SCIENTIST SPOTLIGHT

ARIMA GENOMICS

CELEBRATING SCIENCE & SCIENTISTS

SHILPA GARG, PH.D.

DEPARTMENT OF GENETICS
HARVARD MEDICAL SCHOOL
BOSTON, MA, USA

RESEARCH SNAPSHOT

Research Area	Genome assembly
Species/Sample Type	Human genomic DNA
Arima Product	Arima High Coverage Hi-C kit
Application/Workflow	Phased Genome Assembly

CHROMOSOME-SCALE, HAPLOTYPE-RESOLVED ASSEMBLY OF HUMAN GENOMES

Humans are diploid organisms, typically containing two copies of every chromosome. However, most genome assemblies are represented as haploid and miss heterozygous sequences. To accurately represent diploid genomes, genome assemblies are "phased" to achieve haplotype-level resolution. No one has been able to achieve chromosome-long phasing without using the parent genomes, which are not always available. To overcome this, diploid assembly (DipAsm) combines long-read assembly and Hi-C-based phasing for single individuals to generate a chromosome-scale phased assembly within one day.

Garg S., et al. Nature Biotechnology. 2020. doi: 10.1038/s41587-020-0711-0

RESEARCH QUESTION

Can we accurately produce phased assembly for a single individual?

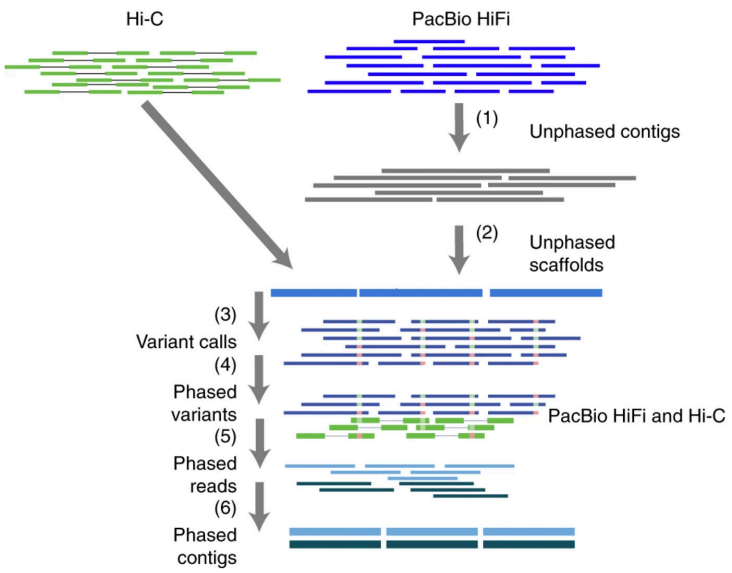
HOW DID ARIMA GENOMICS MAKE A DIFFERENCE?

"The Arima team is easy to work with and very collaborative. The technical support really made the difference in our project."

SCIENTIST SPOTLIGHT

EXPERIMENT OVERVIEW

- Convert genomic DNA from four genomes into a SMRTbell™ (PacBio) library
- Sequence on PacBio Sequel System to yield long high-fidelity (HiFi) reads
- Hi-C sequencing with the Arima High Coverage Hi-C kit
- Phased sequence assembly
    - Scaffold an unphased Peregrine[1] assembly with 3D-DNA[2] or HiRise[3], Call small variants with DeepVariant[4]
    - Phase variants with WhatsHap[5] and HapCUT2[6]
    - Partition the reads and assemble each partition independently with Peregrine again
- Evaluate variant calling accuracy



REFERENCES:

1. Chin C-S, Khalak A. Human Genome Assembly in 100 Minutes. bioRxiv. 2019:705616.
2. Dudchenko O, Batra SS, Omer AD, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92-95.
3. Putnam NH, O'Connell BL, Stites JC, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Research. 2016;26(3):342-350.
4. Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. Nature biotechnology. 2018;36(10):983-987.
5. Martin M, Patterson M, Garg S, et al. WhatsHap: fast and accurate read-based phasing. bioRxiv. 2016:085050.
6. Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Research. 2017;27(5):801-812.

Outline of the phased assembly algorithm, DipAsm.
Assemble HiFi reads into unphased contigs using Peregrine[1]; 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 DeepVariant[4]; phase heterozygous SNP calls with both HiFi and Hi-C data using WhatsHap plus HapCUT2[6]; partition reads based on their phase using WhatsHap[5]; assemble partitioned reads into phased contigs using Peregrine[1].

ACCOMPLISHMENTS / RESULTS / FUTURE DIRECTIONS

DipAsm accurately produces chromosome-long phased assembly using data from PacBio HiFi and Arima High-Coverage Hi-C. In contrast to trio binning, this method does not use pedigree data and can phase *de novo* mutations. The Hi-C portion of the method renders it easier to use and more widely adoptable than alternative methods like Strand-seq.

The *de novo* method is a milestone since generating an assembly without a reference sequence furthers the goals of the Human Genome Reference Project, which aims for unbiased characterization of human genome diversity. Sequencing polymorphic regions at high resolution will also advance personalized, precision medicine.