DELL Technologies

Whitepaper

NEXT-GENERATION SEQUENCING DATA COMPRESSION WITH PETAGENE & DELL EMC POWERSCALE STORAGE

Amplifying Genomic Data Storage Without Disrupting Workflows

Abstract

This whitepaper introduces guidelines for utilizing the PetaGene PetaSuite application for next-generation sequencing (NGS) data compression with Dell EMC PowerScale storage systems. Together, PetaGene and Dell EMC deliver the highest possible storage density for NGS data. The combination also introduces opportunities to accelerate next-generation sequencing analysis pipelines and to optimize data transfers between storage end-points. Life science research directors, system administrators, and bioinformaticians who manage genomic data associated with next-generation DNA sequencing (NGS) workloads are encouraged to read this paper.

October 2020

Table of contents

Revisions

DATE	DESCRIPTION	AUTHOR
August 2017	Initial Release	Glen Otero
May 2020	Updated timings, benchmarks, guidelines.	E. Sasha Paegle

Introduction

Next-Generation Sequencing (NGS) is a combination of laboratory instrumentation technologies and analysis methods to identify patterns in DNA, the code of life, at dramatically increased resolution and quality.

The latest NGS instrumentation produces five times more data than the previous generation of instrumentation. As the adoption of sequencing DNA continues to increase, organizations like the Global Alliance for Genomics and Health (GA4GH) estimate that over 60 million patients will have their Genome (i.e., DNA) sequenced in a healthcare context by 2022 (Birney, 2017). Performing whole genome sequencing for those 60 million patients would require at least 5 exabytes of data storage for the 'raw,' un-analyzed NGS instrument data (Figure 1). The actual amount of data storage needed for 60 million patients is likely to be 3 to 4 times higher if the estimate accounted for the storage required for analysis results, intermediate files, and the data management practices related to NGS workflows.

Placing NGS data on different storage technologies is a universal data management strategy that many healthcare and life science organizations use to optimize usage of data storage while operating on fixed IT budgets. However, as these organizations transition to the latest generation of NGS equipment, they will need to introduce additional data management strategies if they want to keep pace with the accelerating pace of data generation.

Incorporating data compression is a data management strategy that can significantly impact the utility of available data storage resources. PetaGene, a company that specializes in NGS data compression, can reduce the storage footprint of NGS datasets in FASTQ and BAM format on average by 60% while preserving the genotyping accuracy.

This white paper describes best practices for using Petagene PetaSuite, an application for NGS data compression, with Dell EMC PowerScale Scale-Out NAS. Leveraging PowerScale and PetaSuite together provides significant storage capacity savings while also introducing other benefits such as reduced data transfer times and faster analysis times.

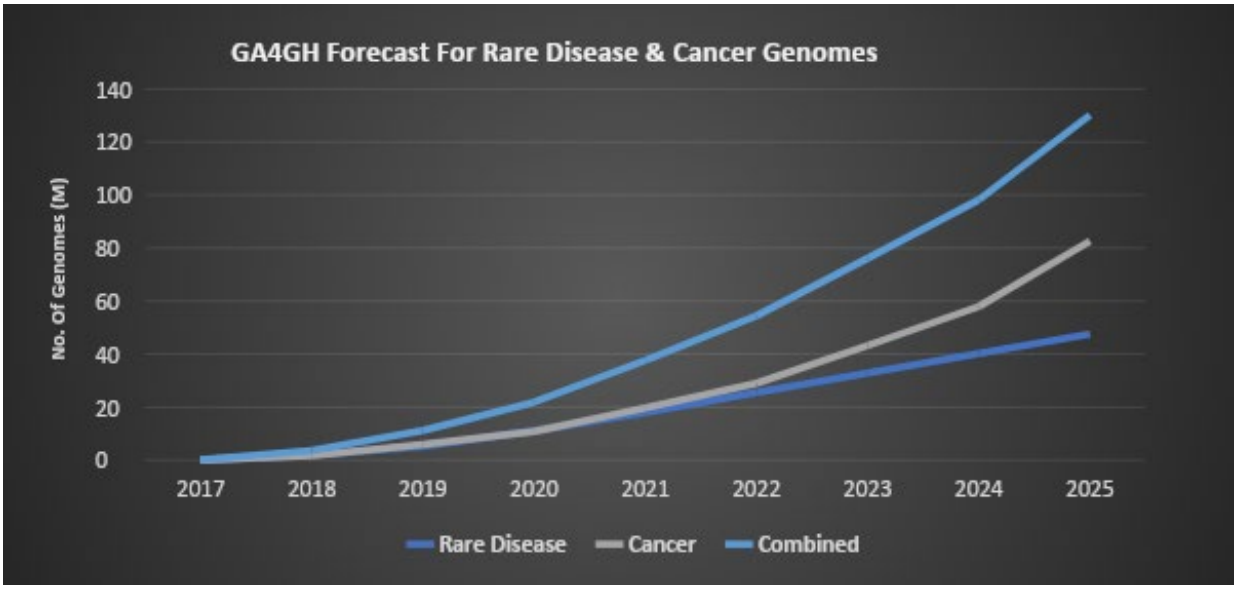

Figure 1.GA4GH FORECAST FOR SEQUENCING 60M PATIENTS BY 2022 REQUIRING 5 EXB OF DATA STORAGE.

Key Components to Amplify Storage of NGS Data

Maximizing the number of genomes stored per terabyte of data storage relies on three core components: data compression software, computing resources to perform data compression within a reasonable timeframe and a storage resource. The benchmarks and guidelines presented combined capabilities from PetaGene and Dell Technologies.

PetaGene PetaSuite

PetaSuite is a data compression capability for genomic data from PetaGene. The toolkit consists of a command-line tool for compression and a user-mode library for read-back of compressed data. The PetaSuite command-line tool, compresses FASTQ, BAM, and CRAM files. The user-mode library, PetaLink, enables any NGS analysis tool, application, or pipeline to access the compressed data files transparently in their native formats.

By servicing requests with virtual files from the file, the system creates a native NGS file transparency without requiring end-users to modify their existing scripts or analysis pipelines. For example, PetaSuite compresses **my.FASTQ.gz** and replaces it with a **my.fasterq** file and removes the original **my.FASTQ.gz** file to recover storage capacity. With the PetaLink library loaded, users and applications access a virtual **my.FASTQ.gz** file in the same manner as the original file. Virtual files do not exist on the filesystem or consume inode resources. The PetaLink library represents virtual files as symbolic links to the underlying PetaSuite fasterq files and very efficiently translates the compressed data to native FASTQ data.

The result is a smaller, compressed file with the same filename and access methods. Decompressing these files is not necessary since the read-back is automatic and transparent when an application accesses the file. The efficiency of PetaLink read-back of PetaGene compressed data is 2-3 times faster than the original BAM or FASTQ data.

Benchmarking exercises used PetaSuite Cloud Edition V1.2.6 installed on a Dell PowerEdge R740 server. See the Appendix for more information on PetaGene PetaSuite.

Dell PowerEdge R740 Server

The Dell PowerEdge R740 Server is part of Dell's 14th Generation server portfolio and designed for HPC and NGS workloads. The 2U Intel Skylake based rack mount server provides an ideal balance between storage, I/O, and application acceleration. During preliminary PetaSuite data compression tests, three different chipsets were evaluated. The Intel® Xeon® Gold 6240 processor provided the best price per performance profile. Descriptions of the other chipsets are in the Appendix.

Dell EMC PowerScale & OneFS

Dell EMC PowerScale is a proven Scale-Out network-attached storage (NAS) solution that can handle the unstructured data prevalent in end-to-end NGS workflows. The PowerScale storage architecture automatically aligns with application requirements for performance, capacity, and economics. As performance and capacity demands increase, both can be scaled non-disruptively, allowing sequencing applications and users to continue working.

The OneFS file system powers the PowerScale storage cluster. For many Life Science organizations, the PowerScale storage with OneFS offers a reliable, easy to manage, and cost-effective storage solution that balances between performance and capacity needed to support NGS secondary analysis and other heterogeneous bioinformatics workflows. OneFS exposes data through multiple network protocols such as CIFS/SMB, NFS, HDFS, and HTTPS and S3 to any number of machines by any number of users leveraging existing authentication services without additional gateway appliances. For example, an PowerScale system accepts raw data from NGS instruments over the Windows SMB protocol. OneFS serves the raw NGS data over the NFS protocol to a secondary analysis pipeline running on a high-performance computing cluster. The secondary analysis pipeline writes its results back to the same PowerScale cluster, where a downstream tertiary analysis workflow using SPARK accesses the results data over HDFS. OneFS streamlines workflows by eliminating the need to move or copy data.

Using PetaSuite and PowerScale for NGS Data Compression

PetaSuite enables the transformation of BAM and FASTQ.gz data to more efficiently compressed formats without disrupting established pipelines that require the original file name and format. For example, upon importing a BAM file, PetaSuite converts it to the PGBAM format, validates that all the read data in the BAM file is preserved, and removes the original BAM file to recover storage space. A virtual BAM file is then made available in the same directory to be used by Linux toolchains, analysis pipelines, and genome browsers, just like the original BAM file. The PetaLink library seamlessly manages access to the PGBAM file via the virtual BAM file pointer.

PetaSuite & PowerScale Compression Scenarios

Data compression, decompression, and high-throughput test cases using sample FASTQ.gz and BAM files identified PetaSuite and PowerScale best practices.
The test environment consisted of a four-node Isilon (PowerScale) H500 storage cluster running OneFS 8.1.2 with default round-robin SmartConnect settings. PetaSuite installed on a Dell PowerEdge R740XD containing two Intel® Xeon® Gold 6240 Processors and 192 GB of RAM accessed and compressed the sample NGS data on the H500 system via NFS v3 over 10 GbE (Gigabit Ethernet) connections. Hyperthreading was enabled, providing 72 compute cores. Sample FASTQ.gz and BAM files representing human whole-genome sequencing (WGS) data were downloaded from the Garvan Institute of Medical Research.

Lossless Compression

Table 1 summarizes the compression results when running PetaSuite lossless compression mode on FASTQ.gz files using a single compute node:

petasuite --compress --v off --dstpath/compressed NA12878_V2.5_Robot_1_R1.FASTQ.gz

For this sample data, PetaSuite compressed 65% of the FASTQ.gz files on average. Each compression job completed in approximately 2 minutes. The example command maintains the original FASTQ.gz files. Reclaiming storage capacity requires the **--replace** option, which removes the original FASTQ.gz file.

PetaSuite compression preserves all the data in a FASTQ.gz file using integrated validation software. To validate the first million reads in a FASTQ.gz file, use the option **--validate quick** (the default). Compress with the **--validate full** option to activate full validation. The **--dstpath** option streams the compressed output files to a target storage destination such as the current file system OneFS, local NVMe drive on the compute node, or a cloud storage bucket hosted at AWS, Azure or GCP.

TABLE 1. FASTQ File Compression Using PetaSuite Lossless Mode

SAMPLE FILES	ORIGINAL FILE SIZE(GB)	COMPRESSED FILE SIZE (GB)	% FILE SIZE REDUCTION	COMPRESSION TIME (min.)
NA12878_V2.5_Robot_1_R1.fastq.gz	36.9	10.6	71.2%	1.9
NA12878_V2.5_Robot_1_R2.fastq.gz	42.8	15.9	62.8%	1.9
NA12878_V2.5_Robot_2_R1.fastq.gz	42.7	12.0	71.9%	2.3
NA12878_V2.5_Robot_2_R2.fastq.gz	53.1	21.6	59.4%	2.3
NA12878_V2.5_Robot_3_R1.fastq.gz	38.6	11.3	70.8%	2.0
NA12878_V2.5_Robot_3_R2.fastq.gz	45.8	17.6	61.5%	2.1
NA12878_V2.5_Robot_4_R1.fastq.gz	39.6	11.1	72.1%	2.1
NA12878_V2.5_Robot_4_R2.fastq.gz	47.9	18.4	61.6%	2.1
AVERAGE	**43.4**	**14.8**	**66.4%**	**2.1**

Table 2 summarizes the compression results when running PetaSuite lossless compression mode on four sample BAM files using a single compute node.

petasuite --compress --v off --dstpath/compressed NA12878_V2.5_Robot_1.bam

PetaSuite reduced the sample BAM files by 80% and averaged 26 minutes per BAM file. Using the --replace option with the four sample BAM files reclaimed 618 GB of storage.

TABLE 2. BAM File Compression Using PetaSuite Lossless Mode

SAMPLE FILES	ORIGINAL FILE SIZE(GB)	COMPRESSED FILE SIZE (GB)	% FILE SIZE REDUCTION	COMPRESSION TIME (min.)
NA12878_V2.5_Robot_1.dedup.realigned .recalibrated.bam	173.67	31.2	82.0%	24.4
NA12878_V2.5_Robot_2.dedup.realigned .recalibrated.bam	204.72	39.1	80.9%	28.9
NA12878_V2.5_Robot_3.dedup.realigned .recalibrated.bam	187.23	33.2	82.3%	24.1
NA12878_V2.5_Robot_4.dedup.realigned .recalibrated.bam	190.9	35.0	81.7%	29.0
AVERAGE	**189.1**	**34.6**	**81.7%**	**26.6**

PetaSuite Bayescal Compression

The BayesCal compression mode calculates a complete posterior estimate of the sequencing error in the data and improves the quality scores associated with sequencing data. To achieve maximal data compression and improve genotyping accuracy compared to lossless compression, use the --bqfilt option.

petasuite -compress -m bqfilt --replace --dstpath/compress NA12878_V2.5_Robot_1_R1.FASTQ.gz

TABLE 3. FASTQ File Compression Using PetaSuite Bayescal Mode

SAMPLE FILES	ORIGINAL FILE SIZE(GB)	COMPRESSED FILE SIZE (GB)	% FILE SIZE REDUCTION	COMPRESSION TIME (min.)
NA12878_V2.5_Robot_1_R1.fastq.gz	36.9	4.9	86.7%	108
NA12878_V2.5_Robot_1_R2.fastq.gz	42.8	8.3	80.7%	158
NA12878_V2.5_Robot_3_R1.fastq.gz	38.6	5.5	85.8%	115
NA12878_V2.5_Robot_4_R1.fastq.gz	39.6	5.1	87.1%	119
AVERAGE	**39.5**	**5.9**	**85%**	**125**

Tables 3 and 4 summarize the results of compressing FASTQ.gz and BAM files with the Bayescal option. PetaSuite compressed 85% of the FASTQ.gz files and required approximately 125 minutes to compress each file. Using the **--Baysecal** option with the sample BAM files achieved nearly [90%} compression savings.

TABLE 4. BAM File Compression Using PetaSuite Bayescal Mode				
SAMPLE FILES	ORIGINAL FILE SIZE (GB)	COMPRESSED FILE SIZE (GB)	% FILE SIZE REDUCTION	COMPRESSION TIME (min.)
NA12878_V2.5_Robot_1.dedup.realigned .recalibrated.bam	173.7	18.45	89.4%	229
NA12878_V2.5_Robot_2.dedup.realigned .recalibrated.bam	204.7	23.97	88.3%	295
NA12878_V2.5_Robot_3.dedup.realigned .recalibrated.bam	187.2	19.58	89.5%	245
NA12878_V2.5_Robot_4.dedup.realigned .recalibrated.bam	190.9	20.9	89.1%	263
AVERAGE	**189.1**	**20.7**	**89.1%**	**258**

The results demonstrate that BayesCal compression of FASTQ.gz files frees up approximately 2X more capacity than lossless compression. However, the time taken to complete BayesCal compression does require additional time. Similarly, BAM files used with the BayesCal compression generates files that are approximately 30% smaller versus lossless compression. Like FASTQ files, using BayesCal compression does require additional time.

Decompression

Reconfiguring analysis pipelines to decompress Petagene compressed NGS data is not necessary. PetaSuite creates virtual FASTQ.gz and BAM files that point to compressed PetaSuite fasterq and PGBAM files. Analysis pipelines can continue to call the original file names. At the same time, PetaLink translates the fasterq and PGBAM files in the background without adding any significant overhead to the analysis pipeline wall-clock time. However, if decompression is necessary, the time decompress is approximately one-fourth of the compression time of the original file.

Best Practices

Organizations are encouraged to apply and adapt the best practices outlined below to their NGS workflows to obtain the maximum benefit from using PetaSuite along with PowerScale storage systems.

Forecast Compression Savings With Your Data

Storage capacity reclaimed from using PetaSuite compression methods does depend on the NGS sequencing instrument on which data is generated and analysis methods used to process the NGS sequencing data. For example, the popular BWA-GATK secondary analysis pipeline developed by the Broad Institute produces multiple output BAM files depending on the required quality parameters. Using PetaSuite lossless mode reclaims 68% storage capacity for the original BAM file, whereas the lossless mode reclaims over 90% storage capacity when compressing a downstream, 'recalibrated' BAM file. Therefore, it is best to compress data sets produced by the organization to achieve a more accurate storage capacity savings forecast.

Ensure File Transparency With Proper Installation and User Environment Configuration

PetaSuite installation is straightforward via an rpm or deb file. By default, the **petalink.so** library is installed in **/usr/lib**. PetaLink starts manually as well as automatically. For a manual startup, specify **LD_PRELOAD=/usr/lib/petalink.so** before starting a PetaSuite command. For example, start a bash instance with PetaLink loaded run:

LD_PRELOAD=/usr/lib/petalink.so bash

This instance of bash and any commands executed from within this bash instance runs with the PetaLink library. This bash instance does not affect other instances of bash or other processes, which means that it can be beneficial to run PetaLink automatically instead.

For automatic startup, modify a startup script to define this environment variable. For example, add the following line to **~/.bashrc**:

export LD_PRELOAD=/usr/lib/petalink.so

Defining the environment variable ensures that PetaLink is loaded whenever a bash shell starts.

PetaSuite utilizes a corpus to aid in compression and decompression. Petagene strongly recommends installing the human corpus at a minimum by using the command:

sudo petasuite_install_corpus human

The corpus is installed in **/opt/petagene/petasuite/species** by default. If the command is run without super-user privileges the corpus is extracted into a directory other than the default system directory by explicitly specifying an installation path:

sudo petasuite_install_corpus human /home/user/path

To use PetaSuite with an alternative location, set the **PETASUITE_REFPATH** environment variable to point to a new path. For example,

export PETASUITE_REFPATH=/home/user/path

The human corpus is 24 GB in size, and there are 70 other species available. To install all species, use:

sudo petasuite_install_corpus all

The installed "auto" species database auto-detects which species the compressed file belongs to. Be aware that installing all the corpora requires 100 GB of storage. Installing individual species corpora is possible. To download and install a specific species corpus other than human, replace "human" with the desired species name.

High Throughput Data Compression

Since compression of large files with PetaSuite can take several hours, we recommend using a Linux cluster to automate the compression of multiple files in parallel. PetaSuite provides support for distributed tasks by splitting up all the files in a directory into separate tasks. For example, if there are 100 files to process in a directory, by specifying the option **--numtasks 4**, PetaSuite can split the files into four tasks of approximately 25 files each. so, the four tasks need to be submitted separately with the **--taskid** number set with arguments like so:

petasuite -c --numtasks 4 --jobid 1 /isilon/petagene

petasuite -c --numtasks 4 --jobid 2 /isilon/petagene

petasuite -c --numtasks 4 --jobid 3 /isilon/petagene

petasuite -c --numtasks 4 --jobid 4 /isilon/petagene

PetaSuite works well with the SLURM job scheduler[1] by detecting the **SLURM_ARRAY_TASK_ID** environment variable and using it for the jobid instead. A distributed PetaSuite job is submitted to SLURM using:

sbatch --arrayt=1-4 --wrap="petasuite -c --numtasks 4 /isilon/petagene"

Note how the **--jobid** option is no longer needed in this case.

[1] PetaSuite does not submit jobs to LSF or other cluster job management systems.

A typical directory may contain a mix of BAM and FASTQ.gz files to be compressed. In the directory listing below there are 10 BAM and FASTQ.gz files totaling nearly 800GB:

$ ls /isilon/petagene

-rwxrwxrwx 1 nfsnobody root 162G Feb 1 2016 NA12878_V2.5_Robot_1.bam

-rwxrwxrwx 1 nfsnobody root 35G Jan 24 2016 NA12878_V2.5_Robot_1_R1.FASTQ.gz

-rwxrwxrwx 1 nfsnobody root 40G Jan 24 2016 NA12878_V2.5_Robot_1_R2.FASTQ.gz

-rwxrwxrwx 1 nfsnobody root 40G Jan 24 2016 NA12878_V2.5_Robot_2_R1.FASTQ.gz

-rwxrwxrwx 1 nfsnobody root 50G Jan 24 2016 NA12878_V2.5_Robot_2_R2.FASTQ.gz

-rwxrwxrwx 1 nfsnobody root 175G Feb 1 2016 NA12878_V2.5_Robot_3.bam

-rwxrwxrwx 1 nfsnobody root 36G Jan 24 2016 NA12878_V2.5_Robot_3_R1.FASTQ.gz

-rwxrwxrwx 1 nfsnobody root 178G Feb 1 2016 NA12878_V2.5_Robot_4.bam

-rwxrwxrwx 1 nfsnobody root 37G Jan 24 2016 NA12878_V2.5_Robot_4_R1.FASTQ.gz

-rwxrwxrwx 1 nfsnobody root 45G Jan 24 2016 NA12878_V2.5_Robot_4_R2.FASTQ.gz

This command was used to submit the batch job to the SLURM job scheduler:

sbatch --array=1-12 --wrap="petasuite -c -m bayescal ---replace --numtasks 20 /isilon/petagene”

The compressed files look like this:

$ ls /isilon/petagene

-rwxrwxrwx 1 nfsnobody root 104G Feb 1 2016 NA12878_V2.5_Robot_1.cram

-rwxrwxrwx 1 nfsnobody root 1.7M Feb 1 2016 NA12878_V2.5_Robot_1.cram.PG.idx

-rwxrwxrwx 1 nfsnobody root 9.9G Jan 24 2016 NA12878_V2.5_Robot_1_R1.fasterq

-rwxrwxr-x 1 nfsnobody root 15G Dec 17 2016 NA12878_V2.5_Robot_1_R2.fasterq

-rwxrwxrwx 1 nfsnobody root 122G Feb 1 2016 NA12878_V2.5_Robot_2.cram

-rwxrwxrwx 1 nfsnobody root 2.0M Feb 1 2016 NA12878_V2.5_Robot_2.cram.PG.idx

-rwxrwxrwx 1 nfsnobody root 12G Jan 24 2016 NA12878_V2.5_Robot_2_R1.fasterq

-rwxrwxrwx 1 nfsnobody root 21G Jan 24 2016 NA12878_V2.5_Robot_2_R2.fasterq

-rwxrwxrwx 1 nfsnobody root 112G Feb 1 2016 NA12878_V2.5_Robot_3.cram

-rwxrwxrwx 1 nfsnobody root 1.8M Feb 1 2016 NA12878_V2.5_Robot_3.cram.PG.idx

-rwxrwxrwx 1 nfsnobody root 11G Jan 24 2016 NA12878_V2.5_Robot_3_R1.fasterq

-rwxrwxrwx 1 nfsnobody root 17G Jan 24 2016 NA12878_V2.5_Robot_3_R2.fasterq

-rwxrwxrwx 1 nfsnobody root 113G Feb 1 2016 NA12878_V2.5_Robot_4.cram

-rwxrwxrwx 1 nfsnobody root 1.9M Feb 1 2016 NA12878_V2.5_Robot_4.cram.PG.idx

-rwxrwxrwx 1 nfsnobody root 11G Jan 24 2016 NA12878_V2.5_Robot_4_R1.fasterq

-rwxrwxrwx 1 nfsnobody root 18G Jan 24 2016 NA12878_V2.5_Robot_4_R2.fasterq

Notice that PetaSuite automatically created the BAM index files after creating the PGBAM files. PetaSuite compressed the directory nearly 4x from 798 GB to 204 GB, or 25% of the original directory size, using the BayesCal modality and removing the original files.

Data Integrity

While reducing storage space up to 90% with lossless compression of NGS data is compelling in terms of reclaiming storage capacity, some organizations may have concerns about data integrity, movement, or deletion in production workloads. However, PetaSuite is designed not to disrupt production workloads.

PetaSuite preserves a perfect representation of the original data in compressed form. PetaSuite retains the read data within the BAM or FASTQ file and the compression wrapper of the original gzip file. Preserving the read data and wrapper creates a bit-for-bit preserved representation of the original data that is verifiable with an MD5 or SHA256 hashing algorithm.

The integrity of the compressed data is automatically checked during compression using a validation step. Validation dynamically decompresses the PetaSuite compressed file and validates it against the original file. If the validation successfully passes, PetaSuite commits and writes the compressed file(s) to storage.

The PetaSuite virtual files are direct replacements of the original files. Replacing a BAM or FASTQ.gz file with a compressed and validated PGBAM or FasterQ file is the same as replacing it with a copy of the original file. With PetaLink enabled, reconfiguring production workloads is not required. The preservation of filenames allows users to confidently compress in place and run their analysis tools and pipelines without disruption or changes in the data.

When to Compress?

The NGS data compression step can be introduced during the data generation, analysis, and archive phases of the data life cycle. Determining when to compress NGS data during its life cycle is a shared responsibility between the IT team and end-users. A shared evaluation of laboratory workflows, analysis pipelines, and data management practices can identify one or more opportunities to introduce NGS data compression.

For example, the NGS laboratory team could compress the FASTQ data as part of their data QA practices during the data generation phase. Besides recapturing active storage capacity, downstream analysis teams would benefit from reduced data transfer times and accelerated analysis pipelines as they use the virtual PetaSuite FASTQ (i.e., FASTERQ) files.

Data analysis teams could introduce lossless mode compression as an explicit step in their analysis pipelines. Research groups conducting large, multi-year population-based studies can minimize the frequency of data transfers between high performance and slow archive storage using this strategy. PetaSuite lossless compression allows the group to maintain virtual BAM files on high-performance storage at a lower cost and quickly perform re-analysis of BAM files as the project adds new samples to the population cohort over time.

PetaSuite compression also plays a role in organization-wide data archiving and data retention policies. For example, a genomics department at a children's hospital could submit native FASTQ and BAM files to a PetaSuite compression service before tiering the NGS data to archive nodes or other storage end-points like ECS object storage or cold cloud storage.

Summary: Amplify Storage of Genomics Data by 10X

Using PetaSuite, along with PowerScale storage, introduces opportunities to reclaim storage capacity, reduce data transfer times, and accelerate analyses.

For the IT organization measuring data storage capacity, "saved" is one way to measure the impact of introducing NGS data compression. However, "saving" a resource like data storage capacity can also imply limiting its use. A more impactful way to data compression is its ability to amplify the amount of *genomic information* stored per terabyte of storage. (Continued on next page)

Consider the Illumina Novaseq 6000, the latest high throughput NGS instrument. At optimal utilization, a Novaseq 6000 requires at least one petabyte (PB) of storage annually. One petabyte of usable storage stores approximately 5000 Human whole-genome sequences (WGS) at 30X coverage in the BAM file format. By applying PetaSuite lossless mode compression to those BAM files, the PowerScale storage capacity for WGS increases by 6X from 5,000 to nearly 30,000 WGS per PB. Using the BayseCal compression mode further amplifies the PowerScale storage capacity to ~50,000 WGS per PB (Figure 2).

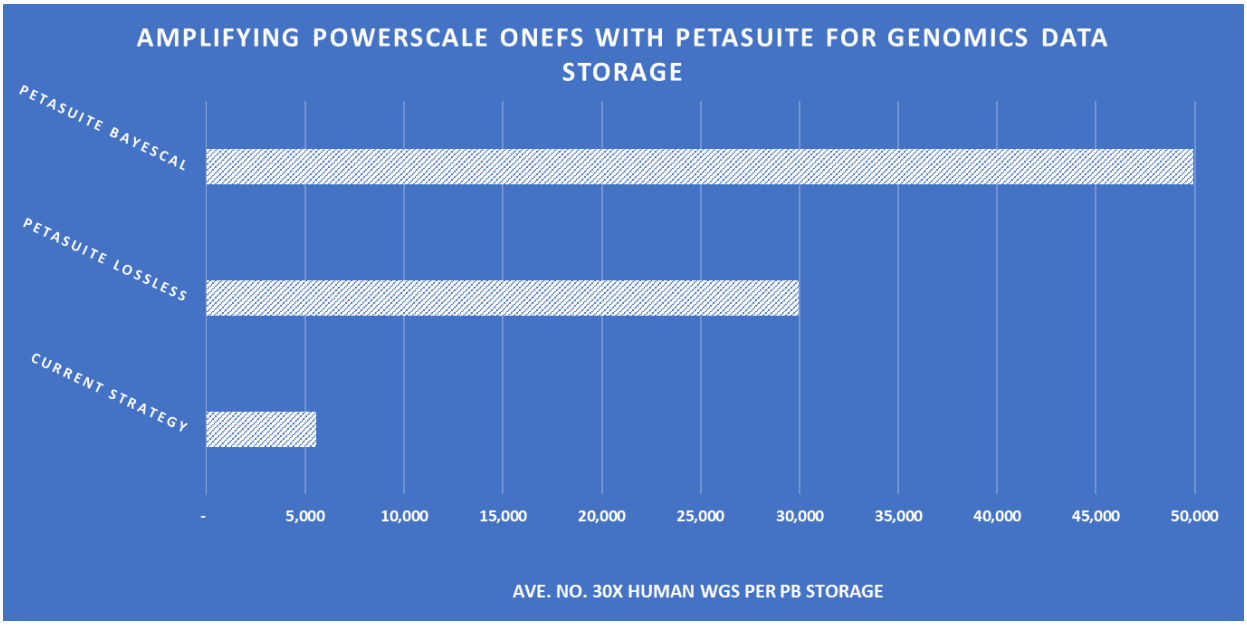

Figure 2. AMPLIFYING POWERSCALE ONEFS WITH PETASUITE

For More Information

To learn more about how your organization can take use of PetaSuite along with Dell EMC PowerScale storage solutions to amplify your genomics data storage requirements, review the references noted in the Appendix and contact your Dell Technologies representative.

Appendix

References

Birney, E. (2017). Genomics in healthcare: GA4GH looks to 2022. Retrieved from https://doi.org/10.1101/203554

PetaGene PetaSuite

Information about PetaGene and PetaSuite are available from www.petagene.com. PetaSuite downloads are available from https://www.petagene.com/downloads/. The PetaSuite software was installed using the recommended default settings.

Dell EMC PowerScale

Dell EMC PowerScale scale-out storage solutions are robust, yet simple to scale and manage, no matter how large your unstructured data environment becomes.

An overview of Isilon Generation 6 Hybrid storage capabilities, features and options are summarized here: https://www.dellemc.com/en-za/collaterals/unauth/data-sheets/products/storage/h16071-ss-isilon-hybrid.pdf

For a detailed description of OneFS, the PowerScale storage file system, see Dell EMC PowerScale OneFS Technical Overview posted here: https://www.emc.com/collateral/data-sheet/h16071-ss-isilon-hybrid.pdf

Dell EMC PowerScale (Isilon) H500 Configuration

DELL EMC ISILON H500 STORAGE CLUSTER: 4TB HDD, 2x 3.2 TB SSD 40GbE/40GbE	
CHASSIS TYPE & NODE COUNT	H500 / 4 Nodes
USABLE (RAW) CAPACITY – TB	192 TB (240 TB)
SSD CAPACITY / NODE	6.2 TB (76.8 TB total)
PROCESSORS / NODE	2.2 GHz, 10-core
MEMORY / NODE	128 GB (1.5 TB total)
FRONT-END NETWORKING	2 x 40 GbE
BACK-END NETWORKING	2 x QSFP+40 GbE Ethernet
OPERATING SYSTEM	ONEFS v8.1.2
ISILON SMARTCONNECT	ROUND ROBIN MODE
SMARTREAD / L3 CACHE	TRUE (enabled, default)
SNAPSHOTS	None
DATA PROTECTION	N+2:1 (default)

Dell PowerEdge R740

A detailed specification of the PowerEdge R740 server is available from https://www.dell.com/en-us/work/shop/povw/poweredge-r740. Preliminary PetaSuite data compression times were generated on a dual socket R740 with 128 GB RAM and 1 TB NVMe flash drive using either the Intel® Xeon® Gold 6240 processor, Intel® Xeon® Gold 8268 processor, or AMD EPYC™ 7742 processor. Hyperthreading was enabled for each test case.

Sample Data

The sample Illumina Hiseq X Ten sequencing data were downloaded from the Garvan Institute Kinghorn Center for Clinical Genomics (The Garvan) in Sydney, Australia. The Garvan was one of the first three organizations in the world to acquire the Illumina HiSeq X Ten sequencing system. To enable the scientific community to assess data quality by an independent laboratory, the Garvan made reference datasets available. The sample WGS data were generated from the popular Coriell Cell Repository NA12878 reference sample, which has been extensively analyzed by the Genome in a Bottle Consortium. The depth of coverage for the sample data averages 36X.

The data can be downloaded from:

https://www.garvan.org.au/research/kinghorn-centre-for-clinical-genomics/sequencing-services/ref-data-tables