ARIMA GENOMICS

The Role of Hi-C in Genome Assembly

Introduction

Since the completion of the Human Genome Project in 2003[1], much of the public believed that the human genome was indeed fully sequenced for the first time. However, this was not entirely true as many regions–particularly highly repetitive, non-coding regions–were not well characterized[2].

Over time, advances in next generation sequencing (NGS), the introduction of third generation sequencing, and the development of methods for sample and library preparation have allowed scientists to "fill the gaps" in the human genome. For example, Dr. Karen Miga and colleagues recently sequenced the human X chromosome from telomere to telomere (T2T) – nearly two decades after initial sequencing of the first human genome was completed in 2001[3]. This incredible feat was achieved thanks to the availability of ultra-long read nanopore technology, single-molecule high-fidelity (HiFi) sequencing technology, and chromosome-spanning connectivity information from the Arima High Coverage Hi-C kit. This process was accompanied by significant algorithmic improvements which resulted not only in improved quality of the de novo assembly but also enabled genome-wide variant phasing. The ability to connect distant variants to the same chromosome molecule promises significant advances to inherited disease research.

Advances in genome assembly techniques have spawned several global collaborations to assemble genomes of a broad range of taxa. For example, the Genome 10K Organization (G10K)[4] has initiated several projects, including the Insect5K, AgPest100, Earth Biogenome Project, Bat 1K, Bird 10K, and its flagship, the Vertebrate Genomes Project (VGP). Arima Genomics' technology is a critical component in achieving VGP's mission to build reference genomes for the ~70,000 extant vertebrate species[5]. Furthermore, the Darwin Tree of Life project is another collaboration leveraging Arima Genomics high coverage Hi-C technology between the Wellcome Sanger Institute, universities, research institutes, museums, and botanical gardens[6] aimed at sequencing all eukaryotic life in Britain and Ireland[7].

In this application note, we discuss the role that Arima High Coverage Hi-C plays in the more complete, high-quality genome assemblies for both human and non-human reference genome assemblies.

1. https://www.genome.gov/human-genome-project
2. https://www.statnews.com/2017/06/20/human-genome-not-fully-sequenced/
3. https://doi.org/10.1038/s41586-020-2547-7
4. https://genome10k.soe.ucsc.edu/about/
5. https://doi.org/10.1101/2020.05.22.110833
6. https://www.darwintreeoflife.org/partners/
7. https://www.darwintreeoflife.org

Assembling High-Quality Reference Genomes

Proper genome reconstruction is often compromised by long repetitive regions that exceed the read length of the sequencing technology. This can lead to ambiguous regions in the genome and an inability to connect the unique portions. With existing approaches, assembling high-quality reference genomes with minimal gaps requires the use of multiple sequencing methods and technologies. Depending on the size and complexity of the organism and the desired assembly accuracy, alternative approaches – for example, high-fidelity long reads supplemented with ultra-long nanopore reads – might be sufficient to sequence through long repetitive regions[8].

In their flagship methods paper, the VGP leverages high-fidelity long reads, synthetic linked reads, optical maps, and Arima-HiC to generate the raw data required for their landmark assembly. The Arima High Coverage Hi-C Kit was and continues to be an invaluable tool to produce high-quality, complete reference genomes. The Hi-C data corrects assembly errors, complements linked reads and optical maps for improved scaffolding of contigs, and provides chromosome-spanning contiguity to the assembly (Figure 1)[9, 10, 11].

The Arima High Coverage Hi-C Kit for Phased Genome Assembly

Phased genome assemblies, or haplotype-resolved genomes, provide a comprehensive picture of the genome and its complex genetic variations. Typically, due to the complexity of chromosome-scale phasing, pedigrees (or trios) had been required to resolve assemblies for distinct chromosomes. In 2013, Selvarag et al. developed an approach for chromosome-scale haplotype phasing using proximity ligation and shotgun sequencing[13].

In 2019, a method was introduced which combines single-cell DNA template strand sequencing (Strand-seq) with high-fidelity long reads to assemble a haplotype-resolved chromosomes, but this method requires active cell cultures and single-cell isolation[14]. Unfortunately, these requirements are often not a viable option for many research labs.

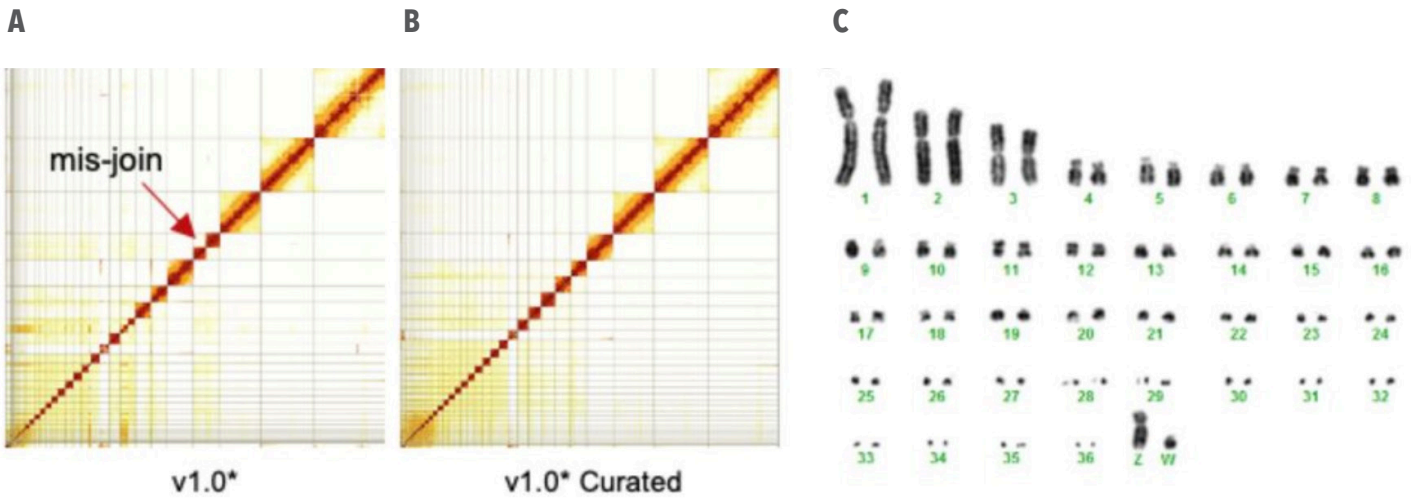

Figure 1. Error correction using manual curation. (A, B) Hi-C interaction heatmaps before and after manual curation, which identified 32 macro/micro-autosomes and ZW sex chromosomes. Grid lines indicate scaffold boundaries. Red arrow, example of a mis-joined scaffold that was corrected during curation. (C) Karyotype of the identified chromosomes (n=36+ZW), consistent with Becak et al. (Anna's Hummingbird)[12].

8. https://doi.org/10.1038/s41586-020-2547-7
9. https://doi.org/10.1101/2020.05.22.110833
10. https://vertebrategenomesproject.org/phase-one
11. https://doi.org/10.1038/s41467-020-20536-y
12. https://doi.org/10.1007/978-3-642-65751-1
13. https://doi.org/10.1038/nbt.2728
14. https://doi.org/10.1101/855049

Recently, Shilpa Garg and colleagues developed "diploid assembly" (DipAsm), which combines Arima High Coverage Hi-C long-range conformation data with high-fidelity (HiFi) long reads to assemble individual chromosomes in one day (Figure 2)[15]. These latest developments in phased genome assembly solve one of the biggest challenges in de novo assembly: most genomes are not homozygous throughout. This study was able to resolve highly polymorphic regions of biological importance, making the process more scalable. Two regions highlighted in particular are human leukocyte antigen (HLA) and killer cell immunoglobulin-like receptors (KIR), which are notoriously difficult to assemble.

The ability to phase the assembly will have profound implications for human genetics. Understanding the long-range linkage of these highly homologous but divergent regions should greatly enhance the ability of researchers to advance precision medicine.

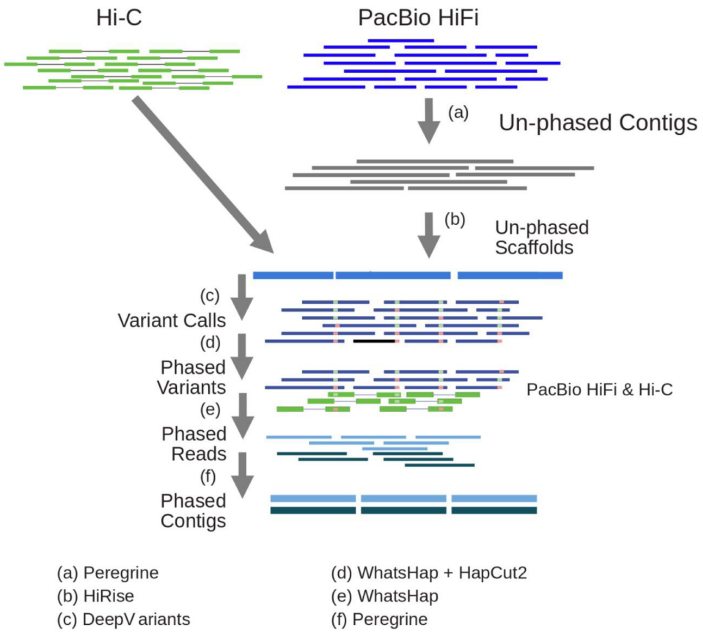

Figure 2. Outline of the phased assembly algorithm, DipAsm. Assemble HiFi reads into unphased contigs using Peregrine1; group and order contigs into scaffolds with Hi-C data using HiRise/3D-DNA (3D de novo assembly)2; map HiFi reads to scaffolds and call heterozygous SNPs using DeepVariant4; phase heterozygous SNP calls with both HiFi and Hi-C data using WhatsHap plus HapCUT26; partition reads based on their phase using WhatsHap5; assemble partitioned reads into phased contigs using Peregrine1.

15. https://doi.org/10.1038/s41587-020-0711-0

Beyond the Reference Genome

Sequence information provides only one dimension of the genome. To derive deeper insights into the characteristics of any organism, whether human, non-human, eukaryotic or plant, it is important to understand many dimensions of the genome. While the Arima High Coverage Hi-C Kit provides significant value in the construction of chromosome-spanning assemblies, the three-dimensional conformation data yields deeper insights into gene folding and, therefore, gene regulation. These insights can supplement future studies, such as mRNA sequencing, by illuminating how differences in gene folding lead to differences in expression between species, tissues, and cells.

Conclusion

The availability of high-quality reference genomes has had a profound impact on the understanding of genome function and species evolution.

As technologies continue to develop, Arima Genomics will continue to expand and refine its technologies to assemble high-quality genomes and understand their function in 3D. In summary, the Arima High Coverage Hi-C Kit supports assembly efforts by:

1. Generating chromosome-scale scaffolds
2. Gap-filling ambiguous regions
3. Correcting mis-scaffolding errors
4. Generating phased assemblies

Arima Genomics is honored to have been part of the technology development spearheaded by prestigious organizations like the Wellcome Sanger Institute, the Genome 10K Organization, and thought leaders worldwide.

ARIMA GENOMICS • 6404 Nancy Ridge Drive, San Diego, CA 92121
Office: 858.275.6463 • info@arimagenomics.com • arimagenomics.com