Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing

Isaac Lee, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Ariel Gershman, Norah Sadowski, Fritz J. Sedlazeck, Kasper D. Hansen, Jared T. Simpson and Winston Timp.

With the proliferation of DNA sequencing technologies, methods have been developed for examining nuclear organization, protein-binding site occupancy, chromatin accessibility and methylation state. Many of these methods rely on the vulnerability of accessible chromatin to enzymatic treatment, for example, DNase I hypersensitive sites sequencing (DNase–seq) and the assay for transposase-accessible chromatin–sequencing (ATAC–seq). One such method, nucleosome occupancy and methylome (NOMe–seq), labels accessible genomic regions using an exogenous MethylC and methyltransferase. Combined with bisulfite conversion, NOMe–seq permits simultaneous evaluation of endogenous cytosine methylation and nucleosome occupancy. This has been adapted to single-cell approaches that have shown the potential of nucleosome footprinting for exploring epigenetic heterogeneity.

Nanopore sequencing is a single-molecule long-read sequencing strategy, which allows collection of data on long reads of unamplified DNA, providing deeper insights into long-range patterns on individual molecules. We and others have previously shown that endogenous CpG methylation can be accurately called with nanopore data. More recently, this technology was applied to exogenous labeling of chromatin accessibility in Saccharomyces cerevisiae, a unicellular eukaryotic model organism without endogenous methylation. Others have demonstrated this application on the PacBio platform even in human cells, although they did not capture endogenous methylation.

To complement bulk and single-cell epigenetic methods, here we present nanopore sequencing of nucleosome occupancy and methylation (nanoNOMe), where we label mammalian cells that have endogenous CpG methylation with exogenous GpC modifications at accessible sites. We are able to take advantage of the long reads (>10 kb) generated by nanopore sequencing to measure methylation and accessibility across stretches of genomic regions at the single-molecule level. This allowed us to evaluate allele-specific epigenetic states across the genome, including regions that are hard to characterize such as repetitive elements and structural variations (SVs). Using this approach, we simultaneously determined phased patterns of native CpG methylation and chromatin accessibility in four different cell lines. The complement of our long-read sequencing methods to burgeoning single-cell methods provides additional insight into the complexity of the epigenome.

Results

Nanopore CpG and GpC methylation calling. We previously demonstrated CpG methylation calling from nanopore sequencing with our software tool nanopolish. The methylation caller outputs log-likelihood ratios (LLRs) for the probability of methylation at a given k-mer, and a threshold cutoff for LLR is used to call CpGs within a k-mer as methylated or unmethylated. Extending the methylation caller to call GpC methylation simultaneously required new training sets, which we generated using combinations of M.SssI (CpG methyltransferase) and M. CviPI (GpC methyltransferase) on unmethylated (PCR amplified) Escherichia coli genomic DNA (gDNA) (Methods). This resulted in samples with methylation at CpG, GpC, both CpG/GpC or completely unmethylated, which we confirmed with Illumina bisulfite sequencing (Supplementary Tables 1 and 2). We then sequenced these samples to generate training data with an average of 1,300× coverage. We plotted the difference in mean current for all 6-mers containing a methylated motif.
Fig. 1 | Overview and assessment of nanoNOMe. a. Receiver operating characteristic curve of methylation calling for a range of LLR thresholds on control samples (GM12878 gDNA modified with Cpg and GpC methylation). AUC, area under curve. b. A schematic of exogenous labeling: intact nuclei were methylated at GpC motifs to serve as the mark for accessibility, and Cpg methylation and GpC accessibility can be simultaneously measured. c. Comparison of the fraction of low sequence complexity regions between WGBS and nanoNOMe that had robust read coverage (coverage between the 5th and 95th percentiles of genome coverage). d, e. Validation of NanoNOMe profiles by pairwise comparison of per-Cpg average methylation from nanoNOMe with WGBS in across the genome (d) and intersections of accessibility peaks from nanoNOMe, DNase–seq and ATAC–seq (e). f. Distribution (left) of observed per-Cpg site methylation frequency in repetitive elements in comparison to random regions across the genome of the same lengths and number (right) of accessibility peaks per 1Mb of repetitive regions in comparison to the entire genome.

(Extended Data Fig. 1a) and found that the deviation is the highest when the methylation occurs on the fifth position along a 6-mer. To benchmark the dual methylation detection, we tested the model on gDNA from the GM12878 human lymphoblast cell line, generating control samples with the same PCR and methyltransferase treatment as the E. coli samples (Supplementary Table 1). After nanopore sequencing, we first confirmed that full methylation does not decrease mappability of the reads in nanopore sequencing (Supplementary Table 2). We tested the performance of nanopopolish with our trained four-state model on these test samples and found high areas under the curve for both Cpg and GpC calls across a range of LLR thresholds (0.91 for Cpg and 0.98 for GpC) (Fig. 1a). We selected LLR cutoffs of –1.5/1.5 for Cpg and –1/1 for GpC methylation, resulting in correct identification of 91% of Cpg calls at the 72% of Cpgs that pass the threshold and 96% of GpCs calls correctly identified at the 93% that pass the threshold (Extended Data Fig. 1b and Supplementary Table 3). This is a conservative estimate of our accuracy because these metrics were calculated with the assumption that the methylated input was 100% methylated, whereas the bisulfite sequencing data indicated incomplete (roughly 96–98%) enzymatic methylation in this testing set. Ambiguously called k-mers were not enriched for any specific sequence (Supplementary Fig. 1b,c) beyond the GCG motif, which is excluded from our analysis. Genome context analysis confirmed that neither the fraction of sites called nor the fraction of accurate calls was dependent on the genomic context (Supplementary Table 4).

Chromatin and DNA methylation profiling with NanoNOMe. With the four-state model in hand, we then adapted NOME–seq’ to nanopore sequencing, exogenously labeling open chromatin with GpC methylation (Fig. 1b). Methylation at cytosines in GCH and HCG contexts were used as measures of chromatin accessibility and endogenous methylation, respectively. We excluded both Cpg and GpC methylation data from GCG contexts, representing 5.6% of GpCs and 24.2% of Cpgs, because of the ambiguity of native methylation or chromatin state informing these locations1. In describing GpCs state, a methylated GpC was interpreted as an accessible mark and unmethylated as inaccessible.

First, we performed nanoNOMe on GM12878 generating 250 gigabases (Gb) (103× coverage) of mapped sequencing data from 15 flowcells (12 minION and three PromethION), with an N50 read length of 14,000bp (Table 1 and Supplementary Table 5). We compared genomic coverage of the resulting nanoNOMe data to whole-genome bisulfite sequencing (WGBS) from a previous study (100× coverage, ENCODE accession ENCSR890UQ0)13 and whole-genome nanopore sequencing of GM12878 (36× coverage, European Nucleotide Archive accession code PRJEB23027)14. We found that the ENCODE WGBS had a GC bias, while nanoNOMe and nanopore whole-genome sequencing coverage were not biased by GC content (Supplementary Fig. 2)15,16.

We then examined regions that are poorly mappable via short reads. We focused on regions of low mappability in WGBS, determined as loci that had ten or more reads with a mapping quality of
less than five. These regions covered 132 Mb of the human genome (roughly 4.5%), consisting of 57,982 distinct regions with an average size of 2.3 kb. The coverage of high-quality reads (mapping quality >20) for nanoNOMe was between the 5th and 95th percentiles of genome coverage (67–116×) for 44% of these regions with a median coverage of 114×. In contrast, only 7% of these regions are between the 5th and 95th percentiles of coverage in WGBS (23–168×), with an abnormally high median coverage of 582× compared to the overall median coverage of 100×. We also examined repetitive elements, which are known to be difficult to map with short-read sequencing, and CpG islands, which are often affected by dropouts due to their high GC content (Supplementary Fig. 2). We observed that nanoNOMe maps to higher fractions of robustly mapped regions than WGBS, especially LINE and CGI (Fig. 1c). Long-read sequencing, and specifically nanoNOMe, does not suffer from mismapping of reads to poorly mappable regions, enhancing our ability to interrogate these sites.

We next assessed the performance of nanoNOMe in simultaneously resolving endogenous cytosine methylation and chromatin accessibility. First, we confirmed that CpG methylation measured from nanoNOMe correlates well with WGBS methylation signal (Pearson correlation of 0.92) (Fig. 1d). We called accessibility peaks based on the frequency of CpG methylation (Methods), and found that out of the 69,305 peaks, 58,742 overlapped with peaks called by ATAC–seq and/or DNase–seq (Fig. 1e and Supplementary Data 1). We then used nanoNOMe to footprint nucleosome positioning by generating metaplots at various genomic contexts. As in Kelly et al., we used computationally predicted CCCTC-binding factor- (CTCF-) binding motifs (from CTCFBSDB v.2.0) that were ≥2 kb away from transcription start sites (TSS) and supported by CTCF chromatin immunoprecipitation coupled with massively parallel DNA sequencing (ChiP–seq) in GM12878, resulting in 6,793 sites (Supplementary Data 2). The methylation and DNA accessibility agreed with gold standard methods (WGBS and MNase-seq, respectively) from previous studies (Extended Data Fig. 2, ENCODE accession numbers ENCSR890UQO and ENCSR000CXP).

To correlate these observations with other epigenetic modifications, we generated metaplots at TSSs with euchromatic (H3K4me3) and heterochromatic (H3K27me3) histone modifications using existing ChiP–seq data on GM12878 (Supplementary Fig. 3, ENCODE accession numbers ENCSR057BWO and ENCSR000AKD). As expected, CpG methylation decreased and GpC accessibility increased at the TSS in promoters with active H3K4me3 marks, in contrast to the high CpG methylation and low accessibility at promoters with repressive H3K27me3 marks. To correlate these epigenetic changes with the transcriptome, we separated the genes by expression quartiles and measured the promoter epigenetic signature, observing this same concordant decrease in CpG methylation and an increase in GpC accessibility at promoters with increasing expression level (Supplementary Fig. 4 and Extended Data Fig. 3 ENCODE accession ENCSR843RJV).

We then characterized epigenetics of annotated repetitive elements in GM12878 using the CpG methylation and accessibility profiles, specifically in LINE, LTR, Alu and MIR, which are the four most abundantly annotated repetitive elements (Fig. 1f). On comparing average methylation distributions of these repetitive regions to randomly shuffled regions, excluding those that overlap with the repetitive elements, we observed that only Alu elements exhibit an increase in methylation. Accessibility peaks were depleted in all repetitive elements, especially in LINE and LTR regions, indicating that repetitive elements have decreased accessibility across the genome in GM12878.

NanoNOMe reveals epigenetic patterns with single-read resolution. Previous studies have demonstrated that nucleosome positioning and DNA accessibility are heterogeneous even within a homogeneous cell population2,14,15, highlighting the importance of probing these features on individual copies of the DNA. However, single-read analysis of nanoNOMe's GpC accessibility is difficult because of the variation in enzymatic efficiency of the M. CviPI GpC methyltransferase, and noise associated with single-molecule measurements. In the control test set, we examined the patterns of incorrect accessibility calls and found that 75% of incorrect calls were singular events, surrounded by correct calls. To remove the isolated noise, we estimated the accessibility of a given site using information from nearby CpG motifs on the same molecule, dampening the isolated erroneous signal (Methods). Briefly, we applied a Gaussian kernel regression on the LLRs of accessibility calls using fixed genomic coordinate bandwidths and estimated accessibility across individual reads. We smoothed the GM12878 nanoNOMe data at CTCF-binding sites and verified that it reduces the frequency of artificial spikes in accessibility, while retaining the ability to footprint nucleosome positioning (Extended Data Fig. 4).

We proceeded to characterize patterns of accessibility and methylation at CTCF-binding sites on individual reads. First, we selected reads that span 2 kb regions centered on the 6,793 CTCF-binding sites with a ChiP–seq peak and the 4,247 binding sites without a ChiP–seq peak and examined runs of accessible and inaccessible calls using run-length encoding methods (Fig. 2a and Extended Data Fig. 5a). We found the length of inaccessible runs correspond to the units of nucleosomes, shown by hotspots of inaccessible runs at 128 bp (mononucleosomes) and 310 bp (dinucleosomes). Examining the length of the inaccessible runs at the center of CTCF-binding sites, we found a higher occurrence of shorter runs (<80 bp), subsequently termed subnucleosomal footprints, indicating CTCF binding (Extended Data Fig. 5b,c). This short length of inaccessibility by regulatory protein binding is consistent with previous findings of protein–DNA interactions via DNase hypersensitivity and X-ray crystallography20–22. We investigated whether the number of GpC sites available proximal to the CTCF-binding site affected our ability to observe this pattern, and found that as few as three GCH reproduce the expected pattern (Supplementary Fig. 5).

| Table 1 | Sequencing statistics of nanoNOMe sequencing data |
|---------|----------------------------------|
| Cell    | Number of flowcells | Number of raw reads (million) | Total raw bases (Gb) | Aligned reads (million) | Aligned bases (Gb) | Average coverage | N50 length |
|---------|----------------------|-------------------------------|---------------------|------------------------|-------------------|-----------------|------------|
| GM12878 | 12 + 3 Plon*         | 32.0                          | 298.3               | 26.4                   | 256.9             | 103             | 14,020     |
| MCF-10A | 9                    | 9.4                           | 81.6                | 7.7                    | 72.4              | 27              | 11,501     |
| MCF-7   | 11                   | 9.0                           | 76.8                | 7.5                    | 69.1              | 26              | 13,025     |
| MDA-MB-231 | 9                 | 8.0                           | 82.4                | 7.0                    | 74.9              | 28              | 13,507     |

*PromethION flowcell (all other were MinION). NanoNOMe was performed on four cell lines using multiple runs of MinION, GridION or PromethION sequencing and pooled to generate one dataset per cell line.
For subsequent analyses, we only considered regions with four or more GpC motifs to predict subnucleosomal footprints.

We classified CTCF-bound reads as inferred by the presence of a subnucleosomal footprint at a CTCF-binding site (Extended Data Fig. 6). Because nanoNOMe does not rely on enrichment or PCR to detect accessibility, the fraction of CTCF-bound reads represents a quantitative estimate of CTCF binding. We found a moderate correlation (r=0.50) between nanoNOMe CTCF-binding fraction and CTCF ChIP-seq signal (Extended Data Fig. 7a). The median bound fraction of nanoNOMe reads at ChIP-seq peaks (0.25) was much higher than the median fraction at locations with no ChIP-seq peak (0.03) (Extended Data Fig. 7b). We stratified the reads based on their nanoNOMe-inferred CTCF-binding states and the presence of ChIP-seq peaks at the motif (Fig. 2b). We found that reads at sites with ChIP-seq peaks have consistently lower methylation and well-positioned nucleosomes even on molecules that were not classified as bound by nanoNOMe. At locations without a ChIP-seq peak that are not bound according to nanoNOMe, we do not see regular nucleosome positioning, agreeing with the literature. Our data finds a small number of reads (2%, 8,893 out of 452,812 reads) indicating bound CTCF and organized nucleosomes by nanoNOMe at sites without ChIP-seq peaks. We attribute the lack of a clear peak at these sites to the low CTCF occupancy, as most sites without peaks have low fractions of bound reads (Extended Data Fig. 7b).

On the TSS of highly transcribed genes, we observed the expected pattern of well-organized nucleosome positioning and longer accessible runs representing nucleosome-depleted regions (Fig. 3a)6. With decreasing expression, CpG methylation increased and accessibility decreased around TSS (1 kb for CpG and 200 bp for GpC) on a single-read level (Supplementary Data 6a). As in subnucleosomal footprint detection, we examined the effect of the number of GpC sites in the TSS window on the ability to distinguish reads with open versus closed promoters. As observed with previous studies, accessibility signal at TSSs is confounded by GC content because promoter GC content affects the gene’s activity (Supplementary Fig. 7). Therefore, we used genes that have >13 GpC sites (20th percentile) within 200 bps of TSS. We used the methylation and accessibility around the TSS to categorize reads into two groups (high and low frequency) for each feature (Supplementary Fig. 6c and Methods). Mean CpG methylation for the two groups was 3% (unmethylated) and 62% (methylated), and GpC groups had mean accessibilities of 20% (inaccessible) and 90% (accessible). Combining the two features resulted in four possible combinatorial epigenetic signatures for each read (Fig. 3b). We observed that with increasing expression, fractions of concordantly active reads (low CpG methylation and high accessibility) increase and conversely inactive (high CpG methylation and low accessibility) reads decrease (Supplementary Fig. 8a and Supplementary Table 6). We also found that genes with euchromatic H3K4me3 histone modification within 1 kb of the TSS have low CpG methylation, and genes with heterochromatic H3K27me3 modification mostly have inaccessible reads (Supplementary Fig. 8b and Supplementary Table 7). Further, the majority of reads on promoter regions with bivalent histone modifications (both H3K4me3 and H3K27me3) have both low CpG methylation and low accessibility, combining the pattern of CpG methylation with H3K4me3 and that of accessibility with H3K27me3.

Using the same subset of genes from the single-read promoter analysis, we examined subnucleosomal footprints within 10 kb of the TSS. We selected regions that have multiple subnucleosomal footprints as candidate protein-binding sites (>210 overlapping and >80 inaccessible runs) (Supplementary Data 3) and performed motif enrichment analysis. Several transcription factors were enriched in the candidate regions, including CTCF, NRF1 and zinc finger proteins, with the strongest enrichment in CTCF-binding sites having a 4× observed to expected ratio. We stratified these reads based on promoter epigenetic signature and calculated the fraction of reads with a subnucleosomal footprint (Supplementary Data 4). In general, reads that had an accessible promoter had a higher fraction of subnucleosomal footprint reads than inaccessible reads, showing that our analysis captures protein-binding events that are associated with active promoter state (Supplementary Fig. 9).
For a specific example, we examined PIM2, a gene that facilitates cell survival and proliferation and is highly expressed in GM12878. PIM2 has a subnucleosomal footprint 1.5 kb downstream of the TSS present only in the reads with an epigenetically active promoter (Fig. 3c). We identified a CTCF-binding motif in this region with a peak in existing CTCF ChIP–seq data. This directly links CTCF binding on the same molecule as an accessible promoter 1.5 kb away.

NanoNOMe identifies allele-specific epigenetic states across the genome. Because nanopore sequencing generates long reads, each read has a greater chance of encountering one or more heterozygous single-nucleotide polymorphisms, which can be used to phase the reads into maternal or paternal origin27. Using existing variant data on GM12878 and both parents 28, we selected heterozygous single-nucleotide polymorphisms and assigned haplotype origin to individual nanoNOMe reads. We were able to confidently determine haplotype assignments on 65% of our sequencing reads; the phased reads covered 86% of the genome to at least ten-times coverage (Extended Data Fig. 8). Using a single assay, we have generated genome-wide allele-specific profiles of DNA methylation.
and accessibility on a human genome. We compared methylation and accessibility near TSSs of autosomal genes, X chromosome inactivated (XCI) genes and X chromosome genes that are known to escape XCI (hereafter referred to as escape genes) using metaplots (Fig. 4a). Genes on the active X chromosome (Xa, maternal allele) were concordantly active with demethylated and accessible promoters and those of inactive X chromosome (Xi, paternal allele) were concordantly inactive with methylated and inaccessible promoters, whereas in autosomal genes and escape genes the two alleles had no significant difference in aggregate (Wilcoxon rank sum test P value > 0.05, Extended Data Fig. 9).

We then found regions that have a significant difference (Benjamini–Hochberg-corrected Fisher’s exact test P value < 0.01 between the two alleles) in methylation or accessibility between

---

**Fig. 4 | Allele-specific methylation and accessibility in GM12878.** a, Methylation and accessibility were separated by parent of origin, and metaplots of TSS methylation and accessibility were generated for each allele. b, Enrichments of DMRs and DARs were calculated at various genomic contexts, showing the enrichment of allele-specific epigenetic patterns in promoters and regulatory elements (Supplementary Table 8). c, Read-level methylation and accessibility plots of ZNF597, an imprinted gene with allele-specific epigenetic patterns. d, Pairwise comparison of CpG methylation in 1kb regions around heterozygous SV breakpoints between alleles with the SV and alleles without the SV, with significance testing by one-direction Wilcoxon rank-sum tests (deletions, $1.3 \times 10^{-6}$ and insertions, $2.0 \times 10^{-5}$).
paternal and maternal alleles, resulting in 9,997 differentially methylated regions (DMRs) and 10,414 differentially accessible regions (DARs) across the genome (Extended Data Fig. 10a and Supplementary Data 5). ATAC-seq is only able to measure allele-specific accessibility at a few of these regions (N = 321, 3% of all DARs), but these did correlate strongly with the nanoNOME accessibility differences (r = 0.76) (Extended Data Fig. 10b). While overlaps between DMRs and DARs were not common (629 overlaps, 6%), they were highly concordant (606 out of 629), that is, an increase in methylation with a decrease in accessibility and vice versa (Supplementary Fig. 10). In the X chromosome, we observed a disproportionate number of hypermethylated Xa DMRs (4,564 hyper- versus 401 hypo-), agreeing with previous findings of Xa hypermethylation (Supplementary Fig. 10c). Further, most (N = 1,050, 80%) DARs had higher accessibility in Xa. We then calculated enrichment of DMRs, DARs and concordant differential regions in different genomic contexts (Fig. 4b and Supplementary Table 8). The enrichment of DMRs with higher Xi methylation near TSS (±500 bps) and the high number of DMRs in gene bodies with hypermethylated Xa agreed with previous findings (Supplementary Fig. 11a and Supplementary Table 8) [30,31]. However, we found that the high number of DMRs in gene bodies was due to their larger size, and hypermethylated Xa DMRs were enriched in enhancers. DARs mostly had higher accessibility in Xa, and this pattern was consistent in all assessed genomic contexts. DARs were enriched in CTCF-binding sites in addition to promoters, suggesting that higher accessibility and consequently increased affinity for CTCF binding, work together to prevent Xc in Xa. Concordant regions with both a DAR and DMR were heavily enriched near TSSs and 90% of them indicated higher activity in Xa (307 out of 339). In autosomes, DMRs, DAR and concordant differential regions all occurred mostly in gene bodies and around TSS, with the highest enrichments around TSSs (Supplementary Fig. 11b,c and Supplementary Table 9).

We then identified genes that had a DMR or a DAR within 500 bps of the TSS (Supplementary Data 6). Of the concordantly differential TSSs, 76% (187) were in the X chromosome and all of these indicated activity in Xa except XIST associated TSSs, a gene known to be specifically active in Xc to promote inactivation of Xo. Out of the 56 autosomal genes, eight were previously identified imprinted genes [32,33]. We plotted ZNF597, one of the eight known imprinted genes, as an example; it had a hypermethylated and less accessible promoter in the maternal copy, indicating that it is active in the paternal allele (Fig. 4c). The ZNF597 gene body exhibited the opposite pattern of methylation, with the active paternal copy fully methylated.

Our long nanopore reads also allow detection of structural variants, large insertions, deletions or transpositions hard to detect with conventional short-read sequencing. We characterized epigenetic consequences of these SVs by comparing epigenetic signals in heterozygous SVs, focusing on large deletions and insertions, the most commonly occurring SV types (Supplementary Data 7). After filtering for heterozygous SVs (Methods), we identified 1,195 deletions and 1,167 insertions, and compared methylation and accessibility near SV breakpoints between the variant and reference alleles (Fig. 4d and Supplementary Fig. 12). Although most of these SVs (80% of deletions and 82% of insertions) do not have a difference in methylation between the alleles, in those that do the variant allele tends to be hypomethylated in deletions (173 hypo versus 65 hyper-) and hypermethylated in insertions (84 hypo versus 131 hyper).

Comparative epigenomic analysis of breast cancer model. Finally, we applied nanoNOME to measure epigenetic differences between three well-characterized breast cell lines: MCF-7 (luminal breast carcinoma, ER+/PR+/HER2−) and MDA-MB-231 (basal breast carcinoma, ER+/PR−/HER2−) as two subtypes of breast cancer, and MCF-10A (fibrocystic disease, ER+/PR−/HER2+) as the normal baseline [34,35]. We generated ≥20x whole-genome coverage of nanoNOME data per cell line (Table 1), and detected DMRs and DARs between normal and cancer cells (Supplementary Data 8). Both of the cancer subtypes had higher numbers of hypomethylated DMRs than hypermethylated DMRs (1.8-fold for MCF-7 and 7.6-fold for MDA-MB-231) suggesting global hypomethylation in the cancer subtypes (Supplementary Fig. 13a,d). While more DARs were accessible in MCF-10A than in the cancer subtypes, the directions were not as skewed as in DMRs (Supplementary Fig. 13b,d). Only a subset of DMRs and DARs coincided at the same genomic loci (8,191 overlapping regions, 11% of DMRs and 6% of DARs), but coinciding DMRs and DARs were highly concordant (r = −0.96, −0.97 and −0.96) (Supplementary Fig. 13c). More concordant regions indicated decreases of accessibility and increase in methylation in the cancer subtypes, especially in MCF-7 (2.6-fold, Supplementary Fig. 13d). Differential epigenic regions were enriched in regulatory regions such as transcription factor binding sites and gene promoter regions (1 kb ±TSS), especially on CTCF-binding sites (Supplementary Fig. 14 and Supplementary Data 9). We specifically inspected the epigenome at the TSS of ER, PR and HER2 as upregulation of these receptors is correlated with cancer aggressiveness (Supplementary Figs. 15–17). We observe a clear change in MCF-7 cells in methylation and accessibility at the PR TSS region and a more subtle change in the ER TSS region, as expected from the upregulation of these genes in that cell line [36].

We also used our long reads to detect SVs and examine the epigenetic features flanking the SV breakpoints calling a total of 18,955 SVs across all three breast lines (Supplementary Table 10 and Supplementary Data 10) [37,38]. Most SVs were singletons (65.9%), with 1,805 SVs occurring in both of the cancer subtypes and not in MCF-10A. While DMRs and DARs were not enriched in regions surrounding SVs (Supplementary Data 9), we did examine the epigenetic state of SVs that occurred only on one cell line (Supplementary Figs. 18 and 19). Within cell lines, we investigated the epigenetic state at heterozygous SVs and found that it was largely the same between the reference allele and the mutated allele (Supplementary Figs. 20 and 21). For example, we found an insertion on chr6:169,976,000 that occurred on both MCF-7 and MDA-MB-231 but not in MCF-10A, which also showed a region 1 kb downstream of the insertion that was hypermethylated and less accessible. These changes in the SV-containing cancer subtypes show a regional correlation between the presence of an insertion and epigenetic alterations (Fig. 5a).

We then used our ability to estimate combinatorial epigenetic states and subnucleosomal footprints on individual reads proximal to genes that were differentially expressed between MCF-10A and MCF-7/MDA-MB-231 (Methods and Supplementary Data 11 and 12). One upregulated gene, ZNF714, has two groups of reads in its TSS for all three cell lines: active (unmethylated and accessible) and inactive (methylated and inaccessible) (Fig. 5b). The two cancer subtypes have more of the reads in the active state, and have subnucleosomal footprints in the promoter, suggesting protein binding. We performed ChiP–quantitative PCR (qPCR) for RNA Pol II Ser5P to see if this represented a transcriptional complex binding and found dramatic enrichment in MCF-7 and MDA-MB-231 compared to MCF-10A (Supplementary Fig. 22). These observations collectively suggest that the upregulation of ZNF714 occurs in conjunction with increased epigenetic activity and subnucleosomal footprints near the TSS.

Discussion

We have leveraged single-molecule nanopore sequencing to directly examine endogenous CpG methylation and chromatin accessibility on long fragments of DNA. Leveraging long reads, we measured epigenetic states at difficult to characterize genomic elements,
for example repetitive elements. In fact, our method is currently limited by the incompleteness of the human genome reference, with large gaps persisting in highly repetitive areas. Long-read technologies have begun to resolve these gaps, enabling nanoNOMe’s interrogation of repetitive regions for example, centromeres. We can also detect SVs with long reads, difficult to detect with short-read sequencing, and examine the epigenome in and around these SVs. Accessibility signals from long reads span multiple protein-binding sites, generating footprints with lengths that allow us to infer the type of protein occupying the region. Using this approach coupled to known CTCF-binding motifs, we have examined the relationship between CTCF-binding and epigenetic patterns of nearby regions. Further, we have combined the ability to predict protein binding with combinatorial promoter epigenetic signatures. We can use these tools to identify differential epigenetic signatures and protein-binding events between different breast cancer cell lines, providing a new window on cancer gene regulation.

With allele-specific data, we can observe parent-of-origin epigenetic features, such as X chromosome inactivation and escape from inactivation, allele-specific activity of imprinted genes and epigenetic differences near heterozygous SVs. We can also explore how imprinting is initiated and controlled by examining the phased epigenome in different tissues and different developmental stages. We can even phase heterozygous mutations with our long reads, and compare the epigenome of alleles with and without mutations. By adding this exogenous layer of information to the DNA itself, we can store information about the epigenetic state of the cell, then read it out along long single molecules using nanopore sequencing. Through the incorporation of additional methyltransferases (for example, EcoGII that methylates adenine to N6-methyladenine), we can take this technique still further, providing a ‘multi-color’ measurement. Paired to longer reads, we can observe further cis interactions of protein-binding and epigenetic states and understand the association of epigenetic features at long genomic distances.

**Online content**
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of...
References

1. Boyle, A. P. et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 132, 311–322 (2008).
2. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).
3. Kelly, T. K. et al. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome Res 22, 2497–2506 (2012).

4. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility and DNA methylation in single cells. Nature 562, 781 (2018).
5. Lai, B. et al. Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. Nature 562, 281–285 (2018).
6. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat. Biotechnol. 37, 925–936 (2019).
7. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods 14, 407–410 (2017).
8. Rand, A. C. et al. Mapping DNA methylation with high-throughput nanopore sequencing. Nat. Methods 14, 411–413 (2017).
9. Shipyoni, Z. et al. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. Nat. Methods 17, 319–327 (2020).
10. Wang, Y. et al. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. Genome Res 29, 1329–1342 (2019).
11. Stergachis, A. B., Debo, B. M., Haugen, E., Churchman, L. S. & Stamatoyannopoulos, J. A. Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science 368, 1449–1454 (2020).
12. Abdullhay, N. J. et al. Massively multiplex single-molecule oligonucleosome footprinting. Preprint at bioRxiv https://doi.org/10.1101/2020.05.20.105379 (2020).
13. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
14. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36, 338–345 (2018).
15. Olova, N. et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. 19, 33 (2018).
16. Ji, L. et al. Methylated DNA is over-represented in whole-genome bisulfite sequencing data. Front. Genet. 5, 341 (2014).
17. Ziebarth, J. D., Bhattacharya, A. & Cui, Y. CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. Nucleic Acids Res. 41, D188–D194 (2013).
18. Gaffney, D. J. et al. Controls of nucleosome positioning in the human genome. PLoS Genet. 8, e1003036 (2012).
19. Valouev, A. et al. Determinants of nucleosome organization in primary human cells. Nature 474, 516–520 (2011).
20. Hesselberth, J. R. et al. Global mapping of protein–DNA interactions in vivo by digital genomic footprinting. Nat. Methods 6, 283–289 (2009).
21. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein–DNA complexes. Genome Biol. 1, REVIEWS001 (2000).
22. Boyle, A. P. et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res 21, 456–464 (2011).
23. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet. 4, e1000138 (2008).
24. Hartl, D. et al. CTCF dinucleotides enhance promoter activity independent of DNA methylation. Genome Res. 29, 554–563 (2019).
25. Pinello, L., Faroumi, R. & Yuan, G.-C. Haystack: systematic analysis of the variation of epigenetic states and cell-type specific regulatory elements. Bioinformatics 34, 1930–1933 (2018).
26. Forneris, O. et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 48, D87–D92 (2020).
27. Gigante, S. et al. Using long-read sequencing to detect imprinted DNA methylation. Nucleic Acids Res. 47, e46 (2019).
28. Eberle, M. A. et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res. 27, 157–164 (2017).
29. Cotton, A. M. et al. Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. Hum. Mol. Genet. 24, 1528–1539 (2015).
30. Hellman, A. & Cherr, A. Gene body-specific methylation on the active X chromosome. Science 315, 1141–1143 (2007).
31. Sharp, A. J. et al. DNA methylation profiles of human active and inactive X chromosomes. Genome Res. 21, 1592–1600 (2011).
32. Jirtle, R. L. Genomic imprinting and cancer. Exp. Cell. Res. 248, 18–24 (1999).
33. Morison, I. M., Ramsay, J. P. & Spencer, H. G. A census of mammalian imprinting. Trends Genet. 21, 457–465 (2005).
34. Holloway, D. L. & Speirs, V. Choosing the right cell line for breast cancer research. Breast Cancer Res. 13, 215 (2011).
35. Messier, T. L. et al. Histone H3 lysine 4 acetylation and methylation dynamics define breast cancer subtypes. Oncotarget 7, 5094 (2016).
36. Subik, K. et al. The expression patterns of ER, PR, HER2, CK5/6, EGFR, Ki-67 and AR by immunohistochemical analysis in breast cancer cell lines. Breast Cancer 4, 35–41 (2010).
37. Sedlacek, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods 15, 461–468 (2018).
38. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. 10, 14061 (2017).
39. Valouev, A. et al. Determinants of nucleosome organization in primary human cells. Nature 474, 516–520 (2011).
40. Hesselberth, J. R. et al. Global mapping of protein–DNA interactions in vivo by digital genomic footprinting. Nat. Methods 6, 283–289 (2009).
41. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein–DNA complexes. Genome Biol. 1, REVIEWS001 (2000).
42. Boyle, A. P. et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res 21, 456–464 (2011).
Methods

GpC methylation model generation for nanopolish. Along with the GpC methylation model, the GpC methylation model was also regenerated to ensure the validity of the method for model generation. dGNA from E. coli K12 MG1655 (ATCC 700926DQ) and genomic DNA from GM12878 lymphoblasts (Coriell Institute) were first sheared to an average fragment size of 8 kb using G-tubes (Covaris catalog no. 520079). The fragmented DNA was PCR amplified to generate unmethylated DNA using the first steps of low input ligation kit SQR-LWP001 (ONT). Samples were end-repaired, deoxyadenosine(dA)-tailed and ligated to amplification adaptors, followed by 11 cycles of PCR amplification. The resulting unmethylated, sheared DNA was methylated with M. SssI (New England Biolabs (NEB) catalog no. M0226) for GpC methylation or M. CviPI (NEB catalog no. M0227) for GpC methylation or both enzymes for GpC+GpC methylation. Two cycles of 4-h methylation were performed for each sample, and for each cycle of treatment the enzyme and methyl donor (S-adenosylmethionine) were replenished at the 2h mark.

Validation of DNA methylation by bisulfite sequencing. Near-complete methylation in the training samples (E. coli) and testing samples (GM12878) were validated by performing WGBS on the Illumina MiSeq platform. NEBNext Ultra library preparation kit (NEB catalog no. E7370) and Zymo EZ DNA methylation-lightning kit (Zymo catalog no. D5300) were used to generate the bisulfite sequencing libraries. DNA from each sample was shared to 300bp fragments using BioRuptor Pico (Diagenode), followed by end-repair and dA-tailing. Methylated universal adaptor (NEB catalog no. E7535) was ligated using the Blunt/TA Kit. The adaptors ligated samples were end-repaired, bisulfite-converted, quenched, and cleaned-up before PCR amplification with multiplexing primers and uracil-tolerant Taq polymerase (KAPA HiFi Uracil+ (Roche catalog no. KK2801)). The resulting DNA sequencing library was sequenced on an Illumina MiSeq device using V2 300-cycle chemistry.

The resulting fastq files were preprocessed by removing adaptor sequences and trimming low quality 3’ ends using Trim Galore v.0.8.3 (https://github.com/FelixKrueger/TrimGalore) with default parameters. Then, data were analyzed using Bismark v.0.19.0 (ref. 37). After alignment, PCR duplicates were removed using Picard tools MarkDuplicates module v2.20.2 (http://broadinstitute.github.io/picard/). Reads were truncated at the 3’ end by two bases at the 5’ end and one base at 3’ end to minimize methylation bias at the ends of reads introduced during the library preparation. The total number of methylated cytosine residues and unmethylated cytosine residues were counted to calculate methylation percentages of the samples.

Cell culture. GM12878 lymphoblast cells were obtained from Coriell Institute and MCF-10A, MCF-7 and MDA-MB-231 breast cells were obtained from ATCC. GM12878 were grown in RPMI 1640 medium (Gibco catalog no. 11875119) supplemented with 15% FBS (Gibco catalog no. 10098037) and 1% penicillin-streptomycin (Gibco catalog no. 15140122). MCF-10A were grown in DMEM F-12 medium (Gibco catalog no. 11320033) supplemented with 10% FBS and 1% penicillin-streptomycin. MCF-7 and MDA-MB-231 were grown in DMEM (Gibco catalog no. 11965118) supplemented with 100 ng ml⁻¹ hydrocortisone (Sigma Aldrich catalog no. H0135) and 1% penicillin-streptomycin.

Nuclease footprinting via GpC methyltransferase. NOME-seq was performed on the cells with adjustments for nanopore sequencing. Cells were collected by trypsinization, then nuclei were extracted by incubating in resuspension buffer (100 mM Tris-Cl pH 7.4, 100 mM NaCl, 30 mM MgCl₂) with 0.25% NP-40 for 5 min on ice. Insoluble debris were collected by centrifugation for 5 min at 500g at 4°C. Nuclei were subjected to a methylation labeling reaction using a solution of 1x M. CviPI Reaction Buffer (NEB), 300 mM sucrose, 96μM S-adenosylmethionine (NEB) and 200 μM of M. CviPI (NEB) in 500 μl volume per 500,000 nuclei. The reaction mixture was incubated at 37°C with shaking on a thermomixer at 1,000rpm for 15 min. S-Adenosylmethione was replenished at 96μM at 7.5 min into the reaction. The reaction was stopped by the addition of an equal volume of stop solution (20mM Tris-Cl, pH 7.9, 600 mM NaCl, 1% SDS, 10 mM disodium EDTA). Samples were treated with proteinase K (NEB) at 55°C for >2h and DNA was extracted via phenol:chloroform extraction and ethanol precipitation. After proteinase K treatment, and in all following steps, samples were handled with care using large orifice pipette tips to avoid excessive fragmentation of DNA.

Nanoformat sequencing. Purified gDNA was prepared for nanopore sequencing following the protocol in the genomic sequencing by ligation kit LSQ-SQK108 (ONT). Samples were first sheared to roughly 10 kb using G-tubes (Covaris); by centrifuging 2–3 mg of fragmented gDNA at 8,000 rpm for 1 min, then adding the tube and centrifuging again. We sheared the DNA to 10 kb because it produces long fragments of DNA while maximizing the yield of nanopore sequencing. Shearing to larger sizes or unsheared DNA may be used to maximize the length of sequenced reads, with the caveat that sequencing yield will drop. In two samples (GM12878 samples 8 and 9), we targeted 20 kb fragments, with an additional step of removing short fragments using the Short Read Eliminator module by Circulomics, following the manufacturer’s specifications. The sheared samples were end-repaired and dA-tailed using NEBNext Ultra II end-repair module (NEB), followed by clean-up using 1X v/v AMPure XP beads (Beckman Coulter). Sequencing adaptors, comprising linker adaptor DNA and motor proteins, were ligated to the end-prepared DNA fragments using Blunt/TA Ligase Master Mix (NEB), followed by clean-up using 0.4x v/v AMPure XP beads and sequencing kit reagents. Libraries were sequenced using the flow cell were loaded onto FLO-MIN106 or PRO-002 flowcells and run on MinION Mk1b, GridION or PromethION sequencers for up to 72h. Data were collected by MinKNOW v.1.14.

Data preprocessing (basecalling, alignment, methylation calling and structural variant calling). Raw current signals were converted to DNA sequences using Gelignment (v. 3.0.3 (ONT)), using the high-accuracy basecalling model (hhmm). gDNA sequences were aligned to hg38 human reference genome without alternative contigs using NGMLR v.0.2.8 with default settings for aligning Oxford nanopore reads (x ont)⁴. GpC and GpC methylation were called using nanopolish v.0.11.1. We used Sniffles v.1.0.11 (ref. 39) with default parameters to detect SVs across each sample and SURVIVOR v.1.0.7 (ref. 40) to obtain a multi-sample VCF file.

Nanopolish methylation training for dual CpG/GpC methylation. To train the methylation calling models we generated nanopore sequencing data for E. coli gDNA treated with M.SssI (to methylate CpGs), M.CviPI (to methylate GpCs), and both M.SssI and M.CviPI (in both contexts methylated). The three datasets were basecalled with Guppy (v.3.0.3) and aligned to the E. coli genome using NGMLR v.0.2.8. The reference genomes for each dataset were then modified by converting Ca to Ms in the appropriate context. We then merged the three reference genomes and three BAM files together and downsampled the alignments to 10% coverage to reduce model training time in the subsequent step. As a part of this preprocessing, we had a dataset with a mixture of reads that have CpG methylation, CpC methylation or both, and matching reference sequences to align each read to indicate the pattern of methylation in each read.

The k-mer states for the GpG/GpC model were trained using ‘nanopolish train’, a new model training method that implements the forward/backward algorithm on a simplified hidden Markov model to calculate the posterior probability of an observed event originating from a k-mer of the reference sequence. The Gaussian distributions were fit as previously described except each observation was weighted by the posterior probability, and we fitted a single Gaussian in all cases rather than Gaussian mixtures. The complete training code is provided in the cppgc_new_train branch (commit c409598) of the nanopolish github repository. Model training was run for ten iterations and the final model was used for the subsequent methylation calling.

Validation of nanopolish calling for dual CpG/GpC methylation. We used the GM12878 methylation controls as the testing set to validate the methylation calling model and choose an appropriate threshold for calling methylation. Receiver operating characteristic curves were generated by applying a range of LLR thresholds to bin the continuous LLR into the binary state of methylated/ unmethylated calls and comparing to the true state for each singularly methylated and the unmethylated data. To choose the LLR for calling CpG methylation, we sampled the LLRs of both the E. coli and GM12878 methylation controls based on the true states, into distributions of methylated CpG calls and unmethylated CpG calls. We then chose the LLR that would allow 5% false calls (that is, the top 5th percentile of LLRs for unmethylated calls and bottom 5th percentile of LLRs for methylated). Then, to make the thresholds symmetric between methylated and unmethylated calls, we averaged the absolute value of the two thresholds for methylated and unmethylated calls, we averaged the absolute value of the two thresholds and applied a ceiling function to the nearest half. As a result, we chose a threshold of 1.5 for calling CpG methylation (LLR <−1.5 is unmethylated and >1.5 is methylated, and values between are uncalled). We set a threshold of 1 for GpC methylation in a similar way.

Accessibility peak finding, one-sample comparison of methylation and structural accessibility, and enrichment analysis of differential epigenetic regions on genomic contexts. For comparison and visualization of bulk methylation and accessibility, estimated profiles of measurements were calculated by fitting locally weighted generalized linear models across the genome for each sample and SURVIVOR v.1.0.7 (ref. 40) to obtain a multi-sample VCF file.
as implemented by biocorporate packages bssge v.1.2.0 (ref. 9). For each CpG methylation site, data in a window of at least 1,000 bps and 50 nearby sites with a maximum gap of 100 kb with tri-cube kernel weights were used to estimate the smoothed methylation frequency. For CpG methylation, the minimum window and number of sites were arbitrarily reduced to 100 bps and ten nearby sites to account for more rapid fluctuations in the accessibility profile due to nucleosome positioning.

To find regions of high accessibility, contiguous regions having smoothed accessibility greater than 99th percentile of the data were selected first. The significance of each accessible region was determined by performing a binomial test of the raw frequency of accessibility, with overall accessibility frequency as the null probability. The probabilities were corrected for multiple testing using the Benjamini–Hochberg correction, and accessibility regions with adjusted p values less than 0.01 and widths greater than 50 bps were determined to be accessibility peaks. Similarly, for DARs, we performed a one-sided Fisher’s Exact test on raw counts of accessible and inaccessible calls. p values were corrected using Benjamini–Hochberg correction and regions with adjusted p values less than 0.01 were determined to be DARs.

To find DMRs between two samples without replicates, the difference of methylation between the two samples was calculated for each CpG site. Then, contiguous regions with differences greater than 99th percentile of the differences were selected as candidates for hypermethylation, and regions with differences less than were first selected as candidates for hypomethylation. Similarly, for DARS, we performed a one-sided Fisher’s Exact test on raw counts of methylated and unmethylated calls on each candidate DAR. p values were corrected using Benjamini–Hochberg correction, and regions with adjusted p values less than 0.01 and widths greater than 100 bps were determined to be significant DMRs.

To calculate the enrichment of DMRs, DARS and concordantly differential regions in various genomic contexts, we first calculated the total width of the genome and the total width of the genomic contexts of interest that contain CpG and GpC data. Then the total number of differential regions was divided by the total width of the genome, which is the expected abundance of differential regions. This was used as the baseline against the total numbers of differential regions in genomic contexts of interest divided by the total widths of the genomic contexts to generate the final values of enrichments. For TSS and small transcription factor binding sites, we used 1000 bps regions centered on the genomic elements.

Comparison of nanoNOMe with conventional methodologies. Bulk next-generation sequencing methodologies comparable to nanoNOMe on GM12878 were used to compare and validate nanoNOMe. WGBS methylation frequencies were obtained from ENCODE accession ENCF8835NTC, normalized MNase-seq signals were obtained from ENCODE accession ENCSR000XCP and normalized DNase-seq signals were obtained from ENCODE accession ENCSR009SEID (ref. 9). ATAC-seq data were obtained from Gene Expression Omnibus ( GEO) accession GSE47753 (ref. 15) and processed using the standard ENCODE pipeline v.78 (ref. 9). Nanopore whole-genome sequencing data were selected as candidates for hypermethylation, and regions with adjusted p values less than 0.01 were determined to be significant DARs.

To discriminate CTCF-binding events from nucleosome-binding events on individual reads, we used lengths of inaccessible runs on centers of CTCF-binding motifs. Lengths were chosen based on the following. To determine an expectation-maximization algorithm implemented by R package Mclust v.5.4.5 (ref. 9). The optimal clustering parameters were determined based on maximum integrated complete-data likelihood and the cluster that had the smallest mean length (54 bps) was chosen as CTCF-binding signal and the other clusters as units of nucleosome-binding signals. This model was applied to classify all inaccessible runs within 25 bps of CTCF-binding sites as CTCF-bound or nucleosome-bound and reads that contained CTCF-bound inaccessible runs were considered to be CTCF-bound reads. To predict protein-binding events outside CTCF-binding sites, we used the model on all inaccessible runs to categorize them to one of the clusters, using the length of the inaccessible run (subnucleosomal footprints) as candidates for protein binding. We selected regions that contained at least ten candidates as the predicted regions of protein-binding events.

To predict combinatorial epigenetic signatures of individual reads on TSS, we used methylation and accessibility in a window around the TSS. On each read, we averaged CpG methylation in the region of each TSS within 10 kb of the TSS. To compare mappability between WGBS, nanopore whole-genome sequencing and nanoNOMe, the numbers of reads aligning to 200 bps bins of the genome were calculated. GC bias of the coverages were determined by calculating the percentages of CG for each of the 200 bps bins and plotting the per-bin coverage against the CG percentage. To compare mappability in specific genomic contexts, a region was considered to be robustly mapped in a dataset if its coverage was within the 5th and 95th percentile of the genome-wide binned coverage. The upper threshold takes into account aberrantly highly mapped regions, while the lower threshold removes low mappability regions.

For comparison of nanoNOMe signals with conventional bulk methods, average methylation was calculated for each CpG and GpC site. To compare nanoNOMe to WGBS, the CpG methylation frequencies for each CpG focus across the genome were compared pairwise between the two methods. To compare nanoNOMe CpG accessibility signal to normalized ATAC-seq and DNase-seq signals, the intersections were determined from accessibility peaks of nanoNOMe, ATAC-seq and DNase-seq.

Methylation frequencies from WGBS and normalized MNase-seq signals at regions surrounding genomic features of interest (CTCF, TSS with respect to expression and histone modifications) were extracted for the generation of the metaplots. For each genomic feature, average methylation frequency and accessibility were aggregated with respect to distance from the feature, following by taking the rolling average with a window of 50 bps. Known TSS and CGI were obtained from ENCODE accession ENCSR000AKB and ENCSR000AKD (ref. 9). TSS was defined as promoted and CGI as subnucleosomal footprint regions. We then separated the reads based on the epigenetic signature of nearby gene promoter(s), and separately assessed the fraction that contained CTCF-binding event at the subnucleosomal footprint region for each group, resulting in protein-bound reads specific to each promoter epigenetic signature. The combinatorial epigenetic signature of reads was determined based on the combination of the cluster assignments.

Read-level subnucleosomal footprint measurement and promoter epigenetic signature estimation were coupled by first estimating regions of protein binding within 10 kb of TSS of a subset of genes. Sites that have ten or more reads with short inaccessible runs in a window less than 80 bps were selected as subnucleosomal footprint regions. We then separated the reads based on the epigenetic signature of nearby gene promoter(s), and separately assessed the fraction that contained CTCF-binding event at the subnucleosomal footprint region for each group, resulting in protein-bound reads specific to each promoter epigenetic signature. The transcription factor enrichment of subnucleosomal footprints was performed using Haystack against the JASPAR transcription factor database (ref. 9).

Haplotype assignment and allele-specific methylation analysis. We obtained genotype information for GM12878 from existing phased Illumina platinum genome data generated by deep sequencing of the cell donors’ familial trio (ref. 9). The bcftools package was used to filter for only variants that are heterozygous in GM12878. The heterozygous GM12878 SNVs were used to identify reads with allele-informative variants and assign the parent of origin for each read using WhatHap v.0.18 (ref. 9). Methylation and accessibility calls on each read were separated based on the haplotype assignments to generate allele-specific profiles of methylation and accessibility.

To identify accurate heterozygous SNVs, we called SNVs on the two alleles separately using Snuﬄes v.1.0.11 and SURVIVOR v.1.0.7 as described above. From the SNVs that were detected heterozygously in methylated SVs by detecting the SNVs that have less than two nonvariant and more than 20 variant reads on only one of the alleles. To remove SVs that are short in length or affected by incorrect alignments, we removed SVs that are shorter than 200 bps and have more than 100 read alignments in one allele. After detecting DMRs and DARS as described above between the two alleles, the DARS were compared with allelic imbalance in GM12878 ATAC-seq (GSE47753). ATAC-seq reads in and around the allele-specific DARS were phased using WhatHap v.0.18 and the annotated GM12878 SNVs. The log of ratio of covariances between alleles in ATAC-seq were compared to the difference in accessibility frequency between alleles, and their correlation was verified by a Pearson correlation coefficient.

Breast cancer cell line analysis. RNA-seq counts of the three cell lines were downloaded from GEO accession GSE75168 (ref. 9) and analyzed using the biocorporate package DESeq2 v.1.24.0 (ref. 9). Using default parameters, differential expression analysis was conducted based on the negative binomial
distribution, comparing MCF-7 and MDA-MB-231 to MCF-10A. Genes were considered to be significantly differentially expressed when the Benjamini–Hochberg corrected P values were less than 0.01.

Of this differentially expressed set, we filtered for genes with five or more differences between normal and cancer lines in the number of reads indicating epigenetically concordant active promoters. Then subnucleosomal footprints were compared on predicted protein-binding regions within 10 kb of the TSS of these genes, and the fraction of reads that have a footprint were calculated for each sample. Those regions that had a difference of footprint fraction $\geq 0.4$ were selected as genes with differences in the promoter epigenetic signatures and protein binding.

Chromatin immunoprecipitation analysis. Cell lines (MCF-10A, MCF-7 and MDA-MB-231) were fixed for 8 min in 1% formaldehyde (Thermo, 28906), and the reaction stopped by quenching the reaction with 2.5 M glycine (SigmaAldrich) to approximately 1-kb fragments. Sheared DNA was clarified for debris by addition of IGEPAL detergent (final concentration 1%), and centrifugation at 15,000 $g$ for 10 min at 4 °C. Then, 100 μl of ProteinG Dynabeads (ThermoFisher, 10003D) were washed in 0.5% BSA/PBS solution, before being incubated overnight with 10 ng of antibody (anti-RNAPser5P; Abcam 5131). Sheared chromatin was incubated with the antibody-bound Dynabeads overnight at 4 °C. Background DNA was removed by four rinses with wash buffer (50 mM HEPES pH 7.5, 1 mM EDTA, 0.7% sodium deoxycholate, 1% IGEPAL, 0.5 M LiCl) and one rinse with hi-salt buffer (10 mM Tris pH 8, 1 mM EDTA, 50 mM NaCl). DNA was eluted from beads by 10 min incubation at 65 °C in elution buffer (50 mM Tris pH 8, 10 mM EDTA, 1% SDS) and cross-links reversed by overnight incubation at 65 °C. Protein was digested by a 2 h proteinase K treatment (final concentration of 0.2 μg/μl) at 55 °C. DNA was isolated using a MinElute kit (Qiagen, 28004). Samples were analyzed using real-time qPCR (Stratagene MX2005P, Agilent) with Fast SYBR Green master mix (Applied Biosystems, 4385612), and normalized against 5% input. The primers used for qPCR amplification were as follows (all 5′ to 3′): GAPDH (fwd:TACTAGCGGTTTTACGGGCG, rev:GGTGGGTAGCAGATGCTC), IL-12a (fwd:GGCACTCCTCTCTCTACGTC, rev:AGCTCGGCCCAAAATGAAAG) and ZNF714 (fwd:TCCCCTTCAGCCATAAGATG, rev:CACGGACCCATTCATAAACC).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
NanoNOMe data for GM12878, MCF-10A, MCF-7 and MDA-MB-231 are available at National Center for Biotechnology Information Bioproject ID PRJNA510783 (http://www.ncbi.nlm.nih.gov/bioproject/510783). Processed single-read data in select regions are deposited in Zenodo (https://zenodo.org/record/3969567) and processed methylation frequency files are available in GEO accession GSE155791.

Code availability
Source code for analysis is available at https://github.com/timplab/nanoNOMe.

References
40. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfi-seq applications. Bioinformatics 27, 1571–1572 (2011).
41. Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulftite sequencing reads to differentially methylated regions. Genome Biol. 13, R83 (2012).
42. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. J. R. Stat. Soc. 289–317 (2016).
43. Patterson, M. et al. Whatshap: weighted haplotype assembly for future-generation sequencing reads. J. Comput. Biol. 22, 498–509 (2015).
44. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).

Acknowledgements
This study was supported by National Human Genome Research Institute (project no. SR01HG009190).

Author contributions
I.L. and W.T. conceived the study. I.L., T.G. and N.S. acquired data. I.L., R.R., T.G., M.M., A.G., F.J.S., K.D.H., J.T.S. and W.T. analyzed and interpreted data. I.L. and W.T. wrote the paper.

Competing interests
W.T. has two patents (8,748,091 and 8,394,584) licensed to Oxford Nanopore Technologies. I.L., T.G., N.S., F.J.S., J.T.S. and W.T. have received travel funds to speak at symposia organized by Oxford Nanopore Technologies. J.T.S. received research funding from Oxford Nanopore Technologies.

Additional information
Extended data is available for this article at https://doi.org/10.1038/s41592-020-01000-7, Supplementary information is available for this article at https://doi.org/10.1038/ s41592-020-01000-7. Correspondence and requests for materials should be addressed to W.T.

Peer review information Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. Nature Methods thanks Jeff Vierstra and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Assessment of CpG and GpC dual methylation calling. The ability of nanopore sequencing to distinguish cytosine methylation at CpG and GpC contexts is shown by a, examining current level shifts depending on the placement of the methylation on a 6-mer (n = 256 unique 6mers for each group). Data are presented as median values, interquartile range (IQR), and 1.5X IQR. The performance of the methylation caller was validated by b, measuring methylation frequencies for calling methylation in samples treated by methyltransferases.
Extended Data Fig. 2 | Bulk NanoNOMe profiles at CTCF binding sites. Metaplots of a, methylation and b, accessibility as a function of distance to CTCF binding motifs in nanoNOMe, WGBS, and MNase-seq agree very closely.
Extended Data Fig. 3 | Pairwise Comparison of Methylation and Accessibility at Gene Promoters. Pairwise scatter plot of average CpG methylation to GpC accessibility for 400 bp regions centered at each gene TSS, colored by its gene expression quartile.
Extended Data Fig. 4 | GpC accessibility kernel estimation on single reads. GpC methylation calls were smoothed using a Gaussian kernel estimator. 

a, Distributions of length of accessible and inaccessible runs and b, metaplot of accessibility near CTCF binding sites before and after the smoothing, along with (c) example of read-level plot of accessibility from a 2 kb region around a CTCF binding site.
Extended Data Fig. 5 | Single-read epigenetic assessment on CTCF binding sites. a, Heatmaps of lengths of runs of accessible chromatin calls on individual reads with respect to distance from CTCF binding sites, separated based on presence of ChIP-seq peaks. b, Density distributions of inaccessible runs at the CTCF binding sites, showing that sites without CTCF binding have long inaccessible runs suggesting nucleosome binding while those with CTCF binding have short inaccessible runs (sub-nucleosomal footprints) suggesting CTCF binding. c, Inaccessible runs were classified as either sub-nucleosomal or nucleosome binding depending on their lengths based on mixed Gaussian models.
Extended Data Fig. 6 | CTCF binding classification. Single-read (a) methylation and (b) accessibility plots on a CTCF binding motif, clustered by the presence of sub-nucleosomal footprint at the binding motif, predicted as events of CTCF protein binding.
Extended Data Fig. 7 | Comparison of protein binding prediction with ChIP-seq. a, The fractions of CTCF-bound reads determined by sub-nucleosomal footprints were compared with ChIP-seq coverage enrichments per CTCF binding motif, showing that the ChIP-seq signal tends to increase with CTCF binding fraction, and b, the distributions of the fractions were stratified by binding motifs with ChIP-seq peaks to those without peaks, showing that sites with ChIP-seq peaks have higher fractions of CTCF binding. Data are presented as median values, interquartile range (IQR), and 1.5X IQR, as well as density distributions.
Extended Data Fig. 8 | Haplotype phasing results on GM12878 nanoNoMe data. a. The number of reads that could be phased into maternal or paternal read based on the presence of heterozygous SNV in the read, showing that 65% of reads could be phased. b. The fractions of the chromosomes that could be phased (the fraction that had > 10x coverage on each allele after phasing) shows on average, 86% of the genome could be phased.
Extended Data Fig. 9 | X-chromosome inactivation promoter comparisons. Methylation and accessibility in 500 bp and 100 bp windows, respectively, centered at TSS compared between maternal and paternal alleles (N = number of genes in the group), a, by plotting and comparing the distributions using boxplots and one-sided Wilcoxon rank-sum test (Data are presented as median values, interquartile range (IQR), and 1.5X IQR, CpG XCI Pat > Mat p-value = 0, GpC XCI Mat > Pat p-value = 1.9e-229), and b, by density plots of the difference in methylation frequencies between the two alleles.
Extended Data Fig. 10 | Differentially methylated and differentially accessible regions between alleles in GM12878. Methylation was compared between the two alleles across the genome to find regions of significant difference and were tested using one-sided Fisher’s exact test, and accessibility peaks were compared by 1) finding peaks of accessibility on each allele separately, 2) selecting peaks that occur exclusively in one allele, 3) and comparing the accessibility frequency between the two alleles in these candidate regions. The detected DMRs and DARs are a, shown as volcano plots, with dashed lines representing thresholds for considering the region as DMR/DAR. b, Examining existing (GEO Accession GSM1155957) ATAC-seq data, we compared allele specific accessible in ATAC-seq peaks that overlapped with a heterozygous SNP. In the 321 DARs detectable via ATAC-seq, we saw high correlation with nanoNOMe (r = 0.76).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a  Confirmed

☐  The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐  A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☒  The statistical test(s) used AND whether they are one- or two-sided

☐ Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐  A description of all covariates tested

☐  A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐  A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐  For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

Give P values as exact values whenever suitable.

☐  For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐  For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒  Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection  minKNOW version 1.14

Data analysis  Trim Galore v0.6.3, Bismark v0.19.0, Picard v2.20.2, Guppy v3.0.3, NGMLR v0.2.8, nanoplot v0.11.1, Sniffles v1.0.11, SURVIVOR v1.0.7, biseq version 1.20.0, Mclust version 5.4.5, WhatsApp version 0.18, ENCODE Pipeline v78, DESeq2 version 1.24.0

Custom code available in Code Accessibility section at https://github.com/timplab/nanoNOMe

For manuscripts utilizing custom algorithms or software that are not central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

NanoNOMe data of GM12878, MCF-10A, MCF-7, and MDA-MB-231 are available at NCBI Bioproject ID PRJNA510783 (http://www.ncbi.nlm.nih.gov/bioproject/510783). Processed single-read data in select regions are deposited in Zenodo (https://zenodo.org/record/9969567) and processed methylation frequency files are available in GEO accession GSE135579.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | Only single cell lines were used to demonstrate the method |
|-------------|---------------------------------------------------------|
| Data exclusions | No data were specifically excluded from the analyses |
| Replication | Data was compared to gold standard methods and data from ENCODE |
| Randomization | Not relevant as we were running individual cell line samples. |
| Blinding | Blinding was not possible as this was a demonstration of methods on cell lines. Our analysis was performed using appropriate control samples. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

- Antikörpers
- Eukaryotic cell lines
- Palaeontology and archaeology
- Animals and other organisms
- Human research participants
- Clinical data
- Dual use research of concern

### Methods

| Sample size | Involved in the study |
|-------------|------------------------|
| Cell line source(s) | GM12878 (Coriell), MCF10A, MCF-7, MDA-MB-231 (ATCC) |
| Authentication | Cell lines obtained directly from ATCC and Coriell and authenticated by them. |
| Mycoplasma contamination | Cell lines not tested for mycoplasma |
| Commonly misidentified lines (See ICLAC register) | None of these lines are on the ICLAC register. |