PetaGene

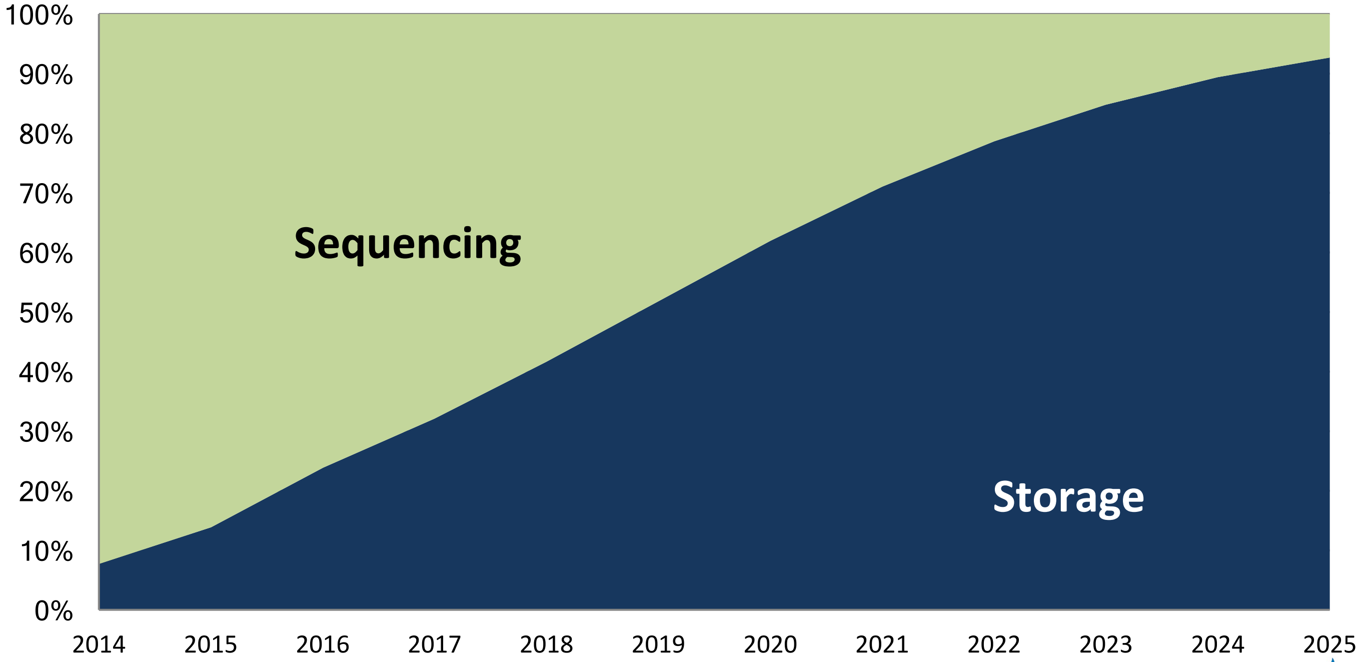
PetaSuite

REDUCING THE SIZE AND COST OF NGS DATA STORAGE AND TRANSFER

Dr Vaughan Wittorff
Co-Founder & Business Development Manager

Motivation

Storage vs Sequencing cost



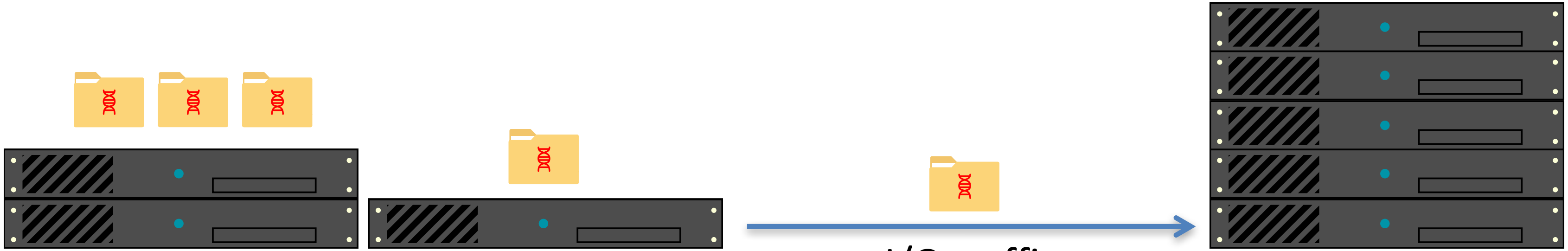
Project PetaGene

- Team of researchers:
  - Dan Greenfield
  - Alban Rrustemi
  - Oliver Stegle (head of Stegle Lab)

- Private and governmental grant funding
- Collaboration with Stegle Group at EMBL-EBI

PetaGene

Smaller, faster, better



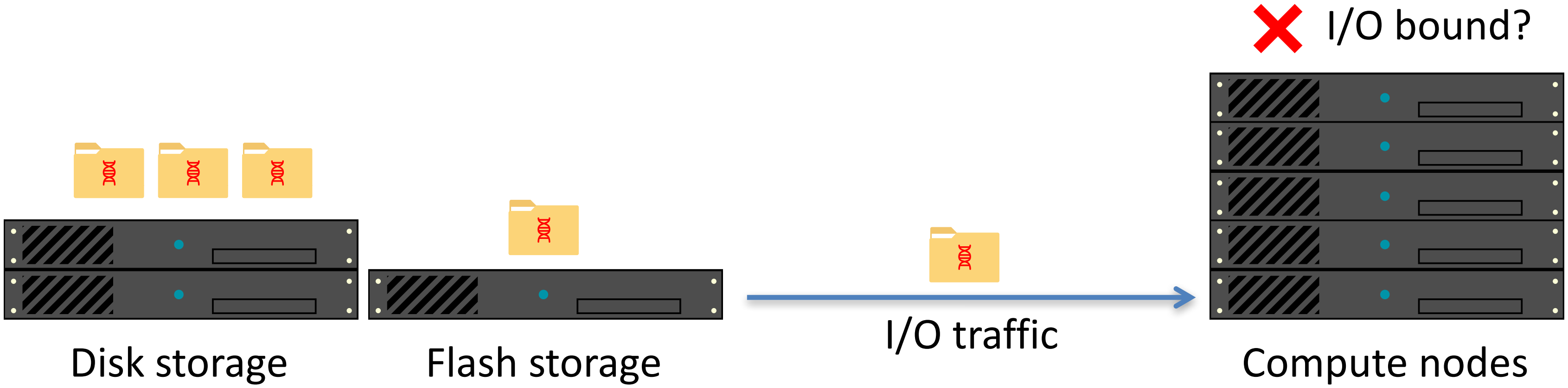
Disk storage

Flash storage

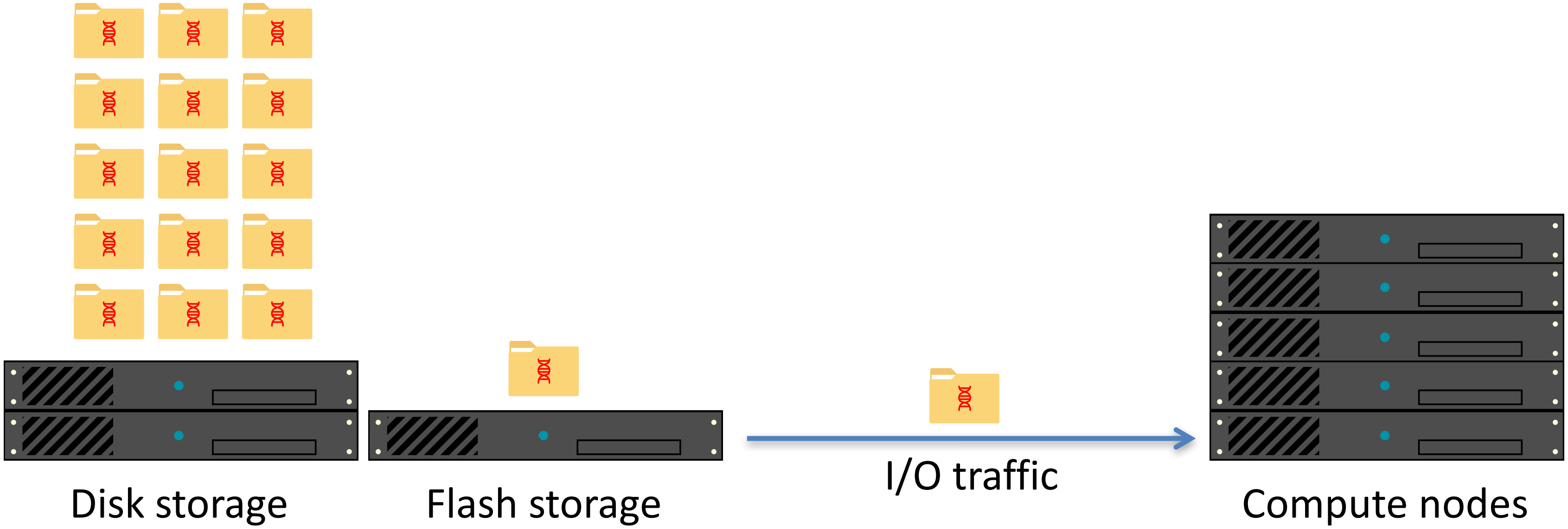
I/O traffic

Compute nodes

Smaller, faster, better

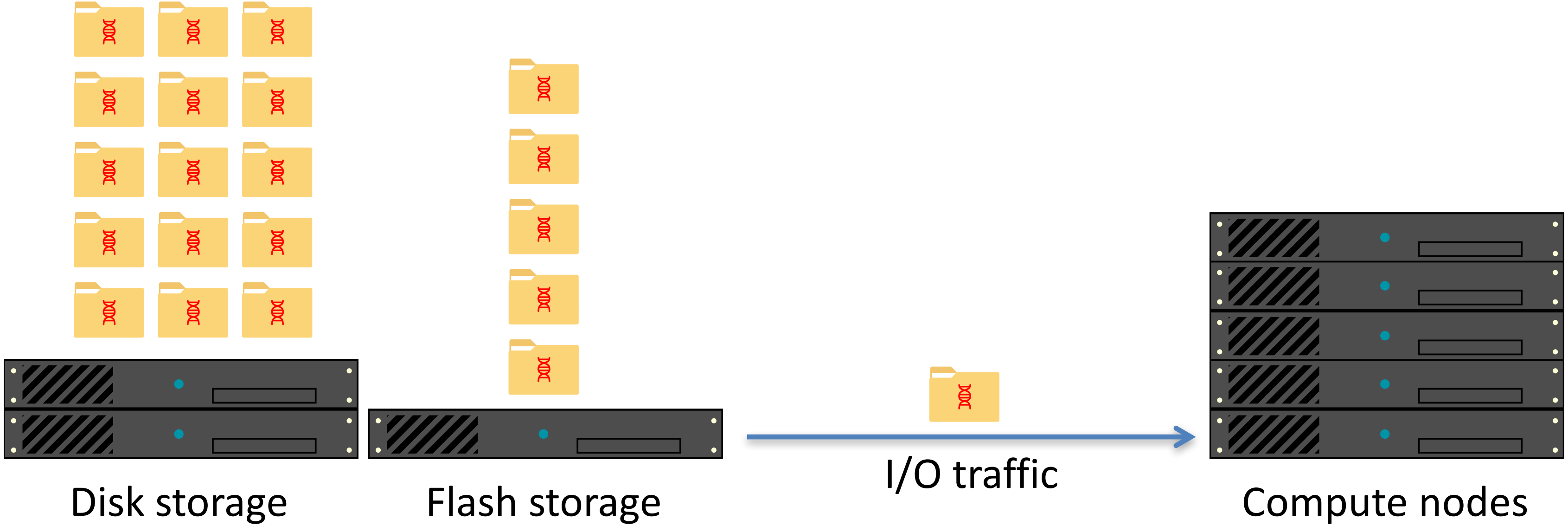


Smaller, faster, better

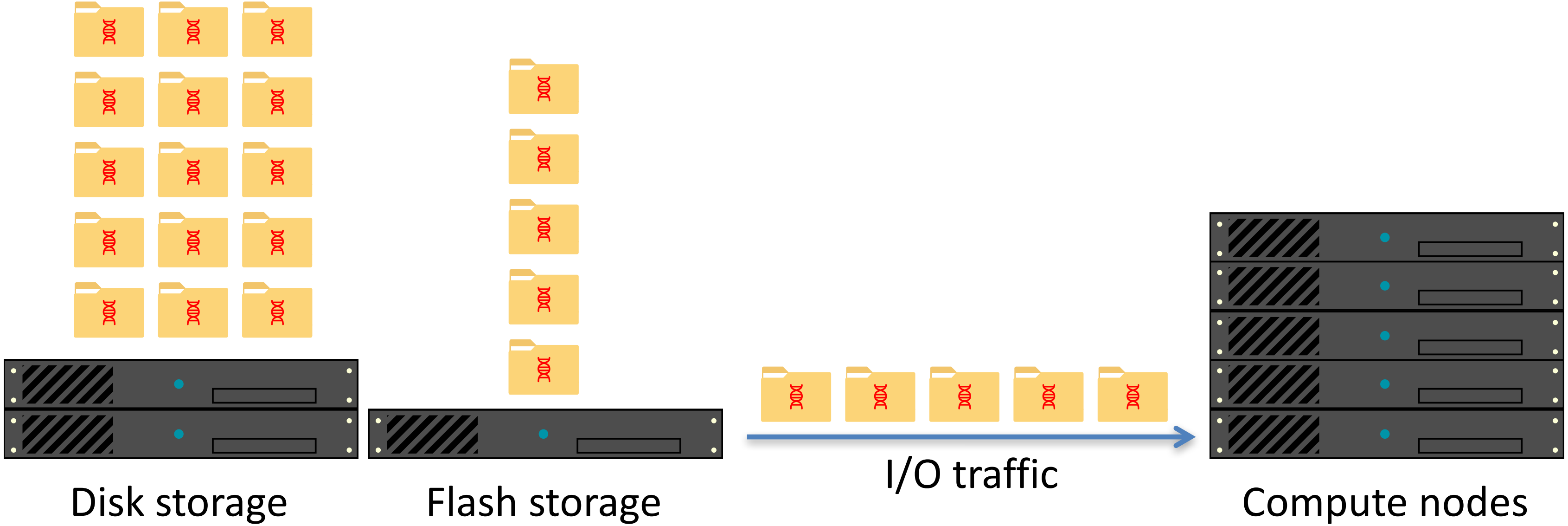


PetaGene

Smaller, faster, better

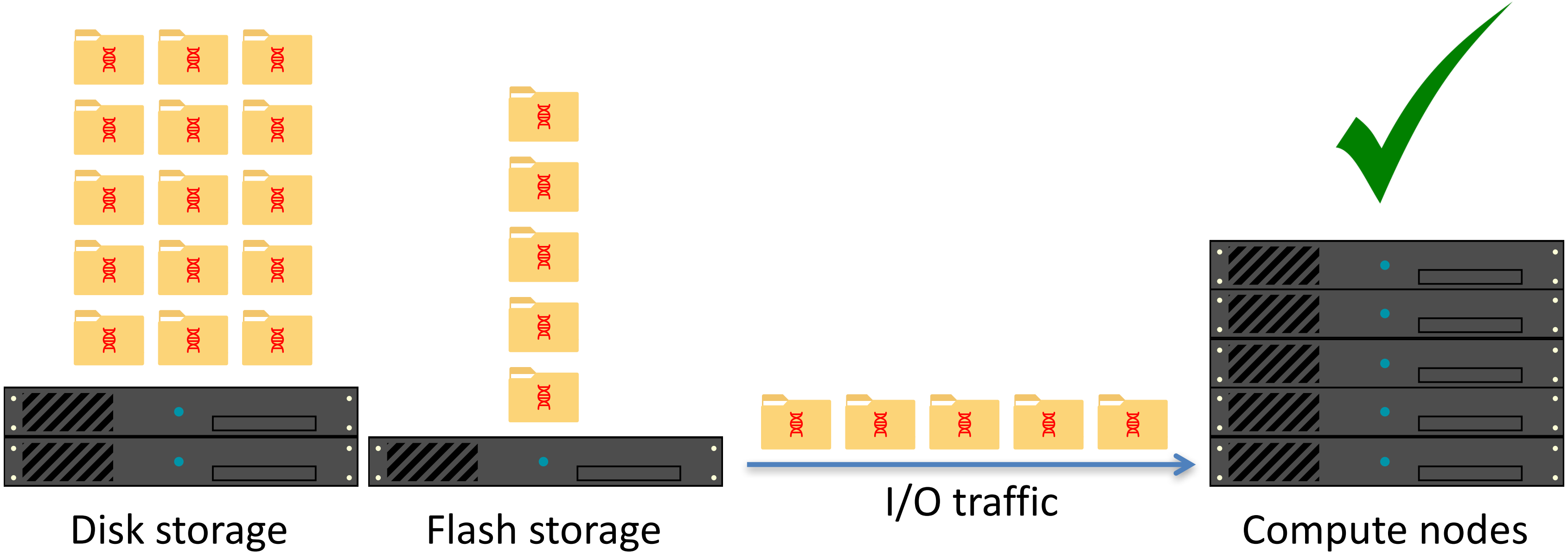


Smaller, faster, better

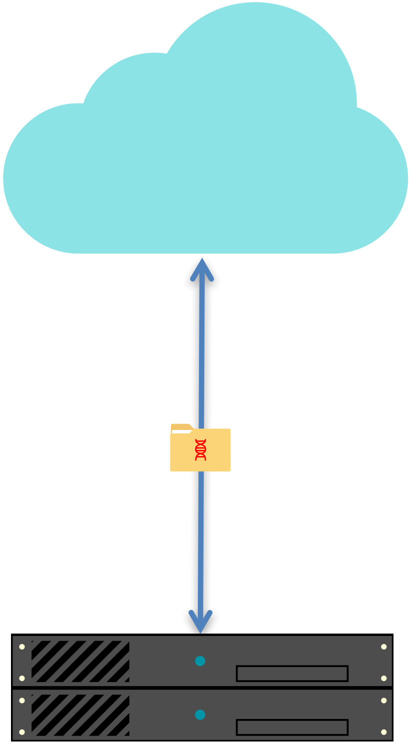


PetaGene

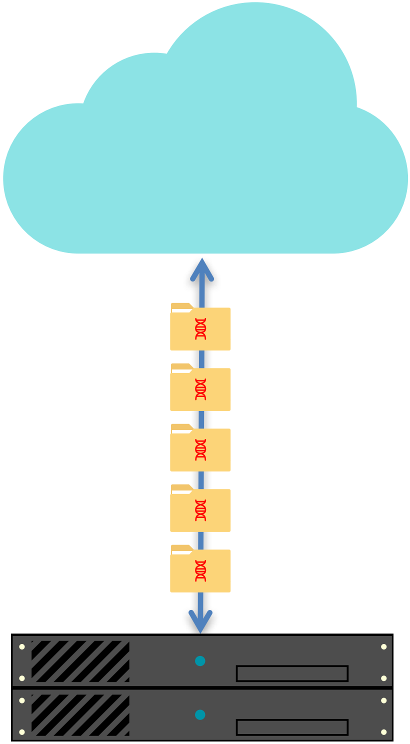
Smaller, faster, better



Smaller, faster, better



Smaller, faster, better



What we do

- Lossless Compression
  - Robust, high performance FASTQ.GZ and BAM Compression
  - Full validation and MD5 matching of FASTQ, FASTQ.GZ and BAM
- Transparent
  - Access compressed files in their native format
  - Access as BAM or FASTQ.GZ at exact same path as before on existing storage
- Accelerated transfers
  - Including streaming compression to/from S3
- BayesCal (optional)
  - Revolutionary Bayesian approach to NGS quality score refinement for FASTQ and BAM files.

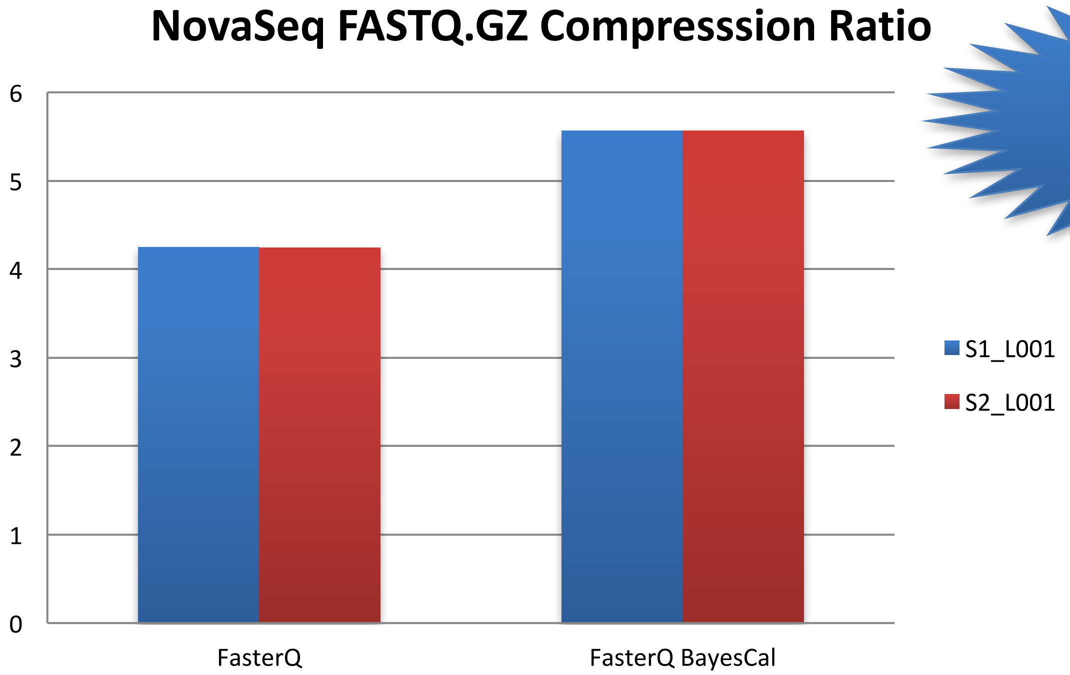
PetaSuite: FasterQ

- Robust 100% lossless compression
- Significantly better compression than CRAM
- Transparent access as FASTQ / FASTQ.gz
- High speed streaming FASTQ compression
- Streaming mode can be used to accelerate FASTQ file transfers

PetaSuite: FasterQ

- 3GB compression memory footprint
- 1GB decompression memory footprint
- 140MBytes/sec compression on a 4-core i7
  - Compared to 17MBytes/sec with GZIP

Outstanding lossless compression



Even better numbers coming out soon

PetaGene CRAM

	Preserves all data fields	Revert to orig. BAM (same MD5)	Storage HW TCO reduction	Direct access as BAM (for tools)	No need to specify a reference
BAM	✓	✓	none	✓	✓
CRAM	X e.g. corner cases MAPQ, CIGAR, MD:Z, and more	X CRAM2BAM has different MD5 in general	2:1	X	X
PetaGene CRAM	✓	✓ preserves every bit of the original BAM	2:1 to 4:1 higher reduction with BayesCal and tiering	✓ via free PetaView library	✓ works even with de-novo aligned BAM

PetaGene

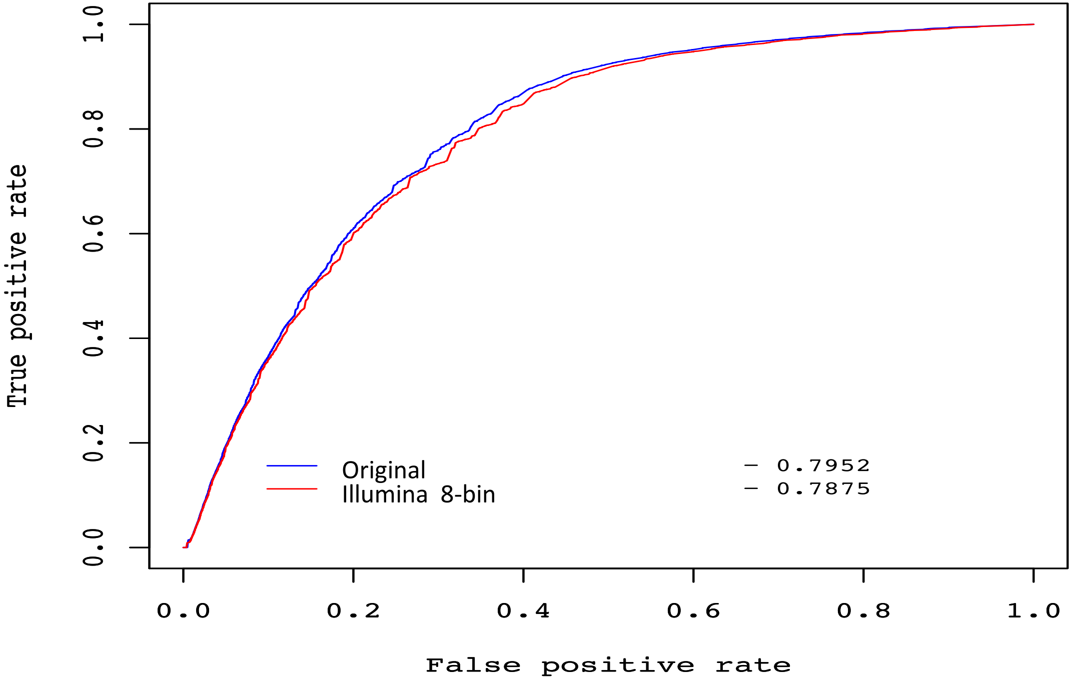
Quality scores

- Example Read:

Sequence bases:	GCAGTATGCCTGGTGTATTTCAGAAACAACCA
Quality scores (QS):	@CCDFDEDFIHHDGGI@GI@FGH?<@A<I?>@

- QS is the estimated probability of an incorrectly sequenced base
- For Illumina reads, QS takes 60-80% of CRAM file
- Generic compression is reaching its limits and these limits are not good enough!

SRR622461 (NA12878) Illumina 8-bin



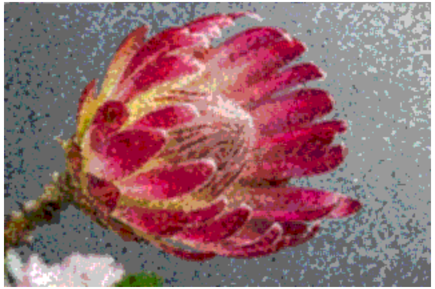
PetaGene

Lossy vs Refined

Original

Lossy →

Quantized

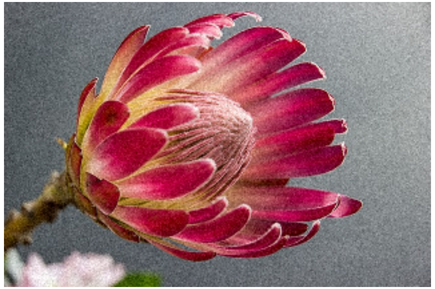


Traditional approach
(Not what we do)

Refined ↔

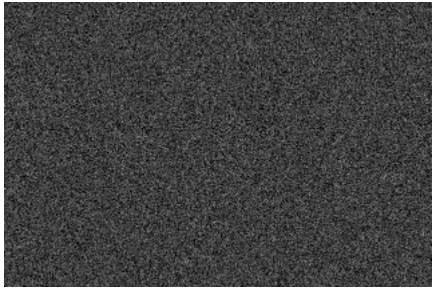
Bayesian approach
with corpus

Refined



+

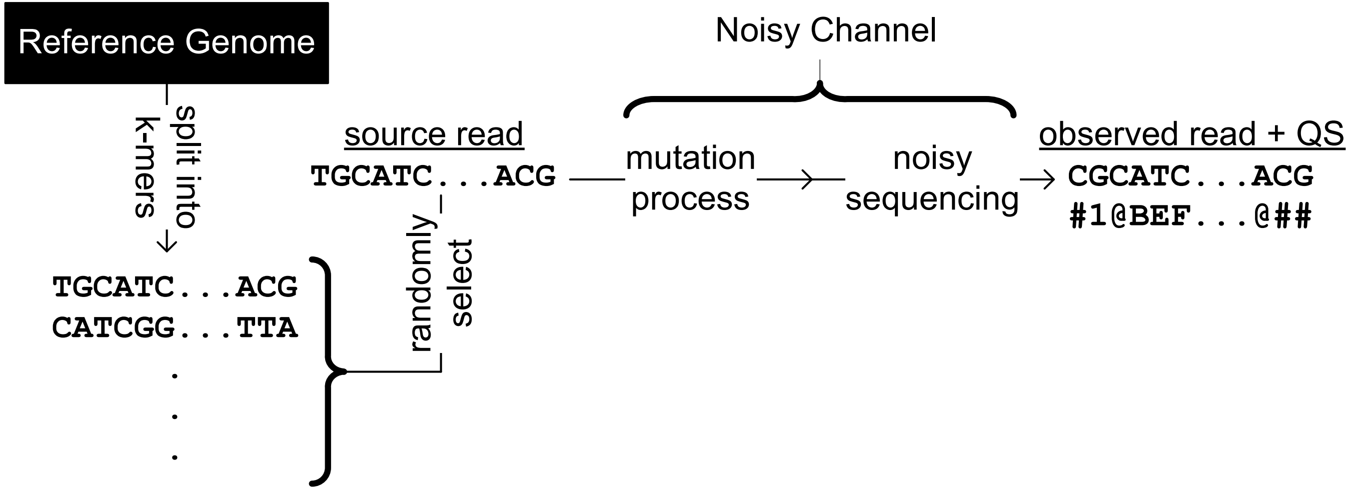
Deltas



PetaSuite: BayesCal

- Bayesian approach to yield a better estimate of sequencing error (i.e. quality score) for each base in a read
- Calculate posterior probability of error given model and prior knowledge (e.g. from a reference genome)
- Leads to better genotyping accuracy

BayesCal = refinement of quality scores

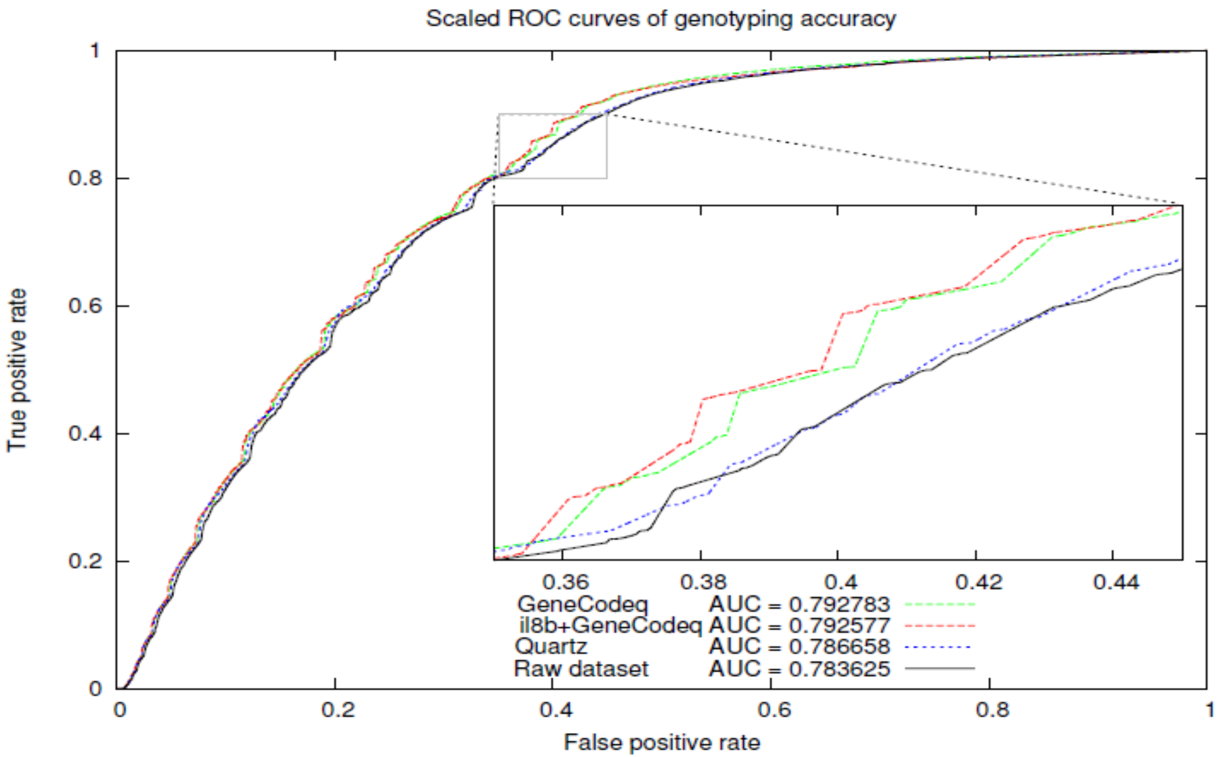


- Sequencing as a Bayesian problem of noisy codeword transmission
- Example source: *k*-mers from a reference genome

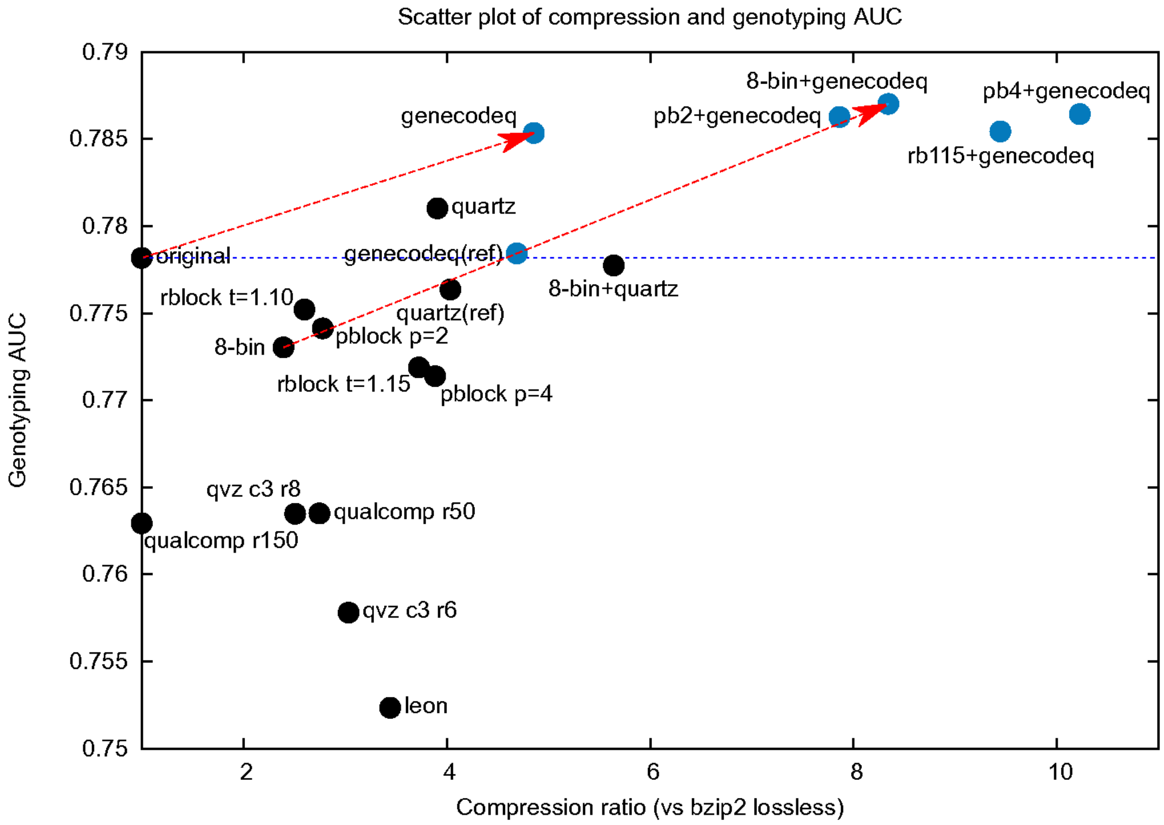
BayesCal with lossless compression

- Works on FASTQ and BAM
- Each read processed completely independently
- Leverages all the quality score information in the read
- Uses a corpus (derived from ref genome) and variants (optional)
- Alignment is not inferred or calculated
- Posterior probability calculated across distribution of all possible source *k*-mers in corpus
- Needs 24GB memory, low memory version forthcoming
- Runs in fraction of time of BQSR or most pipeline stages: (20-40MB/sec on 4-core i7)

Improved quality due to BayesCal (NA12878)

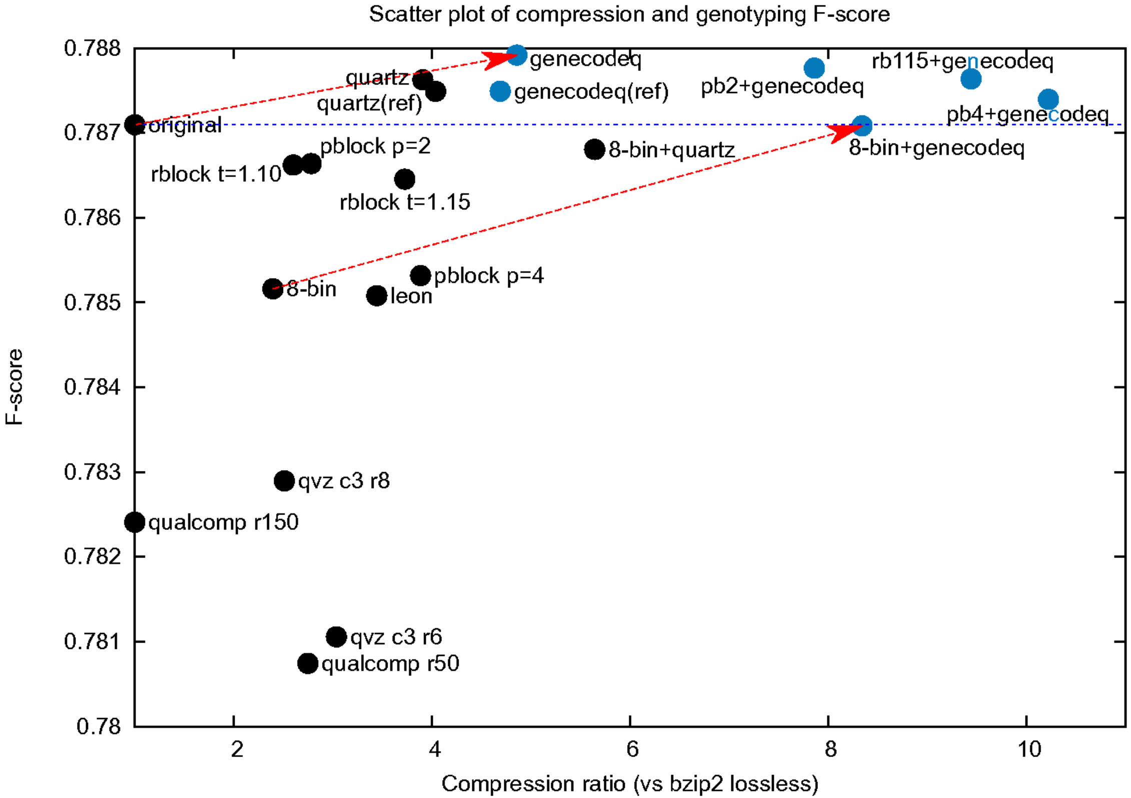


AUC vs Compression Ratio



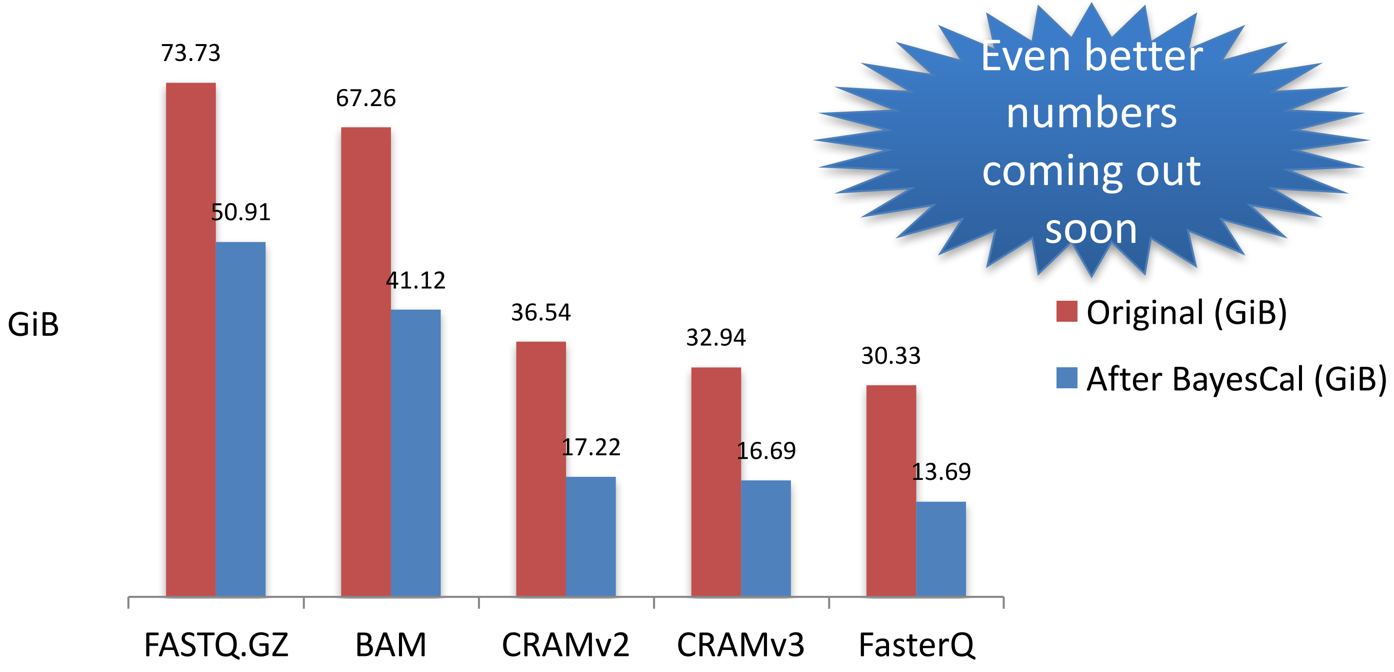
PetaGene

F-Score vs Compression Ratio



PetaGene

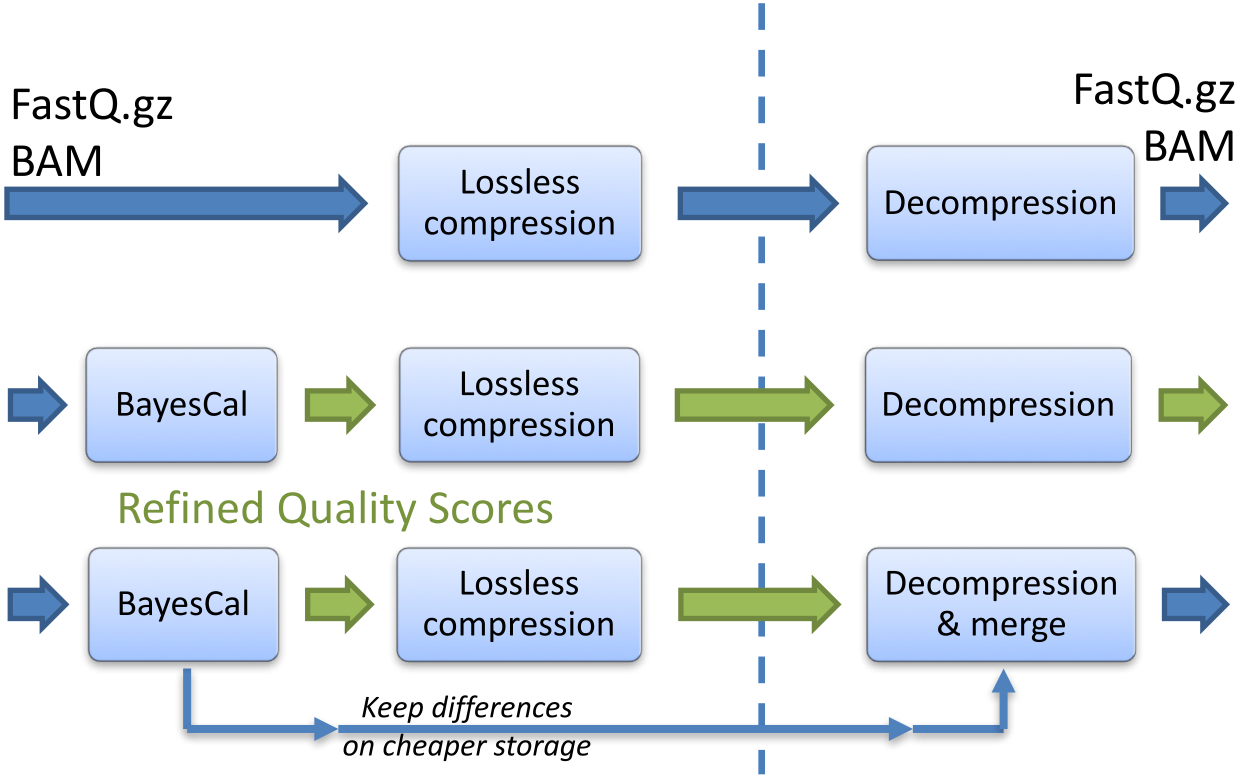
Lossless compression ratios for BayesCal-processed files



Even better numbers coming out soon

April 2017

Comparison of PetaSuite usage modes



FastQ.gz
BAM

Lossless compression

Decompression

FastQ.gz
BAM

BayesCal

Lossless compression

Decompression

Refined Quality Scores

BayesCal

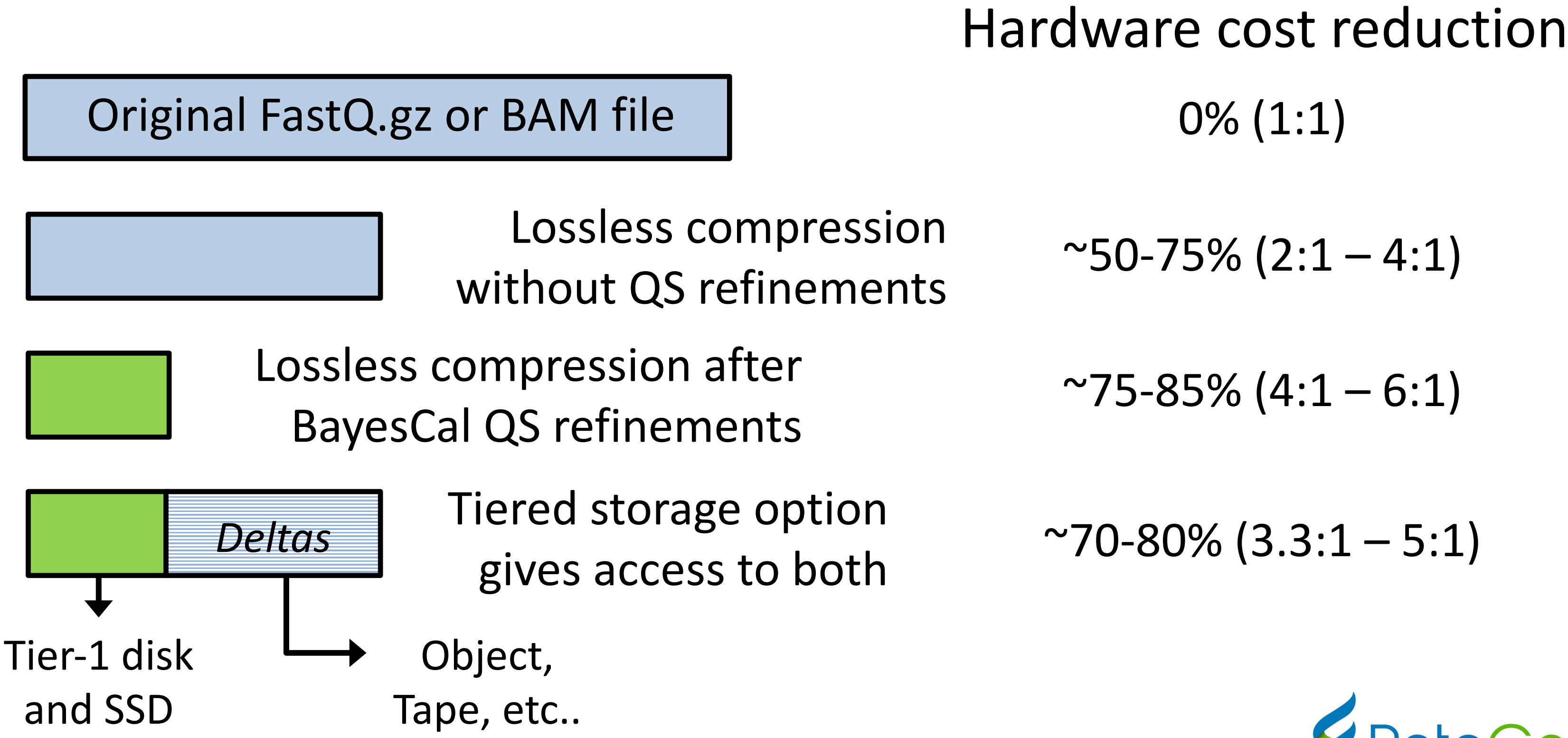
Lossless compression

Decompression & merge

Keep differences on cheaper storage

Refined QS?	Original MD5?	HW TCO saving
X	✓	2:1 - 4:1
✓	X	4:1 – 6:1
✓	✓	3.3:1 – 5:1

Storage savings from PetaSuite



	Hardware cost reduction
Original FastQ.gz or BAM file	0% (1:1)
Lossless compression without QS refinements	~50-75% (2:1 – 4:1)
Lossless compression after BayesCal QS refinements	~75-85% (4:1 – 6:1)
Tiered storage option gives access to both	~70-80% (3.3:1 – 5:1)

PetaView demo

> "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

AstraZeneca

NHS

DNAnexus®

DELL EMC

intel®

Genestack

UNIVERSITY OF CAMBRIDGE

EMBL-EBI

Award winning innovation

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

"The judges chose a new product that could give you millions of dollars worth of storage savings right now, a product that several of our judges wanted to go buy immediately after lunch."

Allison Proffitt, Editorial Director of Bio-IT World

Since last year

- Even faster and better compression
  - Improvements to compression algorithm
  - Multithreaded compression, decompression, validation, and random access
- Exact MD5 preservation
  - Naïve approaches are simple but getting it done right is hard!
- Incredible scalability
  - On top of existing decompression scaling, added support for distributed compression jobs
- Cloud integration
  - Stream compress S3->S3, local->S3, S3->local
  - AWS now, Azure and Google cloud coming
- Autodetect species
  - Instantly autodetects species for optimal compression
  - Support de-novo aligned genomes (e.g. plants)

Summary

- PetaSuite offers powerful tools for:
  - Increasing effective storage capacity
  - Accelerating genomics transfers / WAN acceleration
  - I/O acceleration
  - Improving genotyping accuracy
  - Better utilising tiered storage
- Operates transparently with existing pipelines and storage infrastructure
- We make money by saving our customers money
- No lock-in: all tools for accessing & decompressing data have perpetual free updates for customers

Bias to particular reference?

- Negligible effect of hs37d5 vs much older h16 as source corpus (hs37d5 used for alignment in both)

Reference corpus	**ROC AUC**	**Precision**	**Recall**	**F-SCORE**
Original (no QS refinement)	0.758493	0.873966	0.930355	0.901279
hs37d5 (ref only, no variants)	0.758961	0.874871	0.930250	0.901711
hg16 (ref only, no variants)	0.758570	0.874783	0.930262	0.901670

(Broad Institute recommended GATK pipeline, no-BQSR, NA12878J dataset at 30x, Illumina Platinum set (chr1))

Effect on rare variants?

- Define variants not in *dbSNP* as ‘rare variants’
- Negligible effect on finding true rare variants

Approach	True ‘rare’ SNP variants found (of 46920)	Δ true ‘rare’ SNP variants found	new ‘rare’ SNP variants found
Original	8636	–	–
BayesCal (hs37d5 reference, no variants)	8648	12 (0.13% more)	12
BayesCal (1k Genome h=16 variants)	8638	2 (0.02% more)	23
Illumina 8-bin	8564	–72 (0.83% less)	10

PetaGene

Bias to variants in corpus?

- 14.9 million variants in h=16 corpus
- Negligible effect on false positives

Approach	False positives (of 14.9 million in corpus)	Δ false positives	New false positives
Original	23136	–	–
BayesCal (hs37d5 reference, no variants in corpus)	23171	35 (0.15% more)	45
BayesCal (1000 Genome h=16 variants in corpus)	23240	104 (0.45% more)	138
Illumina 8-bin	22980	-156 (0.67% less)	33
QVZ (3 clusters 0.6 bits/QS)	23272	136 (0.59% more)	666

(Broad Institute recommended GATK pipeline, no-BQSR, SRR622461 dataset at 5x, Illumina Platinum set)

What about sample contamination?

- Highly unlikely to be modified by *PetaGene BayesCal*
- Process *E. coli* dataset with *human* corpus
- Only 0.0045% of reads modified
- Of these reads, e.g. NCBI BLAST expectation value for best *E. coli* match is 7e-7 vs 4e-73 for best human match, indicating this is likely due to contamination of *E. coli* sample with human DNA.
- High specificity suggests that samples are very unlikely to be modified unless related to corpus