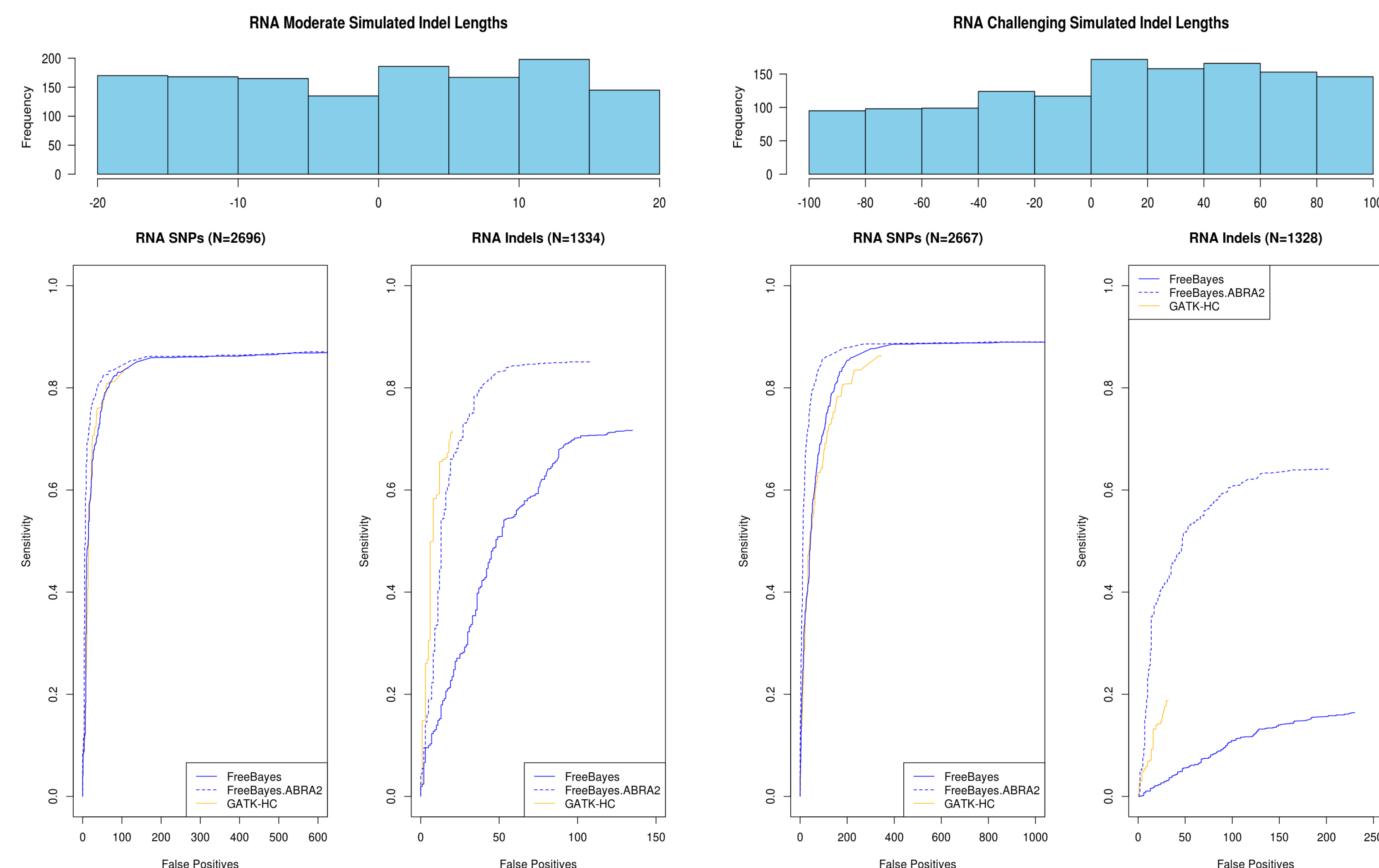


## Background

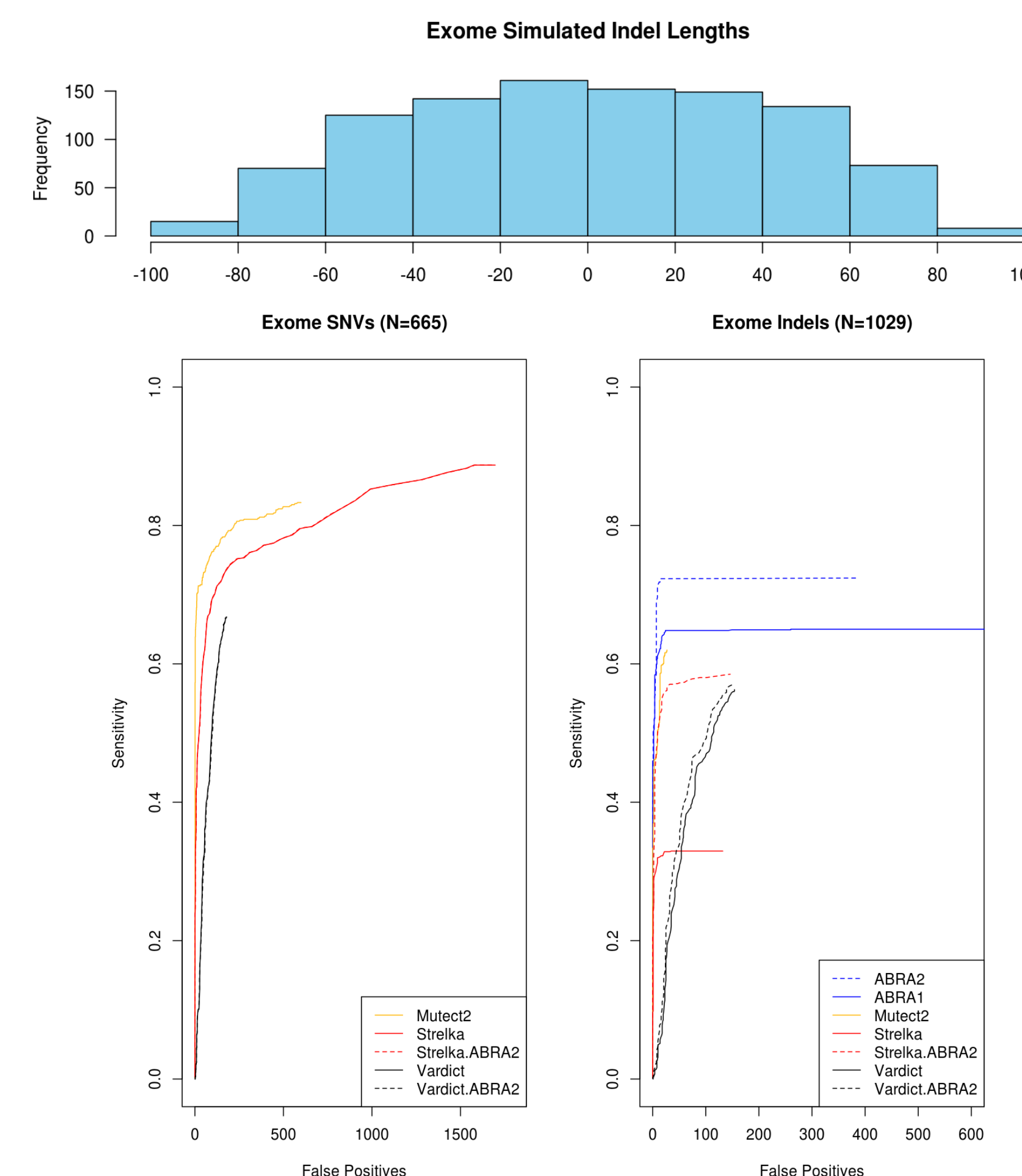
Insertions and deletions in the transcriptome can potentially have significant impact on function and can be clinically actionable. A number of methods have recently been developed that improve indel detection in DNA by utilizing realignment and/or localized assembly to aid in discovering mutations that are more difficult to detect than single nucleotide polymorphisms. While significant effort has been put into these methods in the context of DNA discovery, RNA-Seq has not received the same attention. With this in mind, we have developed ABRA2 which is capable of utilizing splice junction information to augment localized assembly across exon-exon boundaries, thus improving read alignments resulting in improved variant detection in RNA-Seq data. Further, we have improved upon the original ABRA implementation, allowing for improved accuracy in DNA indel detection as well as improved scalability including support of whole genomes.

## RNA Simulation



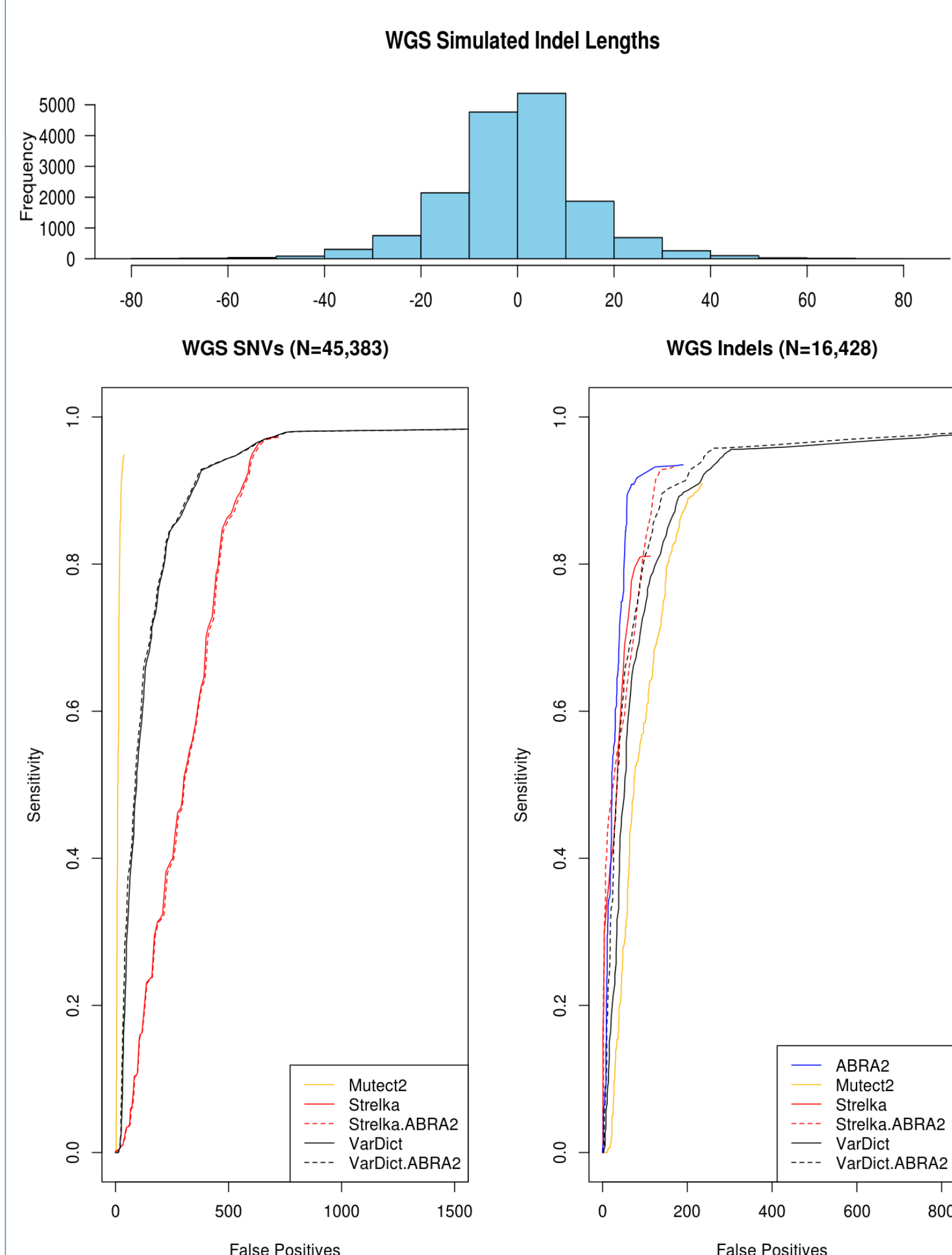
Evaluation of the impact of ABRA2 on variant calling in 2 simulated RNA-Seq datasets. Reads were simulated using a modified version of the BEERs simulator. The dataset on the left includes indels ranging in length from 1 to 19 bases. The more challenging dataset on the right includes indels ranging in length from 1 to 100 bases. The original read mapping was done using STAR.

## Exome Simulation



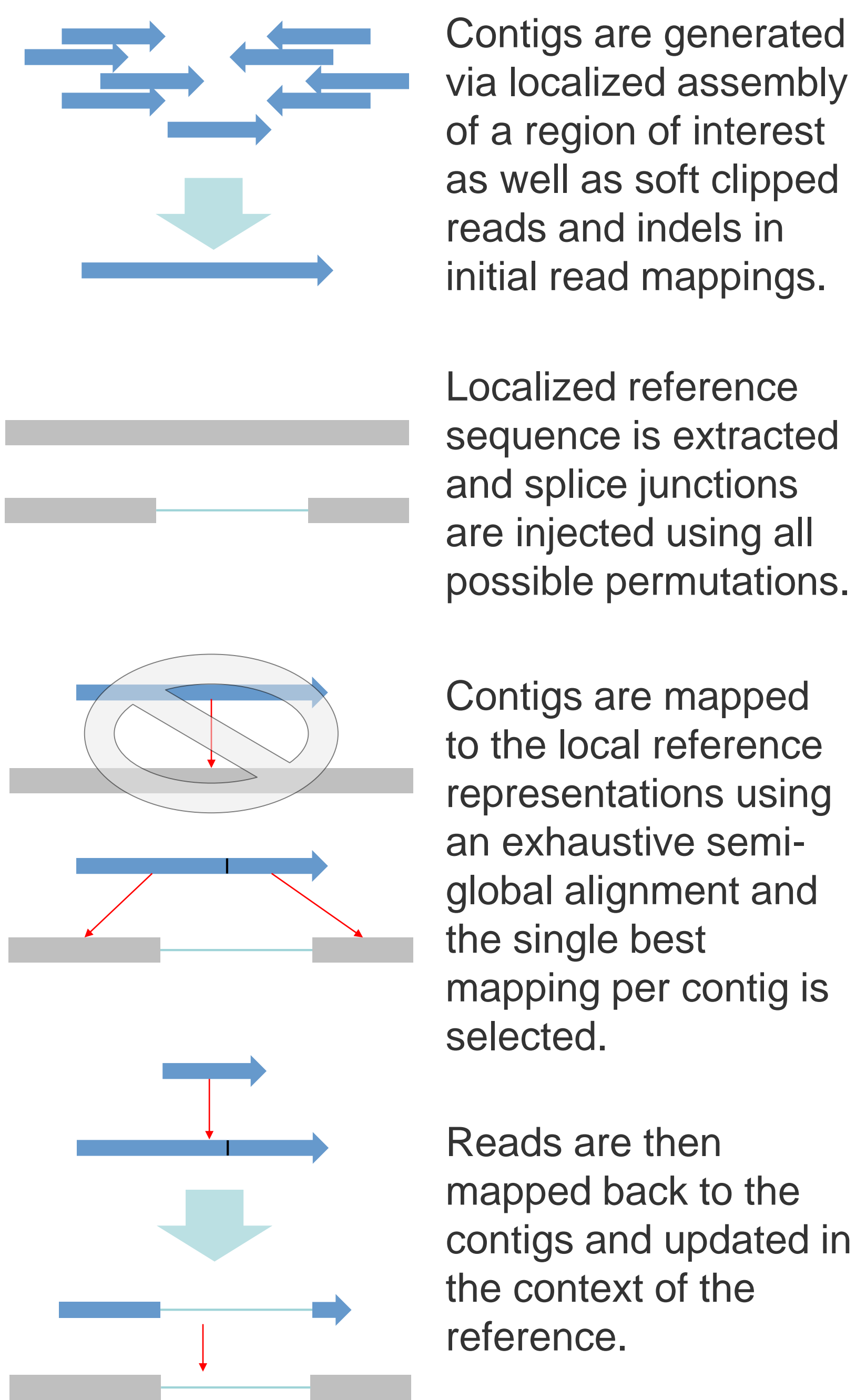
Evaluation of ABRA2 on a challenging exome dataset. An exome from NA12878 was split into a normal and simulated tumor BAM. BAMSurgeon was then used to spike variants into the tumor. Subclone frequencies range from 5% to 35%. ABRA1 and ABRA2 indel calling was done using Fisher's exact test.

## WGS Simulation



Evaluation of ABRA2 on a whole genome tumor/normal pair from the ICGC Dream Somatic Mutation Calling Challenge #5. This dataset contains tumor cellularity of 80% and clone frequency of 50%. The original ABRA implementation was not scalable to whole genomes. ABRA2 indel calling was done via Fisher's exact test.

## Method

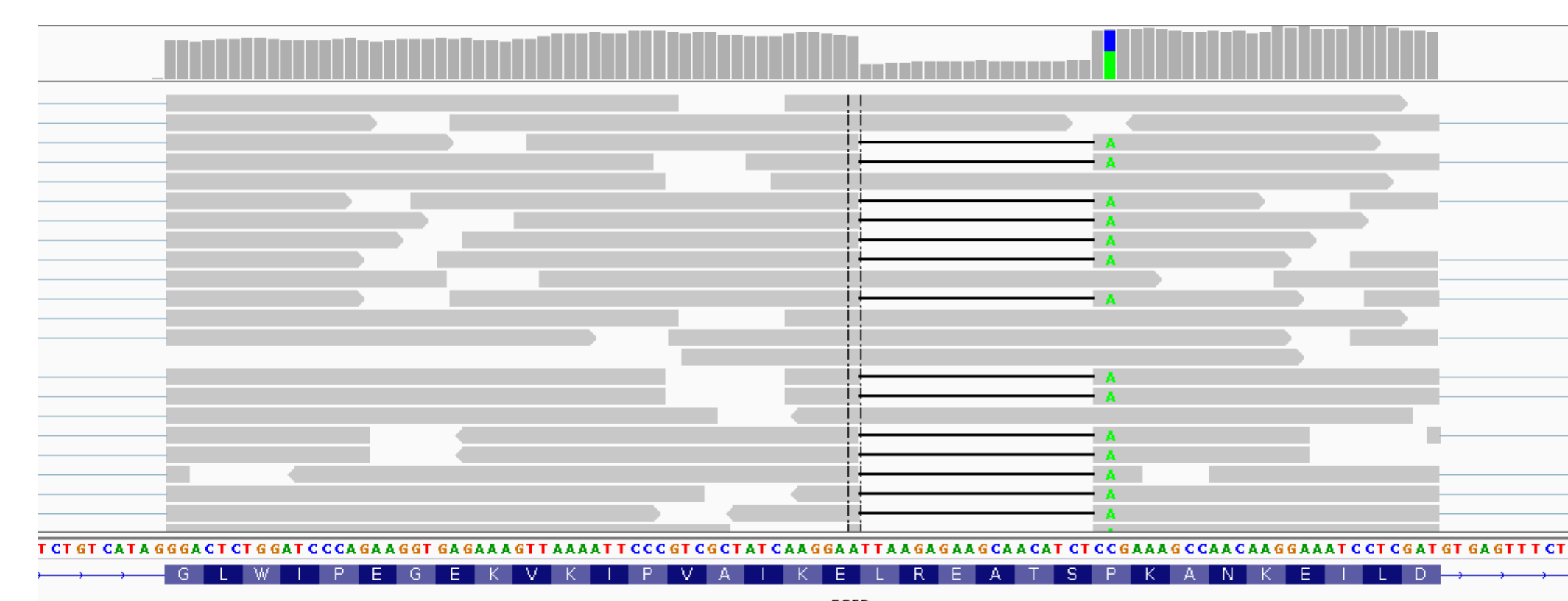
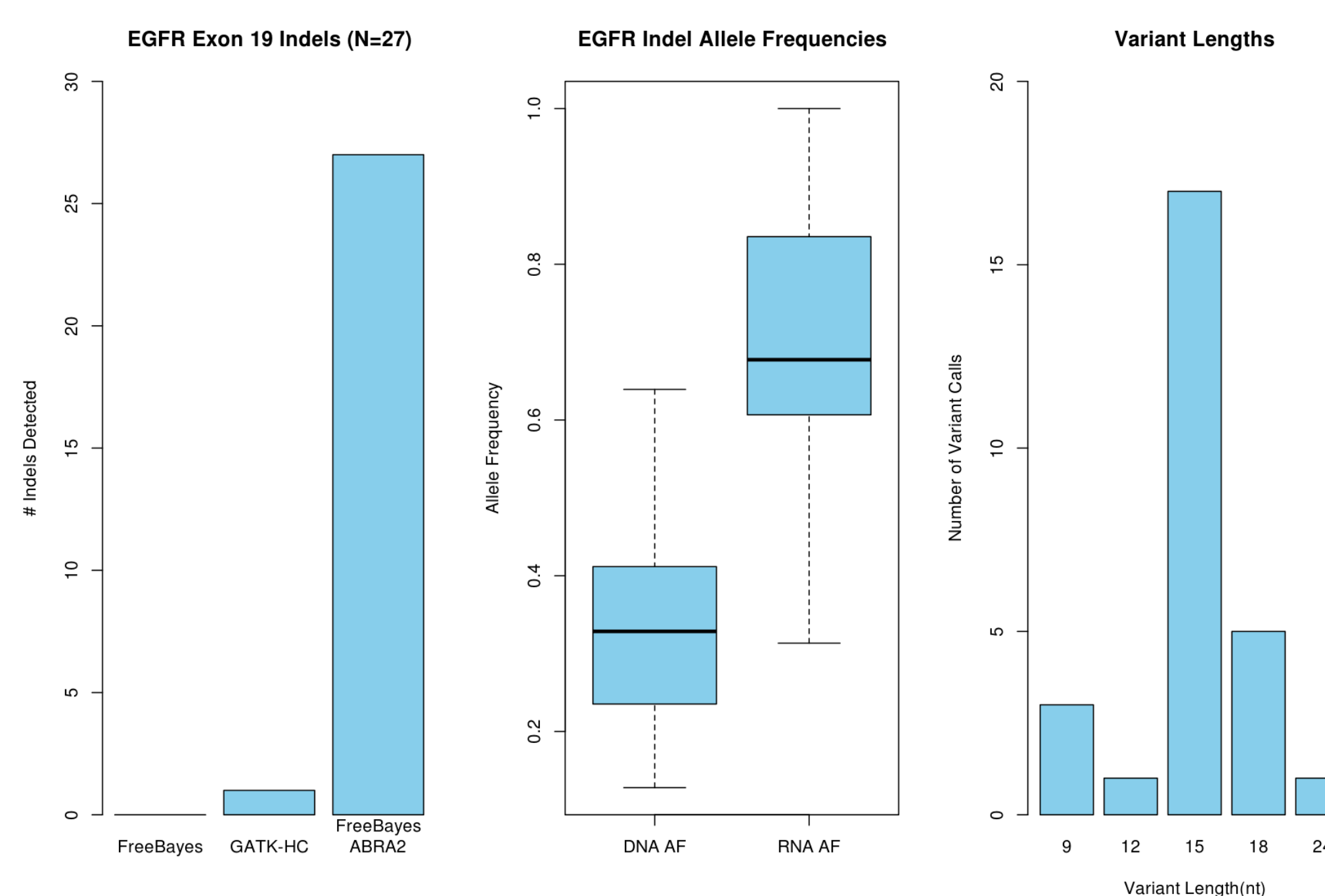


ABRA2 accepts an aligned BAM file(s) as input and outputs realigned BAM file(s)

## Expressed EGFR Indels Identified in RNA-Seq

In frame deletions in Epidermal Growth Factor Receptor (EGFR) have oncogenic potential and can be indicators for Gefitinib or Erlotinib treatment. Deletions such as these ranging in length from 9 to 24 nt were identified from matched tumor/normal DNA in exon 19 for 23 Lung Adenocarcinoma (LUAD) subjects as part of TCGA. Additionally, Ye et al identified 8 complex indels in the same exon with 3 overlapping the TCGA set for a total of 28 cases, of which 27 have available RNA data.

**ABRA2 enables detection of 27 out of 27 EGFR TCGA LUAD exon 19 indels from RNA-Seq alone.**



Example of a complex EGFR variant after ABRA2 realignment in RNA-Seq data. An 18 base deletion and nearby SNV are flanked by neighboring splice junctions on both sides resulting in difficult to map reads.

## Acknowledgements

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

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

## Conclusion

ABRA2 improves upon NGS read alignments in both RNA and DNA providing enhanced detection of indels.

ABRA2 enabled detection of clinically relevant EGFR indels from RNA-Seq in TCGA data.

ABRA2 is freely available at: <https://github.com/mozack/abra2>