Genomic Data Compression and Processing for Large and Growing Databases

Science of Information

EE 25N

Tsachy Weissman

thanks

Idoia
Ochoa

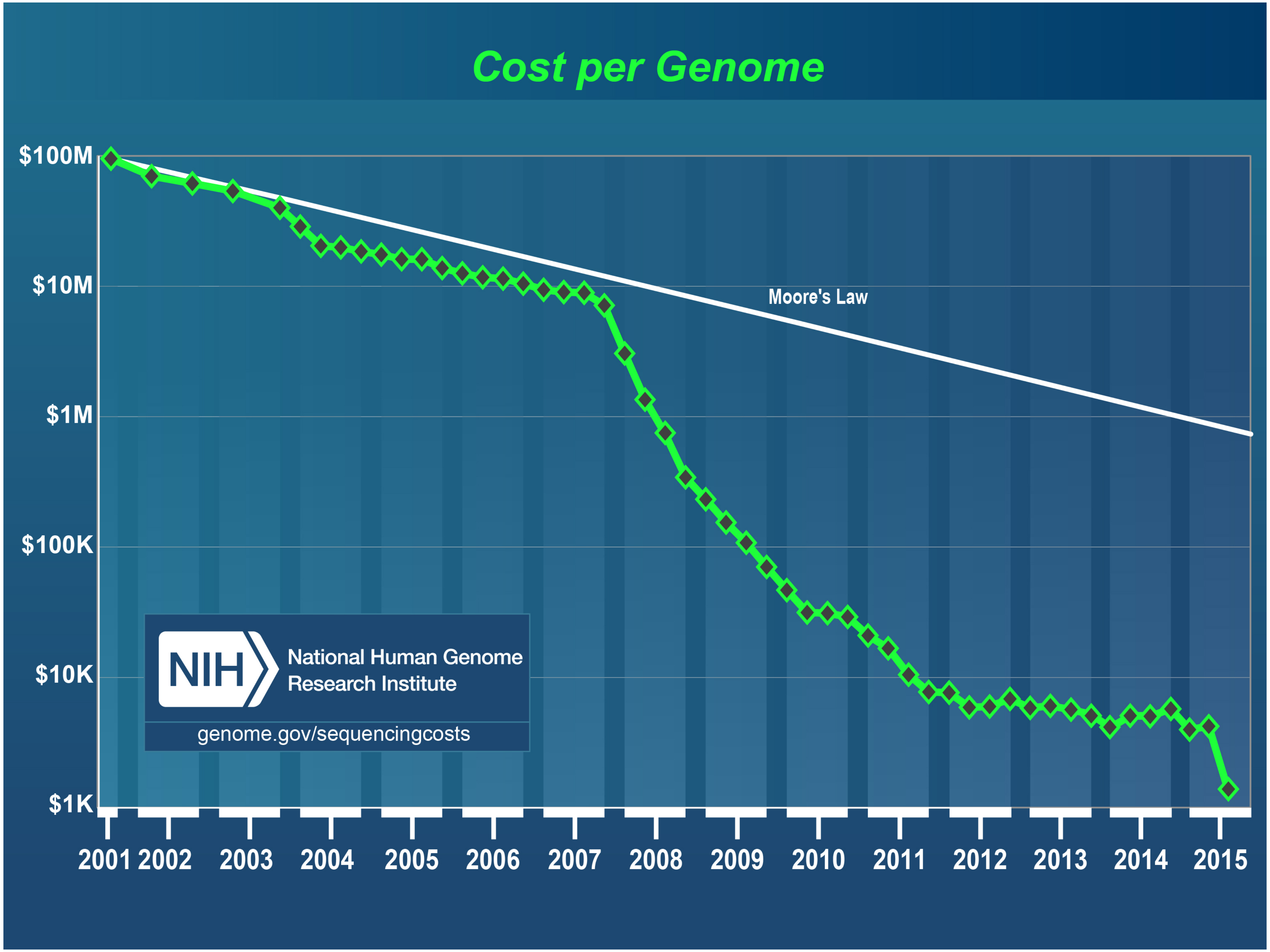
Mikel
Hernaez

Shubham
Chandak

Kedar
Tatwawadi

Dmitri
Pavlichin

a slide we've seen (multiple times..)



why compression of genomic data?

- proxy for reduced cost of storage, communication, computation, processing, etc.
- compression as modelling

2 modes

- lossless
- lossy

(true) story of human genome compression

Single Genome Compression

- **FASTA Compression**: Compression of a single genome
- Human genome can be represented using 2 bits/bp, compressed size $\approx 1GB$
- Specialized compressors: MFCompress[1] : $\approx$ 1.6 bits/bp
- H(human genome): "Entropy" of the human genome

$$H(\text{human genome}) \lesssim 2 \text{ bits/bp} \sim 1 \text{ GB}$$

.

[1] A. J. Pinho and D. Pratas, "MFCompress: a compression tool for FASTA and multi-FASTA data", *Bioinformatics*, vol. 30, no. 1, pp. 117–118, 2014.

Using a reference

- H(human genome) $\lesssim$ 1 GB
- We can do better if we know another genome (reference)
- **Using a reference:** GenomeZip[2] compresses James Watson's genome using: $\approx 2.5MB$

$$H\left(\begin{matrix}\mathit{human}\\\mathit{genome}\end{matrix}\middle|\begin{matrix}\mathit{another}\\\mathit{human}\\\mathit{genome}\end{matrix}\right)\lesssim 2.5MB$$

[2] D. S. Pavlichin, T. Weissman, and G. Yona, "The human genome contracts again", *Bioinformatics*, vol. 29, no. 17, pp. 2199–2202, 2013.

Using a collection

- $H(\text{human genome}) \lesssim 1 \text{ GB}$

- $H\left(\begin{matrix} human \\ genome \end{matrix} \middle| \begin{matrix} another \\ human \\ genome \end{matrix}\right) \lesssim 2.5MB$

- **GTRAC**[3] compressor:

$$H\left(\begin{matrix} human \\ genome \end{matrix} \middle| \begin{matrix} 1K \\ other \\ genomes \end{matrix}\right) \lesssim 1MB$$

- **GTC**[4] compressor:

$$H\left(\begin{matrix} human \\ genome \end{matrix} \middle| \begin{matrix} 27K \\ other \\ genomes \end{matrix}\right) \lesssim 200KB$$

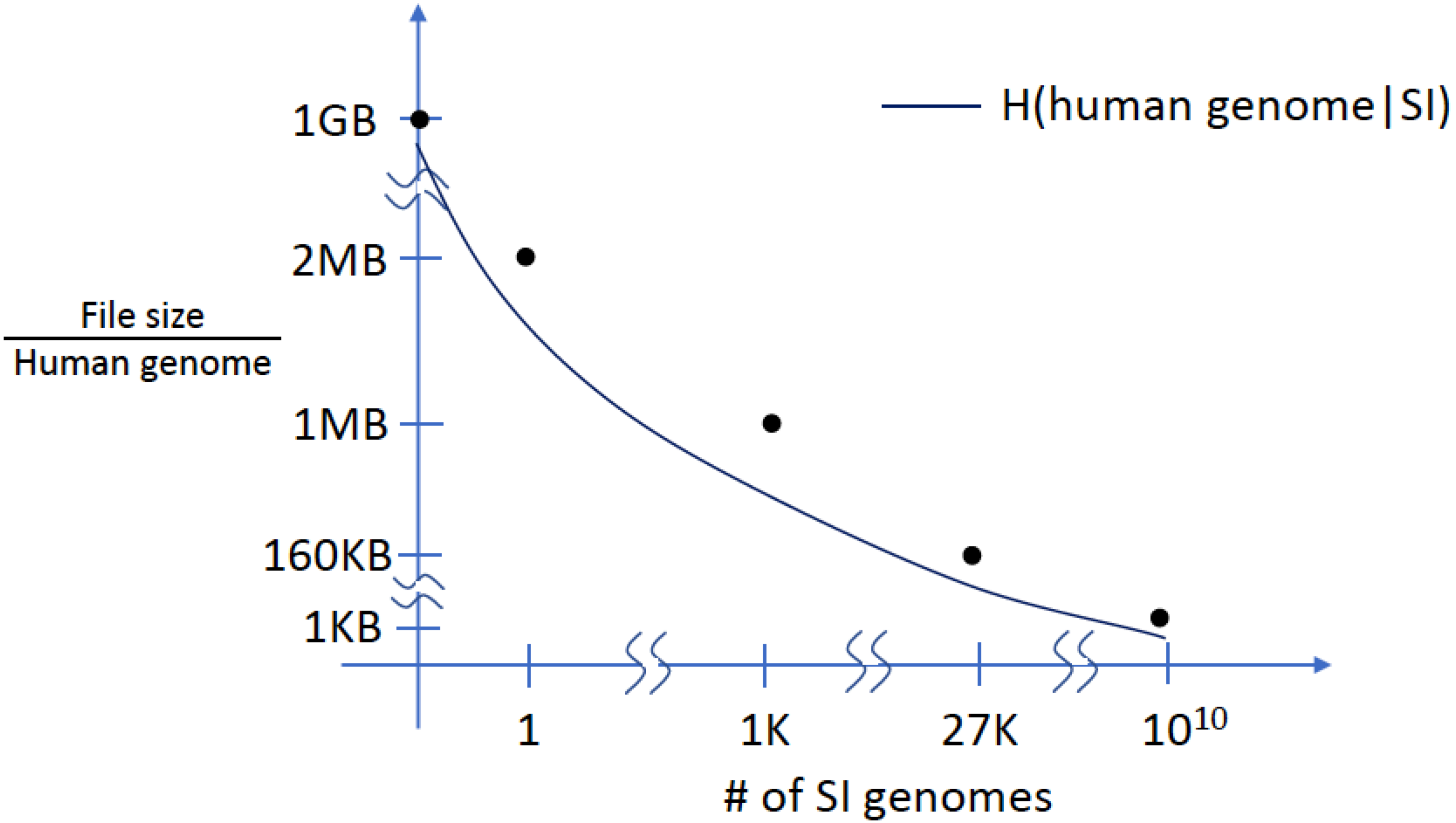
[3] K. Tatwawadi, M. Hernaez, I. Ochoa, *et al.*, “GTRAC: fast retrieval from compressed collections of genomic variants”, *Bioinformatics*, vol. 32, no. 17, pp. i479–i486, 2016.

[4] A. Danek and S. Deorowicz, “GTC: how to maintain huge genotype collections in a compressed form”, *Bioinformatics*, 2018.

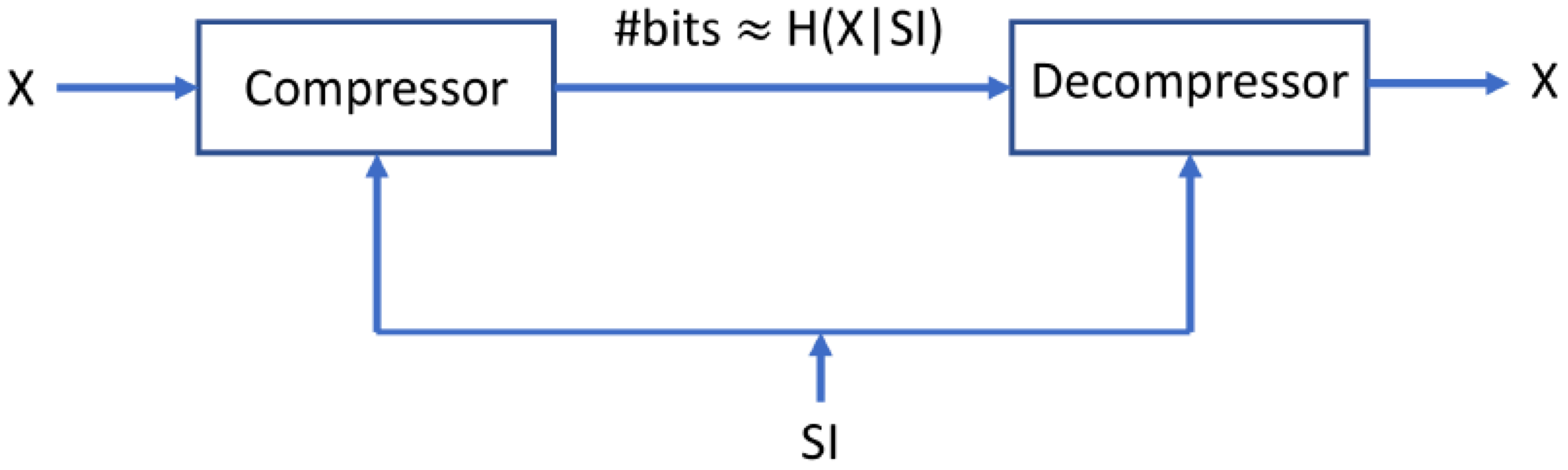
Full Genome Compression

- $H(\text{human genome}) \lesssim 1 \text{ GB}$
- $H\left(\begin{array}{c} \mathit{human} \\ \mathit{genome} \end{array}\middle| \begin{array}{c} \mathit{another} \\ \mathit{human} \\ \mathit{genome} \end{array}\right) \lesssim 2.5MB$
- $H\left(\begin{array}{c} \mathit{human} \\ \mathit{genome} \end{array}\middle| \begin{array}{c} 1K \\ \mathit{other} \\ \mathit{genomes} \end{array}\right) \lesssim 1MB$
- $H\left(\begin{array}{c} \mathit{human} \\ \mathit{genome} \end{array}\middle| \begin{array}{c} 27K \\ \mathit{other} \\ \mathit{genomes} \end{array}\right) \lesssim 200KB$

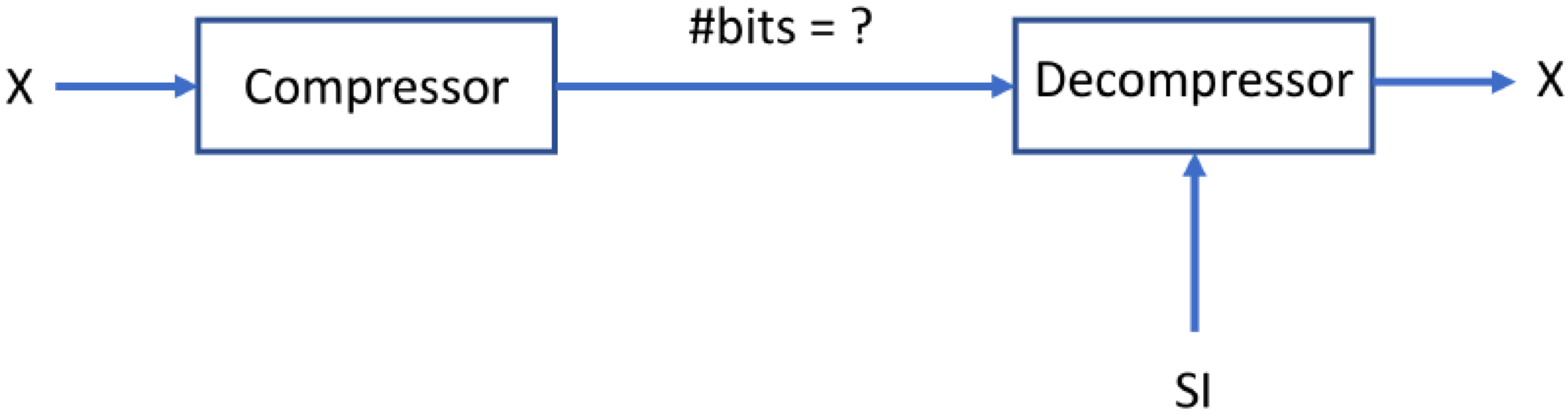
File Size per Genome Vs Database Size



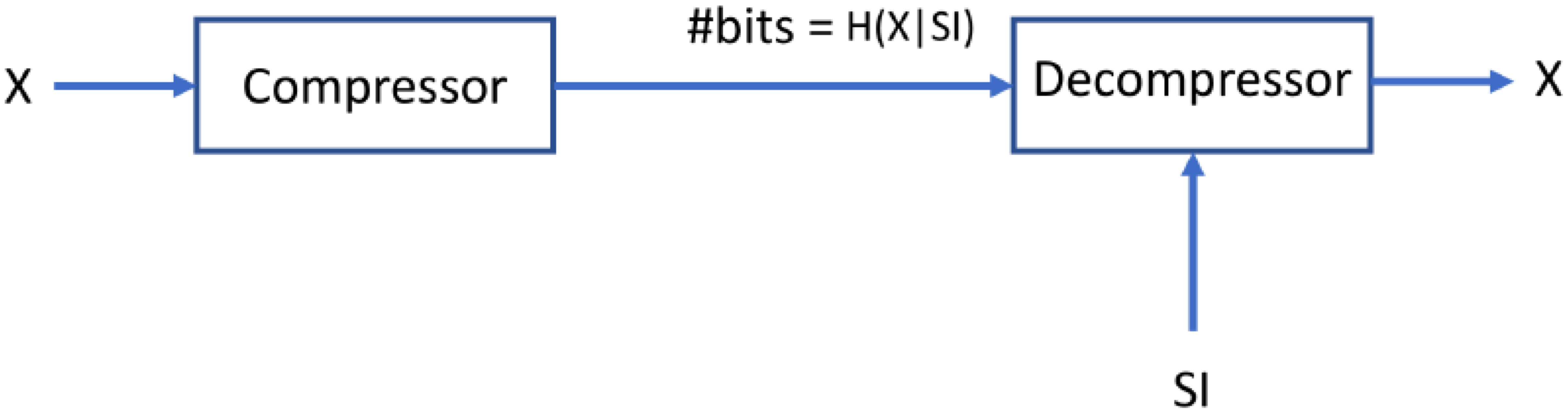
Information Theoretic Perspective



Information Theoretic Perspective

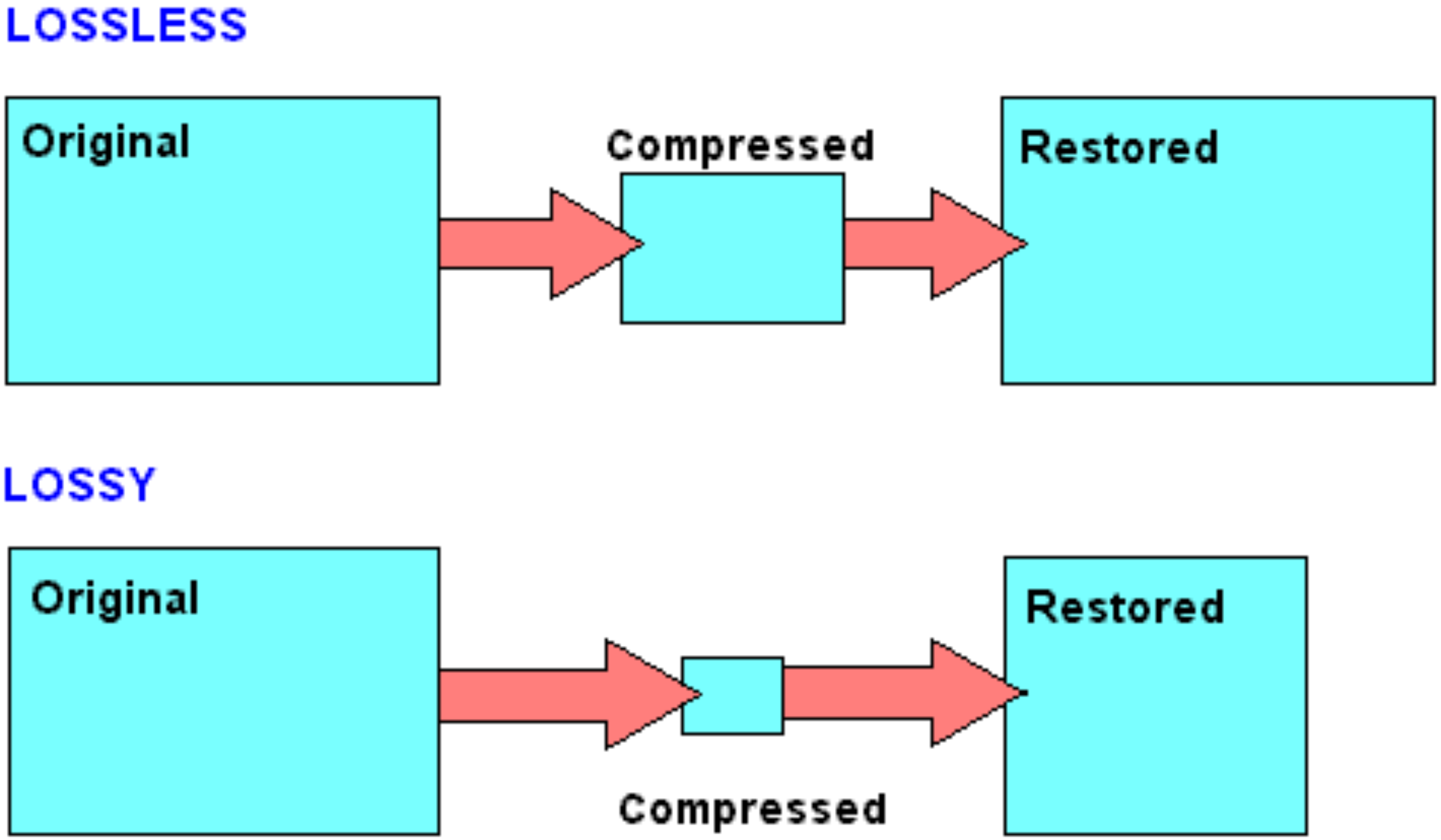


Information Theoretic Perspective



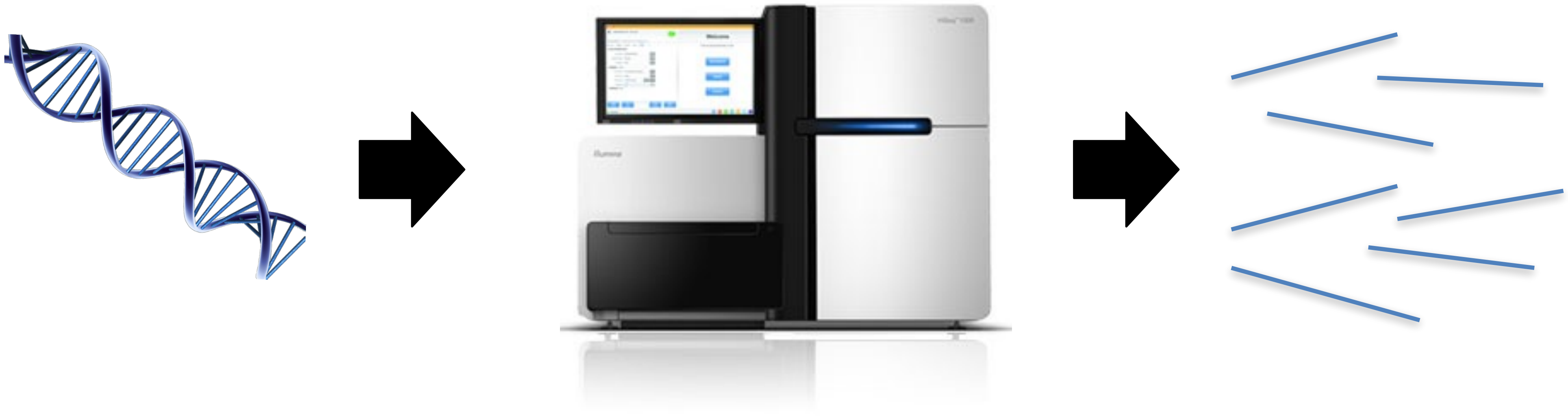
Slepian, David S.; Wolf, Jack K. (July 1973).
"Noiseless coding of correlated information sources"

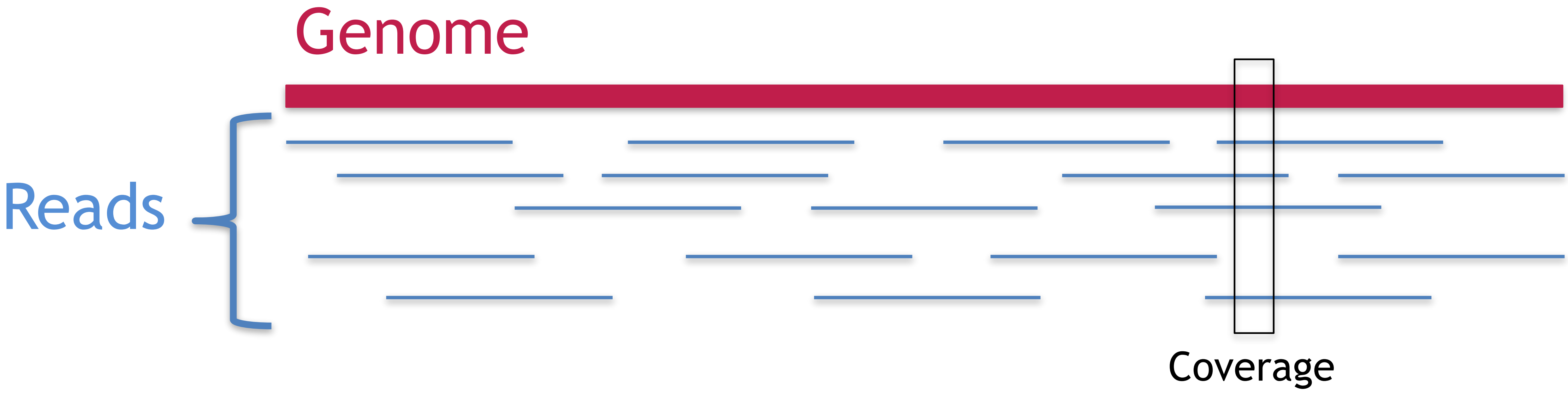
why lossy compression?

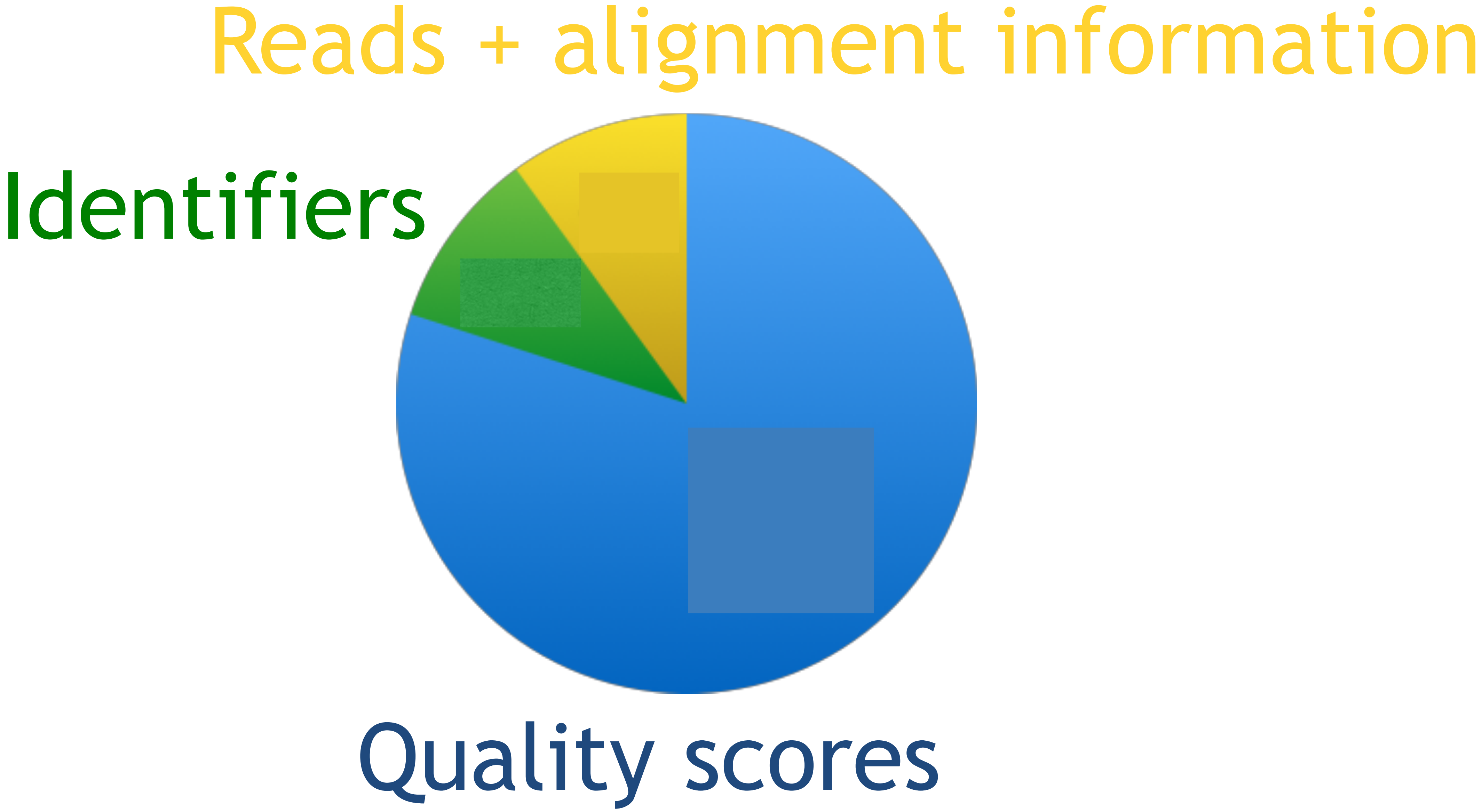


why *lossy* compression of *genomic* data?

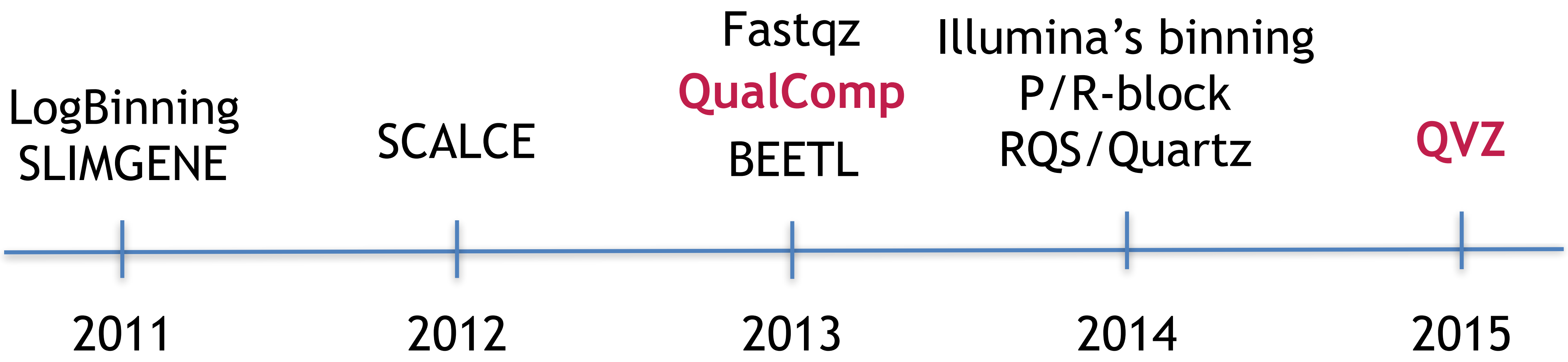
genome sequencing






Reads + alignment information
Identifiers
Quality scores

lossy compressors of quality scores



LogBinning
SLIMGENE

2011

SCALCE

2012

Fastqz
QualComp
BEETL

2013

Illumina’s binning
P/R-block
RQS/Quartz

2014

QVZ

2015

“QualComp: a new lossy compressor for quality scores based on rate distortion theory”

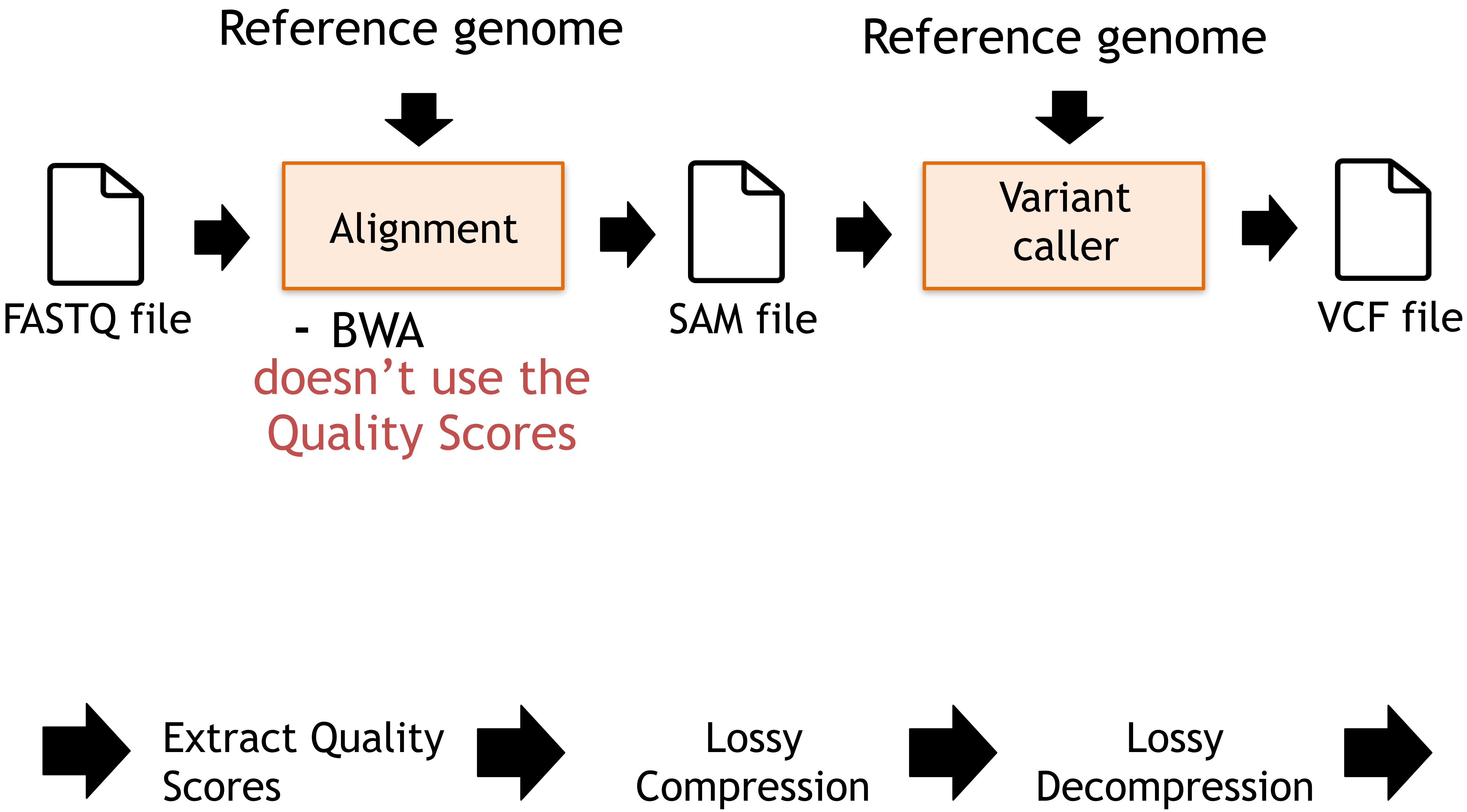
“QVZ: lossy compression of quality values”

how does lossy compression of quality scores affect the inference?

“Effect of lossy compression of quality scores on variant calling”

with Idoia Ochoa, Mikel Hernaez, Rachel Goldfeder and Euan Ashley

Briefings in Bioinformatics, 2017



Reference genome

Reference genome

FASTQ file

Alignment

- BWA

doesn't use the Quality Scores

SAM file

Variant caller

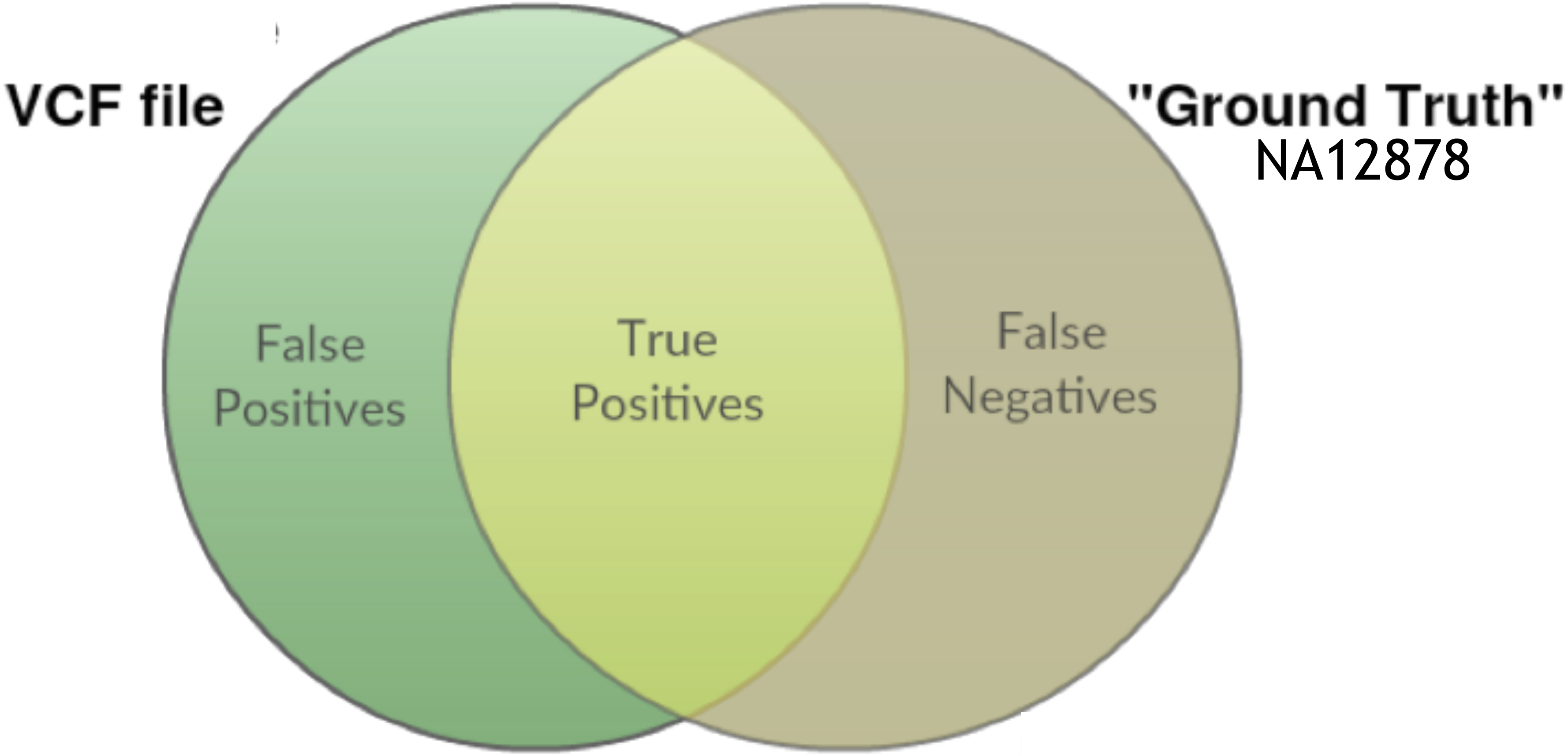
VCF file

Extract Quality Scores

Lossy Compression

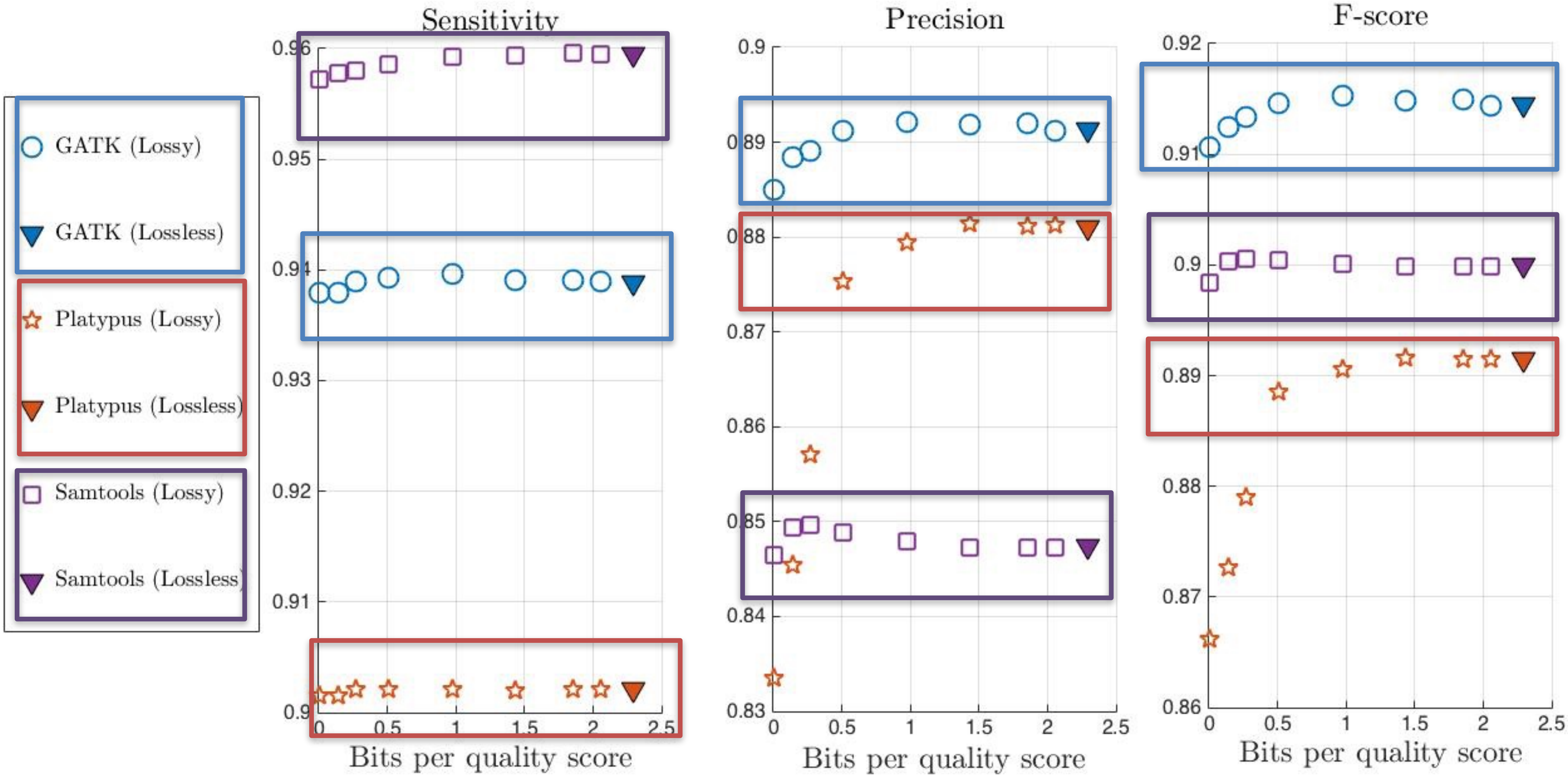
Lossy Decompression

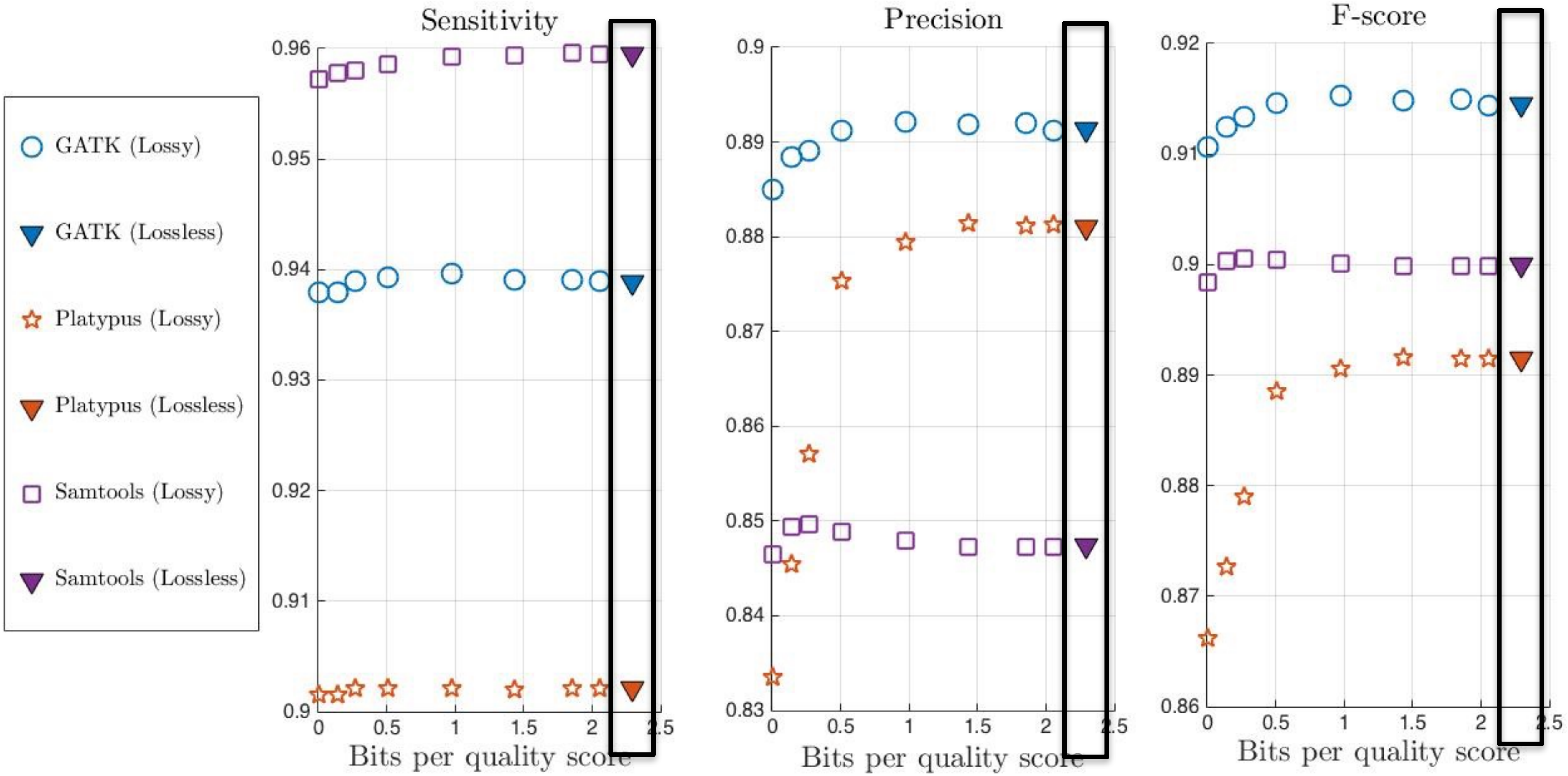
performance metrics



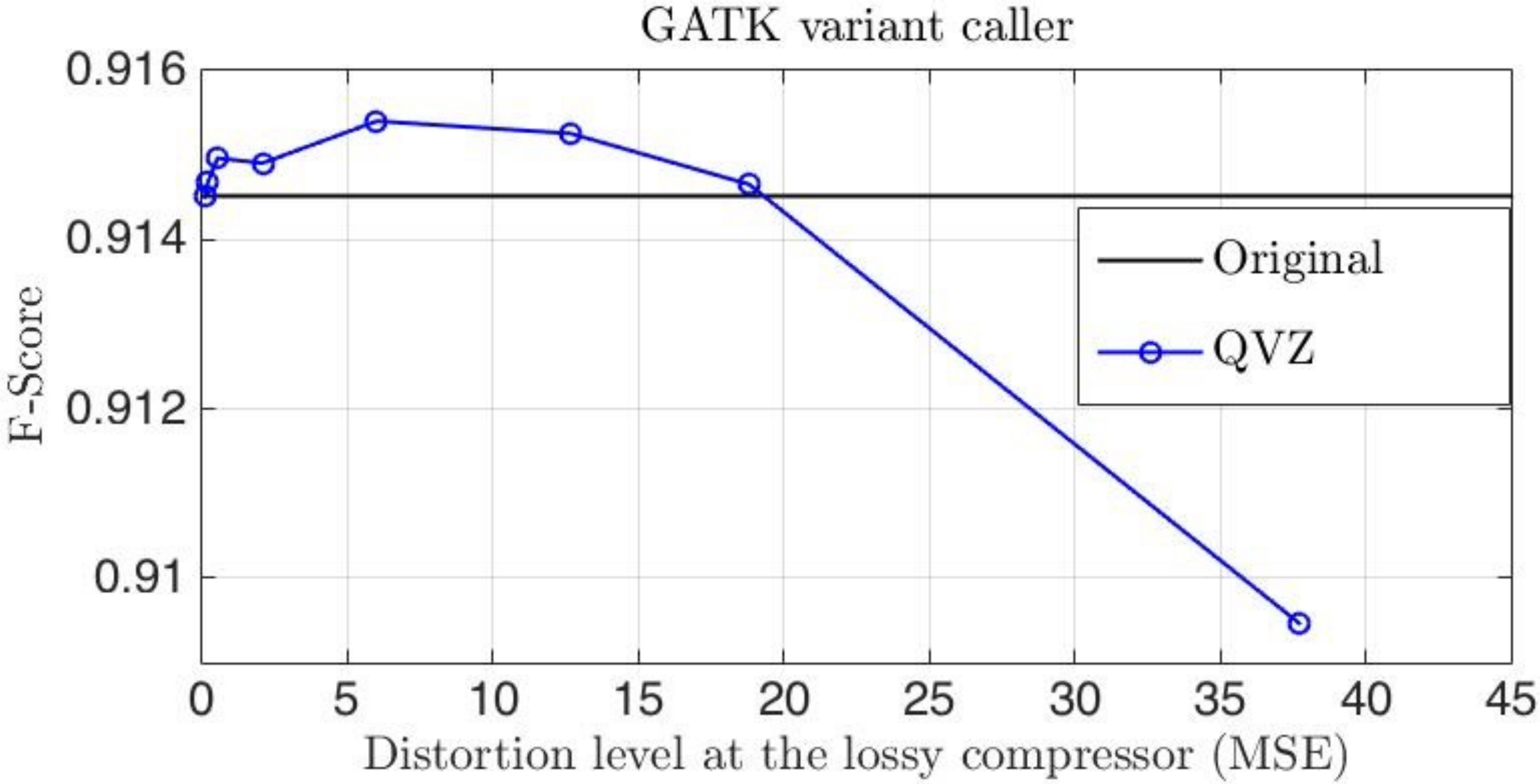
- *sensitivity*: T.P. / (T.P. + F.N.)
- *precision*: T.P. / (T.P. + F.P.)
- *F-score*: Harmonic mean of sensitivity and precision

NA12878, Chr. 11

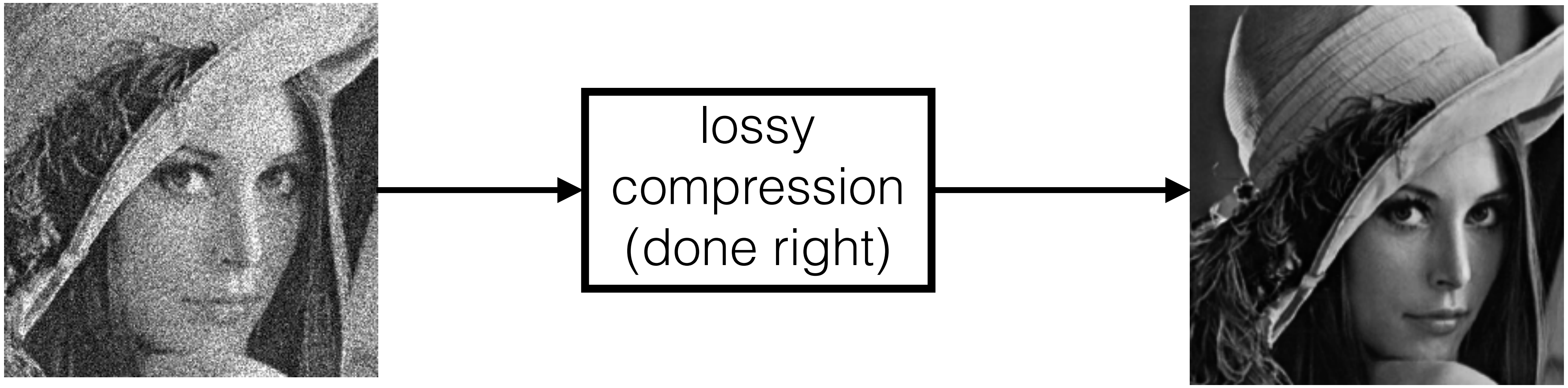



Sensitivity
Precision
F-score
○ GATK (Lossy)
▼ GATK (Lossless)
☆ Platypus (Lossy)
▼ Platypus (Lossless)
□ Samtools (Lossy)
▼ Samtools (Lossless)
0.96
0.95
0.94
0.93
0.92
0.91
0.9
0.9
0.89
0.88
0.87
0.86
0.85
0.84
0.83
0.92
0.91
0.9
0.89
0.88
0.87
0.86
0
0.5
1
1.5
2
2.5
Bits per quality score
Bits per quality score
Bits per quality score

zooming in


GATK variant caller
0.916
0.914
0.912
0.91
F-Score
0
5
10
15
20
25
30
35
40
45
Distortion level at the lossy compressor (MSE)
Original
QVZ

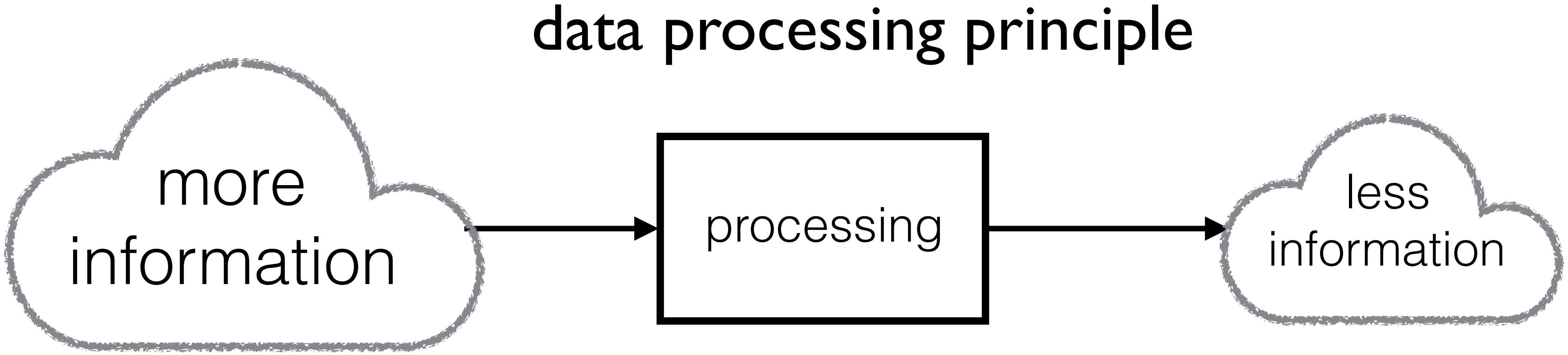
denoising via lossy compression



lossy
compression
(done right)

- “Occam filters”
- “Compresstimation”
- etc.

violating the data processing principle ?



suggests need to lift hood of:

- variant callers
- assemblers
- etc.

thanks