Tech Note

Not All Genome Assemblies are Created Equal

The importance of manual correction to maximizing genome assembly quality

De novo genome assembly technologies have made great strides over the past several years driven by higher accuracy long-read sequencing coupled with restriction enzyme-free proximity ligation methods. These technologies now deliver the highest quality assemblies to date.

At Cantata Bio, we have delivered high-quality assemblies to satisfied customers for close to 2,000 species through our Dovetail® *De Novo* Genome Assembly Services.

> "We have done a few assemblies with Dovetail Genomics, and we're very happy with the results. They have proved to be real experts in achieving chromosome-scale assemblies even from species with highly repetitive genomes. Their customer service is also excellent!"
>
> Wirulda Pootakham, The National Center for Genetic Engineering and Biotechnology (BIOTEC)

Current Pitfalls

However, challenges still exist. While PacBio HiFi and Illumina short read sequencing is straightforward and accessible to most, DNA extraction, library preparation and assembly bioinformatics are frequently challenging and fraught with pitfalls for non-model eukaryotes.

- DNA extraction can be time consuming due to low yields and novel contaminants.
- Proximity ligation library protocols often require species-specific modifications.
- The output from scaffolding software pipelines is not perfect, necessitating downstream manual correction for optimal results.

The Dovetail® Scaffolding Process

Our Dovetail® *De Novo* Assembly workflow has been fully optimized across a broad spectrum of genomes (Figure 1, diploid; and 2, polyploid). Through our extensive experience with a wide array of species, our team is aware of how to avoid the pitfalls described above.

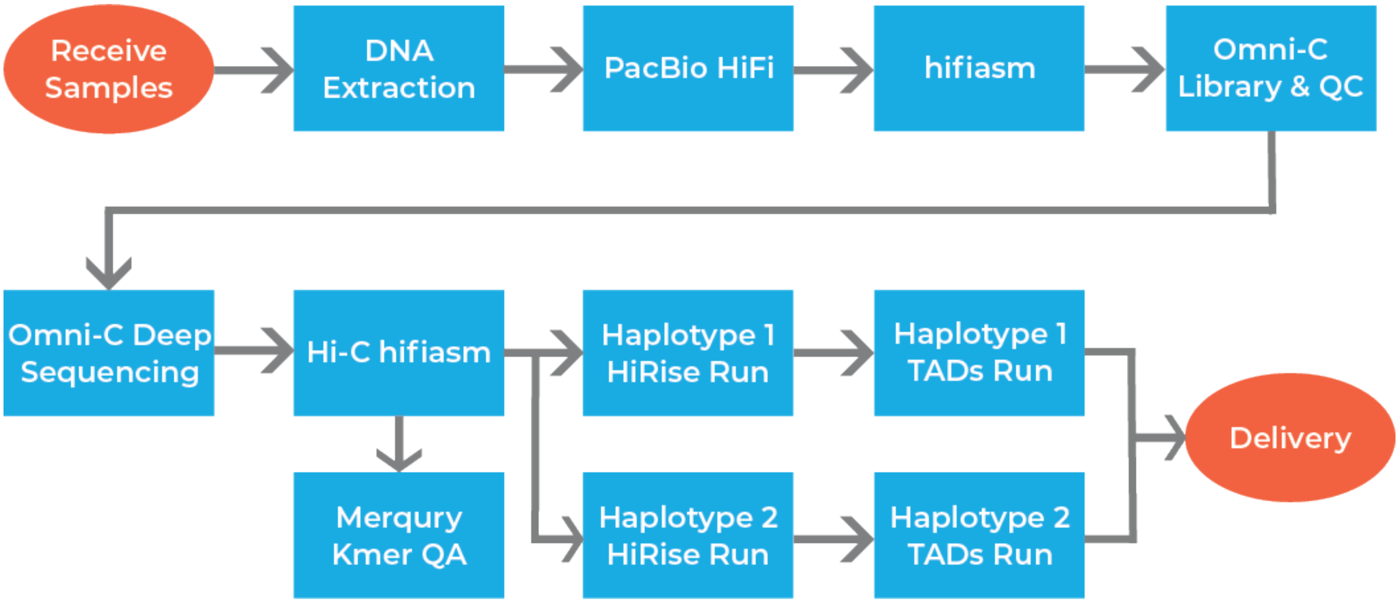

Figure 1 Genome assembly workflow for diploid genomes.

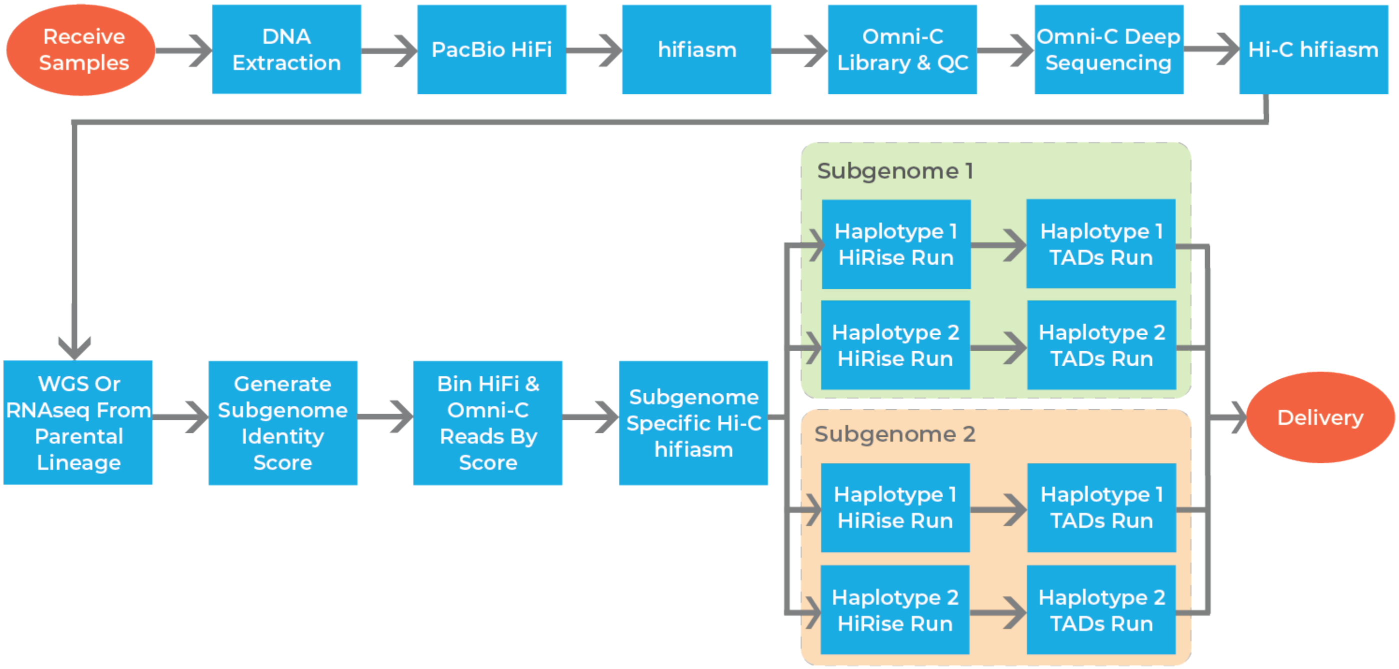

Figure 2 Genome assembly workflow for polyploid organisms.

Scaffolding starts with our proprietary Dovetail® HiRise® Software, designed specifically for use with Dovetail® Omni-C® proximity ligation data (Putnam *et al*, 2016). During the scaffolding process:

1. HiFiasm *de novo* assembly and Dovetail® Omni-C® library reads are used as input.
2. Dovetail Omni-C library sequences are aligned to the draft input assembly.
3. The separations of Dovetail® Omni-C® read pairs mapped within draft scaffolds are analyzed by the Dovetail HiRise Software to produce a likelihood model for genomic distance between read pairs.
4. The model is used to identify and break putative mis-joins, to score prospective joins, and make joins above a certain quality threshold.
5. The output assembly is manually reviewed and corrected by a dedicated and specialized bioinformatician.

It is generally accepted that subsequent manual correction is necessary to ensure optimal assembly accuracy.

> **"...no automated method to date is free from the production of errors, especially during the scaffolding stages."**
>
> Rhie *et al*, 2021

The Dovetail® Scaffolding Difference

Using Juicebox (Durand *et al.* 2016), the raw scaffolded output is reviewed using the Dovetail® Omni-C® link density plot and manual correction is performed by a dedicated bioinformatician. As manual correction is part art, and part science, thorough knowledge of the scaffolding software and extensive experience studying link density plots is required to identify and correct mis-joins in the assembly.

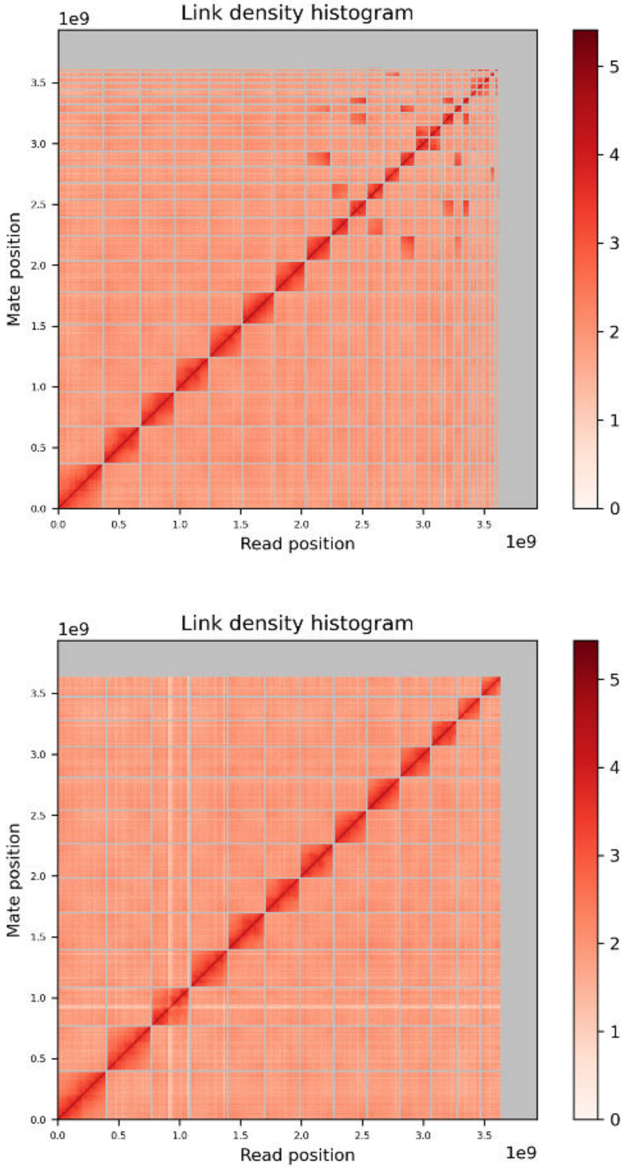

Figure 3 Examples of mis-joins, as evidenced in the link density plot on the top, corrected by the manual correction process. Final corrected assembly is displayed on the bottom.

Mis-joins may be readily evident in the link density plot, however, many issues, such as centromere-telomere inside-out errors, are less obvious to the untrained eye. Moreover, taking the appropriate corrective action when mis-joins are found is a challenge.

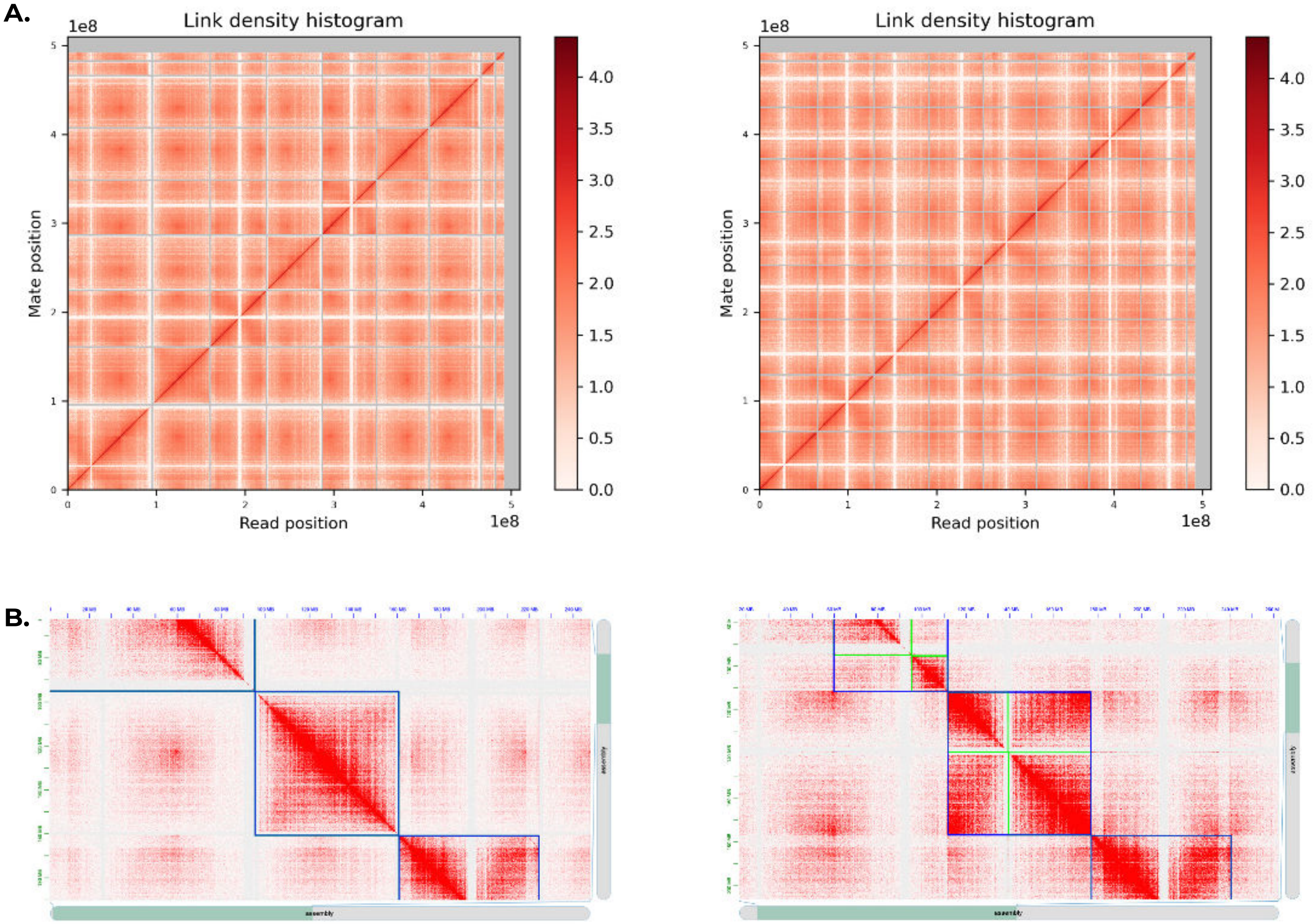

Figure 4 Examples of a centromere-telomere inside-out errors. This error type is not obvious when reviewing the link density plot. **A.** Link density plot prior to (left) and after (right) manual correction. These plots look very similar. **B.** Zooming into a 250 kb region in the #2 scaffold highlights the centromere-telomere inside-out error (left) and the correction after manual adjustment (right).

Conclusion

While *de novo* genome assembly methods have improved greatly, the process still requires significant expertise, specialized knowledge, and experience. Through our Dovetail® Genome Assembly Service, you can rest assured that you will receive the state-of-the-art, including years of density plot pattern recognition experience. Our trained eyes are being used to find anomalies in proximity ligation datasets ensuring your assembly is delivered with the highest accuracy possible.

References

1. Durand, N.C. *et al*. (2016) **Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom**. *Cell Systems*, 3(1):99-101
2. Putnam N.H. *et al*. (2016) **Chromosome-scale shotgun assembly using an in vitro method for long-range linkage**. *Genome Research*, 26(3):342-350
3. Rhie, A. *et al*. (2021) **Towards complete and error-free genome assemblies of all vertebrate species**. *Nature*, 592:737-746