# Exploring Lossy Compression of Gene Expression Matrices

DRBSD-5 Presentation

**Cole McKnight**, Alexandra Poulos, Reed Bender, Jon Calhoun, Alex Feltus

#ClemsonSC19

# Summary

Gene Expression Matrices (GEMs) are a fundamental data type in the genomics domain.

As the size and scope of genomics experiments increase, researchers are struggling to process large GEMs through downstream workflows.

~20M Yeast GEM -> 6GB KINC run (122 samples, ~8,000 genes)
~635MB CCLE GEM -> 1.6TB KINC run (1019 samples, 56,000 genes)
~4.1GB Ath GEM -> N/A (22,000 samples, ~27,000 genes) - cannot process

We developed GEMTrim, a methodology to compress GEMs in order to reduce the resources necessary for processing.

Using GEMTrim resulted in significant space savings while still preserving the biological integrity of the data.

Researchers can apply lossy compression to Gene Expression Matrices to enable the downstream processing of previously inaccessible GEMs.

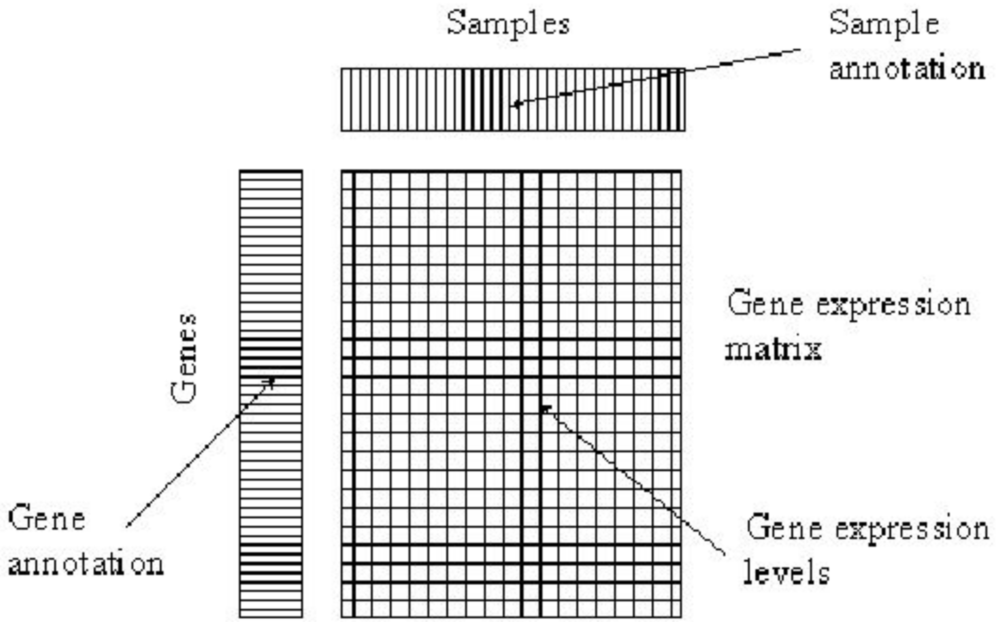
# Background

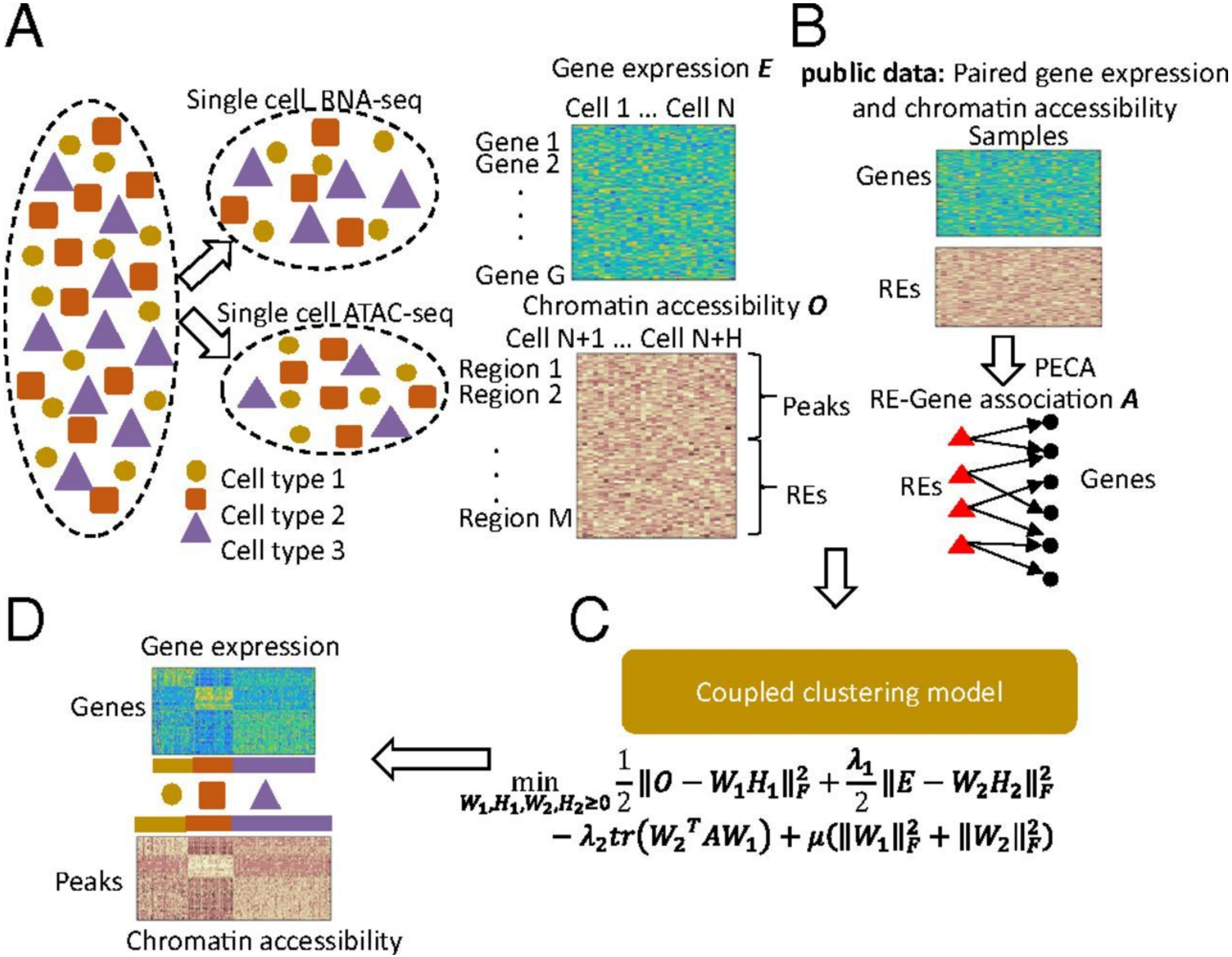
# What are Gene Expression Matrices?

A Gene Expression Matrix(GEM) compares the expression of *m* genes across *n* samples.

GEMs are composed of a *m x n* tab-delimited matrix of floating point expression values.

The first row of a GEM has sample ID headers, the first column has gene ID headers.





$$\min_{W_1,H_1,W_2,H_2\geq 0} \frac{1}{2}\|O - W_1H_1\|_F^2 + \frac{\lambda_1}{2}\|E - W_2H_2\|_F^2 - \lambda_2 tr\left(W_2{}^T A W_1\right) + \mu(\|W_1\|_F^2 + \|W_2\|_F^2)$$

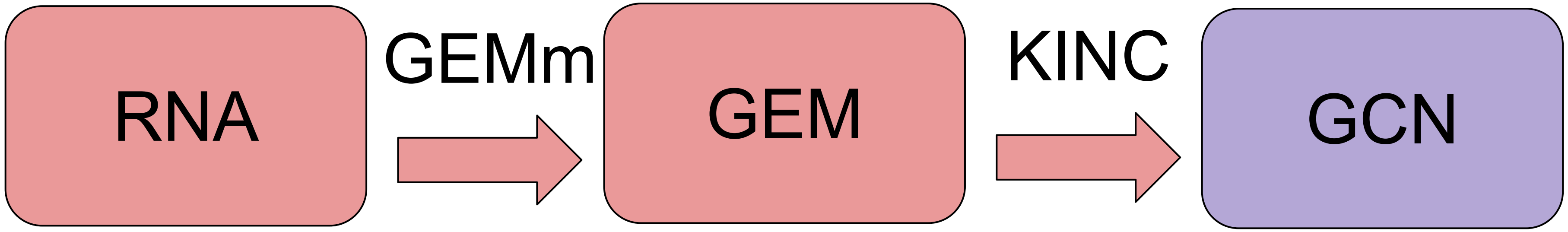
# KINC

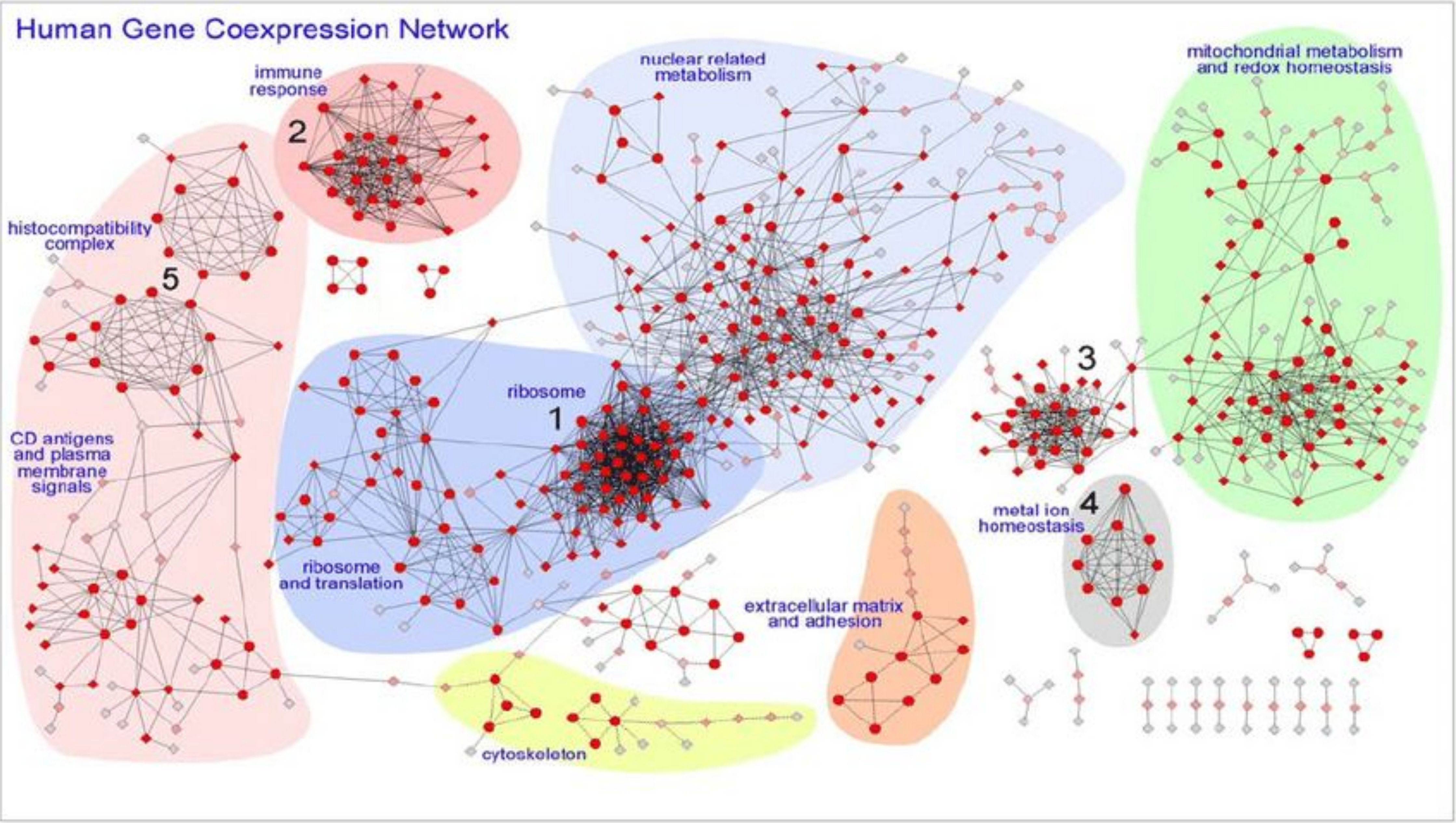
Knowledge Independant Network Construction(KINC)

Used to generate Gene Co-expression Networks(GCNs)

Basic Steps:

- Preparation of input data.
- Pairwise correlation of all genes with every other gene.
- Thresholding of non-significant correlation values.
- Module discovery.




Human Gene Coexpression Network
immune response
2
nuclear related metabolism
mitochondrial metabolism and redox homeostasis
histocompatibility complex
5
ribosome
1
3
CD antigens and plasma membrane signals
metal ion homeostasis
4
ribosome and translation
extracellular matrix and adhesion
cytoskeleton

# What is Lossy Compression?

Two types of compression: lossless and lossy.

Lossless ensures the decompressed data is **identical** to the original.

But do we really **need** identical data?

Lossy compression removes some data during compression, then utilizes a number of methods to **approximate** the data when decompressed.

Using lossy compression results in better compression ratios than even the best lossless methods.

# Basic Idea

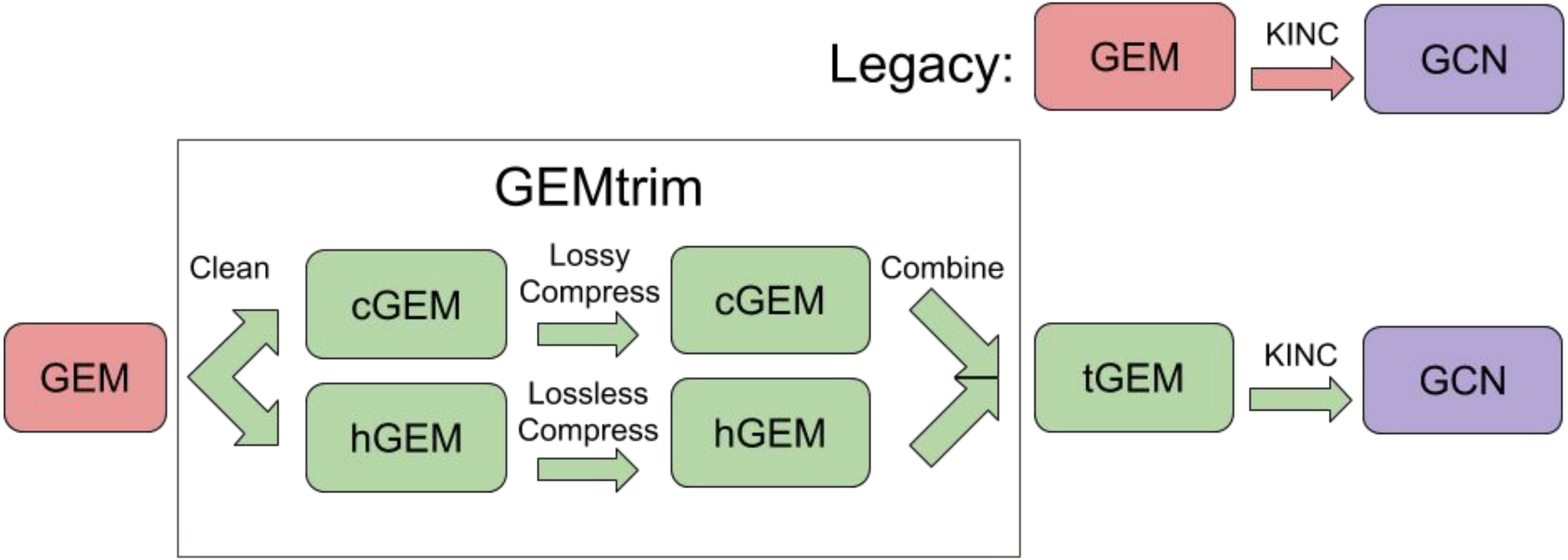
Use lossy compression on GEMs:

- reduce size
- reduce resource requirements for downstream processing

Test if data retains biological integrity

- run GEM through KINC to produce GCN
- compare to control GCN

# Methods



Legacy: GEM → KINC → GCN

GEMtrim

GEM → Clean → cGEM → Lossy Compress → cGEM → Combine → tGEM → KINC → GCN

GEM → Clean → hGEM → Lossless Compress → hGEM → Combine → tGEM → KINC → GCN

# Experiment(Pt 1)

40 initial compression configurations of the Yeast GEM.

3 lossy compressors, 6 lossless compressors.

40 versions of the Yeast GEM were processed through KINC.

Resulting networks were compared with control to determine validity.

Validation metrics:

- # of edges in resulting GCN
- Thresholding value

# Experiment(Pt 2)

Only identical networks “passed”

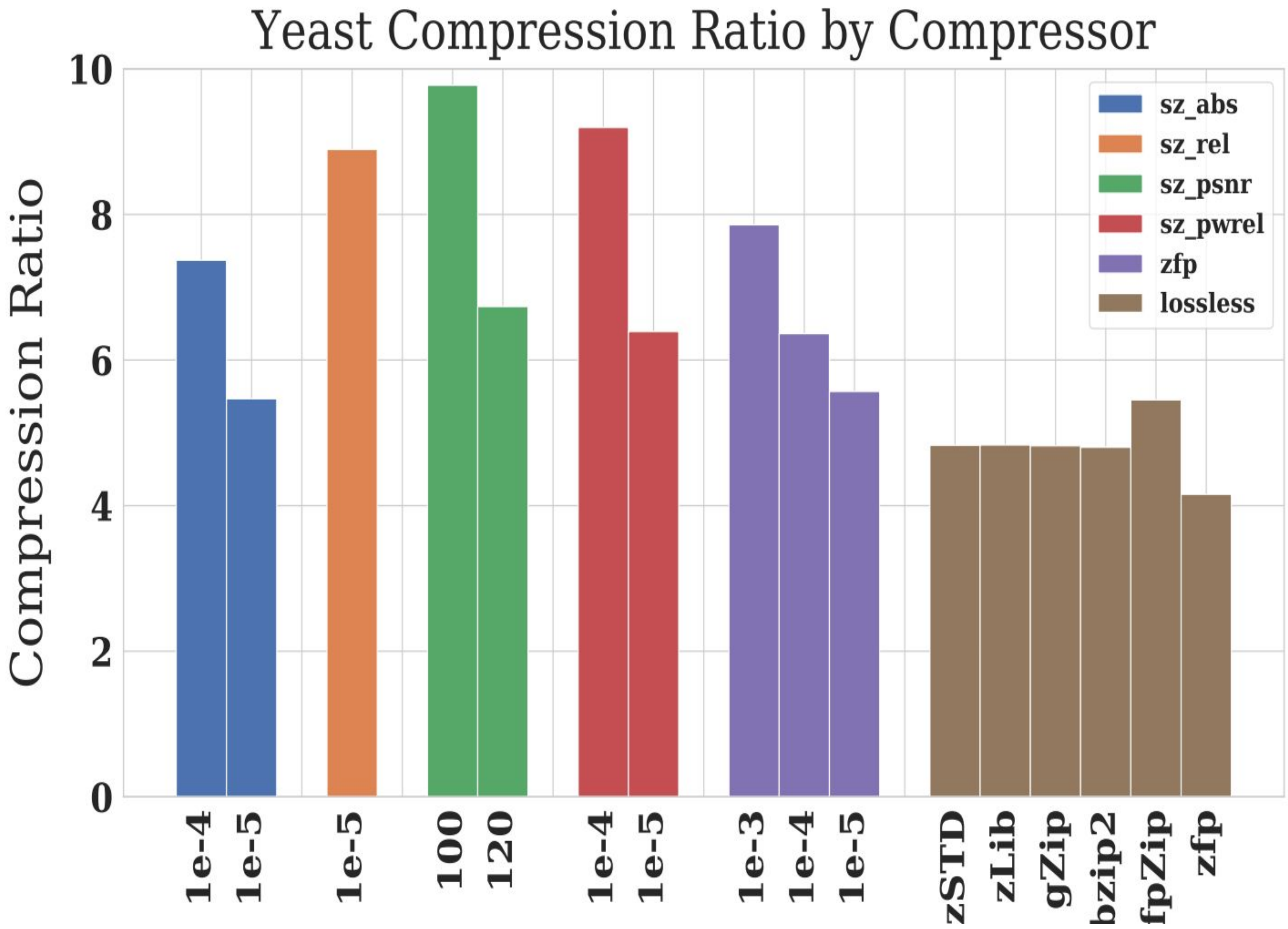
9 compression configurations passed validation of Yeast runs.

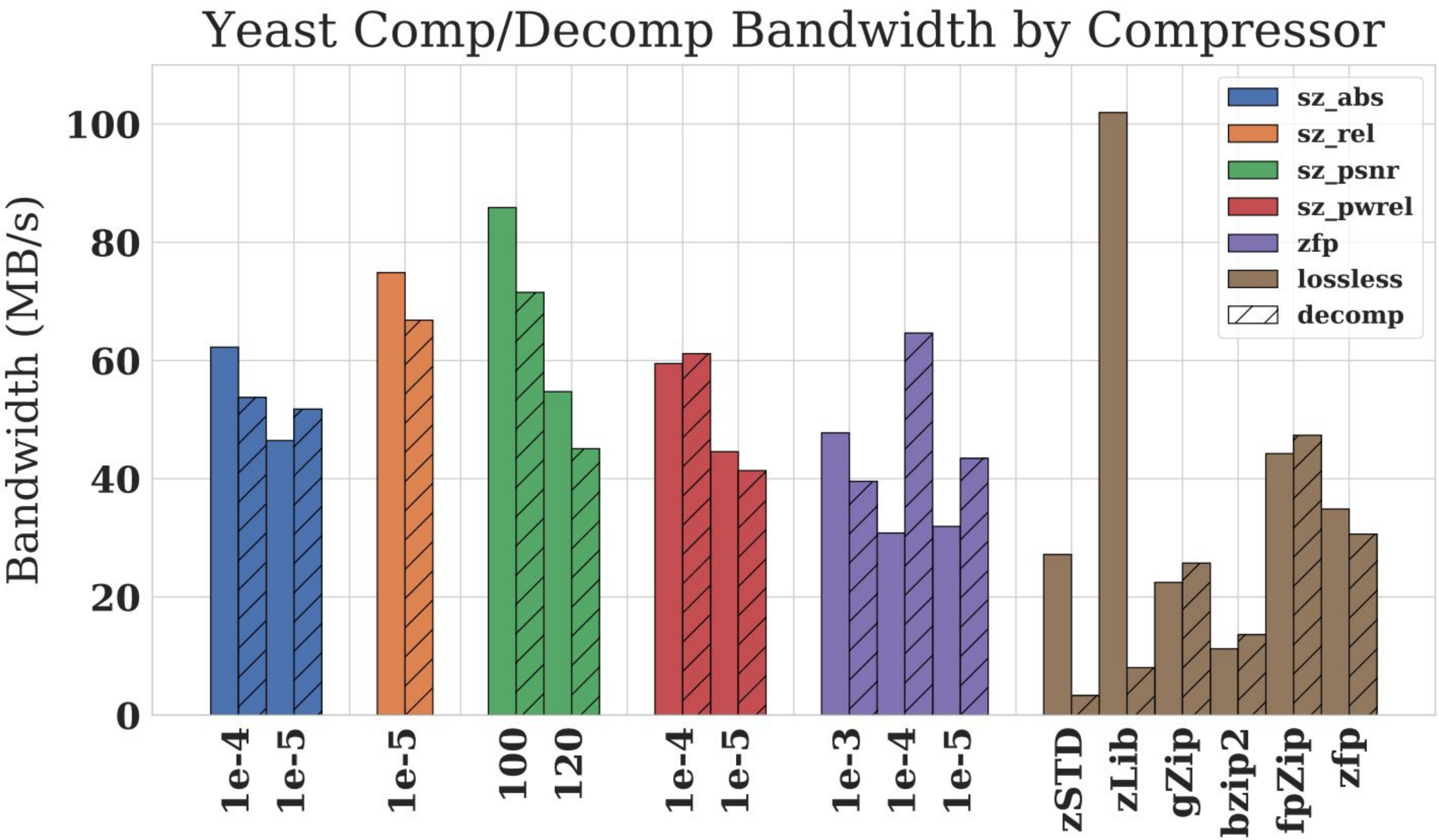
Each of the 9 was used to compress the Cancer Cell Line Encyclopedia(CCLE) GEM.

Performance metrics were taken from both Yeast and CCLE compression runs.

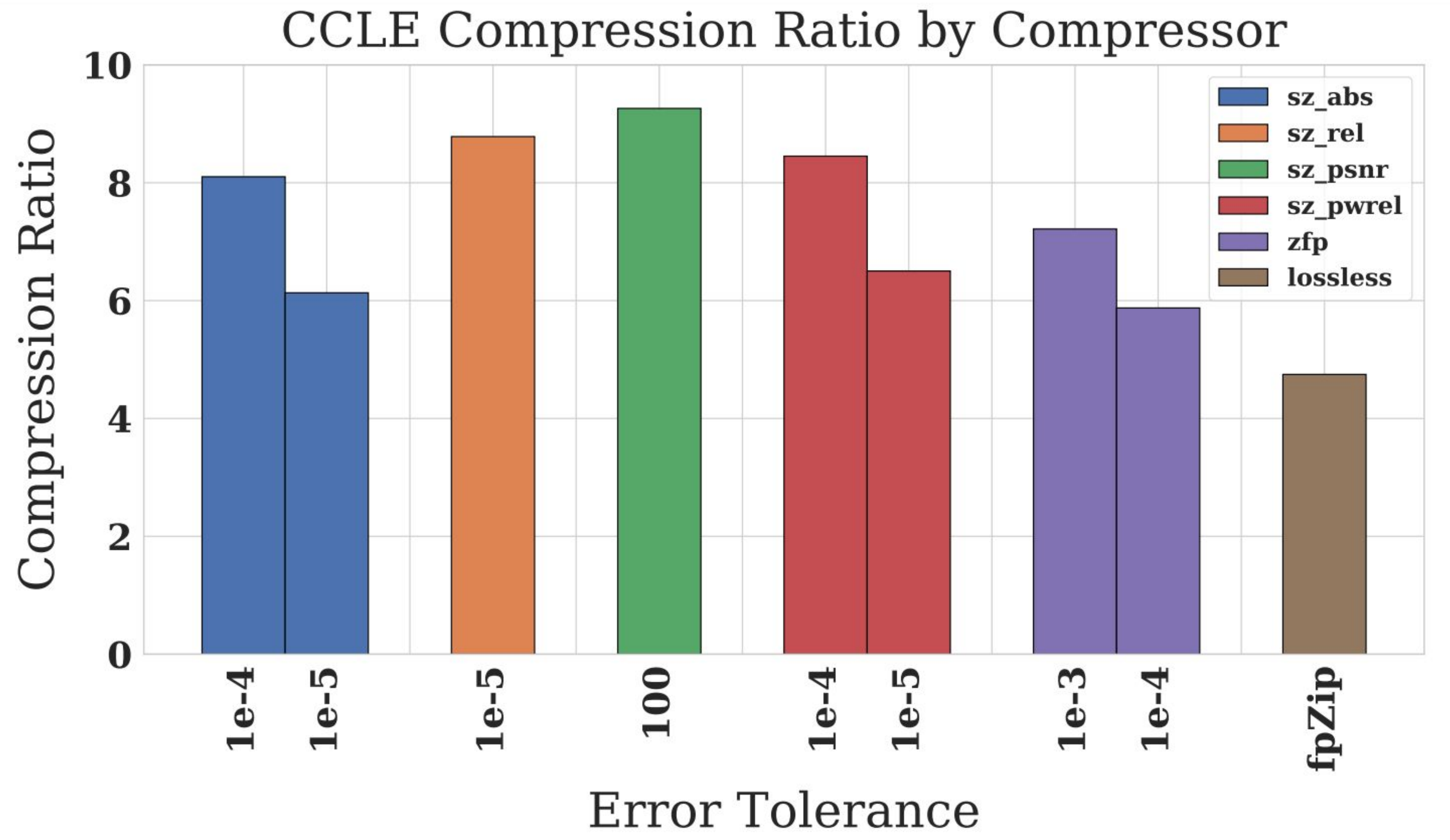
# Results & Discussion

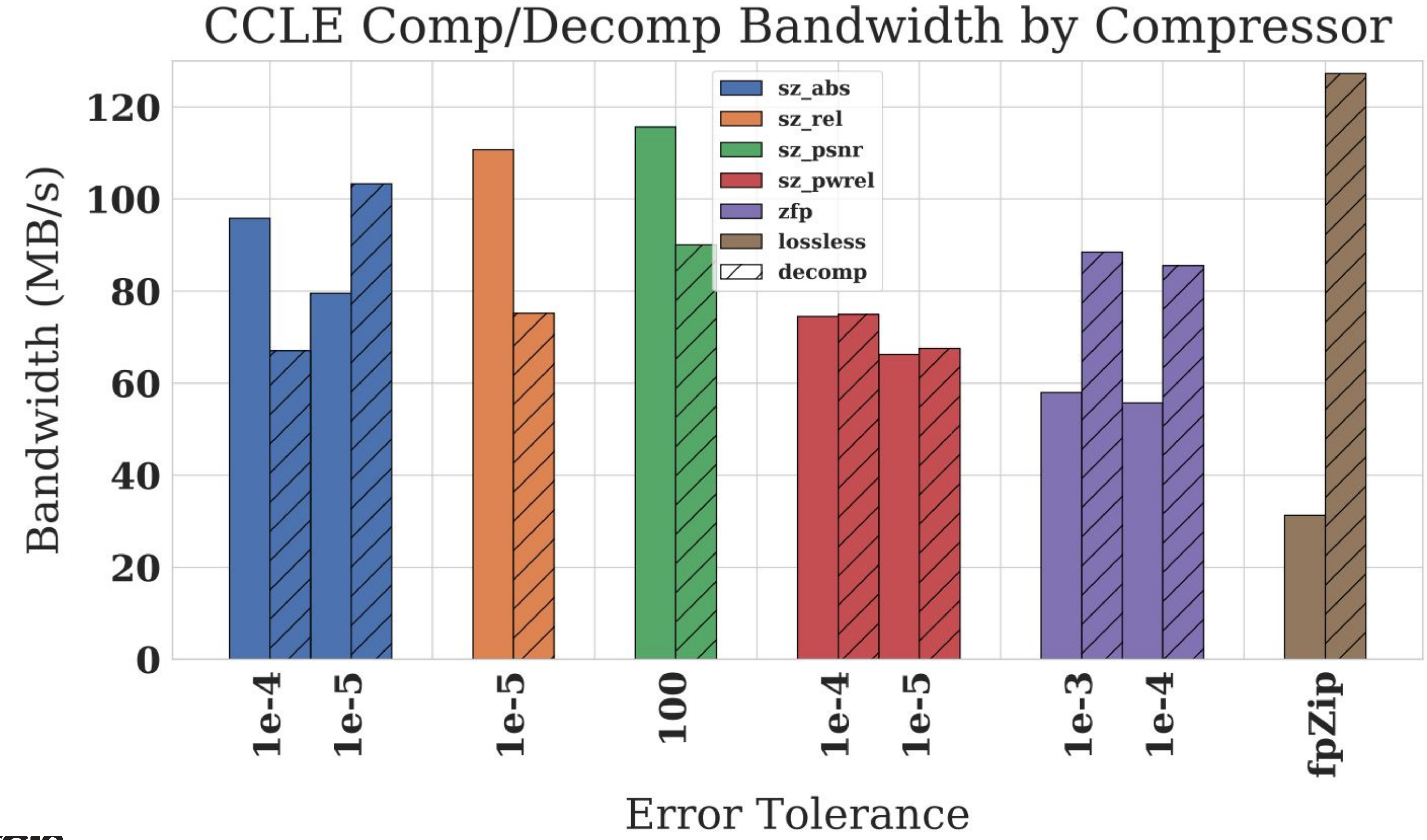
# Yeast


Yeast Compression Ratio by Compressor
Compression Ratio
0
2
4
6
8
10
sz_abs
sz_rel
sz_psnr
sz_pwrel
zfp
lossless
1e-4
1e-5
1e-5
100
120
1e-4
1e-5
1e-3
1e-4
1e-5
zSTD
zLib
gZip
bzip2
fpZip
zfp


Yeast Comp/Decomp Bandwidth by Compressor
Bandwidth (MB/s)
0
20
40
60
80
100
1e-4
1e-5
1e-5
100
120
1e-4
1e-5
1e-3
1e-4
1e-5
zSTD
zLib
gZip
bzip2
fpZip
zfp
sz_abs
sz_rel
sz_psnr
sz_pwrel
zfp
lossless
decomp

# CCLE


CCLE Compression Ratio by Compressor
Compression Ratio
10
8
6
4
2
0
sz_abs
sz_rel
sz_psnr
sz_pwrel
zfp
lossless
1e-4
1e-5
1e-5
100
1e-4
1e-5
1e-3
1e-4
fpZip
Error Tolerance


CCLE Comp/Decomp Bandwidth by Compressor
Bandwidth (MB/s)
0
20
40
60
80
100
120
sz_abs
sz_rel
sz_psnr
sz_pwrel
zfp
lossless
decomp
1e-4
1e-5
1e-5
100
1e-4
1e-5
1e-3
1e-4
fpZip
Error Tolerance

# Results Summary

Utilizing a combination of lossy and lossless compression results in compression ratios up to 9.77× on a Yeast GEM, while still preserving the biological integrity of the data.

Usage of the compression methodology on the Cancer Cell Line Encyclopedia(CCLE) GEM resulted in compression ratios up to 9.26×.

The compression method that resulted in the best compression ratios for both Yeast and CCLE was SZ with the PSNR error bound set to 100.

Compression/decompression bandwidth is not a limiting factor at this point.

# Conclusion

# Future Work

Modify KINC to take compressed data as input - partial decompression at runtime.

Conduct CCLE KINC runs and test validity, see if valid methods for Yeast scale.

Test even larger GEMs.

Test GEMTrim with other workflows, not just KINC.

Develop methodology for determining optimal error bound for a GEM of a given size.

# Acknowledgement

This material is based upon work supported by NSF Grants No. SHF-1910197 and CC*-1659300.

# Key Takeaways

By using GEMTrim, researchers in the Genomics domain may be able to process previously inaccessible GEMs while realizing significant reduction in computational costs.

Lossy compression can be applied to most types of quantitative data.

Can lossy compression be used in your research?

# Questions?