DRAGEN™

DRAGEN Virtual Long-Read Detection (VLRD) Pipeline

Accurate variant calling for segmental duplications

Segmental duplication variant calling

SNPs
Enhanced SNP Accuracy

INDELs
Enhanced INDEL Accuracy

Available both onsite and the Cloud

Overview

Segmental duplications are two or more regions of the genome at least 1 kb in length with >90% sequence similarity. Segmental duplications are important to analyze due to their clinical significance in human disease. Segmental duplications present challenges for variant calling from next-generation sequence (NGS) data because short reads can map to more than one reference location.

The DRAGEN Virtual Long-Read Detection (VLRD) Pipeline is an advanced algorithm that calls variants in segmental duplications from short read sequence data. The DRAGEN VLRD Pipeline has much greater accuracy in segmental duplications than standard variant callers and works by jointly calling all regions that are similar. During mapping and alignment of NGS data, DRAGEN VLRD analyzes all sequence data, even those with low MAPQ scores as seen in segmental duplication regions due to their similarity. The DRAGEN VLRD Pipeline then solves for the four most likely haplotypes that originate in these regions of interest and proceeds to variant calling.

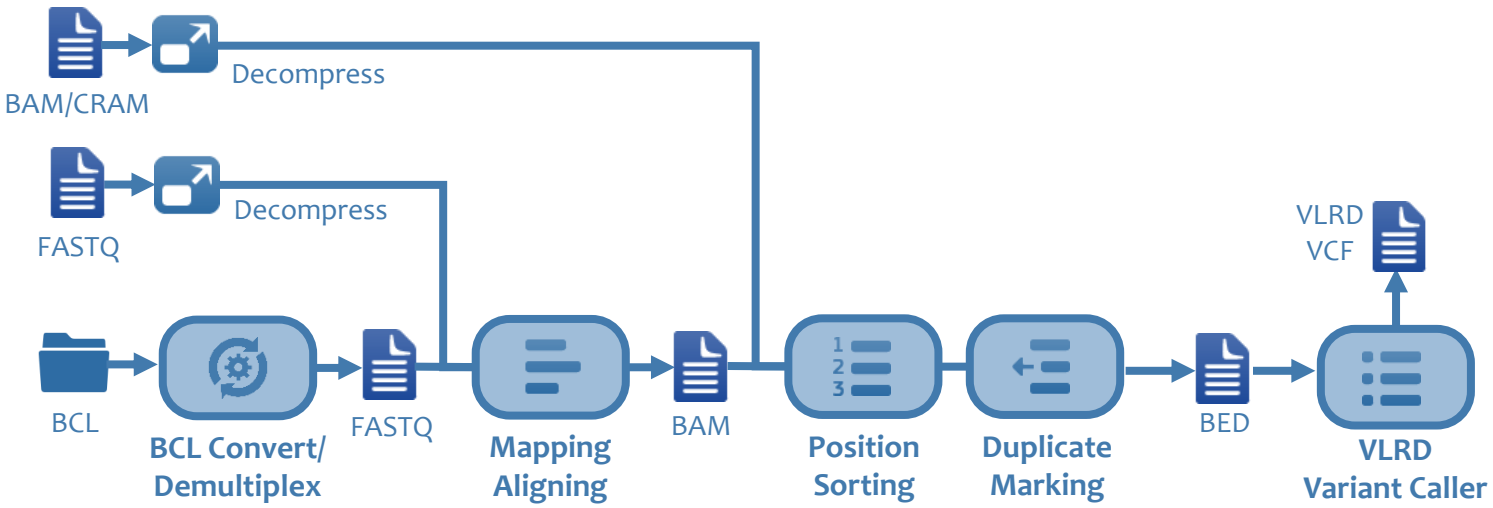
Highlights

- **Comprehensive Variant Calling**—DRAGEN is the only NGS analysis tool with a dedicated algorithm to call variants in segmental duplications.
- **Enhanced efficiency**—Reduces the need for long-read sequencing or validation by resequencing.
- **Easy to Run**—Included as an optional addition to the DRAGEN Germline Pipeline. No separate software needed.
- **Ultra-rapid**—Variant analysis of duplications takes ~20 minutes (whole human genome @ 30x coverage).
- **Improved Accuracy and Sensitivity**—Significant accuracy demonstrated at shorter read lengths than typically necessary for segmental duplication analysis.

edico genome

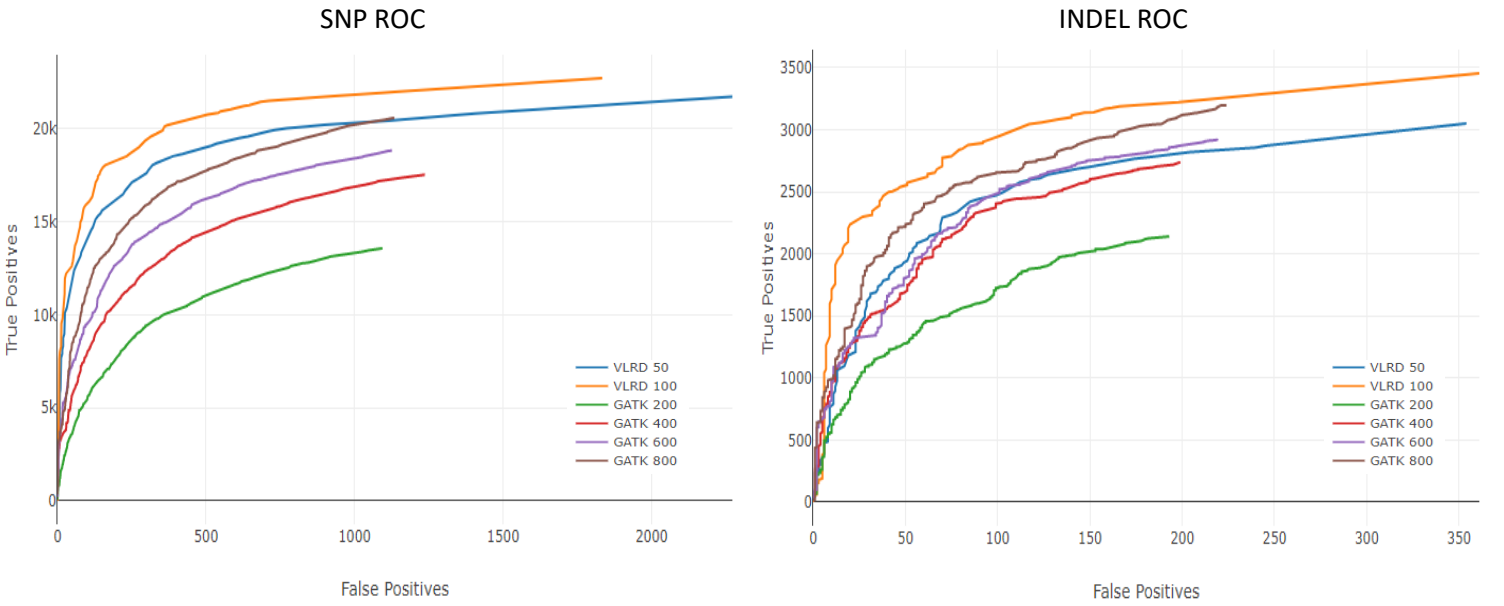
DRAGEN VLRD Pipeline

The DRAGEN VLRD Pipeline accepts FASTQ/BAM/CRAM and produces a VLRD-specific VCF. During mapping and aligning, the DRAGEN VLRD algorithm does not filter out reads with the low MAPQ scores typically found in sequences containing segmental duplications. All reads are considered jointly and are assigned a location based on maximum likelihood estimates. During variant calling, a BED file is used to delineate duplications that are >400 bp. The VLRD algorithm is then run on the duplicate regions and calls SNP and INDEL variants, which are produced in a VLRD-specific VCF file.



Performance: DRAGEN VLRD Pipeline vs. GATK

Longer read length sequencing can mitigate mapping difficulty in segmental duplication regions and improve the accuracy of variant calling. However, the DRAGEN VLRD Pipeline achieves the same accuracy at shorter read lengths than that are necessary for traditional variant callers. The DRAGEN VLRD Pipeline was compared to GATK in genomic regions with ≥98% similarity using the reference genome, hs37d5, at varying read lengths. The DRAGEN VLRD Pipeline shows significant improvement in accuracy at shorter read lengths (50 bp and 100 bp) than GATK at longer read lengths (200 bp, 400 bp, 600 bp, 800 bp).



ROC curves comparing DRAGEN VLRD with the GATK variant caller for two homologous regions. Regions in the hs37d5 reference genome with >98% similarity were analyzed. Random variants were introduced into the genome based on real sample data provided by 10x Genomics and synthetic 100-bp paired end reads were generated from the reference. Results show that the DRAGEN VLRD Variant Caller has better sensitivity and accuracy in duplications compared to GATK.

Pipeline Steps

Input File Formats
- Input BCL, FASTQ or BAM/CRAM
- Output BAM/CRAM or VCF

Position Sorting
- Sorting of reads by position against the reference

Compression / Decompression
- Decompression of FASTQ, BCL, BAM/CRAM
- Gzip and CRAM in and out

Duplicate Marking
- Based on starting position & CIGAR string

Mapping / Aligning
- Single end or paired end reads
- Read lengths from 26 bp to 10 kbp

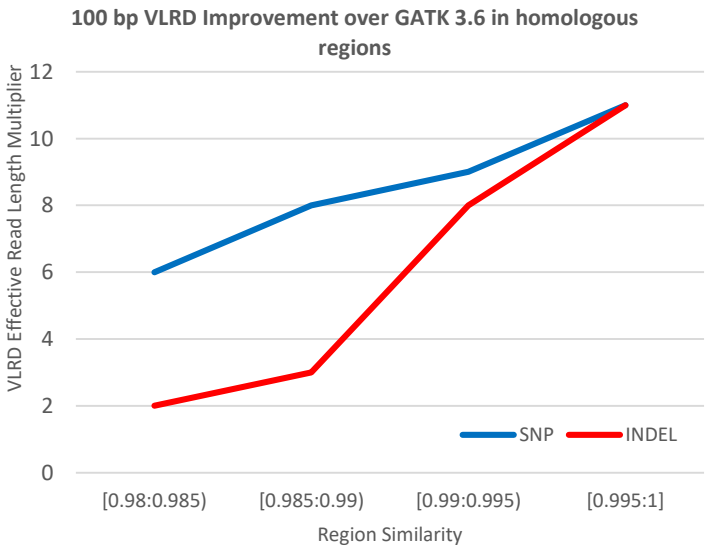
VLRD Variant Calling
- VLRD variant calling of segmental duplicate regions

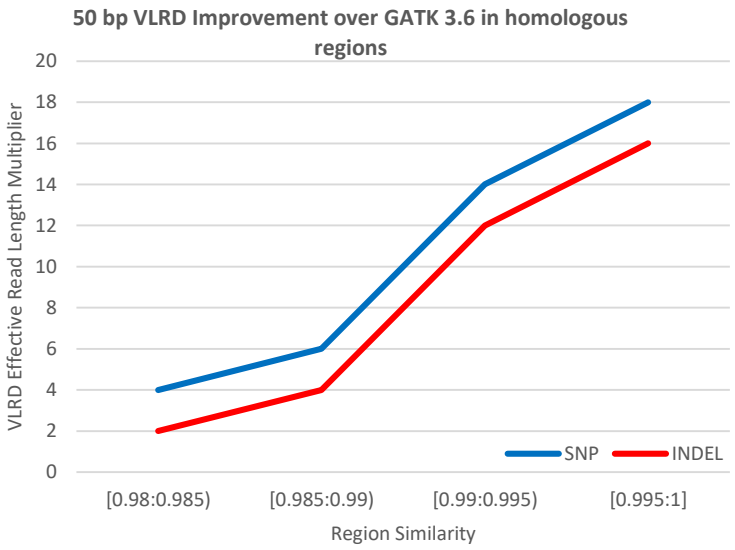
DRAGEN VLRD Pipeline vs. GATK Accuracy at Shorter Read Lengths

We have compared the read lengths of DRAGEN VLRD that are necessary to achieve the same accuracy found in GATK. For regions with similarity > 99%, DRAGEN VLRD achieves greater accuracy at 50 bp read lengths than GATK at 14 times longer (~700 bp). In segmental duplication regions, VLRD is able to achieve a multiplicative read length effect over traditional variant calling. This further demonstrates that DRAGEN VLRD can achieve greater accuracy at industry standard read lengths rather than having to conduct long-read sequencing to analyze regions containing segmental duplications.

DRAGEN VLRD maintains equivalent accuracy to GATK for SNPs and INDELs while using much shorter reads. This table summarize the effective read length multiplier for homologous regions that are 98% - 100% similar, and for datasets where VLRD was run with 50 and 100 bp read lengths. .

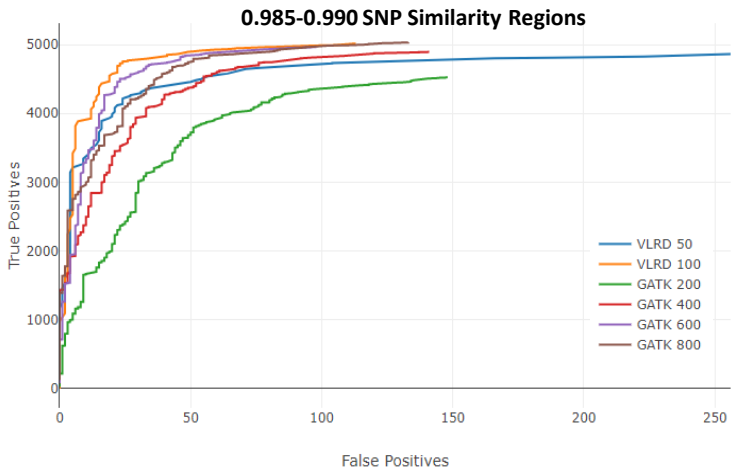
	50 bp		100 bp	
Similarity	**SNP**	**INDEL**	**SNP**	**INDEL**
[0.98:0.985)	4	2	6	2
[0.985:0.99)	6	4	8	3
[0.99:0.995)	14	12	9	8
[0.995:1]	18	16	11	11

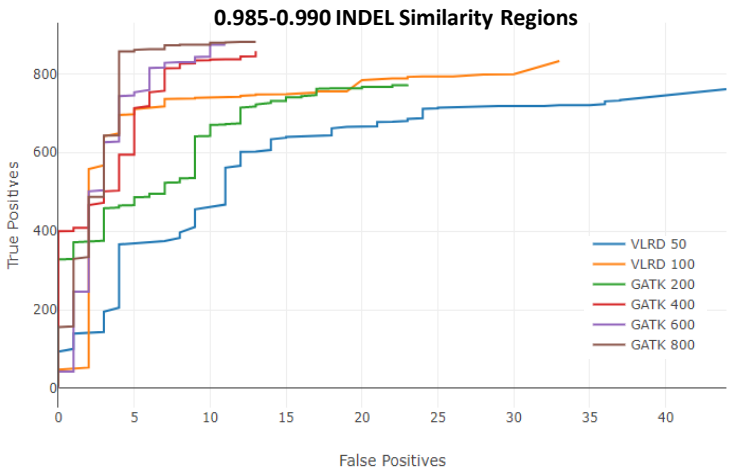


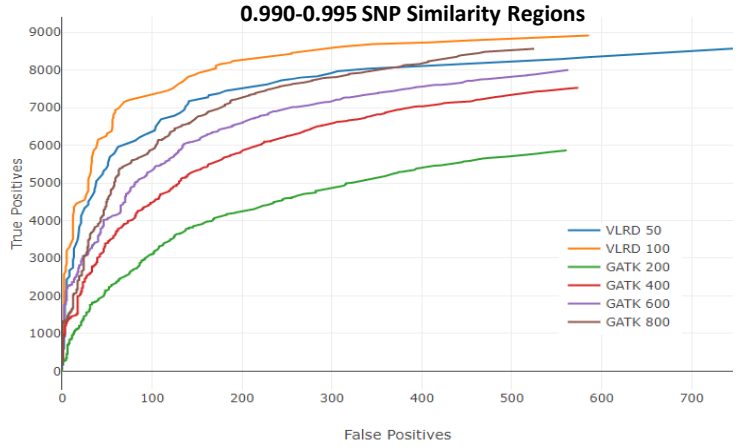


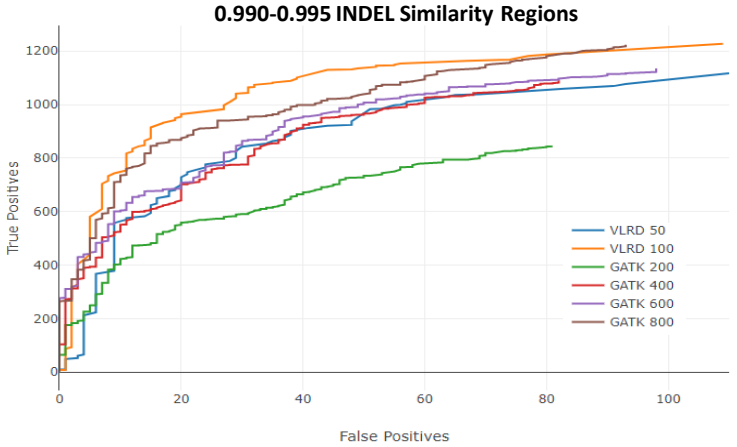
DRAGEN VLRD effective read length multiplier for SNPs and INDELs for two homologous regions at 50 bp read lengths and 100 bp read lengths over GATK. With increased similarity, DRAGEN VLRD maintains equivalent accuracy at 50 bp read lengths to GATK at 18 and 16 times longer read lengths for SNPs and INDELs.

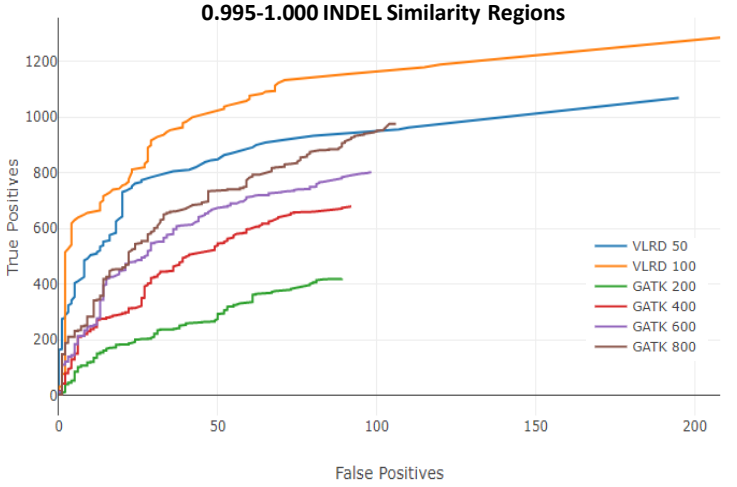
As the similarity between duplicate regions increases, DRAGEN VLRD has increasing accuracy gains compared to the GATK variant caller. In regions with >99% similarity, DRAGEN VLRD with 50bp reads has an effective read length gain of 16x in SNPs and 8x in INDELs over GATK.











About Edico Genome

Edico Genome is the leading secondary analysis solution provider for next-generation sequencing, delivering its powerful DRAGEN Bio-IT Platform to clinical, research and genome centers around the globe. Leveraging field programmable gate array (FPGA) technology, DRAGEN delivers best-in-class accuracy, speed, scalability and costs, enabling customers of all sizes to focus on what matters most – delivering breakthrough results. The comprehensive set of DRAGEN pipelines can be run onsite, in the Cloud or through a seamless hybrid cloud blend, allowing organizations to scale as their throughput fluctuates.

info@edicogenome.com
www.edicogenome.com
@edicogenome