

# Qumulo for Genomic Sequencing

Qumulo delivers a high performance, cloud-native distributed file system that meets the performance and capacity demands for storing, managing and accessing genomic sequencing data, on-prem or in the cloud. Qumulo's file storage efficiently manages billions of small files, and supports a variety of protocols including SMB, NFS, FTP and REST, which means that all phases of the genomic analysis workflow can use the same Qumulo cluster.

## Enterprise-proven performance

Qumulo's file system delivers extreme low latency, IOPS and throughput performance, and handles billions of small files as efficiently as large ones. With Qumulo, researchers can perform their analyses in real-time, which translates into cost efficiencies and faster time to market.

## Efficiency

Qumulo makes 100% of user-provisioned capacity available for file storage, in contrast to legacy scale-up and scale-out NAS that can only support 70% to 80% of usable capacity. Qumulo's efficient use of disk space decreases the data footprint, which not only reduces storage costs, but reduces overall TCO.

## Real-time visibility and control

Qumulo's built-in, real-time visibility and control provides information about what's happening in the storage system, down to the file level, no matter how many files are in the system. System administrators can apply quotas in real time. Insights into storage usage and capacity trends enable organizations to manage ongoing requirements and more proactively plan for future performance and capacity growth.

## Scales across cloud and on-prem

As Qumulo runs the same software on-prem and in the cloud, organizations that want to move some of their genomic analysis workloads can easily scale-across to AWS or GCP. Just point your applications at a cloud instance when you are ready. Qumulo provides high performance and a full set of enterprise features, such as multi-protocol support and real-time visibility.

## Continuous replication

Qumulo uses continuous replication to move data where it's needed, when it's needed. Continuous replication creates a copy of the data in a directory on your primary cluster and transfers it to a directory on a second, target cluster. Continuous replication means you can easily transfer data from your on-prem cluster to your cluster in AWS or GCP to perform your computations, and then transfer the results back to the on-prem storage.

## FEATURES AND BENEFITS

### Performance for Genomics Workloads

Low latency performance with all-NVMe or hybrid options. Supporting mixed I/O performance for billions of small and large files.

### Scale Across

Simple, distributed file storage. Simply add nodes to increase performance or capacity, to scale across on-prem, hybrid or cloud environments, when required. No disruption. No downtime.

### Real-time Visibility

Qumulo eliminates data blindness by providing instant visibility across billions of files. Admins can proactively detect performance and capacity trends as well as access historical data trends.

### Take Control

With Qumulo's REST API, users can build and manage a modern application stack. Easily integrate applications on-prem and in the cloud with modern development tools.

### Customer Delight

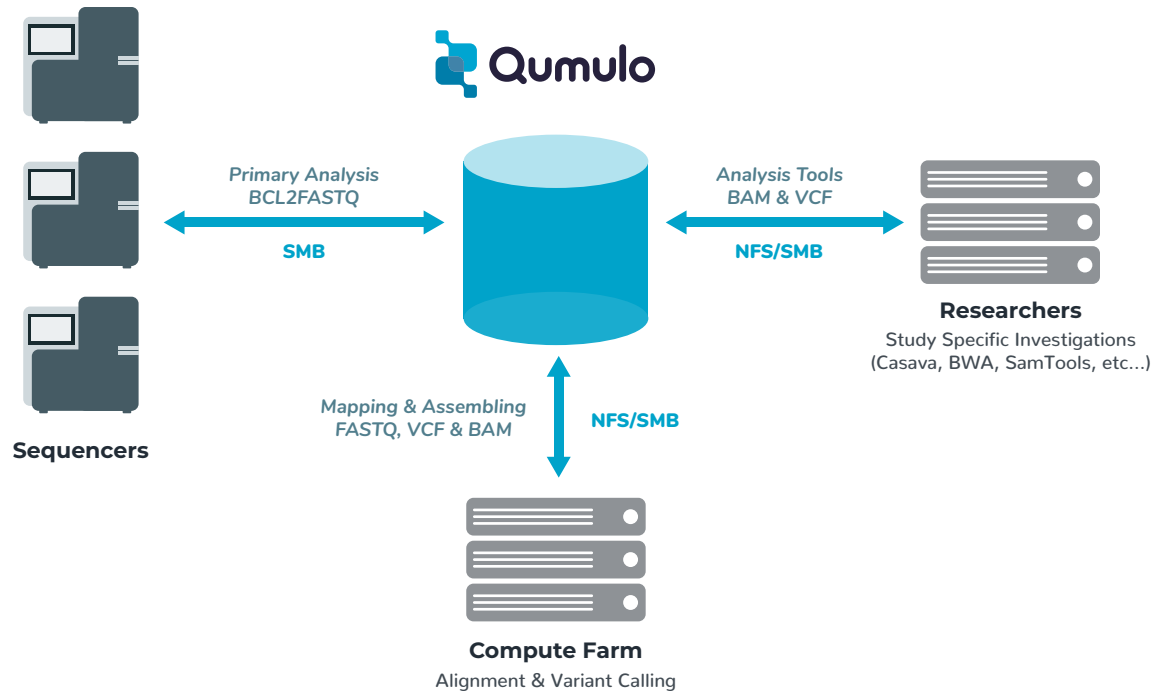
Get help fast from our expert team of storage experts with real-time communication with dedicated Slack channel.

*"Our research organization falls between the cracks for most storage vendors, with giant imaging sets and millions of tiny genetic sequencing scraps. Finding a system that reasonably handled all our complex workflows was difficult, and in the end only Qumulo's file storage was the right fit."*

— **Bill Kupiec**, IT Manager, Department of Embryology, Carnegie Institution for Science

## NEXT-GENERATION SEQUENCING (NGS) WORKFLOW ON-PREM

Below is an example of an on-prem NGS workflow configuration. With Qumulo, organizations can also run a compute farm in the cloud using AWS or GCP.



This example shows the DNA sequencers generating many small BCL files or base calls, which are unordered DNA sequence fragments. A process of demultiplexing assembles BCL files into a FASTQ file, which is a text file that stores the combined output results of the BCL files along with corresponding quality scores.

The compute farm performs alignment and variant calling. In alignment, sequence fragments are quality checked, preprocessed and aligned to a reference genome. A BAM file is a binary file that stores this alignment data. Variant calling looks for differences between the data and the reference genome. Results are stored in a VCF file.

Once these data stores are ready, they can be used for application-specific analysis, which is done by researchers for their own projects. For example, a researcher might be working on a targeted therapy for patients with a tumor that has a specific gene mutation. Researchers may use all the data that is generated contained in the BAM and VCF files. Qumulo provides a distributed file storage system that is suited for all types of genomic data. The file system includes industry-leading small file efficiency, with the powerful performance and throughput to handle all phases of the workflow.

## ABOUT QUMULO

Qumulo is the leader in enterprise-proven hybrid cloud file storage, providing real-time visibility, scale and control of your data across on-prem and cloud. [www.qumulo.com](http://www.qumulo.com)