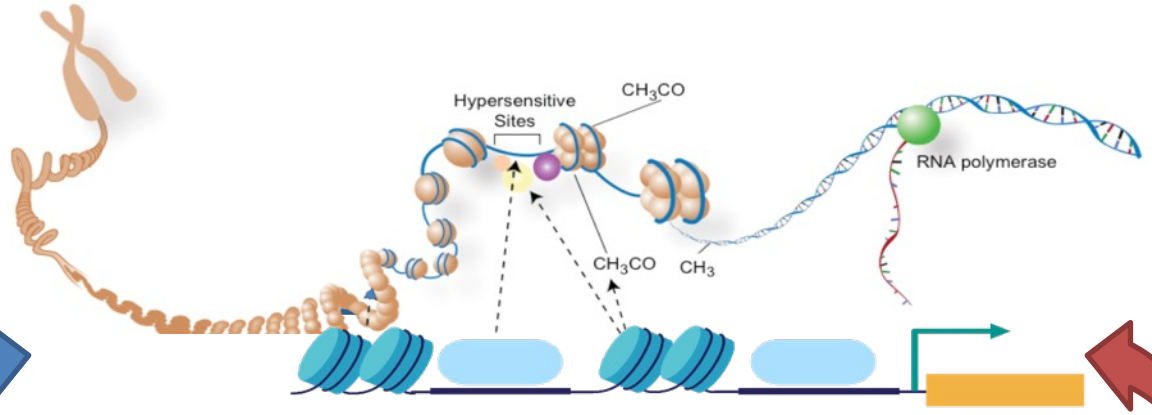
Visualization and interpretation of high-throughput genomics data

Chongzhi Zang
Associate Professor of Genome Sciences
Director of Computational Genomics, UVA Cancer Center
zang@virginia.edu
zanglab.org

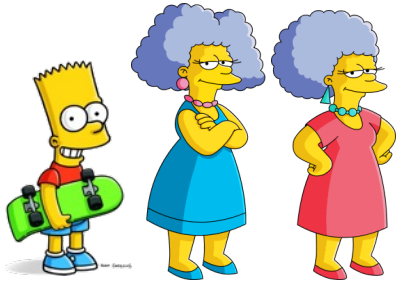
August 26, 2025

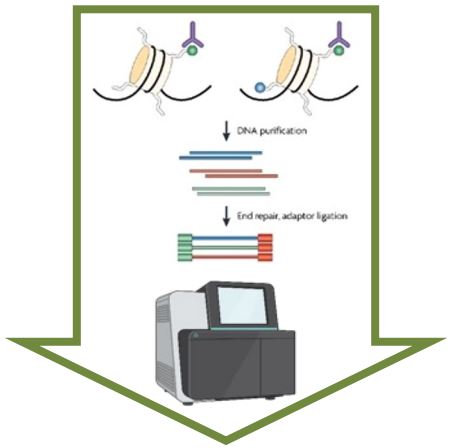
My lab develops computational methods and uses computational approaches to study epigenetics and transcriptional regulation



Method development

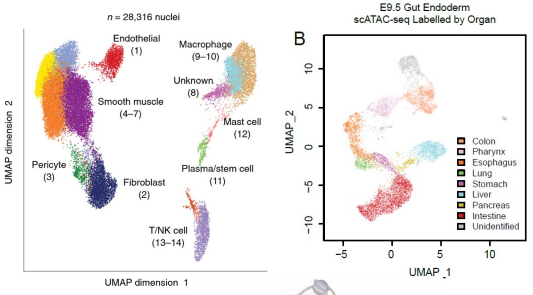
- Stat/math/physics /machine learning models
- Algorithms
- Software tools

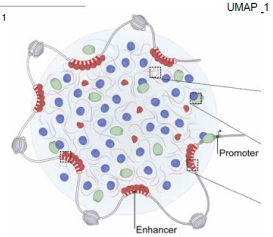




Data-inspired discovery

- Multi-omics data integration
- New models
- New discoveries

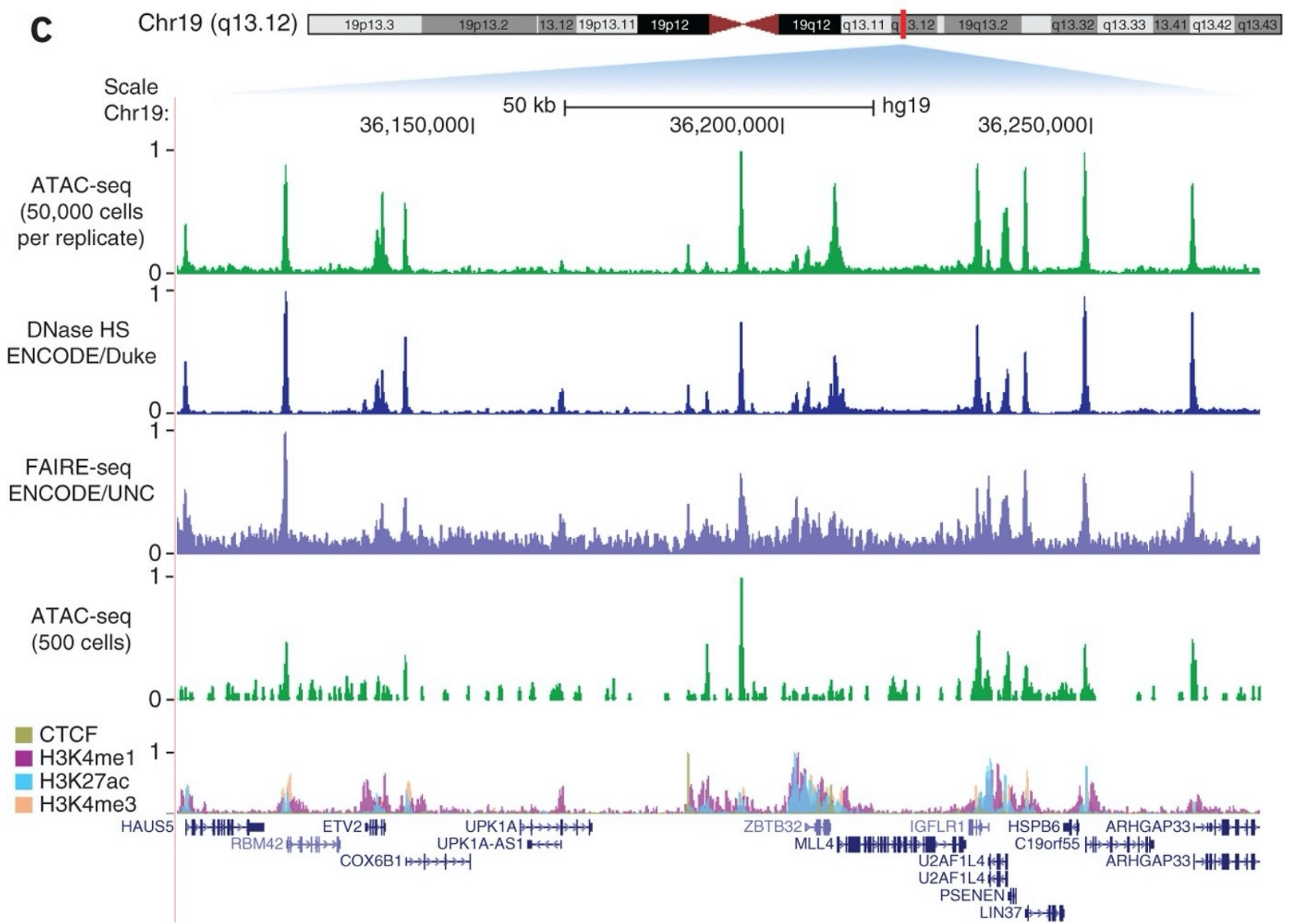




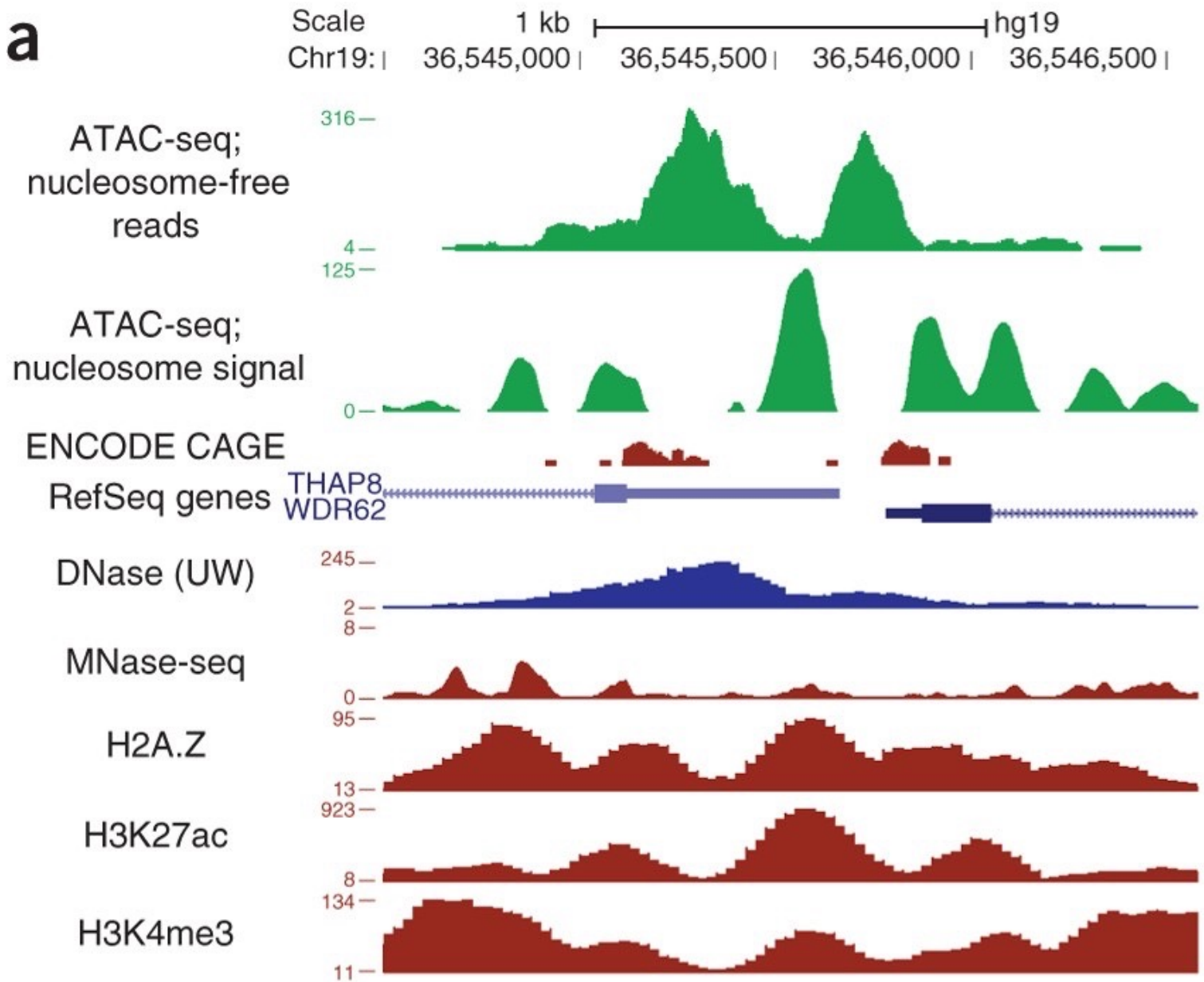

LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE
©marketoonist.com

Learning Objectives

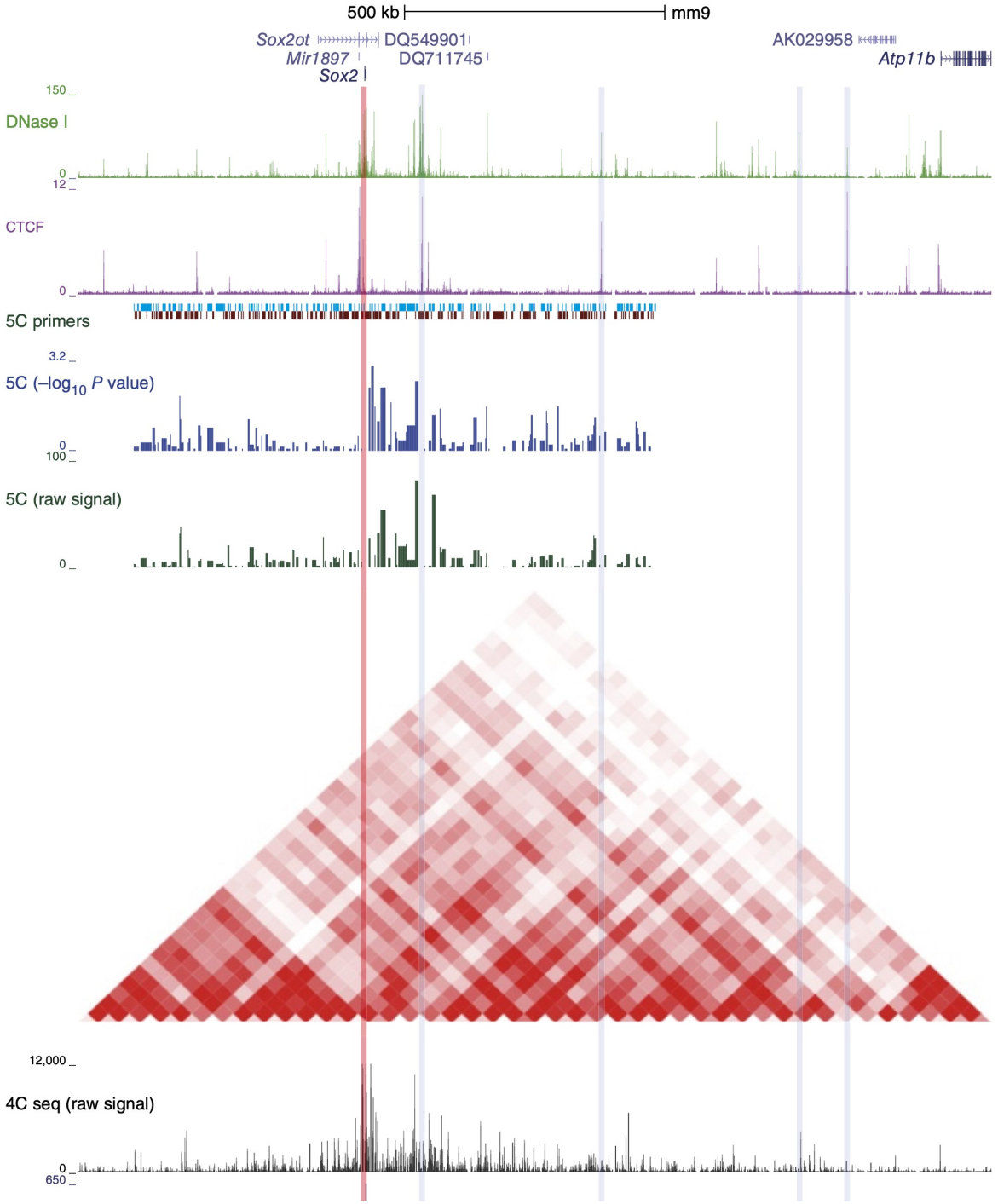
- Understand common types of plots for presenting high-throughput genomics data
- Understand essential elements of effective genomics data visualization
- Learn key principles and tips for data presentation



Buenrostro *et al*. *Nat Methods* 2013

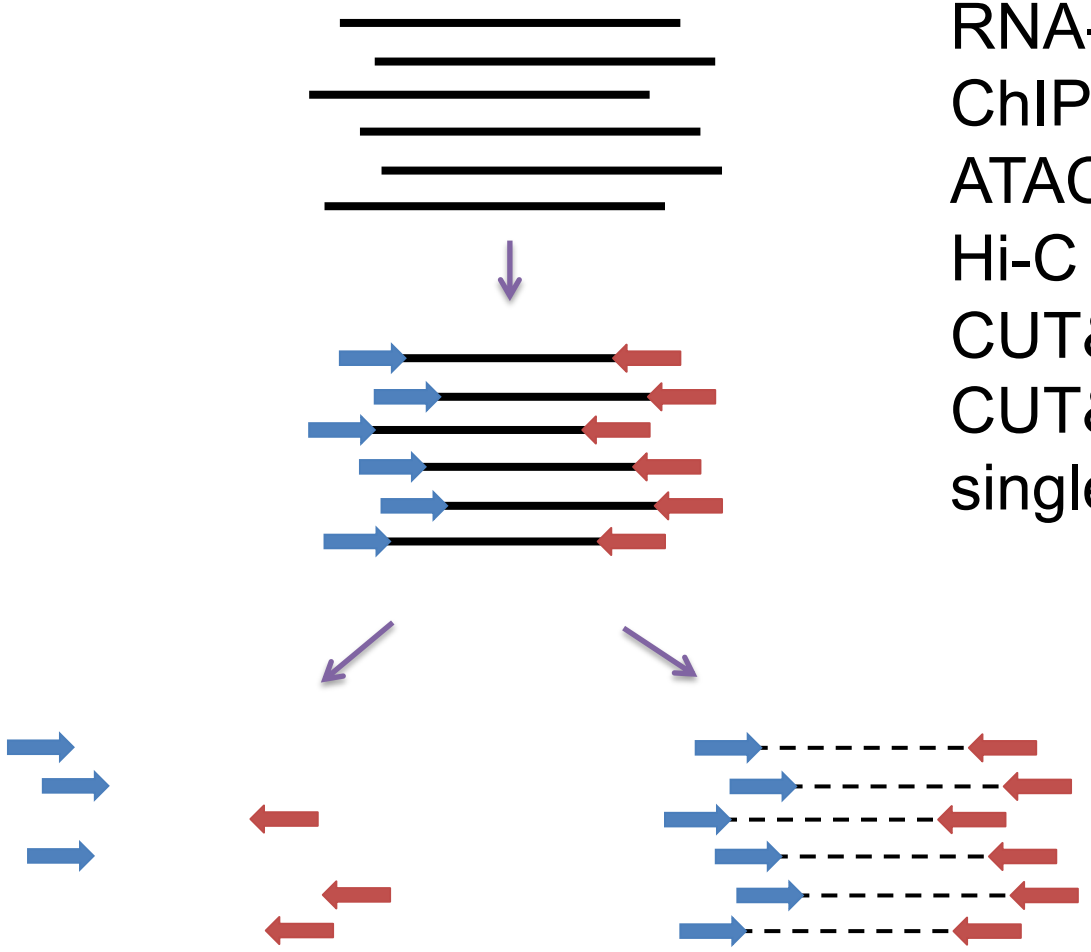


Buenrostro *et al*. *Nat Methods* 2013



Davies *et al*. *Nat Methods* 2017

High-throughput short-read sequencing (Illumina)

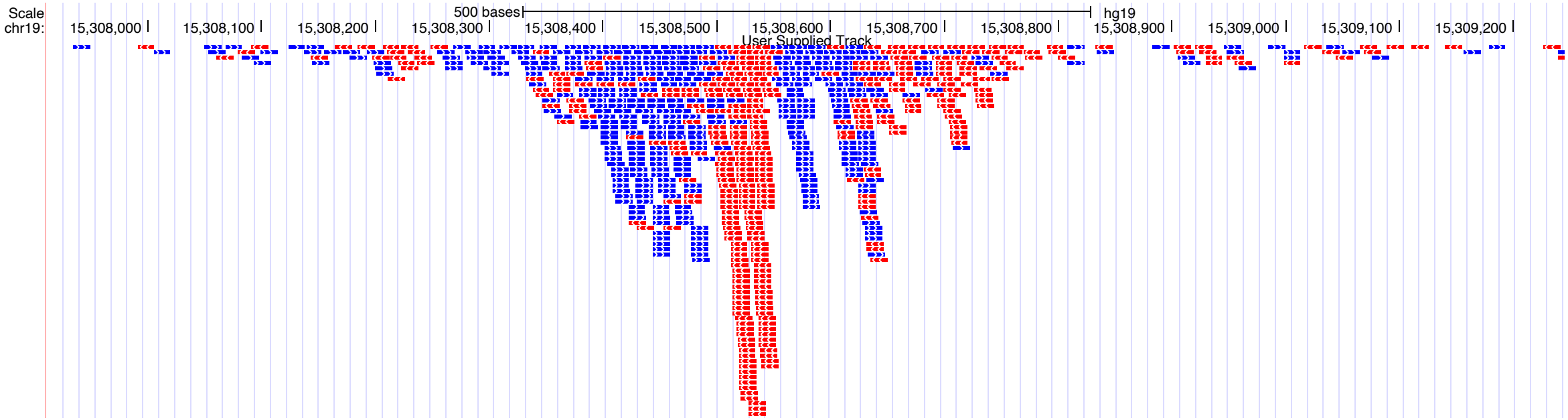


Single-end

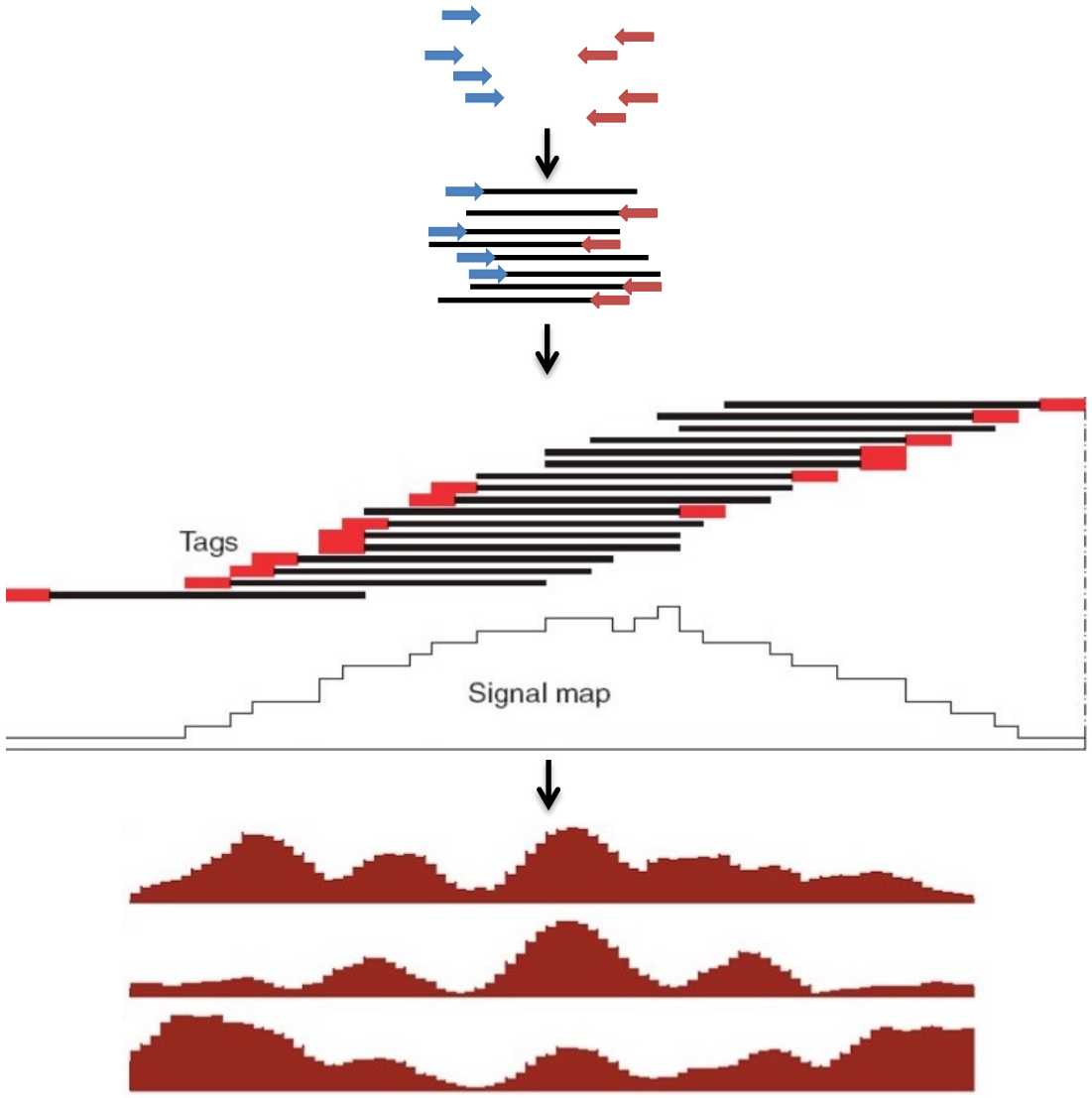
Paired-end

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 fragments piled up



- bedGraph:

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

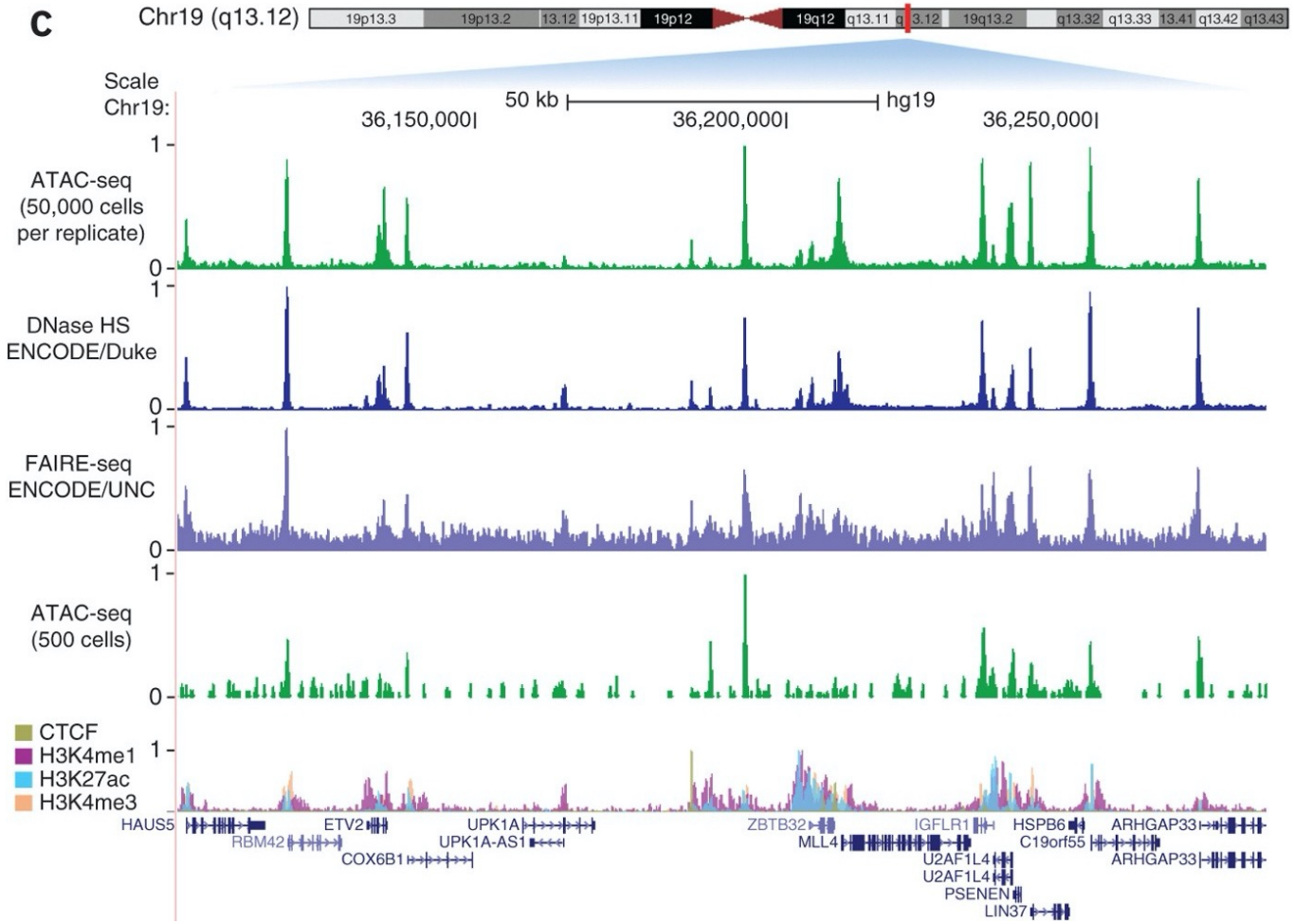
- wiggle:

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

- bigWig: indexed binary format

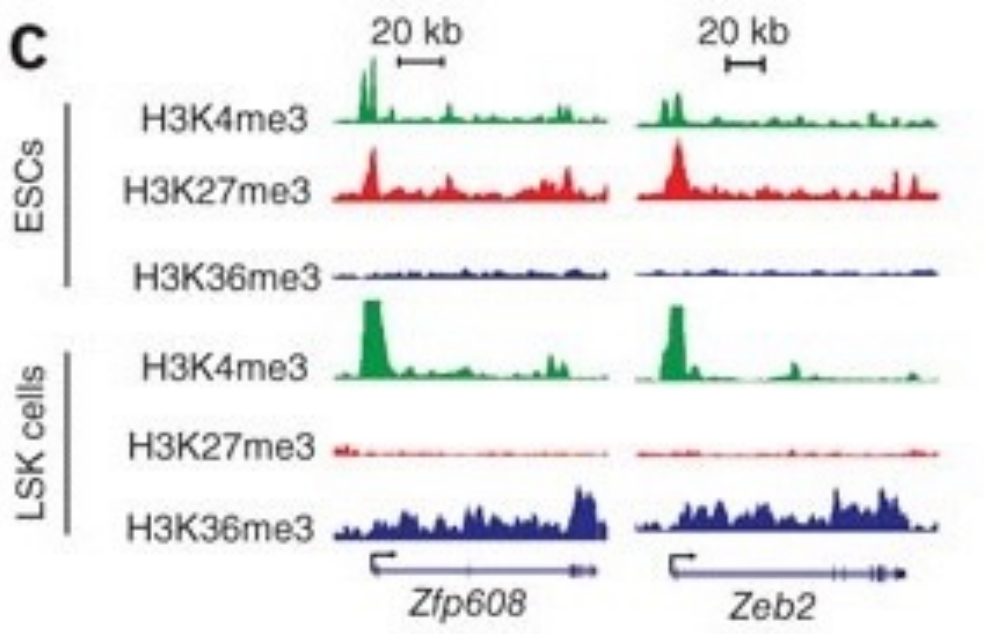
Essential elements in genome browser tracks

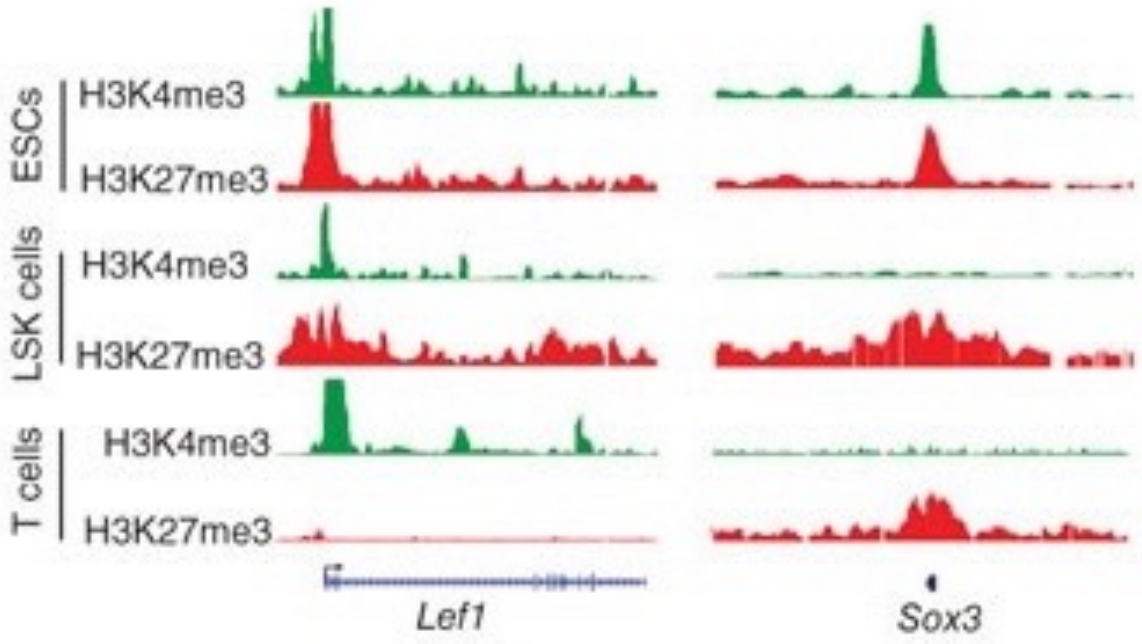
- Chromosomal locations
- Track label
- Track scale
  - x: resolution?
  - y: normalization?



Buenrostro *et al*. *Nat Methods* 2013

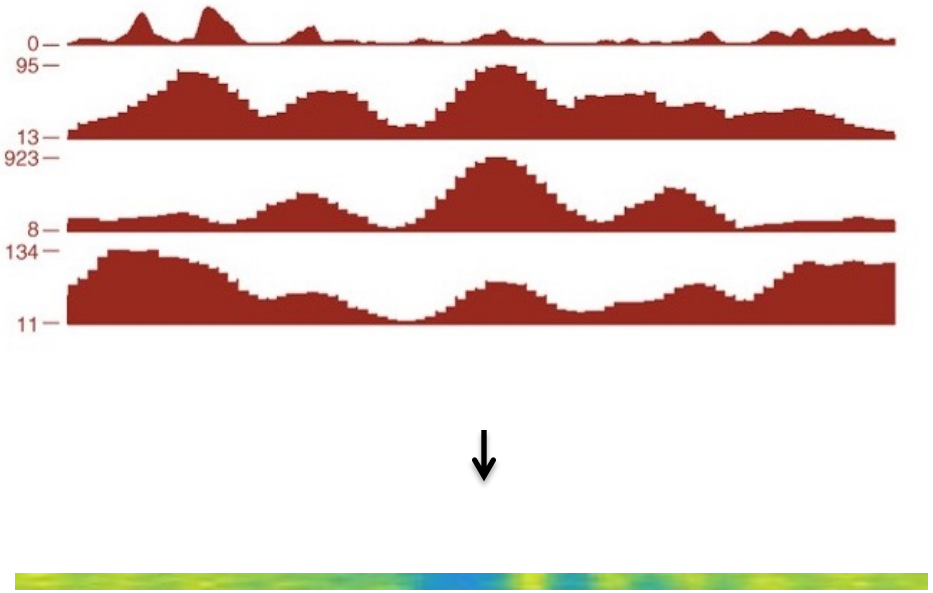
How to integrate patterns observed on signal tracks?

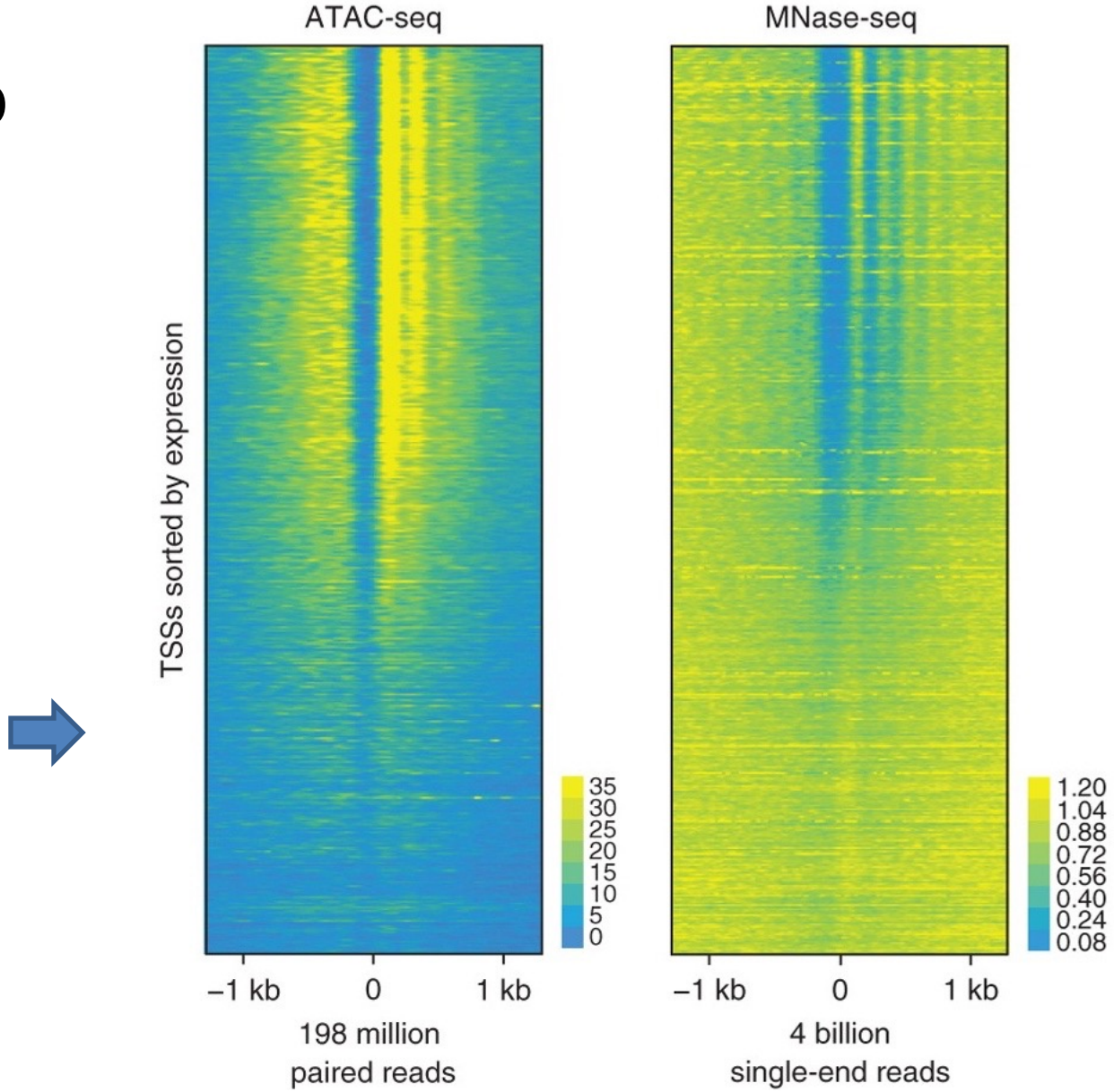




Adli *et al*. *Nat Methods* 2010

Stacked “ripple” heatmap

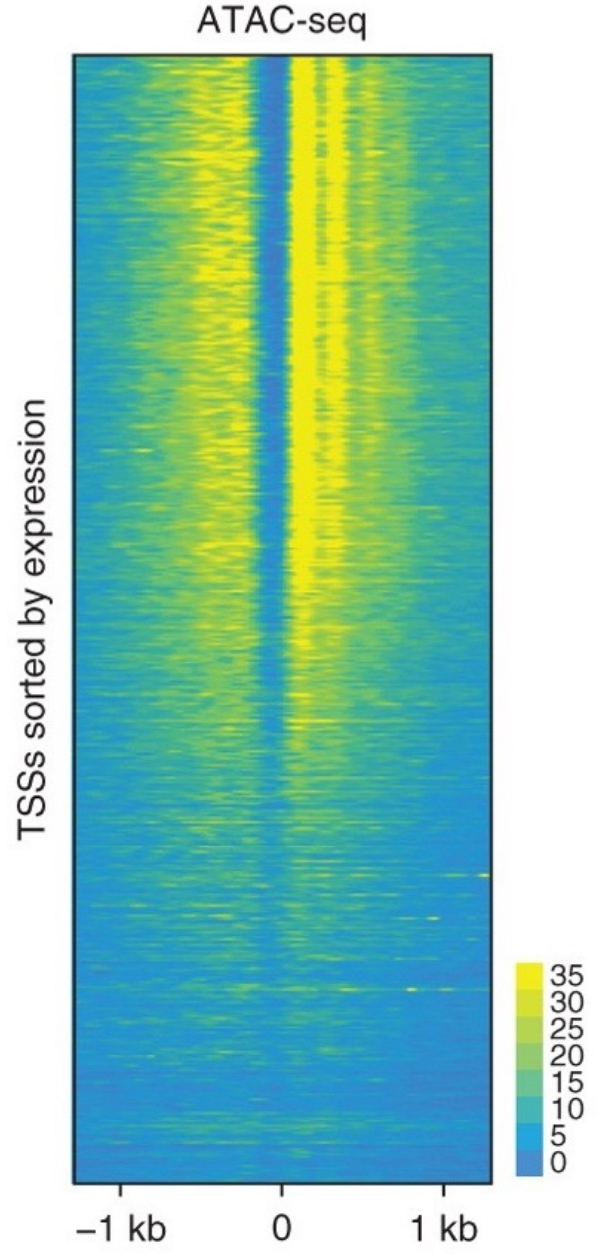




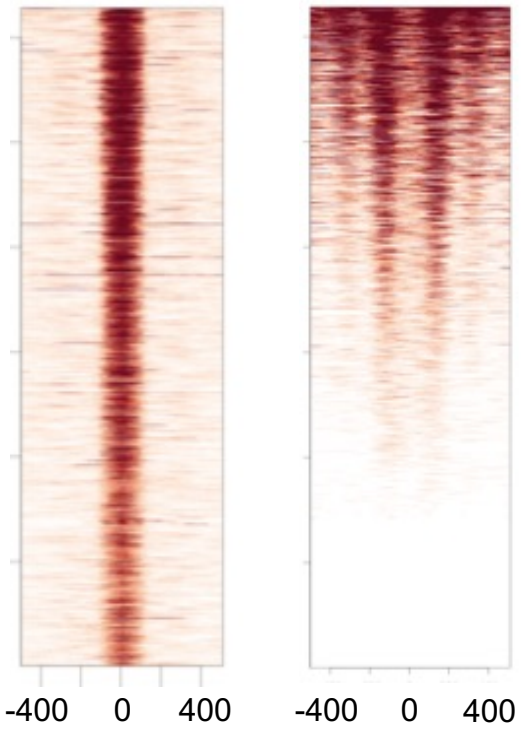
Buenrostro *et al*. *Nat Methods* 2013

Essential elements in a ripple heatmap

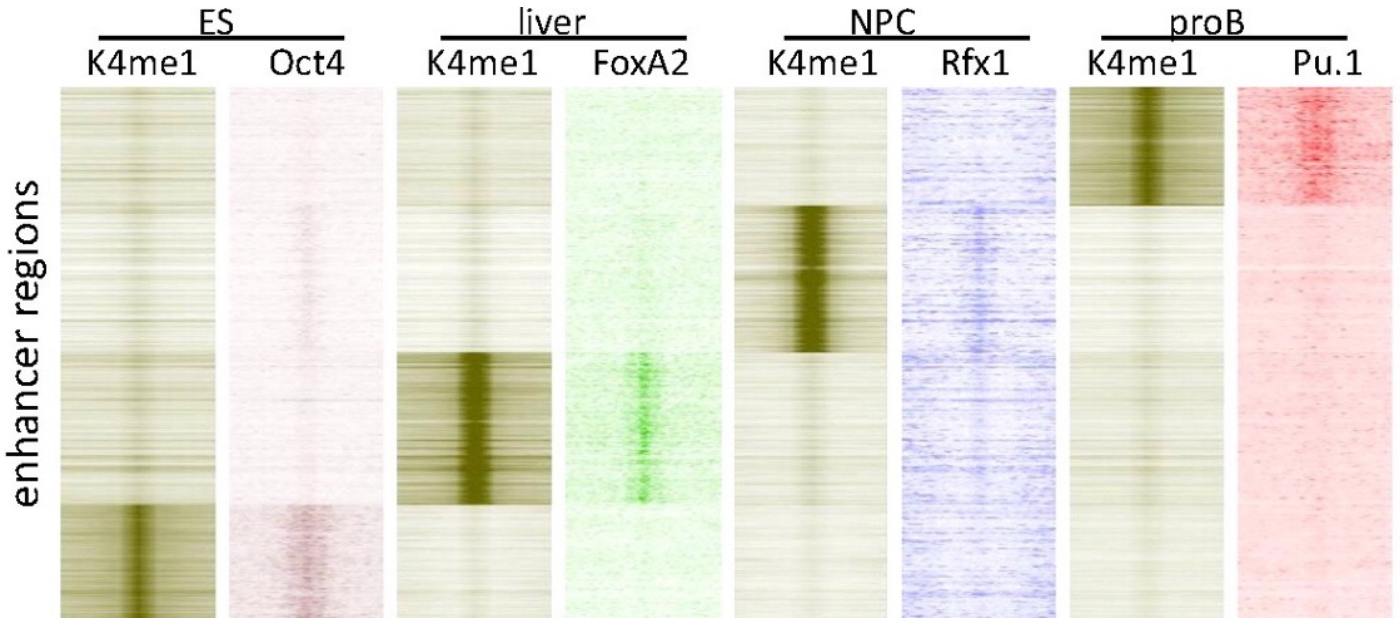
- Heatmap presents 3-dimensional data
- x: What loci/anchor is each row? Range?
- y: What are the rows? How many? How are they ranked?
- h: Data title/label (what signal?) color scale?



Multiple datasets visualization by ripple heatmap

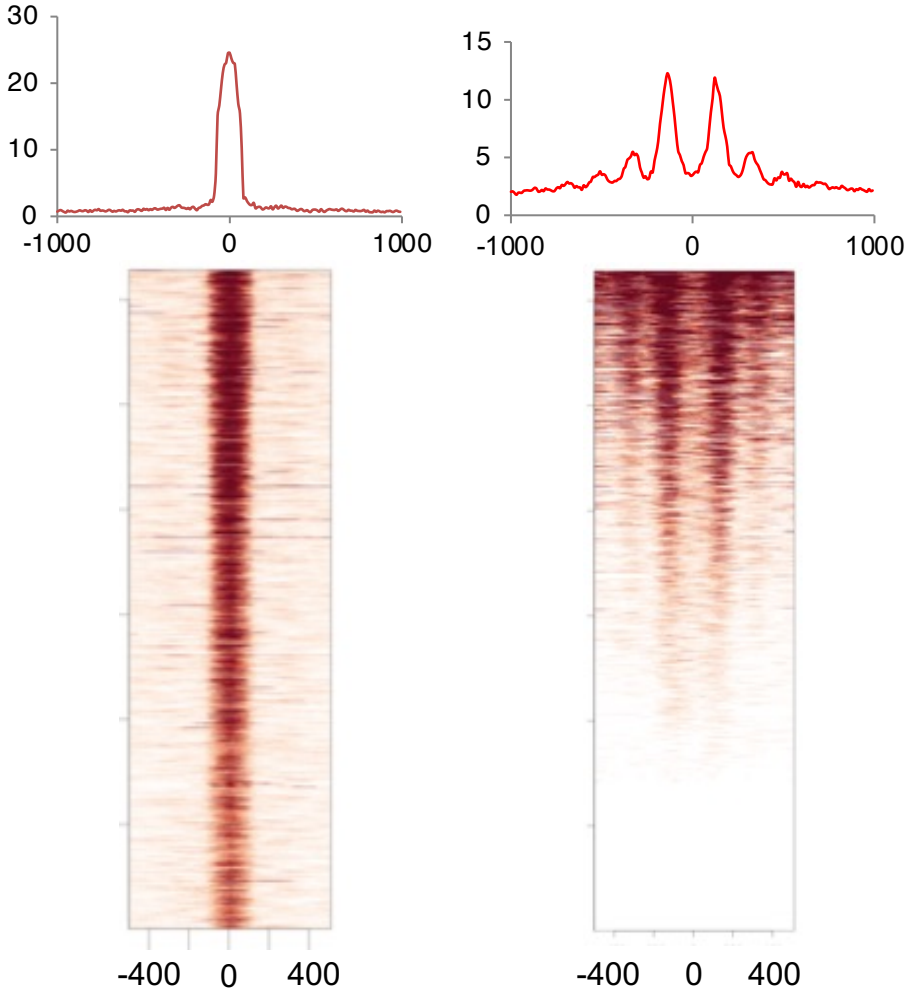


Luyten *et al*. *Genes Dev* 2014



Creyghton *et al. PNAS* 2010

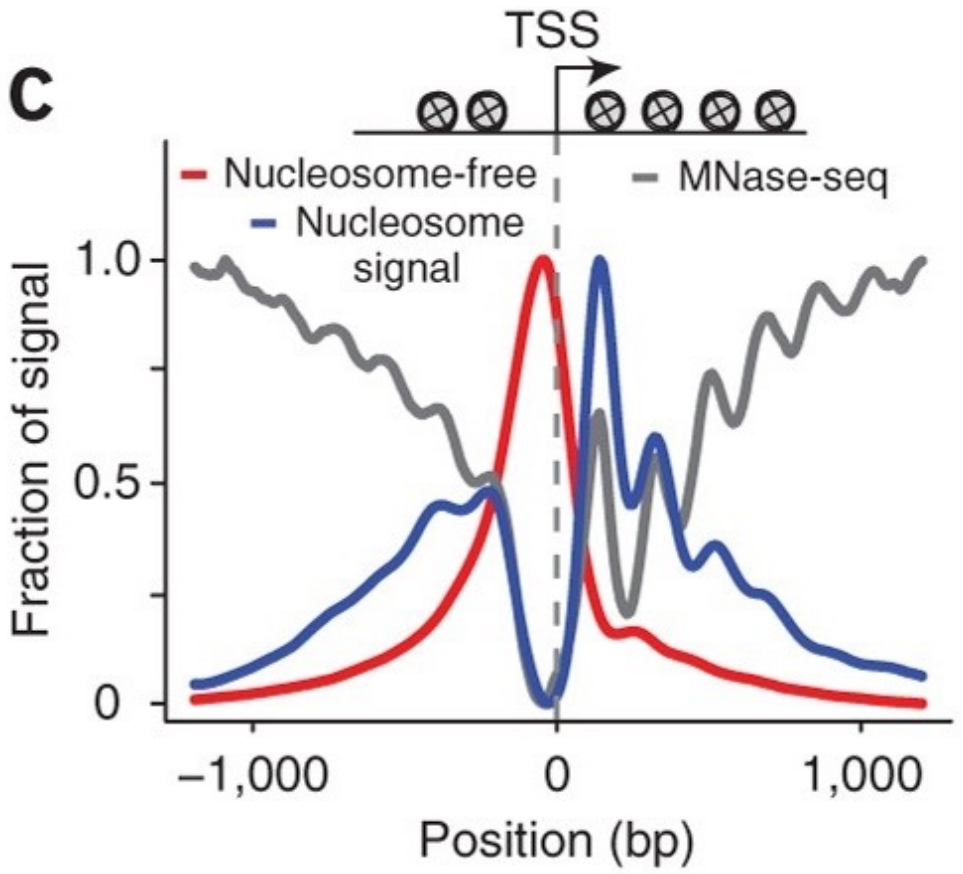
Composite plots



Luyten *et al*. *Genes Dev* 2014

Essential elements in a composite plot

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



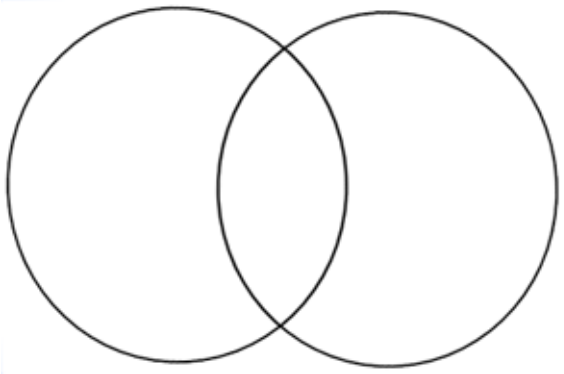
Buenrostro *et al*. *Nat Methods* 2013

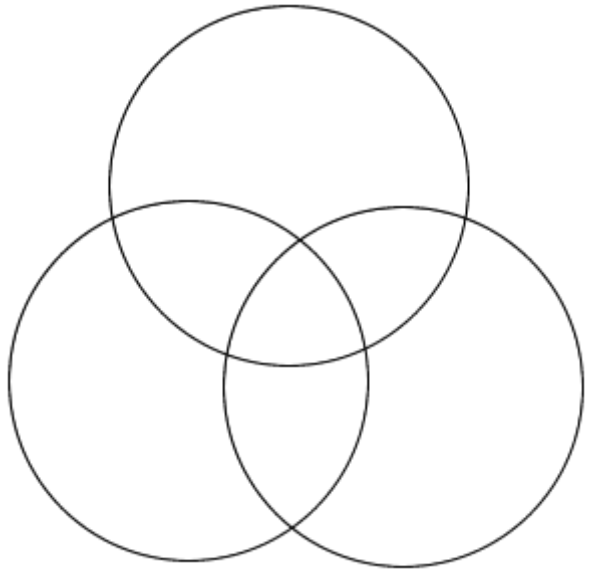
Common misinterpretations of composite plots

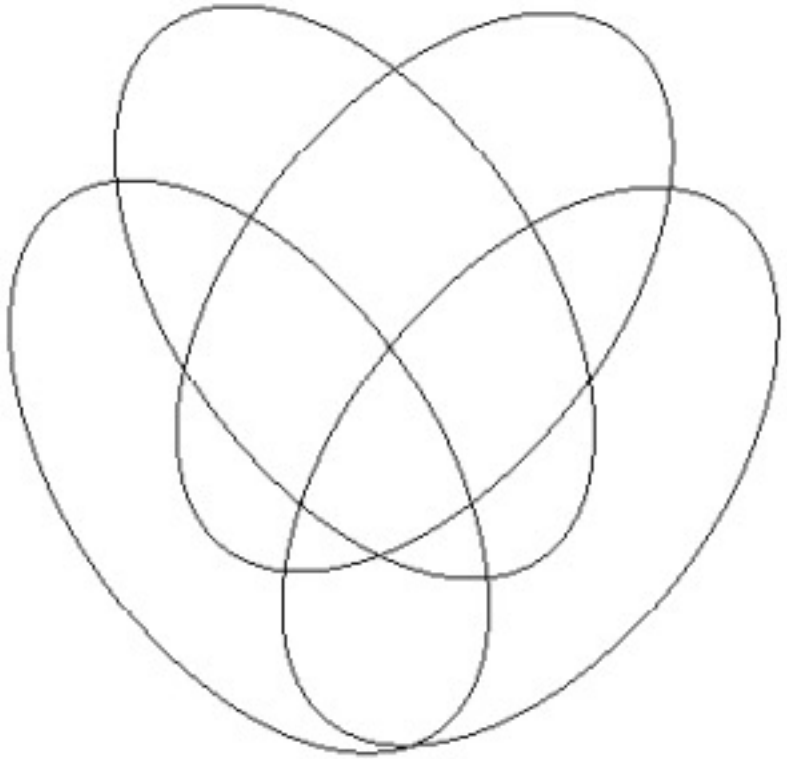
- Caveat 1: A peak in a composite plot may be contribute by only a tiny fraction of regions (not representative of global picture)

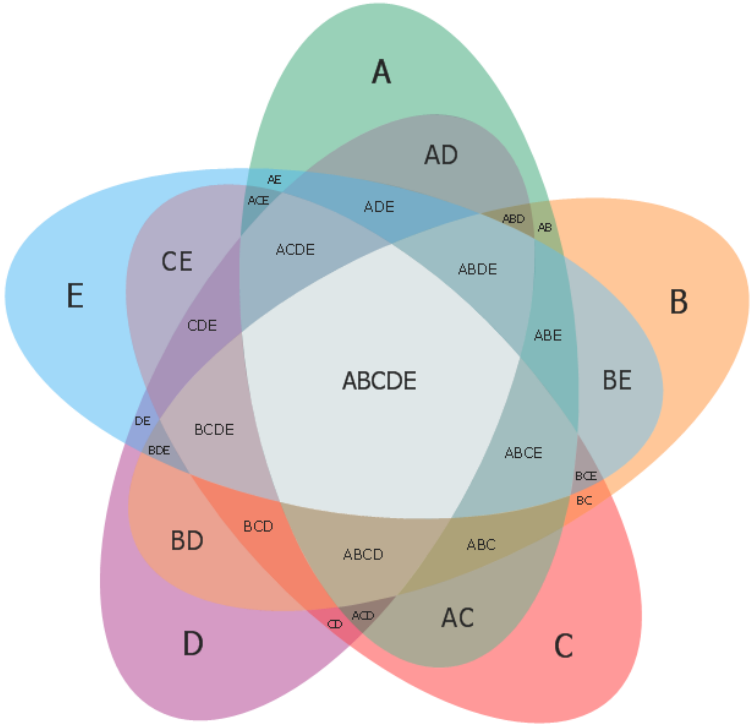
- Caveat 2: A higher peak does not necessarily mean stronger signal or more region coverage

Venn diagram presents yes/no relations between sets

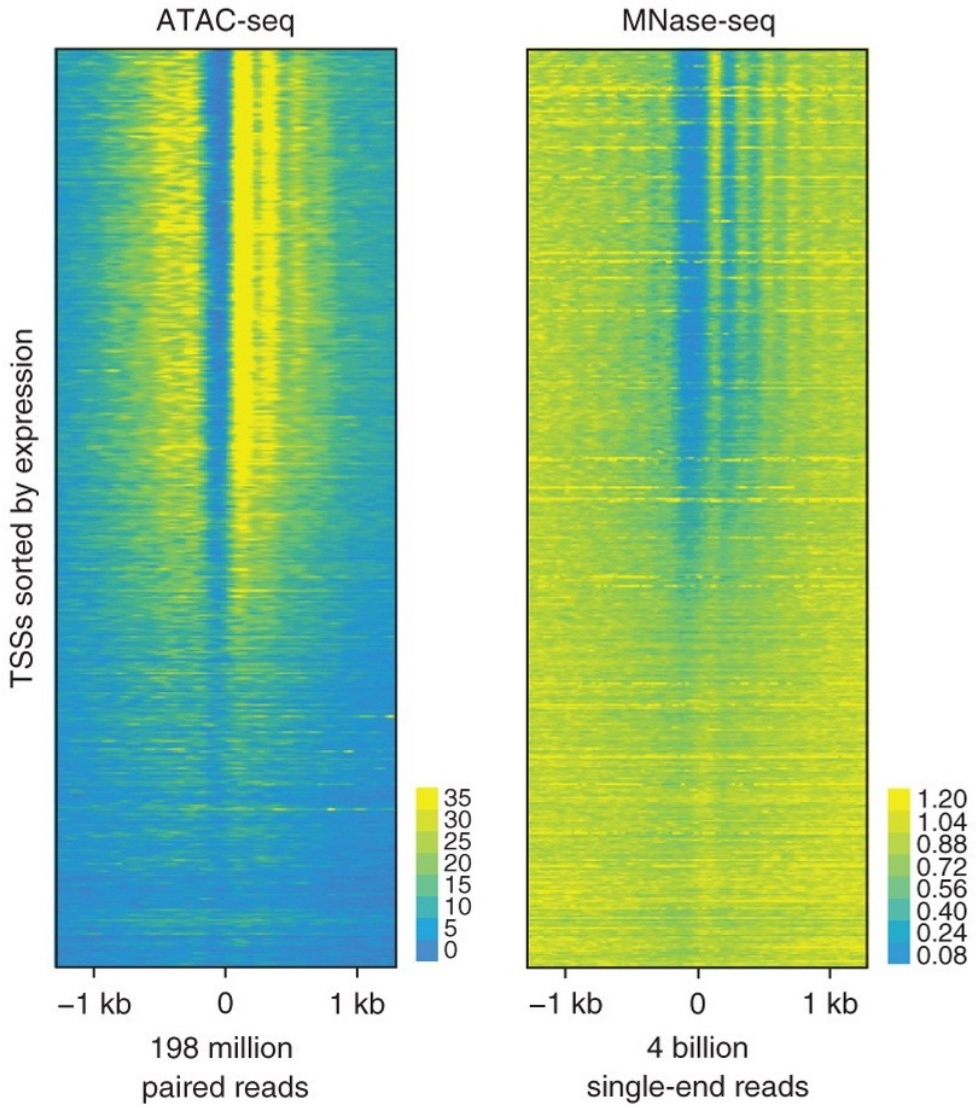


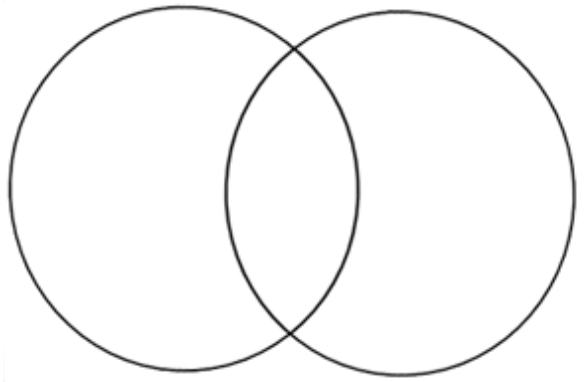


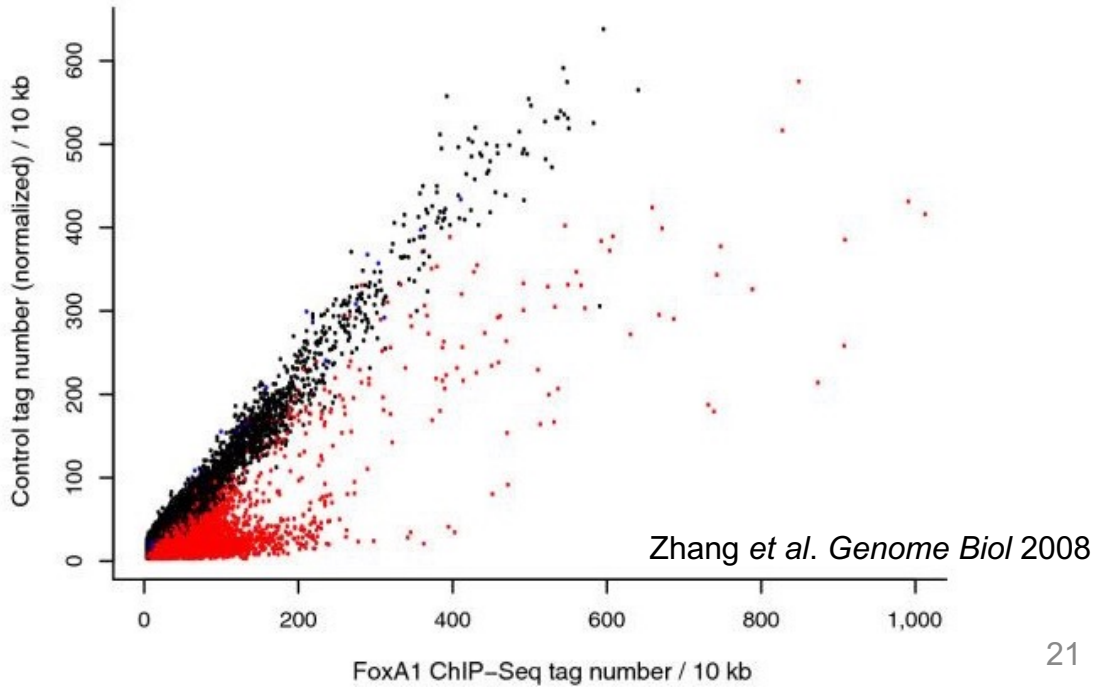




Heatmaps and scatter plots are more informative than a Venn diagram





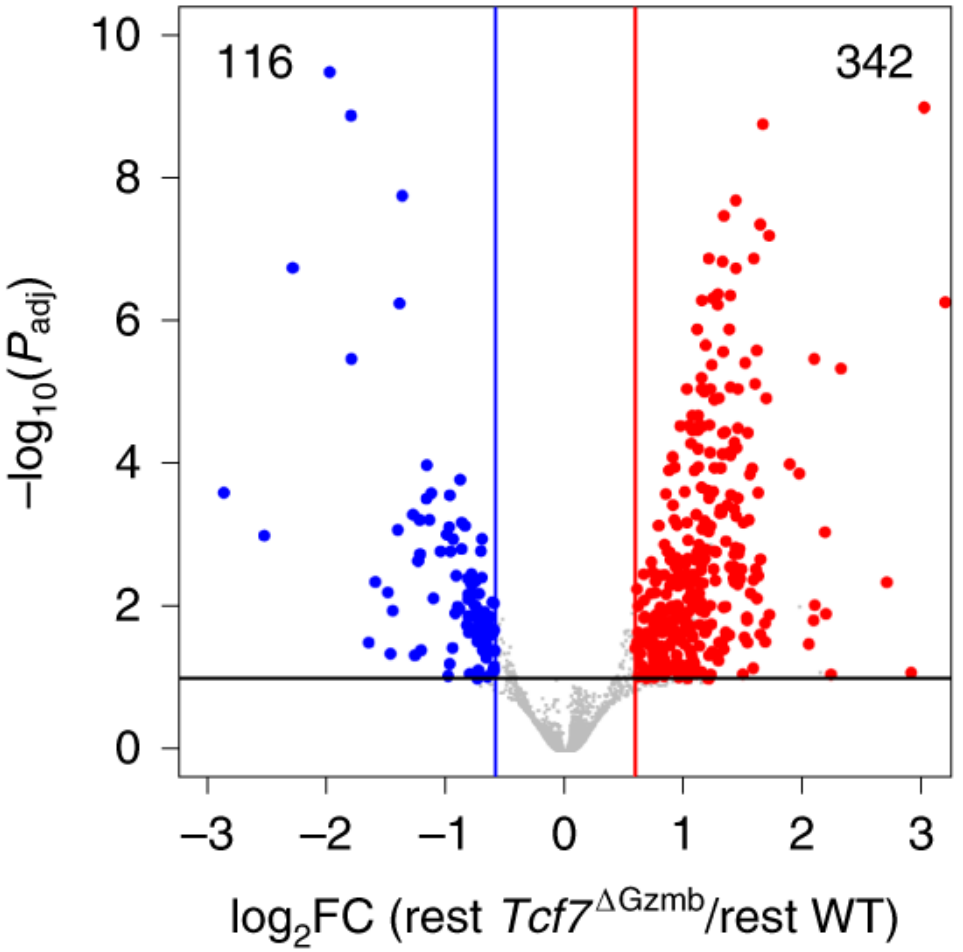


Zhang *et al*. *Genome Biol* 2008

Volcano plot for differential gene expression

- Scatter plot
- 2 dimensions:
  - x: signal strength (e.g., log2 fold change)
  - y: statistical significance (e.g., -log10P)

- Set cutoffs on 2 axes

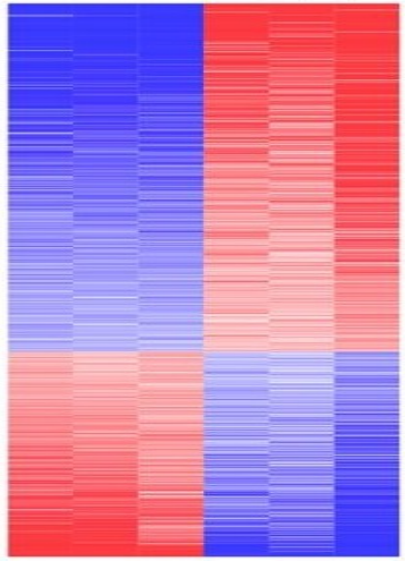


Shan *et al*. *Nature Immunology* 2022

Differential gene expression visualization by heatmap

Differential expression

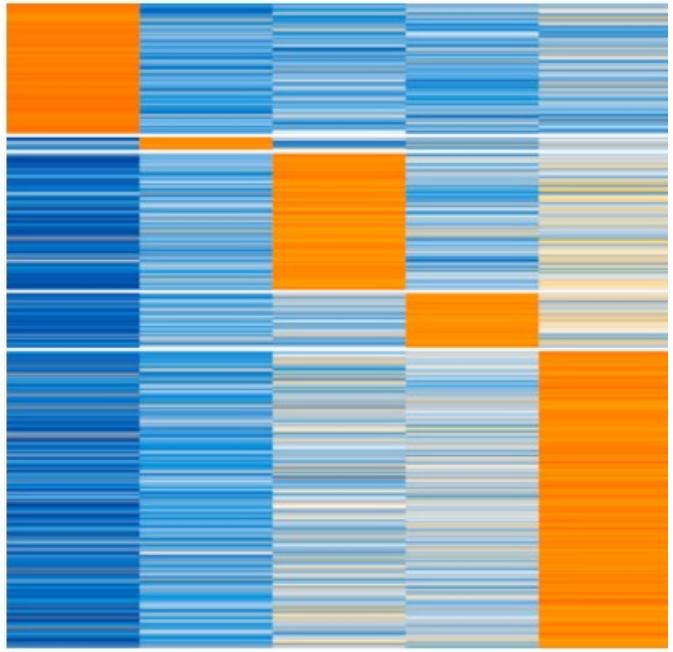
control treated

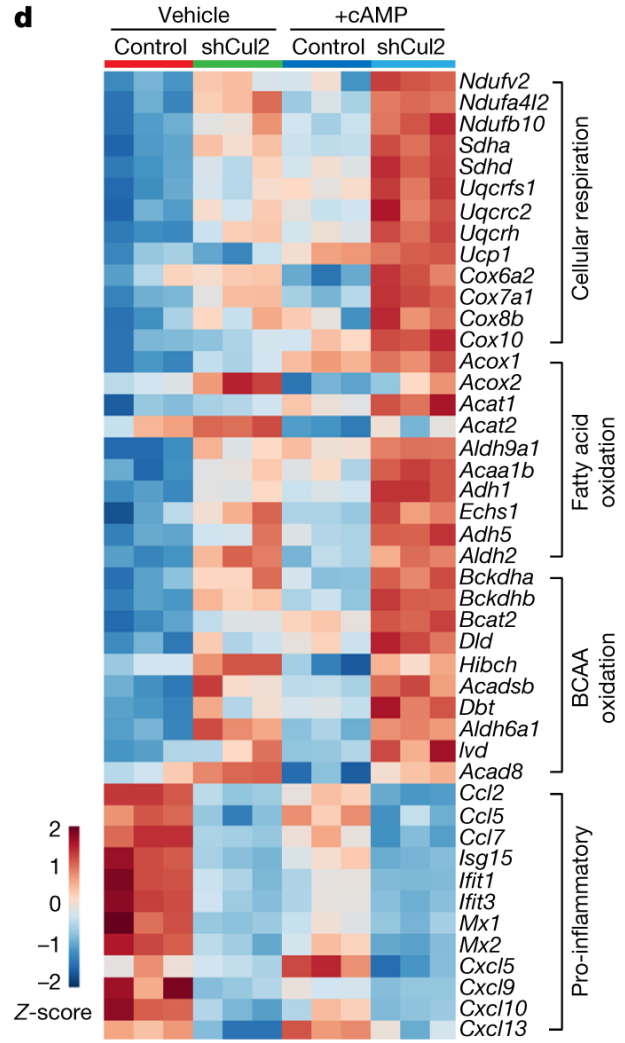


up-regulated

down-regulated

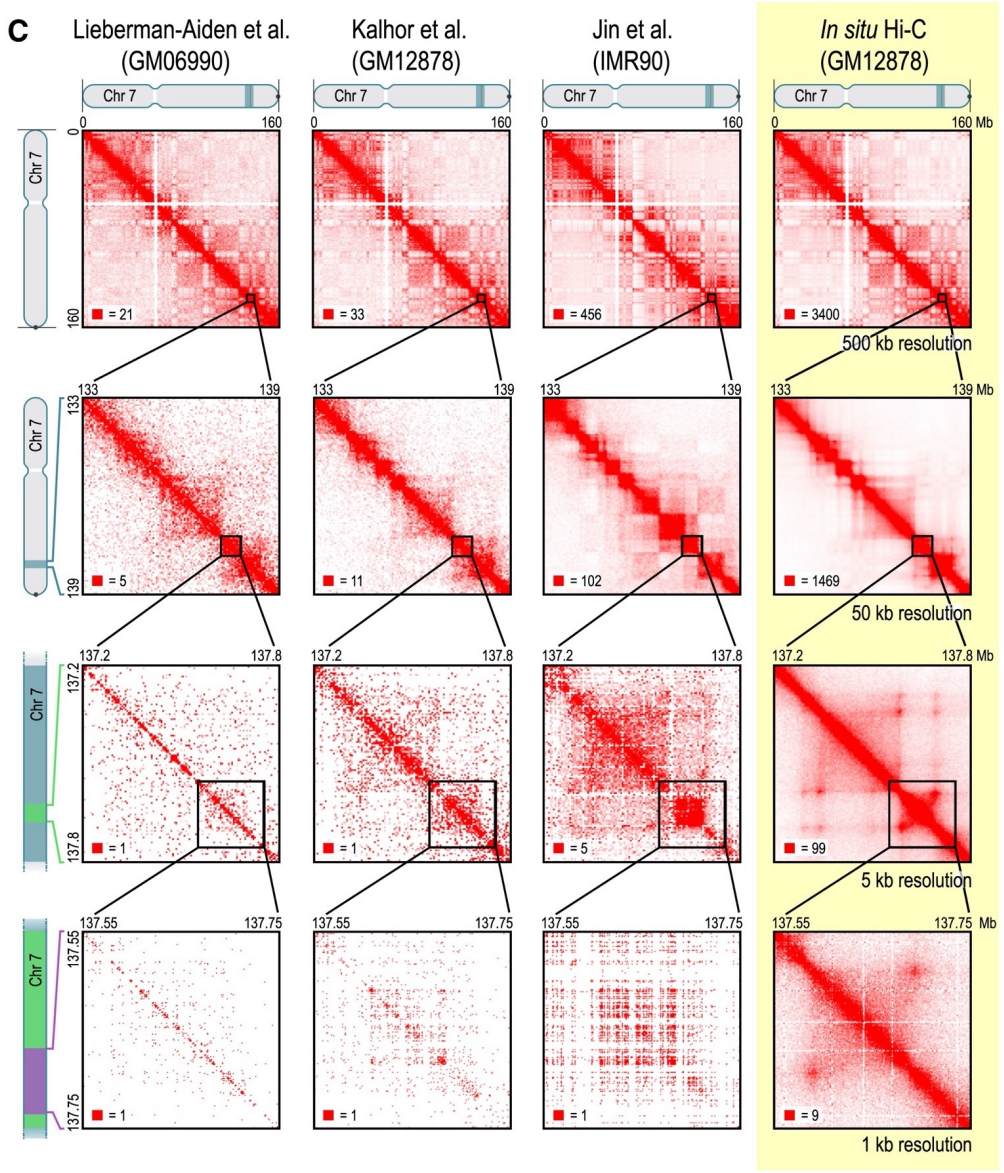
Co-expressed gene clusters





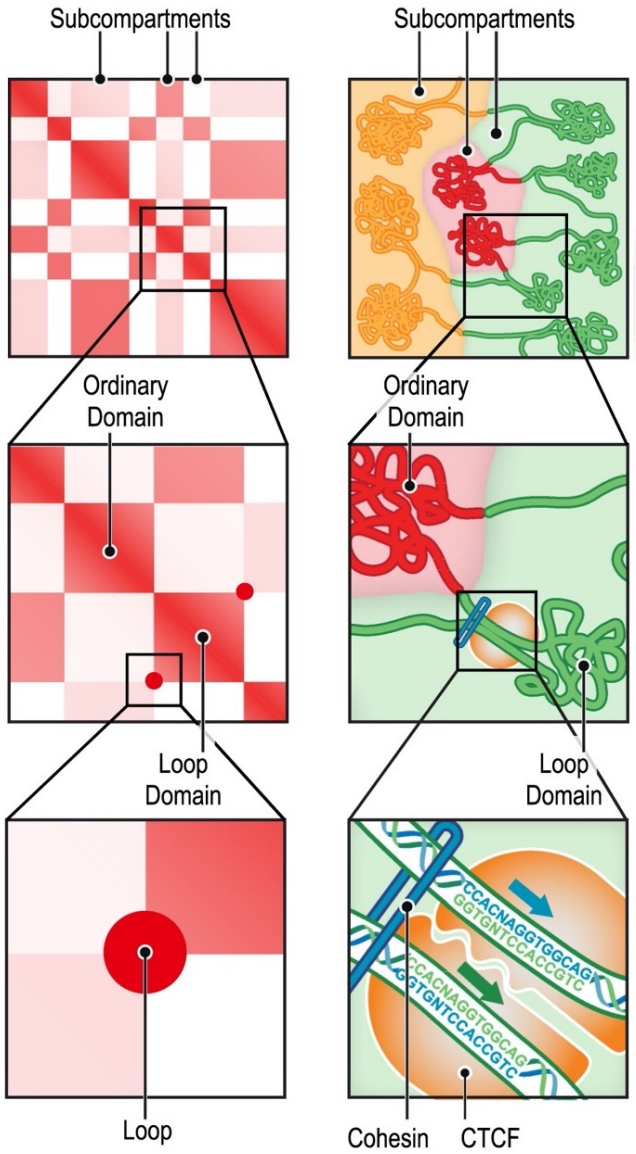
Wang *et al. Nature* 2022

Hi-C contact heatmap for 3D genome interactions



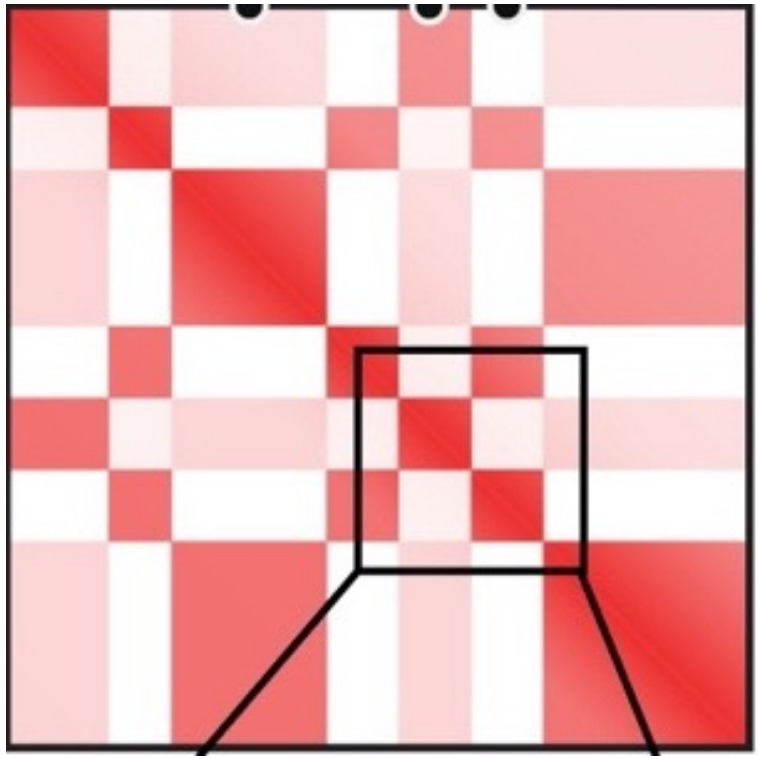
Rao *et al. Cell* 2014

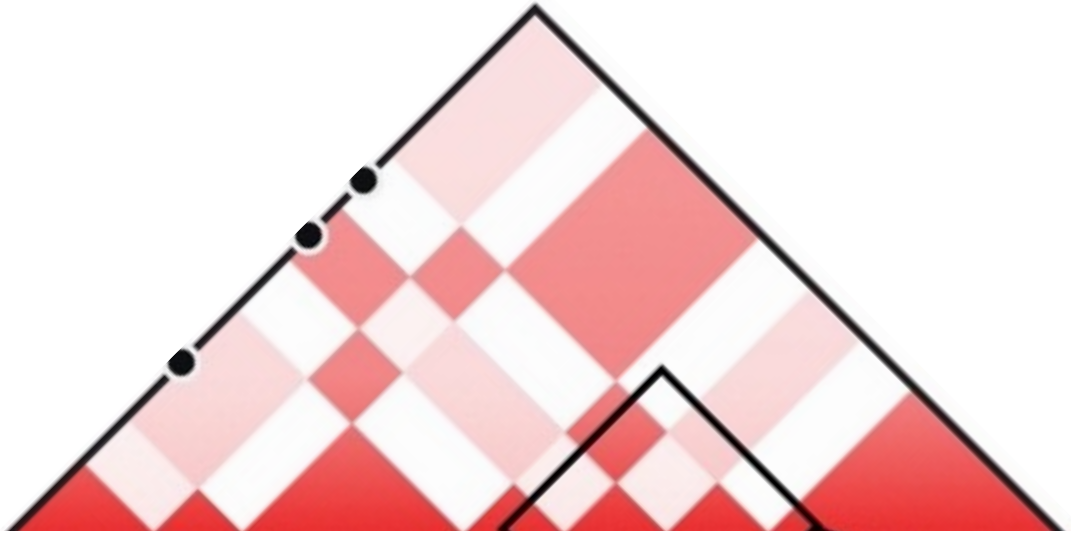
Hi-C contact heatmap for 3D genome interactions



Rao *et al. Cell* 2014

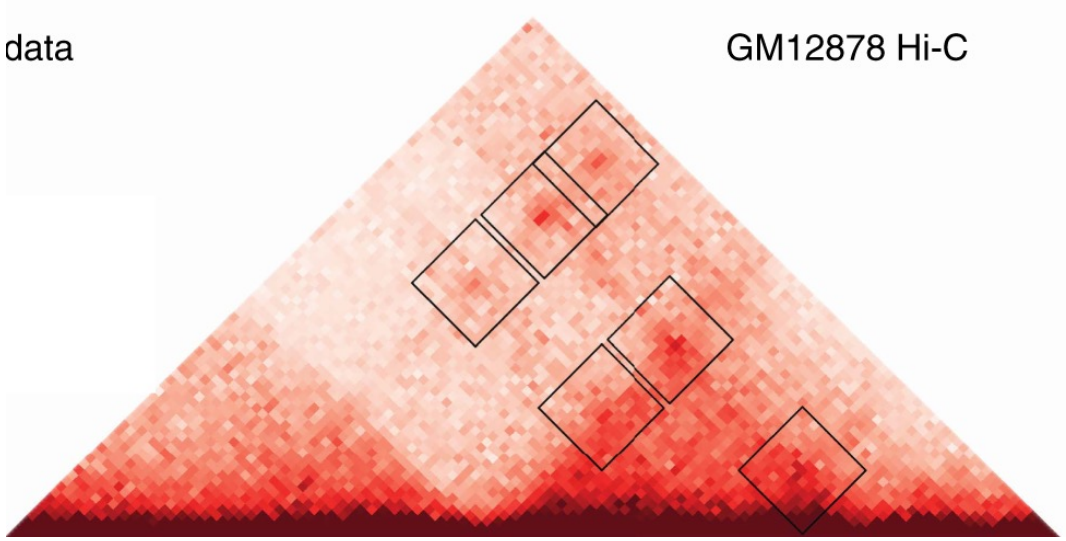
Hi-C contact heatmap for 3D genome interactions

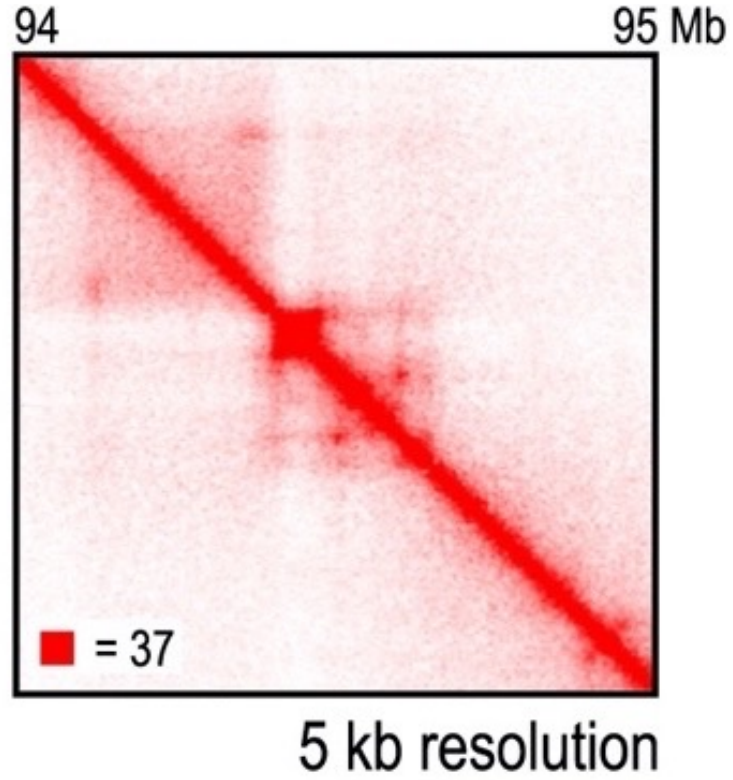


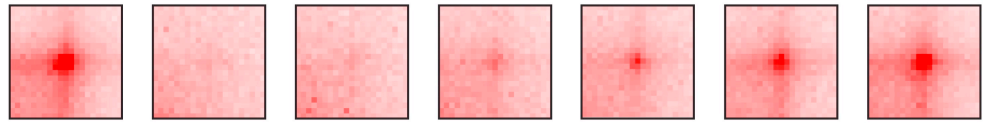


Essential elements in a Hi-C contact heatmap

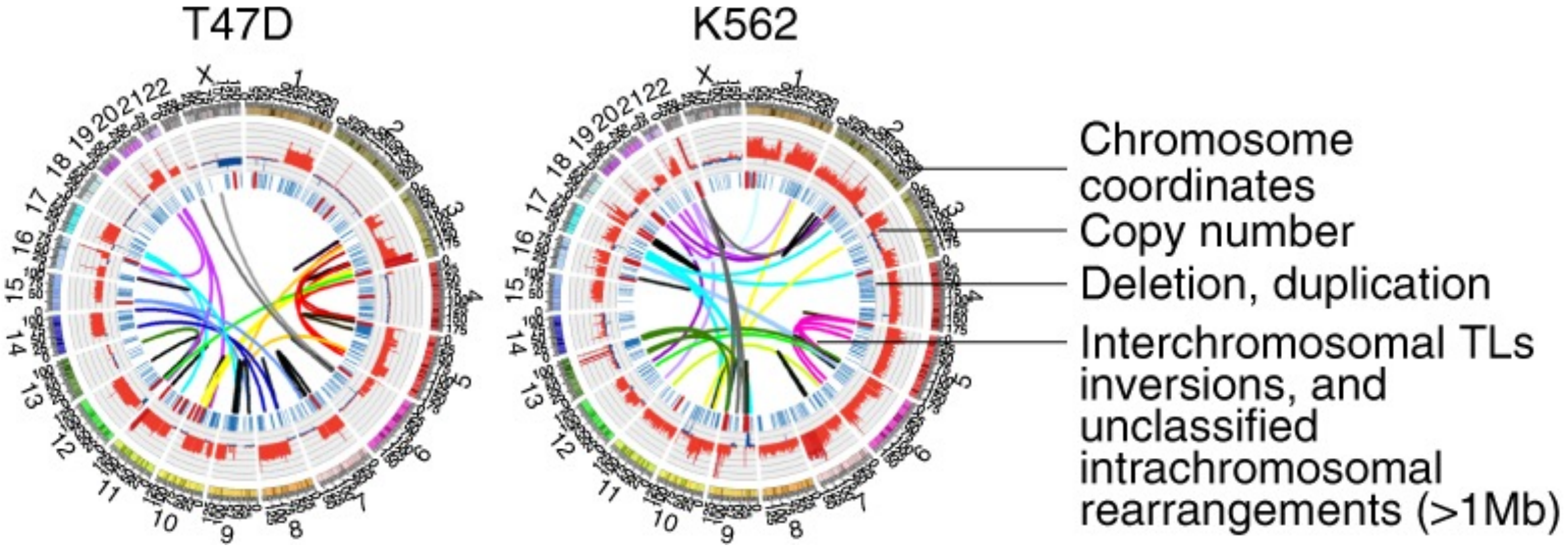
- Scale, scale, scale
- Resolution
- Normalization
- Blocks, stripes, loops (2d peaks)





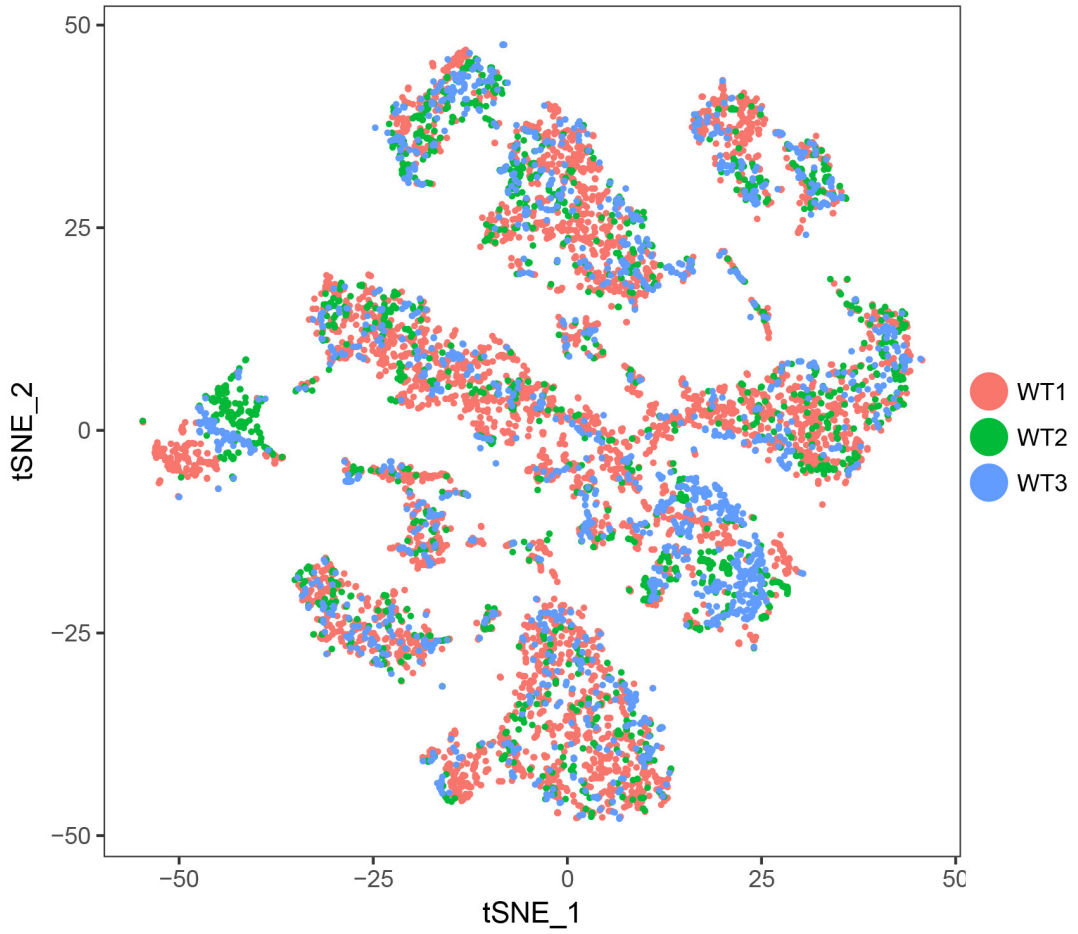


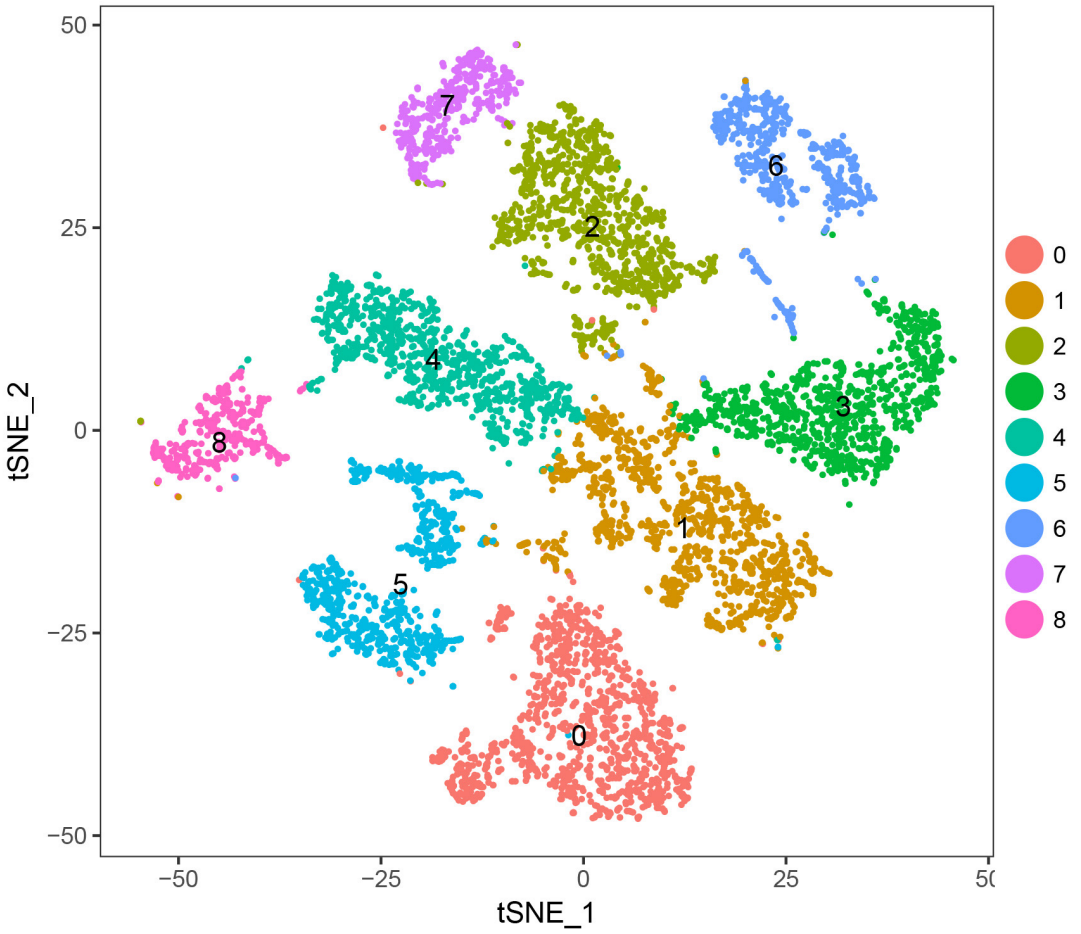
Circos plot integrates multiple layers of genomic information



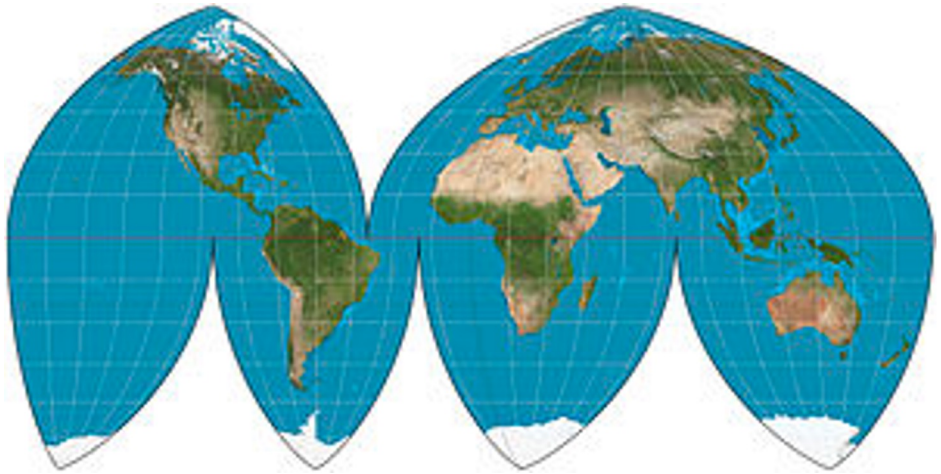
Dixon *et al. Nat Genet* 2018

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

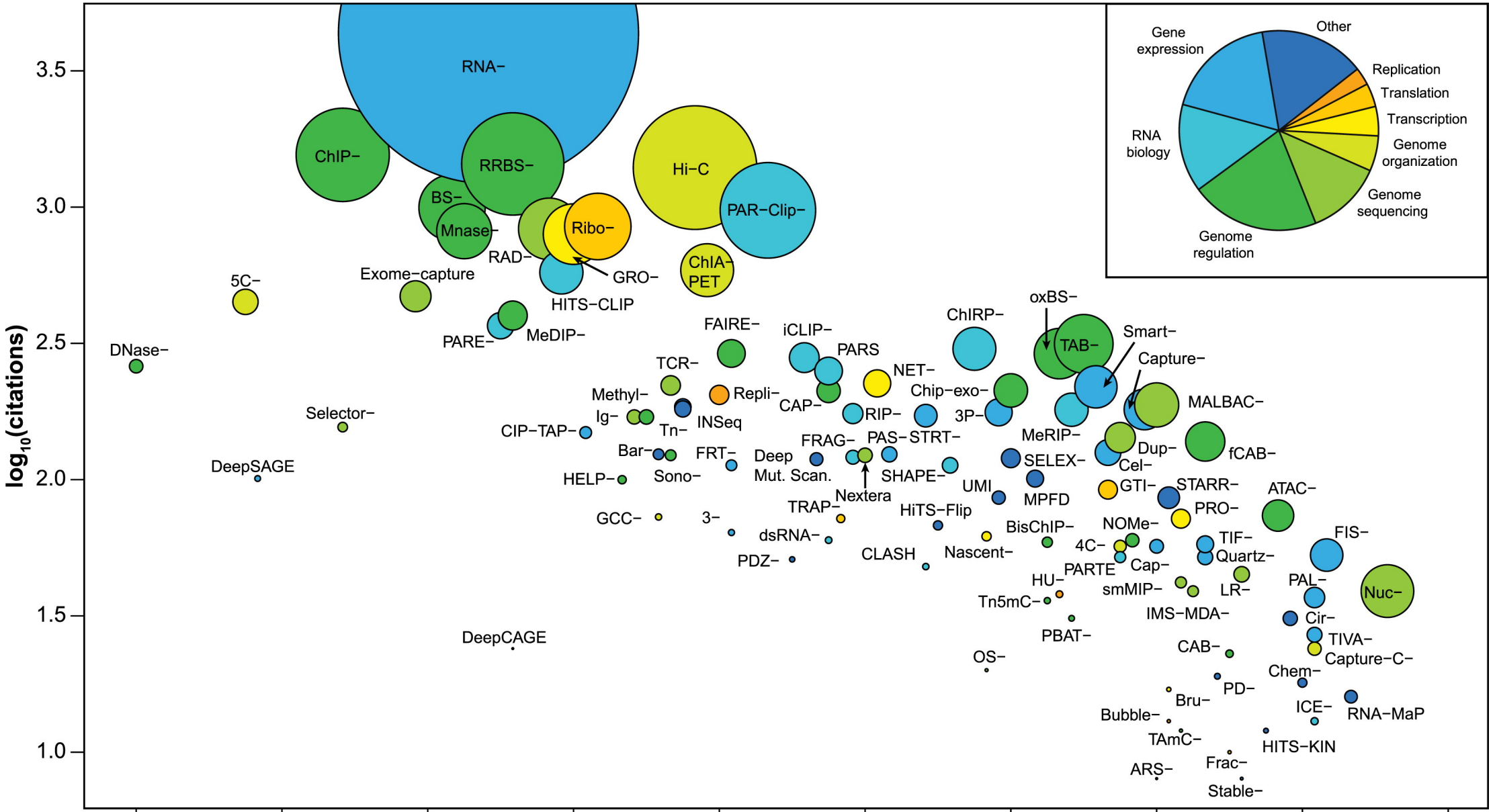




Information may be lost or distorted in dimensionality reduction

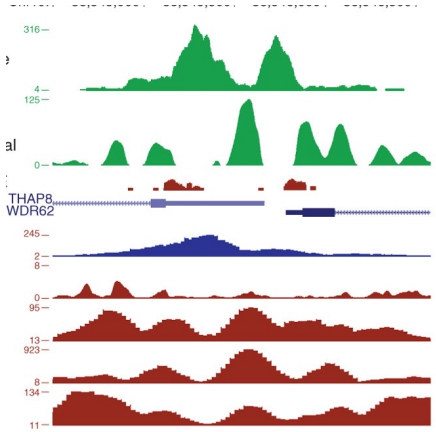
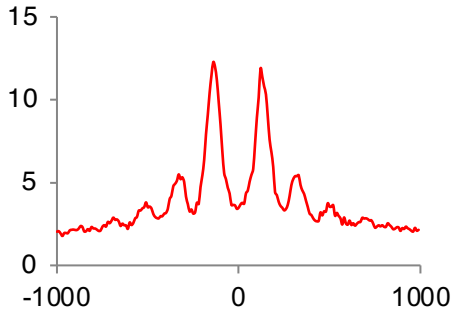
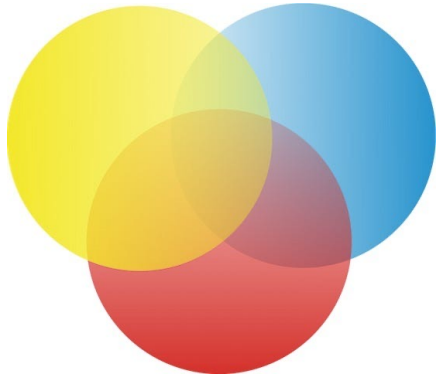
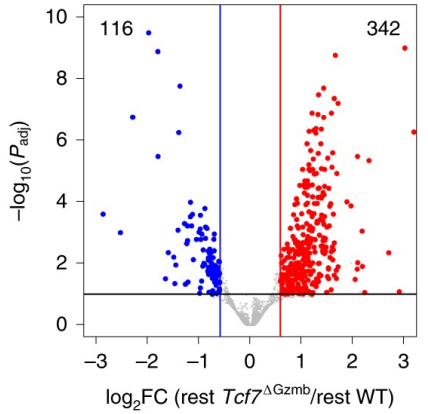
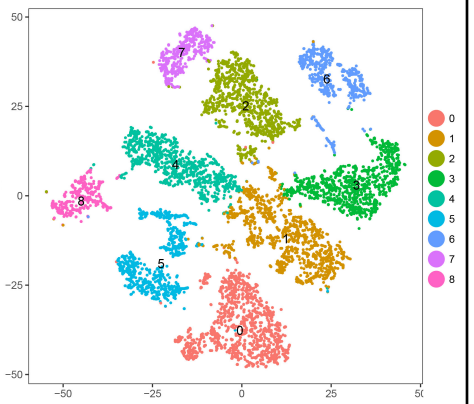
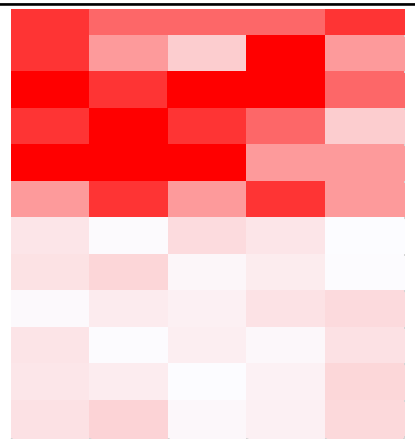
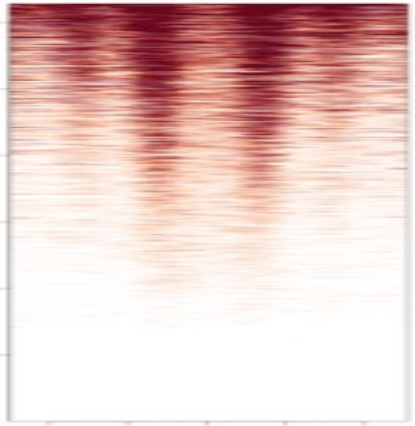
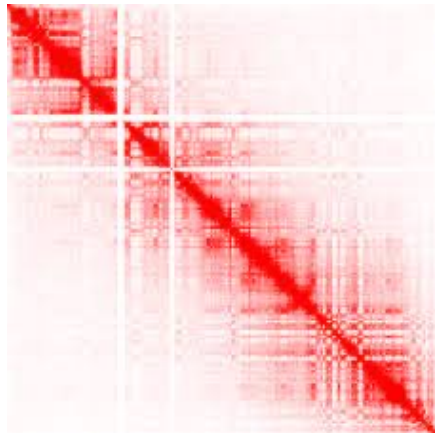
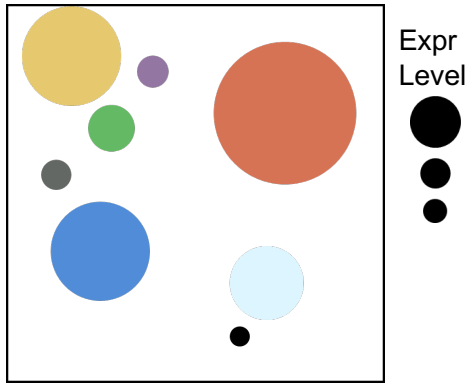
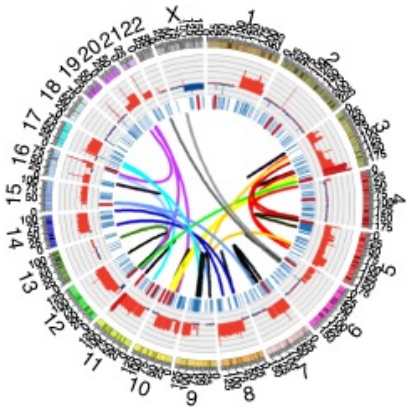


Bubble plots: NGS-based applications (-seq)



Reuter *et al*. *Mol Cell* 2015

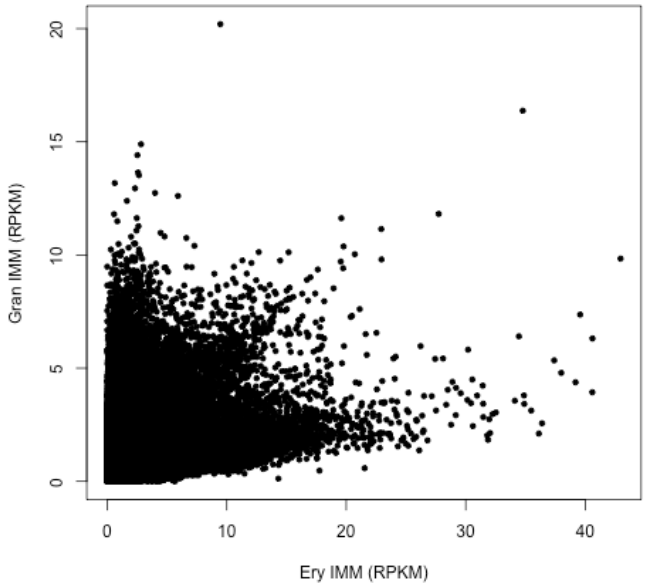
Summary

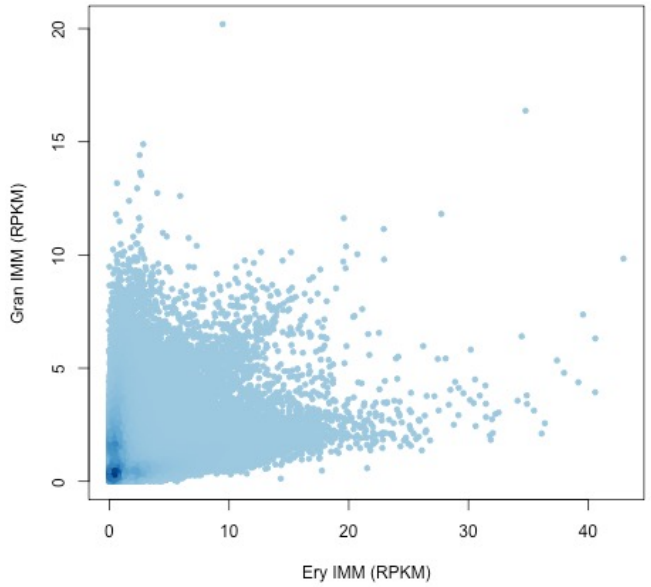
Genome browser track	Composite curve	Venn diagram	Volcano plot	tSNE/UMAP scatter plot
				
Heatmap	**Ripple heatmap**	**Hi-C contact map**	**Bubble plot**	**Circos plot**
				

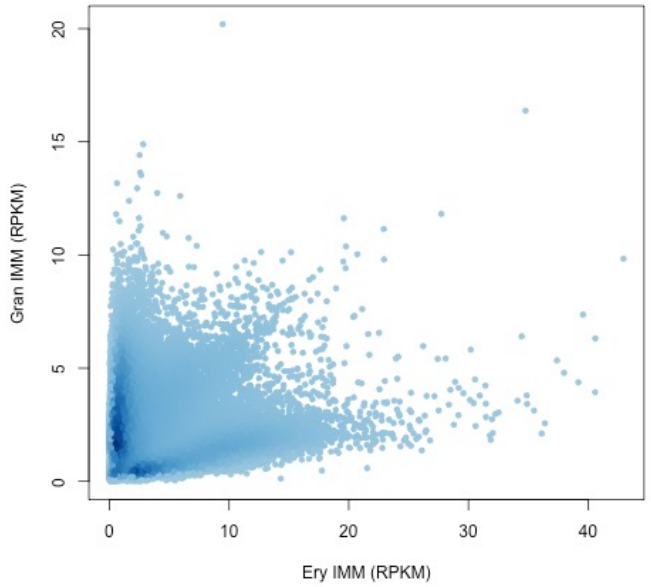
Some tips

Scatter plot

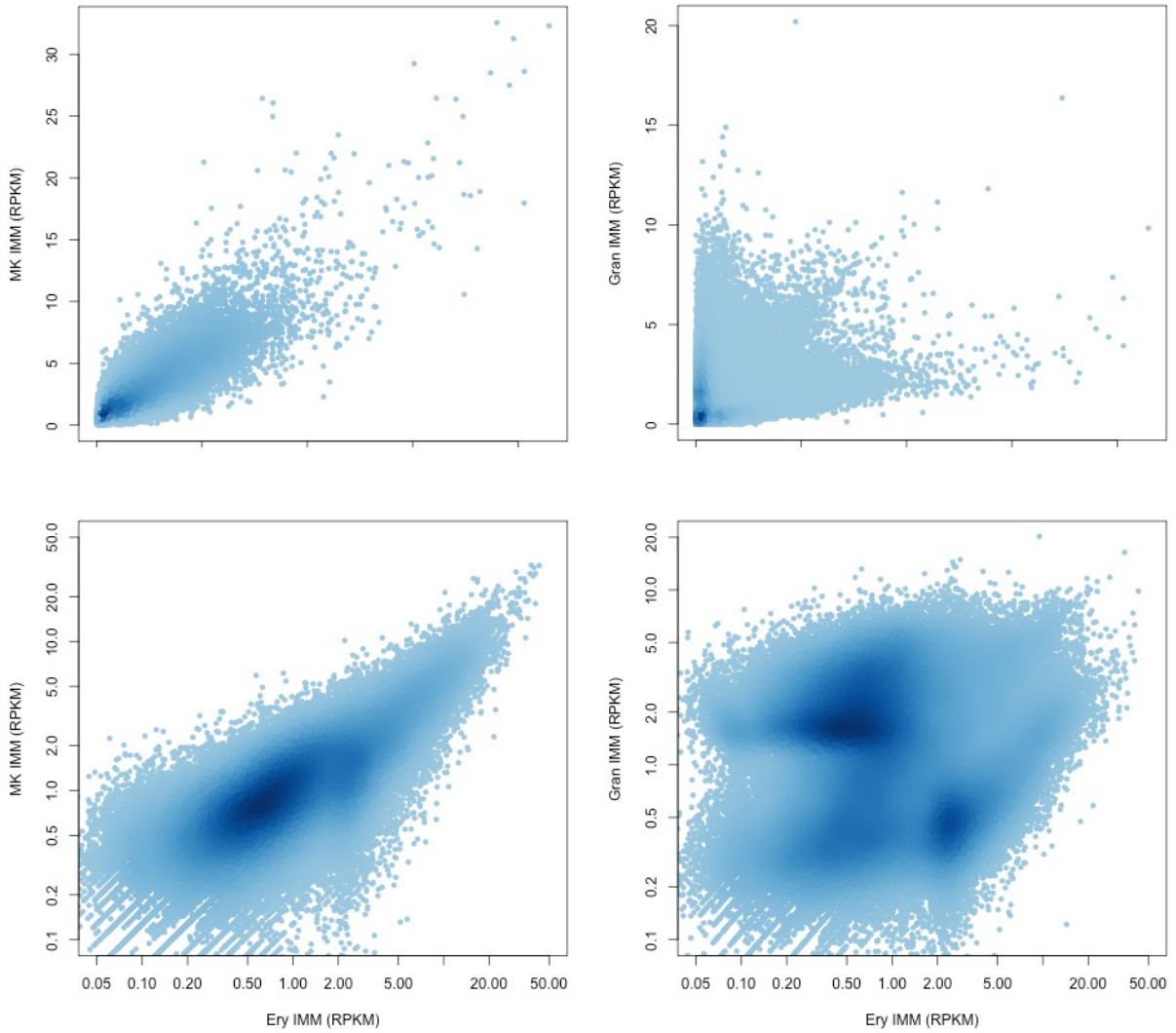
- The density of the data points matters!



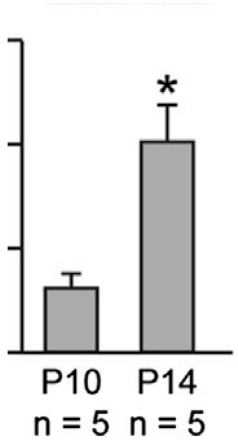


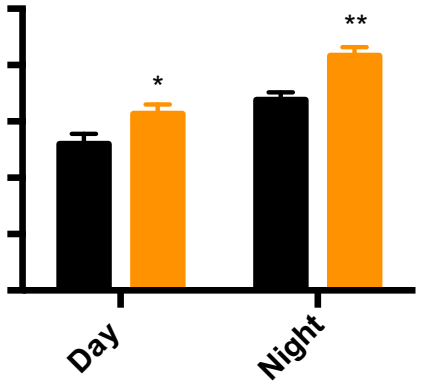


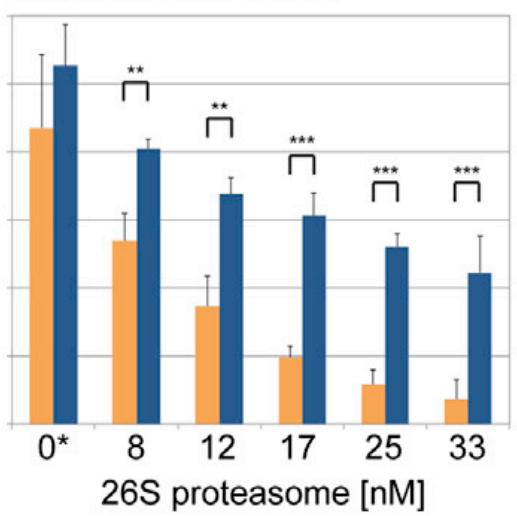
Choose the appropriate axis scale for plotting (linear or logarithm)



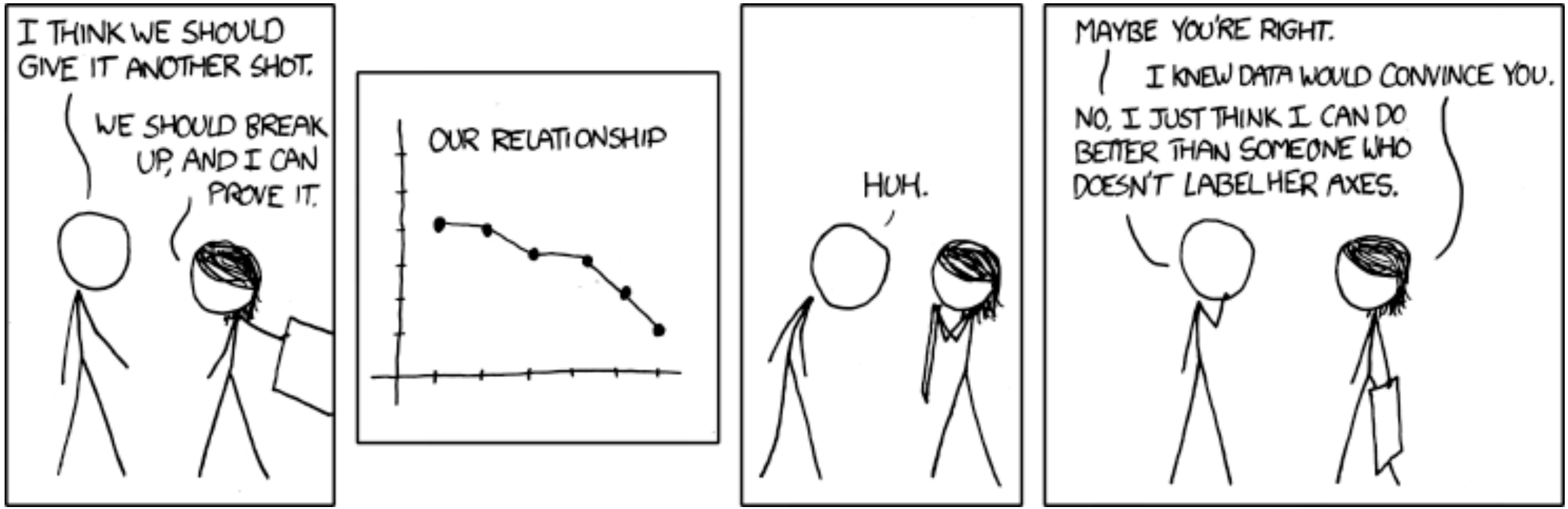
Can you spot any issue with these bar plots?



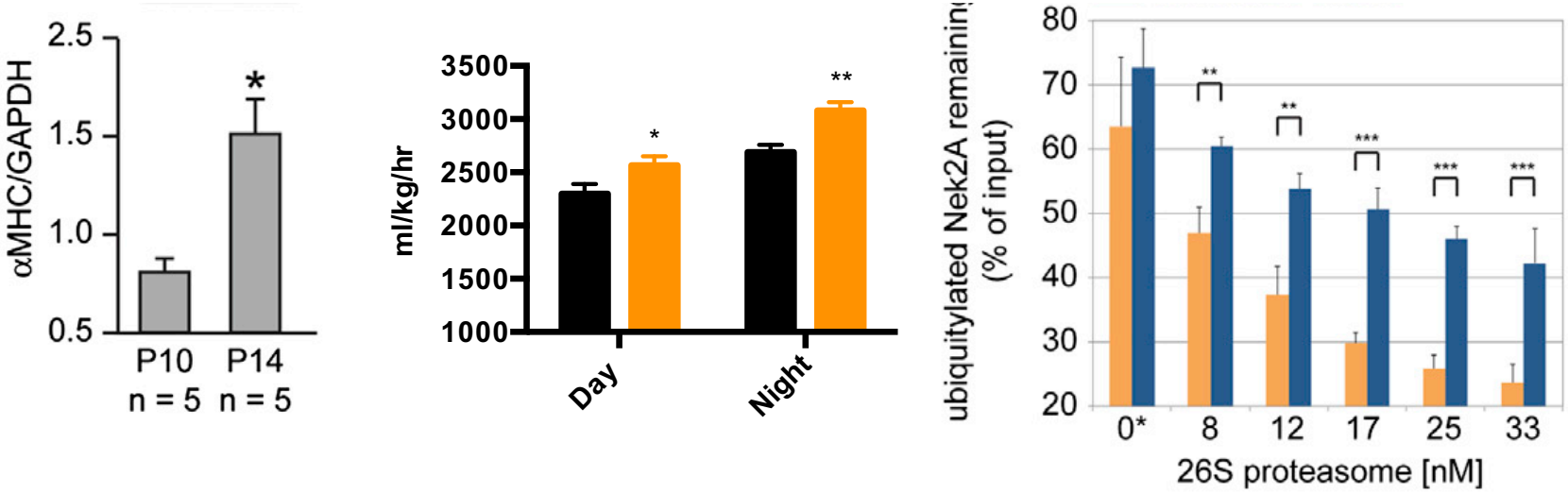




Plots from *Cell* 2014


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.

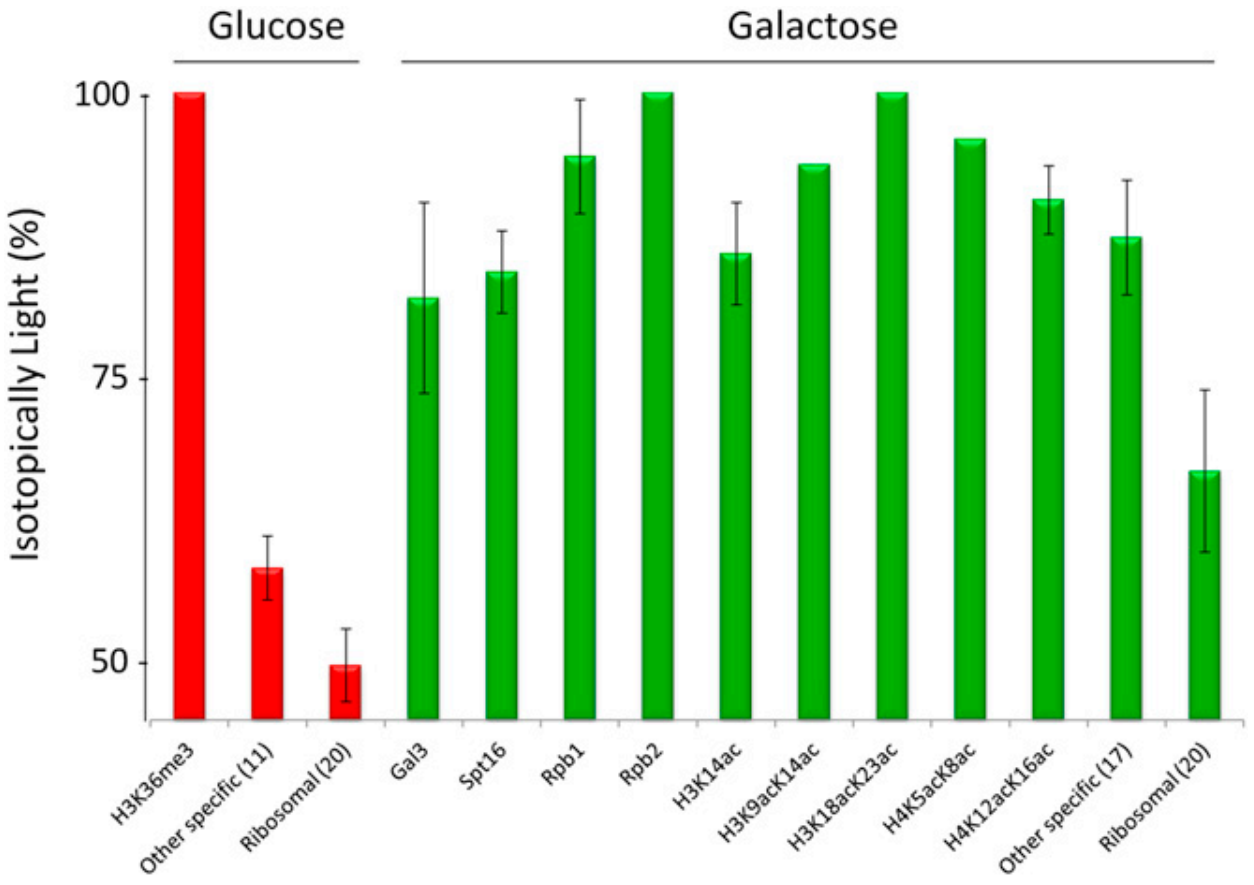
Can you spot any issue with these bar plots?



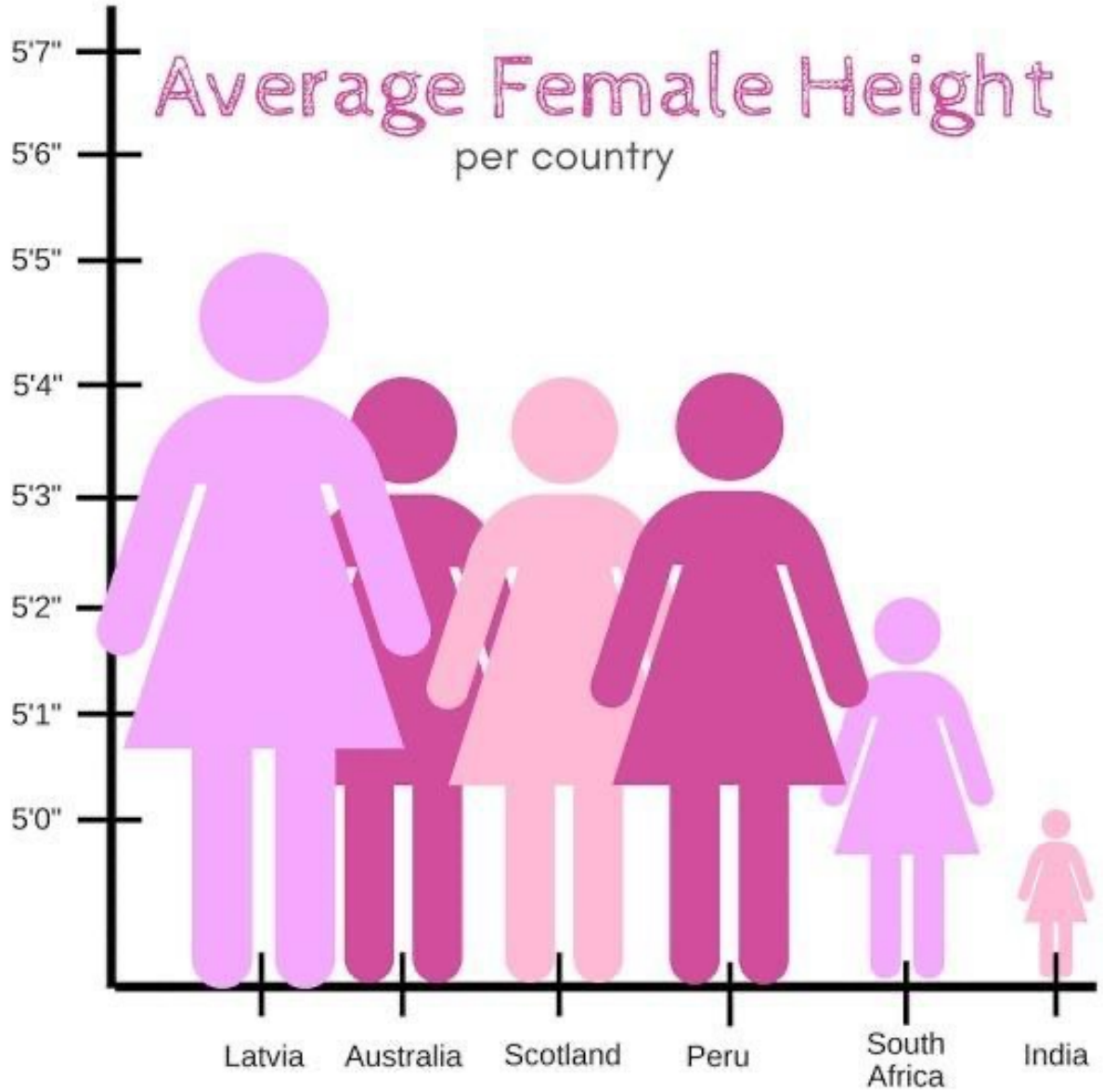
Bar chart y-axis should always start at 0, and on a linear scale.
If the difference is small, box plots or original data points are better.

Plots from *Cell* 2014

Another example

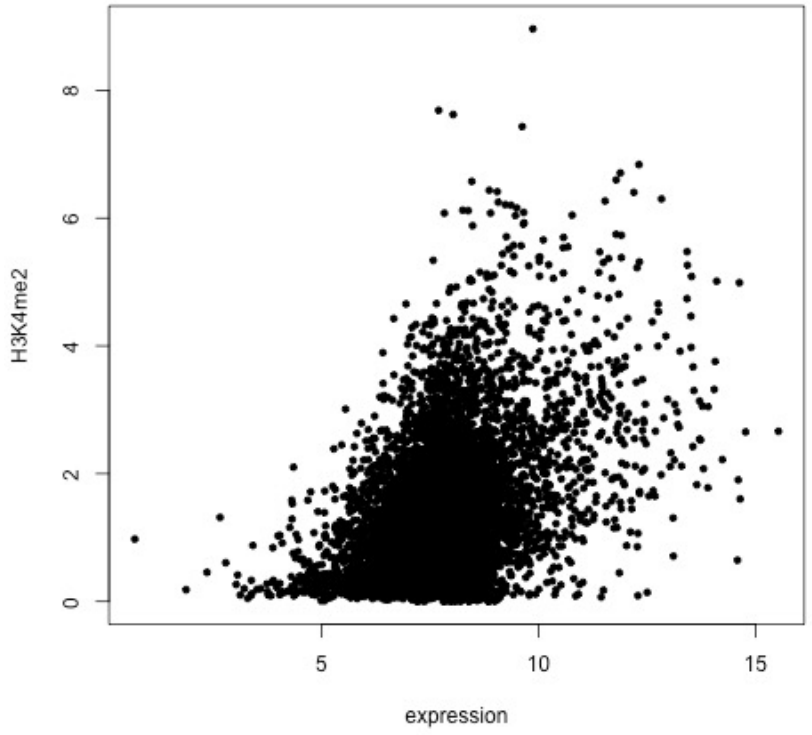


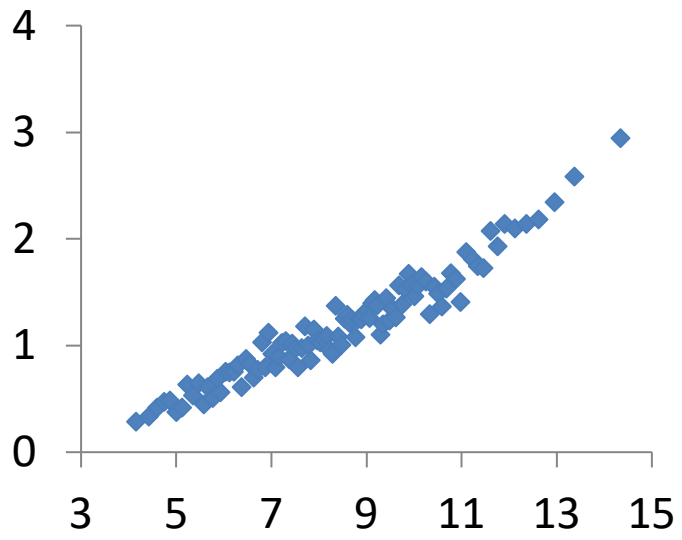
Byrum et al. *Cell Reports* 2012



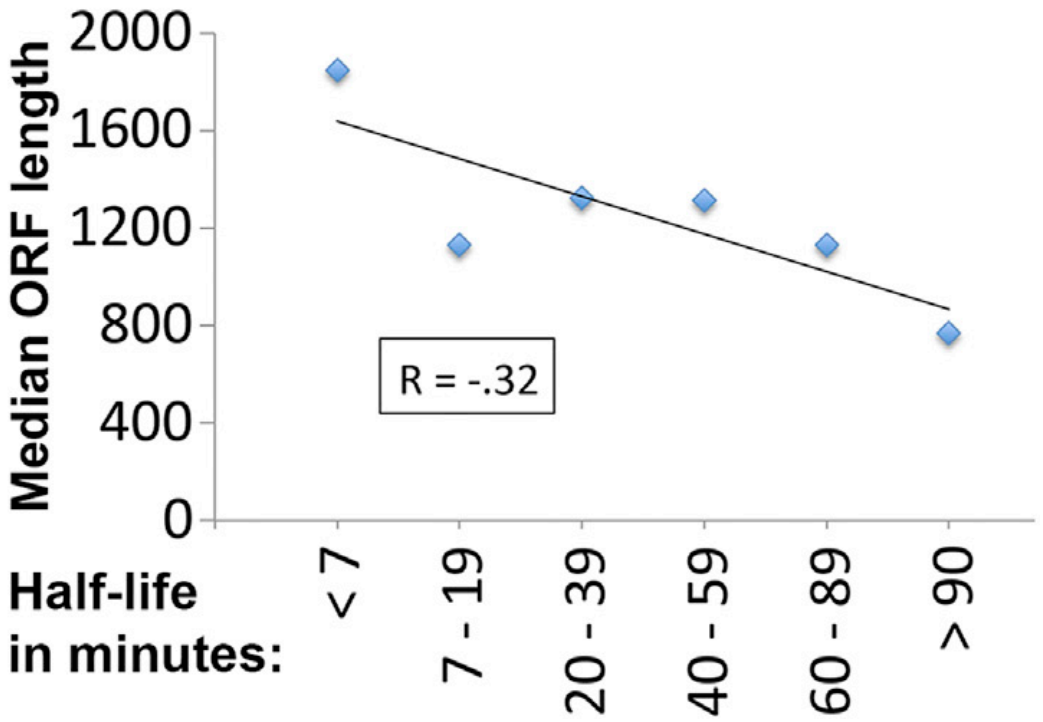
https://venngage.com/blog/bad-infographics/

Scatter plot? Group them if needed

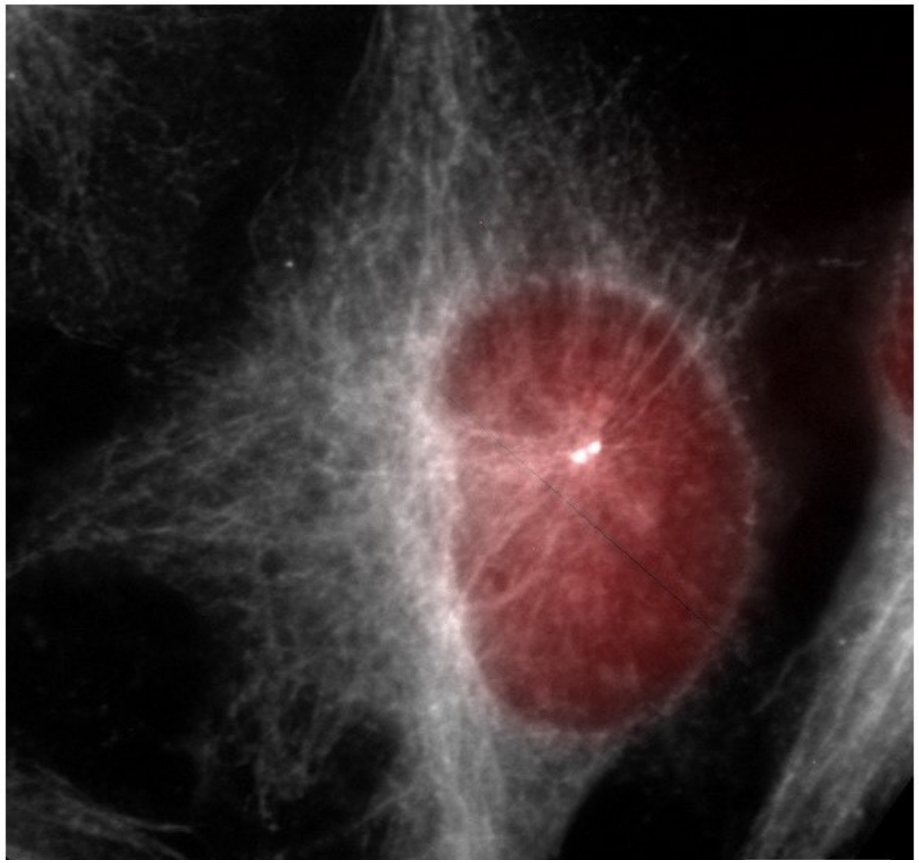




However, grouping should be fair/unbiased

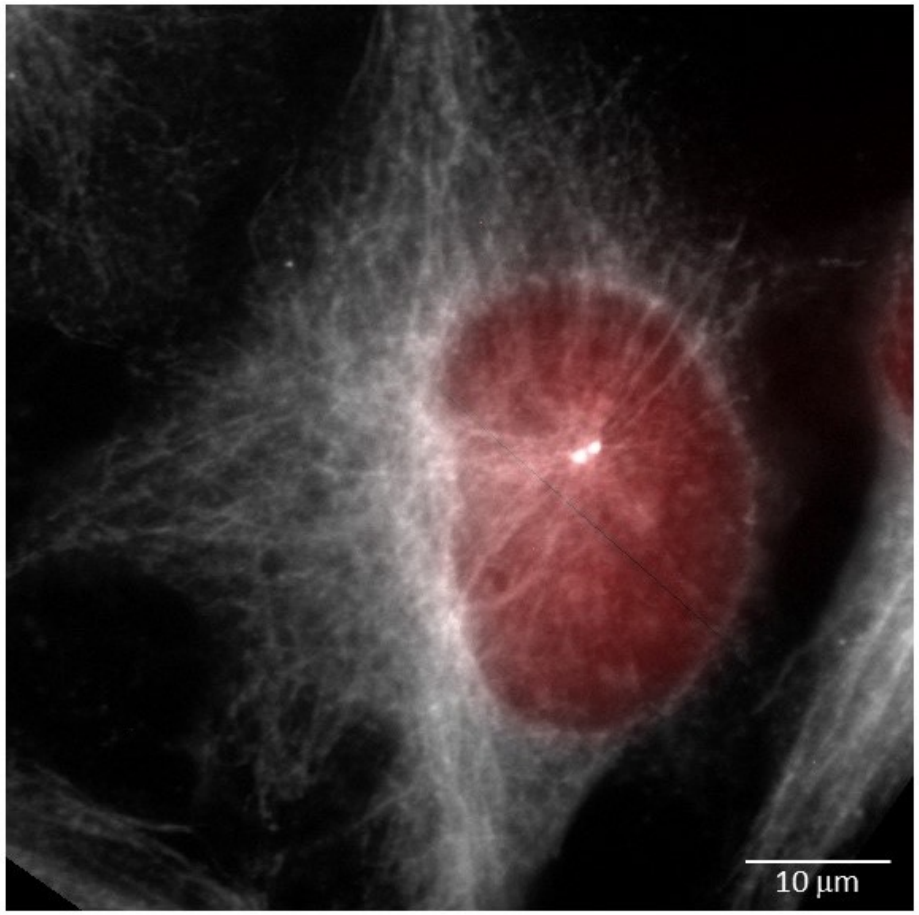


Cell 2014



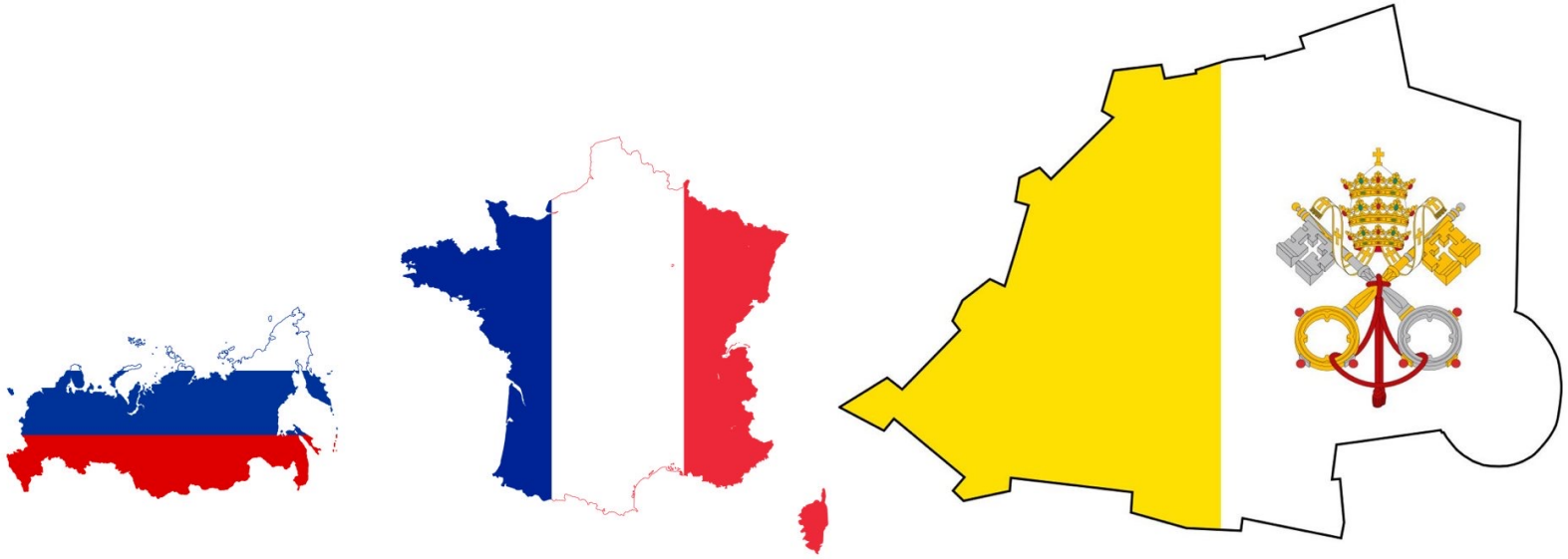
https://twitter.com/HarenLaurence

Scale matters!



https://twitter.com/HarenLaurence

Areas of Russia, France and Vatican compared*



*not to scale

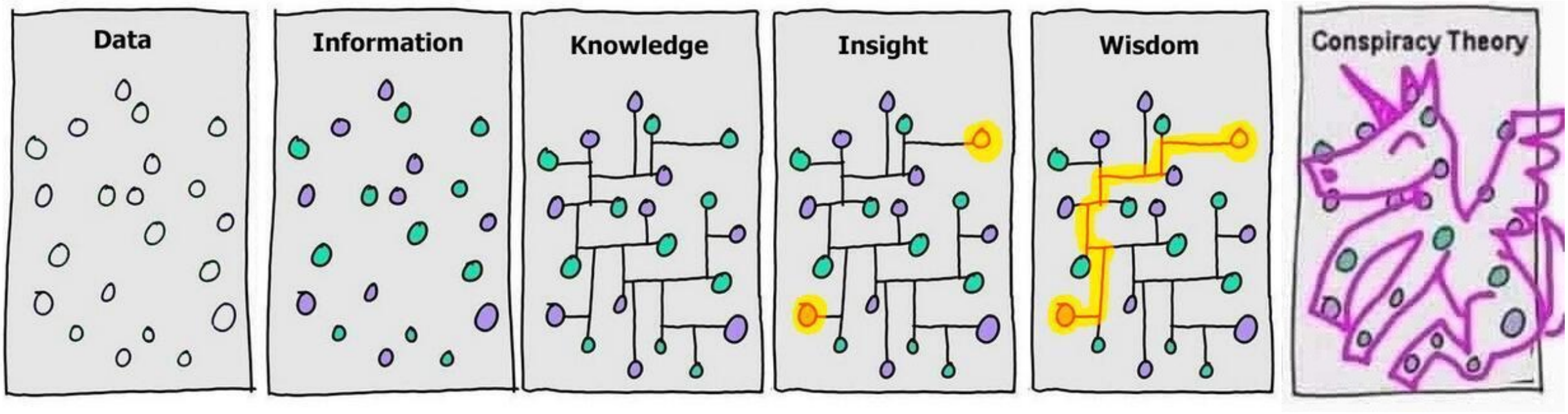
Scale matters. It does!

Twitter/X

Take-Home Tips

- Always check the axes and pay attention to the scale.
- Bar charts should always start at zero, on a linear scale.
- Pay attention to data point density in scatter plots.
- Group data points when needed, but do it in an unbiased way.

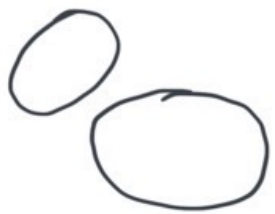
Having the data is not enough; presentation and interpretation matter



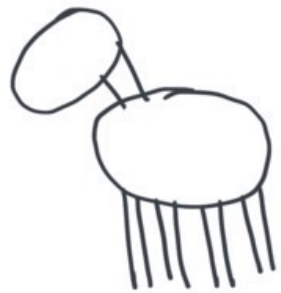
RNA-seq ChIP-seq DNase/ATAC-seq Hi-C Single-cell resolution… …	Gene expression Protein factors Chromatin 3D genome Multi-omics …	Transcriptional regulation, Chromatin organization, …	New insight?	New biology!	Overfitting, overinterpretation...

HOW TO:
DRAW A HORSE

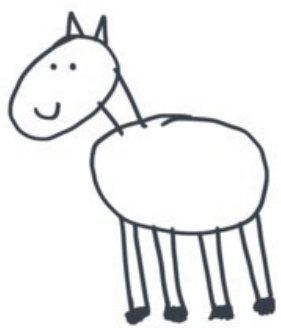
BY VAN OKTOP



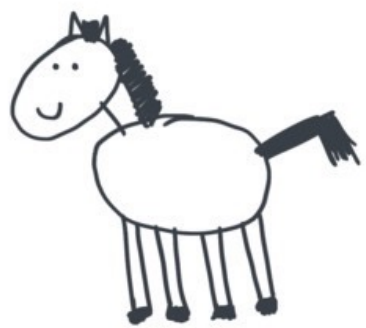
(1) DRAW 2 CIRCLES



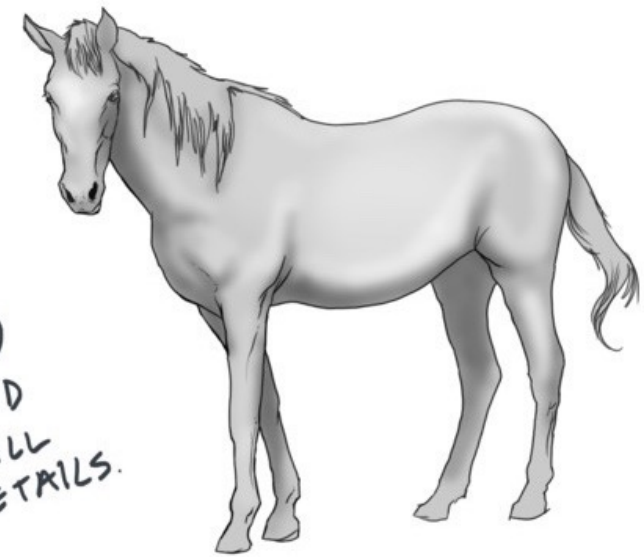
(2) DRAW THE LEGS



(3) DRAW THE FACE



(4) DRAW THE HAIR



(5) ADD SMALL DETAILS.

Thank you very much!

zang@virginia.edu

zanglab.org