Getting at the biological question through topology comparisons

Cory Padilla, Ph.D.
Dovetail Genomics

Preface

Biology biology biology

- What is your biological question?
- Do you have the correct data to answer that question?
- Informatic tools pipelines are a means to an end, make sure the tool your using will help you

There is no "one way"

- There are many tools and many ways to work with data
- Pick a route, give it a go, adjust from there

File structure is paramount

- Learn the format of commonly used data types
- Inspect inputs and outputs
- Get familiar with the data, so when you get asked to do something a tool doesn't do, you know how to get started

Comparisons Can Be Hard…

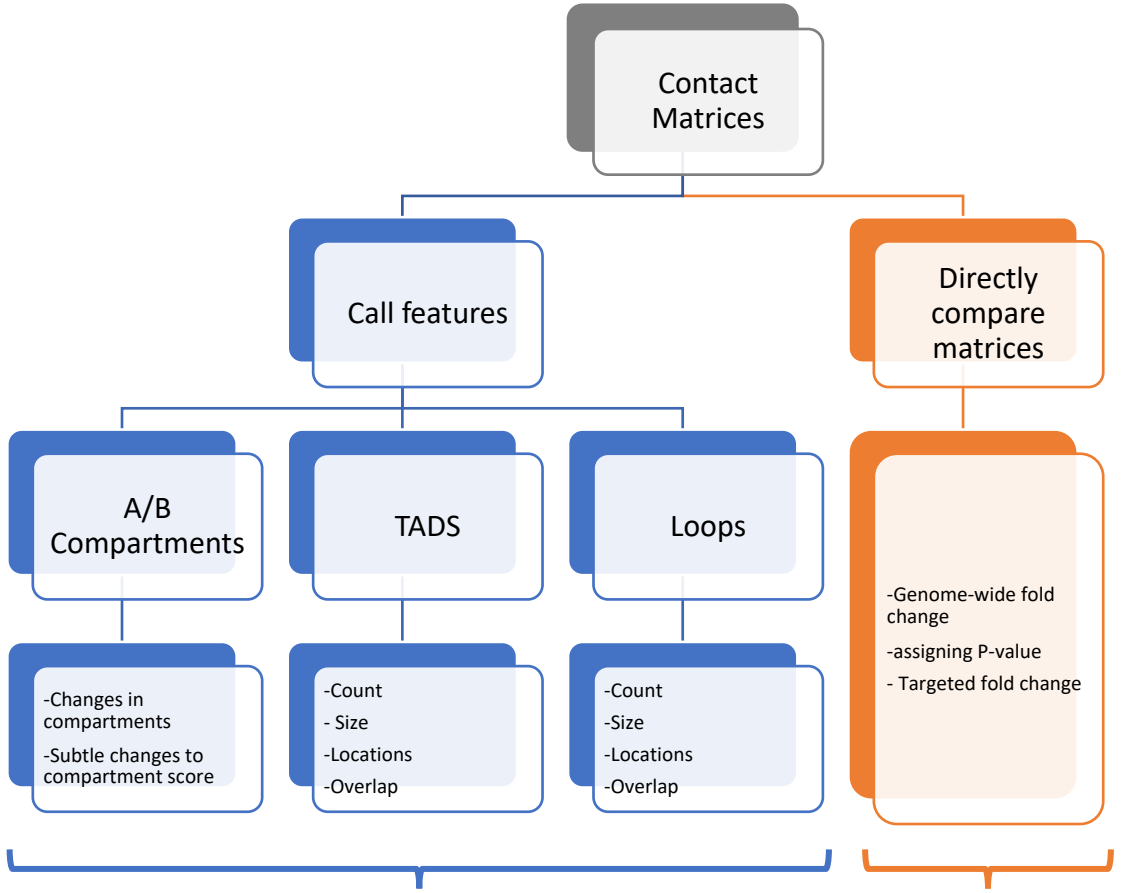
Keep in mind what you are comparing

- Contact matrices
- Particular locus/loci
- Topological features

By what approach(s) do you want to compare

- Fold change
- Feature locations
- Do you have other data to integrate?

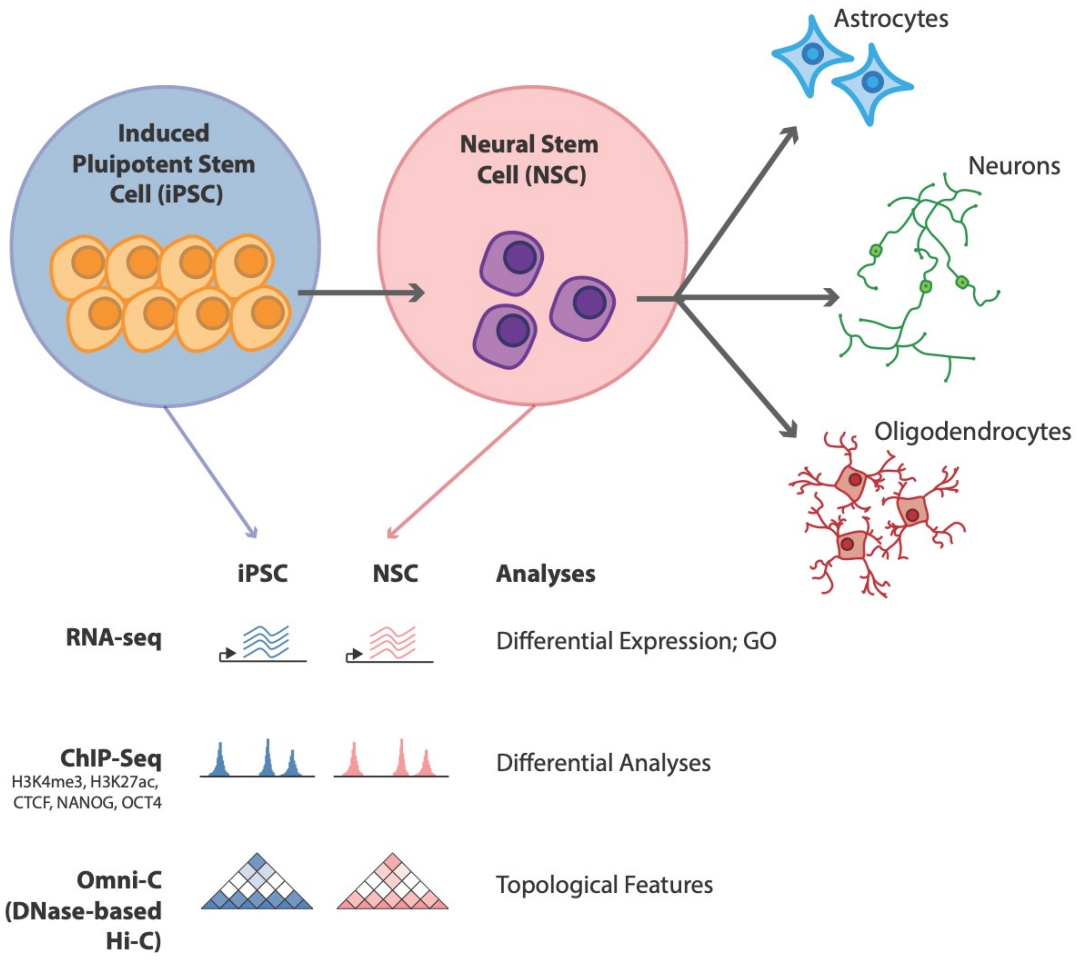
Two fundamental routes to topology comparisons

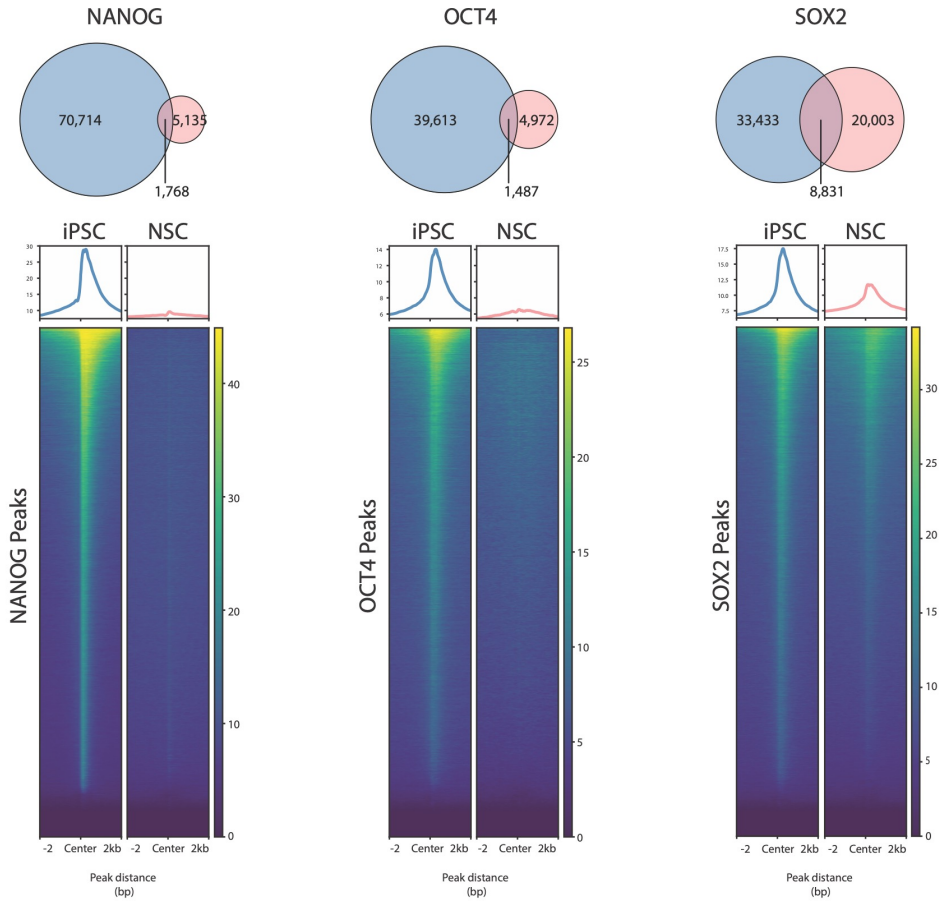


- Straight forward
- Lots of tools
- Less flexible

- Can be complex
- Fewer tools
- Very flexible

Case Study: Neuronal Development



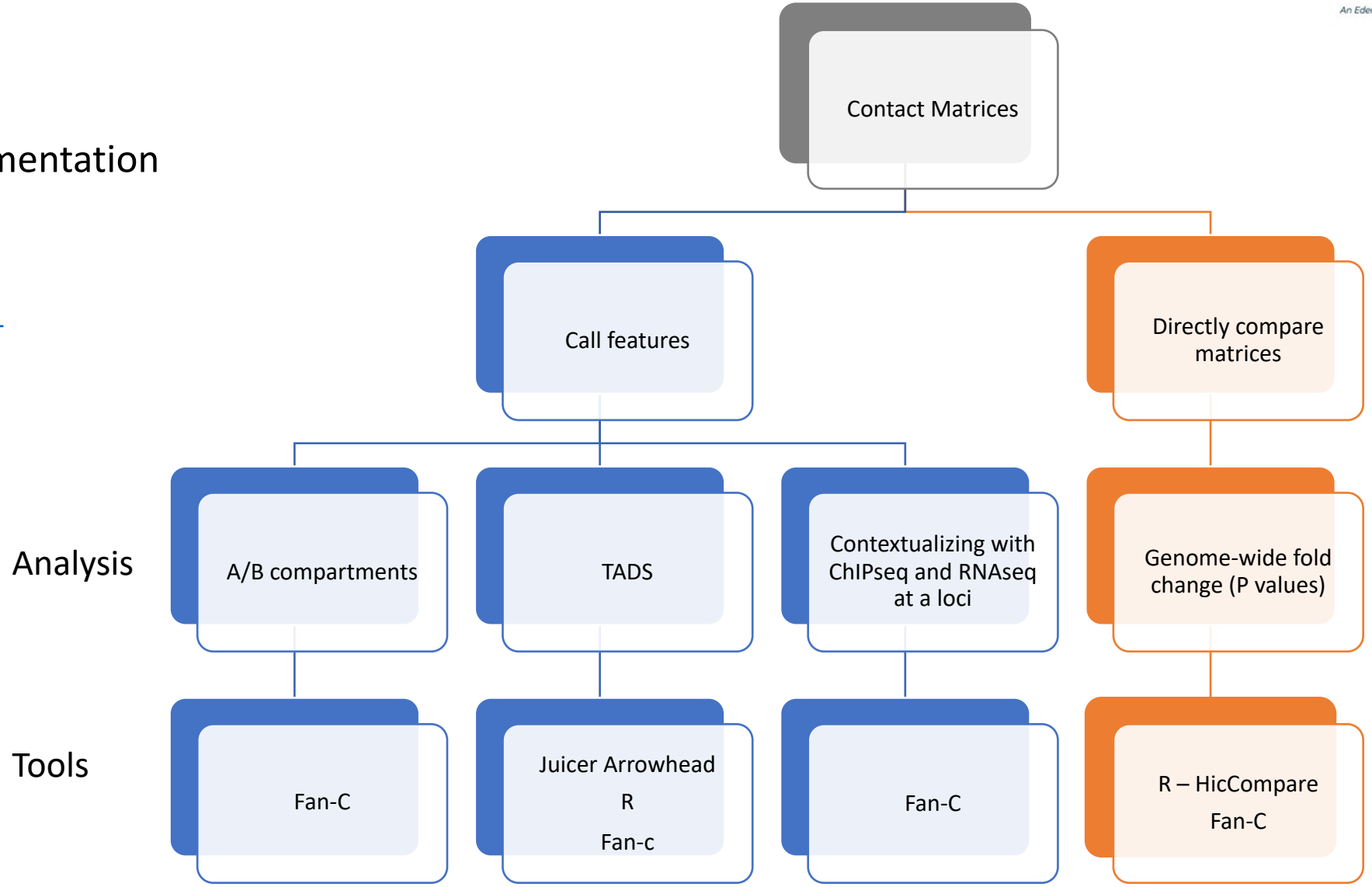


Biological question – Can we link the local topology to the loss of NANOG expression between iPSC and NSC?

How we're going to look at these samples

Tool Repos or Documentation

- R
- Fan-c
- Juicer Arrowhead
- HiCCompare



A/B compartments

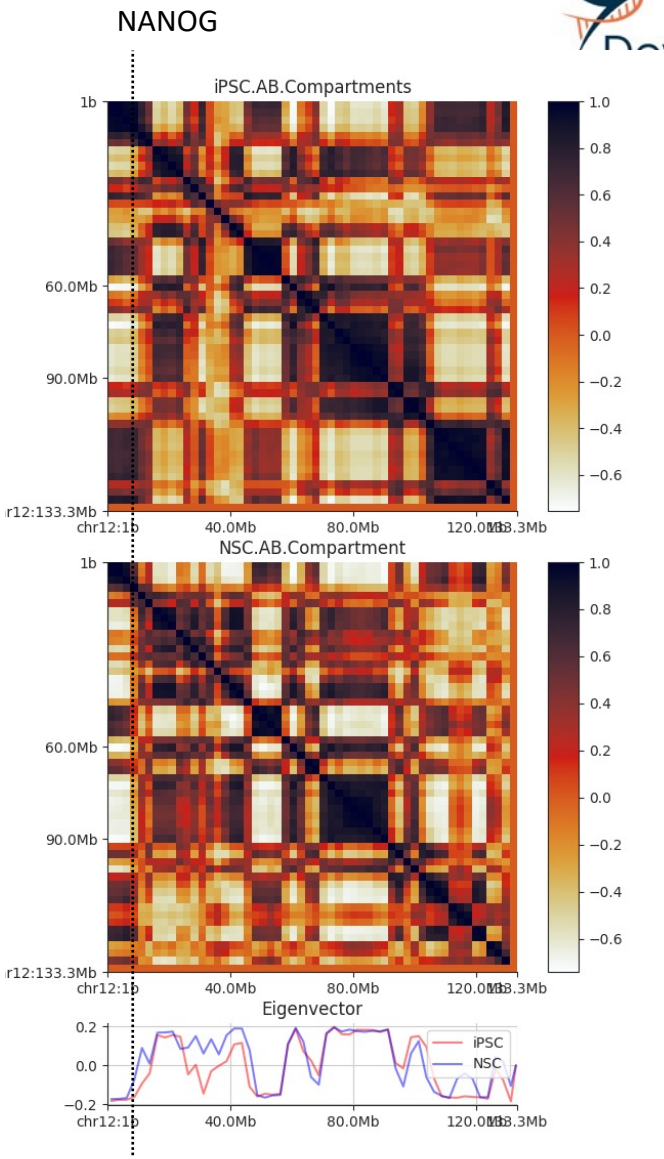
- Input files: Matrices in .hic format
- Tools: fanc

- Command calculate compartment matrix and eigenvectors with fanc

```
fanc compartments -v nsc.ev.txt nsc.merged.hic nsc.ab
fanc compartments -v ipsc.ev.txt ipsc.merged.hic opsc.ab
```

- Plot

```
fancplot -o AB_compare.png \
chr12 \
-p square ipsc.ab --title iPSC.AB.Compartments \
-p square nsc.ab --title NSC.AB.Compartment \
-p line ipsc.ev.txt nsc.ev.txt -l iPSC NSC --title Eigenvector
```



The compartment signal at NANOG is different between iPSC and NSC

TAD Calling

Dovetail GENOMICS
An EdenRoc Sciences Company

Calling

- Input: .hic
- Tool: Juicer Arrowhead
- Command: calling TADs at 25 kbp with Arrow head

```
#Call TAD boundaries with juicer
java juicer ~pathto/juicer.jar arrowhead -r 25000 -k KR ipsc.hic -o output_directory
java juicer ~pathto/juicer.jar arrowhead -r 25000 -k KR nsc.hic -o output_directory
```

- The output is a bedpe file where region 1 and region 2 are the TAD boundaries, we want to plot the entire TAD region, so we need to get the start of region 1 and end of region 2 in order to plot the length of the TAD and characterize them

```
#convert TAD calls to bed format
cut -f 1,2,6 ipsc.TADs.25kb.bedpe > tad.regions.bed
cut -f 1,2,6 nsc.TADs.25kb.bedpe > tad.regions.bed
```

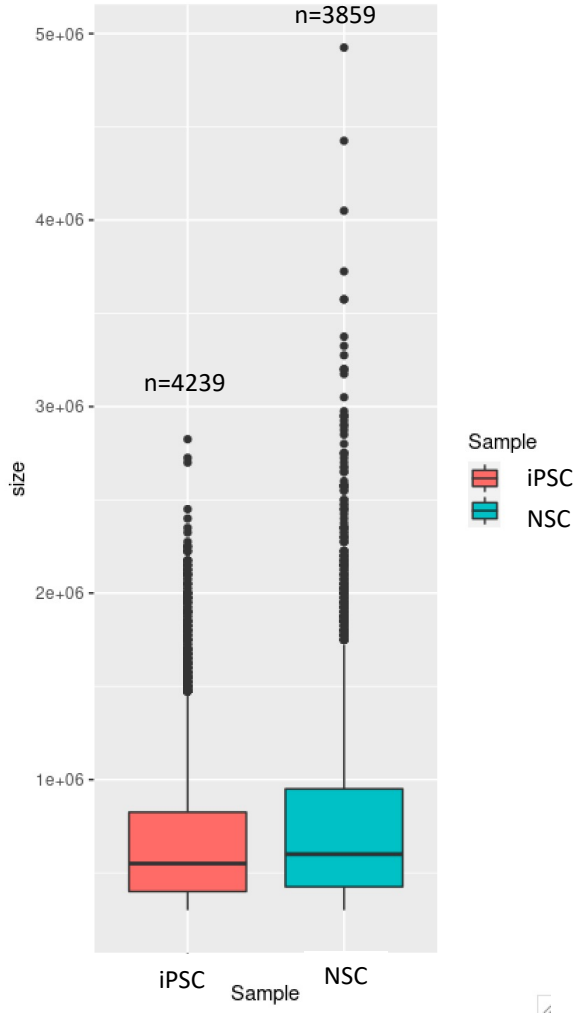
TAD characterization

Characterizing in R

Tools: R

Command:

```
#load libraries
Library(ggplot2)

#load data
ipsc <- read.table("ipsc.tad.regions.bed")
nsc <- read.table("nsc.tad.regions.bed")

#add column for distance
ipsc$dist <- ipsc$V3 - ipsc$V2
nsc$dist <- nsc$V3 - nsc$V2

#count number of TADs
nrow(ipsc) 4239
nrow(nsc) 3859

#charactize TADs size
summary(ipsc$dist) Mean = 683,917 bp
summary(nsc$dist) Mean = 767,459 bp

#add column for sample ID and merge
ipsc$sample <- "ipsc"
nsc$sample <- "nsc"
dat <- rbind(ipsc, nsc)

#plot
ggplot(dat, aes(x=sample, y=dist, fill=sample)) + geom_boxplot()
```



Conclusion we can draw from this:

@25 kbp

iPSC has more – smaller TADs

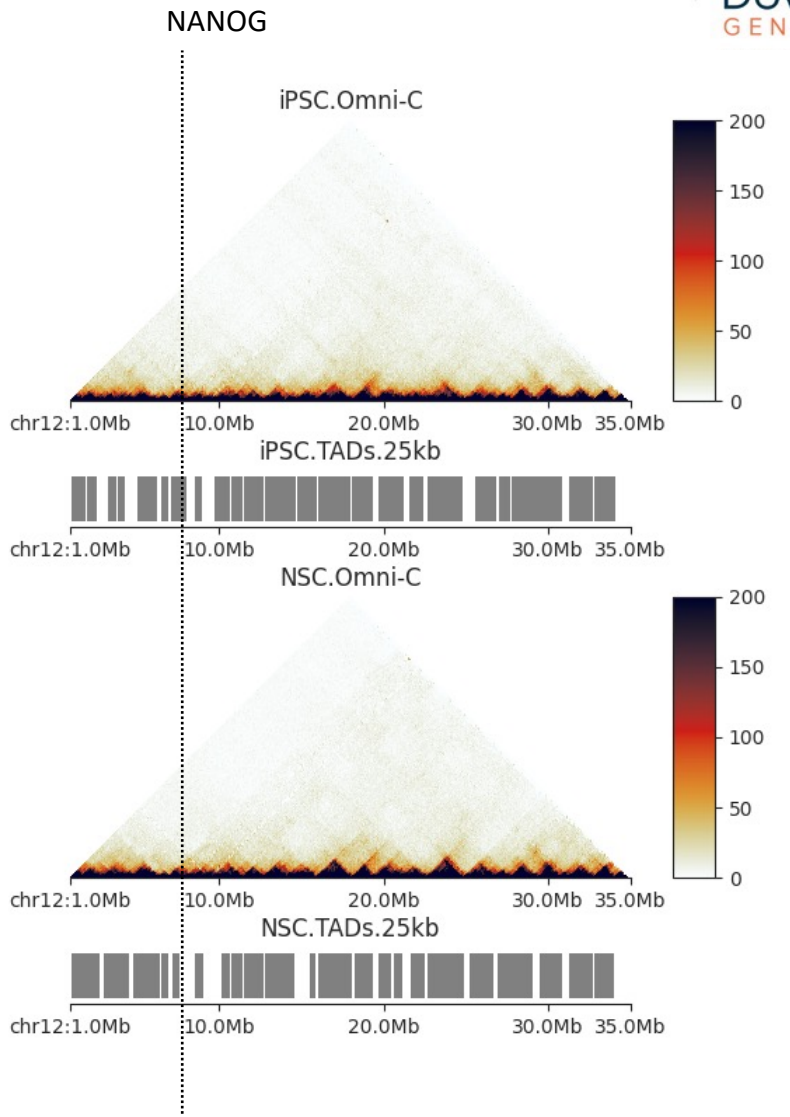
NSC has fewer TADs, but they are larger

TADs –Plotting

Plotting

- Inputs: .hic and TAD bed
- Tools: Fan-C
- Commands

```
fancplot -o TAD_final.png \
chr12:1mb-35mb \
-p triangular o.ipsc.merged.hic@100kb -vmax 200 --title iPSC.Omni-C \
-p layer oipsc.tad.regions.bed --title iPSC.TADs.25kb \
-p triangular o.nsc.merged.hic@100kb -vmax 200 --title NSC.Omni-C \
-p layer nsc.tad.regions.bed --title NSC.TADs.25kb
```



NANOG occurs in a TAD in iPSC, but not in NSC

Contextualizing with other markers

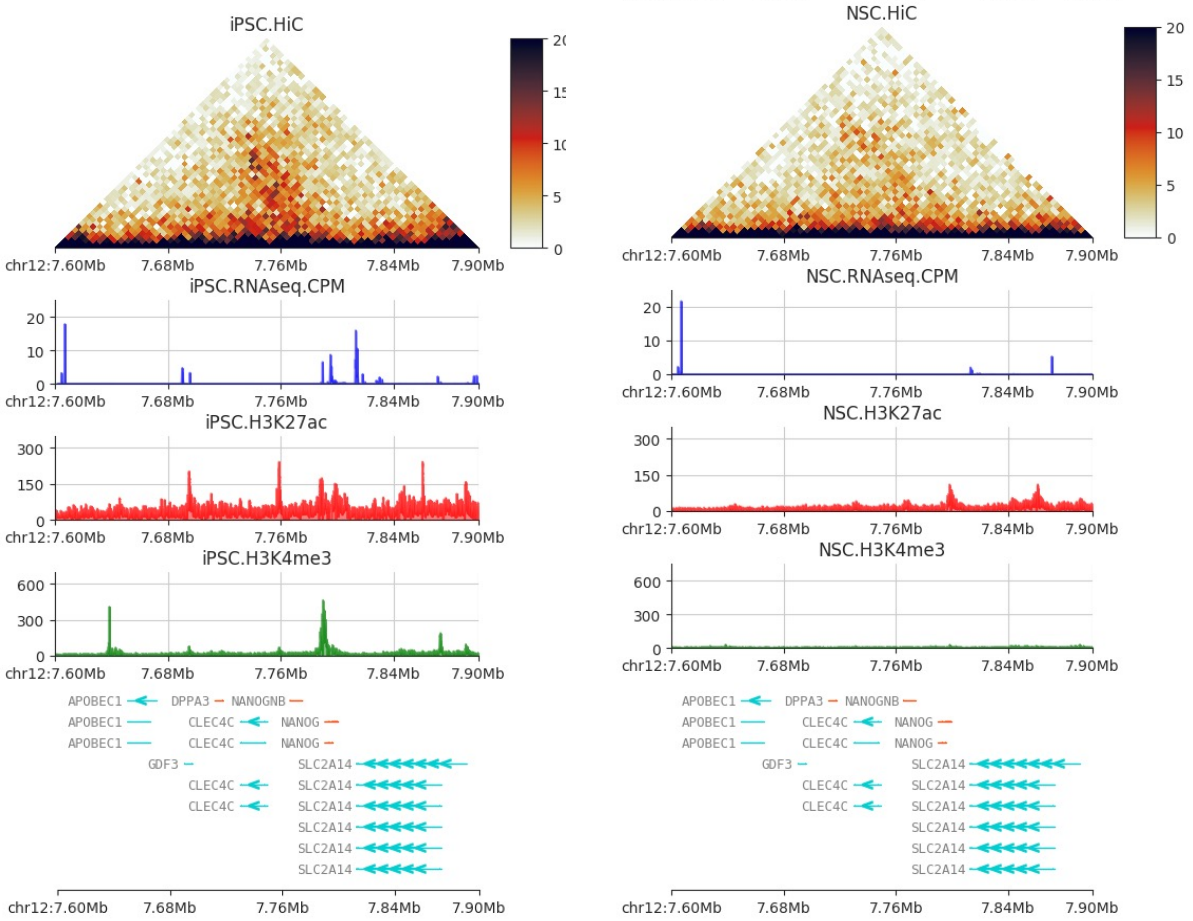
Inputs:

- .hic
- RNAseq: bigwigs
- ChIP-seq: bigwigs
- Gene: GTF

Tools: Fan-C

Commands

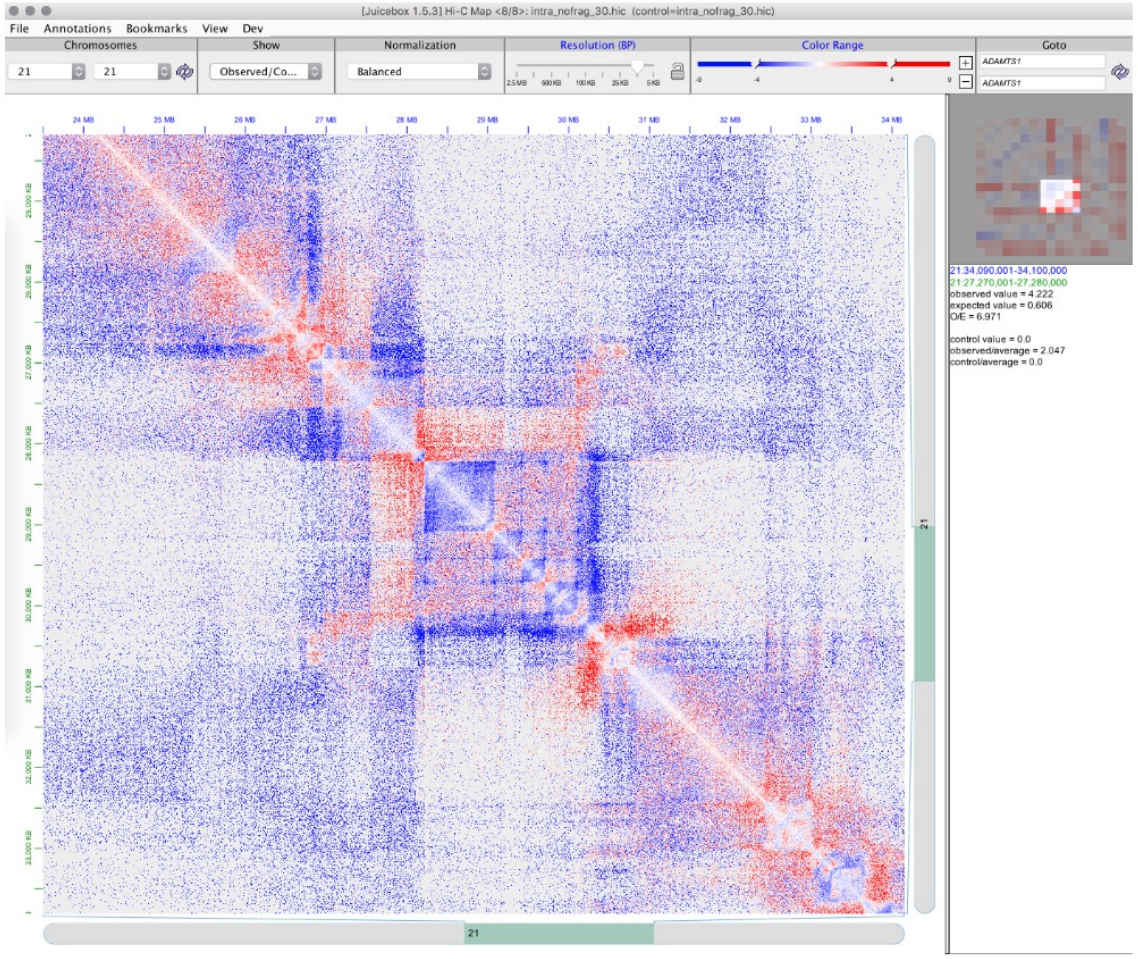
```
fancplot -o zoomed_tracks.test.png \
chr12:7.6mb-7.9mb \
-p triangular o.ipsc.merged.hic@5kb -vmax 20 --title iPSC.HiC \
-p line ipsc_rna_seq.chr12.bigwig -c blue -y 0 25 --title iPSC.RNAseq.CPM
-p line ipsc.H3K27ac.bw -c red -y 0 350 --title iPSC.H3K27ac \
-p line ipsc.H34me3.bw -c green -y 0 700 --title iPSC.H3K4me3 \
-p gene hg38.refGene.gtf
```



Loss of promoter signal and genal loss of contacts between enhancers and promoters in NSC at NANOG

Diff-ing matrices

Select `Observed/Control` to show relative enrichment between the maps.



The Juicebox has a GUI to do this, but let's add some P-values to the image shall we?

Comparing matrices with HiCCompare in R

- Preparing the data from a cool file:

```
cooler dump --join sample.25kb.cool > sample.25kb.cool.txt
```

- Inputs: Upper sparse matrices (the smaller the bin size, the greater the computation time)
- Tools: R and Bioconductor package: HiCCompare
- Commands

```
library(HicCompare)
library(ggplot2)
```

Load the libraries

```
ipsc <- read.table("ipsc.25kb.cool.txt")
nsc <- read.table("nsc.25kb.cool.txt")
```

Load the data

```
ipsc.intra <- ipsc[ipsc$V1 == ipsc$V4,]
nsc.intra <- nsc[nsc$V1 == nsc$V4,]
```

Select for cis interactions only

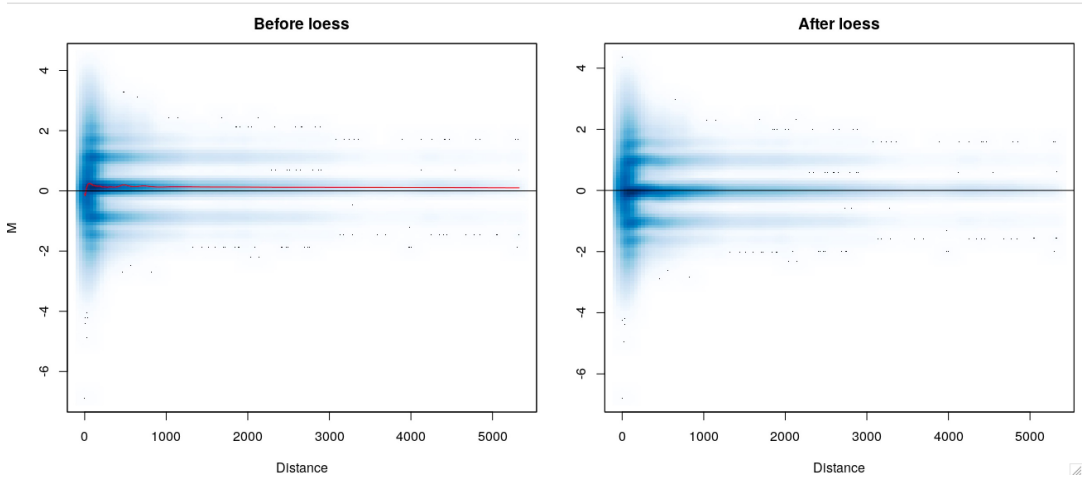
```
ipsc.chr12 <- ipsc.intra[ipsc.intra$V1 == "chr12",]
nsc.chr12 <- nsc.intra[nsc.intra$V1 == "chr12",]
```

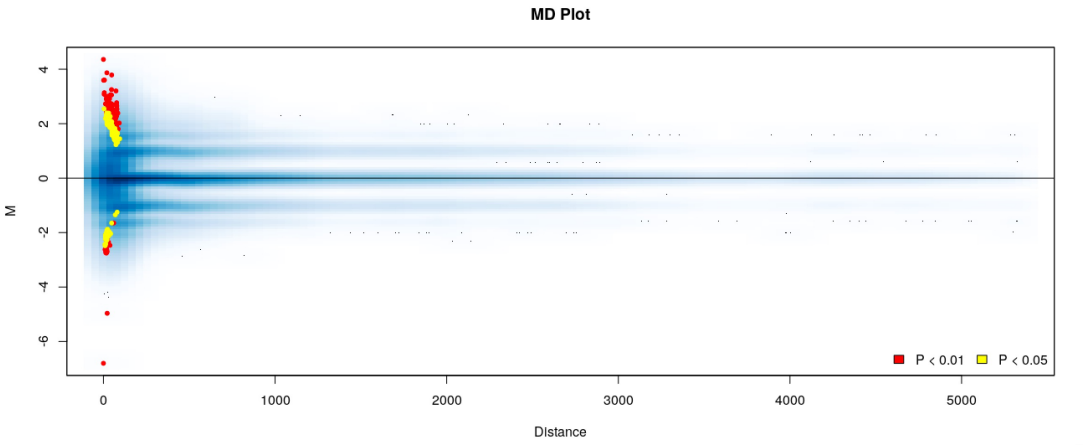
Select for chr12 (where NANOG lives)

Normalizing and assigning a P-Value

Process

```
#Merge tables

combine <- create.hic.table(ipsc.chr12, nsc.chr12, chr = 'chr12')


#Normalize

hic.table <- hic_loess(combine, Plot = TRUE, Plot.smooth = TRUE)


#find sig diffs

hic.table <- hic_compare(hic.table, A.min = 15, adjust.dist =
TRUE, p.method = 'fdr', Plot = TRUE)


#print resulting table

write.csv(hic.table, "chr12.compare.csv", row.names=FALSE)
```

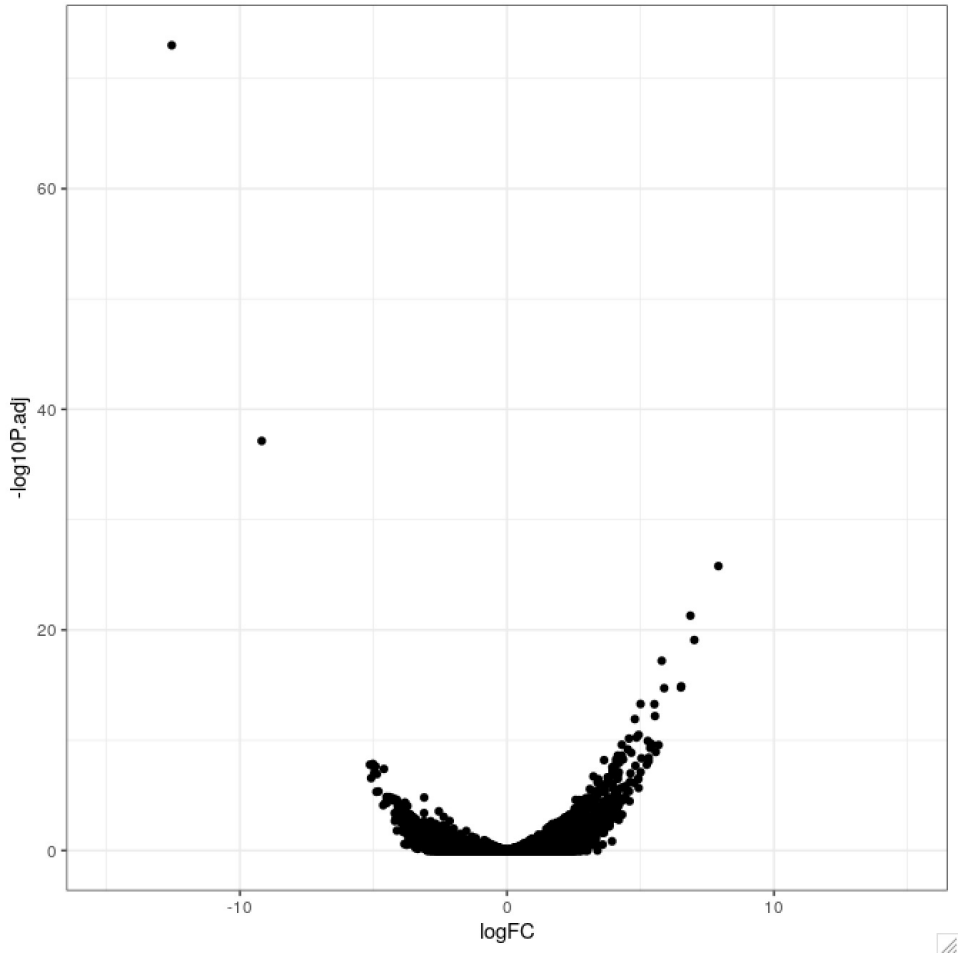




Comparing matrices – something familiar

Plot as volcano plot

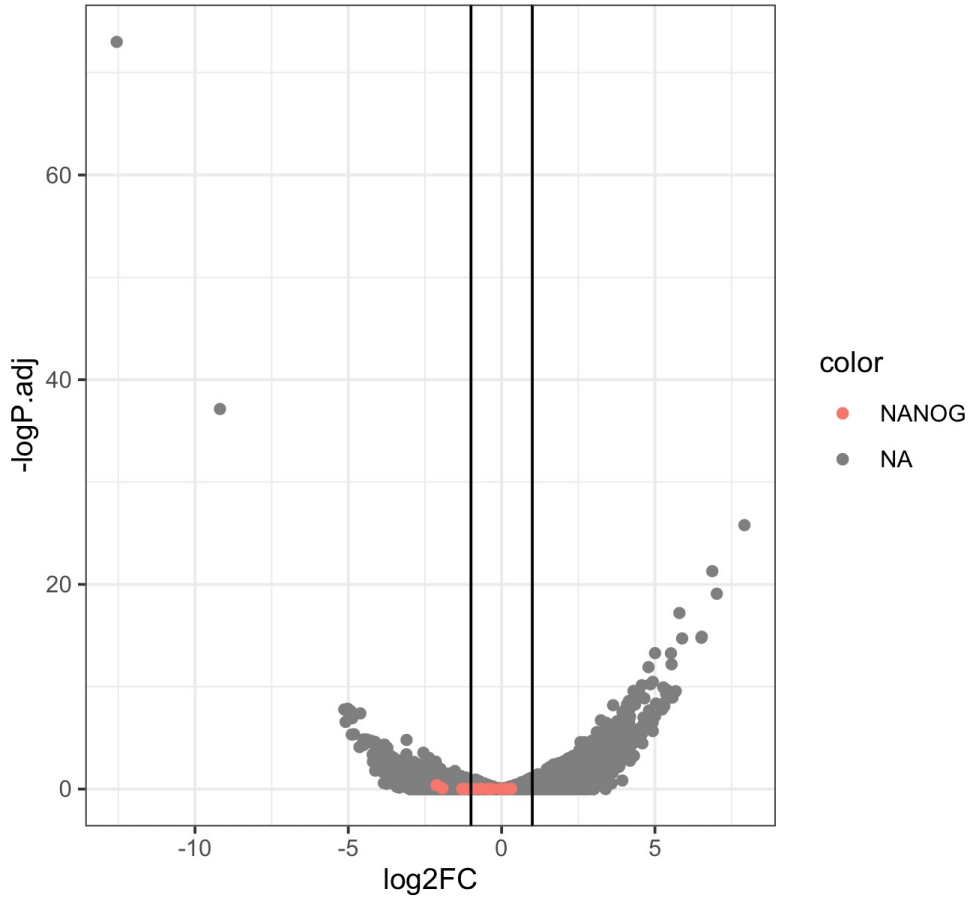
```
#Add a column for –log(p)
dat$log.p.add <- -log(dat$p.adj)

ggplot(dat, aes(x=Z, y=log.p.add)) +
         geom_point() + theme_bw() +
         xlim(-15, 15) +
         labs(x="logFC", y ="-log10P.adj")
```



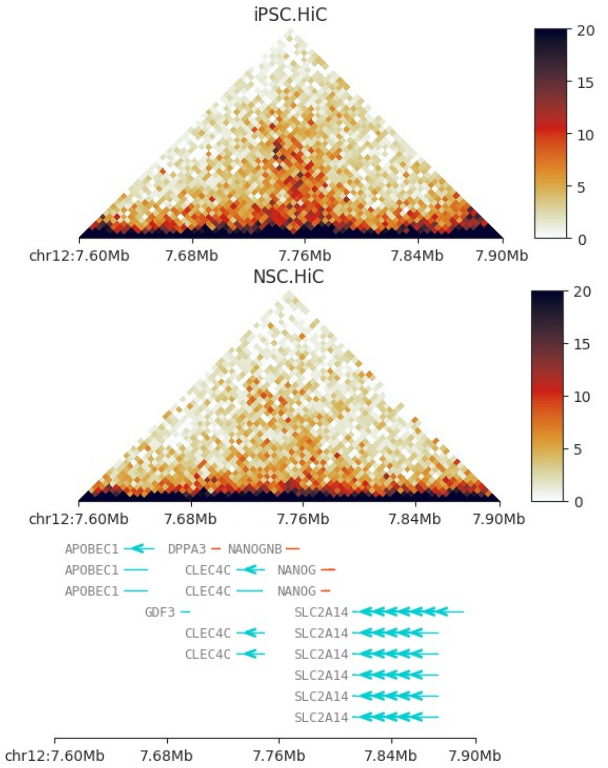
So what about NANOG?

- You can annotate the bins by ROI
- The differences at NANOG do not seem that drastic
- Remember resolution:
  - 25kb vs 5kb

Statistical Difference at 25 kb



Difference at 5 kb



Log fold change matrix with fan-c

Inputs:

- .hic from each sample

Tools: Fanc

Generate log fc matrix (run at both 25 and 5 kb)

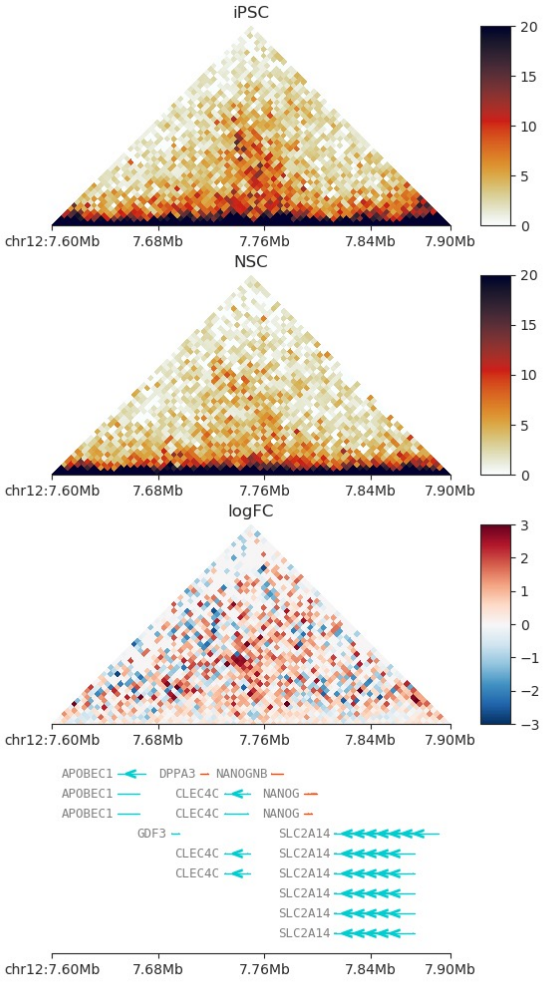
```
fanc –compare -l –Z –I o.ipsc.merged.hic@25kb o.nsc.merged.hic@25kb
logfc.25kb.compare.matrix
```

Generate log fc matrix (run at both 25 and 5 kb)

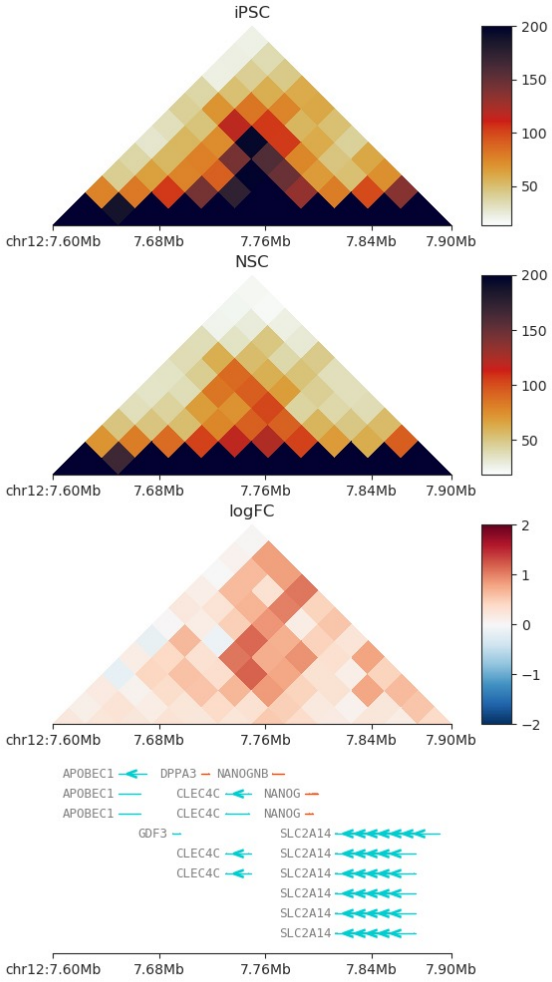
```
Fancplot –o compare.25kb.png –p triangular o.ipsc.merged.hic@25kb  -p
triangular o.nsc.merged.hic@25kb -p triangular
logfc.25kb.compare.matrix –c RdBu_r
```

Log fold change matrix with Fan-C

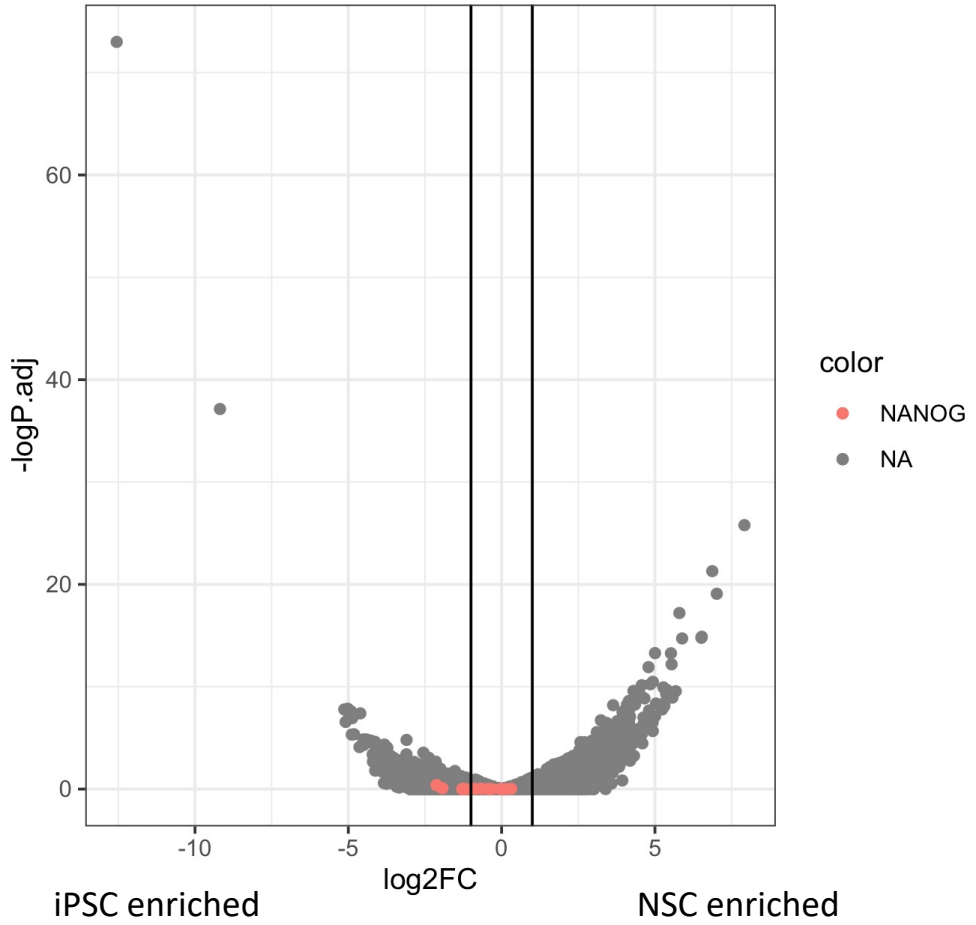
5kb



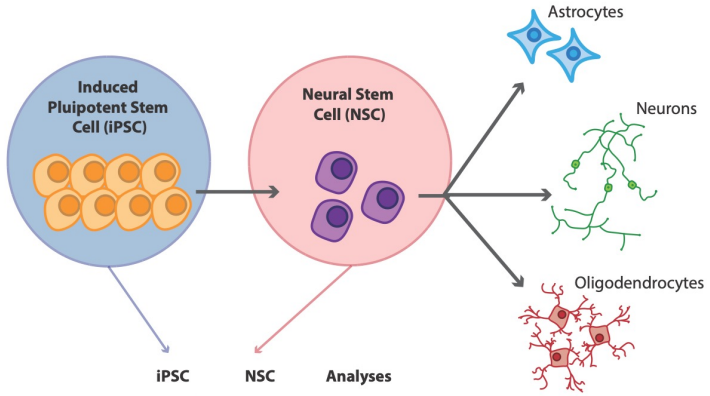
25kb

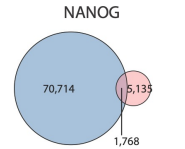


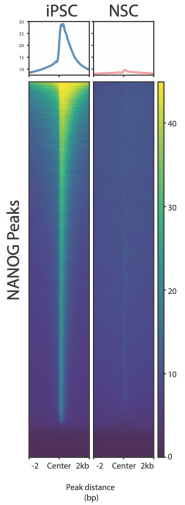
Statistical Difference at 25 kb



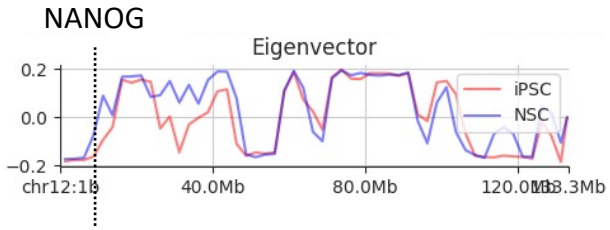
NANOG Topology and loss of pluripotency: Summary





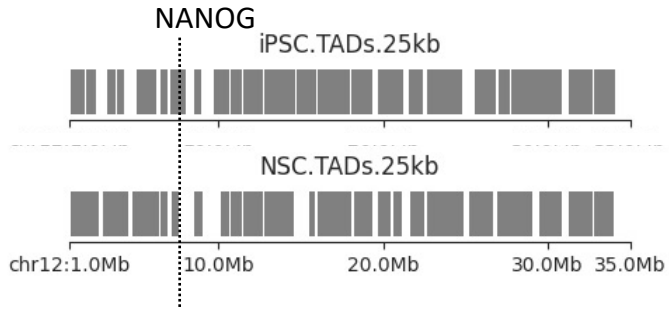


1Mb



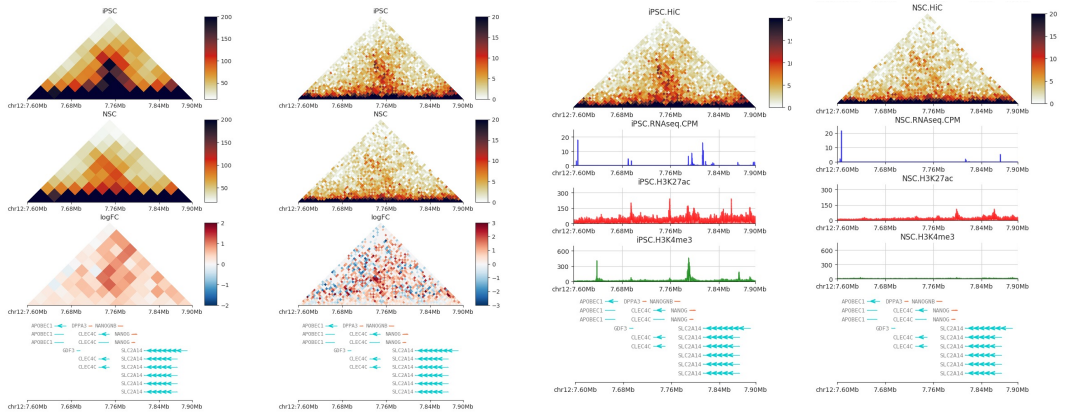
Change in Eigenvector

25kb



Loss of TAD Structure

25kb – 5kb



Clear contact landscape change and loss of E-P interactions linked to loss of NANOG transcription

Some things to keep in mind

Resolution is like looking out a window while you're flying

- Low-res – flying at 30,000 feet - you can see the mountains, but not the houses and roads on the mountains
- High-res – flying close to the ground - you can see the roads and houses, but you won't see all the mountains
- As you call features or look for differences you might to look at different resolutions to find what you're looking for
- Or call features at several resolutions and merge results to get the full picture

Pre-packaged tools and pipelines only get you so far

- Check file types, can you do what the tools are doing without their help?
- When you're using a tool ask what they are doing and how does that relate to the biology you're looking for?

Single-end shotgun, paired-end proximity-ligation

- At its core Hi-C data are just shotgun data, except the distance between pairs isn't the length of the sequenced molecule, but of physical orientation
- Think mate-pair on steroids

Ending with the Preface

Biology biology biology

- What is your biological question?
- Do you have the correct data to answer that question?
- Informatic tools pipelines are a means to an end, make sure the tool your using will help you

There is no "one way"

- There are many tools and many ways to work with data
- Pick a route, give it a go, adjust from there

File structure is paramount

- Learn the format of commonly used data types
- Inspect inputs and outputs
- Get familiar with the data, so when you get asked to do something a tool doesn't do, you know how to get started