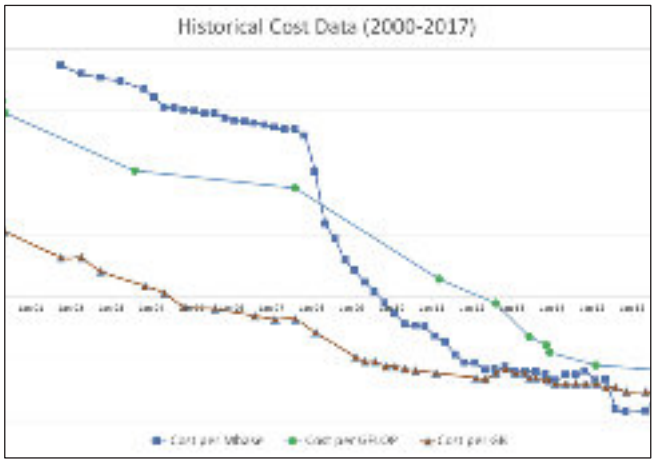
PetaGene
transparent lossless compression

Stop wasting money on genomic data storage.

Use those funds to advance your research.

Storage of genomic data can easily get out of control. First, the primary data files (FASTQ and BAM) are large and considered precious since they require a significant investment in sequencing (FASTQ) and compute (BAM) to generate. Second, there are regulatory and scientific reasons to keep the data safeguarded for long periods of time. And finally, the science is changing rapidly and keeping these data at the ready is important to reanalyzing them or aggregating them with other samples into larger cohorts.

Even for modest projects or departments, you can soon have a petabyte sized storage bill. And, don't expect that bill to go down anytime soon. AWS has not lowered its prices for storage in several years. In fact, the cost of data storage is starting to trend upwards.



Some costs are falling -- storage isn't one of them.

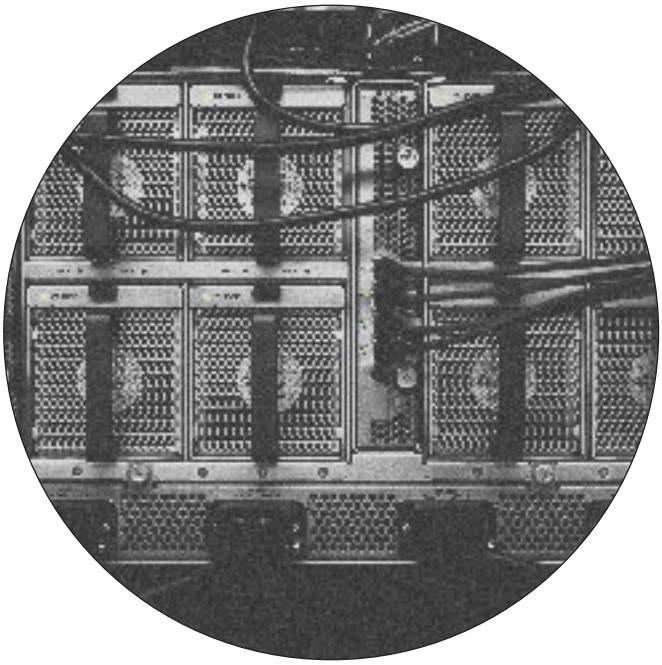
In the two decades since the sequencing of the first human genome, the cost to sequence a megabase has dropped over a million-fold. In the same time, the cost of a gigaFLOP (a billion floating point operations) of compute has dropped 100,000-fold. In contrast, the cost of a GB of disc storage has dropped only 500-fold; and only 5-fold in the last decade[1]. Given the demand for faster, solid state data storage, the unit price of data storage is actually starting to rise. Based on these trends, the cost of data storage will soon become the largest budget line in any genomics project.

[1] https://www.kapwing.com/blog/cloud-costs-arent-actually-dropping-dramatically/

Currently, genomics data is stored as FASTQ.gz or BAM files on a local storage array (on-premise) or in a cloud storage service such as AWS S3. There are different cost structures for these storage options: on-premise storage requires up-front capital and amortized costs over the life of the equipment. With the hidden costs of support and IT personnel, the cost per TB is about $300-500 per year. Cloud services such as S3 are a purely operational cost (no capital) at ~$300/TB/yr.

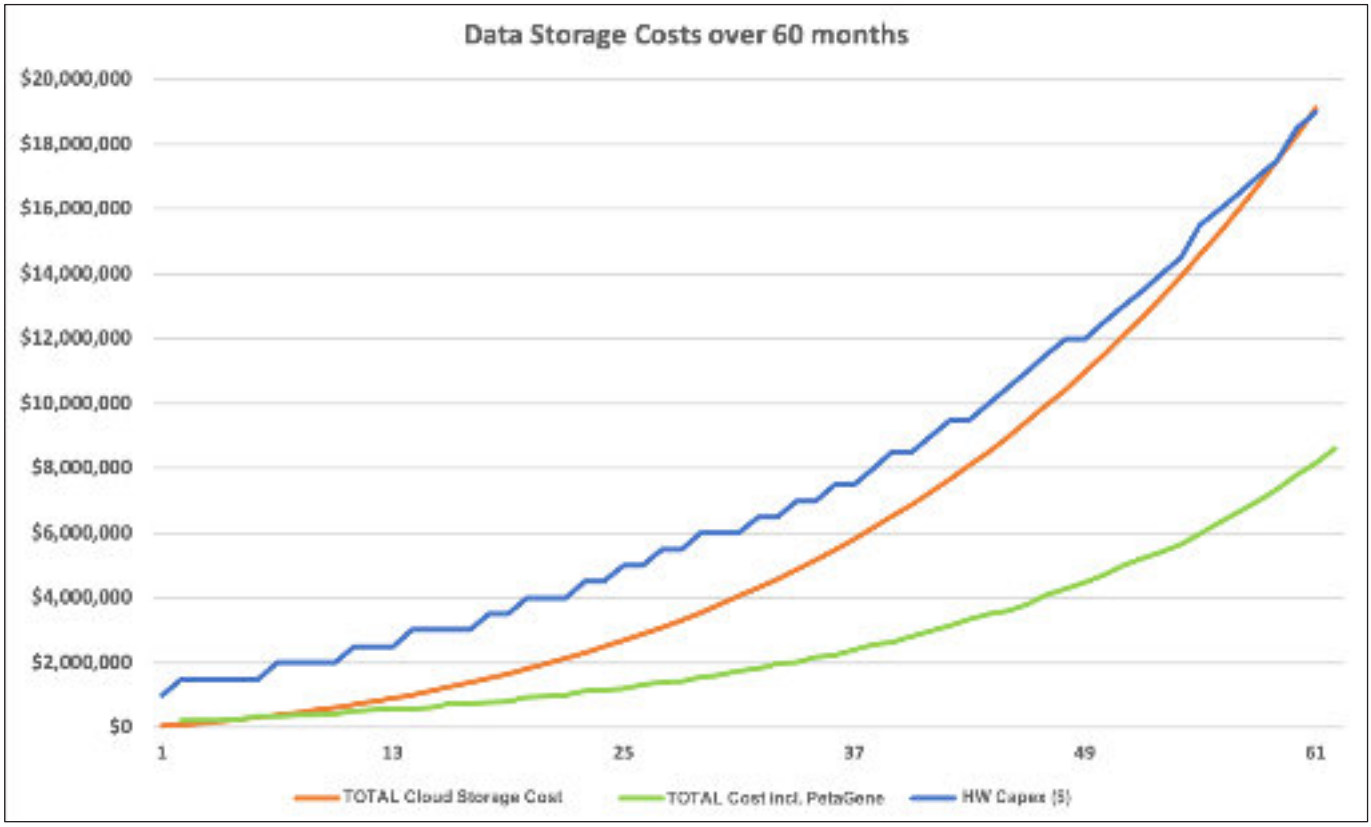
Control your genomic data storage costs

PetaGene genomic data compression can reduce the size of BAM and FASTQ.gz files by 60-90% without loss of any information. Our 100% lossless compression technology is also 100% transparent such that our files do not need to be decompressed to be used by applications or bioinformatics pipelines. In fact, these files look and act just like the original data files. For the popular GATK-Best Practices pipeline, our technology reduces the total data footprint by 75%. These are significant savings, freeing cash to invest in your organization's core objectives.



A real life genomic data storage cost example

Let's take a look at a real use case: a genetic testing company that provides cancer exome panels. This organization does 44,000 tests per year and is growing at 40% year over year. Each test generates 26 GB of data and they need to retain these data for CLIA compliance (2 yrs) and for future product development (>5yrs).



In order to maintain sufficient storage capacity, this company would need to continuously procure, install, and operate over 40 PB of storage hardware (blue line). Given the time required to install, configure, and test this equipment, they need to procure at least 1 quarter ahead of the actual demand, just to be safe. Alternatively, they could use a cloud store in a more elastic mode and bring storage resources on-line just-in-time of demand (orange line). The gap between the blue and orange curves is the economic advantage of using the cloud for data storage in a growing business.

When they use PetaGene compression with a cloud storage strategy, **they save a net 55%** in total storage costs (green line). This is a total cost, including the cloud storage fees and our license fees. The total savings in this example is $10.8M.

If you would like to control your data storage costs and recover funds for business or research goals instead of just handing cash to your chosen cloud storage provider, contact us at sales@petagene.com to arrange a software evaluation.

Ideaspace · Hauser Forum 3 Charles Babbage Rd · Cambridge CB3 0GT · U.K.

www.petagene.com

email: info@petagene.com

PetaGene
transparent lossless compression