

# Compression of Microarray Images and Its Statistical Implications

Rebecka Jörnsten

Bin Yu

Departments of Statistics

Rutgers University and UC Berkeley

<http://www.stat.rutgers.edu/~rebecka>

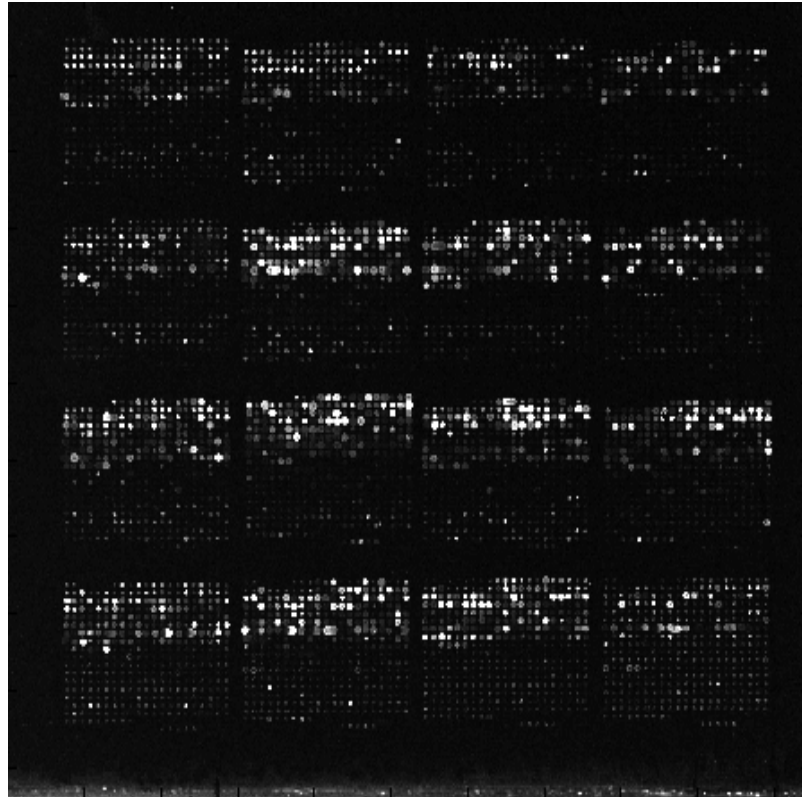
## Outline

1. Microarray Images
2. Information extraction
3. Lossless and progressive compression.
  - Segmented LOCO, and
  - Residual Bitplane Encoding.
4. Results
5. Conclusions

## Microarray Images

Powerful tool for monitoring the expression of thousands of genes simultaneously:

- Identifying gene functions
- dynamics
- pathological context

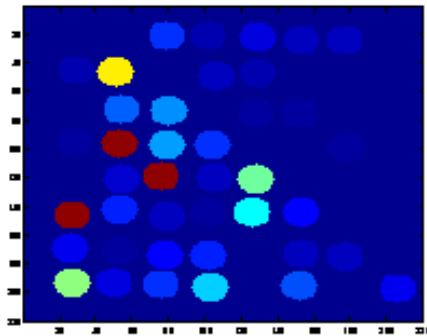


Massive amounts of image data collected in many labs.

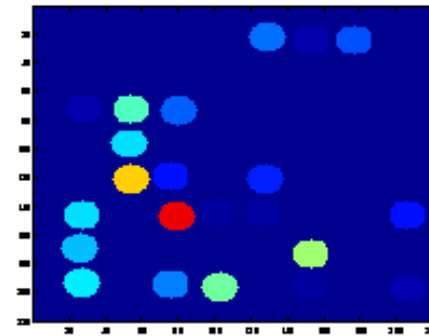
**Current focus:** Development of standards, need for data sharing, data bases.

# Microarray Images

Ideal output

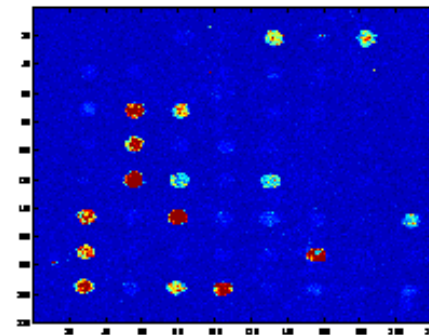
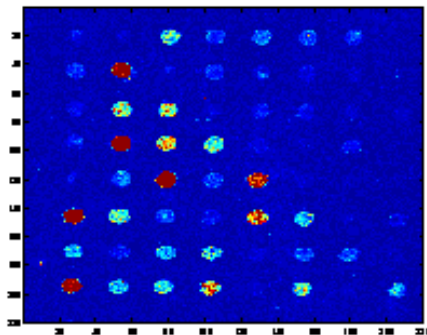


Reference Sample



Sample of Interest

Actual: 10000 spots and **30Mb** (16 bpp), for each array



## Extracting information from the images

1. **Registration:** Where are the spots? (*peak finding*)
2. **Segmentation:** Where did hybridization occur? (*fixed circle, adaptive circle, adaptive shape segmentation*)
3. **Background correction:** Non-specific hybridization. Global, local, sampling, filtering? (*valley-between-peaks, min/max filters*)
4. **Data summary:** Log of background corrected mean intensity ratios.

Comparison study: Yang, Buckley, Dudoit and Speed, UCB technical report (2000).

## The full image data is always kept, because

- Experiments are expensive.
- Methods for image processing are still under development.
- New focus on the development of standards, data sharing, data bases.

However, each image is HUGE (30 MB per scan).

Hence, the need for

**image compression with a data structure to facilitate analysis** to “maximally exploit and share data”.

NHGRI (National Human Genome Research Institute) meeting.

## Microarray Image Compression Scheme

Lossless and lossy compression of microarray images into meaningful data structures for analysis.

- **Lossless**: statistical redundancy reduction. *Predictive lossless coding to take out dependence or redundancy, followed by encoding of residual bitplanes.*
- **Lossy**: irrelevance reduction. *Encode only the most relevant residual bitplanes. Progressive scheme.*

Irrelevance:

- natural images: features not detectable by eyes.
- microarray images: features not useful for statistical inference.

## Compression Scheme

### 1. Segmented LOCO

- Redundancy reduction, up to a maximum per-pixel error  $\delta$ .
- Defines a minimum decodable bitrate.

### 2. Residual Bitplane encoding

- Irrelevance reduction.
- Bit by bit, add more information to image regions.
- Stop at anytime, progressive scheme.

## Compression Scheme - What is Segmented LOCO?

### LOCO, JPEG 2000:

- Causal fixed predictor.
- Context based adaptive predictor (context defined by local gradients).
- Fixed prediction error quantization  $\delta$ .
- Separate encoders for each context.
- Runlength coding in “flat” regions.

### Segmented LOCO:

- Segmentation=Overhead.  
Identifies Background and Regions of Interest (ROIs) (the spots).
- Code background and ROIs separately.
- Variable prediction error quantization  $\tilde{\delta}$ . Pick  $\tilde{\delta}$  as a function of the local SNR.
- Segmented runlength coding: “tunnel under” the ROIs.

## Compression Scheme - Data Structure

**Overhead:** Segmentation map=Quality measures. Gives us spot locations, spot shapes and sizes, spot means and variances.

*All we need if no reprocessing required.*

**Output of Segmented LOCO:** Spots and background image blocks with locally varying maximum error  $\tilde{\delta}$ .

*Allows for subset image reconstruction.*

**Residual Bitplane Encoding:** If necessary, we can reduce the local error further.

*Allows for locally varying degree of loss.*

Each bitplane reduces the maximum error by a factor of 2. Lossless coding =  $\log_2(2\tilde{\delta} + 1)$  bitplanes.

## Compression Scheme - Design parameters

**Goal: efficient lossless encoding *and* good progressive performance.**

That is, the lossy image reconstructions should be good substitutes for the lossless.

We select local maximum errors  $\tilde{\delta}$  as a function of the spot signal-to-noise ratios.

Different ranges of SNR map to different  $\tilde{\delta}$ .

- We get a minimum decodable bitrate of  $\sim 4$  bpp (1. Segmented LOCO).
- We get a lossless compression ratio of  $\sim 1.8 : 1$ . (2. Bit plane encoding).

## Compression Scheme - Lossless

Why such poor lossless performance?

(Compression ratio  $\sim 1.8 : 1$ ).

- The 8 least significant bits are random.
- This puts a 2:1 ceiling on the lossless compression ratio.
- This also suggests we don't need to store the images at full precision. The images are noisy!

## Comparison with other schemes

**LOCO:** Not progressive.

**SPIHT:** Wavelet based, progressive.

**Wavelet+Zerotree+Entropy coding:** Wavelet based, progressive.

Method	Bit-rate (cmp 32 bpp)	Compr. ratio
LZ (gzip)	21.62	1.48:1
SPIHT	19.44	1.65:1
W+ZT+EC	18.64	1.72:1
LOCO	17.28	1.85:1
Segm. LOCO	17.45	1.83:1

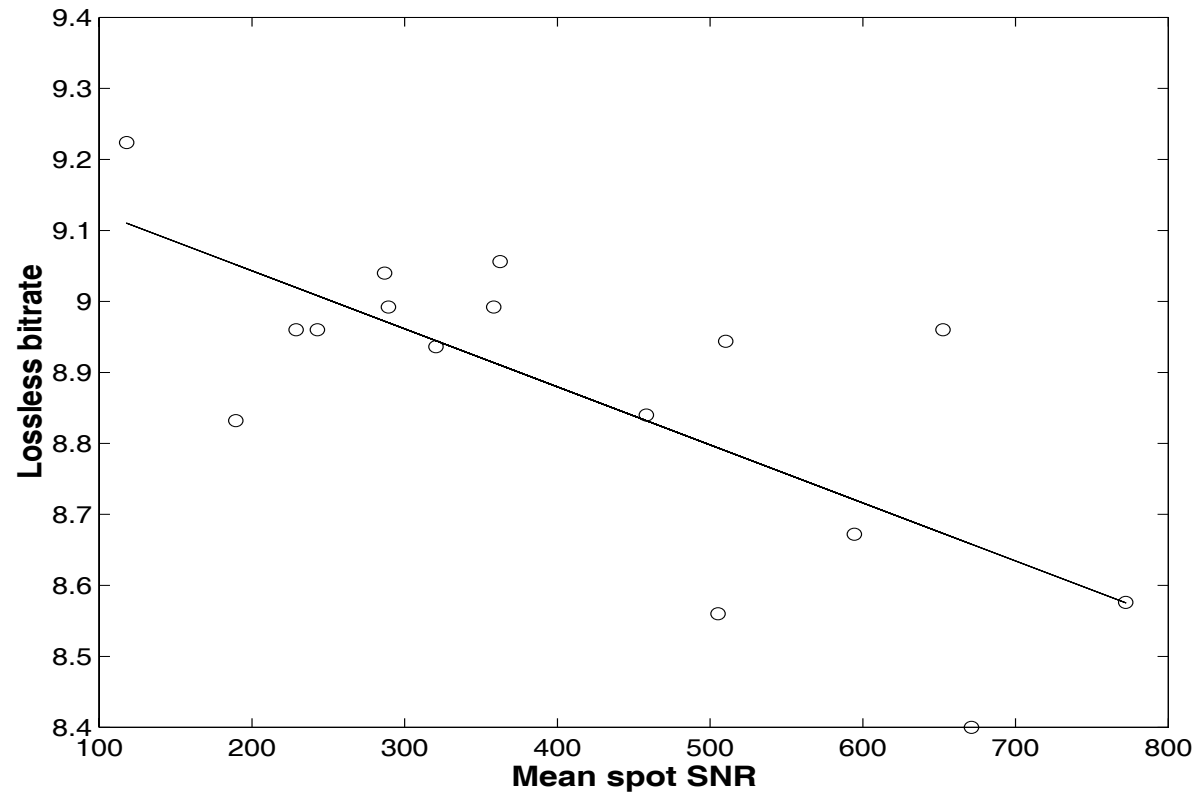
## Comparison with other schemes

SPIHT and W+ZT+EC are state-of-the-art for natural images BUT perform poorly on Microarray images. Why?

**Wavelets:** Wide support leads to smearing.

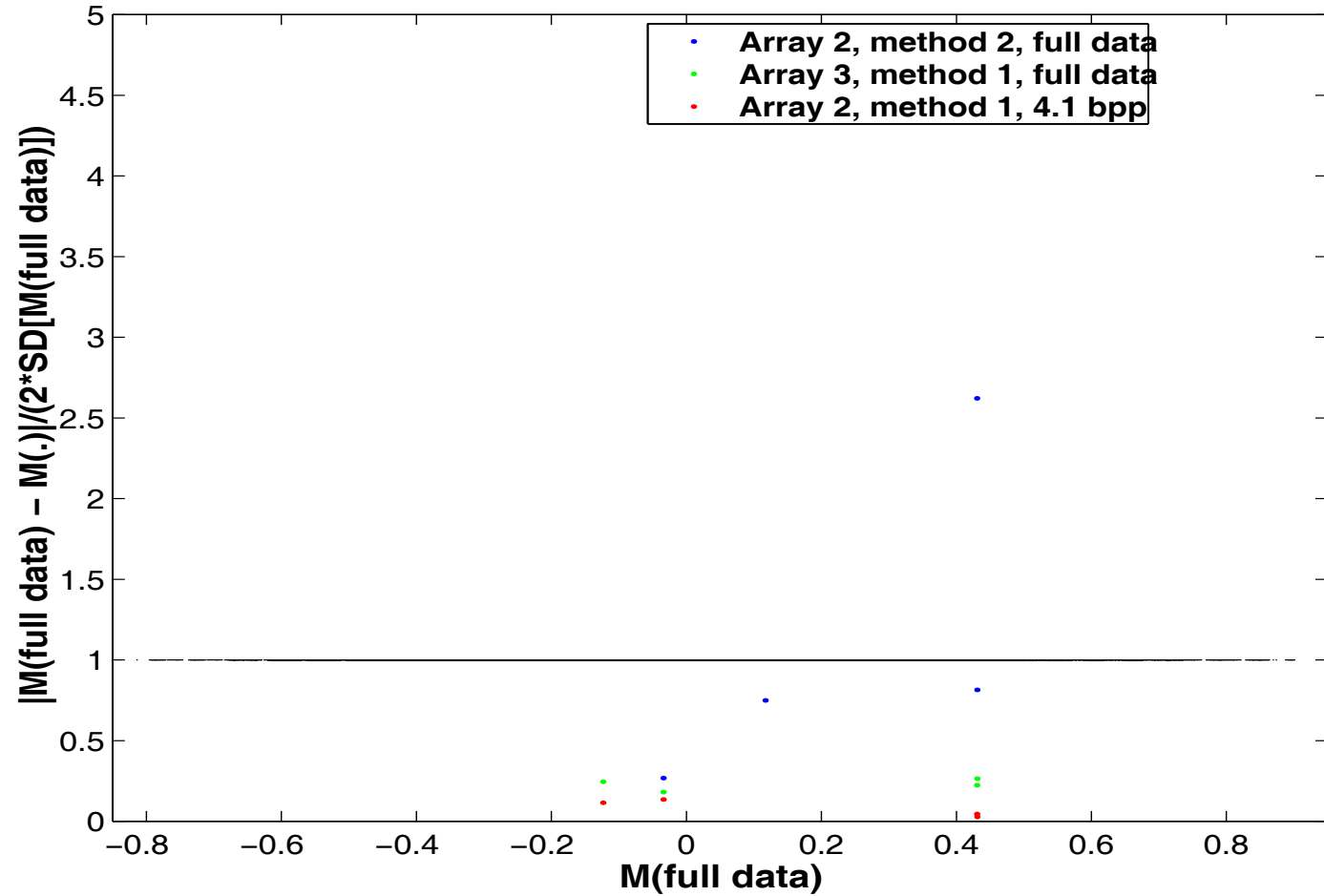
**Zerotree:** Many high-intensity contrast elements (spots) so the zerotree is dominated by non-informative edge structures.

## Lossless Compression and Slide Quality



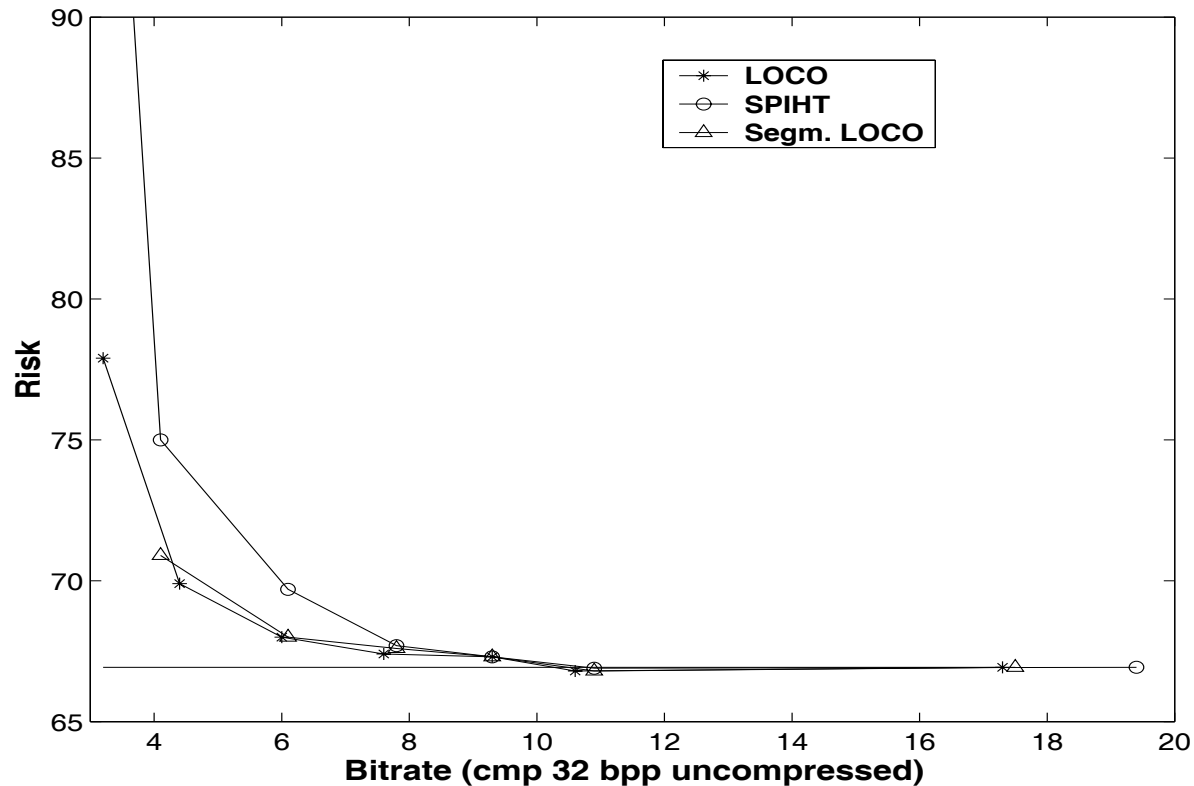
We can sort images in terms of data quality, using the compressed file sizes.

## Lossy Compression Results



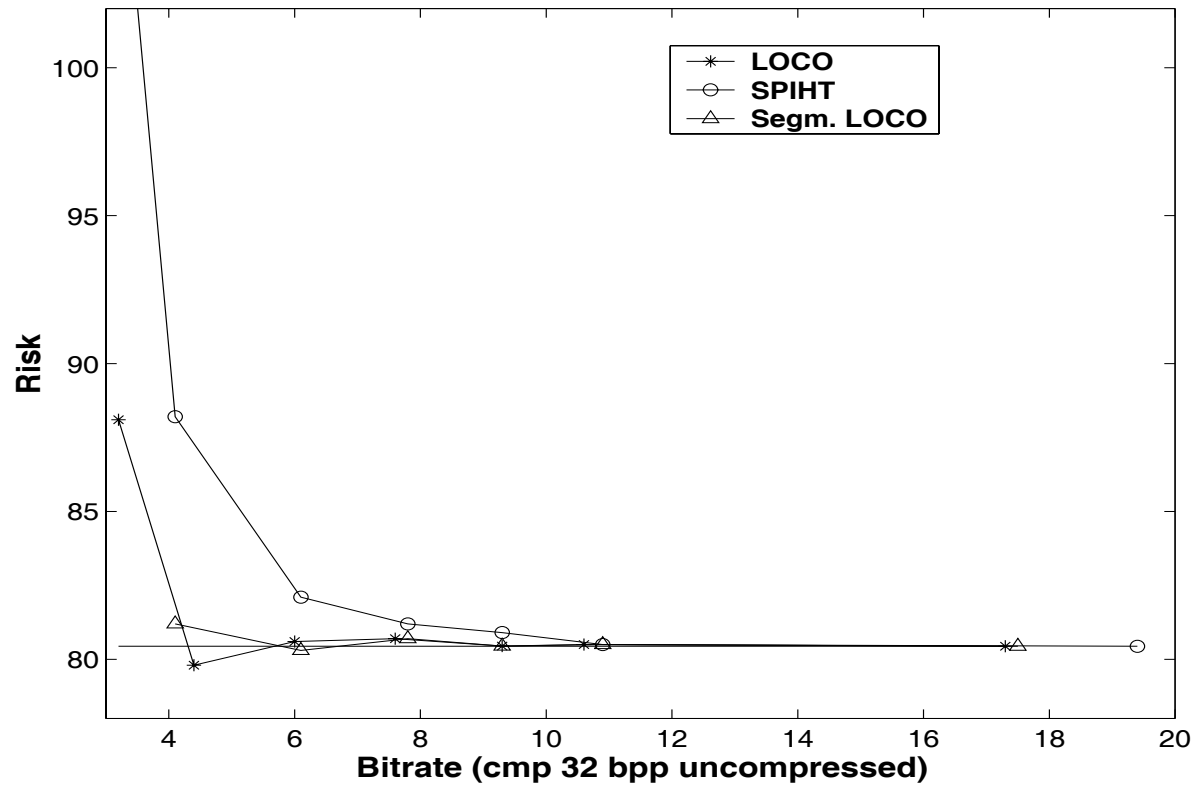
z-scores. Baseline=full data, Blue=Different extraction method,  
Green=Different Array, Red=Lossy compression 4.1bpp (Segm. LOCO)

## Lossy Compression Results



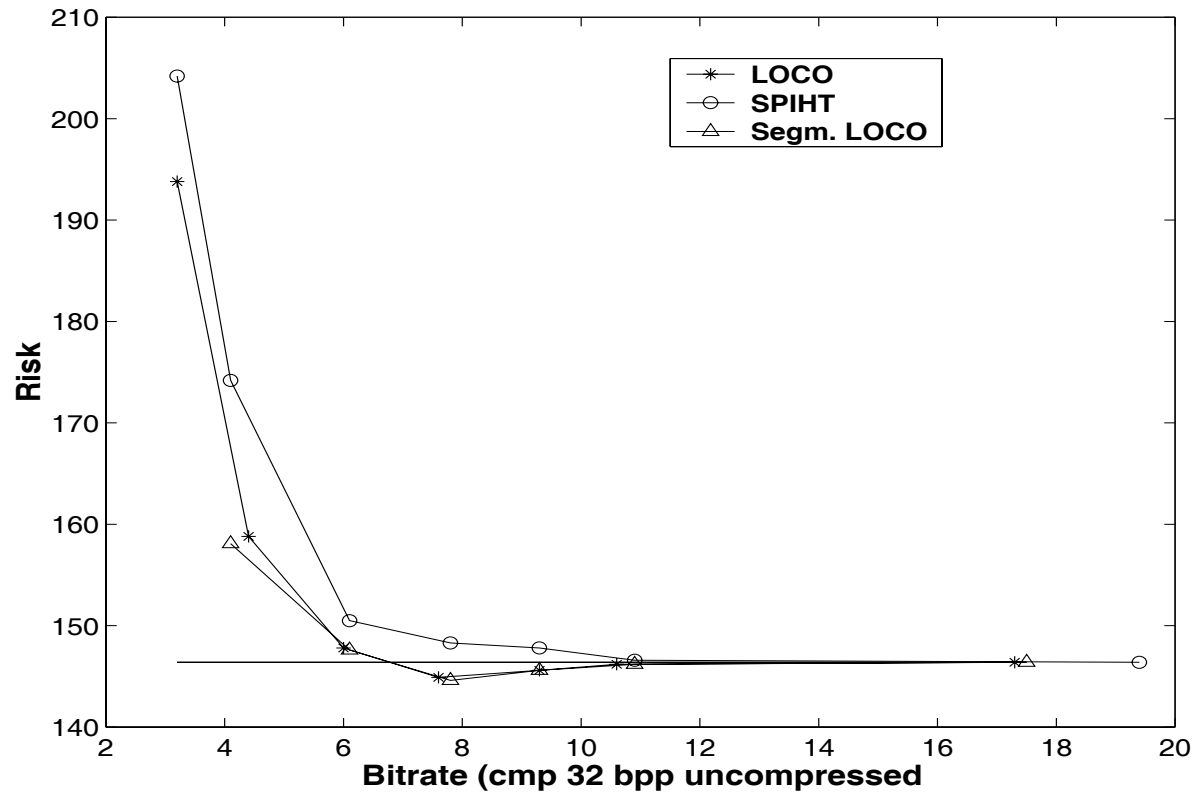
L1-risk as a function of bitrate. Baseline=mean (replicates) expression log-ratios. -Adaptive shape segmentation and background correction via min/max filtering. -A set of arrays with mostly non-differentially expressed genes.

## Lossy Compression Results



A set of noisy arrays. Many differentially expressed genes.

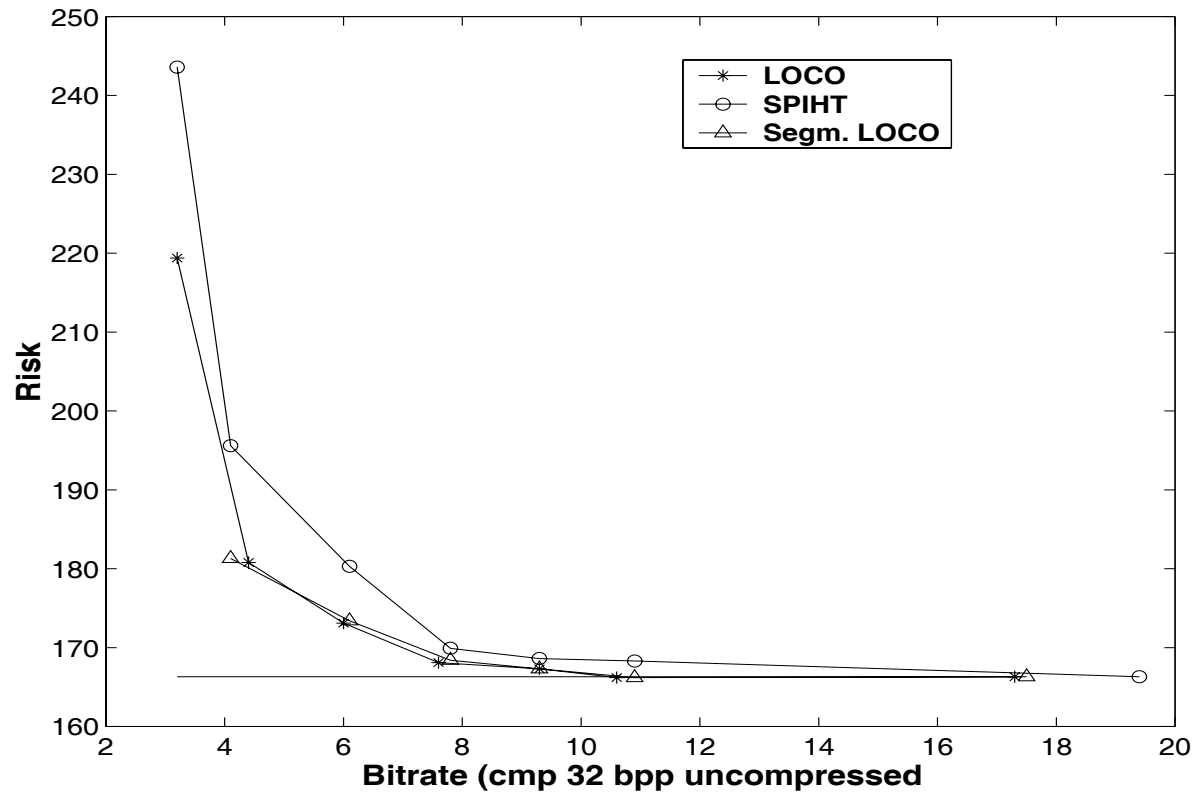
## Lossy Compression Results



Adaptive circle segmentation and background correction via sampling.

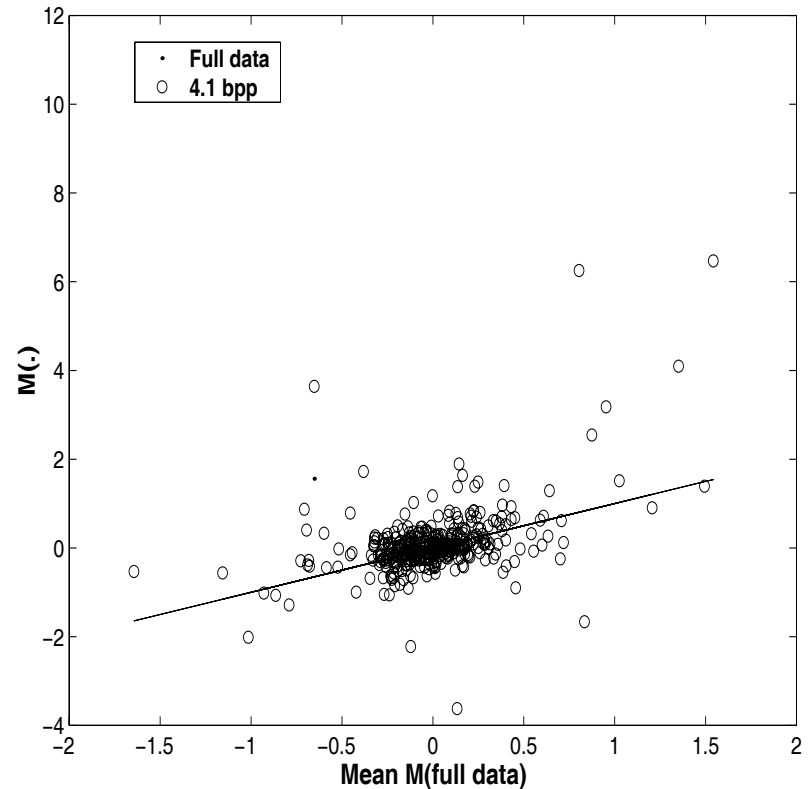
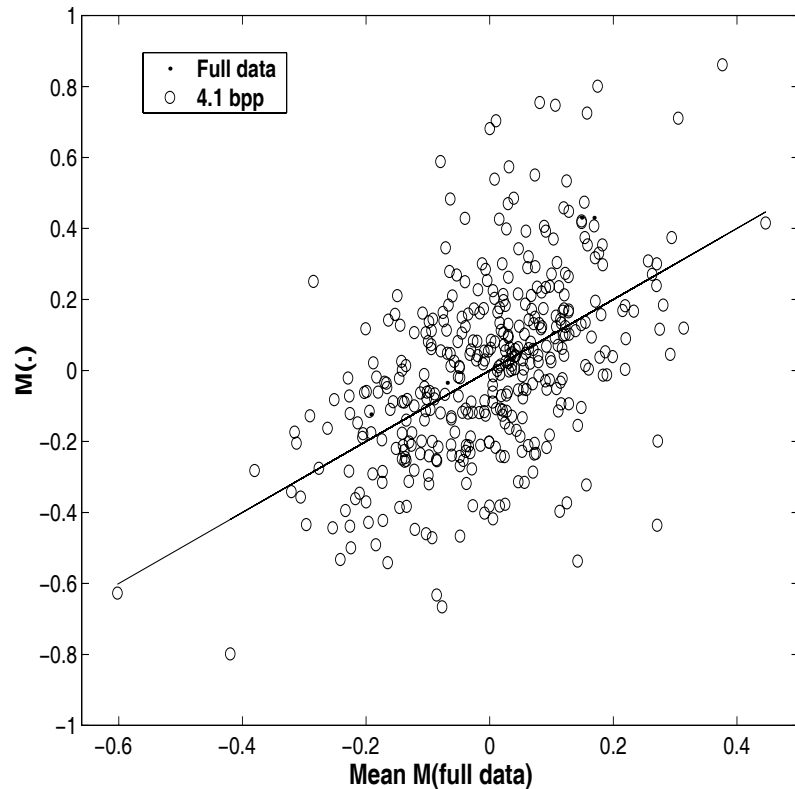
A set of noisy arrays. Many low-intensity spots.

## Lossy Compression Results



A set of arrays with many high-intensity spots.

## Lossy Compression Results



Lossy compression acts as denoising. For larger absolute log-ratios, this results in shrinkage towards the mean (over replicates) log-ratio.

## Conclusions

- Lossless and Progressive Compression of Microarray Images.
- Allows for subset reconstruction, locally varying degree of loss.
- Effect of compression is smaller than array-array variability at  $\sim 4$  bpp.
- Compression  $\simeq$  denoising at rates  $\sim 6-10$  bpp.
- Lossless compression ratio  $\sim 1.8:1$ , lossy compression ratio  $\sim 8:1$ .
- Future Work: Joint coding of images to improve the lossless bitrate, lossy performance.

## Acknowledgments

Dave Nelson, Lawrence Livermore National Labs

Sandrine Dudoit, Biostatistics, UCB

Yee-Hwa Yang, Statistics, UCB

Matthew Callows, Lawrence Berkeley National  
Labs.