Application Note

Don’t Settle for Half a Genome

Dovetail® Assembly Services

Key Takeaways

- Dovetail Assembly Services deliver fully scaffolded, chromosome scale haplotype-resolved assemblies.
- Highly uniform genome coverage of Dovetail® data enables robust assembly quality assessment.
- Capture of parentally inherited genetic variation enables further exploration and deeper biological understanding.

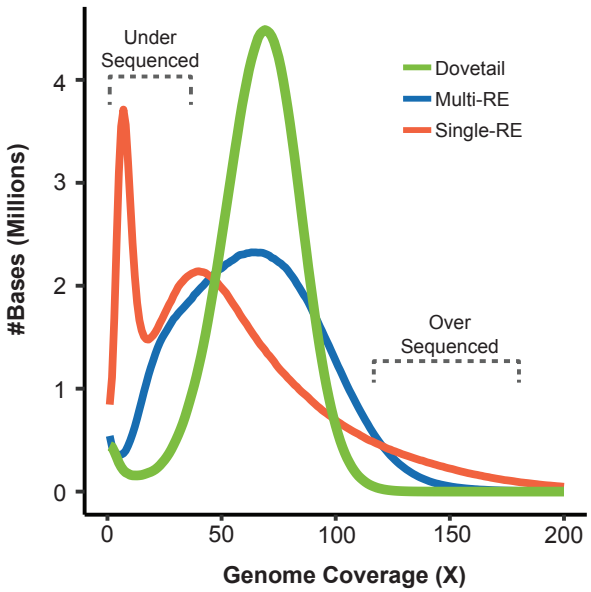

Figure 1. Sequence coverage of Dovetail data compared to single and multiple restriction (RE) enzyme Hi-C. Dovetail data coverage closely approximately coverage achieved by standard shotgun sequencing.

Introduction

The tools available for capturing genomic information have evolved dramatically over the past decade, yet reference genomes primarily remain haploid. A true diploid assembly is an enriched data set that offers a more complete catalogue of genetic variation enabling a deeper biological understanding including species hybridization, allele-specific expression, linkage disequilibrium patterns, and decoding of complex gene families.

Furthermore, it enables complex events to be identified, such as compound mutations, structural rearrangements, and segmental duplications.

Process and Workflow

Dovetail® data is a linked-read, Hi-C-like datatype. As an outstanding feature, it provides highly uniform sequence coverage (Figure 1).

The Dovetail® *de novo* assembly workflow utilizes a unique combination of PacBio HiFi sequencing, for best-in-class base calling accuracy, and Dovetail® data enabling directed phasing and scaffolding for true chromosome-scale diploid genome assemblies (Figure 2).

Sequence contigs, generated using highly accurate PacBio HiFi reads, are fully phased and partitioned using Hi-C integrated Hifiasm (Cheng *et al.*, 2021).

Following partitioning of the reads, the two draft haplotype assemblies are scaffolded using the

Dovetail® HiRise® Scaffolding Software guided by the long-range information captured with the Dovetail® linked-reads.

Dovetail data provides additional advantages, including the ability to compute assembly completeness and consensus quality values (QV), assessed using Merqury (Rhie *et al.*, 2020), and the ability to assess 3D genome organization through mapping topologically associated domains (TADs).

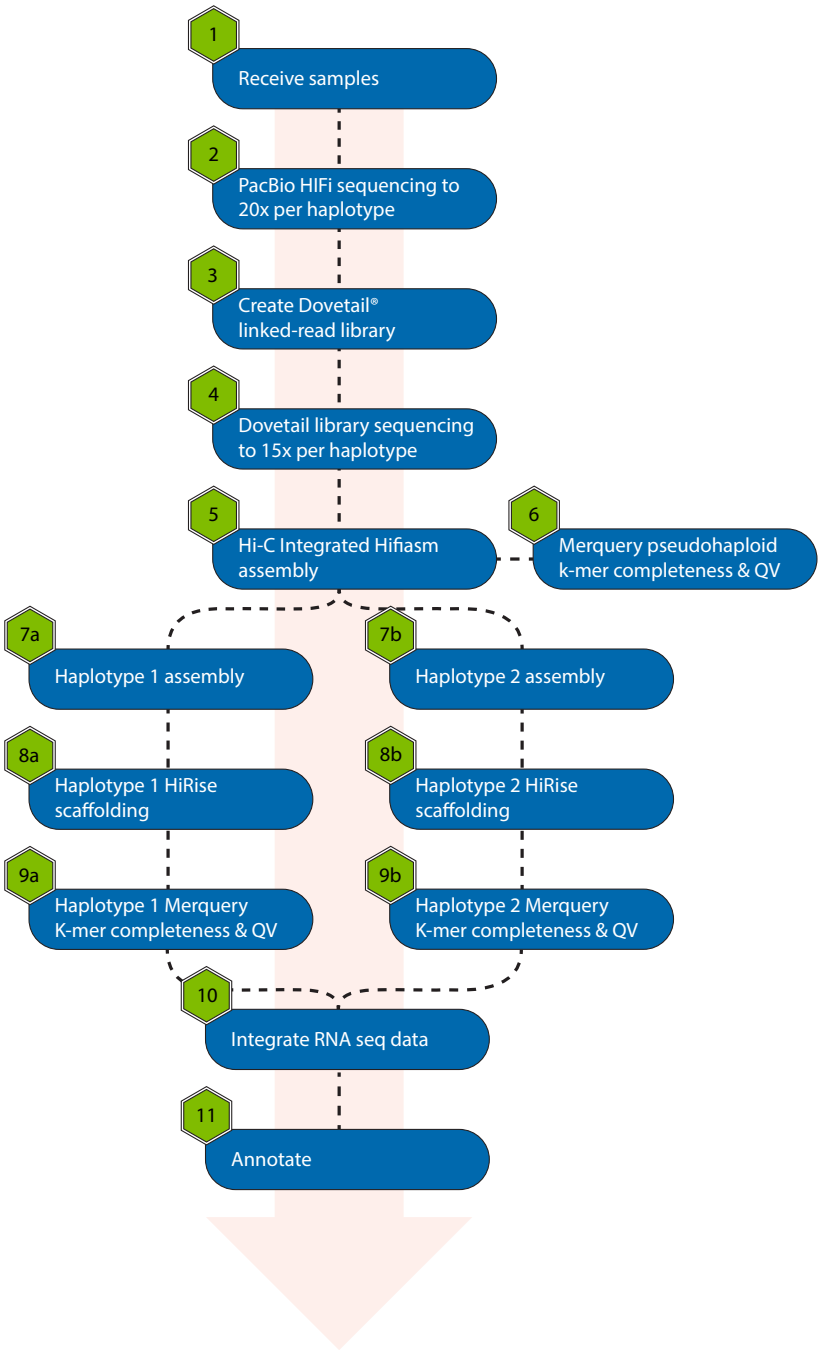

Figure 2. Dovetail *de novo* assembly workflow.
Diagram showing the major steps from receipt of sample to a fully annotated and ready to publish high-quality genome assembly.

Assembly	Completeness	Consensus QV
Haplotype 1	**87.44%**	**49.84**
Haplotype 2	87.38%	50.18
Combined Haplotypes	99.04%	50.01

Table 2 High completeness and consensus quality values (QV) for an Atlantic Tuna diploid assembly. Representative of the general high quality outcomes for diploid species, the QC metrics for the Atlantic Tuna confirm reaching a exceptionally complete, high quality assembly using the Dovetail *de novo* assembly process.

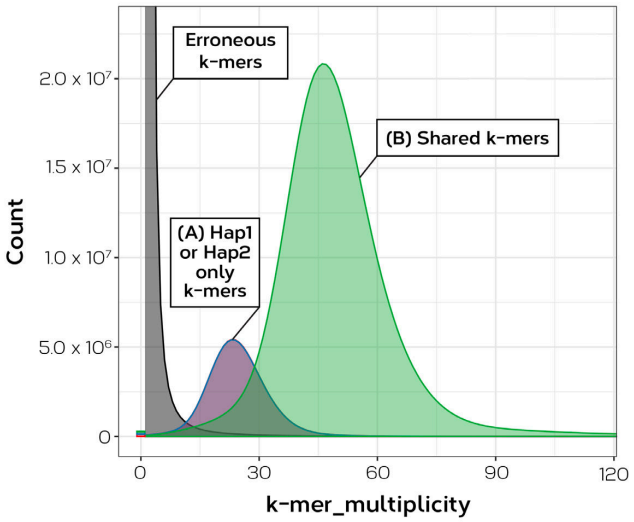

Figure 3. K-mer plot of diploid Atlantic Tuna assemblies. The heterozygous k-mers overlap, generating peak (A) where haplotype specific k-mers overlap. Peak (B) shows Hap1 and Hap2 shared k-mers.

Fully Scaffolded Haplotypes

In the highly heterozygous Atlantic Bluefin Tuna, Merqury analysis shows total assembly completeness exceeding 99%. Haplotype-specific completeness is artificially lower due to a high level of heterozygosity (Table 2). Using Merqury, assembly accuracy assessed through k-mer analysis offers an unbiased, reference-free assessment of assembly completeness and accuracy (Figure 3).

Contact matrices representing HiRise® scaffolded haplotypes of the California Brush Lizard (*Urosaurus nigricaudus*) visually illustrate the accuracy and completeness of the scaffolding process as seen by a lack of off-axis signals indicative of an assembly errors (Figure 4).

By combining high accuracy PacBio HiFi and Dovetail linked-reads, comparable levels of high

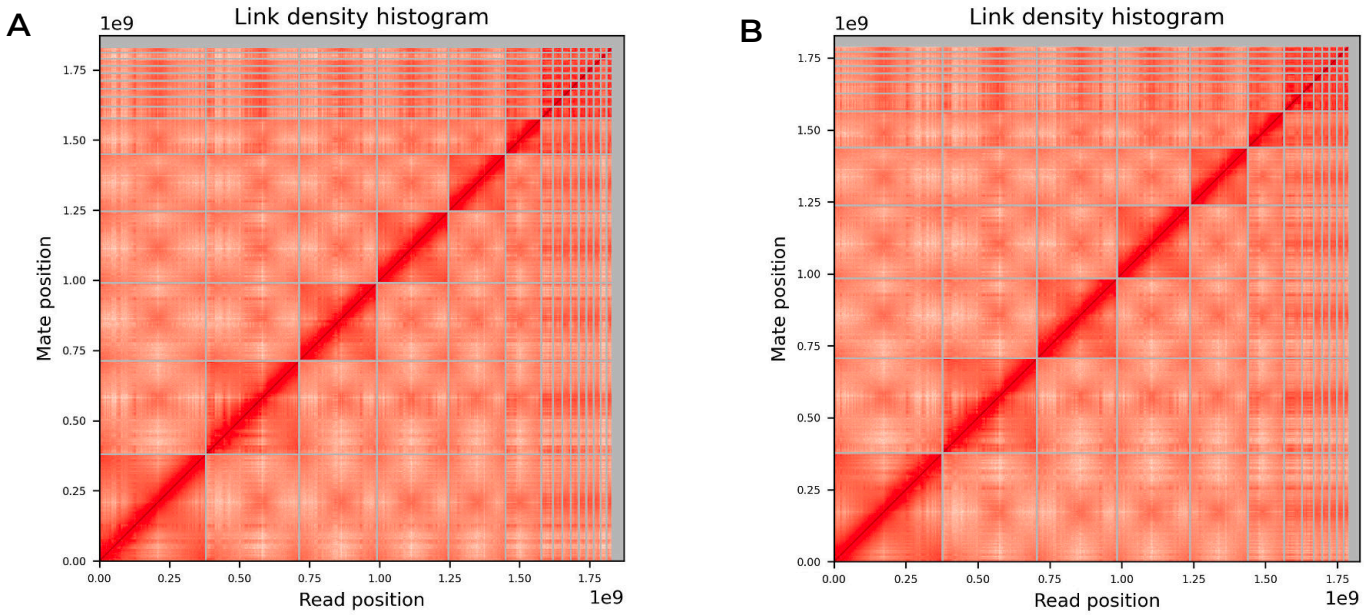

Figure 4. Contact matrix of HiRise® scaffolded haplotype assemblies of California Brush Lizard (*Urosaurus nigricaudus*) A. Contact matrix for haplotype 1. B. Contact matrix for haplotype 2.

Species	Haplotype	Scaffold (Mb) N50	L90	BUSCO (%)	Completeness (%)		Consensus QV
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. Assembly metrics for an assortment of plant and animal species. Scaffold N50 and L90, BUSCO Eukaryote database complete percentage, Merqury completeness for each haplotype and combined assemblies, and QV score across different species.

completeness are demonstrated across a broad assortment of species (Table 3).

Furthermore, scaffold N50 and L90, BUSCO eukaryote database complete percentage, Merqury completeness for haplotype and combined assemblies, and consensus quality values (QV) across the multiple species are all excellent. Of note, we find that BUSCO scores are not as accurate as the k-mer completeness assessment when assessing final quality.

Highly Accurate Diploid Assembly

In summary, Dovetail data exhibits highly uniform genome coverage when compared to restriction enzyme-based (RE) Hi-C libraries and the data output more closely resembles whole genome shotgun data.

The Shotgun-like properties of Dovetail data enables unbiased and near complete whole genome detection of heterozygous SNPs. This, coupled with the capture of long-range information, enables SNPs to be accurately phased on a chromosomal scale with very high accuracy.

References

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

Rhie, A. *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.

For more information, visit https://cantatabio.com/dovetail-genomics/services/genome-assembly/