PetaGene transparent lossless compression **PetaSuite**

대용량 유전체 데이터 압축 솔루션

NGS 데이터를 압축해서 데이터 보관, 유지, 전송 비용을 줄이세요.

Key features

- Cost Savings
- Lossless Streaming Compression
- Storage Optimization
- Faster Transfer
- Transparent Usage
- Speeds up Analysis
- NGS Quality Score Refinement
- No lock-in
- Easy IT Deployment

Cost Savings

- 대용량 유전체 데이터 파일 압축을 통하여 스토리지 증설, 운영 및 백업 비용 약 60% ~ 90% 절감

Speeds up Analysis

- 압축된 파일을 풀지 않고 그대로 분석 가능
- I/O 시간을 줄여 분석 속도 증가

Faster Transfer

- 파일이 압축된 만큼 데이터 전송 시간 최대 90% 단축

No lock-in

- 압축 해제 시 라이선스 불필요

Lossless Streaming Compression

- 무손실 최적화 압축을 통한 안전한 데이터 보존

NGS Quality Score Refinement

- Bayesian 방식 (posterior probability)을 통한 NGS Quality Score 개선
- BayesCal 모드 사용 시 압축률 약 30~70% 증가

Transparent Usage

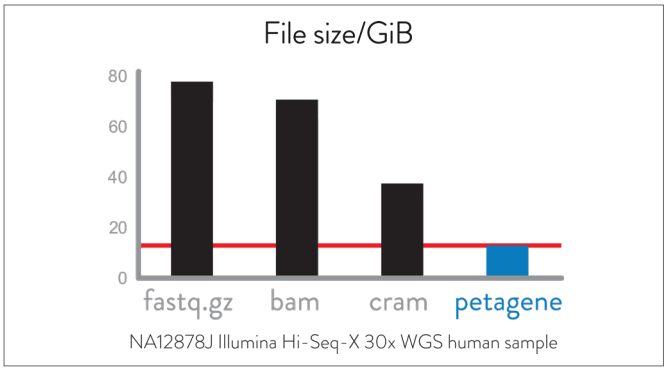
- 압축 파일을 Linux 명령어, 분석 파이프라인 및 유전체 브라우저에서 원본 파일처럼 동일하게 사용

Easy IT Deployment

- 소프트웨어 사용자 수 무제한
- 손쉬운 소프트웨어 설치

NGS 데이터에 특화된 압축 솔루션

PetaSuite는 NGS 데이터에 특화된 압축 솔루션으로써, bam 파일이나 gzip 압축된 fastq 파일을 추가 압축하여 **60~90%나 되는 압축률**을 보여줍니다. PetaSuite는 멀티코어를 활용하여 **290+ Mbytes/sec (4 core – i7, 3GB)**의 속도로 NGS 데이터의 압축을 수행할 수 있는 강력한 솔루션입니다.



NA12878J Illumina Hi-Seq-X 30x WGS human sample

압축 해제 없이 분석 가능

PetaSuite를 이용하여 압축된 BAM 파일은 다시 압축을 해제하지 않고도 IGV나 bamtools, GATK, Picard, Strelka2, PySAM등의 분석 프로그램에서 그대로 활용할 수 있습니다. 파일 크기가 줄어든 만큼 저장공간을 효율적으로 활용할 수 있을 뿐만 아니라 분석 속도도 훨씬 줄어들어 시간, 저장공간 모두를 절약할 수 있습니다.

〈압축파일 그대로 분석 가능한 프로그램 목록〉

Tool	Application type	Tool	Application type
samtools	Toolkit	GATK 4	Pipeline
bamtools	Toolkit	Manta	Variant caller
bcftools	Variant caller	Picard	Toolkit
bedtools	Toolkit	PySAM	Toolkit
BWA-MEM	Mapper	Sambamba	Toolkit
bwa-mem2	Mapper	seqtk	Toolkit
GATK 3	Pipeline	Strelka2	Variant caller

파일 크기가 줄면 비용도 절감

NGS 데이터의 압축은 단순히 저장공간의 확보만을 의미하지 않습니다. PetaSuite를 이용하면 physical 레벨에서 요구되는 더 많은 공간 확보 비용과 스토리지 소유총비용(TCO: total cost of ownership), 백업 공간 확보에 필요한 총 3.5배의 비용을 절감할 수 있습니다.

국외 도입 사례

스웨덴, Gothenburg 임상 유전체 연구소

Gothenburg 임상 유전체 연구소는 후천성 질환, 유전성 질환 및 미생물학 분야의 임상 연구를 수행하고 있으며, 약 50TB에 달하는 WES, WGS, deep panel 데이터를 BAM 파일 형식으로 자체 구축한 스토리지에 보유하고 있었습니다. 이들이 보유한 데이터는 원격으로 접속한 Sahlgrenska 의과대학의 연구자 또는 임상의들에게 Apache 웹서버와 IGV(유전체 데이터 뷰어)를 통해 시각화되어 제공됩니다.
2019년, 연구 트랜드 변화로 인하여 WGS 데이터에 대한 수요가 증가함에 따라, Gothenburg 임상 유전체 연구소장인 Per Sikora는 현재 인프라와 분석 도구 및 파이프라인을 유지하면서 앞으로 생산될 데이터를 저장할 공간을 확보할 수 있는 솔루션을 찾기 시작했습니다.
이들은 PetaSuite를 도입하여 기존 50TB의 데이터를 1/7 수준으로 (6.5TB) 줄일 수 있었습니다(평균 압축률 87%). 또한, 압축된 파일은 별도의 해제 작업이나 마이그레이션 작업 없이 기존 파일 경로와 파일명 그대로 분석 파이프라인에서 이용 가능하여, 인적, 물적 리소스 낭비를 크게 줄일 수 있었습니다.

Case Study: NGS Compression++
Clinical Genomics Gothenburg

Genomics research facility reduces storage footprint to one-seventh of prior level

UNIVERSITY OF GOTHENBURG

PetaGene

The Background

The Clinical Genomics Gothenburg facility at the Sahlgrenska Academy, University of Gothenburg performs clinical research in the areas of acquired disease, hereditary disease and microbiology.

The facility also works with the Sahlgrenska University Hospital to expedite translational research. They use 10X, Illumina and IonTorrent sequencing platforms.

The Challenge

The Clinical Genomics Gothenburg team stores whole exome, whole genome and deep panel clinical data as BAM files on local object storage. These data need to be accessible for ad-hoc remote visualization tasks by researchers and clinicians using Integrative Genomics Viewer (IGV) over an Apache webserver.

The amount of data already in storage (50 TB), combined with the anticipated increase in need associated with the wholesale shift from exome to whole genome sequencing during 2019, was about to place the existing infrastructure under considerable strain. Per Sikora, Head of Facility at Clinical Genomics Gothenburg was tasked with resolving the issue. He set out to find a solution that would create as much capacity as possible in the current infrastructure but not require modifications to existing tools and pipelines.

Clinical Genomics Gothenburg's file sizes before and after compression with PetaSuite

The Solution

Clinical Genomics Gothenburg now compresses its data using PetaGene's PetaSuite compression software and stores it on the same local object storage as previously.

When IGV is used with an Apache server, Apache accesses the data in small chunks which can cause workability challenges for other tools. PetaGene's on-the-fly decompression library had to interact effectively with this Apache access pattern. The University of Gothenburg/Clinical Genomics team was pleased with PetaGene's responsiveness in hardening the PetaSuite product for deployment with Apache. The client found that PetaGene was open, technically adept and effective.

The Results

By using PetaGene's PetaSuite compression software, Clinical Genomics Gothenburg now achieves an average compression ratio of 87%. This means that the university's storage requirements are now less than one-seventh what they would be without compression. The clinicians use IGV to access the data in exactly the same way as they did before and there is no discernible impact on visualization performance. Even during the compression process there was no downtime, and uninterrupted access to the data was maintained throughout – users saw and accessed either an uncompressed file or the equivalent PetaGene virtual file with the original filename in the original location, depending on whether the file had been compressed yet.

The university chose to use the optional BayesCal feature, as their main objective was to compress their data to the maximum extent. However, an additional welcome benefit is an increase in genotyping accuracy across the entire receiver operating characteristic (ROC) curve by using BayesCal.

Per Sikora, Head of Facility commented

"By using PetaSuite compression software for our data we have achieved our primary aim of dramatically increasing our storage capacity. This means that we do not need to spend precious resources on replacing or adding to it. The PetaGene team were responsive to our needs, including managing the to efficiently access the compressed data via Apache server the data first."

+44 (0) 1223 655651 | petagene.com

Customers and Collaborators

National Institute of Health Republic of Korea
NHS
University of Cambridge
King's College London
University of Gothenburg
University du Luxembourg
東京大学 The University of Tokyo
Baylor College of Medicine
Luxembourg Centre for Systems Biomedicine
Princess máxima center pediatric oncology
Sentieon
Genestack
Precision Health Informatics
AstraZeneca
intel
Google Cloud Partner
DELL EMC
NetApp ALLIANCE PARTNER
WEKA.IO World's Fastest File System

소프트웨어 및 생물정보 분석 서비스 문의 **consulting@insilicogen.com** **031.278.0061**