bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure

Netha Ulahannan[1,2 *], Matthew Pendleton[3 *], Aditya Deshpande[1,2,4 *], Stefan Schwenk[5], Julie M. Behr[1,2,4], Xiaoguang Dai[3], Carly Tyer[6], Priyesh Rughani[3], Sarah Kudman[1], Emily Adney[1,2], Huasong Tian[1,2], David Wilkes[1], Juan Miguel Mosquera[1], David Stoddart[5], Daniel J. Turner[5], Sissel Juul[3], Eoghan Harrington[3,], and Marcin Imielinski[1,2,**] ✉**

[1]Weill Cornell Medicine, New York, NY
[2]New York Genome Center, New York, NY
[3]Oxford Nanopore Technologies, New York, NY
[4]Tri-institutional Ph.D. Program in Computational Biology and Medicine, New York, NY
[5]Oxford Nanopore Technologies, Oxford, UK
[6]Oxford Nanopore Technologies, San Francisco, CA
[*]These authors contributed equally.
[**]These authors co-supervised the work.

Abstract

Higher-order chromatin structure arises from the combinatorial physical interactions of many genomic loci. To investigate this aspect of genome architecture we developed Pore-C, which couples chromatin conformation capture with Oxford Nanopore Technologies (ONT) long reads to directly sequence multi-way chromatin contacts without amplification. In GM12878, we demonstrate that the pairwise interaction patterns implicit in Pore-C multi-way contacts are consistent with gold standard Hi-C pairwise contact maps at the compartment, TAD, and loop scales. In addition, Pore-C also detects higher-order chromatin structure at 18.5-fold higher efficiency and greater fidelity than SPRITE, a previously published higher-order chromatin profiling technology. We demonstrate Pore-C's ability to detect and visualize multi-locus hubs associated with histone locus bodies and active / inactive nuclear compartments in GM12878. In the breast cancer cell line HCC1954, Pore-C contacts enable the reconstruction of complex and aneuploid rearranged alleles spanning multiple megabases and chromosomes. Finally, we apply Pore-C to generate a chromosome scale *de novo* assembly of the HG002 genome. Our results establish Pore-C as the most simple and scalable assay for the genome-wide assessment of combinatorial chromatin interactions, with additional applications for cancer rearrangement reconstruction and *de novo* genome assembly.

Chromatin structure | Structural variation | Long read sequencing | *De novo* genome assembly | cancer genomics

Correspondence: *mski@mskilab.org*

Introduction

Mammalian chromatin is hierarchically organized into active and inactive compartments[1,2], topologically associated domains (TADs)[3–5], and loop domains[2,6,7]. A key challenge in epigenomics is to link multi-scale features of chromatin folding with the functional cellular programs that drive transcriptional regulation and cell identity[8,9]. An improved understanding of chromatin folding may also provide insight into the role of the epigenome in driving specific disease states. Previous work in constitutional genetics has linked the presence of inherited non-coding DNA variants, such as CTCF mutations[10] or larger structural variants[11,12], to human limb malformations through chromatin re-structuring[10]. In cancer, complex somatic genomic rearrangements can create neo-TADs[11,13] or hijack enhancers to activate oncogenes[14–18].

Our modern understanding of chromatin structure relies heavily on results obtained from chromosome conformation capture assays (3C, 4C, 5C and Hi-C) which utilize molecular readouts (e.g. PCR, sequencing) to map pairs of loci that spatially interact in the nucleus[1,19]. The general principles underlying these techniques can be modified to target specific nuclear proteins by immunoprecipitation (ChIA-PET[20], HiChIP[21]) or specific loci using nucleic acid hybridization enrichment(ChiC[22]), and they can be adapted to study single-cell samples[23,24]. Most notably, Hi-C enables unbiased mapping of active (A) and inactive (B) chromatin compartments, TADs, cohesin loops, and enhancer-promoter interactions among the billions of interacting locus pairs in these genome-wide contact maps. In addition, Hi-C has been applied to generate reference-grade *de novo* assemblies of mammalian scale genomes[25], identify structural variants in cancer cells[13], and study DNA replication[26,27].

Though analysis of Hi-C and its derivatives have revealed the unprecedented complexity of the 3D genome, the derived features are insufficient to explain many aspects of gene regulation[9,28–30]. The functional 3D state of chromatin, like other cellular machinery (e.g. protein signaling cascades), may not be fully described in terms of simple pairwise interactions. Rather, gene activation or silencing may require the interaction of three or more DNA loci, some of which may co-exist in dynamic and specialized nuclear structures that achieve cooperativity through liquid-liquid phase separation[31,32]. Detection of such higher-order chromatin complexes may be necessary to reveal fundamental links between genome structure and function.

Recent technologies enabling the study of three-way[33] or higher-order chromatin interactions include chromosomal walks (c-walks)[34], genome architecture mapping

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

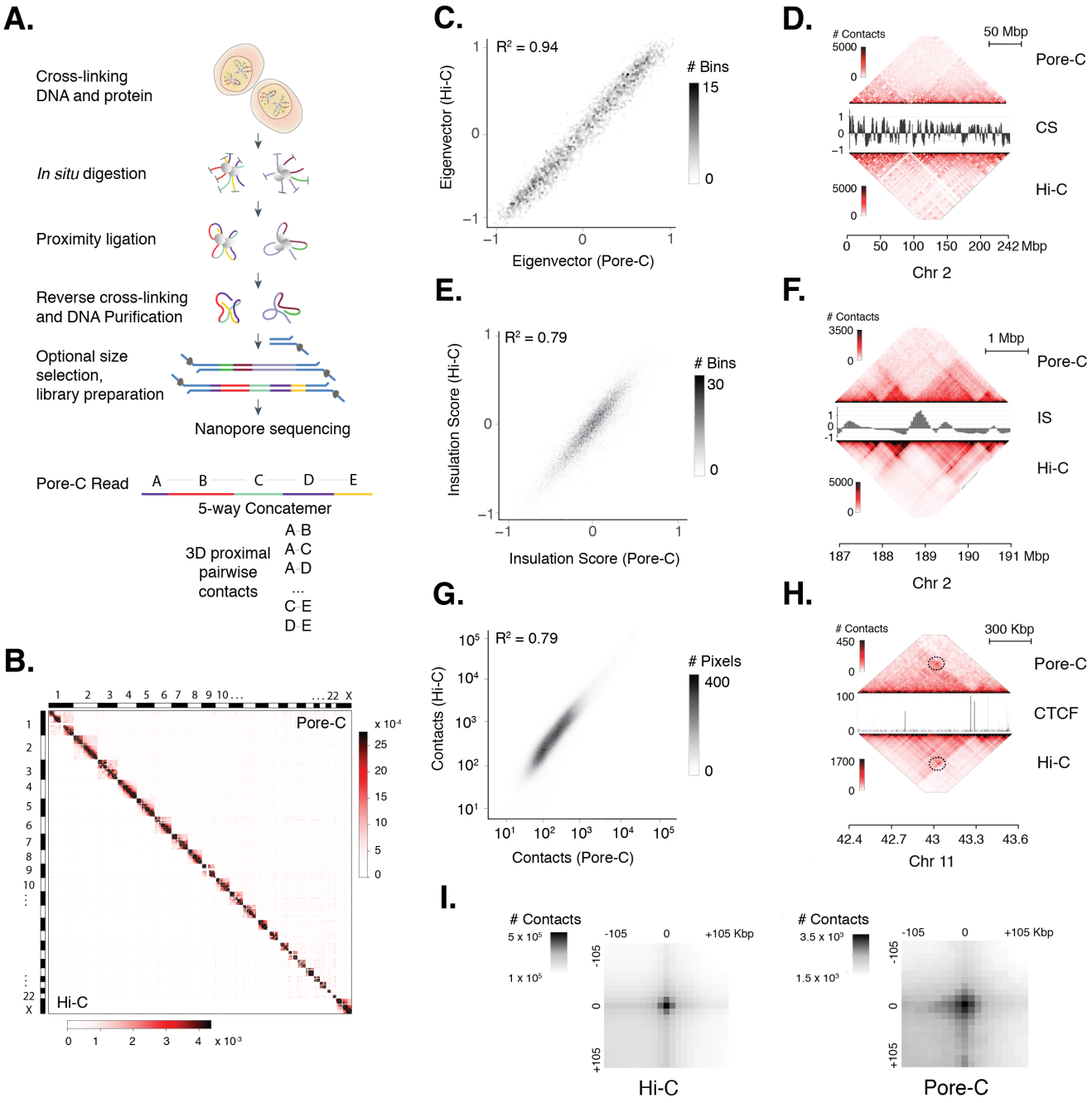

Fig. 1. Pore-C concatemers reflect previously characterized features of chromatin architecture in GM12878 (A) Schematic of the Pore-C protocol: Proximally-ligated concatemers obtained by chromatin conformation capture are directly sequenced using Oxford Nanopore Technologies long-read sequencing platform. A greedy piece-wise alignment algorithm maps each concatemer sequence to a multi-way contact. (B) Comparison of Pore-C (upper triangle) 1 Mbp (virtual pairwise) and Hi-C contact maps[2] (lower triangle) for chromosomes 1-22 and X (hg38) in the cell line GM12878. (C) Comparison of 500 Kbp compartment scores (CS) between Pore-C (1.38 billion virtual pairwise contacts) and Hi-C (~4 billion pairwise contacts) for GM12878[2] (D) Comparison of Pore-C and Hi-C 500 Kbp contact maps for chromosome 2, alongside the CS demonstrating visual correspondence between Pore-C and Hi-C compartment patterns. (E) Comparison of 50 Kbp topologically-associated domain insulation scores (IS) between Pore-C and Hi-C. (F) Comparison of Pore-C and Hi-C 50 Kbp contact maps for a portion of chromosome 2, alongside the IS, demonstrating visual correspondence between Pore-C and Hi-C derived TAD structures. (G) Comparison of raw *cis* off-diagonal contact counts at 500 Kbp resolution between Pore-C and Hi-C. (H) An example of CTCF peaks coinciding with visually recognizable loop anchors (circle) in 25 Kbp Pore-C and Hi-C contact maps. (I) Results of aggregate peak analysis (APA) detecting similar enrichment of Pore-C and Hi-C pairwise data within 100 Kbp of loop anchors. In this map, each 10 Kbp x 10 Kbp pixel represents the total number of contacts detected across the entire loop set in a standard coordinate system centered around each loop anchor.

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(GAM)[35], split-pool recognition of interactions by tag extension (SPRITE)[36], ChIA-drop[37], Tri-C[38], Tethered multiple 3C[39], the concatemer ligation assay (COLA)[33] and Multi-contact 4C (MC-4C)[40]. Among these assays, only a subset (Tri-C, SPRITE, COLA, GAM) have been shown to generate genome-wide maps in mammals, and all methods but SPRITE suffer from very rare representation of (<1%) of contacts with order >3[36]. Though SPRITE contacts can reach a very high order (> 500), many may represent library artifacts arising from very large cross-linked fragments in the library preparation. As a result, they may serve as readouts of gross nuclear features rather than combinatorial chromatin states. In an ideal genome-wide higher-order chromatin assay, multi-way contacts would represent spatial adjacencies between sets of interacting DNA molecules (e.g. cooperative chromatin complexes arising during transcription factor binding).

Previous studies combining targeted chromatin conformation capture with long-read sequencing have revealed that many of the products of proximity ligation are in fact DNA "concatemers" of multiple interacting loci[33,34,37,40,41]. We surmised that chromatin conformation capture could be paired with nanopore sequencing to efficiently assay the combinatorial chromatin structure in human cells (**Fig. 1A**). We hypothesized that this approach would reveal higher-order 3D chromatin structure at both the scale of chromosomal compartments (i.e. multi-megabase) and chromatin loops (kilobase), identify novel 3D structures generated through complex cancer genomic rearrangements, and improve the scaffolding of long read, whole genome sequencing (WGS) derived assemblies.

Results

Pore-C identifies higher-order contacts among DNA concatemers. We coupled nanopore sequencing with chromatin conformation capture to generate Pore-C libraries in an amplification-free manner through *in situ* digestion, and proximity ligation of cross-linked nuclei derived from the EBV-transformed B-lymphocyte cell lines GM12878 and HG002, as well as the breast cancer cell line HCC1954 (**Fig 1A**). For each sequenced Pore-C read, the Pore-C pipeline identifies a minimal subset of candidate alignments that maximally covers the entire Pore-C read. Each of these filtered alignments is then assigned to a restriction fragment based on the position of their midpoint (**Fig 1A**, **Fig S1A**). We refer to the set of filtered alignments associated with a given Pore-C read as a (multi-way) *contact*, and the number of fragments associated with a contact as its *order*.

To compare higher-order contacts with the pairwise contacts produced by Hi-C, we decompose each Pore-C read into all possible pairs, thus producing $\binom{k}{2}$ virtual pairwise contacts per read, where k is the order of the read (**Fig S1A**). Furthermore, we introduce the *pairwise contacts per Gbp* statistic as a measure of the contact density of the raw reads. In total we generated 597 Gbps of sequence data, detecting 145.2 million concatemers, 80.4 million of which had a contact order of 4 or more (**Table S1**, **Table S2**). These concatemers could be decomposed into 2.73 billion virtual pairwise contacts . We found that Pore-C libraries created using the NlaIII restriction enzyme had the highest contact density with an average of 6.5 million contacts per Gbp of sequenced data, closely matching the contact density of Hi-C datasets generated using 100 bp paired-end reads (**Fig S1D**) from the 4D nucleome project. We also observed that, although both the NlaIII and DpnII restriction enzymes were 4-base cutters, the fragment N50 of the NlaIII fragments in GRCh38 was roughly half the DpnII fragment N50 (**Fig S1B**) resulting in a lower contact density for the latter enzyme.

Pore-C virtual pairwise contact maps are concordant with Hi-C. In order to compare Pore-C data to a previously-published "gold standard" GM12878 Hi-C dataset[2], we constructed a contact map for GM12878 using NlaIII-derived virtual pairwise contacts (see above). Visual inspection of the Pore-C virtual contact map revealed previously identified features of GM12878 chromatin structure including A/B compartments (500 Kbp resolution, **Fig 1C**) associated with histone marks known to be predictive of active / inactive chromatin states **(Fig S2D)**, topologically associated domains (TADs) (50 Kbp resolution, **Fig 1F**), and loops (10 Kbp resolution, **Fig 1H-I**), all of which were reflected in the corresponding Hi-C dataset. Additionally, we found a close correlation between Pore-C virtual pairwise contact maps and Hi-C data at the levels of eigenvector compartment scores ($R^2 = 0.94$), TAD insulation scores ($R^2 = 0.79$), and raw contact pixels ($R^2 = 0.79$) (**Fig 1C, E and G**). To assess Pore-C run-to-run variability, the A/B compartment eigenvector score, TAD insulation score, and raw matrix correlations were computed for each of the individual NlaIII and DpnII runs **(Fig S2A-C)** and compared to the Hi-C dataset. We concluded that Pore-C DNA concatemers reflect known previously characterized pairwise features of chromatin architecture.

High-order long-range contacts reflect combinatorial chromatin states. A considerable fraction of Pore-C concatemers represent multi-way (i.e. order > 2) chromatin contacts, with a median order of 7 (**Fig 2A**), and with 78.09%, 10.23%, and 1.04% of these reads harboring a contact order greater than 2, 10, and 20, respectively. In comparison, SPRITE clusters have a median order of 4 with 11.69%, 1.13%, and 0.4% of contacts with order greater than 2, 10, and 20, respectively. Additionally, we noted that SPRITE clusters were relatively depleted in multi-way contacts, and comprised a higher fraction of pairwise contacts (13.8%) and singletons (74.5%) compared to Pore-C (15.6% and 6.3%, respectively). This indicated a marked enrichment (OR = 18.5 for > 2 way; OR = 9.9 for > 10 way) of multi-way contacts in Pore-C as compared to SPRITE.

To explore the properties of multi-way contacts of increasing order, we divided this distribution into several groups based on order (2, 3, 4-5, 6-10, 11-20, 21-50, >50). We analyzed the genomic distance between pairs of monomers from decomposed concatemers, grouped by contact order (**Fig. 2B**). Previous studies have demonstrated high-order

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

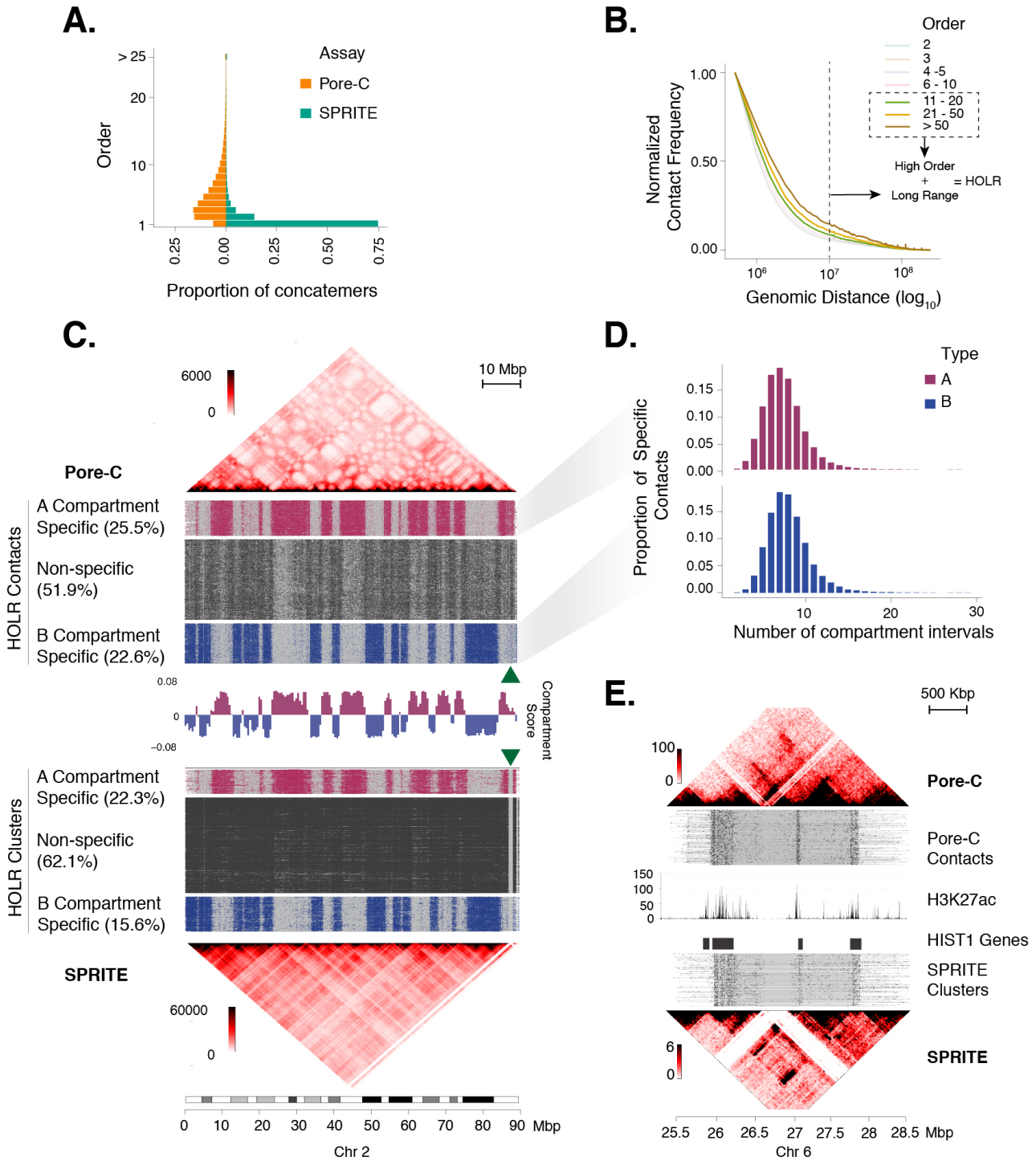

Fig. 2. Pore-C concatemers reveal higher-order long-range contacts (A) Comparison of the distribution of contact order between Pore-C concatemers and SPRITE clusters in GM12878. (B) Contact intensity as a function of linear genomic distance was plotted across all chromosomes for contacts of varying order. High-order concatemers (order > 10) showed a more gradual decay in contact intensity as a function of distance relative to lower orders. The plot depicts the thresholds for high-order (order > 10) long-range (distance > 10 Mbp) (HOLR) contacts used in the following panels. (C) HOLR contacts on chromosome 2p in GM12878 were called A-specific (>80% in A), B-specific (>80% in B), or non-specific according to their overlap with A and B compartment intervals. A and B compartment intervals were defined through segmentation of positive (A compartment) and negative (B compartment) values of the gold-standard Hi-C derived compartment score (middle track)[2]. HOLR contacts are plotted for these corresponding subgroups of Pore-C contacts (top half) and SPRITE cluster (bottom half) alignments. Percentages denote the frequency of A/B compartment-specific contacts among the set of HOLR contacts. These results show a higher fraction of A-specific and B-specific contacts in Pore-C, with additional combinatorial structure apparent in the non-specific Pore-C alignments that is less apparent in the corresponding non-specific SPRITE alignments. The pattern is also reflected in the virtual pairwise contact maps derived from Pore-C (top) and SPRITE (bottom) contacts. The green arrow indicates a peri-centromeric region that is absent in SPRITE but present in Pore-C due to its poor mappability with Illumina short reads. Giemsa banding pattern is shown below for scale and context. (D) Histograms of "compartment order" defined as the number of A or B compartments traversed by compartment specific Pore-C contacts. (E) Pore-C and SPRITE data are shown in the vicinity of a previously identified[36,42,43] focal high-order histone locus body hub which spatially apposes three clusters of *HIST1* genes on human chromosome 6. Pore-C contacts (top track) and SPRITE clusters (bottom track) containing at least one fragment in at least one of the three *HIST1* gene clusters combinatorially connect peaks of active enhancer activity, as delineated by GM12878 H3K27ac ChIPseq (middle track). HOLR: High-order long-range

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

chromatin contacts connect distant genomic regions. This is reflected in a lower decay exponent b in the pairwise interaction probability $P = ad^{-b}$ as a function of genomic distance d for multi-way vs. pairwise contacts[36,37]. To test this, we applied a generalized linear model to determine the effect of contact order on the decay exponent. We found that concatemers in the 11-20 ($P = 0.00138$, Wald test), 21-50 ($P = 7.8 \times 10^{-8}$), and >50 ($P = 7.6 \times 10^{-14}$) groups were associated with a significantly slower distance decay (**Fig. 2B**). We labeled these contacts (order > 10) as "high-order". Upon inspection of the distance-contact plot we labeled contacts with distance greater than 10 Mbp as "long-range".

To explore biological structures revealed by high-order and long-range (HOLR) contacts, we studied the combinatorial connectivity within and between A (active) and B (inactive) compartments revealed by both Pore-C and SPRITE (**Fig 2C**). Using A/B compartment definitions for GM12878 as defined by Rao *et al*[2], we identified A (or B) compartment-specific HOLR contacts as those with 80% alignment to the A (or B) compartment. In Pore-C concatemer libraries, 47.36% of HOLR contacts were compartment specific, with 19.78% and 27.58% demonstrating A- and B-specificity, respectively. In contrast, only 38.62% of SPRITE HOLR clusters demonstrated compartment specificity. We concluded that Pore-C HOLR contacts more accurately reflect direct molecular interactions relative to their SPRITE counterparts.

Plotting the alignment patterns of compartment-specific HOLR Pore-C contacts on chromosome 2p directly revealed the mutually exclusive A and B compartment anatomy (**Fig 2C**). Strikingly, compartment-specific Pore-C HOLR contacts were highly combinatorial, with the median contact intersecting 14 unique compartment intervals (range 2 - 29) with similar combinatorial complexity shown for A- and B–compartment-specific contacts (**Fig 2D**). Interestingly, visual inspection of the alignment patterns of non-compartment-specific HOLR Pore-C contacts revealed additional structures that most resembled B compartment alignments, but was distinct (**Fig 2C**). These results suggest that novel combinatorial compartment structures or inter-compartment dynamics may be revealed through the analysis of non-specific Pore-C HOLR contacts.

We found very subtle evidence of these additional structures among the 62.1 % of SPRITE HOLR clusters that were not compartment-specific; however, this signal was obscured by a substantial background of SPRITE clusters demonstrating uniform alignment along chromosome 2p, suggesting technical noise in the assay. Indeed, comparison of virtual pairwise contact maps across the region for Pore-C and SPRITE (across all concatemers / clusters) showed that SPRITE maps harbored a high burden of non-specific background (**Fig 2C**). In addition, SPRITE maps failed to resolve the compartment structure of a ~1 Mbp peri-centromeric region that is unmappable by short read sequencing (**Fig 2C**, green arrow). Pore-C HOLR alignments were able to firmly place this region into the active compartment of GM12878.

High-order chromatin structure is thought to be important to gene regulation and gene transcription, possibly through the induction of liquid-liquid phase separation enabling cooperative transcription factor kinetics[9,31,44]. Previously, SPRITE analyses provided evidence highof -order chromatin folding at a histone gene cluster on chromosome 6p, associated with the microscopically identifiable nuclear structure known as the histone locus body[36,42,43]. Compared to SPRITE, we observed clearer visualization of combinatorial connections between three clusters of histone genes (**Fig. 2E**) in Pore-C data. These foci of connectivity closely coincided with extended regions of H3K27ac (**Fig. 2E**), consistent with a chromatin configuration bringing multiple enhancers in 3D proximity. As with the compartment scale maps, Pore-C was also able to resolve structure in repetitive genomic regions that were unmappable and hence invisible to the (Illumia-based) SPRITE map (manifesting as white stripes in **Fig. 2E**).

High-order chromatin contacts resolve the allelic structure of a complex cancer rearrangement. Structural DNA variants are common in cancer genomes, yielding complex loci harboring many copies of genomic intervals and rearrangement junctions in *cis* or *trans* phase[45,46]. To explore the phase of highly rearranged cancer loci, we profiled the breast cancer cell line HCC1954 with Pore-C, generating 97.7 Gb of sequence and 677.89 million pairwise contacts with NlaIII and 46.5 Gb of sequence and 211.0 pairwise contacts with DpnII. Analysis of publicly available short read Illumina whole genome sequence (WGS) data of the breast cancer cell line HCC1954 and its paired blood normal derived cell line HCC1954BL using SvAbA (local assembly based somatic junction caller)[47], combined with our junction-balanced genome graph algorithm JaBbA[45], identified a subgraph of amplified rearrangement junctions connecting three loci on chromosomes 9, 12, and 20. Amplified regions in these three loci comprise four discontiguous "islands" of copy number > 10, three of which span more than 300 Kbp.

Deconvolution of the junction-balanced genome graph into distinct alleles yielded multiple possible solutions that were consistent with junction and interval copy number inferred from short-read Illumina WGS. Visualization of the virtual pairwise contact map derived from the Pore-C library demonstrates long-range connectivity between these distinct amplified regions **(Fig 3A, top)**. Integration of Pore-C high-order contacts with the possible alleles traversing this subgraph identified a candidate allele connecting these four intervals in *cis* as the most likely state **(Fig 3A, top)**. This allele was supported by 1,213 Pore-C concatemers with order greater than 10, spanning two or more intervals on the allele, many of which were separated by long distances (>1 Mbp) on the inferred allele. To validate this complex allele, we performed three color metaphase FISH on HCC1954 and the diploid ANA51 cells, designing probes against three discontiguous genomic regions overlapping the reference genomic location of the inferred allele (**Fig 3B, bottom**). As predicted from our Pore-C data, these three probes were proximal and amplified in the HCC1954 cell line, and distinct and diploid in ANA51 (**Fig 3B, top**). These results confirm the presence of a com-

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

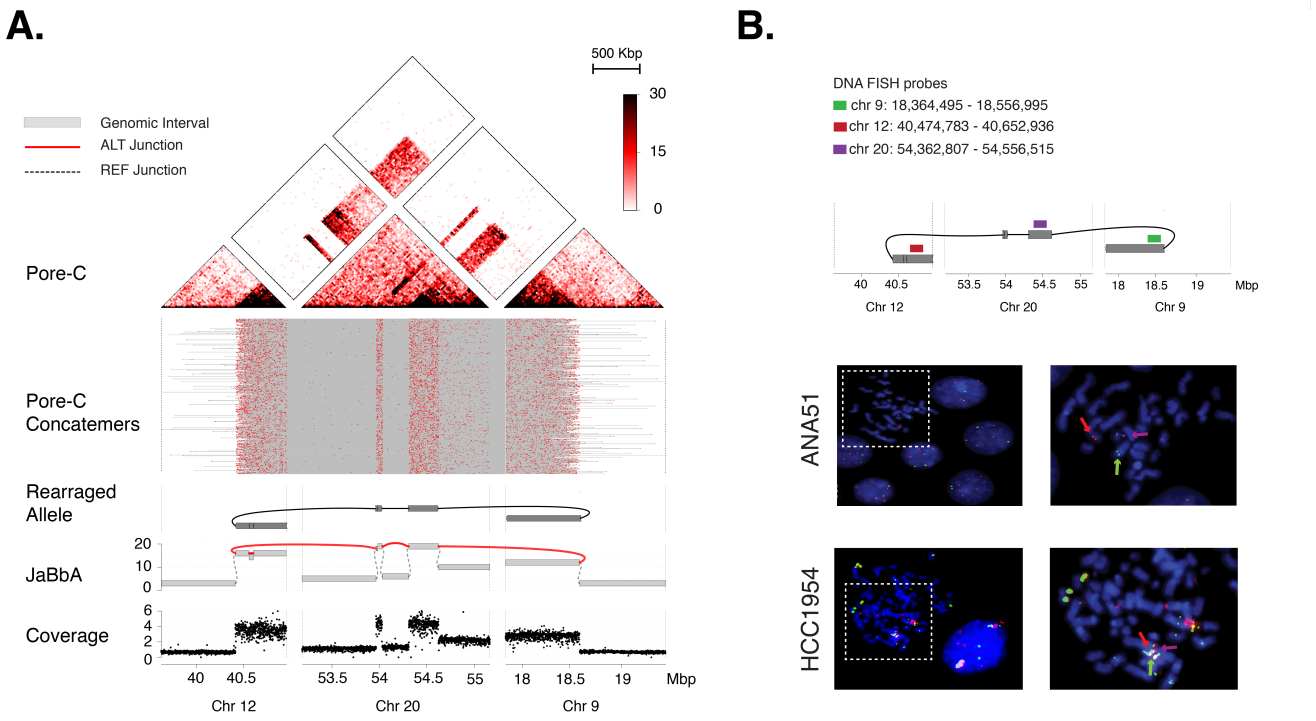

Fig. 3. High-order chromatin contacts resolve the allelic phase of complex genomic rearrangements (A) A complex, multi-chromosomal subgraph connecting chromosomes 9, 12 and 20 in the breast cancer cell line HCC1954 was identified using JaBbA, an algorithm for complex structural variant analysis (https://github.com/mskilab/JaBbA)[45]. Analysis of binned read-depth (bottom track) and junction calls yielded the genome graph (second track from bottom). At each edge, ALT represents an alternative junction that is depicted as a red line and REF represents a reference junction that is depicted as a gray dashed line. The gray bars each represent the genomic intervals of each connected node. Pore-C high-order contacts that mapped to the involved region (top track and heatmap) were used to nominate an amplified multi-junction derivative allele within this graph. The Pore-C alignments and associated contact maps (with each Pore-C read being highlighted in red and with gray lines representing the high-order connectivity between within each Pore-C contact) visually confirm this allele. (B) We designed a multi-color DNA FISH experiment (top) with red, green, and magenta probes each targeting one of three locations on this putative derivative allele. Co-localization of probes yielded a yellow signal in HCC1954 metaphase spreads (bottom), confirming the presence of an amplified multi-junction fusion of chromosomes 9, 12 and 20. We did not observe co-localization of these three probes in a diploid cell line (ANA51).

plex and amplified derivative allele simultaneously connecting portions of chromosomes 9, 12, and 20 in HCC1954.

Pore-C enables *de novo* assembly of chromosomal length scaffolds. In addition to the direct study of chromatin structure, chromatin conformation capture methods can be used to significantly improve the quality of *de novo* genome assemblies. Proximity information from Hi-C has been used to correct and merge contigs[50–52] and even generate reference-grade mammalian genome sequences[25]. To assess the utility of Pore-C reads in scaffolding human genome assemblies, we analyzed the HG002 GIAB reference sample[53]. Whole-genome sequencing (WGS) libraries from HG002 high molecular weight genomic DNA were sequenced to roughly 40-fold average read depth with a read N50 of 22 Kbp using ONT's promethION sequencing platform. We also generated three Pore-C libraries derived using the NlaIII, DpnII, and HindIII restriction enzymes. These were also run on the PromethION, and produced 218, 69, and 18 million virtual pairwise contacts, respectively.

The HG002 WGS dataset was processed using the `redbean` assembler which generated 2,374 contigs with an NG50 of 10.4 Mbp, with the longest contig spanning 82 Mbp (**Fig. 4A**). We then applied `salsa2`[49] to build scaffolds from these contigs. HindIII-derived Pore-C data yielded the longest scaffold (216 Mbp) with a nearly 10-fold increase in the sequence NG50 (98.6 Mbp) relative to the WGS data (**Fig. 4A**). Interestingly, although NlaIII-derived Pore-C libraries generated the highest number of high-order contacts (and hence pairwise contacts per Gbp), this data set only enabled a 3.2-fold improvement in the scaffold NG50 (33.2 Mbp) relative to the original assembly (**Fig. 4A**). DpnII demonstrated intermediate performance.

Visualization of HindIII Pore-C data for sequences derived from chromosome 4 demonstrated a high-frequency of inter-contig contacts (**Fig. 4B**). These were resolved following scaffolding, yielding a scaffold-scaffold contact map with the majority of interactions clustering at the diagonal (**Fig. 4C**). A dot plot revealed a high degree of synteny between HindIII Pore-C scaffolds and the hg38 chromosome 4 (**Fig. 4D**).

Discussion

To our knowledge, Pore-C is the first assay to combine chromosome conformation capture with nanopore sequencing in an amplification-free manner. Our data also represent the only mammalian-scale demonstration of a genome-wide multi-way chromatin conformation capture assay that effi-

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

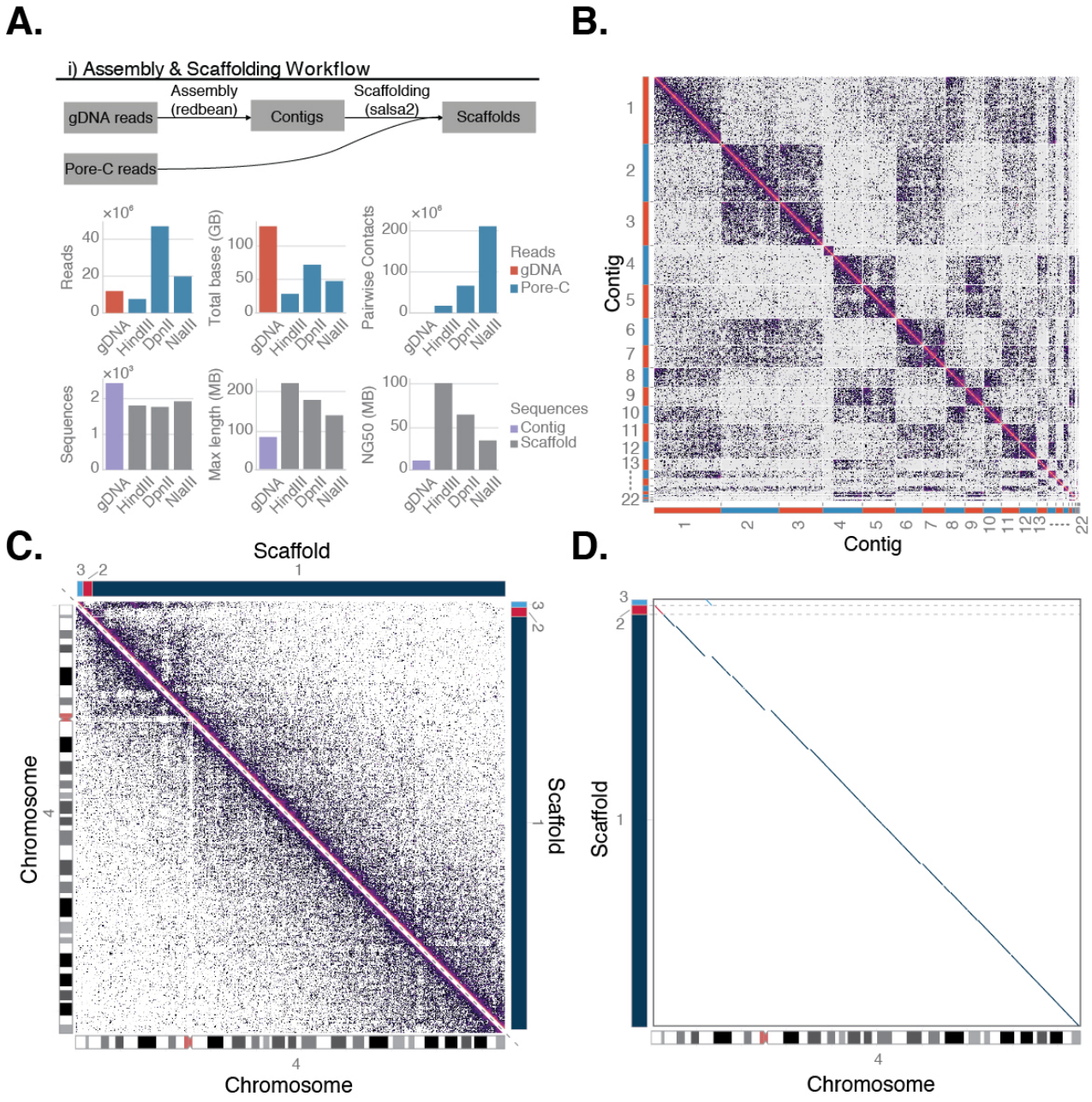

Fig. 4. Pore-C concatemers guide assembly correction and scaffolding (A) Assembly and scaffolding workflow was conducted in a 2-step process. ONT whole genome sequence data for HG002 was assembled using redbean[48], then the resulting contigs were provided to `salsa2`[49] along with pairwise Pore-C data generated from one of three different restriction enzymes to create a final scaffolded assembly. Despite having the fewest pairwise contacts the HindIII dataset created the most continuous scaffold. (B) Contact map showing the inputs to the scaffolding process for the HindIII dataset. Shown are the length-ordered contigs produced by redbean that map to chromosome 4. (C) Contact maps generated by mapping the HindIII Pore-C dataset against the scaffolds (above the diagonal) and the human reference (below the diagonal) for chromosome 4. The chromosome assembles into 3 scaffolds with 175 MB of 190MB in the largest scaffold, and although the centromere is not assembled, the two arms are properly placed in the scaffold relative to each other. (D) A dotplot of the scaffold aligned to GRCh38 shows that although there is a small misplaced scaffold, the assembly is largely correct.

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ciently (i.e. in greater than 1 % of the data) detects contacts with order > 3. Though ChIA-Drop[37] detects high-order contacts, it has been only been applied to *Drosophila melanogaster*, whose genome is an order of magnitude smaller than the human genome. SPRITE does not employ traditional chromatin conformation capture, but applies a ligation-free protocol. Relative to SPRITE, Pore-C shows a substantial (> 12 fold) enrichment of contacts with order > 3 (**Fig. 4D**). Other methods either have only been applied to the analysis of three-way interactions (Tri-C, GAM) and/or show rare representation of order > 3 contacts (COLA).

Through rigorous comparisons with Hi-C, we demonstrate high concordance between the probability of two sequences sharing a Pore-C concatemer and their Hi-C pairwise contact probability. As a result, Pore-C derived pairwise contact maps can be used to identify standard features of chromatin architecture, including A/B compartments, TADs, and loops. Based on our results, a single PromethION of a NlaIII Pore-C library can yield a contact map comprising >1 billion virtual pairwise contacts. These results are comparable to the yield from a ~2 billion read Illumina Hi-C library (based on the ~50% ligation frequency observed in Hi-C libraries). In this way, the Pore-C protocol represents a scalable alternative to a standard Hi-C experiment. Analytic caveats include the non-independence of virtual pairwise contacts[33] arising from the same concatemer, which may require correction (e.g. down-weighting) when determining the effective resolution of a Pore-C contact map.

A key feature of Pore-C is its ability to detect high-order chromatin structure. Our analyses of contact order, A/B compartment connectivity, and histone enhancer hubs demonstrates similar or superior performance of Pore-C relative to SPRITE[36]. We demonstrate increased compartment specificity among Pore-C HOLR contacts relative to SPRITE, which when plotted along the genome allows direct visualization of the A or B compartment boundaries in GM12878. The complex high-order intra-compartment structure highlighted in this work will require further analytic exploration, including dissection of the combinatorial logic underlying the association of certain subsets of these multi-megabase reference genomic regions and their impact on gene regulation or cell identity. Intriguingly, we find additional structure among the ~50% of HOLR Pore-C contacts that do not show A or B compartment specificity. A key question is whether these additional patterns arise from distinct subpopulations of cells (e.g. cell cycle phases) or represent stochastic inter-compartmental crosstalk, like those that dynamically arise during gene regulation. Further algorithmic development and experimental work will be needed to explore these possibilities in Pore-C data, and nominate hubs of interactions (e.g. those arising from superenhancer-associated liquid liquid phase separations) genome-wide in a statistically calibrated manner.

The lower degree of compartment-specificity and overall combinatorial structure observed among HOLR SPRITE contacts relative to Pore-C may be due to technical features of SPRITE's protocol. Since SPRITE employs ligation-free split-pool barcoding of variably-sized fragments of cross-linked chromatin, high-order SPRITE clusters may be effectively sampling very large spatial radius in the nucleus, traversing multiple compartments and/or chromosomes. Indeed, the combinatorial distance decay that we have shown higher-order Pore-C contacts is sharper than what has been previously shown for SPRITE or ChIA-drop[36,37]. By relying on the proximity ligation reaction, high-order Pore-C contacts may be less prone to connecting very spatially distant nuclear regions and thus may provide a more pure readout of spatial adjacency, and hence combinatorial chromatin complexes. Unlike droplet-microfluidic based approaches (ChIA-drop), Pore-C does not suffer from barcode collisions (i.e. from multiple library fragments occupying the same droplet and acquiring the same barcode) which can potentially inflate distance statistics, particularly for "high-order barcodes".

Our results demonstrate the potential of high-order contacts to inform the reconstruction of complex rearranged cancer loci and *de novo* genome assembly. The data derived from high-order chromatin contacts in HCC1954 represent a sort of "ultra-long linked read", in that each contact represents a set of (likely) contiguous sequences that are near each other in sequence. However, while 10X Chromium linked-reads are generated from fragments with average length of 50-80 Kbp, we show that high-order Pore-C contacts link multiple sequences across megabases of DNA. Furthermore, 10X Chromium linked-read libraries often suffer from "barcode collisions", resulting from the fact 5-10 other fragments share a microfluidic droplet and the resulting barcode. These ambiguities require statistical criteria (i.e. multiple fragments spanning a structural variant breakpoint) and / or a minimally divergent reference genome (e.g. grouping read clouds according to mapping location) to resolve, which can impede complex structural variant detection or *de novo* assembly.

The vast majority of sequences contributing to proximally-ligated concatemers, and hence Pore-C contacts, arise in *cis* from a single contiguous molecule. As a result, we conjecture that Pore-C contacts will provide a less ambiguous source of long-range contiguity or phase information than 10X Chromium linked-reads. Though our assembly results demonstrate chromosomal length scaffolds from Pore-C using assembly tools that have been developed for Hi-C (`salsa2`), additional performance improvement will likely result from more specialized algorithms that fully leverage high-order Pore-C contacts to build longer scaffolds, as well as discern long-range phase blocks. Such algorithmic improvements may specifically impact the assembly and phasing of cancer genomes, where copy number alterations and rearrangements conspire to generate many copies of very similar alleles, or in high-ploidy domesticated plants such as wheat where broadly spanning phase information could link rare unique features throughout the genome, and significantly increase the genome assembly contiguity.

Unlike previous approaches combining proximity-based ligation with nanopore sequencing, the Pore-C protocol is amplification-free. As a result, the profiled sequences retain any DNA modifications (e.g. DNA methylation) that may

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

be present on the input DNA molecule. Though we do not explore the analysis of base modifications on Pore-C reads, these signals could theoretically be used to probe the relationship between methylation states and combinatorial chromatin structures. Such an approach would serve as an alternative to Illumina-based bisulfite-sequencing based protocols for 3D methylation profiling (e.g. Hi-Culfite[54], Methyl Hi-C[55]). Additional pipelines will be required to integrate methylation (and other base modifications detectable on nanopore reads) into the analysis of Pore-C contact data. We believe that the development of these and other algorithmic innovations will greatly expand the utility of this powerful and scalable assay for the study of high-order chromatin organization, complex cancer rearrangements, and *de novo* genome assembly.

Methods

Cell culture. Human B lymphocytes (GM12878) [Coriell Institute] and HG002 cells [Coriell Institute] were each maintained in suspension at a concentration greater than 200,000 cells/mL in RPMI-1640 [ATCC 30-2001] supplemented with 15% fetal bovine serum (FBS) and 1% penicillin/streptomycin. HCC1954 [ATCC CRL-2388] human breast cancer cells were maintained using a subculture ratio of 1:4 every three days in RPMI-1640 (ATCC 30-2001) supplemented with 10% FBS and 1% penicillin/streptomycin.

Cross-linking. 10 million cells were washed three times in chilled 1X phosphate buffered saline (PBS) in a 50 mL centrifuge tube, pelleted by centrifugation at 500*xg* for 5 min at 4°C between each wash. Cells were resuspended in 10 mL room temperature 1X PBS 1% formaladehyde [EMD Millipore cat no. 818708] by gently pipetting with a wide bore tip, then incubated at room temperature for 10 min. To quench the cross-linking reaction 527 µL of 2.5 M glycine was added to achieve a final concentration of 1% w/v or 125 mM in 10.5 mL. Cells were incubated for 5 min at room temperature followed by 10 min on ice. The cross-linked cells were pelleted by centrifugation at 500*xg* for 5 min at 4°C.

Restriction enzyme digest. Each cell pellet was resuspended in a mixture of 50 µL of protease inhibitor cocktail [Sigma Aldrich cat no. P8340] in 500 µL of cold permeabilization buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% IGEPAL CA-630) and placed on ice for 15 min. Cells were centrifuged at 500*xg* for 10 min at 4°C after which the supernatant was aspirated and replaced with 200 µL of chilled 1.5X digestion reaction buffer [NEB] compatible with the restriction enzyme used. Cells were centrifuged again at 500*xg* for 10 min at 4°C, then aspirated and re-suspended in 300 µL of chilled 1.5X digestion reaction buffer. To denature the chromatin, 33.5 µL of 1% w/v SDS [Thermo Fisher Scientific cat no. 15553027] was added to each cell suspension and incubated for exactly 10 min at 65°C with gentle agitation then placed on ice immediately afterwards. To quench the SDS 37.5 µL of 10% v/v Triton X-100 [Sigma Aldrich cat no. 93443] was added for a final concentration of 1%, followed by incubation for 10 min on ice. Permeablized cells were then digested with a final concentration of 1 U/µL of either DpnII, NlaIII or HindIII [NEB] brought to volume with nuclease-free water to achieve a final 1X digestion reaction buffer in 450 µL. Cells were then mixed by gentle inversion. Cell suspensions were incubated in a thermomixer at 37°C for 18 hours with periodic < 1000 rpm rotation (< 30 sec every 15 min) to prevent condensation inside the lid.

Proximity ligation and reverse cross-linking. DpnII and NlaIII restriction digests were heat inactivated at 65°C for 20 min with 300 rpm rotation. HindIII digests were chemically inactivated with final concentration of 0.1% w/v SDS at 65°C for 20 min with 300 rpm rotation, then quenched with a final concentration of 1% v/v Triton X-100 [Sigma Aldrich cat no. 93443]. Proximity ligation was set up at room temperature with the addition of the following reagents: 100 µL of 10X T4 DNA ligase buffer [NEB], 10 µL of 10 mg/mL BSA and 50 µL of T4 Ligase [NEB M0202L] in a total volume of 1000 µL with nuclease-free water. The ligation was cooled to 16°C and incubated for 6 hours with gentle rotation.

Protein degradation and DNA purification. To reverse cross-link, samples were treated with 100 µL 20 mg/mL Proteinase K [Thermo Fisher Scientific cat no. 25530049], 100 µL 10% SDS [Thermo Fisher Scientific cat no. 15553027] and 500 µL 20% v/v Tween-20 [Sigma Aldrich cat no. P9416] in a total volume of 2000 µL with nuclease-free water. Samples were incubated in a thermomixer at 56°C for 18 hours with < 1000 rpm rotation (< 30 sec every 15 min) to prevent condensation inside the lid. In order to purify DNA, the sample was transferred to a 5 mL centrifuge tube, rinsing the original tube with a further 200 µL of nuclease-free H_2O to collect any residual sample, bringing the total sample volume to 2.2 mL. DNA was then purified from the sample using a standard phenol chloroform extraction and ethanol precipitation.

Nanopore sequencing. Purified DNA was SPRI size selected for fragments > 1.5 Kbp, then prepared for sequencing using either Oxford Nanopore Technologies SQK-LSK109, and sequenced on ONT's MinION, GridION or PromethION platforms. In total, 27 sequencing runs were conducted generating a total of 449 Gbases of raw sequence from 3 different tissue sources including the human cell lines GM12878 and HG002 as well as the human breast cancer cell line HCC1954. The runs produced a total of 2.4 billion pairwise contacts. They were conducted using either DpnII, HindIII, or NlaIII for restriction enzymes in order to optimize the Pore-C methodology. The complete details for the runs are seen in (**Table S1**).

Pore-C pipeline. We developed a reproducible analytic pipeline for deriving multi-way chromatin contacts from Pore-C concatemers (`https://github.com/nanoporetech/pore-c/`), and a workflow that can be found at (`https://github.com/nanoporetech/Pore-C-Snakemake/`) which uses the snakemake framework[56]. Briefly, the workflow involves alignment of Pore-C

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

reads to a reference genome with Pore-C specific parameters. The resulting alignments are filtered to identify a minimal alignment set that covers each Pore-C read. Each of the resulting filtered alignments is then annotated with the set of reference genome restriction fragments with which it overlaps based on a *in-silico* restriction digest. From those fragment sets, each alignment is assigned to a single fragment based on the position of the alignment midpoint, which also serves to remove direct non-chimeric pairs that are caused by either incomplete digestion of the sample or ligation of cognate free ends. With the assignment of each alignment along a Pore-C read to a single restriction fragment, a multi-way contact is constituted, and this multi-way contact can then be decomposed into pairwise contacts. Decomposition to pairwise contacts involves enumerating the number of multi-way contacts containing a given pair of restriction fragments for every possible pair of restriction fragments in the genome. The fraction of alignments falling a given distance from the predicted *in silico* restriction sites in the GRCh38 reference genome were computed for DpnII, NlaIII and HindIII Pore-C datasets. The pairwise contacts are assigned to genome bins using conventional Hi-C workflows to generate a contact map, such as Hi-C explorer[57] and cooler[58]. The workflow also includes an optional step to output the contact data in the bed format compatible with `salsa2`[49] for genome assembly correction and scaffolding as well as the "medium text format" for conversion to the binary .hic format generated by the `juicer-tools pre` command (`https://github.com/aidenlab/juicer/wiki/Download/`) for use with the juicebox toolkit. Commands and parameters used for each step can be found in (table S2).

Pore-C and Hi-C comparisons. Pore-C concatemers were decomposed into sets of virtual pairwise contacts in order to make correlations with existing Hi-C datasets. The resulting decomposed Pore-C contact matrix could be treated as a Hi-C pairwise contact matrix and explored using existing Hi-C analytic tools. The Pore-C aggregate dataset was composed of 11 NlaIII GM12878 Pore-C sequencing runs containing a total of 1.38 billion virtual pairwise contacts and the Hi-C dataset (4DNFIXP4QG5B)[2] contained 4 billion contacts. Matrix balancing was performed on both the Pore-C and Hi-C contact matrices using the Knight-Ruiz iterative correction algorithm[59] using `cooler balance`. A linear correlation between the Pore-C and Hi-C contact matrices measured three different similarity metrics i) raw contact matrices, ii) compartmental eigenvector scores, identified using `cooltools call-compartments` and iii) TAD insulation scores, calculated using the `cooltools diamond-arrowhead` tools. To further examine any variations between individual runs, the same three comparisons were made between each of the individual Pore-C runs, the complete Hi-C data set, and downsampled subsets of the Hi-C datasets. Downsampled Hi-C datasets were generated to match the contact count of each individual Pore-C run. Raw contact matrix, AB compartment and TAD correlations were performed between the down-sampled sets and the whole set to establish an upper bound for the correlation between any dataset of that size and the whole set. Publicly available ENCODE ChIP-seq peaks for H3K27ac, H3K4me3 and H3K4me1 (ENCSR447YYN) were used to visually confirm the correlation of Pore-C A/B compartmentalization with other existing markers of active / inactive chromatin structure. ENCODE CTCF ChiP-seq peaks (ENCSR000AKB) were utilized to identify loop anchors. The juicer suites APA tool was used to test for aggregate enrichment of Pore-C decomposed pairwise contacts within loop anchors identified in the Rao *et al.* GM12878 Hi-C dataset[2].

Analysis of high-order contacts. The change in pairwise contact frequencies as a function of genomic distance was determined by dividing concatemers into several groups based on order (2, 3, 4-5, 6-10, 11-20, 21-50, >50). The concatemers were then decomposed into pair-wise contacts and the normalized frequency of contacts as a function of genomic distance was plotted. This resulted into curves for each group. To statistically test the gradual decay of higher order contacts, a gamma-Poisson regression model was set up fitting raw frequency counts as: $count \sim log(total\ contacts) + group + log(distance)$ with group comprising of 2-way contacts as the baseline. Wald-test from MASS R package[60] was used to analyze significant deviations of different groups from baseline.

The high order long range (HOLR) contacts were defined if the maximum distance between the contacts for that concatemer, per chromosome, exceeded distance threshold of 10 Mbp (long range) and they belonged to group with order > 10 (high order).

For compartment-specificity analysis, each HOLR concatemer was assigned a label based on its overlap with A/B compartments defined using the compartment eigenvector scores derived from Hi-C dataset[2] for the GM12878 cell line. The number of contacts falling in either compartment was tallied and proportions were calculated for each concatemer for each chromosome. Singleton contacts were filtered out. Focusing on multi-megabase level compartmentalization across the p-arm of chromosome 2, the performance of Pore-C to detect higher order interactions was assessed. For both Pore-C and SPRITE, a contact map using all Pore-C contacts or SPRITE clusters was created. Additionally, for each assay a read pile-up of Pore-C HOLR contacts or SPRITE HOLR clusters for A-specific, B-specific and unspecific HOLR contacts were plotted. The criteria of 80% or more contacts falling in a given compartment was used to determine compartment specificity. Lastly, compartment order for all HOLR A/B specific contacts were determined as the number of A or B compartments intersected by compartment specific Pore-C contacts.

For the analysis pertaining to *HIST1* gene cluster found on chromosome 6, previously described in[42] and visualized using the SPRITE dataset[36], both the SPRITE clusters and Pore-C contacts were filtered such that at least one fragment fell within one of these *HIST1* genes. The resulting read pile-ups for each technology were plotted along with the contact maps derived from each, as well as boxes to demarcate *HIST1* genomic locations.

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We obtained GM12878 SPRITE data from (4DNFI-UOOYQC3)[36]. The tabular contact data was ingested using the R programming language in combination with a set of R/Bioconductor packages (including `GenomicRanges`, `rtracklayer`, `data.table`). All analyses of high order contacts were performed using Imielinski Lab R packages hosted at `https://github.com/mskilab/` (`gUtils`, `bamUtils`, `gTRack`, `gGnome`, `GxG`) to represent, manipulate, analyze, and visualize genomic intervals, high-order contacts, and 1D / 2D genomic tracks.

Validation of rearrangements in HCC1954 using Pore-C. Short read Illumina whole genome sequencing data for the breast cancer cell line HCC1954 and its matched normal HCC1954BL was obtained from the genomic data commons (GDC, https://gdc.cancer.gov/). DNA rearrangement junctions were detected using SvAbA[47] and binned read depth was assessed using `fragCounter` (`http://github.com/mskilab/fragCounter`). Junction and read depth data were input into a junction-balanced genome graph inference algorithm JaBbA (`https://github.com/mskilab/JaBbA`)[45]. JaBbA integrates read depth and structural variation information for a given genome and produces a junction-balanced genome graph by solving a constrained optimization problem[45]. The junction-balanced graph allows querying of non-contiguous regions of genome that become connected through rearrangements through the R package `gGnome` (`https://github.com/mskilab/gGnome`). A sub-graph of amplified rearrangement junctions connecting three loci on chromosomes 9, 12, and 20 was identified. Using the sub-graph, all possible alleles that can give rise to the re-arrangement were combinatorially enumerated. Pore-C concatemer support was assessed for each derived allele. The allele with most concatemer support was selected and validated by FISH.

Fluorescence *in situ* hybridization (FISH). The fusion and amplification event detected between chromosomes 9, 12 and 20 were validated using three-color DNA fluorescence in situ hybridization (FISH). FISH probes were designed against loci of interest using the BAC clones (`https://bacpacresources.org`) RP11-341L12 (directly labeled green), RP11-476D10 (directly labeled red) and RP11-762A13 (directly labeled magenta). The HCC1954 and ANA51 cell lines were treated with Colcemid (0.1 g/ml) for 1 hour to obtain a metaphase preparation. Cells were then fixed in methanol/acetic acid (3:1). The fixed cells were then allowed to dry for 1 week on a microscope slide. Prepped slides were placed in a denaturation solution (70% formamide/2x SSC) for 15 minutes at 73°C proceeded by ethanol dehydration. Directly labeled probes were placed onto the slides and cells on the slide were then denatured for 15 minutes at 74°C proceeded by hybridization for 72 hours at 37°C. After hybridization, slides were washed in 2xSSC/0.3% NP-40, first at 73°C and then at 23°C. Metaphase cells were then stained with DAPI before visualization. A fusion-amplification event was measured as several green, red and magenta signals overlapping per nucleus or chromosome. Prior to use on HCC1954, all BAC clones were validated on metaphase spreads using the diploid ANA51 cell line. The nuclei and metaphase spreads were observed using a fluorescence microscope (Olympus BX51; Olympus Optical, Tokyo, Japan). Cytovision and Fiji software were used for imaging.

Pore-C hybrid genome assembly. HG002 high-molecular weight (HMW) DNA was extracted from 20 million cells using the Puregene extraction kit [Qiagen cat. no 158667]. The sample was size-selected using the Circulomics Short Read Eliminator kit [Circulomics cat no. SS-100-101-01]. The HMW DNA was then sheared using a Megaruptor with a shearing speed of 28 and a library was prepared using a ONT library preparation kit [ONT cat no. SQK-LSK109]. The library was sequenced using the PromethION platform with nuclease flushes performed at 24 and 48 hours. Basecalling was performed using Guppy [version 3.2.2] using high accuracy mode. The sequencing run generated a total of 143.31 Gbp of sequence data with a read N50 of 22 Kbp. Reads shorter than 6 Kbp or with a mean read quality less than Q8 were removed, leaving 112.3 Gbp of raw sequence data. These reads were subjected to adapter trimming with PoreChop (`https://github.com/rrwick/Porechop`) and assembled with redbean[48]. This draft assembly was polished with 3 rounds of Racon[61] followed by a round of Medaka polishing (`https://github.com/nanoporetech/medaka`). Pore-C libraries for each of the three restriction enzymes were generated from HG002 cells and sequenced on individual PromethION flow cells (**Table S2**). The virtual pairwise contacts for each enzyme were converted to a bed format compatible with the `salsa2`[49] tool, which then converted the draft assembly to chromosome-scale scaffolds.

Software Availability. Software used in this paper can be found in the following GitHub repositories:

https://github.com/nanoporetech/pore-c
https://github.com/mskilab/gGnome
https://github.com/mskilab/GxG

ACKNOWLEDGEMENTS

We thank Jane Skok for helpful comments on the manuscript. Marcin Imielinski is supported by a Burroughs Wellcome Fund Career Award for Medical Scientists, Doris Duke Clinical Foundation Clinical Scientist Development Award, Starr Cancer Consortium Award, Melanoma Research Alliance Team Science Award, and National Institutes of Health U24-CA15020.

AUTHOR CONTRIBUTIONS

These contributions follow the Contributor Roles Taxonomy guidelines: https://casrai.org/credit/. Conceptualization: N.U., M.I.; Data curation: N.U, M.P.,E.H., M.I.; Formal analysis: N.U., M.P., A.D., E.H., M.I.; Funding acquisition: S.J., D.T., M.I.; Investigation: N.U., M.P., A.D., S.S., X.D., E.H., S.J., M.I. Methodology: N.U., M.P, A.D., S.S, D.S., C.T., S.J., E.H., M.I.; Project administration: E.H., M.I.; Resources: S.J, M.I.; Software: N.U., M.P., E.H., A.D., M.I. Supervision: S.J., M.I.; Validation: N.U., M.P., A.D, X.D., S.K., E.A., D.W., J.M.M, E.H., S.J., M.I. Visualization: N.U., M.P., A.D., M.I. Writing – original draft: N.U., M.P., M.I. Writing – review & editing: all authors.

COMPETING FINANCIAL INTERESTS

M.P., S.S., X.D., C.T., P.R., D.S., D.T., S.J. and E.H. are employees of, and stock option holders in, Oxford Nanopore Technologies.

Bibliography

1. Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner,

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, oct 2009. ISSN 1095-9203. doi: 10.1126/science.1181369.

2. Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, dec 2014. doi: 10.1016/j.cell.2014.11.021.
3. Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, apr 2012. doi: 10.1038/nature11082.
4. Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A Mirny. Formation of chromosomal domains by loop extrusion. *Cell reports*, 15 (9):2038–2049, may 2016. doi: 10.1016/j.celrep.2016.04.085.
5. Natalia Naumova, Maxim Imakaev, Geoffrey Fudenberg, Ye Zhan, Bryan R Lajoie, Leonid A Mirny, and Job Dekker. Organization of the mitotic chromosome. *Science*, 342(6161):948–953, nov 2013. doi: 10.1126/science.1236083.
6. Johan H Gibcus and Job Dekker. The hierarchy of the 3D genome. *Molecular Cell*, 49(5): 773–782, mar 2013. doi: 10.1016/j.molcel.2013.02.011.
7. Amartya Sanyal, Bryan R Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, sep 2012. doi: 10.1038/nature11279.
8. Ralph Stadhouders, Guillaume J. Filion, and Thomas Graf. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature*, 569:345–354, 2019. ISSN 0028-0836. doi: 10.1038/s41586-019-1182-7.
9. Elizabeth H. Finn and Tom Misteli. A genome disconnect. *Nature Genetics*, 51(8):1205–1206, August 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0476-x.
10. Darío G. Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M. Opitz, Renata Laxova, Fernando Santos-Simarro, Brigitte Gilbert-Dussardier, Lars Wittler, Marina Borschiwer, Stefan A. Haas, Marco Osterwalder, Martin Franke, Bernd Timmermann, Jochen Hecht, Malte Spielmann, Axel Visel, and Stefan Mundlos. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, may 2015. ISSN 10974172. doi: 10.1016/j.cell.2015.04.004.
11. Martin Franke, Daniel M Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, Rieke Kempfer, Ivana Jerković, Wing-Lee Chan, Malte Spielmann, Bernd Timmermann, Lars Wittler, Ingo Kurth, Paola Cambiaso, Orsetta Zuffardi, Gunnar Houge, Lindsay Lambie, Francesco Brancati, Ana Pombo, Martin Vingron, Francois Spitz, and Stefan Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624):265–269, oct 2016. doi: 10.1038/nature19800.
12. Claire Redin, Harrison Brand, Ryan L Collins, Tammy Kammin, Elyse Mitchell, Jennelle C Hodge, Carrie Hanscom, Vamsee Pillalamarri, Catarina M Seabra, Mary-Alice Abbott, Omar A Abdul-Rahman, Erika Aberg, Rhett Adley, Sofia L Alcaraz-Estrada, Fowzan S Alkuraya, Yu An, Mary-Anne Anderson, Caroline Antolik, Kwame Anyane-Yeboa, Joan F Atkin, Tina Bartell, Jonathan A Bernstein, Elizabeth Beyer, Ian Blumenthal, Ernie M H F Bongers, Eva H Brilstra, Chester W Brown, Hennie T Brüggenwirth, Bert Callewaert, Colby Chiang, Ken Corning, Helen Cox, Edwin Cuppen, Benjamin B Currall, Tom Cushing, Dezso David, Matthew A Deardorff, Annelies Dheedene, Marc D'Hooghe, Bert B A de Vries, Dawn L Earl, Heather L Ferguson, Heather Fisher, David R FitzPatrick, Pamela Gerrol, Daniela Giachino, Joseph T Glessner, Troy Gliem, Margo Grady, Brett H Graham, Cristin Griffis, Karen W Gripp, Andrea L Gropman, Andrea Hanson-Kahn, David J Harris, Mark A Hayden, Rosamund Hill, Ron Hochstenbach, Jodi D Hoffman, Robert J Hopkin, Monika W Hubshman, A Micheil Innes, Mira Irons, Melita Irving, Jessie C Jacobsen, Sandra Janssens, Tamison Jewett, John P Johnson, Marjolijn C Jongmans, Stephen G Kahler, David A Koolen, Jerome Korzelius, Peter M Kroisel, Yves Lacassie, William Lawless, Emmanuelle Lemyre, Kathleen Leppig, Alex V Levin, Haibo Li, Hong Li, Eric C Liao, Cynthia Lim, Edward J Lose, Diane Lucente, Michael J Macera, Poornima Manavalan, Giorgia Mandrile, Carlo L Marcelis, Lauren Margolin, Tamara Mason, Diane Masser-Frye, Michael W McClellan, Cinthya J Zepeda Mendoza, Björn Menten, Sjors Middelkamp, Liya R Mikami, Emily Moe, Shehla Mohammed, Tarja Mononen, Megan E Mortenson, Graciela Moya, Aggie W Nieuwint, Zehra Ordulu, Sandhya Parkash, Susan P Pauker, Shahrin Pereira, Danielle Perrin, Katy Phelan, Raul E Piña Aguilar, Pino J Poddighe, Giulia Pregno, Salmo Raskin, Linda Reis, William Rhead, Debra Rita, Ivo Renkens, Filip Roelens, Jayla Ruliera, Patrick Rump, Samantha L P Schilit, Ranad Shaheen, Rebecca Sparkes, Erica Spiegel, Blair Stevens, Matthew R Stone, Julia Tagoe, Joseph V Thakuria, Bregje W van Bon, Jiddeke van de Kamp, Ineke van Der Burgt, Ton van Essen, Conny M van Ravenswaaij-Arts, Markus J van Roosmalen, Sarah Vergult, Catharina M L Volker-Touw, Dorothy P Warburton, Matthew J Waterman, Susan Wiley, Anna Wilson, Maria de la Concepcion A Yerena-de Vega, Roberto T Zori, Brynn Levy, Han G Brunner, Nicole de Leeuw, Wigard P Kloosterman, Erik C Thorland, Cynthia C Morton, James F Gusella, and Michael E Talkowski. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.*, 49(1):36–45, January 2017.
13. Jesse R Dixon, Jie Xu, Vishnu Dileep, Ye Zhan, Fan Song, Victoria T Le, Galip Gürkan Yardımcı, Abhijit Chakraborty, Darrin V Bann, Yanli Wang, Royden Clark, Lijun Zhang, Hongbo Yang, Tingting Liu, Sriranga Iyyanki, Lin An, Christopher Pool, Takayo Sasaki, Juan Carlos Rivera-Mulia, Hakan Ozadam, Bryan R Lajoie, Rajinder Kaul, Michael Buckley, Kristen Lee, Morgan Diegel, Dubravka Pezic, Christina Ernst, Suzana Hadjur, Duncan T Odom, John A Stamatoyannopoulos, James R Broach, Ross C Hardison, Ferhat Ay, William Stafford Noble, Job Dekker, David M Gilbert, and Feng Yue. Integrative detection and analysis of structural variation in cancer genomes. *Nature Genetics*, 50(10):1388–1398, oct 2018. ISSN 1061-4036. doi: 10.1038/s41588-018-0195-8.
14. Stefan Gröschel, Mathijs A Sanders, Remco Hoogenboezem, Elzo de Wit, Britta A M Bouwman, Claudia Erpelinck, Vincent H J van der Velden, Marije Havermans, Roberto Avellino, Kirsten van Lom, Elwin J Rombouts, Mark van Duin, Konstanze Döhner, H Berna Beverloo, James E Bradner, Hartmut Döhner, Bob Löwenberg, Peter J M Valk, Eric M J Bindels, Wouter de Laat, and Ruud Delwel. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*, 157(2):369–381, April 2014.
15. Mark W Zimmerman, Yu Liu, Shuning He, Adam D Durbin, Brian J Abraham, John Easton, Ying Shao, Beisi Xu, Shizhen Zhu, Xiaoling Zhang, Zhaodong Li, Nina Weichert-Leahey, Richard A Young, Jinghui Zhang, and A Thomas Look. MYC drives a subset of high-risk pediatric neuroblastomas and is activated through mechanisms including enhancer hijacking and focal enhancer amplification. *Cancer discovery*, 8(3):320–335, 2018. doi: 10.1158/2159-8290.{CD}-17-0993.
16. Paul A Northcott, Catherine Lee, Thomas Zichner, Adrian M Stütz, Serap Erkek, Daisuke Kawauchi, David J H Shih, Volker Hovestadt, Marc Zapatka, Dominik Sturm, David T W Jones, Marcel Kool, Marc Remke, Florence M G Cavalli, Scott Zuyderduyn, Gary D Bader, Scott VandenBerg, Lourdes Adriana Esparza, Marina Ryzhova, Wei Wang, Andrea Wittmann, Sebastian Stark, Laura Sieber, Huriye Seker-Cin, Linda Linke, Fabian Kratochwil, Natalie Jäger, Ivo Buchhalter, Charles D Imbusch, Gideon Zipprich, Benjamin Raeder, Sabine Schmidt, Nicolle Diessl, Stephan Wolf, Stefan Wiemann, Benedikt Brors, Chris Lawerenz, Jürgen Eils, Hans-Jörg Warnatz, Thomas Risch, Marie-Laure Yaspo, Ursula D Weber, Cynthia C Bartholomae, Christof von Kalle, Eszter Turányi, Peter Hauser, Emma Sanden, Anna Darabi, Peter Siesjö, Jaroslav Sterba, Karel Zitterbart, David Sumerauer, Peter van Sluis, Rogier Versteeg, Richard Volckmann, Jan Koster, Martin U Schuhmann, Martin Ebinger, H Leighton Grimes, Giles W Robinson, Amar Gajjar, Martin Mynarek, Katja von Hoff, Stefan Rutkowski, Torsten Pietsch, Wolfram Scheurlen, Jörg Felsberg, Guido Reifenberger, Andreas E Kulozik, Andreas von Deimling, Olaf Witt, Roland Eils, Richard J Gilbertson, Andrey Korshunov, Michael D Taylor, Peter Lichter, Jan O Korbel, Robert J Wechsler-Reya, and Stefan M Pfister. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*, 511(7510):428–434, jul 2014. doi: 10.1038/nature13379.
17. Russell J H Ryan, Yotam Drier, Holly Whitton, M Joel Cotton, Jasleen Kaur, Robbyn Issner, Shawn Gillespie, Charles B Epstein, Valentina Nardi, Aliyah R Sohani, Ephraim P Hochberg, and Bradley E Bernstein. Detection of enhancer-associated rearrangements reveals mechanisms of oncogene dysregulation in b-cell lymphoma. *Cancer discovery*, 5(10):1058–1071, oct 2015. doi: 10.1158/2159-8290.{CD}-15-0370.
18. Joachim Weischenfeldt, Taronish Dubash, Alexandros P Drainas, Balca R Mardin, Yuanyuan Chen, Adrian M Stütz, Sebastian M Waszak, Graziella Bosco, Ann Rita Halvorsen, Benjamin Raeder, Theocharis Efthymiopoulos, Serap Erkek, Christine Siegl, Hermann Brenner, Odd Terje Brustugun, Sebastian M Dieter, Paul A Northcott, Iver Petersen, Stefan M Pfister, Martin Schneider, Steinar K Solberg, Erik Thunissen, Wilko Weichert, Thomas Zichner, Roman Thomas, Martin Peifer, Aslaug Helland, Claudia R Ball, Martin Jechlinger, Rocio Sotillo, Hanno Glimm, and Jan O Korbel. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nature Genetics*, 49(1):65–74, 2017. doi: 10.1038/ng.3722.
19. Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, feb 2002. ISSN 1095-9203. doi: 10.1126/science.1067799.
20. Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G Y Chew, Phillips Yao Hui Huang, Willem-Jan Welboren, Yuyuan Han, Hong Sain Ooi, Pramila N Ariyaratne, Vinsensius B Vega, Yanquan Luo, Peck Yean Tan, Pei Ye Choy, K D Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Roy Joseph, Haixia Li, Kartiki V Desai, Jane S Thomsen, Yew Kok Lee, R Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin Sung, Edison T Liu, Chia-Lin Wei, Edwin Cheung, and Yijun Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, nov 2009. ISSN 1476-4687. doi: 10.1038/nature08497.
21. Maxwell R Mumbach, Adam J Rubin, Ryan A Flynn, Chao Dai, Paul A Khavari, William J Greenleaf, and Howard Y Chang. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods*, 13(11):919–922, nov 2016. doi: 10.1038/nmeth.3999.
22. Roland Jäger, Gabriele Migliorini, Marc Henrion, Radhika Kandaswamy, Helen E. Speedy, Andreas Heindl, Nicola Whiffin, Maria J. Carnicer, Laura Broome, Nicola Dryden, Takashi Nagano, Stefan Schoenfelder, Martin Enge, Yinyin Yuan, Jussi Taipale, Peter Fraser, Olivia Fletcher, and Richard S. Houlston. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature Communications*, 6(1):6178, may 2015. ISSN 2041-1723. doi: 10.1038/ncomms7178.
23. Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013. ISSN 00280836. doi: 10.1038/nature12593.
24. Longzhi Tan, Dong Xing, Chi-Han Chang, Heng Li, and X Sunney Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405):924–928, aug 2018. ISSN 0036-8075. doi: 10.1126/science.aat5641.
25. Derek M. Bickhart, Benjamin D. Rosen, Sergey Koren, Brian L. Sayre, Alex R. Hastie, Saki Chan, Joyce Lee, Ernest T. Lam, Ivan Liachko, Shawn T. Sullivan, Joshua N. Burton, Heather J. Huson, John C. Nystrom, Christy M. Kelley, Jana L. Hutchison, Yang Zhou, Jiajie Sun, Alessandra Crisà, F. Abel Ponce de León, John C. Schwartz, John A. Hammond, Geoffrey C. Waldbieser, Steven G. Schroeder, George E. Liu, Maitreya J. Dunham, Jay Shendure, Tad S. Sonstegard, Adam M. Phillippy, Curtis P. Van Tassell, and Timothy P. L. Smith. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, 49(4):643–650, April 2017. ISSN 1546-1718. doi: 10.1038/ng.3802.
26. Benjamin D Pope, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L Vera, Yanli Wang, R Scott Hansen, Theresa K Canfield, Robert E Thurman, Yong Cheng, Günhan Gülsoy, Jonathan H Dennis, Michael P Snyder, John A Stamatoyannopoulos, James Taylor, Ross C Hardison, Tamer Kahveci, Bing Ren, and David M Gilbert. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515 (7527):402–405, nov 2014. doi: 10.1038/nature13986.
27. Qian Du, Saul A Bert, Nicola J Armstrong, C Elizabeth Caldon, Jenny Z Song, Shalima S Nair, Cathryn M Gould, Phuc-Loi Luu, Timothy Peters, Amanda Khoury, Wenjia Qu, Elena

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Zotenko, Clare Stirzaker, and Susan J Clark. Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nature Communications*, 10(1):416, jan 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08302-1.

28. Yad Ghavi-Helm, Aleksander Jankowski, Sascha Meiers, Rebecca R. Viales, Jan O. Korbel, and Eileen E. M. Furlong. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature Genetics*, 51(8):1272–1282, August 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0462-3.
29. Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, apr 2019. ISSN 1471-0056. doi: 10.1038/s41576-018-0089-8.
30. Job Dekker and Tom Misteli. Long-Range Chromatin Interactions. *Cold Spring Harbor Perspectives in Biology*, 7(10):a019356, oct 2015. ISSN 1943-0264. doi: 10.1101/cshperspect.a019356.
31. Denes Hnisz, Krishna Shrinivas, Richard A Young, Arup K Chakraborty, and Phillip A Sharp. A phase separation model for transcriptional control. *Cell*, 169(1):13–23, March 2017.
32. Bryan A Gibson, Lynda K Doolittle, Maximillian W G Schneider, Liv E Jensen, Nathan Gamarra, Lisa Henry, Daniel W Gerlich, Sy Redding, and Michael K Rosen. Organization of chromatin by intrinsic and regulated phase separation. *Cell*, 179(2):470–484.e21, October 2019.
33. Emily M. Darrow, Miriam H. Huntley, Olga Dudchenko, Elena K. Stamenova, Neva C. Durand, Zhuo Sun, Su-Chen Huang, Adrian L. Sanborn, Ido Machol, Muhammad Shamim, Andrew P. Seberg, Eric S. Lander, Brian P. Chadwick, and Erez Lieberman Aiden. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences*, 113(31):E4504–E4512, August 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1609643113.
34. Pedro Olivares-Chauvet, Zohar Mukamel, Aviezer Lifshitz, Omer Schwartzman, Noa Oded Elkayam, Yaniv Lubling, Gintaras Deikus, Robert P Sebra, and Amos Tanay. Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. *Nature*, 540(7632):296–300, dec 2016. ISSN 0028-0836. doi: 10.1038/nature20158.
35. Robert A Beagrie, Antonio Scialdone, Markus Schueler, Dorothee C A Kraemer, Mita Chotalia, Sheila Q Xie, Mariano Barbieri, Inês de Santiago, Liron-Mark Lavitas, Miguel R Branco, James Fraser, Josée Dostie, Laurence Game, Niall Dillon, Paul A W Edwards, Mario Nicodemi, and Ana Pombo. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646):519–524, mar 2017. doi: 10.1038/nature21411.
36. Sofia A Quinodoz, Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M Lai, Alexander A Shishkin, Prashant Bhat, Yodai Takei, Vickie Trinh, Erik Aznauryan, Pamela Russell, Christine Cheng, Marko Jovanovic, Amy Chow, Long Cai, Patrick McDonel, Manuel Garber, and Mitchell Guttman. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*, 174(3):744–757.e24, jul 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.05.024.
37. Meizhen Zheng, Simon Zhongyuan Tian, Daniel Capurso, Minji Kim, Rahul Maurya, Byoungkoo Lee, Emaly Piecuch, Liang Gong, Jacqueline Jufen Zhu, Zhihui Li, Chee Hong Wong, Chew Yee Ngan, Ping Wang, Xiaoan Ruan, Chia-Lin Wei, and Yijun Ruan. Multiplex chromatin interactions with single-molecule precision. *Nature*, 566(7745):558–562, feb 2019. ISSN 0028-0836. doi: 10.1038/s41586-019-0949-1.
38. A. Marieke Oudelaar, James O. J. Davies, Lars L. P. Hanssen, Jelena M. Telenius, Ron Schwessinger, Yu Liu, Jill M. Brown, Damien J. Downes, Andrea M. Chiariello, Simona Bianco, Mario Nicodemi, Veronica J. Buckle, Job Dekker, Douglas R. Higgs, and Jim R. Hughes. Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. *Nature Genetics*, 50(12):1744–1751, 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0253-2.
39. Ferhat Ay, Thanh H. Vu, Michael J. Zeitz, Nelle Varoquaux, Jan E. Carette, Jean-Philippe Vert, Andrew R. Hoffman, and William S. Noble. Identifying multi-locus chromatin contacts in human cells using tethered multiple 3c. *BMC Genomics*, 16(1):121, Feb 2015. ISSN 1471-2164. doi: 10.1186/s12864-015-1236-7.
40. Amin Allahyar, Carlo Vermeulen, Britta A M Bouwman, Peter H L Krijger, Marjon J A M Verstegen, Geert Geeven, Melissa van Kranenburg, Mark Pieterse, Roy Straver, Judith H I Haarhuis, Kees Jalink, Hans Teunissen, Ivo J Renkens, Wigard P Kloosterman, Benjamin D Rowland, Elzo de Wit, Jeroen de Ridder, and Wouter de Laat. Enhancer hubs and loop collisions identified from single-allele topologies. *Nature Genetics*, 50(8):1151–1160, aug 2018. ISSN 1061-4036. doi: 10.1038/s41588-018-0161-5.
41. Siddarth Selvaraj, Jesse R Dixon, Vikas Bansal, and Bing Ren. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnology*, 31(12):1111–1118, dec 2013. doi: 10.1038/nbt.2728.
42. Zehra Nizami, Svetlana Deryusheva, and Joseph G Gall. The cajal body and histone locus body. *Cold Spring Harb. Perspect. Biol.*, 2(7):a000653, July 2010.
43. Mark Carty, Lee Zamparo, Merve Sahin, Alvaro González, Raphael Pelossof, Olivier Elemento, and Christina S. Leslie. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. *Nature Communications*, 8:15454, 2017. doi: 10.1038/ncomms15454.
44. Yongdae Shin and Clifford P. Brangwynne. Liquid phase condensation in cell physiology and disease. *Science*, 357:eaaf4382, 9 2017. ISSN 0036-8075. doi: 10.1126/science.aaf4382.
45. Kevin Hadi, Xiaotong Yao, Julie Behr, Aditya Deshpande, Charalampos Xanthopoulakis, Joel Rosiene, Madison Darmofal, Huasong Tian, Joseph DeRose, Rick Mortensen, Emily Adney, Zoran Gajic, Kenneth Eng, Jeremiah A. Wala, Kazimierz O. Wrzeszczynski, Kanika Arora, Minita Shah, Anne-Katrin Emde, Vanessa Felice, Mayu O. Frank, Robert B. Darnell, Mahmoud Ghandi, Franklin Huang, John Maciejowski, Titia De Lange, Jeremy Setton, Nadeem Riaz, Jorge-Reis Filho, Simon Powell, David Knowles, Ed Reznik, Bud Mishra, Rameen Beroukhim, Michael Zody, Nicolas Robine, Kenji M. Oman, Carissa A. Sanchez, Mary K. Kuhner, Lucian P. Smith, Patricia C. Galipeau, Thomas G. Paulson, Brian J. Reid, Xiaohong Li, David Wilkes, Andrea Sboner, Juan Miguel Mosquera, Olivier Elemento, and Marcin Imielinski. Novel patterns of complex DNA rearrangements revealed across thousands of cancer genome graphs. *bioRxiv*, 2019.
46. John Maciejowski and Marcin Imielinski. Modeling cancer rearrangement landscapes. *Current Opinion in Systems Biology*, 1:54 61, 2017. ISSN 2452-3100. doi: 10.1016/j.coisb.2016.12.005.
47. Jeremiah A Wala, Pratiti Bandopadhayay, Noah F Greenwald, Ryan O'Rourke, Ted Sharpe, Chip Stewart, Steve Schumacher, Yilong Li, Joachim Weischenfeldt, Xiaotong Yao, Chad Nusbaum, Peter Campbell, Gad Getz, Matthew Meyerson, Cheng-Zhong Zhang, Marcin Imielinski, and Rameen Beroukhim. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Research*, 28:581–591, 2018. ISSN 1088-9051. doi: 10.1101/gr.221028.117.
48. Jue Ruan and Heng Li. Fast and accurate long-read assembly with wtdbg2. *bioRxiv*, page 530972, jan 2019. doi: 10.1101/530972.
49. Jay Ghurye, Arang Rhie, Brian P. Walenz, Anthony Schmitt, Siddarth Selvaraj, Mihai Pop, Adam M. Phillippy, and Sergey Koren. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Computational Biology*, 15(8):e1007273, August 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007273.
50. Noam Kaplan and Job Dekker. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nature Biotechnology*, 31(12):1143–1147, December 2013. ISSN 1546-1696. doi: 10.1038/nbt.2768.
51. Joshua N. Burton, Andrew Adey, Rupali P. Patwardhan, Ruolan Qiu, Jacob O. Kitzman, and Jay Shendure. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, 31(12):1119–1125, December 2013. ISSN 1546-1696. doi: 10.1038/nbt.2727.
52. Jay Ghurye and Mihai Pop. Modern technologies and algorithms for scaffolding assembled genomes. *PLOS Computational Biology*, 15(6):e1006994, June 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006994.
53. Justin M. Zook, Jennifer McDaniel, Nathan D. Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A. Irvine, Len Trigg, Rebecca Truty, Cory Y. McLean, Francisco M. De La Vega, Chunlin Xiao, Stephen Sherry, and Marc Salit. An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, 37(5):561–566, May 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0074-6.
54. Elena K Stamenova, Neva Durand, Olga Dudchenko, Muhammad S Shamim, Su-Chen Huang, Yiqun Jiang, Ivan D Bochkov, Suhas S. P. Rao, Eric S Lander, Andreas Gnirke, and Erez Lieberman Aiden. The Hi-Culfite assay reveals relationships between chromatin contacts and DNA methylation state. *bioRxiv*, page 481283, nov 2018. doi: 10.1101/481283.
55. Guoqiang Li, Yaping Liu, Yanxiao Zhang, Naoki Kubo, Miao Yu, Rongxin Fang, Manolis Kellis, and Bing Ren. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nature Methods*, 16(10):991–993, oct 2019. doi: 10.1038/s41592-019-0502-z.
56. J. Koster and S. Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, oct 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts480.
57. Fidel Ramírez, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature communications*, 9(1):189, dec 2018. ISSN 2041-1723. doi: 10.1038/s41467-017-02525-w.
58. Nezar Abdennur and Leonid Mirny. Cooler: scalable storage for Hi-C data and other genomically-labeled arrays. *bioRxiv*, page 557660, feb 2019. doi: 10.1101/557660.
59. Philip A. Knight and Daniel Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33(3):1029–1047, 2013. ISSN 02724979. doi: 10.1093/imanum/drs019.
60. W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.
61. Robert Vaser, Ivan Sovic, Niranjan Nagarajan, and Mile Sikic. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, page gr.214270.116, January 2017. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.214270.116.

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Material

Glossary. *Virtual Digest*: A genome sequence informatically fragmented at every restriction site for an enzyme and represented as a set of non-overlapping intervals along the genome.
Virtual Contact Map: A contact map generated from multi-way contact data by counting the number of multi-way contacts supporting the co-observation of each member of every pair of contacts.
Restriction Fragment: A single interval produced by a restriction digest.
Pore-C read: An ONT read generated by sequencing the DNA from a Pore-C library.
Pore-C read alignment: Alignment of Pore-C reads against a genome sequence. Specialized alignment parameters must be be used due to the scrambled nature of the concatemer.
Fragment Assignment: The process by which intervals along a Pore-C read are mapped back to restriction fragments produced by the virtual digest. Can be thought of as a bioinformatic method of decoding what happened during the digestion and ligation steps of the *pore-C protocol*. Spurious alignments must be filtered out during this step.
Pore-C singleton: A Pore-C read where only a single restriction fragment can be assigned. This contains no proximity information and is thus filtered out.
Pore-C multimer: A Pore-C read where multiple restriction fragments can be assigned.
Pore-C pair: A pair of restriction fragments present in the same Pore-C multimer. The standard Pore-C analysis uses all possible pairs of restriction fragments.
Direct pair: a pair of restriction fragments that are adjacent on the read and is roughly equivalent to what a chimeric Hi-C read detects.
Indirect pair: a pair of restriction fragments that are not adjacent on the read (i.e. those with other restriction fragments between them).
Concatemer: A set of restriction fragments ligated together into a single molecule by the Pore-C protocol. Restriction fragments that tend to be close in 3D space will tend to be part of the same concatemer.
Monomer: A single restriction fragment in the context of a concatemer.
Contact order: The number of participating restriction fragments in a single multi-way contact. A single Pore-C run will generate a distribution of contact orders depending on restriction enzyme motif frequency and the read length distribution of the run itself. *Chimeric Junction*: A junction in a concatemer generated by the ligation of two restriction fragment ends that are non-adjacent in the genome. These junctions are used to infer 3D proximity.
Non-Chimeric Junction: A junction in a concatemer generated by the ligation of two restriction fragment ends that are adjacent in the genome. This can be the result of incomplete digestion, the re-ligation of cognate ends of a digestion event or possibly a SNP at the digestion site.
high order long range (HOLR): contacts defined as a multi-way contact containing at least 10 monomers in *cis* whose maximum span is greater than 10 megabases.

Pore-C pipeline. Alignment of raw sequencing data was conducted with bwa using parameters outlined in (**Table S2**). The resulting alignments were then filtered to a subset of alignments that maximally cover the read. Filtering was done using a directed acyclic graph (DAG) traversal formulation of the alignment coverage problem. For each read, a DAG is constituted containing one node for each alignment and two additional nodes which are named "START" and "END". Graph edges are added from START to each of the alignments such that the value of the edge is equal to the bwa bwasw gap penalty for a gap whose size is equal to the distance from the start of the read to the start of the alignment minus the alignment score. Additional edges are added from each alignment to the END node whose values are equal to the bwa bwasw gap penalty of a gap equal to the distance from the end of the alignment to the end of the read. Edges are then added between every pair of alignments A,B such that the edge source is the alignment whose endpoint is closer to the beginning of the read and the sink is the alignment whose endpoint is closer to the end of the read. The edge value is determined based on four relational classes: non-overlapping alignments, dovetailed alignments, contained/containing alignments, co-terminal alignments. First assume that the end of alignment A is closer to the beginning of the read than B. If alignments A and B do not overlap, then the edge value is equal to the gap penalty of the distance from the end of A to the start of B minus the alignment score of B. If A and B do overlap, then the edge value is equal to the gap penalty of an interval of size start of B to the end of A. If alignment A is contained within alignment B, then the edge score from A to B is computed similar to if A and B simply overlap. If a pair of alignments has the same endpoint, then the edge is drawn from the alignment whose start is closer to the start of the read to the second alignment in the pair and the value is equal to the alignment score of the shorter alignment plus a gap penalty the length of the second alignment. The graph is then traversed from the START node to the END node using bellman-ford algorithm to find the collection and order of alignments that has the optimal aggregate alignment score, and therefore the greatest non-redundant coverage of the sequencing read (**Fig S1A**). In addition to removing alignments that are not part of the optimal alignment tilepath, low quality alignments are removed, and additional filtering is done to remove singleton alignments, where a read has only a single alignment and is therefore unable to constitute even a single pairwise contact. For the purpose of creating a balanced contact matrix, we found it necessary to use a balancing

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

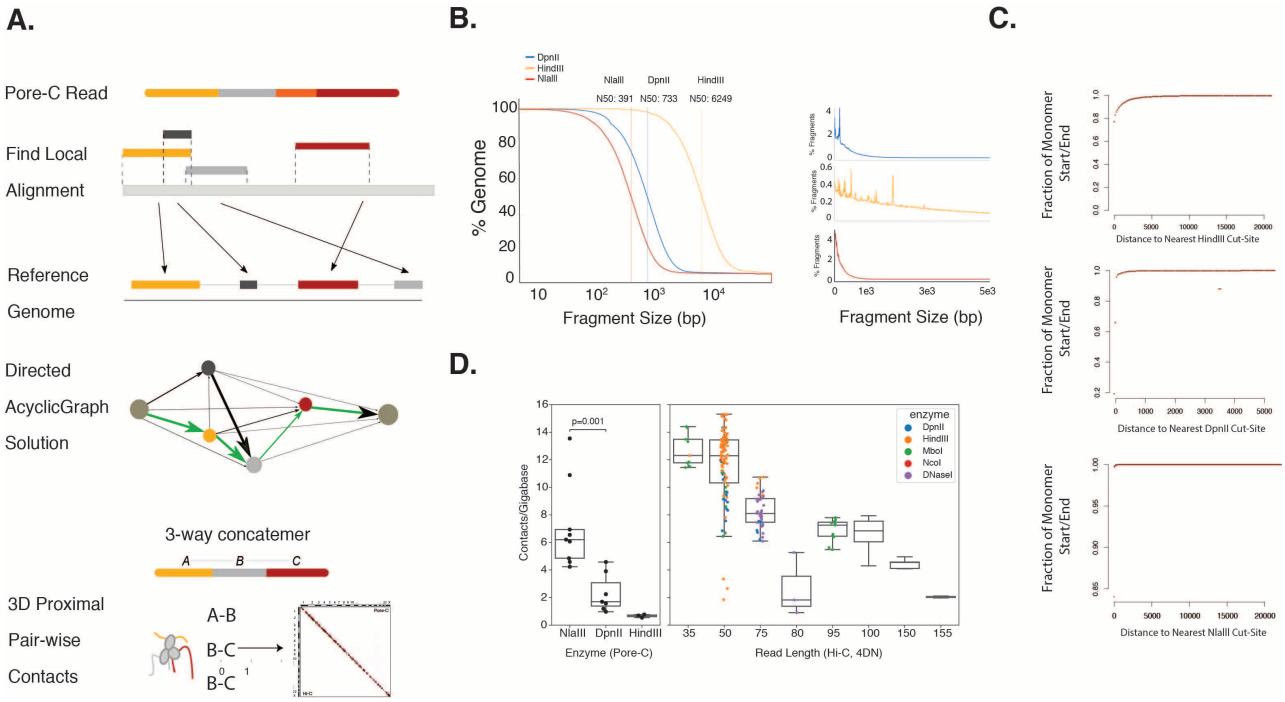

Fig. S1. Optimization of Pore-C Methodology (A) An overview of the directed acyclic graph filtering algorithm. A read is aligned to a reference genome. The DAG is constituted first with nodes representing the read start and end, and then with edges reflecting the combination of the alignment score as well as a gap or overlap penalty between the two involved reads or a given read and either the start or end of the read. Graph traversal from the start to end nodes describes a combination of alignments that maximally cover the read. (B) Comparison of the fragment length distributions generated by the three different restriction enzymes in the human reference GRCh38. (C) The fraction of reads falling a given distance from predicted restriction sites were calculated for each restriction enzyme. Although alignments generated with our workflow are not filtered based on whether they terminate close to a restriction site, alignment ends tend to fall very close to predicted sites. (D) By calculating the number of pairwise contacts generated per Gbp sequenced, head to head comparisons can be made with Hi-C, showing that Pore-C using NlaIII can output roughly as many contacts per gigabase sequenced as 100 bp paired-end Illumina Hi-C reads.

blacklist `https://rdrr.io/bioc/HiCcompare/man/hg38_blacklist.html`. While Pore-C read monomers can confidently align to regions that are generally inaccessible to Hi-C, these regions are often represented by model sequences that differ, by design, in copy number from true biological samples. This results in artificial pileups of alignments that prevent proper matrix balancing and downstream analysis.

Data quality assessment. Practically speaking, read N50 can be used as a simple indicator of a ligation efficiency quality metric for a given library. As a final metric we also calculate "pairwise contacts per Gbp sequenced", which enables users to determine the cost efficiency compared to Hi-C. When compared directly to Hi-C runs from 4DN, NlaIII runs generate a similar contact density to 100 bp paired-end Hi-C runs and approach the contact density of 35 bp paired-end Hi-C runs (**Fig S1B**). Finally, because our published pipeline does not consider the predicted locations of restriction sites in the reference genome as prior data for assignment of alignments to restriction fragments or for filtering, the robustness of our filtering protocol is affirmed by the fact that most of the ends of alignments are located very close to predicted restriction sites (**Fig S1C**).

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

		Reads		Pairwise Contacts		Contact Order Distribution						
		Gbases	*Reads*	*Count*	*Density*			*Bin (Read Count)*				
Aggregate			*(milliions)*	*(millions)*	*(millions/Gbase)*	≤1	*multiway*	*2*	*3*	*4-20*	*21-50*	*>50*
Enzyme												
	DpnII	187	93	432	2.3	46,545,362	46,199,666	18,641,971	10,778,078	16,645,174	133,742	701
	HindIII	49	13	29	0.6	7,173,116	5,597,723	2,655,637	1,375,372	1,564,896	1,792	26
	NlaIII	361	123	2,269	6.3	29,165,039	93,403,875	15,434,646	15,924,516	60,723,171	1,308,255	13,287
	Total	597	228	2,730	4.6	82,883,517	145,201,264	36,732,254	28,077,966	78,933,241	1,443,789	14,014
Sample												
GM12878		265	94	1,443	5.5	27,265,377	66,706,571	13,340,449	12,099,583	40,458,117	800,269	8,153
HCC1954		144	42	889	6.2	10,256,032	32,064,335	4,708,961	5,044,218	21,754,776	550,901	5,479
HG002		188	92	398	2.1	45,362,108	46,430,358	18,682,844	10,934,165	16,720,348	92,619	382
Sample, Enzyme												
GM12878	*DpnII*	60	27	143	2.4	11,625,774	15,765,757	5,236,897	3,705,505	6,808,451	14,793	111
GM12878	*HindIII*	6	2	5	0.7	678,219	882,571	385,260	226,915	270,106	287	3
GM12878	*NlaIII*	198	65	1,295	6.5	14,961,384	50,058,243	7,718,292	8,167,163	33,379,560	785,189	8,039
HCC1954	*DpnII*	46	12	211	4.5	3,535,106	8,078,647	1,346,427	1,290,667	5,323,074	118,012	467
HCC1954	*NlaIII*	98	31	678	6.9	6,720,926	23,985,688	3,362,534	3,753,551	16,431,702	432,889	5,012
HG002	*DpnII*	81	54	78	1.0	31,384,482	22,355,262	12,058,647	5,781,906	4,513,649	937	123
HG002	*HindIII*	43	11	24	0.6	6,494,897	4,715,152	2,270,377	1,148,457	1,294,790	1,505	23
HG002	*NlaIII*	65	27	296	4.6	7,482,729	19,359,944	4,353,820	4,003,802	10,911,909	90,177	236

Table S1. Pore-C sequencing runs in aggregate Summary data aggregated across all runs by either sample, or by sample and enzyme.

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

	Sequencing					Reads			Pairwise Contacts			Contact Order Distribution							
Sample				Run Time		Count	Yield		Count	Density		Median		Bin (% of reads)					
enzyme	*Replicate*	*Run*	*Washes*	*(hours)*	*Platform*	*(millions)*	*(Gbases)*	*N50*	*(millions)*	*(millions/Gbase)*	*% cis*	*all*	*>1*	*≤1*	*2*	*3*	*4-20*	*21-50*	*>50*
GM12878																			
DpnII	OF	e296f	0	12	G	0.8	1.0	1,773	1.7	1.7	42	1	3	51.1	22.9	13.1	12.9	0.0	0.0
	OG	a021f	0	12	G	0.8	1.0	1,589	1.5	1.6	47	1	2	54.4	22.9	12.0	10.6	0.0	0.0
	OH	100db	0	48	G	3.9	6.7	2,196	8.0	1.2	49	1	3	54.7	22.2	11.9	11.2	0.0	0.0
	SF	e2386	0	48	G	5.9	15.2	3,265	35.4	2.3	58	2	3	32.1	20.4	16.5	30.9	0.0	0.0
	SF	69689	2	92	P	12.7	24.7	2,943	53.0	2.1	56	2	3	47.6	18.5	12.6	21.3	0.0	0.0
	SG	83be7	0	72	G	3.2	11.2	4,540	43.8	3.9	51	4	4	20.7	13.3	14.3	51.3	0.3	0.0
HindIII	OJ	278f2	0	6	G	0.9	3.2	6,469	2.2	0.7	41	2	3	49.3	23.6	12.9	14.3	0.0	0.0
	OK	c2347	0	6	G	0.7	3.2	7,304	2.5	0.8	46	2	3	35.5	26.2	16.8	21.5	0.0	0.0
NlaIII	OA	ef1bb	0	8	G	1.0	3.7	5,124	22.8	6.2	39	4	5	22.5	9.3	11.2	55.4	1.6	0.0
	OB	cf93a	0	12	G	1.2	3.8	4,478	26.9	7.0	42	4	5	19.6	9.5	11.4	58.2	1.3	0.0
	OB	78d0d	2	48	P	10.3	33.4	4,659	204.4	6.1	43	4	5	28.5	9.7	10.4	50.3	1.1	0.0
	OB	4a72b	2	48	P	15.0	47.1	4,458	307.5	6.5	43	4	5	23.3	10.6	11.4	53.7	1.0	0.0
	OC	03044	0	48	G	0.9	1.7	2,610	7.4	4.2	39	2	4	33.9	16.9	14.5	34.5	0.1	0.0
	OE	5566f	0	70	G	2.5	13.2	7,873	143.8	10.9	38	6	7	13.8	6.1	7.6	65.5	6.9	0.1
	OL	ace1f	0	24	G	1.0	6.3	8,088	84.5	13.5	40	8	9	2.7	4.1	6.3	76.7	10.1	0.2
	OL	6b7fd	0	24	G	0.9	5.3	8,189	71.9	13.6	40	8	9	2.9	4.3	6.4	76.0	10.2	0.2
	SG	b4e70	0	72	G	6.0	19.7	4,022	119.8	6.1	38	5	5	14.9	8.5	11.4	64.6	0.6	0.0
	SH	cbc40	0	48	G	6.2	17.0	3,542	85.3	5.0	44	4	4	18.6	14.4	15.7	50.9	0.4	0.0
	SH	972fa	2	88	P	20.0	47.1	3,294	220.6	4.7	43	3	4	26.5	15.6	15.2	42.3	0.3	0.0
HCC1954																			
DpnII	SA	7d7ab	0	66	G	1.9	7.6	6,104	36.4	4.8	53	4	5	23.3	13.0	12.6	49.9	1.1	0.0
	SA	ae0b8	0	72	P	4.8	18.1	6,107	77.3	4.3	55	3	5	33.8	11.7	10.9	42.8	0.8	0.0
	SA	e4b64	1	19	P	5.0	20.7	6,533	97.3	4.7	55	3	5	29.9	11.0	10.7	47.2	1.1	0.0
NlaIII	SB	3c36a	0	96	G	5.9	18.6	4,428	128.8	6.9	33	4	5	19.5	11.4	12.8	54.9	1.4	0.0
	SB	6de06	1	91	P	24.8	79.1	4,584	549.0	6.9	34	4	5	22.4	10.8	12.1	53.2	1.4	0.0
HG002																			
DpnII	SC	efd7b	0	69	G	6.7	9.4	1,811	9.2	1.0	51	1	2	59.6	22.4	10.3	7.7	0.0	0.0
	SC	9faf9	0	52	P	47.0	71.9	1,955	68.7	1.0	53	1	2	58.2	22.4	10.8	8.5	0.0	0.0
HindIII	SD	3c3cd	0	69	G	3.8	14.1	6,543	6.3	0.4	64	1	2	60.3	21.2	9.7	8.8	0.0	0.0
	SD	dd128	0	40	P	7.4	28.4	6,850	18.0	0.6	68	1	3	56.7	19.8	10.5	12.9	0.0	0.0
NlaIII	SE	bdb7b	0	69	G	7.1	17.0	3,286	78.1	4.6	36	3	4	27.1	17.0	15.2	40.4	0.3	0.0
	SE	1c575	0	28	P	19.7	47.7	3,300	217.9	4.6	37	3	4	28.2	15.9	14.8	40.8	0.3	0.0

Table S2. Pore-C sequencing runs Per-run statistics and results for all Pore-C sequencing runs used in this publication.

tool	ver	link	parameters	purpose
pore_c refgenome virtual-digest	0.1	https://github.com/nanoporetech/pore-c/	default parameters	generates an in-silico restriction map for assignment of alignments to restriction fragments.
pore_c reads catalog	0.1	https://github.com/nanoporetech/pore-c/	default parameters	
pore_c alignments parse	0.1	https://github.com/nanoporetech/pore-c/	default parameters	conducts alignment filtering from the raw bam file using the fragDAG algorithm.
pore_c to-matrix	0.1	https://github.com/nanoporetech/pore-c/	default parameters	Computes pairwise contacts from multiway contacts in the filtered alignment table data and writes pairwise contacts to the standard pairs file format.
bwa	0.7.17	https://github.com/lh3/bwa/	`bwasw -b 5 -q 2 -r 1 -T 15 -z 10`	alignment of pore-c monomers to a genome
cooler load	0.8.6	https://github.com/mirnylab/cooler/	`--input-copy-status unique -f coo`	creates a 1kb binned cooler file from a pairs file created from a decomposed pore-C multiway contact data set
cooltools diamond-arrowhead	0.2.0	https://github.com/mirnylab/cooltools/	`3 5 10 15 --window-pixels --append-raw-scores`	calculates the insulation score vector used to identify topologically associated domains.
cooltools call-compartments	0.2.0	https://github.com/mirnylab/cooltools/	default parameters	calculates the balanced matrix eigenvectors which indicate the A and B compartments
cooler merge	0.8.6	https://github.com/mirnylab/cooler/	default parameters	
cooler balance	0.8.6	https://github.com/mirnylab/cooler/	`--blacklist blacklist.bed --convergence-policy store_nan --max-iters 10000`	balances the pairwise contact matrix
cooler zoomify	0.8.6	https://github.com/mirnylab/cooler/	default parameters	generates a balanced multi-resolution cooler from a 1kb binned matrix file.
wtdbg2	2.5	https://github.com/ruanjue/wtdbg2/	`--edge-min 4 --force --kmer-psize 19 --kmer-subsampling 4 --preset ont --rescue-low-cov-edges --tidy-reads 5000`	generates a genome assembly from ONT long read sequencing data
racon	1.4.7	https://github.com/lbcb-sci/racon/	`--error-threshold 0.3 --gap -8 --include-unpolished --match 5 --num_corrections 3 --quality-threshold -1 --window-length 1000`	assembly polishing
Medaka	0.9.2	https://github.com/nanoporetech/medaka/	`-b 100 -m r941_min_high`	assembly polishing
SALSA2	2.2	https://github.com/marbl/SALSA/	`--iter 10 --clean yes -p yes`	uses pore-C pairwise data to fix and scaffold the wtdbg2 assembly

Table S3. Commands and parameters used in this study. A table of all the computational tools used in this study to process Pore-C data. These parameters are the ones implemented in the snakemake workflow.

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

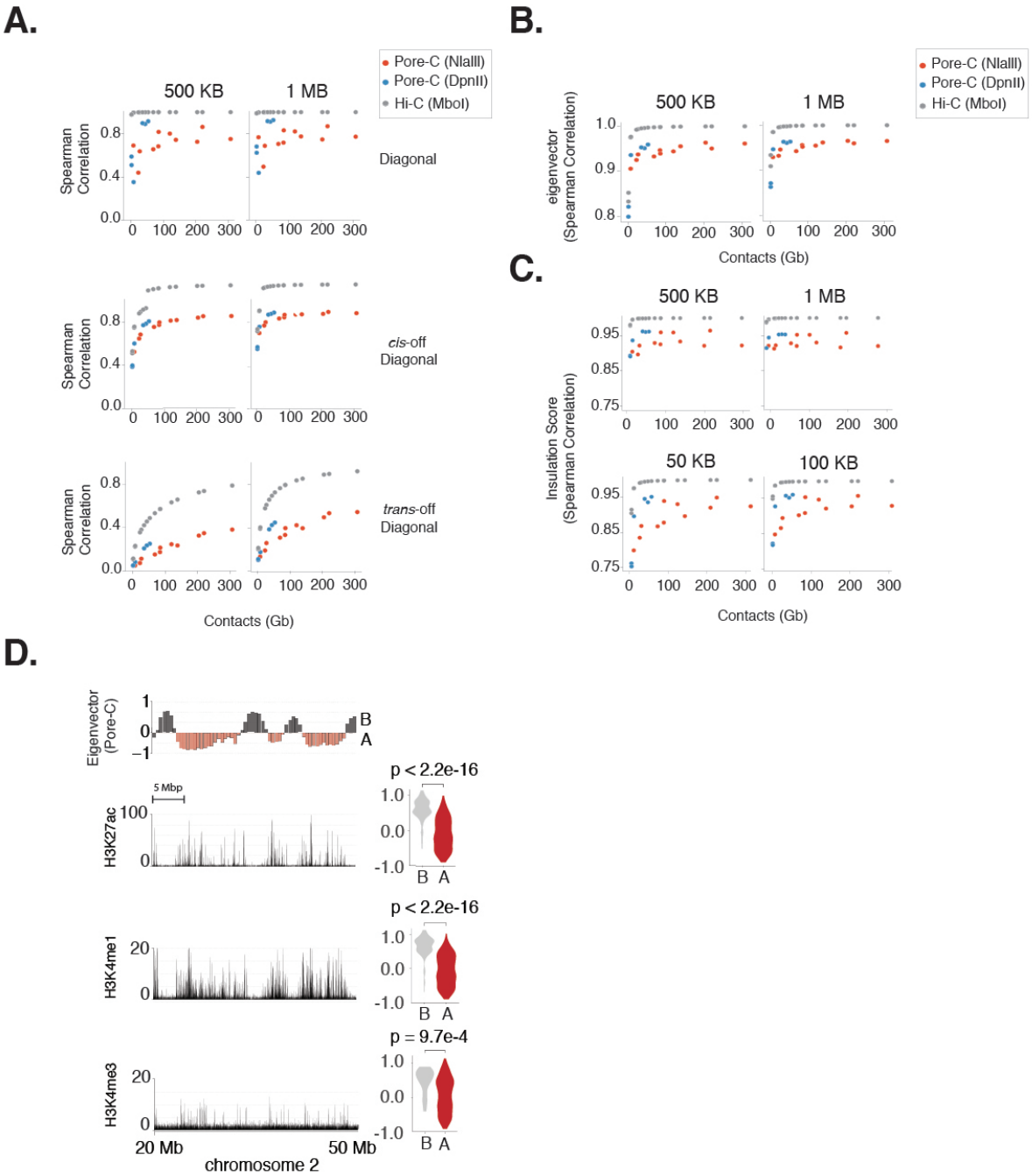

Fig. S2. Genomic Compartmentalization Similarity to the gold standard Hi-C dataset to individual Pore-C sequencing runs. Grey dots represent the upper bound of similarity of a Pore-C run of the indicated size, and is produced by downsampling the gold standard data to the same size as the indicated Pore-C run, and comparing it to the whole Hi-C dataset. Similarity is assessed using (A) Whole unbalanced matrix correlation along the diagonal (top), between all *cis* contacts not on the diagonal (middle), and all *trans* contacts (bottom), (B) using correlation of the compartmental eigenvector values of individual runs to the gold standard Hi-C data set, or (C) Correlation of the insulation scores of the individual runs with the gold standard Hi-C data set. (D) Correlation of the eigenvector compartment score with other indicators of active and inactive chromatin state along a 30 Mbp window in chromosome 2 binned at 500 Kbp resolution, including ChIP-seq peaks for the histone marks H3K27ac, H3K4me3 and H3K4me1. At right, violin plots indicate the correlation of the A and B compartments with the presence of ChIP-seq peaks.

bioRxiv preprint doi: https://doi.org/10.1101/833590; this version posted November 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Word Counts

This section is *not* included in the word count.

Notes on JPhysD article.

- Abstract: 150 words, unreferenced
- Main text: short format 4500 words, long is 8500 words.
- No limit in number of figures

Statistics on word count

```
File: 01_Article_MainText.tex
Sum count: 7661
Words in text: 6524
Words in headers: 108
Words outside text (captions, etc.): 1018
Number of headers: 26
Number of floats/tables/figures: 4
Number of math inlines: 11
Number of math displayed: 0
Subcounts:
  text+headers+captions (#headers/#floats/#inlines/#displayed)
  54+11+0 (1/0/0/0) _top_
  0+1+0 (1/0/0/0) Abstract
  189+1+0 (1/0/0/0) Section: Abstract
  15+0+0 (0/0/0/0) main
  591+1+611 (1/2/0/0) Section: Introduction
  0+1+0 (1/0/0/0) Section: Results
  311+7+0 (1/0/1/0) Subsection: Pore-C identifies higher-order contacts among DNA concatemers
  180+9+0 (1/0/3/0) Subsection: Pore-C virtual pairwise contact maps are concordant with Hi-C
  745+7+226 (1/1/6/0) Subsection: High-order long-range contacts reflect combinatorial chroma
  352+12+181 (1/1/0/0) Subsection: High-order chromatin contacts resolve the allelic structur
  298+9+0 (1/0/0/0) Subsection: Pore-C enables \textit{de novo} assembly of chromosomal lengt
  1070+1+0 (1/0/0/0) Section: Discussion
  0+1+0 (1/0/0/0) Section: Methods
  71+2+0 (1/0/0/0) Subsection: Cell culture
  127+1+0 (1/0/0/0) Subsection: Cross-linking
  254+3+0 (1/0/0/0) Subsection: Restriction enzyme digest
  122+5+0 (1/0/0/0) Subsection: Proximity ligation and reverse cross-linking
  137+5+0 (1/0/0/0) Subsection: Protein degradation and DNA purification
  107+2+0 (1/0/0/0) Subsection: Nanopore sequencing
  319+2+0 (1/0/0/0) Subsection: Pore-C pipeline
  303+4+0 (1/0/0/0) Subsection: Pore-C and Hi-C comparisons
  452+4+0 (1/0/1/0) Subsection: Analysis of high-order contacts
  169+7+0 (1/0/0/0) Subsection: Validation of rearrangements in HCC1954 using Pore-C
  232+5+0 (1/0/0/0) Subsection: Fluorescence \textit{in situ} hybridization (FISH)
  222+4+0 (1/0/0/0) Subsection: Pore-C hybrid genome assembly
  204+2+0 (1/0/0/0) Subsection: Software Availability
  0+1+0 (1/0/0/0) Section: Bibliography
```