Dovetail Genomics Haplotype-Resolved Assembly

Why settle for half a genome when you can have it all?

- **Haplotype-resolved assembly scaffolded up to chromosome-scale**
- **K-mer-based quality assessment leveraging the highly uniform coverage of Omni-C® data**

INTRODUCTION

The tools available for capturing genomic information have evolved dramatically over the past decade, yet human, plant and animal reference genomes remain primarily haploid. A true diploid assembly is an enriched data set that enables a more complete understanding of:

- Hybridization
- *Cis* versus *trans* mutations
- Allele-specific expression
- Structural variation
- LD patterns
- Segmental duplications
- Complex gene families
- And much more.

The newest Dovetail® *de novo* assembly workflow utilizes a unique combination of PacBio HiFi sequencing, for best-in-class base calling accuracy, and Dovetail® Omni-C® directed phasing and scaffolding for true diploid chromosome-scale genome assemblies.

PROCESS & WORKFLOW

Sequence contigs, generated using highly accurate PacBio HiFi reads, are fully phased using Hi-C integrated Hifiasm (*Cheng et al., 2021*) and Omni-C® data, a Hi-C datatype with highly uniform coverage.

The two draft haplotype assemblies are then scaffolded using the Dovetail® HiRise® Scaffolding Pipeline guided by the long-range information captured with Omni-C® read pairs. Omni-C data provides additional advantages, including assembly completeness and consensus quality value (QV), assessed using Merqury (*Rhie et al., 2020*), and topologically associated domains (TADs) mapping.

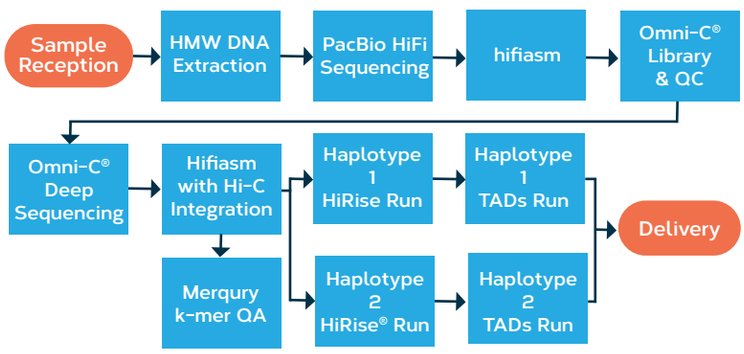

Figure 1. Dovetail Genomics haplotype-resolved assembly workflow.

SHOTGUN-LIKE GENOME COVERAGE OF OMNI-C® DATA

Omni-C® libraries have more even whole genome coverage when compared to restriction enzyme-based (RE) Hi-C libraries and therefore, the data output more closely resembles whole genome shotgun data (Figure 2).

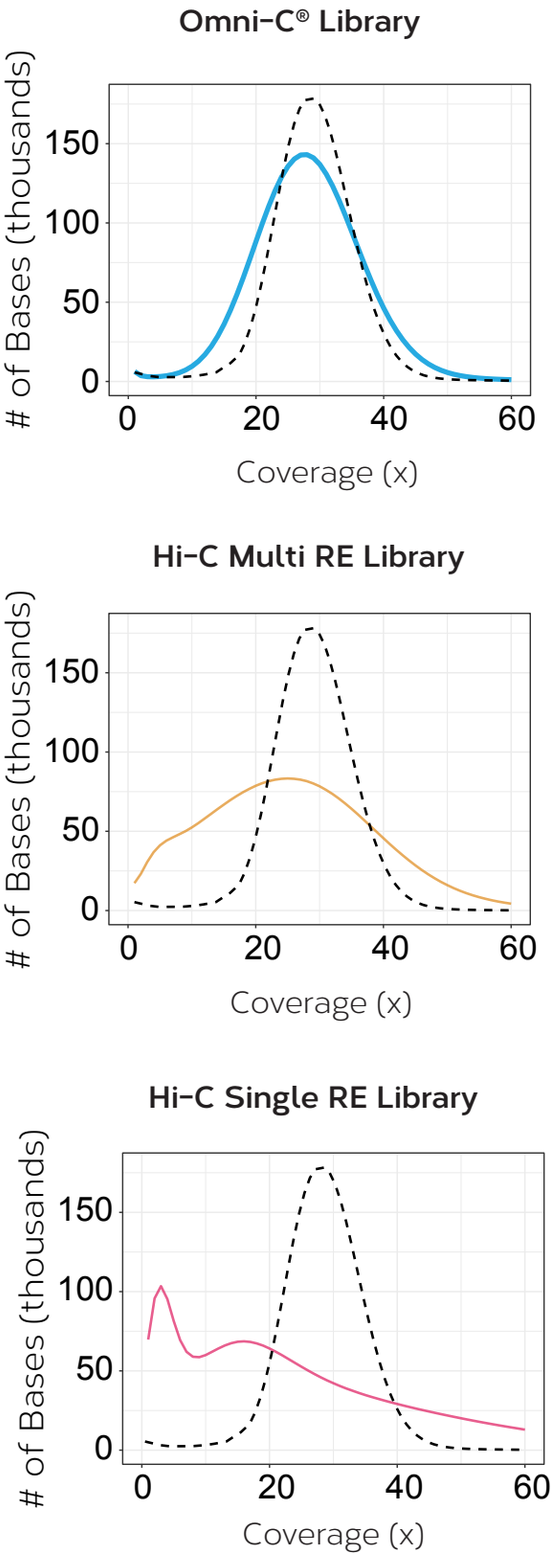

Figure 2. Sequence coverage of Omni-C, multi-RE-Hi-C and single RE-Hi-C libraries. Shotgun distribution is shown with a dotted line. Omni-C coverage is essentially the same as shotgun coverage.

HUMAN DIPLOID ASSEMBLY ACCURACY

Shotgun-like properties of Omni-C® data enable unbiased and near complete whole genome coverage of heterozygous SNPs. This, coupled with the long-range proximity ligation data captured by Omni-C, enables SNPs to be accurately phased on a chromosomal scale with very low switch error rate (Table 1).

Assembly	Hifiasm Hi-C HiRise Hap 1	Hifiasm Hi-C HiRise Hap 2
Number of blocks	1,342	1,166
Total bases in block	3,019,243,135	2,830,692,833
Block N50 size	19,637,744	19,612,443
Longest block size	150,691,192	86,811,354
Switch error rate	**0.107599%**	**0.130775%**

Table 1. Phasing accuracy of human diploid assemblies after HiRise® scaffolding.

ASSEMBLY QUALITY ASSESSMENT WITH MERQURY

Since Omni-C data behaves like shotgun data, Omni-C® k-mers can be used by Merqury to produce an unbiased, reference-free assessment of assembly completeness and accuracy.

In the highly heterozygous Atlantic bluefin tuna, Merqury analysis shows total assembly completeness exceeding 99%. Haplotype-specific competeness is lower due to a high level of heterozygosity. Figure 3 shows the k-mer plot of these assemblies.

Assembly	Completeness	Consensus Quality Value (QV)
Haplotype 1	87.44%	49.84
Haplotype 2	87.38%	50.18
Combined Haplotypes	99.04%	50.01

Table 2. Completeness and QV value of the Atlantic Tuna diploid assemblies.

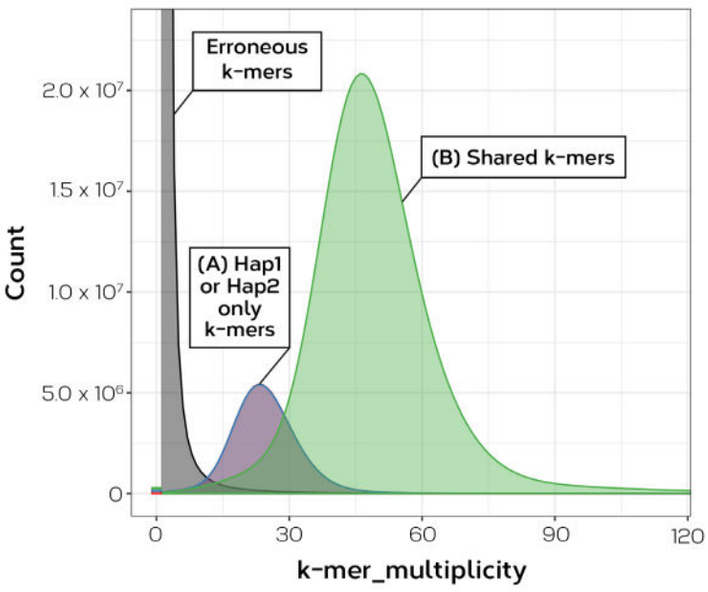

Figure 3. K-mer plot of diploid Atlantic Tuna assemblies. The heterozygous k-mers overlap, generating peak (A) where haplotype specific k-mers overlap. Peak (B) shows Hap1 and Hap2 shared k-mers.

FULLY SCAFFOLDED HAPLOTYPES

Figure 4 shows contact matrices of the HiRise® scaffolded haplotypes of the California Brush Lizard (*Urosaurus nigricaudus*). Both matrices are highly similar and do not show any off-axis signal, which would be indicative of an assembly error.

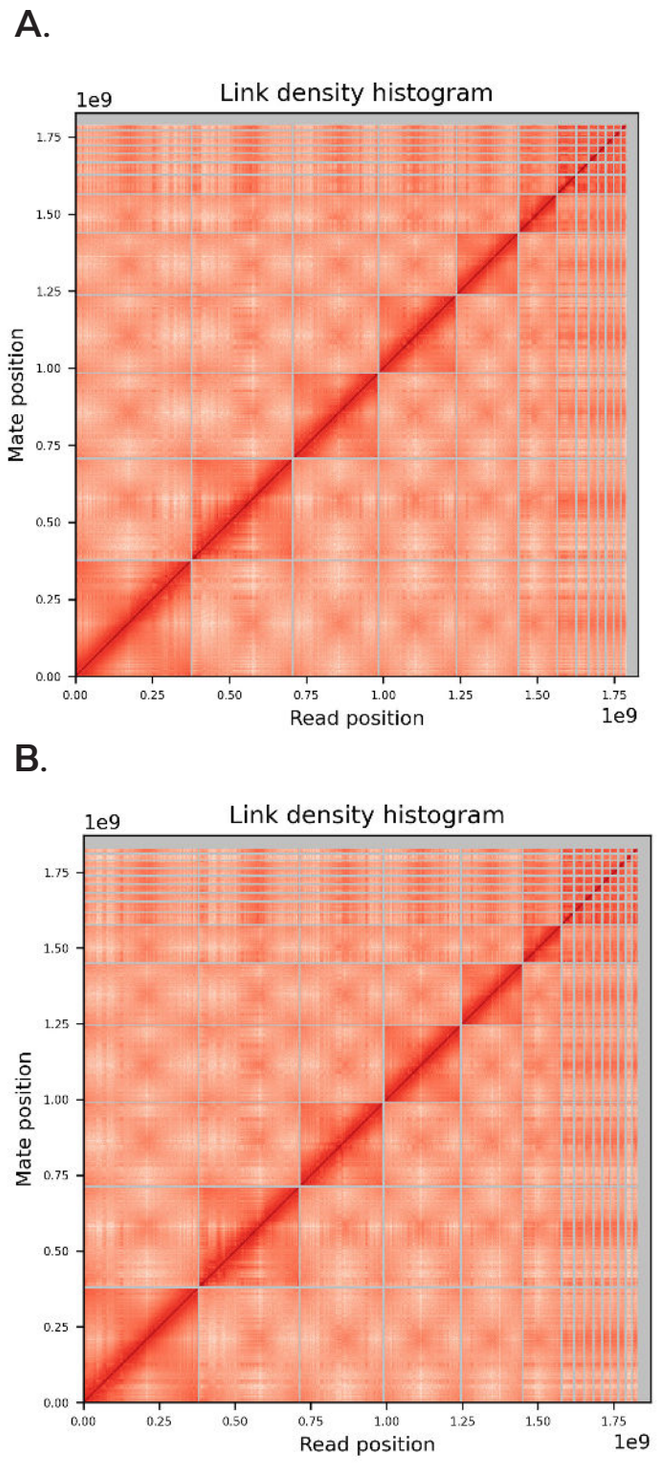

Figure 4. Contact matrix of HiRise® scaffolded haplotype assemblies of CA Brush Lizard (*Urosaurus nigricaudus*) A. Haplotype 1. B. Haplotype 2.

OBTAIN HIGH QUALITY RESULTS FROM THE DOVETAIL® HAPLOTYPE-RESOLVED ASSEMBLY PIPELINE

Combining high accuracy PacBio HiFi and Omni-C reads, Dovetail Genomics reports high completeness for haplotyped-resolved assemblies across multiple species.

Table 3 documents scaffold N50 and L90, BUSCO eukaryote database complete %, Merqury Omni-C® data-based completeness for haplotype and combined assemblies, and QV score across multiple species. Notably, BUSCO scores are not as accurate as the k-mer completeness assessment.

REFERENCES

Cheng, H., Jarvis, E. D., Fedrigo, O. et al. Robust haplotype-resolved assembly of diploid individuals without parental data (2021). https://arxiv.org/abs/2109.04785

Rhie, A., Walenz, B.P., Koren, S. et al. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21, 245 (2020). https://doi.org/10.1186/s13059-020-02134-9

Species	Scaffold (Mb)			BUSCO (%)	Completeness (%)		QV
	Hap	N50	L90				
Homo sapiens (Human)	1	143.5	22	94.90	97.41	99.31	46.42
	2	145.1	20	94.51	94.51		46.93
Dasyurus viverrinus (Eastern Quoll)	1	434.7	8	96.47	97.41	99.41	43.44
	2	625.0	6	96.08	97.51		43.46
Thunnus thynnus (Atlantic Tuna)	1	34.1	22	99.61	87.44	99.04	49.84
	2	34.2	22	99.61	87.38		50.18
Urosaurus nigricaudus (Baja California brush Lizard)	1	277.9	10	97.65	97.74	98.75	45.78
	2	277.4	8	92.55	96.06		45.97
Encelia farinosa (Brittlebrush)	1	71.2	17	100	98.95	99.02	33.99
	2	73.7	17	100	98.94		34.04
Ensete ventricosum (Ethiopian banana)	1	56.4	11	97.65	88.81	98.64	37.10
	2	56.7	9	97.65	88.87		37.78
Euphorbia peplus (Milkweed)	1	31.8	8	99.30	94.03	96.32	49.16
	2	36.2	8	99.30	94.82		50.83

Table 3. Scaffold N50 and L90, BUSCO Eukaryote database complete %, Merqury Omni-C®-based Completeness for each haplotype and combined assemblies, and QV score across different species. (Hap=Haplotype, BUSCO=BUSCO Eukaryotic Ortholog (C), Complete=Omni-C K-mer Completeness, QV=Omni-C K-mer QV Score)