# SCIENTIST SPOTLIGHT

ARIMA GENOMICS

***CELEBRATING SCIENCE & SCIENTISTS***


## KAREN MIGA, PH.D.

**UC SANTA CRUZ GENOMICS INSTITUTE**
**UNIVERSITY OF CALIFORNIA**
**SANTA CRUZ, CA**


## RESEARCH SNAPSHOT

| | |
|---|---|
| **Research Area** | Satellite DNA biology and Genome Assemblies |
| **Species/Sample Type** | human CHM13hTERT cell line (CHM13) |
| **Arima Product** | Arima High Coverage Hi-C kit |
| **Application/Workflow** | HiC |

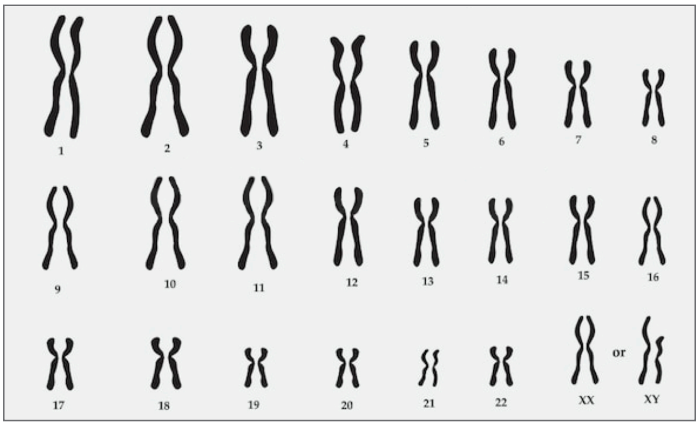
## TELOMERE-TO-TELOMERE ASSEMBLY OF A COMPLETE HUMAN X CHROMOSOME

Although the first human genome sequence was considered complete in 2003, many gaps remain. This work showcases the first end-to-end assembly of a human chromosome and presents a workflow for improved reference genomes with a multi-platform approach using high-coverage ultra-long nanopore sequencing in addition to other long read technologies such as PacBio's continuous long read (CLR), PacBio high-fidelity (HiFi) sequencing, and chromatin conformation capture. The Arima High Coverage Hi-C kit was used as an orthogonal method to confirm expected 3D folding structures of the human X chromosome.


Miga KH, et al. Nature. 2020. doi: **10.1038/s41586-020-2547-7**


## RESEARCH QUESTIONS

How can we resolve the gaps in the human genome?




## HOW DID ARIMA GENOMICS MAKE A DIFFERENCE?

*"We were able to use the Arima-HiC data to provide a secondary orthogonal analysis to see if we were getting the assembly structurally correct… the Arima-HiC data supported previous work demonstrating that the DXZ4 element serves as a TAD boundary. These biological insights and cross-validation analyses were incredibly useful for our team."*

# SCIENTIST SPOTLIGHT

ARIMA GENOMICS

## EXPERIMENT OVERVIEW:

### RECONSTRUCTING THE GENOME

- DNA from CHM13 cells was extracted and prepared for nanopore sequencing.
- Sequenced on 98 MinION flow cells.
- Combined 39X of the ultra-long Nanopore reads with 70X coverage of previously generated PacBio data and assembled the CHM13 genome using Canu.
- Assembly was then iteratively polished by each technology in order of longest to shortest read lengths: Nanopore, PacBio, 10X Genomics Linked-Reads.
- Putative mis-assemblies were identified via analysis of independent linked-read sequencing (10X Genomics) and optical mapping (Bionano Genomics) data and the initial contigs broken at regions of low mapping coverage.
- The corrected contigs were then ordered and oriented relative to one another using the optical map and assigned to chromosomes using the human reference genome.

### FINISHING AND VALIDATION OF THE HUMAN X CHROMOSOME

- Large (>100kb), nearly identical, segmental duplications were manually resolved by identifying ultra-long reads that completely spanned the repeats and were anchored on either side.
- Assembly quality improvements of these difficult regions were evaluated by mapping an orthogonal set of PacBio high-fidelity (HiFi) long reads generated from CHM13 and assessing read-depth over informative single nucleotide variant differences and validated via ddPCR (Bio-rad Laboratories).
- To assemble the human X centromere, a catalog of structural and single nucleotide variants were constructed across a centromeric satellite array (DXZ1) to uniquely tile ultra-long Nanopore reads. This assembly was improved and benchmarked using an automated satellite array assembly method, centroFlye[1]. The resulting centromere assembly was validated using restriction profiles, ddPCR, PacBio HiFi data, and tools designed for tandem repeat evaluation[2].
- To maximize base call accuracy in the final assembly, two rounds of polishing was carried out using each technology: Oxford Nanopore, then PacBio, and finally 10X Genomics Linked-Reads.
- The final polished assembly was further validated with the Arima High Coverage Hi-C kit, confirming the presence of two large superdomains partitioned at the microsatellite repeat DXZ4.

## RESULTS AND FUTURE DIRECTIONS

The manually finished X chromosome assembly is complete, gapless, and estimated to be at least 99.99% accurate.

With a multiplatform approach, the team was able to resolve difficult repetitive regions and confirm TADs using the Arima High Coverage Hi-C kit. The Telomere-to-Telomere (T2T) Consortia continues to work on resolving gaps throughout the human genome, such as those in chromosomes 1, 9, and 16, which are more difficult
because they have huge satellite arrays.

They are also working on automating assemblies of repetitive regions. A primary goal is to build and streamline essential cloud-based protocols and sequencing technologies within the next four years to enable availability of comprehensive, genomic-based, precision medicine.

1. Bzikadze AV, Pevzner PA. Automated assembly of centromeres from ultra-long error-prone reads. Nature biotechnology. 2020;38(11):1309-16.
2. Mikheenko A, Bzikadze AV, Gurevich A, Miga KH, Pevzner PA. TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. Bioinformatics. 2020;36(Supplement_1):i75-i83.