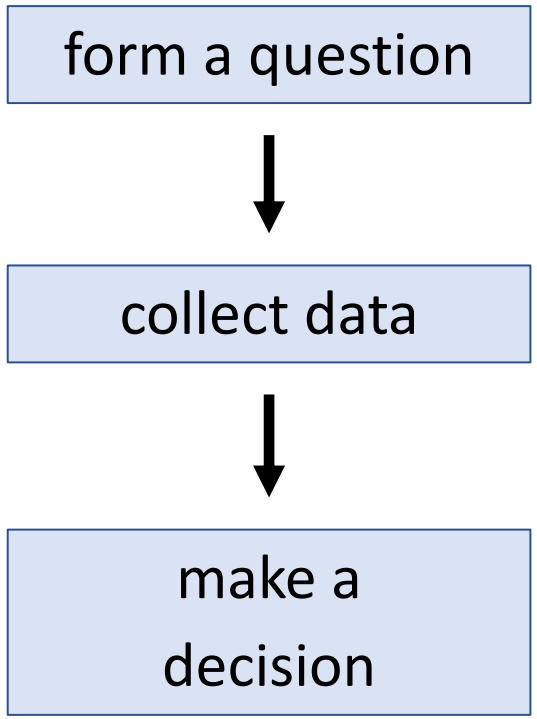
Exploratory Data Analysis

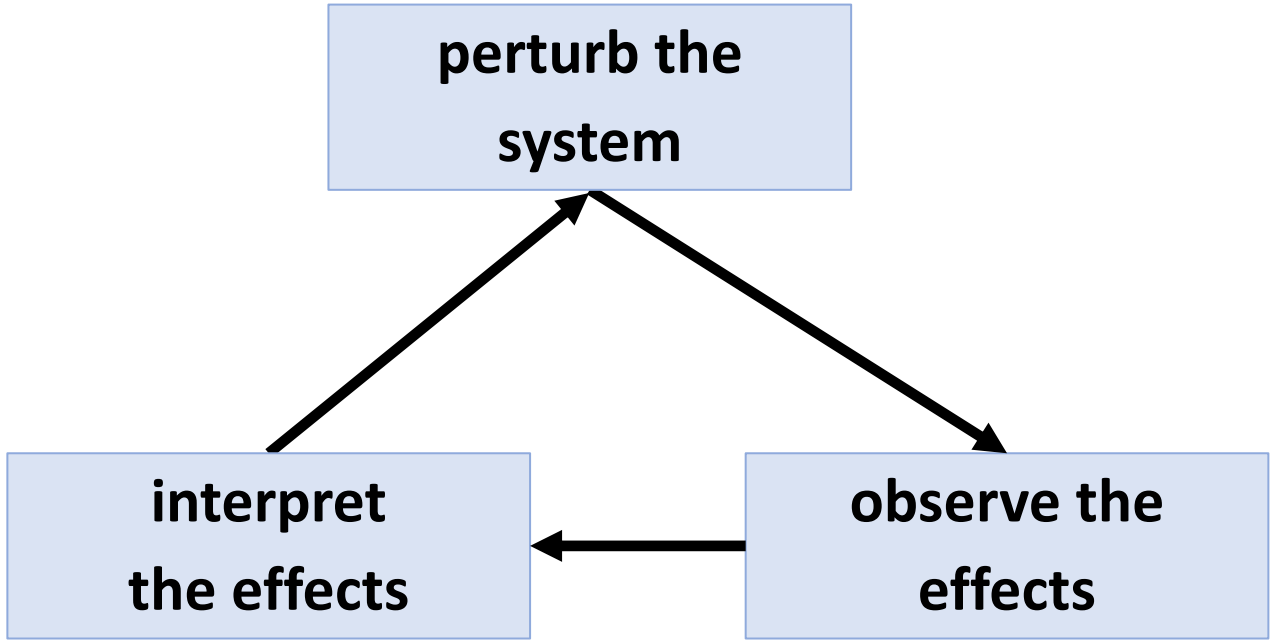
David Quigley, PhD
Assistant Professor
Department of Urology
Department of Epidemiology & Biostatistics
Helen Diller Comprehensive Cancer Center
University of California at San Francisco

Classical statistics: “confirmatory analysis”

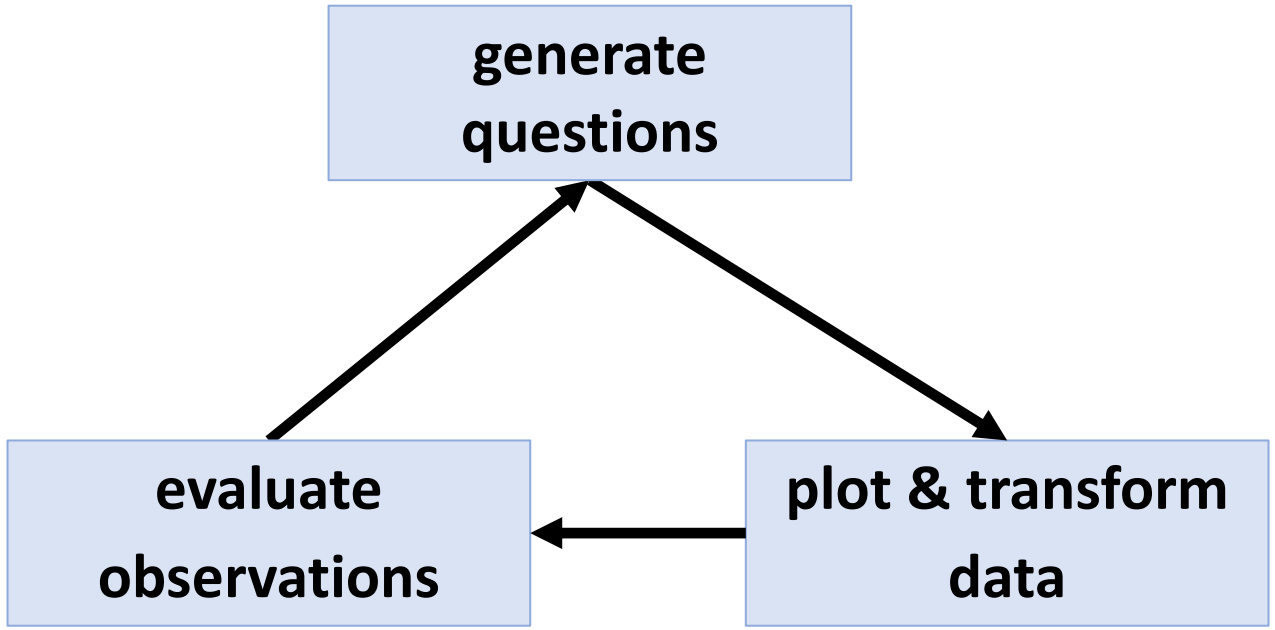


e.g.: randomized clinical trial
drug studies in animal models

Biological research is intrinsically exploratory



Exploratory Data Analysis is iterative



statistical thinking is integral to bioinformatics

hypothesis testing is not the whole process

Zen and the Art of Exploratory Data Analysis

Computational Biologists at work

"EDA is not a formal process with a strict set of rules.

More than anything, EDA is a state of mind."

R for Data Science
Hadley Wickham

Concrete goals of EDA

1) Assess data **quality**

2) Assess **validity** of assumptions

3) Develop hypotheses to **explain** observed data

4) Choose and perform the appropriate tests

Questions during EDA

QUALITY

Are data informative?

QC (peak height, % on-target, # live cells...)
check controls & replicates
batch adjustment
identify uninformative samples

VALIDITY

Is the analysis reasonable?

ranges, summaries of central tendencies, variance
check assumptions

EXPLAIN

What biological model can explain the data?

exploratory analysis (correlation, enrichment, integration)

Quality Control is a pragmatic decision process

QUALITY

trade-off: signal vs. cost

- cells vs. tissue
- assay cost
- prep time
- cheap substrate vs. trial samples
- replicates & study size

What assumptions might we make?

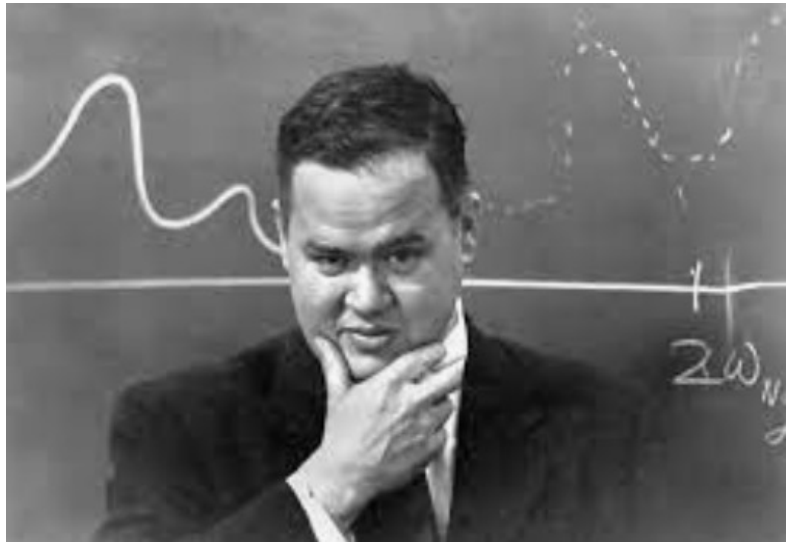
Technical

- expected tissue was sequenced
- library prep succeed
- drug was prepped correctly
- no unexpected batch effects
- CRISPR library contains expected guide distribution

Statistical

- distribution assumptions accurate enough
- no extreme outliers
- missing values distributed at random
- variance and signal are uncorrelated

Statistical wisdom of John Tukey



John Tukey

"It is not enough to look for what we anticipate.
The greatest gains from data come from surprises.
We will usually not be very surprised,
but we should try to be.

The unexpected is best brought to our attention by pictures.
Failing this, as is always to some extent necessary,
more numbers can and do help."

Exploratory Data Analysis as Part of a Larger Whole
John Tukey

Graphical tools for EDA

Summary

box plots & variants
histograms
density plots
clustering
volcano plots

Comparison

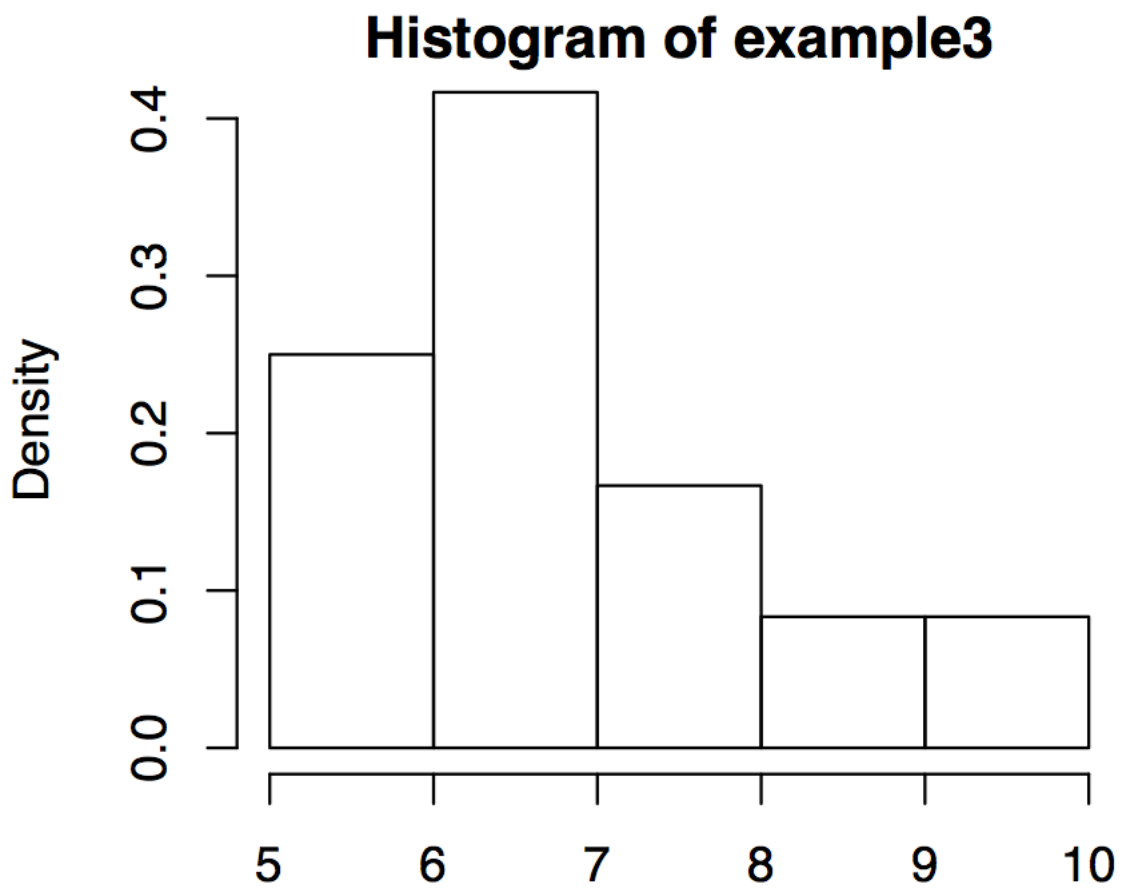
scatter plots
mean - variance

Dimension reduction

Principle Components Analysis
t-SNE
UMAP

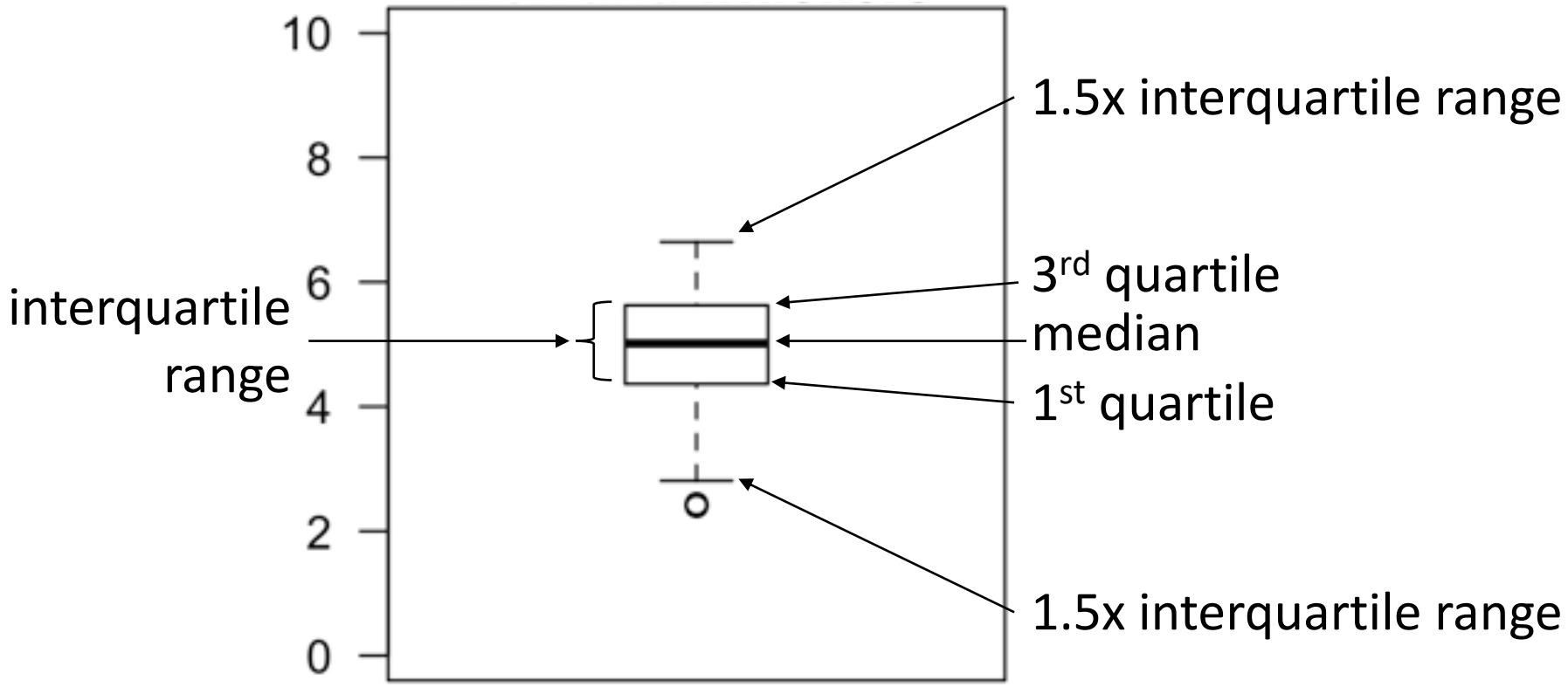
Plotting a histogram

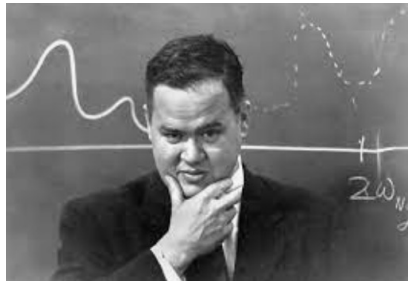
```
example3 = c(5,6,6,7,7,7,7,7,8,8,9,10)
hist(example3, freq = FALSE)
```



Box plots summarize a distribution

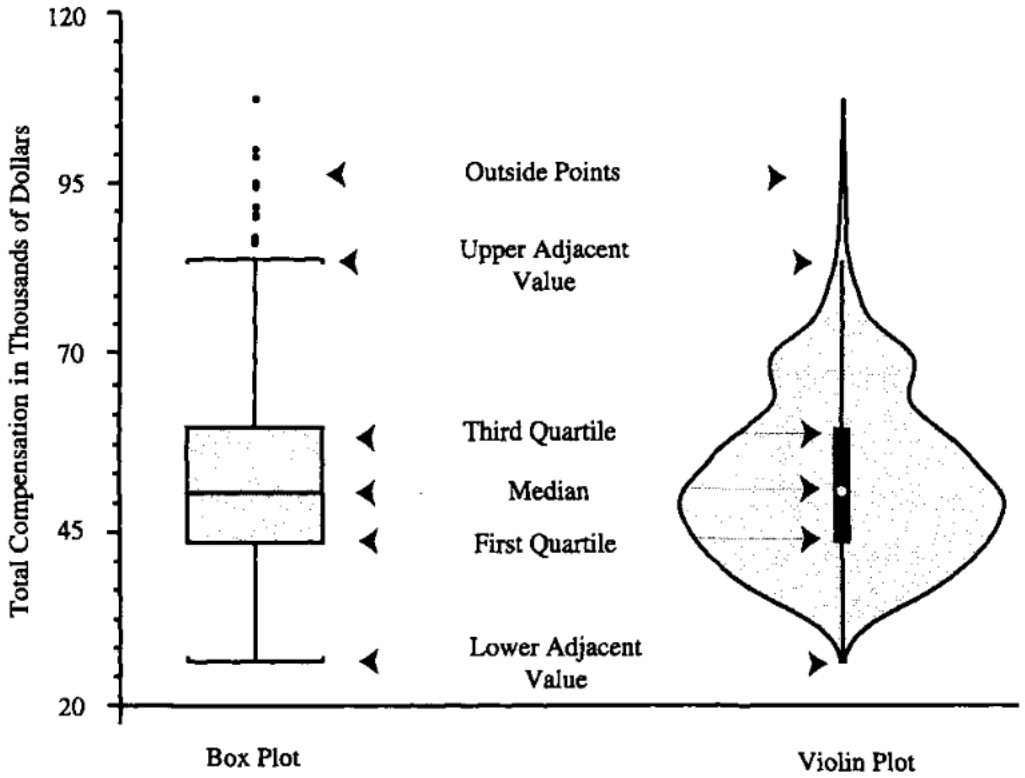
```
boxplot(example3)
```

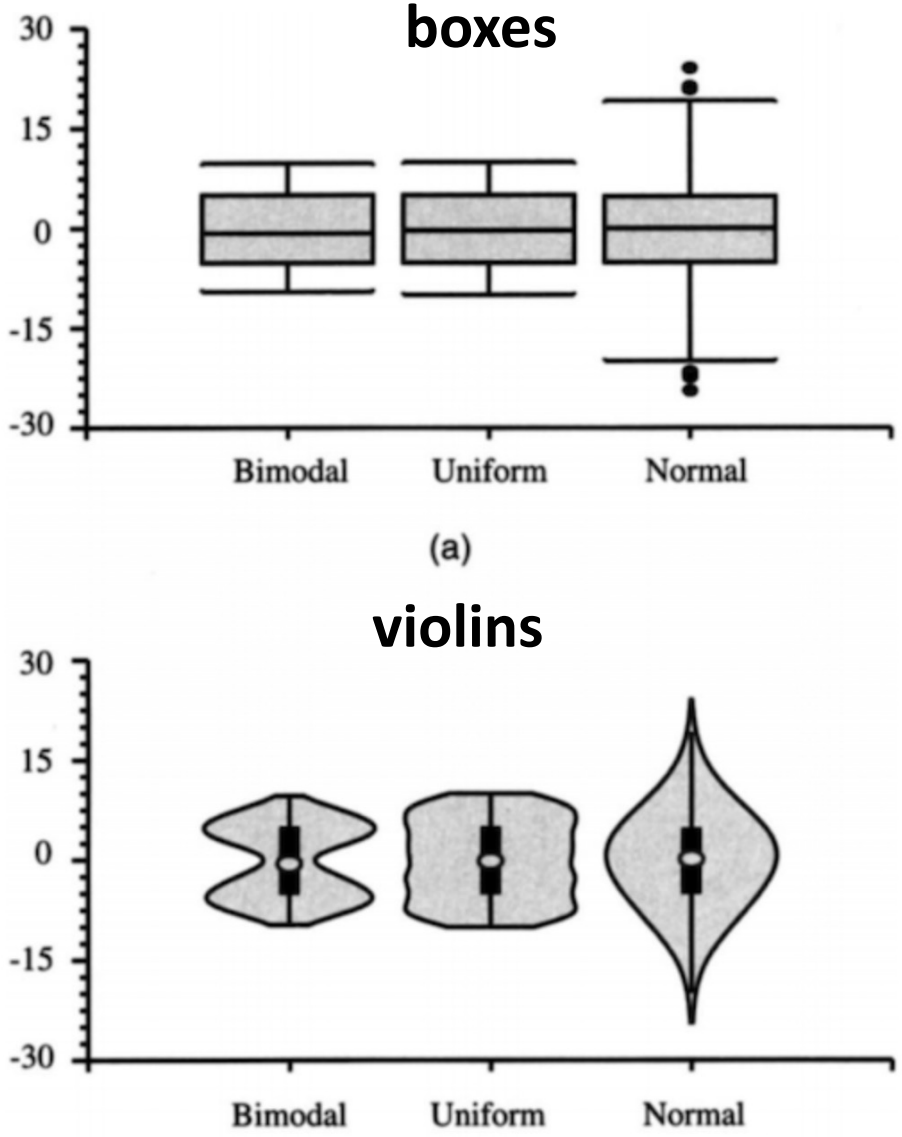




Invented by John Tukey

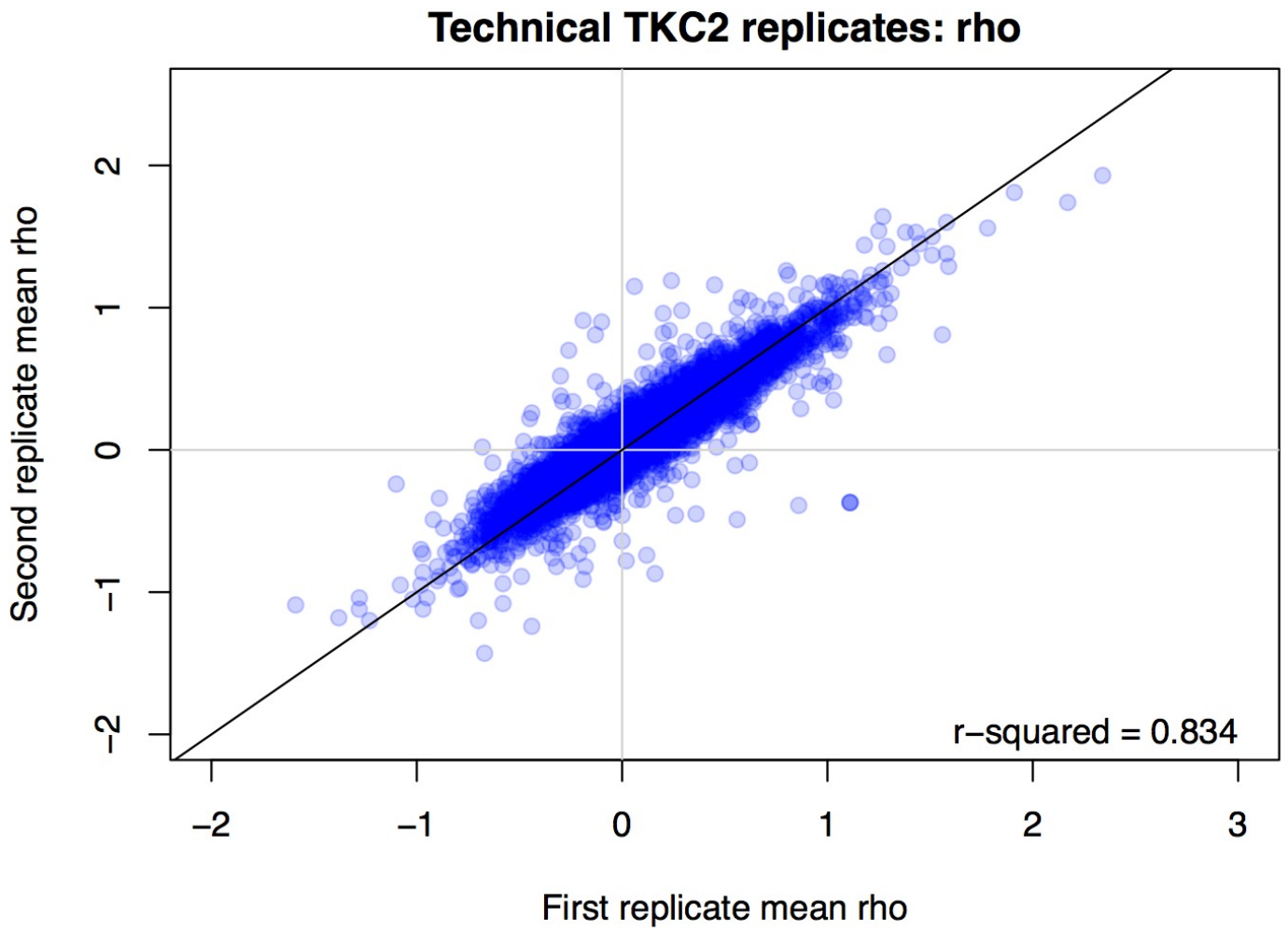
Violin plots may be more informative





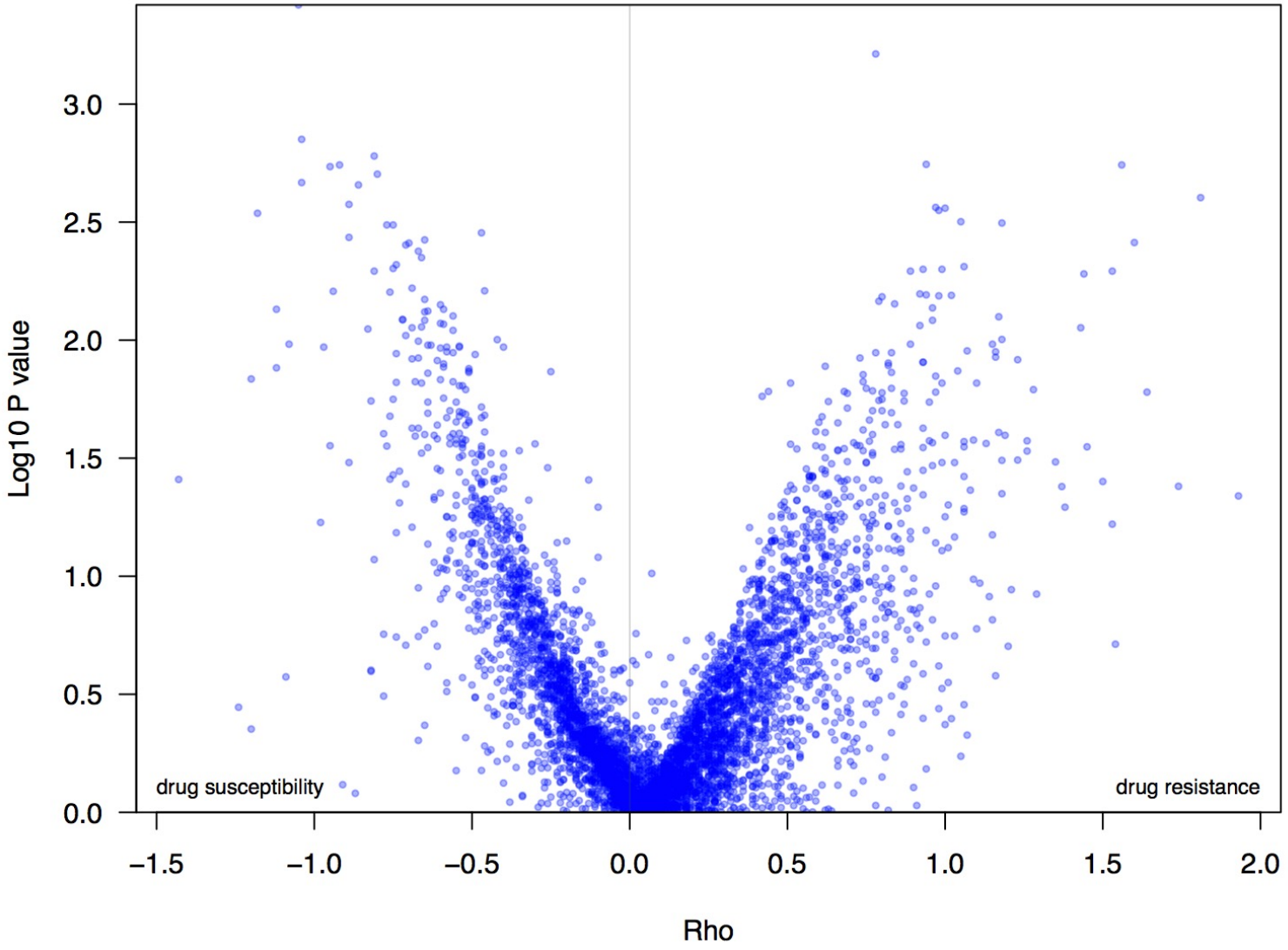
Hintze & Nelson, The American Statistician 1998

Scatter plot: compare technical replicates



Were the replicates consistent?

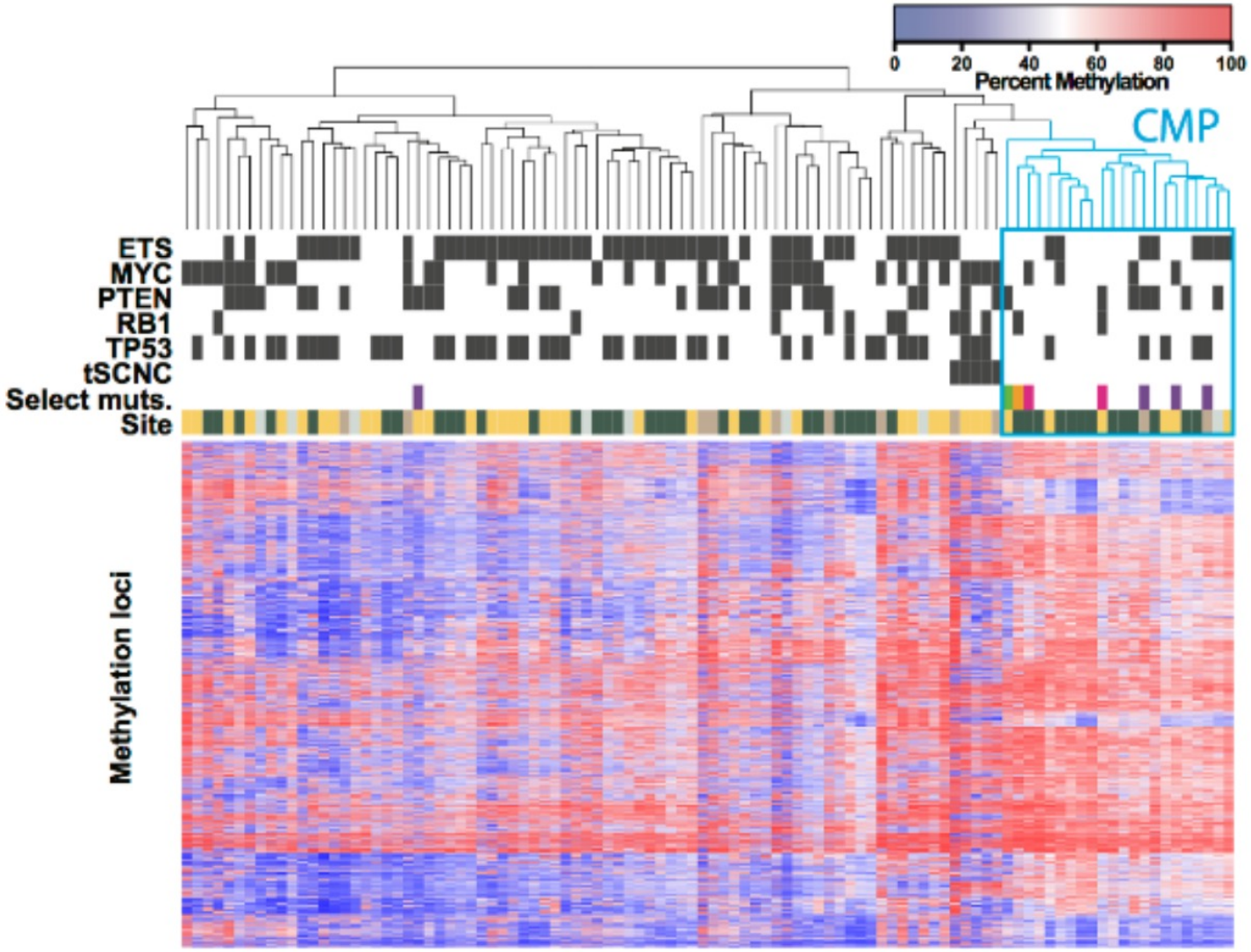
Volcano plot: effect size vs. statistical strength



What's the relationship between effect size and statistical strength?

Where do positive/negative controls lie?

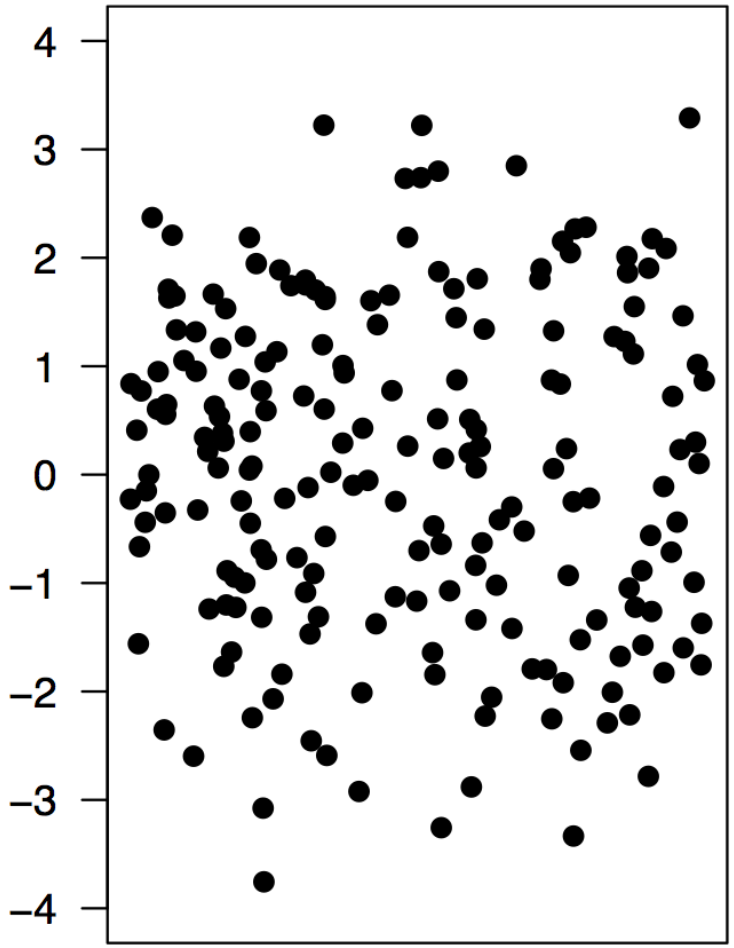
Clustering: which samples are similar to each other?



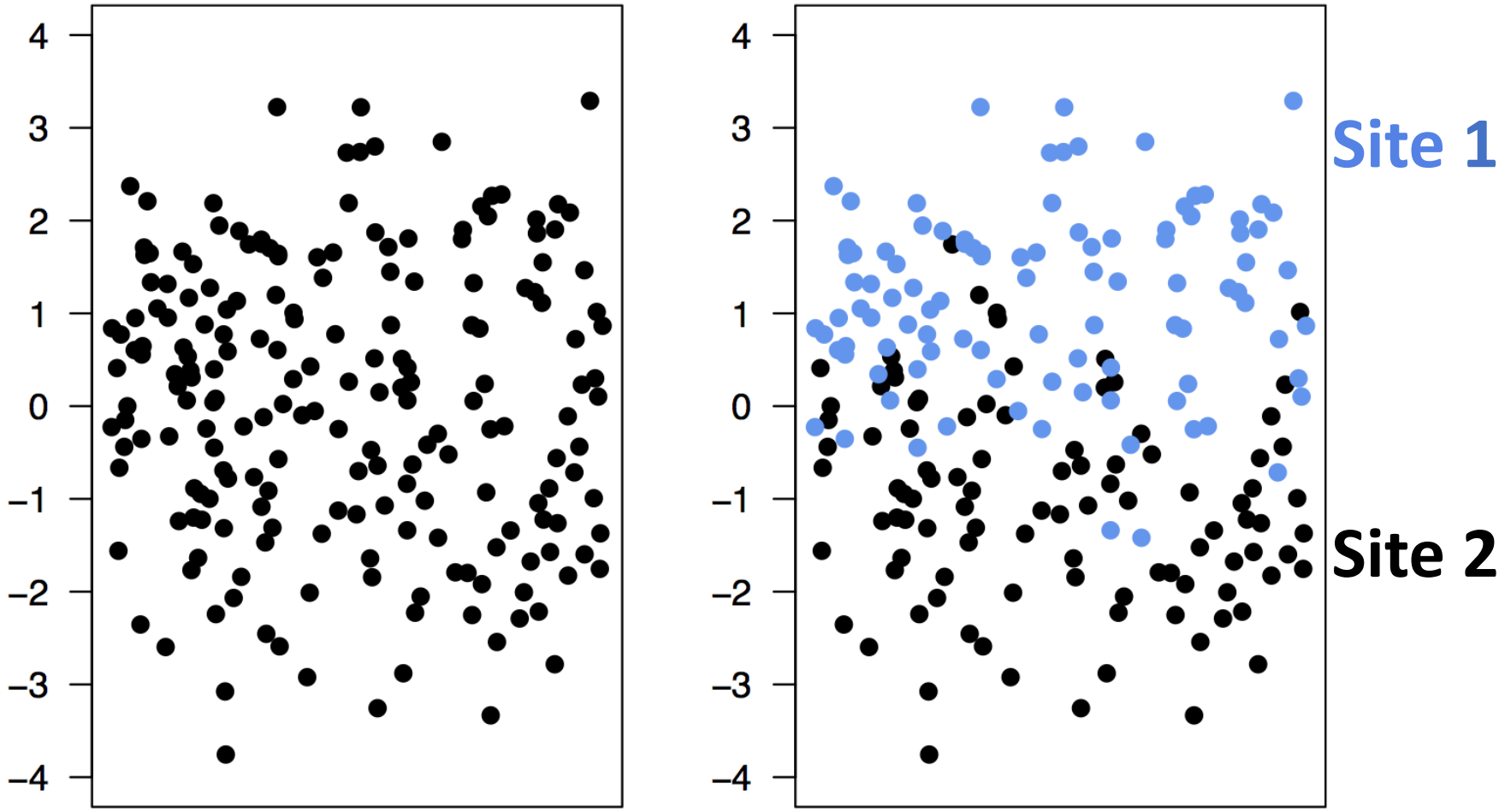
Zhao et al. *Nature Genetics* 2020

Exploratory analysis:
Are there hidden covariates?

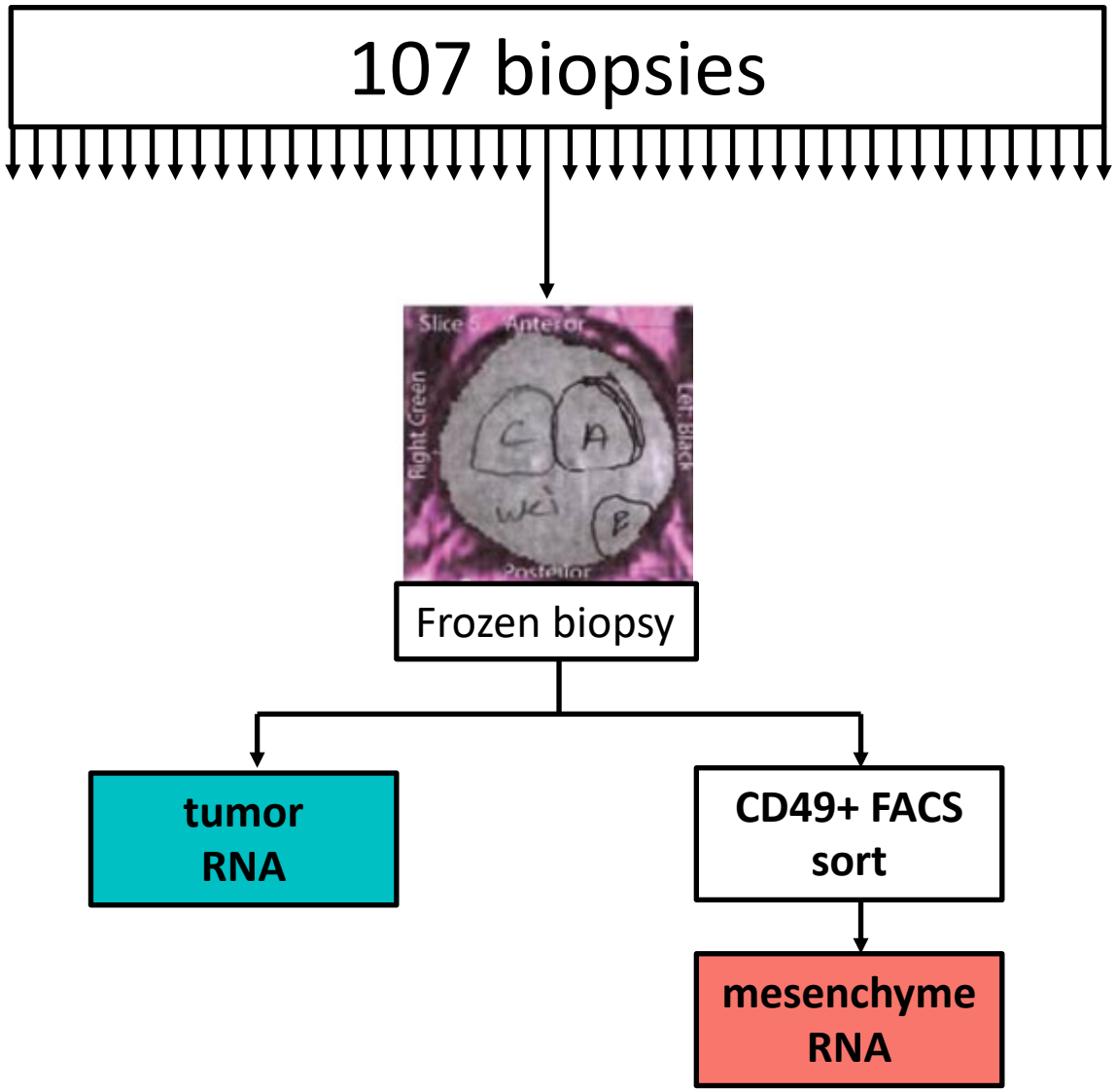
Data collected from two sites



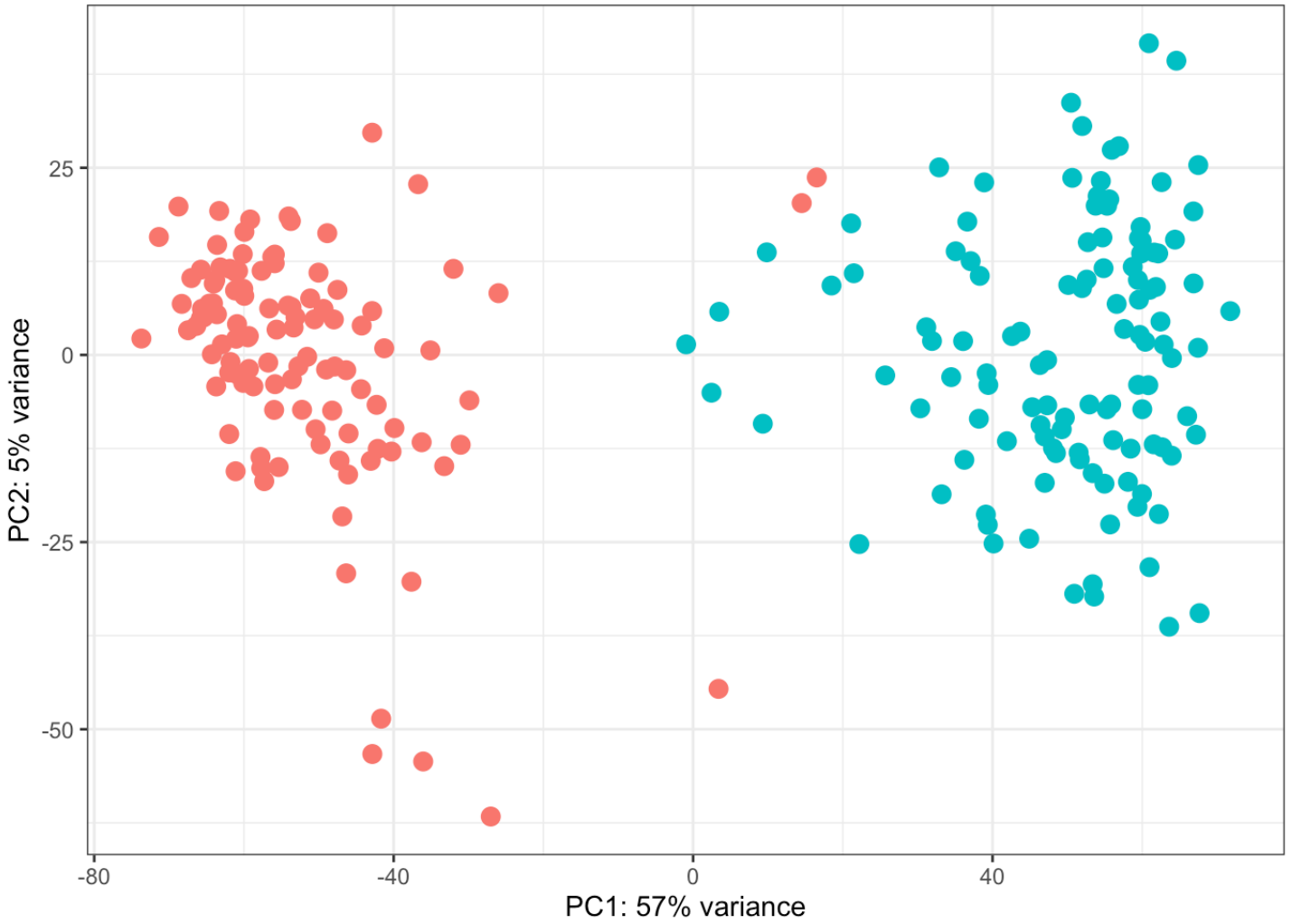
Labeling reveals a batch effect



Real data: total tumor RNA paired with sorted subset



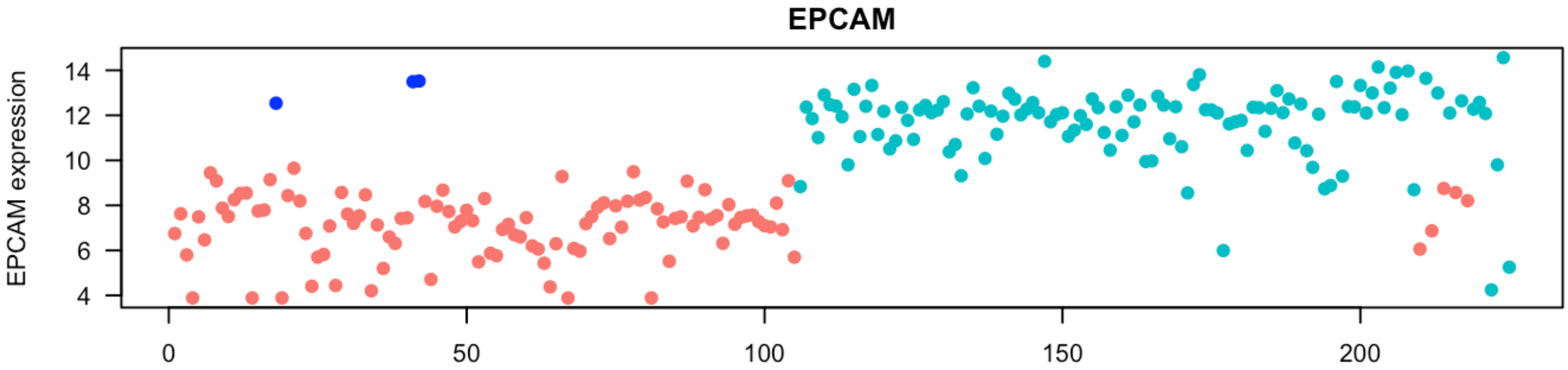
Did the sort work?



snap-frozen tumor sample

CD49+ FACS-sorted cells

The three outliers are EPCAM+, sort likely failed there



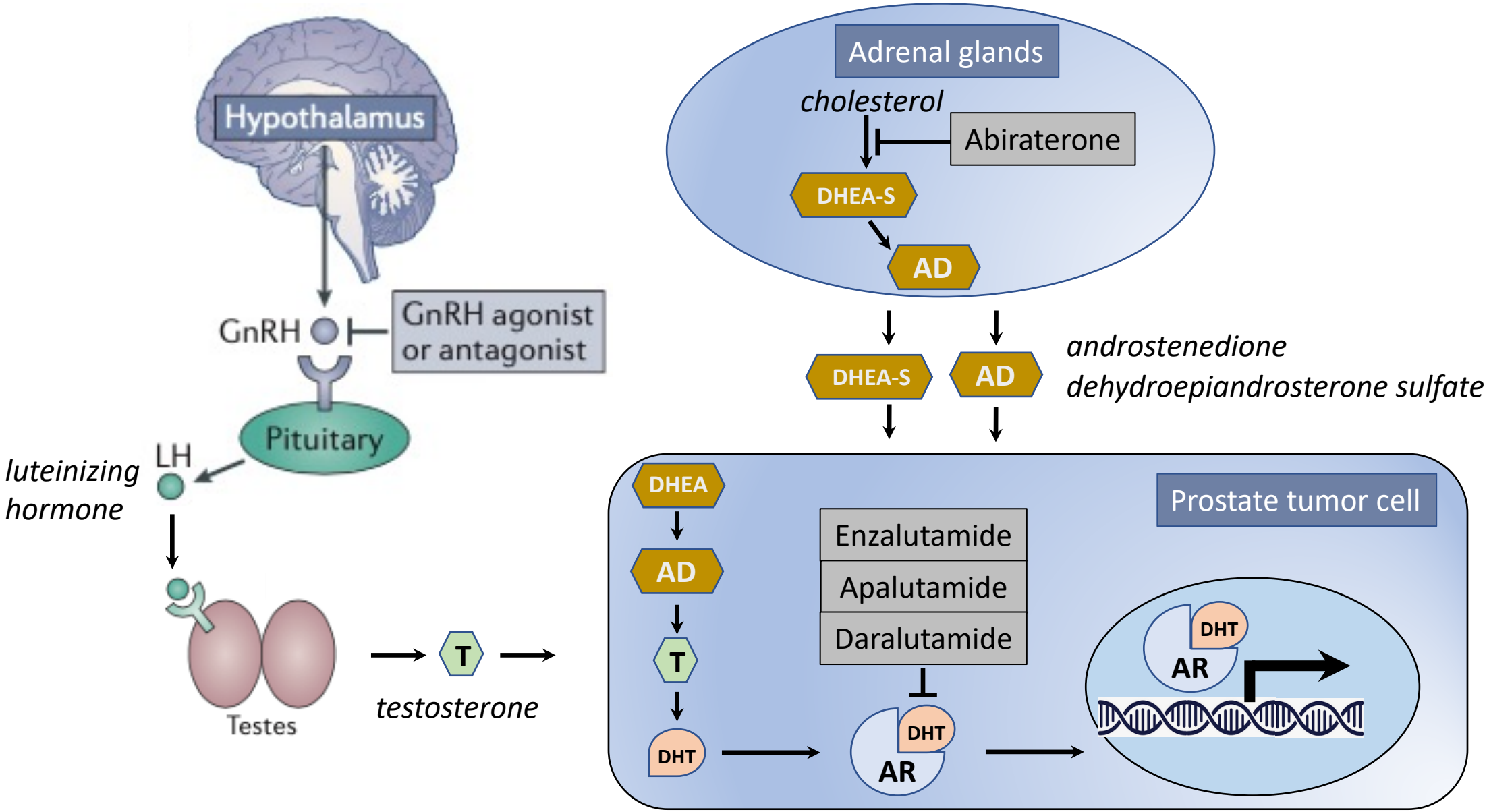
Requires understanding the biology of your question

EPCAM is an epithelial marker

those samples should be EPCAM-negative

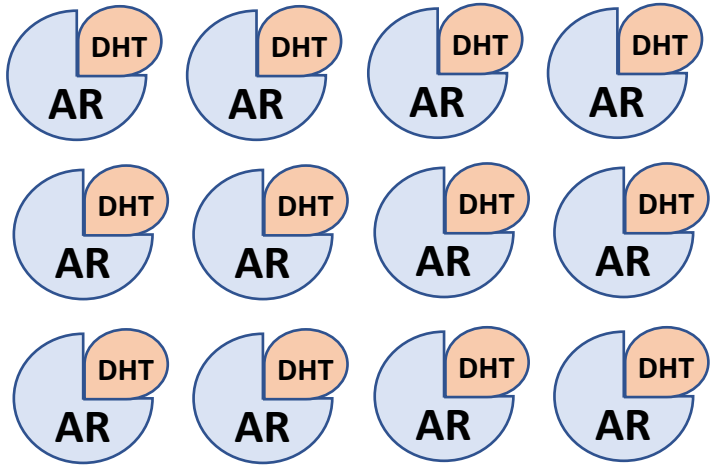
Exploratory data analysis to identify novel drivers of therapy resistance in metastatic castration-resistant prostate cancer

Androgen is the primary drug target for prostate tumors

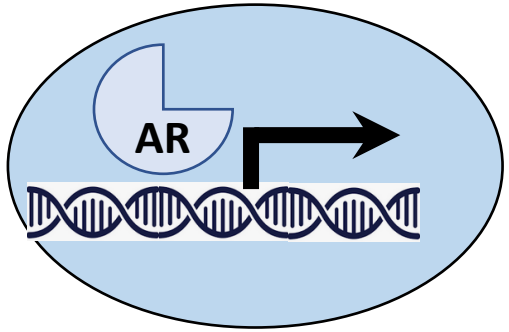


adapted from
Watson et al. Nat. Rev. Cancer 2015

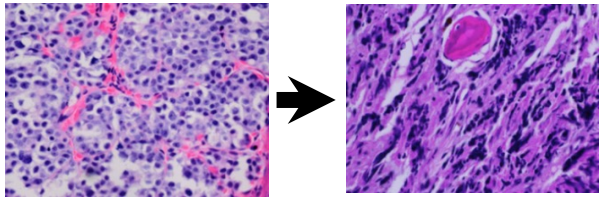
Prostate tumor genome & epigenome response to AR-directed therapy



make more AR

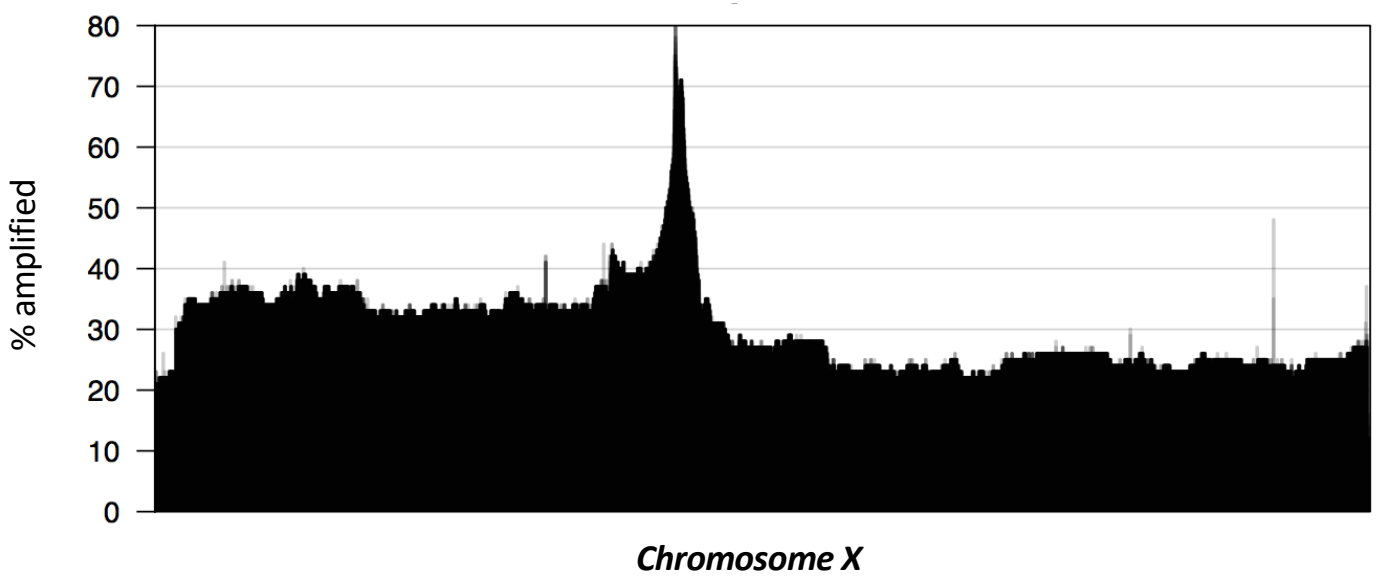


mutate *AR* for ligand independence



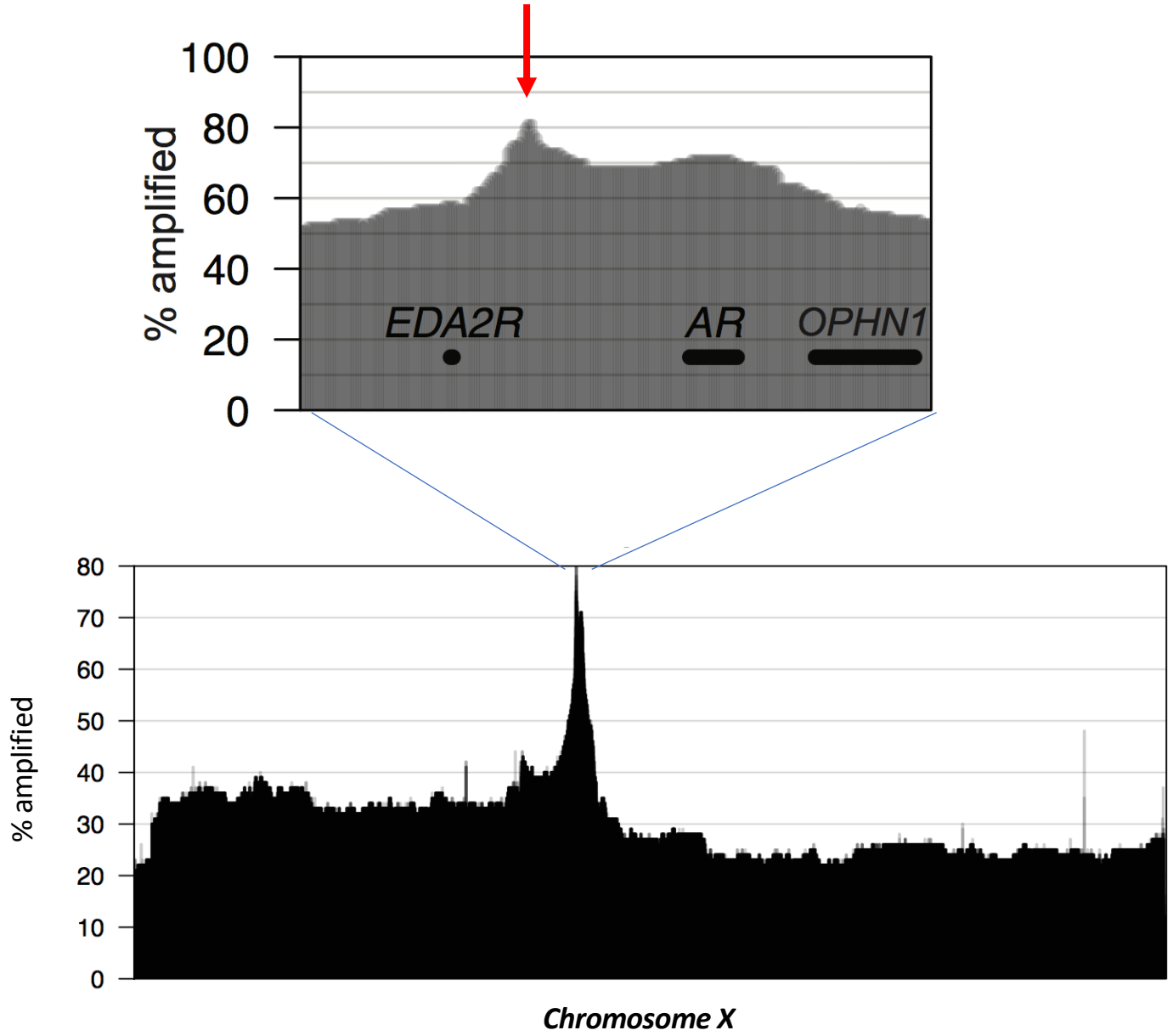
AR-independent phenotype

We generated the first whole genomes for mCRPC and saw amplified DNA



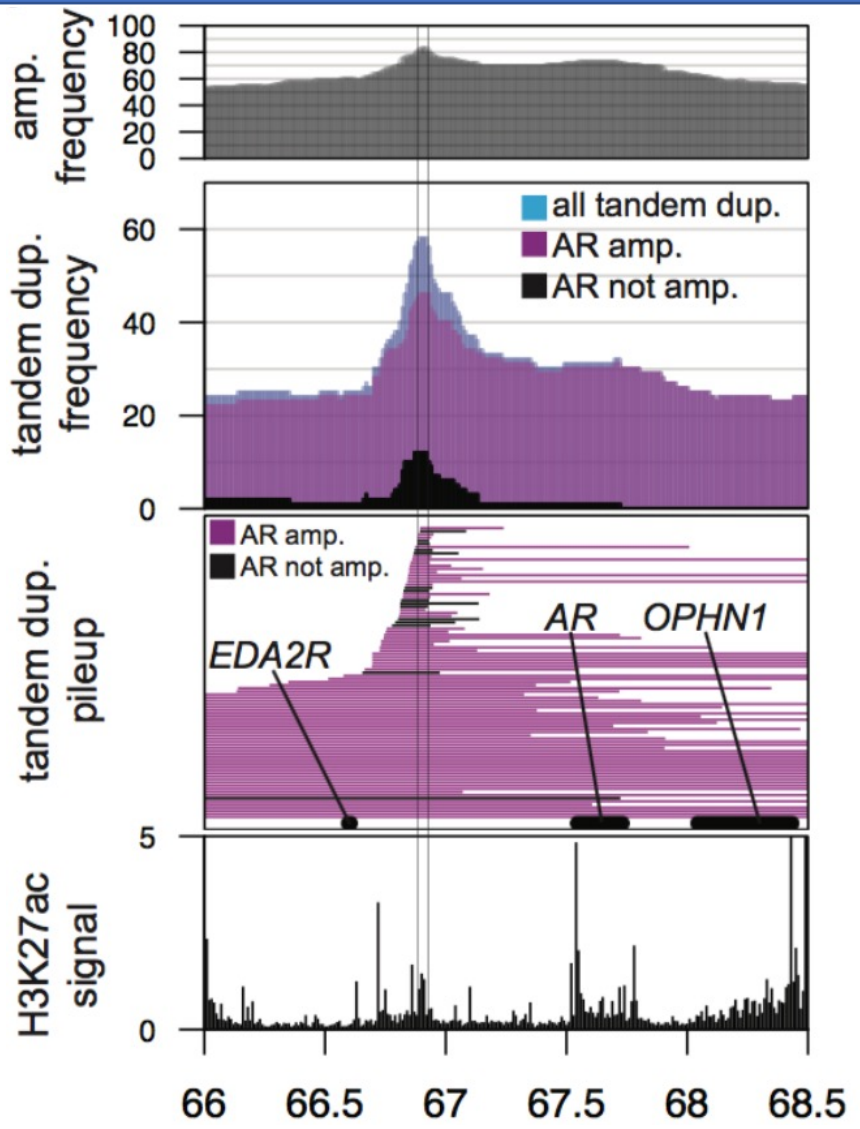
Quigley, et al Cell 2018

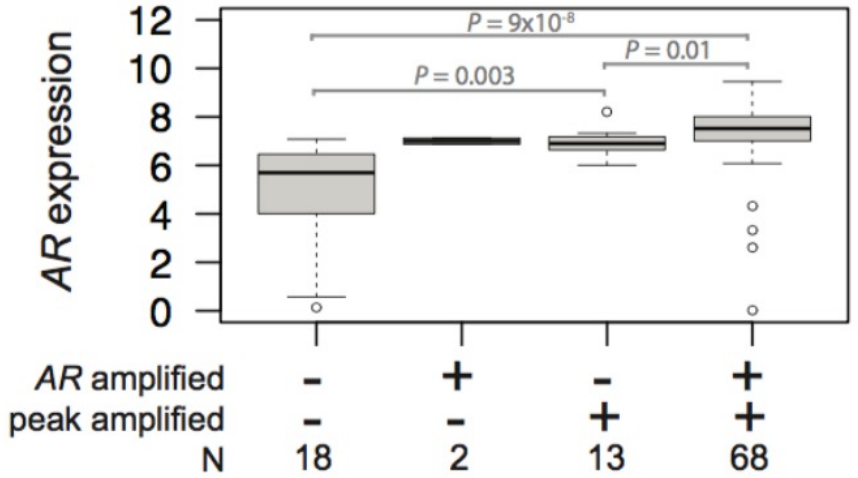
But surprisingly, the peak did not coincide with the *AR* gene itself



Quigley, et al Cell 2018

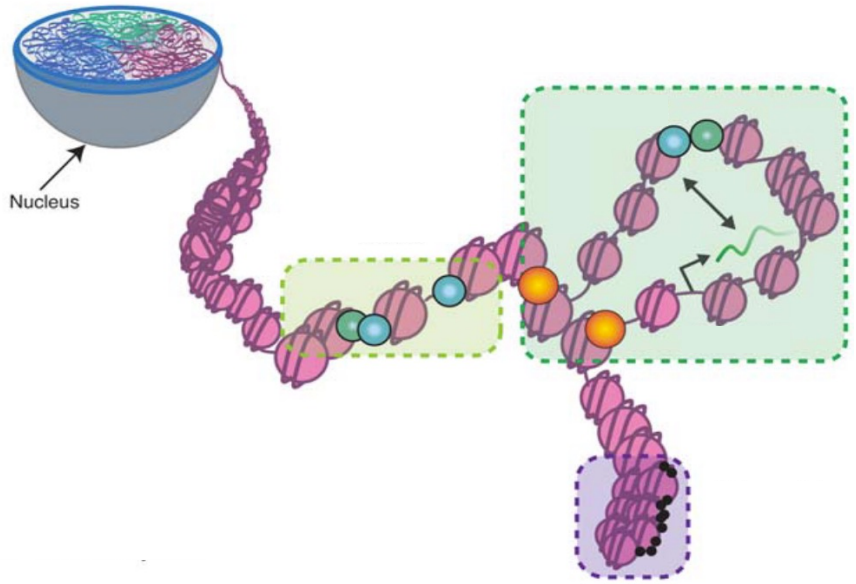
We had discovered a novel enhancer that drives treatment resistance





Quigley, et al Cell 2018

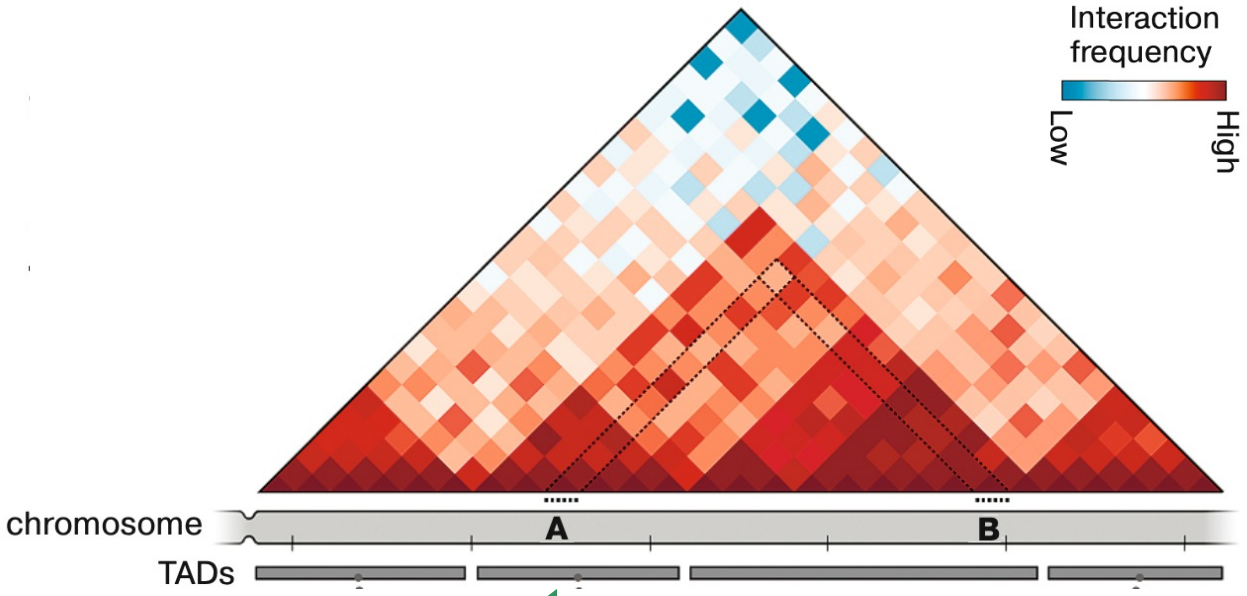
Local DNA regions (TADs) are physically close to each other

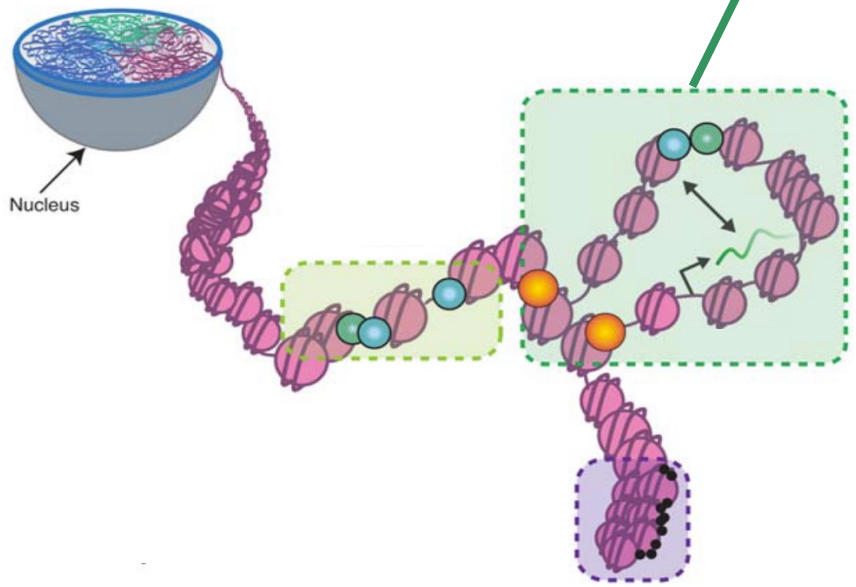


Topologically Associating Domain
(TAD)

Figure adapted from doi: 10.1101/sqb.2016.81.031013
& Razan Biochemistry (Moscow) 2018

Hi-C can identify TADs and chromosomal open vs. closed compartments





Topologically Associating Domain (TAD)

Figure adapted from doi: 10.1101/sqb.2016.81.031013
& Razan Biochemistry (Moscow) 2018

3D structure is linked to transcription and mutagenesis

open chromatin
(alpha compartments)

- location for most genes
- higher transcription levels
- more structural variants

closed chromatin
(beta compartments)

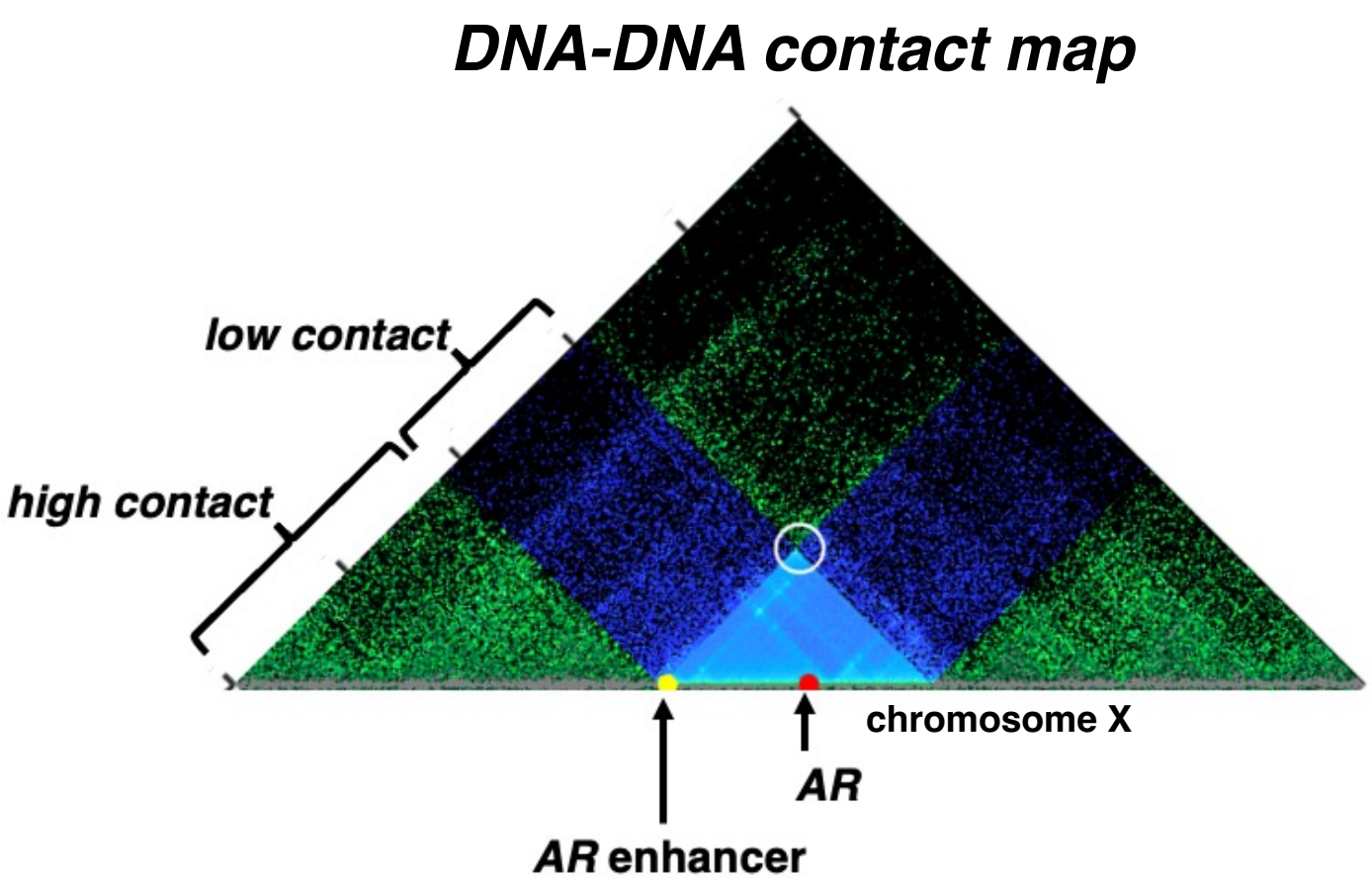
- more mutations
- more partially methylated domains

Zhao et al., in preparation

We performed Hi-C on 80 metastatic prostate tumors

- Does 3D structure help explain structural variants like AR amplification?
- Are there 3D genome subtypes?
- Is 3D structure linked to patient outcome?

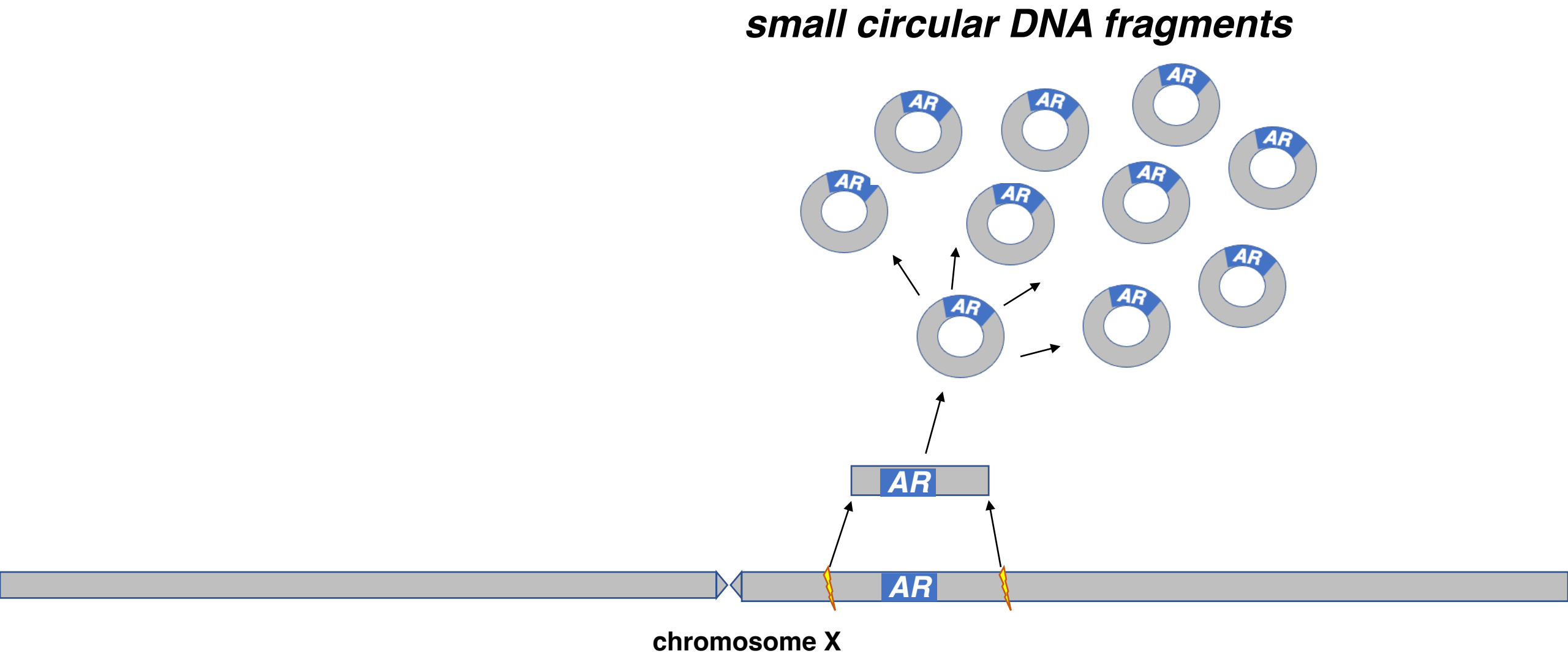
alpha-beta assignment failed around the *AR* in some cases



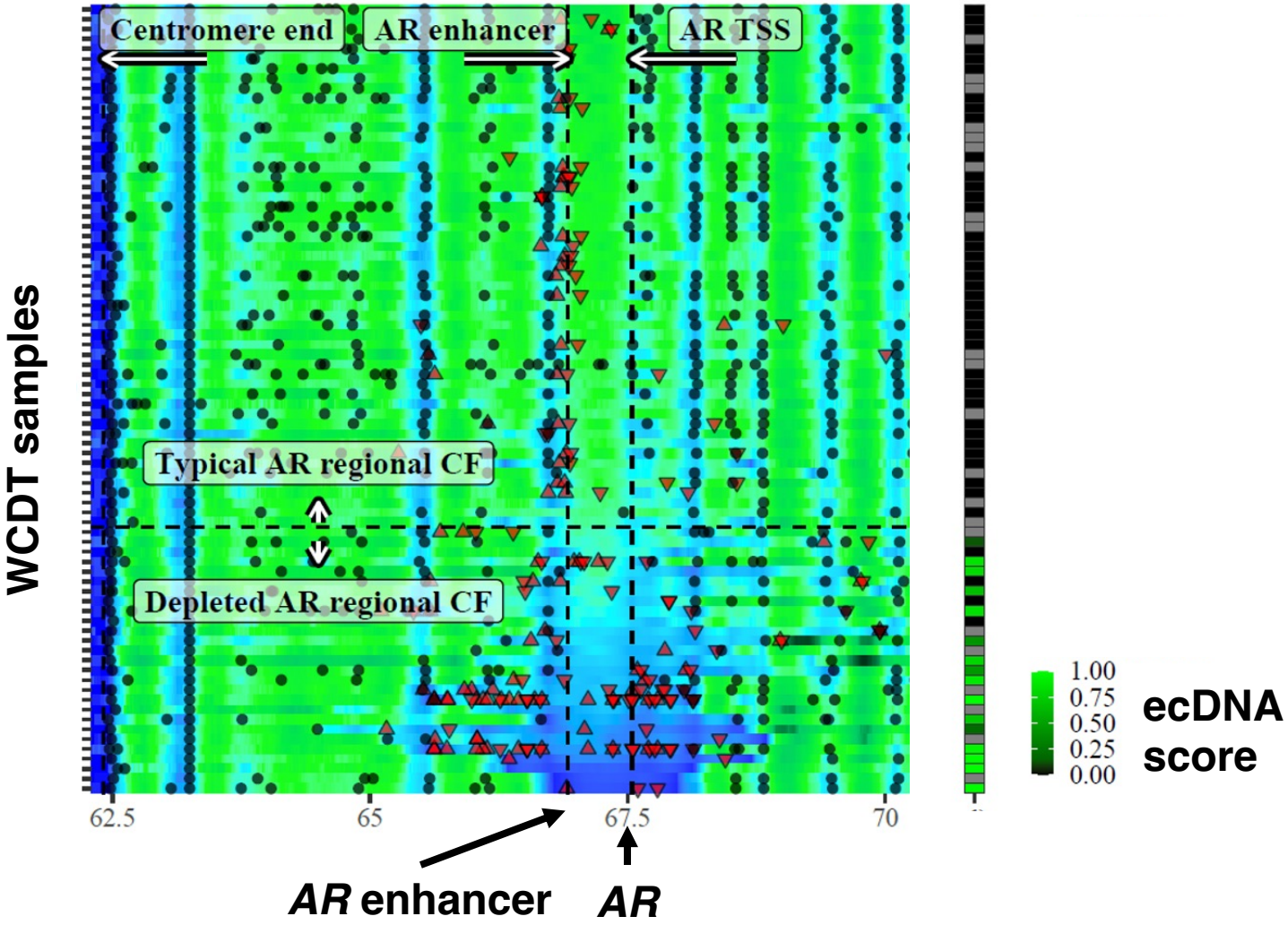
low contact region, alpha/beta failed here.
could this be part of a separate chromosome?

Zhao et al., in preparation

Driver genes can be amplified by extrachromosomal DNA



AR is frequently amplified by ecDNA in mcRPC



Hi-C identifies the presence of an interesting phenotype but doesn't explain it. Structural variant analysis identifies circular DNA structures at these loci

Elements of successful exploratory data analysis

- Know the limitations and of the assays and the analysis. If you're a dry lab researcher, collaborate closely with people who know the assay.
- Start from an unbiased view. When generating or analyzing new data, examine the big picture and broad trends first and then focus on specific questions .
- Use a combination of statistical and graphical tools to ask questions
- Assume there are technical failures or biases; your first job is to identify these, before any biological analysis.
- Study the biology of your system. The better you understand this, the more likely you'll recognize something that is unexpected and potentially interesting.