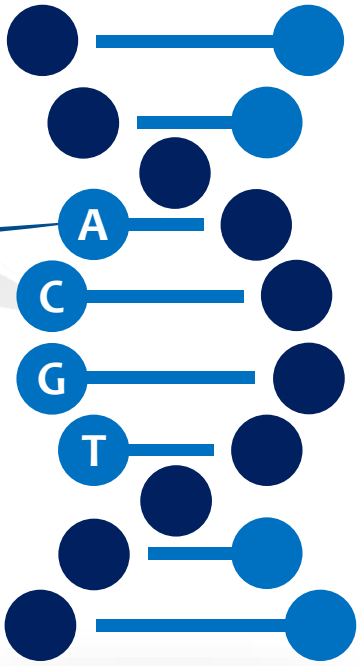
DRAGEN™

DRAGEN Genome Pipeline

Ultra-Rapid Genome and Exome Data Analysis



22min	99.37%	98.31%	Hybrid
FASTQ → VCF	Precision	Sensitivity	Cloud
WHG @ 30x Coverage	SNP + INDEL Combined	SNP + INDEL Combined	Onsite, Cloud, and Hybrid Solution

Overview

The DRAGEN Genome pipeline enables ultra rapid analysis of Next Generation Sequencing (NGS) data, reducing the time required for analyzing a whole genome at 30x coverage from ~10 hours using the current industry standard, BWA-MEM+GATK-HC software, to ~22 minutes, while also improving accuracy for both SNPs and INDELs.

This pipeline harnesses the tremendous power of the DRAGEN Bio-IT Platform and includes highly optimized algorithms for mapping, aligning, sorting, duplicate marking, haplotype variant calling, compression and decompression.

What is DRAGEN?

DRAGEN (Dynamic Read Analysis for Genomics) is a highly reconfigurable Bio-IT Processor which is integrated on a PCIe card and is available in a pre-configured server. DRAGEN can also be integrated directly into sequencing instruments and NGS bioinformatics servers.

The DRAGEN Processor reduces costs related to storage space and IT infrastructure. It's estimated to result in a savings of up to $6 million over four years for customers using Illumina's HiSeq X Ten instruments.

How Does it Work?

DRAGEN takes raw read data produced by a sequencing instrument, such as Illumina's HiSeq X Ten. After variant calling, DRAGEN outputs a standard VCF file ready for tertiary analysis.

The DRAGEN platform includes a fully functional and easy to use graphical user interface (GUI) and Workflow Management System, enabling customers to easily schedule multiple workflow runs, analyze results such as alignment statistics and coverage metrics, compare different pipelines, monitor multiple networked DRAGEN systems and receive updated software releases.

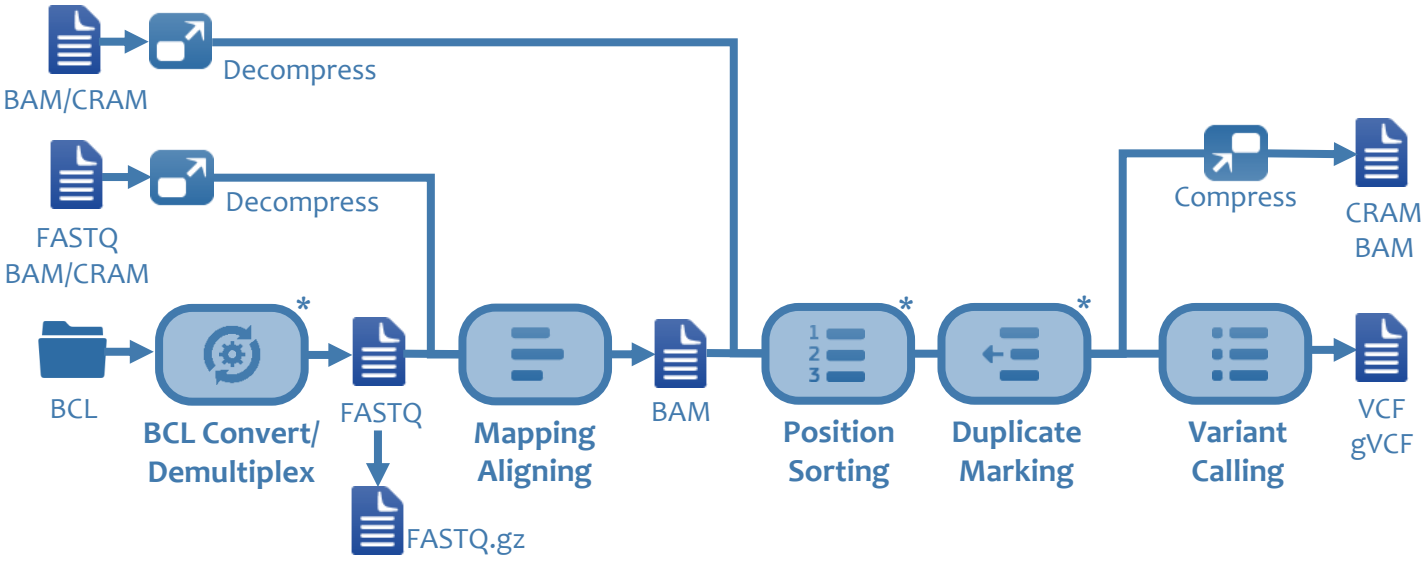
Comprehensive Set of Pipelines

DRAGEN supports pipelines for Whole Genome, Whole Exomes, Targeted Panels, Epigenome / Methylome, RNA-seq / Transcriptome, Microbiome and Cancer Tumor / Normal. DRAGEN is offered as a Platform-as-a-Service (PaaS), enabling customers to license various pipelines according to their needs. All pipelines are available both onsite and in the cloud. Data usage based pricing tiers are available to cater for customers performing benchtop sequencing all the way up to population scale.

edico genome

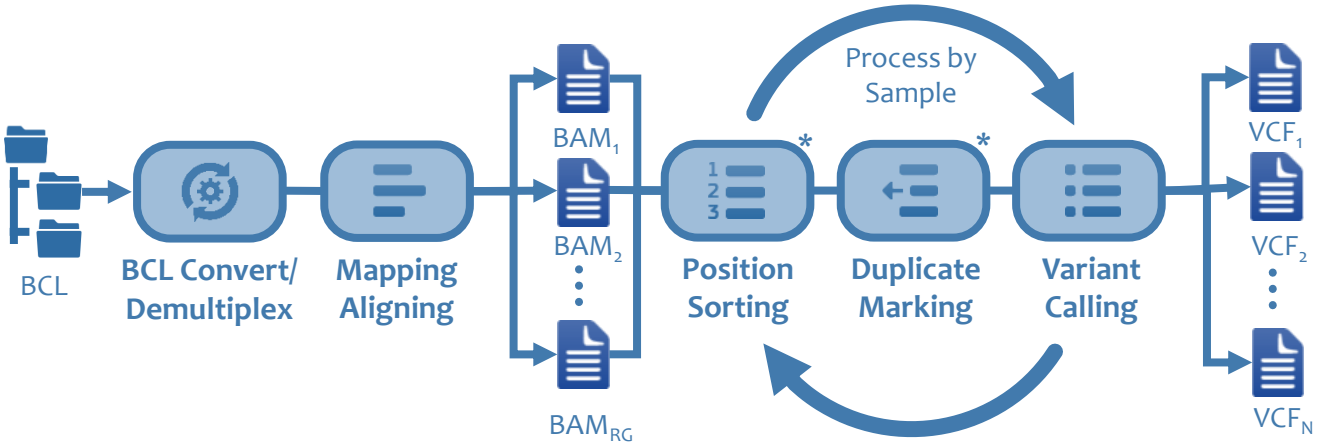
DRAGEN Genome Pipeline

The DRAGEN pipeline offers supreme flexibility of data analysis. DRAGEN can handle multiple input formats and produces industry standard output formats compatible for downstream analysis. DRAGEN can stream BCL data directly from sequencer storage, a solution unique to the DRAGEN pipeline, enabling the customer to go directly from raw sequencing data to an output VCF. DRAGEN can also convert BCL to FASTQ or BAM/CRAM, then proceed with the standard DRAGEN pipeline.



Efficient Multi Sample Processing

DRAGEN is capable of processing BCL data directly, eliminating any FASTQ conversion step. The BCL data is fed directly to the pipeline to produce a unique output VCF file per sample. Intermediate BAM/CRAM files can be generated on demand. To streamline and automate multi sample processing, DRAGEN offers a comprehensive Workflow Management System (WMS). The WMS enables customers to easily schedule multiple workflow runs for any pipeline, as well as adjust or accelerate their own NGS analysis algorithms, pipelines, and applications.



Applications

NICU Diagnostics

Clinical Genomics

Sequencing Centers

Agri Genomics

NIPT NIPS

Pipeline Steps

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

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

BCL Convert/Demultiplex
- BCL conversion to FASTQ
- BCL can also be processed directly

Mapping / Aligning
- Single end or paired end reads
- Supports read lengths from 26 bp to 10k bp

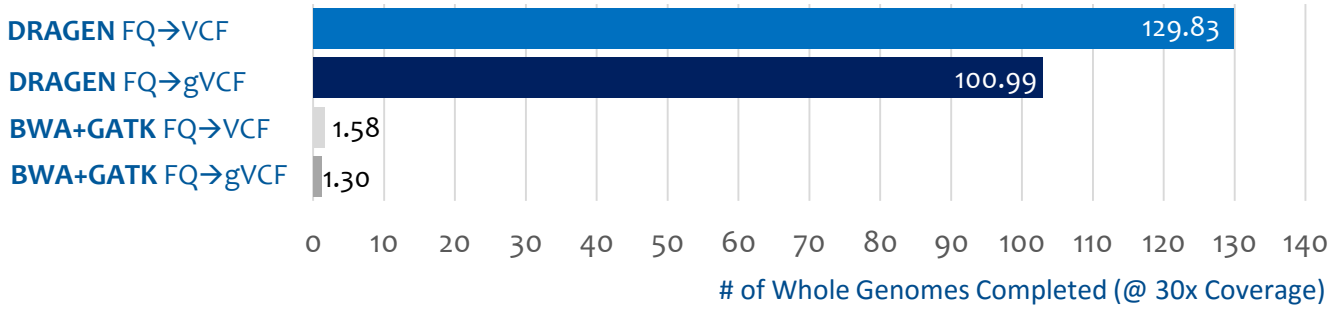
Position Sorting
- Binning by reference range
- Sorting of bins by reference position

Duplicate Marking
- Based on starting position & CIGAR string
- Highest quality duplicate reported

Variant Calling
- Haplotype variant caller with reassembly
- Uses Hidden Markov Model and Smith-Waterman Alignment

* Optional Pipeline Step

Ultra-Rapid Analysis: # Genomes Sequenced in 48 Hours*



Speed: Single Sample Pipeline*

Dataset	Pipeline Configuration	DRAGEN	BWA + GATK-HC	DRAGEN Speed Up
SRA056922 NA12878 @ 30x	FASTQ to VCF	0:22:11	30:20:20	~82X
	FASTQ to gVCF	0:28:31	36:53:29	~78X
Garvan Lane1 NA12878 @ 37x	BCL to VCF	0:26:28	32:18:25	~73X
	BCL to gVCF	0:29:27	38:51:34	~79X

Speed: Multi Sample Pipeline*

Processing of 12 Samples	DRAGEN	Bcl2fastq + BWA + GATK-HC	DRAGEN Speed Up
BCL to FASTQs to VCFs	4:21:43	279:02:36	~70X
BCL to VCFs	3:26:49	N/A**	N/A**

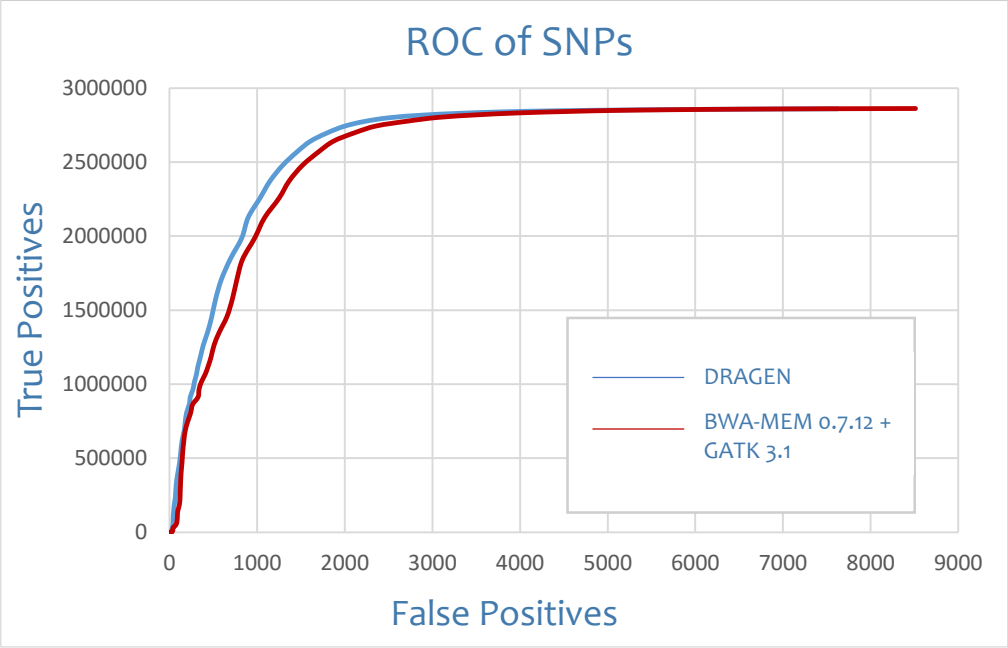
**BCL direct to VCF capability is unique to the DRAGEN pipeline.

Accuracy: Single and Multi Sample Pipelines*

Accuracy	Pipeline	True-Pos	False-Pos	False-Neg	Precision	Sensitivity	F-Measure
SNP + INDEL Combined	DRAGEN	3230927	20398	55603	99.37%	98.31%	98.84%
	BWA-MEM + GATK	3233756	21402	55834	99.34%	98.30%	98.82%
SNP Only	DRAGEN	2859895	7624	25434	99.73%	99.12%	99.43%
	BWA-MEM + GATK	2862995	8511	25593	99.70%	99.11%	99.41%
INDEL Only	DRAGEN	371032	12774	30169	96.67%	92.48%	94.53%
	BWA-MEM + GATK	370761	12891	30241	96.64%	92.46%	94.50%

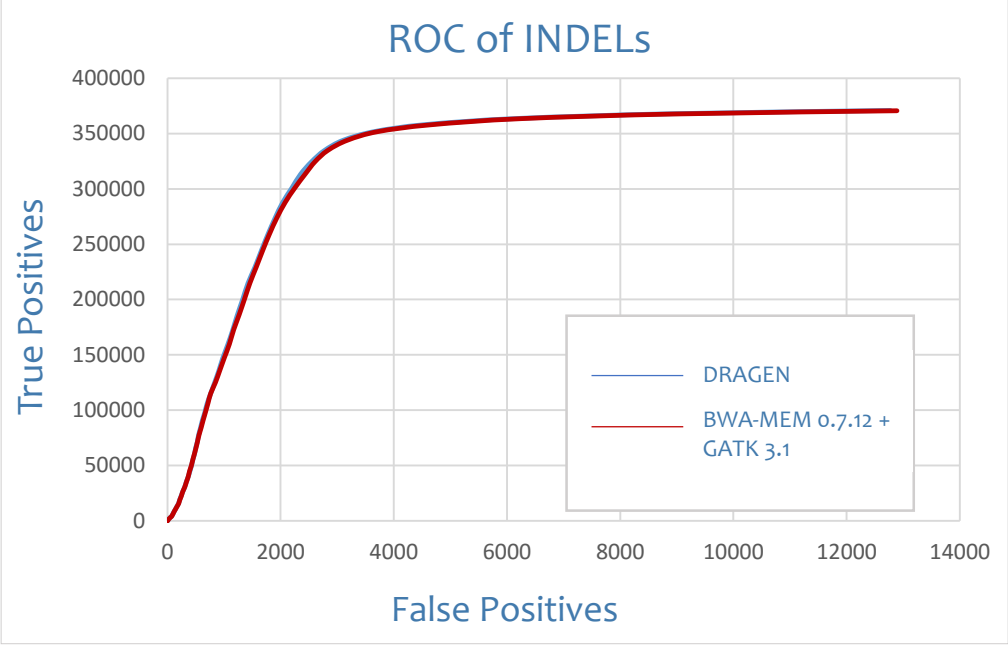
*All DRAGEN results are compared against BWA-MEM 0.7.12 + GATK 3.1 running on comparable servers.

ROC Plots of Variants at 30x Coverage



ROC of SNPs

A SNP (single nucleotide polymorphism) occurs when a single base differs between two genomes, in this case the subject and the reference genome. Use of the NIST Platinum Genome high confidence call set enables performance comparisons between different pipelines. In this ROC plot, a higher count of true positive SNPs and lower count of false positive SNPs is considered better.



ROC of INDELs

An INDEL (insertion or deletion) occurs when bases are inserted or deleted in the subject genome with respect to a reference genome. Use of the NIST Platinum Genome high confidence call set enables performance comparisons between different pipelines. In this ROC plot, a higher count of true positive INDELs and lower count of false positive INDELs is considered better.

About Edico Genome

At Edico Genome, we're helping usher in the new era of personalized medicine by enabling a fundamental change in healthcare with customized treatments and data-driven insights tailored to the individual. At the heart of personalized medicine, DNA sequencing technology is advancing at an even more rapid pace than the cell phone revolution. By increasing the speed and accuracy for NGS data analysis, such as whole genome sequencing (WGS), our computing platform makes it easier to discover links between DNA sequence variations and human disease.

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

edico genome