Tech Note

Dovetail® AssemblyLink™ Kit

A Rapid Path To High Quality Genome Assemblies

Introduction

While long-read sequencing represents a significant improvement over short-read NGS data for genome assembly, it alone is not sufficient to fully scaffold and accurately orient contigs for complex genomes. Hi-C continues to be a key requirement for generating chromosome-scale assemblies. Despite its utility, Hi-C library generation has historically been challenging due to its lengthy and labor-intensive workflow. Additionally, traditional Hi-C data exhibit uneven genomic coverage constraining its use in more advanced genome assembly applications such as resolving haplotypes.

The Dovetail® AssemblyLink™ Kit, built on Dovetail® LinkPrep™ technology, is designed to overcome these challenges. This novel chemistry offers a rapid single-day workflow with minimal sample material, delivering unbiased long-range reads and highly uniform sequence coverage.

Easy Workflow, Rapid Library Generation

The Dovetail AssemblyLink Kit offers a simplified workflow enabling the production of sequencing-ready libraries in a single day (Figure 1). AssemblyLink™ data is compatible with a wide variety of open-source tools such as Salsa2 and YaHS, as well as the Dovetail® HiRise® Scaffolding Software.

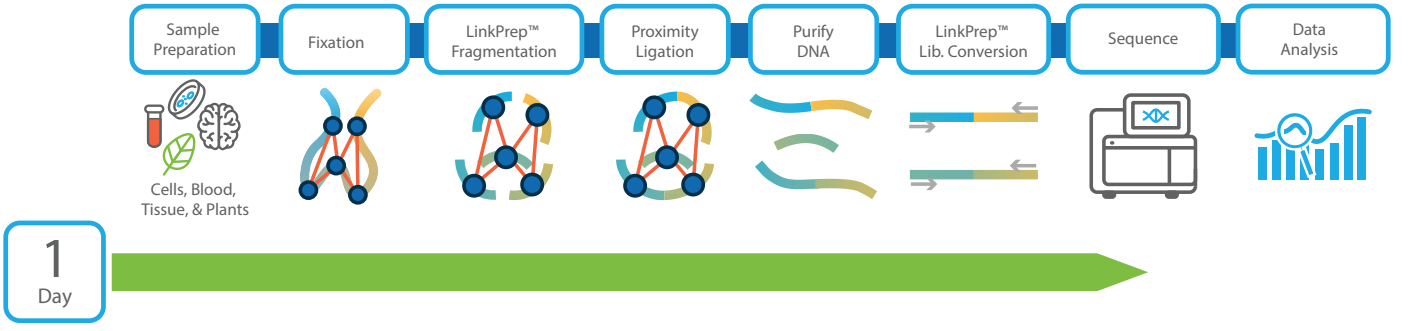

Figure 1. The AssemblyLink™ workflow generates Hi-C libraries in a single day. The workflow starts with sample (cells, blood or tissue) crosslinking. Next, a tagmentation step simultaneously fragments the chromatin and adds adaptor sequences. Proximity ligation creates chimeric molecules linking DNA fragments that are close to each other in the 3D space. Finally, the crosslinks are reversed, and purified DNA has a streamlined library conversion including the removal of shearing and biotin-based enrichment, generating an Illumina-compatible sequence library.

Library Type	% Mapped	% No-Dup	% *Trans*	% *Cis*	% Long-Range *Cis*
AssemblyLink	78.5%	78.4%	11.6%	66.8%	45.1%
Hi-C Multi-RE	73.8%	73.7%	11.5%	62.1%	41.4%
Hi-C Single-RE	68.7%	68.5%	21.5%	47.0%	22.8%

Table 1. The AssemblyLink Kit delivers high-quality libraries as assessed by standard Hi-C library metrics. Mapped, no-dup, *trans*, *cis*, and long-range *cis* read pairs are calculated as a percent of total read pairs in the library. All libraries were generated from GM12878 cells, subsampled to 1 million (2 x 150 bp) read pairs, and processed through the Dovetail Genomics AssemblyLink™ QC pipeline.

The AssemblyLink Kit Delivers High-Quality Hi-C Libraries

AssemblyLink data contains long-range, chimeric read pairs required for genome scaffolding. Moreover, AssemblyLink data contains improved mapping rates with more *cis* pairs, further increasing assembly-informative reads compared to traditional Hi-C approaches (Table 1).

Broad Range Of Samples Supported

The AssemblyLink Kit supports a broad range of sample types including animal tissues and plant leaves. Libraries generated from moth, beetle, hydra, and anemone all show the expected distribution of *cis* read pairs as a function of insert size consistent with a high-quality library suitable for chromosome-scale genome scaffolding (Figure 2). Importantly, for each library, the maximum observed insert size aligns with the longest assembly scaffold, reflecting the assay's ability to span full chromosomes.

All four libraries were prepared from 10 – 20 mg of tissue, the recommended input amount for the AssemblyLink™ assay (Table 2). This reduction in sample input amount, compared to 50 – 200 mg required for alternative Hi-C methods, provides flexibility when working with small specimens and minimal sample material.

Scaffold Draft Assemblies

Using AssemblyLink data, fragmented draft assemblies are elevated to high-quality chromosome-scale assemblies. As an example, an AssemblyLink™ library generated from a plant leaf sample, sequenced to 30X coverage, was used to scaffold a PacBio HiFi draft assembly using Dovetail HiRise Scaffolding Software. The resulting assembly statistics show significant improvements in N50 and L50 values, and a high BUSCO score, consistent with greatly improved assembly contiguity and completeness (Table 3) compared to the initial draft.

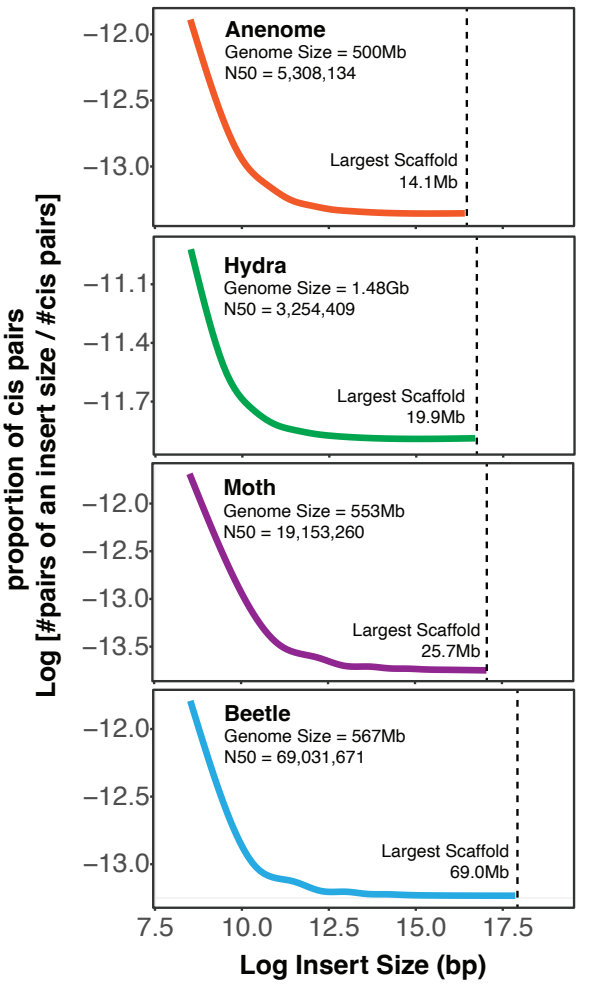

Figure 2. Distance-dependent interaction frequency curves. *Cis* read pairs display the expected decay frequency as linear genomic distance increases. The dashed lines denote the longest scaffold. Library inserts all truncate at the longest scaffold indicative of read pairs spanning long distances up to full chromosomes.

Sample Type	Input Amount
Mammalian Cells	1 million
Mammalian Blood	3 mL
Animal Tissues	10 - 20 mg
Plant Leaves	30 - 60 mg

Table 2. Sample Input Requirements for AssemblyLink Kit

Genome Size	Draft Assembly N50 / L50	Post AssemblyLink Scaffold N50 / L50	BUSCO Score
2.76 Gbp	9.89 Mbp / 83	209.57 Mbp / 6	97.65%

Table 3. PacBio HiFi draft assembly of a plant scaffolded with AssemblyLink data show high N50 and BUSCO Score.

The output assembly was compared to a second assembly produced using the same PacBio HiFi input data but instead scaffolded using Dovetail Omni-C® data. The scaffolding performance observed using the two data sets was highly comparable, confirming that the rapid AssemblyLink™ workflow does not comprise the final assembly (Figure 3).

Capture More Of The Genome, Without The Bias

AssemblyLink data exhibit uniform sequence coverage across the genome compared to traditional Hi-C. Traditional Hi-C's use of restriction enzymes to digest the chromatin results in sequence bias leading to significant under- and over-sequenced portions of the genome (Figure 4). During the AssemblyLink workflow, chromatin is fragmented using Tn5 transposase under conditions driving uniform insertion and tagging which in turn enables the generation of libraries with uniform, unbiased genomic coverage. This more complete view of the genome enables improved haplotype-aware assemblies.

Create Haplotype Resolved Assemblies

The uniform coverage of AssemblyLink data enables more comprehensive heterozygous SNP detection compared to traditional Hi-C. This feature, combined with the long-range information captured in AssemblyLink libraries, facilitates accurate chromosome-scale SNP phasing (Table 4).

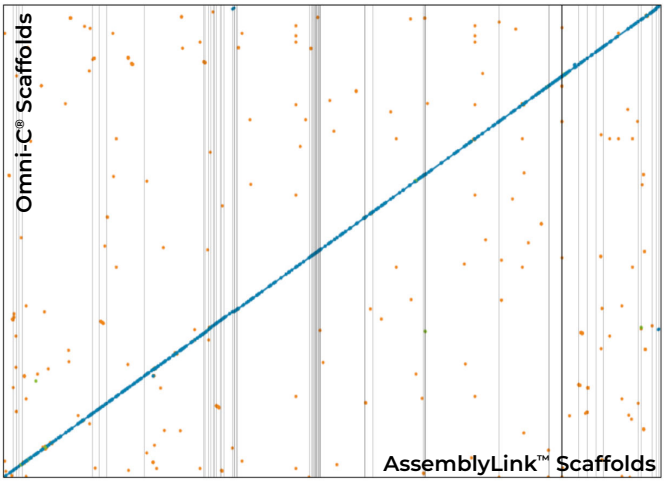

Figure 3. Syntenic dotplots comparing two assemblies of the same plant sample showing that the large scaffolds of the assemblies align perfectly. The x-axis represents the AssemblyLink™ assembly used as the reference, while the y-axis represents the Omni-C® assembly used as the query.

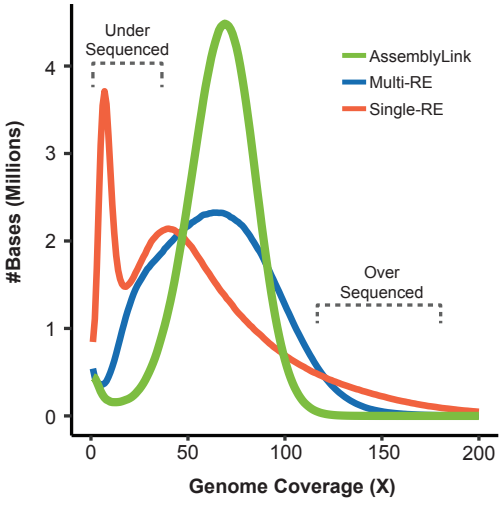

Figure 4. AssemblyLink data provides an unbiased view of the genome. 800M read pairs were aligned for each library type. AssemblyLink data displays uniform coverage distribution with a mean coverage of 65.8 X from 800M read pairs. Traditional Hi-C libraries display a wider, non-Poisson coverage distribution due to RE-based sequence bias.

Library Type	Phased het SNPs in largest phase block	Phased het SNPs in any phase block	Switch error rate	Largest phase block	Chromosomal scale phasing achieved?
Shotgun	77	55.00%	0.042%	7,685 bp	No
AssemblyLink	**147,039**	**95.27%**	**0.156%**	**248.1 Mb**	**Yes**
Hi-C Multi-RE	133,943	67.59%	0.213%	248.3 Mb	Yes
Hi-C Single-RE	113,284	47.09%	0.276%	248.0 Mb	Yes

Table 4. The uniform coverage of AssemblyLink libraries enables phasing of >95% of the heterozygous SNPs with <0.16% switch error rate. SNPs called in cell line GM12878 with DeepVariant (Poplin *et al*. 2018) were matched to the truth set containing high-confidence heterozygous SNPs from the Illumina Platinum Genome across the various library types. AssemblyLink data's uniform coverage enables near-shotgun performance for SNP callability. HapCUT2 (Edge *et al*. 2017) was used to phase the heterozygous SNPs into phase blocks. All input data was normalized to 800M read pairs. The data displayed in the table was computed over Chr1 (the largest chromosome).

The captured phase information enables the generation of high-quality, haplotype-resolved assemblies, offering the most accurate representation of the genome. To highlight this capability, AssemblyLink data was used to generate a fish diploid genome using Hi-C integrated Hifiasm (Cheng *et al*., 2021; Figure 5). Two draft haplotype assemblies were independently generated and scaffolded using the Dovetail HiRise Scaffolding Software.

The contact matrices of the HiRise® scaffolded haplotypes (Figure 5) are highly similar and show a chromosome-scale assembly without off-axis signals indicative of high-quality. BUSCO and Merqury analyses of each haplotype assembly show high assembly completeness and accuracy (Figure 5).

Summary

The AssemblyLink Kit from Dovetail Genomics is a transposase-based next-generation Hi-C approach that delivers high-quality sequence-ready libraries in a single day. This off-the-shelf kit is ready to be used for genome assembly across a broad variety of samples with minimal input requirements. AssemblyLink data offers highly uniform coverage, facilitating genome-wide SNP calling and phasing, enabling the production of haplotype-resolved assemblies.

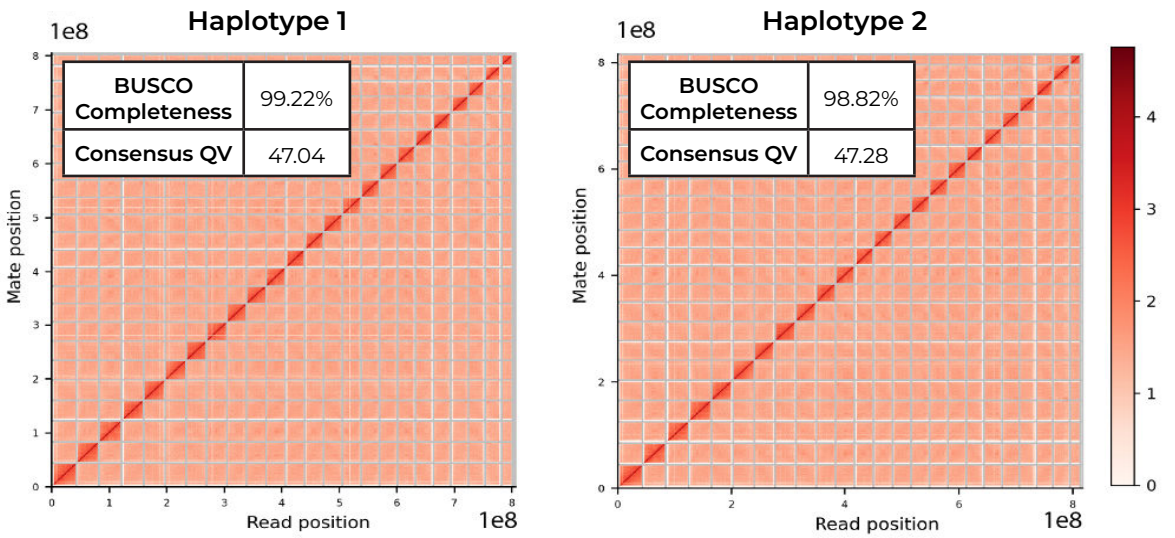

Figure 5. Scaffolding with AssemblyLink data yields chromosome-scale haplotype-resolved genome assemblies. Contact matrices showing a haplotype-resolved fish diploid genome with BUSCO scores and quality values (QV).