Statistical Genomics
Total Request Live

Kasper Daniel Hansen
< khansen@jhsph.edu | www.hansenlab.org >
Department of Biostatistics
Johns Hopkins University

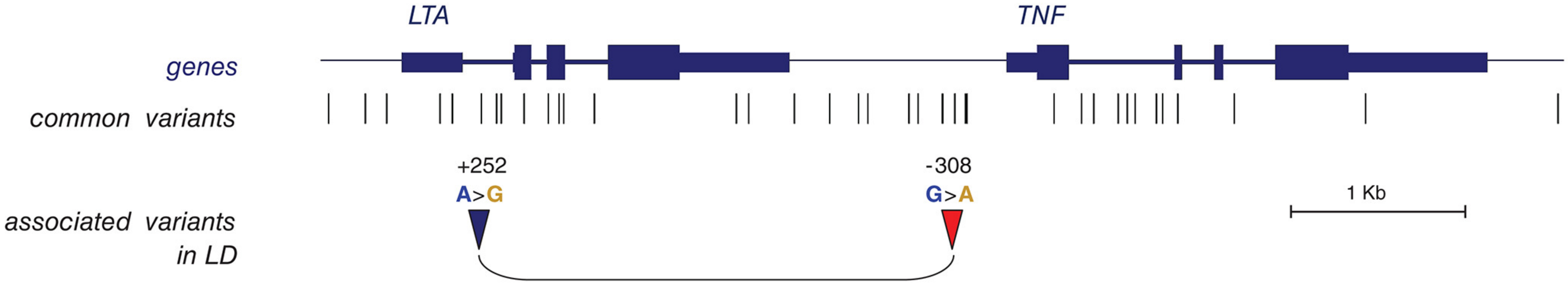
eQTL

Expression quantitative trait loci

In genetics, associating genetic variants with a continuous phenotype (like height) is call a quantitative trait loci (QTL) analysis.

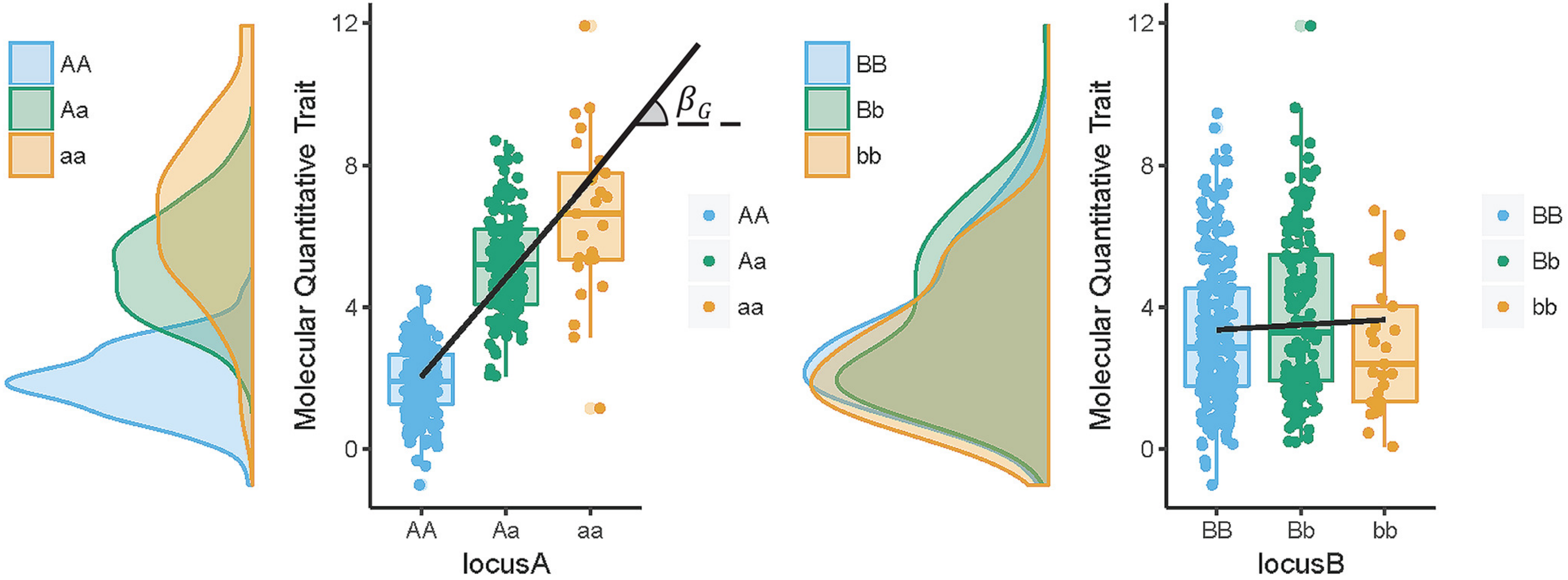
We can do this, where our “phenotype” is the expression of all genes. That is called an expression-QTL analysis.

Some numbers: say we have 1M genotypes and 10,000 SNPs.
That yields 10e11 comparisons. So we only do “local” comparisons.



Single SNP and single gene: relatively straightforward

$$Y = \beta_0 + \beta_G X + \varepsilon \quad \begin{cases} H_0: \beta_G \simeq 0 \\ H_1: \beta_G \neq 0 \end{cases}$$



$\beta_G \neq 0$
locus A is a molQTL

$\beta_G \simeq 0$
locus B is not a molQTL

eQTL analysis

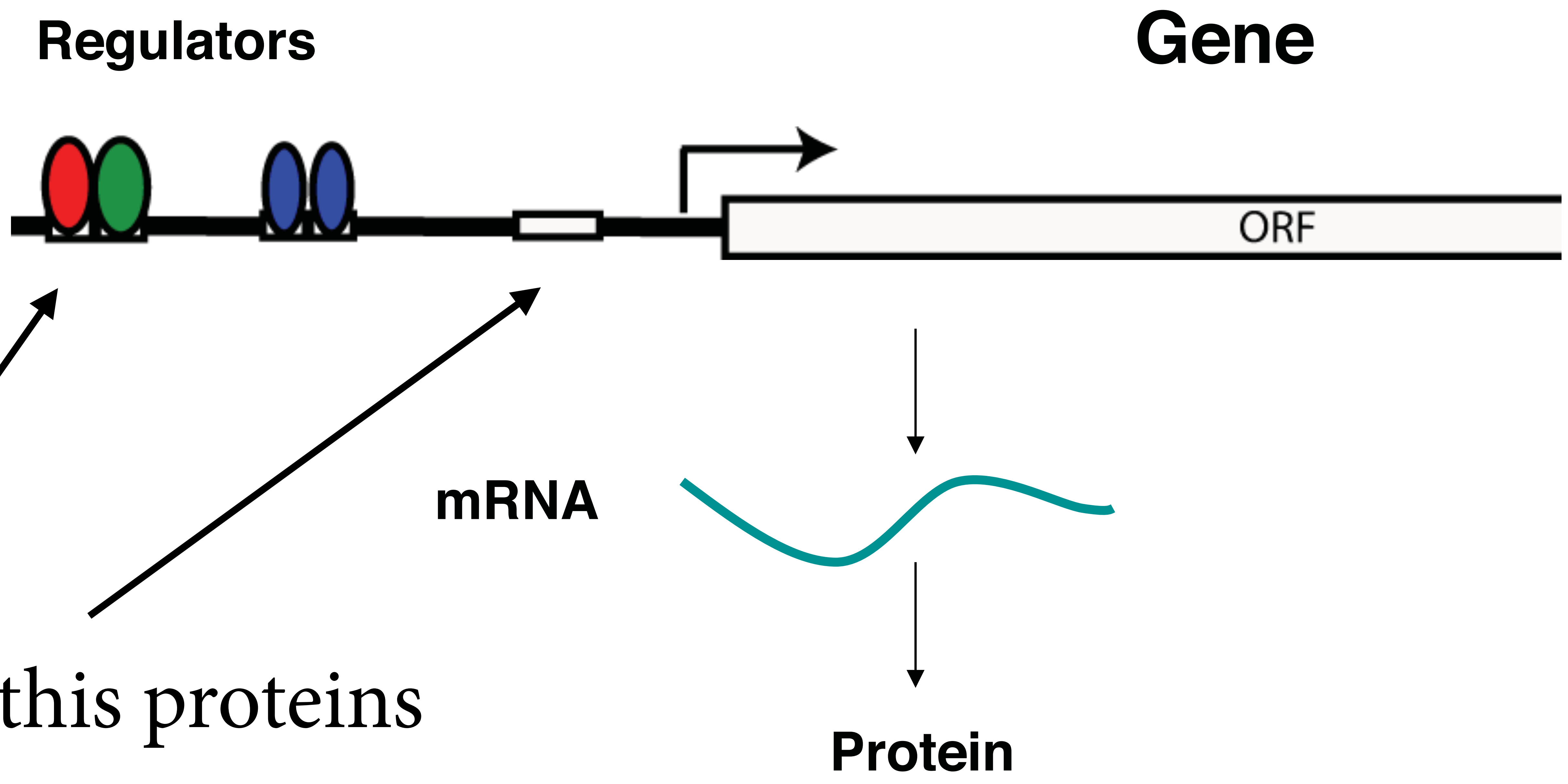
Lots of book keeping. Lots of tests were both the outcome (expression) AND the covariates (the SNPs) vary.

Standard approach: control for “everything” and see if anything “survives”.

Otherwise, there is nothing special.

ATAC (and ChIP) sequencing

Gene regulation



Where / Which / Why do this proteins control gene expression

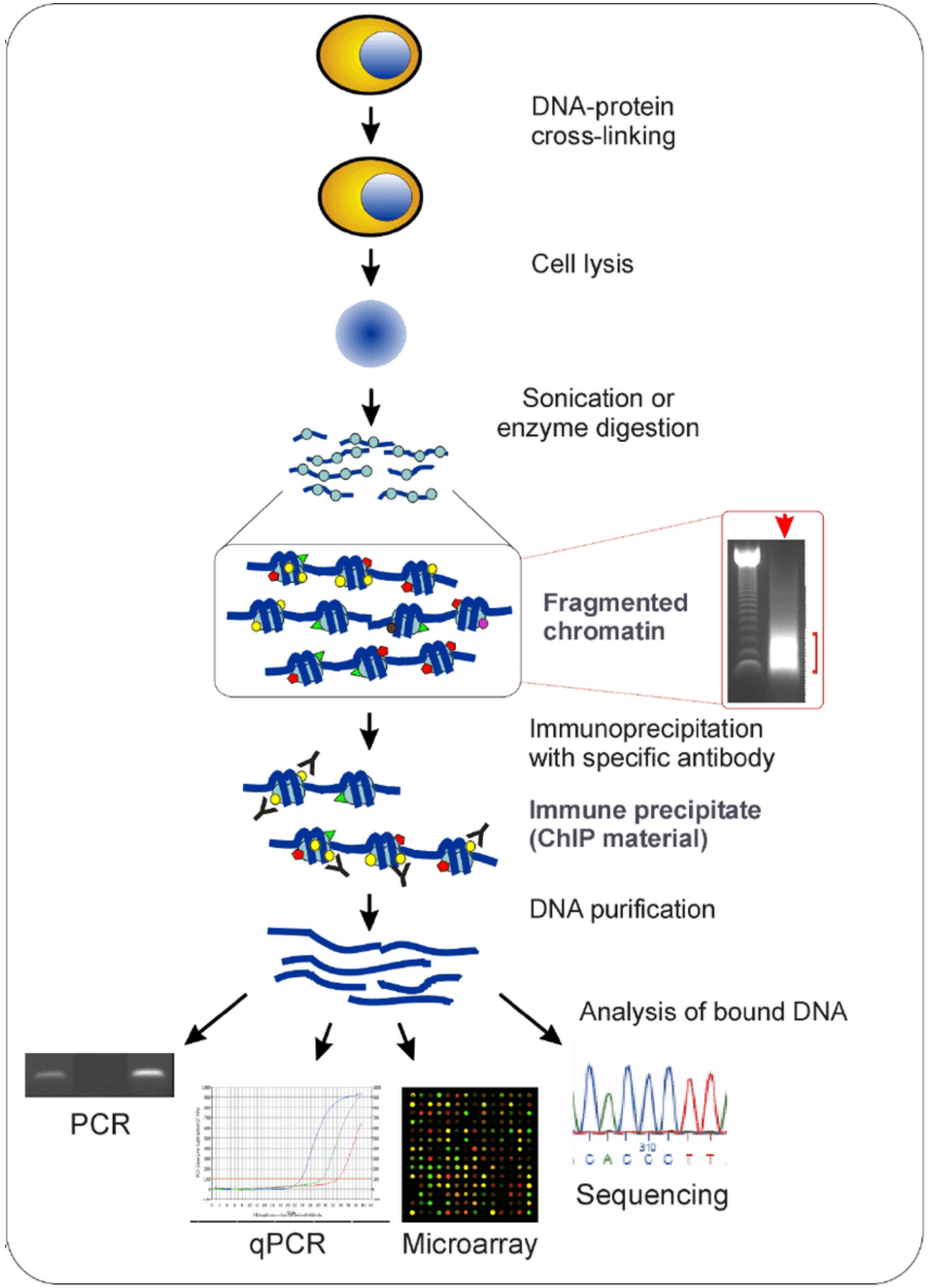
Assays

ChIP: measures binding of a protein to DNA. Used to study

- Transcription factors
- Histone modifications (narrow and broad)

DNAse and ATAC: measures open chromatin
(where can proteins potentially bind)

ChIP (chromatin immunoprecipitation)


DNA-protein cross-linking
Cell lysis
Sonication or enzyme digestion
Fragmented chromatin
Immunoprecipitation with specific antibody
Immune precipitate (ChIP material)
DNA purification
Analysis of bound DNA
PCR
qPCR
Microarray
Sequencing

Cross-linking

Fragmentation

IP enrichment (requires antibody)

Sequencing

Controls

Typical controls

Input (don’t do IP)

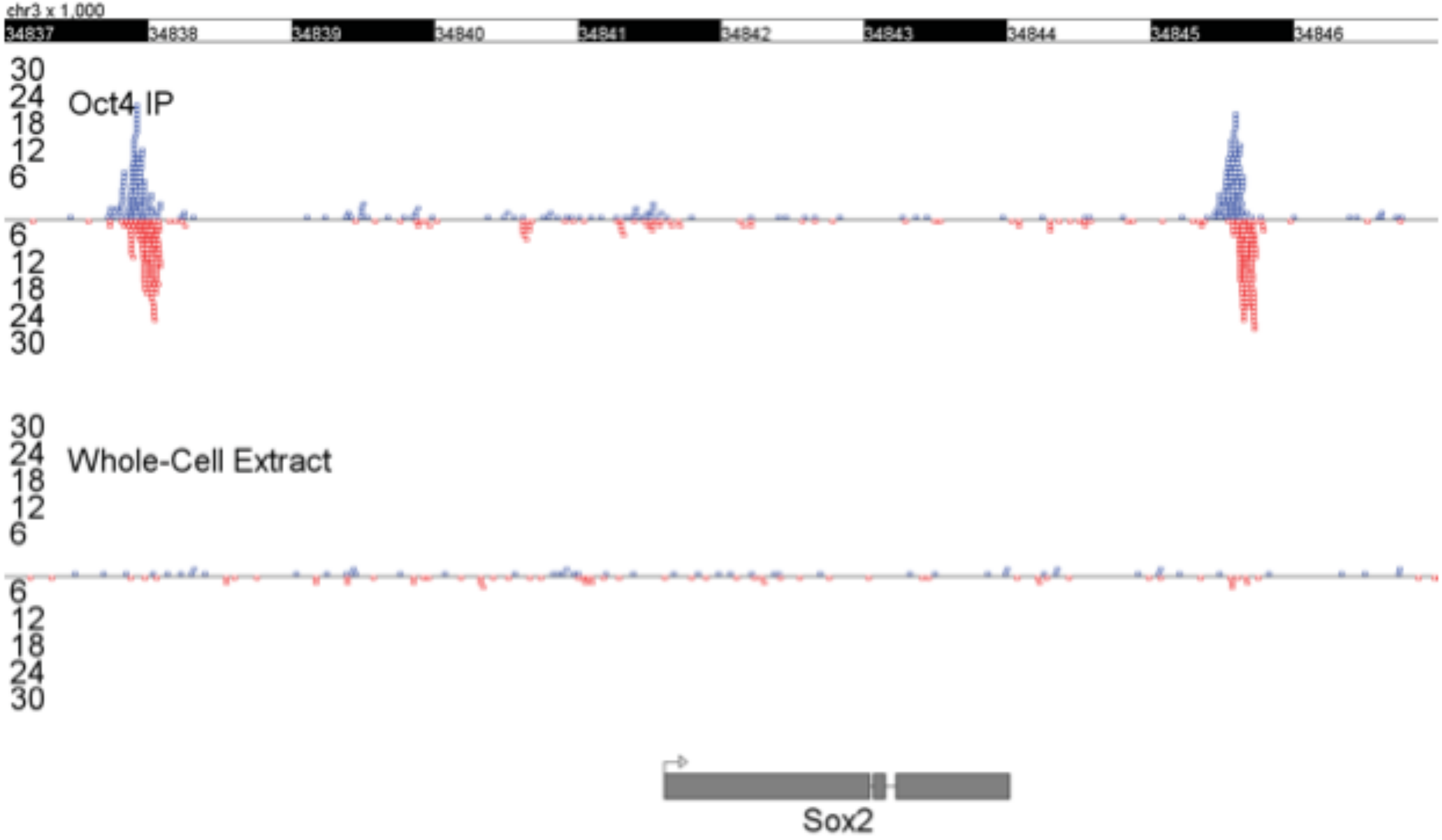
Igf (non-specific antibody)

We see peaks in control samples

Usually we think of the control sample as measuring background peaks, and most peak callers try to separate signal peaks from background peaks.

Some regions of the genome are highly susceptible to background peaks. ENCODE maintains a “black list regions” which is usually excluded.

Data


chr3 x 1,000
34837
34838
34839
34840
34841
34842
34843
34844
34845
34846
30
24
18
12
6
Oct4 IP
6
12
18
24
30
30
24
18
12
6
Whole-Cell Extract
6
12
18
24
30
Sox2

ChIP, some history and drawbacks

ChIP requires a good, specific, antibody. That can be hard to obtain

ChIP requires large amount of material. For this reason, we tend to see ChIP in model organisms or cell lines.

Because of this, ChIP analysis is usually done as “single-condition” (where is the protein bound in this sample) and with few (usually 2) technical replicates.

The first step here is “peak calling”.

On technical replicates: these are usually “growth” replicates of cell lines. In the ChIP literature, this is called “biological replicates”.

Peak calling

Tons of software.

Despite all the work, I don’t consider this a “approximately solved” problem.

But it seems low reward to work on…

IDR

Irreducible discovery rate: Qunhua Li et al (2011) Annals of applied statistics

Question: how do we identify peaks which are “reproducible” between 2 replicates?

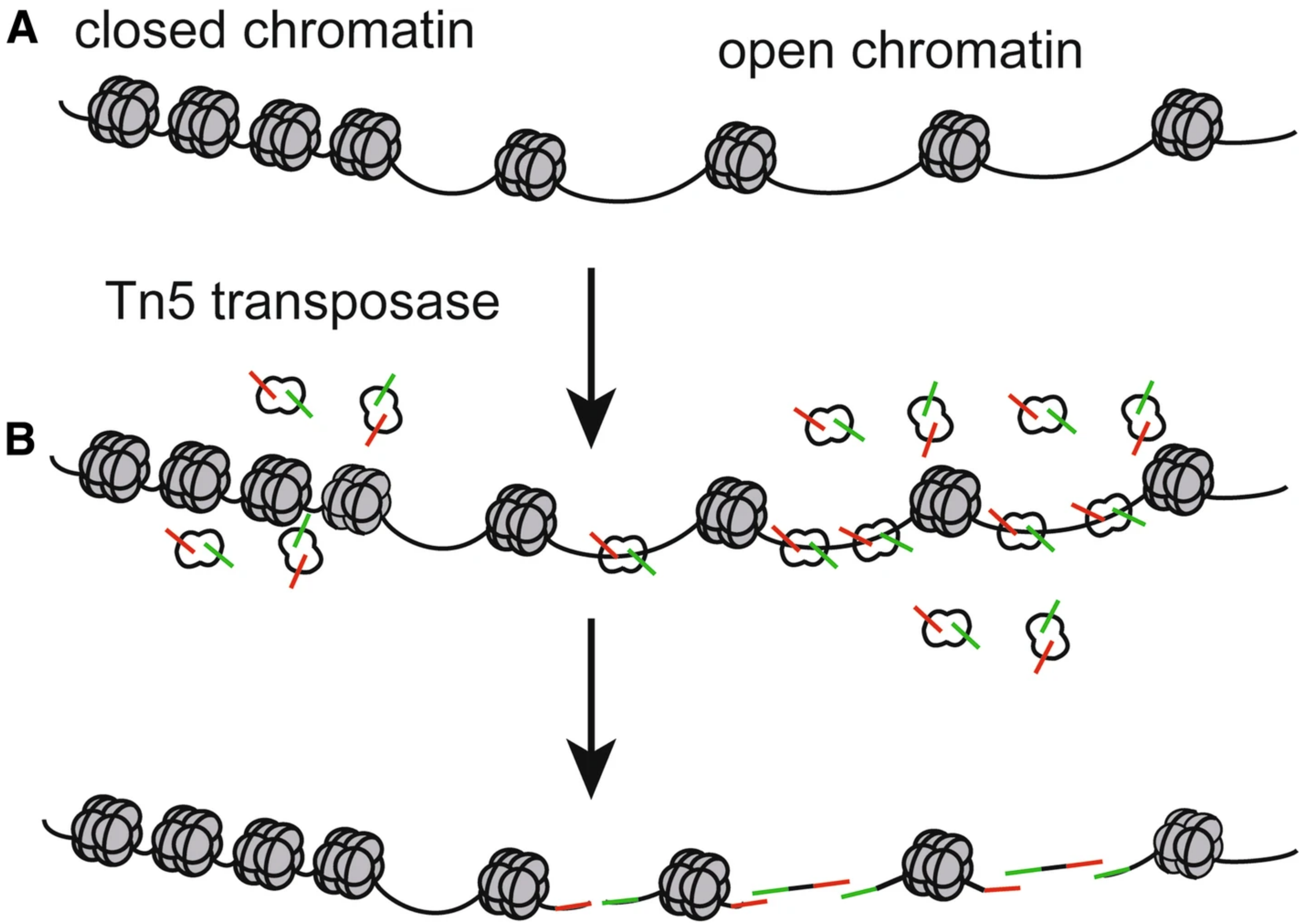
ATAC and DNAse hypersensitivity

Both assays measure open chromatin.

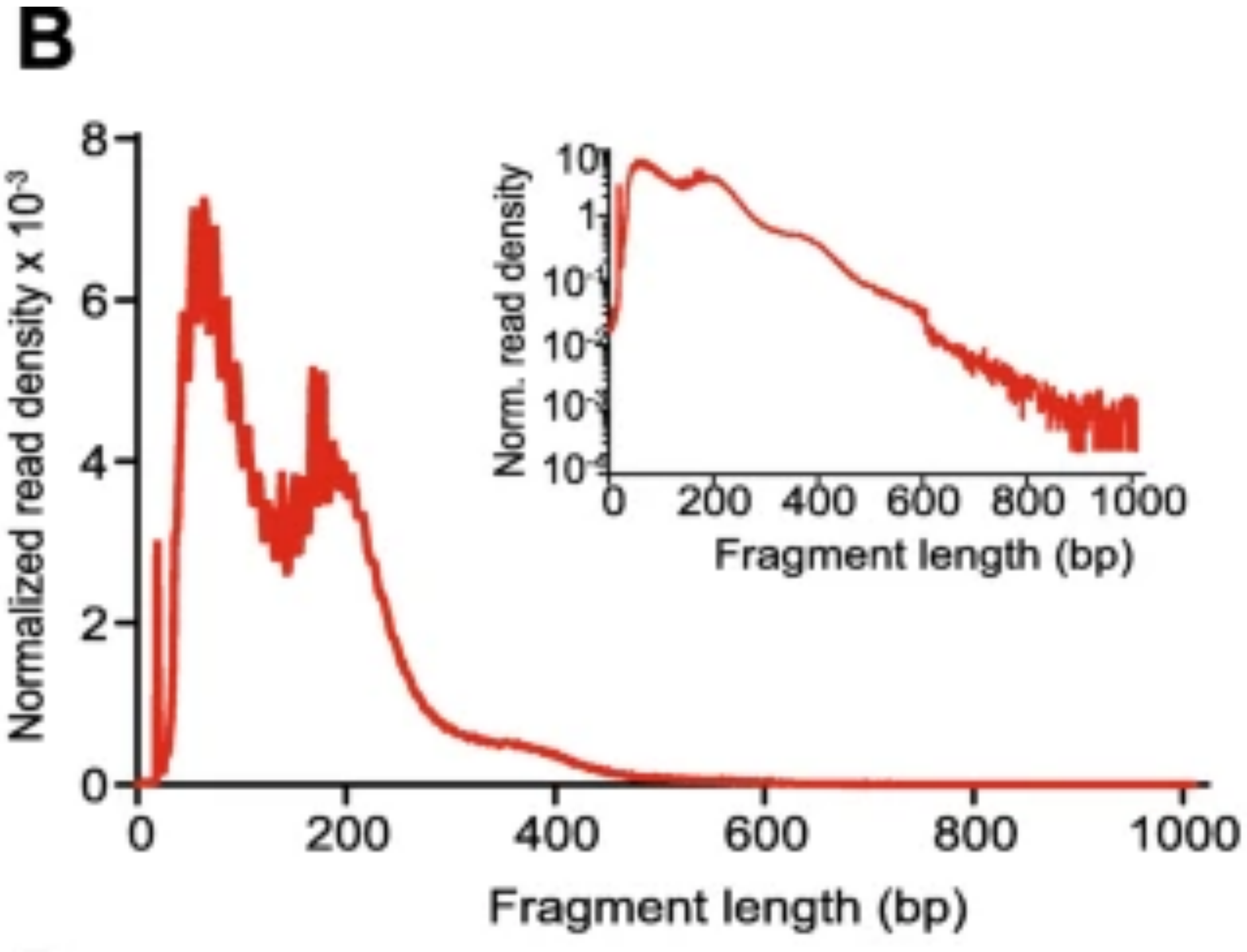
DNAse: classic assay. Requires large amounts of material. REALLY HARD to do well.

ATAC: a modern alternative to DNAse. Easy, low input requirement. Massive game-changer

ATAC assay



Fragment length density plot



Nucleosome free fragmenets, mono-nucleosome fragments, etc

ATAC data tends to have replicates; potential for differential analysis

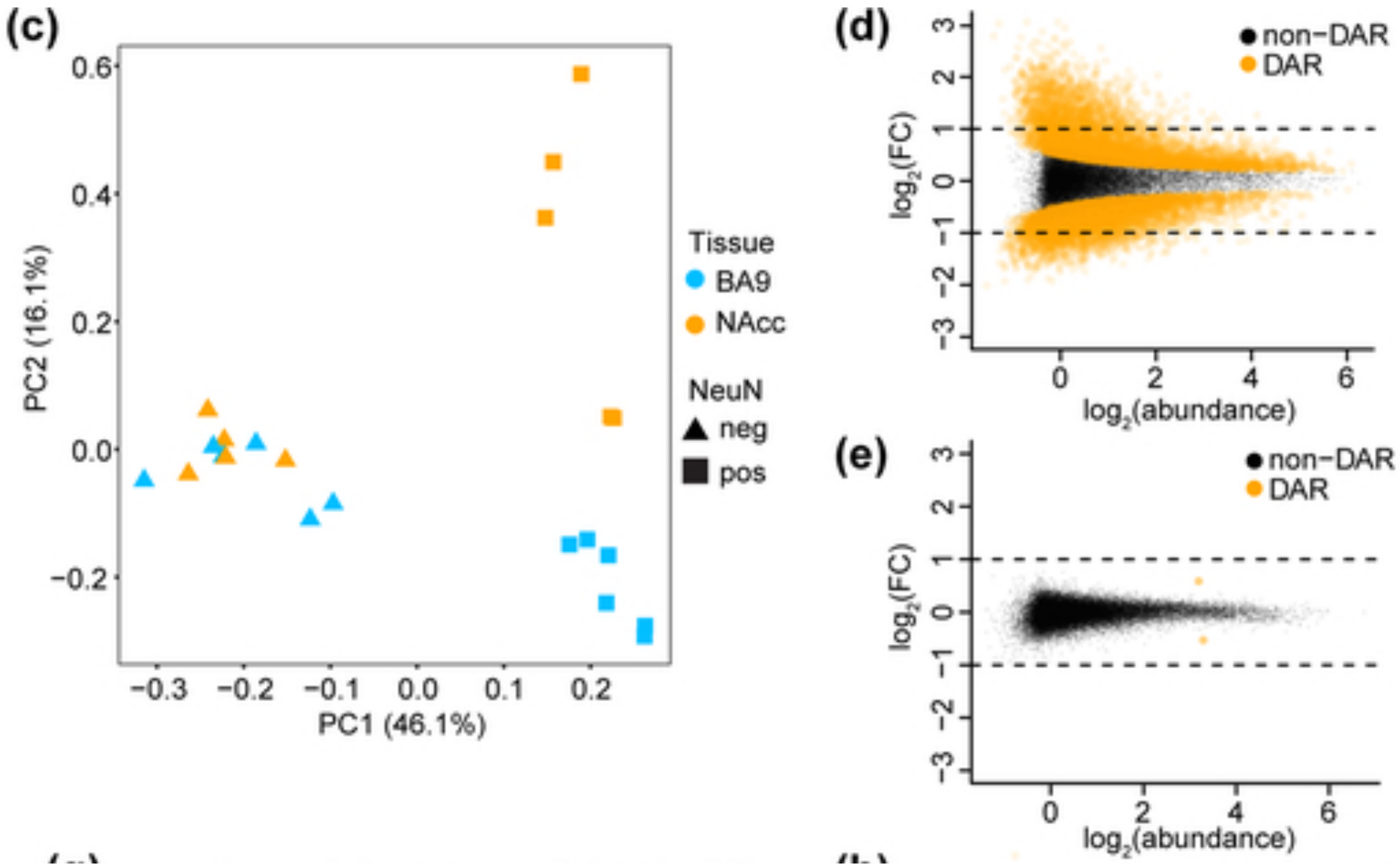
Two approaches:

1) peak calling, followed by differential analysis of peaks
2) side-step peak calling by binning the genome

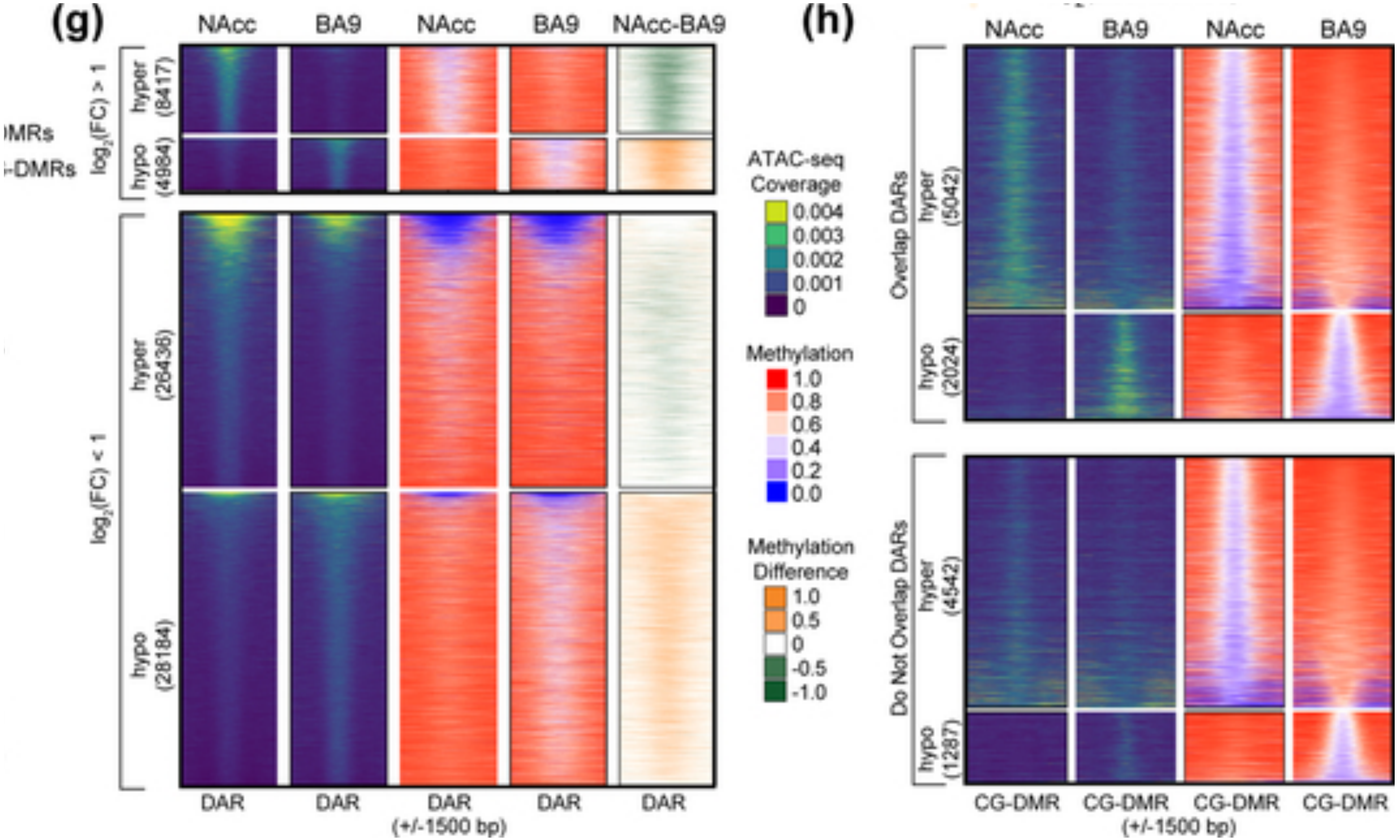
Both creates a features by samples matrix for use with differential expression tools.

Peak calling for ATAC tends to use tools for ChIP without controls.

Some data


(c)
0.6
0.4
0.2
0.0
−0.2
PC2 (16.1%)
−0.3
−0.2
−0.1
0.0
0.1
0.2
PC1 (46.1%)
Tissue
● BA9
● NAcc
NeuN
▲ neg
■ pos
(d)
● non−DAR
● DAR
3
2
1
0
−1
−2
−3
log$_2$(FC)
0
2
4
6
log$_2$(abundance)
(e)
● non−DAR
● DAR
3
2
1
0
−1
−2
−3
log$_2$(FC)
0
2
4
6
log$_2$(abundance)

Some data


(g)
NAcc
BA9
NAcc
BA9
NAcc-BA9
DMRs
-DMRs
log$_2$(FC) > 1
hyper (8417)
hypo (4984)
log$_2$(FC) < 1
hyper (26436)
hypo (28184)
DAR
DAR
DAR (+/-1500 bp)
DAR
DAR
ATAC-seq Coverage
0.004
0.003
0.002
0.001
0
Methylation
1.0
0.8
0.6
0.4
0.2
0.0
Methylation Difference
1.0
0.5
0
-0.5
-1.0
(h)
NAcc
BA9
NAcc
BA9
Overlap DARs
hyper (5042)
hypo (2024)
Do Not Overlap DARs
hyper (4542)
hypo (1287)
CG-DMR
CG-DMR
CG-DMR (+/-1500 bp)
CG-DMR