Okay, for EPIGseq, what was the major contributions:

1. With RNAseq, RPKM and other normalization was not properly justified
2. Correlation using CYs for count level data
3. Magnitude (fold change) vs. Wilcoxon Rank sum test
4. SNR vs. dispersion

Extract patterns

Calculate pairwise CYs correlations

Tally candidate patterns with >= Mt profile with Rt1

Delete patterns with low Mt == No need

Delete profiles with low R to patterns == No need

Model data to estimate dispersion and perform Wilcoxon Rank sum test ??

Delete pattern profiles with (low or high dispersion) or small magnitude (Wilcoxon rank sum Z-statistic < St1

Delete patterns with overlapping profiles to make them mutually exclusive == done

Categorize genes to patterns

Use extracted patterns as seeds

Until no more moves:

Correlate profiles to patterns using CYs similarity

Update patterns with profile with the highest average Spearman rank of correlation with other pattern profiles

Add profile to pattern with the highest correlation (>= Rt2)

Remove pattern profiles with (low or high dispersion) or small magnitude (Wilcoxon rank sum Z-statistic < St2)

Return final assignment

Remove patterns with similar correlation (>=0.8)

Since we have so much of a challenge handling the large matrix, and we will need to filter the data along the process, why don’t we just simply filter out those “non-informative profiles” in the beginning? This will largely reduce the computational time and requirement.