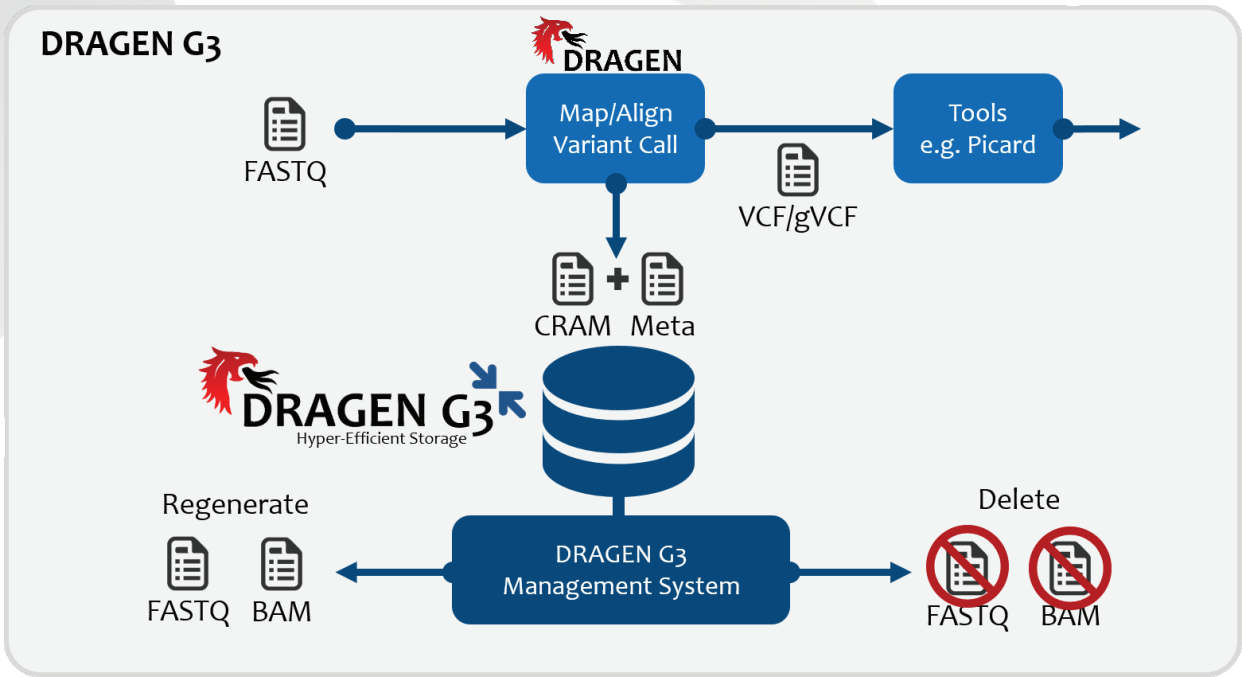
DRAGEN™

DRAGEN G3 Compression

DRAGEN G3 compression technology performs lossless data compression in near real-time. Edico Genome's FPGA-based G3 reduces data size by nearly 70% in lossless mode and 80% in quantized (lossy) mode, significantly reducing the cost to store NGS data.



The NGS Dilemma: Big Data Storage

Next generation sequencing (NGS) is on track to become one of the largest data producers. With new sequencing instruments allowing for higher throughput and increased adoption of personalized medicine, the amount of NGS data generated will soon exceed the capacity to store it. NGS users typically choose to store raw data files (FASTQ), intermediate alignment files (BAM), and variant calling files (VCF) indefinitely. As a result, companies must invest heavily in their IT infrastructures to store the increasing amounts of NGS data.

DRAGEN Solution: G3 Compression

Edico Genome's FPGA-based DRAGEN Bio-IT Platform features G3 for optimized compression and retrieval of NGS data. Among the benefits of G3 are:

- Up to 81% reduction in storage size for 2 FASTQs + BAM, significantly reducing the cost to store genomic data
- Only platform that can directly compress FASTQs from a paired-end run into a single CRAM file
- Compresses a whole genome BAM file in ~ 8 minutes
- Lossless compression for bit-exact file recovery or quantized (lossy) compression for reduced storage space
- Compression and storage available onsite or in the Cloud

edico genome

G3: Clearing the Big Data Bottleneck

To address the challenge of preserving large amounts of NGS data in a cost-effective way, Edico Genome has developed G3 Compression Technology to run on its DRAGEN Bio-IT Platform. G3 provides lossless, near real-time compression of NGS data. G3 utilizes the DRAGEN Workflow Management System (WMS) to provide users with a seamless, automated way to analyze, store, and recall NGS data on a single platform. G3 compression and storage is available both onsite or in the Cloud, providing users with flexibility to suit their particular needs.

G3 Data Compression

DRAGEN accepts any NGS file type and compresses it rapidly with no loss of information. DRAGEN is the only platform that can directly convert FASTQs from a paired-end sequencing run into a single CRAM file, requiring ~ 17 minutes and reducing file size by 45%. Many users choose to save their FASTQ files plus the resulting BAM alignment file. G3 can preserve all these files while reducing storage size by 67%, enough to store three genomes for the cost of one. G3 also features a quantized (lossy) option that achieves even greater storage reduction --- up to 81% compression of 2 FASTQs + BAM --- while still retaining Illumina's recommended level of base quality score information.

G3 File Regeneration

Each time it compresses a file or group of files, DRAGEN creates a metafile that is used to regenerate the original files with 100% bit-accuracy. When the user accesses a G3 compressed file, the original data is regenerated on the fly, providing the same user experience as if working with the original, uncompressed file. The G3 compression system retains all original file information so that users can regenerate the exact data and attributes of the original file, even though only a fraction of the information exists in physical storage. Users realize the cost saving of storing more NGS data without any compromise in their ability to perform analyses.

HiSeqX run of 1 genome (30X coverage at 2 x 150bp) with G3 Compression

Storage Options	Original File Size	CRAM Storage (Lossless/Lossy)	
		CRAM Size	Storage Reduction
2 FASTQs	94 GB	51 GB / 24 GB	**~45% / ~74%**
BAM	83 GB	51 GB / 18 GB	**~39% / ~78%**
2 FASTQs + BAM	177 GB	59 GB / 34 GB	**~67% / 81%**

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