Leveraging constellation mapped read technology for variant detection in complex, medically relevant genes

Katie Larkin[1], Yueyao Gao[1], Michelle Cipicchio[1], Mark Fleharty[1], Victoria Popic[1], Marina DiStefano[1], Sean Hofherr[1], Niall Lennon[1]

[1]Broad Clinical Labs, Broad Institute of MIT and Harvard, Cambridge MA, USA

Challenging genomic regions limit short-read WGS

Whole genome sequencing (WGS) has transformed rare disease diagnostics by enabling rapid and comprehensive detection of pathogenic variants.

However, certain clinically relevant genomic regions remain inaccessible via standard short-read WGS (srWGS) due to:

- **Segmental duplications**
- **Paralogous genes and their pseudogenes**
- **Tandem Repeats**

These regions can contain disease-causing variants that short-read sequencing alone cannot reliably detect.

Clinical labs must use alternative or ancillary assays (MLPA, Long-range PCR, etc) to resolve variants, which is costly and time consuming.

Real-world impact of a paralogous gene: PMS2

Lynch syndrome is an inherited colorectal cancer predisposition caused by pathogenic variants in DNA mismatch repair (MMR) genes (*MLH, MSH2, MSH6, PMS2*).

Pathogenic variants in PMS2 are estimated to account for anywhere from 10-35% of Lynch syndrome cases depending on the study[1], but are frequently underdiagnosed due to high sequence similarity to its pseudogene, *PMS2CL*.

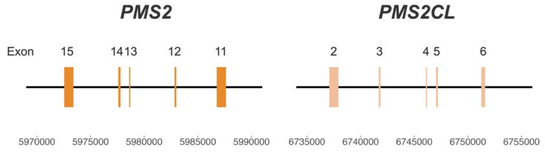

Figure 1. High homology region between PMS2, a Lynch syndrome gene, and its pseudogene, PMS2CL using coordinates from hg38.

Current clinical laboratory practice:

Initial NGS result

Variants identified in the *PMS2/PMS2CL* region cannot be immediately assigned to the correct locus

Multi-step confirmatory testing

- **Long-range PCR -** targeted amplification followed by Sanger sequencing
- **MLPA** - Used if necessary to confirm copy number or large rearrangements

Reporting

- If confirmed, variants in the homologous region of *PMS2* can be returned

MLPA testing is sometimes *still* not high-resolution enough to resolve certain cases, leading to an inconclusive result.

A potential solution

Direct sequencing of native DNA on NovaSeq X Plus

Constellation is a **library-free** sequencing approach that enables genomic DNA to be directly loaded onto a NovaSeq X (NVX) flowcell.

During sequencing, spacial proximity information between neighboring fragments is captured, allowing accurate read mapping even in regions with high sequence similarity.

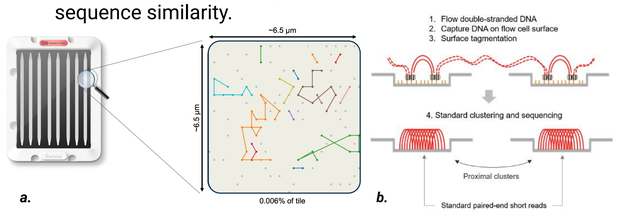

Figure 2. a. Illustration of linked reads as they appear on 0.006% of a flowcell tileb. Overview of the workflow demonstrating proximity information from nearby clusters originating from the same template molecules

Constellation's performance is driven by three key innovations:

1. **Native DNA input** - eliminates fragmentation and amplification prior to sequencing, preserving long-range genomic context
2. **Proximity-based mapping** - reads originating from adjacent locations on the flowcell are used to infer long-range haplotypes
3. **Enhanced alignment algorithms** - specialized computational methods, **like multi-region joint detection (MRJD)** reconstructs sequence contiguity and phase from these spatial clusters

Together, these features enable accurate variant calling and phasing in regions that were previously inaccessible to srWGS.

When benchmarked against the T2T truth set, calling from constellation data is equivalent to or slightly better than srWGS.

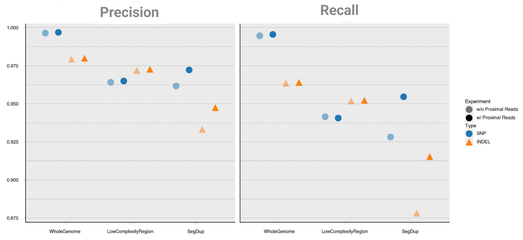

Figure 3. HG002 benchmarking against the T2T v1.1 truth set

What did we do?

To assess the ability of constellation to recover coverage and resolve variants in difficult genomic regions, we reanalyzed a set of 9 clinical samples with known variants, previously analyzed with srWGS, which failed to detect or phase variants in clinically relevant genes including *PMS2*, *PMS2CL*, and *SMN1*.

- For this follow up analysis, DNA was re-sequenced on NVX using constellation chemistry.
- Reads were processed using DRAGEN with the MRJD caller

This analysis focused on both coverage recovery and detection of copy number variation in duplicated regions

Results

Constellation recovered callable territory in difficult to map genes.

	w/o Proximal Reads	w/ Proximal Reads
% Callable *PMS2* Bases	74.03%	**86.43%**

Table 1. Across 40 genomes, constellation boosted the percentage of callable bases in PMS2

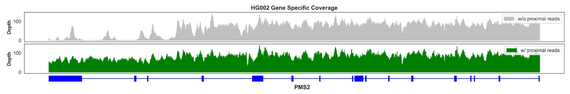

Figure 4. Coverage in of PMS2 in an HG002 genome, with and without proximal reads.

In a next key step, MRJD jointly analyzes reads, using proximity information, across all potential origins, and outputs a phased genotype.

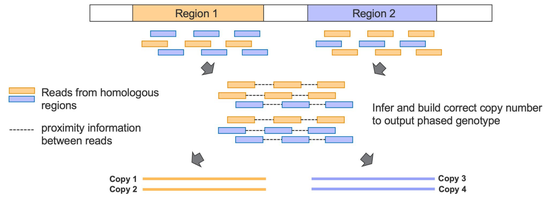

Figure 5. The MRJD workflow.

				srWGS		Constellation	
Sample	Gene of Interest	Event Type	Variant	Variant Called	Variant Phased	Variant Called	Variant Phased
1	*PMS2*	DEL	chr7:5783926-6055303	X	X	✓	✓
2	*PMS2CL*	DUP	chr7:6707807-6751888	X	X	✓	✓
3	*PMS2CL*	SNP	chr7:5977596, G>A	✓	X	✓	✓
4	*SMN1*	DEL	1 copy	✓	X	✓	✓
5	*SMN1*	DEL	1 copy	✓	X	✓	✓
6	*SMN1*	DEL	1 copy	✓	X	✓	✓
7	*SMN1*	DEL	1 copy	✓	X	✓	✓
8	*SMN1*	DEL	0 copy	✓	X	✓	✓
9	*SMN1*	DEL	0 copy	✓	X	✓	✓

Table 2. Nine previously sequenced samples were able to be detected by and phased by constellation where both were not possible with standard short-read sequencing.

Visualizing *PMS2* variants results in samples reanalysed using constellation

Sample 1: Constellation found a previously unidentified 271kb DEL encompassing *PMS2*

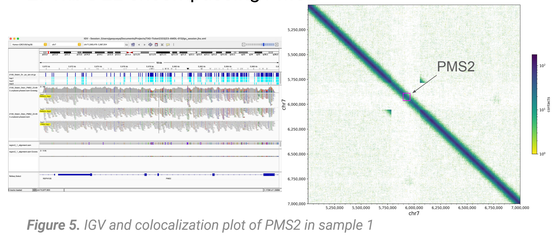

Figure 5. IGV and colocalization plot of PMS2 in sample 1

Sample 2: A variant that previously mapped to *PMS2* was shown instead to have a 44kb DUP encompassing *PMS2CL* that was not detected by MLPA

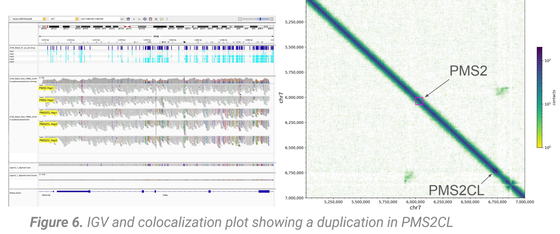

Figure 6. IGV and colocalization plot showing a duplication in PMS2CL

Constellation is a powerful tool

Constellation expands diagnostic yield by resolving variants in regions inaccessible to srWGS.

Eliminates reliance on costly, time-intensive ancillary assays.

Supports scalable rare disease testing while maintaining WGS as a single front-line test.

Advantages of constellation:

- No library preparation - native DNA can be sequenced directly
- Improved variant detection in high-homology and duplicated regions
- Accurate haplotype phasing
- Scalable - easily integrated into existing NVX workflows and DRAGEN pipelines

This approach moves clinical genomics closer to a single-test, whole-genome solution that captures the full spectrum of medically relevant variation.

References

1. Andini, K. D. et al. PMS2-associated Lynch syndrome: Past, present and future. Front Oncol 13, 1127329 (2023).

2. Overcoming the challenges in PMS2 high-homology regions for improved detection of pathogenic variants associated with Lynch syndrome. https://www.illumina.com/content/illumina-marketing/amr/en_US/science/genomics-research/articles/PMS2-small-variant-detection.html.

3. Introducing constellation mapped read technology. https://www.illumina.com/content/illumina-marketing/amr/en_US/science/genomics-research/articles/constellation-mapped-read-technology.html.

Acknowledgements

Special thank you to Yueyao Gao, who performed this analysis, and to the Illumina constellation team, who have supported us throughout this work.

Contact

Katie Larkin
klarkin@broadinstitute.org

Broad Clinical Labs
genomics@broadinstitute.org

broadclinicallabs.org