PetaSuite

REDUCING THE SIZE AND COST OF NGS DATA STORAGE AND TRANSFER

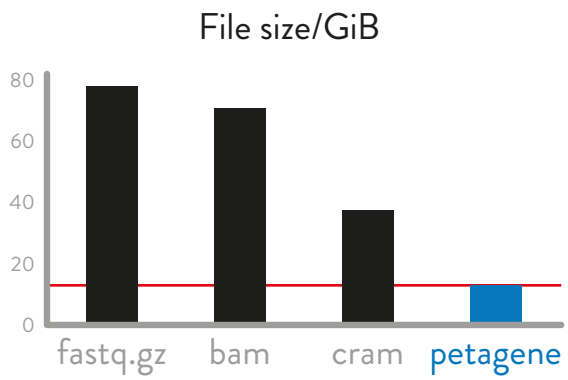
BEST OF SHOW Bio-IT World WINNER

Overview

PetaSuite is a set of scalable 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 6:1 compared to BAM or gzipped FASTQ files. **This is a 96% reduction compared to raw FASTQ files.**



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

Transparent usage

PetaSuite lets researchers 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 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:

Lossless Compression

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

BayesCal (optional)

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

PetaView

Transparently access compressed files in their original format. Efficiently tier NGS data for cost.

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

Improves your 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* your analysis, tools and pipelines, especially in HPC environments.

P.T.O.

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)

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 compress it, 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 compressed 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 FASTQ.gz files, so that it can split out relatively unimportant genomics data onto cheaper storage media. Virtual views of the full and reduced versions are transparently accessible by the user.

Fast, efficient compression

Compression of BAM and FASTQ.gz files at 140+ MBytes/sec (4-core i7) uses 4GB of RAM. **Unlike CRAM, all data is fully preserved, and you do not need to specify a reference for compression or decompression - *not even for BAM.*** The species is automatically detected, for simple and optimal compression.

Accelerated transfers

Streaming compression enables FASTQ.gz or BAM 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 and typically improved.

Improved compression is a side effect, yielding sizes 4-8x smaller than the original. In the NA12878 example shown in the front page plot, the original gzipped FASTQ files are 73.7GiB in size, whereas by combining BayesCal and our Lossless Compression this is reduced to 13.7GiB (5.3x smaller).

Main modes of storage

1) Untiered Lossless:
2-3x compression over BAM.
2-5x compression over FASTQ.gz

2) Untiered BayesCal+Lossless:
Improves genotyping, 4-8x compression over BAM and FASTQ.gz

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

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 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 PetaGene compressed files as BAM virtual files or FASTQ virtual files.