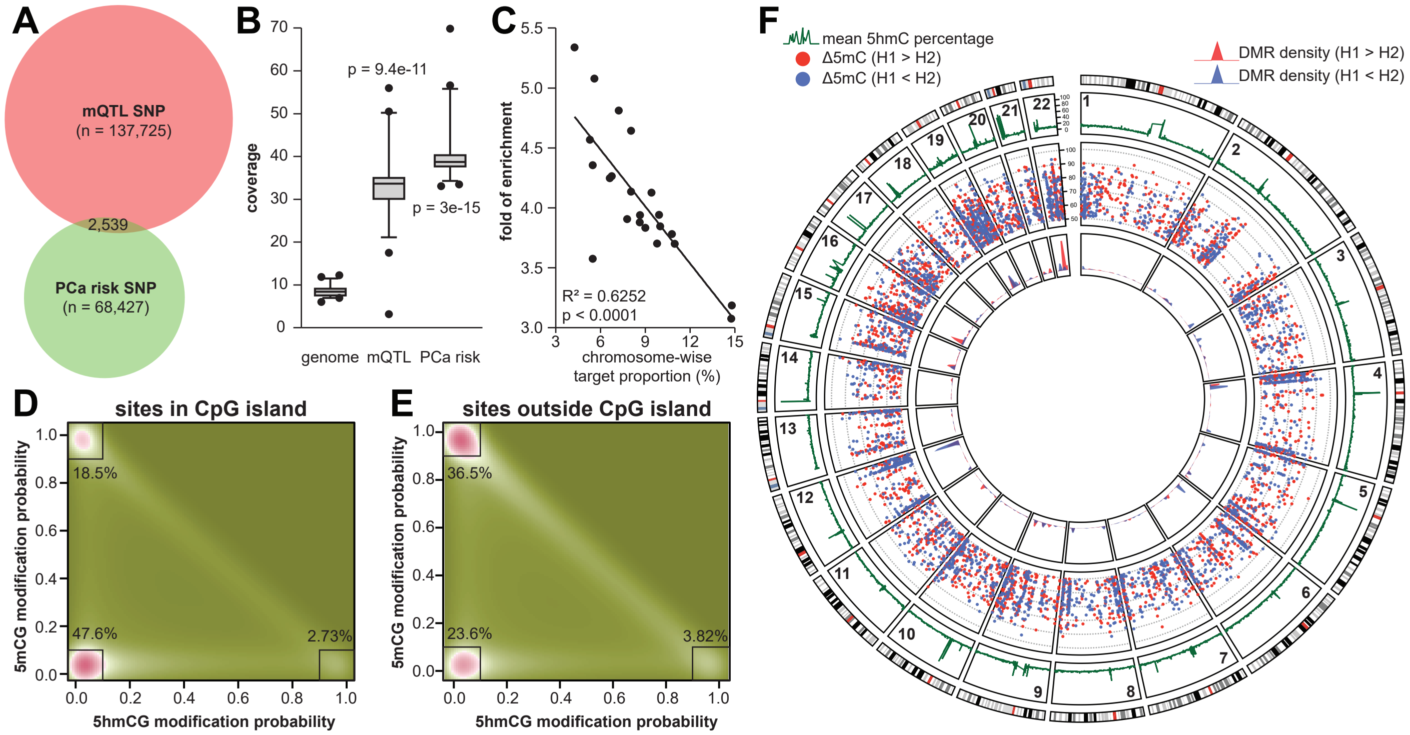
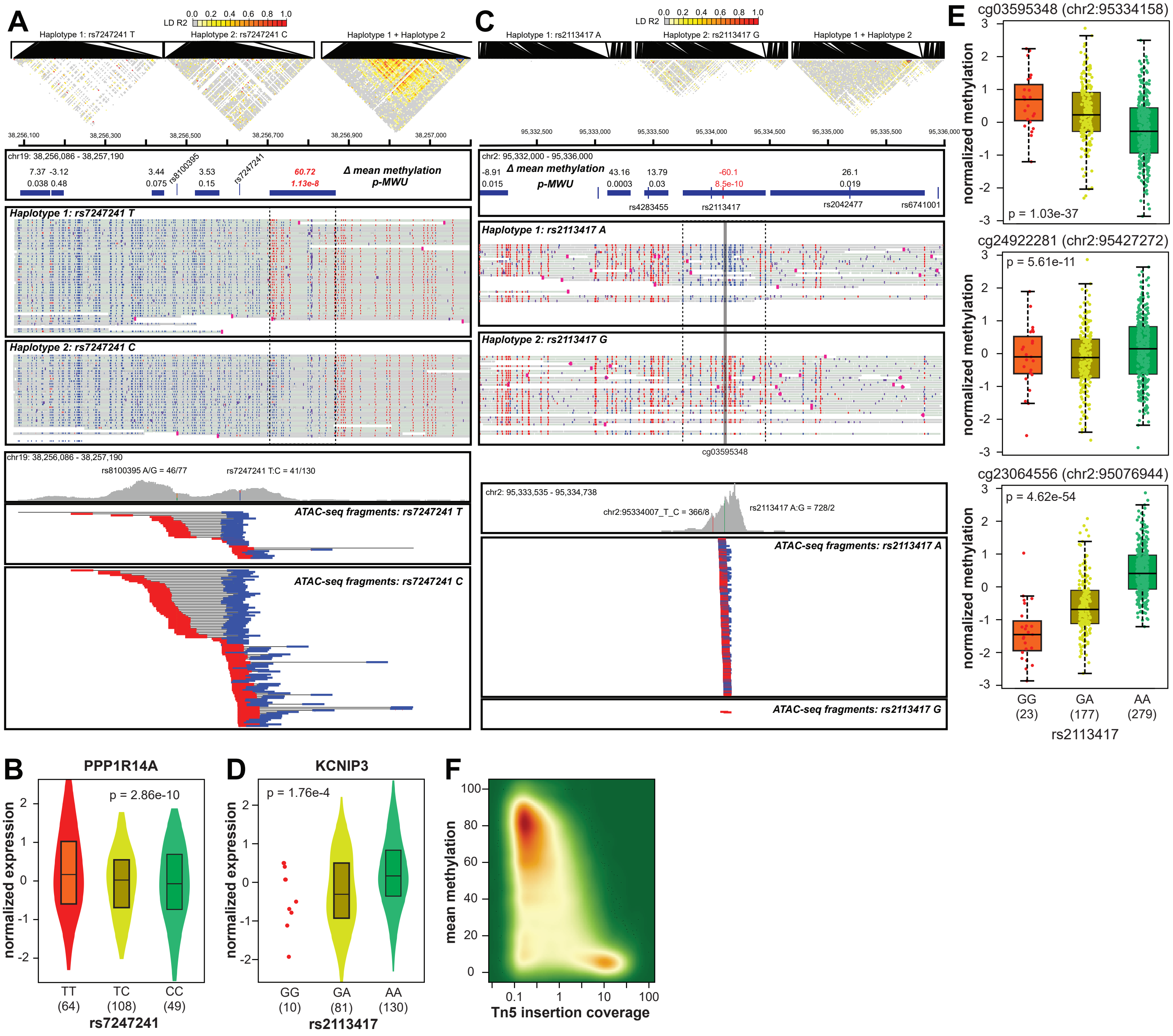
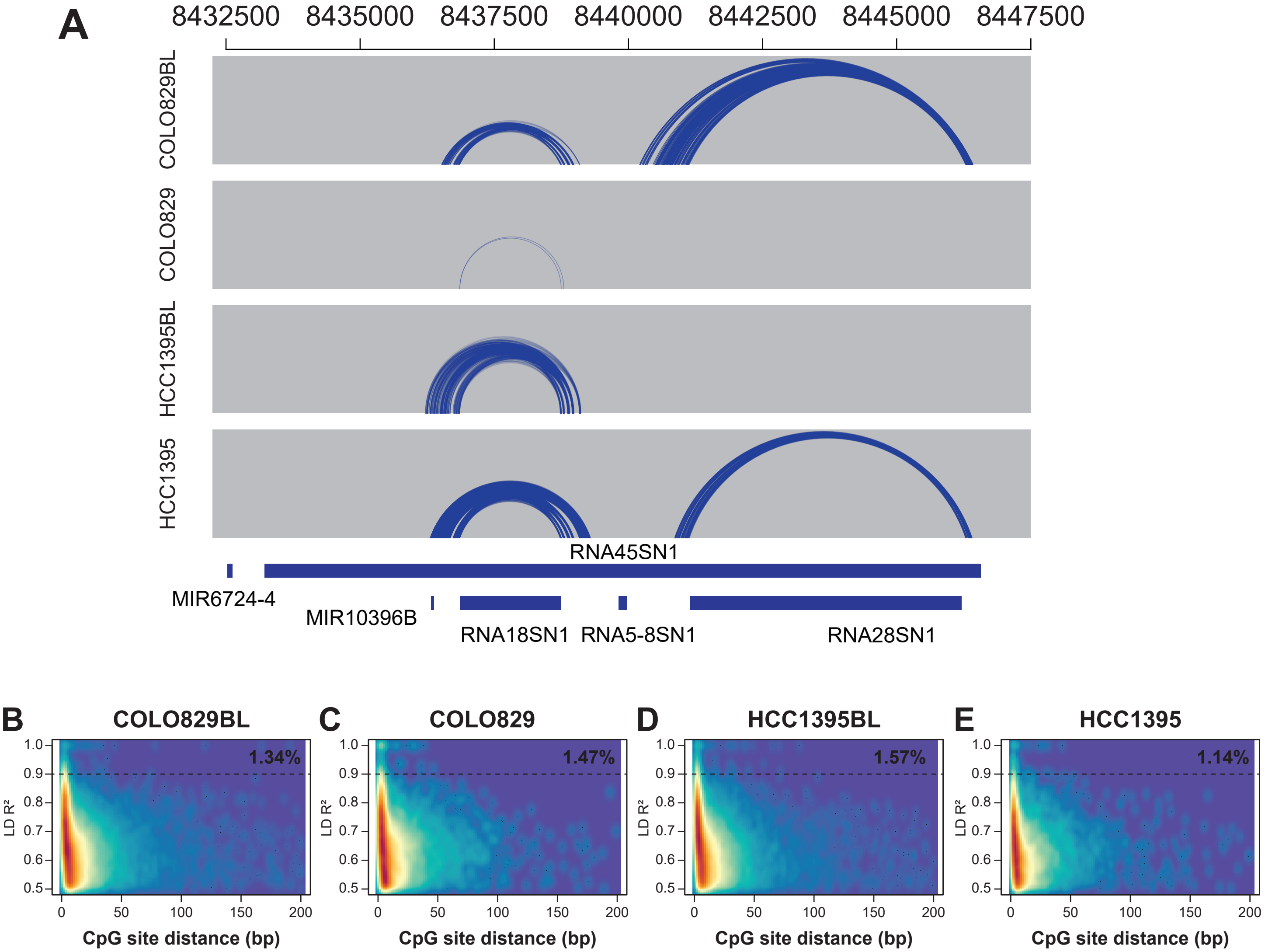

A
mQTL SNP
(n = 137,725)
2,539
PCa risk SNP
(n = 68,427)
B
coverage
70
60
50
40
30
20
10
0
p = 9.4e-11
p = 3e-15
genome
mQTL
PCa risk
C
fold of enrichment
5.5
5.0
4.5
4.0
3.5
3.0
R² = 0.6252
p < 0.0001
3
6
9
12
15
chromosome-wise target proportion (%)
F
mean 5hmC percentage
Δ5mC (H1 > H2)
Δ5mC (H1 < H2)
DMR density (H1 > H2)
DMR density (H1 < H2)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
D
sites in CpG island
5mCG modification probability
5hmCG modification probability
18.5%
47.6%
2.73%
E
sites outside CpG island
36.5%
23.6%
3.82%
0.0
0.2
0.4
0.6
0.8
1.0

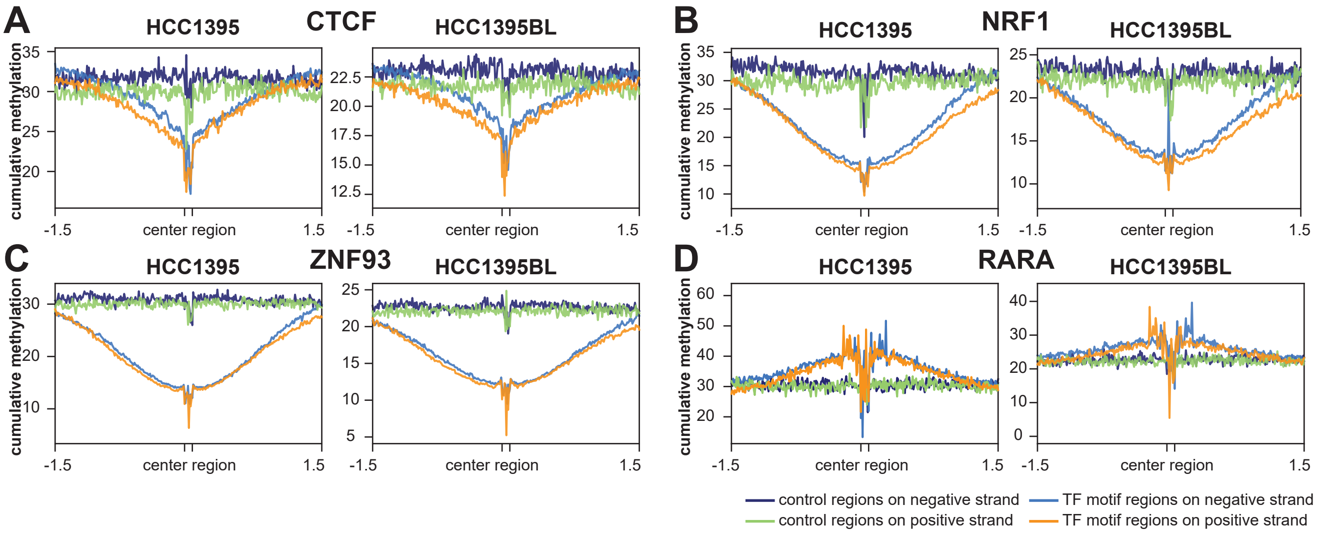
Figure 4


A
LD R2
0.0 0.2 0.4 0.6 0.8 1.0
Haplotype 1: rs7247241 T
Haplotype 2: rs7247241 C
Haplotype 1 + Haplotype 2
38,256,100
38,256,300
38,256,500
38,256,700
38,256,900
38,257,000
chr19: 38,256,086 - 38,257,190
7.37 -3.12
0.038 0.48
3.44
0.075
rs8100395
3.53
0.15
rs7247241
60.72
1.13e-8
Δ mean methylation
p-MWU
Haplotype 1: rs7247241 T
Haplotype 2: rs7247241 C
chr19: 38,256,086 - 38,257,190
rs8100395 A/G = 46/77
rs7247241 T:C = 41/130
ATAC-seq fragments: rs7247241 T
ATAC-seq fragments: rs7247241 C
C
LD R2
0.0 0.2 0.4 0.6 0.8 1.0
Haplotype 1: rs2113417 A
Haplotype 2: rs2113417 G
Haplotype 1 + Haplotype 2
95,332,500
95,333,000
95,333,500
95,334,000
95,334,500
95,335,000
95,335,500
95,336,000
chr2: 95,332,000 - 95,336,000
-8.91
0.015
Δ mean methylation
p-MWU
43.16
0.0003
13.79
0.03
-60.1
8.5e-10
26.1
0.019
rs4283455
rs2113417
rs2042477
rs6741001
Haplotype 1: rs2113417 A
Haplotype 2: rs2113417 G
cg03595348
chr2: 95,333,535 - 95,334,738
chr2:95334007_T_C = 366/8
rs2113417 A:G = 728/2
ATAC-seq fragments: rs2113417 A
ATAC-seq fragments: rs2113417 G
E
cg03595348 (chr2:95334158)
p = 1.03e-37
cg24922281 (chr2:95427272)
p = 5.61e-11
cg23064556 (chr2:95076944)
p = 4.62e-54
normalized methylation
3
2
1
0
-1
-2
-3
GG
(23)
GA
(177)
AA
(279)
rs2113417
B
PPP1R14A
p = 2.86e-10
normalized expression
2
1
0
-1
-2
TT
(64)
TC
(108)
CC
(49)
rs7247241
D
KCNIP3
p = 1.76e-4
normalized expression
2
1
0
-1
-2
GG
(10)
GA
(81)
AA
(130)
rs2113417
F
mean methylation
100
80
60
40
20
0
0.1
1
10
100
Tn5 insertion coverage

Supplementary Figure 1



Supplementary Figure 2


A
CTCF
HCC1395
HCC1395BL
cumulative methylation
center region
B
NRF1
HCC1395
HCC1395BL
C
ZNF93
HCC1395
HCC1395BL
D
RARA
HCC1395
HCC1395BL
control regions on negative strand
control regions on positive strand
TF motif regions on negative strand
TF motif regions on positive strand