SOLUTION BRIEF

Hybrid Cloud Storage for Life Sciences

Challenge:

Technology and collaboration are driving discovery in the realm of genomics research. The first obstacle to better understanding the human genome was sequencing the genome. Now, after several generations of sequencing technology, we are capable of sequencing hundreds of thousands of genomes per year, and that number had been doubling annually—reaching an estimated 1.6 million genomes this year [1].

Generating and using this sequencing data requires increasing amounts of both compute power and storage as well as increased data availability. For example, one Illumina HiSeq X Ten can generate well over 100TB of raw sequencing data each month, but that raw data is not useful until it is consolidated, aligned, and analyzed—all of which further increase storage requirements. And with the increasing pace of genomic research and the advance of personalized medicine, the speed of data access and processing must increase as well. So, today—more than ever before, genomics researchers need highly scalable storage capable of very high throughput rates with simple mechanisms for sharing and distributing data.

- Accelerating pace of data generation
- Insufficient throughput when storing or retrieving data
- Lack of metadata / "search-ability"
- Complicated strategies for collaboration and distribution
- Complex storage environments with lots of small storage silos

Solution:

The majority of genomics research data is "unstructured"—meaning that it is typically stored as a collection of files rather than an entry in a database (structured data). Traditional storage is not the answer to the industry's tremendous growth; legacy architectures have bottlenecks and limitations—particularly with an influx of many small files.

On the other hand, SwiftStack Hybrid Cloud Storage is well suited to meet these genomics data storage requirements while also delivering a superior total cost of ownership (TCO). Using the most modern cloud-native architecture and design, SwiftStack has been built to deliver scale-out storage without the technology lock-in normally associated with legacy storage solutions. SwiftStack can easily support the output of next-generation sequencers (NGS) and meet the availability, throughput, and collaboration needs of today's genomics researchers.

BENEFITS

- Single Namespace
- Manage Massive Growth
- Create Efficient Workflows
- Enable Collaboration
- Keep Data for Decades
- Simple Metadata Search

SOME OF OUR CUSTOMERS

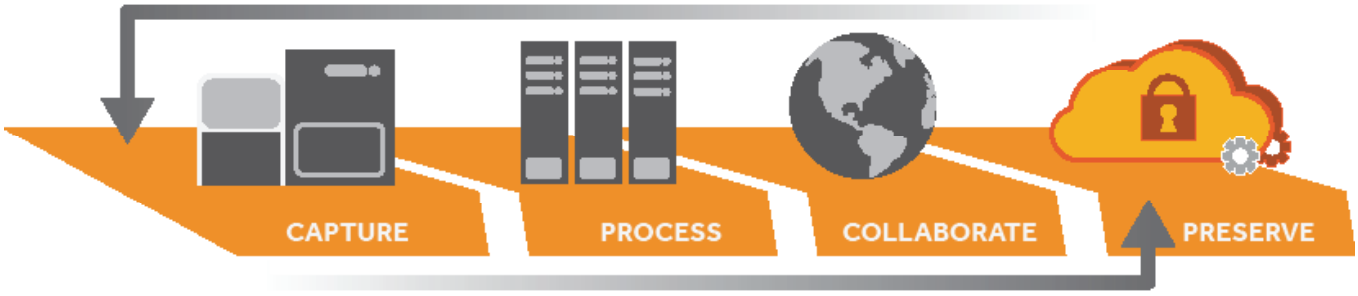
Counsyl

FRED HUTCH
CURES START HERE™

HudsonAlpha
INSTITUTE FOR BIOTECHNOLOGY

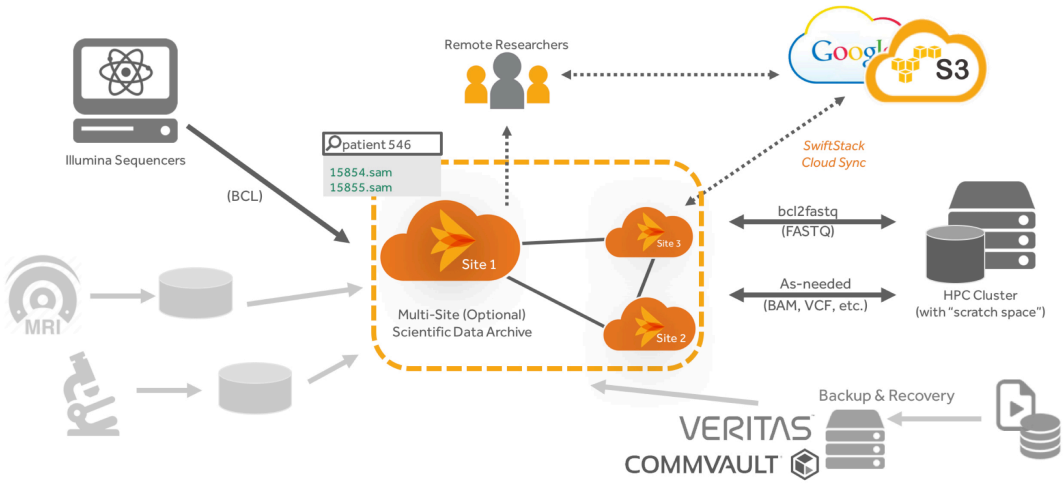
WARWICK
THE UNIVERSITY OF WARWICK

OMRF



Use Case: Storage for Sequencing Workflows

During the sequencing process, there are various storage and compute workflows used to generate, align, analyze, and distribute genomic data. Let's take a look at how SwiftStack fits into each stage of the workflow.



Stage 1: Sequencing

This is the first storage challenge: With a full set of 10 Illumina HiSeq X Ten sequencers, over 4.5 million base call files can be generated from concurrent 3-day runs. The HiSeq X Ten produces up to 13TB across 20 million files in this period. This process requires the following storage characteristics:

Small-file Ingest: SwiftStack is architected to distribute data across many devices in a storage cluster, which provides high throughput ingest of lots of small files over the course of multiple, back-to-back sequencing runs.

Highly durable writes: As data is streaming from multiple sequencers, the storage system must be able to support a continuous ingest rate. Writes are fully durable with SwiftStack—either using multiple object replicas or erasure coding. When a write is acknowledged, the durable write is complete.

Integration: Most sequencer workstations buffer BCL files and then send them to a network CIFS/SMB share, though newer sequencers are adding support for writes to cloud storage APIs. SwiftStack supports the most popular cloud APIs (e.g., S3), is adding native support for NFS and CIFS in 2017, and also offers a "watch folder" feature that can automatically upload BCL files from a sequencer workstation.

Stage 2 & 3: Raw Data Processing and Alignment

Most often, a compute cluster is used to consolidate BCL files into a FASTQ. Many SwiftStack customers take advantage of SwiftStack's massive throughput rates (resulting from concurrent access to small files or segments of large files) to stage BCL files locally to the compute cluster for fast processing and then to upload the FASTQ file into SwiftStack. In most cases, this is managed by a LIMS or similar automation toolchain. For those whose workflows include alignment against a reference genome to produce a BAM file, the same concept holds true. Commonly, FASTQ and/or BAM files are also stored with unique metadata to simplify subsequent searching in the archive.

Stage 4: Analysis

Like in the processing and alignment stages, SwiftStack's ease of access (using automation tools or a simple GUI client) and high throughput rates accelerate and simplify variant calling and other subsequent analysis steps in a research workflow. And if a particular file or set of files needs to be found in a multi-petabyte archive, searching with metadata is extremely fast and dramatically simpler than navigating confusing file folder hierarchies.

Stage 5: Distribution and Collaboration

Whether it's for transferring a FASTQ file from a third-party sequencing lab to a research customer, sharing data between internal departments, or collaborating with other institutions, being able to control and share access to data is critical, and SwiftStack provides several options.

Multi-Region Architecture: SwiftStack storage can span many data centers in a single namespace and distribute data by user-defined policies to make local access simple for distant researchers.

Authentication and Access Control: Internal and/or external users can be granted permission to access "containers" of data in SwiftStack, or—for one-time download access—temporary URL links can be generated to access a single file for a limited time.

Synchronization to the Public Cloud: With SwiftStack's Cloud Sync feature, containers of data can be synchronized to Amazon or Google cloud storage for additional processing or external access.

(1) http://www.technologyreview.com/news/531091/emtech-illumina-says-228000-human-genomes-will-be-sequenced-this-year/

SwiftStack Inc. 333 Bush Street Suite 1650 San Francisco, CA 94104 contact@swiftstack.com swiftstack.com