

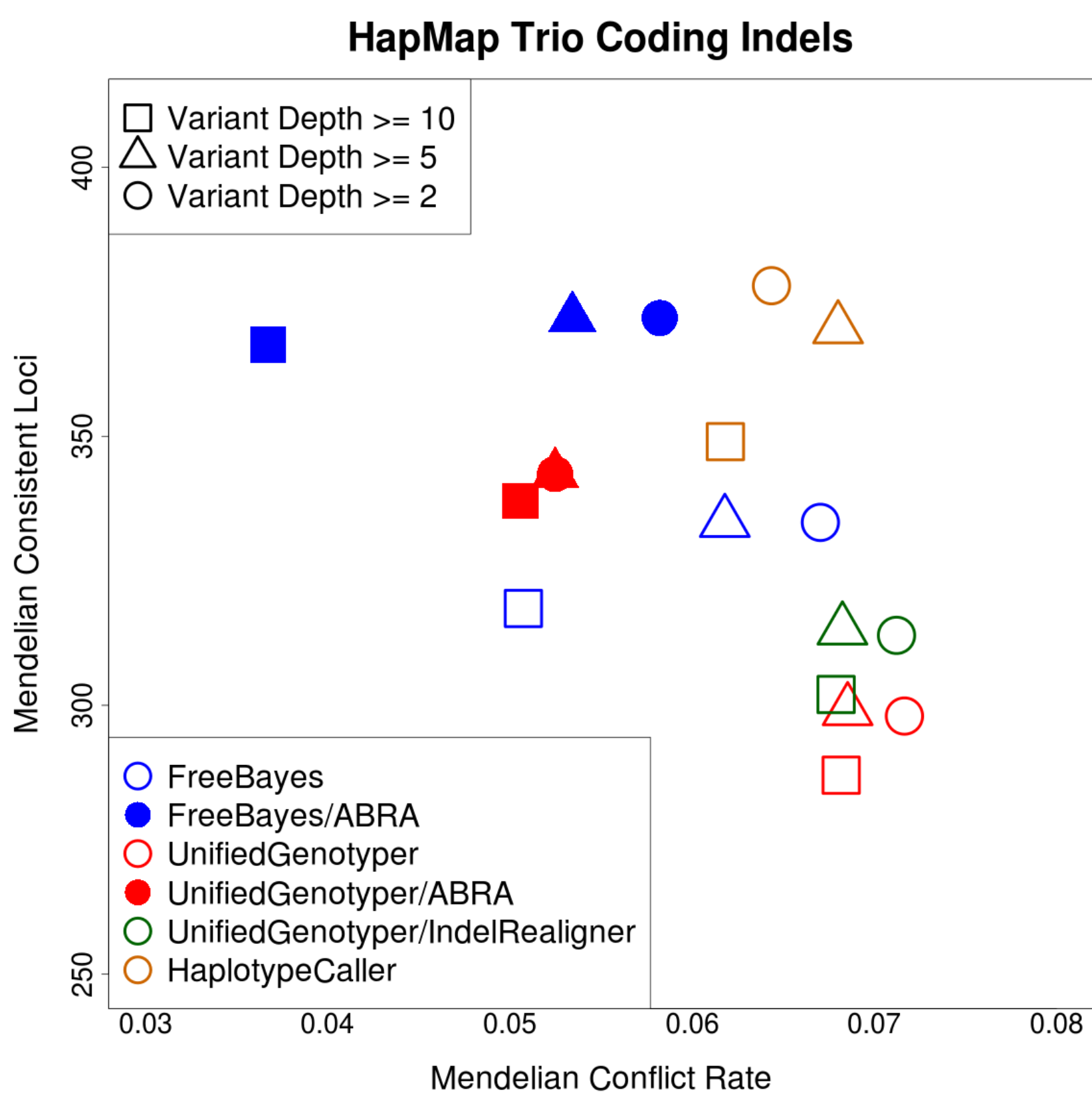
Lisle E. Mose¹, Matthew D. Wilkerson^{1,2}, D. Neil Hayes^{1,3}, Charles M. Perou^{1,2,4} and Joel S. Parker^{1,2}

Background

Variant detection from next generation sequencing (NGS) data is an increasingly vital aspect of disease diagnosis, treatment and research. Commonly used NGS variant analysis tools generally rely upon accurately mapped short reads in order to identify somatic variants and germline genotypes. Existing NGS read mappers have difficulty accurately mapping short reads containing complex variation (i.e. more than a single base change), thus making identification of such variants difficult or impossible. Insertions and deletions (indels) in particular have been an area of great difficulty. (Mills et al., 2011; O’Rawe et al., 2013). Indels are frequent and can have substantial impact on function which makes their detection all the more imperative (1000 Genomes Project Consortium, 2010; Mills et al., 2011)

Here, we present ABRA, an Assembly Based Re-Aligner, which uses an efficient and flexible localized de novo assembly followed by global realignment to more accurately re-map reads. This results in enhanced performance for indel detection as well as improved accuracy in variant allele frequency estimation.

HapMap Trio

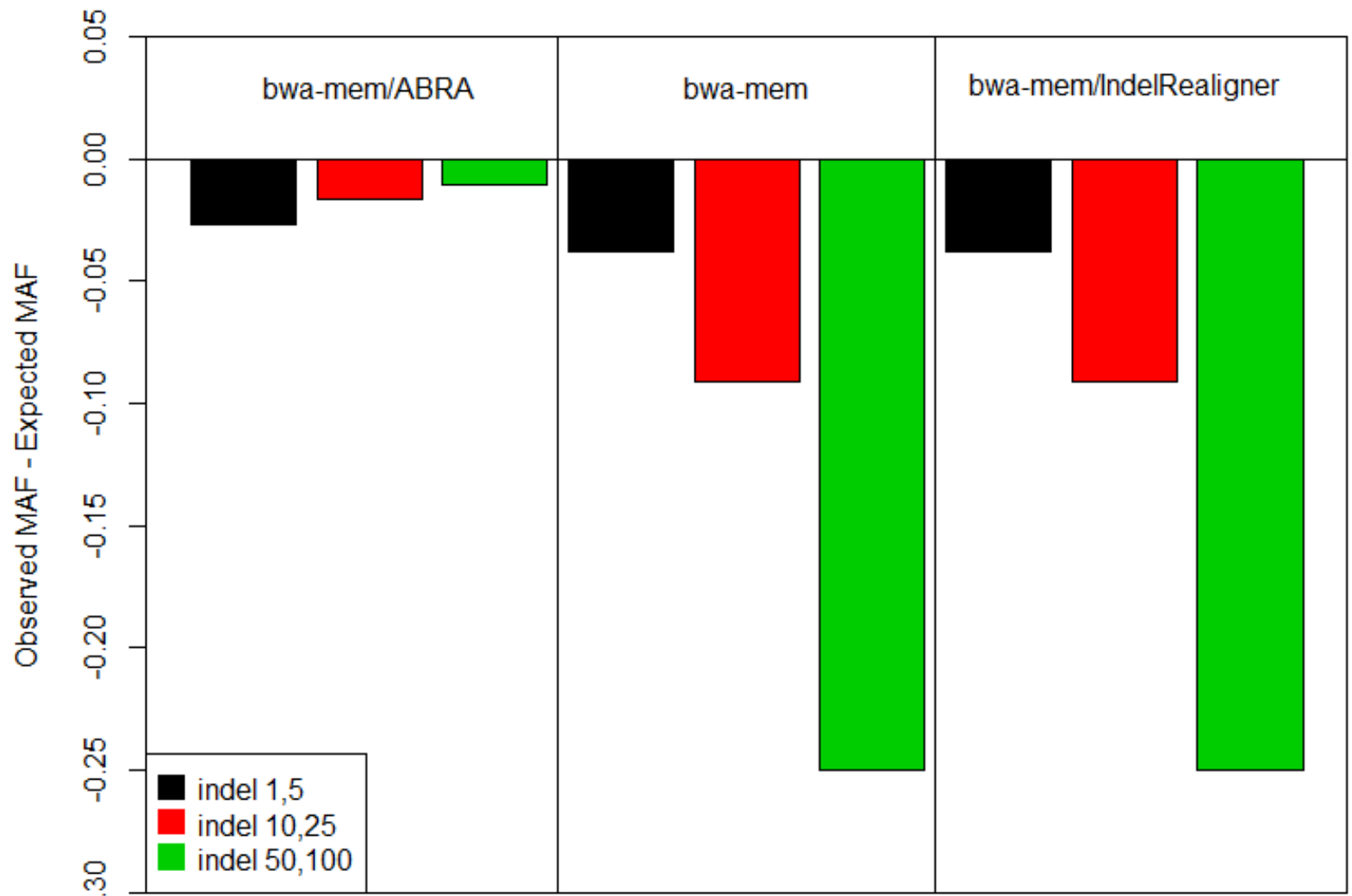
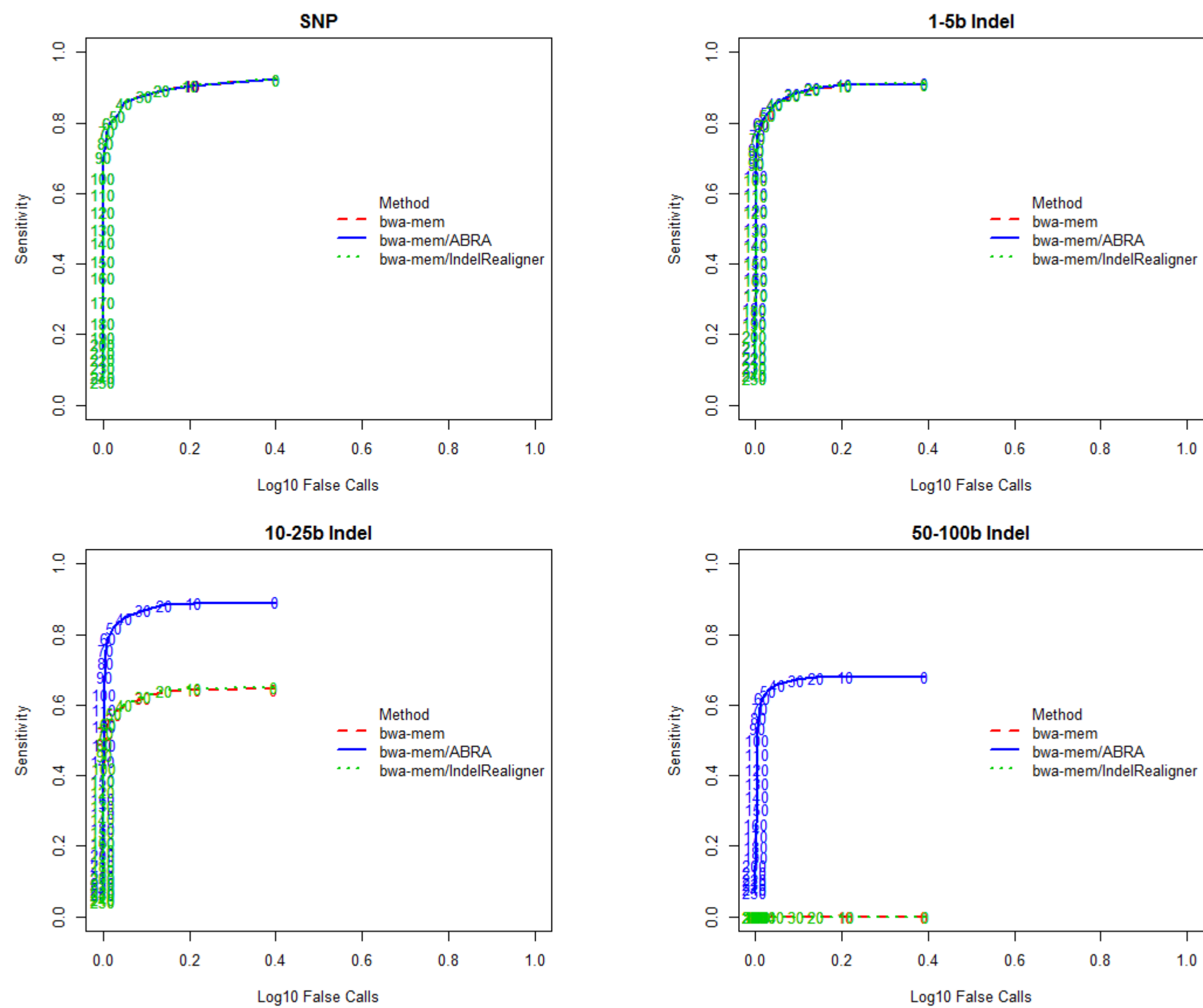


ABRA was applied to exome target regions of a CEPH Hapmap trio of three individuals sequenced to 50x as part of the Illumina Platinum Genomes project. Variants were called with and without ABRA using FreeBayes and UnifiedGenotyper. The GATK’s Haplotype Caller was used to call variants without ABRA and the GATK’s IndelRealigner was applied to UnifiedGenotyper input.

Variant calling against ABRA realigned BAMs enables an increase in the number of Mendelian Consistent Indel Loci detected while simultaneously decreasing the Mendelian Conflict Rate. Pre/post ABRA concordance for Mendelian consistent SNP loci is greater than 99%.

Variant Simulation

We assess ABRA’s impact on somatic variant detection by simulating over 10,000 somatic variants in over 500 samples. SNPs, insertions ranging in length from 1 to 50 bases and deletions from 1 to 100 bases were simulated in single end reads of length 100 bp. Variant calling is performed with Strelka. ABRA enables improved performance in detection of longer indels as well as increased accuracy in Mutant Allele Frequency (MAF) estimation.



TCGA Breast Cohort

Germline variant counts “Pre” and “Post” ABRA

Indel Length	1-29	30-99	100-1,721
Pre-ABRA deletion	100,602	86	0
Post-ABRA deletion	112,729	5,919	4,259
Pre-ABRA insertion	72,831	0	0
Post-ABRA insertion	83,172	1,904	0

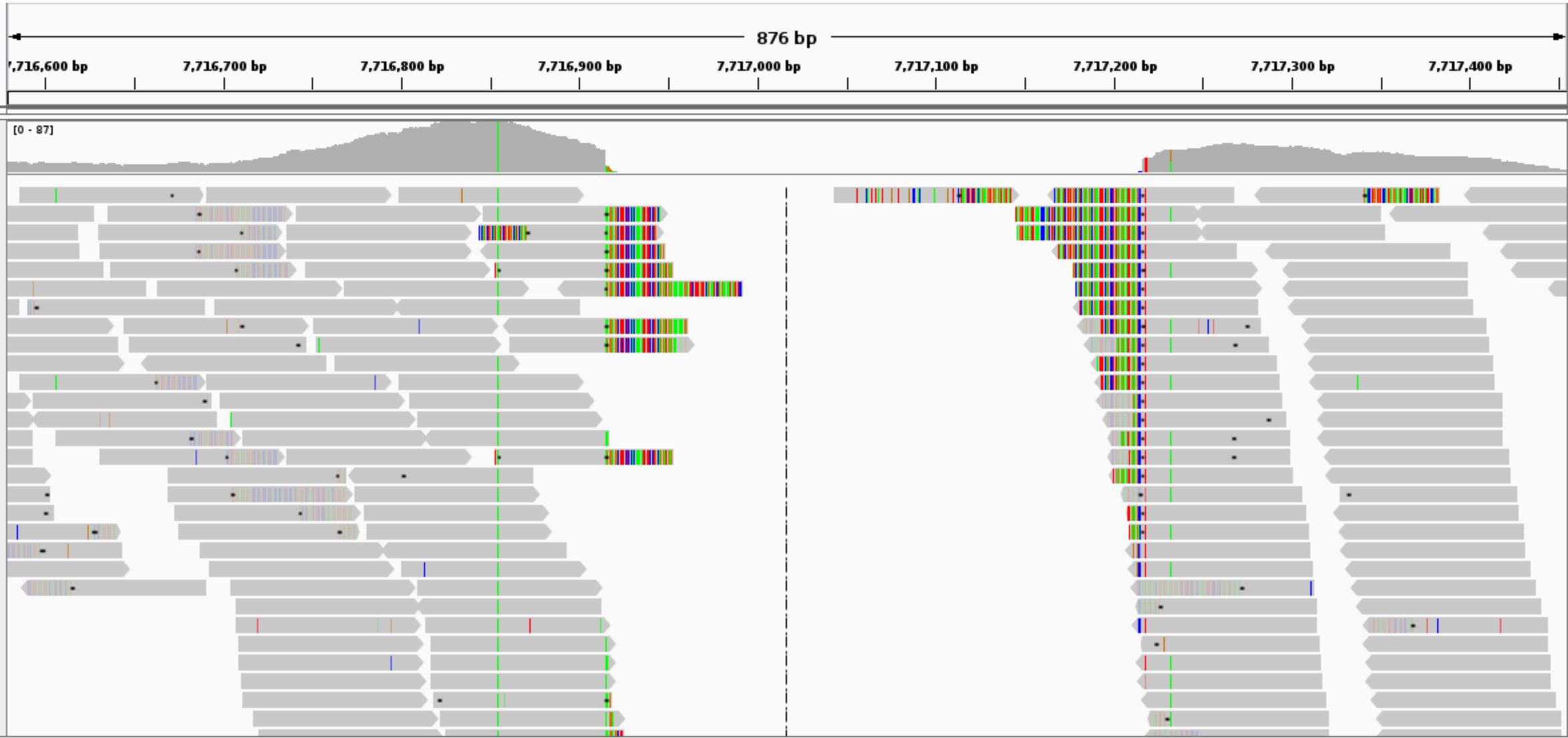
ABRA was applied to 750 breast tumor/normal exome pairs from The Cancer Genome Atlas project. Germline variants were called “Pre” and “Post” ABRA using FreeBayes. Indels novel to ABRA are strongly associated with known SNPs in close proximity (<10kb) as well as ancestry. Several of the most frequently occurring ABRA specific indels appear in dbSNP and were initially discovered in studies utilizing Sanger sequencing.

Somatic variant calling was performed “Pre” and “Post” ABRA using Strelka. An 11% increase in the total number of indels called post-ABRA is observed.

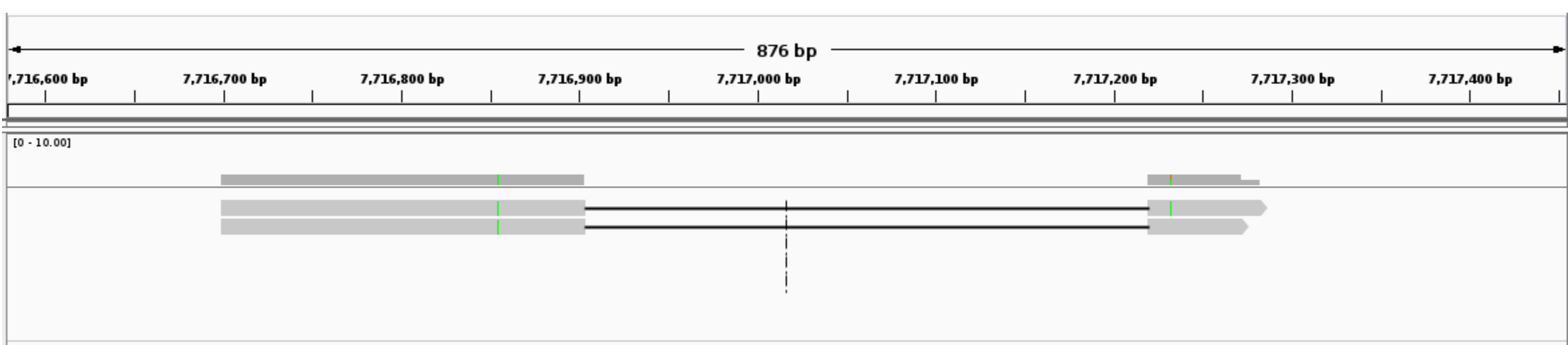
Novel ABRA Somatic Mutation Examples

BRCA2 – 453b deletion	GATA3 – 411b deletion
NOTCH2 - 63b deletion	PIK3R1 – 37b insertion
TP53 – 152b deletion	TP53 – 27b insertion

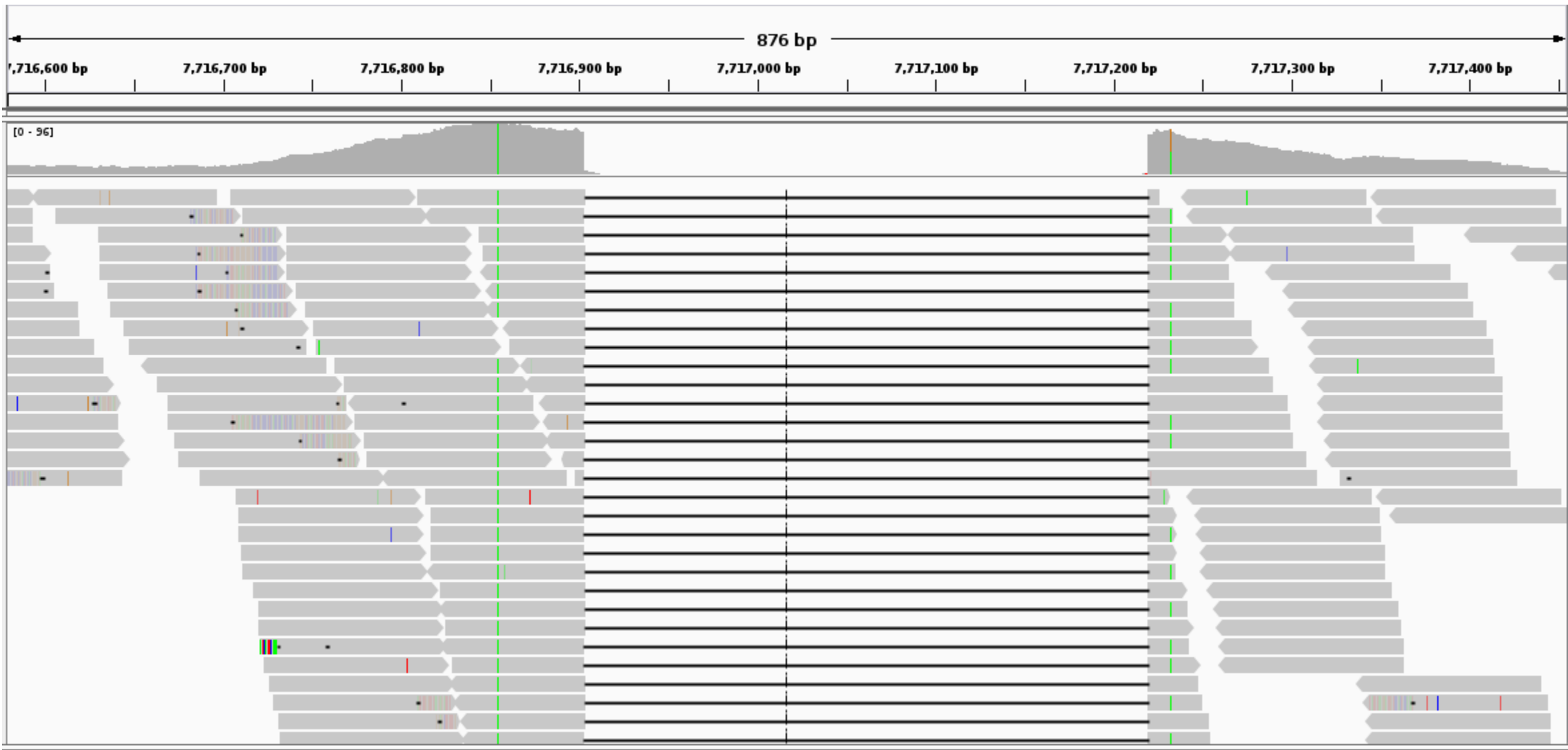
Method



ABRA performs localized assembly on small genomic regions. For exome or targeted sequencing, these regions roughly correspond to capture targets.



After all regions of interest are assembled, reads are globally aligned to the assembled contigs. If a read aligns more closely to an assembled contig than the original reference, that read’s alignment information is updated.



In the example shown here, a 316b deletion is revealed by ABRA. The deletion is bracketed by SNPs within a read length on each side. All 3 variants have corresponding entries in dbSNP. This deletion was found in over 99% of the TCGA Breast samples as well as the HapMap Trio samples.

Acknowledgements

We thank the Cancer Genome Atlas Network for the organization, production, and dissemination of data and results.

We thank Illumina for making the Platinum Genome data available.

This work was supported in part by the North Carolina’s University Cancer Research Fund.

Conclusion

ABRA improves upon NGS read alignments providing enhanced detection of indels and improved variant allele frequency estimation.

ABRA accepts BAM files as input and outputs realigned BAM files, allowing flexibility in downstream analysis.

ABRA can be used with a variety of variant callers for both germline and somatic variant calling.

ABRA is freely available for download at <https://github.com/mozack/abra>

