nature biotechnology

LETTERS

https://doi.org/10.1038/s41587-020-0407-5

Targeted nanopore sequencing with Cas9-guided adapter ligation

Timothy Gilpatrick[1], Isac Lee[1], James E. Graham[2], Etienne Raimondeau[2], Rebecca Bowen[2], Andrew Heron[2], Bradley Downs[3], Saraswati Sukumar[3], Fritz J Sedlazeck[4] and Winston Timp[1,5*]

Despite recent improvements in sequencing methods, there remains a need for assays that provide high sequencing depth and comprehensive variant detection. Current methods[1–4] are limited by the loss of native modifications, short read length, high input requirements, low yield or long protocols. In the present study, we describe nanopore Cas9-targeted sequencing (nCATS), an enrichment strategy that uses targeted cleavage of chromosomal DNA with Cas9 to ligate adapters for nanopore sequencing. We show that nCATS can simultaneously assess haplotype-resolved single-nucleotide variants, structural variations and CpG methylation. We apply nCATS to four cell lines, to a cell-line-derived xenograft, and to normal and paired tumor/normal primary human breast tissue. Median sequencing coverage was 675× using a MinION flow cell and 34× using the smaller Flongle flow cell. The nCATS sequencing requires only ~3 µg of genomic DNA and can target a large number of loci in a single reaction. The method will facilitate the use of long-read sequencing in research and in the clinic.

Targeted sequencing allows investigators to enrich for loci of interest, reducing sequencing costs and labor to achieve high coverage data at desired genomic regions. Deep sequencing coverage is critical for interrogation of methylation patterns or mutation frequency in heterogeneous clinical samples. For next-generation sequencing, leading strategies are amplification or hybridization capture[5], but these do not take advantage of the benefits of newer long-read sequencing technologies, because amplification loses any base modifications present and hybridization capture has yet to be fully optimized for long fragments. Some approaches for long-read enrichment have used PCR to amplify regions of interest and then either sequenced the amplicons directly[1] or cloned into expression plasmids[2] before sequencing, but both of these strategies can be affected by amplification bias and lose information about modified nucleotides. Another method described for target enrichment with nanopore sequencing is Cas9-assisted targeting of chromosome segments (CATCH-seq)[3], wherein regions of interest are excised by dual Cas9 cleavage and then enriched by size selection, but the low recovery from this method required amplification for enrichment from the human genome. Most recently, the enrichment field has had a burgeoning interest in ligating sequencing adapters to cuts with Cas endonucleases, as demonstrated by others for enrichment and methylation status evaluation of the C9orf72 locus[4].

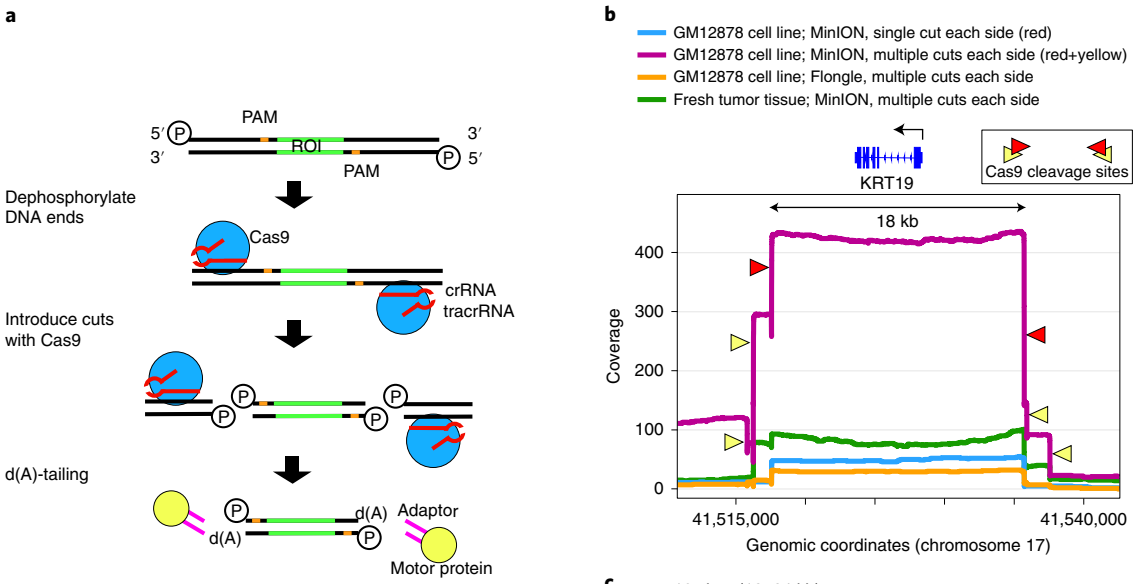
For this work, we enriched by selectively ligating sequencing adapters to fresh cut sites created by an active Cas9/guide (g) RNA–ribonucleoprotein complex (RNP). By dephosphorylating pre-existing DNA ends before cutting with Cas9, we preferentially ligated to the newly produced DNA ends at Cas9 cleavage sites (Fig. 1a). We first validated and tested the nCATS method by comparing our data on the well-characterized GM12878 cell line to both annotated variants[6] and whole-genome bisulfite methylation data[7]. We then applied the enrichment strategy to assess genetic and epigenetic changes in breast cell lines, a breast cancer cell line xenograft and primary patient tissue.

After cleavage with Cas9, the enzyme remains bound to the DNA on the 5′-side of the gRNA[8], resulting in preferential ligation of adapters on to the 3′-side of the cut. We introduced cuts flanking regions of interest (ROIs) to achieve coverage on both strands (Fig. 1a). We targeted 10 sites in our initial panel, with sizes ranging from 12 kb to 24 kb (see Supplementary Table 1). For evaluating single nucleotide mutations, we selected three cancer-associated genes (*TP53*, *KRAS* and *BRAF*) with annotated mutations in the MDA-MB-231 cell line[9]. Five regions for methylation studies (*KRT19*, *SLC12A4*, *GSTP1*, *TPM2* and *GPX1*) and two candidate large deletions (6–8 kb) were selected based on previous whole-genome nanopore data from our lab[10], as well as existing expression data in these breast cell lines[11]. An eleventh region, the *BRCA1* locus, was included in some sequencing runs (see Supplementary Table 1) to test our ability to capture larger regions (>80 kb), and to evaluate this method for sequencing repetitive regions[12].

In our initial experiments, we used one gRNA on either side of each region. We applied this to four cell lines: the well-characterized GM12878 lymphoblast cell line and three breast cell lines (MCF-10A, MCF-7 and MDA-MB-231). Libraries were prepared from 3 µg of starting DNA and run on a MinION flow cell, resulting in coverage ranging from 18× to 846× (see Supplementary Table 2). When libraries were run on the smaller Flongle flow cell, we measured coverage between 8× and 65× (see Supplementary Table 2).

We attributed the variable coverage between regions to differing on-target cutting efficiency and off-target binding of the different gRNAs. Subsequently, we experimented with a combination of multiple gRNAs at the same locus in GM12878, and found this notably improved median coverage. For example, at the *KRT19* locus, coverage with multiple guides increased to 407× versus 47× with single guides (Fig. 1b). Using multiple guides at all loci yielded >400× at all sites from the MinION flow cell (median 680×) and >25× at all sites from a Flongle flow cell (median 34×) (Fig. 1c, and see Supplementary Table 2).

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [2]Oxford Nanopore Technologies, Oxford, UK. [3]Department of Oncology, Johns Hopkins School of Medicine, Baltimore, MD, USA. [4]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [5]Department of Molecular Biology and Genetics, Department of Medicine, Division of Infectious Disease, Johns Hopkins School of Medicine, Baltimore, MD, USA. *e-mail: wtimp@jhu.edu



c

10 sites (12–24 kb)

	GM12878			Breast tumor
	One cut each side	Multiple cuts (two or three) each side		
	MinION	MinION	Flongle	minION
Aligned reads	75,822	235,328	37,715	611,992
On-target reads	1,370	10,854	790	2,702
On-target percentage	1.81	4.61	2.09	0.44
Median coverage	80	680	34	93

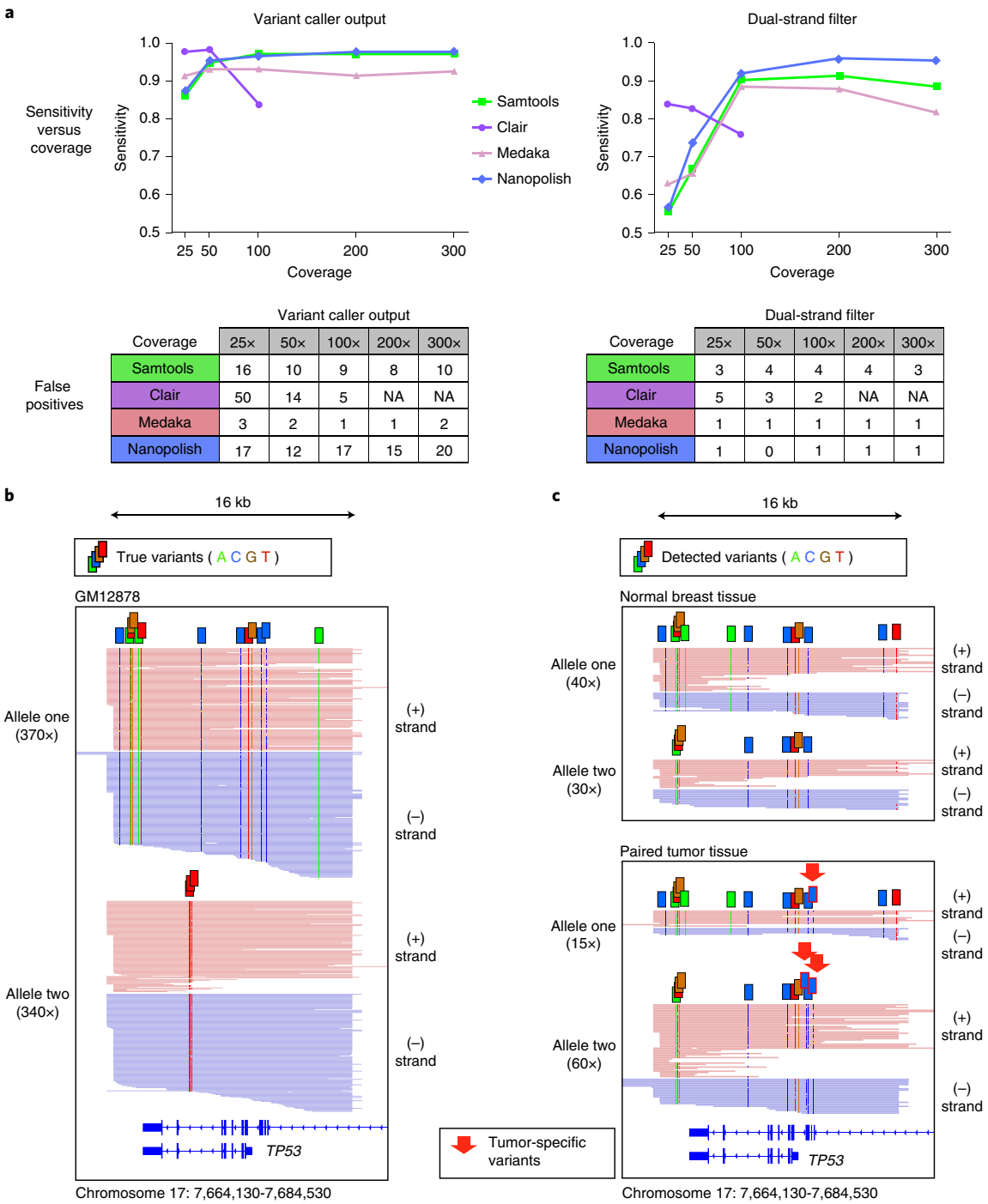
Fig. 1 | Method schematic and coverage data. a, Schematic of Cas9 enrichment operation. First, DNA ends are dephosphorylated, then new cuts are introduced with a Cas9–gRNA complex, and nanopore sequencing adapters are ligated to cuts around the ROI before loading the sample on the nanopore sequencer. **b**, Coverage plots at the *KRT19* gene (enriched area 18 kb) in four separate enrichment experiments: GM12878 with a single gRNA on each side (MinION); GM12878 with three gRNAs on each side (MinION); GM12878 with three gRNAs on each side (Flongle); and fresh tumor tissue with three gRNAs on each side (MinION). **c**, Table showing total aligned read count, on-target reads (within 20 kb of a gRNA site), on-target percentage and median coverage at each of the 10 enriched regions.

From GM12878 MinION data, the percentage of 'on-target' reads was 1.8% with the single gRNA panel and 4.6% with the multi-gRNA panel (see Supplementary Table 2). Genome-wide coverage analysis found the off-target reads to be distributed randomly across the genome, indicating that they result primarily from ligation of nanopore adapters to random breakage points. For example, in the GM12878 cell line with single gRNAs flanking each site, after quality-filtering alignments (MAPping Quality (MAPQ) > 30), there were only two genomic sites outside target regions where coverage reached 25×. Both of these are at repetitive pericentromeric sites and contain reads with lower mapping quality (MAPQ = 30–50), suggesting the increased coverage to be the result of alignment errors in these poorly mappable regions. We did note the occurrence of some off-target cleaving with the inclusion of gRNAs designed to flank the *BRCA1* locus (see Supplementary Fig. 1 and Supplementary Table 3), which we attribute to the abundance of repetitive regions[12] at this locus resulting in increased homology with other genomic loci.

With this new panel of guides, we tested the assay's performance in tissue samples: normal human breast tissue, a breast cancer cell-line-derived xenograft and a human breast tumor/normal pair. In tissue from a reduction mammoplasty (normal) and cell-line-derived mouse xenograft, we measured a median coverage of 162×/312×, and from the paired primary tumor/normal sample with limited input we achieved median coverage of 93×/70× (Fig. 1c, and see Supplementary Table 2).

Nanopore sequencing still has intrinsically high error rates (~5–10%) due to the inability of the base caller to distinguish between some *k*-mers and the difficulty in discriminating signal events in repetitive regions (for example, homopolymers). We explored how the increased coverage from the nCATS protocol would affect the ability to call variants from nanopore data. To simplify analysis, we limited variant calls to single nucleotide substitutions. There are numerous tools that currently exist for calling variants, and we selected four for comparison: (1) the Samtools/Bcftools package[13], which generates genotype likelihoods from alignment data; (2) Clair[14], which uses a deep neural network for variant calling from alignment data; (3) Medaka, a tool from Oxford Nanopore Technologies that also uses a neural network algorithm; and (4) Nanopolish[15], which uses a hidden Markov model to interrogate the raw electrical data as well as the alignment data.

For initial validation, we used the GM12878 cell line and the platinum genome dataset[6] as ground truth for single-nucleotide variants (SNVs). We benchmarked SNVs over the 8 loci without large deletions (total enriched area of 140 kb), which have a total of 174 annotated SNVs. To explore the relationship between coverage and variant calling efficiency, we subsampled the aligned data to coverage of 300×, 200×, 100×, 50× and 25× (see Methods). During



Variant caller output

False positives	Coverage	25×	50×	100×	200×	300×
	Samtools	16	10	9	8	10
	Clair	50	14	5	NA	NA
	Medaka	3	2	1	1	2
	Nanopolish	17	12	17	15	20

Dual-strand filter

Coverage	25×	50×	100×	200×	300×
Samtools	3	4	4	4	3
Clair	5	3	2	NA	NA
Medaka	1	1	1	1	1
Nanopolish	1	0	1	1	1

Fig. 2 | SNVs. a, Plot of sensitivity versus coverage using four tools to call SNVs from enrichment data in GM12878 for a 140-kb region containing 174 annotated SNVs. **b**, Visual representation of high-confidence variants detected by Nanopolish in the MinION data, from GM12878 for the captured region around *TP53*. Reads phased into homologous alleles using WhatsHap. **c**, High-confidence variants identified in primary tissue from a tumor/normal pair; red arrows were used to demarcate tumor-specific variants.

filtering, we selected for reads spanning the region, and maintained balanced coverage between both DNA strands.

We found that at lower coverage data (25× and 50×), the tool Clair had the greatest sensitivity (0.98). However, the model for Clair was trained and assessed on whole genome data only up to 100× coverage; above this coverage it no longer functioned. Medaka showed peak sensitivity of 0.93 at both 50× and 100× coverage, with sensitivity remaining robust at a higher coverage. Samtools variant calling and Nanopolish variant calling both increased in sensitivity up to 200× coverage, at which point they plateaued with sensitivities of 0.97 and 0.98, respectively (Fig. 2a, and see Supplementary Table 4).

One important caveat of the raw output of these variant caller pipelines is the persistence of false positives, limiting the use of this method for new SNV discoveries (Fig. 2a). On inspection, we noted many false positives to occur on only one strand (see Supplementary Fig. 2), suggesting that the base caller has systematic issues with the sequence of *k*-mers on one strand but not on the other. Thus, we

implemented a filter requiring variants to be supported by reads from both strands ('dual-strand filter'). This filter caused a decrease in sensitivity, especially at lower coverage. But, strikingly, this filter eliminated nearly all false-positive variant calls (see Supplementary Table 4), yielding a set of high-confidence variants. The dual-strand filter performed best with 200× coverage using Nanopolish variant calling (sensitivity, 0.96; *F*1 score, 0.97), with the sole false-positive variant existing in a thymidine-dense homopolymer region (see Supplementary Fig. 3). We then applied WhatsHap[16], a weighted haplotype assembler that uses statistical information as well as coverage depth, to assign reads into parental haplotypes based on SNVs detected in long-read data. A graphical depiction of detected variants is shown in Fig. 2b, highlighting the identification and phasing of variants in the captured region of *TP53* in GM12878. All 17 of the annotated SNVs in this region were detected using the dual-strand filtered data, with no false positives.

We then applied this variant-calling pipeline to our data from the MDA-MB-231 cell line to detect cancer-associated mutations. Across the captured regions of three cancer-associated genes (*BRAF*, *KRAS* and *TP53*) Nanopolish called 42 high-confidence SNVs (see Supplementary Table 5), including 2 of the 3 annotated in the COSMIC database for MDA-MB-231 (ref. [17]). The third variant was detected, but at a lower frequency in this aneuploid line, and thereby did not pass dual-strand filtering. Finally, we applied this variant-calling pipeline to the paired tumor/normal breast tissue sample, and phased the reads using WhatsHap[16]. We noticed a strong variation in the number of reads per haplotype in the *TP53* region, implying an imbalanced copy number in tumor cells (Fig. 2c). We examined two other captured regions on the same chromosome and observed similar chromosomal imbalance, with additional mutations in tumor samples (see Supplementary Fig. 4).

We next evaluated CpG methylation, which can be measured from nanopore electrical data[15]. Sites for methylation studies were selected by searching whole-genome nanopore data[10] for differentially methylated promoters between the non-tumorigenic breast cell line MCF-10A and the tumorigenic breast cell lines MCF-7 and MDA-MB-231. Candidate loci were further filtered by comparing with existing RNA-sequencing (RNA-seq) data[11] and genes with prognostic implications in human cancer[18–20].

We use read-level methylation plots to display methylation information from both a MinION sequencing run and a Flongle sequencing run (Fig. 3a, and see Supplementary Fig. 5). Methylation data for one locus (*KRT19)* are shown in Fig. 3a, with four additional genes (*GSTP1, GPX1*, *SLC12A4* and *TPM2)* in Supplementary Fig. 5. We compared nanopore methylation patterns with existing whole-genome bisulfite sequencing (WGBS) data in GM12878 (ref. [7]) using smoothed (loess) line plots (Fig. 3b, and see Supplementary Fig. 5). Directly comparing per-CpG methylation (see Supplementary Fig. 5) at each locus, we observed methylation calls largely clustered at points reflecting completely methylated or unmethylated sites, with an aggregate per-CpG correlation of 0.81 (Pearson's correlation coefficient).

We applied this strategy to our data for breast cell lines, looking for regions with differential methylation at these loci. One gene where we observed differential methylation in breast cell lines is the keratin family member gene *KRT19*. *KRT19* is known to be upregulated in breast cancer[19], and detection of *KRT19* messenger RNA has been used to identify micrometastasis of breast cancer to lymph nodes[21] and to detect circulating tumor cells[20]. We observed that *KRT19* remains largely methylated in the non-tumorigenic MCF-10-A cell line, but becomes hypomethylated in both of the transformed cell lines, MCF-7 and MDA-MB-231 (see Supplementary Fig. 6). This correlates with an observed increased transcript level for *KRT19* in the transformed cell lines (see Supplementary Fig. 7; Gene Expression Omnibus (GEO): GSE75168). Furthermore, we noted that the observed pattern of methylation was largely

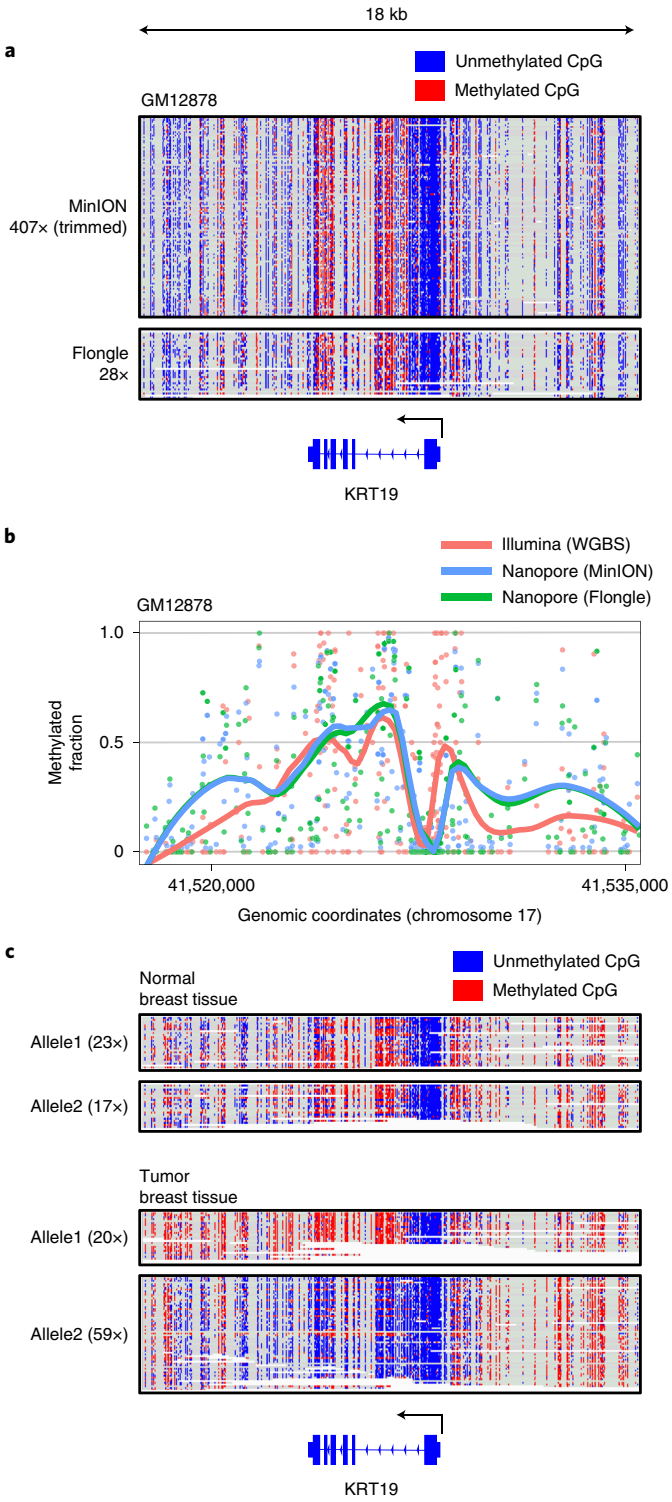


Fig. 3 | Methylation. a, Read-level plots showing methylation patterns in GM12878 from MinION and Flongle data at the *KRT19* locus. **b**, Methylation calls (points) and line plots at the same locus as in **a** showing smoothed (loess) methylation calls from WGBS on the Illumina platform[7], compared with methylation calls from MinION- and Flongle-targeted nanopore sequencing. **c**, Haplotype-phased methylation calls in primary patient tissue and paired tumor at the *KRT19* locus.

maintained in mouse xenografts derived from the MDA-MB-231 cell line (see Supplementary Fig. 8). In evaluation of the paired tumor/normal patient sample, we found that the primary patient

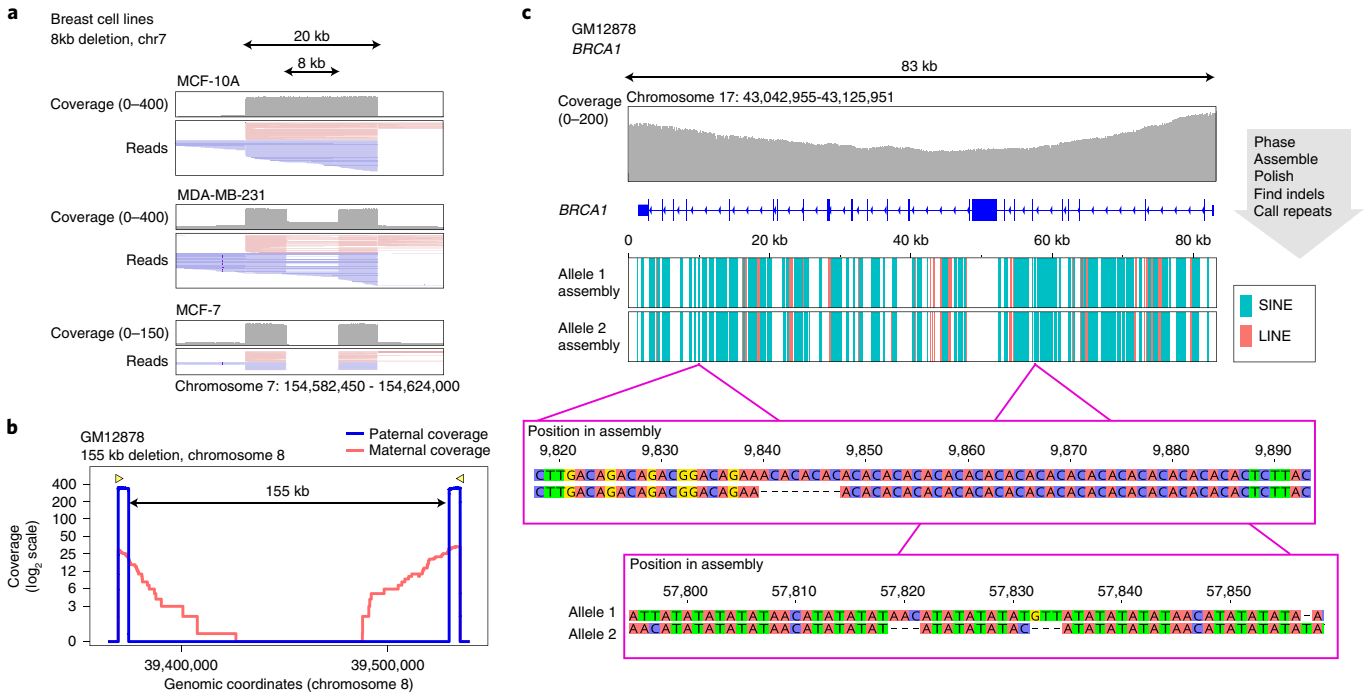


Fig. 4 | Structural variation. a, Reads around an ~8-kb deletion in chromosome 7 present in MCF-7 and MDA-MB-231, and absent in MCF-10A. **b**, Coverage on each parental allele in the region of a large (155-kb) heterozygous deletion in GM12878. **c**, Top: coverage at the *BRCA1* locus from DNA extracted using a Circulomics CBB kit. Middle: LINE and SINE components identified by RepeatMasker on each of the *BRCA1* allele assemblies. Bottom: three indels discovered between *BRCA1* assemblies not annotated in the platinum genome data set for GM12878.

tumor had a dramatic allele-specific hypomethylation of *KRT19* on the haplotype with increased copy number (Fig. 3c, and see Supplementary Fig. 8), in line with evidence suggesting increased expression in tumor cells[19–21]. This unveils a nuance about allele-specific methylation and copy number changes that would be difficult to query without the high-coverage, long-read data achieved by this methodology.

We next applied this method to evaluate structural variations (SVs) by confirming the presence of candidate deletions from whole-genome nanopore sequencing data[10]. We selected two deletions present in the MDA-MB-231 and MCF-7 breast cancer lines and absent in the MCF-10-A cell line, and designed gRNAs to flank breakpoints by ~5 kb. Plotting reads in Integrative Genomics Viewer (IGV) showed both deletions as heterozygous in MDA-MB-231 and homozygous in MCF-7 (Fig. 4a, and see Supplementary Fig. 9). The alignment data were passed to the Sniffles variant caller[22], which identified the breakpoints and zygosity of both deletions in line with our observations (see Supplementary Table 6). We also performed methylation studies on these regions but did not note any difference in methylation patterns between the deleted and intact alleles (see Supplementary Fig. 10).

To explore the use of this method for targeting larger fragments of DNA, we enriched for regions harboring large (>70 kb) heterozygous chromosomal deletions. We identified three large heterozygous deletions in GM12878 from available 10x Genomics data through the Genome In a Bottle (GIAB) Consortium[23], two with sizes of ~70 kb and one of ~155 kb. The gRNAs were designed to flank the deletion breakpoints by 5 kb, resulting in reads of ~10 kb on the deleted allele, and spanning the region between cut sites (80 kb/165 kb) on the intact allele. We again phased the reads into parental alleles[16] and compared read lengths and read counts achieved for each allele. Interestingly, we found that the allele containing the deletion, with the correspondingly shorter distance between the cut sites, demonstrated a number of reads that were an order of magnitude higher (Fig. 4b, and see Supplementary Fig. 11). This reflects a bias against achieving reads >50 kb, probably introduced during DNA purification, library preparation or delivery to the pore. To confirm this size bias, we performed similar parental-allele segregation on sites without SVs and did not observe bias toward either parental allele (see Supplementary Fig. 12). The alignment data were passed to the Sniffles variant caller[22], which identified all 3 of the deletions within 10 nucleotides (nt) from the annotated breakpoints in existing GIAB data (see Supplementary Table 7). We adjusted the Sniffles parameters to call SVs as heterozygous if an allele was supported by even a very low amount (0.1%) of reads, because the imbalance of reads from the two alleles caused the software to initially identify these deletions as homozygous (see Methods).

Finally, we targeted the *BRCA1* gene, because of the well-documented association of this gene with familial breast cancer[12]. *BRCA1* is an attractive target for long-read sequencing because of the abundance of hard-to-map repetitive Alu elements. To capture the entire *BRCA1* gene (distance between flanking gRNAs 83 kb), further adjustments to DNA extraction were needed. Initial MinION sequencing runs from 3 µg of extracted GM12878 DNA resulted in only 10 sequencing reads spanning the entire region with many smaller fragments (see Supplementary Fig. 13). We found that using the Circulomics NanoBind kit for DNA extraction resulted in an increase to almost 30 reads completely spanning the *BRCA1* gene, with coverage between gRNAs ranging from 100× to 200× (Fig. 4c, and see Supplementary Fig. 13). We phased the *BRCA1* reads into haplotypes[16] with new high-confidence variants found using Nanopolish. We then built an assembly for each of the two alleles using the Flye assembler[24] and polished the assemblies using Racon[25] and Medaka (see Methods). This resulted in full-length assemblies for each of the two alleles, within which we identified the presence of short interspersed nuclear elements (SINEs) (for example Alu elements) and long interspersed nuclear elements (LINEs) using RepeatMasker. We compared the two assemblies for variants between alleles and found numerous indels and single nucleotide changes (see Supplementary Table 8) using the minimap2 suite[26].

After filtering for homopolymer regions (which nanopore sequencing is known to still have difficulty resolving[15]), we found 10 indels of at least 3 nt between the two assemblies (see Supplementary Table 8). Seven of these ten were annotated in the platinum genome dataset for GM12878 (ref. [6]). The remaining three unannotated indels are within large repetitive regions (Fig. 4c), making it difficult to map reads from conventional short-read sequencing. To validate these indels, we compared against recently released whole-genome PacBio data for GM12878 (Sequence Read Archive (SRA): SRR9001768 to SRR9001773)[23] which confirmed each of these three small unannotated indels (see Supplementary Fig. 14). Furthermore, as we used nanopore sequencing, we were also able to call CpG methylation on this complex region, and compare methylation profiles between *BRCA1* alleles (see Supplementary Fig. 15).

As a result of the low cost to entry and small footprint of the instrument, this assay has the potential to be widely utilized as a tool for identifying single nucleotide changes, evaluating DNA methylation and studying structural variation. We were even able to apply this to clinical tissue despite the relatively high DNA input requirements (3 µg). We show that SNVs in regions of interest can be queried using the nCATS protocol, although there are persisting limitations, as evinced by the few SNVs not detected by this approach. We found that, by using only high-confidence variants, we could phase nanopore sequencing reads into parental alleles using WhatsHap[16], permitting haplotype resolution of high-coverage nanopore data. As base-calling and variant-calling algorithms continue to improve, we anticipate higher future performance for surveillance and identification of mutations. We also highlight the use of nCATS to detect and validate structural variants. It is only with the advent of long-read sequencing that the great diversity of structural variation in human genomes has been appreciated[27,28], and this method provides a dynamic tool to evaluate genomic rearrangements, including large SVs and hard-to-map repetitive regions[29]. Importantly, as nanopore sequencing interrogates the DNA strand rather than sequencing 'by synthesis', we can simultaneously profile methylation in these loci, providing biologic as well as diagnostic insight into the epigenome, which is commonly disrupted in human neoplasia[30]. In fact, as we show, long reads allow easy phasing of methylation into different alleles, allowing careful exploration of allele-specific epigenetic changes. The high sequencing depth granted by this method is especially useful for characterizing genetically and epigenetically heterogeneous samples typically obtained from clinical samples—giving us insight into the frequency of different mutations and epigenetic changes present.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-0407-5.

Received: 4 June 2019; Accepted: 6 January 2020;
Published online: 10 February 2020

References

1. Karamitros, T. & Magiorkinis, G. Multiplexed targeted sequencing for Oxford Nanopore MinION: a detailed library preparation procedure. *Methods Mol. Biol.* **1712**, 43–51 (2018).
2. Leija-Salazar, M. et al. Evaluation of the detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. *Mol. Genet. Genom. Med.* **7**, e564 (2019).
3. Gabrieli, T. et al. Selective nanopore sequencing of human *BRCA1* by Cas9-assisted targeting of chromosome segments (CATCH). *Nucleic Acids Res.* **46**, e87 (2018).
4. Giesselmann, P. et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat. Biotechnol.* **37**, 1478–1481 (2019).
5. Kozarewa, I., Armisen, J., Gardner, A. F., Slatko, B. E. & Hendrickson, C. L. Overview of target enrichment strategies. *Curr. Protoc. Mol. Biol.* **112**, 7.21.1–7.21.23 (2015).
6. 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).
7. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
8. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
9. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
10. Lee, I. et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/504993 (2018).
11. Messier, T. L. et al. Histone H3 lysine 4 acetylation and methylation dynamics define breast cancer subtypes. *Oncotarget* **7**, 5094–5109 (2016).
12. Welcsh, P. L. & King, M. C. *BRCA1* and *BRCA2* and the genetics of breast and ovarian cancer. *Hum. Mol. Genet.* **10**, 705–713 (2001).
13. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
14. Luo, R. et al. Clair: Exploring the limit of using a deep neural network on pileup data for germline variant calling. Preprint at *bioRxiv* https://doi.org/10.1101/865782 (2019).
15. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
16. Martin, M. et al. WhatsHap: fast and accurate read-based phasing. Preprint at *bioRxiv* https://doi.org/10.1101/085050 (2016).
17. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
18. Martignano, F. et al. GSTP1 methylation and protein expression in prostate cancer: diagnostic implications. *Dis. Markers* **2016**, 4358292 (2016).
19. Kabir, N. N., Rönnstrand, L. & Kazi, J. U. Keratin 19 expression correlates with poor prognosis in breast cancer. *Mol. Biol. Rep.* **41**, 7729–7735 (2014).
20. Wang, X.-M., Zhang, Z., Pan, L.-H., Cao, X.-C. & Xiao, C. *KRT19* and *CEACAM5* mRNA-marked circulated tumor cells indicate unfavorable prognosis of breast cancer patients. *Breast Cancer Res. Treat.* **174**, 375–385 (2019).
21. Noguchi, S. et al. Detection of breast cancer micrometastases in axillary lymph nodes by means of reverse transcriptase-polymerase chain reaction. Comparison between MUC1 mRNA and keratin 19 mRNA amplification. *Am. J. Pathol.* **148**, 649–656 (1996).
22. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
23. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
24. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
25. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
26. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
27. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Preprint at *bioRxiv* https://doi.org/10.1101/193144 (2018).
28. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e19 (2019).
29. Dixon, J. R. et al. Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).
30. Timp, W. & Feinberg, A. P. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat. Rev. Cancer* **13**, 497–510 (2013).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020

Methods

Cell culture and DNA prep. Cell lines were obtained from American Type Culture Collection—MCF10A (CRL-10317), MCF7 (HTB-22) and MDA-MB-231 (HTB-26)—or Coriell institute—CEPH/UTAH Pedigree 1463 (GM12878). Cells were cultured according to recommended protocols. Briefly, all cell lines were maintained at 37 °C in 5% CO_2. The GM12878 cell line was grown in high-glucose RPMI medium supplemented with 10% fetal calf serum, penicillin–streptomycin antibiotics (pen–strep) and L-glutamine. MCF-7 and MDA-MB-231 were grown in hi-glucose Dulbecco's modified Eagle's medium supplemented with 10% fetal calf serum, pen–strep and L-glutamine. MCF-10A cells were grown in hi-glucose Dulbecco's modified Eagle's medium supplemented with 5% horse serum, pen–strep, L-glutamine, epidermal growth factor, insulin, hydrocortisone and cholera toxin. DNA was extracted from the cells using either the MasterPure kit (Lucigen, catalog no. MC85200) or the Nanobind kit (Circulomics, catalog no. NB-900-001-01), and stored at 4 °C until use. DNA was quantified using the Qubit fluorometer (Thermo Fisher Scientific) immediately before performing the assay.

Patient tissue and mouse xenograft. All human samples were collected with appropriate approval from the Johns Hopkins institutional review board. The primary breast tumor was identified as ER/PR$^+$ (estrogen/progesterone receptor positive) by immunohistochemistry and snap frozen. Mouse experiments were conducted with prior approval from Johns Hopkins University Animal Care and Use. Mouse xenografts were generated by injecting 106 ER/PR/HER2$^-$ MDA-MB-231 breast cancer cells into the mammary fat pad of athymic mice. Tumors were collected 6–8 weeks later and frozen immediately as small chunks. The snap-frozen tissue was ground under liquid nitrogen using a CryoMill (Retch) and the DNA extracted using a MasterPure kit (Lucigen, catalog no. MC85200).

Design of gRNAs. The gRNAs were assembled as a duplex from synthetic CRISPR RNAs (crRNAs) (IDT, custom designed) and *trans*-activating crRNAs (tracrRNAs) (Integrated DNA Technologies (IDT), catalog no. 1072532). Sequences are provided in Supplementary Table 1. The crRNAs were designed using IDT's design tool and selected for the highest predicted on-target performance with minimal off-target activity. The gRNA duplex was designed to introduce cuts on complementary strands flanking the region of interest. For methylation studies and SNV studies, the target size between gRNAs was 12–24 kb; for deletions, the gRNAs were designed to flank the suspected breakpoints by ~5 kb.

RNP assembly. Before gRNA assembly, all crRNAs were pooled into an equimolar mix, with a total concentration of 100 µM. The crRNA mix and tracrRNAs were then combined such that the tracrRNA concentration and total crRNA concentration were both 10 µM. The gRNA duplexes were formed by denaturation for 5 min at 95 °C, and then allowed to cool to room temperature for 5 min on a benchtop. RNPs were constructed by combining 10 pmol of gRNA duplexes with 10 pmol of HiFi Cas9 Nuclease V3 (IDT, catalog no. 1081060) in 1× CutSmart Buffer (New England Biolabs, catalog no. B7204) at a final volume of 30 µl (concentration 333 nM), incubated for 20 min at room temperature, then stored at 4 °C until use, up to 2 d.

Cas9 cleavage and library prep. Input DNA, 3 µg, was resuspended in 30 µl of 1× CutSmart buffer (New England Biolabs, catalog no. B7204), and dephosphorylated with 3 µl of Quick CIP enzyme (New England Biolabs, catalog no. M0508) for 10 min at 37 °C, followed by heating for 2 min at 80 °C for calf-intestinal alkaline phosphatase (CIP) enzyme inactivation. After allowing the sample to return to room temperature, 10 µl of the pre-assembled 333 nM Cas9–gRNA complex was added to the sample. In the same tube, 1 µl of 10 mM dATP (Zymo, catalog no. D1005) and 1 µl of Taq DNA polymerase (New England Biolabs, catalog no. M0267) were added for A-tailing of the DNA ends. The sample was then incubated at 37 °C for 20 min for Cas9 cleavage, followed by 5 min at 72 °C for A-tailing. Sequencing adapters and ligation buffer from the Oxford Nanopore Ligation Sequencing Kit (Oxford Nanopore Technologies, catalog no. LSK109) were ligated to DNA ends using Quick Ligase (New England Biolabs, catalog no. M2200) for 10 min at room temperature. The sample was cleaned up using 0.3× Ampure XP beads (Beckman Coulter, catalog no. A63881), washing twice on a magnetic rack with the long-fragment buffer (Oxford Nanopore Technologies, catalog no. LSK109) before eluting in 15 µl of elution buffer (Oxford Nanopore Technologies, catalog no. LSK109). Sequencing libraries were prepared by adding the following to the eluate: 25 µl of sequencing buffer (Oxford Nanopore Technologies, catalog no. LSK109), 9.5 µl of loading beads (Oxford Nanopore Technologies, catalog no. LSK109) and 0.5 µl of sequencing tether (Oxford Nanopore Technologies, catalog no. LSK109). A detailed step-wise description of the enrichment method is available on protocols.io (https://www.protocols.io/view/cas9-enrichment-for-nanopore-sequencing-68ihhue).

Sequencing. Samples were run on a MinION (v.9.4.1) flow cell or a Flongle flow cell (v.9.4.1 pore), using the MK1B or GridION sequencer. Sequencing runs were operated using the MinKNOW software (v.19.2.2). A detailed description of runs (flow cell, gRNAs, sequencer) is provided in Supplementary Table 1.

Analysis. Base calling was performed using GUPPY (v.3.0.3) to generate FASTQ sequencing reads from the electrical data. Reads were aligned to the human reference genome (Hg38) using minimap2 (v.2.17)[26]. Per-nucleotide coverage was determined using Samtools, and clustered using the 'bincov' script of the SURVIVOR (v.1.0.7) software package[31]. On-target reads were defined as those that aligned within 20 kb of a gRNA site. Average coverage per region is the average of coverage of all bases between the innermost gRNA sites, using coverage found by Samtools.

New variant calling was performed using Samtools (v.1.9)[13], Clair (v.2.0.0)[14], Medaka (v.0.10.0) or Nanopolish (v.0.11.1)[15]. For validation, we compared SNV calls with those annotated for GM12878 as part of the platinum genome dataset[6]. To achieve different coverage values for validation of GM12878 data, each region was randomly subsampled using Samtools to achieve 300× coverage with balanced read counts on each strand. The reads were then further subsampled to achieve the lower coverage values of 200×, 100×, 50× and 25×. Sensitivity was calculated as correctly called SNVs (true positives) out of all true SNVs (true positives + false negatives). The *F*1 score is included as a measure of overall test accuracy, calculated as the harmonic mean of precision and recall.

High-confidence variants were generated by an additional filter requiring variants to be supported by reads from both strands. Bam alignment files were split into reads aligning to forward strand and reverse strand, and variant calls were performed on each set of reads separately. Variants were included in the high-confidence set only if they were called in forward strand reads alone, reverse strand reads alone and the complete dataset.

Segregation of reads into parental alleles was performed with WhatsHap (v.0.18)[16], using only new called high-confidence variants. For patient tumor tissue, reads phased into haplotypes with only the variants identified from paired normal tissue.

CpG methylation calling on nanopore data was performed using Nanopolish (v.0.11.1)[15]. Methylation calling on existing WGBS GM12878 data (GEO GSE86765)[7] was performed using the Bismark (v.0.18.2) software tool[32]. The Bismark output files were processed using the bsseq R package (v.3.9)[33] and Pearson's correlation coefficient was calculated using base R. RNA-seq data of MCF-10A, MCF-7 and MDA-MB-231 were downloaded from the GEO (GSE75168) in the form of RNA counts.

Deletions were called using the SV caller Sniffles (v.1.0.11)[22], set to find deletions with a minimum size of 100 bp. In the instance of the very large (>70 kb) heterozygous deletions in GM12878, the allelic size bias caused the ploidy to be incorrectly called as homozygous. To correct this, we used the option '-min_homo_af' set to 99.9, which ensured that a deletion was called as heterozygous if supporting reads for an allele were present at a rate as low as 1 in 1,000.

For assembly of the *BRCA1* region, reads were first split into haplotypes using WhatsHap[16]. A draft assembly for each allele was built using the Flye (v.2.4.2) assembly tool[24], with default parameters for nanopore reads. Draft assemblies were then corrected by using 4 iterative rounds of polishing with the Racon error-correction software (v.1.3.3)[25], with the score for matching bases ('-m') increased to 8 and the score for mismatching bases ('-x') decreased to −6. A final round of polishing was performed using the Medaka consensus tool with default parameters. The assemblies were put under surveillance for indels using the paftools helper script of the minimap2 suite (v.2.17)[26].

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

Data availability

Sequencing data from all non-primary patient samples for this study can be retrieved from the SRA, under the BioProject ID PRJNA531320.

Code availability

The computational code used in all of the analysis is hosted on GitHub (see https://github.com/timplab/Cas9Enrichment, https://github.com/isaclee/nanopore-methylation-utilities).

References

31. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
32. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
33. Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13**, R83 (2012).

Acknowledgements

This work was supported by funding from the National Institutes for Health (grant no. R01 HG009190) (National Human Genome Research Institute).

Author contributions

T.G. and W.T. constructed the study. T.G. performed the experiments. T.G., I.L. and F.S. analyzed the data. T.G., J.G., E.R., R.B. and A.H. developed the method. S.S. and B.D. provided primary breast tissue and generated the mouse xenografts. T.G. and W.T. wrote the paper.

Competing interests

J.G., E.R., R.B. and A.H. are employees of Oxford Nanopore Technologies. W.T. has two patents licensed to Oxford Nanopore Technologies (US patent nos. 8,748,091 and 8,394,584). T.G., I.L., F.S. and W.T. have received travel funds to speak at symposia organized by Oxford Nanopore Technologies.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-0407-5.

Correspondence and requests for materials should be addressed to W.T.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Winston Timp

Last updated by author(s): Jan 1 2020

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 Authors & Referees 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	All data was collected using the commercial MinKNOW software (v19.2.2) , which is provided by Oxford Nanopore Technologies for collecting electrical data during Nanopore sequencing. The raw electrical data was converted to DNA sequence files (fastq) using the Guppy (ver3.0.3) basecaller.
Data analysis	For read alignment Minimap2 (v2.17) Samtools (v1.9) For off-target analysis SURVIVOR (v1.0.7) For single nucleotide variants Bcftools (v1.9) Nanopolish (v0.11.1) Clair (v2.0.0) Medaka (v0.10.0) For allele phasing WhatsHap (v0.18) For methylation : Nanopolish (v0.11.1) Bismark (v0.18.2) BsSeq (v3.9)

For Structural Variation
Sniffles (v1.0.11)
RepeatMasker (open 4.0)
Flye (v2.4.2)
Racon (v1.3.3)

Scripts used to execute these analysis are provided on our gitHub page, as described in the manuscript methods section

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/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

Cas9-Enrichment data are available at NCBI Bioproject ID PRJNA531320 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA531320).

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	We used a single nanopore sequencing flow cell for each sample, all sequencing signal events were used as an independent observation of the cells.
Data exclusions	No data were excluded from these analyses
Replication	There is some variability in the output of total sequencing reads between nanopore sequencing runs
Randomization	Each cell line was considered as a separate experimental group
Blinding	Blinding was not relevant to this study, as it was evaluating chromosomal features of cell lines

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

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	GM12878 cell line was acquired from Coriell Institute for Medical Research. Breast cell lines (MCF-10-A, MDA-MB-231, MCF-7) were received from American Type Culture Collection (ATCC) repository.

Authentication	No additional authentication was performed on cell lines
Mycoplasma contamination	The cell lines were not tested for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	None of the 4 cell lines used are included in the Ver9 update on commonly misidentified cell lines

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	This study involved the use of laboratory mice for generating cell-line-derived xenografts
Wild animals	This study did not involve any wild animals
Field-collected samples	This study did not involve any samples collected from the field
Ethics oversight	Mouse experiments were conducted with prior approval from Johns Hopkins Animal Care and Use Committee (JH-ACUC)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Human subjects were all female. Cancer diagnosis and characterization was performed by the Johns Hopkins pathology department.
Recruitment	Patients were recruited by asking their desire to participate in a research study during the course of their treatment at Johns Hopkins hospital
Ethics oversight	This study was approved by the Johns Hopkins Institutional Review Board (IRB)

Note that full information on the approval of the study protocol must also be provided in the manuscript.