Optimizations of Physical Genome Map Contiguity by *In Silico* Ligation

P Sheth[1], E Chan[2], A Hastie[1], A W Pang[1], T Anantharaman[1], Ž Džakula[1], X Zhou[1], H Sadowski[1], E Cho[1], V Hayes[2], H Cao[1]

[1]BioNano Genomics, San Diego, CA, USA; [2] Garvan Institute of Medical Research, Sydney, Australia

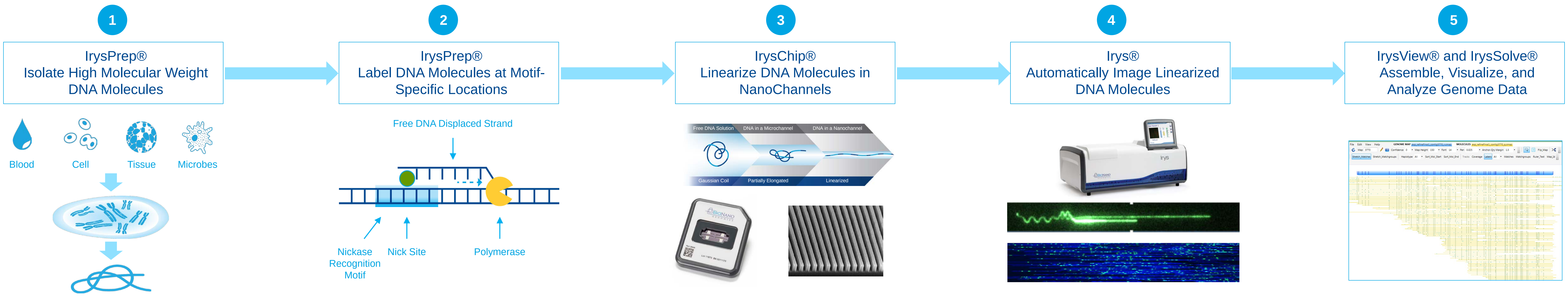
Abstract

High-quality reference genome assembly and finishing continues to be a time-consuming and cost-prohibitive process involving bioinformaticians, biologists, computation time, and manual curation. Assemblies based solely on short-read technologies are often fragmented due to structural complexities such as tandem and interspersed repetitive segments, long-range structural variations (SV), and dispersed segmental duplications. Increased contiguity is essential to understand the true structure of these complex regions. Algorithm improvements that improve automation, reduce manual curation time, and improve data quality and confidence are sorely needed.

We present a computational pipeline (fragileSiteRepair) to effectively *in silico* predict and repair (stitch) nickase-associated fragile sites, improving the overall contiguity and SV detection sensitivity of BioNano Irys® genome maps. This relatively conservative algorithm automates the prediction of fragile sites, scaffolds the genome maps across fragile sites, and applies confidence scores based on single molecule alignments. We present results of fragileSiteRepair from samples with well-characterized references such as human and Arabidopsis.

The BioNano Genomics Irys® System linearizes and molecularly barcodes long DNA molecules, yielding single molecule information contiguous up to megabase lengths, preserving long-range information. These contiguous molecules can span repeats and capture structural information often missed by sequencing platforms. The single molecule information is assembled into genome maps that can scaffold sequencing assemblies, validate the accuracy of the sequences, and anchor the adjacent sequences into the proper order and orientation, improving the quality and accuracy of sequencing data. The long genome map contiguities are further improved by bridging the breaks at fragile sites: fragile site breaks are breaks that occur when modified restriction enzymes introduce nick sites too close in proximity to each other and an unintentional double-stranded break in the DNA molecule occurs.

Methods



1 IrysPrep® Isolate High Molecular Weight DNA Molecules

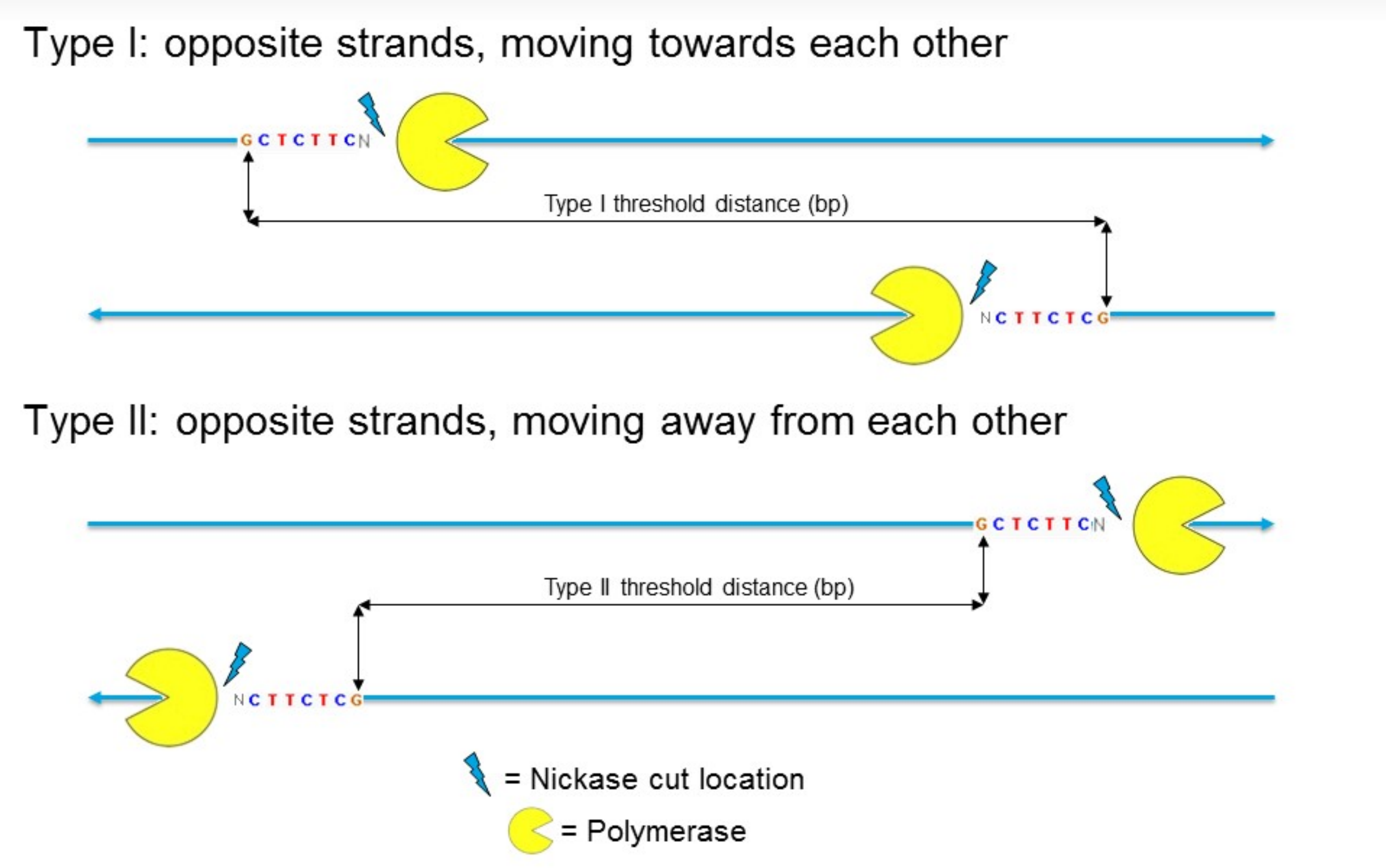
2 IrysPrep® Label DNA Molecules at Motif-Specific Locations

3 IrysChip® Linearize DNA Molecules in NanoChannels

4 Irys® Automatically Image Linearized DNA Molecules

5 IrysView® and IrysSolve® Assemble, Visualize, and Analyze Genome Data

Potential Fragile Sites using Nt.BspQI



Type I: opposite strands, moving towards each other
GCTCTTCN
Type I threshold distance (bp)
NCTTCTCG
Type II: opposite strands, moving away from each other
GCTCTTCN
Type II threshold distance (bp)
NCTTCTCG
= Nickase cut location
= Polymerase

Figure 1: Fragile site-related polymerase activity model by *in silico* prediction

Fragile Site Repair Confirmed by NGS

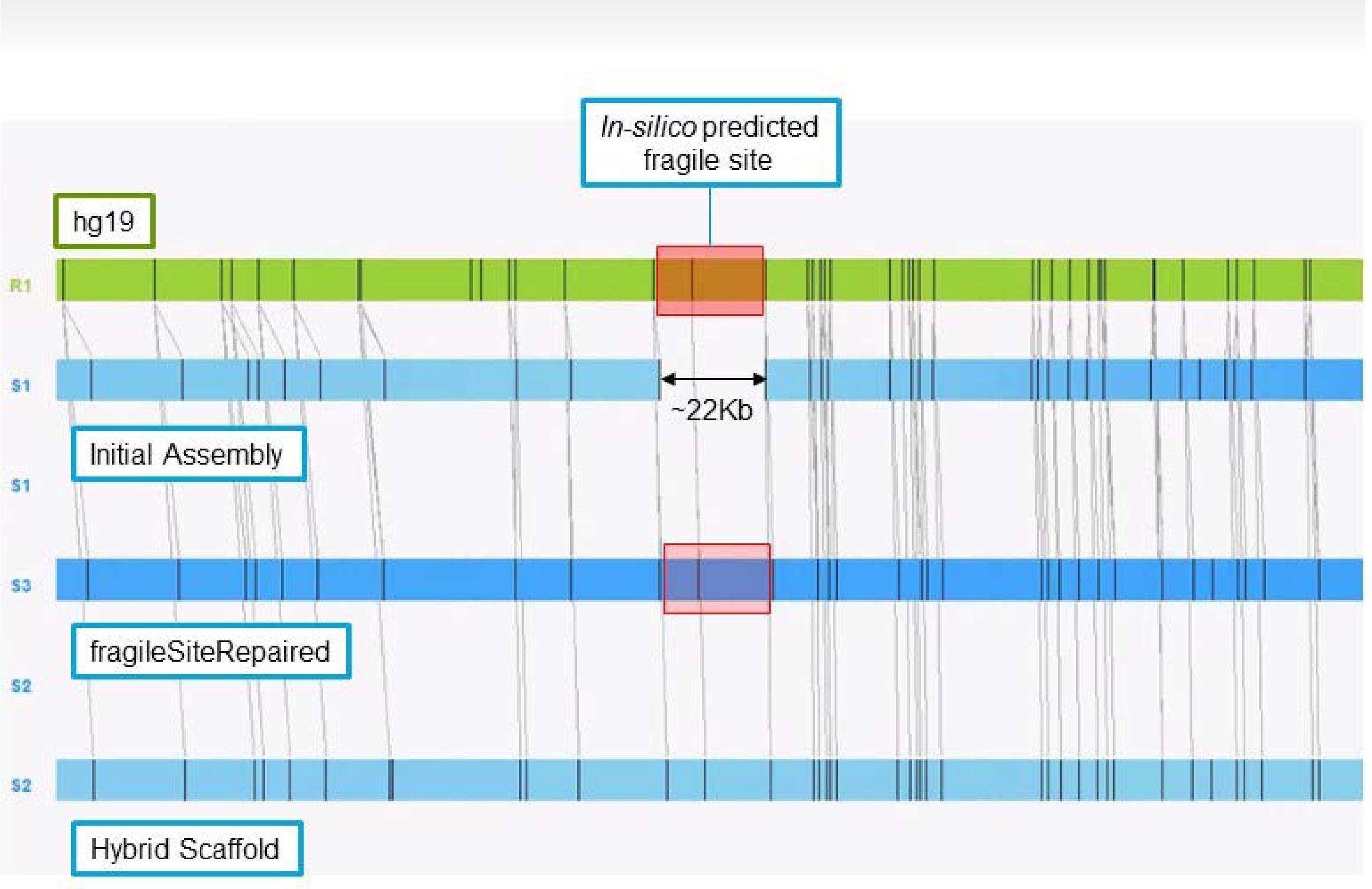


Figure 2: Example cross validation of *in silico* ligation stitch by orthogonal de-novo Hybrid Scaffold with PacBio scaffolds

Fragile Site Repair Results

		GIAB AJ Trio[1]	CEPH Trio[2]	Arabidopsis[3]
Original assembly	Average # maps	1125	1058	208
	Average N50	4.4Mb	4.4Mb	1.0Mb
Fragile site repaired	Average # maps	817	748	85
	Average N50	6.37Mb	6.38Mb	1.54Mb
	Average Compute Time[4]	17min	20min	6min
N50 increase		44%	45%	54%

[1]~96x average coverage [2]~78x average coverage [3]~222x coverage [4]48 cores,128GB RAM

Table 1: Results comparison across GIAB Ashkenazi Trio, CEPH pedigree 1463 Trio, and *Arabidopsis thaliana* samples

Conclusions

The resulting fragile site repaired genome maps are highly contiguous with 25%-100% increase in overall N50 with a minimal computational burden. The validity of repairs (stitches) was confirmed by utilizing orthogonal next-generation sequencing (NGS) data and hybrid scaffold[2] with stitch confidence scores showing a strong positive correlation with true positive stitches. Using fragileSiteRepair, we demonstrate the ability to use NGS information that may be too short in length to be used otherwise for hybrid scaffolding. Future work includes analysis of reduced BioNano and NGS coverages along with implementation of fragileSiteRepair on final assembly contiguity. FragileSiteRepair greatly enhances the ability of next-generation mapping using the Irys® System to reduce manual curation time and generate cost and time-effective near-reference quality genome assemblies with the integration of genome mapping and NGS. See also Posters: P0207, P0702, P0236, and P2078.

Reference

- [1]Kumar-Sinha et. al. 2008, Nature Reviews Cancer 8, 497-511
- [2]Pendleton, M., et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nature Methods (2015); e3454
- [3]Cao, H., et al. Rapid Detection of Structural Variation in a Human Genome using NanoChannel-based Genome Mapping Technology. Giga Science (2014); 3(December 2014): 34
- [4]Lam, E.T., et al. Genome mapping on NanoChannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303