PetaSuite

REDUCING THE SIZE AND COST OF NGS DATA STORAGE AND TRANSFER

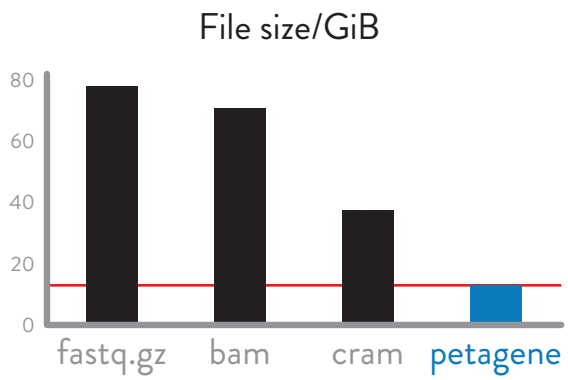
BEST OF SHOW Bio IT World CONFERENCE & EXPO '16 WINNER

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 **10:1 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.**

PetaLink

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

BayesCal (optional)

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

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

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

"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

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 losslessly compress a BAM file, validate that all data in the BAM has been preserved, and remove the original BAM file, PetaLink makes available a high performance virtual BAM file view of the compressed file, with the same filename of the original file, 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.**

Fast, efficient compression

Compression of BAM and FASTQ.gz files at 290+ MBytes/sec (4-core i7) uses 3GB 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. 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.

Bayesian quality score refinement

BayesCal uses a Bayesian approach to calculate a more complete posterior estimation of sequencer error. Genotyping accuracy is preserved across the ROC curve, typically with a signficiant net increase.

Improved compression is a side effect, increasing compression ratios by a further 30-70% compared with straight lossless compression.

PetaGene lossless compression ratios

Source data (human 30x WGS)	Pipeline	PetaGene ratio	PetaGene %savings	CRAM (latest) ratio
FASTQ.gz, HiSeq X		3.0	67%	Not applicable
FASTQ.gz, NovaSeq		4.3	77%	Not applicable
BAM, HiSeq X	BWA-mem only	2.2	55%	1.9
BAM, HiSeq X	GATK	**5.2**	**81%**	1.5
BAM, NovaSeq	Isaac only	2.8	64%	2.3
BAM, NovaSeq	BWA-mem only	3.2	69%	2.4
BAM, NovaSeq	GATK	**10.9**	**91%**	1.5

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

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 PetaLink** to view PetaGene compressed files as BAM virtual files or FASTQ virtual files.