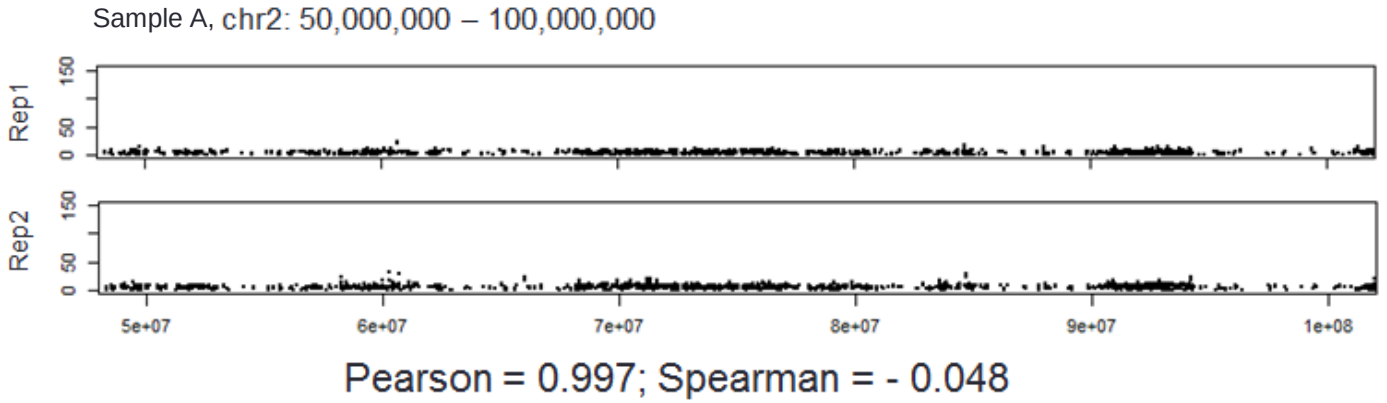
THE HUCK INSTITUTES
OF THE LIFE SCIENCES

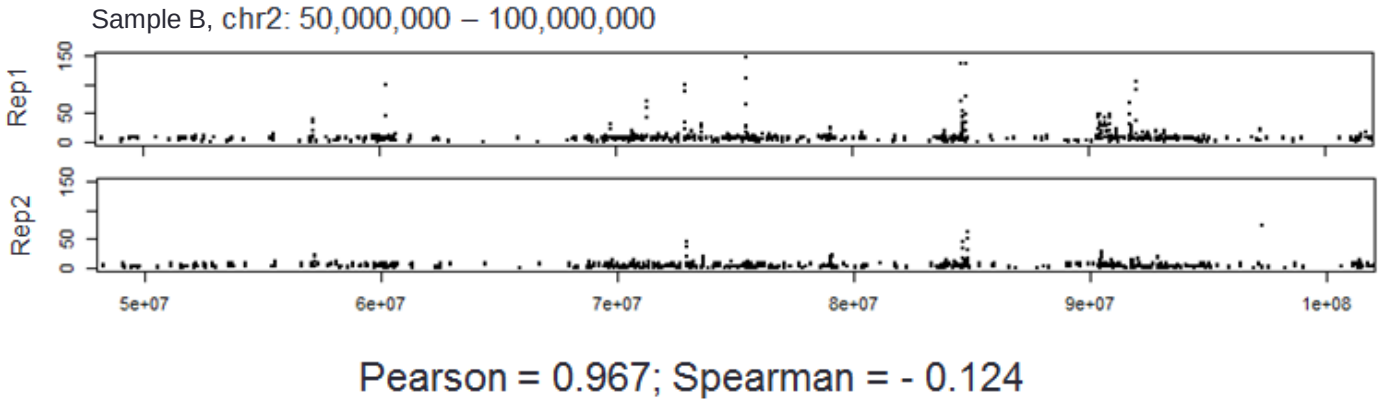
PENNSTATE
1855

Modeling Reproducibility of High Throughput Sequencing Experiment with Tail Dependences

Tao Yang
txy146@psu.edu
Bioinformatics and Genomics
Penn State
9/18/2015

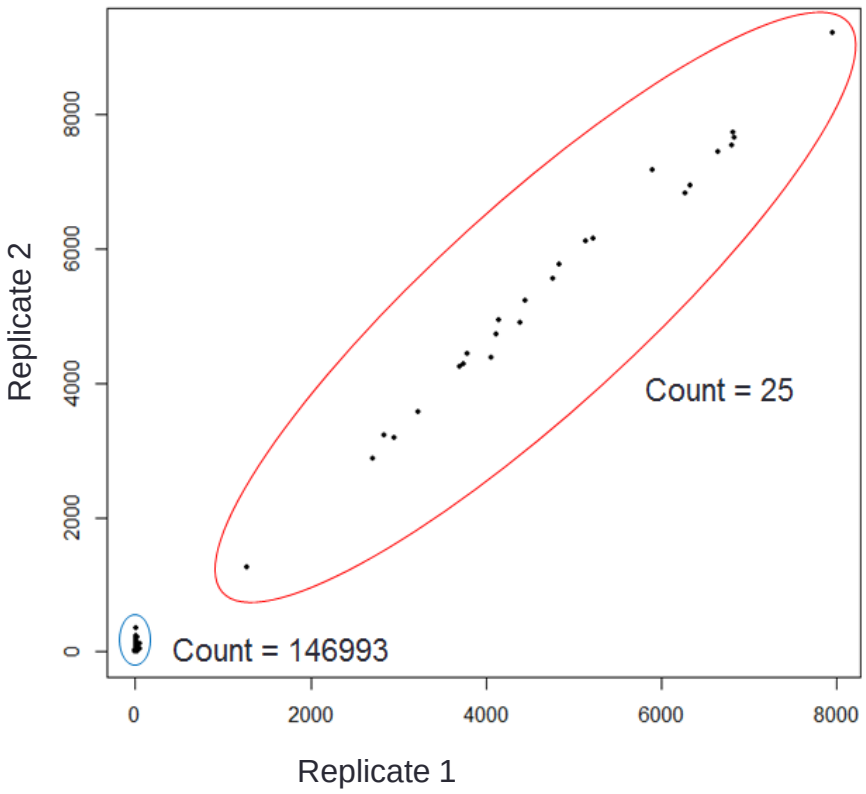
- Pearson and Spearman correlations are misleading



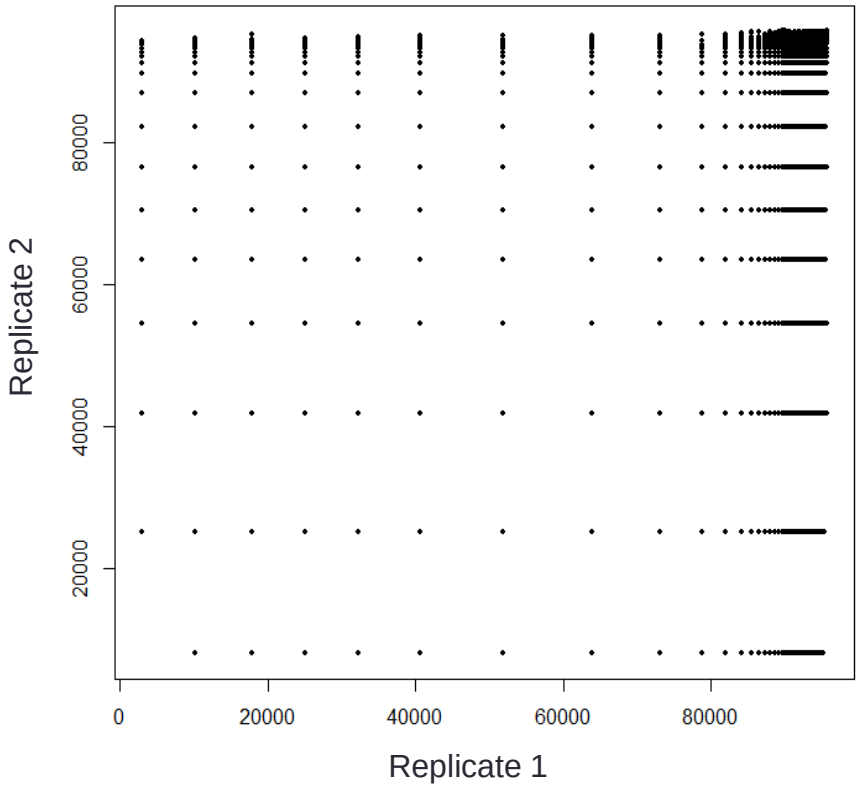


➢ Why Pearson and Spearman correlations fail?

Scatterplot



Rank Scatterplot



Pearson correlation is easily dominated by outliers

(Remove top 25, Pearson = 0.086)

Large amount of rank ties confound the true signal

Method

- Copula
  - Copulas are tools modelling dependence of several random variables
  - Copula models the cumulative density of random variables

Example:

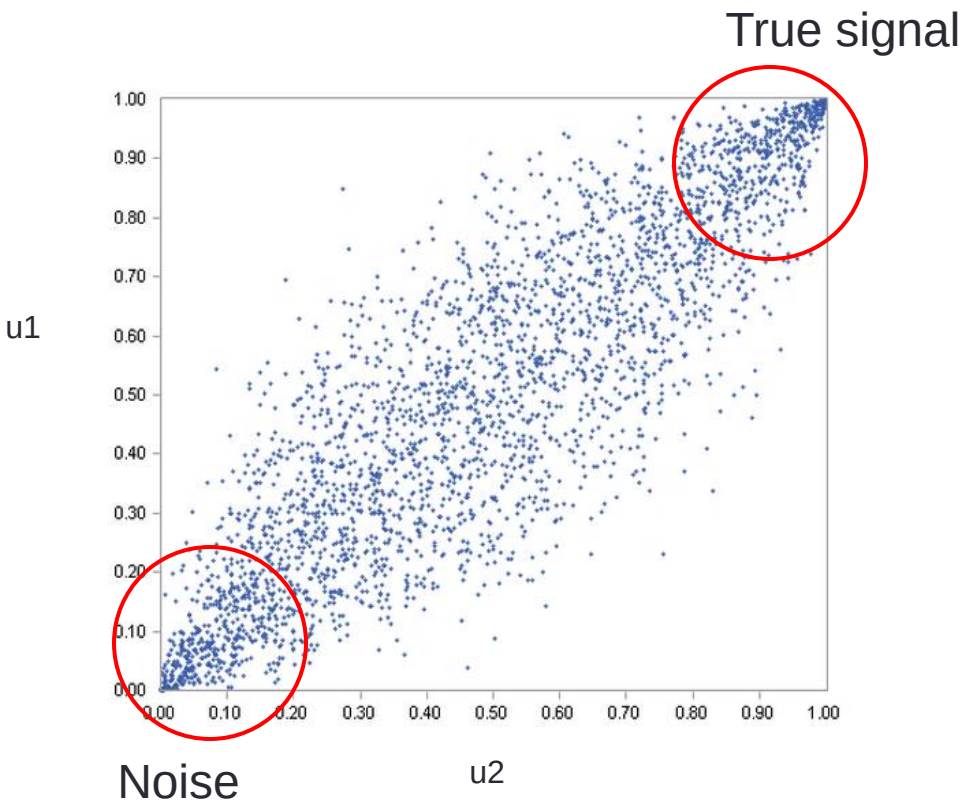
Joint distribution function –

X: number of reads of a 200bp bins

Copula –

U: F(**X**), cumulative density

$$\left(\frac{\#\ of\ reads\ in\ a\ bin}{Total\ \#\ of\ reads}\right)$$



Method

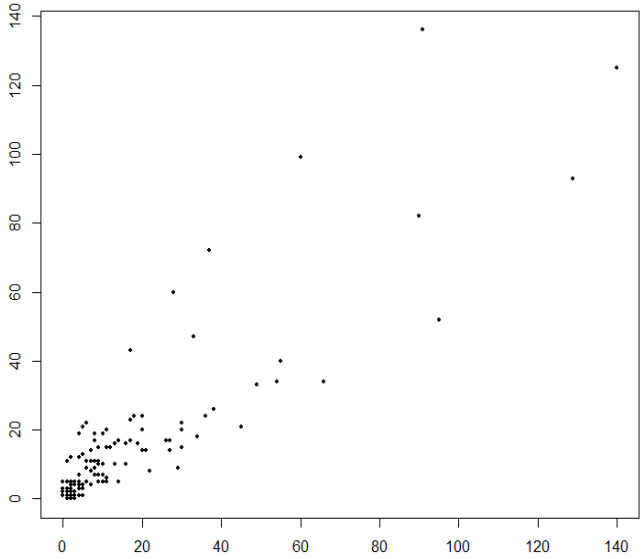
- Gumbel-Clayton mixture model

$$C^{CG} = \pi C_{\beta}^{C}(u,v) + (1-\pi)C_{\theta}^{G}(u,v) \qquad \pi \epsilon [0,1]$$

Upper tail (λ_U)	$(1-\pi)(2-2^{1/\theta})$
Lower tail (λ_L)	$\pi 2^{-1/\beta}$

Result

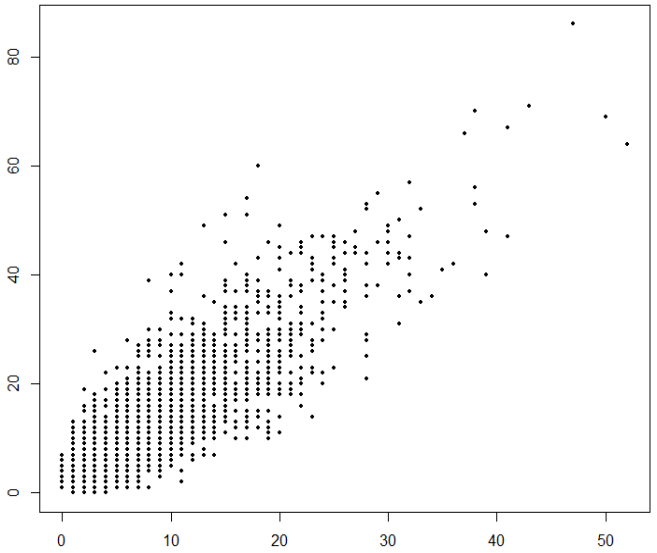
- Our measure tells true difference for TF binding sites data



SydhTfbs_Hepg2_Srebp2

Pearson	0.881
Spearman	0.834

$\lambda_u = 0.326$

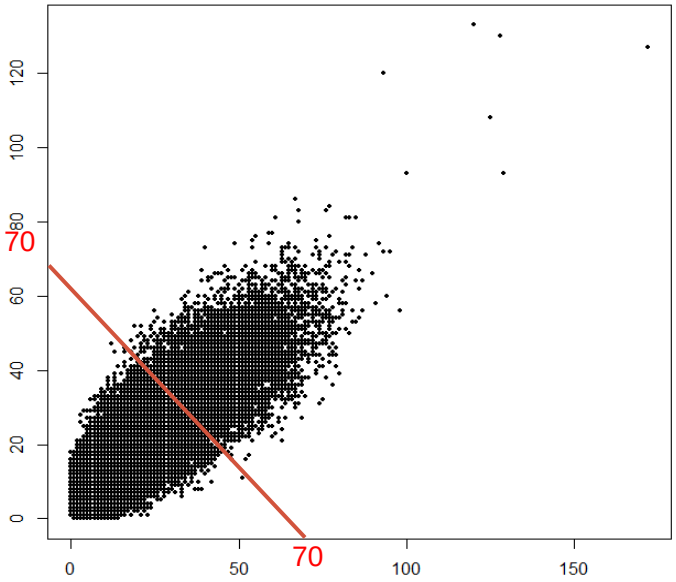


SydhTfbs_Gm12878_Jund

Pearson	0.865
Spearman	0.842

$\lambda_u = 0.611$

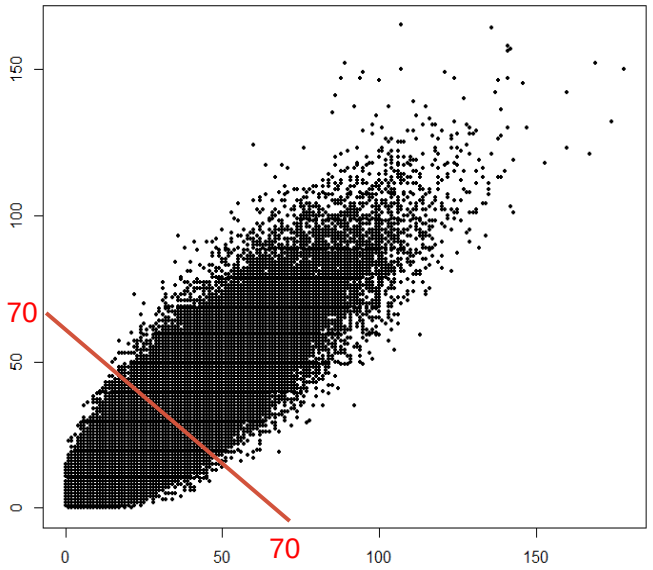
- Our measure better characterizes the histone modification data



BroadHistone_H1hesc_H3k4me3

Pearson	0.885
Spearman	0.824

$\lambda_u = 0.396$

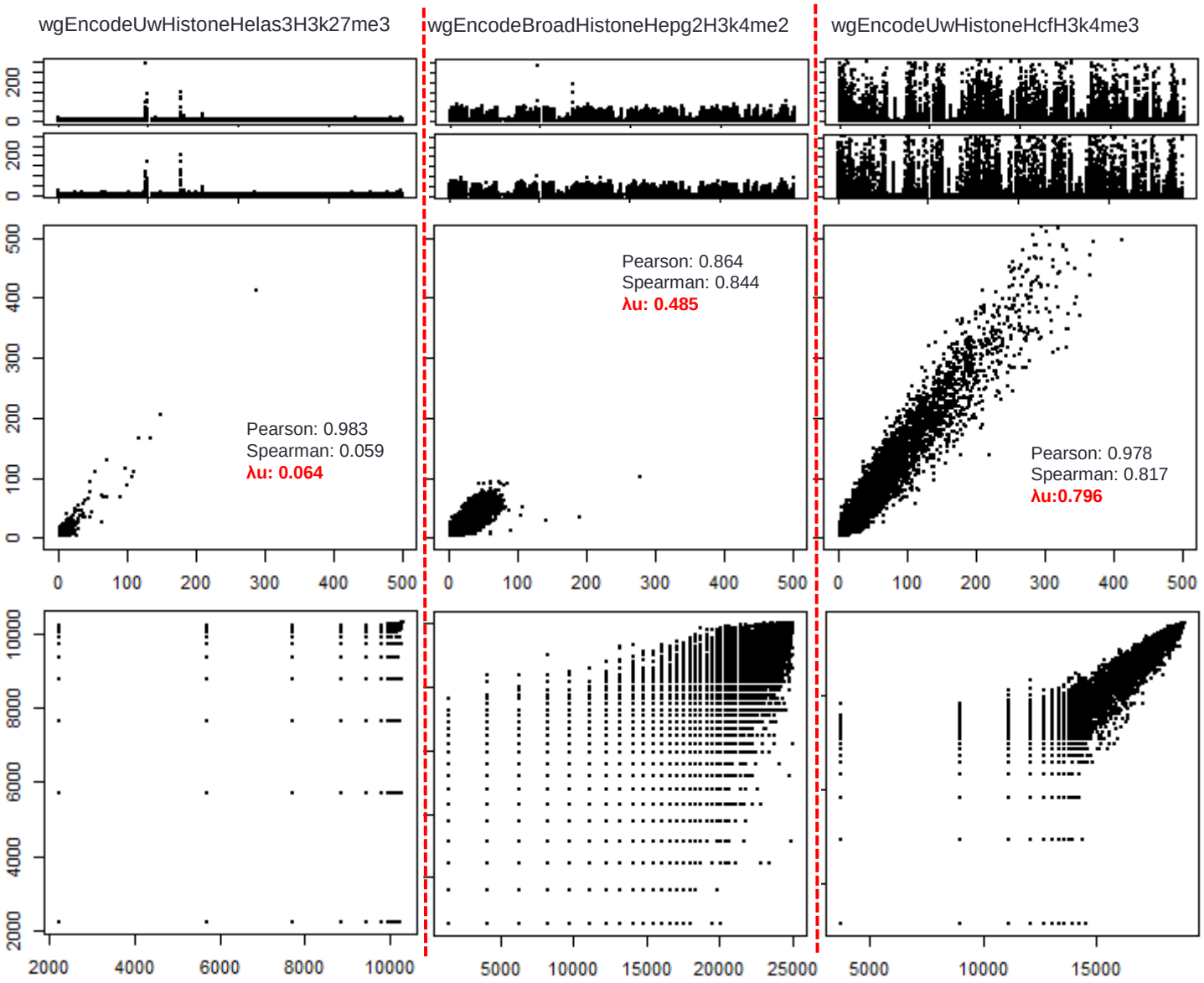


BroadHistone_Hepg2_H3k79me2

Pearson	0.883
Spearman	0.817

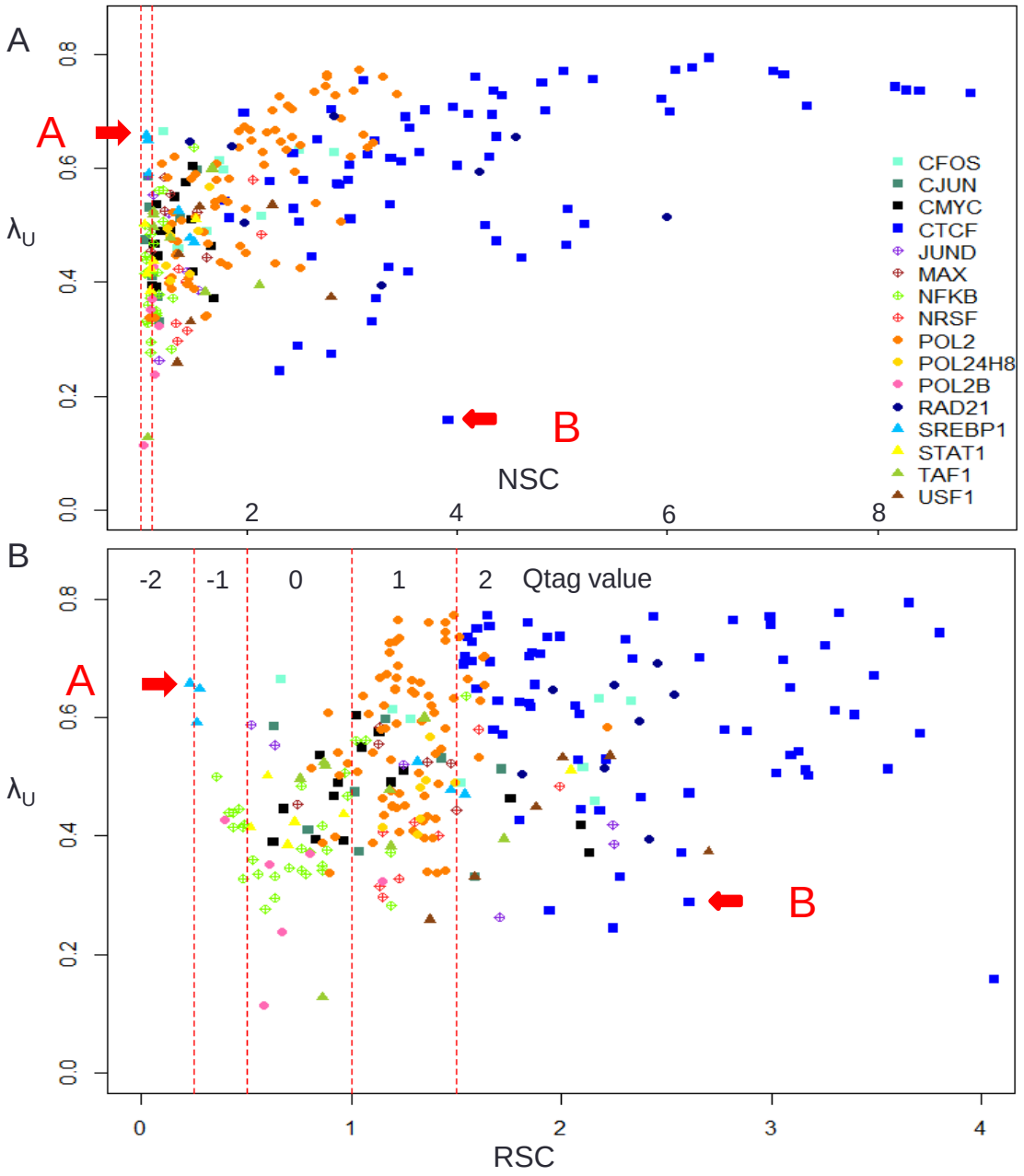
$\lambda_u = 0.700$

- Our measure better characterizes the histone modification data



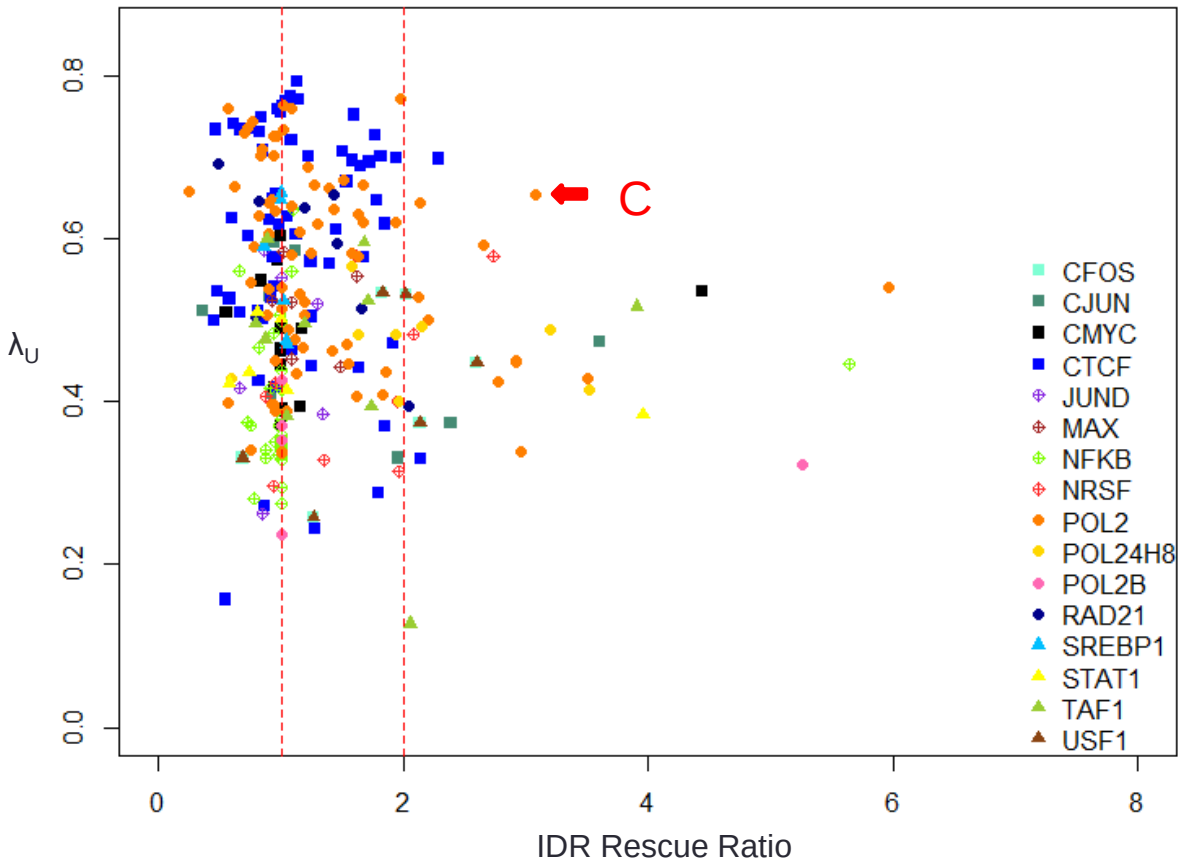
- Comparison with NSC score and RSC score using ENCODE TF data

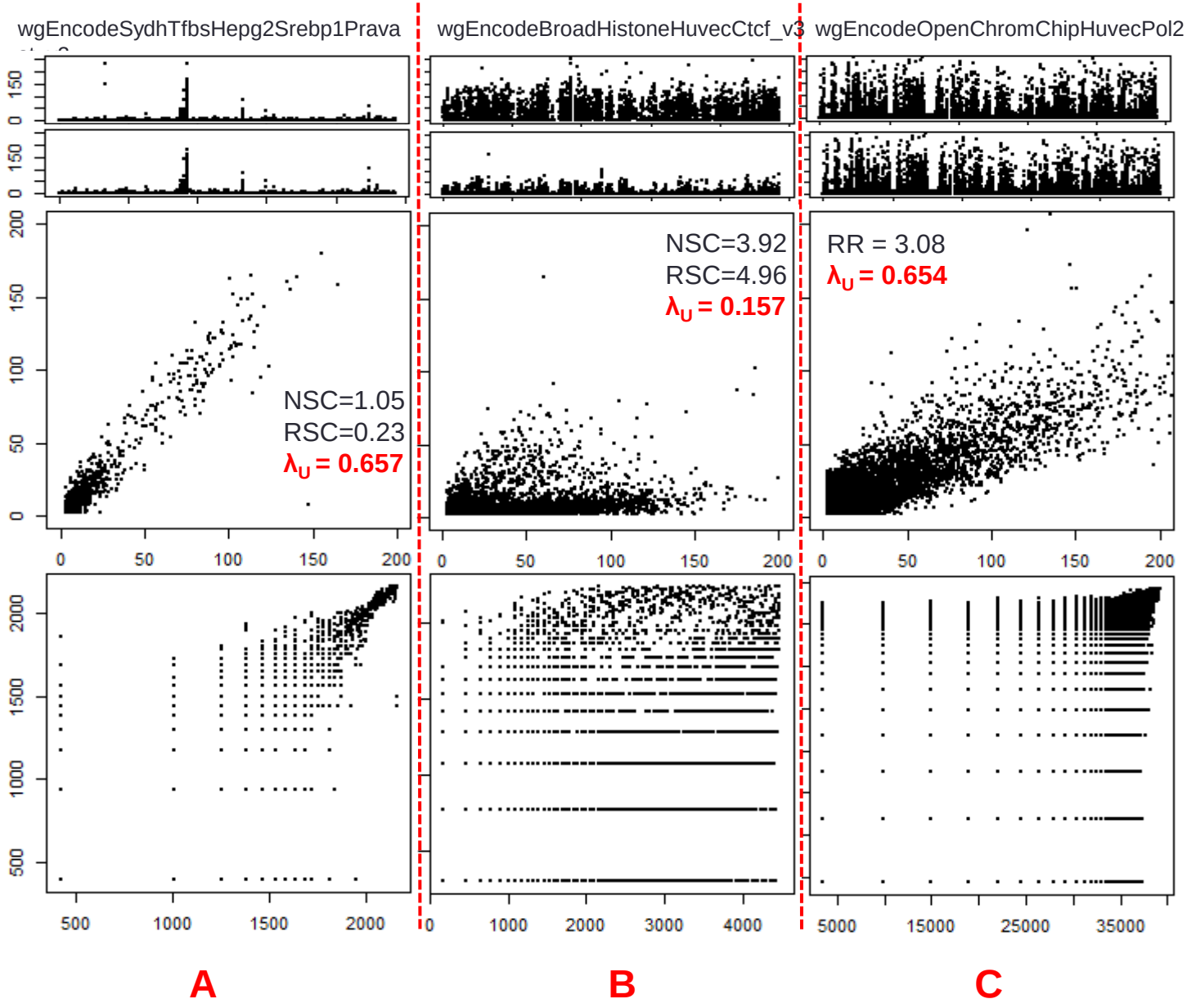
- NSC score less than 1.1 is considered as poorly-enriched with peaks
- RSC score is categorized into 5 Qtag values, the quality increases from -2 to 2
- NSC and RSC are measures for a single replicate, to make it comparable to our measure, we take the average score of the two replicates



➢ Comparison with IDR Rescue Ratio Strategy

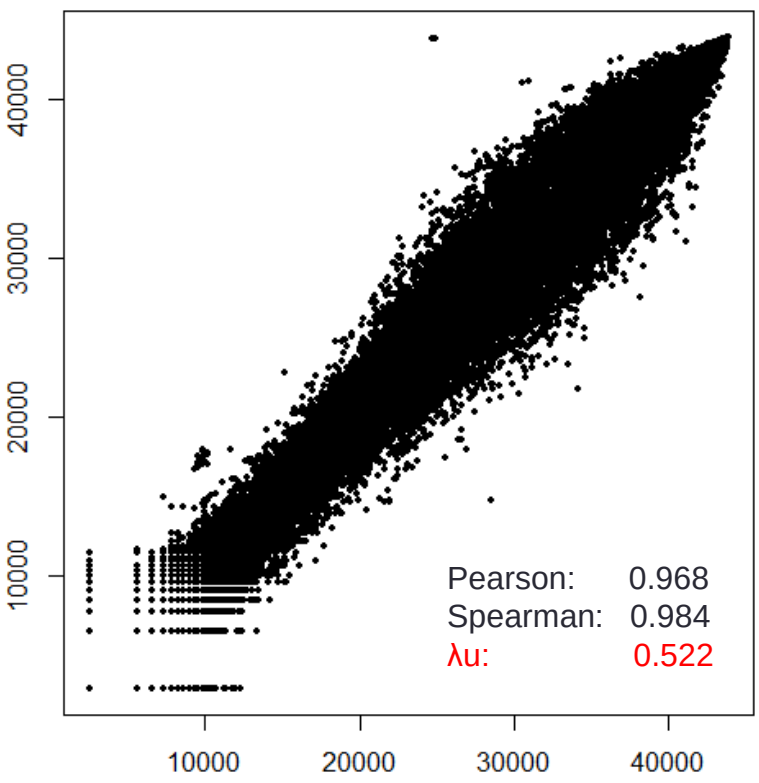
- IDR Rescue Ratio Strategy first generates a pair of pseudo replicates, then uses IDR to call reproducible peaks for both pseudo replicates and biological replicates. The dataset with a ratio greater than 2 will be considered irreproducible



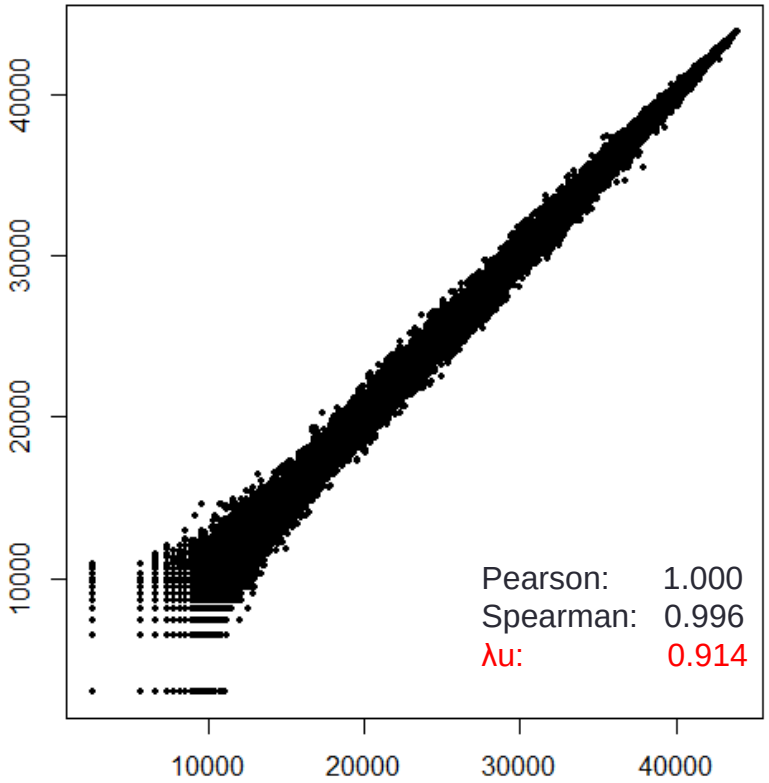

wgEncodeSydhTfbsHepg2Srebp1Prava
wgEncodeBroadHistoneHuvecCtcf_v3
wgEncodeOpenChromChipHuvecPol2
NSC=1.05
RSC=0.23
λU = 0.657
NSC=3.92
RSC=4.96
λU = 0.157
RR = 3.08
λU = 0.654
A
B
C

- Our measure performs well on RNA-seq data

Rank scatterplot



LIV

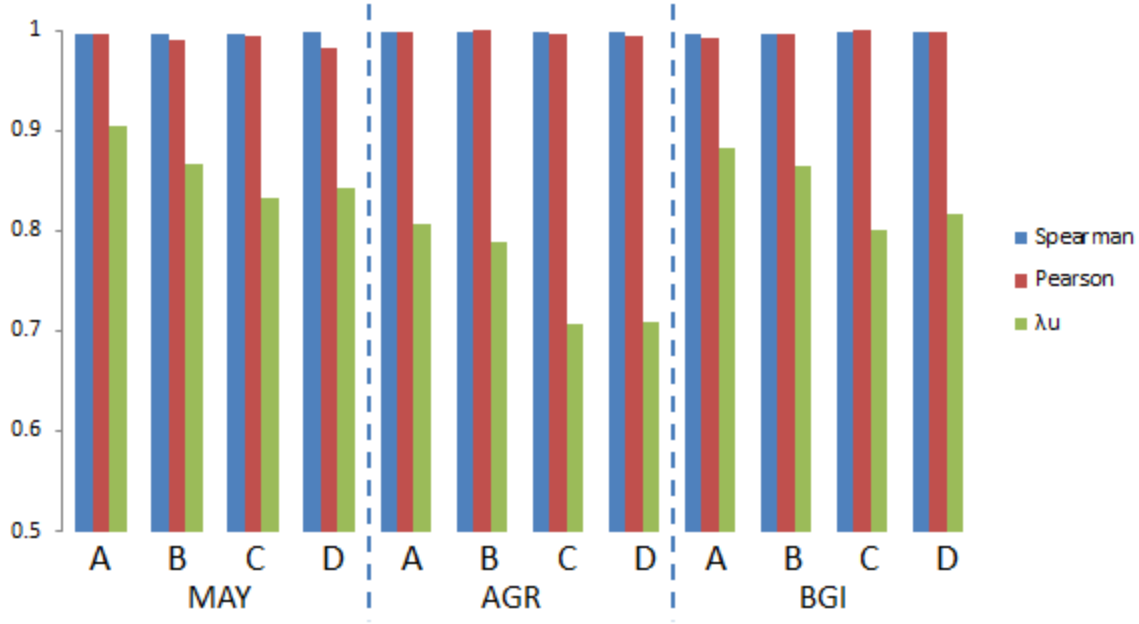


PSU

Pairs of Human universal reference RNA samples sequenced by two institutes

SEQC/MAQC-III consortium. Nature Biotechnology 32, 903–914 (2014)

- Our measure captures the transcriptome complexity



A, B, C, D are four samples used by the SEQC consortium for evaluating the RNA-seq data quality

A: Universal Human Reference RNA
B: Human Brain Reference RNA
C: C = ¾ A + ¼ B
D: D = ¼ A + ¾ B

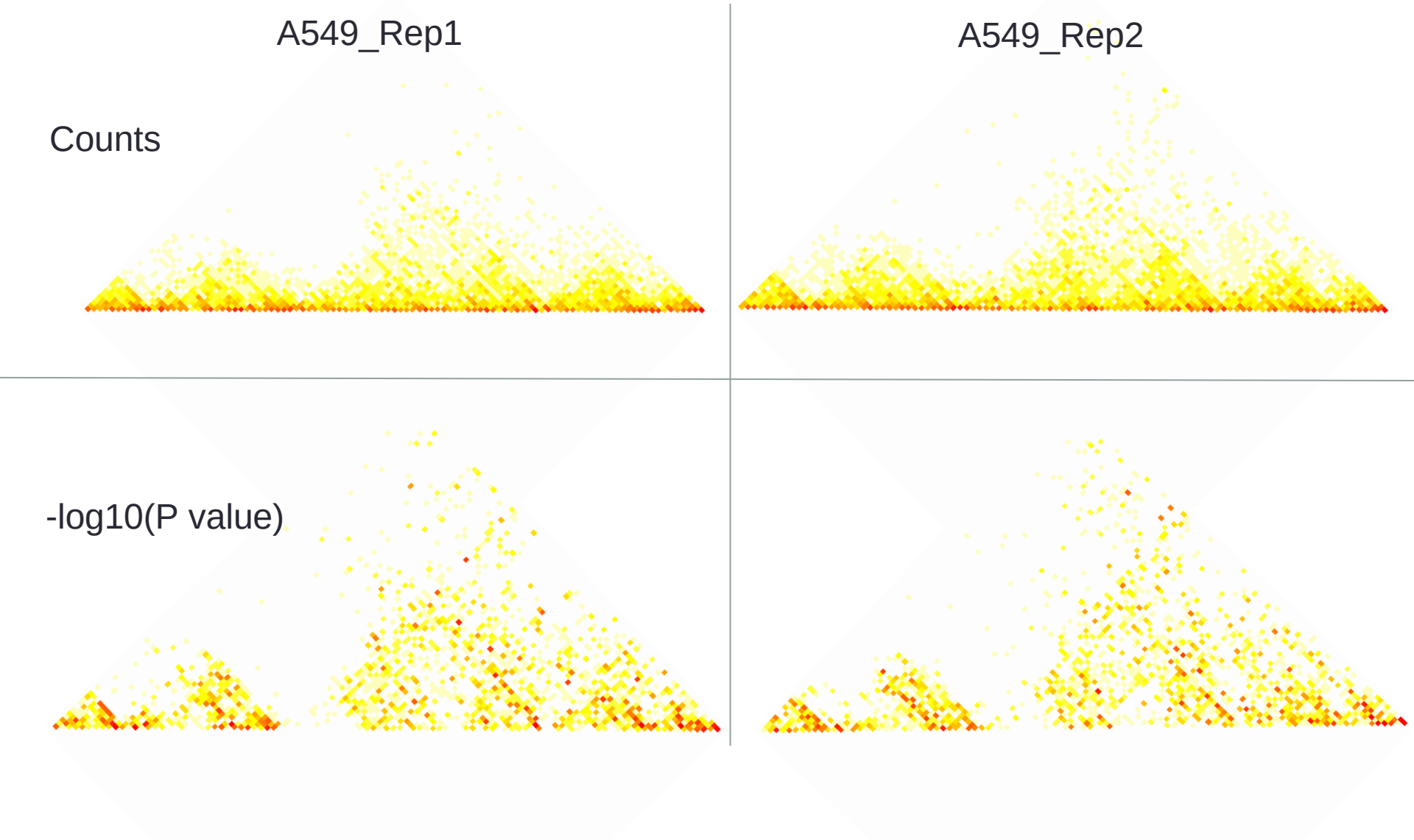
SEQC/MAQC-III consortium. Nature Biotechnology 32, 903–914 (2014)

Conclusions

- Pearson and Spearman correlations could be misleading when measuring the reproducibility of high-throughput data
- Our method models the correlation in noise and signal separately, and is not sensitive to outliers and ties
- Our method is suitable for different kinds of sequencing data (ChIP-seq, RNA-seq)
- R package will be soon released

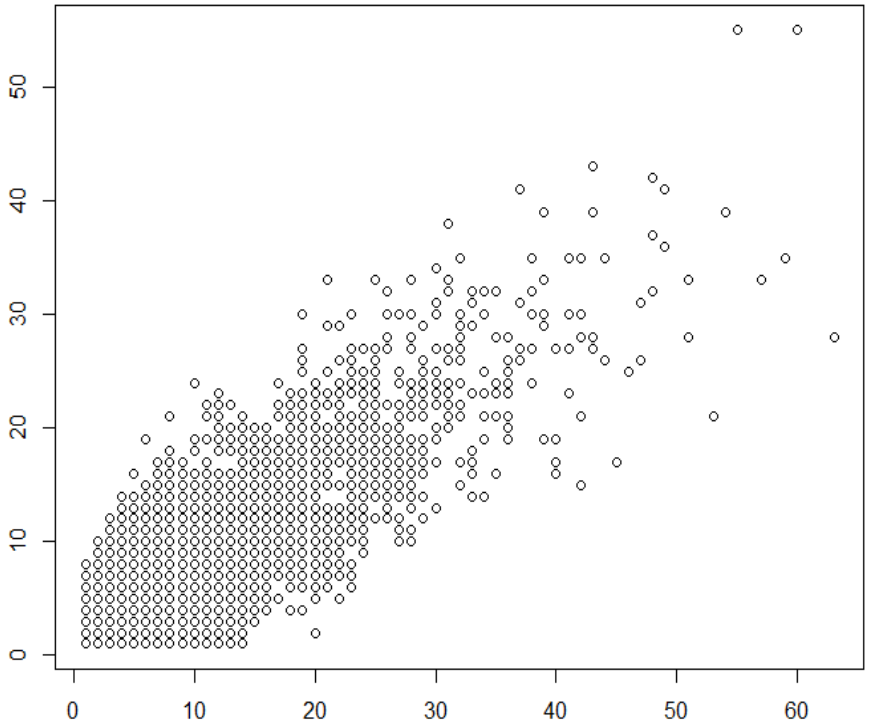
Hi-C reproducibility

Data: Output of Fit-hi-c from Noble lab, only show non-zero counts on Chr22



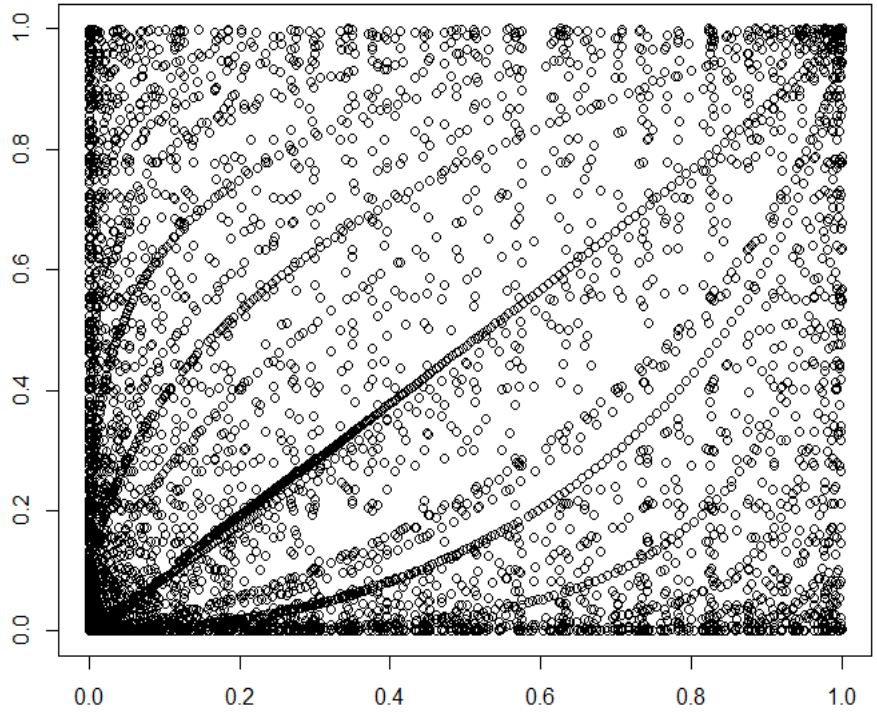
- Without adjusting for distance, counts look more reproducible than p-values

Scatterplot of counts



Pearson = 0.869

Scatterplot of P values

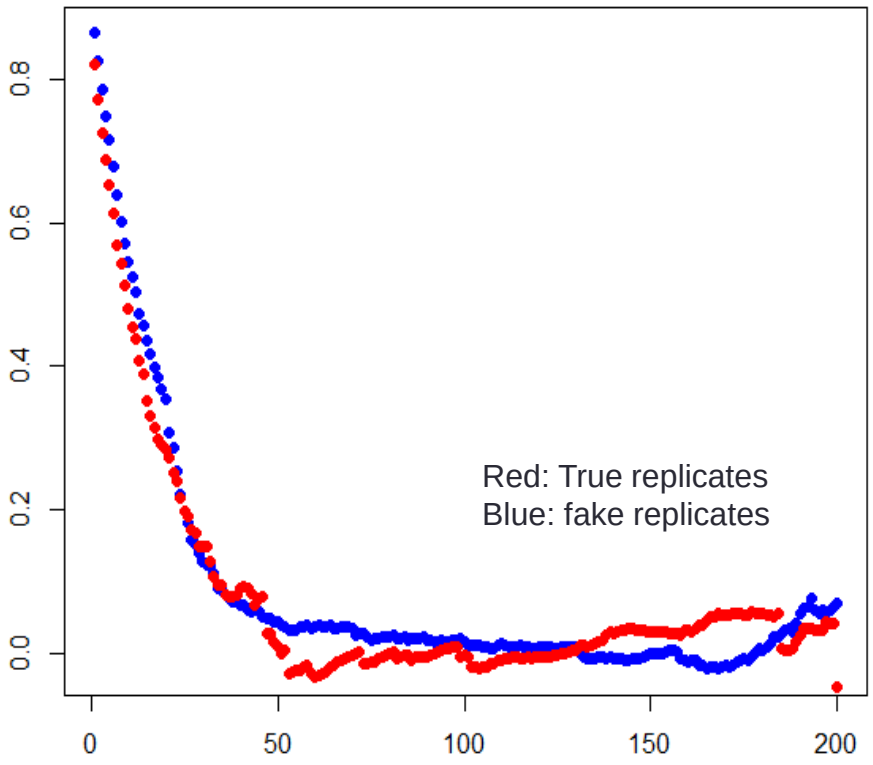


Pearson = 0.416

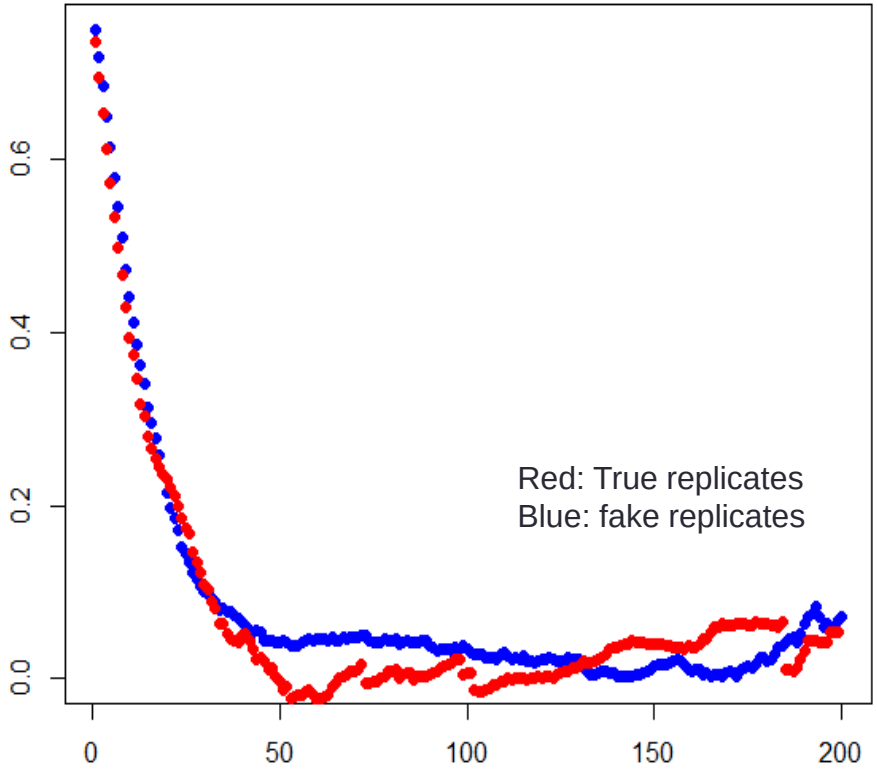
Stratify contacts according to the distance between the two coordinates

- Calculate the Pearson and Spearman correlations based on counts
- Compare true replicates (A549-A549) and fake replicates (A549 – G401)

Pearson vs distance



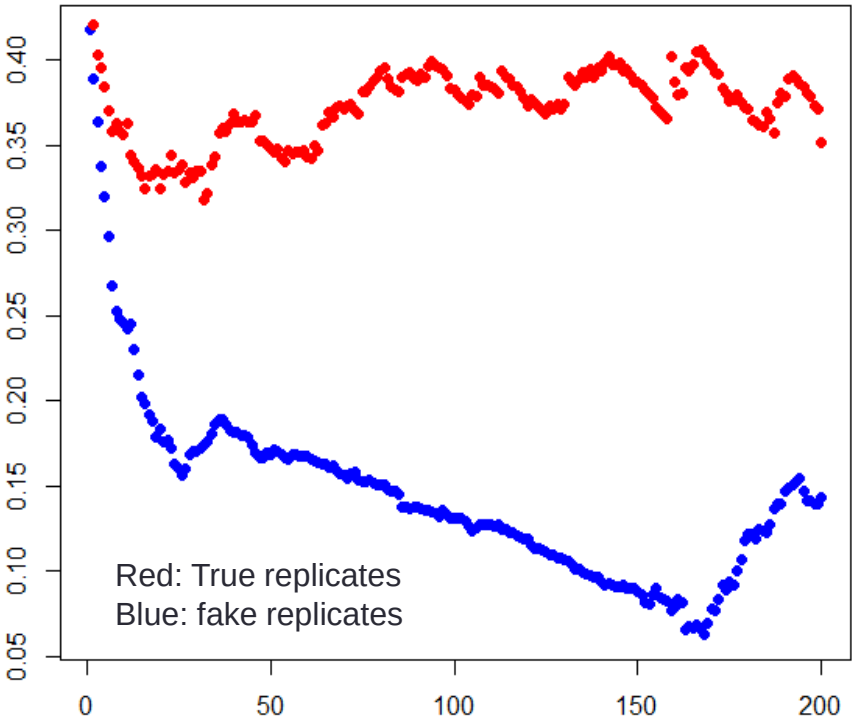
Spearman vs distance



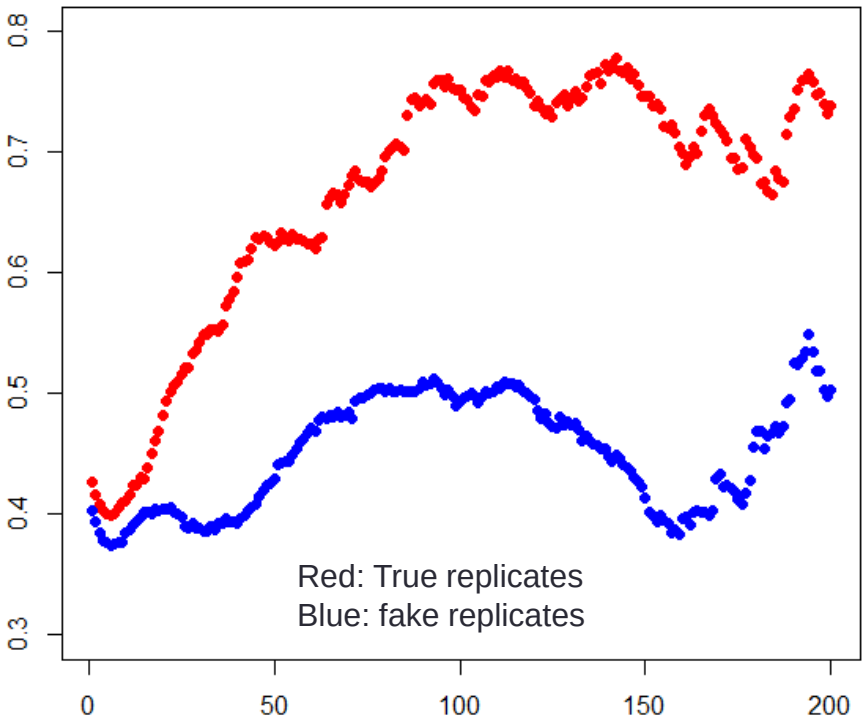
Distance between coordinates (X 80kb)

- Same setting as the previous slide
- Calculate the Pearson and Spearman correlation based on –log10(P values)

Pearson vs distance



Spearman vs distance



Distance between coordinates (X 80kb)

Conclusion

- Reproducibility in Hi-C data depends on the distance between coordinates
- Some preprocessing may be necessary before assessing reproducibility, especially for counts
- The correction of random looping effect in p-value helps establish true reproducibility

Acknowledgement

- Dr. Qunhua Li
- Dr. Feng Yue
- Dr. Ross Hardison’s Lab