

# **Visualization and interpretation of high-throughput genomics data**

Chongzhi Zang  
Associate Professor  
[zang@virginia.edu](mailto:zang@virginia.edu)  
[zanglab.org](http://zanglab.org)

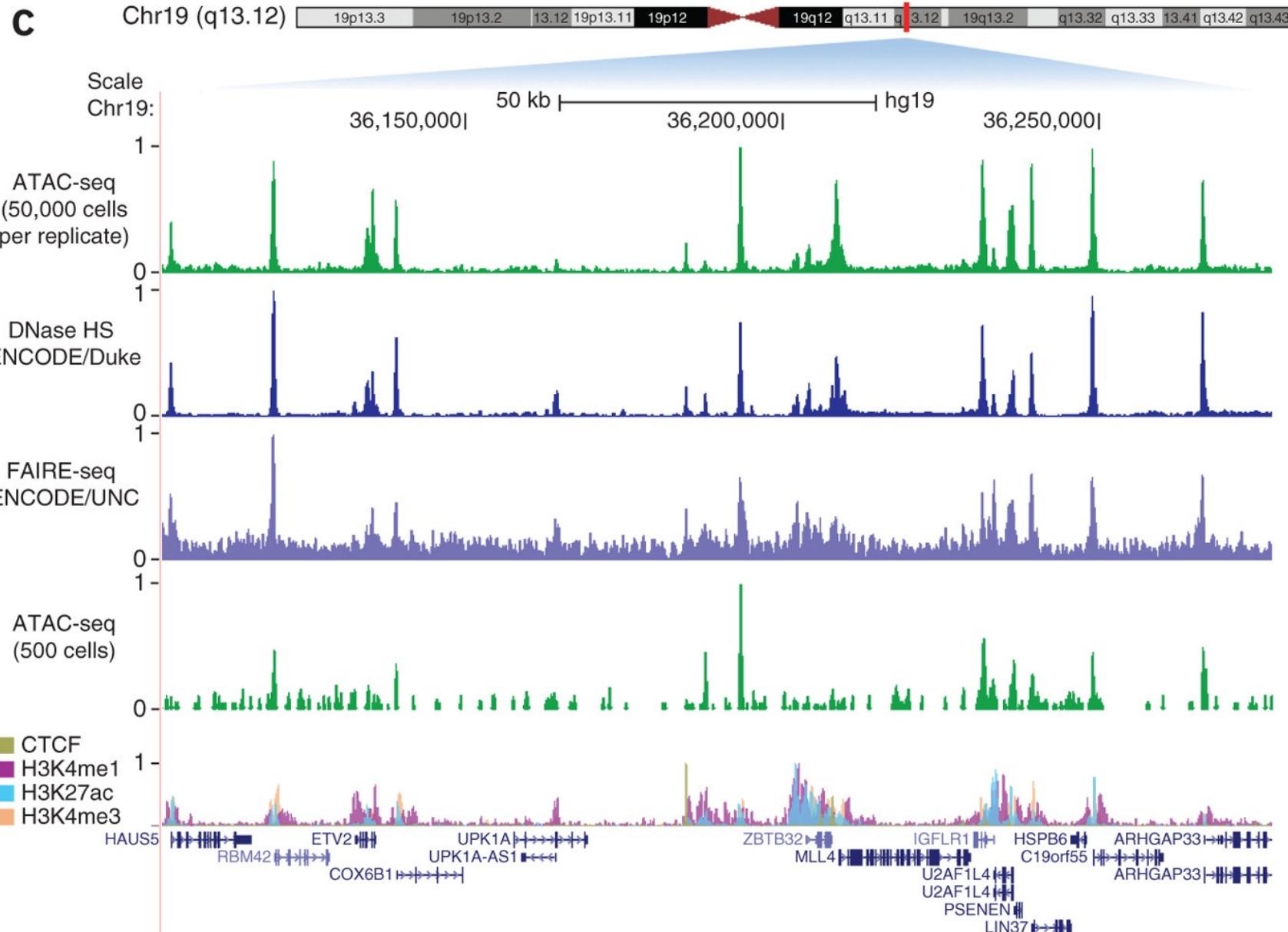
August 23, 2022

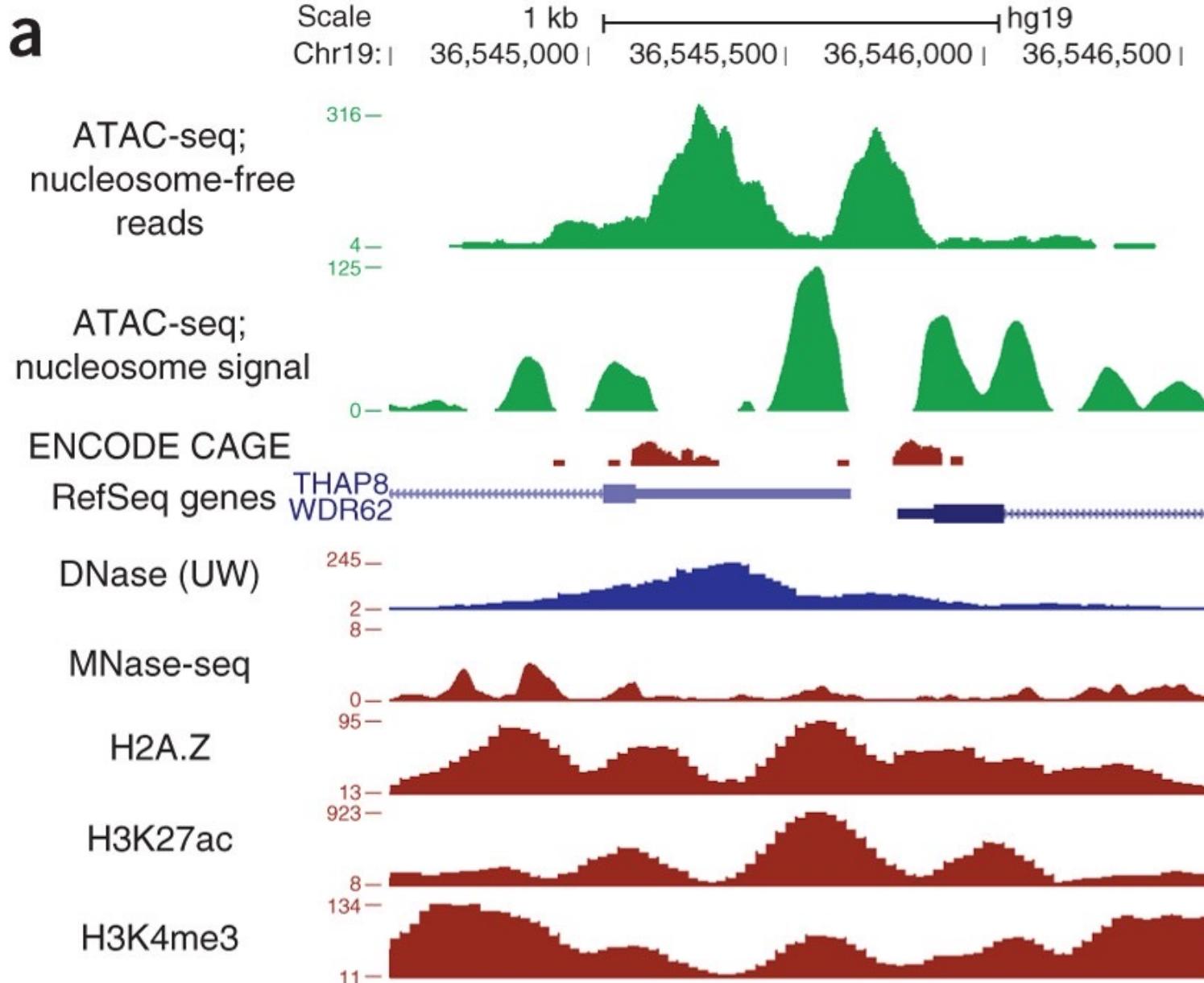
LET'S SOLVE THIS PROBLEM BY  
USING THE BIG DATA NONE  
OF US HAVE THE SLIGHTEST  
IDEA WHAT TO DO WITH



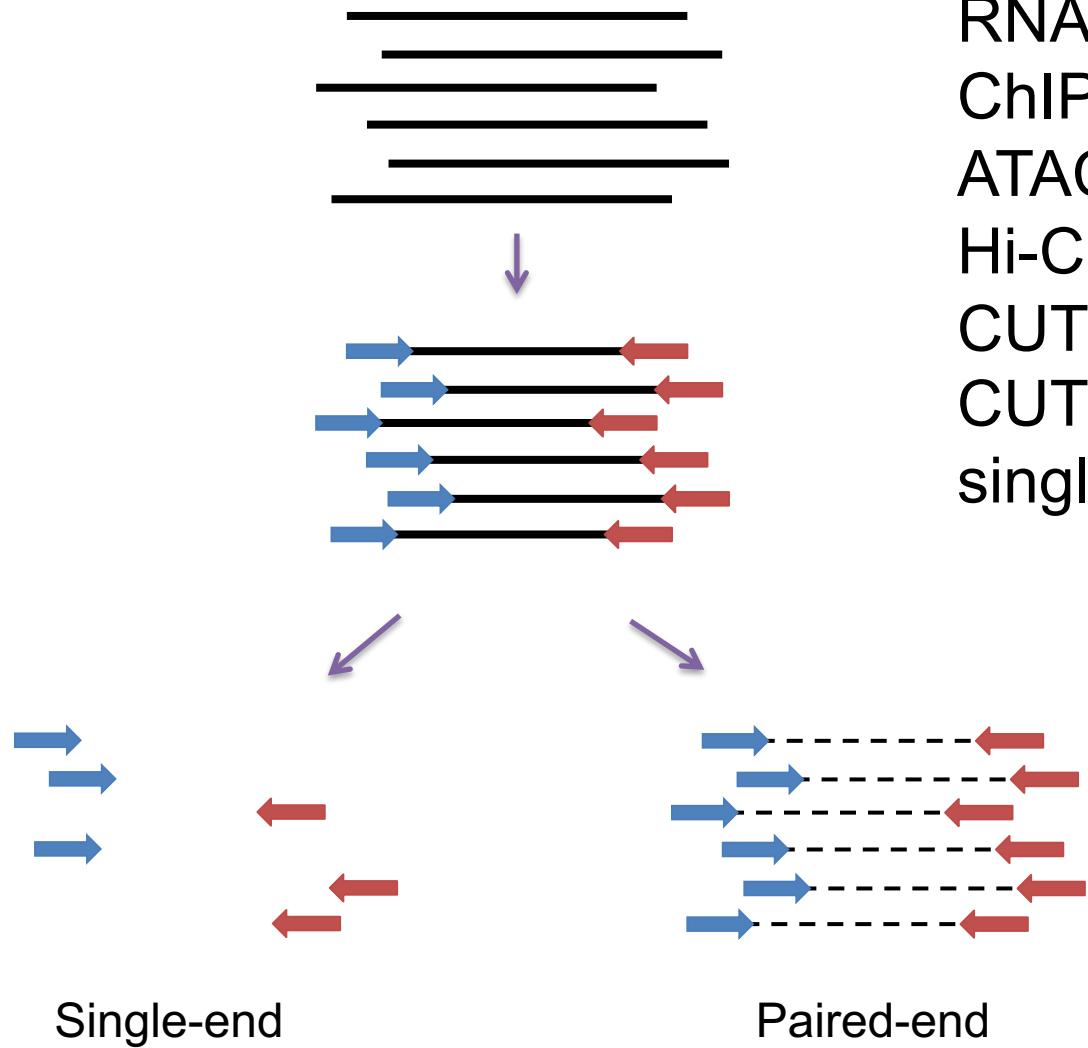
# Learning Objectives

- Be able to read common plots for presenting genomics data
- Understand essential elements in genomics data visualization
- Some tips for data presentation





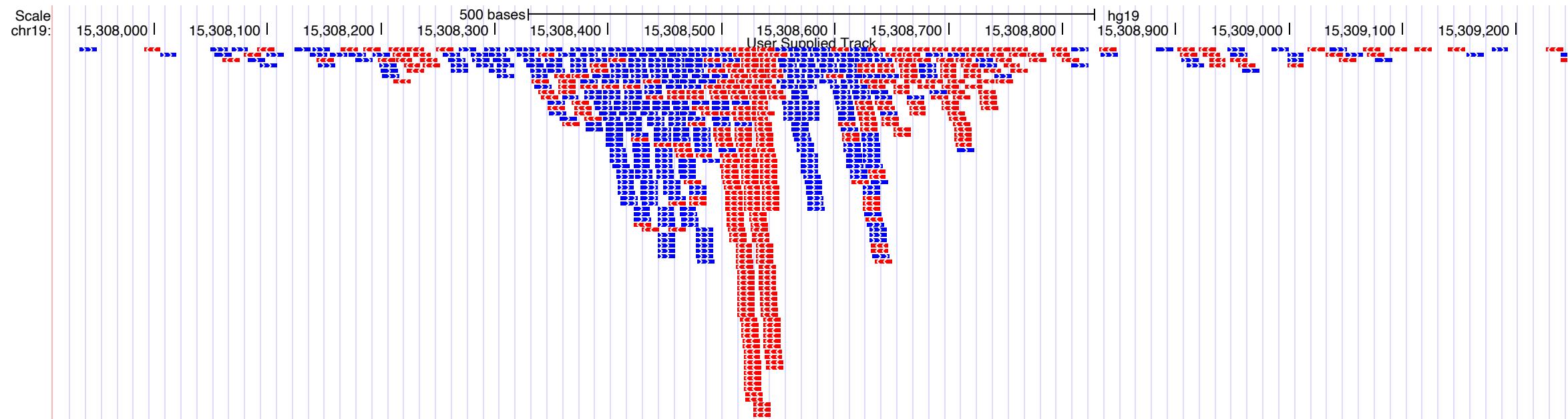
# High-throughput short-read sequencing (Illumina)



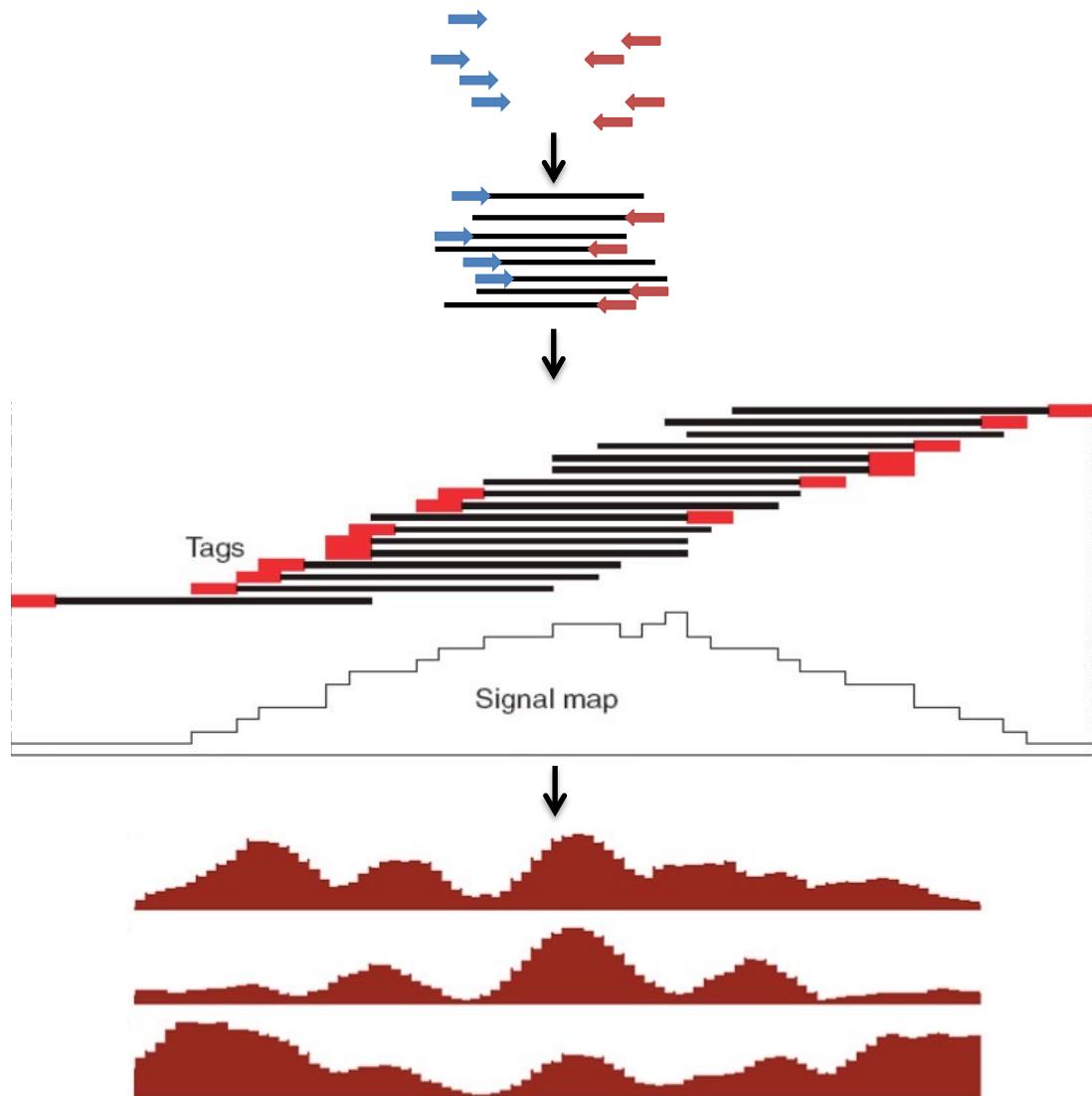
RNA-seq  
ChIP-seq  
ATAC-seq  
Hi-C  
CUT&RUN  
CUT&Tag  
single-cell...



# Original sequence reads are not easy to visualize



# Signal tracks are sequence reads piled up



- bedGraph:

chr4	10344200	10344250	5
chr4	10344250	10344300	10
chr4	10344300	10344350	25
chr4	10344350	10344400	15
chr4	10344400	10344450	8

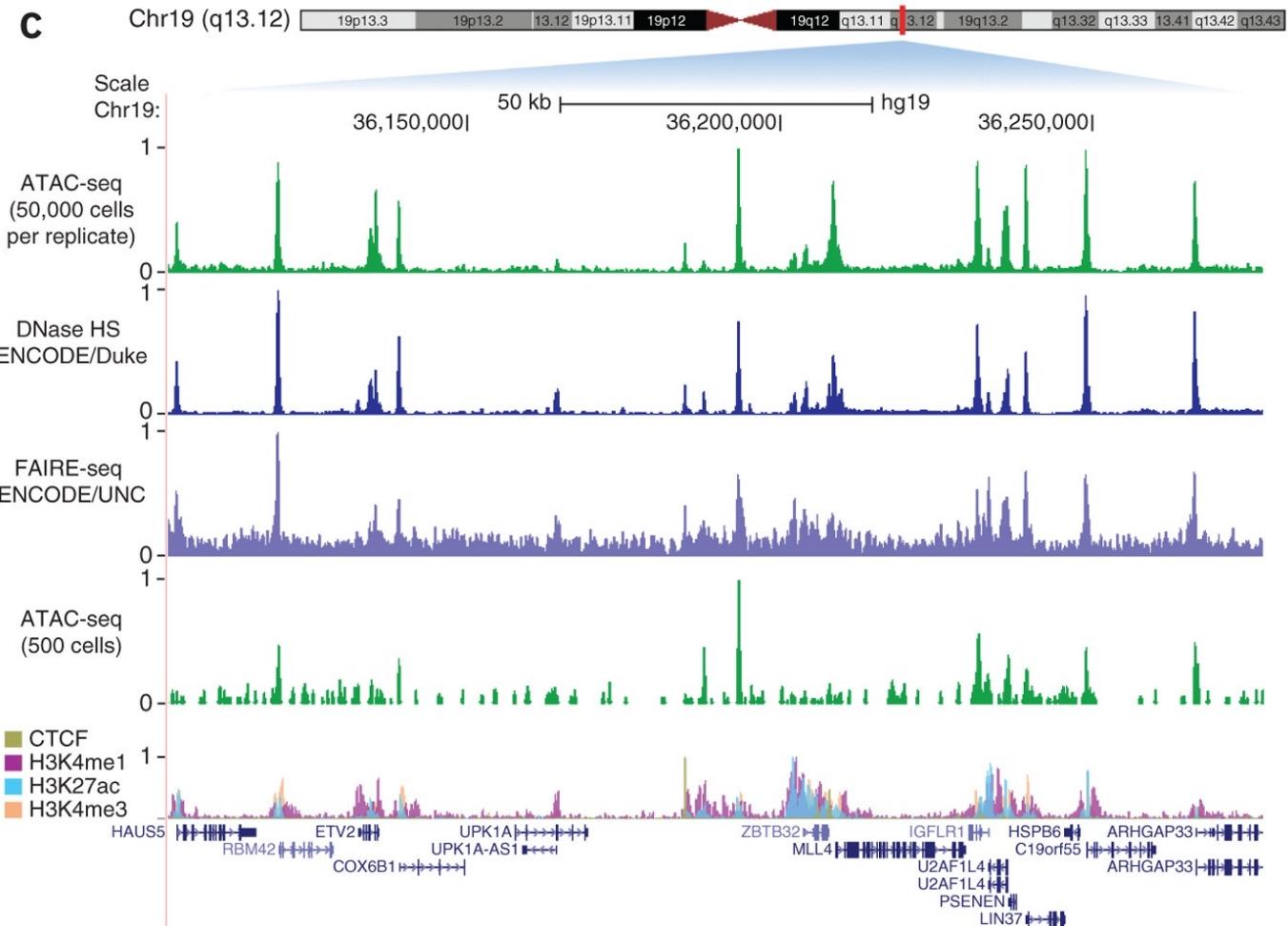
- wiggle:

```
track type=wiggle_0
variableStep chrom=chr4 span=50
10344200 5
10344250 10
10344300 25
10344350 15
10344400 8
```

- bigWig: indexed binary format

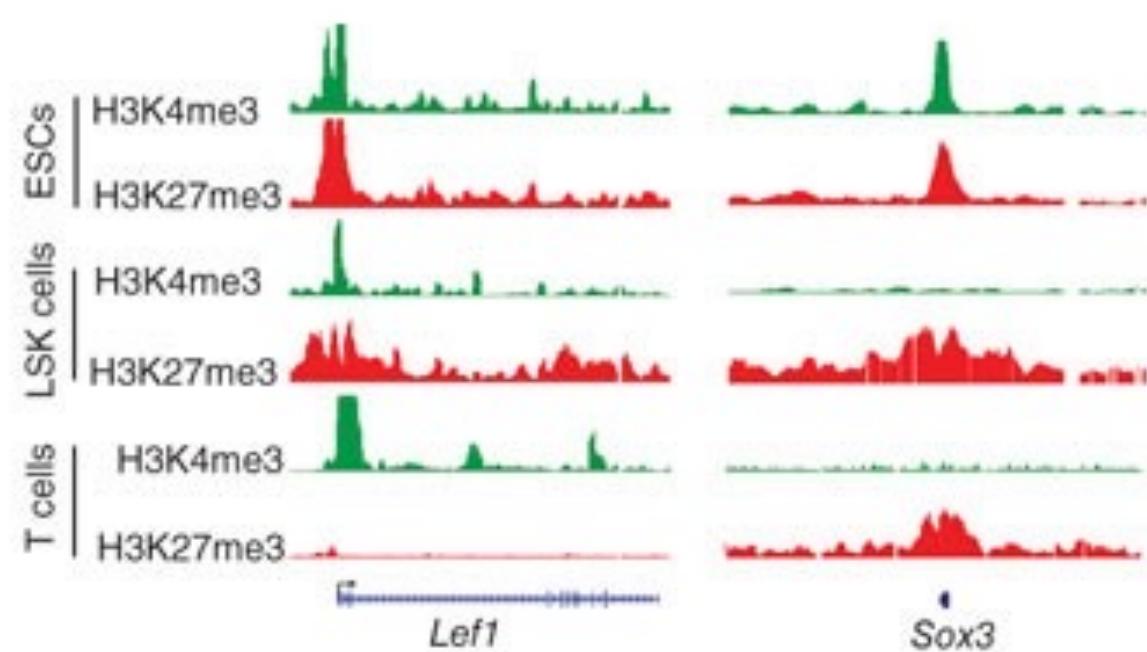
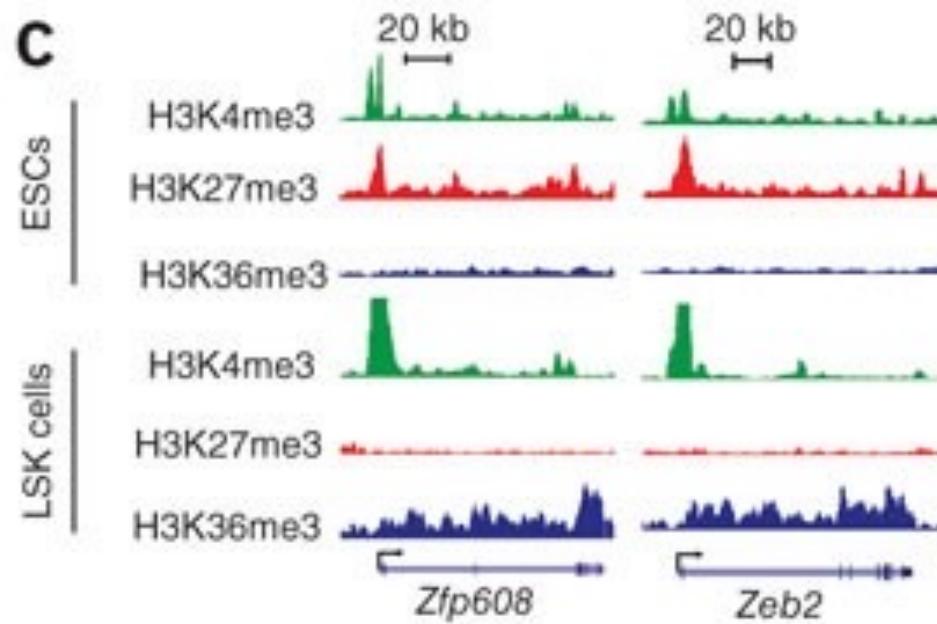
# Essential elements in a ripple heatmap

- Chromosomal locations
- Track label
- Track scale
  - x: resolution?
  - y: normalization?

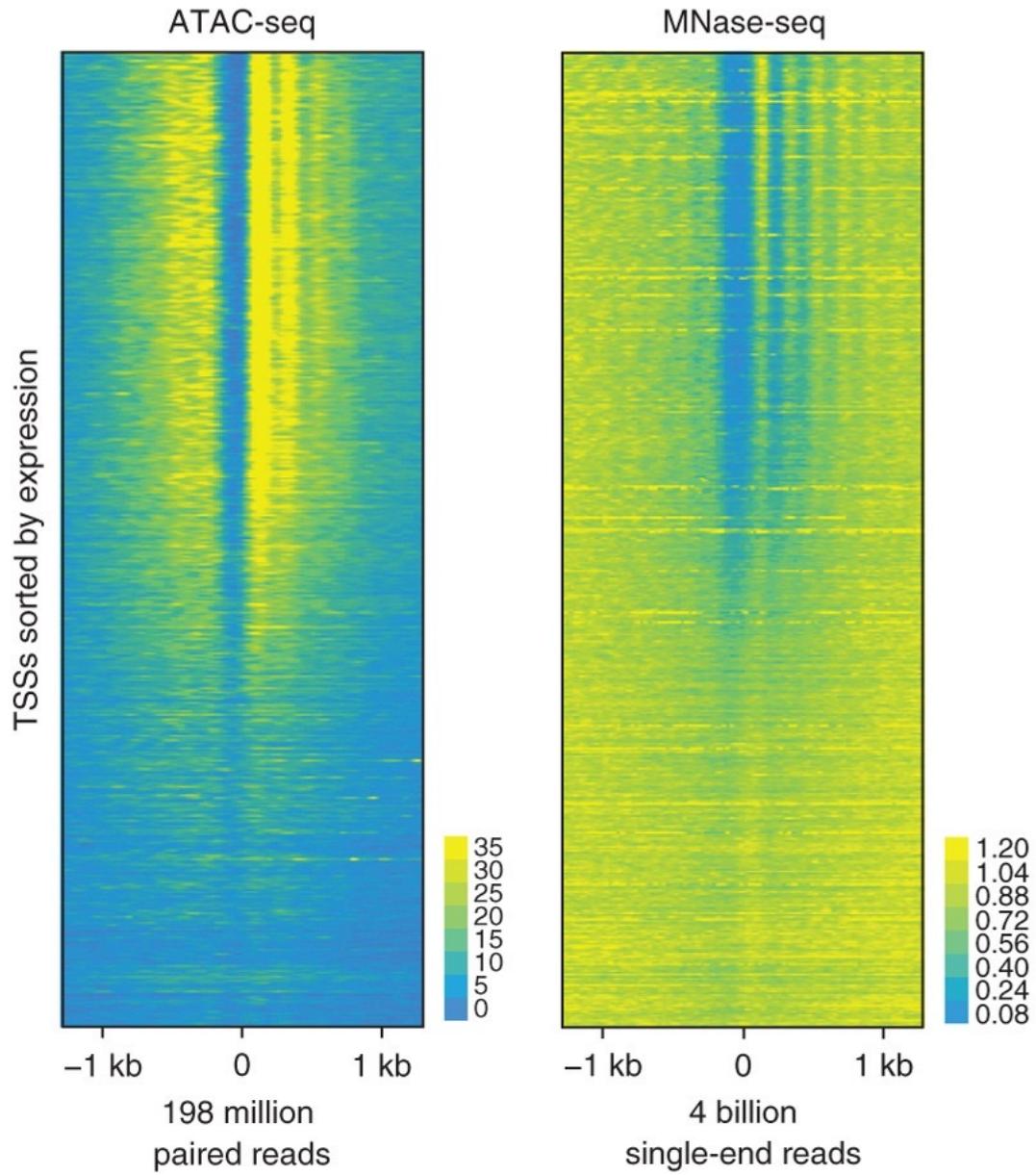
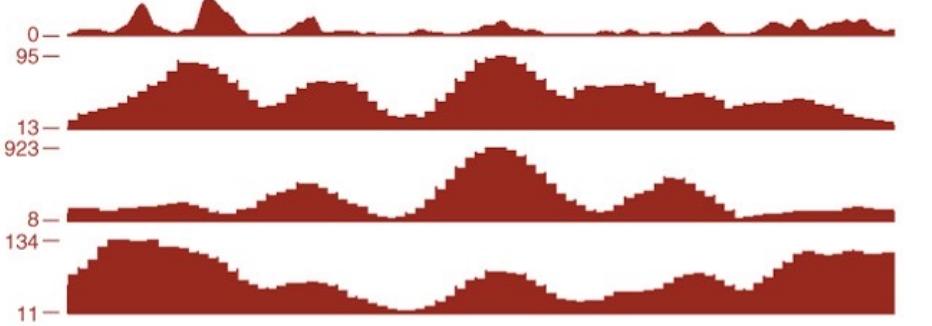


Buenrostro et al. Nat Methods 2013

# How to integrate patterns observed on signal tracks?



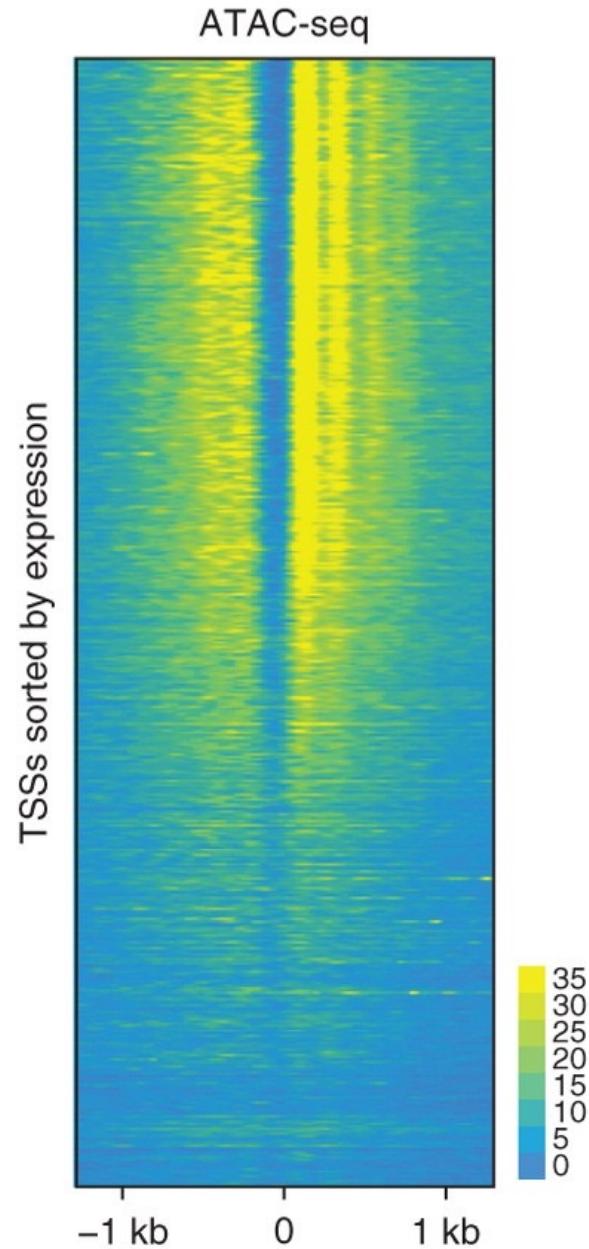
# Stacked “ripple” heatmap



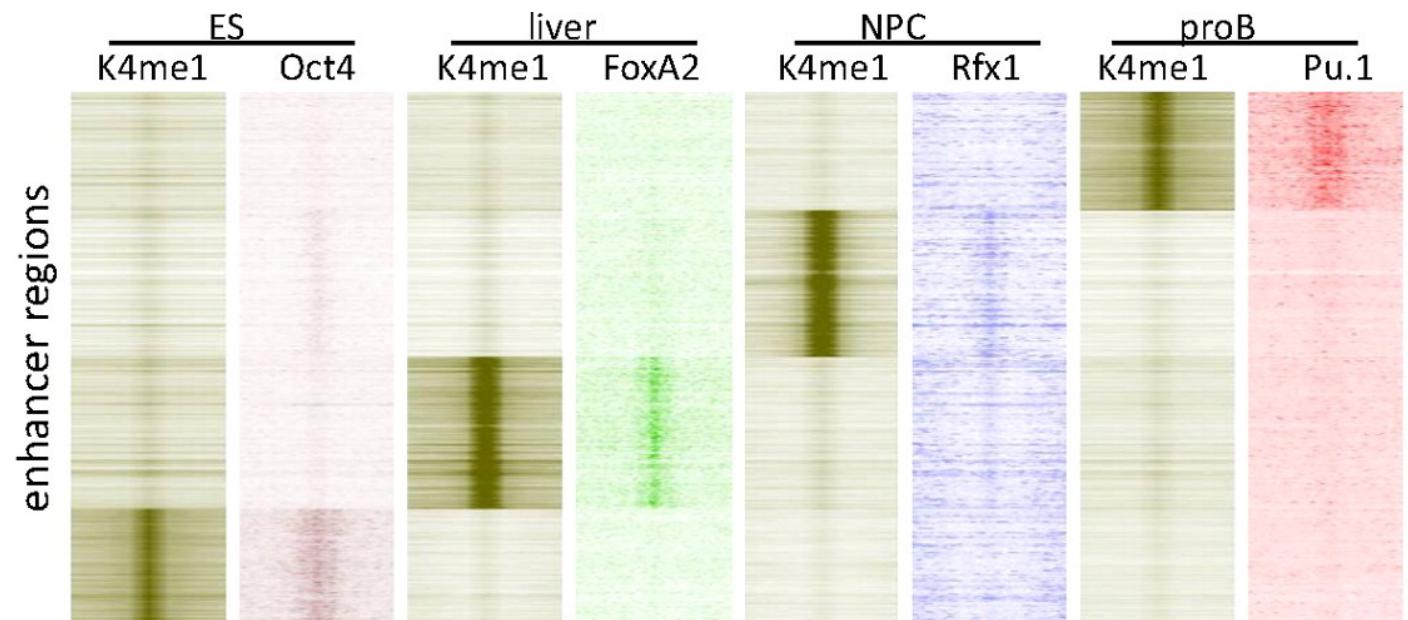
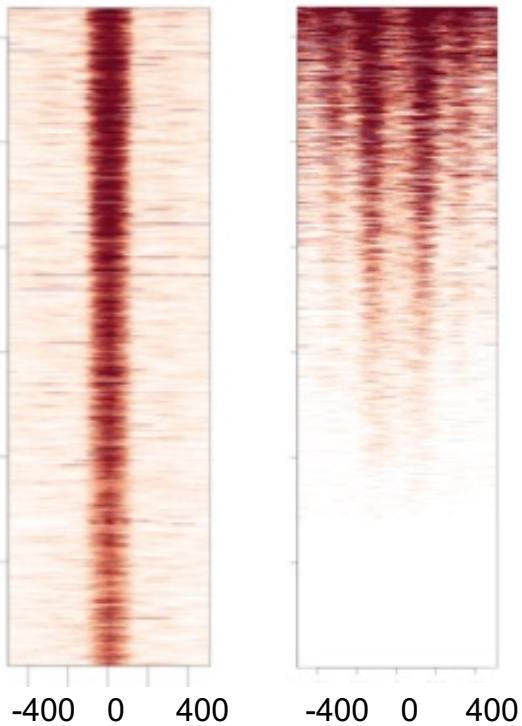
Buenrostro *et al.* *Nat Methods* 2013

# Essential elements in a ripple heatmap

- Heatmap presents 3-Dimensional data
- x: what loci/anchor is each row? Range?
- y: what are the rows? How many? How are they ranked?
- h: Data title/label (what signal?) color scale?

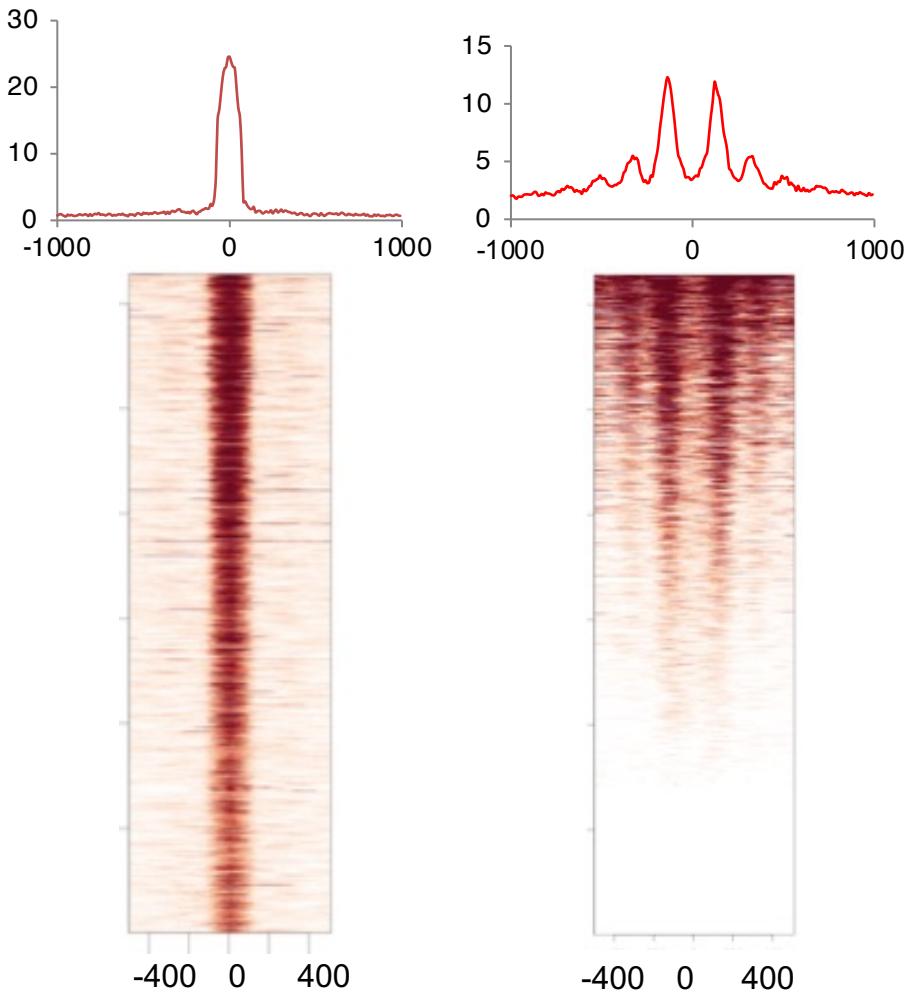


# Multiple datasets visualization by ripple heatmap



Creyghton *et al.* PNAS 2010

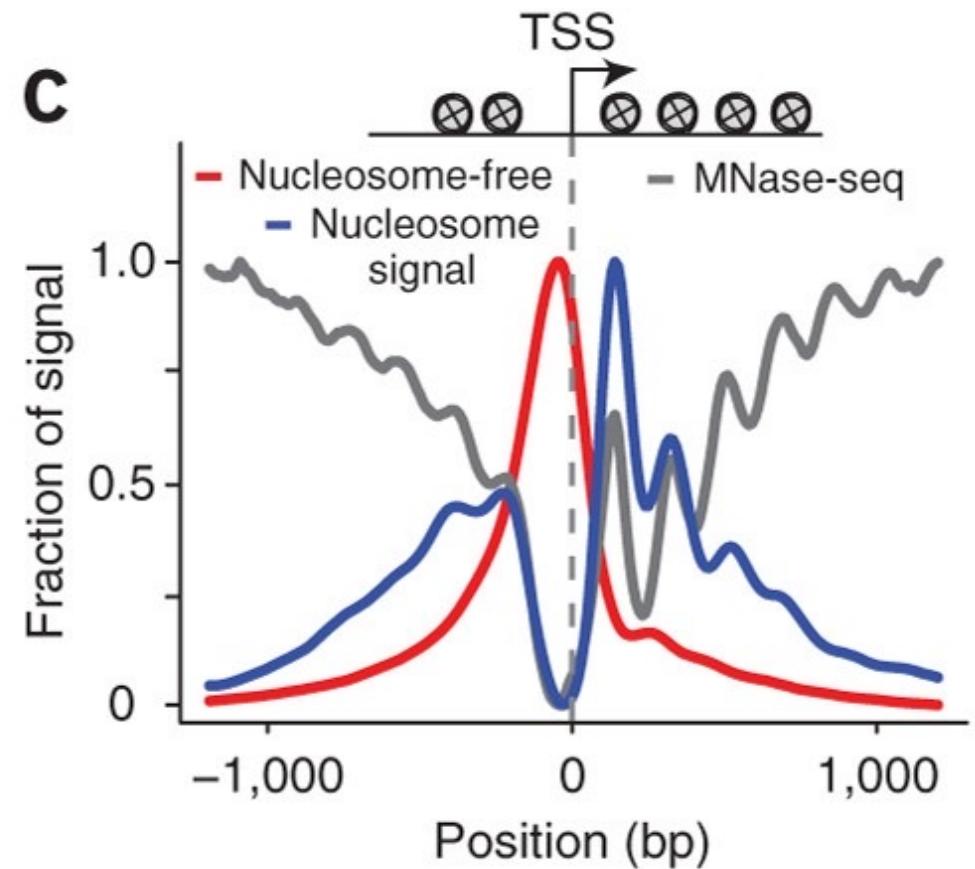
# Composite curve plots



Luyten et al. Genes Dev 2014

# Essential elements in a composite curve plot

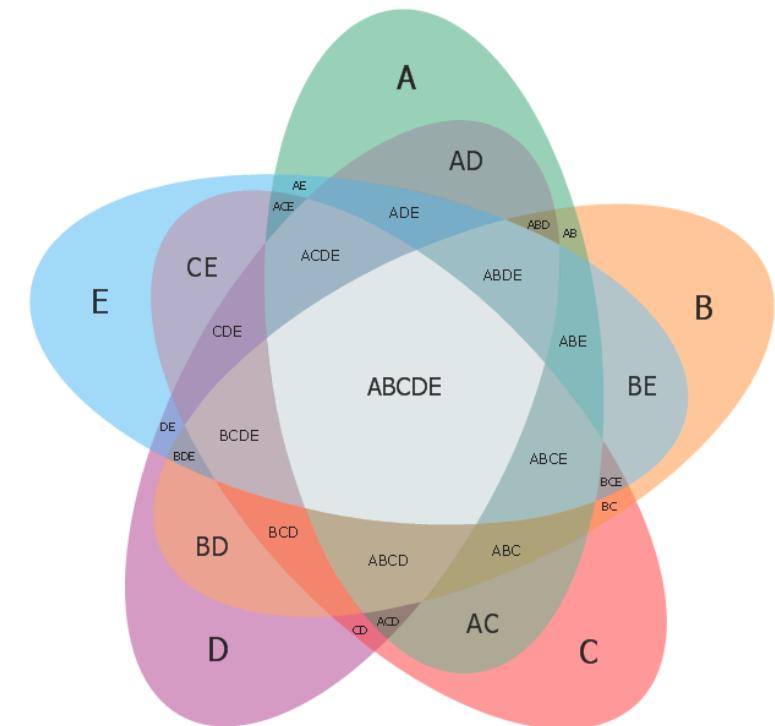
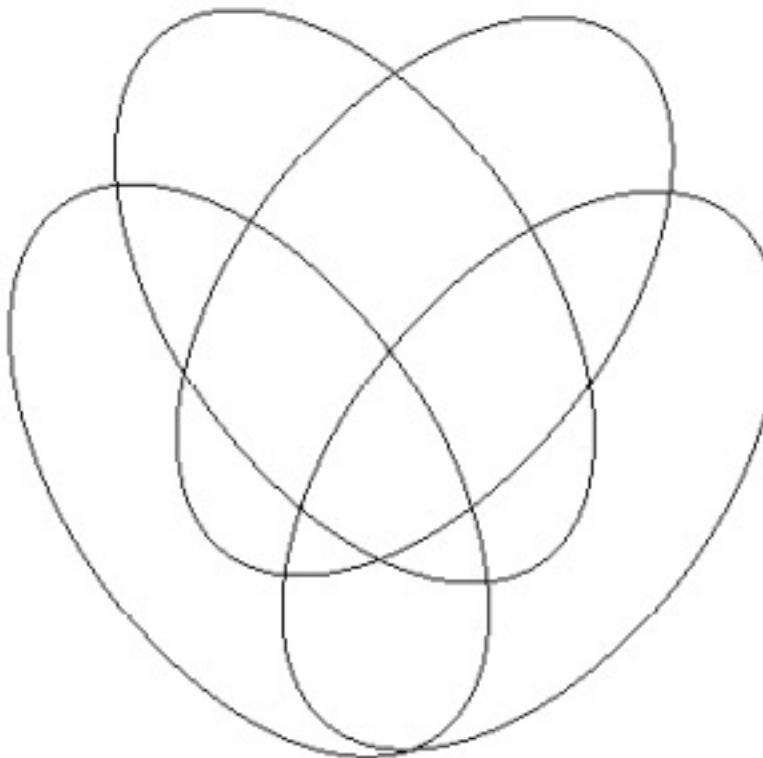
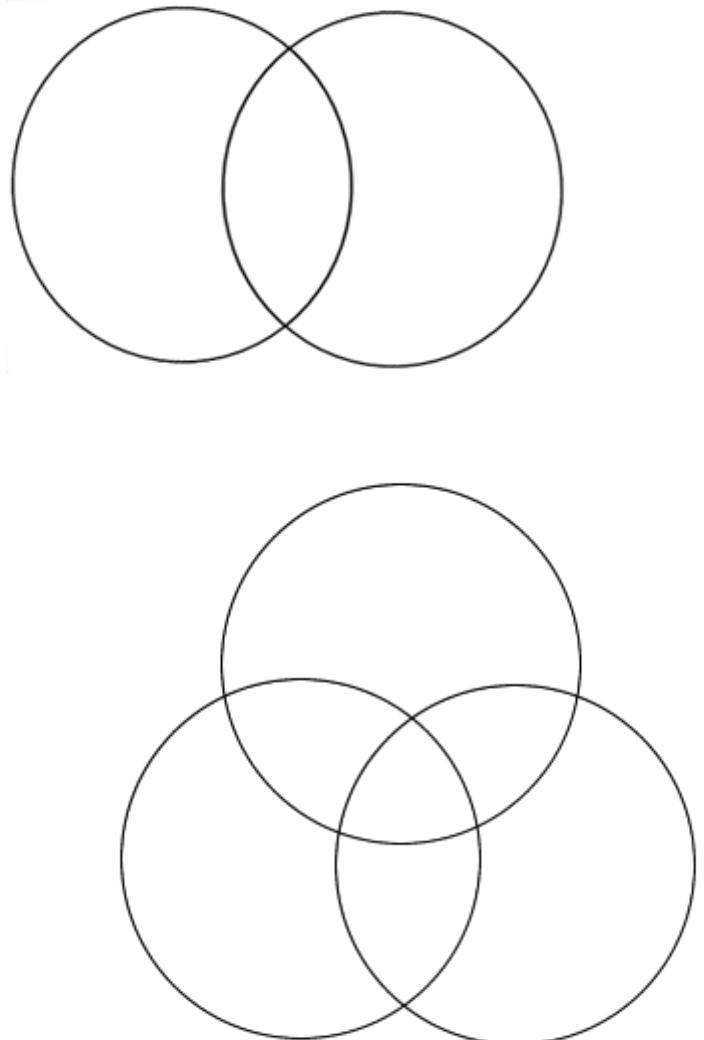
- Data title/label/legend
- Data source (average of what?)
- x: anchor, scale
- y: scale, normalization



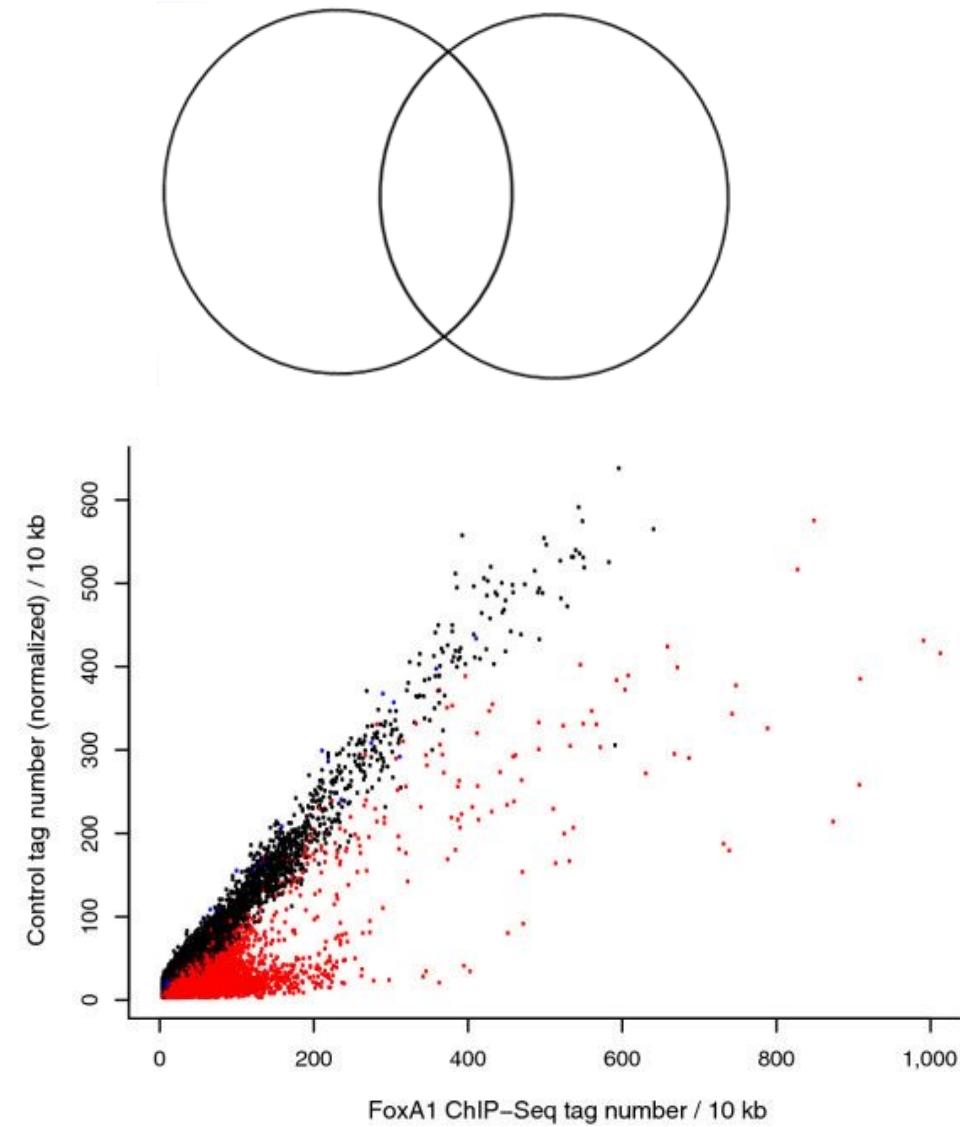
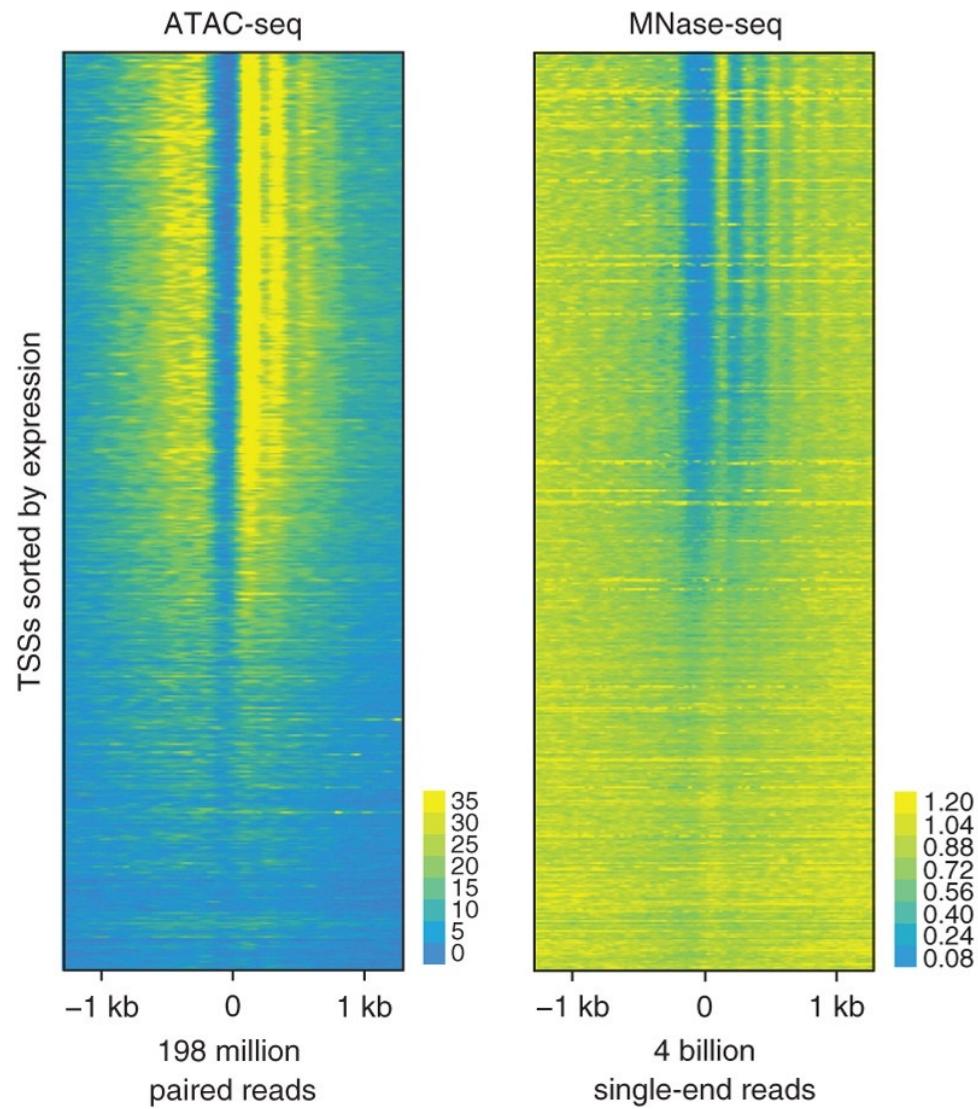
# Common misinterpretations of composite plots

- Caveat 1: A peak in a composite plot may be contribute by only a tiny fraction of regions (not representative of global picture)
- Caveat 2: A higher peak does not necessary mean stronger signal or more region coverage

# Venn diagram presents yes/no relations between sets

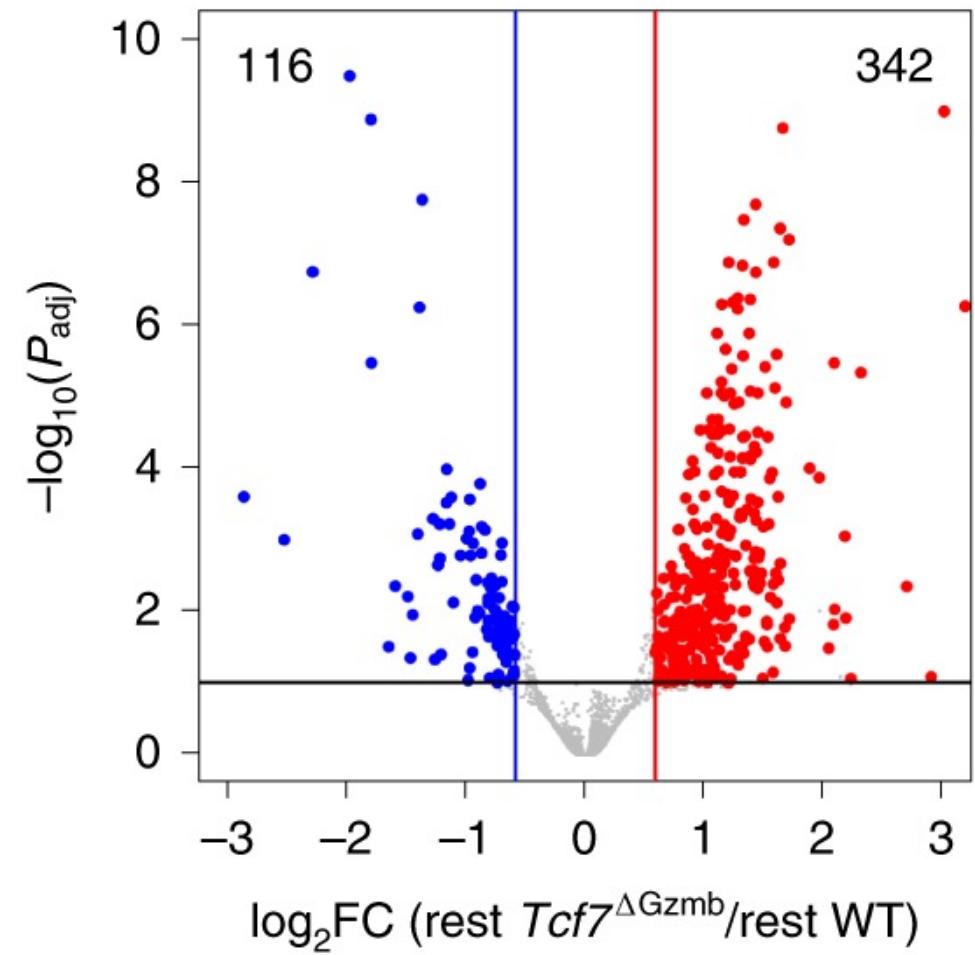


# Heatmaps and scatter plots keep more quantitative info



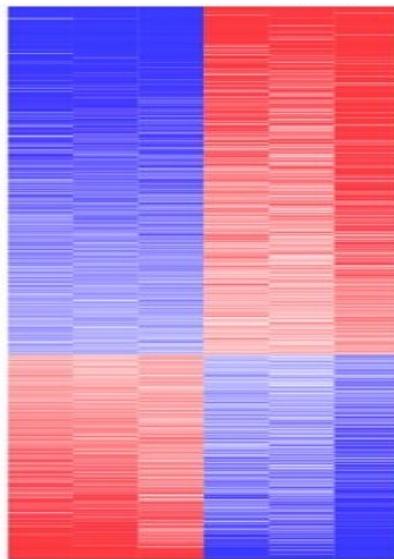
# Volcano plot for differential gene expression

- Scatter plot
- 2 dimension:
  - x: signal strength (e.g., log2 fold change)
  - y: statistical significance (e.g., -log10P)
- Set cutoffs on 2 axes



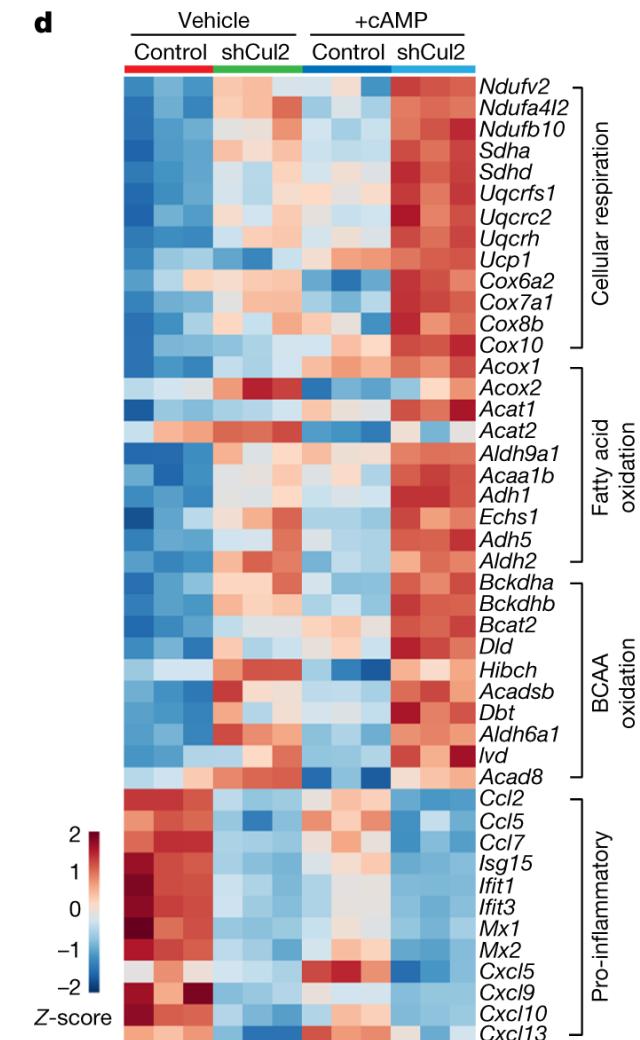
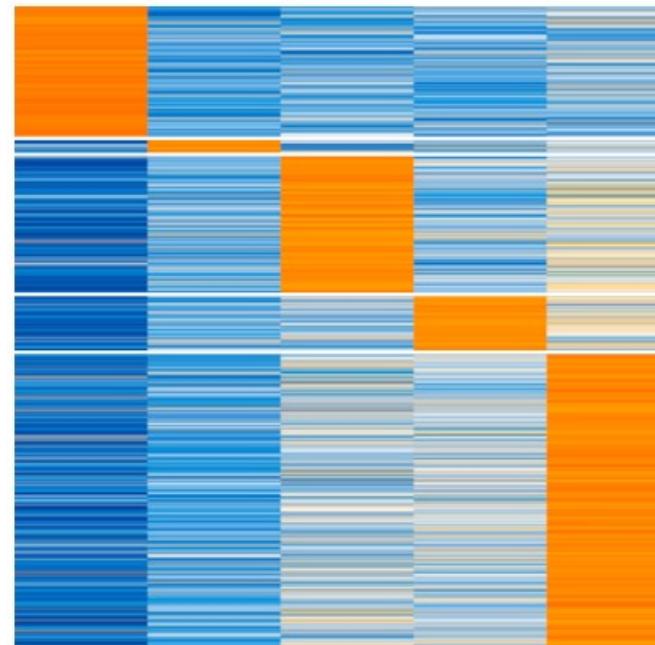
# Differential gene expression visualization by heatmap

Differential expression  
control    treated

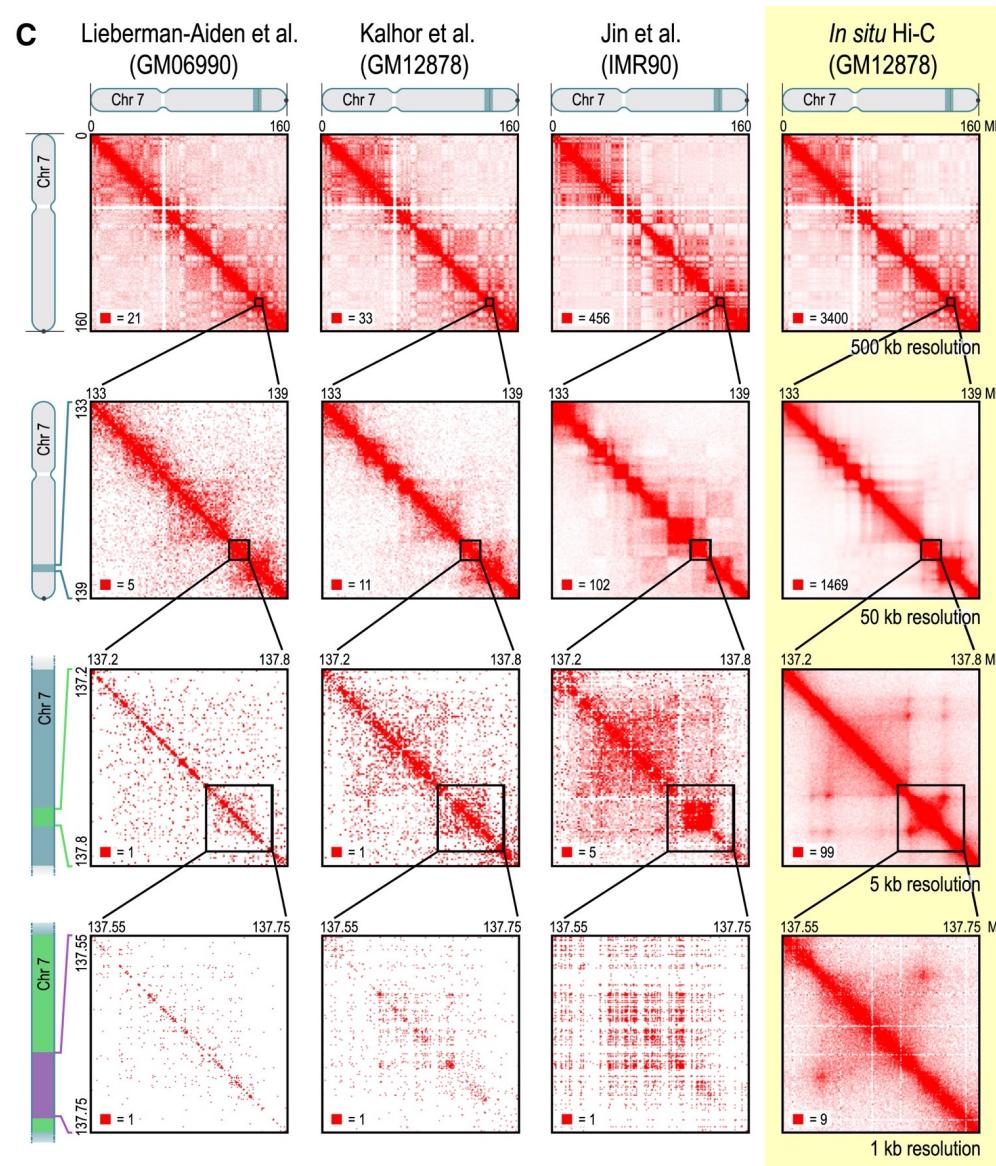


up-regulated  
down-regulated

Co-expressed gene clusters

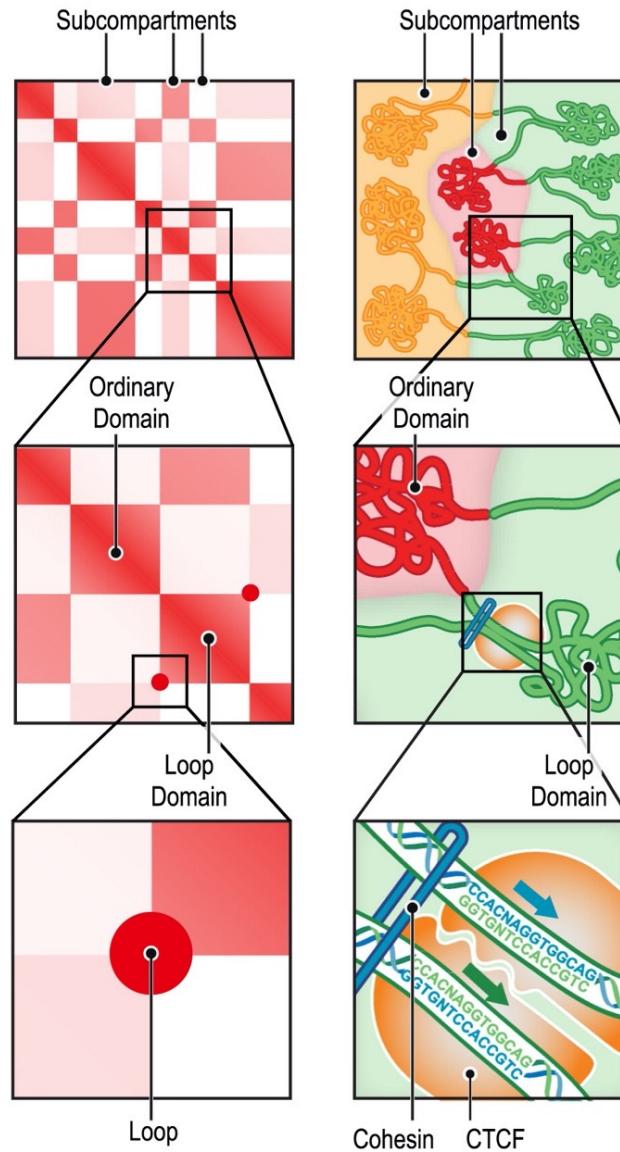


# Hi-C contact heatmap for 3D genome interactions



Rao et al. Cell 2014

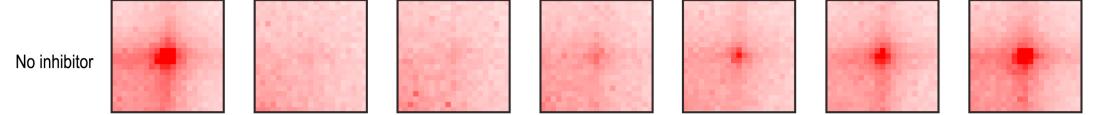
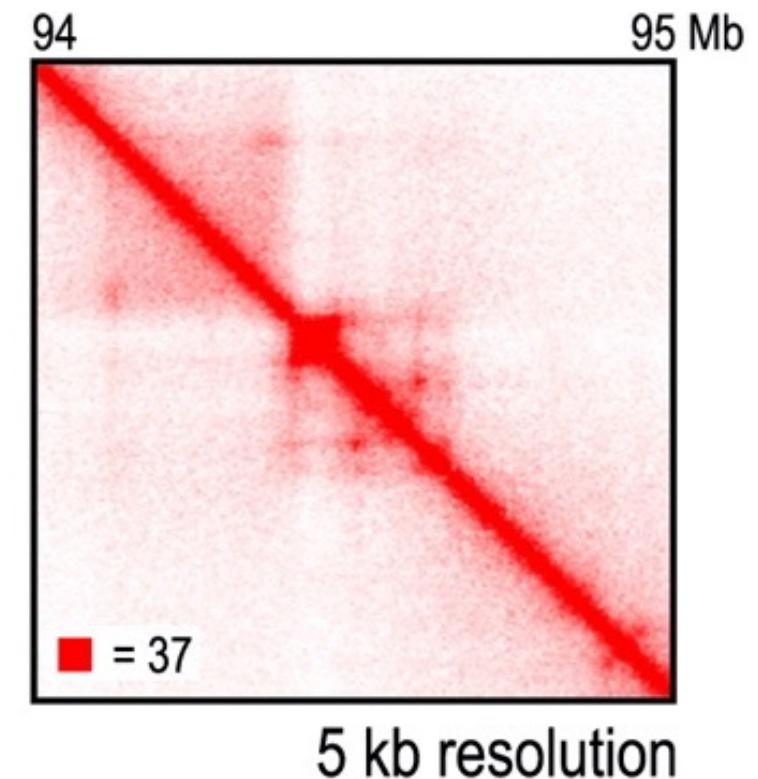
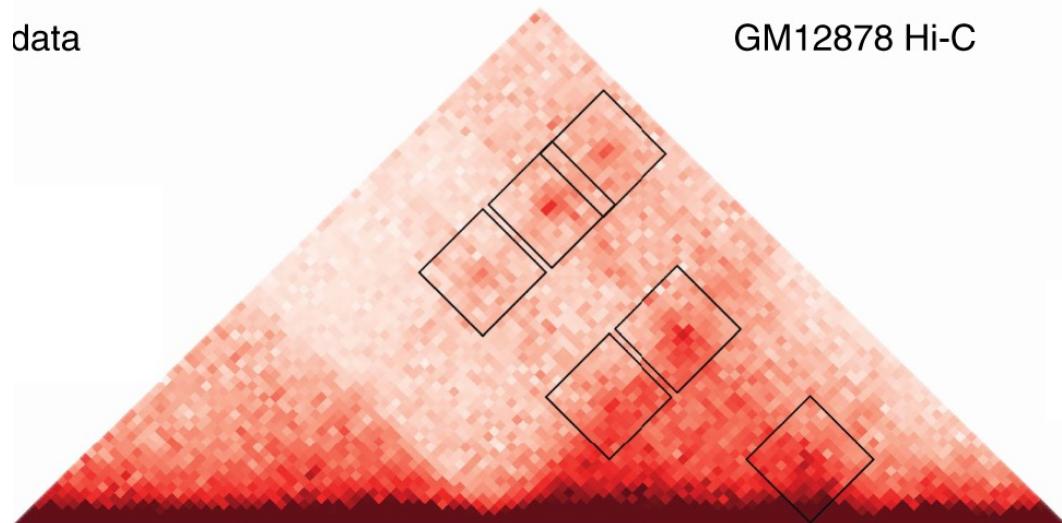
# Hi-C contact heatmap for 3D genome interactions



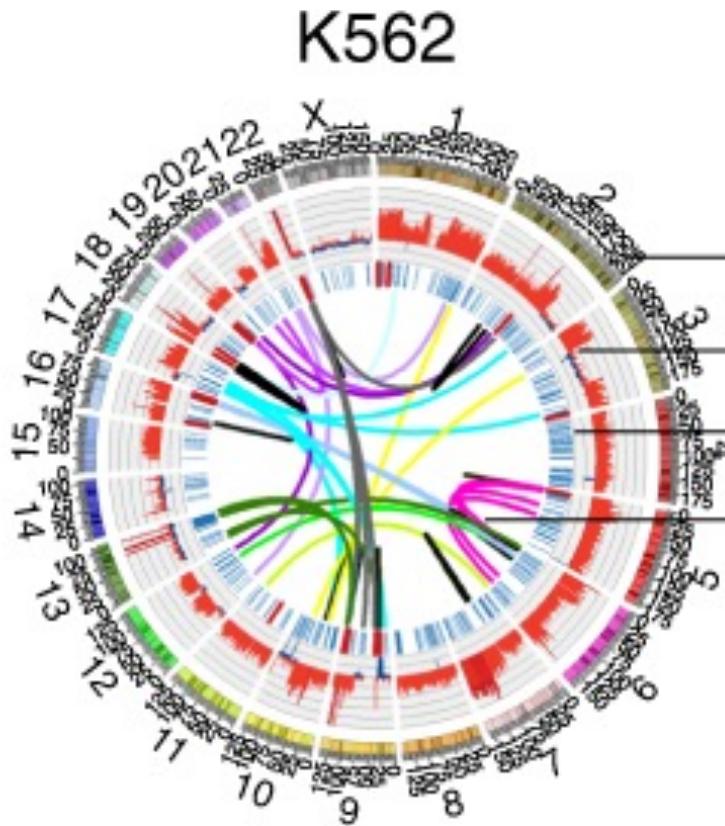
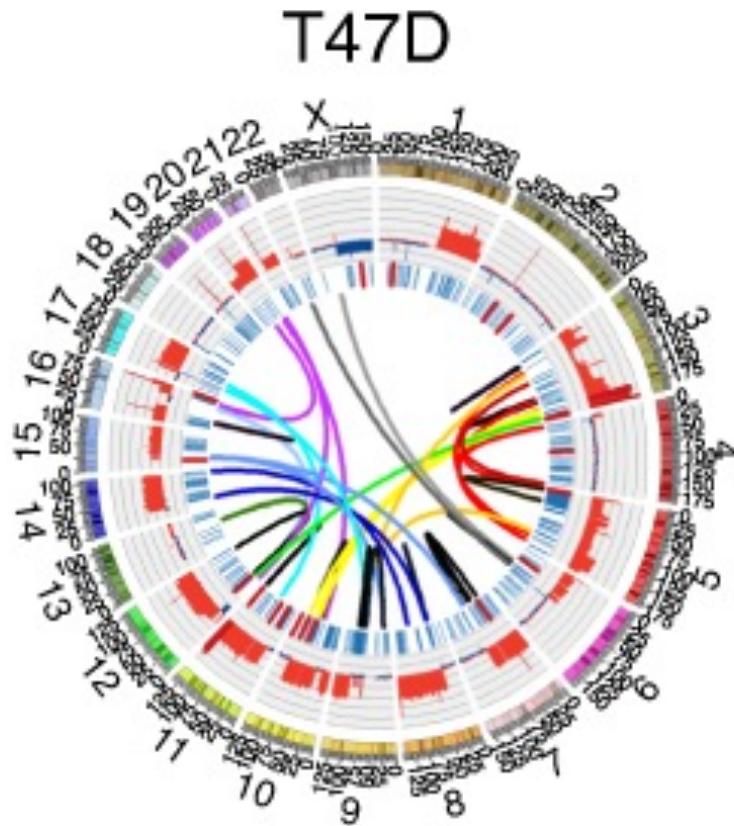
Rao et al. Cell 2014

# Essential elements in a Hi-C contact heatmap

- Scale, scale, scale
- Resolution
- Normalization
- Blocks, stripes, loops (2d peaks)

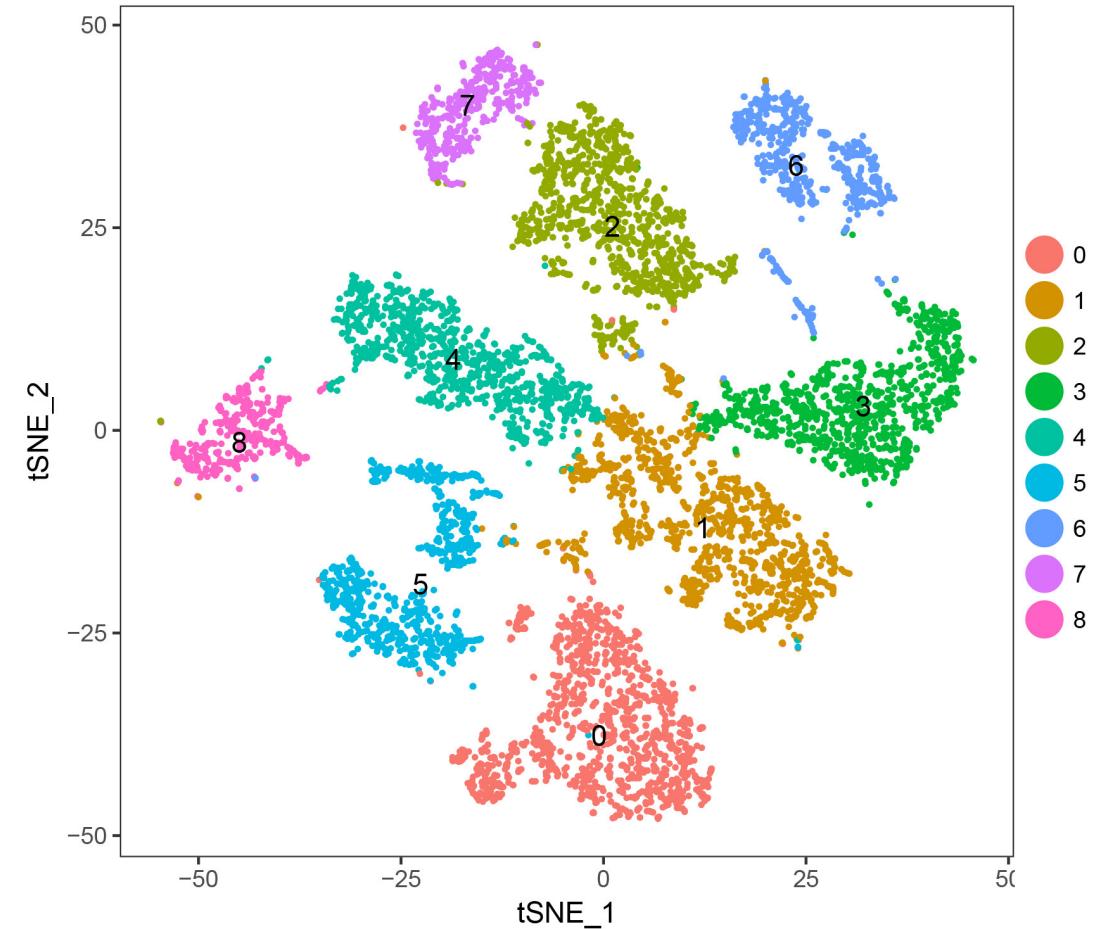
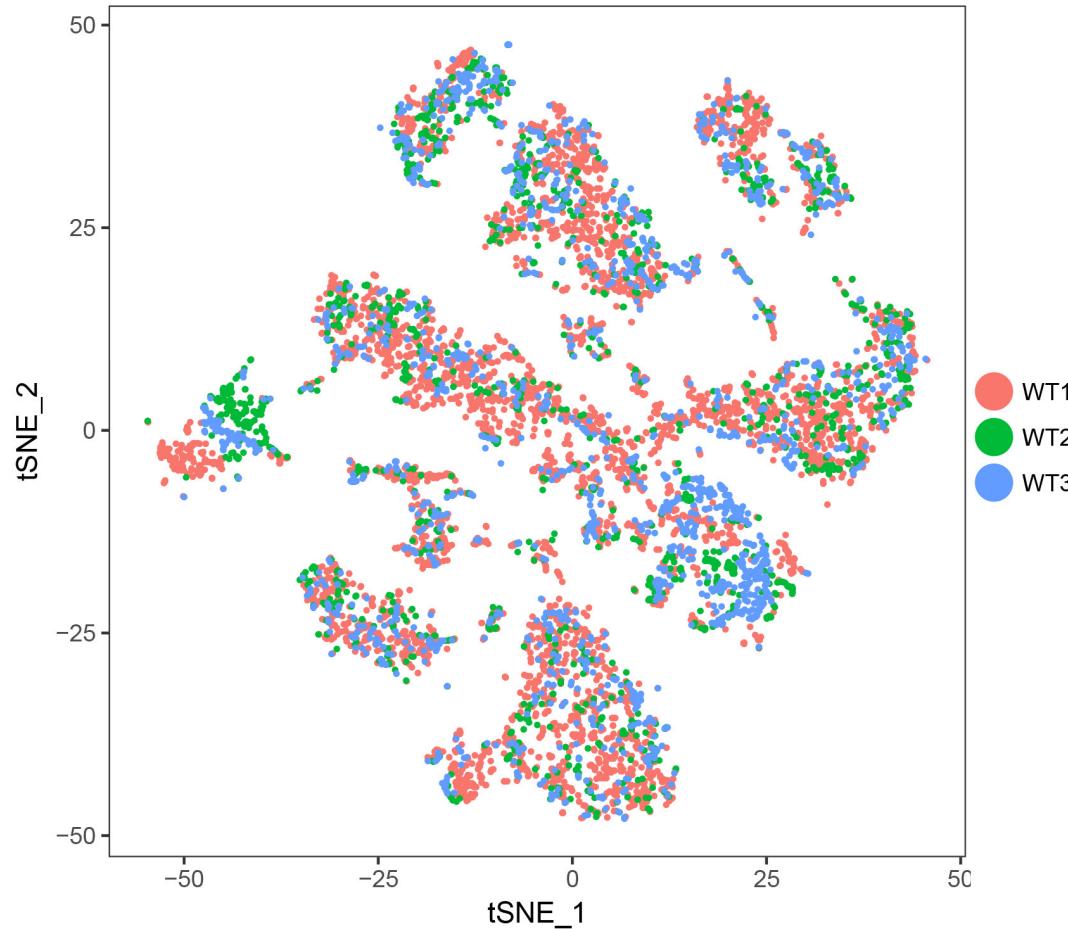


# Circos plot integrates multiple types of genomics data

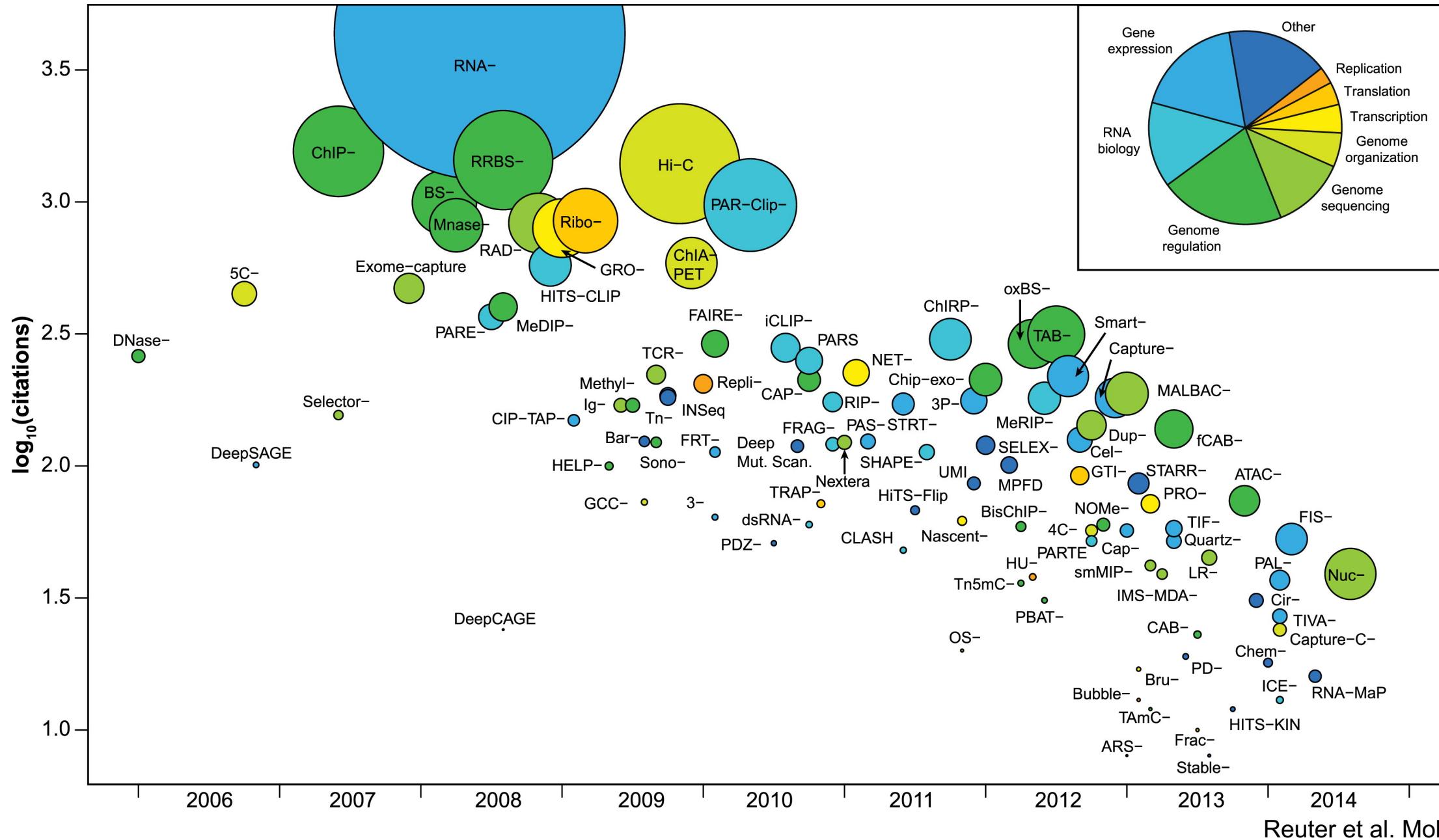


- Chromosome coordinates
- Copy number
- Deletion, duplication
- Interchromosomal TLs  
inversions, and  
unclassified  
intrachromosomal  
rearrangements (>1Mb)

# Single-cell data: clustering vs. t-SNE/UMAP visualization



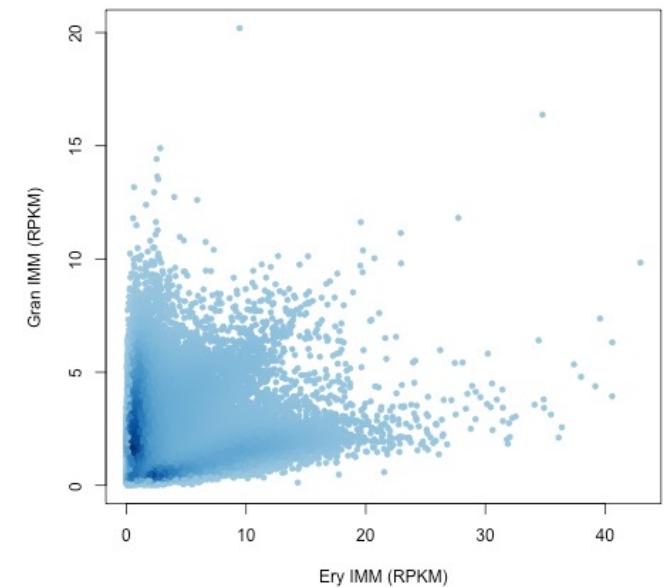
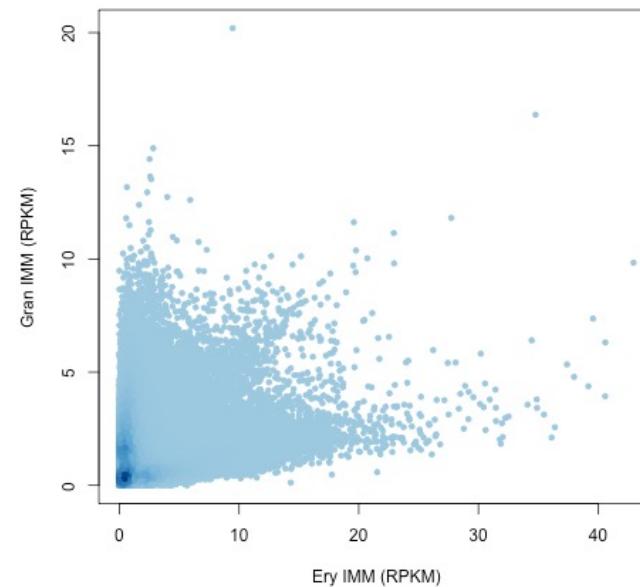
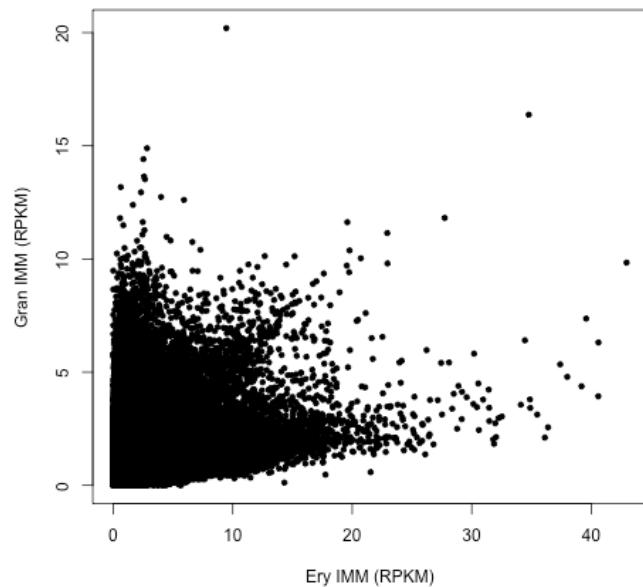
# Bubble plots: NGS-based applications (-seq)



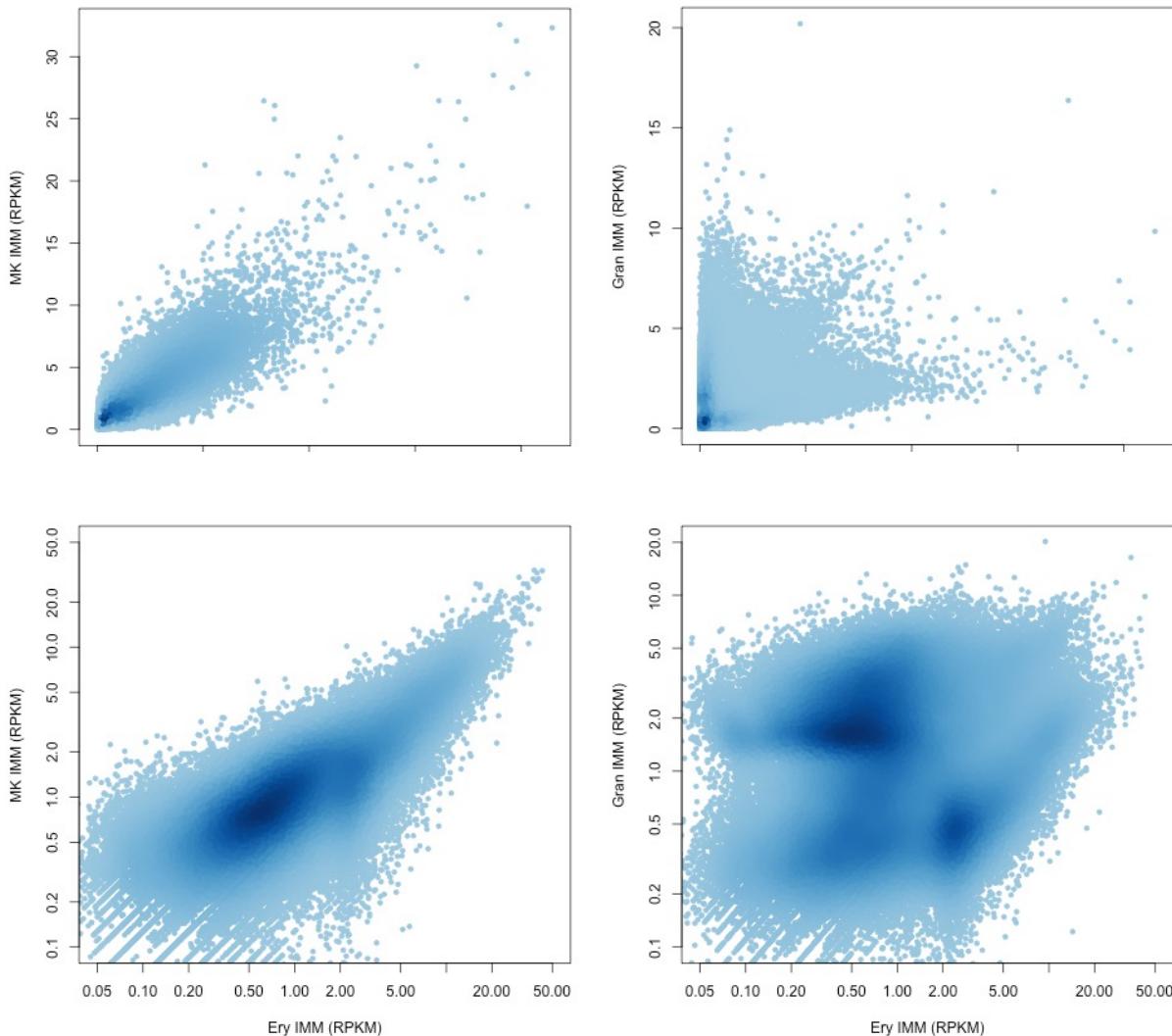
# **Some tips**

# Scatter plot

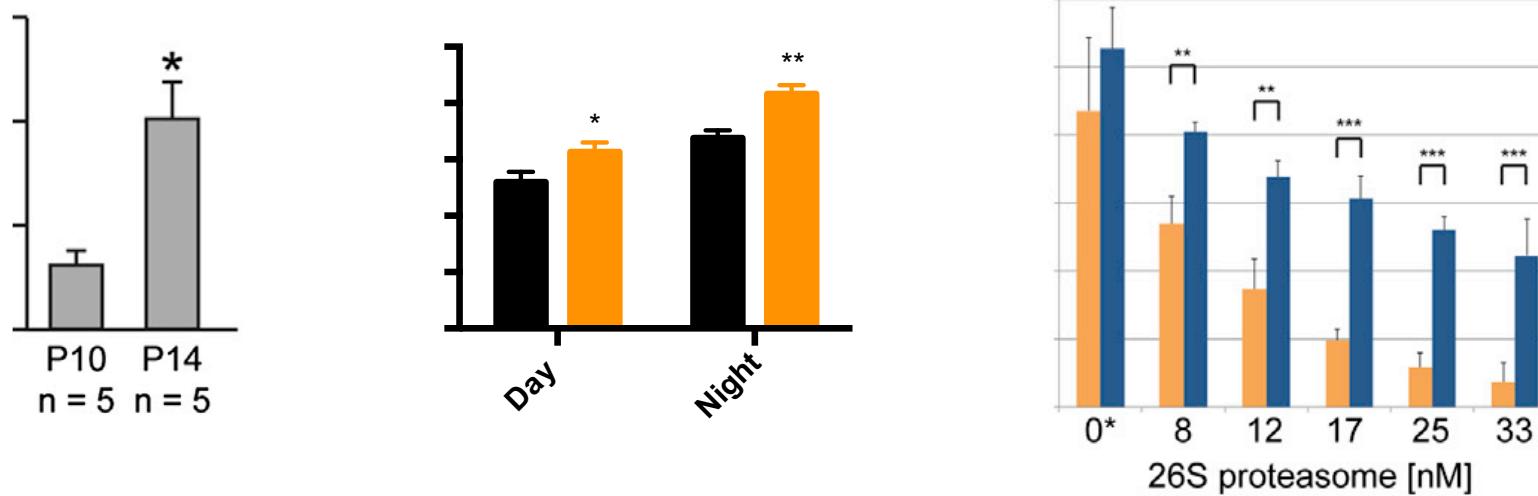
- The density of the data points matters!



# Choose the appropriate scale for plotting (linear or logarithm)



# Bar chart?



Bar charts should always start from 0, and on the linear scale.  
If difference is small, box plots or original dots are better.

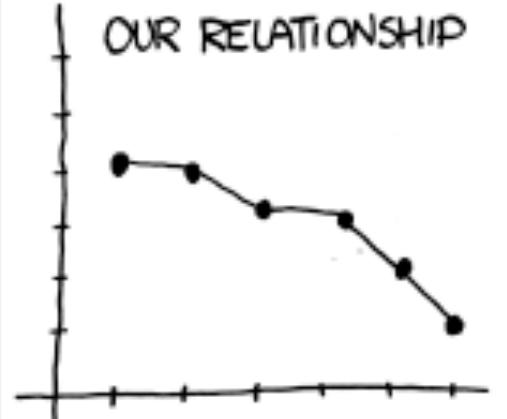
Plots from *Cell* 2014

I THINK WE SHOULD  
GIVE IT ANOTHER SHOT.

{ WE SHOULD BREAK  
UP, AND I CAN  
PROVE IT.



OUR RELATIONSHIP



HUH.

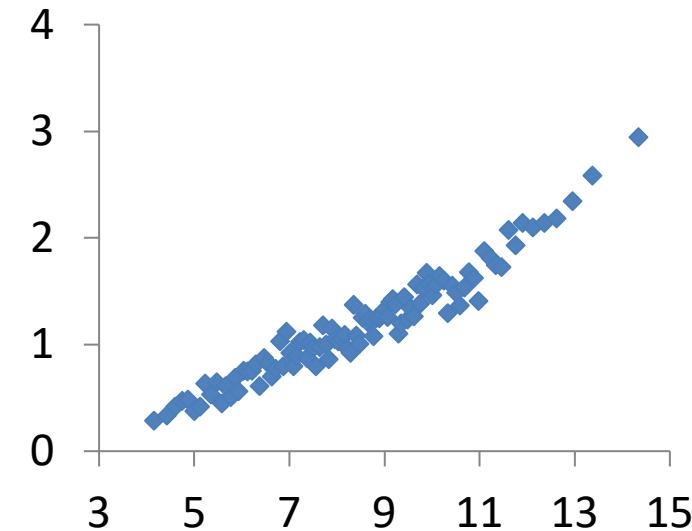
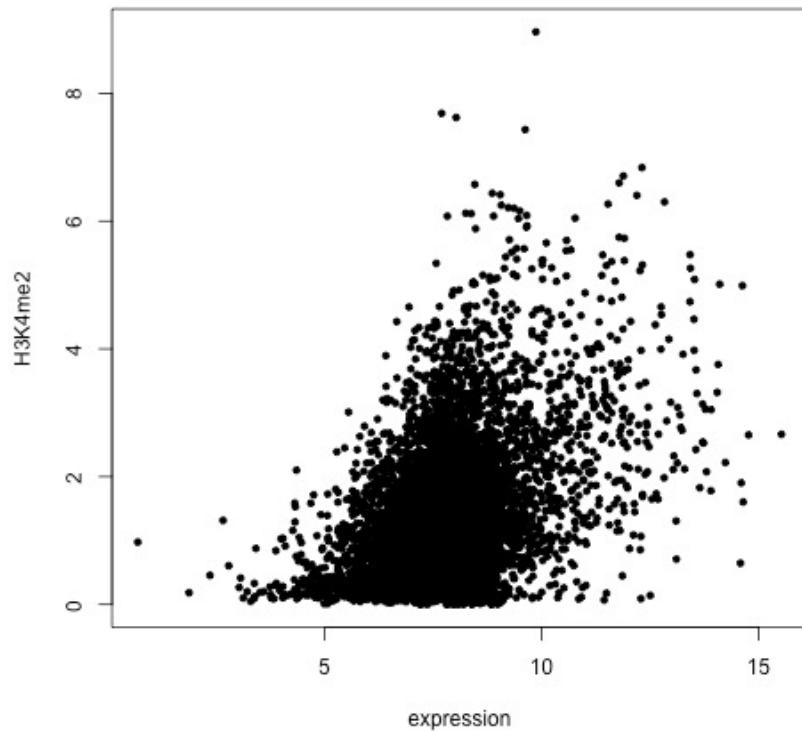


MAYBE YOU'RE RIGHT.

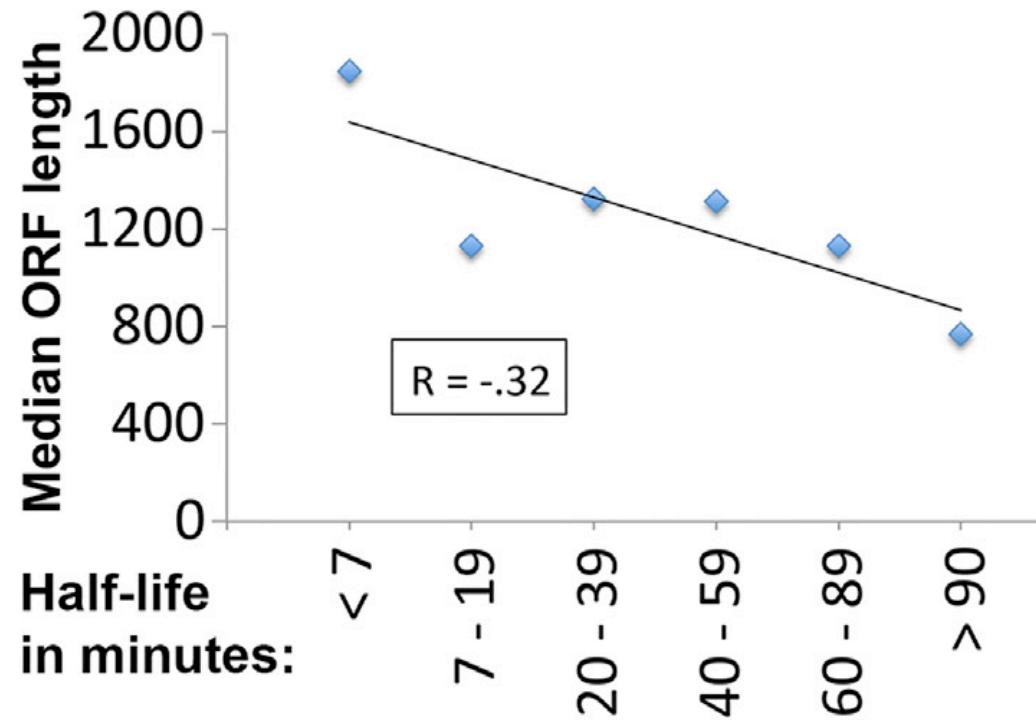
{ I KNEW DATA WOULD CONVINCE YOU.  
NO, I JUST THINK I CAN DO  
BETTER THAN SOMEONE WHO  
DOESN'T LABEL HER AXES.



# Scatter plot? Group them if needed



# However, grouping should be unbiased

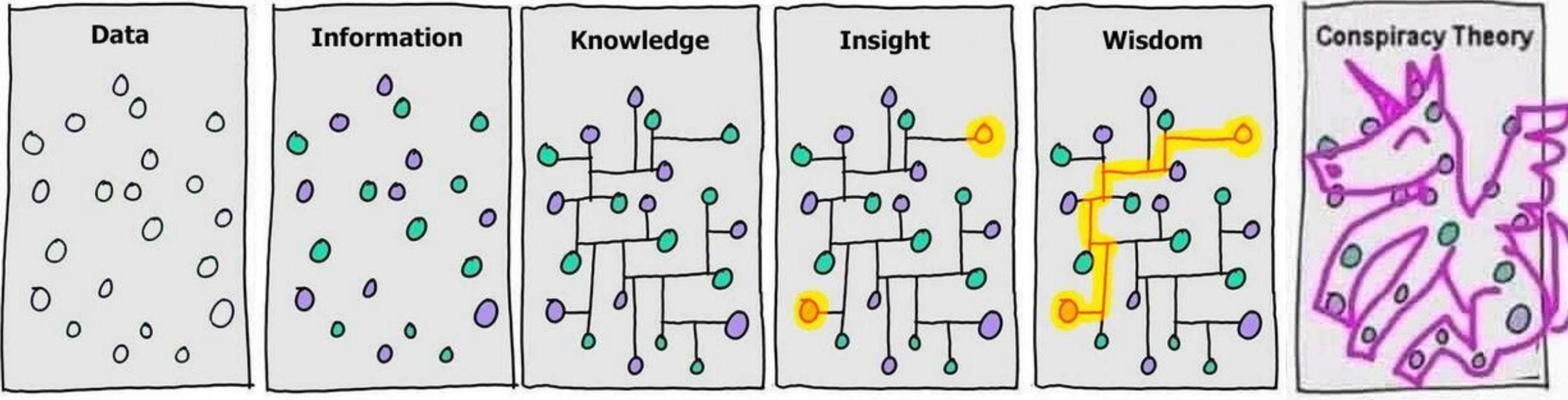


Cell 2014

# Take-Home Tips

- Always read the axes and pay attention to the scales on a figure.
- Bar charts should always start from zero on the linear scale.
- Data point density on a scatter plot is important.
- Group the data points if needed, but do it in an unbiased way.

# Having the data is not enough; presentation and interpretation matter



RNA-seq

ChIP-seq

DNase/ATAC-seq

Hi-C

Single-cell  
resolution...

...

Gene expression

Protein factors

Chromatin

3D genome

Multi-omics

...

Transcriptional  
regulation,

Chromatin

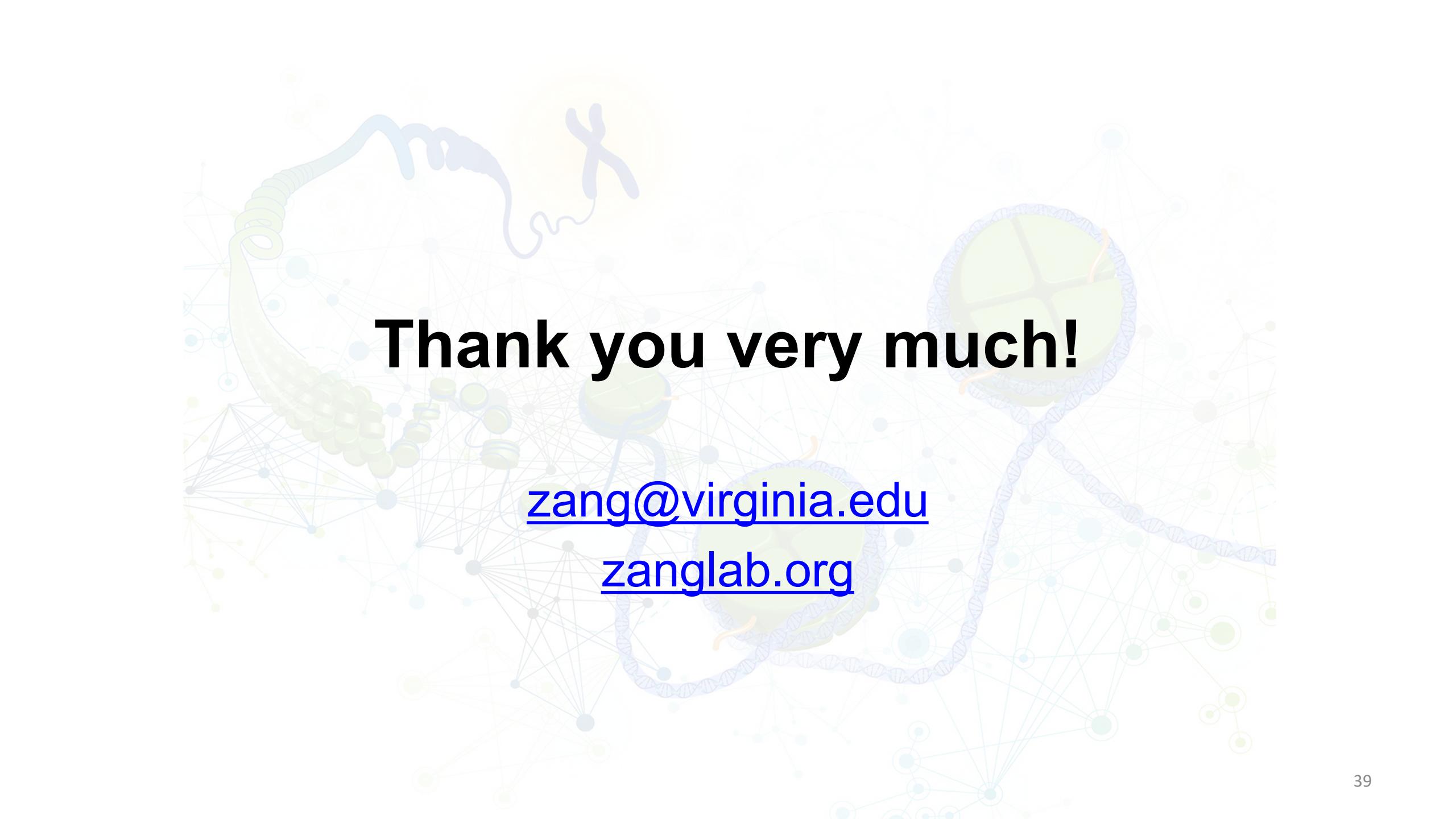
organization,

...

New insight?

New biology!

Overfitting,  
overinterpretation...



# Thank you very much!

[zang@virginia.edu](mailto:zang@virginia.edu)

[zanglab.org](http://zanglab.org)