Weill Cornell Medicine

Oxford NANOPORE Technologies

GENOMIC APPLICATIONS

Pore-C: genome-wide, multi-contact, chromosome conformation capture

Netha Ulahannan[1,2]*, Matthew Pendleton[3]*, Stefan Schwenk[4], Julie Behr[1,2], Aditya Deshpande[1,2], David Dai[3], Priyesh Rughani[3], Sarah Kudman[1], David Wilkes[1], David Stoddart[4], Daniel J. Turner[4], Sissel Juul[3], Eoghan Harrington[3,#], Marcin Imielinski[1,2,#]

[1]Weill Cornell Medicine, New York, NY, USA; [2]New York Genome Center, New York, NY, USA; [3]Oxford Nanopore Technologies Inc, New York, NY, USA, [4]Oxford Nanopore Technologies Ltd, Oxford, UK

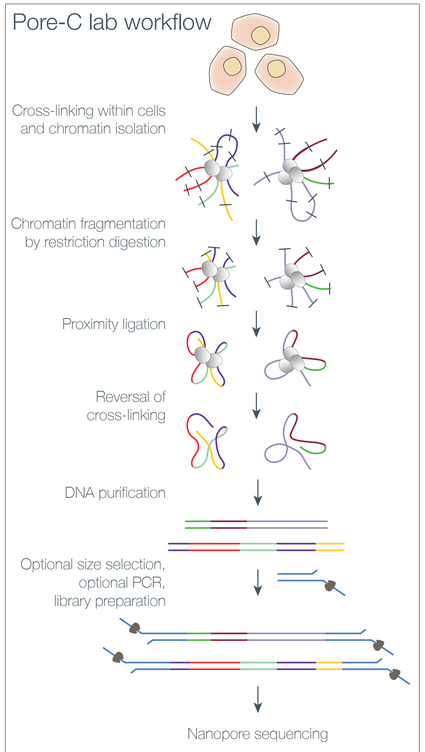
*These authors contributed equally. [#]Corresponding authors

Abstract

The interaction frequency of genomic loci that are in close spatial proximity can be used to predict nuclear organization. Chromatin structure relies on both multi-way interactions and on DNA methylation but it has previously not been possible to view these two phenomena simultaneously. Pore-C was developed to overcome these limitations. The technique uses cross-linking to preserve the spatial orientation of the genome followed by digestion and concatemerization of multiple neighboring segments into longer, chimeric fragments. The long-read capabilities of Nanopore technology allow these fragments to be sequenced in their entirety.

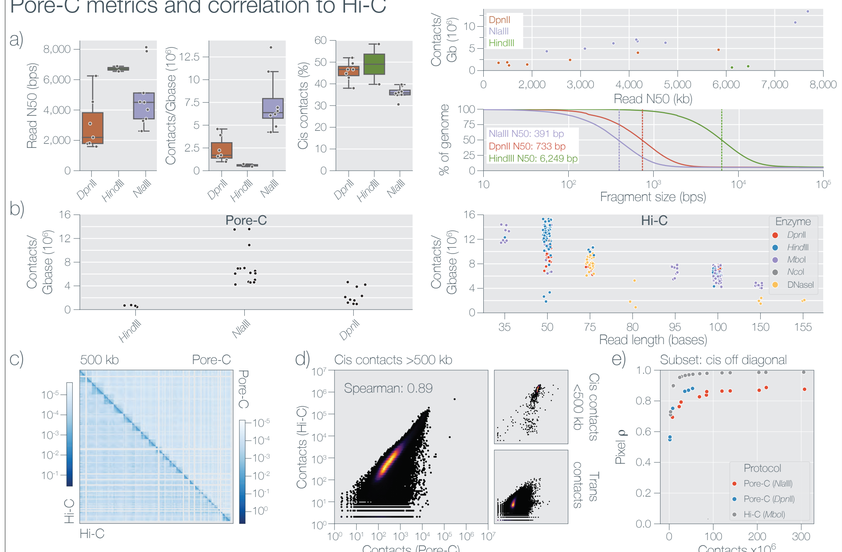
With Pore-C we are able to capture both these multi-way contacts and the methylation status of genomic loci. This information can be used to reveal the genomic compartmentalization previously identified using chromatin conformation capture assays. In addition, we show that Pore-C captures multi-contact hubs and displays broad A/B compartment coherence. We utilize the multi-way nature of Pore-C chromatin concatemers to identify structural rearrangements spanning multiple chromosomes. Furthermore, we demonstrate that even when decomposed into pairwise interactions, Pore-C can be used to assist with genome scaffolding.

Pore-C lab workflow



Genomic DNA is first cross-linked to cellular proteins using formaldehyde, to preserve the spatial proximity of interacting loci. Restriction digestion followed by proximity ligation is used to join cross-linked fragments. These fragments may be size-selected and amplified before library prep and sequencing.

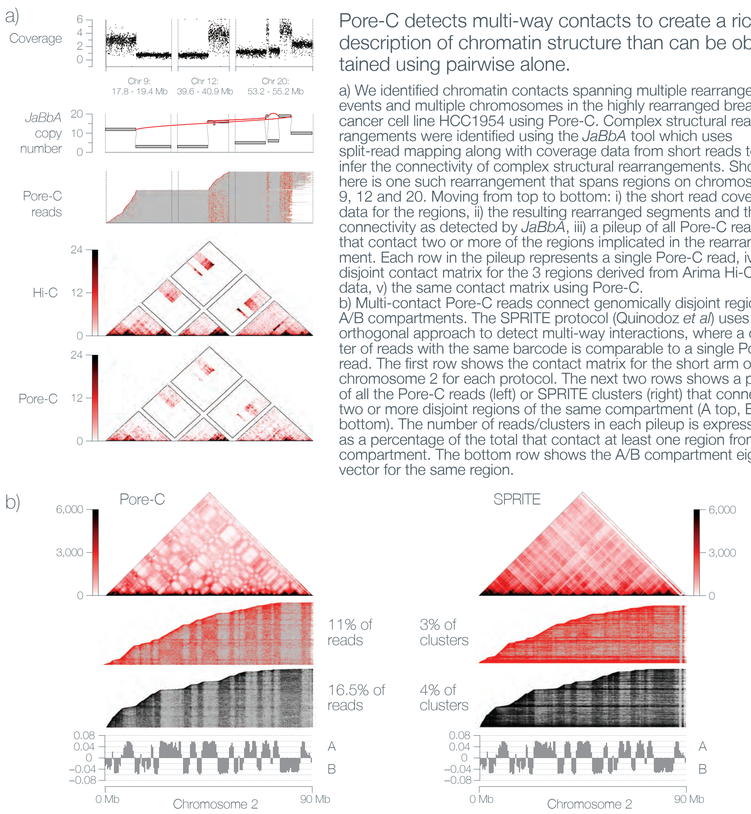
Pore-C metrics and correlation to Hi-C



Pore-C data shows equivalent efficiency and good correlation to Hi-C

a) The efficiency of the Pore-C protocol depends on the fragment sizes produced during restriction digestion and the length of the concatemers produced during re-ligation. We used three restriction enzymes with different fragment distributions: *Nla*III, *Dpn*II and *Hind*III with fragment N50s of 391, 733 and 6,249 bp respectively. We measure protocol efficiency by total number of contacts/gigabase of unfiltered sequence data. Shorter fragment lengths give more fragments per concatemer, so have higher efficiency, but may also have more incorrect alignments, hence the lower proportion of intramolecular (cis) contacts for *Nla*III. The characteristic linear relationship between read N50 and contacts/Gb for each enzyme can also be used to identify samples that have been digested incompletely. b) To compare Pore-C efficiency to that of Hi-C we used metrics from the 4DN data portal to derive contacts/Gb for every Hi-C experiment. A major determinant of Hi-C efficiency is the length of the paired-end read. Pore-C is within the range of the efficiencies of Hi-C. c) A genome-wide balanced contact matrix comparing the merged Pore-C *Nla*III (upper triangle, 1.3 billion contacts) to the Rao *et al. Mbo*I Hi-C dataset (lower triangle, 4 billion contacts, 4DN accession: 4DNFIXP4QG5B). d) A comparison of contact counts (aka pixel correlation) at 500 kb resolution between Pore-C *Nla*III (x-axis) and Rao *et al.* (y-axis) for cis off-diagonal contacts, cis on-diagonal contacts. e) Pixel correlation at 1 Mb resolution for each Pore-C replicate with the full Rao *et al.* dataset. The Hi-C point shows the correlation for a down-sampled version of the Rao *et al.* dataset, representing the maximum possible correlation for a dataset that size. Note that the variation along the x-axis is mostly due to a combination of the sequencing platform (PromethION vs. GridION) and the runtime of the experiment (not all samples were run to completion) and to a lesser extent the number of nuclease washes and protocol efficiency (contacts/Gb).

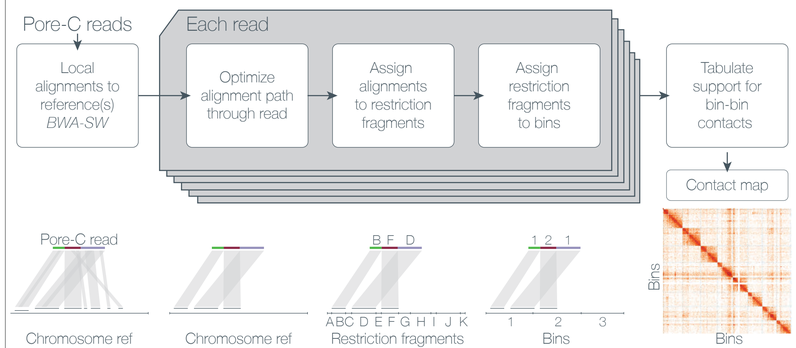
Multiway contacts: resolving complex SVs and comparison to SPRITE



Pore-C detects multi-way contacts to create a richer description of chromatin structure than can be obtained using pairwise alone.

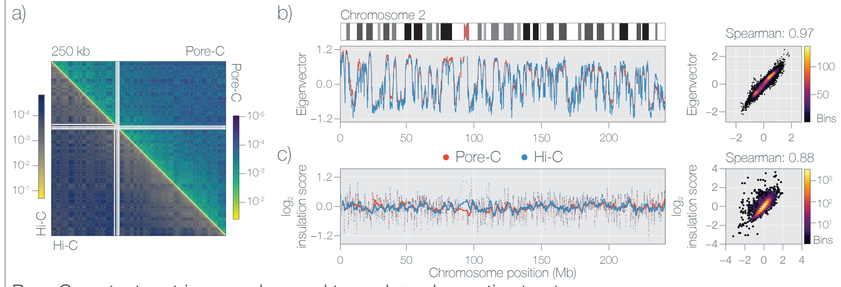
a) We identified chromatin contacts spanning multiple rearrangement events and multiple chromosomes in the highly rearranged breast cancer cell line HCC1954 using Pore-C. Complex structural rearrangements were identified using the *JaBbA* tool which uses split-read mapping along with coverage data from short reads to infer the connectivity of complex structural rearrangements. Shown here is one such rearrangement that spans regions on chromosomes 9, 12 and 20. Moving from top to bottom: i) the short read coverage data for the regions, ii) the resulting rearranged segments and their connectivity as detected by *JaBbA*, iii) a pileup of all Pore-C reads that contact two or more of the regions implicated in the rearrangement. Each row in the pileup represents a single Pore-C read, iv) a disjoint contact matrix for the 3 regions derived from Arima Hi-C data, v) the same contact matrix using Pore-C.
b) Multi-contact Pore-C reads connect genomically disjoint regions of A/B compartments. The SPRITE protocol (Quinodoz *et al*) uses an orthogonal approach to detect multi-way interactions, where a cluster of reads with the same barcode is comparable to a single Pore-C read. The first row shows the contact matrix for the short arm of chromosome 2 for each protocol. The next two rows shows a pileup of all the Pore-C reads (left) or SPRITE clusters (right) that connect two or more disjoint regions of the same compartment (A top, B bottom). The number of reads/clusters in each pileup is expressed as a percentage of the total that contact at least one region from that compartment. The bottom row shows the A/B compartment eigenvector for the same region.

Pore-C bioinformatics pipeline



The concatameric Pore-C reads are first aligned to a reference sequence using *BWA-SW* to identify the separate alignments. Each aligned read is filtered to retain only the minimal collection of alignments that traverse the majority of the read. Following optimisation of the alignment path, each segment of the read is assigned to a restriction fragment, determined through *in silico* digestion of the reference sequence. The reference genome is then divided into equally sized bins and restriction fragments are assigned to their corresponding bin. Finally, the total number of bin-to-bin contacts is calculated from all reads and visualized in a contact map.

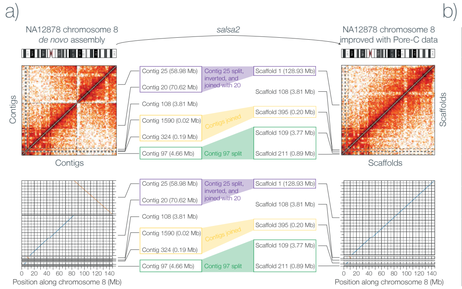
Chromatin structure



Pore-C contact matrices can be used to analyze chromatin structure

The contact matrix provides the information required for analysis of the hierarchical chromatin structure within the cell. We used *cooltools* to compare Pore-C and Hi-C for detection of A/B compartments and topologically associated domains (TADs). a) Balanced contact matrix for Pore-C *Nla*III (upper) and Rao *et al.* Hi-C (lower) for chromosome 2. b) The principal eigenvector of the correlation matrix for a chromosome can be used to divide a chromosome into A (gene dense) and B (gene poor) compartments. i) A plot of the chromosome 2 eigenvector for Hi-C (blue) and Pore-C (red) and ii) the genome-wide correlation of eigenvector values between Pore-C (x-axis) and Rao *et al.* Hi-C (y-axis). c) The diamond insulation score is a metric used to detect potential TAD boundaries in the contact matrix. i) A plot of the diamond insulation score calculated at 50 kb resolution for Hi-C (blue) and Pore-C (red) points show the individual scores per bin while lines show a rolling average of 50 points and ii) the genome-wide correlation of the insulation scores between Pore-C (x-axis) and Rao *et al.* (y-axis).

Improving assembly



b)

	Total data	Coverage	Read N50	
One PromethION flow cell (sheared gDNA)	130 Gb	43x	22 kb	
	Total contigs	Assembly size	Contigs N50	Longest contig
De novo assembly using *wtdbg2*	2,374	2.75 Gb	13.3 Mb	81.6 Mb
	Total data	Read N50	Total contacts	
Pore-C data using *Hind*III One PromethION flow cell	28 Gb	6.9 kb	17 million	
	Total scaffolds	Assembly size	Scaffold N50	Longest scaffold
Scaffolding and correcting using *salsa2*	1,760	2.75 Gb	103 Mb	216 Mb

Pore-C reads from NA12878 were mapped against the contigs produced by de novo whole genome assembly of the same sample (N50 = 36.1 Mb). The *Salsa2* tool uses the resulting contact density map to split, re-orient and join contigs into scaffolds that are consistent with the contact data. Panel a) shows the result for chromosome 8, where we obtained a scaffold spanning ~90% of the chromosome. However, the relative improvement in contiguity provided by Pore-C data is most substantial when the initial assembly is less contiguous. To illustrate this, we combined ~9x Pore-C data with contigs made by sequencing and assembling HG002 gDNA, and achieved a ~8-fold increase in contiguity (panel b).

References

Rao *et al.* Cell 159 (p. 1665-1680) 2014

Quinodoz *et al.* Cell 174 (p. 744–757) 2018