Solution Brief

Simplify Genomic Data Management

Achieve higher compression and faster data transfer

Key Benefits

- Increase collaborative efficiency. Transfer with smaller, more portable files over the NetApp® Data Fabric. Typical full genome file size from PetaGene PetaSuite is 16GiB, versus 65GiB to 85GiB for FASTQ.GZ and BAM formats.
- Use less storage capacity and lower costs. Smaller files use less storage and dramatically reduce storage costs.
- Leverage the flexibility of the cloud. With the NetApp Data Fabric, files can be seamlessly and securely moved to and from the cloud to support cloud-based workflows. Cold data can be tiered to the cloud by using FabricPool, freeing performance tiers for new sequencing projects.
- Maintain interoperability with existing workflows and formats. PetaGene PetaSuite lets researchers and clinicians continue using FASTQ.GZ and BAM file representations in their existing tools and pipelines without needing to decompress first.

Introduction

Raw genomic datasets are huge, and scientists and bioinformaticians have long sought ways to reduce the size of the datasets they work with by using a combination of data compression and reduction techniques. When genomic sequencing was in its infancy, raw sequencer output, around 2TB, was often stored for extended periods while bioinformaticians carried out the complex tasks in assembling and aligning of the sequencing data. With these steps complete, the data could be used in variant calling and interpretation, which are vital steps in understanding gene expression and disease.

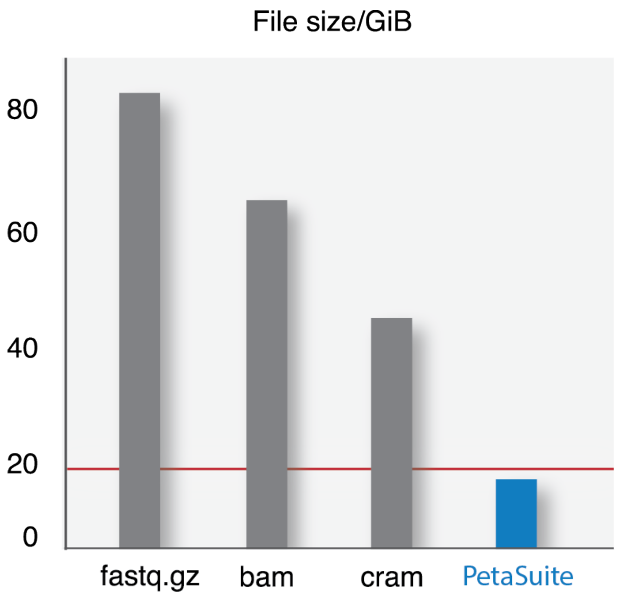

Figure 1) In file size comparison, PetaSuite provides leading data efficiency.

Today this process is highly automated and has been greatly accelerated through a combination of parallel processing and the availability of reference genomes. Work that previously took months or years can now be turned around in little more than a day, and a number of compressed genomic file formats are available that reduce the size of an individually stored genome down to a few tens of gigabytes. This reduction has greatly improved the bioinformatician's ability to work with and transfer data to clinicians quickly and efficiently.

NetApp®

PetaGene
transparent lossless compression
Technology Partner

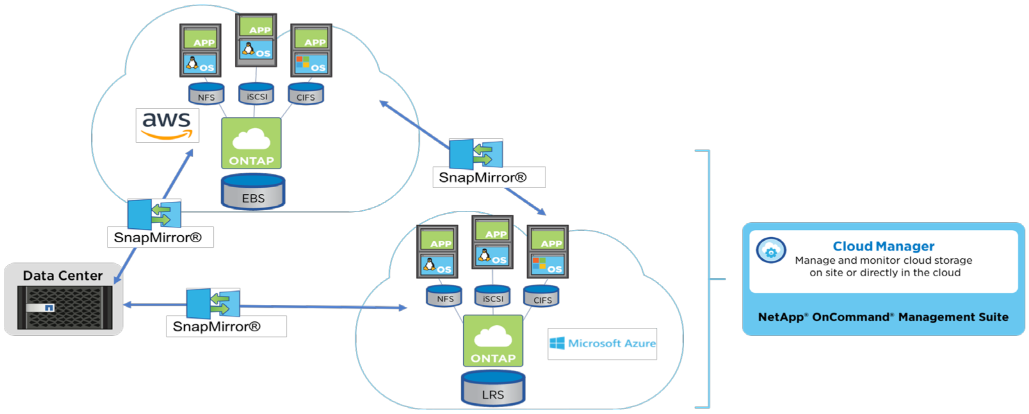

Figure 2: With the NetApp Data Fabric, ONTAP Cloud, and OnCommand Cloud Manager, it's easy to make genomic data available where it's needed. This capability is further enhanced with PetaGene's unique data compression technique, which improves performance and saves money.

The Advent of Precision or Personalised Medicine

Faster sequencing and more compact datasets have increased the number of individual sequences that can be performed. In this new world of personalised or precision medicine, individual patients and even their individual diseases, typically cancers, can be sequenced. (Many tumors have their own genetic makeup that can differ from that of the patient.) This sequencing brings great hope and opportunity for new insight, but it maintains the pressure on data capacity and increases the need to support privacy.

Beyond Storage Efficiency

Unlike generic data reduction techniques, PetaSuite understands the internal structure of genomics files. For lossless storage, PetaSuite offers cost reductions of up to 6:1 when compared with BAM or FASTQ.GZ files. This is a 96% data reduction in comparison with raw FASTQ.GZ files.

ONTAP Cloud and OnCommand Cloud Manager

NetApp ONTAP® Cloud storage software delivers enterprise quality control and protection for genomic data, with the flexibility to simplify the use of public cloud. ONTAP Cloud is a software-only service that offers a universal storage platform with easy-to-use file services (NFS, CIFS) and block services (iSCSI) common across all cloud and on-premises platforms. The SnapMirror® features of ONTAP offer a bandwidth-efficient data replication and transfer mechanism between clouds and to or from the data center.

To simplify the management experience, NetApp also offers OnCommand® Cloud Manager software, a centralised management environment for ONTAP Cloud software that fully supports hybrid storage environments.

Improve Analysis Speed

The PetaView command line file access system is lightweight, and I/O reductions dominate. Therefore using PetaView's on-the-fly random-access client-side decompression can actually speed up analysis, tools, and pipelines, especially in high-performance computing.

Storage Tiering

PetaSuite can exploit tiered storage by identifying and separating out unimportant NGS components to lower-cost tiered storage, while retaining important information in faster storage tiers. This capablity can integrate well with a hybrid FAS Solution and NetApp FabricPool.

About PetaGene

PetaGene was founded in Cambridge, the birth pace of genomics, to address the rapidly growing data management problems of the genomics industry. PetaGene's software enables compression of huge amounts of genomic data without compromising on access or data quality. The company's products go beyond regular data reduction techniques and have twice been recognised by Bio-IT World's Best of Show Award for their industry-leading performance and usability. For more information visit www.petagene.com or email sales@petagene.com.

About NetApp

NetApp is the data authority for hybrid cloud. We provide a full range of hybrid cloud data services that simplify management of applications and data across cloud and on-premises environments to accelerate digital transformation. Together with our partners, we empower global organisations to unleash the full potential of their data to expand customer touchpoints, foster greater innovation, and optimise their operations. For more information, email genomics@netapp.com. #DataDriven