Visualization and interpretation of high-throughput genomics data

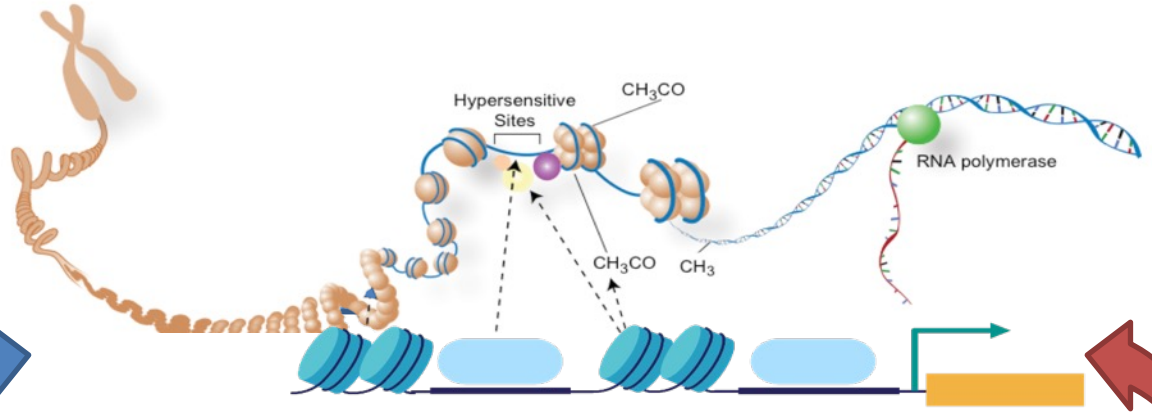
Chongzhi Zang
Associate Professor
zang@virginia.edu
zanglab.org

August 22, 2023

zanglab.org

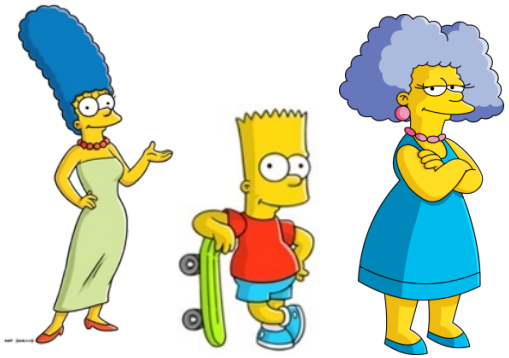
My lab develops computational methods and uses computational approaches to study epigenetics and transcriptional regulation

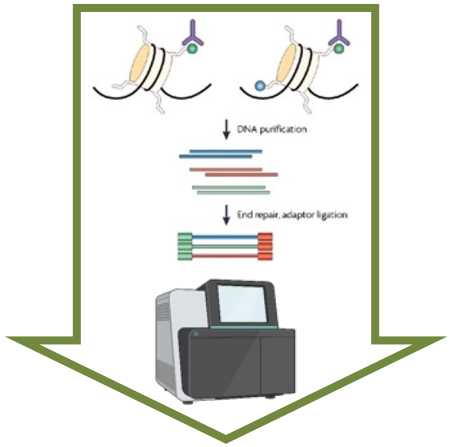
zanglab.org



Method development

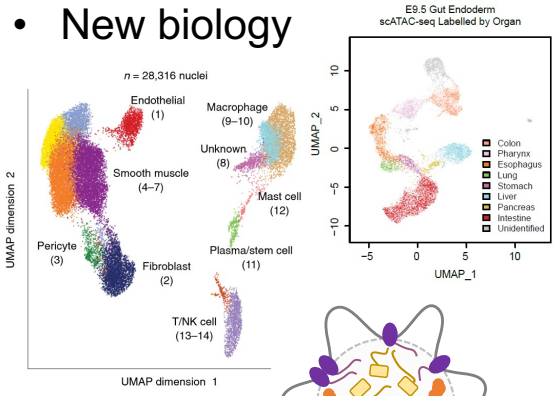
- Stat/math/physics /machine learning models
- Algorithms
- Software tools

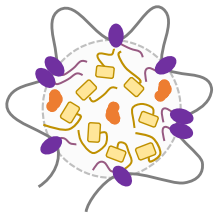




Data-inspired discovery

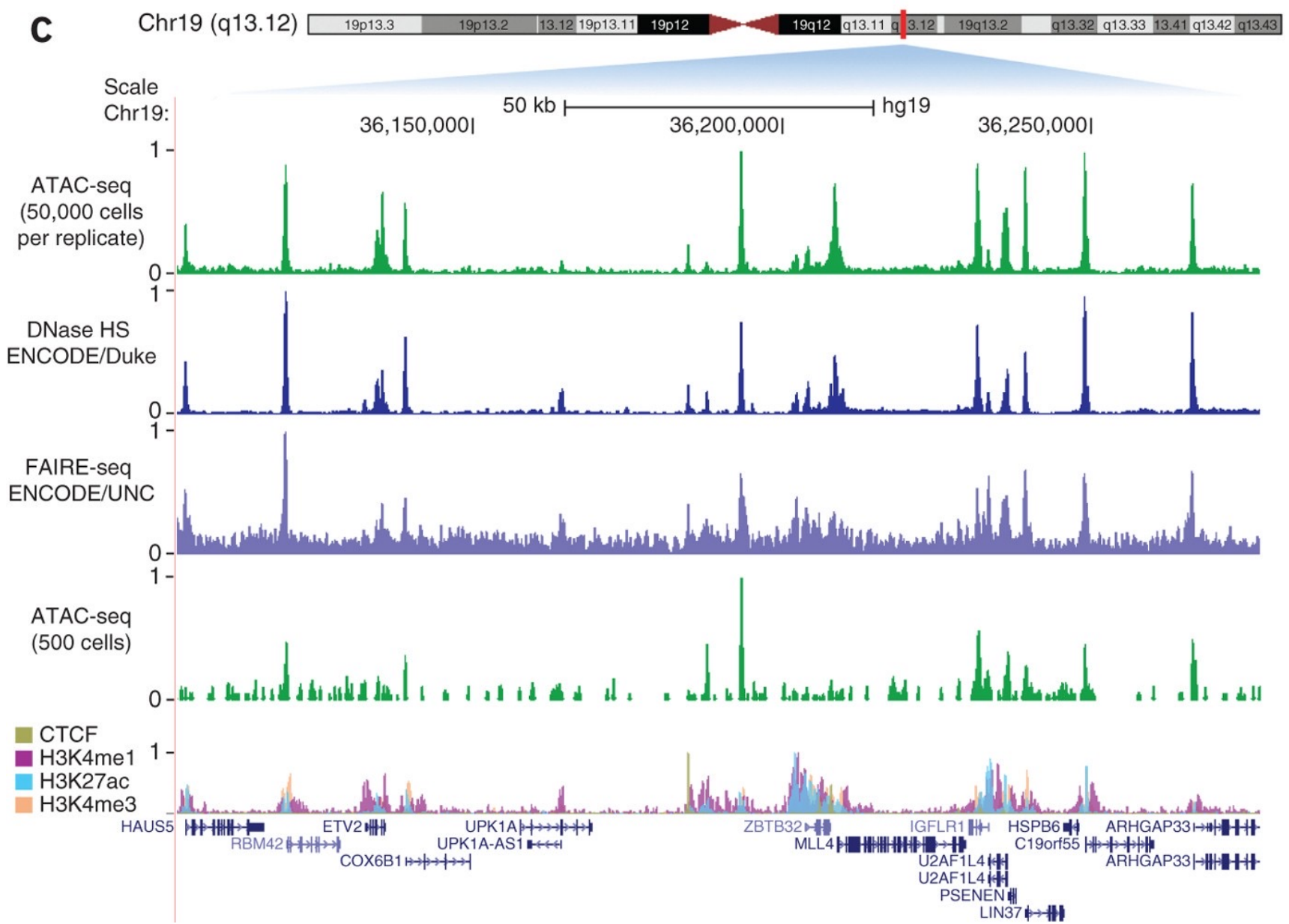
- Multi-modal data integration
- Innovative models
- New biology



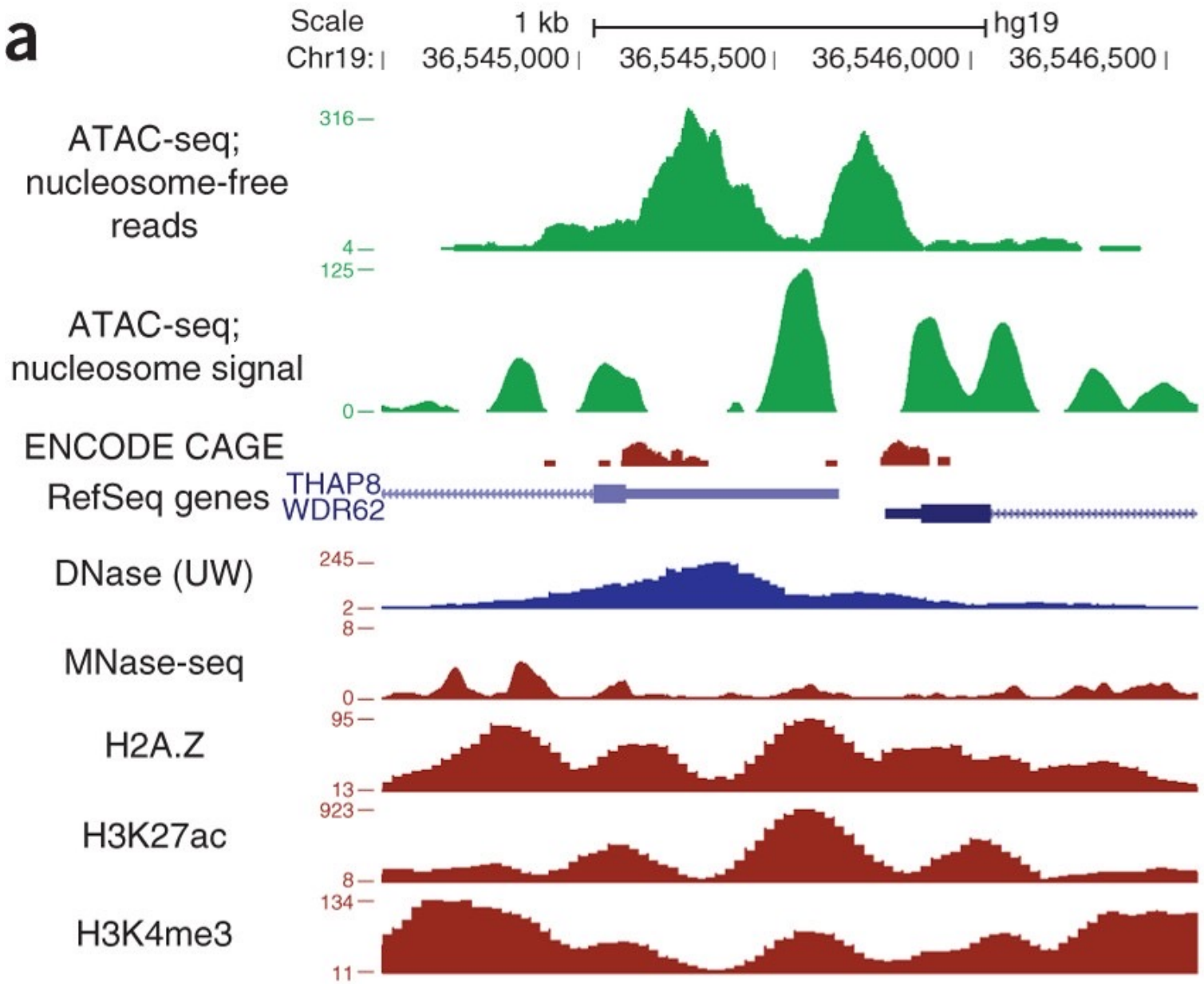


Learning Objectives

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

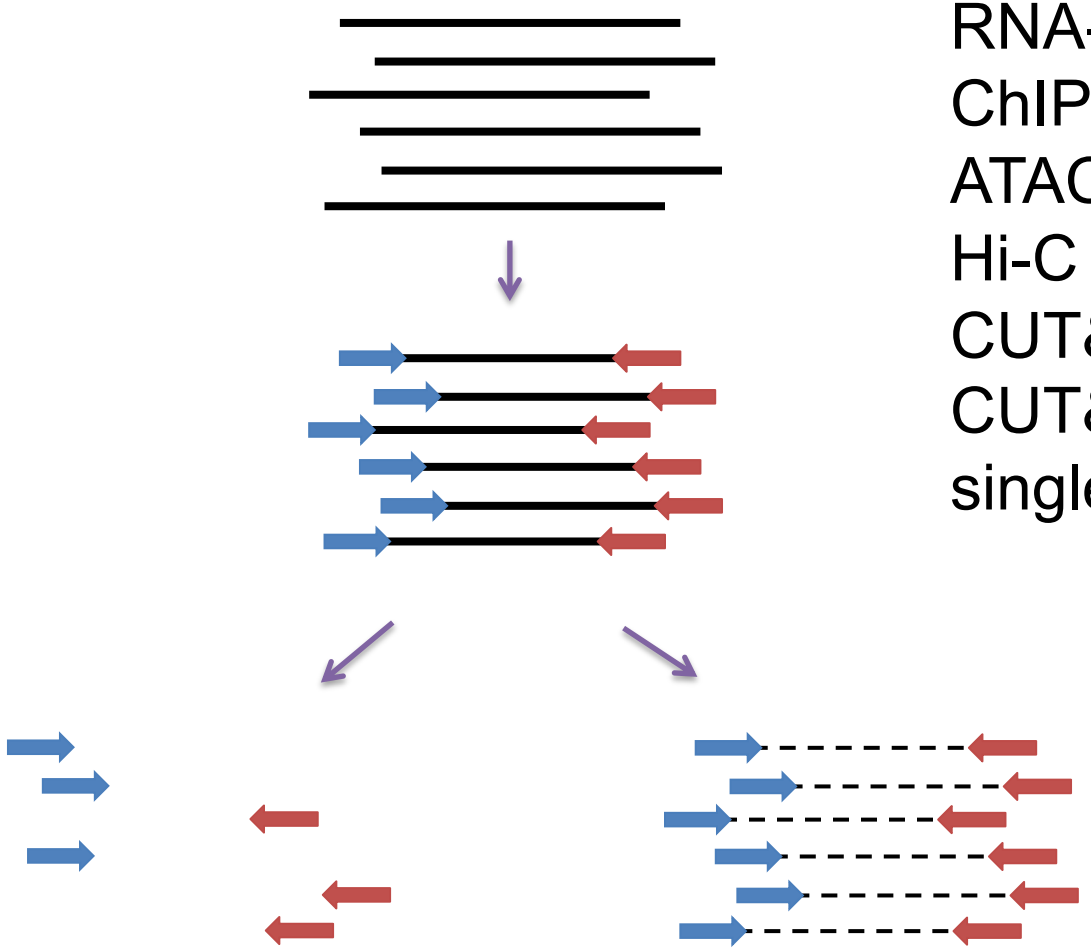


Buenrostro *et al*. *Nat Methods* 2013



Buenrostro *et al*. *Nat Methods* 2013

High-throughput short-read sequencing (Illumina)

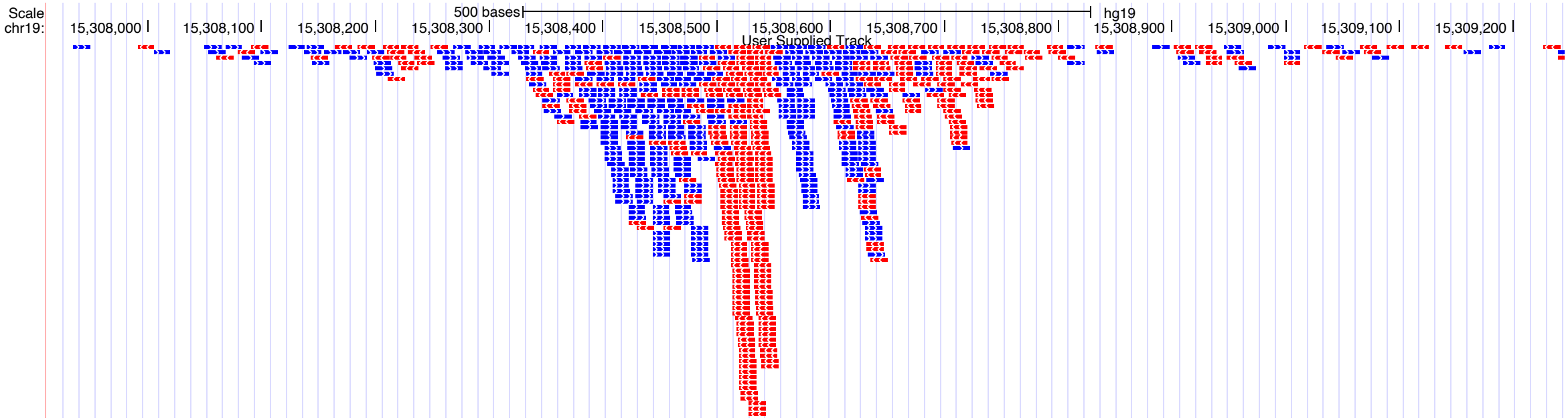


Single-end

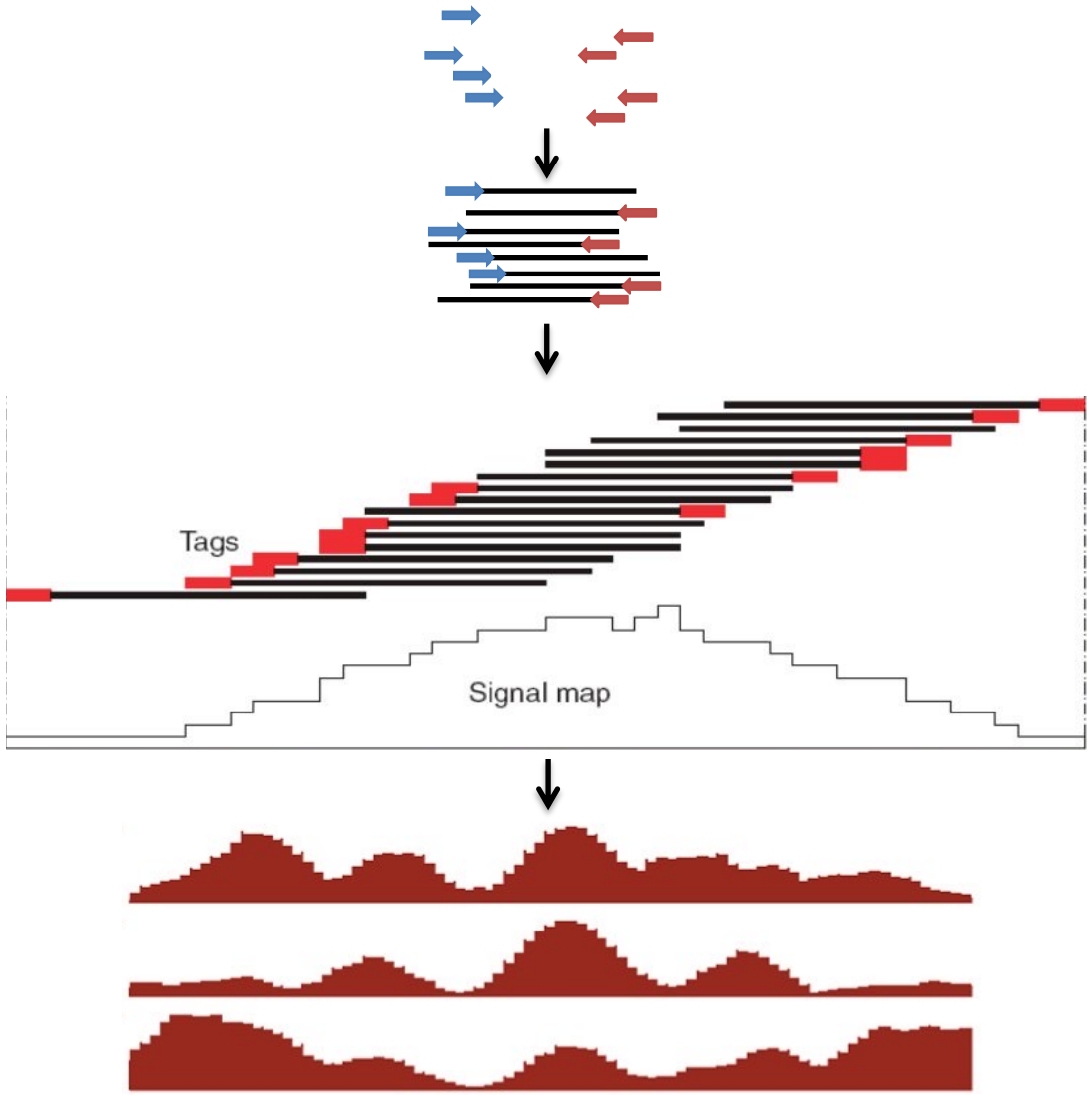
Paired-end

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
```

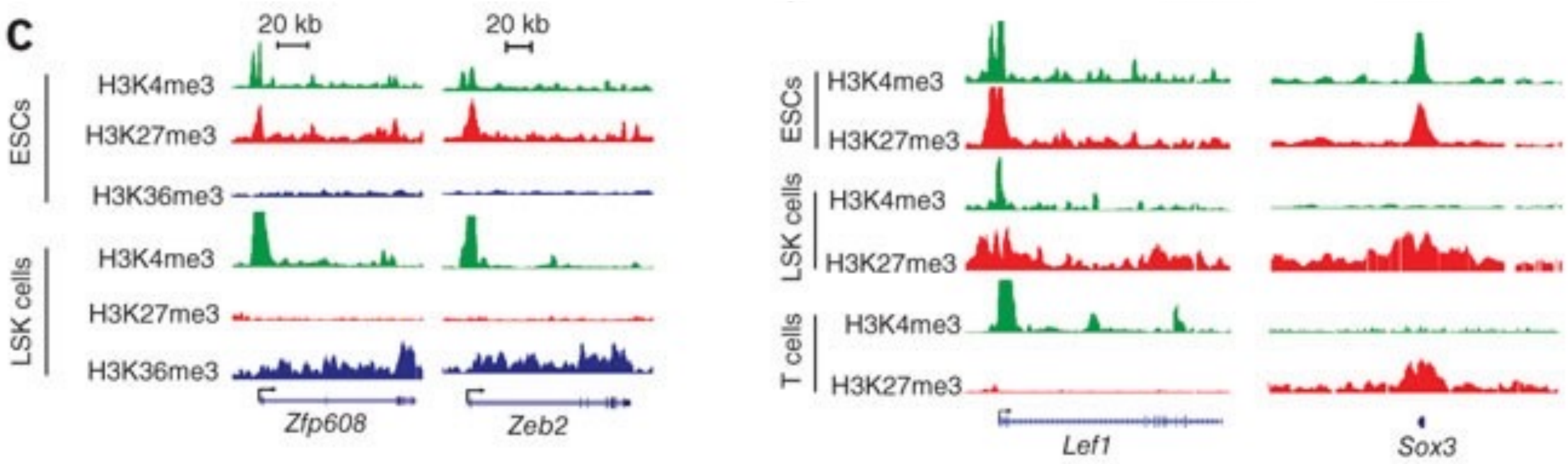
- bigWig: indexed binary format

Essential elements in genome browser tracks

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

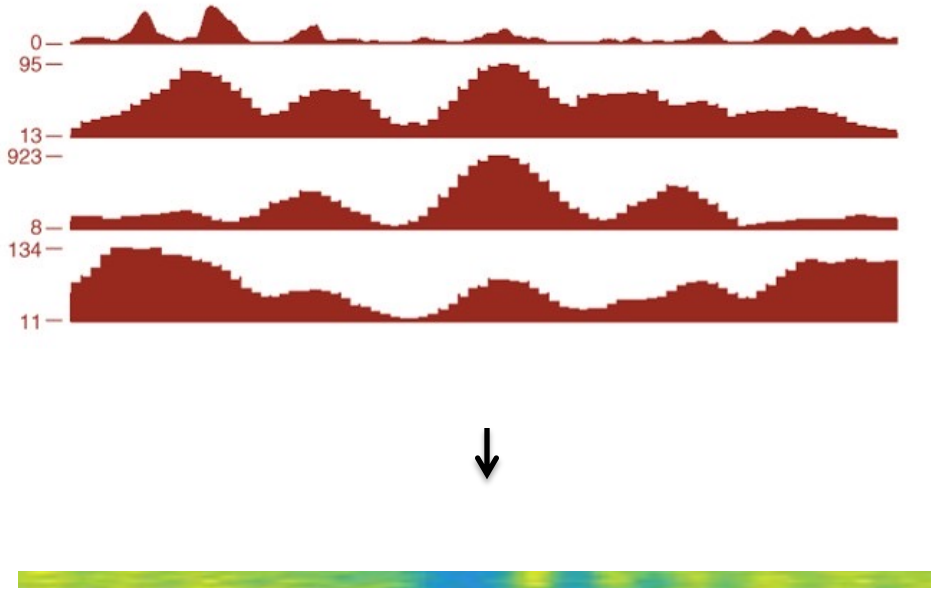
Buenrostro *et al*. *Nat Methods* 2013

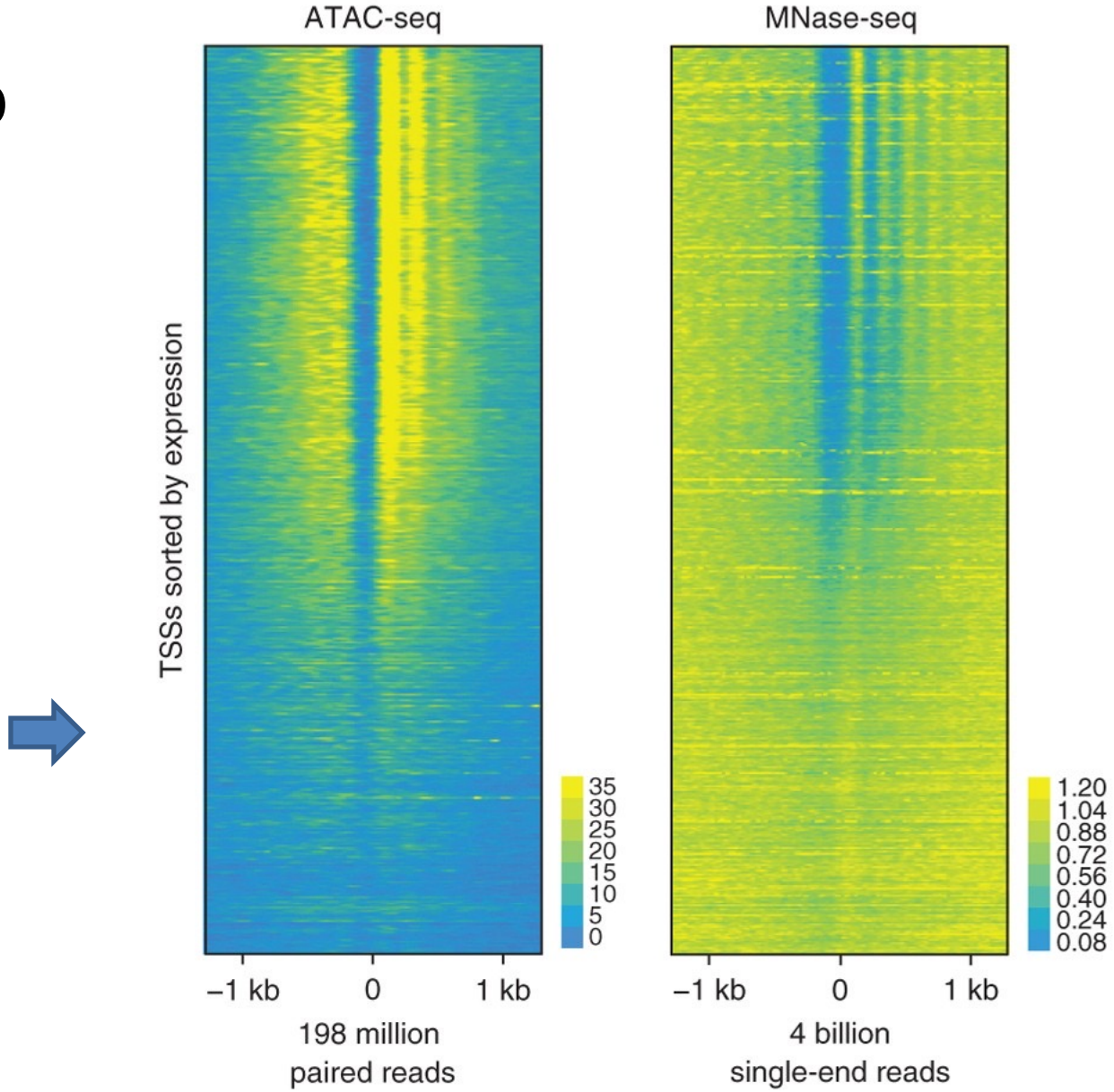
How to integrate patterns observed on signal tracks?



Adli *et al*. *Nat Methods* 2010

Stacked “ripple” heatmap

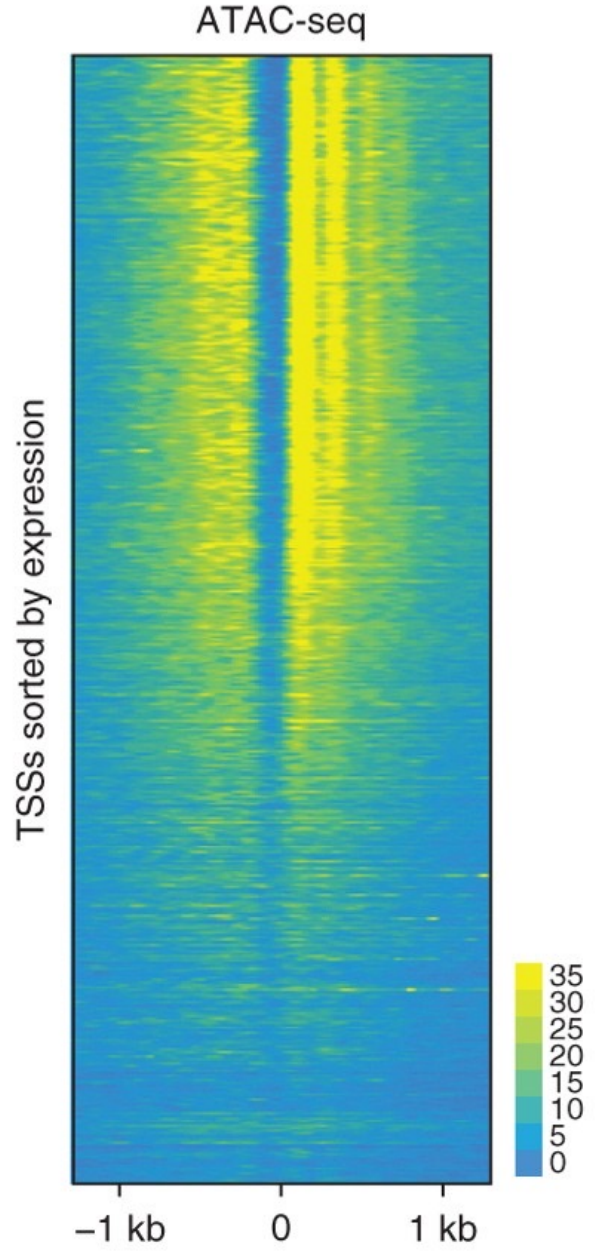




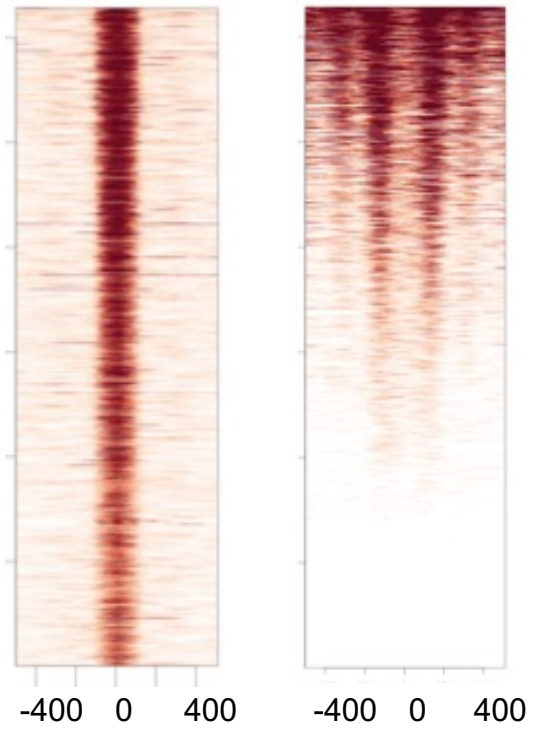
Buenrostro *et al*. *Nat Methods* 2013

Essential elements in a ripple heatmap

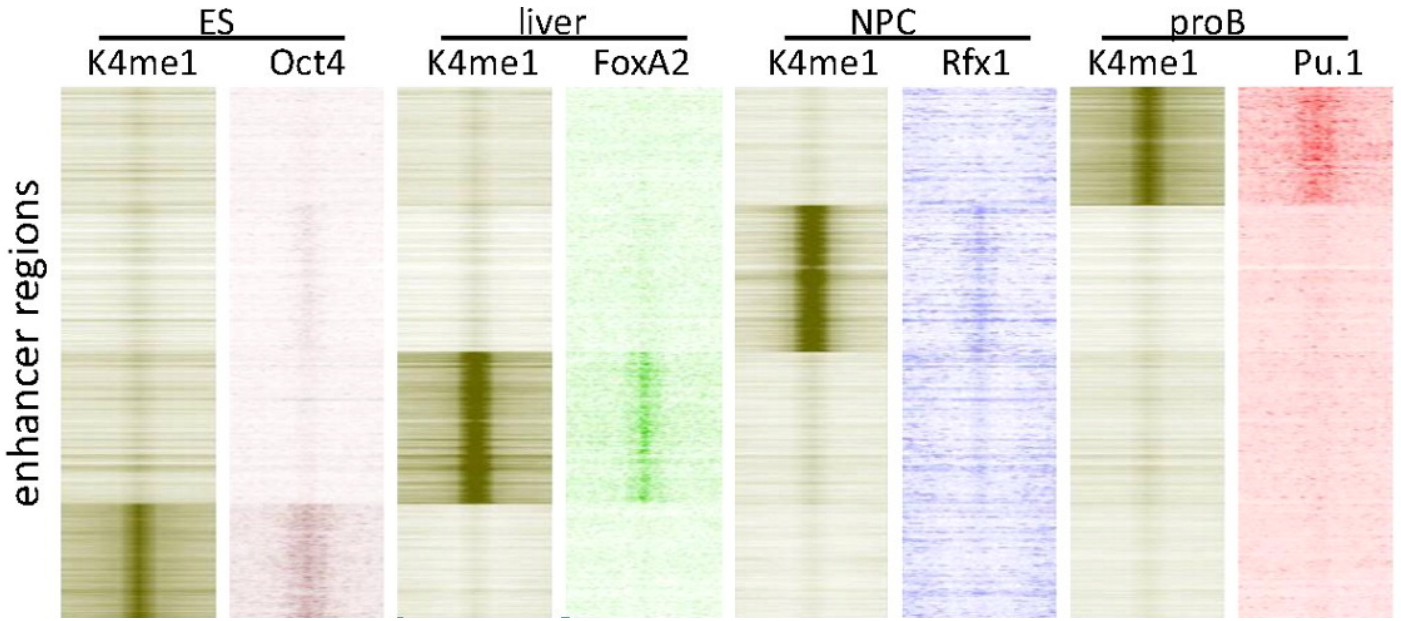
- Heatmap presents 3-dimensional data
- x: What loci/anchor is each row? Range?
- y: What are the rows? How many? How are they ranked?
- h: Data title/label (what signal?) color scale?



Multiple datasets visualization by ripple heatmap

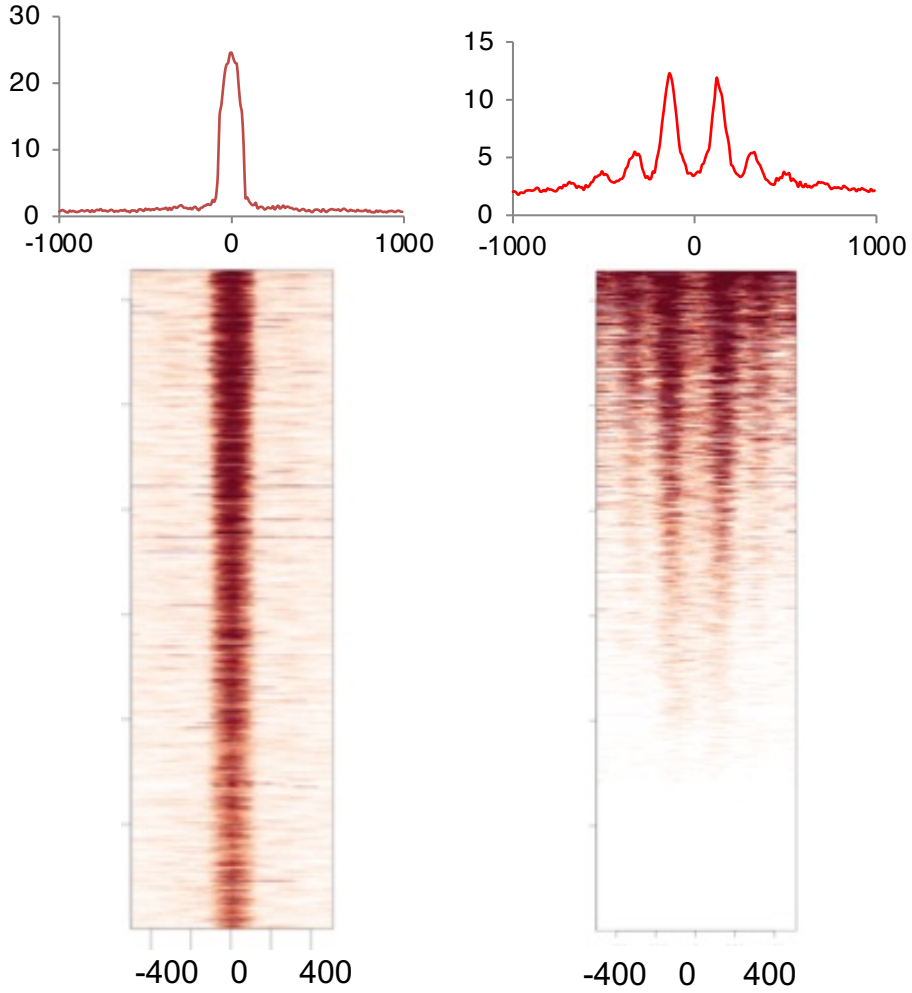


Luyten *et al. Genes Dev* 2014



Creyghton *et al. PNAS* 2010

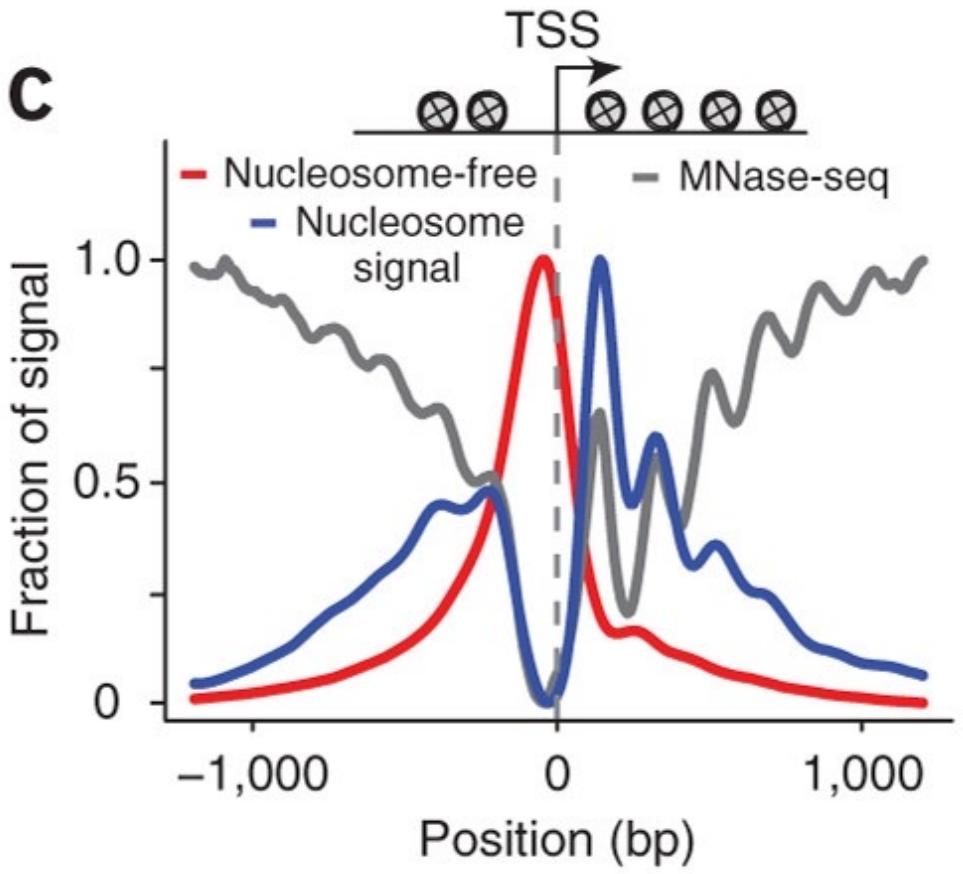
Composite curve plots



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



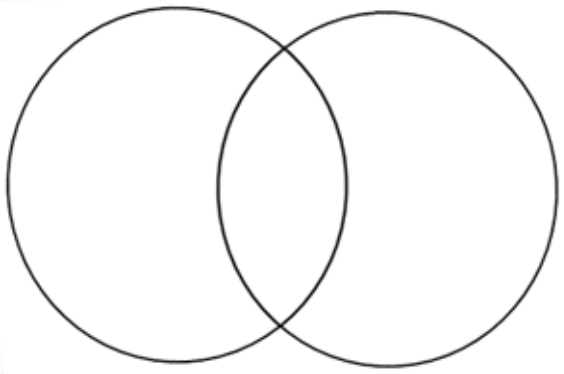
Buenrostro *et al*. *Nat Methods* 2013

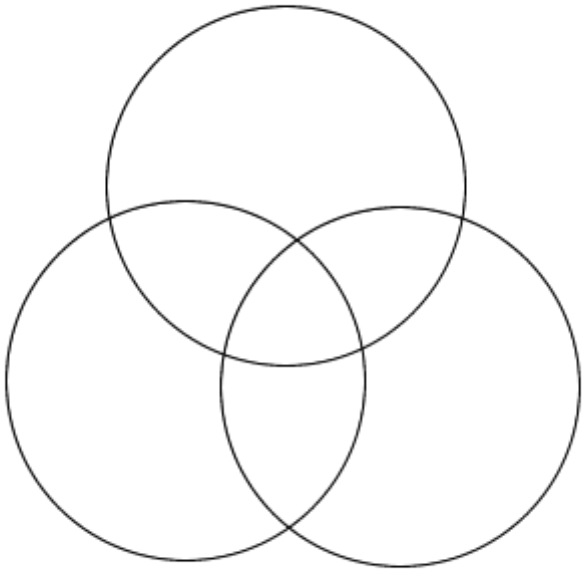
Common misinterpretations of composite plots

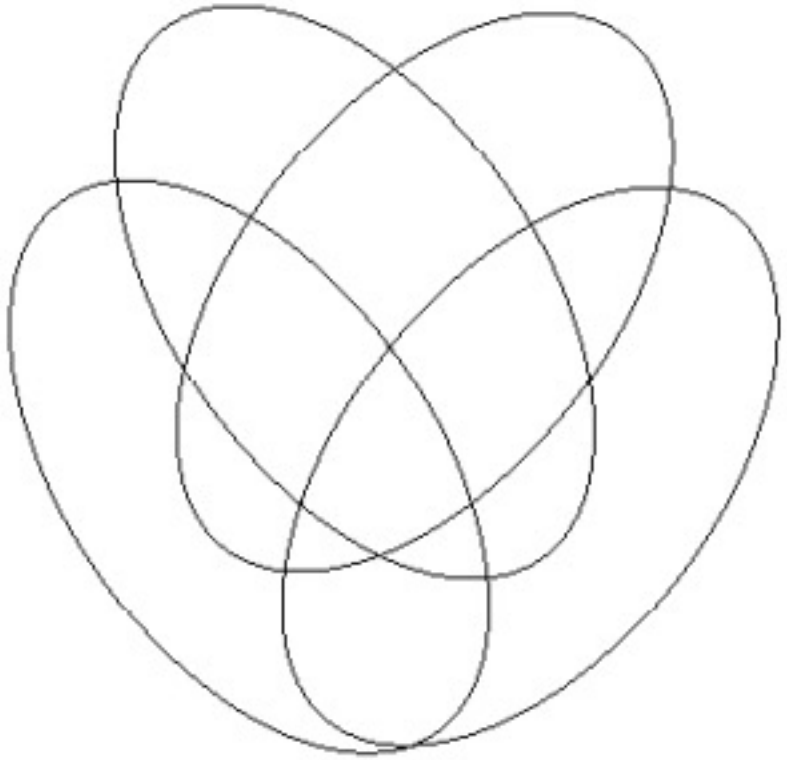
- Caveat 1: A peak in a composite plot may be contributed by only a tiny fraction of regions (not representative of the global picture)

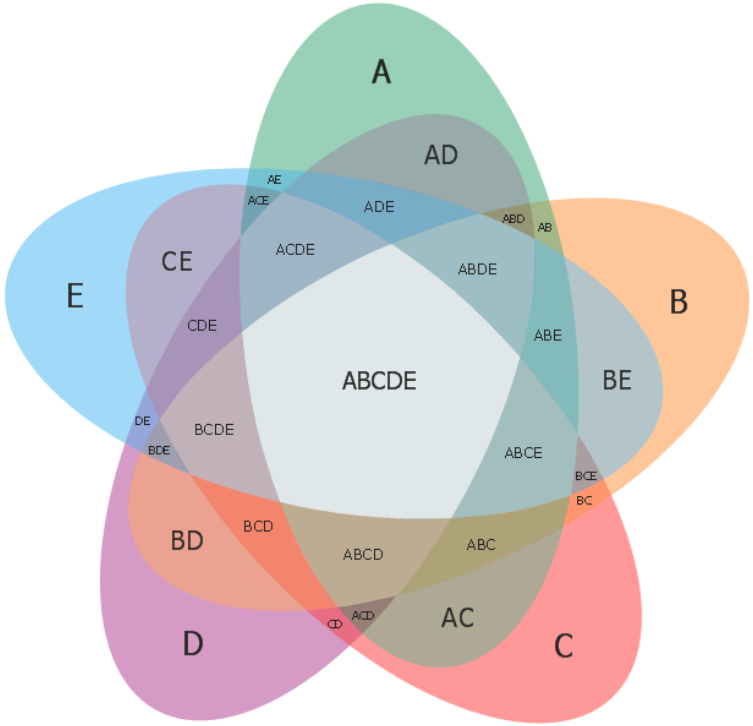
- Caveat 2: A higher peak does not necessarily mean stronger signal or more region coverage

Venn diagram presents yes/no relations between sets

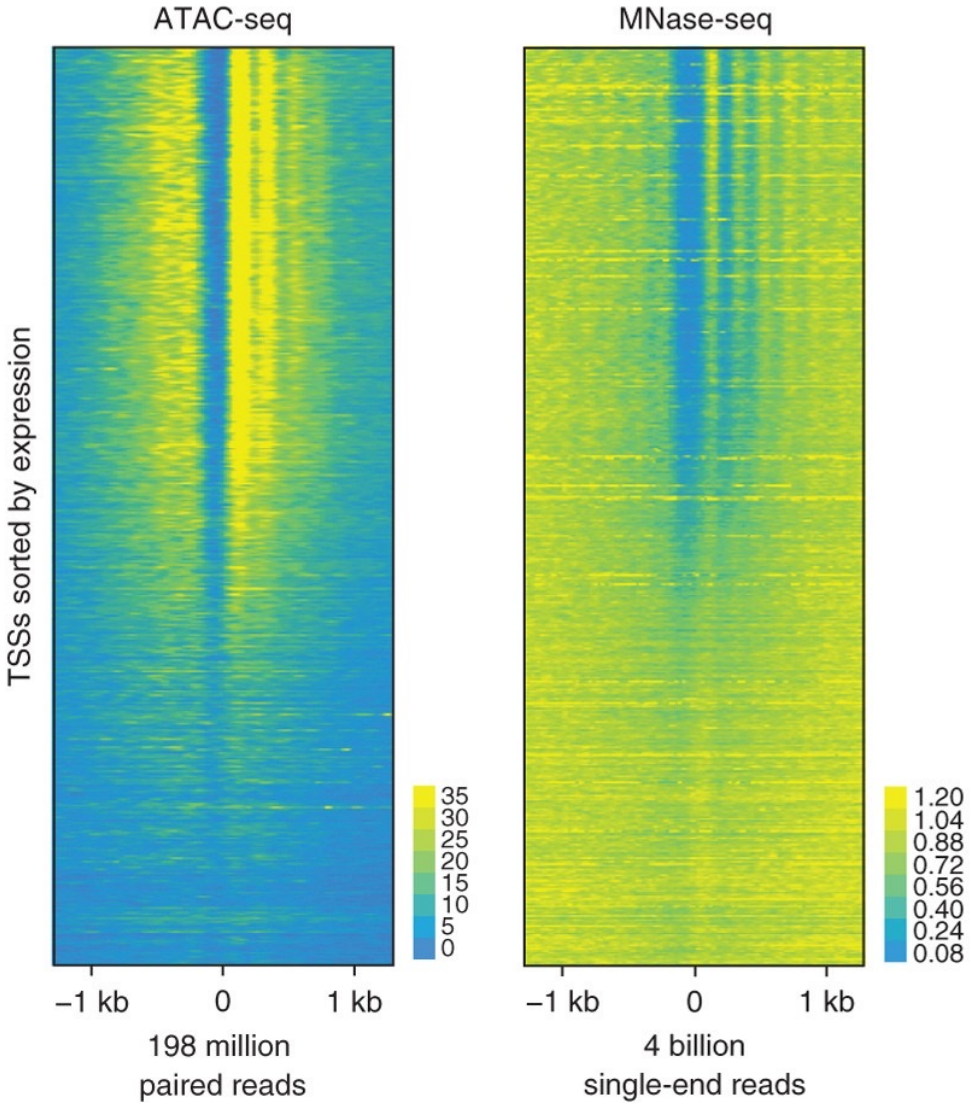


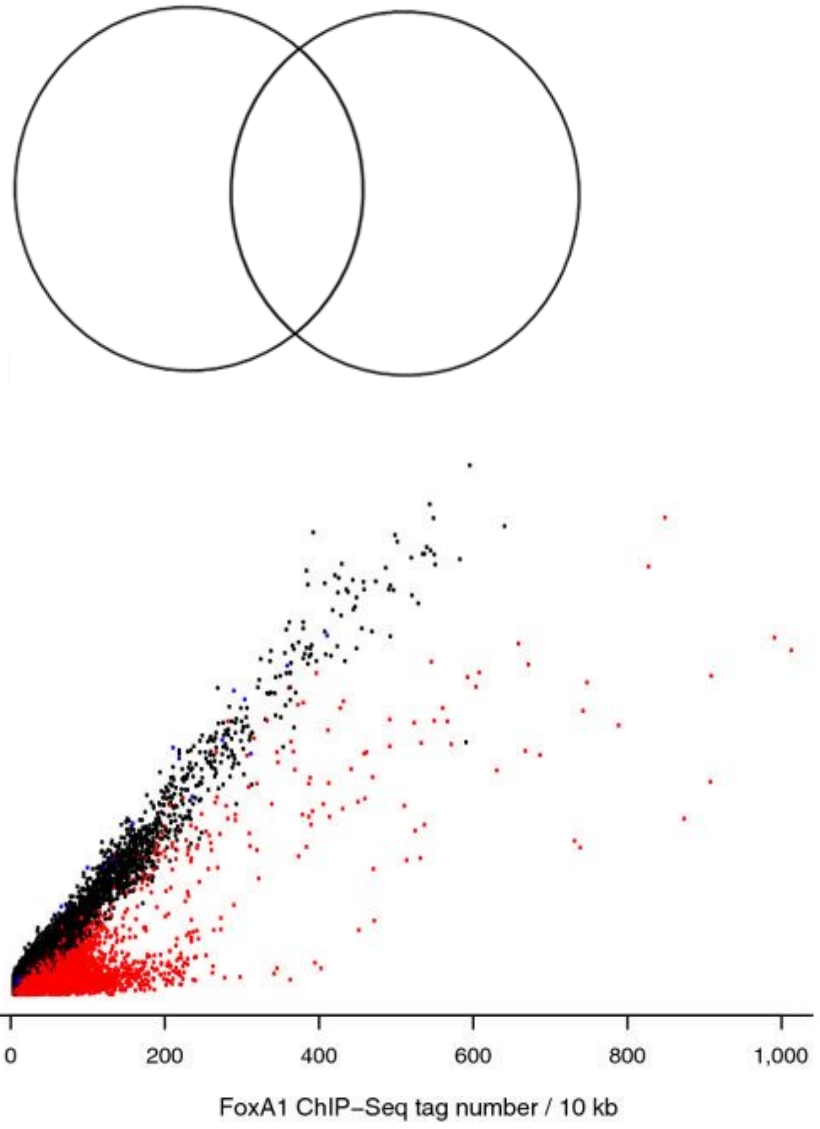






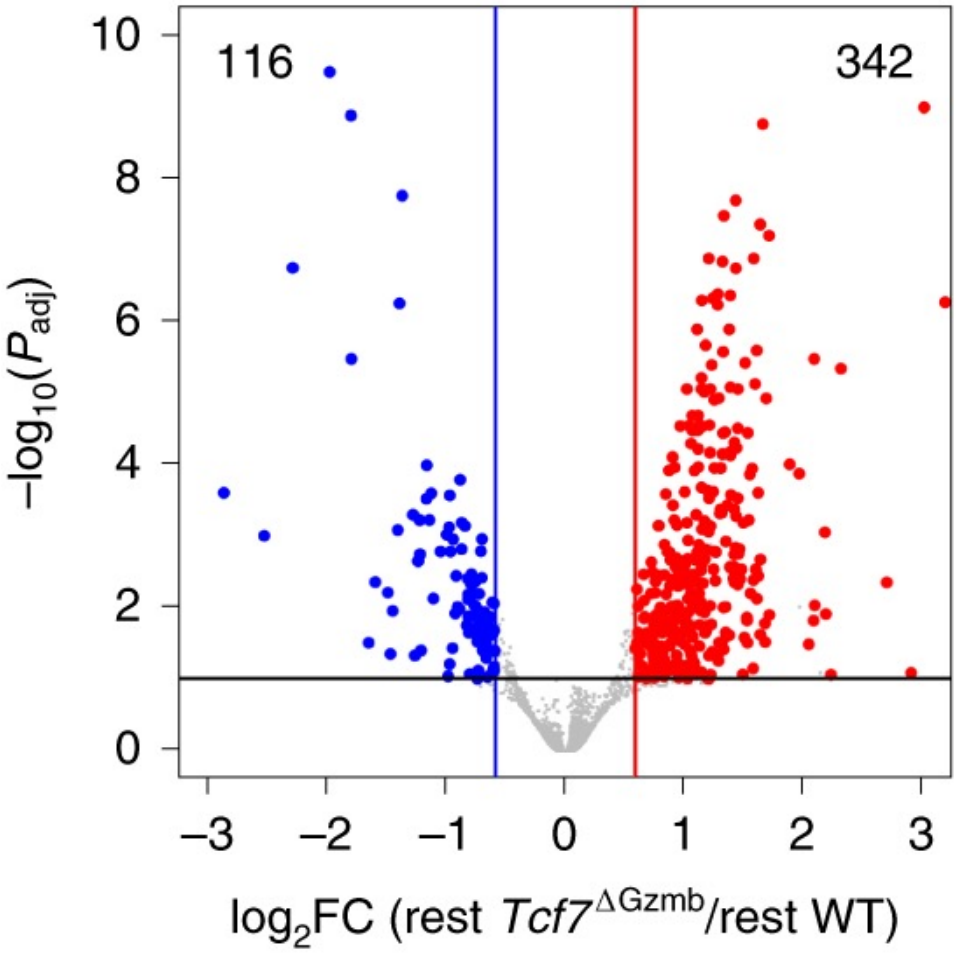
Heatmaps and scatter plots are more informative





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

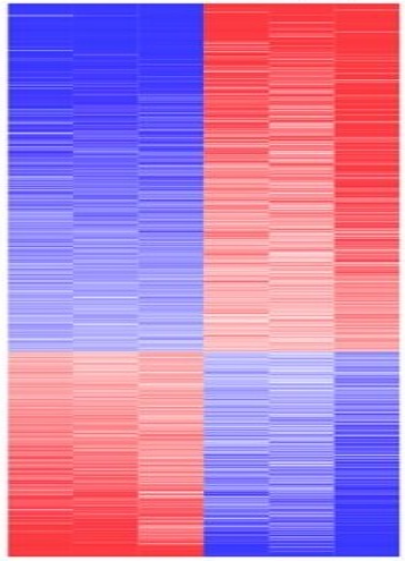


Shan et al. *Nature Immunology* 2022

Differential gene expression visualization by heatmap

Differential expression

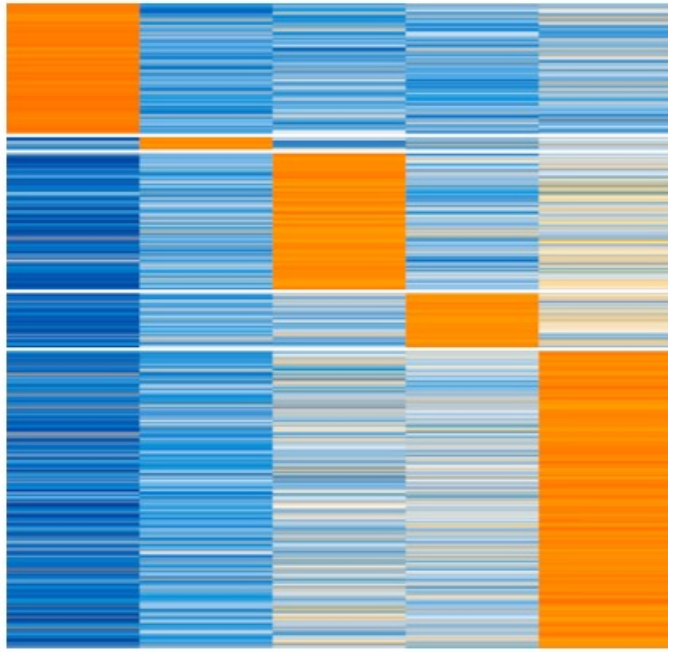
control treated

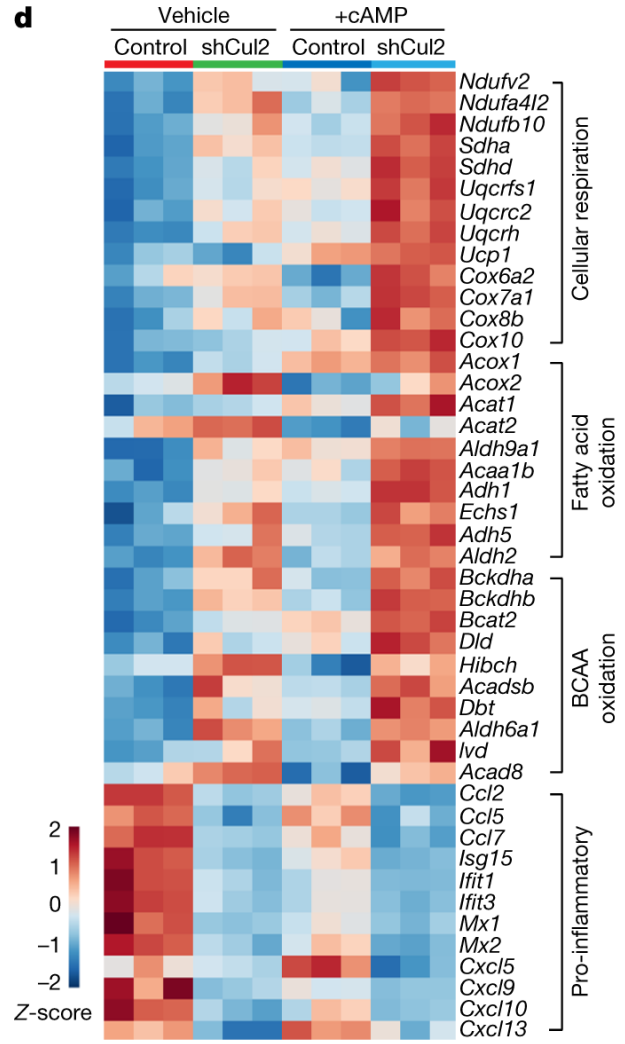


up-regulated

down-regulated

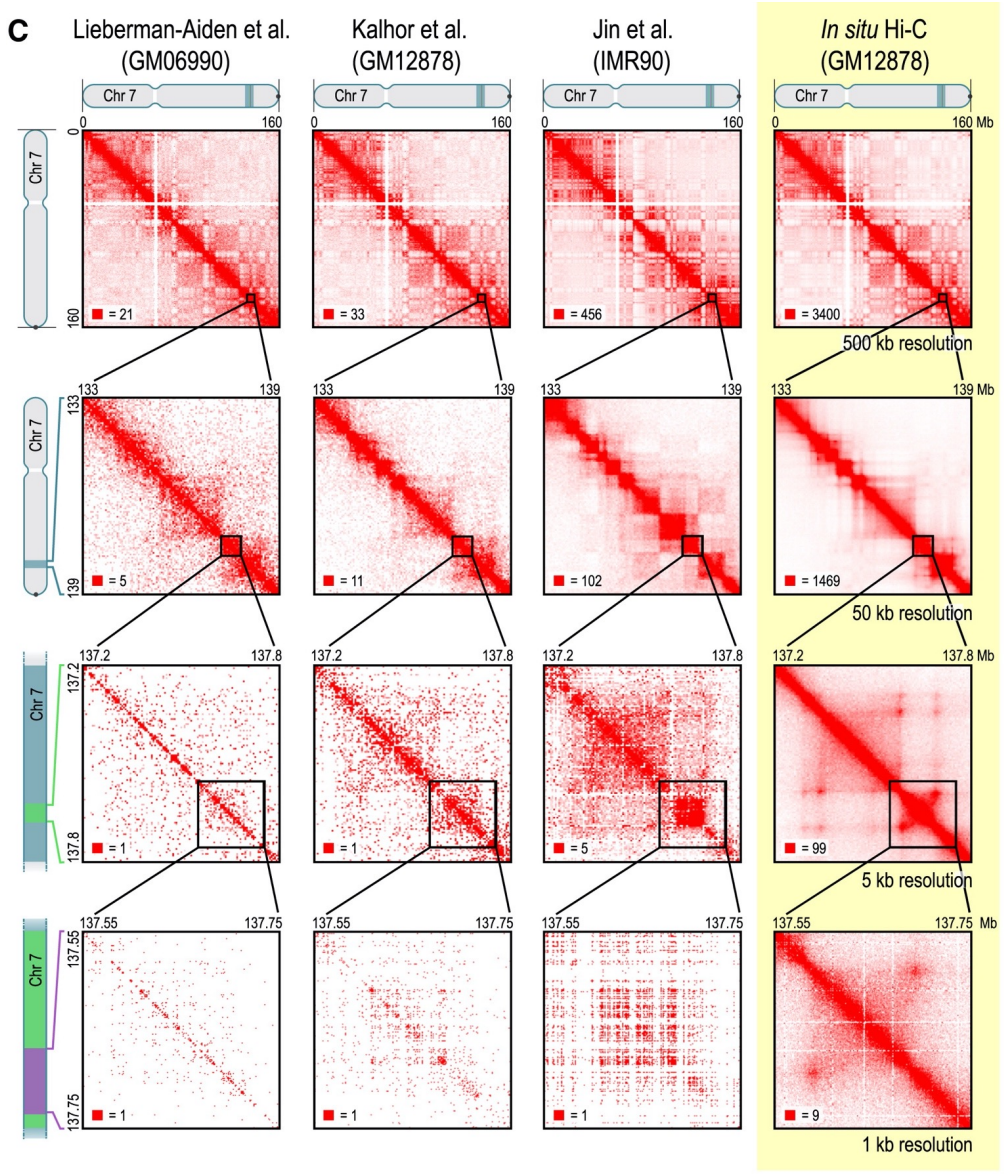
Co-expressed gene clusters





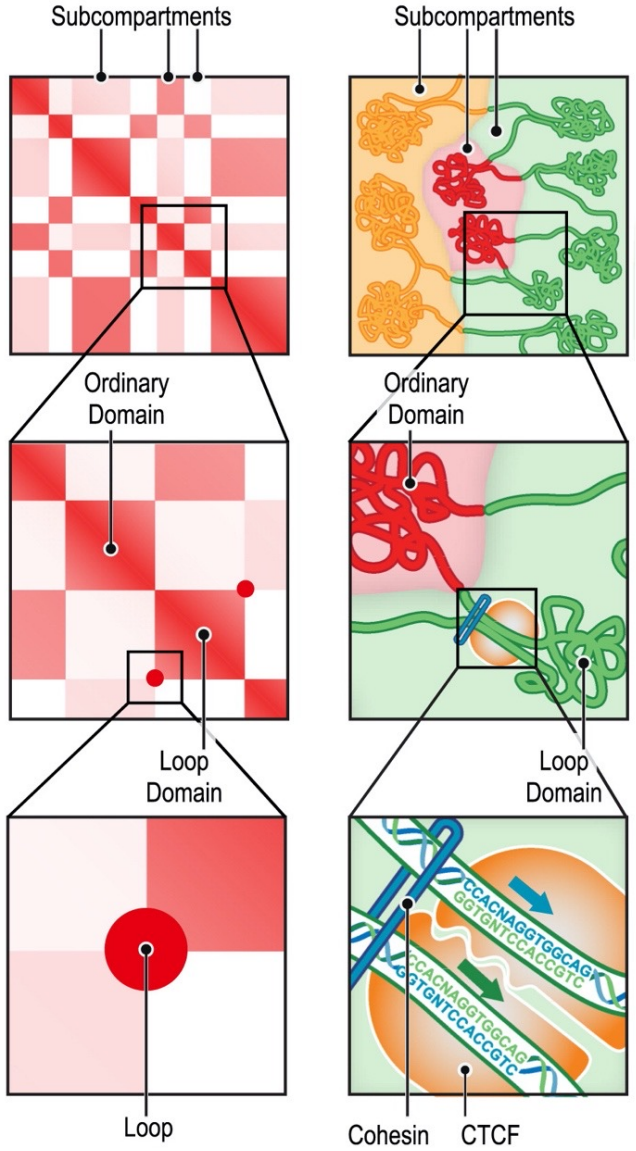
Wang *et al. Nature* 2022

Hi-C contact heatmap for 3D genome interactions



Rao *et al. Cell* 2014

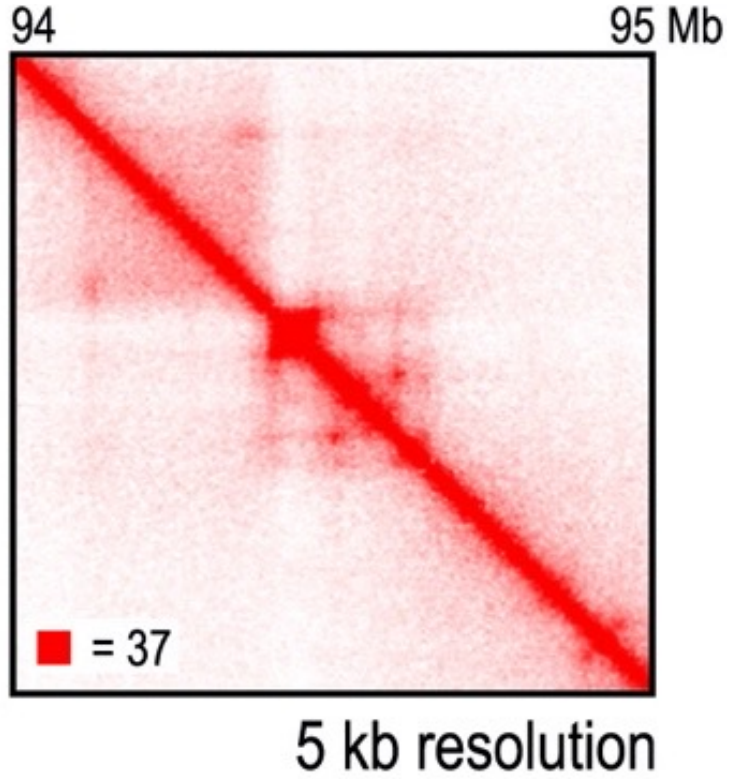
Hi-C contact heatmap for 3D genome interactions

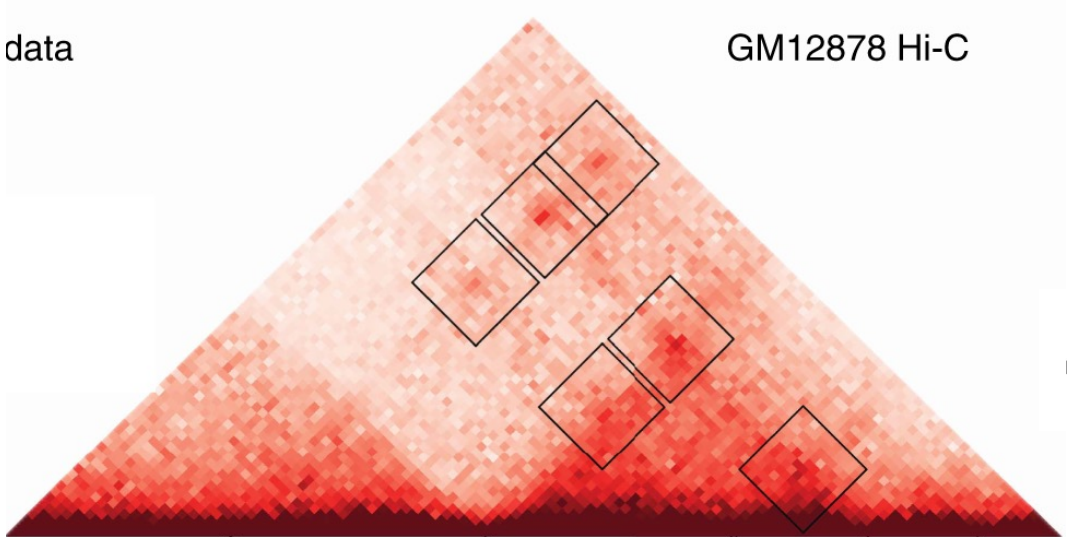


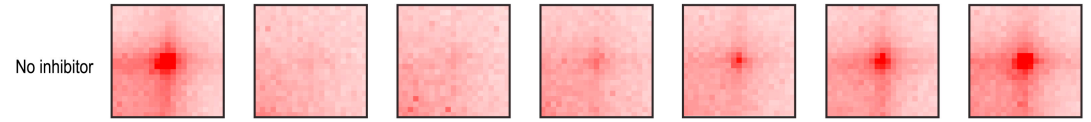
Rao *et al. Cell* 2014

Essential elements in a Hi-C contact heatmap

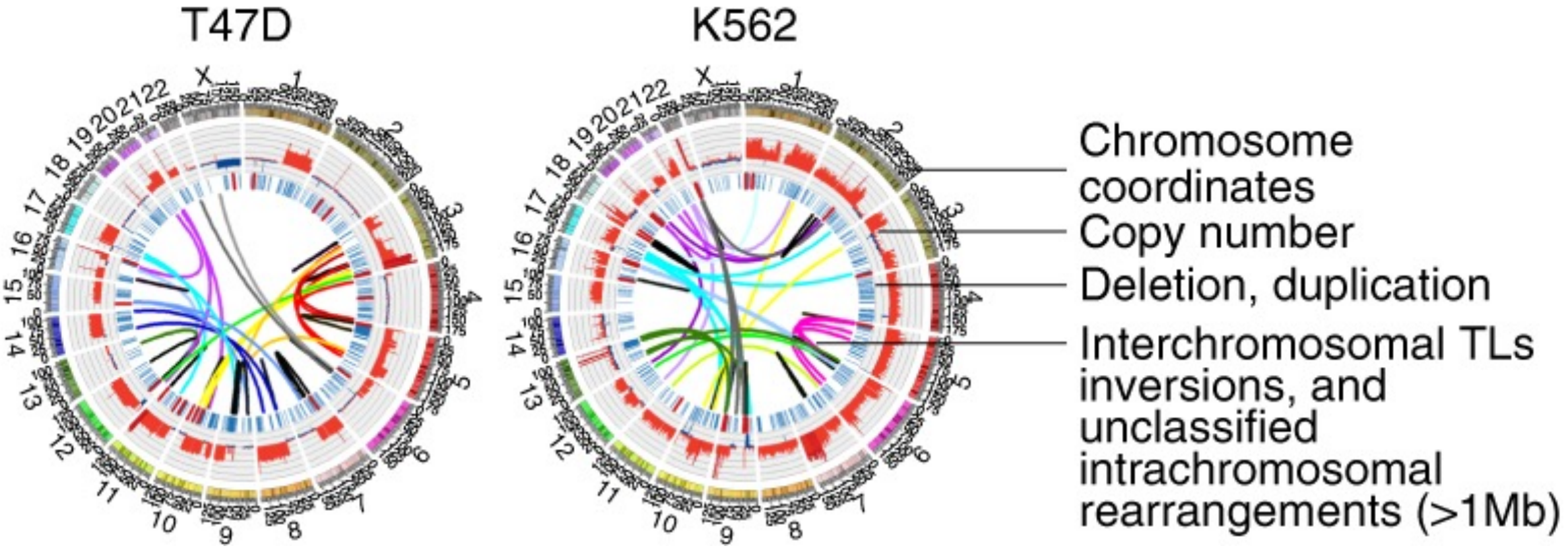
- Scale, scale, scale
- Resolution
- Normalization
- Blocks, stripes, loops (2d peaks)





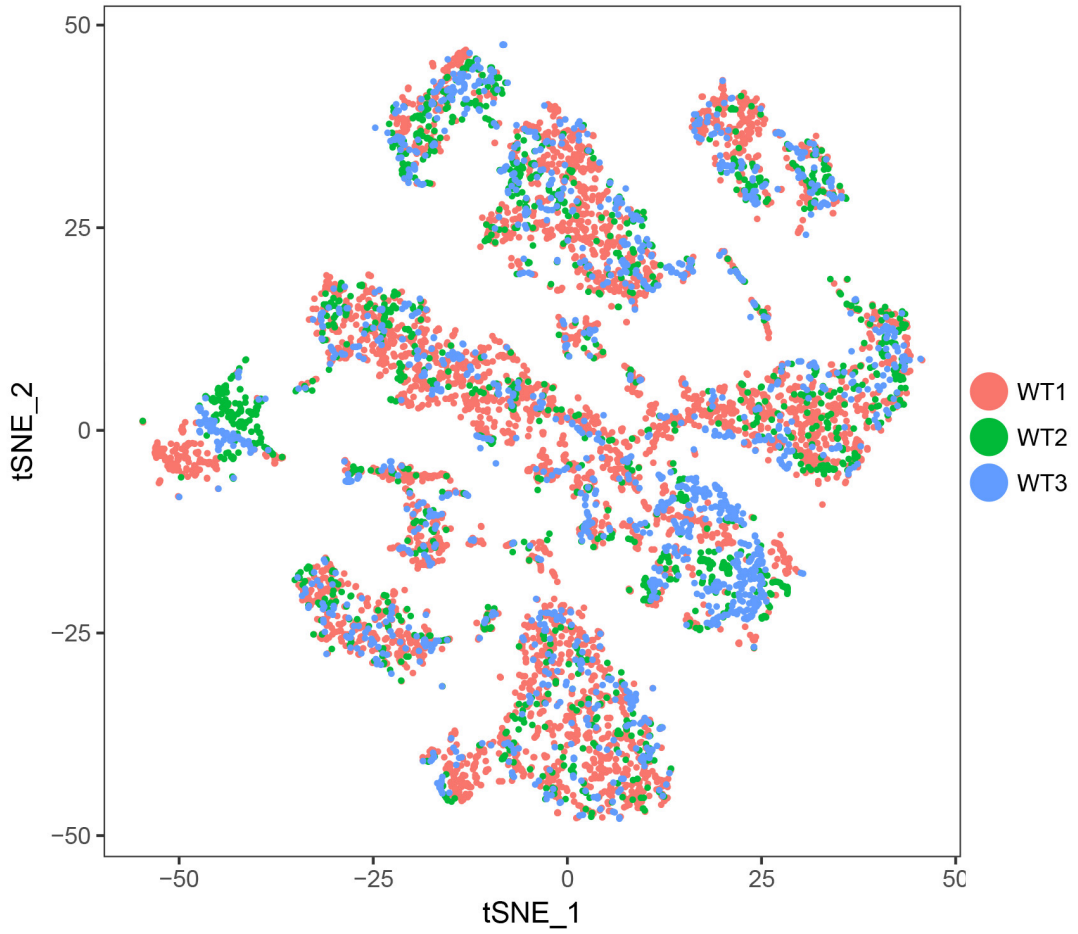


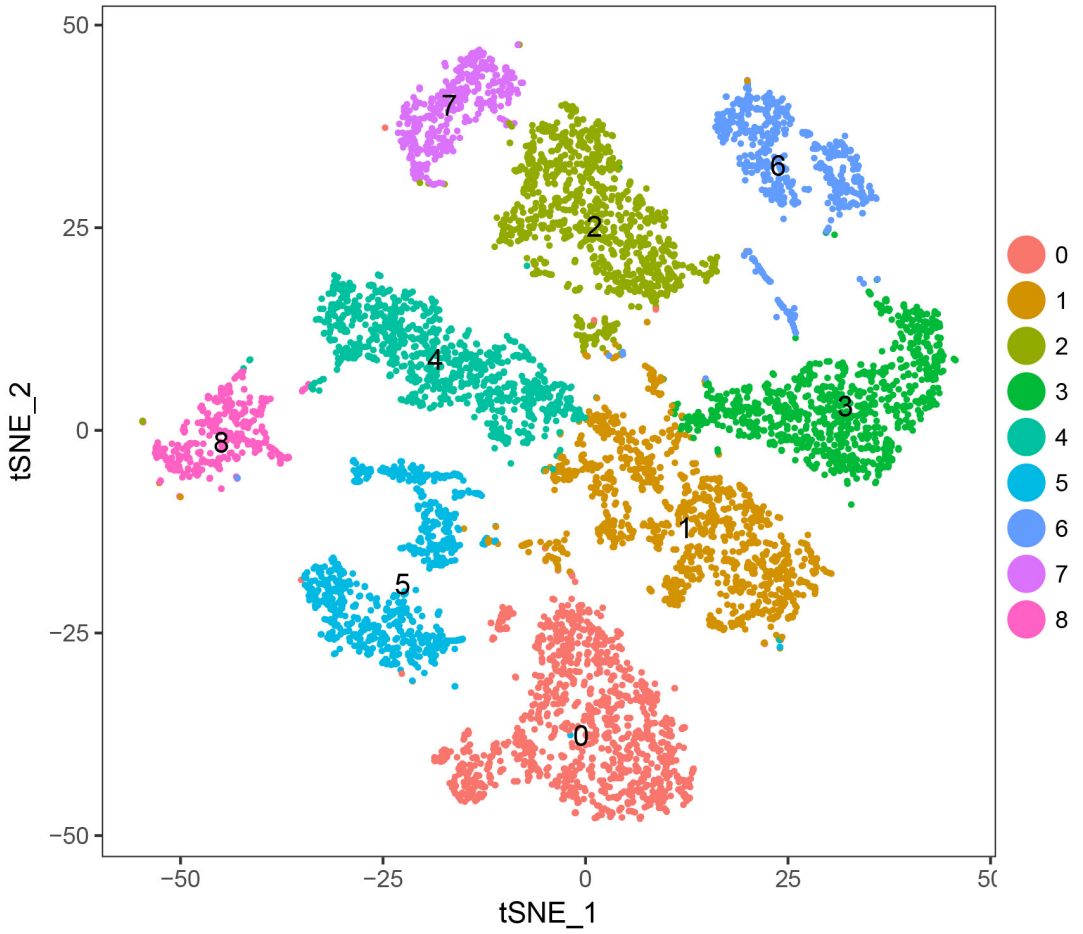
Circos plot integrates multiple types of genomics data



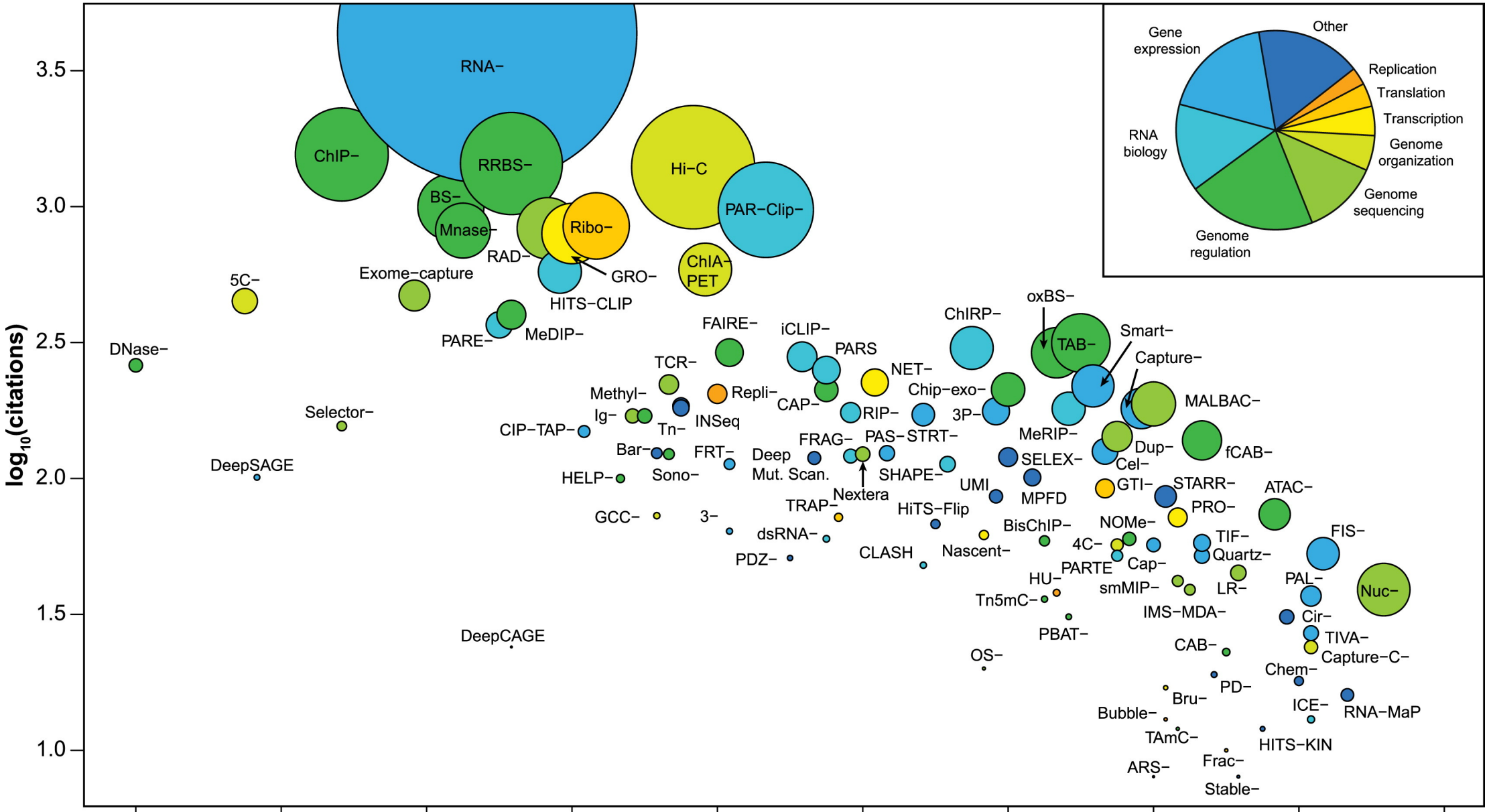
Dixon *et al. Nat Genet* 2018

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





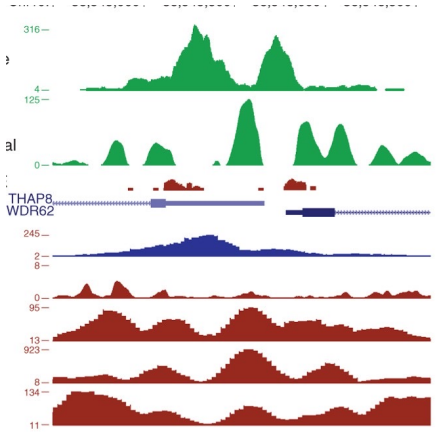
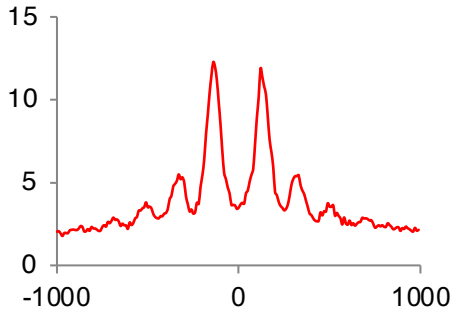
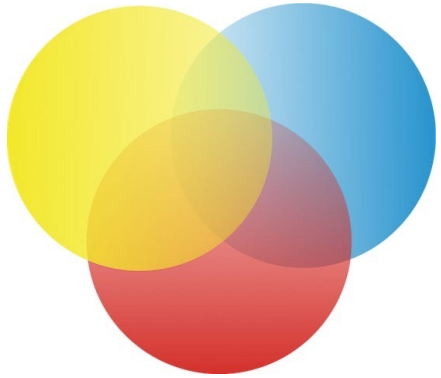
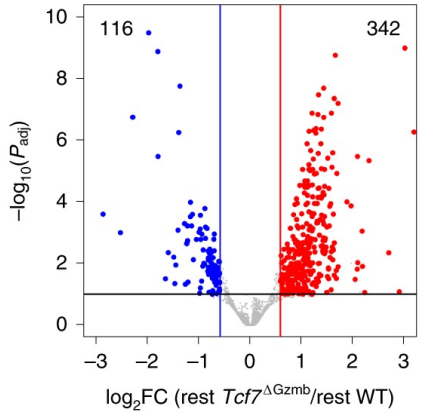
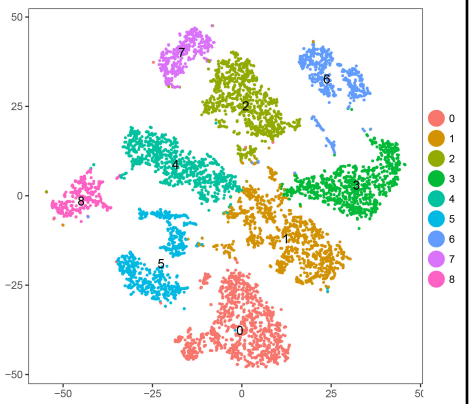
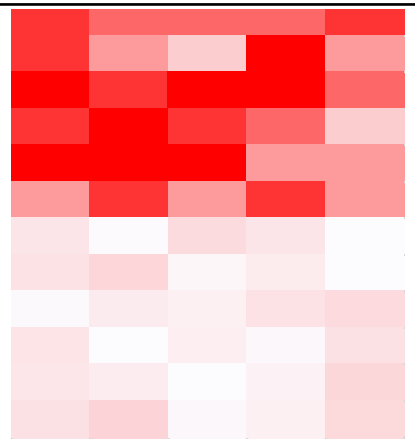
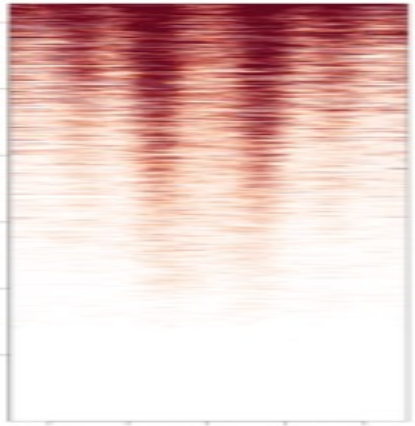
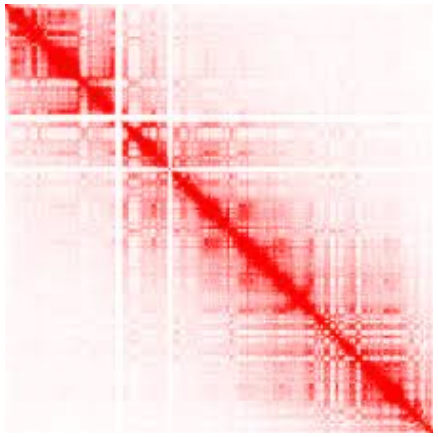
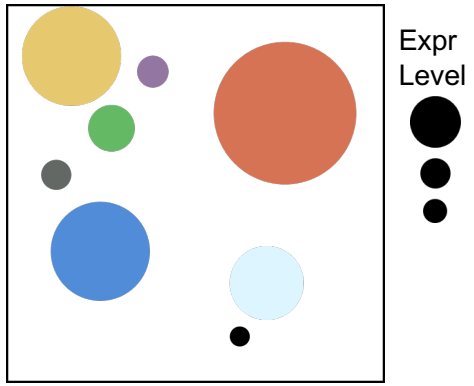
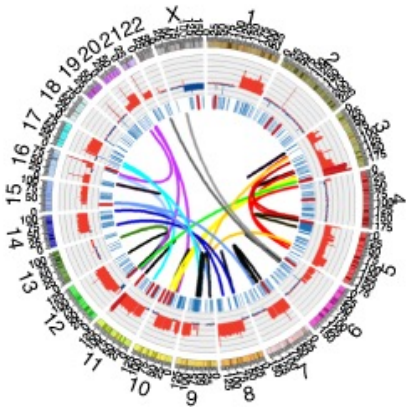
Bubble plots: NGS-based applications (-seq)



Reuter et al. Mol Cell 2015

https://www.youtube.com/watch?v=jbkSRLYSojo

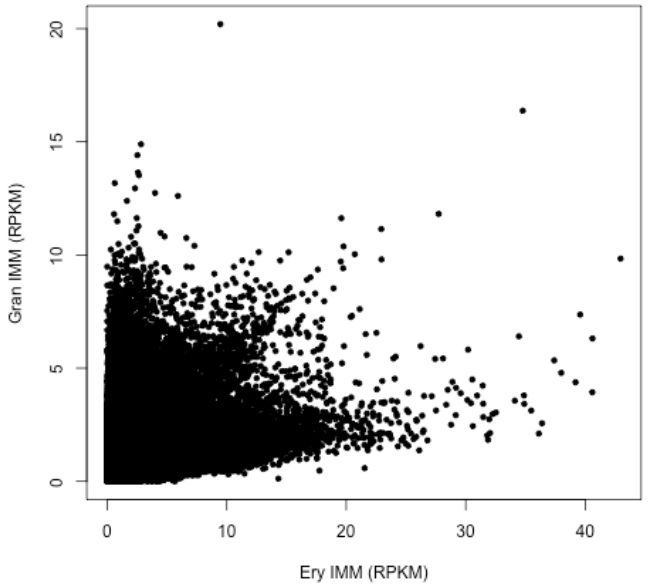
Summary

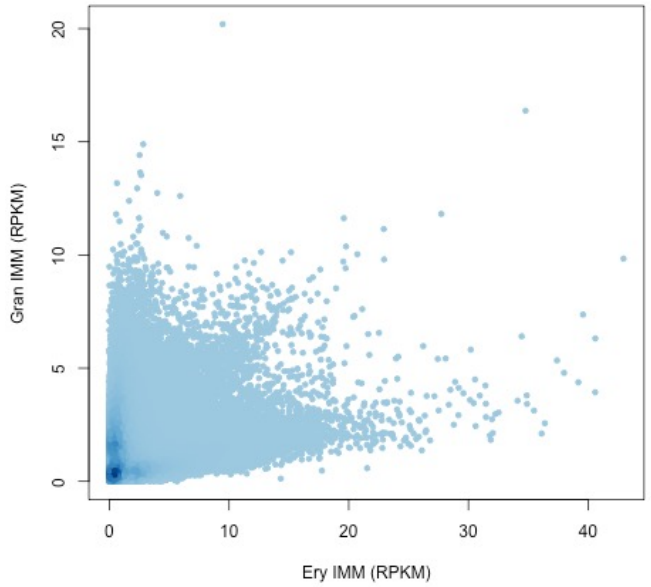
Genome browser track	Composite curve	Venn diagram	Volcano plot	tSNE/UMAP scatter plot
				
Heatmap	**Ripple heatmap**	**Hi-C contact map**	**Bubble plot**	**Circos plot**
				

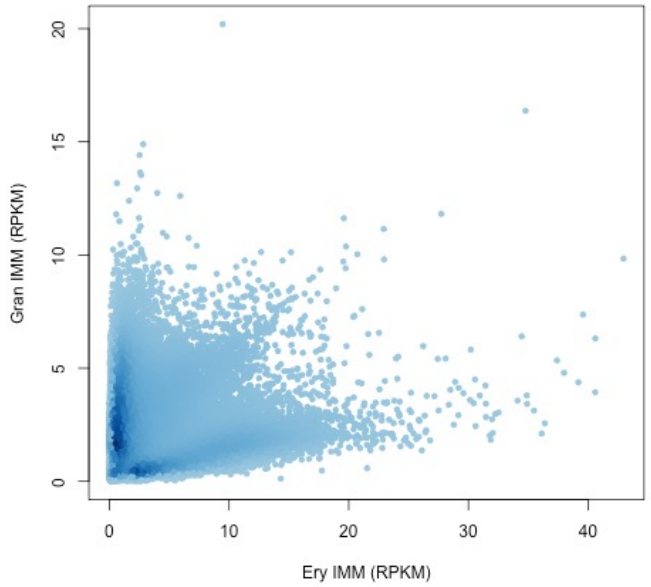
Some tips

Scatter plot

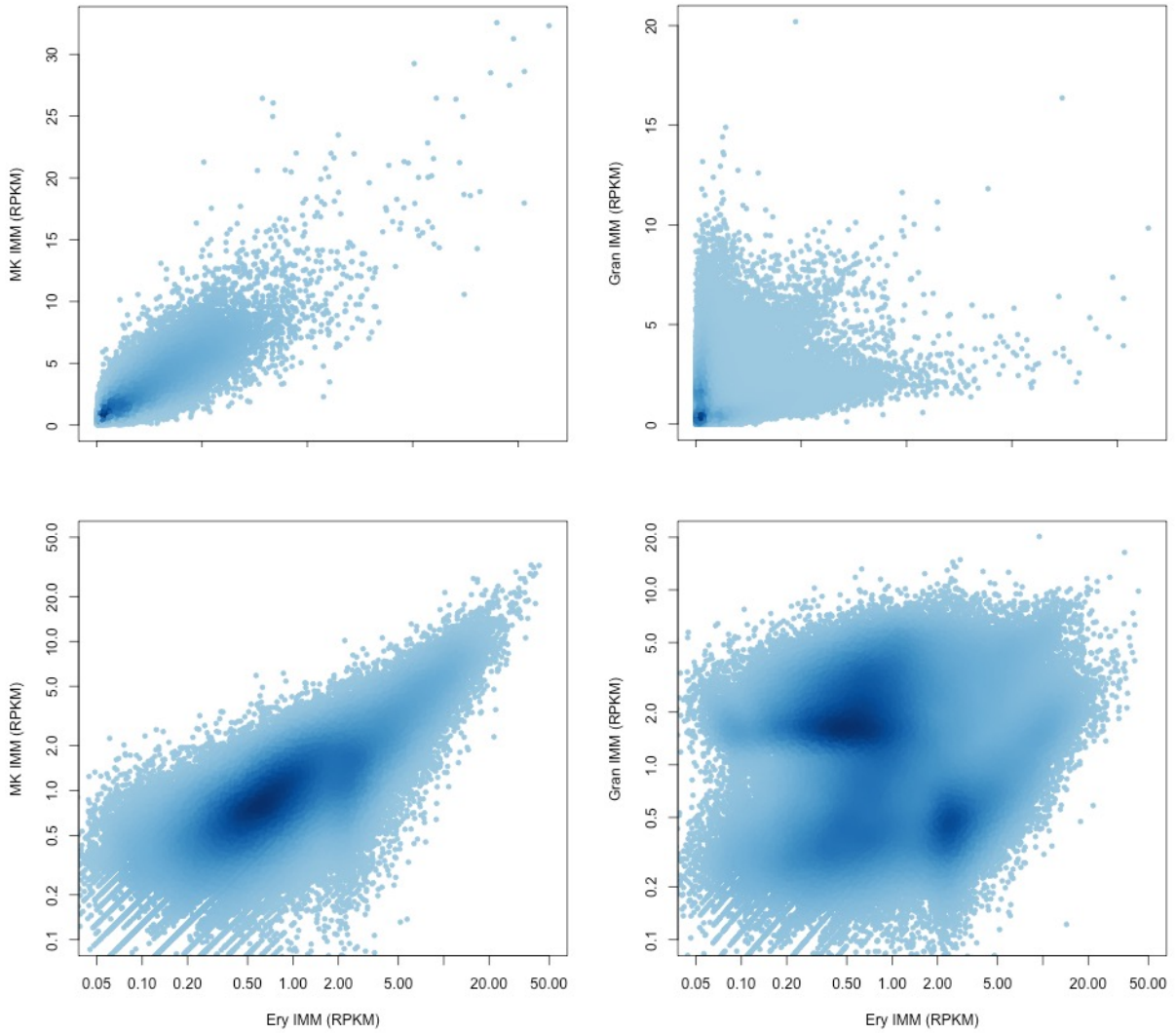
- The density of the data points matters!



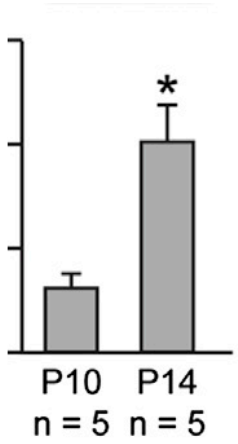


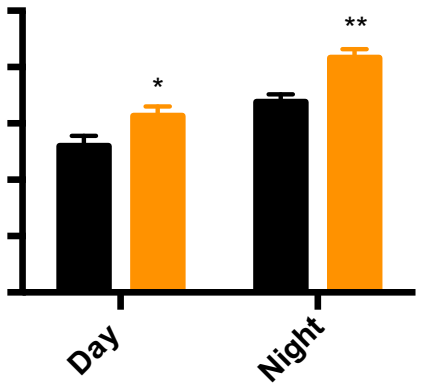


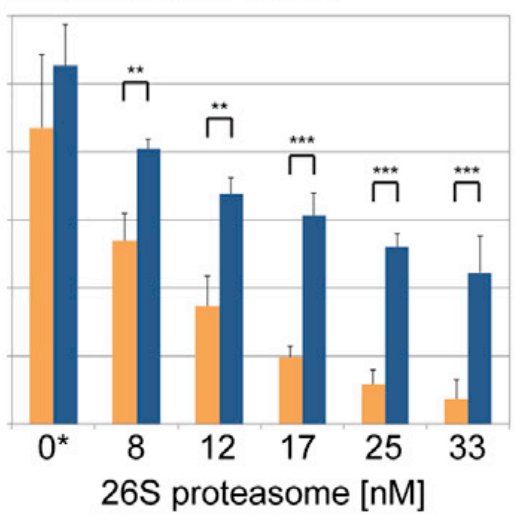
Choose the appropriate scale for plotting (linear or logarithm)



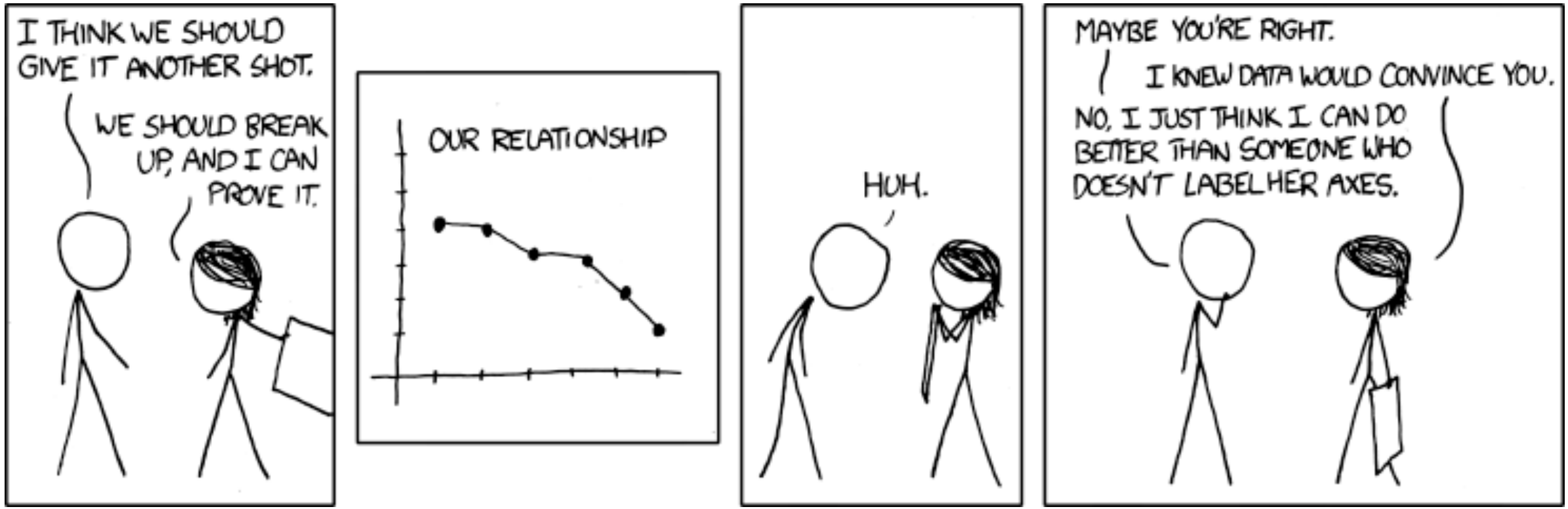
Bar chart?



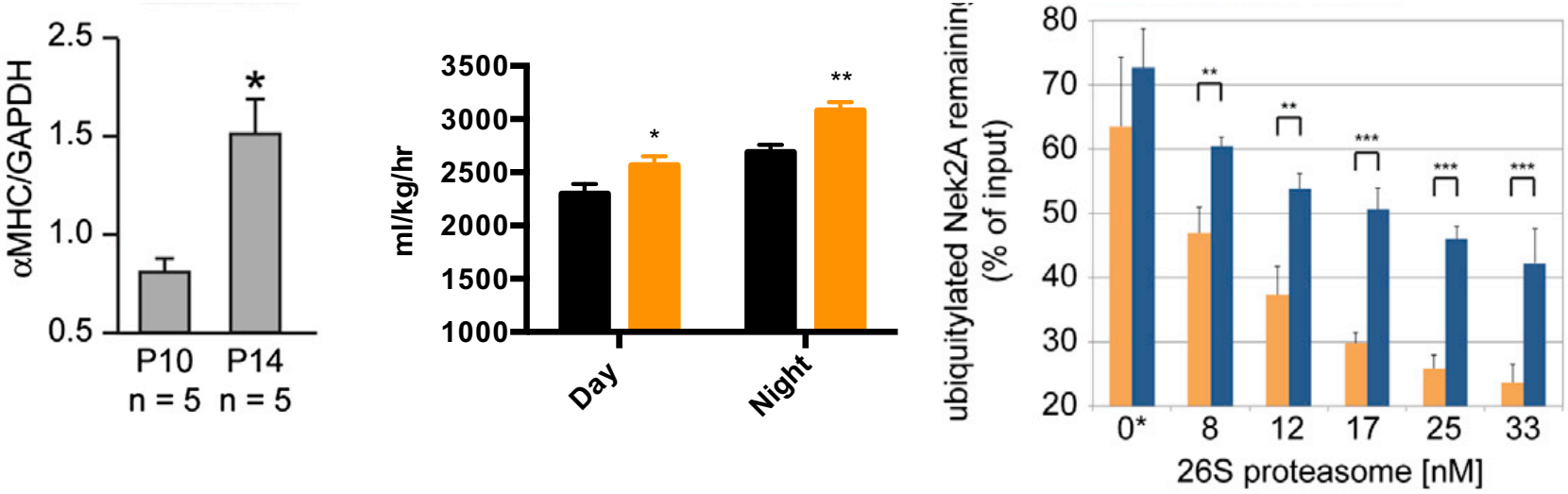




Plots from *Cell* 2014


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.

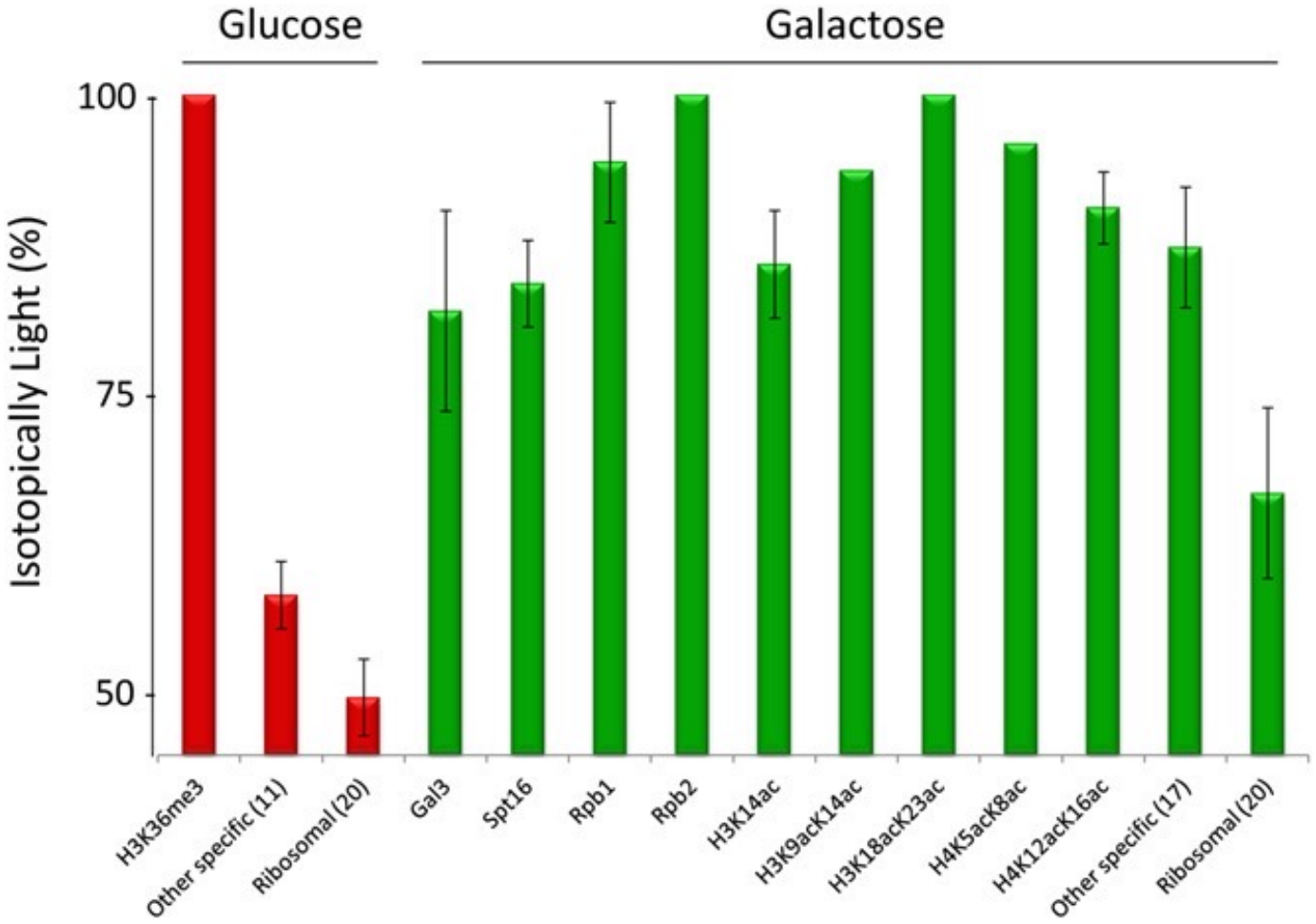
Bar chart



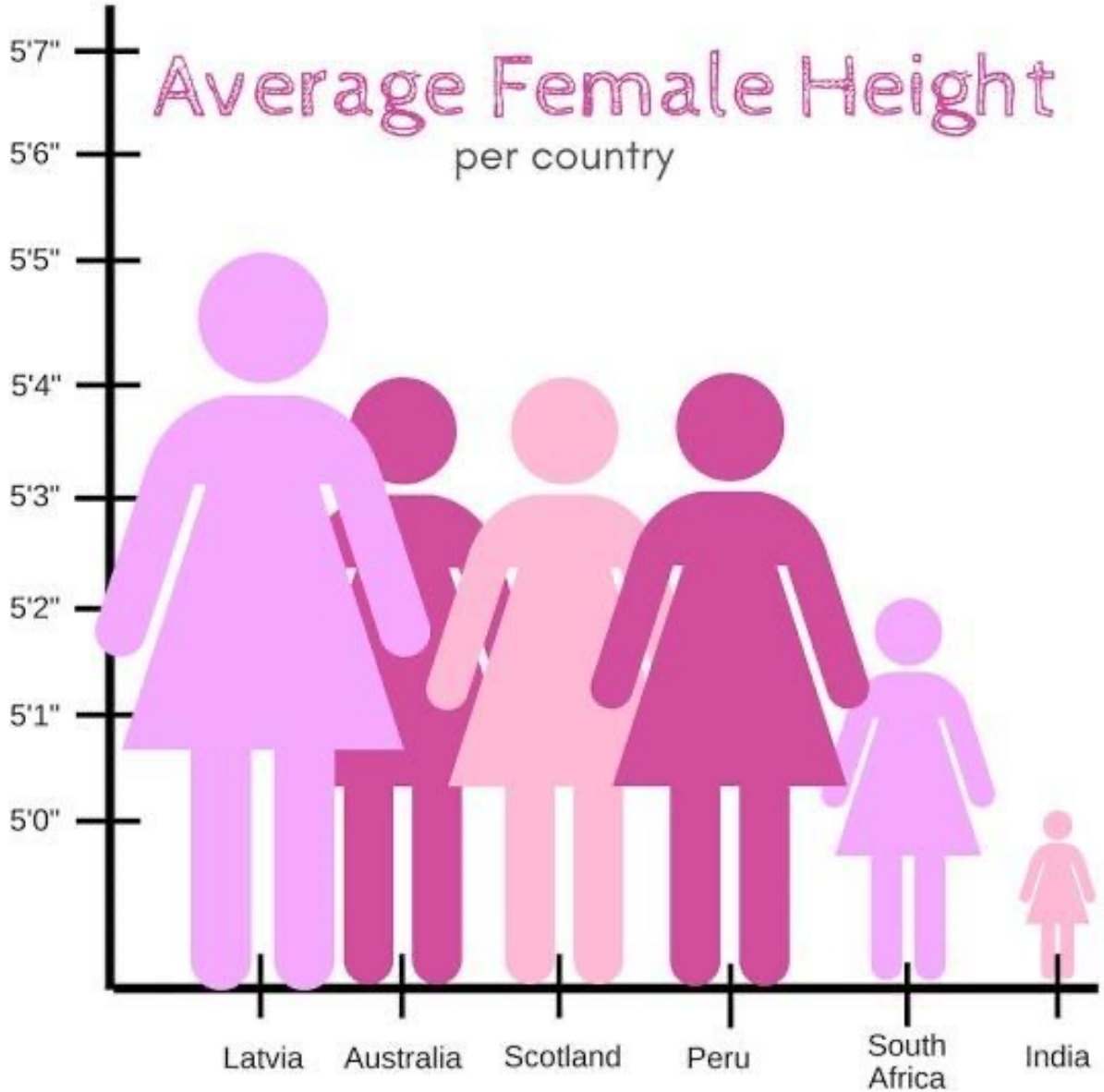
Bar charts should always start from 0, and on the linear scale.
If difference is small, box plots or original dots are better.

Plots from *Cell* 2014

Another example

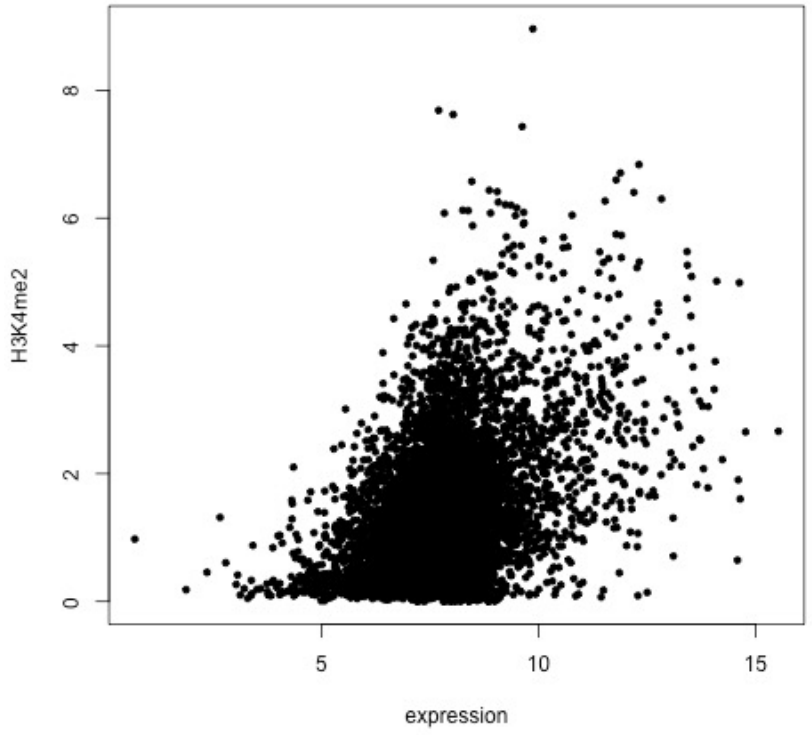


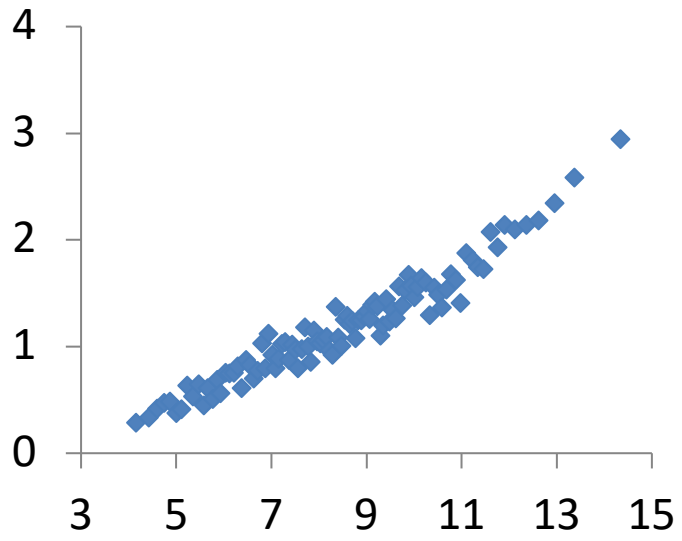
Byrum et al. *Cell Reports* 2012



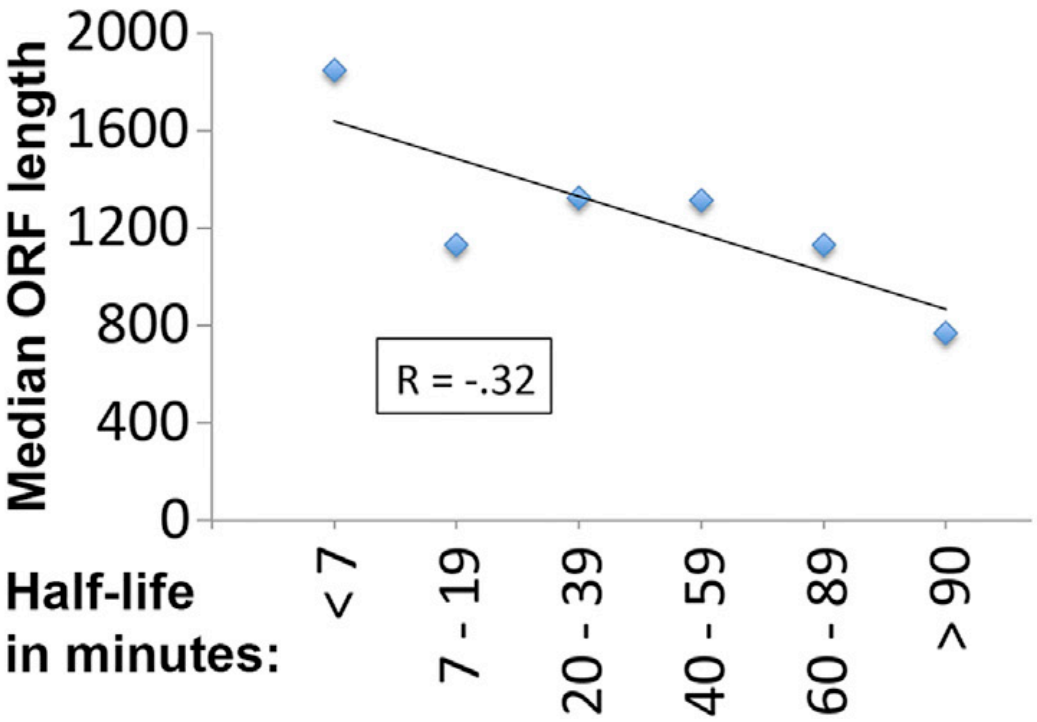
https://venngage.com/blog/bad-infographics/

Scatter plot? Group them if needed

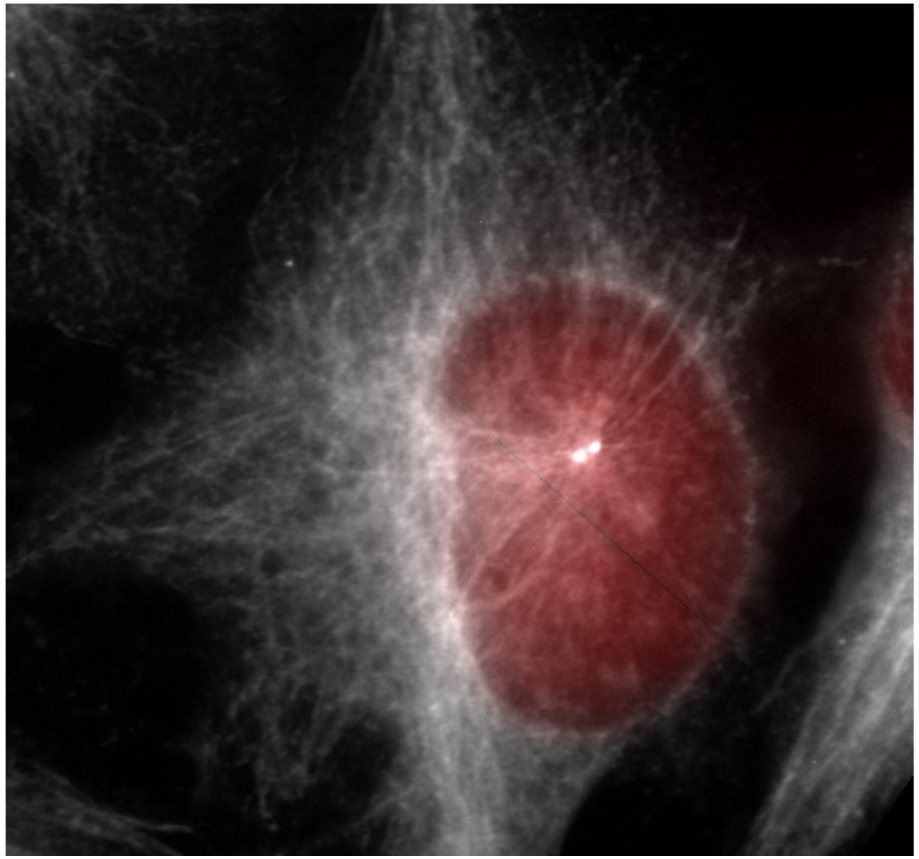




However, grouping should be even

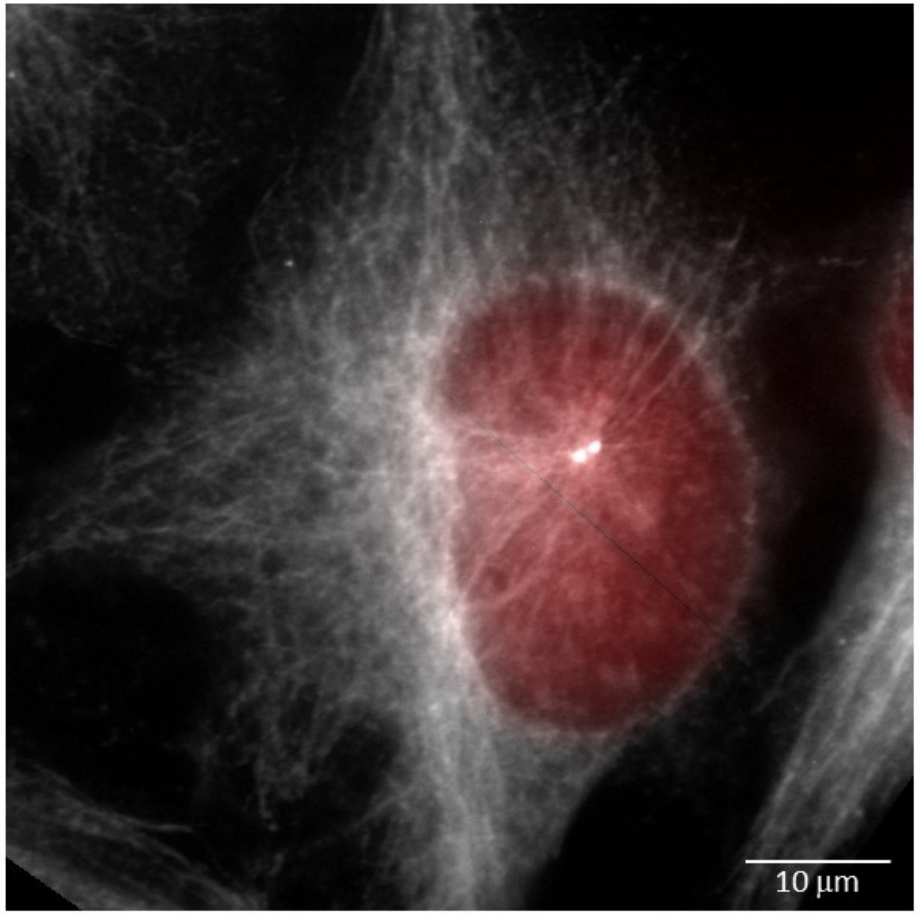


Cell 2014



https://twitter.com/HarenLaurence

Scale bar matters!

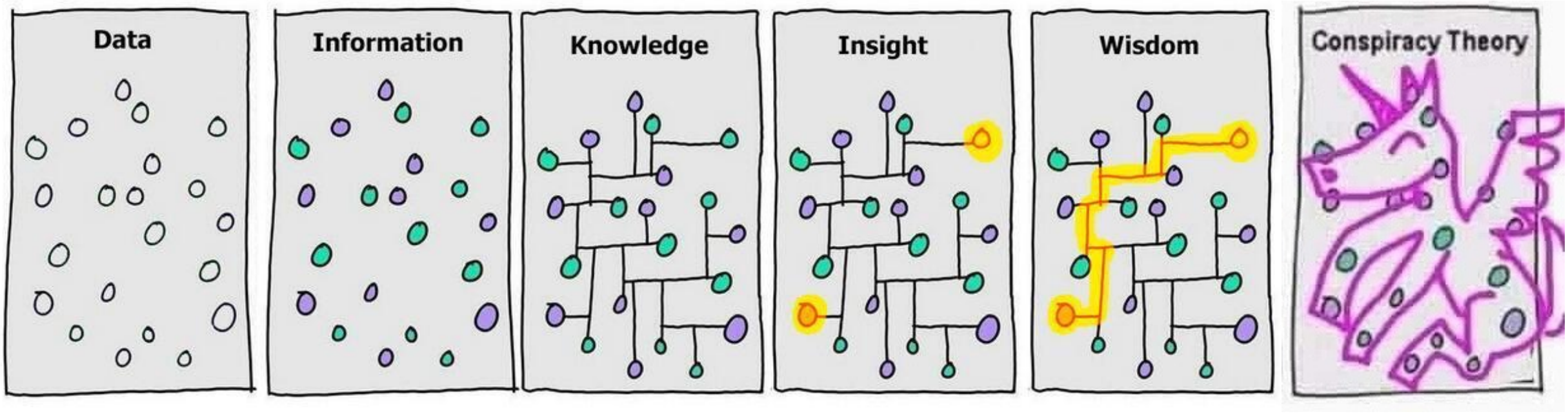


https://twitter.com/HarenLaurence

Take-Home Messages

- 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 even 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?

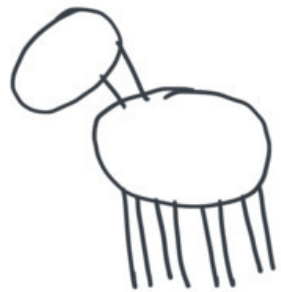
New biology!

Overfitting, overinterpretation...

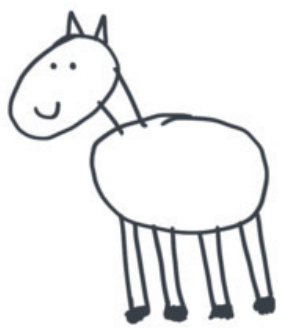
HOW TO:
DRAW A HORSE

BY VAN OKTOP

(1) DRAW 2 CIRCLES

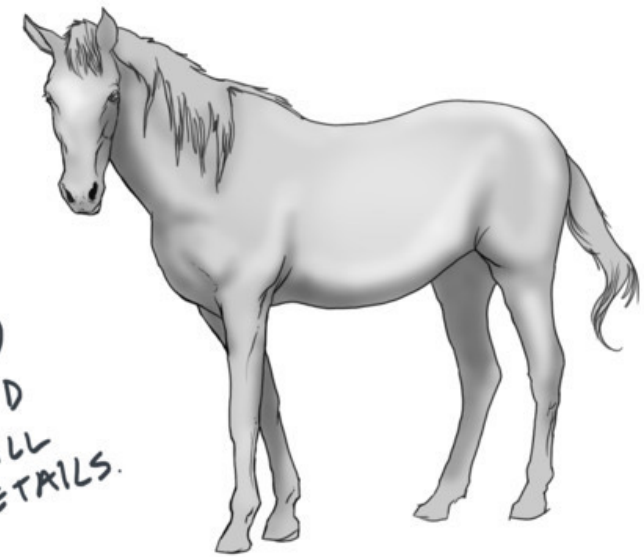


(2) DRAW THE LEGS



(3) DRAW THE FACE

(4) DRAW THE HAIR



(5) ADD SMALL DETAILS.

Thank you very much!

zang@virginia.edu

zanglab.org

zanglab.org