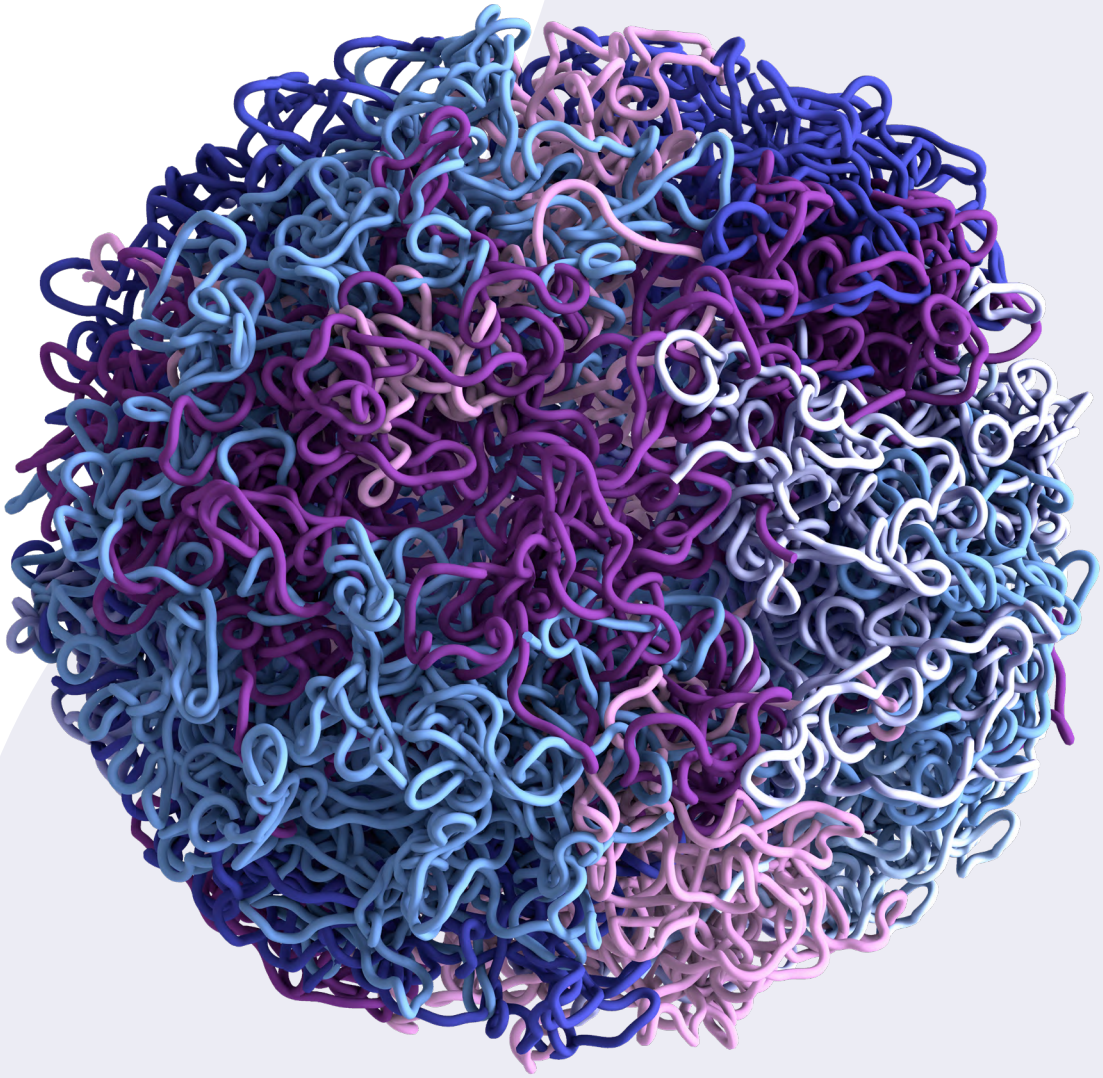
ARIMA GENOMICS

Hi-C for Genome Assembly

Because you can't get phased, chromosome-scale genome assemblies with sequencing alone.



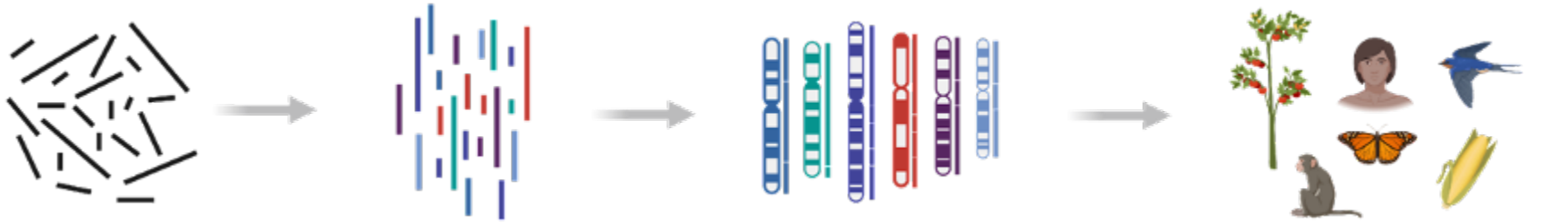
Contents

Chapter 1: Genome Assembly 101

A pile of contigs vs a complete genome

DNA sequencing technologies aren't enough to deliver platinum quality genomes on their own. Adding Hi-C data takes your assembly from a draft to a complete reference genome to drive understanding of genome biology.

Depending on the sequencing technology used, assembly algorithms typically output hundreds to thousands of unmapped contigs as an "assembly". For downstream utility, scaffolding is needed to decipher the order and location of contigs along chromosomal boundaries.



100s to 1000s of unmapped contigs

Scaffolding contigs to location in the genome

Chromosome-scale assembly

Informing conservation, agriculture, and human health

What makes a genome high quality?

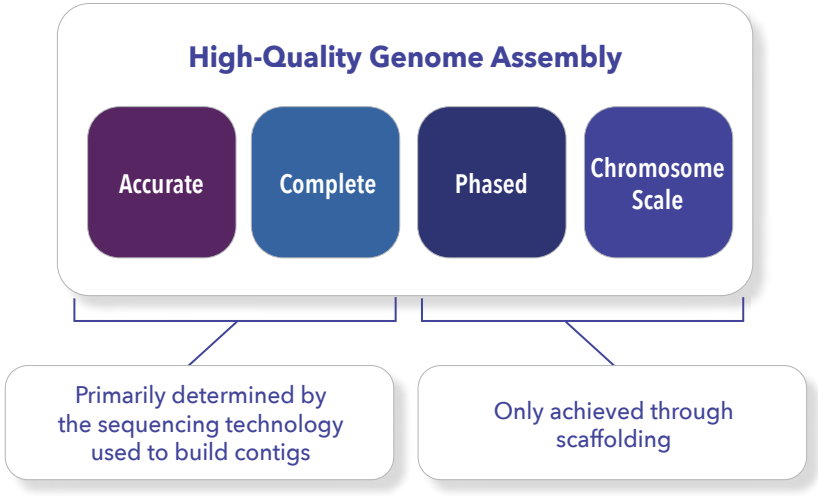
It's no surprise that the quality of a genome assembly will impact the amount and quality of insights that can be gleaned from it. So it's worth asking the question – what makes a genome assembly "high quality"?

Here we will narrow the quality question down to four main areas that separate the draft genomes from the reference-quality genomes.

Accuracy and completeness refer to accurately representing every base in the genome and completely containing all the expressed and not expressed portions of the genome. Both of these are mostly determined at the contig level by the sequencing technology used to read and then assemble the genome into contigs.

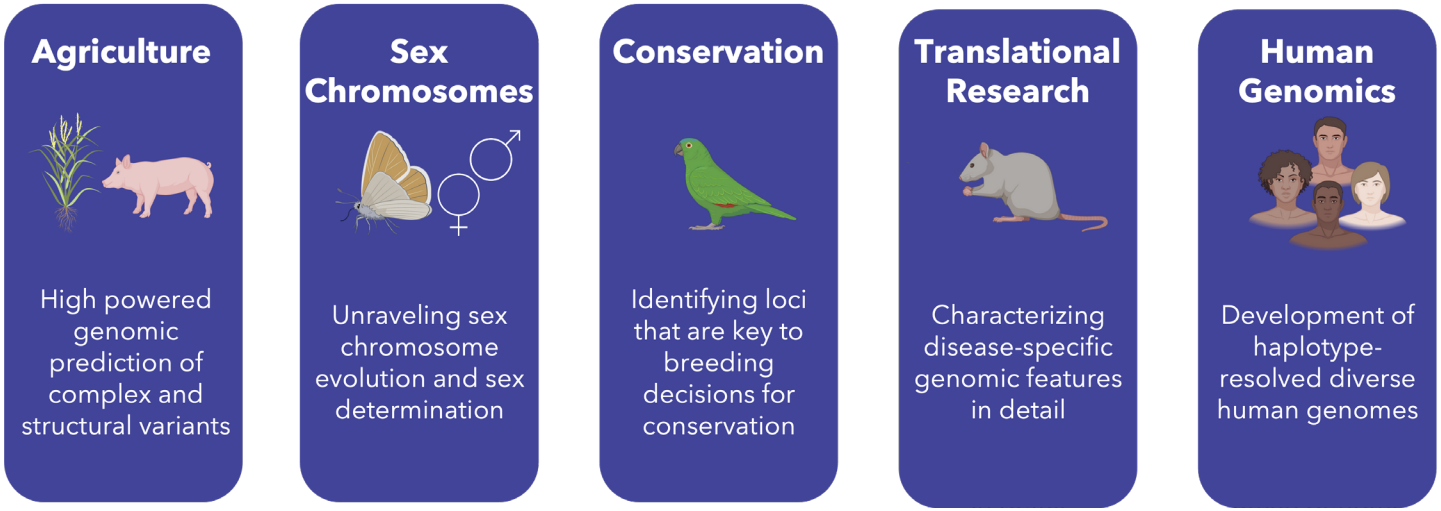
On the other side there's phasing – separating the sequences into their respective haplotypes, and contiguity, meaning there are no gaps in the known sequence across the entire chromosome (also referred to as chromosome-scale contiguity).

While DNA sequencing technologies have made big strides in achieving long, accurate contigs, they still aren't enough to deliver chromosome-scale contiguity while also phasing haplotypes. This is no more apparent than in the recent publication of the first telomere-to-telomere human genome. They used a combination of sequencing technologies and Arima Hi-C data to finally fill all the gaps left from the original human genome project.



Chapter 2: A High-Quality Genome Assembly is a Strong Foundation

To truly understand genome biology, complete genomes are needed. Complete genomes are ones that are accurate, haplotype-resolved, and chromosome-scale. It's possible that DNA sequencing technologies might provide all of the requirements for a reference genome in the future, but for now Hi-C data helps close the gap between contigs and a high-quality genome assembly, enabling groundbreaking discoveries in diverse fields of science.



Agriculture: Ramsay, L., et al. (2021) Genomic rearrangements have consequences for introgression breeding as revealed by genome assemblies of wild and cultivated lentil species. *bioRxiv*.
Sex Chromosomes: Mackintosh, A., et al. (2022) The genome sequence of the lesser marbled fritillary, Benthis ino, and evidence for a segregating neo-Z chromosome. *G3 12:6*.
Conservation: Foster, Y., et al. (2021) Genomic signatures of inbreeding in a critically endangered parrot, the kākāpō. *G3 11:11*.
Translational: Toh, H., et al. (2022) A haplotype-resolved genome assembly of the Nile rat facilitates exploration of the genetic basis of diabetes. *BMC Biology*.
Human Genomics: Nurk, S., et al. (2022) The complete sequence of a human genome. *Science 376*.

“

Arima Genomics is an integral technology partner of the G10K consortium and Phase I of the VGP project. Arima was selected for Phase I based on the quality of their data, proven by their ability to generate reproducible and high-quality data despite variability in input sample quality and quantity. The long-range genomic interactions from Arima-HiC data is an essential component of our current strategy for the generation of chromosome-spanning reference assemblies.”

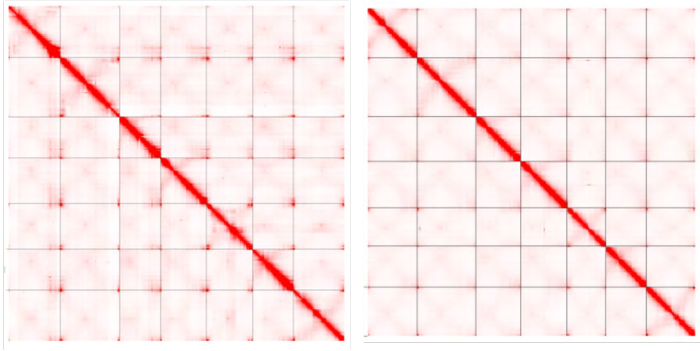
– Gene Myers, PhD
G10K Council Member &
Director, Max-Planck Institute of Molecular
Cell Biology and Genetics, Dresden

Chapter 3: Four Ways Hi-C Data Improves Genome Assemblies

1. Ordering and orienting contigs

By revealing which portions of the DNA in a chromatin structure are in close proximity to each other, interaction probabilities are used to put contigs in the correct order and orientation to give a linear representation of scaffolds in a genome.

In this example, researchers used Hi-C data to place contigs from two different lentil species into the correct order and orientation and build scaffolds to represent the chromosome-scale structure of these genomes.

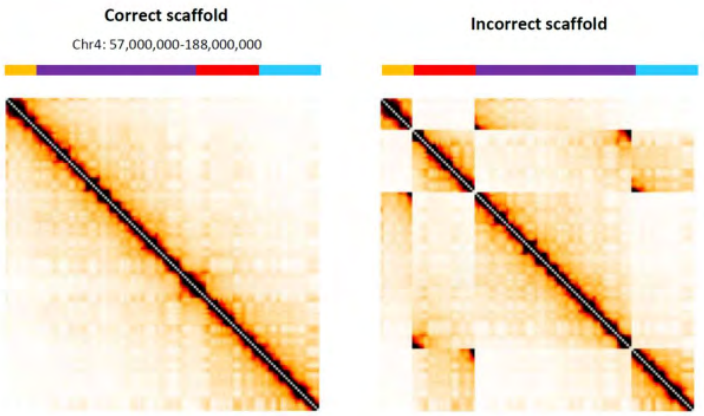


Two lentil genome assemblies with contigs scaffolded by Hi-C data from Ramsay, L., et al (2021).

2. Fixing mis-assemblies and/or identifying structural variation

By visualizing the Hi-C data mapped to sequence data and looking for abnormal (non-linear) patterns that emerge in Hi-C contact maps, structural variation and incorrect assembling of contigs can be identified and manually corrected to improve genome assemblies.

In this example, researchers showed what incorrect assembly of contigs would look like with simulated Hi-C data, demonstrating the value in visualizing genomes with Hi-C contact maps.

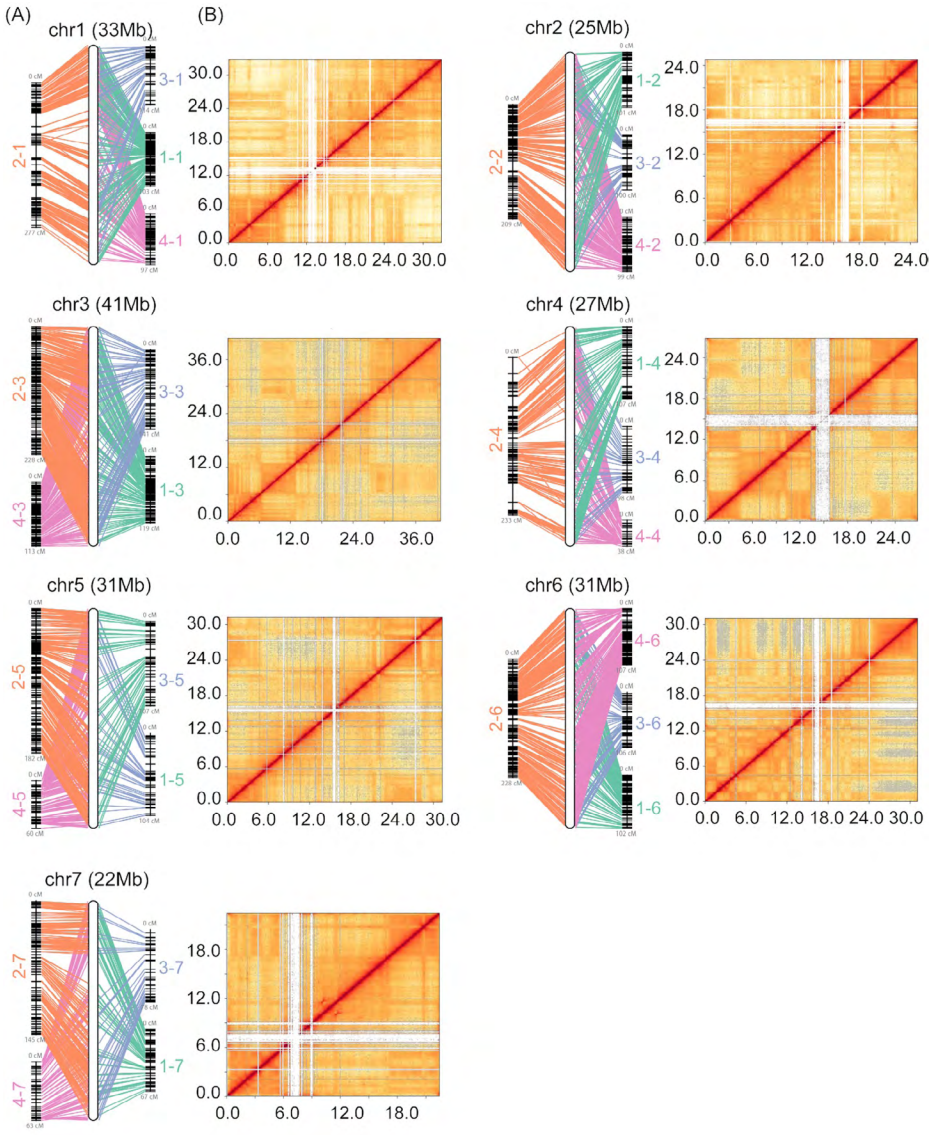


Comparison of a correctly organized Hi-C map with an incorrectly organized map from Oddes, S., et al. (2018).

3. Anchoring contigs to chromosomes

Leveraging 3D information to identify centromeric and telomeric regions, Hi-C data is used to characterize breakpoints and clearly delineate chromosomes in a genome assembly.

In this example, researchers used Hi-C data to directly link their genome sequences into the 7 pseudo-chromosomes expected for a cucumber genome.

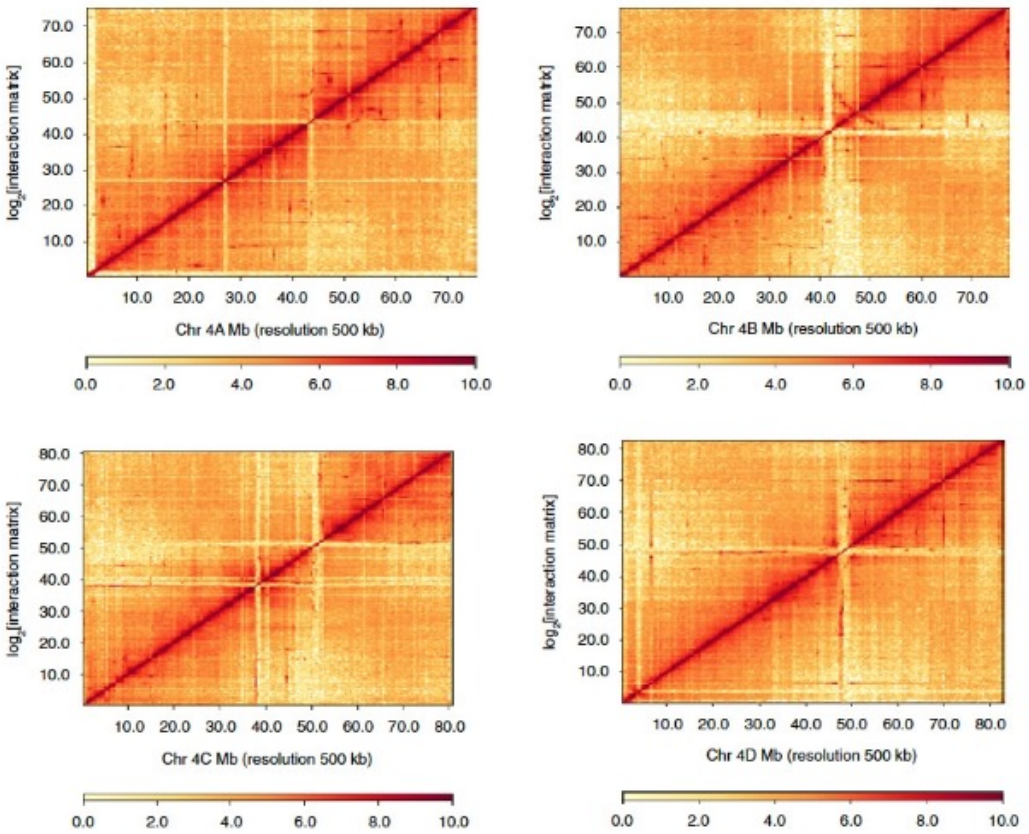


Correlation of genome assembly with genetic maps and Hi-C data for the 7 chromosomes of a cucumber genome from Li, Q., et al. (2019).

4. Phasing haplotypes

Utilizing predictable patterns of intra- vs inter-chromosomal interactions to cluster and scaffold individual haplotypes, Hi-C data helps phase contigs by clustering before scaffolding.

In this example, Hi-C data was used to identify and scaffold the four separate alleles of the wild sugarcane genome.



Hi-C contact maps of four homologous chromosomes in the wild sugarcane genome from Zhang, X., et al. (2019).

Chapter 4: Why Choose Arima Hi-C for Your Genome Assembly

Arima is the chosen Hi-C solution for some of the largest and most diverse genome sequencing consortia because of the quality, consistency, and user-friendly workflows that "just work", regardless of the species being sequenced.

- Fast and user-friendly workflow to go from sample to Hi-C library in 6 hours
- Compatible with the latest assembly and scaffolding pipelines
- Proven performance and quality you can trust while you focus on genome biology

From sample to chromosome-scale genome assembly

Hi-C Prep

Easy to follow, rapid 6-hour protocol

Library Prep

Use Arima Library Prep Module for standard or low input samples

Sequencing

200 million paired-end reads per Gb of genome

Data Analysis

Assemble and scaffold with tools of your choice

Chapter 5: Data Analysis Tools for Hi-C Interpretation

Tools for getting the most out of Hi-C data for genome assemblies

At the heart of using Hi-C data for genome assembly is probability statistics. It turns out there is a tendency for portions of the genome that are nearby in linear sequence space to also be in close proximity to each other in 3D chromatin space compared to sections of the genome that are farther apart in linear space or on different chromosomes. This enables tools to take advantage of these statistical probabilities to infer information and aid in assembling, scaffolding, and phasing genomes.

While this certainly isn't a comprehensive list, we've collected and organized a catalog of software tools that utilize Hi-C data so that you can leverage these bioinformatics workflows for the goals of your genome project.

1. QC, contact maps, and visualization

By far the most common tool encountered when browsing genome papers that utilize Hi-C data is the Juicer Tools and Juicebox software. This set of tools processes Hi-C data and outputs contact maps that have structural features annotated. Originally published in 2016 by Durand, et al., the tools have consistently been updated and maintained as Hi-C data has become more ubiquitous via the GitHub repository.

HiGlass, a web-based tool for exploring genome interaction maps, enables navigation of 2D genomic maps alongside 1D genomic tracks. Originally published in 2018 by Kerpedjiev, et al., this tool dynamically arranges views of several Hi-C datasets to support multiscale contact maps and genomic data track visualization across multiple resolutions, loci, and conditions.

Another suite of tools that include visualization is HiCExplorer, which categorizes itself a set of programs to process, normalize, analyze, and visualize Hi-C data. Originally published in 2018 by Wolff et al., it is intended for users with little bioinformatic background to perform every step in the needed analysis in one workflow.

As part of a suite of tools developed by the High Performance Assembly Group at the Wellcome Sanger Institute, PretextView, is a desktop application for viewing pretext contact maps to aid in detecting scaffolding issues as part of a broader curation strategy outlined in Howe, et al. 2021.

Tools for Hi-C Data

- Juicer Tools & Juicebox Software
- HiGlass
- HiCExplorer
- PretextView

2. Ordering, orienting, and fixing mis-assemblies

Of course, once you have checked the quality and visualized the data, the next important step is to apply the information in the Hi-C data to improve a draft genome assembly. There's no shortage of open source tools for de novo genome scaffolding, but here's a few of the most used ones we've encountered.

3D-DNA, originally published in 2017 by Dudchenko, et al., is a pipeline used to address misjoins, anchor, order, and orient the contigs of a draft assembly. After several iterations of misjoin detection and scaffold building, a series of post-processing steps are used to fix errors and split chromosomes.

SALSA, originally published in 2017 by Ghurye, et al., is an algorithm specifically for scaffolding genome assemblies built with long reads. Another iterative scaffolder, SALSA also detects and adjusts misjoins as part of the process, outputting scaffolds of each iteration as well as a final scaffold file.

YaHS, a more recently developed tool out of the Wellcome Sanger Institute and published currently as a pre-print by Zhou et al., (2022), touts a novel method for building the contact matrices which the authors indicate may improve the assembly accuracy and contiguity, and be more robust to assembly errors.

instaGRAAL, originally published in 2014 as GRAAL by Marie-Nelly, et al., uses a Markov chain Monte Carlo (MCMC) method to find and evaluate the most likely genome given a set of genome-wide contact data through a succession of various operations such as cut, insert, flip, swap, etc.

HiCAssembler, originally published in 2019 by Renschler et al., is from the same development group of HiCExplorer to take the output from that tool and then iteratively assemble the scaffolds into chromosomes.

Tools for Hi-C Data

3D-DNA | SALSA | YaHS | instaGRAAL | HiCAssembler

3. Phasing

Phasing, or the separation of the maternal and paternal haplotypes inherited by an individual, often feels like the last frontier of genome sequencing. Extremely hard to do at the whole genome level, Hi-C data is helping address this need to study human disease as well as to tease apart the polyploid genomes of the most lucrative crop species.

AllHiC, originally published in 2019 by Zhang, et al., enables the chromosome-scale assembly of separate haplotypes in polyploid species. By doing a "pruning" step to remove Hi-C signals between allelic regions, the Hi-C data can be partitioned into haplotypes for downstream scaffolding.

HaploHiC uses Hi-C reads for phasing reads of unknown parental origin. Originally published in 2021 by Lindsly, et al., this tool marks Hi-C reads as haplotype-known or -unknown based on coverage of heterozygous phased SNVs/InDels.

DipAsm is an assembly tool for efficiently generating chromosome-scale, haplotype-resolved human genome assemblies. Originally published in 2020 by Garg, et al., the DipAsm method has been shown to phase >99% of heterozygous sites to 98-88% accuracy across three publicly available human genomes.

Tools for Hi-C Data

AllHiC | HaploHiC | DipAsm

4. Bonus: Hi-C used directly in contig assembly

As the acknowledgment for Hi-C data for scaffolding and phasing has risen, so has the idea of efficiency and taking advantage of Hi-C data directly in the contig assembly pipeline, reducing the steps required to generate a chromosome-scale, phased genome assembly.

HiFiasm, originally published in 2021 by Cheng, et al., has been used across a wide range of species from humans to frogs to strawberries for the generation of mostly phased contig assemblies. The algorithm has recently been updated to include Hi-C data directly in the assembly pipeline, making it more effective for phasing and post-assembly scaffolding.

Tools for Hi-C Data

HiFiasm

Summary

Hi-C data is rich in information to help build and improve genome assemblies for species across the tree of life. Whether you need a tool for publication-quality visualization of a contact map, an assembly with contigs anchored to chromosomes, or phased haplotypes to study disease or a breeding line there are many open-source tools at your fingertips. We hope this tools guide helps you efficiently meet the goals of your genome project and we look forward to including the next generation tools as developers continue to improve and innovate Hi-C data analysis.

Two Ways to Get Started

Genome Assembly Kits

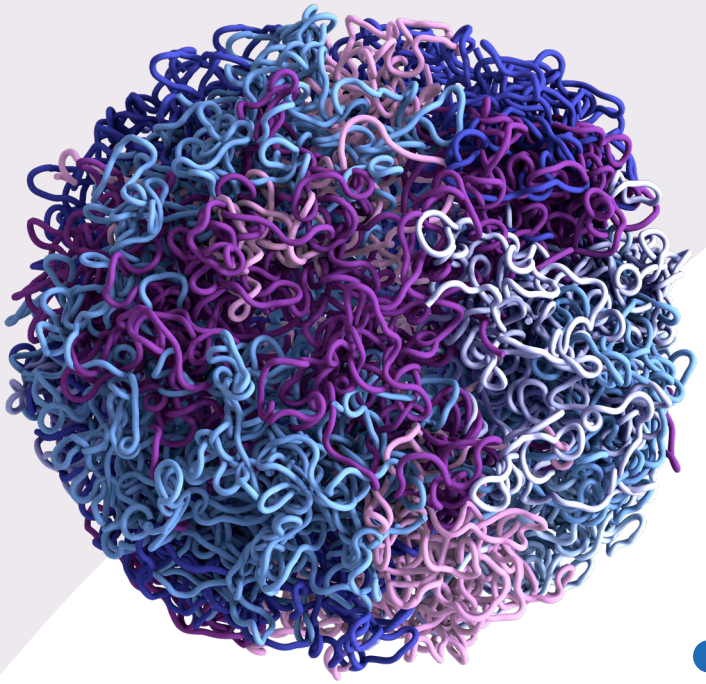
Use our sample and library preparation kits to bring chromosome-scale assemblies to your lab.

Learn About Kits

Genome Assembly Services

Let our scientists share their expertise in sample prep, library construction, and bioinformatics.

Explore Services



ARIMA GENOMICS

arimagenomics.com | team@arimagenomics.com

EB-001-012323