Overcoming Variant Calling Challenges in PMS2 High Homology Regions for Improved Lynch Syndrome Diagnosis using Next Generation Sequencing

Shunhua Han[1], Vitor Ounchic[1], Varun Jain[1], Pavana Anur[2], Francisco M. De La Vega[2], James Han[1]

[1]Illumina, Inc., San Diego, CA.
[2]Tempus Labs, Inc., Chicago, IL.

Summary

- Pathogenetic small variants in *PMS2* can cause lynch syndrome, an autosomal dominant hereditary cancer predisposition syndrome[1]
- Variant calling in *PMS2* gene exons 12-15 is confounded by the presence of high homology pseudogene *PMS2CL*[2]
- We propose a new de novo small variant calling method in paralogous regions that includes two modes. The unique mode that offers >83% precision and >65% recall in *PMS2*, and the high sensitivity mode that enables >98% recall on finding variant alleles in either *PMS2* or *PMS2CL*

Variant calling in *PMS2* exons 12-15 is challenging due to high homology between *PMS2* and *PMS2CL*

PMS2 exons 12-15 regions have low MAPQ due to high homology with a pseudogene *PMS2CL*. Frequent gene conversions in this region also call the accuracy of high MAPQ reads in exons 13-14 into question (Figure 1).

Existing variant calling methods discard low MAPQ reads and therefore have low performance.

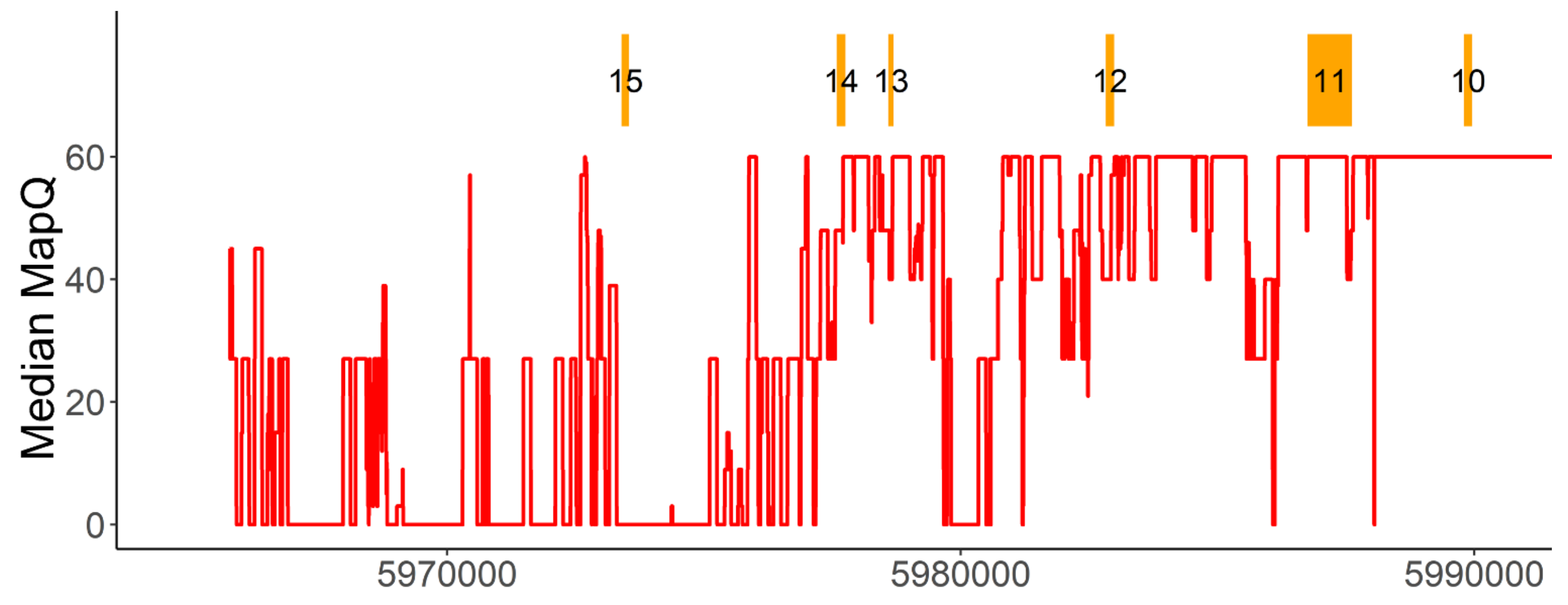
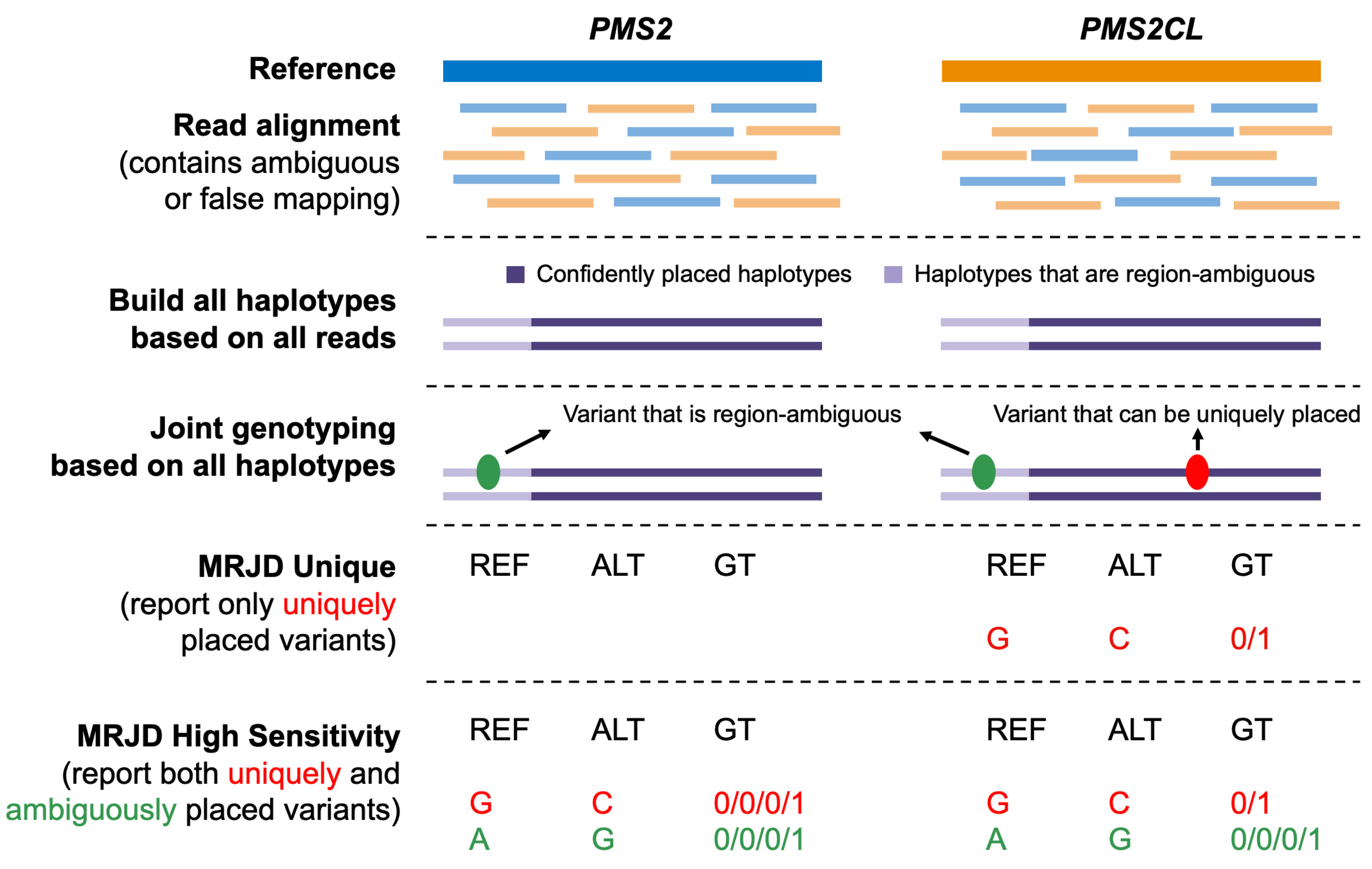


Figure 1. Median MapQ in the PMS2 exon 10-15 region. The median MapQs were computed from WGS data for 2504 unrelated samples from the 1000 Genomes Project (1kGP). X axis represents position in chr7 in 1kbp unit.

A novel WGS-based variant calling method for small variant detection in paralogous regions

We developed a novel method called MRJD (Multi-Region Joint Detection) designed to detect small variants in paralogous regions. Instead of genotyping per region and discarding reads with ambiguous alignment, the new method jointly genotypes all paralogous regions using all reads.

Based on whether the haplotype can be confidently placed, the variant can be uniquely placed or denoted as "region-ambiguous". Because of this, MRJD includes two modes. The unique mode that only reports uniquely placed variants, and the high sensitivity mode that reports both uniquely placed and region-ambiguous variants.



PMS2 *PMS2CL*

Reference

Read alignment (contains ambiguous or false mapping)

■ Confidently placed haplotypes ■ Haplotypes that are region-ambiguous

Build all haplotypes based on all reads

Joint genotyping based on all haplotypes

Variant that is region-ambiguous

Variant that can be uniquely placed

MRJD Unique (report only uniquely placed variants)

	REF	ALT	GT	REF	ALT	GT
MRJD Unique				G	C	0/1
MRJD High Sensitivity	G A	C G	0/0/0/1 0/0/0/1	G A	C G	0/1 0/0/0/1

MRJD High Sensitivity (report both uniquely and ambiguously placed variants)

Figure 2. The Multi-Region Joint Detection (MRJD) method workflow. The method involves building all haplotypes for all paralogous regions using all reads, placing haplotypes in each paralogous region based on prior knowledge, genotyping all paralogous regions jointly based on placed haplotypes, and identifying variants.

Two workflows in clinically relevant paralogous genes based on applications

In this work, we explored two workflows for variant detection in challenging and clinically significant paralogous genes. The first workflow using MRJD Unique mode prioritizes overall accuracy by maximizing the F1 score, while the second using MRJD High Sensitivity mode emphasizes sensitivity.

Research or clinical labs that do not require reflex testing typically prefer the accuracy-driven workflow. In contrast, clinical labs that prioritize identifying all potential clinical cases and possess the capacity for confirmatory testing (e.g., long-range PCR-based tests) tend to favor the sensitivity-driven approach.

The MRJD caller outperforms existing methods and offers high recall in PMS2

We benchmarked MRJD against an orthogonal long-range PCR NGS approach[2] on 150 samples from the public Illumina Polaris diversity panel (>30x coverage using the Illumina NovaSeq 6000 system with PCR-Free library prep and 2x150bp reads). MRJD Unique mode outperforms existing method in the *PMS2* high homology region, particularly on INDEL calling (Figure 3).

Compared to the MRJD Unique mode that prioritizes accuracy, the MRJD High Sensitivity mode enables substantially higher recall at the cost of precision. When comparing with results from merged orthogonal dataset (Figure 4A), the MRJD High Sensitivity mode demonstrates both high recall and precision, indicating low spurious call rate (Figure 4B).

Assessment on independently sequenced 18 cell line samples (~50X coverage using the Illumina NovaSeq 6000 system with PCR-Free library prep and 2x150bp reads) yielded similar performance results compared to the public dataset.

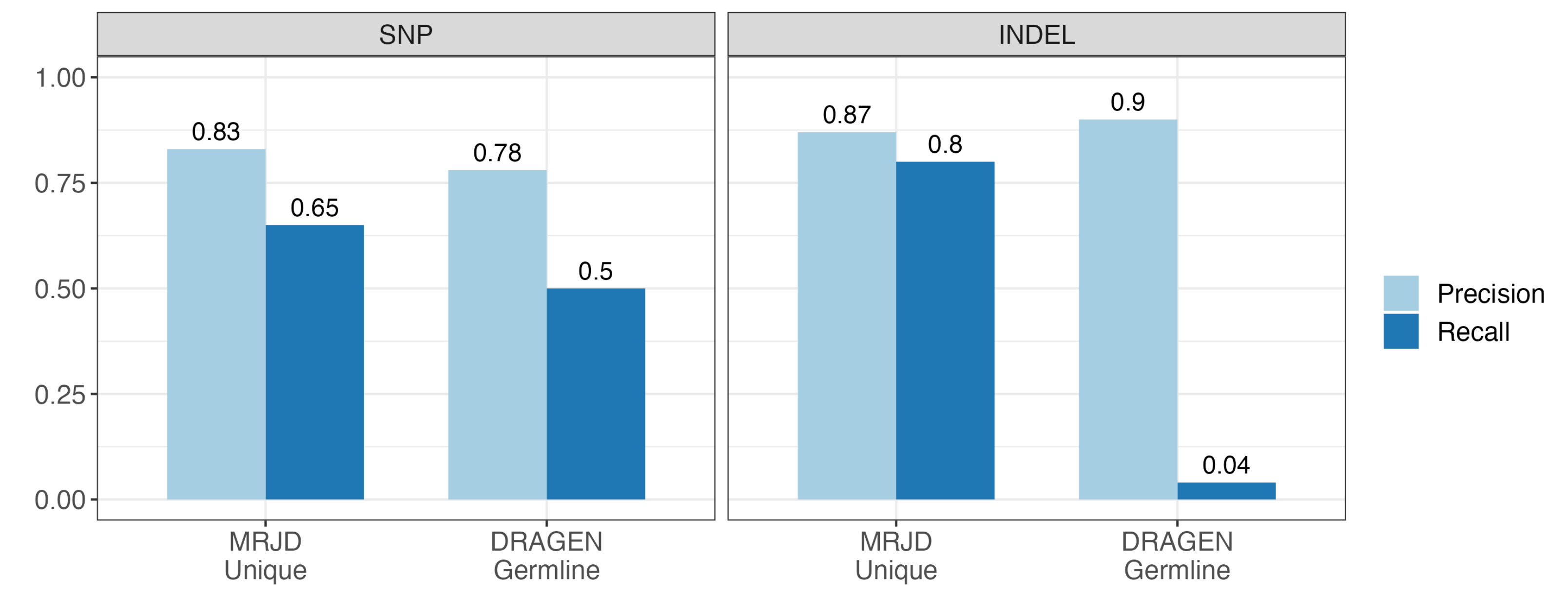


Figure 3. Aggregated SNP and INDEL performance between MRJD Unique and DRAGEN Germline Small Variant Caller on 150 samples from Illumina Polaris diversity panel. hap.py GT exact mode is used to generate benchmark statistics. Homopolymer regions are excluded due to low orthogonal dataset quality.

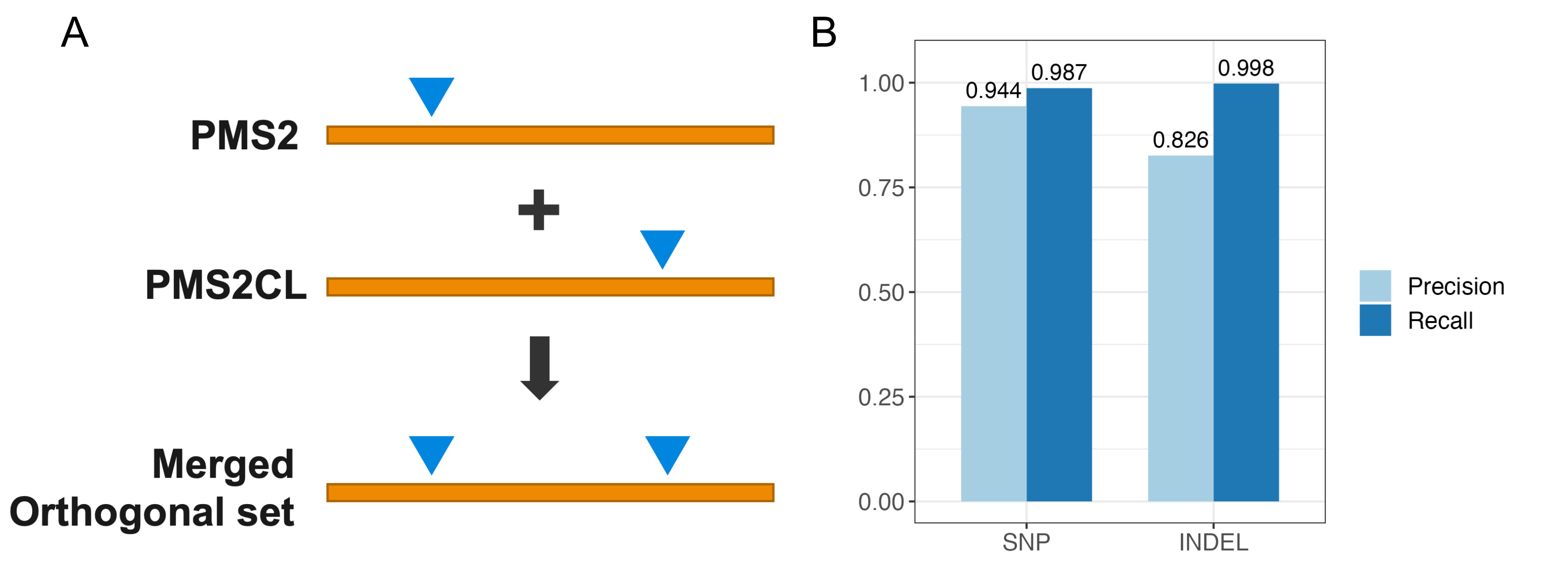


Figure 4. (A) We generated merged orthogonal set from LR-PCR data by merging variants from PMS2 and PMS2CL into one paralogous region (in this case PMS2). (B) We then compared MRJD high sensitivity mode output with merged orthogonal set. RTG Tools ploidy-squash mode is used to generate benchmark statistics. Homopolymer regions are excluded due to low orthogonal dataset quality.

References

1. Lynch et al. "Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications." *Clin. Genet.* (2009)
2. van der Klift et al. "Comprehensive mutation analysis of PMS2 in a large cohort of probands suspected of Lynch syndrome or constitutional mismatch repair deficiency syndrome." *Hum. Mutat.* (2016)
3. Gould et al. "Detecting clinically actionable variants in the 3′ exons of PMS2 via a reflex workflow based on equivalent hybrid capture of the gene and its pseudogene." *BMC Med. Genom.* (2018)

Contact

The software will be available in a future release of DRAGEN. Please contact ffg-info@illumina.com to request early access to the DRAGEN MRJD caller.

For Research Use Only. Not for use in diagnostic procedures.

illumina®