# NGS Compression++

REDUCING THE SIZE AND COST OF NGS DATA STORAGE AND TRANSFER

BEST OF SHOW Bio-IT World Conference & Expo '16 WINNER

BEST of SHOW WINNER Bio-IT World Conference & Expo 18

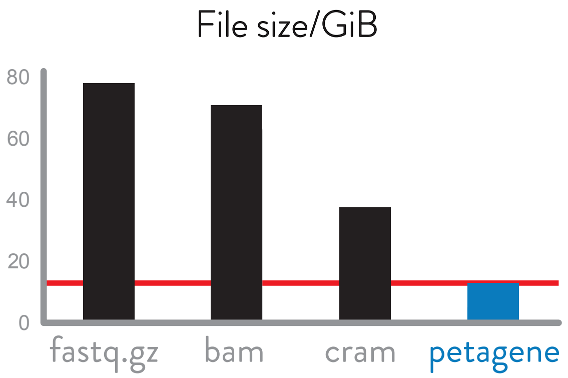
# PetaSuite

## Overview

PetaSuite is a powerful command line compression tool that significantly reduces the size and cost of NGS data for storage and transfer.

## Significant cost reductions

Unlike generic software, PetaSuite understands the internals of genomics files. For lossless storage, PetaSuite offers compression savings of **between 60% and 90% for BAM or gzipped FASTQ files.**



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

## Transparent usage

PetaSuite lets researchers, engineers and clinicians continue using their FASTQ and BAM files in their existing tools and pipelines. It integrates into existing storage infrastructures to provide transparent compression and access.

## Tiered storage

PetaSuite efficiently exploits tiered storage by identifying and pushing unimportant NGS components out to lower-cost tiered storage, while retaining important information in faster storage tiers. This reduces I/O load, and boosts overall speed while also lowering costs.

## Features

### Lossless Compression

Robust, high performance FASTQ.gz and BAM compression. **Full validation and MD5 matching.**

### Transparent Access

The PetaLink library gives transparent access to the compressed files in their original format. Efficiently tier NGS data for cost.

### Your choice of On-Premise or Cloud Storage

Directly stream data to where it's needed, even from object storage. Pipelines written for local use run in the cloud without modification.

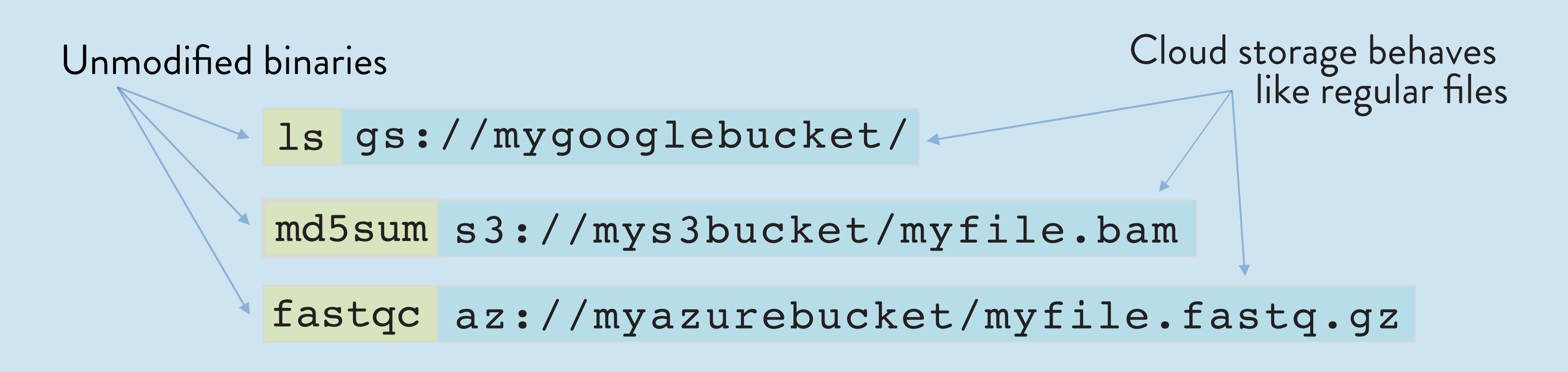
Available for all major Linux distros as .rpm or .deb packages. Customized versions of the Integrative Genomics Viewer (IGV) that support PetaGene compressed files are available for Windows and Mac.

## Improves your analysis speed

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

For more information, please contact us:
info@petagene.com
www.petagene.com

PetaGene
transparent lossless compression


Unmodified binaries

```
ls gs://mygooglebucket/
md5sum s3://mys3bucket/myfile.bam
fastqc az://myazurebucket/myfile.fastq.gz
```

Cloud storage behaves like regular files

## Seamless migration (PetaLink)

PetaLink is a powerful virtual file access system. It enables migration of BAM and FASTQ.gz data to more efficient compression formats. For example, after the PetaSuite binary has been used to (a) losslessly compress a BAM file, (b) validate that all data in the BAM has been preserved, and (c) remove the original BAM file; PetaLink then makes available a high performance virtual BAM file view of the compressed file, with the same filename as the original file, in the same location. **This virtual file can then be used just like the original BAM file by Linux toolchains, pipelines and genome browsers transparently.**

## Fast, efficient compression

Compression of BAM and FASTQ.gz files at 290+ MBytes/sec (4-core i7) uses 3GB of RAM. The species is automatically detected, for simple and optimal compression.

## PetaGene lossless compression ratios

| Source data (human 30x WGS) | Pipeline | PetaGene % savings | CRAM (latest) % savings |
|---|---|---|---|
| FASTQ.gz, HiSeq X | | 67% | Not applicable |
| FASTQ.gz, NovaSeq | | 77% | Not applicable |
| BAM, HiSeq X | BWA-mem only | 55% | 47% |
| BAM, HiSeq X | GATK | **81%** | 33% |
| BAM, NovaSeq | Isaac only | 64% | 57% |
| BAM, NovaSeq | BWA-mem only | 69% | 58% |
| BAM, NovaSeq | GATK | **91%** | 33% |

Note: using PetaGene's optional BayesCal quality score refinement increases the compression ratio by a further 30-70%.

Unlike CRAM, all data is fully preserved, and you do not need to specify a reference for compression or decompression - not even for BAM.

## Cloud integration

**The graphic above shows how simple we make it to access your data whether in the cloud or on local object storage**. Everything can be treated as a regular local file and data is streamed to where you need it *without* downloading it first. On the fly decompression is done at the destination.

We also enable your local pipelines to now run in the cloud without modification.

## Accelerated transfers

Streaming compression to the cloud or local storage enables FASTQ.gz or BAM files to be compressed and transferred in a single operation. On the fly decompression means no waiting to use the data, wherever it is stored. PetaLink can be used to accelerate WAN random access of BAM files such as for interactive Genome Browsers. Smaller files from BayesCal and PetaLink enable faster transfers more generally.

## Our business model – open, no lock in

We make money only if we save our customers money. **We also believe that customers shouldn't be locked in by software,** and for this reason we give a perpetual licence to use the software for free to decompress or access compressed files.

We encourage customers to distribute any PetaGene compressed content. **We freely allow anyone to use PetaLink** to view PetaGene compressed files as BAM virtual files or FASTQ virtual files.

petagene.com