# PetaSuite

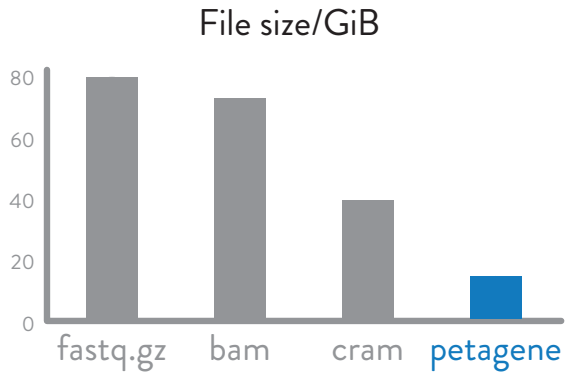
REDUCING THE SIZE AND COST OF NGS DATA
STORAGE AND TRANSFER

## Overview

PetaSuite is a set of complementary software tools that significantly reduce the size and cost of NGS data for storage and transfer.

## Significant cost reductions

Unlike generic storage software, PetaSuite understands the internals of genomics files. For lossless storage, PetaSuite offers cost reductions of up to 4:1 compared to BAM or gzipped FASTQ files.




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

## Transparent usage

PetaSuite lets researchers and clinicians continue using their FASTQ, BAM, and CRAM 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 separating out unimportant NGS components 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.

## Toolkit

PetaSuite consists of several complementary software tools:

### FasterQ

Robust, high performance FASTQ compression for outstanding data reduction and accelerated transfers.

### BayesCal

Revolutionary Bayesian approach to NGS quality score refinement for FASTQ, BAM and CRAM files.

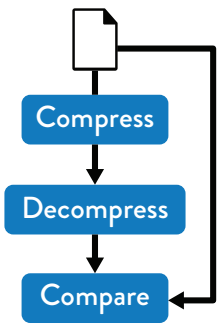
### PetaView

Access CRAM as BAM, FasterQ as FASTQ. Efficiently tier NGS data for cost. Accelerate storage I/O.

Available for all major Linux distros as .rpm or .deb packages.

## Lossless validation

The compression suite has integrated lossless validation to verify that all NGS data is preserved in the compression process.




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

PetaGene
SMALLER, FASTER GENOMICS DATA

"Handling the enormous amount of data we receive from genome sequencing is a huge challenge in our group as we analyse data from more than 10,000 human genomes... PetaGene's solutions allow us to easily store, use and visualise the sequencing data at a fraction of the cost."

**Dr Chris Penkett**

Head of Pipelines for the 10K NIHR Rare Disease Genomes Project
NHS Blood and Transplant & University of Cambridge

## Easy migration

PetaView is a powerful virtual file access system. It enables migration of BAM and FASTQ.gz data to more efficient compression formats. For example, upon importing a BAM file, PetaView can losslessly convert it to CRAM, validate that all data in the BAM has been preserved, and remove the original BAM file. A high performance virtual BAM file view of the CRAM file is then made available in the same directory. This virtual file can then be used just like the original BAM file by Linux toolchains, pipelines and genome browsers transparently.

PetaView also understands the internals of BAM and CRAM files, so that it can split out relatively unimportant genomics data onto cheaper storage media. Virtual views of the full lossless and reduced versions are transparently accessible by the user. This enables fast access to a smaller, improved BayesCal version, and use of the full lossless version when needed.

## Fast, efficient FASTQ compression

FASTQ compression at 140MB/sec (4-core i7), typically smaller than CRAM, uses 4GB of RAM.

## Accelerated transfers

FasterQ streaming compression enables FASTQ.gz files to be compressed, transferred and decompressed in a streaming fashion. PetaView can be used to accelerate WAN random access of BAM files such as for interactive Genome Browsers. Smaller files from BayesCal and PetaView enable faster transfers more generally.

## Bayesian quality score refinement

BayesCal uses a Bayesian approach to calculate a more complete posterior estimation of sequencer error. Genotyping accuracy is preserved or even typically improved.

Improved compression by 2-3x is a side effect. For example, on Illumina Hi-Seq-X 30x WGS human sample NA12878, the original gzipped fastq files are 73.7GiB in size, whereas combining FasterQ and BayesCal this is reduced to 13.7GiB (5.3x smaller).

## Main modes of storage

**1) Untiered Lossless:**
2-3x compression over BAM and FASTQ.gz

**2) Untiered BayesCal:**
Improves genotyping, 5-6x compression over BAM and FASTQ.gz

**3) Tiered lossless:**
Smaller BayesCal version on fast tier, differences stored on slow tier. Transparent access to full lossless and BayesCal versions. Up to 4x overall storage cost reduction.

## Our philosophy

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 make all decompression and accessibility updates available perpetually.

We encourage customers to distribute any PetaGene compressed content. We freely allow anyone to use PetaView to view compressed CRAM files as BAM virtual files, or FasterQ files as FASTQ virtual files.