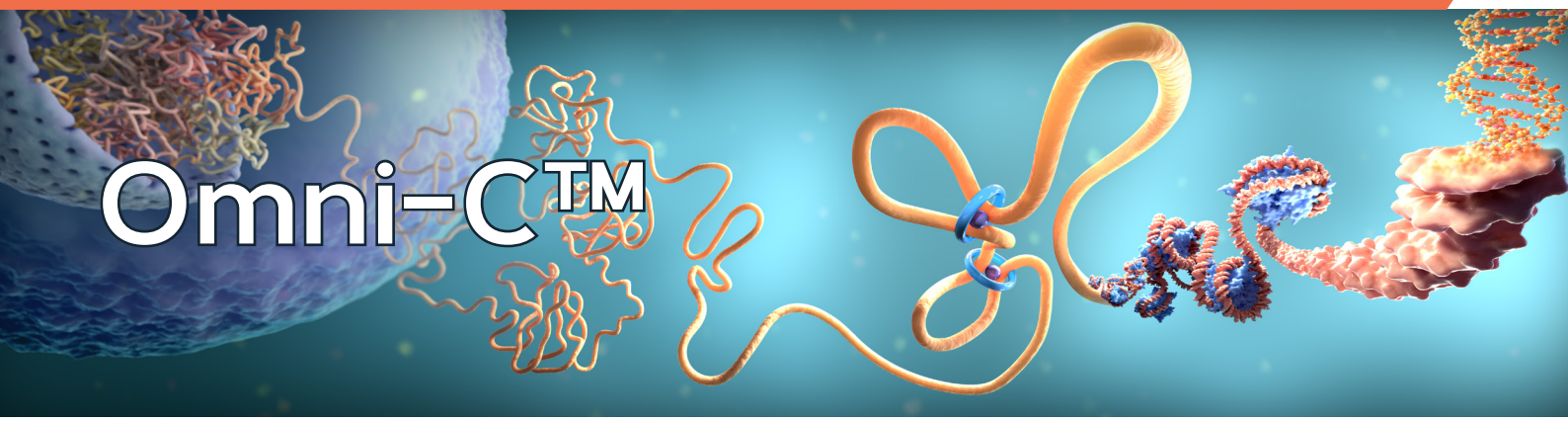
Dovetail GENOMICS

Technical Note



Omni-C™

Introduction

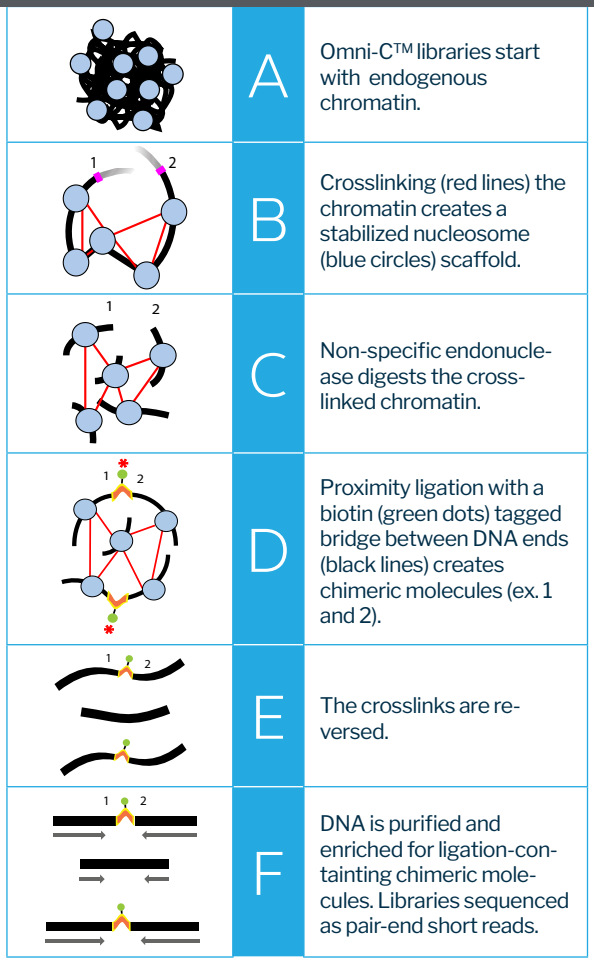
The discovery and adoption of chromatin capture methods have greatly accelerated the study of genome topology and enhanced the quality of genome assemblies. Genome conformation mapping technologies, including Dovetail™ Hi-C, have facilitated an unprecedented view of the three-dimensional (3-D) organization of the genome through sequencing technology. Methods like Hi-C have enabled the study of topological features such as topologically associated domains (TADs) and chromatin looping that are important in gene regulation and epigenetics. Moreover, the 3-D genomic structure can be leveraged for genome assembly by informing the scaffolding of contigs at chromosome scale. While Hi-C has enabled researchers in these areas, there are still notable drawbacks to this method due to the use of restriction enzymes (RE) to digest chromatin. For example, approximately 20% of the mappable human genome is blind to Hi-C due to low RE site density and analyses of Hi-C data are dependent on capturing these RE sites.

Here we introduce Omni-C™, a sequence-independent endonuclease-based Dovetail™ proximity-ligation protocol, which aims to address the limitations of RE-based Hi-C approaches **(Figure 1)**. By employing an endonuclease, Omni-C™ increases the genomic coverage of a proximity-ligation assay, therefore expanding the efficiency of each sequencing run by covering more of the genome and reducing biases imposed by RE site density. As such, Omni-C™ generates libraries where more of the genome is included in analyses and thereby making the data more versatile and unbiased by RE sites.

Product

Omni-C™ is an 8-reaction kit to generate sequence-independent endonuclease-based Hi-C libraries that provides more uniform coverage across the genome. The Omni-C™ assay is a 2-day workflow: Day 1 - Sample Prep & Proximity Ligation; Day 2 - Library Generation resulting in an Illumina ready sequencing library **(Figure 2).** To ensure reaction efficiency,

Figure 1 – Endonuclease-based molecular biology diagram.



The Omni-C™ process start with endogenous chromatin, which is fixed in place (cross-linking) with formaldehyde. After cross-linking, an in-situ chromatin digestion is achieved with an endonuclease. Digested chromatin and release from the cell generating a lysate. The resulting digestion lysate is subject to end-polishing, followed by a ligation of a 28-mer biotinylated oligonucleotide bridge. Then intra-aggregate ligation is carried out, followed by cross link reversal and DNA purification. Libraries are then generated from the purified DNA, with a streptavidin enrichment step on the biotinylated bridge. The result is an Illumina ready Omni-C™ library.

Omni-C™ has three molecular biology-based quality control checks that are predictive of library quality **(Figure 3)**. Omni-C™ is currently validated on mammalian cells and tissues sample inputs, with work in progress to confirm a broad range of sample types **(Figure 4)**. Accompanying the Omni-C™ assay is an open-source library QC tool to assess the library quality. The Omni-C™ workflow generates Hi-C-libraries with uniform coverage that can be easily integrated with open-source analysis tools.

Figure 2 – Omni-C™ 2-day workflow

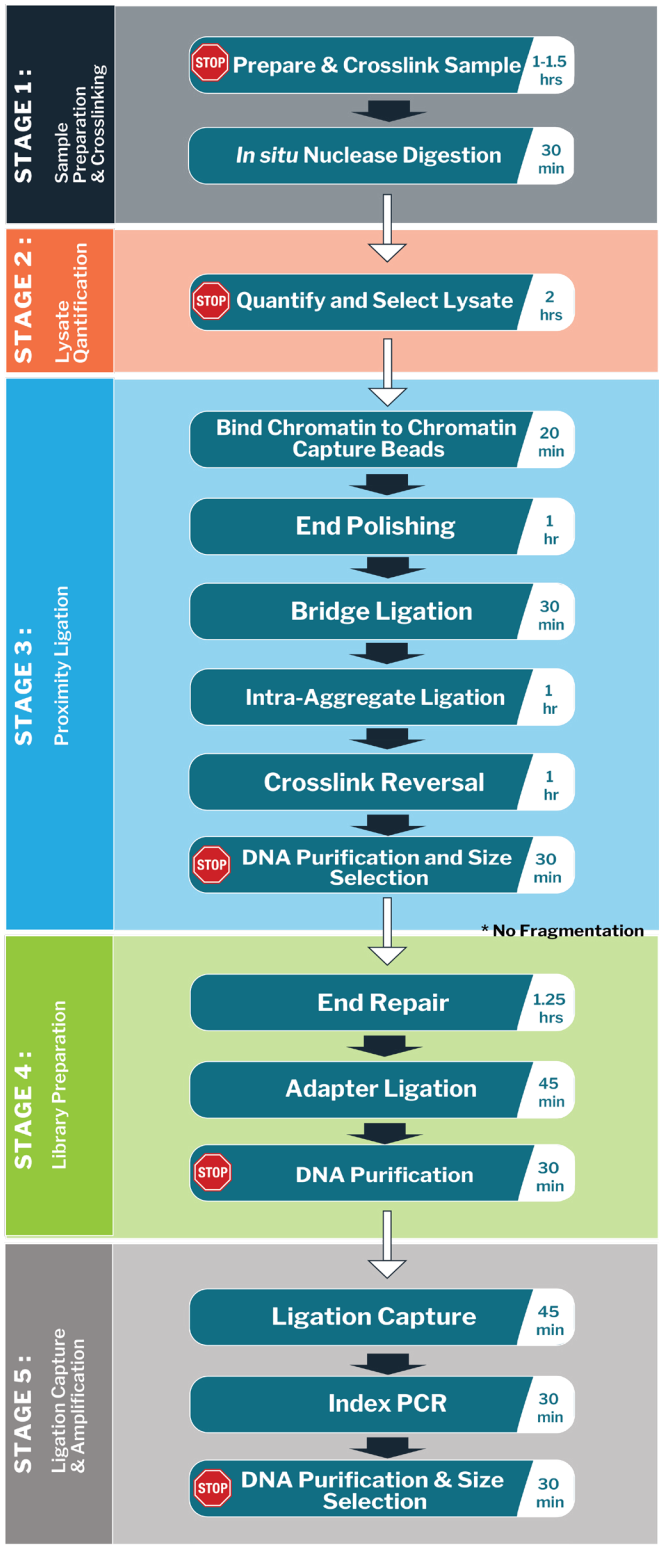
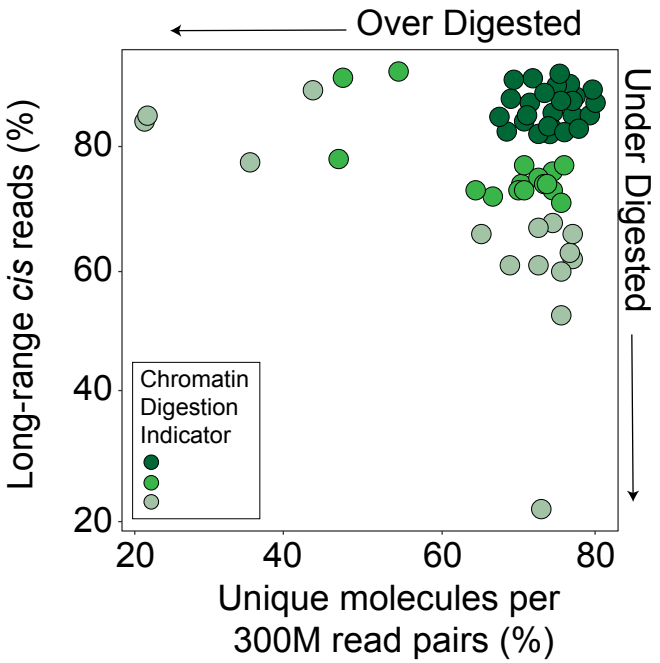


Figure 3 – CDI is predictive of library quality



Omni-C™ libraries were generated across various sample inputs (n=51). The CDI is informative about both library complexity, as the number of complexity is plotted as percent of unique molecules projected per 300 million read pairs, and proportion of long-range cis reads. Samples that are over digested during sample prep produce libraries that are of low complexity while maintaining a high proportion of long-range information, while samples that are under digested are still complex but are deficient in long-range information.

Data Highlights

Quality

The two-day Omni-C™ workflow consistently produces endonuclease-based Hi-C libraries that exhibit high complexity and enrichment of long-range cis reads. The built-in QC steps allow for users to determine library quality before sequencing. The Chromatin Digestion Index (CDI) quantitatively predicts the complexity (in unique molecules per 300 million read pairs sequenced) and the expected proportion of long-range cis reads in for each reaction **(Figure 3)**. Regardless of sample type Omni-C™ generated libraries with the high complexity and long-range information **(Figure 4)**.

Figure 4 – Validation of sample input types

Cells and tissues from human and mice were used as inputs to validate Omni-C™. All libraries were sequenced between 20-40 million 2x150bp read pairs and processed through the Dovetail Genomics Omni-C™ QC pipeline.

A) Long-range cis read pairs are plotted as a percent of total cis reads in the library and B) complexity is plotted as percent of unique molecules projected per 300 million read pairs.

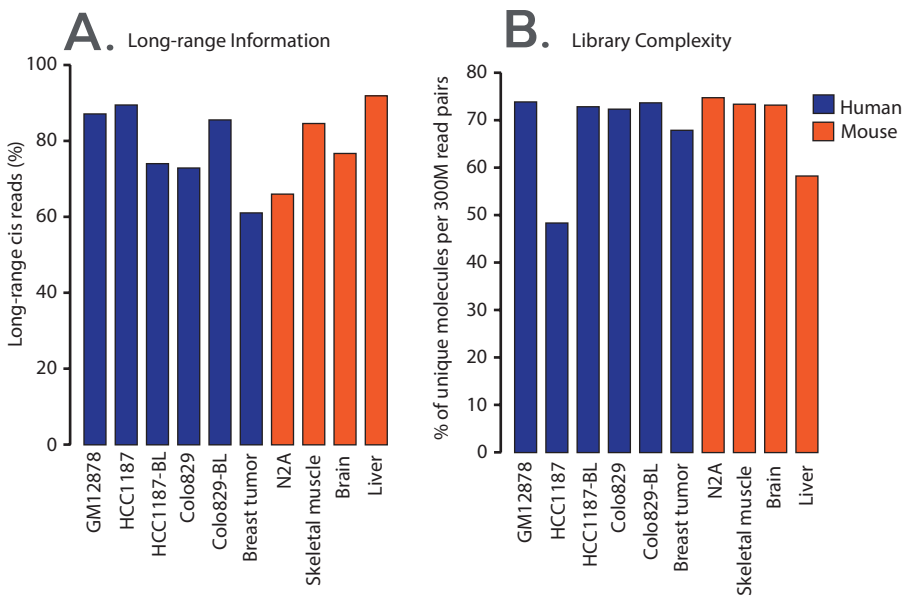


Table 1 – Low input

Omni-C™ libraries were generated from cell and muscle inputs from both human and mice. The resulting libraries were sequenced to 20-40 million read pairs (2x150 bp) and assessed on the Omni-C™ QC pipeline.

Sample	Type	Species	Amount	% Long-range *cis*	% Unique molecules per 300M read pairs
GM12878	Cell	Human	1M	87.1%	82.0%
GM12878	Cell	Human	500K	84.6%	80.7%
GM12878	Cell	Human	250K	84.8%	78.7%
GM12878	Cell	Human	100K	87.1%	81.3%
Breast Tumor	Tissue	Human	10mg	76.7%	73.0%
Skeletal Muscle	Tissue	Mouse	10mg	91.8%	58.3%
Liver	Tissue	Mouse	10mg	82.5%	76.0%
Brain	Tissue	Mouse	10mg	71.8%	69.3%
Brain	Tissue	Mouse	5mg	71.0%	77.3%

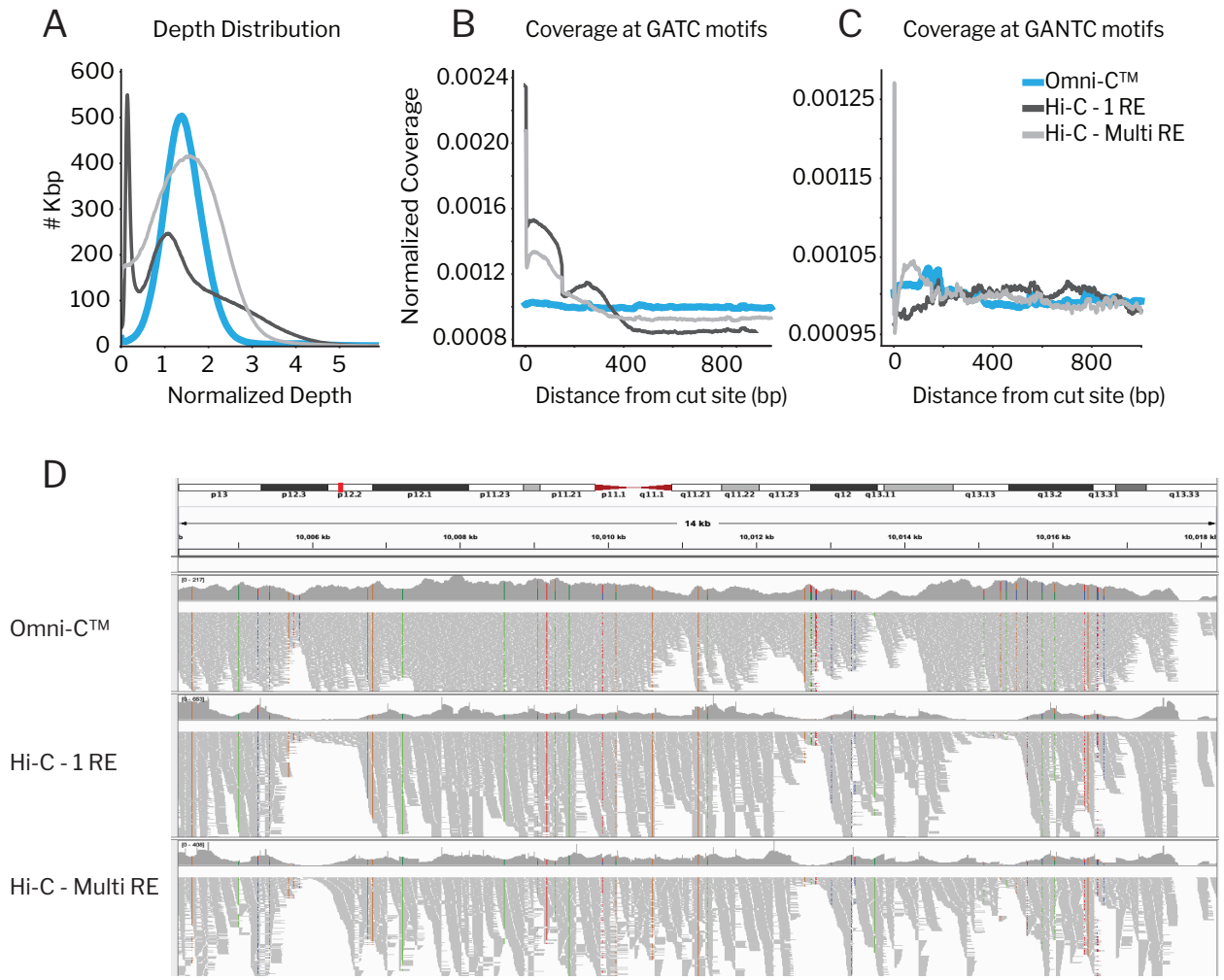
Flexibility

Omni-C™ is also designed to be a flexible assay. The workflow can work with a wide range of starting amounts (number of cells or tissue mass). The normal workflow calls for 1 million cells or 50 mg of tissue, the low input protocol uses starting material as low as 100K cells and 5 mg of tissue. Omni-C™ still yields high complexity libraries at lower starting inputs **(Table 1)**. Omni-C™ is also a dynamic assay by enabling researchers to also perform target enrichments such as hybrid capture, thereby reducing sequence burden and increasing resolution around sites of interest.

Coverage

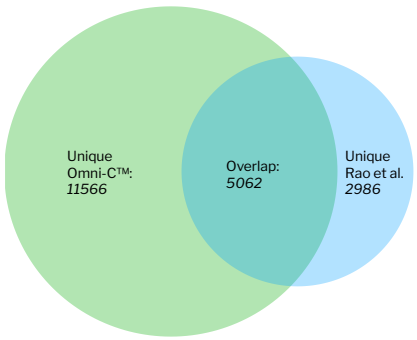
The endonuclease-based Omni-C™ provides superior coverage across the genome **(Table 2)**. Omni-C™ data exhibit a narrow per base coverage histogram when compared to other RE-based Hi-C suggesting the entire genome is being sequenced at the same frequency. Typical Hi-C approaches miss a significant portion of the genome, leading to wider histograms and significant portions of the genome with no coverage at all **(Figure 5a)**. This uneven coverage is also reflected in the pile-up of reads at RE sites, whereas Omni-C™ libraries show no enrichment of reads at RE sites **(Figure 5b)**. As such, Omni-C™ data capture single-nucleotide information in a manner that is independent of RE site proximity. When viewing Omni-C™ data in IGV, it is clear where RE-based Hi-C falls short in coverage and misses SNPs **(Figure 5c)**. The improved coverage that is inherent to Omni-C™ enables a more holistic view of the genome in down-stream analyses, which could include SNP calling and phasing.

Figure 5 – Coverage analysis



Deeply sequenced Omni-C™ (1 billion read-pairs) libraries were compared to RE-based Hi-C libraries for coverage. ***(A)*** *Per base coverage, in Kbps. Coverage at RE sites, GATC* ***(B)*** *and GANTC* ***(C)*** *are plotted as the average of the absolute value both upstream and downstream of RE sites.* ***(D)*** *IVG view of coverage across a 14 Kbp window. Colored vertical lines indicate single nucleotide polymorphisms.*

Figure 6 – Loop detection



Omni-C™ libraries from cell line GM12878 was sequenced to 1.77 billion read pairs and loops were called using HiCCUPs. The resulting loops were then compared to loops found in Rao et al., 2014 (4.9 Billion read pairs).

Table 2 – Contact map resolution at a fixed sequencing depth

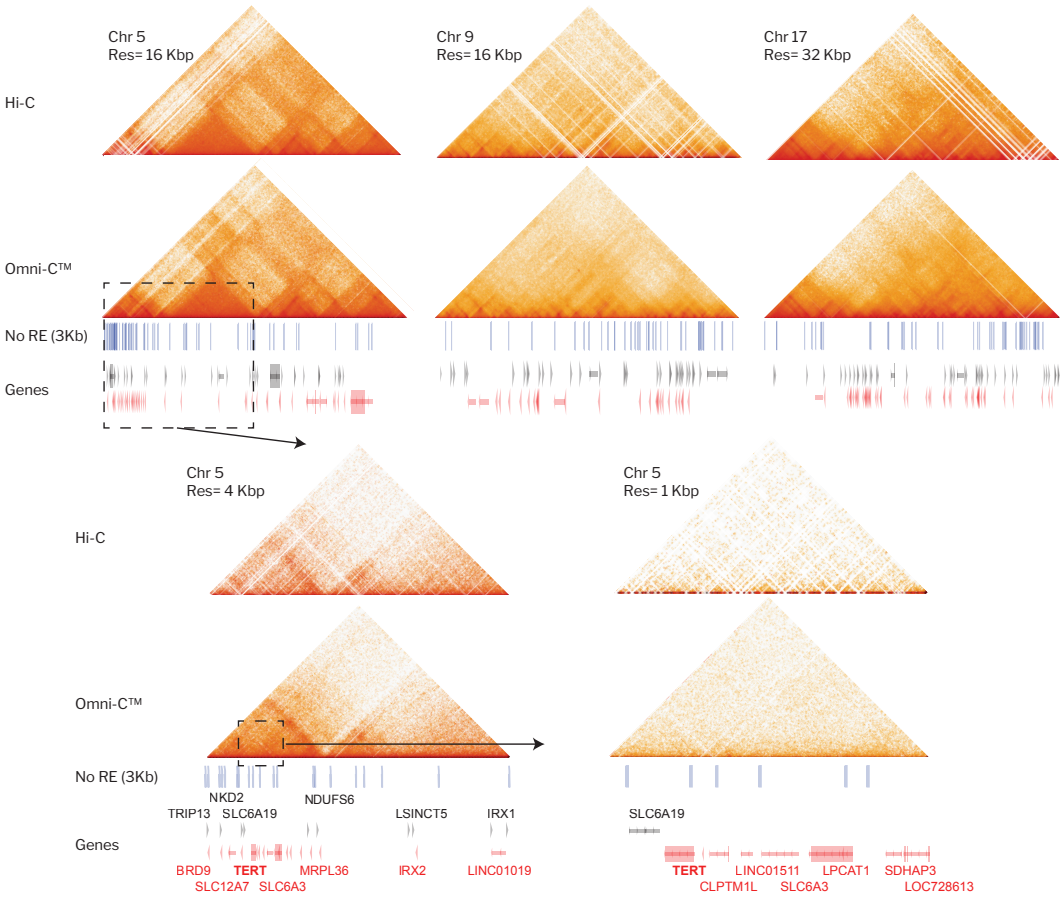
Assay	Resolution	Sequencing Depth (Read Pairs)	Genome Covered (%)
Omni-C™	5 kbp	1 Billion	93.23%
Hi-C	5 kbp	1 Billion	86.30%

Topology

In addition to uniform coverage, Omni-C™ delivers on conformation. Loop calling with Omni-C™ data from GM12878 via HiCCUPS detected 14,552 chromatin loops with 5062 overlapping with Rao et al., 2014, with 3-fold less sequencing **(Figure 6)**. Overlapping loop calls between Omni-C™ and Rao et al., are similar in number to other such comparisons.

The contact matrices generated from Omni-C™ libraries present a more complete view of genome conformation. During contact matrix balancing, areas with low coverage are essentially normalized with zero value in the denominator resulting in contact maps with blank vectors at these low coverage sites. Here we highlight three regions where Omni-C™ reveals topological features interrupted in RE-based Hi-C **(Figure 7)**. TERT, a gene that is vital in telomere maintenance and is often overexpressed in many lung cancers, presents a challenge for RE-based Hi-C contact matrices. The contact matrix generated by Omni-C™ captures a region of chromosome 5, which is known to be a cancer susceptibility locus. The second example (7b) demonstrates a ~5Mbp section on chromosome 9 that lacks sufficient Hi-C coverage to produce an uninterrupted contact map. Omni-C™'s uniformity generates a much more complete contact matrix at this site, which contains genes associated with apoptosis signaling. The last example is chromosome 17. Again, Omni-C™ produces a more comprehensive contact matrix where RE-based Hi-C falls short. Here topology around a potential oncogene, FASN, can now be seen in the Omni-C™ data. FASN is often overexpressed in breast cancers and understanding the looping and conformation impacting FASN, could lead to a better understanding of why this gene is overexpressed in breast cancer.

Figure 7 – Contact matrices from Omni-C™ libraries generate more complete contact matrices

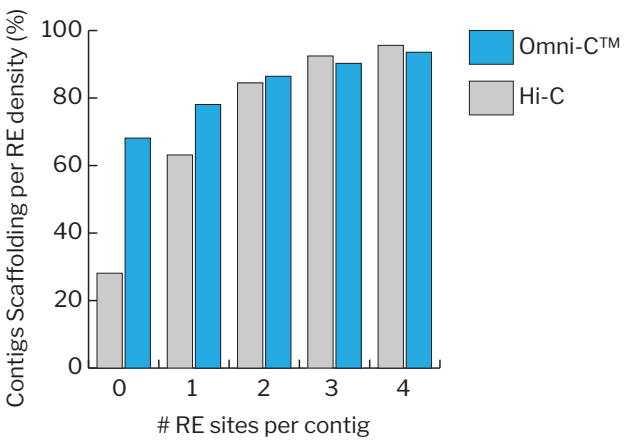


Blank bands in the contact matrix occur during contact matrix balancing in regions were coverage is too low. Low coverage regions cause the contacts to be normalized with a zero value in the denominator which results in blank vectors across these low-coverage regions. The examples are Chr5, a cancer susceptibility locus, which is often over expressed in lung cancer. The second is Chr9 which displays a contact matrix with ~5Mbp region of poor mapping that encompasses the TRAF2 gene that plays a key role apoptotic signaling. The third example is Chr17, containing a suspected oncogene, FASN, which is often over expressed in breast cancer. Blow each comparison are blue boxes denoting 3 Kbp regions that are devoid of RE sites, and gene tracks in black and red arrows. Under these comparisons are zoomed in sections of chr5 that encompasses TERT at 4 Kbp and 1 Kbp resolutions.

Scaffolding

A staple application of proximity ligation data is scaffolding contigs for genome assembly. The ability of Hi-C data to scaffold correctly depends on the RE site density captured within each assembled contig. The analysis on scaffolding a human genome shows RE-dependent Hi-C scaffolding misses contigs that Omni-C™ can include **(Figure 8)**. As Omni-C™ is RE agnostic, it can scaffold contigs more efficiently than Hi-C data, where RE frequency per contig is low. As RE frequency increase RE-based Hi-C and Omni-C™ scaffold at similar rates, which expected as RE-based Hi-C is an efficient means of scaffolding high-quality input assemblies.

Figure 8 – Omni-C™ is more efficient at scaffolding contigs with low RE site density



The human genome, HG38, was cut into contigs of random size. Libraries were made from GM12878 on both Omni-C™ and RE-based Hi-C. The contigs were then scaffolded using HiRise™. The scaffolded contigs were binned into groups by the number of RE sites per contig. Scaffolding efficiency was determined by normalizing the number of contigs scaffolded by the total number of contigs in each RE site group.

Summary

Here we presented data that showcases Omni-C™, an endonuclease-based Hi-C kit, from Dovetail Genomics. The robust protocol provides quality control steps that are predictive of library features before sequencing. The workflow of Omni-C™ does not drastically alter the typical RE-based Hi-C workflow and can incorporate low input samples. Omni-C™ offers uniform coverage across the genome without over representing RE sites. This demonstrated uniformity of coverage provides a more complete view of the genome through proximity ligation.