Dovetail GENOMICS

From Domains to SNPs: A Multi-Resolution View of the Genome

Application Note

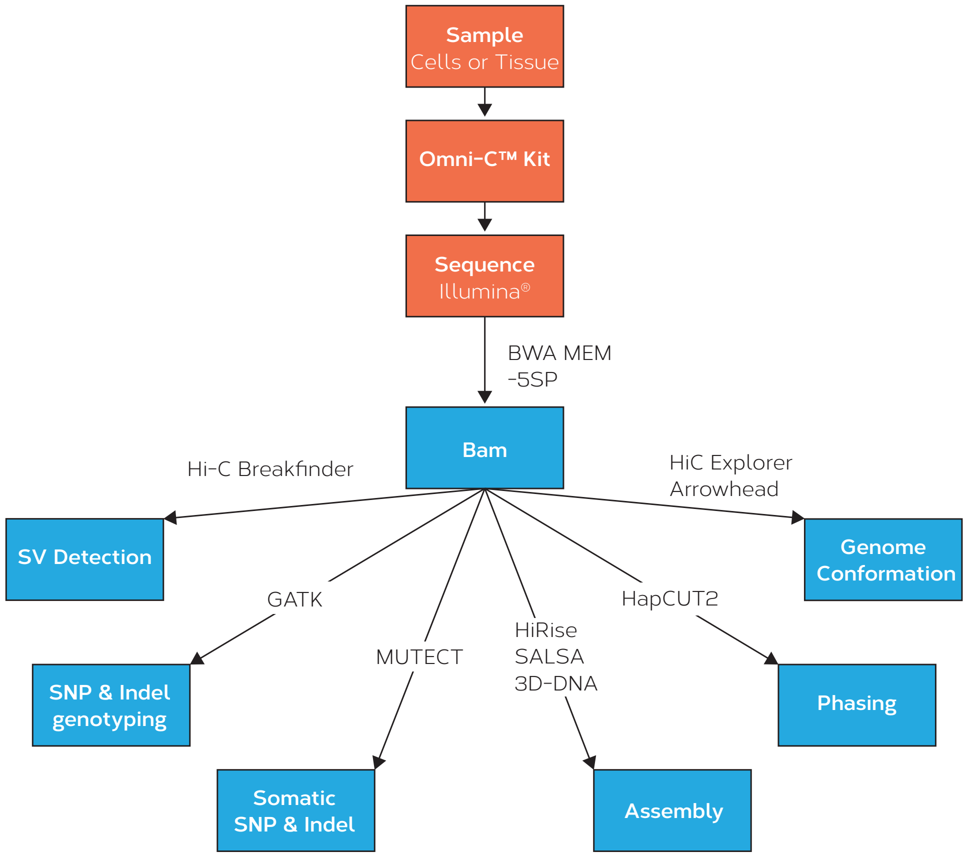
1. Introduction

The discovery and adoption of chromatin capture methods has greatly accelerated the study of genome topology and enhanced the quality of genome assemblies through the creation of long-range linkage information. Commonly, Hi-C is performed using restriction enzymes (RE) to digest chromatin prior to a proximity-based ligation step; however, there are notable drawbacks to this approach. Importantly, RE-based Hi-C is blind to approximately 20% of the mappable human genome due to low RE site density. Here we introduce Omni-C™ technology, a sequence-independent endonuclease-based proximity-ligation assay which addresses the limitations of RE-based Hi-C approaches. Herein we highlight that Omni-C™ libraries uniquely contain proximity-ligation characteristics with the advantage of uniform genome-wide coverage. This expanded data type enables not only chromatin conformation studies, but also applications relying on coverage uniformity, such as SNP calling, chromosome phasing, and structural variant detection **(Figure 1)**.

Figure 1 – Sample to analyses: the rich Omni-C™ data type enables applications beyond the study of genome conformation.

Omni-C libraries are designed to enable a variety of analyses that extend beyond genome conformation. A single Omni-C library enables genome-wide chromatin conformation analyses, genome assembly, haplotype phasing, and small and large variant detection.

Omni-C™ Sample to Analysis Workflow



2. Omni-C libraries enable genome-wide resolution of chromatin interactions.

Omni-C libraries contain all the proximity-ligation characteristics of Hi-C assays but are enriched in long-range *cis* reads compared to other RE-based Hi-C approaches **(Table 1).** When viewed at 1 Mbp and 100 kbp resolutions, the endonuclease approach of Omni-C technology generates data containing significant overlap of chromatin contacts with a standard RE-based Hi-C method (**Figure 2**). However, compared to RE-based methods, an increased number of long-range *cis* reads found in the Omni-C library combined with lack of sequence bias produces a measurable resolution improvement and increases read support for chromatin contacts and looping interactions **(Figure 3)**. Perhaps more importantly, Omni-C libraries enable the most complete view of genome-wide chromatin conformation by dramatically increasing resolution of topological interactions occurring in regions with low RE site density. The example in **Figure 4** shows contact matrices associated with the TERT gene, which is found in a DpnII desert near the telomeric region of chromosome 5. Due to the improved coverage in this region compared to RE-based Hi-C approaches, an Omni-C library provides a more complete contact matrix, revealing a loop associated with TERT that is not visible with RE-based Hi-C.

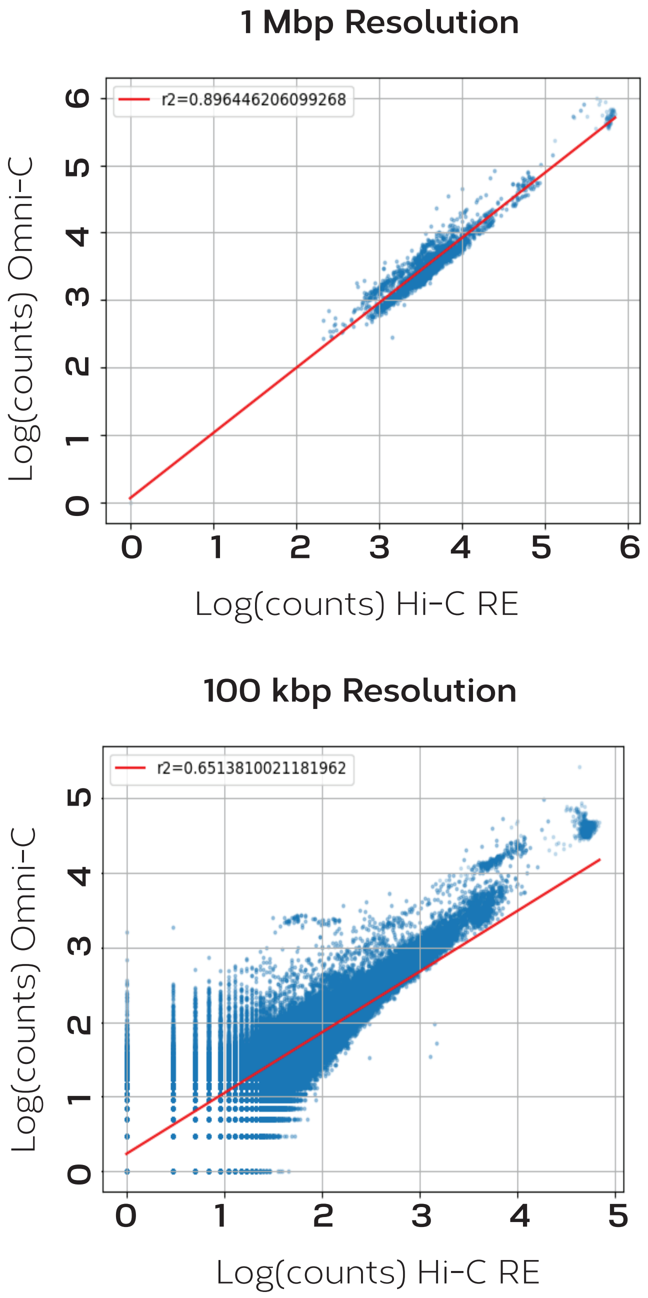
Table 1 – Omni-C libraries display expected proximity ligation characteristics with enriched long-range information.

Library Type	% *cis* <1kbp	% *cis* >1kbp	% unique molecules at 300 M read pairs
Omni-C	5%	95%	73%
Hi-C Multi-RE	20%	80%	69%
Hi-C Single-RE	34%	66%	50%

The enhanced resolution provided by Omni-C™ data may be used to more completely identify higher-order chromatin organization. To exemplify this, **Figure 5** compares contact maps of two distinct mouse tissues: liver and brain. Differential chromatin contacts, evident from the contact map, enables principle component analysis (PC1) to designate A/B compartments within each tissue type. Furthermore, changes in A/B compartments correlate with both histone modifications associated with active transcription (H3K4me3) and gene transcription (RNA-seq). As such, differential chromatin organization between tissue types evident from Omni-C data is highly correlated with tissue-dependent gene expression programs.

Figure 2 – Comparison of chromatin contacts between Omni-C and RE-based Hi-C libraries demonstrates the increased utility of Omni-C technology in chromatin conformation studies.

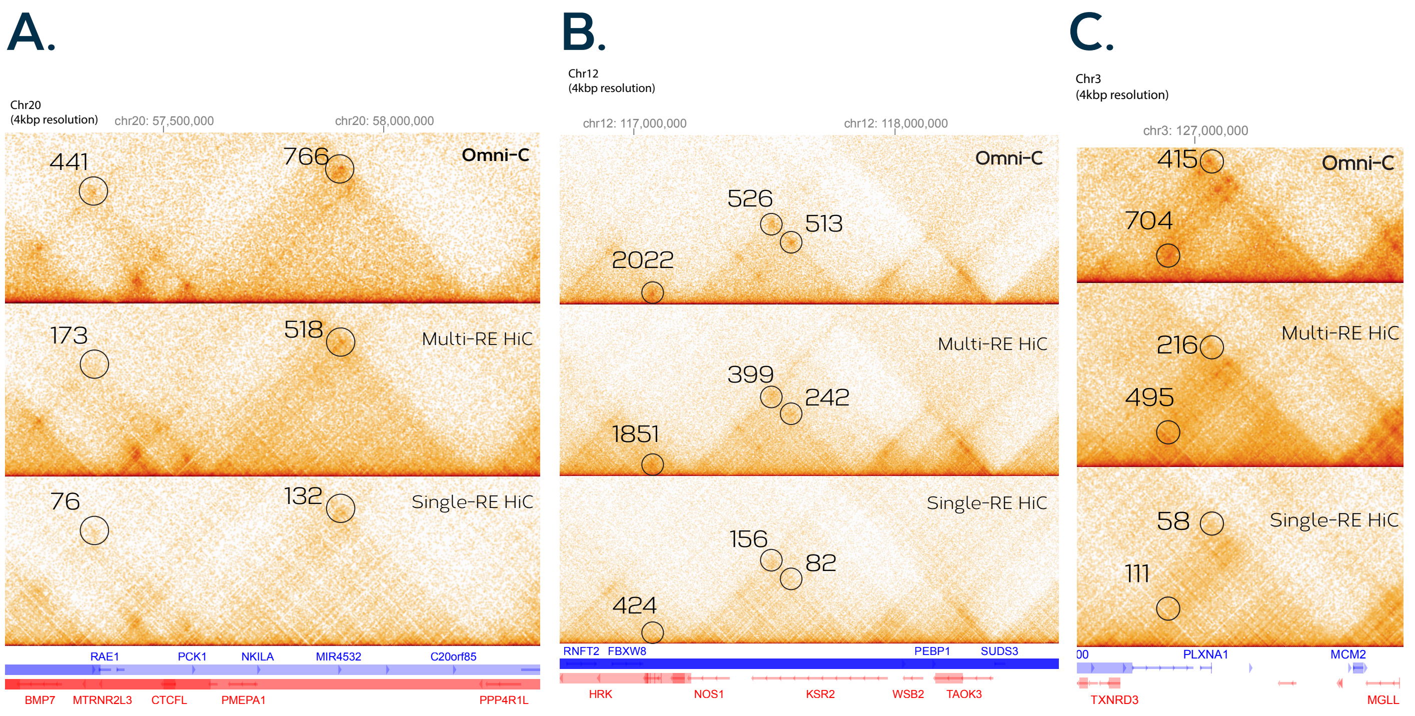
Scatterplots showing correlation between genome-wide contact points of Omni-C and singe RE-based Hi-C data compared at 1 Mbp and 100 kbp resolution. r^2 values indicate degree of correlation between datasets.


1 Mbp Resolution

100 kbp Resolution

Figure 3 – Omni-C technique uniquely provides improved resolution of chromatin conformation and looping interactions with lower sequencing depth.

Contact matrices for Omni-C, multi-RE Hi-C, and single-RE Hi-C libraries are shown at 4 kb resolution from 800 M total reads for each library type. Loops are outlined with circles and the number of raw read supports for each contact is indicated. Notably, Omni-C libraries contain equal or greater read supports genome-wide.



3. Omni-C libraries exhibit shotgun-like coverage distribution.

Omni-C technology's use of a sequence-independent endonuclease for chromatin digestion enables the generation of libraries with shotgun-like coverage qualities. When compared to other RE-based Hi-C approaches, Omni-C libraries exhibit shotgun-like read distribution per base coverage resulting in the most even genome-wide coverage of any proximity ligation methodology **(Figure 6)**. Typical RE-based Hi-C approaches are enriched in coverage around RE sites resulting in a wider coverage distribution, and thus less even coverage, with portions of the genome receiving no coverage at all. In contrast, the lack of sequence bias observed with Omni-C libraries results in a narrower distribution characteristic of a highly uniform genome-wide coverage. This "shotgun-like" coverage translates to less sequencing data required to achieve a given coverage depth thus reducing sequencing costs **(Table 2)**. For example, almost triple the read depth (and cost) is required to cover 80% of the genome at 30X depth in comparing Omni-C and single 4-base RE-based Hi-C libraries. Beyond reducing sequencing costs, an additional benefit to an Omni-C library's uniform coverage is to expand the use of proximity ligation data for applications such as genome-wide SNP detection, low error whole chromosome phasing, and structural variant detection.

4. SNP calling to assembly phasing

The even coverage described for Omni-C data enables genome-wide SNP calling and downstream applications dependent on SNP information. SNP calls identified with GATK show that Omni-C libraries are more sensitive and accurate for SNP detection compared to RE-based Hi-C approaches **(Table 3)**. When compared to the Illumina Platinum Genome truth set (Eberle *et al.,* 2017), Omni-C technology detects 98.6% of the homozygous SNPs with 99.3% precision. By calling SNPs with high fidelity, Omni-C technology can identify the unique nucleotide content from each chromosome pair to

Figure 4 – Omni-C libraries enable the most complete view of chromatin conformation by resolving genomic regions with low RE density.

Contact matrices are shown at two resolutions (8 kbp and 2 kbp shown in left and right columns respectively) for the Omni-C (top row) and single-RE Hi-C (bottom row) libraries each sampled at 800 M read pairs with gene tracks plotted below each contact matrix. This example is a cancer susceptibility locus found in a region of Chr5 containing the TERT gene. Loops for each library type are outlined with circles and the number of raw read supports for each contact is indicated. The Omni-C library provides a more complete view of this low-RE density region through generation of a less pixelated image and clear resolution of loops within the region.

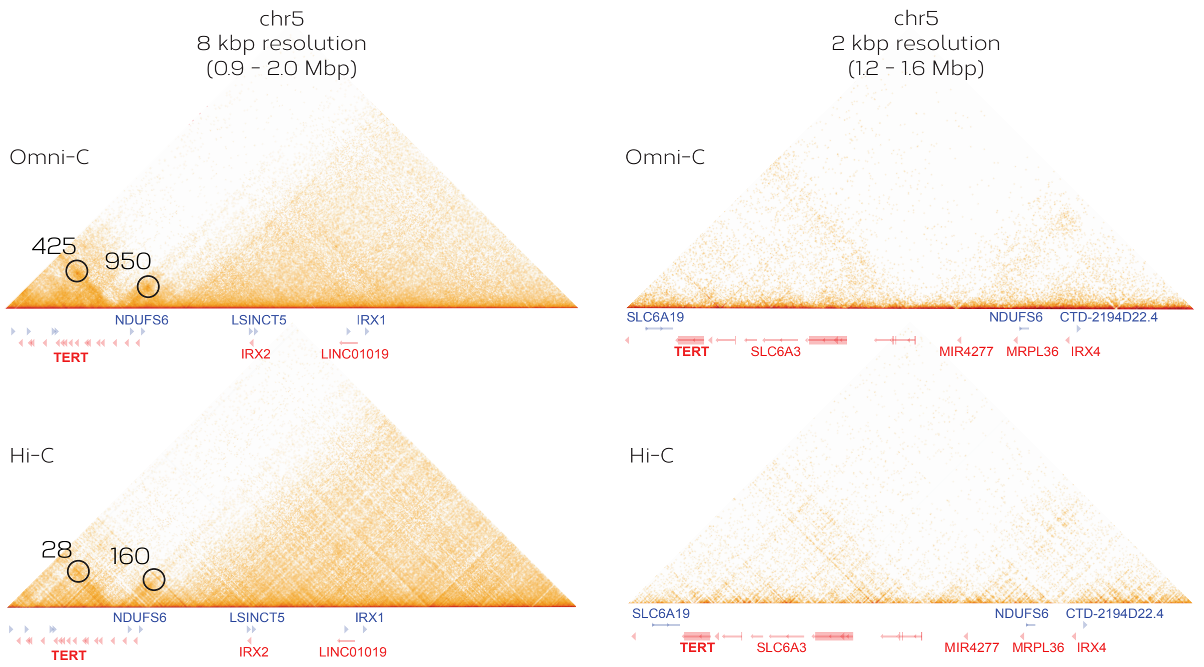


Figure 5 – Differential analysis of chromatin organization derived from Omni-C libraries strongly correlates with histone methylation and gene expression.

Contact matrices were generated from Omni-C libraries on mouse liver and brain tissues. Principle component analysis (PC1) designates the A/B compartments within each tissue type and correlates with the H3K4me3 histone modifications (a mark of active transcription; ENCODE data). Highlighted within the dashed red lines, "Brain A" indicates a change in chromatin topology to a brain-specific A compartment region with respectively increased gene expression (H3K4me3 and RNA-seq).

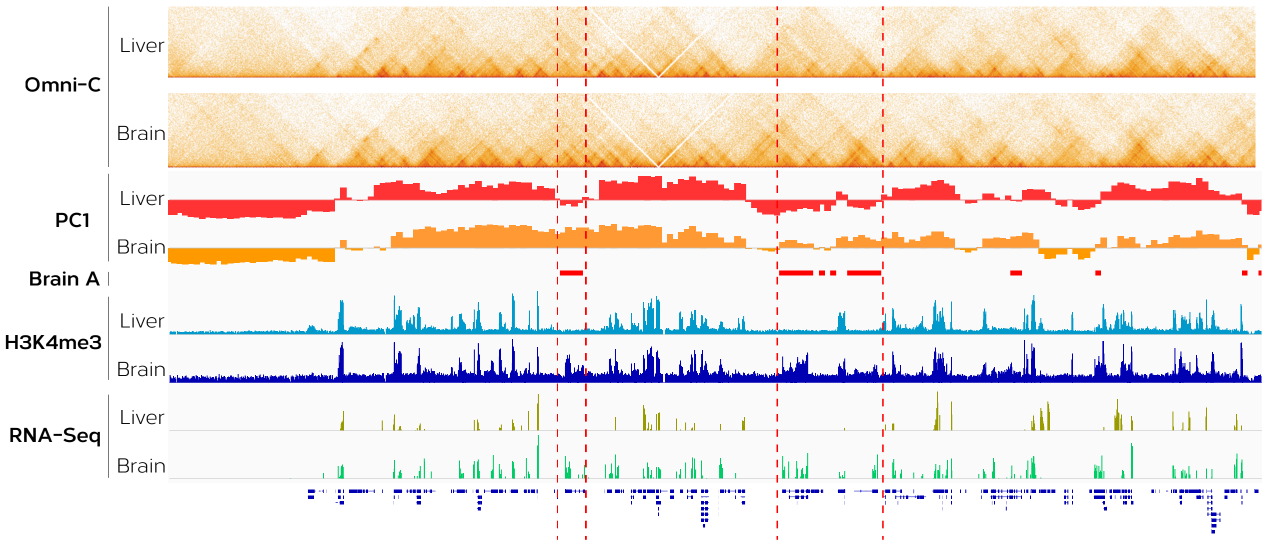


Figure 6 – Omni-C libraries display shotgun-like coverage without the sequence bias inherent to RE-based proximity ligation methods.

Omni-C (blue lines) and RE-based Hi-C (orange – multi-RE; pink – one-RE) libraries were compared to a shotgun library (black dashed line) at 300 M read pairs. A. Histogram plot of per base coverage. B. Coverage at the RE sites GATC (DpnII, MboI, Sau3AI) and GANTC (HinfI) plotted separately as the average of the absolute value both upstream and downstream of RE sites.

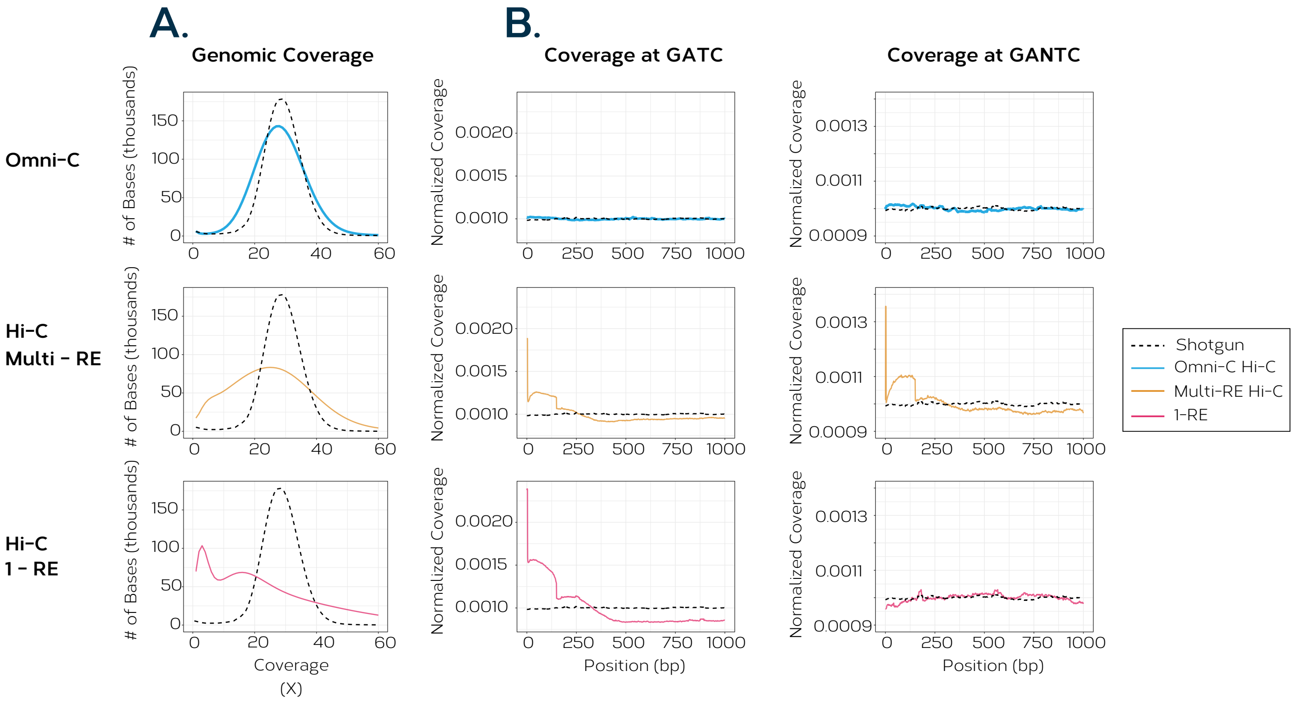


Table 2 – Omni-C library coverage greatly reduces project sequencing costs.

Sequencing requirements to achieve 30X coverage across 80% of the genome are presented for shotgun libraries and different proximity-ligation approaches. Coverage was calculated from an empirical distribution for each library type and the number of read pairs to achieve 30X mean coverage across 80% of the genome was extrapolated. The cost associated with the sequencing depth is based on an Illumina NovaSeq 6000 SP – 300 cycle flow cell (1.6 billion read-pairs, cost $6,600.00 USD) and is relative to the shotgun library (base price $1,816).

Library Type	Actual X coverage needed to achieve 30X across 80% of genome	How many read pairs does that equate to (2x150)	Cost differential relative to shotgun on Nova Seq
Shotgun	41	440 Million	$0
Omni-C	**51**	**543 Million**	**+$424** **(+23%)**
Hi-C Multi-RE	73	782 Million	+$1,411 (+78%)
Hi-C Single-RE	134	1,427 Million	+$4,071 (+224%)

accurately phase assemblies into haplotypes. In fact, it is the combination of proximity ligation and SNP information uniquely found in Omni-C libraries that enables the use of this Hi-C data type for physical linkage of heterozygous SNPs to phase assemblies. Use of proximity ligation information with HapCUT2 enables chromosome phasing end to end whereas a shotgun dataset sampled at the same depth only successfully phased 28.2 kb as the largest phase block. Furthermore, the improved coverage of Omni-C libraries enables correct phasing of 84% of the heterozygous SNPs with 1% switch error rate with only 60X coverage **(Table 4)**. As such, Omni-C technology offers the best possible approach for whole-genome physical phasing using Illumina short reads.

5. Large SVs are captured in Omni-C data

Proximity ligation data can be used to detect and confirm chromosomal rearrangements, like in tumor samples, using open-source software programs. Contact matrices enable the quick visualization of such large structural variants (SVs). We demonstrate this capability in the well characterized breast cancer

Table 3 – Omni-C library SNP call sensitivity and precision approaches shotgun library level.

The Genome Analysis Toolkit (GATK) (McKenna et al., 2010) was used to make homozygous SNP calls on GM12878 libraries, each sampled at 300 M read pairs. Reported values are concordance with the Illumina Platinum Genome high confidence call set, in the intersection of the high-confidence regions for the various library types.

Library Type	True Positive	False Positive	False Negative	Sensitivity	Precision
Shotgun	2,696,291	9,270	6,814	99.7%	99.7%
Omni-C	2,666,339	20,081	36,766	98.6%	99.3%
RE Based Hi-C	2,387,235	33,554	315,870	88.3%	98.6%

Table 4 – Haplotype completeness and accuracy approaches 85% using Omni-C libraries.

HapCUT2 (Edge et al. 2017) was used to assemble GM12878 haplotypes using high confidence heterozygous SNPs from the Illumina Platinum Genome. Each library was sampled at 800 M read pairs.

Library Type	#Variants Phased	% Heterozygous SNPs phased	Switch Error Rate	Largest Phase Block Size	#Chromosomes phased end to end
Shotgun	2,229,492	81.44%	0.0036	28.2 kb	0
Omni-C	2,299,248	84.00%	0.0100	248.01 Mb	23
RE Based Hi-C	1,986,467	72.60%	0.0357	247.95 Mb	23

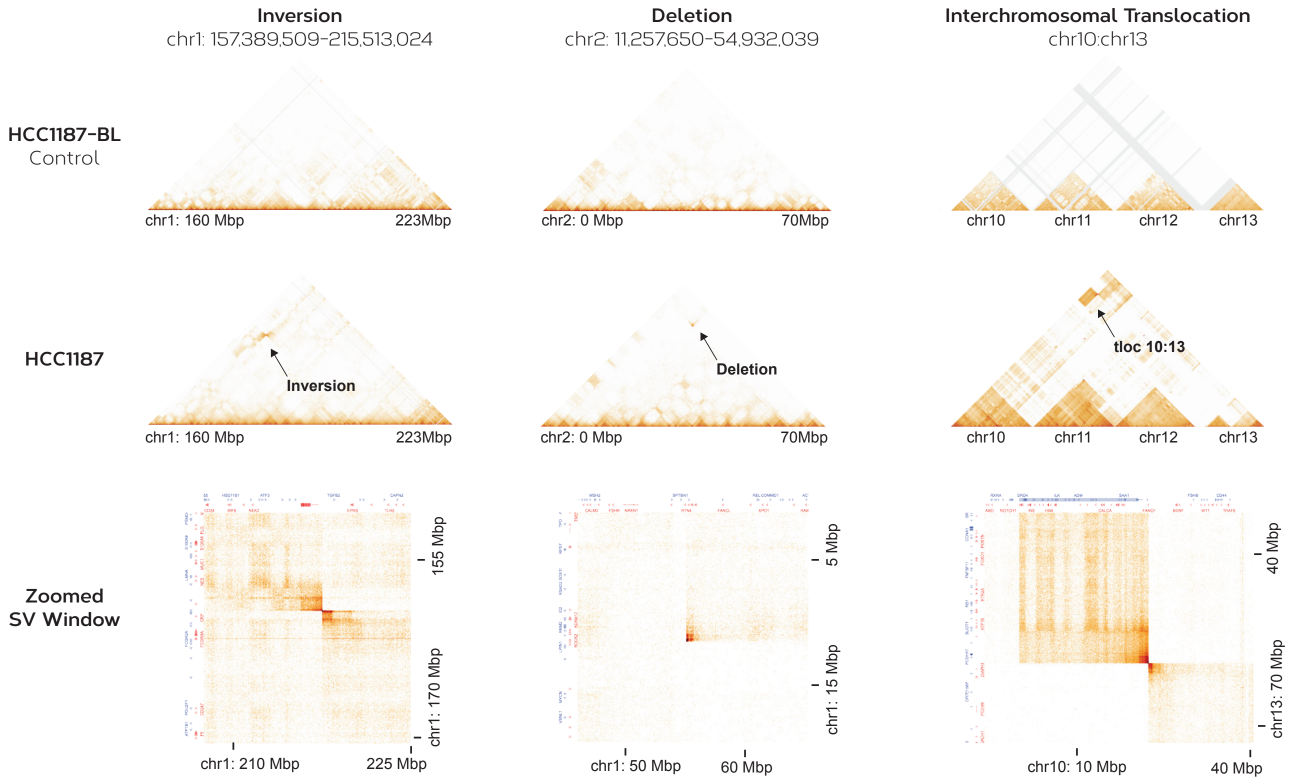
cell line, HCC1187 **(Figure 7)**. Examples of inversions, deletions, and inter-chromosomal translocation SVs are visible in HCC1187 contact matrices and absent in the matched normal cell line, HCC1187-BL. The genes associated with the SV break points can be observed. Moreover, the uniform coverage of Omni-C data should allow for SV break point resolution at single-nucleotide level.

6. Conclusion

We have highlighted the properties and several possible applications of the Omni-C™ technology, a sequence-independent endonuclease-based proximity ligation assay. Omni-C libraries contain the same topological features at low resolution as RE-based Hi-C, with the added benefit of coverage uniformity enabling a true genome-wide view. Omni-C technology's shotgun-like coverage facilitates genome-wide SNP calling in a manner that is independent of RE site location which, in turn, enables effective full chromosome haplotype phasing. We also demonstrate the ability to identify large structural variants using contact matrices generated from Omni-C libraries. Omni-C technology captures genome-wide topology at single nucleotide resolution in a single library prep. Therefore, in addition to being ideal for the study of 3D genome conformation, it is suitable for applications traditionally addressed by shotgun libraries.

Figure 7 – Omni-C captures large structural variants present in cancer samples.

Three examples of large SVs (> 1 Mbp) are shown in the breast cancer cell line HCC1187. These validated SVs (Stephens et al., 2009) are displayed in the contact matrices. The contact matrices on the top row are from the tumor negative control HCC1187-BL, the middle row the tumor positive HCC1187, and the bottom row displays the chromosomal break points mapped to gene tracks showing genes within their vicinity. Break point signal found in HCC1187 (identified with black arrows) are not present in the control HCC1187-BL cell line, indicating these SVs are not an artifact of the proximity-ligation assay.



References

Eberle, MA *et al.* (2017) A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Research 27: 157-164. doi:10.1101/gr.210500.116

Edge, P *et al.* (2017) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Research 27: 801-812. doi:10.1101/gr.213462.116

McKenna, A *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297-1303. doi: 10.1101/gr.107524

Stephens, PJ *et al.* (2009) Complex landscapes of somatic rearrangements in human breast cancer genomes. Nature 462(7276):1005-1010. doi: 10.1038/nature08645

Dovetail GENOMICS

The Genome Unrestricted.